Weierstraß-Institut für
Angewandte Analysis und Stochastik

Leibniz-Gemeinschaft

Clustering using adaptive weights

Vladimir Spokoiny, WIAS and HU (Berlin)
joint with Kirill Efimov and Larisa Hax

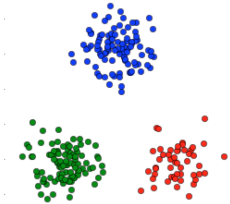
Outline

Data $X_1, \ldots, X_n \in \mathbb{R}^d$.

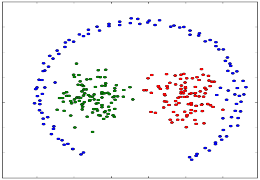
Aim: split into homogeneous groups (clusters).

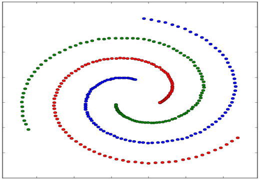
Number and structure/shape of clusters usually unknown.

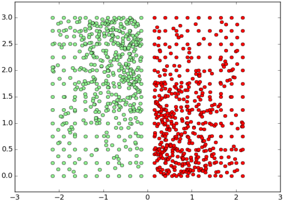
Ideal picture:

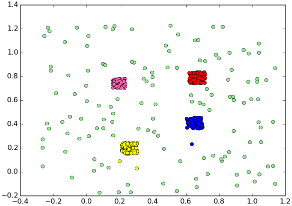


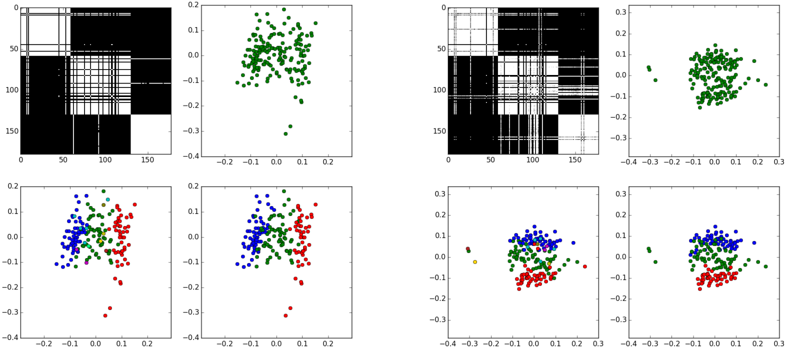
What is a cluster in general?






3.0
2.5
2.0
1.5
1.0
0.5
0.0
−3
−2
−1
0
1
2
3


1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2


0
50
100
150
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.2 −0.1 0.0 0.1 0.2
0.3
−0.4 −0.3 −0.2 −0.1 0.0 0.1 0.2 0.3

- Partitional clustering (k-mean) [MacQueen et al., 1967]. Minimizing the objective function over partitions. Require to fix the number of clusters, hard to implement; cannot deal with non-spherical clusters
- Hierarchical: agglomerative (bottom-up) and divisive (top-down). Irreversibility of the merge decision;
- Density based: cluster = mode of the density, [Ester et al., 1996]. Poor quality of density estimation if $d > 2$;
- Spectral: dimensionality reduction by eigenvalue decomposition of the adjucency matrix; [Ng et al., 2002]. Require a good separation between clusters – spectral gap;
- Affinity propagation: dynamic graphical models by responsibility and availability for each two points; [Frey and Dueck, 2007]. unstable, sensitive to parameter choice.

Aim: an efficient procedure which adapts to unknown cluster structure.

Approach: Describe the cluster structure by an adjacency matrix $W = (w_{ij})$, each w_{ij} means the probability that X_i and X_j are in the same cluster.

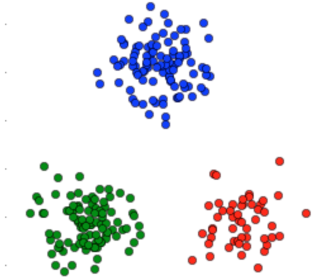
For the standard (partitioned) clustering, W is a block matrix:

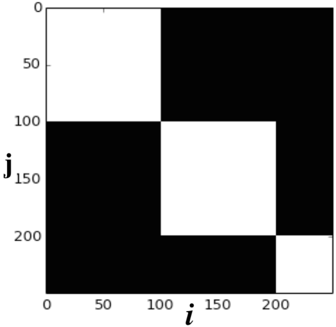
$$w_{ij} = \begin{cases} 1 & i, j \text{ from the same cluster,} \\ 0, & \text{otherwise} \end{cases}$$

The matrix W is recovered from the data by an iterative procedure:

- Initialize with one cluster $\mathcal{C}_i^{(0)}$ per point X_i;
- At each step, increase the locality parameter h_k and recompute the local weights $w_{ij}^{(k)}$ using a statistical test that there is no gap between two local clusters $\mathcal{C}_i^{(k-1)}$ and $\mathcal{C}_j^{(k-1)}$.
- Stop when the bandwidth h_k reaches the global value.

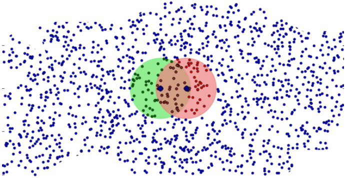
Let $\{X_1, \ldots, X_n\} \subset I\!R^d$ with $d < n$ be the set of all samples X_i.
Example: 250 points, 3 normal clusters (100 + 100 + 50) and the corresponding matrix of weights W.



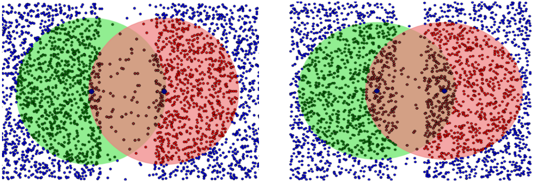


Relaxation: allow a general symmetric $n \times n$ matrix of weights $W = (w_{ij})_{i,j=1,\ldots,n}$ with $w_{ij} \in [0, 1]$.

Homogeneous case:



“Gap” case:



After $k-1$ steps, for each $i \le n$, the cluster $\mathcal{C}_i^{(k-1)}$ is given via weights $w_{ij}^{(k-1)}$, $j \le n$.

At step k, suppose the locality parameter h_k to be fixed and consider any pair (X_i, X_j) with $\|X_i - X_j\| \le h_k$.

Problem: For two local clusters $\mathcal{C}_i^{(k-1)}$ and $\mathcal{C}_j^{(k-1)}$ with $\|X_i - X_j\| \le h_k$, compute the value $w_{ij}^{(k)}$ reflecting the gap between $\mathcal{C}_i^{(k-1)}$ and $\mathcal{C}_j^{(k-1)}$.

Principal idea: check the data density in the overlap $\mathcal{C}_i^{(k-1)} \cap \mathcal{C}_j^{(k-1)}$.

Mass of the overlap $N_{i\wedge j}^{(k)}$:

$$N_{i\wedge j}^{(k)} \stackrel{\text{def}}{=} \sum_{l\neq i,j} w_{il}^{(k-1)} w_{jl}^{(k-1)} \approx \text{\# points in } \mathcal{B}(X_i, h_k) \cap \mathcal{B}(X_j, h_k)$$

Mass of the union $N_{i\vee j}^{(k)}$:

$$N_{i\vee j}^{(k)} \stackrel{\text{def}}{=} N_{i\wedge j}^{(k)} + N_{i\triangle j}^{(k)} \approx \text{\# points in } \mathcal{B}(X_i, h_k) \cup \mathcal{B}(X_j, h_k)$$

where $N_{i\triangle j}^{(k)}$ is the mass of the complementary parts:

$$N_{i\triangle j}^{(k)} \stackrel{\text{def}}{=} \sum_{l\neq i,j:\{\|X_i-X_l\|\leq h_{k-1}\}\triangle\{\|X_j-X_l\|\leq h_{k-1}\}} \left(w_{il}^{(k-1)} + w_{jl}^{(k-1)} \right) .$$

Estimated relative density in the overlap:

$$\widetilde{\theta}^{(k)}_{i\wedge j} = \frac{N^{(k)}_{i\wedge j}}{N^{(k)}_{i\vee j}}$$

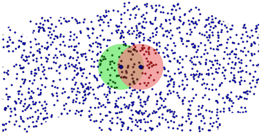
Local homogeneous case corresponds to the nearly uniform distribution:

$$\widetilde{\theta}^{(k)}_{i\wedge j} \approx q^{(k)}_{ij} \stackrel{\text{def}}{=} \frac{\mathrm{Vol}_{\cap}(d_{ij}, h_k)}{2\,\mathrm{Vol}(h_k) - \mathrm{Vol}_{\cap}(d_{ij}, h_k)},$$

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 radii h and the distance d between centers, $d_{ij} = \|X_i - X_j\|$.

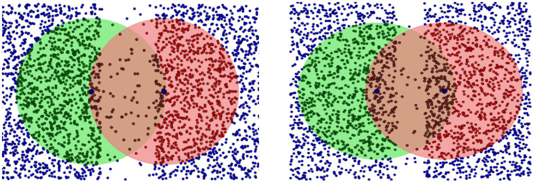
Null (no gap): $\theta^{(k)}_{i\wedge j} = q^{(k)}_{ij}$ vs alternative (a gap) $\theta^{(k)}_{i\wedge j} < q^{(k)}_{ij}$.

Homogeneous case:



Brown area: overlap of two clusters, green and pink - complements.

“Gap” case:



The value $q_{ij}^{(k)}$ depends only on the ratio $t_{ij}^{(k)} = d_{ij}/h_k$ and can be calculated explicitly: $q_{ij}^{(k)} = q(t_{ij}^{(k)})$ with

$$q(t) = 2\frac{B(d+\frac{1}{2},\frac{1}{2})}{B(1-\frac{t}{2},d+\frac{1}{2},\frac{1}{2})} - 1,$$

where $B(a,b)$ is the beta-function, $B(x,a,b)$ is the incomplete beta-function, and d is the space dimension.
We need to test if $\widetilde{\theta}_{i\wedge j}^{(k)} < q_{ij}^{(k)}$. Following to [Polzehl and Spokoiny, 2006], define the test statistic $T_{ij}^{(k)}$

$$T_{ij}^{(k)} = N_{i\vee j}^{(k)}\,\mathcal{K}(\widetilde{\theta}_{i\wedge j}^{(k)}, q_{ij}^{(k)})\big\{\mathbb{I}(\widetilde{\theta}_{i\wedge j}^{(k)} < q_{ij}^{(k)}) - \mathbb{I}(\widetilde{\theta}_{i\wedge j}^{(k)} > q_{ij}^{(k)})\big\},$$

where $\mathcal{K}(\theta,q)$ is the symmetrized Kullback-Leibler divergence:

$$\mathcal{K}(\theta,q) = (\theta-q)\log\frac{\theta(1-q)}{q(1-\theta)}.$$

Parameters:

- A sequence of radii h_k. Fixed from the data to ensure that each ball $\mathcal{B}(X_i, h_k)$ contains nearly $n_k \approx (2d+1)a^k$ points for $a = 2^{1/4}$ and $k = 1, \ldots, K$.
- A parameter λ.
- Localizing kernel $K_{\mathrm{loc}}(u)$; (Default choice – a uniform kernel $K_{\mathrm{loc}}(u) = \mathbb{I}(u \leq 1)$);
- Statistical kernel $K_{\mathrm{stat}}(u)$ (Default choice – a uniform kernel);

Initialization: $k = 0$, for each i and j

$$w_{ij}^{(0)} = K_{\mathrm{loc}}\left(\frac{\|X_i - X_j\|}{h_0}\right).$$

Increase k, recompute

$$w_{ij}^{(k)} = K_{\mathrm{loc}}\left(\frac{\|X_i - X_j\|}{h_k}\right) K_{\mathrm{stat}}\left(\frac{T_{ij}^{(k)}}{\lambda}\right).$$

where

$$T_{ij}^{(k)} = N_{i\vee j}^{(k)} \mathcal{K}\bigl(\widetilde{\theta}_{i\wedge j}^{(k)}, q_{ij}^{(k)}\bigr)\, \mathbb{I}\bigl(\widetilde{\theta}_{i\wedge j}^{(k)} < q_{ij}^{(k)}\bigr)$$

for

$$\widetilde{\theta}_{i\wedge j}^{(k)} = \frac{N_{i\wedge j}^{(k)}}{N_{i\vee j}^{(k)}}.$$

The parameter λ is fixed as the minimal value to ensure that for an artificial sample with one cluster, the procedure ends up with homogeneous weights $w_{ij}^{(K)} = 1$.

Alternatively one can run the procedure with different λ and select one by checking an increase of the sum of weights $\sum_{i,j} w_{ij}^{(K)}$.

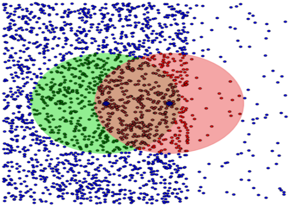
The final clustering decision is made from the weights $w_{ij}^{(K)}$ computed at the last step K.

Propagation: If X_i and X_j are within a homogeneous (spherical) region, then the construction ensures $w_{ij}^{(K)} = 1$.

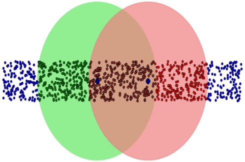
Propagation continues to apply even for many homogeneous regions: $w_{ij}^{(K)} = 1$ for any pairs (X_i, X_j) from the same region.

If $X_i \in \mathcal{C}_i$ and $\mathcal{C}_i$ is separated from all other clusters with a significant gap, then $w_{ij}^K = 0$ for any $X_j \notin \mathcal{C}_i$.

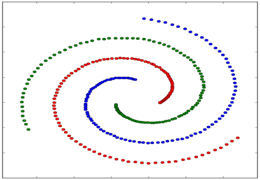
AWC provides the optimal separation rate (minimal margins between clusters) for two or more dense convex (Gaussian like) clusters. see demo



AWC detects automatically sharp edges with a slight gravitation effect: neighbor points are gravitated to (included into) dense clusters.



The propagation property works well along a low dimensional manifold.



The complexity is (almost) dimension free and can be upper bounded by C $n\, n_K^2$, where n_K is the number of screened neighbors of each point X_i at the last step.

For small datasets ($n \leq 2000$) we use $n_K = n$. Then complexity of order n^3 .

For larger n , the value n_K can be bounded to control the total complexity of the procedure.

Outline

Misweighting error via the final computed weights $w_{ij}^{(K)}$: 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:

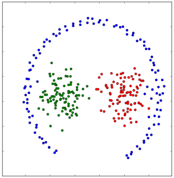
$$e_s = \frac{\sum_{i \neq j} |\widehat{w}_{ij}| \, \mathbb{I}\big(w_{ij}^* = 0\big)}{\sum_{i \neq j} \mathbb{I}\big(w_{ij}^* = 0\big)}, \qquad e_p = \frac{\sum_{i \neq j} |1 - \widehat{w}_{ij}| \, \mathbb{I}\big(w_{ij}^* = 1\big)}{\sum_{i \neq j} \mathbb{I}\big(w_{ij}^* = 1\big)},$$

where w_{ij}^* denote the true weights describing the underlying clustering structure.

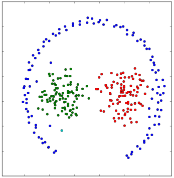
Standard *rand index* R [Rand, 1971] and total error e :

$$R = 1 - \frac{\sum_{i \neq j} |\widehat{w}_{ij}| \, \mathbb{I}\big(w_{ij}^* = 0\big) + \sum_{i \neq j} |1 - \widehat{w}_{ij}| \, \mathbb{I}\big(w_{ij}^* = 1\big)}{\sum_{i \neq j} \mathbb{I}\big(w_{ij}^* = 0\big) + \sum_{i \neq j} \mathbb{I}\big(w_{ij}^* = 1\big)} = 1 - e.$$

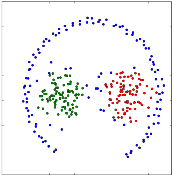
Pathbased



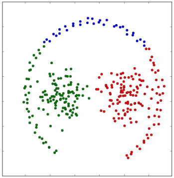
Original clustering



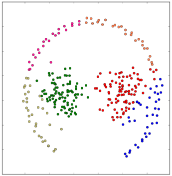
AWC, $\lambda = 5.5$



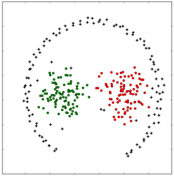
Spectral, $\sigma = 0.1$



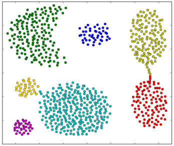
K-means, K=3



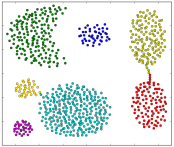
Affinity prop. D=0.5, P=-1464



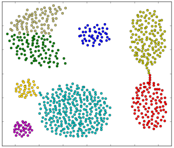
DBSCAN, e=2.1, sp=10



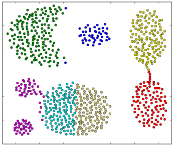
Original clustering



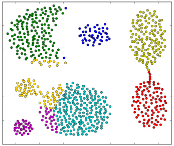
AWC, $\lambda = 4$



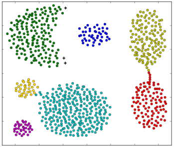
Spectral,



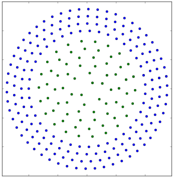
K-means,



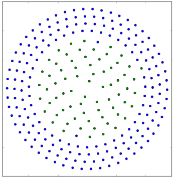
Affinity prop.



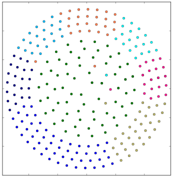
DBSCAN,



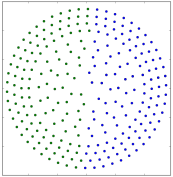
Original clustering



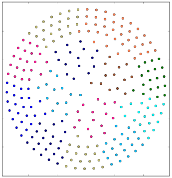
AWC, $\lambda = 2.1$



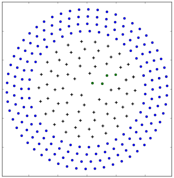
Spectral,



K-means,

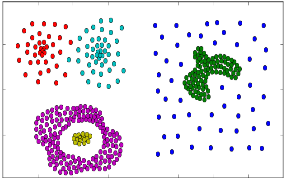


Affinity prop.

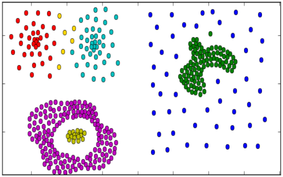


DBSCAN,

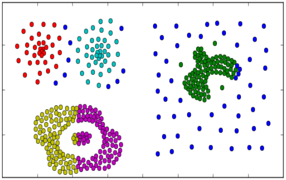
Compound



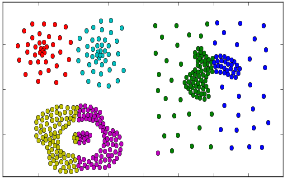
Original clustering



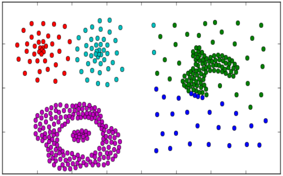
AWC, $\lambda = 3.3$



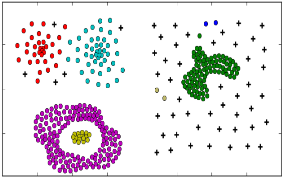
Spectral, $\sigma = 0.1$



K-means, $K = 6$



Affinity prop.



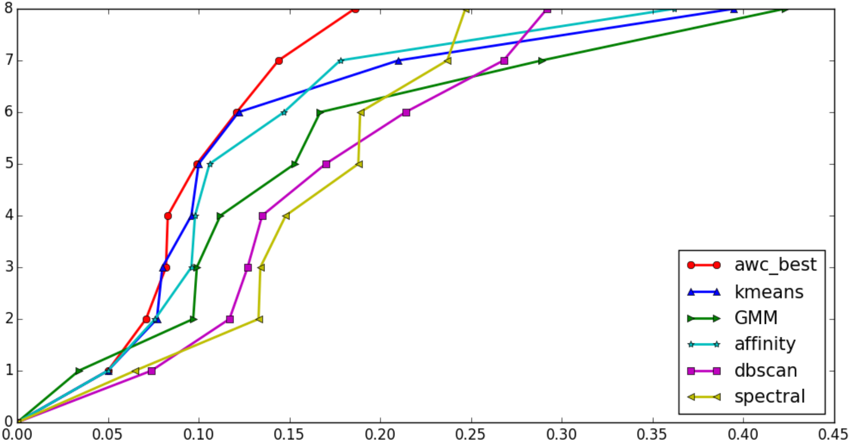
DBSCAN

The data sets are taken from UCI repository.

Data	n	d	$\#clusters$
Iris	150	4	3
Wine	178	13	3
Seeds	210	7	3
Thyroid gland	215	5	3
Ecoli	336	7	8
Olive	572	8	9
Wisconsin	699	9	2
Banknote	1372	4	2

Tabelle: Real world data sets description

Error performance



Data	Error	Algorithm						
		AWC_ best	AWC	k-means	GMM	Affinity	DBSCAN	Spectral
Iris	$e_{\cup}$	0.037	0.037	0.038	0.026	0.038	0.015	0.059
	$e_{\cap}$	0.076	0.076	0.076	0.051	0.076	0.325	0.453
	e	0.05	0.05	0.050	**0.034**	0.05	0.117	0.188
Wine	$e_{\cup}$	0.058	0.092	0.071	0.053	0.071	0.286	0.02
	$e_{\cap}$	0.181	0.191	0.145	0.189	0.145	0.233	0.519
	e	**0.099**	0.125	**0.096**	0.099	**0.096**	0.268	0.189
Seeds	$e_{\cup}$	0.093	0.11	0.164	0.237	0.135	0.199	0.037
	$e_{\cap}$	0.248	0.249	0.301	0.394	0.264	0.479	0.373
	e	**0.144**	0.156	0.21	0.289	0.178	0.292	**0.148**
Thy	$e_{\cup}$	0.08	0.081	0.074	0.127	0.101	0.174	0.151
	$e_{\cap}$	0.077	0.097	0.085	0.071	0.188	0.1	0.331
	e	**0.082**	0.09	**0.08**	0.097	0.147	0.135	0.247
Ecoli	$e_{\cup}$	0.125	0.114	0.08	0.121	0.072	0.137	0.061
	$e_{\cap}$	0.113	0.228	0.201	0.294	0.198	0.259	0.331
	e	0.121	0.145	0.122	0.167	**0.106**	0.17	0.134
Olive	$e_{\cup}$	0.076	0.076	0.097	0.152	0.063	0.052	0.062
	$e_{\cap}$	0.117	0.136	0.114	0.155	0.133	0.462	0.075
	e	0.083	0.087	0.1	0.153	0.076	0.127	**0.065**

Data	Error	Algorithm						
		AWC_ best	AWC	k-means	GMM	Affinity	DBSCAN	Spectral
Wisconsin	$e_\cup$	0.059	0.06	0.103	0.030	0.129	0.073	0.066
	$e_\cap$	0.081	0.137	0.07	0.18	0.071	0.075	0.188
	e	**0.071**	0.102	**0.077**	0.112	0.098	**0.074**	0.133
Banknote	$e_\cup$	0.001	0.001	0.107	0.437	0.094	0.01	0.082
	$e_\cap$	0.367	0.367	0.676	0.409	0.624	0.413	0.389
	e	**0.186**	**0.186**	0.395	0.423	0.362	0.214	0.237

Abbildung: $DS3, \ n = 8000$, AWC result for $\lambda = 15$

More examples

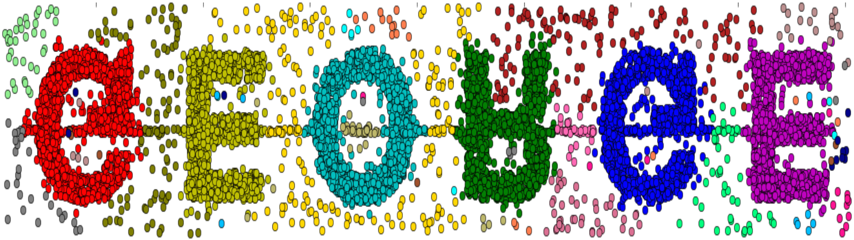


Abbildung: $n = 8000$, AWC result for $\lambda = 15$

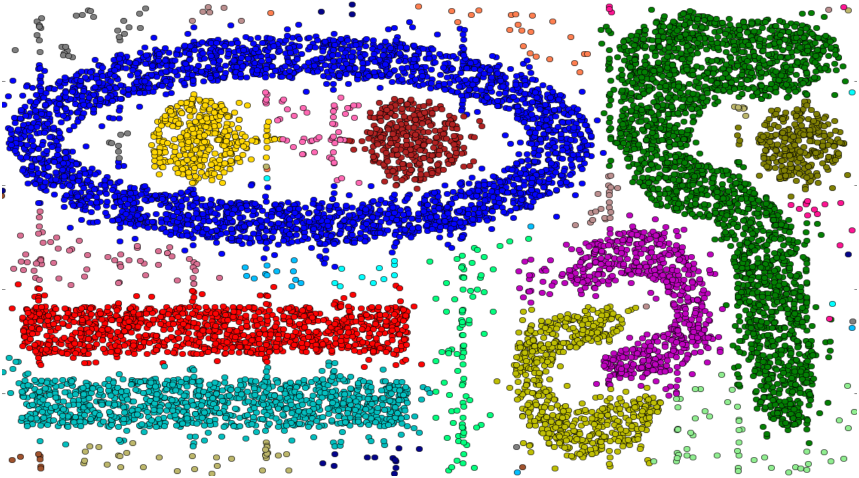


Abbildung: $DS4$, $n = 10000$ points, AWC result for $\lambda = 15$

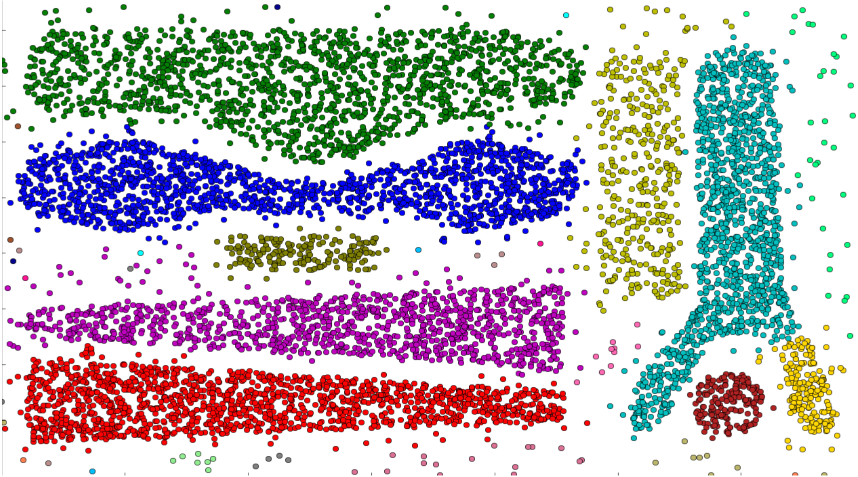


Abbildung: $DS5,\ \ n = 8000$, AWC result for $\lambda = 15$

- New approach to understand clustering using the notions “propagation” and “separation”;
- Structural adaptation using adaptive weights;
- Procedure numerically feasible and applicable even for large data sets
- Optimal separability of convex clusters;
- Procedure is fully adaptive to unknown clustering structure including the number and shape of clusters and the separation distance;
- State-of-the-art performance of a wide range of artificial and real life examples;

- Theoretical study is difficult due to iterative nature of the method. The weights $w_{ij}^{(k-1)}$ from the step $k-1$ depend from the same input data, so empirical process theory for the sums $\sum_j w_{ij}^{(k-1)}$ is not applicable.
- Many attempts to represent each step of the method as gradient decent for some optimization problem – failed so far.
- Similarly, it is unclear whether the procedure can be viewed as a EM algorithm or alternating projections;
- A rigorous theoretical justification of the method is still called for;
- The choice of the only tuning parameter λ is important and matters in complicated examples, the default choice is suboptimal.
- Semisupervised learning (combination of labelled and unlabelled data)
- High dimensional problems, combination with dimension reduction;

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 *Kdd*, volume 96, pages 226–231.

Frey, B. J. and Dueck, D. (2007).

Clustering by passing messages between data points.

science, 315(5814):972–976.

MacQueen, J. et al. (1967).

Some methods for classification and analysis of multivariate observations.

In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

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.

Polzehl, J. and Spokoiny, V. (2006).

Propagation-separation approach for local likelihood estimation.

Probability Theory and Related 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.