



HAL
open science

Une comparaison de quelques méthodes de classification de variables mixtes

Ndèye Niang, Mory Ouattara, Gilbert Saporta

► **To cite this version:**

Ndèye Niang, Mory Ouattara, Gilbert Saporta. Une comparaison de quelques méthodes de classification de variables mixtes. SFC'2023; Rencontres de la Société Francophone de Classification, Jul 2023, Strasbourg, France. pp.115-116. hal-04158375

HAL Id: hal-04158375

<https://cnam.hal.science/hal-04158375v1>

Submitted on 11 Jul 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Une comparaison de quelques méthodes de classification de variables mixtes

Ndèye Niang¹, Mory Ouattara², Gilbert Saporta¹

¹ Cédric-CNAM, Paris, France

ndeye.niang_keita@cnam.fr gilbert.saporta@cnam.fr

² Université de San Pedro, Côte d'Ivoire

ouattara.mory@usp.edu.ci

Résumé

Nous proposons une nouvelle méthode de classification d'un ensemble de variables qualitatives et quantitatives que nous comparons à celles plus anciennes utilisant des mesures de corrélation entre les variables.

Mots-clés

Classification de variables, données mixtes, ACP, coefficient RV

Abstract

We compare several old and recent methods for clustering a set of qualitative and quantitative variables.

Keywords

clustering of variables, mixed data, PCA, RV coefficient

1 Introduction

Des données à grande échelle et hétérogènes sont de plus en plus récoltées dans de nombreux domaines tels que l'environnement, le domaine médical et clinique, etc... Nous nous intéressons ici à l'hétérogénéité des variables. Le traitement simultané d'un mélange de variables quantitatives et qualitatives, que ce soit en analyse factorielle ou en classification, a fait l'objet d'un grand nombre de travaux Tenenhaus [4], Escofier [7], Saporta [5], Kiers [8], Vigneau et Qanari [2], Pagès [9], Chavent [3].

Une question essentielle est d'utiliser des mesures de similarité cohérentes et comparables pour les différents couples des variables possibles. Plusieurs mesures ont été proposées. Le coefficient de corrélation linéaire ou les rapport de corrélation sont d'usage courant, tandis que diverses solutions ont été proposées pour le cas d'un couple de variables qualitatives.

Nous nous intéressons à la classification d'un ensemble de variables qualitatives et quantitatives. Nous proposons une nouvelle méthode que nous comparons celles plus anciennes utilisant des mesures de corrélation entre les variables.

2 Méthodes de classification

Le traitement simultané d'un mélange de J variables quantitatives \mathbf{x}_j et Q qualitatives $\tilde{\mathbf{x}}_q$, que ce soit en analyse factorielle ou en classification repose souvent sur la détermination d'une ou plusieurs variables synthétiques globales ou locales (ie par classe) optimisant le critère suivant introduit par Tenenhaus [4], réutilisé par Escofier [7], puis Saporta [5], Kiers [8] sous le nom de PCAMIX, et Pagès [9].

$$\max_{\mathbf{c}} \left(\sum_{j=1}^J r^2(\mathbf{c}, \mathbf{x}_j) + \sum_{q=1}^Q \eta^2(\mathbf{c}, \tilde{\mathbf{x}}_q) \right) \quad (1)$$

L'algorithme ClustOfVar [3] utilise ce critère pour effectuer une classification d'un ensemble de variables de nature différentes autour de composantes latentes par groupe, étendant la méthode de Vigneau et Qanari [2] introduite pour des variables exclusivement quantitatives.

Le regroupement de variables autour de composantes est une alternative intéressante aux algorithmes directs qui partent d'un tableau de similarités, de dissimilarités ou de distances entre toutes les variables car il optimise simultanément le regroupement et la représentation des classes par une composante, comme dans une approche clusterwise.

Il est fondamental d'utiliser des mesures de similarité cohérentes et comparables dans les trois cas : un couple de variables quantitatives, un couple de variables qualitatives et un couple formé d'une variable quantitative et d'une variable qualitative.

Les coefficients r^2 de corrélation linéaire et η^2 pour le rapport de corrélation sont d'usage courant, tandis que diverses solutions ont été proposées pour le cas d'un couple de variables qualitatives : le chi-deux et ses dérivés comme le carré du coefficient de Tschuprow T^2 [2] ou la plus grande valeur propre de l'AFC du tableau croisant deux variables qualitatives [3].

Les coefficients associés aux variables qualitatives ne sont cependant pas comparables entre eux ni avec un r^2 car leurs distributions dépendent de leurs nombres de moda-

lités. Dans (1) une variable qualitative joue un rôle d'autant plus grand que son nombre de modalités m_q est élevé. Les coefficients RV d'Escoufier [10] entre tableaux engendrés par chaque variable quantitative et par les tableaux d'indicatrices des modalités des variables qualitatives permettent de définir des similarités euclidiennes égales selon les cas à r^2 , $\frac{\eta^2}{\sqrt{m_q-1}}$ ou T^2 (voir [2]).

On peut alors effectuer des classifications hiérarchiques avec l'algorithme de Ward ou des partitions avec les k -means, soit directement sur la matrice des similarités, soit sur les coordonnées obtenues par la formule de Torgerson.

Cette solution élégante mais un peu oubliée souffre quand même d'un défaut : diviser par la racine carrée du degré de liberté ne corrige pas complètement l'effet du nombre de modalités.

Pour cela, il peut être judicieux d'utiliser comme dissimilarité la p -value du test d'indépendance dans l'esprit de l'algorithme de la vraisemblance du lien [1]. Mais on perd les propriétés euclidiennes. De plus, lorsque le nombre d'observations est très grand, les p -value se rapprochent de zéro (*paradox of large samples*) et ne sont plus utilisables.

Nous proposons de les remplacer par les fractiles correspondants de la loi normale standard dans l'esprit des valeurs-test du logiciel SPAD [11].

3 Applications

Ces différentes approches seront comparées, en terme d'indice de qualité externes (indice de Rand) et de distance entre hiérarchies, sur des jeux de données réelles en particulier sur des données relatives à la pollution de l'air intérieur.

Références

- [1] Nicolau, F. Costa and Bacelar-Nicolau, H., Some trends in the classification of variables. *Data Science, Classification, and Related Methods. Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan* Hayashi, C. and Yajima, K. and Bock, H.H. and Ohsumi, N. and Tanaka, Y. and Baba, Y., 1998.
- [2] Qannari, E.M. and Vigneau, E. and Courcoux, Ph., Une nouvelle distance entre variables. Application en classification, *Revue de Statistique Appliquée*, vol. 46, pp. 21-32, 1998.
- [3] Chavent, M. and Kuentz-Simonet, V. and Liquet, B. and Saracco, J., ClustOfVar : An R Package for the Clustering of Variables, *Journal of Statistical Software*, vol. 50, number. 13, pp. 1–16, 2012
- [4] Tenenhaus, M., Analyse en composantes principales d'un ensemble de variables nominales ou numériques, *Revue de Statistique Appliquée* .vol. 25, pp. 39-56, 1977
- [5] Saporta, G., Simultaneous analysis of qualitative and quantitative data, *Atti della XXXV Riunione Scientifica, Societa Italiana di Statistica, Padova, Italy*, vol. 1, pp. 62-72, 1990
- [6] Vigneau, E. and Qannari, E.M., Clustering of variables around latent components, *Communications in Statistics-Simulation and Computation* ,vol. 32, pp. 1131-1150, 2003

- [7] Escoufier, B., Traitement simultané de variables qualitatives et quantitatives en analyse factorielle, *Cahiers de l'Analyse des Données*, vol.4, pp.137-146, 1979
- [8] Kiers, H., Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables, *Psychometrika*, vol. 56, pp. 197-212, 1991
- [9] Pagés, J., Analyse factorielle de données mixtes, *Revue de Statistique Appliquée*, vol.52, number. 4, pp. 93-111, 2004.
- [10] Robert, P. and Escoufier, Y., A unifying tool for linear multivariate statistical methods : the RV-coefficient, *Journal of the Royal Statistical Society, Series C : Applied Statistics*, vol. 25(3), pp. 257–265, 1976.
- [11] Morineau, A., SPAD.N logicielle pour l'analyse statistique des données, *Modulad-Le Monde des Utilisateurs de L'Analyse de Données*, vol. 6, pp. 27-60, 1991