MATH524 – Spring 2025

Problem Set: Week 4

1. (**Kernel Density Estimation with a Fourth Order Kernel**)

   Let $X_1, \ldots, X_n$ be i.i.d. real-valued samples from a distribution with Lebesgue density function $f(x)$. Suppose that f is four times continuously differentiable with $|f^{(r)}(x)| \leq M$ for $0 \leq r \leq 4$.

   Define the fourth order Epanechnikov kernel as $K(u) = \frac{45}{32}\left(1 - \frac{7u^2}{3}\right)(1 - u^2)$ for $-1 \leq u \leq 1$. Recall that the kernel density estimator of $f(x)$ is given by

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i - x}{h}\right).$$

   (a) Show that K is indeed a fourth order kernel by proving the following four statements.

$$\int_{-1}^{1} K(u)\,\mathrm{d}u = 1, \qquad \int_{-1}^{1} uK(u)\,\mathrm{d}u = 0, \qquad \int_{-1}^{1} u^2K(u)\,\mathrm{d}u = 0, \qquad \int_{-1}^{1} u^3K(u)\,\mathrm{d}u = 0.$$

   You may wish to use the fact that $K(u) = K(-u)$.

   Solution: Since $K(u) = K(-u)$, we have that K is an even function. This immediately shows that $\int_{-1}^{1} u^r K(u)\,\mathrm{d}u = 0$ whenever $r \geq 1$ is odd. For the remaining two integrals,

$$\begin{aligned}
\int_{-1}^{1} K(u)\,\mathrm{d}u &= \int_{-1}^{1} \frac{45}{32}\left(1 - \frac{7u^2}{3}\right)(1 - u^2)\,\mathrm{d}u = \int_{-1}^{1} \frac{105u^4}{32} - \frac{75u^2}{16} + \frac{45}{32}\,\mathrm{d}u \\
&= \left[\frac{105u^5}{160} - \frac{75u^3}{48} + \frac{45u}{32}\right]_{u=-1}^{1} = \frac{105}{80} - \frac{75}{24} + \frac{45}{16} = 1, \\
\int_{-1}^{1} u^2K(u)\,\mathrm{d}u &= \int_{-1}^{1} \frac{45u^2}{32}\left(1 - \frac{7u^2}{3}\right)(1 - u^2)\,\mathrm{d}u = \int_{-1}^{1} \frac{105u^6}{32} - \frac{75u^4}{16} + \frac{45u^2}{32}\,\mathrm{d}u \\
&= \left[\frac{105u^7}{224} - \frac{75u^5}{80} + \frac{45u^3}{96}\right]_{u=-1}^{1} = \frac{105}{112} - \frac{75}{40} + \frac{45}{48} = 0.
\end{aligned}$$

   (b) Show that the variance is bounded by

$$\mathbb{V}\left[\widehat{f}(x)\right] \leq \frac{5M}{4nh}.$$

   You may wish to show first that $\int_{-1}^{1} K(u)^2\,\mathrm{d}u = \frac{5}{4}$.

   Solution: By i.i.d. and linearity,

$$\begin{aligned}
\mathbb{V}\left[\widehat{f}(x)\right] &\leq \frac{1}{nh^2}\mathbb{E}\left[K\left(\frac{X_i - x}{h}\right)^2\right] = \frac{1}{nh^2}\int_{x-h}^{x+h} K\left(\frac{u - x}{h}\right)^2 f(u)\,\mathrm{d}u \\
&\leq \frac{M}{nh^2}\int_{x-h}^{x+h} K\left(\frac{u - x}{h}\right)^2 \mathrm{d}u = \frac{M}{nh}\int_{-1}^{1} K(u)^2\,\mathrm{d}u \\
&= \frac{M}{nh}\int_{-1}^{1} \frac{11025u^8}{1024} - \frac{7875u^6}{256} + \frac{15975u^4}{512} - \frac{3375u^2}{256} + \frac{2025}{1024}\,\mathrm{d}u = \frac{5M}{4nh}.
\end{aligned}$$

   (c) Show that the bias is bounded by

$$\left|\mathbb{E}\left[\widehat{f}(x)\right] - f(x)\right| \leq \frac{45Mh^4}{384}.$$

You may wish to use a third order Taylor expansion with Lagrange remainder for f around the point x. Note that $|K(u)| \leq \frac{45}{32}$.

Solution: By Taylor's theorem with Lagrange remainder,

$$\begin{aligned}\left|\mathbb{E}\left[\widehat{f}(x)\right]-f(x)\right| &= \left|\mathbb{E}\left[\frac{1}{h}K\left(\frac{X_i-x}{h}\right)\right]-f(x)\right| = \left|\int_{x-h}^{x+h}\frac{1}{h}K\left(\frac{u-x}{h}\right)f(u)\,\mathrm{d}u - f(x)\right| \\ &= \left|\int_{-1}^{1}K(u)f(x+hu)\,\mathrm{d}u - f(x)\right| \\ &= \left|\int_{-1}^{1}K(u)\left(f(x)+\sum_{r=1}^{3}\frac{f^{(r)}(x)}{r!}(hu)^r+\frac{f^{(4)}(\xi)}{4!}(hu)^4\right)\mathrm{d}u - f(x)\right|\end{aligned}$$

for some ξ between x and $x+hu$. Recalling that K is a fourth order kernel which is bounded by $\frac{45}{32}$,

$$\left|\mathbb{E}\left[\widehat{f}(x)\right]-f(x)\right| = \left|\int_{-1}^{1}K(u)\frac{f^{(4)}(\xi)}{4!}(hu)^4\,\mathrm{d}u\right| \leq \frac{Mh^4}{24}\int_{-1}^{1}|K(u)|\,\mathrm{d}u \leq \frac{45Mh^4}{384}.$$

(d) Use the bias and variance upper bounds to derive an approximate optimal bandwidth which minimizes the mean squared error at each x. You may ignore constants which depend on M, giving your answer as a function of n only.

Solution: We have

$$\mathrm{MSE}\left[\widehat{f}(x)\right] = \mathbb{V}\left[\widehat{f}(x)\right] + \left[\mathbb{E}\left[\widehat{f}(x)\right]-f(x)\right]^2 \leq \frac{5M}{4nh} + \left(\frac{45Mh^4}{384}\right)^2 \lesssim \frac{1}{nh} + h^8.$$

These two terms can be balanced by solving

$$\frac{1}{nh} = h^8 \iff h^9 = \frac{1}{n} \iff h = n^{-1/9}.$$

(e) Using this approximate optimal bandwidth, provide an upper bound on the mean squared error of the kernel density estimator as a function of n only.

Solution: Using the approximate optimal bandwidth gives

$$\mathrm{MSE}\left[\widehat{f}(x)\right] \lesssim \frac{1}{nh} + h^8 \lesssim (n^{-1/9})^8 \lesssim n^{-8/9}.$$

(f) Compare the bias, variance, approximate optimal bandwidth and mean squared error upper bound derived above with those arising from an order 2 kernel, ignoring constants. What do you think would happen in the general case where $p \geq 2$ is even? Give any extra regularity assumptions you require.

Solution: Let $p \geq 2$ be the order of the kernel. The variance does not depend on p up to constants. The bias is on the order of h^p, so is smaller with a higher-order kernel, assuming the density is p times differentiable. Thus the mean squared error is $\frac{1}{nh} + h^{2p}$ giving an approximate optimal bandwidth of $h = n^{\frac{-1}{2p+1}}$ and a final MSE of order $n^{-\frac{2p}{2p+1}}$.

(g) Find u which minimizes $K(u)$ and report the associated value of $K(u)$. Propose a modified estimator $\widetilde{f}(x)$ based on $\widehat{f}(x)$ which satisfies:

i. $\widetilde{f}(x) \geq 0$ for all $x \in \mathbb{R}$ almost surely.

ii. $\mathbb{E}\left[\left(\widetilde{f}(x)-f(x)\right)^2\right] \leq \mathbb{E}\left[\left(\widehat{f}(x)-f(x)\right)^2\right]$ for all $x \in \mathbb{R}$.

Solution: Firstly, the kernel achieves its minima at $u = \pm\sqrt{5/7}$, giving $K(u) = -15/56$. Next take $\widetilde{f}(x) = \max\left\{\widehat{f}(x), 0\right\}$ which is clearly non-negative. If $\widehat{f}(x) \geq 0$ then $\left|\widetilde{f}(x) - f(x)\right| = \left|\widehat{f}(x) - f(x)\right|$, while if $\widehat{f}(x) < 0$ then $\widetilde{f}(x) = 0$ so

$$\left|\widehat{f}(x) - f(x)\right| = f(x) - \widehat{f}(x) = f(x) - \widetilde{f}(x) + \widetilde{f}(x) - \widehat{f}(x) = \left|\widetilde{f}(x) - f(x)\right| + \left|\widehat{f}(x)\right| \geq \left|\widetilde{f}(x) - f(x)\right|.$$

Thus we have $\left|\widetilde{f}(x) - f(x)\right| \leq \left|\widehat{f}(x) - f(x)\right|$ almost surely and the result about MSE follows.

2. (**Influence functions**) Determine the influence function $\varphi(x;\theta)$ for the variance estimator.

Solution: $\varphi(x) = (x - \mathbb{E}\left[X_i\right])^2 - \theta_0$ and

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = \frac{1}{n}\sum_{i=1}^{n}\left[(X_i - \mathbb{E}\left[X_i\right])^2 - \theta_0\right] + o_p(1).$$