# Genomics

Freie Universität Berlin, Institut für Informatik
Knut Reinert, Peter Robinson, Sebastian Bauer
Wintersemester 2012/2013

8. Übungsblatt vom 11. Dezember 2012
Diskussion im kommenden Jahr (to be determined)

---

*Exercise 1.*

In this exercise, you will be guided through the PCA analysis of the "cars" dataset as shown in the lecture. You will need to download the file "cars-fixed04.dat" to do the exercise.

```
# Look at the correlations
library(lattice)
library(gclus)
```

Note: You may need to install these libraries!

Read the data

```
cars = read.csv("cars-fixed04.dat")
head(cars[,8:18])
```

Look at the correlations

```
my.abs = abs(cor(cars[,8:18]))
my.colors = dmat.color(my.abs)
my.ordered = order.single(cor(cars[,8:18]))
cpairs(cars[,8:18],my.ordered,panel.colors=my.colors,gap=.5)
```

Describe the correlation structure of the data? What attributes are most highly correlated?

*Exercise 2.*

Do the pca. First have a look at ?prcomp in R.

```
 cars.pca = prcomp(cars[,8:18], scale.=TRUE)
```

Do the scree plot

```
screeplot(cars.pca,main="Scree Plot",xlab="Components")
```

Make a nicer plot

```
screeplot2<-function(mydata,cor=F,maxcomp=10) {
my.pc<-prcomp(mydata, scale=cor)
k<-min(dim(mydata),maxcomp)
x<-c(0:k)
y<-my.pc$sdev[1:k]*my.pc$sdev[1:k]
```

```
y<-c(0,y)
z<-100*cumsum(y)/sum(my.pc$sdev*my.pc$sdev)

plot(x,z,type="l",xlab="number of dimensions",
 cex.main=1.5, lwd=3, col="red",
 ylim=c(0,100),
 ylab="cumulative percentage of total variance",
  main="Scree plot of variances",
 xaxt="n", yaxt="n")

axis(1,at=x,lwd=2)
axis(2,at=c(0,20,40,60,80,100),lwd=2)
abline(a=100,b=0,lwd=2,lty="dashed",col="orange")
text(x,z,labels=x,cex=0.8,adj=c(1.2,-.1),col="blue")
}


par(mfrow=c(1,2))

screeplot(cars.pca,main="Scree Plot",xlab="Components")

screeplot2(cars[,8:18],cor=T)
```

Question: How many PCs are needed to capture 75% of the variance? 95%?

*Exercise 4.*

In this exercise we will plot the loadings. PC1:

```
load = cars.pca$rotation
sorted.loadings = load[order(load[,1]),1]
Main="Loadings Plot for PC1"
xlabs="Variable Loadings"
dotplot(sorted.loadings,main=Main,xlab=xlabs,cex=1.5,col="red")
```

Question: Interpret PC1. What "meaning" does it have?

```
load = cars.pca$rotation
sorted.loadings = load[order(load[,2]),2]
Main="Loadings Plot for PC2"
xlabs="Variable Loadings"
dotplot(sorted.loadings,main=Main,xlab=xlabs,cex=1.5,col="red")
```

Now interpret PC2. What "meaning" does it have? Have a look at other PCs. Do you see anything interesting?

*Exercise 4.*

Plot the loadings against one another

```
par(mar=c(5.1, 5.1, 2.1, 2.1)) ## increase left margin from 4.1 to 5.1
load = cars.pca$rotation
PC1 = load[order(load[,1]),1]
PC2 = load[order(load[,2]),2]
plot(PC1,PC2,pch=18,col="blue",cex.lab=1.5)
grid()
n<-length(PC1)
arrows(rep(0,n),rep(0,n),PC1,PC2,length=0.1,col="red")
points(0,0,pch=10,col="blue")
```

See any interesting patterns?

*Exercise 4.*

After exploring the "cars" dataset, try to understand the use of PCA in this paper:

Ouyang Z, Zhou Q, Wong WG (2009) ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *PNAS* **106**:21251-21526

(available at www.ncbi.nlm.nih.gov/pubmed/19995984). In particular, read the section TFPCs Provide Information on the Roles of Regulators. Do you agree with the conclusions of the authors? Describe in your own words the "roles" of PC1 and PC2.