Dr. Oldemar Rodríguez R.
Professor University of Costa Rica
www.oldemarrodriguez.com

RSDA: An R Package for Symbolic Data Analysis

This package aims to execute some models on Symbolic Data Analysis. Symbolic Data Analysis was propose by the professor E. DIDAY in 1987 in his paper *"Introduction à l approche symbolique en Analyse des Données"*. Premiére Journées Symbolique-Numérique. Université Paris IX Dauphine. Décembre 1987. A very good reference to symbolic data analysis can be found in "From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis" of L. Billard and E. Diday that is the journal American Statistical Association Journal of the American Statistical Association June 2003, Vol. 98.

The main purpose of Symbolic Data Analysis is to substitute a set of rows (cases) in a data table for an concept (second order statistical unit). For example, all of the transactions performed by one person (or any object) for a single transaction that summarizes all the original ones (Symbolic-Object) so that millions of transactions could be summarized in only one that keeps the customary behavior of the person. This is achieved thanks to the fact that the new transaction will have in its fields, not only numbers (like current transactions), but can also have objects such as intervals, histograms, or rules. This representation of an object as a conjunction of properties fits within a data analytic framework concerning symbolic data and symbolic objects, which has proven useful in dealing with big databases.

More and more we are required to conduct a statistical analysis on *BIG DATA*. In fact, these data sets may be so large that it is necessary to preprocess the data by classifying or reorganizing it into classifications or classes where the number of classes is much smaller than the number of individuals in the original data set. Then, the resulting data set, after the preprocessing will most likely contain symbolic data rather than classical data values. We refer to symbolic data when instead of having a specific, or single value for an observed variable, an observed value, for a given variable, say y_j, may be multi-valued (for example, $y_j = \{16, 21, 35, 40\}$ or $y_j = \{$yellow, white, pink$\}$, it may be interval-valued (e.g., $y_j = [10, 20]$), or it may be modal-valued (e.g., $y_j = \{1$ with probability $0.1, 0$ with probability $0.9\}$). For example, if one is dealing with fuzzy data where the observed variable(s) are represented by an interval of values, then this data would be symbolic data. For any of such symbolic data, it may be inappropriate to use existing data analytic techniques developed for single-valued data. The formal mathematical definition of a symbolic object can be found in in the paper *"From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis"* of L. Billard and E. Diday that is the journal American Statistical Association Journal of the American Statistical Association June 2003, Vol. 98.

In Figure 1 we show an example, note that there is a table with transactions, and each one shows the code of an individual, his/her age, occupation, salary and province. In this case, it was decided to create one Symbolic Object (concept) per province. Then, for instance, the age in San José showed by the interval $[36, 39]$ are the corresponding minimum and maximum ages in San José (shown in the original table). This means that any age in San José shall be included between 36 and 39 years of age, while the occupation in San José is represented by an histogram that says that half of the people in the original table are lawyers, and the other half are doctors.

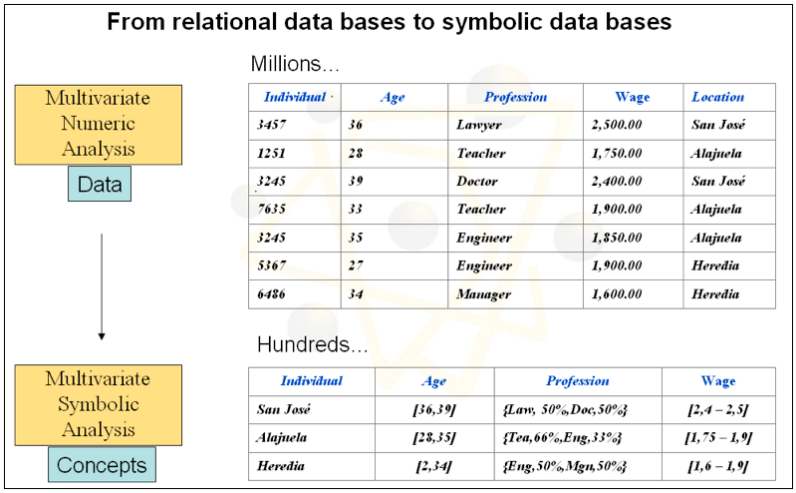


FIGURE 1: Transforming a relational data base into a symbolic table

There are two tool to generate symbolic data from relational data bases, DB2SO (Data Base To Symbolic Objects) in SODAS software (Symbolic Object Data Analysis System) and VSOG (Visual Symbolic Object Generator). For the package RSDA we suppose that the symbolic data table have been created for some of this programs or the table was created directly in a CSV file using expert criterium for some problem.

For example, we will use Ichino's data (oils and fats) that we present in the 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".

	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.

In RSDA Package this table will be read from a CSV file with the following format.

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

If we have a label \$C it means that follows a continuous variable, \$I means an interval variable, \$H means a histogram variables and \$S means set variable. In the first row each labels should be follow of a name to variable and to the case of histogram variables types the names of the modalities (categories). In data rows for continuous variables we have just one value, for interval variables we have the minimum and the maximum of the interval, for histogram variables we have the number of modalities and then the probability of each modality and for set variables we have the cardinality of the set and next the elements of the set.

Once the data have been read with RSDA package we can run a series of analyzes for these kind tables, for example, interval Principal Component Analysis. In this method 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).

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

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

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

Then in the centers method we apply the standard principal components analysis to the matrix (2). To apply this standard principal components we use the matrix of variance–covariance $V^c = (X^c)^t X^c$ and then to compute the interval principal component $[\underline{y_{ik}}, \overline{y_{ik}}]$ using the equations (3) and (4).

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

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

where $\overline{X^c_j}$ 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. If we project the hypercube variable we have that the minimum and the maximum value are given by the equation (5) and (6) respectively (see Rodriguez O. (2012). The Duality Problem in Interval Principal Components Analysis. The 3rd Workshop in Symbolic Data Analysis, Madrid).

$$\underline{r_{ij}} = \sum_{k=1, v_{kj}<0}^{m} \overline{z}^c_{ki} v_{kj} + \sum_{k=1, v_{kj}>0}^{m} \underline{z}^c_{ki} v_{kj}, \tag{5}$$

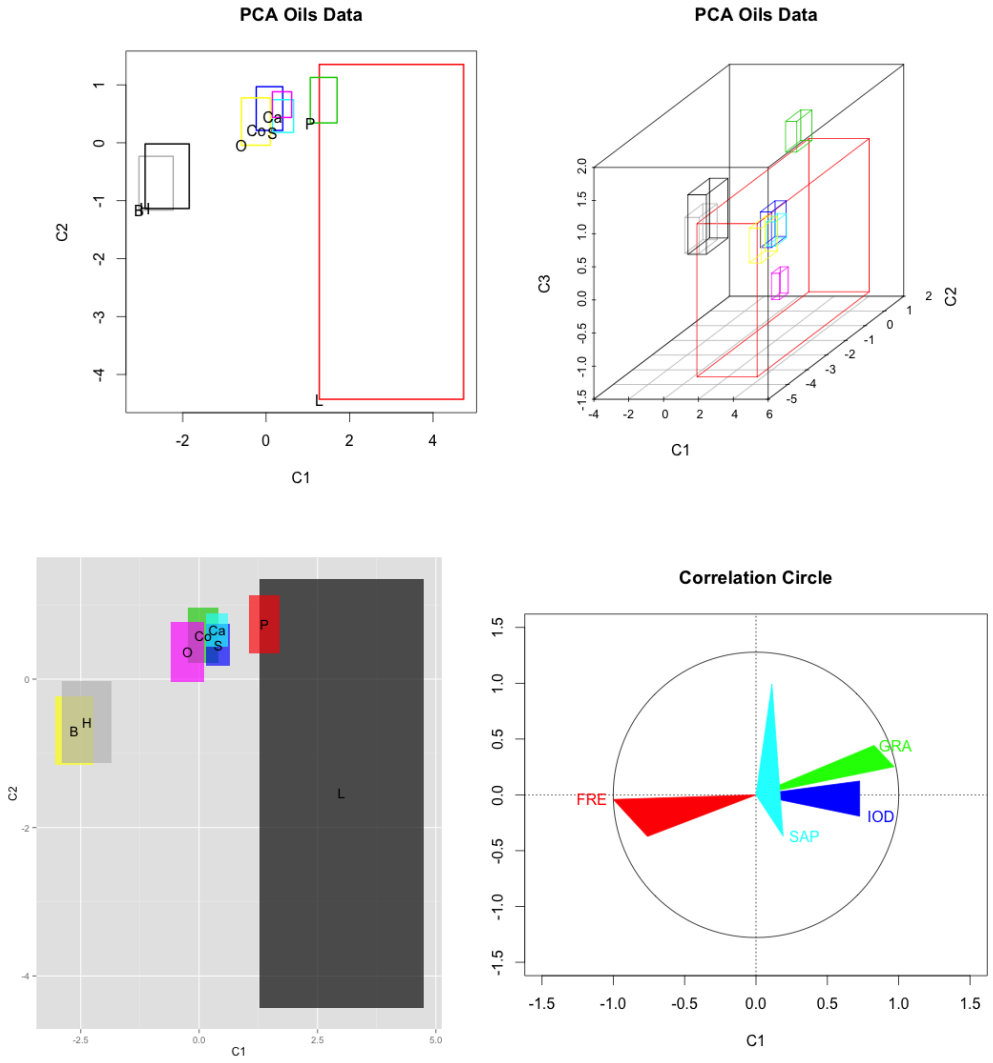
$$\overline{r_{ij}} = \sum_{k=1, v_{kj}<0}^{m} \underline{z}^c_{ki} v_{kj} + \sum_{k=1, v_{kj}>0}^{m} \overline{z}^c_{ki} v_{kj}. \tag{6}$$

In RSDA Package interval PCA can be execute as follows:

```
data(oils)
res<-sym.interval.pca(oils,'centers')
sym.scatterplot(sym.var(res$Sym.Components,1),sym.var(res$Sym.Components,2),
    labels=TRUE,col='red',main='PCA Oils Data')
```

```
sym.scatterplot3d(sym.var(res$Sym.Components,1),sym.var(res$Sym.Components,2),
    sym.var(res$Sym.Components,3),color='blue',main='PCA Oils Data')
sym.scatterplot.ggplot(sym.var(res$Sym.Components,1),
    sym.var(res$Sym.Components,2),labels=TRUE)
sym.circle.plot(res$Sym.Prin.Correlations)
```

Then we get, among other outputs, the following graphs in which we can see the clusters and correlations, but also we can see the variability inside the data represented as rectangles, cubes and slices:



Note: As professor Balasubramanian Narasimhan wrote on my project proposal, the project was "pretty ambitious", I have implemented all the function that I had proposed, but in some of them, due to time constraints, not all algorithms were programmed. For example, to symbolic multidimensional scaling and to symbolic kmeans we still have to program some algorithms to the case of interval distances. However, the project has been very good for me because now I have the basis for further work in this package, package that is much needed by the scientific community in symbolic data analysis.