Statistica & Applicazioni - Vol. VI, n. 2, 2008, pp. 153-180

A SYMBOLIC DATA APPROACH FOR MISSING VALUES TREATMENT IN PRINCIPAL COMPONENT ANALYSIS

Paola Zuccolotto*

SUMMARY

There are two ways in order to completely perform a Principal Component Analysis over a data table with missing values: somehow imputating values to the missing data or excluding some part of the original sample from the analysis. Both these solutions can be rather costly, expecially with datasets having an appreciable number of missing values, but only one or at most two missing on any particular observational unit. An alternative proposal is formulated in this paper using the concept of Symbolic Data.

Keywords: *Principal Component Analysis, missing values, Symbolic Data Analysis, interval-valued data.*

1. INTRODUCTION

The problem of missing data affects many empirical studies, due to a wide range of possible causes.

Nearly all traditional techniques of Data Analysis cannot be performed on incomplete data tables, so that many proposals for statistical analysis with missing data are present in the literature, and in the last decades very recently developed methods have joined the traditional well-known techniques (see for example Little and Rubin, 1987, Allison, 2002, Van Buuren and Eisinga (Eds.), 2003 for some reviews). The procedures for the treatment of missing data can often be different in account of any empirical analysis, as they try to suit the specific peculiarities of the different methods of Data Analysis. Moreover in many empirical analysis, an important role in the choice of the method to use is often played by the basic assumptions about the mechanism generating missing data. A rigorous definition of the different kinds of missing data is due to Rubin (1976), and an informal understanding can be found in Allison (2002) who distinguish between *Missing Completely At Random* (MCAR) and *Missing At Random* (MAR) data and between *ignorable* and *nonignorable* missing data mechanism.

In spite of the great variety of proposals in this field, we can observe that the main part of the traditional approaches to missing data can be roughly summarized in two groups: *deletion* and *imputation*. The former consists in excluding some part

* Quantitative Methods Department - University of Brescia - c.da S. Chiara, 50 - 25122 Brescia, Italy (e-mail: zuk@eco.unibs.it).

of the original sample from the analysis, the latter in replacing missing data with a suitable value, determined on the basis of some statistical procedure. The former involves the loss of information and obviously should not be used when it dramatically reduces sample size; the latter, if the assumptions required by the selected procedure are not verified, can seriously affect the quality of the dataset. Huge problems arise expecially with datasets having an appreciable number of missing values, but only one or at most two missing on any particular observational unit: deleting incomplete units would throw away an intolerably large amount of information, but imputing values could end up by completely altering the real relations among variables.

A complete review of the several methods for missing values treatment is not possible here, nevertheless it's worth recalling that in the last decades researchers have put attention on some further approaches, overcoming the traditional duality deletion-imputation. The most popular are *maximum likelihood estimation (MLE) in presence of missing data* and *multiple imputation*, which have revealed to be promising tools in this context. The former ensures estimates with optimal properties in large samples, but has the strong limitation of requiring a model for the joint distribution of all variables with missing data. The latter, originally proposed by Rubin (1977, 1978), in the last years has become more and more popular, thanks to the improved computational power. It consists in replacing each missing value by m ($m \geq 2$) acceptable values, in order to represent the uncertainty about which value to impute. The m imputations for each missing datum create m complete datasets. Each complete dataset is analyzed using standard Data Analysis techniques, just as if the imputed data were the real ones. Finally all the answers are opportunely combined together so as to obtain a unique result. More details about this procedure are in Rubin (1987).

In this paper an original approach to missing data treatment is proposed, consisting neither in deletion nor exactly in imputation. Every missing value is replaced by an interval covering a reasonable set of possible values, eventually ranging from the minimum to the maximum value admissible for the concerned variable in that context. The resulting dataset is thus composed by single-valued and interval-valued measurements mixed, which has to be properly processed. An approach through Symbolic Data Analysis (SDA, initiated by Diday in 1987) is proposed in order to take into account the intervals corresponding to missing data.

The main purpose of SDA is to extend traditional methods to the analysis of complex data structures. In a symbolic data table the generic element x_{ij} is not necessarily a single quantitative or categorical value, but can be a distribution (histogram-type variables) as well as an interval (interval-type variables) or a set of values linked by some logical rule. Since the definition of the concept of Symbolic Data Analysis, a wide literature has been produced, covering several fields of Data Analysis (see Bock and Diday, 2000, for a review). The definition of dissimilarity measures (Gowda and Diday 1991, Ichino and Yaguchi 1994, De Carvalho 1994, 1996) the development of algorithms for Cluster Analysis (Chavent, 1998) and for tree-structured classifiers (Bravo and García-Santesmases, 1998), the extension to symbolic objects of Principal Component Analysis (Cazes, Chouakria, Diday and Schektman, 1997) and Factorial Discriminant Analysis (Lauro, Verde and Palumbo, 2000) are only few examples.

An alternative approach for the treatment of interval-valued data could refer to the framework of fuzzy data, where additional information can be included through a membership function assigning a weight to each value in the interval. The main references on PCA for fuzzy data include Amato and Palumbo (2004), Giordani and Kiers (2004b), Lauro and Palumbo (2004), D'Urso and Giordani (2005). These techniques are not the object of the present paper, and could be considered for further research and results comparison.

Although the proposed idea of replacing missing values with intervals is quite general, in this paper it is specifically analyzed in the framework of Principal Component Analysis. In Section 2 the basic features of Symbolic Principal Components Analysis are recalled. Section 3 describes the proposed procedure for the treatment of missing values, examining its main qualities with reference to an illustrative example. In Section 4 the problem of the construction of synthetic measures based on the Principal Components is handled in this context and Section 5 shows an application to real data. Section 6 concludes.

2. SYMBOLIC PRINCIPAL COMPONENTS ANALYSIS

With the Symbolic Principal Component Analysis (SPCA) classical dimensionality reduction methods are extended to complex data. Many studies have been centered on this problem, explicitly formalized by Godwa, Diday and Nagabhushan (1995), and then further investigated from different points of view. Cazes, Chouakria, Diday and Schektman (1997) propose two popular methods for SPCA on interval data, called *Vertices Method* and *Centers Method*, later refined by Lauro and Palumbo (2000). The SPCA for histogram data has been proposed by Rodríguez, Diday and Winsberg (2000), with an algorithm also working if the data table has variables of interval-type and histogram-type mixed. Giordani and Kiers (2004a) introduce a method for three-way component analysis of interval data.

In this paper attention is focused on SPCA of interval data, that is on symbolic data tables ${}_S\mathbf{X}$ of the following type:

$$ {}_S\mathbf{X} = \begin{bmatrix} \left[x_{11}^-, x_{11}^+\right] & \cdots & \left[x_{1p}^-, x_{1p}^+\right] \\ \vdots & \ddots & \vdots \\ \left[x_{N1}^-, x_{N1}^+\right] & \cdots & \left[x_{Np}^-, x_{Np}^+\right] \end{bmatrix}. \tag{1} $$

From a geometrical point of view, the difference with the traditional approach is that each subject can be represented by a hyperrectangle with 2^p vertices in the space $\mathbb{R}^p$, instead of by a single point. It follows that the orthogonally projected representations of the subjects are not points in a low-dimensional space, but have themselves a complex shape.

In the next two paragraphs the basic methods for SPCA of interval-valued data, proposed by Cazes, Chouakria, Diday and Schektman (1997) are briefly recalled.

2.1 *Vertices Method*

The Vertices Method SPCA (V-SPCA) performs a classical PCA on the numerical data table $\mathbf{X}_V$ obtained stacking below each other the N matrices $\mathbf{X}_{V_i}$ as follows

$$\mathbf{X}_V = \begin{bmatrix} \mathbf{X}_{V_1} \\ \mathbf{X}_{V_2} \\ \vdots \\ \mathbf{X}_{V_N} \end{bmatrix}$$

where for a generic subject i, $\mathbf{X}_{V_i}$ is a $(2^p \times p)$ matrix containing the 2^p vertices of its hyperrectangle:

$$\mathbf{X}_{V_i} = \begin{bmatrix} x_{i1}^- & x_{i2}^- & \cdots & x_{i(p-1)}^- & x_{ip}^- \\ x_{i1}^- & x_{i2}^- & \cdots & x_{i(p-1)}^- & x_{ip}^+ \\ x_{i1}^- & x_{i2}^- & \cdots & x_{i(p-1)}^+ & x_{ip}^- \\ x_{i1}^- & x_{i2}^- & \cdots & x_{i(p-1)}^+ & x_{ip}^+ \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{i1}^- & x_{i2}^+ & \cdots & x_{i(p-1)}^+ & x_{ip}^+ \\ x_{i1}^+ & x_{i2}^+ & \cdots & x_{i(p-1)}^+ & x_{ip}^+ \end{bmatrix}.$$

Each subject is then represented in a lower-dimensional space by orthogonally projecting all the vertices of its hyperrectangle. When the first two Principal Components (PCs) are chosen, for example, the 2^p points concerning each subject are projected in a plane and that subject is represented by the resulting two-dimensional scattering, which unfortunately has an irregular shape, whose *convex cover* can be defined only by a complex nonlinear function of the PCs.

In the literature the problem is solved by considering on each axis the segment including all the projections and representing the subject as the hyperrectangle (rectangle in the two-dimensional case) built from these segments, as shown in figure 1.

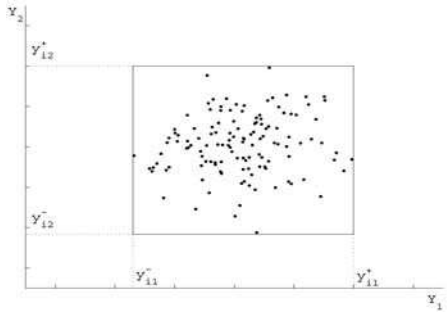

FIGURE 1. - *Projection of the 128 vertices of a seven-dimensional hyperrectangle in the first two PCs' space and two-dimensional resulting rectangle*

Hence the q-dimensional ($q < p$) hyperrectangle corresponding to i-th subject in the first q PCs' space has vertices given by the rows of the $(2^q \times q)$ matrix $\mathbf{Y}_{V_i}$

$$\mathbf{Y}_{V_i} = \begin{bmatrix} y_{i1}^- & y_{i2}^- & \cdots & y_{i(q-1)}^- & y_{iq}^- \\ y_{i1}^- & y_{i2}^- & \cdots & y_{i(q-1)}^- & y_{iq}^+ \\ y_{i1}^- & y_{i2}^- & \cdots & y_{i(q-1)}^+ & y_{iq}^- \\ y_{i1}^- & y_{i2}^- & \cdots & y_{i(q-1)}^+ & y_{iq}^+ \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ y_{i1}^- & y_{i2}^+ & \cdots & y_{i(q-1)}^+ & y_{iq}^+ \\ y_{i1}^+ & y_{i2}^+ & \cdots & y_{i(q-1)}^+ & y_{iq}^+ \end{bmatrix}$$

where the component scores of the vertices in the j-th PC are given respectively by the minimum and the maximum of the component scores y_{ijl} of all the 2^p projected vertices ($l \in L = \{1, 2, \cdots, 2^p\}$):

$$y_{ij}^- = \min_{l \in L} \{y_{ijl}\}$$

$$y_{ij}^+ = \max_{l \in L} \{y_{ijl}\}.$$

The most frequent representation is, as in classical PCA, two-dimensional, so that each subject is denoted by the so called *maximum covering area rectangle* that is necessarily oversized with respect to the real hyperrectangle in $\mathbb{R}^p$. This drawback is in part due to the fact that V-SPCA treats the vertices as simple (independent) points losing any relationship among vertices belonging to the same subject. To overcome this shortcoming Lauro and Palumbo (2000) introduce an approach taking into account the vertices cohesion constraint, by maximizing the variance among subjects instead of the total vertices variance. As will be clear in the next section, the maximum covering area representation can be particularly unpleasant in the context of the strategy for the treatment of missing data which will be introduced later. The more recently proposed representation of the projected vertices via convex hulls, surely interesting in the general perspective of Symbolic Principal Component Analysis, will be especially advantageous in this framework. This alternative representation is illustrated in figure 2.

The convex hull of an arbitrary set is the smallest convex set containing it and, if the original set consists of a finite number of points, the convex hulls is simply a closed convex polygon whose vertices are those points around the periphery of the configuration. An intuitive definition in the two-dimensional case is obtained envisaging the data points as pins in a board: a large elastic band is looped around the pins and released. The band will come to rest forming a polygon: the pins it touches are the extremes (Green, 1981). The computation of convex hulls presents no problems both in the bivariate and in the multivariate case and efficient algorithms are available (see for example Green and Silverman (1979) and Barber, Dobkin and Huhdanpaa (1996)).

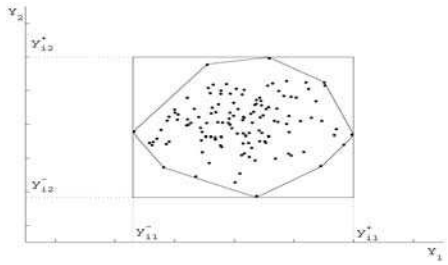

FIGURE 2. - *Projection of the 128 vertices of a seven-dimensional hyperrectangle in the first two PCs' space and two-dimensional resulting convex hull compared with the maximum covering area rectangle*

2.2 *Centers Method*

With V-SPCA the dimension of the matrix $\mathbf{X}_V$ rapidly increases with the number p of variables. To avoid the treatment of large-sized matrices, the Centers Method SPCA (C-SPCA) is based on a $(N \times p)$ matrix $\mathbf{X}_C$, computed from the symbolic data table (1) simply substituting to each interval its midpoint:

$$\mathbf{X}_C = \begin{bmatrix} x_{11}^c & \cdots & x_{1p}^c \\ \vdots & \ddots & \vdots \\ x_{N1}^c & \cdots & x_{Np}^c \end{bmatrix}$$

where

$$x_{ij}^c = \frac{x_{ij}^- + x_{ij}^+}{2} \qquad \forall i,j.$$

The matrix $\mathbf{X}_C$ contains the coordinates of the centers of the N hyperrectangles. Classical PCA is performed on $\mathbf{X}_C$, then all the vertices of each hyperrectangle are projected in the obtained subspace and the lower-dimensional hyperrectangle (rectangle when only the first two PCs are extracted) is again constructed with the segments covering all the projections.

C-SPCA is based upon the assumption that the hyperrectangles should be well represented by their centers, and then the subspace obtained optimizing the projection of the centers should not be too bad for the hyperrectangles either.

3. THE TREATMENT OF MISSING VALUES THROUGH SYMBOLIC PRINCIPAL COMPONENT ANALYSIS

Traditional PCA could be viewed as a particular case of SPCA on interval data. In fact a single valued measurement is to all intents a degenerate interval, with coincident extremes and, from a geometrical point of view, a subject described by p quan-

titative single valued data is a degenerate hyperrectangle, in this case a point, in $\mathbb{R}^p$. When a SPCA is performed on such ''interval data'', a traditional PCA is effectively carried out. This observation justifies the treatment with SPCA of data tables with interval-valued and single-valued measurements mixed.

This paper proposal is to handle missing data with an approach which could be called *Interval Imputation* (InI) and allows to avoid both deletion and imputation. It consists in the replacement of missing data for the variable X_j with a suitable interval. In the worst case of absence of any information about the possible extention of the interval, it could range from the minimum to the maximum value admissible for variable X_j in the examined context, that is $[x_{j,\min}, x_{j,\max}]$. When these two values are not precisely identified, or the obtained interval is judged too wide, $\mu_j - 3\sigma_j$ and $\mu_j + 3\sigma_j$ (where μ_j and σ_j are respectively the mean and the standard deviation of variable X_j, computed on the available data) can reasonably be used, provided they are both admissible values[1]. The following analyses will be carried out on the standardized variables, so that the interval $[-3; +3]$ can replace all missing values since information allowing narrower intervals is not available.

Let $\mathbf{Z}$ be the $(N \times p)$ standardized data matrix, where some elements are missing

$$\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1(p-1)} & z_{1p} \\ z_{21} & \bullet & \cdots & z_{2(p-1)} & z_{2p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \bullet & z_{N2} & \cdots & z_{N(p-1)} & \bullet \end{bmatrix}$$

the corresponding symbolic data table obtained with InI is given by

$${}_S\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1(p-1)} & z_{1p} \\ z_{21} & [z_{2,\min}, z_{2,\max}] & \cdots & z_{2(p-1)} & z_{2p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ [z_{1,\min}, z_{1,\max}] & z_{N2} & \cdots & z_{N(p-1)} & [z_{p,\min}, z_{p,\max}] \end{bmatrix}$$

or, if extreme values are not precisely defined,

$${}_S\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1(p-1)} & z_{1p} \\ z_{21} & [-3, +3] & \cdots & z_{2(p-1)} & z_{2p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ [-3, +3] & z_{N2} & \cdots & z_{N(p-1)} & [-3, +3] \end{bmatrix}.$$

From a geometrical point of view, a unit with single-valued and interval-valued data mixed is a m-dimensional hyperrectangle in $\mathbb{R}^p$ ($m \leq p$): a point, a segment, a rectangle, a cube in $\mathbb{R}^p$, if zero, one, two, three values respectively are missing, and so on.

[1] This choice does not necessarily imply a Gaussian distributional assumption, as the use of a symmetrical interval around the mean, proportional to the standard deviation is justified by the Chebyshev's inequality.

A SPCA can be performed on the symbolic data table ${}_S\mathbf{Z}$. The resulting two-dimensional scatterplot is then composed by units represented by points and rectangles mixed, where the presence of any missing value results in the location of the concerned unit in a variable position inside a given area. The shape and dimension of this area is affected by:

- the number of missing values for the concerned unit,
- the importance of the missing information in the factorial representation.

3.1 *An illustrative example*

An example can better explain how InI works. From a data table $\mathbf{X}$ composed by $N = 40$ units and $p = 5$ variables, some values have been selected and artificially deleted from 5 units, according to different rules, so that three new data tables have been generated:

- the matrix $\mathbf{X}_1$ where each of the 5 units has one missing value, corresponding to the 5 variables

$$\mathbf{X}_1 = \begin{bmatrix} \bullet & x_{12} & x_{13} & x_{14} & x_{15} \\ x_{21} & \bullet & x_{23} & x_{24} & x_{25} \\ x_{31} & x_{32} & \bullet & x_{34} & x_{35} \\ x_{41} & x_{42} & x_{43} & \bullet & x_{45} \\ x_{51} & x_{52} & x_{53} & x_{54} & \bullet \\ x_{61} & x_{62} & x_{63} & x_{64} & x_{65} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{401} & x_{402} & x_{403} & x_{404} & x_{405} \end{bmatrix}$$

- the matrix $\mathbf{X}_2$ where each of the 5 units has two missing values, corresponding to the following pairs of variables: (X_1, X_2), (X_1, X_3), (X_1, X_5), (X_3, X_4), (X_3, X_5)

$$\mathbf{X}_2 = \begin{bmatrix} \bullet & \bullet & x_{13} & x_{14} & x_{15} \\ \bullet & x_{22} & \bullet & x_{24} & x_{25} \\ \bullet & x_{32} & x_{33} & x_{34} & \bullet \\ x_{41} & x_{42} & \bullet & \bullet & x_{45} \\ x_{51} & x_{52} & \bullet & x_{54} & \bullet \\ x_{61} & x_{62} & x_{63} & x_{64} & x_{65} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{401} & x_{402} & x_{403} & x_{404} & x_{405} \end{bmatrix}$$

- the matrix $\mathbf{X}_3$ where each of the 5 units has three or more missing values, corresponding to the following set of variables: (X_1, X_2, X_5), (X_1, X_3, X_4), (X_1, X_3, X_5), (X_1, X_2, X_4, X_5), $(X_1, X_2, X_3, X_4, X_5)$

$$\mathbf{X}_3 = \begin{bmatrix} \bullet & \bullet & x_{13} & x_{14} & \bullet \\ \bullet & x_{22} & \bullet & \bullet & x_{25} \\ \bullet & x_{32} & \bullet & x_{34} & \bullet \\ \bullet & \bullet & x_{43} & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ x_{61} & x_{62} & x_{63} & x_{64} & x_{65} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{401} & x_{402} & x_{403} & x_{404} & x_{405} \end{bmatrix}$$

The standardized data tables $\mathbf{Z}, \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3$ are obtained respectively from $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$. For each data table the standardization is done using the available data. Firstly a PCA has been carried out on the original complete standardized data table $\mathbf{Z}$ in order to get a term of comparison for the results which will be obtained with the different missing values treatment techniques. Secondly two traditional analyses have been carried out on the matrices $\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3$ respectively with listwise and pairwise deletion of the units with missing values. It is useful to recall that deletion is called listwise when it consists simply in the exclusion from the analysis of the units containing any missing value, and pairwise when summary statistics on a given variable are computed using all the cases available for that variate. For example, in the computation of a variance-covariance matrix, means and variances of each variable can be obtained from all values present in that variate and similarly covariances between pairs of variables are computed from all units that have observed value for that pair of variates (it should be noted that in the case of Principal Component Analysis this procedure does not completely solve the problem, as it can help only in calculating and interpreting eigenvalues and eigenvectors, but imputed values are still necessary for obtaining component scores for all the subjects). Both listwise and pairwise deletion have advantages and disadvantages. In particular, listwise deletion usually excludes a larger fraction of the original sample, but it has also appealing features, like the possibility to be used for any kind of statistical analysis, as no special computation methods are required, and moreover it has some attractive statistical properties (see for example Allison, 2002). Pairwise deletion operates a more efficient use of available data, and in some cases more efficient estimates, but some problems could arise. For example calculating the variance-covariance matrix element by element, as described above, destroys some of the normal features of that matrix and can occasionally lead to odd results like, for example, correlations outside the range -1 to $+1$. Devlin, Gnanadesikan and Kettenring (1981) suggest some ways to handle these drawbacks, but anyway extreme caution should be exercised. Moreover the estimated standard errors and test statistics produced by conventional software can be seriously biased, because formulas to obtain consistent estimates (Van Praag, Dijkstra and Van Velzen, 1985) are complex and by now not implemented in any commercial package.

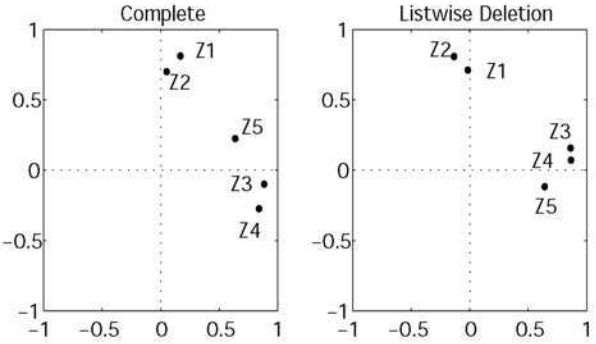

FIGURE 3. - *Two-dimensional factor loadings plots of PCA on the data table without missing (left) and with listwise deletion (right)*

The two-dimensional factor loadings plots of the traditional PCA performed on the complete data table and on the data table obtained deleting the units with missing values are represented in figure 3. The variance accounted for by the first two PCs is respectively 64.27% and 62.55% in the two analyses.

Figure 4 displays the factor loadings plots of PCA performed on the correlation matrix of $\mathbf{Z}_1$, $\mathbf{Z}_2$ and $\mathbf{Z}_3$ with a pairwise deletion implying that the correlation computations involve all cases that have observed values for both variables. As highlighted before, this method provides only eigenvalues and factor loadings, but does not allow the representation of subjects with any missing in the factorial space. The variance accounted for by the first two PCs is respectively 62.04%, 61.22%, 62.42% in the three analyses.

Finally InI is performed using V-SPCA. No information is assumed about minimum and maximum of the variables, so that the corresponding symbolic data tables ${}_S\mathbf{Z}_1$, ${}_S\mathbf{Z}_2$, ${}_S\mathbf{Z}_3$ are obtained replacing the missing data with the interval $[-3, +3]$.

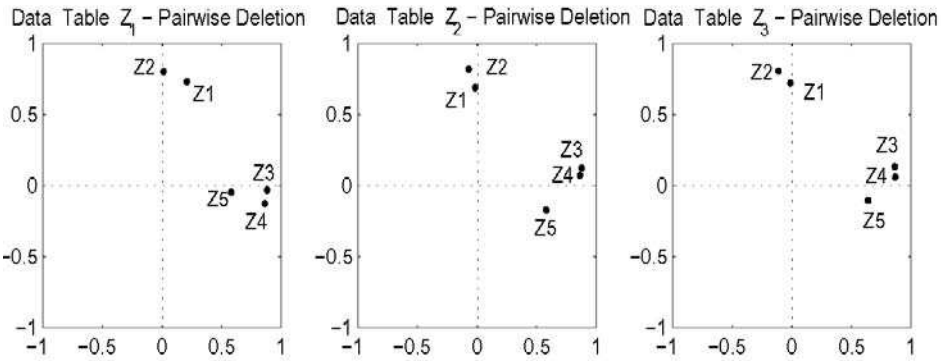

FIGURE 4. - *Two-dimensional factor loadings plots of PCA on* $\mathbf{Z}_1$ *(left),* $\mathbf{Z}_2$ *(middle),* $\mathbf{Z}_3$ *(right) with pairwise deletion*

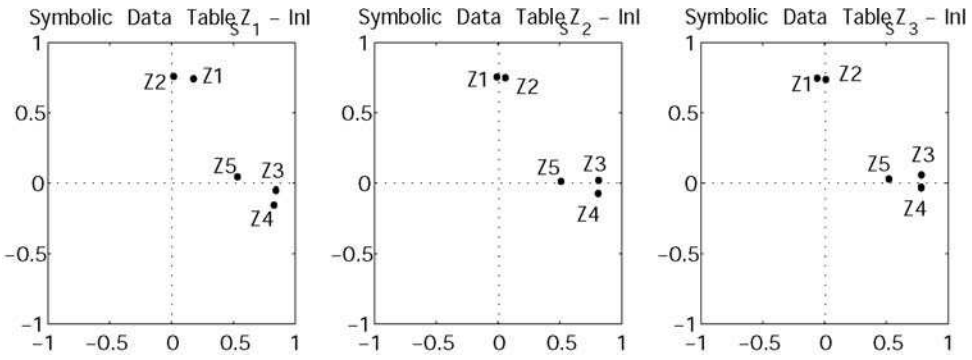

FIGURE 5. - *Two-dimensional factor loadings plots of V-SPCA on* ${}_S\mathbf{Z}_1$ *(left),* ${}_S\mathbf{Z}_2$ *(middle),* ${}_S\mathbf{Z}_3$ *(right)*

The two-dimensional factor loadings plots of V-SPCA performed on the matrices ${}_S\mathbf{Z}_1$, ${}_S\mathbf{Z}_2$ and ${}_S\mathbf{Z}_3$ are represented in figure 5. The variance accounted for by the first two PCs is respectively 57.44%, 54.48%, 52.07% in the three analyses.

The eight factor loadings plots of figures 3, 4 and 5 are quite similar each other, and this means that the two-dimensional factorial solution is substantially correct in all the cases. Nevertheless it is apparent that the presence of missing data destroys some details of the original structure present in the complete data table $\mathbf{Z}$, which cannot be recovered, whatever the method used for the treatment of missing.

The most interesting results of the use of SPCA are

- the possibility to represent *all* the subjects in the factorial space and
- the peculiar representation in the factorial space, taking account of the incomplete information about subjects with missing values.

At a first step the traditional representation via the maximum covering area rectangle is used (figure 6) even if, as previously stated, it can be somehow misleading in this context.

At this point the reason for inadequateness of the maximum covering area rectangles is easy to understand: with the described approach, a given subject i, having m missing values among the p quantitative variables under study, can be viewed as a m-dimensional hyperrectangle in a p-dimensional space. If $m = 0$ or $m = 1$ it is a degenerate hyperrectangle, respectively a point or a segment in $\mathbb{R}^p$.

If the case $m = 0$ is treated in a perfectly coherent manner (the point in $\mathbb{R}^p$ projects onto a point in $\mathbb{R}^2$), the same can't be said for the case $m = 1$, which implies that a segment in $\mathbb{R}^p$ is projected onto a rectangle in $\mathbb{R}^2$.

It follows that if, as very frequently happens, a subject has only one missing value, its two-dimensional representation paradoxically increases its real dimension, in spite of the dimensional reduction drawn by the SPCA.

In general, this problem arises for a given subject when the dimension reduction

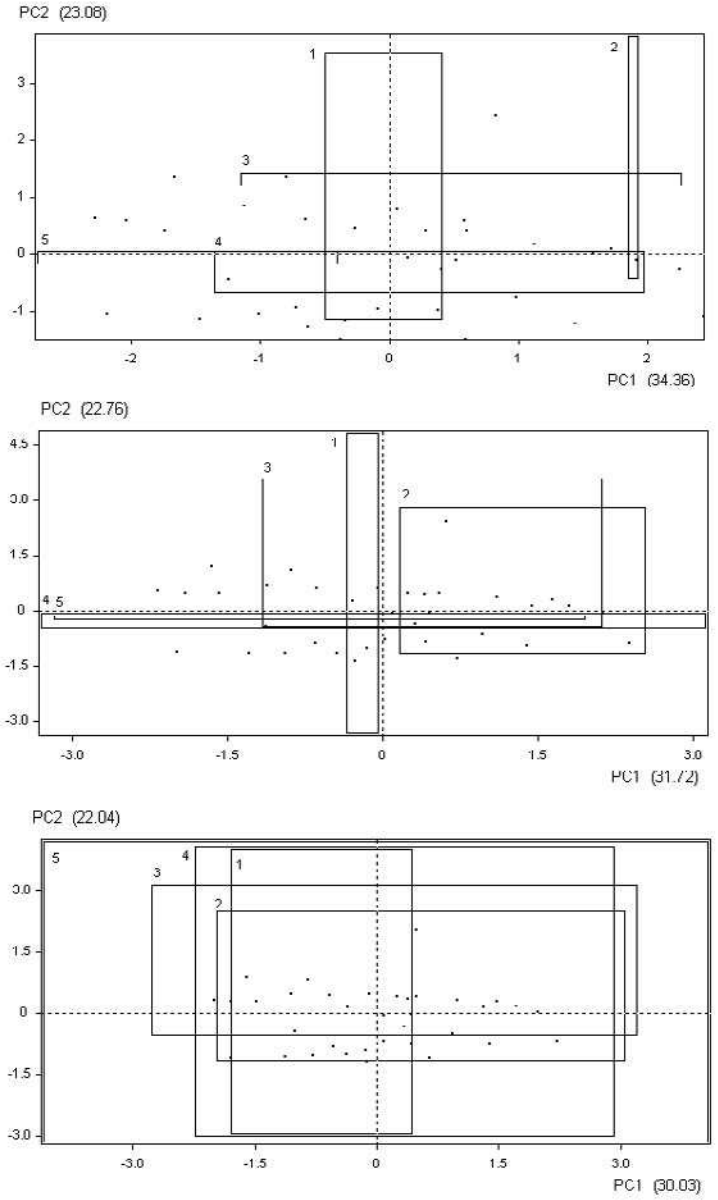

FIGURE 6. - *Rectangular representation of the subjects in the factorial plane, V-SPCA on* ${}_S\mathbf{Z}_1$ *(top),* ${}_S\mathbf{Z}_2$ *(middle),* ${}_S\mathbf{Z}_3$ *(bottom)*

operated through the SPCA projects the hyperrectangles in a space of dimension q, greater than the number m of missing values of the subject ($q > m$). Moreover, even when $q \leq m$, in this context the number of vertices of the hyperrectangles tends to be relatively small, because in most cases each subject has (hopefully) a small

amount of missing. With so few vertices to be projected, the lower-dimensional scattering could be poorly described by the maximum covering area rectangle.

The convex hull of a set of two points (the case of a subject with only one missing, whose degenerate hyperrectangle has only two vertices) is the segment linking the two points themselves and hence a segment in $\mathbb{R}^p$ reasonably projects onto a segment in $\mathbb{R}^2$.

The gain due to the proposed alternative representation can be appreciated in the proposed example, where the two-dimensional convex hulls allow to describe the subjects with irregular, but more fitting shapes, especially in the case of subjects with only one missing (figure 7).

From the inspection of the subjects representation in the factorial space (both in figures 6 and 7), three main remarks can be emphasized:

- subjects with no missing are represented by points, while subjects with any missing are located in an uncertain position inside a give area, delimited by the rectangle/hull;
- a larger number of missing values tends to determine a larger rectangle/hull (it's interesting to observe the fifth unit of matrix $\mathbf{Z}_3$, the borderline case of a subject completely unknown, whose rectangle/hull turns out to cover the entire graphs);
- the shape of the rectangle/hull differs according to the variables where the missing is located. For example when the missing information regards a variable with high score on the first PC, but low score on the second, the resulting rectangle/hull has a large base and a narrow height.

The former feature ensures a reasonable correspondence between number of missing data of a given subject and uncertainty of its location in the factorial space; the latter is probably the most interesting, as it allows for a different treatment of missing information because the variability introduced due to the missing data affects only the PCs where the corresponding variable has high score, leaving the others nearly unaltered. For example subjects 4 and 5 in the data table $\mathbf{Z}_2$ have missing values on two variables with high score on the first PC. This implies a very uncertain location along the first dimension in the factorial space, while along the second dimension their locations exhibit a very little variability.

3.2 *A simulative study about the example*

In order to better understand the performance of the proposed method compared with listwise and pairwise deletion, a simulative study is carried out in the framework of the above described illustrative example. Given the (40×5) data table $\mathbf{X}$, two matrices $\mathbf{X}_{10}$ and $\mathbf{X}_{20}$ are generated randomly positioning in the table respectively $m = 20$ and $m = 40$ missing values (corresponding to the 10% and the 20% of the entire dataset). PCA with listwise deletion, pairwise deletion and InI is carried out on the standardized matrices $\mathbf{Z}_{10}$ and $\mathbf{Z}_{20}$. The procedure is repeated $S = 1000$ times.

Figure 8 shows the boxplots of the variance accounted for by the first two PCs in the various analyses, figures 9 and 10 display the boxplots of the loadings of the 1st

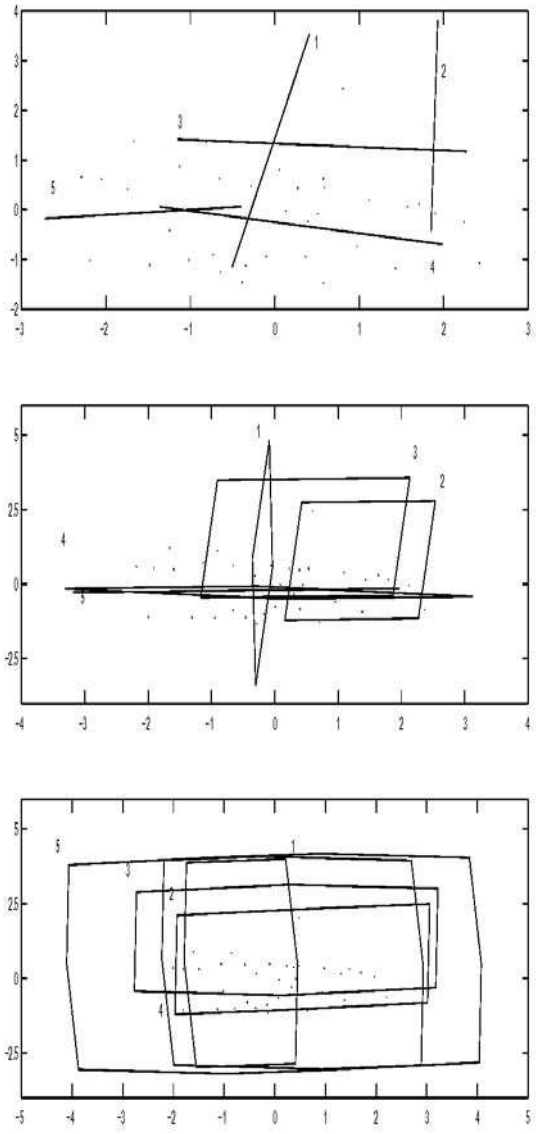

FIGURE 7. - *Convex hulls representation of the subjects in the factorial plane, V-SPCA on* ${}_S\mathbf{Z}_1$ *(top),* ${}_S\mathbf{Z}_2$ *(middle),* ${}_S\mathbf{Z}_3$ *(bottom)*

and 2nd PC for the 5 variables. Finally, figures 11 and 12 represent the boxplots of the correlation of the first two PCs obtained with the different methods[2] and the cor-

[2] In the case of InI, the mean value of the 2^p projected vertices corresponding to each subject is considered.

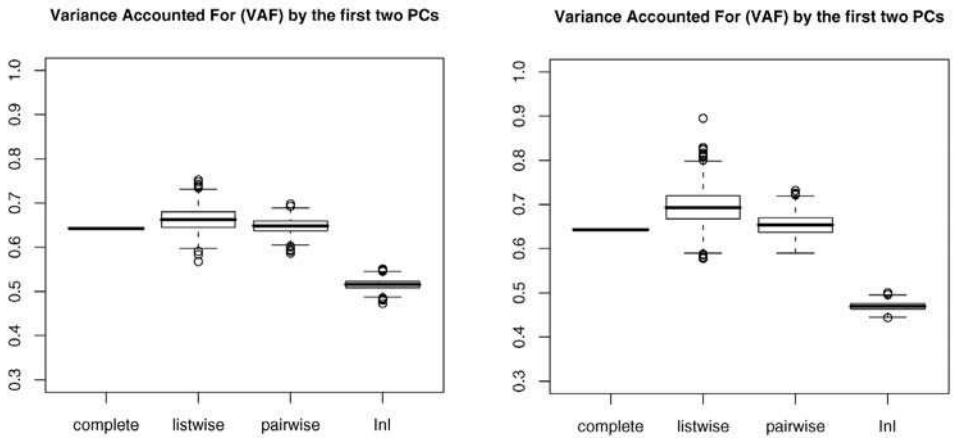

FIGURE 8. - *VAF by the first two PCs - Analysis with 10% (left) and 20% (right) missing values*

responding PCs computed with the original complete data table. Some important observations can be made:

- The VAF by the first two PCs tends to be systematically lower with InI, intuitively due to the variability introduced by the intervals replacing missing data;
- from the point of view of the loadings and of the correlation with the ''true'' PCs, InI well compares with the other methods, often with a better performance than listwise deletion, expecially whit a high number of missing values. This means that the introduction of the intervals does not appreciably affect the global factorial representation.

This simulation study is far from complete: many different data tables (i.e. with different correlation structures in data, different sizes, ...) should be analyzed. Anyway it can give a general idea of how InI performs under different points of view.

4. SYNTHETIC MEASURES BASED ON PRINCIPAL COMPONENTS

A common purpose of PCA is to determine indicators of the features under study, in order to assign to each subject one or more synthetic measures of the extracted characteristics. For example in many analyses the 1st PC is used as an overall synthesis of the phenomenon under study, and the value assumed by a given subject is an index revealing how intense the whole analyzed phenomenon is present on that subject. When the overall explanation cannot be unidimensional, or when a rotation is performed, the 1st and the 2nd PCs are interpretable as two main features, so that each subject can be described by a couple of synthetic measures, referring to different uncorrelated aspects of the same phenomenon.

When missing data are treated with the proposed method, some difficulties arise in the determination of these synthetic measures.

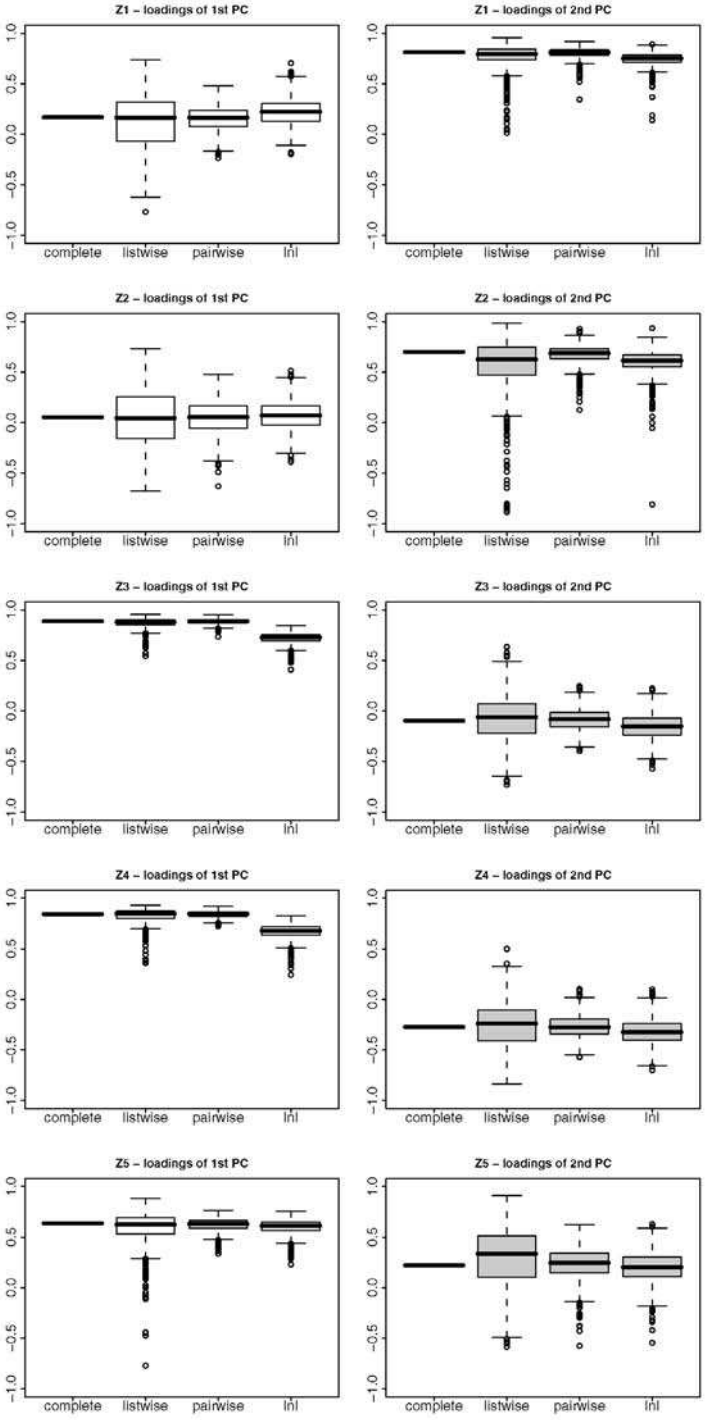

FIGURE 9. - *Loadings - Analysis with 10% missing values*

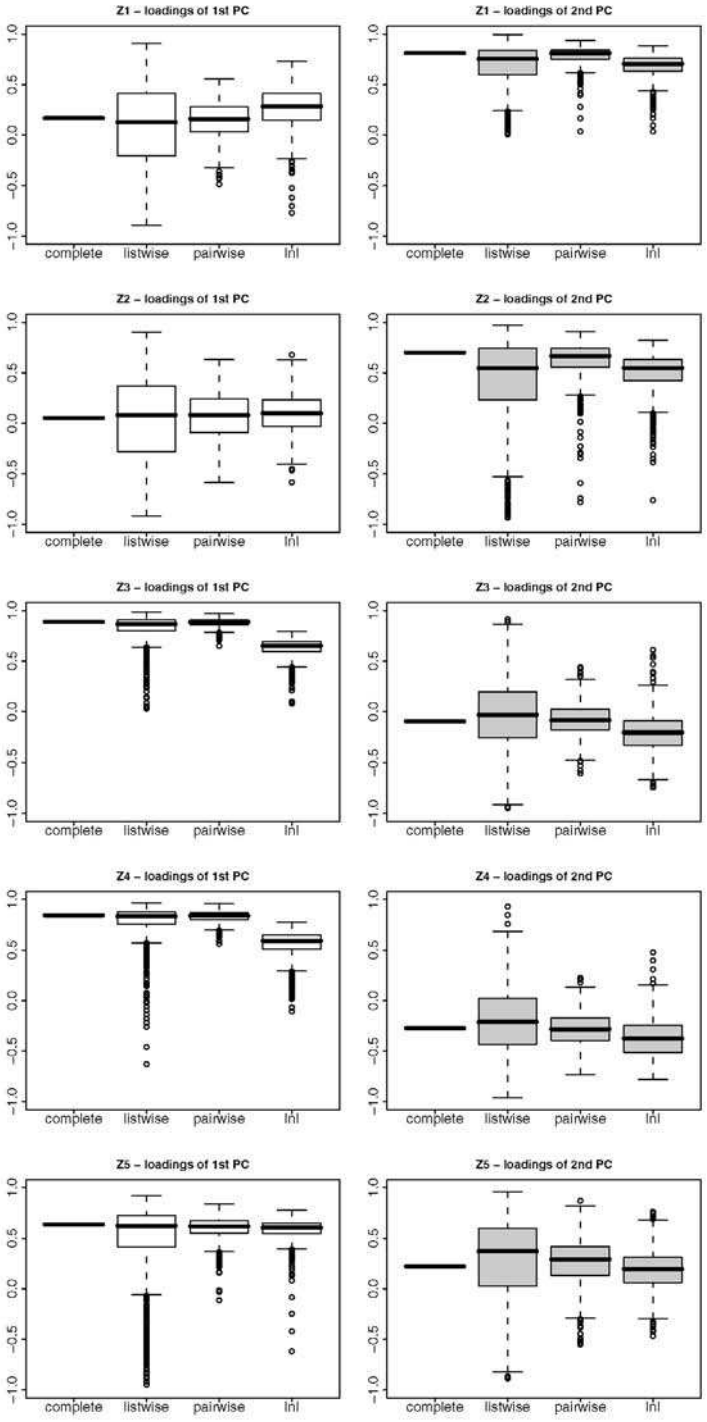

FIGURE 10. - *Loadings - Analysis with 20% missing values*

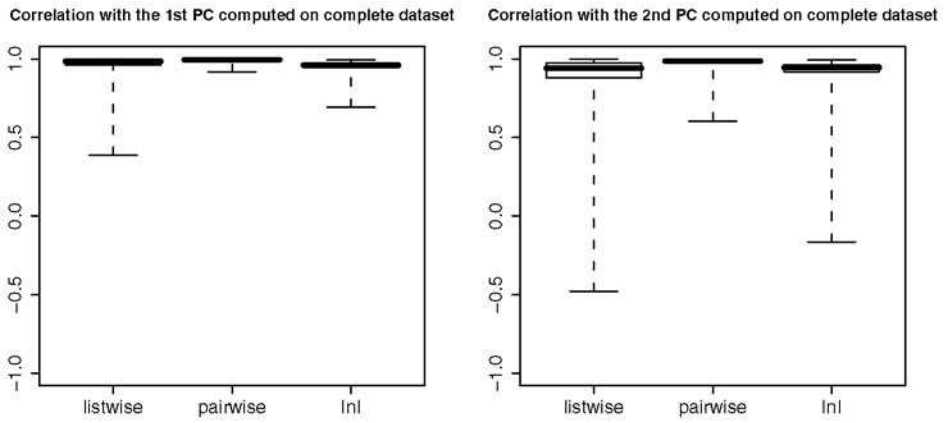

FIGURE 11. - *Correlations with the first two PCs (plot whiskers ranging from minimum to maximum) - Analysis with 10% missing values*

In fact subjects with missing data are described by an interval instead of by a single value. When the attention is focused on the single subjects, an interval-valued description can generally be maintained. If a single reference value should eventually be necessary (e.g. for ranking purposes) the problem could be solved by considering the midpoint of the interval or a mean value computed averaging the projections of all the 2^p vertices of the hyperrectangle. Hence the index assigned to subject i with reference to j-th PC could be alternatively given by

$$y_{ij} = \frac{y_{ij}^{-} + y_{ij}^{+}}{2} \tag{2}$$

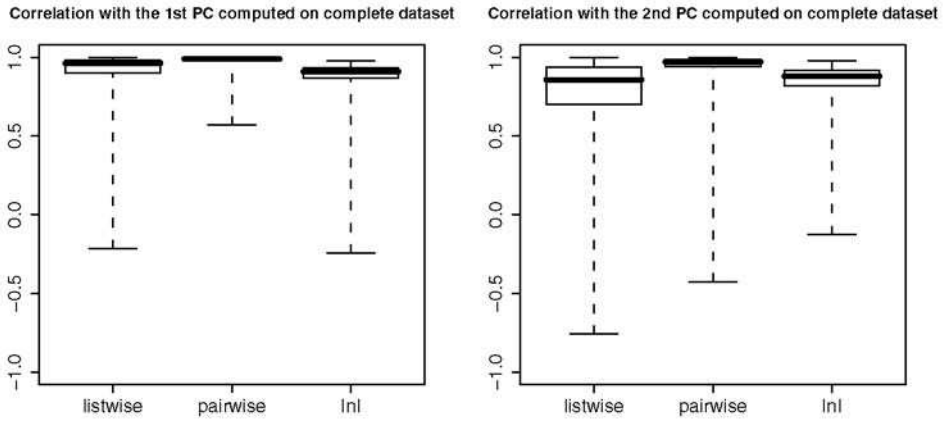

FIGURE 12. - *Correlations with the first two PCs (plot whiskers ranging from minimum to maximum) - Analysis with 20% missing values*

or by

$$y_{ij} = \frac{1}{2^p} \sum_{l=1}^{2^p} y_{ijl} \tag{3}$$

where y_{ijl} is the projection of l-th vertices of i-th hyperrectangle on j-th PC.

The problem is more tricky when the subjects are divided into categories according to a qualitative variable H and a mean index computed within the groups is desired. Averaging the reference values (2) or (3), completely neglecting the intervals, can't be an acceptable solution because the synthetic measure should take into account the uncertainty due to the different presence of subjects affected by missing values in the groups. Two alternatives are possible:

- a weight $\omega \in [0, 1]$ can be assigned to the subjects, in order to take into account the eventual variability of its description with reference to the PC used to evaluate the synthetic measure. The weight could be determined through a strictly decreasing monotone function f of the amplitude of the interval describing the subject along the j-th PC. For example f can be a negative exponential:

$$\omega_{ij} = f(y_{ij}^+ - y_{ij}^-) = e^{-(y_{ij}^+ - y_{ij}^-)}$$

  so that subjects with missing data give a less important contribution in the computation of the mean and, obviously, subjects without missing data have $y_{ij}^+ - y_{ij}^- = 0$ and $\omega_{ij} = 1$. Hence the mean index within category h_k, with reference to j-th PC could be given by

$$\bar{y}_{h_k j} = \frac{\sum_{i \in h_k} y_{ij} \omega_{ij}}{\sum_{i \in h_k} \omega_{ij}}. \tag{4}$$

- A second possibility is to synthetically describe the mean within the groups using an interval-valued index, taking account of the eventual presence of subjects with missing data in the group. The mean interval $[\bar{y}_{h_k j}^-, \bar{y}_{h_k j}^+]$ could be then computed as

$$\bar{y}_{h_k j}^- = \frac{\sum_{i \in h_k} y_{ij}^-}{\sum_{i=1}^N I_i(h_k)} \qquad \bar{y}_{h_k j}^+ = \frac{\sum_{i \in h_k} y_{ij}^+}{\sum_{i=1}^N I_i(h_k)} \tag{5}$$

  where $I_i(h_k)$ is the indicator function

$$I_i(h_k) = \begin{cases} 0 & if \quad i \notin h_k \\ 1 & if \quad i \in h_k \end{cases}$$

If a category h_k is composed by subjects without missing values, $y_{ij}^- = y_{ij}^+$ for all $i \in h_k$, and $\bar{y}_{h_k j}^- = \bar{y}_{h_k j}^+$.

It should be noted that the weighting mechanism introduced in (4) makes the two approaches completely different from the point of view of the computation of the mean, so that it is not ensured that $\bar{y}_{h_k j} \in [\bar{y}_{h_k j}^-, \bar{y}_{h_k j}^+]$. In other words, the synthetic value $\bar{y}_{h_k j}$ obtained by (4) could lie outside the interval $[\bar{y}_{h_k j}^-, \bar{y}_{h_k j}^+]$ computed by (5). Hence the two procedures could be used and compared each other.

5. A CASE STUDY: BUYING AND SELLING ON LINE IN EUROPEAN COUNTRIES

The proposed method is now applied to a real dataset dealing with the attitude toward buying and selling on line of enterprises and individuals within European countries in the year 2004[3].

TABLE 1. - *Observed variables*

Variable	Description
ebuy	Percentage of enterprises having purchased on-line
eirpay	Percentage of enterprises having received on-line payments for Internet sales
esell	Percentage of enterprises having received orders on-line
eturn	Percentage of enterprises' total turnover from e-commerce
ibuy3	Percentage of individuals who ordered goods or services, over the Internet, for private use, in the last 3 months

The 5 observed variables (table 1) are affected by several missing; countries with missing in all variables have been deleted and the analysis is performed on $N = 25$ subjects which can be divided into three groups: North-Centre Europe (NCE), South-Europe (SE), East Europe (EE) (table 2). A 20% of the final dataset is missing: this situation is similar to that inspected with the simulative study described before, so that its results can be used in order to decide how the analysis should be organized. More specifically, when 20% values were missing, we observed an overall bad performance of listwise compared with pairwise deletion and InI. For this reason, to carry out the first-step traditional analysis[4], we choose to compute the correlation matrix with pairwise deletion. The variance accounted for by the first two PCs with this approach is 89.93%. The two-dimensional factor loadings plot is displayed in figure 13 and shows that the important first PC (accounting for a 77.26 % of total variance), summarizes the whole phenomenon.

At a second step InI is performed. The missing data have to be replaced by an interval covering a reasonable range of possible values. In this case the variables are all percentages, so that a natural interval could be $[0, 100]$. Nevertheless such an interval appears largely oversized with respect to the real consistency of the analyzed phenomenon (observed mean, standard deviation, minimum and maximum of the five variables are shown in table 3) and for the generic variable X_j the interval

$$[\max(0; \mu_j - 3\sigma_j), \min(100; \mu_j + 3\sigma_j)]$$

seems to be more appropriate.

[3] Data from EUROSTAT databases (http://epp.eurostat.cec.eu.int).

[4] We recall that, in any case, using a traditional analysis, we renounce to the representation of subjects with missing in the factorial space.

TABLE 2. - *Subjects and missing data*

Country code	Country	Group	Missing data
at	Austria	NCE	-
be	Belgium	NCE	ibuy3
bg	Bulgaria	EE	ibuy3
cz	Czech Republic	EE	ibuy3
dk	Denmark	NCE	-
ee	Estonia	EE	-
fi	Finland	NCE	-
de	Germany	NCE	-
gr	Greece	SE	-
hu	Hungary	EE	eturn
is	Iceland	NCE	ebuy eirpay esell eturn
ie	Ireland	NCE	-
it	Italy	SE	ebuy esell ibuy3
lt	Lithuania	EE	esell
lu	Luxembourg	NCE	ebuy eirpay esell eturn
nl	Netherlands	NCE	ibuy3 eturn
no	Norway	NCE	eirpay
pl	Poland	EE	-
pt	Portugal	SE	-
ro	Romania	EE	ebuy esell ibuy3 eturn
si	Slovenia	EE	eturn
sk	Slovakia	EE	ibuy3
es	Spain	SE	-
se	Sweden	NCE	-
uk	United Kingdom	NCE	-

The SPCA is performed with Vertices Method using the SODAS software[5]. The variance accounted for by the first two PCs is 60.77% (41.57% and 19.20% for the first and the second PC respectively), a sensibly lower value compared with the former approach (an expected result, considering the outcomes of the simulative study). Nevertheless a two-dimensional solution seems again acceptable. The factor loadings plot (figure 14) is quite similar to that obtained with pairwise deletion, with a difference only in the loading of variable ibuy3. The solution obtained with InI reveals

[5] SODAS is the result of the work of 17 teams of 9 countries involved in an European project of EUROSTAT. Its prototype is free, downloadable at http://www.ceremade.dauphine.fr/%7Etouati/sodas-pagegarde.htm.

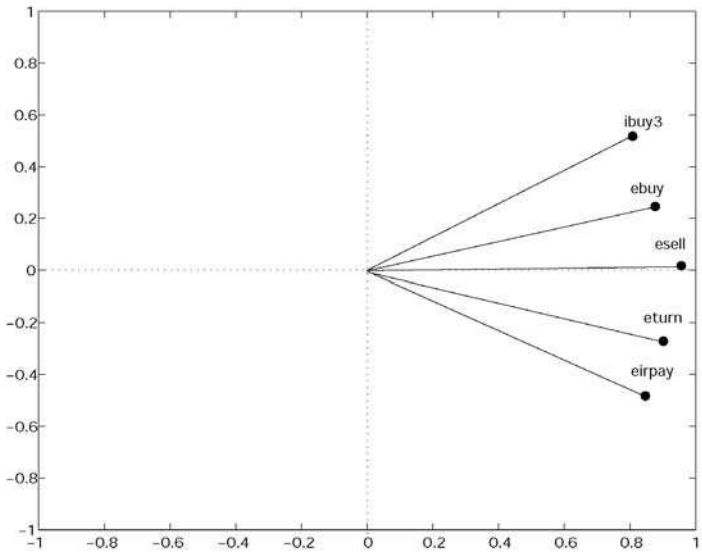

FIGURE 13. - *Factor loadings plot of the first two PCs (PCA with pairwise deletion)*

TABLE 3. - *Observed variables: means, standard deviations, minimum, maximum, intervals replacing missing data*

Variable	Mean	Standard deviation	Min	Max	Interval
ebuy	22.0476	13.7511	3	50	[0, 63.3009]
eirpay	2.5455	1.9477	0	8	[0, 8.3886]
esell	12.6500	7.1991	2	27	[0, 34.2473]
eturn	7.1158	4.9195	1.6	20	[0, 21.8743]
ibuy3	14.8333	11.9826	1	32	[0, 50.7811]

the possible presence of two dimensions: enterprises - first PC - and individuals - second PC.

Figure 15 displays the two-dimensional convex hull representations of the countries, divided in the three groups. It is immediately apparent that countries with many missing values (e.g. Iceland, Luxembourg, Romania, Italy) are characterized by a very large hull, meaning that their information is too poor to allow a sure positioning in the factorial space. Other countries (e.g Netherlands), with a lower number of missing values, are located in a smaller area, allowing some interpretation. For countries with only one missing an almost sure positioning is possible, with easy interpretation, at least along one dimension. Countries with no missing are represented by points.

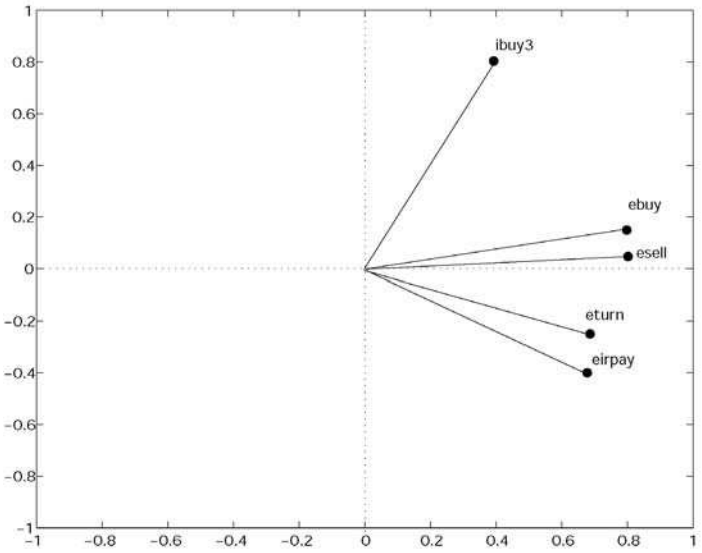

FIGURE 14. - *Factor loadings plot of the first two PCs (V-SPCA)*

With the only exception of Luxembourg and Iceland, whose areas are too large to allow any interpretation, North-Centre European countries exhibit a certain propensity of enterprises towards buying and selling on-line. The situation of individuals is heterogeneous.

A completely different scenery characterizes South European countries, with exceptionally low propension to e-commerce both of enterprises and individuals. An exception is represented by Italy whose location, although not precisely defined, could differ from the others. However sure pronouncements in this sense are not possible.

East Europe landscape looks again various along the individuals' dimension. Enterprises' attitude is generally modest (Romania could constitute an exception,

TABLE 4. - *Synthesis of the mean attitudes within the three groups*

	NCE	**SE**	**EE**
$\bar{y}_{h_k 1}$	1.2070	−1.6826	−1.0274
$\bar{y}_{h_k 2}$	−0.1549	−0.3663	−0.4460
$\bar{y}^{-}_{h_k 1}$	0.3167	−1.8275	−1.5344
$\bar{y}^{+}_{h_k 1}$	1.7858	−0.7625	−0.1156
$\bar{y}^{-}_{h_k 2}$	−0.5442	−0.4325	−0.7611
$\bar{y}^{+}_{h_k 2}$	0.6067	0.3400	0.7178

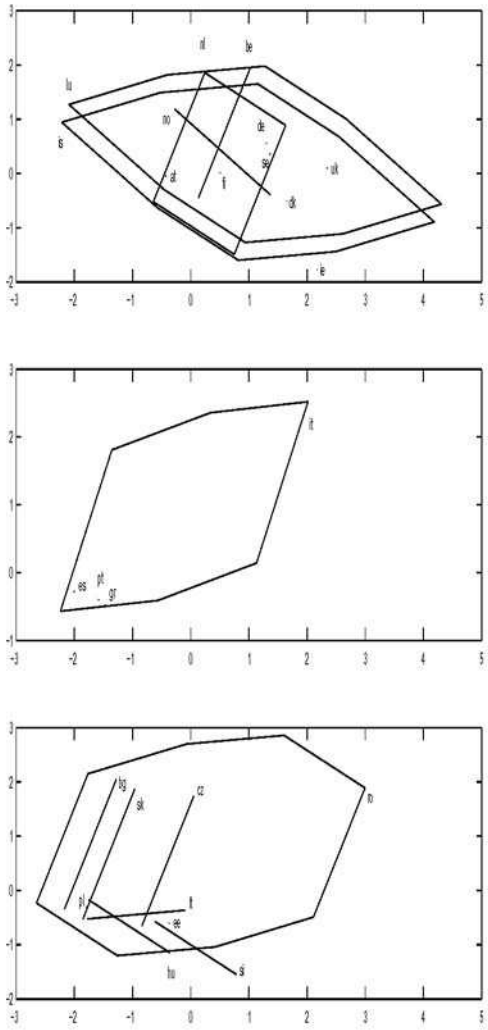

FIGURE 15. - *Convex hulls representation of the countries in the factorial plane, North-Centre Europe (top), South Europe (middle), East Europe (bottom)*

but its area is really very large), while the individual's attitude ranges from low (Slovenia, Hungary, Estonia) to possibly moderately high (Bulgaria, Slovakia, Czech Republic) levels.

The two PCs can be considered synthetic measures of the two main features of the analyzed phenomenon and a ranking of the countries can be made for both the enterprises and the individuals aspect. Following the observations remarked in Section 4, for countries affected by missing data the interval description is maintained, and the midpoint of the interval is taken as a reference value. The two rankings are

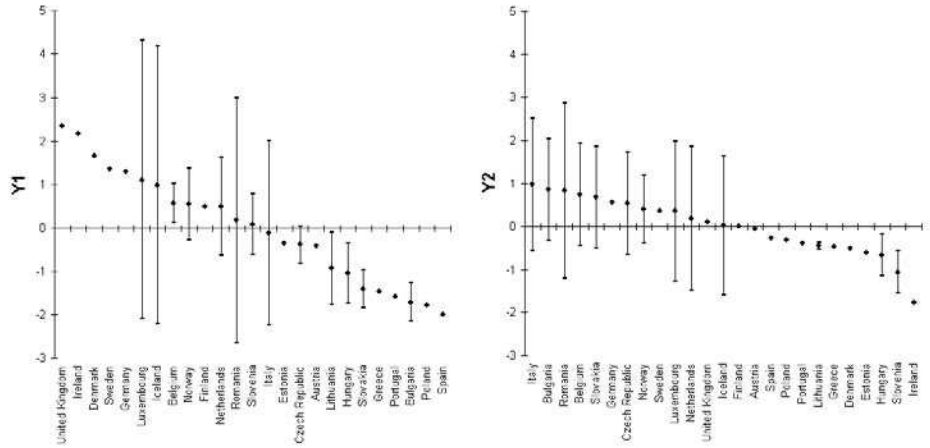

FIGURE 16. - *Ranking of the countries from the point of view of enterprises' (left) and individuals' (right) attitudes toward e-commerce*

presented in figure 16, where, despite the variability affecting certain subjects, the relative positioning of countries can be easily appreciated.

A synthesis of the mean attitudes within the three groups can be obtained, as outlined in Section 4, with the weighted index (4) or with the interval description (5).

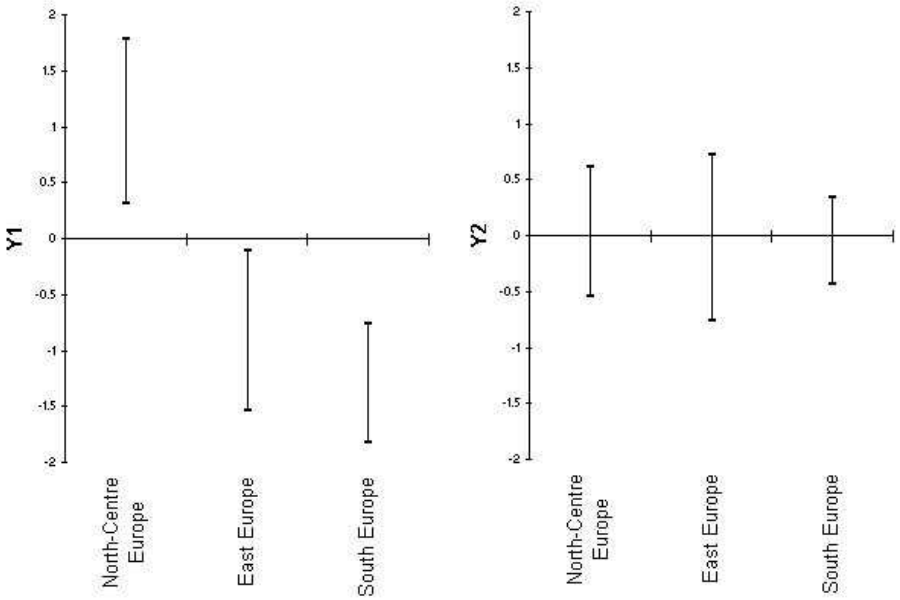

FIGURE 17. - *Interval description of the mean attitudes within the three groups, from the point of view of enterprises (left) and individuals (right)*

The results of the two approaches[6], summarized in table 4 and in figure 17, show that the feature mainly discriminating the three groups is the enterprises' attitude, corresponding to the most informative PC (accounting for a 41.57% of total variance), while the individuals' attitude (however accounting only for a 19.20% of total variance) does not provide interesting differences among groups.

6. CONCLUDING REMARKS

In this paper a strategy for handling missing data is proposed with specific attention to the framework of Principal Component Analysis. It is based on the replacement of each missing value with a suitable interval, so that the main difference with traditional approaches is the possibility to avoid the deletion of subjects with missing, without necessarily passing through imputation. The treatment of a data matrix with interval-valued and single-valued measurements mixed is possible thanks to some recent proposals in the theoretical framework of Symbolic Data Analysis. An illustrative example is examined. Afterward a brief simulated study is used in order to explore how the method works under the points of view of explained variance, PCs formulation, representation of subjects in the factorial space. Even if partial, this study gives an idea of the favorable performance of the proposed method, compared with listwise and pairwise deletion. Finally the procedure is applied to real data describing the attitude of European enterprises and individuals toward buying and selling on line.

In brief, the proposed method has at least four major advantages:

- the replacement of missing data with an interval, covering a wide set of possible values, makes the procedure more ''objective'' than many imputation proposals present in the literature;
- subjects with missing values are maintained in the dataset, and the information they contain, even if partial, is in any case, recovered (also pairwise deletion has this property, but it does not allow to represent incomplete subjects in the factorial space);
- the subjects with missing values can be projected in the factorial space and their particular graphical representation allows to appreciate their peculiar fuzzy condition;
- for a given subject the missing information differently affects its representation in the factorial space, in accordance with the loadings of the variables where the missing is located. In other words the missing data mainly affect the PCs where the corresponding variable has high score.

[6] For the weighted index (4) a negative exponential function $f(y_{ij}^{+} - y_{ij}^{-}) = e^{-(y_{ij}^{+} - y_{ij}^{-})} = \omega_{ij}$ has been used.

RIASSUNTO

Per effettuare un'Analisi delle Componenti Principali su una matrice dei dati affetta da valori mancanti vi sono essenzialmente due strategie: sostituire i valori mancanti con valori opportuni, determinati secondo qualche criterio, o escludere dall'analisi i soggetti incompleti. Entrambe le soluzioni possono risultare costose e rischiose, specialmente quando i valori mancanti sono molti e distribuiti uniformemente nei dati, cioè non concentrati su poche unità. In questo lavoro viene formulata una proposta alternativa per il trattamento dei valori mancanti, che utilizza le recenti tecniche di Analisi dei Dati Simbolici.

REFERENCES

Allison P.D. (2002). *Missing data*, Sage Publications, Thousand Oaks.

Amato S., Palumbo F. (2004). Multidimensional Gap Analysis, *Statistica Applicata*, **16**, 3.

Barber C.B., Dobkin D.P., Huhdanpaa H.T. (1996). The Quickhull Algorithm for Convex Hulls, *ACM Transactions on Mathematical Software*, **22**, 469-483.

Bock H.H., Diday E. (eds.) (2000). *Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data*, Springer Verlag, Heidelberg.

Bravo M.C., García-Santesmases J.M. (1998). Symbolic object description of strata by segmentation trees, in Ph. Nanapoulos et al. (eds.), NNTS'98, 1998, 85-90.

Cazes P., Chouakria A., Diday E., Schektman Y. (1997). Extension de l'analyse en composantes principales à des données de type intervalle, *Revue de Statistique Appliquée*, **45**, 3, 5-24.

Chavent M. (1998). A monothetic clustering algorithm, *Pattern Recognition Letters*, **19**, 989-996.

De Carvalho F.A.T. (1994). Proximity coefficients between Boolean symbolic objects. In Diday E., Lechevalier Y., Shader M. et al. (eds.), IFCS-93, 1994, 387-394.

De Carvalho F.A.T. (1996). Histogrammes et indices de proximité en analyse données symboliques, *Acyes de l'école d'été sur l'analyse des données symboliques*, LISE-CEREMADE, Université de Paris IX Dauphine, 101-127.

Devlin S.J., Gnanadesikan R., Kettenring J.R. (1981). Robust estimation of dispersion matrices and principal components, *Journal of the American Statistical Association*, **76**, 354-362.

Diday E. (1987). Introduction l'approche symbolique en Analyse des Donnés, *Premìere Journées Symbolique-Numérique*, Université de Paris IX Dauphine.

D'Urso P., Giordani P., (2005). A possibilistic approach to latent component analysis for symmetric fuzzy data, *Fuzzy Sets and Systems*, **150**, 285-305.

Giordani P., Kiers H.A.L. (2004a). Three-way component analysis of interval-valued data, *Journal of Chemometrics*, **18**, 253-264.

Giordani P., Kiers H.A.L. (2004b). Principal Component Analysis of symmetric fuzzy data, *Computational Statistics and Data Analysis*, **45**, 519-548.

Godwa K.C., Diday E. (1991). Symbolic clustering using a new dissimilarity measure, *Pattern Recognition*, **24**, 6, 567-578.

Godwa K.C., Diday E., Nagabhushan P. (1995). Dimensionality reduction of symbolic data, *Pattern Recognition Letters*, **16**, 219-223.

Green P.J. (1981). *Peeling Bivariate Data*, in Barnett V. (ed.), *Interpreting Multivariate Data*, John Wiley & Sohns, 3-19.

Green P.J., Silverman B.W. (1979). Constructiong the convex hull of a set of points in the plane, *The Computer Journal*, **22**, 262-266.

Ichino M., Yaguchi H. (1994). Generalized Minkowski metrics for mixed feature type data analysis, *IEEE Transactions on Systems, Man and Cybernetics*, **24**, 4, 698-708.

Lauro N.C., Verde R., Palumbo F. (2000). Factorial discriminant analysis of symbolic objects, in Bock H.H., Diday E. (eds.), *Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data*, Springer Verlag, 212-233.

Lauro N.C., Palumbo F. (2000). Principal Component Analysis if interval data: a symbolic data analysis approach, *Computational Statistics*, **15**, 1, 73-87.

Lauro N.C., Palumbo F. (2004). Principal Component Analysis for non-precise Data, in Vichi M., Monari P., Mignani S., Montanari A. (eds.), *New Developments in Classification and Data Analysis*, Springer Verlag, 173-184.

Little R.J.A, Rubin D.B. (1987). *Statistical Analysis with Missing Data*, John Wiley & Sohns, New York.

Rodríguez O., Diday E., Winsberg S. (2000). Generalization of the Principal Components Analysis to histogram data, *Proc. PKDD2000*, Lyon, France.

Rubin D.B. (1976). Inference and missing data, *Biometrika*, **63**, 581-592.

Rubin D.B. (1977). The design of a generaland flexible system for handling non-response in sample surveys, *Manuscript*.

Rubin D.B. (1978). Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse, *Proceedings of the Survey Research Method section of the American Statistical Association*, 20-34.

Rubin D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sohns, New York.

Van Buuren S., Eisinga E. (eds.) (2003). Special Issue: *Incomplete data: multiple imputation and model-based analysis*, *Statistica Neerlandica*, **57**, 1.

Van Praag B.M.S., Dijkstra T.K., Van Velzen J. (1985). Least-sqaures theory based on general distributional assumption with an application to the incomplete observations problem, *Psycometrika*, **50**, 25-36.