Symbolic PCA of compositional data.

Sun Makosso Kallyth & Edwin Diday

Université Paris Dauphine

COMPSTAT 2010-The 19th International Conference on Computational Statistics.

Context

- We have more and more complex data : sequential, textual, data structured in blocs, . . .
- Problem to analyze this data with usual tool of data analysis.
- Necessity to extend classical methods of data analysis to complex data.

Contribution of symbolic data analysis

- Study efficiently complex data via a superior level of generality (town− >regions, country− >continent, players− >team)
- Variables can be symbolic interval-valued, symbolic multi valued variable, histogram,. . . .
- Output of methodology proposed must have symbolic nature

Compositional Data and histogram data.

$x_1, \ldots, x_m$ m classical variables are compositional if $x_1, \ldots, x_m$ are non negative and

$$x_1 + \ldots + x_m = 1.$$

Symbolic histogram variables are an example of compositional variable. if :
n : number of observations ;
p : number of variables ;
m_j : number of bins of variables ;
$Y_j = (Y_{ij})_{i=1,\ldots,n,\ j=1,\ldots,p}$ is symbolic histogram variable if
$Y_{ij} = \{\xi_j,\ H_{ij}\}$; $\xi_j = \left(\xi_j^{(1)}, \ldots, \xi_j^{(m_j)}\right)$ are bins of variables.
H_{ij} are relatives frequency :

$$H_{ij}^{(1)} + \ldots + H_{ij}^{(m_j)} = 1.$$

Example of Symbolic histogram variable

TABLE: Example of Symbolic histogram variable

Region	GDP in k$ by hab.			Rate of mortality	
Bin	≤ 1 k$	$]1, 20]$ k$	> 20 k$	$\leq$ 0.10	$>$ 0.10
Afrique	0.340	0.660	0.000	0.245	0.755
Alena	0.000	0.333	0.667	1.000	0.000
AsieOrientale	0.067	0.801	0.133	1.000	0.000
Europe	0.000	0.322	0.677	0.742	0.258

$Y_{11} = \{\xi_1, H_{11}\}$ with $\xi_1 = \{]-\infty, 1],]1, 20],]20, +\infty[\}$; $H_{11} = (0.340; 0.660; 0.000)$

Parametric coding.

Let be $\mathcal{D}_j = (\alpha_j, \beta_j)$ domain of all possibles values of bins.
For the first variable (GDP), we have $\alpha_1 = 0,\ \beta_1 = +\infty$;
For the second variable (rate of mortality), we have : $\alpha_2 = 0,\ \beta_2 = 100$;
$\delta_j = \inf_{k_j=1,\dots,m_j} L_{k_j}$, where L_{k_j} is the length of interval $\xi_j^{(k_j)}$.

- If $\xi_j^{(k_j)} =]-\infty, a_j]$ then $\xi_j^{(k_j)} \longrightarrow \xi_j^{(k_j)} =]e, a_j]$ where
$$e = \begin{cases} \alpha_j & \text{if } a_j - \delta_j < \alpha_j \\ a_j - \delta_j & \text{else} \end{cases}.$$
- If $\xi_j^{(k_j)} =]b_j, +\infty[$, then $\xi_j^{(k_j)} \longrightarrow \xi_j^{(k_j)} =]b_j, f_j]$ with
$$f_j = \begin{cases} \beta_j & \text{si } b_j + \delta_j > \beta_j \\ b_j + \delta_j & \text{else} \end{cases}.$$

Parametric coding.

- In the example, $\xi_1^{(1)} =]-\infty, 1]$, $\xi_1^{(2)} =]1, 20]$, $L_2 = 20 - 1 = 19$, we replace $\xi_1^{(1)} \longrightarrow \xi_1^{'(1)} =]\max(1-19, 0), 1] =]0, 1]$ and $\xi_1^{(3)} \longrightarrow \xi_1^{'(3)} =]20, \min(20+19, +\infty)] =]20, \min(39, +\infty)] =]20, 39]$.
- If bins of variables don't have the same unit, we replace each interval $]a', b']$ by an adjusted interval $]a'/(b'-a'); b'/(b'-a')]$.
- Parametric coding assign to one bin a vector of scores $s_j = (s_j^{(1)}, \ldots, s_j^{(m_j)})$, where $s_j^{(kj)}$ is the center of adjusted interval for $k_j = 1, \ldots, m_j$.

Non parametric coding.

Non parametric Coding use as score of bins the rank associated to their bins. In the table of example of histogram data, scores of bins of classes will be $s_j^{(1)} = 1, s_j^{(2)} = 2, \ldots, s_j^{(mj)} = m_j$.

$$s_1^{(1)} = 1,\ s_1^{(2)} = 2;\ s_1^{(3)} = 3;\ s_2^{(1)} = 1,\ s_2^{(2)} = 2.$$

PCA of means of variables.

- Work out means of histogram g_{ij} : $g_{ij} = \sum_{k_j=1}^{m_j} s_j^{(k_j)} H_{ij}^{(k_j)}$:

TABLE: Table of means of histogram variable.

Variable	Y_1	...	Y_p
ω_1	g_{11}	...	g_{1p}
ω_2	g_{21}	...	g_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
ω_n	g_{n1}	...	g_{np}

- Ordinary PCA of the $n \times p$ table of $(g_{ij})_{i=1,\ldots,n;\ j=1,\ldots,p.}$. Let be u_α principal axes of means of variables.

Transformation of $\{s_j; H_{ij}\} = \left\{s_j^{(k)}; H_{ij}^{(k)}\right\}$ in interval $[\underline{x_{ij}}, \overline{x_{ij}}]$ via Tchebychev's rule : if X is random variable, for $t > 0$

$$P\left(X \in [g_{ij} - t\sigma_{ij}, g_{ij} + t\sigma_{ij}]\right) \geq 1 - \frac{1}{t^2} \ \forall t > 0 \tag{2.1}$$

$g_{ij} = \sum_{k_j=1}^{m_j} s_j^{(k_j)} H_{ij}^{(k_j)}$, σ_{ij} is the standard derivation.

TABLE: Histogram transformed into interval via Tchebychev's rule.

Variable $->$	Y_1	Y_2	...	Y_p
ω_1	$[\underline{x_{11}}, \overline{x_{11}}]$	$[\underline{x_{12}}, \overline{x_{12}}]$	...	$[\underline{x_{1p}}, \overline{x_{1p}}]$
ω_2	$[\underline{x_{21}}, \overline{x_{21}}]$	$[\underline{x_{22}}, \overline{x_{22}}]$	...	$[\underline{x_{2p}}, \overline{x_{2p}}]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
ω_n	$[\underline{x_{n1}}, \overline{x_{n1}}]$	$[\underline{x_{n2}}, \overline{x_{n2}}]$	...	$[\underline{x_{np}}, \overline{x_{np}}]$

Representation of dispersion of individual.

- Construction of hypercubes. A hypercube is assimilate by a $2^p \times p$ matrix. For $p = 2$, we have :

$$M_i = \begin{pmatrix} \underline{x_{i1}} & \underline{x_{i2}} \\ \underline{x_{i1}} & \overline{x_{i2}} \\ \overline{x_{i1}} & \underline{x_{i2}} \\ \overline{x_{i1}} & \overline{x_{i2}} \end{pmatrix}$$

- We project the hypercube on principal axes u_α of PCA of means of variable.
- Der termination of $\min$ and $\max$ of 2^p points projected. Then we represent rectangle.

Problem of unit constraint.

Relative fréquency $H_{ij}^{(kj)}$ are compositional data because of unit constraint. Unit constraint (cf. Aitchison (1986)) cause :

1. Spurious correlation
2. Negative biais
3. Lack of normality
4. Instability of variance

Steps of second approach

Usage of angular transformation in second approach $Arsinus(\sqrt{H_{ij}^{(kj)}})$ allows to remove this problem
Steps of second approach are :

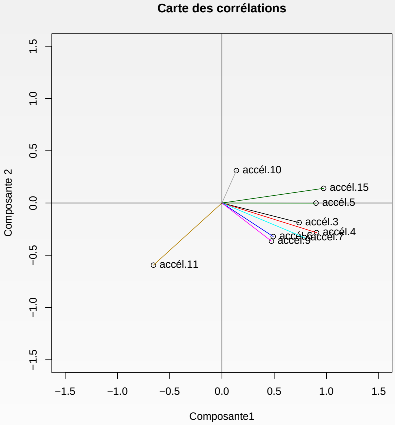
1. Coding of bins
2. Usage of angular transformation $Asin((H_{ij}^{(kj)})^{1/2})$
3. PCA of means of variables
4. Transformation of data into interval by Tchebytchev inequality
5. Construction of hypercube
6. Projection of hypercube on factorial axes

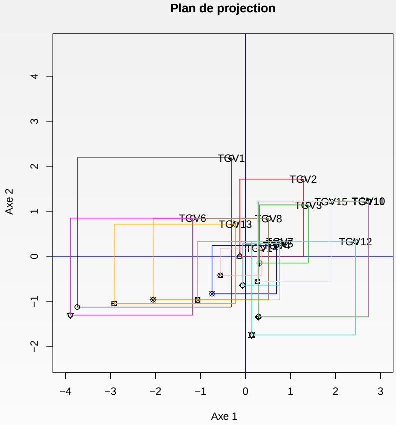
Applications of two approaches on TGV data.

$n = 14$ TGV. Each TGV represents 800.000 values (signal). $p = 9$ variables (Acceleration) captors located in different place on a bridge. Each variable is an histogram with m=20 bins. Objective is To detect anomalies between TGV and characterize them.
We see in two approach mainly 3 groups :

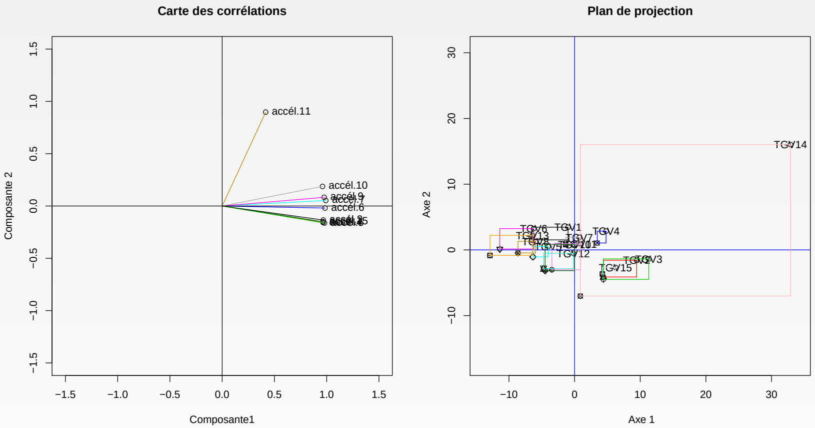
- Groupe 1 : TGV1, TGV6 , TGV13.
- Groupe2 : TGV2, TGV3, TGV10 , TGV11,TGV12, TGV15
- Groupe3 : TGV4, TGV8,, TGV5,TGV7 TGV14

Application of first approach.





Application of second approach.



Conclusion

- Approaches presented improve presented Nagabhushan et al. (2007) methodology, they don't need hypothesis about number of bins of variables.
- Second approach take account unit constraint and seem more robust than the first approach

Bibliography.

[1] Aitchison J.(1986) *The Statistical Analysis of Compositional Data*. London : Chapman and Hall.
[2] Cazes P., Chouakria A., Diday E. et Schektman Y. (1997) : Extension de l'analyse en composantes principales a des données de type intervalle, *Rev. Statistique Appliquee*, Vol. XLV Num. 3 pag. 5-24, France.
[3] Cazes, P. (2002). Analyse factorielle d'un tableau de lois de probabilité. *Revue de Statistique Appliquée*, 50 n° 3, p. 5-24
[4] Diday, E.(1996) : Une introduction à lanalyse des données symboliques, *SFC*,Vannes, France.
[5] Diday E., Noirhomme M. (2008). *Symbolic Data Analysis and the SODAS software*. 457 pages. Wiley. ISBN 978-0-470-01883-5.
[6] Fisher R. A. (1922), On the mathematical foundations of theoretical statistics. Philos. Trans. *Roy. Soc. London* Ser. A 222 309-368.
[7] Makosso Kallyth. S. (2010), Analyse en Composantes Princpales de variables symbolque de type histogramme. Thèse de doctorat, Université Paris IX Dauphine.

[8] Nagabhsushan P. , Kumar P.(2007) : Principal Component Analysis of histogram

Data. *Springer-Verglag Berlin Heidelberg*. EdsISNN Part II LNCS 4492, 1012-1021