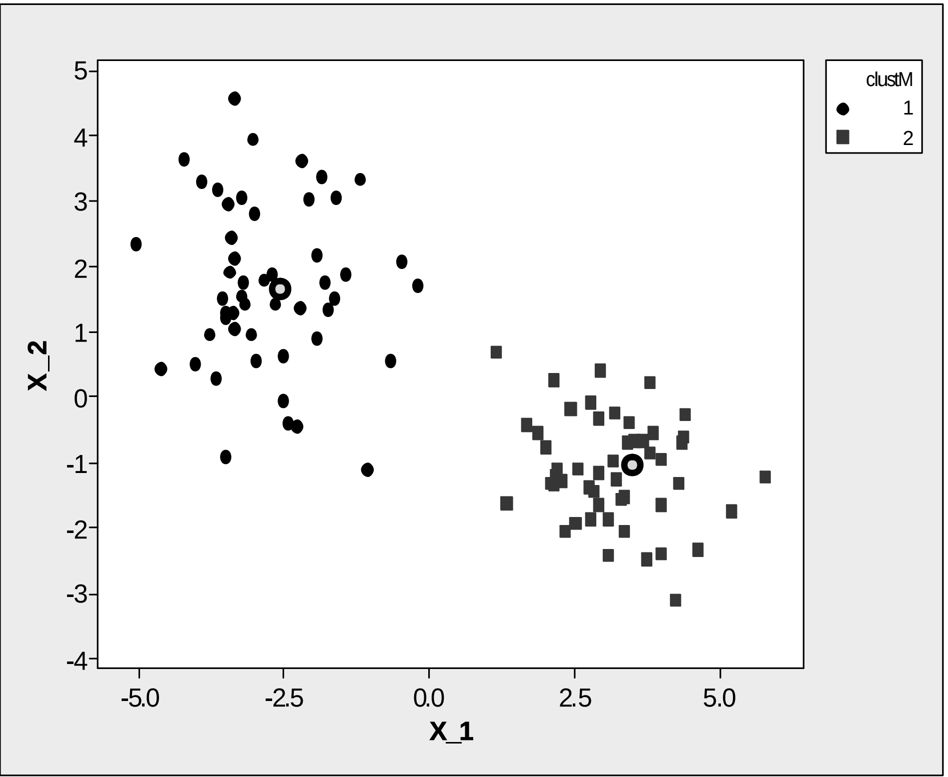
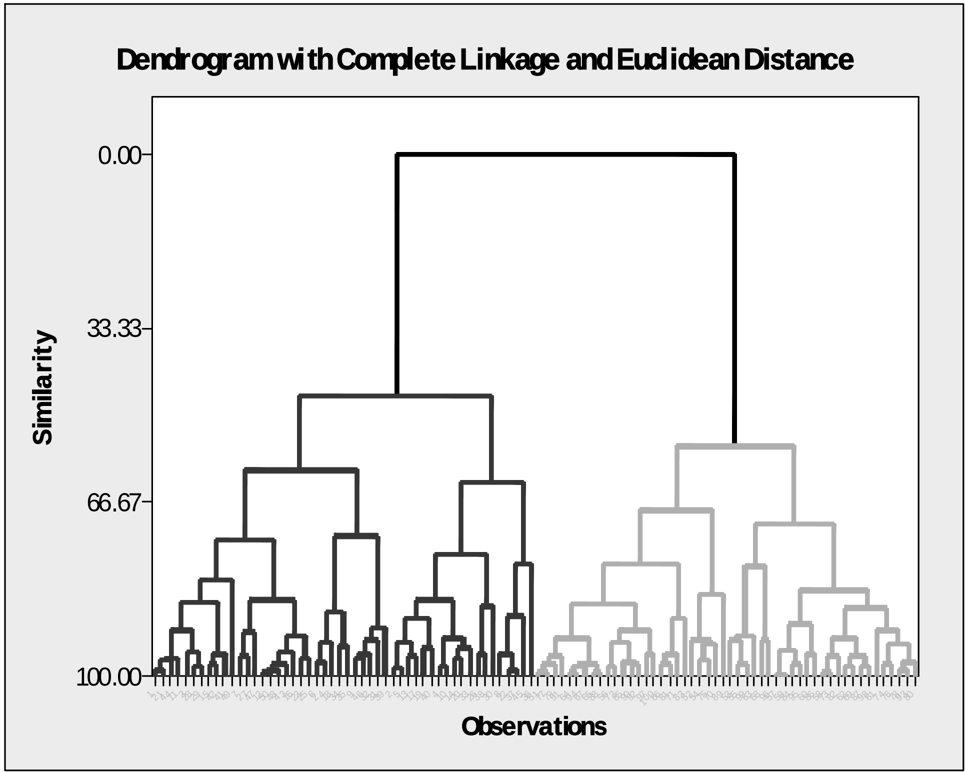
Methods to determine the number of clusters in a data set

Data set: x_i , i=1…N points in R^p
(each coordinate is a feature for the clustering)

Clustering method: e.g. hierarchical with given choices of metric and link function, or k-means with given choice of metric

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





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

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

Squared distances from centroids (within clusters sum of squares). This is what k-means finds a local min for.

$$W(K) = \sum_{j=1\ldots K} d(j)$$

Dissimilarity levels at which clusters are formed.

Low values when the partition is good, and thus K appropriate. BUT this is by construction monotone non-increasing in K (more clusters always means smaller within cluster dissimilarity). Can consider:

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

$$\text{Hartigan index, correction } g(K) = n - k - 1$$

Relative improvement when passing from K to K+1 (with correction). 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 data point is clustered

$$Sil(K) = \frac{1}{N} \sum_{i=1 \ldots N} sil_i$$ Averaging over points, overall quality of the partition

High values when the partition is good, and thus K appropriate. This is not monotone in K.

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]. E 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.

Here we use 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 m=1…M
   - form a perturbed data set X(m), deleting f% of the points at random (resample without replacement (1-f)% of the points).
   - apply the clustering to X(m) obtaining P(K,X(m))

2. Compute the similarities

$$R(K,m) = R(P(K), P(K, X(m))) \quad m = 1...M$$

To observed partition (restrict to X(m))

or

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

Among perturbed partitions (restrict to X(m)'s intersection)

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

High values when the partition is good, and thus K appropriate. This, too, is not monotone in K.

d. Internal reproducibility (predictability)

1. For m=1…M
   - form learn and test data sets L(m), T(m) splitting the points at random
   - apply the clustering to L(m) obtaining P(K,L(m))
   - use P(K,L(m)) to train a supervised classifier
   - create a *predicted* partition P*(K,T(m)) applying the classifier to T(m)
   - apply the clustering to T(m) obtaining P(K,T(m)

2. Compute the similarities

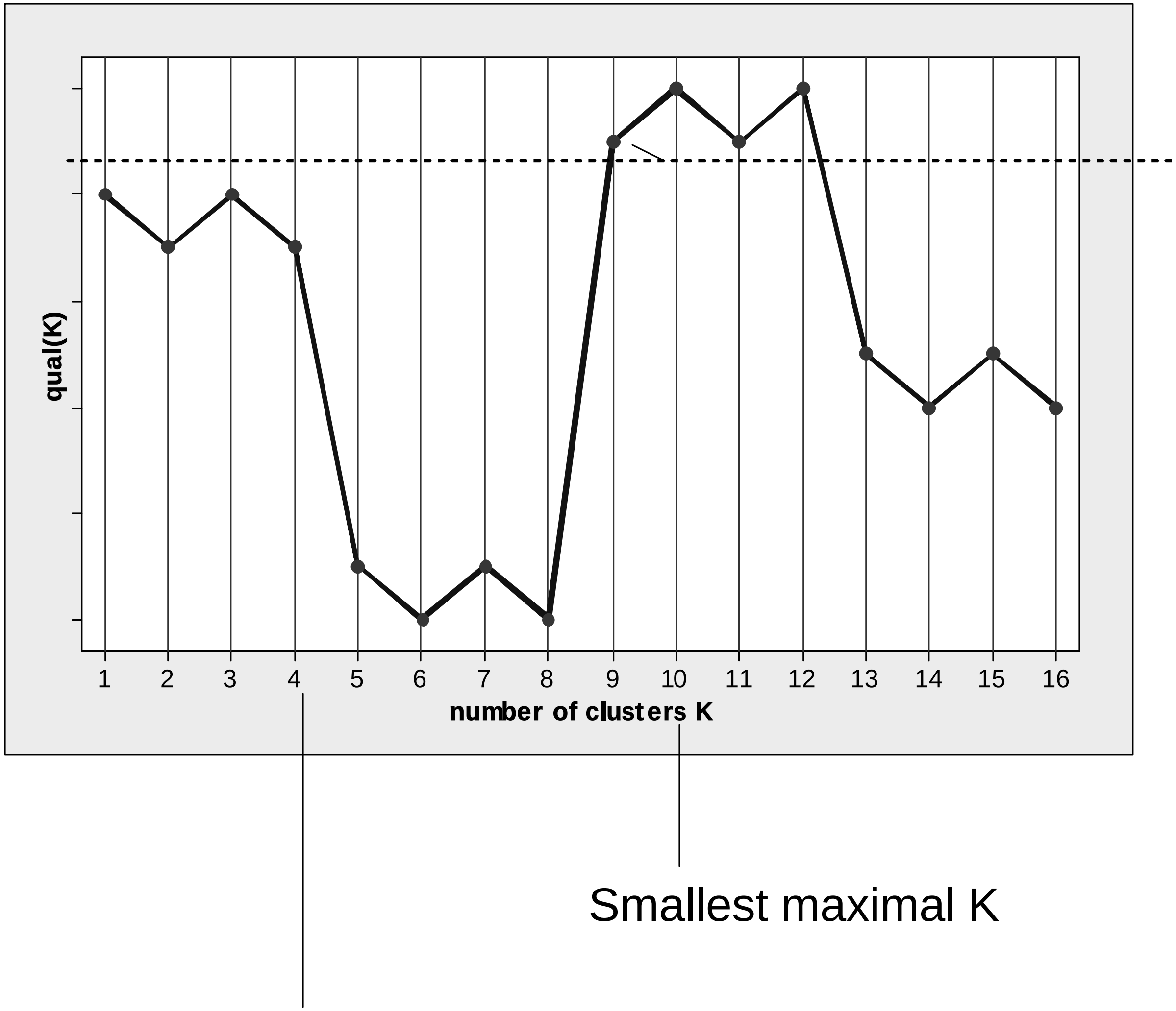
$$R(K,m) = R(P^*(K,T(m)), P(K,T(m))) \quad m = 1...M$$

Among predicted and actual partition of T(m)

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

High values when the partition is good, and thus K appropriate. This, too, is not monotone in K.

i. Choosing K based on simple descriptive criteria.



Smallest K within ***t*** of the maximal K

Smallest maximal K

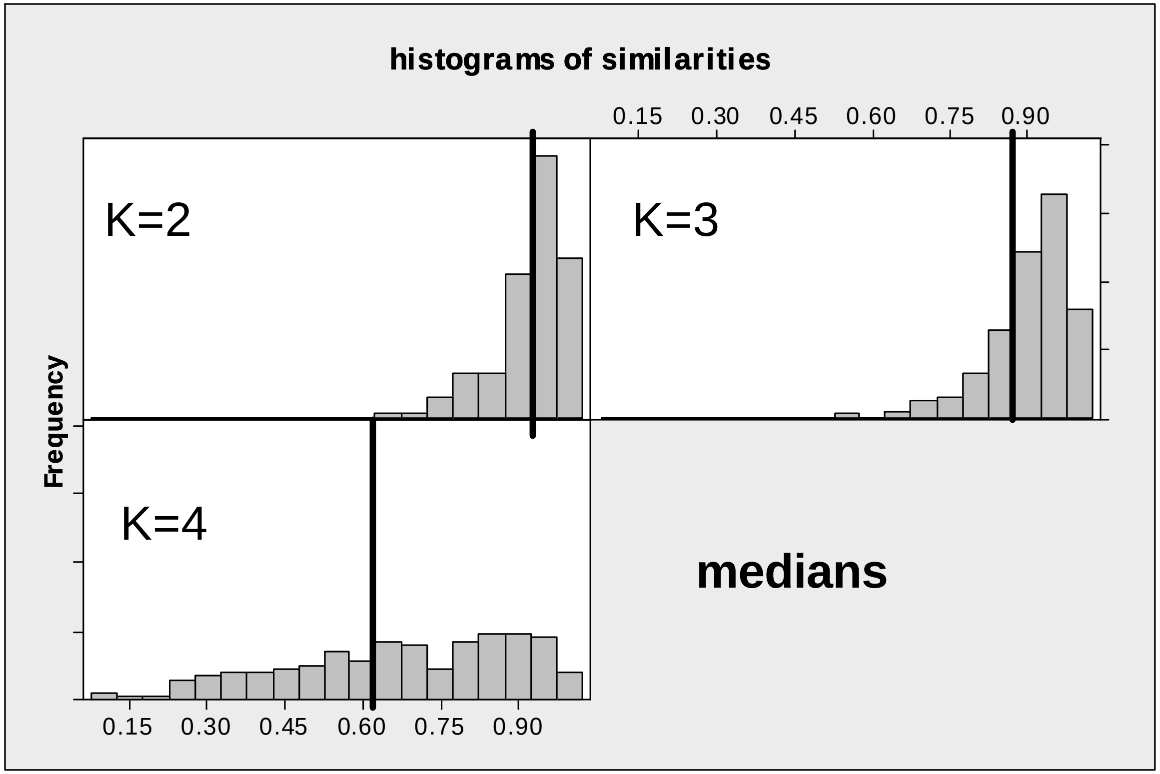
Smallest K after which there is a drop $\geq$ ***t***

For instance:

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

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

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



i. Simulating a no-clustering reference scenario

Chose a null distribution on R^p expressing no-clustering, and

1. For b=1…B

   - draw a data set $X_o(b)$ of size n from the null distribution

   - For K = (1),2… apply the clustering to $X_o(b)$ obtaining $P(K,X_o(b))$

2. Compute the quality statistics

$$qual(K,b) = qual(P(K, X_o(b)) \quad b = 1...B,\ K = (1),2,...$$

   (reproducing the calculations previously described on the actual data set X)

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

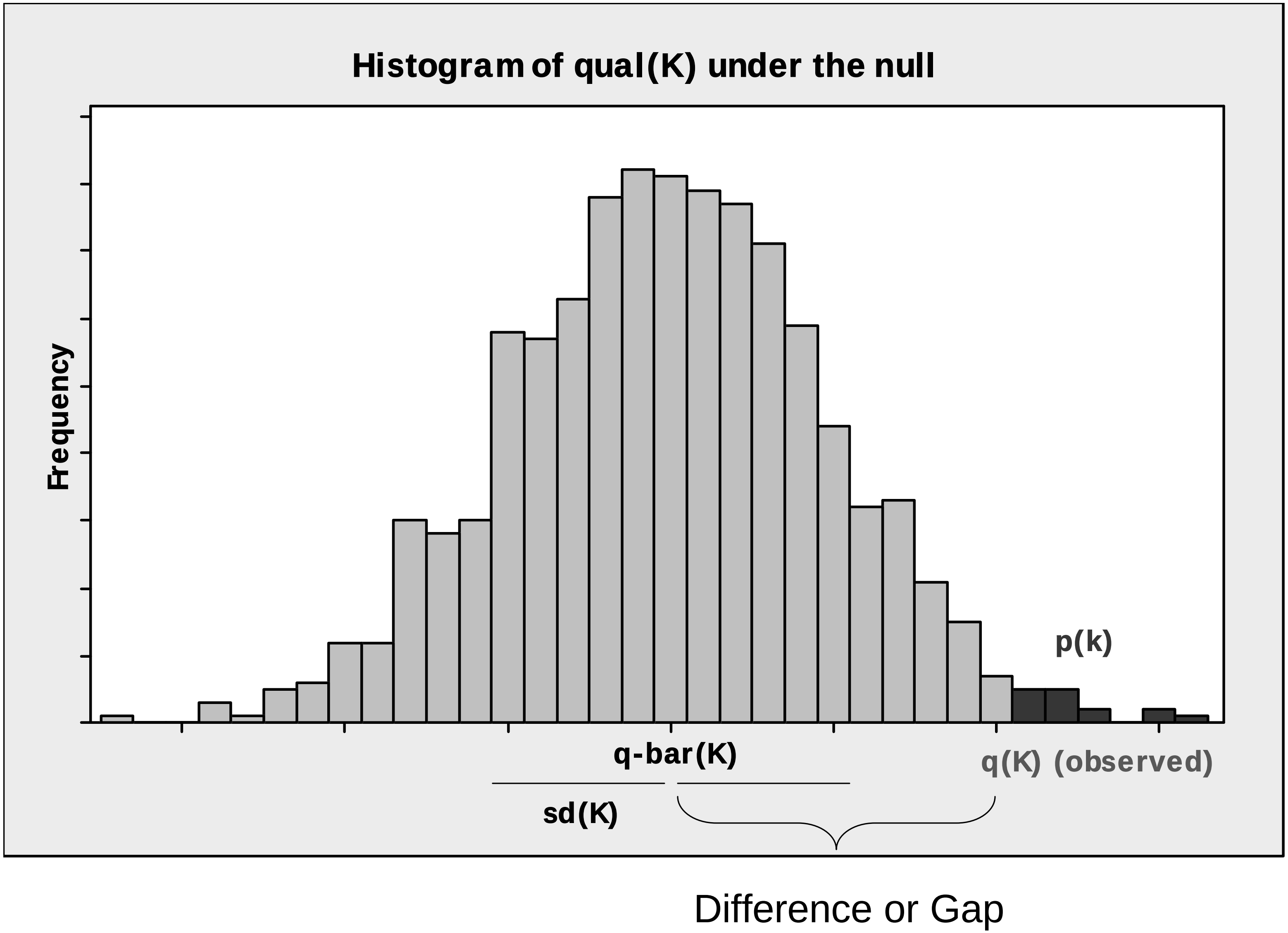
$$\overline{q}(K) = \frac{1}{B}\sum_{b=1...B} qual\,(K,b)$$

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

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

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

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


Histogram of qual(K) under the null
Frequency
p(k)
q-bar(K)
q(K) (observed)
sd(K)
Difference or Gap

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

<u>Gap</u> approach (Tibshirani et al.)

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

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

$$\widetilde{sd}(K) = gsd(K) \quad \text{correction } g = \sqrt{1 + \frac{1}{B}}$$

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

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

<u>CLEST</u> approach (Dudoit et al.)

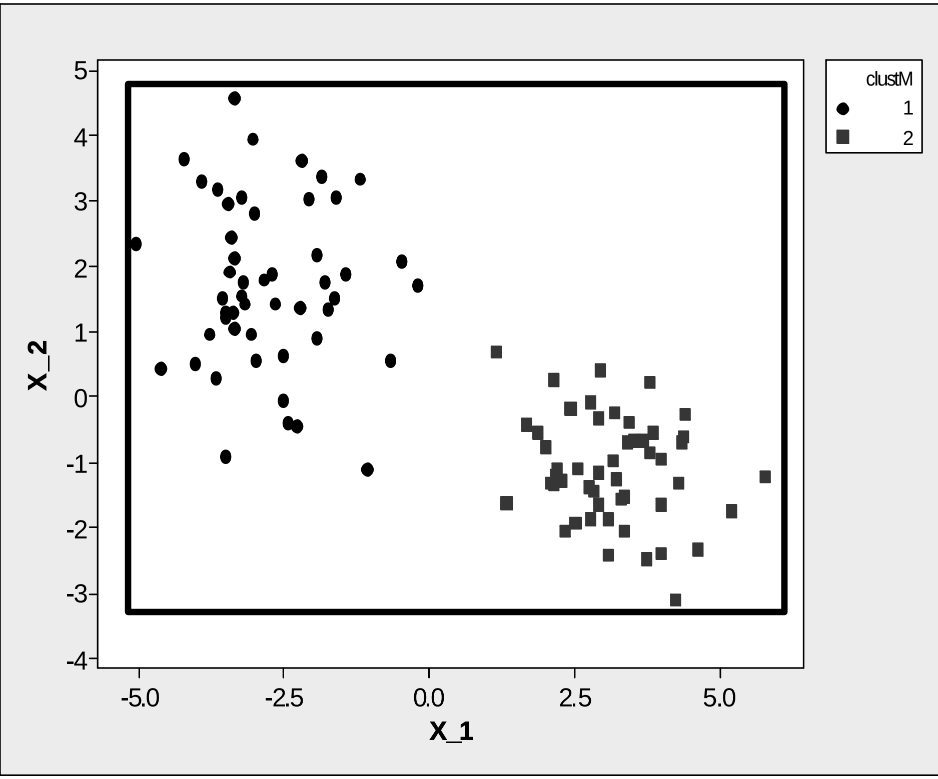
$$qual(K) = \mathrm{Prd}(K)$$

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

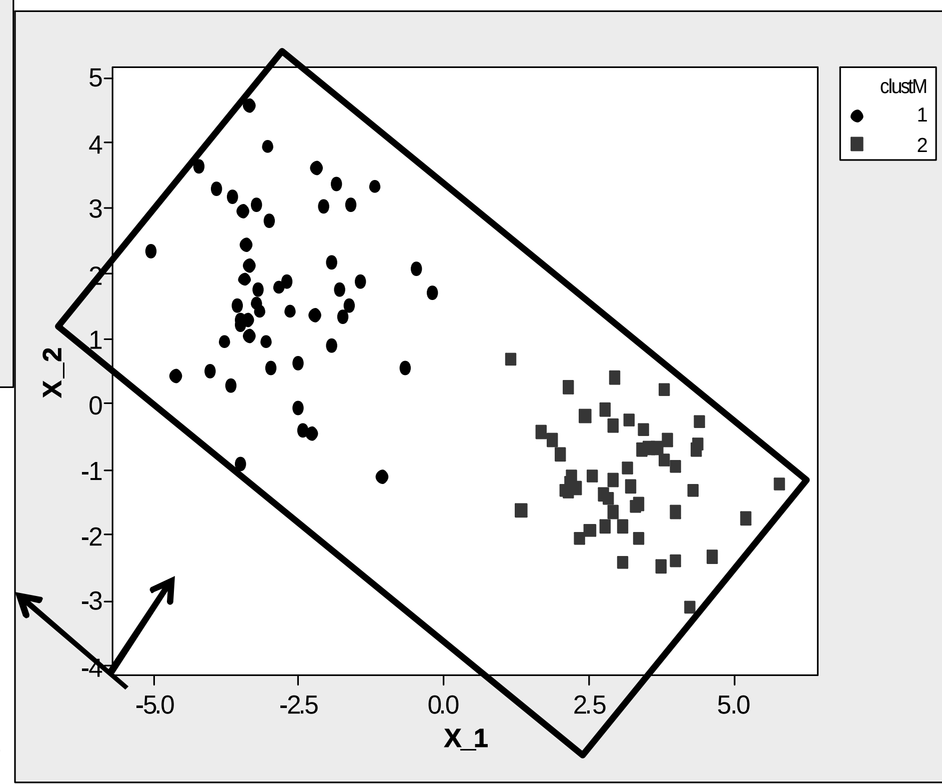
$$\hat{K} : \text{ among those such that } p(K) \leq \boldsymbol{p}, \max_K d(K)$$

Important: how does one select the reference distribution?

Most often used no-clustering scenarii, **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. Technical report, Dept of Biostatistics, Stanford University. More recent reference? [http://www-stat.stanford.edu/~tibs/research.html]

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