Introduction to Cluster Validation

Unsupervised Learning VT 2023

Class note created by Chun partially based on Sect 7.5 in "Introduction to Data Mining", 2nd Ed. by Tan, Steinbach, Karpatne & Kumar

(https://www-users.cs.umn.edu/~kumar001/dmbook/index.php)

Validation of Classification vs Clustering

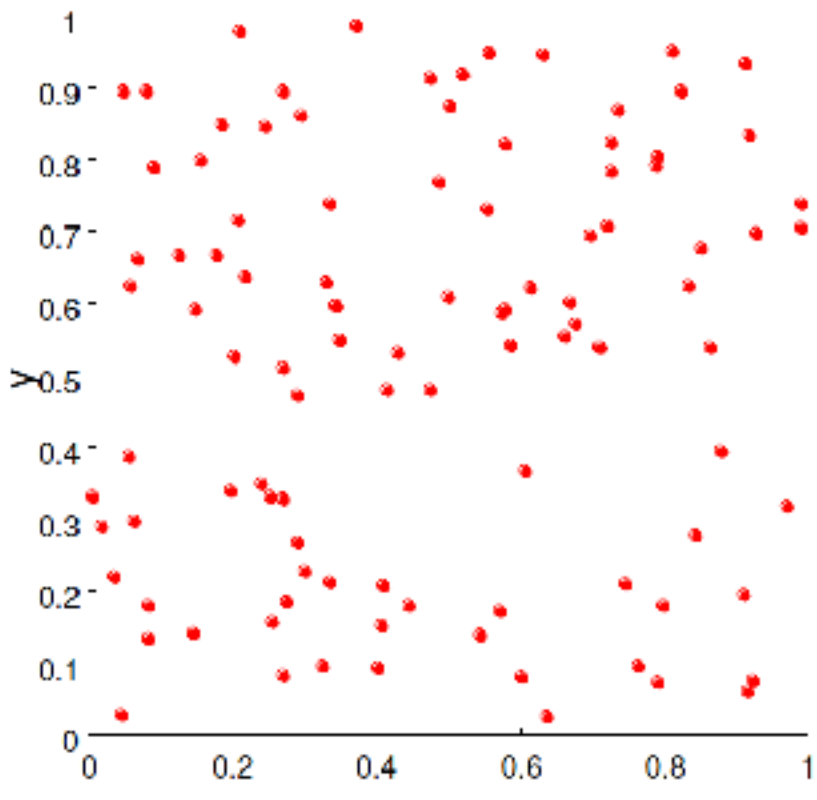
- Classification (supervised learning) comes with labels, trivial to defined classification error
- Clustering (unsupervised learning): How to evaluate the "goodness" of the resulting clusters?

Different Aspects of Cluster Validation

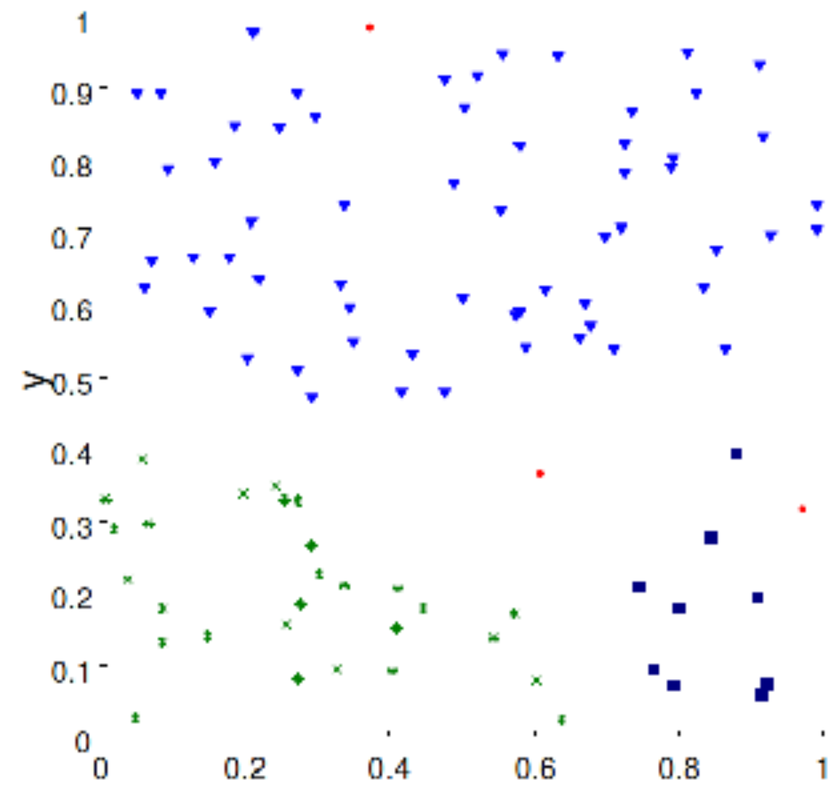
- Determine clustering tendency, i.e., if non-random structure exists, and avoid clustering noise.
- External validation: If one has externally known class labels (essentially supervised learning)
- Internal validation: Only use known internal properties of data, e.g., similarity matrix
- Comparative: Compare clusters from same methods, or clustering results from different methods
- # of clusters?

Examples of Clustering Noises

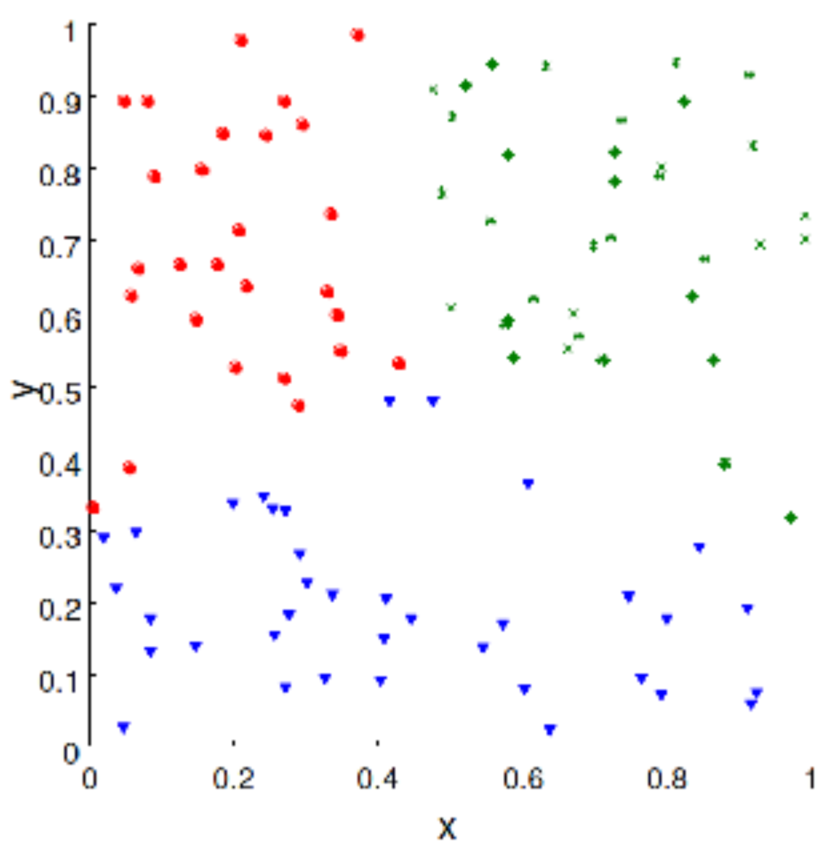
Random Points



DBSCAN

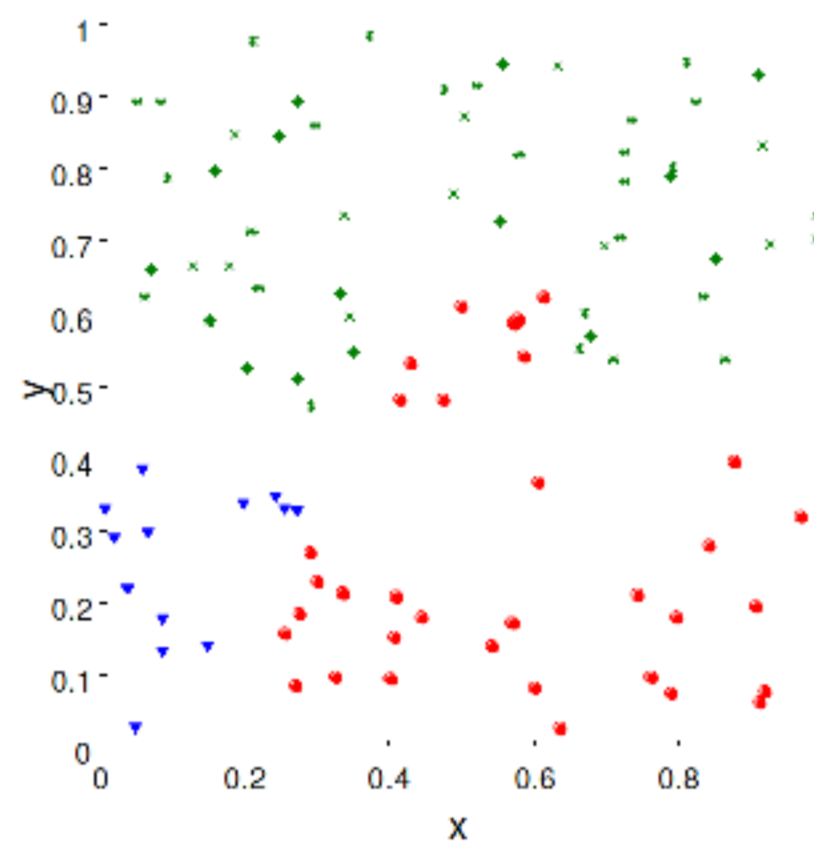


K-means



Complete Link

(Hierarchical Clustering)

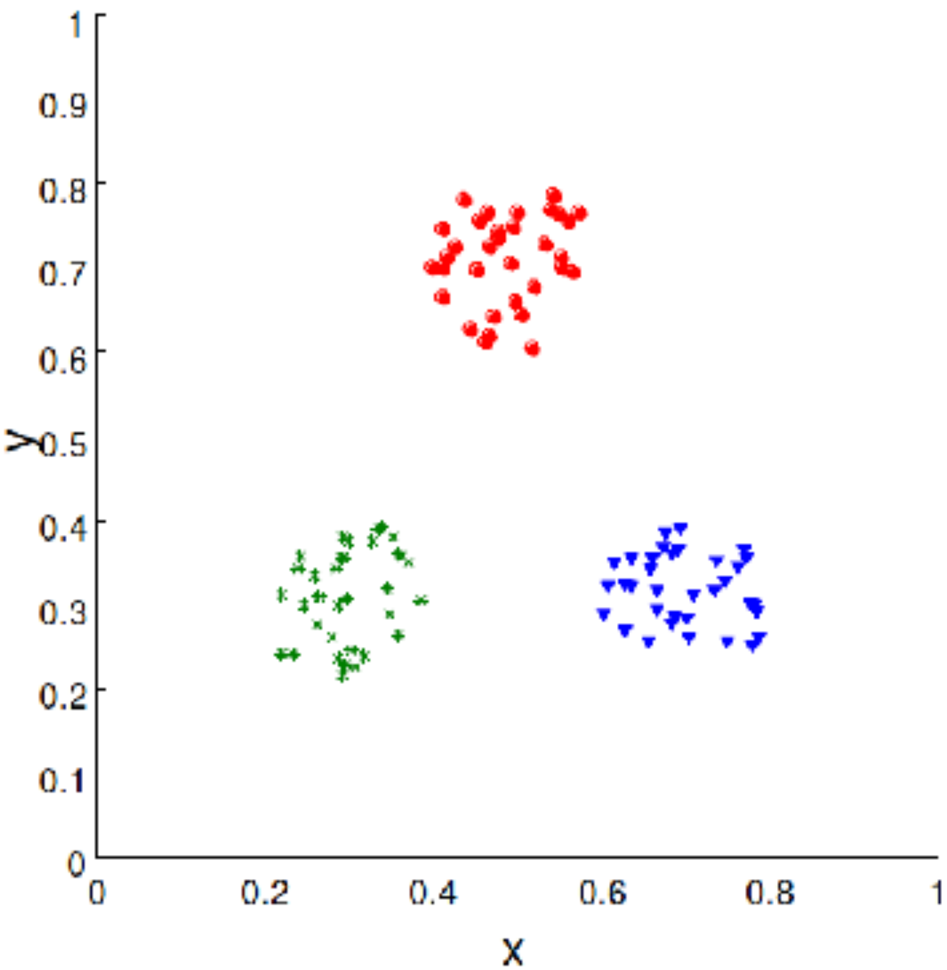


Clustering Validation by Correlation

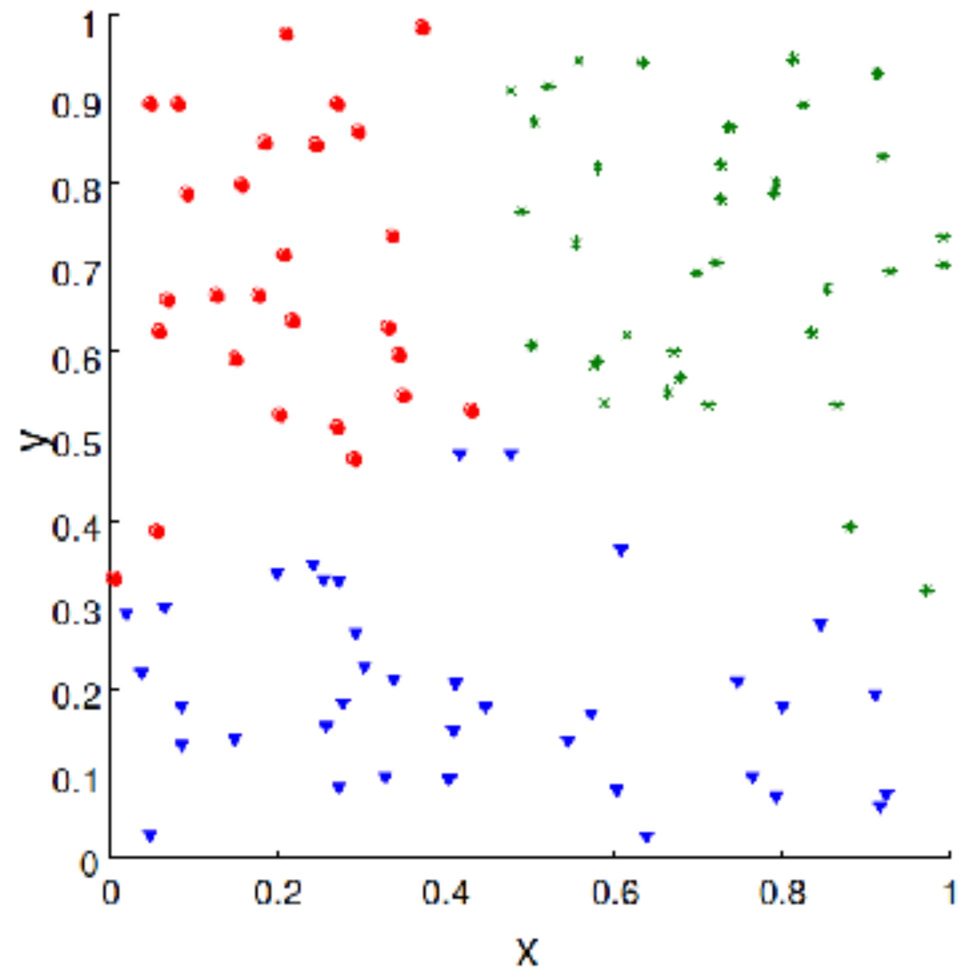
- Internal validation
- Idea: Compare two N-by-N matrices: dissimilarity matrix & Incident matrix
- Incident matrix: (i, j)-th element equals 1 if i-th & j-th data points are in same cluster; 0 otherwise
- Compute Pearson correlation coefficient between the 2, matrices. Both matrices are symmetric, only $N(N\text{-}1)/2$ entries to compare.
- Correlation close to -1 → data points in same cluster are similar
- Allow for hypothesis test
- Not good measure for results from, e.g., density based clustering (why?)

Clustering Validation by Correlation

E.g. Results from k-mean



Random points

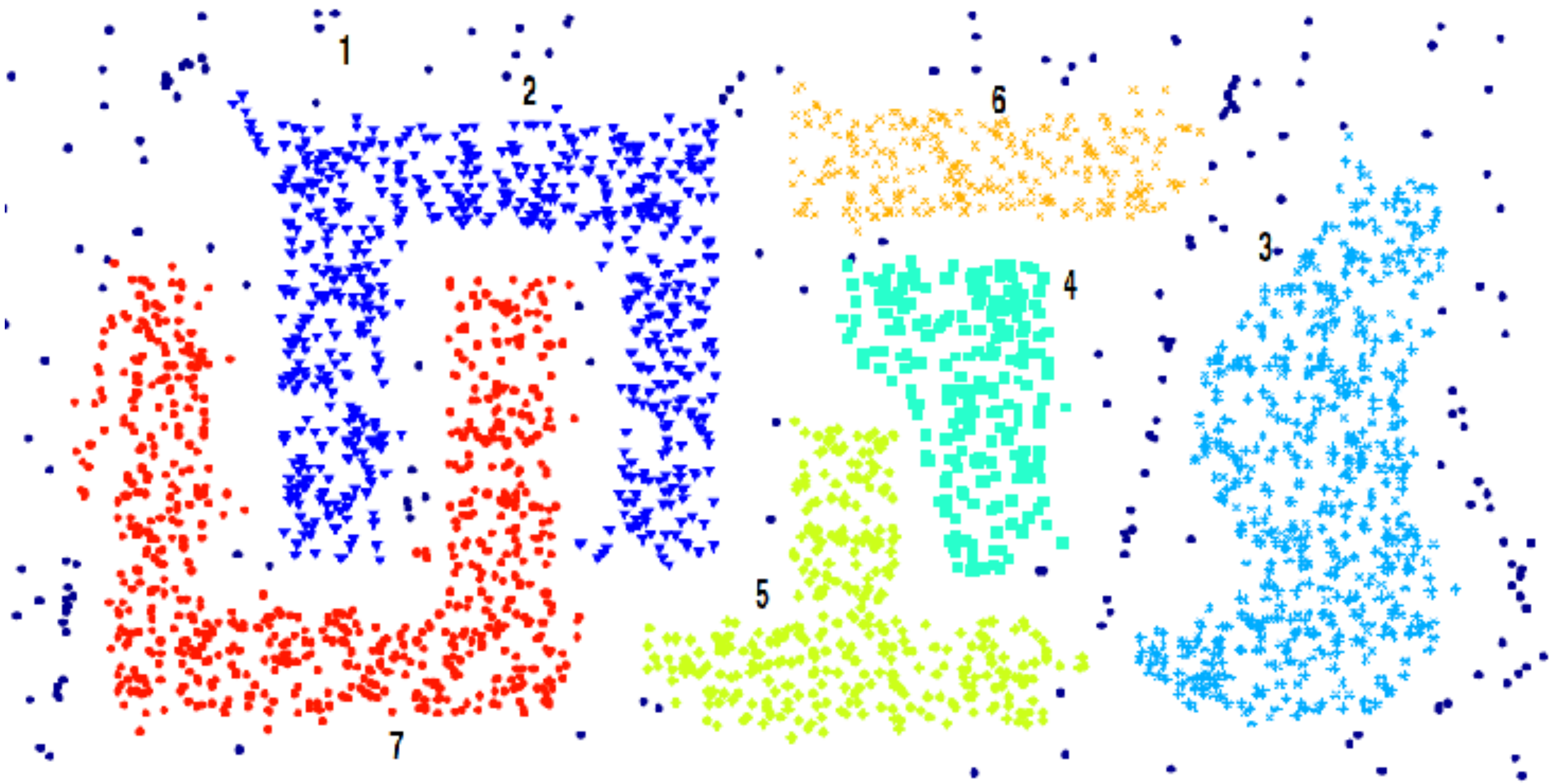


Corr = -0.9235

Corr = -0.5810

Clustering Validation by Correlation

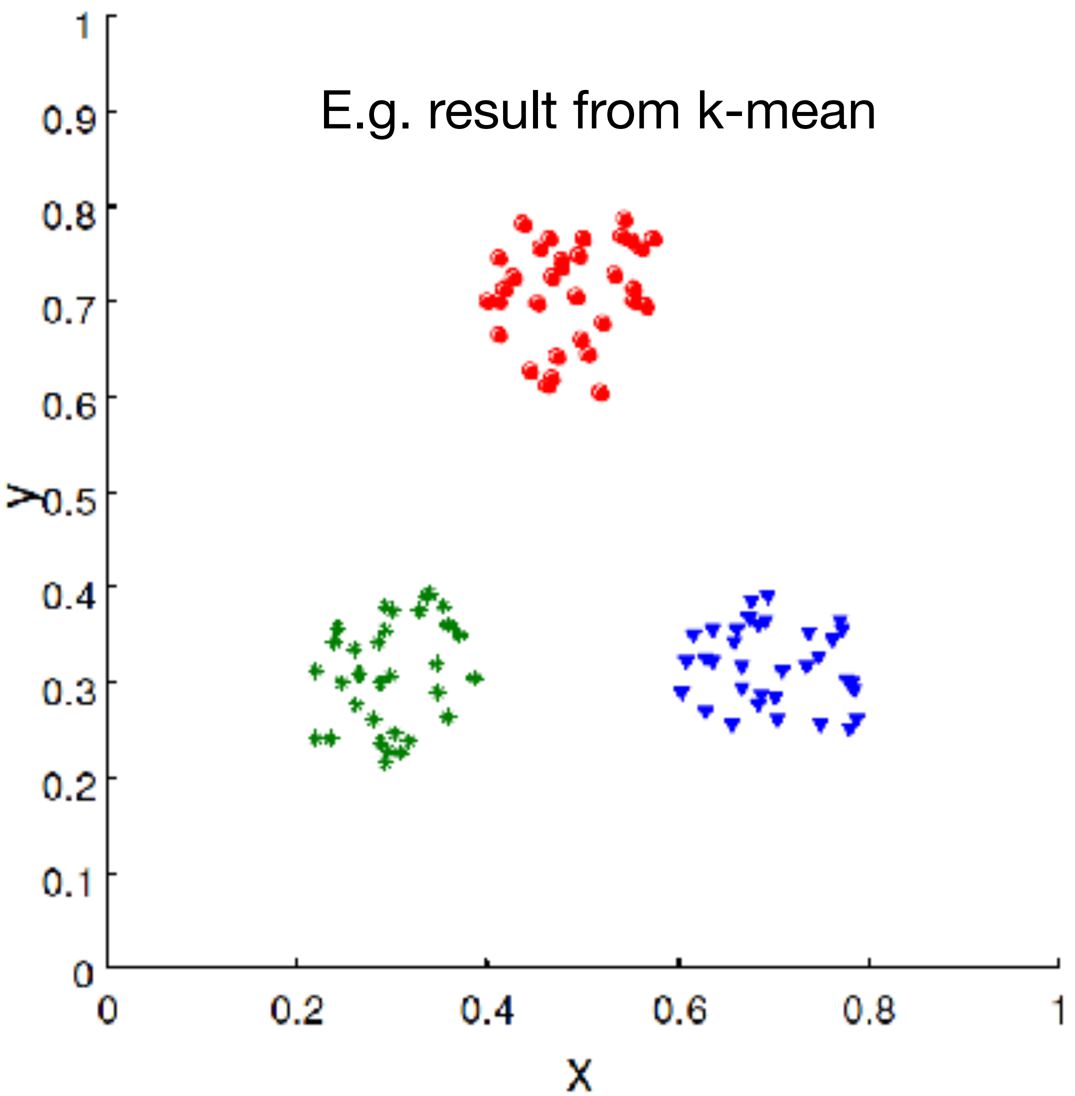
Ex: Any suggestion how to deal with results from, e.g., DBSCAN clustering?

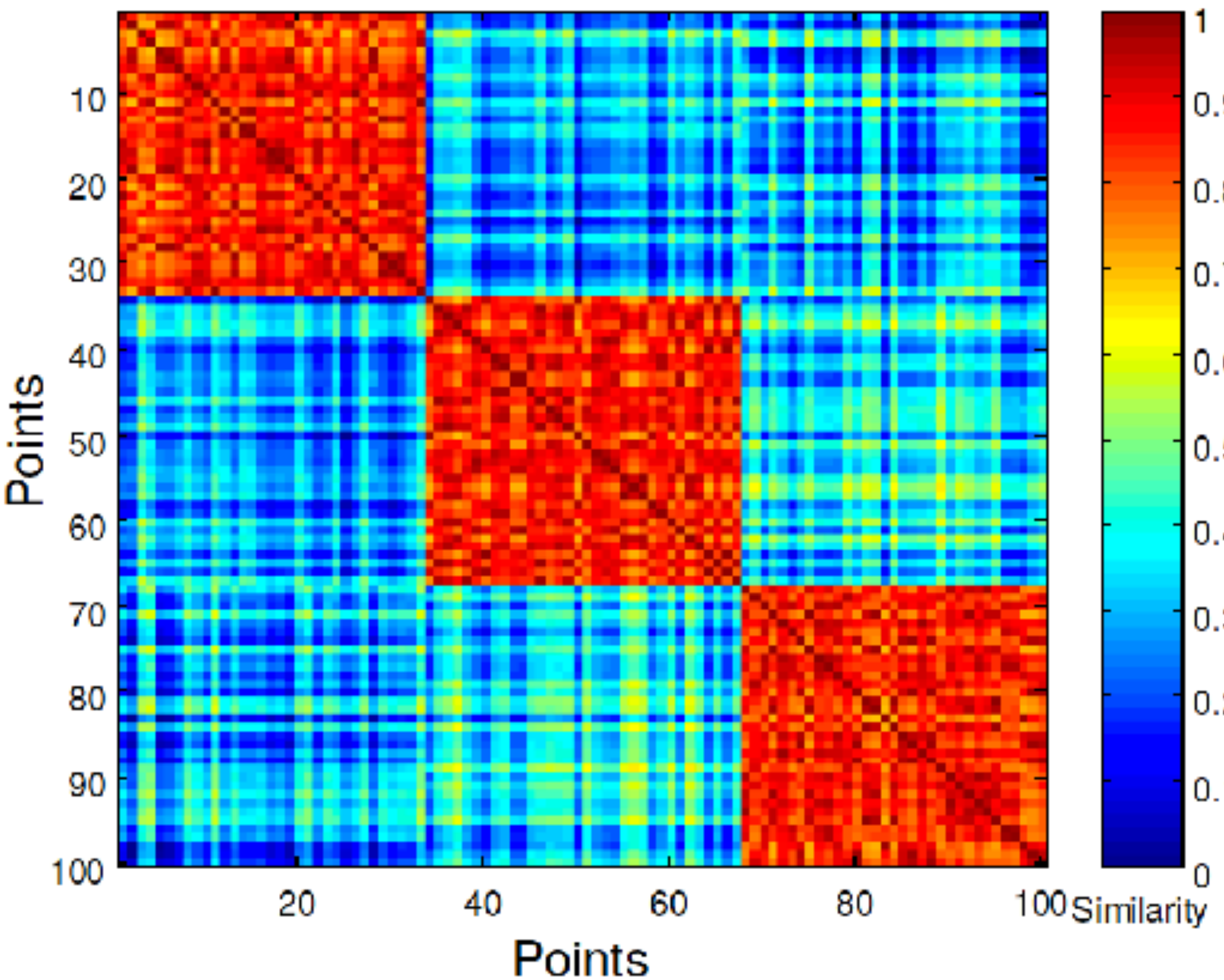


quite good clustering but quite low correlation coef.

Clustering Validation by Similarity Matrix

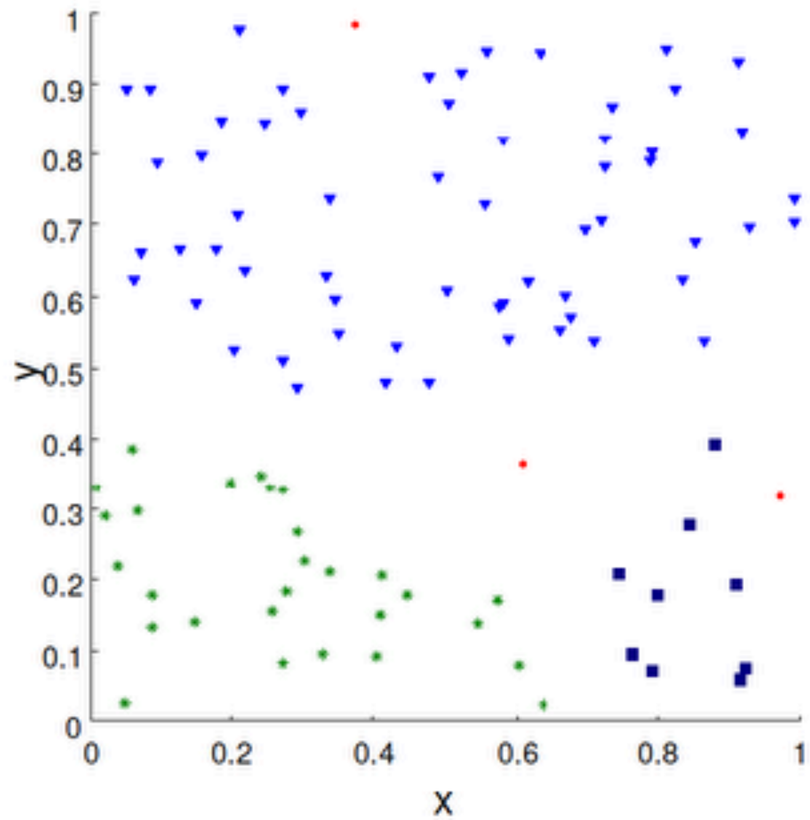
- Similar idea as Correlation method, internal validation
- Order the similarity matrix w.r.t. cluster labels
- Instead of a statistic, it returns a plot!
- Good for inspect clusters from high-dim data
- Same problem for non-spherical clusters

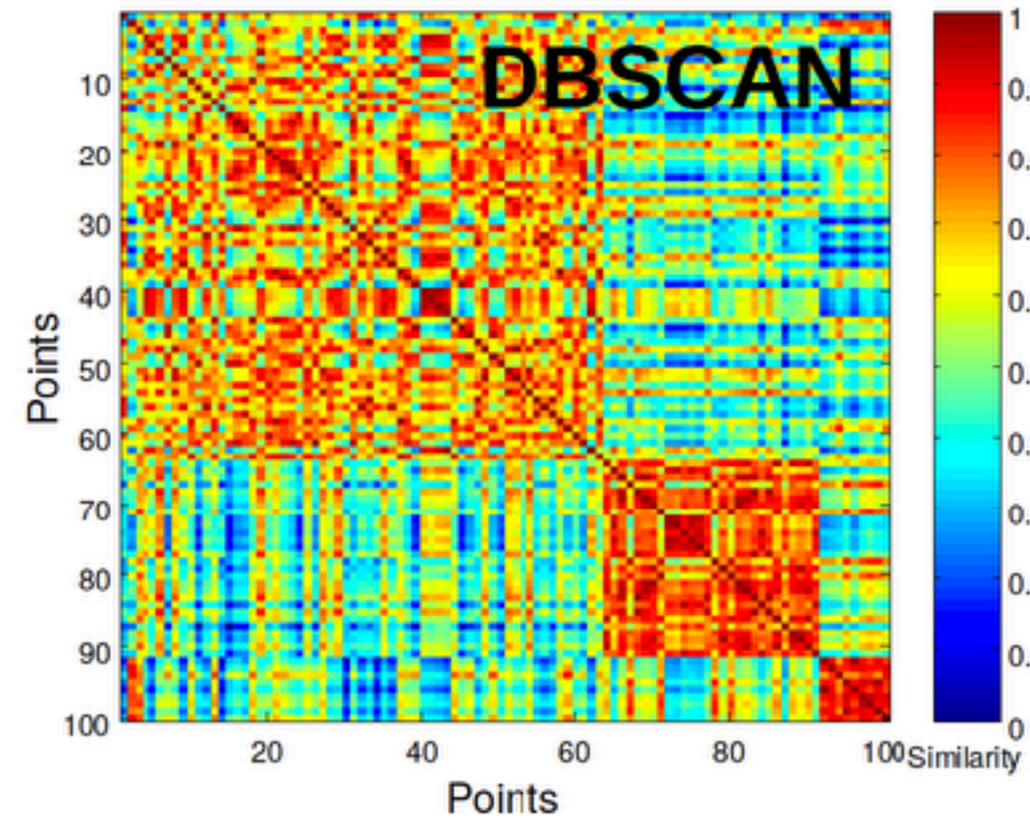


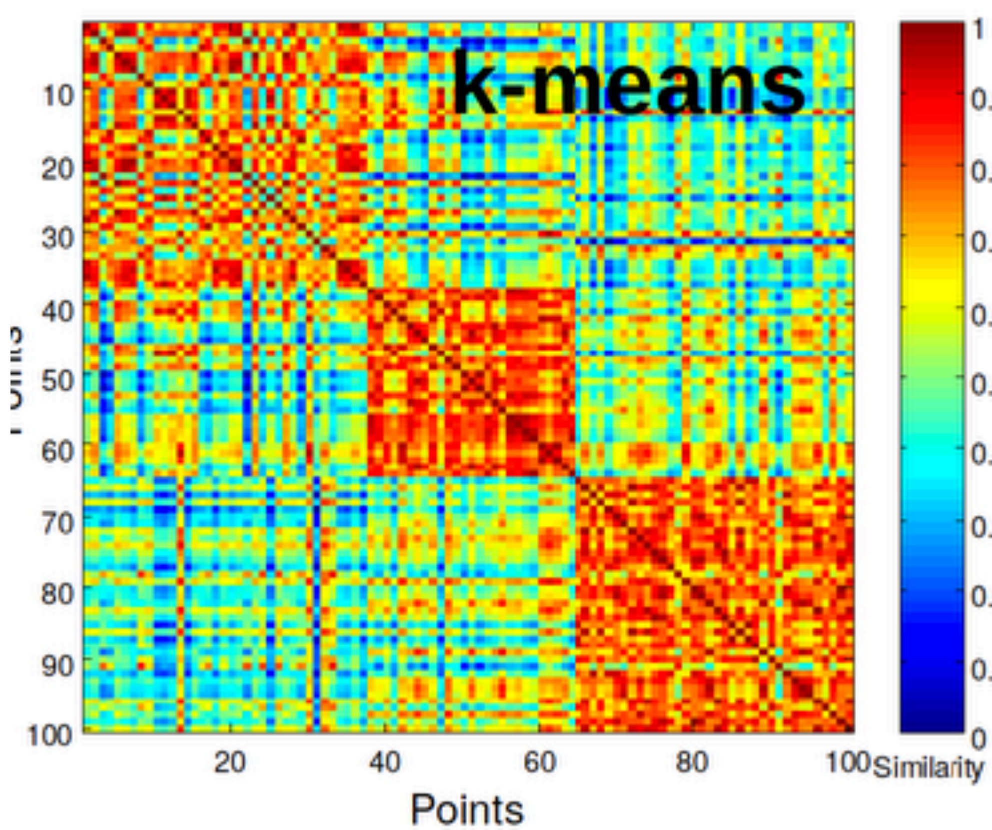


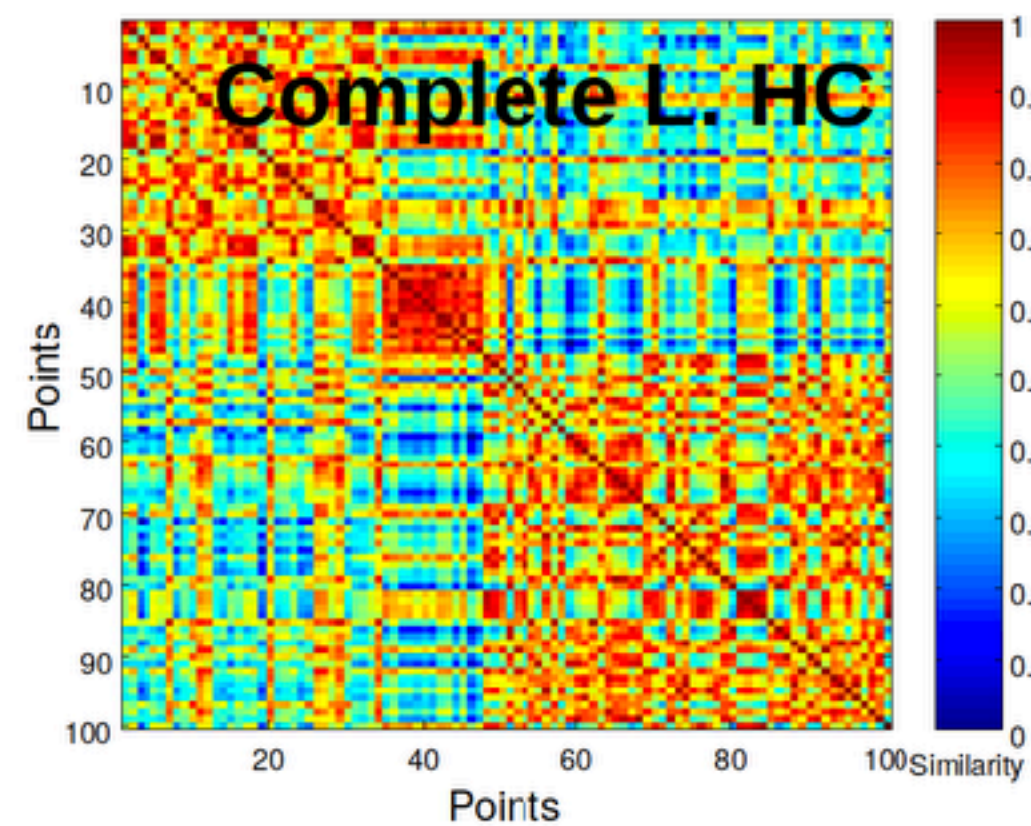
Clustering Validation by Similarity Matrix

How it looks like when clustering random data







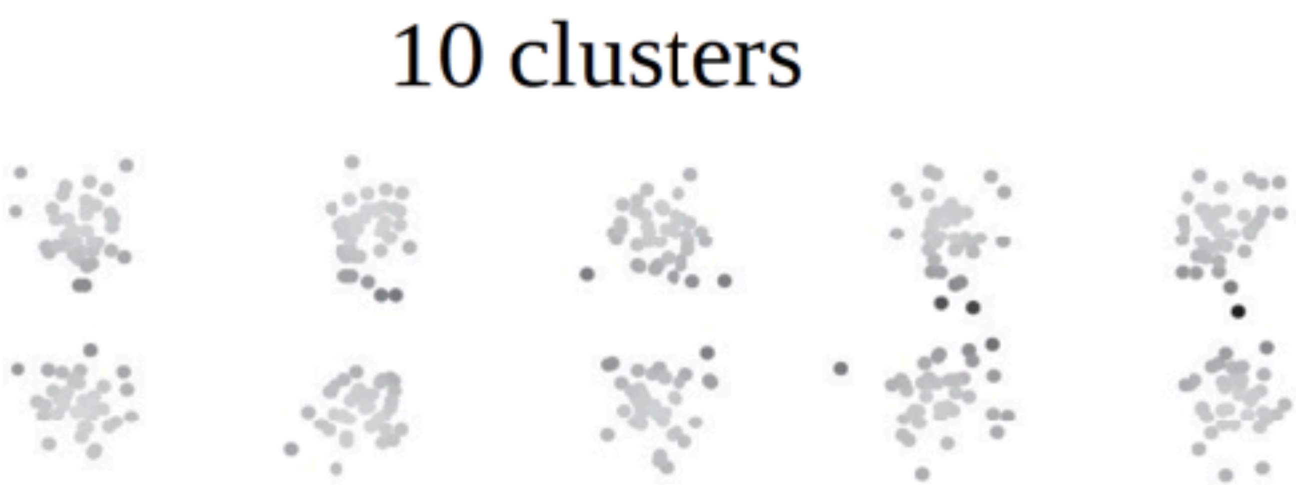


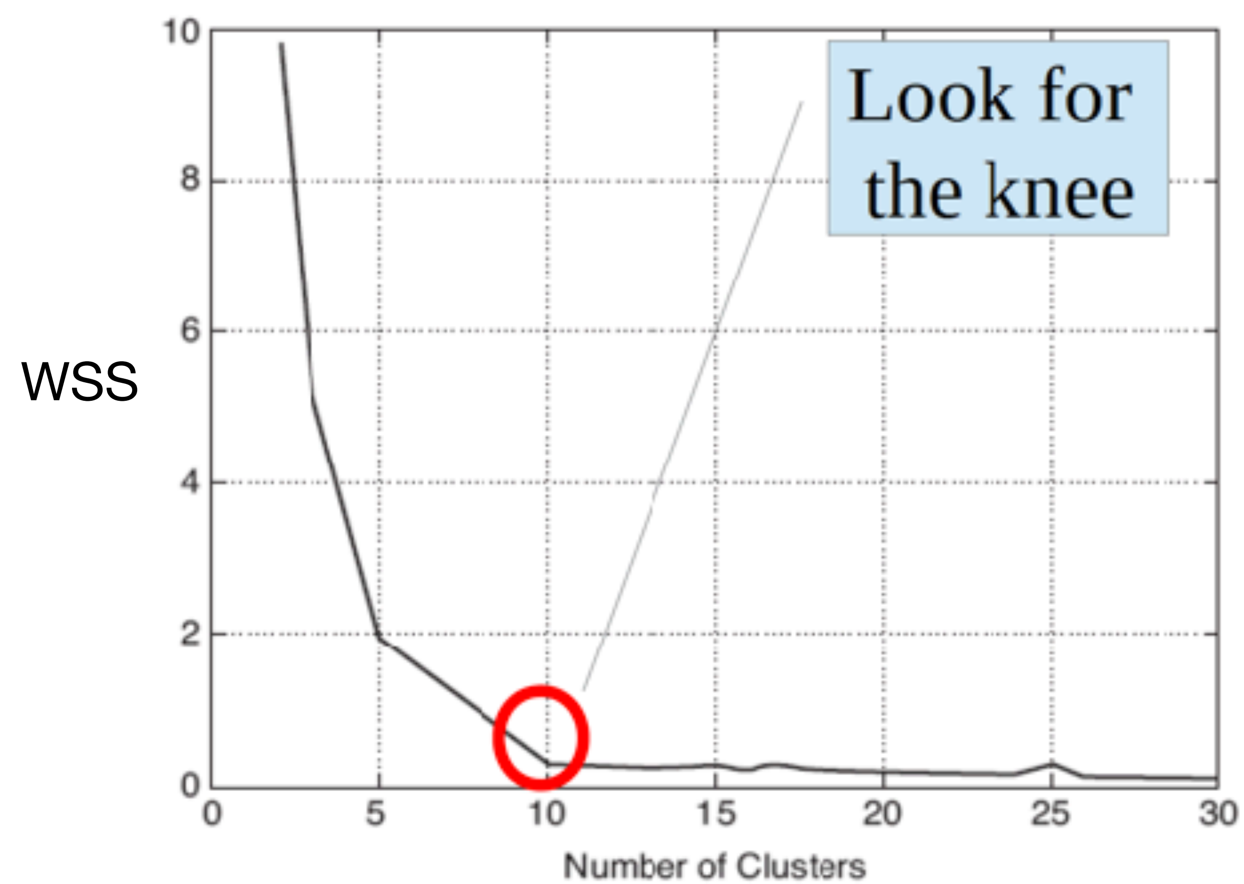
Clustering Validation by Internal Measures

2 internal measures: Within Cluster Sum of Squares (WSS) & Between Cluster Sum of Squares (BSS)

$$\text{WSS} = \sum_{k=1}^{K} \sum_{x_i \in C_k} \sum_{x_j \in C_k} d(x_i, x_j) \qquad \text{BSS} = \sum_{k=1}^{K} \sum_{x_i \in C_k} \sum_{x_j \notin C_k} d(x_i, x_j)$$

Note: Other WSSs & BSSs can be defined that take care of cluster probability, soft clustering & do not scale with # of data!

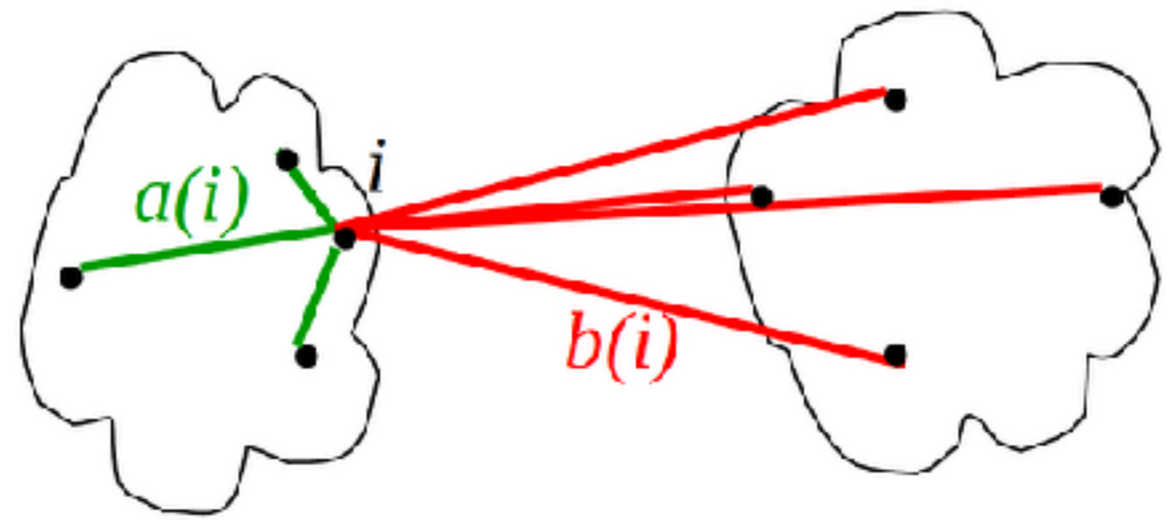




Clustering Validation by Internal Measures: Silhouette Coef.

- Take into account both Cohesion & Separation

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i; i \neq j} d(x_i, x_j)$$



where C_i is the cluster x_i belongs to

$$b(i) = \min_{k \neq i} \boxed{\frac{1}{|C_k|} \sum_{j \in C_k} d(x_i, x_j)}$$

mean distance of x_i to all points in another cluster not containing x_i

Silhouette Value of one point:

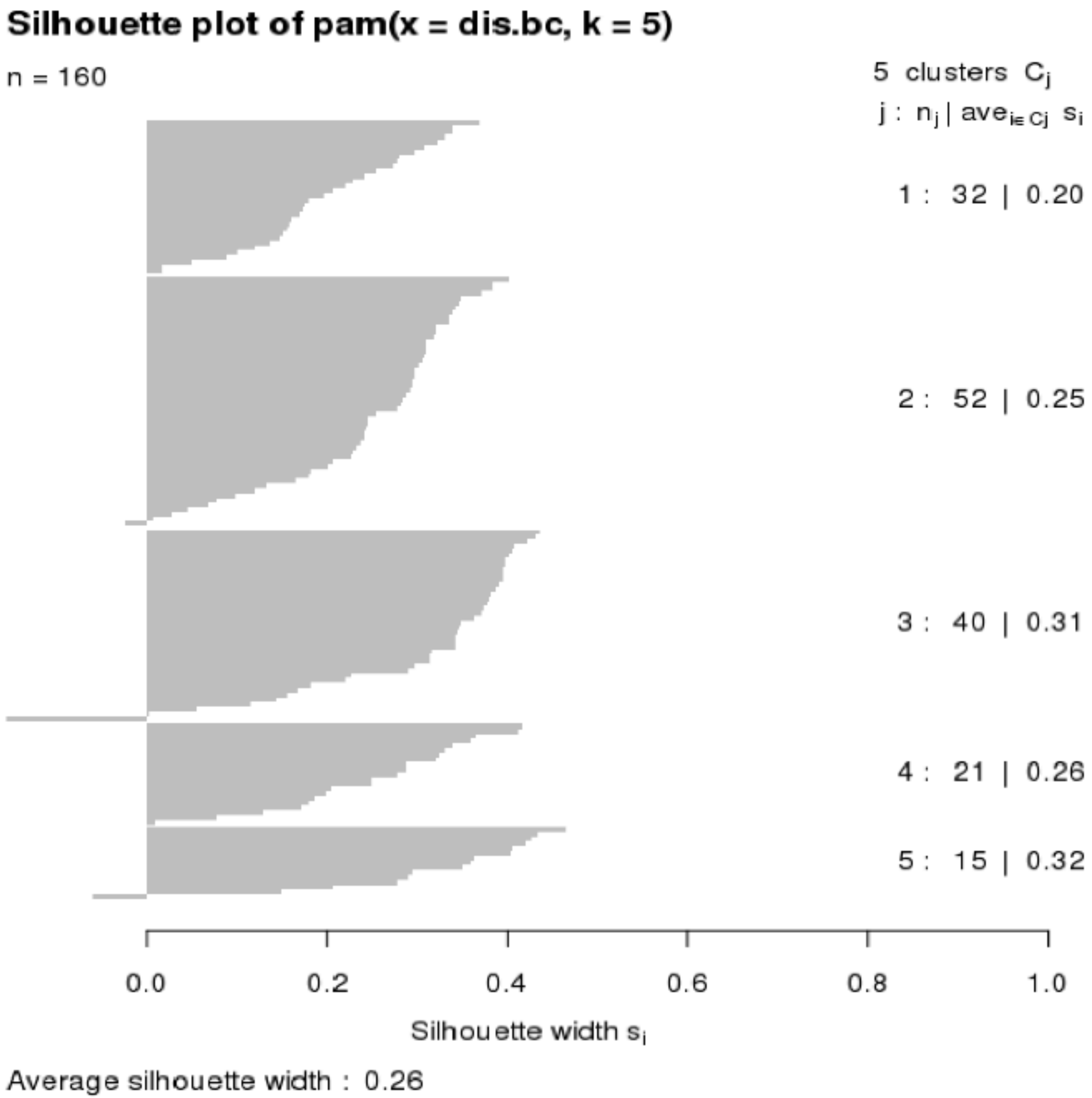
$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

$$-1 \leq s(i) \leq 1$$

$s(i)$ close to 1 for points clustered correctly, i.e., $a(i) << b(i)$

Silhouette Coef. = Average Silhouette value over all data points

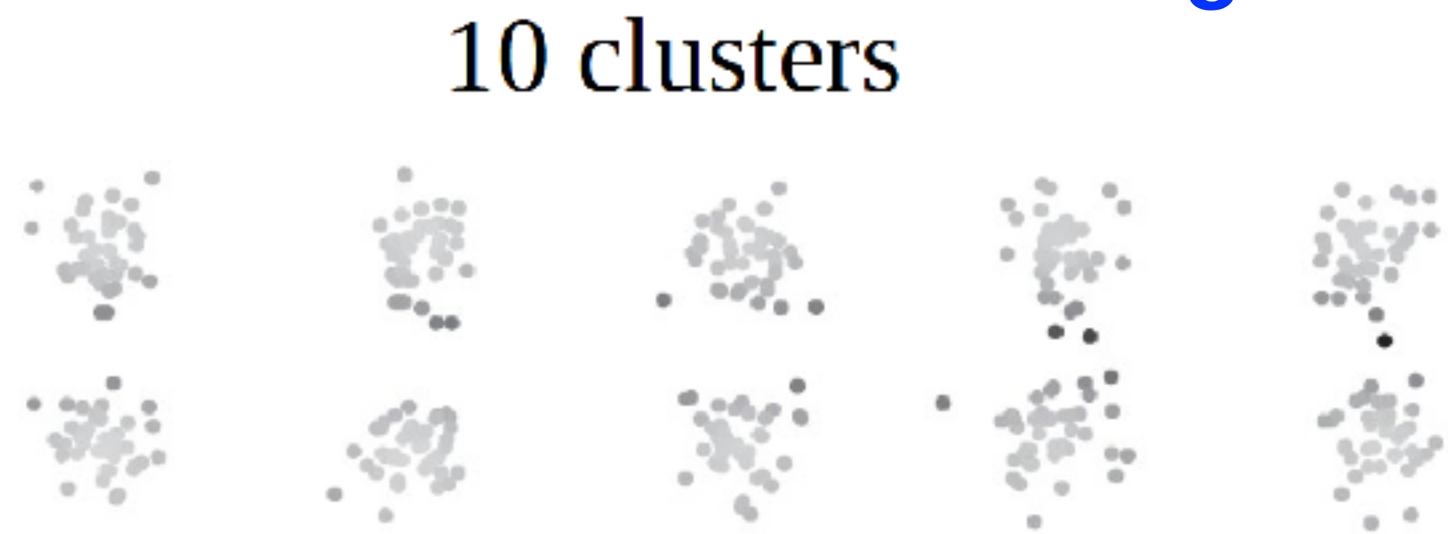
Clustering Validation by Internal Measures: Silhouette Plot

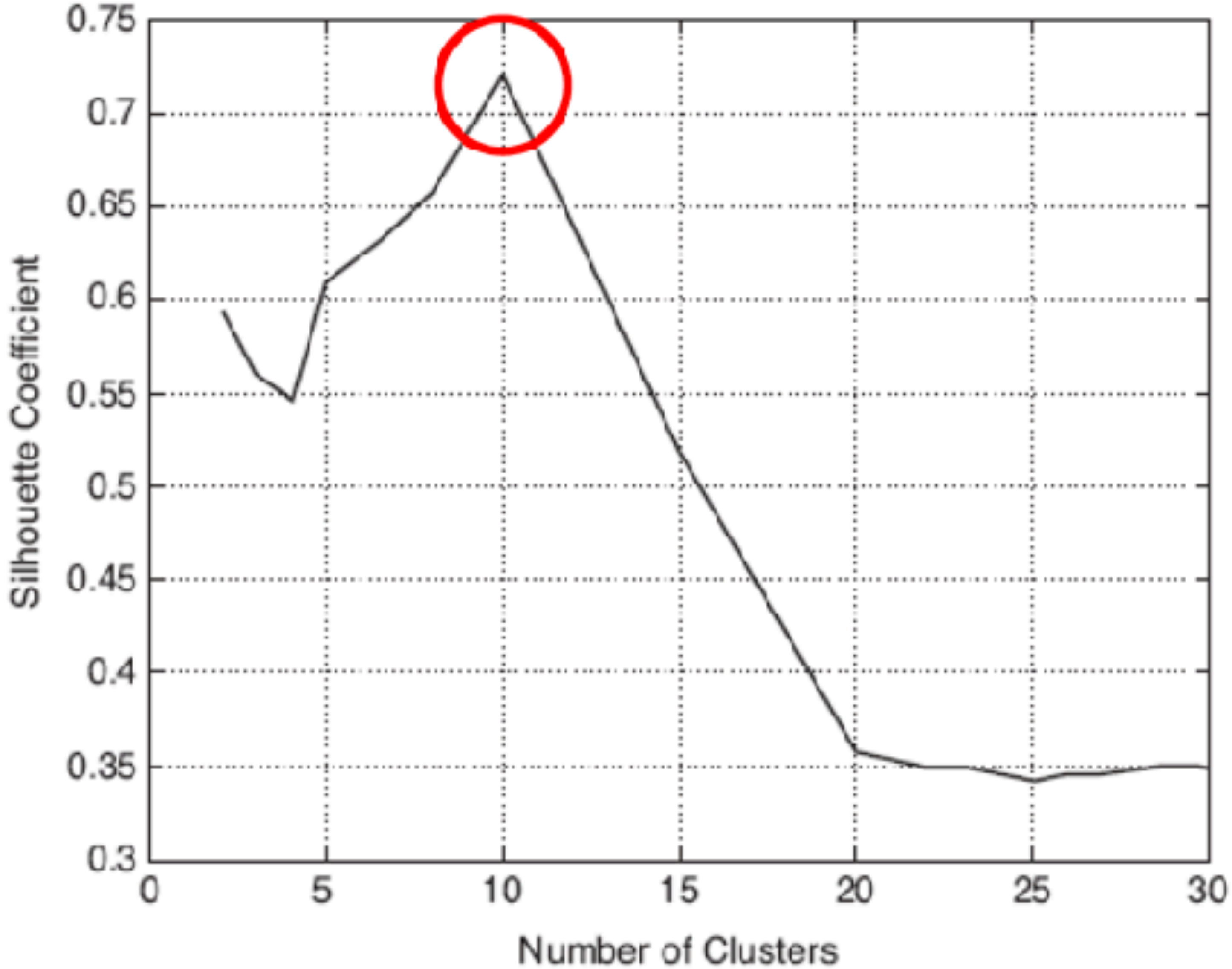


Remarks:

- Provide point-wise, cluster-wise & global diagnoses
- Also work for classification problem!

Clustering Validation by Internal Measures: Choosing # of clusters with Silhouette





Ex:

- Why initially rise when # of cluster is not enough, then drops when # of clusters is too much?
- How does the curve look like for uniform data?
- What if the data does not have discrete structure?
- What if the data has hierarchical structure?

Note: Other validation measures using cohesion & separation - Dunn index, Davies-Bouldin index

Remarks (Full set of clustering problems)

- Over-clustering and under-clustering
- Wrong identification of cluster centers (number & position)
- Incorrect assignment of data point to cluster
- Need soft clustering generalization
- Incorrect noise assignment: background noise & outliers

  (assign cluster points as noise & vice versa)
- local problems, separately investigate cohesion/separation.

Introduction to Dim Reduction Validation

Unsupervised Learning VT 2021

Class note created by Chun based on the paper "A methodology to compare Dimensionality Reduction algorithms in terms of loss of quality" Info. Sci., 270:1 2014 (https://www.sciencedirect.com/science/article/pii/S0020025514001741)

Remark: The paper is quite poorly written and I don't really recommend reading it. However, it contains a good collection of reference for various DR validation methods

Overview

- Distance preserving: Preserve distances from original space to projected space. DR methods relatively easy to validate. This type of DR are quite restrictive to reveal underlying low dim manifold.

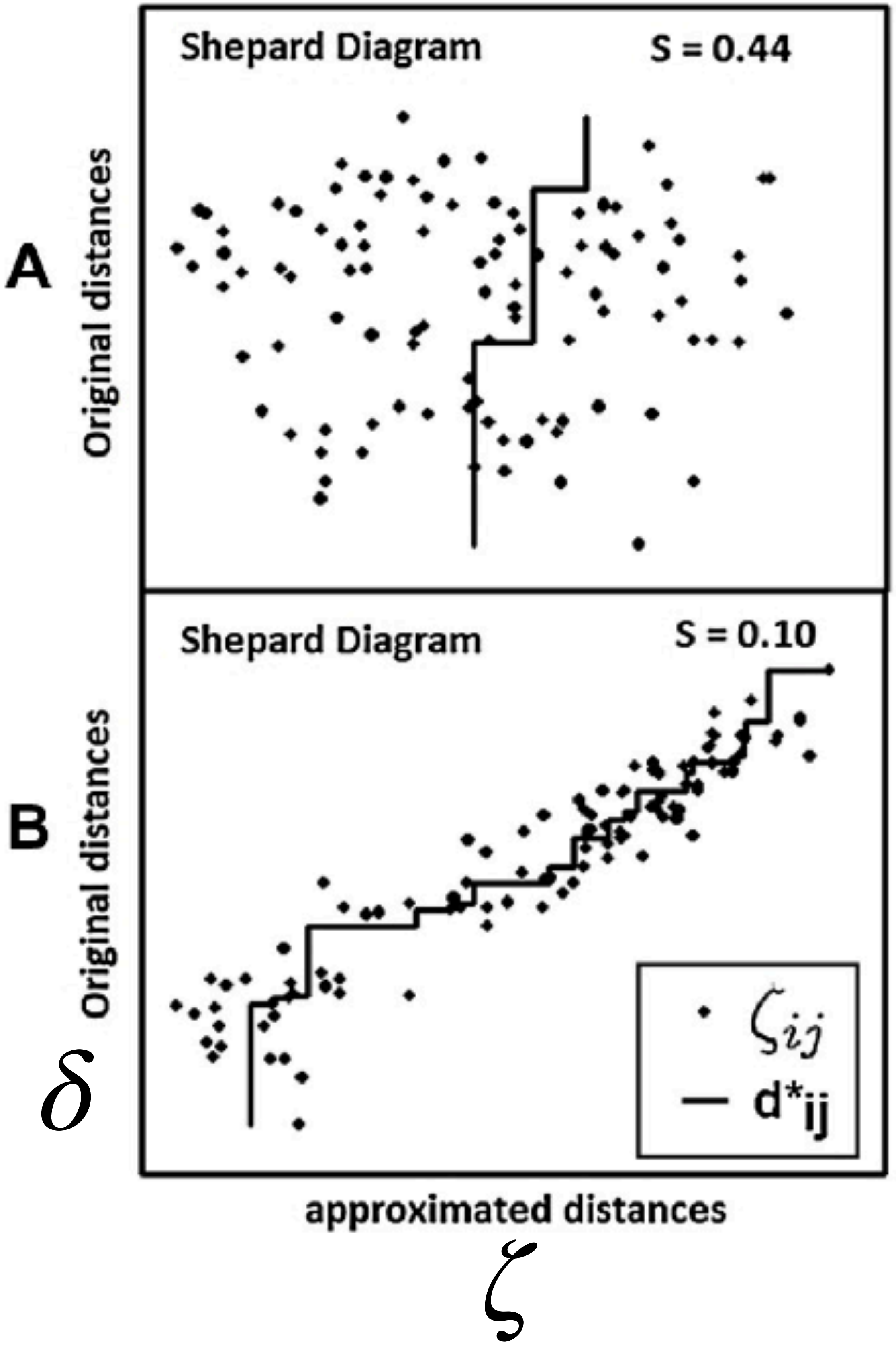
- Topological preserving: Preserve “Topological” properties of underlying manifold. “Topological” properties cannot be summarized in a few indices and relatively difficult to validate. This type of DR are more flexible.

E.g. of Topological Equivalence

- Cluster validation: If DR is for data visualization and perform clustering (e.g. tSNE, UMAP), clustering validation can be performed in projected space.

- Some DR methods come with own validation index (e.g. PCA, MDS, Sammon mapping, ISOMAP, etc.)

Distance Preserving: Shepard Diagam



- Indices are summary statistics of Shepard Diagram
- Simplest validation index: Pearson correlation coef.
- Sammon stress $= \dfrac{1}{\sum_{i<j} \delta_{ij}} \sum_{i<j} \dfrac{(\delta_{ij} - \zeta_{ij})^2}{\delta_{ij}}$
- S stress $= \sqrt{\dfrac{1}{n} \dfrac{\sum_{i<j} (\delta_{ij}^2 - \zeta_{ij}^2)^2}{\sum_{i<j} \delta_{ij}^4}}$
- Compare to monotonic increase relation between original dist. and approx. dist.:

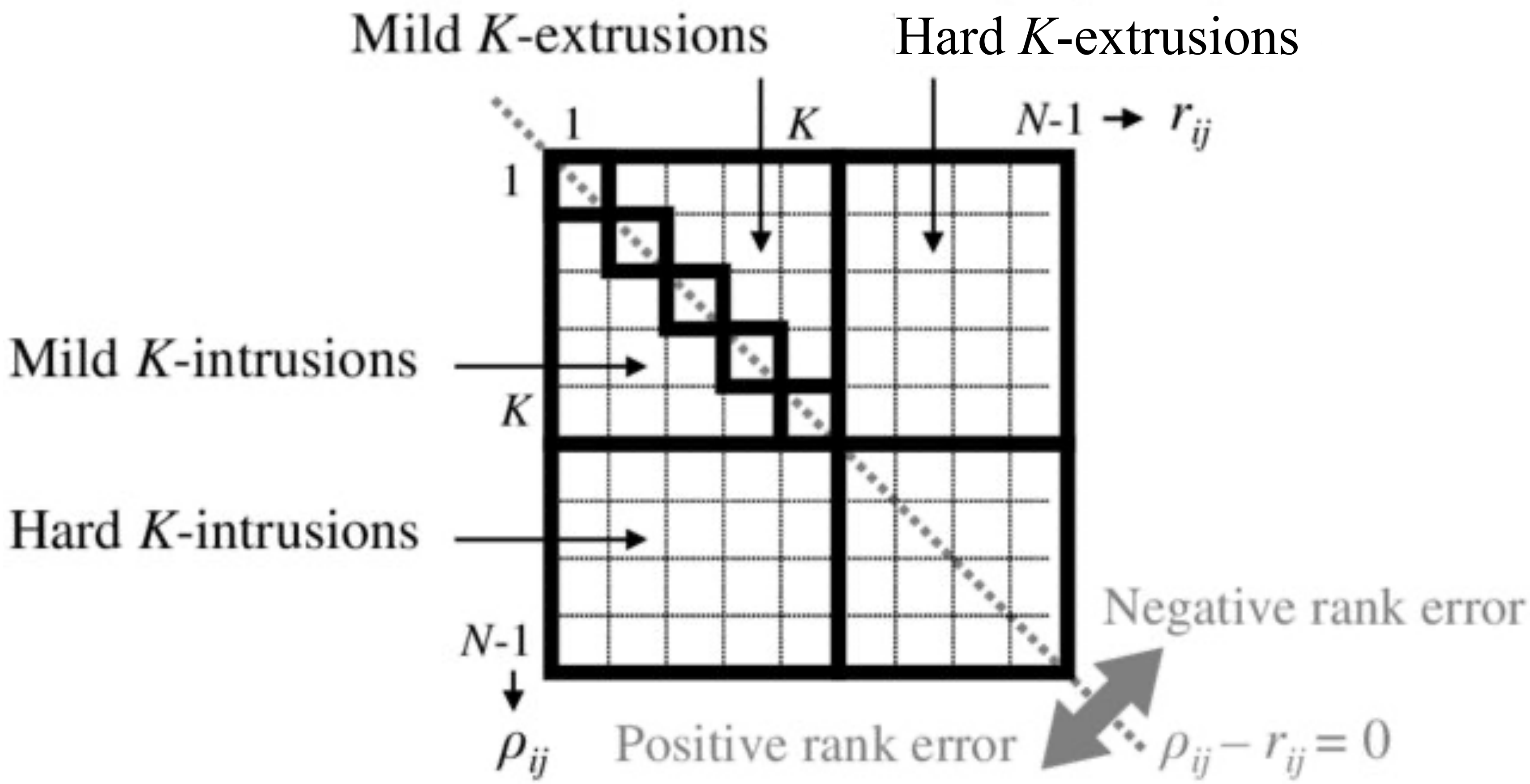
  Kruskal S measure $= \sqrt{\dfrac{\sum_{i,j} (\zeta_{ij} - d_{ij}^*)^2}{\sum_{i,j} \zeta_{ij}^2}}$

d_{ij}^* : monotonic regression

Topological Preserving: Rank based Shepard Diagram

- Shepard diagram with x-axis the distance rank in projected space, y-axis the distance rank in original. space.
- Simplest validation index: Spearman’s rho (Pearson correlation coef. for the rank based Shepard diagram)
- Other validation indices similar to previous page can be defined.

Topological Preserving: Co-Ranking Matrix



ρ_{ij}: distance rank of *j*th point respect to *i*th point in original space

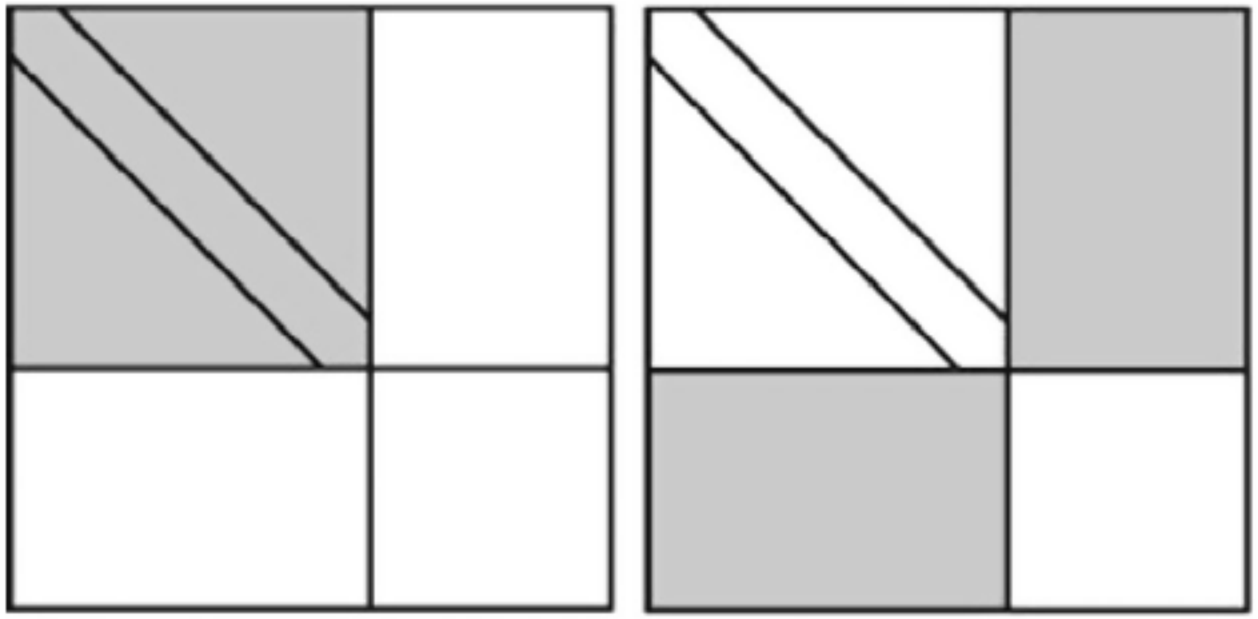
r_{ij}: distance rank for *j*th point respect to *i*th point in projected space

Matrix element Q_{kl}: number of point pairs with $\rho_{ij} = k, r_{ij} = l$ (Ex: What is $\sum_{k,l} Q_{kl}$?)

Extrusion: points leave a neighborhood wrongly $\rho_{ij} < r_{ij}$

Intrusion: points enter a neighborhood wrongly $\rho_{ij} > r_{ij}$

Topological Preserving: Co-Ranking Matrix



Left panel: Lee & Verleysen index for how rank order preserves in a *K*-nearest neighborhood

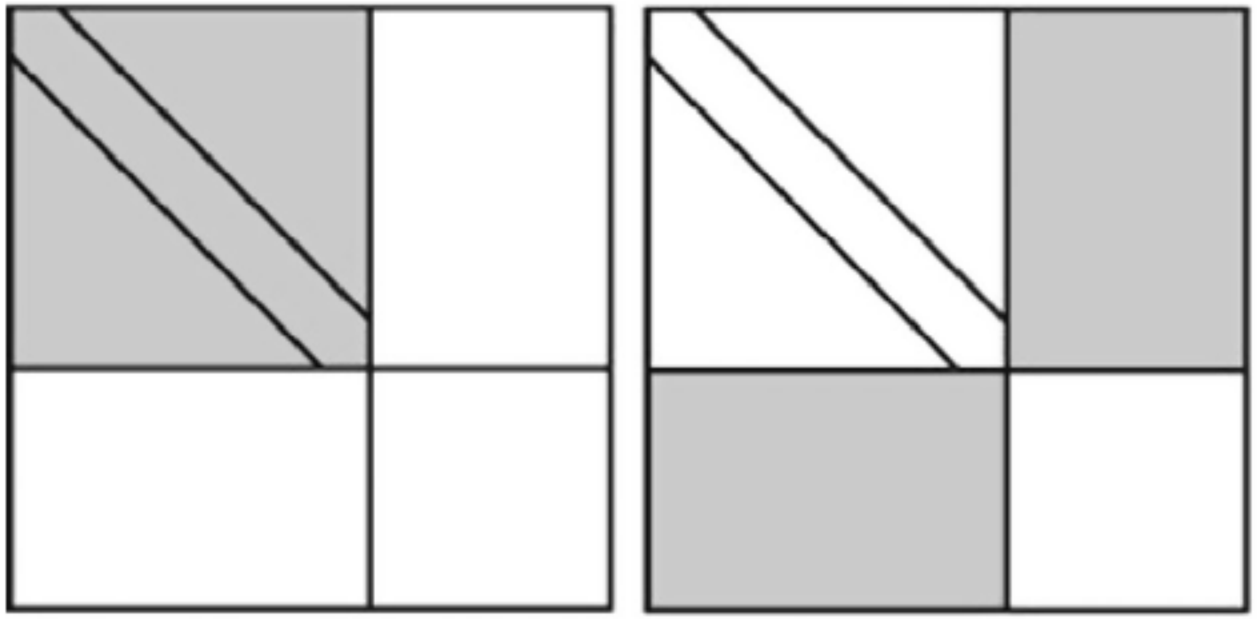
$$Q_{NX}(K) = \frac{1}{Kn} \sum_{k=1}^{K} \sum_{l=1}^{K} Q_{kl}$$

Ex: Can you show that $Q_{NX}(K) = \frac{1}{Kn} \sum_{i=1}^{n} |\Psi_K(i) \bigcap \Psi'_K(i)|$?

where $\Psi_K(i)$ is set of points in original space that are within K-NN from i,
$\Psi_K(i)$ is set of points in projected space that are within K-NN from i,
| . | is the cardinal of the set.

Ex: What is the range of $Q_{NX}(K)$? and what value for perfect rank preserving?

Topological Preserving: Co-Ranking Matrix



Right panel: Venn and Kaski propose the measure of Trustworthiness & Continuity (Only briefly)

Trustworthiness: Data points originally far enter neighborhood after projection (Intrusion)

Continuity: Data points originally close push away after projection, i.e., break continuity (extrusion)