BSG - Homework 7 - Relatedness analysis

HENRY QIU LO & MEYSAM ZAMANI

January 05, 2020

3. (1p) Read the genotype data in (0, 1, 2) format into the R environment. Consult the pedigree information. Are there any documented family relationships for this data set?

```
data <- read.delim("CHD.raw", header = TRUE, sep=" ")
```

The family relationship is documented in the variables 1,3 and 4 which are respectively the family id, the paternal id and the maternal id (there are also other variables that are not genetic variants are the individual id, sex and phenotype):

```
colnames(data)[1]
```

```
## [1] "FID"
```

```
colnames(data)[3]
```

```
## [1] "PAT"
```

```
colnames(data)[4]
```

```
## [1] "MAT"
```

However, in this dataset this information is not stored, because we can see both Paternal ids and maternal ids are empty:

```
#data[3]
#data[4]
```

4. (2p) Compute the Manhattan distance between the inviduals on the basis of the genetic data. Use classical metric multidimensional scaling to obtain a map of the indivuals. Are the data homogeneous? Identify possible outliers.

```
ncols <- ncol(data)
nrows <- nrow(data)
```

```
ncols
```

```
## [1] 28164
```

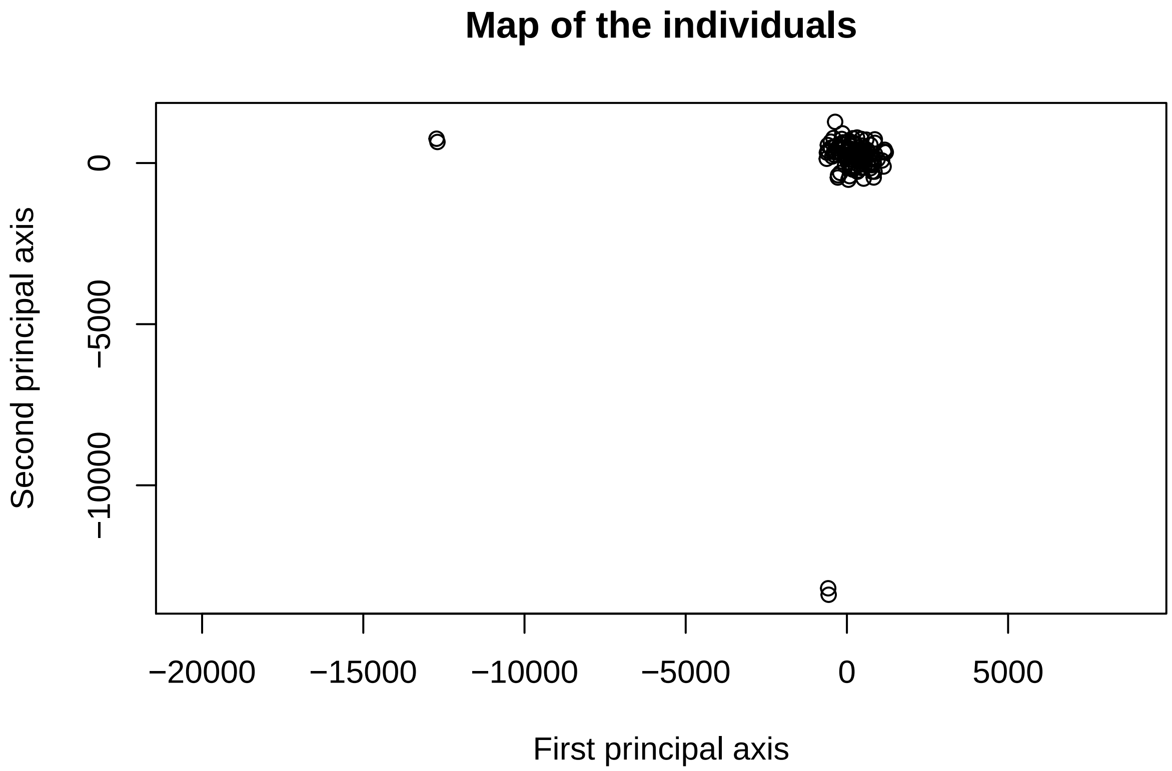
```
nrows
```

```
## [1] 109
```

```
manhattanDist <- as.matrix(dist(data[,7:ncols],method = "manhattan"))

muliDimScaling <- cmdscale(manhattanDist,k=nrows-1,eig=TRUE)
```

Map of the individuals:

```
X <- muliDimScaling$points
plot(X[,2],X[,1], xlab="First principal axis", ylab="Second principal axis", main="Map of the individual
```



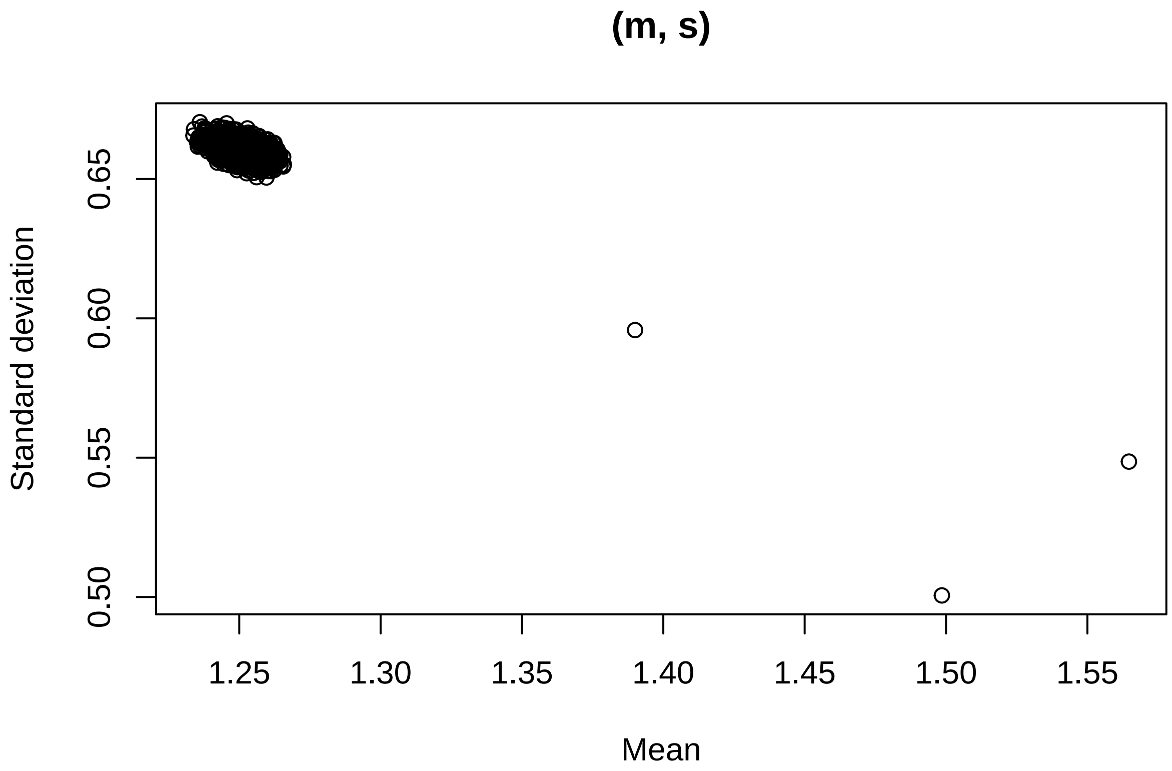
From the map of the individuals we can observe clearly the data is homogeneous in principle, the main part of the individuals are concentrated making a kind of “ball” near the origin, however it seems there are two pairs of outliers that are located far away from the concentration, placed at the left top and right bottom corner.

5. (2p) Compute the average number of alleles shared between each pair of individuals over all genetic variants. Compute also the corresponding standard deviation. Plot the standard deviation against the mean. Do you think there are any pairs with a close family relationship? How many? Identify the corresponding individuals.

```
ibs.mean <- function(x,y) {
  y <- mean(2 - abs(x - y),na.rm=TRUE)
  return(y)
}

ibs.sd <- function(x,y) {
  y <- sd(abs(x-y),na.rm=TRUE)
  return(y)
}

MeanMatrix <- matrix(NA,nrow=nrows,ncol=nrows)
SdMatrix <- matrix(NA,nrow=nrows,ncol=nrows)

famData <- as.matrix(data[,1:6])
genData <- as.matrix(data[,7:ncols])

for(i in 1:nrows) {
  for(j in 1:nrows) {
    MeanMatrix[i,j] <- ibs.mean(genData[i,],genData[j,])
    SdMatrix[i,j] <- ibs.sd(genData[i,],genData[j,])
  }
}

lowerTriMeanMatrix <- MeanMatrix[lower.tri(MeanMatrix)]
lowerTriSdMatrix <- SdMatrix[lower.tri(SdMatrix)]
```

Plot of the standard deviation deviation against the mean:

```
plot(main="(m, s)", lowerTriMeanMatrix,lowerTriSdMatrix,xlab="Mean",ylab="Standard deviation")
```



The plot of the standard deviation deviation against the mean reveals the characteristic clusters that correspond to different family relationships, because the position of these clusters depends on the allele frequencies. This means the 3 points that are far away from the concentration of points have a clear difference in the relationship value, the concentration of points might indicates there is few relation, this means the points well differentiated indicates there some kind of relation between the individuals discribed by that point.

Identifying the individuals:

```
ind1 = c()
ind2 = c()
for(i in 1:nrows) {
  for(j in i:nrows) {
    # The differentiated points have mean value greater than 1.35
    if (MeanMatrix[i,j] != 2 && MeanMatrix[i,j] > 1.35) {
      ind1 <- c(ind1,c(i))
      ind2 <- c(ind2,c(j))
    }
  }
}
```

The individuals that might have some kind of relationship are:

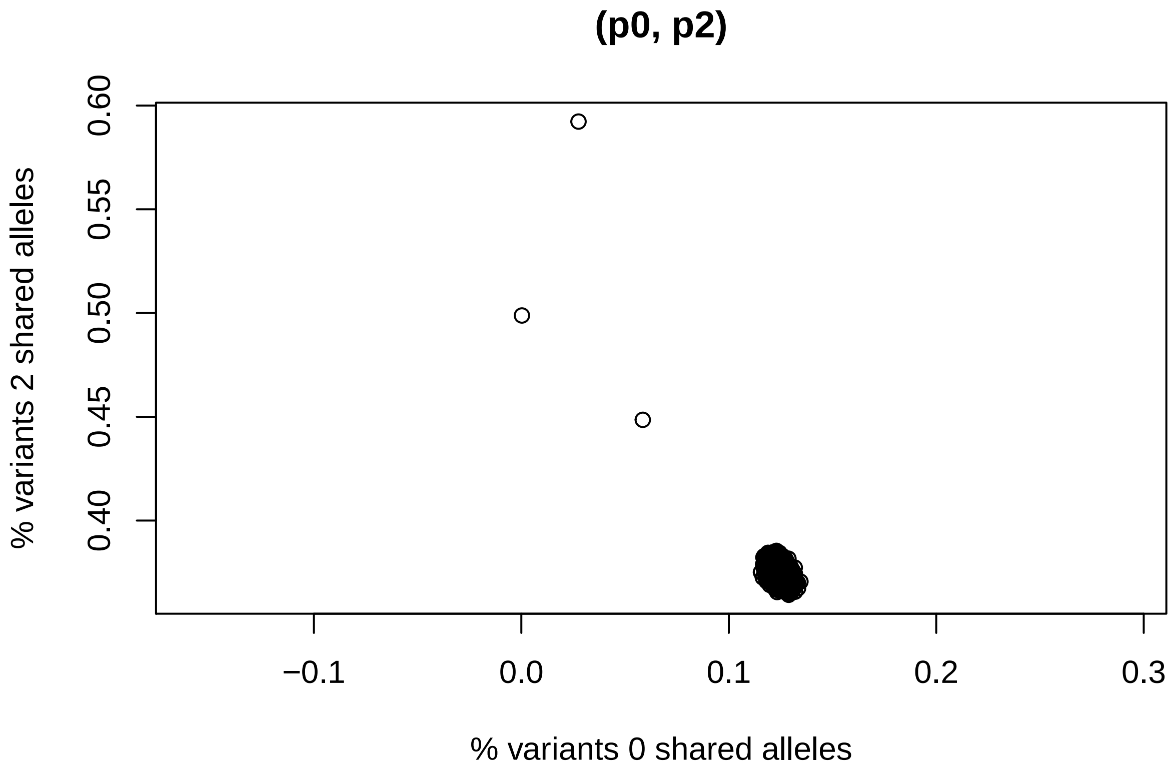
```
for(i in 1:length(ind1)) {
  print("Between individuals: ")
  print(famData[ind1[i],])
  print(famData[ind2[i],])
}
```

```
## [1] "Between individuals: "
##       FID       IID       PAT       MAT       SEX PHENOTYPE
## "NA17981" "NA17981"       "0"       "0"       "2"      "-9"
##       FID       IID       PAT       MAT       SEX PHENOTYPE
## "NA17986" "NA17986"       "0"       "0"       "1"      "-9"
## [1] "Between individuals: "
##       FID       IID       PAT       MAT       SEX PHENOTYPE
## "NA18150" "NA18150"       "0"       "0"       "2"      "-9"
##       FID       IID       PAT       MAT       SEX PHENOTYPE
## "NA17980" "NA17980"       "0"       "0"       "1"      "-9"
## [1] "Between individuals: "
##       FID       IID       PAT       MAT       SEX PHENOTYPE
## "NA17976" "NA17976"       "0"       "0"       "1"      "-9"
##       FID       IID       PAT       MAT       SEX PHENOTYPE
## "NA18116" "NA18116"       "0"       "0"       "2"      "-9"
```

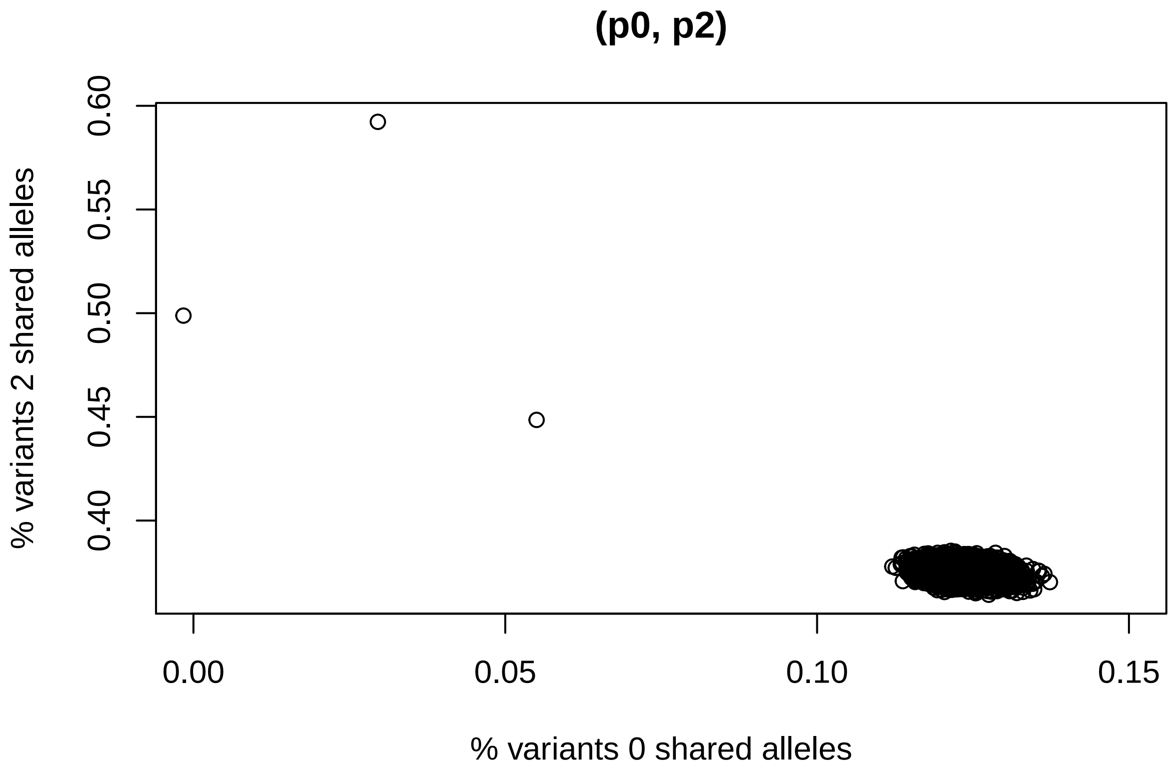
6. (1p) Make a plot of the percentage of variants sharing no alleles versus the percentage of variants sharing two alleles for all pairs of individuals. Do you think there are any pairs with a close family relationship? How many? Identify the corresponding individuals.

```
stats <- function(x,y) {
  aux <- 2-abs(x-y) # number of shared alleles
  n0 <- sum(aux==0,na.rm=TRUE)
  n1 <- sum(aux==1,na.rm=TRUE)
  n2 <- sum(aux==2,na.rm=TRUE)
  n <- sum(!is.na(aux))
  p0 <- n0/n
  p1 <- n1/n
  p2 <- n2/n
  y <- c(p0,p1,p2)
  return(y)
}

Mp0 <- matrix(NA,nrow=nrows,ncol=nrows)
Mp1 <- matrix(NA,nrow=nrows,ncol=nrows)
Mp2 <- matrix(NA,nrow=nrows,ncol=nrows)
for(i in 1:nrows) {
  for(j in 1:nrows) {
    statsofapair <- stats(genData[i,],genData[j,])
    Mp0[i,j] <- statsofapair[1]
    Mp1[i,j] <- statsofapair[2]
    Mp2[i,j] <- statsofapair[3]
  }
}

p0vec <- Mp0[upper.tri(Mp0)]
p2vec <- Mp2[upper.tri(Mp2)]
```

```
plot(main="(p0, p2)", p0vec,p2vec,asp=1,
     xlab="% variants 0 shared alleles",
     ylab="% variants 2 shared alleles")
```



```
plot(main="(p0, p2)", jitter(p0vec,amount=.005),p2vec,xlim= c(0,0.15),
     xlab="% variants 0 shared alleles",
     ylab="% variants 2 shared alleles")
```



Same that before, from the plots we can observe there are three points clearly separated from the concentrated cloud of points. These points are located now at the left hand side of the plot.

Identifying the individuals:

```
ind3 = c()
ind4 = c()
for(i in 1:nrows) {
  for(j in i:nrows) {
    # The differentiated points have value greater than 0.4
    if (Mp2[i,j] != 1 && Mp2[i,j] > 0.4) {
      ind3 <- c(ind3,c(i))
      ind4 <- c(ind4,c(j))
    }
  }
}
```

The individuals that might have some kind of relationship are:

```
for(i in 1:length(ind3)) {
  print("Between individuals: ")
  print(famData[ind3[i],])
  print(famData[ind4[i],])
}
```

```
## [1] "Between individuals: "
##       FID       IID       PAT       MAT       SEX PHENOTYPE
## "NA17981" "NA17981"       "0"       "0"       "2"      "-9"
##       FID       IID       PAT       MAT       SEX PHENOTYPE
## "NA17986" "NA17986"       "0"       "0"       "1"      "-9"
## [1] "Between individuals: "
##       FID       IID       PAT       MAT       SEX PHENOTYPE
## "NA18150" "NA18150"       "0"       "0"       "2"      "-9"
##       FID       IID       PAT       MAT       SEX PHENOTYPE
## "NA17980" "NA17980"       "0"       "0"       "1"      "-9"
## [1] "Between individuals: "
##       FID       IID       PAT       MAT       SEX PHENOTYPE
## "NA17976" "NA17976"       "0"       "0"       "1"      "-9"
##       FID       IID       PAT       MAT       SEX PHENOTYPE
## "NA18116" "NA18116"       "0"       "0"       "2"      "-9"
```

7. (1p) Can you identify any obvious family relationships between any pairs? Argue your answer.

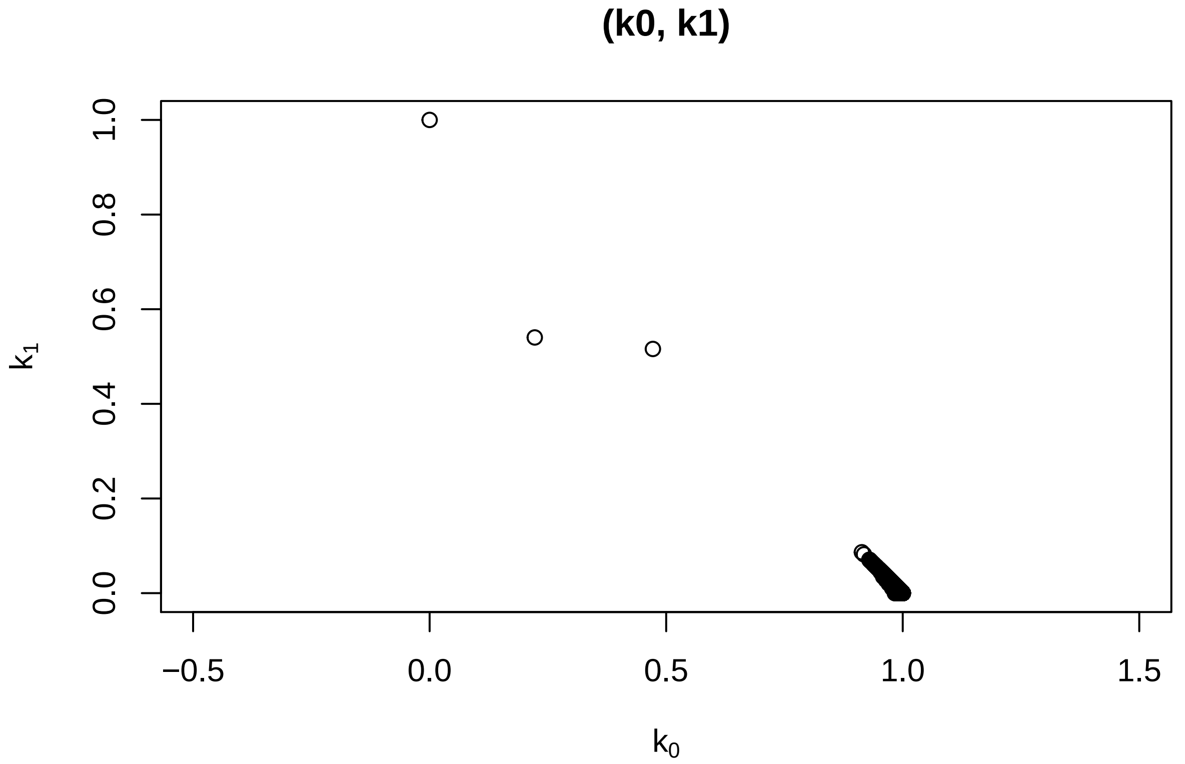
We can observe from the results above the individuals obtained are the same, hence, we can conclude with more confidence there are certainly a obvious family relationships between these three pairs of individuals.

8. (2p) Estimate the Cotterman coefficients for all pairs using PLINK. Read the coeffients into the R environment and plot the probability of sharing no IBD alleles against the probability of sharing one IBD allele. Add the theoretical values of the Cotterman coefficients for standard relationships to your plot.

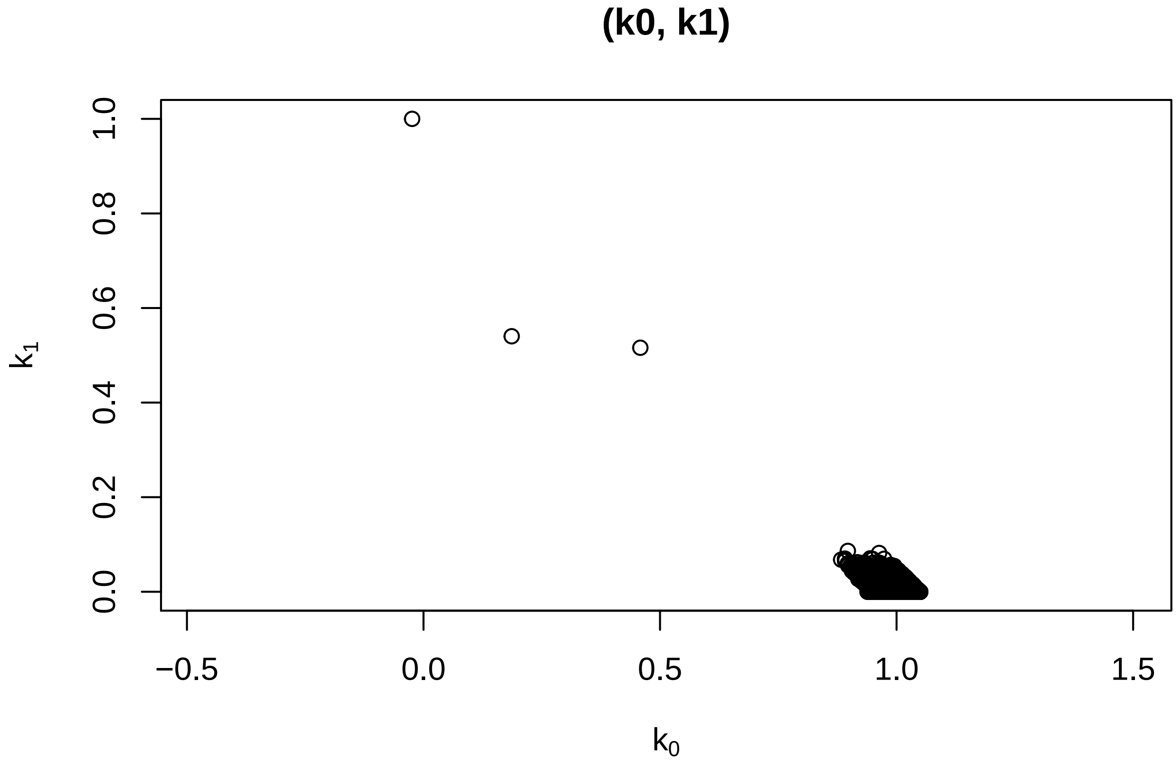
To generate the .genome file we used the command line instruction: plink –bfile CHD –genome –genome-full –out CHD

```
genomeData <- read.table("CHD.genome",header=TRUE)
```

```
plot(main="(k0, k1)", genomeData$Z0,genomeData$Z1,asp=1,xlab=expression(k[0]),ylab=expression(k[1]))
```



```
plot(main="(k0, k1)", jitter(genomeData$Z0,amount=0.05),genomeData$Z1,asp=1,xlab=expression(k[0]),ylab=
```



From the plot we can observe the point placed at top left corner has full probability of sharing one IBD allele. The other two points have nearly 50% of probability to share an allele, then for the remaining points at the right bottom corner the probability of no sharing allele is very high, this coincide with our expectation.

9. (2p) Make a table of pairs for which you suspect that they have a close family relationship, and list their Cotterman coefficients. State your final conclusions about what relationship these pairs probably have.

```
tableMatrix <- matrix(NA,nrow=length(ind3),ncol=4)
for(i in 1:nrow(genomeData)) {
  for (k in 1:length(ind3)) {
    if (genomeData[i,][2] != genomeData[i,][4] && ((genomeData[i,][2] == famData[ind3[k],][2]) && (genom
        || (genomeData[i,][2] == famData[ind4[k],][2]) && (genomeData[i,][4] == famData[ind3[k],][2]))
    ) {
      tableMatrix[k,1] <- paste(famData[ind3[k],][2],famData[ind4[k],][2],sep=',')
      tableMatrix[k,2] <- genomeData[i,7]
      tableMatrix[k,3] <- genomeData[i,8]
      tableMatrix[k,4] <- genomeData[i,9]
    }
  }
}
```

```
library(data.table)
```

```
table <- data.table(tableMatrix)
setnames(table,c("Relations","Z0","Z1","Z2"))
```

```r
table
```

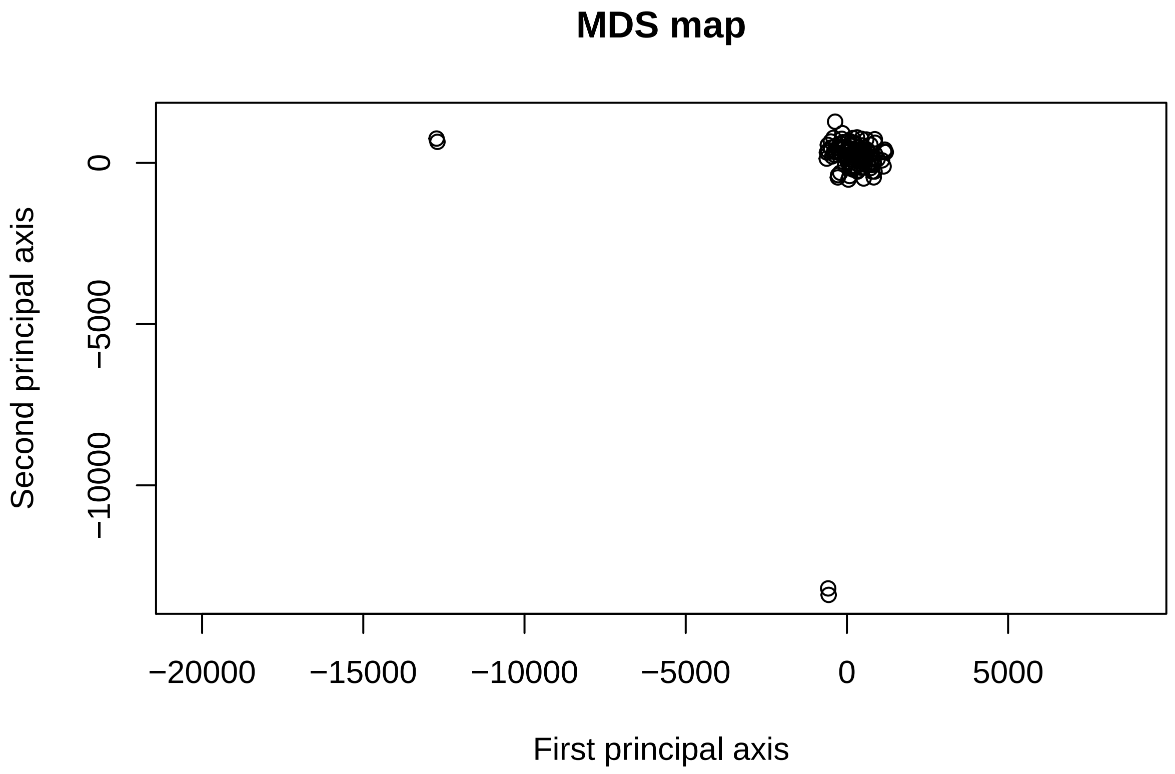
```
##          Relations     Z0     Z1     Z2
## 1: NA17981,NA17986 0.2222 0.5403 0.2376
## 2: NA18150,NA17980 0.4719  0.516 0.0121
## 3: NA17976,NA18116      0      1      0
```

From the table we can see the certainty of the relationship between the individuals with id NA17976 and NA18116 has a 100% of probability of sharing 1 allele. For the other 2 pairs, the probability of sharing 1 allele is also the highest. This indicates us the relationship between them might be of father/mother - son/daughter.

10. (2p) Is there any relationship between the MDS map you made and the relationships between the individuals? Report your findings.

Reviewing the MDS map:

```r
plot(X[,2],X[,1], xlab="First principal axis", ylab="Second principal axis", main="MDS map",asp=1)
```



We can observe clear relationship between the map with the relationships with the individuals, there are clearly a cloud of individuals with simmilar genetic values, and we can observe two pairs of individuals clearly differentiated to the cloud but very simmilar to each other, these must be two of the pairs we found in our relationship analysis. It is logic to think individuals very different to the others but very simmilar to another can be family members.

11. (2p) Which of the three graphics (m, s), (p0, p2) or (k0, k1) do you like best for identifying relationships? Argue your answer.

To identify relationships we (k0, k1) plot, because the information it represents is directly an IBD information, which is our main objective of study. What is more, the plot gives us clearly the probabilities of allele sharing, this gives us more certainty in the estimation of relation between the individuals. For example in our plot we had 3 pairs of suspicion of family relationship, with the plot we can be much more sure which pairs have high probability of being family members, we can also see this probability is clearly higher than the other two pairs, at the same time we can be confident that the cloud of “no related individuals” are no related for sure.