STAT 535, Homework 3

Due date: Nov. 7 Thursday 23:59:59. Submit the homework through Canvas in a PDF file. If the questions involved programming, please include your codes.

1. **(10 pts)** Let $X_1, \cdots, X_n \sim p$, where $p(x) = 2x \cdot I(0 \leq x \leq 1)$. Namely, $X_1, \cdots, X_n$ are from a triangle distribution over $[0, 1]$. Assume that we are using a Gaussian kernel to construct a KDE.

   (a) **(3 pts)** Show that there is a positive number $C_3 > 0$ such that

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

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

   (b) **(3 pts)** Moreover, show that

   $$|\mathbf{bias}(\widehat{p}_n(1))| \geq c_0 > 0$$

   for some constant c_0 that does not depend on h nor n. Namely, the KDE is inconsistent at the point where the density has a jump.

   (c) **(4 pts)** For points $x \in (0, 1)$ show that the derivative of the KDE $\frac{d}{dx}\widehat{p}_n(x)$ is a consistent estimator of the derivative of $p(x)$. Show that the bias and variance converge to 0 when $n \to \infty$ and $h \to 0$.

2. **(10 pts)** Let $X_1, \cdots, X_n \sim p$ and p is an unknown infinitely differentiable 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. Assume that we use a kernel function satisfying

   $$0 = \int xK(x)dx = \int x^2 K(x)dx = \int x^3 K(x)dx = \cdots = \int x^q K(x)dx$$

   and $\int x^{q+1}K(x)dx \neq 0$ for some integer q. Consider the scenario $h \to 0$ as $n \to \infty$.

   This type of kernel function is called a higher-order kernel function. It leads to a KDE with a smaller bias compared to the KDE from a regular kernel function.

   (a) **(8 pts)** At a given point x_0, find out the first order term of the bias in terms of h, i.e., finding out s such that

   $$\mathbf{bias}(\widehat{p}_n(x_0)) = C_1 h^s + o(h^s)$$

   for some constant C_1.

   (b) **(3 pts)** What would happen if the density p only has a derivative up to ℓ-th order?

   (c) **(4 pts)** Although this estimator may have a smaller bias than the usual Gaussian kernel KDE, the density estimator may not be a density function. Explain why.

3. **(10 pts)** In R, the faithful dataset (`faithful`) is a famous dataset with two variables `eruptions` (eruption time) and `waiting` (waiting time). In this question, we will analyze these two variables using the KDE and the kernel regression. Define a kernel function $K(x) = (1 - |x|)I(|x| \leq 1)$; it looks like a triangle so it is also known as a triangle kernel.

   (a) **(3 pts)** Suppose we use the triangle kernel to construct a KDE of variable `waiting`. Under the Silverman's rule for choosing the smoothing bandwidth, plot the estimated density plot.

(b) Suppose we use the triangle kernel to perform a kernel regression with response variable $Y =$`waiting` and covariate $X =$`eruption`. Use a 5-fold cross-validation to choose the smoothing bandwidth (repeat at least 100 times).

i. **(3 pts)** Plot the cross-validation error (under squared distance) versus bandwidth (you need to search within at least $h \in [0.1, 0.5]$).

ii. **(2 pts)** What is the optimal bandwidth you choose?

iii. **(2 pts)** Make a scatter plot of the two variables along with the fitted regression curve.

Note: here is an R script that includes one possible implementation of cross-validation: ***http://faculty.washington.edu/yenchic/17Sp_403/403_17lab9-sol.R***. *The associated lecture note is in:* ***http://faculty.washington.edu/yenchic/17Sp_403/Lec8-NPreg.pdf***.

4. **(10 pts)** Consider a nonparametric regression setting where we observe pairs $(X_1, Y_1), \cdots, (X_n, Y_n)$ and we assume both $X_i, Y_i \in \mathbb{R}$. The local polynomial estimator with q-th order attempts to minimizes

$$\mathsf{LPR}(\beta_0, \beta_1, \cdots, \beta_q; x) = \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)(Y_i - \beta_0 - \beta_1 X_i - \cdots \beta_q X_i^q)^2.$$

The mean can be estimated using the solution $\widehat{\beta}_0$ that minimizes the above criterion.

(a) **(5 pts)** Show that

$$\widehat{\beta}_0(x) = e_1^T(\mathbb{X}^T W(x)\mathbb{X})^{-1}\mathbb{X}^T W(x)\mathbb{Y},$$

where

$$W(x) \in \mathbb{R}^{n\times n} = \mathsf{Diag}\left(K\left(\frac{x - X_1}{h}\right), \cdots, K\left(\frac{x - X_n}{h}\right)\right)$$

and $\mathbb{X}$ is some matrix of the covariates.

(b) **(2 pts)** What will the matrix $\mathbb{X}$ be in this case?

(c) **(3 pts)** Is it a linear smoother?

5. **(10 pts)** Assume that we observe a random sample $(X_1, Y_1), \cdots, (X_n, Y_n)$ with $X_i \in [0, 1] \subset \mathbb{R}$ for each i. Suppose that we are using a regression spline method to estimate the regression function $m(x) = \mathbb{E}(Y_1|X_1 = x)$. The basis we are using is the truncation power basis:

$$h_1(x) = 1, h_2(x) = x, h_3(x) = x^2, h_4(x) = x^3,$$

and

$$h_j(x) = (x - \tau_{j-4})_+^3, \quad j = 5, 6, \cdots, M + 4,$$

where $(x)_+ = \max\{x, 0\}$. In a regression spline (not smoothing spline), the knots

$$\tau_1 < \cdots < \tau_M$$

are chosen by the user so here we assume that these basis are known.

Recall that the estimator $\widehat{m}$ can be written as

$$\widehat{m}(x) = \sum_{j=1}^{M+4} \widehat{\beta}_j h_j(x),$$

for some properly chosen $\widehat{\beta}_j$, where

$$\widehat{\beta}_1, \cdots, \widehat{\beta}_{M+4} = \mathsf{argmin}_{\beta_1, \cdots, \beta_{M+4}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{M+4} \beta_j h_j(X_i) \right)^2 .$$

(a) **(2 pts)** Explain the difference between regression spline and smoothing spline.

(b) **(2 pts)** Show that the estimator can be written as

$$\widehat{m}(x) = H^T(x)\widehat{\beta} = H^T(x)(\mathbb{H}^T\mathbb{H})^{-1}\mathbb{H}^T\mathbb{Y},$$

where $H(x) \in \mathbb{R}^{M+4}$ is a vector of $(M+4)$ elements and $\mathbb{H}$ is an $n \times (M+4)$ matrix. You need to find $H(x)$ and $\mathbb{H}$.

(c) **(2 pts)** Is regression spline a linear smoother?

(d) **(2 pts)** Assume that the noise is homogeneous with $\sigma^2 = Var(Y_1|X_1 = x)$ and the covariates are non-random (fixed design). Find an estimator of σ^2.

(e) **(2 pts)** Because the regression spline does not have a penalty term, what might be a possible problem if we allow $M \to \infty$ when $n \to \infty$ (when we have more observations, we fit more knots)?