CIÊNCIAeNATURA

UFSM

Ci. e Nat., Santa Maria v.42, e16, p. 01-14, 2019
DOI:10.5902/2179460X41662
ISSN 2179-460X

Meio Ambiente Submissão: 26/12/2019 Aprovação: 14/02/2020 Publicação: 15/05/2020

Artigo Original

A performance evaluation in multivariate outliers identification methods

Josino José Barbosa, Anderson Ribeiro Duarte, Helgem Souza Ribeiro Martins

Abstract

Methodologies for identifying multivariate outliers are extremely important in statistical analysis. Outliers may reveal relevant information to variables under investigation. Statistical applications without prior identification of possible extreme values may yield controversial results and induce mistaken decision making. In many contexts, outliers are points of great practical interest. Given this, this paper seeks to discuss methodologies for the detection of multivariate outliers through a fair and adequate comparative technique in their simulation procedure. The comparison considers detection techniques based on Mahalanobis distance, besides a methodology based on cluster analysis technique. Sensitivity, specificity, and accuracy metrics are used to measure the method quality. An analysis of the computational time required to perform the procedures is evaluated. The technique based on cluster analysis revealed a noticeable superiority over the others in detection quality and also in execution time.

Keywords: Multivariate outliers, Simulation, Cluster analysis, Accuracy, Computational time.

[I] Federal University of Ouro Preto, Ouro Preto, Brazil. Federal University of Viçosa, Viçosa, Brazil. josinojba@gmail.com

[II] Federal University of Ouro Preto, Ouro Preto, Brazil. anderson.duarte@ufop.edu.br

[III] Federal University of Ouro Preto, Ouro Preto, Brazil. Federal University of Viçosa, Viçosa, Brazil. helgem@ufop.edu.br

1 Introduction

Statistical analysis of datasets, whether small or large datasets, always requires a careful descriptive study. Such care must be taken before the use of any more sophisticated statistical techniques. In particular, the presence of atypical observations, usually referred to as outliers can greatly deteriorate statistical research.

The assessment of the outliers presence in univariate data is not completely trivial. This investigation becomes even more sophisticated when searching for multivariate outliers, ie, present in datasets with two or more variables being analyzed simultaneously.

According to Hawkins (1980), a seemingly inconsistent observation or inconsistent subgroup of information compared to the rest of dataset is defined as outliers. In the conception of Barnett and Lewis (1994), observations that arouse suspicion that its deviated from the rest of dataset is very large, may come from generation through some mechanism other than the usual nature of the data.

The concept of multivariate outlier refers to $k-$ dimensional subspaces defined by the variables involved in the investigation. According to Jolliffe (2011), some observation can configure a multivariate outlier even not being a univariate outlier when analyzing each of the variables individually.

Several applicated studies use methodologies for detecting outliers. Most of these applications come from studies with data created in productive processes. When these processes shows unusual behavior beyond usual predictability, outliers are generated (Aggarwal, 2017). This evidence illustrates a relevant importance of outliers analysis. Very relevant information about rare (but existing) characteristics around the data under investigation is present and impact the data generation processes.

Establishing strategies for detecting these unusual characteristics leads to a wide range of applications of practical interest. Aggarwal (2017) illustrates several applications as: financial system fraud checks and credit operations; procedures for medical diagnostics; event detection sensors; among many others.

Many studies with proposed methodology to detect multivariate outliers have been developed, some of them based on the identification through the Mahalanobis distance. Rousseeuw and Zomeren (1990) present a method based on the robust estimator MVE (*Minimum volume ellipsoid*). Atkinson and Riani (2002) present an iterative method based on graphical analysis (Atkinson and Riani, 2004; Atkinson et al., 2010). Filzmoser (2005) and Filzmoser et al. (2005) presents a method based on the robust estimator MCD (*Minimum covariance determinant*).

Filzmoser et al. (2008) propose a skillful method for identifying outliers in large datasets. This is a simple implementation strategy based on a data rescaling procedure using the median (MED) and absolute median deviation (MAD).

Van Zoest et al. (2018) presents a method of detecting outliers in urban air quality sensor networks, based on a spatio-temporal classification, focusing on hourly concentrations of NO_2. Kutsuna and Yamamoto (2017) proposes a method for detecting extreme values using binary decision diagrams and the leave-one-out method. Luo et al. (2018) introduce a screening method for outliers based on variograms for medical image vectors. Zhu et al. (2017) proposes a method to detect anomalous trajectories in GPS equipped devices with the help of historical trajectory dataset and popular routes.

Berton et al. (2010) presents a method based on complex networks, the technique uses a random walk together with a dissimilarity index for identification outliers. The study by Valadares et al. (2012) presents an analysis by detecting outliers for multivariate sensor network data. Veloso and Cirillo (2016) present a methodology for identifying outliers based on principal components analysis with chi-square corrected distance samples. Another study based on principal components analysis was presented by Filzmoser et al. (2009). The work presented by Barbosa et al. (2018) proposes a multivariate outlier detection technique performed by cluster analysis.

The relevance of outlier detection procedure is a comprehensive motivation to establish a study object. The various methods already presented are generally validated through simulation techniques. Data are simulated with or without the presence of artificial outliers values and the detection procedure under study is used. In most studies, simulated test data are designed by generating multivariate normal populations with or without artificial outliers by mixing multivariate normal distributions using Monte Carlo simulations.

1.1 Contaminated Multivariate Normal Distribution

The mentioned mixture generates populations whose distribution is usually known as contaminated multivariate normal distribution. For this, consider a random vector $X' = [X_1, X_2, \ldots, X_k] \in \mathbb{R}^k$, with contaminated multivariate normal distribution, its probability density function will be given by:

$$\begin{aligned} f(x_1,x_2,\ldots,x_k) = \quad & (1-\delta)2\pi^{-k/2}|\Sigma_1|^{1/2}exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu}_1)'|\Sigma_1|^{-1}(\underline{x}-\underline{\mu}_1)\right] \\ & +(\delta)2\pi^{-k/2}|\Sigma_2|^{1/2}exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu}_2)'|\Sigma_2|^{-1}(\underline{x}-\underline{\mu}_2)\right] \end{aligned} \tag{1}$$

where $\underline{x} = (x_1, x_2, \ldots, x_k)$ is a multivariate observation of X', $(1-\delta)$ is the probability that the process will be performed by $N_k(\mu_1, \Sigma_1)$, δ is the probability that the process will be performed by $N_k(\mu_2, \Sigma_2)$, Σ_i is a positive definite matrix, $\underline{\mu}_i$ is the mean vector with $i = 1{,}2$ and $0 \leqslant \delta \leqslant 1$.

The usual procedures for studies start from the preset two distinct mean vectors $\underline{\mu}_1$ and $\underline{\mu}_2$ and a covariance matrix $\Sigma = \Sigma_1 = \Sigma_2$ such that:

$$\underline{\mu}_1 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \qquad \underline{\mu}_2 = \begin{bmatrix} \mu \\ \mu \\ \vdots \\ \mu \\ \mu \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & \rho & \rho & \ldots & \rho \\ \rho & 1 & \rho & \ldots & \rho \\ \vdots & & \ddots & & \vdots \\ \rho & \rho & \ldots & 1 & \rho \\ \rho & \rho & \ldots & \rho & 1 \end{bmatrix}$$

where $\mu \in \mathbb{R}$ with $\mu \neq 0$ and $\rho \in [-1{,}1]$ is the correlation coefficient among variables.

Although this is a usual simulation procedure for performing such tests, it is a very unrealistic procedure. A scenario of multivariate distributions with fixed means at all coordinates and constant correlation for all pairwise combinations does not represent realistic situation. The usual procedures are performed in this way due to the lack of effective mechanisms for the simulation of a correlation matrix independent of real data sets. The present study is not intended to question the validity of previous investigations, but to establish a closer analysis of a real situation.

Initially, the $\underline{\mu}_1$ mean vector will not be constant, each μ_j coordinate will be obtained randomly through a univariate $N(0{,}1)$ simulation. Subsequently, the $\underline{\mu}_2$ mean vector will have its coordinates set randomly through a simulation of $N(\pm 2{,}1)$, ie, the standard deviation of each coordinate of the $\underline{\mu}_1$ random vector is fixed, the $\underline{\mu}_2$ mean vector will be around the mean μ_j deviated by 2σ plus or minus. The goal is to create a scenario similar to offset $\mu_j \pm 2\sigma_j$. The definition of adding or subtracting the deviation will also be given by a random draw with a probability of 0.5 for plus and 0.5 otherwise.

Finally, the least realistic part of the test data generation mechanism is the choice of correlation matrices, usually constant. The proposition of an innovative method for correlation matrix generation was proposed by Martins et al. (2020). The Σ matrix illustrated previously can be replaced by a realistic correlation matrix without identical pairwise correlations. The correlation matrix can even be customized according to the comparison interests for outlier detection methods. The use of a realistic correlation matrix allows a comparative study on the efficiency of the most reliable methods through simulated data more similar to the real data.

Thus, the objectives of this work are to compare quality for multivariate detection techniques for three relevant methods, the Mahalanobis distance method proposed by Rousseeuw and Zomeren (1990); the method based on the robust estimator MCD (*Minimum covariance determinant*) proposed by Filzmoser et al. (2005) and the method proposed by Barbosa et al. (2018) by cluster analysis.

2 Methods

2.1 Outliers identification through Mahalanobis distance

The use of Mahalanobis distance, here called MD, is suggested by several authors as a suitable metric for outlier detection procedures in multivariate data. The sample Mahalanobis distance can be defined as:

$$MD_i = \sqrt{(\boldsymbol{x_i} - \bar{\boldsymbol{x}})'\boldsymbol{S}^{-1}(\boldsymbol{x_i} - \bar{\boldsymbol{x}})}$$

where $\boldsymbol{x_i}$ is the $i-$th sample observation of the multivariate data set $\boldsymbol{X}$, $\bar{\boldsymbol{x}}$ is the mean vector of $\boldsymbol{X}$, and $\boldsymbol{S}$ is the matrix of variances and covariances of $\boldsymbol{X}$, given by:

$$\boldsymbol{S} = \frac{\sum_{i=1}^{n}(\boldsymbol{x_i} - \bar{\boldsymbol{x}})(\boldsymbol{x_i} - \bar{\boldsymbol{x}})'}{n-1}\,.$$

The outliers identification based on Mahalanobis distance MD_i, according to Rousseeuw and Zomeren (1990), can be performed using theoretical quantiles of the χ^2 distribution. These authors suggest determining which observations whose Mahalanobis quadratic distance (MD_i^2) are greater than the $1-\alpha$ quantile of the $\chi^2_{(p)}$ distribution, where the degrees of freedom p represents the number of variables considered.

Distance measures, especially Mahalanobis distance, are extremely sensitive to the outliers presence. Extreme values, or an outliers group, can severely influence these metrics for distances. So an immediate question is: “how can a distance easily influenced by outliers be able to identify them?”. The most sensitive parts of this measure, the mean vector and the variance matrix should be treated particularly, calculating them robustly, ie, robust estimation techniques should be used. Among the robust estimators, we can highlight the Minimum Covariance Determinant, here nominated MCD and the Minimum Volume Ellipsoid, here called MVE.

2.1.1 Minimum Covariance Determinant MCD

According to Rousseeuw and Driessen (1999), the minimum covariance determinant MCD is probably the most widely used method for constructing robust estimators since it is a computationally fast algorithm.

The MCD estimator is determined by a subset of size h, which minimizes the determinant of the sample covariance matrix, calculated only under the h points. The location estimate is the mean of these points, while the scatter estimator is proportional to its covariance matrix, where choosing the size of h determines the robustness of the estimator.

With a compromise between robustness and efficiency, the study by Filzmoser et al. (2005) used a value of $h \approx 0.75n$, where n is the sample size.

2.1.2 Minimum Volume Ellipsoid MVE

Rousseeuw and Zomeren (1990) state that the minimum volume ellipsoid MVE is an estimator based on smaller volume ellipsoid able of covering at least k points from the $\boldsymbol{X}$ sample set, where k generally equals the integer part of $[n/2]+1$. This estimator has a breakpoint 0.5 and is equivariant, ie, $T(x_1A+b,\ldots,x_nA+b) = T(x_1,\ldots,x_n)A+b$ and $C(x_1A+b,\ldots,x_nA+b) = A^tC(x_1,\ldots,x_n)A$.

This robust estimator is defined by the (T,C) pair where T is a $p-$dimensional vector of the sample mean and C is the $p-$dimensional sub-matrix of the sample covariance matrix . The robust distance defined for MVE is given by:

$$RD_i = \sqrt{(x_iT(X))C(X)^{-1}(x_iT(X))^t}\ .$$

2.2 Outliers identification through cluster analysis

Let $\mathcal{P}$ be any interest population with the presence of outliers. Consider using the $k-$means cluster analysis procedure, with interest in establishing similarity grouping among the individuals of $\mathcal{P}$ population. The clustering method $k-$means start from the choice of k elements to be centroids, in the usual procedure, this choice is completely random, but nothing restrict the construction of any prior criteria for the choice.

For the random choice of centroids, the method is capable of producing partitions of various formats, without guarantee that two realizations of the same procedure in the same dataset will provide the same partition. On the other hand, some specific centroid choice criteria would be able to provide more stability to the results achieved.

From an implementation standpoint, such randomness can be easily controlled, the choice of centroids is based on a random variable simulation procedure, ie, a random draw. A preset seed for starting the pseudo random number generation process is sufficient to restrain such an unwanted randomness effect. This technique is sufficient to guarantee that given the same dataset in two distinct realizations will always get the same partition.

Once the partition under investigation is established, it is possible to verify the Euclidean distance of each cluster centroid concerning the median of the complete dataset. The growth of the Euclidean distance with respect to a grouping, makes the elements of that grouping tend to be observed as elements of an argument of potential outliers candidates. Specific criteria about the values of this distance can be established to determine outlier grouping.

In particular, the methodology presented by Barbosa et al. (2018) evaluated the standard deviation (s) between cluster centroids and the median of the complete dataset. Clusters whose distance between their centroid and the median exceed $2.5s$ were considered clusters of outliers in the proposed methodology.

In $\mathbb{R}^2$ representation, the geometric visualization is viable and quite didactic to understand the technique. Suppose the dataset divided into four clusters. The figure 1 illustrates this condition.

Figure 1 clearly shows the 4 grouping over $2.5s$ with respect to the median dataset represented in the image by Md. On the other hand, the other groupings are close enough to the median concerning the $2.5s$ criterion. Thus the elements of the 4 grouping are classified by the method as outliers.

It is noteworthy that this technique is dependent on the choice of value k, number of clusters. The study by Barbosa et al. (2018) presents a completely *ad-hoc* choice of $k = {}^{n}\!/_{10}$. It is intuitive to suppose that such a choice tends to behave well in some scenarios, but on the other hand, is completely deficient in others. The discussion of this threshold generates a secondary objective to be achieved by the present study, to verify the effect of different choices of the value k. The simulated experiments that will be presented in this research investigate different k values, in particular k ranging from 2 to ${}^{n}\!/_{\log(n)}$ and choose the best option through some prefixed measure.

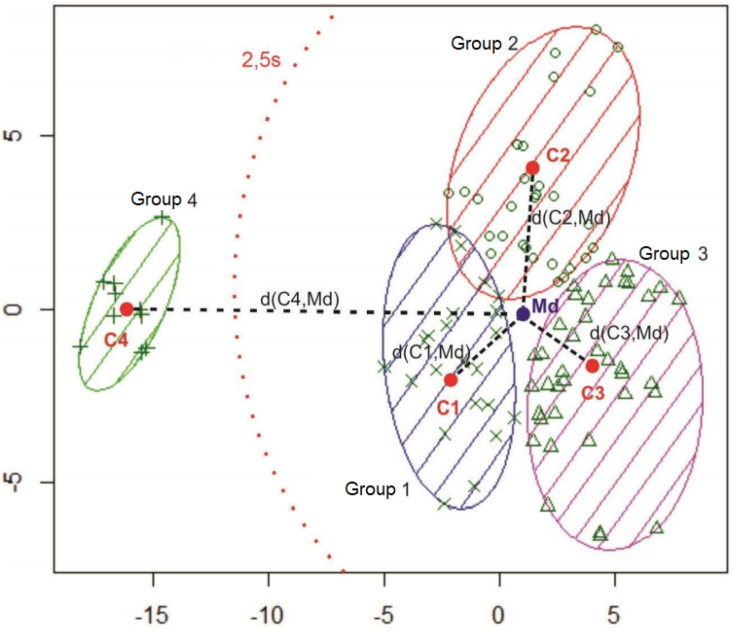

Figure 1: Graphical visualization of outliers identification method by cluster analysis.

3 Results and insights

A procedure for comparative verification will be established between the three mentioned strategies, named MCD (Minimum Volume Determination), MVE (Minimum Volume Ellipsoid) and CAM (Cluster Analysis Method).

Initially, it will be necessary to simulate a sample following the distribution defined in the expression (1). At each generated observation, it will be identified if this value follows the $\underline{\mu_1}$ or $\underline{\mu_2}$ vector of means. Since the generated observation follows the vector of means μ_2, such observation is a potential outlier candidate. It is important to report the fact that an observation generated using the mean vector $\underline{\mu_2}$ does not guarantee that the observation is an outlier, it is just a condition in the distribution to force the possible production of extreme values in the dataset. However, comparing the values identified by the methodology under study with the values supposedly outliers is the fair and appropriate comparative procedure.

The present study will adopt three metrics to assess the quality of the detection procedure and one metric to evaluate computational efficiency in the use of the technique. The quality evaluation will be based on the sensitivity, specificity, and accuracy of the detection procedure. The computational efficiency will be analyzed through the computational time needed to perform the detection procedure.

Given a simulation of n multivariate observations of which some are potential candidates for outliers. Define the following sets: Ω, set of all simulated observations; $\mathbb{O}$, set of *outlier* candidates, ie, observations generated with the mean vector μ_2 and $\mathbb{D}$, set of outliers identified by the method used. Sensitivity is the probability that since an observation belongs to $\mathbb{O}$, it belongs to $\mathbb{D}$, ie $P(\mathbb{D}|\mathbb{O})$. Specificity is the probability that since an observation that not belong to $\mathbb{O}$, it does not belong to $\mathbb{D}$, ie $P(\overline{\mathbb{D}}|\overline{\mathbb{O}})$, where $\overline{A}$ is the complement to the A. Finally, the accuracy is the total ratio of hits between positives and negatives, ie $P[(\mathbb{D}\cap\mathbb{O})\cup(\overline{\mathbb{D}}\cap\overline{\mathbb{O}})]$.

Some doubt may lie in choosing the difference between the mean vectors for sampling and generating artificial outliers to test for detection. Is the experiment conducted in this way sufficient to produce extreme values? Even to ask about the possibility of producing extreme values that would be easily detectable by various methodologies. Choosing μ_1 for non-candidate outliers and μ_2 for candidate outliers (as described in the section 1) in the mixture distribution is justified through the univariate criterion of taking two standard deviations between the means μ_1 and μ_2. Experiments different from this proposition can be seen as future works, as a complementary investigation. It is also worth mentioning that, since the three methods were compared in the same simulation methodology, the comparison is realistic and without bias favoring either method.

Comprehensive experiments with various specifications were performed. Different numbers of variables in the multivariate distribution to generate the sample data were used, the values used were: 5, 10, 20, 50, 100. Additionally, different mix rates δ were used: $0, 0.02, 0.05, 0.10$. Finally, two distinct covariance matrix types, Σ_1, with no fixed correlations, but ensuring all correlations such that $0.3 \leqslant \rho \leqslant 0.6$ and Σ_2 custom, guaranteed 20% high correlations , above 0.6, 30% of average correlations such that $0.3 \leqslant \rho \leqslant 0.6$ and the rest of low correlations, less than 0.3. In all scenarios 500 observations were generated.

The first set of experiments considers mix rate $\delta = 0$. It is noteworthy that in this situation there is no sensitivity measure since this format considers that outliers were not produced in the simulated data. Also, the absence of sensitivity measure makes the accuracy and specificity measures the same. These results can be viewed in the table 1.

Table 1: Experiment performed with mix rate $\delta = 0$ and covariance matrix Σ_1.

accuracy or specificity						
	MCD		MVE		CAM	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	0.9742 (0.0070)	0.9480 - 0.9920	0.9318 (0.0127)	0.8860 - 0.9640	**0.9943** (0.0494)	0.3260 - 1.0000
10	0.9723 (0.0072)	0.9440 - 0.9920	0.9305 (0.0122)	0.8780 - 0.9700	**0.9998** (0.0061)	0.8060 - 1.0000
20	0.9663 (0.0080)	0.9340 - 0.9900	0.9268 (0.0125)	0.8800 - 0.9640	**1.0000** (0.0000)	1.0000 - 1.0000
50	0.8970 (0.0162)	0.8520 - 0.9400	0.8729 (0.0167)	0.8220 - 0.9280	**0.9998** (0.0042)	0.9020 - 1.0000
100	0.7685 (0.0145)	0.7160 - 0.8220	0.6953 (0.0154)	0.6520 - 0.7520	**0.9999** (0.0018)	0.9440 - 1.0000

For the first comparative study, a remarkable superiority of the CAM method can be seen. Higher means are shown in bold in the table 1. The second set of experiments considers mix rate $\delta = 0.02$. These results can be viewed in the table 2.

Table 2: Experiment performed with mix rate $\delta = 0.02$ and covariance matrix Σ_1.

sensitivity						
	MCD		MVE		CAM	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	0.9481 (0.1353)	0.0000 - 1.0000	**0.9699** (0.0973)	0.0000 - 1.0000	0.7889 (0.3886)	0.0000 - 1.0000
10	0.9992 (0.0141)	0.6000 - 1.0000	**0.9995** (0.0130)	0.6000 - 1.0000	0.9799 (0.1314)	0.0000 - 1.0000
20	**1.0000** (0.0000)	1.0000 - 1.0000	**1.0000** (0.0000)	1.0000 - 1.0000	0.9984 (0.0325)	0.0000 - 1.0000
50	**1.0000** (0.0000)	1.0000 - 1.0000	**1.0000** (0.0000)	1.0000 - 1.0000	**1.0000** (0.0000)	1.0000 - 1.0000
100	**1.0000** (0.0000)	1.0000 - 1.0000	0.9832 (0.0636)	0.3000 - 1.0000	**1.0000** (0.0000)	1.0000 - 1.0000
specificity						
	MCD		MVE		CAM	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	0.9779 (0.0067)	0.9571 - 0.9959	0.9336 (0.0121)	0.8878 - 0.9653	**0.9801** (0.0776)	0.0000 - 1.0000
10	0.9756 (0.0069)	0.951 - 0.9939	0.9323 (0.0120)	0.8857 - 0.9653	**0.9988** (0.0148)	0.6918 - 1.0000
20	0.9685 (0.0083)	0.9347 - 0.9878	0.9282 (0.0125)	0.8837 - 0.9612	**0.9998** (0.0046)	0.8735 - 1.0000
50	0.8988 (0.0162)	0.8449 - 0.9408	0.8784 (0.0163)	0.8224 - 0.9265	**1.0000** (0.0000)	1.0000 - 1.0000
100	0.7805 (0.0150)	0.7347 - 0.8367	0.7145 (0.0182)	0.6449 - 0.7735	**1.0000** (0.0000)	1.0000 - 1.0000
accuracy						
	MCD		MVE		CAM	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	**0.9773** (0.0072)	0.9520 - 0.9960	0.9343 (0.0119)	0.8900 - 0.9660	0.9762 (0.0771)	0.0200 - 1.0000
10	0.9761 (0.0068)	0.9520 - 0.9940	0.9337 (0.0117)	0.8880 - 0.9660	**0.9984** (0.0155)	0.6960 - 1.0000
20	0.9691 (0.0081)	0.9360 - 0.9880	0.9296 (0.0123)	0.8860 - 0.9620	**0.9998** (0.0046)	0.8740 - 1.0000
50	0.9008 (0.0159)	0.8480 - 0.9420	0.8809 (0.0159)	0.8260 - 0.9280	**1.0000** (0.0000)	1.0000 - 1.0000
100	0.7849 (0.0147)	0.7400 - 0.8400	0.7198 (0.0181)	0.6520 - 0.7780	**1.0000** (0.0000)	1.0000 - 1.0000

The MVE method presents more consistent results for sensitivity. On the other hand, the MCD and CAM methods are showing increasing improvements as the numbers of variables of database grows. This quality makes it clear that in situations with many variable, the CAM method is very competitive in terms of sensitivity. In the specificity and accuracy analyzes, the CAM method again reports the predominance found in the previous experiment. In particular, the only realization in which the CAM method was not superior, in terms of accuracy, occurred with 5 variables only, and a marginal loss has been verified when the average accuracy is evaluated. The third experimental investigation uses mix rate $\delta = 0.05$. Results can be viewed in the table 3.

Table 3: Experiment performed with mix rate $\delta = 0.05$ and covariance matrix Σ_1.

sensitivity

	MCD		MVE		CAM	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	0.9481 (0.1374)	0.0800 - 1.0000	**0.9706** (0.0979)	0.2000 - 1.0000	0.9306 (0.2090)	0.0000 - 1.0000
10	0.9981 (0.0310)	0.0800 - 1.0000	**0.9986** (0.0273)	0.1600 - 1.0000	0.9908 (0.0846)	0.0000 - 1.0000
20	**1.0000 (0.0000)**	1.0000 - 1.0000	0.9990 (0.0107)	0.7600 - 1.0000	0.9989 (0.0317)	0.0000 - 1.0000
50	**1.0000 (0.0000)**	1.0000 - 1.0000	0.9145 (0.1409)	0.3200 - 1.0000	**1.0000** (0.0000)	1.0000 - 1.0000
100	**1.0000 (0.0000)**	1.0000 - 1.0000	0.6256 (0.1816)	0.1200 - 1.0000	**1.0000** (0.0000)	1.0000 - 1.0000

specificity

	MCD		MVE		CAM	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	0.9822 (0.0060)	0.9621 - 1.0000	0.9354 (0.0119)	0.8821 - 0.9726	**0.9842** (0.0744)	0.0000 - 1.0000
10	0.9790 (0.0065)	0.9537 - 0.9958	0.9342 (0.0122)	0.8926 - 0.9726	**0.9990** (0.0140)	0.6926 - 1.0000
20	0.9721 (0.0078)	0.9453 - 0.9937	0.9318 (0.0126)	0.8863 - 0.9663	**1.0000** (0.0015)	0.9516 - 1.0000
50	0.9043 (0.0165)	0.8316 - 0.9579	0.8881 (0.0168)	0.8295 - 0.9326	**1.0000** (0.0000)	1.0000 - 1.0000
100	0.8007 (0.0162)	0.7474 - 0.8484	0.7111 (0.0190)	0.6568 - 0.7789	**1.0000** (0.0000)	1.0000 - 1.0000

accuracy

	MCD		MVE		CAM	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	0.9805 (0.0091)	0.9260 - 0.9980	0.9372 (0.0120)	0.8880 - 0.9700	**0.9815** (0.0722)	0.0500 - 1.0000
10	0.9800 (0.0065)	0.9160 - 0.9960	0.9374 (0.0118)	0.8560 - 0.9740	**0.9986** (0.0153)	0.6920 - 1.0000
20	0.9735 (0.0074)	0.9480 - 0.9940	0.9351 (0.0120)	0.8920 - 0.9680	**0.9999** (0.0030)	0.9040 - 1.0000
50	0.9091 (0.0156)	0.8400 - 0.9600	0.8895 (0.0173)	0.8220 - 0.9360	**1.0000** (0.0000)	1.0000 - 1.0000
100	0.8107 (0.0154)	0.7600 - 0.8560	0.7069 (0.0238)	0.6440 - 0.7900	**1.0000** (0.0000)	1.0000 - 1.0000

A comparison between the results of tables 2 and 3 shows a great similarity. This makes clear that increasing the δ mix rate from 0.02 to 0.05 does not appear to affect the performance of methodologies in general. The fourth is the last experimental study with correlation matrices Σ_1 uses mix rate $\delta = 0.10$. Results can be viewed in the table 4.

Table 4: Experiment performed with mix rate $\delta = 0.10$ and covariance matrix Σ_1.

sensitivity

	MCD		MVE		CAM	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	0.9156 (0.1901)	0.0000 - 1.0000	**0.9461** (0.1502)	0.0400 - 1.0000	0.9402 (0.1727)	0.0000 - 1.0000
10	**0.9982** (0.0185)	0.6600 - 1.0000	0.9735 (0.1021)	0.3200 - 1.0000	0.9963 (0.0382)	0.1200 - 1.0000
20	**1.0000** (0.0000)	1.0000 - 1.0000	0.7287 (0.2864)	0.1000 - 1.0000	0.9999 (0.0025)	0.9200 - 1.0000
50	**1.0000** (0.0000)	1.0000 - 1.0000	0.3728 (0.1487)	0.1200 - 1.0000	**1.0000** (0.0000)	1.0000 - 1.0000
100	0.4142 (0.1507)	0.1600 - 1.0000	0.4202 (0.0869)	0.1600 - 0.8000	**1.0000** (0.0000)	1.0000 - 1.0000

(continued from table in previous page)

	specificity					
	MCD		MVE		CAM	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	0.9873 (0.0055)	0.9644 - 1.0000	0.9403 (0.0112)	0.8978 - 0.9733	**0.9883** (0.0446)	0.5200 - 1.0000
10	0.9839 (0.0064)	0.9578 - 1.0000	0.9390 (0.0120)	0.8978 - 0.9756	**0.9995** (0.0087)	0.7911 - 1.0000
20	0.9770 (0.0070)	0.9511 - 0.9956	0.9400 (0.0128)	0.8911 - 0.9733	**1.0000** (0.0000)	1.0000 - 1.0000
50	0.9144 (0.0161)	0.8667 - 0.9600	0.8900 (0.0168)	0.8311 - 0.9400	**1.0000** (0.0000)	1.0000 - 1.0000
100	0.7845 (0.0190)	0.7067 - 0.8533	0.7069 (0.0179)	0.6533 - 0.7689	**1.0000** (0.0000)	1.0000 - 1.0000
	accuracy					
	MCD		MVE		CAM	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	0.9801 (0.0204)	0.8720 - 1.0000	0.9409 (0.0165)	0.8400 - 0.9740	**0.9835** (0.0461)	0.5580 - 1.0000
10	0.9853 (0.0059)	0.9600 - 1.0000	0.9425 (0.0135)	0.8780 - 0.9780	**0.9992** (0.0105)	0.7540 - 1.0000
20	0.9793 (0.0063)	0.9560 - 0.9960	0.9188 (0.0292)	0.8280 - 0.9740	**1.0000** (0.0003)	0.9920 - 1.0000
50	0.9229 (0.0145)	0.8800 - 0.9640	0.8383 (0.0228)	0.7720 - 0.9320	**1.0000** (0.0000)	1.0000 - 1.0000
100	0.7474 (0.0269)	0.6680 - 0.8580	0.6783 (0.0208)	0.6180 - 0.7540	**1.0000** (0.0000)	1.0000 - 1.0000

The previous comparison between the results of the tables 2 and 3 could be seen by some analysts as mere coincidence or even some non-systemic noise. Checking the results of the table 4, even with the increase of the mix rate δ to 0.10, the superior performance of the CAM method is shown by the experimental results. The accuracy measure found in the tables 1, 2, 3 and 4 are relevant indications of the quality of the procedure. A simultaneous view of this information for the various values of mix rate δ is shown in figure 2.

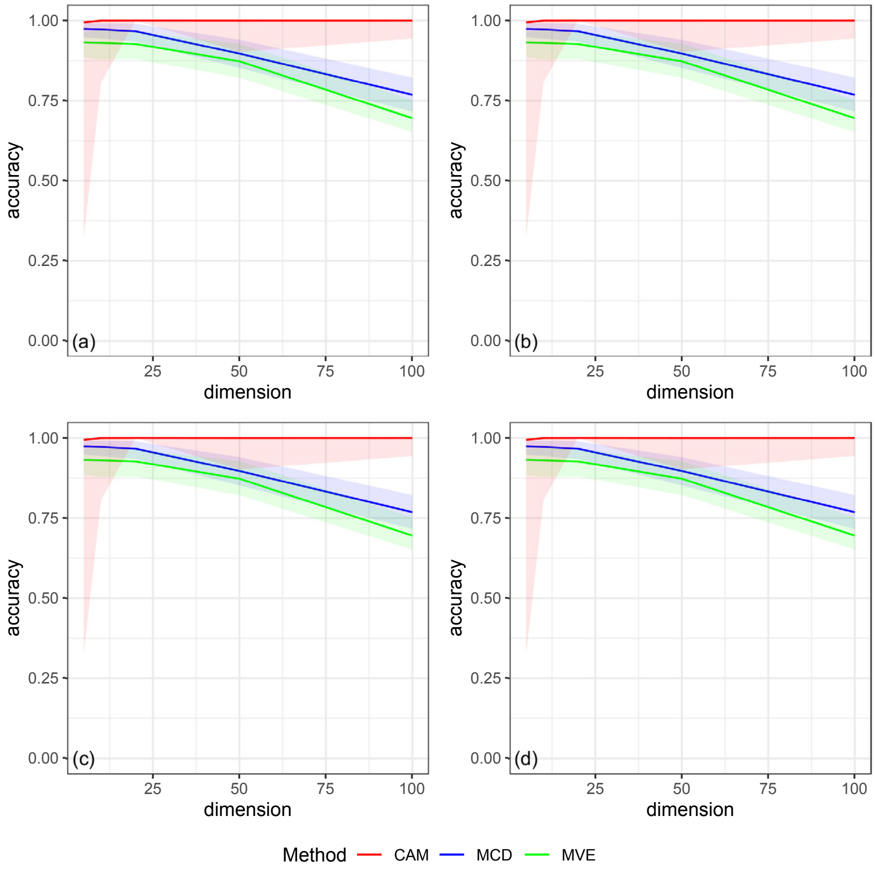

Figure 2: Evolution of accuracy to dimension (number of variables).

The graphs shown in figure 2 show the value of the accuracy and the minimum and maximum limits for each method. In the upper left corner, graph (a) shows values for $\delta = 0$, in the upper right corner, graph (b) shows values for $\delta = 0.02$, in lower left corner, graph (c) shows values for $\delta = 0.05$ and in the lower right corner, graph (d) shows values for $\delta = 0.10$. Graphical analysis clearly illustrates that the behavior of the CAM method does not suffer from an increase in the number of variables. On the other hand, the other methods tend to have their results deteriorated by the dimensional increase. This valence verified for the CAM method is relevant when considering the increasing rate of studies involving a large volume of variables under investigation.

The following experiments consider a new correlation matrix format, Σ_2 custom. Results are shown in table 5.

Table 5: Experiment performed with mix rate $\delta = 0$ and covariance matrix Σ_2.

	accuracy or specifity					
	MCD		MVE		CAM	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	**0.9739** (0.0069)	0.9500 - 0.9920	0.9311 (0.0129)	0.8820 - 0.9660	0.6724 (0.2334)	0.0000 - 1.0000
10	0.9726 (0.0072)	0.9440 - 0.9920	0.9311 (0.0123)	0.8860 - 0.9620	**0.9970** (0.0364)	0.1700 - 1.0000
20	0.9659 (0.0082)	0.9240 - 0.9880	0.9266 (0.0124)	0.8740 - 0.9760	**0.9995** (0.0104)	0.7300 - 1.0000
50	0.8973 (0.0154)	0.8500 - 0.9380	0.8727 (0.0160)	0.8180 - 0.9240	**0.9997** (0.0081)	0.7440 - 1.0000
100	0.7683 (0.0148)	0.7140 - 0.8160	0.6954 (0.0160)	0.6520 - 0.7460	**0.9998** (0.0044)	0.9020 - 1.0000

The custom matrix Σ_2 guarantees 20% high correlations, above 0.6, 30% mean correlations such that $0.3 \leqslant \rho \leqslant 0.6$ and the rest of low correlations, less than 0.3. Again several mix rates will be considered and for a mix rate $\delta = 0$ there is no sensitivity measure. The performance revealed with the Σ_1 matrix is similar to that seen with the Σ_2 matrix. A quick comparison between the tables 1 and 5 illustrates this fact. The tables 6, 7 and 8 show the ratings for different mix rates.

Table 6: Experiment performed with mix rate $\delta = 0.02$ and covariance matrix Σ_2.

	sensitivity					
	MCD		MVE		CAM	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	0.9580 (0.1240)	0.1000 - 1.0000	**0.9746** (0.0915)	0.2000 - 1.0000	0.9255 (0.2084)	0.0000 - 1.0000
10	0.9990 (0.0200)	0.4000 - 1.0000	**0.9993** (0.0164)	0.5000 - 1.0000	0.9710 (0.1570)	0.0000 - 1.0000
20	**1.0000** (0.0000)	1.0000 - 1.0000	**1.0000** (0.0000)	1.0000 - 1.0000	0.9999 (0.0032)	0.9000 - 1.0000
50	**1.0000** (0.0000)	1.0000 - 1.0000	**1.0000** (0.0000)	1.0000 - 1.0000	**1.0000** (0.0000)	1.0000 - 1.0000
100	**1.0000** (0.0000)	1.0000 - 1.0000	0.9873 (0.0517)	0.6000 - 1.0000	**1.0000** (0.0000)	1.0000 - 1.0000
	specificity					
	MCD		MVE		CAM	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	**0.9775** (0.0066)	0.9551 - 0.9959	0.9330 (0.0124)	0.8918 - 0.9653	0.9216 (0.1664)	0.0000 - 1.0000
10	0.9754 (0.0069)	0.9510 - 0.9959	0.9326 (0.0121)	0.8816 - 0.9653	**0.9958** (0.0352)	0.5388 - 1.0000
20	0.9687 (0.0080)	0.9388 - 0.9918	0.9279 (0.0123)	0.8878 - 0.9633	**1.0000** (0.0000)	1.0000 - 1.0000
50	0.8997 (0.0164)	0.8408 - 0.9571	0.8772 (0.0171)	0.8224 - 0.9327	**1.0000** (0.0000)	1.0000 - 1.0000
100	0.7796 (0.0151)	0.7327 - 0.8224	0.7144 (0.0179)	0.6510 - 0.7714	**1.0000** (0.0000)	1.0000 - 1.0000
	accuracy					
	MCD		MVE		CAM	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	**0.9771** (0.0069)	0.9500 - 0.9940	0.9338 (0.0122)	0.8920 - 0.9660	0.9217 (0.1640)	0.0200 - 1.0000
10	0.9759 (0.0067)	0.9520 - 0.9960	0.9340 (0.0118)	0.8840 - 0.9660	**0.9953** (0.0357)	0.5360 - 1.0000
20	0.9694 (0.0079)	0.9400 - 0.9920	0.9294 (0.0120)	0.8900 - 0.9640	**1.0000** (0.0001)	0.9980 - 1.0000
50	0.9017 (0.0161)	0.8440 - 0.9580	0.8797 (0.0168)	0.8260 - 0.9340	**1.0000** (0.0000)	1.0000 - 1.0000
100	0.7840 (0.0148)	0.7380 - 0.8260	0.7199 (0.0177)	0.6520 - 0.7760	**1.0000** (0.0000)	1.0000 - 1.0000

Table 7: Experiment performed with mix rate $\delta = 0.05$ and covariance matrix Σ_2.

sensitivity						
	MCD		MVE		CAM	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	0.9428 (0.1531)	0.0400 - 1.0000	**0.9664** (0.1111)	0.0800 - 1.0000	0.9545 (0.1398)	0.0000 - 1.0000
10	0.9983 (0.0205)	0.4800 - 1.0000	**0.9984** (0.0202)	0.5200 - 1.0000	0.9948 (0.0528)	0.0000 - 1.0000
20	**1.0000** (0.0000)	1.0000 - 1.0000	0.9992 (0.0097)	0.8400 - 1.0000	**1.0000** (0.0013)	0.9600 - 1.0000
50	**1.0000** (0.0000)	1.0000 - 1.0000	0.9262 (0.1381)	0.3200 - 1.0000	**1.0000** (0.0000)	1.0000 - 1.0000
100	**1.0000** (0.0000)	1.0000 - 1.0000	0.6399 (0.1873)	0.1600 - 1.0000	**1.0000** (0.0000)	1.0000 - 1.0000
specificity						
	MCD		MVE		CAM	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	**0.9815** (0.0061)	0.9516 - 0.9958	0.9360 (0.0122)	0.8989 - 0.9684	0.9663 (0.1099)	0.0000 - 1.0000
10	0.9788 (0.0067)	0.9579 - 0.9958	0.9344 (0.0121)	0.8884 - 0.9705	**0.9987** (0.0168)	0.6632 - 1.0000
20	0.9720 (0.0078)	0.9411 - 0.9937	0.9317 (0.0127)	0.8884 - 0.9726	**1.0000** (0.0000)	1.0000 - 1.0000
50	0.9050 (0.0156)	0.8484 - 0.9474	0.8887 (0.0159)	0.8379 - 0.9389	**1.0000** (0.0000)	1.0000 - 1.0000
100	0.8017 (0.0167)	0.7368 - 0.8568	0.7127 (0.0192)	0.6611 - 0.7789	**1.0000** (0.0000)	1.0000 - 1.0000
accuracy						
	MCD		MVE		CAM	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	**0.9796** (0.0100)	0.9240 - 0.9960	0.9375 (0.0122)	0.8840 - 0.9660	0.9657 (0.1060)	0.0500 - 1.0000
10	0.9797 (0.0065)	0.9540 - 0.9960	0.9376 (0.0115)	0.8940 - 0.9720	**0.9985** (0.0172)	0.6580 - 1.0000
20	0.9734 (0.0074)	0.9440 - 0.9940	0.9350 (0.0121)	0.8940 - 0.9680	**1.0000** (0.0001)	0.9980 - 1.0000
50	0.9097 (0.0148)	0.8560 - 0.9500	0.8906 (0.0165)	0.8200 - 0.9400	**1.0000** (0.0000)	1.0000 - 1.0000
100	0.8116 (0.0159)	0.7500 - 0.8640	0.7090 (0.0243)	0.6360 - 0.7900	**1.0000** (0.0000)	1.0000 - 1.0000

Table 8: Experiment performed with mix rate $\delta = 0.10$ and covariance matrix Σ_2.

sensitivity						
	MCD		MVE		CAM	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	0.9203 (0.1777)	0.0000 - 1.0000	0.9503 (0.1359)	0.0800 - 1.0000	**0.9610** (0.0986)	0.0000 - 1.0000
10	**0.9992** (0.0120)	0.6600 - 1.0000	0.9830 (0.0790)	0.3000 - 1.0000	0.9987 (0.0217)	0.3400 - 1.0000
20	**1.0000** (0.0000)	1.0000 - 1.0000	0.7639 (0.2820)	0.1200 - 1.0000	**1.0000** (0.0000)	1.0000 - 1.0000
50	**1.0000** (0.0000)	1.0000 - 1.0000	0.3865 (0.1656)	0.1000 - 1.0000	**1.0000** (0.0000)	1.0000 - 1.0000
100	0.4139 (0.1534)	0.1200 - 1.0000	0.4216 (0.0897)	0.1400 - 0.7800	**1.0000** (0.0000)	1.0000 - 1.0000
specificity						
	MCD		MVE		CAM	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	**0.9871** (0.0054)	0.9667 - 1.0000	0.9396 (0.0116)	0.8978 - 0.9756	0.9759 (0.0857)	0.0000 - 1.0000
10	0.9844 (0.0060)	0.9644 - 1.0000	0.9392 (0.0120)	0.8956 - 0.9778	**0.9998** (0.0036)	0.8956 - 1.0000
20	0.9769 (0.0077)	0.9489 - 0.9978	0.9396 (0.0127)	0.9000 - 0.9800	**1.0000** (0.0000)	1.0000 - 1.0000
50	0.9130 (0.0156)	0.8533 - 0.9644	0.8895 (0.0163)	0.8222 - 0.9422	**1.0000** (0.0000)	1.0000 - 1.0000
100	0.7710 (0.0850)	0.0000 - 0.8556	0.7074 (0.0190)	0.6444 - 0.7756	**1.0000** (0.0000)	1.0000 - 1.0000

(continued from table in previous page)

	accuracy					
	MCD		MVE		CAM	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	**0.9804** (0.0191)	0.8900 - 1.0000	0.9407 (0.0159)	0.8580 - 0.9780	0.9744 (0.0800)	0.1000 - 1.0000
10	0.9859 (0.0056)	0.9340 - 1.0000	0.9436 (0.0122)	0.8880 - 0.9760	**0.9997** (0.0041)	0.9040 - 1.0000
20	0.9792 (0.0069)	0.9540 - 0.9980	0.9220 (0.0284)	0.8480 - 0.9780	**1.0000** (0.0000)	1.0000 - 1.0000
50	0.9217 (0.0140)	0.8680 - 0.9680	0.8392 (0.0240)	0.7680 - 0.9300	**1.0000** (0.0000)	1.0000 - 1.0000
100	0.7353 (0.0710)	0.1000 - 0.8580	0.6788 (0.0220)	0.6140 - 0.7500	**1.0000** (0.0000)	1.0000 - 1.0000

Figure 3 presents the accuracy measures of the tables 5, 6, 7 and 8 and their respective minimum and maximum limits.

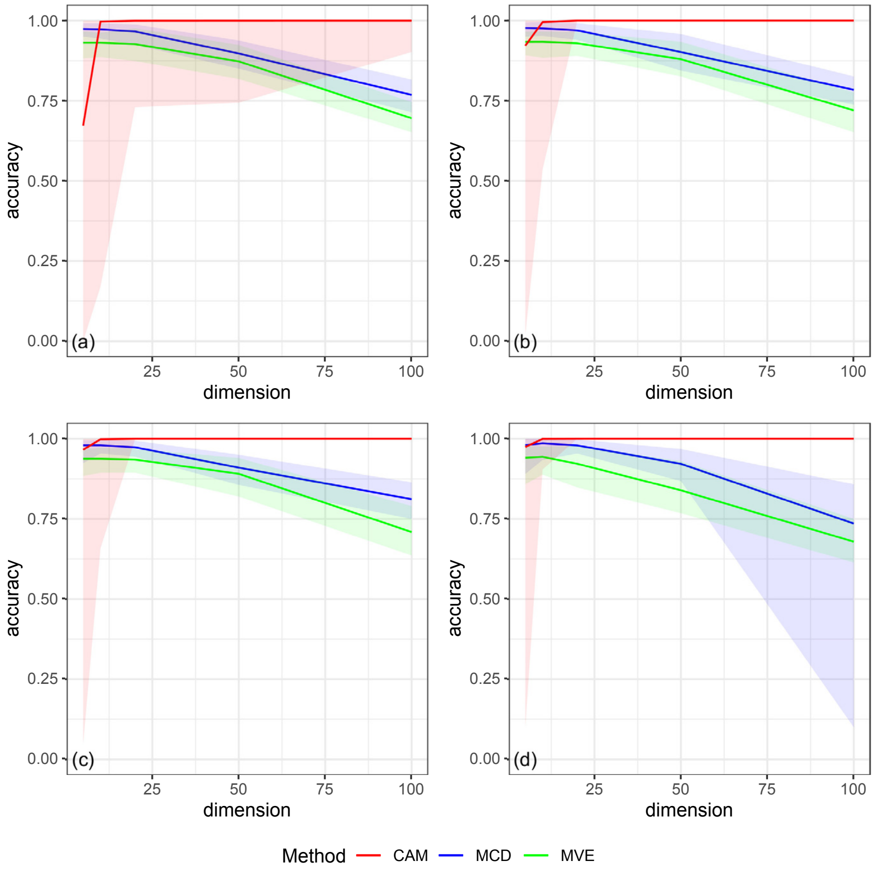

Figure 3: Evolution of accuracy to dimension (number of variables).

Using the Σ_2 matrix, the MVE method yields more consistent results for sensitivity. The MCD and CAM methods increase their quality with the number of variables in the experimental dataset grows. This confirms that in many situations the CAM method is effective with respect to the measure of sensitivity.

Considering the specificity and accuracy, the CAM method again reports the predominance viewed in the previous experiment. In particular, the predominance grows with the number of variables in the dataset under investigation.

In the graphs shown in figure 3 the graph (a) with values for $\delta = 0$ is shown in the upper left corner, the graph (b) with values for $\delta = 0.02$, in the lower left corner, the graph (c) with values for $\delta = 0.05$ and in the lower right corner, the graph (d) with values for $\delta = 0.10$. Graphical analysis again shows that the behavior of the CAM method does not loss from the increase in the number of variables, and that the other methods tend to have worse results with the increase in the number of variables.

A comparison between the results of the tables 6, 7 and 8 and the figure 3 also makes it clear that increasing the δ mix rate from 0.02 to 0.05 and 0.10 does not affect the performance of methodologies in general.

The computational efficiency was evaluated by computational time measurements needed to perform the detection procedure. Results can be viewed through the tables 9 and 10.

Table 9: Cpu time (seconds) in detection procedure with Σ_1 correlation matrix.

	MCD ($\delta = 0$)		MVE ($\delta = 0$)		CAM ($\delta = 0$)	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	0.0966 (0.0061)	0.0937 - 0.1094	**0.0645** (0.0054)	0.0625 - 0.0937	0.1965 (0.1117)	0.0000 - 0.2969
10	0.2202 (0.0049)	0.2031 - 0.2656	0.1414 (0.0034)	0.1406 - 0.1562	**0.0066** (0.0334)	0.0000 - 0.3125
20	0.7201 (0.0058)	0.7031 - 0.7656	0.3203 (0.0078)	0.3125 - 0.3437	**0.0042** (0.0185)	0.0000 - 0.3906
50	4.5833 (0.3317)	4.4835 - 4.5558	1.6504 (0.0495)	1.6028 - 2.5572	**0.0060** (0.0211)	0.0000 - 0.6566
100	18.1032 (0.2012)	18.0143 - 21.2784	6.6741 (0.0633)	6.6089 - 7.2723	**0.0103** (0.0505)	0.0000 - 1.1249
	MCD ($\delta = 0.02$)		MVE ($\delta = 0.02$)		CAM ($\delta = 0.02$)	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	0.0969 (0.0063)	0.0937 - 0.1094	0.0646 (0.0054)	0.0625 - 0.0781	**0.0073** (0.0339)	0.0000 - 0.3125
10	0.2338 (0.0083)	0.2187 - 0.2500	0.1506 (0.0077)	0.1406 - 0.1875	**0.0035** (0.0125)	0.0000 - 0.3437
20	0.7194 (0.0061)	0.7031 - 0.7343	0.3197 (0.0079)	0.3125 - 0.3593	**0.0038** (0.0067)	0.0000 - 0.0156
50	4.4996 (0.0146)	4.4528 - 4.6559	1.6229 (0.0111)	1.6092 - 1.6874	**0.0065** (0.0283)	0.0000 - 0.6249
100	18.0420 (0.0345)	17.9830 - 18.4830	6.6543 (0.0961)	6.5933 - 8.3587	**0.0096** (0.0339)	0.0000 - 1.0468
	MCD ($\delta = 0.05$)		MVE ($\delta = 0.05$)		CAM ($\delta = 0.05$)	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	0.0979 (0.0082)	0.0937 - 0.1855	**0.0658** (0.0079)	0.0616 - 0.1983	0.0960 (0.1230)	0.0000 - 0.4531
10	0.2247 (0.0086)	0.2031 - 0.2500	0.1449 (0.0071)	0.1406 - 0.1875	**0.0192** (0.0573)	0.0000 - 0.3281
20	0.7186 (0.0055)	0.7031 - 0.7500	0.3194 (0.0078)	0.3125 - 0.3437	**0.0092** (0.0218)	0.0000 - 0.3750
50	4.4999 (0.0138)	4.4684 - 4.6090	1.6226 (0.0178)	1.6092 - 2.0780	**0.0114** (0.0070)	0.0000 - 0.0313
100	18.0847 (0.0383)	18.0143 - 8.4518	6.6596 (0.0798)	6.6089 - 8.3744	**0.0180** (0.0067)	0.0000 - 0.0625
	MCD ($\delta = 0.10$)		MVE ($\delta = 0.10$)		CAM ($\delta = 0.10$)	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	0.0984 (0.0101)	0.0937 - 0.1875	**0.0661** (0.0080)	0.0570 - 0.1250	0.1009 (0.1233)	0.0000 - 0.4687
10	0.2195 (0.0040)	0.2031 - 0.2500	0.1413 (0.0034)	0.1250 - 0.1719	**0.0206** (0.0608)	0.0000 - 0.3281
20	0.7180 (0.0068)	0.7031 - 0.7968	0.3201 (0.0080)	0.3125 - 0.3593	**0.0075** (0.0142)	0.0000 - 0.3750
50	4.5194 (0.0161)	4.484 - 4.6715	1.6232 (0.0104)	1.6092 - 1.6561	**0.0100** (0.0080)	0.0000 - 0.0469
100	18.0470 (0.0260)	17.9987 - 8.2330	6.6543 (0.0206)	6.6089 - 6.7964	**0.0149** (0.0077)	0.0000 - 0.0625

Table 9 presents a great quality of the methodology under study. The procedure is extremely competitive in terms of runtime. Note that for 5 variables the times are similar between the three methods, but when the number of variables grows, the computational cost of the CAM method does not present significant growth. The MCD and MVE methods noticeably show a significant increase in execution time as the number of variables grows. Also, it should be noted that the CAM method investigated some different partition quantities k in clusters, therefore, if only a choice of cluster quantities were used the methodology would become even faster.

Table 10: Cpu time (seconds) in detection procedure with Σ_2 correlation matrix.

	MCD ($\delta = 0$)		MVE ($\delta = 0$)		CAM ($\delta = 0$)	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	0.0966 (0.0061)	0.0937 - 0.1094	**0.0645** (0.0054)	0.0625 - 0.0937	0.1965 (0.1117)	0.0000 - 0.2969
10	0.2202 (0.0049)	0.2031 - 0.2656	0.1414 (0.0034)	0.1406 - 0.1562	**0.0066** (0.0334)	0.0000 - 0.3125
20	0.7201 (0.0058)	0.7031 - 0.7656	0.3203 (0.0078)	0.3125 - 0.3437	**0.0042** (0.0185)	0.0000 - 0.3906
50	4.5833 (0.3317)	4.4835 - 4.5558	1.6504 (0.0495)	1.6028 - 2.5572	**0.0060** (0.0211)	0.0000 - 0.6566
100	18.1032 (0.2012)	18.0143 - 21.2784	6.6741 (0.0633)	6.6089 - 7.2723	**0.0103** (0.0505)	0.0000 - 1.1249

(continued from table in previous page)

	MCD ($\delta = 0.02$)		MVE ($\delta = 0.02$)		CAM ($\delta = 0.02$)	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	0.0965 (0.0060)	0.0937 - 0.1094	**0.0647** (0.0055)	0.0625 - 0.0938	0.0979 (0.1195)	0.0000 - 0.2812
10	0.2199 (0.0046)	0.2031 - 0.2500	0.1414 (0.0035)	0.1406 - 0.1719	**0.0243** (0.0683)	0.0000 - 0.3437
20	0.7200 (0.0060)	0.7031 - 0.7500	0.3198 (0.0079)	0.3125 - 0.3437	**0.0091** (0.0140)	0.0000 - 0.3750
50	4.5608 (0.0163)	4.5244 - 4.7053	1.6469 (0.0114)	1.6271 - 1.7670	**0.0127** (0.0026)	0.0080 - 0.0650
100	18.0499 (0.0257)	18.0143 - 18.4361	6.6567 (0.0212)	6.6089 - 6.8276	**0.0187** (0.0065)	0.0000 - 0.0625
	MCD ($\delta = 0.05$)		MVE ($\delta = 0.05$)		CAM ($\delta = 0.05$)	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	0.0966 (0.0061)	0.0937 - 0.1094	**0.0649** (0.0057)	0.0625 - 0.0938	0.0936 (0.1193)	0.0000 - 0.2968
10	0.2195 (0.0041)	0.2031 - 0.2344	0.1415 (0.0035)	0.1406 - 0.1562	**0.0213** (0.0635)	0.0000 - 0.3125
20	0.7190 (0.0057)	0.7031 - 0.7343	0.3196 (0.0078)	0.3125 - 0.3281	**0.0089** (0.0141)	0.0000 - 0.3750
50	4.5972 (0.0740)	4.4842 - 5.2557	1.6586 (0.0304)	1.6066 - 1.7980	**0.0123** (0.0050)	0.0000 - 0.0460
100	18.1093 (0.1743)	18.0143 - 19.8779	6.6657 (0.0609)	6.6089 - 7.1714	**0.0180** (0.0060)	0.0000 - 0.0469
	MCD ($\delta = 0.10$)		MVE ($\delta = 0.10$)		CAM ($\delta = 0.10$)	
var	mean (sd)	min - max	mean (sd)	min - max	mean (sd)	min - max
5	0.0966 (0.0060)	0.0937 - 0.1094	**0.0641** (0.0048)	0.0625 - 0.0938	0.1090 (0.1221)	0.0000 - 0.2969
10	0.2196 (0.0041)	0.2031 - 0.2344	0.1416 (0.0038)	0.1406 - 0.1719	**0.0146** (0.0474)	0.0000 - 0.3125
20	0.7186 (0.0057)	0.7031 - 0.7656	0.3198 (0.0080)	0.3125 - 0.3594	**0.0077** (0.0082)	0.0000 - 0.0469
50	4.5913 (0.0710)	4.4922 - 4.9463	1.6503 (0.0302)	1.6010 - 2.0961	**0.0111** (0.0058)	0.0000 - 0.0469
100	18.0514 (0.0434)	18.0143 - 18.9594	6.6590 (0.0593)	6.6089 - 8.3588	**0.0167** (0.0069)	0.0000 - 0.0469

Table 10 shows that the previous findings from table 9 are actually general and little dependent on the correlation matrix. Again the procedure proves to be competitive when the execution time is evaluated. Dimension growth does not affect the computational cost for the execution of the CAM method. The other methods (MCD and MVE) show a significant increase in execution time as the number of variables increases.

4 Final Remarks

The study presented is broad enough to evaluate the research methodologies. The data simulation procedure through the contaminated normal distribution became more effective with the adaptation of the mean vector and the generation of the most realistic correlation matrix.

The large configurations set of simulation procedure allows the various comparative aspects to be further explored. This enables a fair comparison and allows the potential user to choose some method through the comparative analysis that most closely matches their database under investigation.

The CAM method was more effective in the various comparative forms approached. The accuracy measure, well indicated for such comparison, guarantees the efficiency of the methodology. The results are good even with the significant increase in the number of variables involved in the investigation. Even so, the sensitivity and specificity measurements present very relevant results for the CAM method.

The excellent behavior of the CAM method, even with the increase in the number of variables, without any compromise of the required computational time, is an indication that the technique fits well even for large databases, a subject of great impact today. On the other hand, the alternative methods have their execution time increasing substantially with the growth of the variables number.

The set of results presented illustrates the method's ability to properly diagnose multivariate outliers, as well as non-outliers. And it performs this procedure extremely fast.

The continuity of these studies includes, almost naturally, a better evaluation of possibilities in the criteria for choosing k values through procedures dependent on the data themselves. This can be done through strategies like *machine learning*, or any other kind of data-driven procedure. Adaptive mechanisms for choosing centroids in the $k-$means procedure, as well as the $2.5s$ distance criterion are also subjects that are already under investigation, and which may yield very interesting and useful future results for the statistical community and other researchers.

References

Aggarwal, C. C. (2017). *An Introduction to Outlier Analysis*, Springer International Publishing, pp. 1–34.

Atkinson, A. C., Riani, M. (2002). Forward search added-variable t-tests and the effect of masked outliers on model selection. *Biometrika*, *89*(4), 939–946.

Atkinson, A. C., Riani, M. (2004). The forward search and data visualisation. *Computational Statistics*, *19*(1), 29–54.

Atkinson, A. C., Riani, M., Cerioli, A. (2010). The forward search: Theory and data analysis. *Journal of the korean statistical society*, *39*(2), 117–134.

Barbosa, J. J., Pereira, T. M., Oliveira, F. L. P. (2018). Uma proposta para identificação de outliers multivariados. *Ciência e Natura*, *40*, 1–8.

Barnett, V., Lewis, T. (1994). *Outliers in statistical data*. John Wiley & Sons.

Berton, L., Huertas, J., Araújo, B., Zhao, L. (2010). Identifying abnormal nodes in complex networks by using random walk measure. Em: *IEEE Congress on Evolutionary Computation*, IEEE, pp. 1–6.

Filzmoser, P. (2005). Identification of multivariate outliers: a performance study. *Austrian Journal of Statistics*, *34*(2), 127–138.

Filzmoser, P., Garrett, R., Reimann, C. (2005). Multivariate outlier detection in exploration geochemistry. *Computers & geosciences*, *31*(5), 579–587.

Filzmoser, P., Maronna, R., Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, *52*(3), 1694–1711.

Filzmoser, P., Hron, K., Reimann, C. (2009). Principal component analysis for compositional data with outliers. *Environmetrics: The Official Journal of the International Environmetrics Society*, *20*(6), 621–632.

Hawkins, D. M. (1980). *Identification of outliers*, vol 11. Chapman and Hall.

Jolliffe, I. (2011). *Principal component analysis*, Springer Berlin Heidelberg.

Kutsuna, T., Yamamoto, A. (2017). Outlier detection using binary decision diagrams. *Data mining and knowledge discovery*, *31*(2), 548–572.

Luo, J., Frisken, S., Machado, I., Zhang, M., Pieper, S., Golland, P., Toews, M., Unadkat, P., Sedghi, A., Zhou, H., et al. (2018). Using the variogram for vector outlier screening: application to feature-based image registration. *International journal of computer assisted radiology and surgery*, *13*(12), 1871–1880.

Martins, H. S. R., Duarte, A. R., Oliveira, F. L. P. (2020). Generating custom correlation matrices. *Computational Statistics and Data Analysis (submitted paper)*, pp. 1–20.

Rousseeuw, P. J., Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, *41*(3), 212–223.

Rousseeuw, P. J., Zomeren, B. C. V. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical association*, *85*(411), 633–639.

Valadares, F. G., Aquino, A. L. L., Rabelo, R. A. (2012). Detecção de outliers multivariados em redes de sensores sem fio. Em: *XLIV Simpósio Brasileiro de Pesquisa Operacional*, SBPO.

Van Zoest, V., Stein, A., Hoek, G. (2018). Outlier detection in urban air quality sensor networks. *Water, Air, & Soil Pollution*, *229*(4), 111.

Veloso, M. V. S., Cirillo, M. A. (2016). Principal components in the discrimination of outliers: A study in simulation sample data corrected by pearson's and yates´ s chisquare distance. *Acta Scientiarum Technology*, *38*(2), 193–200.

Zhu, J., Jiang, W., Liu, A., Liu, G., Zhao, L. (2017). Effective and efficient trajectory outlier detection based on time-dependent popular route. *World Wide Web*, *20*(1), 111–134.