IRTG 1792 Discussion Paper 2018-018

Adaptive Nonparametric Clustering

Kirill Efimov *
Larisa Adamyan *
Vladimir Spokoiny *[2]

* Humboldt-Universität zu Berlin, Germany
*[2] Weierstrass Institute Berlin, Germany

This research was supported by the Deutsche Forschungsgemeinschaft through the International Research Training Group 1792 "High Dimensional Nonstationary Time Series".

http://irtg1792.hu-berlin.de
ISSN 2568-5619

International Research Training Group 1792

Adaptive Nonparametric Clustering

Kirill Efimov
Humboldt University Berlin,
IRTG 1792, Spandauer Str. 1,
10178 Berlin, Germany
efimovkq@hu-berlin.de

Larisa Adamyan
Humboldt University Berlin,
IRTG 1792, Spandauer Str. 1,
10178 Berlin, Germany
adamyanl@hu-berlin.de

Vladimir Spokoiny*†
Weierstrass Institute Berlin, Mohrenstr. 39, 10117 Berlin, Germany
IITP RAS, HSE, Skoltech Moscow
spokoiny@wias-berlin.de

Abstract

This paper presents a new approach to non-parametric cluster analysis called Adaptive Weights Clustering (AWC). The idea is to identify the clustering structure by checking at different points and for different scales on departure from local homogeneity. The proposed procedure describes the clustering structure in terms of weights w_{ij} each of them measures the degree of local inhomogeneity for two neighbor local clusters using statistical tests of "no gap" between them. The procedure starts from very local scale, then the parameter of locality grows by some factor at each step. The method is fully adaptive and does not require to specify the number of clusters or their structure. The clustering results are not sensitive to noise and outliers, the procedure is able to recover different clusters with sharp edges or manifold structure. The method is scalable and computationally feasible. An intensive numerical study shows a state-of-the-art performance of the method in various artificial examples and applications to text data. Our theoretical study states optimal sensitivity of AWC to local inhomogeneity.

*Financial support from the Deutsche Forschungsgemeinschaft via the IRTG 1792 "High Dimensional Non Stationary Time Series", Humboldt-Universität zu Berlin, is gratefully acknowledged.

†The research is supported by the Russian Science Foundation (project no. 14 50 00150).

AMS 2000 Subject Classification: Primary 62H30. Secondary 62G10
Keywords: adaptive weights, clustering, gap coefficient, manifold clustering

1 Introduction

Methods for cluster analysis are well established tools in various scientific fields. Applications of clustering include a wide range of problems with text, multimedia, networks, and biological data. We refer to the book Aggarwal and Reddy (2013) for comprehensive overview of existing methods. Here we briefly overview only basic approaches in clustering, their advantages and problems. First we mention the so called *partitional* clustering. These algorithms try to group points by optimizing some specific objective function, thereby using some assumptions on the data structure. The most known representatives of this group of methods are k-means Steinhaus (1956) and its variations. k-means finds a local minimum in the problem of sum of squared errors minimization. The partitional algorithms generally require some parameters for initialization (number of clusters K) and also are nondeterministic by their nature. Also k-means usually produces spherical clusters and often fails to identify clusters with a complex shape. *Hierarchical* methods construct a tree called dendrogram. Each level of this tree represents some partition of data with root corresponding to only one cluster containing all points. The base of the hierarchy consists of all singletons (clusters with only one point) which are the leaves of the tree. Hierarchical methods can be split into *agglomerative* and *divisive* clustering methods by the direction they construct a dendrogram. The main weakness of hierarchical algorithms is irreversibility of the merge or split decisions. *Density-based* clustering was proposed to deal with arbitrary shape clusters, detect and remove noise. It can be considered as a non-parametric method as it makes no assumptions about the number of clusters, their distribution or shapes. DBSCAN Ester et al. (1996) is one of the most common clustering algorithms. DBSCAN estimates density by counting the number of points in some fixed neighborhood and retrieves clusters by grouping dense points. If data contains clusters with a difference in density then it is hard or even impossible to set an appropriate density level. Another crucial point of this approach is that the quality of nonparametric density estimation is very poor if the data dimension is large. *Spectral* clustering methods Ng et al. (2002) explore the spectrum of a similarity matrix, i.e. a square matrix with elements equal to pairwise similarities of data points. Spectral clustering can discover nonconvex clusters. However, the number of clusters should be fixed in advance in a proper way, and the method requires a significant spectral gap between clusters. *Affinity Propagation* (AP) Frey and Dueck (2007) is an exemplar-based clustering algorithm. In the iterative process AP updates two matrices based on simi-

larities between pairs of data points. AP does not require the number of clusters to be determined, whereas it has limitation: the tuning of its parameters is difficult due to occurrence of oscillations. *Graph-based methods* represent each data point as a node of a graph whose edges reflect the proximity between points. For a detailed survey on graph methods we refer to Zhu (2005).

After all, the following main problems arise in clustering algorithms: unknown number of clusters, nonconvex clusters, unbalance in sizes or/and densities for different clusters, stability with changing parameters. For big data the complexity is also very important.

A theoretical study of the clustering problem is difficult due to lack of a clear and unified definition of a cluster. Probably the most popular way of defining a cluster is based on connected level sets of the underlined density Wishart (1969); Hartigan (1975) or regions of relatively high density Seeger (2001). Existing theoretical bounds require to consider regular or r_0-standard clusters which allows to exclude pathological cases Ester et al. (1996); Rigollet (2007). However, this approach cannot distinguish overlapping or connected clusters with different topological properties. This paper does not aim at giving a unified definition of a global cluster. Instead we offer a new approach which focuses on local cluster properties: a cluster is considered as a homogeneous set of similar points. Any significant departure from local homogeneity is called a "gap" and a cluster can be viewed as a collection of points without a gap. An obvious advantage of this approach is that it can be implemented as a family of tests of "no gap" between any neighbor local clusters. The idea is similar to multiscale high dimensional change point analysis where the test of a change point is replaced by a test of a gap; cf. Frick et al. (2014). The method presented in the next section involves the ideas of agglomerative hierarchical (by changing the scale of objects from small to big), density based (by nonparametric test) and affinity propagation (by iteratively updating the weights). The "adaptive weights" idea originates from *propagation-separation approach* introduced in Polzehl and Spokoiny (2006) for regression types models. In the clustering context, this idea can be explained as follows: for every point X_i, the clustering procedure attempts to describe its largest possible local neighborhood in which the data is homogeneous in a sense of spatial data separation. Technically, for each data point X_i, a local cluster $\mathcal{C}_i$ is described in terms of binary weights w_{ij} and includes only points X_j with $w_{ij} = 1$. Thus, the whole clustering structure of the data can be described using the matrix of weights W which is recovered from the data. Usual clustering in the form of a mapping $\boldsymbol{X} \mapsto \mathcal{C}$ can be done using the resulting weights w_{ij}. The weights are computed by the sequential multiscale procedure. The main advantage of the proposed procedure is that it is fully adaptive to the unknown number of clusters, structure of the clusters etc. It applies equally well to convex and shaped clusters of different type and density. We also show that

the procedure does not produce artificial clusters (propagation effect) and ensures nearly optimal separation of closely located clusters with a hole of a lower density between them. The procedure involves only one important tuning parameter λ, for which we suggest an automatic choice based on the propagation condition or on the "sum of weights heuristic". Numerical results indicate state-of-the-art performance of the new method on artificial and real life data sets. The main contributions of this paper are:

1. We propose a new approach to define a clustering structure: a cluster is built by a group of samples with "no gap" inside. This allows to effectively recover the number of clusters and the shape of each cluster without any preliminary information.

2. The method can deal with non-convex and overlapping clusters, it adapts automatically to manifold clustering structure and it is robust against outliers.

3. The proposed procedure demonstrates state-of-the-art performance on wide range of various artificial and real life examples and outperforms the popular competitive procedures even after optimising their tuning parameters. It is computationally feasible and the method applies even to large datasets.

4. The procedure controls the probability of building an artificial cluster in a homogeneous situation.

5. Theoretical results claim an optimal sensitivity of the method in detecting of two or more clusters separated by a hole of a lower density due to multiscale nature of the procedure.

The rest of the paper is organized as follows. Section 2 introduces the procedure starting from some heuristics. Its theoretical properties are discussed in Section 3. The numerical study is presented in Section 4. The proofs are collected in Section 5. Some technical details as well as more numerical examples are postponed to Appendix.

2 Nonparametric Clustering using Adaptive Weights

Let $\{X_1, \ldots, X_n\} \subset I\!R^p$ be an i.i.d. sample from the density $f(x)$. Here the dimension p can be very large or even infinite. We assume for any pair (X_i, X_j) that a known distance (or non-similarity measure) $d(X_i, X_j)$ between X_i and X_j is given, for instance, the Euclidean norm $\|X_i - X_j\|$. This is also the default distance in this paper. The proposed procedure operates with the distance matrix $\bigl(d(X_i, X_j)\bigr)_{i,j=1}^n$ only. For describing the clustering structure of the data, we introduce a $n \times n$ matrix of weights $W = (w_{ij})$, $i, j = 1, \ldots, n$. Usually the weights w_{ij} are binary and $w_{ij} = 1$ means that X_i and X_j

are in the same cluster, while $w_{ij} = 0$ indicates that these points are in different clusters. The matrix W is symmetric and each block of ones describes one cluster. However, we do not require a block structure which allows to incorporate even overlapping clusters. For every fixed i, the associated cluster $\mathcal{C}_i$ is given by the collection of positive weights (w_{ij}) over j. One can consider a more general construction when $w_{ij} \in [0,1]$ and this value can be viewed as probability that the other point X_j is in the same cluster as X_i.

The proposed procedure attempts to recover the weights w_{ij} from data, which explains the name "adaptive weights clustering". The procedure is sequential. It starts with very local clustering structure $\mathcal{C}_i^{(0)}$, that is, the starting positive weights $w_{ij}^{(0)}$ are limited to the closest neighbors X_j of the point X_i in terms of the distance $d(X_i, X_j)$. At each step (or scale) $k \geq 1$, the weights $w_{ij}^{(k)}$ are recomputed by means of statistical tests of "no gap" between $\mathcal{C}_i^{(k-1)}$ and $\mathcal{C}_j^{(k-1)}$; see the next section. Only the neighbor pairs X_i, X_j with $d(X_i, X_j) \leq h_k$ are checked, however the locality (or scale) parameter h_k and the number of scanned neighbors X_j for each fixed point X_i grows in each step. The resulting matrix of weights W is used for the final clustering. The core element of the method is the way how the weights $w_{ij}^{(k)}$ are recomputed.

2.1 Adaptive weights w_{ij}: test of "no gap"

Suppose that the first $k-1$ steps of the iterative procedure have been carried out. This results in collection of weights $\{w_{ij}^{(k-1)}, j = 1, \ldots, n\}$ for each point X_i. These weights describe a local "cluster" associated with X_i. By construction, only those weights $w_{ij}^{(k-1)}$ can be positive for which X_j belongs to the ball $\mathcal{B}(X_i, h_{k-1}) \stackrel{\text{def}}{=} \{x \colon d(X_i, x) \leq h_{k-1}\}$, or, equivalently, $d(X_i, X_j) \leq h_{k-1}$. At the next step k we pick up a larger radius h_k and recompute the weights $w_{ij}^{(k)}$ using the previous results. Again, only points with $d(X_i, X_j) \leq h_k$ have to be screened at step k. The basic idea behind the definition of $w_{ij}^{(k)}$ is to check for each pair i, j with $d(X_i, X_j) \leq h_k$ whether the related clusters are well separated or they can be aggregated into one homogeneous region. We treat the points X_i and X_j as fixed and compute the test statistic $T_{ij}^{(k)}$ using the weights $w_{i\ell}^{(k-1)}$ and $w_{j\ell}^{(k-1)}$ from the preceding step. The test compares the data density in the union and overlap of two clusters for points X_i and X_j. The formal definition involves the weighted empirical mass of the overlap and the weighted empirical mass of the union of two balls $\mathcal{B}(X_i, h_{k-1})$ and $\mathcal{B}(X_j, h_{k-1})$ shown on Figure 2.1. *The empirical mass of the overlap* $N_{i \wedge j}^{(k)}$ can be naturally defined as

$$N_{i \wedge j}^{(k)} \stackrel{\text{def}}{=} \sum_{\ell \neq i,j} w_{i\ell}^{(k-1)} w_{j\ell}^{(k-1)}.$$

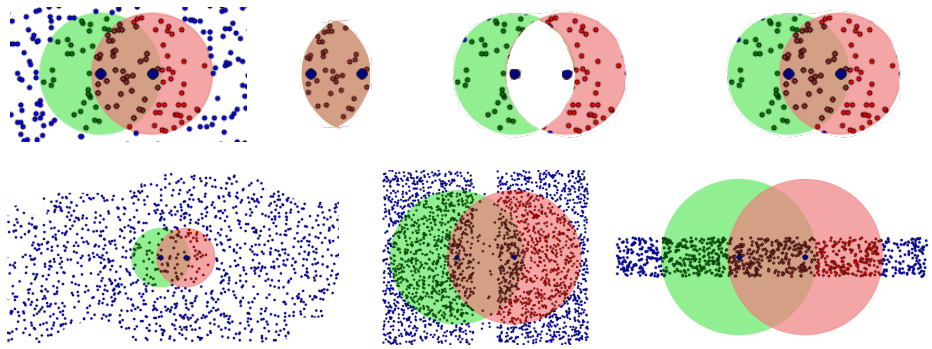

Figure 2.1: Test of "no gap between local clusters". *Top, from left*: Homogeneous case; $N^{(k)}_{i\wedge j}; N^{(k)}_{i\triangle j}; N^{(k)}_{i\vee j}$; *Bottom: from left*: Homogeneous case. "Gap" case. Manifold clustering.

In the considered case of indicator weights $w^{(k-1)}_{ij}$, this value is indeed equal to the number of points in the overlap of $\mathcal{B}(X_i, h_{k-1})$ and $\mathcal{B}(X_j, h_{k-1})$ except X_i, X_j. Similarly, the *mass of the complement* is defined as

$$N^{(k)}_{i\triangle j} \stackrel{\text{def}}{=} \sum_{\ell \neq i,j} \Bigl\{ w^{(k-1)}_{i\ell}\, \mathbb{1}\bigl(X_\ell \notin \mathcal{B}(X_j, h_{k-1})\bigr) + w^{(k-1)}_{j\ell}\, \mathbb{1}\bigl(X_\ell \notin \mathcal{B}(X_i, h_{k-1})\bigr) \Bigr\}.$$

Note that $N^{(k)}_{i\triangle j}$ is the number of points in $\mathcal{C}^{(k-1)}_i$ and $\mathcal{C}^{(k-1)}_j$ which do not belong to the overlap $\mathcal{B}(X_i, h_{k-1}) \cap \mathcal{B}(X_j, h_{k-1})$. Finally, *mass of the union* $N^{(k)}_{i\vee j}$ can be defined as the sum of the mass of the overlap and the mass of the complement:

$$N^{(k)}_{i\vee j} \stackrel{\text{def}}{=} N^{(k)}_{i\wedge j} + N^{(k)}_{i\triangle j}\,.$$

To measure the gap we consider the ratio of these two masses:

$$\widetilde{\theta}^{(k)}_{ij} = N^{(k)}_{i\wedge j} \,/\, N^{(k)}_{i\vee j}\,. \tag{2.1}$$

This value can be viewed as an estimate of the value $\theta^{(k)}_{ij}$ which measures the ratio of the population mass of the overlap of two local regions $\mathcal{C}^{(k-1)}_i$ and $\mathcal{C}^{(k-1)}_j$ relative to the mass in their union:

$$\theta^{(k)}_{ij} \stackrel{\text{def}}{=} \frac{\int_{\mathcal{B}(X_i,h_k)\cap\mathcal{B}(X_j,h_k)} f(u)du}{\int_{\mathcal{B}(X_i,h_k)\cup\mathcal{B}(X_j,h_k)} f(u)du}\,. \tag{2.2}$$

Under local homogeneity one can suppose that the density in the union of two balls is nearly constant. In this case, the value $\theta^{(k)}_{ij}$ should be close to the ratio $q^{(k)}_{ij}$ of the volume of overlap and the volume of union of these balls:

$$q^{(k)}_{ij} \stackrel{\text{def}}{=} \frac{\int_{\mathcal{B}(X_i,h_k)\cap\mathcal{B}(X_j,h_k)} du}{\int_{\mathcal{B}(X_i,h_k)\cup\mathcal{B}(X_j,h_k)} du} = \frac{\mathrm{Vol}_\cap(d_{ij}, h_{k-1})}{2\,\mathrm{Vol}(h_{k-1}) - \mathrm{Vol}_\cap(d_{ij}, h_{k-1})}, \tag{2.3}$$

where $\mathrm{Vol}(h)$ is the volume of a ball with radius h and $\mathrm{Vol}_{\cap}(d,h)$ is the volume of the intersection of two balls with radius h and distance $d_{ij} = d(X_i, X_j)$ between centers. The ratio $\theta_{ij}^{(k)}/q_{ij}^{(k)}$ will be called the *gap coefficient*. If the gap coefficient is significantly smaller than one, this can be treated as a gap between two local clusters at scale k.

The new weight $w_{ij}^{(k)}$ can be viewed as a randomized test of the null hypothesis $H_{ij}^{(k)}$ of no gap between X_i and X_j against the alternative of a significant gap at scale k. This is a composite hypothesis which reads as $\theta_{ij}^{(k)}/q_{ij}^{(k)} \geq 1$ against $\theta_{ij}^{(k)}/q_{ij}^{(k)} < 1$. The construction is illustrated by Figure 2.1 for the homogeneous situation and for the situation with a gap.

To quantify the notion of significance, we consider the statistical likelihood ratio test of "no gap" between two local clusters. The corresponding test statistic can be motivated by the following statistical problem. Let $X_1, \ldots, X_n \in I\!R^p$ be an i.i.d. sample and B, C be two non-overlapping measurable sets in $I\!R^p$. Suppose we are interested to check the relation $I\!P(B) \geq q\bigl\{I\!P(B) + I\!P(C)\bigr\}$ for a given value $q \in (0,1)$ against the one-sided alternative $I\!P(B) < q\bigl\{I\!P(B) + I\!P(C)\bigr\}$. Let

$$S_B \stackrel{\mathrm{def}}{=} \sum_{i=1}^n 1\!\!1(X_i \in B), \quad S_C \stackrel{\mathrm{def}}{=} \sum_{i=1}^n 1\!\!1(X_i \in C).$$

Lemma A.1 from Appendix A shows that corresponding likelihood ratio test statistics can be written as

$$T = (S_B + S_C)\, \mathcal{K}\bigl(\widetilde{\theta}, q\bigr)\, \bigl\{1\!\!1(\widetilde{\theta} \leq q) - 1\!\!1(\widetilde{\theta} > q)\bigr\},$$

where $\widetilde{\theta} \stackrel{\mathrm{def}}{=} S_B/(S_B + S_C)$, and $\mathcal{K}(\theta, \eta)$ is the Kullback-Leibler (KL) divergence between two Bernoulli laws with parameters θ and η:

$$\mathcal{K}(\theta, \eta) \stackrel{\mathrm{def}}{=} \theta \log \frac{\theta}{\eta} + (1-\theta) \log \frac{1-\theta}{1-\eta}\ .$$

It is worth noting that the test statistic T only depends on the local sums S_B and S_C. One can also use the symmetrized version of the KL divergence:

$$\mathcal{K}_s(\theta, \eta) \stackrel{\mathrm{def}}{=} \frac{1}{2}\bigl\{\mathcal{K}(\theta, \eta) + \mathcal{K}(\eta, \theta)\bigr\} = (\theta - \eta) \log \frac{\theta(1-\eta)}{(1-\theta)\eta}\,.$$

Now we apply this construction to the situation with two local clusters. The set B is the overlap of the balls $\mathcal{B}(X_i, h_k)$ and $\mathcal{B}(X_j, h_k)$, while C stands for its complement within the union $\mathcal{B}(X_i, h_{k-1}) \cup \mathcal{B}(X_j, h_{k-1})$. Then the weighted analog of the mass of the overlap S_B is given by $N_{i \wedge j}^{(k)}$, while $S_B + S_C$ is extended to the mass of the union $N_{i \vee j}^{(k)}$ yielding the test statistic $T_{ij}^{(k)}$ of the form

$$T_{ij}^{(k)} = N_{i \vee j}^{(k)}\, \mathcal{K}\bigl(\widetilde{\theta}_{ij}^{(k)}, q_{ij}^{(k)}\bigr)\, \bigl\{1\!\!1(\widetilde{\theta}_{ij}^{(k)} \leq q_{ij}^{(k)}) - 1\!\!1(\widetilde{\theta}_{ij}^{(k)} > q_{ij}^{(k)})\bigr\}. \tag{2.4}$$

Polzehl and Spokoiny (2006) used a similar test of the hypothesis $\theta_{ij}^{(k)} = q_{ij}^{(k)}$ for defining a homogeneous region within an image. In the contrary to that paper, we consider a one sided test with a composite null $\theta_{ij}^{(k)} \geq q_{ij}^{(k)}$. The value $q_{ij}^{(k)}$ from (2.3) depends only on the ratio $t_{ij}^{(k)} = d_{ij}/h_{k-1}$ with $d_{ij} = d(X_i, X_j)$. If $d(X_i, X_j) = \|X_i - X_j\|$, then $q_{ij}^{(k)}$ can be calculated explicitly (Li (2011)): $q_{ij}^{(k)} = q(t_{ij}^{(k)})$ with

$$q(t) = \left(2 \frac{B\left(\frac{p+1}{2}, \frac{1}{2}\right)}{B\left(1 - \frac{t^2}{4}, \frac{p+1}{2}, \frac{1}{2}\right)} - 1 \right)^{-1}, \tag{2.5}$$

where $B(a, b)$ is the beta-function, $B(x, a, b)$ is the incomplete beta-function, and p is the space dimension. The argument $t \in [0, 1)$ and the function can be tabulated.

The famous Wilks phenomenon Wilks (1938) claims that the distribution of each test statistic $T_{ij}^{(k)}$ is nearly χ^2-distributed under the null hypothesis. This justifies the use of family of such tests $T_{ij}^{(k)}$ properly scaled by a universal constant λ which is the only tuning parameter of the method. See below at the end of this section.

2.2 The procedure

This section presents a formal description of the procedure. First we list the main ingredients of the method, then present the algorithm.

A sequence of radii: First of all we need to fix a growing sequence of radii $h_1 \leq h_2 \leq \ldots \leq h_K$ which determines how fast the algorithm will come from considering very local structures to large-scale objects. Each value h_k can be viewed as a resolution (scale) of the method at step k. The rule has to ensure that the average number of screened neighbors for each X_i at step k grows at most exponentially with $k \geq 1$. This feature will be used to show the optimal sensitivity of the method. A specific choice of this sequence is given in Appendix B. Here we just assume that such a sequence is fixed under the following two conditions:

$$n(X_i, h_{k+1}) \leq a\, n(X_i, h_k), \qquad h_{k+1} \leq b\, h_k$$

where $n(X_i, h)$ is the number of neighbors of X_i in the ball of radius h, and a, b are given constants between 1 and 2. Our default choice is $a = \sqrt{2}$, $b = 1.95$ which ensures a non-trivial overlap of any two local clusters at the step k. Decreasing any of the parameters a, b increases the number of steps h_k and thus, the computational time but can improve the separation property of the procedure. Our intensive numerical studies showed that the default choice works well in all examples; no notable improvements can

be achieved by tuning of these parameters. A geometric growth of the values $n(X_i, h_{k+1})$ ensures that the total number of steps K is logarithmic in the sample size n.

Initialization of weights: Define by $h_0(X_i)$ the smallest radius h_k in our fixed sequence such that the number of neighbors of point X_i in the ball $\mathcal{B}(X_i, h_0(X_i))$ is not smaller than n_0, our default choice $n_0 = 2p + 2$. Using these distances we can initialize $w_{ij}^{(0)}$ as

$$w_{ij}^{(0)} = 1\!\!1\big(d(X_i, X_j) \le \max(h_0(X_i), h_0(X_j))\big).$$

Updates at step k: At step k for $k = 1, 2, \ldots, K$, we update the weights $w_{ij}^{(k)}$ for all pairs of points X_i and X_j with distance $d(X_i, X_j) \le h_k$. The last constraint allows us to recompute only $n \times n_k$ weights, where n_k is the average number of neighbors in the h_k neighborhood. The weights $w_{ij}^{(k)}$ at step k are computed in the form

$$w_{ij}^{(k)} = 1\!\!1\big(d(X_i, X_j) \le h_k\big)\ 1\!\!1\big(T_{ij}^{(k)} \le \lambda\big) \tag{2.6}$$

for all points X_i, X_j with $h_{k-1} \ge h_0(X_i)$ and $h_{k-1} \ge h_0(X_j)$. The last constraint guarantees that the weights $w_{ij}^{(k)}$ are computed by the algorithm only when the corresponding balls contain at least n_0 points. The value $T_{ij}^{(k)}$ is the test statistic from (2.4) for the "no gap" test for points X_i and X_j. Here λ is a threshold coefficient and the only parameter for tuning.

The output of the AWC is given by the matrix $W = W^{(K)}$ at the final step K which defines the local cluster $\mathcal{C}(X_i) = (X_j \colon w_{ij} > 0)$ for each point X_i. One can use these local structures to produce a partition of the data into non-overlapping blocks.

Tuning the parameter λ: The parameter λ has an important influence on the performance of the method. Large λ-values result in a conservative test of "no gap" which can lead to aggregation of inhomogeneous regions. In the contrary, small λ increases the sensitivity of the method to inhomoheneity, but may lead to artificial segmentation. This section discusses two possible approaches to fix the parameter λ. The first is not data-driven and depends only on data dimension p. The second approach is based on the "sum-of-weights" heuristics and is completely data driven.

The *propagation* approach which originates from Spokoiny and Vial (2009) suggests to tune the parameter λ as the smallest value which ensures a prescribed level (e.g. 90%) of correct clustering result in a very special case of just one simple cluster. Namely, we tune the parameter λ to ensure that the algorithm typically puts all points into one cluster for the sample uniformly distributed on a unit ball. This is similar to the level condition in hypothesis testing when the procedure is tested under the null hypothesis

of a simple homogeneous cluster. The first kind error corresponds to creating some artificial clusters, where the probability of such events is controlled by the choice of λ. The construction guarantees right performance of the method ($w_{ij}^{(k)} \approx 1$) only in the very special case of locally constant density. However, this situation is clearly reproduced for any local neighborhood lying within a large homogeneous region. Therefore, the propagation condition yields a right performance of the procedure within each homogeneous region.

Another way of looking at the choice of λ called *"sum-of-weights heuristic"* is based on the effective cluster volume given by the total sum of final weights $w_{ij}^{(K)}$ over all i, j. Let $w_{ij}^{(K)}(\lambda)$ be the final weights obtained by the procedure with the parameter λ. Define

$$S(\lambda) \stackrel{\text{def}}{=} \sum_{i,j=1}^{n} w_{ij}^{(K)}(\lambda).$$

Small λ-values lead to artificial clustering with many small blocks of ones and all zeros outside of these blocks. The corresponding $S(\lambda)$ will be small as well. An increase of λ yields larger homogeneous blocks and thus, a larger value $S(\lambda)$. Such behavior is typically observed until λ reaches a reasonable value, then the cluster structure stabilizes and any further moderate increase of λ does not affect $S(\lambda)$. For big λ, the procedure starts to aggregate two or more clusters into one, this leads to a jump in $S(\lambda)$. So, a proposal is to pick up the smallest λ-value corresponding to a plateau in the graph of $S(\lambda)$. In the case of complex cluster structure, one can observe several plateaux, with the corresponding λ-value for each plateau. Then we recommend to check all those λ-values and compare the obtained clustering results afterwards. See Appendix D for some numerical examples.

2.3 Algorithm complexity

The preliminary step of our algorithm requires to fix the sequence of radii $\{h_k\}_{k=0}^{K}$, build the distance matrix and initialize the matrix of weights. The last is updated on each step of the algorithm. Suppose the average number of neighbors for each X_i at step k is n_k. Then finding the first n_K neighbors for each point costs $O(n\, n_K \log n)$. At step k we need to compute $0.5 n\, n_k$ statistics $T_{ij}^{(k)}$. Calculation of all values $N_{i \wedge j}^{(k)}, N_{i \triangle j}^{(k)}$ costs $O(n\, n_k^2)$. As a result the overall complexity of step k is $O(n n_k^2)$. Note that the local nature of the procedure allows to effectively use parallel computations. In our approach the radii h_k are fixed in a way that ensures exponential growth of n_k. It results in $K = O(\log n)$ steps and furthermore, complexity of all steps is determined by the last step: $O(n n_K^2)$. In our experiments the datasets sample sizes are not very large

$n \leq 10000$, therefore we used $n_K = n$. For large datasets, one should use $n_K \ll n$. In this case at the last step we will only catch the local clustering structure for each point and then "recover" the global structure by extracting the connected components.

3 AWC properties

This section discusses some important properties of the AWC method.

3.1 Propagation for regions with a non-constant density

The procedure is calibrated to ensure the propagation within regions with a constant density. It is important to understand how far this property can be extended for a non-constant density. Symmetricity arguments allow to easily extend propagation effect to the case of a linear density. In the univariate case, one can make a further step and show this property for regions with a concave density.

Theorem 3.1. *Let the observations X_i be i.i.d. in $I\!R^p$, let the data density $f(x)$ be supported on a region V. Consider two cases: 1) $p = 1$ and the density $f(x)$ is concave; 2) p is arbitrary and $f(x)$ is linear. If $\lambda > \mathtt{C} \log n$ for some absolute constant $\mathtt{C}$, then with a probability at least $1 - 2/n$, it holds $w_{ij}^{(k)} = 1$ at any step k of the procedure.*

Numerical examples illustrating this and the further results are presented in Section 4.1. The proofs are collected in Section 5.

3.2 Separation with a hole

Now we discuss the "separation" effect between clusters for one particularly important situation, when two homogeneous regions are separated by a hole with slightly smaller density and we compute the weight $w_{ij}^{(k)}$ by (2.6) for two points from different regions each close to the hole; see Figure 2.1 "Gap" case. Let V be a set with a volume $|V|$ and G be a splitting hole with volume $|G|$ such that $V_G \stackrel{\text{def}}{=} V \setminus G$ consists of two disjoint regions. To be more specific, consider two uniform clusters separated by some area (hole) of a lower density. Let f_G denote the density on G and f_{V_G} on the complement $V \setminus G$. We assume the relation $f_G = (1-\varepsilon) f_{V_G}$ for a small value ε. The separation effect would mean that for any pair of points X_i and X_j from different clusters, the statistical test detects this situation leading to a big value of the test statistic $T_{ij}^{(k)}$ and to a vanishing weight $w_{ij}^{(k)}$. The next two theorems answer the following question: what is the smallest depth parameter ε of the hole which enables a consistent and precise separation? First we establish a lower bound.

Theorem 3.2. *Let the data support V contain a fixed hole G, and the data density $f(\cdot)$ be equal to f_1 on the complement $V \setminus G$ and to f_G on G with $f_G = (1-\varepsilon)f_1$. Let $\varepsilon = \varepsilon_n$ as the sample size $n \to \infty$. If $n\varepsilon_n^2 \le \mathtt{C}$ for a fixed constant $\mathtt{C} > 0$, then it is impossible to consistently separate the cases with $\varepsilon = 0$ (no gap) and $\varepsilon = \varepsilon_n$.*

For an upper bound, we need a more specific description of the shape of the region V on which the data is supported. Namely we assume that V is composed of two regions V_1 and V_2 of higher density f_1 separated by a hole G with a slightly smaller density f_G and the volume and shape of all three subregions V_1, V_2, G are nearly the same. The next result heavily uses the multiscale nature of the procedure. Namely we focus on the steps when the bandwidth h_k approaches the global bandwidth h_K. For two points X_i and X_j from different regions this allows to assume that the union of two balls $\mathcal{B}(X_i, h_k)$ and $\mathcal{B}(X_j, h_k)$ contains the whole domain V, while their overlap contains G. We show that for such configuration the computed weights $w_{ij}^{(k)}$ typically vanish provided that $n\varepsilon^2 \ge \mathtt{C}\log(n)$.

Theorem 3.3. *Let a set V be split by a hole G with $\delta = |G|/|V| \ge 1/3$. Let the data density $f(\cdot)$ fulfill $f(x) \le f_G$ for $x \in G$ and $f(x) \ge f_1$ for $x \in V \setminus G$ with $f_G \le (1-\varepsilon)f_1$. Let $X_i \in V_1$, $X_j \in V_2$ be two sample points from different regions and let for some $k \le K$ and the corresponding bandwidth h_k, it holds*

$$\begin{gathered} \mathcal{B}(X_i, h_k) \cup \mathcal{B}(X_j, h_k) = V, \quad \mathcal{B}(X_i, h_k) \cap \mathcal{B}(X_j, h_k) \supseteq G, \\ |V|/3 \le |\mathcal{B}(X_i, h_k) \cap \mathcal{B}(X_j, h_k)| \le |V|/2. \end{gathered} \tag{3.1}$$

If $n\varepsilon^2 \ge \mathtt{C}\log(n)$ for a fixed sufficiently large constant $\mathtt{C}$, then the AWC procedure assigns the weight $w_{ij}^{(k)} = 0$ with high probability.

The conditions (3.1) of the theorem on the shape of the sets V and G can be easily relaxed. In fact we only need that the volume of the union $\mathcal{B}(X_i, h_k) \cup \mathcal{B}(X_j, h_k)$ to be of the order $|V|$ and significantly larger than the volume of the overlap $\mathcal{B}(X_i, h_k) \cap \mathcal{B}(X_j, h_k)$. In its turn, this overlap has to include a massive part of the hole G. The constants 1/3 and 1/2 in the last condition can be replaced by any other two positive constants $c_1 < c_2 < 1$.

3.3 Manifold clustering and high-dimensional data

The procedure is calibrated to ensure the propagation property which means a small probability of artificial clustering for a full dimensional homogeneous region. It appears that this propagation property automatically extends to the case of a low dimensional manifold structure. Suppose that the similarity measure $d(X_i, X_j)$ is based on the

Euclidean distance between X_i and X_j. Let also in a local vicinity of each data point X_i the remaining data concentrate in a small vicinity of a low dimensional linear subspace; see Figure 2.1 bottom right. This implies that the distances $d(X_i, X_j)$ correspond to the effective data dimension $\mathtt{p}_e$ rather than the original dimension p. Here we explain why the propagation property extends to this case. Indeed, the test statistics $T_{ij}^{(k)}$ are built on the base of the distance matrix $(d(X_i, X_j))_{i,j=1}^n$, and in the manifold case, $T_{ij}^{(k)}$ correspond to the effective dimension $\mathtt{p}_e$. The data dimension p does not show up there. There is only one place in the algorithm where the dimension p appears explicitly, namely, in the definition of the function $q(\cdot)$ from (2.5). And this function decreases with p; see Lemma A.3 from Appendix A. Artificial separation can only occur when $\widetilde{\theta}_{ij}^{(k)} < q_{ij}^{(k)}$. Probability of such an event becomes very small in the case of manifold data, because the estimated value $\widetilde{\theta}_{ij}^{(k)}$ corresponds to the data of effective dimension $\mathtt{p}_e$, while the value $q_{ij}^{(k)}$ is computed for the full dimension p. So, one can expect that the propagation effect will be even stronger along a low dimensional manifold. Note however that the arguments do not apply if a low dimensional manifold crosses another manifold of different dimension. Then the procedure indicates a non-homogeneity in the same way as in the case of two close regions with different densities.

The manifold property allows to easily work with high-dimensional data. Suppose that the data dimension p is large but the cluster structure corresponds to a low dimensional manifold of dimension m. And suppose that the distance/similarity matrix $(d(X_i, X_j))_{i,j=1}^n$ also corresponds to this manifold structure. The definition of the adaptive weights does not rely on the dimension p except the definition of the function $q(\cdot)$ from (2.5). We suggest to use the small "effective" dimension m instead of p for computing $q(t)$. If our guess m correctly mimics the effective dimension of the data then the AWC procedure will be properly tuned and preserve all its propagation and separation properties. In Section 4 we show the results of this approach applied on real text data.

4 Numerical examples and evaluation

This section illustrates the performance of AWC by mean of artificial and real datasets.

4.1 Artificial data

First examples serve to illustrate our main theoretical results. We start with the separation result of Theorems 3.2 and 3.3. Figure 4.1 shows the dataset composed of two uniform clusters with density f separated by a hole of lower density $f/2$ shown by vertical lines on the first figure. We fix a point X^* on the boarder of the left cluster marked by red $\times$. One can see that the local cluster of point X^* at original steps spreads to

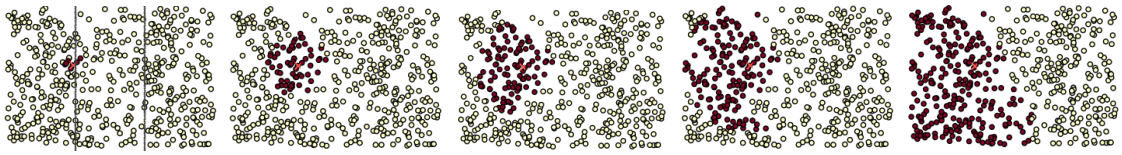

Figure 4.1: Steps 1, 40, 45, 47, 52. Black balls show the cluster for the red point.

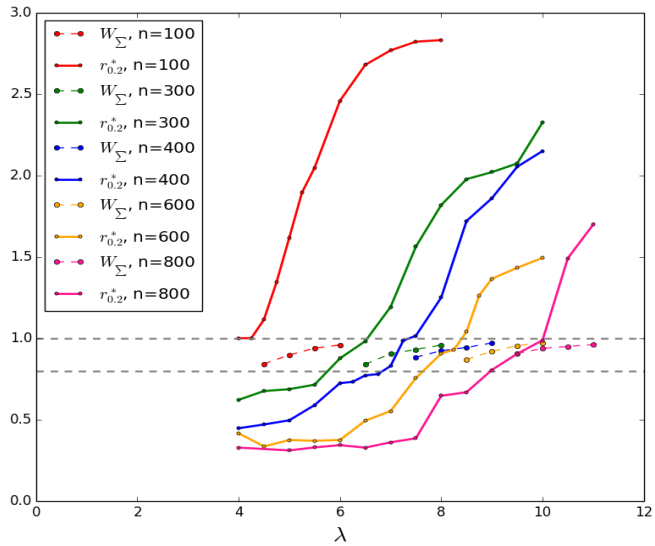

Figure 4.2: $r^*_{0.2}$ for standard Gaussian data (solid lines) and averaged values of $W_\Sigma(\lambda, 1)$ for the uniform data on the unit disk (dashed lines) for different n, λ.

the right until the radius r_k reaches the proper scale to detect the gap by our test. At the final step, the connections from the considered point X^* do not spread over the gap. As a result we have two clusters separated by a hole. More examples on separation with a gap are presented in Appendix D.

One Gaussian cluster Suppose that the data are sampled from a standard normal law $\mathcal{N}(0, I_p)$ in $I\!R^p$. The density in this case is concave only inside a unit ball with center in 0. Therefore, Theorem 3.1 implies the following behavior of AWC: with a high probability, it detects a cluster of points associated with the unit ball. We fix $p = 2$. Let $w_{ij}(\lambda)$ be the final weights of the AWC procedure for a particular realization of the data given by AWC with parameter λ. Define the connectedness coefficient $W_\Sigma(\lambda, r)$ for the ball of radius r:

$$W_\Sigma(\lambda, r) = \frac{\sum_{i,j} w_{ij}(\lambda)\, I\!\!1(\|X_i\| \le r, \|X_j\| \le r)}{\sum_{i,j} I\!\!1(\|X_i\| \le r, \|X_j\| \le r)}.$$

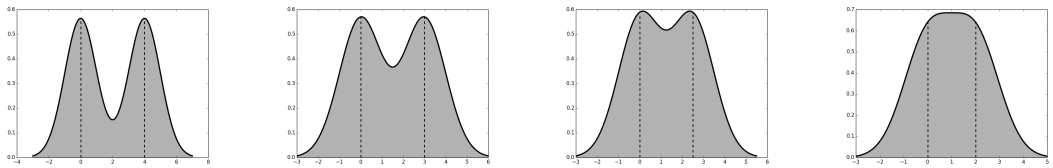

Figure 4.3: Mixture of two normals with variance 1 and distance D between means. *From left*: $D = 4;\ 3;\ 2.5;\ 2$.

Define also the radius $r^* = r^*_\alpha$ by the condition

$$I\!P\bigl(W_\Sigma(\lambda, r^*_\alpha) \geq 1 - \alpha\bigr) = 1 - \alpha.$$

Solid lines in Figure 4.2 show r^*_α for $\alpha = 0.2$, different λ and different sample sizes $n = 100, 300, 400, 600, 800$. In addition, compute the mean of $W_\Sigma(\lambda, 1)$ for the uniform distribution on the unit disk. These values are shown by the dashed lines on Figure 4.2. Comparing the dashed and solid lines of the same color on Figure 4.2 reveals that for a fixed n, the value λ which guaranties 80%-90% connectedness in the case of uniform distribution also guaranties in the case of normal distribution that the radius of central cluster is close to 1. This is in complete agreement with the claim of Theorem 3.1.

Separation for two Gaussian clusters Now we illustrate the separation properties of AWC on the example of two Gaussian clusters. In this case we want to check how AWC can find the possibly small gap between two clusters. Remind that AWC is a fully nonparametric method. A mixture of two Gaussian distribution with nearly the same mean is still unimodal and considered as one cluster. E.g. in the univariate case presented on Figure 4.3, when the distance between means is less than 2 there is no gap between clusters. Let $X_1, \ldots, X_n \in I\!R^2$ be generated from standard normal distribution $\mathcal{N}(0, I\!I_2)$ and $X_{n+1}, \ldots, X_{2n}$ be generated from $\mathcal{N}(D, I_2)$. Select the parameter λ due to suggestion of the previous section to ensure that the radius of detected central cluster is close to 1. Explicitly for $n = 100, 200, 300, 400, 600$ we took $\lambda = 4.2, 6, 6.5, 7.2, 8.3$ correspondingly. Here we are interested in the separation error e_s from (E.1). The ideal cluster separation in this experiment is given separated by the line $(D/2, y)$. Figure 4.4 shows an example of such realization. For each n and distance D we make 200 experiments. The averaged separation error e_{sp} as a function of distance between clusters D is shown on Figure 4.5. One can see that the separation error remains quite high for the distance $D \approx 2$ for all considered sample sizes. At the same time, if the distance D exceeds 3, the procedure starts to separate well the Gaussian clusters without using any prior information about the structure of the underlying density.

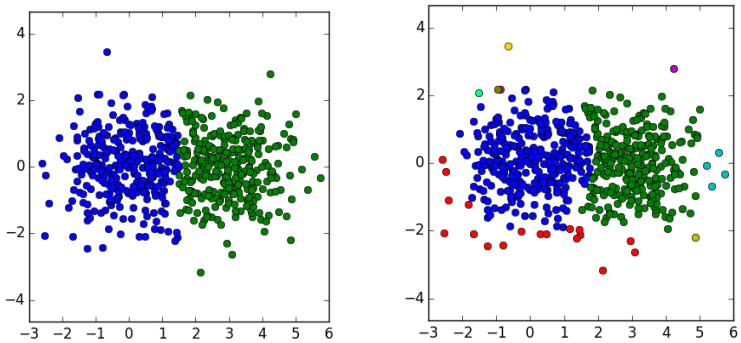

Figure 4.4: Mixture of two normals with $n = 300, D = 3$: ideal clustering vs AWC ($\lambda = 6.5, e_{sp} = .05$)

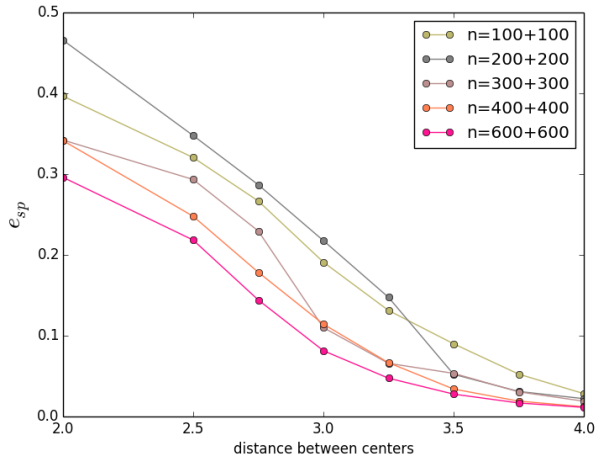

Figure 4.5: Separation error for the mixture of two normals

Performance for benchmark data Next we investigate the performance of AWC by mean of few popular artificial datasets with known cluster structure. The tuning parameter λ of the AWC is selected using "sum-of-weights" heuristics. First we show how AWC finds correct clusters in situations when other popular methods break down. For the comparison we used the Python implementation of the k-means, DBSCAN, spectral clustering and affinity propagation algorithms from scikit-learn tool Pedregosa et al. (2011). Each method requires to fix some tuning parameter(s) and we optimized the choice for each particular example while the AWC is used with the automatic choice. See Appendix C for details.

We consider 3 datasets. The *Pathbased* (300 points), Figure 4.6 top, consists of two clusters with Gaussian distribution surrounded by a circular cluster with an opening. The *Orange* dataset (268 points), Figure 4.6 bottom, is a ball with uniform density surrounded by uniformly distributed sphere with a little bit higher density. *Compound* Zahn (1971) is a dataset consisting of 399 points with various densities, see Figure 4.7. It

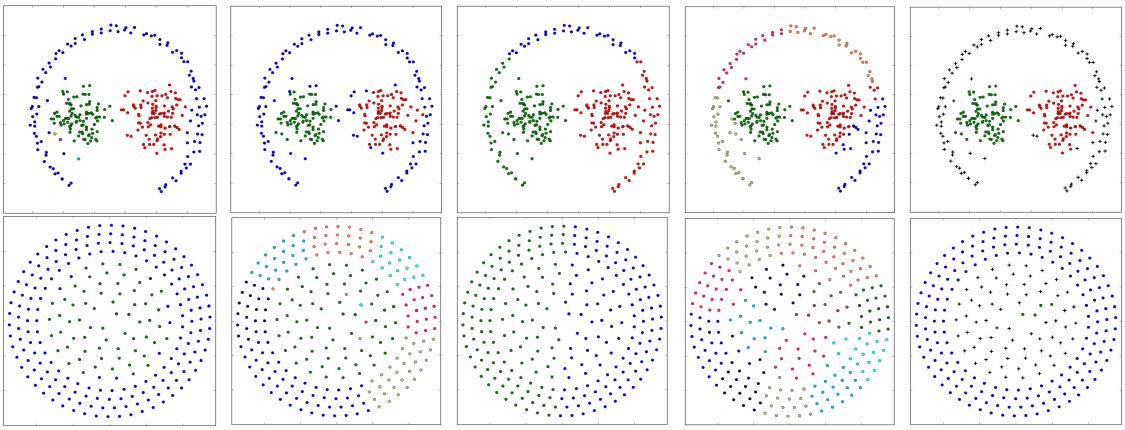

Figure 4.6: Top (pathbased), *from left*: AWC ($\lambda = 4.1$), Spectral ($\sigma = 0.1$), K-means ($K = 3$), Aff. prop. ($D = 0.5$, $P = -1464$), DBSCAN ($\varepsilon = 2.1$, *minsp*=10); Bottom (orange), *from left*: AWC ($\lambda = 2$) , spectral, K-means, affinity propagation, DBSCAN.

contains two nearly normal clusters, one small cluster surrounded by a ring cluster, and a dense cluster inside big sparse one. Figures show best performance of each comparative algorithm after parameter tuning. Each cluster found by the algorithms is represented by its own unique color. Noise points in DBSCAN result are marked by black crosses. One can see that AWC solves all challenges in these datasets such as non-convex clusters, overlapping clusters with different intensities, manifold clustering. Other algorithms even after optimizing can not handle most of them even after parameter tuning.

Other interesting examples are datasets $DS4, DS3$ from Karypis et al. (1999) used for CHAMELEON hierarchical clustering algorithm. The AWC results are shown on Figure 4.8 and we can see that AWC can handle these datasets as well. Many popular within the literature artificial datasets are collected in *https://github.com/deric/clustering-benchmark*. AWC performance on several of them is shown on Figure 4.9. These examples include the following challenges: manifold structure (spiral data), the density which slowly changes inside a cluster, dense clusters with a background of low density, a dense bridge between clusters etc. In all examples AWC does a very good job.

4.2 Text data

This section demonstrates the performance of AWC on text data, where the data dimension is very large. In our experiments we used 9 text datasets from the CLUTO toolkit Karypis (2002) which are widely used in the literature. The basic characteristics of the datasets are summarized in Table 1. The datasets dimension p ranges from 2886 to 10128 which makes these datasets a good benchmark for testing AWC manifold property on high-dimensional data. CLUTO provides already preprocessed datasets. This prepro-

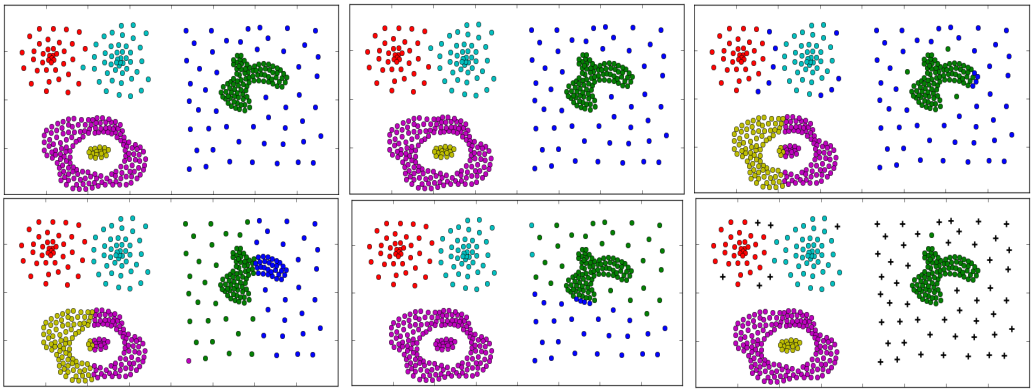

Figure 4.7: Compound. *Top, from left*: Original, AWC ($\lambda = 3.7$), Spectral ($\sigma = 0.1$); *Bottom, from left*: K-means ($K = 6$), Aff. prop. ($D = 0.5$, $P = -737$), DBSCAN ($\varepsilon = 1.48$, *minsp*=3)

Figure 4.8: AWC result for $DS4$ ($n = 10000$) and $DS3$ ($n = 8000$) with $\lambda = 15$.

cessing includes stop-word removal and stemming. In our experiments we represent the documents using the traditional vector space model with TF-IDF transformation: i-th document is presented as vector $X_i = \{x_{ij}\}_{j=1}^{d}$ where

$$x_{ij} = tf_{ij} \times idf_j, \qquad idf_j \stackrel{\text{def}}{=} \log(1+n) - \log(1+n_j) + 1.$$

Here tf_{ij} is the frequency of term j in the document i , n_j is the number of documents which contains the term j and idf_j is the inverse document frequency. The last one reflects how important a word is to a document in a collection. Originally the six datasets *tr11, tr12, tr23, tr31, tr41*, and *tr45* are derived from TREC collections (Text Retrieval Conference, http://trec.nist.gov). The datasets *re0* and *re1* are taken from Reuters-21578 text categorization test collection Lewis (1997). The dataset *wap* is from the WebACE project Boley et al. (1999) where each document corresponds to a web page listed in the subject hierarchy of Yahoo!. For evaluation we used Normalized Mutual Information NMI from Strehl and Ghosh (2002), which is a popular measure for clustering

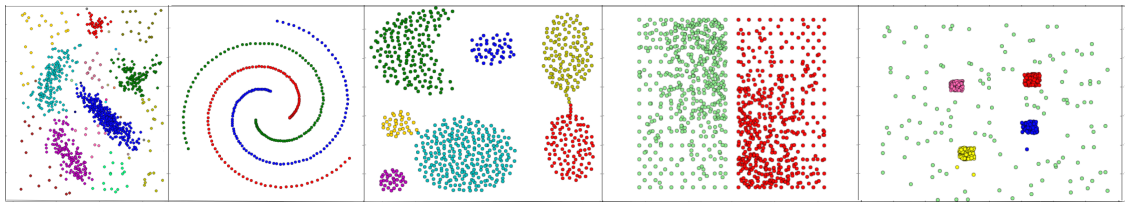

Figure 4.9: AWC result for artificial datasets

accuracy in text data literature. For a true clustering structure $\mathcal{C}^* = \{\mathcal{C}^*_m\}_{m=1}^M$ and some other structure $\mathcal{C} = \{\mathcal{C}_l\}_{l=1}^L$, define $n_{ml} = |\mathcal{C}^*_m \cap \mathcal{C}_l|$, $n^*_m = |\mathcal{C}^*_m|$, $n_l = |\mathcal{C}_l|$ and

$$\mathrm{NMI}(\mathcal{C}, \mathcal{C}^*) = \frac{\sum_{ml} n_{ml} \log \frac{n\, n_{ml}}{n^*_m\, n_l}}{\sqrt{\sum_m n^*_m \log(n^*_m/n) \ \sum_l n_l \log(n_l/n)}}\,.$$

We compared AWC with state-of-the-art algorithms for clustering textual data: Spectral clustering with Normalized Cut (NCut) Shi and Malik (2000); Local Learning based Clustering Algorithm (LLCA) Wu and Schölkopf (2006); Clustering via Local Regression (CLOR), Sun et al. (2008); Regularized Local Reconstruction for Clustering (RLRC) Sun et al. (2009). These methods belong to the group of spectral clustering approaches. The results for these algorithms on our benchmark are taken from the work Sun et al. (2009). All methods were provided with correct number of clusters K and the neighborhood size $k = 40$. For more details about experimental settings we refer to Sun et al. (2009). The LLCA results are obtained after tuning the parameters. We also use the *vcluster* package from CLUTO toolkit Karypis (2002) which provides a bisecting graph partitioning-based algorithm. We used it with the prespecified correct number of clusters K and the neighborhood size $k = 40$. Other parameters were set by default. The results of all methods are presented in Table 1. The maximal two numbers in each row are marked in bold. For CLUTO, after 100 runs, we calculated the best and the worst result among these runs, they are presented in the corresponding column of Table 1 in the form $[\mathrm{NMI}_{worst}, \mathrm{NMI}_{best}]$.

AWC also have a starting neighborhood size $n_0 = 40$ which is similar to other methods. The Euclidean distance was used as similarity measure. Another parameter of AWC is the effective dimension $\mathtt{p}_e$ used in (2.5) for computing the values $q_{ij}^{(k)}$. We just set $\mathtt{p}_e = 2$. In the Table 1 the result of AWC after tuning λ is marked by AWC^*. By AWC_s we marked the result obtained with λ chosen by "sum-of-weights" heuristic. One can see that in most cases "sum-of-weights" heuristic result is close to optimal. Table 1 shows that in 7 out of 9 datasets (*tr11, tr23, tr31, tr45, re0, re1, wap*) the result of AWC is similar to the best result among all considered state-of-the-art algorithms. It is worth mentioning that all methods except AWC were provided with the correct number

Table 1: NMI for real-world text data sets, two best results in each row are in bold.

	Algorithm									
Data	AWC*	AWC_s	NCut	LLCA	CLOR	RLRC	CLUTO	n	d	K
tr11	**0.715**	0.712	0.624	0.620	0.674	**0.723**	[0.637, 0.706]	414	6429	9
tr12	0.652	0.624	0.599	0.622	0.657	**0.737**	[0.632, **0.732**]	313	5804	8
tr23	**0.457**	0.34	0.344	0.298	0.334	0.357	[0.409, **0.446**]	204	5832	6
tr31	**0.641**	0.602	0.457	0.499	0.483	0.534	[0.615, **0.661**]	927	10128	7
tr41	0.639	0.585	0.603	0.622	**0.642**	0.620	[0.639, **0.697**]	878	7454	10
tr45	**0.72**	0.682	0.558	0.585	0.631	0.664	[0.605, **0.708**]	690	8261	10
re0	**0.468**	**0.458**	0.401	0.409	0.426	0.395	[0.367, 0.429]	1504	2886	12
re1	**0.609**	0.583	0.484	0.485	0.498	0.496	[0.555, **0.607**]	1657	3758	24
wap	**0.598**	0.586	0.525	0.542	0.541	0.577	[0.578, **0.611**]	1560	8460	19

of clusters K. In addition all considered algorithms were constructed specially for text data whereas AWC is remained unchanged and all results in this and other sections are obtained by the same algorithm.

5 Proofs

This section presents the proofs of the main results. First we show that the value $\widetilde{\theta}_{ij}^{(k)}$ is a root-n consistent estimator of the value $\theta_{ij}^{(k)}$ for any two neighbor balls $\mathcal{B}(X_i, h_k)$ and $\mathcal{B}(X_j, h_k)$. Unlike standard results from empirical process theory, this bound is dimension free and does not involve any entropy number. The proof mainly uses combinatorial arguments.

Lemma 5.1. *For any $k \le K$ and any $i \ne j$ with $d(X_i, X_j) \le h_k$, let the value $\theta_{ij}^{(k)}$ be defined by* (2.2) *and its estimate $\widetilde{\theta}_{ij}^{(k)}$ by* (2.1)*. Then it holds for a fixed constant $\mathfrak{z}$ on a random set of probability at least $1 - 2\mathrm{e}^{-\mathfrak{z}}$*

$$I\!P\left(N_{i \vee j}^{(k)}\, \mathcal{K}(\widetilde{\theta}_{ij}^{(k)}, \theta_{ij}^{(k)}) > \mathfrak{z}\right) \le 2\mathrm{e}^{-\mathfrak{z}}. \tag{5.1}$$

Proof. Let us fix a step k and a pair of points X_i, X_j with $d(X_i, X_j) \le h_k$. Without loss of generality, we assume $i = 1$ and $j = 2$. Denote

$$\mathcal{B}_{12} \stackrel{\mathrm{def}}{=} \mathcal{B}(X_1, h_k) \cup \mathcal{B}(X_2, h_k), \qquad \mathcal{O}_{12} \stackrel{\mathrm{def}}{=} \mathcal{B}(X_1, h_k) \cap \mathcal{B}(X_2, h_k).$$

Given X_1, X_2 the remaining observations $X_3, \ldots, X_n$ are still i.i.d. from the same distribution. Let also $\mathcal{S}$ be the index subset of the set $\{3, \ldots, n\}$. Introduce the random

event $A_{\mathcal{S}}$ by conditions $X_\ell \in \mathcal{B}_{12}$ for $\ell \in \mathcal{S}$ and $X_\ell \notin \mathcal{B}_{12}$ for $\ell \in \mathcal{S}^c \stackrel{\text{def}}{=} \{3,\ldots,n\} \setminus \mathcal{S}$:

$$A_{\mathcal{S}} \stackrel{\text{def}}{=} \{X_\ell \in \mathcal{B}_{12}, \ell \in \mathcal{S}, X_\ell \notin \mathcal{B}_{12}, \ell \in \mathcal{S}^c\}.$$

After conditioning on X_1, X_2 and on $A_{\mathcal{S}}$, the subsample $\{X_\ell\}_{\ell \in \mathcal{S}}$ is still i.i.d. with the conditional density $f(x)/I\!P(A_{\mathcal{S}} \,|\, X_1, X_2)$. Therefore, the $\xi_\ell = 1\!\!1(X_\ell \in \mathcal{O}_{12})$'s are given $X_1, X_2, A_{\mathcal{S}}$ i.i.d. Bernoulli with the parameter $\theta_{\mathcal{S}} = \theta_{12}^{(k)}$. The deviation bound from Polzehl and Spokoiny (2006) implies for the normalized sum $\widetilde{\theta}_{\mathcal{S}} \stackrel{\text{def}}{=} N_{\mathcal{S}}^{-1} \sum_{\mathcal{S}} \xi_\ell$ with $N_{\mathcal{S}} \stackrel{\text{def}}{=} |\mathcal{S}|$:

$$I\!P\Big(N_{\mathcal{S}} \mathcal{K}(\widetilde{\theta}_{\mathcal{S}}, \theta_{\mathcal{S}}) > \mathfrak{z} \,\Big|\, X_1, X_2, A_{\mathcal{S}}\Big) \le 2\mathrm{e}^{-\mathfrak{z}}; \quad \mathfrak{z} \ge 0.$$

As the right hand-side of this inequality does not depend on X_1, X_2, $\mathcal{S}$, and $A_{\mathcal{S}}$, the bound applies for the joint distribution in the unconditional form yielding (5.1). □

Proof of Theorem 3.1 Suppose that the density function $f(x)$ fulfills one of two theorem conditions. Let also all the weights $w_{ij}^{(m)}$ for $m < k$ computed at the first $k-1$ steps of the algorithm are equal to one. It remains to show that the next step k leads to the same results. Our inductive assumption means that we consider non-adaptive weights $w_{ij}^{(k)}$ which only account to the distance between points X_i, X_j, and X_ℓ for all $\ell \neq i, j$ with $d(X_i, X_\ell) \le h_k$ or $d(X_j, X_\ell) \le h_k$. Now Lemma 5.1 ensures (5.1) for any pair X_i, X_j with $d(X_i, X_j) \le h_k$ and any $k \ge 1$. Also by Lemma A.2, it holds $\theta_{ij}^{(k)} \ge q_{ij}^{(k)}$. For the event $\widetilde{\theta}_{ij}^{(k)} < q_{ij}^{(k)} \le \theta_{ij}^{(k)}$ we are interested in, this implies by convexity of the Kullback-Leibler divergence w.r.t. the first argument that

$$I\!P\Big(N_{i \vee j}^{(k)} \mathcal{K}(\widetilde{\theta}_{ij}^{(k)}, q_{ij}^{(k)})\, 1\!\!1(\widetilde{\theta}_{ij}^{(k)} < q_{ij}^{(k)}) > \mathfrak{z}\Big) \le I\!P\Big(N_{i \vee j}^{(k)} \mathcal{K}(\widetilde{\theta}_{ij}^{(k)}, \theta_{ij}^{(k)}) > \mathfrak{z}\Big) \le 2\mathrm{e}^{-\mathfrak{z}}.$$

This implies a uniform bound: for an absolute constant $\mathtt{C} \le 4$

$$I\!P\left(\max_{i \neq j} \max_{k \ge 1} N_{i \vee j}^{(k)} \mathcal{K}(\widetilde{\theta}_{ij}^{(k)}, q_{ij}^{(k)})\, 1\!\!1(\widetilde{\theta}_{ij}^{(k)} < q_{ij}^{(k)}) > \mathtt{C} \log n\right) \le \frac{2}{n}.$$

Proof of Theorem 3.2 Let V be a set with the volume $|V|$ and G be a splitting hole with volume $|G|$ such that $V_G \stackrel{\text{def}}{=} V \setminus G$ consists of two disjoint regions. Let also the data density be equal to p_G on G and to p_{V_G} on the complement $V \setminus G$. Consider two hypothesis H_0 of "no gap" $p_G = p_{V_G} = 1/|V|$ and H_G of a G-gap with $p_G = (1-\varepsilon) p_{V_G}$. We are interested to understand the conditions which enable us to separate these two hypotheses. Define $\delta = |G|/|V|$, so that $|V_G|/|V| = 1 - \delta$. Then under H_0 the data distribution is uniform on the set V with the density $p_0 = 1/|V|$. Further, under H_G the data density is uniform on G with the density p_G and on its

complement $V \setminus G$ with the density p_{V_G} satisfying

$$|V_G| p_{V_G} + |G| p_G = (1-\delta)|V| p_{V_G} + \delta |V| (1-\varepsilon) p_{V_G} = 1,$$

yielding

$$p_{V_G} = \frac{p_0}{1-\delta\varepsilon}, \qquad p_G = \frac{p_0(1-\varepsilon)}{1-\delta\varepsilon}. \tag{5.2}$$

For the experiment with N observations, the condition of consistent separation between H_0 and H_G is that the total Kullback-Leibler (KL) divergence between two distributions converges to infinity. The KL divergence for the model with N i.i.d. observations is defined as $\mathcal{K}(I\!P_0, I\!P_G) = I\!E_0 \log(dI\!P_0/dI\!P_G)$. As $I\!P_0(G) = |G| p_0 = |G|/|V| = \delta$, it follows by (5.2)

$$\begin{aligned}\mathcal{K}(I\!P_0, I\!P_G) &= N I\!P_0(G) \log \frac{p_0}{p_G} + N I\!P_0(V_G) \log \frac{p_0}{p_{V_G}} \\ &= N\delta \log \frac{1-\delta\varepsilon}{1-\varepsilon} + N(1-\delta)\log(1-\delta\varepsilon) = N\log(1-\delta\varepsilon) - N\delta\log(1-\varepsilon).\end{aligned}$$

If G is a hole of a fixed volume $\delta|V|$ and $\varepsilon = \varepsilon_N \to 0$, then

$$\mathcal{K}(I\!P_0, I\!P_G) = 0.5(\delta - \delta^2) N \varepsilon_N^2 \bigl\{1 + O(\varepsilon_N)\bigr\}$$

and consistent separation between P_0 and P_G is impossible if $N\varepsilon_N^2$ remains bounded by a fixed constant as N grows.

Proof of Theorem 3.3 Now we show that the AWC algorithm does a good job in detecting a gap between two neighbor clusters separated by a hole G of the volume $|G| = \delta|V|$ and the piecewise constant density given by (5.2). Let V consist of three neighbor regions of equal cylindric shape of height h and base radius ρh for some $\rho < 1$. The hole G corresponds to the central part, so that $|G| = |V|/3$ and $\delta = 1/3$. We consider two points X_i, X_j from different side of the hole separated by a distance $\|X_i - X_j\| \ge h_k \ge h$ at the step k. Due to the definition, it is sufficient to show that the corresponding test statistic $T_{ij}^{(k)}$ exceeds λ. We sketch the proof of this fact for the "worst case" situation that the procedure did not gain any structural information during the first $k-1$ steps and all the earlier computed adaptive weights $w_{il}^{(k-1)}$ and $w_{jl}^{(k-1)}$ coincide with the non-adaptive distance based weights, i.e. they are equal to one within the balls of radius h_{k-1} around these points. By the theorem conditions, it holds

$$A_{i \vee j} \stackrel{\text{def}}{=} \mathcal{B}(X_i, h_k) \cup \mathcal{B}(X_j, h_k) = V, \quad A_{i \wedge j} \supset G,$$

and the value $q_{ij}^{(k)} = |A_{i\wedge j}|/|V|$ satisfies $1/3 \le q_{ij}^{(k)} \le 1/2$. As $I\!P(A_{i\vee j}) = I\!P(V) = 1$, it holds

$$\theta_{ij}^{(k)} = I\!P(A_{i\wedge j}) = p_G|G| + p_{V_G}|A_{i\wedge j} \setminus G|\,,$$
$$= \bigl\{|G|(1-\varepsilon) + (|A_{i\wedge j}| - |G|)\bigr\}p_{V_G} = \frac{|A_{i\wedge j}| - \varepsilon|G|}{|V|(1-\delta\varepsilon)} = \frac{q_{ij}^{(k)} - \delta\varepsilon}{1-\delta\varepsilon}$$

yielding

$$q_{ij}^{(k)} - \theta_{ij}^{(k)} = \frac{(1-q_{ij}^{(k)})\delta\varepsilon}{1-\delta\varepsilon} \ge \mathtt{C}\varepsilon$$

with $\mathtt{C} \ge 1/6$. In particular, this means that $\theta_{ij}^{(k)} < q_{ij}^{(k)}$. Also one can bound

$$\mathcal{K}^{1/2}\bigl(\theta_{ij}^{(k)}, q_{ij}^{(k)}\bigr) \ge \mathtt{C}_1\varepsilon \tag{5.3}$$

with a slightly different constant $\mathtt{C}_1$. To show that $\widetilde{\theta}_{ij}^{(k)}$ is significantly smaller than $q_{ij}^{(k)}$, we apply Lemma 5.1. The condition $A_{i\vee j} = V$ implies $N_{i\vee j} = n$ and by Lemma 5.1

$$n\,\mathcal{K}\bigl(\widetilde{\theta}_{ij}^{(k)}, \theta_{ij}^{(k)}\bigr) \le \mathtt{C}_2\log(n)\,. \tag{5.4}$$

If $\widetilde{\theta}_{ij}^{(k)} \le \theta_{ij}^{(k)}$, then $\mathcal{K}\bigl(\widetilde{\theta}_{ij}^{(k)}, q_{ij}^{(k)}\bigr) \ge \mathcal{K}\bigl(\theta_{ij}^{(k)}, q_{ij}^{(k)}\bigr)$. If $\theta_{ij}^{(k)} < \widetilde{\theta}_{ij}^{(k)} \le q_{ij}^{(k)}$, then regularity and convexity of $\mathcal{K}(x,q)$ w.r.t. x, q implies

$$\mathcal{K}^{1/2}\bigl(\widetilde{\theta}_{ij}^{(k)}, q_{ij}^{(k)}\bigr) \ge \mathtt{a}\,\mathcal{K}^{1/2}\bigl(\theta_{ij}^{(k)}, q_{ij}^{(k)}\bigr) - \mathcal{K}^{1/2}\bigl(\widetilde{\theta}_{ij}^{(k)}, \theta_{ij}^{(k)}\bigr)$$

for some fixed constant $\mathtt{a} > 0$; see Polzehl and Spokoiny (2006) for more details. This together with (5.3) and (5.4) implies

$$\mathcal{K}^{1/2}(\widetilde{\theta}_{ij}^{(k)}, q_{ij}^{(k)}) \ge \mathtt{a}\mathtt{C}_1\varepsilon - \sqrt{\mathtt{C}_2 n^{-1}\,\log n} \ge \sqrt{\lambda/n}$$

provided that $\mathtt{a}\mathtt{C}_1\varepsilon\sqrt{n} \ge \sqrt{\mathtt{C}_2\log(n)} + \sqrt{\lambda}$. Together with the bound $\lambda \le \mathtt{C}\log(n)$ this yields a consistent separation $w_{ij}^{(k)} = 1$ under condition $\varepsilon^2 \ge \mathtt{C}n^{-1}\log(n)$.

6 Conclusion

The proposed procedure AWC systematically exploits the idea of extracting the structural information about the underlying data distribution from the observed data in terms of adaptive weights and uses this information for sensitive clustering. The method is appealing and computationally feasible, the numerical results indicate the state-of-the-art performance of the method. Theoretical results show its optimality in separating of neighbor regions.

References

Aggarwal, C. C. and Reddy, C. K. (2013). *Data clustering: algorithms and applications.* CRC Press.

Boley, D., Gini, M., Gross, R., Han, E.-H. S., Hastings, K., Karypis, G., Kumar, V., Mobasher, B., and Moore, J. (1999). Document categorization and query generation on the world wide web using webace. *Artificial Intelligence Review*, 13(5-6):365–391.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD96*, number 34, pages 226–231.

Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *science*, 315(5814):972–976.

Frick, K., Munk, A., and Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):495–580.

Hartigan, J. A. (1975). *Clustering algorithms.* Wiley Series in Probability and Mathematical Statistics. New York etc.: John Wiley Sons. XIII, 322 p. (1975).

Karypis, G. (2002). Cluto-a clustering toolkit. Technical report, DTIC Document.

Karypis, G., Han, E.-H., and Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75.

Lewis, D. D. (1997). Reuters-21578 text categorization test collection, distribution 1.0. *http://www. research. att. com/˜ lewis/reuters21578. html.*

Li, S. (2011). Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, 4(1):66–70.

Lichman, M. (2013). UCI machine learning repository.

Ng, A. Y., Jordan, M. I., Weiss, Y., et al. (2002). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830.

Polzehl, J. and Spokoiny, V. (2006). Propagation-separation approach for local likelihood estimation. *Probab. Theory Relat. Fields*, 135(3):335–362.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Rigollet, P. (2007). Generalization error bounds in semi-supervised classification under the cluster assumption. *J. Mach. Learn. Res.*, 8:1369–1392.

Seeger, M. (2001). Learning with labeled and unlabeled data. Technical report, University of Edinburgh.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905.

Spokoiny, V. and Vial, C. (2009). Parameter tuning in pointwise adaptation using a propagation approach. *Ann. Statist.*, 37(5B):2783–2807.

Steinhaus, H. (1956). Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci*, 1(804):801.

Strehl, A. and Ghosh, J. (2002). Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617.

Sun, J., Shen, Z., Li, H., and Shen, Y. (2008). Clustering via local regression. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 456–471. Springer.

Sun, J., Shen, Z., Su, B., and Shen, Y. (2009). Regularized local reconstruction for clustering. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 110–121. Springer.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

Wishart, D. (1969). Mode analysis: A generalization of nearest neighbor which reduces chaining effects. In Cole, A., editor, *Numerical Taxonomy*, pages 282–311. AP.

Wu, M. and Schölkopf, B. (2006). A local learning approach for clustering. In *Advances in neural information processing systems*, pages 1529–1536.

Zahn, C. T. (1971). Graph-theoretical methods for detecting and describing gestalt clusters. *Computers, IEEE Transactions on*, 100(1):68–86.

Zelnik-Manor, L. and Perona, P. (2004). Self-tuning spectral clustering. In *Advances in neural information processing systems*, pages 1601–1608.

Zhu, X. (2005). Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison.

Appendix A Technical proofs

Lemma A.1. *Let $X_1, \ldots, X_n \in I\!R^p$ be an i.i.d. sample and B, C be two non-overlapping measurable sets in $I\!R^p$. For a given value $q \in (0,1)$, define one sided hypotheses*

$$H_0 : I\!P(B) \geq q(I\!P(B) + I\!P(C)),$$

$$H_1 : I\!P(B) < q(I\!P(B) + I\!P(C)).$$

Then the likelihood-ratio test statistic T for testing the null hypothesis H_0 against the alternative H_1 is given by

$$T = (S_B + S_C)\,\mathcal{K}\bigl(\widetilde{\theta}, q\bigr)\,\bigl\{1\!\!1(\widetilde{\theta} \leq q) - 1\!\!1(\widetilde{\theta} > q)\bigr\},$$

where $\mathcal{K}(\theta, \eta)$ is the Kullback-Leibler (KL) divergence between two Bernoulli laws with parameters θ and η:

$$\mathcal{K}(\theta, \eta) \stackrel{\text{def}}{=} \theta \log \frac{\theta}{\eta} + (1-\theta) \log \frac{1-\theta}{1-\eta}$$

and

$$\widetilde{\theta} = \frac{S_B}{S_B + S_C}\,. \tag{A.1}$$

Proof. Define A as the complement of B and C: $A \stackrel{\text{def}}{=} (B \cup C)^c$. Let also $a = I\!P(A)$, $b = I\!P(B)$, $c = I\!P(C)$, and

$$S_A \stackrel{\text{def}}{=} \sum_{i=1}^n 1\!\!1(X_i \in A), \quad S_B \stackrel{\text{def}}{=} \sum_{i=1}^n 1\!\!1(X_i \in B), \quad S_C \stackrel{\text{def}}{=} \sum_{i=1}^n 1\!\!1(X_i \in C).$$

The log-likelihood $L(a,b,c)$ for the multinomial model with the parameter (a,b,c) reads

$$L(a,b,c) = S_A \log a + S_B \log b + S_C \log c + R,$$

where the remainder R does not depend on a, b, c from $[0,1]$. Now for a fixed $\rho \in [0,1]$, define

$$\widehat{L}(\rho) \stackrel{\text{def}}{=} \sup_{a+b+c=1,\, b=\rho(b+c)} L(a,b,c).$$

Then, the maximum likelihood under null hypothesis H_0 can be written as

$$\widehat{L}_0 \stackrel{\text{def}}{=} \sup_{1>\rho\geq q} \widehat{L}(\rho).$$

Under the alternative

$$\widehat{L}_1 \stackrel{\text{def}}{=} \sup_{0<\rho<q} \widehat{L}(\rho),$$

and the likelihood ratio test statistic T is defined as the difference between $\widehat{L}_1$ and $\widehat{L}_0$:

$$T \stackrel{\text{def}}{=} \widehat{L}_1 - \widehat{L}_0 .$$

Introduce also the quantity

$$\widehat{L} = \sup_{a+b+c=1} L(a,b,c)$$

Optimization under the constraint $a+b+c=1$ yields in view of $S_A + S_B + S_C = n$ that

$$\widehat{L} = S_A \log \frac{S_A}{n} + S_B \log \frac{S_B}{n} + S_C \log \frac{S_C}{n} + R .$$

It is also easy to see that

$$\widehat{L} = \max_{\rho} \widehat{L}(\rho) = \widehat{L}(\widetilde{\theta})$$

with $\widetilde{\theta}$ from (A.3).

Similar optimization under the additional constraint $b = \rho(b+c)$ (see below at the end of the proof)

$$\begin{aligned} \widehat{L}(\rho) &\stackrel{\text{def}}{=} \sup_{a+b+c=1,\, b=\rho(b+c)} L(a,b,c) \\ &= S_A \log \frac{S_A}{n} + (S_B + S_C) \log \frac{S_B + S_C}{n} + S_B \log \rho + S_C \log(1-\rho) . \end{aligned} \tag{A.2}$$

Consider the derivative of $\widehat{L}(\rho)$:

$$\frac{\partial \widehat{L}(\rho)}{\partial \rho} = \frac{S_B}{\rho} - \frac{S_C}{1-\rho} = \frac{S_B - (S_B + S_C)\rho}{\rho(1-\rho)} = \frac{(S_B + S_C)}{\rho(1-\rho)}(\widetilde{\theta} - \rho). \tag{A.3}$$

It follows

$$\begin{aligned} \frac{\partial \widehat{L}(\rho)}{\partial \rho} > 0 &\iff 0 < \rho < \widetilde{\theta} \\ \frac{\partial \widehat{L}(\rho)}{\partial \rho} < 0 &\iff 1 > \rho > \widetilde{\theta} . \end{aligned}$$

To calculate $\widehat{L}_0, \widehat{L}_1$ we need to consider two cases:

$$\begin{aligned} q \le \widetilde{\theta} &\implies \widehat{L}_0 = \widehat{L}, \quad \widehat{L}_1 = \widehat{L}(q) \\ q > \widetilde{\theta} &\implies \widehat{L}_0 = \widehat{L}(q), \quad \widehat{L}_1 = \widehat{L}. \end{aligned}$$

The likelihood ratio test statistic is defined as the difference between $\widehat{L}$ and $\widehat{L}_0$:

$$\begin{aligned} T &\stackrel{\text{def}}{=} \widehat{L}_1 - \widehat{L}_0 = \bigl\{\widehat{L} - \widehat{L}(q)\bigr\}\bigl\{\mathbb{1}(\widetilde{\theta} \le q) - \mathbb{1}(\widetilde{\theta} > q)\bigr\} \\ &= (S_B + S_C)\Bigl\{\widetilde{\theta}\log\frac{\widetilde{\theta}}{q} + (1-\widetilde{\theta})\log\frac{1-\widetilde{\theta}}{1-q}\Bigr\}\bigl\{\mathbb{1}(\widetilde{\theta} \le q) - \mathbb{1}(\widetilde{\theta} > q)\bigr\}. \end{aligned}$$

Note that this test statistic can be written as

$$T = (S_B + S_C)\mathcal{K}\bigl(\widetilde{\theta}, q\bigr)\bigl\{\mathbb{1}(\widetilde{\theta} \le q) - \mathbb{1}(\widetilde{\theta} > q)\bigr\}$$

as required.

It remains to check (A.2). The Lagrange function for this optimization problem reads as follows

$$\mathcal{L}(a,b,c,\nu,\mu) = S_A\log a + S_B\log b + S_C\log c - \nu(a+b+c-1) - \mu(b - \rho(b+c)).$$

The partial derivatives of the Lagrange function are:

$$\begin{cases} \dfrac{\partial\mathcal{L}}{\partial a} = \dfrac{S_A}{a} - \nu = 0 \\ \dfrac{\partial\mathcal{L}}{\partial b} = \dfrac{S_B}{b} - \nu - \mu(1-\rho) = 0 \\ \dfrac{\partial\mathcal{L}}{\partial c} = \dfrac{S_C}{c} - \nu + \mu\rho = 0 \\ \dfrac{\partial\mathcal{L}}{\partial \nu} = a + b + c - 1 = 0 \\ \dfrac{\partial\mathcal{L}}{\partial \mu} = b - \rho(b+c) = 0. \end{cases}$$

These equations can be rewritten as follows:

$$\begin{cases} a = \dfrac{S_A}{\nu} \\ b = \dfrac{S_B}{\nu + \mu(1-\rho)} \\ c = \dfrac{S_C}{\nu - \mu\rho} \\ 1 = a + b + c \\ c = \dfrac{b(1-\rho)}{\rho} \end{cases}$$

Combining second, third and fifth equations

$$\begin{aligned} \frac{S_C}{\nu - \mu\rho} &= \frac{(1-\rho)}{\rho}\frac{S_B}{\nu + \mu(1-\rho)} \\ \mu &= -\nu\frac{\rho(S_B + S_C) - S_B}{\rho(1-\rho)(S_B + S_C)} \end{aligned}$$

and

$$b = \nu^{-1} \frac{S_B}{1 - \frac{\rho(S_B+S_C)-S_B}{\rho(S_B+S_C)}} = \frac{\rho(S_B+S_C)}{\nu}$$
$$c = \nu^{-1} \frac{S_C}{1 + \frac{\rho(S_B+S_C)-S_B}{(1-\rho)(S_B+S_C)}} = \frac{(1-\rho)(S_B+S_C)}{\nu}.$$

It follows from $a+b+c=1$:

$$\frac{S_A}{\nu} + \frac{\rho(S_B+S_C)}{\nu} + \frac{(1-\rho)(S_B+S_C)}{\nu} = 1$$
$$\nu = S_A + S_B + S_C = n.$$

Finally we derive

$$a = \frac{S_A}{n}$$
$$b = \frac{\rho(S_B+S_C)}{n}$$
$$c = \frac{(1-\rho)(S_B+S_C)}{n}$$

which yields the assertion. □

Our next lemma helps to check that in a region with a linear or univariate concave density, the gap coefficient for any two overlapping balls is not smaller than 1. Here we assume that $d(X_i, X_j) = \|X_i - X_j\|$.

Lemma A.2. *Consider the situation with a linear or univariate concave density $f(x)$ for $x \in V$. For any pair $X_i, X_j \in V$ with $d(X_i, X_j) \le h_k$, the value $\theta_{ij}^{(k)}$ from (2.2) fulfills*

$$\frac{\theta_{ij}^{(k)}}{q_{ij}^{(k)}} \ge 1,$$

where the value $q_{ij}^{(k)}$ corresponds to a constant density $f_0(x) \equiv \mathtt{C}$.

Proof. We write h in place of h_k for ease of notation. In the case of a linear density, it holds $\theta_{ij}^{(k)} = q_{ij}^{(k)}$ by symmetricity arguments. In the case of a univariate concave density consider a linear function $g(x)$ such that it coincides with $f(x)$ in points $X_i - h$ and $-X_i + h$: $g(X_i - h) = f(X_i - h), g(-X_i + h) = f(-X_i + h)$. Concavity of $f(x)$ implies

$$f(x) \ge g(x), \qquad x \in A \stackrel{\text{def}}{=} [X_i - h, -X_i + h],$$
$$f(x) \le g(x), \qquad x \in B \stackrel{\text{def}}{=} [-X_i - h, X_i + h] \setminus [X_i - h, -X_i + h].$$

It follows

$$\frac{\int_A f(x)dx}{\int_A f(x)dx + \int_B f(x)dx} \geq \frac{\int_A g(x)dx}{\int_A g(x)dx + \int_B f(x)dx} \geq \frac{\int_A g(x)dx}{\int_A g(x)dx + \int_B g(x)dx} = q.$$

This yields the result. □

One more result specifies the case of manifold structure for the underlying density.

Lemma A.3. *Fix a d_Π-dimensional hyper-plane $\Pi \in I\!R^p$, $p_\Pi < p$. Consider the manifold M in $I\!R^p$ which can be represented as*

$$M = \bigcup_{x \in \Omega} \Pi_x$$

where $\Omega \in I\!R^p$ is a convex set of dimension $p_\Omega \leq p - p_\Pi$ and diameter h, such that subspace of Ω is orthogonal to Π. Π_x is a shifted hyper-plane Π such that $x \in \Pi_x$. Consider two points $O_1, O_2 \in M$ and two balls $\mathcal{B}_1 = \mathcal{B}(O_1, R)$, $\mathcal{B}_2 = \mathcal{B}(O_2, R)$ with radius R and centers in O_1, O_2. Then for $R \gg h$ it holds

$$\frac{V_d((\mathcal{B}_1 \cap \mathcal{B}_2) \cap M)}{V_d((\mathcal{B}_1 \cup \mathcal{B}_2) \cap M)} \approx q_{p_\Pi} > q_p$$

where q_p is equal to $q\left(\frac{|O_1O_2|}{R}\right)$ from (2.5) with the corresponding dimension p, $|O_1O_2|$ is the distance between O_1, O_2, V_p is p-dimensional volume.

Proof. The considered case is represented on Figure 2.1 bottom right. Then

$$\begin{aligned}\frac{V_d((\mathcal{B}_1 \cap \mathcal{B}_2) \cap M)}{V_p((\mathcal{B}_1 \cup \mathcal{B}_2) \cap M)} &= \frac{\int_{x \in \Omega} V_{p_\Pi}((\mathcal{B}_1 \cap \mathcal{B}_2) \cap \Pi_x)\, dx}{\int_{x \in \Omega} V_{p_\Pi}((\mathcal{B}_1 \cup \mathcal{B}_2) \cap \Pi_x)} \approx (R \gg h) \\ &\approx \frac{V_{p_\Omega}(\Omega) V_{p_\Pi}((\mathcal{B}_1 \cap \mathcal{B}_2) \cap \Pi_{O_1})}{V_{p_\Omega}(\Omega) V_{p_\Pi}((\mathcal{B}_1 \cup \mathcal{B}_2) \cap \Pi_{O_1})} = \frac{V_{p_\Pi}(\mathcal{B}_1^{p_\Pi} \cap \mathcal{B}_2^{p_\Pi})}{V_{p_\Pi}(\mathcal{B}_1^{p_\Pi} \cap \mathcal{B}_2^{p_\Pi})} = q_{p_\Pi},\end{aligned}$$

where $\mathcal{B}_1^{p_\Pi}, \mathcal{B}_2^{p_\Pi}$ are the balls in $I\!R^{p_\Pi}$ with radii R and distance between centers $|O_1O_2|$. From equation (2.5) it follows: $d_\Pi < p \Rightarrow q_{p_\Pi} > q_p$. □

Appendix B Fixing the sequence h_k

A sequence h_k ensuring

$$n(X_i, h_{k+1}) \leq a\, n(X_i, h_k),\ h_{k+1} \leq b\, h_k, \tag{B.1}$$

can be fixed as follows. Let us collect for each point X_i the distances $h_\ell(X_i)$ between X_i and its n_ℓ-s neighbor, $\ell = 1, \ldots, M$. In the homogeneous case, all $h_\ell(X_i)$ for a fixed ℓ and different i are of the same order. However, one can often observe a high variability

of such radii in the inhomogeneous situation. Let a set $\{h^*_\ell, \ell \geq 0\}$ be obtained by putting all series $\{h_\ell(X_i),\, \ell = 0, 1, \ldots, K\}$ together in the increasing order. We will select the radii h_k sequentially from this set to ensure the condition (B.1). Set $h_0 = h^*_0$. Equivalently, h_0 is the smallest radius among all $h_0(X_i)$. Then select the largest index ℓ_1 such that

$$\max_i \frac{n(X_i, h^*_{\ell_1})}{n(X_i, h_0)} \leq a$$

and set $h_1 = h^*_{\ell_1}$. The construction of sequences $\{h_\ell\}$ ensures that such $\ell_1 > 1$ exists. Continue in this way. If $h_k = h_{\ell_k}$ is the radius selected at step k, then the next radius h_{k+1} is selected using the largest index $\ell_{k+1} > \ell_k$ such that $h_{k+1} = h^*_{\ell_{k+1}}$ ensures the condition. Stop when h_k reaches the largest possible value h_K. The condition (B.1) can be weaken by just controlling the fraction of points for which the inequality (B.1) can be violated.

Appendix C Other clustering procedures

Here we briefly describe the details how the concurring procedures were implemented. Each clustering method used in the evaluation requires to fix some tuning parameter(s). For each method we optimized the choice for its every parameter by taking the best result over evenly spaced values from the prespecified range. For *k-means clustering* the best result is chosen from 100 algorithm runs for each $k : 1 \leq k \leq 3K$, where K is the true number of clusters taken from the data.

DBSCAN Ester et al. (1996) takes *eps* and *minsp* as the parameter combination to determine dense points, where *eps* is the maximum distance between two samples to be considered in the same neighborhood, and *minsp* is minimum number of points required to form a dense region. For the best result of DBSCAN we evaluated over $eps \in [mindist, maxdist]$ and $minsp \in [1, N]$, where $maxidst(mindist)$ is the maximum(minimum) pairwise distance between the data elements and N is the data set size. DBSCAN can identify points as noise which are colored black on figures. The noise is considered as a separate cluster.

Spectral clustering constructs affinity matrix using either kernel function such the Gaussian (RBF) kernel or a k-nearest neighbors connectivity matrix Zelnik-Manor and Perona (2004). For the first case, the scaling factor σ and *degree* of RBF kernel are tuned by varying over $\sigma \in [mindist, maxdist]$ and *degree* up to 4. For the second case, the parameter for number of neighbors $n \in [1, N]$ is tuned. For each parameter value the best result from 100 runs with random initialization is taken. As a final result the best output of all cases is taken.

Affinity propagation has two parameters for tuning: dumping factor D from [0.5, 1] and preferences P for each point to be chosen as exemplars Frey and Dueck (2007). We set P to a global shared value varying from minimum to maximum value of pairwise similarities (negative Euclidean distance) between data points. The adjustment of these parameters was rather difficult because of high sensitivity of the results to the parameter choice.

Appendix D Examples on separation ability of AWC

Here we consider the case of two dense clusters A and C separated by an area of lower density B. Explicitly we consider a rectangle with sizes 2×3 and three area inside it presented on the Figure D.1. The left and right areas have the same density p, while the area in between has density $f_\varepsilon = (1-\varepsilon)f$, $\varepsilon \in [0,1]$. An example of such generated data with $\varepsilon = 0.3$, $n = 1000$ is shown on the middle plot of the Figure D.1; in the last plot true clusters are labeled by colors.

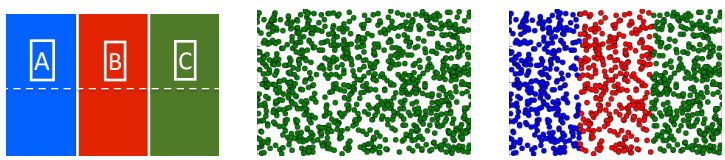

Figure D.1: From left: clusters' areas, realization, true clustering. $\varepsilon = 0.3, n = 1000$

We expect that AWC separates the left and right clusters. In this experiment we are not interested in the behavior of AWC in between. To measure how good AWC separates the clusters A and C we will use the separation error e_s:

$$e_s = \frac{\sum\limits_{i \neq j} |\widehat{w}_{ij}| \, 1\!\!1_{(w^*_{ij}=0)} \, 1\!\!1_{(i,j \in A \cup C)}}{\sum\limits_{i \neq j} 1\!\!1_{(w^*_{ij}=0)} \, 1\!\!1_{(i,j \in A \cup C)}},$$

where w^*_{ij} are true weights and $\widehat{w}_{ij}$ are answer weights of AWC.

Let us fix the overall number of points n and the parameter ε. After running 200 experiments we can calculate the average separation error $e_s(n, \varepsilon)$. For all experiments we count which part of them has error $e_s > 0.1$. For each n the probability having separation error $e_s > 0.1$ as a function of ε is shown on the right plot of Figure D.2. On the left plot of Figure D.2 we show for each number of points n what difference in density ε we can detect such that it guaranties probability of error level $e_s > 0.1$ being less than 0.1. E.g. for $n = 1000$ the value $\varepsilon = 0.47$ guaranties that the probability of $e_s > 0.1$ is less than 0.1. For each n the parameter λ was chosen to have average propagation error e_p equal to 0.1. Hereby run the procedure on data with n points and

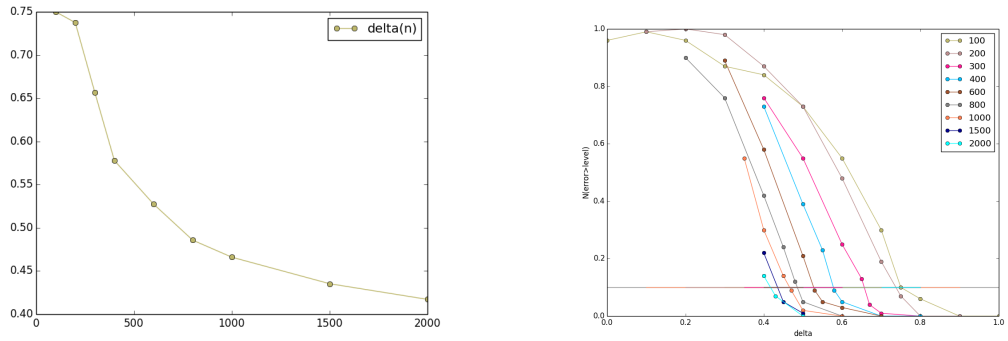

Figure D.2: The smallest density gap $\varepsilon(n)$ yielding $I\!P(e_s > 0.1) \leq 0.1$ for different n and $I\!P(e_s > 0.1)$ for different n and the gap coefficient ε.

$\varepsilon = 0$ and take the minimum λ with propagation error $e_p = 0.1$. True clustering in this case is one cluster containing all points.

Appendix E Experiments on Real World datasets

The quality of the method depends on its separation and propagation ability. The separation quality of the method can be regarded as its ability to separate different clusters. The propagation quality is considered as its ability to aggregate points from the same cluster. As the method is formulated in terms of weights, it is natural to measure these two types of misweighting error via the final computed weights $\widehat{w}_{ij}$: e_s counts all connections (positive weights) between points from different clusters, while e_p indicates the number of disconnecting points in the same cluster:

$$e_s = \frac{\sum\limits_{i \neq j} |\widehat{w}_{ij}| \, 1\!\!1_{(w^*_{ij}=0)}}{\sum\limits_{i \neq j} 1\!\!1_{(w^*_{ij}=0)}}, \qquad e_p = \frac{\sum\limits_{i \neq j} |1 - \widehat{w}_{ij}| \, 1\!\!1_{(w^*_{ij}=1)}}{\sum\limits_{i \neq j} 1\!\!1_{(w^*_{ij}=1)}}, \tag{E.1}$$

where w^*_{ij} denote the true weights describing the underlying clustering structure. The e_p and e_s are just weighted parts of the well known metric for cluster analysis comparison called *Rand index* R Rand (1971). In our notation rand index can be represented as

$$R = 1 - \frac{\sum\limits_{i \neq j} |\widehat{w}_{ij}| \, 1\!\!1_{(w^*_{ij}=0)} + \sum\limits_{i \neq j} |1 - \widehat{w}_{ij}| \, 1\!\!1_{(w^*_{ij}=1)}}{\sum\limits_{i \neq j} 1\!\!1_{(w^*_{ij}=0)} + \sum\limits_{i \neq j} 1\!\!1_{(w^*_{ij}=1)}} \stackrel{\text{def}}{=} 1 - e.$$

Further we will use the *general error* $e \stackrel{\text{def}}{=} 1 - R$ instead of Rand index.

Now consider the behavior of the algorithms on real world data. The data sets are taken from UCI repository Lichman (2013) , except the *Olive* data; see http://www2.chemie.uni-erlangen.de/publications/ANN-book/datasets/oliveoil/. *Iris* data

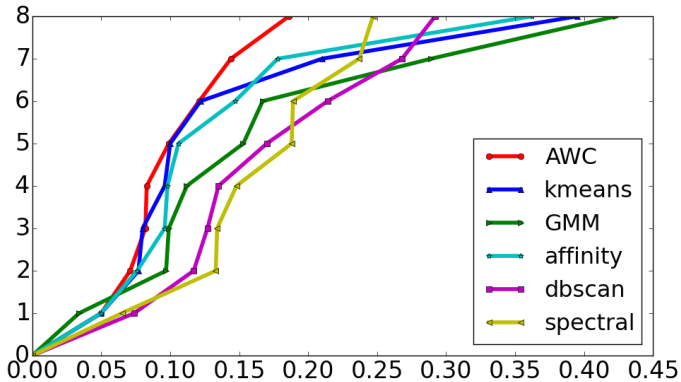

Figure E.1: Comparison on real datasets

set contains 3 type of iris plants. *Wine* data is the result of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. *Seeds* comprise kernels belonging to three different varieties of wheat. *Thyroid* is clinical data used to predict patients thyroid functional state. *Ecoli* data is used for classification of the cellular localization sites of proteins. *Olive* is a group of olive oil samples from nine different regions of Italy. *Wisconsin* stands for Wisconsin Breast Cancer Database designated whether samples are benign or malignant. *Banknote* data set was extracted from images that were taken from genuine and forged banknote-like specimens. The data set sizes n, number of attributes d and clusters K are listed in Table 2.

Similarly to the experiments on artificial data, each algorithm was set with its best parameter configuration minimizing the general error e. Algorithms performances are listed in Table 2. Here for every algorithm only general error e is presented. The graphical interpretation of the Table 2 is shown on Figure E.1. Here x-axis represents the error level and y-axis shows the number of databases. For each clustering algorithm we construct its plot as the function showing for any error threshold a number of databases where the error level is below this threshold. Thus each line is non-decreasing function and the best algorithm is the one lying on the left. One can see that AWC demonstrates the best performance on the majority of databases. The value of the sum of weights statistic $S(\lambda)$ is shown on Figure E.2.

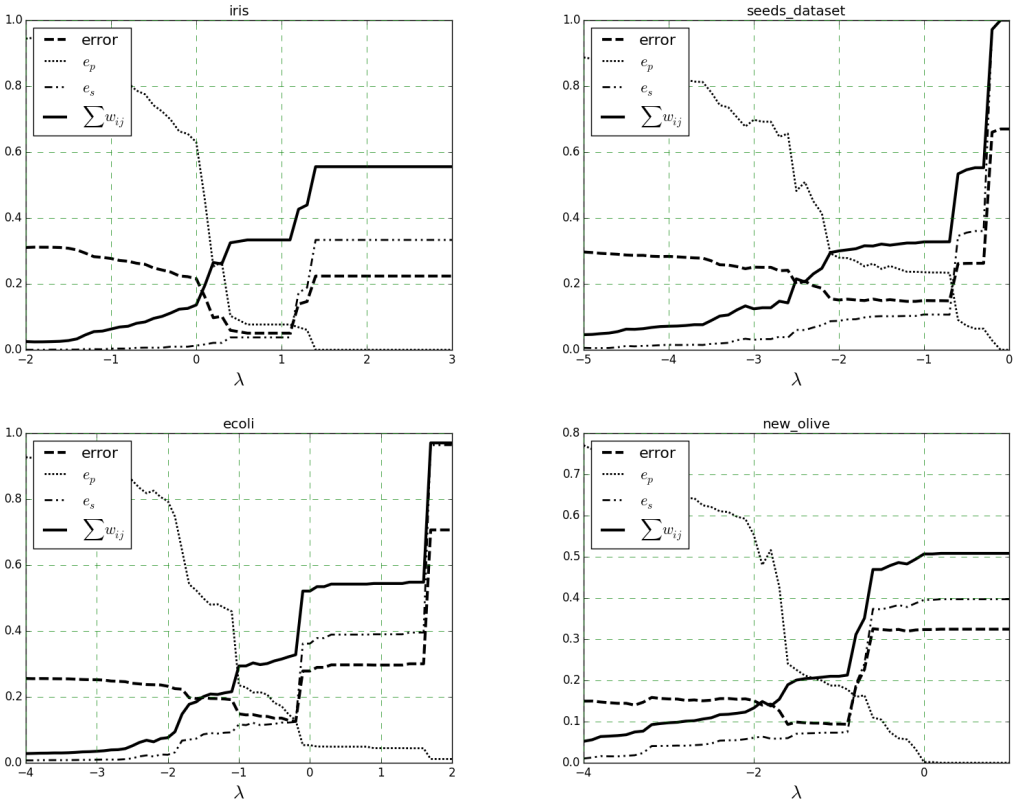

Figure E.2: $S(\lambda)$ Iris (Top-link), Seeds (top-right), Ecoli (bottom-link), Olive(bottom-right)

Table 2: Real world data sets error e for each method, the best two results are in bold

	Algorithm									
Data	AWC	AWC $_{sow}$	k-m	GMM	Affinity	DBSCAN	Spectral	n	d	K
Iris	0.05	0.05	0.050	**0.034**	0.05	0.117	0.188	150	4	3
Wine	**0.101**	0.132	**0.096**	**0.099**	**0.096**	0.268	0.189	178	13	3
Seeds	**0.148**	**0.148**	0.21	0.289	0.178	0.292	**0.148**	210	7	3
Thyroid	**0.089**	**0.089**	**0.08**	0.097	0.147	0.135	0.247	215	5	3
Ecoli	0.125	0.125	0.122	0.167	**0.106**	0.17	0.134	336	7	8
Olive	0.093	0.093	0.1	0.153	0.076	0.127	**0.065**	572	8	9
Wisconsin	**0.067**	**0.070**	**0.077**	0.112	0.098	**0.074**	0.133	699	9	2
Banknote	**0.193**	**0.194**	0.395	0.423	0.362	0.214	0.237	1372	4	2

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

001 "Data Driven Value-at-Risk Forecasting using a SVR-GARCH-KDE Hybrid" by Marius Lux, Wolfgang Karl Härdle and Stefan Lessmann, January 2018.

002 "Nonparametric Variable Selection and Its Application to Additive Models" by Zheng-Hui Feng, Lu Lin, Ruo-Qing Zhu asnd Li-Xing Zhu, January 2018.

003 "Systemic Risk in Global Volatility Spillover Networks: Evidence from Option-implied Volatility Indices " by Zihui Yang and Yinggang Zhou, January 2018.

004 "Pricing Cryptocurrency options: the case of CRIX and Bitcoin" by Cathy YH Chen, Wolfgang Karl Härdle, Ai Jun Hou and Weining Wang, January 2018.

005 "Testing for bubbles in cryptocurrencies with time-varying volatility" by Christian M. Hafner, January 2018.

006 "A Note on Cryptocurrencies and Currency Competition" by Anna Almosova, January 2018.

007 "Knowing me, knowing you: inventor mobility and the formation of technology-oriented alliances" by Stefan Wagner and Martin C. Goossen, February 2018.

008 "A Monetary Model of Blockchain" by Anna Almosova, February 2018.

009 "Deregulated day-ahead electricity markets in Southeast Europe: Price forecasting and comparative structural analysis" by Antanina Hryshchuk, Stefan Lessmann, February 2018.

010 "How Sensitive are Tail-related Risk Measures in a Contamination Neighbourhood?" by Wolfgang Karl Härdle, Chengxiu Ling, February 2018.

011 "How to Measure a Performance of a Collaborative Research Centre" by Alona Zharova, Janine Tellinger-Rice, Wolfgang Karl Härdle, February 2018.

012 "Targeting customers for profit: An ensemble learning framework to support marketing decision making" by Stefan Lessmann, Kristof Coussement, Koen W. De Bock, Johannes Haupt, February 2018.

013 "Improving Crime Count Forecasts Using Twitter and Taxi Data" by Lara Vomfell, Wolfgang Karl Härdle, Stefan Lessmann, February 2018.

014 "Price Discovery on Bitcoin Markets" by Paolo Pagnottoni, Dirk G. Baur, Thomas Dimpfl, March 2018.

015 "Bitcoin is not the New Gold - A Comparison of Volatility, Correlation, and Portfolio Performance" by Tony Klein, Hien Pham Thu, Thomas Walther, March 2018.

016 "Time-varying Limit Order Book Networks" by Wolfgang Karl Härdle, Shi Chen, Chong Liang, Melanie Schienle, April 2018.

017 "Regularization Approach for NetworkModeling of German EnergyMarket" by Shi Chen, Wolfgang Karl Härdle, Brenda López Cabrera, May 2018.

018 "Adaptive Nonparametric Clustering" by Kirill Efimov, Larisa Adamyan, Vladimir Spokoiny, May 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.