Statistical analysis of a hierarchical clustering algorithm with outliers

Nicolas Klutchnikoff, Audrey Poterie, Laurent Rouviere

Statistical analysis of a hierarchical clustering algorithm with outliers

Nicolas Klutchnikoff* Audrey Poterie† Laurent Rouvière‡

Abstract

It is well known that the classical single linkage algorithm usually fails to identify clusters in the presence of outliers. In this paper, we propose a new version of this algorithm and we study its mathematical performances. In particular, we establish an oracle type inequality which ensures that our procedure allows to recover the clusters with large probability under minimal assumptions on the distribution of the outliers. We deduce from this inequality the consistency and some rates of convergence of our algorithm for various situations. Performances of our approach is also assessed through simulation studies and a comparison with classical clustering algorithms on simulated data is also presented.
Keywords: Clustering, single linkage, clustering risk
AMS Subject Classification: 62G07, 62H30

1 Introduction

In unsupervised learning, clustering refers to a very broad set of tools which aim at finding a partition of the data into dissimilar groups so that the observations within each group are quite similar to each other. Considered as one of the most important questions in unsupervised learning, there is a vast literature on this paradigm (Hartigan, 1975, Jain and Dubes, 1988, Duda

*Univ Rennes, CNRS, IRMAR (Institut de Recherche Mathématique de Rennes) - UMR 6625, F-35000 Rennes, France
†Univ Bretagne Sud, LMBA (Laboratoire de Mathématiques Bretagne Atlantique) - UMR CNRS 6205, F-56000 Vannes, France
‡Univ Rennes, CNRS, IRMAR (Institut de Recherche Mathématique de Rennes) - UMR 6625, F-35000 Rennes, France

et al., 2012). Moreover, a lot of various clustering methods have been developed, such as the k-means algorithm (MacQueen, 1967), the hierarchical clustering methods (Johnson, 1967), the spectral clustering algorithms (Ng et al., 2002) or the model-based clustering approaches (McLachlan and Basford, 1988). Clustering plays an important role in explanatory data analysis and has been used in many fields including pattern recognition (Satish and Sekhar, 2006), image analysis (Filipovych et al., 2011), document retrieval, bioinformatics (Yamanishi et al., 2004, Zeng et al., 2012), data compression (Gersho and Gray, 2012). Overall, clustering tools are often used to help users understand the data structure. Furthermore, with the massive increase in the amount of collected and stored data, clustering methods can also be used as dimensionality reduction techniques (Yengo et al., 2014).

In this paper, we consider a mathematical framework close to the one used in Maier et al. (2009), Arias-Castro et al. (2011), Auray et al. (2015). The data are generated according to a mixture of several distributions whose supports are assumed to be disjoint in order to identify the groups. Moreover, we assume that the data is contaminated by outliers, observations that do not belong to any of the supports. Many authors have studied theoretical properties of nearest neighbor graphs, hierarchical and spectral clustering algorithms in a similar context. For instance Maier et al. (2009) provide a deep analysis of k-nearest neighbor graphs while Arias-Castro (2011) study performances of spectral clustering algorithms and single linkage algorithms under assumptions on both distances between supports and number of outliers.

Single linkage algorithm is a hierarchical method which consists in recursively merging the two closest clusters in term of minimal distance. Even if this procedure has many interesting properties, it is well known that it is not robust to the presence of outliers. This lack of robustness comes from two different phenomena. On the one hand, the procedure may wrongly detect small clusters among the outliers. On the other hand, during the recursive clustering procedure, a chain of outliers may merge two clusters which contains observations that belong to two different supports. To circumvent these problems, we propose a simple procedure based on a new analysis of the dendrogram produced by the hierarchical agglomerative clustering in term of minimal distance. This new procedure can be viewed as a simple modification of the classical single linkage algorithm which is both robust to the presence of outliers and well adapted to detect low-dimensional geometrical structures. This last property, shared with spectral clustering, is of prime interest in several modern applications (see Arias-Castro, 2011). Like spectral clustering, our data-driven procedure only requires the knowledge of the number of groups

to identify the clusters with large probability (under mild assumptions on the size of the clusters). The main advantage of our approach, compared to spectral clustering, lies in its time complexity which is drastically better.

The paper is organized as follows. In Section 2, we introduce the mathematical framework, the model assumptions and we define a criterion, called *clustering risk*, to measure the performance of a clustering algorithm. Section 3 describes the single linkage clustering algorithm and shows, through simple examples, that this algorithm often fails to recover the true clusters in the presence of outliers. Then, a robust version of this algorithm which requires the selection of a positive parameter (a radius) to stop the algorithm is proposed. Section 4 provides an oracle inequality which ensures that this procedure is efficient. Some consistency results and rates of convergence are deduced from this inequality. Finally, Section 5 is devoted to highlight the performances of the proposed algorithms in comparison with the single linkage algorithm, the k-means method and the spectral clustering through several synthetic data sets. The proofs are gathered at the end of the paper, in Section 6. The proposed clustering method has been implemented in R and the source code is available at https://github.com/klutchnikoff/robustSL.

2 Mathematical framework

In this section, we define a sufficiently general probabilistic model to generate data which locally lie into low-dimensional structures and which possibly contain outliers. We also specify what we expect from a clustering procedure in this framework and we define a risk to assess the performance of such a procedure.

2.1 Generative model

We are given n independent $[0,1]^D$-valued random variables $X_1, \dots, X_n$ randomly drawn from a distribution $\mathbb{P}$ which can be written as a mixture of $M+1$ distributions $\mathbb{P}_0, \dots, \mathbb{P}_M$. More precisely, for $0 \leq \varepsilon < 1$ and a vector of convex weights $(\gamma_1, \dots, \gamma_M)$, we assume that

$$\mathbb{P} = \varepsilon \mathbb{P}_0 + (1-\varepsilon) \sum_{i=1}^{M} \gamma_i \mathbb{P}_i. \tag{2.1}$$

In this decomposition, ε denotes the weight of outliers contained in the data and $\mathbb{P}_0$ is the distribution of these outliers. The second part provides the distribution for the *actual data* or *non-outlier data*. These data are expanded into M groups, γ_i represents the weight for the i-th group and $\mathbb{P}_i$ its distribution. Let

$$S_i = \text{supp}(\mathbb{P}_i), \quad i \in \{1, \dots, M\}$$

be the (compact) set of all points $x \in \mathbb{R}^D$ for which any neighborhood A of x satisfies $\mathbb{P}_i(A) > 0$. We also define the set

$$S_0 = [0,1]^D \setminus \left(\bigcup_{i=1}^{M} S_i \right)$$

which refers to the support of the outliers.

2.2 Assumptions

Clusters are usually identified by high density regions separated by low density regions. For instance, Hartigan (1975) defines cluster as the connected components of the level sets of the density of the observations. Moreover, the geometry of the cluster often corresponds to low dimensional structures such as manifolds (Arias-Castro, 2011, Arias-Castro et al., 2011). We propose a similar way which is specified in the following assumptions.

(A1) For each $i \in \{1, \dots, M\}$, the set S_i is connected. Moreover

$$\delta = \min_{1 \le i < j \le M} \min\{\|x - y\| : (x, y) \in S_i \times S_j\} > 0,$$

where $\|\cdot\|$ stands for the Euclidean norm.

Assumption **(A1)** ensures that the supports are disjoint and well-separated. This implies that the model is identifiable since the decomposition (2.1) of $\mathbb{P}$ is then unique. Note also that, under this assumption, the dataset $\mathbb{X}_n = \{X_1, \dots, X_n\}$ is splitted into $M+1$ well-defined *groups*: $\mathbb{X}_n \cap S_0$ corresponds to the outliers whereas for $i \ge 1$, $\mathbb{X}_n \cap S_i$ correspond to the *true clusters* we want to recover.

Throughout the paper, for any $0 \le s \le D$, we denote by $\mathcal{H}^s$ the s-dimensional Hausdorff outer measure. We recall that, if s is an integer, this measure agrees

with ordinary "s–dimensional surface area" on regular sets. In particular $\mathcal{H}^D$ is the standard Lebesgue measure on the ambient space $\mathbb{R}^D$. We refer the reader to Evans and Gariepy (2015) for more details on this topic. We also define:

$$s_i = \dim_H(S_i) \quad \text{and} \quad d = \max\{s_i : i \in \{1, \dots, M\}\},$$

where $\dim_H(S_i)$ denotes the Hausdorff dimension of the set S_i, that is the unique real number $s \in [0, D]$ such that $\mathcal{H}^t(S_i) = \infty$ if $t < s$ and $\mathcal{H}^t(S_i) = 0$ if $t > s$. Remark that if S_i is a submanifold of $\mathbb{R}^D$, its Hausdorff dimension s_i corresponds to its classical dimension.

(A2) There exists $\kappa_0 > 0$ such that, for any $A \subseteq S_0$

$$\mathbb{P}_0(A) \leq \kappa_0 \mathcal{H}^D(A).$$

This assumption relates on the distribution of the outliers and can be reformulated as follows: $\mathbb{P}_0$ is absolutely continuous with respect to $\mathcal{H}^D$, with bounded Radon-Nikodym derivative. This implies, in some sense, that the outliers are nowhere dense in S_0 and thus prevents from having clusters that correspond to groups of outliers.

(A3) For any $i \in \{1, \dots, M\}$, there exists $\kappa_i > 0$ such that, for any $A \subseteq S_i$:

$$\mathbb{P}_i(A) \geq \kappa_i^{-1} \mathcal{H}^{s_i}(A).$$

Assumption **(A3)** relates on the distribution of *actual data* and ensures that each $\mathbb{P}_i$ is quite dense on S_i. Note in particular that $\mathbb{P}_i$ can be singular with respect to $\mathcal{H}^{s_i}$ ($\mathbb{P}_i$ is singular with respect to $\mathcal{H}^D$ as soon as $s_i < D$).

Assumptions **(A1)**, **(A2)** and **(A3)** are classical in the clustering setting. The first one guarantees identifiability of the model while the other two highlight differences between outliers and actual data: the former are diffused while the latter are distributed in a dense way into their supports.

Geometric assumptions on the supports S_i are also needed. To state them, we denote by $B(x, r)$ the Euclidean ball centered at $x \in \mathbb{R}^D$ with radius $r > 0$ and by Γ the usual gamma function. Recall that, for any $s > 0$, the function $\eta(s) = \pi^{s/2}\Gamma^{-1}(1 + s/2)$ generalizes to non-integer parameters the volume of the unit ball in dimension s.

(A4) There exists $\kappa_c \geq 1$ such that, for any $i \in \{1, \dots, M\}$, $x \in S_i$ and $0 < r \leq \Delta_i = \operatorname{diam}(S_i)$,

$$\kappa_c^{-1} \leq \frac{\mathcal{H}^{s_i}(S_i \cap B(x, r))}{\eta(s_i) r^{s_i}} \leq \kappa_c.$$

Here $\operatorname{diam}(S_i) = \max\{\|x - y\| : x \in S_i, y \in S_i\}$ denotes the diameter of the support S_i.

Assumption **(A4)** prevents the sets S_i from being "too narrow" in some places. A similar assumption is made by Arias-Castro (2011). Note also that, if S_i is a submanifold that satisfies a *reach* condition, then **(A4)** is automatically fulfilled (see Biau et al., 2007, and references therein).

(A5) For any $i \in \{1, \dots, M\}$, the Hausdorff dimension s_i of S_i agrees with its Minkowski-Bouligand dimension, that is:

$$s_i = \lim_{r \to 0} \frac{\log(N_r(S_i))}{\log(1/r)},$$

where $N_r(S_i)$ denotes the minimal number of open balls of radius r required to cover S_i.

Assumption **(A5)** is needed to obtain sharp bounds on the covering numbers $N_r(S_i), r > 0, i = 1, \dots, M$. Indeed, in general, we only have

$$s_i \leq \liminf_{r \to 0} \frac{\log(N_r(S_i))}{\log(1/r)} \leq \limsup_{r \to 0} \frac{\log(N_r(S_i))}{\log(1/r)} \leq D.$$

Here we assume that the limit inferior matches with the limit superior and that these limits equal s_i. This technical assumption is not too restrictive. We offer two simple generic examples. First, if S_i is a submanifolds of $\mathbb{R}^D$ then **(A5)** holds. Indeed, in this case, the Hausdorff dimension and the Minkowski-Bouligand dimension both match with the usual dimension of S_i. Next, **(A5)** is also satisfied if S_i is a self-similar set. Indeed using Assumption **(A4)** with $r = \Delta_i$, we obtain that $0 < \mathcal{H}^{s_i}(S_i) < +\infty$. This implies that S_i satisfies the *open set condition* which allows us to conclude that both the Hausdorff and the Minkowski-Boulingand dimensions match with the affinity dimension of the self-similar set S_i (see Falconer, 2014, chapter 9).

Last assumption relates on the size of the clusters.

(A6) Let $\gamma_* = \min\{\gamma_i : i \in \{1, \dots, M\}\}$, $\gamma^* = \max\{\gamma_i : i \in \{1, \dots, M\}\}$ and $\not\gamma = \gamma_* - \gamma^*/2$. We assume that:

$$\gamma^* < 2\gamma_* \quad \text{and} \quad 0 \leq \varepsilon < \not\gamma/(1 + \not\gamma).$$

This assumption allows to differentiate actual clusters from the group of outliers. It implies that the sizes of the actual clusters should be of the same order since the largest cluster cannot be twice as large as the smallest one. Moreover, the number of allowed outliers is constrained by the difference in size of the groups: ε could vary from 0 (if $\gamma^* = 2\gamma_*$) to $1/(2M + 1)$ (if $\gamma^* = \gamma_* = 1/M$).

2.3 Clustering risk

We aim at finding a *clustering procedure* that group together the data that lie within the same set S_i, for each $i \in \{1, \dots, M\}$. Regarding the outliers, they can be affected to any other group or garbage into a specific group by the procedure.

A clustering procedure consists of splitting the data $\mathbb{X}_n$ into M disjoint clusters. In other words, a clustering algorithm provides a family of clusters $\mathcal{X} = \{\mathcal{X}_1, \dots, \mathcal{X}_M\}$ such that

$$\begin{cases} \mathcal{X}_i \neq \emptyset, \, i = 1, \dots, M \\ \mathcal{X}_i \cap \mathcal{X}_j = \emptyset, \, 1 \leq i \neq j \leq M \\ \bigcup_{i=1}^{M} \mathcal{X}_i \subseteq \mathbb{X}_n. \end{cases}$$

Observe that the family $\mathcal{X}$ may not cover the whole set $\mathbb{X}_n$, it could be the case if the algorithm reveals some outliers, these outliers are not assigned into one cluster.

In our context, a clustering procedure is efficient if each cluster contains all the observations from (only) one of the supports S_i, $i \in \{1, \dots, M\}$. It means that there exists a unique permutation $\pi \in \Pi_m$ from the set of all permutations of $\{1, \dots, M\}$, such that, for any $i = 1, \dots, M$, the data $\mathbb{X}_n \cap S_i$ are included into $\mathcal{X}_{\pi(i)}$. In this context, we measure the performances of a clustering procedure by the *clustering risk* defined as

$$\mathcal{R}_n(\mathcal{X}) = \mathbb{P}(\forall \pi \in \Pi_M, \exists i \in \{1, \dots, M\}, \, \mathbb{X}_n \cap S_i \not\subseteq \mathcal{X}_{\pi(i)}), \tag{2.2}$$

where $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_M\}$ is the clustering family selected by the clustering procedure. This quantity is the probability that the clustering procedure does not correctly recover one subset of observations from at least one of the S_i's. The smaller this risk, the better the clustering procedure.

3 A robust single linkage algorithm

Many clustering algorithms have been studied in a context similar to our setting. For instance, Maier et al. (2009), Arias-Castro (2011) and Arias-Castro et al. (2011) prove that algorithms based on pairwise distances (k-nearest neighbor graph, spectral clustering...) are efficient as soon as supports $S_i, i = 1, \ldots, M$ are separated enough. The single linkage hierarchical clustering algorithm has also been investigated by Arias-Castro (2011) and Auray et al. (2015). However, it is well known that this algorithm is sensitive to outliers. We propose here to recall the weaknesses of this algorithm in the presence of outliers and to provide a robust version which allows to improve its performances in this context.

3.1 Agglomerative clustering with single linkage

Many hierarchical clustering algorithms rely on the notion of r–connected set of points in $[0,1]^D$, where r is a nonnegative real number. A subset $A \subseteq \mathbb{R}[0,1]^D$ is said to be r–connected, if

$$B(A, r/2) = \bigcup_{a \in A} B(a, r/2)$$

is a connected set, from a topological point of view. In particular $A = \{x, y\}$ is r–connected if $\|x - y\| \leq r$. The single linkage algorithm may be defined with this notion of connected set of points. For any $r \geq 0$, the set $B(\mathbb{X}_n, r/2)$ can be expanded into $M(r) \in \{1, \ldots, n\}$ connected components denoted by $B_m(r), m = 1 \ldots, M(r)$. These connected components provide a partition of $\mathbb{X}_n$ into $M(r)$ clusters defined by

$$\mathcal{Y}_m(r) = B_m(r) \cap \mathbb{X}_n, \quad m \in \{1, \ldots, M(r)\}.$$

The family $\mathcal{Y}(r) = \{\mathcal{Y}_m(r) \colon m \in \{1, \ldots, M(r)\}\}$ provides clusters of the single linkage algorithm with radius r.

We can observe the number of possible families $\mathcal{Y}(r)$ is finite when we let r move in $\mathbb{R}^+$. Indeed, as r increases the clustering process consists in recursively merging clusters. To see that, consider the single linkage distance between two r-connected components $\mathcal{Y}_m(r)$ and $\mathcal{Y}_{m'}(r)$. It is defined as the distance between the two closest members between these components

$$\text{dist}(\mathcal{Y}_m(r), \mathcal{Y}_{m'}(r)) = \inf\{\|X_k - X_l\| : X_k \in \mathcal{Y}_m(r),\, X_l \in \mathcal{Y}_{m'}(r)\}.$$

At the beginning, for $r = \rho_0 = 0$ we have a first family

$$\mathcal{Y}(\rho_0) = \{\mathcal{Y}_m(\rho_0),\, m \in \{1, \ldots, M(\rho_0)\}\}.$$

Observe that if $X_i \neq X_j$ for all $1 \leq i \neq j \leq n$ then $M(\rho_0) = n$ and each cluster $\mathcal{Y}_m(\rho_0)$ corresponds with one observation. Next the two closest clusters are merged according to the (smallest) distance dist$(\cdot)$. Denote by $\rho_1 > 0$ the distance between the two closest clusters in $\mathcal{Y}(\rho_0)$, we obtain the second family

$$\mathcal{Y}(\rho_1) = \{\mathcal{Y}_m(\rho_1), m \in \{1, \ldots, M(\rho_1)\}\}.$$

This process is then recursively repeated until all (distinct) observations belong to a single cluster. We denote by K the (random) number of iterations, observe that $K \leq n - 1$ almost surely.

Remark 3.1 *Let us make some general remarks about this procedure. At every step k with $1 \leq k \leq K$, the new selected radius ρ_k is larger than the previous radius: $\rho_k > \rho_{k-1}$. This radius corresponds to the distance between the two closest clusters belonging to $\mathcal{Y}(\rho_{k-1})$. Moreover, for any $\rho \in [\rho_k; \rho_{k+1}[$ with $0 \leq k \leq K - 1$ and $\rho_K = \infty$, we have*

$$\mathcal{Y}(\rho) = \mathcal{Y}(\rho_k).$$

At the end of the process, we obtain a sequence $\mathcal{Y}(\rho_0), \ldots, \mathcal{Y}(\rho_{K-1})$ of partitions of the data. The aim is to determine how to choose one partition in this sequence. It remains to select one radius in the sequence $\rho_0, \ldots, \rho_{K-1}$. Since the number of clusters is known, a natural way is to choose the radius such that the associated number of clusters is close to M. More precisely, it is usually chosen such that

$$\widehat{\rho}_{n,SL} \in \underset{\rho \in \{\rho_k :\, k \in \{0,\ldots,K-1\}\}}{\text{argmax}} \{M(\rho) \geq M\}.$$

Observe that $\widehat{\rho}_{n,SL}$ exists as soon as each support S_i contains at least one observation. This algorithm is known to be consistent without outlier (if $\varepsilon = 0$) and under assumptions close to ours (Arias-Castro, 2011, Auray et al., 2015).

3.2 Dealing with outliers

In the presence of outliers, clusters in $\mathcal{Y}(\widehat{\rho}_{n,SL})$ may fail to recover supports S_i with large probability. We provide two toy examples to show that.

Example 1. Figure 1 displays clusters obtained by the classical single linkage algorithm on 2 datasets. The first one (data1) contains 2 groups and these two groups are perfectly identified by the algorithm. For the second one (data2), one outlier has been added and we can observe that this single outlier defines one group while all the other observations are in the second group. The performance is clearly affected by this outlier.

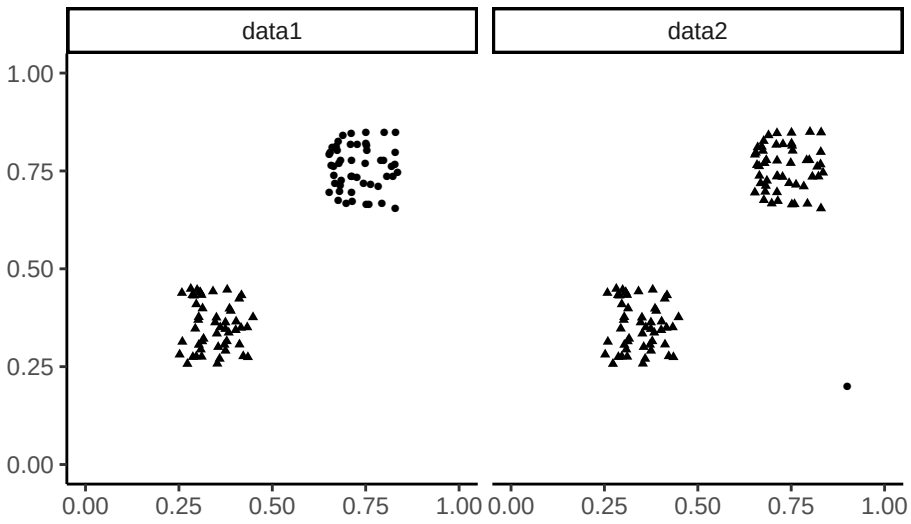


Figure 1: Results of classical single linkage algorithm performed on 2 datasets.

Example 2. We consider data generated according to the following univariate distribution :

$$\mathbb{P} = \frac{1-\varepsilon}{2}(\mathbb{P}_1 + \mathbb{P}_2) + \varepsilon \mathbb{P}_0$$

where $\mathbb{P}_1 = \delta_{-1}$, $\mathbb{P}_2 = \delta_1$, $\mathbb{P}_0 = \mathcal{U}([-3,3])$ and $\varepsilon > 0$. For ε small enough, it is easily seen that assumptions presented in section 2.2 are satisfied. However

simple calculations show that the clustering risk for the the algorithm fails with large probability, more precisely, for any $n > 2$,

$$\mathcal{R}_n(\mathcal{Y}(\widehat{\rho}_{n,SL})) \geq \frac{2}{3} - \frac{8}{3}\frac{1}{(n+1)\varepsilon}.$$

As n increases, the clustering risk tends to $2/3$.

3.3 Robust single linkage clustering

As explained in the last section, the classical single linkage procedure is generally not efficient in presence of outliers. To deal with outliers, Arias-Castro (2011) considers a modified version of this procedure that requires the knowledge of the minimal separation distance δ instead of the knowledge of M. Moreover he assumes that the minimal distance between the outliers and the clusters is at least δ. It is not the case in our model. Since assumptions on our generative model involve that sizes of the clusters should be of the same order, we propose to take account of the cluster sizes in the procedure by considering only the M largest clusters.

Recall that, for a fixed radius $r > 0$, the agglomerative clustering presented at the beginning of Section 3 provides $M(r)$ clusters

$$\mathcal{Y}(r) = \{\mathcal{Y}_m(r), m \in \{1, \ldots, M(r)\}\}.$$

With no loss of generality, we reorder indices of the r-connected components in $\mathcal{Y}(r)$ such that

$$|\mathcal{Y}_1(r)| > |\mathcal{Y}_2(r)| > \ldots > |\mathcal{Y}_{M(r)}(r)|.$$

If $|\mathcal{Y}_m(r)| = |\mathcal{Y}_{m'}(r)|$, then we have a tie. There are several rules for tie breaking, we use the following one: $\mathcal{Y}_m(r)$ is declared "bigger" than $\mathcal{Y}_{m'}(r)$ if

$$\min\{k \in \{1, \ldots, n\} : X_k \in \mathcal{Y}_m(r)\} < \min\{k \in \{1, \ldots, n\} : X_k \in \mathcal{Y}_{m'}(r)\}.$$

This means that tie breaking is done by the smallest index in the cluster.

Our robust single linkage clustering procedure proposes to consider only the M (which is assumed to be known) largest clusters and to merge the other clusters together. Formally, for a fixed value of $r > 0$ we consider the M clusters

$$\mathcal{X}_1(r) = \mathcal{Y}_1(r), \ldots, \mathcal{X}_M(r) = \mathcal{Y}_M(r). \tag{3.1}$$

Other observations, *i.e.* observations that belongs to

$$\mathcal{X}_0(r) = \bigcup_{m=M+1}^{M(r)} \mathcal{Y}_m(r),$$

are not classified and might be considered as outliers. For a fixed value of $r > 0$, this procedure provides the family of clusters $\mathcal{X}(r) = \{\mathcal{X}_1(r), \ldots, \mathcal{X}_M(r)\}$.

Here again, we have to make a safe choice for the radius r. Too small values of r may result in large values $M(r)$. In this case, the M largest clusters could be too small and the clustering procedure may fail to recover all the supports $S_1, \ldots, S_M$. On the opposite, too large values of r may increase:

- the risk to gather observations from different supports $S_i, i = 1, \ldots, M$ in the same cluster;
- and the possibility to obtain clusters defined by outliers.

The algorithm's performance then depends greatly on the choice of r and consequently the radius r has to be defined efficiently based on a data-dependent approach. Here again we use the size of the clusters and we propose to choose the radius in $\{\rho_k \colon k \in \{0, \ldots, K-1\}\}$ which maximizes the size of the M-th cluster:

$$\widehat{r}_n = \max \operatorname*{argmax}_{\rho \in \{\rho_k \colon k \in \{0,\ldots,K-1\}\}} |\mathcal{X}_M(\rho)|. \tag{3.2}$$

Remark 3.2 *The main difference compared to the single linkage clustering is that this algorithm selects the partition which maximizes the size of the M-th cluster. Other observations, i.e. observations in $\mathcal{X}_0(r)$, are not classified and might be considered as outliers.*

4 Main results

The selection procedure (3.2) defines clusters $\mathcal{X}(\widehat{r}_n) = \{\mathcal{X}_m(\widehat{r}_n), m \in \{1, \ldots, M\}\}$ with clustering risk

$$\mathcal{R}_n(\mathcal{X}(\widehat{r}_n)) = \mathbb{P}(\forall \pi \in \Pi_M, \exists i \in \{1, \ldots, M\},\ \mathbb{X}_n \cap S_i \nsubseteq \mathcal{X}_{\pi(i)}(\widehat{r}_n)).$$

The following theorem provides a oracle-type inequality which ensures that this clustering risk is closed to the optimal clustering risk, *i.e.*, the one achieved with the best value of r.

Theorem 4.1 *Assume* **(A1)** *and* **(A6)** *hold. Let* $\eta_0 > 0$ *and* $\eta_1 > 0$ *be such that*

$$\eta_0 = 1 - [(1-\varepsilon)(1+\not\gamma)]^{-1} \quad \text{and} \quad \frac{4\eta_1}{1-\eta_1} = \frac{\gamma_*}{\gamma^*} - \frac{1}{2}.$$

Then for all $\eta \leq \min(\eta_0, \eta_1)$ *and all* $n \geq M$*, the clustering risk for clusters* $\mathcal{X}(\widehat{r}_n)$ *satisfies*

$$\mathcal{R}_n(\mathcal{X}(\widehat{r}_n)) \leq \inf_{r>0} \mathcal{R}_n(\mathcal{X}(r)) + 2M \exp(-\psi(\eta)(1-\varepsilon)\not\gamma\, n) \tag{4.1}$$

where for $\eta > 0$ $\psi(\eta) = (1+\eta)(\log(1+\eta) - 1) + 1 > 0$.

This theorem ensures that the proposed data-driven method (3.2) is efficient for n large enough. Indeed, since $\psi(\eta)(1-\varepsilon)\not\gamma > 0$, inequality (4.1) says that the performance of the proposed procedure is optimal, up to a remainder term which tends to zero at an exponentiel rate. In particular, if there exits a specific value r_n of r such that the clustering risk of $\mathcal{X}(r_n)$ tends to 0 as n increases, Theorem 4.1 implies that the risk of $\mathcal{X}(\widehat{r}_n)$ also tends to 0.

To study the clustering risk of $\mathcal{X}(r)$ for a given value of $r > 0$, we need the following parameters:

$$\mathfrak{a} = \frac{\gamma_*(\kappa^*\kappa_c)^{-1}\eta_*(d)}{1+\not\gamma}, \quad \mathfrak{b} = \eta(D)\kappa_0 \tag{4.2}$$

where

$$\kappa^* = \max_{i\in\{1,\ldots,M\}} \kappa_i \quad \text{and} \quad \eta_*(d) = \min_{0\leq s\leq d} \eta(s) = \min(1, \eta(d)). \tag{4.3}$$

Parameters $\mathfrak{a}$ and $\mathfrak{b}$ measure to some extent the complexity of the problem. Indeed, $\mathfrak{b}$ essentially depends on the density of the outliers through the parameter κ_0. Problems with sparse outliers will correspond to a small value of $\mathfrak{b}$. The second parameter $\mathfrak{a}$ is related to the distribution of the actual data in their supports $S_i, i \in \{1, \ldots, M\}$ and on the regularity of these supports. Regular supports (κ_c small) with a large density of observations (κ^* small) lead to large values of $\mathfrak{a}$. To summarize, difficult problems match with small $\mathfrak{a}$ and/or large $\mathfrak{b}$.

The following theorem controls the clustering risk of $\mathcal{X}(r)$ in terms of the parameters of the model.

Theorem 4.2 *Under assumptions* **(A1)**–**(A6)** *we have, for any* $0 < r < \min(\min_i \Delta_i, \delta)$ *and for any* η *such that* $0 < \eta < \eta_0$

$$\mathcal{R}_n(\mathcal{X}(r)) \leq \Lambda r^{-d} \exp(-\mathfrak{a} n r^d) + n\varepsilon(\mathfrak{b}\varepsilon n r^D)^{\lfloor \frac{\delta}{r} \rfloor} + 2M \exp(-\psi(\eta)(1-\varepsilon) \curlywedge n), \tag{4.4}$$

where Λ *is a positive constant specified in the proof of the Theorem.*

The upper bound in (4.4) is governed by the two first terms since the last term generally tends to zero much faster. Recall that the cluster family $\mathcal{X}(r)$ may fail to recover the true clusters if one of these two conditions is satisfied:

1. Observations in a same support are not r-connected: there exists $i \in \{1, \ldots, M\}$ such that $\mathbb{X}_n \cap S_i$ is not r-connected;
2. Some observations that belong to different supports are r-connected: there is a r-connected path between S_i and S_j for $(i, j) \in \{1, \ldots, M\}^2$ with $i \neq j$.

The first term in the right hand side of (4.4) corresponds to the first conditions. Unsurprisingly, this term is small for large values of r and/or $\mathfrak{a}$. The second term is related to the second condition and, in contrast with the first term, it tends to decrease when r decreases. This second term also depends on the distribution of the outliers. We observe that it is equal to zero without outlier and it increases as the outlier parameter $\mathfrak{b}$ and/or the proportion of outlier ε grows.

Observe also that the minimal distance δ between supports occurs through the exponent $\lfloor \delta/r \rfloor$. If $\mathfrak{b}\varepsilon n r^D < 1$, the second error term decreases as $\lfloor \delta/r \rfloor$ increases. Moreover, we can remark that the upper bound involves two dimensions: the (maximal) Hausdorff dimension d of the support S_i and the dimension D of the outlier space S_0. For fixed values of D, we could obtain slower rates as d increases because it is more difficult to connect observations for large values of d. On the opposite, keeping d constant, rates could be faster when D grows because the probability to connect observations in S_0 decreases. Combining Theorems 4.1 and 4.2 we obtain:

Theorem 4.3 *Under assumptions* **(A1)**–**(A6)** *we have, for all* $n \geq M$*, for any* $0 < r < \min(\min_i \Delta_i, \delta)$ *and all* $\eta \leq \min(\eta_0, \eta_1)$*:*

$$R_n(\mathcal{X}(\widehat{r}_n)) \leq \inf_{r>0} \left\{ \Delta r^{-d} \exp(-\mathfrak{a} n r^d) + n\varepsilon(\mathfrak{b}\varepsilon n r^D)^{\lfloor \frac{\delta}{r} \rfloor} \right\} + 4M \exp(-\psi(\eta)(1-\varepsilon) \curlywedge n).$$

If we intend to prove any consistency results regarding $\mathcal{R}_n(\mathcal{X}(\widehat{r}_n))$, we have to exhibit at least one value of r such that the first terms in this upper bound tends to zero. The following corollary provides sufficient conditions for the consistency of the clustering procedure, *i.e.*, conditions for which we have

$$\lim_{n\to+\infty} \mathcal{R}_n(\mathcal{X}(\widehat{r}_n)) = 0. \tag{4.5}$$

Except for D and d, all parameters (δ, κ^*, ε...) may vary with n. For simplicity, we only let ε vary with n and keep all other parameters fixed in the conditions.

Corollary 4.1 *Under the assumptions of Theorem 4.3, consistency (4.5) holds if either $d < D$ or $D = d$ and $\varepsilon < (\mathfrak{b}\log n)^{-1}$.*

We obtain this result by taking $r^d = D\log(n)/(\mathfrak{a}dn)$. This corollary ensures that consistency holds as soon as the Hausdorff dimension of the supports $S_i, i = 1,\dots,M$ is smaller than the dimension D of the ambient space. When these dimensions match, the proportion of outliers should tend to 0 much faster than $1/\log n$. Observe also that without outlier ($\varepsilon = 0$), convergence occurs for all $d \leq D$. Using similar tools, we can obtain many rates on convergence with respect to the proportion ε of outliers. Some examples are gathered in the following corollary.

Corollary 4.2 *Under the assumptions of Theorem 4.3, there exists two universal constants C_1 and C_2 such that the following propositions hold:*

1. *Few outliers: if $\varepsilon = \exp(-n)$ then:*

$$R_n(\mathcal{X}(\widehat{r}_n)) \leq C_1 n\exp(-C_2 n).$$

2. *Small dimensions of the supports: If $D > d+1$ then*

$$R_n(\mathcal{X}(\widehat{r}_n)) \leq C_1 n\exp(-C_2 n^{1/(d+1)}).$$

3. *Large dimensions of the supports with few outliers: if $d \leq D \leq d+1$ and $\varepsilon = n^{-\beta}$ with $\beta \geq 1 - D/(d+1)$, then*

$$R_n(\mathcal{X}(\widehat{r}_n)) \leq C_1 n^{d/(d+1)}\exp(-C_2 n^{1/(d+1)}).$$

4. *Large dimensions of the supports with many outliers: if* $d \leq D \leq d+1$ *and* $\varepsilon = n^{-\beta}$ *with* $0 < \beta < 1 - D/(d+1)$*, then*

$$R_n(\mathcal{X}(\widehat{r}_n)) \leq C_1 n^{(1-\beta)d/D} \exp(-C_2 n^{1+(\beta-1)d/D}).$$

To summarize, in each of the above situations we obtain an upper bound of the form

$$R_n(\mathcal{X}(\widehat{r}_n)) \leq C_1 n^A \exp\left(-C_2 n^B\right)$$

where A, B are given positive constants that depend on the complexity of the problem. Case 1 exhibit a fast rate when the number of outliers is small. When we have a gap between the ambient dimension D and the dimension of the support of the actual data d (case 2), we observe that the rate is governed by d in the exponential term. Finally, when d gets closer to D and the outlier rate ε is polynomial (cases 3 and 4), only the polynomial term in the bound changes in terms of ε. The rate is better when β is greater than $1 - D/(d+1)$, we obtain slower rates for smaller values of β.

5 Numerical experiments

Here, the proposed clustering algorithm is assessed through several simulation studies. After a description of the models used, the performances of the robust single linkage algorithm (RSL) are assessed and compared with the performances of three classical agglomerative clustering methods, namely the single linkage algorithm (SL), the k-means algorithm (KMeans) and the spectral clustering (SC).

All implementations have been performed in `R` and source code is available at https://github.com/klutchnikoff/robustSL. RSL and SC have been implemented by using the functions `hclust` (package `fastcluster`) and `cutree` (package `stats`). The function `kmeans` (package `stats`) was used to perform KMeans with 20 different random starts (parameter `nstart`). SC has been implemented following Ng et al. (2002) and using the function `specc` (package `kernlab`). The scaling parameter is set to the optimal value provided by `specc` and we consider 20 different random starts for the k-means step of the algorithm. Observe that the four clustering approaches used here require the number M of groups.

The computations have been carried out on a MacBook Pro, 2,4 GHz Intel Core i5 and 16Gb of RAM memory.

5.1 Description of the models

In this study we evaluate the performances of our method on two-dimensional data simulated according to the design introduced in Section 2.1. As stated in Section 4, the performances of the proposed clustering algorithm depend on the complexity of the clustering problem, measured through the parameters $(\mathfrak{a}, \mathfrak{b}, \delta, \epsilon)$. Among them, the most sensitive are the inter-group distance δ and the proportion of outliers ε. Of course, the larger ε and the smaller δ, the more difficult the problem. In the experiments, we consider different values for these two parameters and we use the three different models described below (respectively called the "squares", "concentric-circles" and "sine" models) to simulate data. More precisely, for each model, we consider

- two values of δ, one "small value" which corresponds to an "easy" case and one "large" value for a "tricky" case;
- five values of ε, equally spaced between 0 and 0.2 with a step of 0.05;
- two different sample sizes: $n = 200$ and $n = 500$.

For each model, both groups and outliers are uniformly sampled over their supports $S_i, i = 1, \ldots, M$ and S_0. We now describe the three models.

Squares model. Data are grouped in three distinct squares with similar areas. We use the same weights for each group. Easy and tricky cases correspond to inter-group distances 0.35 and 0.07 respectively, see Figure 2.

Concentric circles. This model consists of 2 groups which correspond to two nested rings with weight 0.4 for the smallest ring and 0.6 for the largest ring. Inter-group distances are fixed to 2.6 (easy) and 1.6 (tricky), see Figure 3. Outliers are generated only between the two rings.

Sine model. This model includes 3 groups with various shapes. The first group is tight and represents the sine curve while the two others are compact groups (squares). We use the same weights for each group. Easy and tricky cases correspond to inter-group distances $\sqrt{(\frac{\pi}{2} - \frac{1}{2})^2 + \frac{1}{4}}$ $(\approx 1,18)$ and $\sqrt{(\frac{\pi}{4} - \frac{1}{2})^2 + \frac{1}{2}}$ $(\approx 0,76)$ respectively, see Figure 4.

Simple calculations show that these models satisfy assumptions **(A1)**-**(A6)** when $\varepsilon < 1/11$ for the "concentric circles" model and $\varepsilon < 1/7$ for the two others. We can also remark that the Hausdorff dimension for all supports equals 2, excepted for the sine group where it equals to 1.

5.2 Performances of the robust single linkage clustering

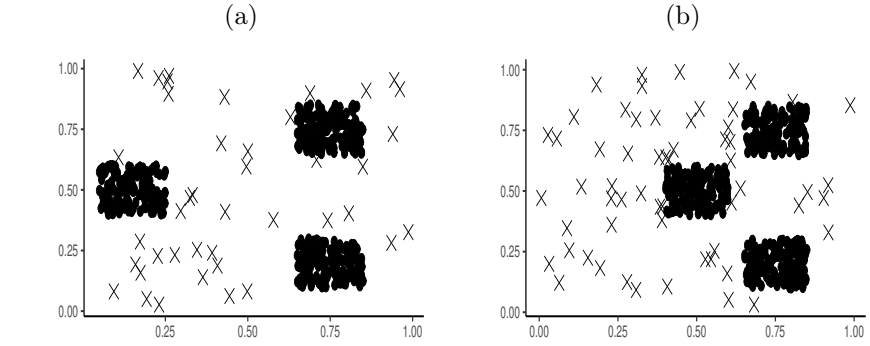
Through various numerical experiments based on the scenarios described in the previous section, the performances of RSL are assessed and compared to those of SL, KMeans and SC, which are part of the classical approaches used in clustering.

In each scenario, the clustering risk (2.2) of each clustering algorithm is approximated by its empirical estimator computed over 1000 Monte Carlo replications

$$\frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{\left\{\forall\, \pi\in\Pi_M\,,\, \exists i=1,\dots,M,\, \mathbb{X}_n^b\cap S_i\not\subseteq \mathcal{X}^b_{\pi(i)}\right\}}, \tag{5.1}$$

where $\mathcal{X}^b = \{\mathcal{X}_1^b, \dots, \mathcal{X}_M^b\}$ denotes the clusters obtained by the procedure on the b-th Monte Carlo sample of data $\mathbb{X}_n^b = \{X_1^b, \dots, X_n^b\}$. Estimator (5.1) corresponds to the proportion of Monte Carlo replications in which the clustering procedure does not correctly recover one subset of observations from at least one of the S_i's.

Figures 2-4 display the three models and the empirical estimate (5.1) of the clustering risk according to ε, n and δ for the four considered algorithms in each model.



	$n = 200$		$n = 500$	
	$\varepsilon = 0$	$\varepsilon = 0.2$	$\varepsilon = 0$	$\varepsilon = 0.2$
	Case (a): δ large			
RSL	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
SL	0.000 (0.000)	0.958 (0.006)	0.000 (0.000)	0.997 (0.002))
KMeans	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
SC	0.000 (0.000)	0.162 (0.012)	0.000 (0.000)	0.145 (0.011)
	Case (b): δ small			
RSL	0.000 (0.000)	0.014 (0.004)	0.000 (0.000)	0.002 (0.001)
SL	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	1.000 (0.000)
KMeans	0.000 (0.000)	0.022 (0.005)	0.000 (0.000)	0.013 (0.004)
SC	0.000 (0.000)	0.604 (0.015)	0.000 (0.000)	0.469 (0.016)

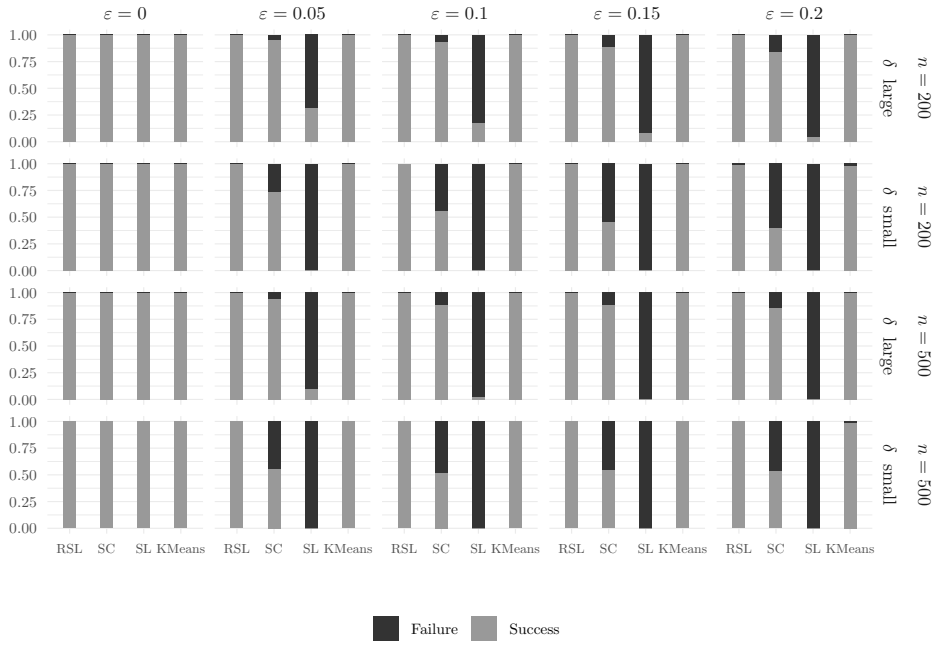
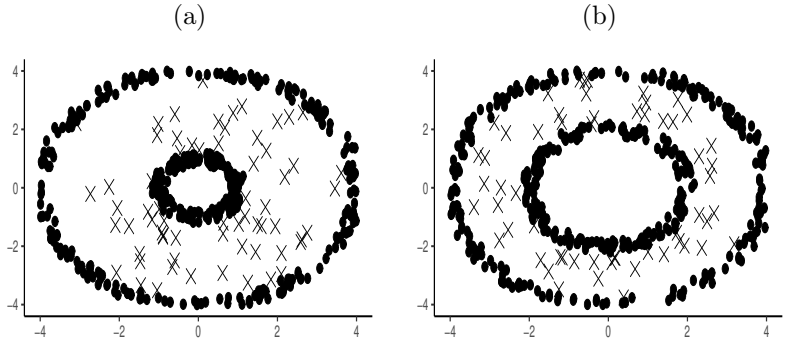


Figure 2: Results in *squares* model. From top to bottom : a sample of $n = 500$ observations with $\varepsilon = 0.1$ and (a) $\delta = 0.35$ (easy) and (b) $\delta \approx 0.07$ (tricky); a table and a barplot displaying the empirical estimate of the clustering risk according to ε, n and δ.



	$n = 200$		$n = 500$	
	$\varepsilon = 0$	$\varepsilon = 0.2$	$\varepsilon = 0$	$\varepsilon = 0.2$
	Case (a): δ large			
RSL	0.001 (0.001)	0.863 (0.011)	0.000 (0.000)	0.342 (0.015)
SL	0.000 (0.000)	0.973 (0.005)	0.000 (0.000)	0.993 (0.003)
KMeans	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
SC	0.000 (0.000)	0.744 (0.014)	0.000 (0.000)	0.192 (0.012)
	Case (b): δ small			
RSL	0.001 (0.001)	0.997 (0.002)	0.000 (0.000)	0.909 (0.009)
SL	0.000 (0.000)	0.999 (0.001)	0.000 (0.000)	1.000 (0.000)
KMeans	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
SC	0.001 (0.001)	0.996 (0.002)	0.000 (0.000)	0.863 (0.011)

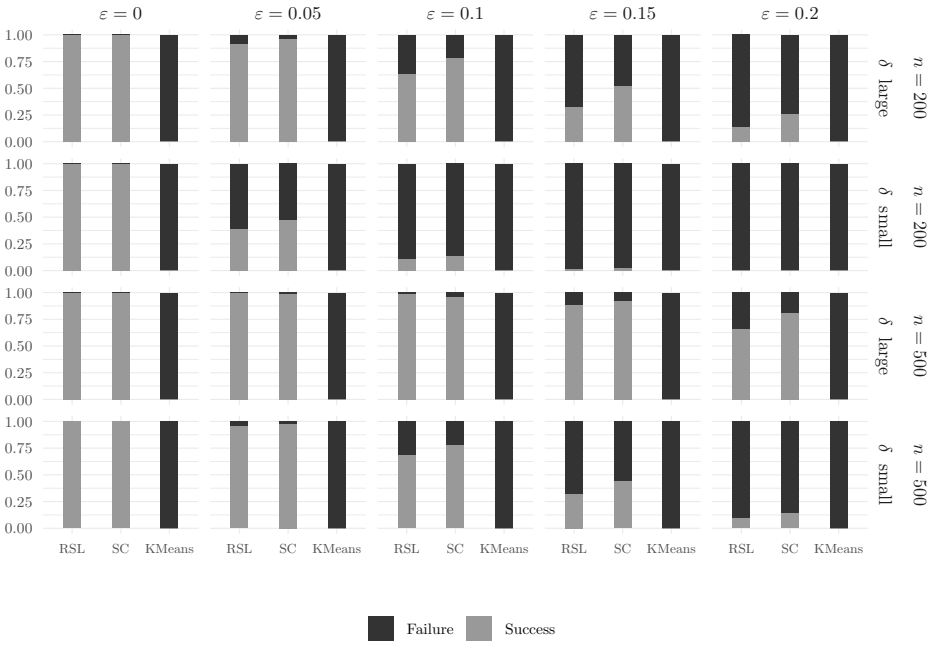
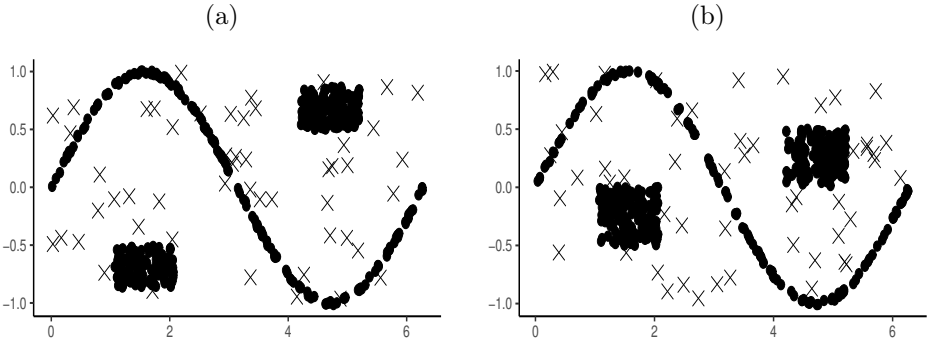


Figure 3: Results in *concentric circles* model. From top to bottom : a sample of $n = 500$ observations with $\varepsilon = 0.1$ and (a) $\delta = 2, 6$ (easy) and (b) $\delta = 1, 6$ (tricky); a table and a barplot displaying the empirical estimate of the clustering risk according to ε, n and δ.



	$n = 200$		$n = 500$	
	$\varepsilon = 0$	$\varepsilon = 0.2$	$\varepsilon = 0$	$\varepsilon = 0.2$
	Case (a): δ large			
RSL	0.001 (0.001)	0.796 (0.013)	0.000 (0.000)	0.307 (0.015)
SL	0.001 (0.001)	0.979 (0.005)	0.000 (0.000)	1.000 (0.000)
KMeans	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
SC	0.009 (0.003)	0.621 (0.015)	0.000 (0.000)	0.208 (0.013)
	Case (b): δ small			
RSL	0.058 (0.007)	0.939 (0.008)	0.000 (0.000)	0.548 (0.016)
SL	0.058 (0.007)	0.994 (0.002)	0.000 (0.000)	1.000 (0.000)
KMeans	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
SC	0.077 (0.008)	0.899 (0.010)	0.002 (0.001)	0.502 (0.016)

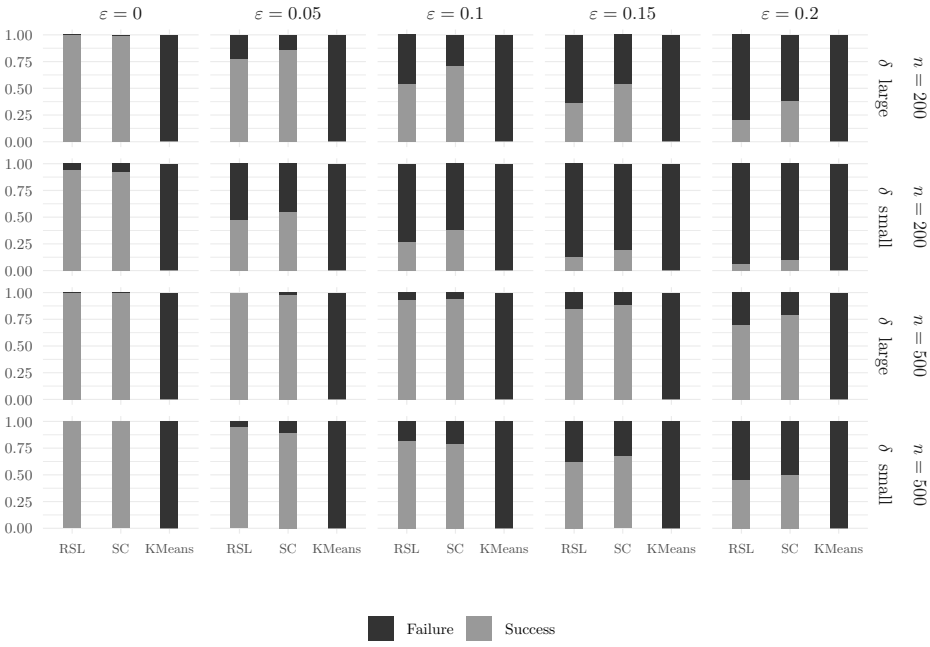


Figure 4: Results in *sine* model. From top to bottom : a sample of $n = 500$ observations with $\varepsilon = 0.1$ and (a) $\delta \approx 1,18$ (easy) and (b) $\delta \approx 0,76$ (tricky); a table and a barplot displaying the empirical estimate of the clustering risk according to ε, n and δ.

First of all, as expected, the estimated clustering risk behaves as an increasing function of ε in all experiments and for all clustering algorithms. Moreover, for a fixed value of ε, the risk of clustering of all methods is higher when the number n of observations and the inter-group distance, δ, are smaller. In all experiments, when there is no outlier (i.e. $\varepsilon = 0$), the clustering risk of both SL and RSL is roughly zero. This result agrees with Theorem 4.1, Arias-Castro (2011) and Auray et al. (2015) which prove that under assumptions close to **(A1)-(A6)** and when $\varepsilon = 0$, SL is consistent and its clustering risk tends quickly to zero. As discussed in Section 3.2, in presence of outliers SL often fails to recover the true clusters and its clustering risk increases quickly with ε. On the contrary, RSL seems less sensitive to the outlier proportion ε. In the numerical experiments, RSL always works better than SL when there are some outliers.

As expected, we can see that KMeans works well with compact groups (see for instance results for the *squares* model) but often fails with non-linearly separable groups (see for instance results for the *sine* model and the *concentric-circles* model) whereas SC is well adapted to non-linearly separable groups but can work quite badly with compact groups (Nadler and Galun, 2007). Contrary to KMeans and SC, RSL is part of the two best methods in each experiment, and so it seems to perform well whatever the shape of the groups. Moreover, observe that compared to KMeans and SC, RSL as well as SL are exact, in the sense that they do not require any random process (for instance the random starts used in KMeans and SC).

Finally, RSL and SC are also compared in terms of time efficiency. Table 1 displays the time required by each algortihm in the *squares* model for various values of n. For large n, RSL is much more faster than SC. Moreover when n increases, time complexity increases much less quicker for RSL than fo SC.

n	RSL	SC
100	1.41	0.10
200	1.56	0.38
500	2.09	3.86
1000	3.37	27.80
2000	3.27	2×10^3
5000	3.94	4×10^4
10000	6.68	3×10^5

Table 1: Time complexity (in second) of RSL and SC for various values of n in the *squares* model.

6 Proofs

6.1 Technical lemmas

Lemma 6.1 *Fix $i = 1, \dots, M$ and $0 < r < \Delta_i$. Under* **(A1)-(A3)-(A4)**, *there exists a positive constant Λ_i such that*

$$\psi_{n,i}(r) = \mathbb{P}(\mathbb{X}_n \cap S_i \text{ is not } r\text{-connected}) \leq \Lambda_i r^{-d} \exp(-\mathfrak{a} n r^d).$$

Proof. We denote by $N_r^*(S_i)$ the minimal number of balls of radius $r > 0$, centered at points of S_i, required to cover S_i. Note that, for any $r > 0$ we have by definition $N_r(S_i) \leq N_r^*(S_i)$. Moreover, using triangle inequality we also have $N_r^*(S_i) \leq N_{r/2}(S_i)$. Using Assumption **(A5)**, this implies that

$$s_i = \lim_{r \to 0} \frac{\log(N_r^*(S_i))}{\log(1/r)}.$$

Thus, there exists a positive constant Λ_i (that depends on S_i) such that, for any $0 < r < \Delta_i$ we have:

$$N_r^*(S_i) \leq 4^{-d} \Lambda_i r^{-s_i} \leq 4^{-d} \Lambda_i r^{-d}.$$

Thus, there exists an index set $\mathcal{L}_i$, whose cardinality is bounded above by $\Lambda_i r^{-d}$, and a family of balls $(B_\ell)_{\ell \in \mathcal{L}_i}$ centered at a points that belong to S_i, with radius $r/4$, which satisfy

$$S_i \subset \bigcup_{\ell \in \mathcal{L}_i} B_\ell.$$

Since, for any $\ell \in \mathcal{L}_i$, we have $(\mathbb{X}_n \cap S_i) \cap B_\ell \neq \emptyset$, there exists $\alpha_\ell \in \{1, \dots, n\}$ such that $X_{\alpha_\ell} \in B_\ell \cap S_i$. Using triangle inequality, this implies that $B(X_{\alpha_\ell}, r/2) \supset B_\ell$. Thus

$$\mathbb{X}_n \cap S_i \subset \bigcup_{\ell \in \mathcal{L}_i} B(X_{\alpha_\ell}, r/2) \quad \text{with} \quad X_{\alpha_\ell} \in \mathbb{X}_n \cap S_i.$$

Thus, $\mathbb{X}_n \cap S_i$ is r-connected. This implies that,

$$\begin{aligned}
\psi_{n,i}(r) &\leq \mathbb{P}\left(\exists \ell \in \mathcal{L}_i, B_\ell \cap (\mathbb{X}_n \cap S_i) = \emptyset\right) \\
&\leq \mathbb{P}\left(\exists \ell \in \mathcal{L}_i,\ \forall k \in \{1, \dots, n\}, X_k \notin S_i \text{ or } (X_k \in S_i, X_k \notin B_\ell)\right) \\
&\leq \sum_{\ell \in \mathcal{L}_i} \left(\mathbb{P}(X \notin S_i) + \mathbb{P}(X \notin B_\ell \mid X \in S_i)\mathbb{P}(X \in S_i)\right)^n \\
&\leq \sum_{\ell \in \mathcal{L}_i} \left(1 - \mathbb{P}(X \in B_\ell \mid X \in S_i)\mathbb{P}(X \in S_i)\right)^n \\
&\leq \sum_{\ell \in \mathcal{L}_i} \left(1 - (1 - \varepsilon)\gamma_i \mathbb{P}_i(X \in B_\ell)\right)^n
\end{aligned}$$

Moreover

$$\begin{aligned}
\mathbb{P}_i(X \in B_\ell) &= \mathbb{P}_i(X \in B_\ell \cap S_i) && \text{from } \mathbf{(A1)} \\
&\geq \kappa_i^{-1} \mathcal{H}^{s_i}(B_\ell \cap S_i) && \text{from } \mathbf{(A3)} \\
&\geq (\kappa_i \kappa_c)^{-1} \eta(s_i) r^{s_i} && \text{from } \mathbf{(A4)} \\
&\geq (\kappa^* \kappa_c)^{-1} \eta_*(d) r^d,
\end{aligned}$$

where κ^* and $\eta_*(d)$ are defined by (4.3). Putting all pieces together we obtain

$$\begin{aligned}
\psi_{n,i}(r) &\leq |\mathcal{L}_i| \big(1 - (1-\varepsilon)\gamma_*(\kappa^* \kappa_c)^{-1} \eta_*(d) r^d\big)^n \\
&\leq \Lambda_i r^{-d} \exp(-\mathfrak{a} n r^d),
\end{aligned}$$

where $\mathfrak{a}$ is defined in (4.2).

Lemma 6.2 *Let $r > 0$ and denote by $\varphi_n(m, r)$ the probability that there exists, in S_0, a path of at least m r-connected observations. If assumptions* **(A1)** *and* **(A2)** *hold, we have*

$$\varphi_n(m, r) \leq n\varepsilon(\mathfrak{b} n \varepsilon r^D)^{m-1}$$

where $\mathfrak{b}$ is defined in (4.2).

Proof. Fix $r > 0$. For any $I \subseteq \{1, \ldots, n\}$ we denote by $\mathcal{A}_I$ the following event: there exists a permutation $i_1 < \ldots < i_m$ of I such that $\|X_{i_j} - X_{i_{j+1}}\| \leq r$ for any $j = 1, \ldots, m-1$. We have:

$$\begin{aligned}
\varphi_n(m, r) &\leq \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ |I| = m}} \mathbb{P}(\mathcal{A}_I \cap \{X_I \subseteq S_0\}) \\
&\leq \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ |I| = m}} \varepsilon^m \mathbb{P}_0(\mathcal{A}_I).
\end{aligned}$$

Now remark that

$$\begin{aligned}
\mathbb{P}_0(\mathcal{A}_I) &\leq m! \mathbb{E}_0 \left(\prod_{j=1}^{m-1} \mathbf{1}_{[0,r]}(\|X_{i_j} - X_{i_{j+1}}\|) \right) \\
&= m! \int_{S_0} \cdots \int_{S_0} \mathbf{1}_{[0,r]}(\|x_1 - x_2\|) \ldots \mathbf{1}_{[0,r]}(\|x_{m-1} - x_m\|) \mathrm{d}\mathbb{P}_0(x_1, \ldots, x_m).
\end{aligned}$$

Note also that, using **(A2)**:

$$\int_{S_0} \mathbf{1}_{[0,r]}(\|x - y\|) \mathrm{d}\mathbb{P}_0(y) \leq \mathbb{P}_0(B(x, r)) \leq \kappa_0 \mathcal{H}^D(B(x, r)) = \kappa_0 \eta(D) r^D.$$

This, combined with Fubini's theorem implies that:

$$\mathbb{P}_0(\mathcal{A}_I) \leq m!(\eta(D)\kappa_0 r^D)^{m-1}.$$

Finally, we obtain:

$$\begin{aligned}\varphi_n(m,r) &\leq \frac{n!}{(n-m)!}\varepsilon^m(\eta(D)\kappa_0 r^D)^{m-1} \\ &\leq n\varepsilon(\mathfrak{b}\varepsilon n r^D)^{m-1}.\end{aligned}$$

Lemma 6.3 *Assume that assumptions* **(A1)** *and* **(A6)** *hold. Define* $N_i = |\mathbb{X}_n \cap S_i|, i \in \{0,\ldots,M\}$ *and for* $0 < \eta \leq \eta_0$ *let*

$$\Omega_\eta = \bigcap_{i=1}^{M} \{(1-\eta)(1-\varepsilon)\gamma_i n < N_i < (1+\eta)(1-\varepsilon)\gamma_i n\}.$$

We have

(i) $\mathbb{P}(\overline{\Omega_\eta}) \leq 2M\exp(-\psi(\eta)(1-\varepsilon)\,\gamma_+ n)$;

(ii) $N_0 < \frac{\gamma_+}{\gamma_*}\min_{i\in\{1,\ldots,M\}} N_i < \min_{i\in\{1,\ldots,M\}} N_i$ *under* Ω_η.

Proof. Since $N_i \sim B(n,(1-\varepsilon)\gamma_i)$, (i) is a direct consequence of (Shorack and Wellner, 1986, page 440). For (ii), observe that $(1-\varepsilon)(1-\eta_0) = 1/(1+\gamma_+)$. Since $0 < \eta \leq \eta_0$, it follows that

$$1-(1-\varepsilon)(1-\eta) \leq (1-\varepsilon)(1-\eta)\,\gamma_+.$$

Thus, under Ω_η,

$$\begin{aligned}N_0 &\leq n - \sum_{i=1}^{M} N_i \leq n\left(1-(1-\varepsilon)(1-\eta)\sum_{i=1}^{M}\gamma_i\right) \\ &\leq n(1-(1-\varepsilon)(1-\eta)) \leq n(1-\varepsilon)(1-\eta)\,\gamma_+ \\ &\leq \frac{\gamma_+}{\gamma_*}n(1-\varepsilon)(1-\eta)\gamma_i < \frac{\gamma_+}{\gamma_*}N_i \quad \forall i = 1,\ldots,M.\end{aligned}$$

Lemma 6.4 *Assume that assumption* **(A6)** *holds. For each* $\eta \leq \min(\eta_0,\eta_1)$ *we have*

$$\frac{1+\eta}{1-\eta}\frac{\gamma^*}{\gamma_*} + \frac{\gamma_+}{\gamma_*} \leq 2.$$

Proof. Let $\eta \leq \min(\eta_0, \eta_1)$, then

$$\begin{aligned}
\frac{1+\eta}{1-\eta}\frac{\gamma^*}{\gamma_*} + \frac{\gamma}{\gamma_*} &= \frac{1-\eta+2\eta}{1-\eta}\frac{\gamma^*}{\gamma_*} + \frac{\gamma}{\gamma_*} \\
&= \frac{\gamma * + \gamma}{\gamma_*} + \frac{1}{2}\frac{4\eta}{1-\eta}\frac{\gamma^*}{\gamma_*} \\
&\leq \frac{\gamma^*/2 + \gamma_*}{\gamma_*} + \frac{1}{2}\left(\frac{\gamma_*}{\gamma^*} - \frac{1}{2}\right)\frac{\gamma^*}{\gamma_*} \\
&= \frac{3}{2} + \frac{\gamma^*}{4\gamma_*} \leq 2.
\end{aligned}$$

Lemma 6.5 *Assume* **(A6)** *holds. For $r > 0$, let*

$$\mathcal{E}(r) = \left\{\exists \pi \in \Pi_M \, \forall i = 1, \ldots, M \;\; \mathbb{X}_n \cap S_i \subseteq \mathcal{X}_{\pi(i)}(r)\right\}.$$

Let $\eta \leq \min(\eta_0, \eta_1)$, then under $\mathcal{E}(r) \cap \Omega_\eta$ we have

1. *$\widehat{r}_n \geq r$ almost surely;*
2. *There exists $\pi \in \Pi_M$ such that, $\forall i = 1, \ldots, M$ $\mathcal{X}_i(r) \subseteq \mathcal{X}_{\pi(i)}(\widehat{r}_n)$.*

Proof. Let $\eta \leq \min(\eta_0, \eta_1)$ and assume that Ω_η is true. We first prove that $\widehat{r}_n \geq r$ with a reductio ad absurdum. Assume that $\widehat{r}_n < r$. Observe that

$$|\mathcal{Y}_M(\widehat{r}_n)| > |\mathcal{Y}_M(r)|$$

by definition of $\widehat{r}_n$. It follows that

$$|\mathcal{Y}_1(\widehat{r}_n)| \geq \ldots \geq |\mathcal{Y}_M(\widehat{r}_n)| > |\mathcal{Y}_M(r)|.$$

Since $\widehat{r}_n < r$, we deduce that one of the $\mathcal{Y}_i(r), i = 1, \ldots, M-1$ contains observations of at least two clusters among $\mathcal{Y}_i(\widehat{r}_n), i = 1, \ldots, M$. It implies that

$$|\mathcal{Y}_1(r)| \geq 2\,|\mathcal{Y}_M(\widehat{r}_n)| > 2\,|\mathcal{Y}_M(r)|. \tag{6.1}$$

Moreover, under $\mathcal{E}(r)$ we have $N_{(1)} \leq |\mathcal{Y}_1(r)| \leq N_{(1)} + N_0$ where $N_i = |\mathbb{X}_n \cap S_i|$ and $N_{(i)}, i = 1, \ldots, M$ are such that

$$N_{(1)} \geq \ldots \geq N_{(M)}.$$

Thus, under $\mathcal{E}(r) \cap \Omega_\eta$, we have from Lemma 6.3

$$|\mathcal{Y}_1(r)| \leq N_{(1)} + N_0 \leq N_{(1)} + \frac{\gamma}{\gamma_*} N_{(M)}.$$

Since $\left|\mathcal{Y}_{(M)}(r)\right| \geq N_M$, we obtain from Lemma 6.4

$$\frac{|\mathcal{Y}_1(r)|}{|\mathcal{Y}_M(r)|} \leq \frac{N_{(1)}}{N_{(M)}} + \frac{\gamma}{\gamma_*} \leq \frac{1+\eta}{1-\eta}\frac{\gamma^*}{\gamma_*} + \frac{\gamma}{\gamma_*} \leq 2$$

which is a contradiction with (6.1). We deduce that $\widehat{r}_n \geq r$ almost surely.

For the second point, observe that since $\widehat{r}_n \geq r$, each $\mathcal{X}_i(\widehat{r}_n), i = 1, \ldots, M$ may be written as an union of clusters in

$$\mathcal{X}_1(r), \ldots, \mathcal{X}_M(r), \mathcal{Y}_{M+1}(r), \ldots, \mathcal{Y}_{M(r)}(r).$$

Moreover for each $i \in \{1, \ldots, M\}$ there exists an unique $j \in \{1, \ldots, M\}$ and a subset $\mathcal{T}(\widehat{r}_n)$ of $\{M+1, \ldots, M(r)\}$ such that

$$\mathcal{X}_i(\widehat{r}_n) = \mathcal{X}_j(r) + \bigcup_{\ell \in \mathcal{T}(\widehat{r}_n)} \mathcal{Y}_\ell(r). \tag{6.2}$$

Indeed if there exists $i \in \{1, \ldots, M\}$ and $1 \leq j \neq j' \leq M$ such that

$$\mathcal{X}_j(r) \cup \mathcal{X}_{j'}(r) \subseteq \mathcal{X}_i(\widehat{r}_n)$$

then $\mathcal{X}_M(\widehat{r}_n)$ may be written as an union of clusters in

$$\{\mathcal{Y}_{M+1}(r), \ldots, \mathcal{Y}_{M(r)}(r)\},$$

and thus $|\mathcal{X}_M(\widehat{r}_n)| \leq N_0$. This is not possible since, by definition of $\widehat{r}_n$ and by Lemma 6.3, we have

$$|\mathcal{X}_M(\widehat{r}_n)| \geq |\mathcal{X}_M(r)| \geq N_{(M)} > N_0.$$

We deduce that (6.2) is true. Therefore there exists $\pi \in \Pi_M$ such that, $\forall i = 1, \ldots, M$ $\mathcal{X}_i(r) \subseteq \mathcal{X}_{\pi(i)}(\widehat{r}_n)$.

6.2 Proof of Theorem 4.1

For $r > 0$, let

$$\mathcal{E}(r) = \left\{\exists \pi \in \Pi_M, \forall i = 1, \ldots, M, \ \mathbb{X}_n \cap S_i \subseteq \mathcal{X}_{\pi(i)}(r)\right\} \tag{6.3}$$

Let $\eta \leq \min(\eta_0, \eta_1)$. Observe that

$$\begin{aligned} 1 - \mathcal{R}_n(\mathcal{X}(\widehat{r}_n)) &= \mathbb{P}\big(\mathcal{E}(\widehat{r}_n)\big) \\ &\geq \mathbb{P}\big(\mathcal{E}(\widehat{r}_n), \mathcal{E}(r), \Omega_\eta\big) \\ &= \mathbb{P}\big(\mathcal{E}(r), \Omega_\eta\big) \end{aligned}$$

where last line comes from Lemma 6.5. We deduce that

$$\begin{aligned}\mathcal{R}_n(\mathcal{X}(\widehat{r}_n)) &\leq 1 - \mathbb{P}(\Omega_\eta, \mathcal{E}(r)) \\ &\leq 1 - \mathbb{P}(\mathcal{E}(r)) + \mathbb{P}(\overline{\Omega_\eta}) \\ &\leq \mathcal{R}_n(\mathcal{X}(r)) + \mathbb{P}(\overline{\Omega_\eta})\end{aligned}$$

and the result follows from Lemma 6.3.

6.3 Proof of theorem 4.2

First observe that for $r > 0$,

$$\begin{aligned}1 - \mathcal{R}_n(\mathcal{X}(r)) &= \mathbb{P}\left(\exists \pi \in \Pi_M \, \forall i = 1, \ldots, M \; \mathbb{X}_n \cap S_i \subseteq \mathcal{X}_{\pi(i)}(r)\right) \\ &\geq \mathbb{P}\left(\exists \pi \in \Pi_M \, \forall i = 1, \ldots, M \; \mathbb{X}_n \cap S_i \subseteq \mathcal{X}_{\pi(i)}(r), \Omega_\eta\right) \qquad (6.4)\end{aligned}$$

where Ω_η is the event defined in Lemma 6.3. Since, under Ω_η, $N_0 < \min_{i \in \{1,\ldots,M\}} N_i$ the event in (6.4) equals

$$\begin{cases} \forall i = 1, \ldots, M, \; \mathbb{X}_n \cap S_i \text{ are } r\text{-connected} \\ \forall i \neq j, \text{ there is no } r\text{-connected path between } \mathbb{X}_n \cap S_i \text{ and } \mathbb{X}_n \cap S_j \\ \Omega_\eta \end{cases}$$

which contains (since $0 < r < \delta$)

$$\begin{cases} \forall i = 1, \ldots, M, \; \mathbb{X}_n \cap S_i \text{ are } r\text{-connected} \\ \text{there is no } r\text{-connected path in } S_0 \text{ with at least } \lfloor \delta / r \rfloor + 1 \text{ observations} \\ \Omega_\eta. \end{cases}$$

We deduce from Lemmas 6.1 and 6.2 that

$$\begin{aligned}\mathcal{R}_n(\mathcal{X}(r)) \leq & \sum_{i=1}^{M} \psi_{n,i}(r) + \varphi_n\left(\left\lfloor \frac{\delta}{r} \right\rfloor + 1, r\right) + \mathbb{P}(\overline{\Omega_\eta}) \\ \leq & \Lambda r^{-d} \exp(-\mathfrak{a} n r^d) + n\varepsilon(\mathfrak{b}\varepsilon n r^D)^{\lfloor \frac{\delta}{r} \rfloor} + \mathbb{P}(\overline{\Omega_\eta}),\end{aligned}$$

where $\Lambda = \sum_{i=1}^{M} \Lambda_i$. Result follows from Lemma 6.3.

References

E. Arias-Castro. Clustering based on pairwise distances when the data is of mixed dimensions. *IEEE Transaction on Information Theory*, 57(3):1692–1706, 2011.

E. Arias-Castro, G. Chen, and G. Lerman. Spectral clustering based on local linear approximations. *Electronic Journal of Statistics*, 5:1537–1587, 2011.

S. Auray, N. Klutchnikoff, and L. Rouvière. On clustering procedures and nonparametric mixture estimation. *Electronic Journal of Statistics*, 9:266–297, 2015.

G. Biau, B. Cadre, and B Pelletier. A graph-based estimator of the number of clusters. *ESAIM Probab. Stat.*, 11:272–280, 2007. ISSN 1292-8100. doi: 10.1051/ps:2007019. URL https://doi.org/10.1051/ps:2007019.

R. Duda, P. Hart, and D. Stork. *Pattern classification*. John Wiley & Sons, 2012.

L. C. Evans and R. F. Gariepy. *Measure theory and fine properties of functions*. CRC press, 2015.

Kenneth Falconer. *Fractal geometry*. John Wiley & Sons, Ltd., Chichester, third edition, 2014. ISBN 978-1-119-94239-9. Mathematical foundations and applications.

R. Filipovych, S. Resnick, and C. Davatzikos. Semi-supervised cluster analysis of imaging data. *NeuroImage*, 54(3):2185–2197, 2011.

Allen Gersho and Robert M Gray. *Vector quantization and signal compression*, volume 159. Springer Science & Business Media, 2012.

J.A. Hartigan. *Clustering Algorithms*. John Wiley, 1975.

A. Jain and R. Dubes. Algorithms for clustering data. 1988.

S. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

J. MacQueen. 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.

M. Maier, M. Hein, and U. Von Luxburg. Optimal construction of k-nearest-neigbor graphs for identifying noisy clusters. *Theoritical Computer Science*, 410: 1749–1764, 2009.

G. McLachlan and K. Basford. *Mixture models: Inference and applications to clustering*, volume 84. Marcel Dekker, 1988.

B. Nadler and M. Galun. Fundamental limitations of spectral clustering. In *Advances in neural information processing systems*, pages 1017–1024, 2007.

A. Y Ng, M. I Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

D. Satish and C. Sekhar. Kernel based clustering and vector quantization for speech segmentation. In *Neural Networks, 2006. IJCNN'06. International Joint Conference on*, pages 1636–1641. IEEE, 2006.

R. Shorack and J. Wellner. *Empirical Processes with Applications to Statistics*. SIAM, 1986.

Y. Yamanishi, J. Vert, and M. Kanehisa. Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, 20(suppl_1):i363–i370, 2004.

L. Yengo, J. Jacques, and C. Biernacki. Variable clustering in high dimensional linear regression models. *Journal de la Societe Française de Statistique*, 155(2): 19, 2014.

E. Zeng, C. Yang, T. Li, and G. Narasimhan. Clustering genes using heterogeneous data sources. In *Computational Knowledge Discovery for Bioinformatics Research*, pages 67–83. IGI Global, 2012.