Problem Set 4

Data-driven bandwidth selection

Least-squares cross-validation

The LSCV-Criterion has the form

$$LSCV(h) = \frac{1}{n^2 h} \sum_{i=1}^{n} \sum_{j=1}^{n} \bar{k}\left(\frac{x_i - x_j}{h}\right) - \frac{2}{n(n-1)h} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} k\left(\frac{x_i - x_j}{h}\right)$$

where where $\bar{k}$ is a convolution kernel. We will use the Gaussian Kernel for k which implies the following convolution kernel:

$$\bar{k}(v) = \frac{exp(\frac{-v^2}{4})}{\sqrt{4\pi}}$$

- Write a function $LSCV(h)$ and simulate a data set with one univariate random variable with $x_i \sim \mathcal{N}(0, 1)$ and $n = 100$.

- Find the value h_{min} that minimizes $LSCV(h)$. Use an appropriate h_{grid} comprised of 50 equidistant points.
- Simulate and repeat the exercise from 1. and 2. 50 times and save the respective minimizing bandwidth in a vector $h_{min,\,all}$.
- Calculate the standard deviation and the mean of $h_{min,\,all}$. Plot the kernel density estimator for the smallest and the largest value of $h_{min,\,all}$ and compare the resulting function. What can you say about the "reliability" of the LSCV method for bandwidth selection?
- Repeat the simulation for $n = 1000$ and compare your results to those from 4. How do your results compare to the theoretical results regarding the relationship between the bandwidth and the asymptotic bias and variance?

Silverman's rule of thumb

Another bandwidth selection tool is Silverman's rule of thumb. *If* the underlying density is normal, then Silverman's rule returns the bandwidth that minimizes the MSE.

$$h_s = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-1/5},$$

where $\hat{\sigma}$ is the estimated standard deviation of the sample.

- Write a function simulating and returning a bimodal Gaussian mixture with $\mu_1, \mu_2, \sigma_1, \sigma_2$ for n observations with on average of $n/2$ observations in each distribution.
- Generate a bandwidth grid on $[0.1, 3]$ with 30, equidistant values.
- Plot three plots with the estimated density function keeping $\mu_1 = 0, \sigma_1 = 1, \sigma_2 = 1$ with h_{LSCV} and h_s for $\mu_2 = 20, \mu_2 = 10, \mu_2 = 1$.