Homework 5 Stat 550

Dr. Scott

November 9, 2023

Due: Friday 11-17
30 Points

Instructions: We will discuss generally in class before due date. You can work in groups, but turn in your own solutions. Due date flexible.
Where to look: `http://www.stat.rice.edu/~scottdw/stat550/HW/`.
Where to turn in: Canvas upload.

1. Problem 6-25 displays a kernel

$$K(t) \propto \exp(-1/(1-t^2)) \qquad -1 < t < 1\,,$$

that has finite support but is infinitely differentiable on the real line.

   a. Use `NIntegrate` in Mathematica to find the correct normalizing constant; then (numerically) compute the variance of this kernel. (The constant given in the book is not correct.)

   b. For the geyser dataset (use `dget('geyser.dat')` to input), plot the default Gaussian KDE using the R function `density`; record the bandwidth given in the x-axis label. Using the bandwidth transformation formula in equation (6.28), overlay the kernel estimate using the new kernel here. Comments?

2. Enter the Swiss Franc data into R. (The `read.table` command is in the file.) The data are 6 measurments on the old 1000 franc bills, 100 genuine and 100 counterfeit. We will use only the 5th and 6th columns for this problem. (Use `as.double(x[,5]) + runif(200,-.05,.05)` to transform into a blurred double vector.)

   a. Load the `ash` library in R to compute the bivariate pdf of the data. Plot the blurred data points and use the `contour(...)`

function with the option `add=T` to superimpose the contours onto the scatterplot. *Hints:* Let the limits of the graphing area be given by `xl=c(7.5,12.75)`, `yl=c(137.5,143.25)`, and choose `kopt=c(2,3)`, which gives the triweight kernel. Use `xlim=xl` and `ylim=yl` when plotting the data so the contours will fit. Let `ab=cbind(xl,yl)` and `nbin=c(200,200)` as inputs to the function `bin2`, which is input to the function `ash2`. Play around with the smoothing parameters in function `ash2`.

b. For this question, repeat part (a) but compute the contours for the genuine (1-100) and counterfeit (101-200) bills separately. Add to the scatter plot.

3. Problem 7-2, with the obvious typo fixed. *Hints:* Write out the integral as the sum of 2 integrals with domains $z < -u$ and $z > -u$. Then add and subtract another integral so the $EZ = 0$ can be computed. Finally, recall $\Phi(-u) = 1 - \Phi(u)$. This useful identity can be used to show the optimal pointwise L_1 and L_2 bandwidths are nearly equal.

Possible research topics for the doctoral students: RKHS, t-SNE/UMAP, Bayesian KDE, NN regression, SOM (self-organizing map), NN density estimation, survival analysis, mixture estimation, adaptive kde.