P-set2(s23) (**For your personal use only. Do not circulate or post.**)

1. (a) The covariance matrix of scaled data is the correlation matrix $\boldsymbol{R}$ of the original data. The first two principal components for the standardized (variance = 1) variables are

$$(\pm 1)Y_1 = 0.37Z_2 + 0.37Z_3 + 0.37Z_4 + 0.39Z_5 + 0.40Z_6 + 0.38Z_7 + 0.35Z_8$$

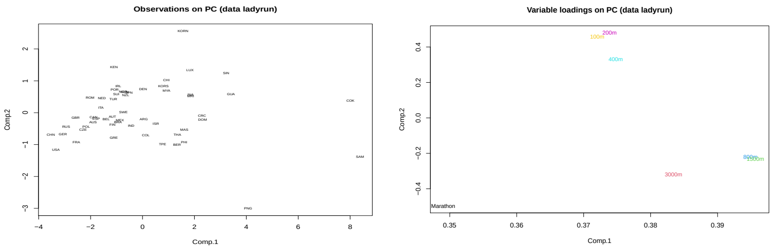
$$(\pm 1)Y_2 = 0.46Z_2 + 0.48Z_3 + 0.33Z_4 - 0.22Z_5 - 0.23Z_6 - 0.32Z_7 - 0.50Z_8$$

where Z_2 = scaled (to variance=1) 100m records, Z_3 = scaled 200m records, ..., Z_8 = scaled Marathon records.

(b) The first two PCs for standardized data are $Y_1 = \pm a_1' X, Y_2 = \pm a_2' X$, where a_1, a_2 are eigenvectors of the first two largest eigenvalues of the correlation matrix $\boldsymbol{R}$.

(c) The percentages of total (scaled data) sample variation explained by the first and second principal components are 81% and 10%, respectively.

(d) i. The left plot below shows the 54 observations in the (PC1, PC2) plane, labeled by country names. Order the countries by their PC1 score values, the resulting is consistent/similar to the performance ranking of the countries by their track records.

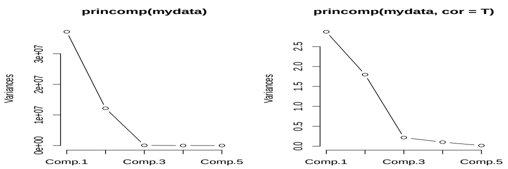


ii. The variable loading on PC plot on the right shows the 7 events in the (PC1, PC2) by their loadings on PC1 and PC2. The PC2 score values sort the events by their distance, separated the short, mid and long distance events. In PC2, the coefficients of the shorter distance events (100m, 200m, 400m) are of opposite signs of the coefficient of longer distance events.

```
### Q1 R commands and partial output ###
ladyrun = read.table("ladyrun22.dat")
colnames(ladyrun)=c("Country","100m","200m","400m","800m","1500m","3000m","Marathon")

### Eigenvalue and eigenvectors ###
eigen(cor(ladyrun[,-1]))
round(eigen(cor(ladyrun[,-1]))$values,2); round(eigen(cor(ladyrun[,-1]))$vectors,2)
eigen(cor(ladyrun[,-1]))$values/sum(eigen(cor(ladyrun[,-1]))$values)

summary(princomp(ladyrun[,-1],cor=T),loading=T)  # PCA

### Q1(d) plots ###
plot(princomp(ladyrun[,-1],cor=T)$scores[,1:2],cex=.3,col=2,pch=16,type="n")
text(princomp(ladyrun[,-1],cor=T)$scores[,1:2],labels=ladyrun$Country,cex=.5)
title(main = "Observations on PC (data ladyrun)")
ladyrun$Country[order(princomp(ladyrun[,-1],cor=T)$scores[,1])]   # rank of Country by PC1

plot(jitter(princomp(ladyrun[,-1],cor=T)$loading[,1:2]),type="n")
text(princomp(ladyrun[,-1],cor=T)$loading[,1:2],labels=(colnames(ladyrun[,-1])),col=7:1,cex=.8)
title("Variable loadings on PC (data ladyrun)")

#dev.copy2pdf()    # save plot to pdf
```

2. (a) Using the original, unscaled data, the first PC summarizes 75% of the variance in the data, while using the scaled data (variable variance 1), two PCs are needed to summarize $\geq 75\%$ of the variance.

(b) The following are the Scree plots of principal values for the original data and the standardized data.



(c) While two components explain most variance in the data in both cases, PCA using unscaled data gives false optimism that the first principal component PC1 explains most variation in the data already (75%). In fact PC1 based on the original data is dominated by the variable `housevalue`, which has much larger magnitude then other variables and is unit-dependent.

```
#### Q2 R commands ##
mydata=read.table("Harman5.txt")
summary(princomp(mydata), loading=T)
summary(princomp(mydata,cor=T), loading=T)
par(mfrow=c(1,2))
screeplot(princomp(mydata),type="l")
screeplot(princomp(mydata,cor=T),type="l")
```

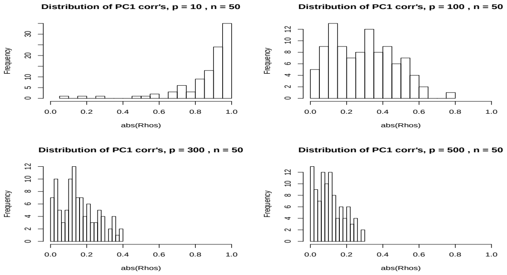
3. (a) $\hat{\boldsymbol{e}}_1$ is the sample estimation of the true $\boldsymbol{v}_1$. We should expect that they are in similar directions, thus we should expect $|\rho| \approx 1$, that is, $\rho \approx 1$ or $\rho \approx -1$. The larger n, then better the approximation should be.

(b) The histogram (top left in the figure below) is concentrated around 1, agree with $|\rho| \approx 1$ expected in (a).

(c) The histogram (top right plot) shows $|\rho|$ moving away from 1, closer to 0, very inconsistent to (a), which means the PCA finds PC vectors deviate away from the true population PCs.

(d) For larger p ($p = 300$ lower left plot, $p = 500$ lower right plot), $|\rho|$ concentrates towards 0 more and more. Therefore sample PC1 is most likely perpendicular to the true one of the population, as the reference paper Theorem 1 claimed.

(e) Here $p = 4000$, $n = 100$. Since $p >> n$, we expect the results will be similar to the ones in (c) and (d). The obtained seemingly important genes could be far from the true important genes in the population.



4. (a) Calculate, show formula used.

i. $Cov(X_1, X_1 + 3X_2 - 2X_3) = Cov(X_1, X_1) + 3Cov(X_1, X_2) - 2Cov(X_1, X_3) = 4 + 0 + 2 = 6 \neq 0$.

ii. $Cov(X_1, X_1 - X_2 + 4X_3) = Cov(X_1, X_1) - Cov(X_1, X_2) + 4Cov(X_1, X_3) = 4 - 0 + (-4) = 6 = 0$.

(b) From $\boldsymbol{X} \sim N_3(\boldsymbol{\mu}, \Sigma)$, we obtain $\begin{pmatrix} X_1 \\ X_3 \end{pmatrix} \sim N\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 4 & -1 \\ -1 & 2 \end{pmatrix}\right)$, marginal distribution of $X_3 \sim N(2, 2)$.

Note that $\begin{pmatrix} 4 & -1 \\ -1 & 2 \end{pmatrix}^{-1} = \frac{1}{7}\begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}$, then

$$f_{X_1|X_3=x_3}(x_1) = \frac{f_{X_1,X_3}(x_1,x_3)}{f_{X_3}(x_3)} = \frac{\frac{1}{2\pi\sqrt{7}} exp\left\{-\frac{1}{2}\begin{pmatrix} x_1-1 \\ x_3-2 \end{pmatrix}^T \begin{pmatrix} 4 & -1 \\ -1 & 2 \end{pmatrix}^{-1} \begin{pmatrix} x_1-1 \\ x_3-2 \end{pmatrix}\right\}}{\frac{1}{\sqrt{2\pi}\sqrt{2}} \exp\left\{-\frac{1}{4}(x_3-2)^2\right\}}$$

$$\text{(simplify)} \qquad = \frac{1}{\sqrt{7\pi}} e^{-\frac{\left(x_1-2+\frac{x_3}{2}\right)^2}{7}} \qquad \Longrightarrow \quad X_1|X_3=x_3 \sim N\left(2-\frac{x_3}{2}, \frac{7}{2}\right)$$

which can be verified by the normal conditional distribution formula (on page 158 in J&W, and lecture notes).

(c) Since $X_1 \perp\!\!\!\perp X_2$, the distribution of $X_1|(X_2=x_2, X_3=x_3)$ is the same as that of $X_1|X_3=x_3$, obtained in (b).

5. (a) The m-factor loading matrix $\boldsymbol{L}_m$ using the principal component method is $\boldsymbol{L}_m = \left[\sqrt{\lambda_1}\boldsymbol{e}_1, \cdots, \sqrt{\lambda_m}\boldsymbol{e}_m\right]$. Thus factor 1 loadings are the same for $m=1$ or $m=2$. The factor loadings for $m=2$ is shown in Table 1 and Table 2 below, corresponding to using the original data with the covariance matrix or using the normalized data with the correlation matrix.

i. Table 1: using original data with the covariance matrix, PC method (by command `princomp`)

```
             [F1 loading]  [F2 loading]
population    128.29     3290.11
schooling       1.48       -0.03
employment    164.23     1155.73
professional   86.34       45.09
housevalue   6095.61     -101.02
```

The $\boldsymbol{\Psi}$ matrix is diagonal with values

```
  population    schooling   employment professional   housevalue
  992273.8     1.011319     177918.8     3720.733     3378808
```

If you use `eigen` command directly or use `prcomp` instead of `princomp`, the two factors will be

```
          [F1 loading]  [F2 loading]
population    134.00    3436.41
schooling       1.54      -0.03
employment    171.53    1207.12
professional   90.18      47.10
housevalue   6366.66    -105.51
```

The estimated $\boldsymbol{\Psi}$ matrix diagonal elements will be

```
  population    schooling   employment professional   housevalue
    6702.29       0.81        54038.12     2858.22      21.77
```

ii. Table 2: using scaled variance=1 data with the correlation matrix, PC method

```
##            [F1 loading]  [F2 loading]
## population   0.58      0.81
## schooling    0.77     -0.54
## employment   0.67      0.73
## professional 0.93     -0.10
## housevalue   0.79     -0.56
```

The $\boldsymbol{\Psi}$ matrix is diagonal with values

```
  population    schooling   employment professional   housevalue
       0.01          0.11         0.02          0.12          0.06
```

(b) The maximum likelihood factor loadings with $m=2$ are in the table below. (R message says $m=3$ is too many.)

```
## Loadings:
##              Factor1 Factor2
## population            0.997
## schooling     0.898
## employment    0.137   0.972
## professional  0.798   0.423
## housevalue    0.962
```

(c) The following are residual matrices (all with m=2 factors), which are

1) Difference between sample covariance matrix $\boldsymbol{S}$ and PC estimates of $L_2L_2' + \Psi$, using original data with covariance matrix.

```
##              population schooling employment professional housevalue
## population            0       -32     328566        14067      41255
## schooling           -32         0        134           16        816
## employment       328566       134          0         7129      79321
## professional      14067        16       7129            0      47368
## housevalue        41255       816      79321        47368          0
```

If you use `eigen` command directly or use `prcomp` instead of `princomp`, the residual matrix will be (just as bad)

```
##              population schooling employment professional housevalue
## population         0.00    -40.50  -19030.20      -428.05     377.72
## schooling        -40.50      0.00     114.93         4.37      -2.31
## employment    -19030.20    114.93       0.00      1101.71   -1071.00
## professional    -428.05      4.37    1101.71         0.00     -61.16
## housevalue       377.72     -2.31   -1071.00       -61.16       0.00
```

2) Difference between sample correlation matrix $\boldsymbol{R}$ and PC estimates $L_2L_2' + \Psi$, using scaled variance=1 data with correlation matrix. $\sum_{i,j}(residual_{ij})^2 = 0.2435$.

```
##              population schooling employment professional housevalue
## population         0.00      0.00       0.00        -0.02       0.01
## schooling          0.00      0.00       0.03        -0.08      -0.05
## employment         0.00      0.03       0.00        -0.04       0.00
## professional      -0.02     -0.08      -0.04         0.00      -0.02
## housevalue         0.01     -0.05       0.00        -0.02       0.00
```

3) Differences between $\boldsymbol{R}$ and ML estimates $L_mL_m' + \Psi$. $\sum_{i,j}(residual_{ij})^2 = 0.0041$.

```
##              population schooling employment professional housevalue
## population            0      0.00       0.00         0.00       0.00
## schooling             0      0.00       0.04        -0.02       0.00
## employment            0      0.04       0.00        -0.01      -0.01
## professional          0     -0.02      -0.01         0.00       0.01
## housevalue            0      0.00      -0.01         0.01       0.00
```

The maximum likelihood estimation does a better job of accounting for the covariances in $\boldsymbol{R}$ than the PC method estimates using scaled data with correlation matrix (by observation and by sum of squared (or absolute value) entries of the residual matrices). The PC method with unscaled data did a terrible job to estimate $\boldsymbol{S}$, due the huge difference in magnitudes and variance of the original variables.

```
### Some Q5 R commands ###
PCcov=princomp(mydata,cor=F).                                       % or prcomp(mydata,scale=F)
PCcov=princomp(mydata,cor=F)
Lcov=PCcov$loading[,1:5]%*%diag((PCcov$sdev))[,1:2]                 % or PCcov$rotation...
round(Lcov,2)
diag(cov(mydata)-Lcov%*%t(Lcov))
hatPsi = diag(diag(cov(mydata)-Lcov%*%t(Lcov)),5,5)
factanal(x = mydata, factors = 2)
```

6. (a) There are $p=3$ variables. The factor model and the dimensions of each term is

$$\boldsymbol{X}_{3\times1} = \boldsymbol{\mu}_{3\times1} + L_{3\times1}\boldsymbol{F}_{1\times1} + \boldsymbol{\varepsilon}_{3\times1}$$

(b) Setup a system of equations and solve for $\ell_i, \psi_i, i=1,2,3$.

$$\begin{cases} \ell_1^2 + \psi_1 = 5 \\ \ell_1\ell_2 = 2 \\ \ell_1\ell_3 = 3 \\ \ell_2^2 + \psi_2 = 6 \\ \ell_2\ell_3 = 6 \\ \ell_3^2 + \psi_3 = 10 \end{cases} \Rightarrow \quad \ell_1 = 1,\ \ell_2 = 2,\ \ell_3 = 3,\ \psi_1 = 4,\ \psi_2 = 2,\ \psi_3 = 1$$

Note that equation 2 times eq. 3 = eq. 5, which gives $\ell_1^2 = 1$, we choose positive loadings $\ell_i > 0$.

(c) $\%Var(X_i)$ explained by the common factor is $\frac{\ell_i^2}{Var(X_i)}$, for $i=1,2,3$, they are $(\frac{1}{5}, \frac{4}{6}, \frac{9}{10}) = (20\%, 67\%, 90\%)$.

(d) The number of variables and the factor model are the same as in (a). Solving

$$\begin{cases} \ell_1^2 + \psi_1 = 5 \\ \ell_1\ell_2 = 2 \\ \ell_1\ell_3 = 3 \\ \ell_2^2 + \psi_2 = 6 \\ \ell_2\ell_3 = 6 \\ \ell_3^2 + \psi_3 = 8 \end{cases} \Rightarrow \quad \ell_1 = 1,\ \ell_2 = 2,\ \ell_3 = 3,\ \psi_1 = 4,\ \psi_2 = 2,\ \psi_3 = -1$$

$\%Var(X_i)$ explained by the common factor for $i=1,2,3$ are $(\frac{1}{5}, \frac{4}{6}, \frac{9}{8}) = (20\%, 67\%, 112.5\%)$. Although the math solution for the factor model exists, but it is unreasonable for the factor explaining more than 100% of the variation and having negative specific factor for X_3. The model is statistically invalid for the given Σ.