A framework approach to initialisation dependent clustering methodologies.

Simon James Chambers

A thesis submitted in partial fulfilment of the
requirements of Liverpool John Moores University
for the degree of Doctor of Philosophy

April 2015

Abstract

Clustering algorithms are commonly used for exploratory data analysis and data mining and used correctly are powerful tools for gaining insights into the underlying structure of data. It is known however that some of these algorithms are dependent upon the parameters with which they start, giving differing results as these vary. Often there is an element of randomness in the initialisation process greatly increasing the difficulty of selecting an appropriately initialised solution.

Effective use of these algorithms depends upon the correct choice of appropriate initialisations, however when exploring new data it is often difficult to objectively obtain values appropriate to the problem. The use of initialisation strategies to maximise the performance of the algorithm are therefore important to ensure solutions identified are both consistent with the structure of the data and reproducible.

This thesis introduces a coherent strategy for dealing with initialisation in the form of chosen parameter selection and randomness. A Separation Concordance (SeCo) framework is developed which uses a dual measure approach to evaluating the solutions from resampling of starting conditions. This SeCo framework also allows for the inference of an appropriate number of partitions within the data and introduces a SeCo map for visualising the solution space.

The performance of these visualisations compared and contrasted with the existing methods in use through an exhaustive series of experiments for both algorithms tested, and is shown to be effective in the selection of a repeatable solution with high concordance to the underlying structure of the data. These results are benchmarked using a range of synthetic and real world data-sets whose composition ranges from trivial to complex.

Acknowledgements

I would like to take this opportunity to express my thanks to my supervisor, Dr Ian Jarman, for the patience, kindness, support and guidance he has given throughout my studies and without which I would have been unable to complete this work. I would also like to thank Prof. Paulo Lisboa for his endless suggestions and insights and Dr Terence Etchells for the original idea which formed the basis for the PhD.

I'd obviously not get away without thanking my Wife, Kinga, for her support and for putting up with me for the last few years and for her ongoing ability to keep me grounded and not take me too seriously, and my parents for their support and encouragement. I'd also like to thank those I shared an office with, Sandra, Younis, Alex and Rizhuan for providing endless distracting conversation and discussion.

Finally I would like to thank Liverpool John Moores University and the School of Computing and Mathematical Sciences for providing the space and funding which has allowed me to pursue a PhD.

Contents

Glossary of Commonly Used Terms

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
Cramers V or CV	Cramers V statistic - a measure of association between nominal variables
Concordance	The measure of association between two variables
Separation Concordance (SeCo)	The combination of two variables for evaluating the separation and concordance of clustering solutions. See SeCo Map.
SeCo Map	A graph using two variables, one for solution separation and the other for between solution concordance to explore solution space.
K-Means	The K-Means clustering algorithm.
ART	Adaptive Resonance Theory - a neural network model for pattern recognition.
ART-2A	A fast implementation of the ART-2 model for grouping continuous real valued data.
ICC	Integrated Cumulative Cramer V - A measure for assessing the number of clusters for a given data set
SSE	Within Cluster Sum of Squared Error - the function optimised by K-Means and a method of evaluating cluster separation
Invariant J	A means of assessing cluster separation using the trace of the between cluster scatter matrices.
Median CV	The median value of the pair-wise Cramer V tests of a given solution with respect to all others.

Introduction 1

Clustering forms part of the larger process of Knowledge Discovery[1] which is collectively the set of techniques used to extract information from data obtaining useful insights into the choices and behaviours of a population. As more data is collected in various forms the minutia of our daily lives are stored and catalogued making the task of getting meaningful understanding from such varied sources becomes ever more difficult.

As the field has advanced, models have been developed allowing researchers to delve into these datasets and tease out relevant nuggets of information, an example of which might be to identify those customers most likely to respond positively or benefit from a marketing strategy. Many of these algorithms rely upon grouping or clustering the data by separating the data elements into cohorts with similar defining characteristics.

Data exploration and analysis makes extensive use of these clustering techniques although some of the algorithms used are dependent either upon their starting conditions or upon the order in which data is fed in. As with any numerical method this should be approached with a critical eye. Examples of such algorithms dependent upon starting conditions are Adaptive Resonance Theory (ART)[2], Decision Tree algorithms[3], partitioning algorithms such as generalised K-Means algorithms[4] or an expectation maximisation based algorithm[5].

It would be very easy to fall into the trap of assuming such algorithms are deterministic in nature with the consequence of applying them in a similar manner to that of regression models in classical statistics. A more considered approach would be to attempt to locate a more optimal solution given some known criteria such as the objective function of the algorithm;

however it has been demonstrated that it is not necessarily possible to obtain a globally optimised solution.

The objective of this thesis is to test a framework approach which can be applied to many such algorithms whose starting conditions play a significant part in determining the final result. The framework will apply a two-factor approach to locating suitable results which are robust and stable against multiple initialisations. The framework will also provide a visualisation of the space of solutions to assist in identifying the chosen solution. This visualisation will present the context of all potentially good solutions which map well onto the underlying data structure.

1.1 Background

This review of the literature comprises three primary elements, an introduction to the problem domain, a summary of the two algorithms explored as part of this thesis and a look at initialisation problems these two algorithms face.

Data mining forms the analysis component of Knowledge Discovery[1] and is an important part of exploratory data analysis where grouping elements together to find structure and clustering algorithms are among the most commonly used. Clustering algorithms broadly fall into one of two different techniques, hierarchical and partitioning with the former having divisive and agglomerative types. Hierarchical algorithms find groups of data by splitting unlike groups of data or grouping like elements respectively and are recursive in nature. Partitioning algorithms are iterative in nature and often attempt to find all partitions simultaneously, [4, 6] with K-Means being widely recognised and used with algorithms implemented natively in common mathematical tools such as R, SPSS, SAS and Matlab.

Representation of data as a set of sensible groups is a fundamentally important aspect of understanding the structure of the data. The easy availability of such algorithms means that it is important to judge an algorithms performance and suitability given a task, not just in respect of whether the result is correct but also in regard to the best theoretically achievable result for the algorithm itself.

Often these algorithms are initialisation dependent and produce different solutions as the starting conditions change, ultimately meaning that some objective measure of suitability is required to differentiate between them. Specifically, it has been known for a long time that although the K-Means algorithm converges its objective function to a local minima it does not necessarily locate a global minima [7]. In practice it is not feasible to locate the best local minima as it is know that finding such a minima is NP-Hard [8] and it is not possible to determine whether a particular solution is global or not.

In an attempt to avoid a local optima, it has been suggested that performing the K-Means algorithm a number of times [7, 9, 10] with different initial conditions would help avoid a solution far from the global optima. However it has been subsequently shown that it is likely that any given solution is a local optima, such that there will be more optimal solutions[11] and the use of even moderately sized datasets can give rise to a scenario where there are many local optima[12].

A question that follows from this is whether it is possible to know whether a particular clustering solution is a good representation of the underlying data structure [11]. Application of K-Means to a dense Gaussian distribution would result in a Voronoi decomposition of the data however this would not necessarily mean that there was any underlying structure to the data simply that the algorithm could partition the elements. It would further be impossible in such a scenario to give any indication as to the correct value for k.

None of this is to say that there is no unique global optima just that it is impractical to locate this even through exhaustive reinitialisation. It may be possible to determine whether the currently chosen solution is close to being globally optimal [5] however this is not the same as being globally optimal.

For the successful application K-Means the common assumption is that it is sufficient to carry out a number of random initialisations followed by selection of the best separated solution, for instance measured by the sum of square distances from cluster prototypes. Various aspects of this process from the choice of separation measures to the best number of clusters k are guided by heuristic methods. Given the age of this particular algorithm it is perhaps astonishing that there is no prescribed method for sampling a single, stable and well separated solution for which repetition of the experiment would obtain a similar solution.

1.2 K-means Algorithm

The K-means algorithm was itself proposed by Steinhaus [13] and Lloyd [14] and was designed with the aim of reducing within cohort variance (Equation 2.1), and initially used as a means of reducing the variance during stratified sampling [15, 16, 17] and Lloyd's algorithm was in this vein looking at the Quantisation of continuous pulse code modulation data.

Variations of these algorithms appeared specifically targeted for the partitioning of more traditional multivariate data and these were subsequently developed to mitigate the question of robustness mentioned above. Examples of this are Forgy [9, 10, 18]. Elements of this included initialisation strategies [18] and adjustments to the method itself[10] and a description of two of the different methods can be found in Algorithms Algorithm 1 and Algorithm 2. The ease of use and implementation along with its practicality, efficiency and easy of use have all been reasons for its continued popularity, despite being in use for over half a century [4].

To some extent these algorithms were a follow on from the others, MacQueen[9] advocated selection of the first k data points as the centroids which is susceptible to the ordering of the data. However Forgy [18] advocated the use of random initial centres, such that each iteration of the method would yield different results. Hartigan-Wong[10] uses an on-line update stage to iteratively reassign elements to their next nearest centroid to see if it lowers the global minima, this is intended to allow the algorithm to get as close to the global optimal as possible.

K-Means is commonly used to refer to Lloyd's algorithm, however it can refer to any generalised procedure using k centroids and minimising the within cluster sum of squares error of a given distance function [17]. Use of the algorithm has been extended to discrete data and to a mix of discrete and categorical [19, 20] and has been linked to maximum likelihood estimation by Diday and Schroeder[17]. K-means can be considered to by a hard assignment method, where each algorithm is given only a single allocation, an extension of this is Fuzzy c-means, which gives an element membership to any given number of clusters[21], such that this is called soft-partitioning.

Steinbach et-al proposed a method called "Bisecting K-Means" [17] which attempts to apply a hierarchical divisive method to K-Means clustering whereby each partition is recursively split into two clusters at each step. Other methods use AIC or BIC to select an appropriate value for k as the method progresses [22]. K-Modes uses the modal value of the data [19] and is suitable for discrete data, K-Medoids uses the median instead of the means[4]. And a Kernel based K-Means is introduced by Scholkopf [23] which depending upon the choice of kernel made within, is capable of detecting arbitrarily shaped data, in contrast to K-Means which is optimised for spherical data. Other kernel methods [6] such as those using Mercer kernels also are available.

1.3 ART2 Algorithm

ART2 [24] is a partitioning algorithm quite different from K-Means, it is not based upon a variance minimisation approach to clustering, although that is its effect, but is part of the ART (Adaptive Resonance Theory) family of algorithms which attempt to model the way the human brain recognises patterns[25]. The original ART algorithm was designed to model binary inputs but later adaptations both extended the model to continuous inputs and simplified the method to recognise the practicalities of implementation [24, 26].

The ART-2A method is a derivative of ART-2 and models the essential dynamics whilst improving computation efficiency and allowing for dynamic learning rates and prototype growth [26]. It has largely been used in Engineering as a fault detection technique Surprisingly it has been often ignored in favour of other better understood and perhaps unsuitable clustering techniques such as K-Means.

Adaptive resonance theory is a powerful family of algorithms covering ART-1 which is a binary classification algorithm, ART-2 and ART-2A for continuous data and ART-3 for mixed data. These algorithms are used in a broad range of applications a sample of which includes Aerospace and

Engineering [27], Chemical analysis [28, 29], Image Segmentation [30, 31], Fault Diagnostics [32, 33], Sensor Analysis [34] and Marketing analysis [35]. These algorithms are also cited as examples of associative systems in Psychology [36].

Unlike in ART2 [24] for its successor ART-2A there is no constraint on the number of prototypes and it can dynamically increase the number of groups. Importantly once the initial training phase is complete the method can continue to identify new structure and allocate prototypes to new data as it arrives, continually assessing whether the existing prototypes are appropriate choices. This plasticity means that the method is highly suitable for situations in which the underlying structure of the data is unknown or subject to change over time.

ART is known to be dependent upon the presentation order of the data with large differences between the clustering solutions presented, not just in the partitions, but also the number of partitions. The revised algorithm is less affected by presentation order than the standard ART2 model [26] however the dependence on this initialisation remains and consequently this variability of solutions does mean that a strategy is required for ensuring that the solution is both robust and reproducible.

1.4 Initialisation

Initialisation of these algorithms can have as significant an effect upon the clustering as can the algorithm itself and a set of poor initialisations can lead to a particularly poor solution. Both the algorithms listed here have been long acknowledged to provide different solutions depending upon the initial conditions set, unfortunately neither algorithm proposed a standard method by which this shortcoming could be addressed. Subsequent to the original publication different approaches have been proposed by others for K-Means, but in the case of ART little has been published in this regard other than to acknowledge that initialisation is important.

K-Means

A variation of the original algorithm for K-Means [18] introduced a variation which selected the initial starting centres at random. This perhaps, is the most commonly used method of initialising K-Means as Hartigan [10] uses the same method of selecting the prototypes at random from within the data. Milligan found that Ward's method yielded good clustering solutions [37, 53].

What these methods have in common though is that they are not purely initialisation techniques but rather are implemented as part of a larger strategy and fall into being part of a hybridised K-Means tool. Bradley and Fayyad [38] presented a method to find a more refined set of starting conditions which it is hoped converge to a better local optima than using random initialisation alone. Indeed, it is suggested that no good method for initialisation of a clustering algorithm exists, so this method is compared to the random initialisation of Forgy. One of the motivations for this method is that it was designed to be used on a sub-sample of a much larger database and to aid the selection of a good set of initialisations in the absence of a large proportion of the data. Hence it is not possible to extrapolate performance for this method to the use of K-Means with a full dataset.

Kaufman and Rousseeuw [39] proposed a method which attempts to identify centrality within the data, such that selecting a central data point, any additional data points give the best reduction in the objective function.

A set of experimental results from Meilă [40] on Expectation Maximisation algorithms showed however that random initialisation performed worse that the other methods of initialisation tested, given the proximity of EM algorithms to K-Means, it is to be expected that these results carry over to the latter family of algorithms.

An alternative method for initialisation which is an extension of Forgy's method, is that of K-Means++ [41] this more recent development attempts to maximise the distance between starting cohorts, whilst still maintaining some elements of the random initialisation. K-means++ uses a method the authors call D^2 weighting which allows the selection of a new data point with a high probability of not having been allocated to an existing cohort, based on the previously selected prototypes. The presented results seemed to show good performance not just in terms of reduction in the time to terminate but also for minimisation of the objective function.

Redmond and Heneghan [42] presented yet another initialisation technique for K-Means, based on kd-trees, essentially using a divisive hierarchical scheme to partition the data into groups and then looking at the density of the leaves and buckets to consider where to place the prototypes. Essentially the greater the distance between an existing seed and a leaf bucket and the greater the density, the more likely the point is to be a candidate prototype, as it is different and contains many points. This approach again showed promising results in locating a solution with a more optimal solution locally than others.

These initialisation strategies whilst they do attempt to maximise the likelihood of getting closer to the global optima, there is no independent guarantee that the solution that they select is the best representative solution for the underlying structure of the dataset.

ART2

There are two aspects to the initialisation of ART2-A, the first is that like K-Means there are two elements which vary during use. The first is the vigilance parameter, which forms a proxy to the number of prototypes to be created, and whilst the number of representative categories can be determined with fine control of this parameter, in practice this is often pre-determined by prior knowledge [43], which for an exploratory analysis is by definition unavailable. Secondarily the order dependence of this ART2-A algorithm also needs controlling for. It has been shown previously for other order dependent algorithms that order dependence becomes less important with adjustments of a thresholding parameter like the vigilance in ART networks [44] and it would make sense that this would hold for Adaptive Resonance Theory also. No work has been done investigating this phenomenon with Adaptive Resonance Theory however, an astonishing oversight given the dependence on precisely this sort of initialisation that the algorithm has.

1.5 Stability

The use of a stability measure to assess the usefulness of a particular solution rather than attempting to condition the inputs has been shown to find good solutions by Ben-Hur [45]. This approach can take advantage of any seeding method the user desires, whilst still observing benefits in a more stable reproducible solution. Given that any solution is likely to suffer from perturbations and noise in the data, however the approach here looks specifically at the selection of a suitable value for k using sub-sampling and compares to the Gap Statistics of Tibshirani [46] rather than comparing the usefulness of a solution to a given reference partition.

A stability measure for comparing cluster solutions fundamentally different to that of Ben-Hur proposed by Lange [47] uses a measure of dissimilarity based on solutions computed for two datasets. This provides insights into comparing clustering solutions on two similar but disjoint datasets, but does not assist in evaluating the stability of reinitialisations for the same data. Steinley [48] proposed a metric for assessing cluster stability using internal and between cluster co-occurrence, adding a penalty for the fitting of more clusters to the data.

None of these methods however provide guidance for the selection of a particular partition set given repeated reinitialisation of an algorithm. Kuncheva et-al [49] use the adjusted rand index [50] to evaluate consensus partitions and while no significant different was noted through the use of a stability measure it should be noted that the use of consensus partitions already has the effect of reducing perturbations due to noise from the solution sets.

This paper introduces an adaptation of the SeCo framework [51] for ART-2A where the dependence on presentation order is used rather than the initial selection of centres as previously to explore the solution space. The new framework is evaluated using a number of publicly available datasets and one synthetic breast cancer dataset to demonstrate the potential for a health and big data scenario. The results section presents an in-depth review of the synthetic breast cancer data and an overview of the results for the other datasets.

1.6 Estimation of k

Specific to the K-Means algorithm is the necessity of identifying which values of k are particularly interesting or useful, and of those how well they represent the underlying structure of the data. This is a non-trivial task and many competing methods have been proposed for solving this problem with varying degrees of success. Part of the problem is the difficulty of knowing in advance the optimal value for k [7, 46] such that recovery of the data structure is complete. Considering this it is important to take a consistent and principled approach to handling this problem.

Hartigan [52] suggested “The number of clusters **K** should not be decided in advance, but the algorithm should be run with several different values of **K**” however the decision of which value to select still remained. A large study by Milligan [53] looked for a robust way to identify a good number of k and whether there were actually partitions with the data. This study compared a number stopping criteria for usefulness in these measures and found that five of the thirty methods compared provided a good indication as to the correct number of partitions within the data. Surprisingly Trace(W), equivalent to the Within Cluster Sum Of Squares performed poorly when compared with other methods.

More recent methods of determining the appropriate number of clusters for a given dataset include Akaike’s and Bayesian Information Criterion [4, 22], Pelleg proposed a method of incorporating it into X-Means, a K-Means extension, but the method can be used externally to provide the parameters for standard K-Means. A further extension of K-Means uses a Gaussian model to hierarchically increase the number of k until the assumption that all the data in each cohort are Gaussian and this compares favourably to the use of BIC[54]. Tibshirani [46] introduced the Gap statistic to estimate the number of clusters using a reference distribution of the data, however whilst this worked will on highly separated data, on poorly separated data it is not clear that this performed better than any other.

These methods form an algorithmic approach to the determining an appropriate value of k, however other methods exist, such as graphical methods like those suggested by Gierl and Schwanenberg [17], where an appropriate criterion is plotted against multiple values of k and a flattening of the criterion as k increases would indicate an appropriate value of k had been reached. A more formulaic approach such as that suggested by Milligan and Cooper [53] would be to use the Calinski and Harabasz statistic, which uses a ratio of the variance of the pooled and within cluster sum of squares to assist the selection.

1.7 Research Objectives and Novel Contribution

An evaluation of the available literature looking at initialisation dependent clustering algorithms fails to reveal a solution which acceptably manages the expected variation in algorithms such as K-Means when using an objective function to differentiate between solutions with different starting conditions. Furthermore, it will be demonstrated that this use of an objective function as an evaluation criteria for solutions can result in the selection of an unstable or non-reproducible solution and that a better methodology is therefore required.

The development and assessment of such a framework using a two-factor approach based on between cluster separability and within cluster stability is therefore the primary aim of this study. This framework will allow for an easily interpreted and understandable visualisation of the space of solutions which should assist in understanding the underlying structure of the data space and the selection of an appropriate representative solution.

Motivation for seeking a single representative solution stems from the recognition that the temptation when using initialisation dependent algorithms is to use them in a manner which yields sub-optimal results. The inherent variation in such algorithms will produce locally optimal solutions which are only stable in that any two “good” solutions will be proximate to each other in their solution [8]. Empirical results presented here show that the quality of solutions can vary substantially in both cluster separation and consistency from one set of random initialisations to another.

Using an exhaustive search to identify a solution proximal to the global optima [11] might not be reproducible and could even differ from a solution obtained using the same process having different initialisation properties or a similar but separate population sample. This is the case even when considering solutions expected to be well separated. Under such conditions it is to be expected that the use of solutions whose stability is under question should be of considerable concern to practitioners using K-Means to obtain insights in data, which are then used to form policy or drive clinical decisions.

Previous work [55, 56, 57] introduced the use of two measures in identifying good cluster solutions and the concept of a SeCo landscape map of solutions. The novel contribution of this thesis will be to build upon the technical foundation of these papers and to benchmark and evaluate the performance of a generic framework for stabilising initialisation dependent clustering, developing the metrics which are used within the landscape map as appropriate for the algorithm being considered.

The two hypotheses being testing throughout this thesis are as follows:

- The current best practice of reinitialising and selecting the best local optima using total within cluster sum of squares (SSE) (see Equation 2.1 on 22) does not provide a solution which maps in a reproducible manner to the underlying structure of the data.
- Adding a second measure to stabilise the solutions will improve the repeatability of the method whilst ensuring that the resultant partition is representative of the underlying data structure.

The proposed plan of research is to investigate the hypotheses above and to test them to see if they hold and if proven to be so to extend these to additional methods whilst investigating the general case. The key objective of which will be to introduce the idea of a simple framework to visualise the structure of many local minima solutions using two objective measures to represent the between cluster separation and cluster stability.

This will aid the user in deciding when it is appropriate to: stop sampling new random initialisations, retrieve a robust solution and provide a guide to the appropriate number of partitions to investigate and which are most likely to represent the true underlying structure of the data space.

Existing good practice indicates the use of a single metric, usually the Total Within Cluster Sum of Squares (SSE), when determining which solution from a set of repeatedly initialised clusterings should be used. The novelty in this method will be to apply two metrics to the results of initialisation dependent clustering algorithms allowing the separation of results into a landscape map. This map will visualise the solution space and identify stable, reproducible solutions.

It will be shown that this is necessary since the highest values of SSE can correspond to significantly different partitions. This necessitates the use of an appropriate additional metric to resolve this ambiguity and identify a single reference partition that is reliably close to the structure of the data.

Further to this, it will be shown that extending the framework approach to other algorithms than K-Means has a beneficial effect in terms of selecting a stable partition of the data and potential numbers of clusters.

1.8 Summary

With K-Means its extensive use has lead to its inherently unstable nature with respect to initial conditions being well known and algorithmic extensions have been proposed to mitigate this susceptibility to perturbation, this is not the case though for the ART2 algorithms where no such research has been found by the author. Despite research done for K-Means however it is still not possible to entirely eliminate the constraint that for any given solution a more global minima exists. For this reason it is important to move away from considering optimisation of the objective function as the aim for the algorithm but rather that the objective is to obtain a sampled solution which is both considered stable and reproducible, but can also be reasonably expected to be representative of the underlying structure of the data. It is the expectation in this study that using the same principled approach to two fundamentally different algorithms will yield positive results for both.

The following sections of this thesis introduce a framework for stabilising the results of an initialisation dependent clustering algorithm and benchmark it against a number of existing, well known datasets, in addition to a number of synthetically generated data whose properties were known.

Clustering Algorithms 2

There are two components to the methodologies used within this research programme, the clustering algorithms for which investigation has been made, and the proposed framework approach to resolve initialisation dependence problems. The first section of this chapter will focus on the clustering algorithms, and the second half the proposed generic framework algorithm.

The clustering algorithms used are K-Means, a variance based clustering algorithm and ART-2A which uses a choice and resonance mechanism for allocating data to the most appropriate prototype. Both these algorithms are different in their operation and whilst they are initialisation dependent the way in which they are so is dissimilar. Choosing two algorithms with different initialisation dependencies allows for testing whether the framework proposed in this thesis has value not just for a given method but more generally.

2.1 K-means

The first algorithm is the K-Means algorithm which was originally proposed by James MacQueen [9] as a method for partitioning N-dimensional data into a number of smaller sets, however perhaps the earliest formulation was in 1950 by Dalenius, and Dalenius and Gurney [15, 16]. This algorithm was used as a competitive learning algorithm in signal processing to model variance, and was later adapted for use as a clustering algorithm. It is a surprisingly simple algorithm which iterates over an input matrix M minimizing an objective function until the matrix of prototypes P becomes stable, this objective function is the called Total Within Cluster Sum of Squared Error (SSE) and is the sum of the squared errors for each of n data-points allocated to each of the given k partitions. The definition is shown in Equation 2.1.

$$\underset{s}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2 \tag{2.1}$$

Where μ_i is the mean of the points in S_i, the data points assigned to the cluster, and $||x_j - \mu_i||^2$ is the distance between an element and the centre of the cohort to which it has been assigned. The original algorithm as defined by Lloyd [14] can be described as below in Algorithm 1 and performs a Voronoi decomposition of the data space where each partition has roughly equivalent shape and volume.

Input X A matrix containing the data
k the number of cohorts into which the data should be grouped

Output P a vector containing the assigned cohorts
C a matrix containing the centres for each cohort

Algorithm 1 Outline of K-Means algorithm

```
Select $k$ prototypes at random from data into $C$
for all $n_i \in N$ do
    Calculate distances from all members of $C$
end for
repeat
    for all $c_i \in C$ do
        Allocate data to $c_i$ if closest prototype
    end for
    Calculate new means for $C$
    for all $n_j \in N$ do
        Calculate distances from all members of $C$
        Allocate $n_j$ to correct prototype
    end for
until No Changes Made Since Last Iteration
```

Given a set of clusters whose centres are well separated and having non-overlapping data points the algorithm will readily find the appropriate partition for the data, however when cohorts overlap or whose shape in higher dimensions has non-uniform variance then the algorithm performs less well. Whilst the algorithm has been shown to locate a local minima it is also known that the algorithm does not necessarily attain a global minimum [11, 52, 58]. Even then the minimum cannot be verified as being the best global minimum. Different methods have been proposed to start the algorithm to avoid local minima; however given the supposition that if a particular solution is a local optima, there is another solution with a lower SSE [12].

It has been shown[12] that even for datasets whose size is not great that the set of local optima is large, and that there is no objective way to evaluate whether a particular solution is globally or locally optimal[12]. This confirms prior hypotheses[52], for which the proscribed solution was to perform many iterations using a discriminating value to determine the best solution, usually the SSE of the partition[7, 52].

Hartigan [52] proposed an updated algorithm using a second phase where online updates occur, comparing each point with the nearest cluster centre to determine if moving the data point to a different centre reduces the total SSE and if this is the case then this point is moved. The Hartigan algorithm is the one which is used whenever the K-Means algorithm is mentioned henceforth without additional qualification.

Input X A matrix containing the data
k the number of cohorts into which the data should be grouped

Output P a vector containing the assigned cohorts
C a matrix containing the centres for each cohort

The difference between Algorithm 1 and Algorithm 2 is that the latter produces solutions with reduced variance of SSE, because repeated initialisations with the same data are likely to produce similar SSE values, so the second algorithm has greater likelihood of hitting a local minima close to the global, although solutions with varying SSE are expected.

As mentioned previously the most common method for selection of a best partition of the results is to repeat the K-Means algorithm multiple times (although the exact number of repetitions is left unspecified) and select the partion whose corresponding SSE is the lowest.

Algorithm 2 K-Means Hartigan-Wong algorithm AS-136[10]

```
Select $k$ prototypes at random from data into $C$
for all $n_i \in N$ do
    Calculate distances from all members of $C$
end for
repeat                                            ▷ Calculate Optimal Transfer
    for all $c_i \in C$ do
        Allocate data to $c_i$ if closest prototype
    end for
    Calculate new means for $C$
    for all $n_j \in N$ do
        Calculate distances from all members of $C$
        Allocate $n_j$ to correct prototype
    end for
                                                  ▷ Calculate Quick Transfer
    for all $n_j \in N$ do
        Check total SSE (T1) if move $n_j$ to alternative cluster (T2)
        if $T1 > T2$ then
            Move point to alternate cluster
        end if
    end for
until No Changes Made Since Last Iteration
```

2.2 ART 2

The second clustering algorithm used is from the Adaptive Resonance Theory (ART) [25] family of algorithms, which are intended to be used in pattern recognition with the intention being to provide a form of model for the cognitive and neural behaviour of the human brain. These algorithms work by not holding any initial prototypes but using a vigilance parameter to control whether newly presented data is given a new prototype or allocated to an existing one.

Initially developed for binary data [24] the first algorithm (usually identified as ART 1) was adapted for continuous inputs by the original authors as ART 2 [24] these algorithms were originally intended for implementation in electronic circuits and used a fixed maximum number of hidden nodes which could represent category prototypes. This last algorithm was later adapted for efficient implementation in software as ART 2-A which in addition to being faster also no longer has the limitation of a fixed number of prototypes.

There are two methods for ART-2A described, the first is the standard ART2-A method proposed by Carpenter and Grossberg [26] followed by a variation on ART-2A [59], both of which produce an output set of two cluster indices for each row M_i and $m \times n$ matrix of observations M and have the stability-plasticity characteristics fundamental to ART-2. A filter is applied to these solutions where prototypes having less than a proportion r of n observations assigned will be discarded and the data points allocated to the nearest alternative prototype. This has the effect of stabilising the number of prototypes produced for a given vigilance parameter ρ over multiple re-initialisations and thus giving fewer but better populated indices. These results are used in the SeCo framework method described here to identify stable and reproducible solutions.

ART 2-A

This original implementation of ART-2A is a derivative of ART-2 which approximates the dynamic nature of the original whilst improving the computational efficiency by an order of magnitude.

The ART2-A algorithm takes as its input a matrix M with m observations of n features, in addition to three control parameters ρ the vigilance parameter, α a threshold variable satisfying the constraint (2.6) and β which controls the learning rate of the method. A fourth parameter θ is referenced in the paper as a means of suppressing noisy signals in the data, but is unused here. The matrix P represents the prototypes and is initialised as an empty set and populated as the method progresses.

Each element of the matrix M is presented to the algorithm in sequence as a non-uniform vector I_0, and transformed with each step being followed for each presentation of the vector to the algorithm over a given number of iterations. Prototypes are dynamically updated throughout using the learning function.

Presentation

$$I = \mathcal{N}\mathcal{F}_0\mathcal{N}I_0 \tag{2.2}$$

where

$$\mathcal{N}x \equiv \frac{x}{||x||} \tag{2.3}$$

and

$$(\mathcal{F}_0 x)_i \equiv \begin{cases} x_i & \text{if } x \geq \theta \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

subject to the constraint

$$0 < \theta \leq \frac{1}{\sqrt{m}} \tag{2.5}$$

where eqs. (2.2) to (2.5) imply that I is nonzero and normalised to unit euclidean length.

$$0 < \alpha \leq \frac{1}{\sqrt{m}} \tag{2.6}$$

Choice and Resonance

The input vector I is compared to each prototype p_j in P using the function (2.7). ART type functions use what are called committed and uncommitted prototypes, where a committed prototype is one which has been previously created, and an uncommitted one is one which hasn't been selected yet. When an uncommitted prototype is selected, it creates a new category to which subsequent presentations of input vectors can be compared.

$$T_i = \begin{cases} \alpha \sum_i I_i & \text{when } p_j \textit{uncommitted} \\ I \cdot p_j & \text{when } p_j \textit{committed} \end{cases} \tag{2.7}$$

To start all the nodes are uncommitted, and the first data point presented forms the first prototype, meaning that the result is dependent upon the order in which data is presented. For each T_j calculated the maximal value is selected, and this becomes a committed node. This node j is tested (2.8) with constraints as in (2.9) if it is a previously committed node, if uncommitted, then the choice remains.

$$T_j \geq \rho^* \tag{2.8}$$

$$0 \leq \rho^* \leq 1 \tag{2.9}$$

When T_j is a previously committed node, the value T_j is equivalent to the cosine distance between the two vectors.

If the selected prototype does not pass the constraint in (2.9) then the value j is reset to the value of an uncommitted prototype.

Learning

Once a prototype has been selected by the above method, then the prototype is updated depending on whether it is a committed or uncommitted node. If a previously committed node then the update is performed as in (2.10) otherwise is it set as $P_j = I$.

$$vP_j^{new} = \mathcal{N}((1-\beta)P_j^{old} + \beta\mathcal{N}I) \tag{2.10}$$

with

$$0 \leq \beta \leq 1 \tag{2.11}$$

It is important to note that the choice of learning parameter is significant as the extreme of each will result in different behaviour. Setting β to 1 will result in ART-2A having fast-learn properties whereas the other extreme will result in learning more like a leader algorithm, with the prototype remaining fixed after commitment. Small values of β will result in a slow-learning rate.

This method will produce varying numbers of prototypes with varying separation for each reinitialisation with the same given parameters, meaning some decision on which solution is most appropriate is necessary.

ART 2-A Variant (Gallant)

The second method for ART-2A [59] uses the same normalisation functions as does the first with L2 Normalisation occurring through eqs. (2.2) to (2.4) differing in the way choice and resonance occur. The use of the parameters (2.12)is also slightly different from the originally proposed method [26] with the α parameter having a slightly different purpose and the β parameter having a changed constraint.

As with the original algorithm, P, which contains the prototype vectors are initialised as empty and populated over time, and the vector I represents the normalised pattern under presentation.

Parameters

$$\begin{array}{ll} \alpha & \text{positive number } \leq \frac{1}{\sqrt{m}} \\ \beta & \text{small positive number} \\ \theta & \text{normalisation parameter, having } 0 < \theta \leq \frac{1}{\sqrt{m}} \\ \rho & \text{vigilance parameter, having } 0 < \rho \leq 1 \end{array} \tag{2.12}$$

Choice and Resonance

Choice of prototype is determined by finding the prototype vector P_j which maximises (2.13), as before any ties at this point are solved arbitrarily.

$$P_j \cdot I \tag{2.13}$$

Following this, P_j is tested to see if it is sufficiently similar to the presented data point (2.14), if it fails this test, then a new prototype is created as before with P_{new} set to I.

$$P_j \cdot I \geq \alpha \sum_i I_i \tag{2.14}$$

If the prototype is similar then it is tested to ensure that the value is greater than the vigilance parameter $P_j \cdot I \geq \rho$, if failing then a new prototype is again created. If the prototype matches then the method proceeds to the learning stage.

Learning

During the learning stage, the prototype is updated to be a combination of the existing prototype and the presented data vector, using β to control the update rate (2.15).

$$P_j = \frac{(1-\beta)P_j + \beta I}{||(1-\beta)P_j + \beta I||} \tag{2.15}$$

This second method has a similar result to the original method, however it also has the benefit that the α parameter provides additional controls over the sparsity of the solution, enforcing a minimum cosine distance below which a new prototype is created, even if an otherwise matching prototype exists.

Differences

The two methods indicated here for ART-2A based clustering are similar in that they both use the Cosine Similarity measure (defined as $\frac{A \cdot B}{\|A\|\|B\|}$, where A and B are unit vectors) to identify how similar partitions are to existing prototypes and determine which is the most appropriate match. The main divergence is the use of the α parameter between each, the original ART-2A uses the alpha parameter in combination with the cosine similarity measure to determine whether the best match is a new prototype or an existing one, whereas the Gallant variation applies a test first to see whether a new prototype is the good option, and if so creates one rather than continuing to the match and resonance step. This latter approach produces a sparser clustering with more numerous prototypes than the first. Both will produce prototypes when ρ is set to zero.

3 Framework Methodology

Proposed here is a stabilisation framework for use with clustering algorithms which have properties such that the output is dependent upon some form of initial condition, although the specifics of implementation may vary with each algorithm. There are two basic components to the framework method, multiple reinitialisation, which depends on how the algorithm is initialisation dependent and solution choice. The framework allows for the selection of a clustering solution which can be considered to be robust and reproducible by producing a landscape map, called the SeCo map of the solution space which can be used to aid in the interpretation of the results.

An important consideration is that for repeated multiple initialisations it is known that some of the partitions will be less well separated than others, and the interest at this point is for well-separated reproducible solutions. When calculating the concordance values used in the map the top 10% of solutions will be selected and then the concordance measure applied pairwise to all these solutions and from these the median concordance is chosen as the most representative concordance value for that solution.

The generic algorithm can be seen in Algorithm 3, where N is the set of variable initialisation parameters and S the top 10% of solutions by separation.

Input X A matrix containing the data
P The initialisation parameters

Output P a matrix containing the assigned cohorts for each solution
K a vector containing the number of prototypes generated for each solution

C a matrix containing all the prototypes for each cohort in each solution

Algorithm 3 Generic framework algorithm

```
for all $n_i \in N$ do
    Select reinitialisation parameters
    Perform clustering algorithm
    Calculate Separation value
end for
Select top 10% of solutions by separation
for all $s_i in S$ do
    Pairwise find the concordance with all other $s_i$
    Calculate the median value of these
end for
```

This gives each of the top 10% of solutions a pair of associated values, which can be plotted on y-x axes (Separation-Concordance) to produce the SeCo map.

This framework therefore uses both an internal measure of separation, and an external measure of cluster stability. A filter is applied to the separation measure to filter out poorly performing solutions; a stability measure is then calculated pairwise for each of the solutions. The median value for each solution is then taken. These two values for each solution form an (x, y) coordinate pair, and can be used to produce a SeCo map of the solution space.

The results from the framework can then be used to evaluate the overall performance of each of these solutions with respect to a particular value of *k* and highlight solutions for which good structure is consistently recovered. Solutions where the recovery of partitions are repeatable will have a high proportion of solutions with greater consistency.

Two clustering algorithms are used during the evaluation of the framework, K-Means and ART-2A, the former being the primary algorithm used throughout and the latter being used to test whether the proposed framework can be extended in a generic fashion. K-Means is a well known and understood algorithm, and if it can be shown that there is a benefit to the SeCo framework for this initialisation dependent clustering method over other methods then this is important. Also it was the variability of solutions from K-Means which provided the impetus for the research, and to the development of the SeCo concept.

ART-2A was selected as the second clustering algorithm for a number of reasons, firstly it is very different from K-Means, not just in the way it operates, but also in the way it is initialised. Because it has incomparable parameters and the random effects are different it made sense to use such an algorithm rather than one similar to K-Means as this is a better test of the generalisability of the SeCo framework.

As discussed above the primary algorithm for evaluation is K-Means, which provided the original motivation for the study as it formed the basis of an

Concordance

There are several indices of agreement between pairs of cluster labels such as cosine similarity and the Jaccard coefficient [45]. In the statistical literature there are also inter-rater agreement measures for known labels such as Cohen's Kappa index. In order to avoid the need for an oracle to set a nominally correct number of clusters [49] a generic index of association, of concordance, must apply to data partitions whose inherent labels are not known in advance, should be normalized and not strongly dependent on the marginal frequencies in each cluster partition [60] and preferably apply to comparisons between partitions with different numbers of clusters.

A suitable measure is Cramer's V-index of concordance [61]. This is a statistical test score to measure the strength of association between two data partitions of the same data set. For a cross-tabulation of n observations representing a partition into p rows and another as q columns, treated as a contingency table with expected entries E_{pq} for independent cluster allocations and observed values O_{pq} , the extent to which one cluster partition predicts the other (i.e. the association between them) is measured by

$$C_v = \sqrt{\frac{\chi^2}{N \times \min(P-1, Q-1)}} \tag{3.1}$$

where,

$$\chi^2 = \sum_{p=1}^{P} \sum_{q=1}^{Q} \frac{(O_{pq} - E_{pq})^2}{E_{pq}} \tag{3.2}$$

An alternative statistical measure of agreement between two partition sets, and also considered, is the Adjusted Rand Index of Hubert and Arabie (ARI-HA) [50]. The measure was adjusted to avoid over inflation due to correspondence between two partitions arising from chance. The Cramér's V-index and ARI-HA have a Pearson correlation coefficient of 0.99, higher than the value 0.95 between the Cramér's V-index and both the unadjusted ARI of Morey and Agresti and the Jaccard coefficient [50].

The Jaccard similarity coefficient where x and y are two positive real vectors, is defined as

$$J(x,y) = \frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)} \tag{3.3}$$

and the adjusted rand index where X is a contingency table, and n_{ij}, a_i and b_j are elements from this table, is defined as

$$ARI(X) = \frac{\sum_{ij}(\frac{n_i j}{2}) - \frac{\sum_i(\frac{a_i}{2})\sum_j(\frac{b_j}{2})}{\frac{n}{2}}}{\frac{1}{2}[\sum_i(\frac{a_i}{2}) + \sum_j(\frac{b_j}{2})] - \frac{\sum_i(\frac{a_i}{2})\sum_j(\frac{b_j}{2})}{\frac{n}{2}}} \tag{3.4}$$

This shows that the two statistical indices are closely related, though not identical, with even better correlation for improved correspondence between partitions.

3.1 Framework for K-Means

When using K-Means the joint optimisation of within-cluster separation and between-cluster stability requires suitable indices to measure these properties. In principle any reasonable performance measures may be applied in the proposed framework, however for this particular algorithm it makes sense to use the Total Within Cluster Sum of Squares Error (SSE) defined in Equation 2.1.

There are two variables to consider when looking at K-Means as an intialisation dependent problem, the number of k to partition the data into, and the initial prototypes from which to start the algorithm. In the case of the former there are methods which attempt to determine the optimal number of k and will be looked at in a later section so are not considered here except to state that they often require multiple reinitialisations to assist in determining where the choice should be made. Given that repeated initialisations are being performed here, the method will be applied directly to many different k. For the initial prototypes methods have been developed which attempt to produce good representative starting parameters such as kmeans++ [41] however no independent criteria are used to evaluate the performance of this initialisation, so they do not solve the problem being attempted here.

Therefore, for a given data set and assumed number of clusters k, the following methodology is proposed:

1. Apply the cluster partition algorithm to a sample of size N_{total} of cluster initialisations, each seeded with k randomly selected points sampled from the full data set i.e. the standard initialisation for K-Means

2. Sort by separation score and select a fraction f by ranked score of ΔSSE, defined as the difference between Total Sum of Squares and the Within Cluster Sum of Squares for a particular solution, returning a working sample of cluster partitions $N_{sample} = N_{total} \times f$ in number

3. Calculate the $N_{sample} \times (N_{sample} - 1)/2$ pair wise concordance indices C_v for the selected cluster partitions and return the median value $med(CV)$ of all pair wise concordance indices for each partition

4. The Separation and Concordance (SeCo) map comprises the 2-dimensional coordinates (ΔSSE, $med(CV)$) for the selected cluster partitions.

5. Once the landscape of cluster partitions has been mapped using the SeCo map, where there is a spread of solutions with similar ΔSSE, choose the solutions with the highest value of $med(CV)$.

Step 2 within the framework process is important to understand and forms part of the selection process for the framework. By selecting the top fraction of results and performing the concordance check on these only, the evaluation is done on those solutions for which good separation has been achieved. This is important because by leaving in the less well separated solutions it is likely that the concordance values for the well separated values would be poorer, and the likelihood is that a less good solution would be selected.

As the assumed number of partitions k is increased the map generates a scatter of points with increasing ΔSSE but with a distribution of $med(CV)$. This shows the stability of each assumed number of clusters when fitting the data structure within the constraints of the particular clustering algorithm.

In each case here the number of reinitialisations is taken to be 500 and through experimentation it is demonstrated that only the top decile by separation need be retained. This results in a working sample size $N_{sample} = 50$ for each value of the assumed cluster number k for general purposes. These parameters can be varied however the total sample size must be sufficient for clear structure to emerge among the cluster solutions in the SeCo map while retaining a small enough fraction of the total initialisations to prevent the map becoming cluttered.

Increasing the number of initialisations in each step adds a computational penalty for a given threshold, as for each result from an initialisation the pairwise concordance with all other results selected. In practice 500 initialisations seems to be a reasonable balance between allowing the cluster structure to appear and computation time. If a particular dataset does not result in structure emerging clearly from the results, then increasing the number of initialisations may help solve this.

3.2 Framework for ART 2-A

For the ART 2-A algorithm the stabilisation framework is different from that proposed for K-Means as the way in which it is initialisation dependent is somewhat different. Where K-Means has a varying control parameter to select the number of partitions and then randomly selects the starting centres from the data, ART-2A has a control parameter which provides a threshold for Similarity and is then dependent upon order presentation.

As before the separation measure chosen here is the total within cluster sum of squares (SSE) (2.1) and is used in combination with Cramers' V statistic [61], a statistical score measuring the strength of the association between discrete variables shown in equation (3.6). As with K-Means any suitable score could be used however.

$$\chi^2 = \sum_{p=1}^{P} \sum_{q=1}^{Q} \frac{(O_{pq} - E_{pq})^2}{E_{pq}} \tag{3.5}$$

$$C_v = \sqrt{\frac{\chi^2}{N \times \min(P-1, Q-1)}} \tag{3.6}$$

For a given dataset the following process is proposed for use with either of the ART 2-A algorithms:

1. Perform ART 2-A algorithm
2. Calculate the within Cluster Sum of Squares
3. Repeat Steps 1–2 until sufficient number of solutions for each generated number of prototypes
4. Select the top 10% of these solutions by the Within Cluster Sum of Squares.
5. Calculate the stability measure, the pairwise Cramers' V statistic for each combination of these solutions.
6. Calculate the median within group Cramers' V statistic for each solution
7. Produce the Separation Concordance (SeCo) map of these solutions.

As with K-Means the selection of the top 10% of solutions is important as it ensures that only well separated solutions are considered when looking at those results which are stable. Step 3 requires waiting until a sufficient number of solutions is found for each different prototype number group, so n=3, n=4 and n=5 prototypes for example. Unlike the value of k in the K-Means algorithm the vigilance parameter for ART 2-A does not directly control the number of partitions within the data, rather it is also influence by the order of presentation for the data. To account for this, the vigilance parameter can be varied for a given dataset and the algorithm is repeated for each of these variations until at least a minimum threshold number has been achieved for that algorithm, but also until a minimum number of solutions has been calculated for each set of solutions with a given number of partitions. As has been previously used with the K-Means algorithm, this was set at 500 for each here.

The pairwise Cramers' V concordance of each of these solutions is calculated within groups of solutions identified by having the same number of prototypes, and the median of these is taken as a representative value of the overall stability of the solution.

Finally the SeCo map is generated as a visualisation of the relative performance of the different groups allowing the user to gauge performance in selecting an appropriate solution for use, the separation of solutions is on the y-axis and the internal concordance on the x-axis. Previous results show that choosing solutions towards the right-most edge is the most effective strategy in obtaining stable, reproducible solutions [51].

3.3 Number of Partitions

The use of a SeCo map to obtain inference about an optimal solution requires that some decision about the appropriate number of partitions to be found within the data. Whilst the SeCo map can provide this information there is no score attached to this information and it is often useful for this information to be presented.

With K-means there are existing methods which can be used determining which is the appropriate number of solutions for a given dataset but these are not necessarily applicable to ART 2-A or other clustering algorithms, and some such as the Gap Statistic [46] use the K-Means algorithm as part of the identification process. For this reason information from the framework has been used to develop a measure for inferring which numbers of partitions are of interest for a given dataset. This measure does not specify a particular value, but rather is intended as a further tool for exploratory data analysis. The measure developed here uses the Integrated Cumulative Cramer V (ICC) of the total stability distribution to determine which of the different numbers of partitions are of interest.

Existing measures for K-Means

For K-Means there are already a number of existing methods for obtaining an inference about the appropriate choice of k and these are detailed below along with some of their pitfalls.

Gap Statistic [46] describes the reduction of the objective function in relation to the expected reduction for a particular value of k and uses a reference distribution estimate to calculate this. Unlike the ICC method this method assumes an underlying structure to the data and bases its results upon the difference between clustering of the reference distribution and the data, not only this but it relies upon K-Means as a component of the calculation and the internal stability of the clustering results is not evaluated.

CLEST [62] uses a prediction based resampling technique to estimate the appropriate number of clusters within the dataset. It does this by dividing the dataset into non-overlapping cohorts (test and train) and iterating through values of k and comparing the predicted labels to those of the training set. It then evaluates these and selects the value of k for which the greatest evidence indicates this best solution. This model uses a null-distribution based upon the uniform distribution in the same manner as the Gap statistic meaning that the results are dependent upon the structure of the data and in cases where the data is non-uniform or non-normal then the method may not perform as well.

Calinski-Harabasz [63] is a pseudo f-statistic measuring

$$CH = \frac{trace(S_p)}{trace(S_w)} \cdot \frac{n-1}{n-K} \tag{3.7}$$

Where n is the number of objects, k the number of cohorts, S_B the between cluster sum of squares and S_W the within cluster sum of squares. The value for k is taken as the solution which maximizes this value. Previous studies have shown this performs well [53] however given that this measure is effectively a one-way ANOVA assumptions of normality and independence hold.

Davies-Bouldin [53] is an internal validity measure of the clustering and is calculated as follows

$$DB = \frac{1}{K}\sum_{i=1}^{k} \min_{j=1...K;i\neq K} \left(\frac{d_i + d_j}{d(c_i, c_j)} \right) \tag{3.8}$$

where k is the number of cohorts, i and j are cluster labels, where d_i and d_j are the respective average distances for each of all objects in each cluster from the centre and $d(c_i, c_j)$ is the distance between these centres. Minimising the value to select the appropriate value of k. This measure uses the data to determine whether a particular solution has cohorts that are both compact and well separated and makes the assumption that the solution which best matches these criteria is the most suitable choice of k.

X-Means [22] is an iterative approach attempting to efficiently search for an appropriate value of k. It does this by attempting to determine where and when a cluster should split, by iteratively allowing the split of each cohort and evaluating the change in performance of the model. It does this by using K-Means to model the cohorts and adding new centres close to existing centres and re-running. If the performance of the model is improved, then this new centre is accepted and the process continues. X-Means relies upon an information criterion (in the case of the paper, BIC) for determining the split or not of a cohort, and assumes an underlying spherical Gaussian structure to the data when determining the maximum likelihood estimate.

Dunn [64] defines an index measuring good separation between clusters and maximising this value allows for inference as to an appropriate value of K. Given that $d(c_i, c_j)$ defines the inter-cluster distance between cluster X_i and X_j and $d(X_m)$ represents the intra-cluster distance the Dunn Index can be computed as below.

$$DI = \min_{i\in 1...K} \left(\min \left(\frac{d(c_i, c_j)}{\max_{m\in 1...K} d(X_m)} \right) \right) \tag{3.9}$$

Silhouette [39] is an index $s(i)$ for each data point, and is a measure of the difference in dissimilarity between the data point and those in the same cluster, compared to those in the nearest cluster, where $b(i)$ is the dissimilarity for of i to the points in the nearest cluster and $a(i)$ is the dissimilarity between the point and those in the same cluster.

> Kaufman and Rousseeuw postulated that by taking the average index across all the data points, for a given value of k and maximising this value, the appropriate value of k can be estimated.
>
> $$s(i) = \frac{b(i) - a(i)}{\max a(i), b(i)} \tag{3.10}$$

The Silhouette, Dunn Index and Davies-Bouldin measure are both internal validity measures which look at the separation of the clusters to assist in the determination of which value of k is most suitable and are dependent upon the structure of the data. Where the data is heavily mixed and/or non-Gaussian looking only at the separation of the partitions will cause these solutions to perform less well.

Integrated Cumulative Cramer V

The SeCo map details the stability and separation of the solutions found for the dataset and the specified values of k. Producing a plot of the proportion of solutions with a given consistency allows the user to gauge the relative performance of the solutions for a given value of K. Further calculating the Integrated Cumulative Cramer V (ICC) gives a metric which can be used to directly compare one against the other.

Calculating the ICC value is as follows:

Algorithm 4 ICC Calculation steps

```
  Calculate the cumulative distribution of the Cramérs' V values for a given value of $k$
  Determine the area under the cumulative distribution using Riemann's method.
  Locate minimum ICC value and determine threshold.
  Select highest value of $k$ for which values of ICC are lower than the upper threshold.
```

The ICC measure shown here uses Riemann integration although any suitable method may be used to calculate the area under the cumulative Cramérs' V distribution for the particular value of k. Lower values of ICC indicate solutions with greater stability with smaller values of k more likely to produce low scores. The intent therefore is to identify higher values of k having low values of ICC; yet selecting the lowest value might not necessarily indicate an optimal solution. The SeCo map shows us that increasing k leads to a reduction in stability even amongst highly separated solutions as data points shift cohorts.

This measure can identify a single value for the number of partitions using the natural variation within the distribution of concordance. To do this an upper threshold of the minimum + 10% ICC is used such that all values falling below this upper bound are now considered. Because the method builds upon the SeCo framework the solution selected is already well separated, thus representing each value of k by its ICC we are able to select the highest value for which the partitions themselves are stable. Doing this provides a model free way of identifying a value for the number of partitions such that the underlying distribution of the data and the shape of the clusters are not a factor in the determination of the structure of the data.

4 Data Description

Adequate benchmarking and testing of the framework developed here requires that testing is implemented against many different datasets to properly evaluate performance. Several publicly available real world datasets whose properties are well known are used in conjunction with synthetic datasets derived from a breast cancer dataset.

The synthetic data is used in several forms with the number of observations in each being adjusted to simulate varying sample density. The sampled data comprised 1076 data points in three dimensions, sampled randomly from a mixture of ten multivariate normal distributions. Other than these the individual datasets are summarised below.

Olive Oil The olive oil dataset comprises 576 samples of a chemometric analysis of Olive Oils [65] with 8 attributes for fatty acids, specifically palmitic, palmitoleic, stearic, oleic, linoleic, arachidic and eicosenoic acid. Two classifications are possible with this dataset, corresponding to the three regions in which the olives were grown, further subdivided into 9 areas.

Iris The Iris dataset is a well known multivariate dataset first introduced by R. Fisher as an example used in discriminant analysis. The dataset consists of 150 samples comprising 4 measurements of various species of Iris, Setosa, Virginica and Versicolor. The attributes consist of the length and width of the Petals and Sepals of each flower.

Wine The wine dataset is from the results of chemical analysis of wines grown in a region of Italy, and consists of 178 observations with 13 attributes for each. The attributes are Alcohol, Malic Acid, Ash, Alcalinity of the Ash, Magnesium, Total Phenols, Flavonoids, Nonflavonoid Phenols, Proanthocyanins, Color Intensity, Hue, OD280 or OD315 of diluted wines and Proline. This dataset is well known and found in the UCI data repository.

Cardiotocography Results from 2126 foetal cardiotocograms with 23 attributes pre-classified by obstetricians into both foetal state (N, S, P) and morphologic pattern (A,B,C,...) [66, 67]. The 23 attributes were automatically processed and respective diagnostic features measured.

Thyroid Data A dataset looking at thyroid function[66, 68] containing 215 observations with 5 features - T3resin, Thyroxin, Triiodothyronine, Thyroidstimulating, TSH_value - with three target classes, Normal, Hyperthyroidism and Hypothyroidism.

The synthetic datasets were used for the purpose of rigorously benchmarking the proposed methodology, the other datasets were intended to validate those results separately.

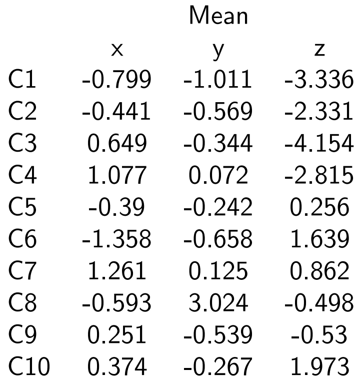
The following sections of this chapter are intended to describe the properties and structure of each of these datasets.

4.1 Synthetic Data

The synthetic breast cancer dataset was generated from an existing breast cancer dataset by randomly sampling from ten different three dimensional Gaussians. The mean and covariance matrices of these are represented in Table 4.1 and Table 4.2 respectively. These parameter settings were chosen to reflect a real world dataset [69] having separated and contiguous cohorts with varying inter mixture. The variables within the dataset were derived from the three principle separating axes[70] from the clustering of a previously studied real-world breast cancer dataset.

Trivial datasets such as Ruspini [71] do not provide a challenge for a reasonably good clustering algorithms due to their well separated cohorts within the data space. Because of this there is a need for a dataset for which an algorithm should not be able to recover the underlying structure of the data due to intermixing of the cohorts. A dataset such as this provides a significant challenge to the algorithm and consequently is a good tool with which to benchmark the algorithm.

In order to properly evaluate the performance of the framework in relation to existing practice, it is necessary to use a dataset for which the true underlying structure is known. For this purpose, a synthetic dataset was generated, whose parameters reflected a real world dataset[69].



		Mean	
	x	y	z
C1	-0.799	-1.011	-3.336
C2	-0.441	-0.569	-2.331
C3	0.649	-0.344	-4.154
C4	1.077	0.072	-2.815
C5	-0.39	-0.242	0.256
C6	-1.358	-0.658	1.639
C7	1.261	0.125	0.862
C8	-0.593	3.024	-0.498
C9	0.251	-0.539	-0.53
C10	0.374	-0.267	1.973

Table 4.1: Generating means for artifical data

Resampling the data multiple times with varying densities in the clusters has the property of producing multiple datasets whose ease of clustering varies for the purpose of testing in- and out-of-sample performance. The initial data produced comprised 1076 observations in three dimensions (x,y,z) with 10 individual cohorts; it is this data that is referred to through out as the Synthetic data, the other datasets are referred specifically as variations of the synthetic data. This dataset forms the cornerstone of the experimental phase as it ensures that the task of separating the data into new partitions is non-trivial as this would render any improvement for the proposed algorithm meaningless if not set within this context.

Visualising the synthetic data using the principle separation axes [70] it is apparent that there is a single well separated cohort (Cluster 8 on this

					Covariance Matrix (i,j)					
	11	12	13	21	22	23	31	32	33	N
C1	0.336	0.044	0.074	0.044	0.371	0.21	0.074	0.21	0.582	64
C2	0.428	0.06	-0.002	0.06	0.123	0.157	-0.002	0.157	0.648	42
C3	0.62	0.023	-0.035	0.023	0.137	0.07	-0.035	0.07	0.446	61
C4	0.366	-0.002	0.076	-0.002	0.043	0.104	0.076	0.104	0.563	32
C5	0.536	0.013	0.031	0.013	0.348	-0.117	0.031	-0.117	0.689	197
C6	0.309	-0.06	-0.055	-0.06	0.245	-0.013	-0.055	-0.013	0.532	131
C7	0.323	0.017	0.027	0.017	0.386	-0.06	0.027	-0.06	0.403	163
C8	0.776	0.033	0.175	0.033	0.491	0.003	0.175	0.003	0.695	97
C9	0.711	-0.025	0.055	-0.025	0.352	-0.081	0.055	-0.081	0.576	106
C10	0.39	-0.097	0.041	-0.097	0.343	-0.014	0.041	-0.014	0.322	183

Table 4.2: Covariance Matrix for artifical data

visualisation), several dense cohorts with varying degrees of inter mixture, and below them, several sparse cohorts, also mixed Figure 4.1. The level of mixture within each of the cohorts is highlighted in Table 4.3, which shows how well separated each generating centre is from the others. Higher values indicate that the centres of the cohorts are further apart, and there is less mixture.

A multidimensional scaling can be performed to visualise these better and can be seen in Figure 4.2 and shows a strong intermixing of the different cohorts; this would imply that it is unlikely that the true data structure would be recovered using a distance based partitioning method such as K-Means. The visualisation therefore indicates that it is important to measure the performance of the framework against that which is actually achievable rather than the underlying structure.

	C1	C2	C3	C4	C5	C6	C7	C8	C9
C2	**0.7805**								
C3	**1.2105**	1.4828							
C4	1.5054	**1.1924**	**1.0687**						
C5	2.4975	1.7636	3.0649	2.3119					
C6	3.3913	2.8294	4.476	3.8029	**1.1757**				
C7	3.2516	2.5575	3.7002	2.7302	**1.2151**	2.2233			
C8	2.9776	2.4341	3.0901	2.4774	2.025	2.6082	2.2314		
C9	2.0388	**1.2969**	2.4543	1.6846	**0.7109**	1.8176	**1.2393**	2.2086	
C10	3.7087	3.0487	4.4727	3.5977	1.2717	1.4141	**1.233**	2.5497	1.6952

Table 4.3: Pairwise indices of c-separation for the artificial data. Coefficients smaller than 1.30 are highlighted in boldface, those around unity or less underlined.

Following this train of thought to its logical conclusion it can be deduced that there is little point in benchmarking the results against an unachiev-

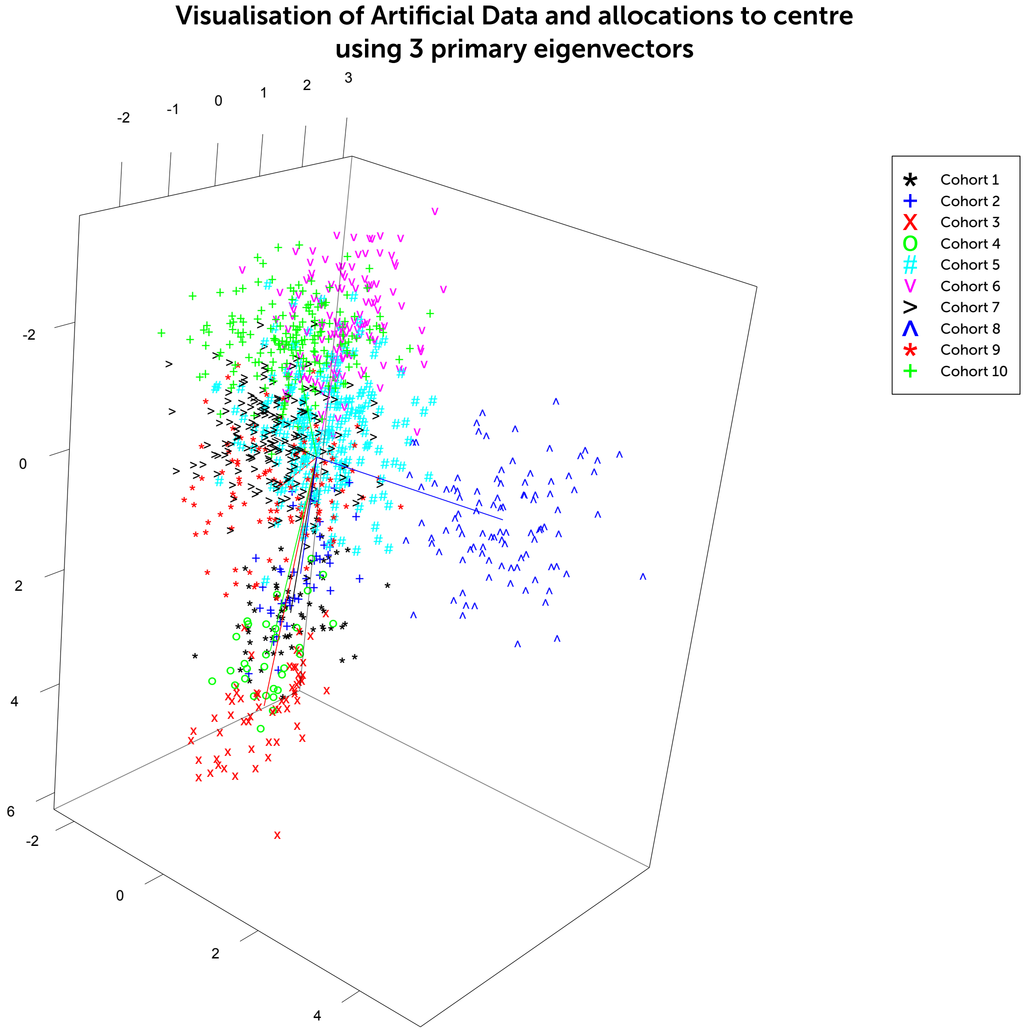


Figure 4.1: Raw Artificial Data, showing the generating cluster structure

able partition set as all that could be said is that one method might be less poorly performing than the other. To this end a reference partition was created to replicate what is theoretically achievable by the partitioning algorithm and to use this as our partition set for benchmarking. For each of the synthetic datasets the reference partition was defined by initialising a single run of the K-Means algorithm around the generating centres of the data, and iterating the algorithm to convergence. In this case the algorithm was Hartigan-Wong, utilising the online update to reduce the minimum sum of squares beyond the classic K-Means algorithm. This yields a partition set which is not an exact Voronoi decomposition of the data around the generating centres.

Having defined the reference partition for the synthetic data a comparison of the synthetic data with the reference partition can be done as shown

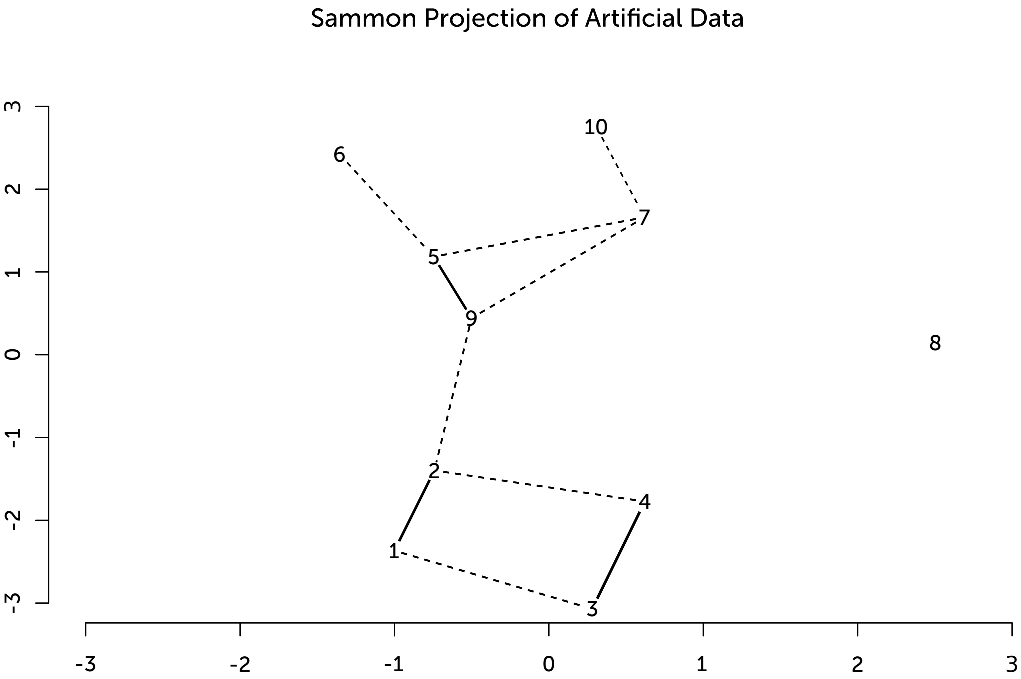
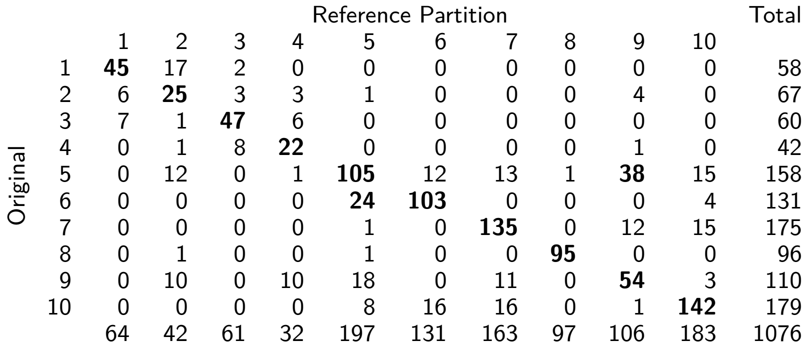


Figure 4.2: Sammon Map of the distances between the generating centres of the synthetic data.



		Reference Partition										Total
		1	2	3	4	5	6	7	8	9	10	
Original	1	**45**	17	2	0	0	0	0	0	0	0	58
	2	6	**25**	3	3	1	0	0	0	4	0	67
	3	7	1	**47**	6	0	0	0	0	0	0	60
	4	0	1	8	**22**	0	0	0	0	1	0	42
	5	0	12	0	1	**105**	12	13	1	**38**	15	158
	6	0	0	0	0	**24**	**103**	0	0	0	4	131
	7	0	0	0	0	1	0	**135**	0	12	15	175
	8	0	1	0	0	1	0	0	**95**	0	0	96
	9	0	10	0	10	18	0	11	0	**54**	3	110
	10	0	0	0	0	8	16	16	0	1	**142**	179
		64	42	61	32	197	131	163	97	106	183	1076

Table 4.4: Original Artificial Data vs Reference Partitions

in Table 4.4. This cross-tabulation shows a high correlation between the reference partition and the original cohorts. An exact match is not expected but does form a “good” partition against which the solutions can be measured.

4.2 Olive Oil Data

The olive oil dataset [65], is comprised of 572 observations with 8 characteristics, relating to fatty acid content of the olive oil. This data corresponds to 3 collection regions, with 9 sub-regions. Four from Southern Italy (North and South Apulia, Calabria and Sicily), three from Umbria (Umbria, East and West Liguria) and two from Sardinia (Inland and Coastal regions).

The projected visualisation of the underlying dataset [70] in Figure 4.3 shows each data point labelled according to the region from which it was obtained, highlighting intermixing of the data for Calabria, North and South Apulia and Sicily.

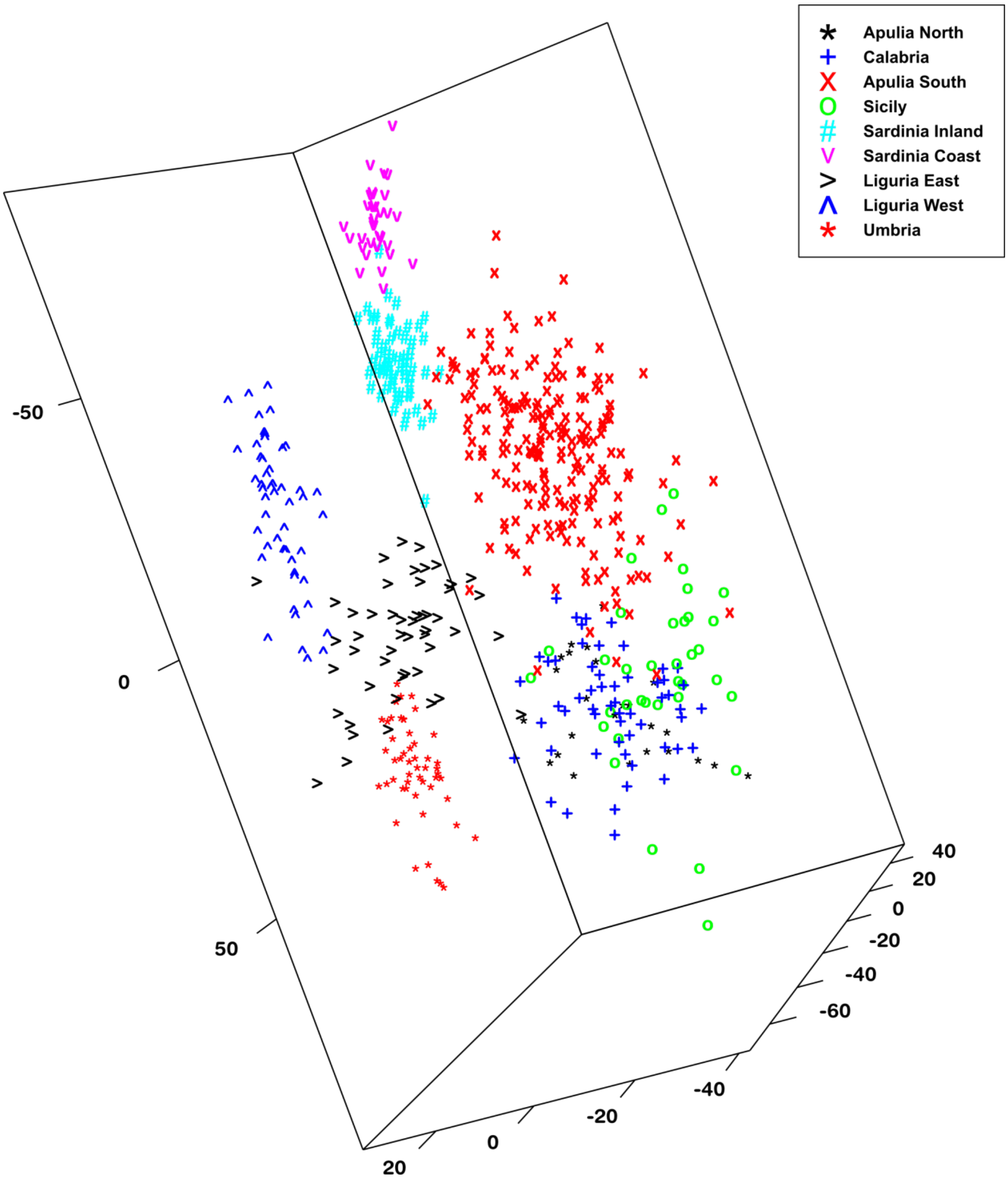


Figure 4.3: Visualisation of the unclustered olive oil data showing the Source Areas

As a result of the structure of the data, as with the Artificial data, there is no possibility for the K-Means algorithm to actually recover the complete structure of the data, however the cohorts whilst exhibiting some mixing should be largely recoverable, albeit with an error associated with it. This makes the dataset good to test the framework approach on.

4.3 Iris Data

The Iris dataset was introduced by Sir Ronald Fisher in 1936 for the purpose of using as an example in explaining discriminant analysis. The dataset comprises 150 data points in four dimensions matching the Sepal and Petal width and height for each observation. Figure 4.4 shows each of the four elements plotted pairwise with the three cohorts present in the data, Setosa, Virginica and Versicolor coloured Black, Red and green respectively.

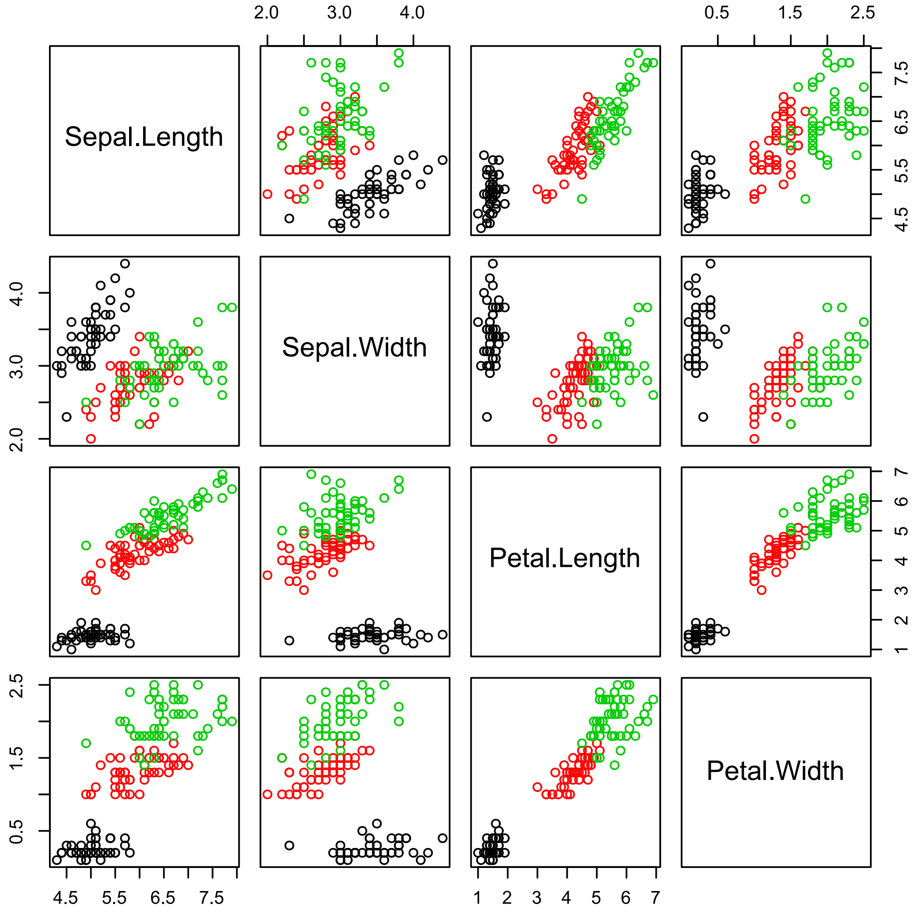


Figure 4.4: Paired scatterplot of each element of the Iris dataset

Also shown is that the three classifications can be separated by most pairs of values into groups of Setosa alone and Virginica and Versicolor together. For this reason it is not often used for clustering problems, as it generally gives high classification rates. This is good for our purposes as the dataset is intended to be used for benchmarking purposes to show the results compare well when using known data.

4.4 Wine Data

This dataset available on the UCI data repository is well known and comprises 178 observations in 13 variables, and was taken from a chemical analysis of wines grown in one region of Italy. Each of the attributes consists of measurements taken from the various wines, which are created using three distinct cultivars. The attributes are Alcohol, Malic Acid, Ash, Alcalinity of the Ash, Magnesium, Total Phenols, Flavonoids, Nonflavonoid Phenols, Proanthocyanins, Color Intensity, Hue, OD280/OD315 of diluted wines and Proline.

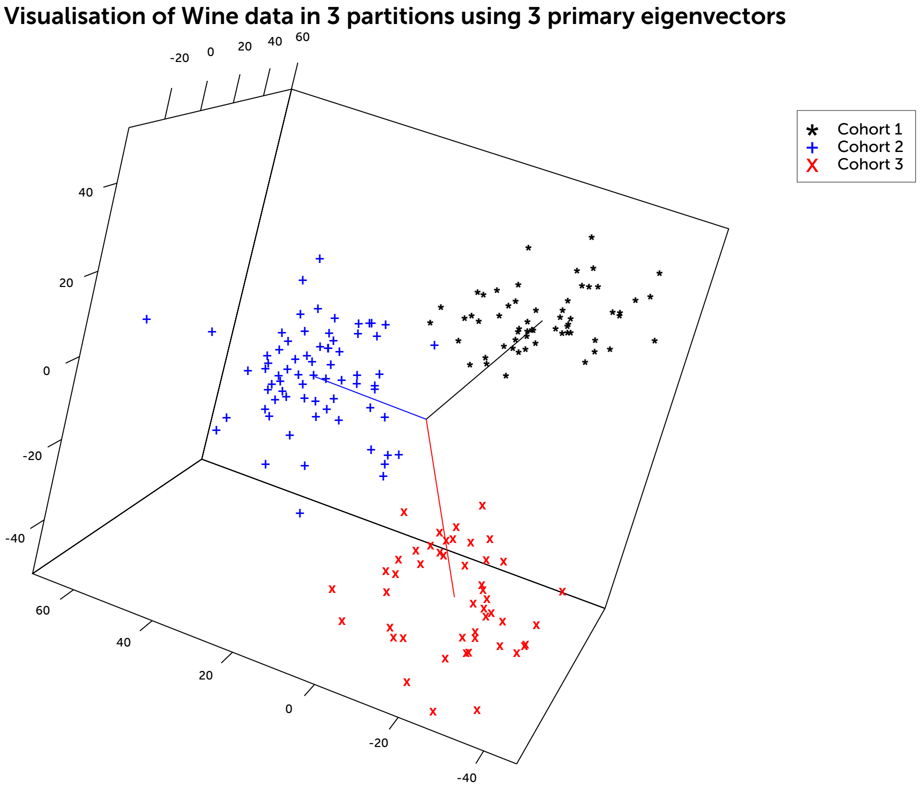


Figure 4.5: Visualisation of the three cultivars of the Wine dataset separated in three dimensions

A three dimensional visualisation of the dataset is shown in Figure 4.5 and it is evident from this visualisation that the cultivars are well separated with the expectation of good classification from K-Means.

4.5 Cardiotocography Data

The Cardiotocography dataset comprises the results of 2126 foetal cardiotocograms each having 21 attributes (shown below) which were automatically processed and the relevant diagnostic features measured prior to being classified into foetal state and morphologic pattern by a team of obstetricians[66, 67]. The size and complexity of this data renders it an ideal candidate for considering the utility of clustering methods on bioinformatics datasets. Previous studies have shown it to be particularly difficult to cluster using variance based techniques [72].

The following variables are found within the dataset:

LB FHR baseline - the heart rate during a longer 10 minute window (beats per minute)

AC Number of accelerations per second - where an acceleration is an abrupt increase in heart rate

FM Number of fetal movements per second

UC Number of uterine contractions per second

DL Number of light decelerations per second

DS Number of severe decelerations per second

DP Number of prolongued decelerations per second

ASTV Percentage of time with abnormal short term variability

MSTV Mean value of short term variability

ALTV Percentage of time with abnormal long term variability

MLTV Mean value of long term variability

Width Width of FHR histogram

Min Minimum of FHR histogram

Max Maximum of FHR histogram

Nmax Number of histogram peaks

Nzeros Number of histogram zeros

Mode Histogram mode

Mean Histogram mean

Median Histogram median

Variance Histogram variance

Tendency Histogram tendency

Two sets of classifications are available for this data

CLASS Foetal Heart Rate pattern class code (1 to 10)

NSP Foetal State Class Code (N = Normal, S = Suspect, P = Pathologic)

Having these two classifications means that this dataset suitable for comparing attempts at identifying gross and macro structure within data.

Using a projection method of visualisation giving a three dimensional representation of this dataset using the three principle separation axes [70] with the underlying data cohorts highlighted. Substantial intermixing is shown between cohorts eight, nine and ten with light mixing between the other cohorts as they spread out. Cohorts two, four, six and seven whilst adjacent are largely separated something which is shown in Figure 4.6. This shows each of the three principle separation axes plotted against each other in a lattice form so that each of the three axes are unobstructed.

This shows that the dataset provides some easier areas to segment, plus an area where poorer performance of the K-Means algorithm is to be expected as shown by results from a previous study on the dataset [72].

4.6 Thyroid Data

A dataset looking at thyroid function[66, 68] containing 215 observations with 5 features - T3resin, Thyroxin, Triiodothyronine, Thyroidstimulat-

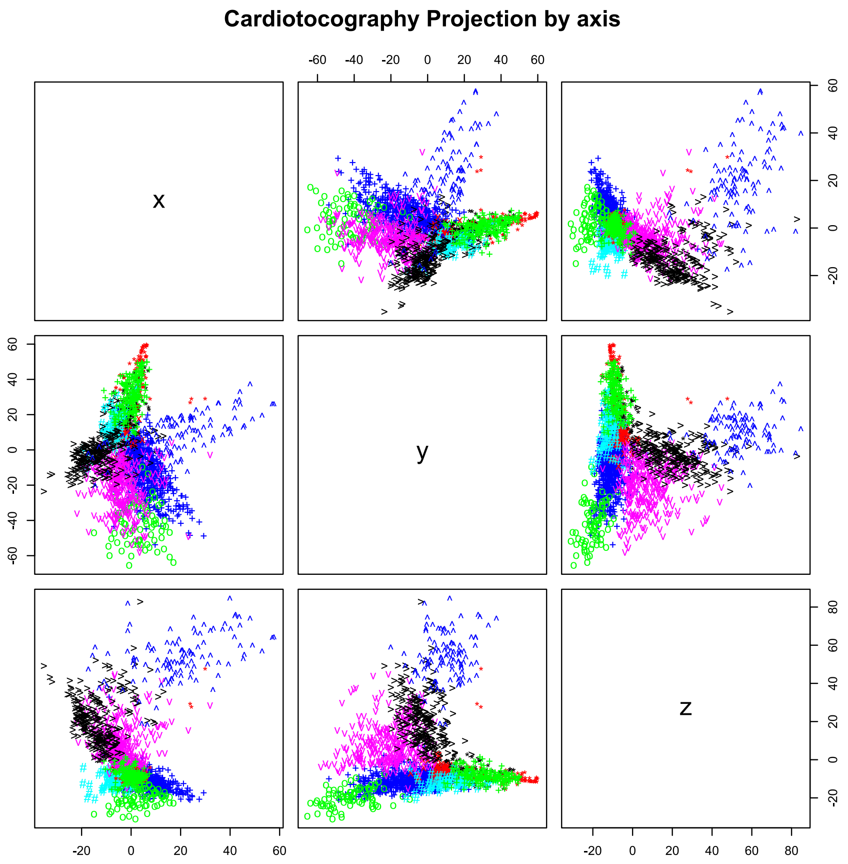


Figure 4.6: Visualisation of cardiotocography dataset projected across three axes, Cohorts 1 through 10 coloured as: Black $*$, Blue $+$, Red X, Green O, Light Blue #, Purple V, Black $>$, Blue , Red $*$and Green $+$ respectively.

ing, TSH_value - with three target classes, Normal, Hyperthyroidism and Hypothyroidism.

This visualisation shows that there are three groups within the data coterminous with each other but visually separable through multiple axes, and as such it should be reasonable to expect a successful clustering of the data.

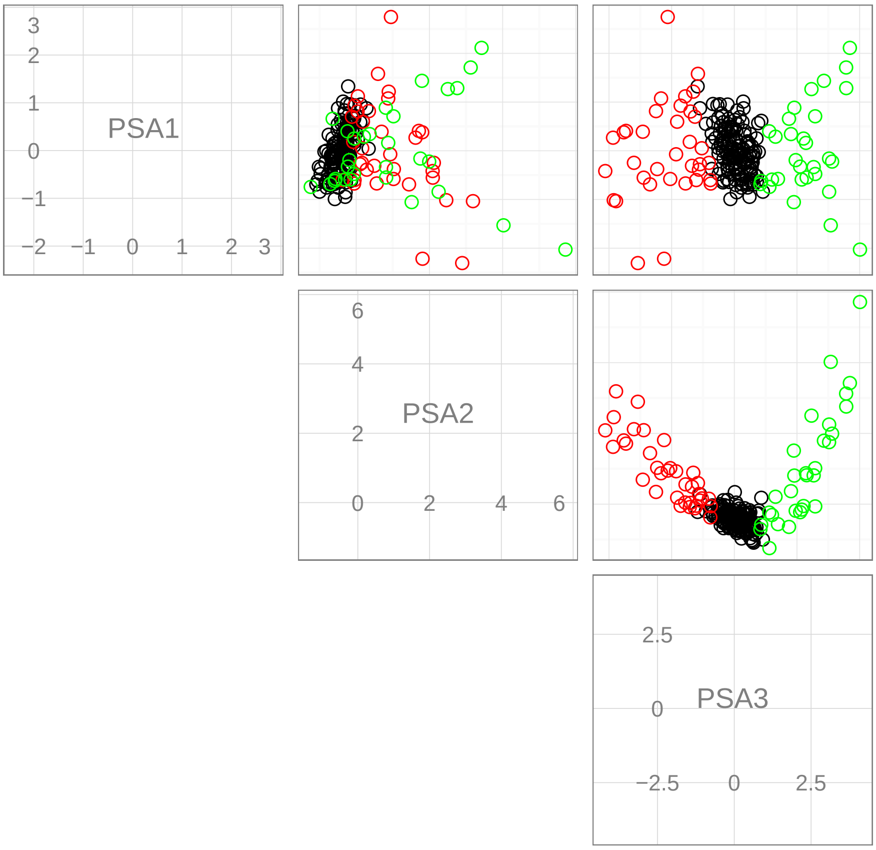


Figure 4.7: Visualisation of Thyroid dataset projected across three axes

5 Experimental Methodology

Having proposed a dual metric framework approach to initialisation dependent clustering problems it is necessary to test and benchmark the performance of the framework both to compare with the expected results of an improvement in performance and stability. This chapter deals with the methods used during the benchmarking process to assess the suitability of the proposed method under certain conditions.

5.1 Reference Partitions

Because of the designed complexity within the synthetic dataset it is not possible to properly recover the generating partitions using a variance based clustering method and is thus limiting in its utility in evaluating the performance of a given clustered solution. To resolve this it was determined that a reference partition could be used rather than the underlying structure of the data as a means of comparing against an achievable partition.

For each of these datasets the reference partitions were created by initialising K-Means on the data and using the original generating centres of the 10 partitions as a source for the prototypes. With the other non-synthetic datasets the reference partition of the data is the original classifications included such as the regions for the Olive Oil data

The exception to this reference partition is for the datasets of varying density as they are samples from the same space, therefore the reference partition was calculated using the densest (10,000 data points) of the datasets and the results then applied to the lower density sets. K-Means was applied to this dense dataset 500 times, and the solution with the lowest within cluster SSE (minimized objective function) used.

5.2 Repetition

The intent of the proposed SeCo framework is to provide a stable reproducible platform for obtaining solutions from an initialisation dependent clustering algorithm, in evaluating the stability of this method it is necessary to evaluate the reproducibility of solutions against other iterations. Robustly benchmarking this requires repeating the method and comparing the solutions obtained with those obtained previously, and given enough samples inference can be made about the relative performance of different solution selection criteria.

Algorithm 5 Algorithm for repeated testing

```
for $i = 1 \rightarrow 100$ do
    $res = framework(data, k, n)$
    $result_i = best(res, t)$
end for
for each $result$ do
    $C_i = CramerV(result_i, reference)$
end for
```

For each method of selecting a solution under test the following basic procedure was followed whereby a complete run of the framework is performed on a given dataset using all the criteria which would be varied under normal usage. This run would be repeated a given number of times N as is aptly demonstrated by the pseudo code shown in Algorithm 5, which shows the procedure for obtaining the results for each parameter set. Following these runs each selection criteria is applied to each result set and the solution is compared to the reference partition.

5.3 Dual Measure Choice

As part of the evaluation process for a given algorithm being used within the SeCo framework the choice of which metrics to use requires investigation to ensure that the results being obtained are both usable and sensible. The concordance measure is of less importance than the separation measure as any adequate metric can be used; here the Cramer's V statistic is used throughout, however evaluating how well separated the solutions are is less well defined, and appropriate choice must be made.

Evaluating the stability of each selection criteria was identified from a set of 100 complete framework results and the selection criteria was calculated both alone and as part of the framework. The purpose behind this was to gain an understanding into structural differences which could be expected in the outputs from the algorithms. In all cases the same sets of initialisations were used to ensure that the only variable factor was the algorithm.

The choices here for separation measure include the objective function of the algorithm in question (in this case K-Means) and Invariant J [73] working on the assumption that using the objective function as the separation metric might result in a solution which is over-fitted. The evaluation was performed for the Synthetic Dataset and real world datasets with the intention of comparing the different methods using the same set of results.

The parameters looked at a range of values for k with a fixed number of re-samples for each of the datasets used for evaluation. The resultant partitions were compared against the reference partitions so as to be able to evaluate the performance of each of the five methods of selecting the best partition.

5.4 Algorithm Choice

The SeCo framework is intended to be used with many algorithms and with algorithms such as K-Means having many slight variations such as “Lloyds“ and “Hartigan-Wong“ a fundamental question was how do these different algorithms affect the solutions. In the case of K-Means the first of these is the simplest, and closest to MacQueen[9], and performs a Voronoi decomposition of the data. The second is the updated algorithm by Hartigan[10] which performs an additional online update stage not present in the original algorithm optimising the objective function further.

These two algorithms have the same objective to minimise the Total Within Cluster Sum of Squares (SSE) but they obtain different results. The first step before continuing with the benchmarking of the framework approach to clustering was to consider the differences which might occur with the use of these two algorithms. Testing both of these algorithms within the framework allows the differences between the two in terms of reproducibility of solution to be adequately quantified.

5.5 Thresholds

The framework uses a thresholding technique whereby the results are top-sliced into those considered well separated and those considered less well separated, with only the best separated solutions being considered on the basis that a less well separated solution will not represent the underlying structure of the data as well. It is necessary however to evaluate this assumption and provide empirical evidence for this.

Experimenting with different threshold levels should allow for evaluation of if this was the case and whether the selection of a different threshold might change the result. Should this be the case then it might be possible to identify an optimal threshold value for a particular solution set. To this end the framework was applied to the same set of initialisations used previously using varied threshold values. Each of these threshold values was considered for each value of k such that a complete picture of the variation associated with the threshold values could be considered.

Each method and value of k were evaluated with thresholds of 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100% of the solutions as ordered by the relevant separation measure. Obviously for the two single measure approaches, the solution was the same for each of the thresholds and was included for comparison against the dual metric approaches.

5.6 Concordance Values

The results of multiple runs of the framework highlighted that multiple solutions with a similar separation criteria had widely differing median concordance and conversely that solutions with the same separation value had different median concordance; it made sense therefore to investigate.

To look into this the concordance of all the solutions for a particular clustering run with the reference partition were calculated with the intent of clarifying if the utility of the dual metric approach remained unclear.

One issue identified was the appropriateness of the selected best solution particularly with respect to the thresholding of the top 10%; for example with SSE there are always many solutions with equivalent SSE value. In this case applying a threshold of 10% of solutions would necessarily mean that some solutions with supposedly equivalent separation are ignored. Therefore with varied thresholds it is not always evident as to which solutions to select, especially if the top 5% to 15% have the same separation measure as you need to select half of them for the top 10%. Selecting at random from those sets of solutions with the same SSE to obtain the desired number of solutions seems an appropriate solution.

5.7 Benchmarking on New Data

In the real world usage classification methods are often used for exploratory analysis however they are also used for the purpose of predicting whether a new data point would belong to a particular class. Therefore in the process of validating any method for classification it often becomes necessary at some point to evaluate the method as a means of predicting the classification of values in a new dataset.

For this purpose test and train datasets are often created so for this purpose ten datasets were generated, using the same specifications as the original data allowing for the performance of the SeCo framework to be evaluated as a predictive tool and whether it produced results which gave better classification on the test data.

5.8 Performance Measures

Two additional measures were used to assist in the evaluation of the performance for the framework and standard approaches to clustering. The first of these is accuracy, the proportion of correctly classified objects for the clustering, and is compared against the reference partitions, the second is affinity, a row level indicator that identifies how often each particular data point is assigned to the same cohort.

This can then be used to determine a dataset-level affinity measure for each method. This gives an indication of how often data points swap cohorts for the whole dataset, and thus providing a measure of the stability of the solutions from the perspective of an individual.

Applied SeCo - K-Means 6

This chapter covers the methods and results specific to benchmarking the framework with the K-Means algorithm and specific extensions of it such as the selection of an appropriate value of k. First however, having described the application of the framework to the K-Means algorithm and described each of the datasets the first step is to describe and discuss the results in terms of what can be seen from these results, to this end the first part of the chapter assumes that the evidence for the use of the framework is conclusive and the second part lays out this evidence.

6.1 Application of the Framework

Each of the Synthetic Data, Olive Oil, Wine, Iris and Cardiotocography datasets are run through the framework in turn with the results discussed below. Also discussed are inferences drawn using the framework as an exploratory analysis tool.

Synthetic Data

This section refers to the original synthetic dataset comprising of 1076 rows of data, with 3 attributes each. For this data it is known that there are ten underlying classifications so the objective is to evaluate how well the framework performs in indicating how well these will be recovered. Figure 4.1 showed the structure of the data with the underlying cohort that each data point was assigned to colour coded. From this it will be possible to compare the result obtained from a solution with the framework and see how they compare.

When applying the framework to the data K-Means is run a set number of times with different initial samples and values of k. From these a SeCo map is produced showing the structure of the results for the data as seen in Figure 6.1. The map shows the results for values of k between 2 and 15 and for which 500 sample initialisations were used. In this case the top 10% of values were selected for inclusion in the map.

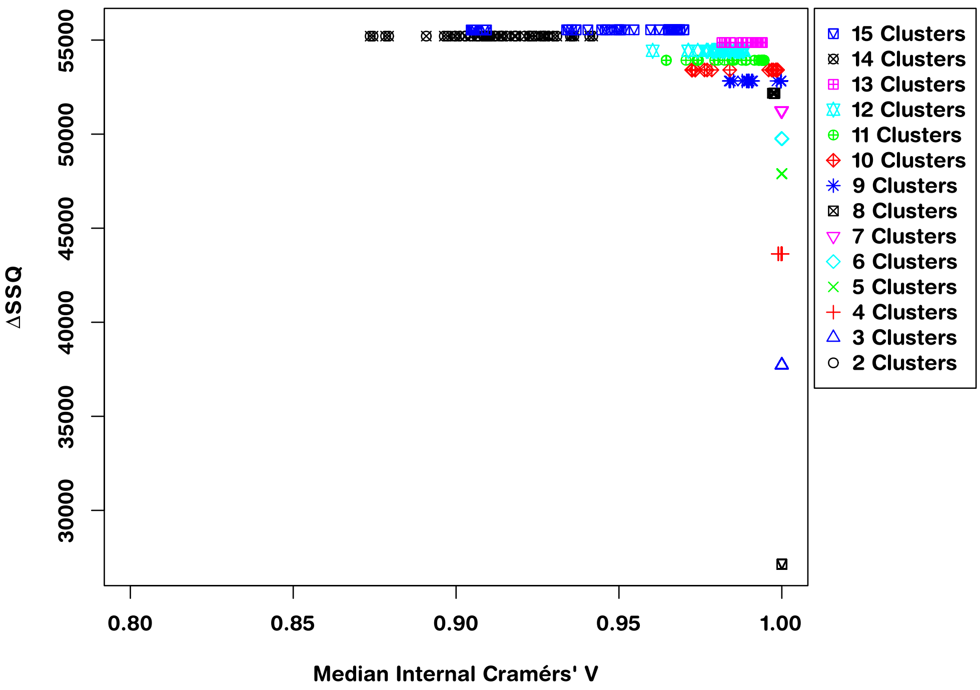


Figure 6.1: SeCo Map for the original synthetic dataset. Showing ΔSSE on y-axis and $med(CV)$ on x-axis

The y-axis represents the ΔSSE which is the difference between the Total Within Cluster Sum of Squared Error and the Total Sum of Squared Error of the data from the mean. The x-axis represents solution stability and is calculated using Cramers V statistic, having the range 0 to 1, and in this case the axis has been truncated from the left to aid readability.

Solutions indicated to the right of the graph as more consistent with the other solutions than those further to the left. Each of the different values of k is coloured distinctly in the map to aid identification of possible solutions. Because a 10% threshold is being used and there are 500 initialisations there are 50 data points representing each value of k.

Given that if a new dataset were being used no prior knowledge of the underlying structure of the data would be available this map well be described as though this were the case. Reading the map it can be seen that that the lower values of k have solutions which are tightly grouped together, and aligned tightly up against the right hand edge of the map. This is the case for $k = 2$–7 and would indicate that the solutions which are selected by the framework are close to identical for these values.

At $k = 8$ there is a slight drop in the concordance as the values start to edge away from the right hand edge however they are still tightly grouped which might indicate that there are possibly two ways to start cutting the data and that the solution is no longer trivial. As k goes up increasing variation is seen not just in the stability of the solution, but also in the ΔSSE value which shows that there are more distinct solutions than before. When $k = 14$ solutions are shown to have a median concordance ranging from ≈ 0.85 to ≈ 0.95.

The known true number of partitions ($k = 10$) does not perform as well as either $k = 8$ or $k = 9$ with two clouds of solutions appearing at $C_v \approx 0.97$ and $C_v \approx 0.99$. The latter group being almost but not quite on the right hand edge of the map.

What is being looked for is a set of solutions which have good consistency so should the analysis be re-run then a solution which is very close to the previous solution would be obtained. Keeping this criteria in mind it can be seen that a definitely good set of solutions would be at $k = 8$; however the map indicates that both $k = 9$ and $k = 10$ are solutions worth taking a look at and considering seriously.

Having selected a solution to consider in this case going with the best value being $k = 8$, the structure of the data can be visualised highlighting the new cohort allocations for comparison with the earlier plot. Figure 6.2 shows this visualisation, and it can be clearly seen that each of the cohorts is now clearly delineated compared to those around it. One distinct cohort of data is separated out from the remainde, corresponding with cluster 8 in the multidimensional scaling map seen in Figure 4.2. The data points in the lower area appear to have been combined together.

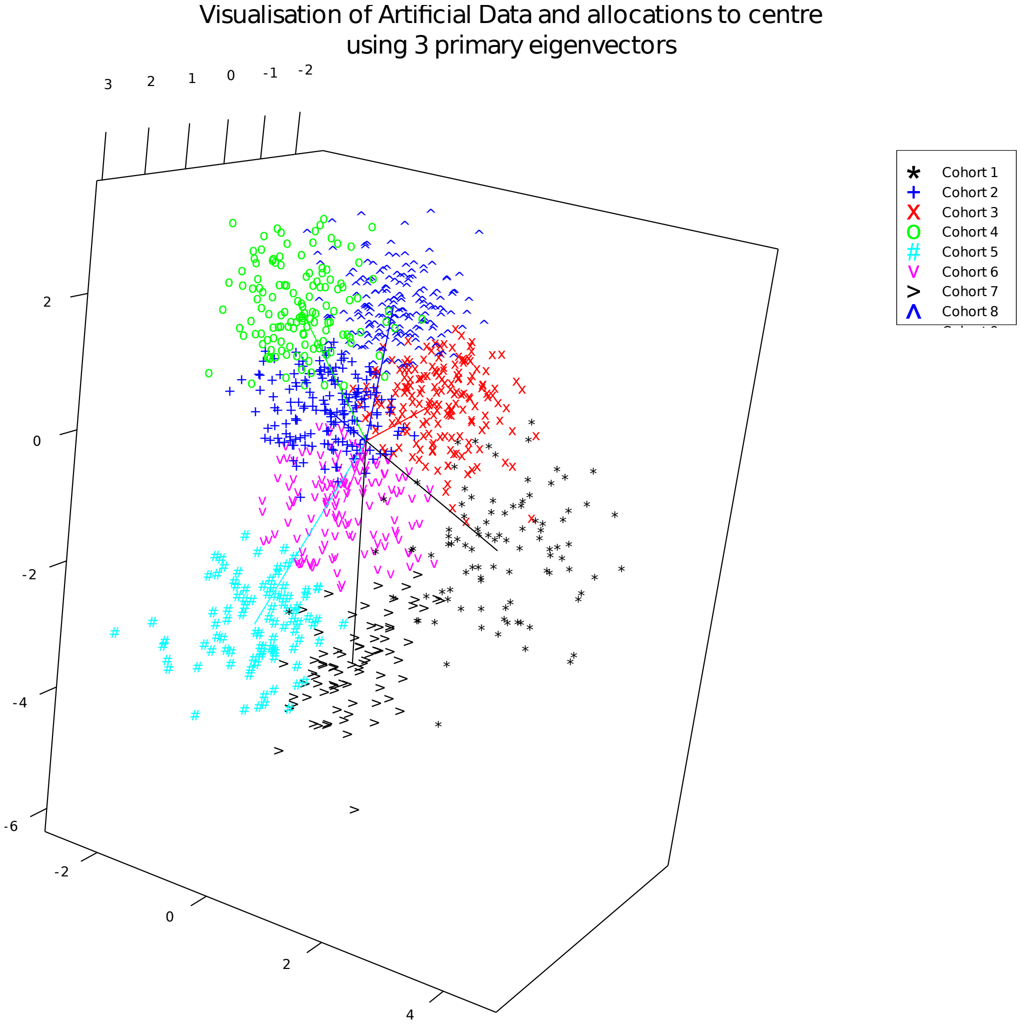


Figure 6.2: Visualisation in 3D of Synthetic Data using 8 cluster solution

				8 Cohort Solution						Total
	Cohort	1	2	3	4	5	6	7	8	
Reference	1	**58**	0	0	0	0	0	0	0	58
	2	**31**	**35**	0	0	0	0	1	0	67
	3	5	0	**55**	0	0	0	0	0	60
	4	0	19	**22**	0	0	1	0	0	42
	5	0	4	0	**153**	1	0	0	0	158
	6	0	0	0	0	**131**	0	0	0	131
	7	0	0	0	1	0	**168**	0	6	175
	8	0	0	0	0	0	0	**96**	0	96
	9	0	**77**	0	**21**	0	12	0	0	110
	10	0	0	0	1	0	0	0	**178**	179
Total	94	135	77	176	132	181	97	184	1076	

Table 6.1: Comparison of the reference partition with the 8 cohort solution selected from SeCo map.

Table 6.1 shows the comparison between the reference solution and our selected best solution. Also even though the comparison is of the 8 cohorts solution to the reference partitions 10 cohorts, there is a solid diagonal indicating a strong association between the two. There is some mixing of the cohorts, most notably cohort 9 in the reference data is split between cohorts 2 and 4 in the solution, and cohort 2 is combined split between cohorts 1 and 2; Cohort 4 is split between 2 and 3.

Having visualised these solutions it is interesting to look at the hierarchical nature of the solutions and Figure 6.3 shows how the data points migrate from one cluster to the next for the best solution in each of the values of k and in this case the eight cohort partition is highlighted in red. It can be seen as early as $k = 3$, that the separate cohort separates from the other data points, and remains so through the entire map.

There is a strong hierarchical structure where k is less than 8 and in $k = 9$ the cohort structure remain largely the same although interestingly the separate cohort splits only to merge again when $k = 10$. At this point there is more mixing of the cohorts as the solution appears to become less stable, and the plot echoes the SeCo map in this regard.

Looking at these results the SeCo framework clearly provides an informative way of visualising the space of solutions and aids in the selection of an appropriate solution for a given dataset.

Olive Oil

Taking a look now at the Olive Oil dataset the SeCo map in Figure 6.4 looks at solutions for values of $k = 2$ to $k = 12$. In this set of results for most values of k solutions are aligned with the right hand edge of the stability axis. This indicates that good solutions can be obtained for all these values of k and that the point at which the K-means algorithm starts to have problems partitioning the data is at $k = 11$. Although for $k = 9$ most solutions are aligned with the right hand side of the axis a small number of solutions have concordance of ≈ 0.95 instead of ≈ 1 as with the other solutions.

Judging from this SeCo map it can be expected that good solutions will be obtainable for all values of k up to $k = 10$ however the known structure of this data hold nine cohorts the subregions. Visualising the data and the allocated cohorts as in Figure 6.5 the results show that the data for $k = 9$ have been well separated out into distinct well defined cohorts.

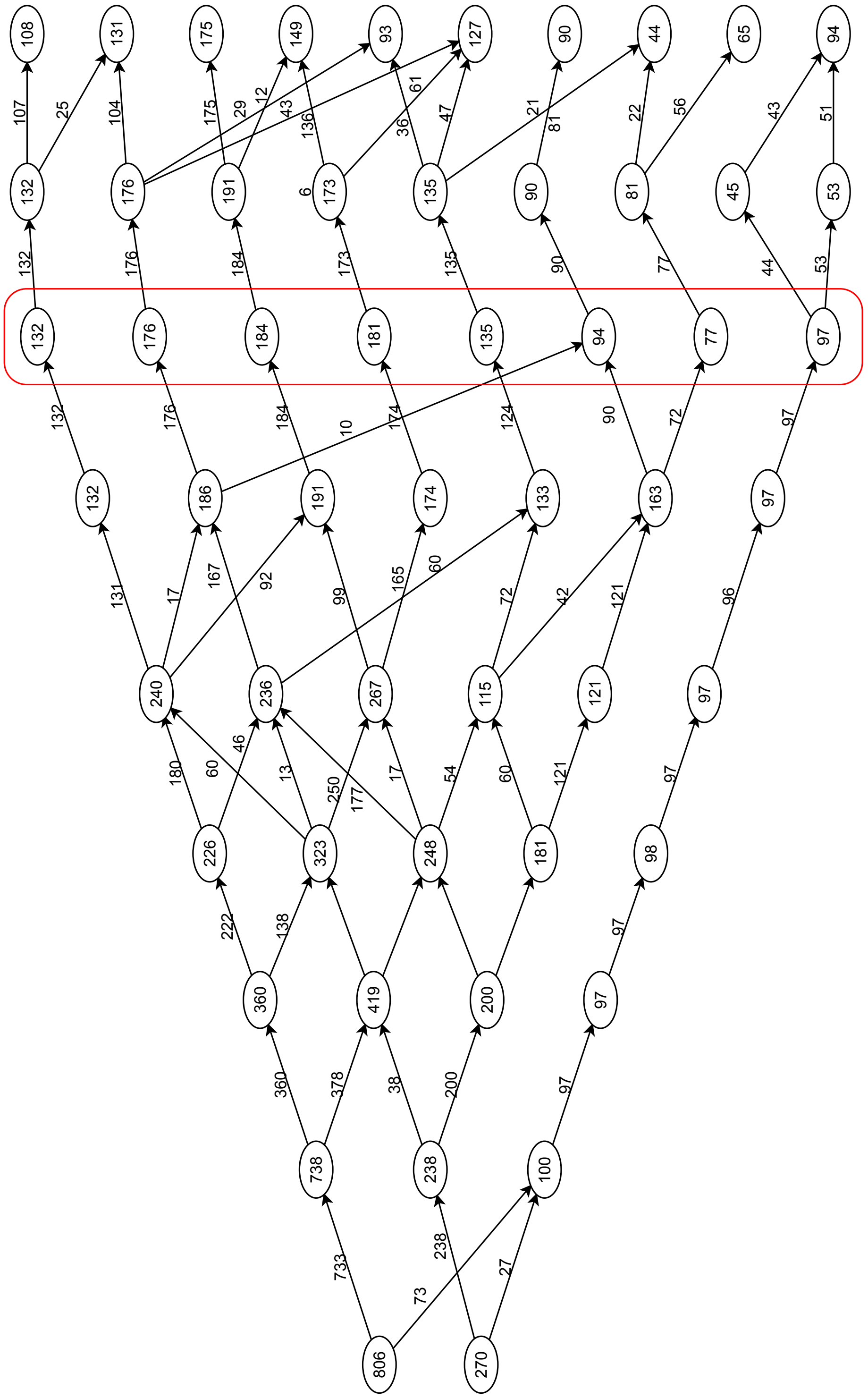


Figure 6.3: Hierarchical structure for clusters 2 to 10 for the artificial data

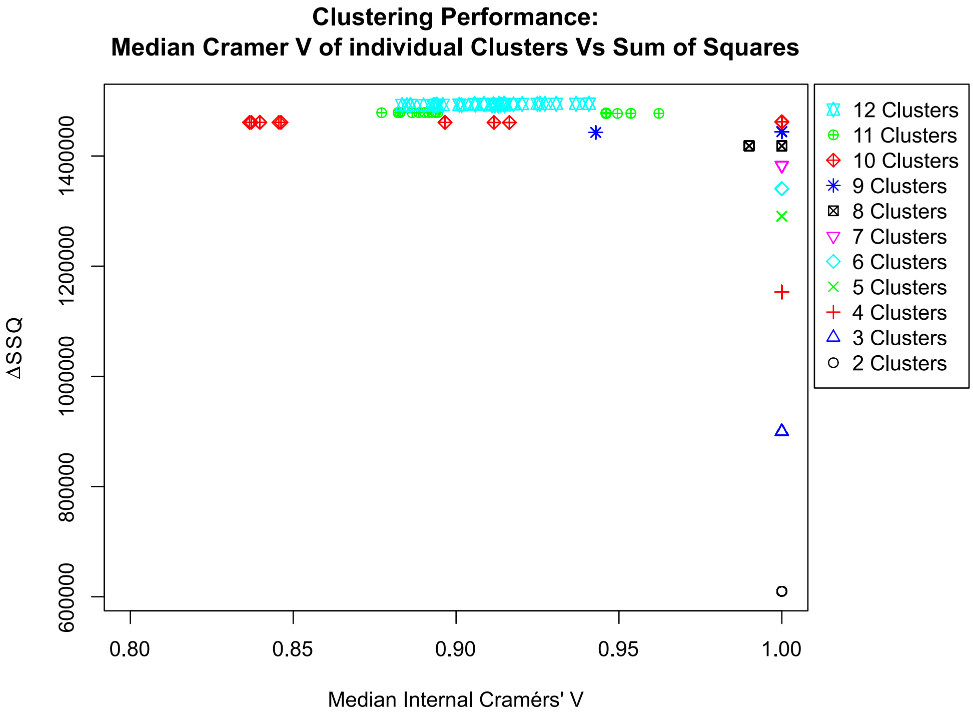


Figure 6.4: SeCo Map for the Olive Oil Dataset

Iris

The iris data has only three partitions compared to the two previous datasets which had ten and nine respectively so as a consequence of this it might be expected that the SeCo map will not have such favourable performance up to higher values of k. This is because the algorithm will be trying to split coherent partitions into much smaller components perhaps giving many potential solutions. Applying the framework gives the SeCo map as can be seen in Figure 6.6 and indeed it can be seen that good solutions are indicated for values of k between 2 and 5 and at $k = 6$ the results start to deteriorate rapidly with this partition set having three distinctly separated clouds of solutions. At $k = 7$ the long continuous clouds of data being shown indicating that even though the solutions have a similar objective function score there are substantially different solutions to be found with varying degrees of concordance.

Despite having three underlying cohorts in the dat, the algorithm seems able to partition the data in a stable manner up until $k = 5$ as K-Means is partitioning the existing cohorts into sub-cohorts developing micro struc-

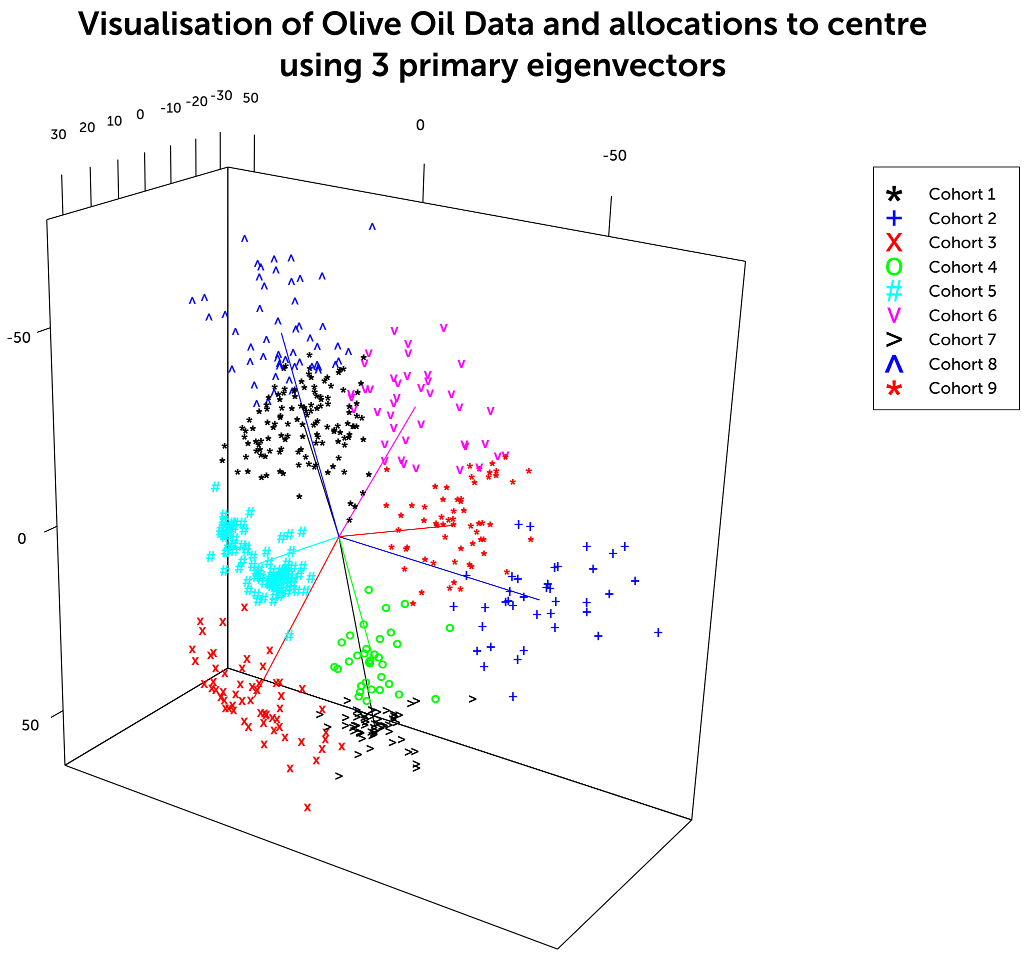


Figure 6.5: Visualisation of the Olive Oil Data showing the assigned cohorts from the selected solution

ture.

Wine

For the wine data the same sort of behaviour occurs that was seen for the Iris data albeit with much greater exaggeration of the effect. Looking at the SeCo map in Figure 6.7 the deterioration of results becomes pronounced at $k = 6$ and as k rises towards $k = 12$ again the solution clouds can be seen trailing back from the right hand edge of the map.

The best solutions for this dataset appear to occur at $k = 3$ coinciding with the underlying number of data partitions and the map clearly shows this.

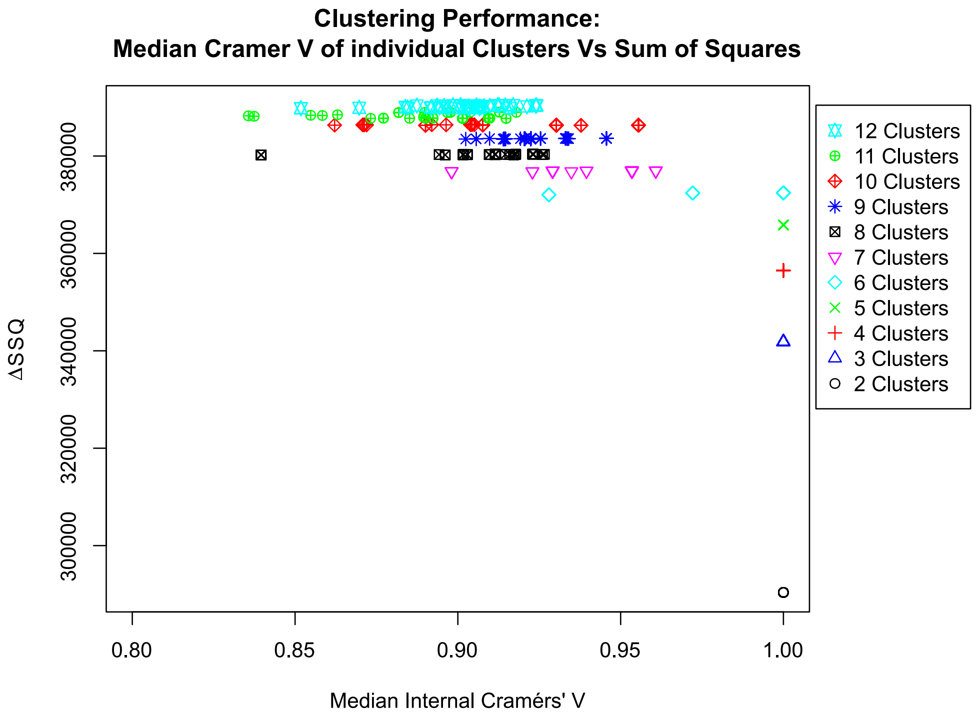


Figure 6.6: SeCo Map for the Iris Dataset

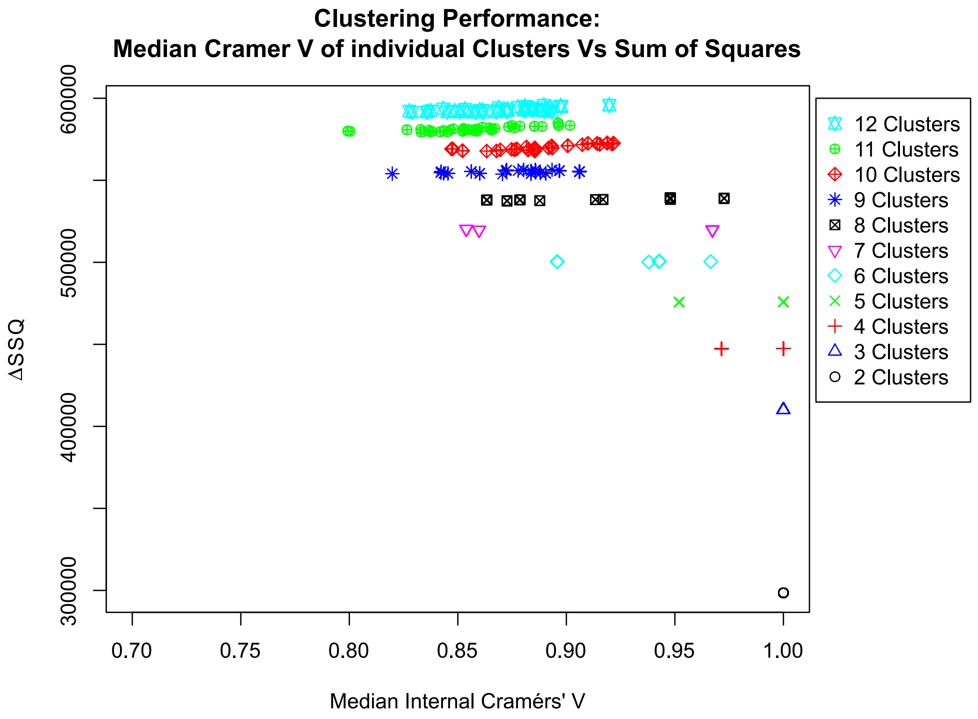


Figure 6.7: SeCo Map for the Wine Dataset

Cardiotocography

With the Cardiotocography dataset previously observed behaviours continue to be seen with the data being segmented easily for lower values of k and as this tends upwards there is increased variation of solution for the same value. For values of k up to eight very high concordance is shown with slight deviation at six cohorts.

At $K = 9$ and above there is an obvious reduction of the within cluster concordance and a greatly increased spread of solutions recovering slightly for $k = 10$ then breaking down further at $k = 11$ and $k = 12$. This implies that the greatest amount of structure is recovered for $k = 8$ and given the complexity of the data as noted above with substantial intermixing of the cohorts this falls in line with expectations.

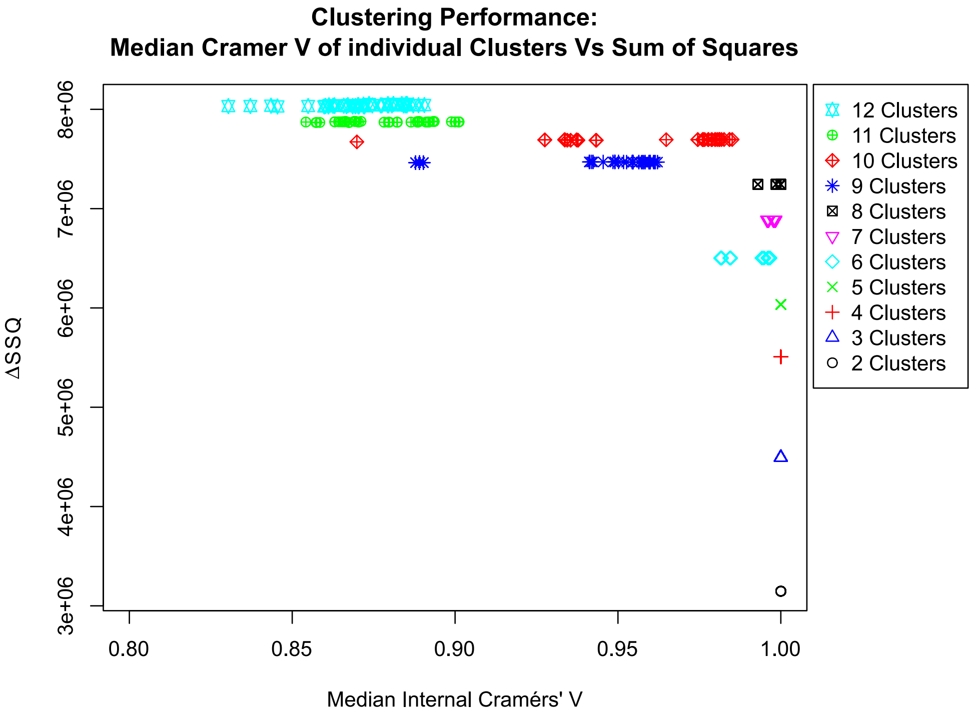


Figure 6.8: SeCo Map for the Cardiotocography Dataset

Summary of Results

The SeCo maps produced by the framework provide valuable insights into the how the K-Means algorithm is partitioning the data and how stable the results are. The use of the stability measure provides substantial information as to which solutions are not likely to yield a good solution as in the Wine data and it is clear that higher values of k will not necessarily provide good results.

It can be noted that the lack of variation within the strata for each value of k indicates whether the structure of the data is being captured with ease for the particular problem and the proliferation of variance in the structure delineates the point at which consideration as to whether refraining from sampling further or not is wise.

Having introduced the framework approach for each of the test datasets and discussed the results the following section presents the results from the multiple experiments done to evaluate the performance of this dual measure approach. This is done in contrast with the existing practice of selecting the solution with the lowest value of the objective function after a number of reinitialisations.

6.2 Benchmarking Methods

During the benchmarking process different tests were performed on the framework which are described in the sections below.

Repeated Testing

After comparing the different methods of selecting a solution the differences between these became evident and further investigation required as the possibility remained that this set of results was not representative of those likely to be achieved in real-world usage. To this end the framework approach was repeated 100 times, as was the method of selecting a sample based on the objective function. The results comprised a comprehensive set of solutions which could be compared to the reference partition and an evaluation of relative performance made.

Performing the calculations this many times would allow for an indication of the stability of each of the methods and also for testing as to significance in terms of differences in concordance between two or more populations.

Concordance

Following on from the previous section it was obvious to investigate what was happening as the solutions were generated for each clustering problem. To look into this the concordance of all the solutions for a particular clustering run with the reference partition were calculated with the intent to clarify if the utility of the dual metric approach remained unclear.

An issue arising from the previous experiment was the appropriateness of the selected best solution for example with SSE there are always many values with equivalent value. So with differing thresholds it is not always evident which solutions should be selected. This is especially the case if the top 5% to 15% have the same separation measure as you need to select half of them for the top 10%, but they are all in theory equivalently good solutions.

This was resolved here by introducing a sampling approach where the values were randomly sampled so that a top 10% threshold would be selected and the values in that top 10% would be representative of all the solutions which could have appeared in there.

Informed by the thresholding experiment the concordances with the reference partition were visualised to allow for easy comparison of the methods. Given that each solution would have a consistent value when compared to the reference partition this visualisation showed what was happening to the solutions when the stability measure was applied at different thresholds.

The utility of this approach was that it would be possible to see if over fitting was occurring with the different metrics and also whether the performance variation of the algorithm was high. Repeating this with slightly different initial conditions it became clear that even selecting the best solution was no guarantee that the solutions would be similar. So from this it became necessary to exhaustively compare all the possible settings and to repeatedly test each possible combination to establish the stability of the results.

Testing of obtained solutions against new data

In the real world usage classification methods are often used for exploratory analysis however they are also used for the purpose of predicting whether a new data point would belong to a particular class. Therefore in the process of validating any method for classification it often becomes necessary at some point to evaluate the method as a means of predicting the classification of values in a new dataset.

For this purpose test and train datasets are often created and here a total of ten datasets were generated, using the same specifications as the original data allowing for the performance of the framework to be evaluated as a predictive tool and whether it produced results which gave better classification on the test data.

6.3 Selection of k

Having established that the use of a SeCo map gives improved stability this allows for use of the information to produce an Integrated Cumulative Cramer V plot (ICC) of the cumulative pairwise Cramérs' V distribution an example of which is shown in Figure 6.9. This plot identifies for which values of k the clustering solutions produce stable solutions indicating the algorithm is identifying the same or very similar solutions for each run.

Figure 6.9 shows the ICC for the well-known Iris dataset the structure of which is well known, and performing the clustering on the two most informative variables (Petal Width and Length) and visualising the results in two dimensions for six cohorts gives the solution shown in Figure 6.15b it can be seen that for the six cohort solutions, one of the three underlying solutions remains separate from the rest, and recovered appropriately.

Integrated Cumulative Cramer V

Looking at each of the datasets in turn, starting with the Iris dataset, Figure 6.6 shows the SeCo map for the Iris Dataset with Figure 6.9 showing the ICC map for the same. The lower panel shows the Integrated Cumulative Cramer V for each of the different numbers of cohorts and the upper panel the actual curve for which the area is being measured from. Showing the upper panel does not provide any significant information above that of the lower panel however and is included here simply for reference purposes as interpretation is best done using the lower panel.

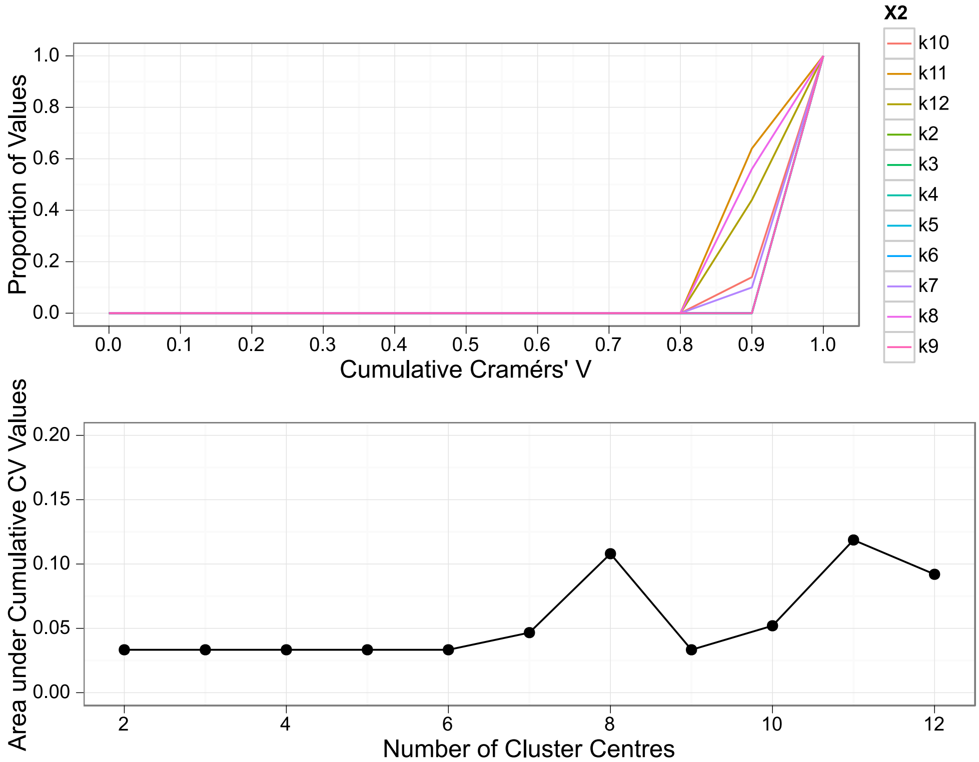


Figure 6.9: Integrated Cumulative Cramer V (ICC) for the Iris Dataset

From this it can be seen that good structure is being recovered for the Iris dataset for any value of k up to six with the performance deteriorating beyond this except for nine which recovers slightly. This is not implying that six cohorts is the best choice here but rather that recovery is being made of interesting structure for all values up to this inclusive and that each of these warrants investigation. Given however that it is already known that there are three cohorts within the data, that the method is highlighting useful values up to six implies that the algorithm is consistently splitting the cohorts in the same way as the value of k increases.

It is important to emphasise that user discretion is best used for this method as it is not intended to be a definitive answer to the number of cohorts, but rather to inform the user of values of interest which warrant further investigation. If pushed then selection of an appropriate

For the cardiotocography dataset in Figure 6.10 it is obvious that for values of k between two and eight consistently good solutions are obtained. This indicates that good structure is being obtained and therefore any value in this range could be used. At k=9 there is a drop in the quality of solutions, indicating that this is not a good choice of k, however for $k = 10$ there is a recovery of performance, followed by further degradation as k continues to rise.

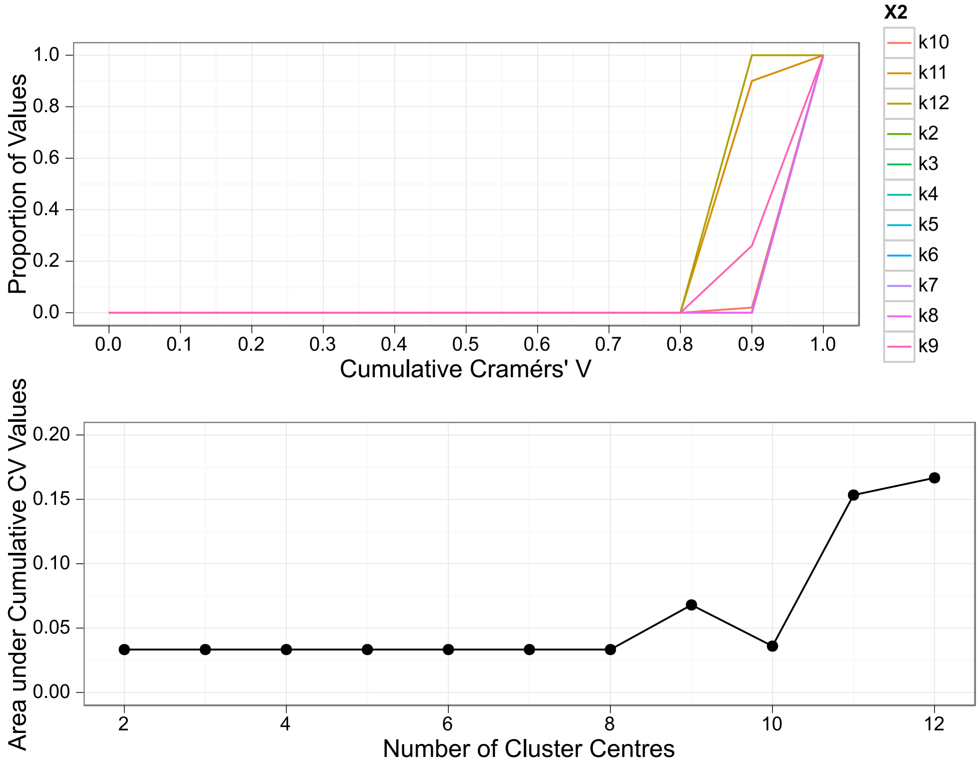


Figure 6.10: Integrated Cumulative Cramer V (ICC) for the Cardiotocography Dataset

For the Olive Oil dataset shown in Figure 6.11 the ICC method shows that structure is being recovered for cohorts between 2 and 9. However following this the recovered partitions are no longer consistent therefore indicating that a value of $k = 9$ is a good choice. The data is split into three regions and nine sub-regions so the method correctly identifies the appropriate value here.

The Wine dataset for which the ICC information is shown in Figure 6.12 a similar pattern is seen to other datasets, except that it indicates that the number of cohorts is lower. After 5 cohorts, there is a four-fold increase in the Integrated Cumulative Cramer V for higher values of k, at which point the measure seems to stabilize.

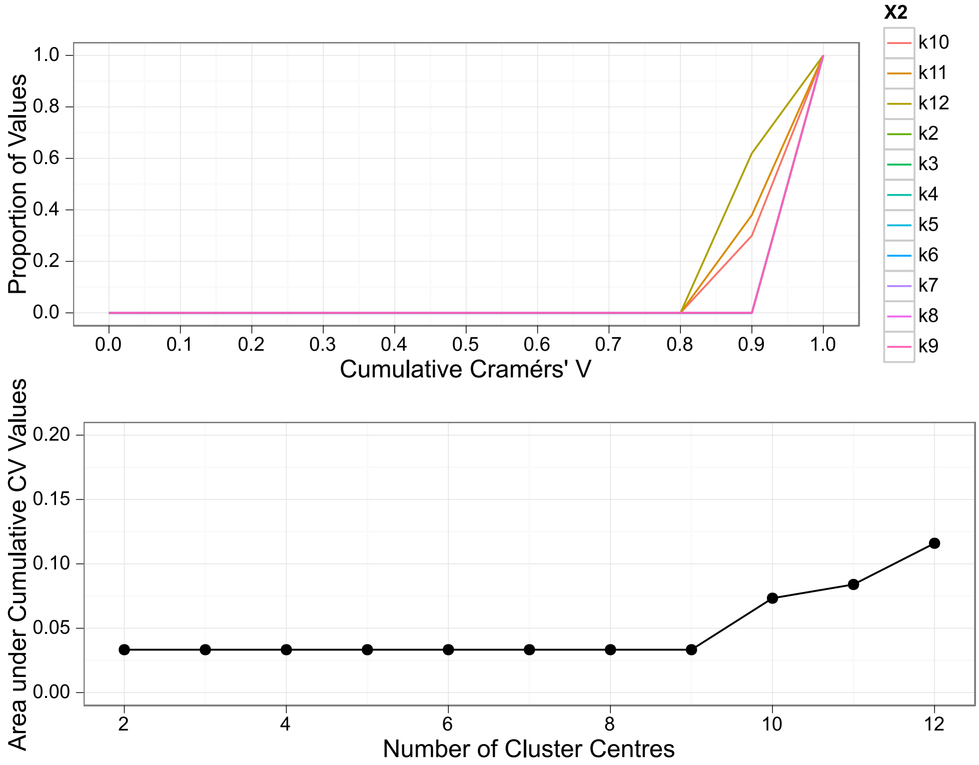


Figure 6.11: Integrated Cumulative Cramer V (ICC) for the Olive Oil Dataset

Finally, for the Artificial Dataset, there are ten underlying cohorts and the Integrated Cumulative Cramer V Figure 6.13 shows that the breakdown in stability structure happens after 11 cohorts.

It is evident from the figures shown here that the ICC method identifies a region of values of k for each dataset for which stable structure is obtained. Identification of a particular number of cohorts as being correct is perhaps an ephemeral objective as partitioning algorithms such as K-Means do not always have the capability to return the underlying cluster structure, especially in the presence of heavy intermixing. It is therefore critical that a partitioning algorithm does return stable structure. The question remains as to the relative performance of this method compared to that of the other standard methods listed above, the first of which is the Gap Statistic [46].

Gap Statistic

The gap statistic uses a bootstrapping technique to build a reference dataset which is then used to determine whether the reduction in within cluster

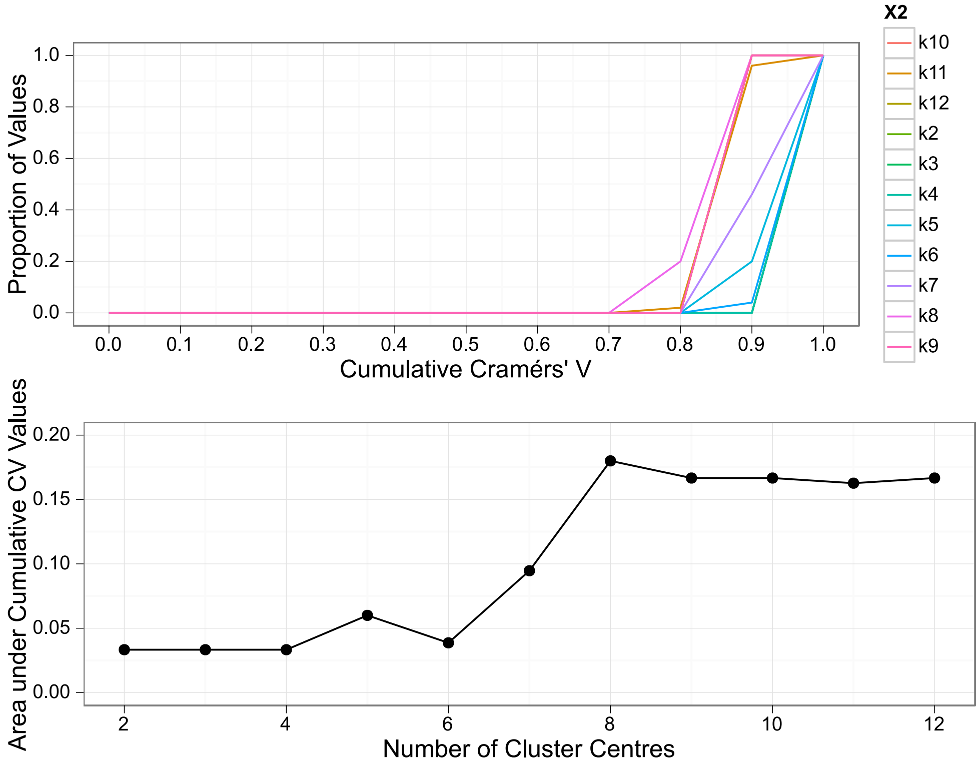


Figure 6.12: Integrated Cumulative Cramer V (ICC) for the Wine Dataset

variance is greater than that which would be expected, it is this reduction in variance which is used to calculate the gap between the expected and observed. This gap can of course be plotted.

Figure 6.14 shows the gap plots for the Artificial, Iris, Wine, Olive Oil and Cardiotocography datasets, and as can be seen there is a different profile of results to that of the ICC plots. The most striking of the five plots is that for the Artificial Data (6.14a) where there is a very noticeable early peak in the data at $k = 4$.

This peak implies that the optimal value of k for the artificial data is four however it is already known from experimental results that very good recovery of the data occurs at $k = 8$, so this is a significant misdirection for the user. Equally for 6.14b, 6.14d, 6.14c and 6.14e it can be seen that there is by contrast to the Artifical dataset no peak at all, and the results tend upwards indefinitely, this implies that the clustering algorithm is incapable of producing a clustering result that produces better than expected results, and that therefore there is no optimal value of k. This misses the richness of the data which is exposed using the framework approach with the SeCo and ICC plots.

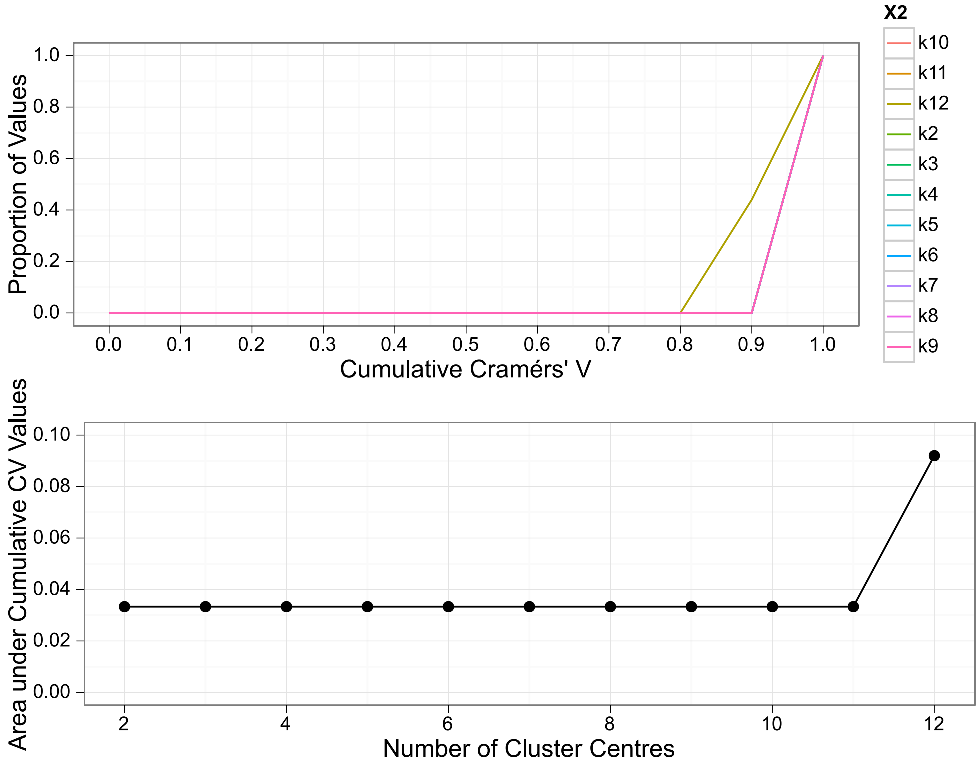


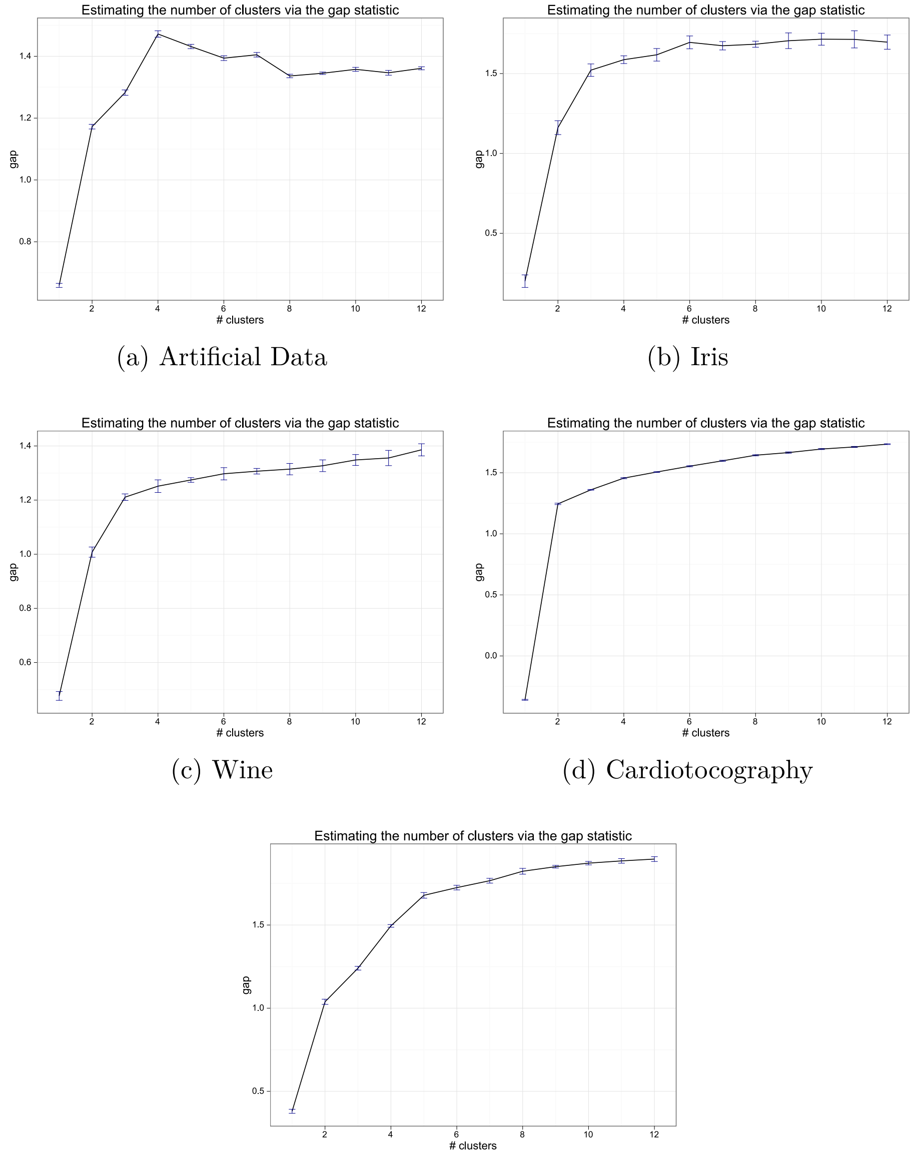
Figure 6.13: Integrated Cumulative Cramer V (ICC) for the Artificial Dataset

The Gap statistic does correctly identify the number of cohorts for the Iris and Wine datasets, however these two are the least noisy datasets used. Where there is substantial mixing of the cohorts or where the assumptions normal structure within the data do not hold it is not able to correctly identify the structure.

It can be seen from these results therefore that the use of the Gap Statistic as an indicator for evaluating an optimum value of k does not necessarily result in an appropriate value being selected. Given that this is a fundamental requirement for the use of a partitioning algorithm, the determination of the number of partitions, it would seem that at least with respect to these datasets, the utility of the gap statistic is in question.

Other methods

CLEST is perhaps the best performing alternative measure for these datasets; it identifies the correct number of cohorts for the Iris dataset and identifies three for the cardiotocography dataset – the number of foetal states.



(a) Artificial Data

(b) Iris

(c) Wine

(d) Cardiotocography

(e) Olive Oil

Figure 6.14: Plots of the Gap Statistic

For the Wine, Olive and Artificial datasets it is unable to identify the correct number of cohorts. For the latter two datasets this is likely because assumptions made about the data do not hold.

Perhaps the most interesting result from the other methods is for X-Means which identifies the correct value for K only for the Wine dataset. It does however identify 8 cohorts for the Iris data and as has been shown above there is stable sub-structure within the Iris dataset when splitting the cohorts. For the other datasets however the assumption of Gaussian structure means that when there is strong mixing between the cohorts or for data where assumptions of normality for the variables do not hold then the method performs poorly.

The remaining measures are internal validity indices looking to evaluate the separation of the partitions, so for data where the assumption that the cohorts are Gaussian in nature does not hold, they do not correctly identify structure.

Of the eight methods proposed here for evaluating the number of clusters within a dataset, the best performance was that of the CLEST method, in that it correctly identifies the underlying values in two of the five cases. The Gap statistic performed similarly but for different datasets.

The remaining methods, X-Means, Dunn Index Silhouette, Calinski-Harabasz metric and the Davies-Bouldin metric each managed to correctly identify the correct number of partitions in only one of the cases. However in every case the ICC method has incorporated the true solution as viable and in the case of the artificial dataset the only option offering a sensible representation of the data.

	Iris	Olive	Cardiotocography	Wine	Artificial
Integrated Cumulative Cramer V	9	9	10	4	11
Calinski-Harabasz	3	5	2	2	4
CLEST	3	5	3	5	4
Davies	2	5	12	3	4
Gap	3	6	9	3	4
Silhouette	2	4	2	3	4
Dunn Index	2	5	10	12	3
X-Means	8	13	15	3	5
TRUE	3	3 / 9	3 / 10	3	10

Table 6.2: Results for different methods of selecting k

Looking at each of the datasets in turn, for the Iris data the underlying structure is of three partitions relating to the three types of plant, the best performing of the methods on this data were the Calinski-Harabasz, CLEST and GAP methods, the Silhouette, Dunn and Davies methods all under predicted and selected two partitions, X-Means highlighted eight as being the correct number of partitions with the ICC method highlighting up to six, or possibly nine solutions as being appropriate.

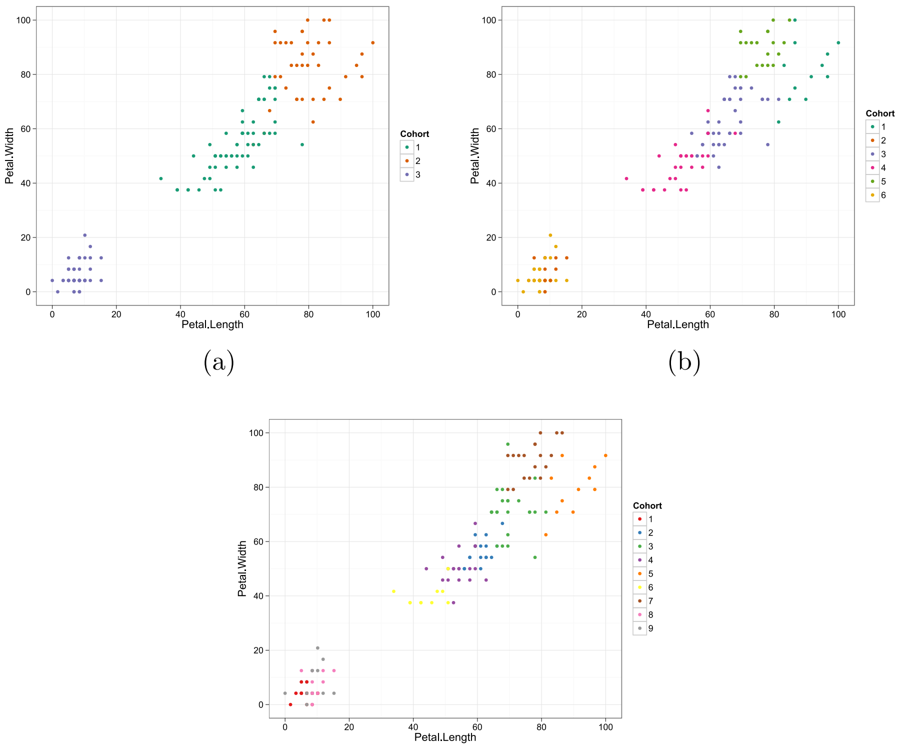
For the Olive Oil dataset, none of the methods other than the ICC highlighted the correct number of solutions as being appropriate, with the Gap and X-Means methods being the furthest out in terms of their result. Again in this case, the ICC method highlighted up to nine solutions as recovering good structure and warranting further investigation.

For the Cardiotocography dataset CLEST was able to identify the smaller three cohort solution as being realistic, the Dunn and the ICC methods were able to identify the ten partition set as being appropriate. None of the other methods were able to correctly identify either the trivial three cluster or more complex ten cluster solutions, the X-Means, Davies and Gap methods performed particularly poorly.

For the Wine dataset, four of the methods were able to identify the correct solution, the Davies, Gap, Silhouette and X-Means methods. The Calinski-Harabasz method identified two solutions as being appropriate, CLEST highlighted five with the ICC method highlighting solutions up to five as being of interest.

Finally for the artificial data, none of the methods other than ICC came close to identifying the correct number of cohorts (ten), with this latter highlighting solutions up to eleven. That the method highlights solutions up to eleven is not a problem, as the method is not proscriptive in its answer, it points the user to those values for which stable structure is highlighted and for which further investigation is necessary. An example of this is to look at the Iris data again, given that it can be easily plotted in two dimensions using the Petal Width and Length, as shown in Figure 6.15.

This plot shows that for the three cohort solution (6.15a), there are well defined cohorts, splitting each of these in two shows (6.15b) that although there is some mixing at the borders, each of these three cohorts has been split into two. Further to this, splitting the original three into a further three parts each (6.15c) gives us another clear split into three, albeit showing some small mixing at the borders as this is a projection from four into two dimensions. This shows that although the ICC does not give a definitive answer for a given dataset, it does allow the user to infer which solutions are consistent and allows for further investigation to be performed to evaluate these solutions individually.



(a)

(b)

(c)

Figure 6.15: Iris data in two dimensions (a) Three Cohorts, (b) Six Cohorts and (c) Nine Cohorts

Summary

Presented in this section were results from a SeCo map derived method of identifying useful values of k and seven alternative methods for identifying the same value. Of these methods some performed better (CLEST, Gap), however even these managed to indicate the correct number of clusters in only two of the five datasets tested upon. Compared to these results was the ICC method which was able to correctly identify the range of results in which the true solution fell in every case. For those users looking for a definitive answer as to which value is the best the results here have highlighted that on complex real world data it is often not possible to approach the problem algorithmically. Rather it is better to use a more considered approach and evaluate each solution on its own merits.

The complexity of some of these datasets does allow some forgiveness in terms of accuracy, for example with the ten cohort solution for the Artificial data it is known from the generating functions that there is by design a substantial intermixing of data points. Under such circumstances it is to be expected for the results of these methods to perhaps return a lower number of cohorts as being appropriate. However in this case none of the other methods were able to identify more than 5 cohorts as being appropriate when it can be seen from a simple visualisation of the data that more can be recovered.

K-Means Framework 7

The use of a framework approach using two measures to find stable solutions first requires that these measures be defined and the results compared to see whether the application of this method to K-Means actually provides a tangible benefit when compared against the standard approach of optimising the objective function. K-Means is often defined as being the algorithm in Hartigan [10] but there are variations on this including the earlier Lloyd's algorithm. Because the online update stage within Hartigan is an additional variance minimisation step it is possible that it coerces the final solution into a poorer local minima it is interesting therefore to consider both algorithms within the framework and evaluate their performance against each other.

7.1 Separation Measure Comparisons

Prior to comparing the performance of the different K-Means algorithms within the framework the choice of an appropriate separation measure is necessary. Applying the framework to test datasets and comparing the resultant best partitions against the reference partitions and against each other. This was done using the Cramer's V statistic.

The experiment is to investigate the performance of the framework using K-Means against the use just a single measure, i.e. just using a separation measure for solution selection. For each k and for each test dataset, the partitions were compared against the reference partitions, and tabulated as shown in Table 7.1 for the synthetic data.

Looking at Table 7.1 the performance of the SeCo framework (SSE/Median

CV) does not appear to be greater than the other methods, including the single measures in the case of the synthetic dataset. For this set of results here, it can be seen that although for the Invariant J/CV combination the dual metric approach is equivalent or better when compared to the other methods for this data. For the SSE/CV combination it is only better once at $k = 10$ and equivalent on all other occasions.

No. Partitions	Best Median CV	Invariant J / Median CV	Invariant J	SSE / Median CV	SSE
2	0.9007	0.9082	0.9888	0.9082	0.9094
3	0.9632	0.9632	0.9632	0.9632	0.9632
4	0.9174	0.9286	0.9286	0.9174	0.9174
5	0.8773	0.8773	0.8773	0.8773	0.8773
6	0.8891	0.8908	0.8787	0.8893	0.8893
7	0.888	0.9182	0.9095	0.8887	0.8881
8	0.917	0.917	0.8505	0.8995	0.8995
9	0.8598	0.8576	0.8063	0.8571	0.8587
10	0.7709	0.8485	0.7644	0.8485	0.8399

Table 7.1: Comparison of the selection criteria, dual measure and single measure, for selecting a partition of the artificial dataset. This shows the Cramers' V statistic for the solution compared to the reference partition. There are five different measures compared. Best Median CV - comparing all solutions generated and selecting the measure with best Median CV. Invariant J & Median CV - using two measures to select a solution. Invariant J - using the Invariant J criterion alone. Finally best Total Within Cluster SSE and the Best SSE & Best Median CV.

In most cases the difference between the different methods of selecting a result is slight such as for $k = 5$ there is no difference irrespective of which method is chosen with the same solution picked. For $k = 4$ in the case of the Invariant J measures the dual metric approach gives the same value as the single and the same for the SSE and SSE/CV solution sets, picking the same solution despite thresholding being used. It can also be seen that using the stability measure on its own gives good results for low numbers of partitions, performing comparably with the SSE/Median CV solution, however as the number of partitions increases this measure performs less well.

This pattern is repeated for the Olive Oil and Biganzoli datasets which give similar findings to those shown above. In the case of the Biganzoli dataset the SSE/CV values were always equivalent to the SSE, and for the Invariant J/CV they were sometimes worse and sometimes better, with a similar pattern being seen for the Olive Oil dataset.

In this case it makes sense to use the SSE of the clustering as the separation criteria, particularly given that this is the function which is being optimised by the algorithm.

7.2 Separation Thresholding

Profiling the solutions for a given set of clusterings and looking at the SSE for each solution it is clear that a fair amount of variation exists in terms of the values obtained, for 8 partitions with the Synthetic Data the mean total SSE is 671.6 with a standard deviation of 27, the histogram of which can be visualised in Figure 7.1. Those values which have lower values of SSE can be considered to be better separated solutions but including them in the pairwise comparisons for concordance with the poorer separated solutions could potentially hide these as they could be dominated by the less well separated solutions, and this would defeat the purpose of the framework.

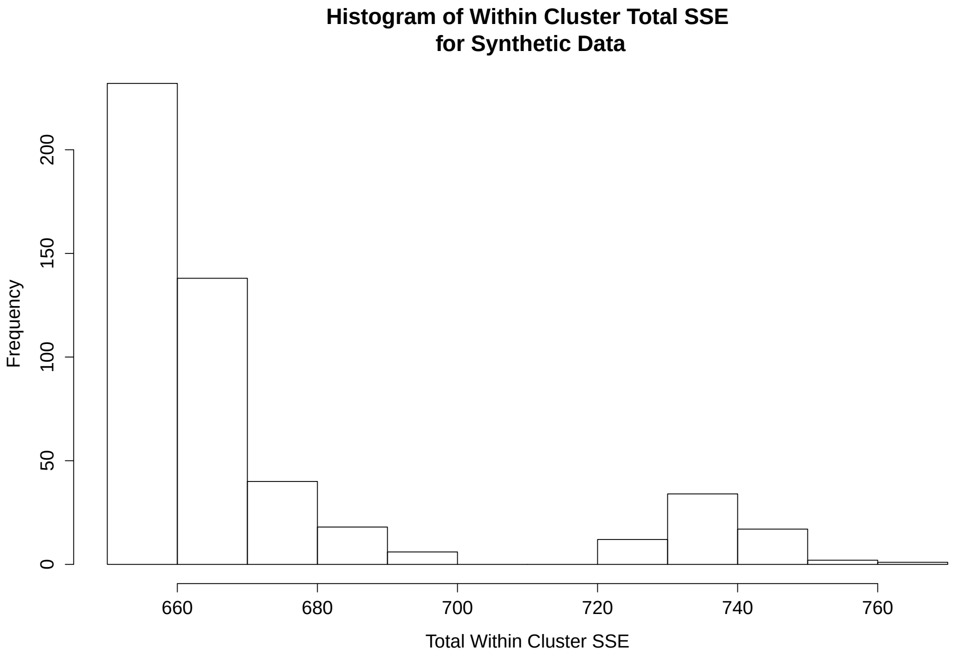


Figure 7.1: Histogram of the Total Within Cluster SSE for 500 runs of K-Means on the synthetic data.

It makes sense therefore to limit the solutions evaluated to those which are known to be well separated, but that still leaves the question at what point should solutions be considered inadequate for inclusion.

By inspection of the results in Table 7.2, it is not immediately obvious which threshold provides the best outcomes. By manually evaluating the results it is possible to select optimal thresholds for each dataset and value of k however this could not easily be done algorithmically and to do so manually would defeat the purpose of a frameworked approach.

Directly following on from this the usefulness of the threshold was considered, in particular answering the question of whether the thresholding level actually makes a substantial impact upon the results, or whether it is of secondary concern to simply thresholding or not.

In addition the performance of each value of k and each set dataset was considered in conjunction with variations in the threshold using values from 0.1 to 1.0 in increments of 0.1.

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
k2	0.9082	0.9082	0.9007	0.9007	0.9007	0.9007	0.9007	0.9007	0.9007	0.9007
k3	0.9632	0.9632	0.9632	0.9632	0.9632	0.9632	0.9632	0.9632	0.9632	0.9632
k4	0.9286	0.9286	0.9286	0.9286	0.9286	0.9286	0.9286	0.9181	0.9181	0.9174
k5	0.8773	0.8773	0.8773	0.8773	0.8773	0.8773	0.8773	0.8773	0.8773	0.8773
k6	0.8908	0.8811	0.8811	0.8817	0.8817	0.8918	0.8918	0.8891	0.8891	0.8891
k7	0.9182	0.9182	0.9182	0.9182	0.9242	0.9242	0.9242	0.9133	0.9199	0.888
k8	***0.917***	***0.8995***	***0.9034***	***0.9196***	***0.9183***	***0.9291***	***0.9291***	***0.9324***	***0.9115***	***0.917***
k9	0.8576	0.9076	0.8633	0.8691	0.8992	0.8402	0.8402	0.8598	0.8598	0.8598
k10	0.8485	0.8399	0.8404	0.7947	0.7959	0.7959	0.7933	0.7701	0.7823	0.7709

Table 7.2: Concordance with Reference Partition for each value of k and each threshold

The results shown in Table 7.2 show concordance with the reference partition for each of the solutions with a given k and threshold value, these values were based on using the dual measure approach. The results indicate that the selection of a particular threshold has made a difference to the performance of the framework however at this point it is arbitrary whether increasing or decreasing any particular threshold to the next point yields an improvement in performance.

For $k = 10$ there is the best indication that setting a threshold has a useful effect but for $k = 5$ there is no effect seen at all. This was the point at which Invariant J started to be no longer considered to be a useful measure for selecting the best solution when compared with SSE.

Essentially for a given dataset where the underlying structure is not known and it is not possible to tune the threshold on a per k basis, it becomes hit and miss as to whether the threshold chosen helps or not. Investigating the distribution of the Invariant J and SSE values for the purpose of looking at whether there was a way to algorithmically select a good threshold for each dataset provided the impetus that lead to looking at overall concordance of solutions.

7.3 Overall Concordance of solutions

By looking at the data from a new perspective by visualising the complete set of results from one iteration of the framework rather than multiple runs it is possible to view the solution space as a distribution rather than as discrete solutions. Using the 500 solutions for each particular value of k a comparison of each set of solutions is made against the reference partition. Having compared each solution against the reference partition it is possible to order them according to the selection criteria, in this case SSE or SSECV and view how the concordance behaves.

Looking at the SSE, Invariant J and SSE/CV and Invariant J/CV values, for a selection of k (k in 8,9 and 10), resulted in a plot such as that in Figure 7.2, where the same data is used for $k = 10$, to investigate what the profile of concordances against the reference partition looked like.

Figure 7.2 shows in the top panel the values ordered by SSE alone, here the best performing solutions are to be found below the top 20% of the best SSE values where solution selection would naturally occur. In the second panel which shows the same solutions ordered by the SSE/CV metric the left hand edge performance has improved as some of the better solutions have moved towards the selection edge.

It is evident from this that applying the stability measure increases the overall performance of the algorithm by $\approx 5\%$, as the value at the left hand edge would indicate. Following on from this positive results for the SeCo approach an investigation into how the algorithm performs when applying the results of the clustering to a new set of datasets was initiated.

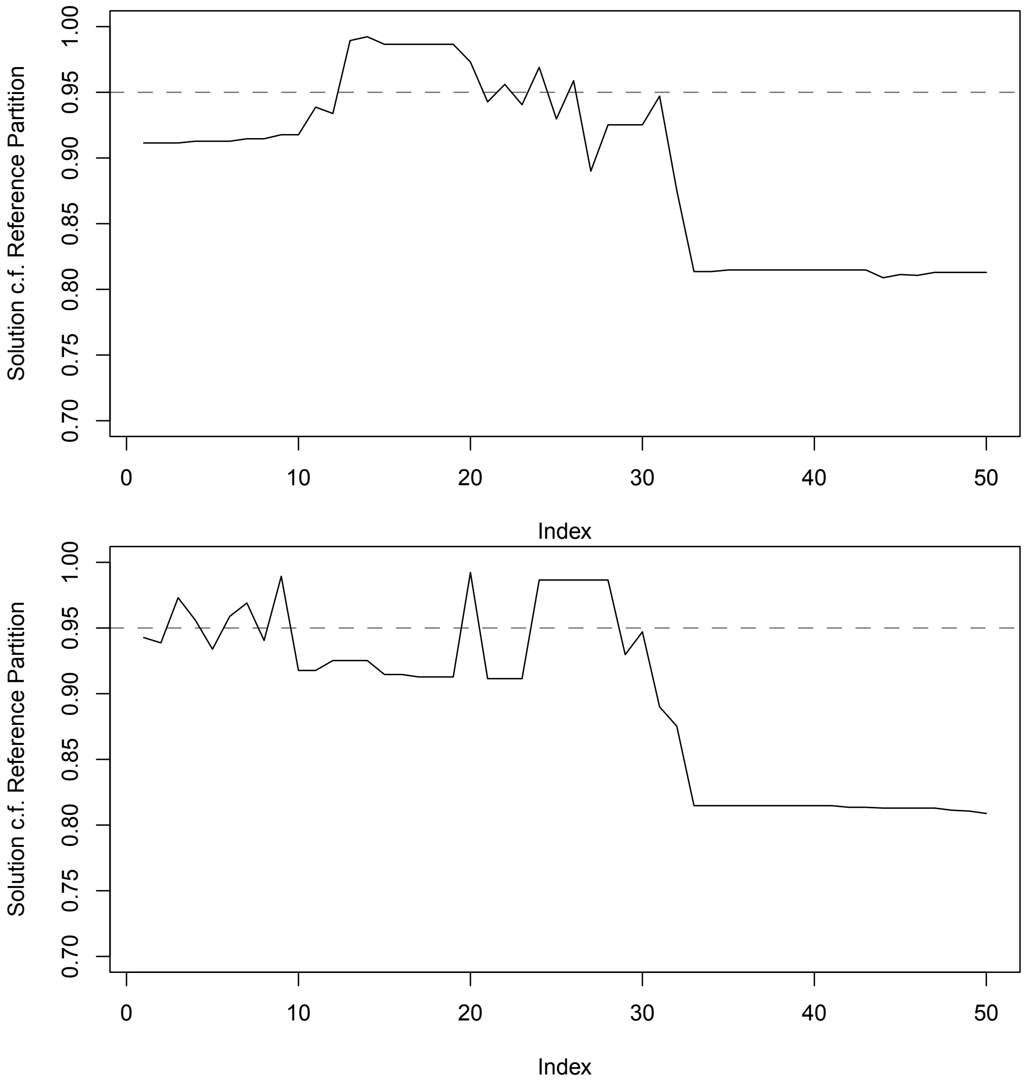


Figure 7.2: Concordance of top 50 solutions for a single set of reinitialisations, top panel indicates solutions selected using SSE alone, bottom panel shows the same solutions reordered through the use of a stability measure, a reference line is shown at 0.95 concordance

7.4 Algorithm Comparisons

Evaluation of the effect of different algorithms on the SeCo Maps was investigated by applying the two algorithms, Lloyds Algorithm 1 and Hartigan Algorithm 2 to the same data using the same initialisations to ensure only the algorithmic difference is evaluated. The process was repeated for each of the real world datasets and the Artificial dataset.

The choice of which K-Means algorithm will clearly make a difference when using the SeCo framework to wrap around it, so it is necessary to first understand how the algorithms affect the results obtained and which algorithm is therefore the most appropriate to use. Hartigans algorithm uses an online update phase to better locate a local minima, the questions to be answered though are all these minima the same and does this online update phase perform better within the framework?

Synthetic Dataset

Looking first at the Synthetic Data there are two maps, one for each of the two methods, as shown in Figure 7.3 and Figure 7.4. A cursory glance might indicate there is no substantial different between the two maps which is certainly the case for lower values of k however at $k = 7$ and above Lloyds method shows greater instability in the concordance values.

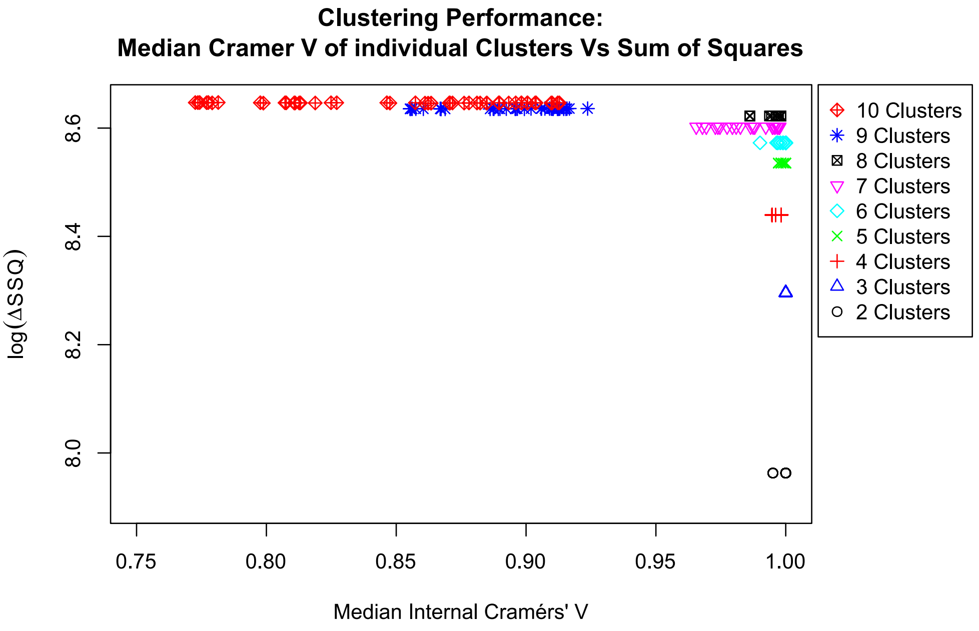


Figure 7.3: SeCo map using batch K-Means and SSE:ArtData

This is to be expected given the online update method adjusts the allocation of the clusters additionally to reduce the SSE so it has prior optimisation towards a given set of solutions. For this reason solutions produced using the online update method have naturally better concordance than those which are using the other method.

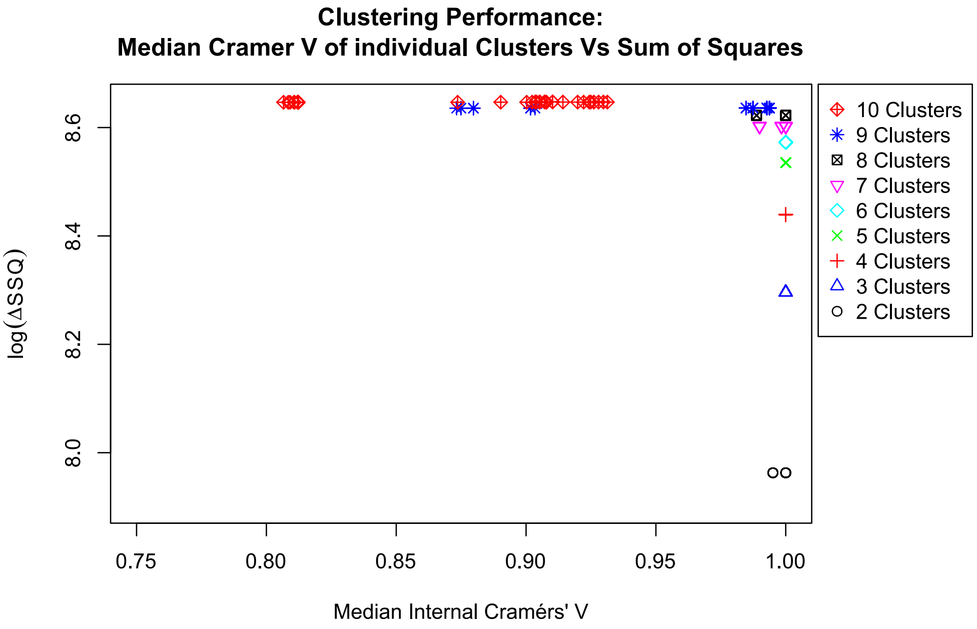


Figure 7.4: SeCo map using online update K-means and SSE:ArtData

Olive Oil Data

Looking comparatively at the Olive Oil data maps with the online and batch algorithms, Figure 7.6 and Figure 7.5 respectively, it can be seen that the batch algorithm performs reasonably well for values of k between 2 and 7 where the results show high concordance for values with similar SSE. At $k = 7$ however, the concordance of the solutions starts to degrade and at $k = 9$ there are shown to be a wide range of solutions whose concordance with each other is diminishing.

For the online update algorithm however a different pattern is seen where the solutions show high concordance for all values of k up to $k = 9$ and with the solutions tightly clustered around the right hand edge of the plot. At $k = 10$ and beyond there is a decline in performance of the solutions from ≈ 0.99 median concordance to between 0.8 and 0.9.

These results are interesting particularly as the underlying structure of the data is already known for the Olive Oil with two ways of partitioning into either Areas or into Regions. The three regions should classify properly every time but for the batch methodology is shown from the SeCo map to not necessarily map the nine areas well. The SeCo map in Figure 7.6 shows that the online update however is able to group the data in a stable way.

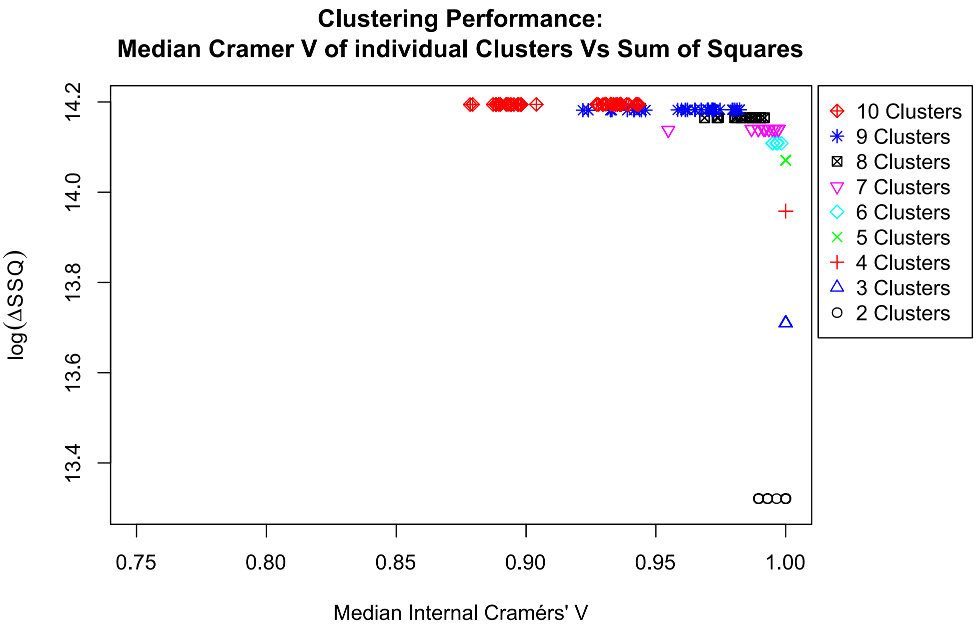


Figure 7.5: SeCo Map using Batch K-Means and SSE: Olive Oil Data

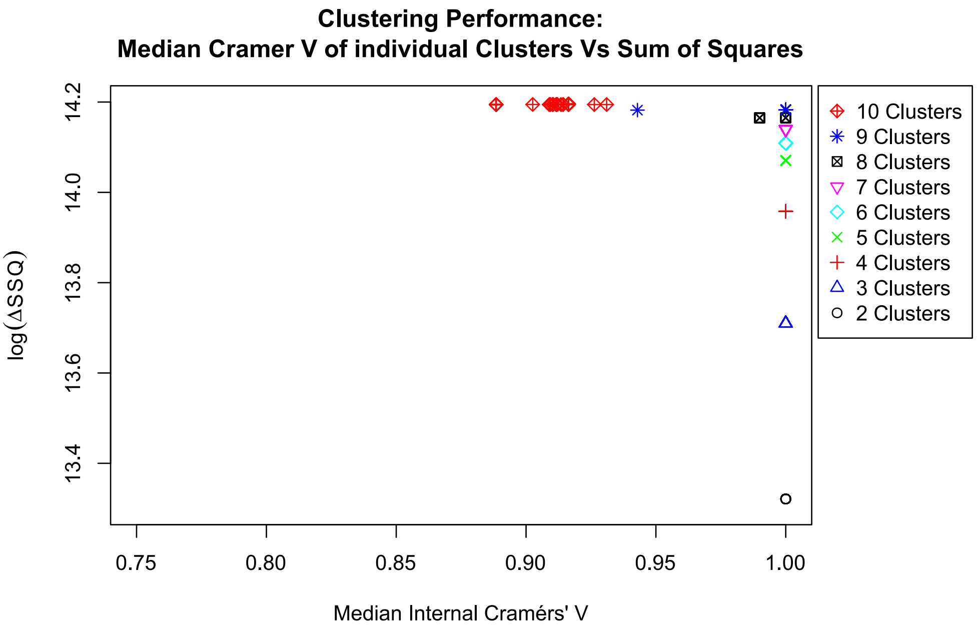


Figure 7.6: SeCo Map using Online K-Means and SSE: Olive Oil Data

Roundup

The results for these two datasets indicate that the use of the online update increases the stability of the algorithm in selecting a solutions and its addition makes it more likely that repetition will yield a good result. This much is evident from the SeCo maps shown in Figure 7.3 to Figure 7.6.

However given that the decline in the concordance of the solutions as k rises it is likely that there will some variation if the process were to be repeated.

7.5 Benchmarking

Having decided upon the level of thresholding, algorithm and separation measure it was necessary to evaluate the reproducibility of the framework given the possibility for variation existing within the chosen solutions. To test this the framework was applied repeatedly to the datasets and compared to the reference partitions to allow comparison of the solutions with the reference partitions.

Applying the framework to the synthetic data 100 times, the best solution as selected by the framework and the best solution from the objective function were selected and then compared to the reference partitions. In this instance different initialisations were used for each run of the algorithm as the purpose of this test was to evaluate the repeatability of the method.

This resulted in nearly 2000 solutions all of which could be compared with the results visualised by histograms in Figure 7.7 and Figure 7.8 of the concordance of each of the solutions with the appropriate reference partition. From this it can be seen that there is little difference between the two for solutions in the range $k = 2$ to $k = 7$ however at $k = 8$ and higher we start to see a change in the distribution of the histograms.

For $k = 8$ there is a spread of values for those solutions selected using just a single measure but for the dual measure approach the range is narrower. On the face of it the performance of the former method appears to be a little better however the bins which are mostly populated are adjacent so the difference is minimal.

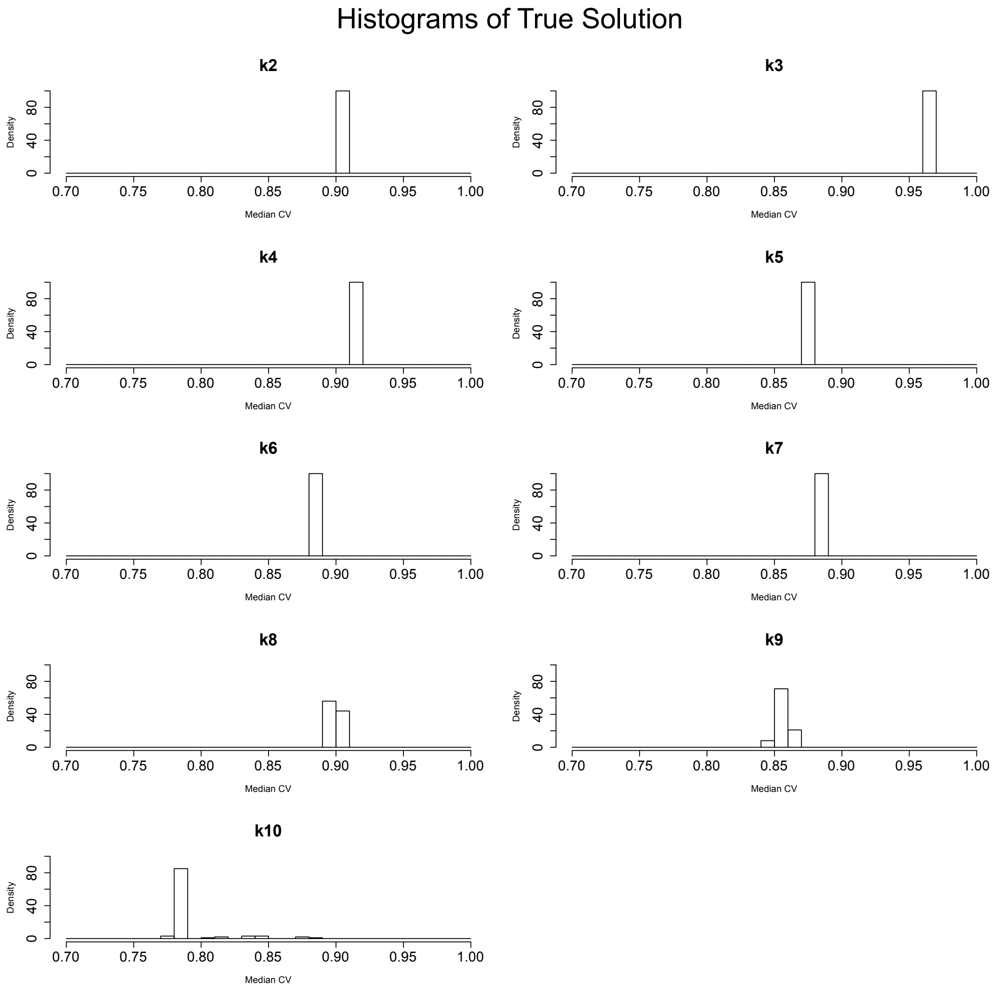


Figure 7.7: Histograms of results from 100 runs of the framework, with solution for best SSE selected from each, compared to reference partition.

Moving to higher values of k again there is a reduction in the range of the values however looking at $k = 10$ there is a substantial improvement in the performance of the dual measure approach. This is notable given that for the Synthetic Data which is used in this case the underlying number of partitions is 10.

So not only is there a reduction in the range of the values but there has been a shift of the values to the right side as overall median concordance has improved. This is good as it means for this particular value of k the framework approach classified better than does the single measure approach. It is important to note however that for lower values of k the results are inconclusive.

These results are particular to the use of the objective function in the

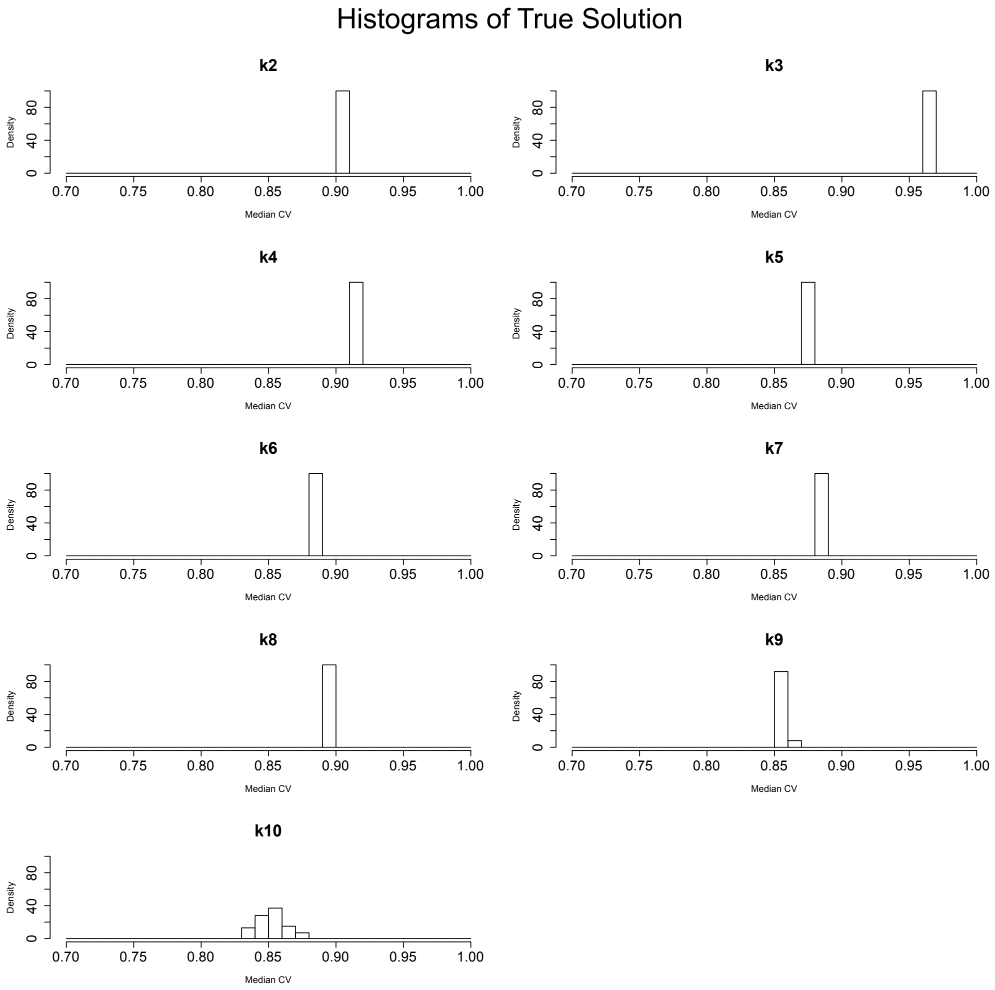


Figure 7.8: Histograms of results from 100 runs of the framework, with solution for best SSE/CV combination selected from each, compared to reference partition.

selection of a best partition however using a different measure for the separation can result in a different conclusion. For example as discussed earlier the invariant J measure could be used as the separation metric producing histograms of the solutions as shown as in Figure 7.9.

Examination of this in combination with previous figures (Figure 7.7, Figure 7.8) shows that Invariant J selects better solutions from 100 runs than do the SSE measures in particular looking at $k = 8$ and it can be seen that the solutions have a higher concordance than for the other measures, as do the $k = 9$ and $k = 10$. However these last two have a wider spread of concordance indicating different solutions are being selected.

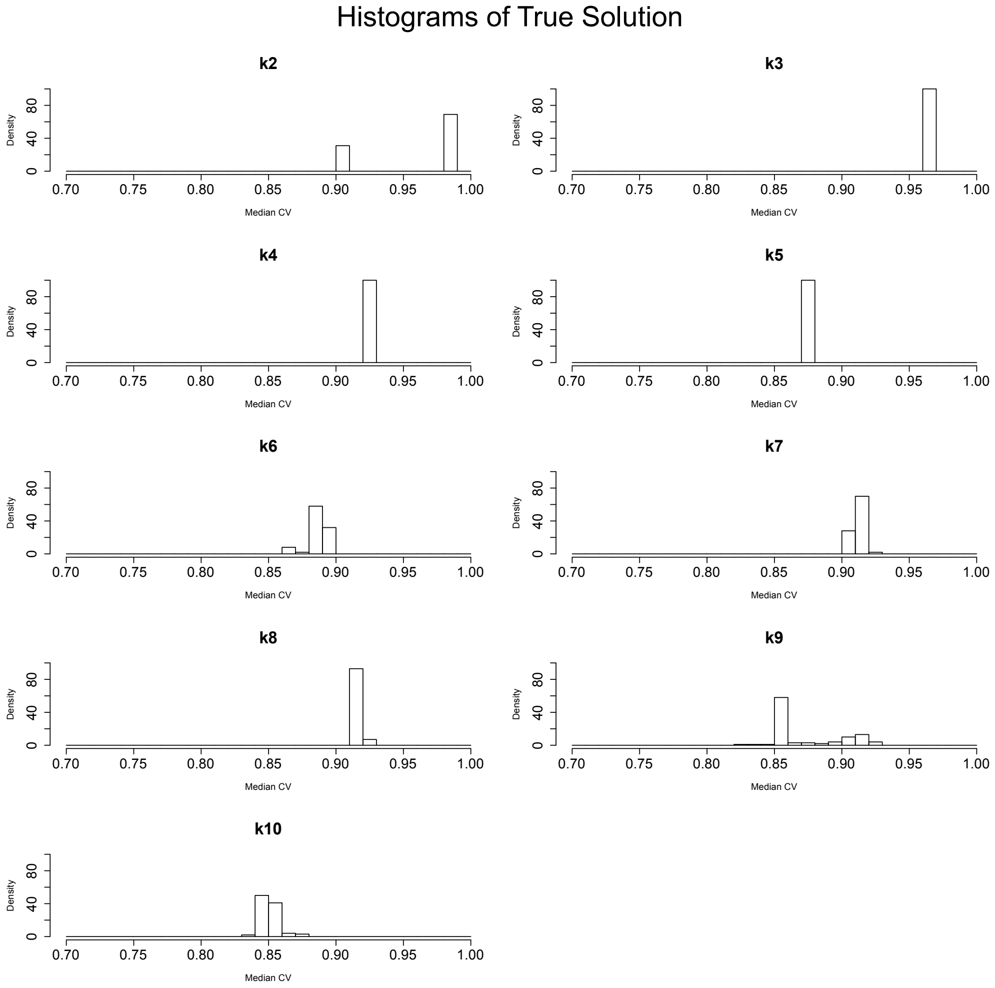


Figure 7.9: Histograms of results from 100 runs of the framework, with solution for best IJ/CV combination selected from each, compared to reference partition.

7.6 Testing of obtained solutions against new data

Ten new Synthetic Datasets were generated such that the solutions from the existing clustering runs could be applied to this new data allowing for direct comparison of the algorithms predictive capability for new data. Should the results provide labels which are representative then applying them to new data would give good concordance with these datasets own reference partitions.

This procedure generated 100 results for each value of k so if the method is appropriate for predicting the likely cohort for a data point then a good concordance would exist between the two. Visualising this through box plots it is expected that a narrow range would indicate good predictive capability. However comparing the SSE to the SSECV solutions showed that selection using a dual measure approach did give an improvement in the concordance, for the interesting values of k it was within 95% confidence intervals thus not significant given the range of values in response.

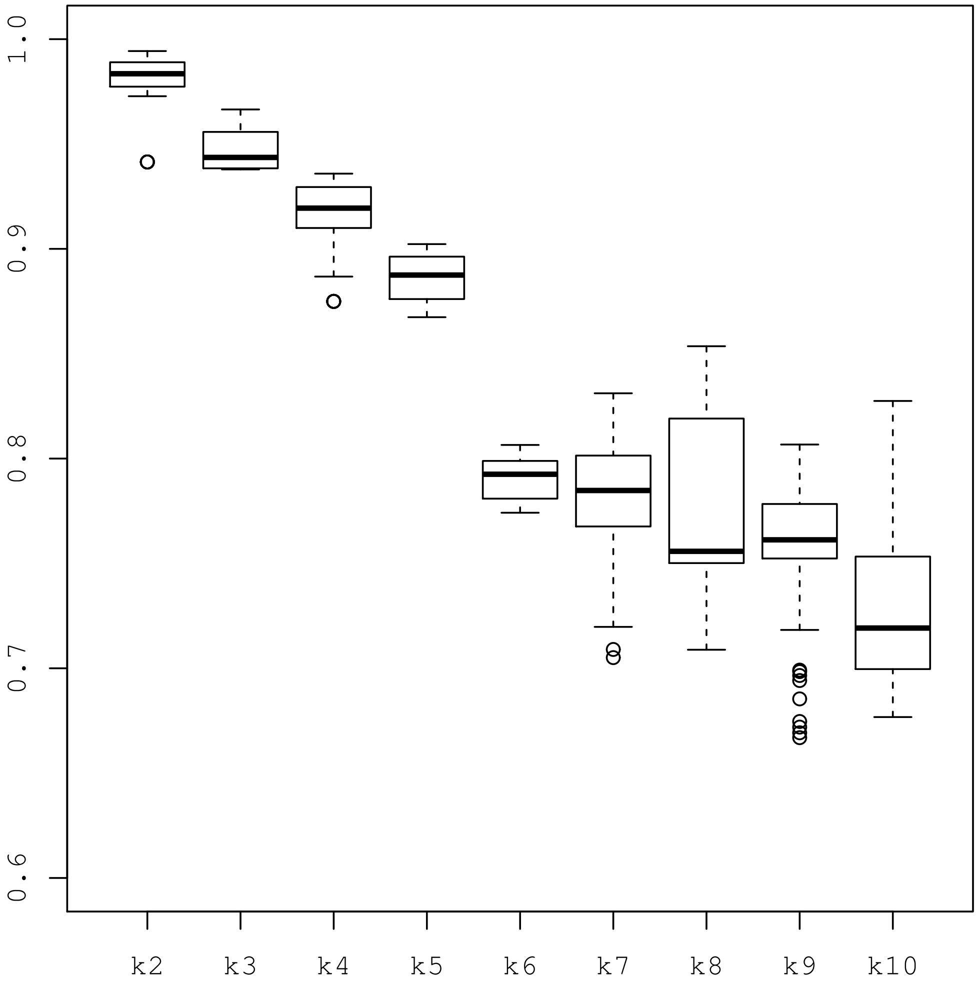


Figure 7.10: Top 10 values compared to Ten new datasets for SSE alone

Figure 7.10 shows the values for SSE and Figure 7.11 shows the result for the dual measure approach, and it can be seen here that although the spread of solutions for the dual measure approach is greater as k increases, the concordance is higher for each of the different values of k.

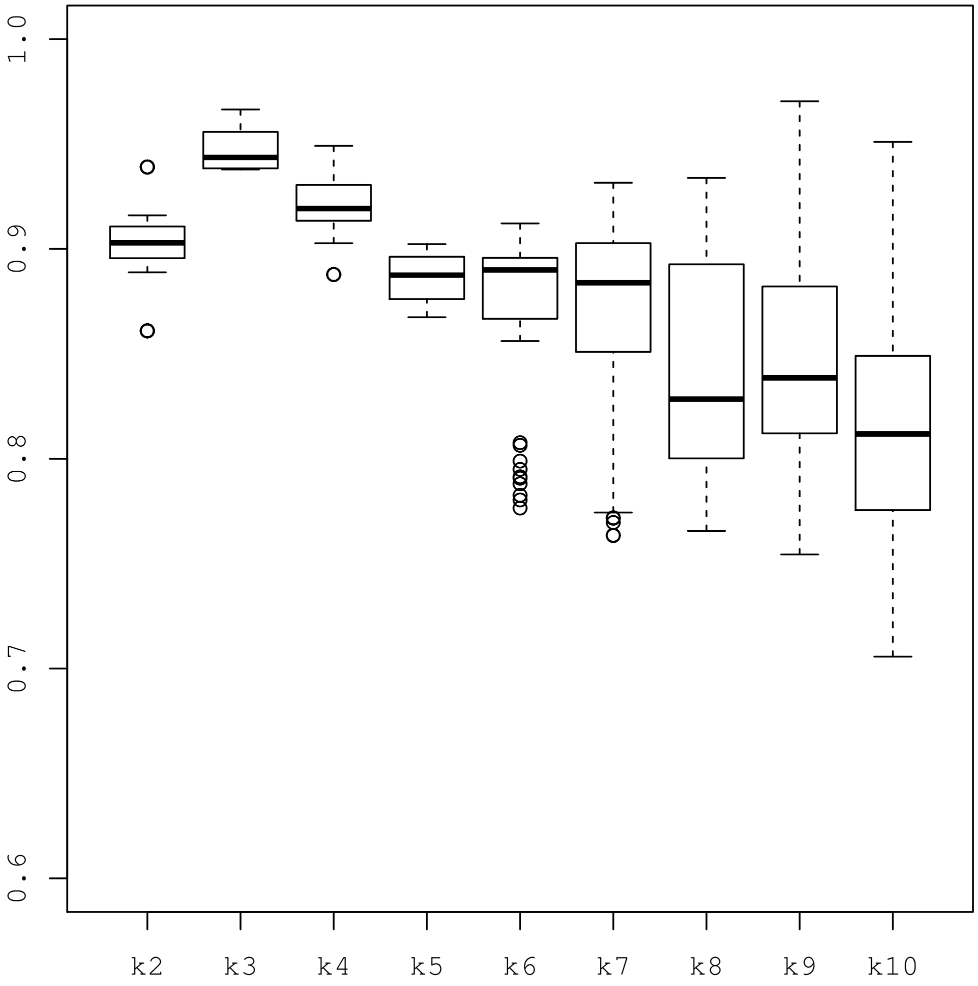


Figure 7.11: Top 10 values compared to Ten new datasets for SSE alone

7.7 High, Medium and Low density data

Finally a comprehensive benchmark was performed to compare the reproducibility of SSE as a single metric for selecting a good partition with that of the SeCo framework approach as the underlying data becomes increasingly sparse. As the previous set of results highlighted applying the framework a single time was insufficient so the framework and SSE were applied to each of the datasets 100 times.

Solutions for $k = 8$, $k = 9$ and $k = 10$ were selected and compared against the reference partitions previously generated, with the repetition allowing the reproducibility and stability of each method to be compared. Figure 7.12 shows the results of one such run for the ten thousand point dataset.

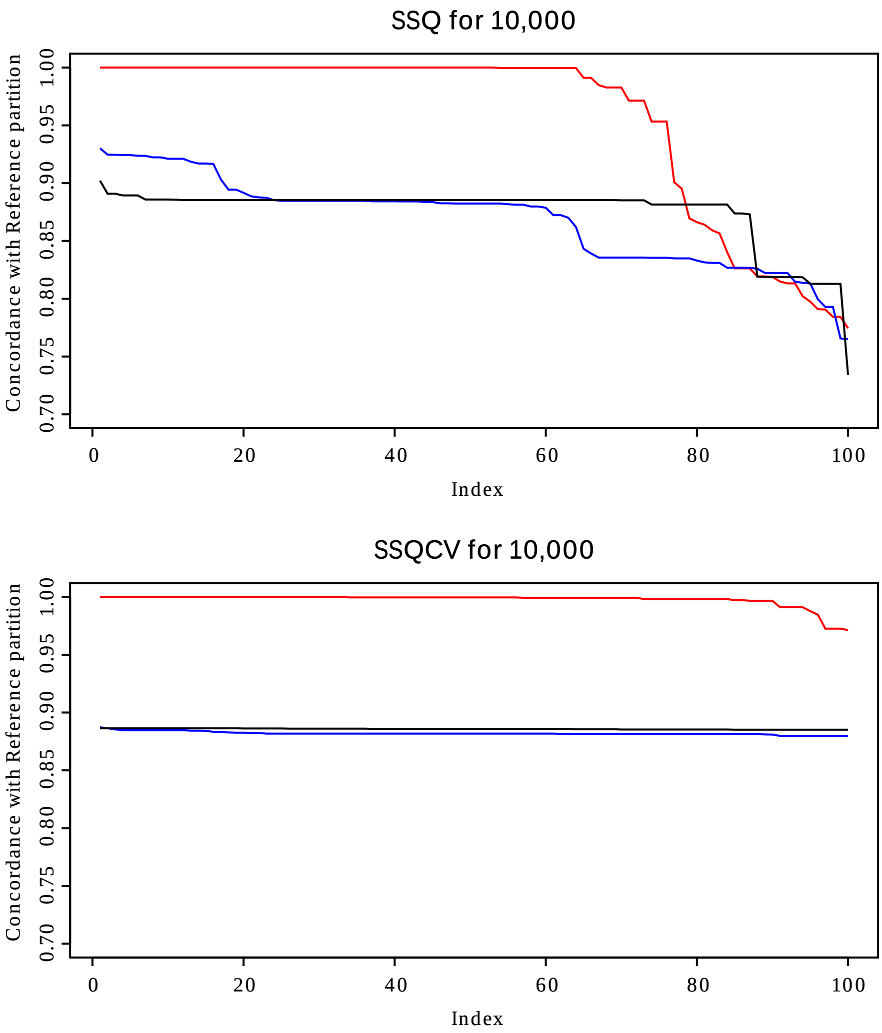


Figure 7.12: Top panel shows SSE for 10k dataset and $k = 8$ (black), $k = 9$ (blue) and $k = 10$ (red) for Synthetic Data. Bottom panel shows SSECV for the same.

These show that for $k = 10$ SSE performs well obtaining near perfect concordance with the reference partition however in approximately forty per cent of cases performance is less good and instead of having concordance of ≈ 1, it is possible for the concordance to drop to below 0.8. For k= 9 greater variation in the results occurs with the best case obtaining 0.925 in 5% of cases for approximately 30% of solutions however the concordance drops to between 0.8 and 0.85 with the remainder settling at 0.875. For $k = 8$ there is little variation in the concordance with the reference partition for most results this is expected as the SeCo map indicates that for $k = 8$ the solutions are very consistent.

Use of the SeCo Framework as shown in the lower panel gives a much different performance profile with solutions exhibiting high levels of consistency throughout. For $k = 9$ the solutions perform equally well and whilst there is no longer the higher peak of 5% of solutions neither is there the reduced concordance for 30% of results. $k = 8$ shows the same concordance as before which is to be expected and corresponds with the amount of variation between solutions indicated by the SeCo map. For $k = 10$ the results are consistent but not showing the drop in concordance for the last 5% of cases as seen before.

Current best practice of using SSE to select a single K-Means partition set from many is shown here to perform less consistently than might be expected and repeated application of this metric has significant potential to produce a sub-optimal result. By contrast using a stability measure in conjunction with SSE is shown to perform consistently and aside from a particular result the pattern is stable. Using the stability measure in conjunction with the separation measure improves the stability and reproducibility for obtaining a solution when using K-Means. In eleven of the twelve benchmark comparisons the SeCo framework performed equivalently to or better than selecting the solution with the lowest SSE alone.

7.8 Cardiotocography Data

To ensure that the results are not applicable to just the Synthetic datasets experiements were repeated using a different dataset. In this case the Cardiotocography dataset was chosen as a real world, complex bioinformatics dataset suited to the task of testing the methods.

Looking at Figure 7.13 it is clear that the results in the lower panel (SSE/CV) show substantially greater stability than those for the single measure approach alone (top panel). This is borne out in the results for the accuracy and affinity across the datasets as shown in Table 7.3 which show the results for the ten cohort solution in all cases. The mean and standard deviation of the classification accuracy for the same 100 runs above are shown along with the mean affinity.

Use of the framework should result in an expected improvement in the stability of solutions i.e. the standard deviation. This is shown to be the case for all six datasets where the standard deviation for the dual measure

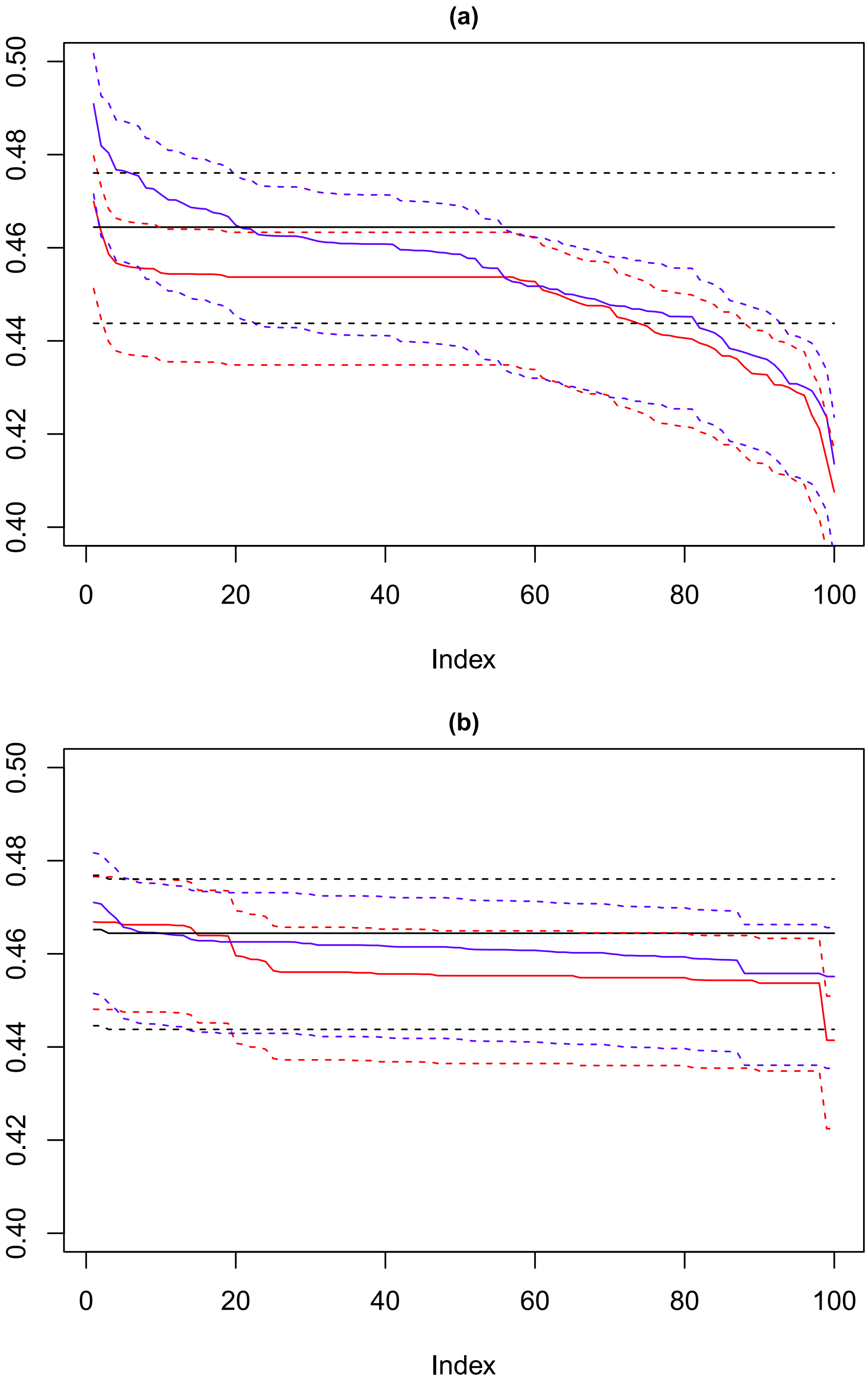


Figure 7.13: Top panel shows SSE performance for Cardiotocography dataset and $k = 8$ (black), $k = 9$ (blue) and $k = 10$ (red). (a) Single Measure (b) Dual Measure Approach.

approach is an order of magnitude lower in four of the six cases and less than half in the remaining two. Equivalent accuracy is returned in four of the six datasets with an improvement in one.

For the 1000 dataset accuracy is lower but this falls in line with previous observations reported above. Comparable affinity is observed for three of the six datasets, with substantially better results for the Cardiotocography and 2,000 Synthetic Dataset.

Current best practice of using SSE to select a single K-Means partition set from many, is shown here to perform less consistently than might be expected, and repeated application of this metric has significant potential to produce a sub-optimal result.

By contrast using a stability measure in conjunction with SSE has been shown to perform consistently and aside from a particular result, the pattern is stable, in that using the stability measure in conjunction with the separation measure improves the stability and reproducibility for obtaining a solution when using K-Means.

Dataset	**Single Measure**		**Dual Measure**	
	Accuracy (Std. Dev)	Affinity	Accuracy (Std. Dev)	Affinity
Artificial 500	0.7758 (0.038)	0.922	0.7701 (0.015)	0.931
Artificial 1,000	0.9263 (0.037)	0.994	0.7773 (0.015)	0.98
Artificial 2,000	0.7332 (0.058)	0.888	0.7345 (0.003)	0.992
Artificial 5,000	0.9079 (0.089)	0.937	0.961 (0.008)	0.989
Artificial 10,000	0.9929 (0.032)	0.993	0.9994 (0.001)	0.999
Cardiotocography	0.3655 (0.017)	0.792	0.3775 (0.002)	0.983

Table 7.3: Summary results for six datasets comparing accuracy and affinity for the single and dual measure approaches.

Adaptive Resonance Theory

8

The use of ART based algorithms within a SeCo framework to control initialisation dependence requires a different approach to the use of a K-Means algorithm as the means of initialisation is very different. The latter randomly selects prototypes from the data to use as initial conditions whereas the former is dependent upon both the order of data presentation and the ρ and α parameters, providing additional complexity. The initial approach to adapting the framework was to keep the process as similar as possible to that of the K-Means approach whilst varying the ρ parameter in a manner similar to that of k and the order of presentation in the same way that the prototypes were sampled for K-Means.

Testing of the framework was done by application to each of the Wine, Synthetic, Olive Oil and Thyroid datasets and a SeCo map being produced for each. As with previous applications of the method[51] the map indicates solutions which are of interest for the user. The maps are produced using parameters of $\rho = 0.95$, $\alpha = \frac{0.5}{sqrt(m)}$ and $\beta = \frac{0.5}{sqrt(m)}$, where m is the number of columns specific to the data set. The latter parameter is a balanced option in terms of the learning rate, with that particular alpha having a reasonably broad matching parameter. The vigilance parameter being set at $\rho = 0.95$ was a deliberate choice based on evidence that the separation of solutions being produced by a range of parameters.

8.1 Framework Approach to ART-2A

Having adapted ART-2A into the framework and applying to the test datasets and evaluating the results it was clear from Separation maps similar to that of Figure 8.1 that results were not consistent with those expected. The vigilance parameter is a proxy for k and directly affects the number of prototypes created, although this is still dependent upon the order of presentation within the data. Looking at Figure 8.1 there is some clearly counterintuitive behaviour.

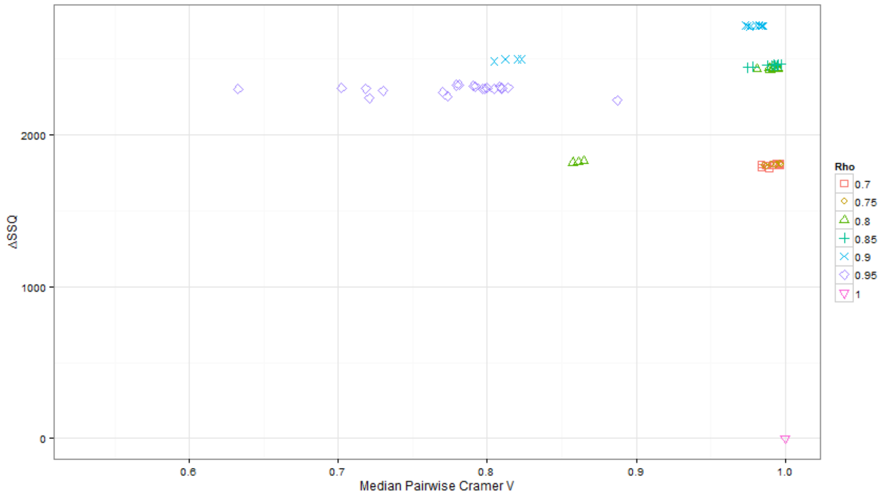


Figure 8.1: SeCo map for Olive Oil data using ART-2A separating by ρ

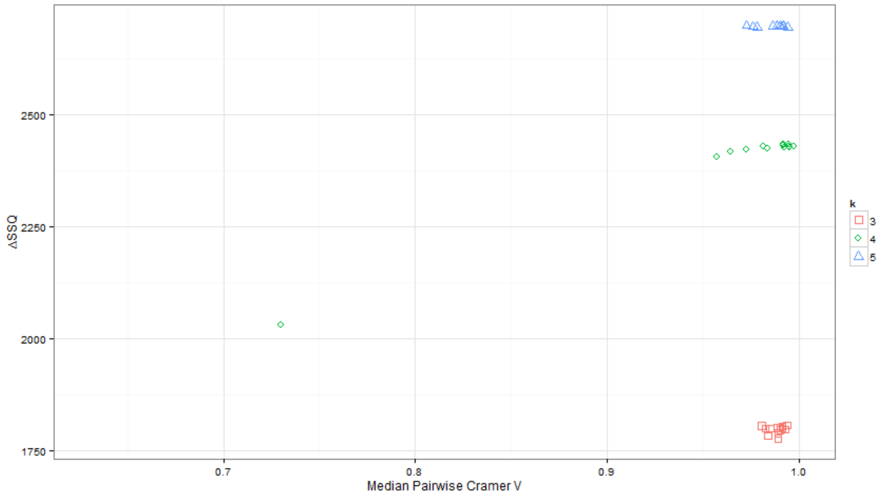


Figure 8.2: SeCo map for Olive Oil data using ART-2A separating by k

The concordance measure is indicating that as the vigilance parameter rises two things happen, there is a drop in concordance as ρ reaches 0.95 which is expected, but that the separation measure indicates less well separated clusters. As the number of partitions increases (ρ being a proxy for number of prototypes) the Total Within Cohort SSE should go down.

This indicated that ρ was not properly acting as a proxy for the value of k as might have been expected, leading to Figure 8.2 refactoring the results so the SeCo map is grouped by the underlying number of clusters irrespective of the vigilance parameter whilst still performing the remaining steps in the framework approach. Figure 8.2 matches the expected map much closer with increasing number of partitions having increasingly better separation.

8.2 Choice of ρ

The number of partitions whilst not directly controlled by the vigilance parameter ρ is strongly influence by it, increasing this value will usually but not always result in an increase in the number of partitions. The choice of an appropriate value for the vigilance parameter is of importance and strategies have been developed which attempt to effectively deal with this[43], however in the context of the framework where the vigilance parameter is varied to increase the number of partitions there will inevitably be an overlap with other solutions whose vigilance parameter is close.

When initially adapting the framework for ART-2A it was anticipated solutions with similar values of ρ would have similar concordance and separation irrespective of the number of cohorts. Given the contrary further understanding relating to the effect ρ has on number of prototypes and the performance and the evidence presented here indicates that a higher vigilance will have a large overlap of the number of cohorts in addition to yielding solutions whose Total SSE is lower even for the same number of clusters compared to lower ρ.

Figure 8.3 shows box plots of the separation measure (within cluster sum of squares) for each set of results with similar number of prototypes, within the set of solutions produced using a given vigilance parameter. As can be seen the higher the vigilance parameter it is often the case that the variance of the within cluster sum of squares is lower for solutions with the same number of prototypes and the overall separation is higher also. For the

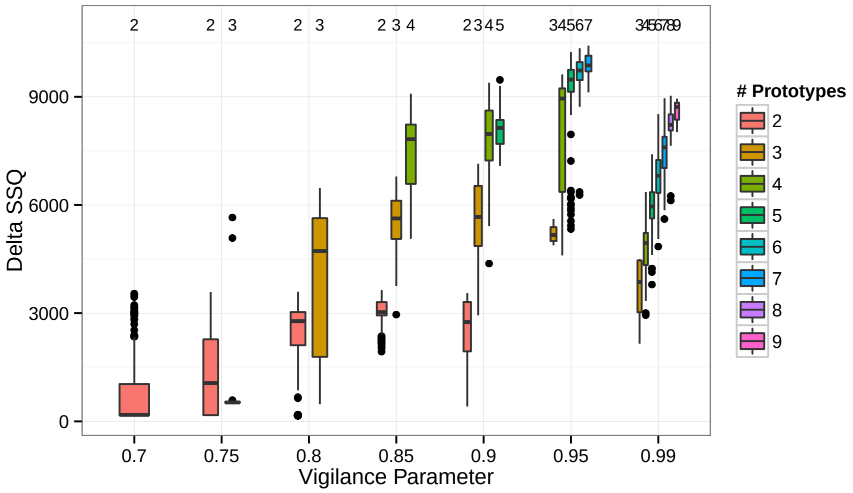


Figure 8.3: Box plots showing stability for the Synthetic Dataset using within cluster sum of squares and showing poorer clustering performance for lower values of ρ

lower vigilance parameters, the expected low number of cohorts is observed with only one or two different classes being generated for $\rho \in 0.7 \ldots 0.8$ which for the most part show that the results are broadly comparable, however the large number of outliers on these indicate that the solutions are quite different.

As the vigilance parameter is increased past 0.8 a rapid increase in the overall separation of the solutions is observed, albeit with a corresponding rise in the variability of separation, until reaching the chosen value of $\rho = 0.95$ where the solutions exhibit low variability of separation with few outlying solutions. This would seem to confirm the choice of vigilance parameter as being prudent.

8.3 SeCo Map

The SeCo maps represent stability on the y-axis in the form of the Within Cluster Sum of Squares (transformed to aid in visualisation[51]) and the internal consistency of the solutions on the x-axis. An ideal solution is one which is both well separated and self-consistent with the other produced solutions, meaning that it will be both robust and reproducible. This property allows repeated application of the framework to the same dataset to produce similar results consistently. Interpreting the map therefore means looking at the right hand edge of the plot where the solutions are most stable, and looking for a group of solutions along that edge.

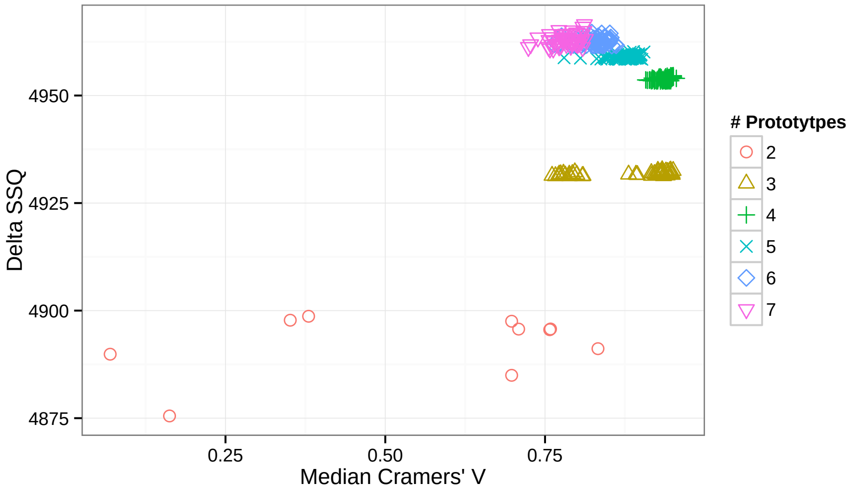


Figure 8.4: SeCo Map of Synthetic data, $\alpha = 0.7$ and $\beta = 0.5$

Using the SeCo map to evaluate the solutions generated by the ART2-A algorithm gives the plot shown in Figure 8.4 which shows the results for the Synthetic Breast Cancer dataset having six different cluster numbers $k \in 2 \dots 7$. It is known that the data was generated using an initial 10 clusters, but that it is not possible to recover the true structure as a result of a combination of sparse and mixed clusters.

Immediately it can be seen that there are two distinct sets of solutions, one for $k = 3$ and one for $k \in 4 \dots 7$, with $k = 2$ showing a grouping with poor separation and poor overall concordance. The solution set for the group $k = 3$ show a consistent block of solutions on the right hand side with another set of less consistent solutions on the left. The most interesting grouping is for those with four prototypes as these have a dense cloud of solutions on the right hand edge of the map all having the best separation, these results are stable and highly reproducible.

The remaining solutions for $k \in 5 \dots 7$ exhibit increased cluster separation but correspondingly a marked decrease in the internal stability of the solutions, this indicates that although the ART2-A algorithm is able to separate the observations into tight groupings, it is not able to do so in a robust manner; meaning that a choice of solutions here is likely to result in an unrepresentative partitioning of the data. The SeCo map therefore indicates that the best choice of solution would be that with the highest internal consistency for $k = 4$ as these have better overall separation and stability than the $k = 3$ block.

It is worth noting here that strong indications are given through the SeCo map as to the appropriate number of partitions for the data, this is particularly important in the context of the ART2-A algorithm as there is no direct mechanism, as with methods like K-Means, for specifying the number of clusters, and the differing partitions are generated using the same tuning parameters. This means that in an exploratory data analysis the map provides useful insights to the structure of the data.

Table 8.1: Summary of results for different datasets

	ρ	Framework (k)	Separation (k)
Wine	0.95	0.907 (3)	0.903 (8)
Olive Oil	0.95	0.905 (5)	0.823 (8)
Synthetic	0.95	0.862 (4)	0.748 (7)
Thyroid	0.95	0.742 (3)	0.742 (3)

Application of the method to the three remaining datasets give the summary shown in Table 8.1 where the solution chosen by the framework for each dataset is compared against the solution with the best separation overall for that dataset, using the Cramers' V of those partitions with the reference partitions for the data. The number of clusters in each solution is indicated in brackets with the Wine data having 3 underlying groups, the Olive Oil having 9 groups, Thyroid 3 groups, and the Synthetic data having 10.

Partitions generated for the Wine dataset are the most similar with both the Framework approach and the best Separated solution having broadly comparable solutions at around 0.9 concordance, this is to be expected as the Wine dataset is relatively easy to partition being comprised of three reasonably distinct groups. The next most complicated dataset the Olive Oil where again the Framework approach works well having a concordance of about 0.9 again, performing slightly better than the k-Means version of the framework on the same data which obtained a concordance of about 0.83, which is equivalent to that selected by the best separated solution by for ART2-A.

The thyroid dataset shows that the best separated solution and the solution chosen by the framework both have the same concordance with the underlying solution as each other, with both having three partitions also. Closer examination of the results revealed that both solutions produced the same partition of the data, having a Cramers' V concordance of 1. The most notable results though are for the Synthetic dataset where the dual measure approach obtained a concordance with the underlying solution of ∼0.86 with the best separated solution getting ∼0.75.

The competing best separated partition was obtained by looking at the complete result set and selecting the single best separated result from here, rather than looking at the best separated result for the same number of partitions as identified by the framework. This was a deliberate choice as in the absence of the framework or any additional measures there is no way to appropriately identify that number of cohorts as being an appropriate choice, so the reasonable approach was to simply select the overall most separated. It would not be appropriate to select from the same value of k as this is already a self selecting group of highly consistent solutions.

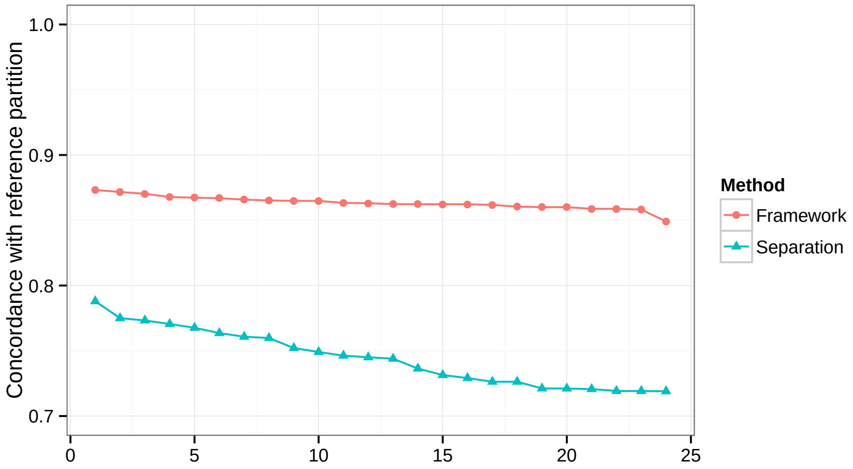


Figure 8.5: Stability analysis of 24 runs of the ART2-A framework on the Synthetic dataset

These results confirm the original hypothesis that a dual measure approach to selecting a solution is preferable to considering just a separation measure alone being equivalent or better in each case for the three datasets of varying complexity examined here. However one of the important features of the framework approach is that the solutions are robust and reproducible, so repeated application of the approach should yield results that are equivalent when comparing with the reference partitions. It is also important to comment that the solutions generated here are produced using the same vigilance parameter, despite having differing numbers of cohorts within the solution space. Without the use of a dual measure approach to evaluating the solutions there is little indication as to which distinct number of partitions best represents the structure of the data.

To test whether the method is both stable and reproducible a check was performed by applying the framework against the data repeatedly, each time selecting the best solution and comparing it with the reference partition, the results of which are shown in Figure 8.5. The concordance with the reference partition is shown on the y-axis with the x-axis, the data points simply being plotted in descending order from left to right. It can be seen here that although there is a slight drop in the concordance with the actual solutions of ∼0.02 the solutions are broadly comparable, whereas the clusters selected by SSE alone perform worse against the reference solution with more variability.

Conclusions 9

A new framework approach for initialisation dependent clustering has been presented combining two performance measures, one for intra-cluster separation and the other to measure inter-cluster stability. These guide the sampling of a single partition of the data following a process of repeated random initialisation to obtain stable reproducible results when using these unstable algorithms. The proposed method was demonstrated to work on K-Means and extended to show a more general proof of concept on ART-2A, part of the Adaptive Resonance Theory family of algorithms.

For the framework implemented using K-Means extensive benchmarking shows that the method shows both stability and reproducibility but also that the approach will give results more consistent with the underlying structure the data over repeated applications than other methods.

9.1 K-Means

Exhaustive experiments have evaluated every aspect of the performance of the framework approach when compared with the results from the standard approach to K-Means clustering showing that mechanistic application of the framework results in high-quality repeatable solutions. It has been shown that for each value of k the solutions produced show a high degree of association internally, ensuring that a repetition of the framework using the same dataset is highly likely to return a solution with high concordance to those previously obtained. This is especially relevant for situations where well-separated cluster partitions can show weak association consequently showing poor concordance among clusters with high values of the separation index.

To test the results were consistent with the underlying structure of the data, multiple datasets were obtained and the framework applied to each of these a hundred times. For each run of both the standard K-Means approach and the Seco Framework the best solution was selected. These solutions were then compared to the reference partitions of the data - representing the best recoverable structure for the data - therefore confirming that it is possible to ensure that the chosen partitions were stable.

For the dual measure approach it was evident that for any given run of the framework it was probable that the result would be consistent with any other run. However when the data was complex the single measure approach has a good chance that the result would not be identical. More importantly if this was the case then the performance of the solution could be good or poor with the user having no indication as to which was the case. For the SeCo framework although there is still the chance of obtaining slightly different solutions these all maintained a consistent concordance with the underlying structure of the data.

Through the use of synthetic and real world datasets it has been shown that sampled solutions are consistently good in their agreement with the known recoverable data structure and that repeatedly using the framework the performance of the dual measure approach is usually better than using the single measure. This is perhaps the fundamental tenet of this thesis showing that consistently good clustering solutions are sampled from within the spectrum of local minima as generated by repeated random initialisation of the algorithm.

A direct example of the applicability of this method is the need, in for example bioinformatics, for consistent assignment of individuals to a single cluster and it has been shown through the use of an affinity measure that the framework approach exhibits a greater likelihood of having individuals assigned to particular cohorts on a consistent basis which cannot be guaranteed with the single measure approach. This means that the framework is a useful tool for obtaining both the gross and micro structure within the data.

Following on from this, existing methods for the identification of an appropriate number of partitions to accurately represent the structure of the data have been shown here to be at least partially ineffectual in correctly identifying the correct value for a range of real world and synthetic datasets with varying complexity. Particularly so in the case of the Synthetic Data where none of the existing methods correctly identified a value close to the true structure of the dataset.

Some of these methods are effective on datasets with low complexity and dimensionality such as the Iris data however results shown here indicate that the methods fail with data containing greater inter-mixture of the cohorts or increased dimensionality. It is perhaps naïve to assume that on real world datasets where there is greater propensity for mixing of the cohorts and for which data does not necessarily fit assumptions of being convex that an automated method could point to a single value of k as being the best.

The method introduced here was tested on multiple datasets and has been shown to provide information regarding the range of values of k for which coherent structure has been recovered. For those datasets tested the performance of this method in identifying the correct number of cohorts is comparable to that of the other methods tested, and it allows inference to be derived in an efficient manner as the calculations are performed as part of the framework analysis. The purpose of this method is not however to specify a particular number of cohorts as being the only solution but to give an indication to the user as to which values of k have stable and repeatedly obtainable structure warranting further investigation as part of a larger exploratory analysis.

Moving beyond stability the framework is shown to provide information about the suitable values of k for which the cluster partition is consistent with the data structure. Existing methods for identifying appropriate values of k to represent the underlying structure of the data have been shown here to have varying reliability. In particular failing for the synthetic dataset where none of the highlighted methods correctly identified a value close to the underlying structure.

Importantly, this suggests that other methods are inconsistent for data sets with substantial mixing between clusters, a typical feature of real-world data. It has been shown previously (Ben-Hur et al. 2002) and here that using a method which does not rely on the structure of the data for determining an optimal number of clusters works well in these circumstances.

This proposed SeCo framework approach when used with K-Means was tested on multiple datasets and has been shown to provide accurate information regarding the range of values of k for which the cluster structure matches that of the data, doing so with computational efficiency compared with current alternatives. It is intended to determine a small range of values of k for which well-separated and stable clustering solutions are obtained and which therefore merit further investigation by expert users as part of a larger exploratory study.

9.2 ART2-A

The initialisation framework used in K-Means was unsuited to the very different ART2-A algorithm but was adapted to assist in the identification of well separated and reproducible solutions It has been shown previously that the use of a dual measure approach to evaluate the suitability of solutions from a clustering algorithm without using any external reference is a robust strategy for dealing with initialisation dependence [51] and the results here have shown that the method provides a credible solution to the problem of initialisation dependence for the ART2-A algorithm. The method has been shown that as with K-Means it allows for the robust identification of reproducible solutions and also to assist in the selection of an appropriate number of solutions to represent the underlying structure of the data.

Results have clearly shown that for the same initial parameters there is a substantial variability in partitions of the data obtained and that simple selection of a solution based on a given separation measure will not result in a reliably good partition.

The results again indicate for ART-2A that the framework approach excels at identifying gross structure (for example identifying the four main groups within the synthetic data, as demonstrated by Figure 4.2) within these datasets however the fine structure is still elusive as shown by the choice of 4 cohorts for the Synthetic Breast Cancer data where the same framework approach using K-Means has identified eight with equivalent concordance to the reference partition. This highlights that appropriate choice of algorithm is still important when considering a clustering of data.

9.3 Summary

A proposed framework approach to initialisation dependent clustering has been introduced and tested using two very different algorithms with very different initialisation problems with the expectation that application of the framework would provide increases in stability and reproducibility for these clustering algorithms across a range of different datasets.

The results here have shown that the SeCo Framework provides a credible solution to the problem of initialisation dependence for clustering algorithms and allows for the robust identification of reproducible solutions and also to potentially assist in the selection of an appropriate number of solutions to represent the underlying structure of the data. The results have clearly shown that for the same initial parameters that there is a substantial variability in the resultant partitions of the data and that simply selecting a solution based on separation will not result in a reliably good partition.

9.4 Future Work

Whilst the results from the ART-2A algorithm show the utility of the framework approach there is still plenty of scope to refine the application of the SeCo framework to this algorithm building on the foundations laid in this thesis. The first avenue for future work is to look into the separation measure used within the framework for ART-2A, currently this is the Between Cluster SSE, which made sense for the K-Means variant as the Total Within Cluster SSE is the objective function to be minimised and the two are related. However for ART-2A no such objective function exists so there is perhaps a better measure which could be used. Given that a lot of cluster separation measures implicitly use the SSE within their calculations, taking a orthogonal approach and using information based measures such as Entropy would be an interesting direction for the research to take.

The second area of work is to look at potential use cases for the ART-2A algorithm itself and developments of the ART-2A algorithm. For example K-Means uses an online update after the algorithm has completed to better optimise the SSE, and it may be possible to develop a similar update mechanism for ART-2A. Another possibility is to apply some elements of the ART-2 algorithm, such as the fixed number of prototypes within the algorithm which could allow for better control over the number of prototypes, as increasing the vigilance parameter can also increase the number of prototypes developed significantly.

Possibly one of the most fertile avenues for research though is to look into using it for clustering big data.

Big Data

Big Data is distinct from the concept of Data Mining as the latter is about manipulating and extracting information from existing stores of data whilst the former is about handling large amounts of data which may never be stored or revisited, but from which useful insights are required. Traditional clustering techniques such as K-Means are not necessarily appropriate for big data which is assessed by the application of the 5 Vs:

- Volume
- Velocity
- Variety
- Veracity
- Value

Within such a context the static nature of K-Means and other hierarchical clustering algorithms mean that once the initial clusters are identified it becomes unwieldy to update such a model for new data as it arrives. To better approach this a dynamic partitioning algorithm which continually updates its prototypes and identifies new groupings of observations as they occur is important, especially for big data and health where new observations are constantly arriving and may even be discarded immediately after use mean that iterative clustering techniques such as K-Means are inappropriate choices. These adaptations provide for a simple and efficient partitioning algorithm implementable in software and with the potential for Big Data applications as a direct result of its dynamic nature [59].

The use of a dual measure approach to initialisation in this framework provides a good starting position to using the clustering algorithm in a big data environment. This allows for robust partitions to be used to cluster incoming data whilst still allowing for the possibility of new cohorts being identified over time and in such a context ART2-A provides a realistic alternative to hierarchical, K-Means or other similar Expectation-Maximisation like algorithms.

ART-2A is capable of rapidly clustering data in either streaming data or batch processing environments however this may be limited if there is a proliferation of prototypes as data arrives. In such circumstances it may be beneficial to the end user to adhere to well defined existing prototypes rather than continuously creating new ones. An interesting avenue of research could be to investigate the effect of dynamic allocations during streaming when compared with more static prototypes and how retaining the plasticity of the algorithm will deal with new cohorts as they are introduced.

The use of ART-2A to cluster data allows for a dynamic clustering where Velocity, Volume and Value are of particular concern, given the ability of this family of algorithms to adapt to new prototypes over time and where learning these new cohorts is important. Identifying data which does not fit existing prototypes is not enough however and whilst this work potentially provides a good starting point for using ART-2A to dynamically cluster big data, there is a need to ensure that such new prototypes are in fact signal and not noise and to evaluate the performance of the method in actual big data scenarios and benchmark the effect of introducing new cohorts into a previously trained structure.

Tables and Figures A

A.1 List of Figures

A.2 List of Tables

References

B

[1] Usama Fayyad et al. “From Data Mining to Knowledge Discovery in Databases”. In: *AI Magazine* 17.3 (1996), pp. 37–54.

[2] Maartje E J Raijmakers et al. “Modeling developmental transitions in adaptive resonance theory.” In: *Developmental science* 7.2 (Apr. 2004), pp. 149–57. ISSN: 1363-755X.

[3] SK Murthy. “Automatic construction of decision trees from data: A multi-disciplinary survey”. In: *Data Mining and Knowledge Discovery* 389 (1998), pp. 345–389. DOI: `10.1023/A:1009744630224`.

[4] Anil K. Jain. “Data clustering: 50 years beyond K-means”. In: *Pattern Recognition Letters* 31.8 (June 2010), pp. 651–666. ISSN: 01678655. DOI: `10.1016/j.patrec.2009.09.011`.

[5] Marina Meilă. “An experimental comparison of model-based clustering methods”. In: *Machine Learning* 42.1 (2001), pp. 9–29. DOI: `10.1023/A:1007648401407`.

[6] Maurizio Filippone et al. “A survey of kernel and spectral methods for clustering”. In: *Pattern Recognition* 41.1 (Jan. 2008), pp. 176–190. ISSN: 00313203. DOI: `10.1016/j.patcog.2007.05.018`.

[7] Douglas Steinley. “K-means clustering: a half-century synthesis.” In: *The British journal of mathematical and statistical psychology* 59.Pt 1 (May 2006), pp. 1–34. ISSN: 0007-1102. DOI: `10.1348/000711005X48266`.

[8] Marina Meilă. “The uniqueness of a good optimum for K-means”. In: *Proceedings of the 23rd international conference on Machine learning - ICML ’06*. New York, New York, USA: ACM Press, 2006, pp. 625–632. ISBN: 1595933832. DOI: `10.1145/1143844.1143923`.

[9] J MacQueen. “Some methods for classification and analysis of multivariate observations”. In: *Proceedings of the fifth Berkeley symposium* 233.233 (1967), pp. 281–297.

[10] J. a. Hartigan et al. “Algorithm AS 136: A K-Means Clustering Algorithm”. In: *Applied Statistics* 28.1 (1979), p. 100. ISSN: 00359254. DOI: `10.2307/2346830`.

[11] Douglas Steinley. “Profiling local optima in K-means clustering: developing a diagnostic technique.” In: *Psychological methods* 11.2 (June 2006), pp. 178–92. ISSN: 1082-989X. DOI: `10.1037/1082-989X.11.2.178`.

[12] Douglas Steinley. “Local optima in K-means clustering: what you don’t know may hurt you.” In: *Psychological methods* 8.3 (Sept. 2003), pp. 294–304. ISSN: 1082-989X. DOI: `10.1037/1082-989X.8.3.294`.

[13] H. Steinhaus. “Sur la division des corp materiels en parties”. In: *Bull. Acad. Polon. Sci* 1 (1956), pp. 801–804.

[14] Stuart P Lloyd. “Least Squares Quantization in PCM”. In: *Information Theory, IEEE Transactions on* 28.2 (1982), pp. 129–137. DOI: `10.1109/TIT.1982.1056489`.

[15] T Dalenius. “The problem of optimal stratification I”. In: *Skandinavisk Aktuarietidskrift* (1950), pp. 203–213.

[16] T Dalenius et al. “The problem of optimal stratification II”. In: *Skandinavisk Aktuarietidskrift* (1951), pp. 133–148.

[17] Hans-Hermann Bock. “Origins and extensions of the k-means algorithm in cluster analysis”. In: *Electronic Journ@l for History of Probability and Statistics* 4.December (2008), pp. 1–18.

[18] E. W. Forgy. “Cluster analysis of multivariate data: efficiency vs interpretability of classifications”. In: *Biometrics* 21 (1965), pp. 768–769.

[19] Zhexue Huang. “A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining”. In: *In Research Issues on Data Mining and Knowledge Discovery*. 1997, pp. 1–8.

[20] Zhexue Huang. “Clustering large data sets with mixed numeric and categorical values”. In: *Asia Conference on Knowledge Discovery and Data*. 1997, pp. 1–14.

[21] James C Bezdek et al. “FCM: The fuzzy c-means clustering algorithm”. In: *Computers & Geosciences* 10.2-3 (Jan. 1984), pp. 191–203. ISSN: 00983004. DOI: 10.1016/0098-3004(84)90020-7.

[22] Dan Pelleg et al. “X-means: Extending K-means with efficient estimation of the number of clusters”. In: *Proceedings of the Seventeenth International Conference on Machine Learning*. San Francisco: Morgan Kaufmann, 2000, pp. 727–734.

[23] Bernhard Schölkopf et al. “Nonlinear Component Analysis as a Kernel Eigenvalue Problem”. In: *Neural Computation* 10.5 (July 1998), pp. 1299–1319. ISSN: 0899-7667. DOI: 10.1162/089976698300017467.

[24] Gail a Carpenter et al. “ART 2: self-organization of stable category recognition codes for analog input patterns.” In: *Applied optics* 26.23 (Dec. 1987), pp. 4919–30. ISSN: 0003-6935.

[25] Stephen Grossberg. “Adaptive Resonance Theory: how a brain learns to consciously attend, learn, and recognize a changing world.” In: *Neural networks : the official journal of the International Neural Network Society* 37 (Jan. 2013), pp. 1–47. ISSN: 1879-2782. DOI: 10.1016/j.neunet.2012.09.017.

[26] Gail a Carpenter et al. “ART 2-A: An adaptive resonance algorithm for rapid category learning and recognition”. In: *Neural Networks* 4.4 (Jan. 1991), pp. 493–504. ISSN: 08936080. DOI: 10.1016/0893-6080(91)90045-7.

[27] Thomas P. Caudell et al. “NIRS: Large scale ART-1 neural architectures for engineering design retrieval”. In: *Neural Networks* 7.9 (1994), pp. 1339–1350. ISSN: 08936080. DOI: 10.1016/0893-6080(94)90084-1.

[28] R. Arimoto et al. “Characterization of Asian Dust during ACE-Asia”. In: *Global and Planetary Change* 52 (2006), pp. 23–56. ISSN: 09218181. DOI: 10.1016/j.gloplacha.2006.02.013.

[29] Cosimo Distante et al. “Odor discrimination using adaptive resonance theory”. In: *Sensors and Actuators, B: Chemical* 69.3 (2000), pp. 248–252. ISSN: 09254005. DOI: 10.1016/S0925-4005(00)00502-5.

[30] Jiaoyan Ai et al. “Artificial Neural Networks - ICANN 2008”. In: *Theoretical Computer Science*. Vol. 5163. 2008. Chap. A New Type, pp. 89–98. ISBN: 9783540875581. DOI: 10.1007/978-3-540-87559-8.

[31] N. C. Yeo et al. “Colour image segmentation using the self-organizing map and adaptive resonance theory”. In: *Image and Vision Computing* 23.12 (2005), pp. 1060–1079. ISSN: 02628856. DOI: 10.1016/j.imavis.2005.07.008.

[32] Hongbo He et al. “Application of Adaptive Resonance Theory neural networks to monitor solar hot water systems and detect existing or developing faults”. In: *Solar Energy* 86.9 (2012), pp. 2318–2333. ISSN: 0038092X. DOI: 10.1016/j.solener.2012.05.015.

[33] Ji He et al. “Initialization of cluster refinement algorithms: a review and comparative study”. In: *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541).* 1. Ieee, 2004, pp. 297–302. ISBN: 0-7803-8359-1. DOI: 10.1109/IJCNN.2004.1379917.

[34] J.R. Whiteley et al. “A similarity-based approach to interpretation of sensor data using adaptive resonance theory”. In: *Computers & Chemical Engineering* 18.7 (1994), pp. 637–661. ISSN: 00981354. DOI: 10.1016/0098-1354(94)85003-8.

[35] Chun Hsien Chen et al. “A strategy for acquiring customer requirement patterns using laddering technique and ART2 neural network”. In: *Advanced Engineering Informatics* 16.3 (2002), pp. 229–240. ISSN: 14740346. DOI: 10.1016/S1474-0346(03)00003-X.

[36] Steven a. Sloman. “The empirical case for two systems of reasoning.” In: *Psychological Bulletin* 119.I (1996), pp. 3–22. ISSN: 0033-2909. DOI: 10.1037/0033-2909.119.1.3.

[37] J.M Peña et al. “An empirical comparison of four initialization methods for the K-Means algorithm”. In: *Pattern Recognition Letters* 20.10 (Oct. 1999), pp. 1027–1040. ISSN: 01678655. DOI: 10.1016/S0167-8655(99)00069-0.

[38] Paul S Bradley et al. “Refining Initial Points for K-Means Clustering One Microsoft Way Refining Initial Points for K-Means Clustering”. In: *Proceedings of the 15th International Conference on Machine Learning (ICML98),* Madison, Wisconsin, 1998, pp. 91–99.

[39] Leonard Kaufman et al. *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley-Blackwell, 1990, p. 368. ISBN: 0471735787.

[40] Marina Meilă. “An experimental comparison of several clustering and initialization methods”. In: *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence* (1998), pp. 386–395.

[41] David Arthur et al. “k-means ++ : The Advantages of Careful Seeding”. In: *SODA ’07 Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Vol. 8. 2007, pp. 1027–1035.

[42] Stephen J. Redmond et al. “A method for initialising the K-means clustering algorithm using kd-trees”. In: *Pattern Recognition Letters* 28.8 (June 2007), pp. 965–973. ISSN: 01678655. DOI: 10.1016/j.patrec.2007.01.001.

[43] Ji He et al. “Modified ART 2A growing network capable of generating a fixed number of nodes.” In: *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council* 15.3 (May 2004), pp. 728–37. ISSN: 1045-9227. DOI: 10.1109/TNN.2004.826220.

[44] Chia-Hui Chang et al. “Categorical data visualization and clustering using subjective factors”. In: *Data & Knowledge Engineering* 53.3 (June 2005), pp. 243–262. ISSN: 0169023X. DOI: 10.1016/j.datak.2004.09.001.

[45] Asa Ben-Hur et al. “A stability based method for discovering structure in clustered data.” In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 17 (Jan. 2002), pp. 6–17. ISSN: 1793-5091.

[46] Robert Tibshirani et al. “Estimating the number of clusters in a data set via the gap statistic”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (May 2001), pp. 411–423. ISSN: 1369-7412. DOI: 10.1111/1467-9868.00293.

[47] Tilman Lange et al. “Stability-based validation of clustering solutions.” In: *Neural computation* 16.6 (June 2004), pp. 1299–323. ISSN: 0899-7667. DOI: 10.1162/089976604773717621.

[48] Douglas Steinley. “Stability analysis in K-means clustering.” In: *The British journal of mathematical and statistical psychology* 61.Pt 2 (Nov. 2008), pp. 255–73. ISSN: 0007-1102. DOI: 10.1348/000711007X184849.

[49] Ludmila I Kuncheva et al. “Evaluation of stability of k-means cluster ensembles with respect to random initialization.” In: *IEEE transactions on pattern analysis and machine intelligence* 28.11 (Nov. 2006), pp. 1798–808. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2006.226.

[50] Douglas Steinley. “Properties of the Hubert-Arabie adjusted Rand index.” In: *Psychological methods* 9.3 (Sept. 2004), pp. 386–96. ISSN: 1082-989X. DOI: 10.1037/1082-989X.9.3.386.

[51] Paulo JG Lisboa et al. “Finding reproducible cluster partitions for the k-means algorithm”. In: *BMC Bioinformatics* 14.Suppl 1 (2013), S8. ISSN: 1471-2105. DOI: 10.1186/1471-2105-14-S1-S8.

[52] John A. Hartigan. *Clustering Algorithms (Probability & Mathematical Statistics)*. John Wiley & Sons Inc, 1975. ISBN: 047135645X.

[53] Glenn W. Milligan et al. “An examination of procedures for determining the number of clusters in a data set”. In: *Psychometrika* 50.2 (June 1985), pp. 159–179. ISSN: 0033-3123. DOI: 10.1007/BF02294245.

[54] Greg Hamerly et al. “Learning the k in k-means”. In: *In Neural Information Processing Systems*. Ed. by Sebastian Thrun et al. Vol. 17. NIPS 16. MIT Press, 2003, pp. 1–8.

[55] I. H. Jarman et al. “Clustering of protein expression data: a benchmark of statistical and neural approaches”. In: *Soft Computing* 15.8 (Apr. 2011), pp. 1459–1469. ISSN: 1432-7643. DOI: 10.1007/s00500-010-0596-9.

[56] I.H. Jarman et al. “Clustering categorical data: A stability analysis framework”. In: *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)* (Apr. 2011), pp. 58–65. DOI: 10.1109/CIDM.2011.5949452.

[57] P J G Lisboa et al. “Discovering hidden pathways in bioinformatics”. In: *Eighth International Meeting on Computational Intelligence Methods in Bioinformatics and Biostatistics*. 2011.

[58] Matus Telgarsky et al. “Hartigan ’ s Method : k-means Clustering without Voronoi”. In: 9 (2010), pp. 820–827.

[59] Stephen I. Gallant. *Neural network learning and expert systems*. A Bradford Book, 1993. Chap. Chapter 7, pp. 147–149. ISBN: 9780262071451.

[60] Alan Agresti. *Categorical Data Analysis*. Wiley, 1990, p. 558. ISBN: 9780471853015.

[61] Harald Cramer. *Mathematical Methods of Statistics. (PMS-9).* Princeton University Press, Mar. 1999. ISBN: 0691005478.

[62] Sandrine Dudoit et al. "A prediction-based resampling method for estimating the number of clusters in a dataset." In: *Genome biology* 3.7 (June 2002), RESEARCH0036. ISSN: 1465-6914.

[63] T. Calinski et al. "A dendrite method for cluster analysis". In: *Communications in Statistics - Theory and Methods* 3.1 (1974), pp. 1–27. ISSN: 0361-0926. DOI: 10.1080/03610927408827101.

[64] J. C. Dunn. "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters". In: *Journal of Cybernetics* 3.3 (Jan. 1973), pp. 32–57. ISSN: 0022-0280. DOI: 10.1080/01969727308546046.

[65] M Forina et al. "Pattern recognition methods in the prediction of Italian olive oil origin by their fatty acid content". In: *Annali di Chimica* 72 (1982), pp. 143–155.

[66] A Frank et al. *UCI Machine Learning Repository.* 2010.

[67] Diogo Ayres de Campos et al. "Sisporto 2.0: A program for automated analysis of cardiotocograms". In: *The Journal of Maternal-Fetal Medicine* 9.5 (Sept. 2000), pp. 311–318. ISSN: 10570802. DOI: 10.1002/1520-6661(200009/10)9:5<311::AID-MFM12>3.3.CO;2-0.

[68] MRN Kousarrizi et al. "An experimental comparative study on thyroid disease diagnosis based on feature subset selection and classification". In: *International Journal of Electrical ...* February (2012).

[69] Dalia M Abd El-Rehim et al. "High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses." In: *International journal of cancer. Journal international du cancer* 116.3 (Sept. 2005), pp. 340–50. ISSN: 0020-7136. DOI: 10.1002/ijc.21004.

[70] P Lisboa et al. "Cluster-based visualisation with scatter matrices". In: *Pattern Recognition Letters* 29.13 (Oct. 2008), pp. 1814–1823. ISSN: 01678655. DOI: 10.1016/j.patrec.2008.05.021.

[71] Enrique H. Ruspini. “Numerical methods for fuzzy clustering”. In: *Information Sciences* 2.3 (July 1970), pp. 319–350. ISSN: 00200255. DOI: `10.1016/S0020-0255(70)80056-1`.

[72] Xiaojun Chen et al. “A feature group weighting method for subspace clustering of high-dimensional data”. In: *Pattern Recognition* 45.1 (Jan. 2012), pp. 434–446. ISSN: 00313203. DOI: `10.1016/j.patcog.2011.06.004`.

[73] Richard O Duda et al. *Pattern Classification and Scene Analysis*. Vol. 7. 4. Wiley, 1973. Chap. 2, p. 482. ISBN: 0471223611. DOI: `10.2307/1573081`.

Publications C

Included here are the two journal papers submitted for publication during the course of the PhD with two conference abstracts also included.

The first two documents abstracts are from the MEDSIP 2012 conference, held in Liverpool, entitled “A Framework Approach to K-Means Clustering” and comments upon the performance of standard K-Means type algorithms compared to the dual measure framework in respect of four datasets. Followed by the IEEE SSCI Conference abstract titled “A framework for initialising a dynamic clustering algorithm: ART2-A”.

Thirdly, the paper “Finding reproducible cluster partitions for the K-Means algorithm” was submitted for a special issue of BMC Bioinformatics, and accepted in September 2012 forr publication in 2013.

Finally, a paper submitted to the Internation Journal of Bioengineering Technology entitled “Inference of number of prototypes with a framework approach to K-means clustering” looks at the use of the framework in assisting the determination of an appropriate value of “K” for the clustering algorithm.