STA 621: Nonparametric Statistics
Breheny

Assignment 1
Due: Tuesday, September 11

Mathematical concepts and derivations

1. Show that, out of all possible CDFs, the empirical CDF $\hat{F}$ maximizes

$$L(F|\mathbf{x}) = \prod_{i=1}^{n} \mathbb{P}_F(x_i).$$

2. Consider the quantile functional $T(F) = F^{-1}(p)$, and let θ denote the true value of $T(F)$. Suppose that F is continuous at θ with positive density $f(\theta)$. Show that, for $x > \theta$,

$$L(x) = \frac{p}{f(\theta)}.$$

Hint: Note that if $G(x) = aF(x) + b$, then $G^{-1}(y) = F^{-1}(\frac{y-b}{a})$. You may also need to brush up on the inverse function theorem, if you have forgotten it.

3. Let $b(\epsilon) = \sup_x |T(F) - T(F_\epsilon)|$, where $F_\epsilon = (1-\epsilon)F + \epsilon\delta_x$. The *breakdown point* of an estimator, ϵ^*, is defined as $\epsilon^* = \inf\{\epsilon : b(\epsilon) = \infty\}$.
   (a) Find the breakdown point of the mean.
   (b) Find the breakdown point of the median.

4. Consider a random variable X that is always positive. Suppose we are interested in the statistical functional $\theta = \int \log(x) dF(x)$.
   (a) What is the plug-in estimator of θ?
   (b) What are the influence and empirical influence functions for θ?
   (c) Suppose instead that we are interested in $\lambda = \log(\mu)$, where $\mu = \mathbb{E}(X)$. What is the plug-in estimator of λ?
   (d) What are the influence and empirical influence functions for λ?
   (e) Derive an asymptotic $1 - \alpha$ nonparametric confidence interval for $\hat{\lambda}$.
   (f) Do $\hat{\lambda}$ and $\hat{\theta}$ converge to the same number?
   (g) Plot the empirical influence functions from parts (b) and (d). Label the point x on the horizontal axis where $L(x) = 0$.
   (h) Briefly, comment on the relative robustness of $\hat{\theta}$ and $\hat{\lambda}$ to outliers.

5. Suppose that there exists a constant C such that the following relation holds for all G:

$$|T(F) - T(G)| \leq C \sup_x |F(x) - G(x)|.$$

Show that $T(\hat{F}) \xrightarrow{\text{a.s.}} T(F)$.

Simulation

6. Generate $X_1, X_2, \ldots, X_{100}$ independent observations and compute a 95 percent global confidence band for the CDF F based on the DKW inequality. Repeat this 1000 times and report the proportion of data sets for which the confidence band contained the true distribution function.

   (a) Carry out the above simulation with data coming from the standard normal $N(0, 1)$ distribution.

   (b) Repeat using data generated from the standard Cauchy distribution.

7. Compare the nonparametric confidence interval for the variance obtained from using the functional delta method to the normal-theory interval:

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}\right],$$

   where s^2 is the (unbiased) sample variance.

   Conduct a simulation study to determine the coverage probability and average interval width of these two intervals.

   (a) Carry out the above simulation with data generated from the standard normal distribution.

   (b) Repeat using data generated from an exponential distribution with rate 1.

   (c) Briefly, comment on the strengths and weaknesses of these two methods.

Application

8. The `R` data set `quakes` contains (among other information) the magnitude of 1,000 earthquakes that have occurred near the island Fiji.

   (a) Estimate the CDF for the magnitude of earthquakes in this region, along with a 95% confidence interval. Plot your results.

   (b) Estimate and provide a 95% confidence interval for $F(4.9) - F(4.3)$.

   (c) Estimate the variance of the magnitude, and provide a nonparametric 95% confidence interval for its value.