Distance de Mahalanobis et ICS pour la détection d'observations atypiques.

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

[1] *TSE-R, Université Toulouse 1 Capitole, 21 allée de Brienne, 31000 Toulouse,*
E-mail: aurore.archimbaud@ut-capitole.fr,
anne.ruiz-gazen@tse-fr.eu
[2] *Department of Mathematics and Statistics, University of Turku,*
20014 Turku, Finlande,
E-mail: klaus.nordhausen@utu.fi

Résumé. Dans cette présentation, nous nous intéressons à la détection non supervisée d'observations atypiques, au sein de données numériques multivariées. Nous considérons plus particulièrement le cas d'une faible proportion d'observations atypiques, comme par exemple dans la détection de fraudes ou de produits défectueux. La distance de Mahalanobis permet de calculer un score associé à chaque observation en prenant en compte la structure de covariances des données. Des scores élevés indiquent de potentiels atypiques. Nous montrons les limites de cette méthode dans le cas où la dimension augmente alors que la structure d'intérêt reste dans un espace de dimension fixe. La méthode ICS (Invariant Coordinate Selection) permet de pallier cet inconvénient en ne sélectionnant que des composantes pertinentes pour la détection d'atypiques. Les résultats seront illustrés sur des exemples simulés et sur des exemples réels à l'aide du package R *ICSOutlier* que nous avons développé.

Mots-clés. Invariant Coordinate Selection, Matrice de covariances robustes, Sélection de composantes.

Abstract. In this presentation, we are interested in detecting outliers in an unsupervised way in multivariate numerical data sets. We focus specifically on the case of a small proportion of outlying observations, like for example fraud or manufacturing faults. The Mahalalanobis distance computes a score for each observation taking into account the covariance structure of the data set. High scores indicate possible outliers. However, the limitation of this method appears if the dimension of the data increases while the structure of interest remains in a fixed dimension subspace. The ICS method (Invariant Coordinate Selection) overcomes this drawback by selecting relevant components for outlier detection. The results will be illustrated on simulated and real data sets through the R package *ICSOutlier* we implemented.

Keywords. Invariant Coordinate Selection, Robust covariance matrix, Components selection.

1 Introduction

La détection d'observations atypiques, c'est-à-dire d'observations dont le comportement diffère de celui de la majorité des autres observations, est un problème important. Tout d'abord, la recherche d'observations atypiques peut être le but d'une analyse statistique, notamment dans le secteur bancaire avec la recherche de fraudes et dans le secteur industriel avec la recherche de produits défectueux. Ensuite, de nombreuses méthodes statistiques sont sensibles à la présence de valeurs atypiques et produisent des résultats contaminés en leur présence. Dans ce cas il est important de détecter ces individus pour qu'ils n'impactent pas les conclusions de l'étude. De nombreuses méthodes de détection de telles observations existent et sont issues du domaine de la statistique mais aussi de l'intelligence artificielle et de l'informatique comme expliqué dans Hodge et Austin (2004), Hadi *et al.* (2009) et Aggarwal (2017).

Dans cette présentation, nous nous focalisons sur des méthodes affines équivariantes utilisant des matrices de variances-covariances classiques ou robustes et proposées dans un cadre multivarié pour la détection d'atypiques. La méthode la plus connue est sans doute la distance de Mahalanobis et ses variantes robustes (voir Rousseeuw et Van Zomeren, 1990, pour l'intérêt d'utiliser une version robuste, Ruiz-Gazen, 2012, pour une présentation simplifiée des concepts de robustesse et Maronna *et al.*, 2006 pour une présentation des estimateurs de matrice de covariances robustes). Toutefois, dans le cas où les observations atypiques sont contenues dans un sous-espace de dimention fixée et que la dimension des données augmente, nous montrons que la distance de Mahalanobis n'est pas adaptée à la détection d'atypiques. Dans ce contexte, nous préconisons la méthode "Invariant Coordinate Selection" proposée dans Tyler *et al.* (2009) et qui généralise la méthode proposée par Caussinus et Ruiz-Gazen (1993). A l'instar de la distance de Mahalanobis, ICS permet de calculer un score d'atypicité mais la sélection de composantes permet de s'affranchir du problème mentionné précédemment lorsque la dimension des données augmente.

2 Limitation de la distance de Mahalanobis

Pour obtenir une propriété théorique de la distance de Mahalanobis dans le cas où les observations atypiques sont contenues dans un sous-espace de dimention fixée et que la dimension des données augmente, nous nous plaçons dans le cadre d'un modèle où la majorité des données suit une distribution Gaussienne et les individus atypiques forment q groupes suivants des distributions Gaussiennes avec des paramètres de position différents de celui du groupe majoritaire.

Plus formellement, soit $\mathbf{X} = (X_1, \ldots, X_p)'$ un vecteur aléatoire réel p-multivarié et supposons que la distribution de $\mathbf{X}$ soit un mélange de $(q+1)$ distributions Gaussiennes avec $q+1 < p$, différents paramètres de position $\boldsymbol{\mu}_h$, pour $h = 0, \ldots, q$, et la même matrice de variance-covariance $\boldsymbol{\Sigma}_W$ définie positive :

$$\mathbf{X} \sim (1-\epsilon)\,\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_W) + \sum_{h=1}^{q} \epsilon_h\,\mathcal{N}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_W) \quad \text{avec} \quad \epsilon = \sum_{h=1}^{q} \epsilon_h < \frac{1}{2} \tag{1}$$

Le paramètre de position est $\boldsymbol{\mu}_{\mathbf{X}} = (1-\epsilon)\,\boldsymbol{\mu}_0 + \sum_{h=1}^{q} \epsilon_h \boldsymbol{\mu}_h$, la matrice de variance-covariance intra est $\boldsymbol{\Sigma}_W$, la matrice de variance-covariance inter est : $\boldsymbol{\Sigma}_B = (1-\epsilon)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_{\mathbf{X}})(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_{\mathbf{X}})' + \sum_{h=1}^{q} \epsilon_h (\boldsymbol{\mu}_h - \boldsymbol{\mu}_{\mathbf{X}})(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{\mathbf{X}})'$, et la matrice de variance-covariance totale est $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_W$.

Pour le modèle (1), on peut définir une distance de Mahalanobis au carré à $\boldsymbol{\mu}_{\mathbf{X}}$ par :

$$d^2(\mathbf{X}) = (\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})$$

ainsi qu'une distance de Mahalanobis robuste au carré par :

$$d_R^2(\mathbf{X}) = (\mathbf{X} - \boldsymbol{\mu}_0)' {\boldsymbol{\Sigma}_W}^{-1} (\mathbf{X} - \boldsymbol{\mu}_0).$$

On introduit les variables aléatoires $\mathbf{X}_{no} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_W)$, où *no* correspond aux observations non atypiques, et $\mathbf{X}_{o,h} \sim \mathcal{N}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_W)$, où *o* correspond aux observations atypiques. On suppose que $\mathbf{X}_{no}$ et $\mathbf{X}_{o,h}$, pour $h = 1, \ldots, q$, sont indépendantes et on s'intéresse au comportement de la différence entre la distance au carré de $\mathbf{X}_o$ et celle de $\mathbf{X}_{no}$, pour chacune des deux distances de Mahalanobis, quand p augmente. La distribution des différences est non triviale et on se concentre donc sur la distribution asymptotique quand p est grand. On obtient, en utilisant le théorème central limite de Lindeberg-Feller, la proposition suivante où E désigne l'espérance (voir Archimbaud *et al.*, 2016b, pour plus de détails).

Proposition 1 *Supposons que q est fixé et que p augmente, la distribution des différences*

$$\frac{1}{2\sqrt{p}} \left(d^2(\mathbf{X}_{o,h}) - d^2(\mathbf{X}_{no}) - E\left(d^2(\mathbf{X}_{o,h}) - d^2(\mathbf{X}_{no}) \right) \right),$$

et celle de

$$\frac{1}{2\sqrt{p}} \left(d_R^2(\mathbf{X}_{o,h}) - d_R^2(\mathbf{X}_{no}) - E\left(d_R^2(\mathbf{X}_{o,h}) - d_R^2(\mathbf{X}_{no}) \right) \right)$$

convergent vers une distribution Gaussienne centrée réduite et les espérances ne dépendent pas de p.

Si on interprète la Proposition 1, il apparaît que, si les atypiques appartiennent à un sous-espace de dimension $q < p$ et que p est grand, alors la probabilité que la distance de Mahalanobis d'un individu atypique excède la distance de Mahalanobis d'un individu généré par la distribution majoritaire est faible car, d'après l'approximation, la variance des différences augmente lorsque p augmente. Ce constat rend la détection des observations atypiques de plus en plus difficile lorsque p augmente et que q est fixe. Il est préférable de projeter les observations dans le sous-espace de dimension q puis de calculer une distance seulement dans ce sous-espace. La méthode ICS permet justement de pallier cet inconvénient en permettant de sélectionner des composantes adéquates.

3 Invariant Coordinate Selection (ICS)

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 affine quivariants de matrices de covariances définies positives. Dans le cas qui nous intéresse ici, *i.e.* la détection d'atypiques en faible proportion, Archimbaud *et al.* (2016b) montrent qu'un couple intéressant de matrices est constitué de la matrice de covariances empirique usuelle COV et de la matrice dite des quatrièmes moments, $\text{COV}_4(\mathbf{X}_n) = 1/((p+2)n)\sum_{i=1}^{n} r_i^2(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ où r_i^2 désigne la distance de Mahalanobis de $\mathbf{x}_i$ évaluée par rapport à $\bar{\mathbf{x}}$ au sens de COV.

Plus formellement, si on considère $\mathbf{X}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$, n observations en p dimensions, $\mathbf{V}_1(\mathbf{X}_n) = \text{COV}_4(\mathbf{X}_n)$ et $\mathbf{V}_2(\mathbf{X}_n) = \text{COV}(\mathbf{X}_n)$, ICS détermine la matrice $\mathbf{B}(\mathbf{X}_n)$ de dimension $p \times p$ et la matrice diagonale $\mathbf{D}(\mathbf{X}_n)$ telle que :

$$\mathbf{V}_1(\mathbf{X}_n)^{-1}\mathbf{V}_2(\mathbf{X}_n)\mathbf{B}(\mathbf{X}_n)' = \mathbf{B}(\mathbf{X}_n)'\mathbf{D}(\mathbf{X}_n).$$

avec la standardisation suivante : $\mathbf{B}(\mathbf{X}_n)\mathbf{V}_1(\mathbf{X}_n)\mathbf{B}'(\mathbf{X}_n) = \mathbf{I}_p$ et $\mathbf{B}(\mathbf{X}_n)\mathbf{V}_2(\mathbf{X}_n)\mathbf{B}'(\mathbf{X}_n) = \mathbf{D}(\mathbf{X}_n)$. $\mathbf{D}(\mathbf{X}_n)$ contient les valeurs propres de $\mathbf{V}_1(\mathbf{X}_n)^{-1}\mathbf{V}_2(\mathbf{X}_n)$ dans l'ordre décroissant et $\mathbf{B}(\mathbf{X}_n) = (\mathbf{b}_1, \ldots, \mathbf{b}_p)'$ les vecteurs propres associés en lignes. Des scores sont obtenus en projettant les observations centrées par rapport à l'estimateur de position $\mathbf{m}_1(\mathbf{X}_n)$ associé à la matrice de dispersion $\mathbf{V}_1(\mathbf{X}_n)$: $\mathbf{Z}_n = (\mathbf{z}_1, \ldots, \mathbf{z}_n)' = (\mathbf{X}_n - \mathbf{1}_n\mathbf{m}_1'(\mathbf{X}_n))\mathbf{B}'(\mathbf{X}_n)$.

Il est intéressant de remarquer que la norme euclidienne des scores centrés correspond exactement à la distance de Mahalanobis évaluée par rapport à $\mathbf{m}_1(\mathbf{X}_n)$ au sens de $\mathbf{V}_1(\mathbf{X}_n)$. Par ailleurs, l'intérêt d'ICS est prouvé thoriquement pour certains mélanges de loi elliptiques (voir Tyler *et al.*, 2009, pour plus de détails). Toutefois, la principale difficulté avec ICS est d'estimer correctement le nombre de composantes pertinentes. Un processus de sélection automatique du nombre de composantes est étudié dans Archimbaud *et al.* (2016b), mais nous nous concentrons ici uniquement sur l'utilisation de l'éboulis des valeurs propres en utilisant la règle du coude. Remarquons que pour le couple COV_4-COV, les valeurs propres correspondent au kurtosis des composantes, ce kurtosis étant ainsi maximisé par la première composante d'ICS.

Après avoir sélectionné le sous-espace de dimension k, un indice d'atypicité est obtenu pour chaque observation $\mathbf{x}_i$ en calculant la norme du vecteur $\mathbf{Z}_{n,k} = (\mathbf{z}_1, \ldots, \mathbf{z}_k)$. Ce score est défini par :

$$ICSD_{\mathbf{V}_1(\mathbf{X}_n)^{-1}\mathbf{V}_2(\mathbf{X}_n)}(\mathbf{x}_i, k) = ||\mathbf{Z}_{n,k}||.$$

Les observations identifiées comme atypiques sont celles dont le score $ICSD_{\mathbf{V}_1(\mathbf{X}_n)^{-1}\mathbf{V}_2(\mathbf{X}_n)}$ dépasse un certain quantile obtenu à partir de simulations sous le modèle normal comme proposé dans Archimbaud *et al.* (2016b).

4 Mise en oeuvre et applications de ICS

Nous avons développé un package R pour la mise en œuvre d'ICS dans le cas de la détection d'observations atypiques, appelé *ICSOutlier* (Archimbaud *et al.*, 2016a). Cette présentation nous permettra d'illustrer l'utilisation de ce nouveau package en insistant sur la comparaison des performances de la distance de Mahalanobis et de la méthode ICS sur des exemples simulés et réels.

La Figure 1 ci-dessous illustre ce point en considérant l'exemple des données HTP qui sont incluses dans le package *ICSOutlier* et qui consistent en 902 observations pour lesquelles on dispose de 88 mesures numériques.

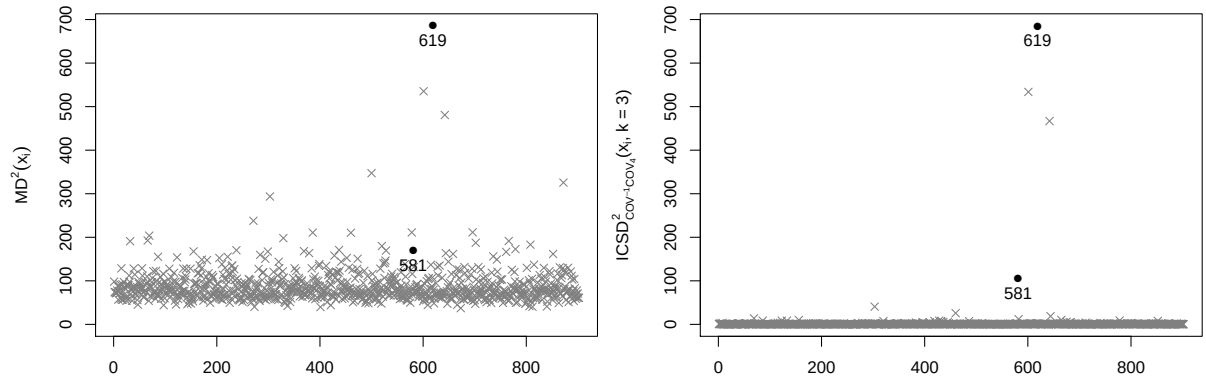

Figure 1: Représentation des distances de Mahalanobis (à gauche) et des scores d'ICS (à droite) pour l'exemple HTP.

Sur cet exemple de données réelles en provenance de l'industrie, deux observations (581 et 619) sont des produits qui ont été vendus mais qui se sont avérés défectueux. On dispose de données récoltées lors du processus de contrôle de qualité en amont de la vente des produits et on se demande si on aurait pu détecter ces deux observations comme atypiques. On voit sur la Figure 1 l'avantage d'ICS sur la distance de Mahalanobis puisqu'ICS, lorsque l'on sélectionne trois composantes, permet de détecter ces deux observations avec un taux de détection acceptable alors que la distance de Mahalanobis n'en détecte qu'une seule.

Parmi les pistes de recherche sur le sujet, nous travaillons actuellement sur l'adaptation de la méthode ICS lorsque le nombre de dimensions est grand devant le nombre d'observations.

Bibliographie

[1] Aggarwal, C. C. (2017), *Outlier Analysis*, Springer.

[2] Archimbaud, A., Nordhausen, K. and Ruiz-Gazen, A. (2016a), ICSOutlier: Outlier Detection Using Invariant Coordinate Selection, *R package version 0.2-0*.

[3] Archimbaud, A., Nordhausen, K. and Ruiz-Gazen, A. (2016b), Multivariate outlier detection with ICS, *https://arxiv.org/abs/1612.06118*.

[4] Caussinus, H., Fekri, M., Hakam, S. and Ruiz-Gazen, A. (2003), A monitoring display of Multivariate Outliers, *Computational Statististics and Data Analysis*, 44(1-2), 237–252.

[5] 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.

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

[7] Hodge, V. J., et Austin, J. (2004), A survey of outlier detection methodologies, *Artificial Intelligence Review*, 22(2), 85–126.

[8] Maronna, R. A., Martin, D. et Yohai, V. (2006). *Robust statistics*. John Wiley & Sons, Chichester.

[9] 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.

[10] Ruiz-Gazen, A. (2012), Robust statistics: a functional approach, *Annals of Institut de Statistiques de l'Université de Paris*, 56(2-3), 49-64.

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