Analyse en axes principaux de variables symboliques de type histogramme.

Sun Makosso Kallyth, Edwin Diday

Analyse en axes principaux de variables symboliques de type histogramme.

Sun Makosso Kallyth & Edwin Diday

CEREMADE Université Paris Dauphine, Paris, Place du Maréchal de Lattre de Tassigny 75775 PARIS Cedex 16, France.

Résumé : Dans cet article nous proposons deux nouvelles approches susceptibles d'effectuer une analyse en composantes principales ACP des variables symboliques de type histogramme. Ces approches sont applicables même quand le nombre de modalité des histogrammes diffère. On associe à chaque variable de type histogramme un ensemble de modalités qui ont chacune une fréquence relative lorsqu'on considère un individu. La première approche compte trois étapes. Elle procède d'abord par un codage des modalités des variables de type histogramme. L'objet de ce codage est d'attribuer des valeurs numériques appelées "scores" aux modalités des variables. A ce titre, on propose un codage paramétrique et un codage non paramétrique. La seconde étape consiste à effectuer une ACP des moyennes des histogrammes puis projette en éléments supplémentaires les sommets d'hypercubes induit par l'inégalité de Tchebychev. Dans la seconde approche, en plus des considérations précédentes, on utilise une transformation normalisatrice des données en guise de prétraitement des variables.

Mots clés : variables de type histogramme, hypercube, transformation angulaire, inégalité de tchebytchev.

Summary : This paper deals with Principal Component Analysis (PCA) where the cells of the input data table are not only numerical values but histograms. Histograms are compositional data. PCA extended to such data table can be used when histogram variables don't have the same number of bins. In this paper, we propose at first two ways for attributing scores to variables. Afterward, an ordinary PCA of mean of variables is achieved. Representation of dispersion of variable is done in using Tchebychev inequality. This inequality allows transforming histogram to interval. Then we project hypercube associated to each observation on principal axes. We also propose usage of angular transformation for removing drawbacks of histograms which are compositional data.

Key words : histogram symbolic variables, hypercube, angular transformation, Tchebytchev inequality.

TABLE 1 – Exemple de variable de type histogramme.

	PIB			Taux de mortalité	
Modalité−− >	$\leq$ 1k$	$]1, 20]$	> 20	≤ 0.1	> 0.1
Afrique	0.340	0.660	0.000	0.245	0.755
Europe	0.000	0.322	0.677	0.742	0.258

1 Introduction.

L'ACP permet de représenter dans un espace de dimension réduite les individus d'une population ou d'un échantillon, de détecter les liaisons entre les variables ainsi que les variables séparant le mieux les individus. Dans la pratique on est de plus en plus confronté à l'étude de tableaux plus complexes. S'ils sont par exemple des tableaux munis de règles de taxonomies ou contiennent des variables de type intervalles, à valeurs multiples ou des variables de type histogramme, leur analyse requiert l'usage de techniques liées à leur nature. C'est ce qui justifie le recours au formalisme des objets symboliques de Diday (1996), Diday et al. (2008). Rodriguez et al. (2001), Nagabhushan et al. (2007), Ichino (2008) ont proposé des approches permettant d'effectuer une ACP de variables de type histogramme. Cet article s'inscrit dans ce cadre.

2 Codage des modalités.

Soit n nombre d'individus, p celui des variables, et soit m_j le nombre de modalités d'une variable histogramme Y_j. La table 1 donne un exemple de variable de type histogramme. Une variable de type histogramme Y_j est un n-uplet dont les éléments pour i=1,. . . ,n sont de la forme $Y_{ij} = \{C_j, H_{ij}\}$ où $C_j = (C_j^{(1)}, \ldots, C_j^{(m)})$ sont les modalités de la variable Y_j et $Hij = (H_{ij}^{(1)}, \ldots, H_{ij}^{(m_j)})$ est un m_j vecteur de fréquences relatives. Les fréquences $H_{ij}^{(kj)}$ sont des données compositionnelles car elles vérifient la relation $\sum_{k_j=1}^{m_j} H_{ij}^{(kj)} = 1$ pour $kj = 1, \ldots, m_j$. La première étape de la méthodologie qui est proposée consiste à coder les modalités des variables histogrammes en leur attribuant des scores.

2.1 Codage paramétrique des modalités des variables.

Soit $D = (\alpha, \beta)$ le domaine contenant l'ensemble des valeurs possibles prises par les modalités. Dans la table 1 la plus petite valeur possible prise par la modalité PIB est $\alpha = 0$. En revanche, $\beta = +\infty$ car β la valeur maximale d'une modalité du PIB n'est majorée par aucune valeur. Dans la table 1 on a par exemple $Y_{11} = \{C_1, H_{11}\}$ avec $C_1 = \{]-\infty, 1],]1, 20],]20, +\infty[\}$; $H_{11} = (0.340; 0.660; 0.000)$. Ensuite, on effectue les opérations suivantes :

1. Soit $\delta = \inf_{k_j=1,\ldots,m_j} L_{k_j}$, L_{k_j} étant la longueur des intervalles des $C_j^{(k_j)}$. Si certaines modalités $C_j^{(k_j)}$sont de longueur infinie i.e de la forme $I =]-\infty, a]$ ou $J =]b, +\infty[$, on remplace I par $I' =]e, a]$ où $e = \begin{cases} \alpha & si\ a - \delta < \alpha \\ a - \delta & sinon \end{cases}$.
De même on remplace J par $J' =]b, f]$ avec $f = \begin{cases} \beta & si\ b + \delta > \beta \\ \beta + \delta & sinon \end{cases}$. Dans la table 1 par exemple, la modalité $C_1^{(2)} =]1, 20]$ a la plus petite longueur $L_2 = 19$. Par conséquent on remplace $C_1^{(1)}$ par $C_1'^{(1)} =]\max(-19, 0), 1] =]0, 1]$ et $C_1^{(3)}$ par $C_1'^{(3)} =]20, \min(39, +\infty)] =]20, 39]$.
2. Si les modalités des différentes variables de type histogramme en jeu n'ont pas la même unité de mesure, on remplace chaque intervalle $]a', b']$ par un intervalle ajusté de la forme $]a'/(b' - a'); b'/(b' - a')]$. Au niveau de l'affectation des scores des modalités, le codage paramétrique que l'on préconise assigne à une modalité un vecteur de scores $S_j = (S_j^{(1)}, \ldots, S_j^{(m_j)})$ où $S_j^{(kj)}$ est égale au centre des intervalles ajustées pour $k_j = 1, \ldots, m_j$.

2.2 Codage non paramétrique des modalités des variable

Ce codage utilise comme score des modalités le rang qui leur est associé. Dans la table 1 par exemple, les scores des modalités des classes seront $S_j^{(1)} = 1, S_j^{(2)} = 2, \ldots, S_j^{(mj)} = m_j$.

3 Approche I : ACP des centres et transformations des histogrammes en intervalles.

La seconde étape après le codage des variables consiste à calculer les moyennes $g_{ij} = \sum_{k_j}^{m_j} S_j^{(k_j)} H_{ij}^{(k_j)}$ de chaque histogramme Y_{ij}. Ensuite on effectue une ACP ordinaire du tableau classique $n \times p$ dont les cellules sont constituées des éléments (gij) i=1,...,n ; j=1,...,n. Soient u_α $\alpha = 1, \ldots, p$ les p premiers vecteurs propres de la matrices des covariances de la matrice des (g_{ij}). Les corrélations entre les composantes principales (z_α) $\alpha = 1, \ldots, p$ de cette ACP et les variables initiales permettront d'interpréter les plans factoriels d'individus.
D'autre part, pour représenter la variabilité des individus sur les plans factoriels du fait de leur nature symbolique, on suggère de transformer les variables histogrammes en intervalles. Dans la première approche que l'on propose, nous utilisons l'inégalité de Tchebychev. D'après Tchebychev, pour tout ensemble de données X et un nombre $t \geq 0$, la proportion des données comprises entre $[M_E - t\sigma_E, M_E + t\sigma_E]$ (où M_E est la moyenne empirique, σ_E est l'écart type empirique) est supérieure ou égale à la valeur $1 - 1/t^2$ i.e $P\left(X \in [M_E - t\sigma_E, M_E + t\sigma_E]\right) \geq 1 - 1/t^2$. C'est dans cette optique que pour une valeur de t déterminée, on transforme chaque H_{ij} en $[c_{ij}, d_{ij}]$ via la règle de Tchebychev. On construit par la suite les hypercubes associés aux individus comme dans la méthode

des sommets de Cazes et al. (1997). Un hypercube se définit par rapport à ses sommets au nombre de 2^p s'il y'a p variables. Soit W_i l'hypercube associé au $i$ème individu. On projette chaque hypercube modélisé par W_i (qui est une matrice $2^p \times p$) sur le αième axe factoriel u_α. Sur cet axe factoriel, on représente ainsi le segment de droite joignant le minimum et maximum des 2^p points projetés. La visualisation des individus sur un plan factoriel permet donc d'obtenir des rectangles.

4 Approche II : Transformation normalisatrice des données.

La seconde approche reprend les considérations précédentes et effectue en plus une transformation normalisatrice des données. Les fréquences relatives sont des variables compositionnelles. Aitchison (1986) a décrit les difficultés que l'on rencontre souvent avec de telles variables (biais négatif, corrélation fallacieuse,...). Or les propriétés d'une ACP s'avèrent robustes et son interprétation est plus aisée si les variables en jeu ont une distribution sous-jacente multinormale car cela entraine, entre autre, la linéarité. Si à défaut d'être linéaire, les variables sont par exemple monotones décroissantes, il est judicieux d'effectuer une transformation normalisatrice des variables. Ces transformations rendent la distribution de fréquence symétrique et permettent aussi de stabiliser la variance en la rendant moins dépendante de la moyenne. Quand les données sont exprimées en proportions, ce problème de stabilisation se pose avec acuité. Sous certaines hypothèses, Bishop. et al. (1975) montrent que ce problème de stabilisation conduit à une équation différentielle du type $f'(p) = 1/(\sigma^2(p))$ où p représente le paramètre à stabiliser, f la fonction normalisatrice, $\sigma^2(p)$ la variance de p. Si ce paramètre p est la proportion d'une loi binomiale, alors cette équation différentielle devient $f'(p) = 1/(\sqrt{p(1-p)})$. La solution d'une telle équation est donnée par $f(p) = 2Arsinus(\sqrt{p}) + Cste$. En d'autres termes, pour normaliser des proportions H_{ij}, une possibilité est d'utiliser la transformation angulaire $Arsinus(\sqrt{Hij})$ de Fisher (1922). Elle permet de normaliser les données et de venir à bout de quelques difficultés liées à la nature compositionnelle des données. D'autres transformations telles que le log ratio de Aitchison (1986) ou les transformations puissances de Box et Cox sont possibles.

5 Application.

La première approche qui consiste à coder les modalités, effectuer une ACP de moyenne et utiliser l'inégalité de Tchebychev a été appliquée à des données de la Banque Mondiale disponible sur le site Web World Perspective. Ces données portent sur cinq régions de la planète : l'Afrique, l'Asie orientale, le Proche et moyen orient, l'Amérique du sud et du centre, l'Alena (Amérique du nord et Mexique) et l'Europe. Ces régions sont décrites

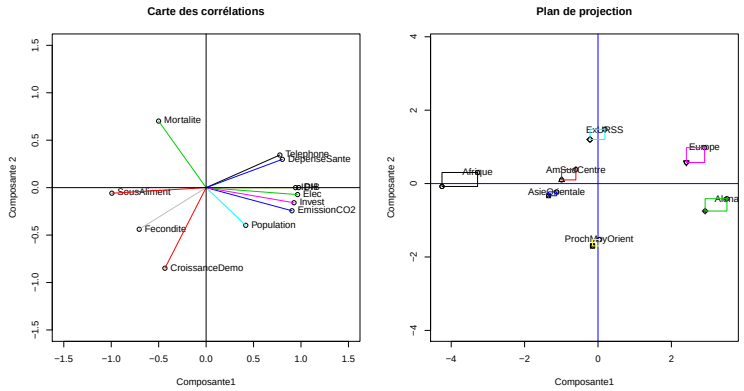

FIGURE 1 – Sorties graphiques de la première approche.

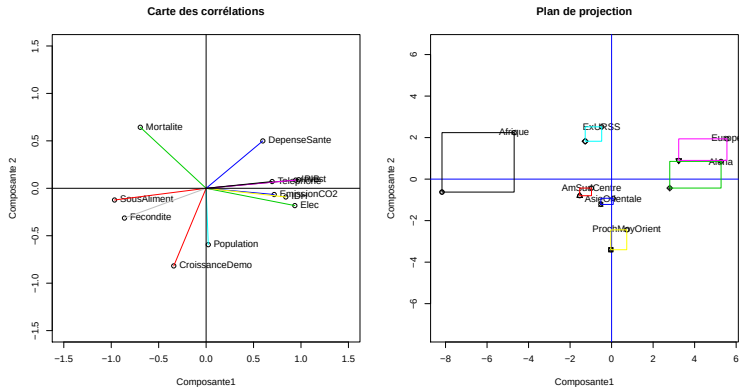

FIGURE 2 – Sorties graphiques de l'approche avec transformation angulaire.

par p=12 variables de type histogrammes. Il s'agit des variables PIB, sous alimentation, Electricité, Emission de CO2, Population, Investissement, Fécondité, Téléphone (nombre de téléphone mobile), Croissance démographique, Taux de mortalité, dépenses en matière de santé. Nous utilisons pour la définition des scores un codage non paramétrique et utilisons une valeur de $t = 1.5$.

La figure 1 illustre les résultats lorsqu'on applique la première approche. On identifie deux groupes de variables. On a d'un côté les moyennes des variables mortalité, sous-alimentation, fécondité et croissance démographique et de l'autre les variables téléphone, dépense danté, PIB, IDH, électricité, investissement, émission de CO_2, Population. Quant au plan factoriel, il oppose les régions sous-développées (Afrique,. . .) des régions développées (Alena et Europe). La seconde approche qui utilise la transformation angulaire donne les mêmes tendances pour ces données (cf. figure 2).

6 Conclusion.

Les approches présentées améliorent celle de Nagabhushan et al. (2007) car elles n'effectuent aucune hypothèse sur le nombre des modalités des variables de type histogramme. La seconde approche a en plus le mérite de prendre en compte la nature compositionnelle des fréquences relatives car elle utilise la transformation angulaire. Toutefois quand le nombre de variables p est très grand, les méthodologies deviennent fastidieuses du point de vue algorithmique et les problèmes de dimensionnalités évoqués par Bellman (1961) peuvent ressurgir.

7 Bibliographie

[1] Aitchison J.(1986) *The Statistical Analysis of Compositional Data*. London : Chapman and Hall.

[2] Bellman, R.E.(1961). *Adaptive Control Processes*. Princeton University Press, Princeton, NJ.

[3] Bishop Y., Feinberg S., Holland P (1975) *Discrete Multivariate Analysis*, Theory and Practice, M.I.T. Press,Cambridge, Mass.

[4] Box G.E.P., and D. R. Cox. (1964). *An analysis of transformations*. J.R. Stat. Soc. B 26 :211-252.

[5] 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.

[6] Cazes, P. (2002). Analyse factorielle d'un tableau de lois de probabilité. *Revue de Statistique Appliquée*, 50 n° 3, p. 5-24

[7] Diday, E.(1996) : Une introduction à l'analyse des données symboliques, *SFC*,Vannes, France.

[8] Diday E. , Noirhomme M. (2008). *Symbolic Data Analysis and the SODAS software*. 457 pages. Wiley. ISBN 978-0-470-01883-5.

[9] Fisher R. A. (1922), On the mathematical foundations of theoretical statistics. Philos. Trans. *Roy. Soc. London* Ser. A 222 309]368.

[10] Ichino M. (2008) : Symbolic PCA for histogram-valued data. *Proceedings IASC* December 5-8, Yokohama, Japan, 5.

[11] Nagabhsushan P. , Kumar P.(2007) : Principal Component Analysis of histogram Data. *Springer-Verglag Berlin Heidelberg*. EdsISNN Part II LNCS 4492, 1012-1021

[12] Rodriguez, O., Diday E., Winsberg S. (2001) : Generalization of the Principal Component Analysis to Histogram Data. *Workshop on Symbolic Data Analysis of the 4th European Conference on Principles and Practice of Knowledge Discovery in Data Bases*, Septembre 12-16, 2000, Lyon,1

[13] Scherrer B. (2008). *Biostatistique. Morin Gaëtan vol. 2. 2nd ed.*