Gangopadhyay

MA750
Project I

(Due Tuesday October 24)

This project is for your individual effort and no collaboration is allowed. You may not seek help from anyone or provide help to anyone. You can contact me if you need clarification on any problem. Some of the problems are open ended, provide as complete a solution as you possibly can. Include your R codes with all problems that involve computation.

1. (40) *Variable Bandwidth Kernel Estimator*

For the last few weeks, we have been discussing the kernel density estimator that utilizes constant amount of smoothing for the whole range of univariate or multivariate data. Thus the estimator does not adopt to local variation in smoothness. It is possible to generalize the kernel estimator that incorporates varying bandwidth allowing less smoothing where the underlying density has more structure, and more smoothing where the density has less structure. There are many ways to accomplish this goal. Here we will consider one useful procedure.

The general formula of one such estimator is given by:

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h(x_i)} K(\frac{x - x_i}{h(x_i)})$$

Since the goal is to smooth less where there is more structure (and more where there is less structure), it is natural to have $h(x_i)$ to vary inversely with the underlying density. Consider the following specific form: $h(x_i) = h \times f(x_i)^{-1/2}$,i.e.,

$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} f(x_i)^{1/2} K(\frac{(x - x_i)f(x_i)^{1/2}}{h})$$

This choice is particularly advantageous as (under certain regularity condition) it can be shown that the estimator has bias $O(h^4)$, rather than the usual $O\left(h^2\right)$, while the variance remains the same $O(\frac{1}{nh})$. Thus the optimal order of $h = O(n^{-1/9})$ leads to the improved rate of convergence of MSE $= O(n^{-8/9})$.

Consider the following proposal for implementating the procedure:

a. Obtain a preliminary estimate of $f(x_i)$,say $\widetilde{f}(x_i)$. The usual choice will be an initial fixed kernel estimator with some optimal choice of bandwidth g.

b. Estimate f as:

$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} \widetilde{f}(x_i)^{1/2} K(\frac{(x - x_i)\widetilde{f}(x_i)^{1/2}}{h})$$

Implement the procedure in R. Also, generate 100 observations from "Claw" density. Use the method to estimate the density based on the sample data. Try different bandwidth h and choose one that looks reasonable. Compare your result against the usual fixed kernel estimator. Try a few different initial estimators $\widetilde{f}$ (different kernels and bandwidths) and comment on its effect on the final estimator $\widehat{f}$.

Note on Claw density: It is a mixture of five normal distributions given by $\frac{1}{2}N(0,1)+\sum_{m=0}^{4}\frac{1}{10}N(\frac{m}{2}-1,\frac{1}{10})$ that creates a challenging density estimation problem. To see the form of the pdf, download the R package "ftnonpar" and use the dclaw command.

plot(dclaw(seq(-3,3,len=1000)),type="l")

Use "rclaw" command to generate samples.

c. Assuming $\widetilde{f}(x_i)^{1/2}$ is fixed, propose an optimal bandwidth selection method for h. Implement your method in R and find the optimal bandwidth for the sample from Claw density in part (b).

2. (30) In the class, we discussed the boundary problem associated with estimating density f. It turns out that the boundary bias gets worse if the objective is to estimate derivatives of the density f. Consider a density $f(x)$ with support on the interval $[0,\infty)$ and let the kernel estimator of the first derivative $f'(x)$ is given by

$$\widehat{f}'(x)=\frac{1}{nh^2}\sum_{i=1}^{n}K'\left(\frac{x-X_i}{h}\right)$$

were kernel K is a symmetric pdf with bounded support on the interval $[-1,1]$.

a. Show that if x is a point on the boundary (as discussed in the class, $x=ph, 0\le p<1$)

$$E\left(\widehat{f}'(x)\right)=\frac{1}{h}a_0^1(p)f(x)-a_1^1(p)f'(x)+\frac{h}{2}a_2^1(p)f''(x)+O(h^2)$$

where $a_j^1(p)=\int_{-1}^{p}u^jK'(u)du$.

b. Think about how in the lectures we resolved the issue of boundary bias for estimating f by coming up with a boundary kernel based on two related kernels K and L. Can you come up with a similar idea to correct for the boundary bias for estimating $f'(x)$? It would be very nice if you can come up with a formula for the boundary kernel (algebra can be a bit tedious), but at least outline your approach clearly and comprehensively.

3. (30) There are other possible approaches of developing kernel estimators that allow local smoothing. Review the following paper and write a short (2-3 pages) summary of the basic idea.

Gangopadhyay and Cheung (2002). Bayesian approach to the choice of smoothing parameter in kernel density estimation. *Journal of Nonparametric Statistics*, page 655-664.