Ngoc Ha

ST 557 - HW 5

Problem 1 ¶

In [1]:
```
track <- read.csv('TrackData.csv')
head(track)
```

Country	Abbrev	X100m.s	X200m.s	X400m.s	X800m.m	X1500m.m	X5000m.m	X10000m.m	M
Argentina	ARG	10.39	20.81	46.84	1.81	3.70	14.04	29.36	
Australia	AUL	10.31	20.06	44.84	1.74	3.57	13.28	27.66	
Austria	AUS	10.44	20.81	46.82	1.79	3.60	13.26	27.72	
Belgium	BEL	10.34	20.68	45.04	1.73	3.60	13.22	27.45	
Bermuda	BER	10.28	20.58	45.91	1.80	3.75	14.68	30.55	
Brazil	BRA	10.22	20.43	45.21	1.73	3.66	13.62	28.62	

(1a)

```r
distances <- track[,3:10]
S <- cov(distances)
R <- cor(distances)
S
R
```

	X100m.s	X200m.s	X400m.s	X800m.m	X1500m.m	X5000m.m	X10000m.
X100m.s	0.12350249	0.20902182	0.43069956	0.016920438	0.03836684	0.17441020	0.40184
X200m.s	0.20902182	0.41557024	0.79905603	0.033115455	0.07788771	0.35913859	0.81171
X400m.s	0.43069956	0.79905603	2.12290020	0.080743131	0.18974209	0.90887976	2.07341
X800m.m	0.01692044	0.03311545	0.08074313	0.004055758	0.00911532	0.04406209	0.10004
X1500m.m	0.03836684	0.07788771	0.18974209	0.009115320	0.02430774	0.11592929	0.26343
X5000m.m	0.17441020	0.35913859	0.90887976	0.044062088	0.11592929	0.64185811	1.41154
X10000m.m	0.40184545	0.81171145	2.07341549	0.100049327	0.26343721	1.41154798	3.26789
Marathon.m	1.68601222	3.54620963	9.47785704	0.473903333	1.24516296	6.89104852	15.73218

	X100m.s	X200m.s	X400m.s	X800m.m	X1500m.m	X5000m.m	X10000m.m	Ma
X100m.s	1.0000000	0.9226384	0.8411468	0.7560278	0.7002382	0.6194618	0.6325389	0
X200m.s	0.9226384	1.0000000	0.8507270	0.8066265	0.7749513	0.6953770	0.6965391	0
X400m.s	0.8411468	0.8507270	1.0000000	0.8701714	0.8352694	0.7786139	0.7872045	0
X800m.m	0.7560278	0.8066265	0.8701714	1.0000000	0.9180442	0.8635939	0.8690489	0
X1500m.m	0.7002382	0.7749513	0.8352694	0.9180442	1.0000000	0.9281140	0.9346970	0
X5000m.m	0.6194618	0.6953770	0.7786139	0.8635939	0.9281140	1.0000000	0.9746354	0
X10000m.m	0.6325389	0.6965391	0.7872045	0.8690489	0.9346970	0.9746354	1.0000000	0
Marathon.m	0.5199490	0.5961837	0.7049905	0.8064764	0.8655492	0.9321884	0.9431763	1

R is more appropriate to use for PCA, as all the variables are on the same scale.

(1b) Eigendecomposition of S

In [3]:
```
eiDecS <- eigen(S)
eiDecS
```

```
eigen() decomposition
$values
[1] 8.991362e+01 1.412626e+00 2.598442e-01 1.094203e-01 2.730060e-02
[6] 1.273280e-02 2.243554e-03 4.455645e-04

$vectors
             [,1]        [,2]         [,3]         [,4]         [,5]
[1,] -0.019865407 -0.21068958 -0.029041979 -0.358784470  0.190181784
[2,] -0.041554499 -0.35892579 -0.018390126 -0.833534544 -0.048582165
[3,] -0.110631838 -0.82786251 -0.377669011  0.396041212 -0.012020033
[4,] -0.005487699 -0.02317490  0.005341591 -0.009568087 -0.011107487
[5,] -0.014386822 -0.04465255  0.050004337 -0.015981502 -0.043222520
[6,] -0.079308444 -0.12996134  0.336448522  0.018873808 -0.909186992
[7,] -0.181098994 -0.29885393  0.848722695  0.134662690  0.364239482
[8,] -0.972787446  0.18080736 -0.141872114 -0.028425488  0.006575083
             [,6]         [,7]          [,8]
[1,]  0.886865894 -0.052444908 -0.0139585779
[2,] -0.409969944  0.062270182 -0.0037828046
[3,] -0.047663812  0.020389912 -0.0094695712
[4,] -0.007204523 -0.261227847  0.9648302746
[5,] -0.067333230 -0.959092660 -0.2622644611
[6,]  0.184076191  0.052548542 -0.0001130819
[7,] -0.068113893  0.045771467  0.0045055042
[8,]  0.003532208 -0.001055127 -0.0008700758
```

(1c) Eigendecomposition of R

In [4]:
```
eiDecR <- eigen(R)
eiDecR
```

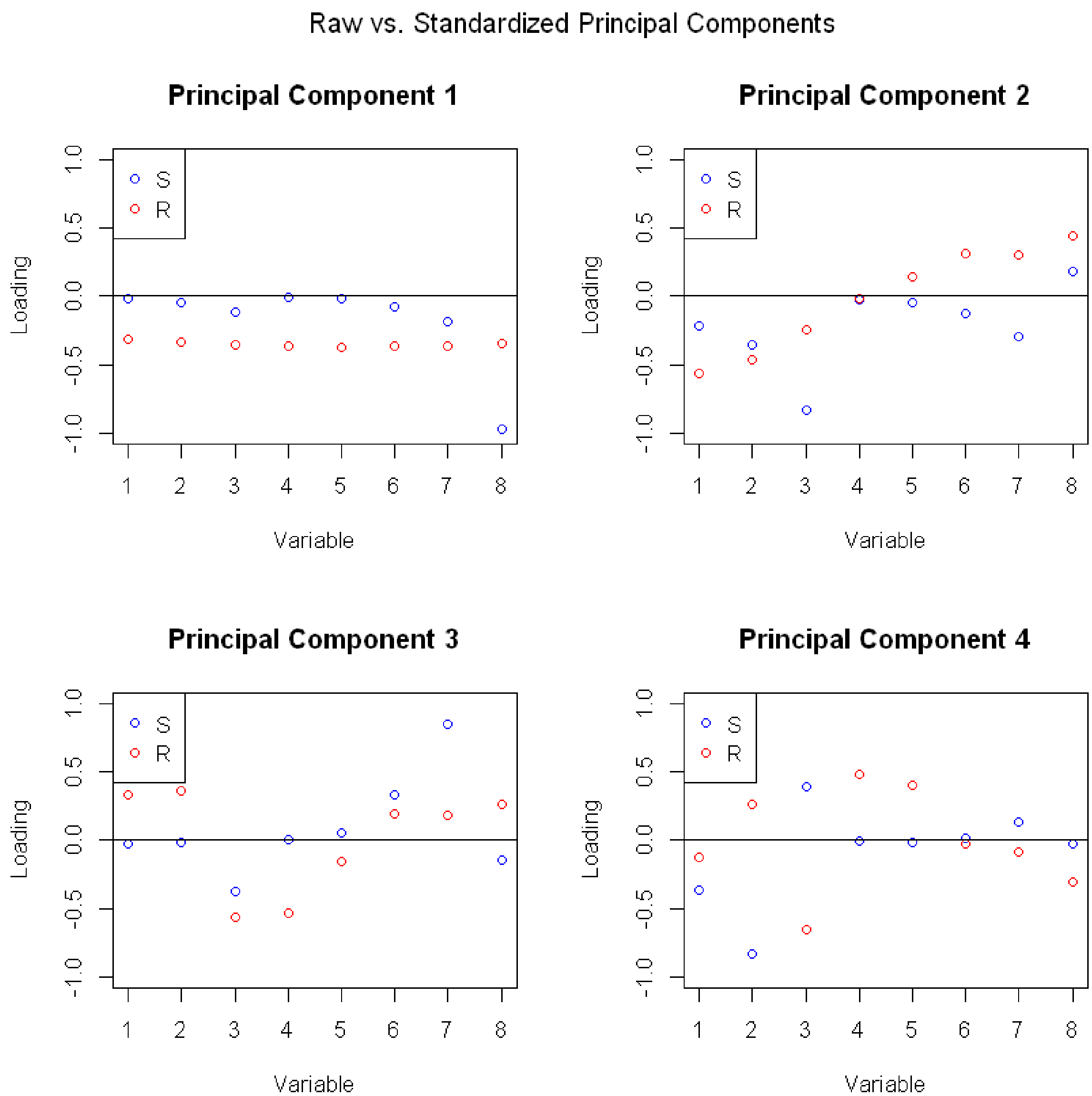
```
eigen() decomposition
$values
[1] 6.62214613 0.87761829 0.15932114 0.12404939 0.07988027 0.06796515 0.04641
953
[8] 0.02260010

$vectors
           [,1]        [,2]       [,3]        [,4]       [,5]       [,6]
[1,] -0.3175565 -0.56687750  0.3322620 -0.12762827  0.2625555 -0.5937042
[2,] -0.3369792 -0.46162589  0.3606567  0.25911576 -0.1539571  0.6561367
[3,] -0.3556454 -0.24827331 -0.5604674 -0.65234077 -0.2183229  0.1566252
[4,] -0.3686841 -0.01242993 -0.5324823  0.47999895  0.5400528 -0.0146918
[5,] -0.3728099  0.13979665 -0.1534427  0.40451039 -0.4877151 -0.1578430
[6,] -0.3643741  0.31203045  0.1897643 -0.02958755 -0.2539792 -0.1412987
[7,] -0.3667726  0.30685985  0.1817517 -0.08006862 -0.1331764 -0.2190168
[8,] -0.3419261  0.43896267  0.2632087 -0.29951213  0.4979283  0.3152849
             [,7]          [,8]
[1,]  0.136241260 -0.1055416752
[2,] -0.112639528  0.0960543222
[3,] -0.002853707  0.0001272032
[4,] -0.238016094  0.0381651151
[5,]  0.610011482 -0.1392909844
[6,] -0.591298850 -0.5466969221
[7,] -0.176871021  0.7967952190
[8,]  0.398822209 -0.1581638575
```

(1d)

In [5]:

```r
par(mfrow=c(2,2), oma=c(0,0,2,0))
for(i in 1:4){
    plot(1:8, eiDecS$vec[,i], xlab="Variable", ylab="Loading", main=paste("Principal Component ", i, sep=""), ylim=c(-1, 1), col = 'blue')
    points(1:8, eiDecR$vec[,i], xlab="Variable", ylab="Loading", main=paste("Principal Component ", i, sep=""), ylim=c(-1, 1), col='red')
    legend("topleft", legend=c("S", "R"), col=c("blue", "red"), pch=21)
        abline(h=0)
}
mtext("Raw vs. Standardized Principal Components", outer=T)
```



(1e)

For S: marathon times (variable 8) dominate the covariances among running times, which makes sense a marathon is much longer than other formats.

(1f)

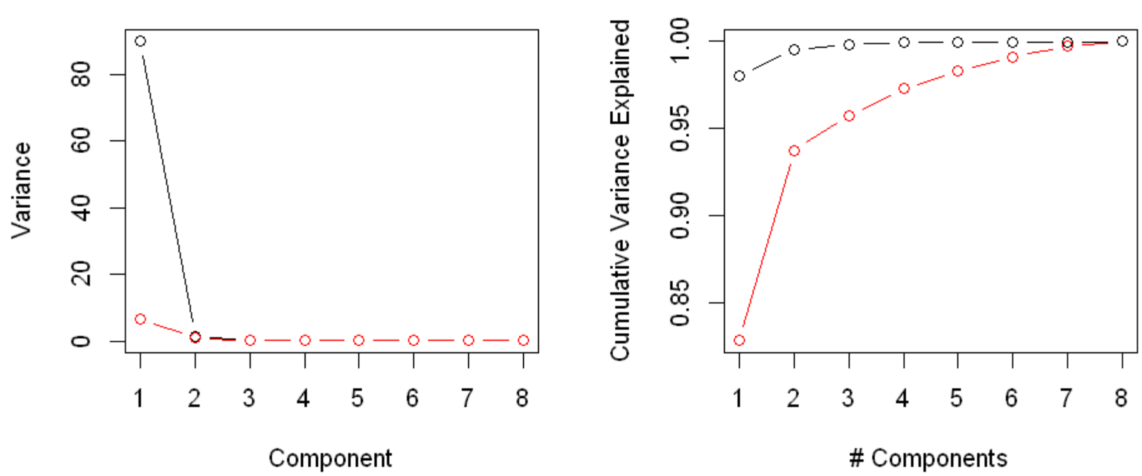
For R: the standardized loadings are roughly equal. This principal component shows how fast or slow a country is in general.

(1g)

The second principal components of R seems to show the contrast between fast marathoners and fast sprinters.

(1h)

In [6]:
```
options(repr.plot.width=8, repr.plot.height=4)
par(mfrow=c(1,2), oma=c(0,0,0,0))
plot(1:8, eiDecS$val, type="b", xlab="Component", ylab="Variance")
lines(1:8, eiDecR$val, type="b", col='red')
plot(1:8, cumsum(eiDecR$val)/sum(eiDecR$val), type="b", col='red', xlab="# Com
ponents", ylab="Cumulative Variance Explained")
lines(1:8, cumsum(eiDecS$val)/sum(eiDecS$val), type="b")
```



(1i)

I'd want to keep the first 3 principal components of R, as they explain 95% of the variance, and the variances seem to taper off starting from the 4th principal component.

Problem 2

In [7]:
```
nyse <- read.csv('NYSEData.csv')
head(nyse)
```

JPMorgan	Citibank	WellsFargo	RoyalDutchShell	ExxonMobil
0.0130338	-0.0078431	-0.0031889	-0.0447693	0.0052151
0.0084862	0.0166886	-0.0062100	0.0119560	0.0134890
-0.0179153	-0.0086393	0.0100360	0.0000000	-0.0061428
0.0215589	-0.0034858	0.0174353	-0.0285917	-0.0069534
0.0108225	0.0037167	-0.0101345	0.0291900	0.0409751
0.0101713	-0.0121978	-0.0083768	0.0137083	0.0029895

(2a)

In [10]:
```
S <- cov(nyse)
eiDecS <- eigen(sampCov)
eiDecS
```

```
eigen() decomposition
$values
[1] 0.0013676780 0.0007011596 0.0002538024 0.0001426026 0.0001188868

$vectors
          [,1]       [,2]        [,3]       [,4]        [,5]
[1,] 0.2228228  0.6252260  0.32611218 -0.6627590  0.11765952
[2,] 0.3072900  0.5703900 -0.24959014  0.4140935 -0.58860803
[3,] 0.1548103  0.3445049 -0.03763929  0.4970499  0.78030428
[4,] 0.6389680 -0.2479475 -0.64249741 -0.3088689  0.14845546
[5,] 0.6509044 -0.3218478  0.64586064  0.2163758 -0.09371777
```

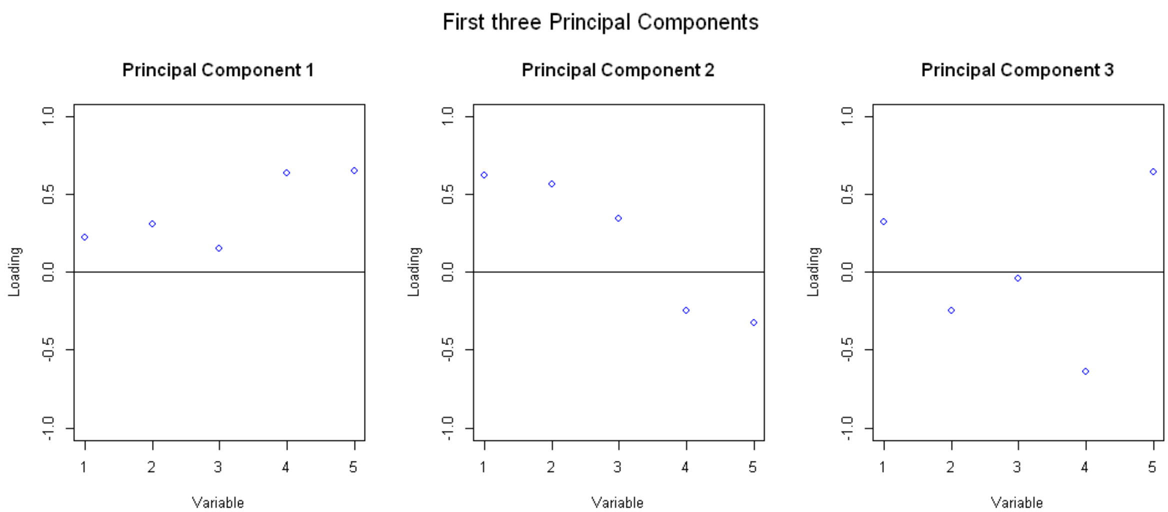
(2b)

In [11]:
```
propVar <- sum(eiDecS$values[1:3])/sum(eiDecS$values)
cat(propVar, "of total variance is explained by the first 3 principal componen
ts")
```

```
0.8988095 of total variance is explained by the first 3 principal components
```

(2c)

In [14]:
```
options(repr.plot.width=9, repr.plot.height=4)
par(mfrow=c(1,3), oma=c(0,0,2,0))
for(i in 1:3){
    plot(1:5, eiDecS$vec[,i], xlab="Variable", ylab="Loading", main=paste("Pri
ncipal Component ", i, sep=""), ylim=c(-1, 1), col = 'blue')
        abline(h=0)
}
mtext("First three Principal Components", outer=T)
```



First PC demonstrates the performance of the market in general. Second PC shows the contrast between stock performance of Finance sector and Energy sector. Third PC focuses on stocks with higher market capitalization (JP Morgan, Shell, Exxon).

Problem 4

In [1]:
```
corr <- as.matrix(read.csv('PhysioData.csv'))
head(corr)
```

	weight	height	physact	ldl	alb	crt	
weight	1.000000000	0.54775881	-0.02803745	0.003570772	0.04673810	0.25399164	-0.14944
height	0.547758814	1.00000000	0.06493509	-0.156549811	0.08819927	0.36485715	-0.29476
physact	-0.028037453	0.06493509	1.00000000	-0.031756690	0.01476370	-0.02632625	-0.00900
ldl	0.003570772	-0.15654981	-0.03175669	1.000000000	0.12453340	-0.13135504	0.19678
alb	0.046738095	0.08819927	0.01476370	0.124533399	1.00000000	0.04428462	-0.06429
crt	0.253991639	0.36485715	-0.02632625	-0.131355039	0.04428462	1.00000000	-0.15481

(4a)

```
In [2]: eiDecCor <- eigen(corr)
```

i. m = 2

```
In [3]: L2 <- cbind(eiDecCor$val[1]*eiDecCor$vec[,1], eiDecCor$val[2]*eiDecCor$vec[,2]
)
errorCov2 <- corr - L2%*%t(L2)
specVars2 <- diag(errorCov2)
L2_table <- data.frame(L2,specVars2)
colnames(L2_table) <- c("Factor 1 Loadings","Factor 2 Loadings","Specific Vari
ances")
L2_table
```

	Factor 1 Loadings	Factor 2 Loadings	Specific Variances
weight	-1.0452745	-0.1403353	-0.11229274
height	-1.3346010	-0.1595277	-0.80660894
physact	-0.1265833	0.1814451	0.95105432
ldl	0.3651604	0.1599542	0.84107255
alb	-0.2255166	0.0941445	0.94027907
crt	-0.8112691	-0.5135738	0.07808445
plt	0.7143848	0.1338270	0.47174474
sbp	0.2741563	-0.7725324	0.32803205
aai	-0.4325669	0.8560635	0.08004119
fev	-1.1450629	0.2331262	-0.36551677
dsst	-0.1974087	0.7093896	0.45779616
atrophy	-0.2404455	-0.4284885	0.75858351

Factor 1: height, weight and fev contribute the most to factor 1. Interpretation: factor 1 correlates with the size of a person.

Factor 2 shows the contrast between **systolic blood pressure** vs. **ankle-to-arm sbp ratio** and **coginite performance**.

ii. m = 3

In [4]:
```
L3 <- cbind(eiDecCor$val[1]*eiDecCor$vec[,1], eiDecCor$val[2]*eiDecCor$vec[,2]
, eiDecCor$val[3]*eiDecCor$vec[,3])
errorCov3 <- corr - L3%*%t(L3)
specVars3 <- diag(errorCov3)
L3_table <- data.frame(L3,specVars3)
colnames(L3_table) <- c("Factor 1 Loadings","Factor 2 Loadings","Factor 3 Load
ings","Specific Variances")
L3_table
```

	Factor 1 Loadings	Factor 2 Loadings	Factor 3 Loadings	Specific Variances
weight	-1.0452745	-0.1403353	0.254226398	-0.17692380
height	-1.3346010	-0.1595277	0.041405944	-0.80832339
physact	-0.1265833	0.1814451	-0.351258636	0.82767169
ldl	0.3651604	0.1599542	0.756781166	0.26835482
alb	-0.2255166	0.0941445	0.606733197	0.57215390
crt	-0.8112691	-0.5135738	-0.007366511	0.07803019
plt	0.7143848	0.1338270	0.299009230	0.38233822
sbp	0.2741563	-0.7725324	0.023902569	0.32746072
aai	-0.4325669	0.8560635	-0.156188604	0.05564631
fev	-1.1450629	0.2331262	0.027012246	-0.36624644
dsst	-0.1974087	0.7093896	0.239375966	0.40049530
atrophy	-0.2404455	-0.4284885	0.326469376	0.65200126

Factor 3 correlates with the **cholesterol level** in the subject's blood.

(4b)

i. m = 2

```
In [5]: res2 <- corr - (L2%*%t(L2)+errorCov2)
res2
```

	weight	height	physact	ldl	alb	crt
weight	0.000000e+00	0.000000e+00	1.040834e-17	-3.903128e-18	1.387779e-17	0.000000e+00
height	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	-5.551115e-17
physact	1.040834e-17	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
ldl	-3.903128e-18	0.000000e+00	0.000000e+00	0.000000e+00	1.387779e-17	0.000000e+00
alb	1.387779e-17	0.000000e+00	0.000000e+00	1.387779e-17	0.000000e+00	0.000000e+00
crt	0.000000e+00	-5.551115e-17	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
plt	2.775558e-17	0.000000e+00	3.469447e-18	0.000000e+00	0.000000e+00	0.000000e+00
sbp	5.204170e-18	0.000000e+00	-1.040834e-17	0.000000e+00	3.469447e-18	-7.806256e-18
aai	0.000000e+00	-1.387779e-17	0.000000e+00	0.000000e+00	1.040834e-17	0.000000e+00
fev	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	5.551115e-17
dsst	0.000000e+00	6.938894e-18	6.938894e-18	3.361027e-18	0.000000e+00	0.000000e+00
atrophy	0.000000e+00	-2.775558e-17	0.000000e+00	1.387779e-17	0.000000e+00	0.000000e+00

ii. m = 3

In [6]:
```
res3 <- corr - (L3%*%t(L3)+errorCov3)
res3
```

	weight	height	physact	ldl	alb	crt
weight	0.000000e+00	0.000000e+00	-3.469447e-18	-3.903128e-18	1.387779e-17	0.000000e+00
height	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	-5.551115e-17
physact	-3.469447e-18	0.000000e+00	0.000000e+00	-6.938894e-18	1.040834e-17	0.000000e+00
ldl	-3.903128e-18	0.000000e+00	-6.938894e-18	0.000000e+00	0.000000e+00	-2.775558e-17
alb	1.387779e-17	0.000000e+00	1.040834e-17	0.000000e+00	0.000000e+00	0.000000e+00
crt	0.000000e+00	-5.551115e-17	0.000000e+00	-2.775558e-17	0.000000e+00	0.000000e+00
plt	2.775558e-17	0.000000e+00	-1.040834e-17	0.000000e+00	0.000000e+00	0.000000e+00
sbp	5.204170e-18	0.000000e+00	-1.040834e-17	0.000000e+00	3.469447e-18	-7.806256e-18
aai	0.000000e+00	-1.387779e-17	0.000000e+00	0.000000e+00	-3.469447e-18	0.000000e+00
fev	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	-5.551115e-17
dsst	-6.938894e-18	6.938894e-18	6.938894e-18	-3.577867e-18	1.387779e-17	0.000000e+00
atrophy	2.775558e-17	-2.775558e-17	0.000000e+00	0.000000e+00	6.938894e-18	0.000000e+00

(4c)

i. m = 2

In [9]:
```
mlfa2 <- factanal(covmat=corr, factors=2, rotation="none")
mlfa2
```

```
Call:
factanal(factors = 2, covmat = corr, rotation = "none")

Uniquenesses:
 weight  height physact     ldl     alb     crt     plt     sbp     aai       f
ev
  0.675   0.084   0.988   0.974   0.990   0.828   0.903   0.801   0.526   0.5
69
   dsst atrophy
  0.883   0.960

Loadings:
        Factor1 Factor2
weight   0.570
height   0.956
physact
ldl     -0.160
alb
crt      0.385  -0.154
plt     -0.310
sbp             -0.438
aai      0.105   0.681
fev      0.610   0.241
dsst             0.340
atrophy  0.121  -0.160

               Factor1 Factor2
SS loadings      1.930   0.889
Proportion Var   0.161   0.074
Cumulative Var   0.161   0.235

The degrees of freedom for the model is 43 and the fit was 0.1927
```

Factor 1 correlates with the **size** of the subject. Factor 2 shows a contrast between **systolic blood pressure** and **arm-to-length blood pressure ratio**.

ii. m = 3

In [10]:
```r
mlfa3 <- factanal(covmat=corr, factors=3, rotation="none")
mlfa3
```

```
Call:
factanal(factors = 3, covmat = corr, rotation = "none")

Uniquenesses:
 weight  height physact     ldl     alb     crt     plt     sbp     aai      f
ev
  0.659   0.097   0.988   0.005   0.969   0.821   0.881   0.800   0.517   0.5
64
   dsst atrophy
  0.884   0.960

Loadings:
        Factor1 Factor2 Factor3
weight   0.584
height   0.935  -0.165
physact
ldl              0.997
alb      0.119   0.124
crt      0.369  -0.135  -0.158
plt     -0.281   0.200
sbp                     -0.437
aai      0.102           0.685
fev      0.613           0.235
dsst                     0.337
atrophy  0.119          -0.160

               Factor1 Factor2 Factor3
SS loadings      1.858   1.106   0.891
Proportion Var   0.155   0.092   0.074
Cumulative Var   0.155   0.247   0.321

The degrees of freedom for the model is 33 and the fit was 0.1187
```

Factor 1 correlates with the **size** of the subject. Factor 2 correlates with **cholesterol level** in the subject's blood. Factor 3 shows a contrast between **systolic blood pressure** and **arm-to-length blood pressure ratio**.

(4d)

i. m = 2

In [11]:
```
mle2Fit <- mlfa2$load %*% t(mlfa2$load) + diag(mlfa2$uni)
mle2Res <- corr - mle2Fit
mle2Res
```

	weight	height	physact	ldl	alb	crt	plt
weight	-2.295501e-06	2.316174e-03	-6.600738e-02	9.474229e-02	-6.364201e-03	3.444910e-02	2.744664e-02
height	2.316174e-03	5.795176e-08	4.071739e-03	-3.665135e-03	4.081529e-04	-8.229091e-03	1.002223e-03
physact	-6.600738e-02	4.071739e-03	-4.414328e-07	-2.110953e-02	5.157061e-03	-3.857461e-02	1.400327e-02
ldl	9.474229e-02	-3.665135e-03	-2.110953e-02	1.515341e-07	1.394198e-01	-6.982532e-02	1.472004e-01
alb	-6.364201e-03	4.081529e-04	5.157061e-03	1.394198e-01	-3.679141e-07	1.442844e-02	-3.435965e-02
crt	3.444910e-02	-8.229091e-03	-3.857461e-02	-6.982532e-02	1.442844e-02	2.907133e-06	-3.956903e-02
plt	2.744664e-02	1.002223e-03	1.400327e-02	1.472004e-01	-3.435965e-02	-3.956903e-02	2.774056e-06
sbp	5.876081e-02	-8.040394e-03	4.604408e-02	-4.377930e-02	8.276838e-03	-3.556760e-02	-7.107719e-03
aai	2.805056e-02	-4.595808e-03	1.118192e-02	-4.186461e-02	-5.241403e-03	1.837698e-02	-2.818814e-02
fev	-1.168937e-02	1.341863e-03	4.041874e-02	3.369120e-02	-7.839537e-04	2.507144e-02	1.605121e-02
dsst	3.688456e-02	-2.267373e-03	-6.381315e-02	6.538309e-03	5.216409e-02	-1.080193e-01	3.696668e-02
atrophy	-4.485594e-03	7.105641e-05	-7.647922e-02	7.147184e-04	4.267490e-02	8.309776e-02	-2.852986e-02

ii. m = 3

```
In [12]: mle3Fit <- mlfa3$load %*% t(mlfa3$load) + diag(mlfa3$uni)
mle3Res <- corr - mle3Fit
mle3Res
```

	weight	height	physact	ldl	alb	crt	plt
weight	1.769072e-06	1.715552e-03	-6.367751e-02	4.930838e-05	-2.203323e-02	3.741820e-02	1.466025e-02
height	1.715552e-03	-1.434591e-07	5.151534e-03	-8.841224e-06	-7.031566e-04	-8.562872e-03	-1.919501e-04
physact	-6.367751e-02	5.151534e-03	-9.850525e-07	1.740750e-05	8.126802e-03	-3.980536e-02	1.700254e-02
ldl	4.930838e-05	-8.841224e-06	1.740750e-05	-1.488487e-07	7.889531e-05	-3.531039e-05	5.173403e-06
alb	-2.203323e-02	-7.031566e-04	8.126802e-03	7.889531e-05	2.104358e-07	2.320952e-02	-5.476550e-02
crt	3.741820e-02	-8.562872e-03	-3.980536e-02	-3.531039e-05	2.320952e-02	2.870971e-08	-2.811321e-02
plt	1.466025e-02	-1.919501e-04	1.700254e-02	5.173403e-06	-5.476550e-02	-2.811321e-02	2.484733e-07
sbp	6.251392e-02	-9.205252e-03	4.505925e-02	-9.551266e-05	1.464274e-02	-3.922948e-02	-2.902823e-04
aai	3.219461e-02	-4.912088e-03	9.644994e-03	-4.221580e-06	3.195008e-04	1.661071e-02	-2.148498e-02
fev	-2.015677e-02	1.870694e-03	4.145989e-02	3.735397e-05	-7.565595e-03	2.440522e-02	1.214759e-02
dsst	3.581256e-02	-3.353576e-03	-6.361512e-02	-1.814167e-04	5.108954e-02	-1.079476e-01	3.622663e-02
atrophy	-5.758711e-03	4.482889e-04	-7.637260e-02	4.381138e-05	4.217393e-02	8.238260e-02	-2.842938e-02

(4e)

```
In [17]: cat("Determinant of PCA residual matrix:", det(res3))
cat("\nDeterminant of MLE residual matrix:", det(mle3Res))
```

```
Determinant of PCA residual matrix: 3.02803e-200
Determinant of MLE residual matrix: 3.220028e-56
```

PCA Factor Analysis has smaller residuals => better for this problem.

(4f)

The factors are similar in both methods (PCA and MLE) for both m = 2 and m = 3.

```
In [ ]:
```