index.Gap(clusterSim)

Tibshirani, Walther and Hastie gap index

Step 1. Cluster the observed data $\boldsymbol{X} = \{x_{ij}\}$, $i = 1, \dots, n$; $j = 1, \dots, m$ (via e.g. any hierarchical clustering method, pam, *k*-means), varying the total number of clusters from $u = 1, \dots, n$, giving within-dispersion measures:

$$W_u = \text{trace}(\mathbf{W}_u),$$

where: $\boldsymbol{W}_u = \sum_r \sum_{i \in C_r} (\boldsymbol{x}_{ri} - \bar{\boldsymbol{x}}_r)\,(\boldsymbol{x}_{ri} - \bar{\boldsymbol{x}}_r)^T$ – within-group dispersion matrix for data clustered into u clusters,
$\boldsymbol{x}_{ri}$ – *m*-dimensional vector of observations of the *i*-th object in cluster *r*,
$\bar{\boldsymbol{x}}_r$ – centroid or medoid of cluster r,
$r = 1, \dots, u$ – cluster number,
u – number of clusters ($u = 1, \dots, n$),
n – number of objects,
m – number of variables,
C_r – the indices of objects in cluster r.

Step 2. Generate *B* reference data sets, using the uniform prescription:

a) generate each reference variable uniformly over the range of the observed values for that variable,

or

b) generate the reference variables from a uniform distribution over a box aligned with the principal components of the data. In detail, if $\boldsymbol{X} = \{x_{ij}\}$ is our $n \times m$ data matrix, assume that the columns have mean 0 and compute the singular value decomposition $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$. We transform via $\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{V}$ and then draw uniform features $\boldsymbol{Z}'$ over the ranges of the columns of $\boldsymbol{X}'$, as in method a) above. Finally we back-transform via $\boldsymbol{Z} = \boldsymbol{Z}'\boldsymbol{V}^T$ to give reference data $\boldsymbol{Z}$,

and cluster each one (using the same clustering method) giving within-dispersion measures W_{ub} ($b = 1, \dots, B$; $u = 1, \dots, n-1$). Compute the (estimated) gap statistic:

$$\text{Gap}(u) = \frac{1}{B}\sum_{b=1}^{B} log\, W_{ub} - log\, W_u$$

Step 3. Compute the standard deviation of $\{log\, W_{ub}\}$, $b = 1, \dots, B$:

$$sd_u = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(log\, W_{ub} - \bar{l}\right)^2},$$

where: $\bar{l} = \frac{1}{B}\sum_{b=1}^{B} log\, W_{ub}$,

and define

$$s_u = sd_u\sqrt{1 + 1/B}$$

Step 4. Finally choose the number of clusters via finding the smallest u such that:

$$\text{Gap}(u) \geq \text{Gap}(u+1) - s_{u+1}\ (u = 1, \dots, n-2)$$

References

Tibshirani R., Walther G., Hastie T. (2001), *Estimating the number of clusters in a data set via the gap statistic*, „Journal of the Royal Statistical Society", ser. B, vol. 63, part 2, 411-423.