STAT 403 Spring 2018
HW08

Nan Tang
1662478

May 30, 2019

Q1

Q1-a

$$
\begin{aligned}
bias(\widehat{g_n}(x_0)) &= E(\widehat{g_n}(x_0)) - g(x_0) \\
&= E(\widehat{p_n}'(x_0)) - p'(x_0) \\
&= \frac{1}{h}E(K'(\frac{x - x_0}{h})) - p'(x_0) \\
&= \frac{1}{h}\int K'(\frac{x - x_0}{h})p(x)dx - p'(x_0)
\end{aligned}
$$

change variable such that $y = \frac{x-x_0}{h}$, then $dy = \frac{dx}{h}$, $K'(\frac{x-x_0}{h}) = K'(y)/h$

$$
bias(\widehat{g_n}(x_0)) = \frac{1}{h}\int K'(y)p(x_0 + hy)dy - p'(x_0)
$$

By Taylor expansion, when h is small,

$$
p(x_0 + hy) = p(x_0) - hy \cdot p'(x_0) + \frac{1}{2}h^2y^2p''(x_0) + \frac{1}{3!}h^3y^3p'''(x_0) + o(h^3)
$$

$$
\begin{aligned}
bias(\widehat{g_n}(x_0)) = \frac{1}{n}[p(x_0)\int K'(y)dy + hp'(x_0)\int yK'(y)dy \\
+ \frac{h^2p''(x_0)}{2}\int y^2K'(y)dy + \frac{h^3p'''(x_0)}{6}\int y^3K'(y)dy + o(h^2)] - p'(x_0)
\end{aligned}
$$

Since K' the derivative of standard normal is and odd function, $\int K'(y)dy, \int y^2K'(y)dy$ are equal to zero.

$\int yK'(y)dy = [yK(y)] - \int K(y)dy = 0 - 1$

$$
\begin{aligned}
bias(\widehat{g_n}(x_0)) &= p'(x_0) + h^2 \frac{p'''(x_0)}{6} \int y^3 K'(y)dy + o(h^2) - p'(x_0) \\
&= C_1 h^2 + o(h^2), \text{ where C1 is a constant}
\end{aligned}
$$

$$
\begin{aligned}
Var(\widehat{g_n}(x_0)) &= Var(\widehat{p_n}'(x_0)) \\
&= \frac{1}{nh^2} Var(K'(\frac{x-x_0}{h})) \\
&\leq \frac{1}{nh^2} E(K'(\frac{x-x_0}{h})^2) \\
&= \frac{1}{nh^2} \int K'(\frac{x-x_0}{h})^2 p(x)dx
\end{aligned}
$$

change variable such that $y = \frac{x-x_0}{h}$, then $dy = \frac{dx}{h}$, $K'(\frac{x-x_0}{h})^2 = K'(y)^2/h^2$

$$
\begin{aligned}
Var(\widehat{g_n}(x_0)) &\leq \frac{1}{nh^3} \int K'(y)p(x_0 + hy)dy \\
&= \frac{1}{nh^3} \int K'(p(x_0)^2 + hyp'(x_0) + o(h))dy
\end{aligned}
$$

Note that $\int K'(y)ydy = 0$ since K is an odd function.

$$
\begin{aligned}
Var(\widehat{g_n}(x_0)) &\leq \frac{1}{nh^3} p(x_0) \int K'(y)^2 dy + o(\frac{1}{nh^3}) \\
&= \frac{C_2}{nh^3} + o(\frac{1}{nh^3})
\end{aligned}
$$

Q1-b

$$
\begin{aligned}
bias(\widehat{p_n}(0)) &= E(\widehat{p_n}(0)) - p(0) \\
&= E(\frac{1}{nh} \sum K(\frac{x_i - 0}{h})) - p(0) \\
&= \frac{1}{h} \int K(\frac{x}{h}) p(x)dx - p(0)
\end{aligned}
$$

Note that $p(0) = 0$ and $p(x) = 2x$ when $0 \leq x \leq 1$, otherwise $p(x) = 0$

$$
\begin{aligned}
bias(\widehat{p_n}(0)) &= \frac{1}{n}\int_0^1 K(\frac{x}{h})2xdx \\
&= \frac{1}{n}\int_0^1 \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2h^2}} 2xdx \\
&= \frac{1}{h} \cdot \frac{2}{\sqrt{2\pi}} \int_0^1 e^{\frac{-x^2}{2h^2}} xdx \\
&= \frac{2}{h\sqrt{2\pi}}[-h^2(e^{-\frac{1}{2h^2}} - 1) + o(h^2)] \\
&= -\frac{2h}{\sqrt{2\pi}} e^{-\frac{1}{2h^2}} + o(h)
\end{aligned}
$$

let $C_3 = -\frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2h^2}}$, $e^{-\frac{1}{h^2}} < 1$, C_3 is a constant.

$$bias(\widehat{p_n}(0)) = C_3 h + o(h)$$

Q1-c

The CDF of $X_i^\star$, $F(x)$ can be represented as

$$
\begin{aligned}
F(x) &= P(X_i^\star \leq x) \\
&= P(Y_i^\star + Z_i \leq x)
\end{aligned}
$$

$Y_i^\star$ is bootstrapped from $X_1 ... X_n$, therefore, probability of $Y_i^\star = X_i$ is equal to $\frac{1}{n}$ (each X_i has equal chance be bootstrapped)

$$
\begin{aligned}
F(x) &= \sum_{i=1}^n \frac{P(Y_i^\star + Z_i \leq x | Y_i^\star = x_i)}{P(Y_i^\star = x_i)} \\
&= \frac{1}{n}\sum_{i=1}^n P(x_i + Z_i \leq x) \\
&= \frac{1}{n}\sum_{i=1}^n P(Z_i \leq x - x_i)
\end{aligned}
$$

$Z_i \sim N(0, h^2)$, therefore, $P(Z_i \leq x - x_i) = \Phi(\frac{x-x_i}{h})$

$$
\begin{aligned}
f_{X_i^\star}(x) &= dF(x)/dx \\
&= (d(\frac{x-x_i}{h})/dx) \cdot \frac{1}{n}\sum_{i=1}^{n} K(\frac{x-x_i}{h}) \text{ where K is pdf of standard normal} \\
&= \frac{1}{nh}\sum K(\frac{x-x_i}{h})
\end{aligned}
$$

Since K is a symmetric function around zero, $K(\frac{x-x_i}{h}) = K(\frac{x_i-x}{h})$.

Therefore, PDF of $X_i^\star = \frac{1}{nh}\sum K(\frac{x_i-x}{h})$ which is equal in density to KDE function.

Q1-d

$$
\begin{aligned}
\widehat{p_n}(x_0) &= \frac{1}{nh}\sum K(\frac{x_i - x_0}{h}) \\
&= \sum K(\frac{x_i - x_0}{h}), \text{ since } h = \frac{1}{n} \\
&= \sum \frac{1}{2} I(-1 \leq \frac{x_i - x_0}{h} \leq 1) \\
2\widehat{p_n}(x_0) &= \sum I(-1 \leq \frac{x_i - x_0}{h} \leq 1) \\
&= \sum I(-h + x_0 \leq x_i \leq h + x_0)
\end{aligned}
$$

$\sum I(-h + x_0 \leq x_i \leq h + x_0)$ is total number of observations that falls in interval $[-h + x_0, h + x_0]$.

For random variables $x_1, x_2...x_n$, let q_n denotes the probability of x_i falls in interval $[-h + x_0, h + x_0]$.

nq_n denotes number of x_i's falls in the interval, which depends on x_0 and h, when $n \to \infty$, $h \to 0$, so nq_n only depends on x_0 when n is large enough.

Let $\lambda(x_0)$ denotes function of x_0 only. When n is large enough, $nq_n \to \lambda(x_0)$, and by law of small number, $x_n \to Poi(\lambda(x_0))$ in distribution.

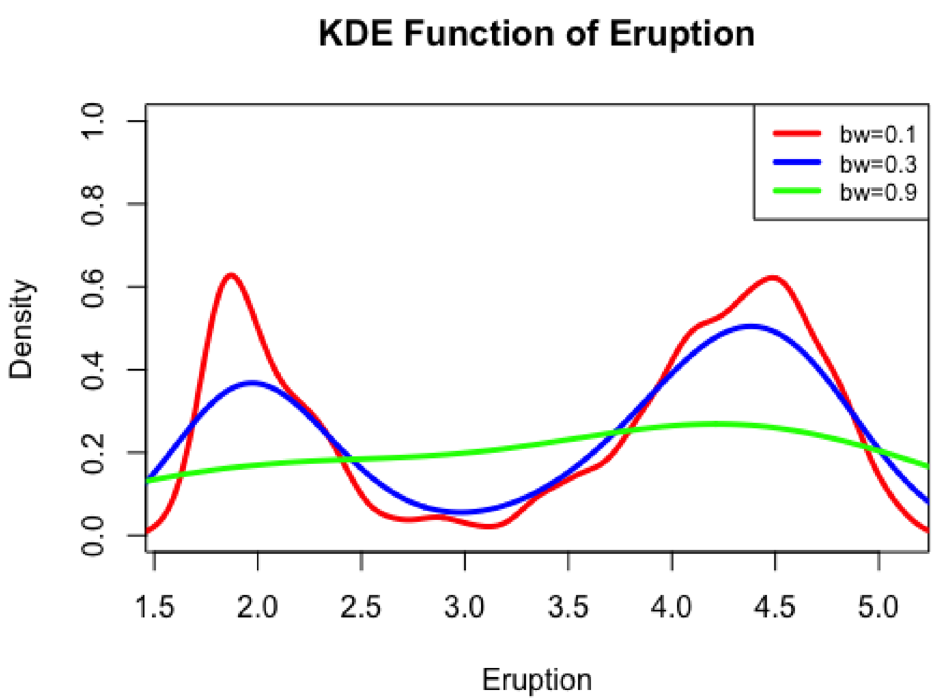
In this case, x_n is $2\widehat{p_n}(x_0)$, therefore, $2\widehat{p_n}(x_0) \to Poi(\lambda(x_0))$ in distribution.

Q2

Q2-a

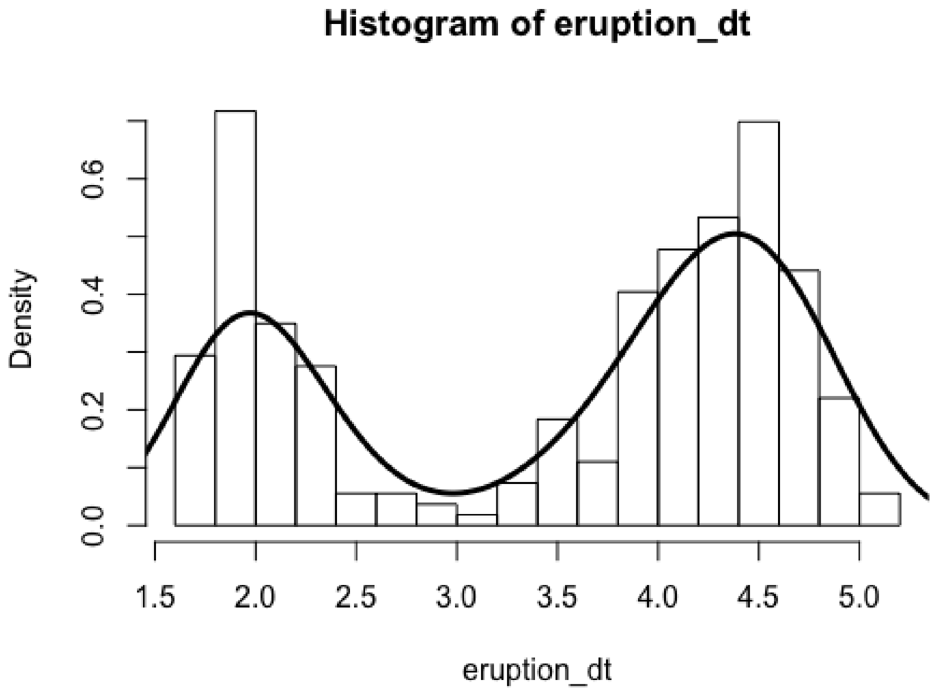
```
eruption_dt <- faithful$eruptions

bwds <- c(0.1, 0.3, 0.9)
colors <- c('red', 'blue', 'green')
plot(1, type="n", xlab="Eruption", ylab="Density", xlim=c(min(eruption_dt), max(eruption
     ylim=c(0, 1), main='KDE Function of Eruption')
for (ii in 1:length(bwds)) {
  erupt_kde <- density(eruption_dt, bw=bwds[ii])
  lines(erupt_kde, lwd=3, col=colors[ii])
}
legend('topright', legend=c('bw=0.1', 'bw=0.3', 'bw=0.9'), col=colors, lwd=3, cex=0.8)
```



Q2-b

```
erupt_kde <- density(eruption_dt, bw=0.3)
hist(eruption_dt, breaks=20, probability=T)
lines(erupt_kde, lwd=3)
```

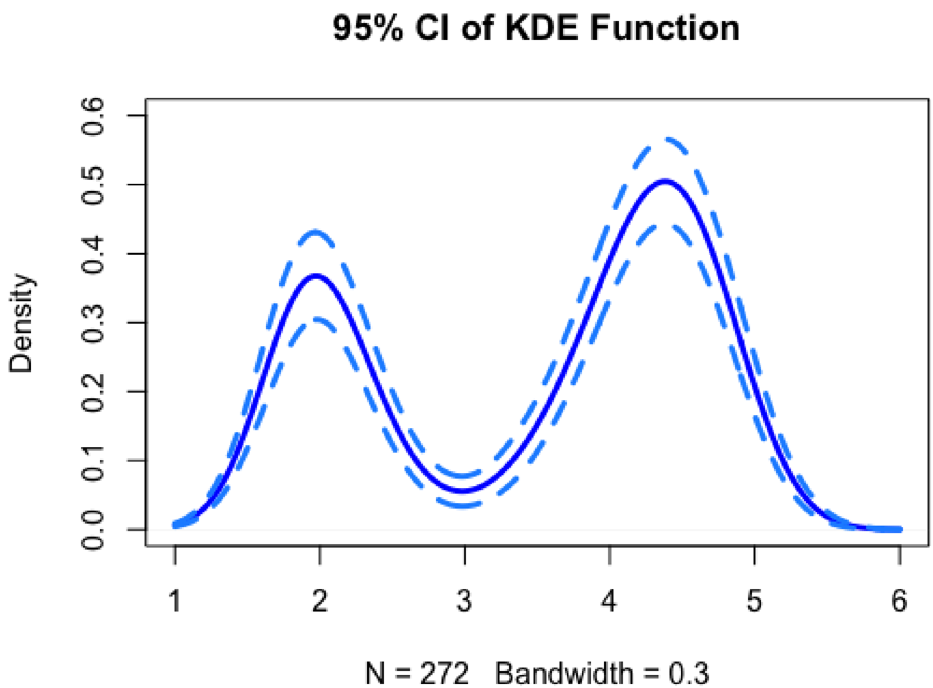


Q2-c

```
erupt_kde <- density(eruption_dt, from=1, to=6, bw=0.3)
n <- length(eruption_dt)
B <- 10000
kde_bt <- matrix(NA, B, length(erupt_kde$x))
for (ii in 1:B) {
  sp_index <- sample(n,n,replace=T)
  sp_bt <- eruption_dt[sp_index]
  sp_kde <- density(sp_bt, from=1, to=6, bw=0.3)
  kde_bt[ii,] <- sp_kde$y
}

bt_sd <- sqrt(diag(var(kde_bt)))

plot(erupt_kde, lwd=3, col="blue", ylim=c(0,0.6),main="95% CI of KDE Function")
lines(x=erupt_kde$x,y=erupt_kde$y+qnorm(0.975)*bt_sd, lwd=3, col="dodgerblue",
      lty=2)
lines(x=erupt_kde$x,y=erupt_kde$y-qnorm(0.975)*bt_sd, lwd=3, col="dodgerblue",
      lty=2)
```

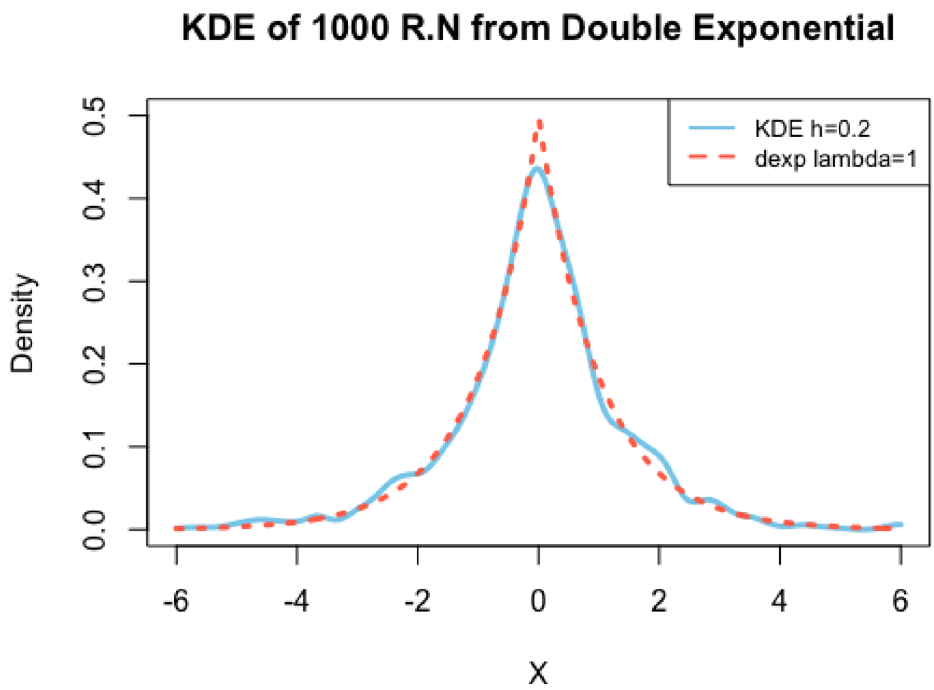


Q3

Q3-a

```
library(smoothmest)
n=1000
dexp_dt <- rdoublex(n, mu=0, lambda=1)
dexp_den <- density(dexp_dt, bw=0.2, from=-6, to=6, n=n)
x_base <- dexp_den$x
dexp_true <- ddoublex(x_base, mu=0, lambda=1)

plot(x=x_base, y=dexp_den$y, type='l', lwd=3, col='skyblue', xlim=c(-6,6),
     ylim=c(0, 0.5), xlab='X', ylab='Density', main='KDE of 1000 R.N from Double Exponen
lines(x=x_base, y=dexp_true, lwd=3, lty=3, col='coral1')
legend('topright', legend=c('KDE h=0.2', 'dexp lambda=1'), col=c('skyblue', 'coral1'),
       lty=c(1,2), lwd=2, cex=0.8)
```



Q3-b

In this case, both bias and variance cause KDE unmatch to true density. We can perceive from the plot in previous question that true density has a sharp turning (peaked bump) near zero, meanwhile, density value around zero is much higher than at other regions. Large bias appears at sharp turning point, while large variance appears at where density is large. Points around zero satisfy both conditions, therefore, large bias and large variance together caused the discrepancy at zero.

Q3-c

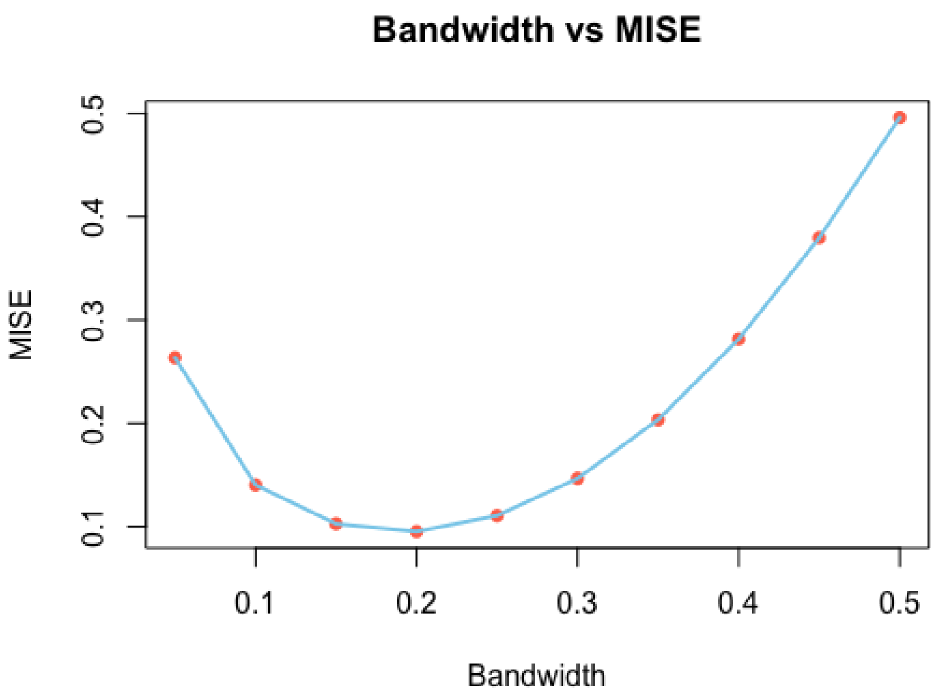
```
N = 10000
bwds <- seq(0.05, 0.5, 0.05)

mise_result <- rep(NA, length(bwds))
for (ii in 1:length(bwds)) {
  bwd <- bwds[ii]
  kde_result <- matrix(NA, nrow=N, ncol=n)
  for (jj in 1:N) {
    temp_den <- density(dexp_dt, from=-6, to=6, n=n, bw=bwd)
    kde_result[jj,] <- temp_den$y
```

```
  }
  mse_result <- colSums((t(t(kde_result) - dexp_true))^2) /N
  mise_result[ii] <- sum(mse_result)
}

plot(x=bwds, y=mise_result, pch=16, xlab='Bandwidth', ylab='MISE',
     main='Bandwidth vs MISE', col='coral1')
lines(x=bwds, y=mise_result, lwd=2, col='skyblue')

opt_bwd <- bwds[which.min(mise_result)]
> opt_bwd
[1] 0.2
```



MISE minimized when choosing bandwidth 0.2