Projection pursuit algorithms to detect outliers

María Inés Stimolo, Pablo Arnaldo Ortiz

Artículo publicado en Cuadernos de Administración
Volumen 33, 2020 – ISSN 0120-3592 / e-ISSN 1900-7205

Cuadernos de
ADMINISTRACIÓN

Vol. 33
Publicación continua
2020

ISSN: 1900-7205 (En línea) | ISSN: 0120-3592 (Impreso)

Articles

Projection pursuit algorithms to detect outliers*

Algoritmos de búsqueda de proyección para detectar valores atípicos

Algoritmos de busca de projeção para detectar valores atípicos

Maria Inés Stimolo [a]
Universidad Nacional de Córdoba - Facultad de Ciencias Económicas, Argentina maria.ines.stimolo@unc.edu.ar
ORCID: http://orcid.org/0000-0001-7277-1638

Pablo Arnaldo Ortiz
Universidad Nacional de Córdoba - Facultad de Ciencias Económicas Argentina ORCID: http://orcid.org/0000-0002-3777-0653

DOI: https://doi.org/10.11144/Javeriana.cao33.ppado

Date received: 26/08/2019
Date accepted: 20/10/2019
Date published: 20/05/2020

Abstract:

In this paper, we compare the methods proposed by Peña and Prieto (2001), and Filzmoser, Maronna, and Werner (2008) to detect outliers in a set of Argentine companies that quote their shares in the Stock Exchange. A significant heterogeneity between observations can be a consequence of the presence of outliers. The detection of outliers is an important task for the statistical analysis since they distort descriptive measures and parameters estimators. There are different multivariate methods to detect outliers, such as distance-based methods and projection pursuit methods.

JEL Codes: C81, M29.
Keywords: outliers, projection pursuit, Kurtosis, Argentinian companies.

Resumen:

En este trabajo se comparan los métodos propuestos por Peña y Prieto (2001) y Filzmoser, Maronna y Werner (2008) para detectar datos atípicos en empresas argentinas que cotizan sus acciones en el Mercado de Valores. La heterogeneidad signifi cativa entre observaciones puede ser una consecuencia de la presencia de datos atípicos. La detección de datos atípicos es importante en el análisis estadístico por su efecto en la distorsión de las medidas descriptivas y en los estimadores de los parámetros. Existen distintos métodos multivariados para detectar datos atípicos, tales como los métodos basados en la distancia o los métodos de búsqueda de proyecciones.

Códigos JEL: C81, M29.
Palabras clave: datos atípicos, búsqueda de proyecciones, curtosis, empresas argentinas.

Resumo:

Este trabalho compara os métodos propostos por Peña e Prieto (2001), e Filzmoser, Maronna e Werner (2008) para detectar dados atípicos em empresas argentinas que cotizam suas ações no Mercado de Valores. A heterogeneidade significativa entre observações pode ser uma consequência da presença de dados atípicos. A detecção de dados atípicos é importante na análise estatística por seu efeito na distorção das medidas descritivas e nos estimadores dos parâmetros. Existem distintos métodos multivariados para detectar dados atípicos, tais como os métodos baseados na distância ou os métodos de busca de projeções.

Códigos JEL: C81, M29.
Palavras-chave: outliers, busca de projeções, curtose, empresas argentinas.

Introduction

Databases often show outliers observations, which present a different behavior from the majority. It is important to detect these observations since they affect the data analysis in different ways. In this respect, Uriel Jiménez and Aldás Manzano (2005) point out:

Author notes

[a] Corresponding author. E-mail: maria.ines.stimolo@unc.edu.ar

i. They could mask the data pattern and distort the results, so the conclusions would be completely different without their presence.
ii. They could affect the normality condition that is necessary in many multivariate techniques.

Outliers have different causes:

- Measurement errors, collection or transcription.
- Intentional errors of response from the respondents.
- Sampling errors: the incorporation of sample statistical units from different populations to the target population.
- Intrinsic heterogeneity: the observed elements belong to the target population, but the inherent variability of the samples differs from the rest in their choices, attitudes or behavior.

Sometimes the detection of outliers is the first step in statistical analysis. Other times, the outliers need to be removed or downweighted; different causes motivate different procedures. The detection of outliers depends on the type of error (or cause) in the data. In the case of errors in measurement or data entry to the base, it is relatively simple to correct them and it is convenient to eliminate the obvious mistakes. However, a controversial question is: What should we do when the outliers derive from the intrinsic heterogeneity of the data?

In this paper we discuss some multivariate methods for detecting outliers. In multivariate methods, there exist two approaches to identify outliers: those based on the distances of the observations in the data center and those projecting the original data. The projection pursuit methods easily identify atypical observations, and they have the advantage that it is not necessary to know the data distribution. However, the disadvantage of the projection pursuit methods is that there are high requirements in terms of computational load, which increments significantly when there is an increase in the variables considered.

The document is organized as follows. The first section describes two algorithms used to detect outliers (Filzmoser et al., 2008; Peña & Prieto, 2001) based on projection pursuit. The second section compares the methods developed applying them in a group of Argentine companies that listed their shares publicly in the period 2004-2012. In final section, we developed the main conclusions, and we describe research limitations and future research work related to this topic.

Algorithms' Description

Mahalanobis' distance from the center of the data is the classical multivariate way of identifying outliers observations far from most others.

Let $\mathbf{x}_1$, $\mathbf{x}_2$...$\mathbf{x}_n$ be a random sample from a normal multivariate distribution $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ where μ is the multivariate location vector and Σ the p x p covariance matrix. The distance between the i-th observation x_i and the location μ, weighted by the covariance Σ is using to detect if the observation x_i is an outlier, it is (1).

$$D_i(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \left[(\mathbf{x}_i - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right]^{1/2} \quad (1)$$

The square Mahalanobis distance D^2_i has Chi-square distribution with p degrees of freedom and the observation i is consider outlier if $D_i^2 > \chi^2_{p,0.95}$ (by setting the squared Mahalanobis distance equal to

certain quantile of Chi-squared distribution it is possible define ellipsoids having the same Mahalanobis distance from the data centre).

When both μ and Σ are unknown it be used estimators, *i.e.* the vector mean $\bar{\mathbf{x}}$ and sample covariance matrix **S**, to estimate the Mahalanobis distance $(v_i^{1/2})$,see (2):

$$\mathbf{v}_i = (\mathbf{x}_i - \bar{\mathbf{x}})\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}) \sim \chi_p^2 \tag{2}$$

The vector mean and the covariance matrix are affected by outliers, besides the Mahalanobis distance relies on the assumption of normality. Therefore, it is affected by outliers and it does not allow identify sets of outliers (Peña, 2002).

An alternative approach is to use robust location and scale estimators, measures resistance against the influence of outlying observations. Maronna (1976) studied affinely equivariant M-estimators for covariance matrices, and Campbell (1980) proposed using the Mahalanobis distance computed using M-estimators for the mean and covariance matrix.

Nevertheless, the distance method approach presents two difficulties: (i) obtaining a reliable robust location estimator, and (ii) determining and classifying the outliers. It is important to find metric separating outliers from regular observations. Rousseeuw (1985) proposed other distance-based algorithm that computes the ellipsoid with the smallest volume or with the smallest covariance determinant that would include at least half of the data points (minimum covariance determinant, MCD). Because these procedures are based on the minimization of certain nonconvex and nondifferentiable criteria, these estimators are computed by resampling.

Rousseeuw and Driessen (1999) get faster algorithm splitting the problem into smaller subproblems (FAST-MCD algorithm).

Others outlier-detection procedures are basing on projections to identify outliers. The underlying motive of these methods is to find suitable projection of the data in which the outliers are readily apparent and can thus be downweighted to yield a robust estimator.

Gnanadesikan and Kettenring (1972) proposed to search for outliers in the direction of the first principal components: the direction of the maximum variability of the data. Although this method provides a correct solution when the outliers are located close to the directions of the principal components, it may fail to identify outliers in the general case.

From that point of view, Stahel (1981) carry on projection pursuit on the data using random directions. They proposed to compute the weight for the robust estimators from the projections of the data onto some directions. These directions were chosen maximizing distances based on robust location and scale estimators, and the optimal values for the distances could also be used to weight each point in the computation of the robust covariance matrix. Stahel (1981) developed a computer approximation based on direction from random subsamples.

Rousseeuw (1993) proposed selecting . observations from the original sample and computing the orthogonal direction of the hyperplane defined by these observations. The maximum over this finite set of directions is used as an approximation to the exact solution.

The disadvantage of the projection pursuit methods is the form to increase the computational burden with the variable number.

Peña and Prieto (2001) proposed an improve examining only the set of 2p directions that maximize or minimize the kurtosis. A small number of outliers would cause heavy tails and lead to a larger kurtosis

coefficient while a large number of outliers would start introducing bimodality and decrease the kurtosis coefficient (Filzmoser et al., 2008).

Projection pursuit methods have a computational time that increase very rapidly in higher dimensions.

Principal components are those orthogonality directions that maximize the variance along each component. It is well-known the method of dimension reduction that seems intuitive to identifying outliers since outliers increase the variance along their respective directions. The outliers appear more visible in principal components space, at least in some direction of maximum variance, than the original data space. Principal components select a small quantity of highly informative components, discarding those are not contribute significant additional information. In this way, the dataset become more computationally tractable without losing a lot of information. Filzmoser et al. (2008) proposed a method based on the principal components properties useful to detect outliers in high dimensions.

In the following sections, we describe the Peña and Prieto (2001) and Filzmoser et al. (2008) methods.

Kurtosis method (Kurt) (Peña & Prieto, 2001)

Given a sample (x_1,....x_n) of a *p*-dimensional random variable *X,* the algorithm consists in projecting each observation onto a set of *2p* directions, which are obtained as the solutions of *2p* simple smooth optimization problems.

1) The original data are rescaled and centred, see (3).

$$z_i = S^{-1/2}(x_i - \bar{x}) \quad i = 1, 2, \ldots, n \tag{3}$$

Where $\bar{x}$ and S_x are the mean and sample variance, respectively.

2) Set $z_i^{(1)} = z_i$, the iteration index *j* compute *p* orthogonal directions that maximize the kurtosis coefficient obtained as a solution of the following optimization problem (4).

$$d_j = \arg\max_d \frac{1}{n}\sum_{i=1}^{n}(d' z_i^{(j)})^4 \quad s.t. \quad d'd = 1 \tag{4}$$

3) Project sample points onto a lower dimension subspace *(p – j)*, orthogonal to the direction d_j. Define (5):

$$v_j = d_j - e_j, \quad Q_j = \begin{cases} I - \dfrac{v_j v_j'}{v_j' d_j} & if \quad v_j' d_j \neq 0 \\ I & otherwise \end{cases} \tag{5}$$

Where e_j denotes the first unit vector, **I** the identity matrix and Q_j is orthogonal. Compute the new values in (6),

$$u_i^{(j)} = \begin{pmatrix} y_i^{(j)} \\ z_i^{(j+1)} \end{pmatrix} = Q_j z_i^{(j)} \quad i = 1, 2, \ldots, n. \tag{6}$$

Where $y_i^{(j)}$ is the first component of $u_i^{(j)}$ which satisfies $y_i^{(j)} = d'_j z_i^{(j)}$ (the univariate projection values) and $y_i^{(j+1)}$ corresponds to the remaining p-j components of $u_i^{(j)}$.

4) Compute $j' = j+1$ and repeat *(2)* y *(3)* up to have p directions: $d_1, d_2, \ldots, d_p$.

5) Repeat (2) and (3) computing p orthogonal directions that minimize the kurtosis coefficient (idem step 2) obtaining $d_{p+1}, d_{p+2}, d_{p+1}, \ldots, d_{2p}$.

6) To determine an outlier in any one of the $2p$ directions ($y_i^{(j)} = d'_j z_i^{(j)}$), we compute a univariate measure rescaling with the median (*med*) and the median absolute deviation (MAD), see (7).

$$r_i = \frac{\left| y_i^{(j)} - med(y^{(j)}) \right|}{MAD(y^{(j)})} > \beta_p \tag{7}$$

Where β is a cut-off chosen to ensure a reasonable level of Type I error and depend on the sample space dimension p. See Table 1.

TABLE 1

Cut-off Values for Univariate Projections			
Sample space dimension p	5.0	10	20
Cut-off value β_p	4.1	6	10

Source: Own elaboration.

7) Define a new sample composed of all observations i if $r_i \prec \beta_p$, and the procedure is applied again to the reduced sample. This is repeated until either no additional observations $r_i \succ \beta_p$ or the number or satisfy remaining would be less than $[(n+p+1)/2]$.

8) Let U denote the set of all observations not labelled as outliers and computed the mean vector $\bar{m}$, the covariance matrix $\bar{s}$; and Mahalanobis distance: $v_i = (x_i - \bar{m})' \bar{S} (x_i - \bar{m})$.

Those observations $i \notin U$ such that $v_i \prec \chi^2_{p-1,0.99}$ are considered not be outliers and included in U. The procedure is repeated until U becomes the set of all observations.

The Kurtosis method is affine equivariant. Peña and Prieto (2001) conclude after several computational experiments to study the practical behaviour of the proposed procedure, that it shows a satisfactory empirical performance, especially for large sample space dimensions and concentrated contaminations.

However other authors discussed the Kurtosis method, they argue important points. The method works well in the presence of scattered outliers or multiple clusters of outliers.

For those cases in which the shape of the contamination is similar to that of the original data, the method can be supplement with other general methods (an alternative approach is to use clustering methods to supplement the general-purpose robust methods). The greatest chance of success comes from use of multiple methods, at least one of which is a general-purpose method such as FAST-MCD and MULTOUT, and at least one of which is meant for clustered outliers, such as Kurt method.

However, several key aspects of the Kurt algorithm proposed are criticized. The standardization in Step 1 uses the classical mean and covariance matrix. It is well known that these estimators are extremely sensitive to outliers, which often leads to labelling outliers as good data points and good points as outliers. The authors answer:

> This problem cannot appear in our method. First, note that the algorithm we propose is affine equivariant, independently of the initial standardization. The kurtosis coefficient is invariant to translations and scaling of the data and a rotation will not affect the maximizes or minimisers. Moreover, we have tried to be careful when defining the operations to generate the successive directions, as well as in the choice of an initial direction for the optimization problems. (Peña & Prieto, 2001, p. 307)

Maximizing and minimizing the kurtosis in Steps 2 and 3. The authors indicates that the kurtosis is maximal (respectively, minimal) in the direction of the outliers when the contamination is concentrated and small (respectively, large). However, this is not always true for an intermediate level of contamination. The authors answer:

> The behaviour of the kurtosis coefficient is particularly useful to reveal the presence of outliers in the cases of small and large contaminations, and this agrees with the standard interpretation of the kurtosis coefficient as measuring both the presence of outliers and the bimodality of the distribution. What is remarkable is that in intermediate cases with α=0.3 the procedure does not break down completely and its performance improves with the sample size. (Peña & Prieto, 2001, p. 307)

Taking $2p$ orthogonal directions in Steps 2 and 3, the chosen directions are still rather arbitrary. Which is the reason to consider the only first p directions that maximize the kurtosis and then p directions that minimize the kurtosis? Why it not proposed to alternate between directions using a procedure that stops once a significant direction is computed? The authors argue "The algorithm we describe does not make use of this feature, and in this sense it is a simpler one to describe and understand, although it may be more expensive to implement" (Peña & Prieto, 2001, p. 307).

Regarding the choice of orthogonal directions, they reply:

> Our motivation to use these orthogonal directions is twofold. On the one hand, we wish the algorithm to be able to identify contamination patterns that have more than one cluster of outliers. The second motivation arises from a property of the kurtosis that implies that in some cases the directions of interest are those orthogonal to the maximization or minimization directions. (Peña & Prieto, 2001, p. 307)

The authors did not explain how they obtained the cut-off values β_p in Table 2 to choice in Step 7. The response:

> The results are unfortunately not totally satisfactory; the reason is the large variability in these values in the simulations[1]. This variability has two main effects –it is difficult to find correct values (huge numbers of replications would be required) and for any set of 100 replications there is a high probability that the resulting values will be far from the expected one. Nevertheless, we agree that these values could be estimated with greater detail, although they do not seem to be very significant for the behaviour of the algorithm, except for contaminations very close to the original sample. (Peña & Prieto, 2001, p. 308)

Sequential determination of outliers in Step 8, the mean and covariance matrix of the good data points are computed and used to decide which outliers can still be reclassified as good observations. This procedure is repeated until no more outliers can be reallocated. It has suggested that Step 8 be applied only once. The authors are a little surprised by the criticism of procedures that determine the outliers sequentially. They replied:

> The statistical literature is full of examples of very successful sequential procedures and, to indicate just one, Peña and Yohai (1999) presented a sequential procedure for outlier detection in large regression problems that performs much better than other nonsequential procedures. (Peña & Prieto, 2001, p. 309)

Method PCOut (Filzmoser et al., 2008)

This algorithm was designed primarily for computational efficiency at high dimension. It consists in two steps: The first one to detect the location outliers and the second one to detect scatter outliers.

1) Rescale the data $\mathbf{X}_{(n,p)}$ using the median (*med)* and the median absolute deviation (MAD), see (8):

$$x_{ij}^{*}=\frac{x_{ij}-med(x_{1j},....,x_{nj})}{MAD(x_{1j},....,x_{nj})} \quad j=1,2,....p \tag{8}$$

Compute the covariance matrix from $\mathbf{X}^{*}$.

2) Compute the eigenvalues and eigenvectors from covariance matrix $\mathbf{X}^{*}$, a semirobust principal component decomposition, and retain only p^{*} eigenvectors whit eigenvalues that represent the 99% of the variance. The matrix of principal components is Z: Z=X*V. where **V** is the matrix of eigenvalues $p^{*}\times p^{*}$. Z is rescaled by the median and the MAD as 8), for i-*th* component, see (9):

$$z_{ij}^{*}=\frac{z_{ij}-med(z_{1j},....,z_{nj})}{MAD(z_{1j},....,z_{nj})} \quad j=1,2,....p^{*} \tag{9}$$

Z*, principal components rescaled is stored for the both phases of the algorithm.

Phase 1: Location outliers.

3) Compute a robust kurtosis weights for each component denoted by w_j in (10).

$$w_j=\left|\frac{1}{n}\sum_{i=1}^{n}\frac{(z_{ij}^{*}-med(z_{1j}^{*},....,z_{nj}^{*})^{4}}{MAD(z_{1j}^{*},....,z_{nj}^{*})^{4}}-3\right| \quad j=1,2,....p^{*} \tag{10}$$

Peña and Prieto (2001) argue in the Kurt method, both small and large values of the kurtosis coefficient can be indicated of outliers. In order to use relative weights is defined $\frac{w_j}{\sum_i w_i}$.

To classify the data between outliers and non-outliers we need to determinate a weighted norm from transformated data Z* but it has not chi quadratic distribution. Z* is similar as a robust Mahalanobis distance (RD_i) (distance from median rescaled by MAD). Therefore, the algorithm used a robust distance transform similar Maronna and Zamar (2012), that helped the empirical distances di to have the same median to the theoretical distance and bring the former somewhat closer χ_p^2. See (11):

$$d_i = RD_i \cdot \frac{\sqrt{\chi^2_{p^*,0.5}}}{med(RD_1,\dots,RD_n)} \quad i = 1,2,\dots,n \tag{11}$$

4) To assign weights to each observation and use it as a measure of outlyingness is calculated the translated bi-weight function w_{1i} (Rocke, 1996), see (12):

$$w_{1i} = \begin{cases} 0, & d_i \geq c \\ \left(1 - \left(\frac{d_i - M}{c - M}\right)^2\right)^2, & M < d_i < c \\ 1, & d_i \leq M \end{cases} \tag{12}$$

Where M is equal to 33⅓ quantil of the distances and *c*, see (13):

$$c = med(d_1,\dots,d_n) + 2{,}5 \cdot MAD(d_1,\dots,d_n) \tag{13}$$

Phase 2: Scatter outliers.

5) Use the step 2 decomposition and calculate the Euclidean norm for the data in non-weighting principal component space (equivalent to the Mahalanobis distance in the original data but faster to compute). After use the Maronna and Zamar (2012) transformation, the distances set is going to use at 6.

6) Determine the weights w_{2i} to each robust distance with the translated biweight function where c^2 is equal to $\chi^2_{p^*,0.99}$, M^2 is equal to $\chi^2_{p^*,0.25}$ and finally we calculate the final weight, see (14):

$$w_i = \frac{(w_{1i} + s)(w_{2i} + s)}{(1+s)^2} \tag{14}$$

Where typically the scaling constant s = 0.25. Outliers are then classified as points they have weight $w_i <$ 0.25. This value implies that if one of the weights equals one the other must be less than 0.0625. If $w_1 = w_2$, *x* is classified as outlier when the common value is less than 0.375.

The computational speed that is the speed t of the computer to process data[2] is an advantage of this algorithm. Using examples an simulated data Filzmoser et al. (2008) infer that PCOut is a competitive outlier detection algorithm regarding detection accuracy as well as computation time.

The comparison with methods in low dimension, using simulated data reveals that PCOut performs well at identifying outliers, with low masked outliers, although it has a higher percent non-outliers that were classified as outliers. It does particularly well for location outliers while Kurt does very poorly, however Kurt does exceptionally well for scatter outliers.

An empirical application

In this paper we applied the detection outliers methods to a sample composed by a set of Argentine companies that quote their shares in the Buenos Aires stock exchange in the period 2004-2012. The database was prepared relying on the data publicly available in the Buenos Aires stock exchange web site (Bolsar, s.f.) including only companies that presented a positive operative ordinary income defined as net sales larger that costs of sales and selling and administrative expenses, and excluding those that belong to the financial and insurance sectors. A total of 744 observations (firms per year) belonging to 111 firms were considered.

The variables used to detect outliers are the following financial reporting indicators to analyse cost-effectiveness and cost behaviour (Anderson, Banker, & Janakiraman, 2003; Banker & Byzalov, 2014).

- *Market to book value.* The ratio of indicates investors' expectations of future abnormal earnings relative to assets in place. It reflects both the magnitude and persistence of sales growth expectations.
- *Current Assets and non-current Assets*
- *Operating income.* It is equals all revenue from the property minus all reasonably necessary operating expenses.
- *Net Revenues.* A company's revenue net of discounts and returns.
- *Net Revenues annual variation coeff.* The annual change in a company's net income.
- *Selling Administrative expenses*[3]. It is the sum of direct, indirect selling expenses and administrative expenses of a company.
- *Selling Administrative expenses. Annual variation coeff.*

The Table 2 and Figure 1 summarize a preliminary descriptive analysis of original sample. The univariate statistical analysis (descriptive measures and boxplots) shows the skewness and outliers.

TABLE 2

Variable	Mean	Standard Deviation	Variation Coeff.	Median	MAD	Min	Max
Market to book value (MBV)	1.10	11.31	10.28	0.73	1.08	-283.00	69.70
Net Revenues Anual variation coeff. (NRC)	1.73	13.28	7.68	1.07	0.16	0.00	357.88
Operating income. (OI)	-0.83	18.13	-21.84	0.10	0.11	-439.55	1.00
Current Assets (CA)	19.19	1.64	0.09	19.25	1.49	12.69	23.78
Non-current Assets (NCA)	19.90	2.14	0.11	20.14	2.42	12.95	24.72
Net Revenues (NR)	19.89	1.91	0.10	20.00	1.66	10.08	24.78
Selling Admin. expenses (S&A)	17.53	3.30	0.19	17.86	1.57	0.00	22.46
Selling Admin. expenses. Anual variation coeff. (S&A_C)	1.12	0.59	0.53	1.09	0.16	0.00	9.68

Source: Own elaboration.

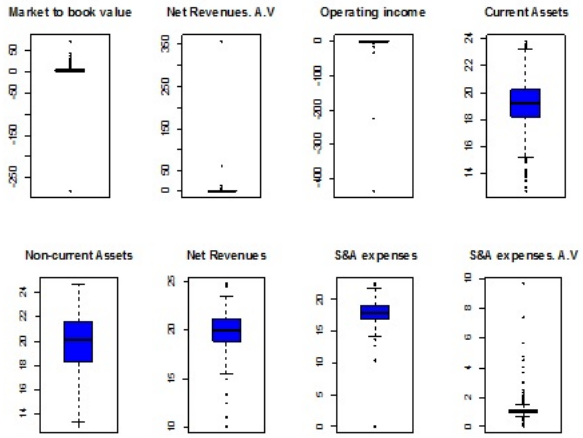

FIGURE 1
Boxplot of original variables
Source: Own elaboration.

We detected outliers using Mahalanobis distance and the both projection pursuit methods presented by Peña and Prieto (2001) and Filzmoser et al. (2008) denominated Kurt and PCOut, respectively. In this work we compare the results defining finally as outliers all the outliers defined using both methods proposed.

We used the Matlab for the Kurt method and the R package *mvoutlier* for the PCOut method (Filzmoser, 2015).

The Table 3 shows the outliers detected by differences algorithms. There were detected 48 outliers (6.5% of the data) by the distance methods (Mahalanobis). By using projection pursuit methods a large outliers were detected. 212 (28.5%) by Kurt method and 203 (27.3%) by PCOut of outliers. The algorithms proposed detected a similar quantity. Nevertheless the 69.8% of Kurt outliers were PCOut outliers (see Table 4).

TABLE 3
Outliers detected

Methods	Outliers detected
Mahalanobis	48 (6.5%)
Kurt	212 (28.5%)
PCOut	203 (27.3%)

Source: Own elaboration.

TABLE 4
Outliers detected comparing methods

		Mahalanobis		PCOut	
		Non- Outliers	Outliers	Non- Outliers	Outliers
Kurt	Non- Outliers	532		477	55
	Outliers	164	48 (6.5%)	64	148 (19.9%)
PCOut	Non- Outliers	524	17		
	Outliers	172	31 (4.2%)		

Source: Own elaboration.

The biplots (Figure 2) shows the detected outliers using different methods in the space of the first and second principal component.

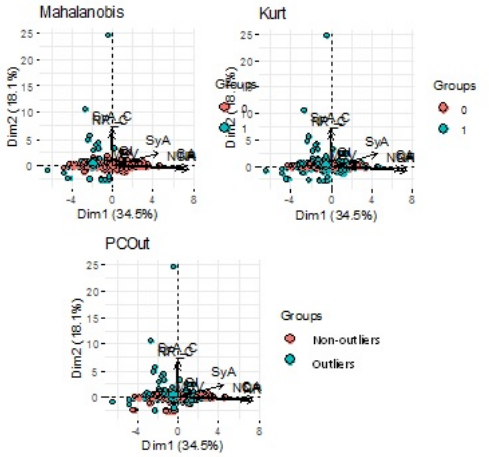

FIGURE 2
Biplot showing outliers identified by Methods
Source: Own elaboration.

For each method it has been calculated the first and second principal components of the non-outliers data and graphic it on biplots which shows different structures of the data (See figure 3).

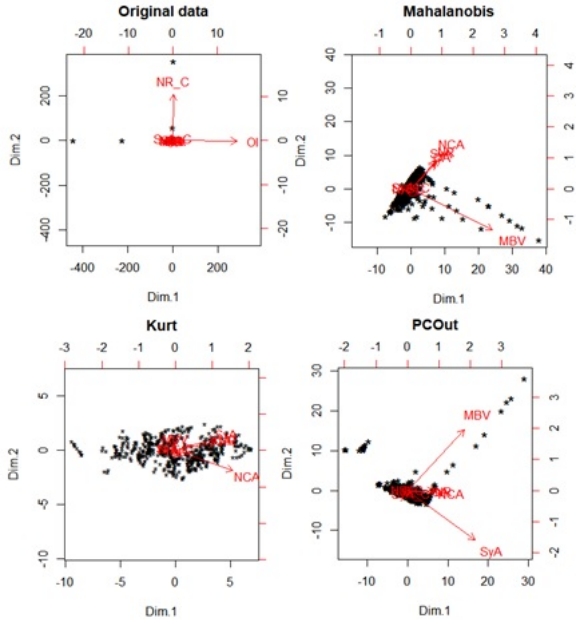

FIGURE 3
Biplot without outliers by Method
Source: Own elaboration.

Table 5 shows descriptive measures of the variables considered in non-outliers database, the mean vector difference, the multivariate variability (total and generalized variance) by method. Multivariate variances are significantly lower in all cases because of excluding outliers distorting the estimates.

TABLE 5
Multivariate descriptive without outliers

	Original	Mahalanobis	Kurt	PCOut
Sample size (n)	744	696	532	541
Mean vector				
Market to book value	1.10	1.33	0.80	1.24
Net Revenues Anual variation coeff	1.73	1.15	1.08	1.07
Operating income	-0.83	0.13	0.11	0.10
Current Assets	19.19	19.27	19.32	19.35
Non-current Assets	19.90	19.90	19.78	19.74
Net Revenues	19.89	20.00	20.16	20.18
Selling Admin. expenses	17.53	18.09	18.24	17.73
Selling Admin. expenses. Anual variation coeff.	1.12	1.10	1.09	1.06
Total variance	655	24	11	33
Generalized variance	8.4E+07	0.010	0.000	0.001

Source: Own elaboration.

A plausible criterion for determining multivariate outliers is to use different methods and consider as those who are simultaneously identified as atypical by them. Particularly, in this application 19.9% of the data were identified as outliers at Kurt and PCOut methods.

The different results and different performance of each method leads us to consider all the outliers detected for these methods. We aggregated the outliers identified by all the methods taking advantage of their performance. I this empirical application, we detected 225 outliers (30.24%). Figure 4 shows in the space of the two first principal components of the sample all the outliers detected for the algorithms.

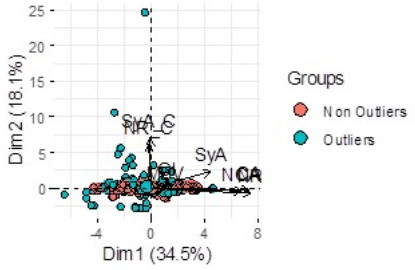

FIGURE 4
Outliers detected by all algorithms
Source: Own elaboration.

The biplot without the outliers detected by all the methods (Figure 5) shows a better performance. Besides we pointed with a circle a set of data that we could study especially because they presented a different behaviour.

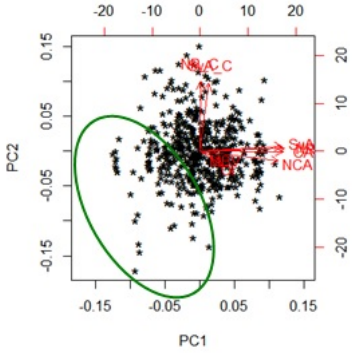

FIGURE 5
Biplot without all the Outliers detected
Source: Own elaboration.

Table 6 and Figure 6 show a descriptive analysis of data without all the Outliers detected. The variables exhibited less skewness and heterogeneity, resulting a data sample more homogeneous.

TABLE 6
Multivariate descriptive without all the Outliers detected

Variable	mean	sd	median	mad	min	max
Market to book value	0.17	1.07	0.02	0.95	-0.86	10.97
Net Revenues Anual variation coeff	0.00	0.95	-0.06	0.78	-2.77	2.91
Operating income	-0.01	0.74	-0.07	0.72	-1.81	1.93
Current Assets	0.03	0.97	0.01	0.93	-2.81	2.49
Non-current Assets	-0.20	0.87	-0.20	1.01	-2.98	1.59
Net Revenues	0.07	0.93	0.05	0.89	-2.21	2.09
Selling Admin. expenses	0.23	1.00	0.15	0.95	-2.43	2.51
Selling Admin. expenses. Anual variation coeff	**0.03**	**1.38**	**0.01**	**0.85**	**-6.79**	**7.55**
Total variance	**8.04**					
Generalized variance	**0.0022**					

Source: Own elaboration.

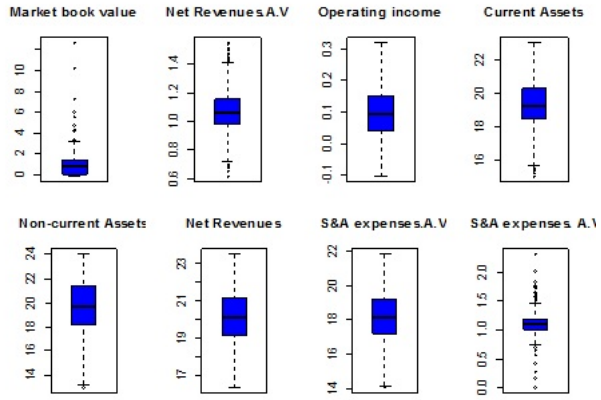

FIGURE 6
Boxplott without all the Outliers detected
Source: Own elaboration.

Conclusions

Outliers distort the results and mask the real data structure, so to detect them is an important task in the multivariate data analysis.

This work presented two pursuit algorithms to detect outliers, they are an example of the different algorithms available. But is not possible to point one algorithm as the better, it depends of the data sample and the algorithms could show similar performance. Each one method has some disadvantage, so we proposed aggregate the outliers detected by different methods (in this paper Kurt and PCOut methods) and use their different performance to improve the outliers detection. Specially the projection pursuit methods that they only search the useful projections, they are not affecting by non-normality and can be widely applied in diverse data situations.

A multivariate outliers detection is important for thorough data analysis, however, the researchers have to decide to exclude the outliers from further analysis or apply robust procedures to reduce the impact of them.

References

Anderson, M., Banker, R., & Janakiraman, S. (2003). Are selling, general, and administrative costs "sticky"? *Journal of Accounting Research, 41*(1), 47-63. https://doi.org/10.1111/1475-679X.00095

Banker, R., & Byzalov, D. (2014). Asymmetric cost behavior. *Journal of Management Accounting Research, 26*(2), 43-79. https://doi.org/10.2308/jmar-50846

Bolsar (s.f). Buenos Aires Stock Exchange. https://www.bolsar.com/VistasDL/PaginaPrincipal.aspx

Campbell, N. A. (1980). Robust procedures in multivariate analysis I: Robust covariance estimation. *Applied statistics*, 231-237. https://doi.org/10.2307/2346896

Filzmoser, P. (2015). Gschwandtner M. mvoutlier: Multivariate outlier detection based on robust methods. R package version 2.0.6.In. Routine available at http://halweb.uc3m.es/esp/Personal/personas/fjp/research.htm

Filzmoser, P., Maronna, R., & Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics & Data Analysis, 52*(3), 1694-1711. https://doi.org/10.1016/j.csda.2007.05.018

Gnanadesikan, R., & Kettenring, J. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 81-124. https://doi.org/10.2307/2528963

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 51-67. https://doi.org/10.1214/aos/1176343347

Maronna, R. A., & Zamar, R. H. (2012). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*. https://doi.org/10.1198/004017002188618509

Peña, D. (2002). *Análisis de datos multivariantes*, vol. 24. Madrid: McGraw-Hill.

Peña, D., & Prieto, F. J. (2001). Multivariate outlier detection and robust covariance matrix estimation. *Technometrics, 43*(3), 286-310. https://doi.org/10.1198/004017001316975899

Peña, D., & Yohai, V. (1999). A fast procedure for outlier diagnostics in large regression problems. *Journal of the American Statistical Association, 94*(446), 434-445. https://doi.org/10.1080/01621459.1999.10474138

Rocke, D. M. (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension. *The Annals of statistics*, 1327-1345. https://doi.org/10.1214/aos/1032526972

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical statistics and applications, 8*, 283-297. https://www.researchgate.net/profile/Peter_Rousseeuw/publication/239666038_Multivariate_Estimation_With_High_Breakdown_Point/links/0deec53137b8cc68aa000000.pdf

Rousseeuw, P. J. (1993). A resampling design for computing high-breakdown regression. *Statistics & probability letters, 18*(2), 125-128. https://doi.org/10.1016/0167-7152(93)90180-Q

Rousseeuw, P. J., & Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics, 41*(3), 212-223. https://doi.org/10.1080/00401706.1999.10485670

Stahel, W. A. (1981). *Robuste schätzungen: infinitesimale optimalität und schätzungen von kovarianzmatrizen*: Eidgenössische Technische Hochschule - ETH. Zürich.

Uriel Jiménez, E., & Aldás Manzano, J. (2005). *Análisis multivariante aplicado: aplicaciones al marketing, investigación de mercados, economía, dirección de empresas y turismo*. Madrid: Thomson.

Notes

* Research paper.

[1] In the papers and the discussion have been conducted a number of computational experiments to study the practical behavior of the proposed algorithm.

[2] It is defined as Millions of Instructions per Second –MIPS–.

[3] The Argentinean GAAP in effect at the time of this study was Resolución Técnica Nº 9 (RT9) of FACPCE. Chapter 5 of RT9 defines as Selling Expenses those related with sales and distribution of products or services rendered by the firm. RT9 Chapter

5 says that Administration Expenses are expenses incurred by the firm in order to carry on its activities but cannot be attributable to any of the following functions: purchasing (procurement), production (operations), selling, research and development, financing of goods or services. The same chapter of RT9 states that net sales (revenues) are to be presented in the income statement and the amount shall exclude returns, discounts and taxes.

Cited as: Stimolo, M. I., & Ortiz, P. A. (2020). Projection pursuit algorithms to detect outliers. *Cuadernos de Administración*, 33. https://doi.org/10.11144/Javeriana.cao33.ppado