Assignment #2 STA355H1S

student number: 1003942326

Yulin WANG

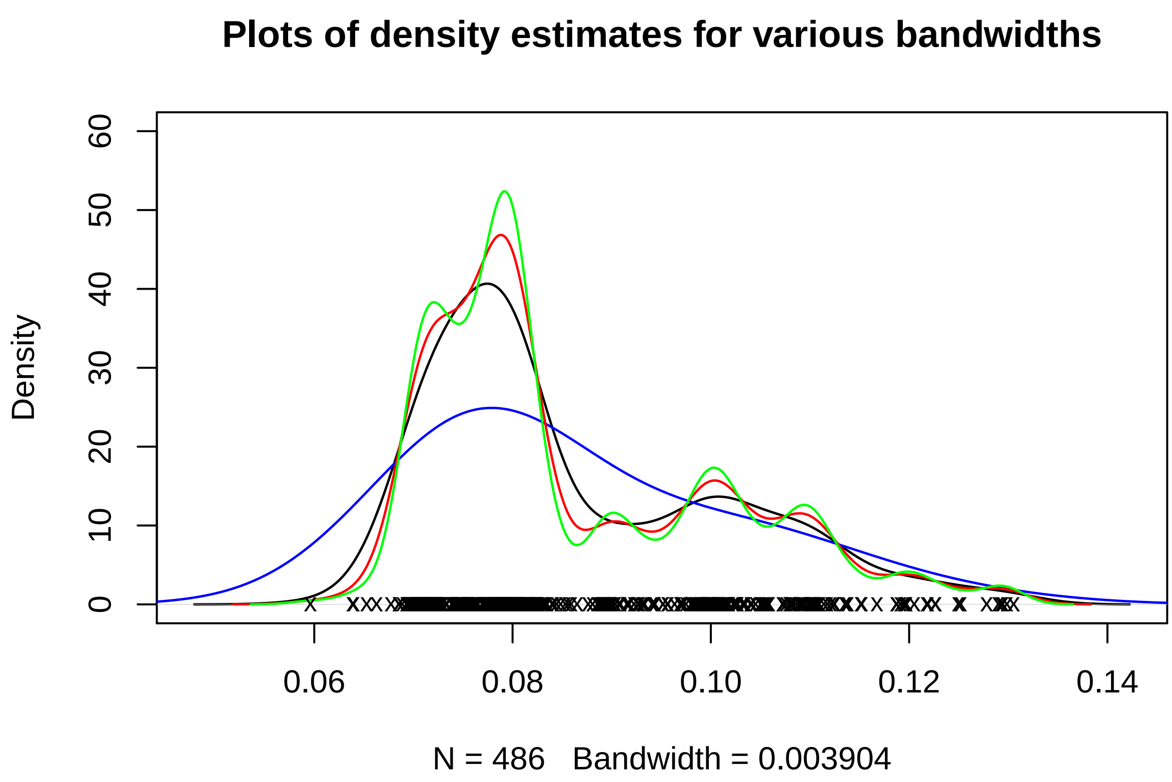
2020-02-13

Question1

(a)

Plots of density estimates for bandwidths 0.003904 (the default value for parameter bw), 0.01, 0.0026, 0.002 are shown below with black, blue, red and green lines respectively.

```r
stamp <- scan("stamp.txt") # load data stamp.txt
plot(density(stamp), lwd=1.2, ylim = c(0, 60), # default bandwidth 0.003904 ->two modes
     main = "Plots of density estimates for various bandwidths")
# try different values of bandwidths(the parameter bw)
lines(density(stamp,bw=0.01), col="blue", lwd=1.2) # 1 mode
lines(density(stamp,bw=0.0026), col="red", lwd=1.2) # 5 modes
lines(density(stamp,bw=0.002), col="green", lwd=1.2) # 7 modes
points(stamp, rep(0,486), pch="x", cex=0.8) # draw points on x-axis
```



- Thus, the local modes become somewhat more evident as the bandwidth decreases, that is, the smaller bandwidth, the more local modes.

- When bandwidth is about 0.0026, the density estimate has 5 modes.
  When bandwidth is about 0.002, the density estimate has 7 modes.

(b)(i)

Since we don't know whether $X_1, ..., X_n$ are distinct or not, so we need to divide this question in to two parts.
For the first part, $X_1, ..., X_n$ are distinct, that is $X_i \neq X_j$ for different $i, j = 1, ...n$, then $X_i - X_j \neq 0$
Since for $X_i - X_j \neq 0$,

$$h^{-1}w(\frac{X_i - X_j}{h}) \to 0 \quad \text{ as } h \downarrow 0$$

So

$$\frac{1}{h}\sum_{j=1}^{n} w(\frac{X_i - X_j}{h}) \to 0$$

Then

$$ln\{\frac{1}{nh}\sum_{j=1}^{n} w(\frac{X_i - X_j}{h})\} \to -\infty$$

Thus

$$\mathcal{L}(h) = \frac{1}{n}\sum_{i=1}^{n} ln\{\frac{1}{nh}\sum_{j=1}^{n} w(\frac{X_i - X_j}{h})\} \to -\infty$$

For the second part, some of $X_1, ..., X_n$ are equal, that is $X_i = X_j$ for some different $i, j = 1, ...n$
Since $X_i - X_j = 0$ and $w(0) > 0$, so

$$\sum_{j=1}^{n} w(\frac{X_i - X_j}{h}) = \sum_{j=1}^{n} w(0) > 0$$

Since $h \downarrow 0$, so $\frac{1}{nh} \to \infty$, then we have

$$\frac{1}{nh}\sum_{j=1}^{n} w(\frac{X_i - X_j}{h}) \to \infty$$

Thus

$$ln\{\frac{1}{nh}\sum_{j=1}^{n} w(\frac{X_i - X_j}{h})\} \to \infty$$

Therefore

$$\mathcal{L}(h) = \frac{1}{n}\sum_{i=1}^{n} ln\{\frac{1}{nh}\sum_{j=1}^{n} w(\frac{X_i - X_j}{h})\} \to \infty$$

In conlusion, after combining these two parts we get that $\mathcal{L}(h) \uparrow \infty$ as $h \downarrow 0$

(b)(ii)

Note that $X_1, ..., X_n$ are distinct, that is $X_i \neq X_j$ for different $i, j = 1, ...n$, then $X_i - X_j \neq 0$

1) Show that $CV(h) \to -\infty$ as $h \downarrow 0$

Since for $X_i - X_j \neq 0$,

$$h^{-1} w(\frac{X_i - X_j}{h}) \to 0 \quad \text{as } h \downarrow 0$$

So

$$\frac{1}{h} \sum_{j \neq i} w(\frac{X_i - X_j}{h}) \to 0$$

Then

$$ln\{\frac{1}{(n-1)h} \sum_{j \neq i} w(\frac{X_i - X_j}{h})\} \to -\infty$$

Thus

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} ln\{\frac{1}{(n-1)h} \sum_{j \neq i} w(\frac{X_i - X_j}{h})\} \to -\infty$$

2) Show that $CV(h) \to -\infty$ as $h \uparrow \infty$

As $h \uparrow \infty$, since $X_i - X_j \neq 0$, then $\frac{X_i - X_j}{h} \to 0$, so $w(\frac{X_i - X_j}{h}) \to w(0)$
Since $w(0) > 0$, then $w(\frac{X_i - X_j}{h}) > 0$, so $\sum_{j \neq i} w(\frac{X_i - X_j}{h}) > 0$
And since $h \uparrow \infty$, then $\frac{1}{(n-1)h} \to 0$, so we have:

$$\frac{1}{(n-1)h} \sum_{j \neq i} w(\frac{X_i - X_j}{h}) \to 0$$

Thus

$$ln\{\frac{1}{(n-1)h} \sum_{j \neq i} w(\frac{X_i - X_j}{h})\} \to -\infty$$

Therefore

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} ln\{\frac{1}{(n-1)h} \sum_{j \neq i} w(\frac{X_i - X_j}{h})\} \to -\infty$$

In conclusion, we've proved that $CV(h) \to -\infty$ as $h \downarrow 0$and $h \uparrow \infty$

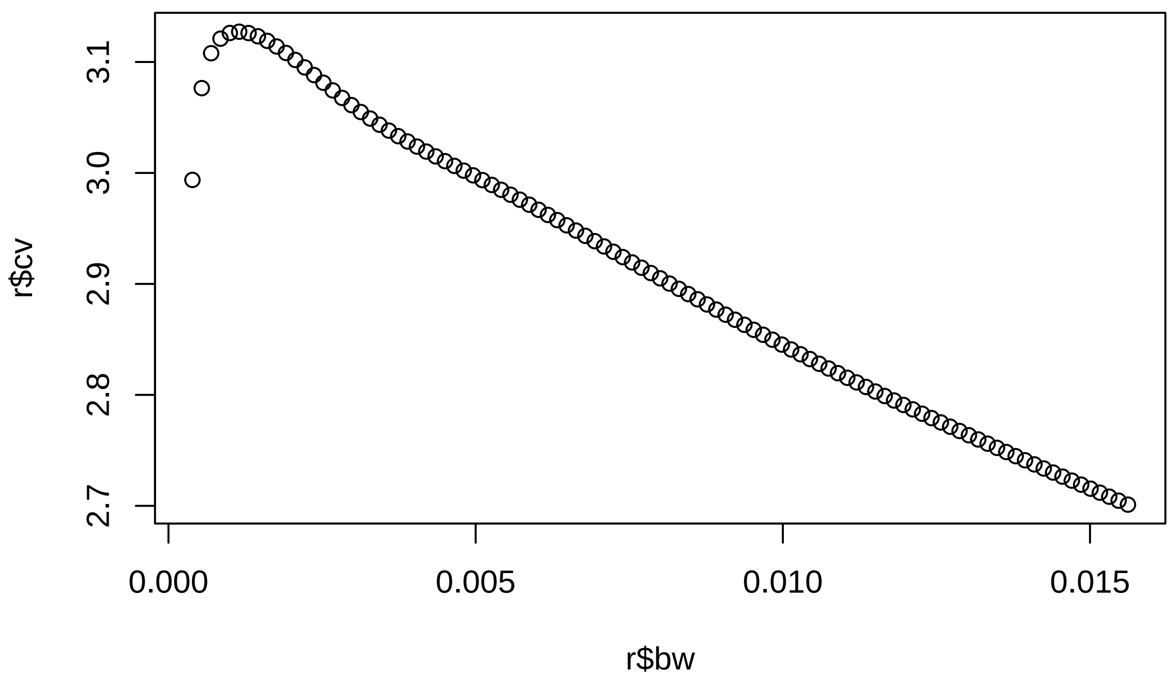
(c)

```
#function kde.cv
kde.cv <- function(x,h) {
             n <- length(x)
             if (missing(h)) {
                r <- density(x)
                h <- r$bw/10 + 3.9*c(0:100)*r$bw/100
                }
             cv <- NULL
             for (j in h) {
                cvj <- 0
                for (i in 1:n) {
                   z <- dnorm(x[i]-x,0,sd=j)/(n-1)
                   cvj <- cvj + log(sum(z[-i]))
```

```
            }
          cv <- c(cv,cvj/n)
          }
        r <- list(bw=h,cv=cv)
        r
}
```

Use above function to:
1) plot CV versus bandwidth

```
r <- kde.cv(stamp)
plot(r$bw,r$cv) # plot of CV versus bandwidth
```



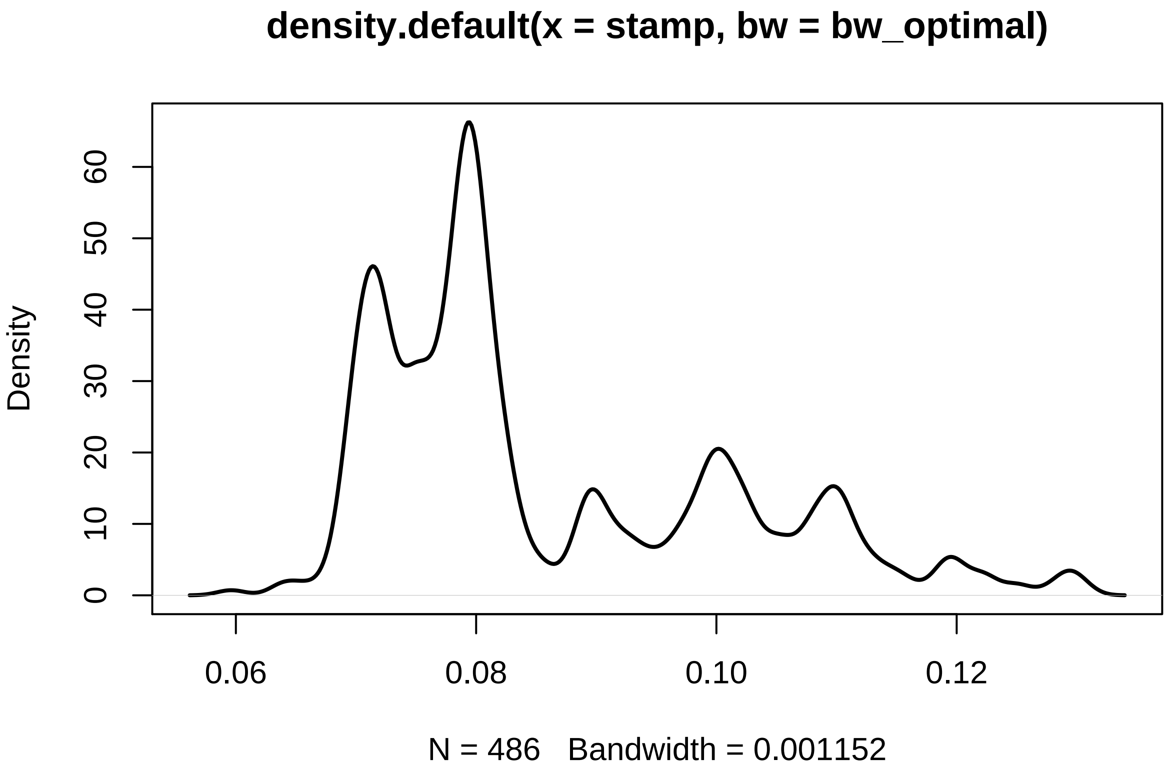
2) get the optimal value of bandwidth that maximizes CV

```
bw_optimal <- r$bw[r$cv==max(r$cv)] # bandwidth maximizing CV
bw_optimal
```

```
## [1] 0.001151797
```

And then use the optimal bandwidth value to estimate the density of the Hidalgo stamp data.

```
plot(density(stamp, bw=bw_optimal), lwd=2)
```



Thus, this density has about 7 modes.

Question2

(a)

Since $\mathcal{L}_F(t) = \frac{1}{\mu(F)} \int_0^t F^{-1}(s)ds$, so $\mathcal{L}_F^{'}(t) = \frac{1}{\mu(F)} \cdot F^{-1}(t)$
Set $\mathcal{L}_F^{'}(t) = 1$, then $\frac{1}{\mu(F)} \cdot F^{-1}(t) = 1$, so $F^{-1}(t) = \mu(F)$
Then we have

$$\begin{aligned} t &= F(\mu(F)) \\ &= F(\mu(F)-) \quad \text{; since F is a continuous distribution function} \\ &= MPS(F) \end{aligned}$$

Thus, $\mathcal{L}_F^{'}(MPS(F)) = 1$

(b)

Since $\mathcal{L}_F(t) = t^{\alpha+1}$, so $\mathcal{L}_F^{'}(t) = (\alpha+1)t^{\alpha}$
And since from part (a) we have $\mathcal{L}_F^{'}(MPS(F)) = 1$, so

$$\begin{aligned} (\alpha+1) \cdot (MPS(F))^{\alpha} &= 1 \\ (MPS(F))^{\alpha} &= \frac{1}{\alpha+1} \\ MPS(F) &= (\frac{1}{\alpha+1})^{\frac{1}{\alpha}} \end{aligned}$$

Thus,

$$MIS(F) = \mathcal{L}_F(MPS(F)) = \mathcal{L}_F[(\frac{1}{\alpha+1})^{\frac{1}{\alpha}}] = (\frac{1}{\alpha+1})^{\frac{\alpha+1}{\alpha}}$$

(c)

Implement the function for computing MPS:

```r
# function MPS
MPS <- function(x) {
  sum(x < mean(x)) / length(x)
}
```

Compute an estimate of MPS(F):

```r
income <- scan("incomes.txt") # load data incomes.txt
MPS(income)
```

```
## [1] 0.69
```

The jackknife standard error for estimating MPS(F) can be evaluated as follows:

```r
mps <- NULL
for (i in 1:200){
  x_i <- income[-i] # data with income[i] deleted
  mps <- c(mps, MPS(x_i))
}
mps_se <- sqrt(199*sum((mps-mean(mps))^2)/200) # jackknife standard error formula
mps_se
```

```
## [1] 0.07811778
```

(d)

Since we have $MIS(F) = \mathcal{L}_F(MPS(F))$ and $\hat{\mathcal{L}}_F(t) = \frac{1}{n\bar{X}} \sum_{i=1}^{\lceil nt \rceil} X_{(i)}$, so:

$$
\begin{aligned}
MI\hat{S}(F) &= \frac{1}{n\bar{X}} \cdot \sum_{i=1}^{\lceil n \cdot MPS(F) \rceil} X_{(i)} \\
&= \frac{1}{n\bar{X}} \cdot \sum_{i=1}^{\lceil n \cdot \frac{1}{n} \sum_{i=1}^{n} I(X_i < \bar{X}) \rceil} X_{(i)} \\
&= \frac{1}{n\bar{X}} \cdot \sum_{i=1}^{\lceil \sum_{i=1}^{n} I(X_i < \bar{X}) \rceil} X_{(i)}
\end{aligned}
$$

Then we can implement the function for computing MIS:

```r
# function MIS
MIS <- function(x){
  x <- sort(x) # sort the data to get order statistics
  n <- sum(x < mean(x))
  sum(x[1:n]) / (length(x)*mean(x))
}
```

Compute an estimate of MIS(F):

```r
MIS(income)
```

```
## [1] 0.3480396
```

The jackknife standard error for estimating MIS(F) can be evaluated as follows:

```r
mis <- NULL
for (i in 1:200){
  x_i <- income[-i] # data with income[i] deleted
  mis <- c(mis, MIS(x_i))
}
mis_se <- sqrt(199*sum((mis-mean(mis))^2)/200) # jackknife standard error formula
mis_se
```

```
## [1] 0.08170359
```