Data set

Setup:

Generate many (e.g. 1e4) i.i.d.r.v.'s from a from a 1:2 mixture of the following two distributions…

The Gaussian distribution #1 with $\mu = 1$ and $\sigma = 0.5$

The Gaussian distribution #2 with $\mu = -1$ and $\sigma = 0.5$

The 1:2 mixture means that $^1/_3$ of r.v.'s come from Gaussian #1 and $^2/_3$ of r.v.'s come from Gaussian #2

Our data set has a pdf which looks like two peaks. Let's explore different ways we can estimate peak separation.

1. Form an empirical cumulative distribution function (ecdf) for your data. As we discussed, this curve still has all the data in it. Find a way to extract peak locations from ECDF alone. There is more than one right answer but try to make the most efficient estimator algorithm of this type. Use bootstrap to find the C.I.

2a. Treat the data as coming from a single distribution with the mixture ratio being yet another parameter and estimate peak separations and the C.I. using an MLE-based approach for this double-peaked distribution. Use MLE to find not just peak locations but also C.I. for each value, then use error propagation to find the C.I. for peaks separation.

2b. Use k-means clustering to find the border location for cluster separation. Use MLE for each cluster to find peak locations, then use error propagation to find the C.I. for peaks separation.

<u>Be extra careful about setting up MLE in this case!</u>

3. Use the method of moments for the 2-Gaussian distribution to find peak separations and use MELE and bootstrap to find the C.I. Compare the results from two techniques for this case.

Note: all C.I. should be at 95% confidence.

Finally, let's consider a data set where the mixing ratio varies from 1:100 to 1:1.

At what mixing ratio can you reject the hypothesis that the data comes from a single Gaussian with 95% confidence if you use…

1. Anderson-Darling test
2. Kolmogorov-Smirnov test