Clustering in the Presence of Background Noise

Shai Ben-David SHAI@CS.UWATERLOO.CA
David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1 CANADA

Nika Haghtalab NIKA@CMU.EDU
Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Abstract

We address the problem of noise management in clustering algorithms. Namely, issues that arise when on top of some cluster structure the data also contains an unstructured set of points. We consider how clustering algorithms can be "robustified" so that they recover the cluster structure in spite of the unstructured part of the input. We introduce some quantitative measures of such robustness that take into account the strength of the embedded cluster structure as well as the mildness of the noise subset. We propose a simple and efficient method to turn any centroid-based clustering algorithm into a noise-robust one, and prove robustness guarantees for our method with respect to these measures. We also prove that more straightforward ways of "robustifying" clustering algorithms fail to achieve similar guarantees.

1. Introduction

Clustering is (usually) aimed to detect groups of similar objects in given datasets. Many common clustering algorithms output a *partition* of the input set. However, it is often the case that datasets that one wishes to cluster contain, on top of groups of similar objects, a significant subset that is unstructured. Such a subset, that may be referred to as noise, tends to disrupt clustering algorithms and make it difficult to detect the cluster structure of the remaining domain points. This problem can be viewed as a *noise robustness* issue.

Are there useful clustering algorithms that are noise robust (w.r.t addition of unstructured data points)? Short reflection reveals that the noise robustness of an algorithm is closely related to its sensitivity to the input data. As an extreme example, it is easy to achieve noise robustness by ignoring the input data and always returning a fixed output. Ackerman et al. (2013) provides some formal tradeoffs between these two desired properties. Roughly stated, their results (for example, Theorem 4.3) show that no algorithm can be both noise robust and responsive to cluster structure of the data (in the language of (Ackerman et al., 2013) these properties are called *robustness to oligarchies* and *separability-detecting*). However, those results consider applying an algorithm with a fixed number-of-clusters parameter. This paper addresses the possibility of overcoming those pessimistic results by allowing clustering algorithms to add to the set of clusters they output an extra subset, serving as a "garbage collector".

An important aspect of clustering, that distinguishes it from other major learning tasks, like classification prediction, is the wide variability of input-output behavior among common clustering algorithms. Different clustering applications employ very different clustering algorithms and there is no single clustering algorithm that is suitable for all. Consequently, solutions to fundamental clustering challenges, like the tradeoff between sensitivity to the input and noise robustness, should be modular in the sense of being applicable to a variety of clustering algorithms. In this paper we propose such a modular solution. In Section 4, we consider a method to transform any centroid-based clustering algorithm to one that outputs a set of clusters augmented by a distinct noise bin. We show that our proposed paradigm can be achieved by employing a simple and efficiently implementable transformation of the input data, after which users can apply any centroid-based clustering algorithm.

Yet another contribution of this paper is the introduction of quantitative measures of noise robustness (Section 5). We consider three aspects of noise robustness for centroid-based clustering algorithms; the degree by which noise can affect the location of the centers of the clusters (or the archetypal cluster representatives), the effect of noise on

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

the cost of the clustering solution (or, the value of the clustering objective function) and the similarity between the clustering of the un-noised data, to its clustering in the presence of the noise.

In our results, we consider a scenario in which the input dataset $\mathcal{X}$ consists of two part: a clusterable subset $\mathcal{I}$, which is also called the un-noised data, and an added noise set $\mathcal{X} \setminus \mathcal{I}$ (the identity of which is not known to the clustering algorithm). We consider two clustering algorithms, the original one, $\mathcal{A}$, and its "robustified" transformation $R_p(\mathcal{A})$ that is obtained by using our paradigm with a tunable parameter p. We examine to what extent well clusterability of $\mathcal{I}$ and mildness properties of $\mathcal{X} \setminus \mathcal{I}$ (in terms of the size and/or diameter of this set, relative to that of $\mathcal{I}$) affect the similarity between the clusterings $\mathcal{A}(\mathcal{I})$ and $R_p(\mathcal{A})(\mathcal{X})$ restricted to $\mathcal{I}$. In Section 6, we prove that our paradigm makes the algorithms noise-robust without sacrificing much of their ability to detect clear cluster structures. The degree of noise-robustness that our paradigm achieves depends on a parameter that can be tuned by the user, depending on the level and structure of the noise expected in the input data. On the other hand, in Section 7, we show that a simple transformation in which $R_p(\mathcal{A})$ is the original algorithm $\mathcal{A}$ with an increased number of output clusters (the extra clusters may be used to handle noise) does not enjoy the same robustness guarantees. In Section 8, we further demonstrate the gap between these two paradigms and use experiments to confirm our theoretical results.

Note that in the interest of space, the proofs of some theorems and lemmas are moved to the supplementary material.

2. Related Work

Previous work on the robustness of clustering methods have mainly focused on two directions. First, developing measures of robustness and examining the performance of traditional clustering algorithms based on these measures. Second, developing clustering algorithms that are robust to noise and outliers.

Various measures of robustness have been developed for examining the robustness characteristics of clustering algorithms to noise (Donoho, 1982; Hampel, 1971; Hennig, 2008). These measures have been used to demonstrate the lack of robustness of some traditional algorithms, where the number of clusters is fixed (Ackerman et al., 2013; Hennig, 2008). That is, they consider the scenario in which a clustering algorithm is used with the same number-of-clusters parameter for both the clean input and the input after the addition of noisy points. In this work, we allow some extra clusters to be used (so as to accommodate the added noisy data points) and introduce and analyze noise robustness measures that allow such flexibility. We show that using the added flexibility of our paradigm, we can overcome some of the limitations that the above papers deemed inevitable (without allowing extra noise-accommodating clusters).

Several methods have been suggested for clustering a potentially noisy dataset (Cuesta-Albertos et al., 1997; Dave, 1993; Ester et al., 1996). One interesting work is the development of the concept of a "noise cluster" in a fuzzy setting by Dave (1991; 1993). In this work, we introduce a novel paradigm for "robustifying" any center-based clustering algorithm. We show that our paradigm generalizes a non-fuzzy variation of the algorithm introduced by Dave (1991). In addition, we prove noise robustness guarantees for our proposed paradigm, guarantees that were not proven in any of the earlier works we are aware of.

Some of the earlier work on noise-robustness of clustering algorithms proposes the use of *trimmming*. Trimming is the natural approach to separating clusterable parts of the data from unstructured ones by fixing some noise-set size (say, as some fraction, α, of the data size) and let the algorithm find the "least structured" α fraction of an input dataset and discard it before applying a clustering algorithm (Cuesta-Albertos et al., 1997; García-Escudero et al., 2008; García-Escudero & Gordaliza, 1999). However, such methods suffer from exponential computational complexity, or have to be compromised for efficient heuristic searches that have no performance guarantees. In this work, we avoid this issue, by developing a paradigm that is of comparable computational complexity to traditional clustering algorithms.

Clustering has been the subject of many other theoretical studies. Ackerman & Ben-David (2009), Balcan et al. (2009), and Bilu & Linial (2012) examined the effect of robustness (to perturbations of the input) on computational complexity of clustering. Ben-David et al. (2006) examined situations where the clustering of the input is similar to the clustering of a random subset of it. Ailon et al. (2009) approximated the k-means objective using extra centers. Our work has a different focus. Our goal is to find clustering algorithms that yield a good clustering on the clusterable portion of a noisy input. Therefore, the questions that we address, our setting, and approach are different from the above studies.

Discussing the details of more relevant previous work requires the definition of few notations, hence, it is delayed to relevant sections.

3. Preliminaries

For a set $\mathcal{X}$ and integer $k \geq 1$, a *k-clustering* of $\mathcal{X}$ is a partition $\mathcal{C} = \{C_1, \ldots, C_k\}$ of $\mathcal{X}$ into k disjoint sets. For a clustering $\mathcal{C}$ of $\mathcal{X}$ and points $x, y \in \mathcal{X}$, we say $x \sim_{\mathcal{C}} y$, if x and y are in the same cluster, otherwise $x \not\sim_{\mathcal{C}} y$. For sets $\mathcal{X}$

and $\mathcal{I}$ such that $\mathcal{I} \subseteq \mathcal{X}$, and a clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$ of $\mathcal{X}$, we denote the *restriction of* $\mathcal{C}$ *to* $\mathcal{I}$ by $\mathcal{C}|\mathcal{I} = \{C_1 \cap \mathcal{I}, \ldots, C_k \cap \mathcal{I}\}$.

For two clusterings $\mathcal{C}$ and $\mathcal{C}'$ of the set $\mathcal{X}$, we define the distance between them as $\Delta(\mathcal{C}, \mathcal{C}')$, the fraction of pairs of domain points which are in the same cluster under $\mathcal{C}$ but in different clusters under $\mathcal{C}'$ or vice-versa. Equivalently, $\Delta(\mathcal{C}, \mathcal{C}') = 1 - i_R(\mathcal{C}, \mathcal{C}')$, where i_R is the Rand index (Rand, 1971). Δ satisfies the triangle inequality.

Let d be a symmetric distance function defined over $\mathcal{X}$ with $d(x, x) = 0$ for all $x \in \mathcal{X}$, and satisfying the triangle inequality. The *diameter* of $\mathcal{X}$, indicated by $\operatorname{diam}(\mathcal{X})$, is defined as the maximum distance between two elements of $\mathcal{X}$. For a clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$, the diameter of $\mathcal{C}$ is defined as $\max_{C_i \in \mathcal{C}} \operatorname{diam}(C_i)$. The *radius* of $\mathcal{X}$ is shown by $\operatorname{rad}(\mathcal{X}) = \min_{c \in \mathcal{X}} \max_{x \in \mathcal{X}} d(c, x)$. Clustering $\mathcal{C}$ is *σ-separable* for $\sigma \geq 1$, if $\min_{x \not\sim_{\mathcal{C}} y} d(x, y) > \sigma \cdot \max_{x \sim_{\mathcal{C}} y} d(x, y)$. Clustering $\mathcal{C}$ is (ρ_1, ρ_2)*-balanced* if for all $i \leq k$, $\rho_1|\mathcal{X}| \leq |C_i| \leq \rho_2|\mathcal{X}|$. We use *$\rho$-balanced* to refer to a clustering that is $(0, \rho)$-balanced.

A *clustering algorithm* is a function $\mathcal{A}$ that takes as input $\mathcal{X}$ and d and returns a clustering $\mathcal{C}$ of $\mathcal{X}$. An *objective function* is a function that takes as input a clustering and outputs a non-negative cost associated with it. An *objective-based clustering algorithm* is an algorithm that produces a clustering that minimizes a specified objective function.

Consider an input set $\mathcal{X}$ drawn from a given space E with distance function d. Throughout this work, let $g : \mathbb{R}^+ \to \mathbb{R}^+$ be any continuous, increasing, and unbounded function. The (k, g)*-centroid algorithm* is an objective-based clustering algorithm with function

$$\Lambda^g_{E,d}(\{C_1, \ldots, C_k\}) = \min_{\mu_1, \ldots, \mu_k \in E} \sum_{i=1}^{k} \sum_{x \in C_i} g(d(x, \mu_i))$$

We refer to μ_i as the *center* of cluster C_i and we define $\mu(x) = \arg\min_{\mu_i \in \{\mu_1, \ldots, \mu_k\}} d(x, \mu_i)$ to be the center closets to x. With a slight abuse of notation we can also define the (k, g)-centroid algorithm as the algorithm that chooses centers $\mu_1, \ldots, \mu_k$ that minimize

$$\Lambda^g_{\mathcal{X},E,d}(\mu_1, \ldots, \mu_k) = \sum_{x \in \mathcal{X}} g(d(x, \mu(x))$$

We remove $\mathcal{X}$ and E from the notation whenever they are clear from the context. Note that for $g(x) = x$ and $g(x) = x^2$, Λ^g_d refers to the k-medians and k-means cost function, respectively.

4. Robustifying Paradigms

We define parameterized robustifying paradigms that transform any clustering algorithm to an algorithm that is more robust to noise to the extent determined by a predefined parameter. We define two robustifying paradigms. Moreover, we establish an equivalence between one of the paradigms and a generalization of an existing algorithm.

A *robustifying parameter*, p, denotes the degree to which an algorithm should be robustified to noise; For example, the number of extra clusters that can be used or a notion of distance beyond which a point is considered an outlier. A *robustifying paradigm*, $R_p(\cdot)$, is a function that takes a clustering algorithm $\mathcal{A}$ and returns a robustified clustering algorithm $R_p(\mathcal{A})$ based on the robustifying parameter p. We refer to $\mathcal{A}$ as the *ground* clustering algorithm of $R_p(\mathcal{A})$.

Since noise, unstructured data, and outlying structures are heterogeneous with respect to the existing data, outliers and noise groups can be considered as separate clusters. Therefore, some statisticians simply recommend increasing the number of clusters when dealing with noisy data (García-Escudero & Gordaliza, 1999). The next paradigm captures robustification as used in this practice.

Definition 1 (p-Increased Paradigm)**.** *The* p-Increased Paradigm *is a robustifying paradigm,* $RI_p(\cdot)$*, that takes as input a* (k, g)*-centroid algorithm and returns a* $(k + p, g)$*-centroid algorithm.*

The next paradigm is parameterized by the distance after which a point should be considered an outlier. To define this paradigm, we first introduce a class of algorithms. Given a space E and distance function d, the *δ-truncated* distance function corresponding to d is the function d' such that $d'(x, y) = \min\{\delta, d(x, y)\}$ for $x, y \in E$. The *(k, g)-δ-truncated algorithm* is an objective based algorithm that, given $\mathcal{X} \subseteq E$, first optimizes the function $\Lambda^g_{\mathcal{X},d'}(\mu'_1, \ldots, \mu'_k)$. For $j \leq k$, let $C'_j = \{x \in \mathcal{X} | j = \arg\min_i d(x, \mu'_i)$ and $d(x, \mu'_j) < \delta\}$ and $C'_{k+1} = \{x \in \mathcal{X} | \min d(x, \mu'_i) \geq \delta\}$. Then the (k, g)-δ-truncated algorithm returns the $(k + 1)$-clustering $\mathcal{C}' = \{C'_1, \ldots, C'_{k+1}\}$. We refer to μ'_i as the center of C'_i for $i \leq k$. With a slight abuse of notation, we define $\mu'(x) = \arg\min_{\mu_i \in \{\mu'_1, \ldots, \mu'_k\}} d(x, \mu'_i)$

Definition 2 (δ-Truncated Paradigm)**.** *The* δ-Truncated Paradigm *is a robustifying paradigm,* $RT_\delta(\cdot)$*, that takes as input a* (k, g)*-centroid algorithm and returns a* (k, g)*-*δ*-truncated algorithm.*

Dave (1991) developed the notion of the *noise prototype*, which is a point equidistant from all points in E. He then introduced a clustering algorithm that performs fuzzy $(k+1)$-means with one center fixed as the noise prototype. In the next definition, we provide a generalization of the non-fuzzy variation of Dave's algorithm for any centroid-based algorithm. Furthermore, we show that the class of algorithms produced in such way is equivalent to the class of algorithms that belong to the δ-Truncated paradigm.

Definition 3. *Let μ^* be defined such that for all $y \in E$, $d(y, \mu^*) = \delta$. The generalized (k, g)-δ-centroid algorithm is an objective-based algorithm with objective function $\Lambda^g_{\mathcal{X}, E\cup\{\mu^*\}, d}(\mu_1, \ldots, \mu_k, \mu^*)$.*

We refer to (k, g)-δ-centroid as δ-*k-median* and δ-*k-means* when $g(x) = x$ and $g(x) = x^2$, respectively.

Theorem 1. *Clustering $\mathcal{C}$ is an optimal (k, g)-δ-centroid clustering of $\mathcal{X}$ if and only if $\mathcal{C}$ is an optimal (k, g)-δ-truncated clustering of $\mathcal{X}$.*

Proof. For convenience, let $\mu_{k+1} = \mu^*$ and $d'(x, y) = \min\{\delta, d(x, y)\}$. We show that the (k, g)-δ-centroid clustering with centers $\mu_1, \ldots, \mu_k, \mu^*$ and the (k, g)-δ-truncated clustering with centers $\mu_1, \ldots, \mu_k$ have the same objective value.

$$\begin{aligned}\Lambda^g_{(E,d')}(\{\mu_1, \ldots, \mu_k\}) &= \sum_{y\in\mathcal{X}} g(\min_{i\in[k]}\{\min\{\delta, d(y, \mu_i)\}\}) \\ &= \sum_{y\in\mathcal{X}} g\left(\min_{i\in[k]}\{\min\{d(y, \mu_{k+1}), d(y, \mu_i)\}\}\right) \\ &= \sum_{y\in\mathcal{X}} g\left(\min_{i\in[k+1]} d(y, \mu_i)\right) \\ &= \Lambda^g_{(E\cup\{\mu^*\},d)}(\{\mu_1, \ldots, \mu_k, \mu^*\})\end{aligned}$$

Moreover, $\mu_1, \ldots, \mu_k, \mu^*$ can induce the same (k, g)-δ-centroid clustering as the (k, g)-δ-truncated clustering induced by $\mu_1, \ldots, \mu_k$. Therefore, $\mathcal{C}$ is an optimal (k, g)-δ-centroid clustering if and only if it is an optimal (k, g)-δ-truncated clustering. □

5. Measures of Robustness

In previous work, robustness to the addition of noise has been measured by comparing the output of the same algorithm on both the un-noised and noisy data. This approach leads to pessimistic results about the possibility of achieving noise robustness. One example of these results is the work of Ackerman et al. (2013), which shows that algorithms that are responsive to the structure of the data are not noise-robust. More precisely, Ackerman et al. (2013) define a k-clustering algorithm $\mathcal{A}$ to be σ-*separability-detecting* if for all $\mathcal{I}$, such that there exists a σ-separable k-clustering $\mathcal{C}$ of $\mathcal{I}$, $\mathcal{A}(\mathcal{I}) = \mathcal{C}$. Then, they show that for any σ-separability-detecting algorithm and any ρ, there is a ρ-balanced σ-separable set $\mathcal{I}$ that is not robust to a noise set of size as small as k.

This approach does not reflect the way clustering is done in practice. For clustering noisy data, one may take special provisions to handle the noise – provisions that are not needed when the data is known to be noise free. For example, if it is known that the input data has k clusters, one may allow the algorithm to create more clusters in an attempt to separate the noise.

One of the take home messages of this paper is that by revising the previous approach to reflect a more practical setting, we can overcome those pessimistic results. To this end, our measures of robustness compare the output of a robustified algorithm (one that can use extra clusters) on noisy data to the output of its corresponding ground algorithm on the unnoised data. More precisely, let $\mathcal{A}$ denote any clustering algorithm (the ground clustering) and $R_p(\cdot)$ denote a robustifying paradigm with parameter p. We use $\mathcal{A}' = R_p(\mathcal{A})$ to denote the robustified algorithm corresponding to $\mathcal{A}$. Given $\mathcal{I} \subseteq \mathcal{X}$, $\mathcal{A}(\mathcal{I})$ denotes the clustering of $\mathcal{I}$ using the ground algorithm, and $\mathcal{A}'(\mathcal{X})$ denotes the clustering of $\mathcal{X}$ by the robustified algorithm. We consider $\mathcal{I}$ to be robust (to $\mathcal{X} \setminus \mathcal{I}$) with respect to the $R_p(\mathcal{A})$ algorithm if certain properties of $\mathcal{A}(\mathcal{I})$ are preserved in $\mathcal{A}'(\mathcal{X})$. In the following definitions, for any $x \in \mathcal{X}$, we use $\mu(x)$ and $\mu'(x)$ to denote the centers of $\mathcal{A}(\mathcal{I})$ and $\mathcal{A}'(\mathcal{X})$, respectively, to which x belongs.

Cluster centers are commonly used to compress and represent data. The distances between points and their corresponding centers can be viewed as the distortion of such a compression. Therefore, it is essential to have clustering algorithms where this value does not grow significantly in the presence of noise. The next definition measures how much this distortion is affected by the addition of noise to the input data.

Definition 4 (α-distance-robust)**.** *A subset $\mathcal{I} \subseteq \mathcal{X}$ is α-distance-robust with respect to $\mathcal{A}'$ if for all $y \in \mathcal{I}$, $d(y, \mu'(y)) \leq d(y, \mu(y)) + \alpha$.*

An algorithm is considered robust, if it separates the input using some low-cost clusters that cover the un-noised data. In the next definition, we capture this property by computing minimal cost of a subset of clusters that covers at least $|\mathcal{I}|$ points in total.

Definition 5 (β-cost-robust)**.** *Let Λ be an objective (cost) function. $\mathcal{I} \subseteq \mathcal{X}$ is* β-cost-robust *with respect to $\mathcal{A}'$ for Λ, if there exists $\mathcal{C}^* \subseteq \mathcal{A}'(\mathcal{X})$, such that $|\bigcup \mathcal{C}^*| \geq |\mathcal{I}|$ and $\Lambda(\mathcal{C}^*) \leq \Lambda(\mathcal{A}(\mathcal{I})) + \beta$.*

Note that if $\mathcal{A}'(\mathcal{X}) = \mathcal{A}(\mathcal{I}) \cup \{\mathcal{X} \setminus \mathcal{I}\}$, then $\mathcal{I}$ is 0-cost-robust to $\mathcal{A}'$.

The next definition measures the degree to which noise affects the structure of a clustering.

Definition 6 (γ-robust)**.** *$\mathcal{I} \subseteq \mathcal{X}$ is γ-robust with respect to algorithm $\mathcal{A}'$, if $\Delta(\mathcal{A}'(\mathcal{X})|\mathcal{I}, \mathcal{A}(\mathcal{I})) \leq \gamma$.*

Lemma 1. *Given a (k, g)-centroid algorithm $\mathcal{A}$ and parameter δ, let $\mathcal{A}' = RT_\delta(\mathcal{A})$. For any $\mathcal{I} \subseteq \mathcal{X}$, such that for all $y \in \mathcal{I}$, $d(y, \mu'(y)) \leq \delta$, $\mathcal{I}$ is $g(\delta)|\mathcal{X} \setminus \mathcal{I}|$-cost-robust to $\mathcal{X} \setminus \mathcal{I}$ with respect to $RT_\delta(\mathcal{A})$ for the Λ^g_d cost function.*

Proof. Refer to the supplementary material. □

6. Robustness of Our Paradigm

In this section, we show guaranteed robustness results for the δ-Truncated paradigm, $RT_\delta(\cdot)$. We prove robustness, distance-robustness, and cost-robustness based on several properties of the underlying structures of $\mathcal{I}$ and mildness properties of $\mathcal{X} \setminus \mathcal{I}$:

RADIUS OF THE BALL COVERING $\mathcal{I}$

The following results guarantee distance-robustness and cost-robustness for the δ-Truncated paradigm. Theorem 2 derives values of δ that render $\mathcal{I}$ robust with respect to the δ-Truncated algorithm, based on the radius of $\mathcal{I}$ and the signal-to-noise ratio of $\mathcal{X}$.

Theorem 2. *For any k and g, let $\mathcal{A}$ be the (k,g)-centroid algorithm. For all $\mathcal{I} \subseteq \mathcal{X}$, let $r = \mathrm{rad}(\mathcal{I})$ and $\lambda = |\mathcal{I}| \,/\, |\mathcal{X} \setminus \mathcal{I}|$ (signal-to-noise ratio). Then for any $\delta \in \left[4r, g^{-1}(\lambda(g(2r) - g(r)))\right)$, $\mathcal{I}$ is $4r$-distance-robust and $g(\delta)|\mathcal{X} \setminus \mathcal{I}|$-cost-robust with respect to $RT_\delta(\mathcal{A})$ for the Λ_d^g cost function.*

Proof. Let $\mathcal{A}' = RT_\delta(\mathcal{A})$ and let $\mathcal{A}'(\mathcal{X}) = \mathcal{C}'$. For all $x \in \mathcal{X}$, let $\mu'(x)$ denote the closest center of $\mathcal{C}'$ to x. Assume on the contrary that there exists $y \in \mathcal{I}$ such that $d(y, \mu'(y)) > 4r$, then for any $y' \in \mathcal{I}$, $d(y', \mu'(y')) > 2r$. Therefore, $\Lambda_{d'}^g(\mathcal{C}') \geq |\mathcal{I}| \cdot g(2r)$. For any clustering $\mathcal{C}''$ that has a center at the center of the r-ball that covers $\mathcal{I}$, $\Lambda_{d'}^g(\mathcal{C}'') \leq |\mathcal{I}| \cdot g(r) + |\mathcal{X} \setminus \mathcal{I}| \cdot g(\delta)$. By the choice of δ, $\Lambda_{d'}^g(\mathcal{C}'') < \Lambda_{d'}^g(\mathcal{C}')$, so $\mathcal{C}'$ is not optimal. Hence, $\mathcal{I}$ is $4r$-distance-robust to $\mathcal{X} \setminus \mathcal{I}$ with respect to $RT_\delta(\mathcal{A})$. Using Lemma 1, $\mathcal{I}$ is also $g(\delta)|\mathcal{X} \setminus \mathcal{I}|$-cost-robust with respect to $RT_\delta(\mathcal{A})$ for cost function Λ_d^g. □

Note that Theorem 2 implies that if $\mathcal{I}$ has radius r and signal-to-noise ratio λ, then for any $\delta \in [4r, r\sqrt{3\lambda})$, $\mathcal{I}$ is $4r$-distance-robust and $\delta^2|\mathcal{X} \setminus \mathcal{I}|$-cost-robust to δ-k-means.

UNDERLYING STRUCTURE OF $\mathcal{I}$

The following results guarantee robustness, distance-robustness and cost-robustness for the δ-Truncated paradigm. Theorem 3 derives values of δ that render $\mathcal{I}$ robust with respect to the δ-Truncated algorithm, based on the underlying structure of $\mathcal{I}$ and the signal-to-noise ratio of $\mathcal{X}$.

Theorem 3. *For any k and g, let $\mathcal{A}$ be the (k,g)-centroid algorithm. Assume that $\mathcal{I} \subseteq \mathcal{X}$ has the following properties: $\mathcal{I}$ can be covered by a (ρ_1, ρ_2)-balanced set of balls, called $B_1, \ldots, B_k$, such that for all $i \leq k$, $\mathrm{rad}(B_i) \leq r$, and for all $i \neq j$, the centers of B_i and B_j are at least $\nu > 4r + 2g^{-1}(\frac{\rho_1+\rho_2}{\rho_1} g(r))$ apart. Let $\delta \in \left[\frac{\nu}{2}, g^{-1}\left(\frac{|\mathcal{I}|}{|\mathcal{X}\setminus\mathcal{I}|}(\rho_1 g(\frac{\nu}{2} - 2r) - (\rho_1+\rho_2)g(r))\right)\right)$, then $\mathcal{I}$ is*

- *0-robust*
- *$min\{\frac{\nu}{2}, g^{-1}(g(\frac{\nu}{2} - 2r) - \frac{\rho_2}{\rho_1} g(r)) + 2r\}$-distance-robust*
- *$g(\delta)|\mathcal{X} \setminus \mathcal{I}|$-cost-robust*

with respect to $RT_\delta(\mathcal{A})$ for the Λ_d^g cost function.

Note that Theorem 3 implies that if $\mathcal{I} \subseteq \mathcal{X}$ can be covered by two balls of radius r whose centers are $10r$ apart, and each covers half of $\mathcal{I}$, and if there is 5% noise, then for any $\delta \in [5r, 8.5r)$, $\mathcal{I}$ is 0-robust, $4r$-distance-robust, and $\delta|\mathcal{X}\setminus\mathcal{I}|$-cost-robust to $RT_\delta(\mathcal{A})$ for the 2-medians function.

We need the next two lemmas to prove Theorem 3. Lemmas 2 and 3 examine the output of the (k,g)-δ-truncated and (k,g)-centroid algorithms when the un-noised data has a well-clusterable underling pattern.

Lemma 2. *Let $\mathcal{A}$ be the (k,g)-centroid algorithm. Assume that $\mathcal{I} \subseteq \mathcal{X}$, $B_1, \ldots, B_k$, and δ are as defined in Theorem 3. Let $\mathcal{A}' = RT_\delta(\mathcal{A})$, then $\mathcal{A}'(\mathcal{X})|\mathcal{I} = \{B_1, \ldots, B_k, \emptyset\}$. Furthermore, for all $y \in \mathcal{I}$, $d(y, \mu'(y)) \leq min\{\nu/2, g^{-1}\left(g(\frac{\nu}{2} - 2r) - \frac{\rho_2}{\rho_1} g(r)\right) + 2r\}$.*

Proof. The conditions of Theorem 3 state that $\mathcal{I}$ can be covered by a (ρ_1, ρ_2)-balanced set of balls, called $\mathcal{B} = \{B_1, \ldots, B_k\}$, such that $\mathrm{rad}(B_i) \leq r$ for all $i \leq k$, and for all $i \neq j$, the centers of B_i and B_j are at least $\nu > 4r + 2g^{-1}(\frac{\rho_1+\rho_2}{\rho_1} g(r))$ apart. Moreover, δ is assumed to be in the range $\left[\frac{\nu}{2}, g^{-1}\left(\frac{|\mathcal{I}|}{|\mathcal{X}\setminus\mathcal{I}|}(\rho_1 g(\frac{\nu}{2} - 2r) - (\rho_1+\rho_2)g(r))\right)\right)$

For $i \leq k$ let b_i represent the center of B_i and D_i represent a ball of radius $\frac{\nu}{2} - r$ centered at b_i. Let $\mathcal{A}'(\mathcal{X}) = \mathcal{C}'$ with centers, $\mu_1', \ldots, \mu_k'$ that minimize $\Lambda_{d'}^g$. Let $\mathcal{D}_1 = \{D_i | D_i \text{ does not cover any } \mu_j'\}$ and $\mathcal{D}_2 = \{D_i | D_i \text{ covers more than one } \mu_j'\}$. Since $D_1, \ldots, D_k$ are pairwise disjoint, $|\mathcal{D}_1| \geq |\mathcal{D}_2|$. Assume in search of a contradiction that $\mathcal{D}_1 \neq \emptyset$. For any $D_i \in \mathcal{D}_1$, for all $y \in D_i$, $d(y, \mu'(y)) \geq \frac{\nu}{2} - 2r$. Consider the following set of $\mu_1'', \ldots, \mu_k''$: If D_j includes exactly one center, μ_i', then let $\mu_j'' = \mu_i'$, otherwise $\mu_j'' = b_j$.

$$
\begin{aligned}
\Lambda_{\mathcal{X},d'}^g(\mu_1'', \ldots, \mu_k'') &\leq \Lambda_{\mathcal{X},d'}^g(\mu_1', \ldots, \mu_k') \\
&\quad + \sum_{D_i \in \mathcal{D}_1} \sum_{y \in B_i} [g(d'(y, \mu''(y))) - g(d'(y, \mu'(y)))] \\
&\quad + \sum_{D_i \in \mathcal{D}_2} \sum_{y \in B_i} [g(d'(y, \mu''(y))) - g(d'(y, \mu'(y)))] \\
&\quad + \sum_{y \in \mathcal{X}\setminus\mathcal{I}} g(d'(y, \mu''(y))) \\
&\leq \Lambda_{\mathcal{X},d'}^g(\mu_1', \ldots, \mu_k') + |\mathcal{D}_1||\mathcal{I}|\rho_1 \left(g(r) - g(\frac{\nu}{2} - 2r)\right) \\
&\quad + |\mathcal{D}_2||\mathcal{I}|\rho_2 g(r) + |\mathcal{X} \setminus \mathcal{I}| g(\delta) \\
&\leq \Lambda_{\mathcal{X},d'}^g(\mu_1', \ldots, \mu_k') \\
&\quad + |\mathcal{D}_1||\mathcal{I}| \left((\rho_1+\rho_2)g(r) - \rho_1 g(\frac{\nu}{2} - 2r)\right) + |\mathcal{X} \setminus \mathcal{I}| g(\delta)
\end{aligned}
$$

$< \Lambda^g_{\mathcal{X},d'}(\mu'_1, \ldots, \mu'_k)$

This forms a contradiction, so without loss of generality let every D_i cover a center μ'_i. For $i \neq j$ and for all $y \in B_i$, $d(y, \mu'_i) \leq \frac{\nu}{2} < d(y, \mu'_j)$. Therefore, $\mathcal{A}'(\mathcal{X})|\mathcal{I} = \{B_1, \ldots, B_k, \emptyset\}$.

For every $C'_i \in \mathcal{C}'$, $|B_i| \cdot \min_{y \in B_i} g(d(y, \mu'_i)) \leq |B_i| g(r) + |C_i \setminus \mathcal{I}| g(\delta)$. Therefore, there exists $y \in B_i$, such that

$$\begin{aligned} g(d(y, \mu'_i)) &\leq g(r) + \frac{|\mathcal{X} \setminus \mathcal{I}|}{|B_i|} g(\delta) \\ &\leq g(r) + \frac{|\mathcal{I}|}{|B_i|} \left(\rho_1 g(\frac{\nu}{2} - 2r) - (\rho_1 + \rho_2) g(r) \right) \\ &\leq g(\frac{\nu}{2} - 2r) - \frac{\rho_2}{\rho_1} g(r) \end{aligned}$$

Hence, for all $y \in \mathcal{I}$, $d(y, \mu'(y)) \leq min\{\nu/2, 2r + g^{-1}\left(g(\frac{\nu}{2} - 2r) - \frac{\rho_2}{\rho_1} g(r)\right)\}$. □

Lemma 3. *Let $\mathcal{A}$ be the (k, g)-centroid algorithm. For any $\mathcal{I}$, if it can be covered with a (ρ_1, ρ_2)-balanced set of k balls, called $\mathcal{B}$, where each ball has radius r and the centers of two different balls are at least $\nu > 4r + 2g^{-1}(\frac{\rho_1 + \rho_2}{\rho_1} g(r))$ apart, then $\mathcal{A}(\mathcal{I}) = \mathcal{B}$.*

Proof. Refer to the supplementary material. □

Proof of Theorem 3. Lemma 2 shows that for every $y \in \mathcal{I}$, $d(y, \mu'(y)) \leq \delta$. By Lemma 1, $\mathcal{I}$ is $g(\delta)|\mathcal{X} \setminus \mathcal{I}|$-cost-robust with respect to $RT_\delta(\mathcal{A})$ for cost function Λ^g_d.

Let $B_1, \ldots, B_k$ be the mentioned balls that cover $\mathcal{I}$. Lemma 2 and Lemma 3 show that $\mathcal{A}'(\mathcal{X})|\mathcal{I} = \{B_1, \ldots, B_k, \emptyset\}$ and $\mathcal{A}(\mathcal{I}) = \{B_1, \ldots, B_k\}$. Therefore, $\mathcal{I}$ is 0-robust with respect to $RT_\delta(\mathcal{A})$.

Lemma 2 shows that for any $y \in \mathcal{I}$, $d(y, \mu'(y)) \leq min\{\frac{\nu}{2}, g^{-1}\left(g(\frac{\nu}{2} - 2r) - \frac{\rho_2}{\rho_1} g(r)\right) + 2r\}$. Therefore, $\mathcal{I}$ is $d(y, \mu'(y)) \leq min\{\nu/2, g^{-1}\left(g(\frac{\nu}{2} - 2r) - \frac{\rho_2}{\rho_1} g(r)\right) + 2r\}$-distance-robust with respect to $RT_\delta(\mathcal{A})$ □

UNDERLYING STRUCTURE OF $\mathcal{I}$ & CONVEXITY OF g

In this section, we restrict our attention to convex cost functions and show that for such functions Theorems 2 and 3 can be strengthened. Note that g is convex in most common clustering, e.g. k-medians and k-means. Throughout this section, we implicitly assume that g is also continuous, increasing, and unbounded.

Theorem 4. *For any k and convex function g, let $\mathcal{A}$ be the (k, g)-centroid algorithm. For all $\mathcal{I} \subseteq \mathcal{X}$, such that $\mathcal{I}$ has a σ-separable, ρ-balanced clustering of diameter s, and for any $\delta > \sigma s/2$, and any*

$$\gamma \leq \frac{\left(\frac{|\mathcal{X} \setminus \mathcal{I}|}{|\mathcal{I}|} + 1\right) g(\delta) + 2g(s)}{g(\sigma s/2)} + 2k\rho^2$$

$\mathcal{I}$ is γ-robust with respect to $RT_\delta(\mathcal{A})$.

We need to develop a few results before proving Theorem 4. The next lemma states an important property of convex functions.

Lemma 4. *For any $x, y \in \mathcal{X}$, a metric distance function d, and a convex function g, $g(d(x, c)) + g(d(y, c)) \geq 2g\left(\frac{d(x,y)}{2}\right)$.*

Proof. In the following, the first inequality holds by the convexity of g and the second inequality holds by the fact that g is increasing and d satisfies the triangle inequality.

$$\begin{aligned} g(d(x, c)) + g(d(y, c)) &\geq 2g\left(\frac{d(x, c) + d(y, c)}{2}\right) \\ &\geq 2g\left(\frac{d(x, y)}{2}\right) \end{aligned}$$

□

The next lemma bounds the number of pairs that are clustered differently in two clusterings based on the distance between the partitions.

Lemma 5. *(Ackerman et al., 2013) Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be two clusterings of $\mathcal{Y}$, where $\mathcal{C}_1$ is ρ-balanced and has k clusters. If $\Delta(\mathcal{C}_1, \mathcal{C}_2) \geq \gamma$, then the number of disjoint pairs $\{x, y\} \subseteq \mathcal{Y}$ such that $x \not\sim_{\mathcal{C}_1} y$ and $x \sim_{\mathcal{C}_2} y$ is at least $\frac{1}{2}(\gamma - k\rho^2)|\mathcal{Y}|$.*

The next two lemmas examine the output of the (k, g)-δ-Truncated and (k, g)-centroid algorithms when g is convex and $\mathcal{I}$ has a desirable internal structure.

Lemma 6. *For any k and convex function g, let $\mathcal{A}$ be the (k, g)-centroid algorithm. For all $\mathcal{I} \subseteq \mathcal{X}$ that has a σ-separable, ρ-balanced k-clustering of diameter s, namely $\mathcal{B}$, and any $\delta > \frac{\sigma s}{2}$, if $\mathcal{A}'$ is $RT_\delta(\mathcal{A})$,*

$$\Delta(\mathcal{B}, \mathcal{A}'(\mathcal{X})|\mathcal{I}) \leq \frac{\frac{|\mathcal{X} \setminus \mathcal{I}|}{|\mathcal{I}|} g(\delta) + g(s)}{g(\sigma s/2)} + k\rho^2$$

Proof. Let $\mathcal{A}'(\mathcal{X}) = \mathcal{C}'$ with centers $\mu'_1, \ldots, \mu'_k$. Let $\Delta(\mathcal{B}, \mathcal{C}') = \gamma$ and assume on the contrary that $\gamma > \frac{\frac{|\mathcal{X} \setminus \mathcal{I}|}{|\mathcal{I}|} g(\delta) + g(s)}{g(\sigma s/2)} + k\rho^2$. Using Lemma 4, for any $\{x, y\} \in \mathcal{I}$ such that $x \not\sim_{\mathcal{B}} y$ but $x \sim_{\mathcal{C}'} y$, $g(d'(x, \mu'_i)) + g(d'(y, \mu'_i)) \geq \min\{2g(\delta), 2g(\frac{\sigma s}{2})\} \geq 2g(\frac{\sigma s}{2})$. Lemma 5 shows that there are at least $\frac{1}{2}(\gamma - k\rho^2)|\mathcal{I}|$ many such disjoint pairs. Therefore,

$$\begin{aligned} \Lambda^g_{d'}(\mathcal{C}') &\geq g(\frac{\sigma s}{2})(\gamma - k\rho^2)|\mathcal{I}| \\ &> g(\frac{\sigma s}{2}) \frac{\frac{|\mathcal{X} \setminus \mathcal{I}|}{|\mathcal{I}|} g(\delta) + g(s)}{g(\sigma s/2)} |\mathcal{I}| \\ &> g(s)|\mathcal{I}| + |\mathcal{X} \setminus \mathcal{I}| g(\delta) \\ &> \Lambda^g_{d'}(\mathcal{B} \cup \{\mathcal{X} \setminus \mathcal{I}\}) \end{aligned}$$

Contradiction. Therefore, $\Delta(\mathcal{B}, \mathcal{A}'(\mathcal{X})|\mathcal{I}) \leq k\rho^2 + \frac{\frac{|\mathcal{X}\setminus\mathcal{I}|}{|\mathcal{I}|}g(\delta)+g(s)}{g(\sigma s/2)}$. □

Lemma 7. *For any k and convex function g, let $\mathcal{A}$ be the (k, g)-centroid algorithm. Let $\mathcal{I}$ have a σ-separable, ρ-balanced clustering of diameter s, namely $\mathcal{B}$. Then, $\Delta(\mathcal{A}(\mathcal{I}), \mathcal{B}) \leq \frac{g(s)}{g(\sigma s/2)} + k\rho^2$*

Proof. Let $\mathcal{A}(\mathcal{I}) = \mathcal{C}$ with centers $\mu_1, \ldots, \mu_k$. Let $\Delta(\mathcal{B}, \mathcal{C}) = \gamma$ and assume, in search of a contradiction, that $\gamma > \frac{g(s)}{g(\sigma s/2)} + k\rho^2$. For any $\{x, y\} \in \mathcal{I}$ such that $x \not\sim_{\mathcal{B}} y$ but $x \sim_{\mathcal{C}} y$, using Lemma 5, $g(d(x, \mu_i)) + g(d(y, \mu_i)) \geq 2g(\frac{\sigma s}{2})$. Lemma 6, shows that there are at least $\frac{1}{2}(\gamma - k\rho^2)|\mathcal{I}|$ many such disjoint pairs. Therefore,

$$\begin{aligned}\Lambda_d^g(\mathcal{C}) &\geq \frac{1}{2}(\gamma - k\rho^2)|\mathcal{I}|2g(\sigma s/2)\\ &> g(s)|\mathcal{I}|\\ &> \Lambda_d^g(\mathcal{B})\end{aligned}$$

This forms a contradiction. Therefore, $\Delta(\mathcal{A}(\mathcal{I}), \mathcal{B}) \leq \frac{g(s)}{g(\sigma s/2)} + k\rho^2$. □

Proof of Theorem 4. Let $\mathcal{B}$ be the σ-separable, ρ-balanced, k-clustering of diameter s that covers $\mathcal{I}$, and let $\gamma' = \frac{\frac{|\mathcal{X}\setminus\mathcal{I}|}{|\mathcal{I}|}g(\delta)+g(s)}{g(\sigma s/2)} + k\rho^2$, and $\gamma'' = \frac{g(\delta)+g(s)}{g(\sigma s/2)} + k\rho^2$. Lemmas 6 and 7 respectively show that $\Delta(\mathcal{B}, \mathcal{A}'(\mathcal{X})|\mathcal{I}) \leq \gamma''$ and $\Delta(\mathcal{B}, \mathcal{A}(\mathcal{I})) \leq \gamma'$. Therefore, $\Delta(\mathcal{A}(\mathcal{I}), \mathcal{A}'(\mathcal{X})|\mathcal{I}) \leq \gamma' + \gamma'' \leq \gamma$. □

7. Non-robustness of the Simplistic Paradigm

A key component of our δ-Truncated paradigm is the use of a "garbage-collecting" cluster. In this section, we show that using the common cost functions with no such "garbage collectors", we can not achieve similar noise robustness performance. More specifically, Theorems 5 and 6 show that for any desired level of robustness and signal-to-noise ratio, there exists $\mathcal{I} \subseteq \mathcal{X}$ with the desired signal-to-noise ratio and a well-clusterable underlying pattern that is not robust with respect to the p-Increased paradigm, as long as $p < |\mathcal{X} \setminus \mathcal{I}|$.

Theorem 5. *Let $\mathcal{A}$ be the k-means algorithm. For any $\alpha, r, \lambda > 0$ there exists $\mathcal{X}$ and $\mathcal{I} \subseteq \mathcal{X}$, such that $\mathrm{rad}(\mathcal{I}) \leq r$, $\mathcal{I}$ can be covered with k balls of arbitrarily small radii, and $\mathcal{X}$ has signal-to-noise ratio $\frac{|\mathcal{I}|}{|\mathcal{X}\setminus\mathcal{I}|} \geq \lambda$. But, for any $p < |\mathcal{X}\setminus\mathcal{I}|$, $\mathcal{I}$ is not α-distance-robust or $\alpha^2(|\mathcal{I}|-|\mathcal{X}\setminus\mathcal{I}|)$-cost-robust with respect to $RI_p(\mathcal{A})$ for the k-means cost function.*

Proof. Let $d_1 = (\alpha + 2r)(\frac{\lambda}{\lambda+1}|\mathcal{X}| + 1)$ and $d_2 = 2(d_1 + 2r) + 1$. For $i \leq k$, let B_i denote a closed ball of radius 0, such that $|B_i| \geq \frac{\lambda}{k(\lambda+1)}|\mathcal{X}|$. Let $B_1, \ldots, B_k$ be evenly placed on a line of length $2r$. For, $i \leq \lfloor \frac{|\mathcal{X}|}{\lambda+1} \rfloor$, let o_i be a point on the line that connects $B_1, \ldots, B_k$, such that $d(o_1, B_1) = d_1$ and $d(o_i, o_{i+1}) = d_2$ (see Figure 1). Let $\mathcal{I} = \bigcup_{i\in[k]} B_i$ and $\mathcal{X} = \mathcal{I} \cup \{o_1, \ldots, o_{\lfloor |\mathcal{X}|/(\lambda+1)\rfloor}\}$. Note that $\mathcal{X}$ and $\mathcal{I}$ are chosen such that $\frac{|\mathcal{I}|}{|\mathcal{X}\setminus\mathcal{I}|} \geq \lambda$.

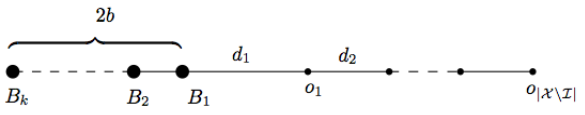


Figure 1. Structure of an unrobust dataset w.r.t $RI_p(\mathcal{A})$.

Let $\mathcal{A}' = RI_p(\mathcal{A})$, $\mathcal{A}'(\mathcal{X}) = \{C_1, \ldots, C_{k+p}\}$, μ_i denote the center of C_i, and $\mu(x)$ denote the closest center to x. Assume (in the hope of finding a contradiction) that for all $i \leq k+p$, $C_i \subseteq \mathcal{I}$ or $C_i \subseteq \mathcal{X}\setminus\mathcal{I}$. Without loss of generality let $C_1 \subseteq \mathcal{I}$, then $d(o_1, \mu_1) \leq d_1 + 2r$. Moreover, for any $C_i \subseteq \mathcal{X} \setminus \mathcal{I}$, such that $|C_i| \geq 2$, if o_j is the left-most or the right-most point of C_i, $d(o_j, \mu_i) \geq d_2/2 > d(o_1, \mu_1)$. Without loss of generality, assume $o_1 \in C_2$. There are two cases:

1. $C_2 = \{o_1, o_j, \ldots\}$: Then the cost of clustering $\mathcal{C}' = \{C_1 \cup \{o_1\}, C_2 \setminus \{o_1\}, \ldots, C_{p+k}\}$ is lower than the cost of $\mathcal{C}$.
2. $C_2 = \{o_1\}$: Let $C_3 \subseteq \mathcal{X} \setminus \mathcal{I}$ be any cluster of size at least 2, and let o_i be its left-most point (such a cluster exists since $p < |\mathcal{X} \setminus \mathcal{I}|$). The cost of clustering $\mathcal{C}' = \{C_1 \cup \{o_1\}, \{o_i\}, C_3 \setminus \{o_i\}, \ldots, C_{p+k}\}$ is lower than the cost of $\mathcal{C}$.

Hence, $\mathcal{C}$ is not an optimal clustering. So, for any optimal $RI_p(\mathcal{A})$ clustering, there exists a cluster C_i such that $\{o_j, y\} \subseteq C_i$ for some $y \in \mathcal{I}$. Then, $d(y, \mu_i) \geq \frac{d_1}{|\mathcal{I}|+1} > \alpha + 2r$. Hence, $\mathcal{I}$ is not α-distance-robust to $\mathcal{X} \setminus \mathcal{I}$ with respect to $RI_p(\mathcal{A})$.

Because there exists $y \in \mathcal{I}$ with $d(y, \mu(y)) > \alpha + 2r$ for all $y' \in \mathcal{I}$, then $d(y', \mu(y')) > \alpha$. So, $\mathcal{I}$ is not $\alpha^2(|\mathcal{I}| - |\mathcal{X}\setminus\mathcal{I}|)$-cost-robust with respect to $RI_p(\mathcal{A})$ for the k-means cost function. □

Note that Theorem 5 implies that for any (k, g)-centroid algorithm, $\mathcal{A}$, any desired level of robustness, and any signal-to-noise ratio, there exists $\mathcal{I}$ with arbitrarily small radius and $\mathcal{X} \supseteq \mathcal{I}$ with that signal-to-noise ratio, that is not robust with respect to the $RI_p(\mathcal{A})$ function, as long as $p \leq |\mathcal{X} \setminus \mathcal{I}| - k$.

The next theorem proves similar results as in Theorem 5 for (structural) robustness, and contrasts results shown in Theorems 3 and 4.

Theorem 6. *Let $\mathcal{A}$ be the k-means algorithm. For any $r, \lambda > 0$, there exists $\mathcal{X}$ and $\mathcal{I} \subseteq \mathcal{X}$, such that $\mathrm{rad}(\mathcal{I}) \leq r$, $\mathcal{I}$ can be covered with k balls of arbitrarily small radii, and $\mathcal{X}$ has signal-to-noise ratio of $\frac{|\mathcal{I}|}{|\mathcal{X} \setminus \mathcal{I}|} \geq \lambda$. But, for any $p < |\mathcal{X} \setminus \mathcal{I}|$, $\mathcal{I}$ is not $(1 - \frac{1}{k})$-robust with respect to $RI_p(\mathcal{A})$.*

Bounded Space

Corollaries 1 and 2 demonstrate the limitations of p-Increased algorithms even when the diameter of the data is bounded.

Corollary 1. *Let $\mathcal{A}$ be the k-means algorithm. For any λ, k and $\nu < \frac{1}{k}$, there exists $\mathcal{I} \subseteq \mathcal{X}$, such that $\mathcal{X}$ has diameter 1 and signal-to-noise ratio $\frac{|\mathcal{I}|}{|\mathcal{X} \setminus \mathcal{I}|} \geq \lambda$, and $\mathcal{I}$ can be covered by k balls that have arbitrarily small radii and their centers are at least ν apart. But for any $p < |\mathcal{X} \setminus \mathcal{I}|$ and any $\alpha \leq \frac{1 - k\nu}{2|\mathcal{X}|(|\mathcal{I}|+1)}$, $\mathcal{I}$ is not α-distance-robust, $(1 - \frac{1}{k})$-robust, or $(\alpha - \nu k)^2(|\mathcal{I}| - |\mathcal{X} \setminus \mathcal{I}|)$-cost-robust with respect to $RI_p(\mathcal{A})$ for the k-means cost function.*

Corollary 2. *Let $\mathcal{A}$ be the k-means algorithm. For any $\lambda > 0$, there exists $\mathcal{I} \subseteq \mathcal{X}$, such that $\mathcal{X}$ has diameter 1 and signal-to-noise ratio $\frac{|\mathcal{I}|}{|\mathcal{X} \setminus \mathcal{I}|} \geq \lambda$, and $\mathcal{I}$ has an arbitrarily small radius, but for any $p < |\mathcal{X} \setminus \mathcal{I}| - k$ and any $\alpha \leq \frac{1}{2(|\mathcal{I}|+1)|\mathcal{X}|}$, $\mathcal{I}$ is not α-distance-robust or $\alpha^2(|\mathcal{I}| - |\mathcal{X} \setminus \mathcal{I}|)$-cost-robust with respect to $RI_p(\mathcal{A})$ for the k-means cost function.*

8. Comparing the Two Paradigms

Here, we use an example to further demonstrate the robustness of the δ-Truncated paradigm compared to the limitations of the p-Increased paradigm. Moreover, we use experiments to back up these our theoretical results.

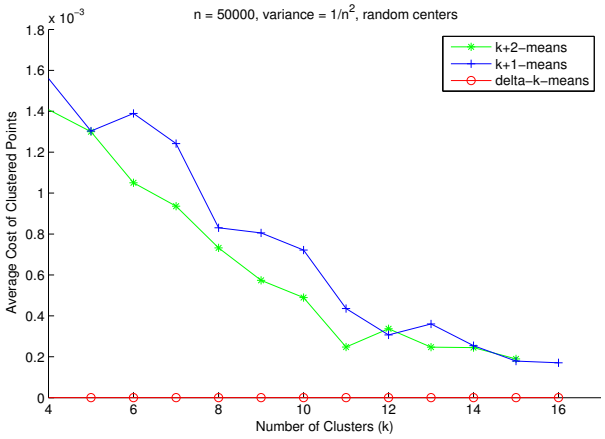
Example 1. *Let $\mathcal{A}$ denote the k-means algorithm. By Corollary 2, there exists $\mathcal{I} \subseteq \mathcal{X}$ such that $|\mathcal{X}| = n$ has diameter 1 and 10% noise, and $\mathcal{I}$ has radius 0, but for any $p < 0.1n - k$, $\mathcal{I}$ is not $5/n^2$-distance robust to $\mathcal{X} \setminus \mathcal{I}$ with respect to $RI_p(\mathcal{A})$. Since $\mathcal{I}$ can be covered by a ball of radius $1/4n^2$, by Theorem 2, for $\delta \in [\frac{1}{n^2}, \frac{3\sqrt{3}}{4n^2})$, $\mathcal{I}$ is $1/n^2$-distance-robust with respect to $RT_\delta(\mathcal{A})$.*

8.1. Experiments

Datasets: For any k, we use $n = 50000$ data points on the unit square. 90% of the points come from k Gaussian distributions with centers selected uniformly at random and standard deviation $= \frac{1}{n}$. Additionally, 10% uniform noise is introduced in the data.

Clustering Algorithms: We run the Lloyd algorithm for $k + p$-means, when $p = 1, 2$. For δ-k-means, we adapt the Lloyd algorithm to calculate the clustering using a δ-truncated distance matrix in every iteration, where δ is set to $\frac{10}{n}$.

Charts: The following graph shows the average cost of a clustering per clustered point, in δ-k-means, $k+1$-means, and $k+2$-means. Let $(\mathcal{C}, \Phi)$ be the δ-k-means clustering. The δ-k-means average cost is $|Cost(\mathcal{C}) \ / \ |\bigcup \mathcal{C}|$. For the $k + p$-means, we find the minimal cost of a collection of clusters that cover at least $|\bigcup \mathcal{C}|$ points, then divide this cost by the number of points that are present in these clusters. The δ-k-means average cost per point is considerably



smaller than that of $k + p$-means. Moreover, this cost is stable throughout the experiments and remains close to the average radius of the cluster. On the other hand, the cost of $k + p$-means fluctuates and is considerably higher than the average radius of the clusters.

9. Concluding Remarks

In this paper, we consider the problem of robustness of clustering algorithm to the addition of unstructured data points (that we termed "noise"). We propose to augment any given center-based clustering algorithm with an efficiently implementable "noise-robustifying" mechanism that creates an additional cluster, used as a "garbage collecting" bucket. We introduce rigorous robustness notions that capture different aspects of robustness that may be desirable for such algorithms. We prove that our algorithmic paradigm indeed guarantees desirable noise robustness, and show that the simple strategy, of just applying the underlying clustering algorithms with extra clusters (to accommodate such noisy data), cannot enjoy similar performance.

Acknowledgments

This research was partially supported by a Google research award.

References

Ackerman, Margareta and Ben-David, Shai. Clusterability: A theoretical study. In Dyk, David A. Van and Welling, Max (eds.), *AISTATS*, volume 5 of *JMLR Proceedings*, pp. 1–8. JMLR.org, 2009.

Ackerman, Margareta, Ben-David, Shai, Loker, David, and Sabato, Sivan. Clustering oligarchies. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, 2013.

Ailon, Nir, Jaiswal, Ragesh, and Monteleoni, Claire. Streaming k-means approximation. In *Advances in Neural Information Processing Systems*, pp. 10–18, 2009.

Balcan, Maria-Florina, Blum, Avrim, and Gupta, Anupam. Approximate clustering without the approximation. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1068–1077. Society for Industrial and Applied Mathematics, 2009.

Ben-David, Shai, Von Luxburg, Ulrike, and Pál, Dávid. A sober look at clustering stability. In *Learning theory*, pp. 5–19. Springer, 2006.

Bilu, Yonatan and Linial, Nathan. Are stable instances easy? *Combinatorics, Probability & Computing*, 21(5): 643–660, 2012.

Cuesta-Albertos, J. A. , Gordaliza, Alfonso, and Matrán, C.˙ Trimmed k-means: An attempt to robustify quantizers. *The Annals of Statistics*, 25(2):553–576, 1997.

Dave, Rajesh N. Characterization and detection of noise in clustering. *Pattern Recognition Letters*, 12(11):657–664, 1991.

Dave, Rajesh N. Robust fuzzy clustering algorithms. In *Proceedings of the 2nd IEEE International Conference on Fuzzy Systems*, pp. 1281–1286, 1993.

Donoho, David L. Breakdown properties of multivariate location estimators. Technical report, Harvard University, 1982.

Ester, Martin, Kriegel, Hans-Peter, Sander, Jörg, and Xu, Xiaowei. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of Knowledge Discovery in Databases*, volume 96, pp. 226–231, 1996.

García-Escudero, Luis Ángel and Gordaliza, Alfonso. Robustness properties of k means and trimmed k means. *Journal of the American Statistical Association*, 94 (447):956–969, 1999.

García-Escudero, Luis Ángel, Gordaliza, Alfonso, Matrán, Carlos, and Mayo-Iscar, Agustin. A general trimming approach to robust cluster analysis. *The Annals of Statistics*, pp. 1324–1345, 2008.

Hampel, Frank R. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, 42(6): 1887–1896, 1971.

Hennig, Christian. Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. *Journal of Multivariate Analysis*, 99(6):1154–1176, 2008.

Rand, William M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.