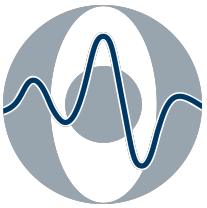
Exercises for
Pattern Analysis
Amir Davari, Dalia Rodriguez Salas, Johannes Bopp, Lina Felsner
Assignment 3, published on June 9, 2019



General Information:

Lecture:	Mon 16:15h – 17:45h and Wed 10:15h – 11:45h (H16)
Exercises:	Tue 14:00 – 15:00, 15:00 – 16:00 (02.151-113) and
	Thu 16:00 – 18:00 (0.01-142, 00.156-113, 02.151b-113)
Certificate:	Oral exam at the end of the semester
Contact:	amir.davari@fau.de & dalia.rodriguez@fau.de

K-Means and the Gap-Statistics

Exercise 1 Preliminary Remarks

In this exercise, we will play with the gap statistics for model selection of k-means clusters. If you like to look at the literature, it can be found in Sec. 14.3.11 of Hastie/Tibshirani/Friedman. If you like to read the original source, you can find the original work by Tibshirani in the literature section of the studon class.

Exercise 2 New Data

We unfortunately have to retire our raccoon, and to wish him all the best for his future carreer outside of our operation. Jokes aside: After the experiments in this exercise, you will likely understand why the raccoon density as we used it is not a good benchmark dataset. Otherwise, just briefly play with it after you finished everything else.

For this exercise, we will resort to a synthetic distribution that consists of random samples from four 2-D Gaussians. It may come handy to move the data creation to a function, in order to test some interesting constellations that illustrate the behavior of the gap statistics.

Exercise 3 Technical Approach

Implement Tibshirani's gap statistic to automatically select the correct number of components of the clustering. In case that you compare Sec. 14.3.11 of Hastie/Tibshirani/Friedman with the original paper, you will notice that the original paper proposes to optionally align the sampling of the reference density with the principal components of the original data. Although this is certainly a smart move for cases of ill-shaped clusters, you may leave it out for this exercise and just follow Sec. 14.3.11.

For each experiment, please produce four plots (either in four individual figures or in one figure with four subplots). These plots are:

(a) the input clusters, color-coded by their membership to the multivariate Gaussians

(b) the k-means result after selecting k with the gap statistics, color-coded by their membership to the k clusters.

(c) the decrease of the within-cluster distance of the original data (i.e., the left part of Fig. 14.11 in Hastie/Tibshirani/Friedman)

(d) the gap statistics, including the standard deviation of the reference clusters (i.e., the right part of Fig. 14.11 in Hastie/Tibshirani/Friedman)

Exercise 4 Experiments

Play with your new method! In which cases is the gap statistic likely to underestimate the number of clusters k?

One aspect that complicates the analysis is that both our data generation and also the k-means clustering are randomized. What might be a reasonable approach to nevertheless measure the performance of the gap statistics?

What are the issues with our raccoon density?

Extra brain teaser (no answer required): can you think of a case where the gap statistic overestimates the number of clusters?

Extra brain teaser (no answer required): assume that all we want from our raccoon is to cluster some of the bright parts of the original image — what might be a reasonable approach to make this happen?

Demonstration of your Results: please book a date on your exercise slot in studOn for Tuesday June 25 or Thursday June 27! The studon registration opens on June 11 at 10am.