How many clusters?

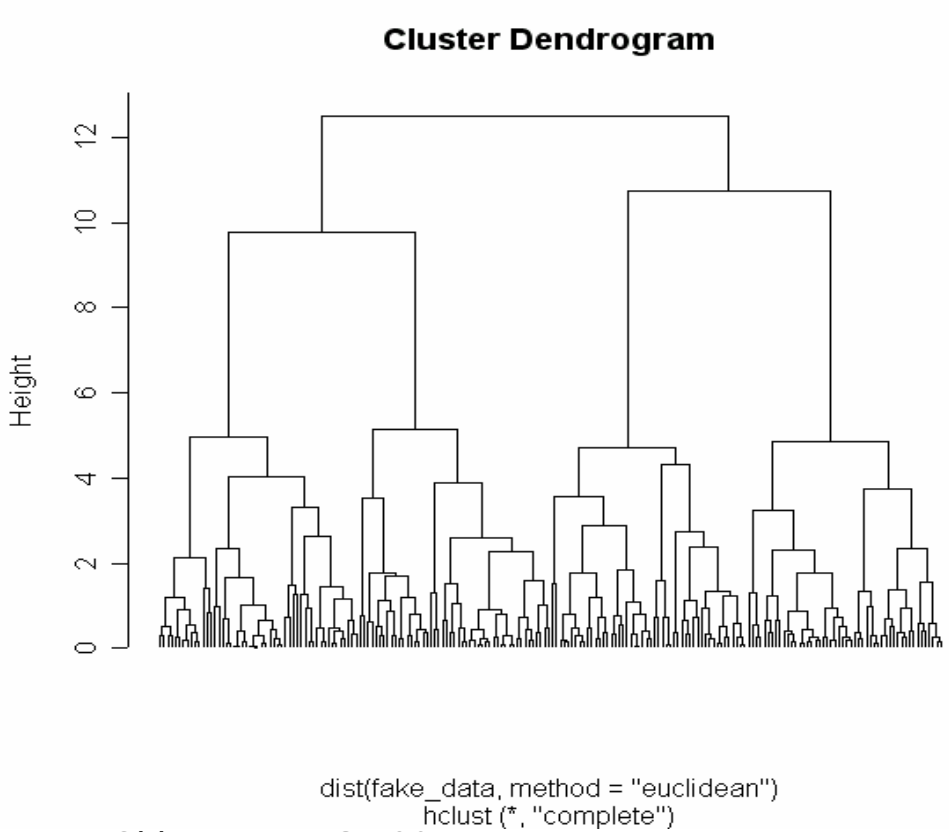
Methods to determine the number of clusters in a data set

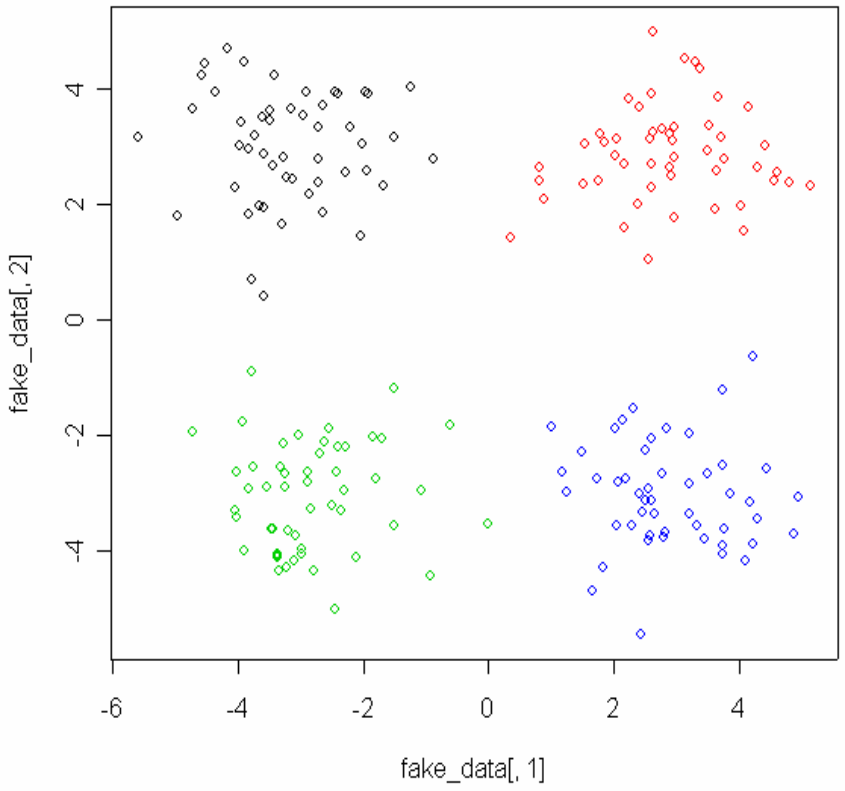
Data set: x_i , i=1…N points in $\boldsymbol{R}^T$ (each coordinate is a feature for the clustering)

Clustering method: hierarchical with given choices of distance (e.g. Euclidean) and link function (e.g. complete); k-means with given choice of distance (e.g. Euclidean); else.

With method and K (# clusters), we obtain a partition of the points: $P(K) = \{C_1 \ldots C_K\}$

For instance, fake 2D data set (n=200, mixture of four $N(\mu_j, I_2)$)





(Hier and Kmeans give identical results)

Define a measure of "quality" of the partition in K clusters:

Using so-called internal indexes, e.g.

a. dissimilarity/distance within the clusters

b. Silhouettes

Or, making *internal use* of a so-called external index, measure

c. Stability of the partition with respect to perturbations by deletion

d. Internal reproducibility (predictability) of the partition

Based on the values of this measure on K =(1),2… use a rule to chose K :

i. The rule can be a simple descriptive criterion

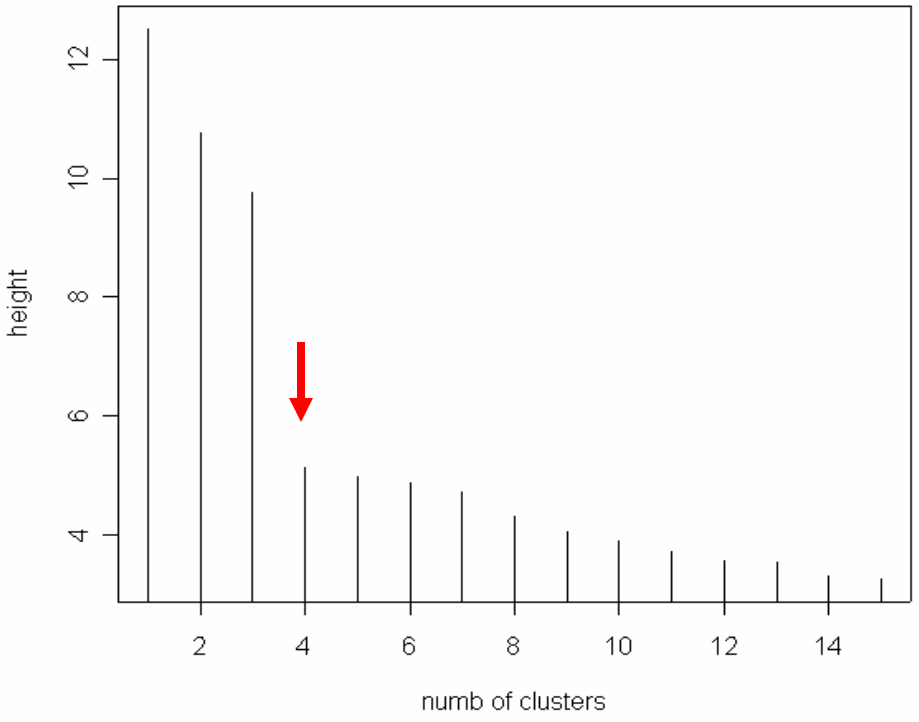
ii. Or it can involve simulating a (null) reference scenario of no-clustering

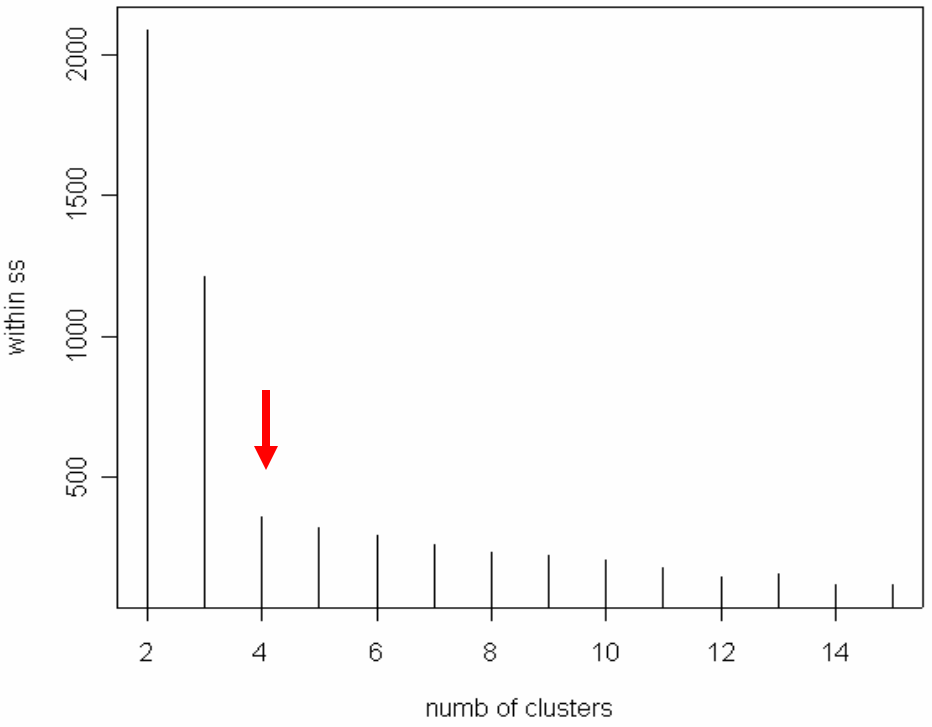
a. Within cluster dissimilarity/distance

Hierarchical: Dissimilarity levels (heights) at which clusters are formed.

K-means: Within clusters sum of squares (what the algorithm finds a local min for).

$$W(K) = \sum_{j=1\ldots K} \sum_{i \in C_j} d^2(x_i; \bar{x}_j)$$

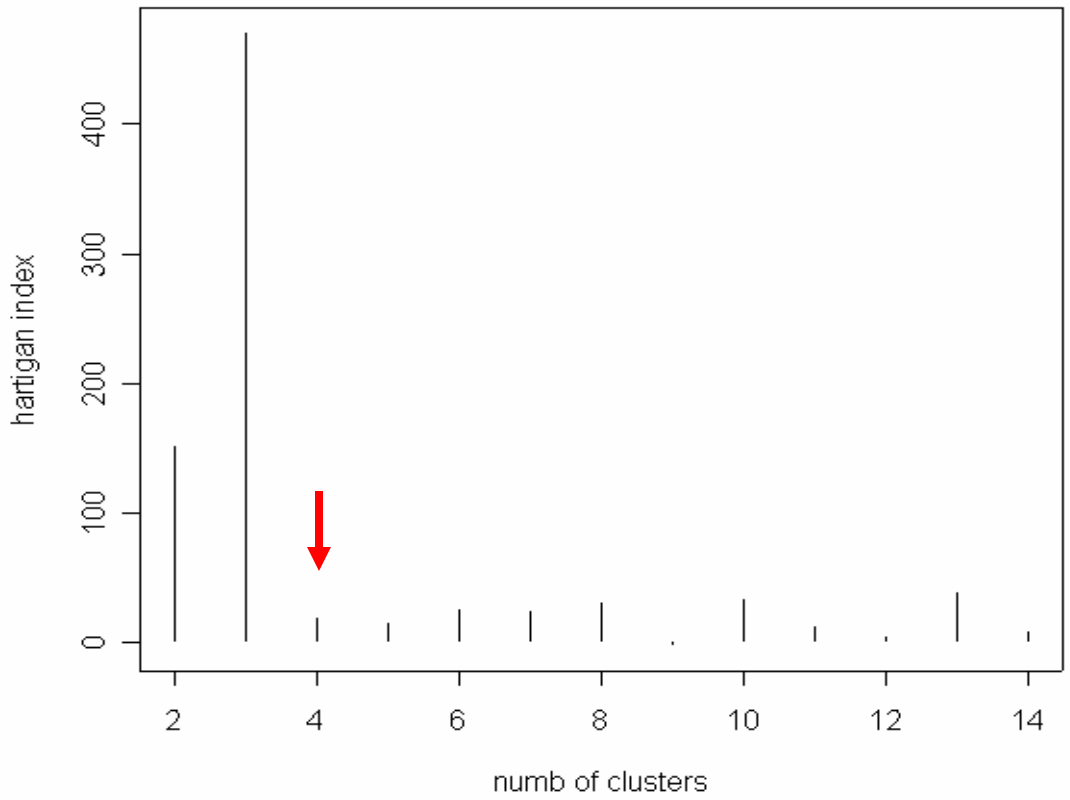




Low values when the partition is good, BUT these are by construction monotone non-increasing (more clusters always makes $\leq$ within cluster dissimilarity); look for “bends”.

$$H(K) = \gamma(K)\frac{W(K) - W(K+1)}{W(K+1)}$$

Hartigan index, correction $\gamma(K) = n - K - 1$



(corrected) relative improvement when passing from K to K +1. High value (right before) followed by low value when the partition is good. NOT monotone.

b. Average Silhouette

$$d_{i,C} = \frac{1}{\#(C)} \sum_{l \in C} d(x_i, x_l)$$

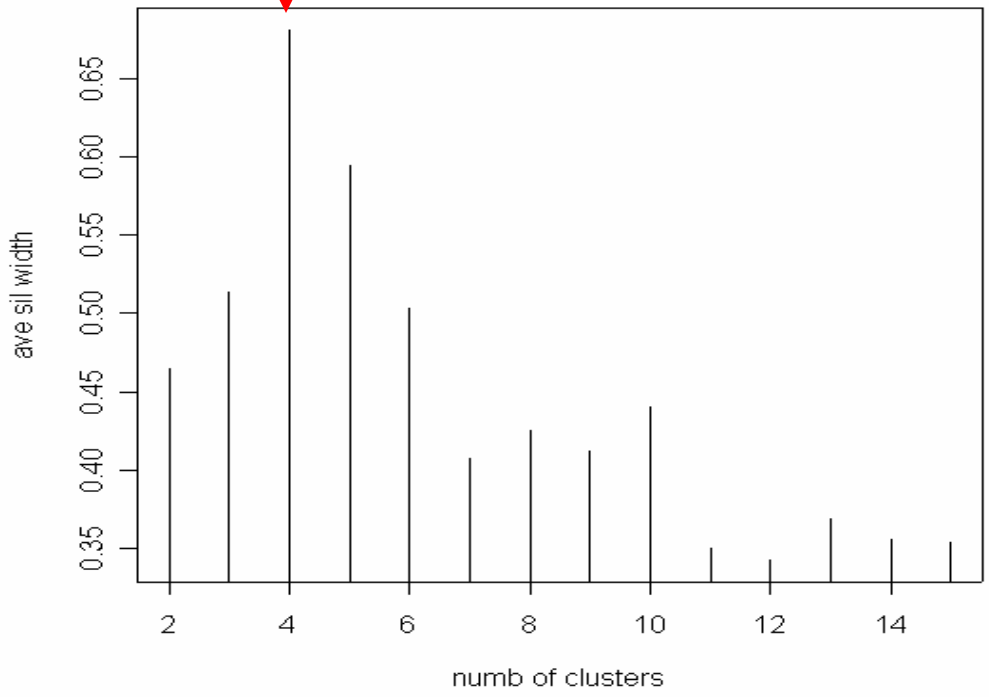
$$a_i = d_{i,C(i)} \qquad b_i = \min_{C \neq C(i)} d_{i,C}$$

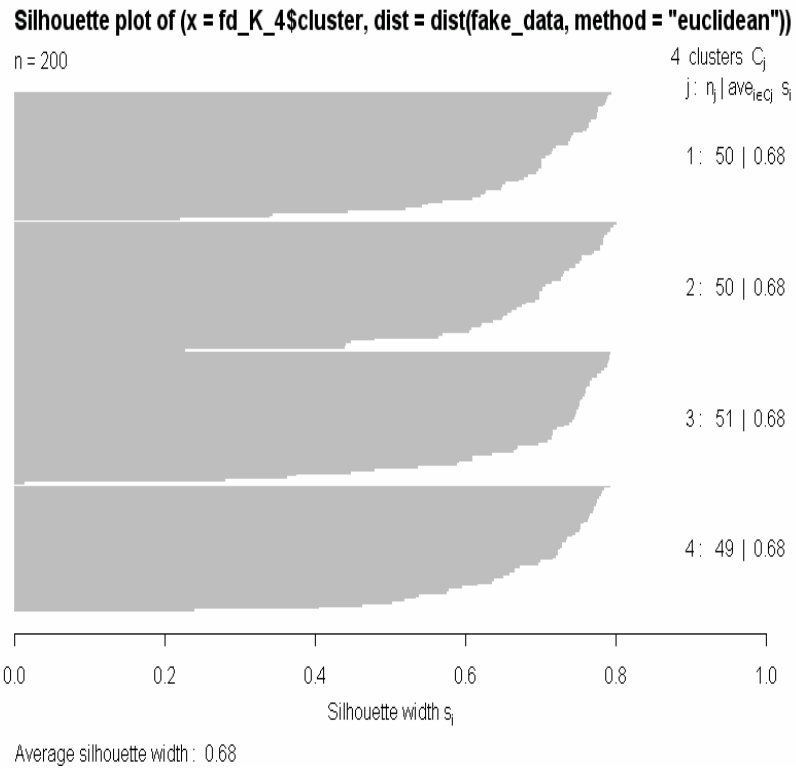
$$Sil_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

How well a point is clustered

$$Sil(K) = \frac{1}{N} \sum_{i=1 \ldots N} sil_i$$

Averaging over points, overall partition quality





High value when the partition is good.
NOT monotone.

External Indexes

Measuring the similarity between two partitions P and Q of the same set of points (but can have different number of clusters), e.g. Rand index

$$Rand = \frac{\#\{(i,l) \text{ together in both P and Q}\}+\#\{(i,l) \text{ NOT together in both P and Q}\}}{\binom{n}{2}}$$

$$R = \frac{Rand - E(Rand)}{Max(Rand) - E(Rand)}$$

Standardizing to a number in [0,1]. Expectation under random partitions. Max depends on the number of clusters in the two partitions.

Can be used to evaluate a $P(K)$ by consistency with a KNOWN partition Q. Useful in several MA data anlyses.

But we can also take another perspective: we adopt an external index (i.e. a measure of similarity between partitions) for internal use… as follows.

c. Stability (to random deletions)

1. For b=1…B
   - form a perturbed data set $\boldsymbol{X}(b)$, deleting f % of the points at random (resample without replacement (1- f)% of the points).
   - apply the clustering to $\boldsymbol{X}(b)$, obtaining $P(K, \boldsymbol{X}(b))$

2. Compute the similarities

$$R(K,b) = R(P(K), P(K, X(b))) \quad b = 1...B$$

To observed partition (restrict to $\boldsymbol{X}(b)$)

or

$$R(K,b,\tilde{b}) = R(P(K, X(b)), P(K, X(\tilde{b}))) \quad b < \tilde{b} = 1...B$$

Among perturbed partitions (restrict to $\boldsymbol{X}(b)$'s intersection)

3. Summarize these similarities, e.g. with their median, to get ***Stb(K)***.

High value when the partition is good. Not monotone in K (can capture nested cluster structure if it exists).

(see Ben-Hur et al., 2002)

d. Internal reproducibility (predictability)

1. For $b = 1 \ldots B$
   - form learn and test data sets $\boldsymbol{L}(b)$, $\boldsymbol{T}(b)$ splitting the points at random
   - apply the clustering to $\boldsymbol{L}(b)$, obtaining $P(K, \boldsymbol{L}(b))$
   - use $P(K, \boldsymbol{L}(b))$ to train a supervised classifier
   - create a *predicted* partition $P^*(K, \boldsymbol{T}(b))$ applying the classifier to $\boldsymbol{T}(b)$
   - apply the clustering to $\boldsymbol{T}(b)$ obtaining $P(K, \boldsymbol{T}(b))$

2. Compute the similarities

$$R(K,b) = R(P^*(K,T(b)), P(K,T(b))) \quad b = 1 \ldots B$$

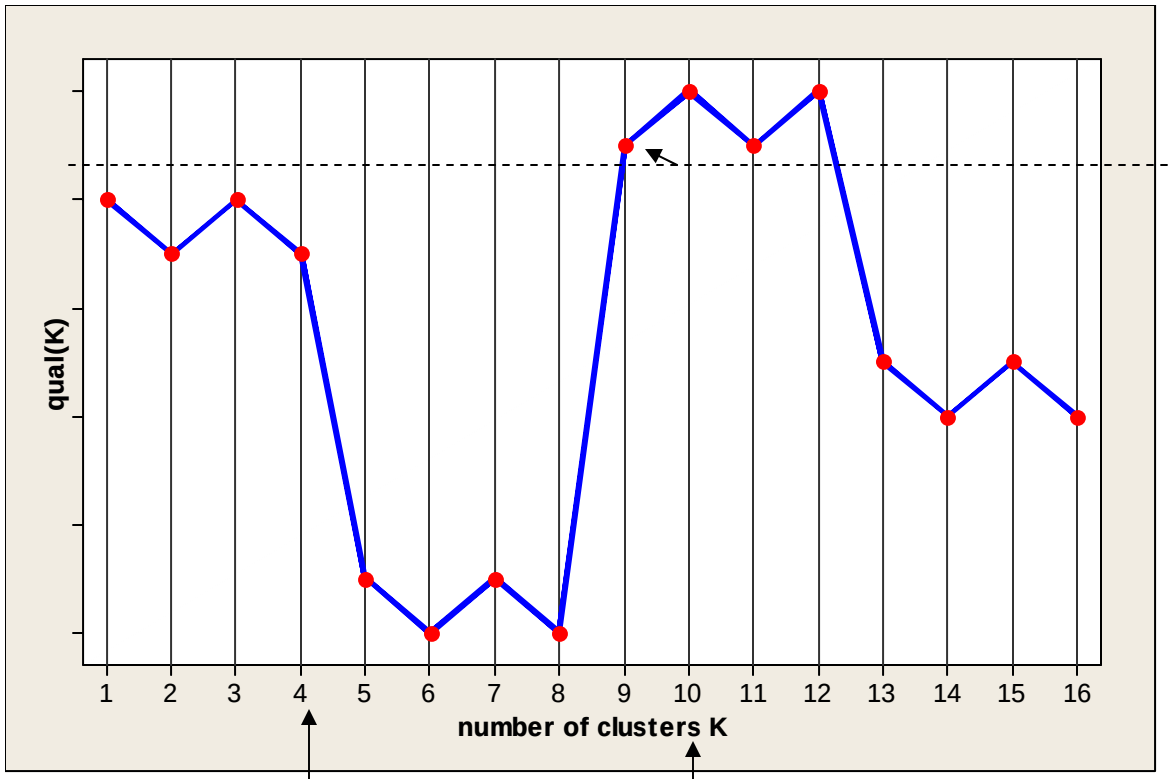
Among predicted and actual partition of $\boldsymbol{T}(b)$

3. Summarize these similarities, e.g. with their median, to get ***Prd(K)***.

High value when the partition is good. Not monotone in K (can capture nested cluster structure if it exists).

(see Dudoit and Fridlyand, 2002)

i. Choosing K based on simple descriptive criteria.



Smallest K within ***t*** of the (smallest) maximal K

Smallest maximal K

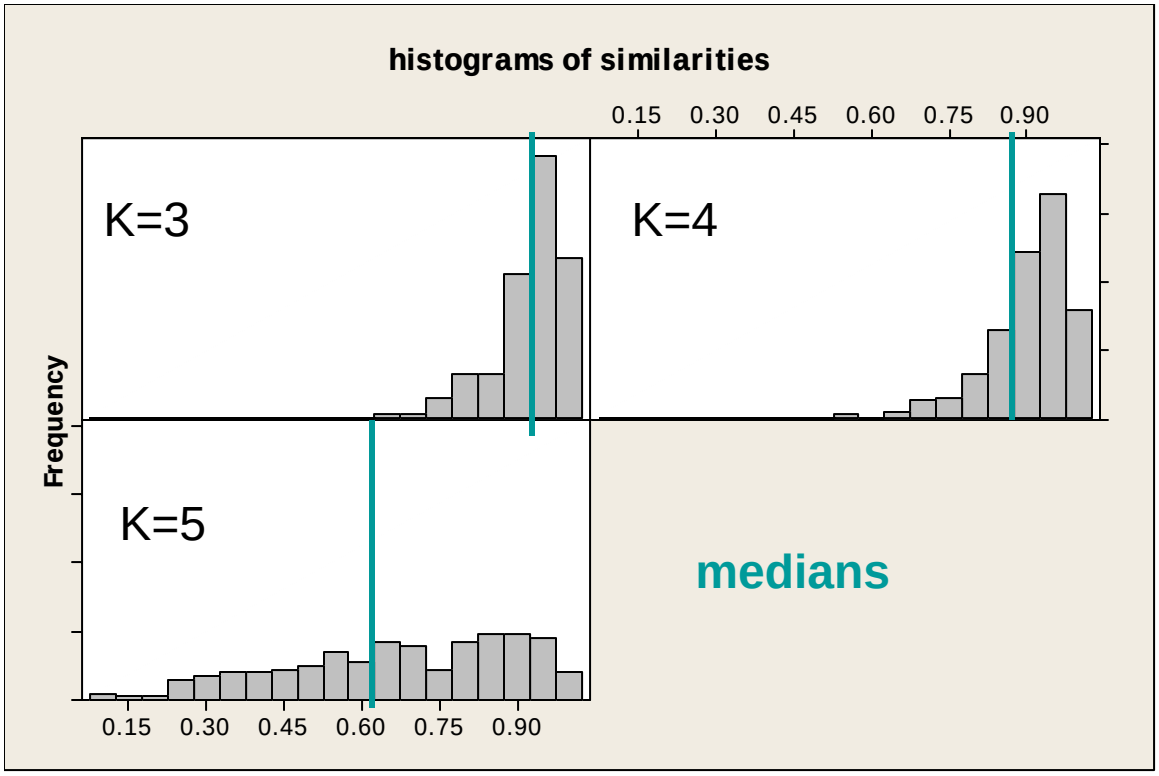
Smallest K after which there is a drop $\geq$ ***t***

For instance:

Silhouette approach $\hat{K}: \max_K Sil(K)$

Hartigan approach $\hat{K}:$ smallest such that $H(K) \leq \eta$ (e.g.10)

Stability approach (Ben-Hur et al.) $\hat{K}:$ smallest such that $Stb(K+1) \leq \sigma$



i. Simulating a no-clustering reference scenario

Chose a null distribution on $\boldsymbol{R}^T$ expressing no-clustering, and

1. For $m = 1 \ldots M$

   - draw a data set $\boldsymbol{X}_o(m)$ of size n from the null distribution

   - For K = (1),2… apply the clustering to $\boldsymbol{X}_o(m)$ obtaining $P(K, \boldsymbol{X}_o(m))$

2. Compute the quality statistics

$$qual(K,m) = qual(P(K, X_o(m)) \quad m = 1 \ldots M,\ K = (1),2,\ldots$$

   (reproducing the calculations previously described on the actual data set $\boldsymbol{X}$)

3. For each K = (1), 2… create summaries

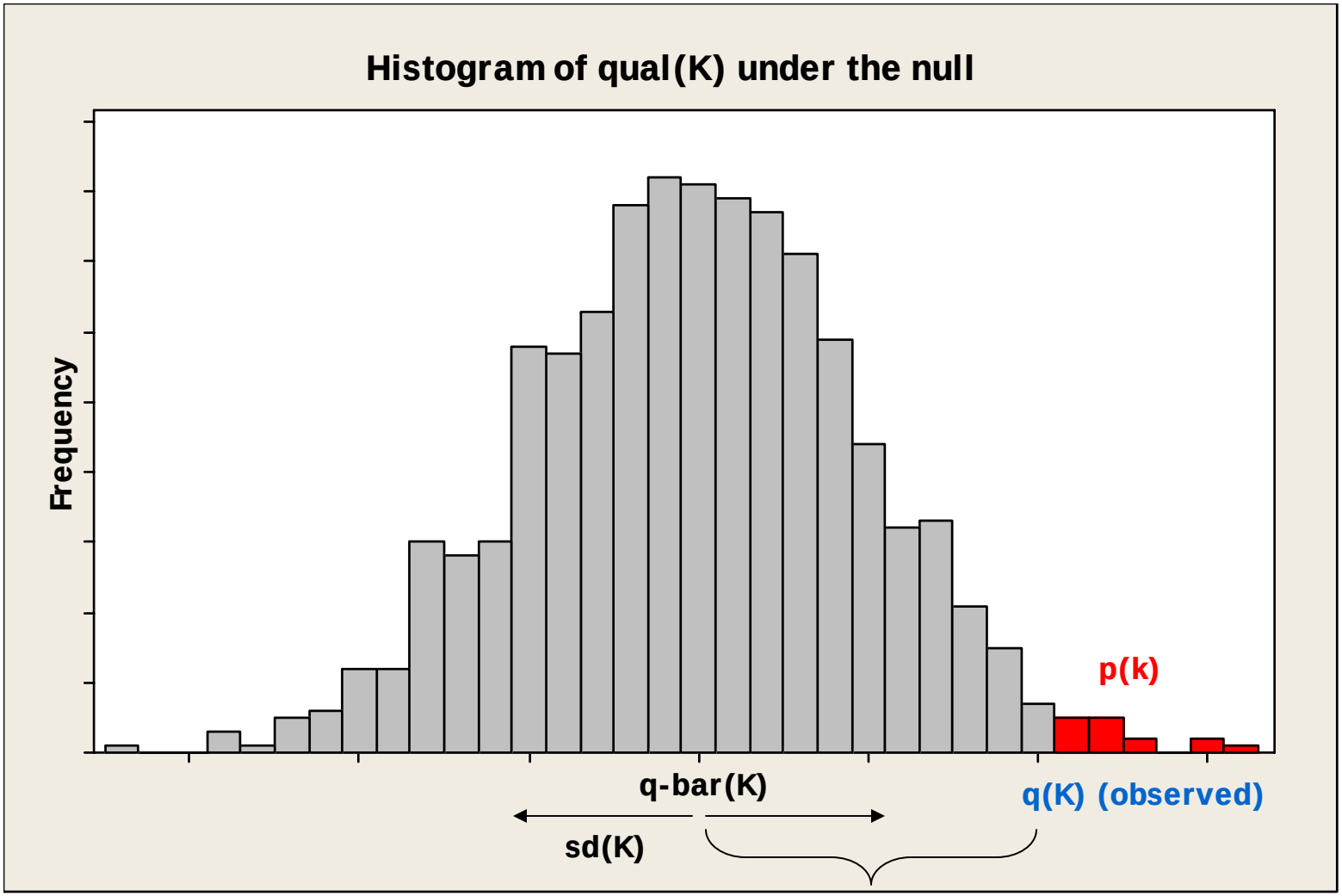
$$\bar{q}(K) = \frac{1}{B} \sum_{b=1 \ldots B} qual\,(K,b)$$

$$sd\,(K) = \sqrt{\frac{1}{B-1} \sum_{b=1 \ldots B} (qual\,(K,b) - \bar{q}(K))^2}$$

$$p(K) = \frac{1}{B} \#\{b : qual\,(K,b) \geq qual\,(K)\}$$

Estimated expected value and variability of the statistic under the null.

Empirical p-value corresponding to the statistic observed on the actual data



Difference or Gap

Now can formulate decision rules for K based on these summaries. For instance

Gap approach (Tibshirani et al., 2001)

$$qual(K) = \log(W(K))$$

$$gap(K) = qual(K) - \bar{q}(K)$$

$$\tilde{sd}(K) = \gamma sd(K) \quad \text{correction } \gamma = \sqrt{1 + \frac{1}{M}}$$

$$K^* : \max_K gap(K)$$

$$\hat{K} : \text{ smallest such that } gap(K) \geq gap(K^*) - \tilde{sd}(K^*)$$

CLEST approach (Dudoit and Fridlyand, 2002)

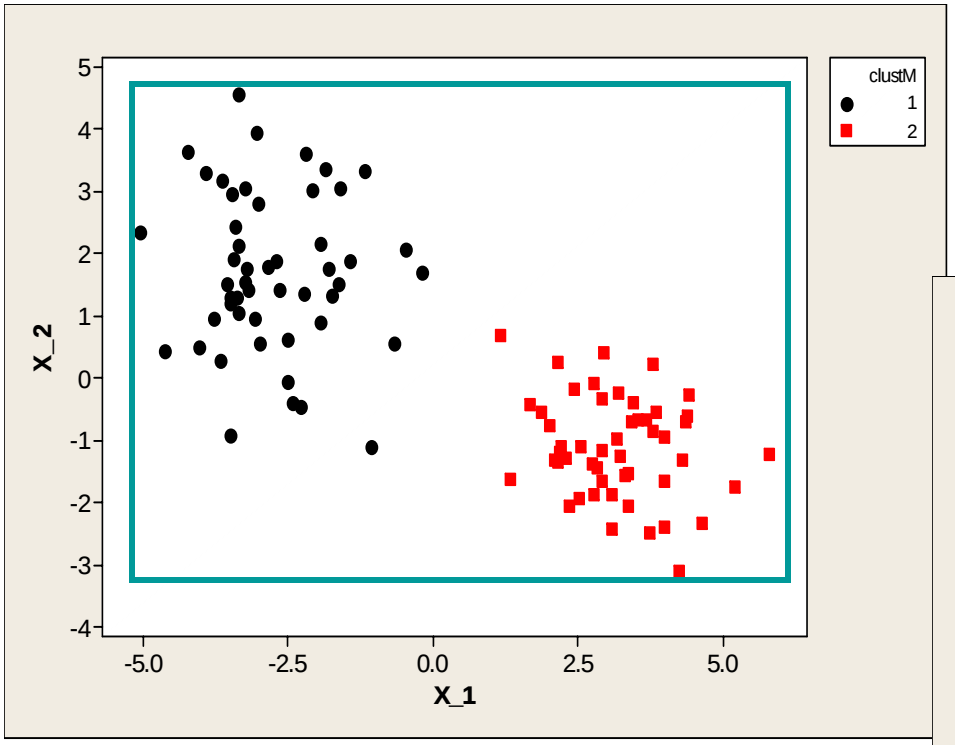
$$qual(K) = \text{Prd}(K)$$

$$d(K) = qual(K) - \bar{q}(K)$$

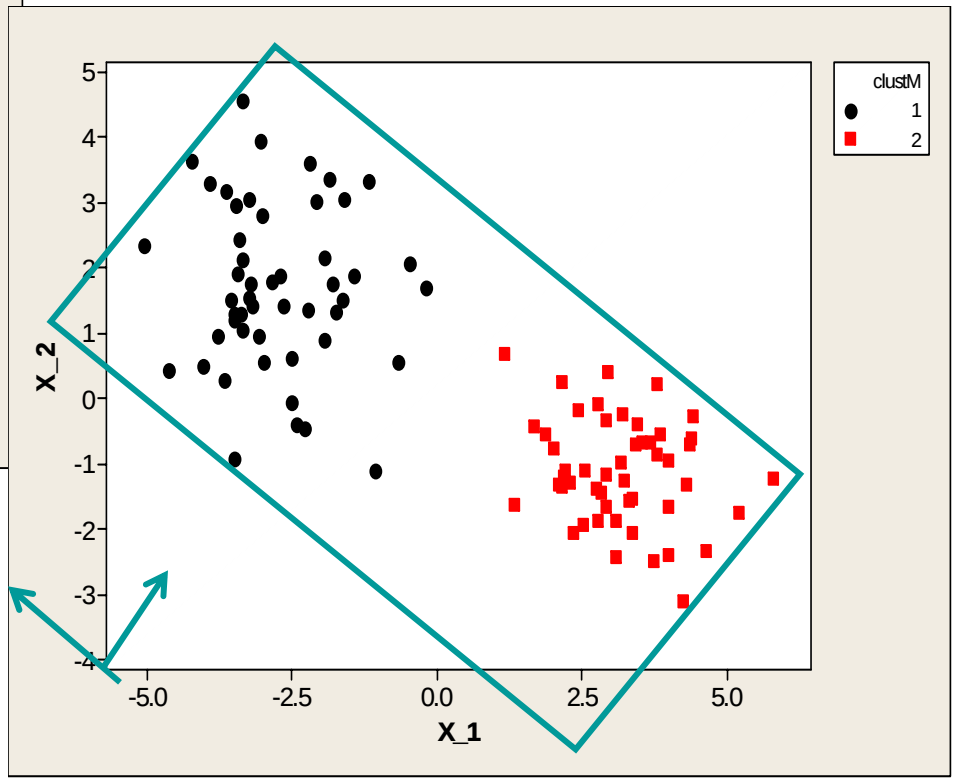
$$\hat{K} : \text{ among those such that } p(K) \leq \pi, \max_K d(K)$$

Important: how does one select the reference distribution?

Most often used no-clustering scenarios, **UNIFORMS**.



On the data (hyper) box, original coordinates



On the data (hyper) box, PCA coordinates – more effective, smaller volume

Useful references:

Ben-Hur A, Elisseeff A, Guyon I (2002) A stability based method for discovering structure in clustered data. *Proceedings of PSB 2002*.

Tibshirani R, Walther G, Hastie. T (2001): Estimating the Number of Clusters in a Dataset via the Gap Statistic*. JRSS-B.*

Dudoit S, Fridlyand J (2002) A prediction based resampling method for estimating the number of clusters in a data set. *Genome Biology*.