Fall 2021 — STAT 2221: Advanced Applied Multivariate Analysis

Homework 1. Due by Thursday, September 16, 2021 at 9:00 a.m. EST

Topic: Density Estimation

Name: Giang Vu

Additional instructions. Add `\newpage` immediately before each problem so that each has its own page. Add `\begin{proof}[Solution.] ... \end{proof}` below each problem for providing your solution. You are welcome to add additional packages to the preamble, but please do not modify the existing commands and formatting.

Problem 1.1. *Rosenblatt's density estimator* is given by

$$\widehat{p}_n(x) = h^{-1}\left[F_n\left(x+\tfrac{h}{2}\right) - F_n\left(x-\tfrac{h}{2}\right)\right],$$

where h denotes bin width and $F_n(x)$ is the empirical cumulative distribution function at $x \in \Re$.

1. Show that this estimator is a kernel density estimator.
2. What type of kernel corresponds to this estimator?
3. Derive the bias and variance of this estimator.
4. Using your results from above, find its mean-integrated squared error (MISE).
5. Apply this kernel density estimator to estimate the density of the 1892 Hidalgo stamp data (on Canvas). What do you notice about the smoothness of the resulting density estimate?

Proof. 1.

$$\begin{aligned}
\widehat{p}_n(x) &= h^{-1}\left[F_n\left(x+\tfrac{h}{2}\right) - F_n\left(x-\tfrac{h}{2}\right)\right] \\
&= \frac{1}{h}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(X_i \le x+\frac{h}{2}\right) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(X_i \le x-\frac{h}{2}\right)\right] \\
&= \frac{1}{nh}\left[\sum_{i=1}^{n}\mathbb{I}\left(X_i \le x+\frac{h}{2}\right) + \mathbb{I}\left(X_i > x-\frac{h}{2}\right)\right] \\
&= \frac{1}{nh}\left[\sum_{i=1}^{n}\mathbb{I}\left(x-\frac{h}{2} < X_i \le x+\frac{h}{2}\right)\right] \\
&= \frac{1}{nh}\left[\sum_{i=1}^{n}\mathbb{I}\left(|X_i| \le x+\frac{h}{2}\right)\right] \\
&= \frac{1}{nh}\left[\sum_{i=1}^{n}\mathbb{I}\left(\left|\frac{X_i-x}{h}\right| \le \frac{1}{2}\right)\right]
\end{aligned}$$

Therefore, we will have this estimator as a kernel density estimator

$$\widehat{p}_n(x) = \frac{1}{nh}\left[\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\right]$$

with

$$K(t) = \mathbb{I}\left(-\frac{1}{2} < t \leq \frac{1}{2}\right)$$

We can check that this kernel is indeed a pdf

$$\int_{-1/2}^{1/2} K(t)dt = \int_{-1/2}^{1/2} dt = \frac{1}{2} + \frac{1}{2} = 1$$

2. This looks like the form of the rectangular kernel.

3. The bias of this estimator is

$$\begin{aligned} bias\{\widehat{p}_n(x)\} &= \mathbb{E}_p\{\widehat{p}_n(x)\} - p(x) = \mathbb{E}_p\left\{\frac{1}{nh}\left[\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\right]\right\} - p(x) \\ &= \frac{1}{h}\mathbb{E}_p\left\{K\left(\frac{X - x}{h}\right)\right\} - p(x) = \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{X - x}{h}\right) p(X)dX - p(x) \end{aligned}$$

Let $u = \frac{X-x}{h}$, $u \in \left(-\frac{1}{2}, \frac{1}{2}\right]$. Then we will have $X = uh + x$ and $\frac{du}{dX} = \frac{1}{h} \rightarrow dX = h(du)$.

$$\rightarrow bias\{\widehat{p}_n(x)\} = \int_{-\frac{1}{2}}^{\frac{1}{2}} K(u)\, p(uh + x)du - p(x)$$

Using Taylor's expansion, when h is small, we have

$$p(uh + x) = p(x) + (uh)p'(x) + \frac{1}{2}(u^2h^2)p''(x) + o(h^2)$$

Therefore,

$$\begin{aligned} bias\{\widehat{p}_n(x)\} &= \int_{-\frac{1}{2}}^{\frac{1}{2}} K(u)\left(p(x) + (uh)p'(x) + \frac{1}{2}(u^2h^2)p''(x) + o(h^2)\right)du - p(x) \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} K(u)\, p(x)du + \int_{-\frac{1}{2}}^{\frac{1}{2}} K(u)\,(uh)p'(x)du + \frac{1}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} K(u)\,(u^2h^2)p''(x)du + \\ &\quad \int_{-\frac{1}{2}}^{\frac{1}{2}} K(u)\, o(h^2)du - p(x) \\ &= p(x)\int_{-\frac{1}{2}}^{\frac{1}{2}} K(u)\, du + hp'(x)\int_{-\frac{1}{2}}^{\frac{1}{2}} uK(u)\, du + \frac{1}{2}h^2p''(x)\int_{-\frac{1}{2}}^{\frac{1}{2}} K(u)\,(u^2)du + \\ &\quad o(h^2)\int_{-\frac{1}{2}}^{\frac{1}{2}} K(u)\, du - p(x) \end{aligned}$$

We can easily calculate

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} K(u)\,du = 1;\ \int_{-\frac{1}{2}}^{\frac{1}{2}} uK(u)\,du = \left(\frac{u^2}{2}\right)\Big|_{-\frac{1}{2}}^{\frac{1}{2}} = 0;\ \int_{-\frac{1}{2}}^{\frac{1}{2}} K(u)\,u^2du = \left(\frac{u^3}{3}\right)\Big|_{-\frac{1}{2}}^{\frac{1}{2}} = \frac{1}{12}$$

So now we have

$$\begin{aligned} bias\{\widehat{p}_n(x)\} &= p(x) + \frac{1}{24}h^2p''(x) + o(h^2) - p(x) \\ &= \frac{1}{24}h^2p''(x) + o(h^2) \end{aligned}$$

The variance of this estimator is

$$\begin{aligned} Var\{\widehat{p}_n(x)\} &= Var\left\{\frac{1}{nh}\left[\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\right]\right\} = \frac{1}{nh^2}Var\left\{K\left(\frac{X_i - x}{h}\right)\right\} \\ &\le \frac{1}{nh^2}\mathbb{E}\left\{K^2\left(\frac{X_i - x}{h}\right)\right\} = \frac{1}{nh^2}\int_{-\infty}^{\infty} K^2\left(\frac{X - x}{h}\right)p(X)d(X) \\ &= \frac{1}{nh}\int_{1/2}^{1/2} K^2(u)\,p(uh + x)d(u) \\ &= \frac{1}{nh}\int_{1/2}^{1/2} K^2(u)\,[p(x) + uhp'(x) + o(h)]d(u) \\ &= \frac{1}{nh}p(x)\int_{1/2}^{1/2} K^2(u)\,d(u) + \frac{1}{n}p'(x)\int_{1/2}^{1/2} uK^2(u)\,d(u) + o\left(\frac{1}{nh}\right)\int_{1/2}^{1/2} K^2(u)\,d(u) \end{aligned}$$

We can easily calculate

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} K^2(u)\,du = 1;\ \int_{-\frac{1}{2}}^{\frac{1}{2}} uK^2(u)\,du = \left(\frac{u^2}{2}\right)\Big|_{-\frac{1}{2}}^{\frac{1}{2}} = 0$$

So now we have

$$Var\{\widehat{p}_n(x)\} \approx \frac{1}{nh}p(x) + o\left(\frac{1}{nh}\right)$$

4. The mean squared error (MSE) of this estimator is

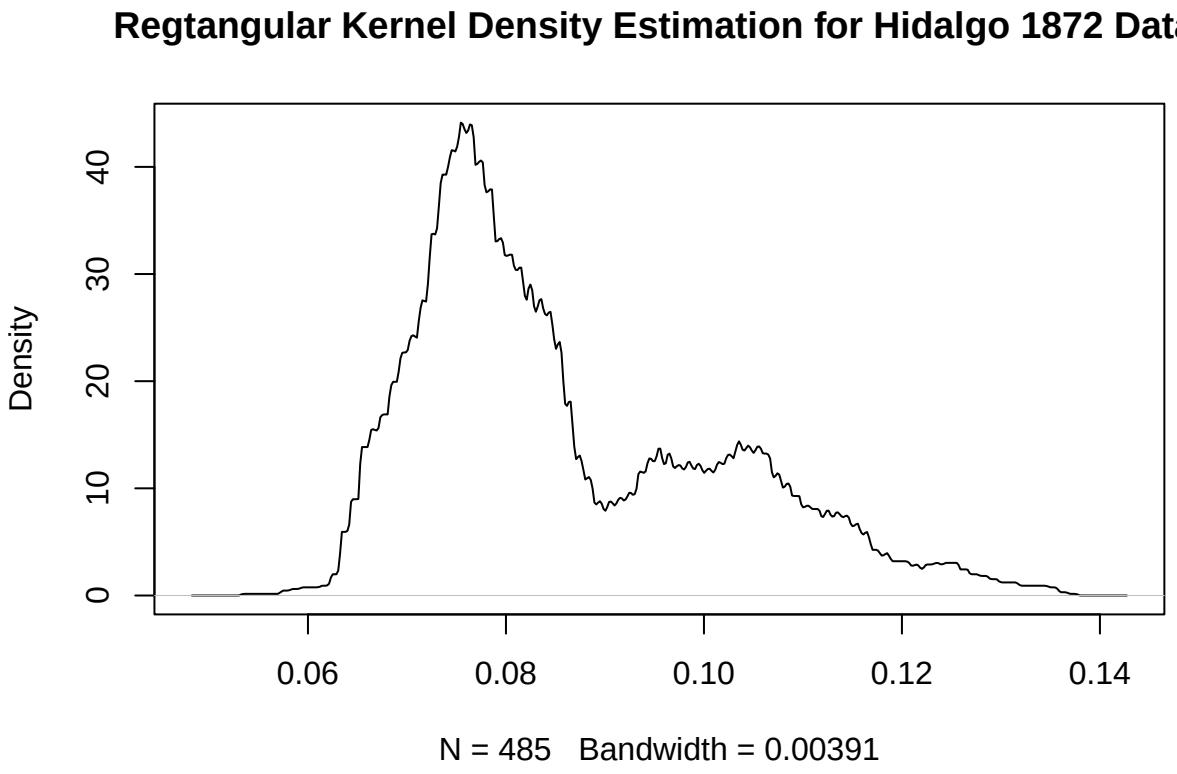
$$\begin{aligned} MSE(\widehat{p}_n(x)) &= bias^2\{\widehat{p}_n(x)\} + Var\{\widehat{p}_n(x)\} \\ &= \frac{1}{576}h^4(p''(x))^2 + o(h^4) + \frac{1}{nh}p(x) + o\left(\frac{1}{nh}\right) \end{aligned}$$

The mean-integrated square error (MISE) is

$$
\begin{aligned}
MISE(\widehat{p}_n(x)) &= IMSE(\widehat{p}_n(x)) = \int MSE(\widehat{p}_n(x))dx \\
&= \int \left[\frac{1}{576} h^4 (p''(x))^2 + o(h^4) + \frac{1}{nh} p(x) + o\left(\frac{1}{nh}\right) \right] dx \\
&= \frac{1}{576} h^4 \int (p''(x))^2 dx + \frac{1}{nh} \int p(x)dx + o(h^4) + o\left(\frac{1}{nh}\right) \\
&= \frac{1}{576} h^4 \int (p''(x))^2 dx + \frac{1}{nh} + o(h^4) + o\left(\frac{1}{nh}\right) \\
&= O(h^4) + O\left(\frac{1}{nh}\right)
\end{aligned}
$$

□

5. The estimated density for the Hidalgo stamp data using regtangular kernel is plotted below using the function density() in R, and default bandwidth (0.00391).
We can see that this kernel density estimate doesn't look too smooth. There is a lot of noise coming from the structure and detail of the underlying density of the data.



Problem 1.2. Let $\mathcal{N}(\mu, \sigma^2)$ denote the normal distribution. Consider the *claw density*

$$p(x) = \frac{1}{2}\mathcal{N}(0,1) + \frac{1}{10}\sum_{k=0}^{4}\mathcal{N}\left(\tfrac{k}{2} - 1, \left(\tfrac{1}{10}\right)^2\right).$$

1. If you are computing with `R`, designate `set.seed(1234)`. Generate $n = 100, 200, 300$ observations from the above density, and estimate the density using the Bartlett–Epanechnikov kernel density estimator. Repeat $1,000$ times at each sample size. Compare the performance of unbiased cross-validation (UCV), biased cross-validation (BCV), and Sheather–Jones plug-in (SJPI) window-width estimators for each simulation. Which window-width estimation method best finds the claws? Discuss.

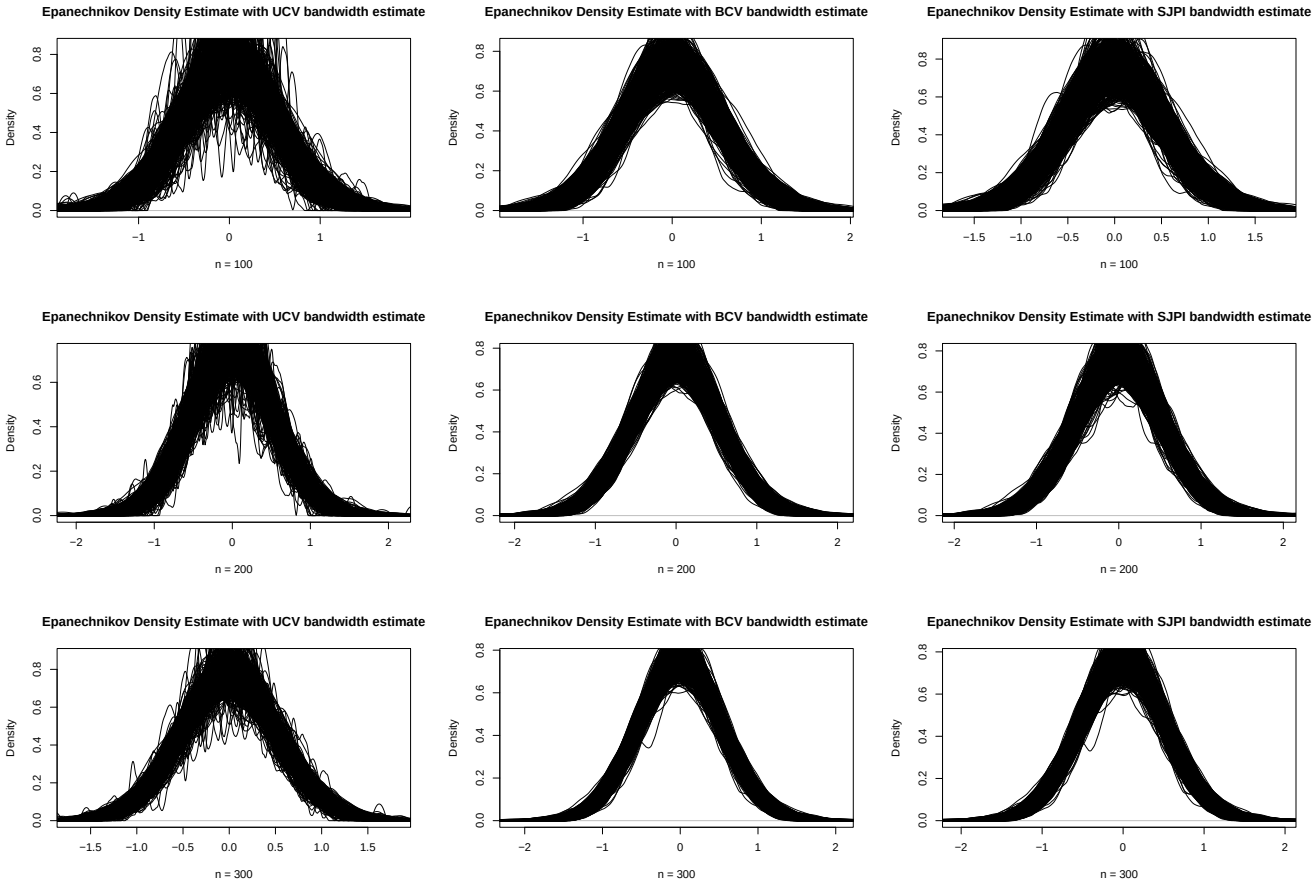
2. Consider the dataset `ushighways` (on Canvas) consisting of the approximate length (in miles) of all 212 U.S. 3-digit interstate highways (spurs and connectors). Compare Bartlett–Epanechnikov kernel density estimates for these data using UCV, BCV, and SJPI.

Proof.

$$\begin{aligned} p(x) &= \frac{1}{2}\mathcal{N}(0,1) + \frac{1}{10}\sum_{k=0}^{4}\mathcal{N}\left(\tfrac{k}{2} - 1, \left(\tfrac{1}{10}\right)^2\right) \\ &= \frac{1}{2}\mathcal{N}(0,1) + \frac{1}{10}\left(\mathcal{N}(-1, 0.01) + \mathcal{N}(-0.5, 0.01) + \mathcal{N}(0, 0.01) + \mathcal{N}(0.5, 0.01) + \mathcal{N}(1, 0.01)\right) \end{aligned}$$

□

1. Below are the plots of Barlett-Epanechnikov kernel density estimates for 1000 samples of size n from the claw density ($n = 100, 200, 300$), with window-width h estimated by 3 methods, UCV, BCV. and SJPI, respectively.
   Looking at the 1000 kernel density estimates together in a plot, we can see that overall, the UCV method gives us the most "claw"-like estimates, with the most detail/noise in the plots. On the other hand, we have the BCV method that provides the smoothest estimates, with the least detail from the original claw density. The SJPI gives us something in between, its estimates show a moderate amount of smoothness as well as detail.
   I also looked at the average bandwidth h estimates from each method across all 1000 iterations, the UCV method indeed gave me the smallest average estimate for h, hence the most noise, while BCV gave me the biggest average estimate for h (smoothest density estimates), and that value for SJPI is in the middle of the other two.



2. Below are the plots of Barlett-Epanechnikov kernel density estimates with window-width h estimated by 3 methods, UCV, BCV. and SJPI, respectively.
We can see that UCV method gives us the most detail/noise about the true density (lowest window-width estimate), while BCV gives us the smoothest density estimate (highest window-width estimate), and SJPI is in the middle of them with moderate smoothness and noise.