Analysis of m sets of symbolic interval variables.

Sun Makosso-Kallyth*,

*CRCHUQ, Plate forme de recherche clinique
10 rue D'Espinay Hôpital SFA, G1L 3L5 Québec, QC, Canada,
sun.makosso-kallyth@crchuq.ulaval.ca.

Abstract. This work presents a new approach to analyze a series of m $n \times p$ tables $X^{(1)}, \ldots, X^{(m)}$ of symbolic interval variables. In this new approach, we firstly define a space of intervals with laws of composition $\oplus$, $\otimes_1$, $\times$. This allows extending this reasoning to the matrices of intervals. Then, we define a $n \times p$ compromise matrix $\overline{X} = \left(\overline{X_{ij}}\right)_{i=1,\ldots,n;\ j=1,\ldots,p,}$ of type intervals, a measure of covariance between interval variables, a new measure of correlation η between interval variables and the product operator $\otimes_2$ between a matrix $n \times p$ of intervals and one p vector u. This way, we achieve a symbolic PCA of compromise. To express the variability of tables $X^{(1)}, \ldots, X^{(m)}$, they are projected on the principal axes of PCA of intervals of compromise. For the interpretation of factorial map, a new measure of correlation η will be used.

1 Introduction

Expert 1				
	Banana	Coffee	Tea	Cocoa
Region 1	[0.9,3.1]	[5.8,6.2]	[6.5,7.5]	[2.1,2.7]
Region 2	[4.8,5.2]	[2.9,3.1]	[2,2]	[3.1,3.4]
Region 3	[5.4,6.6]	[0.8,1.2]	[0.95,1.05]	[2.1,2.3]
Region 4	[6.9,7.9]	[0.75,1.25]	[1.85,2.15]	[1.4,2.0]
Region 5	[1.9,2.6]	[5,5]	[3.6,4.4]	[6.1,6.2]
Region 6	[2.8,3.2]	[3.8,4.9]	[3.6,4.4]	[7.1,8.0]

TAB. 1 – *Example of symbolic interval variable.*

Analysis of m sets of symbolic interval variables.

Advances in information technology lead to more complex data. This complexity is reflected by data in the form of images, histogram or tables structured into blocks of symbolic interval variables.... The formalism of symbolic objects (see Bock and Diday (2000), Billard and Diday (2006), Diday and Noirhomme-Fraiture (2008)) captures this kind of complex data. The purpose of this work is to propose an analysis in principal axes of a series of m tables containing symbolic interval variables. Table 1 is a an example in which an expert assesses some products from six regions in a 0-8 scale. Authors such as Cazes et al. (1997), Palumbo and Lauro (2003), Ichino (2007), Ichino (2008) and Diday et al. (2011) proposed symbolic Principal Component Analysis (PCA) of interval-type variables.This work is an attempt to extend the issue raised by Cazes et al. (1997), Palumbo and Lauro (2003), Ichino (2007), Ichino (2008) and Diday et al. (2011) to multiple tables. In analysis of structured data in blocks, the STATIS (Structuration A Trois Indice de la Statistique) method of L'Hermier des Plantes (1976) (see Lavit (1988)), the Generalized Principal Component Analysis of Casin (2001), the Multiple Factor Analysis (MFA) of Escofier and Pagès (1998), the CCAW (Component Common Analysis and Specific Weight) of Hanafi et al. (2006) used to analyze several tables of ordinary size $n \times p$ (n and p are respectivly the number of individuals and variables). In symbolic data analysis, Corales and Rodriguez (2011) extended the STATIS method for interval-type variables. The method of Corales and Rodriguez (2011) and STATIS have the advantage of dealing with tables that do not necessarily concerne the same number of observations and variables. But the main drawback with these two methods is the lack of direct connection between the representations of individuals and variables with the representation of m tables. Our method overcomes this drawback. We define a space of intervals that we supply with the laws of composition. These laws of composition allow to define a compromise $\overline{X} = \left(\overline{X}_{ij}\right)$ and to carry out after a symbolic interval PCA based on $\min$ and $\max$. The proposed approach generalizes that of Ichino (2007). Indeed, for $m = 1$, $\overline{X} = X^{(1)}$ and PCA of the compromise is equivalent to interval PCA of Ichino (2007). Furthermore, we propose a new measure of correlation η. Finally, to simultaneously analyze the tables $X^{(1)}, \ldots, X^{(m)}$, their variables are projected as supplementary variables onto principal plans of the interval PCA of compromise.

2 Definition of operators.

Let $n,\ p,\ m$ be the number of individuals, variables and tables describing the same number of individuals and the same number of variables. Let $X^{(k)} = \left\{X_{ij}^{(k)}\right\}_{i=1,\ldots,n\ j=1,\ldots,p\ k=1,\ldots,m}$, the array of symbolic interval variables. For $i = 1, \ldots, n,\ j = 1, \ldots, p,\ X_{ij}^{(k)}$ is of form $X_{ij}^{(k)} = \left[a_{ij}^{(k)}, b_{ij}^{(k)}\right]$. For a description of the intervals, we propose a method similar to the algebra of Moore (1966).

2.1 Space of interval.

Let $\mathcal{I}$ be the space of intervals. The algebra proposed by Moore (1966) for a description of the intervals used the centers and radiums. The proposed approach instead uses the $\min$ and $\max$ intervals. Let $\left[a_{ij}^{(k_1)}, b_{ij}^{(k_1)}\right]$ and $\left[a_{ij}^{(k_2)\prime}, b_{ij}^{(k_2)\prime}\right]$ belong to $\mathcal{I}$. We supply in the space $\mathcal{I}$

laws of composition $\oplus$, $\times$ $\otimes_1$, so that :

- addition: $[a_{ij}, b_{ij}] + [a'_{ij}, b'_{ij}] = \left[\min\left\{a_{ij} + a'_{ij}; b_{ij} + b'_{ij}\right\}, \max\left\{a_{ij} + a'_{ij}; b_{ij} + b'_{ij}\right\}\right]$. We define the subtraction of the same way.

- multiplication in a scalar λ: $\lambda \times [a_{ij}, b_{ij}] = [\min\{\lambda a_{ij}; \lambda b_{ij}\}, \max\{\lambda a_{ij}; \lambda b_{ij}\}]$.
- the product of two intervals:

$$\left[a_{ij}^{(k_1)}, b_{ij}^{(k_1)}\right] \otimes_1 \left[a_{ij}^{(k_2)\prime}, b_{ij}^{(k_2)\prime}\right] = \begin{pmatrix} a_{ij}^{(k_1)} \\ \\ b_{ij}^{(k_1)} \end{pmatrix} \times \begin{pmatrix} a_{ij}^{(k_2)\prime} \\ \\ b_{ij}^{(k_2)\prime} \end{pmatrix}$$

$$\left[a_{ij}^{(k_1)}, b_{ij}^{(k_1)}\right] \otimes_1 \left[a_{ij}^{(k_2)\prime}, b_{ij}^{(k_2)\prime}\right] = a_{ij}^{(k_1)} a_{ij}^{(k_2)\prime} + b_{ij}^{(k_1)} b_{ij}^{(k_2)\prime}. \quad (1)$$

- The norm of an interval $\|, \|_{\mathcal{I}}$:

$$\| \left[a_{ij}^{(k_1)}, b_{ij}^{(k_1)}\right] \|_{\mathcal{I}} = \sqrt{\left[a_{ij}^{(k_1)}, b_{ij}^{(k_1)}\right] \otimes_1 \left[a_{ij}^{(k_1)}, b_{ij}^{(k_1)}\right]} = \sqrt{\left(a_{ij}^{(k_1)}\right)^2 + \left(b_{ij}^{(k_1)}\right)^2}$$

The space of intervals $(\mathcal{I}, \oplus, \otimes_1, \|, \|)$ is an euclidian vectorial space.

2.2 Extension to the matrix of intervals.

Previous operators are extending to the matrices of intervals. Let be $\mathcal{I}_n$ a space of n-uplets of intervals. We supply in $\mathcal{I}_n$ laws of composition $\oplus$, $\times$, $\otimes_1$, $\otimes_2$. Then, for $X_{\cdot j_1}^{(k_1)} = \left(\left[a_{ij_1}^{(k_1)}, b_{ij_1}^{(k_1)}\right]\right)_{i=1,\dots,n;}$ and $X_{\cdot j_2}^{(k_2)} = \left(\left[a_{ij_2}^{(k_2)}, b_{ij_2}^{(k_2)}\right]\right)_{i=1,\dots,n} \in \mathcal{I}_n$:

- the product is

$$X_{\cdot j_1}^{(k_1)} \otimes_1 X_{\cdot j_2}^{(k_2)} = \sum_{i=1}^{n} a_{ij_1}^{(k_1)} a_{ij_2}^{(k_2)} + \sum_{i=1}^{n} b_{ij_1}^{(k_1)} b_{ij_2}^{(k_2)} \quad (2)$$

- if $X_{\cdot j_1}^{(k_1)} \in \mathcal{I}_n$, the mean of interval $g_{X_{\cdot j_1}^{(k_1)}}$ is:

$$g_{X_{\cdot j_1}^{(k_1)}} = \left[\underline{g_{X_{\cdot j_1}^{(k_1)}}}, \overline{g_{X_{\cdot j_1}^{(k_1)}}}\right]$$

with

$$\underline{g_{X_{\cdot j_1}^{(k_1)}}} = \frac{1}{n}\sum_{i=1}^{n} a_{ij_1}^{(k)}; \ \overline{g_{X_{\cdot j_1}^{(k_1)}}} = \frac{1}{n}\sum_{i=1}^{n} b_{ij_1}^{(k)}.$$

For example if $X_{\cdot 1} = \begin{pmatrix} [2,4] \\ [-5,6] \\ [9,8] \end{pmatrix}$ then $g_{X_{\cdot 1}} = \left[\frac{1}{3}(2-5+9), \frac{1}{3}(4+6+9)\right] = [2,6]$. The centered vector $X_{\cdot j}^{*(k_1)}$ is : $X_{\cdot j}^{*(k_1)} = \left(X_{ij}^{(k_1)} - g_{X_{\cdot j}^{(k_1)}}\right)_{i=1,\dots,n}$. We have for example:

$$X_{\cdot 1}^{*} = \begin{pmatrix} [2,4]-[2,6] \\ [-5,6]-[2,6] \\ [9,8]-[2,6] \end{pmatrix} = \begin{pmatrix} [0,-2] \\ [-7,0] \\ [7,2] \end{pmatrix}.$$

Analysis of m sets of symbolic interval variables.

The interval $g_{X^*_{\cdot 1}} = [(0-7+7)/3,\ (-2+0+2)/3] = [0,0]$. In general, if $X^{*(k_1)}_{\cdot j_1}$ is a vector centered, $\Longrightarrow g_{X^{*(k_1)}_{\cdot j_1}} = [0,0]$.

Consequently, the covariance of two symbolic interval variables $X^{(k_1)}_{\cdot j_1}$ and $X^{(k_2)}_{\cdot j_2}$ is:

$$Cov\left(X^{(k_1)}_{\cdot j_1}, X^{(k_2)}_{\cdot j_2}\right) = \left(X^{(k_1)}_{ij_1} - g_{X^{(k_1)}_{\cdot j_1}}\right) \otimes_1 \left(X^{(k_2)}_{ij_2} - g_{X^{(k_2)}_{\cdot j_2}}\right) \; i = 1, \ldots, n. \tag{3}$$

In the case of two centered vectors $X^{*(k_1 j_1)}_{\cdot}$ and $X^{*(k_2)}_{\cdot J_2}$, the relationship (3) becomes :

$$Cov\left(X^{(k_1)}_{\cdot j_1}, X^{(k_2)}_{\cdot j_2}\right) = \left(X^{(k_1)}_{ij_1} - [0,0]\right) \otimes_1 \left(X^{(k_2)}_{ij_2} - [0,0]\right) = \left(X^{(k_1)}_{ij_1}\right) \otimes_1 \left(X^{(k_2)}_{ij_2}\right)$$

$$Cov\left(X^{(k_1)}_{\cdot j_1}, X^{(k_2)}_{\cdot j_2}\right) = \begin{pmatrix} a^{(k_1)}_{1j_1} \\ \vdots \\ a^{(k_1)}_{nj_1} \\ \\ b^{(k_1)}_{1j_1} \\ \vdots \\ b^{(k_1)}_{nj_1} \end{pmatrix} \times \begin{pmatrix} a^{(k_2)}_{1j_2} \\ \vdots \\ a^{(k_2)}_{nj_2} \\ \\ b^{(k_2)}_{1j_2} \\ \vdots \\ b^{(k_2)}_{nj_2} \end{pmatrix} = \sum_{i=1}^{n} a^{(k_1)}_{ij_1} a^{(k_2)}_{ij_2} + \sum_{i=1}^{n} b^{(k_1)}_{ij_1} b^{(k_2)}_{ij_2}. \tag{4}$$

The covariance of intervals obtained from equations (3) and (4) induces a variance-covariance matrix W.

– the norm $\|, \|_{\mathcal{I}_n}$ of the interval variable $X^{(k_1)}_{\cdot j_1}$ is defined as :

$$\|X^{(k_1)}_{\cdot j_1}\|_{\mathcal{I}_n} = \sqrt{X^{(k_1)}_{\cdot j_1} \otimes_1 X^{(k_1)}_{\cdot j_1}} = \sqrt{\sum_{i=1}^{n} \left(\left(a^{(k_1)}_{ij}\right)^2 + \left(b^{(k_1)}_{ij}\right)^2 \right)}. \tag{5}$$

The space of interval variables $(\mathcal{I}_n, \oplus, \otimes_1, \|, \|_{\mathcal{I}_n})$ is an euclidian vectorial space.

2.3 Correlation of two symbolic interval variables.

2.3.1 Correlation of two symbolic interval variables proposed by Billard.

Billard (2007) and Billard (2008), developped two ways of correlating interval variables. Billard's correlation is based on the centers of interval $X^{(k_1)}_{\cdot j} = \left(\left[a^{(k_1)}_{ij}, b^{(k)}_{ij}\right]\right)_{i=1,\ldots,n}$. These centers are called

$$\mu^{(k)}_{ij} = \frac{1}{2} \sum_{i=1}^{n} \left(a^{(k)}_{ij} + b^{(k)}_{ij}\right). \tag{6}$$

The average of centers $\mu_{.j}^{(k)}$ and the variance of interval $X_{.j}^{(k)}$ defined by Bertrand and Goupil (2000) are :

$$\mu_{.j}^{(k)} = \frac{1}{2n}\sum_{i=1}^{n}\left(a_{ij}^{(k)} + b_{ij}^{(k)}\right) \tag{7}$$

$$(\sigma_{.j}^{(k)})^2 = \frac{1}{3n}\sum_{i=1}^{n}\left(\left(b_{ij}^{(k)}\right)^2 + b_{ij}^{(k)}a_{ij}^{(k)} + \left(a_{ij}^{(k)}\right)^2\right) - \frac{1}{4n^2}\left[\sum_{i=1}^{n}\left(b_{ij}^{(k)} + a_{ij}^{(k)}\right)\right]^2. \tag{8}$$

Compared to (4), the covariance γ proposed by Billard (2007) is

$$\gamma\left(X_{\cdot j_1}^{(k_1)}, X_{\cdot j_2}^{(k_2)}\right) = \frac{1}{3n}\sum_{i=1}^{n}\mathsf{G}_{j_1}\mathsf{G}_{j_2}\left[\mathsf{Q}_{j_1}\mathsf{Q}_{j_2}\right]^{1/2} \tag{9}$$

where

$$\mathsf{G}_j = \begin{cases} -1, & \textit{if } \mu_{ij}^{(k)} \leq \mu_{.j} \\ 1 & \textit{else} \end{cases} \tag{10}$$

and

$$\mathsf{Q}_j = (a_{ij}^{(k)} - \mu_{.j})^2 + (a_{ij}^{(k)} - \mu_{.j})(b_{ij}^{(k)} - \mu_{.j}) + (b_{ij}^{(k)} - \mu_{.j})^2. \tag{11}$$

The measure of correlation ρ proposed by Billard (2007) is given by the following formula :

$$\rho(X_{.j_1}^{(k_1)}, X_{\cdot j_2}^{(k_2)}) = \frac{\gamma\left(X_{\cdot j_1}^{(k_1)}, X_{\cdot j_2}^{(k_2)}\right)}{\sigma_{\cdot j_1}^{(k_1)}\sigma_{\cdot j_2}^{(k_2)}}. \tag{12}$$

The formula proposed by Billard (2008) is similar with the formula proposed by Billard (2007).

2.3.2 New correlation of two symbolic interval variables.

The formalism used to describe intervals allows us to suggest a new measurement of correlation of interval variables η. The correlation η we propose uses the $\min$ and the $\max$. Let $L_{X_{\cdot j_1}^{(k_1)}}, L_{X.j_2}^{(k_2)}$ two n-vectors obtained from respectively lower and higher values of $X_{\cdot j_1}^{(k_1)}, X_{\cdot j_2}^{(k_2)}$. The correlation η betwen $X_{\cdot j_1}^{(k_1)}, X_{\cdot j_2}^{(k_2)}$ is:

$$\eta\left(X_{\cdot j_1}^{(k_1)}, X_{\cdot j_2}^{*(k_2)}\right) = \frac{Cov\left(X_{\cdot j_1}^{(k_1)}, X_{\cdot j_2}^{(k_2)}\right)}{\sqrt{Cov\left(X_{\cdot j_1}^{(k_1)}, X_{\cdot j_1}^{(k_1)}\right)}\sqrt{Cov\left(X_{\cdot j_2}^{(k_2)}, X_{\cdot j_2}^{(k_2)}\right)}}$$

$$\eta\left(X_{\cdot j_1}^{(k_1)}, X_{\cdot j_2}^{*(k_2)}\right) = r\left(L_{X_{\cdot j_1}^{(k_1)}}, L_{X_{\cdot j_2}^{(k_2)}}\right) = r\left(\begin{bmatrix} a_{1j_1} \\ \vdots \\ a_{nj_1} \\ b_{1j_1} \\ \vdots \\ b_{nj_1} \end{bmatrix}, \begin{bmatrix} a_{1j_2} \\ \vdots \\ a_{nj_2} \\ b_{1j_2} \\ \vdots \\ b_{nj_2} \end{bmatrix}\right) \tag{13}$$

where r is the Pearson's correlation.

For two variables $X_1 = \begin{pmatrix} [11, 11.2] \\ [10.3, 11.3] \\ [11, 11.2] \\ [11.5, 12] \\ [11.1, 11.6] \\ [12, 12.1] \end{pmatrix}$, $X_2 = \begin{pmatrix} [67, 68] \\ [62, 64] \\ [57, 59] \\ [53, 55] \\ [55, 57] \\ [[50, 52]] \end{pmatrix}$, the correlation is

$$\eta\left(X_1, X_2\right) = r \left\langle \begin{pmatrix} 11 \\ 10.3 \\ \vdots \\ 12 \\ 11.2 \\ 11.3 \\ \vdots \\ 12.1 \end{pmatrix}, \begin{pmatrix} 67 \\ 62 \\ \vdots \\ 50 \\ 68 \\ 64 \\ \vdots \\ 52 \end{pmatrix} \right\rangle = -0.625.$$

The correlation proposed by Billard (2007) gives $\rho\left(X_1, X_2\right) = -0.786$

2.4 The operator $\otimes_2$.

Let $Z^{(k)} = X^{(k)} \otimes_2 u_\alpha$ be the product of a matrix $n \times p$ of interval $X^{(k)}$ and an ordinary p vector u_α. $Z^{(k)}$ is defined as follows

$$Z^{(k)} = \left(Z_{ij}^{(k)}\right) = \left(\left[\underline{z_{i\alpha}^{(k)}}, \overline{z_{i\alpha}^{(k)}}\right]\right)_{i=1,\ldots,n;\ \alpha=1,\ldots,p}$$

with

$$\underline{z_{i\alpha}^{(k)}} = \min\left\{\begin{bmatrix} a_{i1}^{(k)} & \ldots & a_{ip}^{(k)} \end{bmatrix} \times \begin{bmatrix} u_{1\alpha} \\ \vdots \\ u_{p\alpha} \end{bmatrix}; \begin{bmatrix} b_{i1}^{(k)} & \ldots & b_{ip}^{(k)} \end{bmatrix} \times \begin{bmatrix} u_{1\alpha} \\ \vdots \\ u_{p\alpha} \end{bmatrix}\right\}$$

$$\underline{z_{i\alpha}^{(k)}} = \min\left\{\sum_{l=1}^{p} a_{il}^{(k)} u_{l\alpha}; \sum_{l=1}^{p} b_{il}^{(k)} u_{l\alpha}\right\}, \quad (14)$$

$$\overline{z_{i\alpha}^{(k)}} = \max\left\{\begin{bmatrix} a_{i1}^{(k)} & \ldots & a_{ip}^{(k)} \end{bmatrix} \times \begin{bmatrix} u_{1\alpha} \\ \vdots \\ u_{p\alpha} \end{bmatrix}; \begin{bmatrix} b_{i1}^{(k)} & \ldots & b_{ip}^{(k)} \end{bmatrix} \times \begin{bmatrix} u_{1\alpha} \\ \vdots \\ u_{p\alpha} \end{bmatrix}\right\}$$

$$\overline{z_{i\alpha}^{(k)}} = \max\left\{\sum_{l=1}^{p} a_{il}^{(k)} u_{l\alpha}; \sum_{l=1}^{p} b_{il}^{(k)} u_{l\alpha}\right\} \quad (15)$$

with $a_{il}^{(k)},\ u_{l\alpha},\ b_{il}^{(k)} \in \mathbb{R}$.

2.5 Definition of the Compromise $X^{(1)}$ and the principal axes.

2.5.1 Compromise.

The compromise $\overline{X}$ is an $n \times p$ table which is representative from the m tables $X^{(1)}, \ldots, X^{(m)}$. The definition of the compromise is given by the following formula :

$$\overline{X} = \frac{1}{m}\left[X^{(1)} \oplus \ldots \oplus X^{(m)}\right] = \frac{1}{m}\sum_{k=1}^{m} X^{(k)}. \tag{16}$$

2.5.2 The principal axes.

Let be $\overline{Z}_\alpha = \overline{X} \otimes_2 u_\alpha, \ _{\alpha=1,\ldots,p}$ where $u_\alpha, \ _{\alpha=1,\ldots,p}$ are the solutions of the following optimisation problem

$$\mathcal{P} : \begin{cases} \max \ Var\left(\overline{Z}_\alpha\right) \\ u_\alpha^t u_\alpha = 1 \\ u_\alpha^t u_\beta = 0, \ _{\alpha \neq \beta}, \end{cases}$$

Let's suppose that $\overline{X}$ is centered and the weight of all n observations is given by $\frac{1}{n}$, then

$$Var\left(\overline{Z}_\alpha\right) = Var\left(\overline{X} \otimes_2 u_\alpha\right) = \frac{u_\alpha^t \otimes_2 \overline{X}^t \otimes_1 \overline{X} \otimes_2 u_\alpha}{n} = u_\alpha^t \mathsf{W} u_\alpha, \tag{17}$$

where $\mathsf{W} = \frac{\overline{X}^t \otimes_1 \overline{X}}{n}$ is the variance-covariance matrix. Therefore,

$$\mathcal{P} : \begin{cases} \max \ u_\alpha^t \mathsf{W} u_\alpha \\ u_\alpha^t u_\alpha = 1 \end{cases}$$

$$\mathcal{P} \Leftrightarrow \max\left\{u_\alpha^t \mathsf{W} u_\alpha - \lambda_\alpha(u_\alpha^t u_\alpha - 1)\right\}$$

where λ_α are the Lagrange's multiplicators. After differenciation in u_α, the solution we obtain is

$$\mathsf{W}u_\alpha - \lambda_\alpha u_\alpha = 0 \Leftrightarrow \mathsf{W}u_\alpha = \lambda_\alpha u_\alpha.$$

$\lambda_\alpha, \ u_\alpha, \ \overline{Z}_\alpha$ represents respectively the α^{th} eigenvalue, eigenvector (or principal axe) and the generalized principal components of the compromise.

2.5.3 Visualizations of interval variables and individuals

After the determination of the generalized principal component $\overline{Z}_\alpha$, we suggest the computation of the correlation $\eta\left(\overline{X}_j, \overline{Z}_\alpha\right)$ between the interval variables of the compromise and the generalized components. The plot of variables is given by the correlation $\eta\left(\overline{X}_j, \overline{Z}_\alpha\right)$ for $j = 1, \ldots, p; \ \alpha = 1, \ldots, p$.

For the representation of individuals, we suggest the projection of each dataset $X^{(1)} \ldots X^{(m)}$ as supplementary element on the principal axes of the compromise via the matrix products $X^{(1)} \otimes_2 u_\alpha, \ldots, X^{(m)} \otimes_2 u_\alpha$.

Analysis of m sets of symbolic interval variables.

2.5.4 The steps of the proposed approach.

Six steps are involved in our approach :

1. Calculate the compromise $\overline{X} = \frac{1}{m}\left[X^{(1)} \oplus \ldots \oplus X^{(m)}\right]$
2. Calculate [1] $\mathsf{W} = \frac{1}{n}\overline{X}^t \otimes_1 \overline{X}$ (or Λ correlation matrix) covariance matrix [2]of the compromise through the relation (4).
3. Calculate the eigenvalues λ_α and eigenvectors u_α of W i.e resolve $\mathsf{W}u_\alpha = \lambda_\alpha u_\alpha$.
4. Calculate the principal components of interval type of compromise $Z_\alpha^{(k)}$ from the product $Z^{(k)} = X^{(k)} \otimes_2 u_\alpha$.
5. Calculate the correlation $\eta\left(\overline{X}_j, \overline{Z}_\alpha\right)$ between the principal components of the compromise and the variables of the compromise (which are both symbolic interval variables).
6. Projection from the $X^{(1)} \ldots X^{(m)}$ in supplementary axes of the compromise via the matrix products $X^{(1)} \otimes_2 u_\alpha, \ldots, X^{(m)} \otimes_2 u_\alpha$.

3 Comparaison with PCA of Ichino (2007) and INTERSTATIS.

3.1 Comparaison with PCA of Ichino (2007).

If we suppose that $m = 1$ table, $\overline{X} = X^{(1)}$ and the method proposed is equivalent to the symbolic PCA interval of Ichino (2007) based on nested recovering. Indeed, Ichino (2007) defined for each observation of the table $X^{(1)}$:

$$\mathsf{M}_{i.} = \begin{pmatrix} a_{i1}^{(1)} \ldots a_{ip}^{(1)} \\ b_{i1}^{(1)} \ldots b_{ip}^{(1)} \end{pmatrix} \tag{18}$$

and

$$\mathsf{M} = \begin{pmatrix} \mathsf{M}_{1.} \\ \mathsf{M}_{2.} \\ \vdots \\ \mathsf{M}_{n.} \end{pmatrix}. \tag{19}$$

$\mathsf{M}_{i.}$ and M are $2 \times p$ and $2n \times p$ usual matrices. Interval PCA of Ichino (2007) consists in computing an ordinary PCA on M. In this regards, Ichino (2007) suggests the usage of the correlation of spearman or the correlation of kendall. Each individual has 2 values. The interval principal components are given by the $\min$ and the $\max$ of these 2 values.

3.2 Comparaison with INTERSTATIS.

STATIS is a 3-index information tool. It is used to analyze multiple data tables. STATIS is capable of analyzing m tables refering to the same number of variables or the same number of individuals (STATIS DUAL). In order to analyze the individuals, the variable and the tables, STATIS defines respectively the *compromise*, the *intrastructure*, and the *interstructure*.

[1] if variables are centered.

[2] $\overline{X}^t$ is the transposed matrix.

INTERSTATIS proposed by Rodriguez (2011) extends the STATIS method to the case of symbolic interval variables. The extension proposed by Rodriguez uses the arithmetic proposed by Moore (1966). If in the methodology that we proposed, we use the product $\otimes_1$ defined in the relationship (1), Rodriguez (2011) used instead the following product $\otimes_3$ proposed by Moore (1966) :

$$\left[a_{ij}^{(k_1)}, b_{ij}^{(k_1)}\right] \otimes_3 \left[a_{ij}^{(k_2)\prime}, b_{ij}^{(k_2)\prime}\right] = [c, d] \tag{20}$$

where

$$\begin{cases} c = \min\left(a_{ij}^{(k_1)} a_{ij}^{(k_2)\prime}, a_{ij}^{(k_1)} b_{ij}^{(k_2)\prime}, b_{ij}^{(k_1)} a_{ij}^{(k_2)\prime}, b_{ij}^{(k_1)} b_{ij}^{(k_2)\prime}\right); \\ d = \max\left(a_{ij}^{(k_1)} a_{ij}^{(k_2)\prime}, a_{ij}^{(k_1)} b_{ij}^{(k_2)\prime}, b_{ij}^{(k_1)} a_{ij}^{(k_2)\prime}, b_{ij}^{(k_1)} b_{ij}^{(k_2)\prime}\right). \end{cases}$$

The results of the covariance and the correlation induced from $\otimes_3$ between two interval variables are intervals. By contrast, the covariance and the correlation induced by $\otimes_1$ are scalars.

3.2.1 Interstructure.

We supposed that tables are centered. Given the individuals in the m datasets, the interstructure compare their spatial distribution by using the $n \times n$ matrices of intervals $\mathsf{V}_k = X^{(k)} \otimes_3 \left(X^{(k)}\right)^t$. In this regards, the interstructure defines the following metric

$$\Delta_{k_1,k_2} = \langle \mathsf{V}_{k_1}, \mathsf{V}_{k_2} \rangle = \mathit{Trace}\left(\mathsf{V}_{k_1} \otimes_3 \mathsf{V}_{k_2}\right) \tag{21}$$

or

$$\Delta^*_{k_1,k_2} = \frac{\mathit{Trace}\left(\mathsf{V}_{k_1} \otimes_3 \mathsf{V}_{k_2}\right)}{\sqrt{\mathit{Trace}\left(\mathsf{V}_{k_1} \otimes_3 \mathsf{V}_{k_1}\right)}\sqrt{\mathit{Trace}\left(\mathsf{V}_{k_2} \otimes_3 \mathsf{V}_{k_2}\right)}}. \tag{22}$$

Δ_{k_1,k_2} and $\Delta^*_{k_1,k_2}$ are intervals because of definition of the product $\otimes_3$. If $a_{ij}^{(k)} = b_{ij}^{(k)}$, $\forall\, i = 1, \ldots, n;\ j = 1, \ldots, p;\ k = 1, \ldots, m$, the relationship (21) and (22) are called coefficient RV of Escoufier (1973).

The center PCA (CPCA) of Cazes et al. (1997) is performed on

$$\Delta = \left(\Delta_{k_1,k_2}\right)_{k_1=1,\ldots,m¿1=1,\ldots,m.} \cdot$$

The interstructure is the correlation circle from the interval CPCA of Δ. Each point represents each data table. Close points mean similar individual configuration.

Compared to our methodology, we can assimilate each table $X^{(k)}$ by its center of gravity $g^{(k)} = \left(\left[\underline{g_{.1}}^{(k)}, \overline{g_{.1}}^{(k)}\right], \ldots, \left[\underline{g_{.m}}^{(k)}, \overline{g_{.m}}^{(k)}\right]\right)$ and we suggest the projecting of the minima and the maxima as supplementary elements on $u_1, \ldots, u_p$ via the product $g^{(k)} \otimes_2 u_\alpha$.

3.2.2 The compromise and the intrastructure.

Regarding the analysis of the individuals, INTERSTATIS proposed by Corales and Rodriguez (2011) defines $\mathsf{V}_k = X^{(k)} \otimes_3 \left(X^{(k)}\right)^t$. The compromise proposed by INTERSTATIS is

$$\mathsf{V} = \sum_{k=1}^{m} \beta_k \mathsf{V}_k \tag{23}$$

where $\beta_k, k = 1, \ldots, m$ are the weights of the V_k tables. Each β_k is a function of the first eigenvector and the first eigenvalue of the CPCA of the interstructure. The study of individuals of INTERSTATIS is given by the CPCA of Cazes et al. (1997) of the compromise. The correlations circle induced from the previous CPCA gives the intrastructure.

Compared to our methodology, the compromise V is a $n \times n$ matrix of intervals. But the compromise we proposed $\overline{X}$ in the relationship (16) is a $n \times p$ matrix of intervals. In our approach, the implicit weighting that we use for each table is $\frac{1}{m}$. INTERSTATIS can also involve tables which have the same number of individuals but different number of variables. In this regards, INTERSTATIS is more general than our approach because our methodology requires the same number of individuals and variables. But in our methodology, we compute only one system of principal axes $u_1, \ldots, u_p$. The drawback of STATIS and INTERSTATIS is the usage of two different systems of axes for the interstructure in one hand, and the compromise and the intrastructure on the another hand.

4 Application.

The proposed approach is implemented to the serie of $m = 3$ tables (6×4). Information in these tables were assessed by 3 experts to evaluate various product from six regions.

Expert 2				
	Banana	Coffe	Tea	Cocoa
Region 1	[0,1]	[4.9,5.1]	[5.8,6.2]	[1.0,4.0]
Region 2	[4.1,4.2]	[3.5,4.1]	[5,5]	[1.8,2.1]
Region 3	[4.8,5.2]	[1.8,2.5]	[0.8,1.2]	[4.0,4.3]
Region 4	[6.9,7.8]	[2,2]	[1.7,2.1]	[2.0,4.8]
Region 5	[4.0,6.3]	[5,5]	[4.4,5.6]	[6.2,7.4]
Region 6	[3.0,3.1]	[4.75,5.25]	[4.6,5.4]	[6.5,7.7]

TAB. 2 – *Appreciation by the expert 2.*

Expert 3				
	Banana	Coffee	Tea	Cocoa
Region 1	[2.9,3.1]	[2.0,3.5]	[6.8,8]	[4.0,4.3]
Region 2	[3.9,4.1]	[3.8,4.2]	[3,3]	[3.0,3.4]
Region 3	[6.8,7.2]	[1,1]	[0.9,1.1]	[2.1,2.2]
Region 4	[0,1]	[1.8,2.2]	[3.8,4.5]	[4.0,4.7]
Region 5	[1.9,2.1]	[6,6]	[5.8,6.2]	[7.0,7.5]
Region 6	[0.9,1.9]	[6.9,7.8]	[5,7.3]	[7.5,7.9]

TAB. 3 – *Appreciation by the expert 3 .*

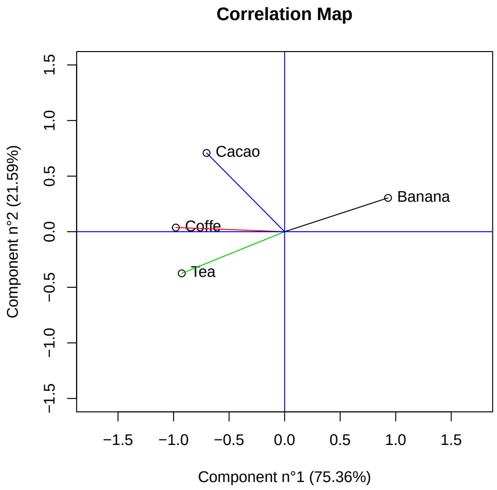

FIG. 1 – *Correlation map beetwen principal component and variables of compromise from η.*

Figure 1 represents the correlation towards η between the variables of compromise and their principal components. It allows to explain the positioning of individuals in Figure 2. Figure 2 shows the individual in rectangular form (because of their symbolic nature) on the principal axes of the compromise. The first two eigenvalues have a cumulative percentage of variability equal to $75.36 + 21.59\% = 96.95\%$. The experts 1 and 2 favourably rated Banana from Region4 and Region3. The expert 3 rated favourably Banana from Region3. Coffee and Tea from Region 1 are well rated by the expert 1 and the expert 2. Expert 3 appreciates coffee, Tea and Cocoa from region5 and region6. The above findings are reversed with the variable banana. The dispersions of the region 1 and the region 2 from the expert 2 are more pronounced.

Analysis of m sets of symbolic interval variables.

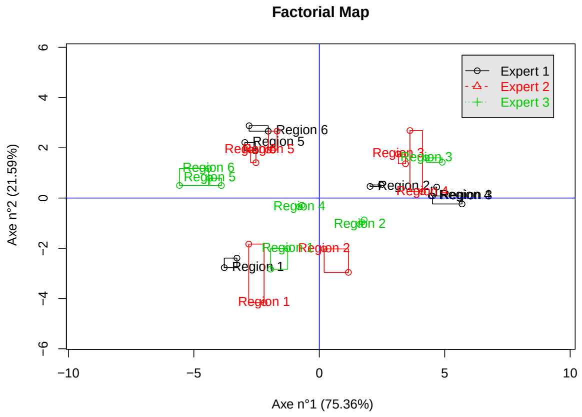

FIG. 2 – *Individual map.*

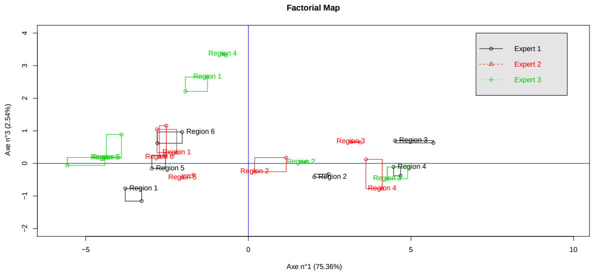

FIG. 3 – *Individual map (Axes 1-3).*

In the figures 3 and 4 we also write down that for the expert 3 the Region4 has a weak value of Coffee and Banana.

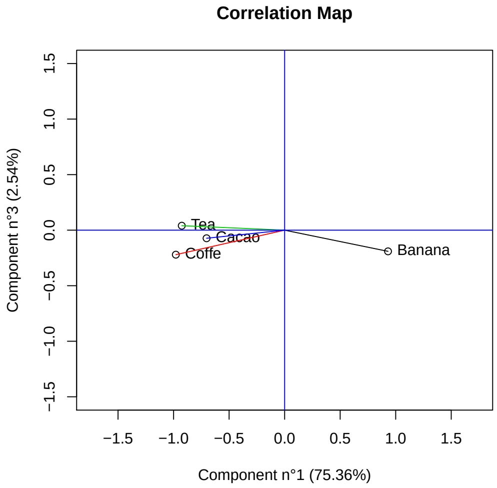

FIG. 4 – *Correlation map (from Axes 1-3).*

5 Conclusion.

The proposed approach allows the simultaneous analysis in principal axes of a set from m tables for the same number of individuals and the same number of symbolic interval variables. For the $m = 1$ table, the method proposed is equivalent to the symbolic PCA interval of Ichino (2007). The use of a compromise and supplementary elements is a common practice in data analysis. Indeed, Benzécri (1973), Escofier and Pagès (1998), Cazes (2002), Makosso-Kallyth and Diday (2010), Makosso-Kallyth and Diday (2012) have used this approach. The uniqueness of the new method is the usage of $m \geq 1$ series of $n \times p$ interval tables. It also proposes a new correlation η. The suggested approach establishes a direct connection between the graphs of individuals and those variables. However, if we assume that the arrays $X^{(1)}, \ldots, X^{(m)}$ are observed in m different time and the time dependence of the structure of tables is complex, the approach may be less robust.

Analysis of m sets of symbolic interval variables.

References

Benzécri, J. P. (1973). *L'Analyse des données. Tome 1 : La taxonomie. Tome 2 : L'analyse des correspondances*. Paris : Dunod.

Bertrand, P. and F. Goupil (2000). Descriptive statistics for symbolic data. In H.-H. Bock and E. Diday (Eds.), *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data.*, pp. 103–124. Springer.

Billard, L. (2007). Dependencies and variation components of symbolic interval-valued data. In P. Brito, P. Bertrand, G. Cucumel, and F. de Carvalho (Eds.), *Selected Contributions in Data Analysis and Classification*, pp. 3–12. Springer.

Billard, L. (2008). Some analyses of interval data. *Journal of Computing and Information Technology 16*(4), 225–233.

Billard, L. and E. Diday (2006). *Symbolic Data Analysis: conceptual statistics and data Mining*. Berlin: Wiley series in computational statistics.

Bock, H.-H. and E. Diday (2000). *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Informations from Complex Data*. Berlin: Springer.

Casin, P. (2001). A generalization of principal component analysis to *K* sets of variables. *Computational Statistics and Data Analysis 35*, 417–428.

Cazes, P. (2002). Analyse factorielle d'un tableau de lois de probabilité. *Revue de Statistique appliquée 50*(3), 5–24.

Cazes, P., A. Chouakria, E.Diday, and Y. Schektman (1997). Extension de l'analyse en composantes principales à des données intervalles. *Revue de Statistique appliquée 45*(3), 5–24.

Corales, D. and O. Rodriguez (2011). Interstatis: The statis method for interval valued data. In *Workshop in Symbolic Data Analysis*, Namur, Belgium.

Diday, E., L. Billard, and A. Douzal-Chouakria (2011). Principal component analysis for interval-valued observations. *Statistical Analysis and Data Mining 4*, 229–246.

Diday, E. and M. Noirhomme-Fraiture (2008). *Symbolic Data Analysis and the SODAS Software*. Chichester: Wiley Interscience.

Escofier, B. and J. Pagès (1998). *Analyses factorielles simples et multiples ; objectifs, méthodes et interprétation*. Paris : Dunod.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics 29*, 751–760.

Hanafi, M., G. Mazerolles, E. Dufour, and E. Qannari (2006). Common components and specific weight analysis and multiple co-inertia analysis applied to the coupling of several measurement techniques. *Journal of Chemometrics 20*, 172–183.

Ichino, M. (2007). Symbolic principal component analysis based on the nested covering. In *56th session of International Statistical Institute*, Lisboa, Portugal.

Ichino, M. (2008). Symbolic pca for histogram-valued data. In *IASC2008, Joint Meeting of 4th World Conference of the IASC and 6th Conference of the Asian Regional Section of the IASC on Computational Statistics & Data Analysis*, Yokohama, Japan.

Lavit, C. (1988). *Analyse conjointe de tableaux quantitatifs*. Paris: Masson.

L'Hermier des Plantes, H. (1976). *Structuration des Tableaux A Trois Indices de la Statistique*. Thèse de doctorat, Université de Montpellier.

Makosso-Kallyth, S. and E. Diday (2010). Analyse en axes principaux de variables symboliques de type histogramme. In *42èmes Journées de Statistique*, Marseille, France.

Makosso-Kallyth, S. and E. Diday (2012). Adaptation of interval pca to symbolic histogram variables. *Advances in Data Analysis and Classification 6*, 147–159.

Moore, R. (1966). *Interval Analysis*. Prentice-Hall.

Palumbo, F. and N. Lauro (2003). A pca for interval-valued data based on midpoints and radii. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, and J. E. Meulman (Eds.), *New Developments in Psychometrics*, pp. 641–648. Springer.