Graphical Abstract

Finding the k in k-means using high density regions

Hugo J. Bello, Domingo Gómez Pérez, Steven Van Vaerenbergh

Highlights

Finding the *k* in *k*-means using high density regions

Hugo J. Bello, Domingo Gómez Pérez, Steven Van Vaerenbergh

- We propose HDR X-Means, a clustering algorithm based on high-density regions derived from estimated probability densities.
- HDR X-Means generalizes classical X-Means and estimate the number of clusters in non-Euclidean domains including circles, cylinders, and spheres.
- The performance of HDR X-Means is evaluated on synthetic and real datasets, demonstrating its flexibility and accuracy, compared with X-Means, DBSCAN, and other clustering techniques.
- We provide theoretical guarantees on clustering errors and demonstrate empirical improvements over existing techniques.

Finding the k in k-means using high density regions

Hugo J. Bello, Domingo Gómez Pérez, Steven Van Vaerenbergh

Universidad de Valladolid, Calle Universidad, s/n, Soria, 42004, Castilla y León, Spain

Universidad de Cantabria, Avd Los Castros s/n, Santander, 39400, Cantabria, Spain

Abstract

Clustering is a fundamental task in unsupervised learning, one that is particularly challenging when the number of clusters k is unknown a priori. Current solutions to partition data are unsatisfactory, due to their lack of consistency and robustness across different data geometries. One of the main issues in our knowledge of clustering is a lack of a way to estimate the number of clusters when similarity measures are based on non-Euclidean metrics. This is a significant limitation because the need to cluster data when the data sources come from specific domains (e.g., circular, spherical). This is overlooked in the literature, despite being a common requirement in real-world applications, such as dates, directions, and locations.

This paper presents a new approach emboided in two algorithms, **HDR X-means** and **HDR++ Merge**, designed to both estimate the number of clusters and perform the corresponding clustering task, by using high-density regions (HDRs). This approach is based on the idea that clusters are regions of high density separated by regions of low density, which makes it robust against outliers.

Keywords: Clustering, Density peaks, X-Means

1. Introduction

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). The most popular clustering algorithm is k-means, which partitions n observations into k clusters in which similarity is measured in terms of Euclidean distance to the cluster center of mass. Improvements to k-means have focused on better initialization methods [1]. However, one key limitation of k-means is that it requires the number of clusters k to be specified a priori, which is often unknown in practice. To address this, Pelleg and Moore [2] proposed X-Means which extends k-means by using the Bayesian Information Criterion (BIC), which balances model fit and complexity, to automatically determine the optimal number of clusters. However, X-Means assumes that clusters are well-modeled by Gaussian distributions, which implies a Euclidean geometry. This assumption may not hold in many real-world scenarios, where data presents structures that are non-Euclidean, such as circular, spherical, or cylindrical geometries. For instance, clustering data that data zones on a sphere (e.g., geographical locations) [3] or circular data (e.g., time of day, angles) [4] requires methods that can handle these specific geometries.

Even if the data is Euclidean, estimating the number of clusters is not considered a solved problem. There are two types of strategies for estimating the number of clusters: Botton-up and top-down. Botton-up strategies start with a large number of clusters and iteratively merge them based on some criteria, while top-down strategies start with a small number of clusters and iteratively split them. X-Means is a top-down approach, as it starts with a small number of clusters and splits them based on the BIC score.

We remark that in practice, it is often set by the user based on domain knowledge or trial and error, so it is important to collect also information about the process to support the user in this task [5]. This additional information has proven very useful, starting the area of the so called **hierachical clustering algorithms**, i.e., a tree of clusters (dendrogram) where each node is a cluster containing its children. This is useful in many ways, allowing to visualize the data in a more intuitive way, and allowing to choose the number of clusters based on a desired level of granularity. Botton-up strategies define **agglomerative clustering algorithms** (AGNES), while top-down strategies define **divisive clustering algorithms** (DIANA). Experiments on the performance of both strategies were performed by [6], who concluded that there is no clear winner between both strategies, but the resulting dendrograms can be quite different.

Regarding the non-Euclidean geometry of the clusters, the most common approach is to study the local density of the data, as in DBSCAN [7], where the data is partitioned by a density threshold, points in low-density regions are discards as noise, and assigns to different clusters disconnected regions of high density. Unfortunately, DBSCAN requires the user to set two parameters: the neighborhood radius ϵ and the minimum number of points $MinPts$ required to form a dense region. Moreover, the computational complexity of DBSCAN is impractical for large datasets, as it is quadratic in the number of points in the worst case. In order to improve the performance of DBSCAN, several variants have been proposed based on the idea of density peaks [8]. These methods identify cluster centers as points that are both densely surrounded and relatively distant from other high-density points. However, these methods still require the user to set parameters and they require a quite large number of points to estimate the density and estimation of density fails in sparse regions [9].

Our new approach, embodied in two algorithms HDR X-Means and HDR++ Merge combines all the advantages of previously mentioned strategies. First, the algorithms estimate the number of clusters and perform clustering in both Euclidean and non-Euclidean domains. Both algorithms rely on high-density regions (HDRs) derived from Estimated Probability Densities. We propose two algorithms, where both start from an initial set of clusters and then refine them based on HDRs:

- **HDR X-Means** (*split–validate*): Clusters are validated by checking non-overlap of estimated HDRs; the selected k is the smallest value for which all HDRs are disjoint.

- **HDR++ Merge** (*seed–merge*): The representatives of each cluster are chosen at random with a suitable distribution and this define a clustering in a k-means fashing. Then, the algorithm *merges* any clusters whose HDRs intersect, repeating until no overlaps remain.

This approach is more amenable to different geometries, as it does not rely on Gaussian assumptions. According to the taxonomy [10, 5, 11], HDR X-Means is an agglomerative clustering algorithm for numerical data, with hard membership and non-overlapping clusters, using a proximity measure based on density estimation. Quite importantly, both algorithms have assured theoretical guarantees on clustering errors, showing that false merging and splitting probabilities decrease with sample size and cluster separation.

Computational experiments shows this approach identifies clusters even in the present of noise, generalizing classical X-Means to settings where geometric separation aligns with regions of high density rather than centroids. The soundness is validated through extensive experiments on synthetic and real datasets [12] demonstrating the method's flexibility and improved accuracy over existing clustering techniques in complex domains. Among their advantages, HDR X-Means and HDR++ Merge do not require large datasets to estimate density. Also, time complexity is more efficient than DBSCAN, as it does not require computing the neighborhood for each point.

As a summary, the main contributions of this paper are:

- We propose a novel approach that estimates the number of clusters using high-density regions derived from estimated probability densities, called **HDR X-Means**. This naturally extends to a clustering algorithm called **HDR++ Merge**.
- We provide theoretical guarantees on clustering errors, showing that false merging and splitting probabilities decrease with sample size and cluster separation.
- The empirical performance of both algorithms is demonstrated on the in-depth benchmark proposed by [12], exhibiting fast execution times, flexibility, and improved accuracy compared to several existing techniques.

.

2. Related Work

There are several surveys on clustering algorithms [10, 5, 11], as mentioned before, so the reader is referred to more details about the specific algorithms with which will be compared to our new proposal.

The foundational work on estimating the number of clusters was done by Pelleg and Moore [2], who proposed the X-Means algorithm. Then, several improvements have been proposed to X-Means, for example, G-means [13] improves the splitting criterion by using the Anderson-Darling test to check if the data in a cluster follows a Gaussian distribution. As mentioned in the introduction, both X-Means and G-means and Gaussian Mixture Models (GMMs) assume that clusters are well-modeled by Gaussian distributions, which implies a Euclidean geometry.

There are several works that extend k-means clustering algorithm to circular data, where clusters are sampling different von Mises distributions [14, 15]. Unfortunately, they still require the user to set the number of clusters k a priori, they are sensitive to initialization, and they are not adapted for more complex geometries, even cylindrical data has not been considered in the literature.

In parallel, several methods have been proposed that rely on density estimation to identify clusters. It started with DBSCAN (Density-Based Spatial Clustering of Applications with Noise) proposed by Ester et al. [7], following with OPTICS (Ordering Points To Identify the Clustering Structure) proposed by Ankerst et al. [16]. Rodriguez and Laio [8] made a breakthrough with a novel clustering method called DensityPeak (DPC) based on the intuition that clusters are regions of high density separated by regions of low density. This method improved over DBSCAN and OPTICS by eliminating the need to set a density threshold and its easy implementation. However, DPC still requires the user to set a cutoff distance parameter, which can result in dividing clusters if several density peaks exist within a single cluster. Another limitations of DPC are its reliance on a large number of points to accurately estimate density, the time consumption in computing distances between all pairs of points, and the lack of tolerance to noise [17].

Even with these limitations, DPC has inspired several variants that aim to address these issues. Han et al. [9] proposed an improved version of DPC called Adaptive Threshold Selection DPC (ATSDPC). This method introduces an adaptive threshold selection mechanism to automatically determine the cutoff distance parameter. Information Entropy Peaks Clustering (IEPC) [18] enhances DPC by incorporating information entropy to better identify cluster centers using dynamic reverse nearest neighbor sequences. Their benchmarks on ten real-world data sets from UCI repository demonstrate improved clustering quality compared to DBSCAN and DPC. This paper adopts a principled alternative based on *highest density regions* (HDRs) [19]. We view clusters as connected components of regions where an estimated, geometry-aware density exceeds a threshold chosen to contain a prescribed probability mass. Building on this idea, we develop two families:

- **HDR X-Means** (*splitvalidate*). Starting from $k_{\max}$, local refinements in the spirit of X-Means [20] are validated by checking non-overlap of estimated HDRs; the selected k is the smallest value for which all HDRs are disjoint.

- **HDR++ Merge** (*seedmerge*). Starting from $k_{\max}$ centres chosen by k-means++, we perform a single Voronoi assignment and then *merge* any clusters whose HDRs intersect, repeating until no overlaps remain.

This produces a monotone decrease in k and avoids repeated local reoptimisation.

3. Preliminaries

Throughout this paper, we will focus on clustering data with certain non-Eucidean geometries, such as circular, spherical, and cylindrical. In these settings, distances wrap with certain known periodicity, clusters form spherical caps, and cylindrical structure mixes periodic with linear behavior.

Thanks to the extensive literature on directional statistics due to the popularity of clustering and easiness to collect data like angles on the circle for directional statistics [3], phases in circular neuroscience [4], or joint angularlinear measurements on the cylinder in spatio-temporal modeling. In these settings, distances wrap (on S^1), clusters form spherical caps (on S^2), and cylindrical structure mixes periodic with linear behaviour, so Euclidean heuristics can mislead.

3.1. Notation

Vectors are bold lowercase ($\mathbf{x}$); $\|\cdot\|$ denotes a norm. Mainly we will use the Euclidean norm, but other norms are considered when appropriate for the geometry at hand. For a data set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$ we write $\widehat{\mu}_X = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ for its empirical mean. All samples are assumed i.i.d. unless explicitly stated.

In the next subsections we review the mathematical foundations of the clustering algorithms used as baselines in our experiments.

3.2. Centroidbased clustering

Classical k-means minimises withincluster squared error

$$\widehat{\sigma}^2 = \frac{1}{n - dk} \sum_{i=1}^{n} \|\mathbf{x}_i - \widehat{\mu}_{c(i)}\|^2, \tag{1}$$

where minimization is over cluster assignments. k-means is algorithmically simple, alternating between assigning points to the nearest cluster centre and updating centres to the empirical means of assigned points. Random initialisation can lead to poor local minima. The strategies like *k-means++* [21], which chooses the first centre uniformly at random and each subsequent centre with probability proportional to the squared distance from the nearest existing centre, provides an $O(\log k)$ approximation with probabilistic guarantee.

XMeans.. XMeans [20] starts with a usersupplied upper bound $k_{\max}$, repeatedly splits clusters by local k-means runs and accepts a split if it improves the Bayesian Information Criterion (BIC).

The BIC score for a clustering into k clusters with means $\{\mu_1, \ldots, \mu_k\}$ and common variance σ^2 is

$$\mathrm{BIC}(X, k) = L(X; k, \mu_1, \ldots, \mu_k, \widehat{\sigma}^2) - \frac{d \cdot (k+1)}{2} \log n$$

where $L(X; k, \mu_1, \ldots, \mu_k, \widehat{\sigma}^2)$ is the loglikelihood and $\widehat{\sigma}^2$ is defined in equation (1).

3.3. Density-based clustering

There is a rich literature on density-based clustering; so we can only review the main ideas of the methods used as baselines.

DBSCAN [22] defines clusters as maximal ε-connected sets of *core points*points that have at least `minPts` neighbours within radius ε.

Let $\mathcal{N}_\varepsilon(\mathbf{x}) = \{\mathbf{y} \in X : \|\mathbf{x} - \mathbf{y}\| \leq \varepsilon\}$ be the ε-neighbourhood of $\mathbf{x} \in X$. A point $\mathbf{x} \in X$ is a *core point* if $|\mathcal{N}_\varepsilon(\mathbf{x})| \geq \texttt{minPts}$. Two points $\mathbf{x}, \mathbf{y} \in X$ are *ε-connected* if there is a sequence of core points $\mathbf{x} = \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m = \mathbf{y}$ such that $\|\mathbf{z}_i - \mathbf{z}_{i+1}\| \leq \varepsilon$ for all $i = 1, \ldots, m-1$. A *cluster* is a maximal set of ε-connected points. We remark that, although the ideas of DBSCAN extend to arbitrary shapes, choosing a single ε can be difficult when clusters have different densities.

OPTICS [16] orders points by reachability distance, producing a reachability plot from which clusters appear as valleys. This removes the need to prespecify a single global ε.

HDBSCAN [23] builds a hierarchy of densityconnected components using mutual reachability distances. A flat clustering is extracted by selecting the most persistent components. We use the efficient implementation of [24].

3.4. High Density regions

A high-density region (HDR) of a sample represents the part of the sample space where the density of data points is highest, often used to identify areas where most of the data is concentrated.

Definition 1. *Let $f(x)$ be the density function of a random variable X on $\mathbb{R}^d$ and $R(\alpha)$ be a region depending on a parameter α,* *then the* $100(1-\alpha)$ ***high density region*** *is the subset* $\mathcal{R}_f(1-\alpha)$ *of the sample* X *such that*

$$\mathcal{R}_f(1-\alpha) = \{x : f(x) \geq f_\alpha\}$$

where f_α *is the largest constant such that* $\mathbb{P}(X \in R(f_\alpha)) \geq 1-\alpha$*.*

The idea of $R(\alpha)$ is to capture the geometric region where the cluster are located, assuming that clusters are regions of high density. Although the natural choice is to select the regions by the level sets of the density function, other choices are possible. This choice is motivated by the fact that the level sets of the density function could be difficult to calculate in practice, especially when the density function is unknown and needs to be estimated from data.

Remark 1. *The regions* $\mathcal{R}_f(1-\alpha)$ *can be estimated under some knowledge of the distribution* f *and a sample* $\mathbf{x}_1, \ldots \mathbf{x}_n$ *that drawn from a random variable* X*. Under the asumption that* f *belongs to the parametric family of distribution* $\{f_\theta : \theta \in \Theta\}$ *and* $R_\theta(\alpha)$*, estimate* θ *using the maximum likelihood estimator*

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(X; k, \theta),$$

where

$$L(X; k, \theta) = \prod_{i=1}^{n} f_\theta(x_i).$$

With the estimated parameter $\hat{\theta}$*, we can* $\hat{\mathcal{R}}_{f_{\hat{\theta}}}(1-\alpha)$ *following definition 1.*

Remark 2. *Given* X *a random variable following a d dimensional multinormal* $\mathcal{N}(\mu, \Sigma)$ *then its* $100(1-\alpha)$ *HDR region is the ellipsoid*

$$\mathcal{R}_f(1-\alpha) = \{(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu) \leq \chi^2_{d,1-\alpha}\}$$

where $\chi_{d,1-\alpha}$ *is the value of the* χ^2 *distribution with d degrees of freedom such that* $P(\chi^2_d \leq \chi_{d,1-\alpha}) = 1-\alpha$

Definition 2. *Let* $x \in \mathbb{R}^d$ *be a point, and let* $\mu \in \mathbb{R}^d$ *be a reference mean vector with associated symmetric positive definite covariance matrix* $\Sigma \in \mathbb{R}^{d \times d}$*. The* ***Mahalanobis distance*** *between* x *and* μ *is defined as:*

$$D_\Sigma(x, \mu) := \sqrt{(x-\mu)^\top \Sigma^{-1} (x-\mu)}.$$

This quantity measures how far x is from μ in terms of the scale and orientation encoded by the covariance matrix Σ. Unlike the Euclidean distance, it accounts for the shape of the distribution: directions with high variance contribute less to the distance, and correlations among components are incorporated via Σ^{-1}.

Equivalently, the squared Mahalanobis distance represents the level surface of a multivariate normal distribution. In particular, the region

$$\left\{x \in \mathbb{R}^d : D_\Sigma(x,\mu)^2 \le c\right\}$$

is an ellipsoid centered at μ, whose size and shape are determined by Σ and the scalar c. For a given confidence level $1-\alpha$, the high-density region *(HDR) of a multivariate Gaussian $\mathcal{N}(\mu,\Sigma)$ is precisely the ellipsoid defined by:*

$$\mathcal{R}(1-\alpha) = \left\{x \in \mathbb{R}^d : D_\Sigma(x,\mu)^2 \le \chi^2_{d,1-\alpha}\right\},$$

where $\chi^2_{d,1-\alpha}$ denotes the $1-\alpha$ quantile of the chi-square distribution with d degrees of freedom.

Example 1. *Let us consider for the sake of this example the measurements* sepal width *and* sepal length *of the well-known* Iris Dataset*. Figure 1 shows the scatter plot of this dataset.*

We will calculate the the 95% HDR associated to the consiered dataset using remark 1. We will assume that the sample follows a binormal distribution $\mathcal{N}(\mu,\Sigma)$. To estimate the parameters we will use the maximum likelihood estimators

$$\hat{\mu} = \overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \begin{pmatrix}5.84\\3.05\end{pmatrix}$$

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^{\mathrm{T}} = \begin{pmatrix}0.68 & -0.04\\-0.04 & 0.19\end{pmatrix}$$

In view of remark 2 and 1, the 95% HDR (estimated) region will be

$$\hat{\mathcal{R}}(1-\alpha) = \{(\mathbf{x}-\widehat{\mu})^T\widehat{\Sigma}^{-1}(\mathbf{x}-\widehat{\mu}) \le \chi^2_{d,1-\alpha} \cong 5.99\}$$

this region is the interior of the ellipse represented in figure 1.

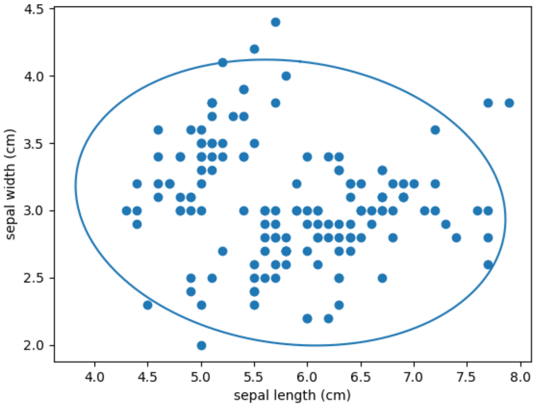

Figure 1: Iris dataset HDR region

3.5. Algorithms definition

Here we present the general version of the High density region X-means algorithm. The procedure assumes that data studied comes from family of distributions $\{f_\theta : \theta \in \Theta\}$, we will explain the particularities of the algorithm for several families of distribution in the following sections.

Algorithm 1 HDR X-Means

Inputs: K_{max} maximun number of clusters, $X = \{x_i\}$ multinormal dataset, $1 - \alpha$ confidence

Output Clusters

```
 1: $k \leftarrow K_{max}$
 2: $stop \leftarrow false$
 3: while $stop = false$ and $k \geq 1$ do
 4:     $Cl \leftarrow \{cl_i\}$ clusters after applying $K$-means to $X$.    ▷ We start applying k means
 5:     $overlap = false$
 6:     for $i = 1$ to $k$ and $j \geq i$ do    ▷ We iterate all pairs of clusters
 7:         $\mathcal{R}_i(1-\alpha) \leftarrow$ High density region of cluster $cl_i$ with probability $1-\alpha$
 8:         $\mathcal{R}_i(1-\alpha) \leftarrow$ High density region of cluster $cl_j$ with probability $1-\alpha$
 9:         if $\mathcal{R}_i(1-\alpha) \cap \mathcal{R}_j(1-\alpha) \neq \emptyset$ then
10:             $k \leftarrow k - 1$
11:             $overlap \leftarrow true$    ▷ If two regions overlap, we remove one cluster
12:             break
13:         end if
14:     end for
15:     if $overlap = false$ then
16:         $stop \leftarrow true$    ▷ if have not found any overlapings we stop the process
17:     end if
18: end while
19: return $Cl$, $k$
```

Notice that in the iterative step (lines 7, 8) we calculate the HDR regions of the clusters cl_i, cl_j using the ideas explained in remark 1

In parallel to HDR X-Means, we introduce *HDR++ Merge*: we first seed up to $K_{\max}$ clusters with k-means++ (one Voronoi assignment), then iteratively merge any clusters whose $(1-\alpha)$ high-density regions (HDRs) overlap. Merging is performed by building a graph on the current clusters with an edge between any pair of intersecting HDRs and contracting connected components in a single pass. Because both methods reason purely in terms of

HDRs, all the theorems and propositions proven for HDR regions apply verbatim to HDR++ Merge as well.

Algorithm 2 HDR++ Merge (k-means++ seeding + HDR-based merging)

```
1: **Input:** $K_{\max}$, dataset $X = \{x_i\}$, confidence $1 - \alpha$
2: **Seeding (k-means++):** choose up to $K_{\max}$ seeds by $D^2$ sampling;
   assign each $x_i$ to the nearest seed to obtain initial clusters $\{cl_j\}$
3: **repeat**
4:     Fit the model family to each $cl_j$ and compute its HDR $\mathcal{R}_j(1-\alpha)$
5:     Build graph $G$ on $\{cl_j\}$ with edge $(i, j)$ iff $\mathcal{R}_i(1-\alpha) \cap \mathcal{R}_j(1-\alpha) \neq \emptyset$
6:     Merge all clusters within each connected component of $G$
7: **until** no edges remain
8: **Output:** final clustering
```

3.5.1. Analysis of the algorithms effectivity

Remark 3. *The idea behind algorithms* **??** *and* **??** *is based in the following principle*

> ***Heuristic.*** *If two normally distributed point clouds cl_i and cl_j are distinct, in the sense that have sufficiently different means and covariance matrices, then their corresponding HDR will not overlap. Otherwise, they will overlap.*

This principle can be seen in action in example **??***, where in the cases in which two cloud points are pieces of the same cluster, the hdr overlap.*

There are two main scenarios in which the previous heuristic, and consequently algorithm **??** *can produce a wrong number of clusters. They both occur in the recursive on lines* **??** *to* **??** *when for two candidate clusters cl_i, cl_j.*

Supposing that $cl_i = \{X_1, \ldots, X_n\}$, $cl_i = \{Y_1, \ldots, Y_n\}$ where $X_1, \ldots, X_n \sim \mathcal{N}(\mu_x, \Sigma_x)$, and $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu_y, \Sigma_y)$, the algorithm may wrongly:

1. *Find that the high density regions $\mathcal{R}_i$ and $\mathcal{R}_j$* ***overlap*** *when in fact the two clusters originate from two distinct normal distributions i. e. $\mu_x \neq \mu_y$ and $\Sigma_x \neq \Sigma_y$. In this case the algorithm* ***will wrongly remove one cluster*** *(merging the two clusters) at line* **??***, producing at least one cluster less than the optimal number, this is what we will call* ***false merging***

2. *Find that the high density regions $\mathcal{R}_i$ and $\mathcal{R}_j$* ***do not overlap*** *when the two clusters originate from the same distribution i. e $\mu_x = \mu_y$ and $\Sigma_x = \Sigma_y$, In this case the algorithm* ***will wrongly regard the clusters as they are, without joining them*** *producing at least one extra cluster than optimal. This type of error will be called* ***false splitting***

In any other case, the algorithm ***will continue correctly guessing the number of clusters****.*

In remark 4 we will point out how theorems 1 and 2 explain the reason behind the described heuristic, and also how the errors described above are unlikely to occur.

Throughout this section we will explore the situations in which (1) and (2) can occur.

The following definition will be useful in this section.

Definition 3. *Given two compact sets K, K' we will denote by $d(K, K')$ the distance between them defined by*

$$d(K, K') = \inf_{x \in K, y \in K'} d(x, y)$$

Notice that by the compactness of K and K' this distance is well defined

The following lemma recalls several well-known properties of the almost sure convergence of random variables. Recall that a sequence of random variables X_n is said to **converge almost everywhere** to X (denoted $X_n \overset{a.s.}{\to} X$) if

$$\mathbb{P}(\lim_{n \to \infty} X_n = X) = 1$$

Lemma 1. 1. *Let X be a random variable on a metric space S. Assume that a map $g : S \to S'$ has the set of discontinuity points D_g such that $\mathbb{P}(X \in D_g) = 0$. If $X_n \overset{a.s.}{\to} X$ then $g(X_n) \overset{a.s.}{\to} g(X)$*

2. *If two sequences of random variables $(X_n)_n$, $(Y_n)_n$ satisfy that $X_n \leq Y_n$ and $Y_n \overset{a.s}{\to} 0$ then $X_n \overset{a.s.}{\to} 0$.*
3. *Given a d-dimensional sample $X_1, \ldots, X_n$ then, taking*

$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$$
$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^T$$

we have that

$$\overline{X} \overset{a.s.}{\to} \mu$$
$$\hat{\Sigma} \overset{a.s.}{\to} \Sigma$$

where $\mu = E[X_i]$ *and* $\Sigma = Var[X_i]$

Proof. (1) This is the continuous map theorem ([25, Theorem 2.3]).

To show (2), notice that if $\lim Y_n(\omega) \to 0$, then since $X_n(\omega) \leq Y_n(\omega)$, we have that $\lim X_n(\omega) \to 0$. This implies that since $\mathbb{P}(\lim Y_n = 0) = 1$ we have that $\mathbb{P}(\lim X_n = 0) = 1$.

(3) is the strong law of large numbers. □

The following is a well known fact about ellipsoids.

Lemma 2. *Given an ellipsoid* $E = \{(\mathbf{x} - c)^t A^{-1}(\mathbf{x} - c) = 1$*, the radius from the center of* E *to its border over each axis* i *coincides with*

$$r_i = \sqrt{\lambda_i}$$

where $\{\lambda_i\}$ *are the eugenvalues of* A*.*

Theorem 1. *Given two* d*-dimensional point clouds* $\mathcal{X} = \{X_1, \ldots, X_n\}$ *and* $\mathcal{Y} = \{Y_1, \ldots, Y_m\}$*, where* $X_i \sim \mathcal{N}(\mu_x, \Sigma_x)$ *and* $Y_i \sim \mathcal{N}(\mu_y, \Sigma_y)$*. Consider* $\hat{\mathcal{R}}_x(1-\alpha)$ *and* $\hat{\mathcal{R}}_y(1-\alpha)$ *their estimated HDR with confidence* $1-\alpha$ *(computed as in remarks 1 and 2). Then*

$$\max\big(0, d(\overline{X}, \overline{Y}) \ - (\hat{R}_x + \hat{R}_y)\sqrt{\chi^2_{d,1-\alpha}}\big) \leq d\big(\hat{\mathcal{R}}_x(1-\alpha), \hat{\mathcal{R}}_y(1-\alpha)\big) \qquad (2)$$
$$\leq \max\big(0, d(\overline{X}, \overline{Y}) \ - (\hat{r}_x + \hat{r}_y)\sqrt{\chi^2_{d,1-\alpha}}\big)$$

where

$$\hat{r}_x = \min_i \sqrt{\hat{\lambda}_i^x} \ \ \hat{r}_y = \min_i \sqrt{\hat{\lambda}_i^y} \qquad (3)$$
$$\hat{R}_x = \max_i \sqrt{\hat{\lambda}_i^x} \ \ \hat{R}_y = \max_i \sqrt{\hat{\lambda}_i^y}$$

And where $\{\hat{\lambda}_i^x\}$*,* $\{\hat{\lambda}_i^y\}$ *are the eugenvalues of the estimated covariance matrices* $\hat{\Sigma}_x$ *and* $\hat{\Sigma}_y$ *respectively.*

Proof. Since $\mathcal{X}$ and $\mathcal{Y}$ are drawn from normal distributions, in view of remarks 1 and 2, we can write

$$\widehat{\mathcal{R}}_x(1-\alpha) = \{(\mathbf{x}-\overline{X})^T\widehat{\Sigma}_x^{-1}(\mathbf{x}-\overline{X}) \leq \chi^2_{d,1-\alpha}\} = \{(\mathbf{x}-\overline{X})^T\left(\widehat{\Sigma}_x\chi^2_{d,1-\alpha}\right)^{-1}(\mathbf{x}-\overline{X}) \leq 1\}$$
$$\widehat{\mathcal{R}}_y(1-\alpha) = \{(\mathbf{x}-\overline{Y})^T\widehat{\Sigma}_y^{-1}(\mathbf{x}-\overline{Y}) \leq \chi^2_{d,1-\alpha}\} = \{(\mathbf{x}-\overline{Y})^T\left(\widehat{\Sigma}_y\chi^2_{d,1-\alpha}\right)^{-1}(\mathbf{x}-\overline{Y}) \leq 1\}$$

We know that the radius in each of the axis of the ellipsoids $\widehat{\mathcal{R}}_x(1-\alpha)$ and $\widehat{\mathcal{R}}_y(1-\alpha)$ can be calculated from the eugenvalues of their matrices $\widehat{\Sigma}_x\chi^2_{d,1-\alpha}$ and $\widehat{\Sigma}_y\chi^2_{d,1-\alpha}$. In fact, by lemma 2 the radii corresponding with the axis i are precisely $\hat{r}^i_x\sqrt{\chi^2_{d,1-\alpha}}$ and $\hat{r}^i_y\sqrt{\chi^2_{d,1-\alpha}}$ writing

$$r^x_i = \sqrt{\hat{\lambda}^x_i}; \;\; r^y_i = \sqrt{\hat{\lambda}^y_i}$$

where $\{\hat{\lambda}^x_i\}$, $\{\hat{\lambda}^y_i\}$ are the eugenvalues of the estimated covariance matrices $\hat{\Sigma}_x$ and $\hat{\Sigma}_y$.

With this in mind we can define the maximun and minimun radii for each ellipsoid $\hat{R}_x, \hat{R}_y, \hat{r}_x, \hat{r}_y$ as in (3). Using this distances, we can define the maximal balls contained in each ellipsoid $B(\overline{X}, \hat{r}_x)$ and $B(\overline{Y}, \hat{r}_y)$, and also the minimal balls containing each ellipsoid $B(\overline{X}, \hat{R}_x)$, $B(\overline{Y}, \hat{R}_y)$. Figure 2 shows an sketch of these definitions.

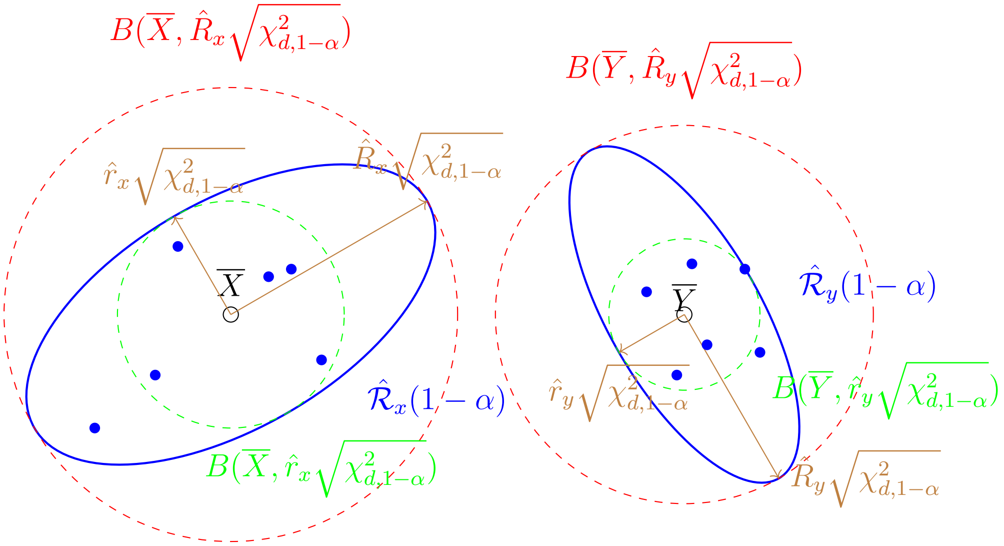
$B(\overline{X}, \hat{R}_x\sqrt{\chi^2_{d,1-\alpha}})$
$B(\overline{Y}, \hat{R}_y\sqrt{\chi^2_{d,1-\alpha}})$
$\hat{r}_x\sqrt{\chi^2_{d,1-\alpha}}$
$\hat{R}_x\sqrt{\chi^2_{d,1-\alpha}}$
$\overline{X}$
$\overline{Y}$
$\hat{\mathcal{R}}_y(1-\alpha)$
$\hat{\mathcal{R}}_x(1-\alpha)$
$\hat{r}_y\sqrt{\chi^2_{d,1-\alpha}}$
$B(\overline{Y}, \hat{r}_y\sqrt{\chi^2_{d,1-\alpha}})$
$B(\overline{X}, \hat{r}_x\sqrt{\chi^2_{d,1-\alpha}})$
$\hat{R}_y\sqrt{\chi^2_{d,1-\alpha}}$

Figure 2: An illustration of the steps in the proof using a two dimensional example

In view of these construction, we obtain that since $\hat{\mathcal{R}}_x(1-\alpha) \subset B(\overline{X}, \hat{R}_y\sqrt{\chi^2_{d,1-\alpha}})$ and $\hat{\mathcal{R}}_y(1-\alpha) \subset B(\overline{Y}, \hat{R}_y\sqrt{\chi^2_{d,1-\alpha}})$ we find

$$d\big(\hat{\mathcal{R}}_x(1-\alpha), \hat{\mathcal{R}}_y(1-\alpha)\big) \geq d\big(B(\overline{X}, \hat{R}_x\sqrt{\chi^2_{d,1-\alpha}}), B(\overline{Y}, \hat{R}_y\sqrt{\chi^2_{d,1-\alpha}})\big) \tag{4}$$

$$= \max\left(0, d(\overline{X}, \overline{Y}) - R_x\sqrt{\chi^2_{d,1-\alpha}} - R_y\sqrt{\chi^2_{d,1-\alpha}}\right) \tag{5}$$

$$= \max\left(0, d(\overline{X}, \overline{Y})\ - (\hat{R}_x + \hat{R}_y)\sqrt{\chi^2_{d,1-\alpha}}\right) \tag{6}$$

Analogously since $B(\overline{X}, \hat{r}_x\sqrt{\chi^2_{d,1-\alpha}}) \subset \hat{\mathcal{R}}_x(1-\alpha)$ and $B(\overline{Y}, \hat{r}_y\sqrt{\chi^2_{d,1-\alpha}}) \subset \hat{\mathcal{R}}_y(1-\alpha)$ we find

$$d\big(\hat{\mathcal{R}}_x(1-\alpha), \hat{\mathcal{R}}_y(1-\alpha)\big) \leq d\big(B(\overline{X}, \hat{r}_x\sqrt{\chi^2_{d,1-\alpha}}), B(\overline{Y}, \hat{r}_y\sqrt{\chi^2_{d,1-\alpha}})\big) \tag{7}$$

$$= \max\left(0, d(\overline{X}, \overline{Y})\ - (\hat{r}_x + \hat{r}_y)\sqrt{\chi^2_{d,1-\alpha}}\right) \tag{8}$$

□

Theorem 2. *Given two d-dimensional point clouds $\mathcal{X} = \{X_1, \ldots, X_n\}$ and $\mathcal{Y} = \{Y_1, \ldots, Y_m\}$, where $X_i \sim \mathcal{N}(\mu_x, \Sigma_x)$ and $Y_i \sim \mathcal{N}(\mu_y, \Sigma_y)$. Consider $\hat{\mathcal{R}}_x(1-\alpha)$ and $\hat{\mathcal{R}}_y(1-\alpha)$ their estimated HDR with confidence $1-\alpha$ (computed as in remarks 1 and 2), and $\mathcal{R}_x(1-\alpha), \mathcal{R}_x(1-\alpha)$ the theoretical HDR calculated from the original distributions $N(\mu_x, \Sigma_x)$ and $N(\mu_y, \Sigma_y)$*

Then

$$d\big(\hat{\mathcal{R}}_x(1-\alpha), \hat{\mathcal{R}}_y(1-\alpha)\big) \overset{a.s.}{\to} d\big(\mathcal{R}_x(1-\alpha), \mathcal{R}_y(1-\alpha)\big) \tag{9}$$

and

$$\max\big(0, d(\mu_X, \mu_Y)\ - (R_x + R_y)\sqrt{\chi^2_{d,1-\alpha}}\big) \leq d\big(\mathcal{R}_x(1-\alpha), \mathcal{R}_y(1-\alpha)\big) \tag{10}$$

$$\leq \max\big(0, d(\mu_X, \mu_Y)\ - (r_x + r_x)\sqrt{\chi^2_{d,1-}}$$

where

$$r_x = \min_i \sqrt{\lambda_i^x} \ \ r_y = \min_i \sqrt{\lambda_i^y} \tag{11}$$

$$R_x = \max_i \sqrt{\lambda_i^x} \ \ R_y = \max_i \sqrt{\lambda_i^y}$$

And where $\{\lambda_i^x\}$, $\{\lambda_i^y\}$ are the eugenvalues of the theoretical covariance matrices Σ_x and Σ_y respectively.

Proof. We will start proving (9). It is clear that $d\big(\hat{\mathcal{R}}_x(1-\alpha), \hat{\mathcal{R}}_y(1-\alpha)\big)$, understood as a function of the two ellipsoids, is continuous over their parameters $\overline{X}, \overline{Y}, \hat{\Sigma}_x$ and $\hat{\Sigma}_y$. Since $\overline{X} \overset{a.s.}{\to} \mu_x$, $\overline{Y} \overset{a.s.}{\to} \mu_y$, $\hat{\Sigma}_x \overset{a.s.}{\to} \Sigma_x$ and $\hat{\Sigma}_y \overset{a.s.}{\to} \Sigma_y$ by the continuous map theorem (lemma 1(1) we deduce (9).

The inequality 11 can be deduced easily following the same steps in the proof of theorem 1 but using the theoretical values μ_x ,Σ_x, μ_y and Σ_y instead of the estimated ones. □

Corollary 1. *Let $X = \{x_1, \dots, x_{n_X}\}$ and $Y = \{y_1, \dots, y_{n_Y}\}$ be two independent samples drawn from the same multivariate normal distribution $\mathcal{N}(\mu, \Sigma) \subset \mathbb{R}^d$. Let $\widehat{\mu}_X, \widehat{\mu}_Y$ denote the empirical means, and let $\widehat{R}_X, \widehat{R}_Y$ be the square roots of the largest eigenvalues of $\widehat{\Sigma}_X$, $\widehat{\Sigma}_Y$, multiplied by $\sqrt{\chi^2_{d,1-\alpha}}$. Then the probability that HDR X-Means falsely separates the clusters satisfies:*

$$\mathbb{P}(\textit{false separation}) \leq \mathbb{P}\left(\chi^2_d > \frac{1}{2\lambda_{\max}} \cdot \frac{n_X n_Y}{n_X + n_Y}(\widehat{R}_X + \widehat{R}_Y)^2 \chi^2_{d,1-\alpha}\right)$$

where $\lambda_{\max}$ is the largest eigenvalue of Σ.

Proof. Since $X, Y \sim \mathcal{N}(\mu, \Sigma)$, the difference of empirical means satisfies:

$$\widehat{\mu}_X - \widehat{\mu}_Y \sim \mathcal{N}\left(0, \Sigma\left(\tfrac{1}{n_X} + \tfrac{1}{n_Y}\right)\right).$$

Let $T := \|\widehat{\mu}_X - \widehat{\mu}_Y\|^2$. Then, taking $Z = \left(\Sigma\left(\tfrac{1}{n_X} + \tfrac{1}{n_Y}\right)\right)^{-1/2}(\widehat{\mu}_X - \widehat{\mu}_Y)$:

$$T = \left(\tfrac{1}{n_X} + \tfrac{1}{n_Y}\right) Z^\top \Sigma Z, \quad Z \sim \mathcal{N}(0, I_d).$$

Bounding using $\lambda_{\max}$:

$$T \leq \left(\tfrac{1}{n_X} + \tfrac{1}{n_Y}\right) \lambda_{\max} \cdot \chi^2_d.$$

The algorithm declares separation if:

$$T > (\widehat{R}_X + \widehat{R}_Y)^2 \cdot \chi^2_{d,1-\alpha}.$$

Combining:

$$\mathbb{P}(\text{false separation}) \leq \mathbb{P}\left(\chi^2_d > \frac{1}{\lambda_{\max}} \cdot \frac{1}{\frac{1}{n_X} + \frac{1}{n_Y}} (\widehat{R}_X + \widehat{R}_Y)^2 \chi^2_{d,1-\alpha}\right),$$

and noting $\frac{1}{\frac{1}{n_X}+\frac{1}{n_Y}} = \frac{n_X n_Y}{n_X+n_Y}$, the result follows. □

Corollary 2. *Let $X = \{x_1, \ldots, x_{n_X}\} \sim \mathcal{N}(\mu_X, \Sigma)$ and $Y = \{y_1, \ldots, y_{n_Y}\} \sim \mathcal{N}(\mu_Y, \Sigma)$ be two independent samples from distinct multivariate normal distributions with common covariance Σ. Define the Mahalanobis distance $\delta = \|\mu_X - \mu_Y\|^2_{\Sigma^{-1}}$ and:*

$$\lambda = \frac{n_X n_Y}{n_X + n_Y} \cdot \frac{\delta}{2}.$$

Let $\widehat{R}_X, \widehat{R}_Y$ be the estimated HDR radii as before. Then, the probability of false merging (i.e., failing to separate two distinct clusters) satisfies:

$$\mathbb{P}(\textit{false merging}) \leq F_{\chi^2_d(\lambda)}\left(\frac{n_X + n_Y}{2 n_X n_Y} (\widehat{R}_X + \widehat{R}_Y)^2 \cdot \chi^2_{d,1-\alpha}\right)$$

where $F_{\chi^2_d(\lambda)}$ is the CDF of the noncentral χ^2 distribution with noncentrality parameter λ.

Proof. We have:

$$\widehat{\mu}_X - \widehat{\mu}_Y \sim \mathcal{N}(\mu_X - \mu_Y, \Sigma(\tfrac{1}{n_X} + \tfrac{1}{n_Y})),$$

so that:

$$T := \|\widehat{\mu}_X - \widehat{\mu}_Y\|^2 \sim \left(\tfrac{1}{n_X} + \tfrac{1}{n_Y}\right) \cdot \chi^2_d(\lambda), \quad \lambda = \frac{\delta}{\frac{2}{n_X} + \frac{2}{n_Y}} = \frac{n_X n_Y}{n_X + n_Y} \cdot \frac{\delta}{2}.$$

The algorithm fails to separate the clusters if:

$$T < (\widehat{R}_X + \widehat{R}_Y)^2 \cdot \chi^2_{d,1-\alpha}.$$

Solving:

$$\mathbb{P}(\text{false merging}) \leq F_{\chi^2_d(\lambda)}\left(\frac{1}{\frac{1}{n_X} + \frac{1}{n_Y}} \cdot (\widehat{R}_X + \widehat{R}_Y)^2 \chi^2_{d,1-\alpha}\right) = F_{\chi^2_d(\lambda)}\left(\frac{n_X + n_Y}{n_X n_Y} \cdot \frac{1}{2}(\widehat{R}_X + \widehat{R}_Y)^2 \chi^2_{d,1-}\right.$$

□

Remark 4. *Theorems 1 and 2, together with Corollary 1, allow us to understand the accuracy of Algorithm* **??** *and the nature of the potential errors it may incur.*

1. *Theorem 2 confirms the heuristic in Remark 3: if two candidate clusters $cl_i = \{X_1, \ldots, X_n\}$ and $cl_j = \{Y_1, \ldots, Y_n\}$ come from different distributions and their means μ_X, μ_Y are sufficiently separated relative to their dispersion, then their HDRs will not intersect. Specifically, the inequality*

$$\max\left(0, d(\mu_X, \mu_Y) - (R_X + R_Y)\sqrt{\chi^2_{d,1-\alpha}}\right) \leq d\left(\mathcal{R}_X(1-\alpha), \mathcal{R}_Y(1-\alpha)\right)$$

guarantees that the distance between the HDRs is positive as long as the separation between means is large enough to outweigh the combined spread of the clusters.

2. *However, as this inequality also shows,* ***high variance*** *(large R_X, R_Y) increases the likelihood that two distinct clusters have overlapping HDRs. This leads to* ***false merging****, i.e., the algorithm failing to separate them. In practice, this is a tolerable error, since large overlap implies that the clusters are statistically difficult to distinguish. Still, it is an important limitation: HDR X-Means becomes less sensitive in high-variance settings.*

3. *Corollary 1 provides a probabilistic estimate of the opposite error:* ***false separation*** *of two clusters that come from the* same *distribution. When two independent samples cl_i, cl_j from $\mathcal{N}(\mu, \Sigma)$ are treated as separate clusters, the separation occurs only if the empirical means $\widehat{\mu}_X, \widehat{\mu}_Y$ are further apart than the sum of the HDR radii. The probability of this event is approximately*

$$\mathbb{P}\left(\|\widehat{\mu}_X - \widehat{\mu}_Y\| > (\widehat{R}_X + \widehat{R}_Y)\sqrt{\chi^2_{d,1-\alpha}}\right).$$

This error becomes more likely in cases of high variance or small sample size, as both contribute to larger fluctuations in the empirical means and HDR size.

4. *Therefore,* ***both types of error****false merging and false separationare exacerbated by* ***high variance****. In the case of false merging, variance increases the chance of overlap; in the case of false separation, it increases the spread of the empirical means and the HDRs. However, the former is more benign in practice, while the latter is theoretically quantifiable and decreases with larger sample sizes.*

Clustering Validation Indices

External Validation Indices

Let:

- a: Number of pairs of data points that are in the same cluster in both the predicted clustering C and the ground truth P.
- b: Number of pairs that are in the same cluster in C but in different clusters in P.
- c: Number of pairs that are in different clusters in C but in the same cluster in P.
- d: Number of pairs that are in different clusters in both C and P.
- $M = \frac{n(n-1)}{2}$: Total number of possible pairs in the dataset of n data points.

These indices compare the clustering result to a ground truth or predefined classification.

- **Rand Index (RI)**:
$$RI = \frac{a+d}{M}$$
where a and d are the number of agreements between clustering and ground truth, and $M = \frac{n(n-1)}{2}$ is the total number of pairs.
- **Jaccard Coefficient (J)**:
$$J = \frac{a}{a+b+c}$$
- **Fowlkes-Mallows Index (FM)**:
$$FM = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$$

Internal Validation Indices

Let:

- n: Total number of data points.
- K: Number of clusters.
- C_i: The i-th cluster.
- n_i: Number of data points in cluster C_i.
- m_i: Centroid of cluster C_i.
- m: Overall mean of the dataset.
- e_i: Average intra-cluster distance (error) for cluster C_i.
- $D(C_i, C_j)$: Distance between clusters C_i and C_j.
- $\Delta(C_l)$: Diameter of cluster C_l, i.e., the maximum distance between any two points in C_l.

These indices evaluate the clustering quality based on intra-cluster compactness and inter-cluster separation.

- **Calinski-Harabasz Index (CH)**:

  $$CH(K) = \frac{\mathrm{Tr}(S_B)}{K-1} \Big/ \frac{\mathrm{Tr}(S_W)}{n-K}$$

  where S_B is the between-cluster scatter matrix and S_W is the within-cluster scatter matrix.

- **Davies-Bouldin Index (DB)**:

  $$DB(K) = \frac{1}{K}\sum_{i=1}^{K} R_i, \quad R_i = \max_{j \neq i}\left(\frac{e_i + e_j}{\|m_i - m_j\|}\right)$$

- **Dunn Index (DI)**:

  $$DI(K) = \min_{i=1,\ldots,K}\left(\min_{j \neq i} \frac{D(C_i, C_j)}{\max_{l=1,\ldots,K} \Delta(C_l)}\right)$$

  where $D(C_i, C_j)$ is the distance between clusters and $\Delta(C_l)$ is the diameter of cluster C_l.

3.6. Comments on the validation

Previously, the validation techniques has been applied to 73 datasets [12]. Based on these techniques, Deng et al. [26] proposed a new technique to infer k.

3.7. Benchmark on the DERIC Clustering Repository

We evaluated the proposed algorithms on the `deric/clustering-benchmark` repository [27], reporting results separately for real-world and simulated datasets.

Performance was summarized using:

- **External (pairwise) indices**: Rand Index **(RI; higher is better)**, Jaccard Index **(J; higher is better)**, and Fowlkes–Mallows **(FM; higher is better)**, which compare pairwise agreements between the predicted and reference partitions. We also report the scikit-learn Fowlkes–Mallows Score **(FMS)**, numerically equivalent to FM in our evaluations.

- **Internal (geometry) indices**: Calinski–Harabasz **(CH; higher is better)**, Davies–Bouldin **(DB; lower is better)**, and Dunn **(Dunn; higher is better)**, which capture cluster compactness and separation without labels.

- **Model order error**: the absolute deviation in the number of clusters, $|k_{err}| = |k_{\text{real}} - k_{\text{predicted}}|$ (smaller is better).

3.7.1. Real-world datasets.

Table 1 indicates a clear division between model-order accuracy and geometric separation. `HDR++ Merge` attains the lowest model-order error ($|k_{\text{err}}| = 10.130$) while also delivering the strongest pairwise co-membership via Fowlkes–Mallows. Among classic mixture/centroid models, `GaussianMixture` achieves the highest Rand Index and `GMeans` yields the best Jaccard. Reachability-based methods excel on internal geometry: `OPTICS` attains the best compactnessseparation profile with the highest Calinski–Harabasz (CH$= 278.411$), the lowest Davies–Bouldin, and the highest Dunn, while maintaining a competitive model-order error. `HDBSCAN` similarly shows strong geometry with modest $|k_{\text{err}}|$.

On the other hand, `IEPC` attains a model-order error comparable to the top group ($|k_{\text{err}}| = 10.391$) and mid-range pairwise/internal scores, whereas

`DensityPeaks` displays high separation as reflected by a large CH (246.374) but lags on pairwise alignment and exhibits the largest model-order error ($|k_{\text{err}}| = 24.478$). Overall, the HDR-based approaches (`HDR++ Merge`, `HDR X-means`) remain competitive across metrics`HDR++ Merge` in particular combining the best $|k_{\text{err}}|$ with leading FMwhile reachability methods dominate purely geometric separation.

Table 1: DERIC real-world datasets External (pairwise) and internal (geometry) indices. Higher is better for RI, J, FM, CH, Dunn; lower is better for DB.

	External (pairwise) indices			**Internal (geometry) indices**			
Algorithm	Rand Index (RI)	Jaccard Index (J)	Fowlkes–Mallows (FM)	Calinski–Harabasz (CH)	Davies–Bouldin (DB) ↓	Dunn Index ↑	$\lvert k_{\text{err}} \rvert$
GaussianMixture	**0.635**	0.300	0.462	123.310	1.775	0.239	15.870
GMeans	0.607	**0.370**	0.528	131.471	1.306	0.320	14.783
HDBSCAN	0.512	0.302	0.476	115.619	0.922	0.598	10.783
DensityPeaks	0.484	0.254	0.432	246.374	1.384	0.255	24.478
XMeans	0.456	0.322	0.513	150.436	1.583	0.252	23.565
OPTICS	0.447	0.331	0.518	**278.411**	0.599	0.774	10.478
DBSCAN	0.441	0.341	0.532	217.153	1.061	0.477	10.696
HDR++ Merge	0.425	0.361	**0.557**	28.131	**0.837**	**0.382**	**10.130**
IEPC	0.423	0.348	0.539	140.860	1.529	0.270	10.391
HDR X-means	0.398	0.328	0.525	70.864	1.278	0.373	11.217

3.7.2. Artificial datasets.

Regarding the synthetic benchmarks (Table 2), `IEPC` attains the highest Rand Index and the smallest average model-order error ($|k_{\text{err}}|$), but its internal geometry is weak, with an extremely large Davies–Bouldin and a very low Dunn, indicating poorly compact/ separated clusters despite many correct pairwise matches. `HDBSCAN` leads the pairwise-agreement metrics (highest Jaccard and Fowlkes–Mallows) and also yields the top Dunn, yet its DB is again extremely large, suggesting over-segmentation on several shapes.

Against this backdrop, the `HDR` methods show complementary strengths with more balanced profiles. `HDR X-means` achieves the *lowest* DB across methods (best compactness) while maintaining competitive RI and one of the larger CH values (second only to XMeans), resulting in a favorable compactnessseparation trade-off. `HDR++ Merge` is competitive on RI (near the upper tier) with a mid-range DB and CH, offering strong pairwise alignment without the extreme geometry instabilities seen in some density/ reachability methods.

Among the non-HDR classical baselines, `XMeans` maximizes CH (very high between-cluster separation) but lags on pairwise indices; `OPTICS` attains the second-lowest $|k_{\text{err}}|$ but shows weaker pairwise and geometry (very large DB). `Density Peaks` posts solid Jaccard/FM and a relatively high Dunn, though with middling DB and model-order error. Overall, `HDR X-means` delivers the most consistent internal geometry (best DB with strong CH), while `HDR++ Merge` provides high pairwise recovery with controlled geometrytogether offering balanced performance across diverse artificial cluster shapes.

Table 2: DERIC simulated datasets External (pairwise) and internal (geometry) indices. Higher is better for RI, J, FM, CH, Dunn; lower is better for DB.

	External (pairwise) indices			**Internal (geometry) indices**			
Algorithm	Rand Index (RI)	Jaccard Index (J)	Fowlkes–Mallows (FM)	Calinski–Harabasz (CH)	Davies–Bouldin (DB) ↓	Dunn Index ↑	$\lvert k_{\text{err}}\rvert$
IEPC	**0.883**	0.257	0.295	5597.602	72879.908	0.092	**345.968**
HDR X-means	**0.852**	0.243	0.294	**12887.043**	**0.539**	0.542	538.905
HDR++ Merge	0.851	0.214	0.279	8605.892	0.612	0.216	539.358
HDBSCAN	0.798	**0.328**	**0.361**	4375.179	1817682.017	**0.649**	542.937
GaussianMixture	0.794	0.249	0.299	4873.272	0.611	0.569	546.263
OPTICS	0.781	0.156	0.207	809.732	14963177.670	0.332	**514.695**
GMeans	0.778	0.240	0.304	5698.737	**0.544**	0.551	541.737
DBSCAN	0.725	**0.286**	**0.335**	1133.809	4846836.619	0.436	542.768
DensityPeaks	0.723	0.279	0.330	2758.988	0.683	0.585	543.853
XMeans	0.550	0.155	0.241	**25799.178**	0.556	0.462	534.800

4. Conclusions

We have introduced HDR X-Means, a novel clustering algorithm that extends the traditional X-Means framework by incorporating High-Density Regions (HDR) for cluster validation and merging decisions. By leveraging HDRs, our method effectively captures the underlying data distribution, allowing for more accurate identification of cluster boundaries and improved handling of non-convex shapes. The algorithms based on simple statistical principles to decide whether to merge or split clusters, relying on the distance between their HDRs. We have provided theoretical guarantees on the convergence and accuracy of our HDR-based distance measure, demonstrating its robustness in various scenarios.

Extensive experiments on both synthetic and real-world datasets demonstrate that HDR X-Means outperforms existing clustering methods in terms of clustering quality and model order accuracy. Our approach shows particular strength in handling datasets with complex geometries and varying densities, where traditional methods often strugle. The methods are easily implementable and computationally efficient, making them suitable for large-scale applications. Future work will explore extensions to other clustering paradigms and investigate the integration of HDR concepts into deep learning-based clustering frameworks.

References

[1] D. Arthur, S. Vassilvitskii, k-means++: the advantages of careful seeding, in: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, SODA '07, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2007, pp. 1027–1035. URL: http://dl.acm.org/citation.cfm?id=1283383.1283494.

[2] D. Pelleg, A. W. Moore, X-means: Extending k-means with efficient estimation of the number of clusters, in: Proceedings of the Seventeenth International Conference on Machine Learning, ICML '00, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000, pp. 727–734. URL: http://dl.acm.org/citation.cfm?id=645529.657808.

[3] K. V. Mardia, P. E. Jupp, Directional statistics, John Wiley & Sons, 2009.

[4] P. Berens, Circstat: A matlab toolbox for circular statistics, Journal of Statistical Software 31 (2009) 1–21. doi:10.18637/jss.v031.i10.

[5] X. Ran, Y. Xi, Y. Lu, et al., Comprehensive survey on hierarchical clustering algorithms and the recent developments, Artificial Intelligence Review 56 (2023) 8219–8264. URL: https://doi.org/10.1007/s10462-022-10366-3. doi:10.1007/s10462-022-10366-3.

[6] S. Wijuniamurti, S. Nugroho, R. Rachmawati, Agglomerative nesting (agnes) method and divisive analysis (diana) method for hierarchical

clustering on some distance measurement concepts, Journal of Statistics and Data Science 1 (2022) 7–11.

[7] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD), Portland, OR, USA, 1996, pp. 226–231.

[8] A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks, Science 344 (2014) 1492–1496. doi:10.1126/science.1242072.

[9] S. Han, X. Zhang, X. Liu, Y. Zheng, J. Qu, Atsdpc: Adaptive two-stage density peaks clustering with hybrid distance based on dispersion coefficient, Expert Systems with Applications (2025) 127639.

[10] G. J. Oyewole, G. A. Thopil, Data clustering: application and trends, Artificial intelligence review 56 (2023) 6439–6475.

[11] A. E. Ezugwu, A. K. Shukla, M. B. Agbaje, O. N. Oyelade, A. José-García, J. O. Agushaka, Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature, Neural Computing and Applications 33 (2021) 6247–6306.

[12] A. Mourer, F. Forest, M. Lebbah, H. Azzag, J. Lacaille, Selecting the number of clusters K with a stability trade-off: an internal validation criterion, CoRR abs/2006.08530 (2020). URL: `https://arxiv.org/abs/2006.08530`. `arXiv:2006.08530`.

[13] G. Hamerly, C. Elkan, Learning the k in k-means, in: Advances in neural information processing systems, volume 16, 2004, pp. 281–288.

[14] A. Banerjee, I. S. Dhillon, J. Ghosh, S. Sra, Clustering on the unit hypersphere using von mises-fisher distributions, Journal of Machine Learning Research 6 (2005) 1345–1382.

[15] K. Hornik, B. Grün, Clustering around circular means, International Journal of Circular Statistics 2 (2014) 3–17.

[16] M. Ankerst, M. M. Breunig, H.-P. Kriegel, J. Sander, Optics: Ordering points to identify the clustering structure, in: Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, ACM, 1999, pp. 49–60.

[17] X. Wei, M. Peng, H. Huang, Y. Zhou, An overview on density peaks clustering, Neurocomputing 554 (2023) 126633. URL: `https://www.sciencedirect.com/science/article/pii/S0925231223007567`. doi:https://doi.org/10.1016/j.neucom.2023.126633.

[18] J. Lu, Z. Chen, J. Shao, C. Wu, Information entropy peaks clustering using dynamic reverse nearest neighbor sequence and 3d decision graph, Expert Systems with Applications 288 (2025) 128197. URL: `https://www.sciencedirect.com/science/article/pii/S0957417425018172`. doi:https://doi.org/10.1016/j.eswa.2025.128197.

[19] R. J. Hyndman, Q. Yao, Estimating and visualizing conditional densities, Journal of Computational and Graphical Statistics 5 (1996) 315–336. doi:10.2307/1390803.

[20] D. Pelleg, A. W. Moore, X-means: Extending k-means with efficient estimation of the number of clusters, in: Proceedings of the Seventeenth International Conference on Machine Learning (ICML), 2000, pp. 727–734.

[21] D. Arthur, S. Vassilvitskii, k-means++: The advantages of careful seeding, in: Proc. 18th ACMSIAM SODA, 2007, pp. 1027–1035.

[22] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proc. 2nd KDD, 1996, pp. 226–231.

[23] R. J. G. B. Campello, D. Moulavi, J. Sander, Density-based clustering based on hierarchical density estimates, PAKDD (2013) 160–172.

[24] L. McInnes, J. Healy, S. Astels, hdbscan: Hierarchical density based clustering, Journal of Open Source Software 2 (2017) 205. doi:10.21105/joss.00205.

[25] A. W. Van der Vaart, Asymptotic statistics, volume 3, Cambridge university press, 2000.

[26] J. Deng, X. Pan, H. Yang, J. Yin, Variational loss ofărandom sampling forăsearching cluster number, in: C. Cao, H. Chen, L. Zhao, J. Arshad, T. Asyhari, Y. Wang (Eds.), Knowledge Science, Engineering and Management, Springer Nature Singapore, Singapore, 2024, pp. 130–143.

[27] T. Barton, Clustering benchmarks, `https://github.com/deric/clustering-benchmark`, 2019. GitHub repository. Last commit on master: 2019-06-17 (fc7aba7). Accessed: 2025-09-25.