Principal Components Analysis for Trapezoidal Fuzzy Numbers

Alexia Pacheco[1] and Oldemar Rodríguez[2]

[1] Costarican Institute of Electricity, San José, Costa Rica apacheco@ice.go.cr
[2] School of Mathematics, University of Costa Rica, San José, Costa Rica
oldemar.rodriguez@ucr.ac.cr

Summary. Scientists in many disciplines face the problem of interpretation of complex structures such as the symbolic data extracted from databases with a significant amount of records; or containing fuzzy numbers based on expert knowledge or partial knowledge originating from incomplete records. Principal Components Analysis (PCA) is most often used to interpret complex patterns as it allows reducing dimensionality and extracting the main characteristics of the data sample, as well as visualization in a two-dimensional plane and in a correlation circle. There is a need to extend this widely used method to the above mentioned data types.

A new method called `PCA-TF` is proposed that allows pererforming PCA on data sets of trapezoidal (or triangular) fuzzy numbers, that may contain also real numbers and intervals. The approach is an extension to fuzzy numbers of the algorithm by Rodríguez [8]. A group of orthogonal axes is found that permits the projection of the maximum variance of a real numbers' matrix, where each number represents a trapezoidal fuzzy number. The initial matrix of fuzzy numbers is projected to these axes by means of fuzzy numbers arithmetic, which yields Principal Components and they are also fuzzy numbers. Based on these components it is possible to produce graphs of the individuals in two-dimensional plane. It is also possible to evaluate the shape of the ordered pairs of fuzzy numbers and visualize the membership function for each point on the z axis over the two-dimensional xy plane. The application is demonstrated on a data sample of students' grades in [4] and is compared to the results of the NN-PCA used there. Also, an important relation between the arithmetics of the intervals and projection formulas for the interval data type is demonstrated.

Key words: Principal Components Analysis, interval, fuzzy numbers, symbolic data, symbolic data analysis.

1 Introduction

Recent developments in informatics and statistics opened a possibility of assessing data types with more complex structures such as symbolic objects [2] extracted from a huge amount of records or fuzzy numbers [6] generated on the basis of expert knowledge or partial knowledge originating from incomplete records. It is therefore necessary to extend the classical methods of data analyses to these new data types. One of the most frequently used methods is the Principal Component Analysis (PCA) as it allows dimension reduction and visualizing the results in a low-dimensional space and as a correlation circle. In the context of symbolic data analysis and fuzzy data analysis there have been a number of attempts to extend this method to cover the new data types, for example, intervals ([1],[8]) and trapezoidal fuzzy numbers basing on neural networks ([4]).

Herein a new approach, PCA-TF, to perform PCA on a trapezoidal fuzzy number matrix is suggested. It considers the theoretical aspects of several independent topics that all operate on interval variable type such as fuzzy numbers, interval arithmetic and symbolic data. An important relation between the arithmetics of the intervals and projection formulas for the interval data type in the context of symbolic data analysis is demonstrated. Relationship between fuzzy numbers and intervals was already noted by Lodwick [7]. Two computer programs were developed to apply and test the method on arbitrary data sets (Edit-PCA-TF, `PCA-TF`). The application is shown on the example of the data matrix with student grades found in [4].

2 The method : PCA-TF

As in case of usual PCA, the objective is to obtain a low-dimensional representation of the objects/individuals with minimum information loss, which facilitates compression of the initial data and extracting the most relevant characteristics. The new development is an extension of the algorithm by Rodríguez ([8],[1]) to fuzzy numbers. A set of orthogonal axes is found that allows projecting the maximum variance of a real matrix that corresponds to the middle points of the mean intervals representing each trapezoidal fuzzy number in the most natural way, as noted by Dubois [3]. The initial fuzzy numbers matrix is projected on these new axes by means of fuzzy numbers arithmetic, which yields the principal components that are fuzzy numbers as well. Basing on these components it is possible to plot the individuals in the principal plane and also appreciate the shape of the ordered fuzzy number pairs. Besides it is possible to visualize the value of the membership function for each point in a two-dimensional space on the Z-axis.

To extend the algorithm by Rodríguez ([8],[1]), first a relationship amongst the equations for the proyection of an interval type variable and the interval arithmetic was proved, especifically in [8], where in Theorem 4.1.1 it is established that if an interval type variable which geometrically is a hypercube defined by the i-th column of the matrix Z over the j-th principal component (in the direction u_i) is proyected, then the maximum and minimum values are defined by the following equations:

$$\underline{r_{ij}} = \sum_{k=1,u_{kj}>0}^{p} \underline{Z_{ki}} * u_{kj} + \sum_{k=1,u_{kj}<0}^{p} \overline{Z_{ki}} * u_{kj}, \tag{1}$$

$$\overline{r_{ij}} = \sum_{k=1,u_{kj}>0}^{p} \overline{Z_{ki}} * u_{kj} + \sum_{k=1,u_{kj}<0}^{p} \underline{Z_{ki}} * u_{kj}. \tag{2}$$

The equations 1 and 2 correspond the matrix operation $(Z_i)^t * u_j$, where Z_i represents an interval type variable and $Z_{ji} = [\underline{Z}_{ji}\overline{Z}_{ji}]$. Rewriting these formulas for the case of an interval matrix $X = Z^t$ $(X_{ij} = Z_{ji})$ and the real numbers matrix U, $y = X_i * u_j$ is obtained, defined by:

$$\underline{y_{ij}} = \sum_{k=1,u_{kj}>0}^{p} \underline{X_{ik}} * u_{kj} + \sum_{k=1,u_{kj}<0}^{p} \overline{X_{ik}} * u_{kj}, \tag{3}$$

$$\overline{y_{ij}} = \sum_{k=1,u_{kj}>0}^{p} \overline{X_{ik}} * u_{kj} + \sum_{k=1,u_{kj}<0}^{p} \underline{X_{ik}} * u_{kj}. \tag{4}$$

Formally the following theorem is established:

Theorem 1. *Let X be an interval matrix defined by:*

$$X = \begin{pmatrix} [\underline{X_{11}}, \overline{X_{11}}] & \ldots & [\underline{X_{1p}}, \overline{X_{1p}}] \\ \vdots & \ddots & \vdots \\ [\underline{X_{n1}}, \overline{X_{n1}}] & \ldots & [\underline{X_{np}}, \overline{X_{np}}] \end{pmatrix}.$$

Let u_j, the j-th column vector of the matrix $U_{p\times p}$ and X_i the i-th row of the matrix X. The Theorem 4.1.1 [8] and modification of eqs. 5 and 6 to calculate the maximum and minimum projections of the vectors $x_i \in X_i$over u_j can be done by using the interval arithmetic to compute $X_i * u_j$, that is:

$$\underline{y_{ij}} = \sum_{k=1,u_{kj}>0}^{p} \underline{X_{ik}} * u_{kj} + \sum_{k=1,u_{kj}<0}^{p} \overline{X_{ik}} * u_{kj}, \tag{5}$$

$$\overline{y_{ij}} = \sum_{k=1,u_{kj}>0}^{p} \overline{X_{ik}} * u_{kj} + \sum_{k=1,u_{kj}<0}^{p} \underline{X_{ik}} * u_{kj}. \tag{6}$$

Proof. The interval arithmetic establishes that given the a and b intervals then:

$$a + b = [\underline{a} + \underline{b}, \overline{a} + \overline{b}], a - b = [\underline{a} - \overline{b}, \overline{a} - \underline{b}]$$

$$\forall c \in I\!R, ca = \left\{ [c\underline{a}, c\overline{a}] \text{ si } c \geq 0,\ [c\overline{a}, c\underline{a}] \text{ si } c < 0. \right.$$

Then $y_{ij} = X_i * u_j = \sum_{k=1}^{p} X_{ik} * u_{kj} = [\sum_{k=1}^{p} \underline{X_{ik} * u_{kj}}, \sum_{k=1}^{p} \overline{X_{ik} * u_{kj}}]$, where

$$X_{ik} * u_{kj} = \begin{cases} \left[u_{kj} * \underline{X_{ik}}, u_{kj} * \overline{X_{ik}}\right] & \text{si } u_{kj} \geq 0, \\ \left[u_{kj} * \overline{X_{ik}}, u_{kj} * \underline{X_{ik}}\right] & \text{si } u_{kj} < 0, \end{cases}$$

therefore

$$\underline{X_{ik} * u_{kj}} = \begin{cases} u_{kj} * \underline{X_{ik}} \text{ si } u_{kj} \geq 0, \\ u_{kj} * \overline{X_{ik}} \text{ si } u_{kj} < 0, \end{cases}$$

$$\overline{X_{ik} * u_{kj}} = \begin{cases} u_{kj} * \overline{X_{ik}} \text{ si } u_{kj} \geq 0, \\ u_{kj} * \underline{X_{ik}} \text{ si } u_{kj} < 0. \end{cases}$$

Giving the following results:

$$\underline{y_{ij}} = \sum_{k=1, u_{kj}>0}^{p} \underline{X_{ik}} * u_{kj} + \sum_{k=1, u_{kj}<0}^{p} \overline{X_{ik}} * u_{kj},$$

$$\overline{y_{ij}} = \sum_{k=1, u_{kj}>0}^{p} \overline{X_{ik}} * u_{kj} + \sum_{k=1, u_{kj}<0}^{p} \underline{X_{ik}} * u_{kj}.$$

which correspond to the equations given in 5 and 6.

This result offered the basis for formulating a theorem for projection of a fuzzy number variable using fuzzy number arithmetic. Before presenting the theorem the following definitions are introduced:

- A trapezoidal fuzzy number Y, represented by $Y = (Y^{(1)}, Y^{(2)}, Y^{(3)}, Y^{(4)})$ and its membership function is defined by:

$$\mu_Y(x) = \begin{cases} 0 & \text{if } x < Y^{(1)}, \\ \frac{x - Y^{(1)}}{Y^{(2)} - Y^{(1)}} & \text{if } Y^{(1)} \leq x < Y^{(2)}, \\ 1 & \text{if } Y^{(2)} \leq x \leq Y^{(3)}, \\ \frac{Y^{(4)} - x}{Y^{(4)} - Y^{(3)}} & \text{if } Y^{(3)} < x \leq Y^{(4)}, \\ 0 & \text{if } Y^{(4)} < x. \end{cases} \quad (7)$$

- The support and core of the trapezoidal fuzzy number Y are defined as $\text{supp}(Y) = [Y^{(1)}, Y^{(4)}]$ and $\text{core}(Y) = [Y^{(2)}, Y^{(3)}]$.
- The α level for the trapezoidal fuzzy number Y, noted by Y_α corresponds to $Y_\alpha = [Y^{(1)} + \alpha(Y^{(2)} - Y^{(1)}), Y^{(4)} - \alpha(Y^{(4)} - Y^{(3)})]$.

Theorem 2. *Let X a trapezoidal fuzzy number matrix defined by:*

$$X = \begin{pmatrix} (X_{11}^{(1)}, X_{11}^{(2)}, X_{11}^{(3)}, X_{11}^{(4)}) & \ldots & (X_{1p}^{(1)}, X_{1p}^{(2)}, X_{1p}^{(3)}, X_{1p}^{(4)}) \\ \vdots & \ddots & \vdots \\ (X_{n1}^{(1)}, X_{n1}^{(2)}, X_{n1}^{(3)}, X_{n1}^{(4)}) & \ldots & (X_{np}^{(1)}, X_{np}^{(2)}, X_{np}^{(3)}, X_{np}^{(4)}) \end{pmatrix}.$$

Let u_j column vector in $I\!R^p$, X_i the i-th row of matrix X, $y_{i\,supp} = supp(X_i) * u_j$, $y_{i\,core} = core(X_i) * u_j$, $y_{i_\alpha} = (X_i)_\alpha * u_j$ and y_i the trapezoidal fuzzy number defined as $(y_i^{(1)}, y_i^{(2)}, y_i^{(3)}, y_i^{(4)}) = (\underline{y_{i\,supp}}, \underline{y_{i\,core}}, \overline{y_{i\,core}}, \overline{y_{i\,supp}})$. Then

$$y_{i_\alpha} = [y_i^{(1)} + \alpha(y_i^{(2)} - y_i^{(1)}), y_i^{(4)} - \alpha(y_i^{(4)} - y_i^{(3)})] = y_{i_\alpha} = (X_i)_\alpha * u_j.$$

Written in another way:

$$\begin{aligned} y_i^{(1)} + \alpha(y_i^{(2)} - y_i^{(1)}) = & \textstyle\sum_{k=1,u_{kj}>0}^{p}(X_{ik}^{(1)} + \alpha(X_{ik}^{(2)} - X_{ik}^{(1)})) * u_{kj} + \\ & \textstyle\sum_{k=1,u_{kj}<0}^{p}(X_{ik}^{(4)} - \alpha(X_{ik}^{(4)} - X_{ik}^{(3)})) * u_{kj}, \end{aligned} \tag{8}$$

$$\begin{aligned} y_i^{(4)} - \alpha(y_i^{(4)} - y_i^{(3)}) = & \textstyle\sum_{k=1,u_{kj}>0}^{p}(X_{ik}^{(4)} - \alpha(X_{ik}^{(4)} - X_{ik}^{(3)})) * u_{kj} + \\ & \textstyle\sum_{k=1,u_{kj}<0}^{p}(X_{ik}^{(1)} + \alpha(X_{ik}^{(2)} - X_{ik}^{(1)}) * u_{kj}. \end{aligned} \tag{9}$$

The proof has been omitted for briefness.

The membership function for a vector of fuzzy numbers was used to plot the principal plane where the value of this function is given on z axis. For the case of two ordered pairs (two dimensions) and given (x_i, y_i) is an ordered pair of trapezoidal fuzzy numbers, the membership function is defined by $\mu_{(x_i,y_i)} = \min(\mu_{x_i}, \mu_{y_i})$.

It was also demonstrated that the intervals PCA is a particular case of the PCA-TF and thus, the classic PCA as well.

Algorithm - Principal Components Analysis for Trapezoidal Fuzzy Numbers (ACP-FT)

This algorithm is an extension to trapezoidal fuzzy numbers of the algorithm suggested by Rodríguez [8]

Inputs

$n =$ Number of individual (number of rows of the data matrix).

$p =$ Number of variables (columns of the data matrix).

$X =$ Data table (matrix of trapezoidal fuzzy numbers).

Outputs

$C =$ Principal components (trapezoidal fuzzy numbers matrix).

$R =$ Correlation between variables and principal componets (interval matrix).

$CAL =$ Quality of representation of individuals (real number matrix).

$CTR =$ Individual contribution and the components (real number matrix).

$INR =$ Individual contribution to the total inertia (real number matrix).

Steps

1 Defuzzify X. The middle point of the mean interval point is calculated. With $i = 1, \ldots, n$ y $j = 1, \ldots, p$, calculate:

$$X^E = ((X^E{}_{ij})) = \bar{E}(X) = (\bar{E}(X_{ij})) = ((\sum_{l=1}^{4} \frac{X_{ij}{}^{(l)}}{4})).$$

2 Calculate the mean and stanrdard deviation for the columns of the matrix X^E.
With $i = 1, \ldots, n$ and $j = 1, \ldots, p$, calculate:

$$\bar{X}_j^E = \sum_{i=1}^{n} \frac{X^E{}_{ij}}{n}, \sigma_j{}^E = \sum_{i=1}^{n} \frac{(X^E{}_{ij} - \bar{X}_j^E)^2}{n}.$$

3 Calculate the matrix $Z = (z_{ij})$.
With $i = 1, \ldots, n$ and $j = 1, \ldots, p$, calculate: $z_{ij} = \frac{1}{\sqrt{(n)}} \frac{X^E{}_{ij} - \bar{X}_j^E}{\sigma_j^E}$.

4 Calculate the matrix $X^c = (X^c{}_{ij})$.
With $i = 1, \ldots, n$ and $j = 1, \ldots, p$, calculate: $X^c{}_{ij} = \frac{1}{\sqrt{(n)}} \frac{X_{ij} - \bar{X}_j^E}{\sigma_j^E}$.

5 Calculate the matrix $V = Z^t Z$.

6 Calculate the first q eigenvectors $u_1, u_2, \ldots, u_q$ of V and its associated eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q \geq 0$.

7 Calculate the principal components C_k. For $k = 1, \ldots, q$ do: $C_k = (X^c)u_k$, $C^c{}_k = Zu_k$

8 Calculate the eigenvector for the matrix ZZ^t using those of the $Z^t Z$ matrix. For $i = 1, \ldots, n$ y $j = 1, \ldots, q$ calculate: $w_{ij} = \frac{1}{\sqrt{\lambda_j}} (\sum_{k=1}^{n} z_{ik} u_{kj})$.

9 Calculate the correlations amongst variables and the principal components:
For $k = 1, \ldots, q$ do: $PR_k = (X^c)^t w_k$, $R_k = (\text{supp}(PR_{ki} \cap [-1, 1])$,

10 Calculate the interpretation parameters:
Individual representation quality for the i individual in the factorial axis j

$$CAL(i, j) = CAL(X_i, u_j) = \frac{(C^c{}_{ij})^2}{\sum_{i=1}^{p} (z_{ij})^2}.$$

Contribution of the i individual to the factorial axis inertia j

$$CTR(i, j) = CTR(X_i, u_j) = \frac{(C^c{}_{ij})^2}{n * \lambda_j}.$$

Contribution of the i to the total inertia

$$INR(i) = INR(X_i) = \frac{1}{\sqrt{n}} \frac{\sum_{i=1}^{p} (z_{ij})^2}{\sum_{i=1}^{p} \lambda_p}.$$

10 End of the algorithm.

3 Application example

To illustrate the ability of the proposed method to provide a condensed view of multidimensional data, let us consider the hypothetical data set shown in Table 1, taked from [4].

Table 1. Student dataset [4]

	M1	M2	P1	P2
TOM	15	fairly good	unknown	[14,16]
DAVID	9	good	fairly good	10
BOB	6	[10,11]	[13,20]	bien
JANE	very bad	very good	19	[10,12])
JOE	(0,0,2,6)	fairly good	[10,14]	14
JACK	(1,1,1,1)	[4,6]	9	[6,9]

Applying the PCA-TF the principal plane and the correlation circle shown in figs. 1 and 2 respectively, are obtained. The plot is generated by drawing the ordered pairs of fuzzy numbers (C_{ij}, C_{ik}), each of which represent the individual i, where j and k are the selected axes for visualization (in figs. 1 and 2,$j = 1$ and $k = 2$).In this way the support rectangle, the core rectangle and the borders , where the membership function experiences a change, can be examined. For the positioning of the individuals a traditional interpretation can be applied.

From the results it stems that the first principal component is related with the behavior in mathematics (variables M1 and M2), while the second one reflects the behavior in physics (variables P1 y P2). That is why Jane, who got the best grades in mathematics, appears in the upper right corner, while Jack having the worst grades - is at the left fringe. The second component has a high negative correlation with P2, therefore those with good grades on this variable are located in the inferior part of the plot and viceversa. Besides it is noted that Jack is represented by a rectangle shape due to the fact that his data values are hard or "crisp" intervals, while the data for Jane consist of fuzzy numbers for the coordinates $C1$ and $C2$, given that half of her grades are fuzzy numbers. Tom's grade in P1 is unknown and therefore it was represented by an interval for the whole range of grades variation, which yields the biggest support intervals for the coordinates among all the students.

In the correlations circle shown in Fig. 2, the correlation between $M1$ and $M2$ can be verified with $C1$, and that of $P2$ with $C2$. $P1$ reflects the huge variation range in Tom's grade in this variable, as his data were modeled as [0,20] interval due to his unknown grade.

In the method proposed by Denœux & Masson [4] the standard iterative gradient descent is proposed and further on, there is no guarantee of the orthogonality of the axes, which is a clear weakness. Meanwhile PCA guarantees

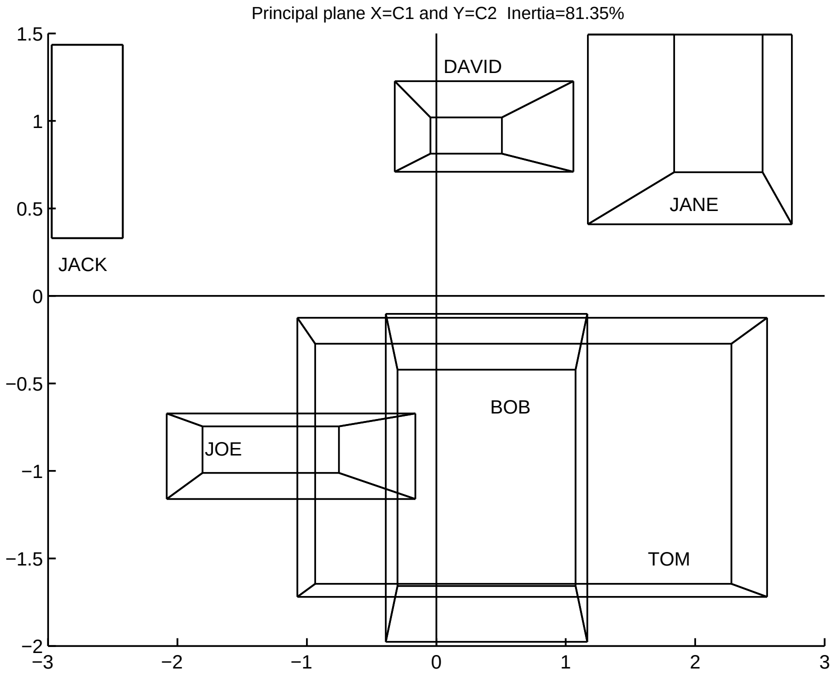

Fig. 1. Principal plane using the first two principal components.

that the principal components are not correlated. Another drawback is that minimizing the function by means of the gradient descent algorithm a minimum local can be found instead of the minimum one, since it's a greedy algorithm. In PCA-TF the axes to which the matrix of fuzzy numbers is projected (which yields principal components) are orthogonal.

A comparison of the results of these two methods applied to the same data showed that they were similar, despite of the fact that in the method by Denœux & Masson the axes were not orthogonal.

PCA-TF was also applyed to fruit juices data set [5] and the results were compared to `NN-PCA` (method proposed by Denœux & Masson) [5]. The components have similar interpretations, but components from PCA-TF are fuzzy numbers and the axes to which the matrix of fuzzy numbers is projected are orthogonal. The details has been omitted for briefness.

Futher work will be oriented to a simulation study to gain knowledge on the performance of the method.

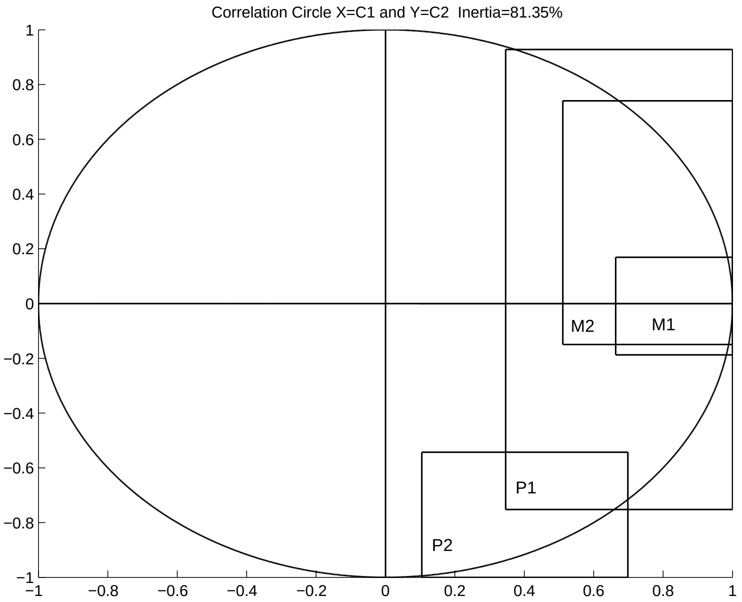

Fig. 2. Correlation circle using the first two principal componets.

4 Conclusions

The study presented herein considered the theoretical aspects of several independent topics that all operate on interval variable type such as fuzzy numbers, interval arithmetic and symbolic data to propose a new method called PCA-TF. An important relationship between the interval arithmetic and the formulas for proyecting interval valued data was also proved (Theorem 1) for that. Relationship between fuzzy numbers and intervals was already noted before by other authors.

The proposed PCA-TF method is an extension to trapezoidal fuzzy numbers of PCA developed in the context of symbolic data analysis for interval data type, applying the fundamental theory of operations on interval data. The PCA-TF method has an advantage of projecting over orthonomal axes and it yields the components which are trapezoidal fuzzy numbers. Both the classical PCA and the interval PCA are particular cases of PCA-TF. Futher work will be oriented to a simulation study to gain knowledge on the performance of the method.

Acknowledgments

The authors dedicate this work to Suzanne Winsberg and thank the Costarican Institute of Electricity for the financial support of the participation in the IFCS 09 held in Dresden, Germany.

References

1. P. Cazes, A. Chouakria, E. Diday, Y. Schektman Extension de l'analyse en composonantes principales à des données de type intervalle. *Rev. Statistique Appliquée Vol. XLV*, 3:5-24, 1997.
2. H-H. Bock, E. Diday, editors. *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data.* Springer, New York, 1999.
3. D. Dubois Possibility theory and statistical reasoning. *Computational Statistics & Data Analysis*, 51:47-69, 2006.
4. T. Denœux, M. Masson Principal Component Analysis of Fuzzy Data using Autossociative Neural Networks. *IEEE Transactions on fuzzy systems*, 12:336-349, 2004.
5. P. Giordani, H. Kiers A Comparison of three methods forPrincipal Component Analysis of Fuzzy Interval data. *Computational Statistics & Data Analysis*,51:379-397, 2006.
6. A. Kaufman, M. Gupta, editors. *Introduction to fuzzy arithmetic: theory and applications.* Van Nostrand Rheinhold, New York, 1991.
7. W. Lodwick, K.D. Jamison Special issue: interfaces between fuzzy set theory and interval analysis. *Fuzzy Sets and Systems*, 135:1-3, 2003.
8. O. Rodríguez, author. *Classification et Modéles Linéaires en Analyse des Données Symboliques.* Université Paris IX-Dauphine, Paris, 2000.