Unsupervised Outlier Detection in Multidimensional Data

Atiq Rehman (atiqjadoon@gmail.com)
Hamad Bin Khalifa University College of Science and Engineering https://orcid.org/0000-0003-0248-7919

Samir Brahim Belhaouari
Hamad Bin Khalifa University College of Science and Engineering

Research

Keywords: nomaly/Outliers Detection, Advanced Statistical Methods, Computationally Inexpensive Methods, High Dimensional Data.

Posted Date: February 25th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-250665/v1

Unsupervised Outlier Detection in Multidimensional Data

Atiq ur Rehman[1*], and Samir Brahim Belhaouari[2]

[1,2]ICT Division, College of Science and Engineering,

Hamad Bin Khalifa University, Doha, Qatar.

Corresponding author: [1*] atrehman2@hbku.edu.qa.

Abstract—Detection and removal of outliers in a dataset is a fundamental preprocessing task without which the analysis of the data can be misleading. Furthermore, the existence of anomalies in the data can heavily degrade the performance of machine learning algorithms. In order to detect the anomalies in a dataset in an unsupervised manner, some novel statistical techniques are proposed in this paper. The proposed techniques are based on statistical methods considering data compactness and other properties. The newly proposed ideas are found efficient in terms of performance, ease of implementation, and computational complexity. Furthermore, two proposed techniques presented in this paper use only a single dimensional distance vector to detect the outliers, so irrespective of the data's high dimensions, the techniques remain computationally inexpensive and feasible. Comprehensive performance analysis of the proposed anomaly detection schemes is presented in the paper, and the newly proposed schemes are found better than the state-of-the-art methods when tested on several benchmark datasets.

Index Terms—Anomaly/Outliers Detection, Advanced Statistical Methods, Computationally Inexpensive Methods, High Dimensional Data.

1 Introduction

An observation in a dataset is considered an outlier if it differs significantly from the rest of the observations. The problem of finding patterns in data that deviate from the expected behavior is called the anomaly detection or the outliers' detection problem. Outliers in data can occur due to the variability in measurements, experimental errors, or noise [1], and the existence of outliers in data makes the analysis of data misleading and degrades the performance of machine learning algorithms [2][3].

Several techniques have been developed in the past to detect outliers in data [4]–[6]. The techniques for outlier detection can be broadly classified as methods based on: (i) Clustering [7], (ii) Classification [8], (iii) Neighbor based [9], (iv) Statistical [10], (v) Information-Theoretic [11], and (vi) Spectral methods [12]. The working of classification-based methods mostly relies on a confidence score, which is calculated by the classifier while making a prediction for the test observation. If the score is not high enough, the observation is not assigned any label and is considered an outlier. Some clustering-based methods identify the outliers by not forcing every observation to belong to a cluster, and the observations that are not assigned to any cluster are identified as outliers. The nearest neighbor techniques are mostly based on a calculation of the distance or similarity measure between the observation and its neighboring observations. Suppose the calculation is greater than a certain threshold, that means that the observation lies far apart from the rest of the observations and is considered as an outlier. Statistical methods usually fit a statistical distribution (mostly normal distribution) to the data and conduct a statistical inference test to see if the observation belongs to the same distribution or not. If not, the observation is marked as an outlier. Information-theoretic techniques use different information theoretic measures for example entropy, relative entropy, etc., to analyze the information content of the data. These techniques are based on an assumption that the outliers or anomalies in the data induce irregularities in the information content. Spectral methods transform the data to a new dimensional space such that the outliers are easily identified and separated from the data in the new space. Furthermore, some outlier detection techniques are also based on geometric methods [13] and neural networks [14].

All the techniques mentioned above are based on some assumptions and all the techniques have some pros and cons. The ideas proposed in this work are based on the novel statistical methods considering data properties like compactness. The aim here is to utilize the power of some statistical methods and to enhance the performance of the outlier detection algorithms in an unsupervised way, while keeping their implementation easy and computationally efficient. The newly proposed methods are based on boxplot adjustment, probability density estimation and neighborhood information. The proposed methods are evaluated using both the synthetic and the real datasets and are found better in scenarios where the traditional approaches fail to perform well. Especially the cases where the data is contaminated with a mixture of different noise distributions.

The rest of the paper is organized as follows: Background (Section 2), Proposed Methods (Section 3), Evaluation

on synthetic datasets (Section 4), Evaluation on real data (Section 5), Comparison with State-of-Art (Section 6) and Conclusions (Section 7).

2 BACKGROUND

After the initial pivotal works for outlier detection based on the distance measure [15], [16], several new methods based on the distance measure were also proposed by different authors in literature [17][18]. The difference between the latterly proposed methods and the previous studies is the use of nearest neighbors in distance calculation. Among the variants of the actual work, either a single distance based on the k^{th} closest neighbor is calculated [19] or the aggregate of the distance of k closets points is calculated [20]. Among other unsupervised outlier detection algorithms are the *local* approaches originated from the concept of Local Outlier Factor [21].

Furthermore, boxplot outlier detection scheme is also one of the fundamental unsupervised approach and the concept of univariate boxplot analysis was first proposed by Tukey et. al. [22]. In a univariate boxplot, there are five parameters specified as: (i) the upper extreme bound (UE), (ii) the lower extreme bound (LE), (iii) the upper quartile Q3 (75th percentile), (iv) the lower quartile Q1 (25th percentile) and (v) the median Q2 (50th percentile). The best way to estimate the extreme boundaries is to estimate Probability Density Function (PDF), $f(x)$, at first step from where the boundaries will be defined, as follows:

$$\begin{cases} UE: \frac{\tau}{2} = P(X > UE) = \int_{UE}^{+\infty} f(x)dx \\ LE: \frac{\tau}{2} = P(X < LE) = \int_{-\infty}^{LE} f(x)dx \end{cases} \quad (1)$$

Where τ is the significance level, the region of suspected outliers is defined for $\tau = 0.05$ and the region of extremely suspected outliers is defined for $\tau = 0.01$. The equation (1) estimates well the boundaries only if the distribution is unimodal, i.e., a distribution that has single peak or at most one frequent value.

However, in a standard boxplot the UE and LE values are computed and well estimated only under the assumption that the PDF is symmetric, as:

$$\begin{cases} LE = Q1 - 1.5(IQR), \\ UE = Q3 + 1.5(IQR). \end{cases} \quad (2)$$

where, the term IQR is defined as the Inter Quartile Range and is given by:

$$IQR = Q3 - Q1. \tag{3}$$

A common practice to identify the outliers in a dataset using a boxplot is to mark the points that lie outside the extreme values, that is, the points greater than UE and less than LE are identified as outliers. This version of outlier detection scheme works well for the symmetric data. However, for skewed data different other schemes are proposed in the literature. For example, different authors have used the semi-interquartile range i.e. $Q3 - Q2$ and $Q2 - Q1$ to define the extreme values as:

$$\begin{cases} LE = Q1 - c_1(Q2 - Q1), \\ UE = Q3 + c_2(Q3 - Q2). \end{cases} \tag{4}$$

Where, c_1 and c_2 are the constants and different authors have adjusted their values differently for example, $c_1 = c_2 = 1.5$ [23], $c_1 = c_2 = 3$ [24] or calculation based on the expected values of the quartiles [25] and few more adjustments to the boxplot for outliers detection are also available, for example [26].

The traditional methods of boxplot for detecting the outliers sometimes fails in situations where the noise in the data is a mixture of distributions, multimodal distribution, or in the presence of small outlier clusters. In this paper, some novel statistical schemes based on (i) the boxplot adjustments, and (ii) a new probability density estimation using k-nearest neighbors' distance vector are proposed to overcome the problem faced by traditional methods. These proposed methods are described in detail in the next section.

3 PROPOSED METHODS

3.1 Boxplot Adjustments using D-k-NN (BADk)

Rather than using the traditional boxplot to identify the outliers from the unique dimensions, one useful idea is to calculate the distance between all the data points considering all the dimensions and use the resulting single dimension distance vector to identify the outliers. The idea of using a single dimension distance vector is useful not only for avoiding problem of sorting data in high dimension but also in terms of computational cost, and can be further enhanced in terms of performance by extending it to consider *k* number of neighbors in the distance calculation. This idea of boxplot adjustment based on the Distance vector considering *k* number of Nearest Neighbors (D-k-NN) is presented here and the resulting extreme values estimation from the modified boxplot are found to be quite useful in identifying the right outliers. Furthermore, the proposed scheme is useful in the cases

where the distribution of the noise is not normal or is a mixture of different distributions, and can identify small outlier clusters in the data.

Consider a case of data in $\mathbb{R}^N$, the N dimensional data can be transformed in a single vector representation using a distance measure for example '*Euclidian distance'*, which computes the distance of each observation in $\mathbb{R}^N$ to its k^{th} closest neighbor, named d_k. The concept of transformation can be summarized as the following function:

$$d_k: \mathbb{R}^N \to \mathbb{R} \tag{5}$$

where, $d_k \in \mathbb{R}$ is a vector containing the distance of each observation in $\mathbb{R}^N$ to its k^{th} closest neighbor and it is used to compute the boxplot with extreme values computed as:

$$\begin{cases} LE_{d_k} = Q1_{d_k} - c_1\left(Q2_{d_k} - Q1_{d_k}\right), \\ UE_{d_k} = Q3_{d_k} + c_2\left(Q3_{d_k} - Q2_{d_k}\right). \end{cases} \tag{6}$$

Similar to the traditional method, now the outliers can be identified as the observations that lie outside the defined extreme values in (6). The constants c_1 and c_2 can either be tuned for better performance or can be kept constant as equal to 1.5 or 3 as suggested by different authors in the literature. The resulting box-whisker plot is easy to implement and can identify the underlying outliers in the dataset accurately.

Furthermore, another useful idea to identify the outliers in a data is to adjust the UE and LE values of a boxplot as follows:

$$\begin{cases} LE = Q1_{d_k} - c_1 \times \sqrt{var\left(X.1_{X<Q2_{d_k}}\right)}, \\ UE = Q3_{d_k} + c_2 \times \sqrt{var\left(X.1_{X\geq Q2_{d_k}}\right)}. \end{cases} \tag{7a}$$

or

$$\begin{cases} LE = Q1_{d_k} - c_1 \times \sqrt{var\left(X.1_{X<Q1_{d_k}}\right)}, \\ UE = Q3_{d_k} + c_2 \times \sqrt{var\left(X.1_{X\geq Q3_{d_k}}\right)}. \end{cases} \tag{7b}$$

where, var is defined as the variance and the quartiles are computed from the distance vector $d_k \in \mathbb{R}$. The extreme values can also be estimated based on the calculation of the separation threshold between centers of two variances, see equation (9), as:

$$\begin{cases} LE = M - c_1 \times var(X.1_{X<M}), \\ UE = M + c_2 \times var(X.1_{X\geq M}). \end{cases} \tag{8}$$

where, M is a value that separate the one-dimension region in order to calculate the variance of two centers. Let $x \in \mathbb{R}$ be any random variable with PDF $f(x)$; and the values of μ_1 and μ_2 are calculated such that:

$$\begin{cases} (\mu_1^*, \mu_2^*) = \underset{(\mu_1,\mu_2)}{\arg} \left[\min_{M,\mu_1,\mu_2} \left[\int_{-\infty}^{M} (x-\mu_1)^2 f(x)dx + \int_{M}^{\infty} (x-\mu_2)^2 f(x)dx \right] \right], \\ Var_1 = \int_{-\infty}^{M} (x-\mu_1)^2 f(x)dx, \\ Var_2 = \int_{M}^{\infty} (x-\mu_2)^2 f(x)dx\,. \end{cases} \tag{9}$$

Both, Var_1 and Var_2 can be partially differentiated with respect to μ_1 and μ_2 respectively, to find the minimum. After simplification the minimization occurs when:

$$\begin{cases} \mu_1 = \frac{E(X_-)}{P(X_-)}, \\ \mu_2 = \frac{E(X_+)}{P(X_+)}, \end{cases} \tag{10},$$

where, $X_- = X.1_{X<M}$ and $X_+ = X.1_{X\geq M}$. and the value of M is calculated as:

$$M = \left[\frac{E(X_-)}{P(X_-)} + \frac{E(X_+)}{P(X_+)} \right] . \frac{1}{2} \tag{11}$$

For further details on the idea proposed in (8)-(11), the readers are referred to [27].

Detecting outliers based on Boxplot is efficient only if the data is unimodal distribution. To overcome the drawbacks of the boxplot estimation, some other statistical methods based on the probability density estimation computed from either the distance vector $d_k \in \mathbb{R}$ or the actual data $D \in \mathbb{R}^N$ are also proposed for outlier's detection, which are discussed below.

3.2 Joint Probability Density Estimation using D-k-NN

The methods proposed in this section compute the distance vector d_k from the actual data and utilize it for estimating some parameters of the joint distribution function. Three different schemes are proposed here which are described as follows:

Scheme 1:

Normal distributions are often used for representing the real value random variables with unknown distributions [28] [29]. The joint probability density function of independent and identically normal distribution is given as:

$$f(x_1, \dots, x_N) = \frac{1}{(\zeta\sqrt{2\pi})^N} e^{\sum_{i=1}^{N} -\frac{1}{2}\left(\frac{x_i-\mu_i}{\zeta}\right)^2} \tag{12}$$

where, ζ is the standard deviation modeled differently in (15) and (17), μ is the mean of the random variable and N is the dimension of the data. Here, some functions based on the normal distribution to identify the outliers in a dataset are proposed. Suppose a two-dimensional dataset $D(x, y)$, we can define a separation threshold T based on the normal distribution for detecting the outliers such that:

$$\begin{cases} Z = \sum_{i=1}^{n} f(x_i, y_i), \\ T = \alpha\, max(Z). \end{cases} \tag{13}$$

where, Z is joint probability distribution function after normalization, n is the total number of observations and the function $f(x, y)$ can be defined as:

$$f(x, y) = \frac{1}{2\pi\zeta^I} e^{-\left(\frac{(x-x_i)^2+(y-y_i)^2}{2\zeta^2}\right)}; i = 1,2, \dots, n.; I = 0,1,2. \tag{14}$$

The σ in equation (14) can be computed as:

$$\zeta = \beta Q3_{d_k}. \tag{15}$$

where, $Q3_{d_k}$ is the third quartile computed from the distance vector d_k as defined in equation (5) and β is a constant value. The points below the threshold value T defined in equation (13) are considered as outliers and the points above T are considered normal inlier data points. The α used in equation (13) is the significance value and it can be used to control the percentage amount of data to be removed as outliers.

Scheme 2:

To better detect the outliers, a better function $f(x, y)$ needs to be constructed in order to weaken the position of the outlier in terms of support and amplitude of the function. Furthermore, another scenario can be defined to detect the outliers based on the threshold defined in (13) by using the below function:

$$f(x,y) = \frac{\zeta^2}{\pi} e^{-\left(\frac{(x-x_i)^2+(y-y_i)^2}{\zeta^2}\right)}; i = 1,2,\dots,N. \tag{16}$$

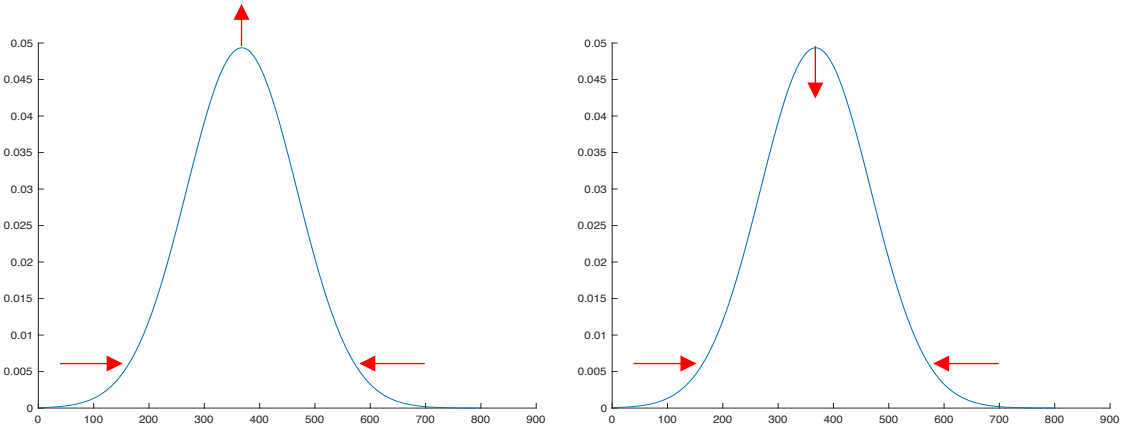
$$\zeta = \frac{\gamma}{(1+d_k)^2} \tag{17}$$

where, k defines the k^{th} closest neighbor for the distance vector and γ is a constant whose value can be adjusted to control the smoothness of the gaussian distribution. The concept is demonstrated in Fig 1, where (a) shows the effect of traditional gaussian approach on compression and (b) shows the effect of proposed scheme 2 on compression.

Both of the above schemes proposed in this section are based on a single gaussian distribution and are expected to work well for the datasets which can be well approximated using a single gaussian distribution. However, if a dataset can be better approximated using multiple gaussians then a better idea is to use a model based on the variable number of gaussians. A new and robust estimation of multiple gaussian distribution is proposed in the next subsection.

Scheme 3:

The scenarios where the data is estimated using a gaussian distribution, the outliers are identified as the points lying on the extreme tails of the gaussian distribution, as shown in Fig 2 (a). However, if the better estimation of underlying data is possible through multiple gaussians, the outliers located at the connecting points of different gaussians might remain unidentified using a single gaussian estimation. In order to identify the outliers existing at



(a) (b)

Fig 1. (a) Effect of traditional gaussian approach on compression. (b) Effect of proposed scheme 2 approach on compression in x and y axes.

the connecting points of the multiple gaussians, an idea based on multiple gaussian estimation is proposed, where a Rejection Area (RA) is defined and computed as:

$$\begin{cases} RA = \{\vec{x}: f(\vec{x}) \leq Cv\}, \\ f(x \in RA) = \tau. \end{cases} \tag{18}$$

where, Cv is defined as a critical value or a threshold value below which is the rejection area or where the outliers are identified, and τ is the significance level. The concept is shown in Fig 2 (b), where as an example a single dimensional data is estimated using two gaussians and the outliers can be identified as the points below Cv.

In order to find the optimum number of gaussians that better approximate the joint probability distribution for a given dataset the sorted values of the vector $d_k$can be utilized. For example, in Fig 3 the graph of sorted values of the vector $d_k$is shown and the best value of number of gaussians can be estimated by taking the value where the graph takes off sharply.

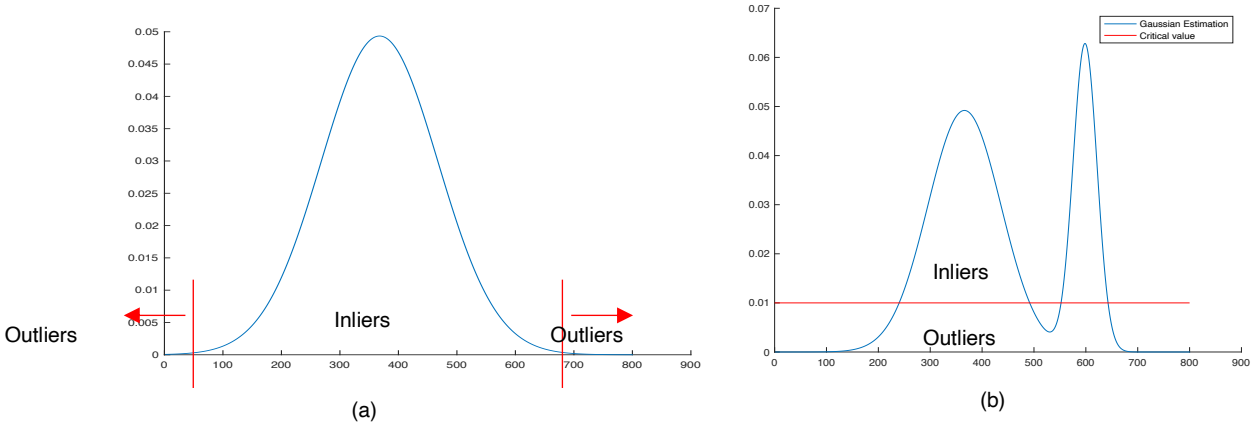


Fig 2. An Example of Gaussian estimation and marking of critical value for outlier detection. (a) Points those lie outside the red boundaries are considered outliers. (b) The points below the critical value are identified as outliers.

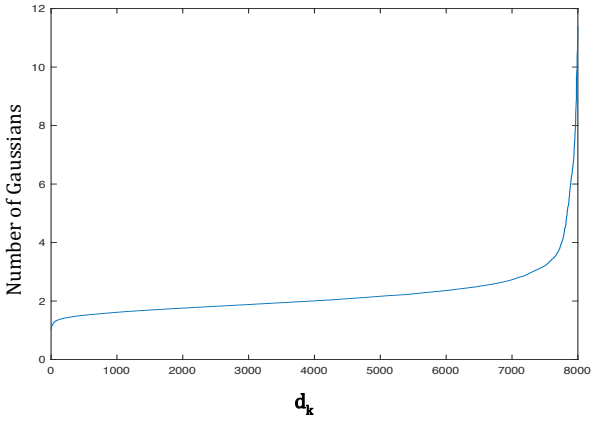


Fig 3. Plot of sorted values of the vector $\boldsymbol{d_k}$

Each estimated gaussian represent a region and for each gaussian inside a region the values of mean and variance can be computed as:

$$\mu_i = \frac{\sum_{i \in R_j} x_i}{n_i}, \text{ j=1, 2, …., m} \tag{19}$$

$$Var(x) = \frac{\sum_{i \in R_j} (x_i - \mu_i)^2}{n_i - 1}, \text{ j=1, 2, ..., m} \tag{20}$$

where R_j represents the jth region, m is the total number of estimated gaussians and n_i is the total number of elements in the respective region. The combined multiple gaussians model is then estimated by:

$$x \sim \sum_{i=1}^{m} \alpha_i N(\mu_i, C_i) \tag{21}$$

where,

$$\alpha_i = \frac{Card(R_i)}{N} \text{ and } \sum \alpha_i = 1. \tag{22}$$

In order to determine the regions, lets define an application S_U that sorts any given sequence $U_i, i = 1, \ldots, N$ such that $U_{S_U(1)} \leq U_{S_U(2)} \leq \cdots \leq U_{S_U(N)}$. For any given data $\vec{X}$, the sorted data can be represented as $\vec{X}_{S_U(i)}$ and suppose that $\overrightarrow{\Delta X}_{S_U(i)}$ represents the difference between two consecutive elements of $\vec{X}_{S_U(i)}$. Similarly, the sorted difference can be represented as $\overrightarrow{\Delta X}_{S_{\Delta X}(S_U(i))}$. In order to define the regions, the elements are grouped together sequentially until $\Delta X_{S_{\Delta X}(S_U(i))} \leq \overrightarrow{\Delta X}_{S_{\Delta X}(S_U(N-m+1))}$, once this condition is not true, start grouping the remaining elements as a new region until all the elements are assigned to a region.

4 Evaluation Using Synthetic Examples

The ability of proposed methods is demonstrated here by the use of some two-dimensional synthetic datasets. The results for each of the proposed method are discussed in the following subsections.

4.1 Evaluation Boxplot Adjustments using D-k-NN (BADk)

Fig 4 (a) shows an example of the data used for evaluating the proposed methods to detect the outliers. From the data shown in Fig 4 (a), it can be seen that the actual data is composed of different clusters with different shapes and is contaminated by a mixture of sinusoidal and gaussian distribution of noise. The aim here is to detect the

noise as outliers and different shape clusters as inliers. The traditional boxplot is used to detect the outliers from the data and the resulting boxplot is shown in Fig 4 (b). It can be seen from the boxplot in Fig 4 (b) that the traditional boxplot is unable to identify any outliers in the data.

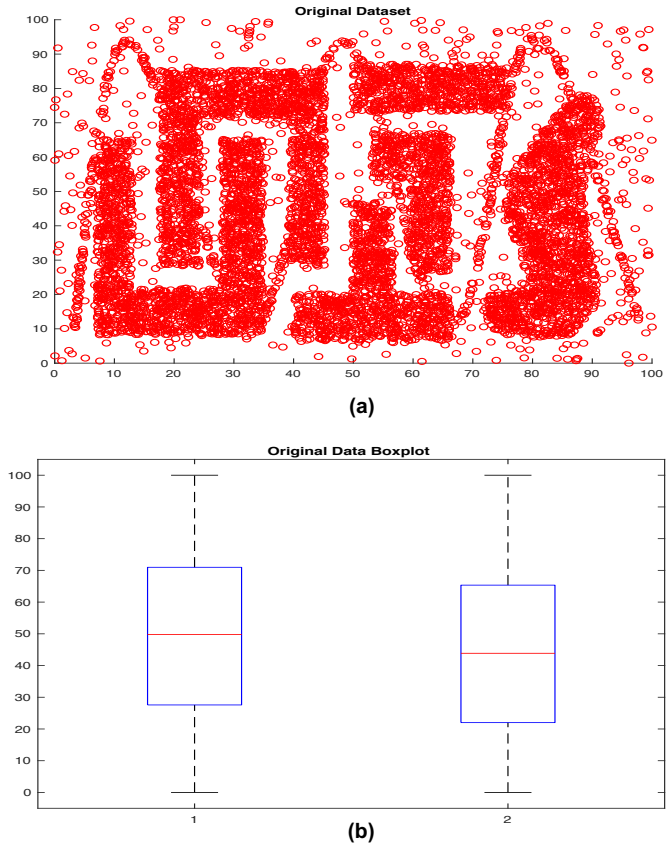


Fig 4. (a) Original data representing clusters with different shapes and a mixture of sinusoidal and gaussian noise. (b) Traditional boxplot showing no outliers/anomalies.

The same dataset is used to evaluate the proposed adjusted boxplot with extreme values define in equation (6) and the results are shown in Fig 5. Different values of k are used to see how it effects the outcome in identifying the outliers. It can be observed from the results shown in Fig 5 that for the smaller values of k only the gaussian noise is identified and while we keep on increasing the value of k the outliers with sinusoidal distribution are also identified.

However, after a certain value of k the data points from the actual clusters (inliers) are also marked as the outliers, while the outliers started to reappear as the inliers. This shows that although the selection of value of k is flexible in this case, still an optimum value of k has to be selected for the optimum performance based on the data. Another

example shown in Fig 6(a) with a different distribution of noise is also tested for evaluating the ability of the proposed method in (6) for outlier detection. The results for this example are shown in Fig 7 using three different values of k. It can be seen from the results in Fig 7 that the selection of value of k is very flexible and still the proposed method performs well in terms of outlier's detection. During all the experiments performed, the values of constants are fixed to $c_1 = c_2 = 1.5$.

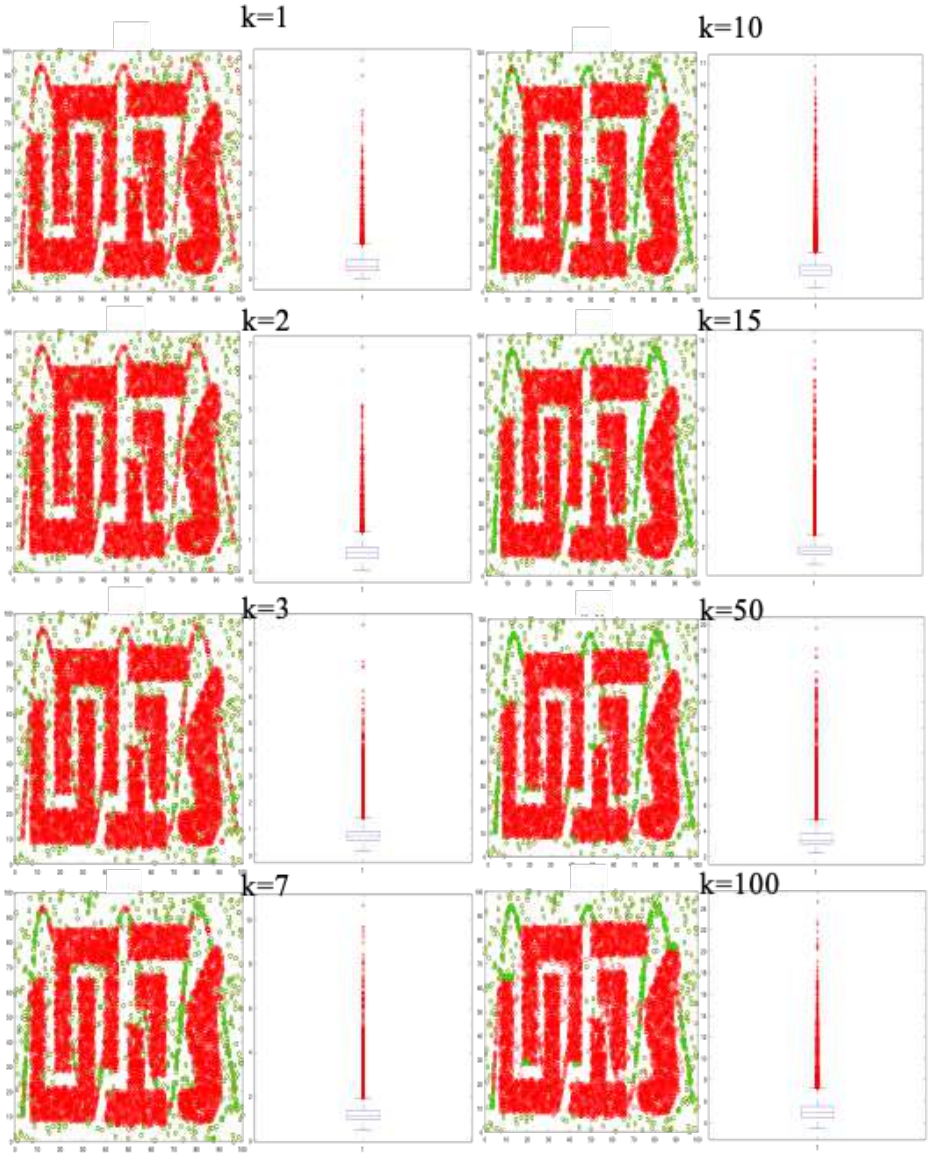


Fig 5. Outliers detection from the data shown in Fig 4(a), using the proposed boxplot with extreme values defined in equation (6) for different values of k. The data in red is identified as the inliers while the data in green is identified as the outliers.

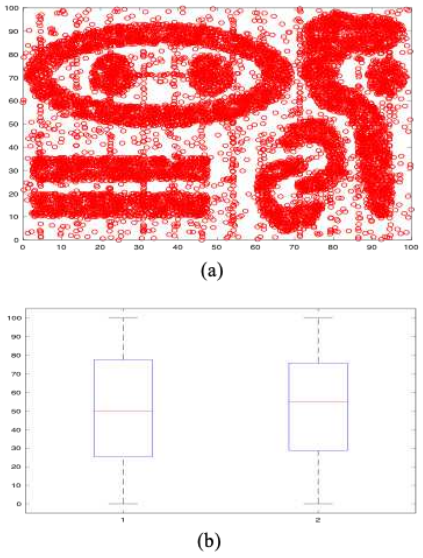


Fig 6. (a) An example of dataset having different shapes of inlier clusters contaminated with noise. (b) Traditional boxplot showing no outliers/ anomalies.

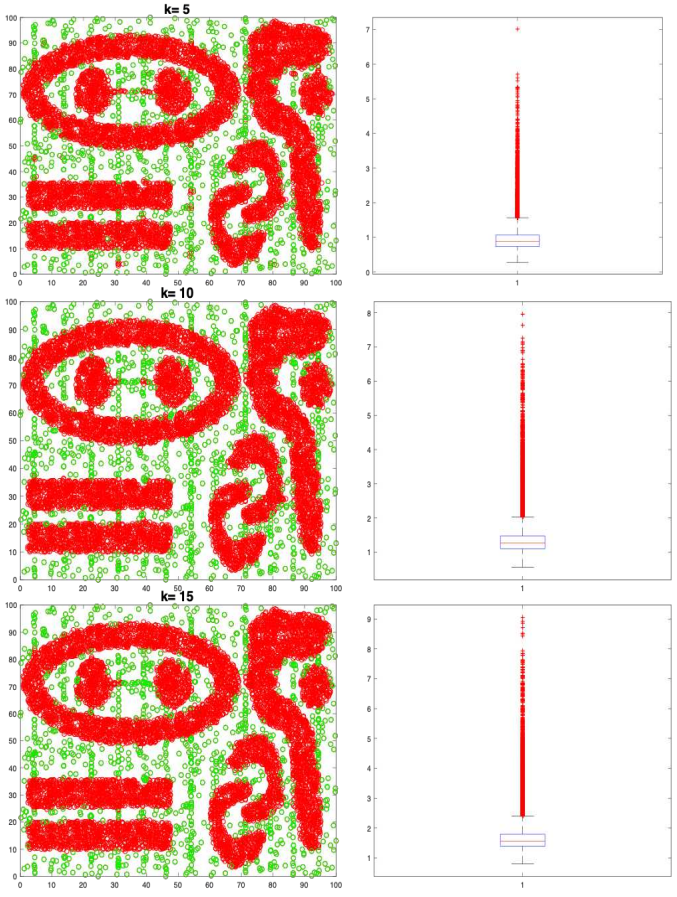


Fig 7. Outliers detection from the data shown in Fig 6 (a), using the proposed boxplot with extreme values defined in equation (6) for different values of k. The data in red is identified as the inliers while the data in green is identified as the outliers.

4.2 Evaluation (Joint Probability Density Estimation) Scheme 1

The density estimation refers to estimation of an unobservable Probability Density Function (PDF) associated with an observable data. The PDF gives an estimate of the density according to which a large population in a data is distributed. In this proposed method, the PDF is computed by placing a gaussian density function at each data point, and then summing the density functions over the range of data, and a threshold value α defines the margin between the inlier data and the outliers. The value of α is computed as a percentage amount of the maximum value of the PDF. The value of σ in equation (14) is computed utilizing the d_k vector as defined in equation (15).

The results for scheme 1 when evaluated using the same example data as shown in Fig 4(a) are given in Fig 8. The example is evaluated using only two different values of α and a fixed value of β. The outliers are shown in green color and the inlier data is shown in red color. Figure 8 also shows the associated 3D plots of the probability density estimations computed using equation (14). For α = 0.1 the proposed method is able to identify the outliers having gaussian distribution only while placing α = 0.3 the proposed method has identified both the gaussian and the sinusoidal outliers in the data. Fig 9 shows the results for the second example with a different noise distribution with fixed values of α and β using equation (14).

4.3 Evaluation (Joint Probability Density Estimation) Scheme 2

The results for scheme 2 proposed in equations (16) - (17) with different value of parameters are shown in Fig 10. It can be observed from the results in Fig 10 that the small value of γ produces sharp density distribution while a larger value of γ produced a smoother distribution. For a small value of γ the inlier data points are also identified as the outliers which is not the case with a comparatively larger value of γ for this particular dataset. However, optimum values for the parameters need to be tuned to get the optimum results using this scheme.

4.4 Evaluation (Joint Probability Density Estimation) Scheme 3

The idea proposed in equation (18) based on the different ways of estimation of gaussians of the distance vector d_k is evaluated on three different synthetic examples having different distribution of noise and the results are shown in Fig 11. It can be seen from the visual results depicted in Fig 11 that this scheme is successful in identifying the outliers of different distributions and even the noisy data that lies in close proximity to the inlier data. The value of G represents the number of gaussians estimated from the d_k.

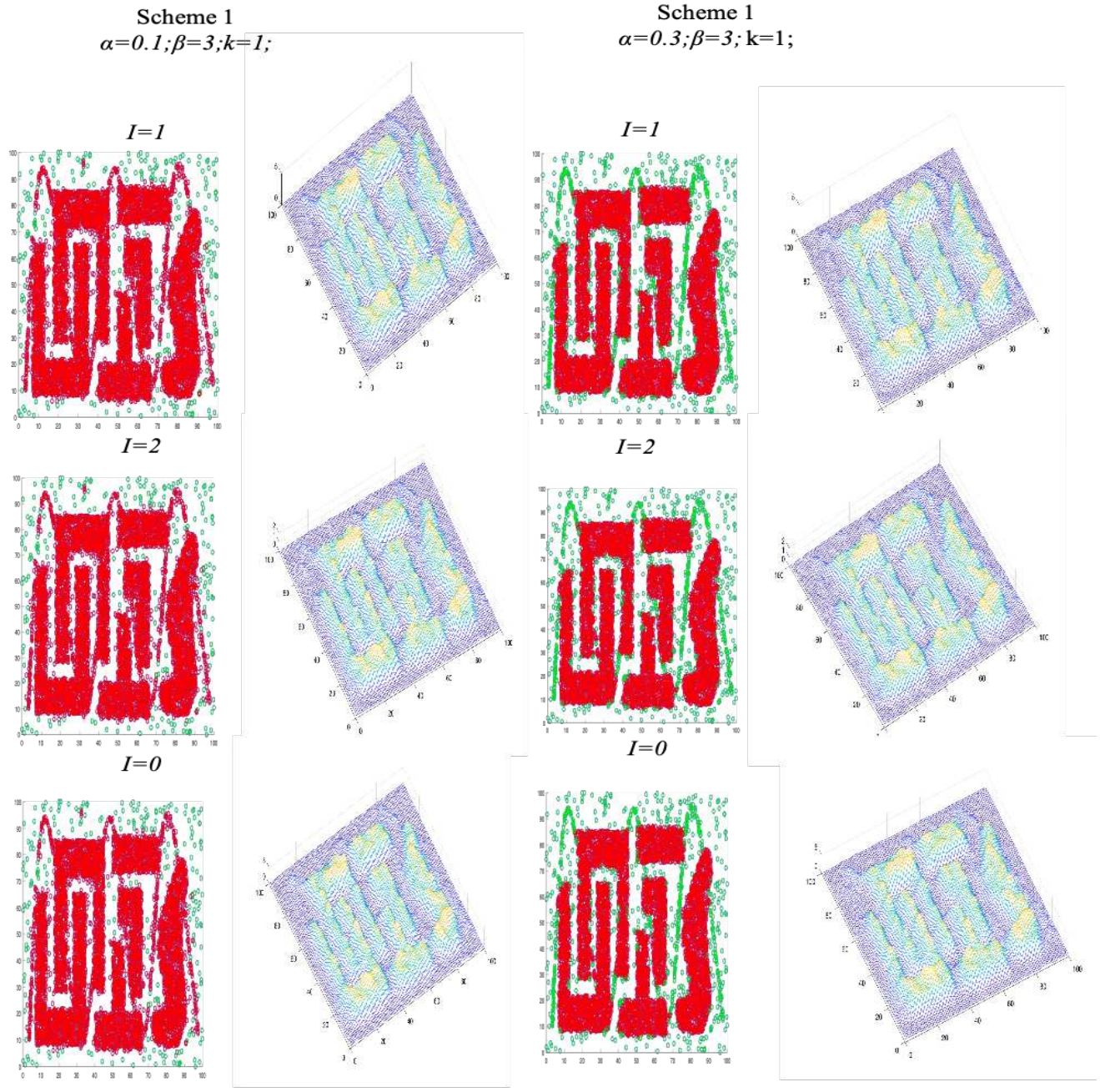


Fig 8. Outlier Detection results using Scheme 1 with two different values of α=0.1 and α=0.3. Outliers are shown in green and the inlier data in red.

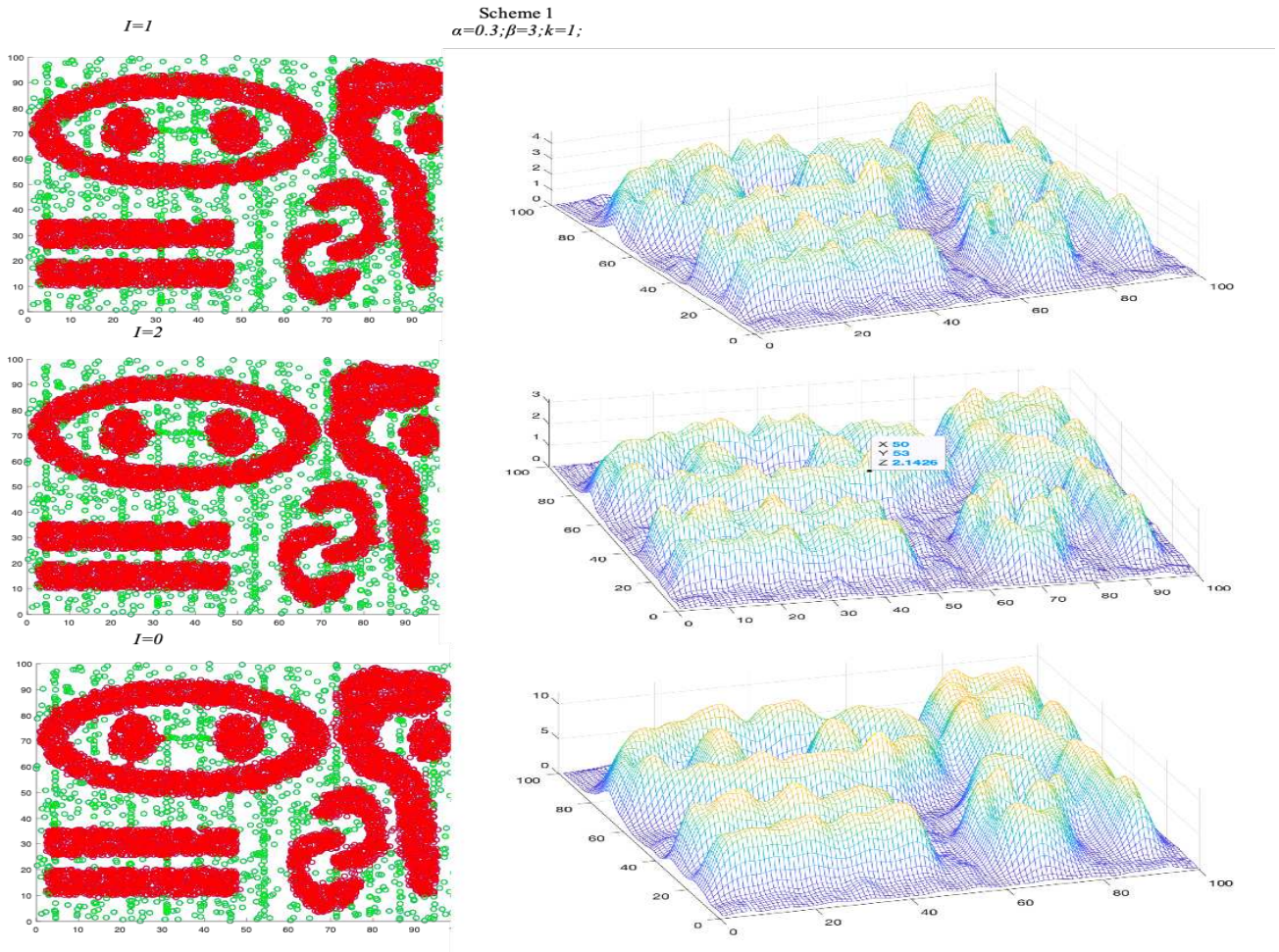


Fig 9. Second example of Outlier Detection results using Scheme 1 with α=0.3, β=3 and k=1. Outliers are shown in green and the inlier data in red. Row 1: I=1, Row 2: I=2 and Row 3: I=0.

5 Evaluation Of Boxplot Adjustments Using A Real Example

The ideas proposed for boxplot adjustments in equation (6) and equation (7) are also evaluated on a real dataset. The dataset used is a subset of the original KDD Cup 1999 dataset from the UCI machine learning repository, the subset used is still a large data containing 95,156 observations and three attributes. The dataset is publicly available online[1]. The ground truth of the dataset used is shown in Fig 12 (a), where the blue data points represent the actual inliers and the yellow points represent the actual outliers. The results for boxplot using extreme values defined in equation (6) are shown in Fig 12 (b) and the achieved value for Area Under Curve (AUC) evaluation parameter is 0.83 for this dataset.

[1] http://odds.cs.stonybrook.edu/smtp-kddcup99-dataset/

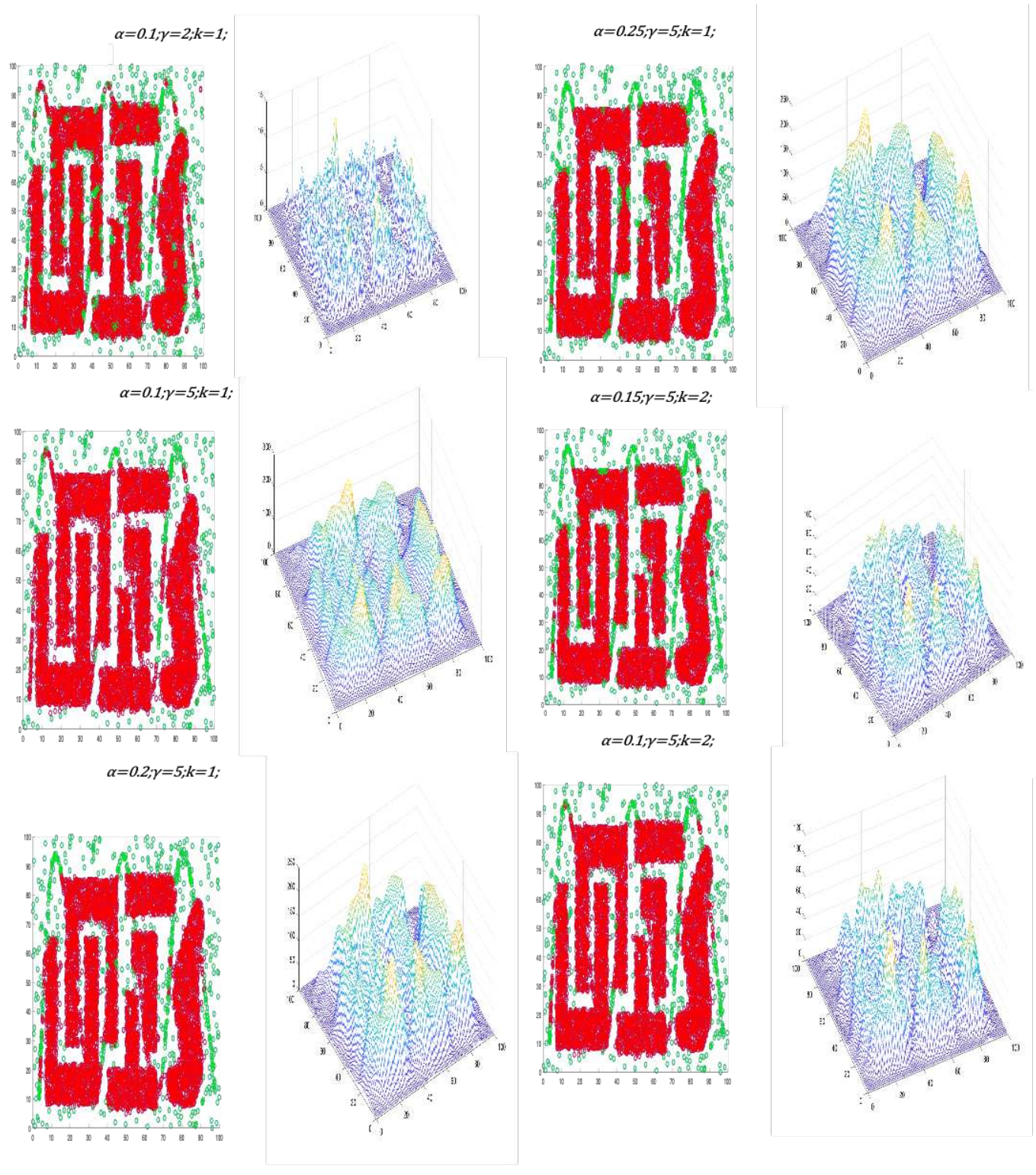


Fig 10. Outlier detection results using scheme 2 with different values of α, γ and k. Green points represent outliers and red points represent the normal data points.

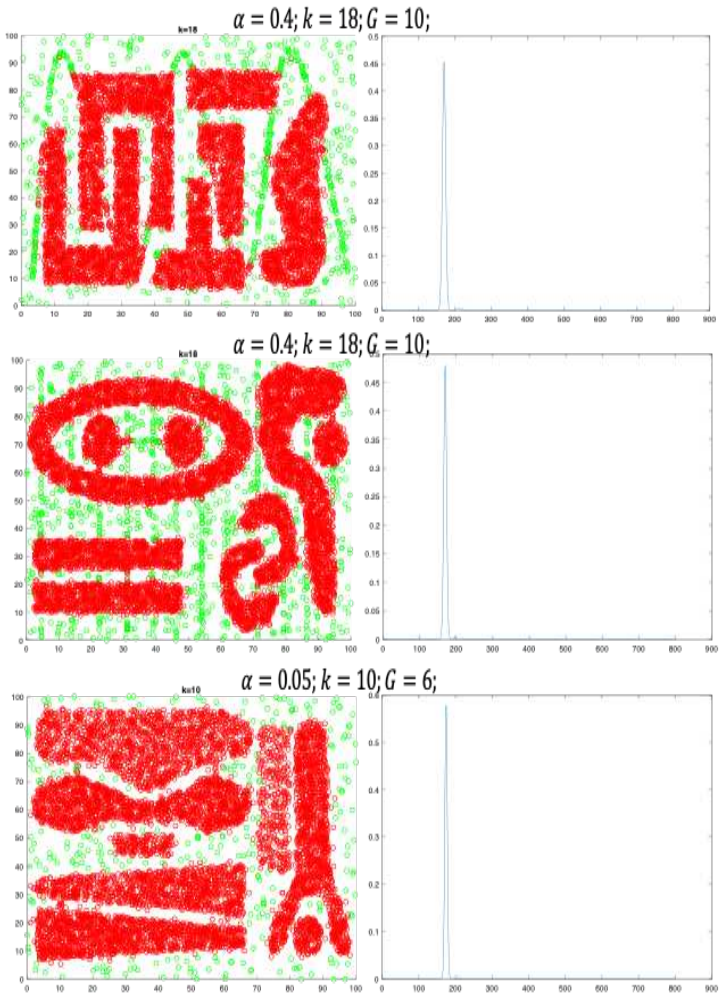


Fig 11. Results achieved on three different datasets with different distribution of noise using the proposed multiple gaussians estimation of d_k. The data points in red are the inliers and the green data points are identified as the outliers.

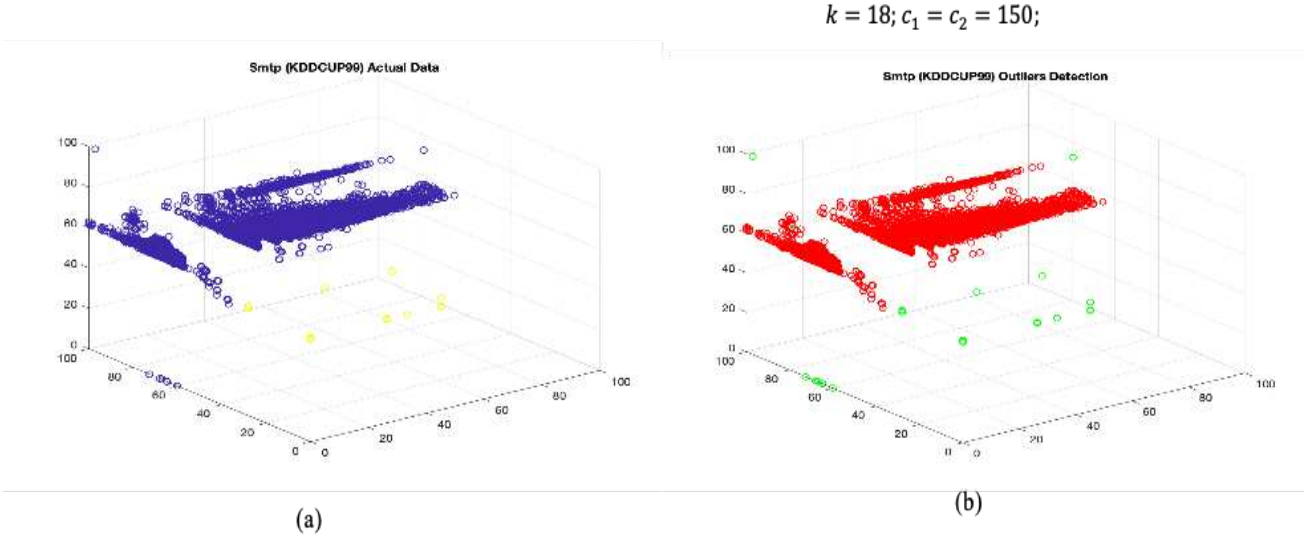


Fig 12. (a) Ground truth of the real data used to evaluate the proposed methods, blue data points are inliers and yellow are the outliers. (b) Results achieved using the boxplot with extreme values proposed in equation 6. Red points are inliers and green are outliers.

The results achieved for the proposed idea in (7) are shown in Fig 13 (a) and (b), respectively for equation 7 (a) and (b). The detected outliers are shown in green color and the inliers are shown in red color. The achieved value of AUC using both equations 7(a) and 7(b) is 0.833.

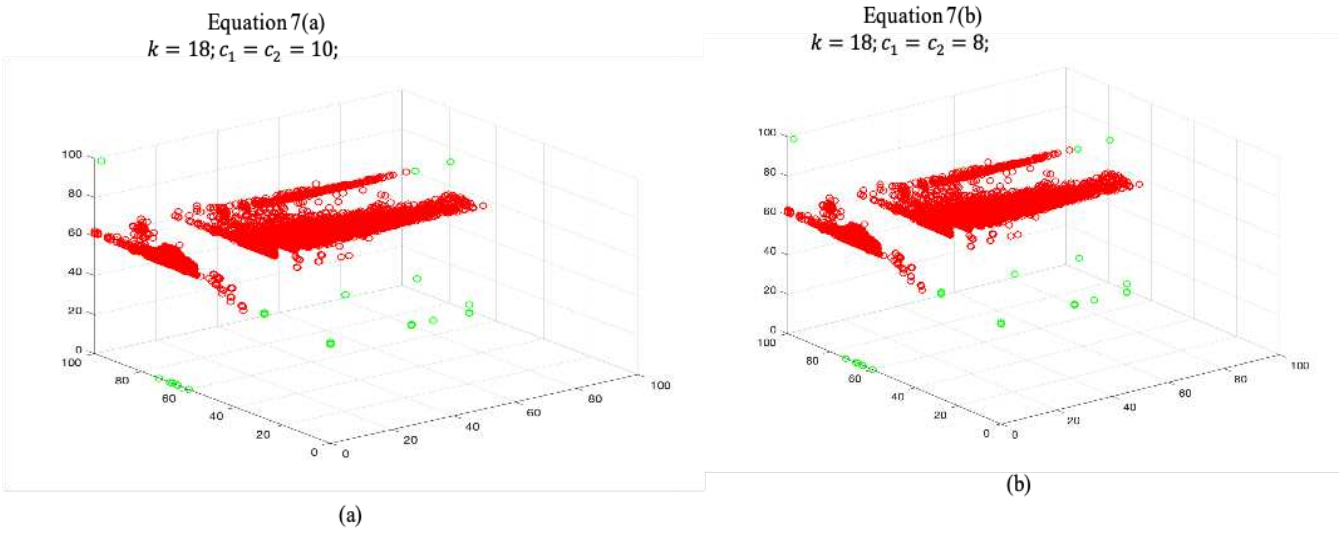


Fig 13. (a) Results achieved using the boxplot with extreme values proposed in equation 7(a). Red points are inliers and green are outliers. (b) Results achieved using the boxplot with extreme values proposed in equation 7(b). Red points are inliers and green are outliers.

6 Comparison With State-Of-Art

The proposed schemes are compared with several state of the art unsupervised outlier detection algorithms of similar kind, using a variety of benchmark datasets reported in [30]. The proposed schemes are found better in most of the cases as compared to the existing algorithms. The algorithms used for comparison include kNN [19], kNN-weight (kNNW) [31][20], Outlier detection using Indegree Number (ODIN) [32], Local Outlier Factor (LOF) [21], Simplified LOF (SLOF) [33], Connectivity based Outlier Factor (COF) [34], Influenced Outlierness (INFLO) [35], Local Outlier Probabilities (LoOP) [36], Local Distance-based Outlier Factor (LDOF) [37], Local Density Factor (LDF) [38], Kernel Density Estimation Outlier Score (KDEOS) [39] and Fast Angle-Based Outlier Detection (FastABOD) [40].

Initially, some of the fundamental outlier detection algorithms are compared with the proposed algorithms using the same three synthetic datasets. To test the unsupervised outlier detection methods, the most popular evaluation measure proposed in literature is based on the Receiver Operating Characteristics (ROC) and is computed as the Area Under Curve (AUC) [30]. The ROC AUC is computed for these three datasets using the proposed schemes

and are compared with some of the fundamental state-of-art algorithms in Table I. Hyperparameters of all the methods are tuned and the best results are reported. It can be seen from the results in Table I that the proposed schemes are performing better than the existing algorithms. As these three datasets are only two dimensional the visual comparison is also possible which is provided in Fig 14. From the visual inspection it is clearer that the newly proposed methods are better than the existing ones in identifying the outliers lying in close proximity to the inliers.

Furthermore, for a more comprehensive comparison, ten more benchmark datasets and twelve state-of-art methods reported in [30] are used and are compared with newly proposed unsupervised outlier detection methods. The results for state-of-art methods and the newly proposed methods for the ten benchmark datasets are given in Table II. It can be seen from Table II that the newly proposed methods are outperforming the existing algorithms in most of the cases. Furthermore, the two newly proposed methods reported in Table II make use of the distance vector only for detection of outliers, so irrespective of the dimensions of the input data the computational complexity of these proposed algorithms remains low.

A visual comparison of proposed methods with the existing state-of-art methods is also provided in Fig 15. From the results in Fig 15, it is clearer that the newly proposed scheme 3 is outperforming rest of the methods in terms of AUC. Furthermore, the required computational cost for the proposed method is also low because of using the d_k vector for outlier detection, instead of using the entire input data dimensions. As the proposed method is using only a single dimension distance vector of outlier detection, this makes it independent of the dimensions of the input data in terms of computational cost, which in turn makes it more feasible for high dimensional data.

TABLE I

COMPARISON USING THREE SYNTHETIC DATASETS

	State-of -the-Art					**Proposed**			
Dataset	KNN	ABOD	FastABOD	COF	LOF	**BADk**	**Scheme 1**	**Scheme 2**	**Scheme 3**
1	0.6222	0.7692	0.7431	0.9081	0.9240	0.9466	**0.9655**	0.9524	0.9574
2	0.5527	0.8526	0.8368	0.9047	0.9527	0.9776	**0.9816**	0.9779	0.9799
3	0.7281	0.8903	0.8951	0.8839	0.9435	0.9514	0.9025	0.9030	**0.9520**

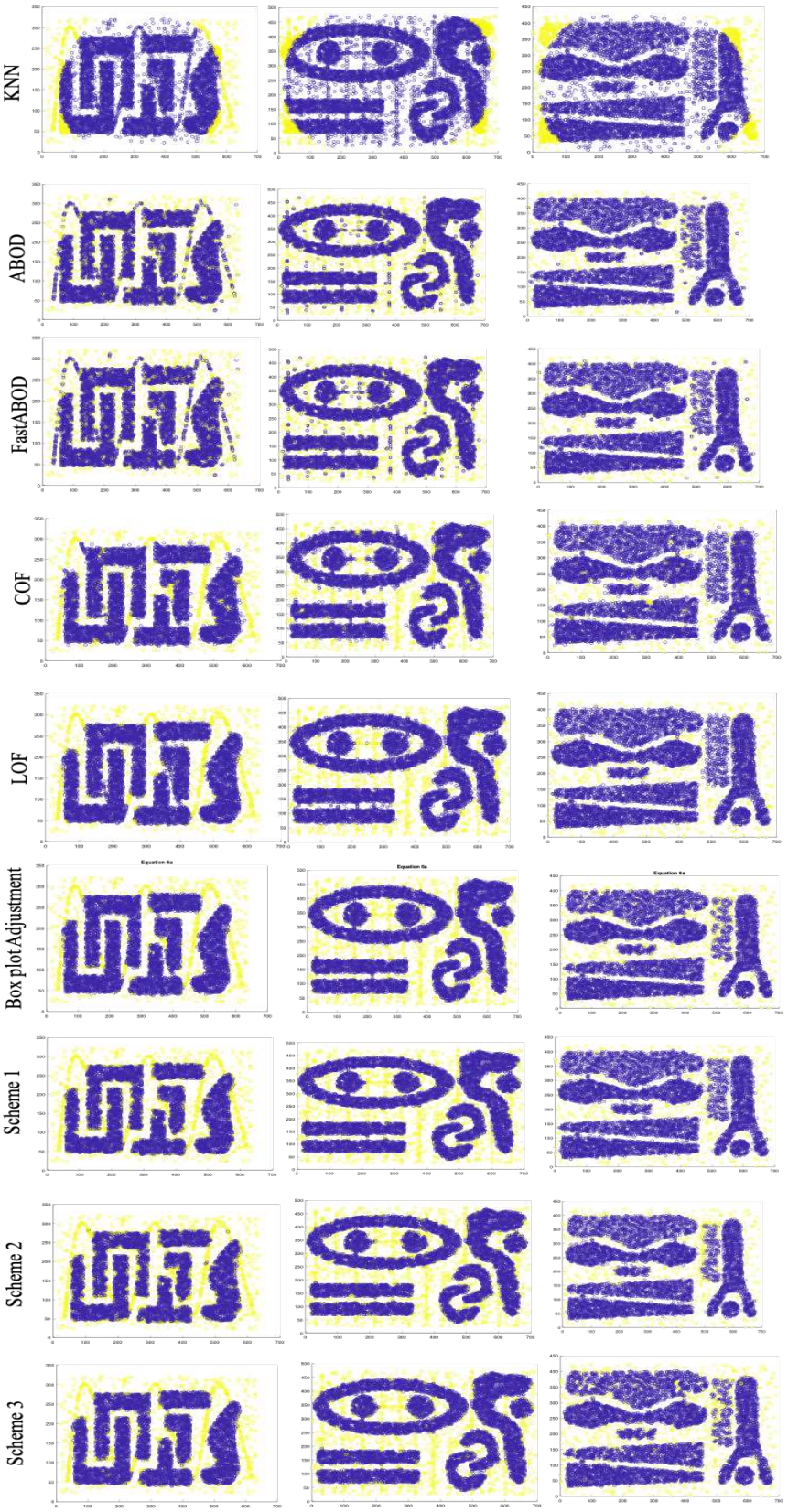


Fig14: Visual comparison for outlier detection using three synthetic datasets. Row 1-5: state-of-art methods and Row 6-9: the newly proposed methods.

TABLE II

ROC AUC VALUES COMPUTED ON DIFFERENT BENCHMARK DATASETS USING STATE-OF-ART ALGORITHMS AND THE PROPOSED SCHEMES

	State-of-art													Proposed
Dataset	**KNN**	**KNNW**	**LOF**	**SLOF**	**LoOP**	**LDOF**	**ODIN**	**FABOD**	**KDEOS**	**LDF**	**INFLO**	**COF**	**BADk**	**Scheme 3**
Arrhythmia	0.7930	0.7674	0.7674	0.7500	0.7500	0.7551	0.7663	0.7715	0.6680	0.8565	0.7950	0.7663	0.8074	**0.8770**
heartdisease	0.8644	0.8311	0.8533	0.7800	0.7977	0.7577	0.8544	0.8911	0.6844	0.8933	0.8333	0.8955	0.8467	**0.9400**
hepatits	0.8955	0.8706	0.9452	0.8806	0.8905	0.8706	0.9228	0.8408	0.8706	**0.9403**	0.9005	0.8955	0.8393	**0.9403**
parkinson	0.9895	0.9895	0.9895	0.9895	0.9895	0.9895	0.9895	0.9895	0.9895	**1**	0.9895	0.9791	0.9583	**1**
spambase40	0.5734	0.5661	0.4738	0.5011	0.4965	0.4796	0.5191	0.4372	0.4766	0.5364	0.4738	0.4994	0.6034	**0.6125**
Glass	0.8748	0.8832	0.8666	0.8650	0.8395	0.7788	0.7292	0.8579	0.7420	0.9035	0.8037	0.8953	0.9122	**0.9293**
pendigit	**0.9868**	0.9854	0.9168	0.9107	0.9039	0.7181	0.9225	0.9610	0.8113	0.9545	0.8908	0.9609	0.9826	0.9819
shuttle	0.9890	0.9861	0.9896	0.9869	0.9869	0.9775	0.9888	0.8381	0.9810	**0.9922**	0.9863	0.9920	0.9865	0.9895
WBC	0.9929	0.9920	0.9906	0.9835	0.9737	0.9582	0.9563	0.9892	0.6023	0.9929	0.9821	0.9863	0.9842	**0.9989**
WPBC	0.5409	0.5319	0.5254	0.5018	0.5018	0.5034	0.5072	0.5341	0.5185	**0.5829**	0.4957	0.5568	0.5427	0.5801

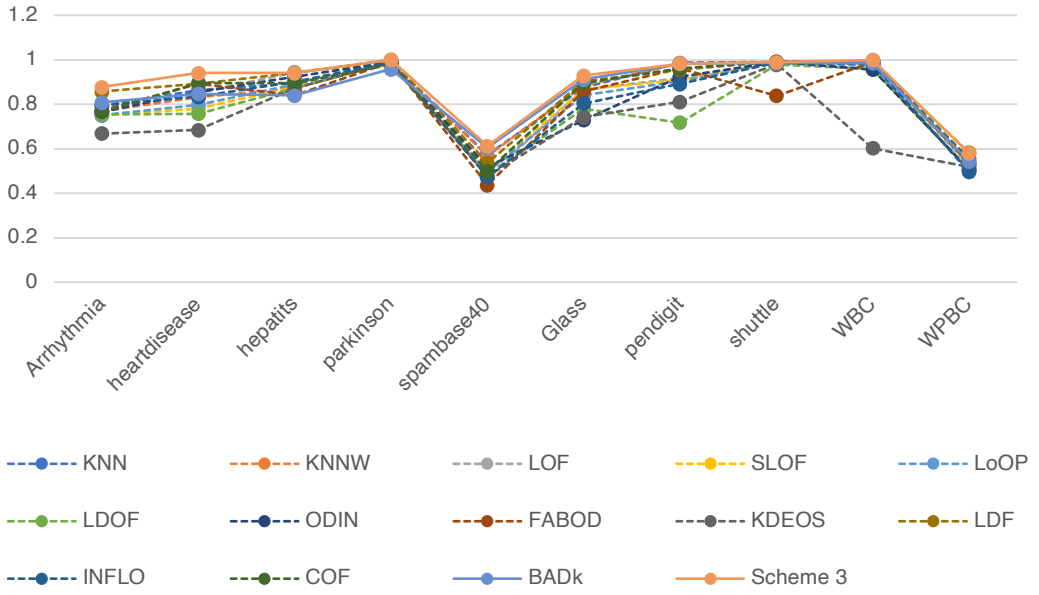


Fig15: Comparison of the proposed schemes with the state-of-art methods using 10 benchmark datasets for outlier detection. Y-axis represents the computed ROC AUC values.

7 CONCLUSIONS

Outlier detection is one of the most important preprocessing steps in data analytics, and for best performance consideration, it is considered a vital step for machine learning algorithms. Different methods are presented in this paper, keeping in view the need for a robust and easy-to-implement outlier detection algorithm. The newly proposed methods are based on novel statistical techniques considering data compactness, which resulted in an added advantage of easy implementation, improved accuracy, and low computational cost. Furthermore, to demonstrate the proposed ideas' performance, ten benchmark multidimensional datasets and three complex synthetic two-dimensional datasets containing the different shapes of clusters contaminated with a mixture of varying noise distributions are used. The proposed methods are found accurate and better in terms of outlier detection as compared to the state-of-art. It is also an observation that some of the fundamental state-of-art methods cannot detect the outliers in scenarios where the outliers are a mixture of two different distributions. Moreover, two of the newly proposed schemes use only a single dimension distance-vector instead of utilizing the entire data dimensions for outlier detection. This makes the proposed methods more feasible and computationally inexpensive, irrespective of the input data's large sizes.

LIST OF ABBREVIATIONS:

ABOD: Angle-Based Outlier Detection.

AUC: Area Under Curve.

BADk: Boxplot Adjustments using D-k-NN.

COF: Connectivity based Outlier Factor.

D-k-NN: Distance vector considering k number of Nearest Neighbors.

INFLO: Influenced Outlierness.

KDEOS: Kernel Density Estimation Outlier Score

LDF: Local Density Factor.

LDOF: Local Distance-based Outlier Factor.

LE: Lower Extreme Bound.

LOF: Local Outlier Factor.

LoOP: Local Outlier Probabilities.

ODIN: Outlier Detection using Indegree Number.

PDF: Probability Density Function.

ROC: Receiver Operating Characteristics.

SLOF: Simplified LOF.

UE: Upper Extreme Bound.

DECLARATIONS:

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable

Availability of data and materials: The datasets analysed during the current study are publicly available at

http://odds.cs.stonybrook.edu/smtp-kddcup99-dataset/

https://www.dbs.ifi.lmu.de/research/outlier-evaluation/DAMI/

Competing interests: Authors declare no competing interest.

Funding: Not applicable.

Authors' contributions: Both the authors have equally contributed in this work.

Acknowledgment: Authors would like to thank Qatar National Library (QNL) for supporting the publication charges of this article.

Authors information:

Atiq Ur Rehman received the master's degree in computer engineering from the National University of Sciences and Technology (NUST), Pakistan, in 2013 and PhD degree in Computer Science and Engineering from Hamad Bin Khalifa University, Qatar in 2019. He is currently working as a Post doc researcher with the College of Science and Engineering, Hamad Bin Khalifa University, Qatar. His research interests include the development of pattern recognition and machine learning algorithms.

Samir Brahim Belhaouri received the master's degree in telecommunications from the National Polytechnic Institute (ENSEEIHT) of Toulouse, France, in 2000, and the Ph.D. degree in Applied Mathematics from the Federal Polytechnic School of Lausanne (EPFL), in 2006. He is currently an associate professor in the Division of Information and Communication Technologies, College of Science and Engineering, HBKU. He also holds and leads several academic and administrator positions, Vice Dean for Academic & Student Affairs at College of Science and General Studies and University Preparatory Program at ALFAISAL university (KSA), University of Sharjah (UAE), Innopolis University (Russia), Petronas University (Malaysia), and EPFL Federal Swiss school (Switzerland). His main research interests include Stochastic Processes, Machine

Learning, and Number Theory. He is now working actively on developing algorithms in machine learning applied to visual surveillance and biomedical data, with the support of several international fund for research in Russia, Malaysia, and in GCC.

REFERENCES

[1] Zhu, Jinlin, Z. Ge, Z. Song, and F. Gao, "Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data," *Annu. Rev. Control*, vol. 46, pp. 107–133, 2018.

[2] G. H. McClelland, *Nasty data: Unruly, ill-mannered observations can ruin your analysis. Handbook of research methods in social and personality psychology*. Cambridge: Cambridge University Press, 2000.

[3] B. Frénay and M. Verleysen, "Reinforced extreme learning machines for fast robust regression in the presence of outliers," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 3351–3363, 2015.

[4] X. Wang, X. Wang, M. Wilkes, X. Wang, X. Wang, and M. Wilkes, "Developments in Unsupervised Outlier Detection Research," *New Dev. Unsupervised Outlier Detect.*, pp. 13–36, 2021.

[5] A. Zimek and P. Filzmoser, "There and back again: Outlier detection between statistical reasoning and data mining algorithms," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 6, p. e1280, 2018.

[6] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, 2009.

[7] B. Angelin and A. Geetha, "Outlier Detection using Clustering Techniques-K-means and K-median," *Proc. Int. Conf. Intell. Comput. Control Syst. ICICCS 2020*, pp. 373–378, 2020.

[8] L. Bergman and Y. Hoshen, "Classification-Based Anomaly Detection for General Data," *arXiv*, 2020.

[9] A. Wahid and C. S. R. Annavarapu, "NaNOD: A natural neighbour-based outlier detection algorithm," *Neural Comput. Appl.*, 2020.

[10] P. D. Domański, "Study on Statistical Outlier Detection and Labelling," *Int. J. Autom. Comput.*, 2020.

[11] Y. Dong, S. B. Hopkins, and J. Li, "Quantum entropy scoring for fast robust mean estimation and improved outlier detection," *arXiv*, 2019.

[12] O. Shetta and M. Niranjan, "Robust subspace methods for outlier detection in genomic data circumvents the curse of dimensionality," *R. Soc. Open Sci.*, vol. 7, no. 2, p. 190714, 2020.

[13] P. Li and O. Niggemann, "Non-convex hull based anomaly detection in CPPS," *Eng. Appl. Artif. Intell.*, vol. 87, no. 103301, 2020.

[14] A. Borghesi, A. Bartolini, M. Lombardi, M. Milano, and L. Benini, "Anomaly detection using autoencoders in high performance computing systems," *CEUR Workshop Proc.*, vol. 2495, pp. 24–32, 2019.

[15] E. Knorr and R. Ng, "A unified notion of outliers: properties and computation," in *In: Proceedings of the 3rd ACM international conference on knowledge discovery and data mining (KDD), Newport Beach*, 1997, pp. 219–222.

[16] E. Knorr and R. Ng, "Algorithms for mining distance-based outliers in large datasets," in *In: Proceedings of the 24th international conference on very large data bases (VLDB), New York*, 1998, pp. 392–403.

[17] G. Wu *et al.*, "A fast kNN-based approach for time sensitive anomaly detection over data streams," in *International Conference on Computational Science*, 2019, pp. 59–74.

[18] R. Zhu *et al.*, "KNN-Based Approximate Outlier Detection Algorithm Over IoT Streaming Data," *IEEE Access*, vol. 8, pp. 42749–42759, 2020.

[19] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *In: Proceedings of the ACM international conference on management of data (SIGMOD), Dallas*, 2000, pp. 427–438.

[20] F. Angiulli and C. Pizzuti, "Outlier Mining in large high-dimensional data sets," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 2, pp. 203–215, 2005.

[21] M. Breunig, H. Kriegel, R. Ng, and J. Sander, "LOF: identifying density-based local outliers," in *In: Proceedings of the ACM international conference on management of data (SIGMOD), Dallas*, 2000, pp. 93–104.

[22] J. W. Tukey, "Exploratoy Data Analysis," *Addison-Wesley Ser. Behav. Sci.*, 1977.

[23] A. C. Kimber, "Exploratory Data Analysis for Possibly Censored Data from Skewed Distributions," *Appl. Stat.*, vol. 39, pp. 21–30, 1990.

[24] L. Aucremanne, G. Brys, M. Hubert, P. J. Rousseeuw, and A. Struyf, "A Study of Belgian Inflation, Relative Prices and Nominal Rigidities using New Robust Measures of Skewness and Tail Weight," *Theory Appl. Recent Robust Methods*, pp. 13–25, 2004.

[25] N. C. Schwertman, M. A. Owens, and R. Adnan, "A Simple More General Boxplot Method for Identifying Outliers," *Comput. Stat. Data Anal.*, vol. 47, pp. 165–174, 2004.

[26] M. Hubert and E. Vandervieren, "An adjusted boxplot for skewed distributions," *Comput. Stat. Data Anal.*, vol. 52, no. 12, pp. 5186–5201, 2008.

[27] S. B. Belhaouari, S. Ahmed, and S. Mansour, "Optimized K-means algorithm," *Math. Probl. Eng.*, vol. 2014, 2014.

[28] N. Distribution, “Encyclopedia.com: https://www.encyclopedia.com/social-sciences/applied-and-social-sciences-magazines/distribution-normal,” *Gale Encyclopedia of Psychology*. .

[29] G. Casella and R. L. Berger, “Statistical Inference,” *(2nd ed.). Duxbury. ISBN 978-0-534-24312-8.*, 2001.

[30] G. O. Campos *et al.*, “On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study,” *Data Minning Knowl. Discov.*, vol. 30, pp. 891–927, 2016.

[31] F. Angiulli and C. Pizzuti, “Fast outlier detection in high dimensional spaces,” in *In: Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD), Helsinki*, 2002, pp. 15–26.

[32] V. Hautamäki, I. Kärkkäinen, and P. Fränti, “Outlier detection using k-nearest neighbor graph,” in *In: Proceedings of the 17th international conference on pattern recognition (ICPR), Cambridge*, 2004, pp. 430–433.

[33] E. Schubert, A. Zimek, and H. Kriegel, “Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection,” *Data Minning Knowl. Discov.*, vol. 28, no. 1, pp. 190–237, 2014.

[34] J. Tang, Z. Chen, A. Fu, and D. Cheung, “Enhancing effectiveness of outlier detections for low density patterns,” in *In: Proceedings of the 6th Pacific-Asia conference on knowledge discovery and data mining (PAKDD), Taipei*, 2002, pp. 535–548.

[35] W. Jin, A. Tung, J. Han, and W. Wang, “Ranking outliers using symmetric neighborhood relationship,” in *In: Proceedings of the 10th Pacific-Asia conference on knowledge discovery and data mining (PAKDD), Singapore*, 2006, pp. 577–593.

[36] H. Kriegel, P. Kröger, E. Schubert, and A. Zimek, “LoOP: local outlier probabilities,” in *In: Proceedings of the 18th ACM conference on information and knowledge management (CIKM), Hong Kong*, 2009, pp. 1649–1652.

[37] K. Zhang, M. Hutter, and H. Jin, “A new local distance-based outlier detection approach for scattered real- world data,” in *In: Proceedings of the 13th Pacific-Asia conference on knowledge discovery and data mining (PAKDD), Bangkok*, 2009, pp. 813–822.

[38] L. Latecki, A. Lazarevic, and D. Pokrajac, “Outlier detection with kernel density functions,” in *In: Proceedings of the 5th international conference on machine learning and data mining in pattern recognition (MLDM), Leipzig*, 2007, pp. 61–75.

[39] E. Schubert, A. Zimek, and H. Kriegel, “Generalized outlier detection with flexible kernel density estimates,” in *In: Proceedings of the 14th SIAM International Conference on Data Mining (SDM), Philadelphia*, 2014, pp. 542–550.

[40] H. Kriegel, M. Schubert, and A. Zimek, “Angle-based outlier detection in high-dimensional data,” in *In: Proceedings of the 14th ACM international conference on knowledge discovery and data mining (SIGKDD), Las Vegas*, 2008, pp. 444–452.

Figures

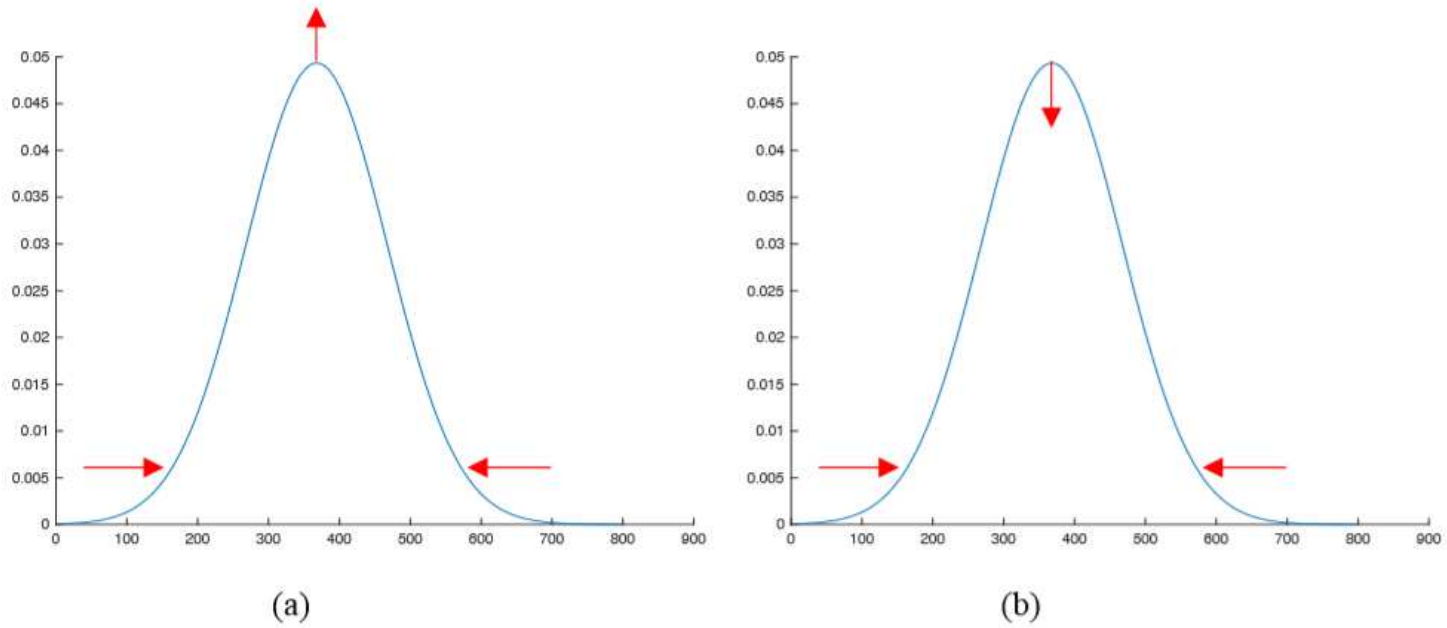


Figure 1

a) Effect of traditional gaussian approach on compression. (b) Effect of proposed scheme 2 approach on compression in x and y axes.

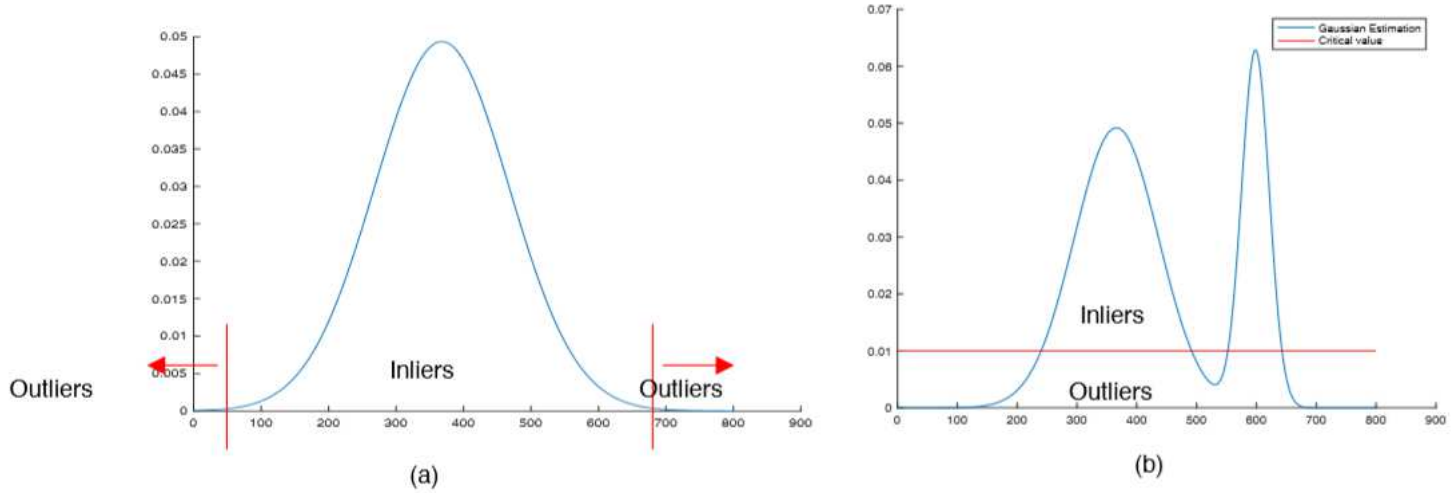


Figure 2

An Example of Gaussian estimation and marking of critical value for outlier detection. (a) Points those lie outside the red boundaries are considered outliers. (b) The points below the critical value are identified as outliers.

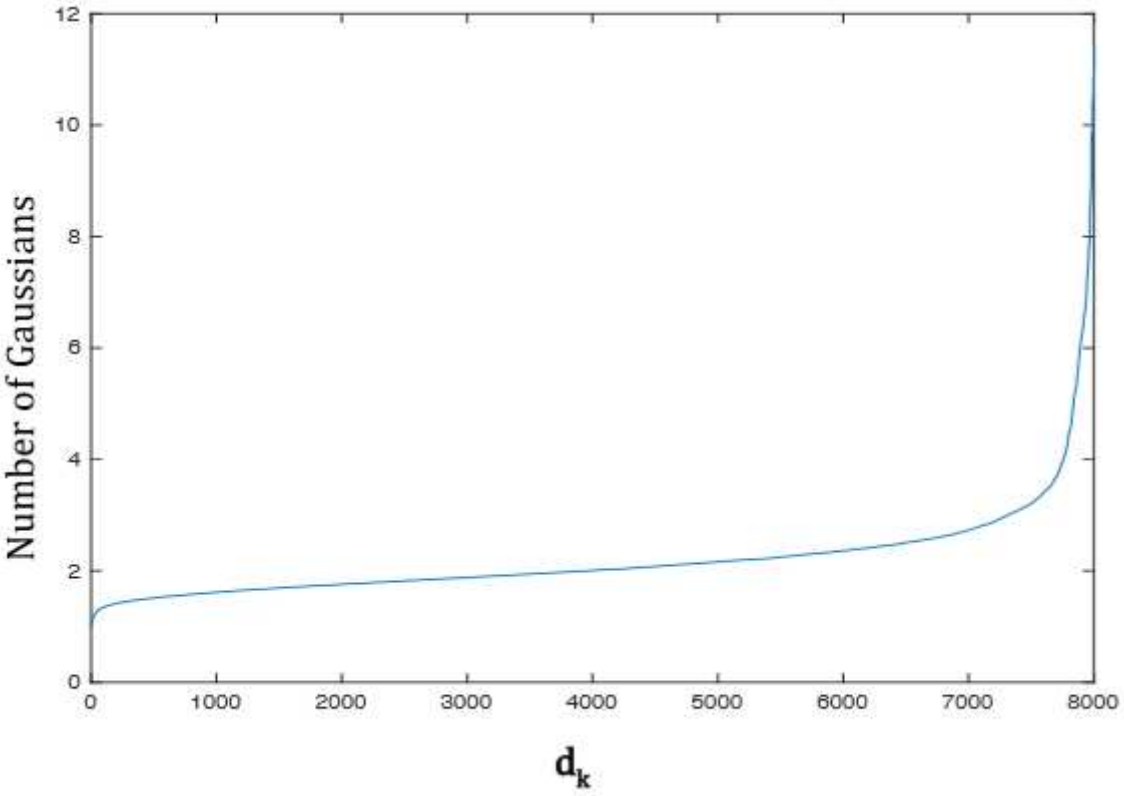


Figure 3

Plot of sorted values of the vector ▯▯

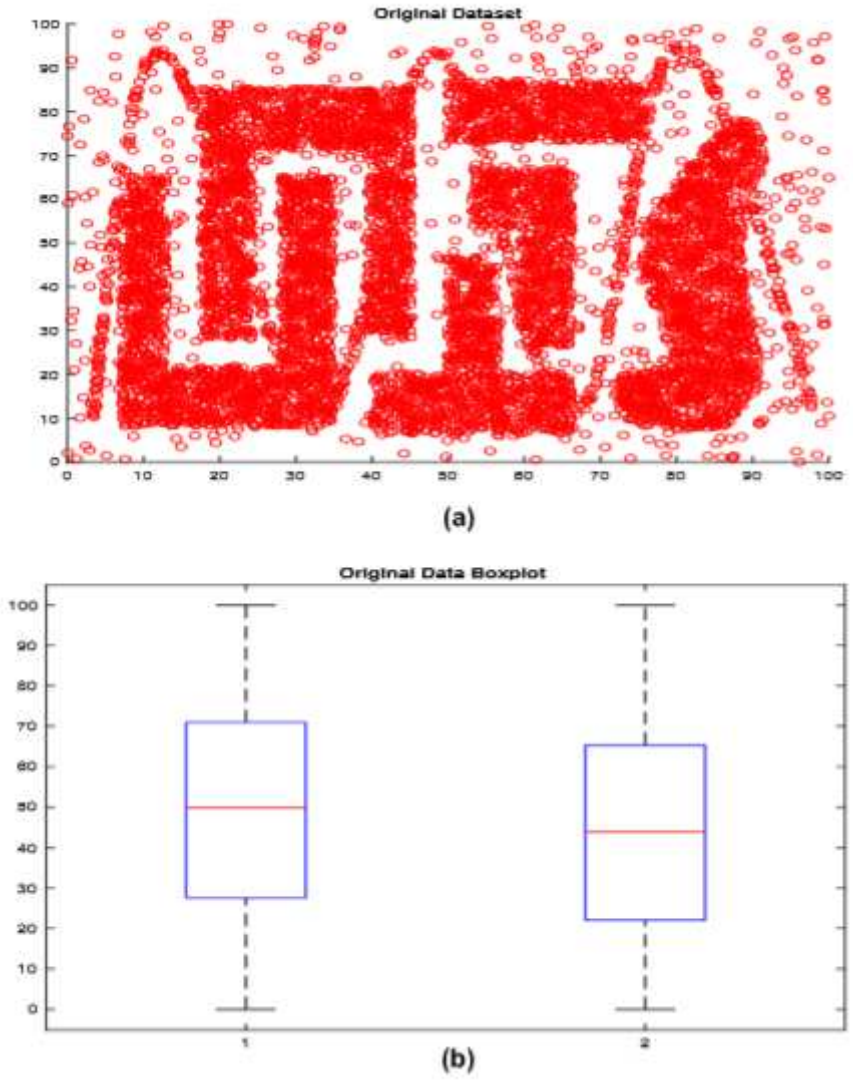


Figure 4

(a) Original data representing clusters with different shapes and a mixture of sinusoidal and gaussian noise. (b) Traditional boxplot showing no outliers/anomalies.

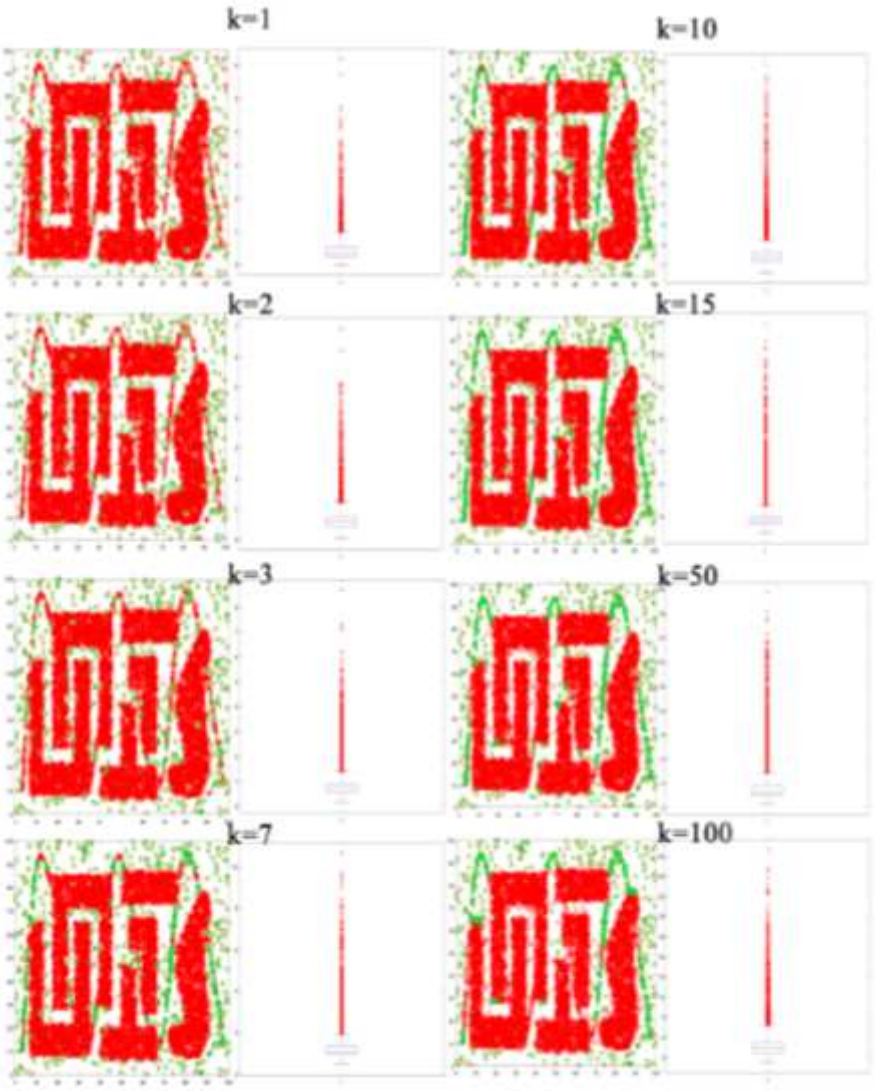


Figure 5

Outliers detection from the data shown in Fig 4(a), using the proposed boxplot with extreme values defined in equation (6) for different values of . The data in red is identified as the inliers while the data in green is identified as the outliers.

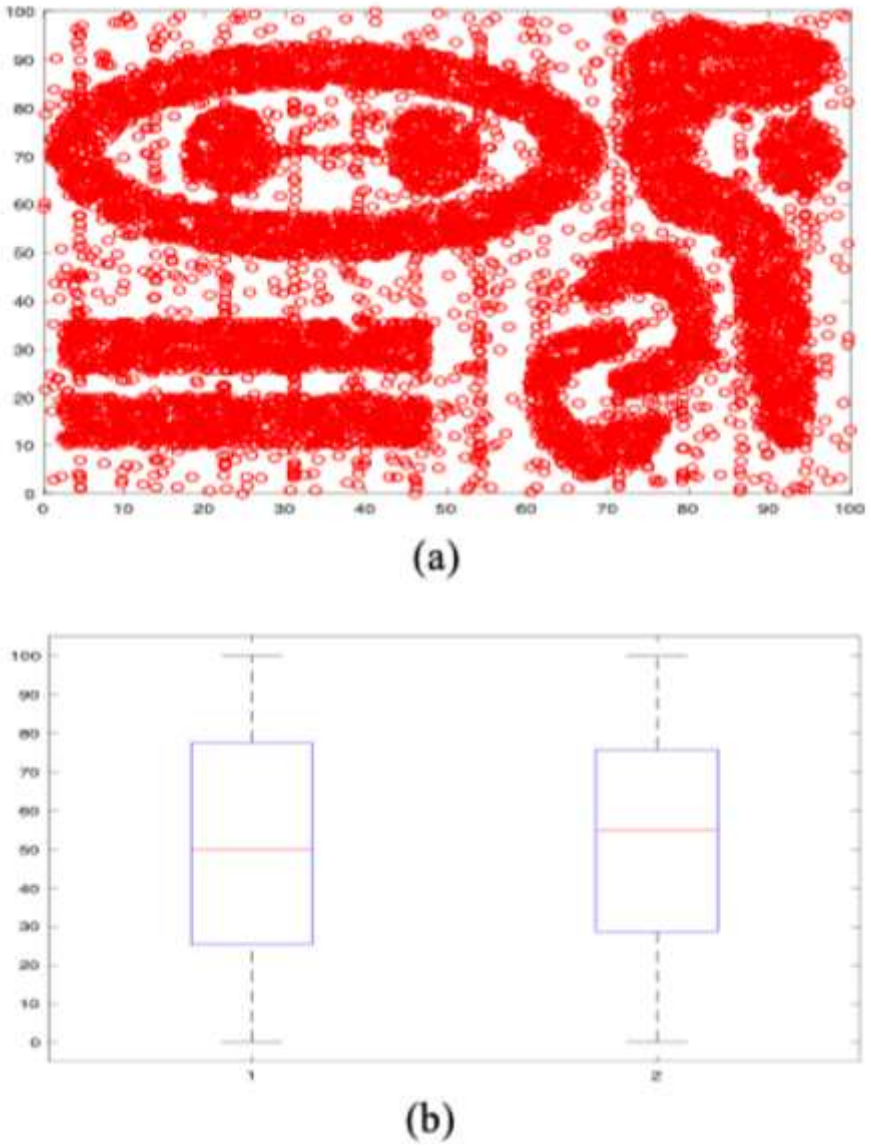


Figure 6

Outliers detection from the data shown in Fig 4(a), using the proposed boxplot with extreme values defined in equation (6) for different values of . The data in red is identified as the inliers while the data in green is identified as the outliers.

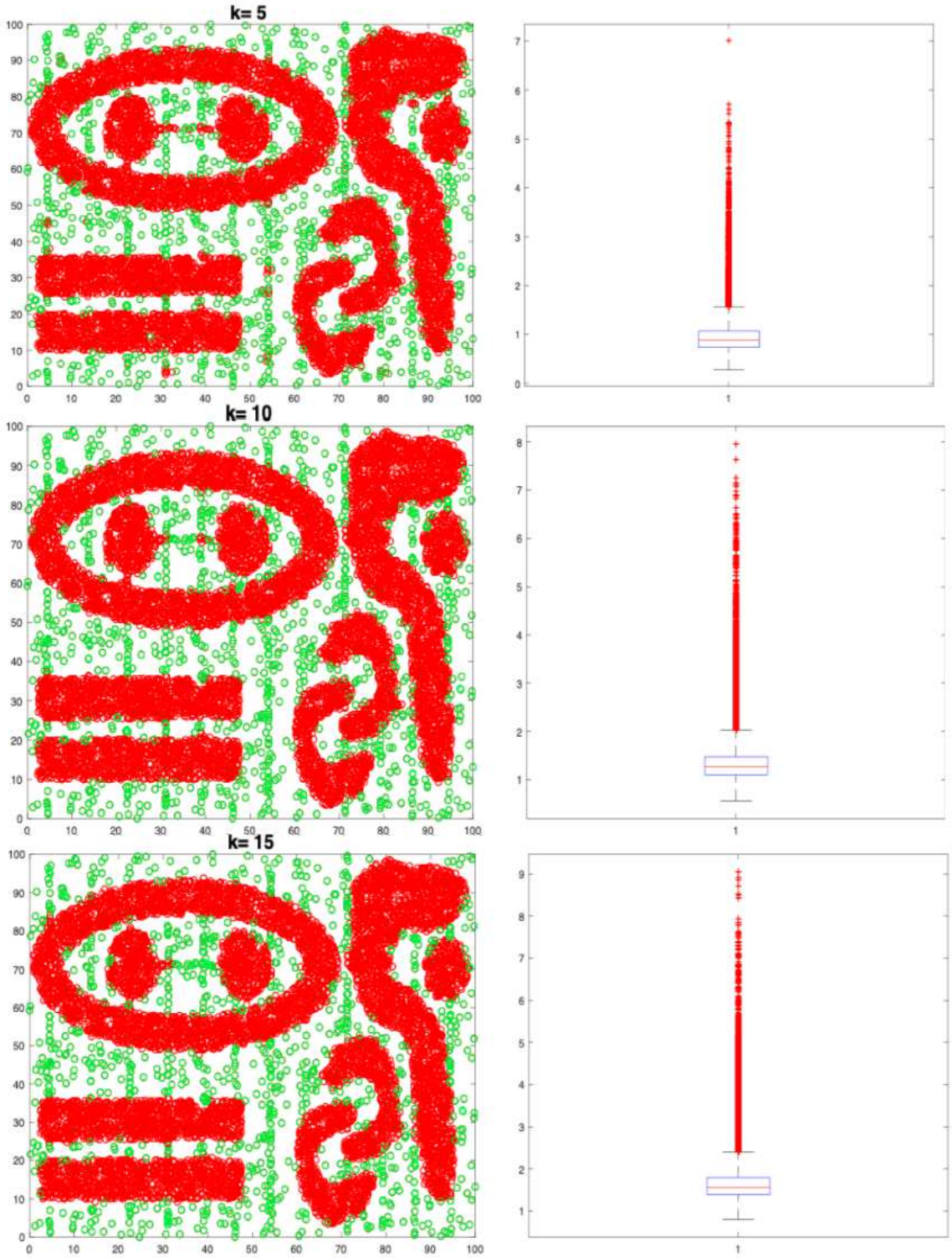


Figure 7

Outliers detection from the data shown in Fig 6 (a), using the proposed boxplot with extreme values defined in equation (6) for different values of . The data in red is identified as the inliers while the data in green is identified as the outliers.

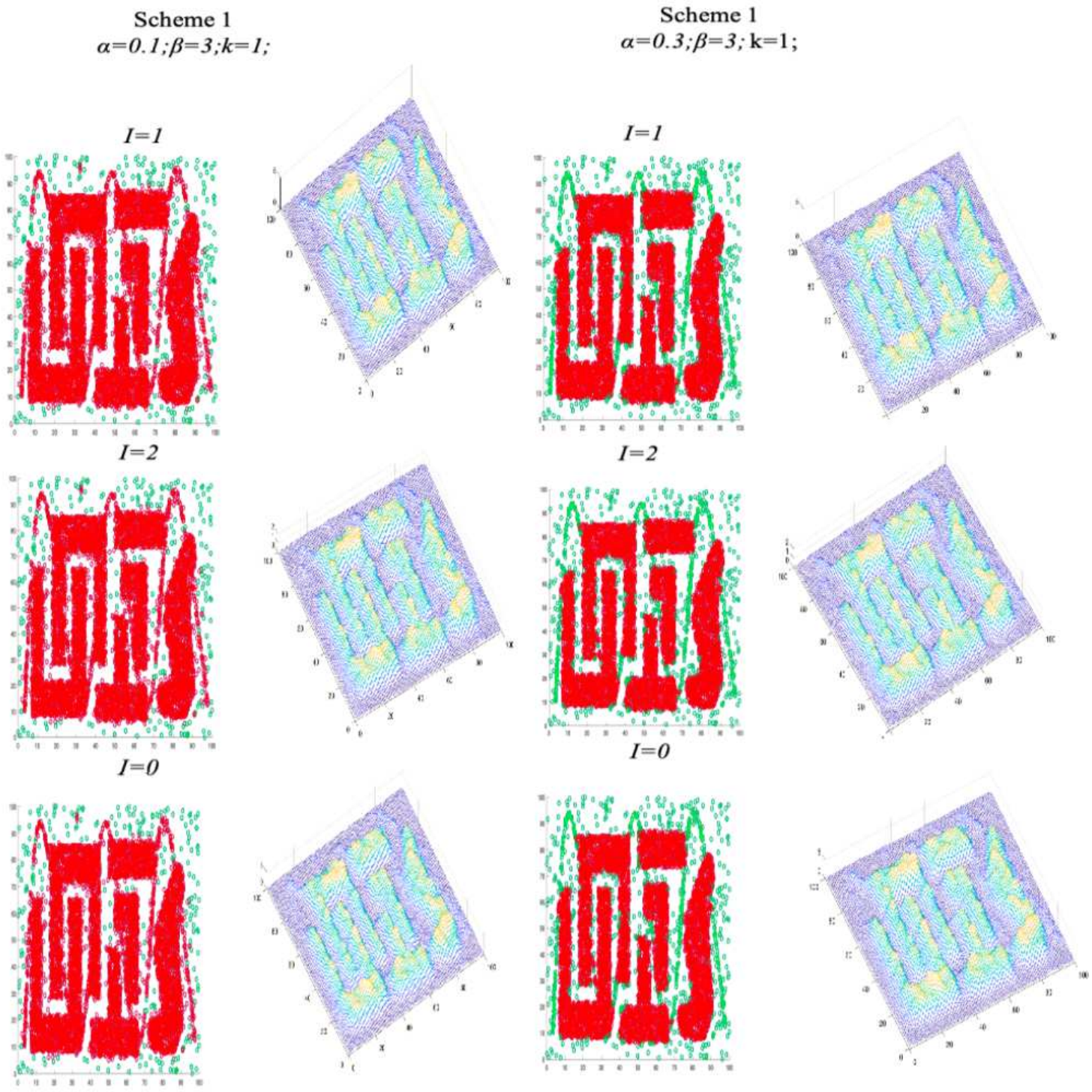


Figure 8

Outlier Detection results using Scheme 1 with two different values of α=0.1 and α=0.3. Outliers are shown in green and the inlier data in red.

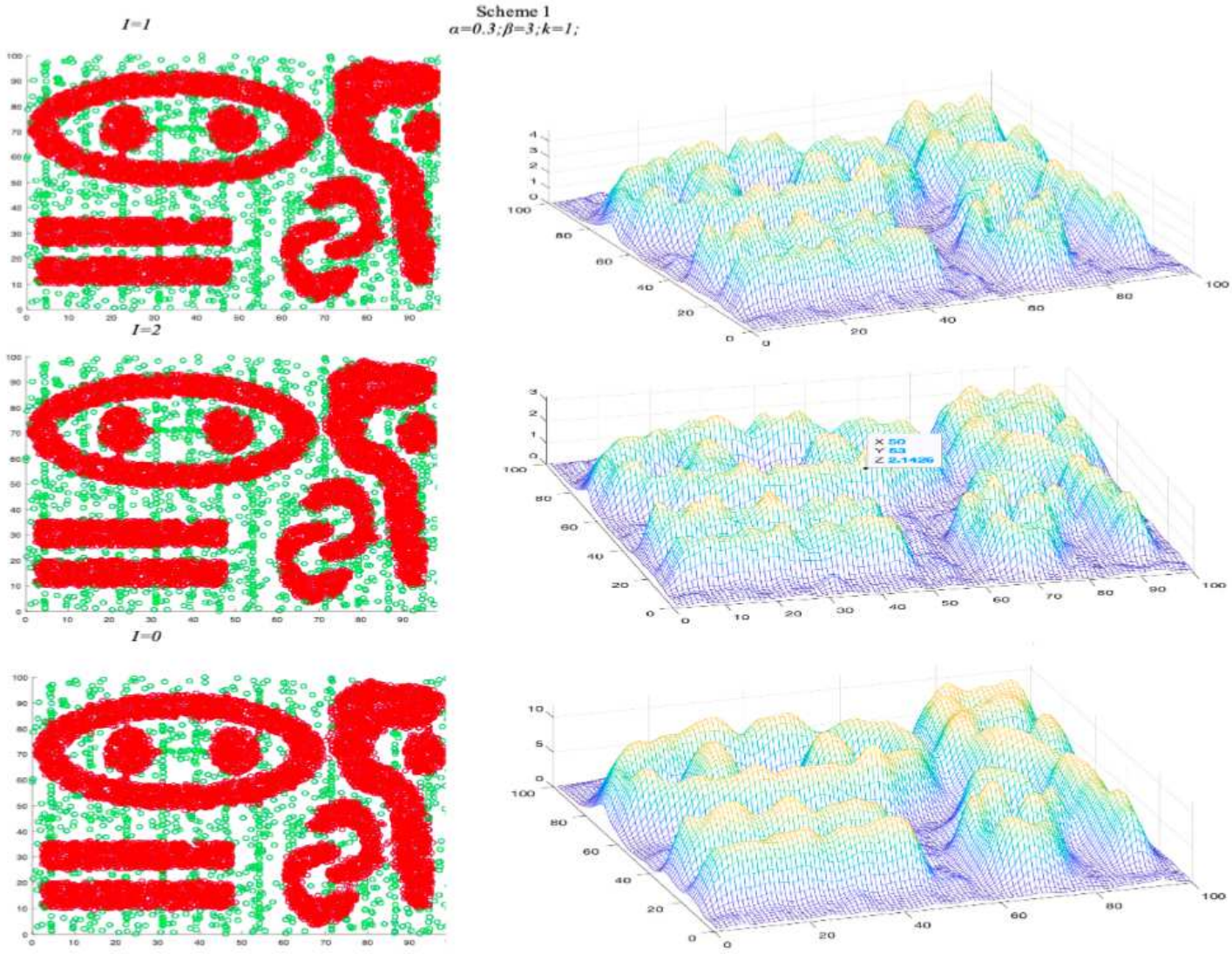


Figure 9

Second example of Outlier Detection results using Scheme 1 with α=0.3, β=3 and k=1. Outliers are shown in green and the inlier data in red. Row 1: I=1, Row 2: I=2 and Row 3: I=0.

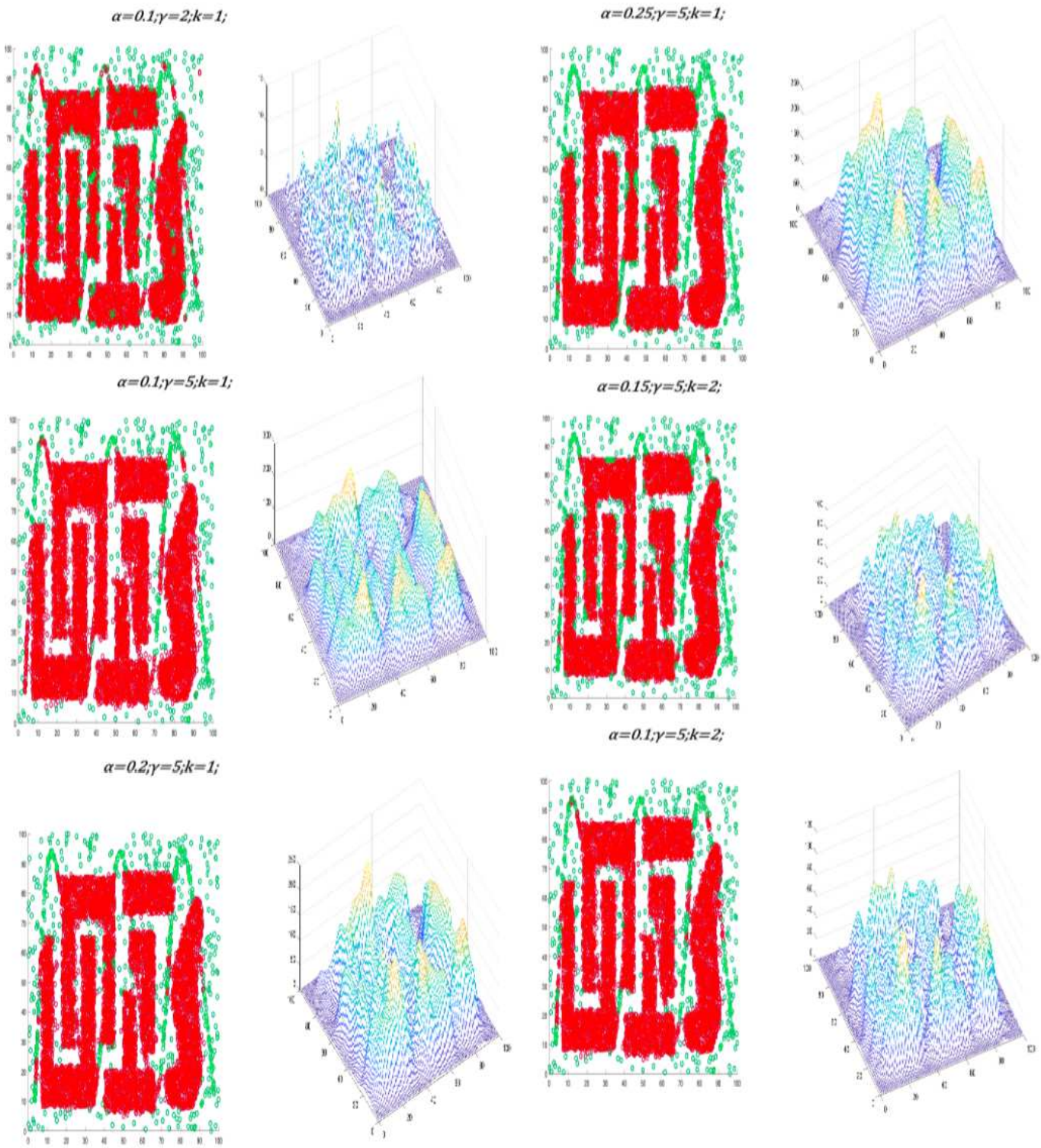


Figure 10

Outlier detection results using scheme 2 with different values of α, γ and k. Green points represent outliers and red points represent the normal data points.

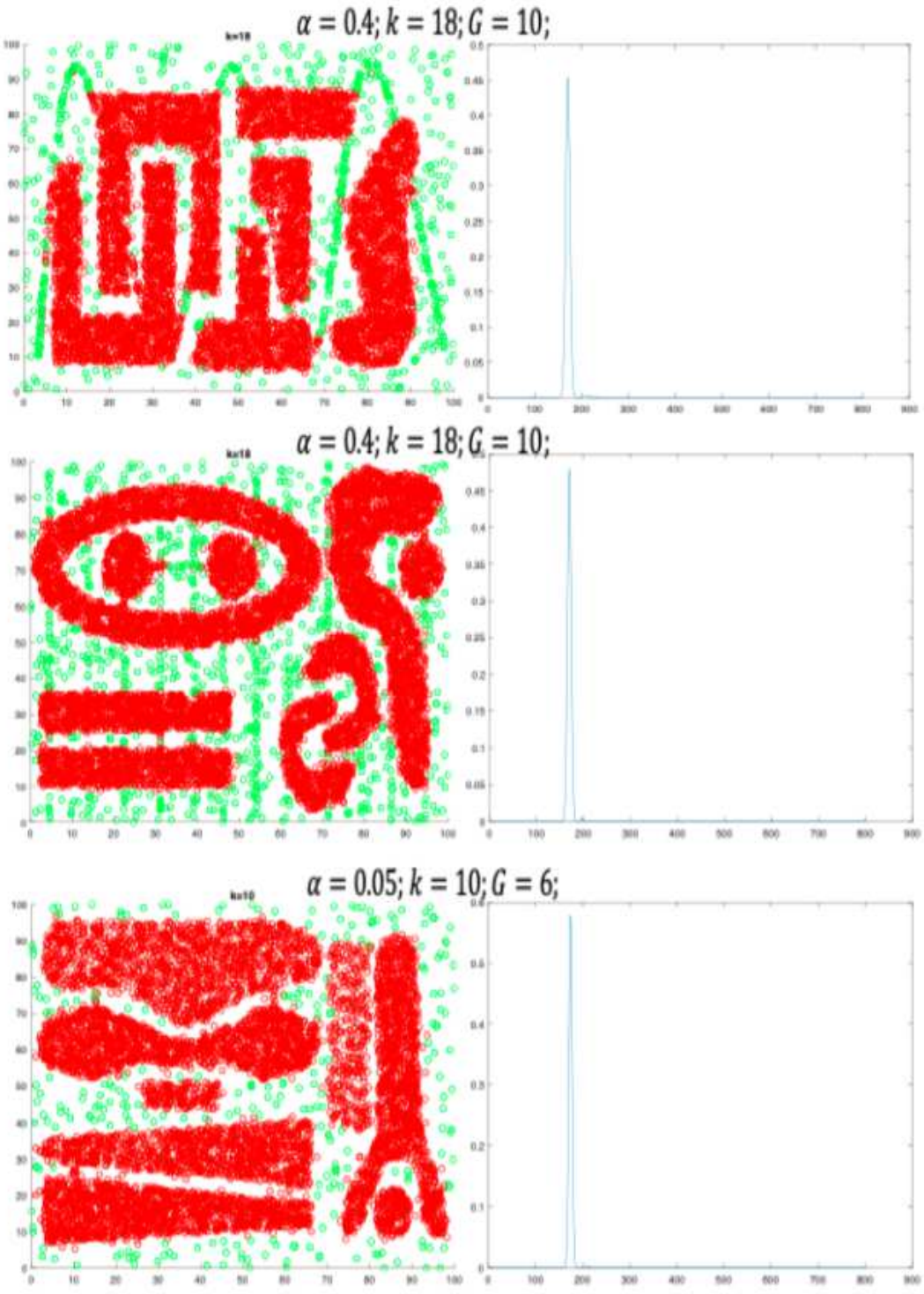


Figure 11

Results achieved on three different datasets with different distribution of noise using the proposed multiple gaussians estimation of P. The data points in red are the inliers and the green data points are identified as the outliers.

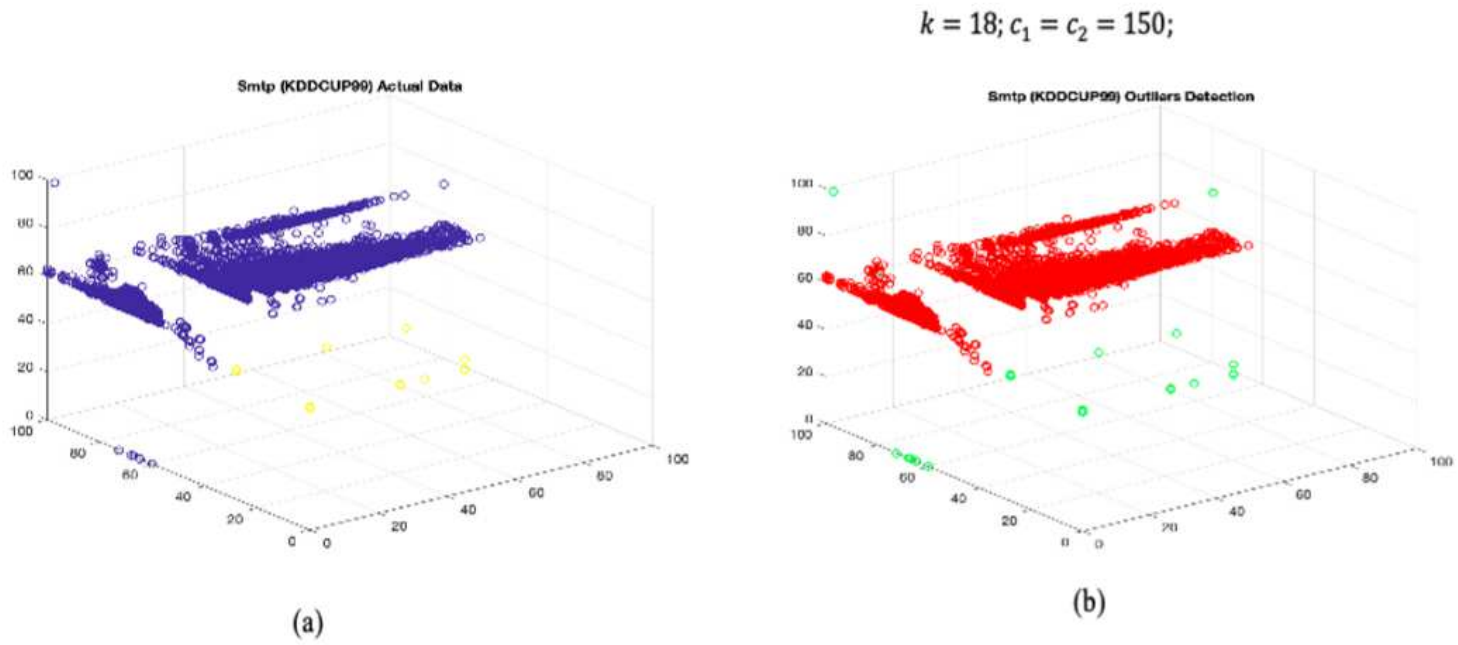


Figure 12

(a) Ground truth of the real data used to evaluate the proposed methods, blue data points are inliers and yellow are the outliers. (b) Results achieved using the boxplot with extreme values proposed in equation 6. Red points are inliers and green are outliers.

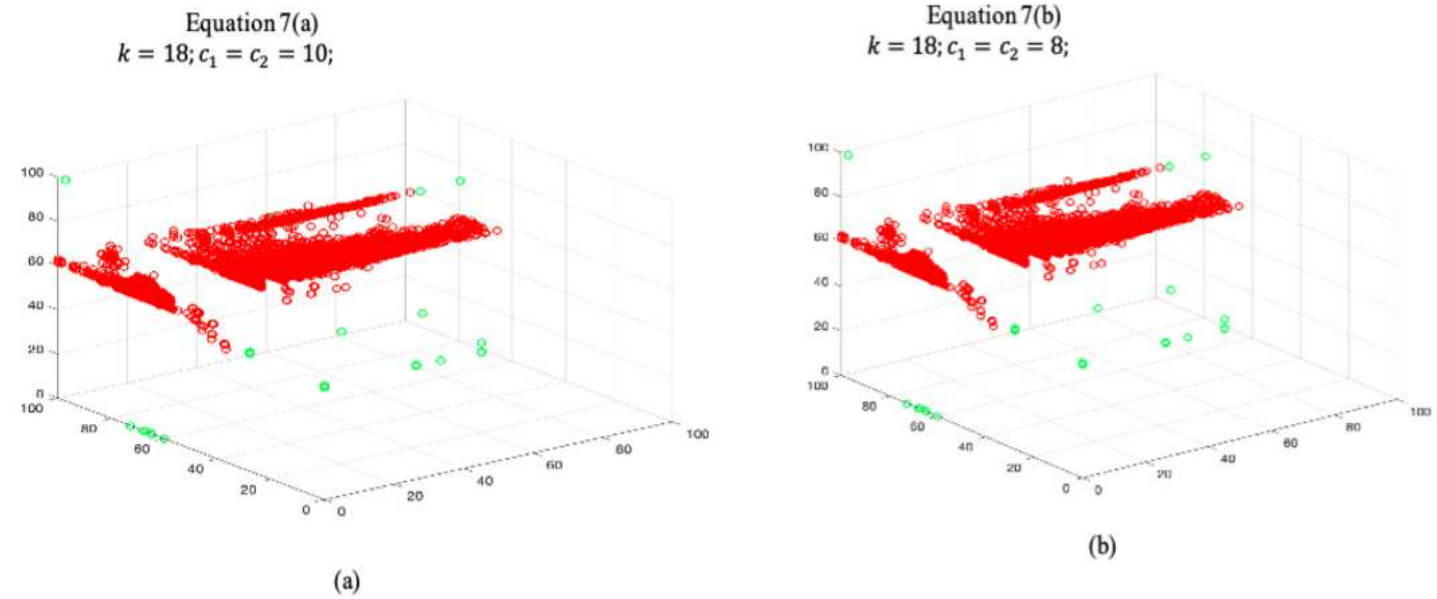


Figure 13

(a) Results achieved using the boxplot with extreme values proposed in equation 7(a). Red points are inliers and green are outliers. (b) Results achieved using the boxplot with extreme values proposed in equation 7(b). Red points are inliers and green are outliers.

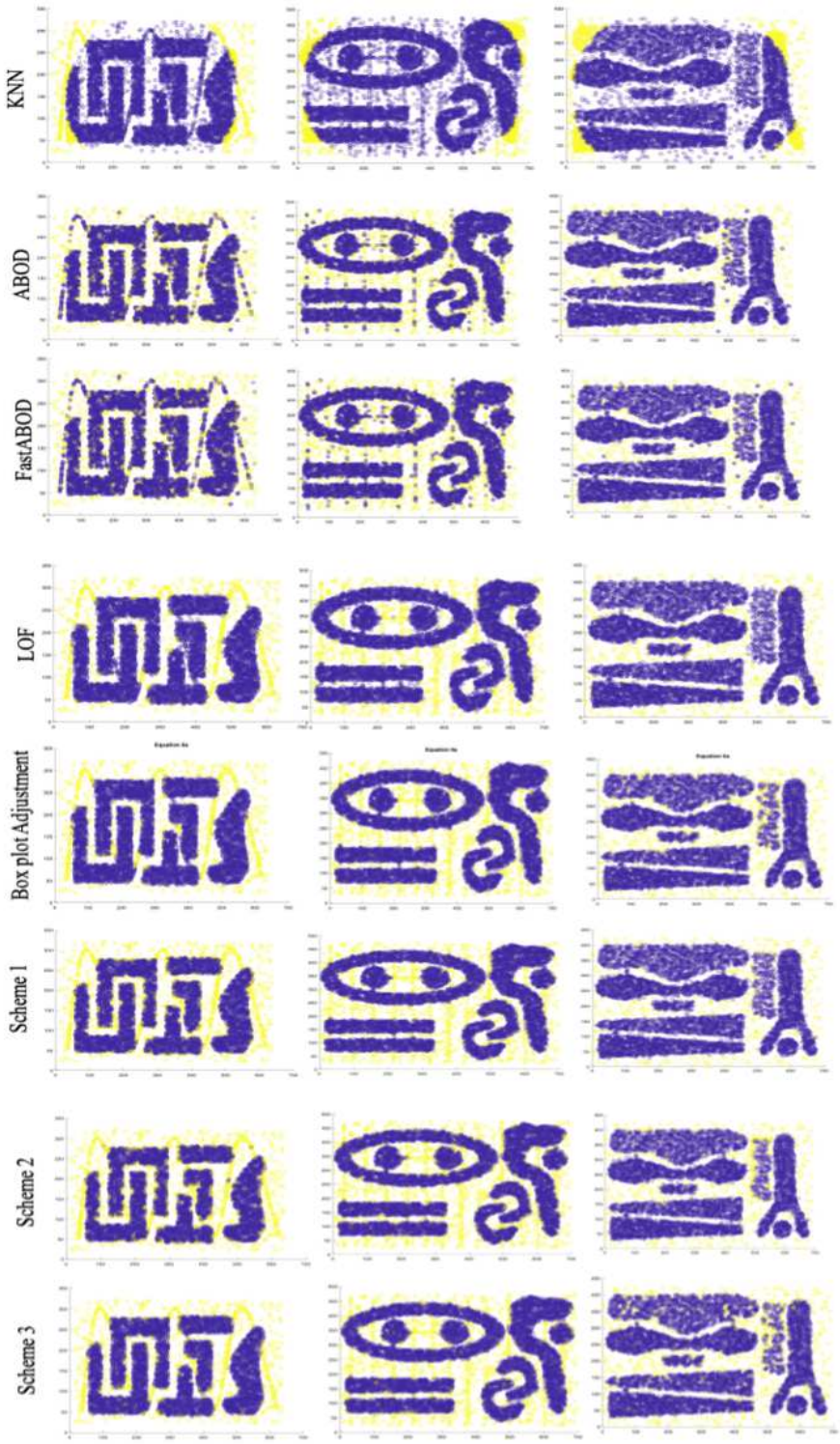


Figure 14

Visual comparison for outlier detection using three synthetic datasets. Row 1-5: state-of-art methods and Row 6-9: the newly proposed methods.

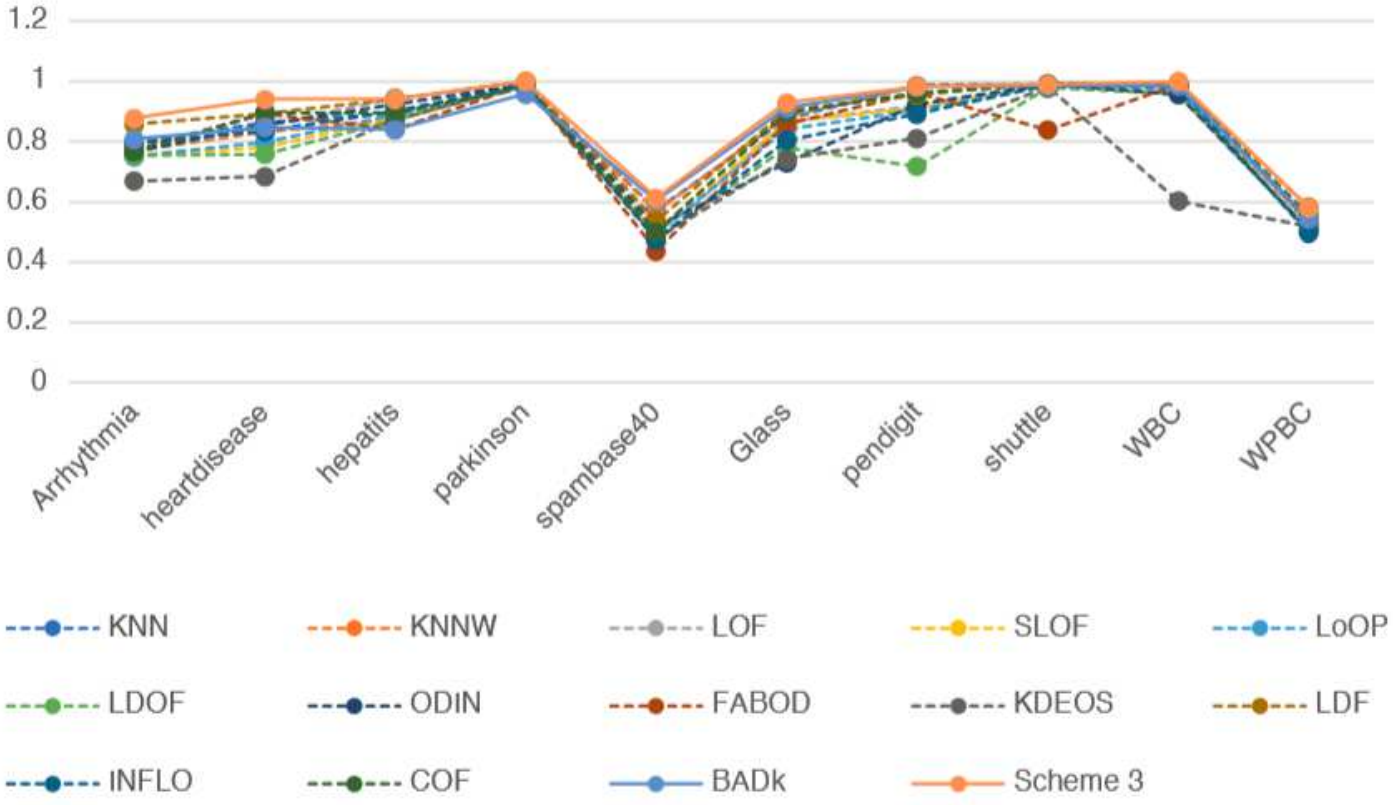


Figure 15

Comparison of the proposed schemes with the state-of-art methods using 10 benchmark datasets for outlier detection. Y-axis represents the computed ROC AUC values.