MATH524 – Spring 2025

Problem Set: Week 3

1. (**Optimal kernel**) Consider the second-order Epanechnikov kernel defined as

$$K_E(x) = \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right)\mathbf{1}_{\{|x|\leq\sqrt{5}\}},$$

and note that $\int |u|^2|K_E(u)|\,\mathrm{d}u = 1$. Let K_0 be another non-negative second-order kernel with $\int |u|^2|K_0(u)|\,\mathrm{d}u = 1$. By considering $e(x) = K_0(x) - K_E(x)$, or otherwise, show that the Epanechnikov kernel always has lower risk than any K_0. That is, $R(K_0) \geq R(K_E)$.

2. (**Linear Smoothers, Cross-validation**)

Let $\{(y_i, x_i) : 1 \leq i \leq n\}$ be a random sample taking values in $\mathbb{R}^2$. A linear smoother is given by

$$\hat{e}(x) = \sum_{i=1}^{n} w_{n,i}(x)y_i, \qquad w_{n,i}(x) = w(x_1, x_2, \cdots, x_n; x).$$

Note that $w_{n,i}(x)$ is only a function of $\{x_i : 1 \leq i \leq n\}$ and not of $\{y_i : 1 \leq i \leq n\}$. Recall that local polynomial regression takes on the following form

$$\hat{e} = \mathbf{e}_0' \arg\min_e \sum_{i=1}^{n}(y_i - p(x_i - x)'e)^2 K_h(x_i - x)$$

where $\mathbf{e}_0$ is the first basis unit vector and $p(x) = (1, x, x^2, \ldots, x^p)'$ is the polynomial basis up to order p.

(a) Show that local polynomial regression estimators can be written as linear smoothers and give the exact form of the "smoothing weights" $w_{n,i}(x)$.

(b) Show the following simplified cross-validation formula holds for local polynomial regression. [1]

$$\mathsf{CV}(c) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{e}_{(i)}(x_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{e}(x_i)}{1 - w_{n,i}(x_i)}\right)^2,$$

where $\hat{e}_{(i)} = \sum_{j\neq i} w_{n,j}(x)y_j$ is the leave-one-out estimator and c denotes a tuning parameter (i.e., a bandwidth h_n for local polynomials).

(c) Providing regularity conditions, show that

$$\frac{\hat{e}(x) - e(x)}{\sqrt{\mathbb{V}[\hat{e}(x)|x_1, x_2, \cdots, x_n]}} \to_d \mathcal{N}(0, 1).$$

where $e(x) = \mathbb{E}[Y|X = x]$

(d) Propose an asymptotically valid 95% confidence interval for $e(x)$, with x fixed. That is,

$$\forall x : \; \liminf_n \mathbb{P}\left[e(x) \in \mathsf{C.I.}(x)\right] \geq 0.95.$$

[1] The following result is useful: for an invertible matrix $\mathbf{A}$ and a column vector $\mathbf{v}$, and $\lambda \neq -1/(\mathbf{v}'\mathbf{A}\mathbf{v})$ the following holds

$$\left(\mathbf{A} + \lambda\mathbf{v}\mathbf{v}'\right)^{-1} = \mathbf{A}^{-1} - \frac{\lambda\mathbf{A}^{-1}\mathbf{v}\mathbf{v}'\mathbf{A}^{-1}}{1 + \lambda\mathbf{v}'\mathbf{A}^{-1}\mathbf{v}}.$$

Is this derived confidence interval equivalent to the uniform confidence band? That is, does it satisfy the following probability expression?

$$\liminf_{n} \mathbb{P}\left[\forall x:\; e(x) \in \mathsf{C.I.}(x)\right] \geq 0.95?$$

Explain your answer.

(e) Conduct the following Monte Carlo experiment. You are free to use inbuilt commands or libraries for matrix operations, dataframe structures, quantile calculations and plotting, but should *not* use any pre-packaged local polynomial regression implementations.

Consider the following DGP

- $x_i \sim \mathsf{Uniform}(-1, 1)$;
- $y_i = 0.3x_i^2 - 1.5x_i^3 + 0.2x_i^4 - 0.002x_i^5 + \varepsilon_i$;
- $\varepsilon_i \sim \mathcal{N}(0, 0.1^2)$,
- Consider the second-order Epanechnikov kernel, $K(u) = \frac{3}{4}(1 - u^2)$ for $-1 \leq u \leq 1$.

The dataset generated by this process is provided as a CSV on Moodle, named `exercise3.csv`.

The first column of the CSV contains the y_i's and the second column contains the x_i's.

i. Consider a degree 3 (p = 3) local polynomial estimator of $\mu(x)$, that is, $\hat{e}(x_i)$. Plot the $\mathsf{CV}(h)$, as a function of h and compute the CV estimator, denoted $\hat{h}_{\mathtt{CV}}$. Use h between 0.5 and 1.0 with 0.1 increments.

ii. Using the data-driven tuning parameter choice $\hat{h}_{\mathtt{CV}}$, plot the following functions of $x \in [-1, 1]$ (in one single graph): (i) the true regression function; (ii) the estimated regression function $\hat{e}(x)$; (iii) the data. (Using a grid of 10 evaluation points should be enough.)