Assignment #2 STA355H1S

due Friday February 14, 2020

Instructions: Solutions to problems 1 and 2 are to be submitted on Quercus (PDF files only) — the deadline is 11:59pm on February 14. You are strongly encouraged to do problems 3 through 6 but these are **not** to be submitted for grading.

1. The Hidalgo stamp data is a (semi-)famous dataset containing thicknesses of 482 postage stamps from the 1872 Mexican "Hidalgo" issue. It is believed that these stamps were printed on different types of papers so that the data can be modeled as a "mixture" of several distributions with the density having between 5 and 7 modes. These data (which have been "jittered" by adding noise) are available on Quercus in a file `stamp.txt`.

(a) Use the `density` function in R to estimate the density. Choose a variety of bandwidths (the parameter `bw`) and describe how the estimates change as the bandwidth changes. How small does the bandwidth need to be for the density estimate to have 5 modes? 7 modes?

(b) One automated approach to selecting the bandwidth parameter h is **leave-one-out cross-validation**. This is a fairly general procedure that is useful for selecting tuning parameters in a variety of statistical problems.

If f and g are density functions, then we can define the Kullback-Leibler divergence

$$D_{KL}(f\|g) = \int_{-\infty}^{\infty} f(x) \ln\left(\frac{f(x)}{g(x)}\right)\, dx.$$

For a given density f, $D_{KL}(f\|g)$ is minimized over densities g when $g = f$ (and $D_{KL}(f\|f) = 0$). In the context of bandwith selection, define $\widehat{f}_h(x)$ to be a density estimator with bandwidth h and $f(x)$ to be the true (but unknown) density that produces the data. Ideally, we would like to minimize $D_{KL}(f\|f_h)$ with respect to h but since f is unknown, the best we can do is to minimize an estimate of $D_{KL}(f\|f_h)$. Noting that

$$\begin{aligned} D_{KL}(f\|f_h) &= -\int_{-\infty}^{\infty} \ln(f_h(x)) f(x)\, dx + \int_{-\infty}^{\infty} \ln(f(x)) f(x)\, dx \\ &= -E_f[\ln(f_h(X))] + \text{constant}, \end{aligned}$$

this suggests that we should try to maximize an estimate of $E_f[\ln(f_h(X))]$, which can be estimated for a given h by the following (leave-one-out) substitution principle estimator:

$$\text{CV}(h) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(\frac{1}{(n-1)h}\sum_{j\neq i} w\left(\frac{X_i - X_j}{h}\right)\right) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(\widehat{f}_h^{(-i)}(X_i)\right)$$

where

$$\widehat{f}_h^{(-i)}(x) = \frac{1}{(n-1)h}\sum_{j\neq i} w\left(\frac{x - X_j}{h}\right)$$

is the density estimate with bandwidth h using all the observations except X_i.

Now suppose we replaced $\widehat{f}_h^{(-i)}(X_i)$ by $\widehat{f}_h(X_i)$ in the formula for CV(h) and maximized

$$\mathcal{L}(h) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(\frac{1}{nh}\sum_{j=1}^{n} w\left(\frac{X_i - X_j}{h}\right)\right) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(\widehat{f}_h(X_i)\right).$$

Maximizing $\mathcal{L}(h)$ or CV(h) over $h > 0$ has the flavour of maximum likelihood estimation. However, selecting the bandwidth by maximizing $\mathcal{L}(h)$ does not work while maximizing CV(h) typically (but not always!) produces a good result.

For (i) and (ii) below, assume that $w(0) > 0$ and for $x \neq 0$, $h^{-1}w(x/h) \to 0$ as $h \downarrow 0$ (which is true for most commonly used kernels).

(i) Show that $\mathcal{L}(h) \uparrow \infty$ as $h \downarrow 0$.

(ii) In the case where $X_1, \cdots, X_n$ are distinct (that is, no tied observations), show that CV(h) $\to -\infty$ as $h \downarrow 0$ and $h \uparrow \infty$.

(c) On Quercus, there is a function `kde.cv` (in a file `kde.cv.txt`) that computes CV(h) for various bandwidth parameters h for the Gaussian kernel. This function has two arguments, the data `x` and a vector of bandwidth parameters `h`; the default for `h` is a vector of 101 bandwidths ranging from $h^*/10$ to $4h^*$ where h^* is the default bandwidth in the R function `density`. The function `kde.cv` can be used as follows:

```
> r <- kde.cv(x)
> plot(r$bw,r$cv) # plot of bandwidth versus CV
> r$bw[r$cv==max(r$cv)] # bandwidth maximizing CV
```

Use this function to estimate the density of the Hidalgo stamp data. How many modes does this density have?

2. Suppose that F is a distribution concentrated on the positive real line (i.e. $F(x) = 0$ for $x < 0$). If $\mu(F) = E_F(X)$ then the **mean population share** of the distribution F is defined as MPS(F) $= F(\mu(F)-) = P_F(X < \mu(F))$. (When F is a continuous distribution, $F(\mu(F)-) = F(\mu(F))$.) For most income distributions, MPS(F) $> 1/2$ with MPS(F) $= 0$ if (and only if) all incomes are equal and MPS(F) $\to 1$ as Gini(F) $\to 1$. Associated with MPS(F) is the **mean income share** MIS(F) $= \mathcal{L}_F(\text{MPS}(F))$ where $\mathcal{L}_F(t)$ is the Lorenz curve

$$\mathcal{L}_F(t) = \frac{1}{\mu(F)}\int_0^t F^{-1}(s)\,ds \quad \text{with} \quad \mu(F) = \int_0^1 F^{-1}(s)\,ds.$$

(a) Suppose that F is a continuous distribution function with Lorenz curve $\mathcal{L}_F(t)$. Show that MPS(F) satisfies the condition

$$\mathcal{L}'_F(\text{MPS}(F)) = 1$$

where $\mathcal{L}'_F(t)$ is the derivative (with respect to t) of the Lorenz curve.

(b) Suppose that $\mathcal{L}_F(t) = t^{\alpha+1}$ for some $\alpha \geq 0$. Find MPS(F) and MIS(F).

(c) Given observations $X_1, \cdots, X_n$ from a distribution F, a substitution principle estimate of MPS(F) is

$$\widehat{\text{MPS}}(F) = \frac{1}{n}\sum_{i=1}^{n} I(X_i < \bar{X})$$

A sample of 200 incomes is given on Quercus in a file `incomes.txt`. Using these data compute an estimate of MPS(F) and use the jackknife to give an estimate of its standard error.

(d) Derive a substitution principle estimator for MIS(F). Using the data in part (c), compute an estimate of MIS(F) and use the jackknife to estimate its standard error. (Hint: Use the following estimator of the Lorenz curve:

$$\widehat{\mathcal{L}}_F(t) = \frac{1}{n\bar{X}}\sum_{i=1}^{\lceil nt \rceil} X_{(i)}$$

where $\lceil x \rceil$ is the smallest integer greater than or equal to x.)

Supplemental problems (not to be handed in):

3. Suppose that $X_1, \cdots, X_n$ are independent random variables with common density $f(x-\theta)$ where f is symmetric around 0 (i.e. $f(x) = f(-x)$) and θ is an unknown location parameter. If $\text{Var}(X_i)$ is finite then the sample mean $\bar{X}$ will be a reasonable estimator of μ; however, if f has heavy tails then $\bar{X}$ will be less efficient than other estimators of θ, for example, the sample median.

An useful alternative to the sample mean is the α-trimmed mean, which trims the smallest α and largest α fractions of the data and averages the middle order statistics. Specifically, if we define $r = \lfloor n\alpha \rfloor$ (where $\lfloor x \rfloor$ is the integer part of x) then the α-trimmed mean, $\widehat{\theta}(\alpha)$, is defined by

$$\widehat{\theta}(\alpha) = \frac{1}{n-2r}\sum_{k=r+1}^{n-r} X_{(k)}.$$

(a) Suppose (for simplicity) that $\lfloor n\alpha \rfloor = \lfloor (n-1)\alpha \rfloor$ and define $\widehat{\theta}_{-i}(\alpha)$ to be α-trimmed mean with $X_{(i)}$ deleted from the sample. Find expressions for $\widehat{\theta}_{-1}(\alpha), \cdots, \widehat{\theta}_{-n}(\alpha)$; in particular, note that

$$\widehat{\theta}_{-1}(\alpha) = \cdots = \widehat{\theta}_{-r}(\alpha) \quad \text{and} \quad \widehat{\theta}_{-(n-r+1)}(\alpha) = \cdots = \widehat{\theta}_{-n}(\alpha)$$

(b) Using the setup in part (a), show that the pseudo-values $\{\Phi_i\}$ are given by

$$\begin{aligned}
\Phi_i &= \frac{n-1}{n-1-2r}X_{(r+1)} - \frac{2r}{(n-2r)(n-1-2r)}\sum_{k=r+1}^{n-r} X_{(k)} \quad \text{for } i = 1, \cdots, r+1 \\
\Phi_i &= \frac{n-1}{n-1-2r}X_{(i)} - \frac{2r}{(n-2r)(n-1-2r)}\sum_{k=r+1}^{n-r} X_{(k)} \quad \text{for } i = r+2, \cdots, n-r-1 \\
\Phi_i &= \frac{n-1}{n-1-2r}X_{(n-r)} - \frac{2r}{(n-2r)(n-1-2r)}\sum_{k=r+1}^{n-r} X_{(k)} \quad \text{for } i = n-r, \cdots, n
\end{aligned}$$

and give a formula for the jackknife estimator of variance of $\widehat{\theta}(\alpha)$. (Think about how you might use this variance estimator to choose an "optimal" value of r.)

4. Suppose that $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are unbiased estimators of a parameter θ and consider estimators of the form

$$\widetilde{\theta} = a\widehat{\theta}_1 + (1-a)\widehat{\theta}_2.$$

(a) Show that $\widetilde{\theta}$ is unbiased for any a.

(b) Find the value of a that minimizes $\text{Var}(\widetilde{\theta})$ in terms of $\text{Var}(\widehat{\theta}_1)$, $\text{Var}(\widehat{\theta}_2)$, and $\text{Cov}(\widehat{\theta}_1, \widehat{\theta}_2)$. Under what conditions would $a = 1$? Can a be greater than 1 or less than 0?

5. A histogram is a very simple example of a density estimator. For a sample $X_1, \cdots, X_n$ from a continuous distribution with density $f(x)$, we define breakpoints $a_0, \cdots, a_k$ satisfying

$$a_0 < \min(X_1, \cdots, X_n) < a_1 < a_2 < \cdots < a_{k-1} < \max(X_1, \cdots, X_n) < a_k$$

and define for $x \in [a_{j-1}, a_j)$:

$$\widehat{f}(x) = \frac{1}{n(a_j - a_{j-1})}\sum_{i=1}^{n} I(a_{j-1} \le X_i < a_j)$$

with $\widehat{f}(x) = 0$ for $x < a_0$ and $x \ge a_k$.

(a) Show that $\widehat{f}$ is a density function.

(b) For a given value of x, evaluate the mean and variance of $\widehat{f}(x)$.

(c) What conditions on $a_0, \cdots, a_k$ are needed for the bias and variance of $\widehat{f}(x)$ to go to 0 as $n \to \infty$?

6. Another measure of inequality based on the Lorenz curve is the **Pietra index** defined by

$$\mathcal{P}(F) = \max_{0 \le t \le 1} \{t - \mathcal{L}_F(t)\}$$

where $\mathcal{L}_F(t)$ is the Lorenz curve.

(a) Show that $g(t) = t - \mathcal{L}_F(t)$ is maximized at t satisfying $F^{-1}(t) = \mu(F)$.

(b) Using the result of part (a), show that

$$\mathcal{P}(F) = \frac{E_F[|X - \mu(F)|]}{2\mu(F)}.$$

(You may assume that F has a density f.)

(c) Give a substitution principle estimator for the Pietra index $\mathcal{P}(F)$ based on the empirical distribution function of $X_1, \cdots, X_n$. Using the data in Problem 2, compute an estimate of $\mathcal{P}(F)$ and use the jackknife to compute an estimate of its standard error.

(d) Suppose that you can assume that the incomes follow a log-normal distribution, that is, $\ln(X_1), \cdots, \ln(X_n)$ are independent $\mathcal{N}(\mu, \sigma^2)$ random variables. Show that the Pietra index for the log-normal distribution is given by

$$\mathcal{P}(F) = 2\Phi(\sigma^2/2) - 1$$

where $\Phi(x)$ is the cdf of a $\mathcal{N}(0, 1)$ distribution.