Packages R pour la détection d'observations atypiques multivariées.

Aurore Archimbaud[a], **Klaus Nordhausen**[b] et **Anne Ruiz-Gazen**[c]

[a]Gremaq (TSE)
Université Toulouse 1 Capitole,
21 allée de Brienne, 31000 Toulouse
aurore.archimbaud@ut-capitole.fr

[b]Department of Mathematics and Statistics
University of Turku
20014 Turku, Finlande
klaus.nordhausen@utu.fi

[c]Gremaq (TSE)
Université Toulouse 1 Capitole,
21 allée de Brienne, 31000 Toulouse
anne.ruiz-gazen@tse-fr.eu

Mots clefs : ACP robuste, distance de Mahalanobis, Invariant Coordinate Selection, Exploratory Projection Pursuit.

La question de la détection d'observations atypiques, autrement dit dont le comportement diffère de celui de la majorité des autres observations, se pose par exemple dans le secteur bancaire, avec la recherche de fraudes, ou dans le secteur industriel, avec la recherche de produits défectueux. De nombreuses méthodes non-supervisées existent en univarié et en multivarié et sont issues du domaine de la statistique mais aussi de l'intelligence artificielle et de l'informatique comme expliqué par exemple dans Hadi et *al.* (2009). Certaines de ces méthodes ont été mises en œuvre dans des packages de R. Dans le poster que nous proposons, nous présenterons et comparerons différentes fonctions extraites des packages *mvoutlier* (Filzmoser et *al.*, 2005), *rrcov* (Todorov and Filzmoser, 2009), *ICS* (Nordhausen et *al.*, 2008) et *REPPlab* (Fischer et *al.*, 2015), adaptées à la détection d'atypiques en multivarié. Les méthodes sous-jacentes sont ou bien des méthodes basées sur des projections révélatrices univariées comme la fonction *EPPlabOutlier* du package *REPPlab* ou bien des méthodes basées sur des estimateurs de matrice de variances-covariances classiques et/ou robustes comme pour la fonction *aq.plot* du package *mvoutlier* ou les fonctions *PcaCov* du package *rrcov* ou *ics* du package *ICS*. Plus précisemment, pour détecter les individus atypiques, nous comparerons quatre méthodes. La première est basée sur le critère de la distance de Mahalanobis robuste tel que détaillé dans Filzmoser *et al.* (2005). La deuxième consiste en un diagnostic graphique tel que proposé par Hubert et *al.* (2005) dans le cadre d'une ACP robuste. La troisième méthode est appelée ICS (Invariant Coordinate Selection) telle que proposée dans Caussinus et Ruiz-Gazen (1993) et étudiée récemment dans Tyler et *al.* (2009). Cette méthode permet de révéler la structure des données en réalisant une diagonalisation simultanée de deux matrices de dispersion plus ou moins robustes. Enfin nous analyserons une méthode de type projections révélatrices (Projection Pursuit en anglais) présentée notamment dans Ruiz-Gazen et *al.* (2010) et qui consiste à projeter les données sur une direction qui maximise un indice de projection telle que le kurtosis par exemple. Le poster présentera ces outils à partir d'exemples et insistera sur les fonctions graphiques des différents packages.

Références

[1] Caussinus, H., Ruiz-Gazen, A. (1993). Projection pursuit and generalized principal component analysis. In *New Directions in Statistical Data Analysis and Robustness* (eds S. Morgenthaler, E. Ronchetti and W. A. Stahel), 35-46, Basel: Birkhäuser.

[2] Filzmoser, P., Garrett, R.G., Reimann, C. (2005), Multivariate outlier detection in exploration geochemistry. *Computers & Geosciences*, *31*, 579-587.

[3] Filzmoser, P., Gschwandtner, M. (2015). mvoutlier: multivariate outlier detection based on robust methods. R package version 2.0.6. http://CRAN.R-project.org/package=mvoutlier

[4] Fischer, D., Berro, A., Nordhausen, K., Ruiz-Gazen, A. (2015). REPPlab: R Interface to EPP-lab, a Java Program for Exploratory Projection Pursuit. R package version 0.9.1. http://CRAN.R-project.org/package=REPPlab

[5] Hadi, A. S., Imon, A. H. M., Werner, M. (2009). Detection of outliers. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(1), 57-70.

[6] Hubert, M., Rousseeuw, P. J., Vanden Branden, K. (2005). ROBPCA: a new approach to robust principal component analysis. *Technometrics*, *47*(1), 64-79.

[7] Nordhausen, K., Oja, H., Tyler, D.E., (2008). Tools for Exploring Multivariate Data: The Package ICS. *Journal of Statistical Software 28*(6), 1-31. URL http://www.jstatsoft.org/v28/i06/.

[8] Ruiz-Gazen, A., Larabi Marie-Sainte, S., Berro, A. (2010). Detecting multivariate outliers using projection pursuit with particle swarm optimization, *COMPSTAT2010*, 89-98.

[9] Todorov, V., Filzmoser, P., (2009). An Object-Oriented Framework for Robust Multivariate Analysis. *Journal of Statistical Software*,*32*(3), 1-47. URL http://www.jstatsoft.org/v32/i03/.

[10] Tyler, D. E., Critchley, F., Dümbgen, L., Oja, H. (2009). Invariant coordinate selection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(3), 549-592.