Computational Statistics (2000) 15:73–87

Principal component analysis of interval data: a symbolic data analysis approach

Carlo N. Lauro[1] and Francesco Palumbo[2]

[1] Dipartimento di Matematica e Statistica
Università "Federico II" – Napoli, Italy
[2] Dipartimento di Istituzioni Economiche e Finanziarie
Università di Macerata – Macerata, Italy

Summary

The present paper deals with the study of continuous interval data by means of suitable Principal Component Analyses (PCA). Statistical units described by interval data can be assumed as special cases of Symbolic Objects (SO) (Diday, 1987). In Symbolic Data Analysis (SDA), these data are represented as hypercubes. In the present paper, we propose some extensions of the PCA with the aim of representing, in a space of reduced dimensions, images of such hypercubes, pointing out differences and similarities according to their structural features.

Keywords: Principal Components, Interval Data, Symbolic Objects

1 Introduction

Statistical methods have been mainly developed for the analysis of single valued variables. However, in real life there are many situations in which the

use of these variables may cause severe loss of information.
Dealing with quantitative variables, there are many cases in which a more complete information can be surely achieved by describing a set of statistical units in terms of interval data.
For example, daily temperatures registered as minimum and maximum values offer a more realistic view on the weather conditions variations with respect to the simple average values. Another example can be given by the financial series. The minimum and the maximum transaction prices, daily recorded for a set of stocks, represent a more relevant information for experts in order to evaluate the stocks tendency and volatility in the same day.

Most widely used approaches to ***interval data*** analysis treat intervals as spread ranges with respect to a central value. The spread is generally assumed as the consequence of a measurement error and is considered as a perturbation in the data. In this paper we do not consider the interval central value, but we only point the attention on the minimum and maximum values. These are evaluated as two different and related aspects of the same phenomenon.

The statistical treatment of interval data has been recently considered in the context of Symbolic Data Analysis (SDA) (Diday, 1996), whose aim is to extend classical statistical methods to the study of more complex data structures with respect to the simple tabular model ***individuals*** × ***variables***. The present paper introduces some new Principal Component Analyses (PCA) techniques in order to visualize and compare structures of interval data. Statistical units described by interval variables can be considered as special cases of symbolic data, in which only quantitative variables are considered. Moreover, the SDA approach for the interval data treatment offers many useful tools that can be helpful in the interpretation of results. For these reasons our approach to interval data representation is presented by adopting the notations and definitions of the SDA domain.

2 Basic concepts and notation

Let Ω be a set of ω_i $(1 \leq i \leq n)$ Symbolic Objects (SO), that are described by p variables or descriptors: $Y = \{y_1, \ldots, y_j, \ldots, y_p\}$.

Nowadays the SDA is based either on numerical treatments of suitably coded SO's followed by symbolic interpretation of results, or on symbolic methods that directly process the ***symbolic*** descriptors.
In the following we are set in the first approach framework in order to analyze SO's described only by quantitative interval variables.

Afterwards, the generic variable y_j represents no more a single valued variable, like in the classical data analysis, but refers to the lower y_j and the upper $\overline{y_j}$ bounds of the interval assumed by the j-th variable. The notation $O_1, \ldots, O_j, \ldots, O_p$ is adopted to define the domains of the p variables. Therefore, the ***symbolic data matrix*** $\mathbf{Y}$ has dimensions $(n \times 2p)$.
A basic kind of SO, denoted as event, is indicated as: $e_j = [y_j = V_j]$, where $V_j \subseteq O_j$. The logical conjunction of events defines the so called ***symbolic assertion object*** $a_i = \wedge_{j=1}^{p} [y_j = V_j^i]$. In order to process, by numerical methods, SO's described by interval variables, they are re-coded combining *min* and *max* values into the so called ***vertices*** data matrix. In the simple case of $p = 2$, the generic SO ω_i description is associated with the i^{th} row of the interval data matrix $\mathbf{Y}$:

$$\mathbf{Y} = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ \underline{y_{i,1}} & \overline{y_{i,1}} & \underline{y_{i,2}} & \overline{y_{i,2}} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

The related vertices coordinates with respect to the new variables z_1 and z_2 - having the same domains of y_1 and y_2, respectively - correspond to the rows of the matrix Z_i:

$$Z_i = \begin{array}{c} \begin{array}{cc} z_1 & z_2 \end{array} \\ \begin{pmatrix} \underline{y_{i,1}} & \underline{y_{i,2}} \\ \underline{y_{i,1}} & \overline{y_{i,2}} \\ \overline{y_{i,1}} & \underline{y_{i,2}} \\ \overline{y_{i,1}} & \overline{y_{i,2}} \end{pmatrix} \end{array}$$

In a geometric view (Figure 1) the generic SO ω_i is represented by a rectangle, having $2^p = 4$ vertices corresponding to all possible (*min*,*max*) combinations.

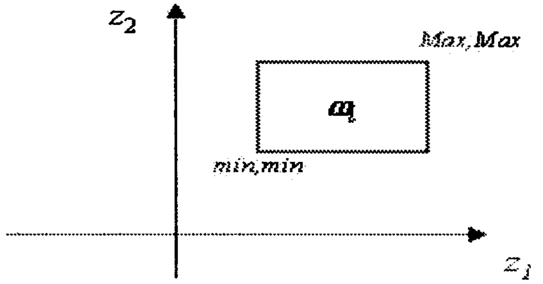

Figure 1: 2-D space SO graphical representation

In the general case of any p variables, each coding matrix Z_i will have 2^p rows and p columns. The stack coding matrix $\mathbf{Z}$ is obtained by superposing the n coding matrices Z_i (with $1 \leq i \leq n$). Matrix $\mathbf{Z}$ has $N = n2^p$ rows and p columns, and represents the numerical coding of the n SO's. Without loss of generality we can assume that z_j variables are standardized.

3 Vertices Principal Component Analysis

The extension of the Principal Component Analysis (PCA) to interval data has been proposed by Cazes, Chouakria, Diday and Schektman (1997) and by Chouakria, Diday and Cazes (1998) as "*Vertices Principal Component Analysis*" (V–PCA).
In the original proposal, the V–PCA consists in performing a classical PCA on the standardized matrix $\mathbf{Z}$. In this way, vertices are elements of the subspace $\mathbb{R}^p$, whereas the p quantitative descriptors are elements of $\mathbb{R}^N$. V–PCA looks for a suitable subspace where to represent SO's and, from a dual point of view, to represent the p variables.
As in classical PCA the optimal sub–space is here spanned by the axes $\mathbf{v}_m$ (with $1 \leq m \leq p$), maximizing the sum of squares of projected vertices coordinates $\psi_m = \mathbf{Z}\mathbf{v}_m$:

$$\psi'_m \psi_m = \mathbf{v}'_m \mathbf{Z}'\mathbf{Z}\mathbf{v}_m \tag{1}$$

with $\mathbf{v}'_m \mathbf{v}_{m'} = 0$ for $m \neq m'$ and $\mathbf{v}'_m \mathbf{v}_{m'} = 1$ for $m = m'$. Therefore, the characteristic equation of V–PCA in $\mathbb{R}^N$ is given by:

$$\frac{1}{N}\mathbf{Z}'\mathbf{Z}\mathbf{v}_m = \lambda_m \mathbf{v}_m \qquad 1 \leq m \leq p \tag{2}$$

where $\mathbf{v}_m$ and λ_m are the generic eigenvector and the generic eigenvalue respectively, associated to the matrix $\frac{1}{N}\mathbf{Z}'\mathbf{Z}$.
Performing the analysis in the $\mathbb{R}^p$ space we define the equation:

$$\frac{1}{N}\mathbf{Z}\mathbf{Z}'\mathbf{w}_m = \lambda_m \mathbf{w}_m \qquad 1 \leq m \leq p \tag{3}$$

that has same non-zero eigenvalues of (2), but different eigenvectors, holding the following relation: $\mathbf{v}_m = \lambda_m^{-1/2}\mathbf{Z}'\mathbf{w}_m$. The principal axes interpretation, also in V–PCA, is accomplished with reference to the variables z_j's having maximal contributions, see Lebart, Morineau and Piron (1995).

In case of normalized variables, contributions are calculated as the squared correlation variable/factor: $\mathrm{CTA}_{j,m} = (\lambda_m^{1/2} v_{j,m})^2 / \lambda_m = v_{j,m}^2$.

The vertices coordinates of the SO ω_i on the principal axes are given by the vector: $\psi_{i,m} = Z_i \mathbf{v}_m$. The representation of ω_i on the generic axis m is given by the *segment* including all vertices projections. Adopting the same criterion in a two dimensional space spanned by the axes m and m', the extreme vertices projections define a rectangle called ***maximum covering area rectangle*** (MCAR).

Even if the MCAR representation of SO's on a plane is coherent with the hypercubes associated to each SO, it does not rarely happen that they are ***oversized*** with respect to the real SO in $\mathbb{R}^p$.
In order to overcome this drawback, Chouakria et al. (1998) proposed to reduce MCAR's dimensions by retaining in the representation only vertices with ***high quality representation.*** The latter is measured in terms of squared cosines criterion: $\mathrm{CRT}_{q,m} = \sum_{j=1}^{p} \left(z_{q,j}v_{j,m}\right)^2 / \sum_j z_{q,j}^2$ $(1 \leq m \leq p)$, where $z_{q,i}$ is the generic vertex of the SO ω_i.

It seems clear that these drawbacks - concerning MCAR's representations of SO's - depend on the optimized criterion in V–PCA. In fact, the V–PCA aims at optimizing the vertices representation at the expense of SO's one; vertices are assumed to be independent and any relationship among vertices belonging to the same SO is lost. The SO's rebuilding - via MCAR's - is done ***ex post*** and consequently the principal axes interpretation cannot be referred to SO's main characteristics.

In the next sections, we propose some new approaches allowing to treat, in a consistent way vertices belonging to the same SO [§4] and to stress SO's ***size and shape*** in graphical representation [§5].

4 Symbolic Objects based PCA

In this section, we introduce a new PCA that takes into account the ***necessary*** vertices cohesion constraint. In this direction, our proposal is to maximize the variance among SO's instead of the total vertices variance, as done in V–PCA. Hereinafter, we call the proposed approach ***Symbolic Object Principal Component Analysis*** (SO–PCA). The method represents the more realistic PCA extension to interval data and it can be extended - without loss of generality - to any numerically handable data structure.
Analogously to classical PCA, SO-PCA is based on the search of axes maximizing the ***differences*** among statistical units which, in this context, are SO's.

The main idea behind the method consists in maximizing the ***between*** SO's variance matrix:

$$\frac{1}{N}\mathbf{Z}'\mathbf{A}\left(\mathbf{A}'\mathbf{A}\right)^{-1}\mathbf{A}'\mathbf{Z},$$

where we denote by $\mathbf{A}$ the boolean matrix $(N \times n)$ describing the belonging of the N vertices to the n SO's.
Looking for the axes that maximize the difference among SO's, the solution in the space $\mathbb{R}^N$ is obtained by the following characteristic equation:

$$\begin{aligned} \frac{1}{N}[\mathbf{Z}'\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{Z}]\tilde{\mathbf{v}}_m &= \\ = \frac{1}{N}\mathbf{Z}'\mathbf{P}_{\mathbf{A}}\mathbf{Z} &= \tilde{\lambda}_m\tilde{\mathbf{v}}_m, \end{aligned} \tag{4}$$

where $\tilde{\mathbf{v}}_m$ is defined under the orthonormality constraints already expressed in (1). Considering that $\mathbf{P}_{\mathbf{A}} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ is an orthogonal projector matrix, this approach represents a special case of the so called PCA with respect to a reference sub–space (D'Ambra and Lauro, 1982), here spanned by the columns of $\mathbf{A}$.
Vertices coordinates, of the generic SO ω_i on the axis m, are given by the vector $\tilde{\psi}_{i,m} = Z_i\tilde{\mathbf{v}}_m$.
The analysis can also be exploited in $\mathbb{R}^p$ and leads to the following eigen-equation:

$$(\mathbf{A}'\mathbf{A})^{-1/2}(\mathbf{A}'\mathbf{Z}\mathbf{Z}'\mathbf{A})(\mathbf{A}'\mathbf{A})^{-1/2} = \tilde{\lambda}_m\tilde{\mathbf{w}}_m.$$

Variable contributions are defined like in V–PCA, whereas, we propose to evaluate the SO's representation considering the whole SO and representing only SO's having high CRT values. In fact, in V–PCA vertices contributions are individually evaluated and the MCAR's are built only with vertices having high contribution. The proposed contribution measure CRT_{ω_i} is obtained by:

$$\mathrm{CRT}_{\omega_i,m} = \sum_{q\in\omega_i}\sum_{j=1}^{p}(z_{q,j}\tilde{v}_{j,m})^2 / \sum_{q\in\omega_i}\sum_{j=1}^{p} z_{q,j}^2 \tag{5}$$

The graphical interpretation of SO's - as in V–PCA - is made by means of *maximum covering area rectangles*. Nevertheless, in our approach the adopted constraints on the vertices make the analysis consistent with the aim of PCA on SO's and offer a more *compact* SO's representation.

The previous treatment of interval data remains in the context of the so called multiple PCA (Escofier and Pagés, 1988) where vertices of each SO are referred to the general mean of each variable.
The partial PCA can be alternatively used to better stress the differences among SO's. The following section shows a partial PCA in which vertices are centered with respect to the *min* value.

5 A PCA on the range transformation of interval data

In the classical PCA, each statistical unit is represented by a point, while in SDA we have to cope with SO's shape and size. In order to take into account these SO's structural elements, we use the *range transformation* $V_j^i = [\overline{y_{i,j}} - \underline{y_{i,j}}]$, that reveals useful information in studying SO's *size* and *shape*.

In SDA, in fact, an important role is played by the *Description Potential* (DP) measure, which is the *hypervolume* of a SO, computed as the Cartesian product $V_1^i \times \cdots \times V_j^i \times \cdots \times V_p^i$ of the p descriptors associated to the Symbolic Object ω_i $(1 \leq i \leq N)$ (De Carvalho, 1992; De Carvalho, 1997). When intervals descriptors need to be normalized, the DP of the assertion a_i is formally defined as:

$$\pi(a_i) = \prod_{j=1}^{p} \mu(\bar{V}_j^i), \tag{6}$$

where $\bar{V}_j^i$ indicates the normalised range with respect to the domain O_j: $\bar{V}_j^i = V_j^i / O_j$. As the DP measure tends to zero if at least one $\bar{V}_j$ is close to zero, we prefer to use the following alternative measure, called *Linear Description Potential* (LDP). The LDP of a_i is defined by De Carvalho (1997) as:

$$\sigma(a_i) = \sum_{j=1}^{p} (\bar{V}_j^i) \tag{7}$$

Let $a_1, \ldots, a_i, \ldots, a_N$ a set of assertion SO's described by p interval descriptors, and $\mathbf{X}$ indicates the $(n \times p)$ matrix having as generic term $x_{ij} = \sqrt{\bar{V}_j^i}$.

The method proposed in this section performs a factorial decomposition of the quantity $\sum_{i=1}^{n} \sigma(a_i)$ allowing a graphical representation that - differently from V–PCA - shows the descriptors influence to the total LDP.

From a geometric point of view, the *range transformation* implies an affine translation of each object, so that the vertices $min = \{\underline{y_{i,1}}, \ldots, \underline{y_{i,j}}, \ldots, \underline{y_{i,p}}\}$ are reported in the origin (see Figure 2).

It is easy to see that, given the orthogonality relationship between couples of sides of each hypercube, the search for a suitable sub-space in order to visualize the *size* and *shape* of each SO, can be simply realized on the *n max* vertices as a PCA with respect to the origin (non centered PCA): $\overline{y_{i,1}}, \ldots, \overline{y_{i,j}}, \ldots, \overline{y_{i,p}}$. Notice that, in this way the *curse of dimensionality*, that affects the V-PCA as well as the SO-PCA, can be eliminated. The total number of points is reduced from $n2^p$ to n.

We refer to this approach as *Principal Component Analysis on the Range Transformation* (RT-PCA) of interval variables. As in this PCA the ver-

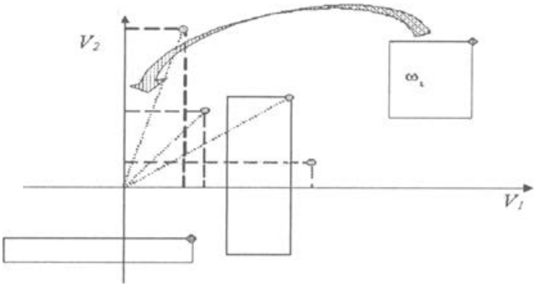

Figure 2: SO's transposition to the origin

tices cohesion is automatically respected, hypercubes and their images can be easily visualized by projecting on the factorial plans the other vertices as supplementary points.

The PCA performed on the matrix $\mathbf{X}$ decomposes the LDP criterion: $\left(\mathrm{tr}(\mathbf{XX}') = \mathrm{tr}(\mathbf{X}'\mathbf{X}) = \sum_i \sigma(a_i)\right)$, according to the following eigenanalysis:

$$\mathbf{X}'\mathbf{X}\mathbf{t}_m = \mu_m \mathbf{t}_m \qquad 1 \leq m \leq p \tag{8}$$

or equivalently:

$$\mathbf{XX}'\mathbf{u}_m = \mu_m \mathbf{u}_m \qquad 1 \leq m \leq p \tag{9}$$

being μ_m the generic eigenvalue ($\sum_m \mu_m = \sum_i \sigma(a_i)$) and $\mathbf{u}_m$ and $\mathbf{t}_m$ are the associated eigenvector in the space $\mathbb{R}^p$ and $\mathbb{R}^n$ respectively. Both analyses are defined under the usual ortho–normality constraints.

The SO ω_i representation in the optimal sub-space $m^* < p$ can be obtained by the matrix $\mathbf{\Phi}$, whose elements are the juxtaposition of the first m^* $(1 \leq m^* \leq p)$ axes:

$$\mathbf{\Phi} = [\, \phi_1 \quad \cdots \quad \phi_m \quad \cdots \quad \phi_{m^*} \,] \tag{10}$$

being $\phi_m = \mathbf{X}\mathbf{t}_m$.

In the RT–PCA the amount of ***information*** contribution associated to each axis is given by the corresponding eigenvalue μ_m. The ratio between the square coordinate and the eigenvector, both with respect to the axis m:

$$\mathrm{CTA}_{i,m} = \frac{\phi^2_{i,m}}{\mu_m} \tag{11}$$

measures the contribution of the SO ω_i to principal axis m. The relative contribution, indicating the quality of the representation, is measured by $\mathrm{CTR}_{i,m} = \sum_i \phi^2_{i,m} / \sum_j x^2_{i,j}$, with $m = \{1, \ldots, m^*\}$.

Having the matrix $\mathbf{X}$ has all positive entries, also the eigenvector $\mathbf{u}_1$ and the factor $\mathbf{t}_1$ have all positive values, so that the first axis is easily interpreted as SO's *size factor*, while the following ones discriminate SO's according to their *shape* features. Their interpretation depends on the contribution (squared coordinates) of the original variables to the axis: $\mathrm{CTA}_{j,m} = \mathbf{t}^2_{j,m}$.

It is worth noticing that in this PCA on the range transformation, SO's can be, more simply, represented by single points (*Max* vertices coordinates). Therefore, close points refer to SO's whose LDP's are mainly influenced by the same variables. In other words, the closeness indicates that SO's are characterized by hypercubes having similar shape and size.

A direction for future research could be the definition of new SO's coding techniques that can directly cope with interval variables. In this framework, we propose to consider, for each SO, the following $(p+2)$characteristics: $[(V_1^i, \ldots, V_j^i, \ldots, V_p^i);\ \mathrm{L}^i;\ \cos\alpha_i]$, where L^i is the lenght of the main diagonal of the i-th hypercube (*i.e.* the euclidean distance between the vertices $\underline{y}_j^i$ and $\overline{y}_j^i$), α_i the angle that the segment L^i forms with the variable y_1. These *variables* include all the SO's characteristics but define a data structure easier to handle.

6 SO-PCA: a mixed strategy

In the above sections, we have considered two different approaches to the analysis of the interval data. The SO-PCA permits to evaluate SO's with respect to their positioning in the space of the recoded descriptors. On the other hand, RT-PCA processes interval data putting in evidence SO's *size* and *shape*. This section provides a mixed strategy combining the SO-PCA [§4] and the RT-PCA [§5] in order to improve SO's representation taking into account their differences in terms of scale and structural (*size* and *shape*) characteristics.
With this aim, we consider the following three–steps approach:

i) perform the RT-PCA of $\mathbf{X}$ in order to extract the principal axes that better represent the size and shape of SO's;

ii) transform $\mathbf{Z}$ as $\hat{\mathbf{Z}} = \mathbf{P}_A\mathbf{Z}$, that permits to take into account the SO vertices cohesion;

iii) perform a PCA on the projections of the rows of $\hat{\mathbf{Z}}$ on $\mathbf{\Phi}$ by $\mathbf{P}_{\mathbf{\Phi}}$ projection matrix (defining $\mathbf{P}_{\mathbf{\Phi}} = \mathbf{\Phi}\left(\mathbf{\Phi}'\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}'$), in order to stress the *size* and *shape* information as extracted by $\mathbf{X}$ in [§5].

The approach leads to the following eigenanalysis:

$$\tilde{\mathbf{Z}}'\mathbf{P}_{\Phi}\tilde{\mathbf{Z}} = \mathbf{Z}'\mathbf{A}\left(\mathbf{A}'\mathbf{A}\right)^{-1/2}\mathbf{P}_{\Phi}\left(\mathbf{A}'\mathbf{A}\right)^{-1/2}\mathbf{A}'\mathbf{Z}\mathbf{s}_m = \rho_m\mathbf{s}_m, \qquad (12)$$

where the diagonal matrix $(\mathbf{A}'\mathbf{A})^{-1}$ has been decomposed in the product $(\mathbf{A}'\mathbf{A})^{-1/2}(\mathbf{A}'\mathbf{A})^{-1/2}$ in order to ensure the symmetry and $\mathbf{s}_m$ and ρ_m are the generic eigenvector and the eigenvalue, respectively, calculated under the orthonormality constraints (1).

The Figure 3 shows an intuitive scheme of the procedure. The upper side represents vertices in the $\mathbb{R}_{\mathbf{Z}}$ space, the two ellipses refer to two different SO's. By means of $\mathbf{P}_{\Phi}$, these are projected into the space $\mathbb{R}_{\mathbf{t}}$, where the V–PCA is performed. The analysis final result is represented by the sub-space spanned by $\mathbf{s}_1$ and $\mathbf{s}_2$ where SO's are represented by means of MCAR's.

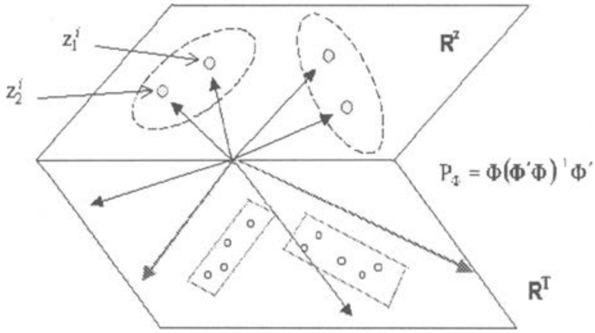

Figure 3: Vertices projection on the structure sub-space

The interpretation of results depends on the choice of $\mathbf{P}_{\Phi}$. In fact, the projection matrix allows putting in evidence different aspects of the considered set of SO's, being $\mathbf{P}_{\Phi}$ a weighting system with respect to the SO's *size* and *shape*. The generic diagonal term of $\mathbf{P}_{\Phi}$ is equal to the quantity in (11): $\phi_i(\phi_i'\phi_i)^{-1}\phi_i' = \sum_m \phi_{i,m}^2/\mu_m$.
Aiming at showing the SO's size we shall include the first principal component ϕ_1 in the definition of $\mathbf{P}_{\Phi}$; alternatively, the shape aspects could be put in evidence by dropping the first principal component.

7 Example on a real data-set

This section shows an example of the proposed methodology on a real data-set. We use the Oils' data set (Ichino, 1988) (reproduced in Table 1) largely used in SDA applications, whose characteristics are well–known to people working on the topic. The data set presents eight different classes of oils described by five variables, we only refer to the four quantitative interval variables: “Specific gravity”, “Freezing point”, “Iodine value” and “Saponification”.

Table 1: Oils Data Matrix

	Spec		*Freezing*		*Iodine*		*Saponifi-*		SO
	gravity		*point*		*value*		*cation*		LDP
Linseed	0.93	0.94	-27	-18	170	204	118	196	1.394
Perilla	0.93	0.94	-5	-4	192	208	188	197	.343
Cotton	0.92	0.92	-6	-1	99	113	189	198	.289
Sesame	0.92	0.93	-6	-4	104	116	187	193	.299
Camellia	0.92	0.92	-21	-15	80	82	189	193	.277
Olive	0.91	0.92	0	6	79	90	187	196	.390
Beef	0.86	0.87	30	38	40	48	190	199	.403
Hog	0.86	0.86	22	32	53	77	190	202	.452

In the following, we confine the presentation to the graphical results of the different approaches proposed in the paper.

In Figure 4, we show the results with respect to the first two axes (first plane), achieved by the V-PCA. Notice that the 88.4% of the total inertia is explained by the first two axes. In the figure, closeness among MCAR’s indicates SO’s mainly influenced by the same descriptors. We cannot give any interpretation of the similarity in size and shape among MCAR’s . As supplementary points we have also represented the variables, even if these should have been represented in the space $\mathbb{R}^N$. However, this simultaneous representation is very useful to the interpretation, provided that variables *must* be only evaluated with respect to their directions.

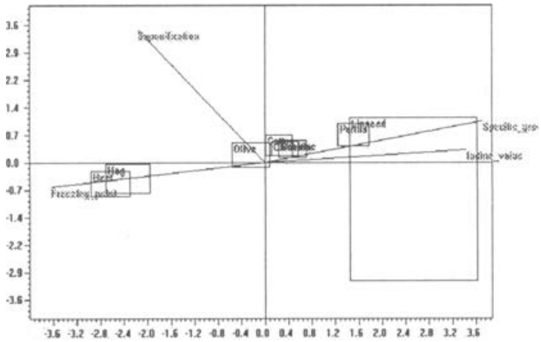

Figure 4: V-PCA: first two axes (88.4%)

SO’s associated to MCAR’s representations on the left side of the plane refer to oils mainly characterized by high values of ***Saponification*** and ***Freezing point***, whereas, on the opposite side we have representations of oils characterized by ***Specific gravity*** and ***Iodine*** high values.

In Figure 5, we represent the result obtained by the SO-PCA approach. The interpretation shall be done following the same rules of the V-PCA. In this case, the total inertia associated to the first plane is equal to `90.7%`. Comparing the two outputs, we note how the vertices cohesion constraints reduce the MCAR's and make the SO's representation more clear. In fact, the MCAR's overlapping is considerably reduced, in particular with respect to the SO's in the axes origin.

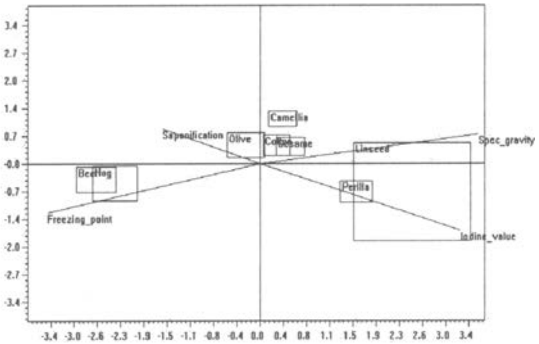

Figure 5: SO-PCA: first two axes (`90.7%`)

It is very important to stress the attention on the interpretation of the MCAR's dimensions. The ***Linseed*** MCAR is represented along the directions of ***Iodine value*** and ***Specific gravity*** variables meaning that high values have been registered for these two variables. Nevertheless, we cannot give any indication on which variables have mainly influenced its size and shape. This could be done by looking at the results of the RT-PCA analysis showed in the rest of the paragraph.

In the two above examples, we gave an interpretation with respect to MCAR's position but to their size and shape. Here, we start analyzing these two aspects by adopting the approach discussed in [§5].
At first we have to compute the variables domains (defined in [§5]) as the ***ranges*** between the lower `minimum` value and the higher `maximum` value for each variable. They resulted equal to: Specific gravity = `0.08`; Freezing point = `65.0`; Iodine value = `168.0`; Saponification = `84.0`.
Sums of normalized ranges define the SO's LDP as showed in the last right column of the table 1.

Looking at variable contributions to LDP, we have the following values:

Specific gravity = 0.775; Freezing point = 0.723; Iodine value = 0.720; Saponification = 1.619. Variables have almost the same contribution but Saponification that has a much larger value (1.619). A large amount of this contribution comes from the Linseed oil.
Figure 6 shows the descriptors positioning with respect to the first two factors. The total amount of LDP explained by these two factors is equal to 96.7%.

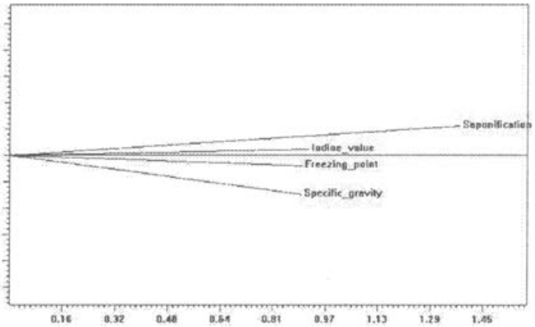

Figure 6: Descriptors contribution to SO's LDP (96.7%)

Looking at Figures 6 and 7 at the same time we note that: the biggest SO is *Linseed Oil* (Fig. 7) and its dimension is mainly determined by *Saponification* descriptors (Fig. 6).

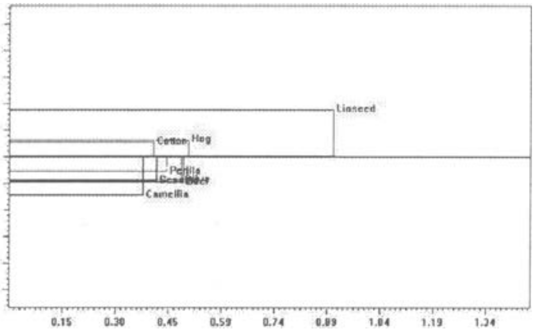

Figure 7: SO's size/shape graphical representation (96.7%)

Dealing with very few SO's, these statements can also be easily verified by looking at the original data table (Table 1). Moreover, *Linseed, Cotton* and *Hog* oils have their LDP mainly influenced by *Saponification* and *Iodine value*, whereas the rest of oils have been mainly influenced by *Freezing point* and *Specific gravity* descriptors.

The last approach proposed is the SO-PCA mixed strategy [§6], in which all the information is involved. Figure 8 shows the related representation, however, in this example, due to the small number of variables, we cannot appreciate very relevant changes in the representation with respect to the V-PCA. This approach, in fact, became useful in case of many variables influencing the SO's structure. Whereas, in the considered example, structures are mainly influenced only by *Saponification* variable.

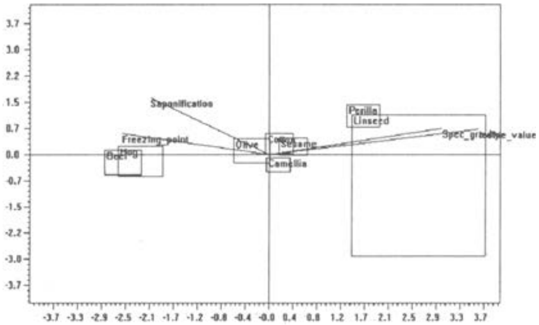

Figure 8: SO-PCA mixed strategy (95.4%)

8 Perspective and conclusion

In the present paper we proposed some new approaches to the PCA for interval data in the SDA definition. From this point of view, we represented the SO's on the factorial planes putting in evidence both structural aspects (*size* and *shape*) and their positioning in the factorial sub-space.

The proposed methodologies represent a first step in the SDA analysis and should be generalized to any typology of symbolic descriptor (*modal* and *multinomial variables*) and, at the same time, to take into account *rules* and *taxonomies* that could be defined in the symbolic data structure (Bock and Diday, 2000).
Another relevant aspect concerns the graphical representation that could be made with more efficient geometric tools than the MCAR. For example, we indicate *convex hulls* and *maximum inertia diagonals* (Verde and De Angelis, 1997).

As a final point, we indicate the axes interpretation that, in order to ensure the necessary consistency, should be made using the SDA language and tools. Also in this direction there are some contributions (Gettler–Summa, 1997) in literature that could be extended to the proposed factorial analyses.

Acknowledgements: This paper has been developed in the ISO-3D Esprit project at DMS of University of Naples "Federico II".
The output in [§7] has been realised with the collaboration of Dr. G. Meccariello (1999).

References

Bock, H. H. and Diday, E. (eds): 2000, *Analysis of Symbolic Data*, Springer. (in press).

Cazes, P., Chouakria, A., Diday, E. and 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., Diday, E. and Cazes, P.: 1998, An improved factorial representation of symbolic objects, *KESDA'98 27-28 April*, Luxembourg.

D'Ambra, L. and Lauro, C. N.: 1982, Analisi in componenti principali in rapporto a un sottospazio di riferimento, *Rivista di Statistica Applicata* **15**(1), 51–67.

De Carvalho, F. A. T.: 1992, *Méthodes Descriptives en Analyse de Données Symboliques*, Thése de doctorat., Université Paris Dauphine, Paris.

De Carvalho, F. A. T.: 1997, Clustering of constrained symbolic objects based on dissimilarity functions, *Indo–French Workshop on Symbolic Data Analysis and its Applications*, University of Paris IX.

Diday, E.: 1987, Introduction á l'approche symbolique en analyse des données, *Journées Symbolique-Numerique*, Université Paris Dauphine.

Diday, E.: 1996, Une introduction á l'analyse des données symboliques, *SFC*, Vannes, France.

Escofier, B. and Pagés, J.: 1988, *Analyse factorielles multiples*, Dunod, Paris.

Gettler–Summa, M.: 1997, Symbolic marking application on car accident reports, *Applied Stocastic Models and Data Analysis*, Vol. Invited and Specialised Session Papers, ASMDA, Anacapri, pp. 299–306.

Ichino, M.: 1988, General metrics for mixed features - the cartesian space theory for pattern recognition, *International Conference on Systems, Man and Cybernetics.*

Lauro, C. and Palumbo, F.: 1998, New approaches to principal components analysis to interval data, *International Seminar on New Techniques & Technologies for Statistics, NTTS'98, 4/6 nov. 1998*, Sorrento, Italy.

Lebart, L., Morineau, A. and Piron, M.: 1995, *Statistique exploratorie multidimensionelle*, Dunod, Paris.

Meccariello, G.: 1999, Analisi in componenti principali per dati ad intervallo, Tesi di Laurea in Statistica Universitá di Napoli "Federico II".

Verde, R. and De Angelis, P.: 1997, Symbolic objects recognition on a factorial plan, *NGUS'97*, Bilbao Spain.