Detecting Outliers with a Non-parametric estimation of the Mahalanobis distance

Catalina Piedrahita Jaramillo

Tesis de Maestría

Asesores:
Henry Laniado, Pablo Saldarriaga

Universidad Eafit
Escuela de Ciencias
Maestría en Ciencias de los Datos y Analítica
Medellín

2023

Detecting Outliers with a Non-parametric estimation of the Mahalanobis distance

Catalina Piedrahita[1], Henry Laniado[2], and Pablo Saldarriaga[1]

[1]Area of Computation and Analytics, Eafit University
[2]Mathematics Department, University of Cauca

Abstract

This paper proposes the creation of a robust version of the Mahalanobis distance for the outlier's identification problem, using robust and non-parametric estimations for the covariance matrix, such as Kendall's Tau and Median Absolute Deviation (MAD), as well as techniques that enhance the numerical properties of the covariance matrix to reduce error during numerical calculations like Ledoit and Wolf's Shrinkage. The performance of the methods is evaluated through simulation of independent normal data, correlated normal data, and real data sets and compared with some methods from the literature. The proposed methods achieve a high percentage of correct identification of outliers and have a low false positive rate for both data types, particularly in the case of correlated normal data

1 Introduction

The outlier identification problem is one important aspect to consider when studying a dataset since it is one of the first steps is to conduct for processing the information prior to analyzing data or creating analytical models. This can be done to separate them from the dataset under analysis to avoid biased estimators or errors in model creation, reduce their weight, apply accommodation techniques, or simply study these data points. It is important to note that outliers do not always represent noise; in fact, they are often of interest to the researcher, for example, in cases of financial fraud or when identifying anomalies in medicine. Noise typically refers to incorrect values, such as "999" instead of "99," values that do not make sense in the context of the data or records with incomplete information (Smiti, 2020). However, outliers can provide information about particular individuals or alert unusual behaviors in the dataset.

Statistician Hawkins (1980) defines outliers as "An observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism." On the other hand, Barnett & Lewis (1978) describe them as "An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs."

Various mechanisms have been proposed for identifying these outliers. According to a classification presented by Smiti (2020), we can speak of four categories: those based on statistical properties, those related to distance, those based on density, and those involving clustering. Each technique has advantages and disadvantages related to the problem and the available information. Statistical methods are generally more suitable when the data distribution is known. In contrast, methods based on distance, density, and clustering do not require this information; they evaluate the discordance or similarity between points quantitatively or geometrically. They make no assumptions about the distribution of the data and do not require prior labeling, making them useful for unsupervised datasets.

Multiple authors have defined the strategies based on distance. For example, Chandola *et al.* (2009) present this approach's general idea: "anomaly detection approach, therefore, is to define a region representing normal behavior and declare any observation in the data that does not belong to this normal region as an anomaly." Smiti (2020) adds, "In these methods, outliers are detected by calculating distances among all data objects based on various distance-related metrics. Afterward, objects that have not enough neighbors are most likely to be outliers." Moreover, Han *et al.* (2012)

express, "Given a set of objects in feature space, a distance measure can be used to quantify the similarity between objects. Intuitively, objects that are far from others can be regarded as outliers."

In general terms, there are various distance-based approaches for identifying anomalous data. Some of these are extensively documented in the literature, such as the Mahalanobis distance, which is often one of the most commonly used tools for identifying outlier points in multivariate data. In this type of data, it is crucial to consider the covariance structure when determining similarity. This is because, in multivariate data, both magnitude and shape must be weighed. In this context, it is not appropriate to identify outliers individually in each variable; a point cannot be considered extreme simply for having an anomalous value in one of its features. In multivariate data, the set of variables describing the data must be considered to conclude which are normal and which are outliers. These challenges arise because the data are interconnected, implying a relationship between location, scale, and orientation, as pointed out by Gnanadesikan & Kettenring (1972) "One may not always be able to define or delineate outliers for one of these purposes (e.g., location) without considering some of the others (e.g., dispersion)." This is why the use of Mahalanobis distance stands out, as it is a distance that considers the correlation between variables.

Another of the significant challenges in identifying outliers in multivariate datasets relies on phenomena such as swamping and masking. Swamping occurs when a normal data point is erroneously labeled as an outlier. This phenomenon arises when the presence of distant outliers from the center of the distribution affects the estimates of the mean and covariance. In contrast, masking manifests when outliers mutually conceal each other; this happens because the presence of multiple outliers close to each other distorts the estimates, making these points appear less extreme than they actually are.

Given this context, one of the crucial challenges is estimating a covariance matrix and a centroid that are robust to the presence of outliers. The goal is to calculate a Mahalanobis distance that allows for more accurate identification of outliers. Several strategies have been proposed to tackle this challenge, such as the Fast Algorithm for the Minimum Covariance Determinant Estimator (Fast-MCD) by Rousseeuw & Driessen (1999), which seeks a robust covariance by identifying the data subset of size h that has the minor determinant. Other notable methods include that of Peña & Prieto (2001), which proposes temporarily eliminating data points that produce maximum and minimum kurtosis projections, then calculating the mean and covariance without them to obtain a more robust Mahalanobis distance, and the method of Sajesh & Srinivasan (2012), which recommends using Falk's Comedian to obtain a robust covariance for calculating the Mahalanobis distance.

In summary, the study of robust and non-parametric techniques for estimating Mahalanobis distance and their application in outlier detection strategies are relevant to the fields of statistics and data analysis—areas that this work aims to address. The present article proposes three approaches for estimating covariance using the robust and non-parametric statistics of Kendall's Tau, and the Median Absolute Deviation (MAD), and the Shrinkage technique of Ledoit & Wolf (2003). In the case of Kendall's Tau, Croux & Dehon (2010) point out in their study on correlation measures that this correlation coefficient proves to be a more resistant method to outliers compared to Pearson and Spearman. On the other hand, MAD is a robust measure of dispersion that, according to Hampel (1974), "is consistent and asymptotically normal with respect to the population standard deviation, meaning it converges to the true value as the sample size increases." As for Shrinkage, Ledoit & Wolf (2003) demonstrated how this method can improve the covariance matrix estimation by reducing its variance, which improves the results of statistical inferences and computational calculations. For these reasons, these statistics and methods are notable for their resistance to the presence of outliers and are proposed in this work to obtain a robust covariance and, consequently, a robust Mahalanobis distance which can be used for outlier detection techniques.

This article is organized as follows: The first section is devoted to explaining the techniques of Fast-MCD by Rousseeuw & Driessen (1999), the Kurtosis of Peña & Prieto (2001), and the Comedian approach of Sajesh & Srinivasan (2012); these techniques serve as the foundation for comparisons with the methods being proposed. Subsequently, the proposed combinations using Kendall's Tau, MAD, and Shrinkage for estimating a robust covariance matrix are described, along with constructing a robust Mahalanobis distance, and the proposed strategy for outlier identification is explained. In section 3, all the results obtained through simulations performed with independent normal data and correlated normal data are presented, as well as the results from applying the methods to real data sets. Finally, section 4 offers some conclusions.

2 Outlier Detection Techniques

In this section, the first part covers the existing robust versions found in the literature, followed by an explanation of the proposed process for robust and non-parametric estimation of the Mahalanobis distance and the outlier detection strategy.

2.1 Robust Methods for Mahalanobis Distance Estimation

Such as it was mentioned previously, one of the most common strategies to achieve a robust method for identifying outliers in multivariate cases is creating a robust version of the Mahalanobis distance. This metric is often highly suitable for this type of data, as it considers both the distance and shape of the data. The following will explain three methods that will be used for comparisons.

2.1.1 Fast MCD: The Fast MCD method is based on the Minimum Covariance Determinant (MCD) proposed by Rousseeuw (1985). This method aims to identify a set of observations that allows estimating a covariance matrix with the smallest possible determinant. Since this process is iterative, its main weakness relies on determining the size of the subsample and the number of repetitions required to achieve the smallest determinant. This challenge led to the introduction of a new version of the process by Rousseeuw & Driessen (1999) a Fast Algorithm for the Minimum Covariance Determinant Estimator (Fast MCD). This methodology is available in various statistical packages such as R or Matlab. It is worth noting that when using this method to identify a dataset that allows obtaining a covariance matrix with the minimum determinant, it is also possible to obtain a mean with less influence from outliers. Therefore, using not only the covariance matrix estimated using Fast MCD but also the mean obtained with Fast MCD as the center of mass is common.

2.1.2 Covariance Matrix by Maximum and Minimum Kurtosis Method: The next method to be used for comparison is the one proposed by Peña & Prieto (2001); in this methodology, similar to Fast MCD, the objective is to obtain robust covariance and mean estimates. The method is based on finding data projections that highlight outliers. However, the directions for these projections are strategically selected, and this is where kurtosis comes into play.

The strategy relies on the observation that in univariate datasets, a minority of outliers elevates kurtosis, indicating the importance of investigating directions with maximum kurtosis. Conversely, in the presence of numerous outliers, low kurtosis and bimodality justify exploring directions with minimal kurtosis (Peña *et al.*, 2013).

The process of the method is then finding 'p' directions that are orthogonal and have maximum kurtosis and 'p' orthogonal directions with minimum kurtosis. Next, exceptional data points in these directions are temporarily removed. Subsequently, the mean and covariance matrix are calculated using only the data that is not considered suspicious of being outliers. Finally, outliers are identified as those data points located at the extremes when using the Mahalanobis distance calculated with these outlier-resistant estimates.

We will refer to this method as 'Kurtosis' in the results section for simplicity.

2.1.3 Comedian Approach: Finally, we introduce the method Sajesh & Srinivasan (2012). This method proposes using the Comedian covariance matrix (Falk, 1997) to estimate a robust version of the Mahalanobis distance. The Comedian matrix is calculated using the following equation:

$$COM(X,Y) = median((X - median(X))(Y - median(Y)) \tag{1}$$

Falk (1997) points out that Comedian is a robust covariance measure because it is based on deviations from the median, making it more resistant to outliers.

However, Sajesh & Srinivasan (2012) note that the Comedian matrix is semidefinite non-positive, making it challenging to estimate its inverse. They use a transformation proposed by Maronna & Zamar (2002) to overcome this obstacle, resolving the issue of non-positive semi-definiteness of the Comedian matrix. Thanks to this transformation, they achieve robust estimates for both the location and dispersion of the data.

This approach results in a robust covariance matrix that can be invertible, essential for estimating the Mahalanobis distance. Furthermore, it is important to highlight that Sajesh & Srinivasan (2012)

also apply a robust approach to estimate a resistant mean, which is fundamental to their robust estimation of the Mahalanobis distance. Once these distances are calculated, they suggest a cutoff point for identifying outliers.

The simulation and results section will use the abbreviation SS for this method.

2.2 Covariance Matrix with Kendall's Tau and Shrinkage

To explain how the robust versions of the Mahalanobis distance proposed in this document were generated, their concept and calculation are introduced.

The Mahalanobis distance is a measure between an observation x_i and the center of the data. (Mahalanobis, 1936)

$$MD_i = ((x_i - \hat{\mu})\hat{S}^{-1}(x_i - \hat{\mu})^T)^{1/2} \tag{2}$$

Where $\hat{\mu}$ represents the sample mean and $\hat{S}$ the sample covariance. This study aims to examine a robust version of this distance using estimators that have higher resistance to the presence of outliers.

For a robust estimate, the mean is replaced by the median because it is a better choice regarding robustness (Cabana *et al.*, 2021), and a robust version replaces the covariance matrix.

We have a data matrix X of size n x p, where n is the sample size, and p is the number of variables, the median is.

$$\hat{\mu}_{med} = (median(x_1), \ldots, median(x_p)) \tag{3}$$

Where median indicates the univariate median and $(x_i) = (x_{1i}, \ldots, x_{ni})$ for all $i = 1, \ldots p$ is the ith column of x.

And the covariance between 2 variables can be estimated using their correlation and standard deviations:

$$S(x_i, x_j) = \rho_{x_i, x_j}\sigma_{x_i}\sigma_{x_j} \tag{4}$$

As demonstrated, there exists a connection between covariance and correlation. While Pearson's correlation is a widely adopted measure, it can be adapted to utilize alternative correlation metrics, given its susceptibility to outliers. Chok (2010) delved into the properties of such correlation measures, including Kendall and others, concluding that Pearson correlation is more sensitive to outliers, thus resulting in less powerful statistical tests in cases where datasets exhibit extreme skewness or kurtosis excess, both of which are characteristics of datasets containing outliers.

Therefore, this work proposes a non-parametric indicator such as Kendall's tau. This indicator can be more robust in the presence of outliers, does not assume any data distribution, and does not require independence assumptions. This way, we obtain that:

$$S_K(x_i, x_j) = \rho k_{x_i, x_j}\sigma_{x_i}\sigma_{x_j} \tag{5}$$

Where Kendall's tau is calculated as (Kendall, 1938):

$$\rho k = \frac{2}{n(n-1)}\sum_{i<j} sgn(x_i - x_j)sgn(y_i - y_j) \tag{6}$$

- n is the number of observations in each variable.
- sgn is the sign function, which takes value +1 if the argument is positive, -1 if the argument is negative, and 0 if the argument is zero.
- x_i and y_i are the ranges of variable 1 and variable 2, respectively, for the i-th observation.

This coefficient quantifies the discrepancy between the number of concordant and discordant pairs.

Once the covariance matrix has been estimated using the Kendall coefficient, we proceed to perform shrinkage on the matrix, with the aim of reducing the estimation error compared to maximum likelihood estimation and obtaining a positive definite and well-conditioned matrix. The procedure for performing the shrinkage on the matrix is the one proposed by Ledoit & Wolf (2003).

$$\begin{aligned} \hat{S}_{KSH} &= (1-\eta)\hat{S}_K + \eta\hat{v}_S I, \\ min_{v_S\eta} &= E[||\hat{S}_{sh} - S||^2], \end{aligned} \tag{7}$$

Where $||A||^2 = trace(AA^T)/p$ then we have:

$$v_S = trace(\hat{S}_K)/p \tag{8}$$

$$\eta = \frac{E[||\hat{S}_K - S||^2]}{E[||\hat{S}_K - v_S I||^2]} \tag{9}$$

2.3 Covariance Matrix with Kendall's Tau and MAD

As shown in (4), it is possible to estimate covariance from correlation and standard deviations. To obtain a robust version of this matrix, it is proposed, similar to case 1, to use the Kendall coefficient but replace the standard deviation with the Median Absolute Deviation (MAD). MAD is a dispersion measure based on the median, making it less sensitive to outliers. Both MAD and the median have a breakdown point of 0.5. Several authors have examined the properties of this indicator. For example, Hampel (1974) mentioned: "The median deviation was rediscovered and derived in Contributions to the Theory of Robust Estimation as the M-estimate of scale with the smallest possible gross-error-sensitivity at the normal (and many other)." Furthermore, Leys *et al.* (2013) highlight that the MAD is immune to the sample size.

The second robust covariance proposal with Kendall and MAD is presented below.

$$S_{KM}(x_i, x_j) = \rho k_{x_i,x_j} MAD(x_i) MAD(x_j) \tag{10}$$

Where MAD is defined as (Falk, 1997):

$$MAD(x) = median(|x - median(x)|) \tag{11}$$

2.4 Covariance Matrix with Kendall's Tau, Shrinkage, and MAD

The third proposed method combines all three strategies. It involves applying Ledoit & Wolf (2003) shrinkage to the S_{KM} covariance, which was estimated using Kendall and MAD.

The estimation of this robust covariance version is then as follows:

$$\begin{aligned} \hat{S}_{KSM} &= (1-\eta)\hat{S}_{KM} + \eta\hat{v}_S I, \\ min_{v_S\eta} &= E[||\hat{S}_{sh} - S||^2], \end{aligned} \tag{12}$$

In this case we will have

$$v_S = trace(\hat{S}_{KM})/p \tag{13}$$

$$\eta = \frac{E[||\hat{S}_{KM} - S||^2]}{E[||\hat{S}_{KM} - v_S I||^2]} \tag{14}$$

2.5 Proposed Outliers Detection Strategy

Once the distances have been estimated with the robust versions of Mahalanobis, from each point to the median of the data, a K-means with k=2 is used to classify the data and identify the outliers. The cluster with the least amount of data is considered as the group of outliers.

Table 1 summarizes the methods evaluated in both simulations and real datasets. Additionally, the analysis incorporates the scenario of estimating the standard Mahalanobis distance and classifying data points using the previously described k-means strategy.

Table 1: Proposed combinations for Mahalanobis distance estimation

Parameter	**KSH**	**KM**	**KSM**	**ME**
Centroid	Median	Median	Median	Mean
Covariance	$\hat{S}_{KSH}$	$\hat{S}_{KM}$	$\hat{S}_{KSM}$	$\hat{S}$

3 Computational Experiments

Next, we detail the procedures for generating independent and correlated datasets simulations with a normal distribution. We present the obtained results and corresponding analyses, comparing the performance of the proposed methods with the reference methods from the literature. We used the F-score indicator to evaluate the methods' performance, which is a harmonic mean between precision and recall. This means that in its scoring it provides a balance between a model's ability to classify positive cases correctly and its ability to avoid false negatives, which is crucial when evaluating models in cases of class imbalance, a characteristic present in these simulations where datasets with different contamination levels were generated: 10%, 20%, and 30%.

For this analysis, the formula used for the F-score is as follows, where equal weight was given to precision and recall.

$$F - score = \frac{2 \times precision \times recall}{precision + recall} \tag{15}$$

The results presented correspond to the average F-score obtained in each scenario. Each scenario was run 50 times, and the F-score was estimated for each of these runs. The tables and graphs shown here reflect the average F-score from the 50 runs of each scenario.

Finally, after the analysis with simulated data, the methods were evaluated using real datasets. In addition to the F-score, the confusion matrix was included as a supplementary metric to assess the methods' performance in this stage.

3.1 Simulated Datasets

3.1.1 Normal Distributed Independent: A normal mix was simulated using the equation on the form $(1-\alpha)N(0,1) + \alpha N(\delta e, \lambda I)$ where α is the level of contamination, δ is the distance of outliers, λ is the concentration of contamination, and e denotes the p-dimensional vector of ones. This model is similar to the used for the study of cases in Peña & Prieto (2001) and Sajesh & Srinivasan (2012).

For the simulations, the values used were: $\alpha \in [0.1, 0.2, 0.3]$, $\delta = 5$, $\lambda \in [0.1, 1]$. The studies were done for cases with p dimension 5 until 50, using data sizes n=500 and 1000. Moreover, each scenario was executed 50 times.

The figure 1 shows a scatterplot of the data to give an example of how it looks when p=2, n=1000, contamination $\alpha = 0.3$, concentration $\lambda = 1$ and 0.1, and distance of outliers $\delta = 5$.

Figure 1 Data example generated with p=2 using two normal mix.

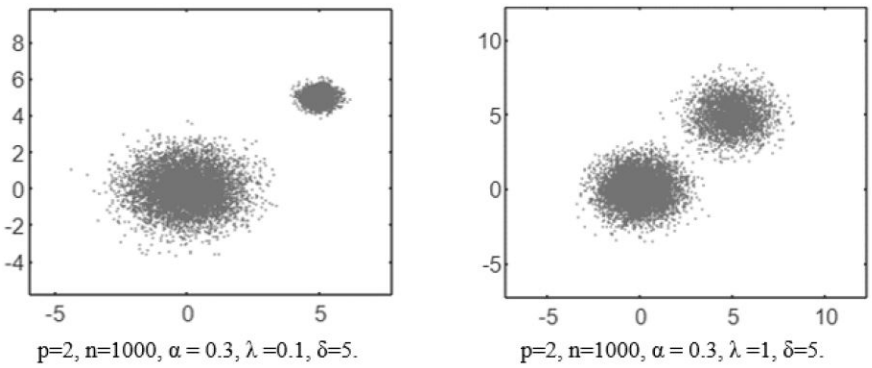
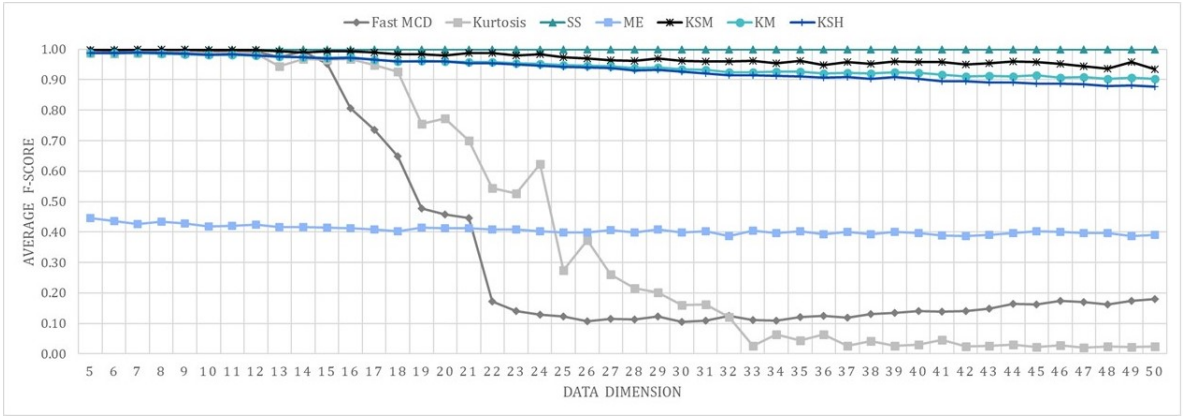


Figure 2 Average F-score of 50 executions with a mixture of normal with $\alpha = 0.3$, n=500, $\lambda = 1$, $\delta = 5$.



It can be observed in figure 2 that the proposed methods KSM, KM, and KSH for estimating covariance for Mahalanobis distance yield excellent results, with a high f-score of around 0.9 for cases with high dimensions like p=50. Among these techniques, the one that combines Kendall, Shrinkage, and MAD (KSM) got the best results consistently, performing between 0.9 and 1 across various dimension

cases. However, it is not superior to the SS method. On the other hand, we can see that Fast-MCD and Kurtosis perform poorly as the data dimension increases. ME consistently yields results around 0.4 for all dimensions, concluding that the method does not successfully identify the outliers.

Table 2 shows that the methods SS, KSM, KM, and KSH exhibit the best performance for all contamination levels in terms of the f-score. Specifically, when the contamination is 10% and 20%, these four techniques display similar behavior, maintaining a high level of performance. Even at a contamination level of 30%, KSM, KM, and KSH only marginally underperform SS, which still retains a high level of effectiveness. On the other hand, in the Fast-MCD and Kurtosis methods, their indicator rapidly decreases as the number of variables increases and high levels of contamination. The method ME consistently exhibits poor performance across all scenarios, regardless of contamination level.

Table 2: Average F-score with independent normal data n=500, $\lambda = 1$, $\delta = 5$.

p	Alpha	Fast MCD	Kurtosis	SS	ME	KSM	KM	KSH
5	0.1	0.9	0.9	1.0	0.6	1.0	1.0	1.0
	0.2	1.0	1.0	1.0	0.5	1.0	1.0	1.0
	0.3	1.0	1.0	1.0	0.4	1.0	1.0	1.0
10	0.1	0.9	1.0	1.0	0.5	1.0	1.0	1.0
	0.2	1.0	1.0	1.0	0.4	1.0	1.0	1.0
	0.3	1.0	1.0	1.0	0.4	1.0	1.0	1.0
15	0.1	0.9	1.0	1.0	0.4	1.0	1.0	1.0
	0.2	1.0	0.9	1.0	0.4	1.0	1.0	1.0
	0.3	1.0	1.0	1.0	0.4	1.0	1.0	1.0
20	0.1	0.9	1.0	1.0	0.3	1.0	1.0	1.0
	0.2	1.0	0.3	1.0	0.4	1.0	1.0	1.0
	0.3	0.5	0.8	1.0	0.4	1.0	1.0	1.0
30	0.1	0.9	0.9	1.0	0.3	1.0	1.0	1.0
	0.2	0.9	0.1	1.0	0.4	1.0	1.0	1.0
	0.3	0.1	0.2	1.0	0.4	0.9	0.9	0.9
40	0.1	0.8	0.5	1.0	0.3	1.0	1.0	1.0
	0.2	0.4	0.0	1.0	0.4	1.0	1.0	1.0
	0.3	0.1	0.0	1.0	0.4	0.9	0.9	0.9
50	0.1	0.7	0.3	1.0	0.3	1.0	1.0	1.0
	0.2	0.2	0.0	1.0	0.3	1.0	1.0	1.0
	0.3	0.2	0.0	1.0	0.4	0.9	0.9	0.9

Figure 3 shows that for the cases when the concentration of contamination is $\lambda = 0.1$ and the level of contamination $\alpha = 0.3$, Fast-MCD cannot identify the outlier for any data dimension, KSM, KM, and KSH have good results for data dimensions between 5 until 17. After that, the performance of the method decreases. For Kurtosis, the good performance is until dimension 32, where we can get an f-score over 0.8. Then, the indicator for this method is less when data increases. In contrast, for SS, the situation is oppositive at the beginning for low dimensions between 5 and 7; the f-score is low under 0.8. However, once the dimension starts increasing, the method successfully identifies the outliers. Similarly, the behavior for ME methods also increases the f-score when dimension increases. It is getting great results when the data size is from 19 to 50.

For this scenario on table 3 where the outliers are less dispersed, the proposed methods of KSM, KM, and KSH correctly identify outliers when the data dimension is less than p=15, and their effectiveness is maintained in larger data sizes up to p=50 if the percentage of contamination is low (10%). However, as the contamination level and data size increase, these techniques show reduced performance in precision and sensitivity. Whereas, for the SS and Kurtosis methods, it is observed that the f-score is 1 in most scenarios, indicating their efficiency in identifying outliers, especially when the contamination percentage is less than 20%. Although, in scenarios with higher contamination (30%), SS is weak when the data size is less than 10, and Kurtosis is weak when the data size is bigger than 30. ME performs poorly in scenarios where the data size is p=20 or less for all contamination levels reviewed. Nevertheless, its capability increases when the data size is bigger than p=20.

Figure 4 presents an overview of the f-score for a particular scenario involving a mixture of normal distribution with $\alpha = 0.3$, n=1000, $\lambda = 1$, and $\delta = 5$. In this case, like the n=500 scenario, SS exhibits the highest performance on the f-score, followed by KSM, KM, and KSH. In the method proposed

Figure 3 Average F-score of 50 executions with a mixture of normal with $\alpha = 0.3$, n=500, $\lambda = 0.1$, $\delta = 5$.

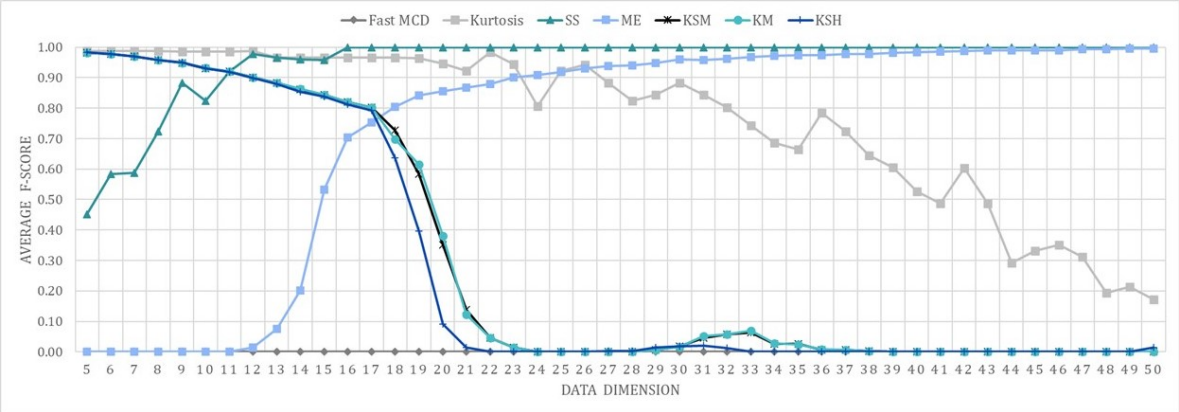


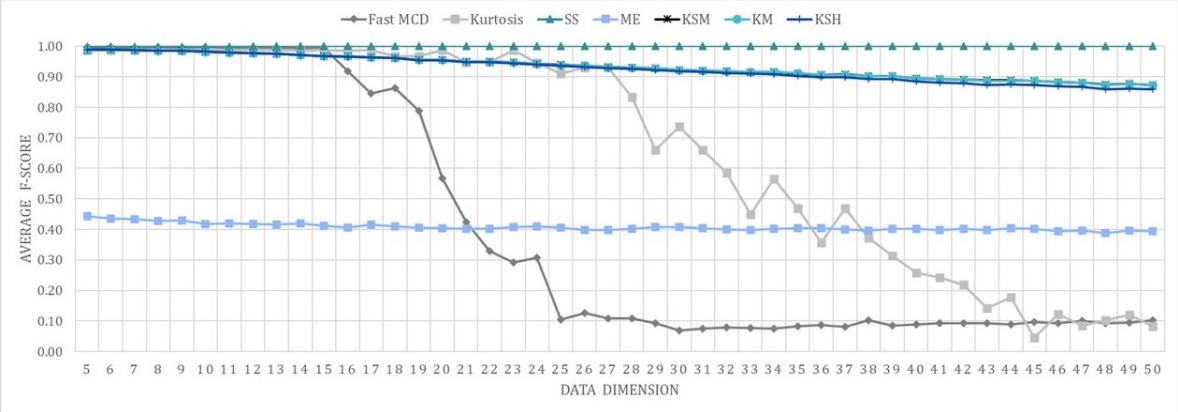
Table 3: Average F-score with independent normal data n=500, $\lambda = 0.1$, $\delta = 5$.

p	Alpha	Fast MCD	Kurtosis	SS	ME	KSM	KM	KSH
5	0.1	0.9	1.0	1.0	0.5	1.0	1.0	1.0
	0.2	1.0	1.0	1.0	0.0	1.0	1.0	1.0
	0.3	0.0	1.0	0.5	0.0	1.0	1.0	1.0
10	0.1	0.9	1.0	1.0	0.0	1.0	1.0	1.0
	0.2	0.0	1.0	1.0	0.0	1.0	1.0	1.0
	0.3	0.0	1.0	0.8	0.0	0.9	0.9	0.9
15	0.1	0.9	0.9	1.0	0.0	1.0	1.0	1.0
	0.2	0.0	1.0	1.0	0.0	1.0	1.0	1.0
	0.3	0.0	1.0	1.0	0.5	0.8	0.8	0.8
20	0.1	0.0	1.0	1.0	0.0	1.0	1.0	1.0
	0.2	0.0	1.0	1.0	0.6	0.9	0.9	0.9
	0.3	0.0	0.9	1.0	0.9	0.4	0.4	0.1
30	0.1	0.0	1.0	1.0	0.3	1.0	1.0	1.0
	0.2	0.0	1.0	1.0	0.9	0.8	0.8	0.7
	0.3	0.0	0.9	1.0	1.0	0.0	0.0	0.0
40	0.1	0.0	0.9	1.0	0.5	1.0	1.0	1.0
	0.2	0.0	0.9	1.0	1.0	0.6	0.6	0.4
	0.3	0.0	0.5	1.0	1.0	0.0	0.0	0.0
50	0.1	0.0	0.9	1.0	0.8	1.0	1.0	1.0
	0.2	0.0	0.9	1.0	1.0	0.0	0.0	0.0
	0.3	0.0	0.2	1.0	1.0	0.0	0.0	0.0

by Sajesh and Srinivasan (2012), which utilizes a comedian approach, the identification of outliers is consistently successful when the data is distributed independently as normal. Regarding the proposed methods, it is noteworthy that their performance on the f-score remains consistently high, over 0.8, across all dimensions. However, there is a decrease in performance as the data dimension increases. The figure reveals interesting patterns by examining the other techniques evaluated in this analysis. Fast-MCD demonstrates a faster performance decrease in this scenario compared to the case with n=500. Conversely, for Kurtosis, the opposite trend emerges, where the performance of the f-score decreases with increasing data size but at a slower rate here compared to the case when n=500. As for ME, the results remain consistent with the scenario involving n=500, with no significant changes observed when all other variables are held constant.

Table 4 summarizes the f-score from simulating an independent normal mixture with $\lambda = 1$, $\delta = 5$, and n = 1000 using different dimensions and contamination levels. In this case, when the dispersion of outliers is higher compared to the scenario with $\lambda = 0.1$, the proposed methods of KSM, KM, and KSH achieve better precision and sensitivity in identifying the atypical data. The performance

Figure 4 Average F-score of 50 executions with a mixture of normal with $\alpha = 0.3$, n=1000, $\lambda = 1$, $\delta = 5$.



of these methods is highly comparable to that of the SS technique. Conversely, the effectiveness of the Fast-MCD and Kurtosis methods diminishes in identifying outliers as the data dimension and contamination percentage increase.

Table 4: Average F-score with independent normal data n=1000, $\lambda = 1$, $\delta = 5$.

p	Alpha	Fast MCD	Kurtosis	SS	ME	KSM	KM	KSH
5	0.1	1.0	1.0	1.0	0.6	1.0	1.0	1.0
	0.2	1.0	1.0	1.0	0.5	1.0	1.0	1.0
	0.3	1.0	1.0	1.0	0.4	1.0	1.0	1.0
10	0.1	0.9	1.0	1.0	0.5	1.0	1.0	1.0
	0.2	1.0	1.0	1.0	0.4	1.0	1.0	1.0
	0.3	1.0	1.0	1.0	0.4	1.0	1.0	1.0
15	0.1	0.9	0.9	1.0	0.4	1.0	1.0	1.0
	0.2	1.0	1.0	1.0	0.4	1.0	1.0	1.0
	0.3	1.0	1.0	1.0	0.4	1.0	1.0	1.0
20	0.1	0.9	1.0	1.0	0.4	1.0	1.0	1.0
	0.2	1.0	0.8	1.0	0.4	1.0	1.0	1.0
	0.3	0.6	1.0	1.0	0.4	1.0	1.0	1.0
30	0.1	0.9	1.0	1.0	0.3	1.0	1.0	1.0
	0.2	1.0	0.1	1.0	0.4	1.0	1.0	1.0
	0.3	0.1	0.7	1.0	0.4	0.9	0.9	0.9
40	0.1	0.9	0.9	1.0	0.3	1.0	1.0	1.0
	0.2	0.6	0.0	1.0	0.4	1.0	1.0	1.0
	0.3	0.1	0.3	1.0	0.4	0.9	0.9	0.9
50	0.1	0.9	0.8	1.0	0.3	1.0	1.0	1.0
	0.2	0.2	0.0	1.0	0.3	1.0	1.0	1.0
	0.3	0.1	0.1	1.0	0.4	0.9	0.9	0.9

In figure 5 similar to figure 3, it is evident that Fast-MCD fails to identify the outliers for any data dimension, indicating limitations in this method when dealing with scenarios featuring low dispersion of atypical points. The performance of the proposed methods, KSM, KM, and KSH, as depicted in Figures 3 and 5, remains consistent, suggesting that the data size does not significantly impact the outcomes of these techniques. Notably, these methods exhibit commendable outlier identification capabilities when the data dimension is less than p=17. In contrast, the performance of SS techniques shows an upward trend with increasing data dimension, a trend also observed in the case of ME. Conversely, for Kurtosis, the situation is reversed; it performs well until p=32, with the f-score exceeding 0.8. However, the method's performance gradually diminishes beyond this point with data dimension. Nevertheless, this decline is comparatively slower than the scenario with the same variables but n=500, as it remains above 0.4.

Figure 5 Average F-score of 50 executions with a mixture of normal with $\alpha = 0.3$, n=1000, $\lambda = 0.1$, $\delta = 5$.

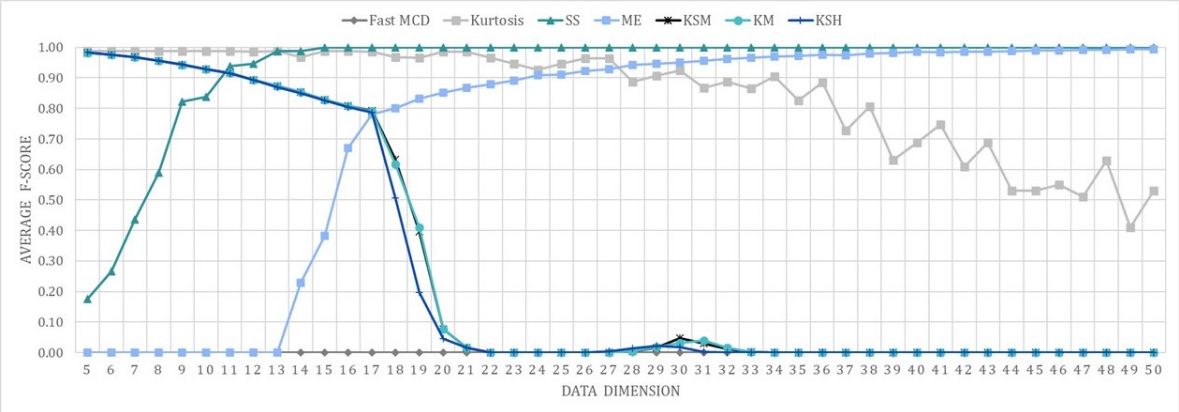


Table 5 consolidates the results of simulations involving independent normal data with less dispersed outliers ($\lambda = 0.1$). Here, it is noteworthy that the proposed KSM, KM, and KSH methods successfully identify outliers in cases where the contamination percentage is low (10%). However, as this contamination percentage and the data dimension increase, the effectiveness of these methods diminishes. Conversely, in this scenario, SS stands out as the method that achieves the best performance. ME shows improved effectiveness as the data dimension increases. Kurtosis also exhibits high effectiveness in this scenario, particularly for cases with dimensions ranging from p=5 to p=40. On the other hand, the Fast-MCD technique demonstrates the least favorable performance in this scenario.

Table 5: Average F-score with independent normal data n=1000, $\lambda = 0.1$, $\delta = 5$.

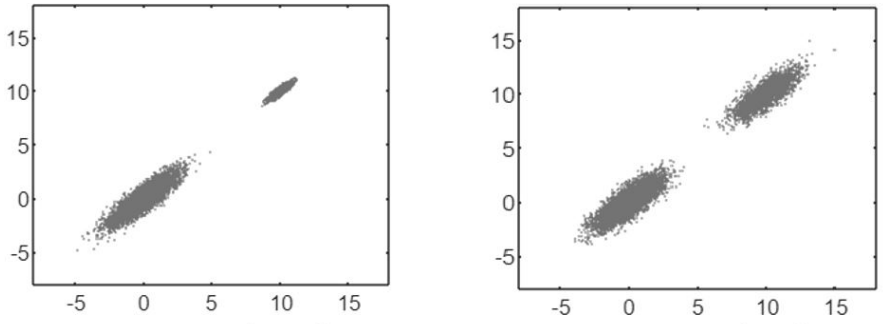
p	Alpha	Fast MCD	Kurtosis	SS	ME	KSM	KM	KSH
5	0.1	0.9	1.0	1.0	0.5	1.0	1.0	1.0
	0.2	1.0	1.0	1.0	0.0	1.0	1.0	1.0
	0.3	0.0	1.0	0.2	0.0	1.0	1.0	1.0
10	0.1	0.9	1.0	1.0	0.0	1.0	1.0	1.0
	0.2	0.0	1.0	1.0	0.0	1.0	1.0	1.0
	0.3	0.0	1.0	0.8	0.0	0.9	0.9	0.9
15	0.1	0.9	1.0	1.0	0.0	1.0	1.0	1.0
	0.2	0.0	1.0	1.0	0.0	1.0	1.0	1.0
	0.3	0.0	1.0	1.0	0.4	0.8	0.8	0.8
20	0.1	0.0	1.0	1.0	0.0	1.0	1.0	1.0
	0.2	0.0	1.0	1.0	0.6	0.9	0.9	0.9
	0.3	0.0	1.0	1.0	0.9	0.1	0.1	0.0
30	0.1	0.0	1.0	1.0	0.3	1.0	1.0	1.0
	0.2	0.0	1.0	1.0	0.9	0.7	0.7	0.7
	0.3	0.0	0.9	1.0	1.0	0.0	0.0	0.0
40	0.1	0.0	0.9	1.0	0.5	1.0	1.0	1.0
	0.2	0.0	0.9	1.0	0.9	0.3	0.3	0.2
	0.3	0.0	0.7	1.0	1.0	0.0	0.0	0.0
50	0.1	0.0	1.0	1.0	0.8	1.0	1.0	1.0
	0.2	0.0	0.9	1.0	1.0	0.0	0.0	0.0
	0.3	0.0	0.5	1.0	1.0	0.0	0.0	0.0

3.1.2 Normal Distributed Correlated: A covariance matrix was simulated for this scenario to get data with high correlation. First, a matrix with p dimension was created using values between 0.1 and 0.65 and multiple by its transposed to obtain a positive semi-definite and symmetric matrix. Using the equation on the form $(1-\alpha)N(0,S)+\alpha N(\delta e, \lambda S)$ was simulated a normal mix. Where α is the level of contamination, δ is the distance of outliers, λ is the concentration of contamination, and e denotes the p-dimensional.

The values used were: $\alpha \in [0.1, 0.2, 0.3]$, $\delta = 10$, $\lambda \in [0.1, 1]$. The studies were done for cases with p dimension 5 until 50, using data sizes n=500 and 1000. Moreover, each scenario was executed 50 times.

Figure 6 shows an example of how the data simulated for dimension p=2, where it can be highlighted how the data has a high correlation between then, both the dataset and the outliers.

Figure 6 Data example generated with p=2 using a covariance matrix for high correlation and two normal mix.



p=2, n=1000, α = 0.3, λ =0.1, δ=10. p=2, n=1000, α = 0.3, λ =1, δ=10.

Figure 7 illustrates that Fast MCD, Kurtosis, and SS performance declines faster than KSM, KM, and KSH as data dimension increases. In this scenario, the proposed methods demonstrate the best results, exhibiting superior performance for KSH up to p=32 and for KM and KSH up to p=44. Among the other techniques with lower efficiency, Kurtosis maintains a relatively slow decline in f-score compared to SS between p=5 and p=26. After that, SS declines slower than Kurtosis and Fast MCD until p=37, where Fast MCD outperforms Kurtosis and SS, achieving a better f-score. Lastly, ME consistently achieves a low f-score of approximately 0.4 across all data dimensions in the simulation.

Figure 7 Average F-score of 50 executions with a mixture of normal with high correlation $\alpha = 0.3$, n=500, $\lambda = 1$, $\delta = 10$.

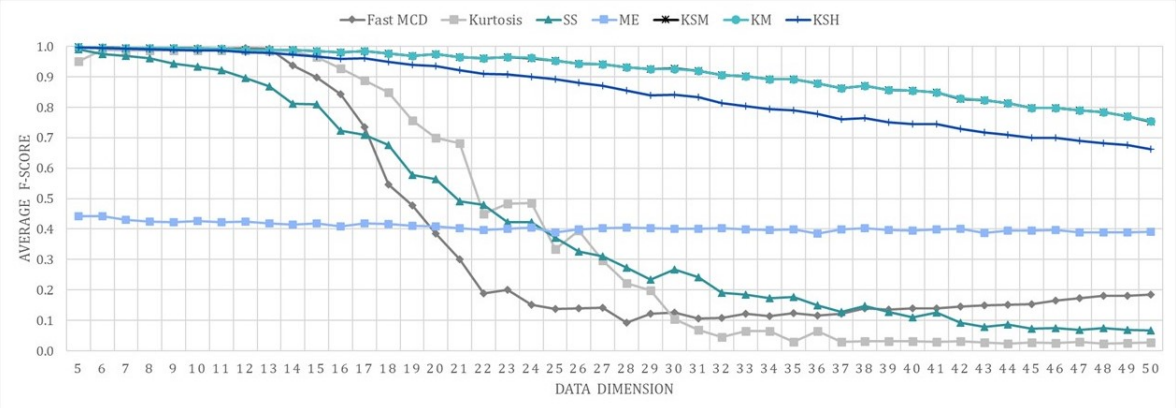


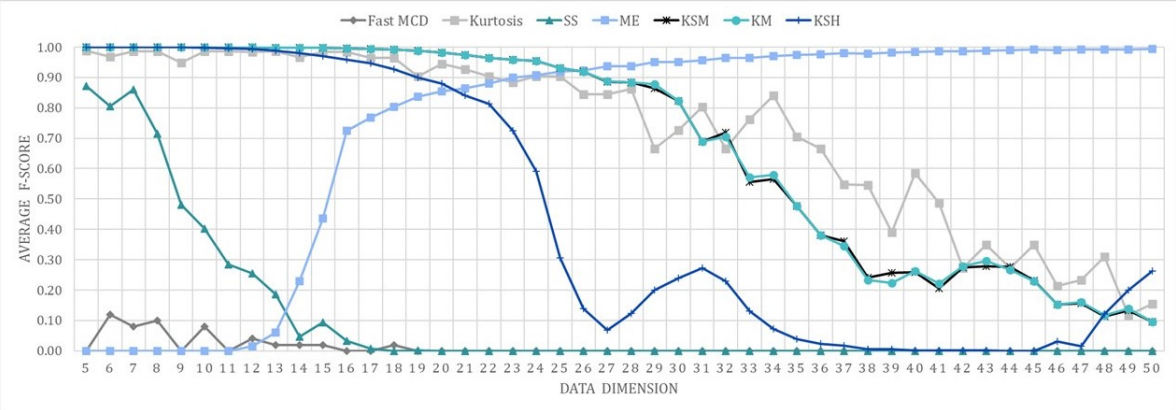
Table 6 reveals that the f-scores of the methods KSM, KM, and KSH demonstrate better outlier identification across multiple data sizes and contamination percentages, with f-scores around 1 and 0.9, indicating a good balance between precision and recall. Moreover, both KSM and KM exhibit the same trend in the f-scores. However, for the other methods, Fast MCD, Kurtosis, and SS, their effectiveness is significant when the data dimension is lower than p=20. Beyond this threshold, their performance diminishes not only due to the percentage of contamination but also as data size increases. On the contrary, the ME approach consistently demonstrates low scores for all scenarios, yielding results around 0.4 in all cases.

In figure 8, we compare the average f-scores of simulations from a scenario with a higher concentration of outliers, unlike in figure 7. Generally, as the data dimension increases, most methods struggle to identify atypical points effectively. However, the ME method stands out as the only one that improves its performance under this condition. For dimensions between p=5 and p=26, the best results are consistently obtained with KSM, KM, KSH, and Kurtosis. In fact, KSM and KM achieve an f-score above 0.8 up to p=30, though beyond p=26, ME becomes the most effective method.

Table 6: Average F-score with correlated normal data n=500, $\lambda = 1$, $\delta = 10$.

p	Alpha	Fast MCD	Kurtosis	SS	ME	KSM	KM	KSH
5	0.1	0.9	1.0	1.0	0.6	1.0	1.0	1.0
	0.2	1.0	1.0	1.0	0.5	1.0	1.0	1.0
	0.3	1.0	1.0	1.0	0.4	1.0	1.0	1.0
10	0.1	0.9	0.9	1.0	0.4	1.0	1.0	1.0
	0.2	1.0	1.0	1.0	0.4	1.0	1.0	1.0
	0.3	1.0	1.0	0.9	0.4	1.0	1.0	1.0
15	0.1	0.9	0.9	1.0	0.4	1.0	1.0	1.0
	0.2	1.0	0.9	1.0	0.4	1.0	1.0	1.0
	0.3	0.9	1.0	0.8	0.4	1.0	1.0	1.0
20	0.1	0.9	0.9	1.0	0.4	1.0	1.0	1.0
	0.2	1.0	0.3	0.9	0.4	1.0	1.0	1.0
	0.3	0.4	0.7	0.6	0.4	1.0	1.0	0.9
30	0.1	0.9	0.9	0.9	0.3	1.0	1.0	0.9
	0.2	0.9	0.0	0.8	0.4	0.9	0.9	0.9
	0.3	0.1	0.1	0.3	0.4	0.9	0.9	0.8
40	0.1	0.8	0.4	0.9	0.3	0.9	0.9	0.8
	0.2	0.3	0.0	0.6	0.4	0.9	0.9	0.8
	0.3	0.1	0.0	0.1	0.4	0.9	0.9	0.7
50	0.1	0.7	0.3	0.7	0.3	0.8	0.8	0.6
	0.2	0.2	0.0	0.3	0.3	0.8	0.8	0.7
	0.3	0.2	0.0	0.1	0.4	0.8	0.8	0.7

Figure 8 Average F-score of 50 executions with a mixture of normal with high correlation $\alpha = 0.3$, n=500, $\lambda = 0.1$, $\delta = 10$.



From table 7, we can see that the KSM and KM methods show high efficiency in identifying outliers for all simulated contamination percentages up to p=30. However, beyond this point, the f-score decreases in all scenarios. For the KSH method, efficiency with different contamination levels extends only up to p=20. Kurtosis is another technique that exhibits good performance in this scenario, achieving effectiveness in identifying atypical data up to p=40, except where the contamination percentage increases to 30%. The ME method proves to be highly effective for identifying outliers in high-dimensional databases. The f-score is high for contamination cases of 30% starting from p=20. On the other hand, the Fast MCD and SS methods do not seem to work very well for correlated data. Fast MCD shows an f-score close to zero in almost all scenarios, and SS has very low values starting from p=15.

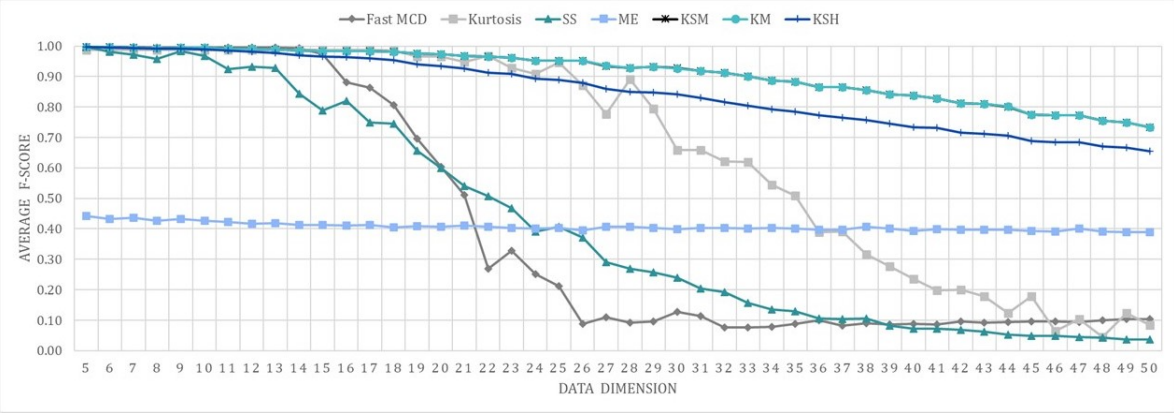
In figure 9, the best performance for most dimensions is achieved with KM and KMS. These methods exhibit the same trend and effectively identify outliers until p=42. The second-best performance is observed with KSH, performing an f-score over 0.8 until p=33. Kurtosis also shows good performance until p=29. On the other hand, Fast MCD and SS decrease their capability in this scenario faster

Table 7: Average F-score with correlated normal data n=500, $\lambda = 0.1$, $\delta = 10$.

p	Alpha	Fast MCD	Kurtosis	SS	ME	KSM	KM	KSH
5	0.1	0.9	1.0	1.0	0.5	1.0	1.0	1.0
	0.2	1.0	1.0	1.0	0.0	1.0	1.0	1.0
	0.3	0.0	1.0	0.9	0.0	1.0	1.0	1.0
10	0.1	0.9	0.9	0.9	0.0	1.0	1.0	1.0
	0.2	0.2	1.0	1.0	0.0	1.0	1.0	1.0
	0.3	0.1	1.0	0.4	0.0	1.0	1.0	1.0
15	0.1	0.6	0.9	1.0	0.0	1.0	1.0	1.0
	0.2	0.1	1.0	0.9	0.0	1.0	1.0	1.0
	0.3	0.0	1.0	0.1	0.4	1.0	1.0	1.0
20	0.1	0.4	0.9	1.0	0.0	1.0	1.0	1.0
	0.2	0.0	1.0	0.5	0.6	1.0	1.0	0.9
	0.3	0.0	0.9	0.0	0.9	1.0	1.0	0.9
30	0.1	0.2	0.9	0.8	0.3	0.9	0.9	0.7
	0.2	0.0	0.9	0.0	0.9	0.9	0.9	0.6
	0.3	0.0	0.7	0.0	1.0	0.8	0.8	0.2
40	0.1	0.1	0.9	0.2	0.5	0.6	0.6	0.4
	0.2	0.0	0.9	0.0	1.0	0.6	0.6	0.1
	0.3	0.0	0.6	0.0	1.0	0.3	0.3	0.0
50	0.1	0.0	0.2	0.0	1.0	0.1	0.1	0.3
	0.2	0.0	0.8	0.0	1.0	0.2	0.2	0.0
	0.3	0.0	0.2	0.0	1.0	0.1	0.1	0.3

than the other mentioned methods. ME, in contrast, does not perform accurate outlier identification in any dimension.

Figure 9 Average F-score of 50 executions with a mixture of normal with high correlation $\alpha = 0.3$, n=1000, $\lambda = 1$, $\delta = 10$.



With table 8, we can corroborate part of the results observed in Figure 9. KSM and KM are the methods that achieve better precision and recall for identifying atypical data in this scenario. KSH shows an f-score around 0.8 up to p=40. Among the other techniques used for comparisons, Fast MCD, Kurtosis, and SS, it is evident that after p=20, their performance declines rapidly for high contamination levels like 20%. The ME method does not successfully identify atypical points for scenarios where $\lambda = 1$, as evidenced in both independent normal data and cases with correlated data.

In figure 10, we can appreciate the trend from the previous scenarios with $\lambda = 0.1$, where ME achieves higher accuracy in identifying outliers as the data dimension increases. Among the other three proposed methods, the best performers are KSM and KM, with an f-score above 0.8 up to p=29. Notably, these two methods consistently exhibit the same trend for correlated data. KSH shows good performance up to p=21, but its effectiveness declines afterward. Kurtosis is a method to highlight here, as it maintains an f-score above 0.4 in both scenarios, correlated data, and independent data. Finally, SS and Fast MCD are the methods that have the worst performance in this scenario, especially

Table 8: Average F-score with correlated normal data n=1000, $\lambda = 1$, $\delta = 10$.

p	Alpha	Fast MCD	Kurtosis	SS	ME	KSM	KM	KSH
5	0.1	1.0	1.0	1.0	0.6	1.0	1.0	1.0
	0.2	1.0	1.0	1.0	0.5	1.0	1.0	1.0
	0.3	1.0	1.0	1.0	0.4	1.0	1.0	1.0
10	0.1	0.9	0.9	1.0	0.5	1.0	1.0	1.0
	0.2	1.0	1.0	1.0	0.4	1.0	1.0	1.0
	0.3	1.0	1.0	1.0	0.4	1.0	1.0	1.0
15	0.1	0.9	1.0	1.0	0.4	1.0	1.0	1.0
	0.2	1.0	1.0	1.0	0.4	1.0	1.0	1.0
	0.3	1.0	1.0	0.8	0.4	1.0	1.0	1.0
20	0.1	0.9	1.0	1.0	0.4	1.0	1.0	1.0
	0.2	1.0	0.8	1.0	0.4	1.0	1.0	1.0
	0.3	0.6	1.0	0.6	0.4	1.0	1.0	0.9
30	0.1	0.9	1.0	1.0	0.3	1.0	1.0	0.9
	0.2	0.9	0.2	0.8	0.4	0.9	0.9	0.9
	0.3	0.1	0.7	0.2	0.4	0.9	0.9	0.8
40	0.1	0.9	0.9	0.9	0.3	0.9	0.9	0.8
	0.2	0.6	0.0	0.6	0.4	0.9	0.9	0.8
	0.3	0.1	0.2	0.1	0.4	0.8	0.8	0.7
50	0.1	0.9	0.7	0.7	0.3	0.8	0.8	0.6
	0.2	0.2	0.0	0.3	0.3	0.8	0.8	0.7
	0.3	0.1	0.1	0.0	0.4	0.7	0.7	0.7

Fast MCD.

Figure 10 Average F-score of 50 executions with a mixture of normal with high correlation $\alpha = 0.3$, n=1000, $\lambda = 0.1$, $\delta = 10$.

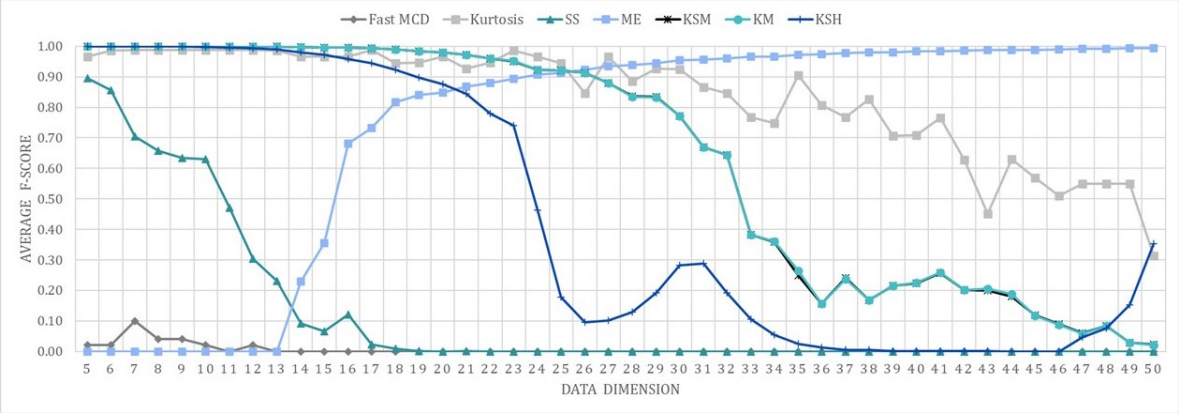


Table 9 shows the consistent trend in the f-score for KSM and KM. These methods effectively identify outliers across scenarios with different contamination levels up to p=30. KSH also exhibits this favorable behavior until p=20, regardless of the contamination percentages. Subsequently, we find ME, which proves to be a robust method as p increases, for 20% and 30% contamination levels. Next, we have Kurtosis, which shows to be a highly effective method in this scenario, successfully identifying outliers in almost all scenarios up to p=40 and, for p=50, when contamination is 10% or 20%. Finally, SS and Fast MCD do not appear to be very effective methods for scenarios with data exhibiting high correlations.

Table 9: Average F-score with correlated normal data n=1000, $\lambda = 0.1$, $\delta = 10$.

p	Alpha	Fast MCD	Kurtosis	SS	ME	KSM	KM	KSH
5	0.1	1.0	1.0	1.0	0.5	1.0	1.0	1.0
	0.2	1.0	1.0	1.0	0.0	1.0	1.0	1.0
	0.3	0.0	1.0	0.9	0.0	1.0	1.0	1.0
10	0.1	0.9	0.9	1.0	0.0	1.0	1.0	1.0
	0.2	0.3	1.0	1.0	0.0	1.0	1.0	1.0
	0.3	0.0	1.0	0.6	0.0	1.0	1.0	1.0
15	0.1	0.7	1.0	1.0	0.0	1.0	1.0	1.0
	0.2	0.1	1.0	0.9	0.0	1.0	1.0	1.0
	0.3	0.0	1.0	0.1	0.4	1.0	1.0	1.0
20	0.1	0.4	1.0	1.0	0.0	1.0	1.0	1.0
	0.2	0.0	1.0	0.8	0.6	1.0	1.0	0.9
	0.3	0.0	1.0	0.0	0.8	1.0	1.0	0.9
30	0.1	0.1	1.0	0.8	0.3	0.9	0.9	0.7
	0.2	0.0	1.0	0.1	0.8	0.9	0.9	0.6
	0.3	0.0	0.9	0.0	1.0	0.8	0.8	0.3
40	0.1	0.0	0.9	0.3	0.4	0.5	0.5	0.4
	0.2	0.0	0.9	0.0	1.0	0.6	0.5	0.1
	0.3	0.0	0.7	0.0	1.0	0.2	0.2	0.0
50	0.1	0.0	0.9	0.0	0.8	0.3	0.3	0.1
	0.2	0.0	0.8	0.0	1.0	0.1	0.1	0.0
	0.3	0.0	0.3	0.0	1.0	0.0	0.0	0.4

3.2 Real Datasets

Three real-world datasets have been used to validate the performance of the proposed methods and make comparisons. These datasets vary in size and levels of contamination; they are the Breast Cancer Wisconsin dataset (Wolberg, 1992), the Wine dataset (Aeberhard & Forina, 1991), and the Ionosphere dataset (Sigillito & Baker, 1989). These datasets are publicly available in the UCI Machine Learning Repository and have been used in other analyses similar to the one performed here. Detailed descriptions and results for each dataset are presented below.

The Breast Cancer Wisconsin dataset (Wolberg, 1992). It is a classification dataset with 9 dimensions and discrete variables, containing 683 data points. The dataset comprises two classes, benign and malignant, with malignant as the outlier class, consisting of 239 instances.

In Table 11 and 12, where the confusion matrices of the models are presented, it can be observed that all of them accurately identify the inlier points, while the challenge lies in labeling the outliers. KSH, followed by KSM, achieved the best f-score. This outcome arises from these methods successfully identifying the highest percentage of correct outliers, with KSH at 29% and KSM at 27%. The total percentage of outliers in the data amounts to 35%; then, they are close. In this scenario, KSH emerges as the superior model, as its false negative rate of 6% is lower than the 8% exhibited by KSM. Moreover, the difference in false positives is minimal, with KSH at 2% and KSM at 1%.

For the methodologies, Fast-MCD and SS F-score are zero. In these cases, the methods could not identify any outliers. The methods Kurtosis and KM were unsuccessful on this dataset because the covariance matrix is not estimated because the matrix was singular to working precision.

Table 10: F-score with Breast W dataset

	F-SCORE						
Dataset	**Fast MCD**	**Kurtosis**	**SS**	**ME**	**KM**	**KSM**	**KSH**
BreastW	0.00	*	0.00	0.76	*	0.86	0.88

For Figure 11, the Mahalanobis distance for the methods with the best f-score was plotted; here, we can see that the unidentified outliers are more on KSM than KSH, the inliers labeled as outliers are more on KSH than KSM, and the identified outliers are more by KSH than KSM. KSH identified 206 outliers and KSM 186.

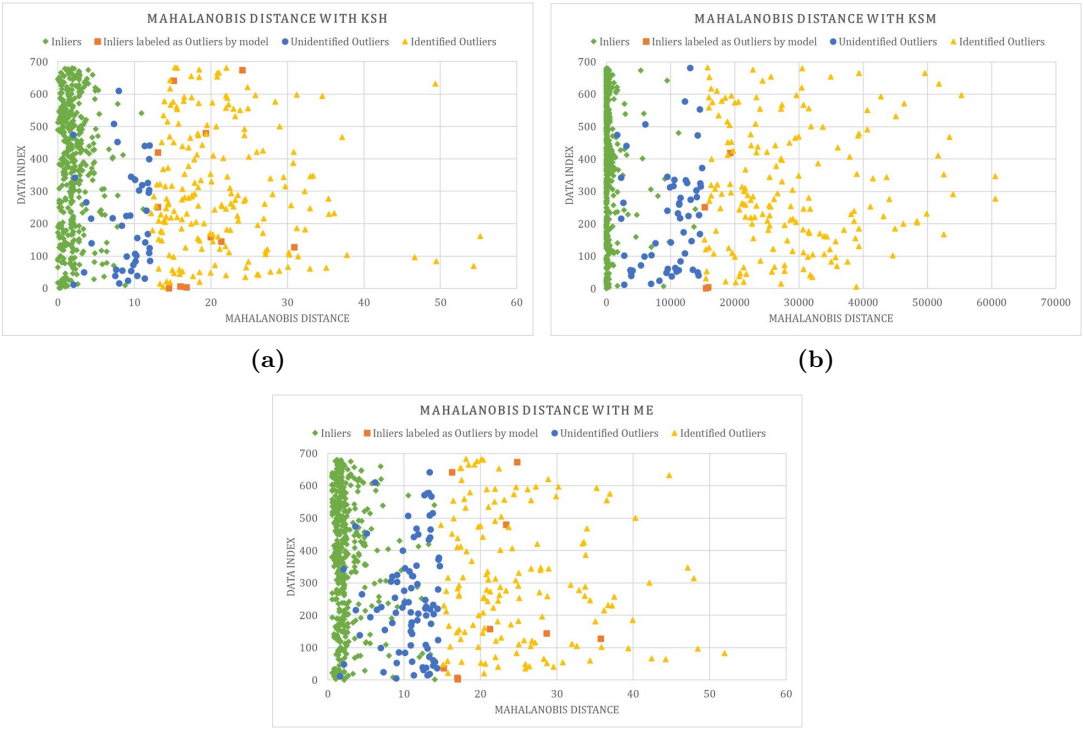
Table 11: Confusion matrices of models Fast-MCD, SS, and ME with Breast W dataset

BreastW		**Predicted Class**					
		Fast MCD		**SS**		**ME**	
		Normal	Outlier	Normal	Outlier	Normal	Outlier
True Class	Normal	65%	0%	65%	0%	64%	1%
	Outlier	35%	0%	35%	0%	13%	22%

Table 12: Confusion matrices of models KSM and KSH with Breast W dataset

BreastW		**Predicted Class**			
		KSM		**KSH**	
		Normal	Outlier	Normal	Outlier
True Class	Normal	64%	1%	63%	2%
	Outlier	8%	27%	6%	29%

Figure 11 BreastW Index data vs. Mahalanobis distance (a)KSH, (b)KSM, (c)ME



(a) **(b)**

(c)

The second dataset consists of 13 measurements on 129 samples of wine, Wine dataset (Aeberhard & Forina, 1991). It has 119 inliers (92%) and 10 outliers (8%).

These data describes chemical analysis data of wines from three Italian cultivars. The analysis determined the quantities of 13 constituents. Classes 2 and 3 serve as inliers, while class 1 is down-sampled to 10 instances to be used as outliers. (Rayana, 2016)

With this dataset, the betters f-score are obtained again with KSH and KM; for this data, it was possible to estimate the covariance matrix with all methods, but the results of the f-score are small. Table 13 shows the information.

Wine dataset presents higher difficulty as the percentage of outliers is low, which may also be interpreted as an imbalanced dataset. Nevertheless, the methods KSM and KSH exhibit a good performance, achieving a significant F-score close to 1. In contrast, the remaining methods show an F-score closer to 0.5. In these methods, the percentage of normal points classified as atypical is high compared to KSM and KSH, indicating the presence of false positives. The researcher's preference or concern regarding such cases depends on specific circumstances, as situations may arise where the presence of false positives incurs high costs or poses problematic implications.

Table 13: F-score with Wine dataset

	F-SCORE						
Dataset	**Fast MCD**	**Kurtosis**	**SS**	**ME**	**KM**	**KSM**	**KSH**
Wine	0.48	0.61	0.42	0.12	0.40	0.71	0.75

Table 14: Confusion matrices of models Fast-MCD, Kurtosis, SS, and ME with Wine dataset

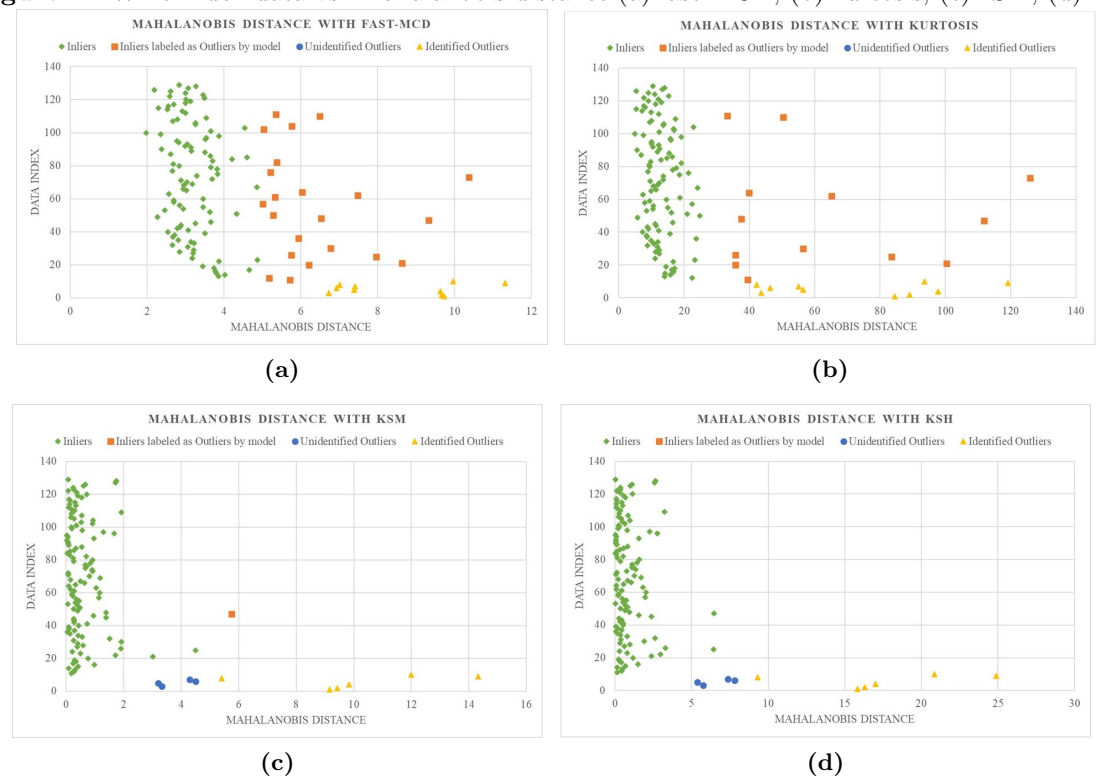
Wine		**Predicted Class**							
		Fast MCD		**Kurtosis**		**SS**		**ME**	
		Normal	Outlier	Normal	Outlier	Normal	Outlier	Normal	Outlier
True Class	Normal	75%	17%	82%	10%	85%	7%	76%	16%
	Outlier	0%	8%	0%	8%	4%	4%	6%	2%

Table 15: Confusion matrices of models KM, KSM, and KSH with Wine dataset

Wine		**Predicted Class**					
		KM		**KSM**		**KSH**	
		Normal	Outlier	Normal	Outlier	Normal	Outlier
True Class	Normal	84%	8%	91%	1%	92%	0%
	Outlier	4%	4%	3%	5%	3%	5%

Figure 12 shows that the results of the Mahalanobis distance calculated with KSM and KSH allow for a slightly clearer visualization of which points are outliers compared to the Fast MCD and Kurtosis methods.

Figure 12 Wine Index data vs. Mahalanobis distance (a)Fast MCD; (b)Kurtosis; (c)KSM, (d)KSH



(a) **(b)**

(c) **(d)**

The third example consists of the Ionosphere dataset (Sigillito & Baker, 1989), which has 351 data and 33 attributes, with a binary classification where the "bad" class are the outliers with 126 points (36%), and "good" class inliers with 225 point (64%).

For this dataset, no model identifies an inlier as an outlier, but there is difficulty in identifying outliers. For instance, Fast-MCD and SS do not label any point as atypical. The two methods that identify the highest number of outliers are ME and KSH, although it is worth noting that for this dataset, no method identifies the majority of the atypical data; the two best practices identify 50% of

Table 16: F-score with Ionosphere Dataset

	F-SCORE						
Dataset	**Fast MCD**	**Kurtosis**	**SS**	**ME**	**KM**	**KSM**	**KSH**
Ionosphere	0.00	*	0.00	0.69	*	0.46	0.66

Table 17: Confusion matrices of models Fast MCD, SS, and ME with Ionosphere dataset

Ionosphere		**Predicted Class**					
		Fast MCD		**SS**		**ME**	
		Normal	Outlier	Normal	Outlier	Normal	Outlier
True Class	Normal	64%	0%	64%	0%	64%	0%
	Outlier	36%	0%	36%	0%	17%	19%

Table 18: Confusion matrices of models KSM, and KSH with Ionosphere dataset

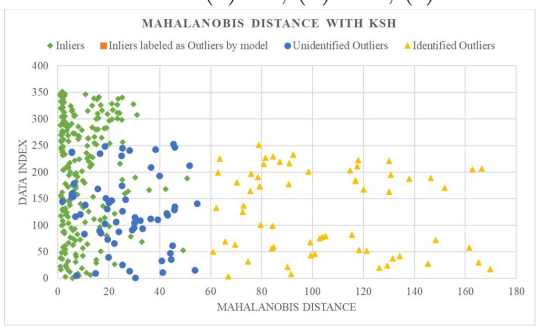
Ionosphere		**Predicted Class**			
		KSM		**KSH**	
		Normal	Outlier	Normal	Outlier
True Class	Normal	64%	0%	64%	0%
	Outlier	25%	11%	18%	18%

these data. Overall, the results of ME and KSH in this test are very similar. Both classify inliers well, identify the same percentage of outliers, and have a similar false negatives rate: 17% for ME and 18% for KSH.

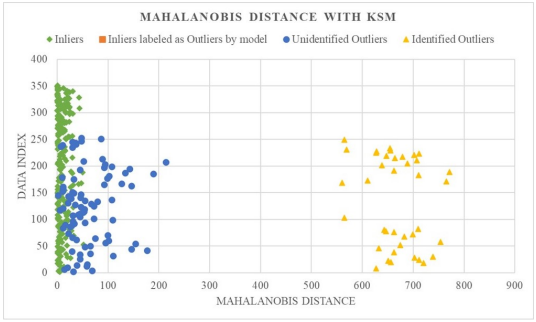
In Figure 13, it can be observed that although the Mahalanobis distance with KSM identifies some of the outliers, it does so only with a small subset of them. In comparison, KSH and ME exhibit a higher identification of outlier points, with ME achieving a better F-score by successfully classifying 5 more points as outliers than KSH.

Figure 13 Ionosphere Index data vs. Mahalanobis distance (a)ME, (b)KSH, (c)KSM

(a)



(b)



(c)

4 Conclusion

Based on the obtained results, it is evident that the proposed methods, involving different combinations of tools such as Kendall's Tau, Median Absolute Deviation, and Ledoit & Wolf Shrinkage for covariance matrix estimation, successfully achieve a Mahalanobis distance that accurately classifies outlier data in scenarios where these outliers exhibit high dispersion. This is evident in cases with $\lambda = 1$, both in sets of independent normal data and correlated normal data, where the KSM, KM, and KSH methods achieve an f-score exceeding 0.8 even in high dimensions. This underscores the favorable trade-off between precision and recall that these techniques exhibit when classifying atypical data, with KSM and KM particularly standing out. While the SS method performs better than those proposed here for data with an independent normal distribution, it is noteworthy that the effectiveness of KSM, KM, and KSH persists in scenarios featuring highly correlated normal data.

Another proposed method involved using standard Mahalanobis distance with k-means to identify outliers. From the conducted simulations, it is concluded that this procedure is successful when the dispersion of outliers is low, i.e., $\lambda = 0.1$, and the dataset dimension is equal to or greater than 20. In these cases, the f-score of this process exceeds 0.8, even in high dimensions. This indicates that the model maintains a balanced performance in accurately identifying positive points and maintaining reasonable precision in its positive predictions in those scenarios.

In the explorations with real data, challenges were encountered in estimating the covariance matrix due to difficulties in obtaining a non-singular and well-conditioned matrix, which hindered the application of certain methods for identifying outliers. Nevertheless, even in these presented scenarios, it was possible to estimate the Mahalanobis distance using the KSH method. This method stood out consistently among those that identified the most outliers while maintaining low false positives and negative rates. This underscores the robustness achievable when calculating covariance using Kendall's Tau and Ledoit & Wolf Shrinkage.

Simulations performed to make comparative studies between methods also led to conclusions about them. For example, the SS method is very robust for identifying outliers in independent normal datasets, as its f-score is over 0.8 at various levels of contamination and contamination concentration. However, its performance was not as effective in dimensions less than 10 when the dispersion of the outliers was low. On the other hand, the kurtosis method showed increased accuracy in correlated normal datasets when the data size was increased, as its f-score was higher in simulated scenarios with n=1000 compared to cases of n=500.

References

Aeberhard, Stefan, & Forina, M. 1991. *Wine*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5PC7J.

Barnett, V, & Lewis, T. 1978. *Outliers in Statistical Data, Series in probability and mathematical statistics; applied probability and statistics*.

Cabana, Elisa, Lillo, Rosa E, & Laniado, Henry. 2021. Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. *Statistical papers*, **62**, 1583–1609.

Chandola, Varun, Banerjee, Arindam, & Kumar, Vipin. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, **41**(3), 1–58.

Chok, Nian Shong. 2010. *Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data*. Ph.D. thesis, University of Pittsburgh.

Croux, Christophe, & Dehon, Catherine. 2010. Influence functions of the Spearman and Kendall correlation measures. *Statistical methods & applications*, **19**, 497–515.

Falk, Michael. 1997. On mad and comedians. *Annals of the Institute of Statistical Mathematics*, **49**, 615–644.

Gnanadesikan, Ramanathan, & Kettenring, John R. 1972. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 81–124.

Hampel, Frank R. 1974. The influence curve and its role in robust estimation. *Journal of the american statistical association*, **69**(346), 383–393.

Han, Jiawei, Kamber, Micheline, & Pei, Jian. 2012. Data mining concepts and techniques third edition. *University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University*.

Hawkins, Douglas M. 1980. *Identification of outliers*. Vol. 11. Springer.

Kendall, Maurice G. 1938. A new measure of rank correlation. *Biometrika*, **30**(1/2), 81–93.

Ledoit, Olivier, & Wolf, Michael. 2003. Honey, I shrunk the sample covariance matrix. *UPF economics and business working paper*.

Leys, Christophe, Ley, Christophe, Klein, Olivier, Bernard, Philippe, & Licata, Laurent. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology*, **49**(4), 764–766.

Mahalanobis, PC. 1936. *On the generalized distance in statistics. Proc. Natl. Inst. Sci. India*.

Maronna, Ricardo A, & Zamar, Ruben H. 2002. Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, **44**(4), 307–317.

Peña, Daniel, & Prieto, Francisco J. 2001. Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, **43**(3), 286–310.

Peña, Daniel, *et al.* 2013. *Análisis de datos multivariantes*. McGraw-Hill España Cambridge.

Rayana, Shebuti. 2016. *ODDS Library. Stony Brook University, Department of Computer Sciences*.

Rousseeuw, Peter J. 1985. Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, **8**(283-297), 37.

Rousseeuw, Peter J, & Driessen, Katrien Van. 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**(3), 212–223.

Sajesh, TA, & Srinivasan, MR. 2012. Outlier detection for high dimensional data using the Comedian approach. *Journal of statistical computation and simulation*, **82**(5), 745–757.

Sigillito, V., Wing S. Hutton L., & Baker, K. 1989. *Ionosphere*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5W01B.

Smiti, Abir. 2020. A critical overview of outlier detection methods. *Computer Science Review*, **38**, 100306.

Wolberg, WIlliam. 1992. *Breast Cancer Wisconsin (Original)*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5HP4Z.