

Title: **Estimating the Number of Clusters in a Dataset via the Gap Statistic**

Author(s): **Robert Tibshirani, Guenther Walther and Trevor Hastie**

Technical Report number (Dept. of Statistics, Stanford Univ.): **208**

Date: **April 2000**

Abstract:

We propose a method (the “Gap statistic”) for estimating the number of clusters (groups) in a set of data. The technique uses the output of any clustering algorithm (e.g. k-means or hierarchical), comparing the change in within cluster dispersion to that expected under an appropriate reference null distribution. Some theory is developed for the proposal and a simulation study that shows that the Gap statistic usually outperforms other methods that have been proposed in the literature. We also briefly explore application of the same technique to the problem for estimating the number of linear principal components.