To Cluster, or Not to Cluster: How to Answer the Question

Andreas Adolfsson
Florida State University
Tallahassee, FL, USA
ada10j@my.fsu.edu

Margareta Ackerman*
San Jose State University
San Jose, CA
margareta.ackerman@sjsu.edu

Naomi C. Brownstein*
Florida State University
Tallahassee, FL
naomi.brownstein@med.fsu.edu

ABSTRACT

Clustering is an essential data mining tool that aims to discover inherent cluster structure in data. For most applications, applying clustering is only appropriate when cluster structure is present. As such, the study of clusterability, which evaluates whether data possesses such structure, is an integral part of cluster analysis. However, methods for evaluating clusterability vary radically, making it challenging to select a suitable measure. In this paper, we perform an extensive comparison of measures of clusterability and provide guidelines that clustering users can utilize to select suitable measures for their applications.

ACM Reference format:
Andreas Adolfsson, Margareta Ackerman*, and Naomi C. Brownstein*. 2016. To Cluster, or Not to Cluster: How to Answer the Question . In *Proceedings of Knowledge Discovery from Data, Halifax, Nova Scotia, Canada, August 13–17 (TKDD'17),* 9 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Clustering is an ubiquitous data analysis tool applied in virtually all disciplines, spanning applications as diverse as bioinformatics, marketing, and image segmentation. Its wide utility is perhaps unsurprising, as its intuitive aim - to divide data into groups of similar items - applies at various stages of the data analysis process, from exploratory data analysis to collaborative filtering.

Despite its popularity, we have barely scratched the surface on many fundamental questions about clustering. Issues as basic as the definition of clustering are being raised [2, 41]. Differences between clustering algorithms are studied to decide which should be used under different circumstances [4–7]. Yet, a more fundamental issue than algorithm selection is when clustering should – or should not – be applied. For most applications, clustering is only appropriate when cluster structure is present in the data. Otherwise, the results of any clustering technique become arbitrary and potentially misleading.

For concreteness, consider a dataset randomly generated from a single Gaussian distribution. Because the data contains only one cluster, further sub-division would be artificial. Most clustering algorithms (e.g. k-means with $k \geq 2$) would find multiple clusters in the data, even though no multi-cluster structure is present.

As such, the application of these data mining tools rely on the presence of inherent structure, rendering *notions of clusterability*, which aim to quantify the degree of cluster structure, integral to cluster analysis. Clusterability analysis should precede the application of clustering algorithms, as the success of any clustering algorithm depends on the presence of underlying cluster structure.

To see how clusterability fits within the clustering process, consider the clustering pipeline depicted in Figure 1.[1] The process begins with data preprocessing, often involving feature selection or extraction. Next, clusterability analysis determines whether the data possesses inherent cluster structure. If the 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 need to be reprocessed. On the other hand, if the data is found to be clusterable, a suitable clustering algorithm may be selected or developed.[2] After the algorithm is executed, the solution is validated by applying clustering quality measures [2, 47], which may result in the selection of an alternate clustering algorithm if a sufficiently high quality clustering has not been found.

Notions of clusterability have been proposed across the computer science and statistics literature, summarized in Section 2. Not unlike clustering algorithms [7], notions of clusterability disagree with each other in surprising ways. A formal analysis by Ackerman and Ben-David [3] reveals that many notions of clusterability are pairwise distinct - despite the fact that they all attempt to evaluate the same characteristic!

The plethora of clusterability methods presents a dilemma: how should one select a clusterability measure suited to their data?[3] Ben-David [17] approaches the problem from a theoretical standpoint, offering several properties that notions of clusterability should satisfy. In particular, he argues that clusterability notions need to be both computationally efficient and effective. (See section 2.1 for more details.)

The effectiveness requirement is complicated by the inherent ambiguity of cluster analysis. As with clustering algorithms, the needs of the application at hand may dictate whether the given data is clusterable. For example, whether we

*These two authors contributed equally.

TKDD'17, Halifax, Nova Scotia, Canada

DOI: 10.1145/nnnnnnn.nnnnnnn

[1] A similar pipeline is presented in the famous survey of clustering algorithms by Xu and Wunsch [57], sans the second 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 [7, 9].
[3] The original "user's dilemma" refers to the problem of selecting a clustering algorithm for a given task. Selecting a notion of clusterability is another dilemma that the clustering user faces and that we address in the current work.

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

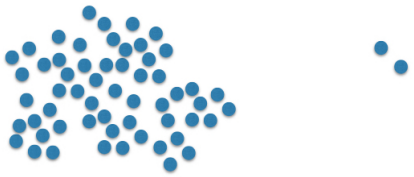

Figure 2: A dataset with ambiguous cluster structure. The outliers can either be ignored or represent a small cluster. Whether or not this data is considered clusterable depends on the needs of the given application.

allow small clusters can change how we evaluate the clusterability of the data in Figure 2; If small clusters are appropriate for the given application, then the data would be clusterable, whereas otherwise it would be unclusterable.[4] Such considerations make room for multiple legitimate clusterability measures, and create the need for guidelines that would help a user determine which notion to choose for their application.

In this paper, we take a practical approach to clusterability, and analyze notions of clusterability for their effectiveness on a variety of datasets. This allows us not only to identify effective notions, but also to discover important differences amongst them that can enable a clustering user to make informed decisions when selecting a clusterability technique for their application.

We now outline the paper. Section 2 presents an overview of clusterability measures, starting with several properties that we use to select measures for our analysis. We then present our simulations, allowing us to discern between approaches to clusterability and determine which may be more appropriate under different circumstances. Next, we show our analysis of these measures of clusterability on real data. We conclude with a summary of our findings and recommendations.

2 MEASURES OF CLUSTERABILITY

Many approaches for measuring clusterability have been proposed in the literature. In the section, we survey the most promising measures, for which we perform an extensive analysis in Section 3. Before delving into the details of the measures, we formalize clusterability and propose several requirements.

[4]For example, distant elements are typically viewed as significant when clustering Phylogenetic data, whereas outliers are often best ignored when clustering is applied to market segmentation. See [5] for a detailed discussion.

2.1 Requirements of clusterability measures

One of the most challenging aspects of cluster analysis is that it is ill-defined [2]. In particular, we do not have a formal definition of clusterability (or even a formal definition of clustering[5]). Recently, Ben-David [17] began tackling the challenge of formalizing clusterability by proposing several interesting properties. In this section, we aim to distill several properties that will help sift through the plethora of clusterability measures in order to identify those that are most likely to be useful in practice. (Two of our properties, the first and third, are related to Ben-David's requirements.)

A *measure of clusterability* is a function that takes in a dataset, and outputs a number that represents its degree of inherent cluster structure.[6] Naturally, this concept is insufficiently detailed, as a function (e.g. a constant function that declares all datasets to be clusterable) can easily contradict our intuition about how a measure of clusterability should behave. An important question remains: What additional requirements are needed for a clusterablility measure to be meaningful? We propose several properties on which we rely to select clusterability measures for our analysis in Section 3:

- *Efficiency*: For practical utility, a measure of clusterability should be efficiently computed.[7]
- *Algorithm independence*: The clusterability measure should not be based on a specific clustering algorithm or objective function.
- *Effectiveness*: The measure of clusterability should be highly accurate in identifying data as clusterable or unclusterable.[8]

The first, and simplest, requirement asks that measures be efficient in practice. They should certainly be computable in polynomial time, but, to have real practical utility, they should run in reasonable time on fairly large datasets. Our second requirement is concerned with the role of clusterability in the clustering pipeline in Figure 1, discussed in the introduction. Notions of clusterability that are based on a specific algorithm ask a different question than the one with which

[5]Many different axioms and properties have been proposed, see, for example [2, 7, 41]. However, we do not yet have a formal definition of clustering, clustering functions, or clusterability.

[6]The output could be a real value, a binary indicator ("clusterable" or "unclusterable"), or a probability based measure, such as a p-value (testing the null hypothesis that the dataset is "unclusterable" against the "clusterable" alternative.)

[7]This requirement relates to Ben-David's [17] third requirement.

[8]This property is related to Ben-David's [17] first requirement. As discussed in Section 1, the inherent ambiguity of clustering necessitates some flexibility on what it means to be "clusterable." Yet, there exist clear examples (such as a single Guassian (unclusterable) or two well-separated Guassians (clusterable)) that let us evaluate the effectiveness of a clusterability notion.

we are concerned here; While we ask "Is this data clusterable?", algorithm-specific notions aim to discover whether the data can be clustered using a particular clustering technique. Furthermore, since different clustering algorithms are apt at identifying distinct types of cluster structure [8, 9], centering a measure of clusterability on a specific algorithm restricts a notion of clusterability from identifying structure that the underlying algorithm cannot capture.

Finally, the third and most challenging requirement is the focus of our work. We first collect a body of existing measures and propose additional measures that satisfy the first two requirements. Next, we empirically compare the performance of these methods on a large number of real and simulated datasets. Our data includes many examples that leave no room for ambiguity, allowing us to determine which clusterability measures are effective. Differences in their behavior on more ambiguous data allow us to identify guidelines that can be used to help clustering practitioners select suitable notions of clusterability for their tasks.

2.2 Effective approaches to clusterability evaluation

Our survey of clusterability notions suggests that a large, practical class of clusterability notions rely on one or more of the following: dimensionality reduction and statistical tests of multimodality, which indicate whether or not multiple clusters are present in the data. The main insight these approaches is the observation that clusterability can be inferred from a one-dimensional view of the data. Briefly, these methods look for separations in the data that would indicate that the presence of separate clusters. In the following two subsections, we briefly review data reduction methods and multimodality tests, before delving into clusterability methods.

2.2.1 Data reduction methods. Contemporary datasets often contain a large number of features, which may greatly outnumber the observations. Due to the computational and theoretical challenges associated with high dimensional data, a popular solution is to reduce the dimension while maintaining the structure of the original data. The reduced dataset informs the clusterability of the original data. This section discusses techniques for reducing data to one dimension.

One of the most famous data reduction methods is principal component analysis (PCA) [39], which projects the data onto independent dimensions that explain the original variance. There are natural connections between PCA and clustering. In fact, the principal components (PC) correspond to the k-means cluster membership indicators [21, 58]. PCA has been recommended to visually inspect for grouping structure [1]. While multiple components are often retained, the first PC, by definition, explains most of the variation in the data [37, 42]. Importantly, PCA is less prone than other data-reduction methods to the curse of dimensionality [37]. Despite its benefits, PCA is not well suited to non-linear structures [37], for which principal curves [33, 55], which produce a non-linear transformation of the data, may be more appropriate.

The set of dissmilarities between each pair of points in a dataset forms an alternative one-dimensional summary. Dissimilarities serve as inputs of many clustering techniques, can be calculated for any dataset, and have been shown to preserve structural features, such as correlation [28]. Euclidian distance is the most common metric, though others are possible, such as the Pearson correlation. Yet, distances are sensitive to the curse of dimensionality, potentially leading to misleading results for data with a large number of features. Additionally, the use of pairwise distances increases the sample size of the summary to nearly the square of the original size; this may be impractical for datasets with a large number of observations.

2.2.2 Multimodality tests. Intuitively, if a dataset contains multiple clusters, then there should be some separation between the clusters. For example, a histogram of the set of pairwise distances would likely show a group of small distances, representing those within clusters, and a group of large between-cluster distances. By contrast, homogenous data would not show such a separation. One could use a statistical test to determine if the set of distances for a given dataset has multiple modes, indicating that there are indeed multiple clusters. Similarly, statistical tests on data reduced by other methods help detect cluster structure. Multimodality tests are employed for other clustering purposes, such cluster splitting, merging, and validation [34, 40, 45, 52].

Numerous statistical tests for multimodality have been developed [44]. Each test provides a p-value, which is the probability of observing the given input or a more extremely multimodal input assuming that the data is generated from a unimodal distribution. If only a single mode is present, then the p-value should be large, indicating that the underlying data is deemed unclusterable. On the other hand, small p-values make us question the original assumption of unimodality and instead conclude that multiple modes (and multiple clusters) are present in the population from which the data was generated.

We discuss two tests: dip and Silverman. Hartigan's dip test [31] rejects the assumption of unimodality if the observed data is sufficiently different from the closest possible uniform distribution. The dip test has been used to calculate the number of clusters, to find suitable clusters, and to test for clusterability [40]. Silverman [53] is based on the kernel density estimate. The technique approximates the empirical distribution of the observed data with a set of Gaussian distributions. If a sufficiently large bandwidth is required to produce a unimodal empirical distribution estimate, then the test concludes that the underlying data distribution is multimodal, comprising a mixture of distinct Gaussian distributions.

One may be tempted to forgo dimension reduction, apply a multimodality test, and conclude that the data is clusterable if the dataset rejects the null hypothesis for unimodality [20]. Unfortunately, the asymptotic behavior of these tests is unknown when the data is multi-dimensional [20, 32, 42, 56]. This severe limitation renders these methods unpredictable for real datasets, most of which have multiple, if not high dimensions, unless the user first reduces the data to one dimension.

We now introduce previous notions of clusterability, starting with those that rely on a combination dimensionality reduction and multimodality tests. We then discuss the Hopkins statistic, which tests for spatial randomness, and introduce several new notions. This section concludes with a brief summary of notions of clusterability that are not well-suited to the goals of the current analysis.

2.3 Clusterability via multimodality

*2.3.1 Dip Test on Pairwise Distances (**Dip-dist**).* Dip-dist [40] tests for clusters in the set of dissimilarities using the Dip test [31]. The lengths of the pairwise distances are sufficient for clusterability analysis without needing to consider how the distances are arranged to form the data. Multiple modes in the distance distribution suggest the presence of multiple clusters.

*2.3.2 Silverman Test on Principal Curve (**Pcurve Silv.**)* To combat the curse of dimensionality, one recommendation is to use Silverman's test[9] for multimodality of principal curves [10]. The first dimension of the principal curve is extracted and Silverman's test is used to determine if that dimension is unimodal or multimodal. A multimodal principal curve suggests that the original, higher dimensional data exhibits cluster structure.

*2.3.3 Silverman Test on Principal Component (**PCA Silv.**)* A linear alternative to the principal curve is to extract the first principal component of the data [10]. This one dimensional transformation explains the maximum amount of variation in the data. The method then applies Silverman's test to this first principal component.[10] [11] A multimodal first principal component reveals that a linear transformation that explains most of the variance of the data containis clusters and hence suggests that the original data is clusterable.

*2.3.4 Classic Methods (**Classic Dip and Classic Silv.**)* While it is known that multimodality tests may be problematic in higher dimensions, we include these methods in our comparisons for completeness. That is, Classic Silverman and Classic Dip conduct, respectively, Silverman's and the Dip test of multimodality on the original, multi-dimensional data.

2.4 Clusterability via Spatial Randomness

Another method by Hopkins [36, 43], (**Hop.**), a test of spatial randomness, tells us if any feature is distributed non-randomly across the dataset. Hopkins compares the distances between a sample of data points and their nearest neighbors to the distances from a sample of pseudo points – with each feature randomly selected from the full dataset – and their nearest neighbors. If the data are not distributed in clusters, then both sets of distances should be similar on average. The Hopkins statistic is calculated based on these distances [36, 43].

Clusterability can be inferred by comparing to a threshold calcuated based on the distribution of the Hopkins statistic. Under the null hypothesis that the data is unclusterable, the test statistic follows a beta distribution with both parameters equal to the number of points selected to sample n [36, 43]. Thus, Hopkins' statistic should be compared to a Beta quantile $q_\alpha(n, n)$. [12] [13] Yet, the choice of n requires caution. According to [43], "if too few points are chosen, then the nearest-neighbor distances chosen will not be representative of the entire distribution of distances. If too many points are chosen, Dubes and Zeng [22] warn that the assumptions about the Beta distribution will be invalid." Previous authors recommend sampling $5 - 10\%$ of the data [16, 43]. In this paper, we calculate the Hopkins statistic using a 10% sampling rate.

2.5 New Clusterability Methods

We combine and compare clusterability measures, as well as proposing additional approaches for studying and evaluating clusterability. This section focuses on an exposition of our additional approaches, while the following two sections provide extensive simulations and results from real datasets.

Since both the dip and Silverman's test are valid statistical tests of multimodality, we propose to use both on each reduced version of the data. Thus, to our knowledge, the following methods below have not been proposed in the literature and need to be briefly described.

*2.5.1 Silverman's test on dissimilarities (**Silv.-dist**).* Rather than using the dip test on the set of pairwise distances [40], we propose to use Silverman's test, with the necessary correction [29].

*2.5.2 Dip test on principal component (**PCA Dip**).* In [10], Instead of using Silverman's test of whether the first principal component is multimodal, this method uses the dip test.

*2.5.3 Dip test on principal curve (**Pcurve Dip**).* Similarly, this method uses the dip test to classify the principal curve of a dataset as unimodal or multimodal.

2.6 Quantifying Efficiency: Runtime

Selecting a suitable clusterability measure involves both qualitative and quantitative considerations. In this work, we focus on qualitative analysis, exploring which measures are most effective and differentiating them based on the types of cluster structures that they identify. Nevertheless, quantitative considerations remain an important part of the process of selecting a suitable measure. There are significant differences in the computational complexity of clusterability techniques that render some of them impractical when the number of elements (n) or the dimension (d) is large.

Classic Dip is linear in n [42]. Hopkins, Classic Silverman, and Dip-Dist have quadratic running time in n. The

[9] Corrections for both Silverman's original test and the dip test have been proposed [19, 29], but only the correction for Silverman is available in standard software, such as *R*.

[10] Prior to PCA, the data is centered about its mean.

[11] Computationally, PCA is performed using the singular value decomposition of the centered data, and the rotated variables are extracted.

[12] The Beta quantile defined as the value such that, assuming the data was generated without clusters, the chance of concluding that the data is clustered, i.e. $P(H < q_\alpha(n, n))$ is $100\alpha\%$. We use a one-sided test because if the data is more spatially random than expected by chance, it would still be unclusterable.

[13] Note that the Hopkins statistic approaches a Gaussian distribution for large samples (e.g. $n > 50$) In this case, one could instead use the threshold $0.5 - z_{1-\alpha}/(2\sqrt{2n+1})$ where n is the number of points sampled.

dimensionality of the data impacts the running time of PCA-based approaches, with PCA dip having asymptotic running time of $O(nd^2 + d^3)$ [23]. Silverman-dist is bounded by a quadric function in n. Finally, PCA Silverman has complexity of $O(n^2 + d^2n + d^3)$.[14]

2.7 Other Clusterability Methods

Many notions of clusterability have been proposed to study clustering from a theoretical standpoint or investigate specific clustering paradigms. As such, there are some notions in the literature that have been omitted from our study, as they are either impractical or otherwise unsuited to our goals.

Measures used for theoretical analysis are often NP-hard to compute [3]. Since we seek notions that are efficient, and thus applicable in practice, we had to omit all such measures from our analysis. Many other notions, based on specific algorithms or objective functions [3, 11, 12, 14, 45, 49], are also omitted from our comparative analysis, as our study is concerned with identifying the presence of any cluster structure, not only that which can be discovered by a specific clustering technique.

Even before delving into the analysis, the effectiveness requirement allowed us to eliminate several notions of clusterability (strict separation [15] and worst pair ratio [24]). Despite their elegance and utility for theoretical analysis, these notions are too strict for practical application due to their high sensitivity to noise and outliers. Lastly, some approaches to clusterability, such as [54], rely on subjective judgment, and do not provide a quantifiable measure. Such measures are also omitted from our analysis.

3 SIMULATIONS

Our extensive simulations evaluate each approach to clusterability using all clusterability tests in Sections 2.3-2.5. The simulations include 35 types of datasets, each generated with the same parameters 1000 times, for a total of 35,000 simulations. The code is found at this link: http://www.cs.sjsu.edu/~ackerman/clusterability.R. Simulations consist of clusters generated from one or more Gaussian or t-distributions and sometimes with a small number of outliers. For example, row (12) of Table 1 describes the results for three bivariate Gaussian clusters, each consisting of 50 points, with means at $(30, 20)$, $(40, 20)$, and $(35, 30)$ and standard deviations of 2. Chaining data is also simulated with one or two lines and two, three, or five circles. Simulations were performed in *R* version 3.3.2. Due to space limitations, the appendix, found at http://www.cs.sjsu.edu/~ackerman/appendix-clusterability.pdf, includes further details with all parameters used in each set of simulations.

In Table 1, we record the the percentage of datasets on which the test yielded a p-value less than 0.05, indicating that the tests rejected the null hypothesis of unimodality at the traditionally used 5% significance level. *High values in Table 1 indicate high values of clusterability, while low values indicate poor clusterability.* For unambiguously unclusterable datasets, the proportion of rejections corresponds to type I error, the rate of erroneously classifying datasets generated without clusters as clusterable. Type I error greatly exceeding 5% indicates that the method is invalid and produces excessive false positives. For unambiguously 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. Higher power is desirable. The ranks of the power values tells us which methods are more likely to capture the structure in clustered data. However, results are complicated by the ambiguous nature of clustering. When a small number of outlying points are present, the decision to classify the data as clusterable depends on whether outliers should be considered as small clusters.

Finally, because fitting a principal curve a large number of times may induce convergence problems, we record the proportion of the time that the principal curve converged properly within 1000 iterations.

3.1 Type I error: Unclusterable Data

First, consider data generated without cluster structure. Principal curve methods were invalid, concluding that most single-cluster datasets were clusterable at a much higher rate than 5%.[15] Hopkins, PCA-Silverman and classic Silverman have type I error around 5% as expected in two dimensions.[16] For all other cases, distance based methods have excessively low type 1 error of less than 1%, indicating that they may be overly conservative. Similarly, Hopkins statistic has very low type 1 error for data with more than 2 dimensions. Overall, all methods except principal curves have reasonably low false positive rates for single Gaussian clusters.

3.2 Performance with outliers and small clusters

When outlying points are introduced to otherwise unclusterable data, one could argue either for or against clusterability. Methods vary in their conclusions: Dip-based methods classify the data as unclusterable, while the Hopkins statistic and Silverman-based methods classify such data as clusterable, considering the outliers as separate clusters. Similarly, rows (8), (9), and (10) feature single t-distributed clusters with 5, 10, and 15 degrees of freedom.[17] Dip-based methods consider the data clusterable less than 10% of the time, even for 5 degrees of freedom, when multiple outliers are likely; Hopkins and Silverman-based methods conclude that the data is clusterable

[14] As shown in [23], PCA Silverman requires $O(nd^2+d^3)$ operations to calculate the principal component. Then it performs Silverman's test, which is bounded by $O(n^2)$. Therefore, the total complexity is $O(n^2 + d^2n + d^3)$.

[15] While principal curves converged properly over 90% of the time on these datasets, they overwhelmingly failed to converge for linear data.

[16] For valid methods, values in Table 1 should be below or reasonably close to 0.05. If the true false positive rate is 5%, then we would expect with 95% confidence that the observed value should be below $0.05 + 1.96 * \sqrt{0.05 * 0.95/1000} \approx 0.064$. Based on this threshold, PCA Silverman has slightly inflated type I error in 50 dimensions, and Classic Silverman has inflated type I error in 3 dimensions. However, because we would expect 5% of the results for unclusterable datasets to exceed this value, it is not unusual to see 2 results with slightly inflated type I error rates. In fact, if we adjust for the total number of comparisons for unclusterable data, then the false positive rates would be compared to a different threshold (0.072) and would not be considered excessive. Additional simulations would be needed to confirm that the false positive rates are indeed controlled.

[17] T-distributions with small degrees of freedom are highly likely to have outliers. As the degrees of freedom increases, the distribution will converge to Gaussian.

a considerable proportion of the time, ranging from 44% to 85%. As expected, the proportion decreases as the degrees of freedom increases and the distribution converges to Gaussian.

Where the dip test is robust to outliers, Silverman's test and the Hopkins statistic allow for small clusters. This finding reflects the inherent ambiguity of clustering; for some applications, small clusters are acceptable, while for others, robustness to outliers is desired. In fact, clustering algorithms display the same phenomenon, where some tend to identify small clusters, while others effectively view such data as outliers [8].

3.3 Power: Clusterable data

When clusters were well-separated, all methods had exactly or nearly 100% power, even in the presence of noise. When two fifty-dimensional clusters were close to each other as in row (20), all methods have nearly perfect power except for Classic Dip with around 70% power. For partially overlapping 50D clusters, e.g. row (21), the power of the Hopkins test drops to 32% and both classic methods drop below 5%. *This indicates that classical methods perform poorly in high dimensions for overlapping clusters.* By contrast, utilizing either PCA or pairwise distances, both Dip and Silverman tests maintain near perfect power to detect the presence of these close or overlapping high dimensional clusters. We also examine two-dimensional data generated from independent T-distributions with 5, 10, and 15 degrees of freedom. All methods have nearly 100% power to detect the t-distributed clusters.

Similarly, most methods had high power (above 80%) to detect data with three or four clusters, except that power for PCA dip and the Hopkins statistic dropped when the separation between clusters decreased. Specifically, classic and distance-based methods, as well as PCA Silverman, considered all and PCA-based methods considered nearly all (95+%) datasets as clusterable. Results are shown in rows 14-19 of Table 1.

3.4 Data with chaining structure

Finally, we examine data with chaining structure, including a single line, two parallel lines, one, two, three, and five concentric circles, and both a line and a circle. For data arranged in one line, classic dip, PCA methods and distance methods did not conclude that the data had multiple clusters. By contrast, Hopkins classified the line as clusterable nearly 40% of the time, and classic Silverman concluded the data had structure over 10% of the time. Surprisingly, all methods except dist-dip considered a single circle as clusterable. Distance Silverman concluded that the single circle had cluster structure about 30% of the time, while PCA, classical methods, and Hopkins concluded the same over 85% of the time. Principal curve methods nearly always failed to converge for data comprising a single line. *Thus, dip-dist may be the only valid method for chaining data.*

While both classic and PCA methods fail to detect the inherent structure of multiple groups of chaining data, distance-based methods and Hopkins continue to detect the clusters. Multiple parallel lines, depicted in row 18, are considered clusterable by distance based methods and reasonably well (87% power) by Hopkins. PCA based methods have less than 12% power, failing to detect the separate lines most of the time. Similarly, distance based methods both have 100% power and Hopkins has nearly 90% power to detect distinct circles, while PCA and classic methods have reduced power for 2 or 3 circles. All methods had high power (≥89%) to detect cluster structure in data consisting of one circle and one line except PCA dip, which only concluded the data was clusterable 20% of the time. *In sum, dip-dist was the most effective method for chaining data. In fact, it was the only method that didn't excessively conclude that data generated to lack groups was clusterable, and it had high power to detect clustered chaining data.*

3.5 Summary

In sum, our simulations indicate that both spatial randomness tests and multimodality tests on one-dimensional reductions are effective and accurate methods of classifying datasets by their level of clusterability. Both clusterable and unclusterable datasets were identified as such in most simulations. Distance-based methods perform well in nearly all scenarios. PCA methods adequately detect structure in simulated data with two or three clusters. In low dimensions, PCA power is not as high as for distance-based methods. Outliers are treated as clusters by all variations of Silverman and the Hopkins statistic. The Hopkins statistic loses signal when clusters touch or overlap. In high dimensions and for chaining data, classical methods are inappropriate. Principal curve methods have excessive false positives and fail to converge for linear data.

4 RESULTS ON REAL DATASETS

In this section, we apply our methods of clusterability evaluation to real datasets from the *R datasets* package.[18] The datasets we present were selected to ensure sufficient sample size and varied dimension. For the sake of completeness, we include all tests, but the reader should recall that some tests may be inappropriate under various conditions. References were examined for evidence of previously known cluster structure. Overall, results of the clusterability tests were consistent with expectations based on the simulations.

Two famous datasets that were known *a priori* to have cluster structure were considered clusterable under all methods. First, the *iris* dataset [26] is known to have three clusters corresponding to three species of iris flowers. Second, the *faithful* dataset [13, 30], which captures eruption duration and waiting time for the Old Faithful geyser, has previously been shown to have two groups [50]. All of the tests conclude that both datasets are clusterable, agreeing with previous knowledge.

Paralleling our simulations, we find that the Hopkins statistics or the Silverman tests may be preferred when small clusters are of interest, while the Dip test may be desired when the application calls for robustness to outlier. The one-dimensional *rivers* dataset [46], which contains the lengths, in miles, of 141 major North American rivers, exhibits inherent cluster structure *if we allow small clusters.* Hopkins method and all methods that use Silverman indicate that the data is clusterable ($p < 0.05$), while all dip-based methods fail to reject

[18] Due to the use of sampling in Hopkins' method, we run the method 100 times for each dataset and report the proportion of p-values less than 0.05.

	Dataset	Dip Dist	Silv. Dist	Hop.	Dip	Silv.	PCA Dip	PCA Silv.	Pcurv Dip	Pcurv Silv.
1.	1 cluster 2D	0.000	0.042	0.057	0.001	0.055	0.001	0.053	0.179	0.346
2.	1 cluster 3D	0.000	0.042	0.012	0.000	0.068	0.002	0.062	0.213	0.444
3.	1 cluster 10D	0.000	0.035	0.000	0.000	0.055	0.003	0.057	0.235	0.585
4.	1 cluster 50D	0.000	0.033	0.000	0.000	0.052	0.002	0.064	0.220	0.764
5.	1 cluster 2D with outlier	0.000	0.987	0.858	0.000	0.890	0.005	0.998	0.136	0.984
6.	1 large cluster 2D with outlier	0.000	0.975	1.000	0.000	0.911	0.001	0.989	0.331	0.980
7.	1 cluster 2D with 3 outliers	0.101	0.976	0.815	0.000	0.954	0.003	0.942	0.016	0.935
8.	1 T-dist cluster with df=5	0.007	0.573	0.852	0.000	0.440	0.000	0.463	0.070	0.490
9.	1 T-dist cluster with df=10	0.000	0.214	0.657	0.000	0.240	0.000	0.282	0.095	0.348
10.	1 T-dist cluster with df=15	0.000	0.117	0.579	0.002	0.209	0.000	0.220	0.098	0.344
11.	2 separated clusters 2D	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
12.	3 close clusters 2D	0.996	1.000	0.789	1.000	1.000	0.801	0.974	0.757	0.826
13.	3 noisy clusters 2D	1.000	0.996	0.989	0.999	0.995	1.000	0.999	1.000	0.983
14.	3 clusters, varied diameters 2D	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
15.	3 clusters, varied density 2D	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
16.	3 separated clusters 2D	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
17.	3 separated clusters 3D	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
18.	2 separated clusters 10D	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
19.	4 separated clusters 10D	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.884
20.	2 close clusters 50D	1.000	1.000	1.000	0.691	0.999	1.000	1.000	1.000	1.000
21.	2 partially overlapping 50D	1.000	1.000	0.445	0.000	0.041	1.000	1.000	0.995	0.997
22.	2 T-dist cluster with df=5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
23.	2 T-dist cluster with df=10	1.000	1.000	0.999	1.000	0.998	1.000	1.000	1.000	0.999
24.	2 T-dist cluster with df=15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
25.	SingleCircle	0.010	0.309	0.837	0.951	0.945	0.909	0.988	0.626	0.775
26.	2 concentric circles	1.000	1.000	0.873	0.533	0.751	0.322	0.472	0.619	0.802
27.	3 concentric circles	1.000	1.000	0.894	0.167	0.486	0.079	0.193	0.607	0.774
28.	5 concentric circles	1.000	1.000	1.000	0.364	0.394	0.159	0.364	0.726	0.895
29.	SingleLine	0.004	0.049	0.378	0.000	0.112	0.000	0.055	0.642	N/A
30.	2 parallel lines	1.000	0.889	1.000	1.000	0.996	0.000	0.055	0.997	0.989
31.	Line+Circle	0.998	0.999	1.000	1.000	0.958	0.209	0.894	0.876	0.982

Table 1: Proportion of datasets classified as clusterable over 1000 runs of each type, for a total of 33000 simulations. Scores denote the proportion of the time that the test concluded that the data was clusterable.

Dataset	Dip Dist	Silv. Dist	Hop.	Classic Dip	Classic Silv.	PCA Dip	PCA Silv.	Pcurve Dip	Pcurve Silv.
faithful	< 0.0001	< 0.0001	1.00	< 0.0001	< 0.0001	0.0017	< 0.0001	< 0.0001	< 0.0001
iris	< 0.0001	< 0.0001	1.00	0.0014	0.0010	< 0.0001	< 0.0001	0.0164	0.0022
rivers	0.2772	< 0.0001	0.92	0.9922	0.0192	0.9922	0.0334	0.9922	0.0291
swiss	< 0.0001	< 0.0001	0.41	0.1386	< 0.0001	0.0001	< 0.0001	< 0.0001	0.0010
attitude	0.9040	0.9598	0.00	0.9113	0.9150	0.6846	0.1534	0.1823	0.2174
cars	0.6604	0.9931	0.19	0.8613	0.3396	0.8320	0.4213	0.7680	0.5866
trees	0.3460	0.2900	0.18	0.0001	< 0.0001	0.8414	0.3675	0.6717	0.2282
USJudgeRatings	0.9938	0.7313	0.69	0.0014	0.0187	0.8550	0.1412	0.4501	< 0.0001
USArrests	0.9394	0.1887	0.01	0.6261	0.0171	0.5545	0.1286	0.0045	< 0.0001

Table 2: Clusterability tests appiled to real data. This table presents the p-values for the each clusterability test on real datasets from the R Datasets package. Recall that $p < 0.05$ signals clusterable data and $p \geq 0.05$ signals that data is unclusterable. The Hopkins value presented is the proportion of the time out of 100 runs that the Hopkins statistic was below the appropriate beta quantile. For the Hopkins results, high values indicate clusterability.

the null hypothesis of lack of structure. Similarly, *swiss* [48], consisting of 6 measures of socio-economic status and fertility for 47 French-speaking nineteenth-century Swiss provinces, illuminated logically pre-existing structure. While Classic Dip

considers the data as unclusterable, and Hopkins considers the data as clusterable 40% of the time, all other tests detect clusters. Results support literature that economic indicators between and within countries may fall into clusters, including a richer cluster much smaller than the other(s) [27, 35].

The remaining datasets lacked previously known structure. Indeed, most tests of clusterability that weren't known or shown to be questionable in simulations provided little or no evidence of clusters. Methods based on distances or PCA concluded that *cars* [25], *attitude* [18], *USArrests* [46], *trees* [51], and *USJudgeRatings* [38], were unclusterable. Hopkins' method agreed for *attitude* and *USArrests*.

Most of the methods that concluded that these remaining datasets without known structure were clusterable were previously shown in the present paper to be questionable. While classic Silverman and principal curve methods declared the *USArrests* [46] and *USJudgeRatings* [38] datasets clusterable, Classic methods may be unpredictable in multiple dimensions (see section 2.2.1), and the principal curve methods had high false positive rates in our simulations. Interestingly, Hopkins considered USJudgeRatings clusterable nearly 70% of the time, trees clusterable 18% of the time, and cars 19% of the time.

The methods with the most logical results include distance dip, distance Silverman, PCA dip, and PCA Silverman. Although classic Dip and Silverman methods appear to produce reasonable conclusions in some famous datasets such as *iris*, they have produced conterintuitive results when classifying other real data, such as *USJudgeRatings* and *USArrests*. This finding, which supports previously known theory about the unpredictability of these tests in multiple dimensions, reflects the importance and value of utilizing dimensionality reduction to evaluate clusterability.

Finally, principal curve methods were highly problematic on real datasets. Even on the famous, well-defined dataset *faithful*, the principal curve failed to converge after 1000 iterations.

5 CONCLUSIONS

The application of clustering algorithms presupposes the existence of cluster structure. Clustering techniques tend to produce some partition for any given dataset, which can lead to invalid conclusions when the data is unclusterable. Consequently, we advocate for the integration of clusterability into cluster analysis, allowing users to determine whether clustering is appropriate for the given data before proceeding with further processing.

Though many approaches to clusterability evaluation have been previously proposed, they vary radically and often result in different conclusions. Here, we perform an extensive analysis of a variety of clusterability methods, identifying which are most effective as well as when certain measures are better suited than others based on the needs of the application at hand.

While other notions of clusterability may also warrant investigation, our paper is the most comprehensive study to date. We compare several approaches, which apply either spatial randomness tests to the original data or multimodality tests to one-dimensional reductions of the data. Extensive simulations allow us to identify effective approaches, as well as differentiate amongst them. Notably, experiments on real data sets parallel the conclusions of our simulations.

Overall, dip-dist and Silv.-dist, as well as PCA dip and PCA Silverman were the most successful methods in our analysis. Below, we discuss several criteria that may one consider when selecting amongst clusterability measures, as well as our findings with respect to each criteria.

Below we summarize several qualitative criteria that can be used to select a suitable clusterability measure for a given application. In addition to these qualitative considerations, quantitative comparison based on the efficiency of these methods could also be integrated, particularly when data is large. See Section 2.6 for a comparison of the methods considered in this analysis based on their computational complexity.

- **False positives:** By proclaiming to discover cluster structure when none is present, methods that exhibit excessive false positives (Type I error) are considered statistically invalid. Both methods that reduce data using the principle curve consistently exhibited inflated Type I error rates, and as such we do not recommend these approaches.
- **Outliers/small clusters:** We discover that clusterability measures vary drastically in their treatment of sparse distant points. While Hopkins and Silverman-based methods treat the points as small clusters, Dip-based methods exhibit outlier robustness.
- **Chaining data:** Dip-dist was the only method that consistently performed well on chaining-type data (concentric circles and parallel lines), able to identify both clusterable and unclusterable structures of these types.
- **High dimensionality:** We tested datasets on up to 50 dimensions. In our experiments, PCA dip, PCA Silverman, Dip-dist, and Silverman Dist did well, suggesting that these methods may be better suited to high dimensional data than the other techniques considered in this analysis.

While our results suggest that some of the methods considered here work well for data of reasonably high dimension, for very high dimensional data (particularly when the dimensions is much greater than the number of elements), additional investigation is desirable. It is possible that simply modifying the data reduction method, such as by using Sparse PCA [59], may be sufficient. This avenue of investigation is left for future work.

We look forward to the widespread application of clusterability tests as part of the clustering process. We close with the following quote to remind of the importance of testing for clusterability before proceeding with further – potentially unnecessary – cluster analysis tasks.

> "There is nothing so useless as doing efficiently that which should not be done at all"
> -Peter F. Drucker

REFERENCES

[1] *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press, 2009.

[2] M. Ackerman and S. Ben-David. Measures of clustering quality: A working set of axioms for clustering. In *Advances in neural information processing systems*, pages 121–128, 2008.

[3] M. Ackerman and S. Ben-David. Clusterability: A theoretical study. *Proceedings of AISTATS-09, JMLR: W&CP*, 5(1-8):53, 2009.

[4] M. Ackerman and S. Ben-David. Discerning linkage-based algorithms among hierarchical clustering methods. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1140, 2011.

[5] M. Ackerman, S. Ben-David, S. Branzei, and D. Loker. Weighted clustering. In *AAAI*, pages 858–863, 2012.

[6] M. Ackerman, S. Ben-David, and D. Loker. Characterization of linkage-based clustering. In *COLT*, pages 270–281, 2010.

[7] M. Ackerman, S. Ben-David, and D. Loker. Towards property-based classification of clustering paradigms. In *Advances in Neural Information Processing Systems*, pages 10–18, 2010.

[8] M. Ackerman, S. Ben-David, D. Loker, and S. Sabato. Clustering oligarchies. *Proceedings of AISTATS-09, JMLR: W&CP*, 31(66fi?!74), 2013.

[9] M. Ackerman and S. Dasgupta. Incremental clustering: The case for extra clusters. In *Advances in Neural Information Processing Systems*, pages 307–315, 2014.

[10] Murat O. Ahmed and Guenther Walther. Investigating the multimodality of multivariate data with principal curves. *Computational Statistics and Data Analysis*, 56(12):4462 – 4469, 2012.

[11] Pranjal Awasthi, Avrim Blum, and Or Sheffet. Stability yields a ptas for k-median and k-means clustering. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 309–318. IEEE, 2010.

[12] Pranjal Awasthi, Avrim Blum, and Or Sheffet. Center-based clustering under perturbation stability. *Information Processing Letters*, 112(1):49–54, 2012.

[13] Adelchi Azzalini and Adrian W Bowman. A look at some data on the old faithful geyser. *Applied Statistics*, pages 357–365, 1990.

[14] M.-F. Balcan, A. Blum, and A. Gupta. Approximate clustering without the approximation. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1068–1077. Society for Industrial and Applied Mathematics, 2009.

[15] M.-F. Balcan, A. Blum, and S. Vempala. A discriminative framework for clustering via similarity functions. In *Proceedings of the 40th annual ACM symposium on Theory of Computing*, pages 671–680. ACM, 2008.

[16] A. Banerjee and R. N. Dave. Validating clusters using the hopkins statistic. In *2004 IEEE International Conference on Fuzzy Systems (IEEE Cat. No.04CH37542)*, volume 1, pages 149–153 vol.1, July 2004.

[17] Shai Ben-David. Computational feasibility of clustering under clusterability assumptions. *arXiv preprint arXiv:1501.00437*, 2015.

[18] S. Chatterjee and B Price. *Regression analysis by example*. John Wiley & Sons, 1991.

[19] M.-Y. Cheng and P. Hall. Calibrating the excess mass and dip tests of modality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):579–589, 1998.

[20] National Research Council. *Discriminant Analysis and Clustering*. The National Academies Press, Washington, DC, 1988.

[21] Chris Ding and Xiaofeng He. K-means clustering via principal component analysis. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 29–, New York, NY, USA, 2004. ACM.

[22] Richard C. Dubes and Guangzhou Zeng. A test for spatial homogeneity in cluster analysis. *Journal of Classification*, 4(1):33–56, 1987.

[23] T. Elgamal and M. Hefeeda. Analysis of PCA Algorithms in Distributed Environments. *ArXiv e-prints*, March 2015.

[24] S. Epter, M. Krishnamoorthy, and M. Zaki. Clusterability detection and initial seed selection in large datasets. In *The International Conference on Knowledge Discovery in Databases*, volume 7, 1999.

[25] Mordecai Ezekiel. methods of correlation analysis. 427 pp., illus. *New York and London*, 1930.

[26] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[27] Garance Genicot and Debraj Ray. Aspirations and inequality. Working Paper 19976, National Bureau of Economic Research, March 2014.

[28] Sarah C. Goslee. Correlation analysis of dissimilarity matrices. *Plant Ecology*, 206(2):279–286, 2010.

[29] Peter Hall and Matthew York. On the calibration of Silverman's test for multimodality. *Statistica Sinica*, 11:515–536, 2001.

[30] Wolfgang Härdle. *Smoothing Techniques: With Implementation in S*. Springer Science & Business Media, 1991.

[31] J. A. Hartigan and P. M. Hartigan. The dip test of unimodality. *Ann. Statist.*, 13(1):70–84, 03 1985.

[32] J.A. Hartigan. Statistical theory in clustering. *Journal of Classification*, 2(1):63–76, 1985.

[33] Trevor Hastie and Werner Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989.

[34] E. S. Helgeson and E. Bair. Non-Parametric Cluster Significance Testing with Reference to a Unimodal Null Distribution. *ArXiv e-prints*, October 2016.

[35] Daniel J. Henderson, Christopher F. Parmeter, and R. Robert Russell. Modes, weighted modes, and calibrated modes: evidence of clustering using modality tests. *Journal of Applied Econometrics*, 23(5):607–638, 2008.

[36] Brian Hopkins and J. G. Skellam. A new method for determining the type of distribution of plant individuals. *Annals of Botany*, 18(70):213–227, 1954.

[37] Peter J. Huber. Projection pursuit. *The Annals of Statistics*, 13(2):435–475, 1985.

[38] John Hartigan. New Haven Register, 1977.

[39] I.T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, 2002.

[40] Argyris Kalogeratos and Aristidis Likas. Dip-means: an incremental clustering method for estimating the number of clusters. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2393–2401. Curran Associates, Inc., 2012.

[41] J. Kleinberg. An impossibility theorem for clustering. *Proceedings of International Conferences on Advances in Neural Information Processing Systems*, pages 463–470, 2003.

[42] Andreas Krause and Volkmar Liebscher. Multimodal projection pursuit using the dip statistic. *Preprint-Reihe Mathematik*, 13, 2005.

[43] Richard G. Lawson and Peter C. Jurs. New index for clustering tendency and its application to chemical problems. *Journal of Chemical Information and Computer Sciences*, 30(1):36–41, 1990.

[44] Xu Ling, Edward J. Bedrick, Timothy Hanson, and Carla Restrepo. A comparison of statistical tools for identifying modality in body mass distributions. *Journal of Data Science*, 12:175 – 196, 2014.

[45] Yufeng Liu, David Neil Hayes, Andrew Nobel, and J. S Marron. Statistical significance of clustering for high-dimension, lowfi?!sample size data. *Journal of the American Statistical Association*, 103(483):1281–1293, 2008.

[46] Donald R McNeil. *Interactive data analysis: a practical primer*. John Wiley & Sons, 1977.

[47] G.W. Milligan. A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, 46(2):187–199, 1981.

[48] Frederick Mosteller and John Wilder Tukey. Data analysis and regression: a second course in statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods*, 1977.

[49] R. Ostrovsky, Y. Rabani, L.J. Schulman, and C. Swamy. The effectiveness of Lloyd-type methods for the k-means problem. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 165–176, 2006.

[50] Daniel Pena and Adolfo Alvarez. Recombining partitions via unimodality tests. DES - Working Papers. Statistics and Econometrics. WS ws130706, Universidad Carlos III de Madrid. Departamento de Estadstica, March 2013.

[51] Thomas A Ryan, Brian L Joiner, Barbara F Ryan, et al. *Minitab student handbook*. Duxbury Press, 1976.

[52] M. Shahbaba and S. Beheshti. Efficient unimodality test in clustering by signature testing. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8282–8286, May 2014.

[53] B. W. Silverman. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society. Series B (Methodological)*, 43(1):pp. 97–99, 1981.

[54] K Szczubialka, J Verd-Andrs, and D.L Massart. A new method of detecting clustering in the data. *Chemometrics and Intelligent Laboratory Systems*, 41(2):145 – 160, 1998.

[55] Robert Tibshirani. Principal curves revisited. *Statistics and Computing*, 2(4):183–190, 1992.

[56] Daniel R. Wells. A monotone unimodal distribution which is not central convex unimodal. *The Annals of Statistics*, 6(4):926–931, 1978.

[57] R. Xu and D. Wunsch. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678, 2005.

[58] Hongyuan Zha, Xiaofeng He, Chris Ding, Ming Gu, and Horst D. Simon. Spectral relaxation for k-means clustering. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 1057–1064. MIT Press, 2002.

[59] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.