ICS et sous-espace de Fisher : généralisation à plus de deux groupes

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

[1] *Toulouse School of Economics, Université de Toulouse Capitole, France, colombe.becquart@tse-fr.eu, anne.ruiz-gazen@tse-fr.eu*
[2] *TBS Business School, France, a.archimbaud@tbs-education.fr*
[3] *Department of Mathematics and Statistics, University of Jyväskylä, Finland, klaus.k.nordhausen@jyu.fi*

Résumé. L'analyse de la structure multidimensionnelle des données telle que l'identification de clusters est souvent rendue complexe par le nombre de dimensions à analyser. Lorsque cette structure est contenue dans un sous-espace, il est ainsi bénéfique de réduire la dimension pour se placer dans ce sous-espace d'intérêt. L'analyse en composantes principales (ACP) est la méthode de référence pour la réduction de dimension. Dans cet article, nous étudions une méthode alternative à l'ACP, appelée "Invariant Coordinate Selection" (ICS). Contrairement à l'ACP, ICS ne repose pas sur la maximisation de la variance mais sur la maximisation/minimisation d'un kurtosis généralisé, et n'est pas uniquement invariante par transformation orthogonale des données mais par toute transformation affine. Plus précisément, ICS consiste à comparer deux matrices de dispersion au travers de leur diagonalisation jointe. La réduction de dimension est obtenue en projetant les données sur les vecteurs propres associés aux plus grandes et plus petites valeurs propres d'ICS. Des travaux empiriques ont montré l'efficacité de la méthode dans le cadre du clustering et de la détection d'anomalies. Certaines propriétés théoriques d'ICS sont aussi connues. En particulier, pour un mélange de distributions elliptiques et sous certaines conditions, une sélection des composantes d'ICS permet de retrouver le sous-espace discriminant de Fisher, quel que soit le choix des matrices de dispersion. Toutefois, les conditions sous lesquelles ce résultat général s'applique ne sont explicites que pour des cas particuliers, tels que des mélanges de deux groupes de même matrice de covariance. L'objectif de cet article est d'explorer plus avant le comportement théorique d'ICS dans le cadre d'un mélange de lois gaussiennes de même matrice de covariance pour un nombre de groupes quelconque k. Les matrices de dispersion considérées sont la matrice de covariance et la matrice basée sur les moments d'ordre 4. Dans ce contexte de $k \geq 2$ groupes, nous étudions les conditions sous lesquelles ICS "fonctionne", c-à.-d. sous lesquelles les composantes associées aux $k-1$ plus grandes et plus petites valeurs propres engendrent le sous-espace de Fisher qui contient les moyennes des groupes. A partir de calculs théoriques et numériques, nous montrons que pour des groupes suffisamment séparés, ces conditions s'expriment essentiellement en fonction des proportions de chaque groupe et que les valeurs des moyennes des groupes ont peu d'influence.

Mots-clés. Clustering, diagonalisation jointe, modèles de mélange, réduction de dimension.

Abstract. Analyzing multidimensional data structures, such as identifying clusters, is frequently complexified by the number of dimensions to be analyzed. When the data structure is contained within a subspace, it is therefore beneficial to reduce the dimension to the subspace of interest. Principal component analysis (PCA) is the reference method. In this article, an alternative method to PCA is studied, it is called "Invariant Coordinate Selection" (ICS). Unlike PCA, ICS is not based on variance maximization but on the maximization/minimization of a generalized kurtosis, and it is invariant not only to orthogonal data transformations but to any affine transformation. Specifically, ICS compares two scatter matrices through their joint diagonalization. Dimension reduction is achieved by projecting the data onto the eigenvectors associated with the largest and smallest eigenvalues obtained with ICS. Empirical work has shown the relevance of applying ICS to clustering and anomaly detection. Some theoretical properties of ICS are also known. In particular, for a mixture of elliptical distributions and under certain conditions, a selection of invariant components recovers the Fisher's linear discriminant subspace for any scatter matrices. However, the conditions under which this general result holds are not always explicit. They are derived only for some special cases, such as mixtures of two groups with the same covariance matrix. The purpose of this article is to further explore the theoretical behavior of ICS in the context of a mixture of Gaussian distributions with the same covariance matrix for any number of groups k. The scatter matrices under consideration are the covariance matrix and the scatter matrix of fourth moments. In this context of $k \geq 2$ groups, we study the conditions under which ICS "works", i.e. under which the components associated with the $k-1$ largest and smallest eigenvalues span the Fisher's linear discriminant subspace containing the group means. Using theoretical and numerical calculations, we show that for sufficiently separate groups, these conditions are mainly related to the group proportions and that the group means has only a minor influence.

Keywords. Clustering, dimension reduction, joint diagonalization, mixture models.

1 Introduction

L'analyse multidimensionnelle vise à explorer les relations complexes entre plusieurs variables d'un ensemble de données. Dans ce contexte, des méthodes statistiques telles que la classification non supervisée et la détection d'anomalies sont utiles dans une multitude de domaines tels que l'industrie, le marketing ou l'économie. Cependant, lorsque la dimension des données est grande (en terme de nombre de variables) comparée à la dimension de la structure d'intérêt, l'identification de cette structure est rendue plus difficile. C'est le cas par exemple si le nombre de groupes en classification non supervisée est petit comparé au nombre de variables. L'espace contenant les centres de groupes est alors de dimension inférieure à celle de l'espace où se trouvent les données. Dans de tels cas, la réduction de la dimension peut s'avérer bénéfique pour se concentrer sur le sous-espace contenant la structure des données.

Invariant Coordinate Selection (ICS) est une méthode de réduction de dimension intro-

duite dans l'article de Tyler et al. (2009). L'article de Caussinus et Ruiz (1990) sur l'Analyse en Composantes Principales Généralisée peut être considéré comme un travail précurseur d'ICS. ICS se rapproche de l'analyse en composantes principales (ACP) dans la mesure où il s'agit d'une transformation ayant pour objectif de réduire la dimension tout en préservant au maximum la structure des données. A la différence de l'ACP, ICS ne se base pas sur la maximisation de la variance, mais sur l'optimisation d'un kurtosis généralisé à travers la diagonalisation jointe de deux matrices de dispersion. ICS se distingue également de l'ACP par les deux propriétés suivantes. Premièrement, les composantes d'ICS sont invariantes (au signe et permutation près) pour toute transformation affine, contrairement aux composantes de l'ACP qui ne sont invariantes que pour des transformations orthogonales. Deuxièmement, cette méthode permet de retrouver le sous-espace discriminant de Fisher lorsque les données suivent un mélange de distributions elliptiques et sous certaines conditions qui ne sont pas faciles à expliciter en dehors de cas particuliers tels qu'un mélange de deux lois gaussiennes (Tyler et al., 2009).

Dans le cadre de la détection d'anomalies, ICS a été étudié sur un plan théorique et empirique par Archimbaud, Nordhausen, et Ruiz-Gazen (2018). Dans le cadre de la classification non-supervisée, les propriétés théoriques d'ICS sont connues dans le cas de deux groupes (Tyler et al., 2009; voir aussi Peña et Prieto, 2001). Pour plus de deux groupes, Alfons et al. (2024) proposent une étude empirique.

Après avoir rappelé le principe d'ICS dans la section 2.1, nous décrivons dans la section 2.2 le comportement bien connu de la méthode dans le cas d'un mélange de deux gaussiennes. L'objectif de notre présentation est d'étudier théoriquement et numériquement le comportement d'ICS lorsqu'il y a plus que deux groupes. L'étude se concentre sur un mélange de lois gaussiennes de même matrice de covariance et sur ICS avec la matrice de covariance et la matrice basée sur les moments d'ordre 4. Cette étude est présentée de manière théorique dans la section 3 et de manière empirique dans la section 4.

2 Invariant Coordinate Selection

Soit Y un vecteur aléatoire de dimension p, et de fonction de répartition F_Y. Une matrice de dispersion de Y, notée $V(F_Y)$, est une matrice symétrique définie positive de dimension $p \times p$ et affine équivariante, c-à.-d $V(F_{AY+b}) = AV(F_Y)A'$ pour toute matrice A non singulière de dimension $p \times p$ et pour tout $b \in \mathbb{R}^p$. On note $\mathcal{P}_p$ l'ensemble de toutes les matrices symétriques définies positives d'ordre p et on simplifie l'écriture de $V(F_Y)$ en V lorsque le contexte est clair.

2.1 Principe général d'ICS

Le principe d'ICS est de comparer deux matrices de dispersion affines équivariantes, notées V_1 et V_2. Si la distribution des données est elliptique, alors toutes les matrices de dispersion affines équivariantes sont proportionnelles. Dans le cas contraire, V_1 et V_2 n'ont pas la même forme et sont donc différentes sur certaines dimensions. Identifier les directions dans lesquelles

elles diffèrent permet de retrouver les déviations par rapport à une distribution elliptique, et donc de révéler la structure des données. Pour effectuer cette comparaison, ICS repose sur la diagonalisation jointe de V_1 et V_2 :

$$H'V_1H = D_1$$
$$H'V_2H = D_2$$

où $V_1, V_2 \in \mathcal{P}_p$, D_1 et D_2 sont des matrices diagonales telles que $D_1^{-1}D_2 = diag(\rho_1, \ldots, \rho_p)$, $\rho_1, \ldots, \rho_p$ étant les valeurs propres de $V_1^{-1}V_2$ triées dans un ordre décroissant, et $H = (h_1, \ldots, h_p)$ est une matrice contenant les vecteurs propres correspondant.

La transformation $Z = H'Y$ présente deux propriétés intéressantes. Premièrement, les variables obtenues sont invariantes sous transformation affine (voir théorèmes 1 et 2 de Tyler et al., 2009). Deuxièmement, elle permet de retrouver le sous-espace discriminant de Fisher lorsque les données suivent un mélange de distributions elliptiques et sous certaines conditions (voir théorèmes 3 et 4 de Tyler et al., 2009). En particulier, Tyler et al. (2009) s'intéressent au cas où le mélange est formé de distributions de même matrice de dispersion, mais dont les centres et/ou fonctions de densité peuvent différer. Les centres sont de dimension p et on suppose qu'ils génèrent un sous-espace de dimension q telle que $0 < q < p$. Le théorème 4 de Tyler et al. (2009) nous apprend que dans ce cas il y a au moins une valeur propre issue de la diagonalisation jointe des deux matrices de dispersion qui est de multiplicité supérieure ou égale à $p - q$. Lorsqu'une valeur propre a une multiplicité égale à $p - q$, et non supérieure, le sous-espace engendré par les vecteurs propres associés aux autres valeurs propres qui sont les plus grandes ou les plus petites, est le sous-espace discriminant de Fisher. Pour réduire la dimension, il suffit alors de projeter les données sur les vecteurs propres associés aux plus grandes et plus petites valeurs propres d'ICS.

2.2 Cas d'un mélange de 2 groupes

Lorsque la distribution est un mélange de deux distributions gaussiennes, le cas où la multiplicité est supérieure à $p - q$ est celui où ICS ne fonctionne pas : toutes les valeurs propres sont égales. En prenant comme matrice de dispersion la matrice de covariance et la matrice de dispersion basée sur les quatrièmes moments, il est montré que le cas décrit précédemment se produit lorsque la proportion d'un groupe est de $(3-\sqrt{3})/6$, soit environ 21% (Tyler et al., 2009). Si la proportion d'un groupe est inférieure à ce seuil, il y a une valeur propre strictement supérieure aux autres qui sont égales entre elles. Inversement, si les deux groupes ont une proportion supérieure à ce seuil, il y aura une valeur propre strictement inférieure aux autres qui seront égales entre elles. Le cas d'un mélange de deux distributions gaussiennes a été approfondi par Archimbaud, Nordhausen, et Ruiz-Gazen (2018) qui recommandent dans un contexte de détection d'anomalies d'utiliser la matrice de covariance, notée COV, pour V_1 et la matrice basée sur les moments d'ordre 4, notée COV_4, pour V_2 :

$$COV(Y) = \mathbb{E}[(Y - \mathbb{E}(Y))(Y - \mathbb{E}(Y))']$$

$$\mathrm{COV}_4(Y) = \frac{1}{p+2}\mathbb{E}[d^2(Y - \mathbb{E}(Y))(Y - \mathbb{E}(Y))']$$

où d^2 est la distance de Mahalanobis au carré :

$$d^2 = (Y - \mathbb{E}(Y))' COV^{-1}(Y)(Y - \mathbb{E}(Y))$$

Avec cette paire, les auteurs démontrent les bonnes performances d'ICS dans un contexte de détection d'anomalies où l'on considère qu'une large proportion des données suit une loi gaussienne tandis qu'une plus faible proportion (les données atypiques) suit une loi gaussienne de moyenne différente. Grâce aux propriétés d'ICS, il suffit dans ce cas de sélectionner la première composante. De plus, les cas où ICS ne fonctionne pas sont assez rares puisqu'ils correspondent uniquement au cas où le groupe des atypiques est en proportion d'environ 21%. Cependant, limiter l'étude théorique à seulement deux groupes est restrictif et nous cherchons à savoir si les résultats obtenus sont généralisables à k groupes. La suite de l'article examine le comportement d'ICS dans un contexte de mélange de $k \geq 2$ distributions gaussiennes, en utilisant COV et COV_4.

3 Généralisation à k groupes

Nous cherchons à obtenir des résultats théoriques lorsqu'il y a $k \geq 2$ groupes, en se concentrant sur un mélange de distributions gaussiennes de même matrice de covariance. Les matrices de dispersion utilisées sont COV et COV_4. Comme dans la section précédente, l'objectif est de trouver les conditions pour lesquelles ICS "fonctionne", c'est-à-dire les cas où toutes les valeurs propres ne sont pas égales. Pour cela, on réécrit le modèle de manière à expliciter la forme de $COV^{-1}\mathrm{COV}_4$ et les valeurs propres associées. Devant les défis posés par les calculs théoriques, nous avons choisi d'utiliser des approximations numériques pour certaines étapes et de visualiser notamment les approximations numériques des valeurs propres.

3.1 Simplification du modèle

Cette sous-section reprend la transformation introduite par Tyler et al. (2009) dans la preuve de son théorème 4, afin de simplifier le modèle sans perte de généralité. Y est un vecteur aléatoire suivant un mélange de k distributions gaussiennes de même matrice de covariance :

$$Y \sim \sum_{j=1}^{k} \alpha_j N_p(\mu_j, \Gamma)$$

où $\alpha_j > 0$ pour $j = 1, \ldots, k$ et $\sum_{j=1}^{k} \alpha_j = 1$, $\mu_j \in \mathbb{R}_p$ pour $j = 1, \ldots, k$, et $\Gamma \in \mathcal{P}_p$.

Soit $M = (\mu_1, ..., \mu_k)$ la matrice contenant les centres de Y. On pose $M_0 = \Gamma^{-\frac{1}{2}}(M - \mu_k \mathbf{1}'_k)$ avec $\mathbf{1}_k$ un vecteur de 1 de dimension k. On note q le rang de la matrice M_0. La décomposition QR de M_0 est :

$$M_0 = PT = P \begin{pmatrix} T_u & 0 \\ 0 & 0 \end{pmatrix}$$

où P est une matrice orthogonale, $T = [t_1, \ldots, t_k]$ et T_u est une matrice triangulaire supérieure de dimension $k - 1 \geq 1$ telle que les $k - 1 - q \geq 0$ dernières lignes sont nulles.
Soit $X = P'\Gamma^{-\frac{1}{2}}(Y - \mu_k \mathbf{1}'_k)$. X est un mélange de k distributions normales de centres $t_1, \ldots, t_k$ et de matrice de covariance l'identité, avec des poids $\alpha_1, \ldots, \alpha_k$. Grâce à la propriété d'invariance affine d'ICS, on considère sans perte de généralité le modèle suivant :

$$X \sim \sum_{j=1}^{k} \alpha_j N_p(t_j, I_p)$$

3.2 Calcul des valeurs propres

A partir du modèle précédent, il est possible d'écrire les matrices COV^{-1} et COV_4 sous les formes suivantes :

$$COV(X)^{-1} = \begin{bmatrix} B & 0 \\ 0 & I_{p-k+1} \end{bmatrix}, \qquad COV_4(X) = \begin{bmatrix} A & 0 \\ 0 & I_{p-k+1} \end{bmatrix},$$

où $A = [a_{ij}]$ et $B = [b_{ij}]$ sont deux matrices $(k-1) \times (k-1)$.

Nous n'explicitons pas les termes de la matrices B qui dans la suite seront calculés numériquement. Pour la matrice A, nous avons:

$$\begin{aligned} a_{mn} &= \frac{1}{p+2}\mathbb{E}\left[\sum_{i=1}^{k-1}\sum_{j=1}^{k-1} x_m^{\mathrm{c}} x_n^{\mathrm{c}} x_i^{\mathrm{c}} x_j^{\mathrm{c}} b_{ij} + x_m^{\mathrm{c}} x_n^{\mathrm{c}} \sum_{i=k}^{p} (x_i^{\mathrm{c}})^2\right] \\ &= \frac{1}{p+2}\left[\sum_{i=1}^{k-1}\sum_{j=1}^{k-1} b_{ij}\mathbb{E}[x_m^{\mathrm{c}} x_n^{\mathrm{c}} x_i^{\mathrm{c}} x_j^{\mathrm{c}}] + (p-k+1)\mathbb{E}[x_m^{\mathrm{c}} x_n^{\mathrm{c}}]\right] \end{aligned}$$

pour $m, n = 1, \ldots, k-1$ et avec $(x_1^{\mathrm{c}}, \ldots, x_p^{\mathrm{c}})^T = X - \mathbb{E}(X)$. Les moments dans l'expression des termes a_{mn} sont calculables explicitement mais pas détaillés ici.

Le produit des deux matrices de dispersion donne :

$$COV^{-1}\mathrm{COV}_4 = \begin{bmatrix} BA & 0 \\ 0 & I_{p-k+1} \end{bmatrix}$$

et nous cherchons les valeurs propres de ce produit. Du fait de la structure du produit, il y a au moins $p - k + 1$ valeurs propres égales à 1. Le calcul des autres valeurs propres conduit à des expressions compliquées, fonctions des termes b_{ij} de la matrice B. Pour les calculer, nous avons recours à des calculs numériques notamment pour l'inverse de la matrice de covariance ainsi que pour les valeurs propres et vecteurs propres de $COV^{-1}\mathrm{COV}_4$.

4 Résultats empiriques et conclusion

L'objectif des calculs numériques est de comprendre le comportement des valeurs propres de $COV^{-1}\text{COV}_4$ pour des mélanges de lois gaussiennes de plus de deux groupes en fonction de la configuration des groupes. Avec la simplification du modèle présentée dans la section 3.1, les matrices de covariances sont égales à l'identité et la configuration des groupes dans le mélange dépend des proportions des différents groupes et de leurs moyennes. L'un des objectifs est de détecter les cas où toutes les valeurs propres sont égales, c'est-à-dire les cas pour lesquels ICS ne fonctionne pas. Dans notre cas, avec les matrices de dispersion choisies, les valeurs propres égales valent 1. On rappelle que dans le cas de deux groupes, le fait qu'ICS ne fonctionne pas ne dépend que de la proportion des groupes et pas des moyennes des groupes.

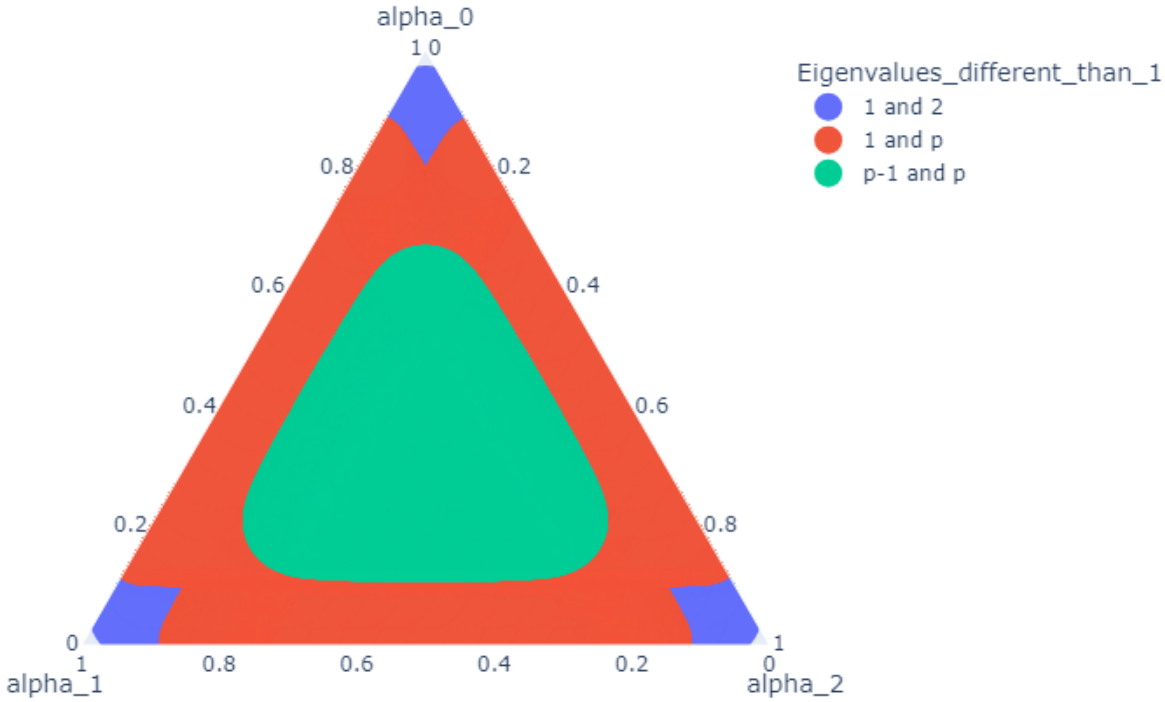

Figure 1: Diagramme ternaire représentant les valeurs propres de $COV^{-1}\text{COV}_4$ selon les proportions des 3 groupes notées alpha_0, alpha_1 et alpha_2. Les moyennes de groupes utilisées pour la génération de ce graphique sont $t_1 = (200, 0, 0, \ldots, 0)$, $t_2 = (400, 100, 0, \ldots, 0)$, $t_3 = (0, 0, \ldots, 0)$, et $p = 13$.

Dans un premier temps, nous étudions des scénarios où les groupes sont bien séparés. Une fois les moyennes de groupes choisies, une grille de proportions de groupes est générée et, pour chaque combinaison, les valeurs propres de $COV^{-1}\text{COV}_4$ sont calculées. Nous avons

répété ce protocole en faisant varier le nombre de groupes de 3 à 11. La Figure 1 illustre les résultats que nous avons obtenus avec 3 groupes et $q = 2$. On rappelle que les valeurs propres sont triées dans un ordre décroissant. Lorsqu'il y a deux groupes de faible proportion, il y a deux valeurs propres supérieures à 1 (zone bleue sur la Figure 1). Dans le cas où il n'y a qu'un groupe de faible proportion, il y a une valeur propre supérieure à 1 et une inférieure à 1 (zone rouge sur la Figure 1). Enfin, dans le cas où il n'y a aucun groupe de faible proportion, les deux valeurs propres différentes de 1 sont inférieures à 1 (zone verte sur la Figure 1). Même s'il n'est pas facile de déterminer précisément le seuil à dépasser pour considérer qu'un groupe à une "faible" proportion, on peut dire que ce seuil est strictement inférieur à 21% et de l'ordre de 18% pour 3 groupes. De plus, même si le graphique ne montre que des cas où il y a 2 valeurs propres différentes de 1, il est possible que les frontières entre les zones bleue, rouge et verte correspondent au cas où ICS ne fonctionne pas. De manière intéressante, nous avons obtenu des résultats très similaires en faisant varier les moyennes de groupes mais aussi des résultats comparables quand on augmente le nombre de groupes. Ainsi, nous faisons la conjecture que, pour des groupes bien séparés, les conditions de bon fonctionnement d'ICS s'expriment d'avantage en fonction des proportions de chaque groupe qu'en fonction des valeurs des moyennes des groupes.

Pour approfondir l'analyse, nous souhaitons utiliser ces calculs numériques pour étudier ICS dans d'autres situations. D'abord, nous examinerons plus en détail les cas impliquant plus de trois groupes dans le mélange de distributions gaussiennes. Ces cas sont néanmoins plus difficile à représenter graphiquement. De plus, nous étudierons les cas où on observe de la colinéarité entre les moyennes des groupes, c-à.-d. les cas où $0 < q < k - 1$.

Bibliographie

Alfons, A., Archimbaud, A., Nordhausen, K. and Ruiz-Gazen, A. *Tandem clustering with invariant coordinate selection*. arXiv:2212.06108 [stat].

Archimbaud, A., Nordhausen, K. and Ruiz-Gazen, A. (2018). ICS for multivariate outlier detection with application to quality control. *Computational Statistics & Data Analysis*, 128, pp. 184-199.

Caussinus, H. and Ruiz, A. (1990), Interesting Projections of Multidimensional Data by Means of Generalized Principal Component Analyses. In: Momirović, K., Mildner, V. (Eds) *Compstat* (pp. 121–126). Physica-Verlag HD.

Peña, D. and Prieto, F. J. (2001). Cluster identification using projections. *Journal of the American Statistical Association*, 96(456), 1433-1445.

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