Problem Set 4 - LSCV

Arno Lindemann, Julius Greiff

December 1, 2021

Simplification of LSCV

$$LSCV(h) = \frac{1}{n^2h}\sum_{i=0}^{n}\sum_{j=0}^{n}\bar{k}(\frac{x_i - x_j}{h}) - \frac{2}{n(n-1)h}\sum_{i=0}^{n}\sum_{j=0,j\neq i}^{n} k(\frac{x_i - x_j}{h})$$

$$= \underbrace{\sum_{i=0}^{n}\sum_{j=0}^{n}(\frac{1}{n^2h}\bar{k} - \frac{2}{n(n-1)h}k)}_{=\mathrm{A}} + \frac{2}{n(n-1)h}\sum_{i=0}^{n} k(\frac{x_i - x_i}{h})$$

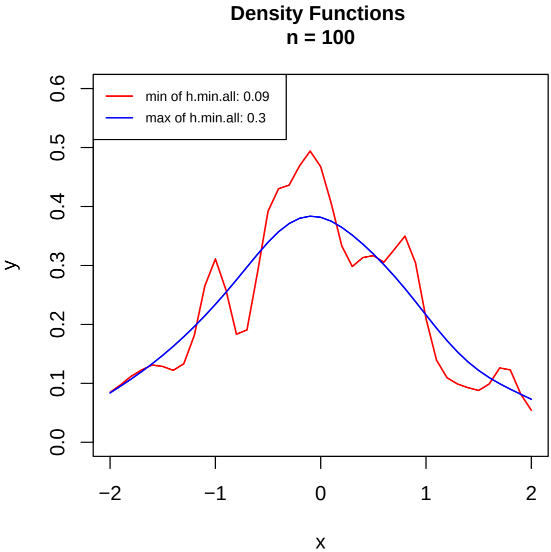
$$= \mathrm{A} + \frac{2}{n(n-1)h}\sum_{i=0}^{n} k(0)$$

$$= \mathrm{A} + \frac{2}{n(n-1)h}\sum_{i=0}^{n}\frac{1}{\sqrt{2\pi}}$$

$$= A + \frac{\sqrt{2}}{(n-1)h\sqrt{\pi}}$$

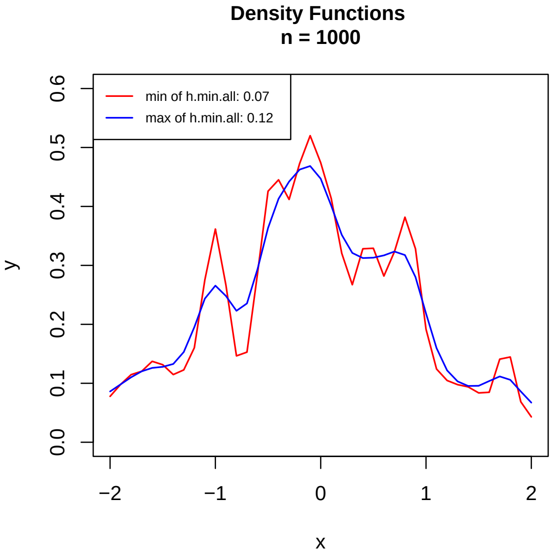
Task 4

- mean of h.min.all = 0.2188
- sd of h.min.all = 0.03889494



Task 5

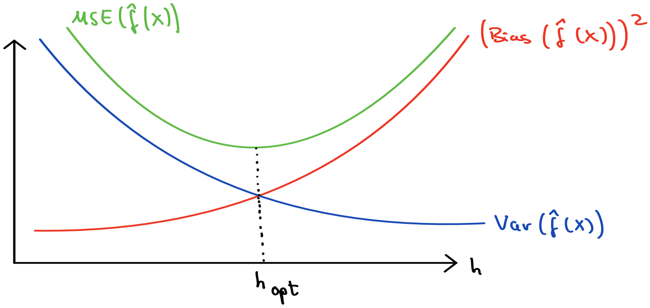
- mean of h.min.all $= 0.0968$
- sd of h.min.all $= 0.01132813$



Dependence of Asymptotic Variance and Bias on Bandwidth Selection

$$MSE(\hat{f}(x)) = (Bias(\hat{f}(x)))^2 + Var(\hat{f}(x))$$

- $(Bias(\hat{f}(x)))^2 = h^2 f''(x) \int z^2 k(z) dz + o(h^2)$
- $Var(\hat{f}(x)) = \frac{1}{nh} f(x) \int (k(z))^2 dz + o(\frac{1}{nh})$



Theoretical Results

Why is h smaller for larger n?

- Bias does not depend on n
- Variance depends on n
- For a larger sample size, we have a smaller variance for each h and thus tend to have a smaller optimal h

Why does the graph not look like a normal distribution for n=1000?

- Crossvalidation Undersmooth: For larger n, h tends to zero too quickly
- For large n, the leave-one-out method hardly makes a difference