

# Which Data Sets are ‘Clusterable’? – A Theoretical Study of Clusterability

Margareta Ackerman  
D.R.C. School of Computer Science  
University of Waterloo  
Waterloo, Ontario, Canada N2L 3G1  
mackerma@uwaterloo.ca

Shai Ben-David  
D.R.C. School of Computer Science  
University of Waterloo  
Waterloo, Ontario, Canada N2L 3G1  
shai@cs.uwaterloo.ca

## ABSTRACT

We investigate measures of the clusterability of data sets. Namely, ways to define how ‘strong’ or ‘conclusive’ is the clustering structure of a given data set. We address this issue with generality, aiming for conclusions that apply regardless of any particular clustering algorithm or any specific data generation model.

We survey several notions of clusterability that have been discussed in the literature, as well as propose a new notion of data clusterability.

Our comparison of these notions reveals that, although they all attempt to evaluate the same intuitive property, they are pairwise inconsistent.

Our analysis discovers an interesting phenomenon; *the more clusterable a data set is, the easier it is (computationally) to find a close-to-optimal clustering of that data*. It has been recently shown that such a property holds with respect to one notion of clusterability, ‘ $k$  - separability’. We show that this phenomenon holds for other notions as well. In particular, we prove that for well clusterable data, using the clusterability notions we discuss, near-optimal clustering can be *efficiently* computed.

Finally, we investigate how hard it is to determine the clusterability value of a given data set. In most cases, it turns out that this is an NP-hard problem.

## 1. INTRODUCTION

Clustering is at the same time a very basic and an immensely useful task. However, in spite of hundreds of clustering papers being published every year, its theoretical foundations are distressingly meager. Clearly, it is very difficult to develop a theory of clustering at a level of generality that will make it relevant across different applications and algorithmic approaches. In this paper we try to take a step in that direction by investigating possible formalizations of the cen-

tral and intuitive notion of *clusterability* of data sets.

The aim of clustering is to uncover meaningful partitions in data; however, not all data sets have meaningful partitions. Clusterability is a measure of clustered structure in a data set. So far, clusterability has been used only peripherally and with no theoretical support. We pioneer a theoretical study of clusterability, comparing and analyzing notions of clusterability. We address this issue with generality, aiming for conclusions that apply regardless of any particular clustering algorithm or any specific data generation model.

We survey several notions of clusterability that have been discussed in the literature (some less explicitly than others), as well as propose a new notion of data clusterability. Our comparison of these notions reveals that, although these notions attempt to evaluate the same intuitive property, and all appear to be reasonable, they are pairwise inconsistent. That is, for each pair of notions, there are data sets that are arbitrarily well-clusterable by one of the notions, but poorly clusterable by the other notion. This illustrates the significance of choosing a specific notion of clusterability for an experiment or theoretical study, as the same results are not necessarily obtainable using a different notion.

Our analysis of these notions gives rise to an interesting phenomenon; *the more clusterable a data set is, the easier it is (computationally) to find a close-to-optimal clustering of that data*. It has been recently shown that such a property holds with respect to one notion of clusterability, ‘ $k$  - separability’. We show that this phenomenon holds for other notions as well. In particular, we prove that for well clusterable data, using the clusterability notions we discuss, near-optimal clustering can be *efficiently* computed.

We investigate how hard is it to determine the clusterability of a data set. The hardness of determining clusterability has practical implications since notions of clusterability (at least the ones presented here) can be used to determine the difficulty of finding a good clustering. For each notion, we find the hardness of determining whether the clusterability of a data set exceeds a given threshold. In most cases, it turns out that this is an NP-hard problem.

We begin by presenting the previous notions and our new notion of clusterability. Next, we prove that for each of these notions, when clusterability is better, it is easier to find a provably near-optimal clustering. We then present the

computational complexity results, followed by the pairwise comparison of the notions.

## 2. FRAMEWORK AND DEFINITIONS

A  $k$ -clustering of data set  $X$  is a  $k$ -partition of  $X$ , that is, a set of  $k$  non-empty, disjoint subsets of  $X$  such that their union is  $X$ . A clustering of  $X$  is a  $k$ -clustering of  $X$  for some  $k \geq 1$ . For  $x, y \in X$  and clustering  $C$  of  $X$ ,  $x \sim_C y$  whenever  $x$  and  $y$  are in the same cluster with respect to  $C$ , and  $x \not\sim_C y$ , otherwise.

We work with data sets in Euclidean space, therefore distances between points are specified implicitly. A notion of clusterability is a function that takes a data set  $X \subseteq \mathbf{R}^m$ , and returns a real value. This function is suppose to represent how ‘strong’ or ‘conclusive’ is the clustering structure of the data set.

A clustering  $C = \{X_1, X_2, \dots, X_k\}$  of  $X$  is *center-based* if there exist points  $c_1 \in X_1, \dots, c_k \in X_k$ , such that for all  $i$ , for all  $x \in X_i$  and all  $j \neq i$ ,  $\|x - c_i\| \leq \|x - c_j\|$ . The set of such points  $c_1, \dots, c_k$  is called a set of *centers* for the clustering  $C$ . The Voronoi partition induced by the centers of a clustering coincides with that clustering partition. While a partition may not have a set of centers inducing it, center-based clustering always do.

Given a loss function  $\mathcal{L}$ , we let

$$OPT_{\mathcal{L},k}(X) = \min\{\mathcal{L}(C) \mid C \text{ a } k\text{-clustering of } X\},$$

the loss of a  $k$ -clustering of  $X$  that minimizes  $\mathcal{L}$ .

Since many of the notions found in the literature are defined with respect to the  $k$ -means loss function, we focus our analysis on clustering with respect to  $k$ -means. We could use the following notions of clusterability with many other common loss functions, and most of our discussion would carry through. In addition, our new notion of clusterability is defined for any loss function where the optimal clustering is center-based. Furthermore, most of our results can be extended from Euclidean space to arbitrary normed vector spaces.

## 3. PREVIOUS WORK

Notions of clusterability have been discussed previously. We start by showing some previously presented notions (then, in Section 4, we add to the list a new notion of clusterability). Since two of the following notions were defined in the context of the  $k$ -means loss function, we start this section by reviewing this definition.

The *center of mass* of data set  $X$  is  $center\text{-}mass(X) = \frac{1}{|X|} \sum_{x \in X} x$ . The  $k$ -means problem is to find a  $k$ -clustering  $C = \{X_1, X_2, \dots, X_k\}$  of a given data set  $X$  that minimizes the  $k$ -means loss function,  $k\text{-}means(C) = \sum_{i=1}^k \sum_{x \in X_i} \|x - center\text{-}mass(X_i)\|^2$ . A  $k$ -means optimal clustering of  $X$  is a  $k$ -clustering that minimizes the  $k$ -means loss over all  $k$ -clusterings of  $X$ . We let  $OPT_k(X) = OPT_{k\text{-}means,k}(X)$ .

### 3.1 Variance ratio clusterability

Variance Ratio measures the ratio of the variance between clusters over the variance of points within clusters. This no-

tion was presented by Zhang [8]. Recall that the variance of  $X$  is  $\sigma^2(X) = \frac{1}{|X|} \sum_{x \in X} \|x - center\text{-}mass(X)\|^2$ . Consider a clustering  $C = \{X_1, X_2, \dots, X_k\}$ . Let  $p_i = \frac{|X_i|}{|X|}$ . Let  $B_C(X) = \sum_{i=1}^k p_i \|center\text{-}mass(X_i) - c\|^2$  denote the *between-cluster variance* of  $C$  and  $W_C(X) = \sum_{i=1}^k p_i \sigma^2(X_i)$  the *within-cluster variance* of  $C$ .

**DEFINITION 1 (VARIANCE RATIO).** *The variance ratio of  $X$  for  $k$  is*

$$VR_k(X) = \max_{C \in \mathcal{C}} \frac{B_C(X)}{W_C(X)},$$

where  $\mathcal{C}$  is the set of  $k$ -means optimal clusterings of  $X$ .

The range of variance ratio is  $[0, \infty)$  and higher values of variance ratio indicate better clusterability. Observe that  $\sigma^2(X) = W_C(X) + B_C(X)$  and  $W_C(X)$  is the loss function that  $k$ -means minimizes divided by  $n$ .

If two clusterings of the same data set  $X$  have the same variance ratio, then they also have the same within-cluster variance and between-cluster variance (since  $\sigma^2(X)$  is constant over all clusterings of  $X$ ). Thus, we let  $W_k(X) = W_C(X)$  and  $B_k(X) = B_C(X)$ , where  $C$  is a clustering that maximizes the between over within ratio over all  $k$ -means optimal clusterings of  $X$ .

### 3.2 Separability clusterability

The separability notion of clusterability captures how sharp is the drop in the loss function when moving from a  $(k-1)$ -clustering to a  $k$ -clustering. This notion was introduced by Ostrovsky, Rabani, Schulman, and Swamy [6].

**DEFINITION 2 (SEPARABILITY).** *A data set  $X$  is  $(k, \epsilon)$ -separable if  $OPT_k(X) \leq \epsilon OPT_{k-1}(X)$ .*

For convenience, we define  $S_k(X)$  to be the smallest  $\epsilon$  such that  $X$  is  $(k, \epsilon)$ -separable. The range of separability is  $[0, 1)$ . A data set has better separability than another data set if it is separable for smaller  $\epsilon$ .

### 3.3 Worst pair ratio clusterability

We call the minimum distance between two points in different clusters of a clustering  $C$  the *split* between the two clusters, and the minimal split between two clusters as the split of  $C$ ; that is,  $split_C(X) = \min_{x \not\sim_C y} \|x - y\|$ . We call the maximum distance between two points within a cluster in  $C$  the *width* of the cluster, and the maximal width of a cluster in  $C$  the width of  $C$ ;  $width_C(X) = \max_{x \sim_C y} \|x - y\|$ . The following definition of clusterability was presented by Epter et al.[3]

**DEFINITION 3 (EPTER WORST PAIR RATIO).** *The Epter worst pair ratio of  $X$  is*

$$Epter\text{-}WPR(X) = \max\left\{\frac{split_C(X)}{width_C(X)} \mid C \text{ a clustering of } X\right\}.$$

We introduce a variation of worst pair ratio that is more in line with the definitions of variance ratio and separability.

**DEFINITION 4 (WORST PAIR RATIO).** *The worst pair ratio of  $X$  for  $k$  is*

$$WPR_k(X) = \max_{C \in \mathcal{C}} \frac{\text{split}_C(X)}{\text{width}_C(X)},$$

where  $\mathcal{C}$  is the set of  $k$ -means optimal clusterings of  $X$ .

Since the two above definitions are very similar, our results apply for both definitions, using essentially the same proofs (for meaningful comparison with other measures, it suffices to modify Epter’s definition to maximize the ratio of split over width over all  $k$ -clusterings of the data set).

The range of worst pair ratio is  $[0, \infty)$  and higher values of worst pair ratio mean better clusterability. In Lemma 1, we prove that if there is a clustering with split greater than width, than that clustering is unique. Therefore, if there exists a  $k$ -means optimal clustering  $C$  with split greater than width, we let  $\text{split}_k(X) = \text{split}_C(X)$  and  $\text{width}_k(X) = \text{width}_C(X)$ .

For some data sets, there is a  $k$ -clustering with the split greater than the width, but no  $k$ -means optimal clusterings with this property. For instance, consider the set of points  $\{0, 10, 19, 19.1, 19.2, \dots, 19.9\} \subseteq \mathbf{R}$ . Then the 2-clustering

$$\{\{0\}, \{10, 19.1, 19.2, \dots, 19.9\}\}$$

has larger split than width. However, the unique  $k$ -means optimal 2-clustering,

$$\{\{0, 10\}, \{19.1, 19.2, \dots, 19.9\}\},$$

has smaller split than width. This illustrates that our definition of worst pair ratio is strictly stronger than Epter’s definition for  $k$ -clusterings.

One of the shortcomings of these notions is their intolerance to noise and outliers. For instance, having only one out of a thousand clusters with high width pulls down the worst pair ratio, suggesting that what may intuitively be well-clusterable data, is poorly-clusterable by this notion of clusterability. Similarly, a small number of non-representative points can drastically decrease the split of a clustering. When using worst pair ratio as a measure of clusterability on real data sets, it may be helpful to preprocess data by removing noise and outliers.

## 4. CENTER PERTURBATION CLUSTERABILITY

In this section, we introduce a new notion of clusterability aiming to capture the clustering robustness to center perturbations. This notion provides a distinctly different perspective at clusterability evaluation than previous notions. In a center-based clustering, each point in a cluster is closer to its own cluster center than to the center of any other cluster. Consider what happens if these centers are slightly perturbed. Are points still going to be closer to the perturbed centers of their clusters? If we re-cluster the data using the perturbed centers, how much does the loss of the

clustering change? If the optimal clustering is “good”, we expect such change to have little effect on clustering loss. That is, if the data set is well-clusterable, the optimal clustering should be robust to (small) center perturbations.

**DEFINITION 5 ( $\epsilon$ -CLOSE).** *Two center-based clusterings,  $C$  and  $C'$  of  $X$ , are  $\epsilon$ -close, if there exist centers  $c_1, c_2, \dots, c_k$  of  $C$ , and centers  $c'_1, c'_2, \dots, c'_k$  of  $C'$ , such that for all  $i \leq k$ ,  $\|c_i - c'_i\| \leq \epsilon$ .*

**DEFINITION 6 (CENTER PERTURBATION CLUSTERABILITY).** *A data set  $X$  is  $(\epsilon, \delta)$ -CP clusterable for  $k$  (for  $\epsilon, \delta \geq 0$ ), if for every clustering  $C$  of  $X$  that is  $\epsilon$ -close to some optimal  $k$ -clustering of  $X$ ,  $\mathcal{L}(C) \leq (1 + \delta)OPT_{\mathcal{L},k}(X)$ .*

In the next section, we show that when a data set has good center perturbation clusterability, it is easy to find clusterings of the data whose loss is close to the loss of an optimal clustering.

There is another interesting way to evaluate data clusterability in terms of center perturbation. Instead of comparing loss function values, one can consider the variability in the resulting clusterings. For instance, the following is a definition of distance between two clusterings  $C = \{C_1, C_2, \dots, C_k\}$  and  $C' = \{C'_1, C'_2, \dots, C'_k\}$  of data set  $X$  as defined by Meila [5]. Let  $m_{i,j} = |C_i \cap C'_j|$ . Then the distance between clusterings  $C$  and  $C'$  is

$$\mathcal{D}(C, C') = 1 - \frac{1}{n} \max_{\{\pi \in \Pi\}} \sum_i m_{i, \pi(i)},$$

where  $n$  is the number of points in the data set and  $\Pi$  is the set of all permutations of  $\{1, 2, \dots, k\}$ . There are many other ways to define the distance between clusterings, see [5] for additional examples.

A data set is  $(\epsilon, \delta)$ - $\mathcal{D}$ -CP clusterable for  $k$  if  $\max(\mathcal{D}(C, C')) = \delta$ , over all pairs  $(C, C')$  where  $C$  is an optimal  $k$ -clustering of  $X$  and  $C'$  is  $\epsilon$ -close to  $C$ . Note that if  $X$  is  $(\epsilon, 0)$ - $\mathcal{D}$ -CP clusterable (that is, the clustering does not change following the center perturbation) then it is  $(\epsilon, 0)$ -CP clusterable. Most of the results comparing CP clusterability to previous notions of clusterability, presented in Section 7, apply also for  $\mathcal{D}$ -CP clusterability.

## 5. CLUSTERING OF WELL-CLUSTERABLE DATA

Our analysis of notions of clusterability gives rise to an interesting phenomenon:

*The more clusterable a data set is, the easier it is, computationally, to find a close-to-optimal clustering of that data.*

Recently, Ostrovsky et al. ([6], 2006) proved that the above property holds for separability clusterability. We prove the property for center-perturbation clusterability and worst pair ratio clusterability. We also show that the property holds for 2-clusterings for variance ratio clusterability. It is an intriguing challenge to explore whether such phenomena extends to other notions of clusterability.

## 5.1 Center perturbation

We show that data sets with better center perturbation clusterability are easier to cluster well. That is, when a data set has better center perturbation clusterability, a provably near-optimal clustering can be computed efficiently. This is in contrast with the situation for arbitrary input, where, for more interesting objective functions (like  $k$ -means) optimal solutions are NP-hard to approximate.

Let  $rad(X)$  denote the radius of the minimum hypersphere that contains all the points in  $X$  (we use the radius of  $X$  to normalize the measure to ensure scale invariance).

**THEOREM 1.** *Given a data set  $X \subseteq \mathbf{R}^m$  on  $n$  points, there exists an algorithm such that, for every fixed  $k \geq 2$  and  $\delta \geq 0$ , if  $X$  is  $(\frac{rad(X)}{\sqrt{\ell}}, \delta)$ -CP clusterable for  $k$ , then the algorithm runs in time polynomial in  $n$ , and outputs a clustering  $C$  of  $X$  with at most  $k$  clusters, such that*

$$\mathcal{L}(C) \leq (1 + \delta)OPT_{\mathcal{L},k}(X).$$

Moreover, this result holds for any loss function  $\mathcal{L}$  where all optimal clusterings are center-based (an optimal  $k$ -clustering is a  $k$ -clustering minimizing  $\mathcal{L}$ ).

We present an algorithm for finding a clustering that is  $\epsilon$ -close to an optimal clustering. The algorithm is based on an algorithm by Ben-David, Eiron, and Simon [1]. If we know the  $(\epsilon, \delta)$ -CP clusterability of  $X$ , then we can lower bound the quality of the clustering found by the algorithm.

Let an  $\ell$ -sequence denote a collection of  $\ell$  elements of  $X$  (not necessarily distinct). The algorithm iterates through all  $k$ -tuples of  $\ell$ -sequences. For each such tuple, it finds the clustering induced by the centers of mass of the  $\ell$ -sequences. It then chooses the clustering with minimal loss.

**ALGORITHM 1.** *Finding near optimal clusterings*

*INPUT: A data set  $X$ ,  $k \geq 1$ ,  $\ell \geq 1$ .*

*OUTPUT: Outputs a clustering  $C_A$  of  $X$  such that  $\mathcal{L}(C_A) \leq \min\{\mathcal{L}(C) \mid C \in \mathcal{C}\}$ , where  $\mathcal{C}$  is the set of all  $k$ -clusterings of  $X$  that are  $\frac{rad(X)}{\sqrt{\ell}}$ -close to any optimal  $k$ -clustering of  $X$ .*

1.  $C_A = \emptyset$
2. for each  $k$ -tuple of  $\ell$ -sequences;
  - (a) find the centers of mass of the  $\ell$ -sequences, call this set  $S$
  - (b) find the clustering  $\hat{C}$  that  $S$  induces on  $X$
  - (c) if  $C_A = \emptyset$  or  $\mathcal{L}(\hat{C}) < \mathcal{L}(C_A)$  then set  $C_A = \hat{C}$
3. return  $C_A$

To prove Theorem 1, we use the following result by Maurey.

**THEOREM 2** (MAUREY, 1981). *For any fixed  $\ell \geq 1$  and each  $x'$  in the convex hull of  $X$ , there exist  $x_1, x_2, \dots, x_\ell \in X$  such that  $\|x' - \frac{1}{\ell} \sum_{i=1}^{\ell} x_i\| \leq \frac{rad(X)}{\sqrt{\ell}}$ .*

Note that  $x_1, x_2, \dots, x_\ell$  are not necessarily distinct from each other. We now prove Theorem 1.

*Proof of Theorem 1*

**PROOF.** By Maurey's result, there is a clustering,  $\hat{C}$ , examined by Algorithm 1, that is  $\frac{rad(X)}{\sqrt{\ell}}$ -close to an optimal clustering of  $X$ . Since Algorithm 1 selects the minimal loss clustering of the ones it reviews,  $\mathcal{L}(C_A) \leq \mathcal{L}(\hat{C})$ . Since  $\hat{C}$  is  $\frac{R}{\sqrt{\ell}}$ -close to an optimal clustering of  $X$ , and  $X$  is  $(\frac{R}{\sqrt{\ell}}, \delta)$ -CP clusterable,  $\mathcal{L}(C_A) \leq \mathcal{L}(\hat{C}) \leq (1 + \delta)OPT_{\mathcal{L},k}(X)$ . The running time of Algorithms 1 is  $O(kmn^{\ell k+1})$ . To see that, observe that there are  $O(n^{\ell k})$   $k$ -tuples of  $\ell$ -sequences, and that for each  $k$ -tuple of  $\ell$ -sequences the algorithm does  $O(kmn)$  operations. Note that, since a clustering induced by  $k$  centers has at most  $k$  clusters, Algorithm 1 returns a clustering with no more than  $k$  clusters (for most common loss functions, including  $k$ -means, any subdivision of clusters improves the loss of a clustering).  $\square$

## 5.2 Worst pair ratio

We now prove that good worst pair ratio clusterability implies that a good clustering can be efficiently computed. First, we show that there is at most one  $k$ -clustering with split strictly greater than width.

**LEMMA 1.** *If there exists a  $k$ -clustering  $C$  of  $X$ , for  $k \geq 2$ , such that  $width_C(X) < split_C(X)$ , then there is only one such clustering.*

**PROOF.** Assume that there are two distinct clusterings  $C$  and  $C'$  of  $X$ , each with exactly  $k$  non-empty clusters, such that  $width_C(X) < split_C(X)$  and  $width_{C'}(X) < split_{C'}(X)$ . If  $split_C(X) = split_{C'}(X)$ , then  $C = C'$ , since each pair of points belong to the same cluster if and only if their distance is less than the split of the clustering. Assume, without loss of generality, that  $split_C(X) < split_{C'}(X)$ . Then every pair of points that belong to the same cluster in  $C$  also belong to the same cluster in  $C'$ . In addition, there is a pair of points that belong to the same cluster in  $C'$  but not in  $C$  (merging two clusters in  $C$ ). So  $C'$  has fewer non-empty clusters than  $C$ , thus it is not a  $k$ -clustering.  $\square$

**THEOREM 3.** *Given a data set  $X \subseteq \mathbf{R}^m$  where  $WPR_k(X) \geq 1$  for some  $k \geq 2$ , we can find the  $k$ -means optimal clustering  $C$  in  $O(n^2 \log n)$  operations, where  $n = |X|$ . Moreover,  $C$  maximizes the split over width ratio over all  $k$ -means optimal clusterings of  $X$ .*

**PROOF.** Let  $C$  be a  $k$ -means optimal clustering that maximizes  $WPR_k(X)$ , over all  $k$ -means optimal clusterings of  $X$ . Then  $width_C(X) < split_C(X)$ . By Lemma 1,  $C$  is the unique clustering with the width strictly small than the split of the clustering.

We can run the single linkage algorithm to recover  $C$ . That is, sort the pairs of points in  $X$  based on pairwise distances, and put pairs of points in the same clusters, starting with the pair with minimal distance and going up the list until exactly  $k$  clusters are formed. Since  $width_C(X) < split_C(X)$ , the

procedure terminates, finding  $C$ , when all edges of length at most  $width_C(X)$  have been marked as within cluster edges. This takes  $O(n^2 \log n)$  operations.  $\square$

Note that we can show the same result for Epter worst pair ratio.

On the other hand, it is unlikely that there is a polynomial-time algorithm for finding a  $k$ -means optimal clustering for any data set  $X$  where  $WPR_k(X) \leq 1$ . Finding a  $k$ -clustering that minimizes the  $k$ -means clustering is an NP-hard problem [2], and as by Theorem 3 the problem can be solved in polynomial time on instances where  $WPR_k(X) \geq 1$ , it follows that the problem is NP-hard for the remaining instances.

### 5.3 Variance ratio

We prove that when variance ratio is good for two clusters, a provably near-optimal clustering can be efficiently computed. We make use of a previous result on separability clusterability. Ostrovsky et al. prove that data with good separability clusterability is easy to cluster.

**THEOREM 4** (OSTROVSKY ET AL., [6]). *Given a  $(2, \epsilon^2)$ -separable data set  $X \subseteq \mathbf{R}^m$ , we can find a 2-clustering with  $k$ -means loss at most  $\frac{OPT_2(X)}{1-\rho}$  with probability at least  $1 - O(\rho)$  in time  $O(nm)$ , where  $\delta = \Theta(\epsilon^2)$  and  $n = |X|$ .*

For arbitrary  $k$ , Ostrovsky et al. [6] provide a similar result.

We now show that variance ratio and separability are equivalent for  $k = 2$ .

**LEMMA 2.**  $VR_2(X) = \frac{1}{S_2(X)} - 1$  for any data set  $X$ .

**PROOF.** Let  $n = |X|$ . We know that  $\sigma^2(X) = W_2(X) + B_2(X)$ , and  $W_2(X) = \frac{OPT_2(X)}{S_2(X)} = S_2(X)\sigma^2(X)$ . Thus,  $VR_2(X) = \frac{B_2(X)}{W_2(X)} = \frac{\sigma^2(X) - W_2(X)}{W_2(X)} = \frac{\sigma^2(X) - S_2(X)\sigma^2(X)}{S_2(X)\sigma^2(X)} = \frac{1}{S_2(X)} - 1$ .  $\square$

Combining Lemma 2 with Theorem 4, implies that a data set with good variance ratio clusterability is easy to cluster well.

**COROLLARY 1.** *Given a data  $X \subseteq \mathbf{R}^m$ , we can find a 2-clustering with  $k$ -means loss at most  $\frac{OPT_2(X)}{1-\rho}$  with probability at least  $1 - O(\rho)$  in time  $O(nm)$ , where  $\delta = \Theta\left(\frac{1}{(VR_2(X)+1)^2}\right)$ .*

## 6. COMPUTATIONAL COMPLEXITY OF CLUSTERABILITY

We analyze the computational complexity of determining the clusterability of a data set. To the best of our knowledge, this is the first computational complexity analysis of

notions of clusterability. The hardness of determining clusterability has practical implications since, as shown in Section 5, notions of clusterability (at least the ones presented here) can be used to determine the difficulty of finding a good clustering. For each previous notion of clusterability that was presented in Section 3, we find the hardness of determining whether the clusterability of a data set exceeds a given threshold.

Our results show for when worst pair ratio is sufficiently good, worst pair ratio clusterability can be found in polynomial time. As discussed above, worst pair ratio is very sensitive to noise and outliers, and thus often assigns low clusterability to intuitively well-clusterable data sets. Therefore, good clusterability for worst pair ratio implies particularly clear clustering structure. As will be shown in Section 7, well-clusterability by worst pair ratio implies well-clusterability by separability and variance ratio, but not vice-versa. For separability and variance ratio clusterability, which can detect clustered structure in a wider range of circumstances, the problem of determining the degree of clusterability is NP-hard.

For future work, it would be interesting to explore the problem of determining variance ratio and separability clusterability in the more flexible framework of property testing. The goal of property testing is to determine whether a given object has a desired property or if the object is close to some other object which has the property. The tester is allowed to make mistakes on both positive and negative assertions, with a certain probability. For more details on property testing, see [4]. Using this setting, we pose the following question: given a notion of clusterability, we ask how hard is it to determine whether the clusterability of a data set surpasses a given threshold or if the data set is similar to some data set that does.

### 6.1 Separability

We prove that determining whether the separability clusterability of a given data set exceeds a given threshold is an NP-hard problem.

**THEOREM 5.** *Given  $X \subseteq \mathbf{R}^m$ , integer  $k \geq 2$ , and  $0 < \epsilon < 1$ , it is NP-hard to determine whether  $X$  is  $(k, \epsilon)$ -separable.*

**PROOF.** The decision version of the  $k$ -means problem is as follows: Does there exist a  $k$ -clustering of  $X$  with  $k$ -means loss at most  $v$ ? This problem is NP-complete for  $k \geq 2$  [2].

A data set  $X$  is  $(2, \epsilon)$ -separable if  $\frac{OPT_2(X)}{OPT_1(X)} \leq \epsilon$ . Suppose that we could determine whether any set  $X \subseteq \mathbf{R}^m$  is  $(2, \epsilon)$ -separable for any  $0 < \epsilon < 1$  in polynomial-time. Then since  $OPT_2(X) \leq \epsilon OPT_1(X)$ , and  $OPT_1(X) = |X|\sigma^2(X)$  can be found in polynomial-time, we can determine whether  $OPT_2(X) \leq \mu$  for any  $\mu$  by checking if  $X$  is  $(2, \epsilon)$ -separable for  $\epsilon = \frac{\mu}{OPT_1(X)}$ . But since determining if  $OPT_2(X) \leq \mu$  for any arbitrary  $\mu > 0$  is NP-hard, determining if  $X$  is  $(2, \epsilon)$ -separable is NP-hard.

To show that the problem is NP-hard for any  $k \geq 3$ , we reduce the problem for  $k = 2$  to the problem for  $k \geq 3$ .

Given  $X$ , add  $k-2$  points sufficiently far away from all points in  $X$  and from each other, so that each one of the new points is its own cluster in the  $k$ -means optimal clustering. Then in any  $k$ -means optimal clustering, the remaining 2 clusters are an optimal 2-means solution for the original data set.  $\square$

## 6.2 Variance ratio

Determining the level of clusterability is also an NP-hard problem for the variance ratio notion of clusterability.

**THEOREM 6.** *Given  $X \subseteq \mathbf{R}^m$ ,  $k \geq 2$ , and  $r > 0$ , it is NP-hard to determine whether  $VR_k(X) \geq r$ .*

**PROOF.** We know that  $\sigma^2(X) = W_k(X) + B_k(X)$ . Then  $VR_k(X) = \frac{B_k(X)}{W_k(X)} = \frac{\sigma^2(X) - W_k(X)}{W_k(X)} = \frac{\sigma^2(X)}{W_k(X)} - 1 = \frac{|X|\sigma^2(X)}{OPT_k(X)} - 1$ .

Thus, if we can tell whether  $VR_k(X) = \frac{|X|\sigma^2(X)}{OPT_k(X)} - 1 \geq r$  for any  $r > 0$ , then we can tell whether  $OPT_k(X) \leq \frac{|X|\sigma^2(X)}{r+1}$ . We can find  $|X|\sigma^2(X)$  in polynomial time. Also, by definition of  $OPT_k(X)$ ,  $OPT_k(X) \leq |X|\sigma^2(X)$ . Thus, by setting  $r = \frac{|X|\sigma^2(X)}{v} - 1$ , we can find whether  $OPT_k(X) \leq v$  for any  $v > 0$ . However, this problem is NP-hard for  $k \geq 2$  [2].  $\square$

## 6.3 Worst pair ratio

We show that whenever the worst pair ratio clusterability is sufficiently good, then it can be determined in polynomial time.

**THEOREM 7.** *Given an integer  $k \geq 2$  and a data set  $X \subseteq \mathbf{R}^m$  where  $WPR_k(X) > 1$ , we can determine the worst pair ratio of  $X$  in polynomial-time.*

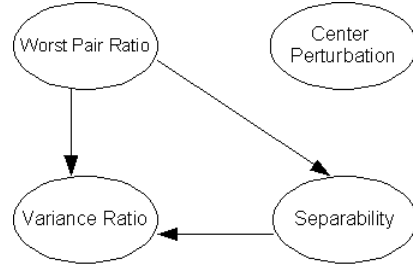
**PROOF.** By Theorem 3, given a data set  $X$  on  $n$  points where  $WPR_k(X) > 1$ , a  $k$ -means optimal clustering  $C$  that maximizes the split over width ratio over all  $k$ -means optimal clusterings of  $X$  can be found in  $O(n^2 \log n)$  operations. We can then find the split and width of  $C$  using  $O(n^2)$  additional operations, thus finding  $WPR_k(X)$ .  $\square$

## 7. COMPARISONS OF CLUSTERABILITY NOTIONS

We perform a pairwise comparison of clusterability notions. Although all these notions attempt to evaluate the same intuitive property, it turns out that none of the notions are equivalent. The following definition assumes that  $\alpha$  and  $\beta$  assign higher values to better clusterable data (otherwise, reverse the direction of the corresponding inequality).

**DEFINITION 7 (IMPLICATION).** *A clusterability notion  $\alpha$  does not imply clusterability notion  $\beta$ , if for infinitely many  $n$ , for any  $t_\alpha \in \text{range}(\alpha)$  and  $t_\beta \in \text{range}(\beta)$  there exists a data set  $X \subseteq \mathbf{R}^m$  on  $n$  points with  $\alpha(X) \leq t_\alpha$  and  $\beta(X) \geq t_\beta$ .*

To prove implication, we present stronger results than are necessary by the above definition. We show that notion  $\alpha$



**Figure 1: Relationships between notions of clusterability.** An arrow from a notion  $\alpha$  to a notion  $\beta$  indicates that  $\alpha$  implies  $\beta$ , and the absence of an arrow indicates that  $\alpha$  does not imply  $\beta$ .

implies notion  $\beta$  by proving that  $\alpha$  clusterability sets a lower bound on  $\beta$  clusterability.

We analyze separability, variance ratio, worst pair ratio, and center perturbation with the  $k$ -means loss function, finding that none of these notions are equivalent; that is, for any pair of notions, at least one of the notions does not imply the other. Figure 1 summarizes our results.

We found that worst pair ratio is a stronger notion of clusterability than variance ratio and separability - it implies these notions, but these notions do not imply it. The only other dependence found is that separability implies variance ratio. We get that worst pair ratio is the strongest of the three notions, followed by separability, followed by variance ratio. Center perturbation is inconsistent with all these notions; it does not imply the other notions, and the other notions do not imply it. This analysis brings out the versatility of the current definitions of clusterability, and enables a comparison of previous results on clusterability.

## 7.1 Separability versus variance ratio

In this section, we explore the relationship between separability and variance ratio. By Lemma 2, the two notions are equivalent for two clusters, that is, separability implies variance ratio and variance ratio implies separability. We provide a characterization of the relationship between these notions for  $k \geq 3$ , and use it to prove that separability implies variance ratio. We then show that for three or more clusters, variance ratio does not imply separability.

We now present the characterization of the relationship between separability and variance ratio for  $k \geq 3$ .

**LEMMA 3.**

$$W_k(X) = S_2(X)S_3(X) \cdots S_k(X)\sigma^2(X),$$

for any  $k \geq 2$  and  $X \subseteq \mathbf{R}^m$ ,

**PROOF.** The result holds for  $k = 2$ , since  $W_2(X) = \frac{OPT_2(X)}{|X|} = \frac{S_2(X)|X|\sigma^2(X)}{|X|} = S_2(X)\sigma^2(X)$ . Assume that the result holds for all  $j < k$ . Then,

$$\begin{aligned}
W_k(X) &= \frac{1}{|X|} OPT_k(X) \\
&= \frac{1}{|X|} S_k(X) OPT_{k-1}(X) \\
&= \frac{1}{|X|} S_k(X) |X| W_{k-1}(X) \\
&= S_2(X) S_3(X) \cdots S_k(X) \sigma^2(X)
\end{aligned}$$

□

**THEOREM 8.**  $VR_k(X) = \frac{VR_{k-1}(X)+1}{S_k(X)} - 1$ , for  $k \geq 3$  and  $X \subseteq \mathbf{R}^m$ ,

**PROOF.** By Lemma 3,  $W_k(X) = S_2(X) S_3(X) \cdots S_k(X) \sigma^2(X)$ . Then,

$$\begin{aligned}
VR_k(X) &= \frac{B_k(X)}{W_k(X)} \\
&= \frac{\sigma^2(X) - W_k(X)}{W_k(X)} \\
&= \frac{\sigma^2(X) - S_2(X) S_3(X) \cdots S_k(X) \sigma^2(X)}{S_2(X) S_3(X) \cdots S_k(X) \sigma^2(X)} \\
&= \frac{1}{S_2(X) S_3(X) \cdots S_k(X)} - 1 \\
&= \frac{1}{S_k(X)} \cdot \frac{\sigma^2(X)}{W_{k-1}(X)} - 1 \\
&= \frac{VR_{k-1}(X) + 1}{S_k(X)} - 1
\end{aligned}$$

□

By the above result, we can show that separability implies variance ratio.

**THEOREM 9.**  $VR_k(X) \geq \frac{1}{S_k(X)} - 1$  for any  $k \geq 2$  and data set  $X \subseteq \mathbf{R}^m$ .

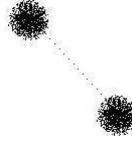
**PROOF.** By Theorem 8, for any  $k \geq 3$ ,

$$VR_k(X) = \frac{VR_{k-1}(X) + 1}{S_k(X)} - 1.$$

Since  $VR_{k-1}(X) \geq 0$ , this implies that  $VR_k(X) \geq \frac{1}{S_k(X)} - 1$ . For  $k = 2$ , the result follows from Lemma 2. □

We now show that variance ratio does not imply separability.

**THEOREM 10.** For any  $\epsilon < 1$ , there is a data set  $X$  and a  $k \geq 3$  such that  $S_k(X) \geq \epsilon$  and  $VR_k(X)$  is arbitrarily high.



**Figure 2:** An example of a data set with good variance ratio clusterability and poor worst pair ratio clusterability.

**PROOF.** It can be shown that there exists a  $k$  such that  $S_{k-1}(X) \geq \epsilon$  for any  $\epsilon < 1$ . Choose one such data set. Then add another point sufficiently far away from all other points in the data set, to increase the between-cluster variance, making  $VR_k(X)$  arbitrarily high. Place the added point sufficiently far so that it has its own cluster in any optimal  $k$ -means and any optimal  $(k-1)$ -means clustering. Therefore, the remaining points have the same clustering in an optimal  $k$ -means clustering as in an optimal  $(k-1)$ -means clustering. The singleton cluster does not effect the  $k$ -means loss or the  $(k-1)$ -means loss, and therefore  $S_k(X) = S_{k-1}(X) \geq \epsilon$ . □

## 7.2 Worst pair ratio versus variance ratio

We show that worst pair ratio implies variance ratio. We then show that variance ratio does not imply worst pair ratio.

**THEOREM 11.** For any  $x, y > 0$ ,  $k \geq 2$ , there exists a data set  $X \subseteq \mathbf{R}^m$  such that  $VR_k(X) \geq y$  and  $WPR_k(X) \leq x$ .

**PROOF.** We first describe an example for  $k \geq 3$ . Consider a set with the following structure of the  $k$ -means optimal clustering: small clusters all but two of which are far apart from each other, so that the between-cluster variance is high and the within-cluster variance is low. Exactly two of the clusters, say clusters  $A$  and  $B$ , are close to each other, and some other cluster has sufficiently high width, so that  $WPR_k(X) = x$ . However, by moving all pairs of clusters, except for  $A$  and  $B$ , further away from each other we increase  $VR_k(X)$  arbitrarily without increasing  $WPR_k(X)$ . Note that we can create data sets where the  $k$ -means optimal clustering has this structure by adjusting the density of the clusters.

For 2-clusterings, consider the example in Figure 2. By increasing the radius of the circles, we can make worst pair ratio arbitrarily low. Moving the circles further away from each other makes variance ratio arbitrarily high, since most within cluster pairs are much closer to each other than most between cluster pairs. To compensate for the larger number of points on the line as the circles move further away from each other, we can increase the density of the circles. □

We prove that when worst pair ratio is sufficiently good, it gives a lower bound to variance ratio clusterability. First, we present an alternative formula for between-cluster variance.

LEMMA 4. Let  $C = \{X_1, X_2, \dots, X_k\}$  be an optimal  $k$ -means clustering of  $X \subseteq \mathbf{R}^m$ , where  $c_i = \text{center-mass}(X_i)$ . Then

$$B_k(X) = \frac{1}{|X|^2} \sum_{i \neq j} |X_i| |X_j| \|c_i - c_j\|^2.$$

PROOF. The between-cluster variance of data set  $X$  is defined as  $\sum_{i=1}^k \frac{|X_i|}{|X|} \|c_i - c\|^2$ , where  $c = \text{center-mass}(X)$ . For a set  $P \subseteq S$  where  $S \subseteq \mathbf{R}^m$ ,  $\sum_{a,b \in P} \|a - b\|^2 = |P| \sum_{a \in P} \|a - p\|^2$ , where  $p = \text{center-mass}(P)$ . The between-cluster variance of  $X$  is the same as the between-cluster variance of a data set  $\bar{X}$ , having exactly  $|X_i|$  points at position  $c_i$  for all  $i \in \{1, 2, \dots, k\}$ . The between-cluster variance of  $\bar{X}$  is  $\sum_{i=1}^k \frac{|X_i|}{|\bar{X}|} \|c_i - c\|^2 = \frac{1}{|\bar{X}|} \sum_{x \in \bar{X}} \|x - c\|^2 = \frac{1}{|\bar{X}|^2} \sum_{a,b \in \bar{X}} \|a - b\|^2 = \frac{1}{|\bar{X}|^2} \sum_{i \neq j} |X_i| |X_j| \|c_i - c_j\|^2$ . Since the between-cluster variance of  $\bar{X}$  is the same as the between-cluster variance of  $X$ , the result holds.  $\square$

THEOREM 12. If  $WPR_k(X) > 1$ , then

$$VR_k(X) > \frac{1}{2n} (WPR_k(X))^2,$$

for all  $X \subseteq \mathbf{R}^m$  where  $n = |X|$ .

PROOF. By Lemma 4,

$$\begin{aligned} B_k(X) &= \frac{1}{n^2} \sum_{i \neq j} |X_i| |X_j| \|c_i - c_j\|^2 \\ &\geq \frac{1}{n^2} \cdot \frac{n}{2} \cdot \text{split}_k(X)^2 \\ &= \frac{1}{2n} \text{split}_k(X)^2 \end{aligned}$$

For within-cluster variability,  $W_k(X) = \frac{1}{n} \sum_{i=1}^k \sum_{a \in X_i} \|a - c_i\|^2 \leq \frac{1}{n} \sum_{i=1}^k \sum_{a \in X_i} \text{width}_k(X)^2 = \text{width}_k(X)^2$ . Therefore,  $VR_k(X) = \frac{B_k(X)}{W_k(X)} \geq \frac{1}{2n} (WPR_k(X))^2$ .  $\square$

### 7.3 Worst pair ratio versus separability

We prove that separability does not imply worst pair ratio. We then show that worst pair ratio implies separability.

THEOREM 13. For any  $x > 0$ ,  $0 < \epsilon < 1$ , and  $k \geq 2$ , there exists a data set  $X$  such that  $S_k(X) \leq \epsilon$  and  $WPR_k(X) \leq x$ .

PROOF. By Lemma 4, there exist data sets with arbitrarily low worst pair ratio and arbitrarily high variance ratio for  $k = 2$ . By Theorem 2, variance ratio and separability are equivalent for  $k = 2$ . Therefore, there is a data set with arbitrarily good separability and arbitrarily poor worst pair ratio for  $k = 2$ . To generalize the example for  $k \geq 3$ , add  $k - 2$  points sufficiently far away from the remaining points and from each other so that they make distinct clusters in the optimal  $k$ -means clustering.  $\square$

We show that worst pair ratio implies separability.

THEOREM 14.  $S_k(X) \geq \frac{n}{2} WPR_k(X)^2$ , for all data sets  $X \subseteq \mathbf{R}^m$  where  $|X| = n$ ,  $k \geq 2$ .

PROOF. It can be shown that the  $k$ -means loss function is equal to

$$k\text{-means}(\{X_1, X_2, \dots, X_k\}) = \sum_{i=1}^k \frac{1}{|X_i|} \sum_{\{x,y\} \subseteq X_i} \|x - y\|^2.$$

Let  $C = \{X_1, X_2, \dots, X_k\}$  be the  $k$ -means optimal clustering of  $X$  maximizing the split over width ratio. By Lemma 1, this clustering is unique. Then,  $OPT_k(X) \leq \sum_{i=1}^k \frac{1}{|X_i|} \sum_{\{x,y\} \subseteq X_i} \text{width}_k(X)^2 < \frac{n}{2} \text{width}_k(X)^2$ . Any  $(k - 1)$ -clustering has at least one within-cluster distance that is at least  $\text{split}_k(X)$  - otherwise, it is a subdivision of  $C$  and therefore has at least  $k$  clusters. Therefore,  $S_k(X) = \frac{OPT_k(X)}{OPT_{k-1}(X)} \leq \frac{n}{2} WPR_k(X)^2$ .

$\square$

### 7.4 Center perturbation versus the other notions

Center perturbation with the  $k$ -means loss function does not imply, and is not implied by the other notions of clusterability; namely, worst pair ratio, variance ratio, and separability. This can be shown using the following property.

DEFINITION 8 (SCALE INVARIANCE). If for all data sets  $A = \{a_1, a_2, \dots, a_n\}$  and  $B = \{b_1, b_2, \dots, b_n\}$ , such that, for all  $i, j$ ,  $\|a_i - a_j\| = c \|b_i - b_j\|$ , for some  $c > 0$ , we have that  $\alpha(A) = \alpha(B)$ , then notion  $\alpha$  is scale-invariant.

We can show that worst pair ratio, variance ratio, and separability are scale-invariant. This means that we can scale a data set (by any non-zero value) while preserving its clusterability by these notions. However, center perturbation is not scale-invariant. Using these observations, it is easy to show that center perturbation is inconsistent with the other three notions of clusterability.

## 8. CONCLUSIONS

In this work, we present a theoretical study of clusterability. We survey some previous notions of clusterability, and present a new notion that captures clustering robustness to center perturbations. Our comparison of these notions shows that, although they attempt to measure the same intuitive property, they are pairwise inconsistent.

Our analysis reveals an interesting property common to these notions of clusterability: The more clusterable is a data set, the easier it is, computationally, to find a provably near-optimal clustering of the data. It is intriguing to figure out how broad this phenomenon is. In particular, it is an interesting challenge to the research community to obtain similar results with other natural notions of clusterability. Such results may help understand when certain clustering algorithms perform well, and show that when they fail to

produce satisfactory clusterings, it is due to insufficient clustering structure in the data.

We explore the hardness of determining the degree of clusterability of a given data set using previous notions of clusterability. Our analysis shows that, for the notions that recognize a wide range of well-clustered data, this is an NP-hard problem. For future work, it would be interesting to explore this problem in the more flexible framework of property testing (for more details, see Section 6).

## 9. REFERENCES

- [1] S. Ben-David, N. Eiron, and H.-U. Simon. “The Computational Complexity of Densest Region Detection.” *J. Comput. Syst. Sci.* 64(1): 22-47, 2002.
- [2] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. “Clustering Large Graphs via the Singular Value Decomposition.” *Machine Learning*, 56:9-33, 2004.
- [3] S. Epter, M. Krishnamoorthy, and M. Zaki. “Clusterability detection and initial seed selection in large datasets.” Technical Report 99-6, Rensselaer Polytechnic Institute, Computer Science Dept., Rensselaer Polytechnic Institute, Troy, NY 12180, 1999.
- [4] O. Goldreich, S. Goldwasser and D. Ron. “Property testing and its connection to learning and approximation.” 37th Annual Symposium on Foundations of Computer Science, *IEEE Comput. Soc. Press*, New York (1996) p. 339-348.
- [5] Marina Meila. “Comparing clusterings – an information based distance.” *Journal of Multivariate Analysis archive*, 98, 5(2007):873-895.
- [6] R. Ostrovsky, Y. Rabani, L.J. Schulman, and C. Swamy. “The Effectiveness of Lloyd-Type Methods for the  $k$ -Means Problem.” *Foundations of Computer Science*, 2006. FOCS '05. 47th Annual IEEE Symposium. Berkeley, CA, USA, Oct. 2006. pp. 165-176.
- [7] G. Pisier, “Remarques sur un résultat non publié de B. Maurey.” *Séminaire d'Analyse Fonctionnelle*. 1980–1981, École Polytechnique, Centre de Mathématiques, Palaiseau, V.1 V.12, 1981.
- [8] Bin Zhang. “Dependence of Clustering Algorithm Performance on Clustered-ness of Data.” Technical Report, 20010417. Hewlett-Packard Labs, 2001.