STAT243-PS4

Jinhui Xu

October 2017

1 Other students

I discuss some problems with Xin Shi.

2 Question 1

2.1 (a)

There is only one copy. Because we can see that data and input share same address.

```
x <- 1:10
f <- function(input){
  print(.Internal(inspect(input)))       #check the address of input
  data <- input
  print(.Internal(inspect(data)))        #check the address of data
        g <- function(param) return(param * data)
        return(g)
}
data<-100
myFun <- f(x)

## @117a12880 13 INTSXP g0c4 [NAM(2)] (len=10, tl=0) 1,2,3,4,5,...
##  [1]  1  2  3  4  5  6  7  8  9 10
## @117a12880 13 INTSXP g0c4 [NAM(2)] (len=10, tl=0) 1,2,3,4,5,...
##  [1]  1  2  3  4  5  6  7  8  9 10
```

2.2 (b)

The size of the serialized object is doubled. The reason is that R store both input and data even though they have same address.

```
x <- rnorm(1e5)
f <- function(input){
  data <- input
        g <- function(param) return(param * data)
        return(g)
}
myFun <- f(x)
object.size(x)

## 800040 bytes

length(serialize(myFun,NULL))
```

```
## [1] 1606454
```

2.3 (c)

When the function contains the command: data=input. myFun can get the value of data even x is removed. However, when we delete that command, myFun need the value of input of f which means that myFun needs the value of x. So if we rm x, there would be an error.

```
x <- 1:10
f <- function(data){
  g <- function(param) return(param * data)
  return(g)
}
myFun <- f(x)
rm(x)
data <- 100
myFun(3)

## Error in myFun(3):  'x'

ls(envir=environment(myFun))

## [1] "data" "g"
```

2.4 (d)

We can use force to force the value of data.

```
x <- 1:10
f <- function(data){
  force(data)
  g <- function(param) return(param * data)
  return(g)
}
myFun <- f(x)
rm(x)
data <- 100
myFun(3)

##  [1]  3  6  9 12 15 18 21 24 27 30
```

3 Question 2

3.1 (a)

When change the a vector of list, I find that the address of the relevant changes and the other one does not change. Therefore, R would create a new vector.

```
list1=list(a=rnorm(1e5),b=rnorm(1e5))
.Internal(inspect(list1))
```

```
## @12b4f14b0 19 VECSXP g0c2 [NAM(2),ATT] (len=2, tl=0)
##   @10ec8a000 14 REALSXP g0c7 [] (len=100000, tl=0) 0.163302,1.5673,0.59027,-1.25237,-0.657689,...
##   @112188000 14 REALSXP g0c7 [] (len=100000, tl=0) -0.316091,0.271516,1.18125,0.0445575,-1.83643,...
## ATTRIB:
##   @10971f640 02 LISTSXP g0c0 []
##     TAG: @103824140 01 SYMSXP g1c0 [MARK,NAM(2),LCK,gp=0x6000] "names" (has value)
##     @12b4f14e8 16 STRSXP g0c2 [] (len=2, tl=0)
##       @103021b98 09 CHARSXP g1c1 [MARK,gp=0x61] [ASCII] [cached] "a"
##       @1032eab98 09 CHARSXP g1c1 [MARK,gp=0x61] [ASCII] [cached] "b"
```

```
list1[[1]][1]<-100
.Internal(inspect(list1))
```

```
## @13d5cb608 19 VECSXP g0c2 [NAM(1),ATT] (len=2, tl=0)
##   @10b52e000 14 REALSXP g0c7 [] (len=100000, tl=0) 100,1.5673,0.59027,-1.25237,-0.657689,...
##   @112188000 14 REALSXP g0c7 [NAM(2)] (len=100000, tl=0) -0.316091,0.271516,1.18125,0.0445575,-1.836
## ATTRIB:
##   @10f603b78 02 LISTSXP g0c0 []
##     TAG: @103824140 01 SYMSXP g1c0 [MARK,NAM(2),LCK,gp=0x6000] "names" (has value)
##     @12b4f14e8 16 STRSXP g0c2 [NAM(2)] (len=2, tl=0)
##       @103021b98 09 CHARSXP g1c1 [MARK,gp=0x61] [ASCII] [cached] "a"
##       @1032eab98 09 CHARSXP g1c1 [MARK,gp=0x61] [ASCII] [cached] "b"
```

3.2 (b)

According to the adddress of two lists before any change, we know that there is no copy-on-change.When the change is made, only the address of the relevant vector changes. Therefore, only a copy of the relevant vector is made.

```
list2=list(a=rnorm(1e5),b=rnorm(1e5))
list2_cp=list2
.Internal(inspect((list2)))
```

```
## @1197c19f8 19 VECSXP g0c2 [NAM(2),ATT] (len=2, tl=0)
##   @1114cf000 14 REALSXP g0c7 [] (len=100000, tl=0) -0.917492,1.34264,0.880094,-0.928486,1.4136,...
##   @112000000 14 REALSXP g0c7 [] (len=100000, tl=0) -1.08646,-0.686031,1.63498,1.22225,-0.340234,...
## ATTRIB:
##   @11841e0b0 02 LISTSXP g0c0 []
##     TAG: @103824140 01 SYMSXP g1c0 [MARK,NAM(2),LCK,gp=0x6000] "names" (has value)
##     @1197c1a68 16 STRSXP g0c2 [] (len=2, tl=0)
##       @103021b98 09 CHARSXP g1c1 [MARK,gp=0x61] [ASCII] [cached] "a"
##       @1032eab98 09 CHARSXP g1c1 [MARK,gp=0x61] [ASCII] [cached] "b"
```

```
.Internal(inspect((list2_cp)))
```

```
## @1197c19f8 19 VECSXP g0c2 [NAM(2),ATT] (len=2, tl=0)
##   @1114cf000 14 REALSXP g0c7 [] (len=100000, tl=0) -0.917492,1.34264,0.880094,-0.928486,1.4136,...
##   @112000000 14 REALSXP g0c7 [] (len=100000, tl=0) -1.08646,-0.686031,1.63498,1.22225,-0.340234,...
## ATTRIB:
##   @11841e0b0 02 LISTSXP g0c0 []
##     TAG: @103824140 01 SYMSXP g1c0 [MARK,NAM(2),LCK,gp=0x6000] "names" (has value)
##     @1197c1a68 16 STRSXP g0c2 [] (len=2, tl=0)
##       @103021b98 09 CHARSXP g1c1 [MARK,gp=0x61] [ASCII] [cached] "a"
##       @1032eab98 09 CHARSXP g1c1 [MARK,gp=0x61] [ASCII] [cached] "b"
```

```
#the address of list2_cp is same with that of list2
list2_cp[[1]][1]<-100
.Internal(inspect(list2_cp))

## @12b2d0a78 19 VECSXP g0c2 [NAM(1),ATT] (len=2, tl=0)
##   @11140b000 14 REALSXP g0c7 [] (len=100000, tl=0) 100,1.34264,0.880094,-0.928486,1.4136,...
##   @112000000 14 REALSXP g0c7 [NAM(2)] (len=100000, tl=0) -1.08646,-0.686031,1.63498,1.22225,-0.34023
## ATTRIB:
##   @14cf41e00 02 LISTSXP g0c0 []
##     TAG: @103824140 01 SYMSXP g1c0 [MARK,NAM(2),LCK,gp=0x6000] "names" (has value)
##     @1197c1a68 16 STRSXP g0c2 [NAM(2)] (len=2, tl=0)
##       @103021b98 09 CHARSXP g1c1 [MARK,gp=0x61] [ASCII] [cached] "a"
##       @1032eab98 09 CHARSXP g1c1 [MARK,gp=0x61] [ASCII] [cached] "b"

#after change, we can find only the address of relevant vector changes
```

3.3 (c)

Notice the change of address after adding a vector into the second list. The address of two lists becomes different, but two original vectors still have original addresses. So the only change is that the second list creates a new vector while other vectors still share original addresses.

```
list3=list(a=list(rnorm(1e5)),b=list(rnorm(1e5)))
list3_cp=list3
.Internal(inspect(list3))

## @11a704b88 19 VECSXP g0c2 [NAM(2),ATT] (len=2, tl=0)
##   @103dd0d28 19 VECSXP g0c1 [] (len=1, tl=0)
##     @110d00000 14 REALSXP g0c7 [] (len=100000, tl=0) -0.304923,0.11101,0.137074,-0.389024,-0.476274,
##   @103dd0de8 19 VECSXP g0c1 [] (len=1, tl=0)
##     @11224c000 14 REALSXP g0c7 [] (len=100000, tl=0) 0.380603,-0.357204,-0.230827,0.886193,-0.540597
## ATTRIB:
##   @111b98758 02 LISTSXP g0c0 []
##     TAG: @103824140 01 SYMSXP g1c0 [MARK,NAM(2),LCK,gp=0x6000] "names" (has value)
##     @11a704c68 16 STRSXP g0c2 [] (len=2, tl=0)
##       @103021b98 09 CHARSXP g1c1 [MARK,gp=0x61] [ASCII] [cached] "a"
##       @1032eab98 09 CHARSXP g1c1 [MARK,gp=0x61] [ASCII] [cached] "b"

.Internal(inspect(list3_cp))

## @11a704b88 19 VECSXP g0c2 [NAM(2),ATT] (len=2, tl=0)
##   @103dd0d28 19 VECSXP g0c1 [] (len=1, tl=0)
##     @110d00000 14 REALSXP g0c7 [] (len=100000, tl=0) -0.304923,0.11101,0.137074,-0.389024,-0.476274,
##   @103dd0de8 19 VECSXP g0c1 [] (len=1, tl=0)
##     @11224c000 14 REALSXP g0c7 [] (len=100000, tl=0) 0.380603,-0.357204,-0.230827,0.886193,-0.540597
## ATTRIB:
##   @111b98758 02 LISTSXP g0c0 []
##     TAG: @103824140 01 SYMSXP g1c0 [MARK,NAM(2),LCK,gp=0x6000] "names" (has value)
##     @11a704c68 16 STRSXP g0c2 [] (len=2, tl=0)
##       @103021b98 09 CHARSXP g1c1 [MARK,gp=0x61] [ASCII] [cached] "a"
##       @1032eab98 09 CHARSXP g1c1 [MARK,gp=0x61] [ASCII] [cached] "b"

list3_cp$b[[2]]<-rnorm(1e5)
.Internal(inspect(list3_cp))
```

```
## @11a667588 19 VECSXP g0c2 [NAM(1),ATT] (len=2, tl=0)
##   @103dd0d28 19 VECSXP g0c1 [NAM(2)] (len=1, tl=0)
##     @110d00000 14 REALSXP g0c7 [] (len=100000, tl=0) -0.304923,0.11101,0.137074,-0.389024,-0.476274,
##   @11a6675c0 19 VECSXP g0c2 [] (len=2, tl=0)
##     @11224c000 14 REALSXP g0c7 [NAM(2)] (len=100000, tl=0) 0.380603,-0.357204,-0.230827,0.886193,-0.5
##     @112310000 14 REALSXP g0c7 [NAM(2)] (len=100000, tl=0) 1.05821,0.760135,0.484417,-1.22833,0.8132
## ATTRIB:
##   @111cdde98 02 LISTSXP g0c0 []
##     TAG: @103824140 01 SYMSXP g1c0 [MARK,NAM(2),LCK,gp=0x6000] "names" (has value)
##     @11a704c68 16 STRSXP g0c2 [NAM(2)] (len=2, tl=0)
##       @103021b98 09 CHARSXP g1c1 [MARK,gp=0x61] [ASCII] [cached] "a"
##       @1032eab98 09 CHARSXP g1c1 [MARK,gp=0x61] [ASCII] [cached] "b"
```

3.4 (d)

Object.size is twice large as the result of gc. I guess that it is because two elements of list is stored in the same address, but object.size estimates the size of list equels to sum of size of each element.

```
gc()

##            used  (Mb) gc trigger  (Mb)  max used  (Mb)
## Ncells  1968924 105.2    3886542 207.6   3886542 207.6
## Vcells 30853507 235.4   97834876 746.5 109412705 834.8

tmp <- list()
x <- rnorm(1e7)
tmp[[1]] <- x
tmp[[2]] <- x
.Internal(inspect(tmp))

## @117682c40 19 VECSXP g0c2 [NAM(1)] (len=2, tl=0)
##   @146800000 14 REALSXP g0c7 [NAM(2)] (len=10000000, tl=0) -0.135277,1.27501,-0.777146,0.638026,-1.7
##   @146800000 14 REALSXP g0c7 [NAM(2)] (len=10000000, tl=0) -0.135277,1.27501,-0.777146,0.638026,-1.7

object.size(tmp)

## 160000136 bytes

gc()

##            used  (Mb) gc trigger  (Mb)  max used  (Mb)
## Ncells  1968862 105.2    3886542 207.6   3886542 207.6
## Vcells 30853415 235.4   97834876 746.5 109412705 834.8
```

4 Question 3

Notice that in the original code,firstly, the if else is not necessary at all. So I directly calculate q without if else. Secondly I replace three nested for loops with simple computation of vector

```
ll <- function(Theta, A) {
  sum.ind <- which(A==1, arr.ind=T)
  logLik <- sum(log(Theta[sum.ind])) - sum(Theta)
  return(logLik)
}
########################################
##original code########################
oneUpdate <- function(A, n, K, theta.old, thresh = 0.1) {
  theta.old1 <- theta.old
  Theta.old <- theta.old %*% t(theta.old)
  L.old <- ll(Theta.old, A)
  q <- array(0, dim = c(n, n, K))
##############the following part would be revised
  for (i in 1:n) {
    for (j in 1:n) {
      for (z in 1:K) {
        if (theta.old[i, z]*theta.old[j, z] == 0){
        q[i, j, z] <- 0 } else {
           q[i, j, z] <- theta.old[i, z]*theta.old[j, z] /Theta.old[i, j]
         }
       }
     }
  }
################
  theta.new <- theta.old
  for (z in 1:K) {
  theta.new[,z] <- rowSums(A*q[,,z])/sqrt(sum(A*q[,,z]))
  }
  Theta.new <- theta.new %*% t(theta.new)
  L.new <- ll(Theta.new, A)
      converge.check <- abs(L.new - L.old) < thresh
  theta.new <- theta.new/rowSums(theta.new)
  return(list(theta = theta.new, loglik = L.new,converged = converge.check))
}

########################################
##revised code#########################
oneUpdate_new <- function(A, n, K, theta.old, thresh = 0.1) {
  theta.old1 <- theta.old
  Theta.old <- theta.old %*% t(theta.old)
  L.old <- ll(Theta.old, A)
  q <- array(0, dim = c(n, n, K))
###################begin.revised part
  for (z in 1:K) {
       q[ , , z] <- theta.old[, z]%*%t(theta.old[ , z]) /Theta.old
  }
###################end.revised part
  theta.new <- theta.old
  for (z in 1:K) {
  theta.new[,z] <- rowSums(A*q[,,z])/sqrt(sum(A*q[,,z]))
  }
  Theta.new <- theta.new %*% t(theta.new)
  L.new <- ll(Theta.new, A)
      converge.check <- abs(L.new - L.old) < thresh
  theta.new <- theta.new/rowSums(theta.new)
  return(list(theta = theta.new, loglik = L.new,converged = converge.check))
```

```
}
# initialize the parameters at random starting values
temp <- matrix(runif(n*K), n, K)
theta.init <- temp/rowSums(temp)
#compare the time used
system.time(out <- oneUpdate(A, n, K, theta.init))
system.time(out_new <- oneUpdate_new(A, n, K, theta.init))
all.equal(out,out_new)
```

The results are same while the time used by revised code decreases.

5 Question 4

Notice that in the function FYKD, the for loop is aim to generate vector x. However, the algorithm only need the first k value of vector x. So we can only calculate that part rather than entire vector.

```
PIKK <- function(x, k) {
x[sort(runif(length(x)), index.return = TRUE)$ix[1:k]]
}
FYKD <- function(x, k) {
  n <- length(x)
#in the original code , the following code is to generate entire n values in vector x
  for(i in 1:n) {
     j = sample(i:n, 1)
     tmp <- x[i]
     x[i] <- x[j]
     x[j] <- tmp
  }
return(x[1:k])  # while we only need first k values
}

#so revised code do not calculate the latter part of the vector
FYKD_new <- function(x, k) {
  n <- length(x)
  for(i in 1:k) {
     j = sample(i:n, 1)
     tmp <- x[i]
     x[i] <- x[j]
     x[j] <- tmp
  }
return(x[1:k])
}
```

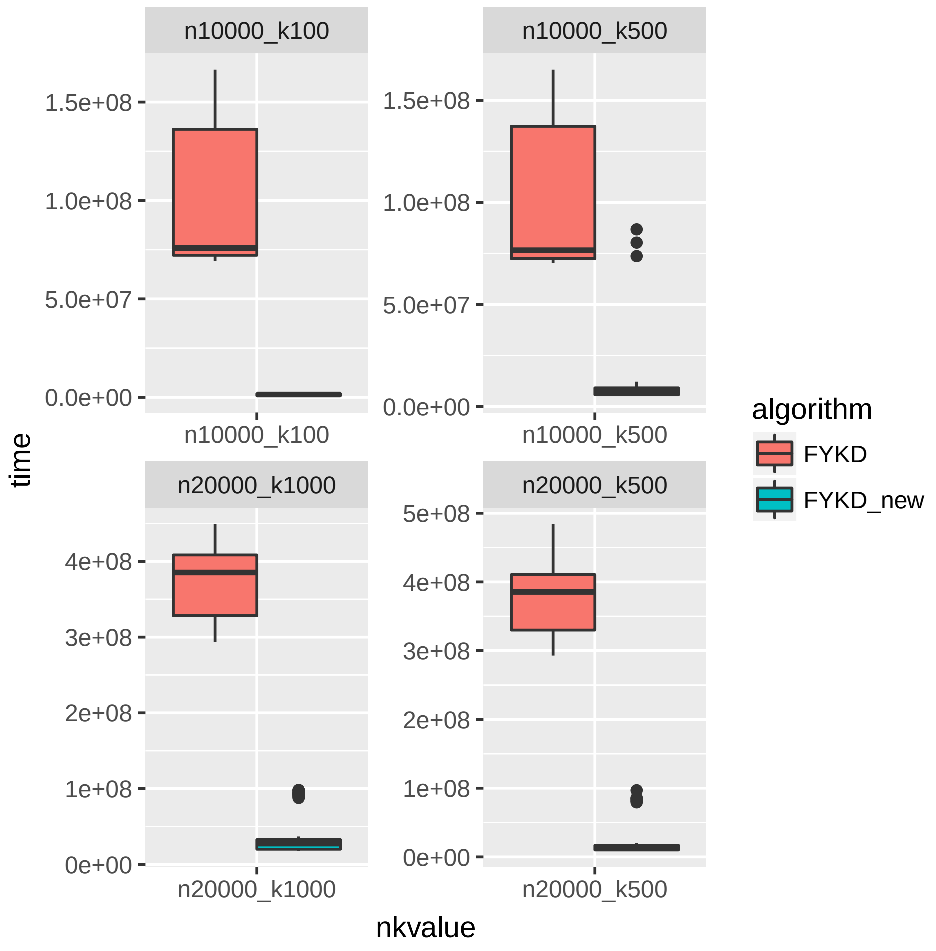
Then compare the revised code and original code in term of time needed to calculate the result.

```
####calculate time from different value of n,k. Find that microbenchmark does 100 simulations.
x=rnorm(10000)
FYKD_newtime_10000_500<-microbenchmark(FYKD_new(x,500))$time
FYKD_time_10000_500<-microbenchmark(FYKD(x,500))$time
FYKD_newtime_10000_100<-microbenchmark(FYKD_new(x,100))$time
FYKD_time_10000_100<-microbenchmark(FYKD(x,100))$time
x=rnorm(20000)
FYKD_newtime_20000_500<-microbenchmark(FYKD_new(x,500))$time
```

```
FYKD_time_20000_500<-microbenchmark(FYKD(x,500))$time
FYKD_newtime_20000_1000<-microbenchmark(FYKD_new(x,1000))$time
FYKD_time_20000_1000<-microbenchmark(FYKD(x,1000))$time
```

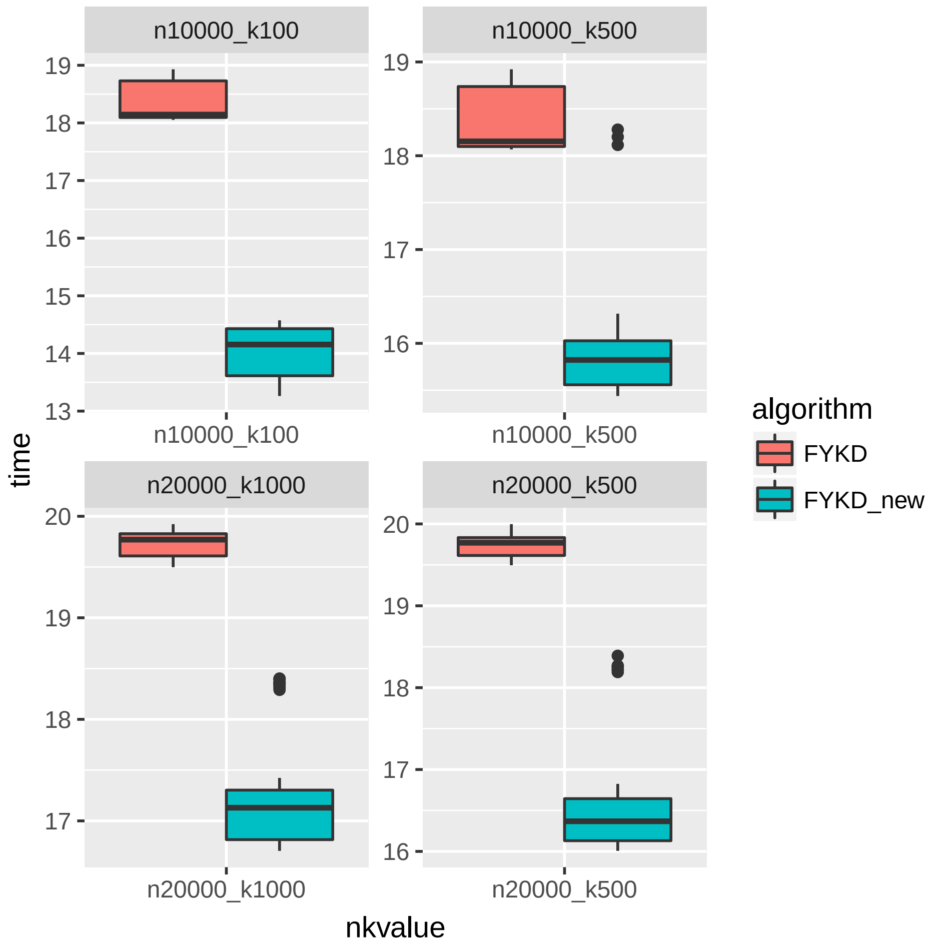
```
#####create dataframe needed to do a boxplot
data1=cbind(rep('n10000_k500',100),rep('FYKD_new',100),FYKD_newtime_10000_500)
data2=cbind(rep('n10000_k500',100),rep('FYKD',100),FYKD_time_10000_500)
data3=cbind(rep('n10000_k100',100),rep('FYKD_new',100),FYKD_newtime_10000_100)
data4=cbind(rep('n10000_k100',100),rep('FYKD',100),FYKD_time_10000_100)
data5=cbind(rep('n20000_k500',100),rep('FYKD_new',100),FYKD_newtime_20000_500)
data6=cbind(rep('n20000_k500',100),rep('FYKD',100),FYKD_time_20000_500)
data7=cbind(rep('n20000_k1000',100),rep('FYKD_new',100),FYKD_newtime_20000_1000)
data8=cbind(rep('n20000_k1000',100),rep('FYKD',100),FYKD_time_20000_1000)
data=rbind(data1,data2,data3,data4,data5,data6,data7,data8)
colnames(data)<-c('nkvalue','algorithm','time')
data=as.data.frame(data)
data[,3]=as.numeric(as.character(data[,3])) #change factor to numeric
```

```
###do a boxplot
p<-ggplot(data=data, aes(x=nkvalue,y=time))+geom_boxplot(aes(fill=algorithm))
p+ facet_wrap(~ nkvalue, scales="free")
```



It is obviously that new algoritnm costs much less time. To view a clearer plot, I do a log function on data.

```
###go a log function on data[,3],and do boxplot on data_log
data_log=data
data_log[,3]=log(data[,3])
p<-ggplot(data=data_log, aes(x=nkvalue,y=time))+geom_boxplot(aes(fill=algorithm))
p+ facet_wrap(~ nkvalue, scales="free")
```



We can find that revised algorithm performs better as n increases or k decreases.