Comparaison de méthodes statistiques multivariées pour la détection d'observations atypiques.

Aurore Archimbaud[1], Klaus Nordhausen[2] & Anne Ruiz-Gazen[3]

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

[2] *Department of Mathematics and Statistics, University of Turku, 20014 Turku, Finlande, E-mail: klaus.nordhausen@utu.fi*

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

Résumé. Dans cette présentation, nous nous intéressons à la détection d'observations atypiques, comme par exemple des fraudes ou des produits défectueux, au sein de données numériques multivariées. Différentes méthodes non-supervisées basées sur l'analyse de matrices de variances-covariances classiques ou robustes existent dans la littérature statistique. Notre objectif est de comparer trois de ces méthodes : la distance de Mahalanobis, la méthode ICS (Invariant Coordinate Selection) et l'ACP robuste avec son diagnostic graphique. Ces méthodes conduisent chacune à des scores qui sont calculés pour toutes les observations, avec des scores élevés associés aux éventuelles observations atypiques. Nous montrons en particulier que seule la méthode ICS permet la sélection de composantes pertinentes pour la détection d'atypiques ce qui constitue un avantage si le nombre de variables non pertinentes pour caractériser les atypiques est élevé. Les résultats seront illustrés sur des exemples simulés et sur des exemples réels.

Mots-clés. ACP robuste, distance de Mahalanobis, Invariant Coordinate Selection.

Abstract. In this presentation, we are interested in detecting outliers, like for example fraud or manufacturing faults, in multivariate numeric data sets. Several non-supervized methods that are based on robust and non-robust covariance matrix estimators exist in the statistical literature. Our aim is to compare three methods: the Mahalanobis distance, the ICS method (Invariant Coordinate Selection) and the robust PCA method with its diagnostic plot. Each of these methods leads to scores computed for all the observations where high scores are associated with potential outliers. We show in particular that the ICS method is the only one that permits a selection of the relevant components for detecting outliers. This is an advantage when the data are very noisy. The results will be illustrated on simulated and real data sets.

Keywords. Invariant Coordinate Selection, Mahalanobis distance, robust PCA.

1 Introduction

La détection d'observations atypiques, c'est-à-dire dont le comportement diffère de celui de la majorité des autres observations, est un problème qui se pose dans de nombreux domaines, notamment dans le secteur bancaire avec la recherche de fraudes et dans le secteur industriel avec la recherche de produits défectueux. Les méthodes de détection existantes sont issues du domaine de la statistique mais aussi de l'intelligence artificielle et de l'informatique comme expliqué dans Hodge et Austin (2004) et Hadi *et al.* (2009). Dans cette présentation, nous considérons trois méthodes statistiques qui ont été proposées dans un cadre multivarié pour la détection d'observations atypiques. Il s'agit du critère de la distance de Mahalanobis (voir Rousseeuw et Van Zomeren, 1990), de la méthode appelée "Invariant Coordinate Selection" qui a été proposée dans Tyler *et al.* (2009) et qui généralise la méthode proposée par Caussinus et Ruiz-Gazen (1993) et enfin de l'ACP robuste avec son diagnostic graphique tel que proposé par Hubert *et al.* (2005). Toutes ces méthodes permettent de prendre en compte la structure de covariance des données en se basant sur des estimateurs de matrice de variances-covariances classiques et/ou robustes. Toutefois, elles ne sont pas identiques et notre objectif est de discuter leurs différences.

2 Présentation des méthodes

Les trois méthodes présentées utilisent des estimateurs de matrice de variances-covariances robustes c'est-à-dire peu sensibles à la présence de valeurs atypiques. Les caractéristiques essentielles des estimateurs robustes sont la B-robustesse qui correspond à une fonction d'influence bornée et un haut de point de rupture qui permet de ne pas obtenir un résultat complètement insensé même en présence d'une forte proportion d'individus atypiques (voir Ruiz-Gazen, 2012, pour une présentation simplifiée de ces concepts). De nombreux estimateurs robustes de position et de variances-covariances ont été proposés dans la littérature statistique avec notamment les M-estimateurs qui sont B-robustes mais dont le point de rupture décroît avec le nombre de variables et l'estimateur de déterminant minimum appelé MCD qui est B-robuste et à haut point de rupture (voir Maronna *et al.*, 2006, pour plus de détail). Dans ce résumé, nous considérons uniquement des cas où la proportion d'observations atypiques est faible, de l'ordre de 2 ou 3% maximum. La conséquence de cette hypothèse est qu'il n'est pas nécessaire de considérer des estimateurs à haut point de rupture. Notons que dans certains contextes notamment industriels, cette hypothèse est tout à fait acceptable.

2.1 Méthode basée sur la distance de Mahalanobis

Il s'agit de calculer pour chaque observation sa distance de Mahalanobis au centre de la distribution. Ainsi, si on note $y_1, \ldots, y_n$, n observations caractérisées par p variables quantitatives (les y_i sont des vecteurs colonnes), pour un estimateur μ_n de position et un

estimateur Σ_n de matrice de variances-covariances, on définit la distance de Mahalanobis au carré d'une observation y_i à μ_n par :

$$\mathrm{MD}^2_{\mu_n,\Sigma_n}(y_i) = (y_i - \mu_n)^t \Sigma_n^{-1} (y_i - \mu_n).$$

Rousseeuw et Van Zomeren (1990) proposent d'utiliser cette distance comme score pour mesurer l'atypicité des observations en utilisant des estimateurs de position et d'échelle robustes. Le critère pour décider du seuil au-delà duquel une observation est considérée comme atypique est basé sur la normalité des observations et correspond au quantile d'une loi du χ^2. L'idée d'utiliser des estimateurs robustes au lieu de la moyenne et de la matrice de variances-covariances classiques vient du fait que les estimateurs usuels sont sensibles aux observations atypiques et peuvent donc être contaminés, ce qui peut conduire à un effet dit de masque où les observations atypiques ne sont plus détectées comme telles. Dès lors que l'on utilise des estimateurs de position et d'échelle vérifiant la propriété d'affine équivariance, le score obtenu est invariant par transformation affine des données.

2.2 Invariant Coordinate Selection

La méthode a été définie dans Tyler *et al.* (2009) et généralise des résultats de Caussinus et Ruiz-Gazen (1993) et Caussinus *et al.* (2003). Il s'agit d'effectuer la diagonalisation conjointe de deux estimateurs de matrices de variances-covariances. Dans le cas qui nous intéresse ici, *i.e.* la détection d'atypiques, les deux matrices sont respectivement la matrice de variances-covariances empirique usuelle et une matrice robuste telle que définie par exemple dans Caussinus et Ruiz-Gazen (1993). Plus précisemment, si on note à nouveau $y_1, \ldots, y_n$, n observations caractérisées par p variables quantitatives, et que l'on considère V_1 et V_2 deux estimateurs de variances-covariances avec V_1 l'estimateur de matrice de variance empirique classique et V_2 un estimateur plus robuste, la méthode ICS consiste à diagonaliser conjointement les deux estimateurs V_1 et V_2 en calculant des vecteurs $l_1, \ldots, l_p$ tels que

$$V_1\, l_i = \lambda_i\, V_2\, l_i\,, \quad \text{pour} \quad i = 1, \ldots, p.$$

Les valeurs propres λ_i sont classées par ordre décroissant ($\lambda_1 \geq \ldots \geq \lambda_p$) et on normalise généralement les vecteurs l_i de telle sorte que

$$l_i'\, V_2\, l_i = 1 \;\text{ pour }\; i = 1, \ldots, p \quad \text{et} \quad l_i'\, V_2\, l_j = 0 \;\text{ pour }\; i = 1, \ldots, p \;\text{ et }\; i \neq j.$$

On note $L = (l_1, \ldots l_p)$ la matrice $p \times p$ qui contient les vecteurs l_i en colonnes. Pour une observation y_i et un estimateur μ_n de position, on obtient les composantes z_i par ICS en suivant Caussinus et Ruiz-Gazen (1993) et en calculant $z_i' = (y_i - \mu_n)'L$. Remarquons à ce niveau que dans Tyler *et al.*, les composantes d'ICS différent de celle présentées ici à un coefficient près sur chaque composante. Par ailleurs, notons qu'avec la définition proposée par Caussinus et Ruiz-Gazen (1993), les distances euclidiennes entre observations

calculées à partir de l'ensemble des composantes d'ICS correspondent à la distance de Mahalanobis au sens de la matrice robuste V_2. L'idée de la méthode est d'obtenir, à partir de la comparaison de deux estimateurs de variances-covariances, des projections orthogonales des observations qui révèlent une structure telle que la présence d'atypiques dans le cas où on compare un estimateur robuste à un estimateur non-robuste. Dans le cas de distributions elliptiques, les deux estimateurs vont estimer les mêmes paramètres et aucune structure ne sera détectée. Par contre, en présence d'observations atypiques, l'estimateur robuste va se différencier de l'estimateur non-robuste et la méthode ICS va en quelque sorte pointer dans la ou les directions où se situent les observations atypiques. La méthode est invariante par transformation affine au sens que les composantes obtenues par projection des observations initiales restent inchangées par transformation affine des données. A partir de ces composantes, on peut calculer un score qui correspond à la distance à l'origine des individus mesurés par les composantes. Caussinus *et al.* (2003) ont préconisé de ne conserver dans ce calcul de score que les composantes associées aux valeurs propres significativement supérieures à 1, les autres valeurs propres étant supposées associées à du bruit.

2.3 ACP robuste et diagnostic graphique

Comme précisé dans Jolliffe (2002), l'ACP n'est pas adaptée à la détection d'individus atypiques. Toutefois, il n'est pas rare de détecter des atypiques sur les premiers ou les derniers axes d'une ACP. Hubert *et al.* (2005) proposent d'utiliser une ACP avec un estimateur de variances-covariances robuste et un diagnostic graphique associé. Une fois choisi le nombre de composantes à retenir, le graphique consiste en un diagramme de dispersion. La distance à l'origine de chaque observation calculée à partir des composantes conservées et pondérées par l'inverse de leur variance est représentée en abscisse. Pour chaque observation, sa distance à l'origine dans l'espace orthogonal à celui engendré par les composantes retenues est représentée en ordonnée. On associe ainsi deux scores à chaque observation et les observations associées à des score élevés sont considérées comme atypiques. Cette méthode, contrairement aux deux précédentes présente l'inconvénient de ne pas être invariante par transformation affine des données. Ainsi, les résultats diffèrent selon que les données sont préalalement standardisées ou pas.

Il est important de remarquer que lorsque V_1 est l'estimateur de matrice de variances-covariances usuel et V_2 est un estimateur robuste, la méthode ICS, telle que proposée par Tyler *et al.* (2009), est équivalente à une ACP robuste lorsque l'estimateur V_2 est calculé sur des données rendues sphériques c'est-à-dire transformées de telle sorte que leur matrice de variances-covariances usuelle soit égale à l'identité. La méthode ICS n'est donc pas très différente d'une ACP robuste réalisée sur des données standardisées.

3 Comparaison et Perspective

Cette présentation permettra de comparer les trois méthodes de détection d'observations atypiques multivariées présentées précédemment sur des exemples simulés. La méthode ICS permet la sélection de composantes qui mettent en évidence les observations atypiques alors que la distance de Mahalanobis prend en compte l'ensemble des composantes et donc éventuellement des composantes qui bruitent le résultat et empêchent le repérage des atypiques. Ce point sera illustré sur des exemples simulés dans le cas par exemple où une seule variable est informative pour expliquer l'atypicité des individus mais que l'on dispose de nombreuses variables supplémentaires correspondant à du bruit. L'ACP robuste quant à elle permet de sélectionner des composantes. Toutefois, l'objectif d'une ACP n'est pas la détection d'observations atypiques et il est bien connu que de telles observations peuvent être associées aux dernières composantes de l'analyse. Ainsi, on doit continuer d'analyser l'ensemble des axes lorsque l'on effectue une ACP robuste et le diagnostic graphique représente non seulement le score sur les premières composantes mais aussi un score prenant en compte les dernières composantes. L'utilisation de la méthode ICS apparaît donc comme la mieux adaptée à la recherche d'atypiques même si les deux autres méthodes peuvent conduire à des résultats similaires comme l'atteste notre expérience du traitement de nombreux fichiers de données réelles mais confidentielles.

Parmi les pistes de recherche sur le sujet, nous travaillons actuellement sur le problème de la détection d'observations atypiques lorsque le nombre de dimensions est grand, en particulier devant le nombre d'observations. Dans ce cas, les méthodes précédentes doivent être adaptées. L'ACP robuste telle que proposée par Hubert *et al.* (2005) est déjà adaptée à ce contexte de grande dimension en utilisant notamment en amont de l'analyse une méthode de type projections révélatrices. L'adaptation de la méthode ICS est un travail en cours.

Bibliographie

[1] Caussinus, H., Fekri, M., Hakam, S. and Ruiz-Gazen, A. (2003), A monitoring display of Multivariate Outliers, *Computational Statistics and Data Analysis*, 44(1-2), 237–252.
[2] Caussinus, H. and 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: Birkhuser.
[3] Hadi, A. S., Imon, A. H. M. et Werner, M. (2009), Detection of outliers, *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1), 57–70.
[4] Hodge, V. J., et Austin, J. (2004), A survey of outlier detection methodologies, *Artificial Intelligence Review*, 22(2), 85–126.
[5] Hubert, M., Rousseeuw, P. J. et Vanden Branden, K. (2005), ROBPCA: a new approach to robust principal component analysis, *Technometrics*, 47(1), 64–79.

[6] Jolliffe, I. (2002), *Principal component analysis*, John Wiley & Sons, Ltd.
[7] Maronna, R. A., Martin, D. et Yohai, V. (2006). *Robust statistics*. John Wiley & Sons, Chichester.
[8] Rousseeuw, P. J. et Van Zomeren, B. C. (1990), Unmasking multivariate outliers and leverage points, Journal of the American Statistical Association, 85(411), 633–639.
[9] Ruiz-Gazen, A. (2012), Robust statistics: a functional approach, *Annals of Institut de Statistiques de l'Universit de Paris*, 56(2-3), 49-64.
[10] Tyler, D. E., Critchley, F., Dmbgen, L. et Oja, H. (2009), Invariant coordinate selection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3), 549–592.