Generalizations of Principal Components Analysis

Oldemar Rodríguez[1], Edwin Diday[2], and Suzanne Winsberg[3]

[1] University Paris IX–Dauphine, CEREMADE. Pl. Du Ml de L. de Tassigny. 75016 Paris, FRANCE.
rodriguez@ceremade.dauphine.fr
[2] University Paris IX–Dauphine, CEREMADE. Pl. Du Ml de L. de Tassigny. 75016 Paris, FRANCE.
diday@ceremade.dauphine.fr
[3] IRCAM, 1 Place Igor Stravinsky, F–75004, Paris, FRANCE.
Suzanne.Winsberg@ircam.fr

Abstract. In [2, Cazes, Chouakria, Diday and Schektman (1997)], they proposed the Centers and the Tops Methods to extend the known principal components analysis method, PCA, to a particular kind of symbolic objects characterized by multi–valued variables of interval–type. Nevertheless they utilize the classical circle of correlation to represent the variables. In this paper we derive duality relations for PCA of interval–type data and we propose a method to compute the symbolic correlation circle using the duality relations.
Also, in this article we propose an algorithm for PCA when the variables are histogram type. This algorithm also works if the data table has variables of interval type and histogram type mixed. If all the variables are interval type it produces the same output as the one produced by the algorithm of the Centers Method.

1 Introduction

Nowadays we often need to perform data analysis (such as principal components, discriminant analysis, discriminant analysis, regression, multidimensional scaling, etc.) on enormous data sets, so large that it makes standard or classical analysis extremely difficult to implement and interpret. To overcome these difficulties it may be necessary and useful to aggregate the data into summary–type classifications or classes, where the number of classes is drastically smaller than the number of single individuals in the original data set.
For example suppose a study involves several cities (or regions, countries, etc.) classified by occupation, age and gender. It may be useful to merge the data for each region, retaining the identifying classifications of "occupation", "age", "income" and "gender". We may wish to describe and analyze underlying concepts such as unemployment and we may also want to query the data set relating to the absence or presence of certain occupations. In these (and related examples) the aggregation process gives rise to symbolic data rather than classical data values on some if not all of the variables describing each symbolic object or observation of the aggregated data set. Most likely, symbolic data methods may have been an integral part of the aggregation procedure.
When we refer to symbolic data, we mean that rather than having a specific Y_j value, an observed value for Y_j can be multi–value, for example $\sum_j = \{2, 5, 7, 9\}$ or $\sum_j = \{$yellow,white,red$\}$, it may be interval–value, for example $\sum_j = [15, 20]$ or it may be modal value, for example $\sum_j = \{1$ with probability $0.1, 0$ with probability $0.9\}$ etc. Moreover, the values of one variable may depend logically or functionally on the values of another variable which are encoded by rules, such as, if the weight of an individual is less than 55 kg the height of the person is less than 180 cm (See [1, Bock and Diday (2000)] for detailed description of symbolic data).
In this paper we present a duality theorem for Centers Principal Components Analysis [2, Cazes, Chouakria, Diday and Schektman (1997)] which is an extension of classical principal component analysis, PCA, to

interval data. Using the duality theorem we present an improved algorithm for Centers PCA. Furthermore we use the duality theorem to develop an efficient algorithm to compute the minimum and the maximum correlation r_{jk} and $\overline{r_{jk}}$ between the $j-$th variable and the $r-$th principal component.
We then extend Centers PCA to the case where X_j, $j = 1, 2, \ldots, n$ are data of histogram type. Our extension of Centers PCA to histogram–type includes the case where the data is of mixed types, histogram, interval, classical (single–value) as well as the cases where the data is of any one or two of these types of data.

2 The duality problem in Centers Method

In the Center Method for interval data the input is m symbolic objects $S_1, S_2, \ldots, S_m$ describe by n interval variables $X^1, X^2, \ldots, X^n$ like we show in the equation (1).

$$\begin{pmatrix} S_1 \\ \vdots \\ S_m \end{pmatrix} = \begin{pmatrix} X_{S_1 1} & \cdots & X_{S_1 n} \\ \vdots & \ddots & \vdots \\ X_{S_m 1} & \cdots & X_{S_m n} \end{pmatrix} = \begin{pmatrix} \left[\underline{x_{11}}, \overline{x_{11}}\right] & \cdots & \left[\underline{x_{1n}}, \overline{x_{1n}}\right] \\ \vdots & \ddots & \vdots \\ \left[\underline{x_{m1}}, \overline{x_{m1}}\right] & \cdots & \left[\underline{x_{mn}}, \overline{x_{mn}}\right] \end{pmatrix}. \tag{1}$$

The idea of the centers method is to transform the matrix presented into (1) in the following matrix (2):

$$X^c = \begin{pmatrix} x_{11}^c & x_{12}^c & \cdots & x_{1n}^c \\ x_{21}^c & x_{22}^c & \cdots & x_{2n}^c \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1}^c & x_{m2}^c & \cdots & x_{mn}^c \end{pmatrix} = \begin{pmatrix} \frac{\underline{x_{11}} + \overline{x_{11}}}{2} & \frac{\underline{x_{12}} + \overline{x_{12}}}{2} & \cdots & \frac{\underline{x_{1n}} + \overline{x_{1n}}}{2} \\ \frac{\underline{x_{21}} + \overline{x_{21}}}{2} & \frac{\underline{x_{22}} + \overline{x_{22}}}{2} & \cdots & \frac{\underline{x_{2n}} + \overline{x_{2n}}}{2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\underline{x_{m1}} + \overline{x_{m1}}}{2} & \frac{\underline{x_{m2}} + \overline{x_{m2}}}{2} & \cdots & \frac{\underline{x_{mn}} + \overline{x_{mn}}}{2} \end{pmatrix}. \tag{2}$$

Then in the Centers Method we conduct a standard PCA to the data matrix defined in (2). To apply this standard PCA [3, Chouakria (1998)] uses the matrix of variance–covariance $V^c = (X^c)^t X^c$ and then to compute the interval principal component $[\underline{y_{ik}}, \overline{y_{ik}}]$ [2, Cazes, Chouakria, Diday and Schektman (1997)] propose equations (3) and (4).

$$\underline{y_{ik}} = \sum_{j, u_{jk} < 0} \left(\overline{x_{ij}} - \overline{X_j^c}\right) u_{jk} + \sum_{j, u_{jk} > 0} \left(\underline{x_{ij}} - \overline{X_j^c}\right) u_{jk}, \tag{3}$$

$$\overline{y_{ik}} = \sum_{j, u_{jk} < 0} \left(\underline{x_{ij}} - \overline{X_j^c}\right) u_{jk} + \sum_{j, u_{jk} > 0} \left(\overline{x_{ij}} - \overline{X_j^c}\right) u_{jk}. \tag{4}$$

where $\overline{X_j^c}$ is the mean of the column j–th of the matrix X^c, and $u = (u_{1k}, u_{2k}, \ldots, u_{nk})$ is the $k-$th eigenvector of V^c.
However [2, Cazes, Chouakria, Diday and Schektman (1997)] utilize the classical circle of correlation to represent the variables. The correlation between the variables and the principal components are not symbolic, because they compute the standard correlations between the centers of gravity of variables and the principals components. It is well known that in standard principal component analysis method we can compute the correlation between the variables and the principal components using the duality relations starting from the coordinates of the individuals in the principal plane. Also we can compute the coordinates of the individuals in the principal plane using duality relations starting from the correlation between the variables and the principal components.
We shall to center and reduce the matrix X^c in order to work with correlations as we show in (5) where $\overline{X_j^c}$ and σ_j^c are the mean and the variance of the j–th column of the matrix X^c respectively:

$$z_{ij} = \frac{1}{\sqrt{m}} \frac{x_{ij}^c - \overline{X_j^c}}{\sigma_j^c}. \tag{5}$$

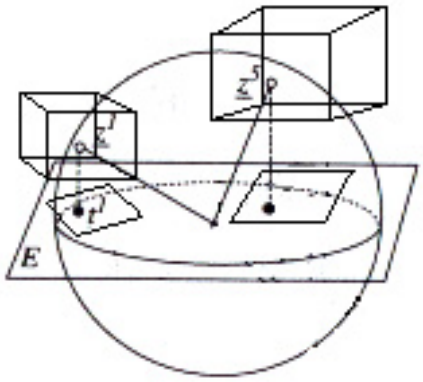

Fig. 1. Projection of the hypercube variables.

Then we will work with the matrix $Z = (z_{ij})_{\substack{i=1,2,\ldots,m \\ j=1,2,\ldots,n}}$. If we denote z^j the column j–th of the matrix Z, so we have $(z^j)^t \cdot z^i = R(z^j, z^i) \leq 1$ then the center of the hypercube variable is always inside of the radius one circle. We illustrate that in Figure 1. Let $\overline{z}_{ij}^c = \frac{1}{\sqrt{m}} \frac{\overline{x_{ij}} - \overline{X_j^c}}{\sigma_j^c}$ and $\underline{z}_{ij}^c = \frac{1}{\sqrt{m}} \frac{\underline{x_{ij}} - \overline{X_j^c}}{\sigma_j^c}$.
The inertia matrix ZZ^t is symmetrical, its eigenvectors are orthonormal and its eigenvalues are all positive. We denote by $v_1, v_2, \ldots, v_q$ the q eigenvectors of ZZ^t associated with the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q > 0$. We also denote by $V = [v_1|v_2|\cdots|v_q]$ the matrix of size $m \times q$ that has as columns the eigenvectors of ZZ^t. It is well known that we can compute the coordinates of the variables in the circle of correlation by Z^tV, then we can compute the coordinate of the i–th column of X^c (point centre–variable) on j–th principal component (in the direction of v_j) by the equation (6):

$$r_{ij} = \sum_{k=1}^{m} z_{ki} v_{kj}. \tag{6}$$

Like Z is the matrix X centered and reduced. The number r_{ij} also represents the correlation between the center of gravity of the interval–variable X^i and the j–th principal component.
If we project the hypercube variable defined by the i–th column of Z on the j–th principal component (on the direction of v_i), then we have that the minimum and the maximum value are given by the equation (7) and (8) respectively:

$$\underline{r_{ij}} = \sum_{k=1, v_{kj}<0}^{m} \overline{z}_{ki}^c v_{kj} + \sum_{k=1, v_{kj}>0}^{m} \underline{z}_{ki}^c v_{kj}, \tag{7}$$

$$\overline{r_{ij}} = \sum_{k=1, v_{kj}<0}^{m} \underline{z}_{ki}^c v_{kj} + \sum_{k=1, v_{kj}>0}^{m} \overline{z}_{ki}^c v_{kj}. \tag{8}$$

To prove that, let be $\widehat{z}_j = (\widehat{z}_{1j}, \widehat{z}_{2j}, \ldots, \widehat{z}_{mj}) \in Z_H^j$ (the hypercube defined by the j-th column of Z) then $\widehat{z}_{ij} \in [\underline{z}_{ij}^c, \overline{z}_{ij}^c]$ for all $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, q$. We denote by $p\widehat{z}_{ij}$ the projection of $\widehat{z}_j$ on the axis factorial with direction v_i.
Since $\widehat{z}_{ij} \in [\underline{z}_{ij}^c, \overline{z}_{ij}^c]$ we have (9) and (10):

$$\underline{z}^c_{ki} v_{kj} \leq \widehat{z}_{ki} v_{kj} \leq \overline{z}^c_{ki} v_{kj} \text{ if } v_{kj} \geq 0, \tag{9}$$

$$\underline{z}^c_{ki} v_{kj} \geq \widehat{z}_{ki} v_{kj} \geq \overline{z}^c_{ki} v_{kj} \text{ if } v_{kj} \leq 0. \tag{10}$$

By definition $p\widehat{z}_{ij} = \sum_{k=1}^{m} \widehat{z}_{ki} v_{kj}$ then:

$$p\widehat{z}_{ij} = \sum_{k=1}^{m} \widehat{z}_{ki} v_{kj} = \sum_{k=1, v_{kj}>0}^{m} \widehat{z}_{ki} v_{kj} + \sum_{k=1, v_{kj}<0}^{m} \widehat{z}_{ki} v_{kj}.$$

So, using (9) and (10) we get:

$$p\widehat{z}_{ij} \leq \sum_{k=1, v_{kj}<0}^{m} \underline{z}^c_{ki} v_{kj} + \sum_{k=1, v_{kj}>0}^{m} \overline{z}^c_{ki} v_{kj} = \overline{r_{ij}},$$

and analogously we have:

$$p\widehat{z}_{ij} \geq \sum_{k=1, v_{kj}<0}^{m} \overline{z}^c_{ki} v_{kj} + \sum_{k=1, v_{kj}>0}^{m} \underline{z}^c_{ki} v_{kj} = \underline{r_{ij}}.$$

Hence, we have proved that $p\widehat{z}_{ij} \in [\underline{r_{ij}}, \overline{r_{ij}}]$ and we also have that r_{ij}, $\overline{r_{ij}}$ are the projection of some vertex of the hypercube. Then we have proved that the value of $\underline{r_{ij}}$ and $\overline{r_{ij}}$ are given by the equation (7) and (8) respectively.
But like we show in Figure 1, there are some points of the projection that are out of circle of radius one, so if we want to give to $\underline{r_{ij}}$ and $\overline{r_{ij}}$ the meaning of minimum and maximum correlation respectively we have to compute them as in equations (11) and (12):

$$\underline{r_{ij}} = \max \left[\sum_{k=1, v_{kj}<0}^{m} \overline{z}^c_{ki} v_{kj} + \sum_{k=1, v_{kj}>0}^{m} \underline{z}^c_{ki} v_{kj}, -1 \right]. \tag{11}$$

$$\overline{r_{ij}} = \min \left[\sum_{k=1, v_{kj}<0}^{m} \underline{z}^c_{ki} v_{kj} + \sum_{k=1, v_{kj}>0}^{m} \overline{z}^c_{ki} v_{kj}, 1 \right]. \tag{12}$$

There are some very well known relations of duality between the eigenvectors of ZZ^t and Z^tZ, it is known that both matrices have the same q strictly positive eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_q$ and if we denote by $u_1, u_2, \ldots, u_q$ the first q eigenvectors of Z^tZ, then the relations between the eigenvectors of ZZ^t and Z^tZ are shows in the equations (13) and (14):

$$u_\ell = \frac{Z^t v_\ell}{\sqrt{\lambda_\ell}} \text{ for } \ell = 1, 2, \ldots, q. \tag{13}$$

$$v_\ell = \frac{Z u_\ell}{\sqrt{\lambda_\ell}} \text{ for } \ell = 1, 2, \ldots, q. \tag{14}$$

With these ideas we propose two algorithms to apply a PCA that extend the one propose in [2, Cazes, Chouakria, Diday and Schektman (1997)] in order to produce a symbolic circle of correlation. We also propose an third algorithm to improve the time of the execution by considering which matrix is smaller ZZ^t or Z^tZ.

ALGORITHM 1: PRINCIPAL COMPONENT ANALYSIS WITH ZZ^t

Input :

- m =number of symbolic objects.
- n =number of symbolic variables.
- The symbolic data table $X = \begin{pmatrix} \left[\underline{x_{11}}, \overline{x_{11}}\right] & \left[\underline{x_{12}}, \overline{x_{12}}\right] & \cdots & \left[\underline{x_{1n}}, \overline{x_{1n}}\right] \\ \left[\underline{x_{21}}, \overline{x_{21}}\right] & \left[\underline{x_{22}}, \overline{x_{22}}\right] & \cdots & \left[\underline{x_{2n}}, \overline{x_{2n}}\right] \\ \vdots & \vdots & \ddots & \vdots \\ \left[\underline{x_{m1}}, \overline{x_{m1}}\right] & \left[\underline{x_{m2}}, \overline{x_{m2}}\right] & \cdots & \left[\underline{x_{mn}}, \overline{x_{mn}}\right] \end{pmatrix}$.

Output :

- The symbolic correlation between the variables and the principal components in the following matrix:

$$R = \begin{pmatrix} \left[\underline{R}(X^1, Y^1), \overline{R}(X^1, Y^1)\right] & \cdots & \left[\underline{R}(X^1, Y^n), \overline{R}(X^1, Y^n)\right] \\ \vdots & \ddots & \vdots \\ \left[\underline{R}(X^n, Y^1), \overline{R}(X^n, Y^1)\right] & \cdots & \left[\underline{R}(X^n, Y^n), \overline{R}(X^n, Y^n)\right] \end{pmatrix}.$$

- The symbolic matrix with the first q principal components:

$$Y = \begin{pmatrix} \left[\underline{y_{11}}, \overline{y_{11}}\right] & \left[\underline{y_{12}}, \overline{y_{12}}\right] & \cdots & \left[\underline{y_{1q}}, \overline{y_{1q}}\right] \\ \left[\underline{y_{21}}, \overline{y_{21}}\right] & \left[\underline{y_{22}}, \overline{y_{22}}\right] & \cdots & \left[\underline{y_{2q}}, \overline{y_{2q}}\right] \\ \vdots & \vdots & \ddots & \vdots \\ \left[\underline{y_{m1}}, \overline{y_{m1}}\right] & \left[\underline{y_{m2}}, \overline{y_{m2}}\right] & \cdots & \left[\underline{y_{mq}}, \overline{y_{mq}}\right] \end{pmatrix}.$$

Step 1: Compute the matrix $X^c = (x_{ij}^c)_{\substack{i=1,2,\ldots,m \\ j=1,2,\ldots,n}}$ by:

$$x_{ij}^c = \frac{\underline{x_{ij}} + \overline{x_{ij}}}{2}.$$

Step 2: Compute the matrix $Z = (z_{ij})_{\substack{i=1,2,\ldots,m \\ j=1,2,\ldots,n}}$ by:

$$z_{ij} = \frac{1}{\sqrt{m}} \frac{x_{ij}^c - \overline{X_j^c}}{\sigma_j^c}.$$

Step 3: Compute the matrix $\underline{Z} = (\underline{z}_{ij})_{\substack{i=1,2,\ldots,m \\ j=1,2,\ldots,n}}$ and $\overline{Z} = (\overline{z}_{ij})_{\substack{i=1,2,\ldots,m \\ j=1,2,\ldots,n}}$ by:

$$\underline{z}_{ij} = \frac{1}{\sqrt{m}} \frac{\underline{x_{ij}} - \overline{X_j^c}}{\sigma_j^c},$$

$$\overline{z}_{ij} = \frac{1}{\sqrt{m}} \frac{\overline{x_{ij}} - \overline{X_j^c}}{\sigma_j^c}.$$

Step 4: Compute $H = ZZ^t$.

Step 5: Compute the first q eigenvectors $v_1, v_2, \ldots, v_q$ of H and the associated eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q > 0$.

Step 6: For $i = 1, 2, \ldots, n$

	GRA	FRE	IOD	SAP
Linsed (L)	$[0.93, 0.935]$	$[-27, -18]$	$[170, 204]$	$[118, 196]$
Perilla (P)	$[0.93, 0.937]$	$[-5, -4]$	$[192, 208]$	$[188, 197]$
Cotton (Co)	$[0.916, 0.918]$	$[-6, -1]$	$[99, 113]$	$[189, 198]$
Sesame (S)	$[0.92, 0.926]$	$[-6, -4]$	$[104, 116]$	$[187, 193]$
Camellia (Ca)	$[0.916, 0.917]$	$[-25, -15]$	$[80, 82]$	$[189, 193]$
Olive (O)	$[0.914, 0.919]$	$[0, 6]$	$[79, 90]$	$[187, 196]$
Beef (B)	$[0.86, 0.87]$	$[30, 38]$	$[40, 48]$	$[190, 199]$
Hog (H)	$[0.858, 0.864]$	$[22, 32]$	$[53, 77]$	$[190, 202]$

Table 1. Oils and Fats data table.

Step 6.1: For $j = 1, , 2, \ldots, q$ compute

$$\underline{R}(X^i, Y^j) = \max \left[\sum_{k=1, v_{kj}<0}^{m} \overline{z}_{ki} v_{kj} + \sum_{k=1, v_{kj}>0}^{m} \underline{z}_{ki} v_{kj}, -1 \right].$$

$$\overline{R}(X^i, Y^j) = \min \left[\sum_{k=1, v_{kj}<0}^{m} \underline{z}_{ki} v_{kj} + \sum_{k=1, v_{kj}>0}^{m} \overline{z}_{ki} v_{kj}, 1 \right].$$

Step 7: For $i = 1, 2, \ldots, n$

Step 7.1: For $j = 1, 2, \ldots, q$ compute

$$u_{ij} = \frac{1}{\sqrt{\lambda_j}} \left(\sum_{k=1}^{m} z_{ki} v_{kj} \right).$$

Step 8: For $i = 1, 2, \ldots, m$

Step 8.1: For $j = 1, 2, \ldots, q$ compute

$$\underline{y_{ij}} = \sum_{k=1, u_{kj}<0}^{n} \overline{z}_{ik} u_{kj} + \sum_{k=1, u_{kj}>0}^{n} \underline{z}_{ik} u_{kj}$$

$$\overline{y_{ij}} = \sum_{k=1, u_{kj}<0}^{n} \underline{z}_{ik} u_{kj} + \sum_{k=1, u_{kj}>0}^{n} \overline{z}_{ik} u_{kj}$$

Step 9: END of the algorithm.

Example 1. To illustrate the symbolic circle of correlations we use Ichinos'data (oils and fats) that we present in Table 1. Each row of the data table refers to a class of oil described by 4 quantitative interval type variables, "Specific gravity", "Freezing point", "Iodine value" and "Saponification".

The symbolic correlations that we got using the Algorithm 1 are presented in Table 2 and the classical correlations between the gravity center of variables and the gravity center of the principal components (for the centers method) are presented in the Table 3. Notice that with this method the classic correlations are always contained in the interval that represents the symbolic correlation.
The symbolic circle of correlation to the oils and fats data obtained by the Duality Center Method is shown in Figure 2. The principal plane obtained by the Duality Center Method and corresponding to this correlation circle is presented in Figure 3 and principal components are presented in the Table 4. Note that the first

	PC1	PC2	PC3	PC4
GRA	$[0.827, 1.000]$	$[-0.443, -0.265]$	$[-0.038, 0.087]$	$[-0.238, -0.084]$
FRE	$[-1.000, -0.760]$	$[0.044, 0.372]$	$[-0.428, -0.220]$	$[-0.288, 0.019]$
IOD	$[0.726, 1.000]$	$[-0.124, 0.191]$	$[-0.565, -0.401]$	$[-0.024, 0.161]$
SAP	$[-1.000, 0.190]$	$[-1.000, 0.371]$	$[-0.442, 0.163]$	$[-0.231, 0.325]$

Table 2. Symbolic correlations between the variables and principal components with Duality Center Method.

	PC1	PC2	PC3	PC4
GRA	0.9210665	-0.3537703	0.0246894	-0.1608524
FRE	-0.9130654	0.2080771	-0.3238118	-0.1347643
IOD	0.8724116	0.0337627	-0.4827661	0.0685206
SAP	-0.7354523	-0.6613331	-0.1397354	0.0471425

Table 3. Classical correlations between the variables and principal components with Duality Center Method.

principal component separates hog and beef from the other oils and fats especially from linseed and perilla oils and is related to differences in freezing point, codine value and specific gravity. The second principal component separates linseed oil from the others due to the wide range of saponification values for this type of oil.

The next algorithm extends the algorithm proposed in [2, Cazes, Chouakria, Diday and Schektman (1997)], it works with the same variance–covariance matrix than [3, Chouakria (1998)], but we introduce some steps to compute the symbolic correlation using duality relations in order to plot the symbolic circle of correlation.

ALGORITHM 2: PRINCIPAL COMPONENT ANALYSIS ALGORITHM WITH Z^tZ

Input :

- m =number of symbolic objects.
- n =number of symbolic variables.
- The symbolic data table $X = \begin{pmatrix} [\underline{x_{11}}, \overline{x_{11}}] & [\underline{x_{12}}, \overline{x_{12}}] & \cdots & [\underline{x_{1n}}, \overline{x_{1n}}] \\ [\underline{x_{21}}, \overline{x_{21}}] & [\underline{x_{22}}, \overline{x_{22}}] & \cdots & [\underline{x_{2n}}, \overline{x_{2n}}] \\ \vdots & \vdots & \ddots & \vdots \\ [\underline{x_{m1}}, \overline{x_{m1}}] & [\underline{x_{m2}}, \overline{x_{m2}}] & \cdots & [\underline{x_{mn}}, \overline{x_{mn}}] \end{pmatrix}$.

Output :

	PC1	PC2	PC3	PC4
L	$[1.275, 4.733]$	$[-1.353, 4.428]$	$[-1.025, 1.289]$	$[-0.989, 0.989]$
P	$[1.059, 1.701]$	$[-1.128, -0.343]$	$[-1.508, -1.046]$	$[-0.134, 0.334]$
Co	$[-0.236, 0.399]$	$[-0.969, -0.213]$	$[-0.170, 0.368]$	$[-0.246, 0.204]$
S	$[0.154, 0.658]$	$[-0.745, -0.179]$	$[-0.027, 0.342]$	$[-0.369, 0.028]$
Ca	$[0.151, 0.613]$	$[-0.881, -0.437]$	$[0.807, 1.204]$	$[0.113, 0.538]$
O	$[-0.594, 0.100]$	$[-0.775, 0.043]$	$[0.019, 0.545]$	$[-0.645, -0.101]$
B	$[-3.046, -2.226]$	$[0.234, 1.162]$	$[-0.392, 0.152]$	$[-0.530, 0.193]$
H	$[-2.900, -1.841]$	$[0.020, 1.135]$	$[-0.729, 0.171]$	$[-0.105, 0.720]$

Table 4. Principal components with Duality Center Method.

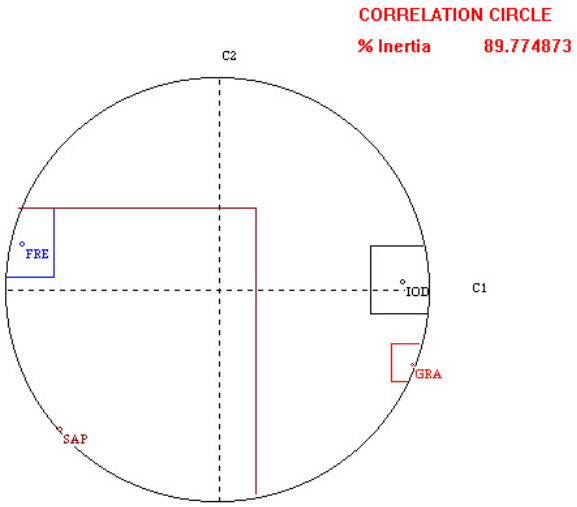

Fig. 2. Symbolic circle of correlation with Duality Centers Method.

– The symbolic correlation between the variables and the principal components in the following matrix:

$$R = \begin{pmatrix} [\underline{R}(X^1, Y^1), \overline{R}(X^1, Y^1)] & \cdots & [\underline{R}(X^1, Y^n), \overline{R}(X^1, Y^n)] \\ \vdots & \ddots & \vdots \\ [\underline{R}(X^n, Y^1), \overline{R}(X^n, Y^1)] & \cdots & [\underline{R}(X^n, Y^n), \overline{R}(X^n, Y^n)] \end{pmatrix}.$$

– The symbolic matrix with the first q principal components:

$$Y = \begin{pmatrix} [\underline{y_{11}}, \overline{y_{11}}] & [\underline{y_{12}}, \overline{y_{12}}] & \cdots & [\underline{y_{1q}}, \overline{y_{1q}}] \\ [\underline{y_{21}}, \overline{y_{21}}] & [\underline{y_{22}}, \overline{y_{22}}] & \cdots & [\underline{y_{2q}}, \overline{y_{2q}}] \\ \vdots & \vdots & \ddots & \vdots \\ [\underline{y_{m1}}, \overline{y_{m1}}] & [\underline{y_{m2}}, \overline{y_{m2}}] & \cdots & [\underline{y_{mq}}, \overline{y_{mq}}] \end{pmatrix}.$$

Step 1: Compute the matrix $X^c = (x_{ij}^c)_{\substack{i=1,2,\ldots,m \\ j=1,2,\ldots,n}}$ by:

$$x_{ij}^c = \frac{\underline{x_{ij}} + \overline{x_{ij}}}{2}.$$

Step 2: Compute the matrix $Z = (z_{ij})_{\substack{i=1,2,\ldots,m \\ j=1,2,\ldots,n}}$ by:

$$z_{ij} = \frac{1}{\sqrt{m}} \frac{x_{ij}^c - \overline{X_j^c}}{\sigma_j^c}.$$

Step 3: Compute the matrix $\underline{Z} = (\underline{z}_{ij})_{\substack{i=1,2,\ldots,m \\ j=1,2,\ldots,n}}$ and $\overline{Z} = (\overline{z}_{ij})_{\substack{i=1,2,\ldots,m \\ j=1,2,\ldots,n}}$ by:

$$\underline{z}_{ij} = \frac{1}{\sqrt{m}} \frac{\underline{x_{ij}} - \overline{X_j^c}}{\sigma_j^c},$$

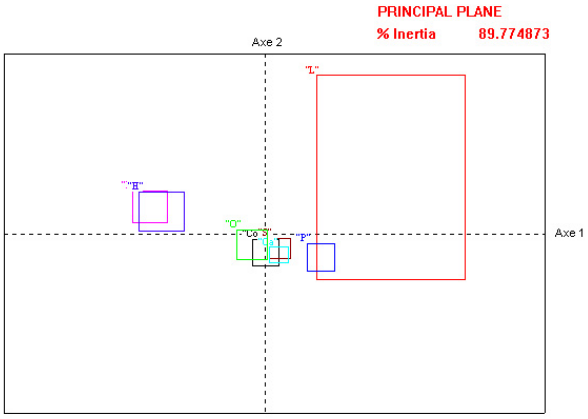

Fig. 3. Symbolic principal plane with Duality Centers Method.

$$\overline{z}_{ij} = \frac{1}{\sqrt{m}} \frac{\overline{x_{ij}} - \overline{X_j^c}}{\sigma_j^c}.$$

Step 4: Compute $R = Z^t Z$.

Step 5: Compute the first q eigenvectors $u_1, u_2, \ldots, u_q$ of R and the associated eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q > 0$.

Step 6: For $i = 1, 2, \ldots, m$

Step 6.1: For $j = 1, 2, \ldots, q$ compute

$$\underline{y_{ij}} = \sum_{k=1, u_{kj}<0}^{n} \overline{z}_{ik} u_{kj} + \sum_{k=1, u_{kj}>0}^{n} \underline{z}_{ik} u_{kj}$$

$$\overline{y_{ij}} = \sum_{k=1, u_{kj}<0}^{n} \underline{z}_{ik} u_{kj} + \sum_{k=1, u_{kj}>0}^{n} \overline{z}_{ik} u_{kj}$$

Step 7: For $i = 1, 2, \ldots, m$

Step 7.1: For $j = 1, 2, \ldots, q$ compute

$$v_{ij} = \frac{1}{\sqrt{\lambda_j}} \left(\sum_{k=1}^{m} z_{ik} u_{kj} \right).$$

Step 8: For $i = 1, 2, \ldots, m$

Step 8.1: For $j = 1, 2, \ldots, q$ compute

$$\underline{R}(X^i, Y^j) = \max \left[\sum_{k=1, v_{kj}<0}^{m} \overline{z}_{ki} v_{kj} + \sum_{k=1, v_{kj}>0}^{m} \underline{z}_{ki} v_{kj}, -1 \right].$$

$$\overline{R}(X^i, Y^j) = \min \left[\sum_{k=1, v_{kj}<0}^{m} \underline{z}_{ki} v_{kj} + \sum_{k=1, v_{kj}>0}^{m} \overline{z}_{ki} v_{kj}, 1 \right].$$

	PC1	PC2	PC3	PC4
GRA	$[-1.000, -0.827]$	$[-0.443, -0.265]$	$[-0.038, 0.087]$	$[-0.238, -0.084]$
FRE	$[0.760, 1.000]$	$[0.044, 0.372]$	$[-0.428, -0.220]$	$[-0.288, 0.019]$
IOD	$[-1.000, -0.726]$	$[-0.124, 0.191]$	$[-0.565, -0.401]$	$[-0.024, 0.161]$
SAP	$[-0.190, 1.000]$	$[-1.000, 0.371]$	$[-0.442, 0.163]$	$[-0.231, 0.325]$

Table 5. Symbolic correlations between the variables and principal components with Center Duality Method.

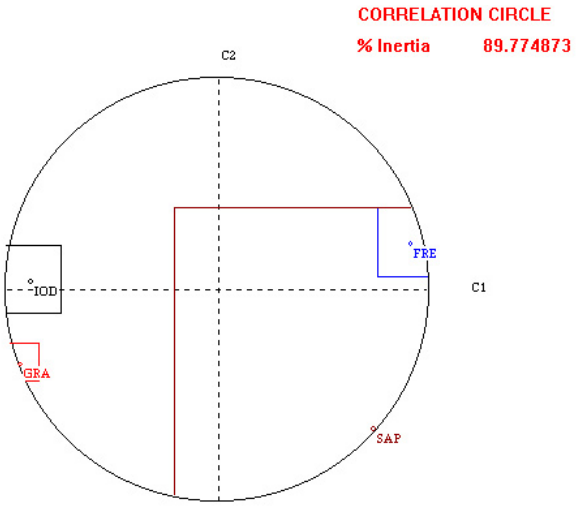

Fig. 4. Symbolic circle of correlation with Duality Centers Method starting with Z^tZ.

Step 9: END of the algorithm.

Example 2. To illustrate the duality symbolic circle of correlations using Z^tZ we use again Ichinos'data (oils and fats) that we have presented in the Table 1. The symbolic correlations that we got using the algorithm 2 are presented in Table 5.

The symbolic circle of correlation to the oils and fats data got with data of the table 5 is shown in Figure 4. The principal plane corresponding to this correlation circle is presented in Figure 5. Note that the results are the same as those obtained from the first algorithm except that the first axis is reversed.
The size of the matrix ZZ^t is $m \times m$ while the size of Z^tZ is $n \times n$, sometimes ZZ^t is very big and Z^tZ is very small, in this case is better to use the algorithm 2 than the algorithm 1, or inversely Z^tZ is very big and ZZ^t is very small then it is faster the algorithm 1 than the algorithm 2. Hence, considering if $m \leq n$ or not we propose the algorithm 3.

ALGORITHM 3: PRINCIPAL COMPONENT ANALYSIS OPTIMAL ALGORITHM

Input :

- m =number of symbolic objects.
- n =number of symbolic variables.

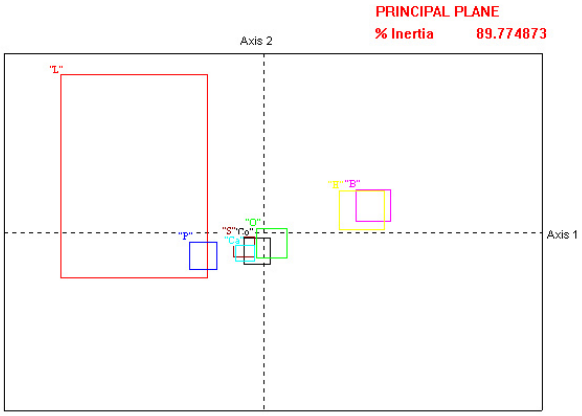

Fig. 5. Symbolic principal plane with Duality Centers Method starting with Z^tZ.

– The symbolic data table $X = \begin{pmatrix} [\underline{x_{11}}, \overline{x_{11}}] & [\underline{x_{12}}, \overline{x_{12}}] & \cdots & [\underline{x_{1n}}, \overline{x_{1n}}] \\ [\underline{x_{21}}, \overline{x_{21}}] & [\underline{x_{22}}, \overline{x_{22}}] & \cdots & [\underline{x_{2n}}, \overline{x_{2n}}] \\ \vdots & \vdots & \ddots & \vdots \\ [\underline{x_{m1}}, \overline{x_{m1}}] & [\underline{x_{m2}}, \overline{x_{m2}}] & \cdots & [\underline{x_{mn}}, \overline{x_{mn}}] \end{pmatrix}$.

Output :

– The symbolic correlation between the variables and the principal components in the following matrix:

$$R = \begin{pmatrix} [\underline{R}(X^1, Y^1), \overline{R}(X^1, Y^1)] & \cdots & [\underline{R}(X^1, Y^n), \overline{R}(X^1, Y^n)] \\ \vdots & \ddots & \vdots \\ [\underline{R}(X^n, Y^1), \overline{R}(X^n, Y^1)] & \cdots & [\underline{R}(X^n, Y^n), \overline{R}(X^n, Y^n)] \end{pmatrix}.$$

– The symbolic matrix with the first q principal components:

$$Y = \begin{pmatrix} [\underline{y_{11}}, \overline{y_{11}}] & [\underline{y_{12}}, \overline{y_{12}}] & \cdots & [\underline{y_{1q}}, \overline{y_{1q}}] \\ [\underline{y_{21}}, \overline{y_{21}}] & [\underline{y_{22}}, \overline{y_{22}}] & \cdots & [\underline{y_{2q}}, \overline{y_{2q}}] \\ \vdots & \vdots & \ddots & \vdots \\ [\underline{y_{m1}}, \overline{y_{m1}}] & [\underline{y_{m2}}, \overline{y_{m2}}] & \cdots & [\underline{y_{mq}}, \overline{y_{mq}}] \end{pmatrix}.$$

Step 1: If $m \leq n$ then we apply algorithm 1 else we apply algorithm 2.
Step 2: END of the algorithm.

3 Extension for principal component analysis to histogram data

To extend PCA to histogram type data we develop the idea first proposed in [5, Diday (1998)]. We represent each histogram–individual by a succession of k interval–individuals (the first one included in the second one, the second one included in the third one and so on) where k is the maximum number of modalities taken by some variable in the input symbolic data table.
Instead of representing the histograms in the factorial plane, we are going to represent the Empirical Distribution Function F_Y defined in [1, Bock and Diday (2000)] associated with each histogram. In other words

if we have a histogram variable Y on a set $E = \{a_1, a_2, \ldots\}$ of objects with domain $\mathcal{Y}$ represented by the mapping $Y(a) = (U(a), \pi_a)$, for $a \in E$, where π_a is frequency distribution, then in the algorithm we will use the function $F(x) = \sum_{i \,/\, \pi_i \leq x} \pi_i$ instead of the histogram.

Definition 1. Let $X = (x_{ij})_{\substack{i=1,2,\ldots,m \\ j=1,2,\ldots,n}}$ be a symbolic data table with continuous, interval and histogram type variables, and let be $k = \max\{s$, where s is the number of modalities of $Y^j\}$, $j = 1, 2, \ldots, n$ where Y^j is of histogram type (If all the variables are interval type then $k = 1$). We define the vector–succession of intervals associated with each cell of X as:

1. if $x_{ij} = [a, b]$ then the associated column–vector of intervals is:

$$x_{ij}^{\downarrow} = \begin{bmatrix} [a,b] \\ [a,b] \\ \vdots \\ [a,b] \end{bmatrix}_{k \times 1}.$$

2. If $x_{ij} = (1(p_1), 2(p_2), \ldots, s(p_s))$ with $s \leq k$ (histogram) then the associated column–vector of intervals is:

$$x_{ij}^{\downarrow} = \begin{bmatrix} [0, p_1] \\ [0, p_1 + p_2] \\ \vdots \\ \left[0, \sum_{w=1}^{s} p_w\right] \end{bmatrix}_{k \times 1}.$$

3. If $x_{ij} = a$ then the associated column–vector of intervals is:

$$x_{ij}^{\downarrow} = \begin{bmatrix} [a,a] \\ [a,a] \\ \vdots \\ [a,a] \end{bmatrix}_{k \times 1}.$$

Definition 2. Let $X = (x_{ij})_{\substack{i=1,2,\ldots,m \\ j=1,2,\ldots,n}}$ be a symbolic data table with variables type continuous, interval and histogram. We define the matrix $X^{\downarrow} = (x_{ij}^{\downarrow})$ for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$. It is important to note that $X^{\downarrow}$ has $m \cdot k$ rows (k like in the previous definition) and n columns.

Example 3. If $X = \begin{bmatrix} [1,3] & (1(0.2), 2(0.3), 3(0.5)) \\ [7,9] & (1(0.8), 2(0.1), 3(0.1)) \end{bmatrix}$ then

$$X^{\downarrow} = \begin{bmatrix} [1,3] & [0.0000, 0.2000] \\ [1,3] & [0.0000, 0.5000] \\ [1,3] & [0.0000, 1.0000] \\ [7,9] & [0.0000, 0.8000] \\ [7,9] & [0.0000, 0.9000] \\ [7,9] & [0.0000, 1.0000] \end{bmatrix}.$$

The idea is to apply the algorithm 3 to the matrix $X^{\downarrow}$. With this Principal Components Analysis we can find the *shape* of the "individual–histogram" in the principal plane, However all the individual–histograms will be projected in about the same position around the origin. So in addition we apply another principal component analysis in order to find a good *cluster structure* to the individual–histogram. Therefore we will apply a classical principal component analysis to the matrix presented in the followings definitions.

Definition 3. Let $X = (x_{ij})_{\substack{i=1,2,\ldots,m \\ j=1,2,\ldots,n}}$ be a symbolic data table with variables type continuous, interval and histogram. We define the associated row–vector with each cell of X as:

1. If $x_{ij} = [a, b]$ then the row–vector associated is:
$$\overrightarrow{x_{ij}} = \left[\frac{a+b}{2}\right]_{1\times 1}.$$
2. If $x_{ij} = (1(p_1), 2(p_2), \ldots, s(p_s))$ where s is number of modalities of the j–th variable, then the associated row–vector is:
$$\overrightarrow{x_{ij}} = [p_1, p_2, \ldots, p_s]_{1\times s}.$$
3. If $x_{ij} = a$ then the associated row–vector is:
$$\overrightarrow{x_{ij}} = [a]_{1\times 1}.$$

Definition 4. Let $X = (x_{ij})_{\substack{i=1,2,\ldots,m \\ j=1,2,\ldots,n}}$ be a symbolic data table with variables type continuous, interval and histogram. We define the matrix $X^{\rightarrow} = (\overrightarrow{x_{ij}})$ of m rows and $p = \sum_{j=1}^{n} s_j$ columns, where

$$s_j = \begin{cases} \text{number of modalities of the variable} & \text{If the variable } j \text{ is histogram type,} \\ 1 & \text{If the variable } j \text{ is interval type,} \\ 1 & \text{If the variable } j \text{ is continue type.} \end{cases}$$

Example 4. If $X = \begin{bmatrix} [1,3] & (1(0.2), 2(0.3), 3(0.5)) \\ [7,9] & (1(0.8), 2(0.1), 3(0.1)) \end{bmatrix}$ then

$$X^{\rightarrow} = \begin{bmatrix} 2 & 0.2 & 0.3 & 0.5 \\ 8 & 0.8 & 0.1 & 0.1 \end{bmatrix}.$$

The idea of the algorithm is to apply a principal components analysis to the matrix $X^{\downarrow}$ to find the shape of the individual–histogram, and then to apply another principal component analysis to the matrix $X^{\rightarrow}$. Using this last principal components, we will translate each individual–histograms to find the cluster structure of individual–histograms in the principal plane.

ALGORITHM 4: HISTOGRAM PRINCIPAL COMPONENTS ANALYSIS

Input :

- m =number of symbolic objects.
- n =number of symbolic variables.
- The symbolic data table $X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$.

Output :

– The symbolic matrix with the first q principal components:

$$Y = \begin{pmatrix} y_{11}^{\downarrow} & y_{12}^{\downarrow} & \cdots & y_{1q}^{\downarrow} \\ y_{21}^{\downarrow} & y_{22}^{\downarrow} & \cdots & y_{2q}^{\downarrow} \\ \vdots & \vdots & \ddots & \vdots \\ y_{m1}^{\downarrow} & y_{m2}^{\downarrow} & \cdots & y_{mq}^{\downarrow} \end{pmatrix},$$

where (k like in definition 1):

$$y_{ij}^{\downarrow} = \begin{bmatrix} \left[\underline{y_{ij}^{1}}, \overline{y_{ij}^{1}}\right] \\ \left[\underline{y_{ij}^{2}}, \overline{y_{ij}^{2}}\right] \\ \vdots \\ \left[\underline{y_{ij}^{k}}, \overline{y_{ij}^{k}}\right] \end{bmatrix}.$$

Step 1: Compute the matrix $X^{\downarrow}$ of the definition 2.
Step 2: Apply the algorithm 3 taking as input $X^{\downarrow}$. It will produce the matrix:

$$\widehat{Y}^{\downarrow} = \begin{pmatrix} \widehat{y}_{11}^{\downarrow} & \widehat{y}_{12}^{\downarrow} & \cdots & \widehat{y}_{1q_1}^{\downarrow} \\ \widehat{y}_{21}^{\downarrow} & \widehat{y}_{22}^{\downarrow} & \cdots & \widehat{y}_{2q_1}^{\downarrow} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{y}_{m1}^{\downarrow} & \widehat{y}_{m2}^{\downarrow} & \cdots & \widehat{y}_{mq_1}^{\downarrow} \end{pmatrix},$$

where (k like in definition 1):

$$\widehat{y}_{ij}^{\downarrow} = \begin{bmatrix} \left[\underline{\widehat{y}_{ij}^{1}}, \overline{\widehat{y}_{ij}^{1}}\right] \\ \left[\underline{\widehat{y}_{ij}^{2}}, \overline{\widehat{y}_{ij}^{2}}\right] \\ \vdots \\ \left[\underline{\widehat{y}_{ij}^{k}}, \overline{\widehat{y}_{ij}^{k}}\right] \end{bmatrix}.$$

for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, q_1$ with $q_1 \leq n$.
Step 3: Compute the matrix $X^{\rightarrow}$ of the definition 4.
Step 4: Apply a classical principal component analysis to the matrix $X^{\rightarrow}$. It will produce the matrix:

$$\widetilde{Y}^{\rightarrow} = \begin{pmatrix} \widetilde{y}_{11} & \widetilde{y}_{12} & \cdots & \widetilde{y}_{1q_2} \\ \widetilde{y}_{21} & \widetilde{y}_{22} & \cdots & \widetilde{y}_{2q_2} \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{y}_{m1} & \widetilde{y}_{m2} & \cdots & \widetilde{y}_{mq_2} \end{pmatrix},$$

where $q_2 \leq p = \sum_{j=1}^{n} s_j$ (s_j like in definition 4):
Step 5: $q = \min(q_1, q_2)$.
Step 6: Compute the first q principal components:

$$Y = \begin{pmatrix} y_{11}^{\downarrow} & y_{12}^{\downarrow} & \cdots & y_{1q}^{\downarrow} \\ y_{21}^{\downarrow} & y_{22}^{\downarrow} & \cdots & y_{2q}^{\downarrow} \\ \vdots & \vdots & \ddots & \vdots \\ y_{m1}^{\downarrow} & y_{m2}^{\downarrow} & \cdots & y_{mq}^{\downarrow} \end{pmatrix},$$

using the translation:

$$y_{ij}^{\downarrow} = \begin{bmatrix} \left[\underline{y_{ij}^1}, \overline{y_{ij}^1}\right] \\ \left[\underline{y_{ij}^2}, \overline{y_{ij}^2}\right] \\ \vdots \\ \left[\underline{y_{ij}^k}, \overline{y_{ij}^k}\right] \end{bmatrix} = \begin{bmatrix} \left[\underline{\widehat{y}_{ij}^1} + \widetilde{y}_{ij}, \overline{\widehat{y}_{ij}^1} + \widetilde{y}_{ij}\right] \\ \left[\underline{\widehat{y}_{ij}^2} + \widetilde{y}_{ij}, \overline{\widehat{y}_{ij}^2} + \widetilde{y}_{ij}\right] \\ \vdots \\ \left[\underline{\widehat{y}_{ij}^k} + \widetilde{y}_{ij}, \overline{\widehat{y}_{ij}^k} + \widetilde{y}_{ij}\right] \end{bmatrix}$$

Step 7: End of the algorithm.

3.1 The interpretation

To explain how to interpret the Histogram Principal Components Analysis we will use one small example. The interpretation of the position of the histogram–individual in the principal plane is the same as in the classical principal component analysis situation. We shall explain the interpretation of the succession of rectangles that represents each individual.

Example 5. Let be

$$X = \begin{array}{|c|c|c|} \hline & \text{VAR-1} & \text{VAR-2} \\ \hline \text{Ind1} & (1(0.1), 2(0.4), 3(0.5)) & (1(0.2), 2(0.3), 3(0.5)) \\ \hline \text{Ind2} & (1(0.7), 2(0.2), 3(0.1)) & (1(0.8), 2(0.1), 3(0.1)) \\ \hline \end{array}.$$

If we apply the Histogram Principal Components Analysis to the previous data table we get the principal plane that we show in the Figure 6.
The smallest rectangle of the projection of the individual–1 (Ind1) represents the probability that individual–1 takes the modality 1 for the variable 1 or the modality 1 for the variable 2. The size of the rectangle agrees with the representation of the individual–1 in the histogram, because the value of the modality 1 for the variable 1 is 0.1 and the value of the modality 1 for the variable 2 is 0.2, i.e. the mean for the modality 1 is 0.15. The second rectangle of the projection of the individual–1 represents the probability that individual–1 takes the modality 1 or the modality 2 for the variable 1, or the probability that individual–1 takes the modality 1 or modality 2 for the variable 2. The size of the second rectangle also agrees with the representation of individual–1 in the histogram, because the value of the empirical distribution function for the modality 2 of the variable 1 is 0.5 and the value of the empirical distribution function for the modality 2 of the variable 2 is also 0.5. The third rectangle of individual–1 represents the probability 1, that is the probability that individual 1 takes any of the modalities.
The smallest rectangle of the projection of individual–2 (Ind2) is bigger than the smallest rectangle of the projection of the individual–1 (see Figure 6); it is consistent with the interpretation, because the probability for individual–2 to take the modality 1 for the variable 1 is 0.7 and the probability for individual–2 to take the modality 1 for the variable 2 is 0.8, i.e. the mean of taken the modality 1 is 0.75. This value is bigger than the same value for individual–1 that is 1.5; that's why, the smallest rectangle of the projection of "Ind1" is smaller than the smallest rectangle of the projection of "Ind2". For the same reasons the second rectangle of the projection of "Ind1" is smaller than the second rectangle of the projection of "Ind2".

Example 6. In this example we present the execution of the algorithm 4 with the household consumption in Great Britain data, this symbolic data table partially presented in (15). In the data matrix of 15 variables and 25 individuals.

$$X = \begin{bmatrix} [1,4] & (1(0.09), 2(0.90)) & (1(0.86), 2(0.07), 3(0.05), 4(0.00)) & \cdots \\ [1,6] & (1(0.12), 2(0.87)) & (1(0.71), 2(0.11), 3(0.10), 4(0.06), 5(0.00), 6(0.00)) & \cdots \\ [1,6] & (1(0.23), 2(0.76) & (1(0.82), 2(0.07), 3(0.09), 4(0.00), 5(0.00)) & \cdots \\ [1,4] & (1(0.19), 2(0.80)) & (1(0.76), 2(0.06), 3(0.10), 4(0.05) & \cdots \\ [1,6] & (1(0.22), 2(0.77)) & (1(0.89), 3(0.00), 3(0.81), 4(0.01)) & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \quad (15)$$

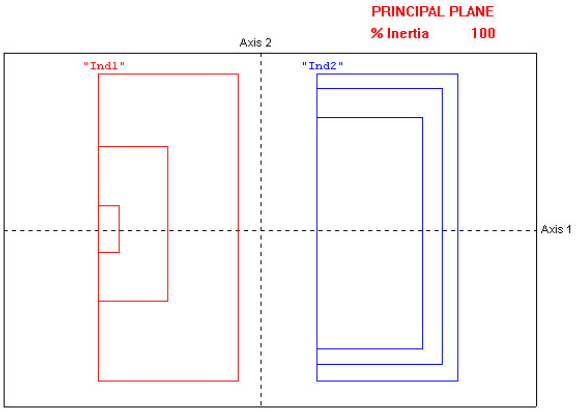

Fig. 6. Histogram Principal Component Plane.

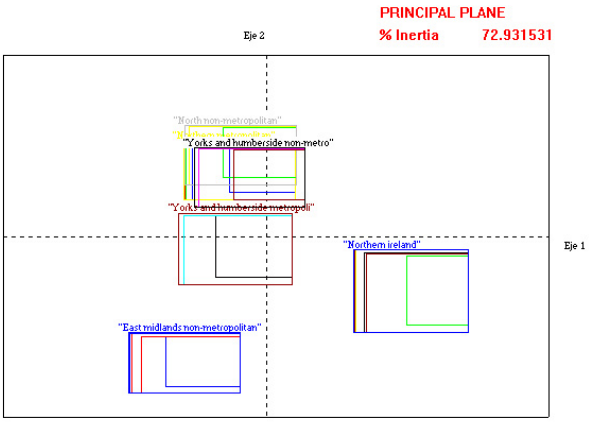

Fig. 7. Principal plane with data of continuous, interval and histogram type.

Applying the algorithm 4 proposed before we get the principal plane of Figure 7.
The individuals “East midlands non-metropolitan” and “Northern Ireland” are isolated and the individuals “North non-metropolitan”, “Yorks and Humberside metropolis”, “Yorks and Humberside non-metro” and “East midlands non-metropolitan” are grouped.

4 Conclusion

Using a duality theorem for Centers PCA we have developed a method to compute the symbolic correlation circle in an efficient manner. The symbolic correlation circle permits the avaluation of the maximum and minimum values of the correlation of each variable with each principal component. Moreover, we have extended the Centers PCA for interval–type data to deal with variables of mixed–type that is they may be a combination of histogram–type, interval–type or single–value or any one or two of the above. This extension of PCA permits the use of this valuable tool for data reduction to be used in complex situations

with data sets that are preprocessed and aggregated because the very size of the original data sets wold prohibit the use of classical PCA. We have ilustrated these techniques with aggregated data sets (the oils and fats data and the household consumption in Great Britain), data sets too large to permit implementing PCA on the original data. The PCA of these aggregated sets using our techniques permits the study of the relationships among the objecs and the variables.

References

1. Bock H-H. and Diday E. (eds.) *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data.* Springer Verlag, Heidelberg, 425 pages, ISBN 3-540-66619-2, 2000.
2. Cazes P., Chouakria A., Diday E. et Schektman Y. *Extension de l'analyse en composantes principales à des données de type intervalle*, Rev. Statistique Appliquée, Vol. XLV Num. 3 pag. 5-24, Francia, 1997.
3. Chouakria A. *Extension des méthodes d'analysis factorialle à des données de type intervalle*, Thèse de doctorat, Université Paris IX Dauphine, 1998.
4. Diday E. *Introduction l'approche symbolique en Analyse des Données*. Première Journées Symbolique-Numérique. Université Paris IX Dauphine. Décembre 1987.
5. Diday E. *L'Analyse des Données Symboliques: un cadre théorique et des outils*. Cahiers du CEREMADE, 1998.
6. Rodríguez R. *Classification et Modèles Linéaires en Analyse des Données Symboliques*, Thèse de doctorat, Université Paris IX Dauphine, 2000.