Exercies 8 - Alon Goodman & Ran Hassid

Q1

a

נתונות n תצפיות iid. האומד הוא $S(\vec{x}) = \hat{\theta}$

האומד ללא תצפית i הוא $\hat{\theta}_{(-i)}$

נגדיר $\tilde{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{(-i)}$

א. נניח $S(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$

$$\tilde{\theta}_i = n \cdot \frac{1}{n}\sum_{i=1}^{n} f(x_i) - (n-1)\cdot \frac{1}{n-1}\cdot \sum_{j\neq i} Fx_j = \sum_{i=1}^{n} f(x_i) - \sum_{j\neq i} f(x_j) = f(x_i)$$

קיבלנו ש $\tilde{\theta}_i = f(x_i)$ ו $\vec{x}$ הם iid לכן גם $\tilde{\theta}_i$ הם iid.

b

ב.

$$\frac{1}{n}\cdot\sum_{i=1}^{n}\tilde{\theta}_i = \frac{1}{n}\left(\sum_{i=1}^{n} n\hat{\theta} - (n-1)\hat{\theta}_{(-i)}\right) = \frac{1}{n}\left[n^2\hat{\theta} - (n-1)\sum_{i=1}^{n}\hat{\theta}_{(-i)}\right] =$$

$$n\hat{\theta} - \frac{n-1}{n}\cdot\sum_{i=1}^{n}\hat{\theta}_{(-i)} = \hat{\theta} + (n-1)\hat{\theta} - (n-1)\cdot\frac{\sum\hat{\theta}_{(-i)}}{n} =$$

$$\hat{\theta} + (n-1)\left(\hat{\theta} - \frac{\sum_{i=1}^{n}\hat{\theta}_{(-i)}}{n}\right) = \hat{\theta} - \widehat{bias}_{Jk}\,\hat{\theta}$$

c

ג.

$$\hat{Var}(\bar{\tilde{\theta}}_i) = \frac{1}{n} Var(\tilde{\theta}_i) = \frac{\sum_{i=1}^{n}(\tilde{\theta}_i - \bar{\tilde{\theta}}_i)^2}{(n-1)n} = \frac{\sum(\tilde{\theta}_i - \hat{\theta} + bias_{JK}(\hat{\theta}))^2}{(n-1)n} =$$

$$= \frac{\sum((n-1)\hat{\theta} - (n-1)\hat{\theta}_{(-i)} + bias_{JK}(\hat{\theta}))^2}{(n-1)n} =$$

$$= \frac{\sum_{i=1}^{n}((n-1)(\hat{\theta} - \hat{\theta}_{(-i)}) + bias_{JK}(\hat{\theta}))^2}{(n-1)n} =$$

$$= \frac{\sum_{i=1}^{n}[(n-1)(\hat{\theta} - \hat{\theta}_{(-i)}) + (n-1)(\frac{\sum \hat{\theta}_{(-i)}}{n} - \hat{\theta})]^2}{(n-1)n} =$$

$$= \frac{\sum[(n-1)(\hat{\theta} - \hat{\theta}_{(-i)} + \bar{\hat{\theta}}_{(-i)} - \hat{\theta})]^2}{(n-1)n} =$$

$$= \frac{\sum[(n-1)(-1)(\hat{\theta}_{(-i)} - \bar{\hat{\theta}}_{(-i)})]^2}{(n-1)n} = \frac{(n-1)\cdot\sum[\hat{\theta}_{(-i)} - \bar{\hat{\theta}}_{(-i)}]^2}{n} = Var_{JK}(\hat{\theta})$$

d

ד. ניתן להסיק ששונות JK טובה אם c היא פונקציה לינארית כדי להשיג אומדן
טוב יותר. נראה והתנאי הוא אומדן לא לינארי שונות JK מתאימה במידה באומדן.
כמו שראינו בתרגיל קודם התנאי סמוך על הפונקציה c כפונקציה של גודל המדגם
והאלגוריתם מושפע לא רק על תרגילים ולא על n תרגילים

Q2

a

```
ex8dataq2 <- read.csv("~/Desktop/Ran/D year/semester b/hishov statisti/exercies/HW8/e
x8data1.csv")
#View(ex8dataq2)
```

```
var_cov_metrix <- var(ex8dataq2)
eigen_vec <- eigen(var_cov_metrix)
max_eigen_vec_hat <- max(eigen_vec$values)
```

```r
n.size <- 500
B <- 1000

result_vec <- numeric(B)

for (b in 1:B) {
  rows.sample.b <- sample.int(n = 500,size = 500,replace = TRUE)
  b_data <- ex8dataq2[rows.sample.b,]

  var_cov_metrix.b <- var(b_data)
  eigen_vec.b <- eigen(var_cov_metrix.b)
  max_eigen_vec_hat.b <- max(eigen_vec.b$values)

  result_vec[b] <- max_eigen_vec_hat.b
}
```

Pivot 95% CI

```r
alpha.pivot <- 2*max_eigen_vec_hat - quantile(result_vec,0.975)
beta.pivot <- 2*max_eigen_vec_hat - quantile(result_vec,0.025)

c(alpha.pivot,beta.pivot)
```

```
##    97.5%     2.5%
## 3.967878 6.567409
```

Percentiles 95% CI

```r
quantile(result_vec,c(0.025,0.975))
```

```
##     2.5%    97.5%
## 4.218732 6.818263
```

b

```r
var.b <- var(result_vec)
bais.b <- mean(result_vec) - max_eigen_vec_hat
```

```
## The variance estimator is = 0.4698
## The bias estimator is = -0.00042
```

c

```r
result_vec_jk <- numeric(500)

for (i in 1:500) {
  jk_data.i <- ex8dataq2[-i,]

  var_cov_metrix.i <- var(jk_data.i)
  eigen_vec.i <- eigen(var_cov_metrix.i)
  max_eigen_vec_hat.i <- max(eigen_vec.i$values)

  result_vec_jk[i] <- max_eigen_vec_hat.i
}
```

```
## The JK variance estimator is = 0.00097
## The JK bias estimator is = 3e-05
```

We got a smaller variance & bias in the JK methods.

It is make sense because in JK methods we take out one sample. So, the final result does not change dramatically

Q3

$Z=(X,Y)$

$$T_1 = \int x \, d\hat{F}(z), \quad T_2 = \int y \, d\hat{F}(z), \quad T_3 = \int xy \, d\hat{F}(z), \quad T_4 = \int x^2 \, d\hat{F}(z), \quad T_5 = \int y^2 \, d\hat{F}(z)$$

$$\int h(x) \, d\hat{F} = \frac{1}{n} \sum_{i=1}^{n} h(x_i)$$

$$\widehat{corr}(X,Y) = a(T_1, T_2, T_3, T_4, T_5) = \frac{T_3 - T_1 \cdot T_2}{\sqrt{(T_4 - T_1^2)(T_5 - T_2^2)}} =$$

$$= \frac{\frac{1}{n}\sum x_i y_i - \frac{1}{n}\sum x_i \cdot \frac{1}{n}\sum y_i}{\sqrt{[\frac{1}{n}\sum x_i^2 - (\frac{1}{n}\sum x_i)^2][\frac{1}{n}\sum y_i^2 - (\frac{1}{n}\sum y_i)^2]}} = \frac{\frac{1}{n}\sum x_i y_i - \bar{x} \cdot \bar{y}}{\sqrt{(\frac{\sum x_i^2}{n} - \bar{x}^2)(\frac{\sum y_i^2}{n} - \bar{y}^2)}} = \frac{\frac{\sum x_i y_i - n\bar{x}\bar{y}}{n}}{\sqrt{(\frac{\sum x_i^2 - n\bar{x}^2}{n})(\frac{\sum y_i^2 - n\bar{y}^2}{n})}} =$$

$$\frac{\sum x_i y_i - n\bar{x}\bar{y}}{(\sum x_i^2 - n\bar{x}^2)(\sum y_i^2 - n\bar{y}^2)} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}$$

$$\sum (x_i - \bar{x})^2 = \sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2 = \sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum x_i^2 - n\bar{x}$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y} = \sum x_i y_i - \bar{y}\sum x_i - \bar{x}\sum y_i + n\bar{x}\bar{y} =$$

$$= \sum x_i y_i - n\bar{y}\bar{x} - n\bar{x}\bar{y} + n\bar{x}\bar{y} = \sum x_i y_i - n\bar{x}\bar{y}$$

Q4

```r
ex8data2 <- read.csv("~/Desktop/Ran/D year/semester b/hishov statisti/exercies/HW8/ex
8data2.csv")
#View(ex8data2)
```

a

```
k_nn <- function(xy_data,x_new,k){
  nn.index <- order(rank(x = abs(xy_data$x - x_new),ties.method = "random"))
  nn.index <- nn.index[1:k]

  y_new <- sum(xy_data$y[nn.index])/k
  return(y_new)
}
```

```
k_options <- c(3,15,100)
PRESS_per_k <- numeric(3)

for (i in 1:3) {
  results_for_k <- numeric(1500)
  for (j in 1:1500) {
    y_new_j <- k_nn(xy_data = ex8data2[-j,],x_new = ex8data2$x[j],k = k_options[i])
    results_for_k[j] <- (ex8data2$y[j] - y_new_j)^2
  }

  PRESS_per_k[i] <- sum(results_for_k)
}
```

```
##                    3       15      100
## PRESS_per_k 324.4492 257.9184 329.2284
```

The K that minimizing the PRESS is 15.

b

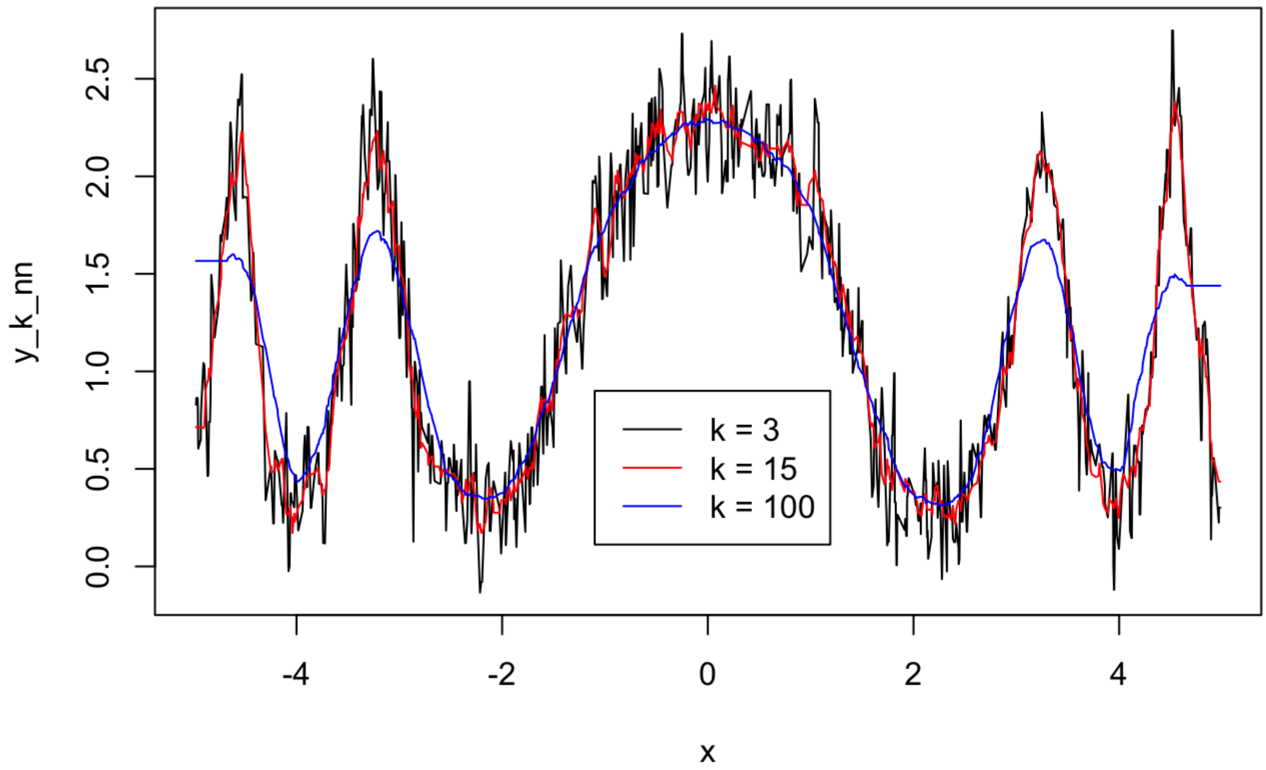
```
k_options <- c(3,15,100)

Q4.b.df <- data.frame(x_new = numeric(1000), k_3 = numeric(1000), k_15 = numeric(1000
), k_100 = numeric(1000))


for (i in 1:1000) {
  x_new_Q4 <- runif(n = 1,min = min(ex8data2$x),max = max(ex8data2$x))
  Q4.b.df[i,1] <- x_new_Q4
  for (k in 1:3) {
    y_new_k <- k_nn(xy_data = ex8data2,x_new = x_new_Q4, k = k_options[k])
    Q4.b.df[i,k+1] <- y_new_k
  }
}
```

```
Q4.b.df <- Q4.b.df[order(Q4.b.df$x),]

plot(x = Q4.b.df$x_new, y = Q4.b.df$k_3,type = 'l', xlab = "x", ylab = "y_k_nn")
legend(x = -1.1, y = .9, legend = c("k = 3", "k = 15", "k = 100"),col = c("black", "r
ed", "blue"),lty = 1)
lines(x = Q4.b.df$x_new, y = Q4.b.df$k_15, col='red')
lines(x = Q4.b.df$x_new, y = Q4.b.df$k_100, col='blue')
```



We can see that this is a variance-bias trade off problem.

It can be seen that for k = 3 we got a very flexible model with large variance and small bias.

For k = 100 we get a model with a small variance but with a large bias.

For k = 15 we get a combination of variance and bias neither small nor large.

In this case, according to the data in section A, it seems that we will prefer the model with k = 15, which gives us a better combination of variance and bias than the other models.

This is consistent with the results in section A where we saw that k = 15 minimizing the PRESS.