STAT/Q SCI 403: Introduction to Resampling Method
Spring 2018
Homework 08

Instructions:

- You have to submit all your answers in a single PDF file generated by either LaTeX or *Rmarkdown*.
- You may use the LaTeXtemplate `HW_template.tex` to submit your answer.
- For questions using R, you have to attach your code in the PDF file. If the question ask you to plot something, you need to attach the plot in the PDF as well.
- If the question asks you to show a figure, the clarity of the figure will also be graded.
- The total score of this homework is 10 points but at most 8 points will be counted in the score. Namely, you can lose 2 points without any penalty. You will receive an extra credit if you answer all questions correctly.
- Questions with ♠ will be difficult questions.

Questions:

1. Let $X_1, \cdots, X_n \sim p$ an unknown smooth density function. Let $\widehat{p}_n(x) = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)$ be the KDE. In the first three sub-questions ((a)–(c)), we use the Gaussian kernel. i.e., $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

   (a) **(1 pt; ♠)** Assume p is infinitely differentiable. Let $g(x) = p'(x)$ be the derivative of $p(x)$ and $\widehat{g}_n(x) = \widehat{p}'_n(x)$ be the derivative of the KDE. For a fixed point x_0, assume $p(x_0) > 0$ and the third derivative $p'''(x_0) \neq 0$. Show that there are two positive constants $C_1, C_2 < \infty$ such that

   $$\textbf{bias}(\widehat{g}_n(x_0)) = C_1 h^2 + o(h^2), \quad \mathsf{Var}(\widehat{g}_n(x_0)) \leq \frac{C_2}{nh^3} + o\left(\frac{1}{nh^3}\right).$$

   (b) **(1 pt; ♠)** Assume that $p(x) = 2x \cdot I(0 \leq x \leq 1)$. Namely, $X_1, \cdots, X_n$ are from a triangle distribution over $[0, 1]$. Show that there is a positive number $C_3 < \infty$ such that

   $$\textbf{bias}(\widehat{p}_n(0)) = C_3 h + o(h).$$

   Namely, the bias at the *boundary* point is higher than the interior point. This phenomena is known as the *boundary bias*.

(c) **(1 pt; ♠)** Here is the method of *smooth bootstrap*. Given the original sample $X_1, \cdots, X_n$. As the usual bootstrap, we first sample with replacement to obtain $Y_1^*, \cdots, Y_n^*$. Then we generate IID random noises $Z_1, \cdots, Z_n \sim N(0, h^2)$ that are independent of $Y_1^*, \cdots, Y_n^*$. The final bootstrap sample is

$$X_1^*, \cdots, X_n^*, \quad \text{where } X_i^* = Y_i^* + Z_i.$$

Show that given the original sample $X_1, \cdots, X_n$ being fixed, each point in the smooth bootstrap sample, say X_1^*, has a PDF the same as $\widehat{p}_n$. Namely, show that the PDF of X_i^* given $X_1, \cdots, X_n$ being fixed is $\widehat{p}_n$.

(d) **(1 pt; ♠)** In this question, assume that we are using the uniform kernel. i.e., $K(x) = \frac{1}{2} I(-1 \le x \le 1)$. Assume p is infinitely differentiable and we choose $h = \frac{1}{n}$. Show that

$$n \cdot \widehat{p}_n(x_0) \xrightarrow{D} \mathsf{Poi}(\lambda(x_0))$$

for some $\lambda(x_0)$ depending on x_0.

Hint: *Law of small number.* Let $X_1, \cdots, X_n, \cdots$ be a sequence of random variables such that $X_n \sim \mathsf{Bin}(n, q_n)$ is a binomial distribution with parameter n, q_n. If $n \cdot q_n \to \lambda$, then $X_n \xrightarrow{D} \mathsf{Poi}(\lambda)$.

2. In this question, we will analyze the dataset `faithful`, a built-in dataset in R. We focus on the variable `eruptions`.

(a) **(1 pt)** Apply the KDE with three smoothing bandwidth $h = 0.1, 0.3, 0.9$. Show the three estimated density curves in the same plot. Remember to indicate which curve corresponds to which smoothing bandwidth.

(b) **(1 pt)** Apply the KDE with smoothing bandwidth $h = 0.3$ and compare to the density histogram (with `breaks = 20`).

(c) **(1 pt)** Apply the KDE with smoothing bandwidth $h = 0.3$. Use the bootstrap to construct a 95% confidence band.

3. In this question, we will use Monte Carlo Simulation to analyze the errors of the KDE. In particular, we will draw samples from the double exponential distribution with rate parameter $\lambda = 1$. Namely, the density we are drawing from is

$$p(x) = \frac{1}{2} e^{-|x|}.$$

The sample size we are using is $n = 1000$.

(a) **(1 pt)** Draw a single sample from that double exponential distribution. Apply the KDE with smoothing bandwidth $h = 0.2$. Show the KDE and the true density curve in the same plot.

(b) **(1 pt)** We see a huge difference between the KDE and the true density curve at $x = 0$. Explain why this occurs. Hint: is this from bias? or is it from the variance? and why?

(c) **(1 pt)** Analyzing the MISE for h within the range 0.05 to 0.5 using Monte Carlo Simulation with at least $N = 10000$. For simplicity, when evaluating the MISE, we only compare the density $p(x)$ and the KDE $\widehat{p}_n$ within the range $x \in [-6, 6]$. Show the MISE versus smoothing bandwidth h in a single plot. About what value of h will the MISE be minimized?