Computational Statistics (2006) 21:343-363
DOI 10.1007/s00180-006-0267-6

Principal Component Analysis on Interval Data

Federica Gioia and Carlo N. Lauro

Dipartimento di Matematica e Statistica
Università degli Studi di Napoli
"Federico II".

Summary

Real world data analysis is often affected by different types of errors as: measurement errors, computation errors, imprecision related to the method adopted for estimating the data.
The uncertainty in the data, which is strictly connected to the above errors, may be treated by considering, rather than a single value for each data, the interval of values in which it may fall: the interval data. Statistical units described by interval data can be assumed as a special case of Symbolic Object *(SO)*. In Symbolic Data Analysis *(SDA)*, these data are represented as boxes. Accordingly, purpose of the present work is the extension of Principal Component analysis *(PCA)* to obtain a visualisation of such boxes, on a lower dimensional space pointing out of the relationships among the variables, the units, and between both of them. The aim is to use, when possible, the interval algebra instruments to adapt the mathematical models, on the basis of the classical *PCA*, to the case in which an interval data matrix is given. The proposed method has been tested on a real data set and the numerical results, which are in agreement with the theory, are reported.
Keywords: interval-valued variable , interval algebra, *PCA*, visualization.

1 Introduction

The statistical modelling of many problems must account in the majority of cases of "errors" both in the data and in the solution. These errors may be for example: *measurement errors, computation errors, errors due to uncertainty in estimating parameters.* Interval algebra provides a powerful tool for determining the effects of uncertainties or errors and for accounting them in the final solution.

Interval mathematics deals with numbers which are not single values but sets of numbers ranging between a maximum and a minimum value. Those sets of numbers are the sets all possible determinations of the errors. A form of interval algebra appeared for the first time in the literature in (Burkill 1924), (Young 1931); then in (Sunaga 1958). Modern developments of such an algebra were started by R.E. Moore (Moore 1966). Main results may be found in (Alefeld & Herzerberger 1983), (Kearfott & Kreinovich 1996), (Neumaier 1990).

The methods which have been proposed for treating errors in the data, may be also applied to different kind of data that in real life are of interval type. For example:

- Financial data; e. g., (opening value and closing value in a session)
- Customer satisfaction data (expected or perceived characteristic of the quality of a product).
- Tolerance limits in quality control.
- Confidence intervals of estimates from sample surveys.
- Query on a database.

It is known that statistical methods have been primarily developed for single-valued variables. However, in real life there are many situations in which the adoption of single-valued variables cause a loss of information. This have prompted the development of new methodologies of statistical analysis for treating interval-valued variables, that is variables that may assume not just a single value on the unit on which they have been measured, but an interval of values. Statistical indexes for interval-valued variables have been defined in (Canal & Pereira 1998) as scalar statistical summaries. These scalar indexes, may cause loss of information inherent in the interval data For preserving the information contained in the interval data many researchers and in particular Diday and his school of Symbolic Data Analysis (SDA) have developed some methodologies for interval data which provide interval index solutions that sometimes appear oversized as they include unsuitable elements. An approach, which is typical for handling imprecise data, is proposed by (Marino & Palumbo 2003). The centre and the radius of each considered interval and the relations between these two quantities are taken into account. An alternative approach for treating interval-valued variables is proposed in (Gioia & Lauro 2005).The methodology consists in using both the interval algebra and the optimization theory.

Methods for Factorial Analysis and in particular for Principal Component Analysis *(PCA)* on interval data, has been proposed by (Cazes et al. 1997), (Chouakria 1998), (Chouakria et al. 1998), (Gioia F. 2001), (Lauro & Palumbo 2000), (Lauro et al. 2000), (Palumbo & Lauro 2003), (Rodriguez 2000). Statistical units described by interval data can be assumed as a special case of Symbolic Object *(SO)*. In Symbolic Data Analysis *(SDA)*, these data are represented as boxes. The purpose of the present work is the extension of Principal Component analysis *(PCA)* to obtain a visualisation of such boxes, on a lower dimensional space pointing out the relationships among the variables, the units, and between both of them. The approach that we propose, having previously analysed the applicability of the interval algebra tools (Alefeld & Herzberger 1983), (Neumaier 1990), (Kearfott & Kreinovich 1996) is to adapt the mathematical models, on the basis of the classical *PCA*, to the case in which an interval data matrix is given. With difference to other approaches proposed in the literature that work on scalar recoding of the intervals using classical tools of analysis, we make extensively use of the interval algebra tools combined with some optimization techniques. The introduced methodology, named *Interval Principal Component Analysis (IPCA)* will embrace classical *PCA* as special case.

In section 2 of the present work some definitions, notations and main results of the interval algebra are introduced. In section 3 the *IPCA* methodology is presented. In section 4 and section 5 the interpretation of the obtained interval solutions and some numerical results on a real data set are presented.

2 Definitions notations and basic facts

2.1 Interval algebra

An interval $[a,b]$ with $a \leq b$, is defined as the set of real numbers between a and b:

$$[a,b] = \{x / a \leq x \leq b\}$$

Degenerate intervals of the form $[a,a]$, also named *thin* intervals, are equivalent to real numbers.The symbols $\in$, $\subset$, $\cup$, $\cap$, will be used in the common sense of set theory. For example by $[a,b] \subset [c,d]$ we mean that interval $[a,b]$ is included as a *set* in the interval $[c,d]$.Furthermore it is $[a,b]=[c,d] \Leftrightarrow a=c\,,\ b=d$.

Let $\Im$ be the set of intervals. Thus $I \in \Im$ then $I=[a,b]$ for some $a \leq b$. Let us introduce an arithmetic on the elements of $\Im$. The arithmetic will be an extension of real arithmetic. If $\bullet$ is one of the symbols +, -,·, /, we define arithmetic operations on intervals by:

$$[a,b] \bullet [c,d] = \{x \bullet y / \quad a \leq x \leq b, c \leq y \leq d\} \qquad (2.1.1)$$

except that we do not define $[a,b]/[c,d]$ if $0 \in [c,d]$.

The sum, the difference, the product, and the ratio (when defined) between two intervals is the set of the sums, the differences, the products, and the ratios between any two numbers from the first and the second interval respectively.

Let us write an equivalent set of definitions in terms of formulas for the endpoints of the resultant intervals.

Let $[a,b]$, $[c,d]$ be elements of $\Im$, it is:

$$[a,b]+[c,d]=[a+c,\ b+d]$$
$$[a,b]-[c,d]=[a-d,b-c] \qquad (2.1.2)$$
$$[a,b]\times[c,d]=[min(ac,ad,bc,bd),\ max(ac,ad,bc,bd)]$$

if $0\notin[c,d]$, then

$$[a,b]/[c,d]=[a,b]\times[1/d,1/c]$$

It can be easily proved that the addition and the product in *(2.1.2)* are associative and commutative. Real numbers *0* and *1* can be both regarded as units for addition and for product respectively. Other properties may be found in (Moore 1966).

Definition 2.1.1

A rational expression $F(X_1,X_2,...,X_n)$ *in the intervals* X_1, X_2, ..., X_n, *is a finite combination, with the interval arithmetic operations, of* X_1, X_2, ..., X_n *and a finite set of constant intervals.*

Theorem 2.1.1

If $F(X_1,X_2,...,X_n)$ *is a rational expression in the intervals* X_1, X_2, ..., X_n, *then*

$$X_1' \subset X_1,...,\ X_n' \subset X_n \Rightarrow F(X_1',X_2',...,X_n') \subset F(X_1,X_2,...,X_n).$$

for every set of interval numbers X_1, X_2, ..., X_n *for which the interval arithmetic operations in F are defined.*

From Theorem 2.1.1 follows that, computing a finite number of interval arithmetic operations, it is possible to *bound* the range of values of a real rational function over interval of values for each of its arguments.

Proposition 2.1.1

If $F(x_1, \cdots, x_n)$ *is a real rational function in which each variable* x_i *occurs only once and only at the first power, then the corresponding interval expression* $F(X_1,X_2,...,X_n)$ *will compute the actual range of values of F for* x_i *in* X_i *:*

$$F(X_1,X_2,...,X_n)=\{\,y\,/\,y=F(x_1,x_2,...,x_n),\ \ x_i\in X_i\,,i=1,...,n\}.$$

2.2 Interval matrices

Definition 2.2.1

An n×n interval matrix is the following set:

$$X^I=\left[\underline{X},\overline{X}\right]=\left\{X:\underline{X}\le X\le\overline{X}\right\} \qquad (2.2.1)$$

where $\underline{X}$ *e* $\overline{X}$ *are n×n matrix which verify:*

$$\underline{X}\le\overline{X}$$

The inequalities are understood to be component wise.

Introducing the centre matrix and the radius matrix:

$$X_c = \frac{1}{2}\left(\underline{X} + \overline{X}\right) \text{ , } \Delta = \frac{1}{2}\left(\overline{X} - \underline{X}\right)$$

the *(2.2.1)* may be expressed as follow:

$$X^I = \left[X_c - \Delta \text{ , } X_c + \Delta\right].$$

Definition 2.2.2

An $n\times n$ interval matrix X^I is called symmetric if:

$$X^I = X_s^I$$

where:

$$X_s^I = \left[\frac{1}{2}\left(\underline{X} + \underline{X}^T\right), \frac{1}{2}\left(\overline{X} + \overline{X}^T\right)\right]$$

From the definition follows the characterisation:

X^I is symmetric $\Leftrightarrow$ $\underline{X}$ and $\overline{X}$ are symmetric

Hence a symmetric interval matrix may contain *non-symmetric* matrices.

Let us indicate by $M_{np}(R)$ the set of interval matrices of order $n\times p$. An interval matrix $X^I \in M_{np}(R)$ will be represented, in analogy to the case of scalar matrices, by its components as follow: $X^I = (X_{ij})$, where X_{ij} is an interval.

Definition 2.2.3

- *Let $X^I = (X_{ij})$, $Y^I = (Y_{ij}) \in M_{np}(R)$. Then:*

$$X^I \pm Y^I := (X_{ij} \pm Y_{ij})$$

defines the sum interval matrix and the difference interval matrix respectively.

- *Let $X^I = (X_{ij}) \in M_{nr}(R)$ and $Y^I = (Y_{ij}) \in M_{rp}(R)$. Then:*

$$X^I Y^I := \left(\sum_{v=1}^{r} X_{iv} Y_{vj}\right)$$

defines the product interval matrix.

In particular:

let $X^I = (X_{ij}) \in M_{nr}(R)$ and $\boldsymbol{u}^I \in M_{r1}(R)$ (interval vector of r interval components), it is:

$$X^I \boldsymbol{u}^I = \left(\sum_{v=1}^{r} X_{iv} u_v\right)$$

- *Let $X^I \in M_{np}(R)$ and K be an interval. Then:*

$$KX^I = X^I K := (KX_{ij}).$$

2.3 Interval eigenvalues and interval eigenvectors

Given an interval data matrix $X^I \in M_{np}(R)$, a lot of research has been done in characterizing solutions of the following interval eigenvalues problem:

$$X^I u^I = \lambda\, u^I \qquad (2.3.1)$$

which has interesting properties (Deif 1991a), (Rhon 1993), and serves a wide range of applications in physics and engineering.

More in details, the interval eigenvalue problem *(2.3.1)* is solved by determining two sets λ_α^I and u_α^I given by:

$$\lambda_\alpha^I = [\lambda_\alpha(X) : X \in X^I] \quad and \quad u_\alpha^I = [u_\alpha(X) : X \in X^I] \qquad \alpha = 1,\Lambda\,,r$$

where $(\lambda_\alpha(X), u_\alpha(X))$ is an eigenpair of $X \in X^I$. The couple $(\lambda_\alpha^I, u_\alpha^I)$will be the *α-th* eigenpair of X^I and it represents the set of all *α-th* eigenvalues and the set of the corresponding eigenvectors of all matrices belonging to the interval matrix X^I.

Definition 2.3.1

For $x \in R^n$ *the vector* $z = sign\, x$ *may be defined as:*

$$z_i = \begin{cases} 1 & if \quad x_i \geq 0 \\ -1 & if \quad x_i < 0 \end{cases} \qquad i=1,\ldots,n$$

S=diag(sgn x) will indicate the diagonal matrix with *sgn x* on the principal diagonal.

The above definitions are necessary to enunciate the following theorem (Deif 1991a) which gives an important instrument for calculating the eigenvalues of an interval matrix.

Let X^I be an $n{\times}n$ real interval matrix, X_c and ΔX its centre and radius matrix respectively, and let $u_\alpha(X_c)$ $\alpha=1,\ldots,n$, be the eigenvectors of X_c.

Theorem 2.3.1

If X^I *is symmetric and if* $S^\alpha = diag\ (sgn\ u_\alpha(X_c))$, $(\alpha=1,\ldots n)$ *calculated for* X_c *is constant on* X^I, *then the eigenvalue* λ_α *of* X, $X \in X^I$ *ranges over the interval:*

$$\lambda_\alpha^I = [\underline{\lambda}_\alpha(X_c - S^\alpha \Delta X S^\alpha)\ ,\ \overline{\lambda}_\alpha(X_c + S^\alpha \Delta X S^\alpha)],\quad \alpha = 1,\ldots,n \qquad (2.3.2)$$

Theorem 2.3.1 consents to calculate the interval λ_α^I in which the *α-th* eigenvalue of the matrix X, $X \in X^I$ lies. The theorem gives an interesting interpretation of λ_α^I which may be regarded as the α*-th* eigenvalue of the given interval matrix. The further novelty lies in the fact that while previously the problem of the search of the bounds for λ_α^I was limited to their "estimate", now the approach is different inasmuch as it is possible to determine exactly the interval λ_α^I without recurring to any approximation. The interval eigenvectors may be computed by solving a linear programming problem as described in (Seif et al. 1992):

Theorem 2.3.2

A necessary and sufficient condition for $\boldsymbol{u}_\alpha(X)$ to be an eigenvector of X corresponding to $\lambda_\alpha(X)$ is:

$$-\Delta X\,|\boldsymbol{u}_\alpha(X)| \le \left(\lambda_\alpha(X)I - S^\alpha X_c S^\alpha\right)|\boldsymbol{u}_\alpha(X)| \le \Delta X\,|\boldsymbol{u}_\alpha(X)| \qquad (2.3.3)$$

where I is the unitary matrix and $\underline{\lambda}_\alpha(X) \le \lambda_\alpha(X) \le \overline{\lambda_\alpha}(X)$.

To obtain bounds for the components of $\boldsymbol{u}_\alpha(X)$, we write *(2.3.3)* as:

$$\begin{bmatrix} \lambda_\alpha(X)I - S^\alpha X_c S^\alpha - \Delta X \\ S^\alpha X_c S^\alpha - \Delta X - \lambda_\alpha(X)I \end{bmatrix} |\boldsymbol{u}_\alpha(X)| \le 0$$

where $\underline{\lambda}_\alpha(X) \le \lambda_\alpha(X) \le \overline{\lambda_\alpha}(X)$.

To compute lower and upper bounds for $\boldsymbol{u}_\alpha(X)$ we minimize and maximize $|u_{i\alpha}|$ subject to *(2.3.3)* for *α=1,...,n-1*, while keeping $|u_{in}|$ equal to unity. This type of constrained optimization problems is known as Linear Parametric Programming Problems, the solution of which will be obtained via numerical technique. Bounds for $\boldsymbol{u}_\alpha(X)$ are readily obtained by multiplying those for $|\boldsymbol{u}_\alpha(X)|$ by the matrix S^α.

2.4 Interval singular values

The *interval singular values* of an interval matrix X^I can be computed directly from the eigenvalue problem for the matrices $X^t X$, $X \in X^I$ (Deif 1991b).
Thus the problem of computing the interval singular values of X^I becomes the following: given an interval matrix X^I with central matrix $X_c \in \boldsymbol{R}^{n\times p}$, find a description of the set:

$$\Sigma = \left\{\sigma : X^T X\boldsymbol{u} = \sigma^2 \boldsymbol{u}, \quad \boldsymbol{u} \ne 0, \quad X \in X^I\right\}$$

Rather than compute bounds for set Σ, to confine our self to the single interval singular values of X^I:

$$\sigma^I_\alpha,\ \alpha = 1,\ldots,p, \quad \forall X \in X^I$$

the following three assumption must be introduced:

Assumption 1

$sign(\boldsymbol{u}_\alpha(X))$, $\alpha=1,...,p$, is invariant for each $X\in X^I$, and therefore equals $sign(\boldsymbol{u}_\alpha(X_c))$ evaluated at the centre matrix X_c.

Assumption 2

$$\left|\delta X\boldsymbol{u}_\alpha\right| < 2\left|X_c\boldsymbol{u}_\alpha\right|$$

where $\left|\delta X\right| \leq \Delta X$

Assumption 3

$sign(X_c\boldsymbol{u}_\alpha)$, $\alpha=1,...p$, is invariant for each $X\in X^I$ and is therefore equal to $sign(X_c\boldsymbol{u}_\alpha(X_c))$, evaluated at the centre matrix X_c.

Conditions for the validity of Assumptions 1,2,3 may be found in (Deif & Rohn 1994). Indicating by:

$$S_1^\alpha = diag(sign(\boldsymbol{u}_\alpha)) \quad , \quad S_2^\alpha = diag(sign(X_c\boldsymbol{u}_\alpha))$$

it can be proved:

Lemma

Values of δX which extremize the singular value σ_α of the matrix $X_c+\delta X$, $\forall\left|\delta X\right| \leq \Delta X$ are given by:

$$\delta X = \pm S_2^\alpha \Delta X S_1^\alpha .$$

Theorem 2.4.1

Under some Assumptions 1,2,3, the squared singular values σ^2 of $\mathrm{X}_\mathrm{c} + \delta X$, $\forall\left|\delta X\right| \leq \Delta X$, range over the interval:

$$\lambda_\alpha^I = \left[\underline{\lambda}_\alpha , \overline{\lambda}_\alpha\right] \quad \alpha = 1,...,r$$

where:

$$\underline{\lambda}_\alpha = \lambda_\alpha\left(X_c^T X_c - 2\left(S_1^\alpha \Delta X^T S_2^\alpha X_c\right)_s + S_1^\alpha \Delta X^T \Delta X S_1^\alpha \right)$$

$$\overline{\lambda}_\alpha = \lambda_\alpha\left(X_c^T X_c + 2\left(S_1^\alpha \Delta X^T S_2^\alpha X_c\right)_s + S_1^\alpha \Delta X^T \Delta X S_1^\alpha\right)$$

Thus, once the singular values an interval matrix X^I have been computed, a description of the set:

$$\Sigma = \left\{\sigma : X^T X\boldsymbol{u} = \sigma^2\boldsymbol{u}, \quad \boldsymbol{u} \neq 0, \quad X \in X^I\right\}$$

is provided and, in particular, a description of the set of the eigenvalues of any matrix of the form X^TX, when $X\in X^I$, is computed.

3 Principal component analysis on interval data

Let us consider an interval data matrix of n units on which p interval-valued variables $X_1{}^I, X_2{}^I, \dots, X_p{}^I$, *with* $X_j{}^I = \left(X_{ij} = \left[\underline{x}_{ij}, \overline{x}_{ij}\right]\right)_i$, $i = 1, \dots, n$, have been observed:

$$X^I = \begin{bmatrix} \left[\underline{x}_{11}, \overline{x}_{11}\right] & \cdots & \left[\underline{x}_{1j}, \overline{x}_{1j}\right] & \cdots & \left[\underline{x}_{1p}, \overline{x}_{1p}\right] \\ \mathrm{M} & & \mathrm{M} & & \mathrm{M} \\ \left[\underline{x}_{i1}, \overline{x}_{i1}\right] & \cdots & \left[\underline{x}_{ij}, \overline{x}_{ij}\right] & \cdots & \left[\underline{x}_{ip}, \overline{x}_{ip}\right] \\ \mathrm{M} & & \mathrm{M} & & \mathrm{M} \\ \left[\underline{x}_{n1}, \overline{x}_{n1}\right] & \cdots & \left[\underline{x}_{nj}, \overline{x}_{nj}\right] & \cdots & \left[\underline{x}_{np}, \overline{x}_{np}\right] \end{bmatrix} \qquad (3.1)$$

X^I may be visualized as a set of n boxes in a p-dimensional space.

The task is to extend to X^I Principal Component Analysis to obtain a visualisation, on a lower dimensional space, of the relationships among the variables, the units, and between both of them.

The aim is to use, when possible, the interval algebra instruments to adapt the mathematical models, on the basis of the classical PCA, to the case in which an interval data matrix is given. Let us suppose that the interval-valued variables have been previously standardized (see Appendix).

It is known that the classical *PCA* on a real matrix X, in the space spanned by the variables, solves the problem of determining $m \leq p$ axes $\boldsymbol{u}_\alpha$, $\alpha = 1, \dots, m$ such that the sum of the squared projections of the point-units on $\boldsymbol{u}_\alpha$ is maximum:

$$\boldsymbol{u}_\alpha' X'X \boldsymbol{u}_\alpha = Max \qquad 1 \leq \alpha \leq m$$

under the constraints: (3.2)

$$\begin{cases} \boldsymbol{u}_\alpha' \boldsymbol{u}_\beta = 0 & for\ \alpha \neq \beta \\ \boldsymbol{u}_\alpha' \boldsymbol{u}_\beta = 1 & for\ \alpha = \beta \end{cases}$$

The above optimization problem may be reduced to the eigenvalue problem:

$$X'X \boldsymbol{u}_\alpha = \lambda \boldsymbol{u}_\alpha, \qquad 1 \leq \alpha \leq m \qquad (3.3)$$

When the data are of interval type, X^I may be substituted in *(3.3)* and the interval algebra may be used for the products; equation *(3.3)* becomes an *interval eigenvalue problem* of the form:

$$\left(X^I\right)' X^I \boldsymbol{u}_\alpha^I = \lambda^I \boldsymbol{u}_\alpha^I \qquad (3.4)$$

which has the following interval solutions:

$$[\lambda_\alpha(Z) : Z \in (X^I)'X^I] \ , \ [u_\alpha(Z) : Z \in (X^I)'X^I] \qquad \alpha = 1,\Lambda, p \qquad (3.5)$$

i.e., the set of *α-th* eigenvalues of any matrix *Z* contained in the interval product $(X^I)'X^I$, and the set of the corresponding eigenvectors respectively. The intervals in *(3.5)* may be computed by *Theorem 2.3.1*.

Using the interval algebra for solving problem *(3.4)*, the *interval solutions* will be computed but, refer to worse, those intervals are *oversized* with respect to the *intervals of solutions* that we are searching for as it will be discussed below.

For the sake of simplicity, let us consider the case *p=2*, thus two interval-valued variables:

$$X_1{}^I = \left(X_{i1} = \left[\underline{x}_{i1}, \overline{x}_{i1}\right]\right), i = 1,\dots,n, \ X_2{}^I = \left(X_{i2} = \left[\underline{x}_{i2}, \overline{x}_{i2}\right]\right), i = 1,\dots,n$$

have been observed on the *n* considered units.

$X_1{}^I$ and $X_2{}^I$ assume an *interval of values* on each statistical unit: we do not know the exact value of the components x_{i1} or x_{i2} for *i=1,...n*, but only the *range* in which this value falls. In the proposed approach the task is to contemplate *all possible values* of the components x_{i1} , x_{i2} each of which in its own interval of values $X_{i1} = \left[\underline{x}_{i1}, \overline{x}_{i1}\right], X_{i2} = \left[\underline{x}_{i2}, \overline{x}_{i2}\right]$ for *i=1,...n*. Furthermore for each different set of values $x_{11},x_{21},\dots,x_{n1}$ and $x_{12},x_{22},\dots,x_{n2}$, where $x_{ij} \in \left[\underline{x}_{ij}, \overline{x}_{ij}\right] i = 1,\Lambda, n, j = 1,2$, a different cloud of points in the plane is univocally determined and the *PCA* on that set of points must be computed. Thus, with *interval PCA (IPCA)* we mean to determine the *set* of solutions of the classical *PCA* on each set of point-units, set which is univocally determined for any different choice of the point-units each of which in its own rectangle of variation.

Therefore, the *interval of solutions* for which we are looking for are the set of the *α-th axes*, each of which maximize the sum of square projections of a set of points in the plane, and the set of the *variances* of those sets of points respectively.

This is equivalent to solve the optimization problem *(3.3)*, and so the eigenvalue problem *(3.4)* for each matrix $X \in X^I$.

In the light of the above considerations, the background in approaching directly the interval eigenvalue problem *(3.4)*, comes out by observing that the following inclusion holds:

$$(X^I)'X^I = \left\{XY \ / \ X \in (X')^{\mathrm{I}}, Y \in X^I\right\} \supset \left\{X'X \ / \ X \in X^I\right\} \qquad (3.6)$$

this means that in the interval matrix $(X^I)'X^I$ are contained also matrices which *are not* of the form *X'X*. Thus the *interval eigenvalues* and the *interval eigenvectors* of *(3.4)* will be *oversized* and in particular will *include* the set of all eigenvalues and the set of the corresponding eigenvectors of any matrix of the form $X'X$ contained in $\left(X^I\right)'X^I$.

This drawback may be solved by computing an interval eigenvalue problem considering in place of the product:

$$(X')^I X^I = \left\{ XY \quad / \quad X \in (X')^{\mathrm{I}}, Y \in X^I \right\}$$

the following set of matrices:

$$\Theta^I = \left\{ X'X \quad / \quad X \in X^I \right\}$$

i.e., the set of all matrices given by the product of a matrix multiplied by its transpose.

For computing the *α-th* eigenvalue and the corresponding eigenvector of set Θ, that will still be denoted by λ^I_α u^I_α, *Theorem 2.4.1* may be used.

It is important to remark that *Theorem 2.4.1* may be applied under strong hypotheses[1] on the input matrix as described in §2.4. When the above hypotheses are not verified, considering that the variables have been previously standardized, the eigenvalues and eigenvectors of the *correlation interval matrix* may be computed by *Theorem 2.3.1* which is subject to a reduced number of hypotheses than *Theorem 2.4.1.* The *correlation interval matrix* will be indicated by: $\Gamma^I = (corr^I_{ij})$ where $corr^I_{ij}$ is the *interval of correlations* between X_i^I, X_j^I (Gioia & Lauro 2005). Notice that while the *ij-th* component of Γ^I is the *interval of correlations* between X_i^I, X_j^I, the *ij-th* component of $(X')^I X^I$ is an interval which includes that interval of correlations and contains also redundant elements.

It is important to remark that $\Theta^I \subset \Gamma^I$, then the eigenvalues/eigenvectors of Γ^I will be also oversized with respect to those of Θ^I.

The *α-th interval axis* or *interval factor* will be the *α-th* interval eigenvector associated with the *α-th* interval eigenvalue in decreasing order[2].

The *orthonormality* between pairs of interval axes must be interpreted according to:

$$\forall u_\alpha \in u^I_\alpha \text{ such that } u'_\alpha u_\alpha = 1, \quad \exists\, u_\beta \in u^I_\beta \text{ with } \alpha \neq \beta \text{ such that } u'_\beta u_\beta = 1 \quad /$$

$$u'_\alpha u_\beta = 0$$

Thus two interval axes are orthonormal to one another if, taking a unitary vector in the first interval axis there exists a unitary vector in the second one so that their scalar product is zero.

[1] The method works with intervals which are *small* with respect to the *ratio* between the radius and the coordinate of the centre of each interval. Empirically it has been observed that the above ratio must be approximately of 2-3%.

[2] Considering that the *α-th* eigenvalue of Θ is computed by perturbing the *α-th* eigenvalue of $(X^c)'X^c$, the ordering on the interval eigenvalues is given by the natural ordering of the corresponding scalar eigenvalues of $(X^c)'X^c$.

In the classical case the importance explained by the *α-th* factor is computed by: $\lambda_\alpha / \sum_{\beta=1}^{p} \lambda_\beta$. In the interval case the importance of each interval factor is the interval:

$$\left[\frac{\underline{\lambda}_\alpha}{\underline{\lambda}_\alpha + \sum\limits_{\substack{\beta=1 \\ \beta\neq\alpha}}^{p} \overline{\lambda}_\beta} \;,\; \frac{\overline{\lambda}_\alpha}{\overline{\lambda}_\alpha + \sum\limits_{\substack{\beta=1 \\ \beta\neq i}}^{p} \underline{\lambda}_\beta} \right] \qquad (3.7)$$

i.e., the set of all *ratios of variance* explained by each real factor $\boldsymbol{u}_\alpha$ belonging to the interval factor $\boldsymbol{u}_\alpha^I$. The analytical form of the bounds in *(3.7)* has been computed by considering the following chain of equalities:

$$f(\lambda_1, \lambda_2, \Lambda\ , \lambda_p) = \frac{\lambda_\alpha}{\sum\limits_{\beta=1}^{p} \lambda_\beta} = \frac{1}{1 + \frac{1}{\lambda_\alpha} \sum\limits_{\substack{\beta=1 \\ \beta\neq\alpha}}^{p} \lambda_\beta}$$

f has been transformed into a real rational function in which each variables occurs only once and at the first power; therefore according to Proposition 2.1.1, the corresponding interval expression $f\left(\lambda_1^I, \lambda_2^I, \Lambda\ \lambda_p^I\right)$ will compute the actual range of values of *f* for $\lambda_\alpha \in \lambda_\alpha^I \quad \forall\, \alpha = 1, \Lambda\ , p$.

Analogously to what already seen in the space $\boldsymbol{R}^p$, in the space spanned by the units ($\boldsymbol{R}^n$), the eigenvalues and the eigenvectors of the set:

$$(\Theta')^I = \left\{ XX' \quad / \quad X \in X^I \right\}$$

must be computed; the *α-th* interval axis will be the *α-th* interval eigenvector associated with the α-th interval eigenvalue in decreasing order.

Also in this case, *Theorem 2.4.1* on the interval matrix $(X')^I$ may be used if all its hypotheses are satisfied, otherwise the eigenvalues/eigenvectors of the standardized interval matrix $(SS')^I$:

$$(SS')^I = ((ss'_{ij})^I) \quad where \quad (ss'_{ij})^I = [\underline{ss'}_{ij}, \overline{ss'}_{ij}]$$

(see Appendix for details) may be computed.

Considering that $(\Theta')^I \subset (SS')^I$, the eigenvalues/eigenvectors of $(SS')^I$ will be oversized with respect to those of $(\Theta')^I$.

It is known that a real matrix and its transpose have the same eigenvalues and the corresponding eigenvectors connected by a particular relationship. Let us indicate again with $\lambda_1^I , \lambda_2^I ,\Lambda , \lambda_p^I$ the interval eigenvalues of Σ' and with $\boldsymbol{v}_1^I$, $\boldsymbol{v}_2^I,...,\boldsymbol{v}_p^I$ the corresponding eigenvectors, and let us see how the above relationship applies also for the "interval" case. Let us consider for example the *α-th* interval eigenvalue λ_α^I and let $\boldsymbol{u}_\alpha^I$, $\boldsymbol{v}_\alpha^I$, be the corresponding eigenvectors of Θ^I and $(\Theta')^I$ associated with λ_α^I respectively.

Taking an eigenvector of some $X'X \in \Theta^I$: $\boldsymbol{v}_\alpha \in \boldsymbol{v}_\alpha^I$, then:

$$\exists\, \boldsymbol{u} \in \boldsymbol{u}_\alpha^I \quad / \quad \boldsymbol{u}_\alpha = k_\alpha X' \boldsymbol{v}_\alpha \qquad (3.8)$$

where the constant k_α is introduced for the condition of unitary norm of the vector $X'\boldsymbol{v}_\alpha$.

4 Representation and interpretation

4.1 Units

From classical theory, given an $n \times p$ real matrix X we know that the *α-th* principal component $\boldsymbol{c}_\alpha$ is the vector of the coordinates of the n units on the *α-th* axis. Two different approaches may be used to compute $\boldsymbol{c}_\alpha$:

1) $\boldsymbol{c}_\alpha$ may be computed by multiplying the standardized matrix X by the *α-th* computed axis $\boldsymbol{u}_\alpha$: $X\boldsymbol{u}_\alpha$,
2) from the relationship *(3.8)* among the eigenvectors of $X'X$ and XX', $\boldsymbol{c}_\alpha$ may be computed by the product $\sqrt{\lambda_\alpha} \cdot \boldsymbol{v}_\alpha$ of the *α-th* eigenvalue of XX' with the corresponding eigenvector.

When an $n \times p$ interval-valued matrix X^I is given, the *interval coordinate* of the *i-th* interval unit on the *α-th* interval axis, is a representation of an interval which comes out from a linear combination of the original intervals of the *i-th* unit by p interval weights; the weights are the interval components of the α-th interval eigenvector. A box in a bi-dimensional space of representation, is a rectangle having for dimensions the *interval coordinates* of the corresponding unit on the pair of computed interval axis. For computing the *α-th* interval principal component $\boldsymbol{c}_\alpha^I = \left(c_{1\alpha}^I, c_{2\alpha}^I, \Lambda, c_{n\alpha}^I\right)$ two different approaches may be used:

1) compute by the interval row-column product: $\boldsymbol{c}_\alpha^I = X^I \boldsymbol{u}_\alpha^I$,
2) compute the product between a constant interval and an interval vector: $\boldsymbol{c}_\alpha^I = \sqrt{\lambda_\alpha^I} \cdot \boldsymbol{v}_\alpha^I$.

In both cases, the interval algebra product is used thus, the *i-th* component $c_{i\alpha}^{I}$ of c_{α}^{I} will *include* the interval coordinate, as it has been defined above, of the *i-th* interval unit on the *α-th* interval axis.
We refer to the first approach, for computing principal components, when the theorem for solving the eigenvalue problems (for computing v_α^I) cannot be applied if its hypotheses are not verified .Classical *PCA* gives a representation of the results by means of graphs, which permit us to represent the units on projection planes spanned by pairs of factors. The methodology (*IPCA*), that we have introduced, permit us to visualize on planes how the coordinates of the units vary when each component, of the considered interval-valued variable, ranges in its own interval of values, or equivalently when each point-unit describes the boxes to which it belongs.
Indicating with U^I the interval matrix whose *j-th* column is the interval eigenvector $\boldsymbol{u}_\alpha^{\mathrm{I}}$ $(\alpha=1,...p)$, the coordinates of all the interval-units on the computed interval axis are represented by the interval product $X^I U^I$.

4.2 Interval variables

In the classical case, the coordinate of the *i-th* variable on the *α-th* axis is the correlation coefficient between the considered variable and the *α-th* principal component. Thus variables with greater coordinates (in absolute value) are those which best characterize the factor under consideration.
Furthermore, the standardization of each variable makes the variables, represented in the factorial plane, fall inside the correlation circle.
In the interval case the interval coordinate of the *i-th* interval-valued variable on the *α-th* interval axis is the interval correlation coefficient (Gioia & Lauro 2005) between the variable and the *α-th* interval principal component. The interval variables in the factorial plane however, are represented, not in the circle but in the rectangle of correlations. In fact, computing all possible pair of elements, each of which in its own interval correlation, may happens that pairs with the coordinates that are not in relation one another would be also represented; i.e. pairs of elements which are correlations of different *realizations* of the two single-valued variables for which the correlation would be considered.
The interval coordinate of the *i-th* interval-valued variable on the first two interval axes $\boldsymbol{u}_\alpha^I\ \boldsymbol{u}_\beta^I$, namely, the interval correlation between the variable and the first and second interval principal component respectively, will be computed according to the procedure in (Gioia & Lauro 2005) and indicated as follow:

$$corr((X\boldsymbol{u}_\alpha)^I, X_i^I) = \left[\underline{corr}(\boldsymbol{u}_\alpha, i)\ ,\ \overline{corr}(\boldsymbol{u}_\alpha, i)\right]$$

$$corr((X\boldsymbol{u}_\beta)^I, X_i^I) = \left[\underline{corr}(\boldsymbol{u}_\beta, i), \overline{corr}(\boldsymbol{u}_\beta, i)\right]$$

Naturally the rectangle of correlations will be restricted, in the representation plane, to its intersection with the circle with centre in the origin and unitary radius.

4.3 Contributions

In the case of single-valued variables, the weight of the *i-th* unit on the variability of the *α-th* axis, named *absolute contribution*, is given by:

$$\frac{c^2_{i\alpha}}{\sum_{h=1}^{n} c^2_{h\alpha}} \qquad (4.3.1)$$

where c^2_{ij} is the squared coordinate of the *i-th* unit and $\sum_{h=1}^{n} c^2_{hj}$ is the variance of the projected units on the *α-th* axis respectively. In the case of interval-valued variables, *(4.3.1)* must be considered as a function *g* of $c_{i\alpha}$ which may be transformed as follow:

$$g(c_{1\alpha}, c_{2\alpha}, \Lambda, c_{n\alpha}) = \frac{c^2_{i\alpha}}{\sum_{h=1}^{n} c^2_{h\alpha}} = \frac{1}{1 + \frac{1}{c^2_{i\alpha}} \sum_{\substack{h=1 \\ h \neq i}}^{n} c^2_{h\alpha}}$$

Proposition 2.1.1 applies to function *g*, thus the interval:

$$\left[\frac{\underline{c}^2_{i\alpha}}{\underline{c}^2_{i\alpha} + \sum_{\substack{h=1 \\ h \neq i}}^{n} \bar{c}^2_{h\alpha}} , \frac{\bar{c}^2_{i\alpha}}{\bar{c}^2_{ij} + \sum_{\substack{h=1 \\ h \neq i}}^{n} \underline{c}^2_{h\alpha}} \right] \qquad (4.3.2)$$

is the set of all absolute contributions of the *i-th* unit on the *α-th* axis varying the squared projections $c^2_{i\alpha}$ in their interval of values. Interval *(4.3.2)* is the *interval absolute contribution* of the *i-th* interval unit on the *α-th* interval axis.

The contribution of the *j-th* variable on the *α-th* axis may be analogously computed. Interval indexes for the quality of representation on that axis might be calculated substituting the denominator in *(4.3.1)* with the sum of squared coordinates of the units or of the variables. This procedure however would not furnish a good solution for measuring the "quality" of the reconstruction of the original data matrix. To this purpose the introduction of the *singular value decomposition* for interval matrices is necessary.

5 Numerical results

This section shows an example of the proposed methodology on a real data set: the Oil data set (Ichino 1988) (the table below). The data set presents eight different classes of oils described by four quantitative interval-valued variables: "Specific gravity", "Freezing point", "Iodine value" "Saponification".

	Spec. gravity		*Freezing point*		*Iodine value*		*Saponification*	
Linseed	0.93	0.94	-27	-18	170	204	118	196
Perilla	0.93	0.94	-5	-4	192	208	188	197
Cotton	0.92	0.92	-6	-1	99	113	189	198
Sesame	0.92	0.93	-6	-4	104	116	187	193
Camellia	0.92	0.92	-21	-15	80	82	189	193
Olive	0.91	0.92	0	6	79	90	187	196
Beef	0.86	0.87	30	38	40	48	190	199
Hog	0.86	0.86	22	32	53	77	190	202

***Table2**: The interval data set*

The first step of the *IPCA* consists in calculating the following interval correlation matrix:

	Spec.gravity	*Freezing point*	*Iodine value*	*Saponification*
Spec.gravity	[1.00,1.00]			
Freezing point	[-0.97,-0.80]	[1.00,1.00]		
Iodine value	[0.62,0.88]	[-0.77,-0.52]	[1.00,1.00]	
Saponification	[-0.64,-0.16]	[0.30,0.75]	[-0.77,-0.34]	[1.00,1.00]

***Table2**: The interval-correlation matrix*

The interpretation of the interval correlations must take into account both the location and the span of the intervals. Intervals containing the zero are not of interest because they indicate that "everything may happen". An interval with a radius smaller than that of another one is more interpretable. In fact as the radius of the interval correlations decreases, the stability of the correlations improves and a better interpretation of the results is possible. In the considered example, the interval correlations are well interpretable because all intervals do not contain the zero, thus each pair of interval-valued variables are positively correlated or negatively correlated. For example we observe a strong positive correlation between Iodine and Specific gravity and a strong negative correlation between Freezing point and Specific gravity. At equal lower bounds, the interval correlation between Iodine value and Freezing point is more stable than that between Iodine value and Saponification.

Eigenvalues and explained variance:
λ_1=[2.45,3.40], *Explained Variance on the 1st axes:* [61% , 86%]
λ_2=[0.68,1.11], *Explained Variance on the 2nd axes:* [15% , 32%]
λ_3=[0.22,0.33], *Explained Variance on the 1st axes:* [4% , 9%]
λ_4=[0.00,0.08], *Explained Variance on the 1st axes:* [0% , 2%].

The choice of the eigenvalues and so of the interval principal components may be done using the interval eigenvalue-one criterion [1,1]. In the numerical example, only the first principal component is of interest because the lower bound of the corresponding eigenvalue is greater than 1. The second eigenvalue respects the condition of the interval eigenvalue-one partially and, moreover, it is not symmetric with respect to 1. Thus the representation on the second axis is not of great interest even though the two first eigenvalues reconstruct most part of the initial variance. Thus, the second axis is not well interpretable.

Interval variables representation:
The principal components representation is made analysing the correlations among the interval-valued variables and the axes, as illustrated below:

	Spec.gravity	*Freezing point*	*Iodine value*	*Saponification*
Correlations Variables/1st axes	[-0.99 , -0.34]	[0.37 , 0.99]	[-0.99 , -0.20]	[-0.25 , 0.99]
Correlations Variables/2nd axes	[-0.99 , 0.99]	[-0.99 , 0.99]	[-0.99 , 0.99]	[-0.99 , 0.99]

***Table3**: Interval-correlations Variables/Axes*

The first axis is well explained by the contraposition of the variable *Freezing point*, on the positive quadrant, with respect to the variables *Specific gravity* and *Iodine value* on the negative quadrant. The second axis is less interpretable because all the correlations vary from -0.99 and 0.99.[1]
Here below, the graphical results achieved by *IPCA* on the input data table are shown. In Figure 1 the graphical representation of the units is presented; in Figure 2 only the two variables: *Specific gravity* and *Freezing point* are represented:

[1] The absolute contributions on the first axes vary from the interval [0 , 0.91] for *Linseed* and the interval [0,0.16] for *Sesame*, this reflect the "size" of the individuals on the first axes.

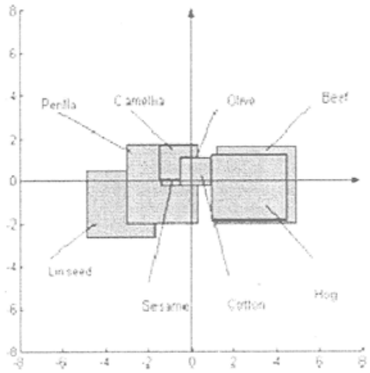

Figure 1: Representation of the units on the 1° factorial plane

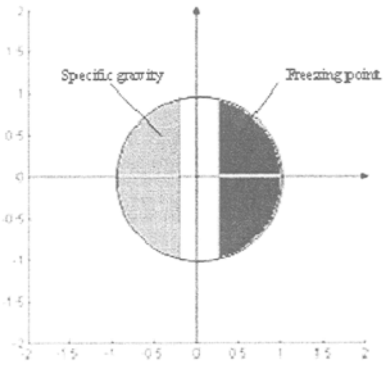

Figure 2: Representation of the variables on the 1° factorial plane

The objects (Fig 1) have a position on the first axis which is strictly connected to the "influence" that the considered variables have on that axis. It can be noticed that *Beef* and *Hog* are strongly influenced by *Saponification* and *Freezing point*; on the contrary *Linseed* and *Perilla* are strongly influenced by *Specific gravity* and *Iodine value*. The other oils Camilla and Olive, are positioned in the central zone so they are not particularly characterized by the interval-valued variables.
It is important to remark that the different oils are characterized not only by the positions of the boxes but also by their *size* and *shape*. A bigger size of a box with respect to the first axis, remarks a greater variability of the characteristics of the oil represented by the first axis. However also the shape and the position of the box can give information on the variability of the characteristics of the oil, with respect to the first and second axis.
Computational cost: the computational cost of each optimization problem reefers to the cost of a *constrained nonlinear optimization* or *nonlinear programming problem*. For computing the correlation matrix, $p \times p$ optimization problems must be solved. The computational cost for computing the *j-th* eigenvector reefers to the cost of a *linear parametric programming problem.*

Appendix

Given two single-valued variables: $X_r = (x_{ir}), X_s = (x_{is})$, $i = 1,\dots,n$, it is known that the correlation between X_r and X_s may be computed as follow:

$$corr(X_r, X_s) = h(x_{1,r},\dots x_{n,r}; x_{1,s},\dots x_{n,s}) = \frac{cov(X_r, X_s)}{\sqrt{var(X_r)}\sqrt{var(X_s)}} \qquad (1)$$

Let us consider now the following interval-valued variables:

$$X_r^I = \left(X_{ir} = \left[\underline{x}_{ir}, \overline{x}_{ir}\right]\right) \quad , \quad X_s^I = \left(X_{is} = \left[\underline{x}_{is}, \overline{x}_{is}\right]\right)_i \qquad i = 1,\dots, n$$

the *interval correlation* is computed as follow (Gioia & Lauro 2005):

$$Corr(X_r^I, X_s^I) = \left[\min_{\substack{x_{ir}\in X_{ir} \\ x_{is}\in X_{is} \\ i=1,\dots,n}} h(x_{1,r},\dots,x_{n,r}; x_{1,s},\dots,x_{n,s}) \ , \ \max_{\substack{x_{ir}\in X_{ir} \\ x_{is}\in X_{is} \\ i=1,\dots,n}} h(x_{1,r},\dots,x_{n,r}; x_{1,s},\dots,x_{n,s}) \right]$$

where $h(x_{1,r},\dots,x_{n,r}; x_{1,s},\dots,x_{n,s})$ is the function in *(1)*.

Analogously , given the single-valued variable X_r, the *standardized* $S_j=(s_{ir})_i$, of X_r is given by:

$$s_{ir} = \frac{x_{ir} - \overline{x}_r}{\sqrt{n}\cdot\sigma_r^2}, \quad i = 1,\dots,n \qquad (2)$$

where $\overline{x_r}$ and σ_r^2 are the mean and the variance of X_r respectively.

When an interval-valued variable X_r^I is given, following the same approach of (Gioia & Lauro 2005), the component s_{ir} in *(2)*, for each $i=1,\dots,n$, transforms into the following function:

$$s_{ir}(x_{ir},\dots x_{nr}) = \frac{x_{ir} - \overline{x}_r}{\sqrt{n}\cdot\sigma_r^2} \qquad (3)$$

as x_{ir} varies in $\left[\underline{x}_{ir}, \overline{x}_{ir}\right]$, $i=1,\dots,n$. The *standardized interval* component s_{ir}^I of X_r^I may be computed by minimizing/maximizing function *(3)*, i.e. calculating the following set:

$$s_{ir}^I = \left[\min_{\substack{x_{ir}\in X_{ir} \\ i=1,\dots,n}} s_{ir}(x_{ir},\dots x_{nr}) \ , \ \max_{\substack{x_{ir}\in X_{ir} \\ i=1,\dots,n}} s_{ir}(x_{ir},\dots x_{nr}) \right] \qquad (4)$$

s_{ir}^I in *(4)* is the interval of the standardized component s_{ir} that may be computed when each component x_{ir} ranges in its interval of values. For computing the interval standardized matrix S^I of an $n\times p$ matrix X^I, interval *(4)* may be computed for each $i=1,\dots,n$ and each $r=1,\dots,p$. Given a real matrix X and indicating by S the standardized of X, it is defined the product matrix: $SS'=(ss'_{ij})$. Given an interval matrix X^I, the product of S^I by its transpose will not be computed by the interval matrix product $(S')^I S^I$ but by minimizing/maximizing each component of SS' when x_{ij} varies in its interval of values. The interval matrix $(SS')^I = ((ss'_{ij})^I)$ is:

$$(ss'_{ij})^I = \left[\min_{\substack{x_{ij} \in X_{ij} \\ i=1,\dots n}} ss'_{ij}(x_{ij},\dots x_{nj}) \ , \ \max_{\substack{x_{ij} \in X_{ij} \\ i=1,\dots n}} ss'_{ij}(x_{ij},\dots x_{nj}) \right]$$

References

Alefeld, G. & Herzerberger, J. (1983), 'Introduction to Interval computation', *Academic Press*, New York.

Billard, L. & Diday, E. (2002), 'Symbolic regression Analysis', *Proceedings IFCS*. In Krzysztof Jajuga et al (EDS.): Data Analysis, Classification and Clustering Methods Heidelberg, Springer-Verlag.

Billard, L. & Diday, E. (2000), 'Regression Analysis for Interval-Valued Data', in: *Data Analysis, Classification and Related Methods* (eds.H.-H. Bock and E. Diday), Springer, 103-124.

Burkill, J. C. (1924), 'Functions of Intervals', *Proceedings of the London Mathematical Society*, **22**, 375-446.

Canal, L. & Pereira, M. (1998), 'Towards statistical indices for numeroid data', in: *Proceedings of the NTTS'98* Seminar, Sorrento Italy.

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

Chouakria, A. (1998), 'Extension des méthodes d'analyse factorielle à des données de type intervalle', Paris IX Dauphine.

Chouakria, A., Diday, E. & Cazes, P. (1998), 'An improved factorial representation of symbolic objects', in: *KESDA'98* April, Luxembourg.

Deif, A.S. (1991a), 'The Interval Eigenvalue Problem', *ZAMM* **71**, 1.61-64, Akademic-Verlag Berlin.

Deif, A.S. (1991b), 'Singular Values of an Interval Matrix', *Linear Algebra and its Applications* **151**, 125-133.

Deif, A.S. & Rohn, J. (1994), 'On the Invariance of the Sign Pattern of Matrix Eigenvectors Under Perturbation', *Linear Algebra and its Applications* **196**, 63-70.

Gioia, F. (2001), 'Statistical Methods for Interval Variables', *Ph.D. thesis*, Dip. di Matematica e Statistica-Università di Napoli "Federico II", in Italian.

Gioia, F. & Lauro, C. (2005), 'Basic Statistical Methods for Interval Data', *Statistica Applicata*, **17** (1). In press.

Kearfott, R. B. & Kreinovich, V. (Eds.) (1996), 'Applications Of Interval Computations', *Kluwer Academic Publishers.*

Lauro, C.N. & Palumbo, F. (2000), 'Principal component analysis of interval data: A symbolic data analysis approach', *Computational Statistics*, **15** (1), 73–87.

Lauro, C.N., Verde, R. & Palumbo, F. (2000), 'Factorial methods with cohesion constraints on symbolic objects', in: *IFCS'00.*

Marino, M. & Palumbo, F. (2003), 'Interval arithmetic for the evaluation of imprecise data effects in least squares linear regression', *Statistica Applicata*, **3**.

Moore, R.E. (1966), 'Interval Analysis', *Prentice Hall*, Englewood Cliffs, NJ.

Neumaier, A. (1990), 'Interval methods for systems of equations', *Cambridge University Press*, Cambridge.

Palumbo, F. & Lauro, C.N. (2003), 'A PCA for interval valued data based on midpoints and radii', in: *New developments in Psychometrics*, Yanai H. et al. eds., Psychometric Society, Springer-Verlag, Tokyo.

Rodriguez, O. (2000), 'Classification et Modeles Lineaires en Analyse des Donnes Symboliques'. *Doctoral Thesis*, Universite de Paris Dauphine IX.

Rhon, J. (1993), 'Interval Matrices : Singularity and real eigenvalues', *SIAM J, Matrix Anal Apply*, **14**, 82-91.

Seif, N.P., Hashem, S. & Deif, A. S. (1992), 'Bounding the Eigenvectors for Symmetric Interval Matrices', *ZAMM* **72**, 233-236.

Sunaga, T. (1958), 'Theory of an Interval Algebra and its Application to Numerical Analysis', *Gaukutsu Bunken Fukeyu-kai*, Tokyo.

Young, R.C. (1931), 'The algebra of many-valued quanties', Math. Ann. **104**, 260-290.