Homework 3

STAT 547, Fall 2019

You are encouraged to discuss the homework questions with classmates or the instructor, but you must write and submit your individual copy. Please write down the name of the persons with whom you discussed the homework, and submit your homework in a pdf file through Canvas.

1. Let $\{W(t),\, 0 \le t \le 1\}$ be a Brownian motion (Wiener process). The Brownian bridge $\{B(t),\, 0 \le t \le 1\}$ is the Brownian motion conditioned on $W(1) = 0$ and can be represented as $B(t) = W(t) - tW(1)$. Derive the Karhunen–Loève representation of the Brownian bridge $B(t)$.

2. Simulate a sample of 50 realizations of a). the Brownian motion and b). the Brownian bridge. Each curve should have 1000 support points. Show the trajectories on two separate plots and include your R code.

3. Let $X_1, \ldots, X_n$ be a sample of i.i.d. real-valued random variables sharing a distribution with an unknown density f supported on a compact interval $[a, b]$. The kernel density estimate (KDE) of $f(x_0)$ at $x_0 \in [a, b]$ is

   $$\hat{f}(x_0) = \frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x_0),$$

   where $K_h(\cdot) = h^{-1}K(\cdot/h)$, $K(\cdot)$ is a kernel function, $h > 0$ is the bandwidth. Write down an intuitive argument for why the KDE "works". [Hint: consider a uniform kernel K]

4. Analyze the Lake Acidity data in the `gss` package of R. The data were extracted from the Eastern Lake Survey of 1984 conducted by the United States Environmental Protection Agency, concerning 112 lakes in the Blue Ridge. To gain access to the data, type the following commends in R:

   ```
   library(gss)
   data(LakeAcidity)
   ```

   For more information check the help document about this data set.

5. Perform a nonparametric regression on the calcium concentration (Y) against surface ph level (X).

(a) Show a KDE and a dot plot of the ph levels.

(b) Compare the results of local polynomial estimator, smoothing spline, regression spline and penalized spline. Manually vary the tuning parameters, including bandwidth, the number of knots, and the penalty λ on the second derivative of the regression curve. For each smoother identify a parameter setting that i). oversmooths (the estimate is too smooth), ii) undersmooths (the estimate is too rough), and iii) smooths appropriately. Show the graphs and your code.

(c) Write a brief summary of your data analysis.