Stability-based methods to discover structures in bio-molecular data

Giorgio Valentini
{valentini}@dsi.unimi.it

DSI - Dipartimento di Scienze dell'Informazione

Università degli Studi di Milano

Motivations and objectives

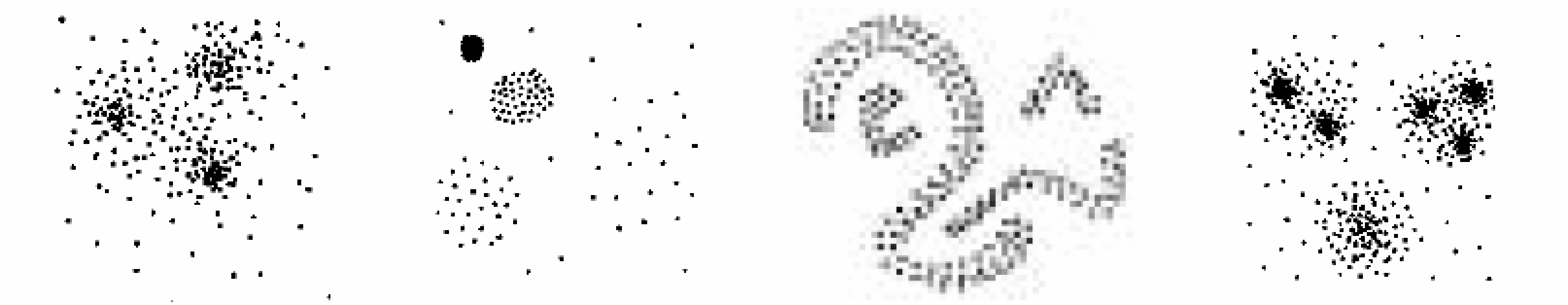
Motivations:

- Several molecular biology problems require unsupervised analysis of bio-molecular data
- Most clustering algorithms find structures in bio-molecular data even if no structure is present in the data.
- Clustering solutions need to be evaluated and validated
- Classical validity indices are biased towards specific clustering algorithms (Jain et al. 1999)
- Multiple structures may be simultaneously present into the data
- We need to estimate the statistical significance of clustering solutions

Objectives:

- Development of stability-based methods designed to discover structures in clustered bio-molecular data:
  - Assessment of the reliability of a given clustering solution
  - Model order selection
  - Assessment of the statistical significance of clustering solutions
  - Discovery of multiple and hierarchical structures in the bio-molecuar data

Clustering



- Grouping a set of data objects into clusters
- Cluster: a collection of data objects:
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Clustering is an ***unsupervised method*** (no labeled examples)

Typical usage:

As a *stand-alone tool* to get insights into data distribution

As a *preprocessing step* for other algorithms

Bioinformatics application examples of clustering

- Inferring unknown gene functions from clusters
- Discovering functionally related sets of genes
- Discovering new subclasses of diseases
- Discovering regulatory networks
- …

For instance, clustering permits us to

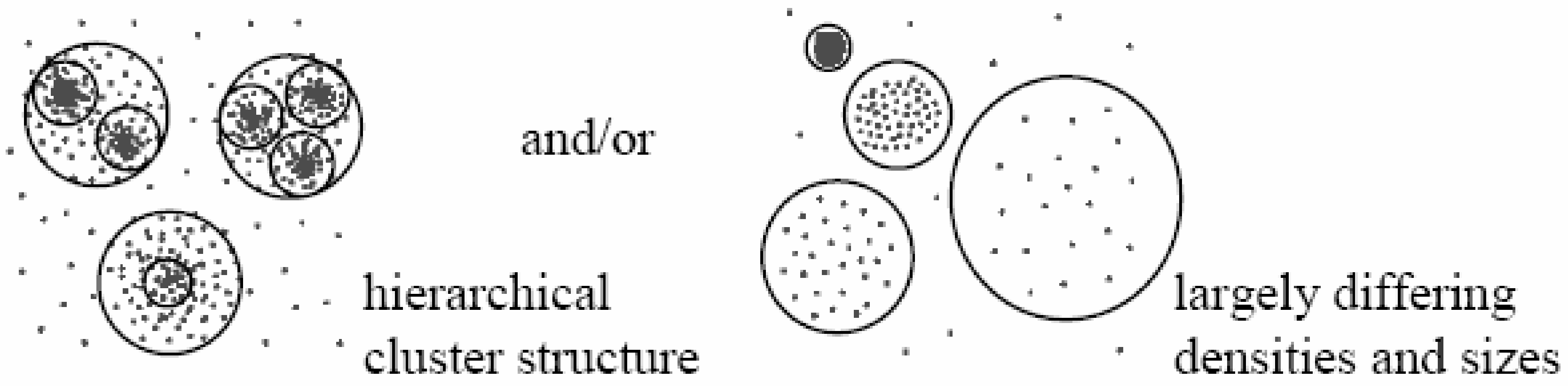
– group together genes that respond similarly across several experimental conditions

– group together different examples with similar expression patterns across the whole genome

Major problems with clustering algorithms

- What about the reliability of a given clustering solution?
  - What about the “natural” number of clusters?
  - What about the reliability of each individual cluster inside a given clustering?
- What about the statistical significance of a given clustering solution?
- What about (possible) multiple structures present in the data?

How many clusters in the data?


and/or
hierarchical cluster structure
largely differing densities and sizes

Model order selection through stability-based procedures

The main idea: *a clustering is considered reliable if it is approximately maintained across multiple perturbations.*

Induce *random perturbations* on a data set by:

- *subsampling*, (*BenHur et al*, 2002);
- *noise injection*, (*Mc Shane et al*, 2003),
- *random projections* (*Smolkin and Gosh*, 2003).

then apply a *clustering algorithm*

A general stability based procedure to estimate the reliability of a given clustering

1. Randomly perturb the data many times according to a given perturbation procedure.
2. Apply a given clustering algorithm to the perturbed data
3. Apply a given clustering similarity measure (e.g. Jaccard similarity) to multiple pairs of k-clusterings obtained according to steps 1 and 2.
4. Use the similarity measures to assess the stability of a given clustering.
5. Repeat steps 1 to 4 for multiple values of k and select the most stable clusterings as the most reliable.

Similarity measures between clusterings (1)

A k-clustering: $C = < A_1, A_2, \ldots, A_k >$

may be represented through a *pairwise similarity matrix M:*

$$M_{ij} = \begin{cases} 1 & \text{if } x_i \text{ and } x_j \text{ belong to the same cluster, } i \neq j \\ 0 & \text{otherwise} \end{cases}$$

Consider two clusterings $C^{(1)}$ and $C^{(2)}$ and the corresponding $M^{(1)}$ and $M^{(2)}$ matrices. Using:

N_{ij} = number of entries for which both $M^{(1)}$ and $M^{(2)}$ have respectively values i and j , i,j ε {0,1},

we can compute *similarity indices* between clusterings

Similarity measures between clusterings (2)

Matching coefficient:

$$M(C^{(1)}, C^{(2)}) = \frac{N_{00} + N_{11}}{N_{00} + N_{11} + N_{01} + N_{10}}$$

Jaccard coefficient:

$$J(C^{(1)}, C^{(2)}) = \frac{N_{11}}{N_{11} + N_{01} + N_{10}}$$

Fowlkes and Mallows coefficient:

$$F(C^{(1)}, C^{(2)}) = \frac{N_{11}}{\sqrt{(N_{01} + N_{11})(N_{10} + N_{11})}}$$

Using the distribution of the similarities to estimate the stability (1)

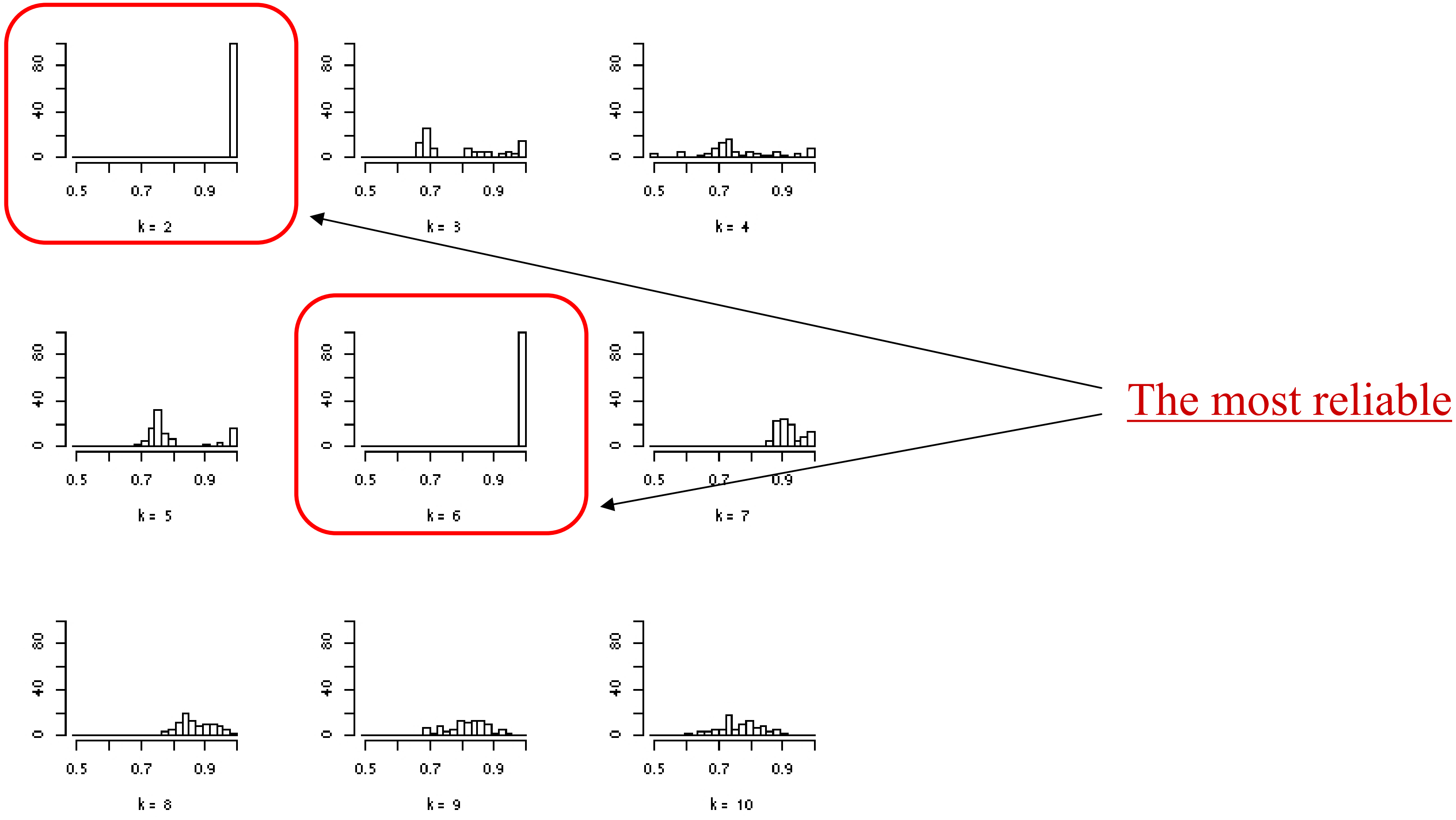
We may consider multiple pairs of clusterings obtained from pairs of perturbed data and then measure their pairwise similarities, using the *sim* coefficients previously defined:

$$S_{kj} = sim\left(C(\rho^{(1)}_{kj}(D),k), C(\rho^{(2)}_{kj}(D),k)\right), \quad 1 \leq j \leq n$$

Then we could compare the distribution of S_k, for different numbers of clusters *k (Ben Hur et al. 2002)*

Using the distribution of the similarities to estimate the stability (2)

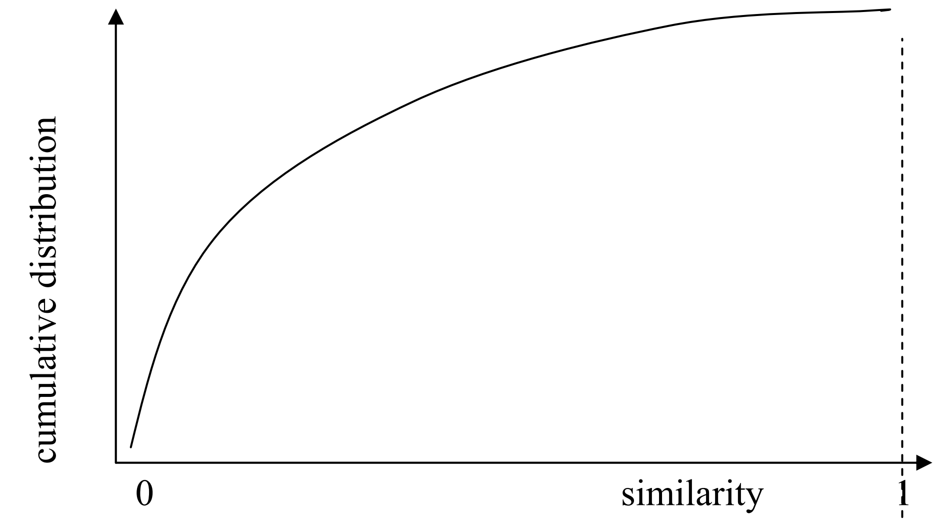
The intuitive idea is that if S_k is concentrated close to 1, the corresponding clustering is stable with respect to a given controlled perturbation and hence it is reliable:



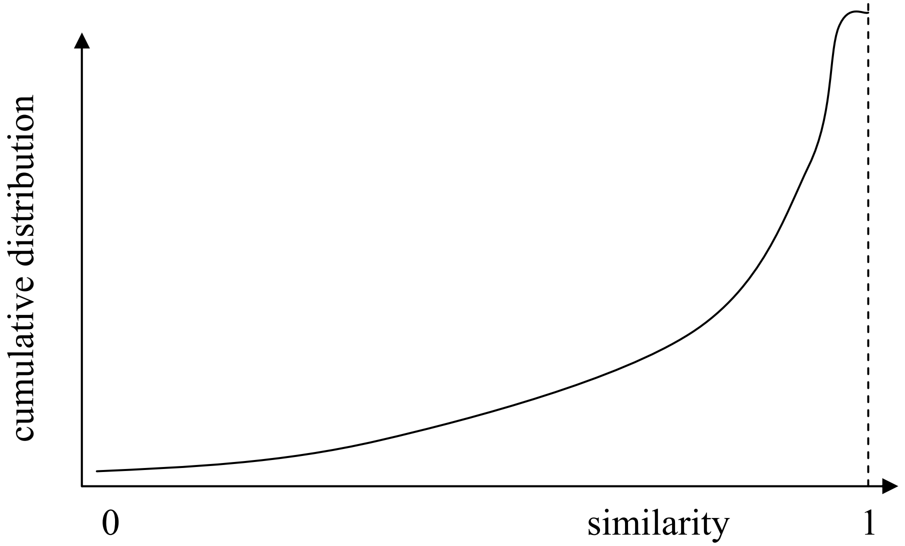
The most reliable

A quantitative estimate of clustering stability (1)

- The main idea: consider the *cumulative distribution* of the similarities:



Similarities are spread across multiple values: the clustering is unstable



Similarities are cumulated close to 1: the clustering is stable

Observe that the area below the cumulative distribution is smaller …

A quantitative estimate of clustering stability (2)

- S_k ($0 \leq S_k \leq 1$) is the random variable that represents the similarity between two k-clusterings, and $f_k(s)$ is its density function.

$$F_k(\bar{s}) = \int_{-\infty}^{\bar{s}} f_k(s)ds \qquad g(k) = \int_0^1 F_k(s)ds$$

- g(k) is a parameter of concentration (*Bertoni and Valentini*, 2007)

If g(k) ~ 0 → The clustering is very reliable

If g(k) ~ 1 → The clustering is very unreliable

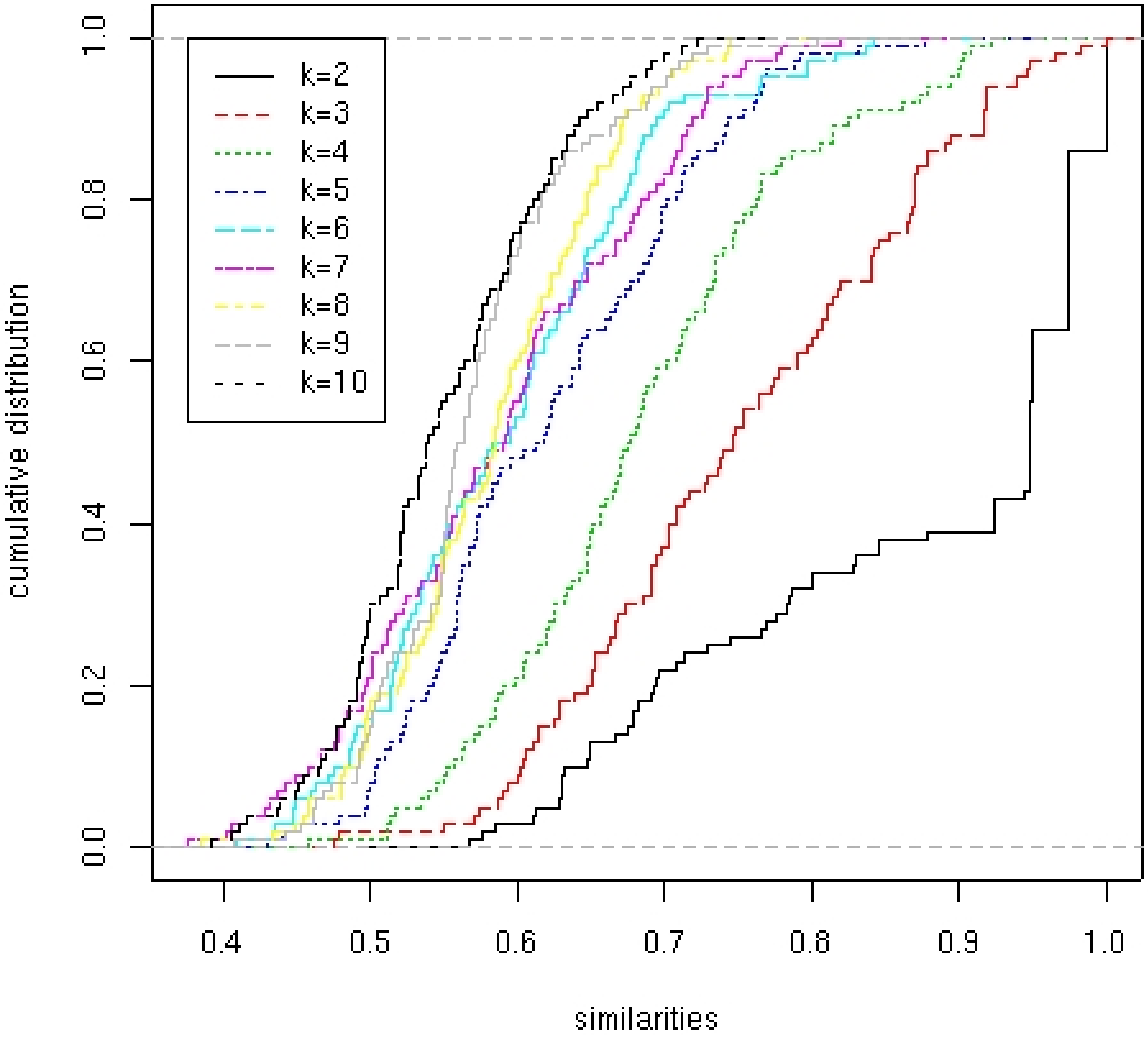
Equivalently $E[S_k]$ can be used as a good index of the reliability of the k-clusterings. Indeed:

$$\left.\begin{aligned} E[S_k] &= \int_0^1 s f_k(s)ds = \int_0^1 sF'_k(s)ds = 1 - \int_0^1 F_k(s)ds = 1 - g(k) \\ Var[S_k] &= E[S_k^2] - E[S_k]^2 \leq E[S_k] - E[S_k]^2 = g(k)(1 - g(k)) \end{aligned}\right\}$$

- $E[S_k]$ may be estimated through the empirical means ξ_k: $\xi_k = \sum_{j=1}^{n} \frac{S_{kj}}{n}$ where

$$S_{kj} = sim\left(\mathcal{C}(\rho_{kj}^{(1)}(D), k), \mathcal{C}(\rho_{kj}^{(2)}(D), k)\right), \quad 1 \leq j \leq n$$

Cumulative distribution of similarities for different number of clusters k



Empirical means are computed for different numbers of clusters k

Assessment of the most reliable clusterings

- We may perform a sorting of the ξ_k : $(\xi_2, \xi_3, \ldots, \xi_{H+1}) \overset{sort}{\rightarrow} (\xi_{p(1)}, \xi_{p(2)}, \ldots, \xi_{p(H)})$

  $\xi_{p(1)} \geq \xi_{p(2)} \geq \ldots \geq \xi_{p(H)}$

- The *p(1)*-clustering is the most reliable, while *p(H)* is the least reliable: $0 \leq \xi_i \leq 1$provides a *stability score* of the obtained clusterings

- Exploiting this ordering we would establish: *which are the significant clusterings discovered in the data* ?

- Making assumtpions about the distribution of the reliability scores (normal distribution) a χ^2- *based statistical test* has been proposed (*Bertoni* and *Valentini*, 2007).

A χ^2-based method to estimate the significance of the discovered clusterings (1)

- Perform a sorting of the ξ_k : $(\xi_2, \xi_3, \ldots, \xi_{H+1}) \overset{sort}{\longrightarrow} (\xi_{p(1)}, \xi_{p(2)}, \ldots, \xi_{p(H)})$

  p is the index permutation such that: $\xi_{p(1)} \geq \xi_{p(2)} \geq \ldots \geq \xi_{p(H)}$

- For each *k*-clustering consider the random variable S_k (recall that $E[S_k]$ is the stabiliy index)
- For all *k* and for a fixed threshold t^o consider :

  $B_k = I(S_k > t^o)$ (*Bernoulli* random variable , *I* the indicator function)

$\Downarrow$

$$X_k = \sum_{j=1}^{m} B_k^j \sim \mathrm{B}(\mathrm{m}, \theta_k), \quad \theta_k = Prob(I(S_k > t^o))$$

X_k is distributed according to a binomial distribution

A χ^2-based method to estimate the significance of the discovered clusterings (2)

$$\frac{X_k - m\theta_k}{\sqrt{m\theta_k(1-\theta_k)}} \sim \mathrm{N}(0,1)$$

Hypothesis: All the binomial populations are drawn form the same distribution, that is the clusterings are all equally reliable ⟺ $\forall k \in \mathrm{K}, \theta_k = \theta$

⇓

$$Y = \sum_{k \in K} \frac{(X_k - m\hat{\theta})^2}{m\hat{\theta}(1-\hat{\theta})} \quad \text{with} \quad \hat{\theta} = \frac{\sum_{k \in K} X_k}{|K| \cdot m}$$

Y distributed according to a χ^2 with |K|-1 degrees of freedom

A χ^2-based method to estimate the significance of the discovered clusterings (3)

- Using the previous *Y* statistic we can test the following alternative hypotheses:

  - *Ho: all the θ_k are equal to θ (the considered set of k-clustering are equally reliable)*

  - *Ha: the θ_k are not all equal between them (the considered set of k-clustering are not equally reliable)*

- If $Y \geq \chi^2_{\alpha,|K|-1}$ we may reject the null hypothesis at α significance level, that is we may conclude that with probability 1-α the considered proportions are different, and hence that at least one *k*-clustering significantly differs from the others.

A χ^2-based method to estimate the significance of the discovered clusterings (4)

An *iterative procedure* to detect the significant number(s) of clusterings:

1. Consider the ordered vector $\xi = (\xi_{p(1)}, \xi_{p(2)}, \ldots, \xi_{p(H)})$

2. Repeat the χ^2-based test until no significant difference is detected or the only remaining clustering is *p(1)* (the top-ranked one). At each iteration, if a significant difference is detected, remove the bottom-ranked clustering from ξ.

Output: set of the remaining (top sorted) *k*-clusterings that correspond to the set of the estimate stable number of clusters (at α significance level).

Drawbacks of the χ^2- *based statistical test*

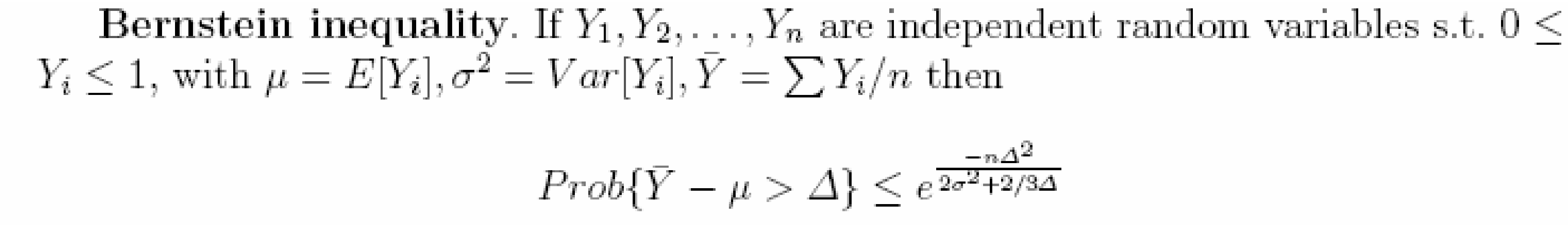
1. A priori assumptions about the distribution of the similarity values needed to estimate the reliability of the obtained clusterings
2. Test results depend on the choice of user-defined parameters.

An alternative approach based on the *Bernstein inequality:*

1. No assumptions about the distribution of the similarity values.
2. No requirements of any user-defined additional parameters

<u>It may be applied to a large range of bioinformatics problem</u>

A test of hypothesis based on *Bernstein inequality* to estimate the significance of the discovered clusterings (1)



Bernstein inequality. If $Y_1, Y_2, \ldots, Y_n$ are independent random variables s.t. $0 \leq Y_i \leq 1$, with $\mu = E[Y_i], \sigma^2 = Var[Y_i], \bar{Y} = \sum Y_i/n$ then

$$Prob\{\bar{Y} - \mu > \Delta\} \leq e^{\frac{-n\Delta^2}{2\sigma^2+2/3\Delta}}$$

The above inequality is used to build up the following **test of hypothesis**:

H_0 : *p(1)*-clustering is not more reliable than *p(r)*-clustering, $2 \leq r \leq H$,
that is $\mathrm{E}[S_{p(1)}] \leq \mathrm{E}[S_{p(r)}]$

H_a : *p(1)*-clustering is more reliable than *p(r)*-clustering,
that is $\mathrm{E}[S_{p(1)}] > \mathrm{E}[S_{p(r)}]$

We apply an *iterative procedure* estimating the reliability of the first ranked clustering with respect to the last (H-ranked), then to the H-1, H-2, … until a significant difference is detected or until it only remains the first top-ranked clustering.

A test of hypothesis based on *Bernstein inequality* to estimate the significance of the discovered clusterings (2)

Given the following random variables: $P_i = S_{p(1)} - S_{p(i)}$ and $X_i = \xi_{p(1)} - \xi_{p(i)}$

Considering the first and last ranked clusterings H_0 becomes:

$$\mathrm{E}[S_{p(1)}] \leq \mathrm{E}[S_{p(H)}] \longrightarrow \mathrm{E}[S_{p(1)}] - \mathrm{E}[S_{p(H)}] = \mathrm{E}[P_H] \leq 0$$

Fixing a parameter $\Delta \geq 0$, if H_0 is true, using the *Bernstein inequality,* we have:

$$Prob\{X_H \geq \Delta\} \leq Prob\{X_H - E[P_H] \geq \Delta\} \leq e^{\frac{-n\Delta^2}{2\sigma^2 + 2/3\Delta}}$$

Considering a measured value $\hat{X}_H$ of the random variable X_H, setting $\hat{X}_H = \Delta$, the probability of type I error is:

$$P_{err}\{X_H \geq \hat{X}_H\} \leq e^{\frac{-n\hat{X}_H^2}{2\sigma_H^2 + 2/3\hat{X}_H}}$$

If $P_{err}\{X_H \geq \hat{X}_H\} < \alpha$ *we reject the null hypothesis H_0:*

a significant difference between the two clusterings is detected at α significance level

A test of hypothesis based on *Bernstein inequality* to estimate the significance of the discovered clusterings (3)

We perform an *iterative procedure*, exploiting the ordering of the stability scores: if H_0 is rejected for the *p(H-r+1)*-clustering, then we consider the *p(H-r)*-clustering, estimating by union bound the type I error:

$$P_{err}(H-r) = Prob\{\bigvee_{H-r\le i\le H} X_i \ge \hat{X}_i\} \le \sum_{i=H-r}^{H} Prob\{X_i \ge \hat{X}_i\} \le \sum_{i=H-r}^{H} e^{\frac{-n\hat{X}_i^2}{2\sigma_i^2+2/3\hat{X}_i}}$$

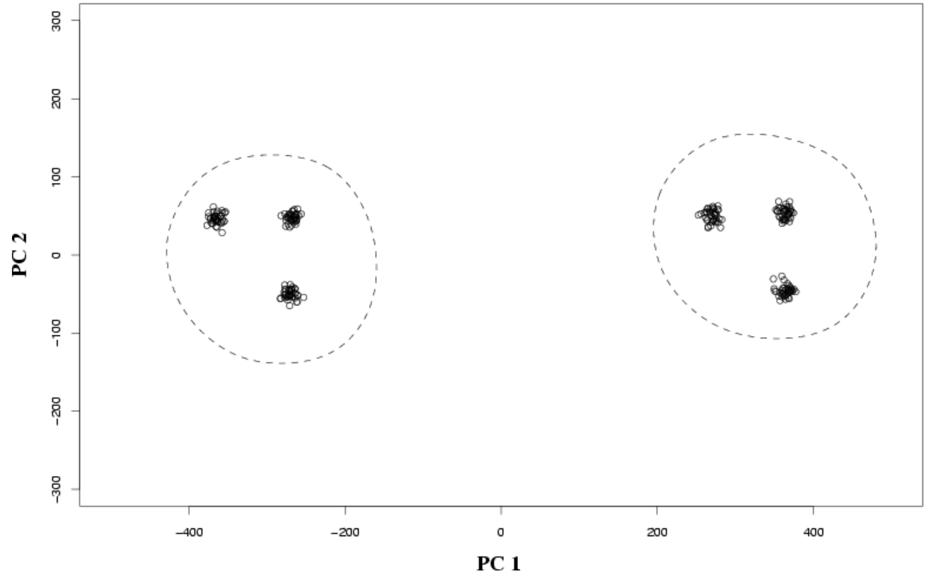
The iterative procedure stops if one of these 2 cases succeeds:

1. H_0 is rejected till to $r = H\text{-}2$. ➡ { The only reliable clustering at α significance level is the top ranked one, i.e. *p(1)*

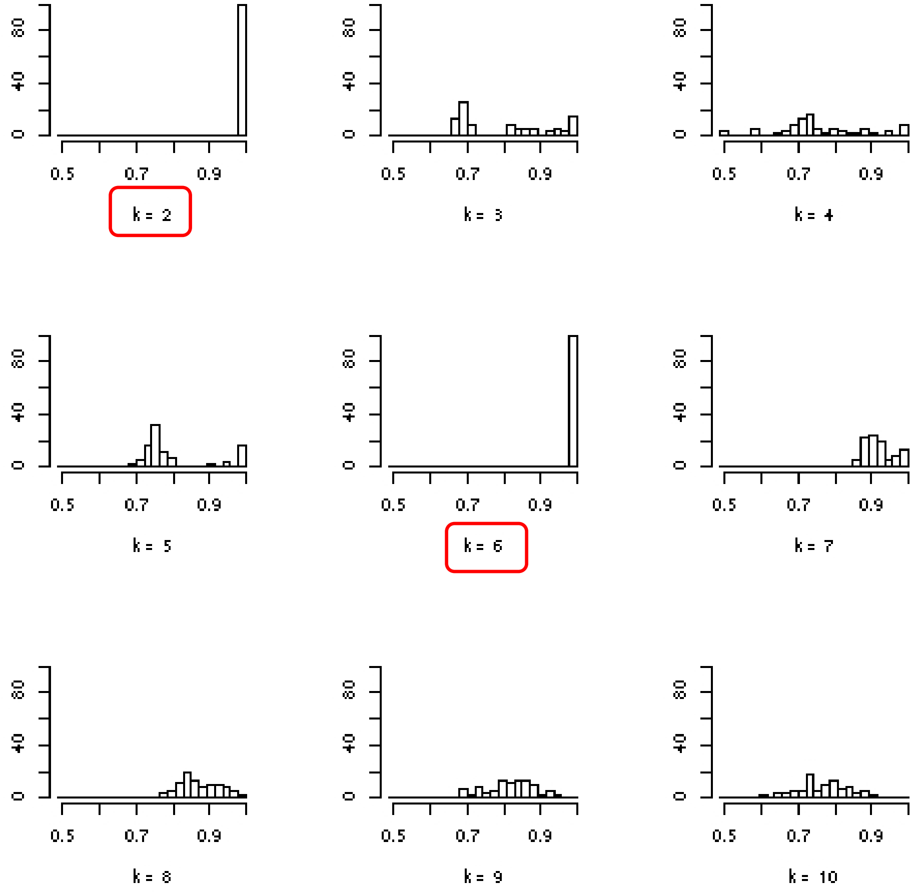
2. H_0 cannot be rejected for a $r < H\text{-}2$. ➡ { *p(r+1), p(r+2),…, p(H)* clusterings are significantly less reliable than *p(1)*.

Note the for case 2 we cannot state that there is no significant difference between the first r top-ranked clusterings (Bernstein inequality is not guaranteed to be tight)

Experiments with high dimensional synthetic data (I)



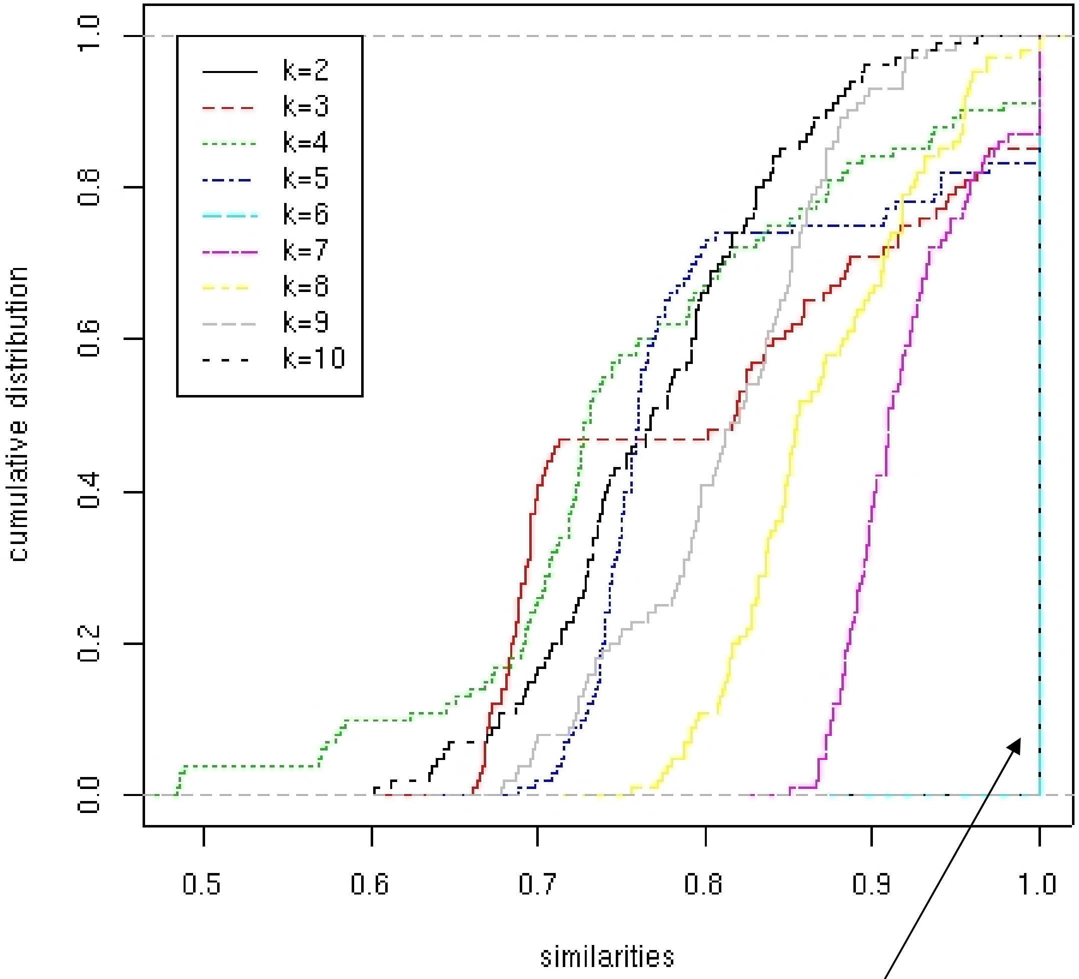
Histograms of the similarity measures obtained by applying PAM clustering to 100 pairs of PMO projections from 1000 to 471-dimensional subspaces (ε=0.2):



- 1000-dimensional synthetic data
- data distributed according to a multivariate gaussian distribution
- 2 or 6 clusters of data (as highlighted by the PCA projection to the two principal components)

Experiments with high dimensional synthetic data (II)

Empirical cumulative distribution of the similarity measures for different k-clusterings



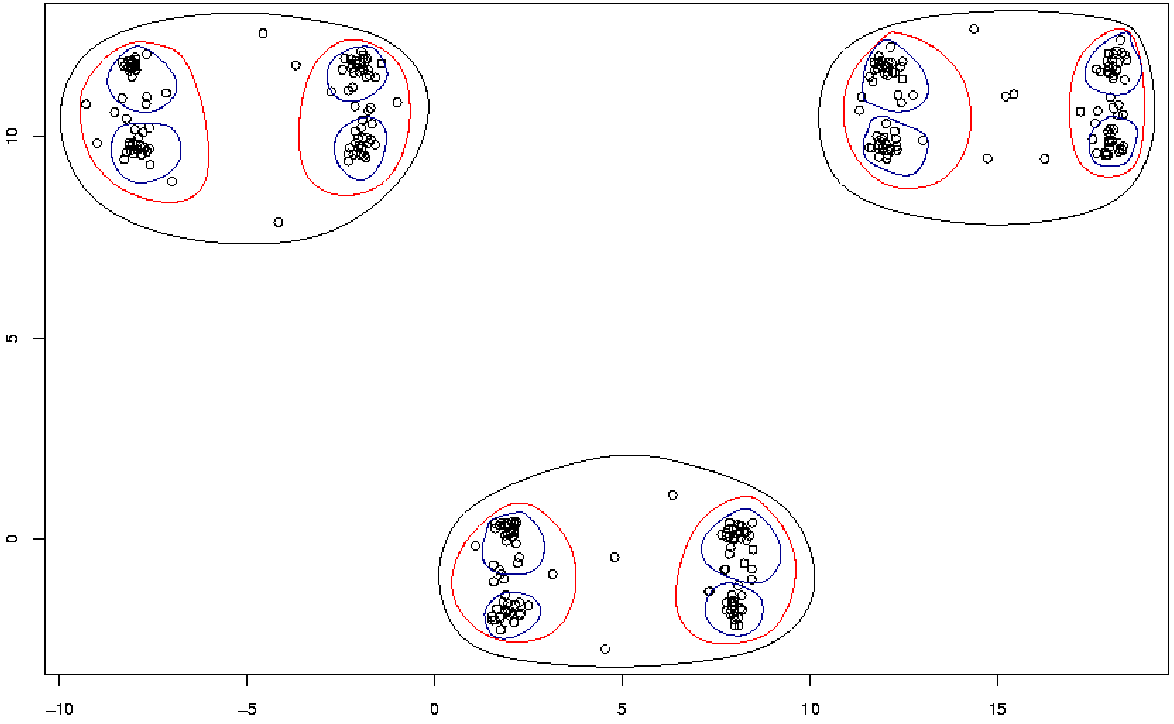
Similarity

Sorting according to the means

k	p-value	mean	variance
2	----	1.0000	0.0000
6	1.0000	1.0000	0.0000
7	3.6e-06	0.9217	0.0016
8	7.8e-10	0.8711	0.0033
9	8.7e-14	0.8132	0.0042
5	1.4e-15	0.8090	0.0104
3	1.1e-16	0.8072	0.0157
10	8.5e-17	0.7715	0.0056
4	5.7e-20	0.7642	0.0158

2 and 6 clusters are selected at 0.01 significance level

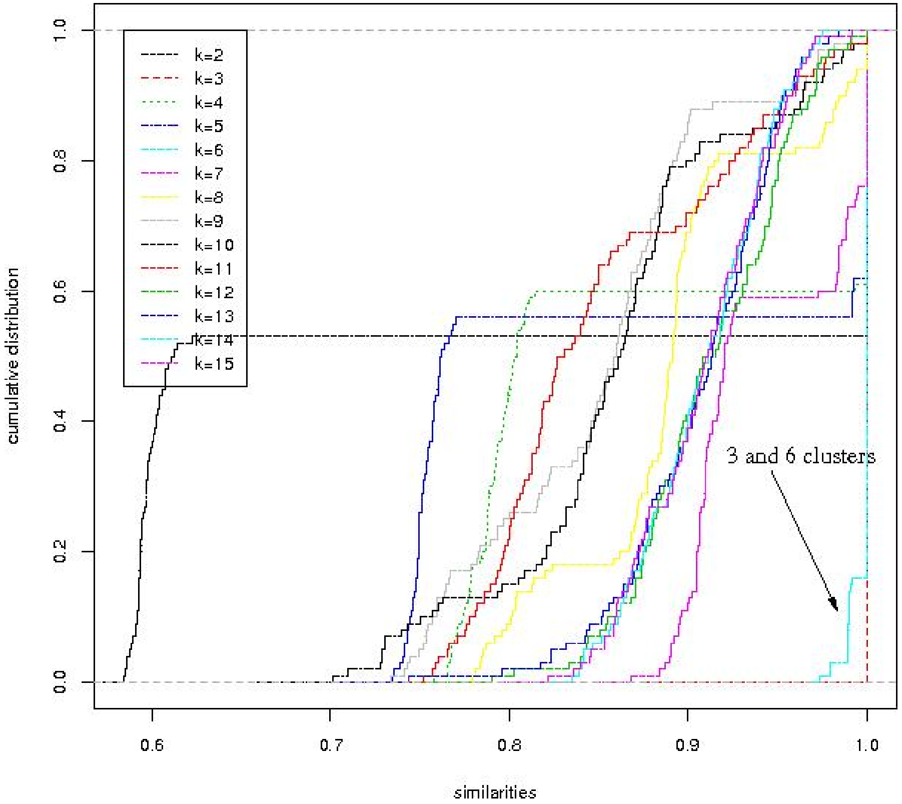
Discovering multi-level structures with stability based methods and statistical tests (1)



- Synthetic data generated using the *clusterv* R package (*Valentini*, 2006)

- A three-level hierarchical structure: 3 (black lines), 6 (red), 12 (blue) clusters

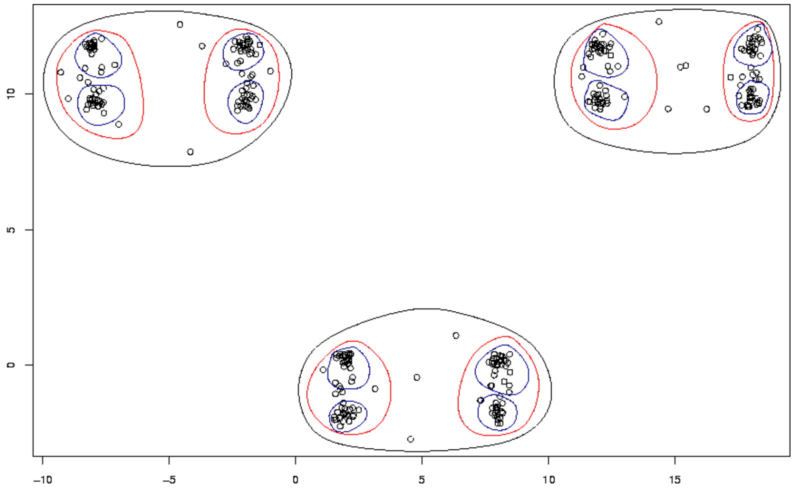
Discovering multi-level structures with stability based methods and statistical tests (2)



Test	Structures discovered (10^{-5} sign.level)
χ^2	3-clustering 6-clustering
Bernstein ind.	3-clustering 6-clustering 7-clustering
Bernstein	3-clustering 6-clustering 7-clustering 12-clustering

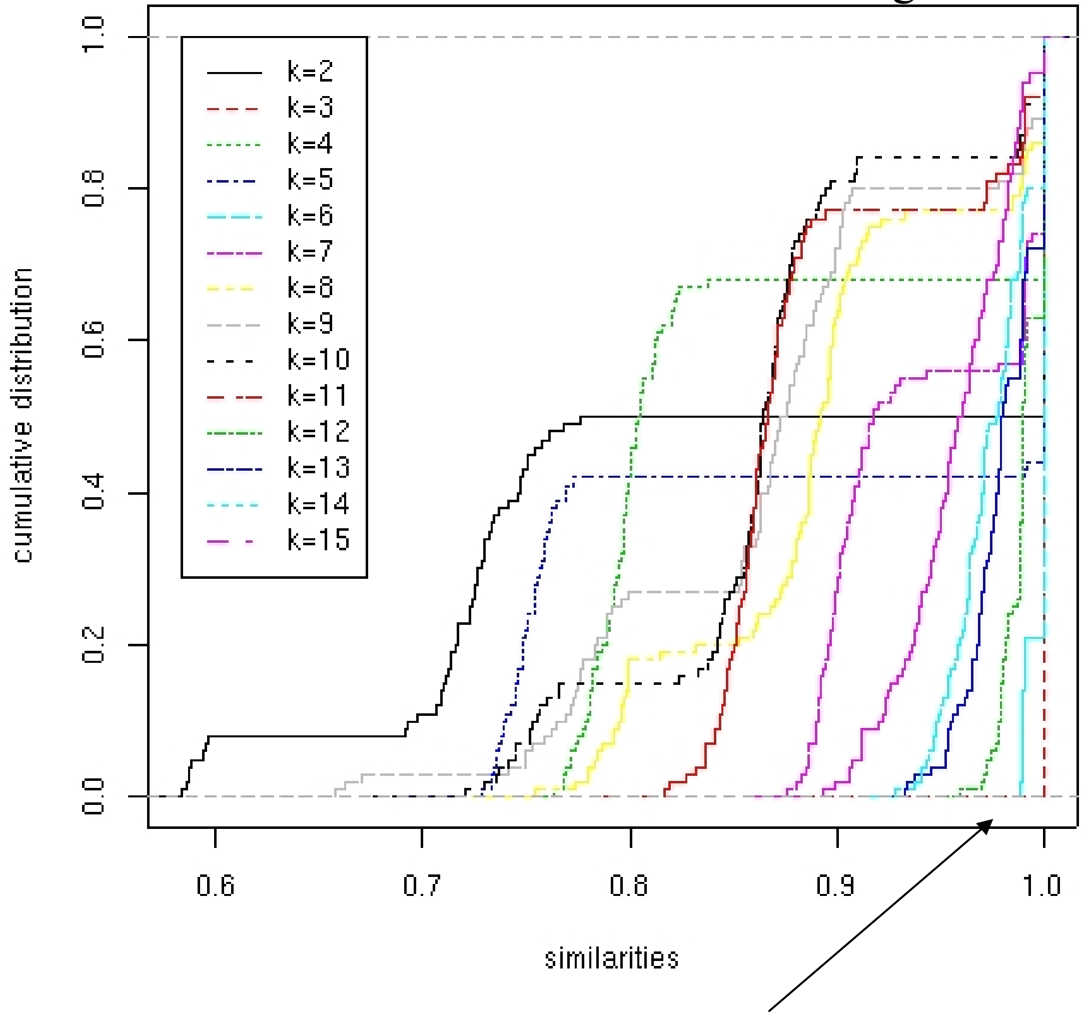
- *Bernstein* test more sensitive to multiple structures
- Drawback: false positives

Discovering … (3): PAM clustering



k	p-value	mean	variance
3	--------	1.0000	0.0000e+00
6	1.0000e+00	0.9979	1.6185e-05
12	1.0000e+00	0.9907	8.0657e-05
13	6.9792e-03	0.9809	2.8658e-04
14	2.2928e-06	0.9754	3.3594e-04
15	0.0000e+00	0.9580	6.8150e-04
7	0.0000e+00	0.9435	2.3055e-03
8	0.0000e+00	0.8954	4.6829e-03
5	0.0000e+00	0.8947	1.5433e-02
11	0.0000e+00	0.8897	3.2340e-03
9	0.0000e+00	0.8706	6.9421e-03
10	0.0000e+00	0.8691	5.0763e-03
4	0.0000e+00	0.8609	9.3463e-03
2	0.0000e+00	0.8532	2.3234e-02

Empirical cumulative distribution of the similarity measures for different k-clusterings

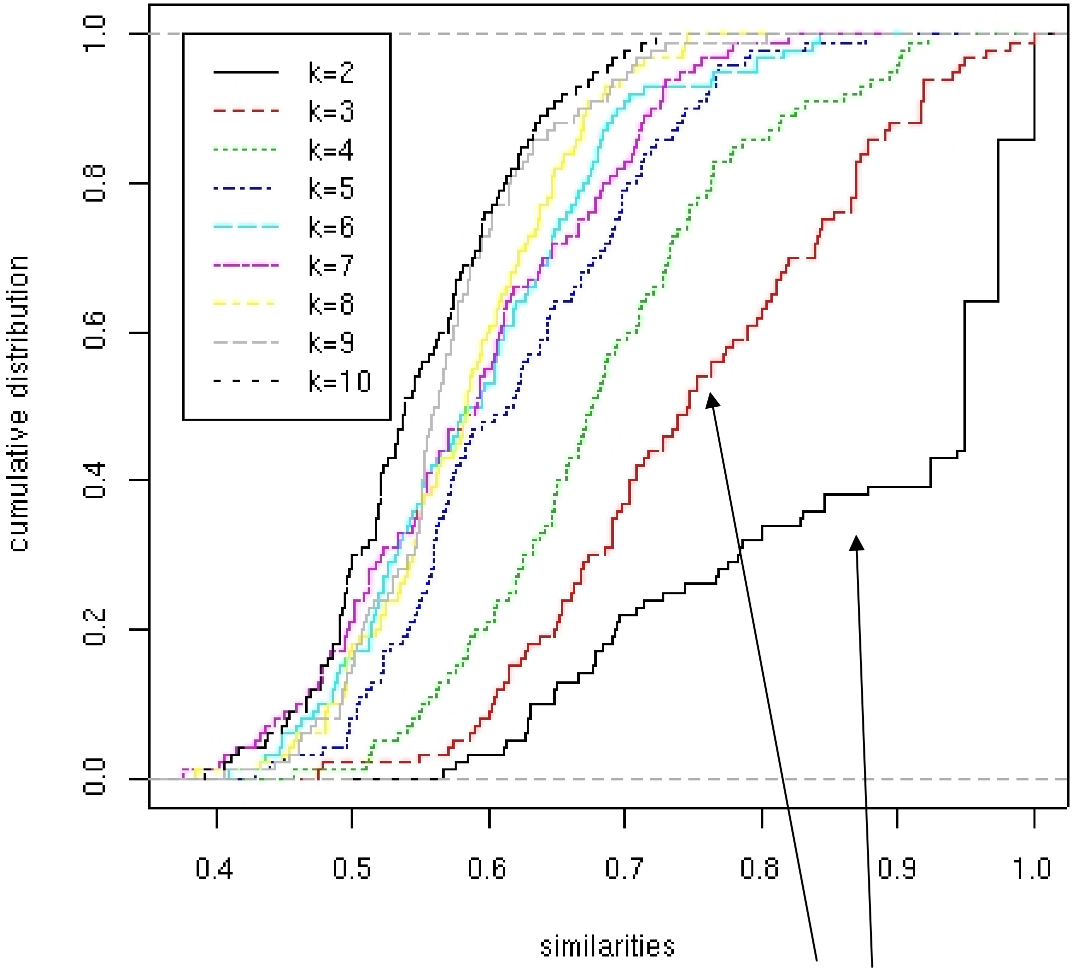


3,6 and 12 clusters are selected at 0.01 significance level

Discovering significant structures in bio-molecular data

(Leukemia data, Golub et al. 1999)

Empirical cumulative distribution of the similarity measures for different k-clusterings

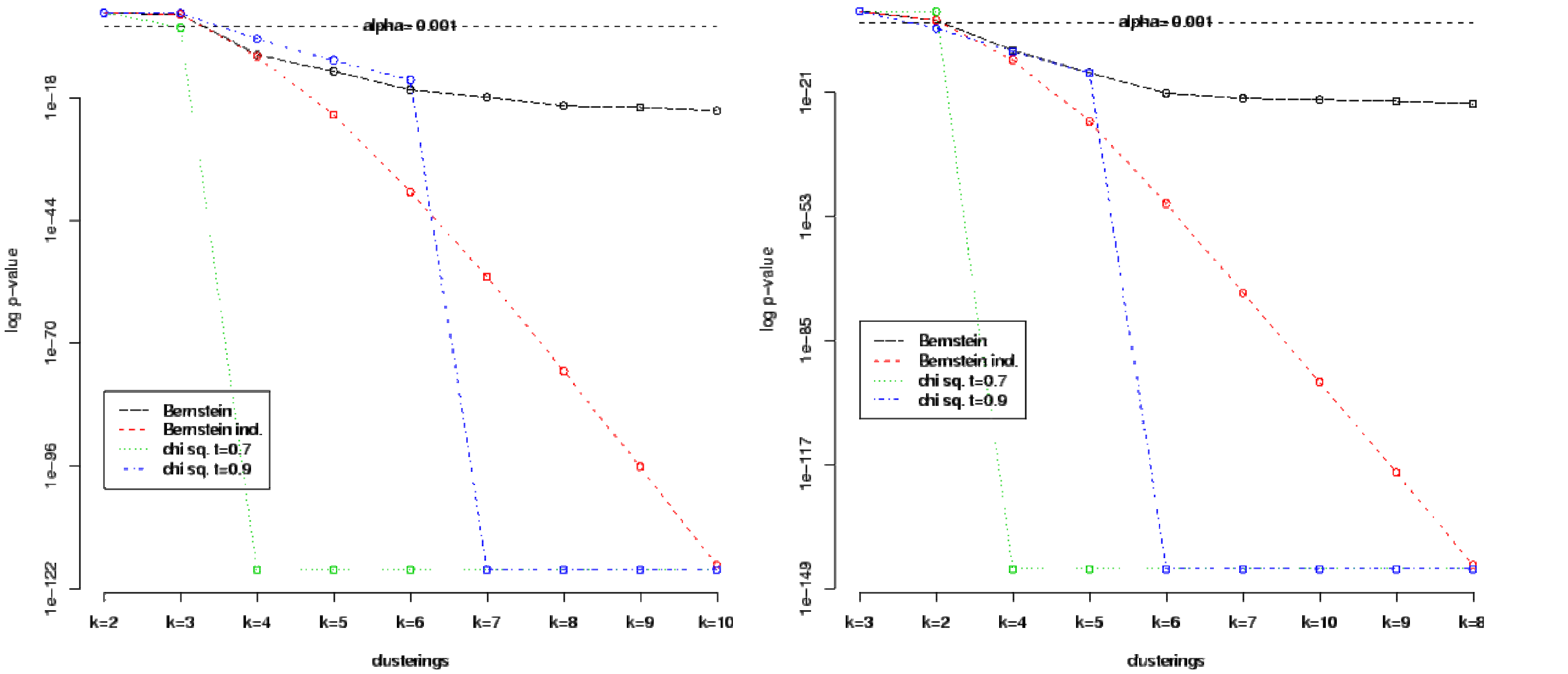


		Similarity
k	p-value	mean
2	------------	0.8664
3	1.056e-04	0.7521
4	1.216e-08	0.6850
5	1.055e-12	0.6196
6	3.932e-14	0.5922
7	1.763e-14	0.5878
8	2.373e-15	0.5822
9	2.757e-16	0.5690
10	1.629e-17	0.5491

2 and 3 clusters are selected at 10^{-5} significance level

- C-mean clustering
- Perturbation: random projections

Comparison of the results between χ^2 and Bernstein test (I)



Leukemia data set :

- 2,3 clusterings detected both by χ^2 and *Bernstein* test
- *Bernstein* p-values decrease more slowly w.r.t. to χ^2 (better sensitivity)
- *Bernstein ind.* is in between *Bernstein* and χ^2

Comparison of the results between χ^2 and Bernstein test (II)

Lymphoma data (*Alizadeh et al.*, 2000): 2,3-clusterings (DLBCL,FL,CLL):

- Hierarchical clustering; perturbation through random susbsampling

Test	**Structures discovered (0.001 sign.level)**
χ^2 (t=0.9)	2-clustering
χ^2 (t=0.7)	2-clustering
Bernstein ind.	2-clustering 3-clustering
Bernstein	2-clustering 3-clustering

Bernstein test more sensitive to multiple structures underlying the data

Comparison with other methods

Methods	**Class. risk** (Lange et al., 2004)	**Gap statistic** (Tibshirani et al. 2001)	**Clest** (Dudoit et.al., 2002)	**Figure of Merit** (Levine& Domany, 2001)	**Model Explorer** (BenHur et al. 2002)	**Random proj. χ^2** (Bertoni Valentini 2006)	**Random proj. Berns.**	**“True” number k**
Data set								
Leukemia (Golub et al., 1999)	k=3	k=10	k=3	k= 2,8,10	k=2	k=2,3	k=2,3	k=2,3
Lymphoma (Alizadeh et al, 2000)	k=2	k=4	k=2	k=2,9	k=2	k=2	k=2,3	k=2,(3)

What can we do with stability based methods and associated statistical tests?

- *Assessment of the reliability* of a given clustering solution;
- *Model order selection*, that is the discovery of the "natural" number of clusters in the data;
- *Estimate of the statistical significance* of a given clustering solution;
- *Discovery of multiple structures* underlying the data, i.e. the detection of multiple reliable clustering solutions at a given significance level.

R software implementing stability based methods

- *Mosclust* (R package)
  Downloadable from:
  http://homes.dsi.unimi.it/~valenti/SW/mosclust

- It requires the R package *clusterv*, downloadable from:
  http://homes.dsi.unimi.it/~valenti/SW/clusterv

Clusterv implements also stability indices to assess the reliability of each individual cluster and the membership of the examples to each cluster.

References

- *A. Bertoni, G. Valentini*, Model order selection for bio-molecular data clustering. *BMC Bioinformatics* 8(Suppl.3), 2007.
- *A. Bertoni, G. Valentini*, Randomized maps for assessing the reliability of patients clusters in DNA microarray data analyses. *Artificial Intelligence in Medicine* 37, 2006.
- *G. Valentini, F.Ruffino*, Characterization of Lung tumor subtypes through gene expression cluster validity assessment, *RAIRO - Theoretical Informatics and Applications*, 40, 2006.
- *G. Valentini*, Mosclust: a software library for discovering significant structures in bio-molecular data. *Bioinformatics* 23(3), 2007.
- *G. Valentini*, Clusterv: a tool for assessing the reliability of clusters discovered in in DNA microarray data. *Bioinformatics* 22(3), 2006.

All the listed papers are downloadable from my home page:

http://homes.dsi.unimi.it/~valenti/pub.html