Homework 5

ST557

Nick Sun

Question 1

For this problem, we are analyzing the track athlete data for 55 different questions. There are 8 race data columns: 100m, 200m, 400m, 800m, 1500m, 5000m, 10000m, and Marathon.

Part a.

First we can obtain the sample covariance and correlation matrices.

	100m.s	200m.s	400m.s	800m.m	1500m.m	5000m.m	10000m.m	Marathon.m
100m.s	0.1235025	0.2090218	0.4306996	0.0169204	0.0383668	0.1744102	0.4018455	1.6860122
200m.s	0.2090218	0.4155702	0.7990560	0.0331155	0.0778877	0.3591386	0.8117114	3.5462096
400m.s	0.4306996	0.7990560	2.1229002	0.0807431	0.1897421	0.9088798	2.0734155	9.4778570
800m.m	0.0169204	0.0331155	0.0807431	0.0040558	0.0091153	0.0440621	0.1000493	0.4739033
1500m.m	0.0383668	0.0778877	0.1897421	0.0091153	0.0243077	0.1159293	0.2634372	1.2451630
5000m.m	0.1744102	0.3591386	0.9088798	0.0440621	0.1159293	0.6418581	1.4115480	6.8910485
10000m.m	0.4018455	0.8117114	2.0734155	0.1000493	0.2634372	1.4115480	3.2678936	15.7321815
Marathon.m	1.6860122	3.5462096	9.4778570	0.4739033	1.2451630	6.8910485	15.7321815	85.1381467

	100m.s	200m.s	400m.s	800m.m	1500m.m	5000m.m	10000m.m	Marathon.m
100m.s	1.0000000	0.9226384	0.8411468	0.7560278	0.7002382	0.6194618	0.6325389	0.5199490
200m.s	0.9226384	1.0000000	0.8507270	0.8066265	0.7749513	0.6953770	0.6965391	0.5961837
400m.s	0.8411468	0.8507270	1.0000000	0.8701714	0.8352694	0.7786139	0.7872045	0.7049905
800m.m	0.7560278	0.8066265	0.8701714	1.0000000	0.9180442	0.8635939	0.8690489	0.8064764
1500m.m	0.7002382	0.7749513	0.8352694	0.9180442	1.0000000	0.9281140	0.9346970	0.8655492
5000m.m	0.6194618	0.6953770	0.7786139	0.8635939	0.9281140	1.0000000	0.9746354	0.9321884
10000m.m	0.6325389	0.6965391	0.7872045	0.8690489	0.9346970	0.9746354	1.0000000	0.9431763
Marathon.m	0.5199490	0.5961837	0.7049905	0.8064764	0.8655492	0.9321884	0.9431763	1.0000000

Since the data are not all in the same units, 3 of the races are measured in seconds and 5 of them are measured in minutes, we should opt to use the correlation matrix **R** instead of **S** since it is standardized.

Part b.

Let's get the eigenvalues and eigenvectors of **S**.

```
## 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
```

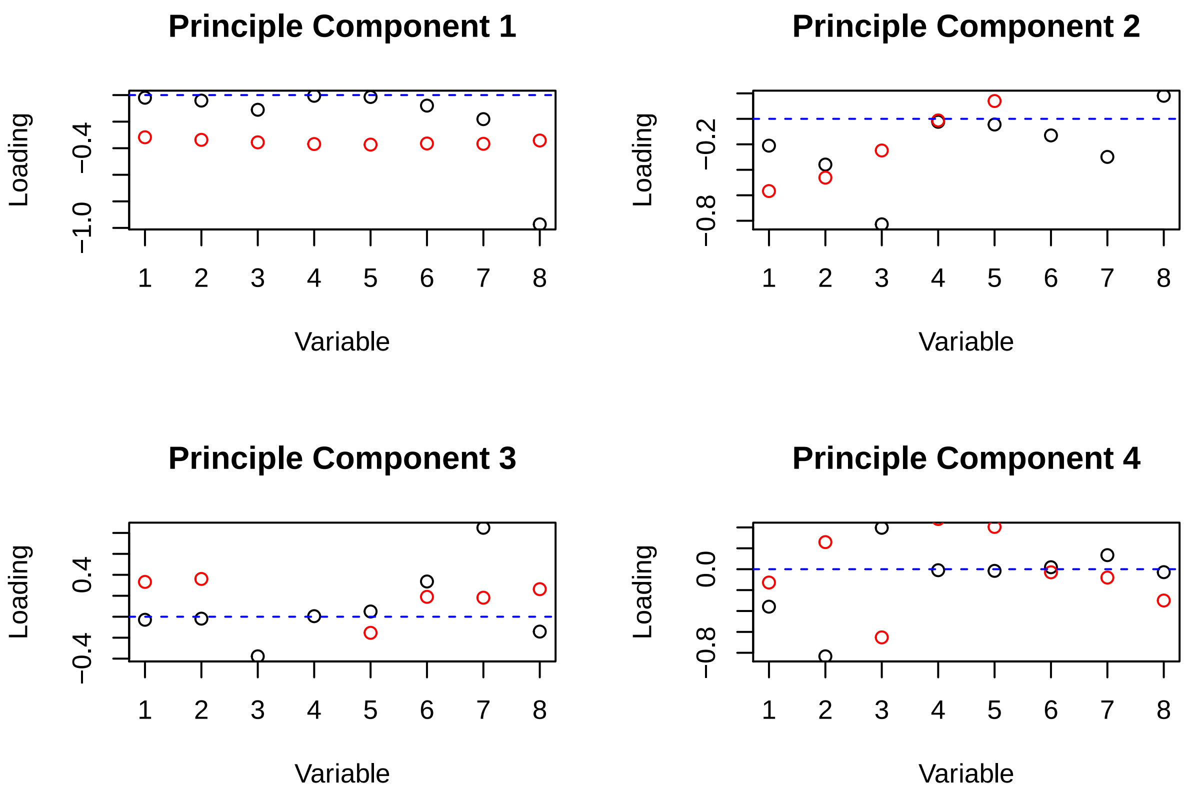
Part c.

Here are the eigenvalues and eigenvectors of **R**.

```
## eigen() decomposition
## $values
## [1] 6.62214613 0.87761829 0.15932114 0.12404939 0.07988027 0.06796515
## [7] 0.04641953 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
```

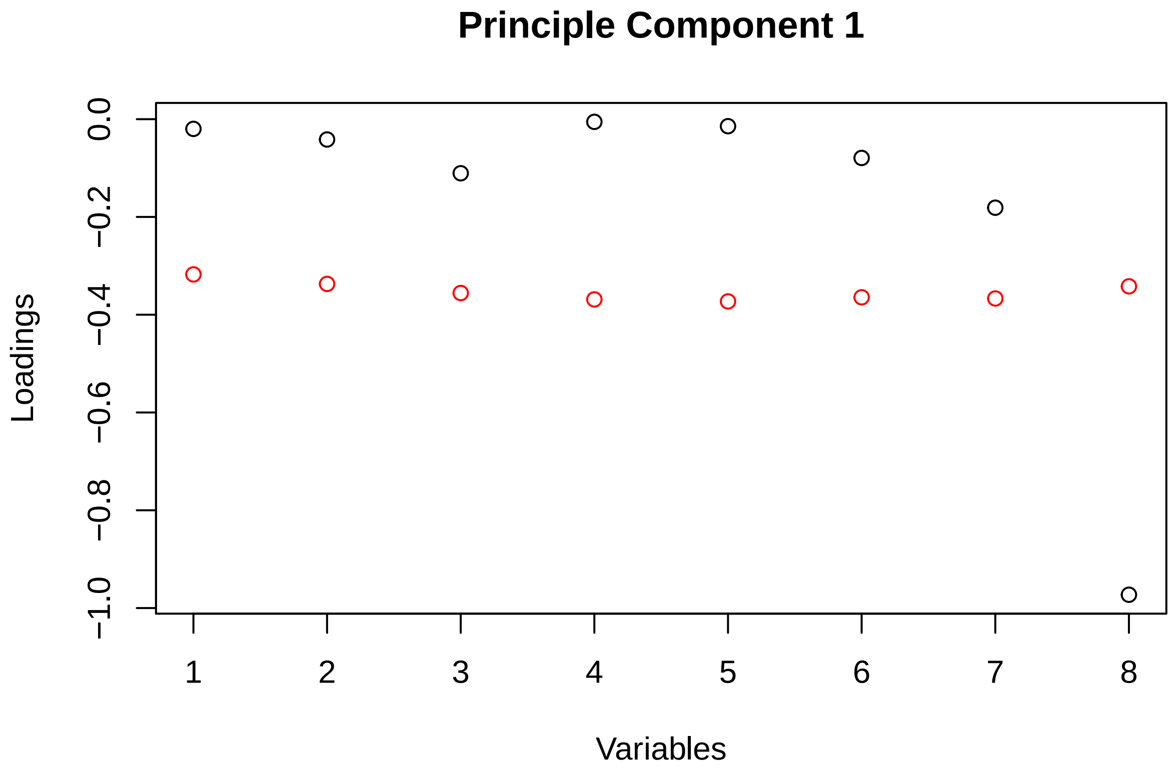
Part d.

Let’s construct a loading plot for the first four principle components of S.



Here the loadings from the correlation matrix are labeled in **red** and the loadings from the covariance matrix are colored **black**.

Part e.

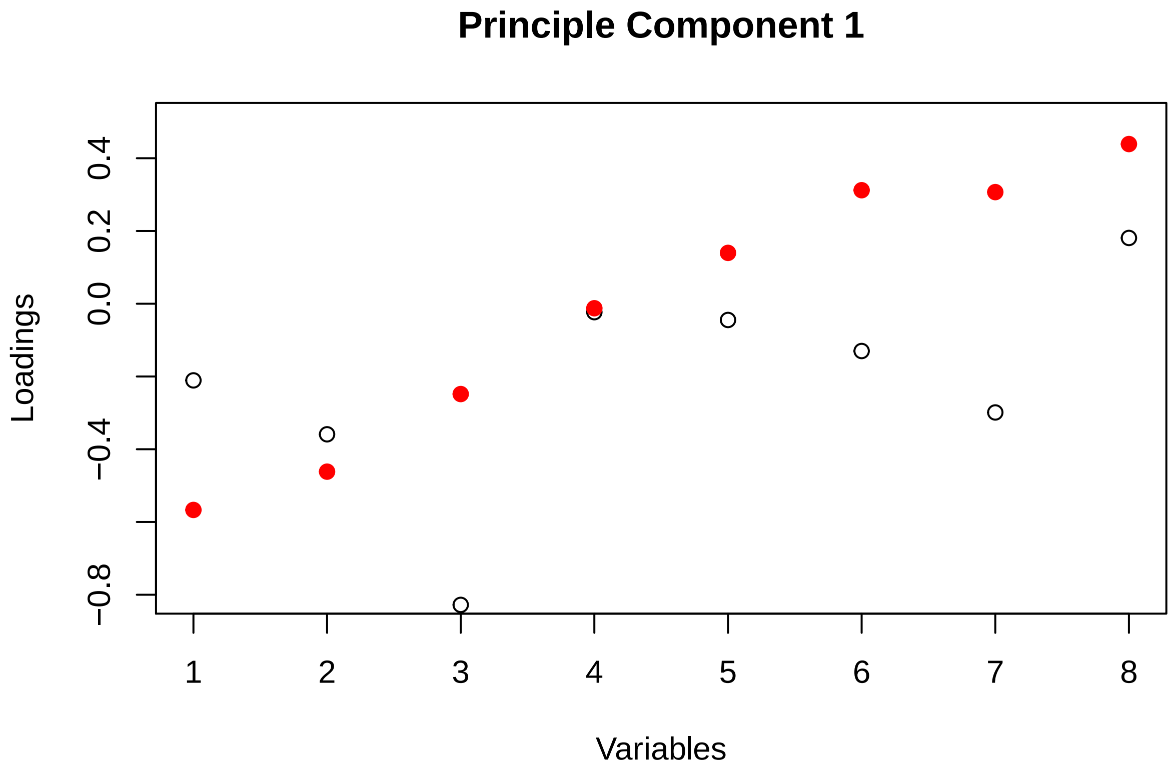


The interpretation of the loadings for the first principle component of **S** are a large emphasis on all the races except for the marathon (or conversely, that the marathon is by far the most important race, depending upon how we consider the sign). A country with a high value on this principle component will likely have high times on the seven shortest races and lower times on the marathon as compared to the global average.

Part f.

The interpretation of the loadings for the first principle component of **R** would be a relatively equal weighting on all 8 races. A country with a high value on this principle component will likely have higher times than average on all of these races.

Part g.



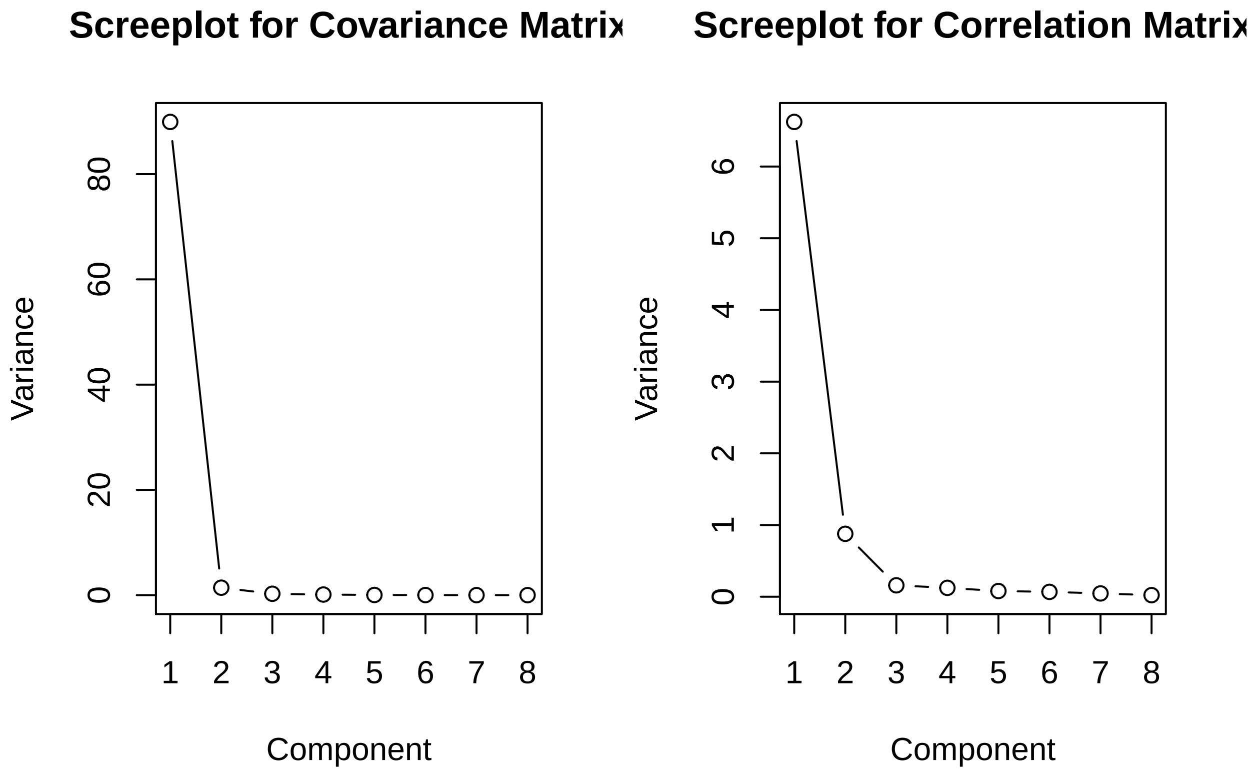
The red dots are the loadings from the correlation matrix.

We can see that there is an inverse weighting of the shorter races compared to the longer races. A country with a high score on this principle component is a country of average sprinters and relatively slow marathon runners.

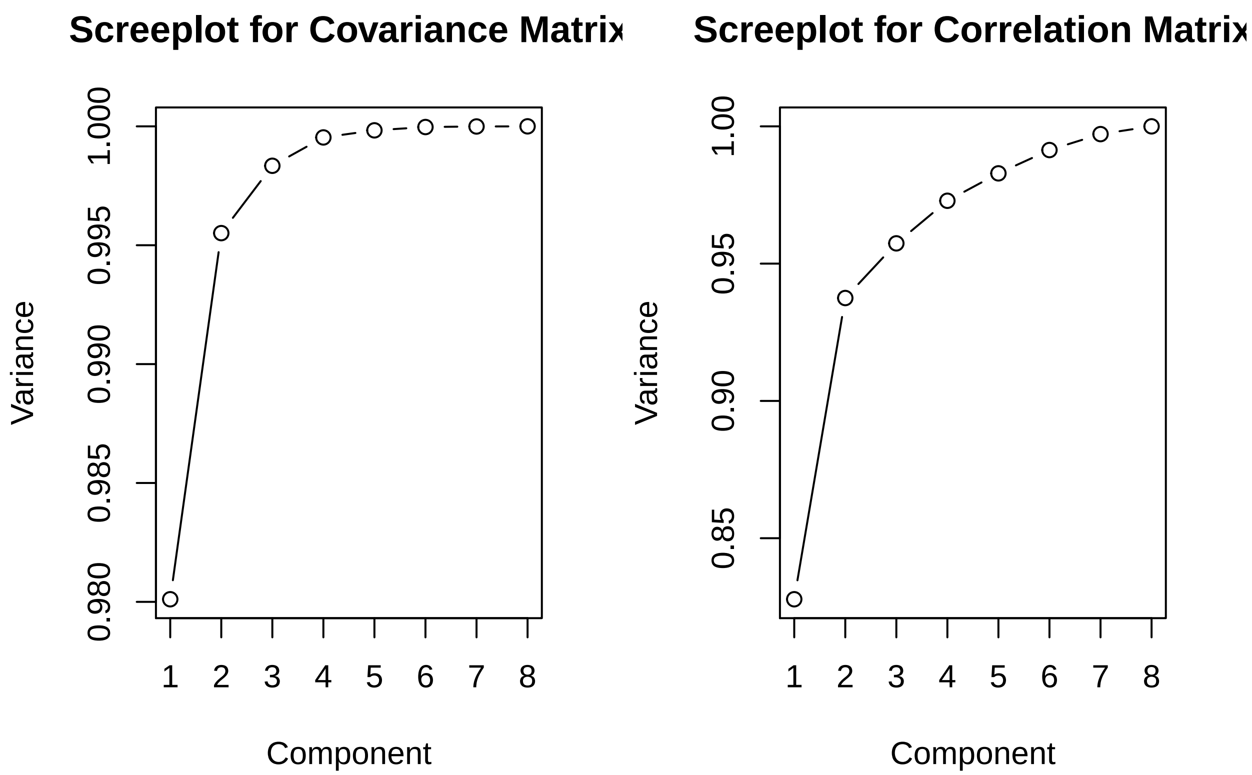
Part h.

Let's make a scree plot for **S** and **R**.

Here is the one for **S**.



Let's also make a cumulative variance plot for both **S** and **R**.



The elbow on the correlation matrix appears to be at 3, so that would suggest not go beyond three principle components. We can see from the cumulative variance plot as well that the first two principle components explain almost 95% of the variation in the data, so we can actually keep just those two and be alright.

Question 2

For this problem, we will be analyzing data from the New York Stock Exchange. 103 weeks worth of data is provided here for 5 stocks where the first three stocks are for financial companies and the last two stocks are from energy/petroleum.

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

Part a.

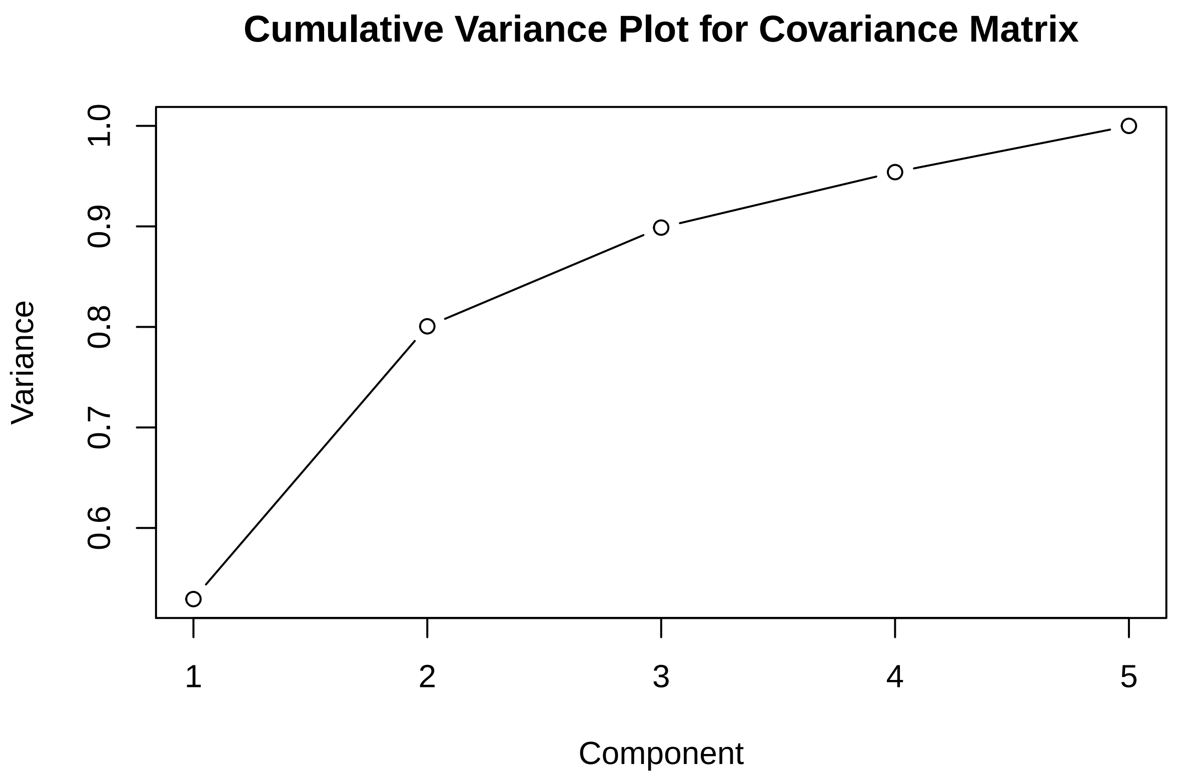
Let's start analyzing this dataset by performing a principle components analysis on the covariance matrix.

```r
(nyse_s_eig <- eigen(cov(nyse_data)))
```

```
## 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
```

Part b.

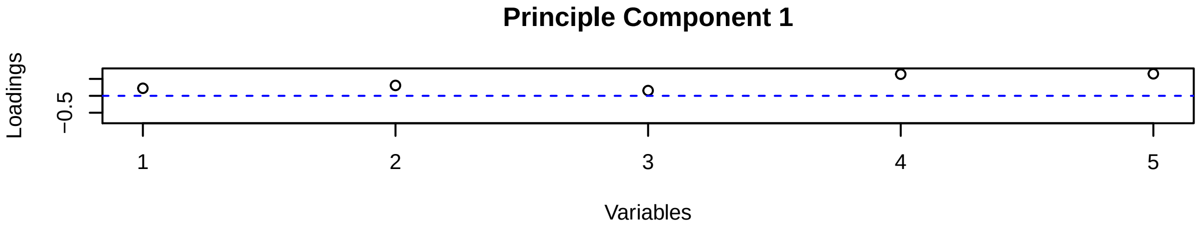
The proportion of the total sample variance that is explained by the first three sample components can be visualized by a scree plot and compute directly using the eigenvalues.

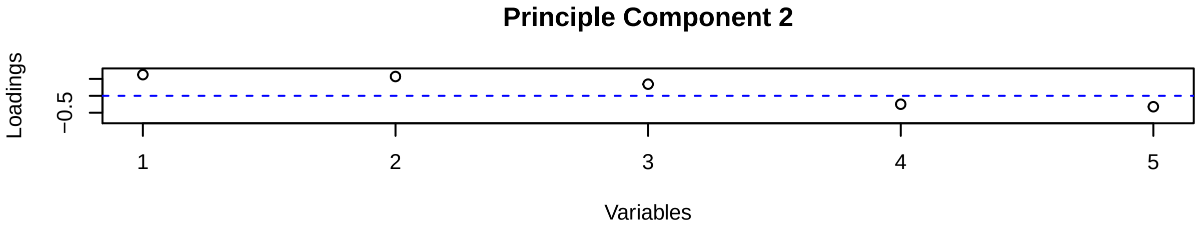


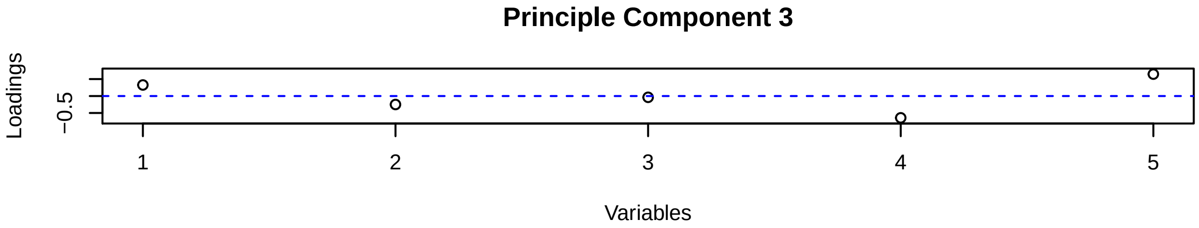
Or computing this directly, we get 89.88% of the total variance is explained by the first three principle components..

Part c.

Let's take a look at the loadings for the first three principle components.







These loading plots helps us with the following interpretations:

- For PC1, there appears to be a relatively equal weighting of all 5 variables. A week that has a high value of this PC means that all the stocks did better than average.
- For PC2, there is a higher load on the first three stocks than the other two. In other words, weeks that have a high value of this PC had higher than average prices on the financial stocks and a dip in the oil stocks.
- For PC3, there is no easy interpretation in my opinion. There is high loading on Exxon and JPMorgan, but there is no connection that I know of between these companies other than the fact that they are some of the largest companies on the exchange.

Question 3

Here we have $\tilde{\mathbf{L}}$ which is a *(p x m)* matrix of factor loadings where the columns are the eigenvectors multipled by the square root of their respective eigenvaleus.

We also have $\tilde{\boldsymbol{\psi}}$ which is a diagonal matrix of the specific variances $\tilde{\psi}_j = s_{jj} - \sum_{k=1}^{m} \tilde{l}_{jk}^2$

Part a.

Consider $(\mathbf{S} - (\tilde{\mathbf{L}}\tilde{\mathbf{L}}^T + \tilde{\boldsymbol{\psi}}))$.

The first m diagonal elements of this matrix will be equal to 0 since we will have $s_{jj} - (\tilde{l}_{jj}^2 + s_{jj} - \tilde{l}_{jj}^2) = 0$ for those elements.

Therefore, the sum of the squared elements of $(\mathbf{S} - (\tilde{\mathbf{L}}\tilde{\mathbf{L}}^T + \tilde{\boldsymbol{\psi}}))$ is $\leq$ the sum of the squared elements of $(\mathbf{S} - \tilde{\mathbf{L}}\tilde{\mathbf{L}}^T)$.

Part b.

Knowing that $\mathbf{S} = \sum_{j=1}^{p} \lambda_j e_j e_j^T$, let's rewrite $\mathbf{S} - \tilde{\mathbf{L}}\tilde{\mathbf{L}}^T$ in a similar form.

We know that $\tilde{\mathbf{L}}$ is a p x m matrix so $\tilde{\mathbf{L}}\tilde{\mathbf{L}}^T$ will be a p x p matrix that can be rewritten as $\sum_{j=1}^{m} \lambda_j e_j e_j^T$ where e_j will be the eigenvector j.

Then we have $\mathbf{S} - \tilde{\mathbf{L}}\tilde{\mathbf{L}}^T = \sum_{j=1}^{p} \lambda_j e_j e_j^T - \sum_{j=1}^{m} \lambda_j e_j e_j^T = \sum_{j=m+1}^{p} \lambda_j e_j e_j^T$

Part c.

Now we have to calculate trace $\left(\mathbf{S} - \tilde{\mathbf{L}}\tilde{\mathbf{L}}^T \left(\mathbf{S} - \tilde{\mathbf{L}}\tilde{\mathbf{L}}^T\right)^T\right)$.

This quantity is equal to:

$$\begin{aligned} \text{trace}\left(\sum_{j=m+1}^{p} (\lambda_j e_j e_j^T)^2\right) &= \sum_{j=m+1}^{p} \lambda_j^2 e_{jj}^2 \\ &= \sum_{j=m+1}^{p} \lambda_j^2 \end{aligned}$$

The last equality comes from the fact that if an eigenvector is squared, it equals 1.

Part d.

Combining our results from part (a) and part (c), we have finally shown that the sum of the squared elements of $(\mathbf{S} - (\tilde{\mathbf{L}}\tilde{\mathbf{L}}^T + \tilde{\boldsymbol{\psi}})) \leq \sum_{i=m+1}^{p} \lambda_i^2$

Question 4

For this problem, we will analyze some physiological data from a sample of adults who are at least 65 years old.

Part a.

Let's begin by performing a principle components factor analysis on the correlation matrix.

Here is a factor analysis with $m = 2$ with the loadings for the first 2 components displayed:

	X1	X2
weight	-0.6703222	-0.1119984
height	-0.8558640	-0.1273154
physact	-0.0811764	0.1448073
ldl	0.2341731	0.1276558
alb	-0.1446212	0.0751346
crt	-0.5202573	-0.4098717
plt	0.4581266	0.1068043
sbp	0.1758132	-0.6165407
aai	-0.2774001	0.6832050
fev	-0.7343154	0.1860528
dsst	-0.1265959	0.5661479
atrophy	-0.1541949	-0.3419670

Here are the specific variances:

Table 1: Table continues below

weight	height	physact	ldl	alb	crt	plt	sbp	aai
0.5381	0.2513	0.9724	0.9289	0.9734	0.5613	0.7787	0.589	0.4563

fev	dsst	atrophy
0.4262	0.6635	0.8593

Here is factor analysis with *m = 3*, again showing the loadings for the first three factors:

	X1	X2	X3
weight	-0.6703222	-0.1119984	0.2333167
height	-0.8558640	-0.1273154	0.0380004
physact	-0.0811764	0.1448073	-0.3223682
ldl	0.2341731	0.1276558	0.6945372
alb	-0.1446212	0.0751346	0.5568304
crt	-0.5202573	-0.4098717	-0.0067606
plt	0.4581266	0.1068043	0.2744162
sbp	0.1758132	-0.6165407	0.0219366
aai	-0.2774001	0.6832050	-0.1433424
fev	-0.7343154	0.1860528	0.0247905
dsst	-0.1265959	0.5661479	0.2196877
atrophy	-0.1541949	-0.3419670	0.2996178

And here are the specific variances:

```
##    weight    height   physact       ldl       alb       crt       plt 
## 0.4836878 0.2498436 0.8685200 0.4464850 0.6633794 0.5612918 0.7034086 
##       sbp       aai       fev      dsst   atrophy 
## 0.5884861 0.4357331 0.4255508 0.6151874 0.7695116
```

I go over the interpretation of these loadings in **part e.** as part of the discussion of PCFA vs MLFA.

Part b.

Let's get the residual matrix for these factor analyses.

Here is the residual matrix for *m = 2*:

	weight	height	physact	ldl	alb	crt	plt	sbp	
weight	0.0000000	-0.0402050	-0.0662336	0.1748394	-0.0417897	-0.1406534	0.1696076	0.0592603	-0.02
height	-0.0402050	0.0000000	0.0138953	0.0601230	-0.0260110	-0.1325953	0.1109309	-0.0029647	-0.07
physact	-0.0662336	0.0138953	0.0000000	-0.0312329	-0.0078562	-0.0092065	0.0127151	0.1057341	-0.04
ldl	0.1748394	0.0601230	-0.0312329	0.0000000	0.1488084	0.0427977	0.0758695	0.0073150	-0.08
alb	-0.0417897	-0.0260110	-0.0078562	0.1488084	0.0000000	-0.0001600	-0.0060653	0.0550413	-0.06
crt	-0.1406534	-0.1325953	-0.0092065	0.0427977	-0.0001600	0.0000000	0.1273027	-0.1622632	0.08
plt	0.1696076	0.1109309	0.0127151	0.0758695	-0.0060653	0.1273027	0.0000000	0.0163149	-0.02
sbp	0.0592603	-0.0029647	0.1057341	0.0073150	0.0550413	-0.1622632	0.0163149	0.0000000	0.14
aai	-0.0220425	-0.0774361	-0.0439287	-0.0807854	-0.0604241	0.0897712	-0.0247630	0.1401281	0.00
fev	-0.1351319	-0.0277578	0.0156104	0.0843566	-0.0547046	-0.0829590	0.1368117	0.1322215	-0.10
dsst	0.0358219	-0.0152872	-0.1240172	-0.0418203	0.0079228	0.0198041	0.0142077	0.2089510	-0.21
atrophy	-0.0767750	-0.0539766	-0.0452959	0.0610405	0.0511678	-0.0660179	0.0452226	-0.1231650	0.10

Here is the residual matrix for m = *3*:

	weight	height	physact	ldl	alb	crt	plt	sbp	
weight	0.0000000	-0.0490711	0.0089803	0.0127923	-0.1717076	-0.1390760	0.1055817	0.0541421	0.01
height	-0.0490711	0.0000000	0.0261454	0.0337304	-0.0471707	-0.1323384	0.1005030	-0.0037983	-0.07
physact	0.0089803	0.0261454	0.0000000	0.1926639	0.1716483	-0.0113859	0.1011782	0.1128058	-0.09
ldl	0.0127923	0.0337304	0.1926639	0.0000000	-0.2379310	0.0474932	-0.1147228	-0.0079208	0.01
alb	-0.1717076	-0.0471707	0.1716483	-0.2379310	0.0000000	0.0036045	-0.1588686	0.0428263	0.01
crt	-0.1390760	-0.1323384	-0.0113859	0.0474932	0.0036045	0.0000000	0.1291579	-0.1621149	0.08
plt	0.1055817	0.1005030	0.1011782	-0.1147228	-0.1588686	0.1291579	0.0000000	0.0102952	0.01
sbp	0.0541421	-0.0037983	0.1128058	-0.0079208	0.0428263	-0.1621149	0.0102952	0.0000000	0.14
aai	0.0114016	-0.0719891	-0.0901377	0.0187712	0.0193932	0.0888022	0.0145725	0.1432726	0.00
fev	-0.1409159	-0.0286998	0.0236021	0.0671386	-0.0685087	-0.0827914	0.1300088	0.1316777	-0.09
dsst	-0.0154349	-0.0236354	-0.0531969	-0.1944016	-0.1144060	0.0212894	-0.0460782	0.2041318	-0.18
atrophy	-0.1466808	-0.0653622	0.0512914	-0.1470552	-0.1156685	-0.0639923	-0.0369974	-0.1297377	0.15

Part c.

Now let's do a maximum likelihood factor analysis based on the correlation matrix for m = *2* and m = *3*.

We'll do m = *2* first.

```
##
## Call:
## factanal(factors = 2, covmat = as.matrix(physio_data), roatation = "none")
##
## Uniquenesses:
##  weight  height physact     ldl     alb     crt     plt     sbp     aai
##   0.675   0.084   0.988   0.974   0.990   0.828   0.903   0.801   0.526
##     fev    dsst atrophy
##   0.569   0.883   0.960
##
## Loadings:
##         Factor1 Factor2
## weight   0.569
## height   0.956
## physact
## ldl     -0.159
## alb
## crt      0.395  -0.126
## plt     -0.308
## sbp             -0.443
## aai              0.686
## fev      0.592   0.283
## dsst             0.342
## atrophy  0.132  -0.151
##
##                Factor1 Factor2
## SS loadings      1.900   0.918
## Proportion Var   0.158   0.077
## Cumulative Var   0.158   0.235
##
## The degrees of freedom for the model is 43 and the fit was 0.1927
```

Now m = *3*:

```
##
## Call:
## factanal(factors = 3, covmat = as.matrix(physio_data), roatation = "none")
##
## Uniquenesses:
##  weight  height physact     ldl     alb     crt     plt     sbp     aai
##   0.659   0.097   0.988   0.005   0.969   0.821   0.881   0.800   0.517
##     fev    dsst atrophy
##   0.564   0.884   0.960
##
## Loadings:
##         Factor1 Factor2 Factor3
## weight   0.563   0.144
## height   0.945
## physact
## ldl     -0.235   0.967
## alb              0.151
## crt      0.407          -0.108
## plt     -0.316   0.122
## sbp                     -0.441
## aai                      0.695
## fev      0.579           0.304
## dsst                     0.339
## atrophy  0.139          -0.145
##
##                Factor1 Factor2 Factor3
## SS loadings      1.895   1.016   0.945
## Proportion Var   0.158   0.085   0.079
## Cumulative Var   0.158   0.243   0.321
##
## The degrees of freedom for the model is 33 and the fit was 0.1187
```

I go over the interpretation of the loads for these factors in **part e.** when deciding which of the methods produces better results.

Part d.

Let's get the residual matrices for these maximum likelihood factor analyses.

	weight	height	physact	ldl	alb	crt	plt	sbp	
weight	-0.0000023	0.0023162	-0.0660074	0.0947423	-0.0063642	0.0344491	0.0274466	0.0587608	0.02
height	0.0023162	0.0000001	0.0040717	-0.0036651	0.0004082	-0.0082291	0.0010022	-0.0080404	-0.00
physact	-0.0660074	0.0040717	-0.0000004	-0.0211095	0.0051571	-0.0385746	0.0140033	0.0460441	0.01
ldl	0.0947423	-0.0036651	-0.0211095	0.0000002	0.1394198	-0.0698253	0.1472004	-0.0437793	-0.04
alb	-0.0063642	0.0004082	0.0051571	0.1394198	-0.0000004	0.0144284	-0.0343597	0.0082768	-0.00
crt	0.0344491	-0.0082291	-0.0385746	-0.0698253	0.0144284	0.0000029	-0.0395690	-0.0355676	0.01
plt	0.0274466	0.0010022	0.0140033	0.1472004	-0.0343597	-0.0395690	0.0000028	-0.0071077	-0.02
sbp	0.0587608	-0.0080404	0.0460441	-0.0437793	0.0082768	-0.0355676	-0.0071077	0.0000004	-0.02
aai	0.0280506	-0.0045958	0.0111819	-0.0418646	-0.0052414	0.0183770	-0.0281881	-0.0229162	0.00
fev	-0.0116894	0.0013419	0.0404187	0.0336912	-0.0007840	0.0250714	0.0160512	0.0460829	-0.00
dsst	0.0368846	-0.0022674	-0.0638131	0.0065383	0.0521641	-0.1080193	0.0369667	-0.0104249	-0.02
atrophy	-0.0044856	0.0000711	-0.0764792	0.0007147	0.0426749	0.0830978	-0.0285299	0.0010284	0.01

	weight	height	physact	ldl	alb	crt	plt	sbp	
weight	0.0000018	0.0017156	-0.0636775	0.0000493	-0.0220332	0.0374182	0.0146603	0.0625139	0.03
height	0.0017156	-0.0000001	0.0051515	-0.0000088	-0.0007032	-0.0085629	-0.0001920	-0.0092053	-0.00
physact	-0.0636775	0.0051515	-0.0000010	0.0000174	0.0081268	-0.0398054	0.0170025	0.0450592	0.00
ldl	0.0000493	-0.0000088	0.0000174	-0.0000001	0.0000789	-0.0000353	0.0000052	-0.0000955	-0.00
alb	-0.0220332	-0.0007032	0.0081268	0.0000789	0.0000002	0.0232095	-0.0547655	0.0146427	0.00
crt	0.0374182	-0.0085629	-0.0398054	-0.0000353	0.0232095	0.0000000	-0.0281132	-0.0392295	0.01
plt	0.0146603	-0.0001920	0.0170025	0.0000052	-0.0547655	-0.0281132	0.0000002	-0.0002903	-0.02
sbp	0.0625139	-0.0092053	0.0450592	-0.0000955	0.0146427	-0.0392295	-0.0002903	0.0000004	-0.02
aai	0.0321946	-0.0049121	0.0096450	-0.0000042	0.0003195	0.0166107	-0.0214850	-0.0228456	0.00
fev	-0.0201568	0.0018707	0.0414599	0.0000374	-0.0075656	0.0244052	0.0121476	0.0467982	0.00
dsst	0.0358126	-0.0033536	-0.0636151	-0.0001814	0.0510895	-0.1079476	0.0362266	-0.0111851	-0.02
atrophy	-0.0057587	0.0004483	-0.0763726	0.0000438	0.0421739	0.0823826	-0.0284294	0.0011930	0.01

Part e.

Which method produced better results?

One way to decide would be to analyze which analyses produced factors that make for better stories. This is the "recent MBA graduate" way to do it.

The factors that we considered in our Principle Components factor analysis were:

	X1	X2	X3
weight	-0.6703	-0.112	0.2333
height	-0.8559	-0.1273	0.038
physact	-0.08118	0.1448	-0.3224
ldl	0.2342	0.1277	0.6945
alb	-0.1446	0.07513	0.5568
crt	-0.5203	-0.4099	-0.006761
plt	0.4581	0.1068	0.2744
sbp	0.1758	-0.6165	0.02194
aai	-0.2774	0.6832	-0.1433
fev	-0.7343	0.1861	0.02479
dsst	-0.1266	0.5661	0.2197
atrophy	-0.1542	-0.342	0.2996

Some possible interpretations for these factors:

- Factor 1 has relatively high loadings for weight, height, and forced expiratory volume. This factor therefore could be related heavily with the *size* of a person.
- Factor 2 has relatively high loadings for observed variables related to systolic blood pressure, so this factor maybe related to *overall heart health*.
- Factor 3 has relatively high loadings for LDL and ALB, which suggests that this factor is related to *diet*.

In my opinion, theese are some pretty devent **WOW** factors.

Examining the factors from the maximum likelihood method, we have that

```
##
## Loadings:
##         Factor1 Factor2
## weight   0.569
## height   0.956
## physact
## ldl     -0.159
## alb
## crt      0.395  -0.126
## plt     -0.308
## sbp             -0.443
## aai              0.686
## fev      0.592   0.283
## dsst             0.342
## atrophy  0.132  -0.151
##
##                Factor1 Factor2
## SS loadings      1.900   0.918
## Proportion Var   0.158   0.077
## Cumulative Var   0.158   0.235
```

```
##
## Loadings:
##         Factor1 Factor2 Factor3
## weight   0.563   0.144
## height   0.945
## physact
## ldl     -0.235   0.967
## alb              0.151
## crt      0.407          -0.108
## plt     -0.316   0.122
## sbp                     -0.441
## aai                      0.695
## fev      0.579           0.304
## dsst                     0.339
## atrophy  0.139          -0.145
##
##                Factor1 Factor2 Factor3
## SS loadings      1.895   1.016   0.945
## Proportion Var   0.158   0.085   0.079
## Cumulative Var   0.158   0.243   0.321
```

We actually see some parallels between the factors from different methods. The first factor for both ML analyses both load heavily on weight, height, and forced expiratory volumne, similar to the principle components factor analysis. The second factor when $m = 2$ seems to be based around blood pressure, indicating that this is also related to heart health. When $m = 3$, this heart health factor appears third, and the second factor instead puts a lot of load on LDL so this factor might be more related to cholesterol specifically, as opposed to overall diet.

A second way to judge the better approach would be looking at the sum of squared entries for the residual matrices.

From Problem 3, we know that an easy way to do this is to just calculate the trace of the product of the residual matrix and its transpose.

Here is that calculation for $m = 3$, first for the Principle Components factor analysis.

```
## [1] 1.427706
```

Here is that calculation for the maximum likelihood method:

```
## [1] 0.1459336
```

We see that the sum of the squared entries of the residual matrix are much smaller for the MLE method than the PC method, while producing similarly intelligible factors that tell interesting stories.

Part f.

Are the factors from the two methods similar for the $m = 2$ models?

As we explored above, I would say *yes*.

Both analyses produced factors that seemed to be based around the latent factors of *body size*, *diet/cholesterol*, and *heart health/blood pressure* I would say either approach generates some "useful" analyses.