STAT 5372 Assignment 4

(Lectures 14–16)

Mathematical concepts and derivations

1. If $\hat{f}(x)$ denotes the density estimate of $f(x)$, show that, for the integrated squared error (ISE),

$$L(f, \hat{f}) = \int \{\hat{f}(x) - f(x)\}^2 dx,$$

we have the following expression for the mean ISE (MISE):

$$\mathbb{E}L(f, \hat{f}) = \int b(x)^2 dx + \int v(x) dx,$$

where $b(x) = \mathbb{E}\hat{f}(x) - f(x)$ and $v(x) = \mathbb{V}\hat{f}(x)$ denote the bias and variance of $\hat{f}(x)$.

2. Consider the histogram density estimate based on a sample $x_i, \ldots, x_n$, discussed in Lecture 14, where for $x \in B_x = [\ell_x, \ell_x + h)$:

$$\hat{f}(x) = \frac{y_x}{nh}, \qquad \text{where } y_x \text{ is the number of the } x_i \text{ that fall in bin } B_x.$$

Recalling that

$$\mathbb{E}\hat{f}(x) = \frac{p_x}{h} = \frac{1}{h}\int_{\ell_x}^{\ell_x+h} f(y) dy,$$

it can be shown that the bias and variance of $\hat{f}(x)$ satisfy:

$$b(x) = \frac{h}{2}f'(x) - f'(x)(x - \ell_x) + O(h^2), \qquad v(x) = \frac{f(x)}{nh} + O\left(\frac{1}{n}\right).$$

These expressions follow by using the two-term Taylor series expansion of $f(y)$ about the point x: $f(y) = f(x) + f'(x)(y - x) + O(h^2)$.

(a) By ignoring the terms of $O(h^2)$ and smaller, use the stated Taylor series expansion of $f(y) \approx f(x) + f'(x)(y - x)$ in the calculation of p_x, to show that:

$$\mathbb{E}\hat{f}(x) = \frac{p_x}{h} \approx f(x) + \frac{f'(x)}{2}[h - 2(x - \ell_x)].$$

Deduce from this that: $b(x) \approx f'(x)h/2 - f'(x)(x - \ell_x)$.

(b) Using the above results, prove the Theorem on Slide 14, that provided $f(x)$ is continuous at x, then $\hat{f}(x)$ is a consistent estimate of $f(x)$ under the 2 conditions: (i) $h \to 0$, and (ii) $nh \to \infty$ as $n \to \infty$.

3. For a kernel density estimate $\hat{f}(x)$, we showed in Lecture 15 that the approximate MISE (AMISE) is given by:

$$\mathbb{E}\int \{\hat{f}(x) - f(x)\}^2 dx \approx \frac{1}{4}\sigma_K^4 h^4 R(f'') + \frac{R(K)}{nh} \equiv \text{AMISE}(h),$$

where we use the notation:

$$\sigma_K^2 = \int x^2 K(x) dx, \qquad \text{and for any function } g(x), \quad R(g) = \int g(x)^2 dx.$$

Note that the first term is an approximation to the integrated squared bias and the second an approximation to the integrated variance of $\hat{f}(x)$. More precisely, one can show that:

$$b(x) = \frac{\sigma_K^2}{2} f''(x) h^2 + o(h^2), \qquad v(x) = \frac{R(K)}{nh} f(x) + o\left(\frac{1}{nh}\right).$$

(a) Prove the Theorem on Slide 12 that, under the same conditions as for histograms in Question 2(b), $\hat{f}(x)$ converges in probability to $f(x)$.

(b) Verify the claim made in Slide 13, that based on AMISE(h) the optimal bandwidth is:

$$h_* = \left(\frac{R(K)}{n\sigma_K^4 R(f'')}\right)^{1/5},$$

and that AMISE$(h_*) = O(n^{-4/5})$, i.e., show that AMISE$(h_*)n^{4/5}$ is constant.

(c) How does the variance of $\hat{f}(x)$ change as a function of x?

(d) How does the variance of $\hat{f}(x)/f(x)$ change as a function of x?

(e) It is sometimes claimed that methods using an adaptive bandwidth (in which h changes as a function of x) correct for the tendency of fixed-bandwidth estimators to have high variance in regions with little data. Are such claims referring to the variance of the density itself, or the relative accuracy of the density?

Simulation

4. When we discussed the bootstrap, we noted that drawing random samples from $\hat{F}$, the empirical CDF, is equivalent to resampling the original data with replacement. Suppose that instead, we wish to draw random samples from $\hat{f}$, a kernel density estimate. Write a function called `rdensity` that produces random samples from an estimated density. Your function only needs to work for the case of Gaussian kernels, but it does need to work for arbitrary bandwidths. The function should take three arguments:

n: the desired number of draws from $\hat{f}$ (sample size),

d: a fitted density object as returned by the function, and

x: the original data (strictly speaking, you do not need the original data, but it is convenient to be able to access it).

(a) Explain the idea/principle behind your function, and submit the code.

(b) Apply your function to generate a simulated realization from your kernel density estimate in Question 5(a) below. Does your realization look similar to this estimate?

Application

5. The course website contains the dataset `nhanes.txt` that lists the triglyceride levels of 3,026 adult women.

   (a) Obtain a kernel density estimate for the distribution of triglyceride levels in adult women and plot it. You are free to decide on whatever kernel and bandwidth you like, but describe which ones you used.

   (b) Obtain a parametric density estimate assuming that triglyceride levels follow a normal distribution and overlay this density estimate with your estimate from (a).

6. Try to obtain a kernel density estimate for the dataset `nerve-pulse.txt` (waiting times between nerve pulses) on the course website, with bandwidth chosen by (unbiased) cross-validation. You will receive a warning message, and your estimate will appear to be clearly incorrect!

   (a) What's going on? What is causing this problem?

   (b) Fix the problem and obtain a reasonable-looking estimate of the density.

7. The course website contains the dataset `clouds.txt`: measurements of the rainfall from 26 "Unseeded" clouds (ignore the "Seeded" column). Most clouds gave off very little precipitation, so the density near 0 is high. Standard kernel density approaches produce an estimate of the density that is high near zero and — this is the problem — also below zero. Obviously, rainfall cannot be negative, so this estimate is unappealing. For (a)-(c) below, estimate the density according to the described approach, and plot the resulting estimate (you may overlay all your answers into a single plot or keep them separate; either way is fine).

   (a) Estimate the density using the standard approach, ignoring the boundary problem.

   (b) Estimate the density by taking a log transformation of the data, fitting the density on this scale, then transforming the estimated density back to the original scale. (Note: specify exactly the equation that you used to transform back!)

   (c) Estimate the density by taking the standard approach and "reflecting" the estimated density that lies in $(-\infty, 0)$ about 0.