R-Blatt 4: Visualisierung statistischer Konzepte

Statistical Aspects (09-202-2413)

Janne Pott

Last compiled on 15 September, 2022

Session Setup

```
rm(list = ls())
time0<-Sys.time()

source("../sourceFile.R")
setwd(pathToExercise)

knitr::opts_chunk$set(echo = TRUE)
```

Verwandschaftsmatrix

Bitte laden Sie das Objekt *verwandtschaft.RData* mittels *load()*. Das R-Objekt enthält zwei Datensätze: *genotypes*, eine Genotyp-Matrix von 10 Personen mit 30,000 SNPs, und *allelfreq*, ein Vektor der Allelfrequenzen pro SNP aus *genotypes* bezüglich Allel B.

a) Die Verwandtschaftsmatrix kann man mittels Matrix-Operation bestimmen. Definieren Sie dazu eine Hilfsmatrix h, die auf die Allelfrequenzen und Anzahl der SNPs adjustiert ist, und bilden Sie das Matrixprodukt $H = h^T * h$.

$$h_{m,i} = \frac{(g_{m,i} - 2 * p_{m,B})}{\sqrt{M * 4 * p_{m,B} * p_{m,A}}}$$

b) Warum gilt:

$$\hat{H}_{i,i} \approx 0.5$$

c) Wie viele paarweise Verwandtschaften (von Grad 1,2, ... , unverwandt) beobachten Sie?
d) Welche Familienstruktur könnte die beobachteten Verwandtschaftsbeziehungen erklären?

XY-Plots

Bitte laden Sie das Objekt *XYPlots.RData* mittels *load()*. Das R-Objekt enthält zwei Datensätze: *daten*, eine Matrix von 300 Personen und 300 gonosomalen SNPs (*X:i* bzw. *Y:i*), wobei die SNPs sowohl als Genotypen und Intensitäten pro Allel gegeben sind (AA = 0, AB = 1, BB = 2), und *samples*, ein Tabelle der gemessenen Individuen aus *daten* mit Geschlechtsinformationen.

Die Heterozygosität eines Samples ist der Anteil der AB Genotypen. Das genetische Geschlecht kann typischerweise über die Chr. X Heterozygosität bestimmt werden, da Männer in der Regel hier nur haploid

sind. Zusätzlich können die Bindungsintensitäten von X- und Y-SNPs mit berücksichtigt werden. In wenigen Ausfällen ist die Intensität der X- bzw. Y-SNPs zu verrauscht um eine Aussage zu treffen. Diese werden dann als *unknown* klassifiziert.

a) Für den XY-Plot brauchen wir die Gesamtintensitäten pro Sample. Bilden Sie daher zuerst den Mittelwert der Intensität für Allel A und B pro SNP und bilden Sie dann die jeweilige mittlere Intensität aller X-SNPs und aller Y-SNPs pro Sample!
b) Zusätzlich brauchen wir die Heterozygosität auf Chromosom X. Bestimmen Sie dazu pro Sample die Genotyphäufigkeiten (AA, AB, BB) aller SNPs von Chromosom X und berechnen Sie den Anteil von AB an allen Genotypen!
c) Sie sollten nun ein Objekt mit den Variablen *ID*, *X-Intensität*, *Y-Intensität, X-Heterozygosität*, *sex_datenbank* und *sex_computed* pro Sample haben. Erzeugen Sie nun folgende drei Plots und markieren Sie in diesen Plots Ausreißer (widersprüchliche Geschlechtsangabe, zu hohe/niedrige Intensitäten, auffällige Heterozygosität):
    i. X-Intensität – Y-Intensität
    ii. X-Intensität – X-Heterozygosität
    iii. Y-Intensität – X-Heterozygosität

Principal Component Analysis (PCA)

In dieser Aufgabe sollen die Eigenvektoren von genetischen Daten erzeugt werden. Dafür wird ein Teil der Daten von 1000 Genomes (1KG, Phase 1, release 3) verwendet. Zusätzlich zu R wird hier PLINK verwendet, da dieses Programm effizienter große Datenmengen verarbeiten kann.

Diese Aufgabe ist als Tutorial aufgebaut: jeder Schritt wird angefangen und soll von Ihnen vervollständigt werden.

Als erstes soll der Pfad zu PLINK definiert werden. Als kleiner Test wird Plink einmal aufgerufen.

- Falls 127 zurückggeben wird, hat R die Plink .exe nicht gefunden - bitte Pfad prüfen!

```
plink_call<-pathToPLINK2

# test if plink can start
system(plink_call)
```

Datenvorbereitung - SNPs filtern

Überprüfen Sie in R, ob alle SNPs von *mySNPs.txt* in 1KG sind. Hierfür sollte man am besten das *1KG_PCA.bim* File verwenden (tab-delimited). Recherchieren Sie, was in den einzelnen Spalten des .bim Files steht. Nutzen sie zum Einlesen der Befehl *fread()* aus dem Paket *data.table*. Filtern Sie nach den SNPs in der Schnittmenge und erstellen Sie ein gefiltertes Text-File mySNPs_filtered.txt!

Hinweis: Es sollten am Ende 206,233 SNPs sein!

```
myTab<-read.table(paste0(pathToData,"mySnps.txt"))
dim(myTab)
```

```
## [1] 224458      1
```

```
rslist<-fread(paste0(pathToData,"1KG_PCA.bim"),sep="\t",stringsAsFactors=F)
dim(rslist)
```

```
## [1] 498586      6
```

```
head(rslist)
```

```
##    V1         V2 V3      V4 V5 V6
## 1:  1  rs3094315  0  752566  G  A
## 2:  1 rs12184325  0  754105  T  C
## 3:  1  rs3131969  0  754182  A  G
## 4:  1 rs12562034  0  768448  A  G
## 5:  1 rs11240777  0  798959  A  G
## 6:  1 rs11579015  0 1036959  C  T
```

```
# to do: filter for SNPs in overlap and save as mySNPs_filtered.txt
```

Datenvorbereitung - Samples filtern

Erstellen Sie eine Sample Liste mit Individuen aus Asien, Afrika und Europa! Nutzen Sie hierfür das **1KG_PCA.fam** File (space-delimited). Wir wollen eine möglichst große Menge an Samples, aber jeder Herkunft sollte gleich oft vorhanden sein! Ziehen Sie zufällig aus der jeweiligen Teilmenge und speichern Sie ihre Liste als **mySamples.txt** ab!

Hinweis: Es sollten am Ende 3*246 Individuen sein!

```
fam.data<-read.table(paste0(pathToData,"1KG_PCA.fam"),stringsAsFactors=F,sep=" ")
dim(fam.data)
```

```
## [1] 1092    6
```

```
head(fam.data)
```

```
##        V1          V2 V3 V4 V5 V6
## 1 HG00096 EUR_HG00096  0  0  1 -9
## 2 HG00097 EUR_HG00097  0  0  2 -9
## 3 HG00099 EUR_HG00099  0  0  2 -9
## 4 HG00100 EUR_HG00100  0  0  2 -9
## 5 HG00101 EUR_HG00101  0  0  1 -9
## 6 HG00102 EUR_HG00102  0  0  2 -9
```

```
ethno<-substr(fam.data$V2,1,3)
table(ethno)
```

```
## ethno
## AFR AMR ASN EUR
## 246 181 286 379
```

```
# to do: choose randomly individuals from AFR, ASN and EUR to obtain
#        same sample size per ethnicity and save as mySamples.txt
```

Datenvorbereitung - SNPs prunen

Jetzt prunen Sie die SNPs mit PLINK, d.h. Sie prüfen, welche SNPs in hohem LD miteinander sind. Folgende Parameter sollten Sie setzen: a. Input:–bfile 1KG_PCA b. SNPs einschränken: –extract mySNPs_filtered.txt c. Samples einschränken: –keep mySamples.txt d. LD-Pruning-Parameter festlegen: –indep-pairwise 50 5 0.2 e. Output: –out pruned_filter

Was bedeuten die drei Zahlen hinter dem –indep-pairwise Befehl?

Hinweis: Es sollten am Ende 117,351 SNPs sein.

```
call1<-paste(plink_call,
             "--bfile ",pathToData,"/1KG_PCA",
             "--extract ",pathToData,"/mySnps_filtered.txt",
             "--keep ",pathToData,"/mySamples.txt",
             "--indep-pairwise 50 5 0.2",
             "--out ",pathToData,"/pruning_filter")
system(call1)
```

Datenvorbereitung - Datensatz erstellen

Erstellen Sie jetzt mit PLINK ein neues .bed-File, dass nur noch die geprunten SNPs und die gewünschten Samples (aus 2) enthält (–bfile, –extract, –keep, und –make-bed).

```
call2<-paste(plink_call,
             "--bfile ",pathToData,"1KG_PCA",
             "--extract ",pathToData,"pruning_filter.prune.in",
             "--keep ",pathToData,"mySamples.txt",
             "--make-bed --out ",pathToData,"pruned_data")
system(call2)
```

PCA berechnen

Jetzt kann mit den neuen Files die PCA ausgerechnet werden (–bfile, –pca, –out):

```
call3<-paste(plink_call,
             "--bfile ",pathToData,"pruned_data",
             "--pca --out ",pathToData,"pca_out")
system(call3)
```

PCA auswerten

Laden Sie beide Outputs der PCA in R ein! Wie sind die Daten aufgebaut?

Erstellen Sie einen Plot der ersten beiden Vektoren mit Ethnien-Färbung! Was kann man daraus schließen?

Berechnen Sie den Anteil der erklärten Varianz a. durch den ersten Eigenvektor und b. durch die ersten beiden Eigenvektoren!

Was würden Sie erwarten, wenn alle 4 Ethnien in die Analyse eingeflossen wären? Wo würden Sie die Amerikaner einordnen? Rechnen Sie das nach!

```
pca2values<-read.table(paste0(pathToData,"pca_out.eigenval"))$V1
pca2vector<-read.table(paste0(pathToData,"PCA/pca_out.eigenvec"),stringsAsFactors=F,sep="\t")

# to continue ...
```