Robust Sparse k-means Clustering Against STOF Observations

Basso Madjoukeng, Ariel; Frénay, Benoît; Kenmogne, Edith Belise

DOI:
10.36227/techrxiv.174114537.78354861/v1

Publication date:
2025

Document Version
Version revue par les pairs

Robust Sparse k-means Clustering Against STOF Observations

Ariel Basso Madjoukeng[1*], Edith Bélise Kenmogne[2] and Benoît Frénay[1]

[1]Faculty of Computer Science, University of Namur, Rue Grandgagnage 21, Namur, 5000, Belgium.

[2]Computer Science, University of Dschang, 96, Dschang, Cameroon.

*Corresponding author(s). E-mail(s): ariel.bassomadjoukeng@unamur.be;

Abstract

k-means is a clustering algorithm used to group observations into clusters. Due to the multidimensionality of datasets, interpreting clustering results has become increasingly challenging. In response, sparse clustering variants have emerged, allowing each feature to be weighed. In the sparse ***k***-means algorithm, feature weights are computed based on the values of their associated observations. However, the sparse ***k***-means algorithm is known to be sensitive to outliers. Hence, robust sparse ***k***-means variants have emerged, performing sparse ***k***-means while detecting outliers. In numerous real-world cases, data entry or measurement errors can lead to poorly collected values for a feature, making them significantly different from other values in that feature. Due to dataset multidimensionality, these observations are often not detected as outliers by existing robust approaches. This negatively impacts the evaluation of feature weights, biases the interpretability of results, and leads to poor clustering quality. To fill this gap, this paper introduces a new robust sparse ***k***-means framework consisting of a new robust initialization and a detection method of these observations. The proposed robust initialization method shows robustness in terms of the observations chosen as initial centers; The proposed sparse ***k***-means shows an improvement in feature selection, interpretability and clustering quality compared to other robust variants on several real and synthetic datasets.

Keywords: Clustering, Sparse ***k***-means, Robustness, Interpretability, STOF observation.

1 Introduction

Clustering is a machine learning task aiming to group observations in clusters. In this field, several algorithms exist, among which k-means [1], DBSCAN [2] and others. Due to its simplicity and its efficiency, the k-means algorithm has become one of the most popular clustering algorithm [3]. It consists of grouping observations in clusters while minimizing the distance between observations in each cluster. In real-world applications, for several reasons, interpreting clustering results can be a complex task. The first reason is the multidimensionality of modern datasets. In fact, in many application areas, data are multidimensional (data high number of features), which increases clustering complexity and makes visualization more difficult. Second, many datasets contain observations that are significantly different from others, known as outliers. Several studies have shown the significant impact of outliers on clustering results. Faced with the issues related to outliers, robust k-means variants have been developed, enabling simultaneously clustering and outliers removal [4, 5, 6]. To avoid the issues related to interpretability, the sparse k-means algorithm has been introduced [7]. It is a variant of the k-means algorithm that performs clustering by weighting features.

Feature weights typically represent their relative importance in the clustering process. According to feature weight, it is possible to determine whether the feature is relevant for the clustering task or not, enabling thus to mitigate the issues related to dimensionality.

In the sparse k-means algorithm and its robust variants, feature weights are computed based on their observed values. At each iteration, feature values are multiplied by their computed weights. However, measurement or input errors can lead to incorrect recordings for one or a few features of some observations. Naively removing outliers, as done in the literature, may fail to identify these observations as outliers, since only a few feature values are abnormal and most of their features have plausible values. In such cases, these observations are included in feature contribution calculations, even for features where they are abnormal. This is problematic, as it biases the contribution of those features, affects all feature observations over iterations, and leads to poor results.

This paper addresses the challenge posed by observations that are locally aberrant with respect to a specific feature without being global outliers. Since these observations are not outliers in the context of the entire dataset, they are referred to as **Strange Observation Values for a Feature** (STOF). To address them, this paper introduces a new variant of the sparse k-means that includes an improved initialization technique and a method for detecting STOF observations. This approach enables the simultaneous discovery of clusters and STOF observations, while robustifying the feature weight computation process and enhancing the interpretability of clustering results.

The rest of this paper is organized as follows: Section 2 provides a brief review of clustering and sparse k-means variants. Section 3 introduces STOF observations, discusses their impact on sparse k-means clustering, and explains the necessity of proposing a new robust sparse k-means variant. Section 4 presents the novel robust sparse k-means framework and its associated methods. Section 5 presents the data, experiments and the results. The GitHub repository for this work is available at https://github.com/abassodev/rkstof.

2 Related Works

This section is a summary of the k-means, as well as related initialization and outlier detection methods. The sparse k-means is also presented.

2.1 The k-means algorithm

Given a dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{n}$ of n observations $\mathbf{x}_i \in \Re^d$, the k-means aims to divide these n observations into k distinct groups in such a way that the distance between observations in a cluster is minimal, and conversely, that the distance between clusters is maximal [1]. k-means algorithm begins by initializing k points as initial centers using a chosen initialization method. Then, it computes the distance from each point to the k centers and assigns each point to its nearest center. Next, the group centers are updated as the means of their respective groups. These steps are repeated iteratively until the clusters converge. In this algorithm, the number of clusters k is a meta-parameter. Several approaches exist to determine the optimal number (e.g., Gap Statistic, Elbow, ESG [8, 9, 10]).

2.2 Initialization methods

Initialization is the first step in the k-means algorithm. It is of the utmost importance, as the initial points will serve as clustering centers. Poor initialization can lead to poor or even incorrect clustering results. One of the first initialization techniques is the random initialization [1]. It randomly selects k points from the dataset to serve as centers. The main limitation of this method is the fact that it can sometimes produce suboptimal results and be sensitive to outliers. To compensate for this limitation, David et al. [11] propose the k-means++ initialization. It selects the first centroid randomly, then successively chooses the other centroids with a probability proportional to the square of their distance from the already selected centroids. Resulting in faster convergence and better clusters. It is sometimes sensitive to outliers, hence Mohammad et al. [12] propose the ROBIN initialization. It is a robust initialization method that selects initial centers located in a dense area that are far from each other as well. To determine dense areas, they use the Local Outlier Factor (LOF) [5]. Hence, this algorithm assures that none of the initial centers are outliers.

2.3 Outlier detection methods

Several outlier detection methods exist. Some are linked to the k-means, and others none. Among the not linked methods, they are Local Outlier Factor (LOF) [5], Interquartile Range (IQR) [13] and others. Introduced by Breunig et al., LOF is an outlier detection method that identifies abnormal data points by assessing how much they deviate locally from their neighbors. It is based on local density given by k nearest neighbors of an observation. With this method, an observation is an outlier if its density is significantly lower than its neighbor density. The IQR is a metric used to quantify statistical dispersion, representing how spread out the data is. It has been used as an outlier detection tool in numerous areas. It consists of calculating the interquartile range of a dataset (the difference between the third and the first quartile). The algorithm then uses a threshold based on the IQR to identify outliers. In general, a value is considered an outlier if it is less than $Q_1 - 1.5 \times IQR$ or greater than $Q_3 + 1.5 \times IQR$.

Khan et al. [4] propose a robust clustering algorithm to handle outliers and non-spherical data. They use the winsorization method, which consists of replacing extreme values with the nearest value within a specified percentile range. Generally, these methods are used as preprocessing methods. As the main goal of the k-means is to group observations while minimizing intra-cluster inertia, recent outlier detection methods are directly linked to the algorithm.

Hence, Shrifan et al. [14] propose an outlier detection algorithm linked to the k-means, which performs outlier detection at each iteration based on the IQR.

Ahmed et al., propose an outlier detection method linked to the k-means algorithm called ODC (Outlier Detection and Clustering Improvement). At each iteration, it computes the distance between each observation to its centroid, and compared to the mean distance between all the dataset points to their centroids. If the distance between an observation and its centroids is p time greater than the mean dataset distance, it is detected as an outlier. By removing outliers, existing approaches minimize intra-cluster inertia. However, due to the dimensionality of datasets, interpretability issues persist. The sparse k-means clustering has been proposed.

2.4 The sparse k-means algorithm

Proposed by Witten et al. [7], the sparse k-means addresses the dimensionality issues that k-means suffers from. At each step, features are multiplied by a computed weight. The goal is to reduce to zero the features with lower weights. By the end of the algorithm, the less contributing features get a weight of zero. The objective function is given:

$$\begin{aligned} &\max \sum_{j=1}^{d} w_j \left[\sum_{i=1}^{n} (x_{ij} - \boldsymbol{\mu_j})^2 - \sum_{p=1}^{k} \sum_{i \in \mathcal{C}_p} (x_{ij} - z_{pj})^2 \right] \\ &\text{s.t. } \sum_{j=1}^{d} w_j^2 \leq 1, \sum_{j=1}^{d} |w_j| \leq s, \forall j : w_j \geq 0 \end{aligned} \tag{1}$$

where w_j is the weight of the jth feature, s is a meta-parameter, $\boldsymbol{\mu_j}$ is the mean of the jth feature and $\mathbf{z}_p$ is the center of the cluster $\mathcal{C}_p$. Sparse k-means is written following [15] as:

1. initialize centers and feature weights $w_j = 1/\sqrt{d}$;
2. perform k-means, multiplying each feature j by $\sqrt{w_j}$;
3. keeping centers fixed, update w_j as

$$w_j = \frac{\text{sign}(\gamma_j)\ (|\gamma_j| - \Delta)}{\sqrt{\sum_{j=1}^{d} \text{sign}(\gamma_j)\ (|\gamma_j| - \Delta)^2}}, \tag{2}$$

   where

$$\gamma_j = \sum_{i=1}^{n} (x_{ij} - \boldsymbol{\mu_j})^2 - \sum_{p=1}^{k} \sum_{i \in \mathcal{C}_p} (x_{ij} - z_{pj})^2; \tag{3}$$

4. repeat step 2 and 3 until $\sum_{j=1}^{d} \left| \frac{w_j^r - w_j^{r-1}}{w_j^{r-1}} \right| < 10^{-4}$ where w_j^r represents the weights after the current iteration and w_j^{r-1} the weights at the previous iteration,

This algorithm has been very successful. Yet the k-means, it is sensitive to the outlier observations, hence several robust sparse k-means variants have emerged.

2.5 Robust sparse *k*-means variants

Faced with the sensitivity of the sparse k-means to outliers, several studies have been conducted. Brodinová et al. [16] proposed a robust variant of the sparse k-means that aim to enhance the robustness of the sparse k-means against outliers. At each step, the weight of the observations is calculated using the LOF. With an observation weight, compared to a threshold (i.e., 0.5), their approach determines whether an observation is an outlier or not. Outliers are removed and no longer used at the next iteration. Their studies allow the sparse k-means to simultaneous group observations, detect the outliers and informative variables.

Kondo et al. [17] shown the impact of missing values for the sparse k-means. They propose a robust variant of sparse k-means against the missing values based on the trimmed k-means. The existing approaches focus on the robustness of the sparse k-means in terms of outlier detection, missing values, or noisy features. But neither take account of STOF observations nor enhance the calculation of feature weights. It sometimes happens that some features, which should not contribute, do not have a zero contribution. Conversely, some features have a low contribution simply because they have been biased by these observations. The next section presents STOF observations and explains why it is important to take them into account.

3 STOF Observations

A strange observation value for a feature (STOF) is an observation that deviates from others in one or some feature without being an outlier. In several multidimensional datasets, it occurs that some observations deviate from others in one or several particular features. Due to the dimensionality, these deviations are not observable. To illustrate STOFs, we used a robust variant of the k-means algorithm, especially the ODC algorithm. After applying this algorithm, grouped observations and removing outliers, a cluster was selected, and its numerical observations were displayed. Figure 1 shows the cluster observations after applying the ODC on the Glass and Breast Cancer datasets. In this figure, for the Glass dataset, in the feature BA, all the observations, except one, have a zero value. For the Breast Cancer, in feature Concavity, all the observations except two have a value different from zero. The observations are bordered, and the corresponding values that deviate on a particular feature are double-bordered in red. STOFs are not detected as outliers by robust k-means variants, but deviate from the others on some particular features.

In the sparse k-means algorithm, feature weight is computed according to the Equation 2, making it dependent on the feature observations. In this case, by calculating the feature weight of the BA feature, it will be different to zero, but near to zero (i.e., 3.19×10^{-34}) value only because this non-null value is used. Reciprocally, for the Breast cancer dataset, by using the two null values during the computing of the weight of the Concavity feature, it will decrease, only because two null values are used. Hence, STOFs are able to disrupt the calculation of feature contributions. Contrary to outliers, it is not easy to remove them because deleting a value in a given feature will create missing values. It is imperative to find a way to stop their effect on feature weight calculation without deleting them.

To date, to the best of our knowledge, no method in the literature has taken this aspect of outliers into account. Hence, the next section proposes a new variant of the sparse k-means that is robust against this type of observation.

4 Proposed method

The proposed sparse k-means consists of a new initialization and a STOFs detection algorithm designed to work together. This section presents the different algorithms and the final sparse k-means.

4.1 A new initialization algorithm: robust *k*-means++ against STOF observations

The first step in the k-means is to initialize the centers. If these points are poorly chosen, this will have an impact on the algorithm. Since traditional initialization techniques do not take STOF observations into account, it is imperative to develop a new, more robust initialization. k-means++ [11] is renowned for its execution speed, but it lacks robustness. The proposed RKAS

# Si	# K	# Ca	# Ba
72.87	0.7	9.23	0
73.81	0.35	9.42	0
70.57	0.08	11.64	0
71.15	0.08	10.79	0
72.19	0.81	13.24	0
69.81	0.58	13.3	3.15
70.16	0.12	16.19	0
72.67	0.1	11.52	0
74.45	0	10.99	0
73.21	0	14.68	0
73.08	0	14.96	0
72.02	0.06	14.4	0

# area_mean	# smoothne...	# compactn...	# concavity_...
736.9	0.1257	0.1555	0.2032
372.7	0.1006	0.05743	0.02363
349.6	0.08792	0.04302	0
227.2	0.09138	0.04276	0
302.4	0.09699	0.1294	0.1307
832.9	0.09831	0.1556	0.1793
526.4	0.06251	0.01938	0.001595
508.8	0.08739	0.03774	0.009193
2250	0.1094	0.1914	0.2871
1311	0.1141	0.2832	0.2487

Fig. 1 STOF observation for BA feature in Glass dataset (top) and Concavity mean in Breast Cancer dataset (bottom).

(robust k-means ++ against STOF observations) initialization method therefore modifies it. For each point generated by a classical k-means++ algorithm, a verification is performed to ensure that the points obtained are not STOF observations according to all the datasets. This algorithm uses a function called `STOF`, which is detailed in Algorithm 2 in Section 4.2.

Note that Algorithm 1 only performs the initialization of centers the actual clustering is handled by Algorithm 3.

4.2 STOFs detection algorithm

The aim of the k-means algorithm is to minimize intra-cluster inertia. Clustering is better if all the points present in a cluster are very close to their center, i.e., if $\sum_{p=1}^{k} \sum_{\mathbf{x}_i \in \mathcal{C}_p} d(\mathbf{x}_i, \mathbf{z}_p)^2$ for close to zero. So, in this study, the error contribution of a point in a cluster is defined as the ratio of the distance between this point and its center by the sum of the distances of all other points except this point. The error contribution of a point $\mathbf{x}$ in a cluster $\mathcal{C}$ with center $\mathbf{z}$ is:

$$\text{contribution}(\mathbf{x}, \mathcal{C}) = \frac{d(\mathbf{x}, \mathbf{z})}{\sum_{i \in \mathcal{C} | \mathbf{x}_i \neq \mathbf{x}} d(\mathbf{x}_i, \mathbf{z})}. \quad (4)$$

Reciprocally, the contribution to the error of a point in a cluster with respect to a feature is defined as

$$\text{contribution}_j(\mathbf{x}, \mathcal{C}) = \frac{(x_j - z_j)^2}{\sum_{i \in \mathcal{C} | \mathbf{x}_i \neq \mathbf{x}} (x_{ij} - z_j)^2}. \quad (5)$$

Based on the above equations, an efficient algorithm to detect STOF observations, called ESD (efficient STOFs detection algorithm) is proposed.

Generally, outlier detection and robust k-means variants [18, 5, 6, 16], used a meta-parameter p called contamination rate. It represents the estimated proportion of abnormal data points in the dataset. As others, this algorithm used the contamination rate as a meta-parameter. In statistics, a measure called the *percentile* is used to describe the distribution of values in a dataset. Since the contamination rate estimates the proportion of observations that may be considered as STOF, the percentile can be used to calibrate the detection criterion based on the distribution of error contributions. Hence, an observation is considered as STOF in a feature if its error contribution to this feature is greater than the (100-p)th percentile of the contribution of others in the feature. Algorithm 2 presents the proposed STOFs detection method. Although the contamination rate is a meta-parameter, this study provides an empirical analysis to determine a suitable default value. Section 4.4 details this empirical analysis.

4.3 A robust sparse k-means clustering against STOF observations

This part of the paper presents RSKC-STOF (robust sparse k-means clustering against STOF observations) that leverages ESD and RKSA algorithms to mitigate the impact of STOFs during clustering. Having written the previous algorithms, the final step is to write a complete, robust sparse k-means algorithm wrapping them together. The sparse k-means assigns a weight to each feature that represents its respective contribution: the weight w_j of a feature at an iteration is given by Equation (2) where γ_j is defined in Equation (3). As STOF observations should not impact the calculation of feature weights, the proposed approach modifies γ_j values to remove their influence. Indeed, γ_j is a function of the feature mean $\boldsymbol{\mu_j}$, the center feature value z_{pj} and the observation values x_{ij}. One need to compute γ'_j based on the feature mean $\boldsymbol{\mu'_j}$, the center feature value z'_{pj} and the observation values x'_{ij} where STOF observations with respect to feature j have been removed. The expression of the weight of each feature j becomes

$$w_j = \frac{\text{sign}(\gamma'_j)(|\gamma'_j| - \Delta)}{\sqrt{\sum_{j'=1}^{d} \text{sign}(\gamma'_{j'})(|\gamma'_{j'}| - \Delta)^2}}. \quad (6)$$

Notice that STOF observations are not removed from the dataset, they are just not used when calculating feature contributions. Algorithm 3 shows the new sparse k-means.

In algorithm 3, STOF observations are not suppressed, they just do not participate in the feature weighting calculation.

Algorithm 1 robust k-means++ against STOF observations

Require: dataset $\mathcal{D}$, number of clusters k
Ensure: initial set of centers $\mathcal{Z}$

randomly select the first center $\mathbf{z}_1$ from the dataset $\mathcal{D}$
initialize $\mathcal{Z} = \{\mathbf{z}_1\}$ and $i = 2$
while $i < k$ **do**
  compute the distance $d(\mathbf{x})$ for each point $\mathbf{x} \in \mathcal{D}$ to its nearest center in $\mathcal{Z}$
  select $\mathbf{x}_i \in \mathcal{D}$ as new candidate center with probability proportional to $d(\mathbf{x_i})^2$
  if $\neg$STOF$(\mathcal{D}, \mathbf{x}_i)$ **then**
    add $\mathbf{x}_i$ to $\mathcal{Z}$
    i++
  end if
end while
return initial set of centers $\mathcal{Z}$

Algorithm 2 ESD

Require: clusters $\mathcal{C}$, contamination rate p
Ensure: STOF observations

Initialize clustering
repeat
  for each cluster $\mathcal{C}_p \in \mathcal{C}$ **do**
    for each feature j **do**
      for each observation $\mathbf{x}_i$ in cluster $\mathcal{C}_p$ **do**
        $\mathcal{C}_p^{(\mathbf{x}_i)}$ = cluster $\mathcal{C}_p$ deprived of $\mathbf{x}_i$
        contributions$_j$ = error contributions of all points in cluster $\mathcal{C}_p^{(\mathbf{x}_i)}$ for feature j with Equation (5)
        contribution$_j(\mathbf{x}_i)$ = error contribution of $\mathbf{x}_i$ in cluster $\mathcal{C}_p^{(\mathbf{x}_i)}$ for feature j with Equation (5)
        if contribution$_j(\mathbf{x}_i) >$ Percentile$_{100-p}$(contributions$_j$) **then**
          add observation $\mathbf{x}_i$ to the list of identified STOF observations for the jth feature
        else
          observation is not STOF for this feature
        end if
      end for
    end for
  end for
until convergence
return the list of STOF observations for each feature

4.4 Determination of the default contamination rate

To determine the optimal meta-parameters, various studies in clustering generally proceed through empirical studies [19, 20, 10, 5, 7, 21]. To determine the default contamination rate, this study aligns with others and provides an empirical evaluation of the proposed ESD on two real datasets (Iris and Lymphography). A commonly used criteria to assess the quality of clustering methods, is the intra-cluster inertia [19, 5]. The intra-cluster inertia is generally considered an indicator of the quality of clustering: the lower it is, the more compact the clusters are [6, 19].

Because STOF observations deviate from others and the centroid, they increase intra-cluster inertia. Thus, by detecting and addressing them, the intra-cluster inertia will progressively decrease based on the relevance of the detected observations.

Algorithm 3 Sparse k-means Clustering Algorithm

Require: Dataset $\mathcal{D}$, number of clusters k

Ensure: partition of $\mathcal{D}$ into k clusters and weights w_j for each feature j of $\mathcal{D}$.

1: initialize k initial centers using algorithm 1
2: initialize $w_1 = ... = w_p = \frac{1}{\sqrt{d}}$.
3: **repeat**
4: performing k-means on the scaled data, i.e., by multiplying each feature j by $\sqrt{w_j}$.
5: detect STOF using ESD in each cluster and save them
6: keeping centers fixed, apply Equation (6) to calculate the weighting of each feature deprived of STOF observations
7: **until** $\sum_{j=1}^{d} \left| \frac{w_j^r - w_j^{r-1}}{w_j^{r-1}} \right| < 10^{-4}$
8: **return** the cluster with the weight of each feature

Hence, for this empirical evaluation, the ESD algorithm was applied in different datasets, detected STOFs at varying contamination rates (1% to 10%), and excluded them when computing intra-cluster inertia. The aim is to assess the variation in intra-cluster inertia and the number of STOFs detected across this range of contamination rates.

Figure 2 and 3 present the variation of the number of STOFs detected, the intra-cluster inertia, on the Iris and Lymphography datasets.

Based on these figures, it is empirically observable that, using a contamination rate of 0%, the intra-cluster inertia is generally high at this point. When the contamination rate increases to 1%, the number of observations identified as STOFs slightly increases, and the intra-cluster inertia begins to decrease. At 2%, the number of detected STOFs rises more noticeably, and there is a significant drop in intra-cluster inertia. Beyond this 2%, although the number of STOFs continues to grow, the decrease in intra-cluster inertia becomes much less pronounced. This suggests that most of the problematic observations are already identified by this point, representing around 2% of the dataset, which remains reasonable for these datasets.

Therefore, in this study, a default contamination threshold of 2% is adopted. However, this threshold remains flexible and can be adjusted depending on the characteristics of the dataset or the specific goals of the analysis, as other outlier detection methods.

5 Experiments

This section evaluates the proposed method on several datasets. The proposed algorithms were implemented in Python, and all experiments were conducted on a 64-bit Windows 11 platform, equipped with a Core i7 processor and 32 GB of RAM. Three evaluations are conducted for the initialization methods, the STOFs detection algorithm and the sparse k-means framework respectively.

5.1 Datasets

For this study, real and synthetic datasets were used. The real datasets were the follow:

- Iris dataset: is a small dataset, containing 150 samples, each with four features, representing the characteristics of three flower species: Setosa, Versicolor, and Virginica.
- Lymphography dataset: consists of 148 samples with 18 categorical features, used to classify lymphatic system conditions into four categories: normal, metastases, malign lymph, and fibrosis.
- Wine dataset: it contains the results of chemical analyses of wines from three distinct regions in Italia, comprising 178 samples with 13 features.
- Diabetes dataset: it consists of 442 instances with 10 features, reported in two classes. It captures medical and demographic information from patients with their diabetes status.
- Breast Cancer dataset: consists of 569 samples with 30 features, related to characteristics of breast cancer cases (corrected from the original, which incorrectly mentioned diabetes).

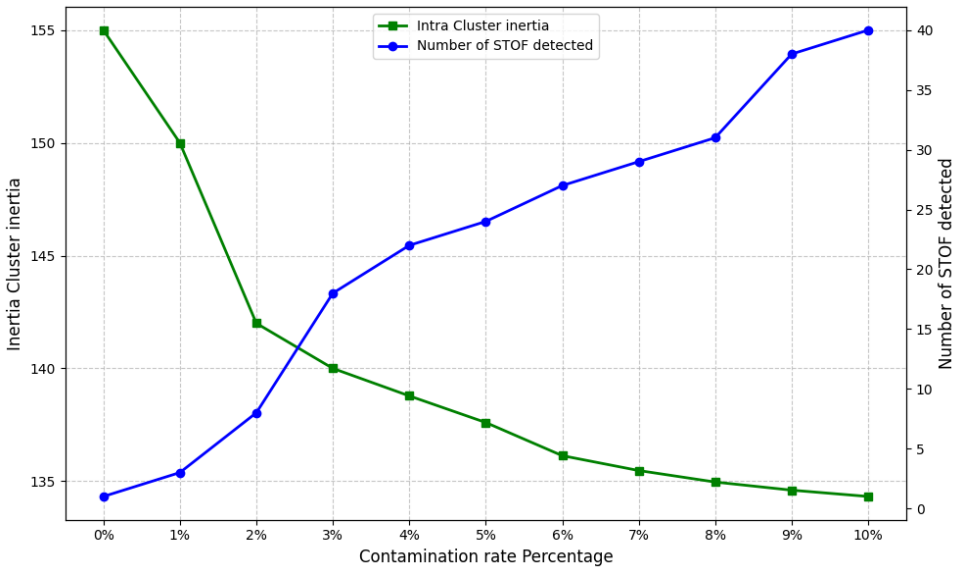

Fig. 2 Variation of intra-cluster inertia, number of STOFs detected, with respect to the contamination rate on Iris dataset.

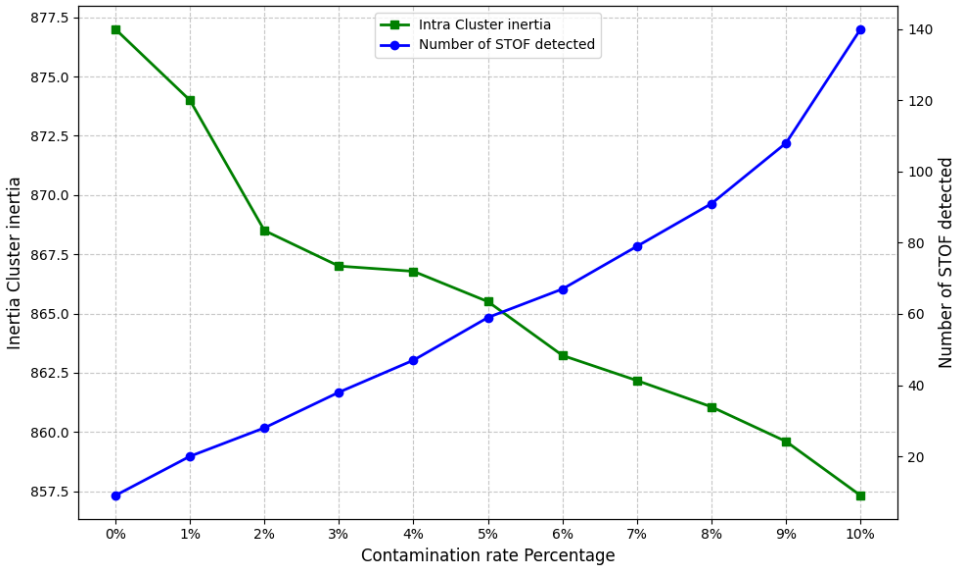

Fig. 3 Variation of intra-cluster inertia and number of STOFs detected with respect to the contamination rate on the Lymphography dataset.

For the synthetic datasets, six reference datasets were generated. The first three (X_1, X_2, X_3) with different sizes served as reference datasets for evaluating the initialization (RKAS) and STOFs detection (ESD) methods, while the other three (Y_1, Y_2, Y_3) were used to evaluate the final sparse k-means algorithm. Table 1 presents the information about these datasets.

To robustly evaluate the ESD and RKAS algorithms, Monte Carlo simulations were employed, 50 variants for each reference dataset (X_1, X_2, X_3,Y_1, Y_2, Y_3) were generated respecting their specific characteristics.

The synthetic datasets were generated such that each observation naturally belongs to a specific cluster [22, 23]. This was achieved by sampling data points from multivariate normal

Table 1 Description of X_i and Y_i datasets

Dataset	Features	Observations	Zero Features
X_1	4	500	0
X_2	40	2000	0
X_3	200	30,000	0
Y_1	50	600	2
Y_2	100	10,000	7
Y_3	200	9,000	12

distributions $\mathcal{N}(\boldsymbol{\mu}, \sigma^2)$, where $\boldsymbol{\mu}$ is the mean and σ is the standard deviation. After generating the datasets, STOFs and outliers were injected following the framework proposed by Pei et al. [22].

Pei et al. [22], define multiple levels of difficulty for introducing outliers. The first level corresponds to a scenario where no outliers are explicitly added, and anomalies result solely from the randomness of the data generation process. The second level involves introducing a small percentage of outliers, randomly placed within the regions occupied by clusters. In this study, both types of outliers were considered, with STOFs and outliers injected at different proportions. For each of the 50 variants, parameters were randomly sampled as follows:

- Number of features and observations: sampled according to the size in Table 1.
- Proportion of STOFs and outliers: sampled between 0% and 15%.
- Standard deviation (σ): sampled between 1 and 20.

5.2 Initialization technique evaluation

The primary objective of the proposed RKAS initialization technique is to ensure that the initial centroids do not include STOF observations. In the context of robust initializations, it is crucial that the initial centers do not exhibit outliers. To evaluate the effectiveness of RKSA, it is essential to consider scenarios where STOFs and outliers are present. The generated variants of X_1, X_2, X_3 and two real datasets were used.

The proposed initialization algorithm and the ROBIN algorithm were executed consecutively on the 50 variants of the generated datasets. For each execution, the number of outliers and STOFs selected as initial centroids was recorded. Table 2 presents the accuracy, with 95% of confidence, of the proportion of STOFs ($P(\neg STOF)$) and outliers ($P(\neg O)$) that are not selected as initial centers by the algorithms in the different datasets.

From this table, it can be observed that RKAS is more robust than ROBIN initialization in avoiding STOF observations in initial centers. It is also robust against outliers. In contrast, the ROBIN algorithm often selects STOF observations as centroids, but generally robust against outliers, highlighting the superiority of RKAS in terms of STOFs.

Table 2 Success rate of RKSA and ROBIN in avoiding outliers and STOF.

Dataset	**Method**	$P(\neg O)$	$P(\neg STOF)$
X_1	RKSA	100%	91.50±3.50%
	ROBIN	100%	81.46±8.54%
X_2	RKSA	100%	78.75±8.91%
	ROBIN	100%	67.25±9.92%
X_3	RKSA	100%	83.75±7.25%
	ROBIN	100%	66.18±10.35%
Breast C	RKSA	100%	100%
	ROBIN	100%	100%
Wine	RKSA	100%	100%
	ROBIN	100%	100%

5.3 STOFs detection evaluation

The robustness of the proposed ESD algorithm is evaluated based on its performance on STOFs detection. Hence, this algorithm and three other outlier detection algorithms were tested on previously established datasets. The datasets X_1, X_2 and X_3 were used. For each dataset, the percentage of STOF observations correctly identified by these algorithms was reported. Table 3 presents the accuracy on the different datasets at 95% of confidence. ESD algorithm performs best because it detects STOFs on features independently, in contrast to the classic outlier detection algorithms.

5.4 Evaluation of the Final algorithm

The effectiveness of the proposed robust sparse k-means algorithm is evaluated based on four key

Table 3 Accuracy of several algorithms in STOFs detection at 95% of confidence.

Dataset	Method	Accuracy
X_1	ESD	71.12±3.47%
	LOF	33.29±3.96%
	ODC	44.12±4.96%
	IQR	39.43±2.61%
X_2	ESD	68.26±5.78%
	LOF	45.69±3.69%
	ODC	40.16±5.37%
	IQR	37.24±2.78%
X_3	ESD	81.16±5.39%
	LOF	50.11±5.22%
	ODC	57.94±2.45%
	IQR	47.07±2.03%

aspects: clustering quality, robustness in feature selection in the presence of STOF observations, and interpretability.

To evaluate robustness in feature selection, the proposed sparse k-means and other variants were applied to the 50 generated variants of the datasets Y_1, Y_2, Y_3, and the pertinence of feature weights was assessed at each step. Originally, these datasets contain some features with all values set to zero as describe in the Table 1. These features are not supposed to play any role in the algorithm. However, a very small proportion of STOFs and outliers was intentionally added to them. The effectiveness of each method in detecting zero features across various datasets is evaluated. Table 4 shows the accuracy of different approaches in identifying zero features at 95% confidence across 50 generated datasets. The proposed sparse k-means demonstrates greater robustness in feature selection compared to other approaches.

For clustering quality evaluation, the proposed robust sparse k-means variant and other sparse k-means variants were executed on several real datasets (Iris, Wine, Breast Cancer, Diabetes). Each time, the ratio between the Within-Cluster Sum of Squares (WCSS) and the Between-Cluster Sum of Squares (BCSS) [19] was computed. A lower value of this ratio generally indicates better clustering quality. This criterion is particularly emphasized in sparse k-means, as feature values are multiplied by their respective weights at each iteration. Table 5 presents the $WCSS/BCSS$ ratio achieved by each algorithm, along with the number of removed observations. It highlights that, although RSKC-STOF does not remove any observations, it consistently produces high-quality clusters.

For interpretability, the proposed sparse k-means and others were executed in several real datasets, and at the end of the algorithms, the feature weight was computed. Tables 6, 7 and 8, illustrate the feature weights obtained by RSKC-STOF and that of Brodinová et al. on various datasets. The results show that RSKC-STOF identifies a higher number of zero features compared to Brodinová et al. By detecting STOFs, RSKC-STOF reduces extremely low feature contributions to zero, offering a dual advantage: enhanced interpretability of results and improved axis selection during visualization.

Table 4 Percentage of correct Zero Feature detections by different algorithms on several datasets.

Dataset	Method	P (Zero Features)
Y_1	Sparse k-means [7]	18.53±2.09%
	Brodinová et al. [16]	53.15±4.85%
	RSKC-STOF	89.81±3.19%
Y_2	Sparse k-means [7]	31.15±3.95%
	Brodinová et al. [16]	69.83±2.17%
	RSKC-STOF	87.57±3.43%
Y_3	Sparse k-means [7]	12.45±3.55%
	Brodinová et al. [16]	49.48±4.06%
	RSKC-STOF	65.73±4.27%

6 Conclusion and future works

Interpretability and feature selection are important tasks for clustering. This paper has introduced a variant of the sparse k-means algorithm aiming to simultaneously detect clusters, outliers and STOFs, while strengthening the feature selection. The proposed RSKC-STOF approach aims to make the feature selection more robust, while proposing a new initialization and new methods for detecting outliers and STOF observations. RSKC-STOF is geared towards improving the performance of the popular sparse k-means algorithm for real-world data, which may potentially contain both outliers and STOF observations. Although this approach improves the robustness and interpretability of the sparse k-means algorithm, it has certain limitations. It is primarily designed for

Table 5 $WCSS/BCSS$ ratio and number of deleted observations on several datasets.

Dataset	Method	Deleted Observation	$WCSS/BCSS$
Iris	Sparse k-means [7]	0	0.082
	Brodinová et al. [16]	6	0.068
	RSKC-STOF	0	0.042
Breast Cancer	Sparse k-means [7]	0	0.41
	Brodinová et al. [16]	27	0.38
	RSKC-STOF	0	0.35
Wine	Sparse k-means [7]	0	0.19
	Brodinová et al. [16]	7	0.17
	RSKC-STOF	0	0.13
Diabetes	Sparse k-means [7]	0	0.23
	Brodinová et al. [16]	6	0.22
	RSKC-STOF	0	0.18

Table 6 Comparison of feature weights on the Iris dataset.

	Final Feature Weights			
Brodinová et al.	4.51×10^{-7}	3.7×10^{-15}	1.0	1×10^{-6}
RSKC-STOF	4.91×10^{-7}	0.0	1.0	1×10^{-6}

STOF observations and may not be effective in handling missing values, which can also introduce bias in variable selection evaluation. Furthermore, this method relies on a contamination threshold determined empirically. This threshold can be adjusted by the user based on their needs and analysis objectives. However, this poses a significant limitation when the user lacks in-depth knowledge of their data, and the default threshold (2%) is not optimal for their specific application. A potential direction for future research would be to develop adaptive techniques based on statistical concepts, allowing for an analytical and automatic determination of the contamination threshold. Also, we plan to enhance the proposed RSKC-STOF to handle missing values.

References

[1] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[2] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[3] Iliyas Karim Khan, Hanita Daud, Nooraini Zainuddin, Rajalingam Sokkalingam, Abdussamad, Abdus Samad Azad, Mudassar Iqbal, Mudasar Zafar, Atta Ullah, Musarat Elahi, and Ahmad Abubakar Suleiman. Exploring k-means clustering efficiency: Accuracy and computational time across multiple datasets. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 62(3):57–69, 2024.

[4] Iliyas Karim Khan, Hanita Binti Daud, Nooraini Binti Zainuddin, Rajalingam Sokkalingam, Abdussamad, Abdul Museeb, and Agha Inayat. Addressing limitations of the k-means clustering algorithm: Outliers, non-spherical data, and optimal cluster selection. *AIMS Math*, 9:25070–25097, 2024.

[5] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[6] Mohiuddin Ahmed and Abdun Naser Mahmood. A novel approach for outlier detection and clustering improvement. In *2013 IEEE 8th Conference on Industrial Electronics and Applications (iciea)*, pages 577–582. IEEE, 2013.

[7] Daniela M Witten and Robert Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010.

[8] Iliyas Karim Khan, Hanita Daud, Nooraini Zainuddin, and Rajalingam Sokkalingam. Standardizing reference data in gap statistic for selection optimal number of cluster in k-means algorithm. *Alexandria Engineering Journal*, 118:246–260, 2025.

[9] Iliyas Karim Khan, Hanita Binti Daud, Nooraini Binti Zainuddin, Rajalingam Sokkalingam, Muhammad Farooq, Muzammil Elahi Baig, Gohar Ayub, and

Table 7 Comparison of Weights on the Wine Dataset.

Final Feature Weights - Brodinová et al.					
5.36×10^{-33}	3.06×10^{-24}	2.76×10^{-31}	4.19×10^{-19}	4.69×10^{-15}	7.53×10^{-24}
4.44×10^{-22}	7.08×10^{-31}	1.07×10^{-25}	2.89×10^{-24}	2.80×10^{-29}	2.55×10^{-24}
1.0					
Final Feature Weights with RSKC-STOF					
0.0	0.0	0.0	3.88×10^{-19}	4.69×10^{-15}	5.92×10^{-25}
5.84×10^{-23}	0.0	0.0	6.44×10^{-25}	0.0	0.0
1.0					

Table 8 Comparison of Weights on the Breast Cancer Dataset.

Final feature weights - Brodinová et al.					
2.68×10^{-38}	9.66×10^{-21}	1.38×10^{-11}	2.16×10^{-2}	2.98×10^{-42}	0.0
1.89×10^{-32}	2.74×10^{-34}	2.30×10^{-39}	3.64×10^{-28}	4.18×10^{-41}	2.21×10^{-21}
1.14×10^{-10}	1.02×10^{-49}	2.38×10^{-41}	9.46×10^{-40}	4.04×10^{-43}	5.68×10^{-58}
1.95×10^{-54}	4.18×10^{-17}	1.36×10^{-19}	2.07×10^{-10}	9.99×10^{-1}	2.49×10^{-40}
1.50×10^{-31}	9.54×10^{-30}	9.54×10^{-33}	4.03×10^{-36}		
Final feature weights with RSKC-STOF					
0.0	0.0	8.29×10^{-14}	3.55×10^{-4}	0.0	0.0
0.0	0.0	0.0	0.0	5.06×10^{-19}	0.0
1.56×10^{-12}	8.71×10^{-20}	0.0	0.0	0.0	0.0
0.0	0.0	5.60×10^{-4}	3.68×10^{-4}	2.01×10^{-15}	9.99×10^{-1}
0.0	6.73×10^{-8}	0.0	0.0		

Mudasar Zafar. Determining the optimal number of clusters by enhanced gap statistic in k-mean algorithm. *Egyptian Informatics Journal*, 27:100504, 2024.

[10] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the royal statistical society: series b (statistical methodology)*, 63(2):411–423, 2001.

[11] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

[12] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed J Zaki. Robust partitional clustering by outlier and density insensitive seeding. *Pattern Recognition Letters*, 30(11):994–1002, 2009.

[13] Frederik Michel Dekking, Cornelis Kraaikamp, Hendrik Paul Lopuhaä, and Ludolf Erwin Meester. *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer Science & Business Media, 2006.

[14] Nawaf HMM Shrifan, Muhammad F Akbar, and Nor Ashidi Mat Isa. An adaptive outlier removal aided k-means clustering algorithm. *Journal of King Saud University-Computer and Information Sciences*, 34(8):6365–6376, 2022.

[15] Avgoustinos Vouros and Eleni Vasilaki. A semi-supervised sparse k-means algorithm. *Pattern Recognition Letters*, 142:65–71, 2021.

[16] Šárka Brodinová, Peter Filzmoser, Thomas Ortner, Christian Breiteneder, and Maia Rohm. Robust and sparse k-means clustering for high-dimensional data. *Advances in Data Analysis and Classification*, 13:905–932, 2019.

[17] Y Kondo, M Salibian-Barrera, and R Zamar. A robust and sparse k-means clustering algorithm; 2012. *Available: arXiv preprint arXiv*, 12016082.

[18] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

[19] Ariel Basso Madjoukeng, Edith Belise Kenmogne, and Benoît Frenay. Fast k-means with stable instance sets. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2024.

[20] Robert L Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

[21] Braulio V Sánchez Vinces, Erich Schubert, Arthur Zimek, and Robson LF Cordeiro. A comparative evaluation of clustering-based outlier detection. *Data Mining and Knowledge Discovery*, 39(2):13, 2025.

[22] Yaling Pei and Osmar Zaïane. A synthetic data generator for clustering and outlier analysis. 2006.

[23] Lucija Petricioli, Luka Humski, and Mihaela Vranić. Preserving clusters in synthetic data sets based on correlations and distributions. *Electronics*, 14(11):2230, 2025.