An Effective and Efficient Approach for Clusterability Evaluation

Andreas Adolfsson
AN EFFECTIVE AND EFFICIENT APPROACH FOR CLUSTERABILITY EVALUATION

By

ANDREAS ADOLFSSON

A Thesis submitted to the Department of Computer Science in partial fulfillment of the requirements for the degree of Master of Science

2016
Andreas D. Adolfsson defended this thesis on May 25, 2016.
The members of the supervisory committee were:

Margareta Ackerman  
Professor Co-Directing Thesis

Naomi Brownstein  
Professor Co-Directing Thesis

Sonia Haiduc  
Committee Member

Gary Tyson  
Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the thesis has been approved in accordance with university requirements.
# TABLE OF CONTENTS

List of Figures .............................................................................................................. iv
List of Tables ................................................................................................................... vi
Abstract ........................................................................................................................... vii

<table>
<thead>
<tr>
<th>CHAPTER 1: INTRODUCTION</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1: Previous work</td>
<td>3</td>
</tr>
<tr>
<td>1.2: Multimodality tests</td>
<td>4</td>
</tr>
<tr>
<td>1.3: Comparison with previous notions</td>
<td>5</td>
</tr>
</tbody>
</table>

| CHAPTER 2: METHOD FOR CLUSTERABILITY EVALUATION: OUR APPROACH                           | 6 |

<table>
<thead>
<tr>
<th>CHAPTER 3: SIMULATIONS</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1: Simulation details</td>
<td>10</td>
</tr>
<tr>
<td>3.2: Visualizing the data</td>
<td>13</td>
</tr>
<tr>
<td>3.3: Summary of results</td>
<td>23</td>
</tr>
</tbody>
</table>

| CHAPTER 4: RESULTS ON REAL DATASETS                                                     | 24 |

| CHAPTER 5: CONCLUSIONS AND DISCUSSION                                                    | 34 |

References ..................................................................................................................... 35

Biographical Sketch ........................................................................................................ 38
LIST OF FIGURES

Figure 1: Clustering pipeline. This figure shows the feedback pathways in cluster analysis and role of clusterability in this process. ................................................................. 1

Figure 2: Illustrative examples demonstrating the foundation of our approach for clusterability evaluation ........................................................................................................ 7

Figure 3: A sample projection of the Single Cluster 2D dataset, which corresponds to row A in Table 1. ........................................................................................................ 14

Figure 4: This figure shows 2 of the dimensions existing in the Single Cluster 3D dataset, corresponding to row M in Table 1. ........................................................................ 15

Figure 5: This figure shows 2 of the dimensions existing in the Single Cluster 10D dataset, corresponding to row N in Table 1. ................................................................. 15

Figure 6: This figure shows the Single Cluster with Outlier 2D dataset, which corresponds to row O in Table 1. .......................................................................................... 16

Figure 7: This figure portrays the Single Large Cluster with Outlier 2D dataset, which corresponds to row P in Table 1. .................................................................................. 16

Figure 8: This figure shows the Single Cluster with Three Scattered Outliers 2D dataset, corresponding to row Q in Table 1. ................................................................. 17

Figure 9: This figure shows the Two Separated Clusters 2D dataset, which corresponds to row B in Table 1. ..................................................................................................... 17

Figure 10: This figure depicts two of the dimensions from the Two Separated Clusters 10D dataset, which corresponds to row I in Table 1. ......................................................... 18

Figure 11: This figure shows the Three Separated Clusters 2D dataset, which corresponds to row G in Table 1. .......................................................................................... 18

Figure 12: This figure shows two dimensions of the Three Separated Clusters 3D dataset, corresponding to row H in Table 1. ................................................................. 19

Figure 13: This figure shows the Three Close Clusters 2D dataset, corresponding to row C in Table 1. .......................................................................................... 19

Figure 14: This figure shows the Three Noisy Clusters 2D dataset, which corresponds to row D in Table 1. .......................................................................................... 20

Figure 15: This figure shows the Three Clusters with Different Diameters 2D dataset, which corresponds to row E in Table 1. ................................................................. 20

Figure 16: This figure shows the Three Clusters with Different Densities 2D dataset, which corresponds to row F in Table 1. ................................................................. 21
Figure 17: This figure depicts two of the dimensions from the Four Separated Clusters 10D dataset, which corresponds to row J in Table 1. 

Figure 18: This figure shows the Ten Separated Clusters 2D dataset, which corresponds to row K in Table 1. 

Figure 19: This figure shows the Ten Close Clusters 2D dataset, which corresponds to row L in Table 1. 

Figure 20: This figure shows the Faithful dataset with natural cluster structure. Both Dip and Silverman tests captures this with P values of 0 indicating that faithful is clusterable. 

Figure 21: This figure shows 2D projections of the famous Iris dataset, which possesses clear cluster structure. Our method for clusterability evaluation attains a p-value of 0 using both the dip and Silverman tests, indicating that iris is clusterable. 

Figure 22: This figure shows the Rivers dataset. Using the Dip test we capture the weak cluster structure with a p value of 0.2772, while Silverman test can be used to capture the small outlying clusters with a p value of 0. 

Figure 23: This figure depicts the Swiss dataset’s 2D projections suggesting the data is clusterable. Both of our methods capture this clusterability with P values 0. 

Figure 24: Two-dimensional projections of the USArrests data, suggesting that the data is inherently unclusterable, as confirmed by the p-values of 0.9394 and 0.1897 for the dip and Silverman tests, respectively. 

Figure 25: Two-dimensional projections of the Attitude dataset suggesting inherently unclusterable structure. Our method captures this inherent structure, yielding p-values of 0.9040 and 0.9449 for the Dip and Silverman tests, respectively. 

Figure 26: Two-dimensional projections of the USJudgeRatings data, depicting an unclusterable structure. Our methods come to the same conclusion with p=0.9938 for the Dip test and p=0.7451 for Silverman. 

Figure 27: This figure displays the Cars dataset, showing no distinct clusters. Our method agrees with p=0.6604 for the dip test and p=0.9931 for the Silverman test, both indicating the data to be unclusterable. 

Figure 28: This figure depicts the two-dimensional projections of the Trees data. No strong cluster structure is apparent, and our results, p = 0.346 for dip and p=0.3235 for Silverman, agree that no structure exists.
LIST OF TABLES

Table 1: Proportion of datasets classified as clusterable over 1000 runs for each type, for a total of 17000 simulations. For Dip and Silverman, high scores indicate highly clusterable data, and low scores indicate lower cluster structure. 24

Table 2: Real Data p-values. This table presents the p values for the dip and Silverman multimodality tests on real datasets from the R Datasets package. Recall that p < 0.05 signals clusterable data, which larger values of p signal that data is unclusterable. 25
ABSTRACT

Clustering is an essential data mining tool that aims to discover inherent cluster structure in data. As such, the study of clusterability, which evaluates whether data possesses such structure, is an integral part of cluster analysis. Yet, despite their central role in the theory and application of clustering, current notions of clusterability fall short in two crucial aspects that render them impractical; most are computationally infeasible and others fail to classify the structure of real datasets.

In this thesis, we propose a novel approach to clusterability evaluation that is both computationally efficient and successfully captures the structure in real data. Our method applies multimodality tests to the (one-dimensional) set of pairwise distances based on the original, potentially high-dimensional data. We present extensive analyses of our approach for both the Dip and Silverman multimodality tests on real data as well as 17,000 simulations, demonstrating the success of our approach as the first practical notion of clusterability.
CHAPTER 1
INTRODUCTION

Clustering is a ubiquitous data analysis tool, applied in virtually all disciplines. The goal of clustering is to uncover meaningful cluster structure in data. As such, the application of this data mining tool relies on the presence of inherent structure, making notions of clusterability, which aim to quantify the degree of cluster structure in data, integral to cluster analysis. Evaluation of clusterability plays a central role in both clustering theory and practice. Clusterability analysis should precede the application of clustering algorithms, because the success of any clustering algorithm depends on the presence of cluster structure.

To see how clusterability fits within the clustering process, consider the clustering pipeline depicted in Figure 1. The process begins with data preprocessing, often involving feature selection or extraction. Next, clusterability analysis determines whether the data possesses sufficient cluster structure. If it is discovered that data does not possess sufficient cluster structure to be meaningfully partitioned, then clustering may not be suitable for the given data, or the data may be reprocessed.

On the other hand, if the data is found to be clusterable, a suitable algorithm may be selected or developed. After the algorithm is executed, the solution is validated by applying clustering quality measures [1, 30], which may result in the selection of an alternate algorithm.

In the theoretical analysis of clustering, data clusterability is an essential consideration, as many desirable characteristics of clustering algorithms arise only when the data is sufficiently clusterable. Since the results of clustering methods are often inconsequential when data fails to

---

1 A similar pipeline is presented in the famous survey of clustering algorithms by Xu and Wunsch [35], sans the third step. Figure 1 shows how clusterability fits within the clustering process.
2 Note that multiple methods should be considered at this step, because different algorithms are apt at identifying different types of cluster structures [5, 7].
possess inherent cluster structure, it suffices to study how algorithms behave on clusterable sets. Clusterability notions have been used to show that, when data possesses inherent cluster structure, various algorithms are computationally efficient [12, 32], produce the desired output [11, 7], and possess other desirable properties, such as robustness to small adversarial sets [6]. While prior notions of clusterability help gain insight into the behavior of clustering techniques, they fall short from a practical standpoint. Perhaps the most significant hindrance to the application of prior notions is that most are NP-hard to compute [2]. The few exceptions that can be computed in polynomial time fail to capture the structure of real datasets, as they are designed to identify exceptionally well-structured data that rarely occur in practice, as detailed in Section 1.1.

Computational infeasibility and failure to identify the structure of real datasets inhibit the practical application of prior notions of clusterability. Further, inability to apply these notions in practice ultimately limits the implications of theoretical results relying on these notions, making it difficult to gauge whether or when these notions capture how real data is structured. Thus, both clustering theory and practice stand to benefit from a practical notion of clusterability. Lastly, many previous notions of clusterability are based on specific algorithms or objective functions [32, 11, 12, 8, 2, 9]. This effectively inverts the clustering pipeline, requiring that we choose an algorithm before we determine whether data possesses sufficient structure to be meaningfully clustered. Further, since different clustering algorithms identify distinct types of cluster structure [7, 6], relying on a specific algorithm restricts a notion of clusterability to identifying structure that the underlying algorithm can capture. In this thesis, we propose the first practical notion of clusterability. Our method of clusterability evaluation describes the structure of real sets, runs is low polynomial time, and is independent of any algorithm or objective function. In an effort to develop this notion, we make an important paradigm shift. Until now, notions of clusterability were typically developed in the theoretical realm [32, 11, 12], with the aim of identifying when algorithms exhibit interesting behavior. Ours is the first notion developed through extensive data analysis.

By developing a clusterability measure that can be computed quickly and captures the structure of real data, we allow users to start utilizing clusterability for what is arguably its primary purpose: to discover whether data possesses sufficient cluster structure to meaningfully apply clustering techniques. Further, theoretical analysis relying on our notion of clusterability benefits from its ability to capture the structure of real sets.

The key to our approach lies in the discovery that a dataset's histogram of dissimilarities reveals a lot of information about its inherent cluster structure. Namely, the presence of multiple modes
in the (one-dimensional) set of pairwise distances indicates that the original (possibly high-dimensional) data possesses inherent cluster structure. We show how tests of multimodality, namely the Dip and Silverman tests, can be applied on the set of pairwise dissimilarities to determine whether the underlying data is clusterable. Notably, our approach is computationally efficient, being quadratic in the input size, which is a radical improvement over standard NP-hard approaches.

The remainder of the thesis is organized as follows. We begin with an overview of previous work on clusterability and an introduction to tests of multimodality. We then present a detailed description of our approach and several illustrative examples. Next, we describe our extensive simulations and findings on real data. We conclude with a discussion and future work.

1.1 Previous work

Due to their central role in cluster analysis, many different notions of clusterability have been proposed. Most of these notions rely on a specific, often NP-hard, objective functions. For instance, one elegant notion defines data as clusterable when clusterings of near-optimal cost possess near-optimal structure [11]. Another similar notion considers the ratio of the optimal k-means solution over the optimal solution with k-1 clusters to attain a clusterability score [32]. Other notions that rely on optimal solutions for NP-hard objective functions were considered in [8, 2] and [9].

Another class of clusterability notions consists of those that can be efficiently computed, but are too strict for most practical applications. For example, Epter [20] defines data as clusterable when the minimum between-cluster separation exceeds the maximum in-cluster distance. Another elegant notion, by Balcan et al [12], defines data to be clusterable when each element is closer to all elements in its cluster than to all other data.

Both the computationally infeasible notions of clusterability and those that are too strict for practical applications helped enrich our understanding of clustering and played a key role in many recent findings [12, 3, 32, 6, 11, 4]. However, they leave open the challenge of developing a realistic notion for evaluating the degree of inherent cluster structure in data, which is the problem we tackle in the current work.

Detecting the number of clusters is a related problem, where the typical approach involves the analysis of clustered data, requiring that we select and execute an algorithm before determining whether multiple clusters are present [16]. This inverts the clustering pipeline, typically requiring numerous runs of both the algorithm and a validity index, but most importantly, this approach may fail to detect multiple clusters if an inappropriate algorithm is chosen, as different clustering algorithms are apt at detecting different cluster structures [5, 6, 7]. As such, selecting
an inappropriate algorithm may occlude the presence of multiple clusters in the data, leading to invalid conclusions.

Another related measure that can be viewed as a method for evaluating clusterability is stability [15], which evaluates the consistency of an algorithm's behavior on different samples of the same data. Like typical approaches for detecting the number of clusters, whether data is found to be clusterable using stability depends on the chosen algorithm. By contrast, we seek to test whether data is clusterable independently of any choice of clustering algorithm. This allows one to determine whether clustering is warranted.

Despite its prominence as a central concept in clustering, very little work has been done studying clusterability as an independent construct. The first theoretical study of clusterability by Ackerman and Ben-David [2] compared previous notions, showing that they are pairwise inconsistent; for each pair of notions considered, there was data for which one of the notions evaluated as well-clusterable, while another indicated that the data is poorly clusterable. This highlights the importance of treating notions of clusterability with care, as notions which seem natural at first glance are often radically different from one another. Ackerman and Ben-David [2] found that whenever data is clusterable using any of the studied notions, there is an algorithm that can be used to efficiently cluster the data. Daniely et al [19] recently studied this phenomenon, proposing the hypothesis, also known as the CDNM thesis, that “clustering is difficult only when it does not matter.” Ben-David [14] more carefully considered this hypothesis, concluding that the CDNM thesis is far from being substantiated.

In an effort to explore the CDNM hypothesis, Ben-David [14] brought up the open problem of developing novel notions of clusterability, proposing four properties that these notions should satisfy. Among these four requirements are the following two which pertain to the clusterability evaluation component of the pipeline in Figure 1: (1) It should be reasonable to assume that most (or at least a significant proportion of) the inputs one may care to cluster in practice satisfy the clusterability notion and (2) There exists an efficient algorithm for testing clusterability; namely, given an instance, the algorithm determines whether it satisfies the clusterability requirement or not. No previous methods satisfy both of these conditions, and we have no evidence that any satisfy (1). In this thesis, we introduce the first approach to clusterability evaluation that satisfies these two requirements.

1.2 Multimodality tests

Determining whether a dataset has two or more modes is an integral component of our approach. A number of tests for multimodality have been developed. One popular test by Silverman [34] is based on the kernel density estimate. The general idea involves approximating
the empirical distribution of the observed data with a set of Gaussian distributions of a specified bandwidth. If a sufficiently large bandwidth is required to produce a unimodal estimate, then the test concludes that the data distribution is multimodal. Another, the Dip test, compares the empirical distribution to a uniform density and rejects the assumption of unimodality if the data is sufficiently different from the closest possible uniform distribution [25]. Tests of multimodality have been applied to studies of carbon dioxide emissions, income levels, body size, cognitive processes, and archeological artifacts [26, 35, 23, 13, 18].

1.3 Comparison with previous notions
When proposing a new technique, it is common practice to compare it against previous methods. However, for clusterability analysis there is a major difficulty - none of the previous notions are practical. The biggest challenge is that, as discussed in Section 1.1 of this thesis, most prior notions of clusterability are NP-hard (which explains their exclusive use in the theoretical literature).

For the few that can be computed in polynomial time, it is easy to see that they often fail to detect that data is clusterable. Consider, for example, Epter's [20] notion that defines data as clusterable when the minimum between-cluster separation exceeds the maximum in-cluster distance, and the notion by Balcan et al. [12] that defines data to be clusterable when each element is closer to all elements in its cluster than to all other data. Both of these notions are highly sensitive to noise. As such, they fail to discover cluster structure when noise is present, such as the data depicted in Figure 2d. Note also that variations of these notions that explicitly provide the number of outliers would fail when these values cannot be accurately provided by the user, as is typically the case. Furthermore, these notions are unable to identify cluster structure when clusters are close together (e.g. Figure 2c), which our method does successfully.
CHAPTER 2
METHOD FOR CLUSTERABILITY EVALUATION: OUR APPROACH

Developing measures of clusterability has thus far been primarily considered in the theoretical context, often with the goal of demonstrating that specific algorithms possess desirable characteristics. We take a fresh perspective at the development of clusterability measures, proposing a new approach for studying and evaluating clusterability. This section focuses on an exposition of our approach, while the following two sections provide extensive simulations and results from real datasets.

The main insight in our approach is the observation that clusterability can be inferred from a one-dimensional view of the pairwise distances. That is, the lengths of the pairwise distances are sufficient for this analysis, without the need for considering how these distances are arranged to form the data.

To illustrate this approach, we begin with two extreme cases. First, consider highly clusterable data, consisting of two very well-separated, dense regions as depicted in Figure 2b. This dataset consists of two types of pairwise distances: very long ones, corresponding to between-cluster distances, and very short ones, corresponding to within-cluster distances. The histogram of pairwise distances depicted in the right column of Figure 2b reveals a sharp drop-off, clearly delineating the breaking point between in-cluster and between-cluster distances and resulting in two clear modes.

In general, when the underlying data is clusterable, it leads to a multimodal distribution of the pairwise distances. However, it is important to note that the modes do not represent distinct clusters. Instead, they tend to group together in-cluster and between-cluster distances.

Now, consider an example of unclusterable data, such as the random data generated using a single Gaussian distribution, depicted in Figure 2a. For this data, the histogram of pairwise distances is unimodal.

While the below examples consider clear cases of clusterable and unclusterable data, our findings include a large number of experiments that confirm this phenomenon: The degree of clusterability is captured in the histogram of its pairwise distances, allowing us to determine the degree of clusterability by counting the number of modes found in this histogram. In particular, multiple modes indicate that data is clusterable, while a single mode signals that data is not clusterable.
Figure 2: Illustrative examples demonstrating the foundation of our approach for clusterability evaluation.
Figures 2c-f illustrate additional key examples, showing that our findings persist in the presence of noise (Figure 2d), varied cluster diameters (Figure 2e), different cluster sizes (Figure 2f), and even when the separation between clusters is small (Figure 2c). Analysis of clusterable data consistently uncovers a fall-off after the first mode, followed by one or more additional modes. In stark contrast, unclusterable sets consistently yield unimodal dissimilarity distributions. The next two sections analyze many additional datasets.

Figure 2 includes the Dip and Silverman p-values for each of the depicted datasets. Figure (a) depicts unclusterable, random data generated using a single Gaussian distribution, along with the corresponding histogram of the underlying pairwise dissimilarities, which consists of a single mode. In contrast, Figure (b) shows data that possesses clear clustering structure, comprised of two dense, well-separated, clusters, with a corresponding bi-modal histogram of pairwise dissimilarities. Figures (c) through (f) demonstrate the robustness of our approach in the presence of noise, when cluster diameters and sizes vary, and even when cluster separation is small.

While the method of constructing a dendrogram of pairwise distances remains of independent interest, we propose applying multimodality tests that provide concrete recommendations and eliminate the need for creating dendrograms. To this end, we utilize statistical tests of multimodality on the set of pairwise distances, which are inherently one-dimensional. In particular, we use the Dip and Silverman tests. The following sections demonstrate the application of both simulated and real data.

The Dip and Silverman tests each provide a p-value, which indicates the probability of seeing the given input or a more extremely multimodal input if the data is unimodal. If only a single mode is present, then the p-value should be large. This indicates that the underlying data is not clusterable. On the other hand, small p-values make us question the original assumption of unimodality and instead conclude that multiple modes are present in the population. Consequently, this lets us conclude that the underlying data is clusterable. Note that both tests run in low polynomial time, where Silverman's test is quadratic, while the Dip test is linear in the number of pairwise distances [28].

The user should a priori choose a cut-off score, or significance level, to determine the point below which datasets would be determined to be clusterable. For example, the significance level of 0.05 is common. In this case, over the long term, we would expect 5% of unclusterable datasets to be (incorrectly) classified as clusterable. One could choose a smaller significance level at a cost of lower statistical power to correctly classify truly clusterable datasets as
clusterable. By convention, we use the significance level of 0.0, where values below 0.05 allow us to conclude that the distribution of pairwise distances has multiple modes. The p-values attained for the dip and Silverman multimodality tests are displayed next to each dataset, signaling that data is clusterable whenever p<0.05 and indicating that data is unclusterable otherwise. Note that both multimodality tests correctly identify the data in Figure 2a as unclusterable, with p-values greater than 0.05, and the other datasets in this figure as clusterable, by having p<0.05. The following two sections present extensive experimental results on both simulations and real datasets, confirming the validity of our approach.
We now describe our extensive simulations for evaluating our approach to clusterability using both the dip and Silverman tests. The simulations consist of 17 types of datasets, where each type was generated with the same parameters 1000 times, for a total of 17,000 simulations. For example, row (c) of Table 1, found in chapter 4, describes the results for all simulations with 3 clusters with small separation in 2D, for which we have selected parameters that generate 3 bivariate Gaussian clusters each consisting of 50 points, with means at (30,20), (40,20), and (35,30). One simulation of this type is displayed in Figure 2c. Note that many of our tests are in 2D to facilitate visual classification of the data as clusterable or unclusterable. For high dimensional data, we examine all 2D projections.

We generate 1000 different datasets using these parameters. For each dataset, we run the statistical Dip and Silverman tests. In Table 1, the percent of datasets on which each of the remaining tests has a p-value less than 0.05 (that is, the percent of time that the tests rejected the null hypothesis of unimodality at the traditionally used 0.05 significance level\(^3\)). As such, high values in Table 1 indicate high values of clusterability, while low values indicate poor clusterability.

### 3.1 Simulation details

We determine the integrity of our approach for determining the clusterability of a dataset through simulation of distinct datasets containing varying levels of inherent cluster structure. Using our understanding of the predisposed cluster structure of each simulated set, we demonstrate the strength of our analysis by comparing the empirical results of each respective set to its expected results. The simulated data is generated via R software version 3.2.2. Gaussian distributions are used for each cluster. Each scenario is run 1000 times. The rate of p-values below the selected significance level (0.05) is recorded in Table 1 from chapter 4. For unclusterable datasets, this rate represents Type I error, while for clusterable datasets, the rate represents statistical power.

We now discuss how each type of data summarized in chapter 4, Table 1 was generated:

---

\(^3\) For unclusterable datasets, the proportion of rejections corresponds to type I error, or the rate of erroneously classifying unclusterable datasets as clusterable. For clusterable datasets, the proportion of rejections corresponds to the statistical power, or ability of the test to correctly classify clusterable sets as having cluster structure.
• **Rows (a) of Table 1:** We generate datasets containing a single bivariate Gaussian cluster with mean and standard deviation of 100 and 2, respectively, for each independent dimension. The simulation describes a circular dataset with expected unclusterable data structure.

• **Row (m) of Table 1:** We next simulate a single cluster in 3 dimensions in which all three dimensions have mean 100 and standard deviation 2.

• **Row (n) of Table 1:** To test the performance of our method on high dimensional data, we simulate data from a single 10 dimensional cluster of 50 points. Each independent dimension is generated as a Gaussian random variable with mean sampled from a uniform (30, 40) distribution and a fixed standard deviation of 2. As shown in Table 1, both dip and Silverman tests reject the null hypothesis of multimodality less than 5% of the time, indicating that the tests have adequately low rates of erroneously classifying unclusterable datasets even in high dimensions.

• **Row (o) of Table 1:** Next, we tested the robustness of the tests to outliers. As before, we simulated a single bivariate Gaussian cluster consisting of 50 points with mean 50 and standard deviation 2 in each dimension. We simply added a single bivariate Gaussian observation with standard deviation 2 and mean randomly sampled from a uniform (60, 65) distribution. 100% of the p-values were above the significance level of 0.05.

• **Row (p) of Table 1:** We also verify the effect of density on outliers by expanding the larger cluster size from 50 to 250, and moving outliers further away until recognized as individual clusters. As the main cluster grows, we expect the effect of the outlier to diminish, and the dip test reflects this drop in relevance from 100% to only 60% of the p-values below 0.05. Silverman’s test does not reflect this decrease, instead maintaining 100% of datasets deemed as clusterable.

• **Row (q) of Table 1:** Third, we generate a similar single cluster with three outliers. The outliers are individually generated from bivariate Gaussian distributions with the standard deviation of 2. The mean of each outlier is randomly sampled from (40, 55) to (45, 60), (65, 65) to (70, 70), (65, 45) to (70, 50), respectively. Here we see similar results to the original outlier data, as Dip proves unclusterable results of 8.2% clusterability, yet the Silverman Test display 97.6% of the P-values below the significance level of 0.05. Our approach successfully accounts for the outliers as its own small cluster when arranged with structure, or noise when structure is absent.
• **Row (b) of Table 1**: We simulated two well-separated clusters using 50 point bivariate Gaussian distributions, each dimension centered with the mean values between 30 and 50 and a standard deviation of 2. The p-values of the Dip and Silverman tests both have all 1000 results below the significance level of 0.05 and classifying the dataset as clusterable.

• **Row (g) of Table 1**: Next we generate three separated Gaussian clusters. The centers of each 50 point 2-dimensional cluster are located at (35, 40), (65, 40), (50, 60). Then, the tri-variate clusters had standard deviation of 2. The multimodality tests both confirm the expected results with all 1000 p-values below the significance level of 0.05, indicating clusterability.

• **Row (h) of Table 1**: Expanding upon the dimensions of the prior dataset, we observe tri-variate clusterable structure. Maintaining the Gaussian distributed clusters, we now center all dimensions with mean values of 20, 40, and 60 for each cluster, respectively.

• **Row (c) of Table 1**: Many prior notions of clusterability fail when clusters are not well separated. We simulated data with three nearby Gaussian clusters each consisting of 50 points with a standard deviation of 2, centered around (30, 20), (40, 20), (35, 30). Our approach successfully yields clusterable results on all 1000 simulations.

• **Row (d) of Table 1**: One notorious test for prior notions of clusterability is noise robustness, so we generated data consisting of the same three well separated clusters discussed earlier (yielding clusterable results) additionally, we concatenates 80 points of noise with 2 Gaussian generated dimensions centered around (50, 50) with a standard deviation of 30; effectively producing 35% noise in the dataset. Both Silverman and Dip tests confirmed clusterable results all 1000 runs.

• **Row (f) of Table 1**: Next, we explore the effects of varying cluster density and spread, Simulating a dataset with three Gaussian clusters centered around (35, 40), (65, 40), and (50, 70). First, we generate all clusters with standard deviation of 2, and clusters consisting of 100, 75, and 50 points, respectively.

• **Row (e) of Table 1**: Then, we generate three clusters with 50 points, but standard deviation 1, 3, and 5, respectively. In each case, the desired result is clusterable - and our approach encapsulates this as all 1000 simulations yielded P-values below the significance level of 0.05.

• **Row (i) of Table 1**: We also generate high-dimensional clusterable data, simulating two well separated 10-dimensional clusters. Using the same approach as simulating a single 10 dimensional cluster, we create 10 independent dimensions of 50 points each for
each cluster. Dimensions centered randomly between 10 and 20 with a standard
deviation of 2 for the first cluster, and dimensions with a mean randomly generated from
the uniform (30, 40) distribution and a standard deviation of 2 for the second cluster. All
p-values from both the Dip and Silverman tests were below the 0.05 significance level,
indicating the data is clusterable, as expected.

- **Row (j) of Table 1**: Expanding on this set, we simulate four 10-dimensional, well
separated clusters. The first two clusters were as described in the preceding paragraph.
The third is a 10-dimensional Gaussian cluster consisting of 50 points and a standard
deviceation of 2, each respective dimension centered between randomly selected centers
from 50 to 60. The fourth is a similarly generated cluster with dimensions having
standard deviation 2 and mean randomly distributed from 70 to 80 each. Our approach
proves clusterable on this high dimensional data, as the p-values for all 1000 simulations
are below the significance level 0.05. Thus, our solution is robust to high dimensional
data.

- **Rows (k) and (l) of Table 1**: Finally, with all prior test results proving reliant, we
generated 2 distinct datasets consisting of ten clusters in 2 dimensions each. These sets
contain clusterable structure of well-separated clusters, and nearby clusters, respectively.
Each dataset is constructed using a standard deviation of 2 with 30 points per cluster.
The ten well-separated clusters are centered around (35,40), (65,40), (50,60), (50,15),
(80,20), (10,60), (15,25), (35,80), (85,90), and (95,50). The ten clusters in close vicinity
of each other are centered around (30,20), (40,20), (35,30), (25,10), (25,30), (35,10),
(50,30), (30,40), (20,20), and (45,40). Over the 1000 simulations, the well separated
dataset results are all below the significance level of 0.05 indicating very clusterable data.
The nearby cluster simulation results indicates 97.9% clusterability for Dip, and 100%
clusterability for Silverman, proving our notion of clusterability maintains integrity over
many clusters.

### 3.2 Visualizing the data

In this section, we observe the data in graphical form along with the histograms generated by
the pairwise dissimilarities. The figures feature 2, 3, and 10 dimensional data ranging from 1 to
10 clusters. As there are no previous practical notions of clusterability, visualizing the data was
helpful for assessing the utility of our new approach. As such, most datasets were generated in
2D. The data is reproducible by following the directions in the previous section. Each simulation
is generated 1000 times, and we present one of these simulations for each of the 17 types of data we study. Observe that because we generate each of the 17 types of data 1000 times, it is unrealistic to include all visualizations. However, we display one visualization for each of the 17 types, showing 2D projections for the high-dimensional data. 6 of the 17 types appear in Figure 2. In particular, rows (a) through (f) in Table 1 correspond to the same rows in Figure 2. The complete set of 2D projections for Table 1 appear below.

Figure 3: A sample projection of the **Single Cluster 2D** dataset, which corresponds to row A in Table 1.
Figure 4: This figure shows 2 of the dimensions existing in the **Single Cluster 3D** dataset, corresponding to row M in Table 1.

Figure 5: This figure shows 2 of the dimensions existing in the **Single Cluster 10D** dataset, corresponding to row N in Table 1.
Figure 6: This figure shows the **Single Cluster with Outlier 2D** dataset, which corresponds to row O in Table 1.

Figure 7: This figure portrays the **Single Large Cluster with Outlier 2D** dataset, which corresponds to row P in Table 1.
Figure 8: This figure shows the **Single Cluster with Three Scattered Outliers 2D** dataset, corresponding to row Q in Table 1.

Figure 9: This figure shows the **Two Separated Clusters 2D** dataset, which corresponds to row B in Table 1.
Figure 10: This figure depicts two of the dimensions from the Two Separated Clusters 10D dataset, which corresponds to row I in Table 1.

Figure 11: This figure shows the Three Separated Clusters 2D dataset, which corresponds to row G in Table 1.
Figure 12: This figure shows two dimensions of the **Three Separated Clusters 3D** dataset, corresponding to row H in Table 1.

Figure 13: This figure shows the **Three Close Clusters 2D** dataset, corresponding to row C in Table 1.
Figure 14: This figure shows the **Three Noisy Clusters 2D** dataset, which corresponds to row D in Table 1.

Figure 15: This figure shows the **Three Clusters with Different Diameters 2D** dataset, which corresponds to row E in Table 1.
Figure 16: This figure shows the Three Clusters with Different Densities 2D dataset, which corresponds to row F in Table 1.

Figure 17: This figure depicts two of the dimensions from the Four Separated Clusters 10D dataset, which corresponds to row J in Table 1.
Figure 18: This figure shows the **Ten Separated Clusters 2D** dataset, which corresponds to row K in Table 1.

Figure 19: This figure shows the **Ten Close Clusters 2D** dataset, which corresponds to row L in Table 1.
3.3 Summary of results

We examine the results summarized in Table 1, beginning with clusterable data that consists of multiple Gaussian clusters. Rows (b) through (l) of Table 1 summarize these results, with different number of clusters, ranging from 2 to 10, either 2, 3, or 10 independent dimensions, varying degrees of separation between the clusters, and different cluster sizes and diameters. On most clusterable datasets, both Dip and Silverman correctly identified that the data is clusterable 100% of the time, as represented with a score of 1 in the table. In rows (c), (d) and (l) we find scores of 0.979 to 1.000 for dip test, and 0.997 to 1.000 for the Silverman test. This means that on 3 out of 1000 datasets with “3 close clusters in 2D,” the dip test incorrectly identified the data as unclusterable (and similarly for the other two cases). In summary, the power for our simulations ranged from 97.9% to 100%.

Consider Table 1(a) and (m)-(q), summarizing the results of our simulations for single cluster data. Table 1(a), (m) and (n) all consist of a single Gaussian distribution, in dimensions 2, 3, and 10 respectively. Both Dip and Silverman generally conclude that the given data is unclusterable with excellent accuracy, namely 100% for the Dip test and over 95% for Silverman.

When outliers are introduced to unclusterable data, we observe a disparity between Dip and Silverman. In particular, rows (o), (p), and (q) represent a single Gaussian cluster and a small number of outliers. Here we find that the dip test identifies such data as unclusterable, while Silverman classifies it as clusterable. Where the dip test is robust to outliers, the Silverman test allows for small clusters. This goes back to the inherent ambiguity of clustering; for some applications, small clusters are acceptable, while for others, robustness to outliers is desired. In fact, the same phenomenon is observed with clustering algorithms, where some tend to identify small clusters, while others effectively view such data as outliers [6].

In sum, our simulations indicate that the use of multimodality tests on the set of pairwise distances is an effective and accurate method of classifying datasets by their level of clusterability. Both clusterable and unclusterable datasets were identified as such in almost all simulations. Different behavior in the presence of outliers allows users to select which test to use based on how they prefer to treat small clusters.

Lastly, we note that we would have liked to compare our new notion of clusterability with previous ones; however, in this respect there is a major difficulty. As discussed in Section 1.1, prior notions are impractical due to unreasonable computational complexity or unrealistic requirements for clusterable data.
CHAPTER 4
RESULTS ON REAL DATASETS

In this section, we demonstrate the success of our method of clusterability evaluation by analyzing real datasets from the *R datasets* package, and compare the Dip and Silverman tests on these sets.

We present clusterable and unclusterable sets and the corresponding p-values from applying the Dip and Silverman tests on the sets of their pairwise distances. To determine whether our approach was able to evaluate clusterability appropriately in these real datasets, we manually judge the clusterability of the data through two-dimensional projections of the original data. All of the datasets we present are freely available within *R* and were selected to ensure sufficient sample size and varied dimensionality. We provide examples of multi-dimensional data for applicability to real-world problems and 2D examples for ease of visualization.

<table>
<thead>
<tr>
<th>Data</th>
<th>Dip</th>
<th>Silv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. 1 cluster 2D</td>
<td>0.000</td>
<td>0.038</td>
</tr>
<tr>
<td>b. 2 separated clusters 2D</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>c. 3 close clusters 2D</td>
<td>0.997</td>
<td>1.000</td>
</tr>
<tr>
<td>d. 3 noisy clusters 2D</td>
<td>1.000</td>
<td>0.997</td>
</tr>
<tr>
<td>e. 3 clusters, varied diameters 2D</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>f. 3 clusters, varied density 2D</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>g. 3 separated clusters 2D</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>h. 3 separated clusters 3D</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>i. 2 separated clusters 10D</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>j. 4 separated clusters 10D</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>k. 10 separated clusters 2D</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>l. 10 close clusters 2D</td>
<td>0.979</td>
<td>1.000</td>
</tr>
<tr>
<td>m. 1 cluster 3D</td>
<td>0.000</td>
<td>0.042</td>
</tr>
<tr>
<td>n. 1 cluster 10D</td>
<td>0.000</td>
<td>0.036</td>
</tr>
<tr>
<td>o. 1 cluster 2D with outlier</td>
<td>0.000</td>
<td>0.985</td>
</tr>
<tr>
<td>p. 1 large cluster 2D with outlier</td>
<td>0.000</td>
<td>0.965</td>
</tr>
<tr>
<td>q. 1 cluster 2D with 3 outliers</td>
<td>0.082</td>
<td>0.976</td>
</tr>
</tbody>
</table>

Table 1: Proportion of datasets classified as clusterable over 1000 runs for each type, for a total of 17000 simulations. For Dip and Silverman, high scores indicate highly clusterable data, and low scores indicate lower cluster structure.
Table 2: Real Data p-values. This table presents the p values for the dip and Silverman multimodality tests on real datasets from the R Datasets package. Recall that p < 0.05 signals clusterable data, which larger values of p signal that data is unclusterable.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Dip</th>
<th>Silv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>iris</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>swiss</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>faithful</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>rivers</td>
<td>0.2772</td>
<td>0.0000</td>
</tr>
<tr>
<td>trees</td>
<td>0.3460</td>
<td>0.3235</td>
</tr>
<tr>
<td>USAJudgeRatings</td>
<td>0.9938</td>
<td>0.7451</td>
</tr>
<tr>
<td>USArrests</td>
<td>0.9394</td>
<td>0.1897</td>
</tr>
<tr>
<td>attitude</td>
<td>0.9040</td>
<td>0.9449</td>
</tr>
<tr>
<td>cars</td>
<td>0.6604</td>
<td>0.9931</td>
</tr>
</tbody>
</table>

We begin by considering the *Faithful* dataset [24, 10], which captures waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park. This famous data consists of 272 observations on 2 variables, numeric eruption time in minutes and numeric waiting time to next eruption. As depicted below, this real, two-dimensional dataset possesses exceptionally clear cluster structure. Furthermore, both the Dip and Silverman test agree with this conclusion (p<0.0001).

*Figure 20: This figure shows the *Faithful* dataset with natural cluster structure. Both Dip and Silverman tests captures this with P values of 0 indicating that faithful is clusterable.*
Next, we consider the famous *iris* dataset [22], depicted in Figure 21, which is known to have three clusters represented by three species of the iris flowers. The dataset gives the measurements in centimeters for the sepal length and width and petal length and width, respectively, for 50 flowers from each of the 3 species. Running the Dip test on the set of pairwise distances of the *iris* data, we obtain a p-value less than 0.0001 for both the Dip and Silverman tests indicating that both methods indeed classify the data as clusterable.

![Figure 21](image)

*Figure 21: This figure shows 2D projections of the famous *Iris* dataset, which possesses clear cluster structure. Our method for clusterability evaluation attains a p-value of 0 using both the dip and Silverman tests, indicating that *iris* is clusterable.*
The *rivers* dataset [29] contains the lengths, in miles, of 141 major rivers in North America. Depicted below in Figure 22, the dataset exhibits some inherent cluster structure if we allow small clusters. Using our approach, the Silverman test yields \( p < 0.0001 \). The Dip test results in a \( p = 0.2772 \), which is above our threshold of 0.05 and indicates a lack of strong cluster structure. Confirming our finding from simulated data, this suggests that the Silverman test may be more appropriate when smaller clusters are of interest, while the Dip test may be desired when the application calls for large clusters and considers small clusters to be better interpreted as outliers.

![Figure 22: This figure shows the Rivers dataset. Using the Dip test we capture the weak cluster structure with a p value of 0.2772, while Silverman test can be used to capture the small outlying clusters with a p value of 0.](image)

The *Swiss* dataset [31] consists of 6 standardized fertility measures and socio-economic indicators represented as percentages for each of 47 French-speaking provinces of Switzerland at about 1888. Our method produces low p-values (\( p < 0.0001 \)) for both the Dip and Silverman tests suggesting the data contains clusterable structure. The 2D projections below, visualizing Figure 23, confirms that this data is indeed clusterable.
Figure 23: This figure depicts the Swiss dataset’s 2D projections suggesting the data is clusterable. Both of our methods capture this clusterability with $P$ values 0.

We now turn to data that appears unclusterable, starting with USArrests [29], depicted in Figure 4. It contains 50 observations and 4 variables: arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973, as well as percent urban population. Our findings yield $p = 0.9394$ for Dip and $p = 0.1897$ for the Silverman test, both of which are above 0.05, implying that the data is not considered clusterable by our methods.
Figure 24: Two-dimensional projections of the USArrests data, suggesting that the data is inherently unclusterable, as confirmed by the p-values of 0.9394 and 0.1897 for the dip and Silverman tests, respectively.

The next dataset we consider, attitude [17], comes from a survey of the clerical employees of a large financial organization, aggregated from the questionnaires of the approximately 35 employees for each of 30 (randomly selected) departments. The data consists of 30 observations on 7 variables, such as their evaluation of handling of employee complaints and opportunity to learn. The data, visually lacking strong cluster structure as shown in Figure 25 below, is in fact classified as unclusterable, using our tests on pairwise distances, yielding a p-value of 0.9040 for the Dip test and $p = 0.9449$ for the Silverman test.
Figure 25: Two-dimensional projections of the Attitude dataset suggesting inherently unclusterable structure. Our method captures this inherent structure, yielding p-values of 0.9040 and 0.9449 for the Dip and Silverman tests, respectively.

We now turn to the analysis of higher dimensional unclusterable data, considering the dataset USJudgeRatings [27] represents lawyers' ratings of state judges in the US Supreme Court, containing 43 observations on 12 numeric variables. As shown below, this dataset appears to be unclusterable. Our methods make the same conclusion with p=0.9938 for the Dip test and p=0.7451 for Silverman. Figure 26 below visualizes the many dimensions of this dataset.
The next data set we consider is the 2-dimensional data set, cars [21], consisting of 50 elements and two variables, representing speed and stopping distance. This data was recorded in the 1920s. This data set is interesting, having an oval shape, within which data is fairly evenly distributed.

Figure 26: Two-dimensional projections of the USJudgeRatings data, depicting an unclusterable structure. Our methods come to the same conclusion with $p=0.9938$ for the Dip test and $p=0.7451$ for Silverman.
distributed, particularly possessing no obvious distinct clusters. Our methods similarly identify this data as unclusterable, with a $p=0.6604$ for the dip test and $p=0.9931$ for the Silverman test. See Figure 27 below for the visualization of this data.

![Figure 27: This figure displays the Cars dataset, showing no distinct clusters. Our method agrees with $p=0.6604$ for the dip test and $p=0.9931$ for the Silverman test, both indicating the data to be unclusterable.](image)

The trees [33] data set is depicted in Figure 28 below. This dataset provides measurements of the girth, height and volume of timber in 31 felled black cherry trees. The tests results are $p = 0.346$ for the Dip test, and $p=0.3235$ for Silverman. Both test results are above 0.05, showing that both modality tests classify the data as unclusterable.
Figure 28: This figure depicts the two-dimensional projections of the Trees data. No strong cluster structure is apparent, and our results, $p = 0.346$ for dip and $p=0.3235$ for Silverman, agree that no structure exists.
CHAPTER 5
CONCLUSIONS AND DISCUSSION

In this thesis, we introduce the first practical method for clusterability evaluation that (1) captures how real data is structured, and (2) is computable in low polynomial time. Our approach applies multimodality tests to the (one-dimensional) set of pairwise distances based on original -- potentially high-dimensional -- data, facilitating easily interpretable results based on sound statistical theory. We demonstrate the utility of our approach on extensive simulations and real data found in the R datasets package.

We emphasize that until now, no experimental study of this type has been done. Previously, notions of clusterability were presented without justification, relying entirely on the formulation of these notions being intuitive. The approach presented in this thesis is the first to be supported by empirical results.

Our development of an effective and efficient approach for evaluating clusterability enables several important areas of investigation. Firstly, it allows us to investigate the CDNM thesis, which asks whether “clustering is difficult only when it doesn’t matter.” This would involve exploring Ben-David’s [14] remaining desirable characteristics for notions of clusterability, investigating which algorithms are best suited for uncovering the underlying cluster structure when our approach detects the presence of such structure.

Another avenue for future work is closer investigation of statistical tests of multimodality, such as the application of alternate tests and development of novel multimodality methods that would allow for more fine-grained clusterability analysis. Lastly, we look forward to the application of this new approach for clusterability analysis to practical applications as well as for theoretical investigation, where it will for the first time allow for bridging theory and practice by testing theoretical results on real data.
REFERENCES


PERSONAL
Andreas Adolfsson
Department of Computer Science
Florida State University
Tallahassee, FL 32304

EDUCATION
Master of Science in Computer Science
Florida State University, Tallahassee, FL
November 2017
Bachelor of Science in Computer Science
Florida State University, Tallahassee, FL
November 2015
Minors in Mathematics

TECHNICAL SKILLS
Programming Languages: Java, C, C++, C#, SQL, Prolog, MIPS, R, UNIX, LaTeX, and HTML
Platforms: Windows (95, 98, 2000, XP, Vista, 7, 8, 10), UNIX, Linux, and Mac OS
Tools: MS Office, g++Dreamweaver, and LaTeX

RELATED EXPERIENCE
Research Assistant, June 2015 - Present
Florida State University, Tallahassee, FL
• Developing empirical theory for quantifying clusterability by parsing pairwise distances and calculating statistical data
• Exploring the effects of inherent structure of arbitrary datasets on clustering behavior by generating numerous graphical and numerical works
• Discovering existing statistical measures for determining modality by exploring published research papers

TEAMWORK & INTERESTS
Scrum Master - Experienced with agile development methods and scrum meetings
Actor – Practiced 3 years of public speaking and team building exercises
Theatre Ambassador – Contributed 3 years of theatre, light, and sound management
Caretaker – Assisted special needs and partially paralyzed individuals for 2 years

LANGUAGES
• English | Native Speaker
• Swedish | Native Speaker
• Norwegian | Basic communication
• ASL | Basic communication