Stat 274/374: Nonparametric Inference

Assignment 3

Due: Thursday, November 15, 2018

1. *Chicago Crime* (20 points)

   Chicago crime data is available at https://data.cityofchicago.org/Public-Safety/Crimes-Map/dfnk-7re6. The dataset contains information on the reported incidents of crime that have occurred in the City of Chicago over the past year. You can download a copy of this dataset as a CSV file by clicking the 'Export' icon on the top right of the web page. In this dataset, columns `LATITUDE` and `LONGITUDE` represent the location of reported crime cases.

   (a) Estimate a one-dimensional density for each of the two continuous variables `LATITUDE` and `LONGITUDE` using a kernel density estimator. Note that you will need to take care of the entries with missing value. Choose the amount of smoothing using cross validation and make a plot of cross-validation scores versus bandwidths. For each variable, plot your density estimate with the optimal bandwidth.

   (b) Estimate a two-dimensional density for the two variables using a kernel density estimator. Choose the amount of smoothing using cross validation or the Normal reference rule. Use a contour or a heatmap to visualize your estimate. Overlap your plot on a map of Chicago, if possible.

2. *Normal Means and Penalization* (20 points)

   Consider a normal means model $X_i = \theta_i + \sigma_n \epsilon_i$ where $\epsilon_i \sim N(0,1)$, for $i = 1, 2, \ldots, n$. Let $\widehat{\theta}_\lambda$ be the *ridge-regression estimator*, defined as

$$\widehat{\theta}_\lambda = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (X_i - \beta_i)^2 + \lambda \|\beta\|^2 \right\},$$

   where $\|\beta\|^2 = \sum_{i=1}^{n} \beta_i^2$.

   (a) Derive the bias-variance decomposition of the risk of $\widehat{\theta}_\lambda$.

   (b) What value of λ minimizes the risk?

   (c) Derive the optimal value λ^* using SURE.

   (d) Now let $\sigma_n^2 = \sigma^2/n$. Is the estimator $\widehat{\theta}_{\lambda^*}$ asymptotically minimax over the ball $\Theta(c) = \{\theta : \|\theta\|^2 \leq c^2\}$? Explain.

3. *Risk, Baseball and Cars* (20 points)

(a) Suppose that $Z_i \sim N(\theta_i, \sigma^2)$ for $i = 1, \ldots, n$ are all independent. Define the risk of an estimator $\widehat{\theta}$ of θ to be

$$R(\widehat{\theta}, \theta) = \mathbb{E}_\theta \sum_{i=1}^{n} (\widehat{\theta}_i - \theta_i)^2.$$

Suppose that $n \geq 3$. The James-Stein estimator

$$\widehat{\theta}^{JS} = \left(1 - \frac{(n-2)\sigma^2}{\sum_{i=1}^{n} Z_i^2}\right) Z$$

can be viewed as shrinking the data toward 0. For any $v \in \mathbb{R}^n$, we can define

$$\widehat{\theta}^{JS,v} = v + \left(1 - \frac{(n-2)\sigma^2}{\sum_{i=1}^{n} (Z_i - v_i)^2}\right)(Z - v)$$

as a generalized James-Stein estimator that shrinks the data towards v. Show that, for any v and θ,

$$R(\widehat{\theta}^{JS,v}, \theta) < n\sigma^2.$$

What choice of v leads to the smallest risk?

(b) An estimator $\widehat{\theta}$ is said to be *admissible* if there is no other estimator $\widetilde{\theta}$ such that

$$R(\widetilde{\theta}, \theta) \leq R(\widehat{\theta}, \theta) \text{ for all } \theta$$

with strict inequality at at least one θ.

When $n = 1$, we can show that $\widehat{\theta} = Z$ is admissible. (*Take this as a matter of fact, which you don't need to show.*) With $n = 1$, consider the linear estimator $\widehat{\theta}^{a,b} = aZ + b$. For each of the following conditions, decide if the estimator $\widehat{\theta}^{a,b}$ is admissible: (i) $a = 0$, (ii) $a < 0$, (iii) $a > 1$, (iv) $a = 1$ and $b \neq 0$.

When $n \geq 3$, is the estimator $\widehat{\theta} = Z$ admissible? Is the estimator $\widehat{\theta}^{JS,v}$ admissible?

(c) In Table 1, we have the data of 18 Major League Baseball players' batting average during the 1970 season. Column 2 shows the batting averages through their first 45 official at bats and column 3 the batting averages over the remainder of the season. The problem is to predict each player's batting average in column 3 using only the data from column 2. Let Y_i be the batting average of player i, $i = 1, \ldots, 18$ $(n = 18)$ after $m = 45$ at bats. Perform the following transform

$$Z_i \triangleq f_m(Y_i) \triangleq \sqrt{m} \arcsin(2Y_i - 1)$$

so that Z_i is approximately normal with nearly unit variance and mean $\theta_i = f_m(p_i)$. Now estimate θ_i's using the MLE, the generalized James-Stein estimator with $v_i = 0$ and $v_i = \overline{Z}$ respectively. Compute the mean squared errors and compare.

Table 1: 1970 batting averages for 18 Major League players

i	Player	Y_i=AVG for 1st 45	p_i=AVG for rmd season
1	Clemente (Pitts)	.400	.346
2	F. Robinson (Batt)	.378	.298
3	F. Howard (Wash)	.356	.276
4	Johnstone (Cal)	.333	.222
5	Berry (Chi)	.311	.273
6	Spencer (Cal)	.311	.270
7	Kessinger (Chi)	.289	.263
8	L. Alvarado (Bos)	.267	.210
9	Santo (Chi)	.244	.269
10	Swoboda (NY)	.244	.230
11	Unser (Wash)	.222	.264
12	Williams (Chi)	.222	.256
13	Scott (Bos)	.222	.303
14	Petrocelli (Bos)	.222	.264
15	E. Rodriguez (KC)	.222	.226
16	Campaneris (Oak)	.200	.285
17	Munson (NY)	.178	.316
18	Alvis (Mil)	.156	.200

(d) Now suppose that we are also interested in the proportion of imported automobiles in Chicago. Of the first 45 samples recorded we find 9 to be foreign-made. We can either estimate the true proportion using the average 9/45, or combine this estimation problem with the 18 players' batting averages and use the James-Stein estimator. Which do you think gives a smaller risk? Comment.

4. *James-Stein Estimator* (20 points)

   Let $\theta_i = 1/i^2$ for $i = 1, \ldots, n$. Take $n = 1000$. Let $Z_i \sim N(\theta_i, 1)$ for $i = 1, \ldots, n$. Compute the risk of the maximum likelihood estimator (MLE) $\widehat{\theta}_i = Z_i$. Compute the risk of the estimator $\widetilde{\theta} = (bZ_1, \ldots, bZ_n)$. Plot this risk as a function of b. Find the optimal value b_*. Now conduct a simulation. For each run of the simulation, find the (modified) James-Stein estimator $\widehat{b}Z$ where

   $$\widehat{b} = \left[1 - \frac{n}{\sum_i Z_i^2}\right]_+$$

   where $[x]_+ = \max(0, x)$. You will get one $\widehat{b}$ for each simulation. Compare the simulated values of $\widehat{b}$ to b_*. Also, compare the risk of the MLE and the James-Stein estimator (the latter obtained by simulation) to the Pinsker bound $\sigma^2 c^2/(\sigma^2 + c^2)$. Explain your findings.

5. *Justifying Your Means* (20 points)

   The file `assn3-prob5-data.txt` contains values $X_i \sim N(\theta_i, \sigma^2)$ for $i = 1, \ldots, 1000$. Estimate θ, using any method of your choice. In your problem solutions, give a detailed description of your estimator. Send your estimates $\widehat{\theta}_i$ (for $i = 1, \ldots, 1000$) by email to `nonparametric16fall@gmail.com`, in the form of an attached text file with name `assn3-prob5-<your_cnet_id>.txt`. Use the same format as the data file (one value per line). We will compute the risk of the estimates, and the points you receive for the problem will depend on your relative ranking in the class.