doi.org/10.26398/IJAS.0033-013

PRINCIPAL COMPONENT ANALYSIS FOR INTERVAL DATA: COMMON APPROACHES AND VARIATIONS

Alfonso Iodice D'Enza[1], Viviana Schisa, and Francesco Palumbo

Department of Political Sciences, Università degli Studi di Napoli Federico II, Naples, Italy

Abstract. *In real life there are many kinds of phenomena that are better described by interval bounds than by single-valued variables. In fact, intervals take into account the imprecision due to measurement errors. When there is information about the imprecision distribution the fuzzy data coding is used to represent the imprecision. In this paper, we first review the main dimension reduction techniques for interval-valued data and then we propose a midpoints and radii-based approach. In particular, an alternative pre-processing and Procrustean rotation of the traditional midpoints and radii approach is proposed.*

Keywords: *Interval-Valued Data; Principal Component Analysis.*

1. INTRODUCTION

Unsupervised learning is a general definition that refers to problems where no observable response is available, yet it is implicitly assumed that some underlying quantitative/qualitative latent response(s) provide a meaningful synthesis of the available data. Therefore, unsupervised learning methods aim to gather information from the analyzed data by estimating one (or more) latent responses using the observed attributes.

In modern applications, however, the information at hand can hardly be coded into a classic *observations* by *attributes* structure; hence, more complex data coding and structures that can be referred to as symbolic data aid in better exploiting the information from data. Symbolic data encompasses interval-valued variables, multi-valued variables and other set-valued variables, where data cells contain sets of categories, ranges (intervals), or weights (Billard and Diday, 2000; Bock and Diday, 2000; Diday, 1988). Therefore, interval data are a special kind of symbolic data, in which only quantitative variables are considered and represent a viable coding of information concerning complex phenomena. An interval data matrix has a complex structure because each cell does not contain a single value

[1] Corresponding author: Alfonso Iodice, D'Enza email: iodicede@unina.it

but a pair of values (center and range, or min-max). This type of data can derive from several sources: in chemometrics, it is possible to study mineral concentrations in food products in different experimental situations; in meteorology, the daily temperature, humidity, and wind speed can be recorded in different places; in environmental sciences, concentrations of pollutants can be recorded in various locations; in finance, the exchange rate euro-dollar varies throughout the day; in medicine, daily systolic and diastolic pressure, heart frequency and temperature fluctuate in a range. In all the above cases, it is often more interesting to consider the minimum and maximum values for each variable rather than the average value, which would cause the loss of information. In fact, ranges refer to the variability of phenomena measured within each observation, and different application fields could benefit from interval data analysis, such as behavioral analysis, weather forecasts, statistical quality control, and financial analysis. Several other conditions exist in the numerical coding of variables, where interval-valued variables can represent the real world better than single-valued variables. A very appropriate condition arises when data are naturally interval-valued: this is typical when describing living species or giving a product specification (Billard and Diday, 2003; Billard and Diday , 2007; Bock and Diday, 1999; D'Esposito et al., 2012).

The *fuzzy sets theory* offers alternative approaches to account for the variability within each statistical unit. The extensive related scientific literature introduced diverse definitions of fuzzy numbers, yet some of the definitions some are closely related to the concepts of uncertainty and intervals (readers may refer to Dubois, 1980; Dubois and Prade, 1993). The so-called interval algebra represents the first attempt to deal with complex data coded as interval-valued variables: in particular, interval-valued algebra was introduced to deal with the round-off issue of the fixed-point computer processors. Due to the specific nature of the problem, interval-valued algebra was meant to deal with small intervals; for that reason, it was of little or non-use in statistical analysis (see, e.g., Ferson et al., 2002; Marino and Palumbo, 2002).

Symbolic Data Analysis (SDA) (Diday, 1988) aims to extend classical unsupervised (and, supervised) methods to complex data structures. Since the publication of the book edited by Bock and Diday (2000), several authors contributed to the growth of the SDA framework, that, as of today, covers several topics with hundreds of published scientific papers on clustering methods (Billard and Diday , 2019), dimensionality reduction techniques (Nagabhushan et al., 1995), decision trees (Mballo and Diday, 2005).

This paper focuses on Principal Component Analysis (PCA, see, e.g., Jolliffe, 2005, for a thorough discussion) for interval data, in particular on the most used approaches such as Centers PCA and Vertices PCA (respectively C-PCA and V-PCA, Cazes et al., 1997). Furthermore, an enhancement for Midpoints and Radii Principal Component Analysis (MR-PCA Palumbo and Lauro, 2003) is proposed that exploits a symbolic variance-based pre-processing, and a Procrustean analysis-based rotation of the data structures. In particular, the pre-processing takes into account the center/range covariance structure, and the rotation of the ranges preserves the general center/ranges geometric structure. The methods are applied to synthetic and real data sets. It is worth noting that all of the reviewed methods can also be helpful to analyze data coded as fuzzy numbers via the triangular membership function.
The paper is structured as follows: in Section 2 we briefly recall the general concepts for PCA of single-valued quantitative data; Section 3 discusses interval-valued data pre-processing alternatives. Section 4 reviews the best known in the literature PCA methods for interval-valued data; in Section 5 the so-called midpoint and radii PCA is presented and reconsidered; in Section 6 we show the results of our method applied to simulated data and the facial recognition data set. The last Section is for discussion and it concludes the paper.

2. PCA FOR SINGLE-VALUED DATA

The best-known linear dimension reduction method for single-valued quantitative variables is the Principal Component Analysis (PCA, see, e.g., Jolliffe, 2005, for a thorough discussion). PCA serves different tasks, from data visualization to feature extraction and data compression. In general, starting from p variables, the PCA aims to define $q \ll p$ *principal components* (PC's) that are linear combinations of the p original variables; and explain most of the original variability.
Formally, let $\mathbf{X}$ be the $(n \times p)$ data matrix, with statistical units on rows and single-valued variables on columns. We assume $\mathbf{X}$ to be pre-transformed, centered, and standardized for simplicity and without loss of generality. The transformed data matrix can be re-written, therefore, as

$$\mathbf{X} = \mathbf{A}\mathbf{B}' + \mathbf{E}$$

where $\mathbf{A}$ $(n \times q)$ and $\mathbf{B}$ $(p \times q)$ are the so-called components scores and component loadings matrices, respectively; $\mathbf{E}$ is the $(n \times p)$ residual matrix. For a specified value of q, the analysis goal is to find the matrices $\mathbf{A}$ and $\mathbf{B}$ that minimize the

squared residuals.
Based on the Eckart and Young (1936) theorem, upon defining

$$\begin{aligned} \mathbf{A} &= \mathbf{U}\mathbf{D}_\lambda \\ \mathbf{B} &= \mathbf{V}, \end{aligned}$$

with $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{D}_\lambda$ obtained via the singular value decomposition of $\mathbf{X}$; it results in that

$$\mathbf{A}\mathbf{B}' \approx \mathbf{X}$$

that is the best rank-q approximation of $\mathbf{X}$, in the least-squares sense. In other words, the PCA solution boils down to the SVD of the centered and standardized data matrix. Due to the relation between the SVD and the eigenvalue decomposition (EVD) of $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}\mathbf{X}'$, the singular vectors in $\mathbf{U}$ and $\mathbf{V}$ are the eigenvectors of

$$\begin{aligned} \mathbf{X}\mathbf{X}' &= \mathbf{U}\mathbf{D}_\lambda^{*2}\mathbf{U}' \\ \mathbf{X}'\mathbf{X} &= \mathbf{V}\mathbf{D}_\lambda^{2}\mathbf{V}' \end{aligned}$$

and $\mathbf{D}_\lambda^{*2}$, $\mathbf{D}_\lambda^2$ are diagonal matrices of eigenvalues, that are the singular values of $\mathbf{X}$, squared. Also, $\mathbf{D}_\lambda^{*2}$ and $\mathbf{D}_\lambda^2$ have the same non-zero elements.
Due to the correspondence between the above eigendecompositions and the singular value decomposition, the component matrices $\mathbf{A}$ and $\mathbf{B}$ can be obtained by either the SVD of $\mathbf{X}$ or the EVD of $\mathbf{X}'\mathbf{X}$, to get $\mathbf{V}$ and $\mathbf{D}_\lambda$ and then use the transition formula to get $\mathbf{U} = \mathbf{X}\mathbf{V}\mathbf{D}_\lambda$.
In PCA, the point-wise assessment of the obtained solution is measured by the so-called absolute contributions and relative contributions (or squared cosine). In particular, the absolute contribution $ctr\_abs_{i\alpha}$ of a point i to the α-th PC, indicates the influence of that point on the solution. The relative contribution $ctr\_rel_{i\alpha}$ of the point i on the α-th PC measures the quality of the representation of the point on that axis. Formally, let $a_{i\alpha}$ be the score of i on the α-th PC, then the absolute contribution is

$$ctr\_abs_{i\alpha} = \frac{a_{i\alpha}^2}{n\lambda_\alpha}.$$

Note that, since λ_α is the variance explained by the α-th component, and since $a_\alpha = \frac{1}{n}\sum_{i=1}^{n} a_{i\alpha} = 0$, then the quantity $\frac{a_{i\alpha}^2}{n}$ is, in fact, the contribution of the i-th

point to the variability of the α-th PC. The relative contribution $ctr\_rel_{i\alpha}$ represents the ratio between the length (squared norm) of the projection of i on the α-th PC (that is, $a_{i\alpha}$) and the full-dimensional length of observation i, which is given by the squared norm of the i-th row of $\mathbf{X}$, indicated by $\mathbf{x}_i$. Formally,

$$ctr\_rel_{i\alpha} = \frac{\|a_{i\alpha}\|^2}{\|\mathbf{x}_i\|^2}.$$

Both the absolute and relative contribution indexes take value in $[0,1]$.

3. INTERVAL-VALUED DATA CODING

Let $[x]$ be a closed interval in $\mathbb{R}$, such that $[x] \subset \mathbb{R}$, and that $[x] \equiv [\underline{x}, \overline{x}]$, with $\overline{x} \geq \underline{x}$. Then an interval-valued variable $[X]$ is defined as $[X] = [x]_1, [x]_2, \cdots [x]_i, \cdots [x]_N$. Statistical methods to deal with interval-valued variables require that intervals are defined through mathematical entities that are numerically tractable.
The *interval algebra* approach postulates that the knowledge about the interval is limited to its extremes values: min and max, which have been denoted with $\underline{x}$ and $\overline{x}$, respectively. The generic interval value $[x]_i$ is defined as

$$[x]_i = [\underline{x}_i, \overline{x}_i] \qquad\qquad i = 1, \ldots, N.$$

However, under the interval algebra paradigm, $[x]_i$ can equivalently be represented in the center x_i^c and the range x_i^r (also called midpoints-*radii*) notation. So the interval $[x]_i \equiv \{x_i^c, x_i^r\}$, where

$$\begin{aligned} x_i^c &= \frac{1}{2}(\underline{x}_i + \overline{x}_i) \\ x_i^r &= \frac{1}{2}(\overline{x}_i - \underline{x}_i). \end{aligned} \tag{1}$$

Note that, for sake of simplicity, we refer to the *radius* as range, which is half of the min-max range.

3.1 VARIANCE AND COVARIANCE FOR INTERVAL VALUED DATA

Just like for PCA, in interval data PCA, one wants to pre-process (or transform) the data prior to the analysis, and the data transformation must be consistent with the intrinsic nature of the interval-valued variables at hand. Therefore, a crucial point to extend the PCA applicability to interval-valued variables is the definition of a proper mean and deviation (squared deviation) for interval data and

the definition of distance between intervals. The variance can be defined starting from Hausdorff's distance between two intervals. Given two generic intervals coded in the center and range notation (1), the squared distance between $\{x_i^c, x_i^r\}$ and $\{x_{i'}^c, x_{i'}^r\}$ is defined as (Ferson et al., 2007)

$$d([x]_i, [x]_{i'})^2 = (|x_i^c - x_{i'}^c| + |x_i^r - x_{i'}^r|)^2.$$

Let $[\mathbf{X}]$ be the raw interval data matrix containing *min* and *max* values for p interval-valued variables registered for n statistical units, it has dimensions $(n \times 2p)$. Furthermore, let $\tilde{\mathbf{X}}^c$ and $\tilde{\mathbf{X}}^r$ be the midpoint and range matrices $(n \times p)$; in the following $\mathbf{X}^c$ and $\mathbf{X}^r$ denote the centered versions, that is:

$$\mathbf{X}^c = \left(\tilde{\mathbf{X}}^c - \frac{1}{n}\mathbf{1}\mathbf{1}'\tilde{\mathbf{X}}^c\right), \qquad \mathbf{X}^r = \left(\tilde{\mathbf{X}}^r - \frac{1}{n}\mathbf{1}\mathbf{1}'\tilde{\mathbf{X}}^r\right).$$

A measure of overall variability can be obtained starting from the matrix $\hat{\mathbf{X}}$ obtained by juxtaposing $\mathbf{X}^c$ and $\mathbf{X}^r$:

$$\hat{\mathbf{X}} = \begin{bmatrix} \mathbf{X}^c & \mathbf{X}^r \end{bmatrix}$$

and by considering:

$$\mathbf{V}_{\mathbf{X}} = \text{diag}(\mathbf{\Sigma}_{\hat{\mathbf{X}}}),$$

where $\mathbf{\Sigma}_{\hat{\mathbf{X}}}$ is the variance and covariance matrix for $\hat{\mathbf{X}}$:

$$\hat{\mathbf{X}}'\hat{\mathbf{X}} = \mathbf{\Sigma}_{\hat{\mathbf{X}}} = \frac{1}{2n}\begin{bmatrix} \mathbf{X}'^c\mathbf{X}^c & \mathbf{X}'^c\mathbf{X}^r \\ \mathbf{X}'^r\mathbf{X}^c & \mathbf{X}'^r\mathbf{X}^r \end{bmatrix}.$$

In particular, Palumbo and Lauro (2003) first defined the variability as the mean of the squared Hausdorff distances between each interval in a set of n intervals and the mean interval:

$$\begin{aligned} \sigma^2 &= \frac{1}{N}\sum_{i=1}^{N}(x_i^c + x_i^r)^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(x_i^{c2} + x_i^{r2} + 2|x_i^c||x_i^r|\right), \end{aligned} \tag{2}$$

so that in matrix notation:

$$\mathbf{VAR}_{\mathbf{X}} = (\mathbf{X}'^c\mathbf{X}^c) + (\mathbf{X}'^r\mathbf{X}^r) + (|\mathbf{X}'^c\mathbf{X}^r| + |\mathbf{X}'^r\mathbf{X}^c|).$$

Note that the interval data coding is equivalent to a triangular symmetric fuzzy coding: in fact, Giordani and Kiers (2004) proposed the variance

$$\sigma^2 \quad = \quad = \frac{1}{N}\sum_{i=1}^{N}(x_i^c + \lambda x_i^r)^2,$$

where λ is a weight; under the symmetric triangular fuzzy number condition, λ is a constant equal to .5; in other words, the ranges have a weight equal to .5 with respect to the *min*/*max* coding. Therefore, the interval and symmetric triangular fuzzy coding are equivalent.

More recently, Le-Rademacher and Billard (2012) introduced the definition of symbolic mean and the symbolic covariance matrix $\mathbf{\Sigma}_{symb}$ for interval-valued variables; the latter results from the variability of the interval-valued variables, where the interval values are considered as a whole. Let $[X] = ([X]_1, ..., [X]_p)$ be a p-variate interval-valued random variable and $[\underline{x}_{ij}, \bar{x}_{ij}]$ the interval-valued realization of the j-th variable for the i-th observation $(j = 1, ..., p;\ i = 1, ..., n)$, where $\underline{x}_{ij} \leq \bar{x}_{ij}$. Then the *symbolic* mean $\bar{w}_j$ and the *symbolic* variance $\sigma^2_{j_{symb}}$ of $[x]_j$ are defined according to the following formulae:

$$\bar{w}_j \quad = \quad \frac{1}{2n}\sum_{i=1}^{n}(\underline{x}_{ij} + \bar{x}_{ij}), \tag{3}$$

$$\sigma^2_{j_{symb}} \quad = \quad \frac{1}{3n}\sum_{i=1}^{n}(\underline{x}_{ij}^2 + \underline{x}_{ij}\bar{x}_{ij} + \bar{x}_{ij}^2) - \left(\frac{1}{2n}\sum_{i=1}^{n}(\underline{x}_{ij} + \bar{x}_{ij})\right)^2. \tag{4}$$

It is worth noting that $\bar{w}_j$ corresponds to the mean of the midpoints for the j-th variable. The variance (4) can be extended to the bivariate case for the covariance for j and j':

$$\begin{aligned}\sigma_{jj'_{symb}} = \frac{1}{6n}\sum_{i=1}^{n}[&2(\underline{x}_{ij} - \bar{w}_j)(\underline{x}_{ij'} - \bar{w}_{j'}) + (\underline{x}_{ij} - \bar{w}_j)(\bar{x}_{ij'} - \bar{w}_{j'}) \\ &+ (\bar{x}_{ij} - \bar{w}_j)(\underline{x}_{ij'} - \bar{w}_{j'}) + 2(\bar{x}_{ij} - \bar{w}_j)(\bar{x}_{ij'} - \bar{w}_{j'})]\end{aligned} \tag{5}$$

Then, the *symbolic covariance matrix* $\mathbf{\Sigma}_{symb}$ has diagonal and extra diagonal elements (4) and (5), respectively.

4. PRINCIPAL COMPONENT ANALYSIS FOR INTERVAL-VALUED VARIABLES

This section is devoted to illustrating three of the most widespread principal component analysis (PCA) methods for interval-valued data. Before going into

details, it is worth pinpointing that points in the reduced space cannot consistently represent statistical units described by interval variables. They are represented as segments in $\mathbb{R}$, parallelograms in $\mathbb{R}^2$, parallelepipeds when in $\mathbb{R}^3$, and parallelotopes (generally hyper-rectangles) in $\mathbb{R}^p$ when $p > 3$ (Palumbo and Irpino, 2005). Consequently, interval PCA must account for aspects other than the location.

4.1 CENTERS PRINCIPAL COMPONENT ANALYSIS

Given the previously defined $\mathbf{X}_c$ and $\mathbf{X}_r$, the standardization is given by

$$\mathbf{Z}^c = \mathbf{X}^c \mathbf{V}_c^{-1/2}$$
$$\mathbf{Z}^r = \mathbf{X}^r \mathbf{V}_c^{-1/2}$$

where $\mathbf{V}_c = \frac{1}{n}\text{diag}(\mathbf{X}'^c\mathbf{X}^c)$; therefore, both matrices share the same scaling operator, that is, $\mathbf{V}_c^{-1/2}$.

C-PCA (Cazes et al., 1997; Chouakria et al., 1998) is a PCA of $\mathbf{Z}^c$, with $\mathbf{A}$ being the centers' scores and $\mathbf{B}$ the loadings. The interval bounds are then represented as supplementary points. Therefore, each observation is represented as a d-dimensional hyper-rectangle, obtained by joining the corresponding interval bound projections.

To show how the hyper-rectangles are obtained, consider the component matrix $\mathbf{B}$, and let the positive and negative loadings stored in the non-zero elements of $\mathbf{B}^+$ and $\mathbf{B}^-$, respectively. Formally:

$$b^+_{j\alpha} = \begin{cases} b_{j\alpha} & \text{if } b_{j\alpha} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$b^-_{j\alpha} = \begin{cases} b_{j\alpha} & \text{if } b_{j\alpha} \leq 0 \\ 0 & \text{otherwise} \end{cases}$$

In matrix notation, the bounds of the component scores matrix are given by:

$$\begin{aligned} \underline{\mathbf{A}} &= (\mathbf{Z}^c + \mathbf{Z}^r)\mathbf{B}^- + (\mathbf{Z}^c - \mathbf{Z}^r)\mathbf{B}^+ \\ \overline{\mathbf{A}} &= (\mathbf{Z}^c + \mathbf{Z}^r)\mathbf{B}^+ + (\mathbf{Z}^c - \mathbf{Z}^r)\mathbf{B}^-. \end{aligned}$$

The interpretation of C-PCA is straightforward, as it resembles the single-valued PCA results in terms of plots (observation and loadings) and in terms of quality of the representation (absolute and relative contributions). The drawback of the C-PCA approach is that it does not preserve the nature of the data because it transforms each interval into a single value: its center. While it is possible to represent the hyper-rectangles on the factorial map, the solution is not based on the complete information at hand, as it does not depend on the ranges.

4.2 VERTICES PRINCIPAL COMPONENT ANALYSIS

In order to account for the centers and the ranges simultaneously, V-PCA replaces the interval data matrix that contains the min and max values by transforming the i^{th} row of the matrix $[\mathbf{X}]$ into a *vertices* numeric matrix $\mathbf{Y}_i$, $i = 1, ..., n$. In other words, for p variables, the rows of $\mathbf{Y}_i$ contain one of the 2^p vertices of the i^{th} hyperrectangle:

$$\mathbf{Y}_i = \begin{bmatrix} \underline{x}_{i1} & \underline{x}_{i2} & \cdots & \underline{x}_{ip} \\ \overline{x}_{i1} & \underline{x}_{i2} & \cdots & \underline{x}_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{x}_{i1} & \overline{x}_{i2} & \cdots & \overline{x}_{ip} \end{bmatrix}.$$

Stacking the $\mathbf{Y}_i$'s together, one obtains the $(n2^p \times p)$ vertices data matrix $\mathbf{Y}$ (Cazes et al., 1997), that is

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_n \end{bmatrix}. \tag{6}$$

V-PCA is a PCA on the centered and scaled version of $\mathbf{Y}$.

V-PCA maximizes the explained variability that characterizes vertices; therefore, using the vertices data re-coding, the variables' minima and maxima are considered irrespectively of the observations they are associated with. The analysis maximizes the explained variability among the vertices and not among the units. Note that V-PCA implicitly considers the interval widths by considering all the vertices of each hyperrectangle. For each observation and each component, all the vertices can be derived using (6).

V-PCA contributions are calculated as the squared correlation between the j-th variable and the α-th factor

$$ctr\_abs_{j,\alpha} = \frac{(\lambda^{1/2} u_{j,\alpha})^2}{\lambda_\alpha} = u_{j,\alpha}^2.$$

The vector $\psi_{i,\alpha} = \mathbf{y}_i \mathbf{u}_\alpha$ gives the vertices coordinates of each statistical unit on the principal axes. The representation of the i-th unit on the generic axis α is given by the segment that includes all vertices projections. Adopting the same criterion in a two-dimensional space spanned by the first two principal components, the extreme vertices projections define a rectangle called Maximum Covering Area Rectangle (MCAR). The main problem of this kind of representation is the inevitable MCARs oversize that depends on the leakage of any relationship between

the vertex and its corresponding interval data unit in the analysis (Le-Rademacher and Billard, 2012). The observation unit reconstruction via MCAR is done ex-post, and consequently, the interpretation of the principal components cannot be referred to its main characteristics (Lauro and Palumbo, 2000).
Furthermore, it is worth pinpointing that the number of rows of $\mathbf{Y}$ (6) increases exponentially as the number p of variables increases. For example, suppose we deal with $n = 16$ observation units and $p = 12$ interval-valued variables, then, for this relatively small data set, $\mathbf{Y}$ has dimension 65536×12, which results hard to handle. Actually, thanks to the PCA properties, the number of rows does not represent an insurmountable computational limit. Nevertheless, many variables soon make the PCA of $\mathbf{Y}$ unfruitful because of the huge number of vertex points that refer to every single unit and that inevitably tend to cause over-sized MCARs (Giordani and Kiers, 2006). To (partially) address this issue, getting more consistent MCARs concerning the units' variations, Chouakria et al. (1998) proposed to retain only those vertices having a satisfactory quality of the representation index in the considered q dimensional subspace. Each vertex is a single point, hence its corresponding index is measured in terms of the squared cosines criterion

$$ctr\_rel_{l,\alpha} = \frac{\sum_{j=1}^{p} (z_{l,j} u_{j,\alpha})}{\sum_j z_{l,j}^2}, \tag{7}$$

where l indicates a generic vertex, and $1 \leq l \leq 2^p$. However, the choice of the cut-off level for $ctr\_rel_{l,\alpha}$ remains arbitrary, although it strongly affect the final solution and the consequent results interpretation.

5. MIDPOINTS AND RADII PRINCIPAL COMPONENT ANALYSIS: RECONSIDERED

The *midpoints* and *radii* PCA (MR-PCA, Palumbo and Lauro, 2003) solution takes into account ranges (radii), centers (midpoints) and the inter-connection between centers and ranges. In their paper, the authors refer to the centers and ranges as midpoints and radii, respectively. Without loss of generality, to be consistent with the rest of the article, midpoints and radii are named centers and ranges from now on.
MR-PCA performs two independent analyses of $\mathbf{X}'^{c}\mathbf{X}^{c}$ and $\mathbf{X}'^{r}\mathbf{X}^{r}$, which, however, are not sufficient to cover the whole variability in the interval data matrix. So Palumbo and Lauro's (2003) method provides an additional step to take into account the covariance attributable to the inter-connection between centers and ranges. Towards that end, the MR-PCA exploits a Procrustean rotation (Kiers and

Groenen, 1996) to maximize the congruence coefficient (Tucker, 1951) between the first right eigenvector of the standardized ranges matrix $\mathbf{Z}^r$ and the first right eigenvector of standardized centers matrix $\mathbf{Z}^c$. The method starts with two independent PCAs on $\mathbf{Z}^c$ and $\mathbf{Z}^r$, and upon rotating the ranges, the coordinates of the corresponding range are projected on the centers factorial map. The obtained configuration of points depends, therefore, on both centers *and* ranges.

In Palumbo and Lauro (2003), the MR-PCA pre-processing step, the standardization, is the same as C-PCA. While this kind of standardization has been adopted in the literature (see, e.g. Cazes et al., 1997; Giordani and Kiers, 2004), it does not take into account the variability of the ranges, as the centers' variability standardizes both centers and ranges.

This article proposes a two-fold improvement to the Palumbo and Lauro's MR-PCA: (*i*) standardization of centers and ranges via the SDA-based approach described in subsection 3.1; (*ii*) an enhanced Procrustean rotation. In particular: the symbolic variance as a scaling operator is more appropriate, as outlined by Le-Rademacher and Billard (2012); the updated rotation procedure of the ranges maximizes the average congruence (Tucker, 1951) between all the pairwise considered eigenvectors (not just the first one) of the standardized centers matrix and the corresponding ones of the standardized ranges matrix.

More specifically, the symbolic standardized versions of the centers and ranges are:

$$\begin{aligned}\mathbf{Z}^c &= \mathbf{X}^c\mathbf{S}^{-1} = \left(\tilde{\mathbf{X}}^c - \frac{1}{n}\mathbf{1}\mathbf{1}'\tilde{\mathbf{X}}^c\right)\mathbf{S}^{-1};\\ \mathbf{Z}^r &= \mathbf{X}^r\mathbf{S}^{-1} = \left(\tilde{\mathbf{X}}^r - \frac{1}{n}\mathbf{1}\mathbf{1}'\tilde{\mathbf{X}}^r\right)\mathbf{S}^{-1},\end{aligned}$$

where $\mathbf{S}$ is a diagonal matrix with general term $s_{jj} = \sqrt{\sigma^2_{jj_{symb}}}$, for $j = 1,\ldots,p$. The rotation matrix $\mathbf{T}$ is such that $\mathbf{Z}^c$ is as close as possible to the rotated version of $\mathbf{Z}^r$. Therefore the definition of $\mathbf{T}$ is obtained solving the following constrained optimization problem:

$$\begin{aligned}\min_{\mathbf{T}}: \quad & tr\left(\mathbf{Z}^c - \mathbf{Z}^r\mathbf{T}\right)\left(\mathbf{Z}^c - \mathbf{Z}^r\mathbf{T}\right)' \\ & tr\left(\mathbf{Z}^c\mathbf{Z}^{c\prime}\right) + tr\left(\mathbf{Z}^r\mathbf{T}\mathbf{T}'\mathbf{Z}^{r\prime}\right) - 2tr\left(\mathbf{Z}^{c\prime}\mathbf{Z}^r\mathbf{T}\right) \quad s.t. \quad \mathbf{T}\mathbf{T}' = \mathbf{I}.\end{aligned} \tag{8}$$

It is easy to see that the problem in 8 can be re-written as

$$\max_{\mathbf{T}}: tr\left(\mathbf{Z}^{c\prime}\mathbf{Z}^r\mathbf{T}\right) \qquad s.t. \qquad \mathbf{T}\mathbf{T}' = \mathbf{I},$$

which is equivalent to the maximize the correlation coefficient among the columns of the centers matrix and the corresponding columns of the ranges matrix

$$\sum_{j=1}^{p} \frac{\mathbf{t}_j' \mathbf{Z}^{r\prime} \mathbf{z}_j^c}{(\mathbf{t}_j' \mathbf{Z}^{r\prime} \mathbf{Z}^r \mathbf{t}_j)^{1/2} (\mathbf{z}_j^{c\prime} \mathbf{z}_j^c)^{1/2}},$$

where $\mathbf{t}_j$ and $\mathbf{z}_j^c$ are the j-th column of the $n \times p$ matrices $\mathbf{T}$ and $\mathbf{Z}^c$, respectively. The solution of the optimization problem in Formula 9 is obtained via the iterative procedure proposed by Kiers and Groenen (1996). The convergence is guaranteed as the target function increases at each iteration; local optima can be avoided using different initialization. The procedure is detailed in Algorithm 1.
The coordinates for ranges and centers are

$$\mathbf{\Psi}^c = \mathbf{Z}^c \mathbf{U}_{(d)}^c \quad \text{and} \quad \mathbf{\Psi}^r = \mathbf{Z}^r \mathbf{T} \mathbf{U}_{(d)}^r,$$

where $\mathbf{U}_{(d)}^c$ and $\mathbf{U}_{(d)}^r$ are the first d columns of the eigenvector matrices of $\mathbf{Z}^{r\prime}\mathbf{Z}^r$ and $\mathbf{Z}^{r\prime}\mathbf{Z}^r$, respectively.

Algorithm 1: Monotonically convergent algorithm for orthogonal congruence rotation

1 $\mathbf{W} := \mathbf{Z}^{r\prime}\mathbf{Z}^c (diag(\mathbf{Z}^{c\prime}\mathbf{Z}^c))^{-1/2}$ (with general column $\mathbf{w}_j$).
2 $\mathbf{C} := \mathbf{Z}^{r\prime}\mathbf{Z}^r$.
3 $\rho :=$ largest eigenvalue of $\mathbf{C}$.
4 Choose $\mathbf{T}_c$ (as an orthonormal initialization of $\mathbf{T}$); if $\mathbf{W}'\mathbf{T}_c$ has negative diagonal elements, multiply the corresponding columns of $\mathbf{T}_c$ by -1.
5 $f := tr\mathbf{W}'\mathbf{T}_c (diag(\mathbf{T}_c'\mathbf{C}\mathbf{T}_c))^{-1/2}$.
6 $f^{old} := f$
For $j := 1$ *to* p
(I). $p_j = \mathbf{t}_\mathbf{j}^{c\prime} \mathbf{C} \mathbf{t}_\mathbf{j}^c$.
(II). $q_j = \mathbf{w}_j' \mathbf{t}_j^c$.
(III). If $q_j \neq 0$, $\mathbf{u}_j := {p_j}^{-3/2} q_j (\mathbf{C}\mathbf{t}_j^c - \rho \mathbf{t}_j^c) - 2p_j^{-1/2} q_j^{-1} \mathbf{w}_j' \mathbf{w}_j {\mathbf{t}_j}^c - p_j^{-1/2} \mathbf{w}_j$.
(IV). If $q_j = 0$, $\mathbf{u}_j := \mathbf{0}$.
7 Find $\mathbf{P}$ and $\mathbf{Q}$ by means of SVD: $\mathbf{U} = \mathbf{P}\mathbf{D}\mathbf{Q}'$
8 $\mathbf{T} = -\mathbf{P}\mathbf{Q}'$
9 If $\mathbf{W}'\mathbf{T}$ has negative diagonal elements, multiply the corresponding columns of $\mathbf{T}$ by -1
10 $f := tr\mathbf{W}'\mathbf{T}(diag(\mathbf{T}'\mathbf{C}\mathbf{T}))^{-1/2}$
11 if $f < f^{old} + \varepsilon * |f|$, where ε is a small positive constant, consider the algorithm converged, else $\mathbf{T}_c := \mathbf{T}$ and go to step 6.

The projection of the i^{th} range on the α^{th} component is

$$\psi_{i,\alpha} = [(\psi^c_{i\alpha} - \psi^r_{i\alpha}), (\psi^c_{i\alpha} + \psi^r_{i\alpha})],$$

where $\psi^c_{i,\alpha}$ and $\psi^r_{i,\alpha}$ are the coordinates of the i^{th} center and range on the α^{th} axis. Let the covariance matrix for projected centers and ranges in the d-dimensional subspace be:

$$\mathbf{\Omega} = \mathbf{\Psi}'^c\mathbf{\Psi}^c + \mathbf{\Psi}'^r\mathbf{\Psi}^r + |\mathbf{\Psi}'^c\mathbf{\Psi}^r| + |\mathbf{\Psi}'^r\mathbf{\Psi}^c|,$$

then $tr(\mathbf{\Omega})$ will be used to calculate the percentage of variability explained according to d:

$$I_{(d)} = \frac{tr(\mathbf{\Omega})}{total\ inertia} \times 100, \tag{9}$$

where the total inertia is represented by the trace of the global variance and covariance matrix (3) calculated on $\mathbf{Z}^c$ and $\mathbf{Z}^r$.
For interval-valued data the analogue of the PCA relative contributions is given by

$$ctr\_rel_{i\alpha} = \frac{\sum_\alpha (|\psi^c_{i,\alpha}| + |\psi^r_{i,\alpha}|)^2}{\sum_{j=1}^{p} (|z^c_{i,j}| + |z^r_{i,j}|)^2}.$$

6. RESULTS

In this section a simulation study is carried out to assess the explained variability performance of MR-PCA for different scenarios. Then a real data comparative review of the interval PCA approaches is presented.

6.1 SIMULATION

The simulation set up is derived from the one introduced by Giordani and Kiers (2004) with some differences. The generated data structures refer to 18 observations and four interval-valued variables: $V_1, \cdots, V_4$. In particular, data were generated according to the following simulation scheme:

- centers are generated from a multivariate random Gaussian variable considering three different mean vectors, each mean vector for six units;
- two different correlation structures:
  - S1: positive correlation between V_1 and V_2 and a negative correlation between V_3 and V_4
  - S2: mild positive correlation between V_1 and V_3 and a high negative correlation between V_2 and V_4

- $\alpha = (0.2, 0.5, 0.8)$ noise levels ;
- $\tau = (0.8, 0.6, 0.4)$ proportion of variability due to the centers (e.g. $\tau = 0.8$ means that 80% of total variability is due to centers and 20% to ranges);
- For each combination of α and τ, 100 data sets are generated.

Note that $V1$ and $V2$ are considered signal, whereas $V3$ and $V4$ are considered noise, therefore data structures under the S1 scenario are supposed easier to reconstruct, because of the pairwise correlations characterizing signal only and noise only variables. Data structures in scenario S2 are harder to reconstruct because of the signal/noise correlations. Table 1 summarizes the simulation main results. In particular, the average and standard deviation of $I_{(d)}$ (in full dimensions, see Formula 9), the average and standard deviation of the number of iteration needed for the Procrustean rotation. Also, we report for each combination of scenario, α and τ, the proportion of inertia due to centers and ranges (see the inertia decom-

Tab. 1: Simulation results for S1 (signal/signal and noise/noise correlations) and S2 (signal/noise correlations). Average and sd explained inertia, average and sd of iteration counts; proportion of inertia due to centers and ranges

	τ	α	av. $I_{(d)}$.	sd $I_{(d)}$	av. iter.	sd. iter.	in. cen.	in. ran.
		0.2	99.39	0.895	3.15	1.635	0.89	0.01
	0.8	0.5	98.65	1.604	3.45	2.615	0.82	0.02
		0.8	98.80	1.872	2.92	1.895	0.77	0.03
		0.2	98.70	1.805	3.37	2.135	0.78	0.03
S1	0.6	0.5	99.07	1.932	3.29	2.129	0.75	0.03
		0.8	98.89	2.110	3.24	1.990	0.73	0.04
		0.2	98.35	2.681	3.45	2.105	0.62	0.08
	0.4	0.5	98.92	2.486	3.30	2.028	0.66	0.06
		0.8	98.77	2.296	3.25	1.546	0.68	0.06
		0.2	98.67	0.986	2.74	1.488	0.88	0.01
	0.8	0.5	98.64	1.359	3.02	2.025	0.82	0.02
		0.8	99.01	1.914	3.03	1.795	0.77	0.03
		0.2	96.54	1.789	3.50	1.667	0.75	0.03
S2	0.6	0.5	98.59	2.064	3.29	1.805	0.75	0.03
		0.8	98.94	2.359	2.68	1.384	0.73	0.04
		0.2	95.22	2.564	3.75	2.564	0.60	0.07
	0.4	0.5	98.74	2.408	3.37	1.790	0.65	0.06
		0.8	98.37	2.572	2.84	1.606	0.68	0.06

position in Formula 2).
From the results we see that the average value of $I_{(d)}$ never drops below the 95%, that is, the reconstruction of data structure is always satisfactory. In particular, in scenario S1 neither α nor τ affect the results. In scenario S2, as expected, the method performance slightly decreases with the proportion of variance due to centers (τ).

6.2 APPLICATION: FACE RECOGNITION DATA SET

The face recognition data set refers to 27 man faces images; for each face, a sequence of 1000 images was recorded to calculate the distances between six pairs of informative points; hence altogether, there are 1000×6 values for each face. Then, for each image, the six variables (distances) are summarized into *min-max* interval data. Therefore, the considered data set (Douzal-Chouakria et al., 2011; Le-Rademacher and Billard, 2012) contains six interval-valued variables for 27 units. The interval-data coding allows for considering both the variability among the units (through the centers) and the uncertainty within each unit (through the range). It is worth remarking that any analysis on the whole data set, based on single-valued variables, should consider that the 27000 observations are not independent since, for each considered unit, there are 1000 replications. Figure 1 summarizes the variables by identifying both points of interest and distances, and the variables are referred to.

For the sake of comparing C-PCA and V-PCA with MR-PCA results, the centers and ranges matrices were standardized by the centers' deviations before the analysis for C-PCA and V-PCA. In the MR-PCA, data were standardized according to the symbolic variance (formula 4). As a consequence, the results map for C-PCA

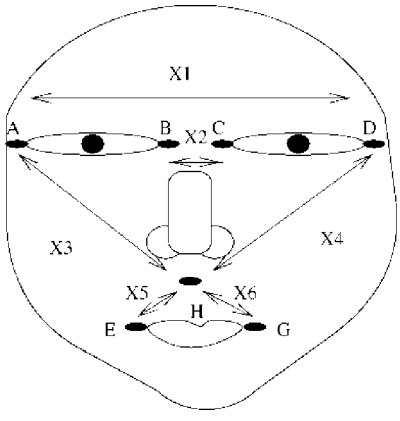

Fig. 1: Variables for face recognition. (Fig. 5 in Douzal-Chouakria et al., 2011)

Tab. 2: Symbolic correlation matrix.

	AD	BC	AH	DH	EH	GH
AD	1.000	0.684	0.387	0.639	-0.003	0.167
BC	0.684	1.000	0.269	0.456	0.242	0.272
AH	0.387	0.269	1.000	0.702	-0.319	-0.644
DH	0.639	0.456	0.702	1.000	-0.449	-0.286
EH	-0.003	0.242	-0.319	-0.449	1.000	0.676
GH	0.167	0.272	-0.644	-0.286	0.676	1.000

and V-PCA differ from the ones in Douzal-Chouakria et al. (2011).

Figures 2 and 3 show the C-PCA and V-PCA observation maps, respectively. Pursuant to the prior equal standardization of the data structures, the two configurations are directly comparable coherently. For the sake of brevity, and considering that the two first dimensions explain most of the variability, comments just refer to the 2-dimensional solutions. The explained inertia is similar considering the first factorial plan: 80.49% for C-PCA, 79.37% for V-PCA. It is worth noting that, in V-PCA, observations are mostly represented by larger rectangles than the

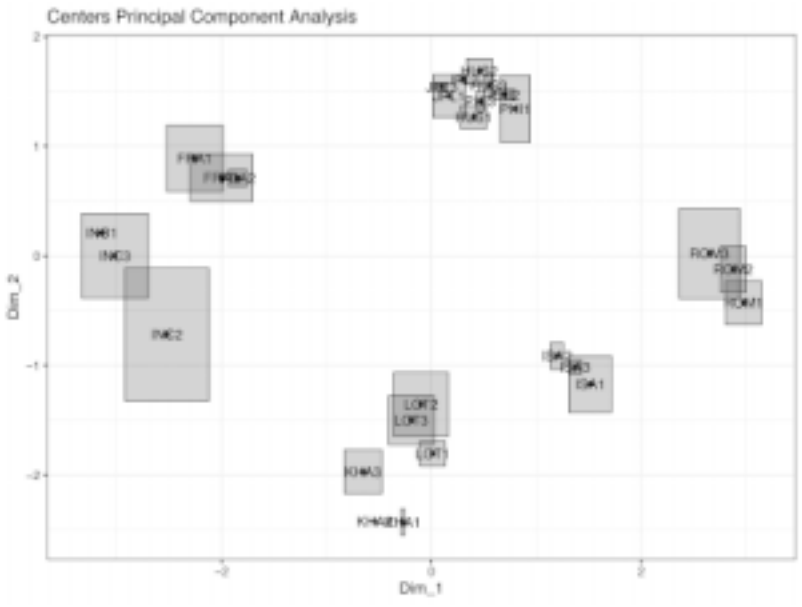

Fig. 2: C-PCA: observations map (80.49% explained inertia). Centers- based configuration; supplementary representation of ranges

C-PCA representation. Lauro and Palumbo (2000) underlined that the active role of minima and maxima in V-PCA accounts for the within-unit variations in the analysis, leading to oversized unit representations. In contrast, in C-PCA, vertices are just projected on the centers-based map as supplementary points: the oversizing effect is limited, but the internal unit variations have no active role in the analysis.

To illustrate the MR-PCA results, it is worth starting from the maps of the loadings in Figure 4. The data structures are standardized according to the square roots of the diagonal values of the symbolic covariance matrix, and Table 2 reports the corresponding correlation matrix. The two sides of the Figure refer to centers and ranges, respectively. The arrows represent the loadings, and the spread of the maps is proportional to the overall variability accountable to centers and ranges, respectively. The more significant proportion of variability comes from the centers. The left-hand side plot in Figure (4) helps to understand the centers' correlation structure and explain the units positioning in Figure 6. In a complementary manner, the right-hand side plot allows for explaining the ranges correlation structure and

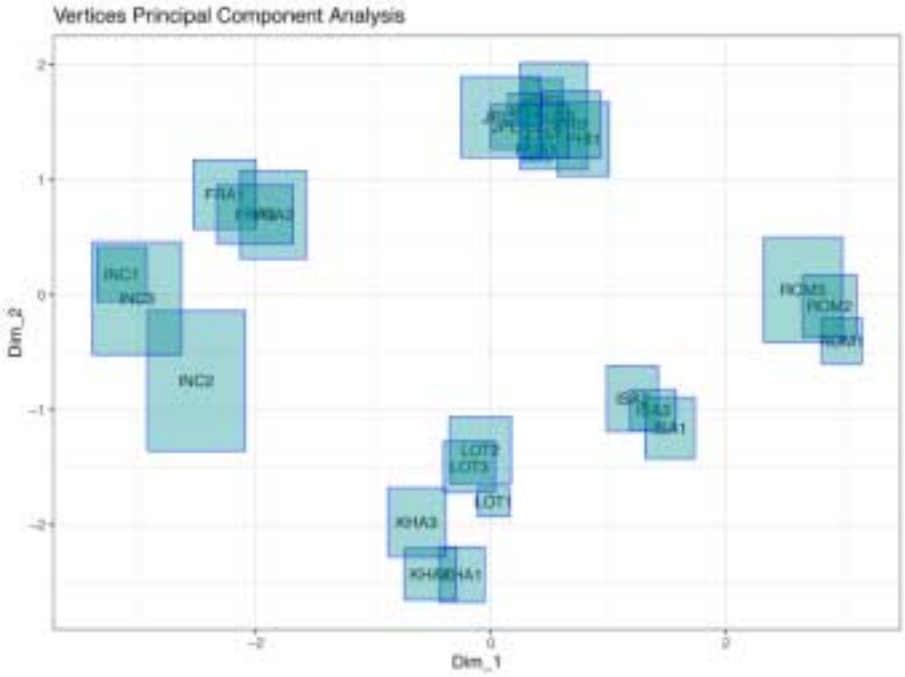

Fig. 3: V-PCA: observations map (79.37% explained inertia). Maximum covering area rectangle representation

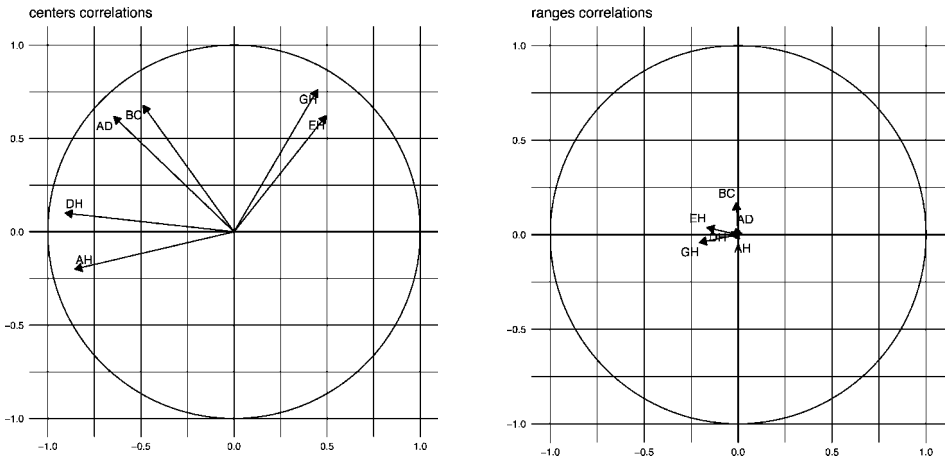

Fig. 4: Symbolic covariance-based standardization: two-dimensional maps of the centers (LHS) and ranges loadings.

their orientations and lengths in the plot.

In Figure 6 rectangles' transparency refers to the quality of the representation. It is measured by the ratio between the length of the projections of the rotated range and the original length of the standardized range. In this example, the centers-ranges correlation structure allows for high values of quality of the representation on the first two dimensions. In particular, it is very high for $HUS1$ and $INC2$, while the worst represented is $KHA3$ that is represented by a segment. Looking at the MR-PPCA map, the ranges orientations on the factorial plan allow tracing back the variations to the source variables: e.g., $INC2$ and $ROM3$ ranges have the same direction meaning that they depend on the same variables; $ROM1$ and $ROM2$ ranges have opposite orientations. Returning on the un-rotated ranges representations and having at hand the ranges correlations map, it is possible to identify the variables that have mainly affected the range length and orientations. Looking again at the $KHA3$ range, it is evident that it has been penalized by the rotation that led to poor representation. This interpretation is allowed only by MR-PCA, representing a significant added value for the method.

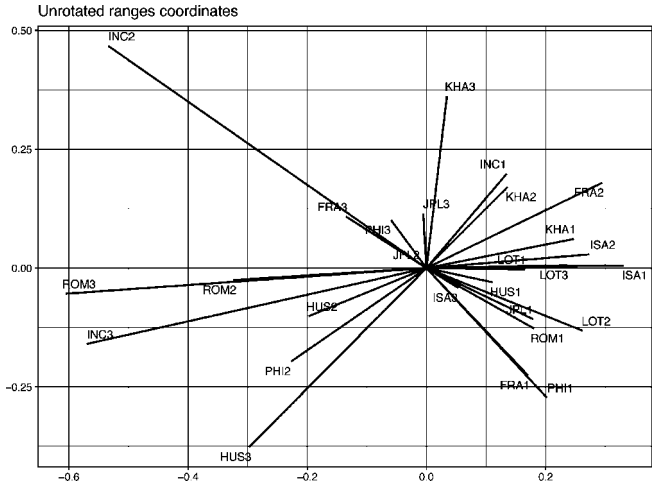

Fig. 5: Un-rotaded ranges representation

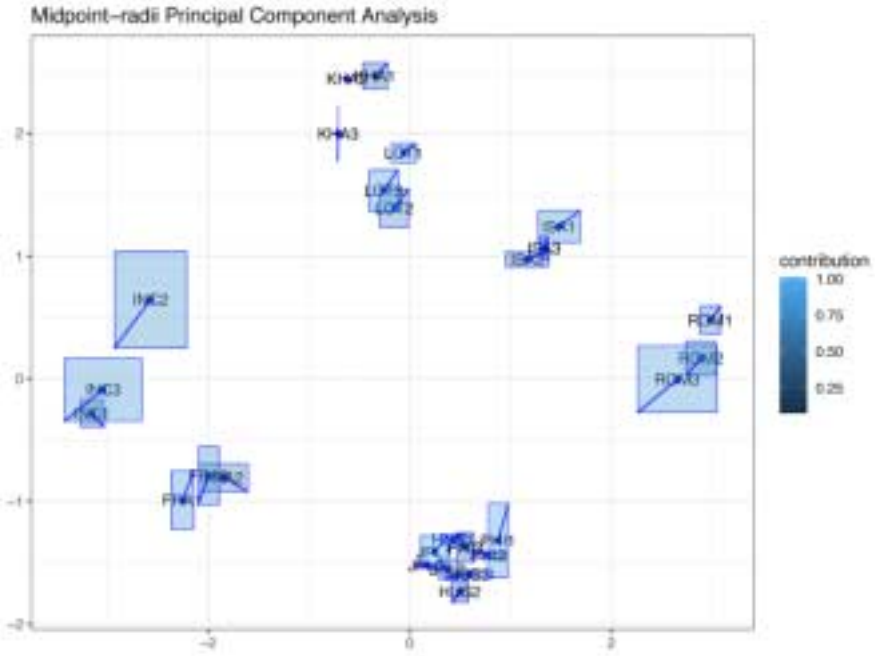

Fig. 6: MR-PCA: observations map (98.23% explained inertia). Pro- crustean rotation of ranges. Darker rectangles have lower relative contri- butions (*ctr_rel*)

7. CONCLUSION

Principal component analysis for complex data structures represents a topic becoming ever more relevant as the total amount of available data increases. Several approaches were proposed in the literature to perform a consistent factorial analysis on interval data when dealing with uncertainty in continuous data. This paper considered three among the most used methods that exploit the PCA on interval-valued data matrices: C-PCA, V-PCA, and MR-PCA. Moreover, starting from Palumbo and Lauro's proposal (2003), it proposes some novelties in the method. In particular, it introduces in the technique the symbolic variance (Le-Rademacher and Billard, 2012) as a measure of variability that allows considering centers and ranges contribution to the total variability. A quite extensive simulation study and an example on a real data set demonstrated the MR-PCA capability in summarizing the information from interval-valued variables.

Data dimensionality and complexity may refer to the number of statistical units and the number of variables. The most recent contributions in analyzing large and massive data sets are often aimed at jointly facing both issues: dimensionality reduction and observations clustering. It is reasonable to assume that integration between dimensionality reduction and clustering could be one of the following challenges in the interval data analysis domain.

Interval data analysis, and PCA in particular, has been applied in an increasing number of fields, such as fault detection (Harkat et al., 2019; Lahdhiri and Taouali, 2021), process monitoring (Ait-Izem et al., 2018), and energy consumption analysis (Gatto and Drago, 2020). Despite the renewed interest, there is lack of specific software for interval data: an R package implementing unsupervised learning and visualization methods for interval-valued data is in the works. Furthermore, supplementary material for reproducible results is available as a Github repository[2].

[2] https://github.com/alfonsoIodiceDE/Interval_data_project

REFERENCES

Ait-Izem, T., Harkat, M.F., Djeghaba, M. and Kratz, F. (2018). On the application of interval pca to process monitoring: A robust strategy for sensor fdi with new efficient control statistics. In *Journal of Process Control*, 63: 29–46.

Billard, L. and Diday, E. (2000). Regression analysis for interval-valued data. In *Data Analysis, Classification, and Related Methods*, 369–374. Springer, Berlin, Heidelberg. 2 https://github.com/alfonsoIodiceDE/Interval_data_project

Billard, L. and Diday, E. (2003). From the statistics of data to the statistics of knowledge: Symbolic data analysis. In *Journal of the American Statistical Association*, 98: 470–487.

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley Series in Computational Statistics. John Wiley & Sons, Chichester.

Billard, L. and Diday, E. (2019). *Clustering methodology for symbolic data*. John Wiley & Sons, Chichester.

Bock,£H.H. and Diday, E. (1999). *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. Springer Science & Business Media.

Bock, H.H. and Diday, E., eds. (2000). *Analysis of Symbolic Data*. Springer Verlag, Hiedelberg.

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

Chouakria, A., Diday, E. and Cazes, P. (1998). Vertices principal components analysis with an improved factorial representation. In *Advances in data sci- ence and classification*, 397–402. Springer, Berlin, Heidelberg.

D'Esposito, M., Palumbo, F. and Ragozini, G. (2012). Interval archetypes: a new tool for interval data analysis. In *Statistical Analysis and Data Mining*, 5 (4): 322–335.

Diday, E. (1988). The symbolic approach in clustering and related methods of data analysis. In *Proceedings of IFCS, Classification and Related Methods of Data Analysis, 1988*, 673–384.

Douzal-Chouakria, A., Billard, L. and Diday, E. (2011). Principal component analysis for interval-valued observations. In *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4 (2): 229–246.

Dubois, D.J. (1980). *Fuzzy sets and systems: theory and applications*, vol. 144. Academic press.

Dubois, D. and Prade, H. (1993). Fuzzy numbers: an overview. In *Readings in Fuzzy Sets for Intelligent Systems*, 112–148.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. In *Psychometrika*, 1 (3): 211–218.

Ferson, S., Ginzburg, L., Kreinovich, V., Longpré, L. and Aviles, M. (2002). Computing variance for interval data is np-hard. In *ACM SIGACT News*, 33 (2): 108–118.

Ferson, S., Kreinovich, V., Hajagos, J., Oberkampf, W., and Ginzburg, L. (2007).

Experimental uncertainty estimation and statistics for data having interval uncertainty. In *Sandia National Laboratories, Report SAND2007-0939*, 162.

Gatto, A. and Drago, C. (2020). Measuring and modeling energy resilience. In *Ecological Economics*, 172: 106527.

Giordani, P. and Kiers, H.A.L. (2004). Principal component analysis of symmetric fuzzy data. In *Computational Statistics and Data Analysis*, 45: 519–548.

Giordani, P. and Kiers, H.A. (2006). A comparison of three methods for principal component analysis of fuzzy interval data. In *Computational Statistics & Data Analysis*, 51 (1): 379–397.

Harkat, M.F., Mansouri, M., Nounou, M. and Nounou, H. (2019). Fault detection of uncertain nonlinear process using interval-valued data-driven approach. In *Chemical Engineering Science*, 205: 36–45.

Jolliffe, I. (2005). Principal component analysis. In *Encyclopedia of statistics in behavioral science*.

Kiers, H.A. and Groenen, P. (1996). A monotonically convergent algorithm for orthogonal congruence rotation. In *Psychometrika*, 61 (2): 375–389.

Lahdhiri, H. and Taouali, O. (2021). Interval valued data driven approach for sensor fault detection of nonlinear uncertain process. In *Measurement*, 171: 108776.

Lauro, C.N. and Palumbo, F. (2000). Principal component analysis of interval data: a symbolic data analysis approach. In *Computational statistics*, 15 (1): 73–87.

Le-Rademacher, J. and Billard, L. (2012). Symbolic covariance principal component analysis and visualization for interval-valued data. In *Journal of Com- putational and Graphical Statistics*, 21 (2): 413–432.

Marino, M. and Palumbo, F. (2002). Interval arithmetic for the evaluation of im- precise data effects in least squares linear regression. In *Statistica Applicata, Italian Journal of Applied Statistics*, 14 (3): 277-291.

Mballo, C. and Diday, E. (2005). Decision trees on interval valued variables. In *The electronic journal of symbolic data analysis*, 3 (1): 8–18.

Nagabhushan, P., Gowda, K.C. and Diday, E. (1995). Dimensionality reduction of symbolic data. In *Pattern recognition letters*, 16 (2): 219–223.

Palumbo, F. and Irpino, A. (2005). Multidimensional interval-data: metrics and factorial analysis. In *Proceedings ASMDA 2005*.

Palumbo, F. and Lauro, C.N. (2003). A PCA for interval valued data based on midpoints and radii. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, and J. Meulman, eds., *New developments in Psychometrics*. Psychometric Soci- ety, Springer-Verlag, Tokyo.

Tucker, L.R. (1951). A method for synthesis of factor analysis studies. *Tech. rep.*, Educational Testing Service Princeton Nj.

Dichiarazione Sostitutiva di Atto Notorio
(ART.47 E ART.19 D.P.R.28.12.2000 N.445)

Il sottoscritto IODICE D'ENZA ALFONSO nato a Napoli il 18/07/1977 , residente a Napoli, Via Pontano, 3, Codice Fiscale: DCDLNS77L18F839F, con riferimento alla domanda di partecipazione per il conseguimento dell'abilitazione scientifica nazionale alle funzioni di professore universitario di Prima Fascia nel settore concorsuale 13/D1- Statistica (Decreto Direttoriale n. 553 del 26 febbraio 2021), consapevole che le dichiarazioni mendaci sono punite ai sensi degli artt. 483, 495, 496, del codice penale e delle leggi speciali in materia,

DICHIARA SOTTO LA PROPRIA RESPONSABILITÀ CIVILE E PENALE DI ESSERE COAUTORE DEL SEGUENTE LAVORO FRUTTO DELL'ATTIVITÀ DI RICERCA CONGIUNTA:

Iodice D'Enza Alfonso, Schisa Vivana, Palumbo Francesco (2021). Principal component analysis for interval data: common approaches and variations. STATISTICA APPLICATA, vol. 33, p. 249-270, ISSN: 2038-5587, doi: 10.26398/IJAS.0033-013

- che la redazione del lavoro è frutto di una stretta e paritetica collaborazione fra gli autori che ne condividono l'impostazione, le ipotesi di ricerca ed i risultati conseguiti;

- ai fini dell'individuazione dell'apporto individuale nei lavori in collaborazione, che l'apporto individuale di Alfonso Iodice D'Enza si è maggiormente esplicitato nelle sezioni 2, 4 e 5, in coerenza con le ricerche e le tematiche affrontate e desumibili dal curriculum e dal percorso scientifico del candidato.

Letto, confermato e sottoscritto.

Napoli, 29 maggio 2023

Alfonso Iodice D'Enza