Week 4. 3D plots and animation, correlation heatmap and partial correlation

3D scatterplots and animation in R. Correlation heatmap, partial correlation and its analysis for stock prices returns (finance.yahoo.com). The multidimensional analysis of osteometric skeleton measurements.

R codes: `iris3D`, `stock.price, skelet`, `mah, parsROC`

Data: `iris` (built-in), `Iris.pptx`, `Most active stocks prices.csv`, `GoldMeasures.pdf,Goldman.csv`

3D plots and 360° animation

Section 4.1.1

The built-in function `persp` for 3D graphics

R function `mn3`

Famous R. Fisher iris discrimination analysis: `Iris.pptx`

`iris3D`

Correlation heatmap

Section 6.6.2

The time series analysis of stock prices.

The stock return

$$r_t = \frac{p_t - p_{t-1}}{p_{t-1}}.$$

From statistical perspective, it's easier to deal with log transformed stock price $\ln p_t$. Then the difference on the log scale is close to r_t, that is,

$$\Delta_t = \ln p_t - \ln p_{t-1} = \ln \frac{p_t}{p_{t-1}} = \ln \left(1 + \frac{p_t - p_{t-1}}{p_{t-1}}\right) \simeq \frac{p_t - p_{t-1}}{p_{t-1}} = r_t$$

due to calculus approximation

$$\ln(1 + x) \simeq x.$$

How to visualize the correlation in the multivariable data set? If $\mathbf{X}$ is the data matrix with n rows (observations) and m columns (variables or features) use cor(X) to obtain $m \times m$ correlation coefficient matrix between features.

The built-in `image` function and correlation heatmap – color represents value.

See the R function `cimcorSP`

The R function `cor(X)`.

stock.price

The time series analysis of 87 daily stock prices, `stock.price`.

Partial correlation

Section 4.4.2 and 6.6.2

If $\mathbf{x}_i$ is feature i and $\mathbf{x}_j$ is feature j high correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$ may be due to the presence of common factor.

Example 1 *Let $X = Z + U$ and $Y = Q + U$ where Z, Q, and U are independent and U is the common factor. Show that the correlation between X and Y can be high due to the presence of the common factor U.*

Solution. Without loss of generality we assume that all random variables have zero mean and $var(Z) = var(Q) = 1$ but $var(U) = \sigma^2$. Then

$$\begin{aligned} cov(X,Y) &= cov(Z+U, Q+U) = cov(Z,Q) + cov(Z,U) + cov(U,Q) + cov(U,U) \\ &= 0 + 0 + 0 + var(U) = \sigma^2, \\ var(X) &= var(Z) + var(U) = 1 + \sigma^2, \\ var(Y) &= var(Q) + var(U) = 1 + \sigma^2. \end{aligned}$$

Thus we have

$$\rho = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} = \frac{\sigma^2}{1+\sigma^2}.$$

The variance σ^2 reflects contribution of the common factor. Letting the contribution to infinity we obtain

$$\lim_{\sigma^2 \to \infty} \rho = 1.$$

Conclusion: correlation between X and Y becomes high due to the presence of a common factor.

Every time you see an expected high correlation it may be caused by the presence of a common factor that influences both variables.

Example: Revenue and truck drives (common factor is time).

Example 2 *High correlation between blood pressure and poor heart functioning may be explained by a common factor: unhealthy lifestyle. After conditioning on the lifestyle the correlation drops significantly.*

How condition on the common factor? Answer: **partial correlation**.

Definition 3 *The conditional or partial correlation is the correlation between X and Y in the **conditional** bivariate distribution of (X, Y) conditioned on Z, and can be computed as*

$$\rho_{XY|Z} = -\frac{\rho^{12}}{\sqrt{\rho^{11}\rho^{22}}},$$

where ρ^{ij} are the (i,j)th elements of the inverse correlation matrix,

$$\begin{bmatrix} 1 & \rho_{XY} & \rho_{XZ} \\ \rho_{XY} & 1 & \rho_{YZ} \\ \rho_{XZ} & \rho_{YZ} & 1 \end{bmatrix}^{-1} = \begin{bmatrix} \rho^{11} & \rho^{12} & \cdot \\ \rho^{12} & \rho^{22} & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}.$$

*Alternatively, partial correlation coefficient can be computed as the correlation between **residuals** from regressions X on Z and Y on Z. This is what we did in Revenue vs Truck drivers example.*

Example 4 *If in the above example where, $X = Z + U$ and $Y = Q + U$, we have*

$$\rho_{XY|U} = 0$$

because conditional on U (fixed) X and Y are independent because Z and Q are independent.

How to compute partial correlation:

1. Compute correlation matrix $\mathbf{R} = \{r_{ij}\}$ via `cor` function.
2. Inverse $\mathbf{R}^{-1} = \{r^{ij}\}$ where the superscript indicates the element of the inverse matrix.
3. Compute

$$R_{ij}^{(p)} = -\frac{r^{ij}}{\sqrt{r^{ii} r^{jj}}}$$

and set $R_{ii}^{(p)} = 1$.

4. Interpretation $R_{ij}^{(p)}$ is the partial (pure) correlation between predictors i and j conditional on other features.

`cimcorSP(job=3)`

Correlation and partial correlation heatmap for stock returns

Section 6.6.2

Return of the stock is

$$r_t = \frac{X_t - X_{t-1}}{X_{t-1}},$$

where X is the price. From a statistical perspective, it is easier to use an approximation

$$r_t \simeq \ln \frac{X_t}{X_{t-1}}$$

$$\frac{X_t}{X_{t-1}} = 1 + \frac{X_t - X_{t-1}}{X_{t-1}}.$$

Use approximation

$$u \simeq \ln(1 + u)$$

and therefore

$$r_t \simeq \ln\left(1 + \frac{X_t - X_{t-1}}{X_{t-1}}\right) = \ln \frac{X_t}{X_{t-1}}$$

See function `stock.price`

The R package pheatmap

See `stock.price(job=1.01)`

Skeleton correlations

Archeology: human remains

`https://web.utk.edu/~auerbach/GOLD.htm`

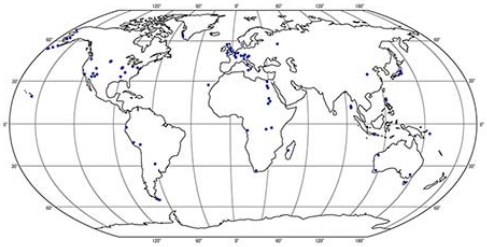
https://web.utk.edu/~auerbach/GOLD.htm

BENJAMIN M. **AUERBACH**, Ph.D.

GOLDMAN OSTEOMETRIC DATA SET

HOME PUBLICATIONS DATA SETS COURSES RESEARCH

PLEASE READ THIS PAGE BEFORE DOWNLOADING THE DATA SET



ABOUT THE DATA

The Goldman Data Set consists of osteometric measurements taken from 1538 human skeletons dating from throughout the Holocene. Links to the data are at the bottom of this page. Links to sample provenience and dating information may be found in links below. Measurements were taken bilaterally from four of the long bones: humerus, radius, femur, and tibia. Three measurements were additionally obtained from the pelvis. Sex and age were estimated as well from pelvic observations (see Auerbach and Ruff, 2004, and Auerbach and Ruff, 2006, for a description of the methods employed).

Dr. Benjamin Auerbach obtained all of these data during three solo research trips beginning in September 2001 and ending in July 2003. These trips were made possible as part of a research fellowship provided by generous funding from the Joanna Jackson Goldman Memorial Prize; the data set is named after this organization. Dr. Auerbach continues to be grateful to the staff and curators at the institutions in the United Kingdom, Italy, Austria, Germany, France, Belgium, the United States of America, and Japan for allowing permission to work with their collections. A list of source institutions may be found below.

These data are now made available to researchers for download. The data are available below as a Microsoft Excel legacy file and as a comma-separated text file. A supporting document describing the osteometric measurements is also provided as a PDF file. Additional provenience information for the samples is currently available by request from Dr. Auerbach.

ALL RESEARCHERS WHO PLAN TO USE THESE DATA ARE WELCOMED AND ENCOURAGED TO CONTACT DR. AUERBACH. ANY USE OF THESE DATA IN PRESENTED OR PUBLISHED RESEARCH CARRIES THE STIPULATION THAT THE SOURCE OF THE DATA BE CITED. ACCEPTABLE CITATIONS FOR THE DATA INCLUDE THE REFERENCE OF THE DATA'S ANALYSIS (AUERBACH & RUFF, 2004 or 2006) AND OF THIS WEB SITE.

ADDITIONAL INFORMATION ABOUT THE DATA

Description of the data set and measurements: PDF file

General dates of the sites sampled: Excel file

SPECIAL THANKS TO THE FOLLOWING INSTITUTIONS FOR PERMITTING ACCESS TO THEIR COLLECTIONS:

American Museum of Natural History, New York
Cleveland Museum of Natural History, Cleveland
Duckworth Collection, The University of Cambridge (LCHES), Cambridge
Institut Royal des Sciences Naturelles de Belgique, Brussels
Kent State University Department of Anthropology, Kent
Kyoto University (Kyodai), Kyoto
Musee de l'Homme, Paris
Muzeo Nationale di Antropologia e Etnologia, Florence
National Museum of Natural History (Smithsonian Institution), Washington, D.C.
Natural Histsory Museum, London
Naturhistorisches Museum, Vienna
Staatssammlung fur Anthropologie und Palaeoanatomie, Munich
Webb Osteology and Archaeology Collection, University of Kentucky, Lexington

THE GOLDMAN OSTEOMETRIC DATA SET

A GUIDE TO THE MEASUREMENTS

COLUMN HEADING	MEASUREMENT	DESCRIPTION
Inst	Collection housing remains	AMNH – American Museum of Natural History CMNH – Cleveland Museum of Natural History DC – Duckworth Collection IRSN – Institut Royal des Sciences Naturelles de Belgique KSU – Kent State University KU – Kyoto University (Kyodai) MdH – Musee de l'Homme MNdAE – Muzeo Nationale di Antropologia e Etnologia NHM – Natural History Museum (London) NM – Naturhistorisches Museum SfAP – Staatssammlung fur Anthropologie und Palaeoanatomie WOAC – Webb Osteology and Archaeology Collection
ID		Museum accession identifier (these are based on either computerized accession records at collections, or on labels from bone boxes or skeletal remains)
Sex	Male or female, determined from os coxae (occasionally with cranial characteristics)[1]	0 = Male 0? = Probable male 1 = Female 1? = Probable female
Age	Age range based on pubic symphysis and auricular surface (also known age if cadaveric)[2]	Generally: 20-22; 22-25; 25-30; 30-40; 40-50; 50+
NOTE		For most skeletons, site of origin location (most are archaeological site name). If hand-written notes were found with individual skeletons, these are transcribed here.
Location		Modern country of origin or state in case of United States

`GoldMeasures.pdf, Data Goldman.csv.` See function `skelet`

Homework 4

Presentation matters.

1. (10 points). Compute the 3×3 partial correlation matrix for Revenue and truck drivers example using two methods: by inverse correlation matrix and correlation of residuals. Make sure that the two matrices coincide. Interpret the result in layman terms.

2. (5 points). Explain in layman language false correlation referring to Example 1.

3. (10 points). (a) Remove the columns that have -1 or 1 correlation with others as follows from `skelet_2.pdf`. (b) Apply option `use="complete.obs"` when computing the regular correlation matrix and then regularize it by adding 10^{-20} to the diagonal elements. (c) Compute and display the partial correlation matrix. Use your own breaks and colors (see `Rcolor.pdf`) to cover the range of correlation coefficients from -1 to 1. (d) Display the partial correlation matrix using `pheatmap` package. (d) Make your interpretation and conclusion. Save the last heatmap in large size png format file.

Solutions

1.

```
> truckHW4=function()
{
dump("truckHW4","c:\\QBS124\\truckHW4.r")
da=read.csv("c:\\QBS124\\truckR.data.csv")
n=nrow(da);ti=1:n
X=as.matrix(cbind(da,ti))
R=cor(X)
iR=parR=parRES=solve(R)
for(i in 1:3)
for(j in 1:3)
parR[i,j]=-iR[i,j]/sqrt(iR[i,i]*iR[j,j])
diag(parR)=1
print("Partial correlation using R matrix inverse:")
print(parR)
diag(parRES)=1
for(i in 1:3)
for(j in 1:3)
if(i>j)
{
x=X[,-c(i,j)]
y1=X[,i];y2=X[,j]
res1=lm(y1~x)$residuals
res2=lm(y2~x)$residuals
parRES[i,j]=parRES[j,i]=cor(res1,res2)
}
print("Partial correlation using cor on residuals:")
print(parRES)
paste("Partial cor^2 between Revenue and truck dr = parRES[1,2]^2 =",parRES[1,2]^2)
}
```

```
> truckHW4()
[1] "Partial correlation using R matrix inverse:"
truc.dr revenue ti
truc.dr 1.0000000 0.2585552 0.4995997
revenue 0.2585552 1.0000000 0.6802236
ti 0.4995997 0.6802236 1.0000000
[1] "Partial correlation using cor on residuals:"
truc.dr revenue ti
truc.dr 1.0000000 0.2585552 0.4995997
revenue 0.2585552 1.0000000 0.6802236
ti 0.4995997 0.6802236 1.0000000
[1] "Partial cor^2 between Revenue and truck dr = parRES[1,2]^2 = 0.066850803911374"
```

Partial correlation removes the impact of time/trend from Revenue and Truck drivers variables. The coefficient of determination 7% reflects the true contribution of truck drives to explanation of the variance of Revenue.

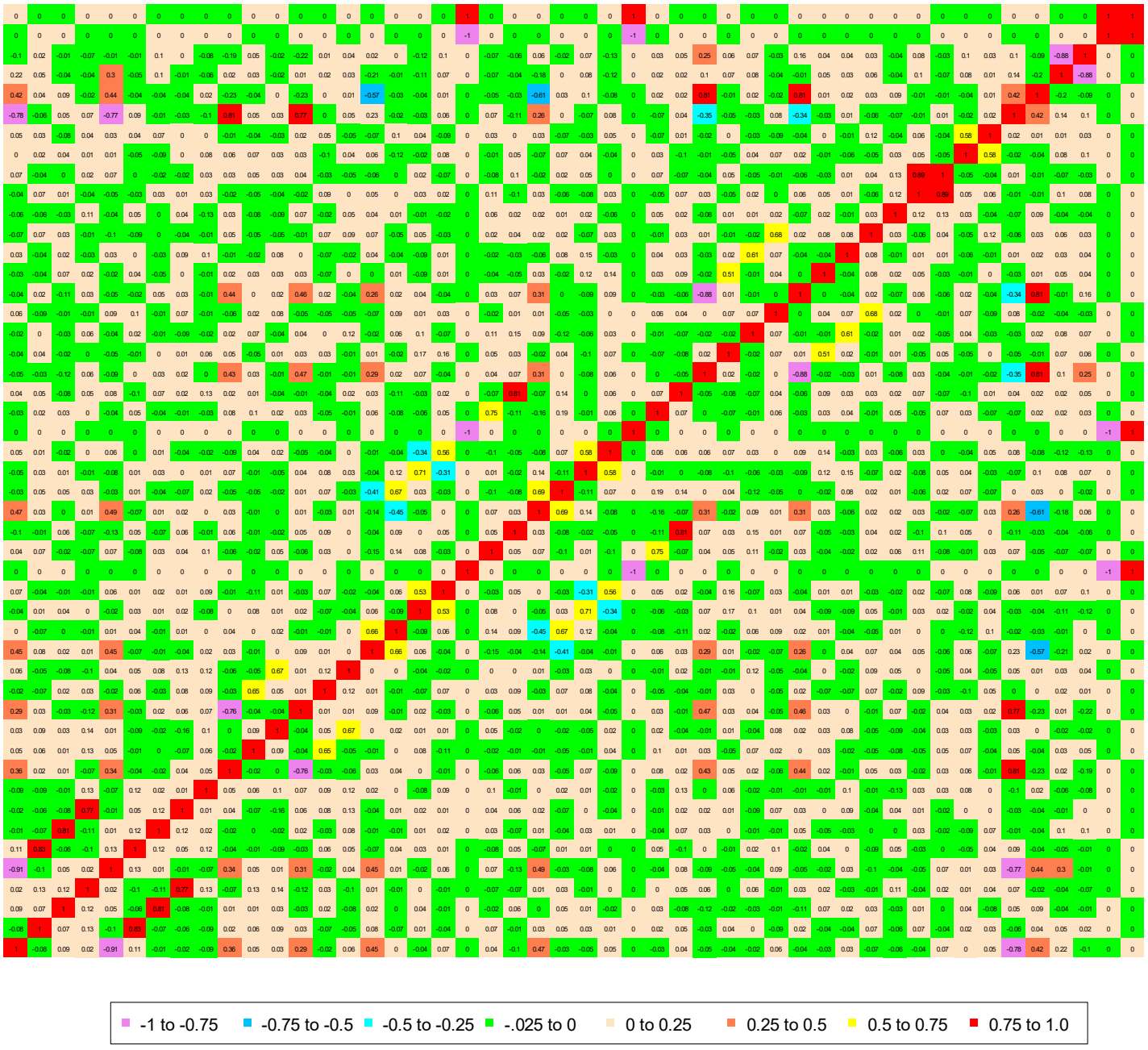
2. False correlation between two random variables is due to the presence of the common factor that influences both variables. When the contribution of this factor increases correlation coefficient approaches 1.

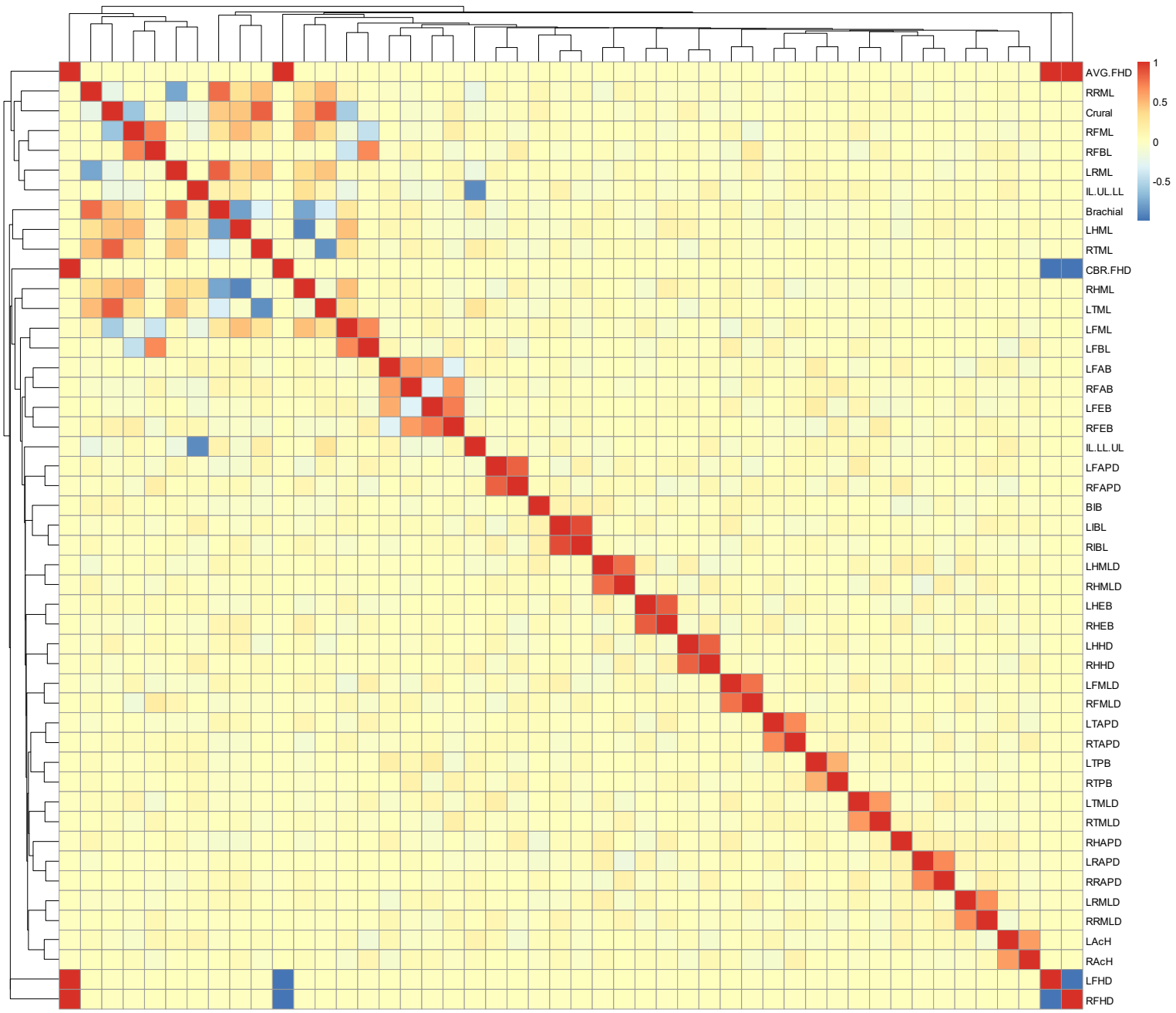
3.

```
hw21_4=function(job=1)
{
dump("hw21_4","c:\\QBS124\\hw21_4.r")
d=read.csv("c:\\QBS124\\Goldman.csv",stringsAsFactors=F)
sex=as.numeric(as.vector(d[,3]))
d=as.matrix(d[,18:ncol(d)])
nm=names(d)
nc=ncol(d);nr=nrow(d)
for(i in 1:nc)
if(sum(is.na(d[,i]))==nr) {alln=i;break}
d=d[,-c(alln,5,50,51)]
nm=nm[-c(alln,5,50,51)]
n=ncol(d)
R=cor(d,use="complete.obs")
diag(R)=rep(1+20^-10,n)
iR=parR=solve(R)
for(i in 1:n)
for(j in 1:n)
parR[i,j]=-iR[i,j]/sqrt(iR[i,i]*iR[j,j])
diag(parR)=rep(1,n)
if(job==1) # my own png
{
png("c:\\QBS124\\slelet_3.png",height=1000,width=1200)
par(mfrow=c(1,1),mar=c(0,1,2,1))
cl=c("violet","deepskyblue","cyan","green","bisque","coral","yellow","red")
image(1:n,1:n,breaks=c(-1,-.75,-.5,-.25,0,.25,.5,.75,1),ylim=c(-5,n+1),xlim=c(-2,n+.5),
        col=cl,ylab="",xlab="",parR,axes=F)
text(1:n,rep(0.5,n),nm,adj=1,cex=.75,srt=45)
text(rep(.3,n),1:n,nm,adj=1,cex=.75)
for(i in 1:n)
for(j in 1:n)
text(i,j,round(parR[i,j],2),cex=.5)
mtext(side=3,paste("Correlation heatmap of",n,"osteometric measurements taken from 1538 human skeletons"),cex=2)
br=c("-1 to -0.75","-0.75 to -0.5","-0.5 to -0.25","-.025 to 0","0 to 0.25","0.25 to 0.5","0.5 to 0.75","0.75 to 1.0")
legend(5,-1.75,br,col=cl,pch=15,horiz=T,cex=1.25)
dev.off()
```

```
}
if(job==2)
{
library(pheatmap)
parR=as.data.frame(parR)
names(parR)=nm
pheatmap(parR)
}
}
```

Partial correlation heatmap of 48 osteometric measurements taken from 1538 human skeletons




AVG.FHD
RRML
Crural
RFML
RFBL
LRML
IL.UL.LL
Brachial
LHML
RTML
CBR.FHD
RHML
LTML
LFML
LFBL
LFAB
RFAB
LFEB
RFEB
IL.LL.UL
LFAPD
RFAPD
BIB
LIBL
RIBL
LHMLD
RHMLD
LHEB
RHEB
LHHD
RHHD
LFMLD
RFMLD
LTAPD
RTAPD
LTPB
RTPB
LTMLD
RTMLD
RHAPD
LRAPD
RRAPD
LRMLD
RRMLD
LAcH
RAcH
LFHD
RFHD
1
0.5
0
-0.5