```
PCA.centers.SDA, PCA.mrpca.SDA, PCA.spaghetti.SDA,
PCA.spca.SDA, PCA.vertices.SDA (symbolicDA)
```

Principal component analysis for symbolic interval-valued data

PCA (Principal Component Analysis) aims to visualize, synthesize and compare units onto factorial spaces with minimum loss of information (for example minimum distortion of the distance between original data). Whereas units are represented by points, it is sufficient to just take care of their position in space. On the other hand, symbolic objects described by interval valued variables, represented as boxes in a multidimensional space, needs to be visualized, synthesized and compared onto the factorial spaces, taking care not only of their location but also by their size and shape [Diday and Noirhomme-Fraiture (2008), p. 279-280].

There are six main algorithms for PCA for symbolic interval-valued data:

a) vertices,

b) centers,

c) spca,

d) midpoints and radii,

e) spaghetti,

f) interval-valued PCA.

The vertices algorithm [Cazes et. al. (1997); Diday and Noirhomme-Fraiture (2008)] looks for the best representation of the symbolic objects on a factorial plan, by optimizing the total variance criterion of the all vertices of the hyperrectangles. This algorithm matrix of hyperrectangles are represented by their vertices, so instead of interval-valued data we get matrix M_H defined as follows:

$$\mathbf{M}_H = \begin{bmatrix}
\underline{v_{1,1}} & \underline{v_{1,2}} & \cdots & \underline{v_{1,\,m-1}} & \underline{\underline{v_{1,\,m}}} \\
\underline{v_{1,1}} & \underline{v_{1,2}} & \cdots & \underline{\underline{v_{1,\,m-1}}} & v_{1,\,m} \\
\underline{v_{1,1}} & \underline{v_{1,2}} & \cdots & \underline{v_{1,\,m-1}} & \underline{\underline{v_{1,\,m}}} \\
\underline{v_{1,1}} & \underline{v_{1,2}} & \cdots & v_{1,\,m-1} & v_{1,\,m} \\
\underline{\cdots} & \underline{\cdots} & \cdots & \underline{\cdots} & \underline{\cdots} \\
v_{1,1} & v_{1,2} & \cdots & v_{1,\,m-1} & v_{1,\,m} \\
\underline{v_{2,1}} & \underline{v_{2,2}} & \cdots & \underline{v_{2,\,m-1}} & \underline{\underline{v_{2,\,m}}} \\
\underline{v_{2,1}} & \underline{v_{2,2}} & \cdots & \underline{\underline{v_{2,\,m-1}}} & v_{2,\,m} \\
\underline{v_{2,1}} & \underline{v_{2,2}} & \cdots & \underline{v_{2,\,m-1}} & \underline{\underline{v_{2,\,m}}} \\
\underline{v_{2,1}} & \underline{v_{2,2}} & \cdots & v_{2,\,m-1} & v_{2,\,m} \\
\underline{\cdots} & \underline{\cdots} & \cdots & \underline{\cdots} & \underline{\cdots} \\
v_{2,1} & v_{2,2} & \cdots & v_{2,\,m-1} & v_{2,\,m} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\underline{v_{n,1}} & \underline{v_{n,2}} & \cdots & \underline{v_{n,\,m-1}} & \underline{\underline{v_{n,\,m}}} \\
\underline{v_{n,1}} & \underline{v_{n,2}} & \cdots & \underline{\underline{v_{n,\,m-1}}} & v_{n,\,m} \\
\underline{v_{n,1}} & \underline{v_{n,2}} & \cdots & \underline{v_{n,\,m-1}} & \underline{\underline{v_{n,\,m}}} \\
\underline{v_{n,1}} & \underline{v_{n,2}} & \cdots & v_{n,\,m-1} & v_{n,\,m} \\
\underline{\cdots} & \underline{\cdots} & \cdots & \underline{\cdots} & \underline{\cdots} \\
v_{n,1} & v_{n,2} & \cdots & v_{n,\,m-1} & v_{n,\,m}
\end{bmatrix}$$

where: v_{ij} – realization of *j*-th symbolic variable for *i*-th symbolic object; symbols: $\_$ and $^{-}$ mean lower and upper bound of interval.

Each symbolic object is represented by 2^m rows of matrix $\mathbf{M}_H$. This matrix has the size of $n \cdot 2^m \times m.$

The vertices algorithm (V-PCA) decomposes the correlation matrix associated to the vertices of hyperrectangles. In the original proposal V-PCA consists of performing a classic PCA on the standardized **Z** matrix. In this way, vertices are elements of the subspace $\mathbf{R}^m$, whereas the *m* quantitative descriptors are elements of $\mathbf{R}^N$. V-PCA looks for a suitable subspace to represent symbolic objects and represent variables.

As in the classic PCA optimal subspace is spanned by the axes maximizing the sum of squares of projected vertices coordinates $\boldsymbol{\psi}_m = \mathbf{Z}\mathbf{v}_m$:

$$\boldsymbol{\psi}'_m \boldsymbol{\psi}_m = \mathbf{v}'_m \mathbf{Z}' \mathbf{Z} \mathbf{v}_m$$

with $\mathbf{v}'_m \mathbf{v}_m = 0$ for $m' \neq m$ and $\mathbf{v}'_m \mathbf{v}_m = 1$ for $m' = m.$

The characteristic equation of V-PCA in $\mathbf{R}^N$ is given by:

$$\frac{1}{N}\mathbf{Z}'\mathbf{Z}\mathbf{v}_m = \lambda_m \mathbf{v}_m \quad 1 \le m \le p$$

where: $\mathbf{v}_m$ and λ_m are the generic eigenvector and the generic eigenvalue respectively associated with the matrix $\frac{1}{N}\mathbf{Z}'\mathbf{Z}$.

The principal axes interpretation is performer with reference to the variables z_j's having maximal contribution.

Centers PCA (CPCA) decomposes the correlation matrix of the centers of intervals and then project on the factorial axis the vertices as supplementary points. Let *p* be the interval variables and the *n* the number of observations. The analysis is performed on the matrix coding the centers that has *n* rows and *p* columns.

The characteristic equation of CPCA in $\mathbf{R}^n$ is given by:

$$\frac{1}{n}\tilde{\mathbf{M}}'\tilde{\mathbf{M}}\mathbf{v}_m = \lambda_m \mathbf{v}_m \quad 1 \le m \le p$$

where: $\mathbf{v}_m$ and λ_m are the generic eigenvector and the generic eigenvalue respectively, associated to the matrix $\tilde{\mathbf{M}}'\tilde{\mathbf{M}}$, $\tilde{\mathbf{M}}$ is the matrix **M** standardized by the standard deviation of the centers.

SPCA was proposed by Lauro and Palumbo (2000) to overcome drawbacks of vertices PCA for symbolic data. SPCA technique allows the maximization of the inter-object variance being, in this sense, similar to CPCA. The SPCA method is based on the maximization of the between symbolic objects variance matrix:

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

where: **A** is a Boolean matrix (*N* x *n*) describing the belonging of the *N* vertices to the *n* symbolic object representations.

The axes of maximum inertia are obtained in the space $\mathbf{R}^N$ as solutions to the following characteristic equation:

$$\frac{1}{N}\left[\mathbf{Z}'\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{Z}\right]\tilde{\mathbf{v}}_m = \frac{1}{N}\mathbf{Z}'\mathbf{P}_{\mathbf{A}}\mathbf{Z} = \tilde{\lambda}_m \tilde{\mathbf{v}}_m$$

where: $\tilde{\lambda}_m$ are the eigenvalues of the matrix $\mathbf{Z}'\mathbf{P}_{\mathbf{A}}\mathbf{Z}$ and $\tilde{\mathbf{v}}_m$ are the associated eigenvectors (for $m = 1, \ldots, M$, with *M* the maximum number of non-null eigenvalues), defined under the orthonormality constraints.

Midpoints and radii PCA (MRPCA) is an interval algebra based method. In this approach PCA is generalized to interval variables, described by midpoints and radii, by maximizing the projections of distances between any interval and the mean interval. The generic term of the data matrix is given by the standardized interval:

$$\tilde{d}_{ij} = \frac{d\left([y]_{Iij},[\mu]_{Ij}\right)}{\sqrt{\mathrm{var}(j)\cdot n}} = \frac{\left|m_{ij}-\overline{m}_j\right| + \left|r_{ij}+\overline{r}_j\right|}{\sqrt{\mathrm{var}(j)\cdot n}}$$

The full matrix of the correlation is defined as follows:

$$\left[\left(\mathbf{M}'\Sigma^{-1}\mathbf{M}\right)+\left(\mathbf{R}'\Sigma^{-1}\mathbf{R}\right)+\left(\mathbf{M}'\Sigma^{-1}\mathbf{R}+\mathbf{R}'\Sigma^{-1}\mathbf{M}\right)\right]$$

The distance formulation allows PCs calculation according three steps:

1. A PCA of the midpoints matrix:

$$\mathbf{M\Sigma^{-1}u}_k^m = \lambda_k^m \mathbf{u}_k^m$$

where: $\mathbf{u}_k^m$ and λ_k^m $(1 \le k \le p)$ are defined under the usual orthonormality constraints.

2. A PCA of the radii matrix:

$$\mathbf{R\Sigma^{-1}u}_k^r = \lambda_k^r \mathbf{u}_k^r$$

with the same orthonormality constraints on $\mathbf{u}_k^r$ and λ_k^r $(1 \le k \le p)$.

3. An interval reconstruction step by the projection of suitably rotated ranges into the midpoints space spanned by the eigenvectors associated with the PCa of the midpoints matrix.

A radii rotation matrix is obtained by maximizing the Tucker congruence coefficient between midpoints and radii:

$$f(\mathbf{T}) = \sum_l \frac{\mathbf{t}_l'\mathbf{M}'\mathbf{r}_l}{\left(\mathbf{t}_l'\mathbf{M}'\mathbf{Mt}\right)^{1/2}\left(\mathbf{r}_l'\mathbf{r}_l\right)^{1/2}}$$

Under the constraint $\mathbf{T}'\mathbf{T} = \mathbf{I}$.

Spaghetti PCA, which is a hybrid approach, was proposed by Irpino (2006). The *Spaghetti* PCA decomposes the correlation matrix of the main diagonal of the hyperrectangles representing multidimensional interval data.

The multidimensional boxes are represented by their main diagonals (*md*$_i$) that are considered as segments of uniform points described in the following way:

$$md_i(t) = \left\{\begin{array}{c}\cdots \\ \underline{y}_{ij} + t\left(\overline{y}_{ij} - \underline{y}_{ij}\right) \\ \cdots\end{array}\right. = \left\{\begin{array}{c}\cdots \\ m_{ij} + r_{ij}(2t-1) \\ \cdots\end{array}\right. \quad j = 1,\ldots,p,\ 0 \le t \le 1$$

The technique decomposes the correlation matrix of the descriptors of such segments according to the following formulas, for the computation of the mean, standard deviation, covariance and correlation:

a) mean

$$\mu_{ij} = \frac{1}{n} \sum_{i} \int_{0}^{1} \left[m_{ij} + r_{ij} (2t-1) \right] dt = \frac{1}{n} \sum_{i} m_{ij}$$

b) standard deviation

$$\sigma_j = \sqrt{{}^{1}\!/\!_{n} \sum_{i} \int_{0}^{1} \left[m_{ij} + r_{ij} (2t-1) \right]^2 dt - \mu_j^2} = \sqrt{{}^{1}\!/\!_{n} \sum_{i} \left(c_{ij}^2 + {}^{1}\!/\!_{3}\, r_{ij}^2 \right) - \mu_j^2}$$

c) covariance

$$Cov\left(Y_j ; Y_k\right) = {}^{1}\!/\!_{n} \sum_{i} \int_{0}^{1} \left[m_{ij} + r_{ij} (2t-1) \right] \left[m_{ik} + r_{ik} (2t-1) \right] dt - \mu_j \mu_k =$$
$${}^{1}\!/\!_{n} \sum_{i} \left(m_{ij} m_{ik} + {}^{1}\!/\!_{3}\, r_{ij} r_{ik} \right) - \mu_j \mu_k$$

d) correlation

$$Corr\left(Y_j ; Y_k\right) = \frac{Cov\left(Y_j ; Y_k\right)}{\sigma_j \sigma_k}$$

Considering the standardized main diagonals as:

$$\tilde{y}_i(t) = \frac{md_i(t) - \boldsymbol{\mu}}{\boldsymbol{\sigma}} = \begin{cases} \cdots \\ \dfrac{\left[\underline{y}_{ij} + t\left(\overline{y}_{ij} - \underline{y}_{ij} \right) \right] - \mu_j}{\sigma_j} \\ \cdots \end{cases} = \begin{cases} \cdots \\ \dfrac{\left[m_{ij} + r_{ij} (2t-1) \right] - \mu_j}{\sigma_j} \\ \cdots \end{cases} \quad j = 1, \ldots, p\,, \ 0 \le t \le p$$

The method decomposes the correlation matrix of the main diagonals:

$$\frac{1}{n} \tilde{\mathbf{Y}}' \tilde{\mathbf{Y}} \mathbf{u}_m = \lambda_m \mathbf{u}_m \quad 1 \le m \le p$$

under the usually orthonormality constraint on the $\mathbf{u}_m$'s.

Interval-valued PCA (IPCA) was proposed by Gioia and Lauro (2006). It is known that the classical PCA on a real matrix **Y**, in the space spanned by the variables, solves the problem of determining *m*≤*p* axes $\mathbf{u}_\alpha$, $\alpha = 1, \ldots, m$ such that the sum of the squared projections of the point units on $\mathbf{u}_\alpha$ is maximum:

$$\mathbf{u}_\alpha \mathbf{Y}' \mathbf{Y} \mathbf{u}_\alpha = \max \ 1 \le \alpha \le m$$

This optimization problem may be reduced to the eigenvalues problem:

$$\mathbf{Y}' \mathbf{Y} \mathbf{u}_\alpha = \lambda \mathbf{u}_\alpha \quad 1 \le \alpha \le m$$

When the data are of interval type ($\mathbf{Y}^I$) this equation can becomes an *interval eigenvalue problem* of the form:

$$\left(\mathbf{Y}^{\mathbf{I}} \right) \mathbf{Y}^{\mathbf{I}} \mathbf{u}_\alpha^I = \boldsymbol{\lambda}^{\mathbf{I}} \mathbf{u}_\alpha^I$$

which has the following interval solutions:

$$\left[\lambda_{\alpha}(\mathbf{Z}) : \mathbf{Z} \in \left(\mathbf{Y^I}\right)^{\mathbf{T}} \mathbf{Y^I}\right],\ \left[u_{\alpha}(\mathbf{Z}) : \mathbf{Z} \in \left(\mathbf{Y^I}\right)^{\mathbf{T}} \mathbf{Y^I}\right]\ \alpha = 1, \ldots, p$$

The interval eigenequation decomposed by IPCA is defined as follows:

$$\mathbf{\Gamma^I u}_{\alpha}^{I} = \lambda^{I} \mathbf{u}_{\alpha}^{I}$$

where $\mathbf{\Gamma^I}$ is the correlation interval matrix between $\mathbf{Y}_i^{\mathbf{I}}, \mathbf{Y}_j^{\mathbf{I}}$.

References:

1. Cazes P., Chouakira A., Diday E., Schektman Y. (1997), *Extension de l'analyse en composantes principales à des données de type intervalle*. "Revue de Satistique Apliquée" XIV(3), p. 5-24.

2. Diday E., Noirhomme-Fraiture M. (Eds.) (2008), *Symbolic data analysis and the SODAS software*, John Wiley & Sons, Chichester.

3. Gioia F., Lauro N.C. (2006), *Principal component analysis on interval data*, "Computational Statistics", 21(2), p. 343-363.

4. Lauro N.C., Palumbo F. (2000), *Principal component analysis of interval data: a symbolic data analysis approach*, "Computational Statistics", Vol. 15, No. 1, p. 73-87.

5. Irpino A. (2006), "*Spaghetti" PCA analysis: An extension of principal component to time dependent data*, "Pattern Recognition Letters" Vol. 27, Issue 5, p. 504-513.