LA MÉTHODE ICS POUR DÉTECTER DES OBSERVATIONS ATYPIQUES EN MULTIVARIÉ

Anne Ruiz-Gazen (*), Aurore Archimbaud (*) et Klaus Nordhausen (**)

(*) Toulouse Scool of Economics, Université Toulouse 1 Capitole
(**) Institute of Statistics & Mathematical Methods in Economics, TU Wien

anne.ruiz-gazen@tse-fr.eu

Mots-clés : Analyse en Composantes Principales, Affine invariance, Distance de Mahalanobis, ICSOutlier et ICSShiny, Point de rupture, Qualité des données,

Résumé

La méthode ICS (Invariant Coordinate Selection) est une méthode multivariée qui permet de détecter la présence d'observations atypiques ou de groupes au sein de données caractérisées par des variables quantitatives. Cette méthode est basée sur la diagonalisation conjointe de deux estimateurs de matrices de dispersion. Lorsque les estimateurs sont équivariants par transformation affine, ICS est une méthode invariante pas transformation affine. Elle conduit à des composantes dites invariantes et telles que la distance euclidienne entre observations calculées à partir des composantes est équivalente à une distance de Mahalanobis dans l'espace des variables de départ. La distance de Mahalanobis est une méthode classique pour la détection d'observations atypiques. On montre toutefois qu'elle rencontre des difficultés dans le cas où les observations sont en grande dimension mais que les observations atypiques engendrent un espace de faible dimension. ICS permet de pallier le problème en autorisant la sélection de composantes en amont du calcul de distance. Dans le cas où les individus atypiques sont contenus dans un espace de faible dimension, les propriétés théoriques d'ICS montrent, notamment pour des mélanges de lois elliptiques, que l'espace d'intérêt est engendré par les composantes associées aux plus grandes ou/et aux plus petites valeurs propres issues de la diagonalisation simultanée des estimateurs de dispersion. L'objectif du présent article est d'étudier l'influence du choix des estimateurs de matrices de dispersion sur les résultats d'ICS en fonction de la proportion d'observations atypiques au sein des données. Pour certains de ces estimateurs, il est possible d'obtenir des résultats théoriques tandis que pour d'autres nos analyses se basent sur des exemples simulés. On montre en particulier que le point de rupture des estimateurs de dispersion a un impact sur les résultats mais que des estimateurs à point de rupture nul peuvent aussi être envisagés. L'utilisation des récents packages R, ICSOutlier et ICSShiny permet aussi des comparaisons sur des exemples réels.