ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Bachelor Thesis Business Analytics and Quantitative Marketing

A Critical Evaluation of Stability-based Measures for the Selection of the Number of Clusters

Name student: M.Y.M Chow

Student ID number: 449580

Supervisor: M. van de Velden

Second assessor: C. Cavicchia

Abstract

This research paper aims to critically evaluate the validity of clustering stability for the selection of the number of clusters. By means of simulated examples it is shown that the stability-based CV_a, CV_v (Wang, 2010) and CC (Monti, Tamayo, Mesirov, & Golub, 2003) measures consistently outperform the popular non-stability-based KL, Silhouette, GAP and JUMP measures. However, the objective of this paper is not to establish superiority for stability-based selection measures, but to provide a complete picture of their strengths and weaknesses. Theory highlights some potential pitfalls for the CV_a, CV_v and CC measures and for clustering stability in general, which are confirmed in simulated examples. Most striking is the existence of stable but meaningless clusters, indicating that a stable solution cannot be guaranteed to be valid. It is concluded that, while there are compelling arguments for the use of stability-based measures for the selection of the number of clusters, they alone are not sufficient to validate the clustering. Therefore, they should be used as a part of a wider strategy for the identification of the number of clusters. In case one decides to make use of stability-based selection measures, the conditions at hand determine which of the measures, CV_a, CV_v or CC, should be employed.

July 4, 2020

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Table of Contents

1 Introduction 1

2 The validity of clustering stability 3

3 Methodology 5

3.1 Cross validation measures 5

3.2 Consensus clustering 7

3.3 Cross validation measures versus consensus clustering 8

3.4 Non-stability-based selection measures 10

3.5 Clustering algorithms 11

4 Experimental design 12

4.1 The validity of clustering stability 12

4.1.1 Four simulated examples (Wang, 2010) 12

4.1.2 Unequal cluster sizes and correlated features 13

4.1.3 More clusters 14

4.1.4 A stable but meaningless clustering 14

4.2 Cross validation measures versus consensus clustering 15

4.2.1 Stability scores for components 15

4.2.2 Visualisation by heat maps 15

4.2.3 Small sample size 15

4.3 Data pre-processing 16

5 Results 16

5.1 The validity of clustering stability 16

5.1.1 Four simulated examples (Wang, 2010) 16

5.1.2 Unequal cluster sizes and correlated features 19

5.1.3 More clusters 20

5.1.4 A stable but meaningless clustering 21

5.2 Cross validation measures versus consensus clustering 23

5.2.1 Stability scores for components 23
5.2.2 Visualisation by heat maps 24
5.2.3 Small sample size 25

6 Practical implications 26

7 Conclusion and discussion 28
7.1 Conclusion 28
7.2 Discussion 29

References 31

A Data sets 33
A.1 The validity of clustering stability 33
A.2 Cross validation measures versus consensus clustering 35

B Programming codes 35

1 Introduction

Cluster analysis describes the task of grouping similar observations together in clusters. The characteristics of these clusters may then yield valuable insight in the underlying distribution of data. The formation of the clusters depends on the employed clustering algorithm. A wide variety of clustering algorithms exist, utilizing various objective functions. Whether a given algorithm is appropriate depends on the objective of the study and the structure of the data. Moreover, the selection of the number of clusters remains a fundamental issue in the field of cluster analysis. This is mainly attributed to the absence of an objective measure for the comparison of different clusterings of the same data set (Wang, 2010). While some heuristics exists, these methods remain subjective. One clustering might be better according to one metric, while the reverse is true according to another metric. As there is no definite answer for which metric is uniformly the best, it is important that measures for the selection of the number of clusters are developed and evaluated against each other.

Wang (2010) proposed two selection methods based on the notion of clustering stability, which builds on the idea that a valid clustering should be robust against randomness in sampling. Such approaches are attractive as they can be applied to any type of clustering algorithm. While the idea is intuitively clear, theoretical and practical justification of stability-based selection methods appear to be lacking. The question arises whether clustering stability is a valid tool for the selection of the number of clusters:

R1: To what extend is clustering stability a valid measure for the selection of the number of clusters?

This research question is answered from a theoretical and empirical perspective. Firstly, the theoretical background of clustering stability is examined to identify its main advantages and pitfalls. To evaluate validity from an empirical perspective, the performance of several stability-based methods is evaluated. Specifically, the cross validation measures CV_a and CV_v (Wang, 2010) and consensus clustering (CC) (Monti et al., 2003) are considered.

To evaluate the competitiveness of the stability-based selection measures, I compare their selection performance to that of several popular non-stability-based selection measures in four simulated examples. These examples are taken from Wang (2010), who selected these examples to illustrate the competitiveness of his cross validation measures. Therefore, it is expected that the stability-based selection measures perform well in these examples, providing a biased view. The aim of this paper is not to establish superiority for stability-based selection measures, but to provide a complete picture of their strengths and weaknesses. This allows for a more considered use of the CV_a, CV_v and CC measures for the selection of the number of clusters. Based on the theoretical pitfalls of clustering stability, new data sets are generated on which the selection performance of the stability-based measures is evaluated. In this way, the practical implications of their theoretical pitfalls can be assessed.

This research question is scientifically relevant, as it is yet unclear whether stability-based measures should be used for the selection of the number of clusters. This paper aims to fill this gap in the literature and provides a thorough picture of clustering stability as a measure for the selection of the number of clusters. Existing research often failed to provide an unbiased picture, by stressing either the advantages or the weaknesses of a certain selection measure. It is important that researchers include both sides of the story in order to provide an unbiased picture of the measure in question.

Even though the CV_a, CV_v and CC measures are all based on the notion of clustering stability, they each take a different approach in computing and evaluating the stability scores. One measure might be more suitable in a certain data set, while another measure is more suitable in another. For a more considered use of the CV_a, CV_v and CC measures, it is important that the circumstances in which one of the measures is preferred are identified. Therefore, the second research question is posed:

R2: How do the CV_a, CV_v and CC measures compare to each other?

Again, this research question is answered from a theoretical and empirical perspective. The methodology behind the measures is compared to highlight their differences. Based on these theoretical differences, new data sets are generated. The performance of the stability-based measures are then compared on these data sets to assess the practical implications of their theoretical differences. This research question is scientifically relevant, as it is not yet clear how clustering stability should be implemented in practice. To yield more insight in this problem, stability-based measures should be critically evaluated against each other.

On one hand, the results support the validity of clustering stability for the selection of the number of clusters, as the stability-based measures consistently outperform the non-stability-based measures. On the other hand, several circumstances were identified in which stability-based selection measures produced misleading results. Most concerning is the existence of stable but meaningless clusterings, which implies that stable solutions cannot be guaranteed to be valid. This provides a counter example to the validity of clustering stability for the selection of the number of clusters. Stability-based measures alone are thus not sufficient to validate a clustering and they should be used as a part of a wider strategy for the identification of the number of clusters. In case one chooses to utilize stability-based selection measures, the conditions at hand determine which of the measures, CV_a, CV_v or CC, should be employed.

The remainder of this paper is structured as follows. The theoretical validity of clustering stability and related literature are explored in Section 2. The methods of this paper are then described in Section 3. This concerns the CV_a, CV_v and CC measures, the non-stability-based measures and the employed clustering algorithms. Section 4 describes how the data sets are simulated and is divided into two sub-sections, corresponding to the research questions. This is followed by the results in Section 5, which follows the same structure as Section 4. The practical implications of the results are then discussed in Section 6. Finally, a conclusion and discussion are presented in Section 7.

2 The validity of clustering stability

The general idea behind clustering stability is that a good clustering should be robust against randomness in sampling. When random samples from a population result in similar clusterings, it can be concluded with more confidence that these clusterings represent the actual structure of the data (Monti et al., 2003). Various implementations exist for the selection of the number of clusters based on an algorithm's stability. These approaches differ in the way they generate perturbed samples and how they compute and evaluate the stability scores. In general, stability-based methods select the number of clusters, k, by minimizing the estimated instability. Instability increases when the incorrect number of clusters is chosen. When k is set too high, the true clusters are randomly split. When k is set too low, the true clusters are randomly merged. This induces instability. Intuitively, the use of stability-based selection methods seems compelling. Furthermore, clustering stability is flexible and can be applied to any type of clustering algorithm. Amongst others, Ben-Hur, Elisseeff, and Guyon (2001), Lange, Roth, Braun, and Buhmann (2004) and Wang (2010) have found promising results for their proposed stability-based measures for the selection of the number of clusters in both simulated and real life examples.

The use of stability-based selection measures is not without challenges. While it is true that an unstable solution is "bad", it is not necessarily true that a stable solution is "good". A simple example illustrates that a stable solution could be meaningless: a clustering in which each cluster contains a single observation is perfectly stable, yet completely meaningless. Furthermore, a clustering may sometimes only be stable due to the inflexibility of the employed clustering algorithm (Hennig, 2007). This is in line with the findings of Handl, Knowles, and Kell (2005), who found that the tendency of the k-means algorithm to construct spherical clusters lead to a stable but meaningless clustering of a data set with elongated clusters. The existence of stable but meaningless clusterings implies that clustering stability alone is not enough to validate the clustering and thus cannot reliably be used to select the number of clusters. On the other hand, Von Luxburg (2010) found, for the k-means algorithm, that the values of k that lead to stable solutions have desirable properties, such that stability-based methods can be used to identify the number of clusters. These contrasting conclusions highlight the importance for further research in this field.

Research has highlighted several circumstances in which stability-based measures, previous to the CV$_a$, CV$_v$ and CC measures, produced misleading results. For example, previous stability-based selection measures were found to perform badly when cluster sizes were unequally distributed and when the features exhibited high correlations (Krieger & Green, 1999). Furthermore, when the true number of clusters was relatively large, previous stability-based measures were found to be biased towards fewer clusters (Breckenridge, 2000). As it is not uncommon for real data sets to exhibit these features, it is important that selection measures can deal with them. The question arises whether more recently proposed stability-based selection measures, such as CV$_a$, CV$_v$ and CC, can deal with such features in the data or whether stability-based measures are in general inappropriate in this context.

Ben-David, Von Luxburg, and Pál (2006) found that the stability of a given algorithm is not determined by clustering parameters. Instead, stability depends only on the algorithm's objective function and in case it has a unique global optimum the algorithm is asymptotically stable. The stability of the algorithm thus does not depend on the correctness of the selected k and a stable clustering could be constructed with the wrong number of clusters. This casts some serious doubt on the validity of clustering stability for the selection of the number of clusters. However, the mean of the limiting distribution of the re-scaled instabilities can be shown to be dependent on k (Shamir & Tishby, 2008). This implies that, in the limit, different values of k lead to different values of re-scaled instabilities. Hence, if the sample size is large enough, clustering stability could be used to identify the number of clusters.

From a theoretical point of view, there are compelling arguments for clustering stability as a measure for the selection of the number of clusters. Yet, many potential pitfalls have been identified that suggest that clustering stability alone is not sufficient to determine the number of clusters. While stability-based selection measures may perform well under certain conditions, a stable solution cannot be guaranteed to be valid. Clustering stability should therefore not be used in isolation, but as a component in a broader strategy for the selection of the number of clusters. In this way, the risk of accepting false outcomes is minimized.

3 Methodology

The following notation is used throughout the text. Data sets contain N observations, which have to be grouped into $k = \{2,3,...,K\}$ clusters. K is some pre-specified maximum and individual clusters are denoted by $h = \{1,2,...,k\}$.

3.1 Cross validation measures

CV$_a$ and CV$_v$ select the number of clusters by minimizing the algorithm's instability. The data set is partitioned into three equal-sized subsets: two training sets (Z_1 and Z_2) and a validation set (Z_3). Cluster analysis is performed on Z_1 and Z_2, resulting in two clusterings. Z_3 is then used to validate the performed analysis by measuring the agreement between the two clusterings. Let $\psi(i,Z)$ denote a classifier that is trained on $Z \in \{\mathrm{Z}_1,\ Z_2\}$

and that assigns observation i in Z_3 to cluster h. For each combination of observations i and j in Z_3, I compare $\psi(i,Z_1)$ with $\psi(j,Z_1)$ and $\psi(i,Z_2)$ with $\psi(j,Z_2)$. Such predictions of cluster membership are easily obtained for k-means clustering. For spectral clustering, however, cluster membership must be computed via the k-nearest neighbours algorithm. This algorithm assigns the observation to the cluster that is most common among its k nearest neighbours. In the following, the number of nearest neighbours is set to 10. Other values are possible, but this is not expected to have a significant influence on the selection performance (Wang, 2010). A stable clustering algorithm should then either assign i and j to the same cluster or to different clusters according to both classifiers, such that the clusterings are similar. When the clusterings are not in accordance, the estimated instability score is increased with one point. An instability score is computed for all possible values of k and the number of clusters is selected by minimizing the instability over k.

The clusterings and the estimated instability depend on how the data are split. To reduce the variability in estimation, 100 random splits are considered. There are two approaches for summarizing the results over the 100 splits: cross validation with voting (CV_v) and cross validation with averaging (CV_a). Cross validation with voting considers each split of the data individually. For each split, CV_v casts a vote on the k that minimizes instability. The k that receives the most votes is selected. Cross validation with averaging, on the other hand, computes the average instability over all splits for each k. k can then be selected by simply minimizing the average instability, resulting in CV_a1. For a finer evaluation of the differences in average instability, one could compute the standard deviations. k is then selected as the largest k for which the average instability minus twice the standard deviation is smaller than all average instabilities of $k' < k$, resulting in CV_a2. A consequence is that CV_a2 selects larger k's than CV_a1 when the magnitudes of the average instabilities are comparable.

Wang (2010) showed that both CV_a and CV_v reach asymptotic selection consistency, such that the probability of selecting the optimal number of clusters converges to one, when the data are properly split. This result, however, depends on how Wang (2010) defined "the optimal number of clusters". He defined it as the number of clusters that minimizes the algorithm's instability. As this is precisely the objective of the cross validation measures, the result merely implies that the measures converge to the global optimum. This result favours the use of CV$_a$ and CV$_v$ over other selection measures that could terminate at local optima. However, it does not say anything about their ability to select the true number of clusters as a stable clustering may not necessarily be meaningful.

3.2 Consensus clustering

Consensus clustering (CC) assesses the agreement, or consensus, between clusterings of perturbed samples of the data set. This paper considers 100 perturbed samples, corresponding to the 100 splits for the cross validation measures. The perturbed samples are created by randomly taking 70% of the observations. While sub-sampling leads to smaller data sets, no significant adverse effect has been found on the selection performance of consensus clustering (Monti et al., 2003). The consensus between the clusterings is assessed by means of the consensus matrix M. M is a symmetric N X N matrix with elements $M_{(i,j)}$ that represent the proportion of clusterings in which observations i and j are assigned to the same cluster. $M_{(i,j)}$ is referred to as the consensus index of pair (i, j) and takes values between zero and one. Perfect consensus then corresponds to all elements of M being either zero or one. The rows and columns of M are arranged such that observations that belong to the same cluster are adjacent. In case of perfect consensus, M is a block-diagonal matrix in which each block of ones represents a cluster and is surrounded by zeros.

The consensus matrix M can be used to visualize the clusterings. So-called heat maps are created by applying a color gradient to the zero to one range of real numbers, in which zero corresponds to white and one corresponds to dark red. In practice cluster membership is often unknown in advance, such that the order of the rows and columns in M is unknown. Hierarchical clustering with M as the similarity matrix can be used to find the ordering of the observations. In particular, this paper employs the optimal leaf-ordering algorithm that was proposed by Bar-Joseph et al. (2003). The resulting dendogram ensures that observations with the highest consensus indices are adjacent such that the block-diagonal nature of the heat map is maximized. In addition, the consensus matrix can be used to compute stability scores for components, groups of observations that may or may not correspond to the clusters. These scores are computed as the average consensus indices of all pairs of observations that belong to the component. This is particularly useful in the context of overlapping clusters.

To select the number of clusters, a consensus matrix is constructed for each k, resulting in $M(k)$. Recall that perfect consensus corresponds to all elements of M being zero or one. The best k then corresponds to the "cleanest" matrix $M(k)$. This is formally evaluated by the consensus distribution, which asses how the values $M(k)_{(i,j)}$ are distributed on the zero to one range. For this reason, the empirical cumulative distribution function $\mathrm{F}_e()$ is constructed. $\mathrm{F}_e(t)$ is computed as the proportion of consensus indices that are smaller or equal to t, for $0 \leq t \leq 1$. In case of perfect consensus $\mathrm{F}_e()$ resembles a step function. A gradually climbing $\mathrm{F}_e()$ reflects the lack of consensus, as it demonstrates the many fractional elements in M. To quantify the difference in the shapes of $\mathrm{F}_e()$ for the different k, the area below the curves, $A(k)$, is considered. Before reaching the true k, increasing k leads to an increase in $A(k)$ as M will contain more zero elements. This is because observations, that were previously wrongfully put in the same cluster, are now separated into their true clusters. However, after reaching the true k, increasing k will not lead to any further increase in $A(k)$ as this causes the true clusters to be randomly split. This leads to more fractional elements in M. This behaviour is summarized by $\Delta(k)$ which denotes the relative increase in $A(k)$ with respect to the largest $A(k')$ for $k' < k$. The number of clusters is then chosen as the largest k for which $\Delta(k)$ is significantly larger than zero. This may provide a range of possible k's. However, with the inspection of the heat maps it is often possible to select the single best k.

3.3 Cross validation measures versus consensus clustering

This paper considers two approaches for the selection of the number of clusters based on clustering stability: the cross validation measures and consensus clustering. This choice is motivated by their different approaches for the computation and assessment of the stability scores. The cross validation measures estimate the algorithm's instability and minimize this. Consensus clustering evaluates the instability scores of all pairwise observations simultaneously by means of the consensus distribution. Conventional stability-based measures often take a similar approach as the cross validation measures, such that the results can be generalised to a wide range of stability-based selection measures.

The cross validation measures compute a single instability score for the entire clustering and can therefore not be used to evaluate the stability of components of the data. Consensus clustering, on the other hand, can be used to identify unstable components. This is useful in the context of overlapping clusters, as observations on the border are often found to be less stable. Lord, Willems, Lapointe, and Makarenkov (2017) found that the removal of such unstable observations lead to a better recovery of the true number of clusters, advocating the use of consensus clustering.

Another advantage that consensus clustering has over the cross validation measures is that it can be used to produce a visualisation of the clustering, thereby yielding more insights. However, the inspection of the heat maps is also often needed to select k, as the decision rule of consensus clustering can be vague. The decision rules for CV_a and CV_v, on the other hand, are stricter. They select k by minimizing the estimated instability, but do not take into account the sizes, shapes and boundaries of the clusters. Even though not formally, consensus clustering does account for these features in the heat maps. A distinction can also be made between the decision rules of CV_a and CV_v. The first makes use of averaging, while the latter makes use of voting to summarize the results over the splits. A deviant value for the estimated instability in a single split is more likely to affect the result of CV_a than that of CV_v. The latter is thus more robust against outlying clusterings.

The estimation of instability by means of cross validation presents another issue. Firstly, the splitting of the data causes the data set for the cluster analysis to be three times as small as the original data set. As the selected k is dependent on sample size, the selection of k based on the estimated instability of the sub-samples is questionable. An additional problem is introduced when the sample size becomes too small. In this case, taking small subsets may come at the cost of the underlying data structure. Consensus clustering avoids both problems as it takes sub-samples containing 70% of the observations of the original data set, which is considerably more than a third.

Lastly, while asymptotic selection consistency is established for the cross validation measures, such a result lacks for consensus clustering. The latter may thus terminate at a local optimum.

3.4 Non-stability-based selection measures

To evaluate the competitiveness of the stability-based measures, their selection performance is compared to that of several popular non-stability-based measures. The following presents a an overview of the non-stability-based measures that are employed in this paper.

The Silhouette measure (Rousseeuw, 1987) evaluates how similar an observation is to the cluster that it belongs to compared to other clusters. The higher the Silhouette score, the more similar the observations are to the cluster they belong to. The number of clusters is then selected by maximizing the Silhouette score. This measure is often used as a benchmark for newly proposed selection measures. The KL measure (Krzanowski & Lai, 1988) is based on the within-group sum-of-squares objective function. The authors find promising results for their criterion but note that its performance is limited to distance-based data sets, for which the sum-of-squares objective function is appropriate. Furthermore, the criterion might terminate at a local optimum. The GAP measure (Tibshirani, Walther, & Hastie, 2001) is based on the deviation of the within-cluster sum of squares with the expected value under the null hypothesis of no obvious clustering. The number of clusters is selected as the smallest k for which the $\text{GAP}(k) \geq \text{GAP}(k+1) - s_{k+1}$, where s_{k+1} denotes the sample standard deviation. When this condition is not satisfied, the algorithm selects $k = 1$. The authors find that the GAP measure performs well when the uniform reference in the principal component orientation is used. This is the approach that is employed in this paper. Lastly, the JUMP measure (Sugar & James, 2003) minimizes the average distance between an observation and the closest cluster center, also known as the distortion. The JUMP score captures the change in distortion when k is reduced by one and selects the number of clusters by maximizing the JUMP score. The authors find promising results for their measure but note that the distortion curve may be monotone. In this case, optimization is not sensible as it selects a corner solution. Both the GAP and JUMP measures employ bootstrap estimation techniques, which means that they generate perturbed samples by randomly sampling observations with replacement from the original data set.

Each of the authors have illustrated great selection performance for their proposed measures, which is why these measures were chosen as benchmarks for the stability-based measures. To allow for a fair comparison between the selection measures, the number of bootstrap samples for the GAP and JUMP measures is set to 100. This corresponds to the 100 splits for the cross validation measures and the 100 sub-samples for consensus clustering.

3.5 Clustering algorithms

Two clustering algorithms are employed in this paper: a distance-based algorithm and a non-distance-based algorithm. K-means clustering is distance-based and starts with k random observations as starting points. The algorithm then assigns the remaining observations to the cluster whose center is the closest. The resulting clusters have the property that the sum of squares from the observations to their assigned cluster centers is minimized. Due to its simplicity and efficiency, k-means has become a popular clustering method and many variations have been developed. This paper utilizes the algorithm proposed by Hartigan and Wong (1979), which is superior to other k-means algorithms. However, as it is a greedy heuristic approach, it may terminate at a local optimum. Therefore, 20 random restarts for the starting points are considered.

Spectral clustering is not based on distance, but based on the connectivity of the observations. The observations are connected by edges to form a similarity graph, which models the local neighbourhood relationships between the observations. The similarity graph is then used to compute the graph Laplacian, a lower-dimensional representation of the similarity graph. Finally, a standard clustering algorithm is applied to the eigenvectors of the graph Laplacian. An advantage of spectral clustering is that it makes no assumptions on the statistics of the clusters. It can therefore successfully find clusters with non-convex shapes, in contrast to k-means. Various spectral algorithms exist, but the majority of these algorithms have no theoretical backbone for their validity. This paper utilizes the algorithm proposed by Ng, Jordan, and Weiss (2002), which is supported by both theory and practice.

Agglomerative hierarchical clustering is used to determine the optimal order of the rows and columns of the consensus matrix M. The algorithm starts with each observation as a cluster. In each iteration, the most similar clusters are merged until all observations belong to the same cluster. In this paper, M is used as the similarity matrix. Furthermore, average linkage is used, such that the similarity between two clusters is computed as the average similarity between the pairs of observations. The algorithm returns a dendogram with ordered observations. However, many possible orderings exist, as the algorithm is sensitive to noise. The optimal leaf ordering algorithm (Bar-Joseph et al., 2003) is employed to find the single best ordering of the observations. It ensures that the most similar observations are adjacent, such that the most diagonal heat map can be attained.

4 Experimental design

4.1 The validity of clustering stability

4.1.1 Four simulated examples (Wang, 2010)

To evaluate the competitiveness of the stability-based measures, their selection performance is compared to that of several non-stability-based measures in four simulated examples. The true number of clusters is known such that the performance of the measures can be compared objectively. To illustrate the flexibility of the selection measures, both distance-based and non-distance based data sets are considered. This distinction regards the structure of the data. In distance-based data sets clusters are based on the proximity of the observations. In non-distance-based data sets clusters are based on the connectivity of the observations. The distinction is clear from Figures 4 and 5 of Appendix A.1.

Examples 1 and 2 are distance-based. Example 1 consists of two elongated clusters of 100 observations in a three-dimensional space. It is generated by setting $x_1 = x_2 = x_3 = \mathrm{t}$, where t takes 100 equally spaced values between -0.5 and 0.5. Gaussian noise with mean 0 and standard deviation 0.1 is then added to all three features. This is the first cluster. The second cluster is created in the same way but adds a value of 10 to each feature at the end. Example 2 consists of four non-Gaussian clusters of 100 observations in a ten-dimensional space. The first two dimensions of the clusters are sampled from four bi-variate exponential distributions with location parameters (4,4), (4,-4), (-4,4) and (4,4). The distributions have scale parameter 1 and are independent. The remaining eight dimensions are noises sampled from a standard exponential distribution. Plots of the two examples can be found in Figure 4 of Appendix A.1.

Examples 3 and 4 are non-distance-based. Example 3 is the two moon example, which consists of two clusters of 100 observations in a two-dimensional space. The shapes of the clusters are similar to that of a waxing and waning crescent moon. The data set is constructed with the *shapes.two.moon()* function in the **clusterSim** package in R, with shape parameters 0 and 2 and radii between 1.9 and 2. Example 4 is the bull's eye example, which consists of two clusters in a two-dimensional space. The data resemble a bull's eye with an inner and an outer ring. The first consists of 80 observations, the latter of 240 observations. The *BullsEye* data set can be found in the **Mixall** package in R. Plots of the two examples can be found in Figure 5 of Appendix A.1.

Note that these data sets are not exact replications of those that appeared in Wang (2010). Important information such as the number of observations, shape parameters and the dependence in the features was missing. Especially for examples 3 and 4 much information was missing. Yet, the main characteristics of the data are the same.

4.1.2 Unequal cluster sizes and correlated features

Krieger and Green (1999) found that previous stability-based measures performed badly when cluster sizes were unequally distributed and when features exhibited high correlations. To evaluate whether the more recent CV$_a$, CV$_v$ and CC measures also suffer from these pitfalls, new data sets are generated. These data sets consist of 192 observations in four clusters in a ten-dimensional space. The cluster means[2] are chosen such that their pairwise Euclidean distances are larger than ten. This ensures that the clusters are well-separated. The observations are generated as $\mu_i + \sigma S_i \epsilon_i$ for cluster $i = \{1, 2, 3, 4\}$. σ is a vector of scaling constants that are set to 0.7, S governs the correlation between the features and ϵ are independent samples from $N(0, 1)$. In the base case, clusters are of equal size and features are uncorrelated, such that S is set to the identity matrix. In the case of unequal cluster sizes, the first cluster contains 114 observations, while the three remaining clusters each contain 26 observations. In the case of correlated features, S is a matrix with ones on the diagonal and off-diagonal elements sampled from $N(1, 1)$. Plots of these data sets can be found in Figure 6 of Appendix A.1.

[2] μ_1 = [1,1,1,1,1,1,1,1,1,1], μ_2 = [2,3,2,2,5,7,2,2,8,2], μ_3 = [0,-2,8,4,5,1,6,0,2,3], μ_4 = [-1,6,-3,2,6,5,4,0,0,-1]

4.1.3 More clusters

Breckenridge (2000) found that previous stability-based measures were biased towards fewer clusters when the true number of clusters was large. To evaluate whether the more recent CV$_a$, CV$_v$ and CC measures also suffer from these pitfalls, new data sets are generated. These data sets consist of $m = \{5, 6, 7, 8, 9, 10\}$ equally sized clusters with uncorrelated features. To allow for a fair comparison, the number of observations is fixed at 400. The clusters thus contain 80, 67, 57, 50, 44, and 40 observations, respectively. The clusters are generated in a similar manner to example 1. Set $x_1 = x_2 = x_3 =$ t, where t takes equally spaced values between -0.5 and 0.5. Gaussian noise with mean 0 and standard deviation 0.1 is then added to the three features. These clusters are then moved by adding or subtracting 5, 10 or 15 to or from the features, such that the clusters are well-separated. Figure 7 of Appendix A.1 illustrates where the clusters are located in the first two dimensions.

4.1.4 A stable but meaningless clustering

Handl et al. (2005) and Hennig (2007) found that previous stability-based measures produced misleading results when the shape of the data caused the clustering algorithm to converge to sub-optimal solutions. Handl et al. (2005) showed that the k-means algorithm constructed a stable but meaningless clustering of a data set with elongated clusters. The authors concluded that this was the result of k-means' tendency to form spherical clusters. To evaluate the validity of this criticism, a new data set with elongated clusters is generated. Recall that example 1 also contains elongated clusters. However, from the left plot in Figure 4 of Appendix A.1 it is clear that these clusters are far apart such that their shapes become negligible. The new data set consists of two elongated clusters with little spatial separation such that their shapes are more influential. The clusters reside in a two-dimension space and each contain 100 observations. This data set is generated as follows: for the first feature, 200 independent values are sampled from $N(0, 2)$. The second feature is set to 1 for the first 100 observations and set to 0 for the second 100 observations. The values of the second feature are then jittered to induce some randomness. A plot of this data set can be found in Figure 8 of Appendix A.1.

4.2 Cross validation measures versus consensus clustering

4.2.1 Stability scores for components

Theory highlights that consensus clustering can be used to compute stability scores for components of the data. This is particularly useful when clusters are overlapping, which is the case in the Iris data set. The Iris data set is available in R and consists of 150 observations that are recorded in four dimensions: sepal width, sepal length, petal width and petal length. The three species of irises, *setosa*, *versicolor* and *virginica*, are each represented with 50 observations. The true number of clusters is three, but as the second and third species are overlapping, many selection methods fail to distinguish the two (Sugar & James, 2003). A plot of the Iris data set can be found in Figure 10 of Appendix A.2.

4.2.2 Visualisation by heat maps

An advantage that consensus clustering has over the cross validation methods is that it can be used to produce a visualisation of the clustering. This is particularly useful in the context of high-dimensional data sets, for which the structure of the data might not be immediately clear from plotting the data. Heat maps then provide an intuitive two-dimensional visualisation of the clustering, irrespective of the dimensionality of the data. As an illustration, the data set with the four equally sized clusters in ten dimensions and uncorrelated features, described in section 4.1.2, is considered. From the left plot in Figure 6 of Appendix A.1, it is clear that plotting the data yields little insight into the underlying structure of the data.

4.2.3 Small sample size

Some concerns were raised about the use of cross validation estimation techniques in small samples. To evaluate the empirical validity of this concern, some variations of example 2 are considered. Example 2 consists of 400 observations in four equally sized clusters. Sub-samples are then taken to construct three data sets with 200, 100 and 48 observations, respectively. The clusters are assumed to be of equal size, such that they each contain 50, 25 or 12 observations. Plots of the three data sets can be found in Figure 9 of Appendix A.2.

4.3 Data pre-processing

Before clustering, the data sets must be checked for missing values, errors and outliers. In the Iris data set no missing values, errors or outliers were detected. For the simulated data sets such issues are absent by construction. Secondly and more importantly, the features are re-scaled to ensure that they receive equal weights in the clustering solution. For each data set in this paper, it is assumed that the features are of equal importance such that re-scaling is imperative. Otherwise, large-scaled features wrongfully receive higher weights. Several re-scaling methods exist, but those that divide by the range have been found to yield the best recovery of the data structure (Milligan & Cooper, 1988). For each data set, the features are re-scaled as follows:

$$\frac{X - min(X)}{max(X) - min(X)} \tag{1}$$

5 Results

5.1 The validity of clustering stability

5.1.1 Four simulated examples (Wang, 2010)

The selection performance of the stability-based CV$_a$, CV$_v$ and CC measures is compared to that of the non-stability-based Silhouette (SIL), KL, GAP and JUMP measures in four simulated examples to evaluate the competitiveness of clustering stability for the selection of the number of clusters. Each of the examples is repeated 50 times. For the distance-based examples, k-means clustering is employed. These results are summarized in Table 1. For the non-distance-based examples, spectral clustering is employed. These results are summarized in Table 2. The tables illustrate how often the proposed measures select a specific k in the 50 runs. The true number of clusters is marked in bold.

Example 1 illustrates a low-dimensional data set with well-separated elongated Gaussian clusters. From Table 1a it is clear that all proposed measures, with the exception of the JUMP measure, consistently select the correct number of clusters. It is likely the elongated nature of the clusters that hinders the JUMP measure from selecting the correct k. Example 2

Table 1: Number of selected clusters for the distance-based examples

k	**2**	3	4	5	6	7	8	9	10
SIL	50	0	0	0	0	0	0	0	0
KL	50	0	0	0	0	0	0	0	0
GAP	50	0	0	0	0	0	0	0	0
JUMP	0	0	0	0	31	1	15	1	2
CC	50	0	0	0	0	0	0	0	0
CV_a1	50	0	0	0	0	0	0	0	0
CV_a2	5	15	27	2	1	0	0	0	0
CV_v	50	0	0	0	0	0	0	0	0

(a) Example 1

k	2	3	**4**	5	6	7	8	9	10
SIL	0	0	49	1	0	0	0	0	0
KL	0	0	46	2	0	0	0	0	2
GAP	0	15	35	0	0	0	0	0	0
JUMP	0	0	48	0	0	0	0	0	2
CC	0	0	50	0	0	0	0	0	0
CV_a1	1	0	49	0	0	0	0	0	0
CV_a2	0	0	32	18	0	0	0	0	0
CV_v	0	0	50	0	0	0	0	0	0

(b) Example 2

is slightly more complex and illustrates a high-dimensional data set with well-separated non-Gaussian clusters. The GAP measure diminishes in selection performance, while the other measures perform well. The stability-based measures perform well in both distance-based examples, suggesting that clustering stability is a competitive measure for the selection of the number of clusters. From the distance-based examples it is clear that the stability-based measures are most competitive with the Silhouette and KL measures.

Table 2: Number of selected clusters for the non-distance-based examples

k	**2**	3	4	5	6	7	8	9	10
SIL	0	0	0	0	0	0	50	0	0
KL	0	0	0	0	0	0	0	50	0
GAP	50	0	0	0	0	0	0	0	0
JUMP	45	0	5	0	0	0	0	0	0
CC	49	1	0	0	0	0	0	0	0
CV_a1	50	0	0	0	0	0	0	0	0
CV_a2	47	3	0	0	0	0	0	0	0
CV_v	50	0	0	0	0	0	0	0	0

(a) Example 3

k	**2**	3	4	5	6	7	8	9	10
SIL	0	0	0	0	0	0	0	50	0
KL	0	0	0	1	0	0	0	49	0
GAP	0	0	0	0	0	0	0	0	0
JUMP	50	0	0	0	0	0	0	0	0
CC	50	0	0	0	0	0	0	0	0
CV_a1	50	0	0	0	0	0	0	0	0
CV_a2	50	0	0	0	0	0	0	0	0
CV_v	50	0	0	0	0	0	0	0	0

(b) Example 4

Examples 3 and 4 are non-distance-based and concern low-dimensional data sets with non-Gaussian clusters that take unconventional shapes such as half moons and annuli. The KL and Silhouette measures now fail to select the correct number of clusters and lose their competitive edge. This was expected as they select k based on the optimization of some within-cluster similarity measure, which is inappropriate for non-distance-based data sets. While the GAP and JUMP measures perform reasonably well in example 3, their performance degrades in example 4. The GAP measure consistently fails to detect a clustering, selecting $k = 1$. Moreover, the distortion curve for example 4, depicted in Figure 1, is monotonically decreasing. As Sugar and James (2003) pointed out, optimizing the JUMP measure may not be sensible in this case. Therefore, the JUMP measure may have only accidentally selected two clusters as it happens to be a corner solution.

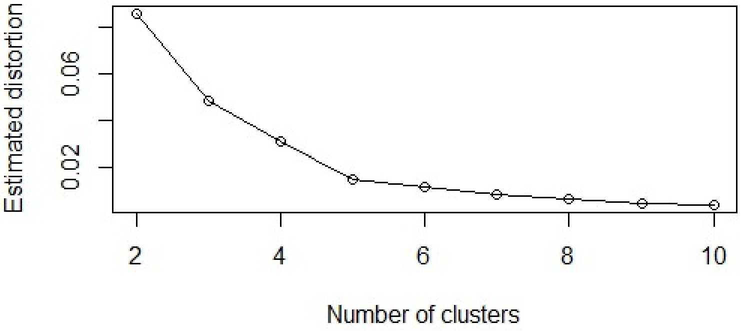

Figure 1: Monotone distortion curve for example 4 (JUMP measure)

The stability-based measures consistently select the correct number of clusters in both the distance-based and the non-distance-based examples, demonstrating their flexibility. The non-stability-based methods seem to be less robust. From the results it is clear that the stability-based measures outperform the already established and generally accepted non-stability-based selection measures. While no formal conclusions about the validity of clustering stability can be drawn from these results, they do suggests that clustering stability is a good, or at least a competitive, measure for the selection of the number of clusters. Furthermore, it is noted that CV_a2 seems to be biased towards more clusters. This is a natural consequence of its decision rule. The more simplistic CV_a1 seems to be preferred.

Recall that the examples are not exact replications of those that appeared in Wang (2010). Consequently, the results are not exactly the same. The slight variation in data sets reveals features of the selection measures that were previously left uncovered. For instance, example

4 illustrates one of the major pitfalls of the JUMP measure. Namely that its distortion function may be monotone, in which case a corner solution is selected. Nonetheless, as the key characteristics of the data sets were preserved, the main conclusion remains the same. The stability-based measures outperform the non-stability-based measures in all four simulated examples. These results support the validity of clustering stability for the selection of the number of clusters. They do not, however, imply that stability-based measures are always appropriate for the selection of the number of clusters. Considering that the examples were taken from Wang (2010), who selected them to illustrate the competitiveness of his cross validation measures, it is not surprising that the stability-based measures perform well. The results paint a too optimistic picture that cannot necessarily be generalized to other data sets. To provide an unbiased picture of clustering stability as a measure for the selection of the number of clusters, some cases in which previous stability-based selection measures failed to select the correct number of clusters are considered

5.1.2 Unequal cluster sizes and correlated features

Previous stability-based measures were found to perform badly when cluster sizes were unequally distributed or when features were correlated (Krieger & Green, 1999). To evaluate whether this is also the case for the CV_a, CV_v and CC measures, new data sets were generated. These data sets consist of 192 observations in four clusters in a ten-dimensional space. The data sets are distance-based and k-means is employed as the clustering algorithm. Note that the differences in cluster size only concern the number of observations that they entail and not the area that they take up. As the k-means algorithm is concerned with the distance between the observations (Hartigan & Wong, 1979), the clustering solution is not likely to be affected by the differences in the cardinalities of the clusters or the correlation between the features. The differences in the performance of the selection measures can thus be entirely contributed to the features of the data. Each of the examples is repeated 50 times. Table 3 summarizes the hit rates of the stability-based measures, i.e. the proportion of runs in which they select the correct number of clusters.

All three measures perform well in the case of equal cluster sizes and uncorrelated features. When the cluster sizes are unequally distributed, all three measures degrade in selection

Table 3: Hit rates of the stability-based measures for three data sets

	Equal cluster sizes & Uncorrelated features	Unequal cluster sizes	Correlated features
CC	1.00	0.00	1.00
CV_a1	1.00	0.80	0.00
CV_v	1.00	0.86	0.00

performance. CV_a1 and CV_v still perform reasonably well, but CC consistently fails to detect the correct number of clusters. On the other hand, when the features are correlated, CC performs well, while CV_a1 and CV_v perform badly. The results highlight that each approach has its own strengths. Future research should look into the precise characteristics of each of the measures that allows them to perform well in these circumstances.

The results for the cross validation methods are consistent with those of Krieger and Green (1999). However, CC performs differently than expected. A possible explanation is that consensus clustering takes a different approach in computing and evaluating stability scores compared to conventional stability-based measures. Conventional methods tend to estimate some instability score and minimize this, like the cross validation measures do. As Krieger and Green (1999) employ a rather conventional method, their results are expected to be more in line with those of CV_a1 and CV_v.

5.1.3 More clusters

To evaluate whether CV_a, CV_v and CC are biased towards fewer clusters when the true number of clusters is large, data sets with $m = \{5, 6, 7, 8, 9, 10\}$ equal-sized clusters are considered. The data sets are distance-based and k-means clustering is used. Each case is repeated 50 times. The hit rates of each of the measures, i.e. the proportion of runs in which they select the correct number of clusters, can be found in Table 4.

For $h = 5$ all measures consistently select the correct number of clusters. However, as h increases the performance of CC starts to worsen, followed by that of CV_a2, CV_a1 and CV_v. The measures fail to select the correct number of clusters and are biased towards fewer clusters, which is in accordance with the results of Breckenridge (2000). Even CV_a2 which was previously biased towards more clusters is now biased towards fewer clusters.

Table 4: Hit rates of the stability-based measures for various numbers of clusters

m	5	6	7	8	9	10
CC	1.00	0.30	0.00	0.00	0.00	0.00
CV$_a$1	1.00	1.00	1.00	0.04	0.00	0.00
CV$_a$2	1.00	0.80	0.22	0.00	0.00	0.00
CV$_v$	1.00	1.00	1.00	1.00	0.98	0.00

Intuitively, these results make sense. When the true number of clusters is large, the clustering algorithm may consistently merge the same clusters together, such that a clustering with less clusters appears to be stable. Moreover, instability scales with k (Von Luxburg, 2010). As the number of clusters increases, there are more possibilities of assigning each observation to a cluster. More possibilities lead to more uncertainty and thus higher instability. Therefore, lower values of k may result in less instability. As none of the measures normalize the instability scores prior to optimizing over k, they are expected to be biased towards fewer clusters. Normalization of the instability scores is thus important, especially when the number of clusters is potentially large. Researchers have however not yet reached consensus on how normalization should be implemented in practice (Von Luxburg, 2010), highlighting the importance of further research on this matter.

5.1.4 A stable but meaningless clustering

A data set with two elongated clusters with little spatial separation was generated to evaluate the criticism of Handl et al. (2005). This data set is distance-based and k-means clustering is used. The example is repeated 50 times. Table 5 summarizes how often each measure selected a specific value of k. The true number of clusters is marked in bold.

Contrary to what is expected from earlier results, all three stability-based measures perform well in the data set with elongated clusters. The disparity in results follows from the fact that Handl et al. (2005) did not re-scale the data prior to clustering, causing k-means to falsely assign more weight to the large-scaled feature. The resulting clustering was thus invalid. However, it was stable as the same incorrect clustering was formed in each perturbed sample. When the re-scaling step is omitted, the CV$_a$, CV$_v$ and CC measures also fail to

TABLE 5: Number of selected clusters in the data set with two elongated clusters

	k	**2**	3	4	5	6	7	8	9	10
Re-scaled	CC	50	0	0	0	0	0	0	0	0
	CV$_a$1	50	0	0	0	0	0	0	0	0
	CV$_v$	50	0	0	0	0	0	0	0	0
Not re-scaled	CC	0	2	13	24	11	0	0	0	0
	CV$_a$1	7	2	0	0	0	0	0	0	41
	CV$_v$	22	4	0	0	0	0	0	0	24

select the correct number of clusters and choose $k = 10$ or $k = 5$ instead.

Figure 2 illustrates the cluster membership of the not re-scaled observations for $k = 10$. Each colour corresponds to a cluster that is formed by the k-means algorithm. The true clusters correspond to the two horizontally elongated shapes. Clearly, k-means assigns more weight to the horizontal axis as it is recorded on a larger scale. As a result spherical clusters are formed. The outer orange, blue, dark green and violet clusters are particularly problematic as they contain observations from both elongated shapes. These clusters make it impossible to recover the true structure of the data, even when the characteristics of the formed clusters are inspected. This example illustrates how clustering stability could produce misleading results. The clustering is stable but meaningless.

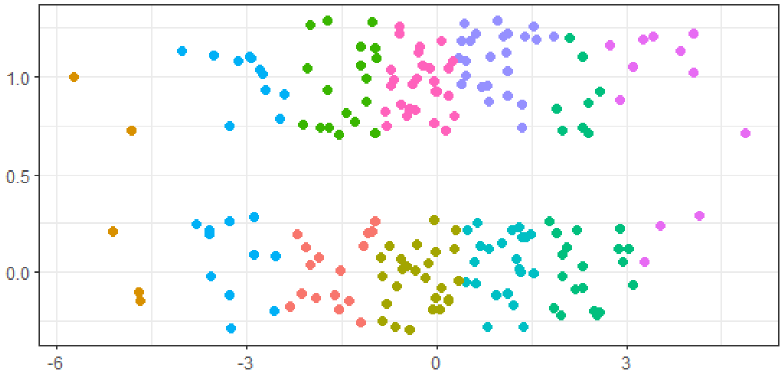

FIGURE 2: Cluster membership for $k = 10$

This result is not attributed to the fact that clustering stability is bad in itself. Instead, it is the lack of adequate data pre-processing steps that cause the measures to select the wrong number of clusters. This highlights the importance of making the right choices in each step of the model selection process. The right conditions are crucial for the selection performance of the stability-based selection measures (Von Luxburg, 2010). Nonetheless, this result still presents a counter example for the validity of clustering stability for the selection of the number of clusters. It shows that a stable solution cannot be guaranteed to be valid and may thus not correspond to the structure of the data. While something can be said for the use of stability-based selection methods, it is clear that they alone are not sufficient for the selection of the number of clusters. It is therefore recommended that the stability-based analysis is complemented by other selection methods.

5.2 Cross validation measures versus consensus clustering

5.2.1 Stability scores for components

The Iris data, in which the *versicolor* and *virginica* species are overlapping, is used to illustrate why it is useful to compute stability scores for components of the data. The data set is distance-based and k-means is used as the clustering algorithm. The example is repeated 50 times. The choices of the stability-based selection measures are summarized in Table 6. The true number of clusters is marked in bold.

TABLE 6: Number of selected clusters in the Iris data set

k	2	**3**	4	5	6	7	8	9	10
CC	0	50	0	0	0	0	0	0	0
CV_a1	50	0	0	0	0	0	0	0	0
CV_v	50	0	0	0	0	0	0	0	0

CC consistently selects the correct number of clusters, while CV_a1 and CV_v select two clusters, thereby merging the two overlapping species. In an effort to improve the selection performance of the cross validation measures, the stability score of each species is computed. The *setosa* species exhibit the highest stability, with an average consensus score of 1. This is expected as the *setosa* species is distinct from the other two species. For the *versicolor* and *virginica* species the stability scores are 0.8533 and 0.6176, respectively.

17 observations are found on the boundary of the *versicolor* and *virginica* clusters. These

observations induce instability as the algorithm is unable to detect to which species they truly belong. When these 17 observations are isolated as a separate component, the *versicolor* and *virginica* species become much more stable with stability scores of 0.9559 and 0.9485, respectively. On the other hand, the component of 17 observations has a mere score of 0.6231. Following Lord et al. (2017) these 17 observations are removed from the data. As expected, the selection performance of the cross validation measures improves. From Tables 6 and 7 it is clear that the hit rates of CV$_a$1 and CV$_v$ increase from 0.00 to 0.28 and 0.76, respectively.

TABLE 7: Number of selected clusters in the Iris data set after removing the 17 observations

k	2	**3**	4	5	6	7	8	9	10
CV$_a$1	36	14	0	0	0	0	0	0	0
CV$_v$	13	37	0	0	0	0	0	0	0

5.2.2 Visualisation by heat maps

Consensus clustering can be used to visualise the clustering, irrespective of the dimensionality of the data. The data set with four equally sized clusters in ten dimensions is used as an illustration. The heat maps corresponding to k = 2, 3, 4 and 5 are displayed in Figure 3. Recall that the rows and columns of the heat maps correspond to the observations, for which the ordering is determined by the dendogram to achieve the most diagonal heat map.

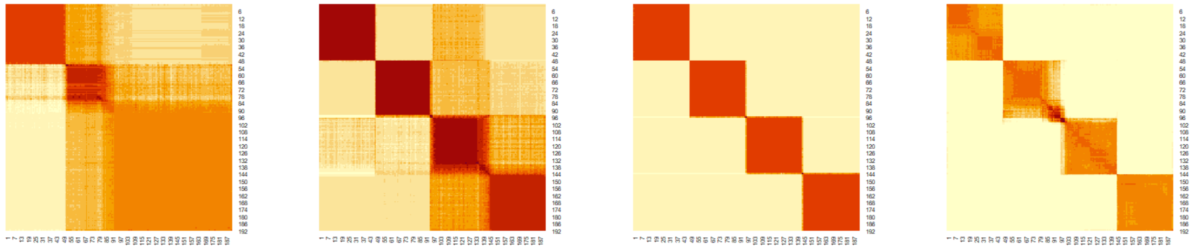

FIGURE 3: Heat maps of $M(2)$, $M(3)$, $M(4)$ and $M(5)$

From Figure 3 it is clear that four clusters should be chosen, as $M(4)$ produces the "cleanest" heat map. The diagonal blocks that correspond to the four clusters are opaque and little to no noise is detected on the off-diagonal elements. The heat map shows that the four clusters are of equal size and that they exhibit equal instabilities. Each of the clusters is identified, such that they must be well-separated. From the remaining heat maps it is clear that the true clusters are merged or split, when the wrong number of clusters is chosen. In $M(2)$ the last three clusters are merged, in $M(3)$ the last two clusters are merged, and in $M(5)$ the second cluster is split. Each of these operations induce instability, resulting in more noise on the off-diagonal elements.

From this illustration it is clear that heat maps can provide a visualisation of high-dimensional clusterings. This is not to say that heat maps are not useful for low-dimensional clusterings. Even if the data structure is apparent from plotting, heat maps may help in the decision of the number of clusters, yield insight on how the clusterings evolve as k increases and illustrate the instabilities of individual clusters.

5.2.3 Small sample size

To evaluate whether cross validation estimation techniques are problematic in small samples, data sets with 400, 200, 100 and 48 observations are considered. The data sets are distance-based and k-means is used as the clustering algorithm. Each case is repeated 50 times. The hit rates of the stability-based measures, i.e. the proportion of runs in which they select the correct number of clusters, can be found in Table 8.

TABLE 8: Hit rates of the stability-based measures for various sample sizes

N	400	200	100	48
CC	1.00	1.00	0.96	0.98
CV$_a$1	1.00	1.00	0.82	0.00
CV$_v$	1.00	1.00	0.94	0.00

As the sample size N decreases, the performance of both CV$_a$1 and CV$_v$ starts to degrade. For $N = 100$ they still perform reasonably well, but for $N = 48$ they consistently fail to select the correct number of clusters. CC, on the other hand, still performs well. These results suggest that splitting of the data into small sub-samples is inappropriate for small samples and that cross validation estimation techniques are problematic. As CC takes 70% of the observations of the original data set, it encounters less problems.

Note that this example is distance-based and has well-separated clusters. The underlying structure is therefore clear-cut and well-preserved when small sub-samples are taken. In non-distance-based data sets or data sets with little spatial separation between the clusters, taking small sub-samples may come at the costs of the underlying data structure. In these cases, the underlying data structure cannot be detected such that the clustering is formed in an arbitrary manner. Cross validation estimation techniques are then more problematic and larger sample sizes than are suggested by these results may be required. With the rise of new technologies, however, data sets have grown to contain many observations such that this concern for the cross validation measures becomes negligible. This pitfall of the cross validation measures thus seems to be more theoretical than it is practical.

From Tables 3, 4, 5, 6, 7 and 8 it is also clear that the performance of $\mathrm{CV}_a 1$ worsens at a faster rate than that of CV_v. This is attributed to the fact that the voting strategy of CV_v is more robust against outlying clusterings compared to the averaging strategy of CV_a, such that the first performs better under various conditions.

6 Practical implications

Cluster analysis has applications in various fields of studies. In marketing, the technique is often used to form representative groupings of products, customers or markets. With the problem of clustering, the problem of selecting the appropriate number of clusters follows.

One of the primary applications of cluster analysis in marketing is market segmentation. It aims to divide a heterogeneous market, one that is characterized by divergent demands, into smaller homogeneous segments based on the differences in product preferences across the segments (Smith, 1956). The characteristics of the segments may then aid in gaining a better understanding of buyer behaviour. The groupings are often formed based on consumer choice tasks and individual characteristics. The first of which concerns human behaviour, which is inherently unstable. As a consequence, the data may contain outliers and high variability. Moreover, small differences in behaviour may correspond to distinct groups such that the clusters have little spatial separation or are even overlapping. This complicates the clustering task. In addition, such data sets often exhibit correlations in the features. For instance, the number of visits to a store and total expenditure are likely to be positively correlated. The use of the cross validation measures for the selection of the number of clusters may then be problematic. Furthermore, such data sets often contain a large number of clusters that may be of unequal size. Large clusters correspond to the main target markets, while small clusters correspond to niche markets. Stability-based selection measures must then be used with care. While the previous results have illustrated the consequences of each of these challenges in isolation, it is yet unclear what the implications are when data sets contain multiple of these challenges simultaneously. As this is often the case for real data sets, it is important that future researchers look into this matter. Despite these challenges, cluster analysis is still used in practice for market segmentation. For instance, Furse, Punj, and Stewart (1984) used cluster analysis to successfully identify six distinct search patterns among purchasers of new automobiles.

In addition, cluster analysis could be used for market structure analysis. It is crucial for firms to understand their competitive landscape as it guides their strategic decision making. By clustering products, firms can evaluate their current competitive position in the market. In this way, firms can determine whether their offerings are uniquely positioned and who their main competitors are. Moreover, this technique can help to identify gaps in the market, which can be used to guide the development of new product offerings. Amongst others, Srivastava, Leone, and Shocker (1981) found that clustering methods yield inherently interpretable groupings that could be used to guide the strategic decisions of managers. Cluster analysis could also be used for the problem of test market selection. Test marketing concerns the practice of evaluating a certain marketing activity, for example a product launch or an advertising strategy, in a limited geographical area (Green, Frank, & Robinson, 1967). The success of the test market program depends on whether the responses in the test market can be generalized to a greater area. Cluster analysis could be used to identify homogeneous markets, such that the results can be generalized within clusters. In this way, the number of test market programs that have to be executed can be reduced, thereby reducing costs.

While it is interesting to study the performance of stability-based selection measures in real data sets, it is not within the scope of this paper. Those type of analyses are specific to the example at hand and cannot be generalized to other cases. The objective of this research paper was to provide a clear picture of the strengths and weaknesses of the stability-based selection measures to guide a more considered use. Simulated examples were employed, as they allow for a controlled environment, such that the results can be generalized. This is not to say that the application of the stability-based selection measures in real life examples is not interesting. On the contrary, many applications of cluster analysis exist and future research should look into the performance of stability-based selection measures in these instances.

7 Conclusion and discussion

7.1 Conclusion

The objective of this paper was to identify the strengths and weaknesses of stability-based selection measures. Based on the results, it can be concluded that there are compelling arguments for the use of stability-based measures for the selection of the number of clusters. The three stability-based measures outperformed several popular non-stability-based measures in four simulated examples. These examples illustrated the flexibility of the stability-based measures and highlight that clustering stability is a competitive measure for the selection of the number of clusters. However, this paints a too optimistic picture. The stability-based measures cannot be expected to always perform well. Indeed, when features were correlated, the cross validation measures were found to perform badly, while consensus clustering performed well. On the contrary, when clusters were of unequal size, the cross validation measures performed reasonably well, while the performance of consensus clustering deteriorated. Furthermore, when the true number of clusters grew large, all three measures were biased towards fewer clusters. These results are concerning as unequal cluster sizes, correlated features and a large number of clusters are not uncommon in real data sets. Yet, most concerning is the existence of stable but meaningless clusterings. While this result followed from a lack of data pre-processing steps, it shows that a stable clustering is not necessarily valid. This is a counter example for the validity of clustering stability for the selection of the number of clusters. It is concluded that clustering stability is not a valid measure for the selection of the number of clusters. However, previous results suggest that, given the right conditions, stability-based measures could be helpful in selecting the number of clusters.

Therefore, they should not be used in isolation but as a component of a broader strategy for the selection of the number of clusters. In this way, the risk of accepting wrong results can be minimized.

To guide a more informed use of stability-based selection measures, the cross validation measures were compared to consensus clustering. From the results it is concluded that cross validation estimation techniques are problematic in small samples, such that consensus clustering is preferred. Other advantages of consensus clustering include the visualisation of the clusterings and the computation of stability scores for components of the data. The first is especially useful in high-dimensional data sets and yields additional insights. The latter allows for the identification of unstable components. It was found that, when these components were removed, the recovery of the true number of clusters improved. For the cross validation measures, CV_v was found to be more robust than CV_a.

Consensus clustering can be used for a wide range of purposes, favouring it over the cross validation measures. However, its decision rule is somewhat vague, especially when $\Delta(k)$ converges to zero. In this case it is unclear for which k $\Delta(k)$ is significantly larger than zero. The inspection of the heat maps is then needed to select k. Visual inspection is rather subjective, which is undesirable as it may cause researchers to unconsciously favour results that support their own assumptions. On the contrary, the cross validation measures provide a stricter decision rule but are unable to yield additional insights about the clusterings. Neither of the approaches can thus be concluded to be uniformly the best. Depending on the conditions at hand, one of the approaches may be preferred. If it is unclear which one is preferred, it is recommended that both are implemented and that their results are compared.

7.2 Discussion

This paper contributes to the literature in several ways. Firstly, it provides an unbiased view of clustering stability for the selection of the number of clusters. This is important as it guides the choice of future researches of which measure to use for the selection of the number of clusters. Previous research often emphasized either the strengths or the weaknesses of the measures in question, thereby failing to provide an unbiased view. Secondly, it provides a formal evaluation of the validity of clustering stability for the selection of the number of clusters. It is made clear that stability-based measures can be used for the selection of the number of clusters. However, they should be used with care and in combination with other selection methods. Lastly, the comparison of the cross validation measures with consensus clustering has not been considered in previous literature. This study yields valuable insights on how clustering stability should be implemented in practice. This lays a foundation for future research.

Due to time constraints, this paper evaluated three stability-based measures: CV_a, CV_v and CC. These methods were chosen as they cover a wide range of existing stability-based selection measures, such that the results can be generalized. In fact, other variations exists for the computation and evaluation of instability scores. It is important that future researchers objectively compare all these measures to obtain a better idea of how clustering stability should be implemented in practice for the selection of the number of clusters. This paper also employed two clustering algorithms: k-means and spectral clustering. The choice of these algorithms was motivated by their simplicity, popularity and efficiency. In practice, other clustering algorithms exist that may be more suitable in certain data sets. From the results it is clear that the stability-based measures may select the wrong number of clusters due to the inflexibility of the employed clustering algorithm. Therefore it is recommended that future research looks into the selection performance of the stability-based measures in combination with other more flexible clustering algorithms.

Lastly, this paper identified some challenges for the use of stability-based selection measures. These challenges were investigated in isolation and it is yet unclear what the consequences are when the data contains multiple of these challenges simultaneously. This is likely to be the case in real data sets. While the examples in this paper were designed to incorporate some characteristics of real data sets, they are still rather stylized examples. Future research should look into the selection performance of the stability-based measures in real data sets.

References

Bar-Joseph, Z., Demaine, E. D., Gifford, D. K., Srebro, N., Hamel, A. M., & Jaakkola, T. S. (2003). K-ary clustering with optimal leaf ordering for gene expression data. *Bioinformatics*, *19*(9), 1070–1078.

Ben-David, S., Von Luxburg, U., & Pál, D. (2006). A sober look at clustering stability. In *International conference on computational learning theory* (pp. 5–19).

Ben-Hur, A., Elisseeff, A., & Guyon, I. (2001). A stability based method for discovering structure in clustered data. In *Biocomputing 2002* (pp. 6–17). World Scientific.

Breckenridge, J. N. (2000). Validating cluster analysis: Consistent replication and symmetry. *Multivariate Behavioral Research*, *35*(2), 261–285.

Furse, D. H., Punj, G. N., & Stewart, D. W. (1984). A typology of individual search strategies among purchasers of new automobiles. *Journal of consumer research*, *10*(4), 417–431.

Green, P. E., Frank, R. E., & Robinson, P. J. (1967). Cluster analysis in test market selection. *Management science*, *13*(8), B–387.

Handl, J., Knowles, J., & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, *21*(15), 3201–3212.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108.

Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, *52*(1), 258–271.

Krieger, A. M., & Green, P. E. (1999). A cautionary note on using internal cross validation to select the number of clusters. *Psychometrika*, *64*(3), 341–353.

Krzanowski, W. J., & Lai, Y. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, 23–34.

Lange, T., Roth, V., Braun, M. L., & Buhmann, J. M. (2004). Stability-based validation of clustering solutions. *Neural computation*, *16*(6), 1299–1323.

Lord, E., Willems, M., Lapointe, F.-J., & Makarenkov, V. (2017). Using the stability of objects to determine the number of clusters in datasets. *Information Sciences*, *393*, 29–46.

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of classification*, *5*(2), 181–204.

Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, *52*(1-2), 91–118.

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

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Shamir, O., & Tishby, N. (2008). Cluster stability for finite samples. In *Advances in neural information processing systems* (pp. 1297–1304).

Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. *Journal of marketing*, *21*(1), 3–8.

Srivastava, R. K., Leone, R. P., & Shocker, A. D. (1981). Market structure analysis: hierarchical clustering of products based on substitution-in-use. *Journal of Marketing*, *45*(3), 38–48.

Sugar, C. A., & James, G. M. (2003). Finding the number of clusters in a dataset: An information-theoretic approach. *Journal of the American Statistical Association*, *98*(463), 750–763.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423.

Von Luxburg, U. (2010). *Clustering stability: an overview.* Now Publishers Inc.

Wang, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika*, *97*(4), 893–904.

Appendix A Data sets

The following contains an overview of the employed data sets in this paper. The plots are only generated for the first two dimensions, but this often provides enough intuition about the structure of the data sets. The data sets in Figure 6 are an exception, in which all ten dimensions are required to grasp the structure of the data. This is however not possible and the plots of the first two dimensions are included for completeness.

A.1 The validity of clustering stability

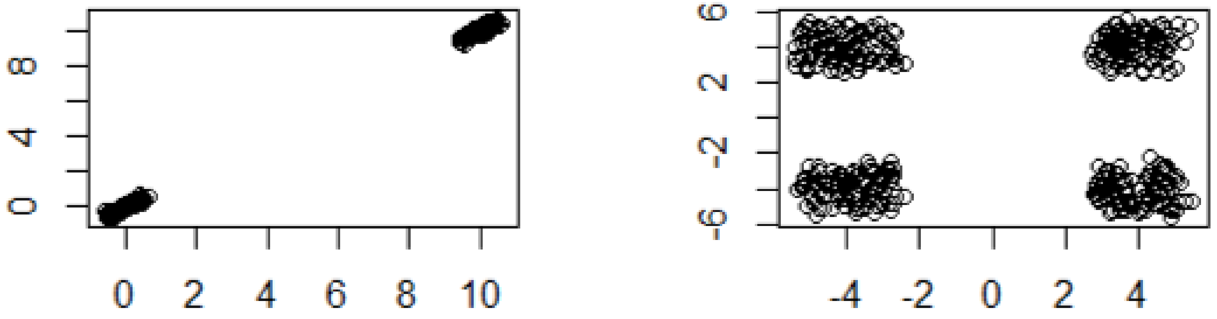

Figure 4: Distance-based examples (1 and 2)

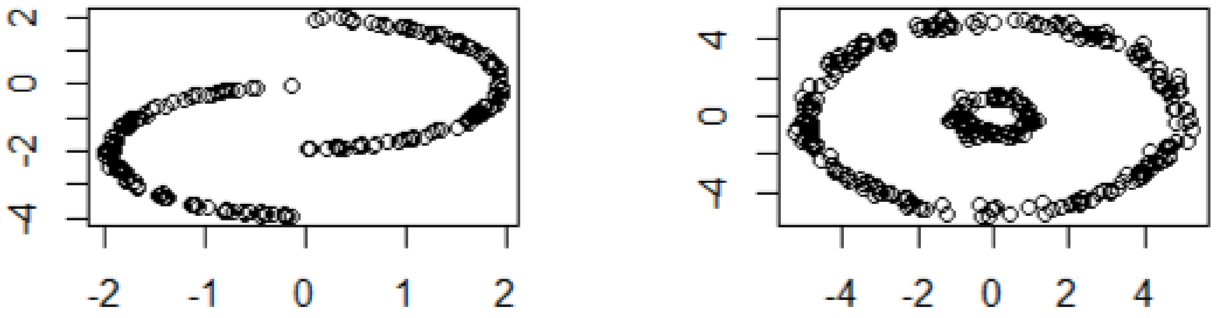

Figure 5: Non-distance-based examples (3 and 4)

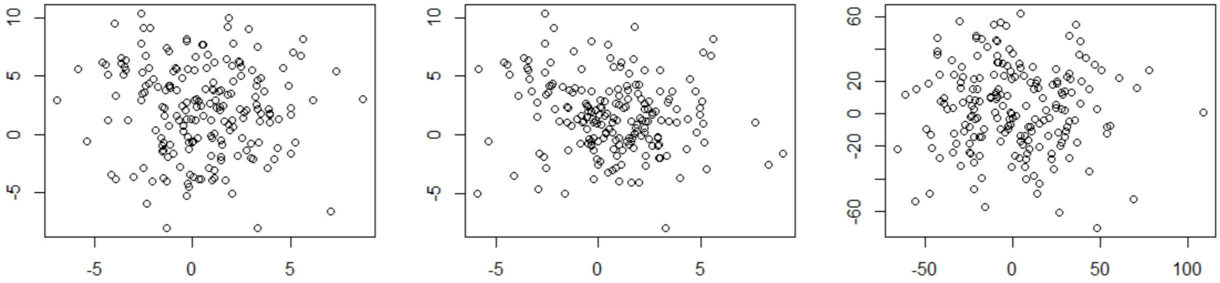

FIGURE 6: Data sets with equal cluster sizes and independent features, unequal cluster sizes and independent features, and equal cluster sizes and dependent features (from left to right)

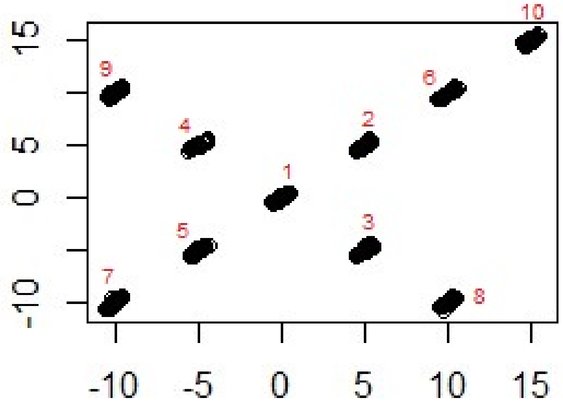

FIGURE 7: Data sets with $m = \{5, 6, 7, 8, 9, 10\}$ clusters

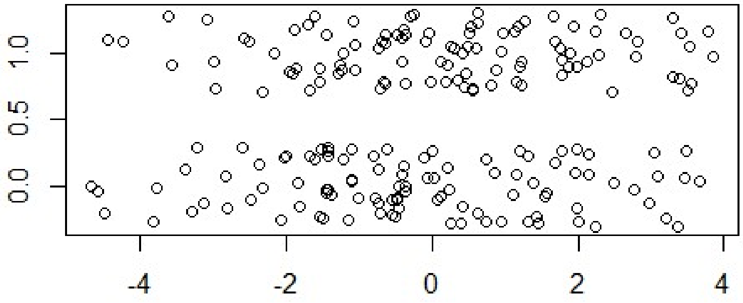

FIGURE 8: Data sets with two elongated clusters (stable but meaningless)

A.2 Cross validation measures versus consensus clustering

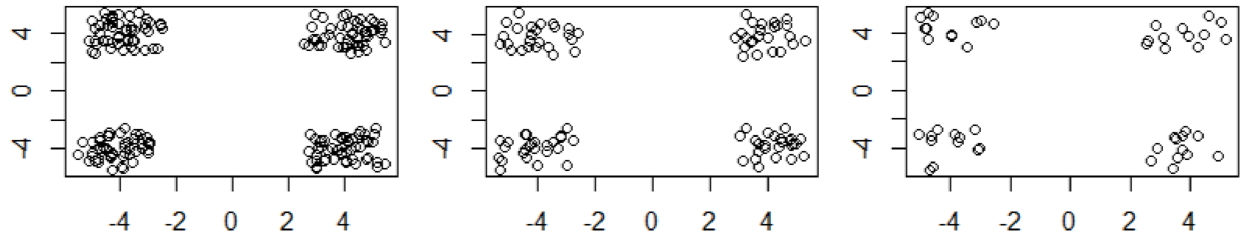

FIGURE 9: Data sets with sample sizes $N = \{200, 100, 48\}$

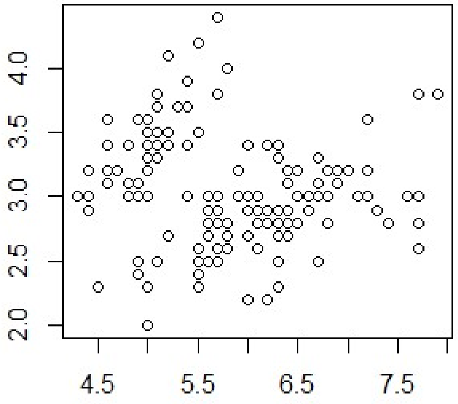

FIGURE 10: The Iris data set with overlapping clusters

Appendix B Programming codes

This simulation study was implemented in R. In total 9 scripts were written to execute the simulations. Script 1 contains auxiliary functions and functions for the implementation of the non-stability-based selection measures. Script 2 contains functions for the cross validation measures. Script 3 contains functions for consensus clustering. These functions are utilised in scripts 4 to 9, in which the various data sets are generated and used to evaluate the performance of the selection measures. The ordering of these scripts is in line with the ordering in which the results are presented in this paper.

A more precise description of the scripts and the exact programming codes can be found in the attached zip file.