ENGINEERING MATHEMATICS
AND COMPUTING LAB

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Mathematical Clustering Based on Cross-Sections in Medicine: Application to the Pancreatic Neck

Saskia Haupt, Nassim Fard-Rutherford, Philipp D. Lösel, Lars Grenacher, Arianeb Mehrabi, Vincent Heuveline

Preprint No. 2020-01

Preprint Series of the Engineering Mathematics and Computing Lab (EMCL)

$\frac{\partial u}{\partial t} - \nu\Delta u$
$+ (u \cdot \nabla) u + \frac{1}{\rho}\nabla p$
$\frac{\partial u}{\partial t}$

http://emcl.iwr.uni-heidelberg.de

Preprint Series of the Engineering Mathematics and Computing Lab (EMCL)
ISSN 2191–0693
Preprint No. 2020-01

Affiliation of the Authors

Saskia Haupt[a,1], Nassim Fard-Rutherford[b], Philipp D. Lösel[a,c], Lars Grenacher[d], Arianeb Mehrabi[e], Vincent Heuveline[a]

[a] *Engineering Mathematics and Computing Lab (EMCL), Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Germany*
[b] *Zentralinstitut für diagnostische und interventionelle Radiologie, Stadtklinikum Ludwigshafen, Ludwigshafen, Germany*
[c] *Heidelberg Institute for Theoretical Studies (HITS), Heidelberg, Germany*
[d] *Conradia Radiologie München, Munich, Germany*
[e] *General, Visceral and Transplantation Surgery, Heidelberg University, Germany*
[1] *Corresponding Author: Saskia Haupt, saskia.haupt@uni-heidelberg.de*

Impressum

Heidelberg University
Interdisciplinary Center for Scientific Computing (IWR)
Engineering Mathematics and Computing Lab (EMCL)

Im Neuenheimer Feld 205,
69120 Heidelberg
Germany

Mathematical Clustering Based on Cross-Sections in Medicine: Application to the Pancreatic Neck

Saskia Haupt, Nassim Fard-Rutherford, Philipp D. Lösel, Lars Grenacher,

Arianeb Mehrabi, Vincent Heuveline

April 22, 2020

Abstract

In the context of current surgical techniques, the classification of 3D organs based on two-dimensional cross-sections is a decisive and still challenging task. The goal of this paper is to explore an approach to address this problem. By this means, the expectation is to go further in the direction of patient-specific surgery. Based on two-dimensional image data, we analyze different clustering results assuming specific evaluation criteria. By doing so, a determination of the most appropriate number of clusters is possible. As an example, we use this method to classify the shape of the neck of the pancreas of humans, which is relevant for different types of distal pancreatectomy. Hereby, scaling issues of the available data are a key point. Therefore, an overall protocol needs to care for comparable data.

1 Introduction

A common goal in data analysis is to cluster a given data set into different groups. This has become increasingly important in different application fields, for example in medicine in order to group different human organs based on their shapes.

The general idea of clustering is to partition a given data set $\mathcal{X} \subset \mathbb{R}^d$ into $k \in \mathbb{N}$ different groups, called clusters $C_i, i = 1, \dots, k$. The latter are represented by their centroids $c_i, i = 1, \dots, k$ and the partition is denoted by $\mathcal{C}$.

Hereby, two main goals are pursued: Points within one cluster should be close together. Simultaneously, two different clusters should be far from each other. There are different ways to quantify these distances, e.g. for intra-cluster distances we can use the average distance of different points in this cluster or the average distance of the points to the centroid of the cluster. Further, for inter-cluster distances we can choose the minimal distance of the centroids or the average distance to the nearest neighbor and so on. Hereby, the distances can be measured in arbitrary metrics. It is important to mention that these goals should be both pursued at the same time. Otherwise, the optimal number of clusters may always be one, if we only want to maximize the inter-cluster distances. If we instead only want to minimize the intra-cluster distances, it may happen that we get as many clusters as we have data points.

In different clustering approaches, different metrics are defined and the resulting clustering is chosen in such a way that we go further in the direction of minimum intra-cluster distances and maximum inter-cluster distances. As it turns out no metric is significantly better than the other, but not all of the mentioned metrics can be optimized. In order to obtain better results, one should have a closer look at supervised learning approaches.

2 Clustering with k-means++ Algorithm

In this section, we have a closer look at the following objective function

$$\phi(\mathcal{C}) = \sum_{x \in \mathcal{X}} \min_{c_i \in \mathcal{C}} \|x - c_i\|^2,$$

which should be minimized by choosing the k centroids c_i of the corresponding clusters $C_i, i = 1, \dots, k$ in an appropriate way.

Since this is an combinatorial optimization problem, it is NP-hard to solve it exactly. A possibility to overcome this problem is to use an iterative scheme which yields a locally optimal solution. This is done by the k-means clustering algorithm [1] consisting of an initialization and an update procedure. In the standard k-means algorithm by Lloyd [6], the initial centroids are chosen in an arbitrary way. An extension of this algorithm, called k-means++ algorithm, developed by Arthur and Vassilvitskii [1] aims at finding better starting values and then uses the remaining update routine of Lloyd's algorithm to minimize ϕ. This results in Algorithm 1, whereby $D(x)$ denotes the shortest distance from a data point x to the closest centroid c_k that is already chosen.

Algorithm 1: k-means++

```
begin
    /* Initialization of centroids                                          */
    Choose an initial centroid $c_1$ uniformly at random from $\mathcal{X}$
    for $i \in \{2, \dots, k\}$ do
        choose the centroid $c_i$ by selecting $c_i = x' \in \mathcal{X}$ with probability $P(x) \sim D(x)^2$
    /* Update scheme                                                        */
    while $\mathcal{C}$ changes do
        for $i \in \{1, \dots, k\}$ do
            $C_i \longleftarrow$ set of points in $\mathcal{X}$ that are closer to $c_i$ than they are to $c_j$ for all $j \neq i$
        for $i \in \{1, \dots, k\}$ do
            $c_i \longleftarrow$ centroid of all points in $C_i$: $c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$
```

Characteristics. The algorithm is widely used in practice due to its speed and relative accuracy. No clustering will be repeated during the course of the algorithm since the potential function ϕ is monotonically decreasing. Further, the algorithm will always terminate because there are at most k^n possibilities of determining clusters, where n denotes the number of samples. This implies the convergence of the k-means++ algorithm, but a global convergence can not be guaranteed as the results depend on the starting values. In practice, only a few iterations are usually required which makes the k-means algorithm very fast and therefore attractive.

Unfortunately, there are many examples for which the algorithm generates arbitrarily bad clusterings, meaning that there is unsatisfying accuracy. However, by choosing appropriate starting values using k-means++, desired approximation can be guaranteed. The algorithm remains fast and simple. Arthur and Vassilvitskii [1, Section 3] proved that the total error compared to the ground truth clustering in expectation is at most $\mathcal{O}(\log k)$ which is a tight bound. In fact, they proved that this holds only after the initialization of the algorithm above. The update step can then only decrease ϕ.

Since the solution of k-means++ depends on the starting values, it is possible that only local minima are reached. This is related to the fact that the set of data should not contain too many outliers since the algorithm can not detect them and the resulting clusters would probably be displaced. In order to find a good clustering, the algorithm has to be replicated and the solution with the lowest total sum of distances, meaning the minimizer of $\phi(\mathcal{C})$ over all replications has to be chosen.

Another disadvantage arises from the fact that the number of clusters k has to be chosen a priori as the k-means++ algorithm needs k as an input. Since the number of clusters is often not known in applications, there are several strategies to find an appropriate number of clusters only based on the data. This is done by choosing different metrics for the intra- and inter-cluster distances and try to find the best possible clustering result among different numbers of clusters based on these metrics. The goal of this paper is to examine the most appropriate number of clusters for our application of clustering shapes of human organs by the example of the pancreas. We will have a closer look at some of these distance functions in the next section.

3 Clustering Evaluation

As already mentioned, the `k-means++` algorithm gets the number of clusters k as an input. From an application point of view, the best possible number of clusters is often not known a priori. Indeed, one goal of this paper is to evaluate different clustering criteria and based on this, determine an appropriate number of clusters. Obviously, the value of ϕ is not a good choice as a clustering evaluation value since ϕ decreases as k increases. This is due to the fact that the more centroids of clusters exist, the smaller the distances to the centroids become. In the following, we will explore four different clustering evaluation criteria, which are based on the relation of intra- and inter-cluster distances. Hereby, the definition of the distance functions varies among the different criteria.

3.1 Silhouette Coefficients

The first way to choose k is by using silhouette values. These values compare for each data point the distance to the centroid and the distances to centroids of other clusters.

Description. For this purpose, generalized distances have to be defined in the following way, according to Rousseeuw [7, Section 2] and Ester and Sander [4, pp. 65–66]: Let o be any object in the data set $\mathcal{X}$, and denote by A the cluster to which it has been assigned. Then, the following quantities can be computed:

$$a(o) = \frac{1}{|A|} \sum_{x \in A} \|o - x\|^2,$$

i.e. the average distance of o to all other objects of its own cluster A.
Considering now an arbitrary cluster C_i different from A yields:

$$d(o, C_i) = \frac{1}{|C_i|} \sum_{x \in C_i} \|o - x\|^2,$$

i.e. the average distance of o to all objects of C_i.
After computing $d(o, C_i)$ for all clusters $C_i \neq A$, the smallest of those numbers is selected and denoted by:

$$b(o) = \min_{C_i \neq A} d(o, C_i).$$

The cluster B for which this minimum is attained is called the neighbor of object o. This is due to the fact that it would be the closest competitor if o could not be accommodated into cluster A.

Definition 1 (Silhouette Value $s(o)$)**.** Using these distances, the silhouette value $s(o)$ of an object $o \in A$ is defined in the following way:

$$s(o) = \begin{cases} 0 & \text{if } a(o) = 0, \\ \dfrac{b(o) - a(o)}{\max\{a(o), b(o)\}} & \text{else.} \end{cases}$$

It holds that

$$-1 \leq s(o) \leq 1 \qquad \forall \text{ objects } o \in A.$$

Characteristics. The following situations can be helpful for obtaining a better intuition of the meaning of $s(o)$ assuming $a(o) \neq 0$:

1. $s(o) \approx 1$: o is 'well-clustered',
2. $s(o) \approx 0$: o lies equally far away from both, A and B, considered as an 'intermediate case',
3. $s(o) \approx -1$: o is 'misclassified'.

Therefore, $s(o)$ measures how well object o matches the clustering at hand, meaning the closer $s(o)$ is to 1, the better is the assignment of o to its cluster.

Plotting the points in the several clusters over the silhouette values for each of them, yields a *silhouette plot*. Comparing several of these plots shows which division into clusters is probably most natural. Thus, it is an indication of which number of clusters should be chosen. To be more precise, the average over all $|C_i|$ silhouettes of a cluster C_i, $i \in \{1, \dots, k\}$ is a measure for the quality of the cluster. It is called the *silhouette width* of C_i and is defined by:

$$s_{C_i} = \frac{1}{|C_i|} \sum_{o \in C_i} s(o).$$

Definition 2 (Silhouette Width $s_k(\mathcal{X})$)**.** The silhouette width for the entire data set $\mathcal{X}$ for a given number of clusters k is defined by:

$$s_k(\mathcal{X}) = \frac{1}{|\mathcal{X}|} \sum_{C_i \in \mathcal{C}} \sum_{o \in C_i} s(o).$$

The larger the silhouette width for the whole set of data, the better is the clustering. Thus, $s_k(\mathcal{X})$ is computed for all possible k and the maximum silhouette width is chosen and called the *silhouette coefficient*.

Definition 3 (Silhouette Coefficient SC)**.** The *silhouette coefficient* SC is defined as the maximum silhouette width. In formulas, this reads:

$$SC = \max_k s_k(\mathcal{X}), \quad \text{where } k \in \{1, \dots, n\}.$$

The SC is a useful measure of the amount of clustering structure that has been discovered by the respective classification algorithm.

Kaufman and Rousseeuw [5, Table 4 on p. 88] propose the following interpretation of the silhouette coefficient:

SC	Proposed Interpretation
0.71 – 1.00	A strong structure has been found.
0.51 – 0.70	A reasonable structure has been found.
0.26 – 0.50	The structure is weak and could be artificial; please try additional methods on this data set.
$\leq$ 0.25	No substantial structure has been found.

3.2 Caliński-Harabasz Criterion

Next, we will give an introduction in the clustering evaluation concept of Caliński and Harabasz [2] established in 1974. This criterion is based on the so called *variance ratio criterion*. Hereby, the within-cluster variance should be minimized, whereas the between-cluster variance should be maximized. We give the definition of those measures.

Definition 4 (Within-Cluster Variance CV_W)**.** Considering a division into k clusters $C_i, i = 1, \dots, k$ of a given data set $\mathcal{X} \in \mathbb{R}^d$. Each cluster contains $n_i, i = 1, \dots, k$ points, where the centroids are denoted by $c_i, i = 1, \dots, k$. Then, the *within-cluster variance* CV_W is defined as

$$CV_W = \sum_{i=1}^{k} \sum_{o \in C_i} \|o - c_i\|^2,$$

where the inner sum is performed over all objects $o \in C_i$.

Definition 5 (Between-Cluster Variance CV_B)**.** Using the notations above and denote the overall mean of the data set by $\bar{c}$, then the *between-cluster variance* CV_B is defined analogously:

$$CV_B = \sum_{i=1}^{k} n_i \|c_i - \bar{c}\|^2.$$

Having these two definitions by hand, we can define the Caliński-Harabasz criterion.

Definition 6 (Caliński-Harabasz Criterion)**.** The *Caliński-Harabasz criterion* CH_k for k clusters, also called *variance ratio criterion*, is given by:

$$CH_k = \frac{CV_B}{CV_W} \cdot \frac{|\mathcal{X}| - k}{k - 1},$$

where $|\mathcal{X}|$ denotes the number of elements of the data set $\mathcal{X}$.

Thus, for choosing the best possible number of clusters k, we compute for each clustering result the Caliński-Harabasz criterion value and plot it against k. Then, k is determined in such a way that the Caliński-Harabasz criterion value is maximized or at least has a strong increase compared to lower values of k. If there are several local maxima, Caliński and Harabasz [2] suggest to choose the smallest of those k in order to keep the computational costs at a minimum.

3.3 Davies-Bouldin Index

The following subsection describes the Davies-Bouldin index, which is first introduced by Davies and Bouldin [3] in 1979. The latter serves as a basis for this subsection. The index is used to compare relative goodness of different numbers of clusters. However, a good value does not mean that the considered division in a given number of clusters is appropriate in absolute values.

The Davies-Bouldin index is based on the ratio of a measure of scatter S_i within a cluster C_i and a measure of separation $M_{i,j}$ of clusters C_i and C_j, i.e. the relation of within-cluster and between-cluster distances.

Definition 7 (Measure of Scatter S_i)**.** We consider a data set $\mathcal{X} \subset \mathbb{R}^d$ partitioned into k clusters $C_i, i = 1, \dots, k$ with centroids c_i. Then the measure of scatter S_i for cluster C_i is defined by

$$S_i = \frac{1}{|C_i|} \sum_{o \in C_i} \|o - c_i\|_p,$$

where $o \in C_i$ is an object in cluster C_i.

Often, $p = 2$, such that the norm is the Euclidean norm. Hereby, it is important that this choice of distance function matches with the choice of metric done in the clustering itself in order to obtain meaningful results.

Definition 8 (Measure of Separation $M_{i,j}$)**.** Using the notations of the definition of S_i, the measure of separation $M_{i,j}$ of two clusters $C_i, C_j \in \mathcal{C}$ is given by

$$M_{i,j} = \|c_i - c_j\|_p.$$

In order to obtain a good clustering in relative terms, S_i should be as small as possible and $M_{i,j}$ as large as possible. Further, the resulting measure for the within-to-between cluster ratio should be symmetric and non-negative. With this, the Davies-Bouldin index and the corresponding within-to-between cluster ratio are defined as follows.

Definition 9 (Davies-Bouldin Index)**.** The Davies-Bouldin index introduced in [3] for clusters $C_i, C_j \in \mathcal{C}$ is given by

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} D_{i,j},$$

where $D_{i,j}$ describes the above mentioned ratio of scattering within clusters and separation between clusters:

$$D_{i,j} = \frac{S_i + S_j}{M_{i,j}}.$$

If we consider different clustering results, i.e. different numbers of clusters k, the clustering results with the smallest Davies-Bouldin index is the best. Thus, this index can be used to determine the value of k in the `k-means++` algorithm by plotting it against the number of clusters k. Then, k is chosen such that this index is minimized. However, how good this clustering is in absolute values, can not be determined using this evaluation method.

3.4 Gap Criterion

The last criterion we focus on here is the gap criterion which is a formalization strategy of the *elbow method*. It is first published by Tibshirani et al. [8] in 2000.

Let W_k be an arbitrary error measure for a clustering approach given by

$$W_k = \sum_{i=1}^{k} \frac{1}{2|C_i|} D_i,$$

where

$$D_i = \sum_{o,o' \in C_i} d_{oo'}$$

denotes the sum of pairwise distances for all objects o, o' in cluster $C_i, i = 1, \ldots, k$. If this distance is the Euclidean distance, it corresponds to the measure of scatter S_i introduced in the Davies-Bouldin index.

If the error measure W_k is plotted against the number of clusters k, it often turns out that the error measure at some point decreases rapidly and after that flattens remarkable. Tibshirani et al. [8] show that such an „elbow“ is a marker for the best possible number of clusters k. To formalize this, the gap value is introduced which compares $\log(W_k)$ with „its expectation under an appropriate null reference distribution of the data“ [8].

Definition 10 (Gap Value $\text{Gap}_{|\mathcal{X}|}(k)$)**.** Using the above error measure W_k, the *gap value* $\text{Gap}_{|\mathcal{X}|}(k)$ is given by:

$$\text{Gap}_{|\mathcal{X}|}(k) = \mathbb{E}^*_{|\mathcal{X}|} \left[\log\left(W_k\right)\right] - \log\left(W_k\right),$$

where the expectation $\mathbb{E}^*_{|\mathcal{X}|}$ is determined by a sampling of size $|\mathcal{X}|$ from the reference distribution and $\log(W_k)$ is computed using the data set $\mathcal{X}$.

Then, the gap criterion works in the following way: We choose k in such a way that the gap value is maximized. We note that this criterion works for arbitrary clustering algorithms using arbitrary distances.

4 Application to the Pancreas

Exemplarily, we apply the `k-means++` algorithm to cluster the form of a human organ, in particular the neck of the pancreas. As a basis, we use computed tomography (CT) image data of three cross-sections of 100 different pancreases of healthy patients. Hereby, we need mathematical quantities describing the shape of the pancreatic neck, which is recorded by these images. As the shape of an object is independent of its size, we need a mathematical quantity reflecting this. In other words, the required quantity has to be scale-invariant in order to ensure comparable data, meaning that a relative quantity is needed. We propose to compare the area of the circumscribed circle relative to the area of the cross-section of the pancreas. We compute this relative area $A_{\mathrm{rel,i}}, i = 1, 2, 3$ in the following way:

$$\begin{aligned} A_{\mathrm{rel},i} &= \frac{A_{\mathrm{pancreas},i}}{A_{\mathrm{circle},i}}, \\ &= \frac{A_{\mathrm{pancreas},i}}{\pi \cdot (0.5 \cdot d_{\mathrm{pancreas},i})^2}. \end{aligned} \tag{1}$$

This computation is firstly based on the measured diameter $d_{\mathrm{pancreas},i}$, which is the largest distance of two boundary points of the pancreas. Secondly, the area of the cross-section of the pancreas $A_{\mathrm{pancreas},i}$ is computed using tools from image segmentation. To allow comparison, the images of the pancreases are scaled such that $d_{\mathrm{pancreas},i} = 1$ respectively.

Since the number of clusters is not known a priori, we examine the above mentioned evaluation criteria in order to obtain estimates for the most appropriate value of k.

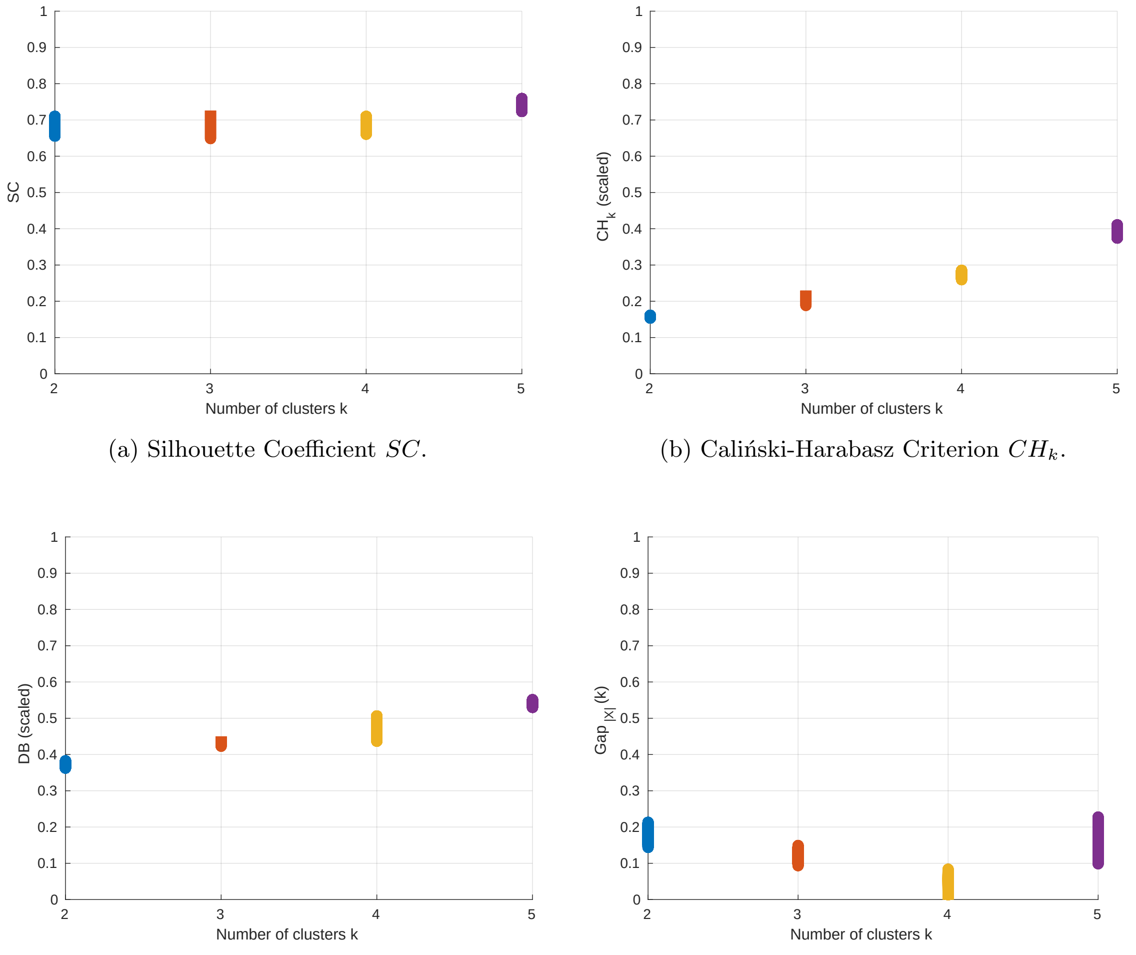
5 Numerical Results

We will present the results of the clustering evaluation methods described in Section 3, which we apply to the data set of 100 data points for each of the three cross-sections of the pancreas. As the underlying clustering method, we choose the k-means++ algorithm illustrated in Section 2. The data is given by the computed relative area proposed in Equation (1).

After the numerical computation of the different evaluation methods, we are doing a post-processing step by scaling the results of the different methods to the unit interval $[0, 1] \in \mathbb{R}$. As in some methods, we are maximizing and in others we are minimizing the evaluation criterion value, the corresponding methods are rewritten in such a way that the most appropriate number of clusters k always is given by the maximum value of the corresponding evaluation criterion. This is all done in such a way that we are able to compare the different evaluation methods directly.

The four mentioned clustering evaluation criteria are implemented in Matlab using the function `evalclusters`. As the example data set consists of only 100 data points, a division into more than 5 clusters seems not to be meaningful. Therefore, we restrict ourselves to the evaluation of clustering results up to 5 clusters. We pointed out in Section 2 that the `k-means++` algorithm depends on the choice of the initial centroids, which is why it is not a deterministic algorithm. Thus, the results for the evaluation of the clusterings may vary as well. In the numerical experiments, we replicated the `k-means++` algorithm 1000 times. For all slices, the silhouette coefficient and the Caliński-Harabasz criterion suggest to use $k = 5$ clusters, which is the maximum number of clusters we considered. However, The Davies-Bouldin criterion suggests to use $k = 2$ clusters and for the gap criterion, there is no consistent proposed number of clusters to be chosen. We illustrate the distribution of the values for the different criteria in Figures 1— 3.

As a second approach, we combine all of the three cross-sections by introducing a three-dimensional data vector $A_{\mathrm{rel}} = (A_{\mathrm{rel},1}, A_{\mathrm{rel},2}, A_{\mathrm{rel},3}) \in \mathbb{R}^3$. Like before, we perform the `k-means++` algorithm 1000 times and evaluate the respective clustering results using the four above mentioned evaluation criteria. The corresponding results are given in Figure 4.



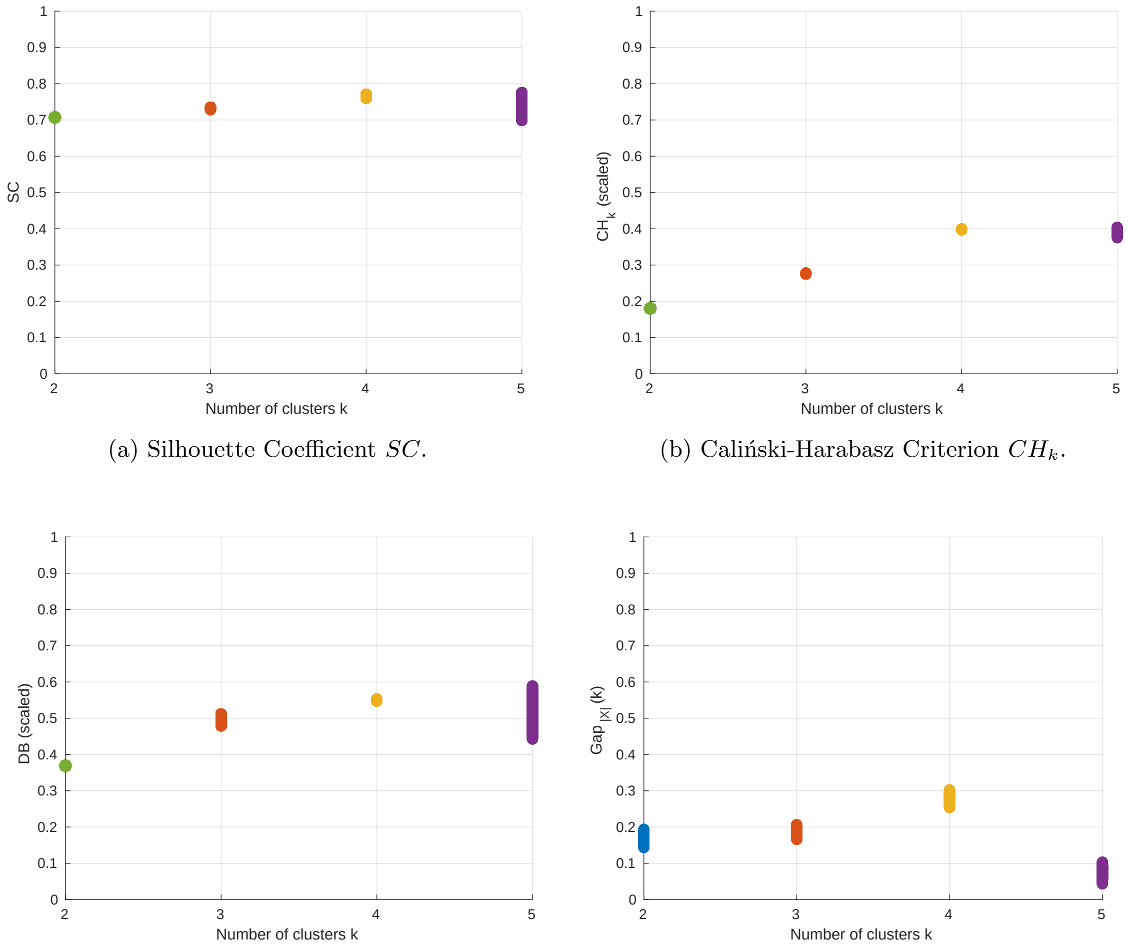
(a) Silhouette Coefficient SC.

(b) Caliński-Harabasz Criterion CH_k.

(c) Davies-Bouldin Criterion DB.

(d) Gap Criterion $\text{Gap}_{|\mathcal{X}|}(k)$.

Figure 1: Results for four clustering evaluation criteria computed for $k = 2, \ldots, 5$ for the exemplary first cross-section of the pancreas image data with 1000 reruns.



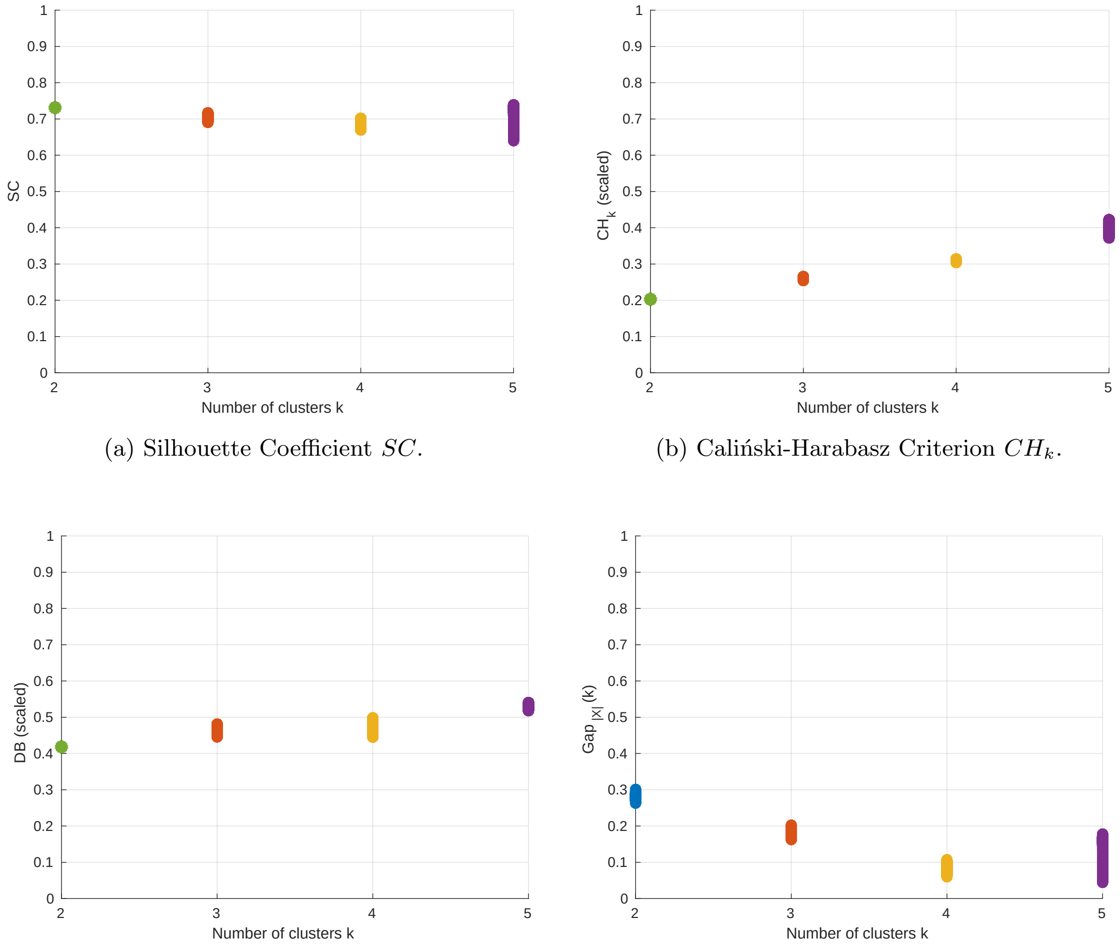
(a) Silhouette Coefficient SC.

(b) Caliński-Harabasz Criterion CH_k.

(c) Davies-Bouldin Criterion DB.

(d) Gap Criterion $\text{Gap}_{|\mathcal{X}|}(k)$.

Figure 2: Results for four clustering evaluation criteria computed for $k = 2, \ldots, 5$ for the exemplary second cross-section of the pancreas image data with 1000 reruns.



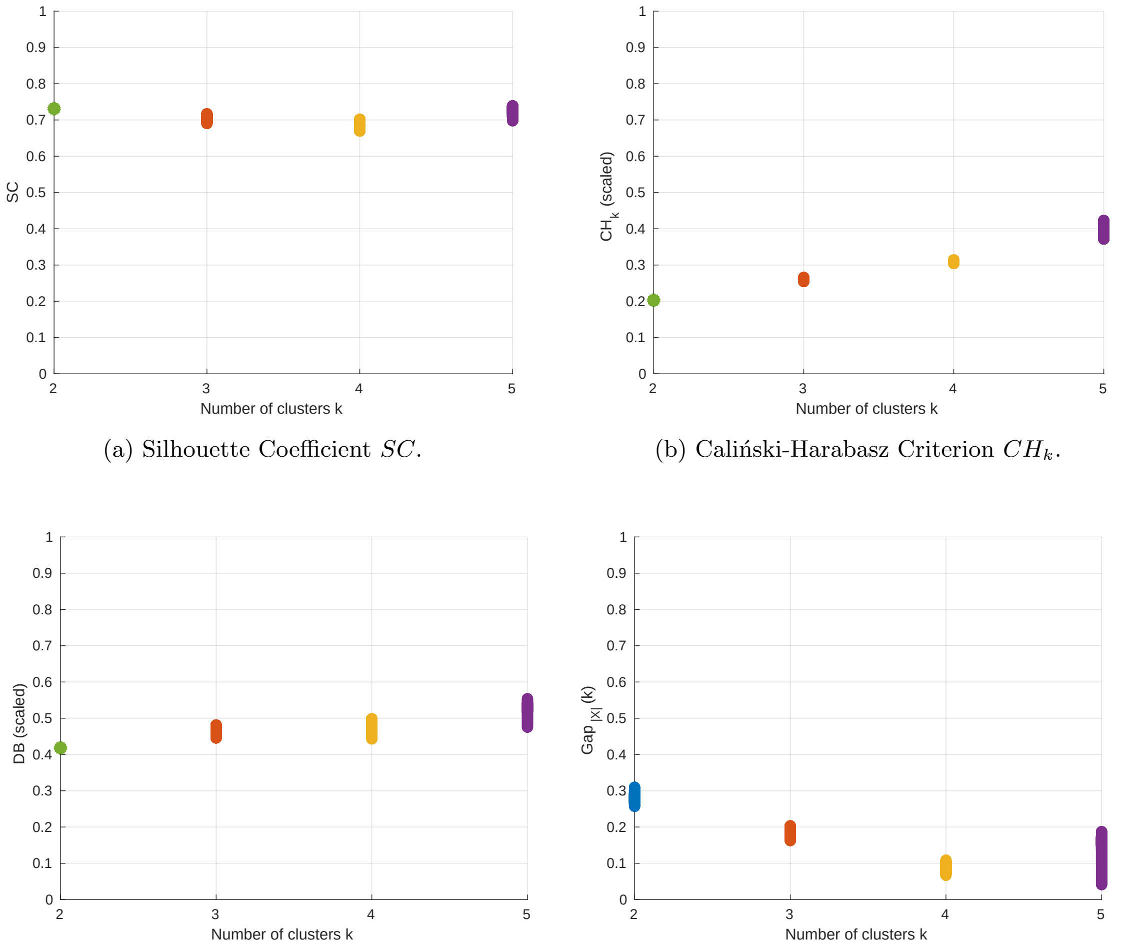
(a) Silhouette Coefficient SC.

(b) Caliński-Harabasz Criterion CH_k.

(c) Davies-Bouldin Criterion DB.

(d) Gap Criterion $\mathrm{Gap}_{|\mathcal{X}|}(k)$.

Figure 3: Results for four clustering evaluation criteria computed for $k = 2, \dots, 5$ for the exemplary third cross-section of the pancreas image data with 1000 reruns.



(a) Silhouette Coefficient SC.

(b) Caliński-Harabasz Criterion CH_k.

(c) Davies-Bouldin Criterion DB.

(d) Gap Criterion $\mathrm{Gap}_{|\mathcal{X}|}(k)$.

Figure 4: Results for four clustering evaluation criteria computed for $k = 2, \ldots, 5$ for the example of the 3D data A_{rel} of the pancreas image data with 1000 reruns.

6 Conclusion

The goal of this work was to explore different clustering evaluation criteria for the classification of the shape of different solid human organs.

We illustrated the underlying `k-means++` algorithm, followed by descriptions of four different evaluation methods consisting of the silhouette coefficients, the Caliński-Harabasz index, the Davies-Bouldin index and the gap criterion. Thereby, no evaluation measure works better than the others since all of them use a different underlying metric for describing the intra-cluster and inter-cluster distances.

The methods were exemplarily applied to a data set which corresponds to the description of three cross-sections of 100 healthy human pancreases. In this case, we introduced a relative quantity, called relative area, to cluster the data. This was an essential step to overcome scaling issues and to care for comparable data. In summary, the clustering using the `k-means++` algorithm was possible. The silhouette coefficient for each clustering shows that the underlying data has a reasonable structure. However, the determination of an appropriate number of clusters is a non trivial task.

To further improve the surgical implications, the use of supervised learning approaches, like classification methods, could be helpful. The main difference to clustering is that we introduce labels which are assigned to each feature. By doing so, the found classes could be interpreted, which is not possible with a clustering approach.

References

[1] D. Arthur and S. Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[2] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[3] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

[4] M. Ester and J. Sander. *Knowledge discovery in databases.* Springer, Berlin ; Heidelberg [u.a.], 2000.

[5] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley series in probability and mathematical statistics. John Wiley, New York, 1990.

[6] S. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inf. Theor.*, 28(2):129–137, September 2006.

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

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

Preprint Series of the Engineering Mathematics and Computing Lab

recent issues

No. 2019-02 Nils Schween, Nico Meyer-Hübner, Philipp Gerstner, Vincent Heuveline: A time step reduction method for Multi-Period Optimal Power Flow problems

No. 2019-01 Philipp Gerstner, Martin Baumann, Vincent Heuveline: Analysis of the Stationary Thermal-Electro Hydrodynamic Boussinesq Equations

No. 2018-02 Simon Gawlok, Vincent Heuveline: Nested Schur-Complement Solver for a Low-Mach Number Model: Application to a Cyclone-Cyclone Interaction

No. 2018-01 David John, Michael Schick, Vincent Heuveline: Learning model discrepancy of an electric motor with Bayesian inference

No. 2017-07 Martin Baumann, Fabian Gebhart, Oliver Mattes, Sotirios Nikas, Vincent Heuveline: Development and implementation of a temperature monitoring system for HPC systems

No. 2017-06 Simon Gawlok, Philipp Gerstner, Saskia Haupt, Vincent Heuveline, Jonas Kratzke, Philipp Lösel, Katrin Mang, Mareike Schmidtobreick, Nicolai Schoch, Nils Schween, Jonathan Schwegler, Chen Song, Martin Wlotzka: HiFlow3 – Technical Report on Release 2.0

No. 2017-05 Nicolai Schoch, Vincent Heuveline: Towards an Intelligent Framework for Personalized Simulation-enhanced Surgery Assistance: Linking a Simulation Ontology to a Reinforcement Learning Algorithm for Calibration of Numerical Simulations

No. 2017-04 Martin Wlotzka, Thierry Morel, Andrea Piacentini, Vincent Heuveline: New features for advanced dynamic parallel communication routines in OpenPALM: Algorithms and documentation

No. 2017-03 Martin Wlotzka, Vincent Heuveline: An energy-efficient parallel multigrid method for multi-core CPU platforms and HPC clusters

No. 2017-02 Thomas Loderer, Vincent Heuveline: New sparsing approach for real-time simulations of stiff models on electronic control units

No. 2017-01 Chen Song, Markus Stoll, Kristina Giske, Rolf Bendl, Vincent Heuveline: Sparse Grids for quantifying motion uncertainties in biomechanical models of radiotherapy patients

No. 2016-02 Jonas Kratzke, Vincent Heuveline: An analytically solvable benchmark problem for fluid-structure interaction with uncertain parameters

No. 2016-01 Philipp Gerstner, Michael Schick, Vincent Heuveline, Nico Meyer-Hübner, Michael Suriyah, Thomas Leibfried, Viktor Slednev, Wolf Fichtner, Valentin Bertsch: A Domain Decomposition Approach for Solving Dynamic Optimal Power Flow Problems in Parallel with Application to the German Transmission Grid

No. 2015-04 Philipp Gerstner, Vincent Heuveline, Michael Schick : A Multilevel Domain Decomposition approach for solving time constrained Optimal Power Flow problems

No. 2015-03 Martin Wlotzka, Vincent Heuveline: Block-asynchronous and Jacobi smoothers for a multigrid solver on GPU-accelerated HPC clusters

The responsibility for the contents of the working papers rests with the authors, not the Institute. Since working papers are of a preliminary nature, it may be useful to contact the authors of a particular working paper about results or caveats before referring to, or quoting, a paper. Any comments on working papers should be sent directly to the authors.

$\frac{\partial u}{\partial t} = \nu\Delta u + (u \cdot \nabla) u + \frac{1}{\rho}\nabla p$
$\frac{\partial u}{\partial t}$