Clustering methods: Part 3

Cluster validation

Pasi Fränti

10.5.2017

Machine Learning

University of Eastern Finland

UNIVERSITY OF EASTERN FINLAND

Part I:

Introduction

Cluster validation

Supervised classification:

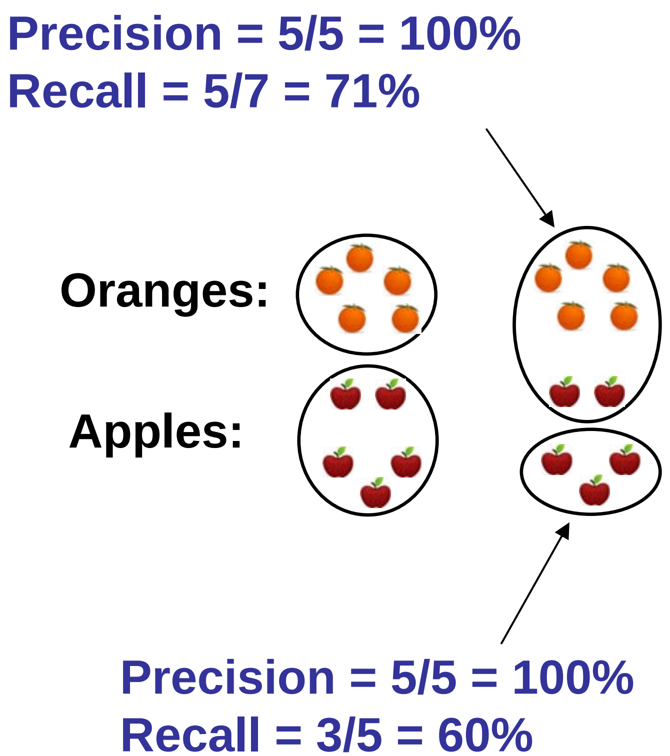
- Ground truth class labels known
- Accuracy, precision, recall

Cluster analysis:

- No class labels

Validation need to:

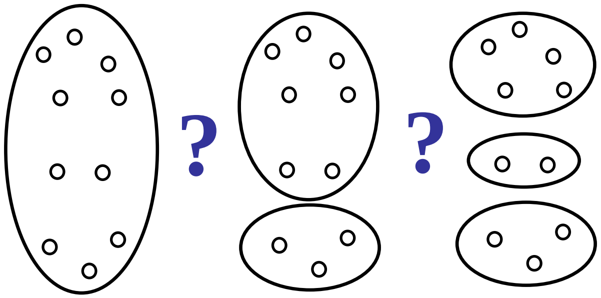
- Compare clustering algorithms
- **Solve the number of clusters**
- Avoid finding patterns in noise



Measuring clustering validity

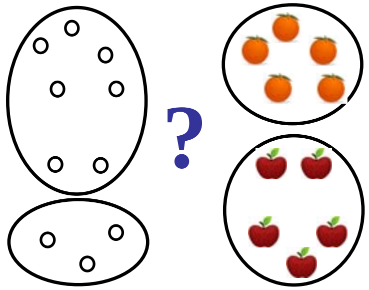
Internal Index:

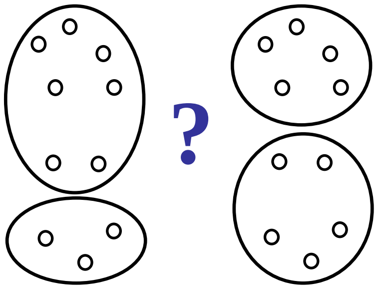
- Validate *without* external info
- With different number of clusters
- Solve the number of clusters



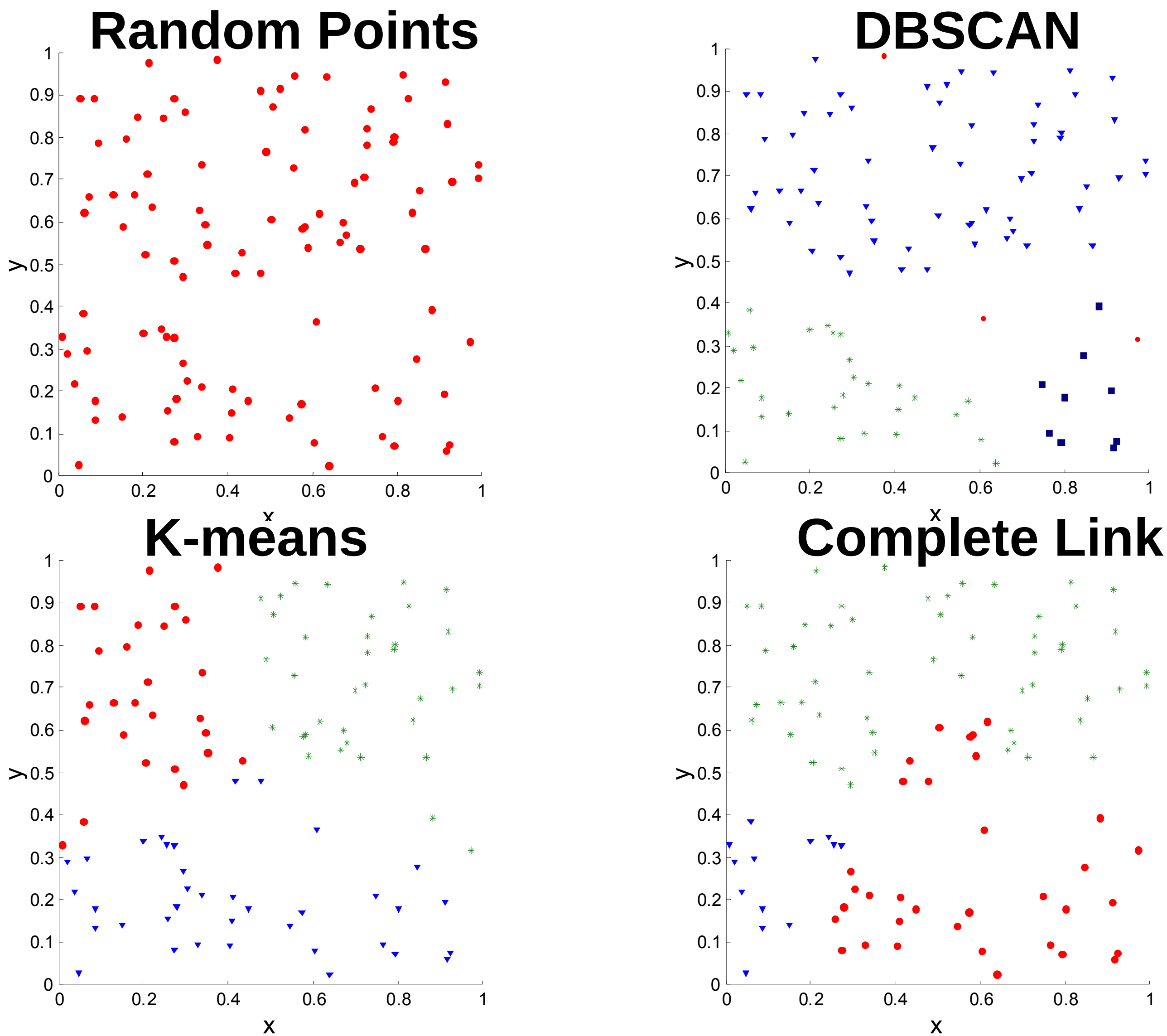
External Index

- Validate against ground truth
- Compare two clusters: (how similar)





Clustering of random data

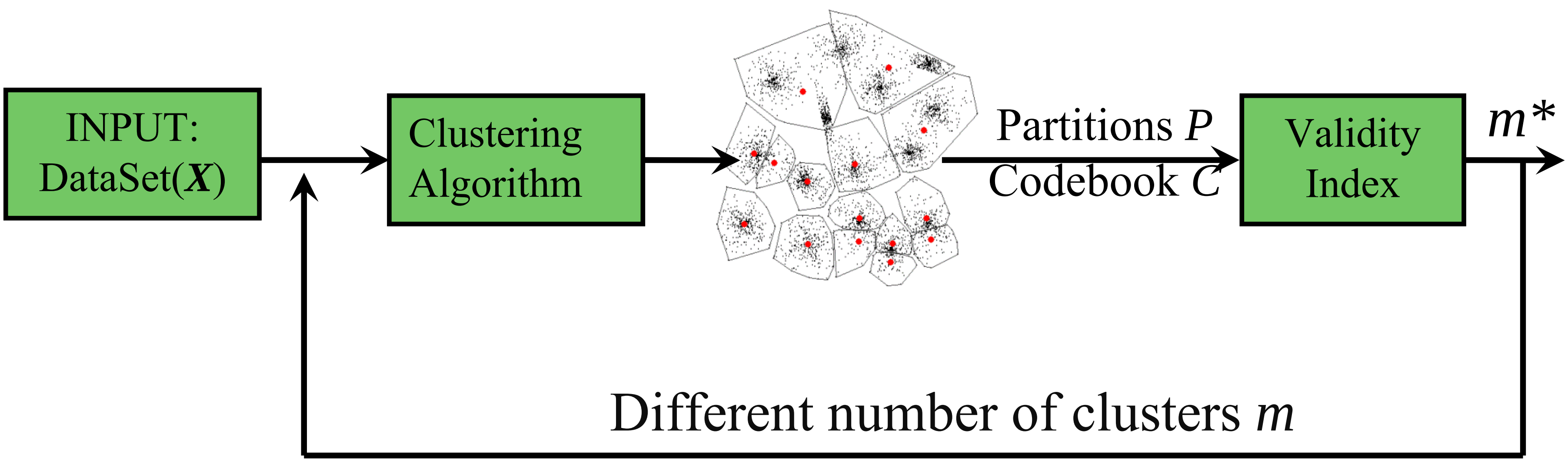


Cluster validation process

1. Distinguishing whether non-random structure actually exists in the data (one cluster).
2. Comparing the results of a cluster analysis to external ground truth (class labels).
3. Evaluating how well the results fit the data *without* reference to external information.
4. Comparing two different clustering results to determine which is better.
5. Determining the number of clusters.

Cluster validation process

- **Cluster validation** refers to procedures that evaluate the results of clustering in a **quantitative** and **objective** fashion. [Jain & Dubes, 1988]
  - How to be "quantitative": To employ the measures.
  - How to be "objective": To validate the measures!



Part II:

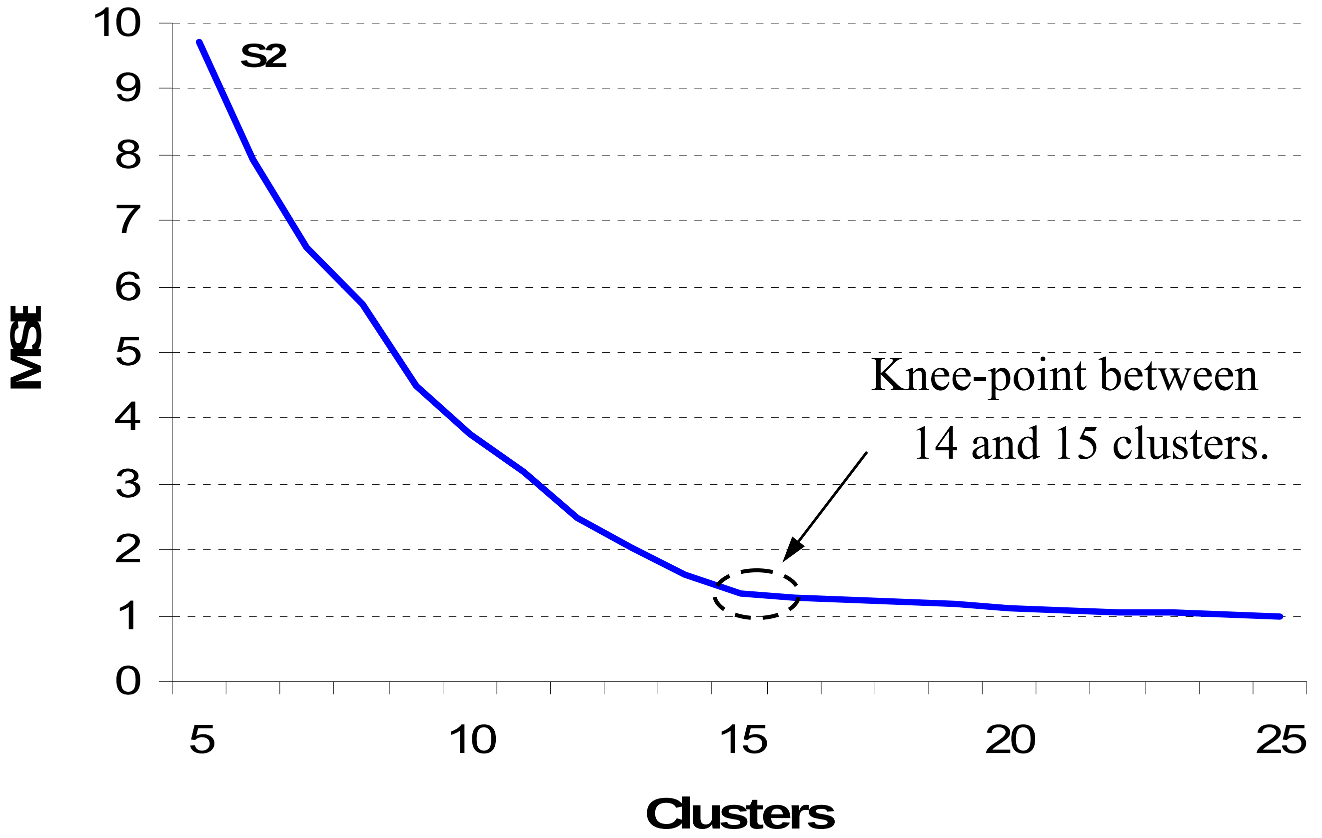
Internal indexes

Internal indexes

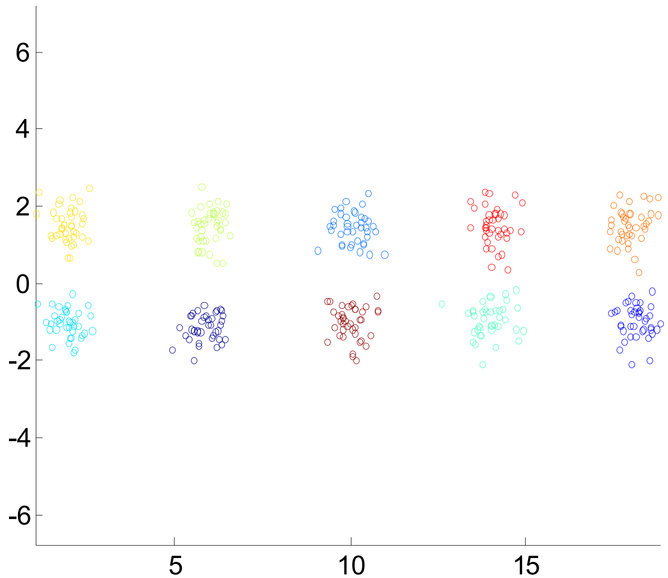
- Ground truth is rarely available but unsupervised validation must be done.
- Minimizes (or maximizes) internal index:
  - Variances of within cluster and between clusters
  - Rate-distortion method
  - F-ratio
  - Davies-Bouldin index (DBI)
  - Bayesian Information Criterion (BIC)
  - Silhouette Coefficient
  - Minimum description principle (MDL)
  - Stochastic complexity (SC)

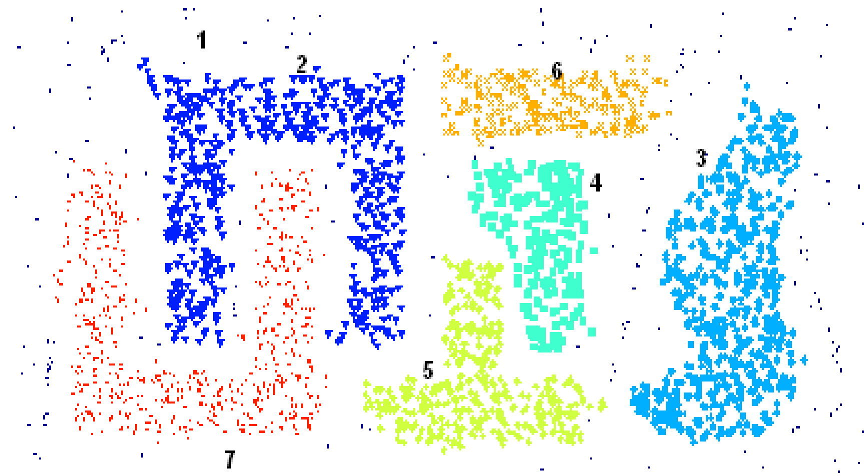
Sum of squared errors

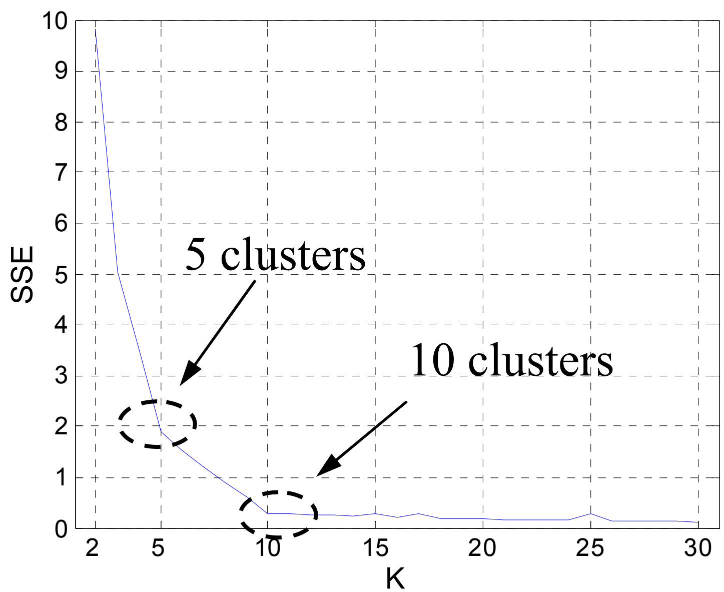
- The more clusters the smaller the value.
- Small knee-point near the correct value.
- But how to detect?

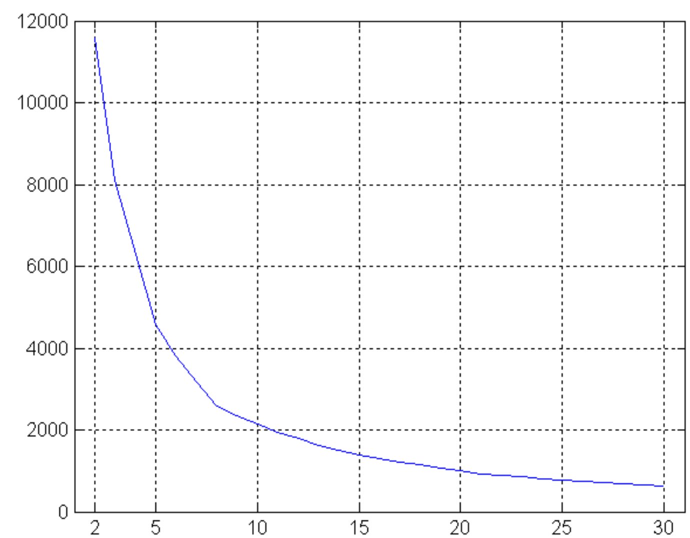


Sum of squared errors



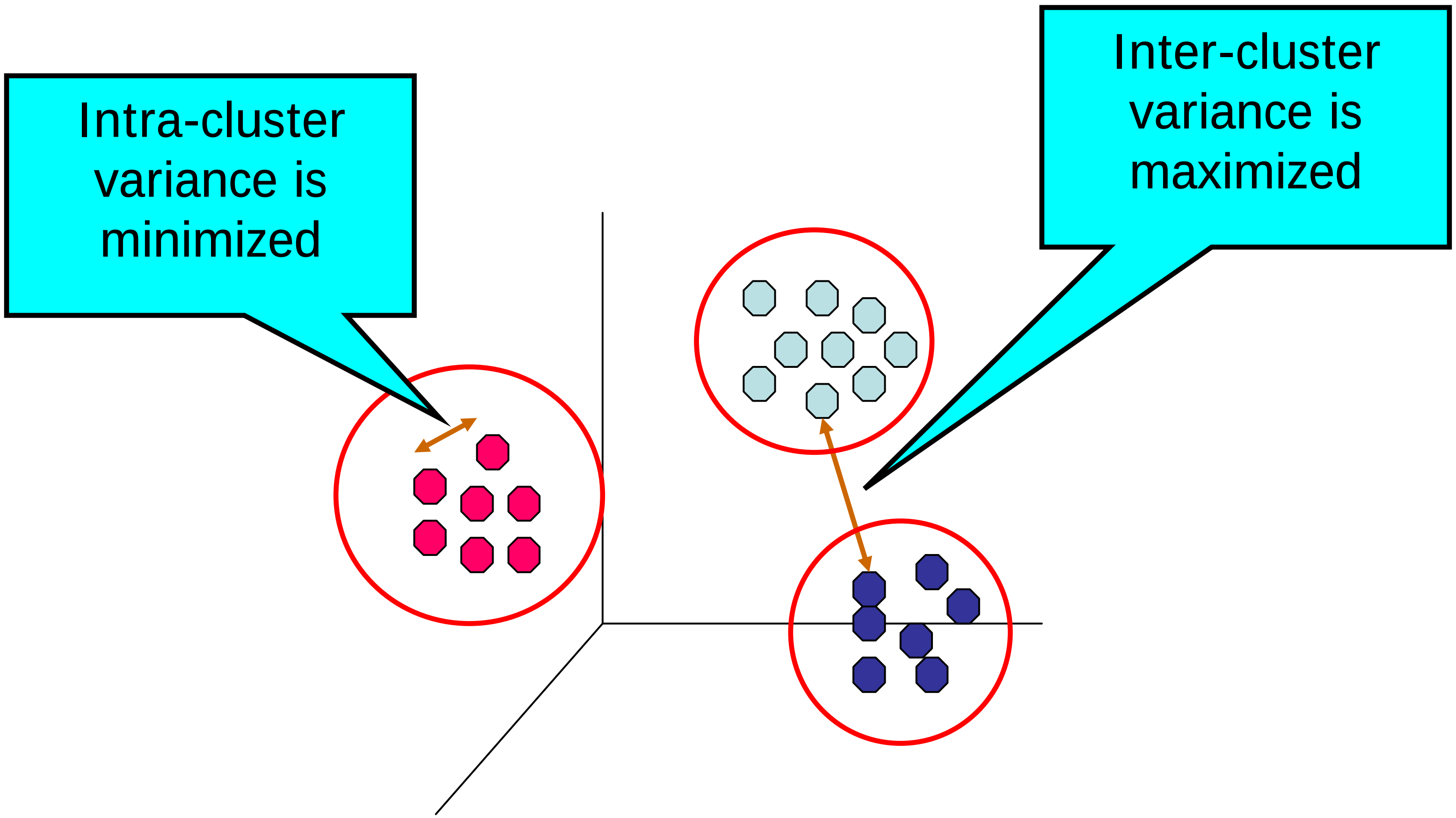






From TSE to cluster validity

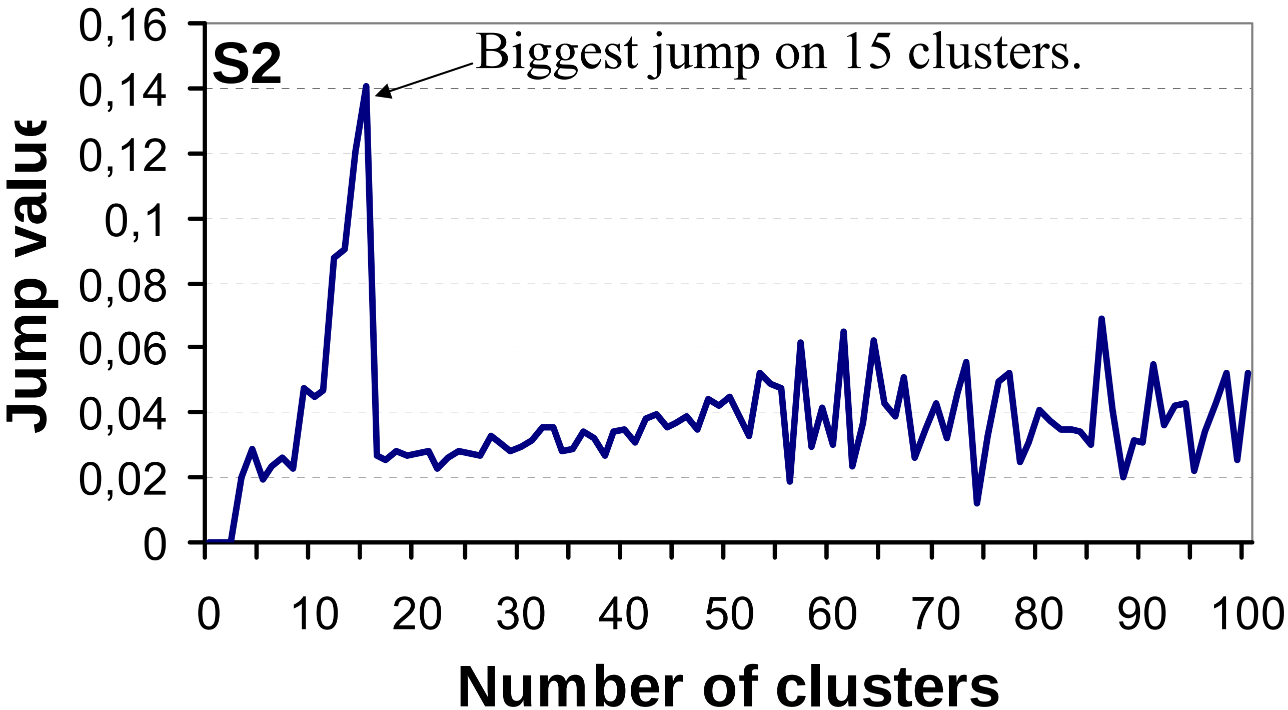
- Minimize within cluster variance (TSE)
- Maximize between cluster variance



Jump point of TSE
(rate-distortion approach)

First derivative of powered TSE values:

$$J(k) = TSE(k)^{-d/2} - TSE(k-1)^{-d/2}$$



Cluster variances

Within cluster: $$SSW(C,k) = \sum_{i=1}^{N} \| x_i - c_{p(i)} \|^2$$

Between clusters: $$SSB(C,k) = \sum_{j=1}^{k} n_j \| c_j - \bar{x} \|^2$$

Total Variance of data set:

$$\sigma(X) = \underbrace{\sum_{i=1}^{N} \| x_i - c_{p(i)} \|^2}_{\text{SSW}} + \underbrace{\sum_{j=1}^{k} n_j \| c_j - \bar{x} \|^2}_{\text{SSB}}$$

WB-index

- Measures ratio of between-groups variance against the within-groups variance
- WB-index:

$$F = \frac{k \cdot \sum_{i=1}^{N} \| x_i - c_{p(i)} \|^2}{\sum_{j=1}^{k} n_j \| c_j - \bar{x} \|^2} = \frac{k \cdot SSW}{\sigma(X) - SSW}$$

SSB

Sum-of-squares based indexes

- *SSW* / k ---- Ball and Hall (1965)
- $k^2|W|$ ---- Marriot (1971)
- $\frac{SSB / k-1}{SSW / N-k}$ ---- Calinski & Harabasz (1974)
- log(*SSB/SSW*) ---- Hartigan (1975)
- $d\log(\sqrt{SSW/(dN^2)})+\log(k)$ ---- Xu (1997)

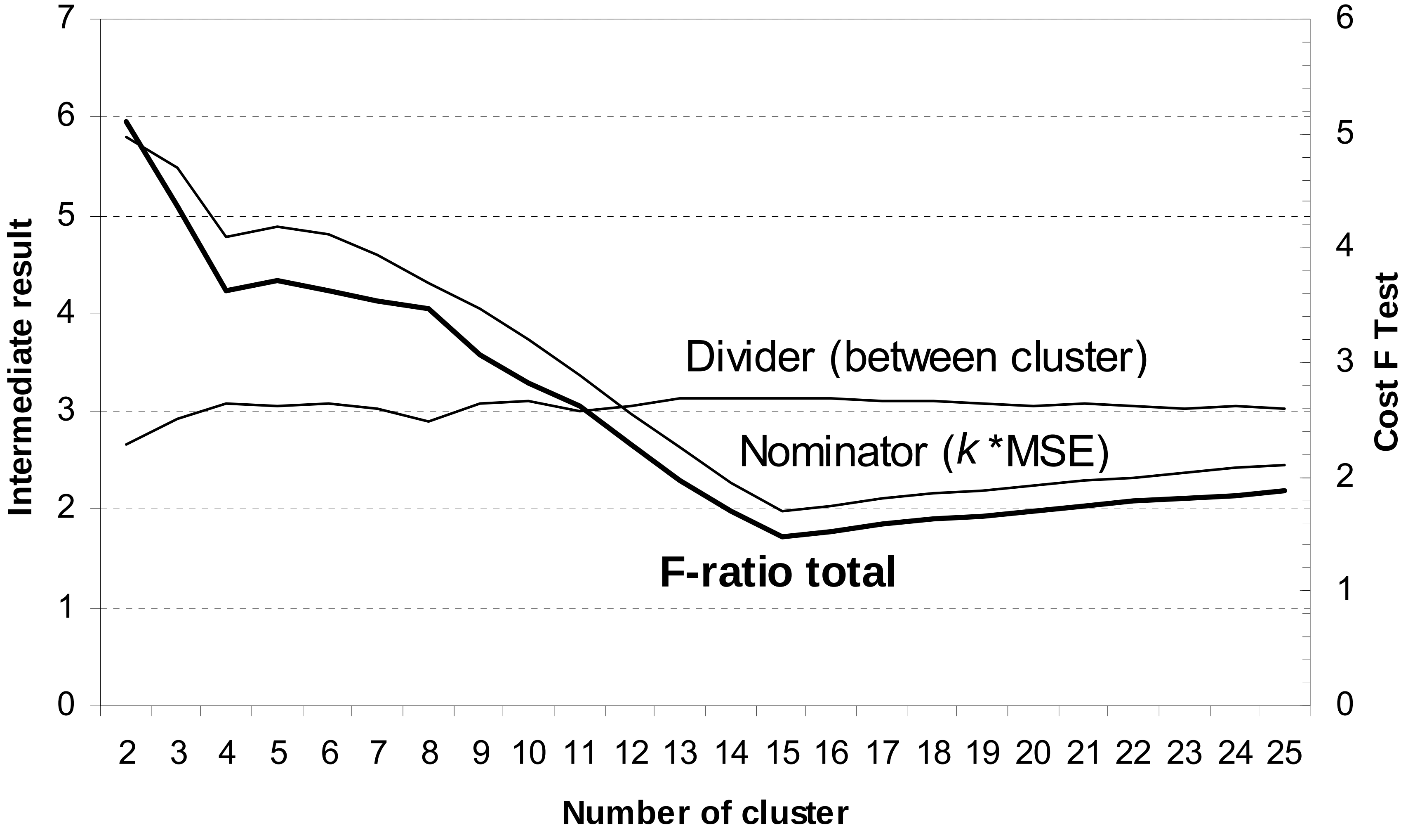
(*d* = dimensions; *N* = size of data; *k* = number of clusters)

SSW = <u>S</u>um of <u>s</u>quares **<u>w</u>ithin** the clusters (=*TSE*)

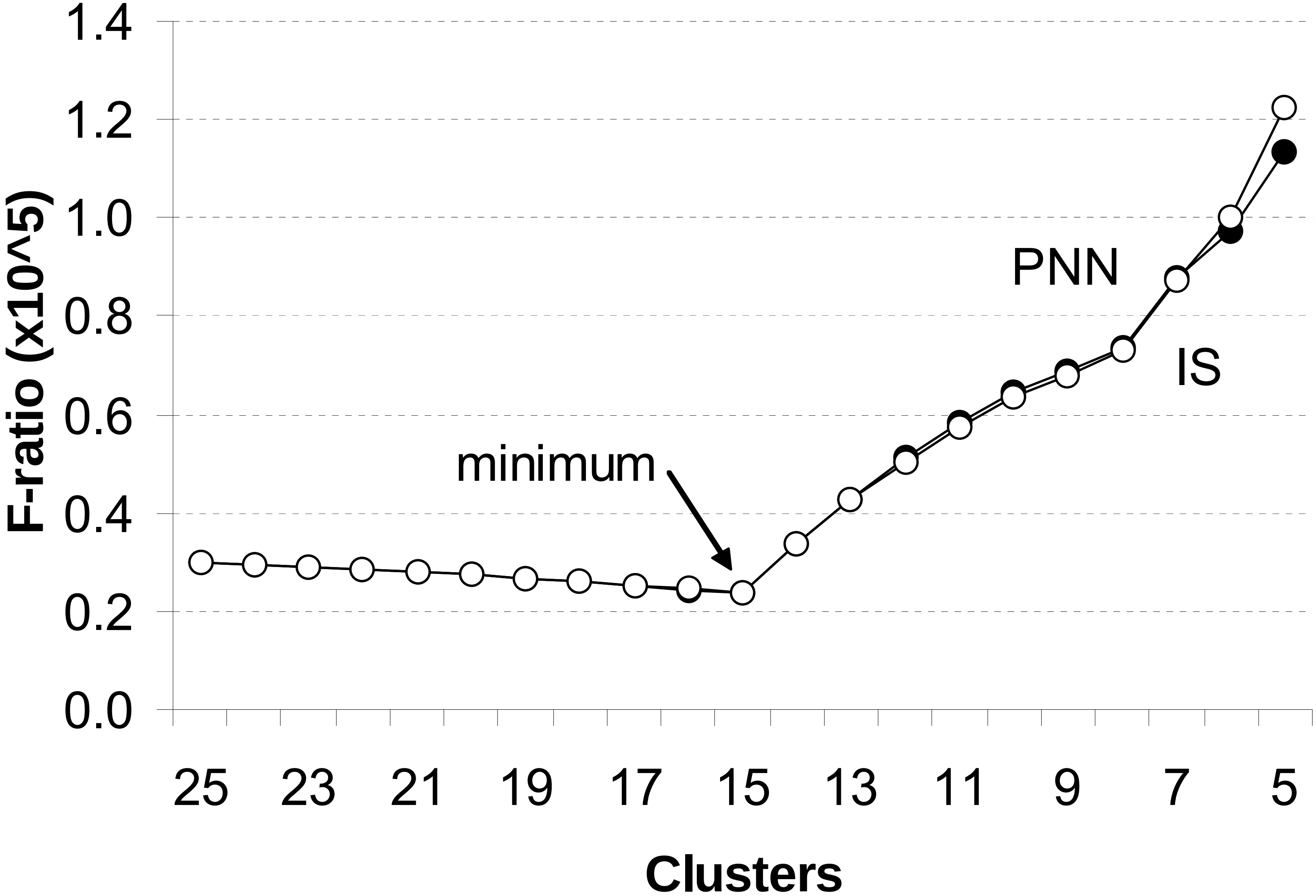
SSB = <u>S</u>um of <u>s</u>quares **<u>b</u>etween** the clusters

Calculation of WB-index

(called also F-ratio / F-test)

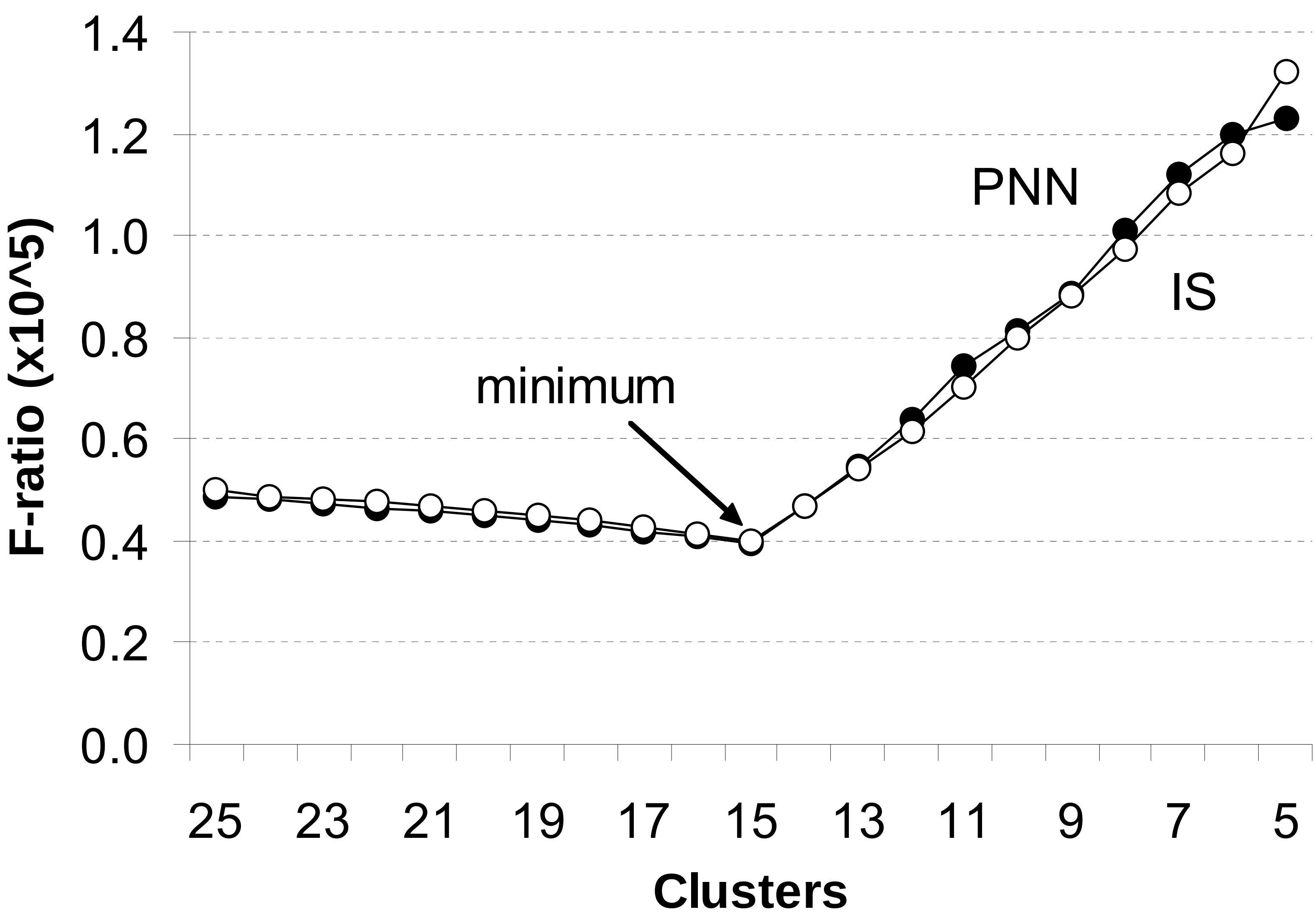


Dataset S1



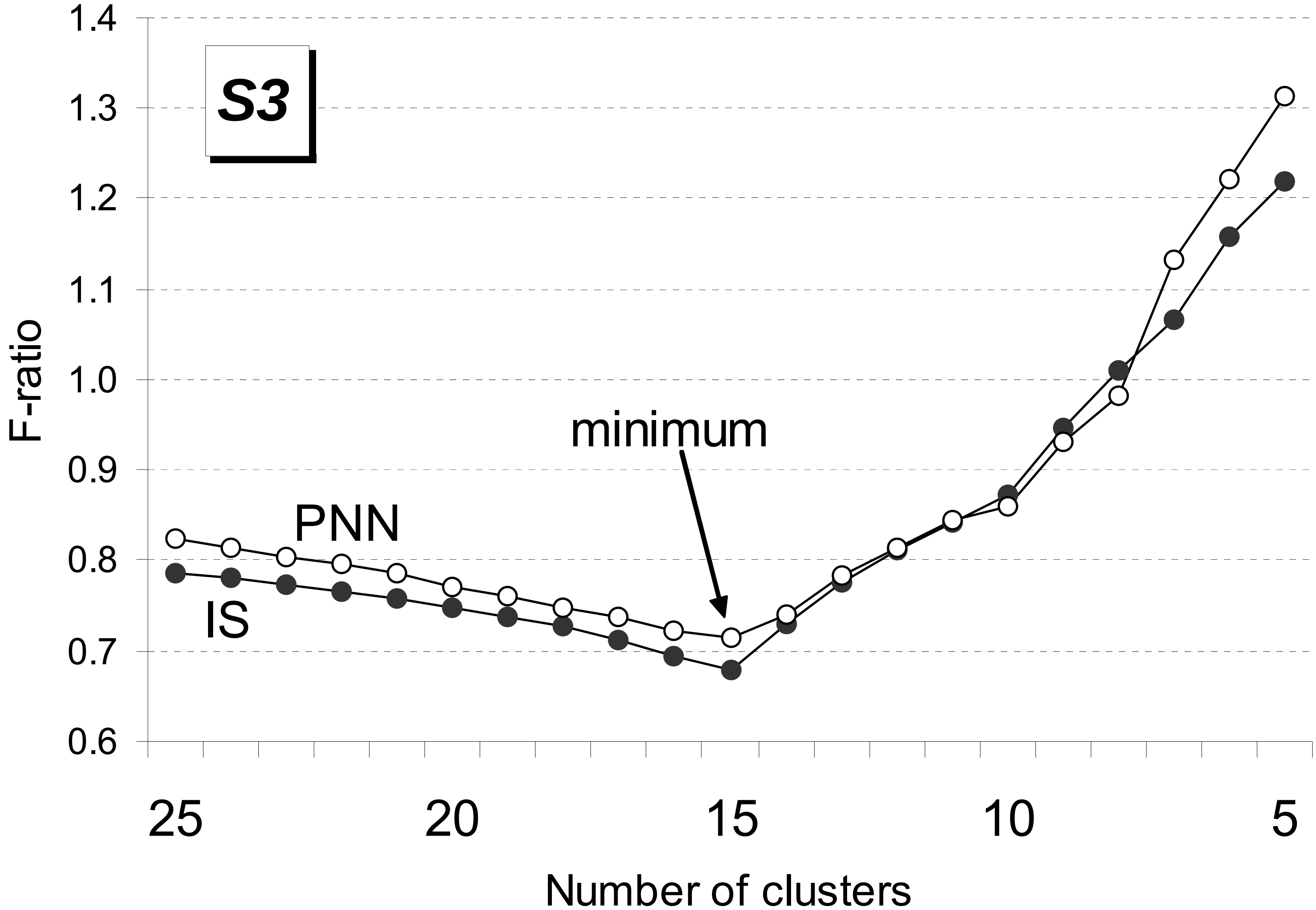
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
F-ratio (x10^5)
PNN
IS
minimum
25
23
21
19
17
15
13
11
9
7
5
Clusters

Dataset S2

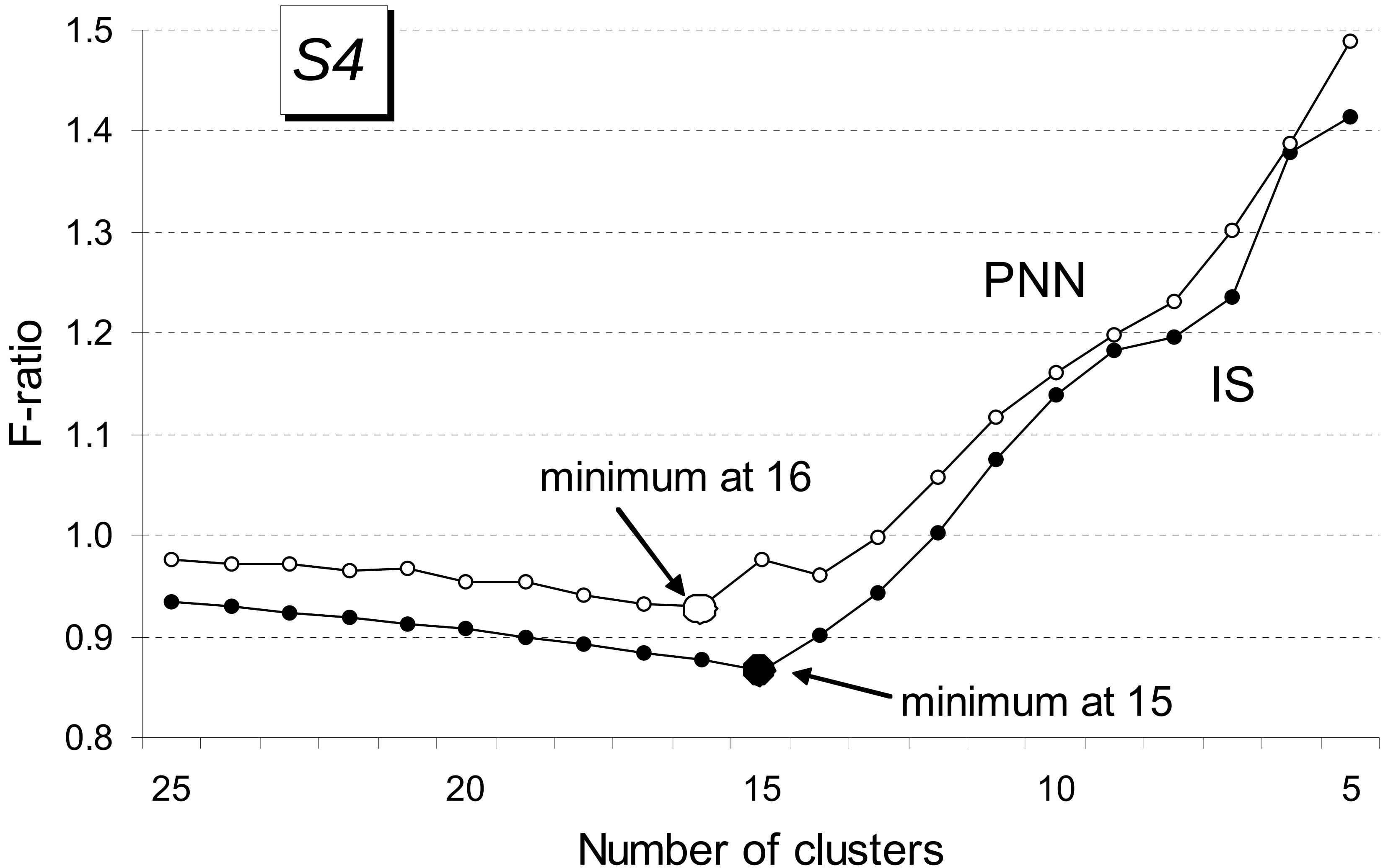


1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
F-ratio (x10^5)
PNN
IS
minimum
25
23
21
19
17
15
13
11
9
7
5
Clusters

Dataset S3

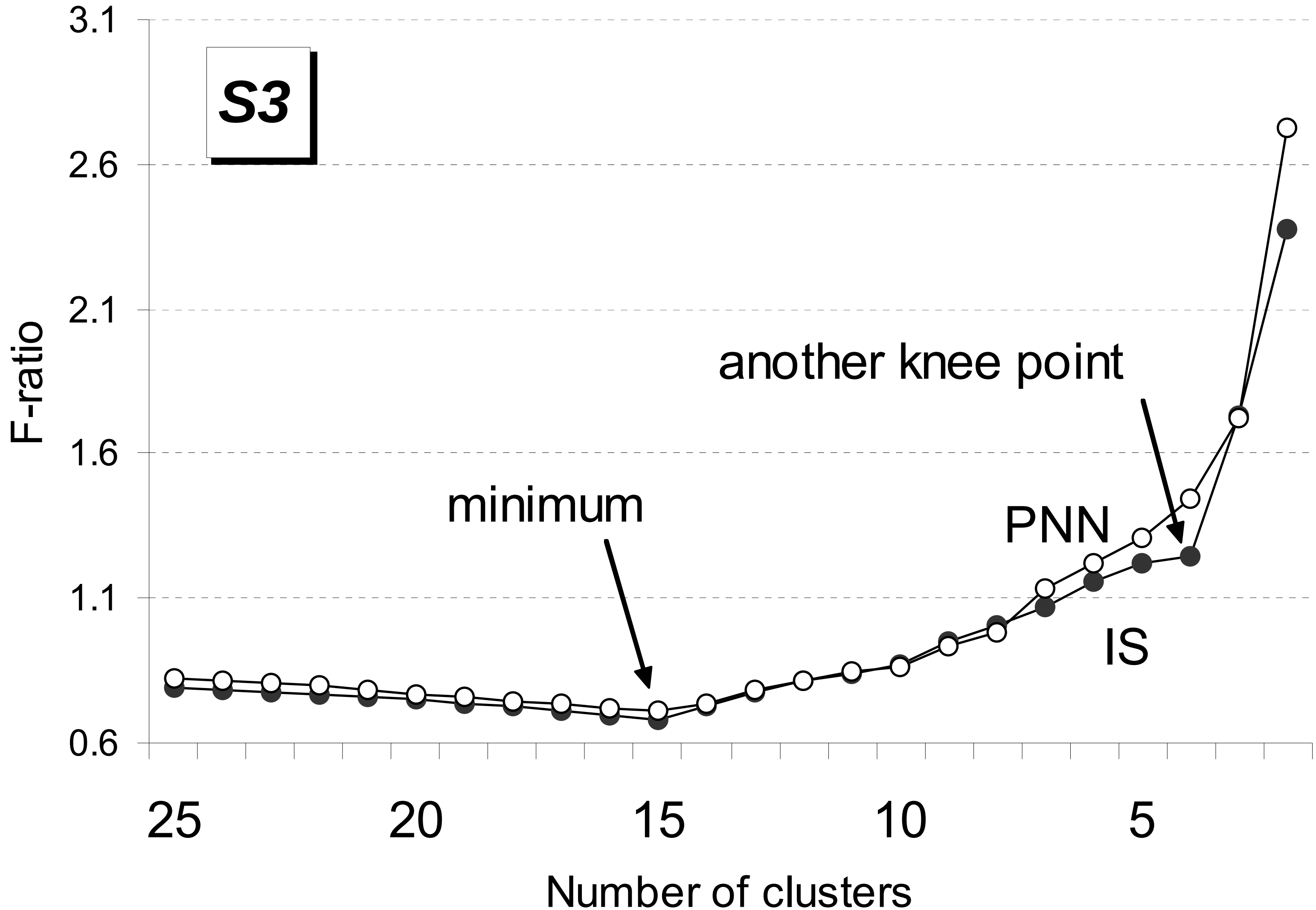

S3
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
F-ratio
minimum
PNN
IS
25
20
15
10
5
Number of clusters

Dataset S4

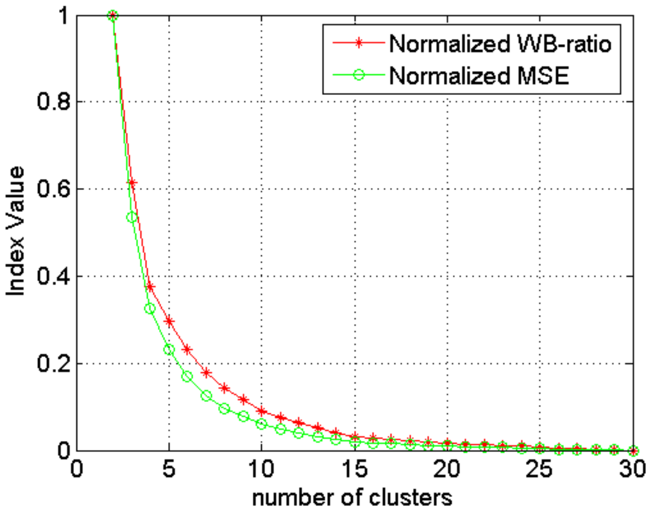


S4
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
F-ratio
PNN
IS
minimum at 16
minimum at 15
25
20
15
10
5
Number of clusters

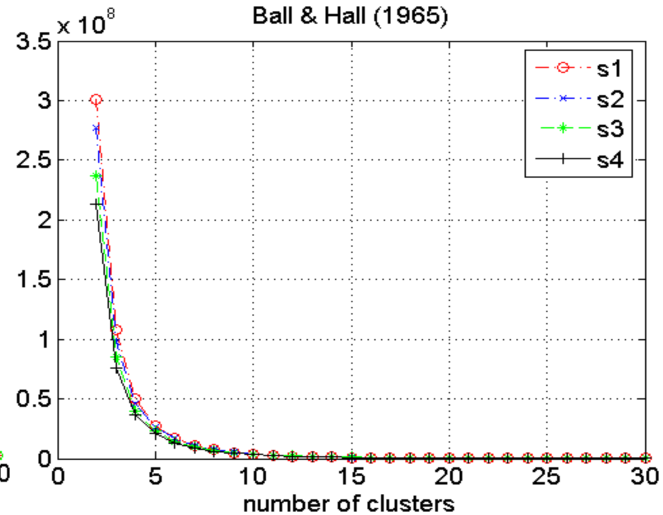
Extension for S3



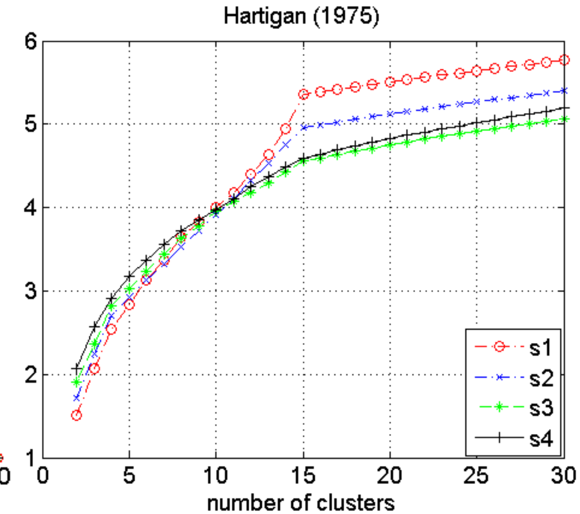
Sum-of-square based index



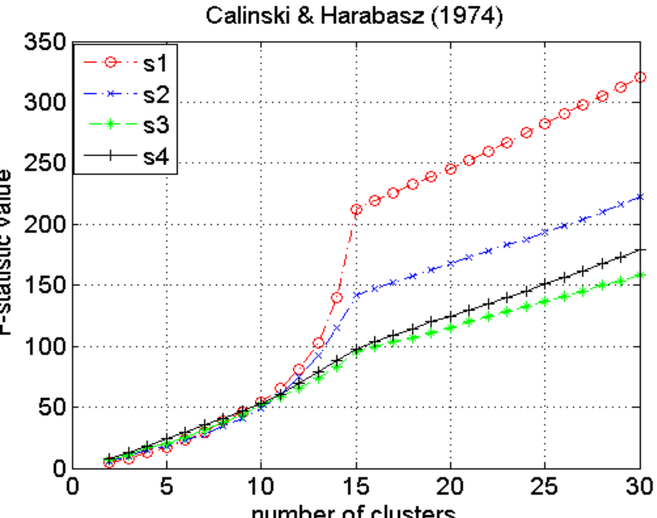
SSW / SSB & MSE



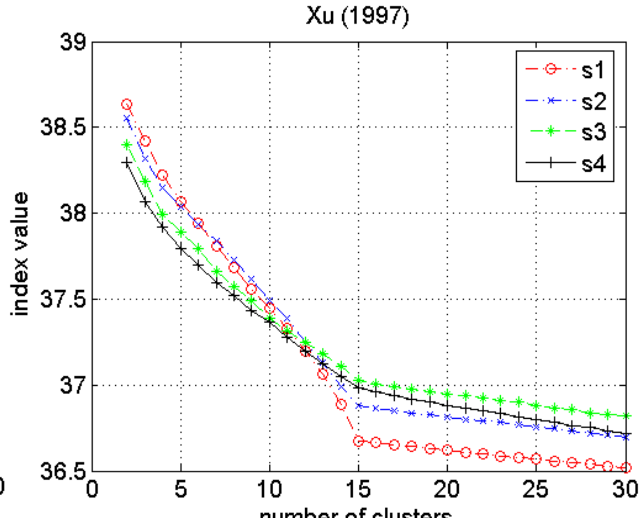
SSW / m



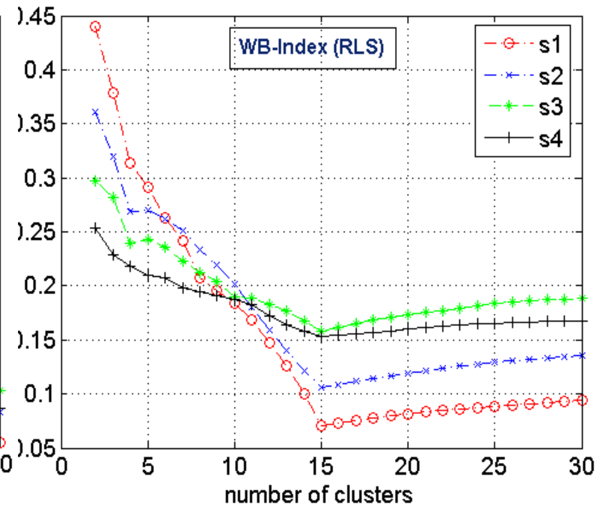
log(SSB/SSW)



$$\frac{SSB / m-1}{SSW / n-m}$$



$$d\log(\sqrt{SSW/(dn^2)})+\log(m)$$



m* SSW/SSB

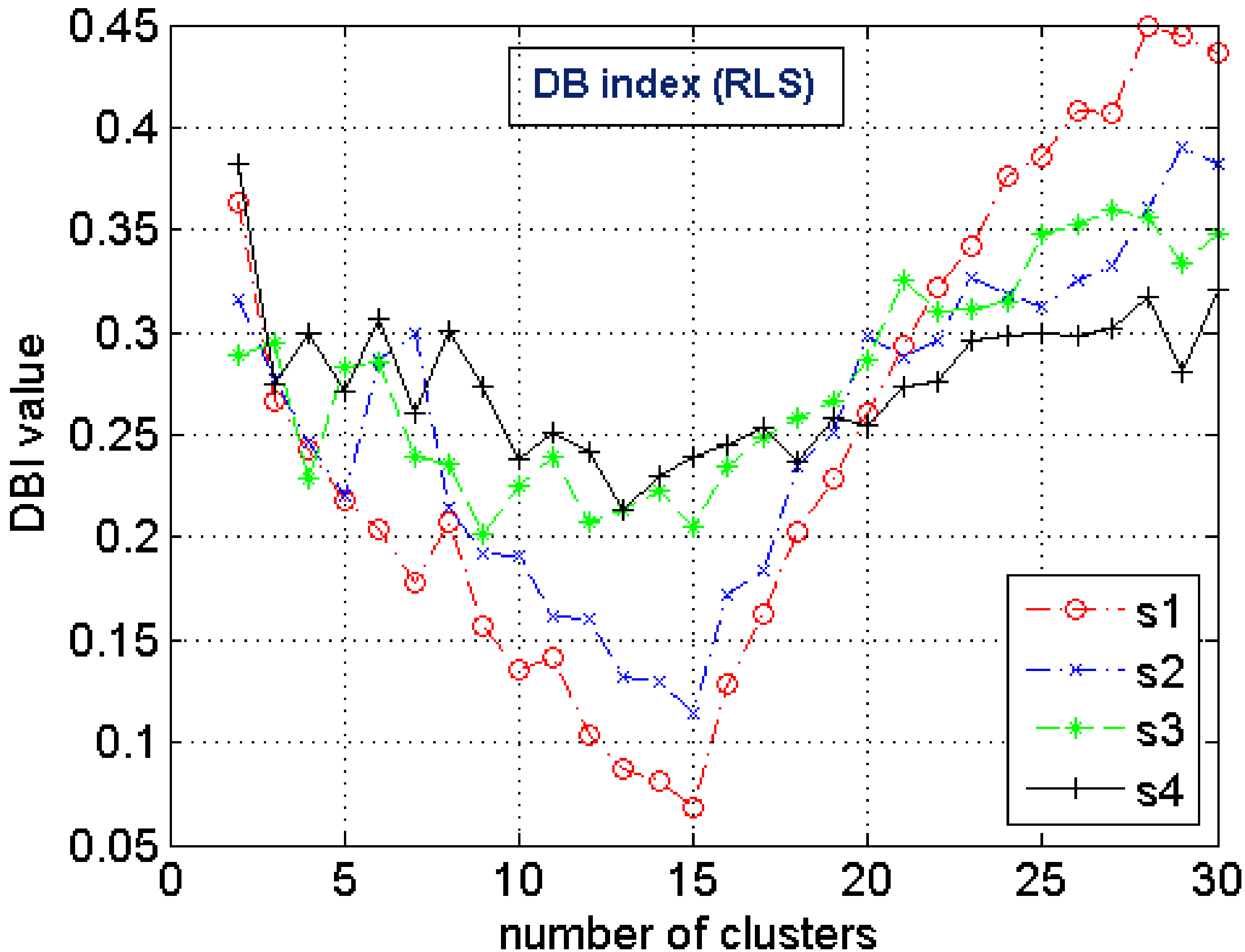
Davies-Bouldin index (DBI)

- Minimize intra cluster variance
- Maximize the distance between clusters
- Cost function weighted sum of the two:

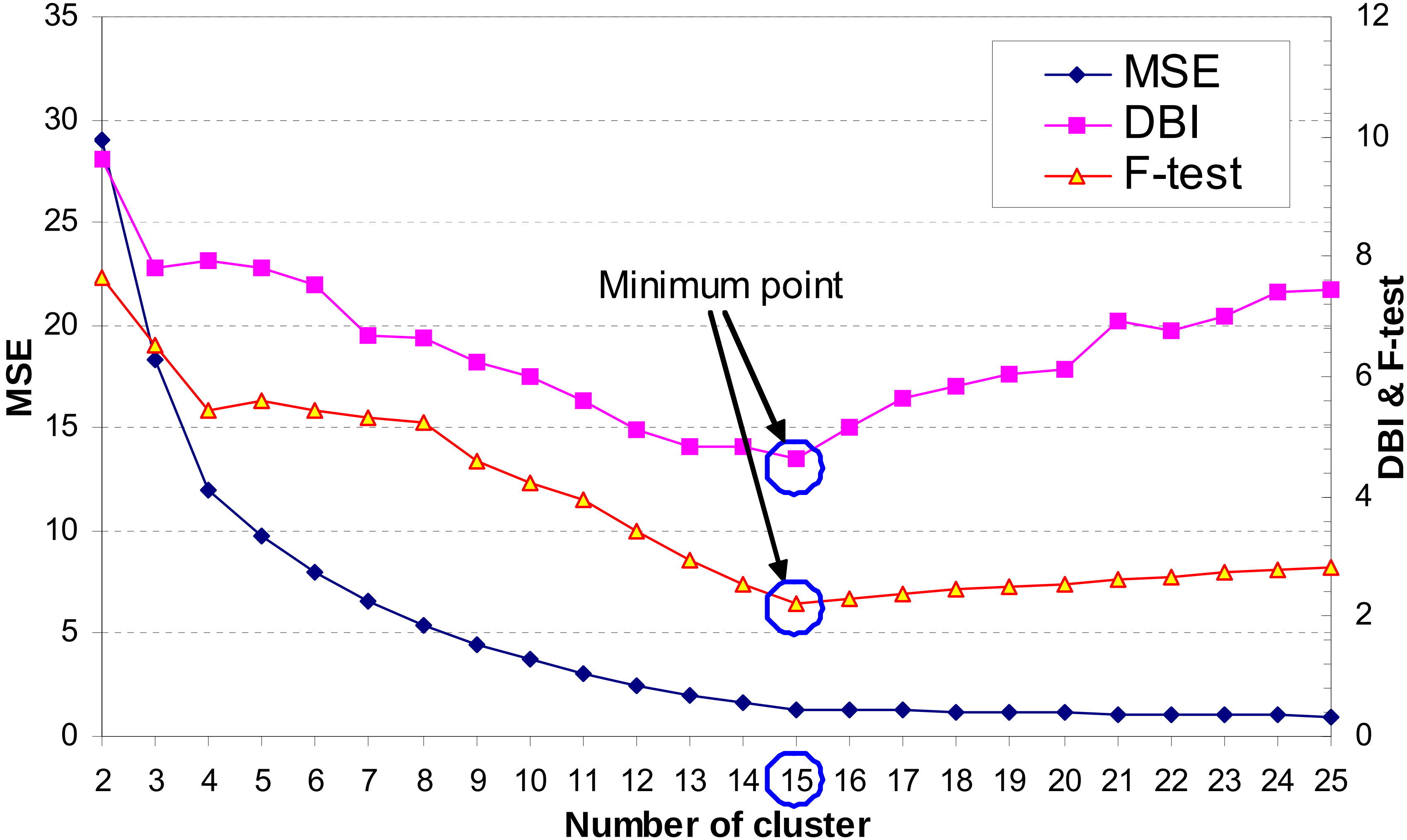
$$R_{j,k} = \frac{MAE_j + MAE_k}{d(c_j, c_k)}$$

$$DBI = \frac{1}{M} \sum_{j=1}^{M} \max_{j \neq k} R_{j,k}$$

Davies-Bouldin index (DBI)



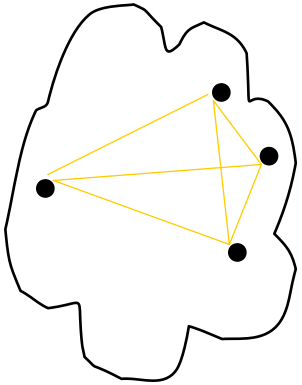
Measured values for S2



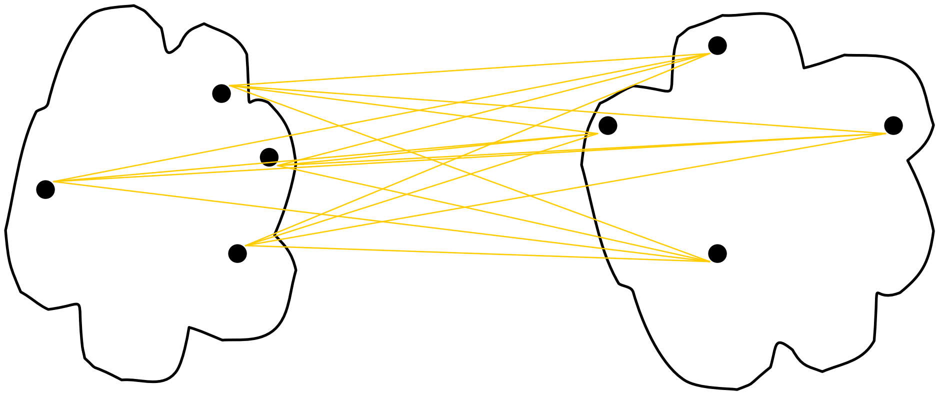
Silhouette coefficient

[Kaufman&Rousseeuw, 1990]

- **Cohesion**: measures how close objects are in a cluster
- **Separation**: measure how separated the clusters are



cohesion



separation

Silhouette coefficient

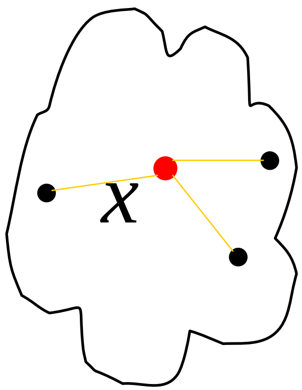
- *Cohesion* $a(x)$: average distance of x to all other vectors in the same cluster.
- *Separation* $b(x)$: average distance of x to the vectors in other clusters. Find the minimum among the clusters.
- *silhouette* $s(x)$:

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

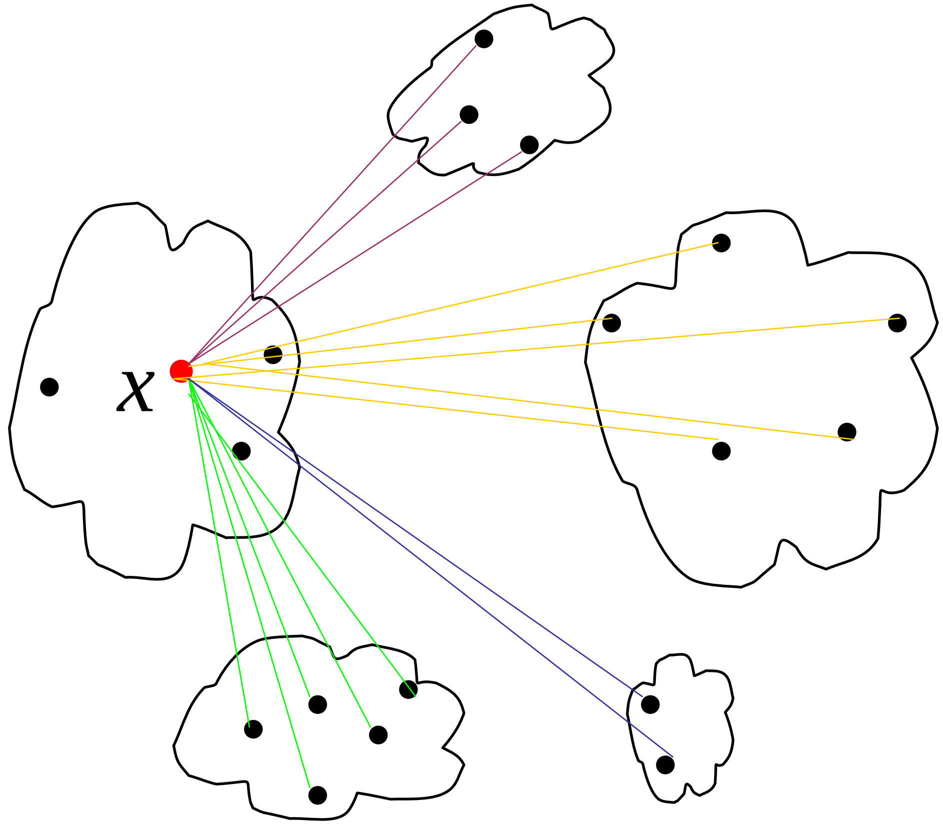
- $s(x) = [-1, +1]$: -1=bad, 0=indifferent, 1=good
- Silhouette coefficient (SC):

$$SC = \frac{1}{N}\sum_{i=1}^{N} s(x)$$

Silhouette coefficient (SC)



cohesion

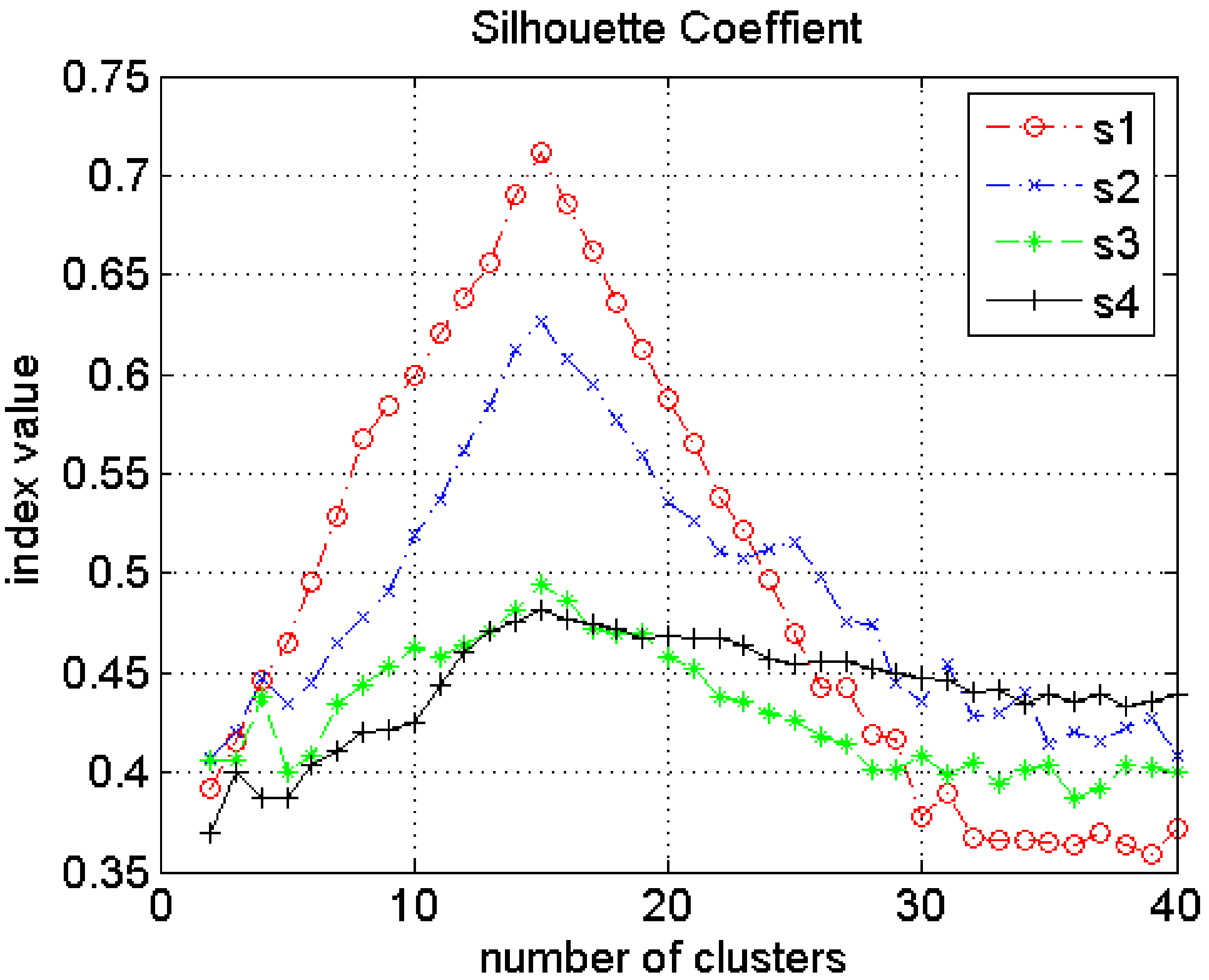


separation

$a(x)$: average distance in the cluster

b(x): average distances to others clusters, find minimal

Performance of SC



Bayesian information criterion (BIC)

Formula for GMM

$$BIC = L(\theta) - \frac{1}{2} m \log n$$

L(θ) -- log-likelihood function of all models;

n -- size of data set;

m -- number of clusters

Under spherical Gaussian assumption, we get :

Formula of BIC in partitioning-based clustering

$$BIC = \sum_{i=1}^{m} \left(n_i \log n_i - n_i \log n - \frac{n_i * d}{2} \log(2\pi) - \frac{n_i}{2} \log \Sigma_i - \frac{n_i - m}{2} \right) - \frac{1}{2} m \log n$$

d -- dimension of the data set

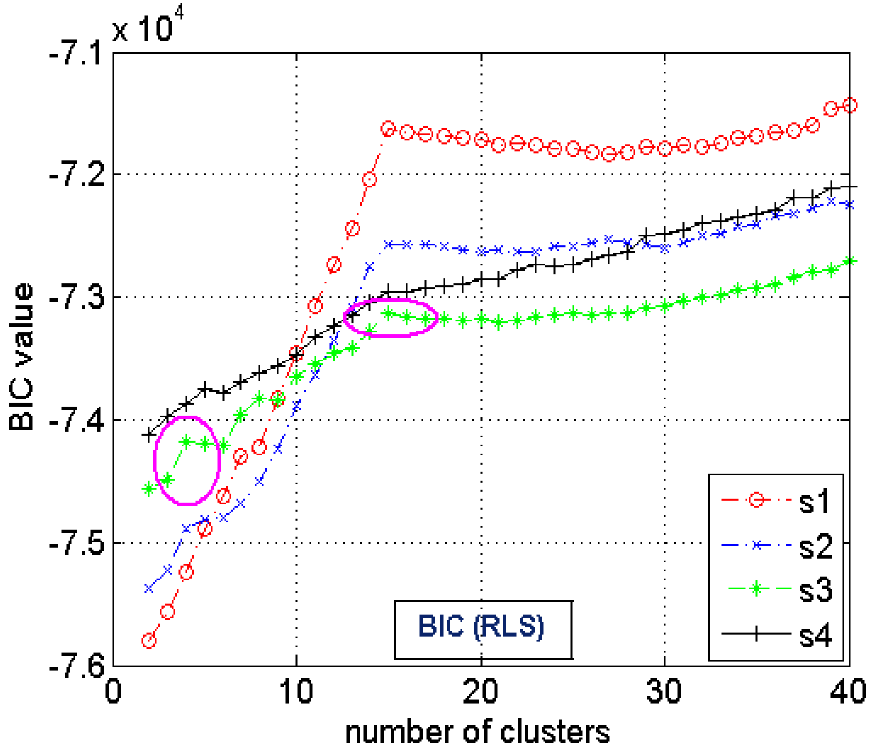
n_i -- size of the i^{th} cluster

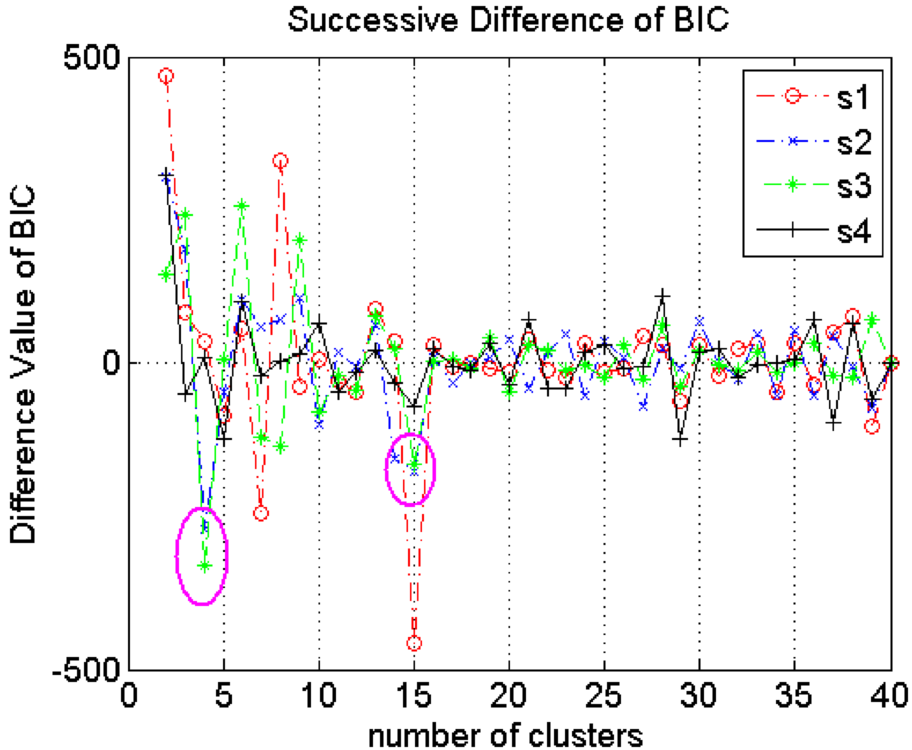
Σ_i -- covariance of i^{th} cluster

Knee Point Detection on BIC

Original BIC = *F*(*m*)

$SD(m) = F(m-1) + F(m+1) - 2 \cdot F(m)$





Internal indexes

Table B.1: Formulas for internal indexes

Name	Formula
SSW	$SSW = \frac{1}{N}\sum_{i=1}^{N} \left\Vert x_i - C_{p_i} \right\Vert^2$
SSB	$SSB = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1, j \neq i}^{M} \left\Vert C_i - C_j \right\Vert^2$
Calinski-Harabasz index	$CH = \frac{SSB/(M-1)}{SSW(N-M)}$
Hartigan	$H_M = (\frac{SSW_M}{SSW_{M+1}} - 1)(N - M - 1)$ $or : H_M = \log(SSB_M / SSW_M)$
Krzanowski-Lai index	$diff_M = (M-1)^{2/D} SSW_{M-1} - M^{2/D} SSW_M$ $KL_M = \lvert diff_M \rvert \ / \ \lvert diff_{M+1} \rvert$
Ball&Hall	$BH_M = SSW_M / M$
Xu-index	$Xu = D \log(\sqrt{SSW_M / (DN^2)}) + \log M$
Dunn's index	$Dunn = \sum_{i=1}^{M} \frac{\max(\left\Vert x_j - C_i \right\Vert^2)_{j \in C_i}}{}$
Davies&Bouldin index	$R_{ij} = \frac{S_i + S_j}{d_{ij}}, i \neq j$ $where : d_{ij} = \left\Vert C_i - C_j \right\Vert^2, S_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \left\Vert x_j - C_i \right\Vert^2$ $and, R_i = \max_{j=1,\ldots,M} R_{ij}, i = 1, \ldots, M$ $DBI = \frac{1}{M} \sum_{i=1}^{M} R_i$

Internal indexes

Silhouette Coefficients	$a(x_i) = \frac{1}{n_m - 1} \sum_{j=1, j \neq i}^{n_m} \Vert x_i - x_j \Vert^2_{x_i, x_j \in C_m}$ $b(x_i) = \min_t \{\frac{1}{n_t} \sum_{j \in C_t} \Vert x_i - x_j \Vert^2\}_{x_i \notin C_t}$ $s(x_i) = \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))}$ $SC = \frac{1}{N} \sum_{i=1}^{N} s(x_i)$ $b(x_i) = \min \{\sum_{t \neq m} \Vert C_t - C_m \Vert^2\}_{x_i \notin C_t} (SC'2008)$
RMSSTD	$RMSSTD = \frac{\sum_{\substack{k=1,\dots,M \\ d=1,\dots,D}} \sum_{i=1}^{n_{kd}} (x_i - \overline{x^d})^2}{\sum_{\substack{k=1,\dots,M \\ d=1,\dots,D}} (n_{kd} - 1)}$
R-square	$RS = \frac{SST - SSW}{SST} = \frac{\sum_{d \; 1,\dots,D} \sum_{i=1}^{n_d} (x_i - \overline{x^d})^2 - \sum_{\substack{k=1,\dots,M \\ d=1,\dots,D}} \sum_{i=1}^{n_{kd}} (x_i - \overline{x^d})^2}{\sum_{d=1,\dots,D} \sum_{i=1}^{n_d} (x_i - \overline{x^d})^2}$
Bayesian Information Criterion	$BIC = L * N - \frac{1}{2} M(D+1) \sum_{i=1}^{M} \log(n_i)$
Xie-Beni	$XB = \frac{\sum_{i=1}^{N} \sum_{k=1}^{M} u_{ik}^2 \Vert x_i - C_k \Vert^2}{N \min_{t \neq s} \{\Vert C_t - C_s \Vert^2\}}$
Partition Coefficient	$PC = \sum_{i=1}^{N} \sum_{k=1}^{M} u_{ik}^2 / N$
Partition Entropy	$PE = -(\sum_{i=1}^{N} \sum_{k=1}^{M} u_{ik} \log(u_{ik})) / N$

Soft partitions

Comparison of the indexes

K-means

Data Set / ValidityCri	R15	A7	s1	s2	s3	s4	D31	birch1	Iris	wine	Control	Image	wdbc	yeast	bridge	zernike
Ball&Hall (knee min)	15	25	30	57	58	63	53	108	10	11	21	40	22	37	62	38
Calinski-Harabsz (max)	24	4	15	20	6	16(2)	39	198	2	2	2	2	2	2	2	2
Hartigan (diff knee min)	15	25	30	57	58	49	53	108	10	10	20	40	21	28	62	38
Wb-index (min)	24	28	15	20	19	20	39	198	2	2	2	3	2	3	2	2
Krzanowski-Lai index (mi	18	19	15	12	63	50	25	86	9	2	2	23	2	29	33	39
Xu-index (min)	24	28	15	20	17	17	39	117	12	13	24	48	23	38	64	44
Dunn (max)	5	23	4	15	10	15	53		12	2	19	3	3	2	38	36
DBI (min)	24	4	15	13	4	13	23	93	2	2	2	2	2	23	2	2
SC (max)	17	3	15	17	13	17	34	103	2	2	2	2	2	2	2	6
SCI * (max)	NA	4	15	17	17	17	NA	95	2	2	2	2	2	2	NA	NA
XieBeni (min)	3	4	15	4	4	4	4	4	2	2	2	2	2	2	2	5
ORI_IBIC (first max)	NA	4	NA	20	4	5	NA(4)	26	NA(4)	NA	2	NA(3)	NA	NA	NA	NA
Angle-based	NA	4	8	15	4	3	3	199	8	NA	2	20	NA	NA	NA	NA
DiffBIC	NA	4	11	15	15	5	NA	4	NA	NA	NA	NA	NA	NA	NA	NA
# of clusters	15	7	15	15	15	15	31	100	3	3	6	7	2	10	NA	10

Comparison of the indexes

Random Swap

ValidityCri \ Data Set	R15	A7	s1	s2	s3	s4	D31	birch1	Iris	wine	Control	Image	wdbc	yeast	bridge	zernike
Ball&Hall (knee)	24	26	69	68	66	68	51	99	9	11	21	45	18	8	62	42
Calinski-Harabsz (max)	15	25	15	15	2	15	32	199	2	2	2	2	2	4	2	2
Hartigan (max knee)	23	26	69	68	66	68	51	99	9	11	21	40	18	8	62	36
Wb-index (min)	15	25	15	15	15	15	32	199	2	2	2	2	2	4	2	4
Krzanowski-Lai index (mi	15	14	15	15	15	15	16	84	10	2	9	20	2	19	35	2or36
Xu-index (min)	15	25	15	15	15	15	32	102	12	13	24	48	23	38	64	44
Dunn (max)	8	25	15	15	63	20	31		2	2	19	33	21	36	51	26
DBI (min)	8	4	15	16	15	15	30	99	2	2	2	2	2	5	2	2
SC (max)	15	4	15	15	15	15	30	102	2	2	2	2	2	4	2	6
SCI * (max)	8	4	15	15	15	15	31	102	2	2	2	2	2	8	2	21
XieBeni (min)	15	3	15	15	4or15	15	30	99	2	10	6	2	2	4	2	25
ORI_IBIC (max knee)	8	4	15	15	15	15	31	4	5	2	NA	2	2	2	2	2
Angle-based	8	4	15	4	4	5	11	4	11	12	14	23	10	25	47	36
DiffBIC	8	4	14	15	15	5	12	4	2	2	NA	1	1	1	1	1
# of clusters	**15**	**7**	**15**	**15**	**15**	**15**	**31**	**100**	**3**	**3**	**6**	**7**	**2**	**10**	**NA**	**10**

Part III:

Stochastic complexity for binary data

Stochastic complexity

- Principle of minimum description length (MDL): find clustering *C* that can be used for describing the data with minimum information.
- Data = Clustering + description of data.
- Clustering defined by the centroids.
- Data defined by:
  - which cluster (partition index)
  - where in cluster (difference from centroid)

Solution for binary data

$$SC = \sum_{j=1}^{M} n_j \sum_{i=1}^{d} h\left(\frac{n_{ij}}{n_j}\right) + \sum_{j=1}^{M} -n_j \log\left(\frac{n_j}{N}\right) + \frac{d}{2}\sum_{j=1}^{M} \log \max(1, n_j)$$

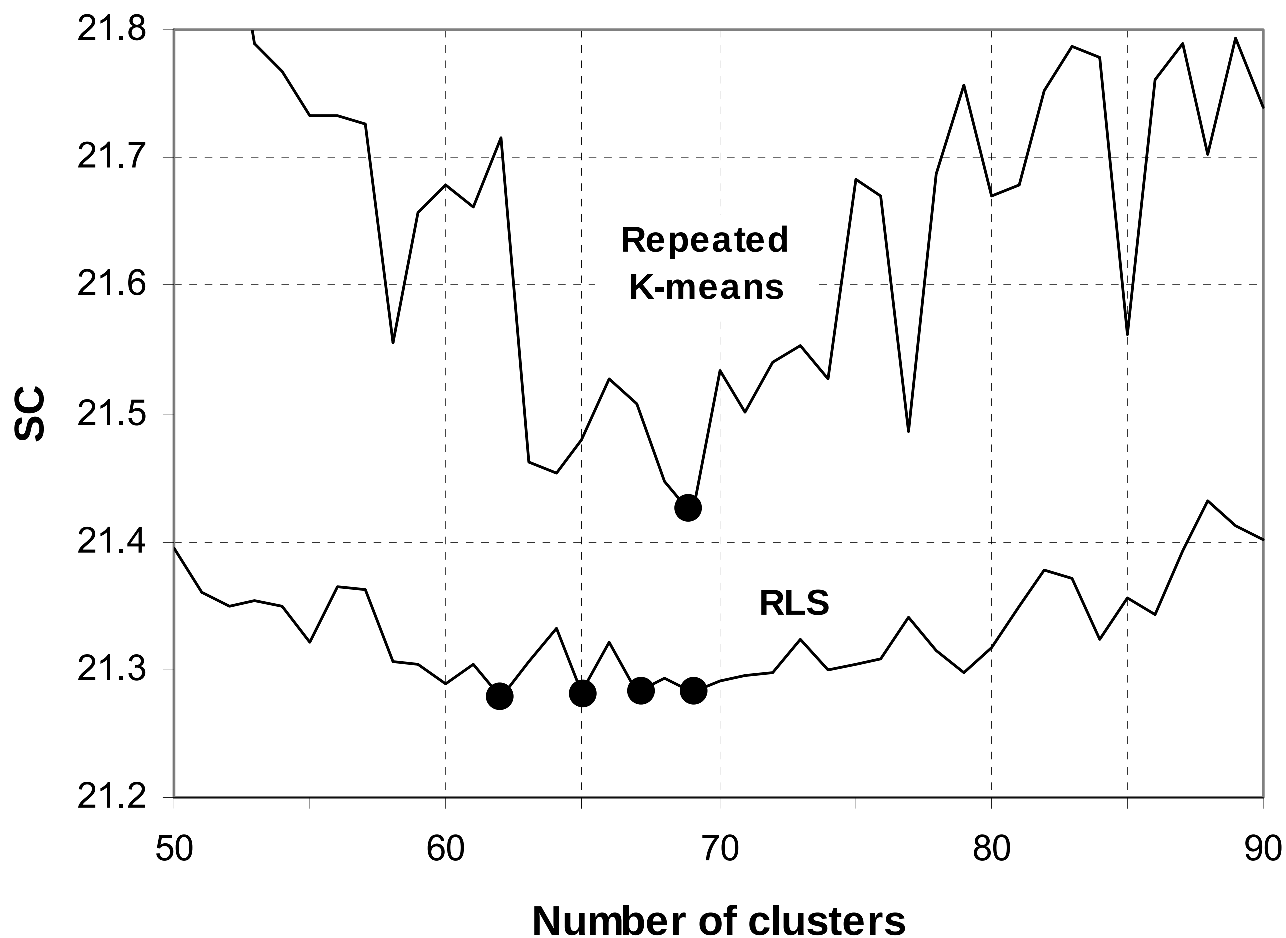
where

$$h(p) = -p\log(p) - (1-p)\log(1-p)$$

This can be simplified to:

$$SC \approx \sum_{j=1}^{M} n_j \sum_{i=1}^{d} h\left(\frac{n_{ij}}{n_j}\right) + \sum_{j=1}^{M} -n_j \log n_j + \frac{d}{2}\sum_{j=1}^{M} \log \max(1, n_j)$$

Number of clusters by stochastic complexity (SC)

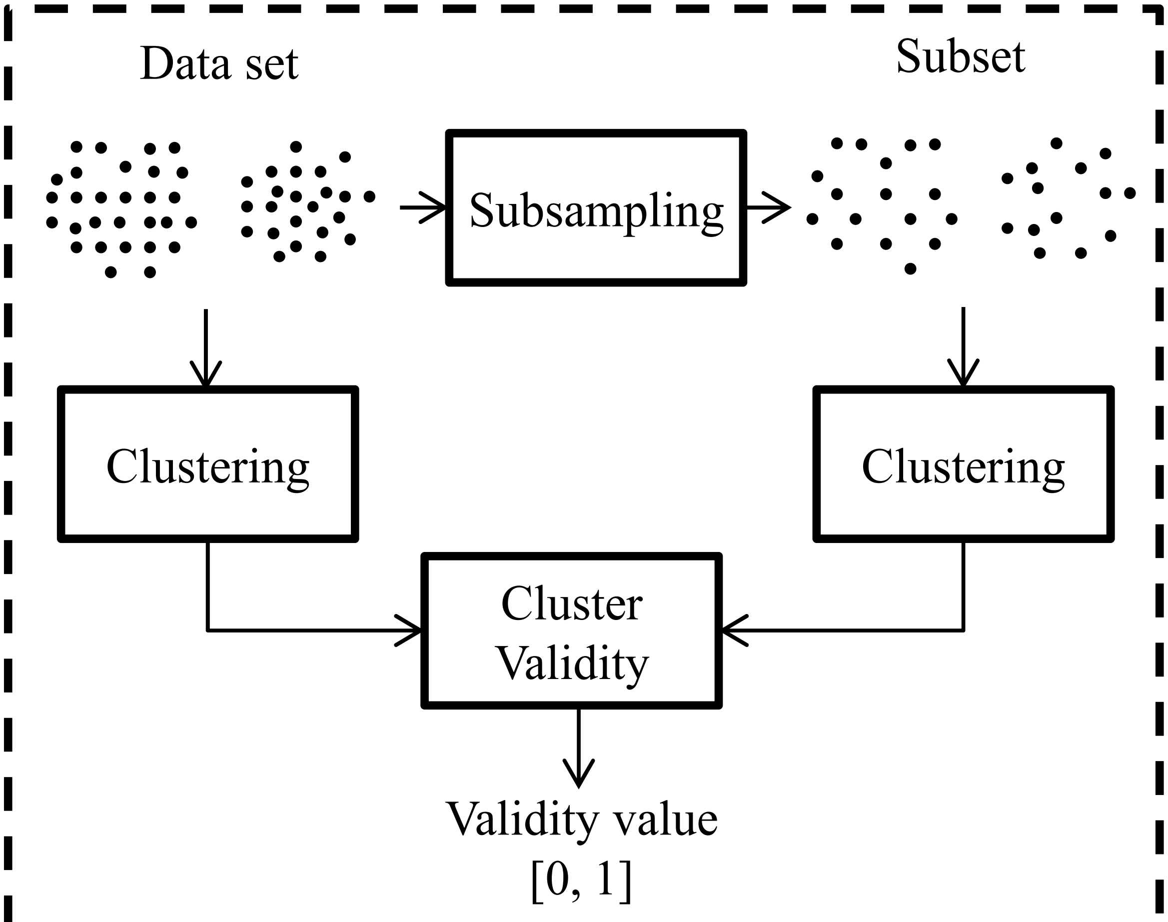


Part IV:

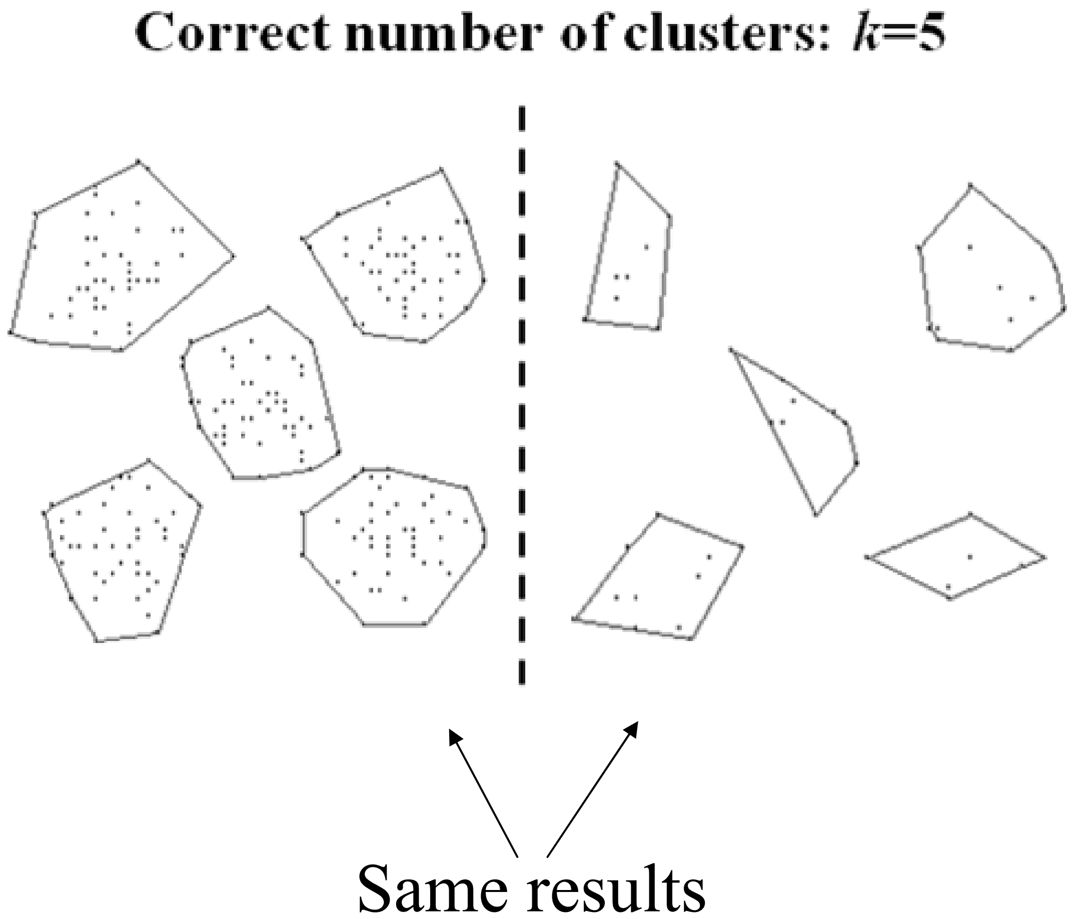
Stability-based approach

Cross-validation

Compare clustering of full data against sub-sample

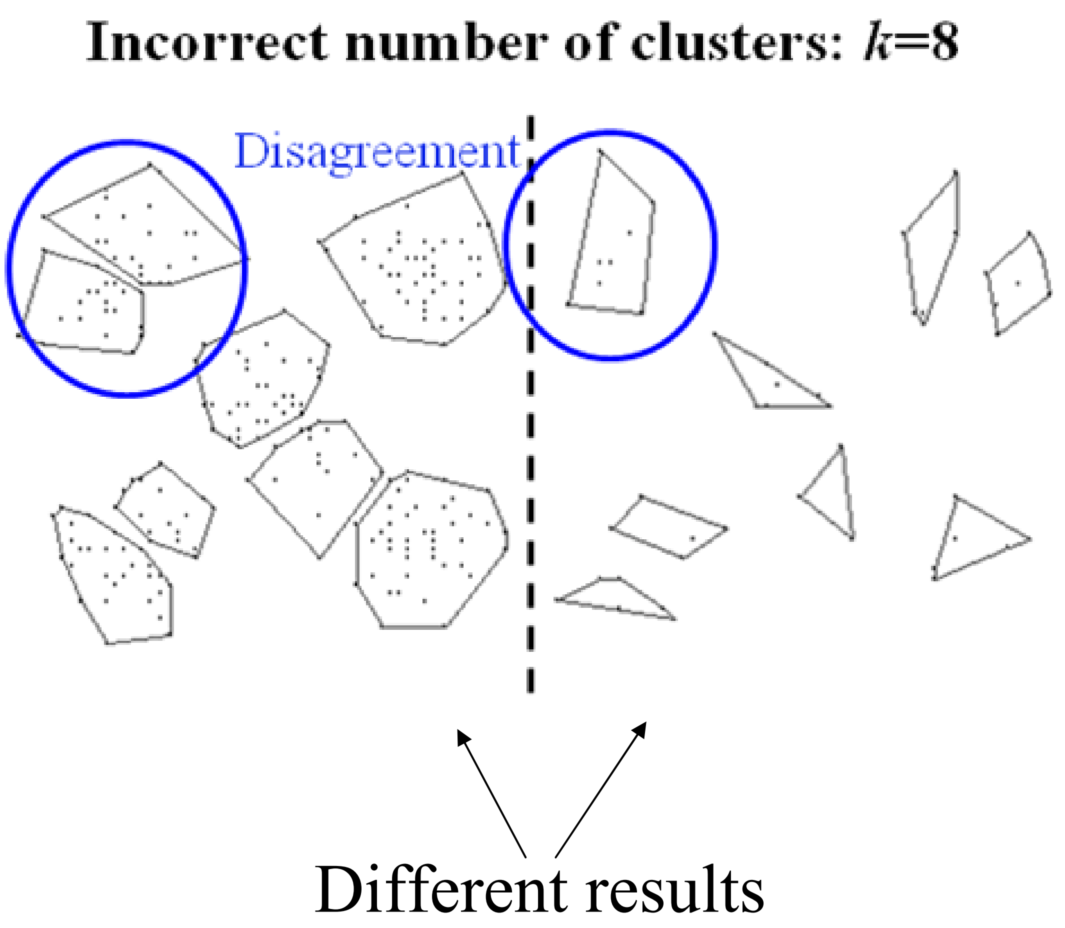


Cross-validation: Correct



Cross-validation

Incorrect



Stability approach in general

1. Add randomness
2. Cross-validation strategy
3. Solve the clustering
4. Compare clustering

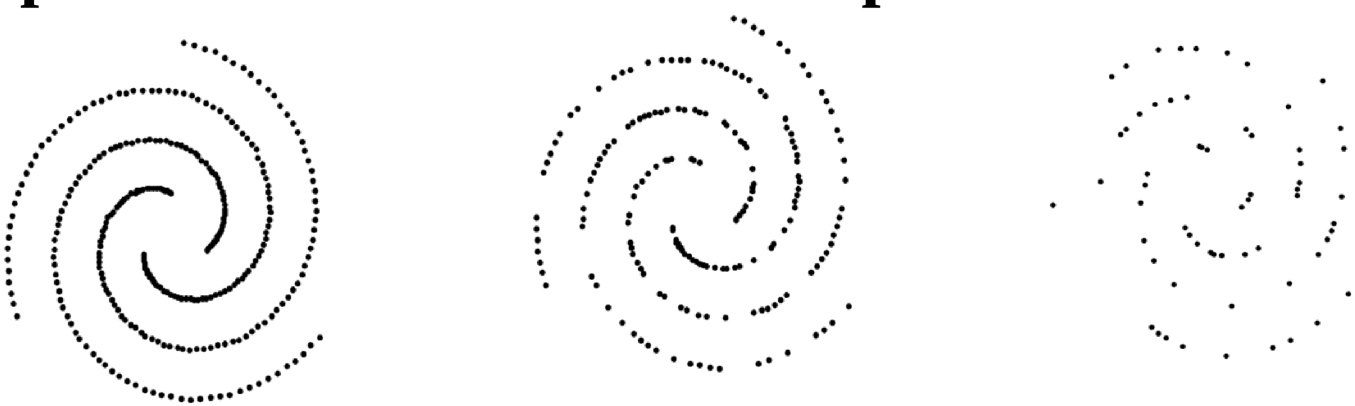
Adding randomness

- Three choices:
    1. Subsample
    2. Add noise
    3. Randomize the algorithm
- What subsample size?
- How to model noise and how much?
- Use k-means?

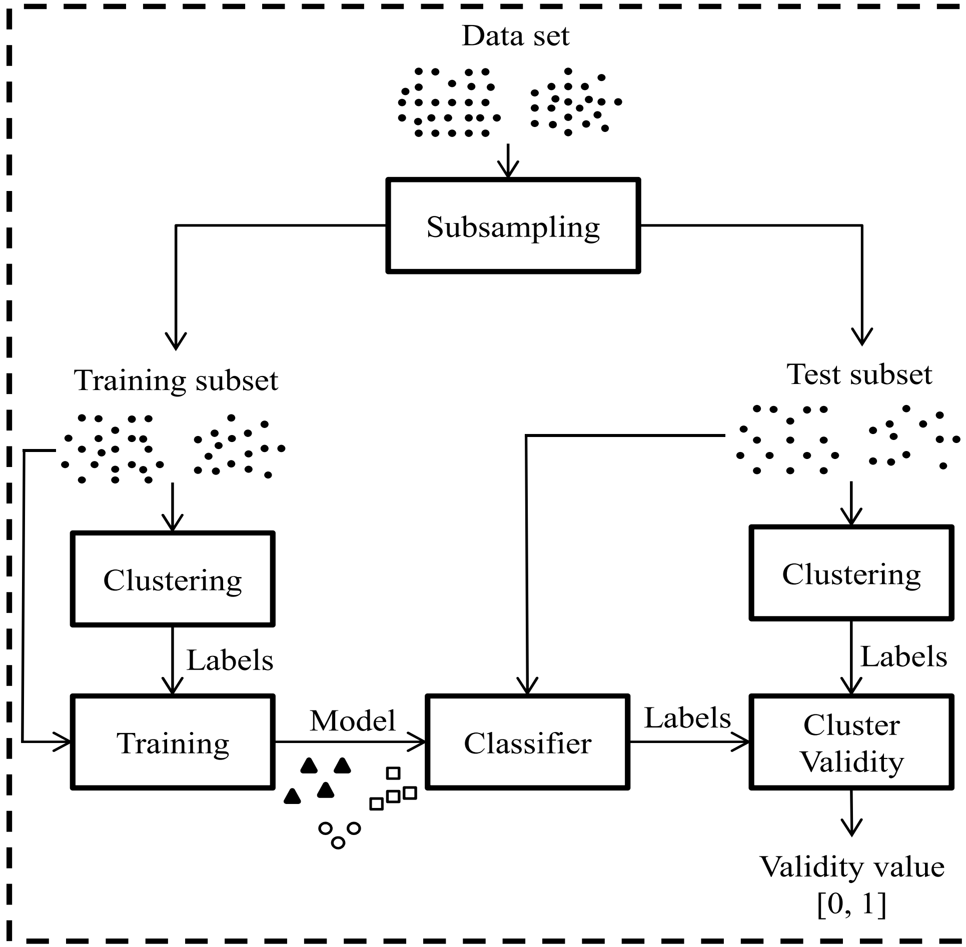
Sub-sample size

- Too large (80%): same clustering always
- Too small (5%): may break cluster structure
- Recommended 20-40%

Spiral dataset **60% subsample** **20% subsample**



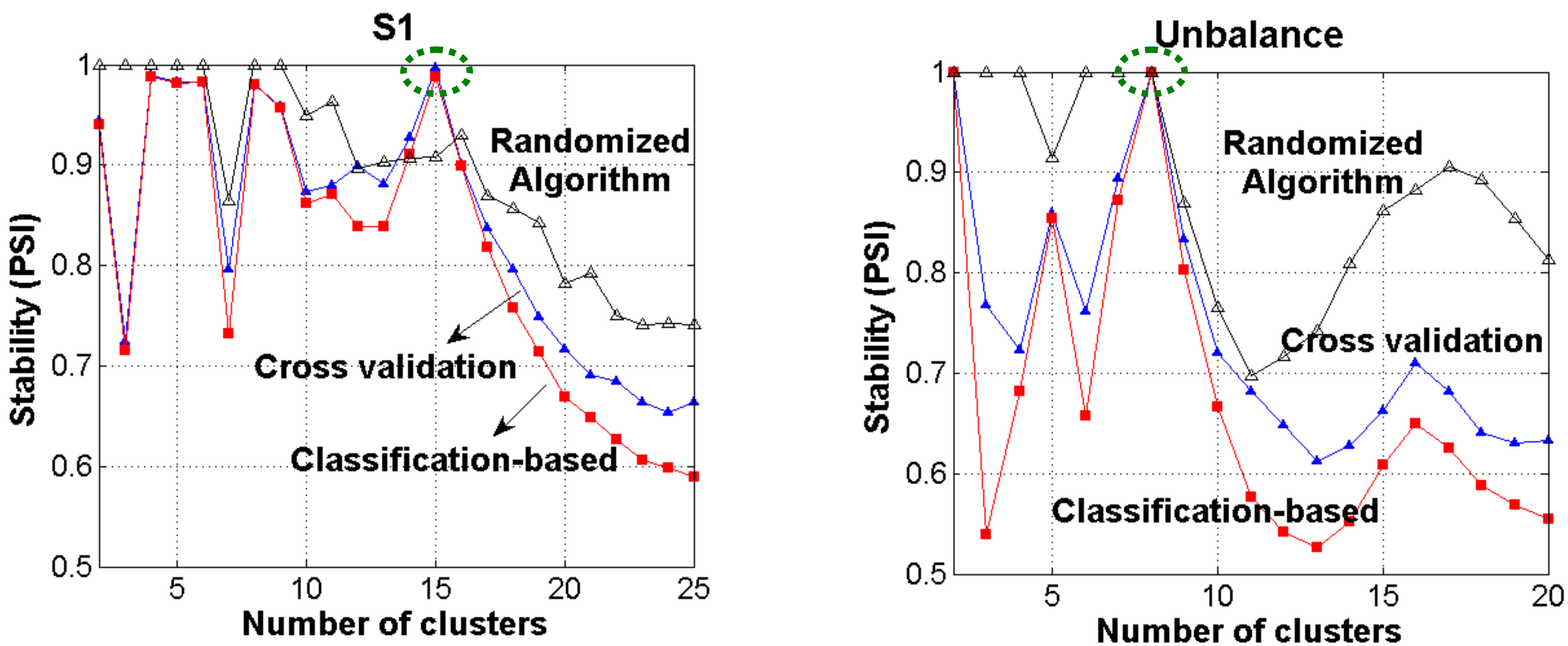
Classification approach



Does not really add anything more.
Just makes process more complex.

Comparison of three approaches

- Cross-validation works ok
- Classification also ok
- Randomizing algorithm fails



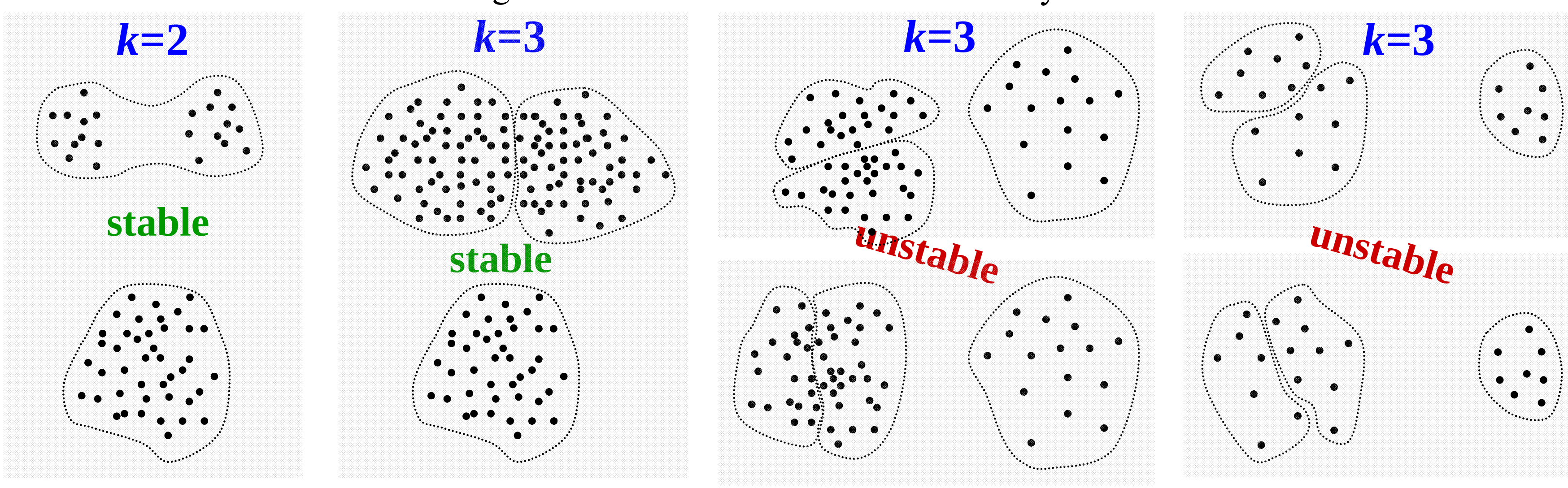
Problem

Stability can also come from other reasons:

- Different cluster sizes
- Wrong cluster model

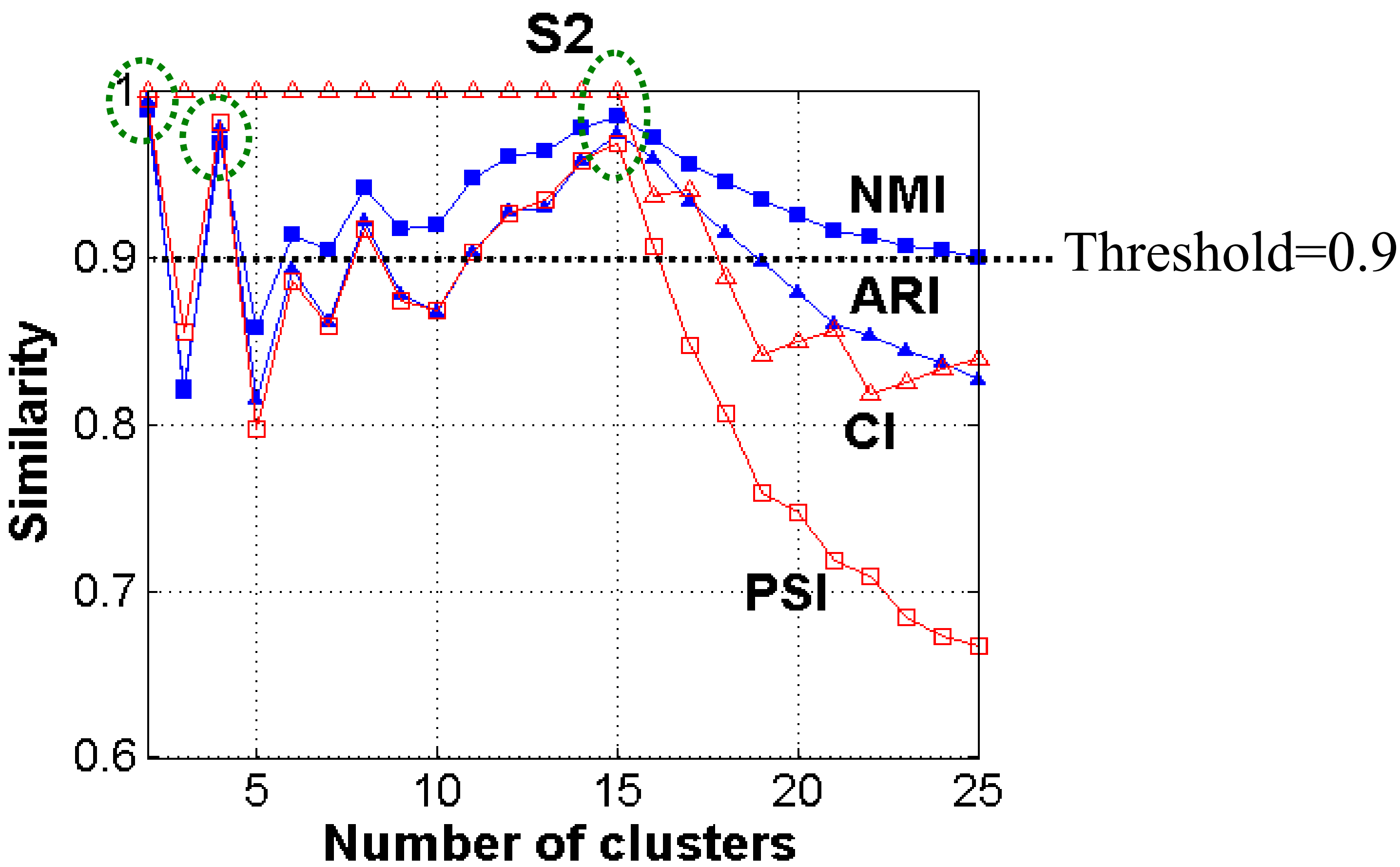
Happens when *k*<*k***

Too few clusters different size | **Too many clusters** wrong model | **Too many clusters** different density | **Too many clusters** different size



Solution

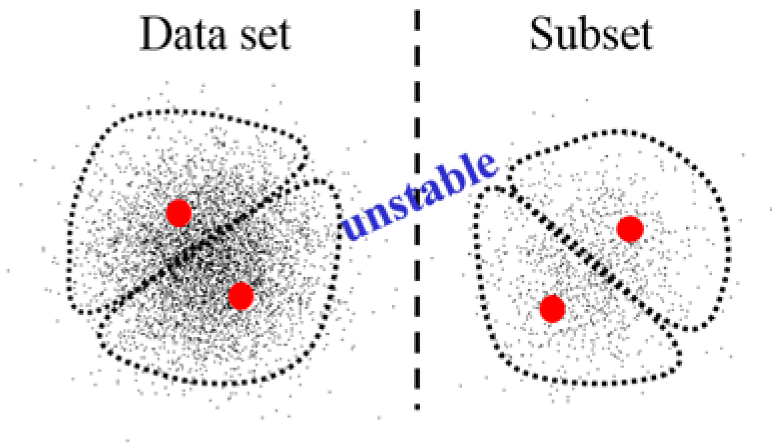
Instead of selecting *k* with maximum stability, select <u>last</u> *k* with stable result.



Effect of cluster shapes

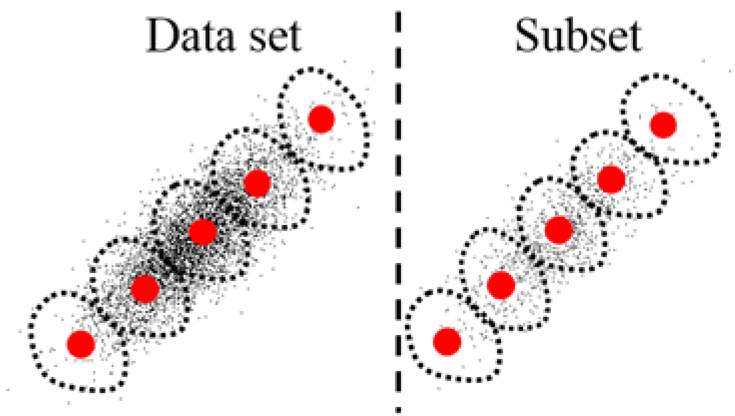
Correct model:

- Works ok.



Wrong model:

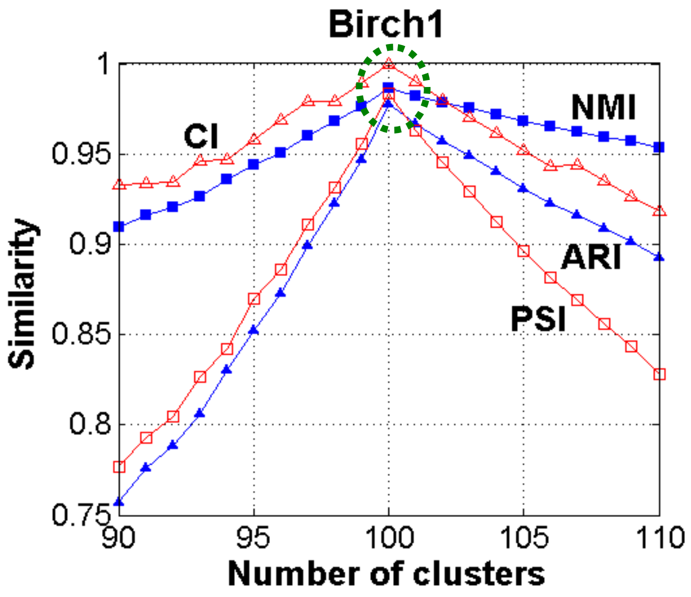
- Elliptical cluster
  Minimizing TSE would
  find 5 spherical clusters

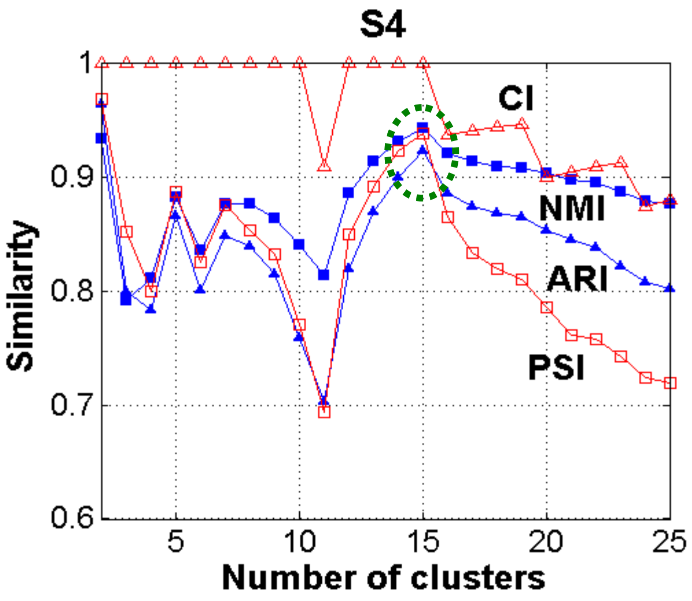


Which external index?

Does not matter much

- This is not: RI
- These all ok: ARI, NMI, PSI, NVD, CSI
- CI cares only allocation: sometimes too rough.

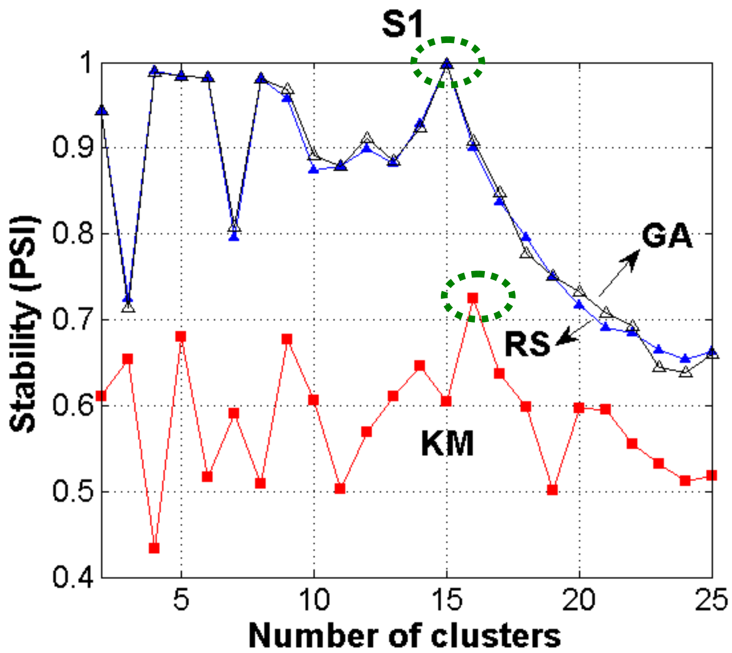


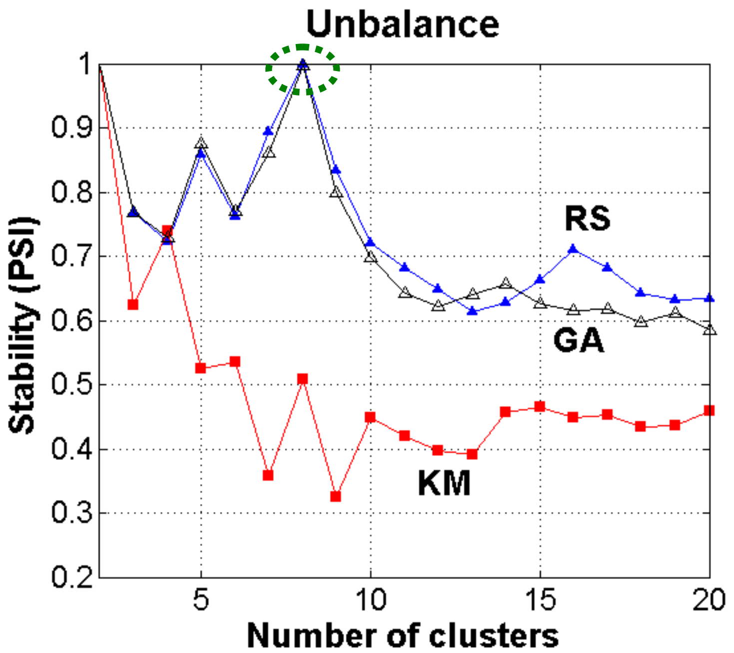


Does algorithm matter?

Yes it does.

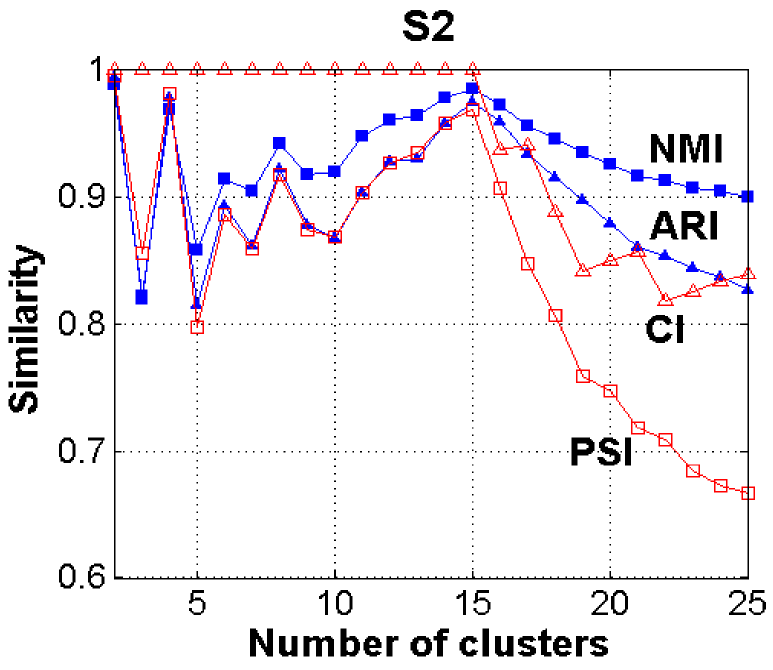
- Ok: Random Swap (RS) and Genetic Algorithm (GA)
- Not: K-means (KM)





Summary

- The choice of the cross-validation strategy not critical
- Last stable clustering instead of global maximum
- The choice of external index is not critical
- Good clustering algorithm required (RS or GA)



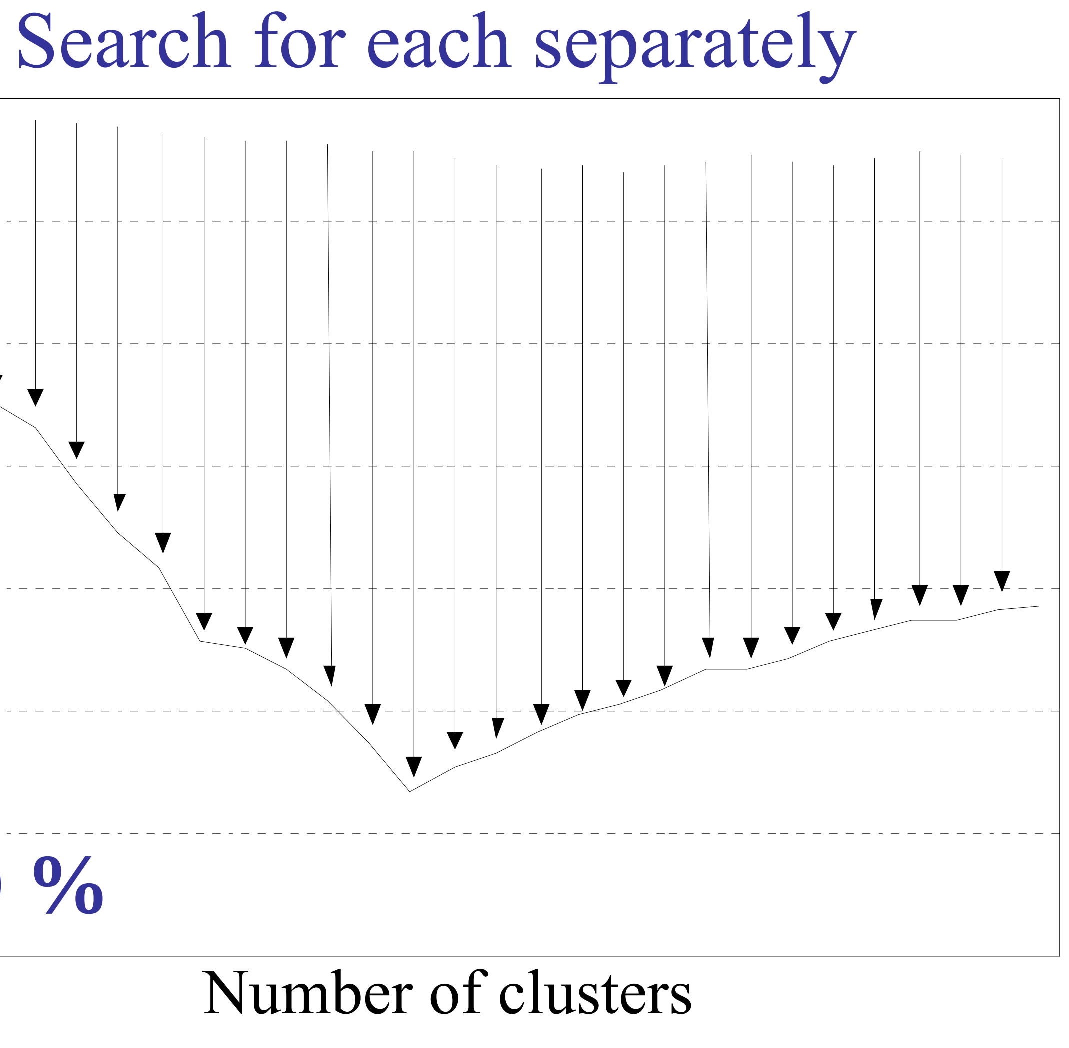
Part V:

Efficient implementation

Strategies for efficient search

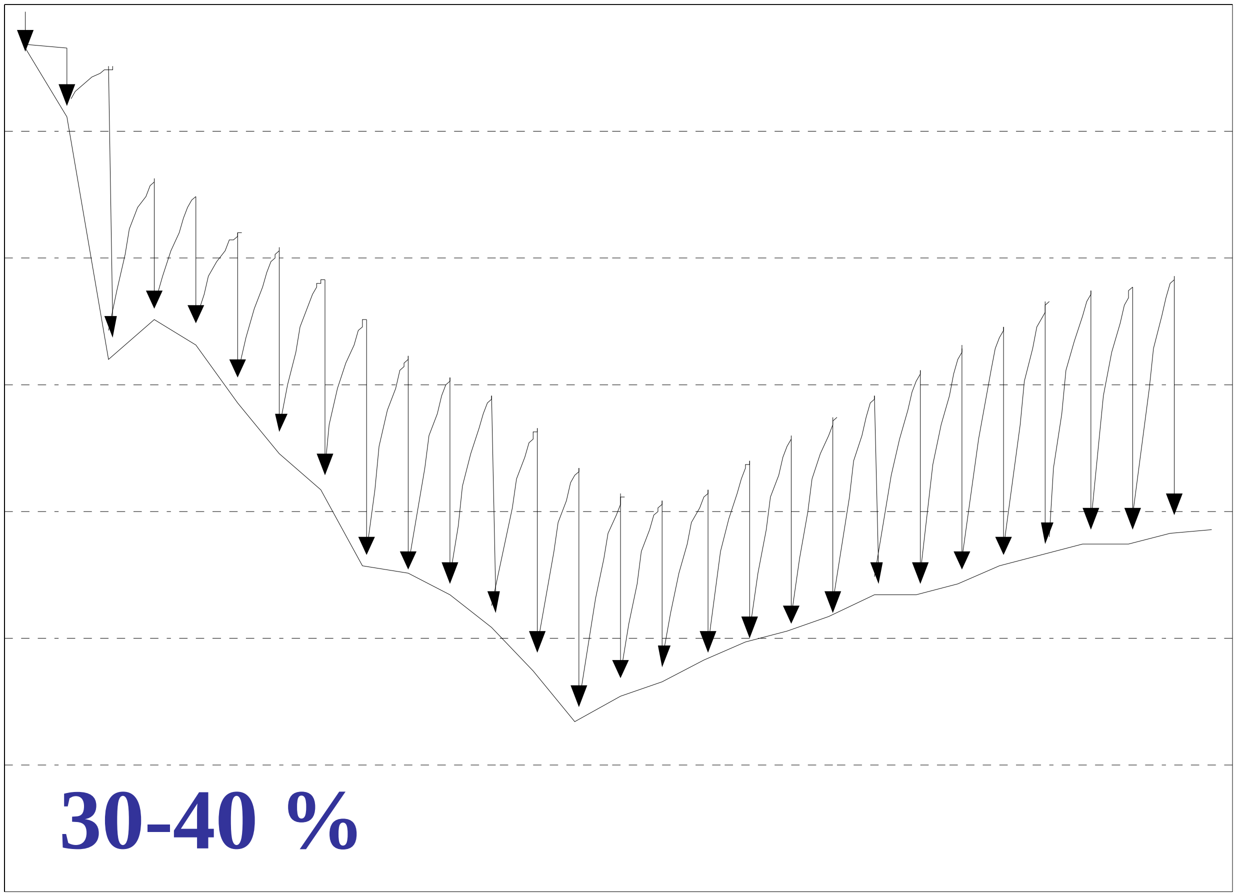
- **Brute force**: solve clustering for all possible number of clusters.
- **Stepwise**: as in brute force but start using previous solution and iterate less.
- **Criterion-guided search**: Integrate cost function directly into the optimization function.

Brute force search strategy



Stepwise search strategy

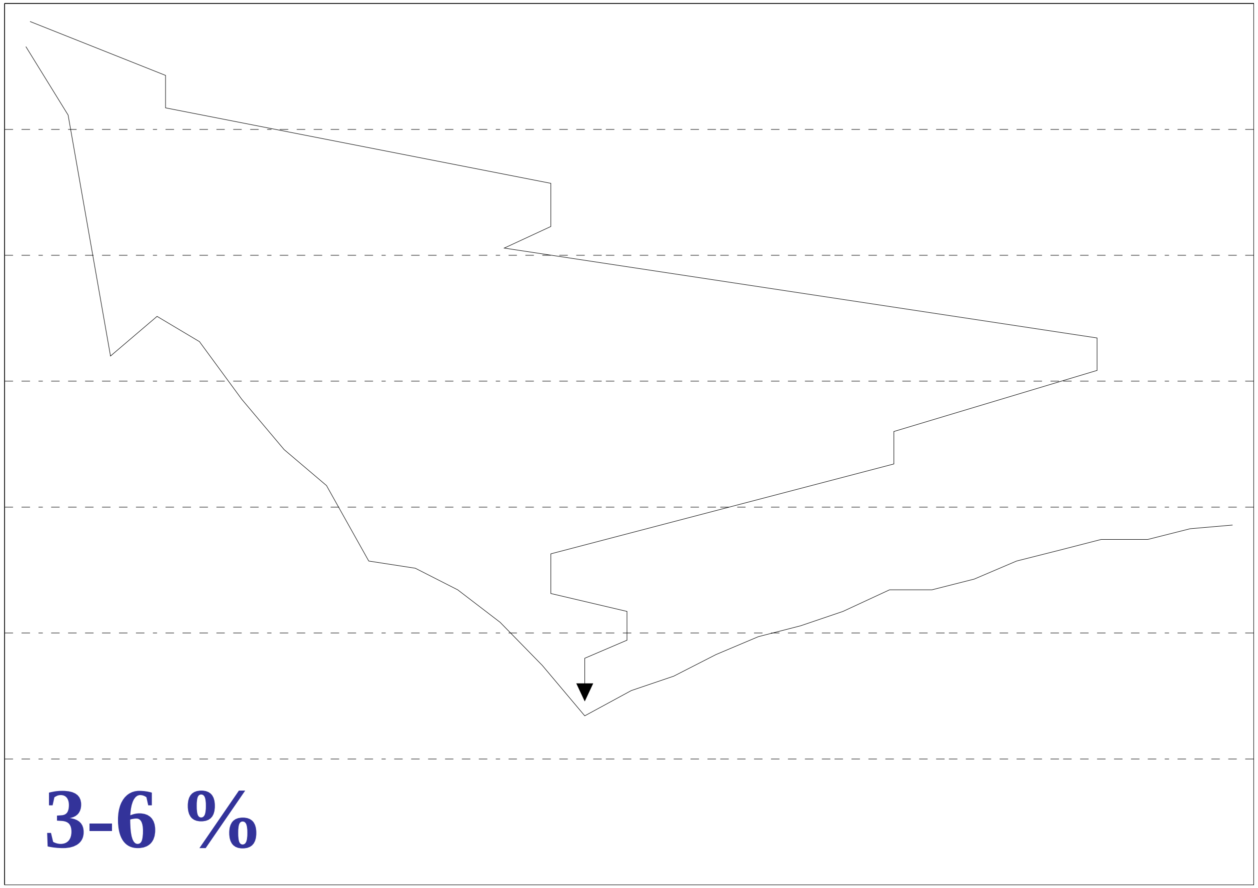
Start from the previous result



Number of clusters

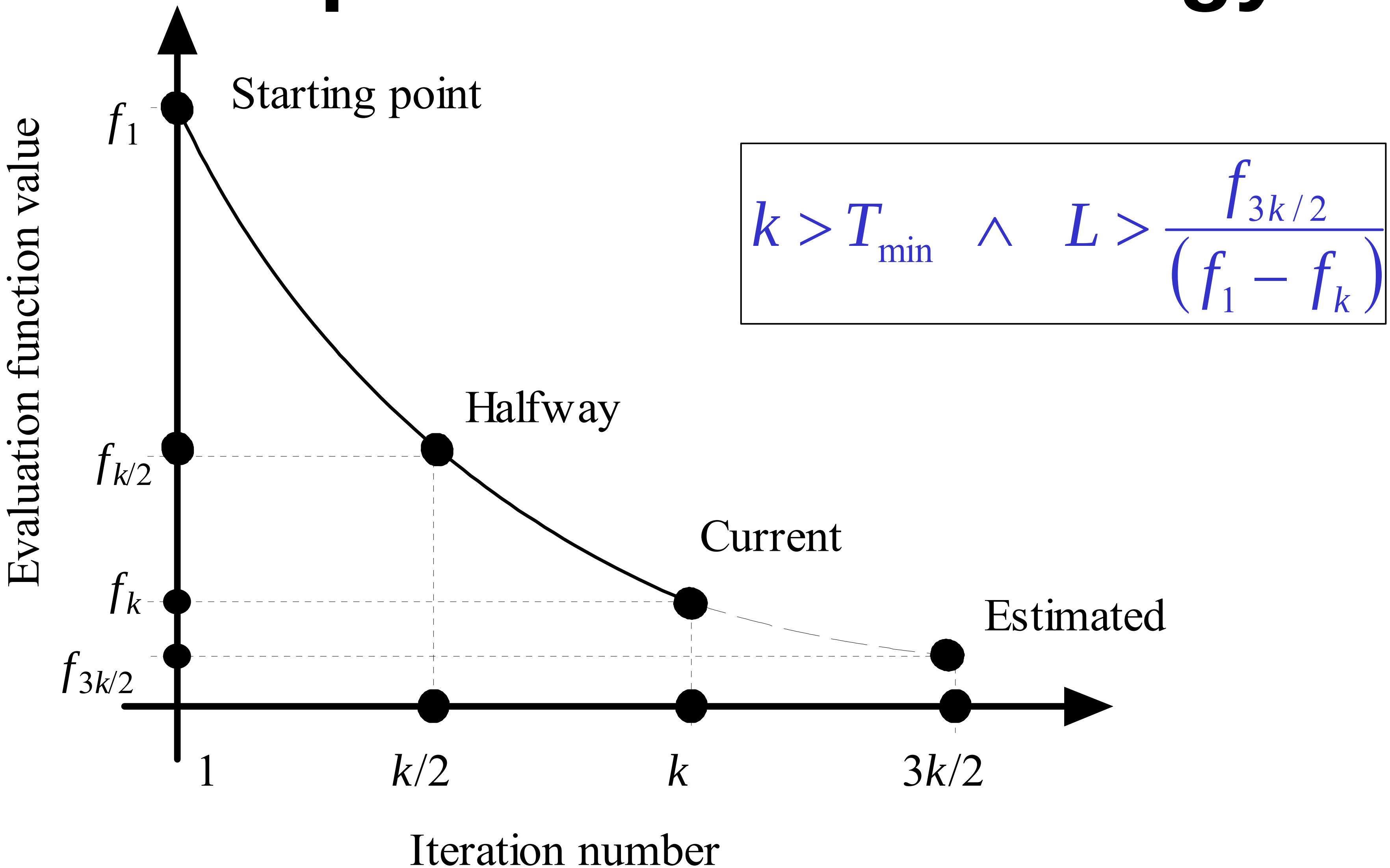
Criterion guided search

Integrate with the cost function!



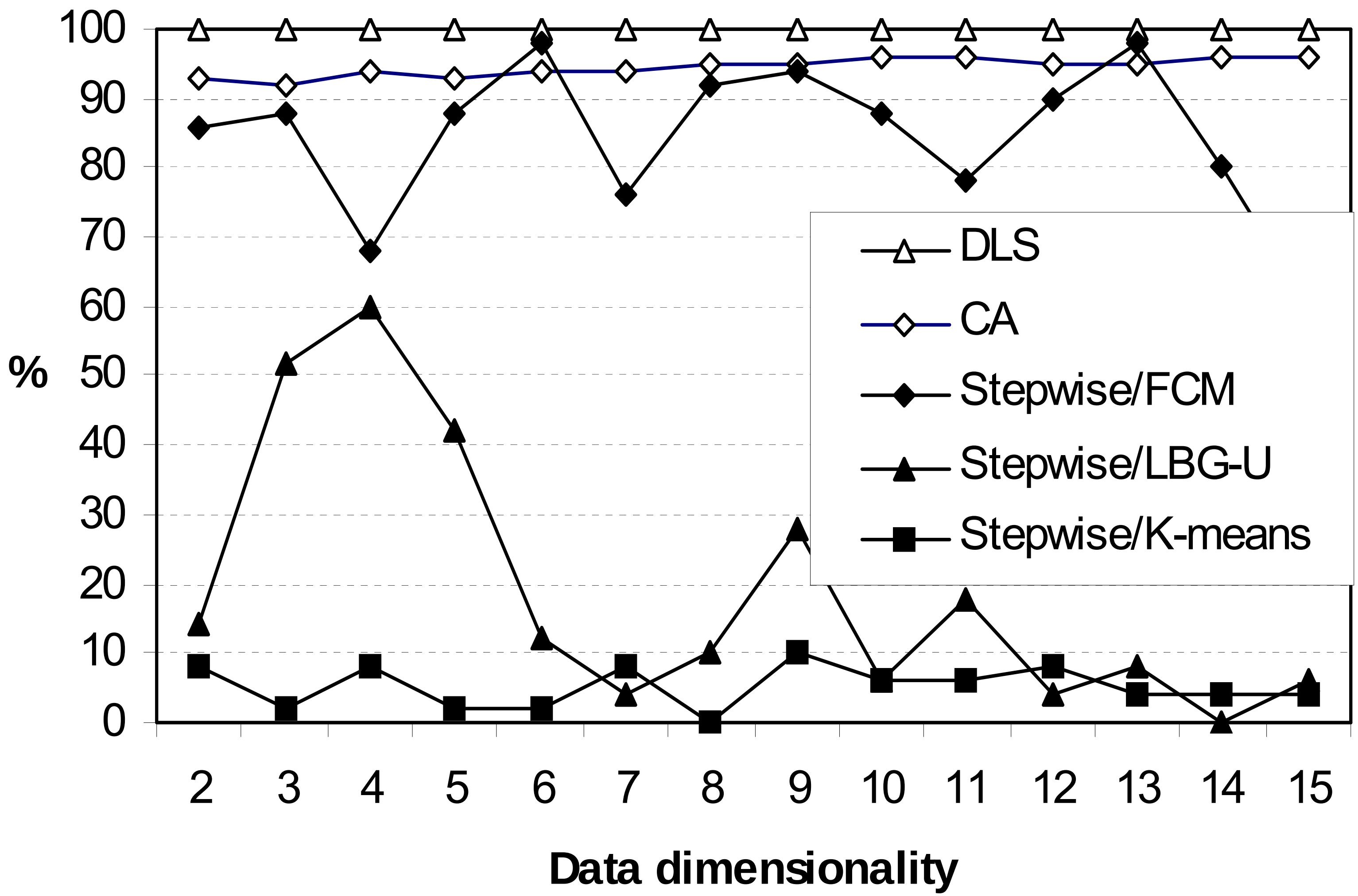
Number of clusters

Stopping criterion for stepwise search strategy



$$k > T_{\min} \;\wedge\; L > \frac{f_{3k/2}}{(f_1 - f_k)}$$

Comparison of search strategies



Open questions

Iterative algorithm (K-means or Random Swap) with criterion-guided search

… or …

Hierarchical algorithm ???

Potential topic for MSc or PhD thesis !!!

Part VI:

External indexes

Pair-counting measures

The number of pairs that are in:

Same class **both** in *P* and *G*.

$$a = \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}(n_{ij}-1)$$

Same class in *P* but different in *G*.

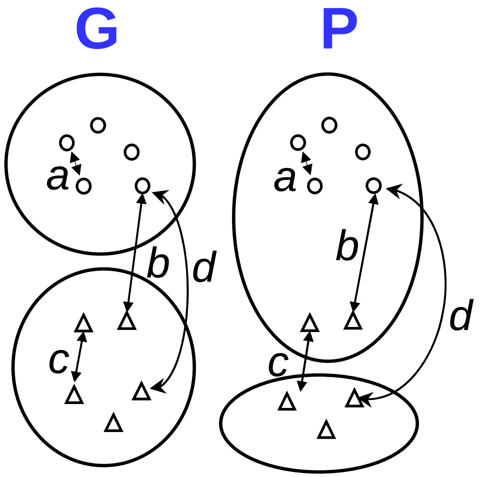
$$b = \frac{1}{2}\left(\sum_{j=1}^{K'} m_j^2 - \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2\right)$$

Different classes in *P* but same in *G*.

$$c = \frac{1}{2}\left(\sum_{i=1}^{K} n_i^2 - \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2\right)$$

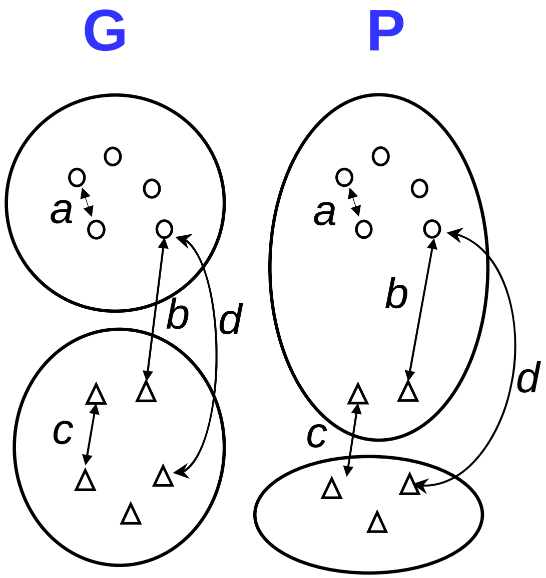
Different classes **both** in *P* and *G*.

$$d = \frac{1}{2}\left(N^2 + \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2 - \left(\sum_{i=1}^{K} n_i^2 + \sum_{j=1}^{K'} m_j^2\right)\right)$$



Rand index

[Rand, 1971]



$$RI(P,G) = \frac{a+d}{a+b+c+d}$$

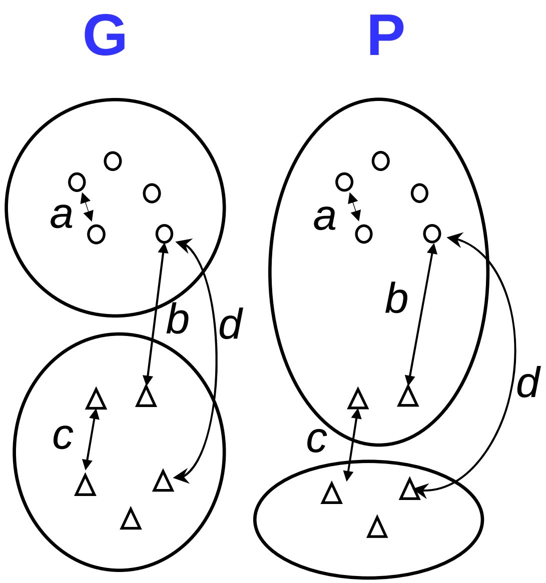
a = 20 b = 24

d = 72 c = 20

Rand index = (20+72) / (20+24+20+72) = 92/136 = **0.68**

Rand and Adjusted Rand index

[Hubert and Arabie, 1985]



$$ARI = \frac{RI - E(RI)}{1 - E(RI)}$$

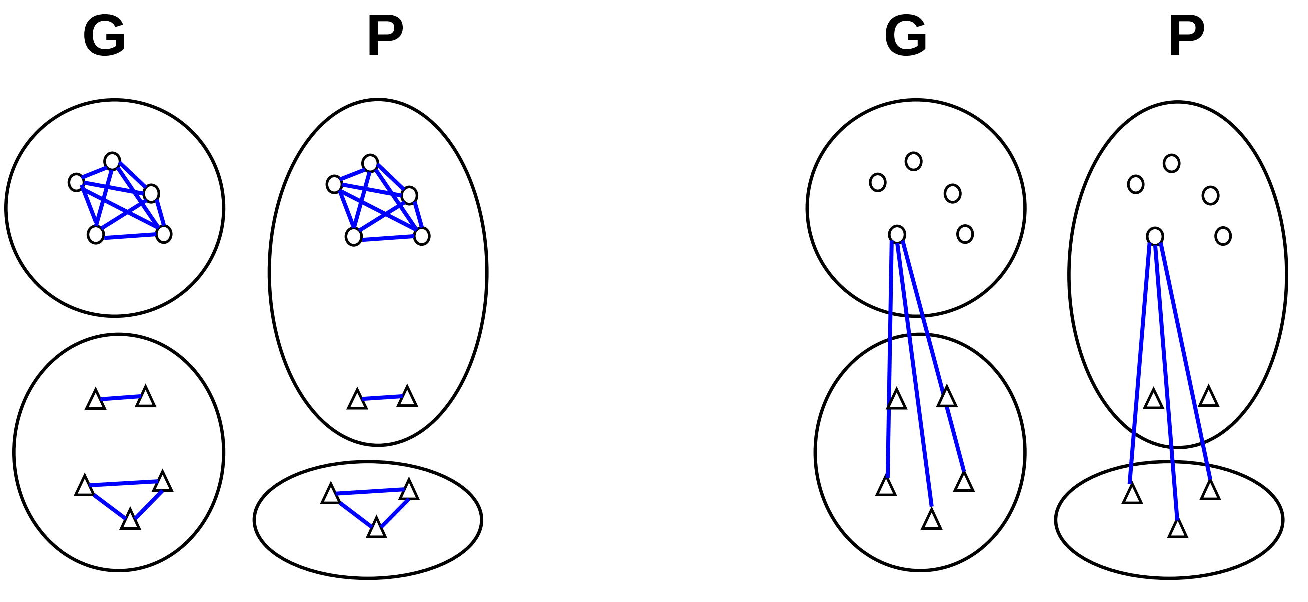
a = 20 b = 24

d = 72 c = 20

Adjusted Rand = (to be calculated) = **0.xx**

Rand statistics

Positive examples

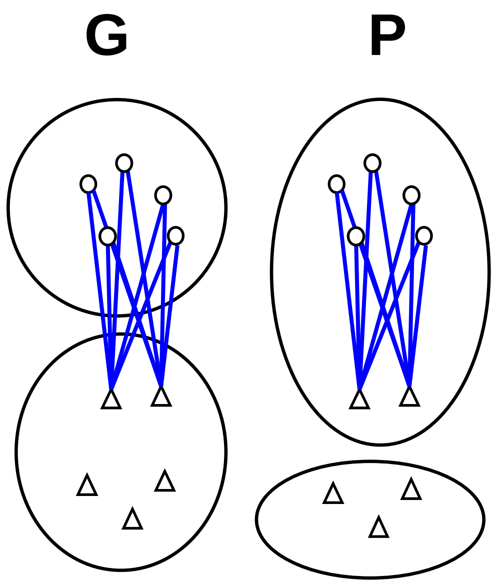


a = 20

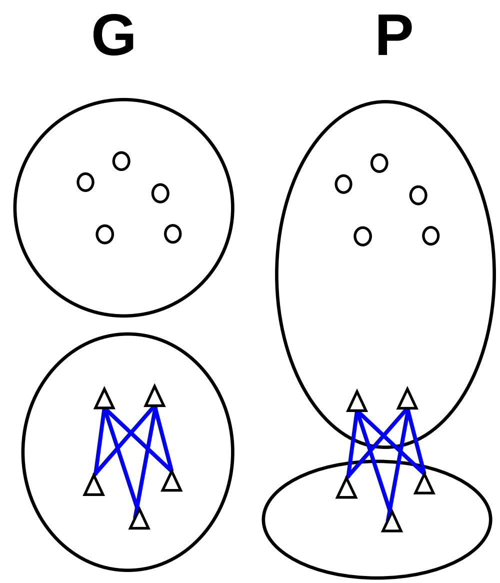
d = 72

Rand statistics

Negative examples



$b = 24$



$c = 20$

External indexes

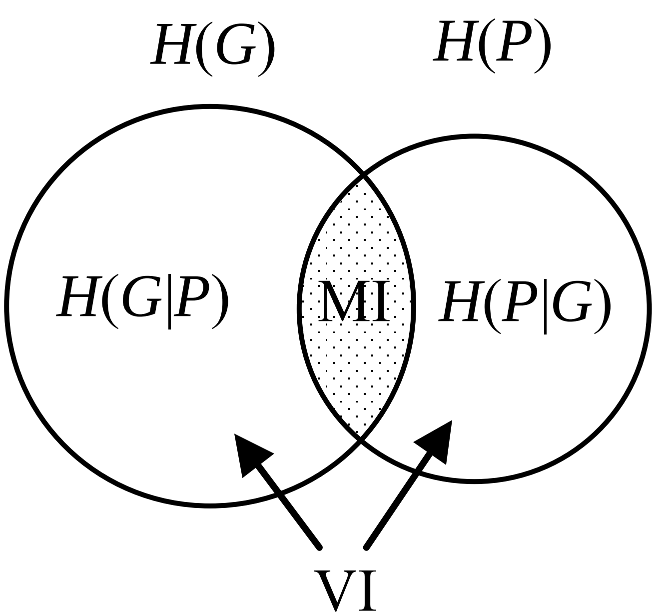
- Pair counting
- Information theoretic
- Set matching

Information-theoretic measures

- Based on the concept of entropy
- *Mutual Information* (MI): the shared informatio:

$$MI(P,G)=\sum_{i=1}^{K}\sum_{j=1}^{K'} p(P_i,G_j)\log\frac{p(P_i,G_j)}{p(P_i)\,p(G_j)}$$

- *Variation of Information* (VI) is complement of MI



Set-matching measures

Categories

- Point-level
- Cluster-level

Three problems

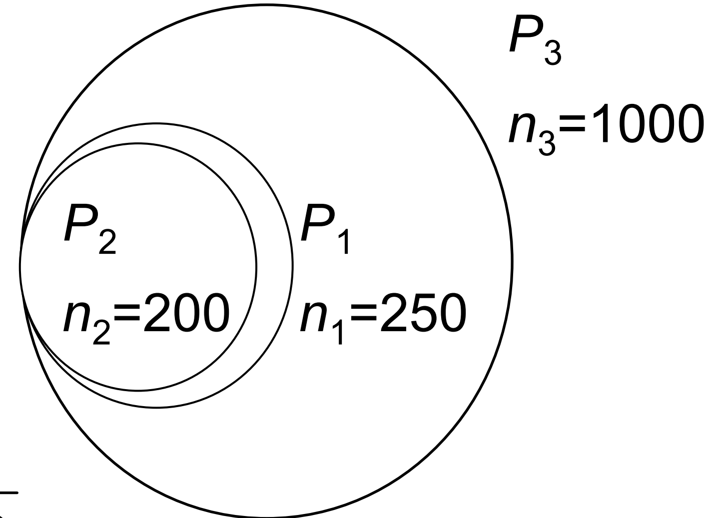
- How to measure the similarity of two clusters?
- How to pair clusters?
- How to calculate overall similarity?

Similarity of two clusters

Jaccard $$J = \frac{|P_i \cap G_j|}{|P_i \cup G_j|}$$

Sorensen-Dice $$SD = \frac{2|P_i \cap G_j|}{|P_i| + |G_j|}$$

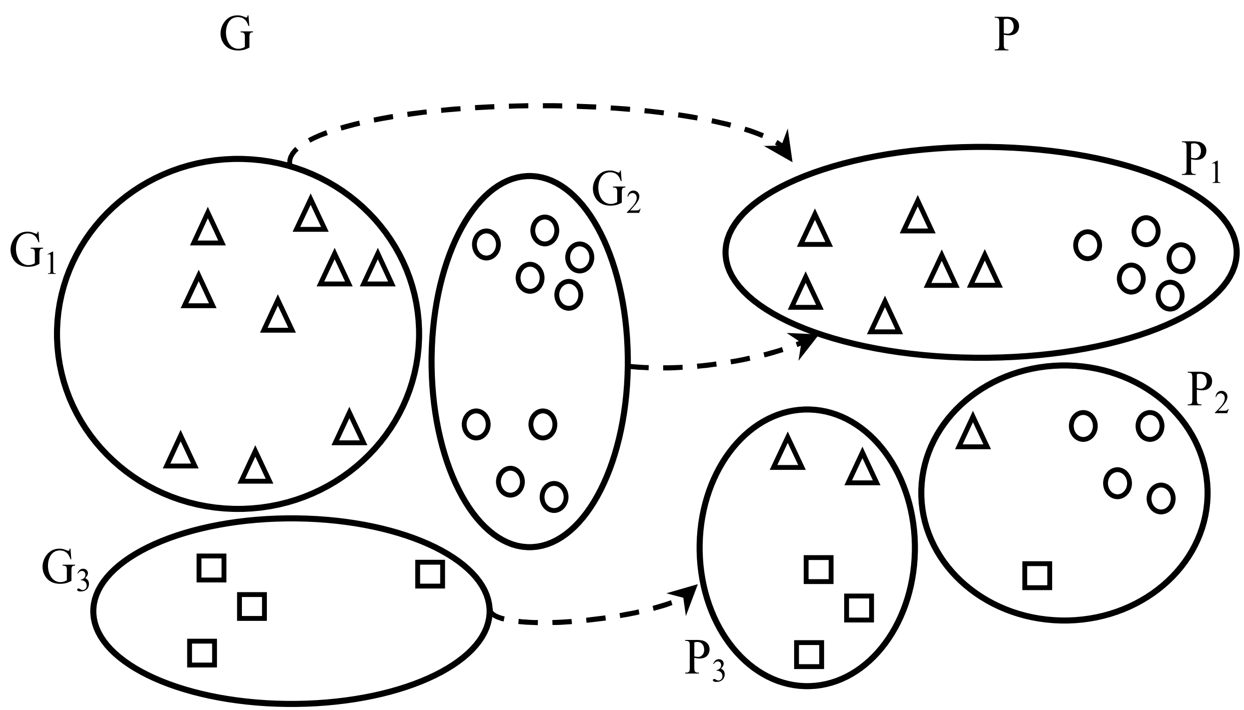
Braun-Banquet $$BB = \frac{|P_i \cap G_j|}{\max(|P_i|, |G_j|)}$$



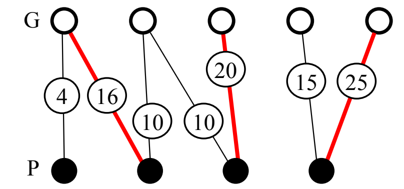
Measure:	P_1, P_2	P_1, P_3
Criterion H / NVD / CSI	200	250
J	0.80	0.25
SD	0.89	0.40
BB	0.80	0.25

Matching

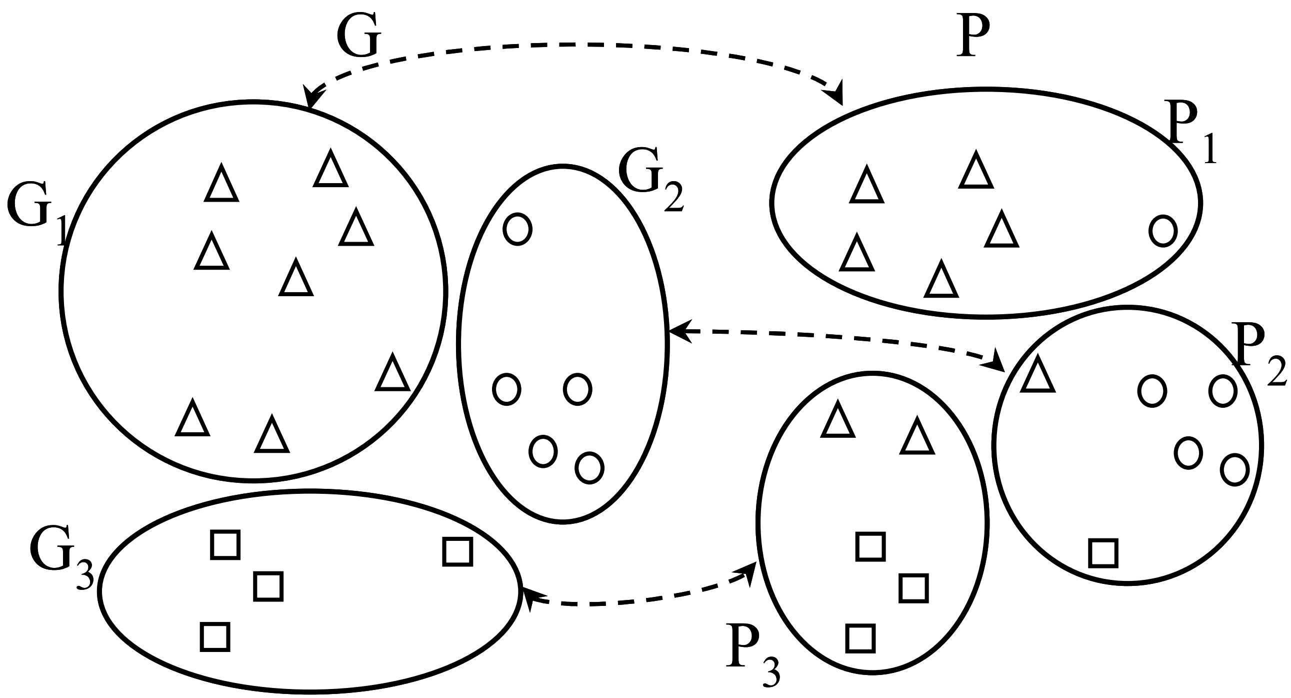
Every cluster is mapped to the cluster with maximum overlap



Pairing

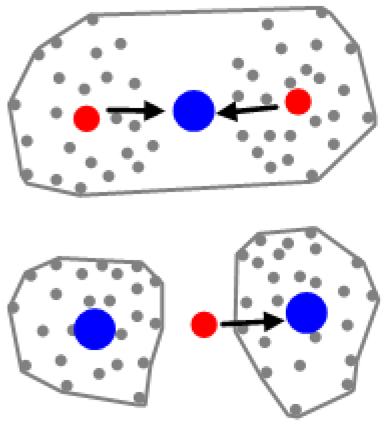


Optimal pairing by Hungarian algorithm
or greedy pairing



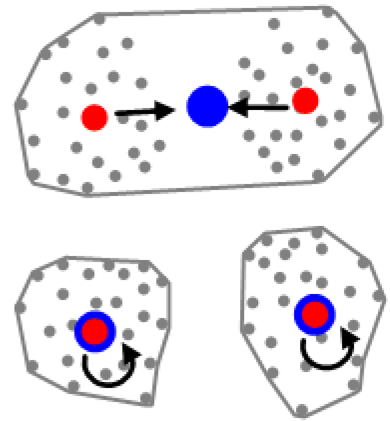
Matching vs. Pairing

3-vs-3 clusters



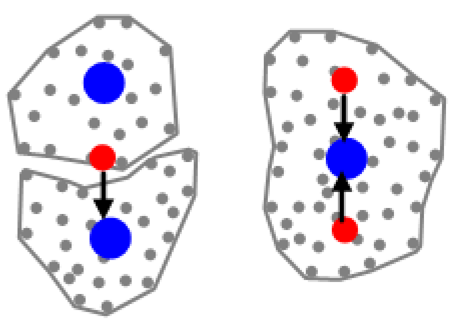
Matching=75%, Pairing=50%

3-vs-4 clusters



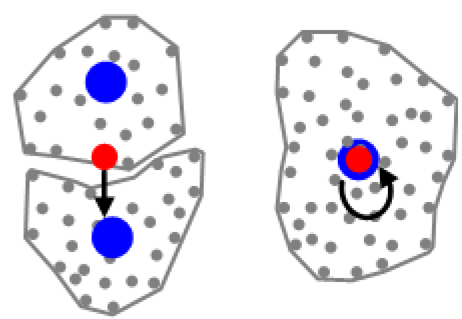
Matching=87%, Pairing=75%

3-vs-3 clusters



Matching=75%, Pairing=50%

3-vs-2 clusters



Matching=87%, Pairing=75%

Summary of matching

	Pairing/ Matching	Matching criterion	Algorithm
FM	Matching	SD	One-way
CH	Pairing	$\lvert P_i \cap G_j \rvert$	Greedy
NVD	Matching	$\lvert P_i \cap G_j \rvert$	Two-way
Purity	Matching	$\lvert P_i \cap G_j \rvert$	One-way
PSI	Pairing	BB	Optimal
CI	Matching	Centroid distance	Two-way
CSI	Matching	Centroid distance	Two-way
CR	Pairing	Centroid distance	Greedy

Overall similarity

	Total summation	Range	Normalization
FM	similarity of matched clusters	[0, 1]	N
CH	Shared objects	[0, 1]	N
NVD	Shared objects in both directions	[0, 1]	$2N$
Purity	Shared objects in one direction	[0, 1]	N
PSI	Normalized similarity of paired clusters	[0, 1]	K
CI	Orphan clusters	[0, K-1]	-
CSI	Shared objects in both directions	[0, 1]	$2N$
CR	Unstable clusters	[0, 1]	K

Normalized Van Dongen

Closely related to Purity and CSI

(Assumed that matching is symmetric)

$$NVD = 1 - \frac{\sum_{i=1}^{K} n_{ij} + \sum_{j=1}^{K'} n_{ji}}{2N} = 1 - \frac{2\sum_{i=1}^{K} n_{ij}}{2N} =$$

$$1 - \frac{\sum_{i=1}^{K} n_{ij}}{N} = CH = 1 - Purity = 1 - CSI$$

Pair Set Index (PSI)

M. Rezaei and P. Fränti, "Set matching measures for external cluster validity", *IEEE Trans. on Knowledge and Data Engineering*, August 2016.

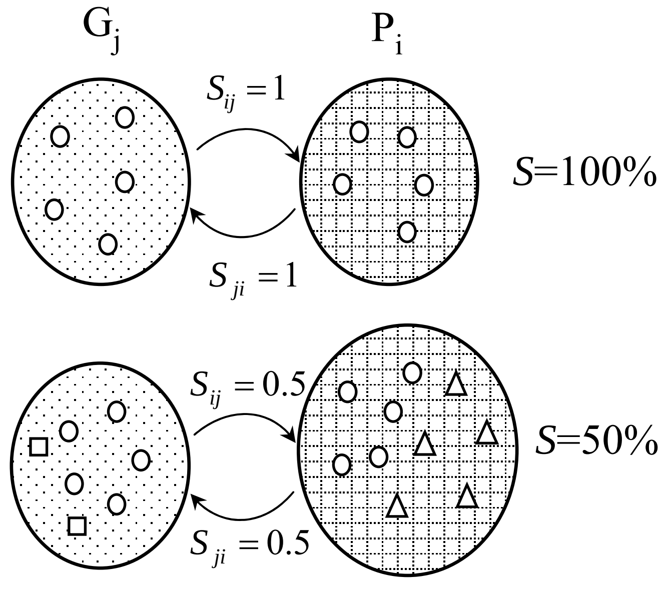
– Similarity of two clusters:

$$S_{ij} = \frac{n_{ij}}{\max(|P_i|,|G_j|)}$$

– Total similarity:

$$S_{PG} = \sum_{i} S_{ij}$$

– Pairing by Hungarian:



Pair Set Index (PSI)

Correction for chance

$$Max(S) = \min(K, K')$$

$$E(S) = \sum_{i=1}^{\min(K,K')} \frac{n_i \times (m_i / N)}{\max(n_i, m_i)}$$

size of clusters in *P* : $n_1 > n_2 > \ldots > n_K$
size of clusters in *G* : $m_1 > m_2 > \ldots > m_{K'}$

$$Transformation: \begin{cases} Max(S) \to 1 \\ E(S) \to 0 \end{cases}$$

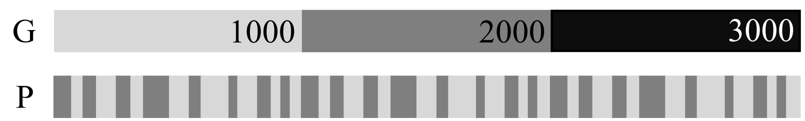
$$PSI = \begin{cases} \dfrac{S - E}{\max(K, K') - E} & S \geq E, \max(K, K') > 1 \\ 0 & S < E \\ 1 & K = K' = 1 \end{cases}$$

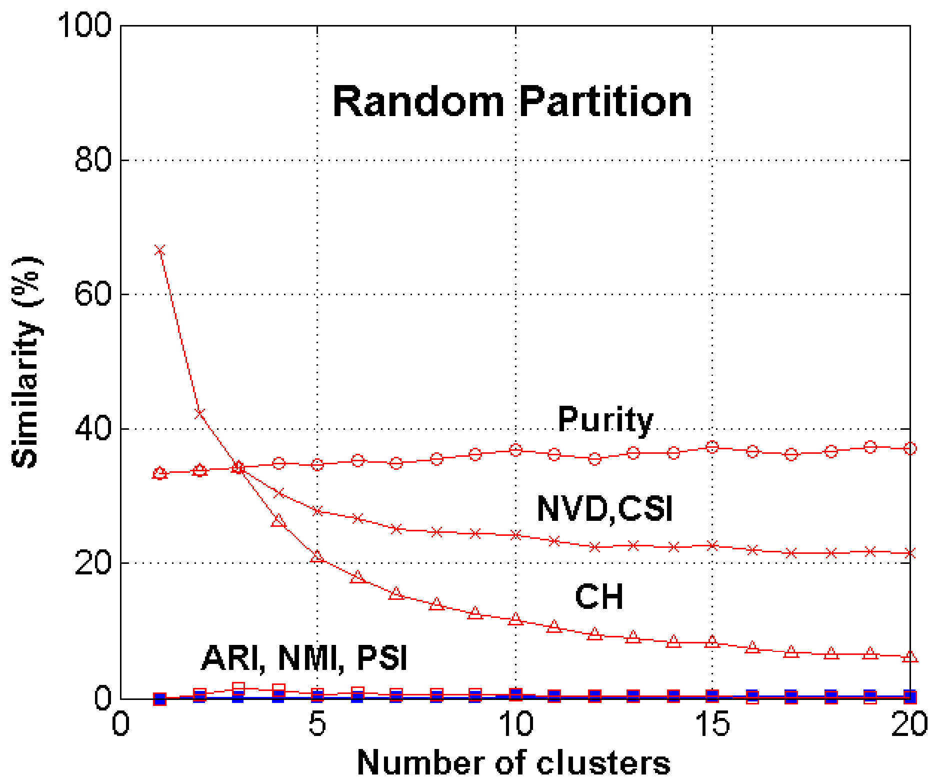
Properties of PSI

- Symmetric
- Normalized to number of clusters
- Normalized to size of clusters
- Adjusted
- Range in [0,1]
- Number of clusters can be different

Random partitioning

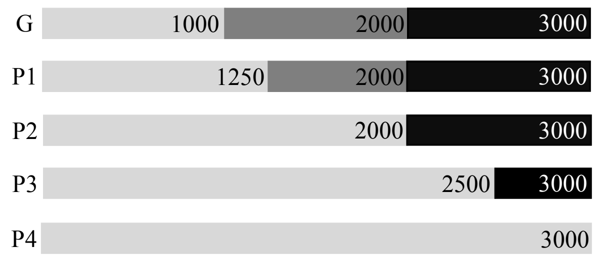
Changing number of clusters in P from 1 to 20

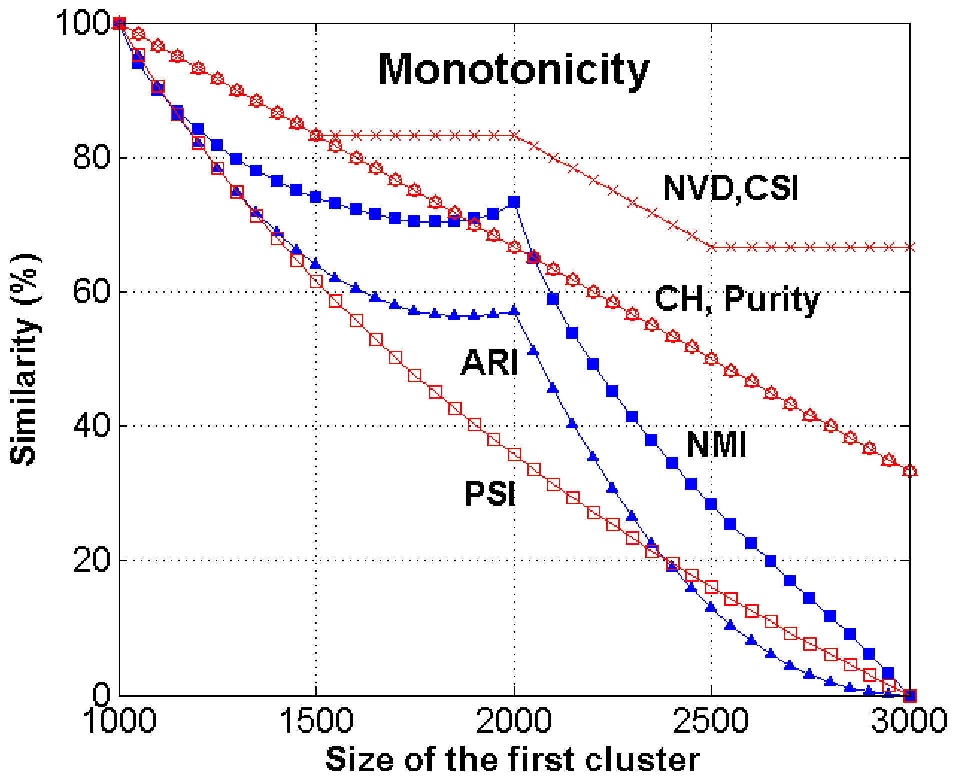




Monotonicity

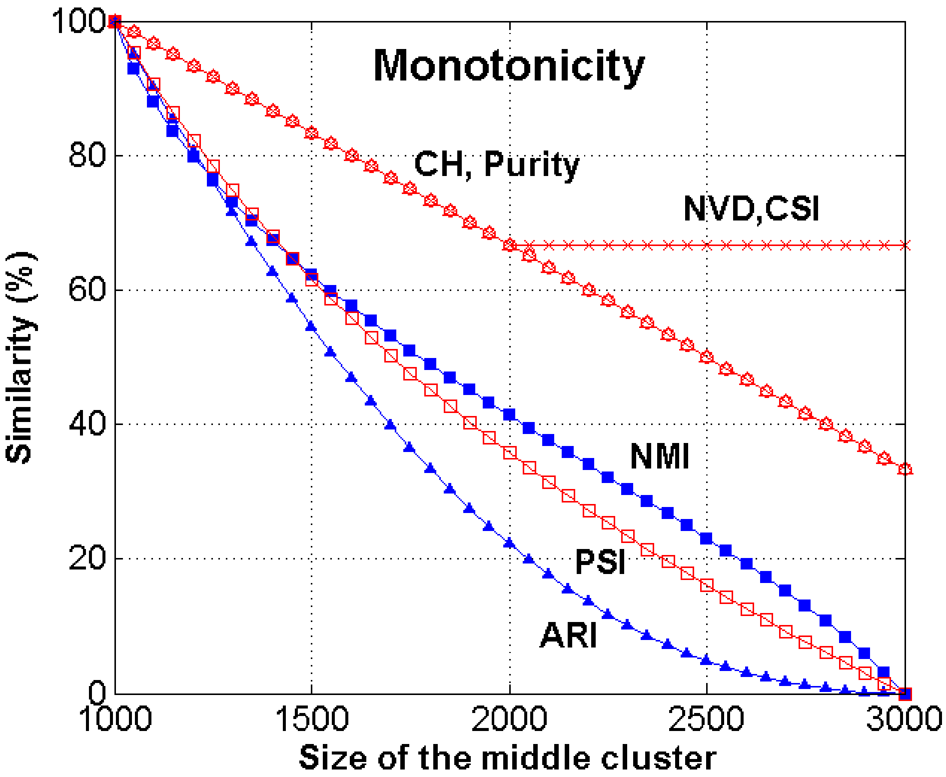
Enlarging the first cluster





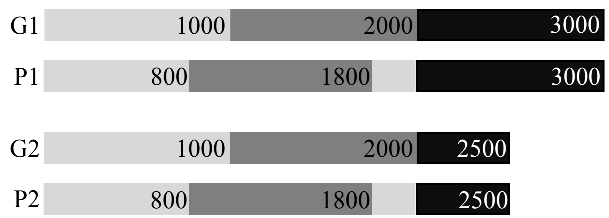
Monotonicity

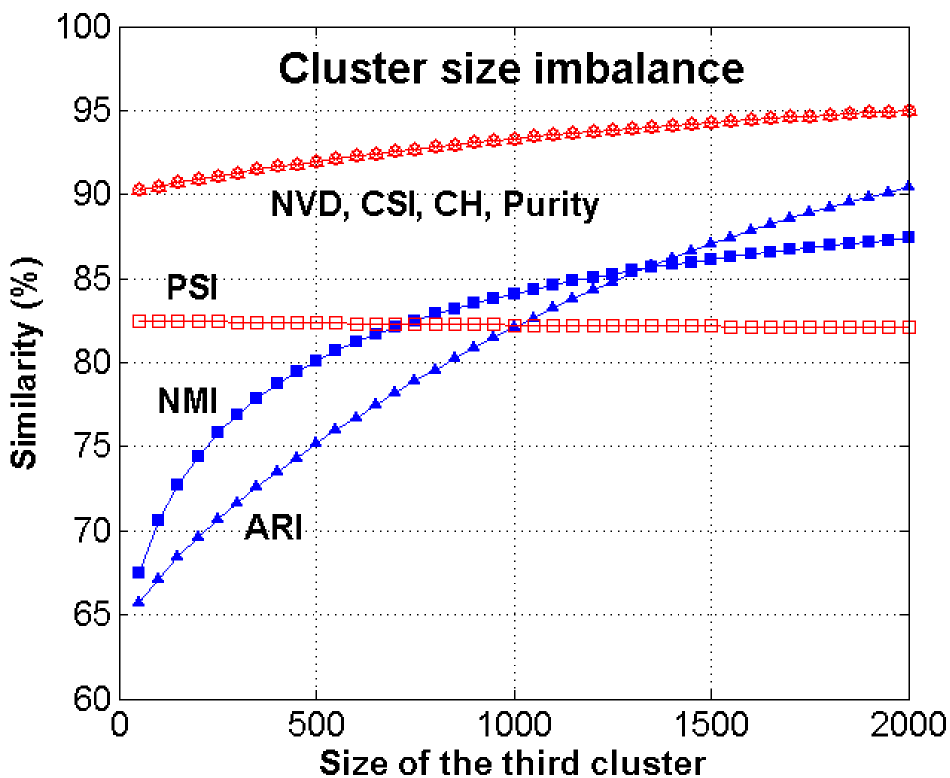
Enlarging the second cluster



Cluster size imbalance

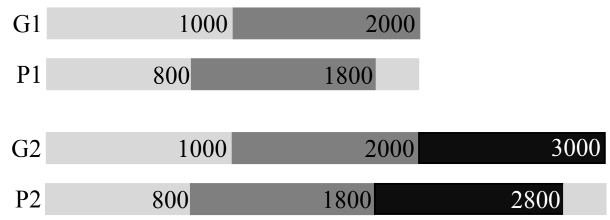
Same error in first two clusters

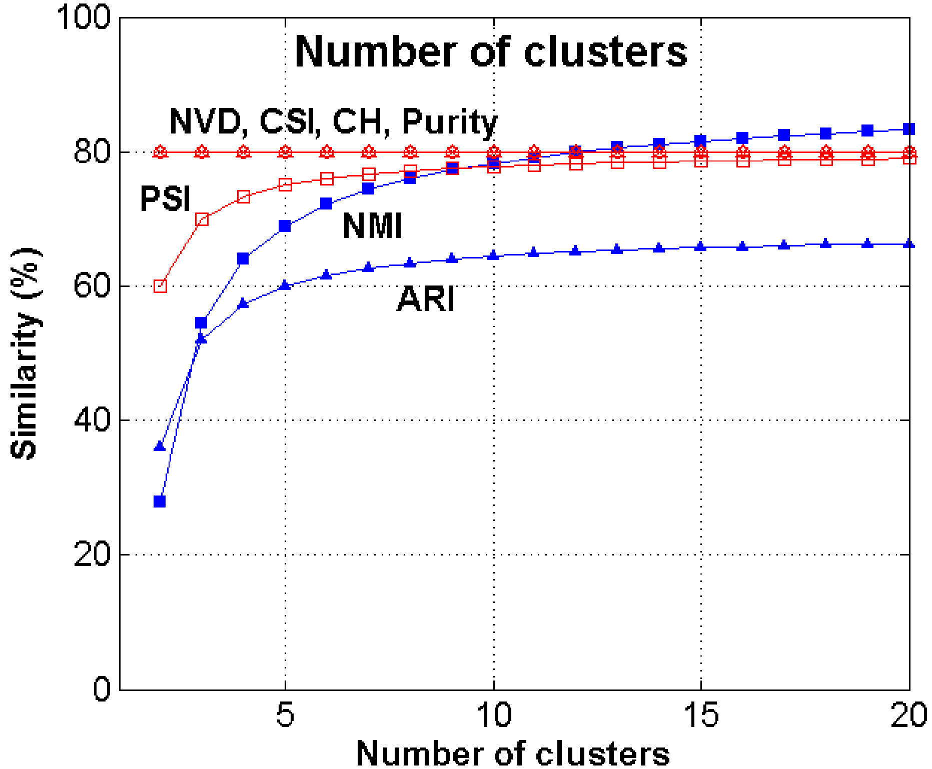




Number of clusters

Always 200 errors; k varies

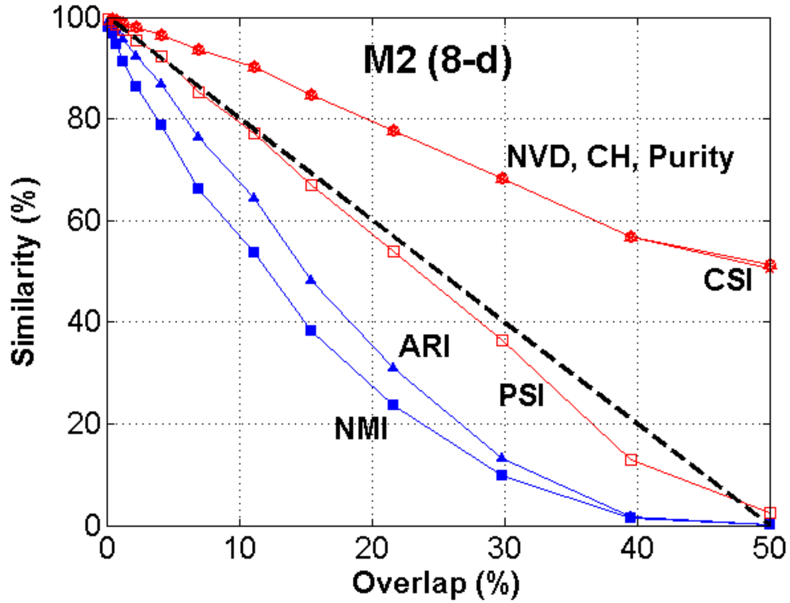




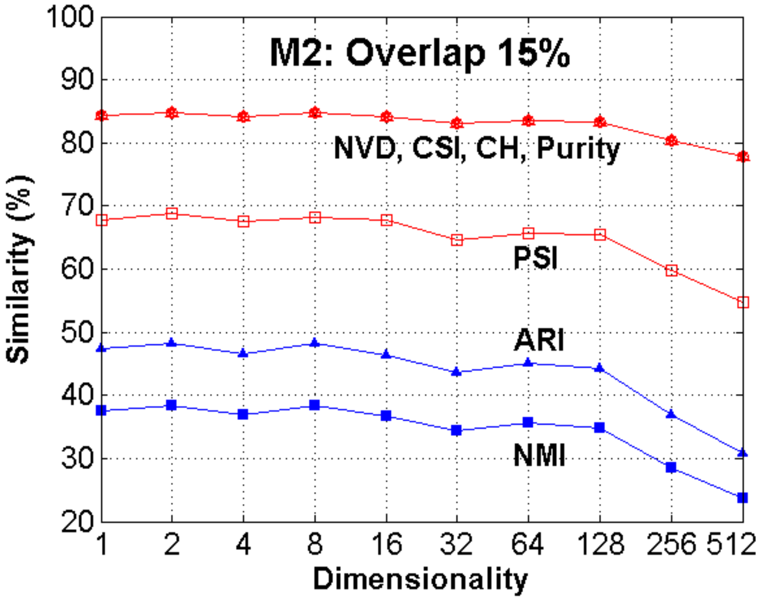
Overlap and dimensionality

Two clusters with varying overlap and dimensions

Overlap varies



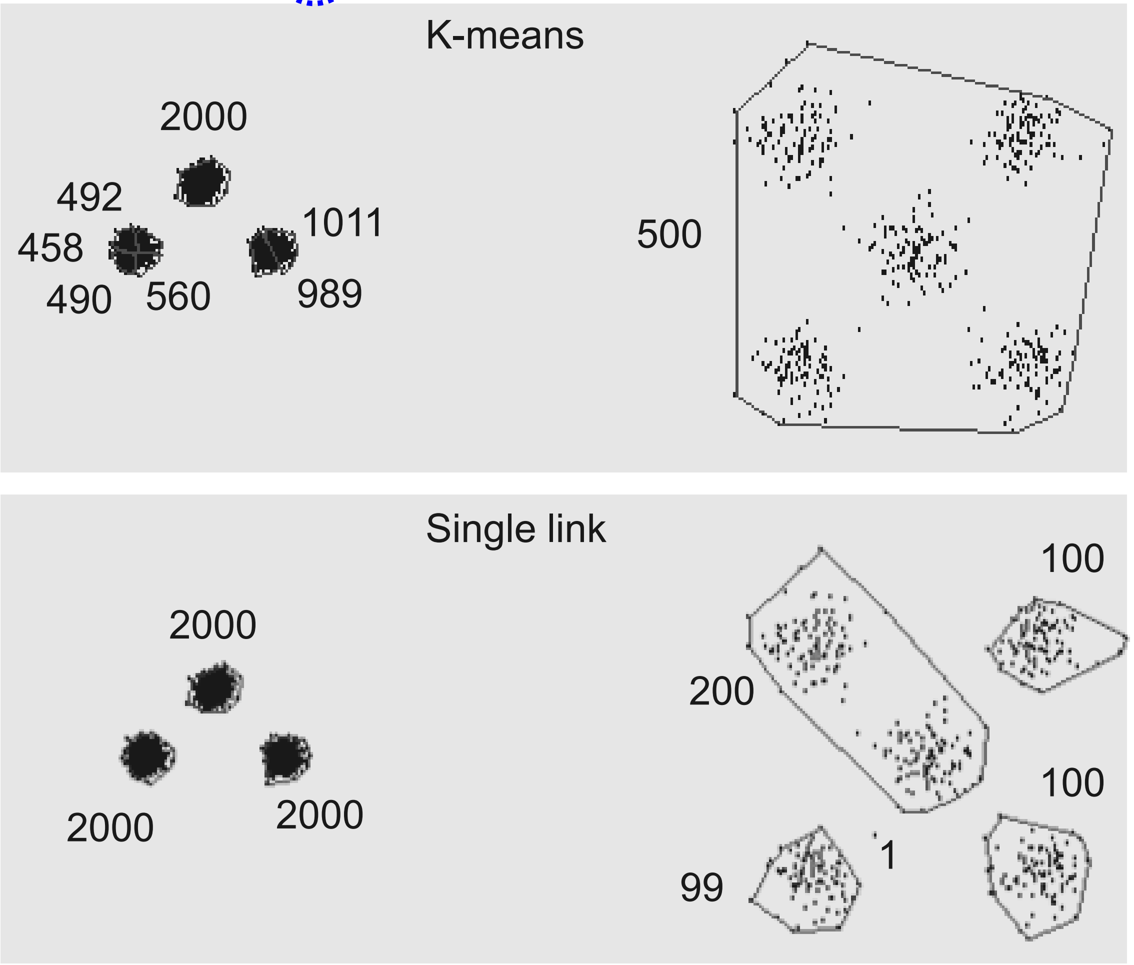
Dimensions



Unbalance

Algorithms	External indexes			
	ARI	NMI	NVD	PSI
RS	1.00	1.00	1.00	1.00
AC	1.00	1.00	1.00	1.00
SL	1.00	0.99	0.99	0.78
KM	0.66	0.77	0.78	0.18

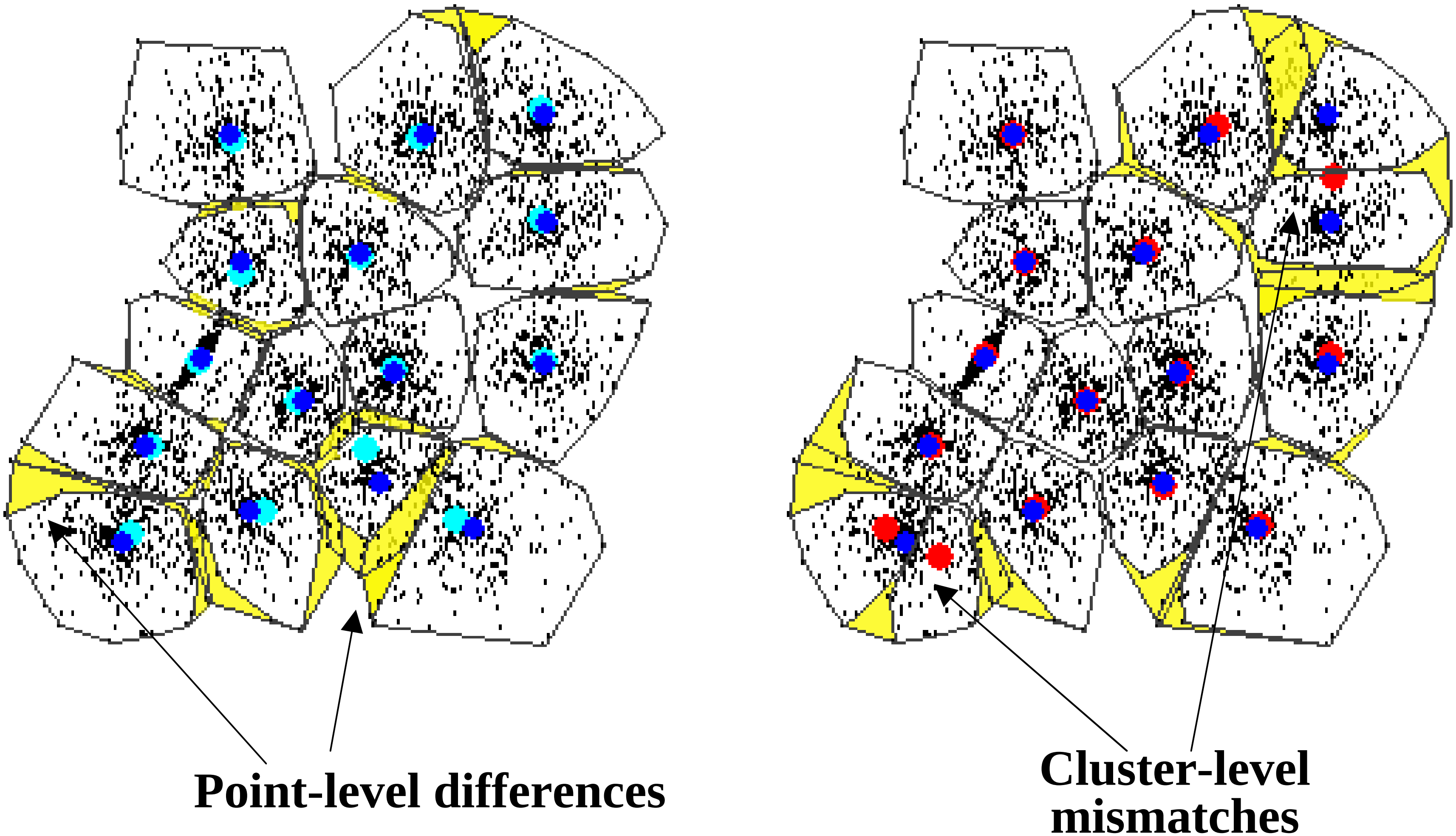
Unrealistic high



Part VII:

Cluster-level measure

Comparing partitions of centroids



Centroid index (CI)

[Fränti, Rezaei, Zhao, Pattern Recognition, 2014]

Given two sets of centroids C and C', find nearest neighbor mappings ($C \rightarrow C'$):

$$q_i \leftarrow \arg\min_{1 \le j \le K2} \left\| c_i - c'_j \right\|^2, \quad \forall i \in [1, K1]$$

Detect prototypes with no mapping:

$$orphan(c'_j) = \begin{cases} 1, & q_i \neq j \quad \forall i \\ 0, & \text{otherwise} \end{cases}$$

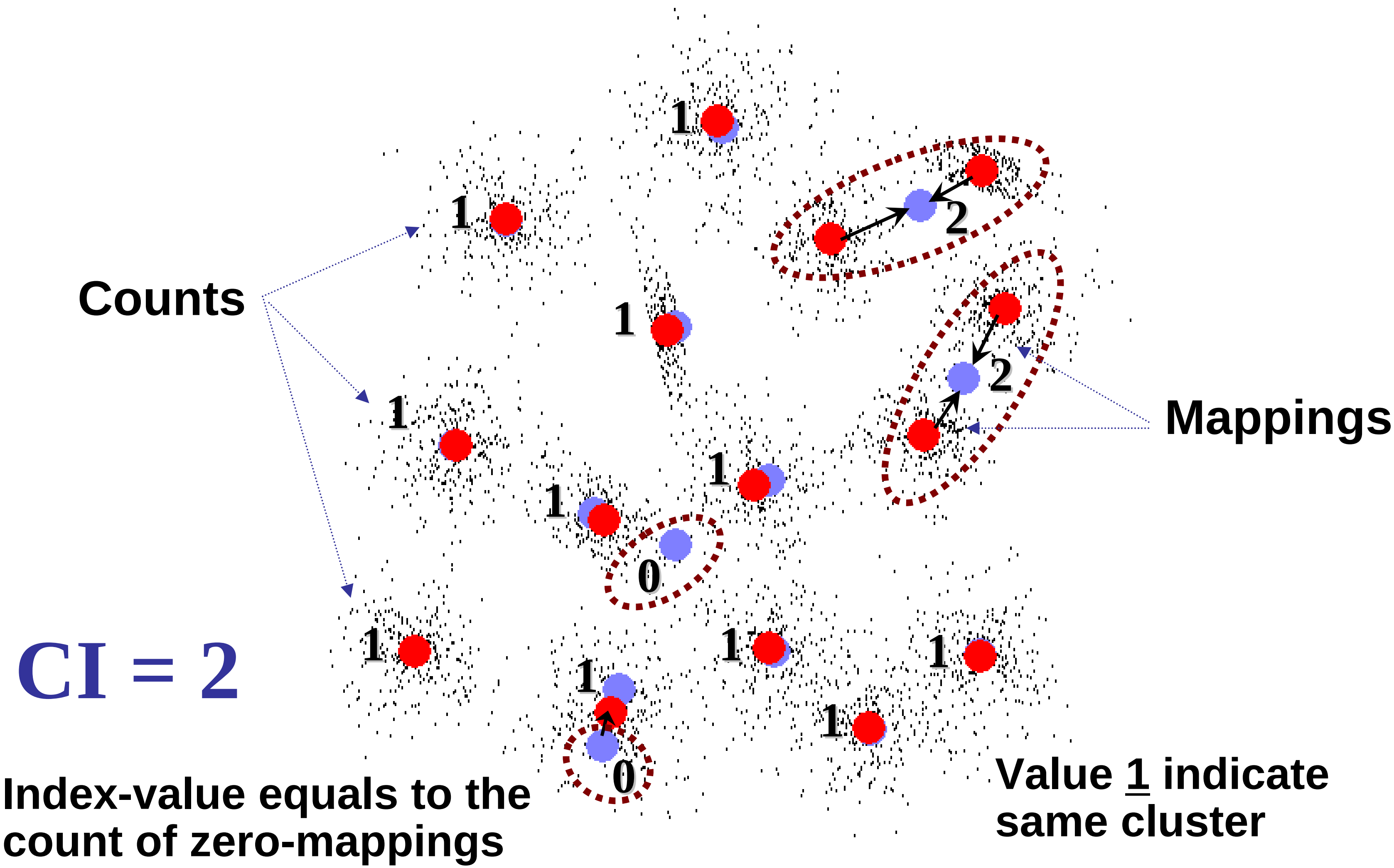
Centroid index:

$$CI_1(C, C') = \sum_{j=1}^{K2} orphan(c'_j)$$

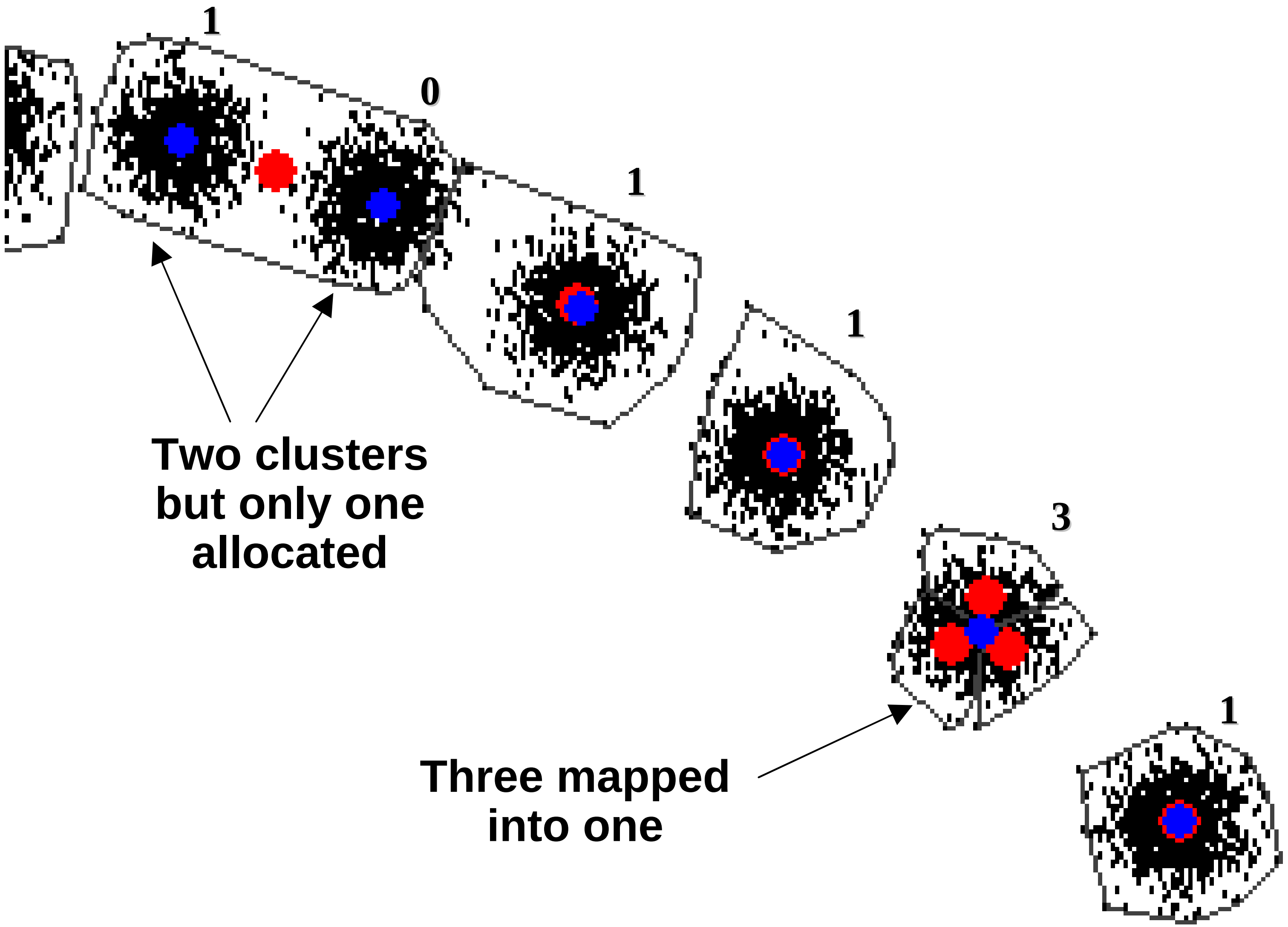
Number of zero mappings!

Example of centroid index

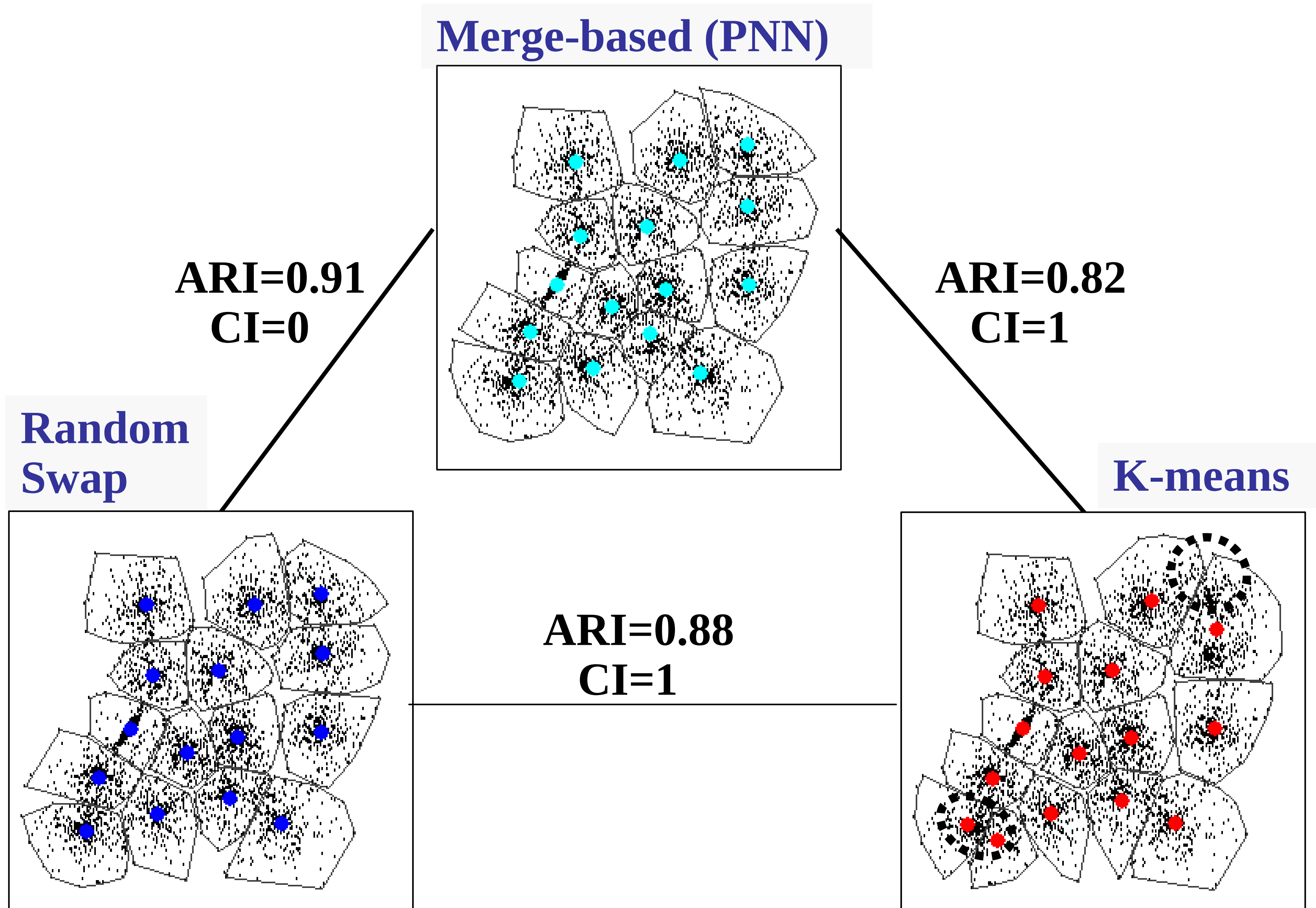
Data S_2



Example of the Centroid index



Adjusted Rand vs. Centroid index



Centroid index properties

- Mapping is not symmetric ($C \rightarrow C' \neq C' \rightarrow C$)
- Symmetric centroid index:

$$CI_2(C, C') = \max\{CI_1(C, C'), CI_1(C', C)\}$$

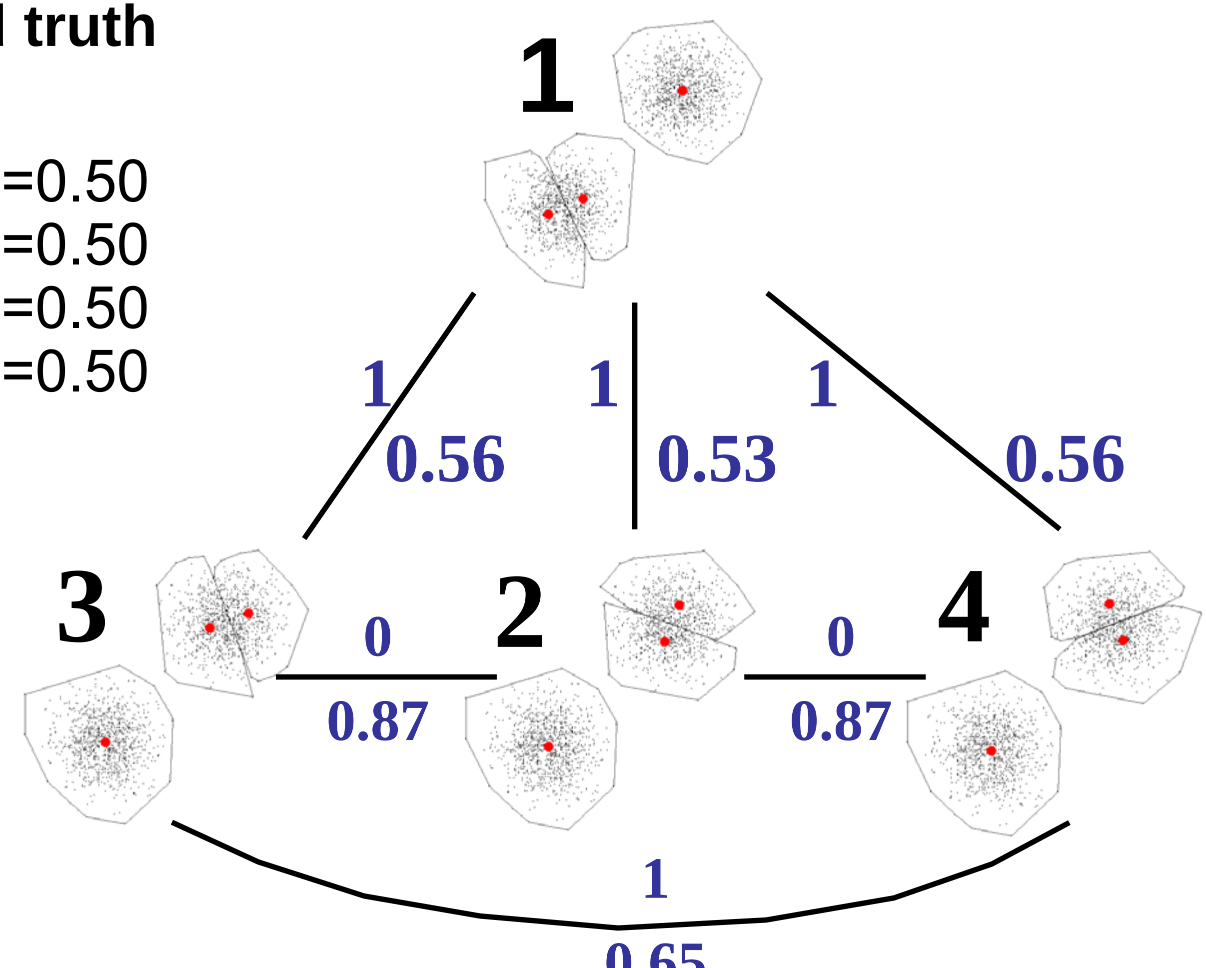
- Pointwise variant (Centroid Similarity Index):
  - Matching clusters based on CI
  - Similarity of clusters

$$CSI = \frac{S_{12} + S_{21}}{2} \text{ where } S_{12} = \frac{\sum_{i=1}^{K_1} C_i \cap C_j}{N} \quad S_{21} = \frac{\sum_{j=1}^{K_2} C_j \cap C_i}{N}$$

Centroid index

Distance to ground truth (2 clusters):

1 ↔ GT   CI=1   CSI=0.50
2 ↔ GT   CI=1   CSI=0.50
3 ↔ GT   CI=1   CSI=0.50
4 ↔ GT   CI=1   CSI=0.50



Mean Squared Errors

Data set	Clustering quality (MSE)							
	KM	RKM	KM++	XM	AC	RS	GKM	GA
Bridge	179.76	176.92	173.64	179.73	168.92	164.64	164.78	161.47
House	6.67	6.43	6.28	6.20	6.27	5.96	5.91	5.87
Miss America	5.95	5.83	5.52	5.92	5.36	5.28	5.21	5.10
House	3.61	3.28	2.50	3.57	2.62	2.83	-	2.44
$Birch_1$	5.47	5.01	4.88	5.12	4.73	4.64	-	4.64
$Birch_2$	7.47	5.65	3.07	6.29	2.28	2.28	-	2.28
$Birch_3$	2.51	2.07	1.92	2.07	1.96	1.86	-	1.86
S_1	19.71	8.92	8.92	8.92	8.93	8.92	8.92	8.92
S_2	20.58	13.28	13.28	15.87	13.44	13.28	13.28	13.28
S_3	19.57	16.89	16.89	16.89	17.70	16.89	16.89	16.89
S_4	17.73	15.70	15.70	15.71	17.52	15.70	15.71	15.70

Adjusted Rand Index

Data set	Adjusted Rand Index (ARI)							
	KM	RKM	KM++	XM	AC	RS	GKM	GA
Bridge	0.38	0.40	0.39	0.37	0.43	0.52	0.50	1
House	0.40	0.40	0.44	0.47	0.43	0.53	0.53	1
Miss America	0.19	0.19	0.18	0.20	0.20	0.20	0.23	1
House	0.46	0.49	0.52	0.46	0.49	0.49	-	1
Birch $_1$	0.85	0.93	0.98	0.91	0.96	1.00	-	1
Birch $_2$	0.81	0.86	0.95	0.86	1	1	-	1
Birch $_3$	0.74	0.82	0.87	0.82	0.86	0.91	-	1
S_1	**0.83**	1.00	1.00	1.00	1.00	1.00	1.00	1.00
S_2	**0.80**	0.99	0.99	**0.89**	0.98	0.99	0.99	0.99
S_3	**0.86**	0.96	0.96	0.96	0.92	0.96	0.96	0.96
S_4	**0.82**	0.93	0.93	0.94	**0.77**	0.93	0.93	0.93

Normalized Mutual information

Data set	Normalized Mutual Information (NMI)							
	KM	RKM	KM++	XM	AC	RS	GKM	GA
Bridge	0.77	0.78	0.78	0.77	0.80	0.83	0.82	1.00
House	0.80	0.80	0.81	0.82	0.81	0.83	0.84	1.00
Miss America	0.64	0.64	0.63	0.64	0.64	0.66	0.66	1.00
House	0.81	0.81	0.82	0.81	0.81	0.82	-	1.00
Birch $_1$	0.95	0.97	0.99	0.96	0.98	1.00	-	1.00
Birch $_2$	0.96	0.97	0.99	0.97	1.00	1.00	-	1.00
Birch $_3$	0.90	0.94	0.94	0.93	0.93	0.96	-	1.00
S_1	**0.93**	1.00	1.00	1.00	1.00	1.00	1.00	1.00
S_2	**0.90**	0.99	0.99	**0.95**	0.99	0.93	0.99	0.99
S_3	**0.92**	0.97	0.97	0.97	0.94	0.97	0.97	0.97
S_4	**0.88**	0.94	0.94	0.95	**0.85**	0.94	0.94	0.94

Normalized Van Dongen

Data set	Normalized Van Dongen (NVD)							
	KM	RKM	KM++	XM	AC	RS	GKM	GA
Bridge	0.45	0.42	0.43	0.46	0.38	0.32	0.33	0.00
House	0.44	0.43	0.40	0.37	0.40	0.33	0.31	0.00
Miss America	0.60	0.60	0.61	0.59	0.57	0.55	0.53	0.00
House	0.40	0.37	0.34	0.39	0.39	0.34	-	0.00
Birch $_1$	0.09	0.04	0.01	0.06	0.02	0.00	-	0.00
Birch $_2$	0.12	0.08	0.03	0.09	0.00	0.00	-	0.00
Birch $_3$	0.19	0.12	0.10	0.13	0.13	0.06	-	0.00
S_1	**0.09**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S_2	**0.11**	0.00	0.00	**0.06**	0.01	0.04	0.00	0.00
S_3	**0.08**	0.02	0.02	0.02	0.05	0.00	0.00	0.02
S_4	**0.11**	0.04	0.04	0.03	**0.13**	0.04	0.04	0.04

Centroid Index

Data set	C-Index (CI_2)							
	KM	RKM	KM++	XM	AC	RS	GKM	GA
Bridge	74	63	58	81	33	33	35	0
House	56	45	40	37	31	22	20	0
Miss America	88	91	67	88	38	43	36	0
House	43	39	22	47	26	23	---	0
$Birch_1$	7	3	1	4	0	0	---	0
$Birch_2$	18	11	4	12	0	0	---	0
$Birch_3$	23	11	7	10	7	2	---	0
S_1	**2**	0	0	0	0	0	0	0
S_2	**2**	0	0	**1**	0	0	0	0
S_3	**1**	0	0	0	0	0	0	0
S_4	**1**	0	0	0	**1**	0	0	0

Centroid Similarity Index

Data set	Centroid Similarity Index (CSI)							
	KM	RKM	KM++	XM	AC	RS	GKM	GA
Bridge	0.47	0.51	0.49	0.45	0.57	0.62	0.63	1.00
House	0.49	0.50	0.54	0.57	0.55	0.63	0.66	1.00
Miss America	0.32	0.32	0.32	0.33	0.38	0.40	0.42	1.00
House	0.54	0.57	0.63	0.54	0.57	0.62	---	1.00
Birch $_1$	0.87	0.94	0.98	0.93	0.99	1.00	---	1.00
Birch $_2$	0.76	0.84	0.94	0.83	1.00	1.00	---	1.00
Birch $_3$	0.71	0.82	0.87	0.81	0.86	0.93	---	1.00
S_1	**0.83**	1.00	1.00	1.00	1.00	1.00	1.00	1.00
S_2	**0.82**	1.00	1.00	**0.91**	1.00	1.00	1.00	1.00
S_3	**0.89**	0.99	0.99	0.99	0.98	0.99	0.99	0.99
S_4	**0.87**	0.98	0.98	0.99	**0.85**	0.98	0.98	0.98

High quality clustering

	Method	MSE
GKM	Global K-means	164.78
RS	Random swap (5k)	164.64
GA	Genetic algorithm	161.47
RS_{8M}	Random swap (8M)	161.02
GAIS-2002	GAIS	160.72
+ RS_{1M}	GAIS + RS (1M)	160.49
+ RS_{8M}	GAIS + RS (8M)	160.43
GAIS-2012	GAIS	160.68
+ RS_{1M}	GAIS + RS (1M)	160.45
+ RS_{8M}	GAIS + RS (8M)	160.39
+ PRS	GAIS + PRS	160.33
+ RS_{8M} +	GAIS + RS (8M) +	160.28

Centroid index values

Main algorithm:	RS_{8M}	GAIS 2002			GAIS 2012				
+ Tuning 1	×	×	RS_{1M}	RS_{8M}	×	RS_{1M}	RS_{8M}		RS_{8M}
+ Tuning 2	×	×	×	×	×	×	×	×	
RS_{8M}	---	19	19	19	23	24	24	23	22
GAIS (2002)	23	---	0	0	14	15	15	14	16
+ RS_{1M}	23	0	---	0	14	15	15	14	13
+ RS_{8M}	23	0	0	---	14	15	15	14	13
GAIS (2012)	25	17	18	18	---	1	1	1	1
+ RS_{1M}	25	17	18	18	1	---	0	0	1
+ RS_{8M}	25	17	18	18	1	0	---	0	1
+ PRS	25	17	18	18	1	0	0	---	1
+ RS_{8M} + PRS	24	17	18	18	1	1	1	1	---

Summary of external indexes

(existing measures)

Table 1: External Cluster Validation Measures.

	Measure	Notation	Definition	Range
1	Entropy	E	$-\sum_i p_i(\sum_j \frac{p_{ij}}{p_i}\log\frac{p_{ij}}{p_i})$	$[0,\log K']$
2	Purity	P	$\sum_i p_i(\max_j \frac{p_{ij}}{p_i})$	(0,1]
3	F-measure	F	$\sum_j p_j \max_i[2\frac{p_{ij}}{p_i}\frac{p_{ij}}{p_j}/(\frac{p_{ij}}{p_i}+\frac{p_{ij}}{p_j})]$	(0,1]
4	Variation of Information	VI	$-\sum_i p_i\log p_i - \sum_j p_j \log p_j - 2\sum_i\sum_j p_{ij}\log\frac{p_{ij}}{p_ip_j}$	$[0, 2\log\max(K,K')]$
5	Mutual Information	MI	$\sum_i\sum_j p_{ij}\log\frac{p_{ij}}{p_ip_j}$	$(0,\log K']$
6	Rand statistic	R	$[\binom{n}{2} - \sum_i\binom{n_{i\cdot}}{2} - \sum_j\binom{n_{\cdot j}}{2} + 2\sum_{ij}\binom{n_{ij}}{2}]/\binom{n}{2}$	(0,1]
7	Jaccard coefficient	J	$\sum_{ij}\binom{n_{ij}}{2}/[\sum_i\binom{n_{i\cdot}}{2}+\sum_j\binom{n_{\cdot j}}{2}-\sum_{ij}\binom{n_{ij}}{2}]$	[0,1]
8	Fowlkes and Mallows index	FM	$\sum_{ij}\binom{n_{ij}}{2}/\sqrt{\sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2}}$	[0,1]
9	Hubert Γ statistic I	Γ	$\frac{\binom{n}{2}\sum_{ij}\binom{n_{ij}}{2}-\sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2}}{\sqrt{\sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2}[\binom{n}{2}-\sum_i\binom{n_{i\cdot}}{2}][\binom{n}{2}-\sum_j\binom{n_{\cdot j}}{2}]}}$	(-1,1]
10	Hubert Γ statistic II	Γ'	$[\binom{n}{2} - 2\sum_i\binom{n_{i\cdot}}{2} - 2\sum_j\binom{n_{\cdot j}}{2} + 4\sum_{ij}\binom{n_{ij}}{2}]/\binom{n}{2}$	[0,1]
11	Minkowski score	MS	$\sqrt{\sum_i\binom{n_{i\cdot}}{2}+\sum_j\binom{n_{\cdot j}}{2}-2\sum_{ij}\binom{n_{ij}}{2}}/\sqrt{\sum_j\binom{n_{\cdot j}}{2}}$	$[0,+\infty)$
12	classification error	ε	$1-\frac{1}{n}\max_\sigma\sum_j n_{\sigma(j),j}$	[0,1)
13	van Dongen criterion	VD	$(2n-\sum_i\max_j n_{ij}-\sum_j\max_i n_{ij})/2n$	[0, 1)
14	micro-average precision	MAP	$\sum_i p_i(\max_j \frac{p_{ij}}{p_i})$	(0,1]
15	Goodman-Kruskal coefficient	GK	$\sum_i p_i(1-\max_j\frac{p_{ij}}{p_i})$	[0,1)
16	Mirkin metric	M	$\sum_i n_{i\cdot}^2+\sum_j n_{\cdot j}^2-2\sum_i\sum_j n_{ij}^2$	$[0,2\binom{n}{2})$

Note: $p_{ij}=n_{ij}/n,\ p_i=n_{i\cdot}/n,\ p_j=n_{\cdot j}/n$.

Literature

1. G.W. Milligan, and M.C. Cooper, “An examination of procedures for determining the number of clusters in a data set”, *Psychometrika*, Vol.50, 1985, pp. 159-179.
2. E. Dimitriadou, S. Dolnicar, and A. Weingassel, “An examination of indexes for determining the number of clusters in binary data sets”, *Psychometrika*, Vol.67, No.1, 2002, pp. 137-160.
3. D.L. Davies and D.W. Bouldin, "A cluster separation measure “, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 224-227, 1979.
4. J.C. Bezdek and N.R. Pal, "Some new indexes of cluster validity “, *IEEE Transactions on Systems, Man and Cybernetics*, 28(3), 302-315, 1998.
5. H. Bischof, A. Leonardis, and A. Selb, "MDL Principle for robust vector quantization“, *Pattern Analysis and Applications*, 2(1), 59-72, 1999.
6. P. Fränti, M. Xu and I. Kärkkäinen, "Classification of binary vectors by using DeltaSC-distance to minimize stochastic complexity", *Pattern Recognition Letters*, 24 (1-3), 65-73, January 2003.

Literature

7. G.M. James, C.A. Sugar, "Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach". *Journal of the American Statistical Association*, vol. 98, 397-408, 2003.
8. P.K. Ito, Robustness of ANOVA and MANOVA Test Procedures. In: Krishnaiah P. R. (ed), *Handbook of Statistics 1: Analysis of Variance*. North-Holland Publishing Company, 1980.
9. I. Kärkkäinen and P. Fränti, "Dynamic local search for clustering with unknown number of clusters", *Int. Conf. on Pattern Recognition (ICPR'02*), Québec, Canada, vol. 2, 240-243, August 2002.
10. D. Pellag and A. Moore, "X-means: Extending K-Means with Efficient Estimation of the Number of Clusters", *Int. Conf. on Machine Learning* (ICML), 727-734, San Francisco, 2000.
11. S. Salvador and P. Chan, "Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms", *IEEE Int. Con. Tools with Artificial Intelligence* (ICTAI), 576-584, Boca Raton, Florida, November, 2004.
12. M. Gyllenberg, T. Koski and M. Verlaan, "Classification of binary vectors by stochastic complexity ". *Journal of Multivariate Analysis*, 63(1), 47-72, 1997.

Literature

13. M. Gyllenberg, T. Koski and M. Verlaan, "Classification of binary vectors by stochastic complexity ". *Journal of Multivariate Analysis*, 63(1), 47-72, 1997.
14. X. Hu and L. Xu, "A Comparative Study of Several Cluster Number Selection Criteria", *Int. Conf. Intelligent Data Engineering and Automated Learning* (IDEAL), 195-202, Hong Kong, 2003.
15. Kaufman, L. and P. Rousseeuw, 1990. Finding Groups in Data: An Introduction to Cluster Analysis. *John Wiley and Sons, London*. ISBN: 10:0471878766.
16. [1.3] M.Halkidi, Y.Batistakis and M.Vazirgiannis: Cluster validity methods: part 1, SIGMOD Rec., Vol.31, No.2, pp.40-45, 2002
17. R. Tibshirani, G. Walther, T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *J.R.Statist. Soc. B*(2001) 63, Part 2, pp.411-423.
18. T. Lange, V. Roth, M, Braun and J. M. Buhmann. Stability-based validation of clustering solutions. *Neural Computation*. Vol. 16, pp. 1299-1323. 2004.

Literature

19. Q. Zhao, M. Xu and P. Fränti, "Sum-of-squares based clustering validity index and significance analysis", *Int. Conf. on Adaptive and Natural Computing Algorithms (ICANNGA'09*), Kuopio, Finland, LNCS 5495, 313-322, April 2009.
20. Q. Zhao, M. Xu and P. Fränti, "Knee point detection on bayesian information criterion", *IEEE Int. Conf. Tools with Artificial Intelligence (ICTAI)*, Dayton, Ohio, USA, 431-438, November 2008.
21. W.M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, 66, 846–850, 1971
22. L. Hubert and P. Arabie, "Comparing partitions", *Journal of Classification*, 2(1), 193-218, 1985.
23. P. Fränti, M. Rezaei and Q. Zhao, "Centroid index: cluster level similarity measure", *Pattern Recognition*, 47 (9), 3034-3045, September 2014, 2014.
24. M. Rezaei and P. Fränti, "Set matching measures for external cluster validity", *IEEE Trans. on Knowledge and Data Engineering*, 28 (8), 2173-2186, August 2016.
25. M. Rezaei and P. Fränti "Can the number of clusters be solved by external index?", (submitted)