Cluster stability analysis based on the assessment of individual clusters

Patrice Bertrand *

ENST Bretagne (France)

* joint work with **G. Bel Mufti and L. El Moubarki**

CEFI ESSEC (Tunisie)

Tübingen, July 2007

Partition stability

Several approaches:

- Clustering cross validation
- Effects of small changes in the data set:

  a) Adding a noise

  b) Different sub-sampling schemes

  c) Random projections

  ...

Number of clusters: Roberts (1997), Levine and Domany (2001), Tibshirani, Walther and Hastie (2001), Tibshirani, Walther, Botstein *et al.* (2001), Ben-Hur *et al.* (2002), T. Lange *et al.* (2004), ...

Asymptotic case: Krieger and Green (1999), Ben David, von Luxburg and Pal (2006).

Outline

1) Introduction

2) Cluster stability w.r.t. cluster isolation and cluster cohesion

3) Illustration on artificial and real data sets

4) Partial Membership

5) Some conclusions and perspectives

Our notations

- $\mathcal{X}$ **set of objects of the data set**
- $\mathcal{X}'$ **sample drawn i.i.d. from** $\mathcal{X}$
- A_k k**-partitioning algorithm**
- r **sampling ratio**

Stability based on sampling the data set

Levine & Domany (2001), Ben-Hur *et al.* (2002), ...

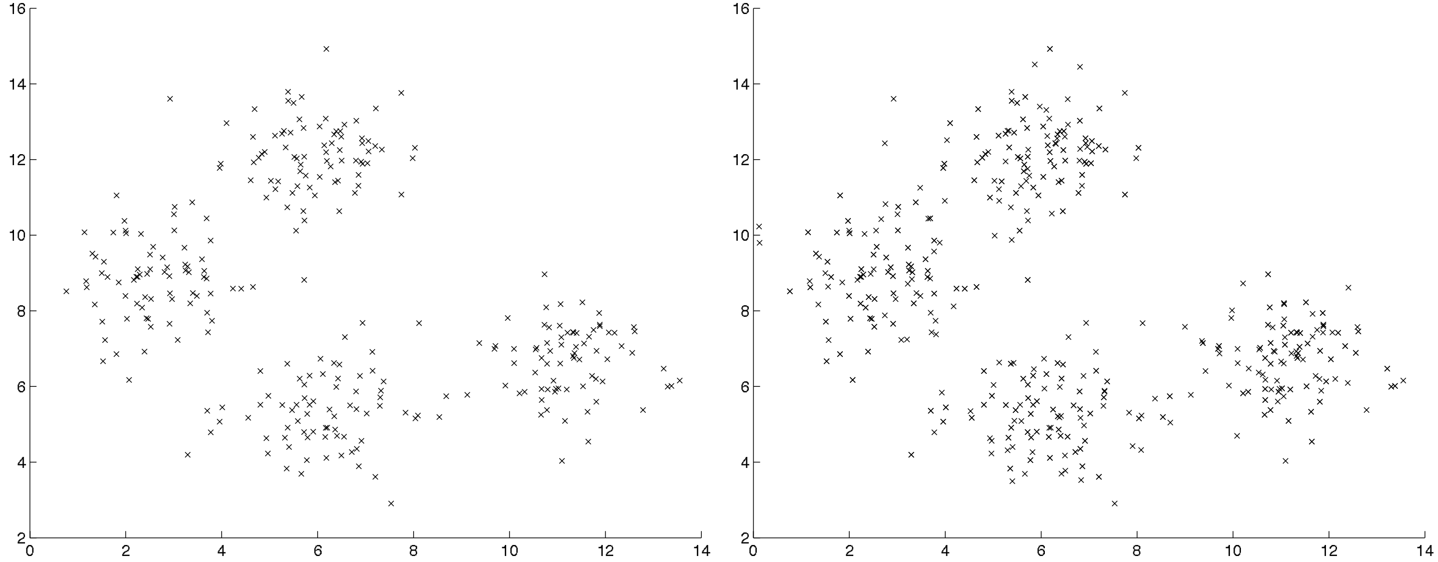
- *Procedure:*

  1. **Using a large sampling ratio ($0.9 > r > 0.7$), draw i.i.d. two samples $\mathcal{X}'_1$ and $\mathcal{X}'_2$ from $\mathcal{X}$.**
  2. **Comparison of the partitions $\mathsf{A}_k(\mathcal{X}'_1)$ and $\mathsf{A}_k(\mathcal{X}'_2)$.**
  3. **Repeat N times ($N \geq 100$) step 1. and step 2.**

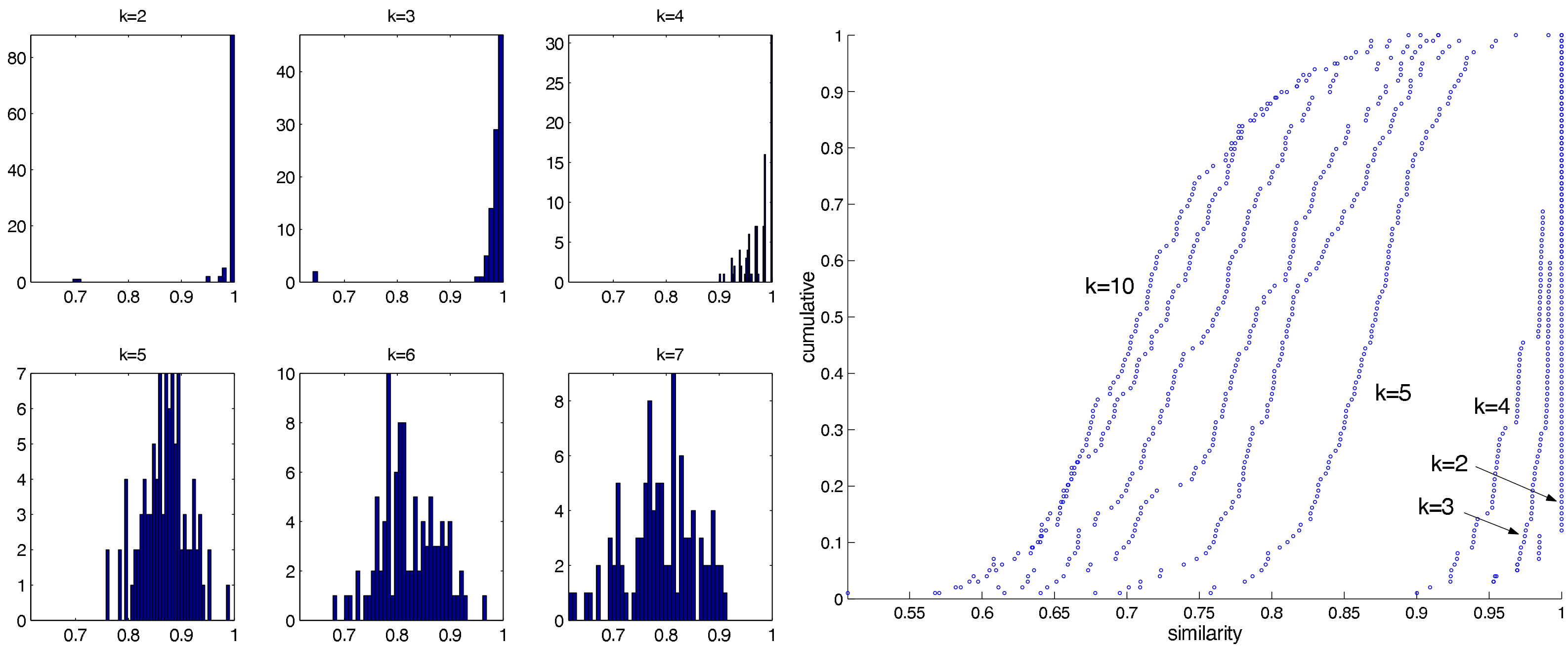
Cluster stability w.r.t. $\mathcal{X}$ if $\mathsf{A}_k(\mathcal{X}'_1)$ and $\mathsf{A}_k(\mathcal{X}'_2)$ are similar for most of the pairs of samples $\mathcal{X}'_1$ and $\mathcal{X}'_2$.

- *Alternative:* **replace $\mathcal{X}'_2$ by $\mathcal{X}$.**

Artificial data set



Correlation similarity



Asymptotic results

Empirical approach: Krieger and Green (1999)

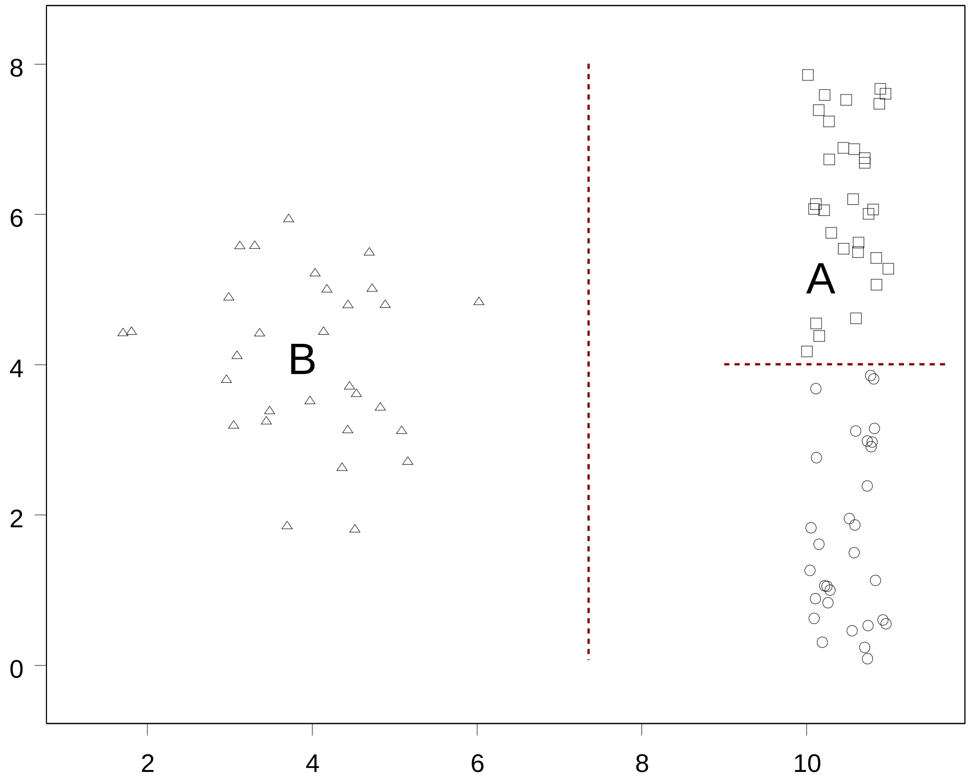
Theoretical proofs: Ben David, von Luxburg and Pal (2006)

For large sample size, stability is fully determined by the behavior of the objective function minimized by the clustering algorithm:

- **If the objective function has a unique global minimizer, the algorithm is stable;**
- **Otherwise, the algorithm is unstable.**

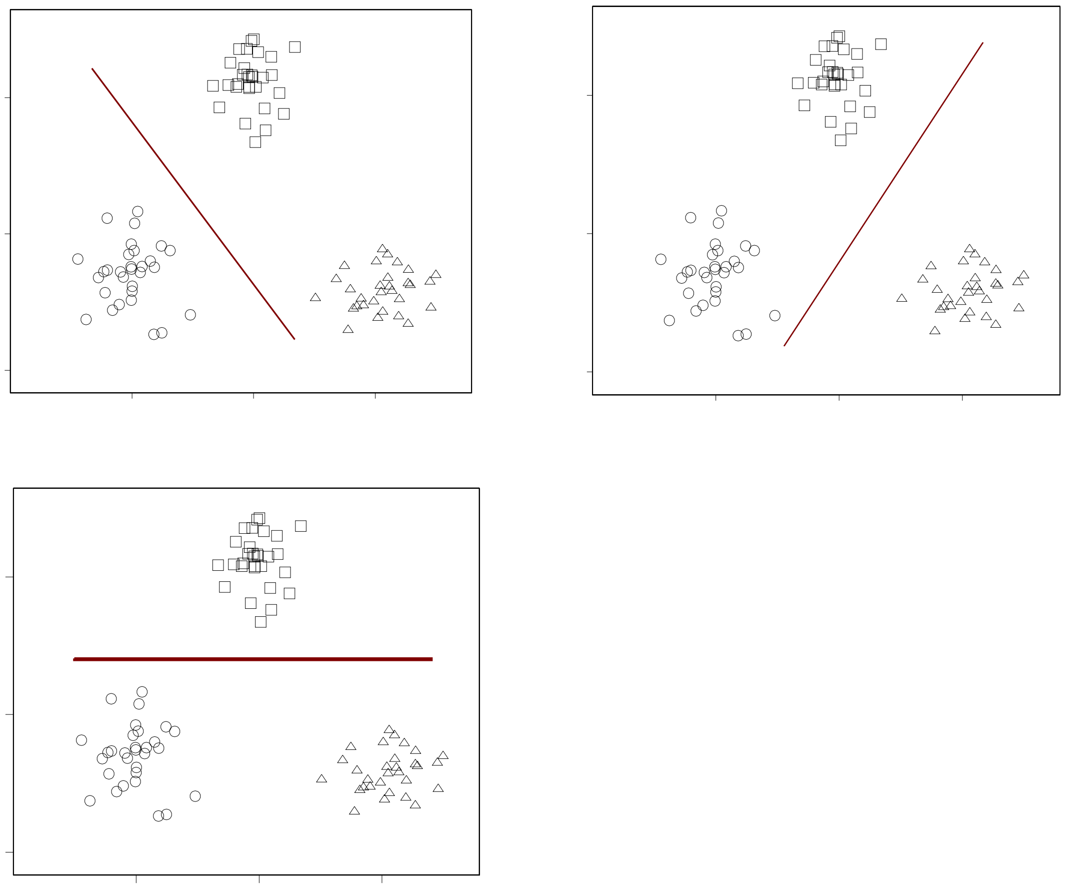
Example of a unique minimizer

A mixture of one gaussian distribution and one uniform distribution.



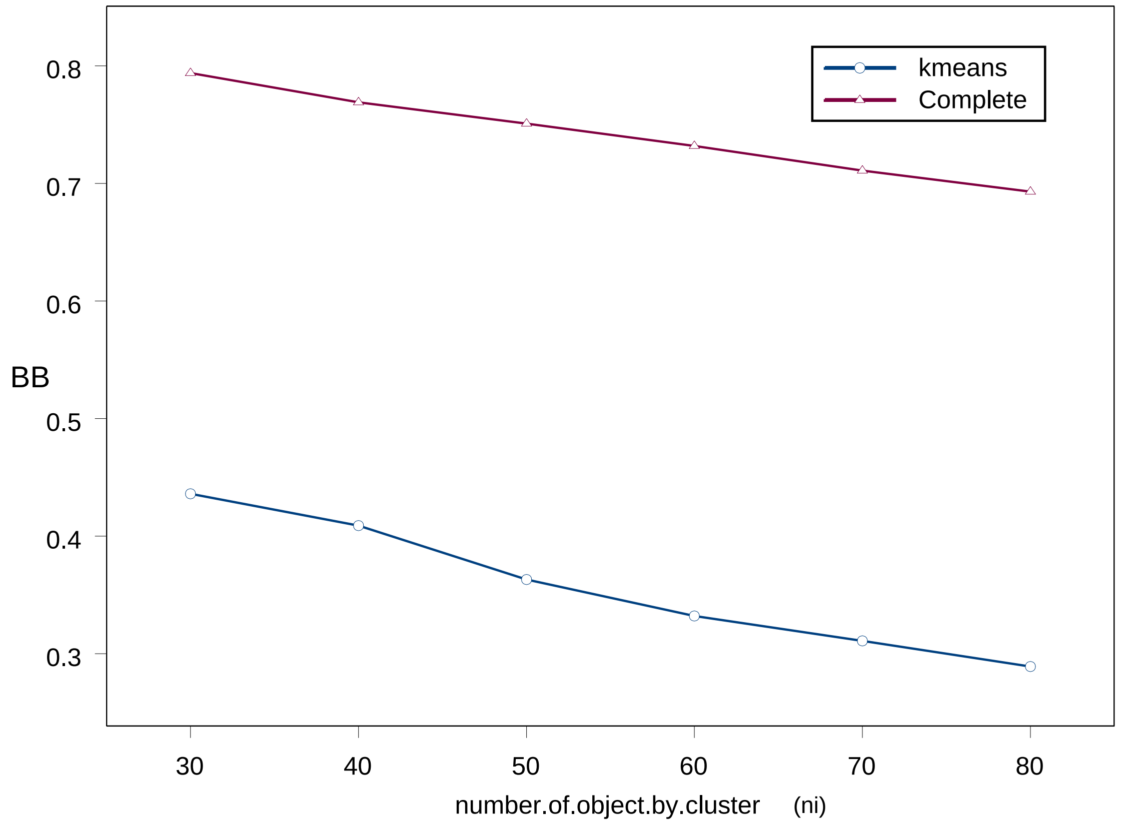
Example of instability from symmetry

A mixture of three gaussians



Example (continuing)

Stability measure of the partition vs. size of the data set



Some issues

For which purpose?

a) *well separated and/or homogeneous clusters;*

b) *cluster interpretability.*

Examples: data compression, data dissection.

How to identify 2 types of unstability?

a) *Several global minimizers;*

b) *When n is moderately large and some clusters are adjacent.*

How to asses stability values?

a) *Testing a null hypothesis of absence of structure;*

b) *Comparing stability values for different parameter values.*

A proposal for measuring cluster stability w.r.t. cohesion and isolation

Bertrand and Bel Mufti (2006)

(a) **Cohesion of a single cluster**

(b) **Isolation of a single cluster**

(c) **Stability of a single cluster**

(d) **The same three characteristics for a partition**

(e) **Influence of an individual object**

1. Stability measures

Perturbation by proportionate stratified sampling

- *Each perturbed data set is a sample.*
- n_C **= size of any cluster** C **in** $\mathcal{P}$ **;**
- n'_C **= size of** $C \cap \mathcal{X}'$
- **Sampling ratio:**

  $\boxed{r > 0.7}$
- **Proportionate stratified sampling:**

$$n'_C := \lfloor r n_C \rfloor \text{ so } n' \approx rn.$$

Isolation of a cluster

- **Isolation of cluster C:**

  "If two objects of $\mathcal{X}'$ are not clustered together by $\{C, \mathcal{X} \setminus C\}$, then they are not in the same cluster of $\mathcal{Q} = \mathsf{A}_k(\mathcal{X}')$."

- Measures to assess association rules;

- **Loevinger's measure** of rule $E \Rightarrow F$:

$$L(E \Rightarrow F) = 1 - P(E \cap \neg F)/P(E)P(\neg F)$$

- N samples are necessary to faithfully estimate the isolation of C:

$$\mathcal{X}'_1, \ldots, \mathcal{X}'_N.$$

- **Stability Measure:**

$$t^{is}(C, \mathcal{X}') = 1 - \frac{n'(n'-1)m_{(\mathcal{X}';C,\overline{C})}}{2n'_C(n'-n'_C)\, m_{(\mathcal{X}')}}\,,$$

where:

$m_{(\mathcal{X}')}$ **= number of pairs of (sampled) objects that are clustered together by** $\mathcal{Q} = \mathsf{A}_k(\mathcal{X}')$

$m_{(\mathcal{X}';C,\overline{C})}$ **= number of previous pairs for which exactly one of the two objects belongs to** C**.**

- $\bar{t}_N^{is}(C)$ **= average of** $t^{is}(C, \mathcal{X}')$ **for** N **samples** $\mathcal{X}'_i$**.**

Isolation between two clusters

- $\bar{t}_N^{is}(C, B)$: Isolation between cluster C and cluster B

  "**If** an object is in $\mathcal{X}' \cap C$ and another one in $\mathcal{X}' \cap B$, **then** they remain not clustered together by $\mathcal{Q}$."

- $\bar{t}_N^{is}(C) =$ weighted mean of $\bar{t}_N^{is}(C, B)$ for $B \in \mathcal{P}$

Isolation of a partition

- $\bar{t}_N^{is}(\mathcal{P})$: Isolation of all the clusters of $\mathcal{P}$

  "**If** two objects of $\mathcal{X}'$ are not clustered together by $\mathcal{P}$, **then** they remain not clustered together by $\mathcal{Q}$."

- $\bar{t}_N^{is}(\mathcal{P}) =$ weighted mean of $\bar{t}_N^{is}(C)$ for $A \in \mathcal{P}$

Other cluster features

- $\bar{t}_N^{co}(C)$: **Cohesion of cluster C.**

  "If two objects of $\mathcal{X}'$ belong to C,
  then they remain clustered together by $\mathcal{Q}$".

- $\bar{t}_N^{co}(\mathcal{P})$: **Cohesion of partition $\mathcal{P}$.**
  $\bar{t}_N^{co}(\mathcal{P})$ = **weighted mean of** $\bar{t}_N^{co}(C)$ **for** $A \in \mathcal{P}$

- **Stability of a cluster C**
- **Stability of a partition $\mathcal{P}$**

Self learning the number of samples

- **General notation:** $$\bar{t}_N(C) = \frac{1}{N}\sum_{i=1}^{N} t(C, \mathcal{X}'_i)$$

Which value of N should be choosed?

- **The central limit theorem**
- **Length of the approximate 95%-confidence interval**

p-value of a stability measure.

Internal criterion Jain and Dubes (1988), Gordon (1994).

- *Step 1.* **Define a null hypothesis H_0 that specifies the absence of cluster structure for the data set under investigation;**
- *Step 2.* **Estimate the probability significance (p-value), under H_0, of the observed value of the measure of stability by performing a Monte Carlo test**

Random position hypothesis. **The n points of the data set $\mathcal{X}$ are equally likely in a region (convex hull of the data set).**

"Optimal number" of clusters.

- k is an optimal number of clusters when partitional stability is a local maximum.
- refinement:
  - Stability of isolation and cohesion, separately.
  - Stability of a partition can be interpreted as a weighted average of the stability of its clusters.
  - p-value of each stability measure.

2. Comparison with other validation measures

- **The index of Calinski and Harabasz (1974):**

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)}$$

$B(k)$ **and** $W(k)$ **: between and within cluster sums of squares of the partition, respectively.**

- **The index of Krzanowski and Lai (1985):**

$$KL(k) = |\frac{DIFF(k)}{DIFF(k+1)}|$$

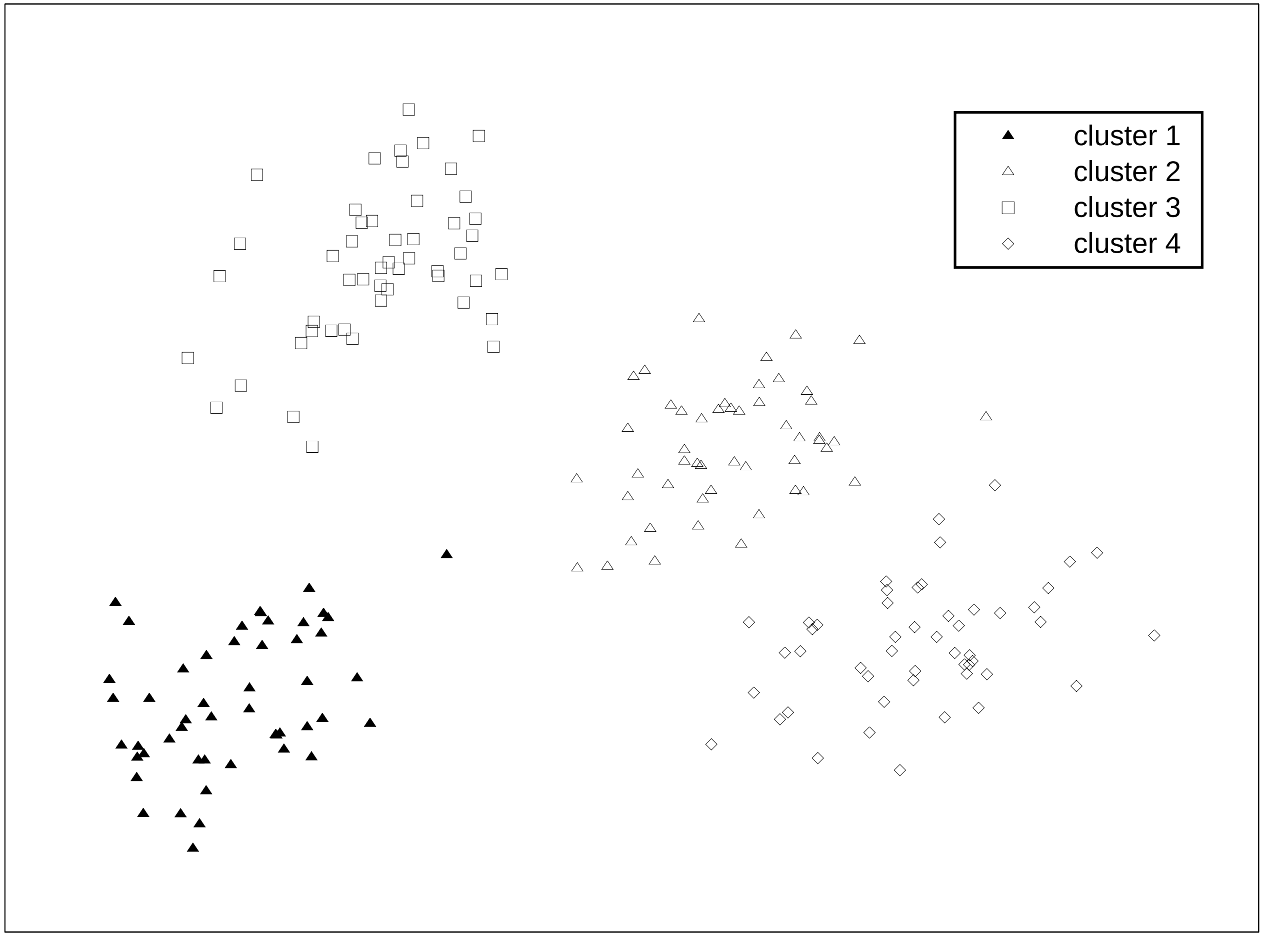
$$DIFF(k) = (k-1)^{2/p}W(k-1) - (k)^{2/p}W(k),$$

p **= number of features in the data set.**

- **The Gap statistic (2001):**

$$Gap(k) = E^{\star}[log(W(k))] - log(W(k))$$

Artificial data set



		Number of clusters (k)				
Index		**2**	**3**	**4**	**5**	**6**
$CH(k)$		**145**	**414**	**580***	**494**	**446**
$KL(k)$		**.26**	**3.36**	**3.89**	**1.39**	**5.95 ***
$Gap(k)$		**0.17**	**0.82**	**1.05 ***	**0.96**	**0.89**
$BBM(k)$		**.779**	**.958**	**.992 ***	**.914**	**.816**
P-value of $BBM(k)$	$(\%)$	$48-61$	$2.4-6.8$	$0-1$	$0-4.5$	$2.5-9.2$

* indicates the optimal number of clusters

Stability measures and p-values

		Isolation		*Cohesion*		*Stability*	
			%		%		%
Cluster	**1**	**.990**	$0-1$	**.980**	$0-5$	**.986**	$0-1$
	2	**.984**	$0-1$	**.992**	$0-2$	**.987**	$0-1$
	3	**1.**	$0-1$	**1.**	$0-1$	**1.**	$0-1$
	4	**.994**	$0-1$	**.996**	$0-2$	**.995**	$0-1$
Partition		**.992**	$0-1$	**.992**	$0-1$	**.992**	$0-1$

Stability measures and p-values

(5-partition)

		Isolation	%	*Cohesion*	%	*Stability*	%
Cluster	**1**	**.993**	$0-1$	**.939**	$0-1$	**.973**	$0-1$
	2	**.993**	$0-1$	**.936**	$0-1$	**.972**	$0-1$
	3	**.989**	$0-5$	**.873**	$1-13$	**.945**	$0-8$
	4	**.696**	$32-49$	**.798**	$48-65$	**.716**	$34-50$
	5	**.727**	$29-47$	**.980**	$1-9$	**.777**	$22-39$
Partition		**.915**	$0-4.5$	**.913**	$0-1$	**.914**	$0-4.5$

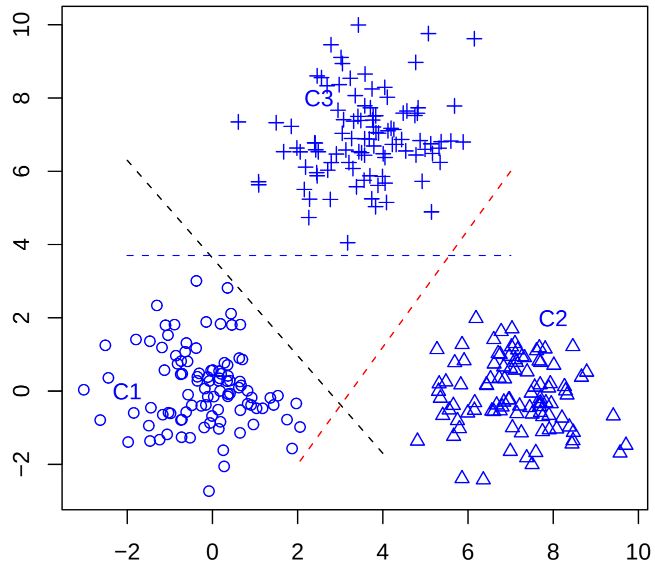
Iris data

		Number of clusters (k)				
Index		**2**	**3**	**4**	**5**	**6**
$CH(k)$		**795.7**	**1211.2**	**1266.4**	**1358.7** *	**1154.4**
$KL(k)$		**4.83**	**6.01** *	**1.30**	**1.12**	**1.19**
$Gap(k)$		**.68**	**1.28**	**1.48**	**1.61** *	**1.39**
BBM(k)		**.992** *	**.959**	**.881**	**.900**	**.870**
P-value of BB**M**(k)	$(\%)$	$.3 - 3.4$	$6.7 - 11.9$	> 34	$5.2 - 9.4$	$4.9 - 9.6$

* **indicates the optimal number of clusters**

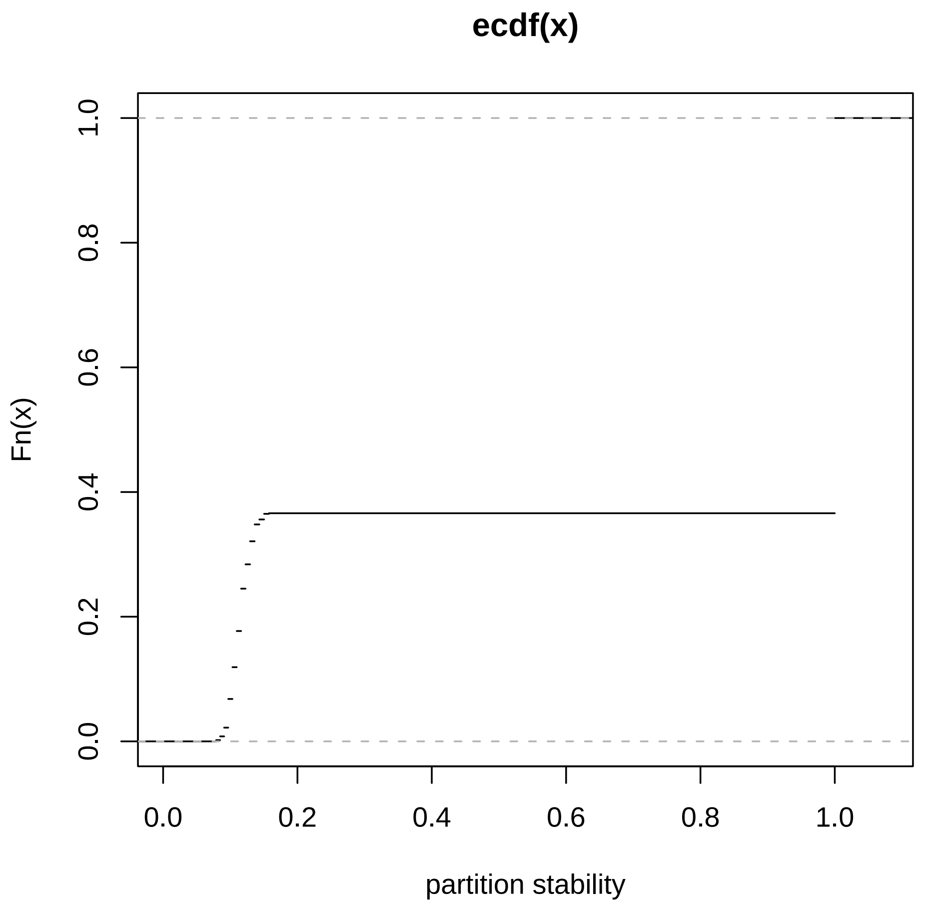
Characterizing different types of unstability

Data set #1: 3 symmetrical Gaussians
Partition: 2 clusters



$$n = 300$$

3 symmetrical gaussians (continuing)



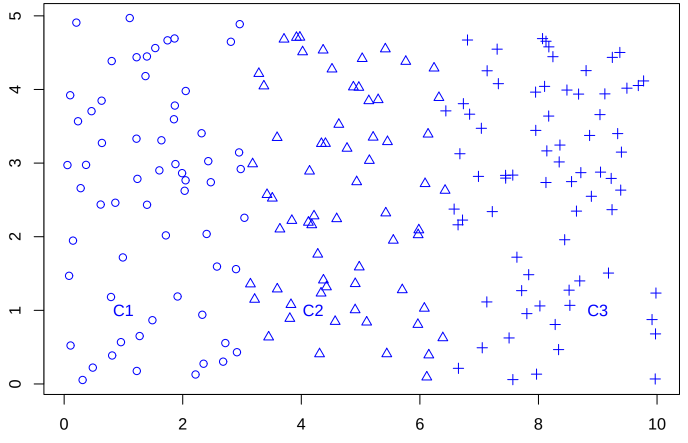
Stability measures

	C_1	$\{C_2, C_3\}$	Partition
Cohesion	1	0.594	0.675
Isolation	0.676	0.676	0.676
Stability	0.731	0.639	0.675

Based on 1000 bootstrapped samples:

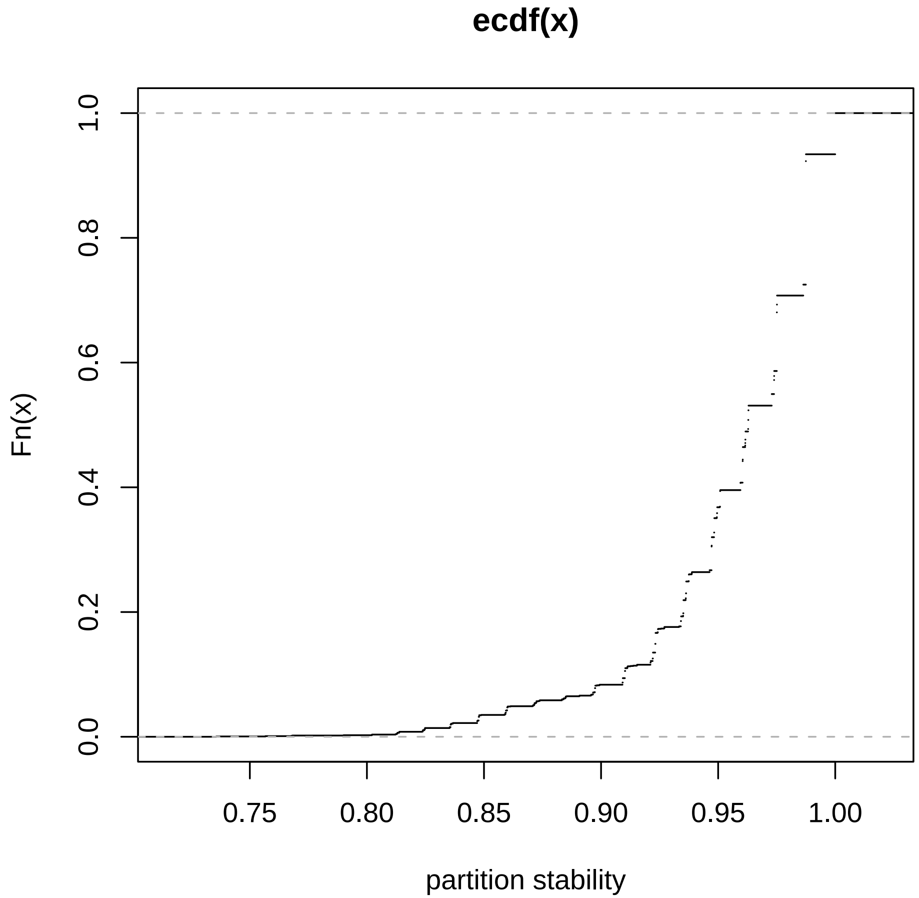
- $IC_{95\%} = [\,0.433, 1\,]$

Data set #2: Uniform data set
Partition: 3 clusters



$$n = 300$$

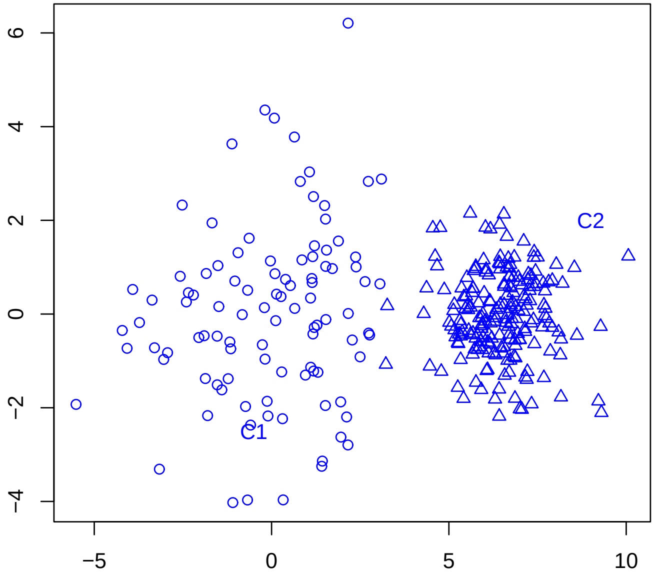
Uniform data set (continuing)



Stability measures

	C_1	C_2	C_3	Partition
Cohesion	0.951	0.879	0.928	0.919
Isolation	0.936	0.877	0.940	0.918
Stability	0.949	0.877	0.936	0.918

Data set #3: 2 Gaussians with different variances
Partition: 2 clusters



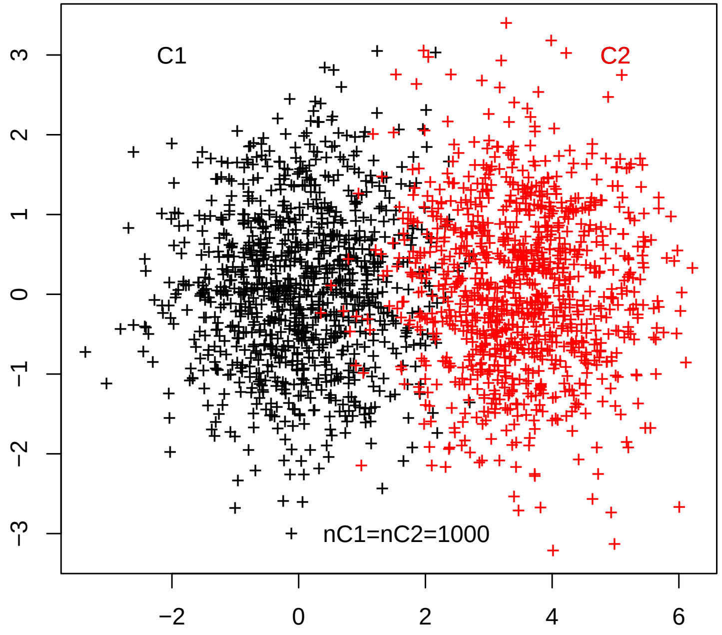
$$n = 300$$

Stability measures

	C_1	C_2	Partition
Cohesion	0.961	1	0.992
Isolation	0.984	0.984	0.984
Stability	0.980	0.991	0.988

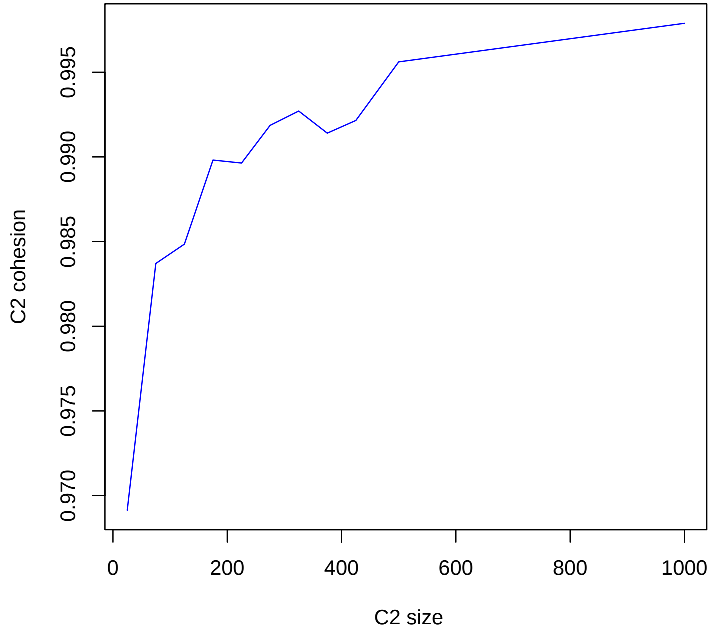
Data sets #4: 2 Gaussians with same variances
Partition: 2 clusters

Cluster sizes are increasing from 25 to 425 by step of 25, and then take values 500 and 1000.



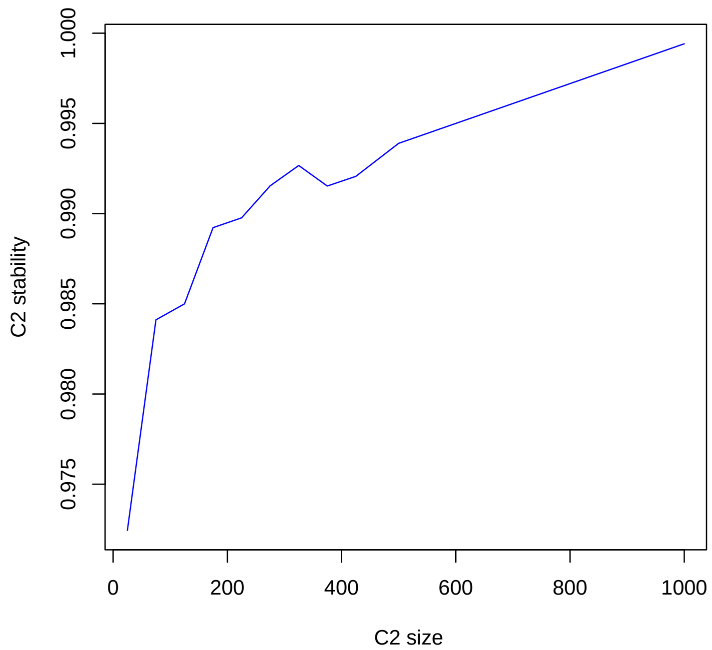
$$50 \leq n \leq 1000$$

Data sets #4 (continuing)



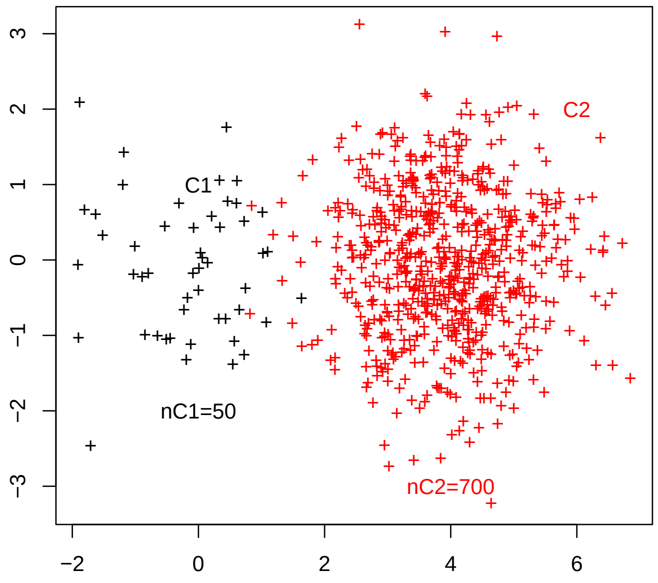
Cohesion of C_2 versus data size

Data sets #4 (continuing)



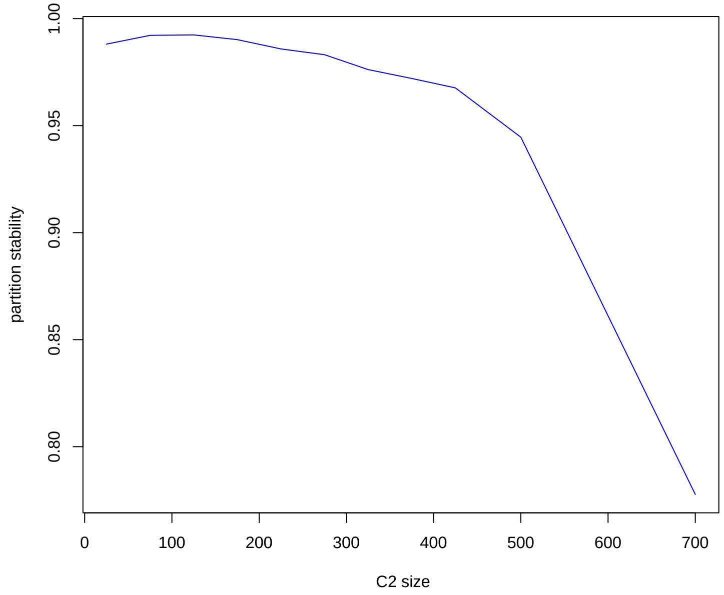
Stability of C_2 versus data size

Data sets #5: 2 Gaussians with same variances
Partition: 2 clusters
Only C_2 size increasing from 25 to 700



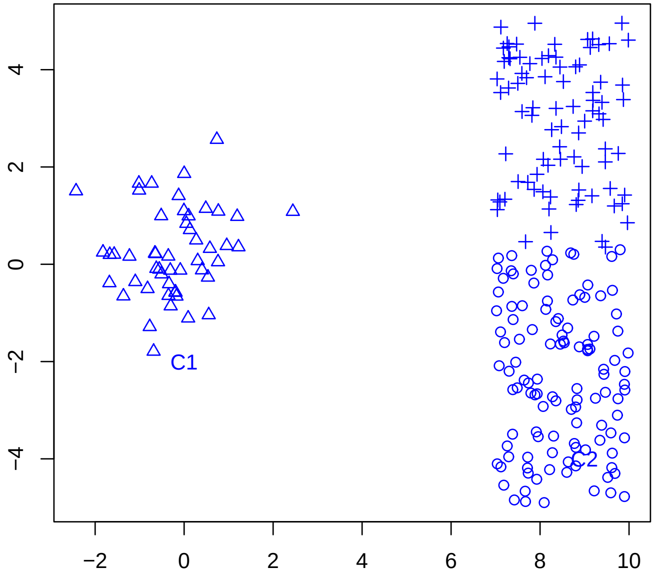
$$25 \leq |C_2| \leq 700$$

Data sets #5 (continuing)



Partition stability versus C_2 size

Data set #6: Mixture of 1 Gaussian and 1 uniform law
Partition: 3 clusters



$$n = 200$$

Stability measures

	C1	C2	C3	Partition
Cohesion	1 (0 %)	0.953 (59 %)	1 (0 %)	0.984 (3 %)
Isolation	1 (0 %)	0.976 (16 %)	0.976 (14 %)	0.984 (4 %)
Stability	1 (0 %)	0.968 (28 %)	0.983 (9 %)	0.984 (3 %)

Two individual scores

$$
\begin{aligned}
J &= \{1, \ldots, N\}, \\
J(x) &= \{j \in \{1, \ldots, N\} : x \in \mathcal{X}'_j\}, \\
\mathcal{P}(x) &= \{z \in X : x \text{ and } z \text{ are clustered together in } \mathcal{P}\}, \\
\mathcal{P}^\star(x) &= \mathcal{P}(x) \setminus \{x\}.
\end{aligned}
$$

- **Partial Membership:** $\widehat{M}(x, A) = \dfrac{1}{|J(x)|} \displaystyle\sum_{j \in J(x)} \dfrac{|\ \mathcal{P}^\star_j(x) \cap A\ |}{|\ \mathcal{P}^\star_j(x)\ |}$

- **Partial Filiation:** $\widehat{F}(x, A) = \dfrac{1}{|J(x)|} \displaystyle\sum_{j \in J(x)} \dfrac{|\ \mathcal{P}^\star_j(x) \cap A\ |}{a^*}$

- **Decomposition:** $\bar{t}^{\,is}_N(A) = 1 - \frac{1}{\widehat{p}} \sum_{x \in A} \frac{|J(x)|}{\sum_{x \in A} |J(x)|} \widehat{cF}(x, A)$

Membership scores of intermediary points

Iris data

Objects (x)	Cluster	Iris cluster	$\lvert J(x) \rvert$	1	2	3
#84	2	1	405	.27	.73	0
#120	1	2	397	.88	.12	0
#122	2	2	387	.22	.78	0
#124	1	2	376	.70	.30	0
#127	1	2	403	.88	.12	0
#128	1	2	394	.68	.32	0
#134	1	2	398	.68	.32	0
#139	1	2	391	.88	.28	0
#150	2	2	391	.07	.93	0

Filiation scores of intermediary points
Iris data

Objects (x)	Cluster	Iris cluster	$\|J(x)\|$	1	2	3
#84	2	1	405	.27	.78	0
#120	1	2	397	.89	.13	0
#122	2	2	387	.22	.83	0
#124	1	2	376	.75	.30	0
#127	1	2	403	.88	.15	0
#128	1	2	394	.75	.29	0
#134	1	2	398	.73	.32	0
#139	1	2	391	.88	.14	0
#150	2	2	391	.05	.98	0

Some conclusions and perspectives:

In the view of exploratory data analysis,

- For all values of n, the interpretation of stability values is easier with:
  a) Stability measures that concern isolation and cohesion for each cluster;
  b) Cumulative distribution function of the partitional stability measure.
- If a the cohesion of a cluster is assessed to be large, then its dispersion is certainly larger than its neighbors dispersions, but the converse is not true.
- Individual scores and small groups of outliers.
- Assuming "clusters of equal sizes", stability seems to be more informative for small and medium size data sets.