Clustering with the Average Silhouette Width

Fatima Batool

Department of Statistical Science
University College London
Gower Street, London WC1E 6BT, United Kingdom

Christian Hennig

Dipartimento di Scienze Statistiche "Paolo Fortunati"
Universita di Bologna
Bologna, Via delle belle Arti, 41, 40126, Italy

Abstract

The Average Silhouette Width (ASW) is a popular cluster validation index to estimate the number of clusters. The question whether it also is suitable as a general objective function to be optimized for finding a clustering is addressed. Two algorithms (the standard version OSil and a fast version FOSil) are proposed, and they are compared with existing clustering methods in an extensive simulation study covering known and unknown numbers of clusters. Real data sets are analysed, partly exploring the use of the new methods with non-Euclidean distances. The ASW will be shown to satisfy some axioms that have been proposed for cluster quality functions. The new methods prove useful and sensible in many cases, but some weaknesses are also highlighted. These also concern the use of the ASW for estimating the number of clusters together with other methods, which is of general interest due to the popularity of the ASW for this task.

Keywords: axiomatic clustering, distance-based clustering, partitioning around medoids, number of clusters

*A detailed definition of data generating processes used in the simulation study is in an electronic supplement.

Email addresses: `fatiuma.batool.14@ucl.ac.uk` (Fatima Batool), `christian.hennig@unibo.it` (Christian Hennig)

2010 MSC: 62H30

---

1. Introduction

Cluster analysis is a central task in modern data analysis. It is applied in diverse areas such as public health, machine learning, psychology, archaeology, genetics, computer vision, and text analytics to just name a few. There is a wide variety of clustering methods (Hennig et al. (2015)), and it has been argued that there is no universally best approach, and that the cluster analysis approach needs to be chosen taking into account what kinds of clusters are required, which depends on domain knowledge and the aim of clustering (Von Luxburg et al. (2012); Hennig (2015)).

Many cluster analysis methods such as k-means (Lloyd (1982)) and Partitioning Around Medoids (PAM; Kaufman and Rousseeuw (1987)) are defined by optimizing an objective function over all partitions of the data set into a fixed number of k clusters. The objective function formalizes the quality of the clustering. The PAM objective function, for example, sums up distances of all observations to the center of the cluster to which they were assigned, and a good clustering is one for which this is small. Many of these objective functions are not suitable for finding an optimal number of clusters k, because they will automatically improve if k is increased due to more degrees of freedom in optimization.

Therefore, so-called cluster validation indexes that can be meaningfully optimized over k are often used together with such partitioning clustering methods in order to find an optimal k. Many such indexes have been proposed in the literature (see Halkidi et al. (2015); Arbelaitz et al. (2012)). Some of these are for fixed k equivalent to objective functions of partitioning methods such as k-means, e.g., the Calinski and Harabasz index (Caliński and Harabasz (1974)). Some others are not in this way connected to a specific partitioning method. One popular such index is the Average Silhouette Width (ASW) (Rousseeuw (1987)). The ASW achieved overall very good results in the extensive simulation study of Arbelaitz et al. (2012). In Kaufman and Rousseeuw (1990) it is suggested for finding the number of clusters with PAM, but in fact its definition is not directly connected to any specific partitioning method, and it can be seen as a general distance-based approach to assess the quality of a clustering.

The ASW is an intuitive and simple measurement of cluster quality that does not rely on statistical model assumptions. Given that it is widely used and trusted for comparing the qualities of clusterings produced by various clustering methods over different numbers of clusters, it seems natural to investigate optimal ASW quality clustering not only over k but also for fixed k in order to integrate the problem for fixed k and the problem of finding the best k. This idea is explored here. We treat the idea with an open mind and do not attempt to suggest that optimum ASW clustering is an optimal clustering method in any other sense than optimizing the ASW (which can be seen as desirable on its own terms); rather what we do is to show both potential and problems with this approach. The problems are also relevant to the use of the ASW for just choosing k, for which it is of widespread use despite a far from comprehensive evaluation and theoretical basis. To our knowledge, up to now using the ASW also for choosing a clustering with fixed k has only been explored by Van der Laan et al. (2003), where a modification of the PAM algorithm called PAMSil was proposed that looks for a local medoid-based optimum of the ASW. Rousseeuw (1987), where the ASW was originally introduced, mentions a possibility of its optimisation for finding a clustering in a side remark. Exploring the new clustering method based on the ASW, we also to some extent explore strengths and weaknesses of the popular use of the ASW as a method to estimate the number of clusters.

In Section 2 we introduce optimum ASW clustering and propose two algorithms for it. In Section 3 we show that the ASW fulfils some axioms that have been proposed for clustering quality measures in Ackerman and Ben-David (2009). In Section 4 we run an extensive simulation study to explore the performance of optimum ASW clustering compared to other well established clustering methods. There is also an experiment regarding outliers. Section 5 applies optimum ASW clustering to a number of real data sets with and without given "true" clustering, also illustrating the use of optimum ASW for non-Euclidean dissimilarities. A conclusion is given in Section 6.

2. Methodology

2.1. Notation and basic definitions

Let $X = \{x_1, \ldots, x_n\}$ be a data set of n objects from a space $\mathcal{X}$, d be a dissimilarity or distance over $\mathcal{X}$. The triangle inequality is not really necessary here, although the intuition behind the concepts cluster separation and homogeneity may look dubious if for example for points x_1, x_2, x_3 it is possible that $d(x_1, x_2)$ and $d(x_2, x_3)$ are very small, but $d(x_1, x_3)$ is very large. We deal with clusterings that are partitions, i.e., non-overlapping and exhaustive. A partition can equivalently be expressed by labels $l(1), \ldots, l(n) \in \mathbb{N}_k = \{1, \ldots, k\}$ where $l(i) = r \Leftrightarrow x_i \in C_r$, $i \in \mathbb{N}_n$, and cluster sizes are denoted by $n_r = \sum_{i=1}^n 1(l(i) = r)$, $r \in \mathbb{N}_k$.

Definition 2.1. *The silhouette width for an observation $x_i \in X$ is*

$$s_i(\mathcal{C}, d) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, \tag{1}$$

where

$$a(i) = \frac{1}{n_{l(i)} - 1} \sum_{\substack{l(i)=l(j) \\ i \neq j}} d(x_i, x_j) \quad and \quad b(i) = \min_{r \neq l(i)} \frac{1}{n_r} \sum_{l(j)=r} d(x_i, x_j)$$

in case that $n_r > 1$ for $l(i) = r$. Otherwise $s_i(\mathcal{C}, d) = 0$.

*The **Average Silhouette Width** (ASW) of a clustering $\mathcal{C}$ is*

$$\bar{S}(\mathcal{C}, d) = \frac{1}{n} \sum_{i=1}^n s_i(\mathcal{C}, d).$$

$a(i)$ is the average distance of x_i to points in the cluster to which it was assigned, and $b(i)$ is the average distance of x_i to the points in the nearest cluster to which it was not assigned. A large value of $s_i(\mathcal{C}, d)$ means that $b(i)$ is much larger than $a(i)$, and that consequently x_i is much closer to the observations in its own cluster than to the neighboring one. Given that clusters are meant to be homogeneous and well separated, larger values of s_i and $\bar{S}$ indicate better clustering quality, and an optimal clustering (for example with optimal k if various values of k are compared) in the sense of the ASW is one that maximizes $\bar{S}$. A more in-depth heuristic motivation of silhouette widths along with some examples is given by Rousseeuw (1987), which focuses on the graphical display of individual silhouette widths but also introduces the ASW for assessment of the whole clustering.

Here are some simple properties of the ASW. $-1 \leq s_i \leq 1$ always, and the same holds obviously for the ASW. In fact an ASW of 0 can be seen as a rather bad value, because it means that on average observations are not closer to the observations in their own cluster than to the observations in the closest other cluster. However, a random clustering with more than 2 clusters cannot normally be expected to achieve $\bar{S} = 0$, because $b(i)$ is computed by minimizing over clusters to which an observation does not belong, and on average this minimum can be expected to be smaller than $a(i)$ if the observations in the same cluster are as randomly chosen as those in the other clusters.

If there are well separated and compact subsets in the data, taking these as the clusters will make the vast majority of s_i and consequently $\bar{S}$ substantially larger than zero, and this will be better than having k close to its maximum value n (in which case many one-point clusters will lead to $s_i = 0$ for many i). Putting two or more such subsets together in a cluster will have a detrimental impact on the corresponding $a(i)$-values, damaging $\bar{S}$ in turn, so that the optimum value of $\bar{S}$ will not normally occur at a k lower than the number of well separated subsets either. There is a possible exception to this though. Putting two neighboring clusters together can make their separation from the rest, i.e., the corresponding $b(i)$-values, much higher, so that occasionally data subsets that have some separation from each other put together produce a better ASW-value, if their separation from the rest is much stronger. This can be seen as a general pitfall of the ASW, particularly when it comes to estimating k. See also Hennig and Lin (2015), where maximizing the ASW over k (for given k, clusterings were produced by PAM) produces an apparently too low value of $k = 2$.

The ASW cannot be computed for $k = 1$, so it cannot directly be used to decide whether the data set as a whole is homogeneous and whether there should be any clustering at all. Hennig and Lin (2015) suggest to compare ASW values on the data to ASW values from clustering homogeneous "null model" data without clustering to see whether the data have a significant clustering structure.

Definition 2.2. *A clustering $\mathcal{C}^*$ is an* **optimum ASW clustering** *for dissimilarity d if*

$$\bar{S}(\mathcal{C}^*, d) = \max_{\mathcal{C}} \bar{S}(\mathcal{C}, d), \tag{2}$$

where $\mathcal{C} \in \mathcal{P}(X)$ with $|\mathcal{C}| \geq 2$. Define $\mathcal{C}_k^*$ as optimum ASW clustering out of the clusterings with $|\mathcal{C}| = k$.

As with other clustering principles that are defined as optimizing an objective function, finding a global optimum will be computationally infeasible for all but the smallest data sets. We therefore propose two algorithms to find local optima that are hopefully close to the global one.

2.2. The OSil algorithm

The following algorithm improves an initialisation by changing the cluster membership of the point that improves the ASW most at any given stage until no further improvement can be found. As the ASW will always be improved and there are only finitely many clusterings, the algorithm will converge in a finite number of steps. We call this the OSil (optimum silhouette) algorithm.

OSil algorithm

- Set k_{min} as the minimum number of clusters and k_{max} as the maximum number of clusters.
- Input: dissimilarity $d(x_i, x_h)$, $\forall i \neq h \in \mathbb{N}_n$.
- For every $k \in \{k_{min}, \ldots, k_{max}\}$:
  1. Start with initialisation clustering $\mathcal{C}_k^0$ defined by $l^0(\mathcal{X}, k) = (l^0(1), \ldots, l^0(n))$, for $i \in \mathbb{N}_n: \ l^0(i) \in \mathbb{N}_k$. Let $q = 0$.
  2. Compute $\bar{S}(\mathcal{C}_k^q, d)$.
  3. For all pairs (i, r) such that $l^0(i) \neq r$, $i \in \mathbb{N}_n, \ r \in \mathbb{N}_k$, assign $l^{(i,r)}(i) = r, \ l^{(i,r)}(j) = l^0(j)$ for all $j \neq i$, and denote the so obtained clustering as $\mathcal{C}_k^{(i,r)}$.
  4. Compute $f^{(i,r)} = \bar{S}(\mathcal{C}_k^{(i,r)}, d)$.
  5. $(h, s) = \underset{(i,r)}{\arg\max} \ f^{(i,r)}$ (we recommend to constrain (i, r) so that $\mathcal{C}_k^{(i,r)}$ still has k nonempty clusters, but not using this constraint would be an alternative if for given k a clustering with k nonempty clusters is not necessarily required),
  6. If $f^{(h,s)} \leq \bar{S}(\mathcal{C}_k^q, d)$: $q = q + 1$, $\mathcal{C}_k^{(q)} = \mathcal{C}_k^{(h,s)}$, and go to Step 2. Otherwise stop and give out $\mathcal{C}_k = \mathcal{C}_k^{(q)}$ as final solution for number of clusters k.

Give out $\mathcal{C}_{k^*} = argmax_{k \in \{k_{min}, \ldots, k_{max}\}} \bar{S}(\mathcal{C}_k, d)$ as final clustering.

It is advisable to run the algorithm more than once from different initialisations, and then to use the solution that achieves the best ASW-value. We ran some simulations comparing different possible ways of initialisation, particularly initialisation by already existing clustering methods, see Batool and Hennig (2019). Good solutions over a variety of setups can be achieved initializing six times with k-means, PAM, average linkage, single linkage, Ward's method and model-based clustering (using the standard settings of the R-package `mclust` (Scrucca et al. (2017)), which we recommend here.

2.3. FOSil - An approximation algorithm for bigger data sets

The OSil algorithm is computationally expensive, because it considers all possible combinations of cluster and observation swaps for each iteration. For large data sets it can be very slow. A simple way to construct a faster algorithm for larger data sets is to run the OSil algorithm on a subset of observations, and then to assign all remaining observations to the clusters in such a way that the ASW is maximized for every observation separately. We call this the FOSil (Fast OSil) algorithm.

FOSil algorithm

- Set k_{min} as the minimum number of clusters and k_{max} as the maximum number of clusters.
- Set a sample size n_s and a number M of samples to be drawn.
- Input: dissimilarity $d(x_i, x_h)$, $\forall i \neq h \in \mathbb{N}_n$.
- For every $k \in \{k_{min}, \dots, k_{max}\}$:
  1. For $m = 1, \dots, M$:
     (a) Choose a random sample S_m of size n_s from X. Let $I_{S_m} \subset \{1, \dots, n\}$ be the set of indexes of the observations in S_m. Let d_m be d constrained to the elements of S_m.
     (b) Run the OSil algorithm on S_m, d_m with number of clusters k, potentially starting from several initialisations (see the discussion at the end of Section 2.2). Call the resulting clustering $\mathcal{C}_{S_m,k}$. For a label vector $(l^*_{S_m,k}(1), \dots, l_{S_m,k}(n))$ define $l_{S_m,k}(i) = r$ if $x_i \in C_r$ in $\mathcal{C}_{S_m,k}$ for $i \in I_{S_m}$.
  2. Choose $S = \underset{S_m \in \{S_1,\dots,S_M\}}{\arg\max} \bar{S}(S_m, d_m)$.
  3. Calculate the cluster memberships for the points in $S' = X \setminus S$ by maximizing the ASW. For all $x_i \in S'$ and $C_r \in \mathcal{C}_{S,k},\ i \in \{2, \dots, n\} \setminus I_S,\ r \in \{1, \dots, k\}$:
     (a) Consider the clustering $\mathcal{C}^{(i,r)}_{S,k}$ of $S \cup \{x_i\}$ defined by putting $x_i \in C_r$ and otherwise leaving $\mathcal{C}_{S,k}$ unchanged. Let d^* denote d constrained to the elements of $S \cup \{x_i\}$.
     (b) Compute $f^{(i,r)} = \bar{S}(\mathcal{C}^{(i,r)}_{S,k}, d^*)$.
     (c) $(h, s) = \underset{(i,r)}{\arg\max}\ f^{(i,r)}$; $l_{S,k}(h) = s$.
  4. Give out C_k as defined by $l_{S,k}$ as the final solution for number of clusters k.

Give out $\mathcal{C}_{k^*} = argmax_{k \in \{k_{min},\dots,k_{max}\}} \bar{S}(\mathcal{C}_k, d)$ as final clustering.

We used $M = 25$ (larger M does not seem to improve matters much) and $n_s = 0.2n$, although for larger data sets n_s can be chosen smaller; the absolute size of n_s will often matter more than its ratio to n. Alternatively, one could choose 20 observations times the maximum number of clusters of interest. As in Section 2.2, we initialize the clustering of the S_m by k-means, PAM, average linkage, single linkage, Ward's method and model-based clustering. See Figure 1 for a comparison of computing times of OSil and FOSil.

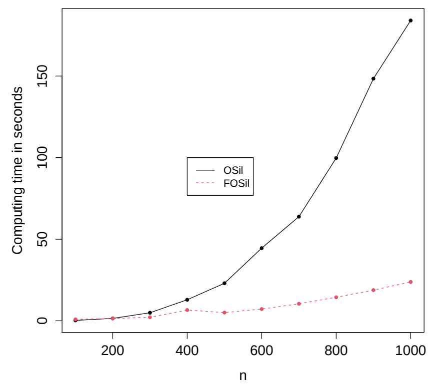


Figure 1: Computing times of OSil and FOSil on two-dimensional data sets with equally sized Gaussian clusters centered at (0,0), (0,1), (1,0), and (1,1), with standard deviations 0.1 in both independent dimensions for n between 100 and 1000.

There are some possible variations of the FOSil algorithm. For example, when assigning the points in S' to the clusters, one may use for classifying a new point the ASW including all points already assigned rather than just S plus the new point. Also, points in S' may be assigned to the clusterings $\mathcal{C}_{S_m,k}$ from all M subsamples before comparing the results from the different subsamples. Both of these can be expected to improve the clustering but will take more computation time than FOSil. We do not explore them here.

3. Axiomatic characterisation of the ASW

The ASW has originally been introduced as a heuristic concept. Despite its popularity and good results in some studies, to this day there has not been much theoretical investigation of its characteristics. We apply an axiomatic approach by Ackerman and Ben-David (2009) to the ASW. The general idea is to characterize a desirable behavior of a reasonable clustering method (or clustering quality measure (CQM)) by certain theoretical axioms and then to check whether for a given method or CQM these axioms are fulfilled. This approach goes back to at least Rubin (1967). Some further early work was done by Jardine and Sibson (1968), Fisher and Ness (1971), and Wright (1973).

More recently, Kleinberg (2003) defined three apparently intuitive axioms for clustering functions (i.e., clustering methods), "scale invariance" (the clustering does not change if all dissimilarities are multiplied by the same constant), "consistency" (if dissimilarities are changed in such a way that all within-cluster dissimilarities are either made smaller or unchanged, and all between-cluster dissimilarities are either made larger or unchanged, the clustering does not change), and "richness" (for all possible clusterings there is a dissimilarity that will produce them). He then proved that it is impossible for any clustering function to fulfil all three of them. Ackerman and Ben-David (2009) argued that particularly the consistency axiom is not as desirable as Kleinberg (2003) had claimed. They proposed versions of the three axioms that do not apply to clustering functions but rather to CQMs such as the ASW, which we will use here. In this way, all three can be fulfilled. Some other relaxations of Kleinberg's axioms have been proposed by Zadeh and Ben-David (2009), Correa-Morris (2013), and Carlsson and Mémoli (2013).

3.1. Definitions and axioms

Let $X = \{x_1, \cdots, x_n\}$ be the data set as before and d a dissimilarity on X. For a given partition $\mathcal{C}$ of X, write $x_i \sim_{\mathcal{C}} x_j$ if x_i and x_j are in the same cluster in $\mathcal{C}$. Clustering $\mathcal{C}$ is called non-trivial if not either $\mathcal{C} = \{X\}$ (only the whole data set is a cluster) or $\mathcal{C} = \{\{x_1\}, \ldots, \{x_n\}\}$ (all n singletons are the clusters). A clustering quality measure (CQM) Π takes the pair (X, d) and a clustering $\mathcal{C}$ over (X, d) and returns a non-negative real number, where a larger number is interpreted as a higher cluster quality.

Definition 3.1. *For a dissimilarity d over X and a positive real η, the scalar multiplication of d with η is defined for every pair $x_i, x_j \in X$, as $(\eta \cdot d\ )(x_i,\ x_j)$ $= \eta d(x_i, x_j)$.*

Definition 3.2. *A dissimilarity d' is called a $\mathcal{C}$-**transformation** of d, if $d'(x_i, x_j) \leq d(x_i, x_j)$ for all $x_i \sim_{\mathcal{C}} x_j$ and $d'(x_i, y_j) \geq d(x_i, x_j)$ for all $x_i \not\sim_{\mathcal{C}} x_j$ for all $i, j \in \mathbb{N}_n$.*

Definition 3.3. *Two clusterings $\mathcal{C}$ and $\mathcal{C}'$ of X are **isomorphic** ($\mathcal{C} \approx_d \mathcal{C}'$ if there exists a distance-preserving isomorphism $\phi : X \mapsto X$,such that for all $x, y \in X : x \sim_{\mathcal{C}} y \Leftrightarrow \phi(x) \sim_{\mathcal{C}'} \phi(y)$.*

Ackerman and Ben-David (2009) state their axioms as follows.

Axiom 1. ***CQM scale invariance:*** *A CQM Π is scale invariant if for all $\eta > 0$, and every $\mathcal{C}$ of (X, d), $\Pi(\mathcal{C}, (X, d)) = \Pi(\mathcal{C}, (X, \eta \cdot d))$.*

Axiom 2. ***CQM consistency:*** *A CQM Π is consistent if for every clustering $\mathcal{C}$ over (X, d), $\Pi(\mathcal{C}, (X, d')) \geq \Pi(\mathcal{C}, (X, d))$ holds, provided that d' is a $\mathcal{C}$-transformation of d.*

Axiom 3. ***CQM richness:*** *A CQM Π is rich for every possible non-trivial clustering $\mathcal{C} \in \mathcal{S}(X)$ of X there exist a dissimilarity d over X such that $\mathcal{C} = \arg\max_{\mathcal{C}} \Pi(\mathcal{C}, (X, d))$.*

Axiom 4. ***Isomorphism invariance:*** *A CQM Π is isomorphism-invariant if for allclusterings $\mathcal{C}, \mathcal{C}'$ over (X, d) where $\mathcal{C} \approx_d \mathcal{C}' : \Pi(\mathcal{C}, (X, d)) = \Pi(\mathcal{C}', (X, d))$.*

If a clustering method is defined by maximizing a CQM (such as the ASW), CQM richness is identical to the richness axiom in Kleinberg (2003), and CQM scale invariance implies the scale invariance axiom in Kleinberg (2003). CQM consistency is weaker than Kleinberg (2003)'s consistency for clustering methods.

The axioms are justified as follows. Scalar multiplication should not affect the grouping structure of a data set, as all dissimilarities are modified in the same way. Regarding CQM consistency, a $\mathcal{C}$-transformation improves both the within-cluster homogeneity and the between-cluster separation, so the existing clustering should be rated as of higher quality by the CQM. Kleinberg (2003)'s consistency demands more, but it is hard to justify that the transformation is not allowed to lead to another even better clustering, see Ackerman and Ben-David (2009). The rationale behind richness is that if any non-trivial clustering cannot be achieved by constructing a dissimilarity for which this clustering is optimal, non-optimality of that clustering is an artefact of the CQM (or the clustering method) rather than a defect of the clustering itself. Ackerman and Ben-David (2009) show that CQM scale invariance, CQM richness, and CQM consistency form a consistent set of axioms. They added isomorphism invariance in order to stop some functions that would be unreasonable as CQMs from fulfilling all axioms.

3.2. Characterisation of the ASW

We now prove CQM scale invariance, CQM richness, and CQM consistency for the ASW. Isomorphism invariance is trivially fulfilled because the ASW depends on dissimilarities only.

Proposition 3.4. *The ASW is a scale invariant CQM.*

Proof: If d is replaced by $\eta\dot{d}$, for all i both $a(i)$ and $b(i)$ are multiplied by η, and therefore s_i does not change.

Theorem 3.5. *The ASW is a consistent CQM.*

Proof: Let d' be a $\mathcal{C}$-transformation of d, and $a'(i)$, $b'(i)$, $s'_i = s_i(\mathcal{C}, d')$, $\bar{S}' = \bar{S}(\mathcal{C}, d')$ denote the corresponding quantities from the definition of the ASW based on d'. By Definition 3.2: $d'(x_i, x_j) \leq d(x_i, x_j)$ for all $x_i \sim_{\mathcal{C}} x_j$, and $\min_{x_i \not\sim_{\mathcal{C}} x_j} d'(x_i, y_j) \geq \min_{x_i \not\sim_{\mathcal{C}} x_j} d(x_i, x_j)$. This implies for all $i \in \mathbb{N}_n$:

$$a'(i) \leq a(i),\ b'(i) \geq b(i). \tag{3}$$

Show for all i:

$$\frac{b'(i) - a'(i)}{\max\{a'(i), b'(i)\}} - \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \geq 0, \tag{4}$$

which is equivalent $s'_i(\mathcal{C}, d) \geq s_i(\mathcal{C}, d)$, and implies CQM consistency, as $\bar{S}'$ and $\bar{S}$ average these.

There are four possible cases:

$$\text{Case I: } \max\{a(i), b(i)\} = a(i), \qquad \max\{a'(i), b'(i)\} = a'(i) \tag{5}$$

$$\text{Case II: } \max\{a(i), b(i)\} = a(i), \qquad \max\{a'(i), b'(i)\} = b'(i) \tag{6}$$

$$\text{Case III: } \max\{a(i), b(i)\} = b(i), \qquad \max\{a'(i), b'(i)\} = a'(i) \tag{7}$$

$$\text{Case IV: } \max\{a(i), b(i)\} = b(i), \qquad \max\{a'(i), b'(i)\} = b'(i). \tag{8}$$

Check whether (4) holds for each of these:

Case I: (4) amounts to

$$\frac{b'(i) - a'(i)}{a'(i)} - \frac{b(i) - a(i)}{a(i)} \geq 0 \Leftrightarrow$$
$$\frac{b'(i)}{a'(i)} - \frac{b(i)}{a(i)} \geq 0.$$

This follows from (3).

Case II: (4) amounts to

$$2 - \frac{a'(i)}{b'(i)} - \frac{b(i)}{a(i)} \geq 0,$$

and due to (6) both $\frac{a'(i)}{b'(i)}$ and $\frac{b(i)}{a(i)}$ are ≤ 1.

Case III: $b(i) \geq a(i)$ and $a'(i) \geq b'(i)$ imply $a'(i) \geq b'(i) \geq b(i) \geq a(i)$, which is only compatible with (3) if they are all equal, and $s'_i = s_i(\mathcal{C}, d)$.

Case IV: (4) amounts to

$$\frac{b'(i) - a'(i)}{b'(i)} - \frac{b(i) - a(i)}{b(i)} \geq 0 \Leftrightarrow$$
$$\frac{a(i)}{b(i)} - \frac{a'(i)}{b'(i)} \geq 0.$$

This follows from (3).

Theorem 3.6. *The ASW is a rich CQM.*

Proof: Consider every possible non-trivial clustering $\mathcal{C}$ and construct a distance function d for it such that no other clustering $\mathcal{C}'$ is as good or better in terms of the ASW.

Case I: All clusters in $\mathcal{C}$ contain more than one object. Construct d by setting $d(x_i,x_i)=0$, $d(x_i,x_j)=1$ if $x_i \sim_{\mathcal{C}} x_j,\ i\neq j$, and $d(x_i,x_j)=2$ if $x_i \not\sim_{\mathcal{C}} x_j$ for all $i,j\in X$. Then for all $i\in\mathbb{N}_n$:

$$a(i)=1,\ b(i)=2,\ s_i(\mathcal{C},d)=0.5,\ \bar{S}(\mathcal{C},d)=0.5.$$

Consider any other clustering $\mathcal{C}'$ (with quantities in the definition of the ASW denoted as $a'(i),b'(i)$). As long as $\mathcal{C}'$ contains no one-point clusters, either in $\mathcal{C}'$ there are $i\neq j$ with $x_i\sim_{\mathcal{C}} x_j$ and $d(x_i,x_j)=2$, or $i\neq j$ with $x_i\not\sim_{\mathcal{C}} x_j$ and $d(x_i,x_j)=1$, or both. This results in $a'(i)\geq 1$, $b'(i)\leq 2$, and $s_i(\mathcal{C}',d)\leq 0.5$ with strict inequality for at least one i, therefore $\bar{S}(\mathcal{C}',d)<0.5$. One-point clusters $\{x_i\}\in\mathcal{C}'$ yield $s_i(\mathcal{C}',d)=0$ by definition, and no other $s_j(\mathcal{C}',d)$ can be larger than 0.5, hence again $\bar{S}(\mathcal{C}',d)<0.5$, showing that all $\bar{S}(\mathcal{C}',d)<\bar{S}(\mathcal{C},d)$ for all $\mathcal{C}'\neq\mathcal{C}$.

Case II: $\mathcal{C}$ contains t one-point clusters, $n>t>0$, w.l.o.g, $\{x_1\},\ldots,\{x_t\}$. Construct d by setting $d(x_i,x_i)=0$, $d(x_i,x_j)=1$ if $x_i\sim_{\mathcal{C}} x_j,\ i\neq j$, $d(x_i,x_j)=2$ if $x_i\not\sim_{\mathcal{C}} x_j$, and at least one of i and j is larger than t, otherwise $d(x_i,x_j)=2+\frac{1}{2n^2}$.

Observe for $i=1,\ldots,t:\ s_i(\mathcal{C},d)=0$, for $i=t+1,\ldots,n:\ s_i(\mathcal{C},d)=0.5$, therefore $\bar{S}(\mathcal{C},d)=\frac{(n-t)0.5}{n}$.

For $t=1$, if $x_t\sim_{\mathcal{C}'} x_i$ for any $i\neq t$, then $s_1(\mathcal{C}',d)=0$ because $b'(1)=a'(1)=2$, and $s_1(\mathcal{C}',d)=0$ also otherwise. Any difference between $\mathcal{C}$ and $\mathcal{C}'$ will make at least for one $i>1:\ s_i(\mathcal{C}',d)<0.5$, because for at least one $i>1:\ x_i\sim_{\mathcal{C}'} x_j$ where $x_i\not\sim_{\mathcal{C}} x_j$, leading to $a'(i)>1, b'(i)\leq 2$, or $\{x_i\}\in\mathcal{C}'$ with $s_i(\mathcal{C}',d)=0$, and for all $i:\ s_i(\mathcal{C}',d)\leq 0.5$, so

$$\bar{S}(\mathcal{C}',d)<\frac{(n-1)0.5}{n}=\bar{S}(\mathcal{C},d).$$

If $t\geq 2$, consider first the case that there is no pair (i,j) where $i\leq t$ and $j>t$ with $x_i\sim_{\mathcal{C}'} x_j$. Because $\mathcal{C}\neq\mathcal{C}'$, there must be either (i,j) with both $i,j\leq t$ and $x_i\sim_{\mathcal{C}'} x_j$, or (i,j) with both $i,j>t$, $x_i\not\sim_{\mathcal{C}} x_j$, and $x_i\sim_{\mathcal{C}'} x_j$, or $x_i\sim_{\mathcal{C}} x_j$, and $x_i\not\sim_{\mathcal{C}'} x_j$. In all these cases, for $i>t$ always $s_i(\mathcal{C}',d)\leq 0.5$ as before, and for $i\leq t:\ a'(i)\geq 2, b'(i)\leq 2$, therefore $s_i(\mathcal{C}',d)\leq 0$. Therefore still

$\bar{S}(\mathcal{C}', d) \le \frac{(n-t)0.5}{n}$.

Now show "<", which is required to make $\mathcal{C}$ the unique maximiser of $\bar{S}$. If there exists (i,j) with both $i,j \le t$ and $x_i \sim_{\mathcal{C}'} x_j$, $s_i(\mathcal{C}', d) < 0$ because $a'(i) > 2, b'(i) \le 2$, thus $\bar{S}(\mathcal{C}', d) < \frac{(n-t)0.5}{n}$. If there exists (i,j) with both $i,j > t$, $x_i \not\sim_{\mathcal{C}} x_j$, and $x_i \sim_{\mathcal{C}'} x_j$, $a'(i) > 1, b'(i) \le 2$, therefore $s_i(\mathcal{C}', d) < 0.5$ and again $\bar{S}(\mathcal{C}', d) < \frac{(n-t)0.5}{n}$. If there exists (i,j) with both $i,j > t$, $x_i \sim_{\mathcal{C}} x_j$, and $x_i \not\sim_{\mathcal{C}'} x_j$, then $a'(i) \ge 1, b'(i) < 2$, therefore $s_i(\mathcal{C}', d) < 0.5$ and again $\bar{S}(\mathcal{C}', d) < \frac{(n-t)0.5}{n}$.

The last situation to consider is $t \ge 2$ where there exists (i,j) where $i \le t$ and $j > t$ with $x_i \sim_{\mathcal{C}'} x_j$. Then

$$2 + \frac{1}{2n^2} \ge a'(i) \ge 2, 2 + \frac{1}{2n^2} \ge b'(i) \ge 2,$$

so that $s_i(\mathcal{C}', d) \le \frac{1}{2n^2}$, and this holds for all $i \le t$ (in fact, $s_i(\mathcal{C}', d) \le 0$ as before unless $x_i \sim_{\mathcal{C}'} x_m$ with any $x_m,\ m \ge t$). Therefore, $\sum_{i \le t} s_i(\mathcal{C}', d) \le \frac{t}{2n^2}$. Furthermore, let $n'_j = |C|$, where $x_j \in C \in \mathcal{C}'$. With that,

$$2 \ge a'(j) \ge \frac{2 + (n'_j - 2)}{n'_j - 1} = 1 + \frac{1}{n'_j - 1},$$

$b'(j) \le 2$, so that

$$s_j(\mathcal{C}', d) \le \frac{2 - (1 + \frac{1}{n'_j - 1})}{2} = 0.5 - \frac{1}{2(n'_j - 1)}.$$

For $j > t$ so that there is not any $m \le t$ with $x_j \sim_{\mathcal{C}'} x_m$, still $s_j(\mathcal{C}', d) \le 0.5$. Therefore,

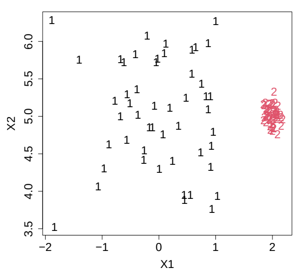
$$\sum_{i=1}^{n} s_i(\mathcal{C}', d) \le 0.5 - \frac{1}{2(n'_j - 1)} + (n - t - 1)0.5 + \frac{t}{2n^2} < (n-t)0.5$$

because $\frac{t}{2n^2} < \frac{1}{2(n'_j - 1)}$. Overall $\bar{S}(\mathcal{C}', d) < \frac{(n-t)0.5}{n}$, finishing the proof.

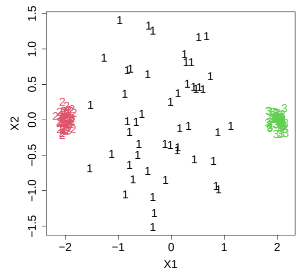
4. Simulation study

We have run a comprehensive simulation study comparing OSil, FOSil, and PAMSil with a number of well established clustering methods from the literature. We simulated from a variety of data generating processes (DGPs) with different characteristics that are listed in Table 1 and illustrated in Figure 2. A detailed description can be found in the supplementary material. 500 data sets were simulated from every DGP. The case of the number of clusters fixed at the true number is investigated as well as the case of an estimated number of clusters. We consider the achieved values of the ASW (the optimisation of which can be of interest in its own right) and the recovery of the "true" clusters using the Adjusted Rand Index (ARI; Hubert and Arabie (1985)).

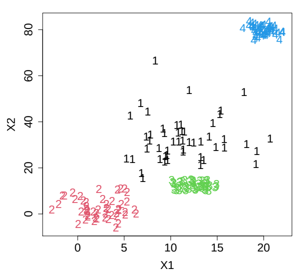
The involved clustering methods are: k-means (Lloyd (1982), Hartigan and Wong (1979), *kmeans*-function in R using `nstart=100` to stabilize the results; apart from this default parameters were used everywhere), k-medoids/Partitioning Around Medoids (PAM; Kaufman and Rousseeuw (1987), *pam*-function in R-package *cluster*; Maechler et al. (2017)), average and single linkage (Sokal and Michener (1958)), Ward's method (Ward Jr (1963), all three incorporated in function *agnes* in R-package *cluster*), spectral clustering (algorithm by Ng et al. (2001) implemented as *specc* in R-package *kernlab*; Zeileis et al. (2004)), and Gaussian mixture model-based clustering (Fraley and Raftery (1998)), using function *Mclust* in R-package *mclust*; Scrucca et al. (2017)). For PAMSil we have used the standalone C code written by the authors Van der Laan et al. (2003). The function *silhouette* from R-package *cluster* was used for computing ASW-values outside OSil.



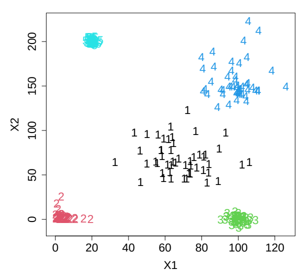
(a) DGP 1 ($k = 2$)



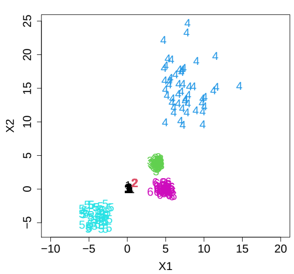
(b) DGP 2 ($k = 3$)



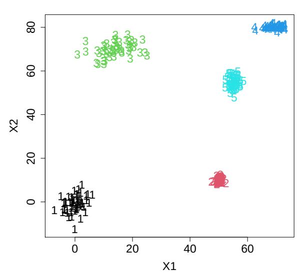
(c) DGP 3 ($k = 4$)



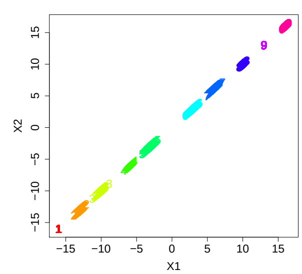
(d) DGP 4 ($k = 5$)



(e) DGP 5 ($k = 6$)



(f) DGP 6 ($k = 6$)



(g) DGP 7 ($k = 10$)

Figure 2: Plots of the first two dimensions of exemplary data sets generated by the DGPs 1-7 with true clusters indicated by colors.

Table 1: DGPs used in the simulation study. k: number of clusters, p: number of dimensions, n: number of observations. See the Appendix for precise descriptions.

DGP	k	p	Distributions	Cluster size	n
DGP1	2	2	Spherical Gaussians	50	100
DGP2	3	2	Spherical Gaussians	50	150
DGP3	4	2	t, uniform and two Gaussians	50	200
DGP4	5	2	F, χ^2, t, skew Gaussian, Gaussian	50	250
DGP5	6	2	six different distributional shapes	50	300
DGP6	5	5	five Gaussians with different within-cluster dependence structures	50	250
DGP7	10	500	10-cluster structure on 1-d hyperplane in 500-d space.	50	500
DGP8	7	60	Experiment of Van der Laan et al. (2003) simulated genes defining 3 patient groups	varying	500
DGP9	3	1000	First 100 dimensions informative spherical Gaussian	40	120

Estimating k was done by optimizing the ASW with two exceptions. Gaussian mixture model-based clustering was used with the BIC, which is the standard choice for mixtures. Exemplary for many other possible combinations of clustering method and method to estimate the number of clusters, we also ran k-means with the gap statistic (Tibshirani et al. (2001)) as implemented in the R-function *clusGap*.

4.1. Results and discussion

Results of the simulations for the DGPs 1-3 are in Table 2; for the DGPs 4-6 in Table 3, and for DGPs 7-9 in Table 4. These tables report average ARI and ASW and their estimated standard errors over the simulation runs for both fixed and estimated k as well as the percentage of runs out of those for estimated k in which the true k was estimated.

Table 2: Simulation results for DGPs 1-3. PPR is the percentage of runs in which the true k was estimated.

	Fixed k				Estimated k				
Methods	ASW	SE	ARI	SE	ASW	SE	ARI	SE	PPR
DGP1									
k-means	0.665	0.001	0.808	0.004	0.665	0.001	0.796	0.004	92
gap-k-means					0.665	0.001	0.760	0.005	70
PAM	0.665	0.001	0.834	0.003	0.666	0.001	0.816	0.004	92
average	0.622	0.004	0.687	0.016	0.655	0.001	0.812	0.007	72
Ward's	0.654	0.001	0.936	0.004	0.656	0.001	0.891	0.006	86
Single	0.410	0.007	0.167	0.016	0.545	0.005	0.785	0.013	29
BIC-mixture	0.646	0.001	0.993	0.001	0.643	0.001	0.991	0.001	99
spectral	0.643	0.003	0.923	0.005	0.656	0.001	0.906	0.009	89
PAMSil	0.667	0.001	0.848	0.006	0.668	0.001	0.816	0.008	87
OSil	0.666	0.001	0.847	0.004	0.667	0.001	0.812	0.005	86
FOSil	0.659	0.001	0.785	0.007	0.662	0.001	0.803	0.005	94
DGP2									
k-means	0.711	0.001	0.844	0.004	0.719	0.001	0.805	0.005	39
gap-k-means					0.714	0.001	0.801	0.002	39
pam	0.711	0.001	0.85	0.003	0.719	0.001	0.809	0.004	38
average	0.671	0.004	0.773	0.013	0.711	0.001	0.821	0.005	26
Ward's	0.696	0.002	0.934	0.005	0.708	0.001	0.844	0.006	30
single	0.348	0.021	0.404	0.023	0.611	0.008	0.818	0.014	11
BIC-mixture	0.652	0.002	0.863	0.002	0.682	0.002	0.991	0.001	91
spectral	0.628	0.013	0.898	0.012	0.700	0.002	0.877	0.006	50
PAMSil	0.710	0.001	0.859	0.004	0.721	0.001	0.804	0.004	22
OSil	0.712	0.001	0.856	0.004	0.722	0.001	0.806	0.004	25
FOSIL	0.705	0.002	0.67	0.015	0.714	0.001	0.815	0.006	48
DGP3									
k-means	0.674	0.002	0.815	0.005	0.764	0.001	0.318	0.001	0.2
gap-k-means					0.692	0.001	0.816	0.006	38
pam	0.702	0.001	0.912	0.001	0.765	0.001	0.319	0.001	0.2
average	0.644	0.001	0.647	0.004	0.764	0.001	0.325	0.000	0
Ward's	0.689	0.001	0.96	0.003	0.762	0.001	0.328	0.001	0
single	0.54	0.005	0.464	0.009	0.75	0.002	0.341	0.004	0.6
BIC-mixture	0.676	0.001	0.996	0.000	0.664	0.002	0.964	0.002	53
spectral	0.563	0.009	0.329	0.001	0.761	0.001	0.836	0.007	0
PAMSil	0.703	0.001	0.913	0.002	0.768	0.001	0.320	0.001	0
OSil	0.703	0.001	0.91	0.003	0.768	0.001	0.322	0.000	0
FOSil	0.699	0.001	0.502	0.002	0.767	0.001	0.32	0.000	0

Table 3: Simulation results for DGPs 4-6. PPR is the percentage of runs in which the true k was estimated.

	Fixed k				Estimated k				
Methods	ASW	SE	ARI	SE	ASW	SE	ARI	SE	PPR
DGP4									
k-means	0.727	0.006	0.887	0.007	0.791	0.002	0.962	0.003	65
gap-k-means					0.728	0.004	0.814	0.010	47
pam	0.818	0.000	0.99	0.000	0.818	0.000	0.99	0.000	100
average	0.808	0.002	0.975	0.003	0.816	0.000	0.99	0.001	93
Ward's	0.817	0.000	0.992	0.001	0.817	0.000	0.992	0.001	99
single	0.694	0.005	0.859	0.005	0.777	0.002	0.965	0.003	42
BIC-mixture	0.8	0.001	0.98	0.001	0.761	0.003	0.956	0.002	37
spectral	0.645	0.010	0.962	0.003	0.785	0.002	0.885	0.006	58
PAMSil	0.818	0.000	0.993	0.000	0.818	0.000	0.993	0.000	97
OSil	0.818	0.000	0.993	0.000	0.818	0.000	0.993	0.000	98
FOSil	0.817	0.000	0.987	0.002	0.817	0.000	0.99	0.001	99
DGP5									
k-means	0.659	0.004	0.769	0.007	0.723	0.001	0.745	0.013	35
gap-k-means					0.626	0.005	0.565	0.009	2
pam	0.743	0.001	0.957	0.003	0.745	0.000	0.975	0.001	82
average	0.581	0.001	0.342	0.006	0.700	0.000	0.232	0.010	0
Ward's	0.717	0.000	0.775	0.000	0.726	0.001	0.766	0.005	48
single	0.572	0.002	0.581	0.007	0.68	0.001	0.191	0.008	2
BIC-mixture	0.696	0.001	0.778	0.001	0.731	0.001	0.821	0.001	4
spectral	0.646	0.006	0.766	0.005	0.725	0.001	0.798	0.012	33
PAMSil	0.748	0.000	0.995	0.001	0.748	0.000	0.995	0.001	95
OSil	0.747	0.001	0.974	0.003	0.748	0.000	0.988	0.001	87
FOSIL	0.747	0.001	0.785	0.005	0.748	0.000	0.815	0.001	92
DGP6									
k-means	0.649	0.007	0.739	0.009	0.77	0.003	0.852	0.008	28
gap-k-means					0.773	0.002	0.589	0.008	7
PAM	0.865	0	1	0	0.865	0	1	0	100
average	0.865	0	1	0	0.865	0	1	0	100
Ward's	0.865	0	1	0	0.865	0	1	0	100
Single	0.865	0	1	0	0.865	0	1	0	100
BIC-mixture	0.865	0	1	0	0.865	0	1	0	100
spectral	0.618	0.023	0.881	0.016	0.797	0.007	0.834	0.016	43
PAMSil	0.865	0	1	0	0.865	0	1	0	100
OSil	0.865	0	1	0	0.865	0	1	0	100
FOSil	0.865	0	0.980	0	0.865	0	0.98	0	100

Table 4: Simulation results for DGPs 8-10. PPR is the percentage of runs in which the true k was estimated.

	Fixed k				Estimated k				
Methods	ASW	SE	ARI	SE	ASW	SE	ARI	SE	PPR
DGP7									
k-means	0.754	0.003	0.767	0.004	0.825	0.002	0.863	0.004	21
gap-k-means					0.724	0.003	0.649	0.006	3
PAM	0.921	0.001	1	0	0.921	0.001	1	0	100
average	0.921	0.001	1	0	0.921	0.001	1	0	100
Ward's	0.921	0.001	1	0	0.921	0.001	1	0	100
Single	0.92	0.001	1	0	0.921	0.001	1	0	99.6
BIC-mixture	0.921	0.001	1	0	0.860	0.001	0.896	0	0
spectral	0.584	0.031	0.798	0.017	0.830	0.011	0.854	0.032	14.6
PAMSil	0.921	0.001	1	0	0.921	0.001	1	0	100
OSil	0.921	0.001	1	0	0.921	0.001	1	0	100
FOSil	0.921	0.001	1	0	0.921	0.001	1	0	100
DGP8									
k-means	0.178	0.004	0.823	0.010	0.231	0.000	0.949	0.005	47
gap-k-means					0.180	0.003	0.747	0.010	16
PAM	0.104	0.003	0.528	0.009	0.118	0.003	0.591	0.008	20
average	0.242	0.000	0.999	0	0.242	0.000	0.999	0	100
Ward's	0.242	0.000	1	0	0.242	0.000	1	0	100
single	0.186	0.001	0.703	0.007	0.217	0.000	0.522	0.016	12
BIC-mixture	0.191	0.000	0.695	0.006	0.203	0.000	0.785	0.006	24
spectral	0.081	0.004	0.546	0.011	0.220	0.000	0.546	0.012	16
PAMSil	0.242	0.000	1	0.000	0.243	0.000	0.907	0.013	91
OSil	0.242	0.000	0.999	0	0.255	0.000	0.230	0.018	22
FOSIL	0.240	0.000	0.996	0.005	0.247	0.000	0.609	0.018	68
DGP9									
k-means	0.488	0.006	0.827	0.013	0.56	0.001	0.865	0.010	69
gap-k-means					0.074	0.003	0.654	0.006	6
PAM	0.573	0	1	0	0.573	0	1	0	100
average	0.573	0	1	0	0.573	0	1	0	100
Ward's	0.573	0	1	0	0.573	0	1	0	100
Single	0.573	0	1	0	0.573	0	1	0	100
BIC-mixture	0.573	0	1	0	0.573	0	1	0	100
spectral	0.544	0.005	0.956	0.007	0.569	0.001	0.966	0.006	92
PAMSil	0.573	0	1	0	0.573	0	1	0	100
OSil	0.573	0	1	0	0.573	0	1	0	100
FOSil	0.573	0	1	0	0.573	0	1	0	100

As could be expected, the methods for (at least locally) optimizing the ASW achieve the best ASW values, although sometimes PAM and occasionally some other methods find solutions with competitive ASW values. Occasionally, PAM can find a slightly better ASW value than FOSil, which often loses some optimization power compared with OSil and PAMSil. OSil yields sometimes higher ASW values than PAMSil, but sometimes slightly worse; for achieving the highest possible ASW value, it is advisable to run both.

Regarding the recovery of the modelled clusters, OSil and PAMSil clearly do the best job for DGP 5 compared to the other methods both for estimated and fixed k. They are best by a small margin for DGP 4 and up with the best for DGPs 6, 7, and 9, and for fixed k only for DGP 8. FOSil drops a bit in comparison in some of these. DGP 4 and 5 in particular are difficult for most standard methods because of the different distributional shapes of the clusters, with variance-covariance structures also strongly varying. As could be expected, Gaussian mixture model-based clustering is best in the setups with all or mostly Gaussian within-cluster distributions. For DGP 3 and estimated k, it is the only method that does a good job.

The performance of PAMSil and OSil for DGPs 1-3 is surprisingly relatively better for fixed k than for estimated k, contrasting with the popularity of the ASW for estimating k. In DGP 8, where Ward's method and average linkage perform best, PAMSil, OSil, and FoSil sometimes find solutions with better ASW for wrong values of k (often $k = 2$) when estimating k, which diminishes their ARI-performance for estimated k. This is in line with the skepticism that can be found in Hennig and Lin (2015) regarding naively maximizing the ASW for estimating k, see also Section 5.3. The fact that OSil can find a better ASW for another than the modelled k more often than PAMSil, leads to a disappointingly low ARI here.

Regarding the other methods and the recovery of the modelled clusters, Ward's method performs very well across the board with exception of DGP 5, where it particularly drops for estimated k. It is almost always better than k-means despite being worse at optimizing the k-means objective function. Probably the hierarchical structure helps it to adapt better to the varying cluster shapes; similarly PAMSil can occasionally do better than OSil even where the value of the objective function ASW is worse.

The other methods have some mixed results, with spectral, k-means and single linkage performing mostly clearly behind the top methods. PAM has some drops in quality but performs generally well and often similar to PAMSil and OSil; it achieves the best ASW values out of the methods that do not attempt to optimize it. The results of average linkage are overall with one exception on a slightly lower level than PAM. The Gaussian mixture drops in quality in DGPs 5 and 8, but is good in the other setups. The gap statistics with k-means achieves a good result for DGP 3 but does worse than k-means with the ASW otherwise.

Overall PAMSil and OSil do a good job, particularly with the mixed distribution shapes and fixed k, with none of the two clearly superior to the other. Some more caution is required when estimating k. The simulation shows that there are situations in which these methods are best, and therefore they can be seen as a valuable addition to the cluster analysis toolbox without generally outperforming the competition. Particularly in situations with Gaussian distributions only, Gaussian mixture-based clustering is to be preferred. Ward's method presents itself as a good allrounder. The computationally less intensive FOSil can sometimes not keep up with OSil and PAMSil, but it has some very good results estimating the number of clusters, better than OSil and PAMSil in DGPs 1, 2, and 4.

4.2. An experiment with outliers

The ASW and OSil do not rely on any distributional assumption, and it is therefore of interest whether they are less affected by outliers than methods that rely on a normal distribution assumption or on sums of squares such as k-means or Gaussian mixture model-based clustering. Unfortunately, as shown in Hennig (2008), as a method for estimating the number of clusters, the ASW is not perfectly robust. An outlier that is very far from the rest of the data can prompt the ASW to select a solution with two clusters, where one cluster is just the outlier, and the other cluster is everything else taken together. The reason for this is that if k is made larger, and less strongly separated data subsets that intuitively still qualify as "clusters" are separated, the corresponding $b(i)$-values in the ASW are much smaller than if these subsets are put together and there is another cluster far away from them, which then is the closest (for $k = 2$ in fact the only) other cluster.

The ASW can however add some robustness in case that the outliers are not that extreme. We generated two more data sets from DGP 6, one with 250 and one with 1000 observations, for which most methods did a very good job, see Table 3. We added outliers, one at $(70, 70, 30, 0, 55)$ in some distance from the clusters but not outside the original data range in any variable, and another one at $(-100, -100, -100, -100, -100)$, outlying on all variables. Four datasets, namely with 250 and 1000 original points, plus the first outlier only and two outliers, are considered. The original clustering pattern is very clear (see Figure 2), and it would be desirable to still find this structure with outliers present. OSil with estimated k isolates each outlier in a one-point cluster, leaving the original clustering intact, which seems sensible. Standard hierarchical methods do the same if the ASW is used to decide the number of clusters. FOSil needs the outliers in the random subsample used for initial clustering to arrive at the same result. Otherwise it assigns the outliers to the nearest original cluster, still leaving the original structure intact, but causing some very large within-cluster distances. The same happens with Gaussian model-based clustering for three of the four data sets. With 1000 observations and the first outlier only, the first outlier is merged with a small part of an original cluster, which is split up. k-means splits original clusterings in the presence of outliers in its solutions with $k = 5, 6$, and 7; the gap statistic chooses $k = 5$ anyway with one outlier, but $k = 1$ with two outliers. PAM and spectral clustering also avoid one-point-clusters and split original clusters and join parts with outliers if present. In these examples, the ASW both for choosing k and used by OSil and FOSil for given k treats the outliers better than competing methods, due to its ability to isolate well separated one-point (or very small) clusters.

5. Applications

Section 5.1 is devoted to four genetic data sets using the Euclidean distance, which come with ground truth information. Sections 5.2 and 5.3 explore the application of OSil, FOSil, and PAMSIL to data represented by other dissimilarity measures than the Euclidean distance.

5.1. Clustering single cell RNA sequencing data

Clustering of single cell RNA sequencing (scRNA-seq) data is a vital field. It is of interest in its own respect, and it can be used as first step for further analysis. Since much of the downstream analysis is based on clustering, the final conclusions may be strongly affected by it. Definition or discovery of new cell types via clustering is an important area of research in the field. Many different studies have been conducted on various organs either during development or at fixed time to discover several new putative cell sub-populations using novel clusters, for instance, in early embryonic development (Biase et al. (2014), Goolam et al. (2016)) or various regions of the brain (Zeisel et al. (2015)). We consider scRNA-seq data clustering by the proposed methods for a number of published data sets for which the true cell types were originally identified by the authors.

The data sets are listed in Table 5. We followed Lun et al. (2016) normalizing them. All data sets were represented by principal components (PCs). Euclidean distances on these were used for clustering. scRNA-seq data are typically of low sample size and high dimensionality, and dimension reduction is routinely applied. The number of principal components was generally chosen as q so that from q to $q+1$ PCs there was still a substantial drop in explained variance per PC, but from $q+1$ PCs onward every PC would only account for low percentages of the variance with little further drop from one PC to the next. For $j > 0$, variance PC $q+j$ was only bigger than variance PC $q+j+1$ by substantially less than factor 2.The maximum number allowed for the estimation of number of clusters was 12.

Table 5: scRNA-seq data sets

Data	n	number of genes	number of PCs	k
Yan et al. (2013)	90	20214	2	7
Biase et al. (2014)	49	25737	3	3
Goolam et al. (2016)	124	41324	4	5
Kolodziejczyk et al. (2015)	704	38577	5	3

Yan et al. (2013) is a study of human embryonic development. The authors identified 7 cell types (development stages) as oocyte (3 samples), zygote (3 samples), 2-cell (6 samples), 4-cell (12 samples), 8-cell (20 samples), lateblast (30 samples), and morula (16 samples). The first two principal components were used, which explain 80% of the variance.

Biase et al. (2014) studied the cell fate decision during early embryo development. There are 1-cell (9 samples), 2-cell (20 samples), and 4-cell (20 samples) embryos. The first three principal components were used. They explain 43% of the variance.

Goolam et al. (2016) studied pre-implantation development. There are five distinct cell types, 2-cell (16 samples), 4-cell (64 samples), 8-cell (32 samples), 16-cell (6 samples) and 32-cell (6 samples). The first four principal components were used, which explain 60% of the variance.

Kolodziejczyk et al. (2015) studied mouse embryonic stem cell growth under different culture conditions. The three culture conditions are serum (250 cells), 2i (295 cells) and 2ai (159 cells). The first five principal components were used, which explain 36% of the variance.

The first two PCs of all data sets are plotted in Figure 3. It can be seen here is that for all data sets except the one of Biase et al. (2014), the "true" clusters have clearly separated subgroups. One could therefore expect a reasonable cluster analysis method to choose more clusters than the given "true" groups, although this does not happen for the data of Yan et al. (2013), because some "true" clusters are not well separated. In case of Kolodziejczyk et al. (2015) it is actually known that there are meaningful sub-populations within each culture condition. These are subsets of the true clusters, so similarity of solutions to the true clusters as measured by the ARI is still relevant.

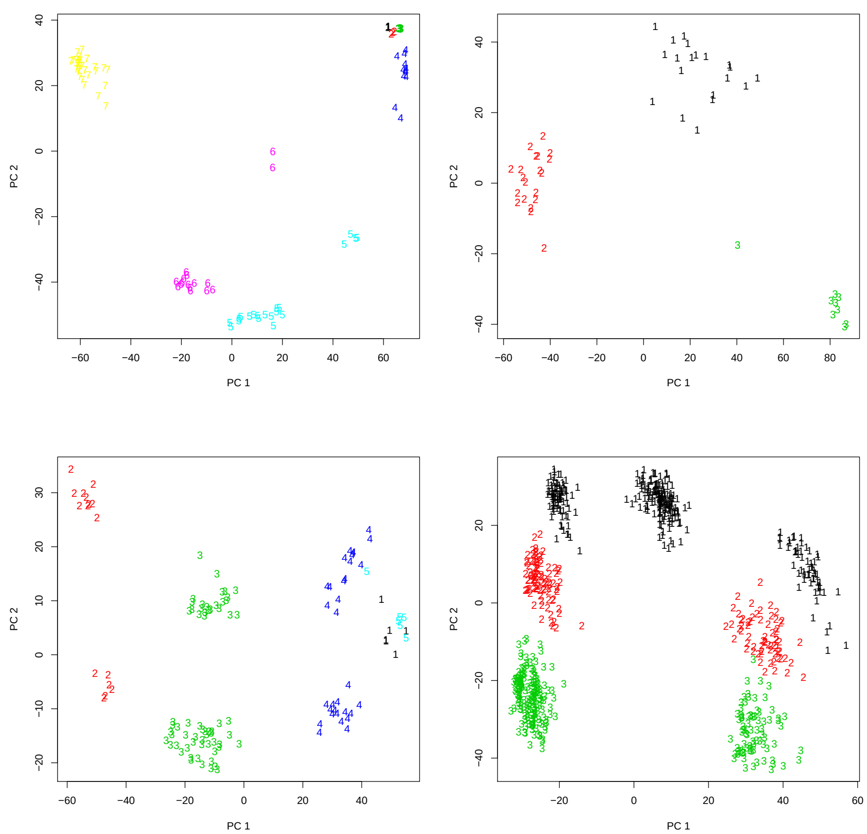


Figure 3: Plots of first two PCs with true cell types: (a) Yan et al. (2013) data, (b) Biase et al. (2014) data, (c) Goolam et al. (2016), (d) Kolodziejczyk et al. (2015) data.

Table 6: Clustering results for the scRNA-seq data sets of Yan et al. (2013) and Biase et al. (2014) for all methods included in the comparison with k fixed as the true known k, for estimated k, and averaged over all $k \in \{2, \dots, 12\}$. Best ARI values in every column are boldfaced.

	Yan et al. (2013) data					
	estimated k			fixed $k = 7$		ave. all k
Method	ASW	ARI	$\widehat{k}$	ASW	ARI	ARI
k-means	0.827	**0.796**	6	0.489	0.515	0.617
gap-k-means	0.803	0.791	5			
PAM	0.803	0.791	5	0.778	**0.894**	0.722
average	0.827	**0.796**	6	0.766	0.795	0.671
Ward's	0.827	**0.796**	6	0.778	**0.894**	0.693
BIC-mixture	0.578	0.651	8	0.766	0.795	0.663
spectral	0.827	**0.796**	6	0.716	0.842	0.565
PAMSil	0.827	**0.796**	6	0.807	0.795	0.746
OSil	0.827	**0.796**	6	0.778	**0.894**	**0.750**
FOSil	0.801	0.685	3	0.592	0.632	0.644

	Biase et al. (2014) data					
	estimated k			fixed $k = 3$		ave. all k
Method	ASW	ARI	$\widehat{k}$	ASW	ARI	ARI
k-means	0.765	**1.000**	3	0.765	**1.000**	0.581
gap-k-means	0.765	**1.000**	3			
PAM	0.765	**1.000**	3	0.765	**1.000**	0.575
average	0.765	**1.000**	3	0.765	**1.000**	0.683
Ward's	0.765	**1.000**	3	0.765	**1.000**	0.619
BIC-mixture	0.560	0.902	4	0.765	**1.000**	0.612
spectral	0.765	**1.000**	3	0.765	**1.000**	0.535
PAMSil	0.765	**1.000**	3	0.765	**1.000**	0.682
OSil	0.765	**1.000**	3	0.765	**1.000**	**0.751**
FOSil	0.765	**1.000**	3	0.765	**1.000**	0.716

Table 7: Clustering results for the scRNA-seq data sets of Goolam et al. (2016) and Kolodziejczyk et al. (2015) for all methods included in the comparison with k fixed as the true known k, for estimated k, and averaged over all $k \in \{2, \dots, 12\}$. Best ARI values in every column are boldfaced.

	Goolam et al. (2016) data					
	estimated k			fixed $k = 5$		ave. all k
Method	ASW	ARI	$\widehat{k}$	ASW	ARI	ARI
k-means	0.632	0.544	5	0.632	0.544	0.493
gap-k-means	0.698	**0.571**	7			
PAM	0.698	**0.571**	7	0.623	0.528	0.458
average	0.698	**0.571**	7	0.566	**0.842**	**0.624**
Ward's	0.690	0.566	7	0.624	0.543	0.518
BIC-mixture	0.598	0.432	9	0.632	0.544	0.400
spectral	0.698	**0.571**	7	0.617	0.562	0.436
PAMSil	0.698	**0.571**	7	0.632	0.544	0.590
OSil	0.698	**0.571**	7	0.632	0.544	0.583
FOSil	0.661	0.522	6	0.632	0.544	0.500
	Kolodziejczyk et al. (2015) data					
	estimated k			fixed $k = 3$		ave. all k
Method	ASW	ARI	$\widehat{k}$	ASW	ARI	ARI
k-means	0.525	0.451	7	0.389	0.295	0.336
gap-k-means	0.420	0.347	12			
PAM	0.528	0.524	7	0.465	**0.434**	0.423
average	0.506	0.478	12	0.465	0.389	0.399
Ward's	0.526	**0.531**	7	0.465	0.389	0.413
BIC-mixture	0.442	0.432	9	0.415	0.147	0.418
spectral	0.499	0.359	4	0.415	0.147	0.345
PAMSil	0.529	**0.531**	7	0.465	0.432	**0.438**
OSil	0.529	**0.531**	7	0.465	0.389	0.421
FOSil	0.528	0.525	7	0.465	0.432	0.412

Tables 6 and 7 show ASW and ARI results for most of the methods also compared in the simulation study (single linkage yields inappropriate results). We show ASW, ARI, and the estimated number of clusters $\widehat{k}$ for the solutions optimizing the ASW. The ARI here is the most important result, because in reality the assumption will usually be that k is unknown. We also show ASW and ARI values at fixed "true" k. As this is often not a reasonable "data analytic" number of clusters, we also show the average ARI over all k between 2 and 12 in order to investigate whether the methods based on optimizing the ASW for fixed k can find good clusters also at other numbers k.

Results differ between data sets. The Biase et al. (2014) data (Table 6) are easiest to cluster. Almost all methods find the true clustering and estimate $k = 3$ correctly, except the BIC-mixture. Some more differentiation occurs at the averaged ARI over all k. OSil yields the best result here, followed by FOSil.

For the Yan et al. (2013) data (Table 6), OSil is best in all respects, but for estimated k many other methods find the same solution. FOSil drops somewhat in quality compared with the other ASW-optimizing methods.

For the Goolam et al. (2016) data (Table 7), most methods estimate a higher k than the true $k = 5$ here, as was to be expected. k-means with ASW delivers $\widehat{k} = 5$, but the corresponding clustering is worse than most others in terms of the ARI. For estimated k, the best clustering is found by OSil, PAMSil, PAM, average linkage, k-means with the gap statistic, and spectral clustering. Regarding fixed k, average linkage (which is up with PAMSil and OSil regarding estimated k) produces a solution that is far superior to everything else, also lifting its average ARI over all k to the top spot. PAMSil and OSil follow on the next positions.

For the Kolodziejczyk et al. (2015) data (Table 7) and estimated k, OSil, PAMSil, and Ward do best. The BIC-mixture and gap/k-means are much worse. For fixed k, PAM is best, only very narrowly over PAMSil and FOSil. There are three clustering solutions with almost the same ASW 0.465 (rounded) at the true $k = 3$, namely the one found by PAM, the one found by PAMSil and FOSil, and the third one by OSil, average linkage and Ward. All these are clearly better than the solutions found by k-means, model-based clustering, and spectral clustering. Regarding the average ARI over all k, PAMSil is best, followed by PAM and OSil. This is the largest of the data sets, and FOSil, which is meant for larger data sets, loses less quality here compared with OSil than for the smaller data sets, and is substantially faster.

Considering ASW-values, as in the simulation study, PAMSil is sometimes better and sometimes worse than OSil. Overall, in terms of optimizing the ASW, PAMSil does a good job, and in practice one may run them both and then pick the one that yields the better ASW. The ASW does a generally good job picking the number of clusters; almost all ARI values with k estimated by optimizing the ASW are clearly better than the ARI at true k or averaged over k, and the ASW seems to be far more suitable here than model-based clustering with the BIC, probably because the BIC tries to approximate non-Gaussian clusters by more than one Gaussian component. The BIC always yields the largest $\widehat{k}$ here leading to a usually weak ARI.

OSil and PAMSil show the best performances overall. Regarding estimated k they are the only methods that find the best solution for all four data sets, with some good results for average linkage, Ward, and PAM. These three each miss the best solution just once, but do occasionally much worse averaged over all k.

5.2. Species delimitation of Veronica plants

The Veronica data set analysed here is from Martinez-Ortega et al. (2004). It gives genetic information about 207 individual Veronica plants of sub-genus *Pentasepalae* from the Iberian peninsula and Morocco. The aim of clustering these is to discover and delimit different species of such plants. The plants are characterized by 583 variables. These contain genetic information, which was obtained using AFLP-technology ("amplified fragment length polymorphism"). This detects the presence or absence of certain characteristics ("markers") of DNA fragments (AFLP bands from 61 to 454 bp). The variables can take the values 0 (absence) and 1 (presence). As joint presences are the key information and far more relevant than joint absences, we computed a Jaccard distance matrix (see, e.g., Shi (1993)) on the individuals. Involving also further information, Martinez-Ortega et al. (2004) gave a "true" species classification into eight species. Figure 4 shows a 2-dimensional classical multidimensional scaling (MDS) representation of the data. The left side shows the true species, which look very much in line with the data. Running OSil and PAMSil on the data, estimating the number of clusters, delivers $\widehat{k} = 8$ clusters that are exactly identical to those true clusters. The ASW picks the same solution out of the average linkage and complete linkage dendrogram. As this is distance data, mclust and k-means are not available in their standard form. The ASW was in Kaufman and Rousseeuw (1990) proposed for use together with PAM, but with PAM it suggests 7 clusters, merging somewhat counter-intuitively true species 5 and 8. The reason is that the 8-cluster PAM solution, shown on the right side of Figure 4, is even worse in terms of the ASW. The PAM objective function for this solution is better than for the true 8-cluster solution, so the problem is not that the PAM algorithm would not find a global optimum. Inspecting individual distances, it turns out that according to the PAM objective function it is better to fit the 48 members of true species 3 by two centroids, sacrificing the only 4 members of true species be merging them with a part of true cluster 3.

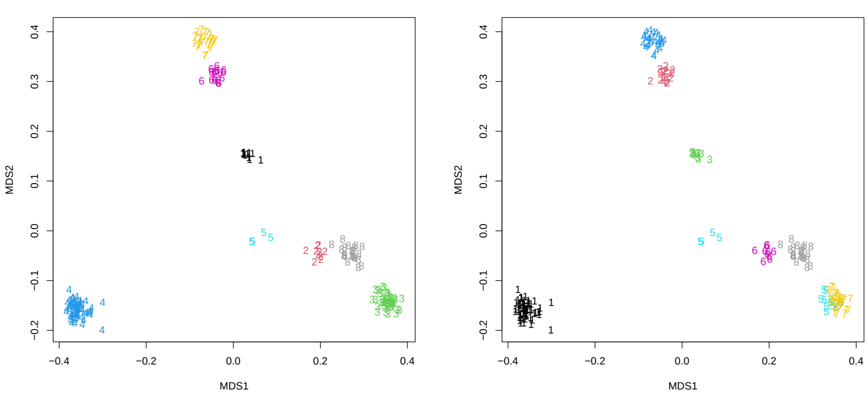


Figure 4: 2-dimensional multidimensional scaling representation of the Veronica plants data. Left side: "True" species, identical to the solution found by OSil and PAMSil. Right side: Solution found by PAM for 8 clusters.

Optimizing the ASW also for fixed number of clusters $k = 8$ stops this from happening, because the large distance within PAM-8 cluster 5 spoils the homogeneity ($a(i)$) values in the silhouette, whereas separating true clusters 5 and 3 is good for the separation ($b(i)$). The OSil solution is not only better in line with the "true" species, but also with the impression given by the MDS plot, and going through individual distances.

This is an example how optimizing the ASW can be beneficial, particularly compared to PAM, for a non-Euclidean distance. Other methods (e.g., average and complete linkage) contain the true 8-species solution in the dendrogram, but require the ASW to pick $k = 8$.

5.3. France rainfall data clustering

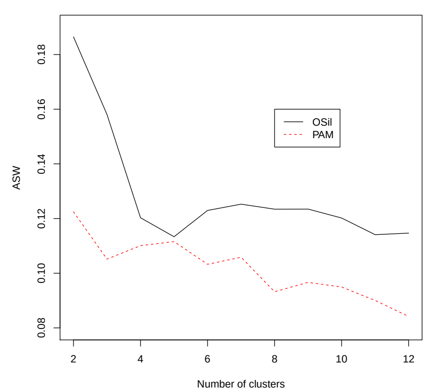


Figure 5: France rainfall data: ASW for clusterings from 2 to 12 clusters by OSil and PAM.

Finding spatial or temporal patterns in climate data sets based on statistical techniques is of crucial importance for climatologists. The data analysed here is taken from Bernard et al. (2013), who clustered 92 French weather stations based on rainfall for the three months of fall, September to November from 1993 to 2011, considering weekly maxima of hourly precipitation, resulting in time series of length 288. Based on subject matter considerations, they proposed a specific distance measure, the F-madogram, and then ran a PAM clustering on the data. They used the ASW in a rather exploratory manner. For PAM, the optimum number of clusters is $k = 2$ (see Figure 5), however the authors are also interested in more than two clusters, and they also show and interpret solutions for $k = 5$ and $k = 7$, which are local optima of the ASW. As Figure 5 shows, OSil yields consistently higher values of the ASW. The two-cluster solution of OSil just separates two outlying weather stations from the rest, which is not very useful.

This points to a weakness of the ASW, which has a tendency to favor $k = 2$ or a low k if one or more clearly separated clusters (potentially small) can be found, ignoring less strongly separated structure elsewhere, see Section 4.2.

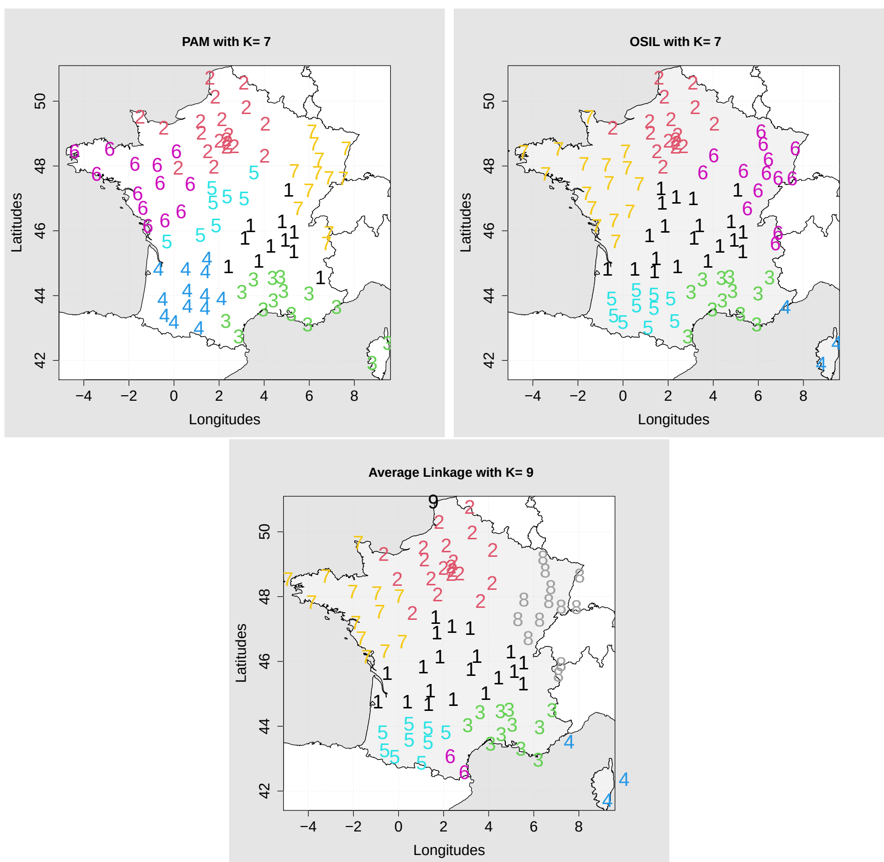


Figure 6: France rainfall data: Clusterings for PAM and OSil ($k = 7$) and average linkage ($k = 9$).

A simpler approach is to look for local optima, and OSil delivers a local optimum at $k = 7$ (see Figure 5). Figure 6 shows the 7-cluster solutions for PAM (as used in Bernard et al. (2013)), OSil, and (for comparison) a local ASW optimum at $k = 9$ of average linkage. There is no true clustering given for these data, but one can look at individual distances to assess the different clusterings.

OSil cluster 4 has just three stations, two of them on Corsica, with average within-cluster distance (awcd) 0.095. OSil cluster 3 has an awcd of 0.091. PAM cluster 3 is mainly a merger of these two, and has a much larger awcd of 0.101. There is an average distance of 0.115 between OSil clusters 3 and 4, whereas the average distance between PAM clusters 3 and 1 is just 0.111. This means that OSil clusters 3 and 4 actually seem well separated and leaving them separated is more convincing that merging them, as PAM does. On the other hand, OSil cluster 1 is much bigger than the roughly corresponding PAM clusters 1 and 5. The awcds of these two PAM clusters are 0.085 and 0.086, respectively. OSil cluster 1 has awcd 0.093, which is admittedly bigger, however three of the seven PAM clusters and only two OSil clusters have awcds larger than that (the largest awcd of any cluster of both clusterings is 0.101 of PAM cluster 3). The average distance to the closest different cluster for OSil cluster 1 is 0.108, and not only is this larger than the average distance 0.098 between PAM cluster 1 and 5, also both are closer on average (0.107 and 0.104) to the closest of the remaining five PAM-clusters. Overall the OSil clustering has a smaller average within-cluster distance, and a larger average between-cluster distance, and it looks overall more convincing. OSil focuses more on homogeneity and separation, whereas PAM focuses more on representation by the best centroid, which is arguably not that important in this application.

We ran some other distance-based clustering methods on these data. PAMSil, as usually, performed similar to OSil, with an ASW optimum at $k = 2$ and a local optimum at $k = 7$, however the ASW maximum found by OSil was not found and PAMSil's solution is a bit different and slightly worse in ASW. All tried out methods had their ASW optimum at $k = 2$. Figure 6 shows the local

ASW optimum at $k = 9$ for average linkage (only local optimum for $k \leq 12$, which was the largest k we tried). The smaller clusters 6 (2 stations) and 9 (just one station) in this solution have average distances to the closest other cluster of 0.113 and 0.108, smaller than the corresponding value between OSil clusters 3 and 4, so they are hardly strongly outlying, and arguably not that useful. Complete and single linkage do not yield very convincing solutions either.

Lacking "true" cluster information (as is typically the case in real applications), such arguments elaborate in what sense the OSil method can achieve something valuable for these data that is not achieved by the other methods. This is only achieved after accounting for a weakness of the ASW for estimating k (for which it is regularly used), namely that it can get stuck at $k = 2$ because of large within-cluster distances that hide smaller but still relevant separation at a higher level of k. In real data analysis, higher local optima should also be explored, be it with OSil, or be it where the ASW is used with other methods for choosing k.

6. Conclusion

We introduce the OSil and FOSil methods for optimum ASW clustering. The ASW is shown to fulfil the desirable axioms for clustering quality measures proposed by Ackerman and Ben-David (2009). From our experiments, the ability of OSil to find "true" clusters is good, but there are exceptions. It was the strongest method for DGPs containing clearly separated clusters with differing spreads and sizes. For several further models, including those with higher dimensionality, it performed well, and in line with most other methods. Results for the scRNA-seq data were overall best. An issue, highlighted in Sections 4.2 and 5.3, but also present in some simulations, is that the ASW as an estimator of the number of clusters k can be tempted to choose a too low k if this allows for very large distances between "neighboring" clusters, which can hide structure that is still meaningful but characterized by somewhat lower between-cluster distances. This is a problem not only with OSil, but with the widespread general use of the ASW as a criterion to estimate k. It is advised to consider locally optimal values of k regarding the ASW on top of the global optimum where the global optimum is at $k = 2$ or very low. A bootstrap scheme to estimate the number of clusters with the ASW correcting for bias in favor of low k has been proposed in Hennig and Lin (2015). On the other hand, the applications to non-Euclidean data show that OSil can achieve sensible results where PAM and other methods have difficulties. Also the ASW did better finding good clusterings at estimated k for the scRANA-seq data than the BIC with a Gaussian mixture model.

The PAMSil algorithm by Van der Laan et al. (2003) is a good approximation to OSil, often finding the same optimum ASW, sometimes worse, sometimes even better. For larger data sets, the FOSil algorithm based on subsetting has been proposed, which however occasionally results in considerable quality loss. As long as the optimisation of the ASW is of interest in its own right, FOSil still usually achieves a better ASW-value than other clustering methods apart from the slower OSil and PAMSil.

Software

An R package is available at the first author's Github site at `https://github.com/bfatimah`. Code for PAMSil was built on the standalone C-code by Van der Laan et al. (2003).

Acknowledgement

The research of the first author was funded by the Commonwealth Scholarship Commission.

References

Ackerman, M., Ben-David, S., 2009. Measures of clustering quality: A working set of axioms for clustering, in: Advances in neural information processing systems, pp. 121–128.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Perez, J.M., Perona, I., 2012. An extensive comparative study of cluster validity indices. Pattern Recognition 46, 243–256.

Batool, F., Hennig, C., 2019. Clustering by optimizing the average silhouette width. arXiv preprint arXiv:1910.08644 .

Bernard, E., Naveau, P., Vrac, M., Mestre, O., 2013. Clustering of maxima: Spatial dependencies among heavy rainfall in france. Journal of Climate 26, 7929–7937.

Biase, F.H., Cao, X., Zhong, S., 2014. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell rna sequencing. Genome research 24, 1787–1796.

Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Communications in Statistics-Theory and Methods 3, 1–27.

Carlsson, G., Mémoli, F., 2013. Classifying clustering schemes. Foundations of Computational Mathematics 13, 221–252.

Correa-Morris, J., 2013. An indication of unification for different clustering approaches. Pattern Recognition 46, 2548–2561.

Fisher, L., Ness, J.W.V., 1971. Admissible clustering procedures. Biometrika 58, 91–104.

Fraley, C., Raftery, A.E., 1998. How many clusters? which clustering method? answers via model-based cluster analysis. The Computer Journal 41, 578–588.

Goolam, M., Scialdone, A., Graham, S.J., Macaulay, I.C., Jedrusik, A., Hupalowska, A., Voet, T., Marioni, J.C., Zernicka-Goetz, M., 2016. Heterogeneity in oct4 and sox2 targets biases cell fate in 4-cell mouse embryos. Cell 165, 61–74.

Halkidi, M., Vazirgiannis, M., Hennig, C., 2015. Method-independent indices for cluster validation and estimating the number of clusters, in: Hennig, C., Meila, M., Murtagh, F., Rocci, R. (Eds.), Handbook of Cluster Analysis. CRC Press, pp. 595–618.

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

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

Hennig, C., 2015. What are the true clusters? Pattern Recognition Letters 64, 53–62.

Hennig, C., Lin, C.J., 2015. Flexible parametric bootstrap for testing homogeneity against clustering and assessing the number of clusters. Statistics and Computing 25, 821–833.

Hennig, C., Meila, M., Murtagh, F., Rocci, R., 2015. Handbook of Cluster Analysis. CRC Press, Boca Raton, USA.

Hubert, L., Arabie, P., 1985. Comparing partitions. Journal of Classification 2, 193–218.

Jardine, N., Sibson, R., 1968. The construction of hierarchic and non-hierarchic classifications. The Computer Journal 11, 177–184.

Kaufman, L., Rousseeuw, P., 1987. Clustering by means of medoids. Amsterdam: North-Holland.

Kaufman, L., Rousseeuw, P.J., 1990. Finding groups in data: an introduction to cluster analysis. volume 344. John Wiley & Sons.

Kleinberg, J.M., 2003. An impossibility theorem for clustering, in: Advances in neural information processing systems, pp. 463–470.

Kolodziejczyk, A.A., Kim, J.K., Tsang, J.C., Ilicic, T., Henriksson, J., Natarajan, K.N., Tuck, A.C., Gao, X., Bühler, M., Liu, P., et al., 2015. Single cell rna-sequencing of pluripotent states unlocks modular transcriptional variation. Cell stem cell 17, 471–485.

Van der Laan, M., Pollard, K., Bryan, J., 2003. A new partitioning around medoids algorithm. Journal of Statistical Computation and Simulation 73, 575–584.

Lloyd, S., 1982. Least squares quantization in pcm. IEEE Transactions on Information Theory 28, 129–137.

Lun, A.T., McCarthy, D.J., Marioni, J.C., 2016. A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. F1000Research 5.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., 2017. cluster: Cluster Analysis Basics and Extensions. R package version 2.0.6.

Martinez-Ortega, M.M., Delgado, L., Albach, D.C., Elena-Rossello, J.A., Rico, E., 2004. Species boundaries and phylogeographic patterns in cryptic taxa inferred from aflp markers: Veronica subgen. pentasepalae (scrophulariaceae) in the western mediterranean. Systematic Biology 29, 965–986.

Ng, A.Y., Jordan, M.I., Weiss, Y., 2001. On spectral clustering: Analysis and an algorithm, in: NIPS, pp. 849–856.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65.

Rubin, J., 1967. Optimal classification into groups: an approach for solving the taxonomy problem. Journal of Theoretical Biology 15, 103–144.

Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E., 2017. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R Journal 8, 205–233.

Shi, G.R., 1993. Multivariate data analysis in palaeoecology and palaeobiogeography-a review. Palaeogeography, Palaeoclimatology, Palaeoecology 105, 199–234.

Sokal, R.R., Michener, C.D., 1958. A statistical method for evaluating systematic relationships. Univesity Kansas Science Bulletin 38, 1409–1438.

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

Von Luxburg, U., Williamson, R.C., Guyon, I., 2012. Clustering: Science or art?, in: Proceedings of ICML Workshop on Unsupervised and Transfer Learning, pp. 65–79.

Ward Jr, J.H., 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58, 236–244.

Wright, W.E., 1973. A formalization of cluster analysis. Pattern Recognition 5, 273–282.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., et al., 2013. Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. Nature structural & molecular biology 20, 1131.

Zadeh, R.B., Ben-David, S., 2009. A uniqueness theorem for clustering, in: Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence, AUAI Press. pp. 639–646.

Zeileis, A., Hornik, K., Smola, A., Karatzoglou, A., 2004. kernlab —an s4 package for kernel methods in r. Journal of Statistical Software 11, 1–20.

Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al., 2015. Cell types in the mouse cortex and hippocampus revealed by single-cell rnaseq. Science 347, 1138–1142.