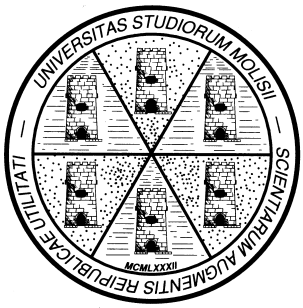
Università degli Studi del Molise
Facoltà di Economia
Dipartimento di Scienze Economiche, Gestionali e Sociali
Via De Sanctis, I-86100 Campobasso (Italy)



ECONOMICS & STATISTICS DISCUSSION PAPER
No. 13/03

A least squares approach to Principal Component Analysis for interval valued data

by

Pierpaolo D'Urso
University of Molise, Dept. SEGeS

and

Paolo Giordani
"La Sapienza" University, Rome, Dept. of Statistics, Probability, and Applied Statistics

A least squares approach to Principal Component Analysis for interval valued data

Pierpaolo D'Urso

Dipartimento di Scienze Economiche, Gestionali e Sociali,

Università degli Studi del Molise, Via De Sanctis, 86100 Campobasso, Italy.

E-mail: durso@unimol.it, pierpaolo.durso@uniroma1.it

Paolo Giordani [1]

Dipartimento di Statistica, Probabilità e Statistiche Applicate,

Università degli Studi di Roma "La Sapienza", P.le A. Moro, 5 - 00185 Roma, Italy.

E-mail: paolo.giordani@uniroma1.it

Abstract: Principal Component Analysis (PCA) is a well known technique the aim of which is to synthesize huge amounts of numerical data by means of a low number of unobserved variables, called components. In this paper, an extension of PCA to deal with interval valued data is proposed. The method, called *Midpoint Radius Principal Component Analysis* (MR-PCA) recovers the underlying structure of interval valued data by using both the midpoints (or centers) and the radii (a measure of the interval width) information. In order to analyze how MR-PCA works, the results of a simulation study and two applications on chemical data are proposed.

Keywords: Principal Component Analysis, Least squares approach, Interval valued data, Chemical data

[1] Corresponding author.

1. Introduction

In conventional data analysis, the variables are represented by single valued vectors (numerical vectors). However, in several substantive applications, the utilization of single valued variables may bring about a heavy loss of information. For example, in chemometrics, we may study the mineral concentrations of food products analyzed at different times or in different experimental situations; in meteorology, we may consider the daily temperature, humidity and wind speed registered in different places; in environmetrics, we may refer to the pollutant concentrations of SO_2, CO, NO, NO_2, O_3 recorded at various places; in finance, we may examine the daily rate of exchange between Euro-Dollar or Euro-Sterling; in medicine, we may make reference to daily systolic and diastolic pressure, pulse rate, temperature of patients; and so on. In all the previous cases, it is more interesting to take into account the minimum and maximum values registered in the considered period rather than the average one because they offer more detailed and complete information about the phenomenon under examination.

We can formalize an interval valued datum as $x_{ij}=[\underline{x}_{ij},\overline{x}_{ij}]$, $i=1,\ldots,I$; $j=1,\ldots,J$, where x_{ij} represents the j-th interval valued variable observed on the i-th observation unit; $\underline{x}_{ij}$ and $\overline{x}_{ij}$ denote, respectively, the lower and upper bounds of the interval; in particular, they represent the minimum and maximum values registered for the j-th interval valued variable with respect to the i-th observation unit. Notice that, in the general case of J interval valued variables, each observation unit can be represented geometrically by a hyperrectangle in $\Re^J$ having 2^J vertices. The set of the 2^J vertices corresponds to all the possible (lower bound, upper bound) combinations. In particular, in $\Re$ (J=1) the generic object is represented by a segment; in $\Re^2$ (J=2), it is represented by a rectangle with 2^2=4 vertices, and so on. See also [1-2]. Moreover, see, in the fuzzy data framework, [3-4].

In many real situations, as it happens with traditional single valued data, it is desirable to compress interval valued data losing relevant information as little as possible. When the data set is numerical valued, this can be done by means of Principal Component Analysis (PCA). Let **X** be the numerical data matrix of order $(I\times J)$. We have

$$\mathbf{X}=\mathbf{AB}'+\mathbf{E} \qquad (1)$$

where **A** is the component scores matrix of order $(I \times P)$, **B** is the component loadings matrix of order $(J \times P)$, **E** $(I \times J)$ is the residual matrix and P ($<J$) is the number of extracted components. Notice that $\mathbf{AB}'$ provides the best approximation of rank P of the original data matrix **X**.

The popularity of PCA (and its three-way extensions) in chemistry is recognized. For instance, it is used for second-order calibration, fluorescence spectroscopy, chromatography, food quality evaluation (see, e.g., [5-7]). In this paper, we propose suitable extensions of conventional Principal Component Analysis (PCA) when the data are intervals.

The paper is organized as follows. In the next section we recall Vertices Principal Component Analysis (V-PCA) and Centers Principal Component Analysis (C-PCA). They are, probably, the two most popular methods which detect the underlying structure of interval valued data. In section 3, we propose our method. In section 4, we propose how to plot the observation units in the obtained low dimensional space. In section 5, we give the results of a simulation study carried out in order to compare our method with C-PCA and V-PCA. Finally, in section 6, we show two applications of our method on two chemical data sets.

2. Vertices Principal Component Analysis (V-PCA) and Centers Principal Component Analysis (C-PCA).

Vertices Principal Component Analysis (V-PCA) and Centers Principal Component Analysis (C-PCA) are multi-step procedures which aim at detecting the underlying structure of two-way interval valued data sets [1-2]. Let us consider to deal with I observation units characterized by J interval valued variables. The data are stored in the interval valued matrix **X**, the generic element of which is $x_{ij} = [\underline{x_{ij}}, \overline{x_{ij}}]$, i=1,…,I , j=1,…,J .

The first step of both methods consists of replacing the interval valued matrix by a single valued one. In V-PCA, this is done by transforming the original data matrix of order $(I \times J)$ into a numerical one of order $(I2^J \times J)$. In the original interval valued matrix **X**, the generic i-th row pertains to the i-th observation unit. Each row is transformed into the submatrix ${}_i\mathbf{X}$ of order $(2^J \times J)$ in which each row refers exactly to each vertex of the hyperrectangle associated to the generic i-th observation unit. Thus, with regard to observation unit i, if we have $J = 2$ variables, the coding procedure leads to

$$
{}_i\mathbf{X} = \begin{pmatrix} \underline{x_{i1}} & \underline{x_{i2}} \\ \underline{x_{i1}} & \overline{x_{i2}} \\ \overline{x_{i1}} & \underline{x_{i2}} \\ \overline{x_{i1}} & \overline{x_{i2}} \end{pmatrix} \tag{2}
$$

and the new data matrix is

$$
\mathbf{X}_{V-PCA} = \begin{pmatrix} {}_1\mathbf{X} \\ \vdots \\ {}_I\mathbf{X} \end{pmatrix}. \tag{3}
$$

V-PCA is nothing but performing PCA on (3). It should be clear that V-PCA is computationally cumbersome when the data size is huge because the number of rows of the matrix in (3) is exponentially related to the number of variables. However, the computation of the component loadings matrix can be simply obtained, because it is based on the eigendecomposition of the cross-products matrix $\mathbf{X}'_{V-PCA}\mathbf{X}_{V-PCA}$, which can be easily computed. In fact, it can be shown that

$$
\mathbf{X}'_{V-PCA}\mathbf{X}_{V-PCA} = 2^{J-2}\begin{bmatrix} 2\sum_{i=1}^{I}\left(\underline{x}_{i1}^2 + \overline{x}_{i1}^2\right) & \sum_{i=1}^{I}\left(\underline{x}_{i1} + \overline{x}_{i1}\right)\left(\underline{x}_{i2} + \overline{x}_{i2}\right) & \cdots & \sum_{i=1}^{I}\left(\underline{x}_{i1} + \overline{x}_{i1}\right)\left(\underline{x}_{iJ} + \overline{x}_{iJ}\right) \\ \sum_{i=1}^{I}\left(\underline{x}_{i2} + \overline{x}_{i2}\right)\left(\underline{y}_{i1} + \overline{x}_{i1}\right) & 2\sum_{i=1}^{I}\left(\underline{x}_{i2}^2 + \overline{x}_{i2}^2\right) & \cdots & \sum_{i=1}^{I}\left(\underline{x}_{i2} + \overline{x}_{i2}\right)\left(\underline{x}_{iJ} + \overline{x}_{iJ}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{I}\left(\underline{x}_{iI} + \overline{x}_{iI}\right)\left(\underline{x}_{i1} + \overline{x}_{i1}\right) & \sum_{i=1}^{I}\left(\underline{x}_{iI} + \overline{x}_{iI}\right)\left(\underline{x}_{i2} + \overline{x}_{i2}\right) & \cdots & 2\sum_{i=1}^{I}\left(\underline{x}_{iJ}^2 + \overline{x}_{iJ}^2\right) \end{bmatrix}. \tag{4}
$$

See, for further details, [2].

In C-PCA, each element of the interval valued matrix $\mathbf{X}$ is replaced by the midpoint (or center) of the associated interval. Thus, we get the $(I \times J)$-matrix

$$
\mathbf{X}_{C-PCA} = \begin{pmatrix} m_{11} & \cdots & m_{1J} \\ \vdots & \ddots & \vdots \\ m_{I1} & \cdots & m_{IJ} \end{pmatrix}, \tag{5}
$$

where $m_{ij}=\frac{\overline{x_{ij}}+\underline{x_{ij}}}{2}$, for i=1,…,I and j=1,…,J. Now, classical PCA is performed on $\mathbf{X}_{C-PCA}$.

Using both methods, we can represent each observation unit as a low dimensional hyperrectangle. When the loadings are columnwise orthonormal, the scores provide the projection of the observation units in the low dimensional space spanned by the loadings. With respect to V-PCA, the scores give the coordinates of the vertices for all the observation units in such low dimensional space. Unfortunately, the projected vertices do not define exactly a hyperrectangle. This problem can be solved by considering the Maximum Covering Area Rectangle (MCAR), that is considering the hyperrectangle (the rectangle if we extract P=2 components) which encloses all the projected vertices. If we are performing C-PCA, the low dimensional hyperrectangles can be found by noting that the coordinates of the midpoints are enclosed between the lower and upper bounds and that the principal components are linear functions of the m_{ij}'s. See, for further details, [1-2].

So far, we recalled the two most popular methods to analyze the underlying structure of interval valued data sets. Further extensions of PCA for interval valued data can be found in [8-9]. In the next section we provide a different approach to the problem.

3. Principal Component Analysis for Interval Valued Data

3.1. The model

We already noticed that each observation unit can be seen as a hyperrectangle in $\Re^J$. Let **M** be the midpoints matrix as given in (5), the generic element of which is the midpoint of the associated interval. Moreover, let **R** be the radii matrix of order $(I \times J)$

$$\mathbf{R}=\begin{pmatrix} r_{11} & \cdots & r_{1J} \\ \vdots & \ddots & \vdots \\ r_{I1} & \cdots & r_{IJ} \end{pmatrix}, \tag{6}$$

the generic element of which is $r_{ij}=\frac{\overline{x_{ij}}-\underline{x_{ij}}}{2}$, for $i=1,\ldots,I$ and $j=1,\ldots,J$. Thus, the radius is the

half-width of an interval.

The principal component model for interval valued data is given by:

$$\mathbf{M}=\mathbf{M}^{*}+\mathbf{E}_{M}, \tag{7}$$

$$\mathbf{M}^{*}=\mathbf{A}_{M}\mathbf{B}', \tag{8}$$

$$\mathbf{R}=\mathbf{R}^{*}+\mathbf{E}_{R}, \tag{9}$$

$$\mathbf{R}^{*}=\mathbf{A}_{R}\mathbf{B}', \tag{10}$$

$$\mathbf{M}+\mathbf{R}\mathbf{H}_{k}=\mathbf{M}^{*}+\mathbf{R}^{*}\mathbf{H}_{k}+\mathbf{Z}_{k} \qquad k=1,...,K, \tag{11}$$

where, $\mathbf{M}^{*}$ and $\mathbf{R}^{*}$ are the matrices of order $(I\times J)$ of the estimated midpoints and radii, respectively. $\mathbf{A}_{M}$ and $\mathbf{A}_{R}$ are, respectively, the component scores matrices for the midpoints and for the radii of order $(I\times P)$, $\mathbf{B}$ is the component loadings matrix of order $(J\times P)$. Finally $\mathbf{E}_{M}$, $\mathbf{E}_{R}$ and $\mathbf{Z}_{k}$ are residual matrices of order $(I\times J)$. We refer to the model in (7)-(11) as *Midpoint Radius Principal Component Analysis* (MR-PCA).

In MR-PCA, we assume that different scores are determined for the midpoints and the radii, whereas the loadings are the same. Thus, the proposed model is based upon the assumption that the midpoints and the radii are modelled by means of the same components. It follows that the MR-PCA model can be seen as a special case of Simultaneous Component Analysis with invariant Pattern (SCA-P). SCA is a generalization of PCA proposed in [10-12] when observations on the same variables have been registered in more than one population. Instead of analyzing the observations separately, the idea is to find components that explain as much variance as possible in all populations simultaneously. In SCA-P [12], a special version of SCA, the number of component scores matrices is equal to the number of populations, while one common loadings matrix is constructed. It is well known that the same components can be always extracted from different populations and, indeed, in the MR-PCA method, from the midpoints and radii matrices. However, one can argue whether such components are able to synthesize simultaneously both the midpoints and the radii in a satisfactory way. In Section 5, we aim at answering this question. In particular, we will give the results of a simulation study carried out in order to assess whether our method recovers the underlying structure in the data better than C-PCA and V-PCA.

It is fruitful to observe that the model in (7)-(11) is the same as the one for fuzzy data proposed in [4]. It is recognized that the analysis of interval valued data can be considered as a sub-domain of fuzzy set theory. Specifically, an interval valued number can be seen as a fuzzy number with the so-called rectangular membership function. See, for instance, [13]. However, in spite of all that, we

prefer to treat differently interval valued data and fuzzy data. This is based upon the assumption that methods suitable for fuzzy data should be constructed in such a way that the role of the midpoints or centers (whose membership function values are maximal) must be emphasized. This does not hold in the interval valued data framework where all the points in the interval are considered on the same foot.

So far, we showed how to model the midpoints and the radii. Following a least squares approach, the optimal component matrices $\mathbf{A}_M$, $\mathbf{A}_R$ and $\mathbf{B}$ are then obtained by minimizing a suitable loss function, which compares the observed and theoretical data as given in (7)-(11). Such a loss function is based on the distance measure between observed and theoretical interval valued data that will be proposed in the next subsection.

3.2. The distance measure

In order to compare, in the least squares sense, two observation units described by J interval valued variables, we suggest to consider all the vertices of the two hyperrectangles pertaining to the observation units involved. Thus, we get the following squared distance:

$$\Delta^2(i',i'') = \sum_{k=1}^{K} \left\| \left(\mathbf{m}_{i'} + \mathbf{r}_{i'} * \mathbf{h}_k\right) - \left(\mathbf{m}_{i''} + \mathbf{r}_{i''} * \mathbf{h}_k\right) \right\|^2, \tag{12}$$

in which $\mathbf{m}_{i'}$ and $\mathbf{m}_{i''}$ denote, respectively, the i'-th and the i'' rows of $\mathbf{M}$ and $\mathbf{r}_{i'}$ and $\mathbf{r}_{i''}$ those of $\mathbf{R}$. The symbol $*$ denotes the Hadamard product, that is the elementwise product of two matrices (vectors) of the same order. The vectors, $\mathbf{h}_k$, $k = 1,\ldots,K$, where $K=2^J$, help us to define every vertex of the hyperrectangle associated to each observation unit separately. In fact, their elements are equal to ± 1 in order to refer exactly to every vertex. The vectors $\mathbf{h}_k$, $k = 1,\ldots,K$, are the rows of a new matrix, say $\mathbf{H}$, of order $(K \times J)$. If $J=3$, we get:

$$\mathbf{H}=\begin{bmatrix} -1 & -1 & -1 \\ -1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & -1 \\ 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ -1 & 1 & 1 \end{bmatrix}. \tag{13}$$

Using $\mathbf{h}_1$ (first row of $\mathbf{H}$) we get the vector of the lower bounds pertaining to the *i*-th observation unit:

$$\left(\underline{x_{i1}} \quad \underline{x_{i2}} \quad \underline{x_{i3}}\right) = \mathbf{m}_i + \mathbf{r}_i * \left(-1 \quad -1 \quad -1\right) = \mathbf{m}_i - \mathbf{r}_i . \tag{14}$$

Analogously to (14), by means of $\mathbf{h}_5$, we obtain the vector of the upper bounds:

$$\left(\overline{x_{i1}} \quad \overline{x_{i2}} \quad \overline{x_{i3}}\right) = \mathbf{m}_i + \mathbf{r}_i * \left(1 \quad 1 \quad 1\right) = \mathbf{m}_i + \mathbf{r}_i . \tag{15}$$

In order to compare a set of *I* observation units characterized by *J* interval valued variables, and a set of estimated ones, the use of (12) leads to the following squared distance:

$$\Delta^2 = \sum_{k=1}^{K} \left\| \left(\mathbf{M} + \mathbf{R}\mathbf{H}_k\right) - \left(\mathbf{M}^* + \mathbf{R}^*\mathbf{H}_k\right) \right\|^2 , \tag{16}$$

where $\mathbf{H}_k$, $k = 1,\ldots,K$, are diagonal matrices whose diagonal elements are equal to those of the vectors $\mathbf{h}_k$, $k = 1,\ldots,K$.

It can be easily seen that the matrices $\mathbf{H}_k$, $k = 1,\ldots,K$, satisfy the following properties that will be very useful in order to simplify (16):

$$\mathbf{H}_k\mathbf{H}_k = \mathbf{I}_J , \; k = 1,\ldots,K, \tag{17}$$

$$\sum_{k=1}^{K} \mathbf{H}_k = \mathbf{0}_J . \tag{18}$$

Let us consider the *k*-term of the sum in (16). After a little algebra we get

$$\left\|\left(\mathbf{M}+\mathbf{R}\mathbf{H}_k\right)-\left(\mathbf{M}^*+\mathbf{R}^*\mathbf{H}_k\right)\right\|^2=\left\|\mathbf{M}-\mathbf{M}^*\right\|^2+tr\left(\mathbf{H}_k\mathbf{R}'\mathbf{R}\mathbf{H}_k\right)+tr\left(\mathbf{H}_k\mathbf{R}^{*'}\mathbf{R}^*\mathbf{H}_k\right)+$$
$$-2tr\left(\mathbf{H}_k\mathbf{R}'\mathbf{R}^*\mathbf{H}_k\right)+2tr\left[\left(\mathbf{M}'\mathbf{R}+\mathbf{M}^{*'}\mathbf{R}^*-\mathbf{M}'\mathbf{R}^*-\mathbf{R}'\mathbf{M}^*\right)\mathbf{H}_k\right]. \quad (19)$$

By taking into account (17), (19) can be simplified as

$$\left\|\left(\mathbf{M}+\mathbf{R}\mathbf{H}_k\right)-\left(\mathbf{M}^*+\mathbf{R}^*\mathbf{H}_k\right)\right\|^2=\left\|\mathbf{M}-\mathbf{M}^*\right\|^2+\left\|\mathbf{R}-\mathbf{R}^*\right\|^2+2tr\left[\left(\mathbf{M}'\mathbf{R}+\mathbf{M}^{*'}\mathbf{R}^*-\mathbf{M}'\mathbf{R}^*-\mathbf{R}'\mathbf{M}^*\right)\mathbf{H}_k\right]. \quad (20)$$

Upon substituting (20) into (16) and considering (18), we obtain

$$\Delta^2=\sum_{k=1}^{K}\left\|\left(\mathbf{M}+\mathbf{R}\mathbf{H}_k\right)-\left(\mathbf{M}^*+\mathbf{R}^*\mathbf{H}_k\right)\right\|^2=2^J\left\|\mathbf{M}-\mathbf{M}^*\right\|^2+2^J\left\|\mathbf{R}-\mathbf{R}^*\right\|^2\cong\left\|\mathbf{M}-\mathbf{M}^*\right\|^2+\left\|\mathbf{R}-\mathbf{R}^*\right\|^2=\widetilde{\Delta}^2. \quad (21)$$

We notice that (21) is the matrix generalization of the distance between two vectors of fuzzy numbers proposed in [14].

3.3. The solution

The solution of the model is obtained by means of the Singular Value Decomposition (SVD) of the matrix $\mathbf{Y}=\begin{bmatrix}\mathbf{M}\\ \mathbf{R}\end{bmatrix}$ of order $(2I\times J)$ that is decomposed as $\mathbf{PDQ}'$ where $\mathbf{P}$ and $\mathbf{Q}$ are matrices containing the unit length singular vectors of $\mathbf{Y}$ and $\mathbf{D}$ is the diagonal matrix displaying the singular values of $\mathbf{Y}$ in decreasing order. The best *P*-rank decomposition of $\mathbf{Y}$ is $\mathbf{P}_P\mathbf{D}_P\mathbf{Q}_P'$ where $\mathbf{P}_P$ and $\mathbf{Q}_P$ are matrices containing the first *P* columns of $\mathbf{P}$ and $\mathbf{Q}$, and $\mathbf{D}_P$ is the diagonal matrix displaying the first *P* singular values of $\mathbf{Y}$. In fact, according to (8) and (10), (21) can be written as

$$\widetilde{\Delta}^2=\left\|\mathbf{M}-\mathbf{A}_M\mathbf{B}'\right\|^2+\left\|\mathbf{R}-\mathbf{A}_R\mathbf{B}'\right\|^2=\left\|\begin{bmatrix}\mathbf{M}\\ \mathbf{R}\end{bmatrix}-\begin{bmatrix}\mathbf{A}_M\\ \mathbf{A}_R\end{bmatrix}\mathbf{B}'\right\|^2=\left\|\mathbf{Y}-\mathbf{A}\mathbf{B}'\right\|^2 \quad (22)$$

where $\mathbf{A} = \begin{bmatrix} \mathbf{A}_M \\ \mathbf{A}_R \end{bmatrix}$. We then get the following solution:

$$\mathbf{A}_M = \mathbf{P}_P^1 \mathbf{D}_P \tag{23}$$
$$\mathbf{A}_R = \mathbf{P}_P^2 \mathbf{D}_P \tag{24}$$
$$\mathbf{B} = \mathbf{Q}_P \tag{25}$$

where $\mathbf{P}_P^1$ and $\mathbf{P}_P^2$, contain, respectively, the first I rows and the last I rows of $\mathbf{P}_P$.

We notice that the solution in (23)-(25) does not guarantee that the estimated radii are non negative. If negative estimated radii occur, two possible approaches can be adopted.

The first approach simply consists of replacing negative estimated radii by zero. This is done under the assumption that possible negative estimates of the radii correspond to a lack of uncertainty. Thus, for any practical purpose, negative estimated radii can be set equal to zero.

The second approach is based on the following rowwise Alternative Least Squares (ALS) algorithm that updates the rows of the component matrices $\mathbf{A}_M$, $\mathbf{A}_R$ and $\mathbf{B}$ by minimizing the loss function in (22) subject to the non-negativity constraint on $\mathbf{R}^*$, as given in (10). Notice that the updates of all the rows of $\mathbf{A}_M$ do not affect the non-negativity of the estimated radii. Thus, in order to reduce computation time, the updating of the entire matrix $\mathbf{A}_M$ should be considered. The algorithm involves the following steps:

Step1 (Initialization):

Consider a feasible solution such that $\mathbf{A}_R\mathbf{B}' \geq \mathbf{0}$. It can be randomly generated from the uniform distribution in $[0, z]$, $z > 0$ or it can be obtained considering (24) and (25) by replacing negative values with random values from the uniform distribution in $[0, z]$, $z > 0$.

Step 2 (Updating):

Update $\mathbf{A}_M$ and all the rows of $\mathbf{A}_R$ ($\mathbf{a}_{Ri}$'s) and $\mathbf{B}$ ($\mathbf{b}_j$'s) by solving, for each row of each matrix, the following constrained problem:

$$\begin{aligned} &\text{minimize} \quad \widetilde{\Delta}^2 = \left\|\mathbf{M} - \mathbf{A}_M\mathbf{B}'\right\|^2 + \left\|\mathbf{R} - \mathbf{A}_R\mathbf{B}'\right\|^2 \\ &\text{subject to} \quad \mathbf{A}_R\mathbf{B}' \geq \mathbf{0}. \end{aligned} \tag{26}$$

Step 3 (Convergence):

Check whether the loss function decreased less than a pre-specified percentage with respect to the previous function value. If the decrease is negligible, conclude that the algorithm has converged; otherwise go to Step 2.

The problem in (26) can be solved by means of active sets algorithms as described in [15]. In particular, the LSI algorithm can be adopted. It solves the following problem:

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{x}) = \|\mathbf{C}\mathbf{x} - \mathbf{d}\|^2 \\ &\text{subject to} \quad \mathbf{S}\mathbf{x} \geq \mathbf{t}. \end{aligned} \tag{27}$$

where $\mathbf{C}$ is a $(n_1 \times n_2)$-matrix, $\mathbf{x}$ a n_2-vector, $\mathbf{d}$ an n_1-vector, $\mathbf{S}$ a $(n_3 \times n_2)$-matrix, $\mathbf{t}$ a n_3-vector. The updating of the generic *i*-th row of $\mathbf{A}_R$ is obtained by setting $\mathbf{x} = \mathbf{a}'_{Ri}$, $\mathbf{C} = \mathbf{B}$, $\mathbf{d} = \mathbf{r}'_i$ (where $\mathbf{r}_i$ is the *i*-th row of $\mathbf{R}$), $\mathbf{S} = \mathbf{B}$ and $\mathbf{t} = \mathbf{0}_J$. The updating of the generic *j*-th row of $\mathbf{B}$ is obtained by setting $\mathbf{x} = \mathbf{b}'_j$, $\mathbf{C} = \mathbf{A}$, $\mathbf{d} = \mathbf{y}^j$ (where $\mathbf{y}^j$ is the *j*-th column of $\mathbf{Y}$), $\mathbf{S} = \mathbf{A}_R$ and $\mathbf{t} = \mathbf{0}_I$. Finally, since $\mathbf{A}_M$ does not affect the estimated radii, $\mathbf{A}_M$ is updated by solving an ordinary regression problem. We thus have $\mathbf{A}_M = \mathbf{M}\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}$.

In fact, whenever a matrix or a row is updated, the loss function to minimize does not increase. The expression in (22) has a lower bound and therefore the function value converges to a stable value. We notice that the above procedure does not guarantee that the global optimum is attained.

3.4. The goodness of fit index

In order to evaluate the goodness of fit of the model, we suggest to compare the estimated and observed values by considering the following performance index:

$$F = \left(1 - \frac{\|\mathbf{M} - \mathbf{M}^*\|^2 + \|\mathbf{R} - \mathbf{R}^*\|^2}{\|\mathbf{M}\|^2 + \|\mathbf{R}\|^2} \right) 100. \tag{28}$$

The index takes values from 0 to 100. Values near 100 show that the model fits the data well.

Remark 1: Preprocessing

In the MR-PCA method, if it is desirable to preprocess the data, we suggest to standardize each midpoint score by subtracting the mean and dividing by the standard deviation of the variable at hand. The radii can be preprocessed by dividing them by the standard deviation of the related midpoints.

4. Graphical representation of the observation units

In several cases, it can be convenient to plot the entities of the observation unit mode onto the low dimensional space spanned by the component loadings matrix **B**. This provides a representation of each observation unit as a low dimensional hyperrectangle in $\Re^P$. In the MR-PCA method, as well as for classical PCA and, indeed, for the C-PCA and V-PCA methods, in order to have an adequate plot, **B** must be columnwise orthonormal. If such a property does not hold, the plot is distorted because the axes have unequal length. See, for further details, [16].

The non-iterative algorithm guarantees that **B** is columnwise orthonormal taking into account that, in the SVD decomposition, $\mathbf{Q}_P$ contains the first P unit length singular vectors of the matrix on which the SVD is performed. If negative estimates of the radii occur, the iterative algorithm must be run. In this case, the optimal loadings matrix **B** is not columnwise orthonormal. Thus, we find a transformation matrix **T** such that $\hat{\mathbf{B}}=\mathbf{BT}$ is columnwise orthonormal provided that **A** is postmultiplied by $(\mathbf{T}')^{-1}$, that is $\hat{\mathbf{A}}=\mathbf{A}(\mathbf{T}')^{-1}$. In fact, this procedure does not modify the fitting of the model taking into account that

$$\hat{\mathbf{A}}\hat{\mathbf{B}}'=\mathbf{A}(\mathbf{T}')^{-1}(\mathbf{BT})'=\mathbf{AB}'. \tag{29}$$

It is well known that the matrix **T** can be found, for instance, by means of the Gram-Schmidt orthonormalization procedure.

To simplify the notation, let us suppose that **B** is columnwise orthonormal. Going into detail, two plotting procedures can be proposed. The first one consists of projecting all the vertices pertaining to each hyperrectangle in the low dimensional space spanned by **B**. The matrices of the estimated vertices of the *i*-th hyperrectangle, say $\mathbf{V}_i$, of order $(K \times J)$, can be written as

$$\mathbf{V}_i = \mathbf{1}\mathbf{M}_i^* + \mathbf{H}\mathbf{R}_i^* \tag{30}$$

where **H** was introduced in Section 3, $\mathbf{M}_i^*$ and $\mathbf{R}_i^*$ are diagonal matrices whose main diagonals are the *i*-th rows of, respectively, $\mathbf{M}^*$ in (7) and $\mathbf{R}^*$ in (10), and **1** is a matrix with unit elements of order $(K \times J)$. The matrix

$$_i\mathbf{A} = \mathbf{V}_i\mathbf{B} \tag{31}$$

contains the coordinates of the vertices pertaining to the *i*-th hyperrectangle in the low dimensional space spanned by **B**. As one may expect, the union of the points in (31) does not define a low dimensional hyperrectangle. The problem can be solved by considering the hyperrectangle which encloses all the projected vertices. With respect to the *i*-th observation unit and the *p*-th component, we have the lower and the upper bounds, respectively, as

$$\underline{_i a_p} = \min\left(_i\mathbf{a}^p\right), \tag{32}$$

$$\overline{_i a_p} = \max\left(_i\mathbf{a}^p\right), \tag{33}$$

where $_i\mathbf{a}^p$ denotes the *p*-th column of $_i\mathbf{A}$.

In order to find exact low dimensional hyperrectangles we also propose the following procedure. Let $\mathbf{H}^C$ be the matrix of order $(2^P \times P)$ whose role is to define every vertex of the low dimensional hyperrectangles in $\Re^P$, similarly to **H** with respect to the hyperrectangles in $\Re^J$. The matrix **H** refers to the variable space, whereas $\mathbf{H}^C$ refers to the component space. As the rows of $\mathbf{A}_M$ and $\mathbf{A}_R$ provide the coordinates of, respectively, the midpoints and the radii in the low dimensional space spanned by **B**, the vertices of the low dimensional hyperrectangle pertaining to the generic *i*-th observation unit are

$$_i\mathbf{A} = \mathbf{1}^C \mathbf{A}_i^M + \mathbf{H}^C \mathbf{A}_i^R, \tag{34}$$

where $\mathbf{A}_i^M$ and $\mathbf{A}_i^R$ are diagonal matrices whose main diagonal elements are, respectively, those of $\mathbf{a}_{M\,i}$, the *i*-th row of $\mathbf{A}_M$, and $\mathbf{a}_{R\,i}$, the *i*-th row of $\mathbf{A}_R$. Finally $\mathbf{1}^C$ is a matrix of order $\left(2^P \times P\right)$ whose elements are 1's. It is worth to notice that the elements of $\mathbf{A}_i^R$ can be negative. It follows that the visualization tool loses the information about the signs of the elements in $\mathbf{A}_i^R$, i=1,…,I, but it has no effects from a graphical point of view. However we suggest, using the rotational freedom of MR-PCA, to find, if it exists, a columnwise orthonormal rotation matrix **W** (it very often suffices to use a diagonal matrix whose elements are ± 1) such that the coordinates of the radii are non negative provided that the rotation is compensated by postmultipling the loadings by $\left(\mathbf{W}'\right)^{-1}$.

5. Simulation study

In this section we give the results of a simulation study carried out in order to compare MR-PCA to C-PCA and V-PCA. Specifically, the simulation study aims at answering whether the compromise structure obtained by means of the MR-PCA method recovers better than C-PCA and V-PCA the underlying structure (the component loadings **B**) in the interval valued data. Notice that, in the C-PCA method, only the information pertaining to the midpoints is used in recovering the underlying structure of the interval valued data set involved. The information about the width of the intervals plays a relevant role just afterwards. In fact, the radii help us in determining the size of the low dimensional hyperrectangles associated to the observation units. Therefore, the simulation study also offers a comparison between MR-PCA and classical PCA applied on the midpoints.

Moreover, we investigate about the cases in which negative estimated radii occur. In these cases we apply the iterative algorithm. In fact, we study the computation time and the decrease of fit with respect to the solution of the non iterative procedure.

For each simulated data set, we randomly generate from the uniform distribution in $[0,1]$ the known component loadings matrix, say $\widetilde{\mathbf{B}}$, and the component scores matrices for the midpoints, say $\widetilde{\mathbf{A}}_M$, and for the radii, say $\widetilde{\mathbf{A}}_R$. Six levels of noise (n=0.1,0.3,0.5,1.0,1.5,2.0) are added to the obtained data. Therefore, we get

$$\mathbf{M}=\widetilde{\mathbf{A}}_M\widetilde{\mathbf{B}}'+n\mathbf{N}_M, \tag{35}$$

$$\mathbf{R}=\widetilde{\mathbf{A}}_R\widetilde{\mathbf{B}}'+n\mathbf{N}_R, \tag{36}$$

where $\mathbf{N}_M$ and $\mathbf{N}_R$ are randomly generated matrices from the uniform distribution in $[0,1]$ for which the following relations hold:

$$\left\|\widetilde{\mathbf{A}}_M\widetilde{\mathbf{B}}'\right\|^2=\left\|\mathbf{N}_M\right\|^2, \tag{37}$$

$$\left\|\widetilde{\mathbf{A}}_R\widetilde{\mathbf{B}}'\right\|^2=\left\|\mathbf{N}_R\right\|^2. \tag{38}$$

By means of (37)-(38), we are able to quantify exactly the level of added noise according to the values of the parameter n. We construct data sets with three different numbers of observation units (I=12,18,24) and variables (J=6,9,12) and we consider three relative sizes of the radii with respect to the midpoints (r=0.25,0.5,1.0). Finally, we use P=2,3 components. For each condition (a combination of the values pertaining to the five design variables), we generate ten data sets. Therefore, the simulation study is done on 3.240 data sets.

It remains to show how to evaluate whether one method works noticeably better than the other ones. Let $\mathbf{B}_C$, $\mathbf{B}_V$ and $\mathbf{B}_{MR}$ be the estimated component loadings matrix obtained performing, respectively, C-PCA, V-PCA and MR-PCA. To deal with the rotational freedom in the models, we first transform $\mathbf{B}_C$, $\mathbf{B}_V$ and $\mathbf{B}_{MR}$ in such a way that they are as similar as possible to $\widetilde{\mathbf{B}}$. In fact, we seek the matrices $\mathbf{W}_C$, $\mathbf{W}_V$ and $\mathbf{W}_{MR}$ for which, respectively, $\left\|\mathbf{B}_C\mathbf{W}_C-\widetilde{\mathbf{B}}\right\|^2=\min_{\mathbf{W}}\left\|\mathbf{B}_C\mathbf{W}-\widetilde{\mathbf{B}}\right\|^2$, $\left\|\mathbf{B}_V\mathbf{W}_V-\widetilde{\mathbf{B}}\right\|^2=\min_{\mathbf{W}}\left\|\mathbf{B}_V\mathbf{W}-\widetilde{\mathbf{B}}\right\|^2$ and $\left\|\mathbf{B}_{MR}\mathbf{W}_{MR}-\widetilde{\mathbf{B}}\right\|^2=\min_{\mathbf{W}}\left\|\mathbf{B}_{MR}\mathbf{W}-\widetilde{\mathbf{B}}\right\|^2$. The above minimization problems are simple regression problems that can be easily solved.

Let $\widetilde{\mathbf{B}}_C=\mathbf{B}_C\mathbf{W}_C$, $\widetilde{\mathbf{B}}_V=\mathbf{B}_V\mathbf{W}_V$ and $\widetilde{\mathbf{B}}_{MR}=\mathbf{B}_{MR}\mathbf{W}_{MR}$. In order to compare the estimated component loadings matrices with the known generated one, we consider the following index known as proportion of recovery (PR) measure (see, e.g., [17]):

$$PR=\left(1-\frac{\left\|\widetilde{\mathbf{B}}_*-\widetilde{\mathbf{B}}\right\|^2}{\left\|\widetilde{\mathbf{B}}\right\|^2}\right)100, \tag{39}$$

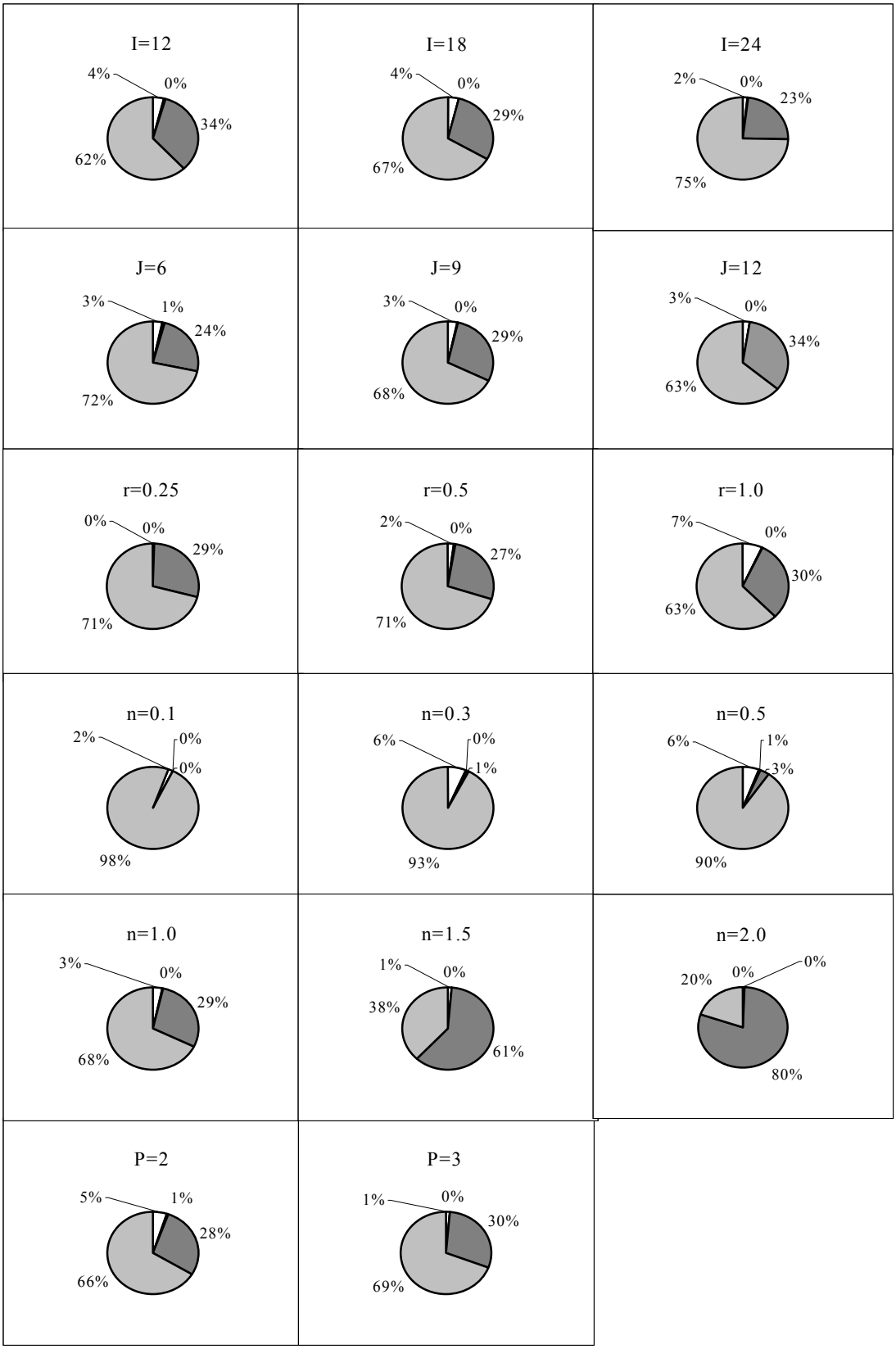
where $\widetilde{\mathbf{B}}_*$ denotes the rotated estimated component loadings matrix. The index in (39) takes values from 0 to 100. Values near to 100 show that the method at hand recovers the known component loadings matrix $\widetilde{\mathbf{B}}$ well.

5.1. Performance of the model

We assume that one method works noticeably better than the remaining two when the *PR* value differs more than 5% with respect to the ones pertaining to the remaining two methods. The results are displayed in Figure 1 for each level of each design variable separately. On average, in 28.8% of cases, MR-PCA recovers the underlying structure of the data better than the other methods, whereas C-PCA and V-PCA only in, respectively, 3.1% and 0.3% of cases.

Going into detail, we notice that MR-PCA works noticeably better than C-PCA except when the level of added noise is small. The size of the radii and the number of extracted components do not affect the results. The recovering performance of MR-PCA is better when the number of observation units decreases or the number of variables increases. The MR-PCA method performs very well when the level of noise added is higher than 0.5. In fact, on average, the *PR* values pertaining to the MR-PCA method are considerably higher in 56.2% of cases and C-PCA in 1.7% (V-PCA only in 0.2%).

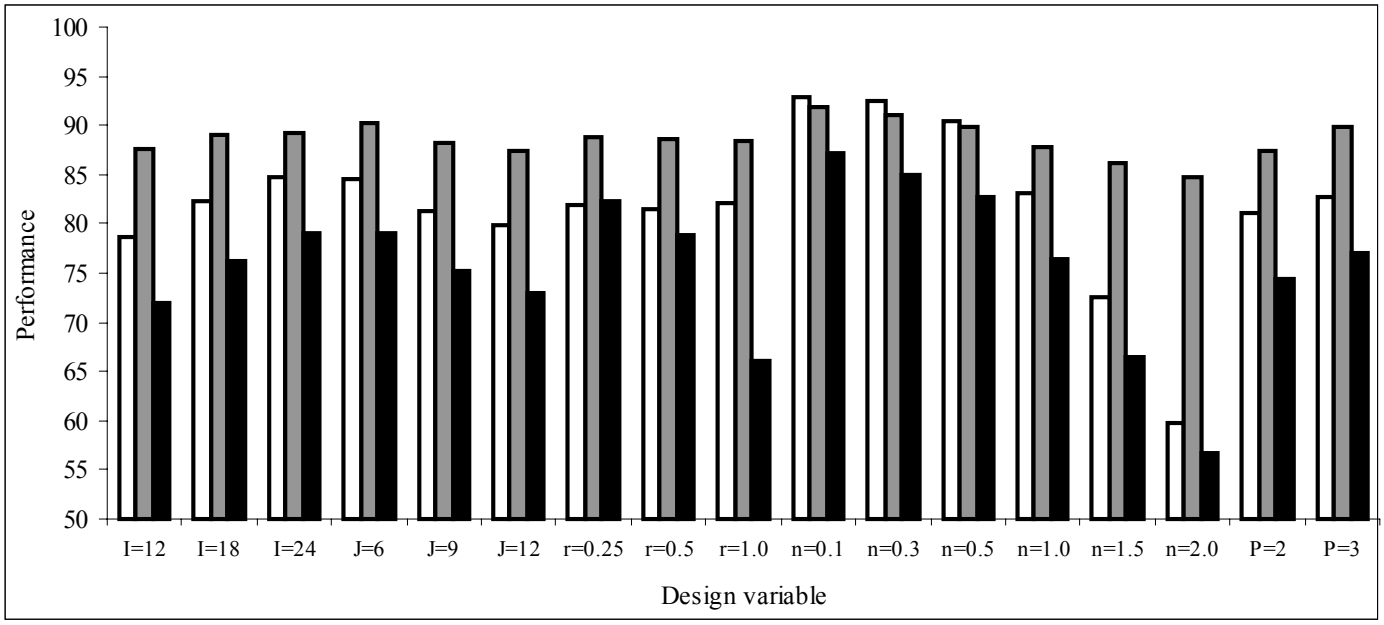
Figure 1: ***Percentage of times in which one method recovers the known component loadings matrix noticeably better than the others***



Note: White slices = C-PCA works better, Black slices = V-PCA works better, Dark Grey slices = MR-PCA works better, Light Grey slices = The methods work equally well.

Further details are given by considering Figure 2, which gives the average recovering performance of C-PCA, V-PCA and MR-PCA in terms of the PR index, for each level of each design variable separately. According to Figure 2, we can observe that the average *PR* value pertaining to the V-PCA method is always the lowest except for $r = 0.25$ when it is slightly higher than that pertaining to C-PCA. The average *PR* values pertaining to the MR-PCA method are very often higher than those pertaining to C-PCA. Only when the levels of added noise are small, the average C-PCA values are slightly higher. The average *PR* value for MR-PCA is 88.6% versus 81.9% for C-PCA and 75.7% for V-PCA. With respect to the MR-PCA method, the average values seem to not depend on the number of observation units, the number of variables and the size of the radii. The highest difference between the performance of MR-PCA and that of the remaining methods is attained when the level of added noise is the highest (n=2.0).

Figure 2: ***Recovering performance of C-PCA (or PCA) in white, V-PCA in black and MR-PCA in grey, in terms of the PR index.***



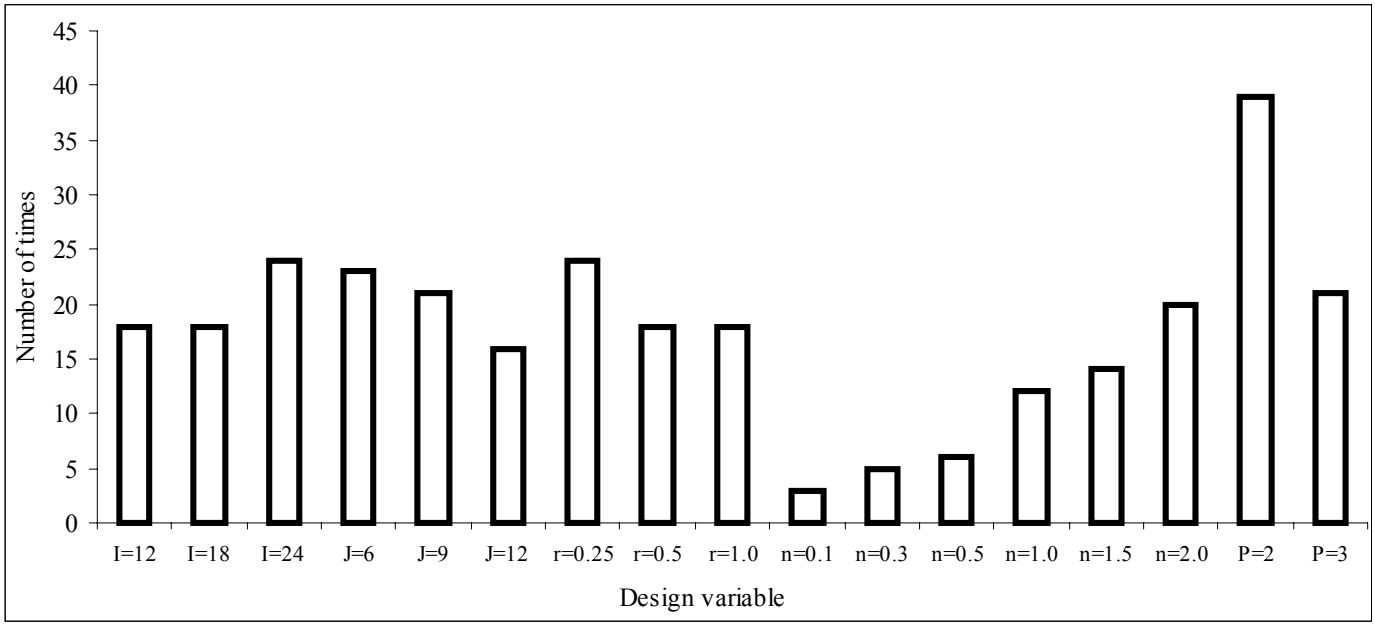
Summing up, we can argue that the MR-PCA method works better than C-PCA and V-PCA especially when the level of added noise is high. In the remaining conditions, the MR-PCA method should be preferable. However, the different levels of the remaining design variables slightly affect the performance of the method.

5.2. Negative estimated radii

In this subsection, we aim at studying the occurrences in which the estimates of the radii are negative by means of the non-iterative algorithm given in Section 3.3.

Out of 3.240 randomly generated data sets, we obtain negative estimates only 60 times (1.9% of cases). Further details can be found in Figure 3.

Figure 3: ***Number of times in which negative estimates occur***

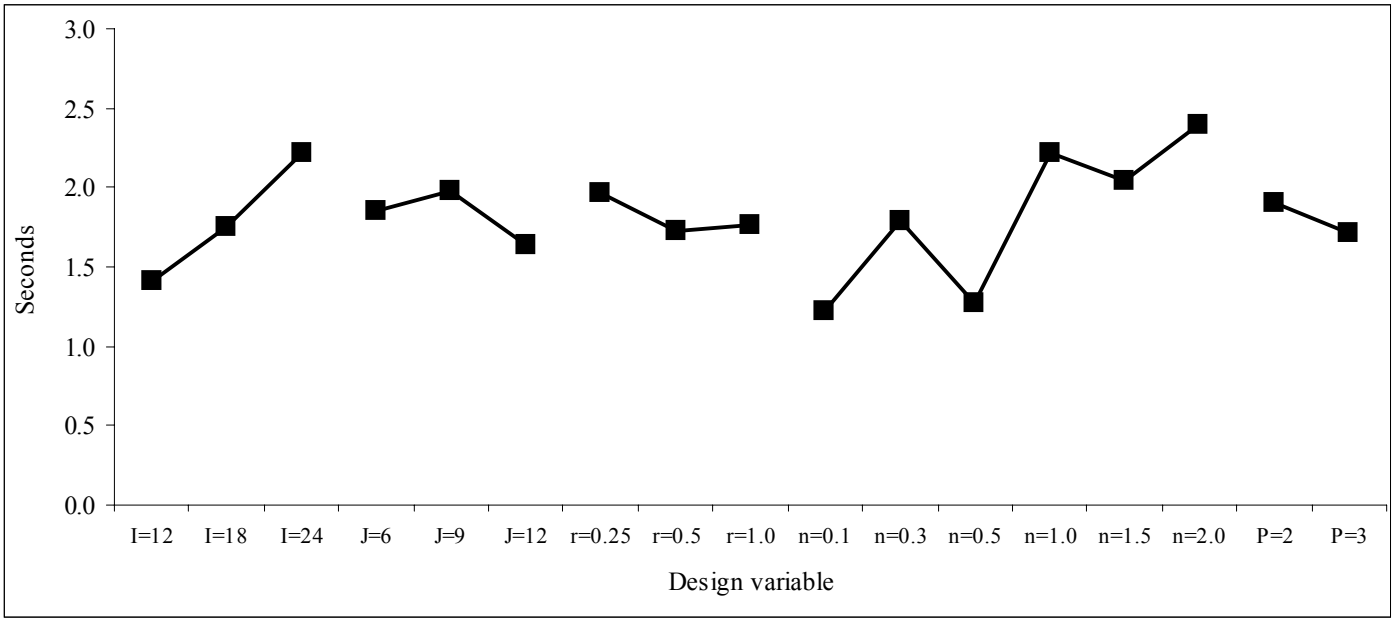


The number of times in which negative estimates of the radii occur increases when the number of observation units increases or the number of variables decreases. As it should be expected, when the size of the radii is small, the risk of negative estimates increases. In fact, as the generated radii are nearer to zero, the possibility to obtain negative estimates increases. Analogously, when the level of added noise increases or the number of extracted components decreases, the number of times in which negative estimates occur increases. In both cases, the fit of the model decreases and, as a consequence, the risk of irregular estimates – in particular, negative estimates - increases.

Whenever negative estimates of the radii occur, we run the row-wise ALS algorithm given in Section 3.3 considering as starting point the non-iterative solution and replacing negative component loadings and component scores for the radii by randomly generated numbers from the uniform distribution in $[0,1]$ in order to deal with a feasible solution. The need for the iterative algorithm leads to a negligible decrease of fit (0.02% on average according to the index in (28)).

The computation time of the algorithm is very low. In fact, the average computation time is 1.8 seconds. It seems to be affected only by the number of observation units and by the level of added noise as we can observe in Figure 4.

Figure 4: ***Computation time of the iterative algorithm***



To sum up, the possibility of obtaining negative estimates of the radii is very low. In such unlucky occurrences, the iterative algorithm works very well: it converges quickly and the decrease of fit is very small.

6. Applications

In this section, we give the results of two applications of MR-PCA on chemical data. In the first one, the solution is obtained running the non-iterative algorithm whereas, in the second one, the iterative algorithm is necessary because negative estimates of the radii occur.

6.1 Portuguese mineral water data

In this subsection the MR-PCA method is performed on a data set describing I=5 Portuguese mineral waters characterized by J=6 interval valued variables. In particular, the variables are

mineral concentrations of HCO_3^-, Cl^-, Na^+, Ca^{2+} and SiO_2 (mg/l) and the PH value. The data are summarized in Table 1.

Table 1: ***Portuguese mineral water data***

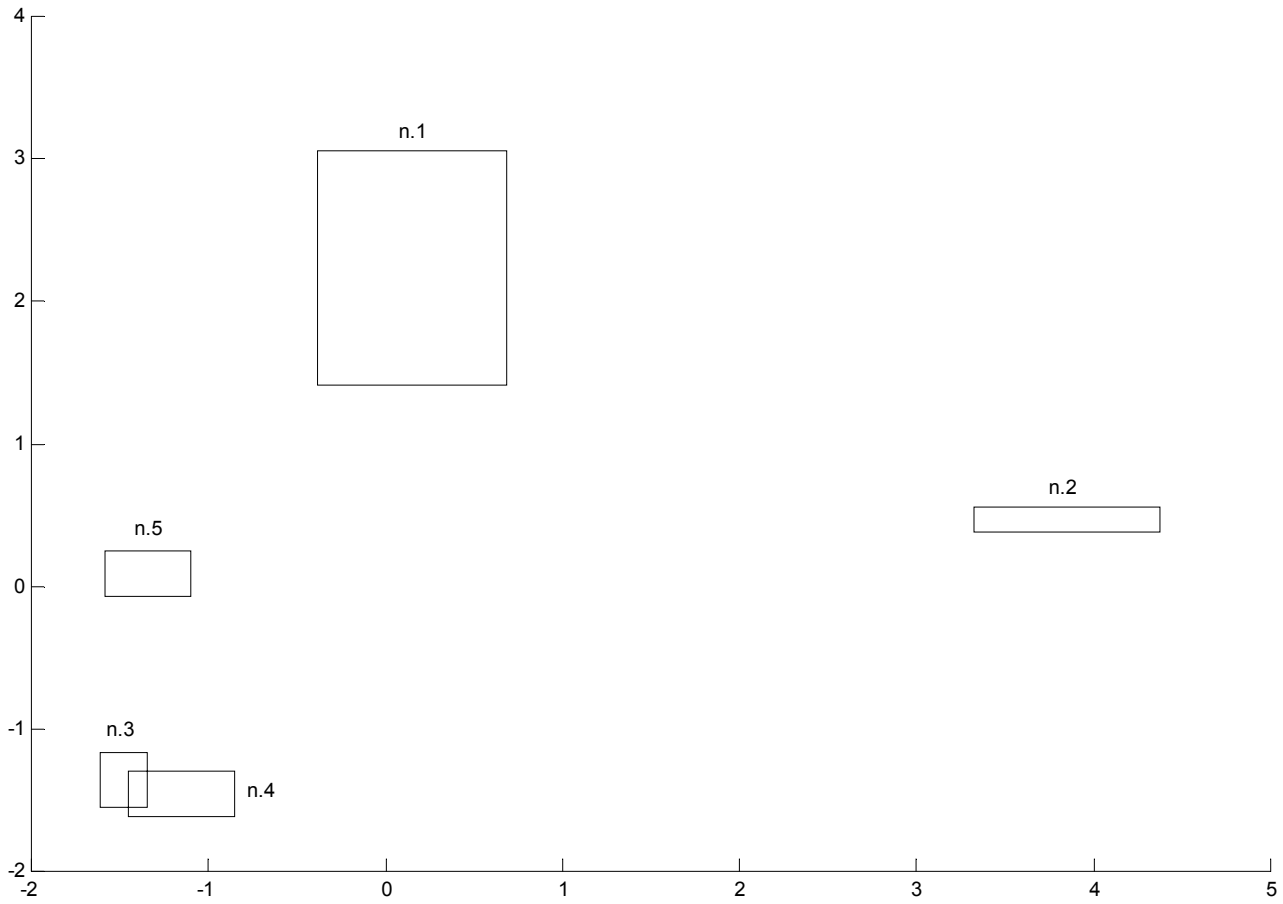
Water	HCO_3^-	Cl^-	Na^+	Ca^{2+}	SiO_2	PH
n.1	[21,41]	[7,9]	[10,16]	[3,4]	[23,29]	[6.1,6.5]
n.2	[113,119]	[16.5,17.5]	[10.3,10.7]	[15,21]	[13.7,14.9]	[6.7,7.1]
n.3	[2.2,4.2]	[3.6,4]	[2.8,3.8]	[0.01,1.01]	[5.8,7.8]	[5.71,5.81]
n.4	[8,11.6]	[4.1,4.7]	[2.8,3.6]	[1.9,2.9]	[5.8,6.8]	[5.6,6]
n.5	[4.6,5]	[6.6,7.4]	[5.4,5.6]	[0.72,0.84]	[16.7,18.3]	[5.4,5.8]

After preprocessing the data as described in Remark 1, the MR-PCA method is performed extracting P=2 components. The non-iterative procedure is run. In fact, the estimates of the radii are non negative. We decide to extract P=2 components because the goodness of fit is very high (following (28), 96.6%) and the solution is easily interpretable. Notice that each variable plays a relevant role in exactly one component. In order to interpret the solution, the matrix of the varimax-rotated component loadings and the low dimensional representation of the waters in the low dimensional space spanned by the (orthonormal) loadings are given, respectively, in Table 2 and Figure 5. Notice that the low dimensional configuration is obtained using (34) and that the component scores matrix for the radii has all non negative elements.

Table 2: ***Component loadings matrix***

Mineral	PC1	PC2
HCO_3^-	0.50	-0.05
Cl^-	0.46	0.05
Na^+	0.16	0.62
Ca^{2+}	0.53	-0.14
SiO_2	-0.10	0.77
PH	0.47	0.11

Figure 5: ***Low dimensional representation of the Portuguese waters***



From Table 2, we can easily assess the role of the original variables in describing the extracted components. The first component is strongly related to the concentrations of HCO_3^-, Cl^- and Ca^{2+} and the values of PH. The remaining variables (Na^+, SiO_2) have a negligible influence. On the contrary, Na^+ and SiO_2 help us to define the second component. In fact, high second component scores for the midpoints depend on high values of Na^+ and SiO_2.

The low dimensional representation (as rectangles because $P=2$) of the waters is consistent with the above interpretation of the components. The position of Water n.2 can be explained by considering the values of PH and, above all, HCO_3^-, Cl^- and Ca^{2+} which are sensibly higher than the ones pertaining to the remaining waters. The same comments hold with respect to the second component and Water n.1. In fact, its position reflects the values of Na^+ and, above all, SiO_2. Notice that Waters n.3 and n.4 are overlapped. This can be explained by the observed interval valued scores of Na^+, SiO_2 and PH for Waters n.3 and n.4 which have not empty intersections.

Figure 5 provides useful information by considering the size of each rectangle which gives a measure of the uncertainty associated with each observation unit. Water n.1 is characterized by the biggest rectangle. In fact, the radii of the intervals pertaining to Water n.1 are the highest ones

except for Ca^{2+}. Water n.2 has the highest first component score for the radii. It depends on the radius of Ca^{2+}. Thus, the component scores for the radii (e.g. the size of the low dimensional hyperrectangles) can be explained by considering the widths of the interval valued variables, which play a relevant role in interpreting the component at hand. It follows that the aim of the low dimensional configuration of the observation units is to provide information about not only the similarities among the observation units but also the uncertainty associated to each observation unit. Hence, the structure of the observation unit mode entities (as low dimensional hyperrectangles) in the principal space can be detected considering their position, their size and, eventually, their overlap.

6.2. Greek wine data

In this subsection we illustrate the application of the MR-PCA method on Greek wine samples [7]. Specifically, the data set involved describes I=33 Greek red and white commercial wines from the 1998 vintage characterized by J=9 mineral concentrations (K^+, Na^+, Ca^{2+}, Mg^{2+}, Fe^{3+}, Cu^{2+}, Mn^{2+}, Zn^{2+}, P^{5+}). The available information for the *i*-th wine with respect to the *j*-th mineral concentration is the interval $\left(m_{ij}-r_{ij}, m_{ij}+r_{ij}\right)$ where m_{ij} denotes the mean value and r_{ij} the standard deviation.

In [7] two different PCA's based on red wines and, separately, on white wines are performed. The PCA's were carried out considering the midpoints only. It can be observed, as it is already noticed in [7], that the features of red and white wines are very similar. With respect to the red wines, $P=3$ components are extracted: the first one is highly related to Fe^{3+}, Cu^{2+}, and Na^+, whereas the second and the third ones to Ca^{2+} and Mg^{2+}. Similar results were obtained from PCA of white wines.

In [7], the authors did not describe how they preprocessed the data. Here, we preprocess them according to the procedure described in Remark 1.

After preprocessing the data, we perform the MR-PCA method. We extract P=3 components. As some estimates of the radii take negative values, we run the iterative algorithm in order to find a feasible solution. According to the index in (28), the goodness of fit of the model is 71.8% (58.9% if P=2 and 79.1% if P=4). We rotate the components to simple structure. The varimax-rotated component loadings matrix is summarized in Table 3.

Table 3: ***Component loadings matrix***

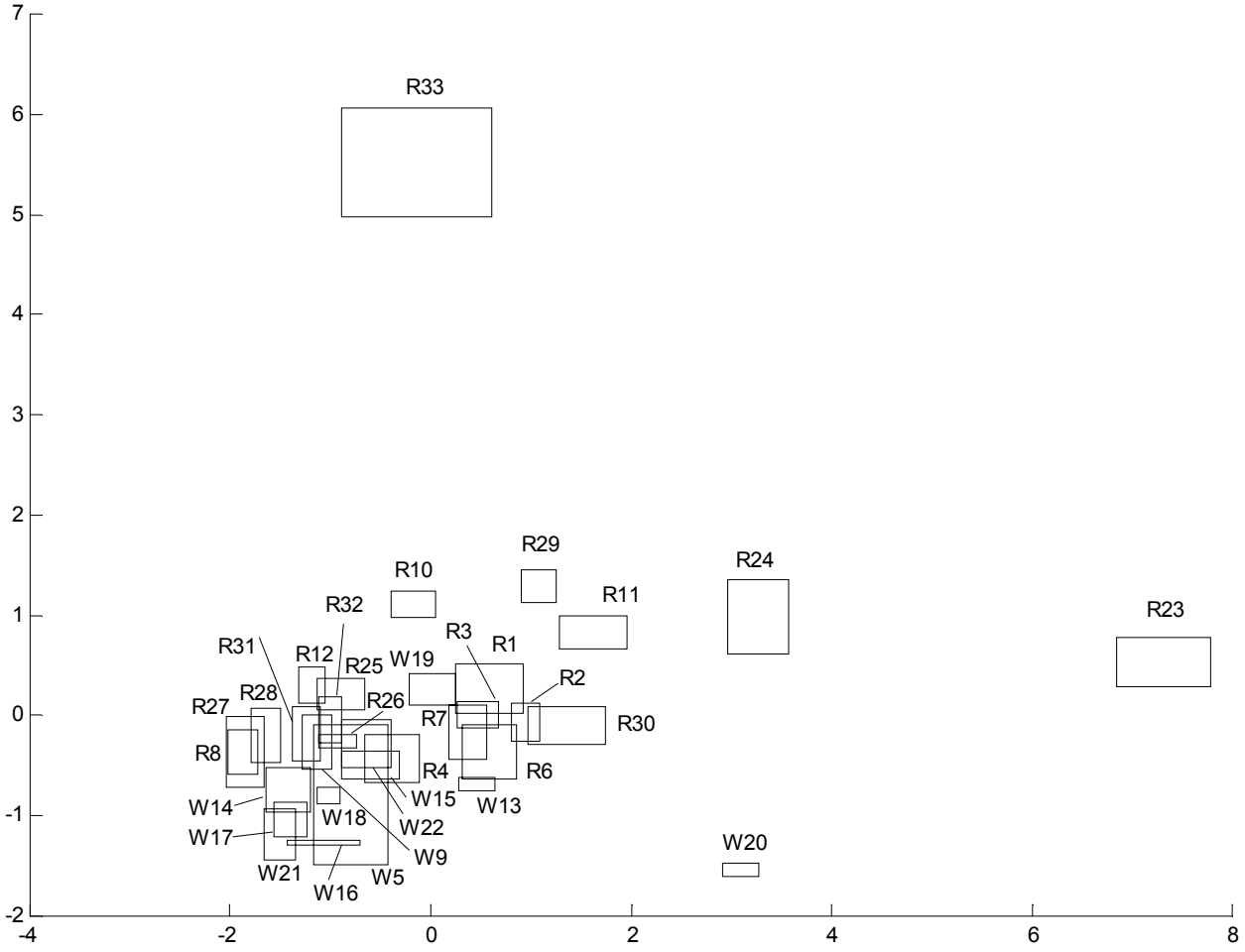
Mineral	PC1	PC2	PC3
K^+	0.02	0.39	0.39
Na^+	-0.09	-0.07	0.68
Ca^{2+}	0.49	-0.25	-0.01
Mg^{2+}	0.39	0.27	-0.19
Fe^{3+}	0.42	0.25	-0.06
Cu^{2+}	-0.03	0.79	0.03
Mn^{2+}	0.41	-0.16	0.16
Zn^{2+}	0.48	-0.04	0.02
P^{5+}	0.14	0.03	0.56

From Table 3, we can observe that the first component is positively related to high values of Ca^{2+}, Mg^{2+}, Fe^{3+}, Mn^{2+} and Zn^{2+}. The second component loadings show the importance of Cu^{2+} and, to a lesser extent, K^+, Mg^{2+}, Fe^{3+}, Ca^{2+} (the last one is inversely related to the axis). Finally, except for K^+, the minerals Na^+ and P^{5+} that play a slight role in describing the first two components are strictly related to the third one.

In Figures 6 and 7, we provide the low dimensional representations of the wines according to (34). Notice that the component scores matrix for the radii has all non negative elements. Each wine is recognized by a letter ('W' if it is a white wine, 'R' if it is red) and by a number (from 1 to 33 following [7]). In Figure 6, we plot the wines on the two dimensional space spanned by the first and the second components. We can observe that the white wines are on the low side and the red ones on the high side. Thus, the second component can be interpreted as the type of wine. Just one white wine ('W19' from Kefalinia) is not consistent with the interpretation of the second component.

Further information about the interpretation of the axes is provided by Figure 7 in which the wines are represented with respect to the second and third components. The third component seems to be related to the geographical position of the production areas. More specifically, positive scores pertain to wines the geographical origin of which is South Greece or Greek islands while negative scores pertain to wines from North Greece. Three areas of production are not consistent with the above distinction. In fact, the scores of the wines from Rapsani (in the geographical area above Peloponnese, belonging to the North Greece group) are higher than 0 and the wines from the island of Crete (three out of four) and those from Mantinia (two out of three) have negative scores. We also notice that the wines produced in the island of Kefalinia have negative scores but this is consistent with the island position.

Figure 6: ***Low dimensional representation of the Greek wines (PC1 vs PC2)***



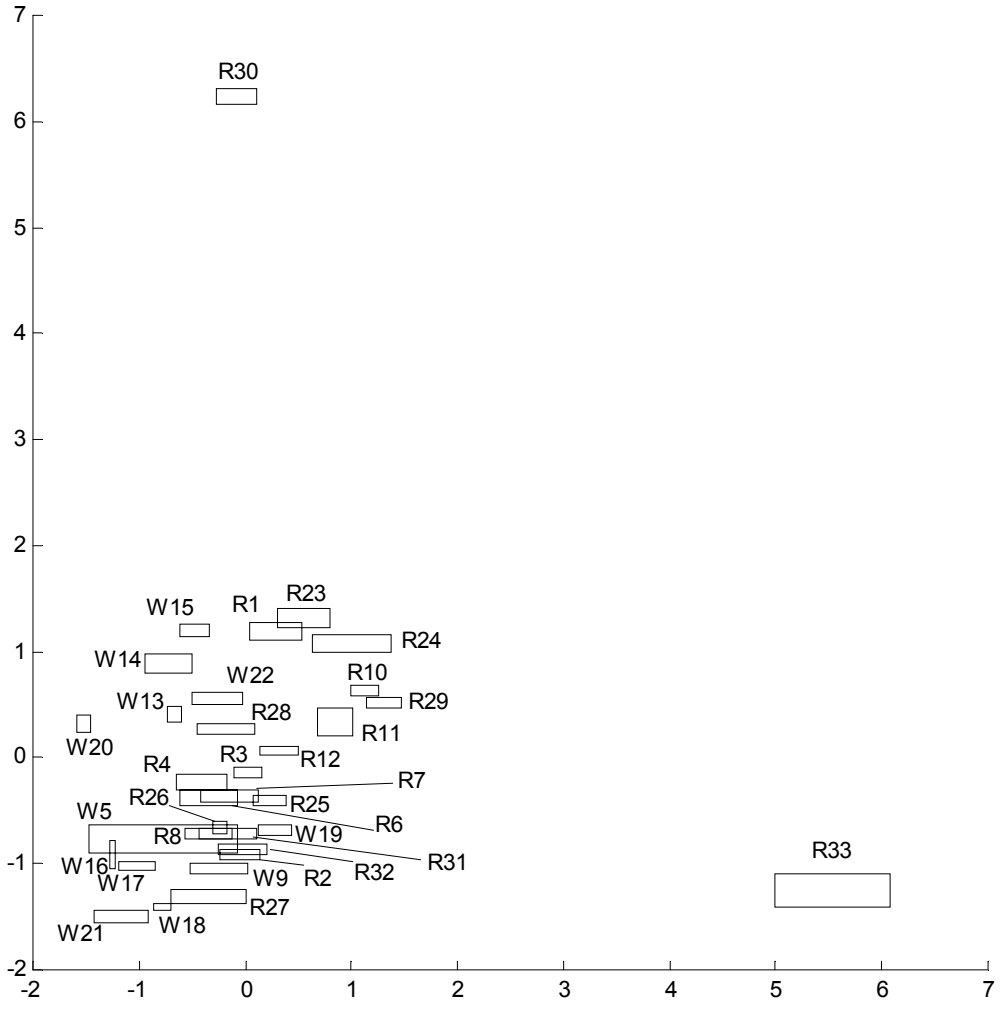
Finally, we inspect Figure 6 again, and see that the first component seems to not be discriminated by geographical origin, variety or type. It only provides a measure of the dissimilarities among the (red and white) wines with respect to the mineral concentrations that play a relevant role in interpreting this component. For instance, notice that 'R23' is very far from the remaining wines. This can be explained considering that its observed scores are the highest ones among all the wines. In fact, this observation unit can be considered an outlier.

Let us now consider the uncertainty associated to the red and white wines in the obtained low dimensional space. With respect to the first component, we observe that the highest low dimensional scores for the radii pertain to 'W5', 'W16', 'R23', 'R30' and 'R33'. By considering the original data set and the loadings (Table 3) we can state that the size of 'W5' and 'R23' depends on high radii of Mg^{2+} (for 'R23', also of Zn^{2+}); the component score for the radii of 'W16' is related to Mn^{2+}. The component scores for the radii of 'R30' and 'R33' are affected by Fe^{3+} ('R30' also by Ca^{2+}).

Again, ‘W5’ and ‘R33’ have high second component scores for the radii. The score of ‘R33’ can be explained by the radius of Fe^{3+} and that of ‘W5’ by the radii of Mg^{2+} and Cu^{2+}. The component score for the radii of ‘W5’ is related to Cu^{2+} (in the data set, the associated uncertainty is the highest one). As well as for ‘W5’ and ‘R33’, the second component score for the radii of ‘R24’ is high. It depends on the radius of Fe^{3+}.

With respect to the uncertainty associated to the third component, three wines can be well distinguished: ‘W5’ and ‘R11’ (both wines have high radii of P^{5+}), ‘W16’ and ‘R33’.

Figure 7: ***Low dimensional representation of the Greek wines (PC2 vs PC3)***



In conclusion, this application shows that the components are able to distinguish the wines according to the type (red or white) of wine (second component) and to the geographical position (third component) whereas the first component reflects the chemical differentation of wines. Moreover, the size of the hyperrectangles in the low dimensional space provides a measure of the uncertainty associated to the registered mineral concentrations. Specifically, for each component,

the scores for the radii give information about the uncertainty of the mineral concentrations the role of which in interpreting the component at hand is relevant

7. Final remarks

In this paper we proposed a Principal Component method for two-way interval valued data matrices (observation units × interval valued variables) based on a least squares approach. The suggested method, called Midpoint Radius Principal Component Analysis (MR-PCA), is capable to find the underlying structure of the interval valued data by using the midpoints and the radii information. Moreover, in order to analyze how our method works, the results of a simulation study and two chemical data applications have been shown.

In future works, it will be interesting to extend our data reduction method to a three-way framework, in order to synthesize three-way data arrays (i.e. observation units × interval valued variables × occasions) by means of three-way methods such as Tucker3 [18] and PARAFAC [19]. In this respect, it will be also attractive to analyze the situation in which the occasions are times, i.e. time intervals, extending the so called Dynamic Factor Analysis [20, 21] to interval valued time arrays. Moreover, other possible extensions may concern the suggestion, in the two-way as well as in the three-way framework, to make use of interval arithmetic (see, e.g. [22,23]) in suitably generalizing principal component methods to deal with interval valued data. These will be the main issues of our future research in this field.

References

[1] Bock, H.H., and Diday, E., (Eds.), Analysis of symbolic data: exploratory methods for extracting statistical information from complex data, Springer Verlag, Heidelberg, 2000.

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

[3] D'Urso, P., and Giordani, P., Fitting of fuzzy linear regression models with multivariate response, International Mathematical Journal, 3 (2003) 655-664.

[4] Giordani, P., and Kiers, H.A.L., Principal Component Analysis of symmetric fuzzy data, Computational Statistics and Data Analysis (2003) in press.

[5] Guo, Q., Wu, W., Massart, D.L., Boucon, C., and de Jong, S., Feature selection in principal component analysis of analytical data, Chemometrics and Intelligent Laboratory Systems, 61

(2002) 123-132.
[6] Estienne, F., Matthijs, N., Massart, D.L., Ricoux, P., and Leibovici, D., Multi-way modelling of high-dimensionality electroencephalografic data, Chemometrics and Intelligent Laboratory Systems, 58 (2001) 59-72.
[7] Kallithraka, S., Arvanitoyannis, I.S., Kefalas, P., El-Zajouli, A., Soufleros, E., and Psarra, E., Instrumental and sensory analysis of Greek wines; implementation of principal component analysis (PCA) for classification according to geographical origin, Food Chemistry, 73 (2001) 501-514.
[8] Lauro, C., and Palumbo, F., Principal component analysis of interval data: a symbolic data analysis approach, Computational Statistics, 15 (2000) 73-87.
[9] Palumbo, F., and Lauro, C., A PCA for interval-valued data based on midpoints and radii. In: 'New Developments on Psychometrics: Proceedings of the International Meeting of the Psychometric Society (IMPS 2001)'. Springer Verlag, Tokyo, 2003.
[10] Millsap, R.E., and Meredith, W., Component analysis in cross-sectional and longitudinal data, Psychometrika, 53 (1988) 123-134.
[11] Kiers, H.A.L., and ten Berge, J.M.F., Alternating least squares algorithms for Simultaneous Components Analysis with equal component weight matrices for all populations, Psychometrika, 54 (1989) 467-473.
[12] Kiers, H.A.L., and ten Berge, J.M.F., Hierarchical relations between methods for simultaneous component analysis and a technique for rotation to a simple simultaneous structure, British Journal of Mathematical and Statistical Psychology, 47 (1994) 109-126.
[13] Lodwick, W.A., and Jamison, K.D., Special issue:interfaces between fuzzy set theory and interval analysis, Fuzzy Sets and Systems, 135 (2003) 1–3.
[14] D'Urso, P., and Gastaldi, T., A least-squares approach to fuzzy linear regression analysis, Computational Statistics and Data Analysis, 34 (2000) 427-440.
[15] Lawson, C.L., and Hanson, R.J., Classics in applied mathematics, Vol. 15, SIAM, Philadelphia, PA, 1995.
[16] Kiers, H.A.L., Some procedures for displaying results from three-way methods, Journal of Chemometrics, 14 (2000) 151-170.
[17] Timmerman, M.E., and Kiers, H.A.L., Three-way component analysis with smoothness constraints, Computational Statistics and Data Analysis, 40 (2002) 447-470.
[18] Tucker, L.R., Some mathematical notes on three-mode factor analysis, Psychometrika, 31 (1966) 279-311.
[19] Harshman, R.A., Foundations of the PARAFAC procedure: models and conditions for an 'exploratory' multi-mode factor analysis, UCLA Working papers in phonetics, 16 (1970) 1-84.
[20] Coppi, R., Blanco, J., Camaño, G., and Corazziari, I., Descomposición Factorial y Regresiva de la Variabilidad de un Array a tres Vias, Quantum, 4, 10 (1999) 81-107.
[21] Coppi, R., and D'Urso, P., The Dual Dynamic Factor Analysis Model, in: Classification, Automation, and New Media (eds. W. Gaul, G. Ritter), Springer-Verlag, Heidelberg (2002) 47-58.
[22] http://www.cs.utep.edu/interval-comp/main.html (Web site regarding interval computations topics and interval computations researchers).
[23] Alefeld, G., and Herzberger, J., Introduction to interval computations, Academic Press, New York, 1983.