Flexible parametric bootstrap for testing homogeneity against clustering and assessing the number of clusters

Christian Hennig
Department of Statistical Science, University College London, United Kingdom
c.hennig@ucl.ac.uk

Many cluster analysis methods deliver a clustering regardless of whether the dataset is indeed clustered or homogeneous, and need the number of clusters to be fixed in advance. Validation indexes such as the Average Silhouette Width are popular tools to measure the quality of a clustering and to estimate the number of clusters, usually by choosing the number of clusters that optimizes their value. Such indexes can be used for testing the homogeneity hypothesis against a clustering alternative by exploring their distribution, for a given number of clusters fitted by a given clustering method, under a null model formalising homogeneous data. The same approach can be used for assessing the number of clusters by comparing what is expected under the null model with what is observed under different numbers of clusters. Many datasets include some structure such as temporal or spatial autocorrelation that distinguishes them from a plain Gaussian or uniform model, but cannot be interpreted as clustering. The idea is to specify a null model for data that can be interpreted as homogeneous in the given application, which captures the non-clustering structure in the dataset by some parameters, which are estimated from the data, and then bootstrapping a cluster validity index can be used for testing homogeneity against a clustering alternative and for assessing the number of clusters. Applications will be presented.
Keywords: cluster analysis; average silhouette width; Bayesian Information Criterion.