STAT 32950

Robert Winter

HW #2

1. $X \sim N_3(\mu, \Sigma)$, $\mu = (1, -3, 2)$,

$$\Sigma = \begin{bmatrix} 4 & 0 & -1 \\ 0 & 7 & 0 \\ -1 & 0 & 2 \end{bmatrix}.$$

1a(i) $Y_1 = X_1 + 3X_2 - 2X_3$. $X_1 \perp\!\!\!\perp Y_1$?

- $\mathbb{E}[Y_1] = \mathbb{E}[X_1 + 3X_2 - 2X_3]$
  $= \mathbb{E}[X_1] + 3\mathbb{E}[X_2] - 2\mathbb{E}[X_3]$
  $= 1 + 3(-3) - 2(2)$
  $= -12$
- $\mathbb{E}[X_1 Y_1] = \mathbb{E}[X_1(X_1 + 3X_2 - 2X_3)]$
  $= \mathbb{E}[X_1^2 + 3X_1X_2 - 2X_1X_3]$
  $= \mathbb{E}[X_1^2] + 3\mathbb{E}[X_1X_2] - 2\mathbb{E}[X_1X_3]$
  $= (\mathrm{Var}(X_1) + \mathbb{E}[X_1]^2) + 3(\mathrm{Cov}(X_1, X_2) + \mathbb{E}[X_1]\mathbb{E}[X_2])$
  $\quad - 2(\mathrm{Cov}(X_1, X_3) + \mathbb{E}[X_1]\mathbb{E}[X_3])$
  $= (4 + 1^2) + 3(0 + (1)(-3)) - 2(-1 + (1)(2))$
  $= 5 + 3(-3) - 2(1)$
  $= -6$
- $\mathrm{Cov}(X_1, Y_1) = \mathbb{E}[X_1 Y_1] - \mathbb{E}[X_1]\mathbb{E}[Y_1]$
  $= -6 - (1)(-12)$
  $= 6$
- Since $\mathrm{Cov}(X_1, Y_1) = 6 \neq 0$, $X_1 \not\perp\!\!\!\perp Y_1$.

1a(ii) $Y_2 = X_1 - X_2 + 4X_3$. $X_1 \perp\!\!\!\perp Y_2$?

- $\mathbb{E}[Y_2] = \mathbb{E}[X_1 - X_2 + 4X_3]$
  $= \mathbb{E}[X_1] - \mathbb{E}[X_2] + 4\mathbb{E}[X_3]$
  $= 1 - (-3) + 4(2)$
  $= 12$

→

- $\mathbb{E}[X_1 Y_2] = \mathbb{E}[X_1(X_1 - X_2 + 4X_3)]$
$= \mathbb{E}[X_1^2 - X_1X_2 + 4X_1X_3]$
$= \mathbb{E}[X_1^2] - \mathbb{E}[X_1X_2] + 4\mathbb{E}[X_1X_3]$
$= (\mathrm{Var}(X_1) + \mathbb{E}[X_1]^2) - (\mathrm{Cov}(X_1, X_2) + \mathbb{E}[X_1]\mathbb{E}[X_2])$
$+ 4(\mathrm{Cov}(X_1, X_3) + \mathbb{E}[X_1]\mathbb{E}[X_3])$
$= (4 + 1^2) - (0 + (1)(-3)) + 4(-1 + (1)(2))$
$= 5 - (-3) + 4(1)$
$= 12$
- $\mathrm{Cov}(X_1, Y_2) = \mathbb{E}[X_1Y_2] - \mathbb{E}[X_1]\mathbb{E}[Y_2]$
$= 12 - (1)(12)$
$= 0.$
- Let $a, b \in \mathbb{R}$. Then
$aX_1 + bY_2 = aX_1 + b(X_1 - X_2 + 4X_3)$
$= (a+b)X_1 - bX_2 + 4bX_3$,
which is a linear combination of X_1, X_2, X_3.
Since $X \sim N_3(\mu, \Sigma)$, $aX_1 + bY_2 = (a+b)X_1 - bX_2 + 4bX_3$ has normal distribution $\forall a, b \in \mathbb{R}$. Thus, (X_1, Y_2) has bivariate normal distribution.
- Since $(X_1, Y_2) \sim N_2(\quad)$ and $\mathrm{Cov}(X_1, Y_2) = 0$, we conclude $X_1 \perp\!\!\!\perp Y_2$.

1b. Derive $\mathbb{E}[X_3 | X_1 = x_1]$, $\mathrm{Var}(X_3 | X_1 = x_1)$.

- Let $\mu_1 = 1$, $\mu_{(23)} = (-3, 2)$, $\Sigma_{11} = 4$, $\Sigma_{(1)(23)} = \begin{bmatrix} 0 & -1 \end{bmatrix}$, $\Sigma_{(23)(1)} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$, $\Sigma_{(23)(23)} = \begin{bmatrix} 7 & 0 \\ 0 & 2 \end{bmatrix}$, so
$$\mu = \begin{bmatrix} \mu_1 \\ \mu_{23} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{(1)(23)} \\ \Sigma_{(23)(1)} & \Sigma_{(23)(23)} \end{bmatrix}.$$
- Then
$$\begin{bmatrix} X_2 \\ X_3 \end{bmatrix} | X_1 = x_1 \sim N_2\left(\mu_{23} + \Sigma_{(23)(1)}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{(23)(23)} - \Sigma_{(23)(1)}\Sigma_{11}^{-1}\Sigma_{(1)(23)}\right)$$
$$= N_2\left(\begin{bmatrix} -3 \\ 2 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix}\left(\tfrac{1}{4}\right)(x_1 - 1), \begin{bmatrix} 7 & 0 \\ 0 & 2 \end{bmatrix} - \begin{bmatrix} 0 \\ -1 \end{bmatrix}\left(\tfrac{1}{4}\right)\begin{bmatrix} 0 & -1 \end{bmatrix}\right)$$
$$= N_2\left(\begin{bmatrix} -3 \\ 2 \end{bmatrix} + \begin{bmatrix} 0 \\ -1/4 \end{bmatrix}(x_1 - 1), \begin{bmatrix} 7 & 0 \\ 0 & 2 \end{bmatrix} - \frac{1}{4}\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

$$= N_2\left(\begin{bmatrix}-3\\2\end{bmatrix} + \begin{bmatrix}0\\-\frac{1}{4}x_1+\frac{1}{4}\end{bmatrix}, \begin{bmatrix}7&0\\0&2\end{bmatrix} - \begin{bmatrix}0&0\\0&\frac{1}{4}\end{bmatrix}\right)$$

$$= N_2\left(\begin{bmatrix}-3\\\frac{9}{4}-\frac{1}{4}x_1\end{bmatrix}, \begin{bmatrix}7&0\\0&\frac{7}{4}\end{bmatrix}\right)$$

- So,
$$X_3 | X_1 = x_1 \sim N\left(\frac{9}{4} - \frac{1}{4}x_1, \frac{7}{4}\right).$$
- That is,
$$\mathbb{E}[X_3 | X_1 = x_1] = \frac{9}{4} - \frac{1}{4}x_1$$
and
$$\mathrm{Var}(X_3 | X_1 = x_1) = \frac{7}{4}.$$

1c. Specify the distribution of $X_1 | X_3 = x_3$, & write $f_{X_1|X_3=x_3}(x_1)$.

- Since $X \sim N_3(\mu, \Sigma)$, $aX_1 + bX_2 + cX_3$ has normal distribution $\forall a, b, c \in \mathbb{R}$. In particular, $aX_1 + 0X_2 + cX_3 = aX_1 + cX_3$ has normal distribution $\forall a, c \in \mathbb{R}$. So, (X_1, X_3) has bivariate normal distribution.
- So,
$$X_1 | X_3 = x_3 \sim N\left(\mu_1 + \frac{\sigma_1}{\sigma_3}\rho_{13}(x_3 - \mu_3), \sigma_1^2(1-\rho_{13}^2)\right)$$
$$= N\left(1 + \frac{\sqrt{4}}{\sqrt{2}} \cdot \frac{-1}{\sqrt{4}\sqrt{2}}(x_3 - 2), 4\left(1 - \left(\frac{-1}{\sqrt{4}\sqrt{2}}\right)^2\right)\right)$$
$$= N\left(1 + \frac{-1}{2}(x_3 - 2), 4\left(1 - \frac{1}{8}\right)\right)$$
$$= N\left(1 - \frac{1}{2}x_3 + 1, 4 \cdot \frac{7}{8}\right)$$
$$\Rightarrow X_1 | X_3 = x_3 \sim N\left(2 - \frac{1}{2}x_3, \frac{7}{2}\right)$$

- So, $f_{X_1|X_3=x_3}(x_1) = \frac{1}{\sqrt{2\pi \cdot \frac{7}{2}}} e^{-\frac{1}{2 \cdot \frac{7}{2}}\left(x_1 - \left(2 - \frac{1}{2}x_3\right)\right)^2}$

$$\Rightarrow f_{X_1|X_3=x_3}(x_1) = \frac{1}{\sqrt{7\pi}} e^{-\frac{1}{7}\left(x_1 - \left(2 - \frac{1}{2}x_3\right)\right)^2}$$

1d. Specify the distribution of $X_1 | (X_2, X_3) = (x_2, x_3)$. What is $f_{X_1|(X_2,X_3)=(x_2,x_3)}(x_1)$?

- Let $\mu_1 = 1$, $\mu_{23} = (-3, 2)$, $\Sigma_{11} = 4$, $\Sigma_{(1)(23)} = [0, -1]$, $\Sigma_{(23)(1)} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$, $\Sigma_{(23)(23)} = \begin{bmatrix} 7 & 0 \\ 0 & 2 \end{bmatrix}$, so

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_{23} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{(1)(23)} \\ \Sigma_{(23)(1)} & \Sigma_{(23)(23)} \end{bmatrix}.$$

- Then

$$\begin{aligned} X_1 | (X_2, X_3) = (x_2, x_3) &\sim N\left(\mu_1 + \Sigma_{(1)(23)} \Sigma_{(23)(23)}^{-1} \left(\begin{bmatrix} x_2 \\ x_3 \end{bmatrix} - \mu_{23}\right), \right. \\ &\qquad \left. \Sigma_{11} - \Sigma_{(1)(23)} \Sigma_{(23)(23)}^{-1} \Sigma_{(23)(1)}\right) \\ &= N\left(1 + [0 \ \ -1] \begin{bmatrix} 1/7 & 0 \\ 0 & 1/2 \end{bmatrix} \left(\begin{bmatrix} x_2 \\ x_3 \end{bmatrix} - \begin{bmatrix} -3 \\ 2 \end{bmatrix}\right), \right. \\ &\qquad \left. 4 - [0 \ \ -1] \begin{bmatrix} 1/7 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \end{bmatrix}\right) \\ &= N\left(1 + \begin{bmatrix} 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} x_2 + 3 \\ x_3 - 2 \end{bmatrix}, \right. \\ &\qquad \left. 4 - \begin{bmatrix} 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 0 \\ -1 \end{bmatrix}\right) \\ &= N\left(1 + -\tfrac{1}{2}(x_3 - 2), \ 4 - \tfrac{1}{2}\right) \\ &= N\left(1 - \tfrac{1}{2} x_3 + 1, \ \tfrac{7}{2}\right) \end{aligned}$$

$$\Rightarrow \boxed{X_1 | (X_2, X_3) = (x_2, x_3) \sim N\left(2 - \tfrac{1}{2} x_3, \tfrac{7}{2}\right)}$$

- So, $f_{X_1|(X_2,X_3)=(x_2,x_3)}(x_1) = \frac{1}{\sqrt{2\pi \cdot \frac{7}{2}}} e^{\frac{-1}{2 \cdot \frac{7}{2}} (x_1 - (2 - \frac{1}{2} x_3))^2}$

$$\boxed{f_{X_1|(X_2,X_3)=(x_2,x_3)}(x_1) = \frac{1}{\sqrt{7\pi}} e^{-\frac{1}{7}(x_1 - (2 - \frac{1}{2} x_3))^2}}$$

STAT 32950: Homework 2

Robert Winter

Table of Contents

1 Exercise 2: Basic Principal Component Analysis

Use the same dataset `ladyrun24.dat` on national track records for women as used in Assignment 1. The nominal variable `Country` should be excluded in numerical calculations.

1.1 Part (a)

Determine the first two principal components for the scaled data with variable variance $= 1$, express the two principal components as linear combinations of the scaled variables. State the meaning of the (scaled) variables that the principal components consist of. (Recall that a PC variable is unique up to a multiple of ± 1).

As shown in the R output below, the first principal component of the scaled data has the form

$$\begin{aligned} PC1 = {} & 0.372 \times 100\text{m} + 0.374 \times 200\text{m} + 0.375 \times 400\text{m} + 0.395 \times 800\text{m} \\ & + 0.396 \times 1500\text{m} + 0.383 \times 3000\text{m} + 0.349 \times \text{Marathon}, \end{aligned}$$

and the second principal component of the scaled data has the form

$$\begin{aligned} PC2 = {} & 0.458 \times 100\text{m} + 0.480 \times 200\text{m} + 0.331 \times 400\text{m} - 0.221 \times 800\text{m} \\ & - 0.231 \times 1500\text{m} - 0.318 \times 3000\text{m} - 0.497 \times \text{Marathon}. \end{aligned}$$

Each of the above principal components may alternatively be expressed by multiplying all coefficients by a factor of -1, essentially "flipping" the direction of the PC vector to point in the opposite direction.

```
PC = princomp(dplyr::select(ladyrun, select = -Country), cor = T)
summary(PC, loading = TRUE, digits = 3)
```

```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4    Comp.5
Standard deviation     2.387740 0.8584325 0.53721402 0.33824062 0.2934077
Proportion of Variance 0.814472 0.1052723 0.04122842 0.01634382 0.0122983
Cumulative Proportion  0.814472 0.9197443 0.96097276 0.97731657 0.9896149
                            Comp.6      Comp.7
Standard deviation     0.225255770 0.148174712
Proportion of Variance 0.007248595 0.003136535
Cumulative Proportion  0.996863465 1.000000000

Loadings:
```

```
          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
d100m      0.372  0.458  0.149  0.526  0.155  0.568
d200m      0.374  0.480         0.111        -0.750 -0.204
d400m      0.375  0.331 -0.487 -0.508 -0.456  0.200
d800m      0.395 -0.221 -0.148 -0.377  0.769  0.121 -0.156
d1500m     0.396 -0.231  0.425 -0.140        -0.143  0.750
d3000m     0.383 -0.318  0.477        -0.377  0.140 -0.598
Marathon   0.349 -0.497 -0.553  0.534 -0.137 -0.146
```

By scaling the data such that each track distance has variance 1, we are essentially normalizing the national records for each distance so that their deviations from the mean are on the same scale. For example, consider the 100m dash compared to the marathon run: while elite athletes can complete the former in a matter of seconds, the latter is run over the course of more than two hours. As such, national records for the marathon can vary by minutes, but national records for the 100m dash may vary by only fractions of a second — since the race is so short, there is not as much time for differences in performance to accumulate. Scaling each variable to unit variance ensures that the multi-minute variability of marathon records does not overpower the decisecond variability of 100m dash records in calculating principal components.

1.2 Part (b)

Compare the two principal components PC1 and PC2 with the eigenvectors of the sample correlation matrix (which you obtained in Question 1(f) in Assignment 1).

In Assignment 1, Question 1(f), we found that the eigenvectors of the correlation matrix of the `ladyrun` data were as follows:

```r
eigen(cor(dplyr::select(ladyrun, -Country)))$vectors %>% round(3)
```

```
      [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]
[1,] -0.372 -0.458 -0.149  0.526 -0.155  0.568  0.083
[2,] -0.374 -0.480 -0.074  0.111 -0.092 -0.750 -0.204
[3,] -0.375 -0.331  0.487 -0.508  0.456  0.200  0.074
[4,] -0.395  0.221  0.148 -0.377 -0.769  0.121 -0.156
[5,] -0.396  0.231 -0.425 -0.140  0.082 -0.143  0.750
[6,] -0.383  0.318 -0.477 -0.075  0.377  0.140 -0.598
[7,] -0.349  0.497  0.553  0.534  0.137 -0.146  0.033
```

Observe that the first two columns of the output above, which correspond to the first and second eigenvectors of the correlation matrix, are identical to the first and second principal components we identified in Part (a) (though with opposite signs, as principal components are unique up to a multiple of ± 1).

1.3 Part (c)

What are the percentages of total (scaled data) sample variation explained by the first and second principal components you described in Part (a)?

As shown in the code output of Part (a), the first principal component explains roughly 81.4% of the total variation in the (scaled) data. The second principal component explains an incremental 10.5% of the total variation in the (scaled) data. So, together, the first two principal components explain nearly 92.0% of the total variation in the (scaled) data.

1.4 Part (d)

Plots and interpretations.

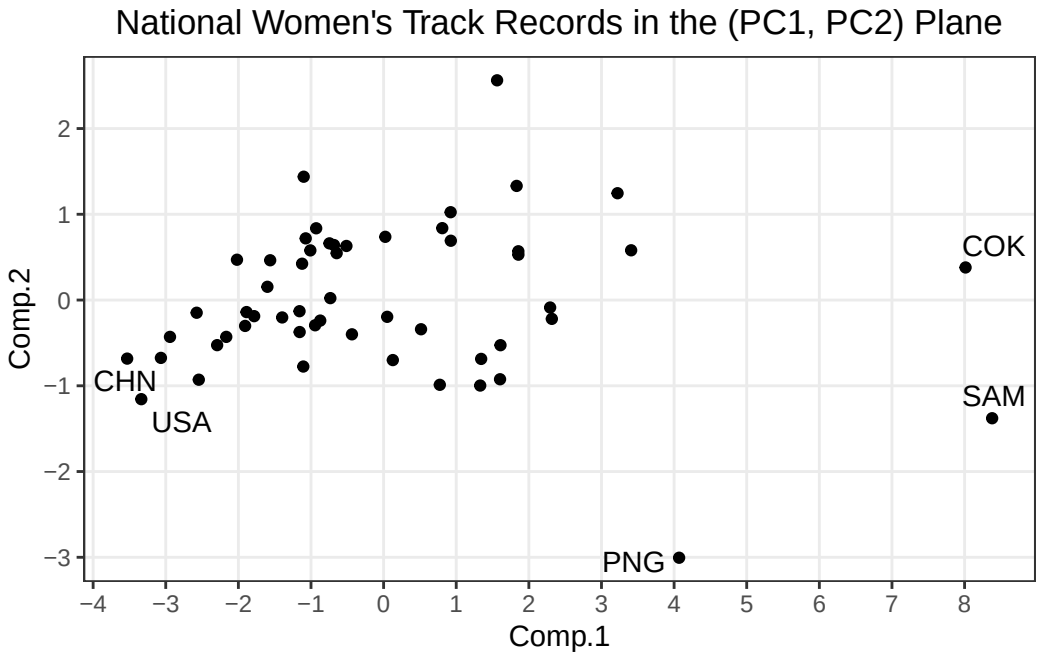
1.4.1 Part (i)

Now every observation has its coordinates in the space of principal components (PC1, ..., PC7) obtained using the scaled data (of variable variance 1). Construct a two-dimensional scatterplot of the 54 observations in the (PC1, PC2) plane. Compare the values of PC1 scores of the countries of the athletes with their approximate ranking (i.e., goodness or levels of overall performance) in track records.

Below, we plot the 54 countries' national women's track records in the (PC1, PC2) plane. In particular, notice that we have labeled the three countries with the largest values of the first principal component (Samoa, the Cook Islands, and Papua New Guinea), as well as the two countries with the smallest values of the first principal component (China and the US). In the excerpt from the data table printed below the scatterplot, notice that the US and China have significantly faster records across all distances compared to Samoa, the Cook Islands, and Papua New Guinea. This suggests that the first principal component captures a country's overall standing across all distances in the dataset, with lower PC1 scores corresponding to faster national records across distances, and higher PC1 scores corresponding to slower national records across distances.

```
ggplot(as.data.frame(PC$scores), aes(x=Comp.1, y = Comp.2,
                                     label = ladyrun$Country)) +
  theme_bw() +
  geom_point() +
  geom_text_repel(aes(label = ifelse(Comp.1 < -3.2 | Comp.1 > 4,
                                as.character(ladyrun$Country), ""))) +
  scale_x_continuous(breaks = c(-10:15),
                     minor_breaks = NULL) +
  scale_y_continuous(minor_breaks = NULL) +
```

```
  ggtitle("National Women's Track Records in the (PC1, PC2) Plane") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
# Rank countries by value of PC1, call out the highest and lowest
rankingPC1 = order(PC$scores[,1], decreasing = T)
ladyrun[rankingPC1,] %>% filter(Country %in% c("SAM", "COK", "PNG", "USA",
↪  "CHN"))
```

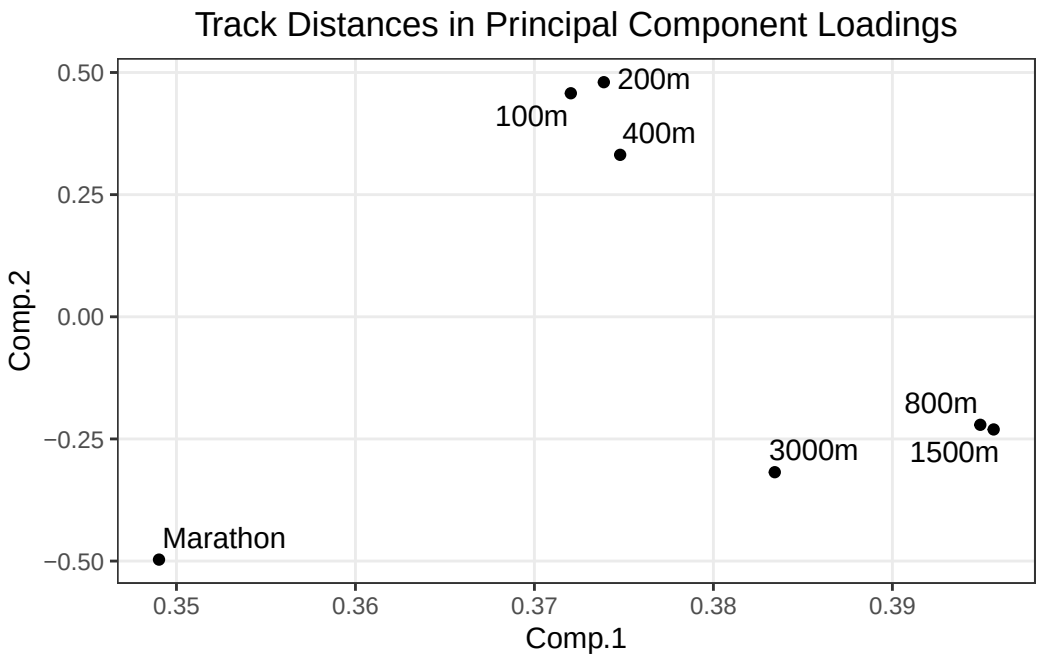
```
  Country d100m d200m d400m d800m d1500m d3000m Marathon
1     SAM 12.38 25.45 56.32  2.29   5.42  13.12   191.58
2     COK 12.52 25.91 61.65  2.28   4.82  11.10   212.33
3     PNG 11.29 23.12 55.18  2.24   4.62  10.21   221.14
4     USA 10.49 21.34 48.70  1.93   3.92   8.43   139.60
5     CHN 10.79 22.01 45.14  1.93   3.84   8.10   139.65
```

1.4.2 Part (ii)

Now every variable has its loading coefficients on each of the principle components (PC1, ..., PC7) obtained using the scaled data (of variable variance 1). Construct a two-dimensional scatterplot of the 7 original variables (all but the `Country` variable) in the (PC1, PC2) plane. Comment on the pattern of the variable loadings (coefficients) in PC2.

Below, we plot the seven race distance variables in the (PC1, PC2) plane using their principal component loadings. Observe that short-distance races (the 100m, 200m, and 400m dashes) all have PC2 loadings greater than 0. Meanwhile, the middle- and long-distances races (the 800m, 1500m, 3000m, and marathon runs) all have PC2 loadings less than 0. This suggests that the second principal component captures a "distance level" (i.e., short vs. middle- or long-distance) component of the data.

```
dists = c("100m", "200m", "400m", "800m", "1500m", "3000m", "Marathon")

ggplot(as.data.frame(PC$loadings[1:7,]),
       aes(x = Comp.1, y = Comp.2, label = dists)) +
  theme_bw() +
  geom_point() +
  geom_text_repel() +
  scale_x_continuous(minor_breaks = NULL) +
  scale_y_continuous(minor_breaks = NULL) +
  ggtitle("Track Distances in Principal Component Loadings") +
  theme(plot.title = element_text(hjust = 0.5))
```



2 Exercise 3: Scaling Effects in Principal Component Analysis

Download data `Harman5.txt`. The measurements are on five socioeconomic variables for each of 12 census tracts (years ago). Use the data to conduct Principal

Component Analysis, using the original data as well as scaled data (by making each variable of variance 1, equivalent to using the correlation matrix.

```
Q3PCAcov = princomp(harman5, cor = FALSE) # raw data
Q3PCAcor = princomp(harman5, cor = TRUE) # scaled data
```

2.1 Part (a)

In each of the two cases, how many principal components are needed to effectively summarize at least 75% of the variability in the data?

As shown below, when we conduct principal component analysis on the raw (non-scaled) data, the first principal component alone explains roughly 75.3% of the total variation in the data.

```
summary(Q3PCAcov, loadings = TRUE)
```

```
Importance of components:
                            Comp.1       Comp.2       Comp.3       Comp.4
Standard deviation     6099.7809499 3488.9515515 2.360509e+02 5.097917e+01
Proportion of Variance    0.7525993    0.2462211 1.127059e-03 5.256792e-05
Cumulative Proportion     0.7525993    0.9988204 9.999474e-01 1.000000e+00
                            Comp.5
Standard deviation     7.210752e-01
Proportion of Variance 1.051711e-08
Cumulative Proportion  1.000000e+00

Loadings:
             Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
population            0.943  0.332
schooling                                 1.000
employment            0.331 -0.943
professional                       -1.000
housevalue    0.999
```

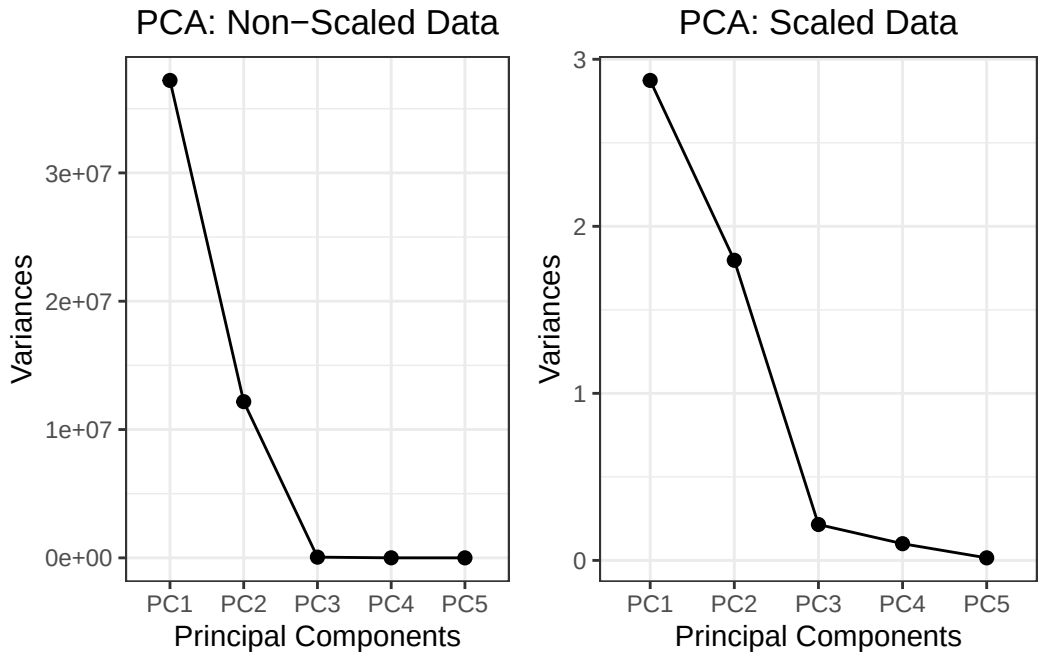
On the other hand, as shown below, when we conduct principal component analysis on the scaled data, two principal components are necessary to explain more than 75% of the total variation in the data. In particular, the first principal component explains 57.5% of the total variation in the data, and the second principal component explains an additional 35.9% of the variation, such that the two components together explain 93.4% of the variation.

```r
summary(Q3PCAcor, loadings = TRUE)
```

```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4      Comp.5
Standard deviation     1.6950851 1.3403955 0.46350500 0.31612348 0.123512644
Proportion of Variance 0.5746627 0.3593320 0.04296738 0.01998681 0.003051075
Cumulative Proportion  0.5746627 0.9339947 0.97696211 0.99694893 1.000000000

Loadings:
             Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
population    0.343  0.602         0.204  0.689
schooling     0.453 -0.406  0.689 -0.354  0.175
employment    0.397  0.542  0.248        -0.698
professional  0.550        -0.664 -0.500
housevalue    0.467 -0.416 -0.140  0.763
```

2.2 Part (b)

Plot two scree plots, one from PCA based on the original data, one based on the standardized data.



2.3 Part (c)

Compare and comment, based on the coefficients (loadings) of the first principal component in each case, and using the results in Parts (a) and (b). Which analysis result is better? Why?

As shown in Part (a), when we perform principal component analysis using non-scaled data, the first principal component ("PC1") is comprised almost entirely of the house value variable. When we perform principal component analysis using scaled data, however, PC1 is comprised of a mixture of all five variables in the dataset, with loadings on each ranging from 0.343 to 0.550.

This discrepancy is a result of the vastly different scales of the five variables. For instance, `housevalue` values are on the fourth order of magnitude (i.e., have values around $10,000$), while `schooling` values are on the first order of magnitude (i.e., have values around 10). When we do not scale the data, variation in `housevalue` dominates the overall variation in the data, and so PC1 practically coincides with the `housevalue` axis. We can also see this in the fact that it just takes PC1 to explain at least 75% of the variation in the non-scaled data, as shown in Part (b). When we normalize the data so that deviations from each variable's mean are on the same scale, deviations from the means of variables other than `housevalue` no longer seem so minuscule. Variation in these variables now contributes to the direction of PC1. Since the total variation in the data is no longer "artificially" dominated by `housevalue`, it now takes a second principal component to explain at least 75% of the variation in the data, as shown in Part (b).

Ultimately, the PCA on the scaled data is the more useful of the two. By putting every variable on the same scale of variability, this analysis allows for dimensionality reduction that reflects the variation in every variable, rather than just the variables that happen to be measured on greater orders of magnitude.

3 Exercise 4: Simulation on Consistency and Sparsity of High Dimensional PCA

Let $X \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$, and v_1 be an eigenvector of the largest eigenvalue of Σ. (Assuming the largest eigenvalue of Σ is unique, strictly larger than other eigenvalues.)

3.1 Part (a)

Consider a sample of size n drawn from $\mathcal{N}_p(\mathbf{0}, \Sigma)$ with $n > p$. Suppose you conducted a principal component analysis on the sample data and obtained the first sample principal component $Y = \hat{e}_1^T X$. According to your reasoning, what should be the approximately value of the correlation (i.e., Pearson correlation coefficient) $\rho = \mathrm{Corr}(\hat{e}_1, v_1)$? And, what should be the approximate value of $|\rho|$?

The first principal component of a dataset *is*, by definition, (± 1 times) a unit-length eigenvector corresponding to the largest eigenvalue of the covariance matrix of that dataset. If we are working with a random sample of data, rather than an entire population, then the first principal component of that sample should be *approximately* (± 1 times) a unit-length eigenvector corresponding to the largest eigenvalue of the covariance matrix of the population from which the sample is drawn. In this case, the loadings of the first sample principal component are $\hat{e}_1$, and a unit eigenvector of the largest eigenvalue of Σ is v_1, meaning that we should have $\hat{e}_1 \approx \pm v_1$. Thus, we would expect $\rho = \text{Corr}(\hat{e}_1, v_1) \approx \pm 1$, and $|\rho| \approx 1$.

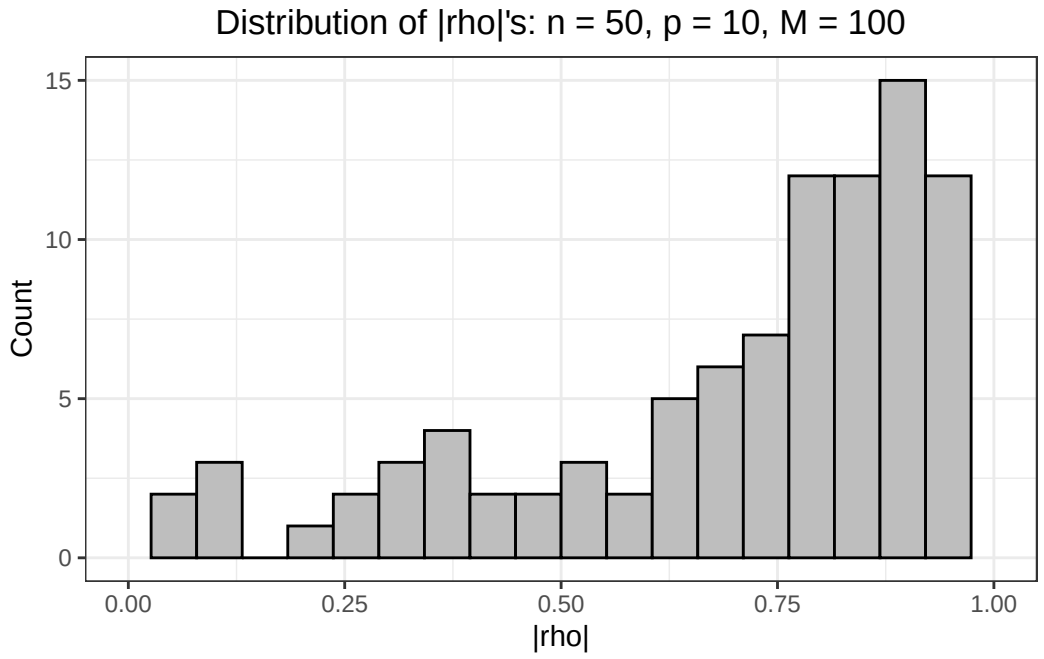
3.2 Part (b)

Construct a non-trivial covariance matrix $\Sigma \neq cI_p$ (and non-diagonal) for $p = 10$. Draw a sample of size $n = 50$ from $\mathcal{N}_p(\mathbf{0}, \Sigma)$ using your Σ. Compute ρ defined in Part (a). Repeat the above construction and sample draw for 100 times, obtain ρ values $\{\rho_1, \dots, \rho_{100}\}$. Plot a histogram for $|\rho_i|$ using the $100\rho_i$'s. Does the distribution of the values of ρ's agree with your hypothesis in Part (a)?

First, we write a function that will handle the simulations throughout this exercise:

```
rhosim = function(n, p, M){
  # Function to handle simulations in Parts (b)-(d)

  set.seed(41) # favorite number

  # Random covariance matrix Sigma
  C = replicate(p, rnorm(p))
  myCov = C %*% t(C)

  # Simulation for drawing M rhos
  rhos = rep(0, M)
  for(i in 1:M){
    data = mvrnorm(n = n, mu = rep(0,p), Sigma = myCov)
    samS = cov(data)
    rhos[i] = cor(eigen(samS)$vector[,1], eigen(myCov)$vector[,1])
  }

  return(rhos)
}
```

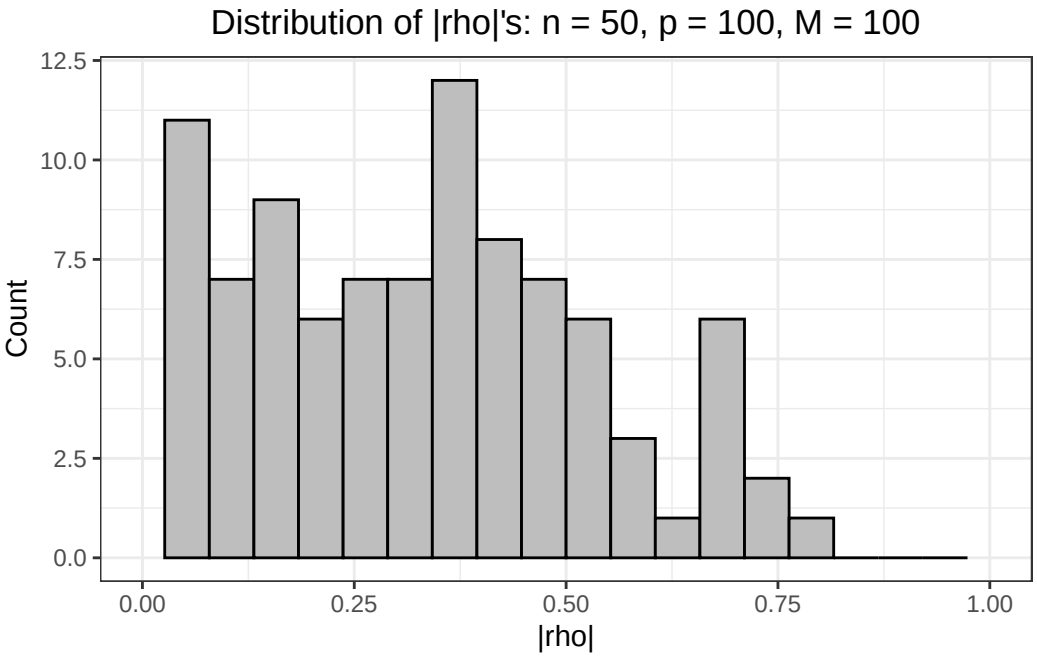
Below, we generate $\{\rho_1, \dots, \rho_{100}\}$ for $n = 50$, $p = 10$ and plot the $|\rho_i|$'s in a histogram. Observe that most of the mass of the distribution is around 1 (or at least between 0.75 and 1), which aligns with our prediction in Part (a).



3.3 Part (c)

Conduct a similar simulation as in Part (b), now with $p = 100$ (keep $n = 50$). Plot the histogram of the $|\rho_i|$'s. Is this histogram similar to that in Part (b)? Does the distribution of $|\rho|$'s agree with your hypothesis in Part (a) this time? What does your result mean in terms of the goodness of the first sample principal component?

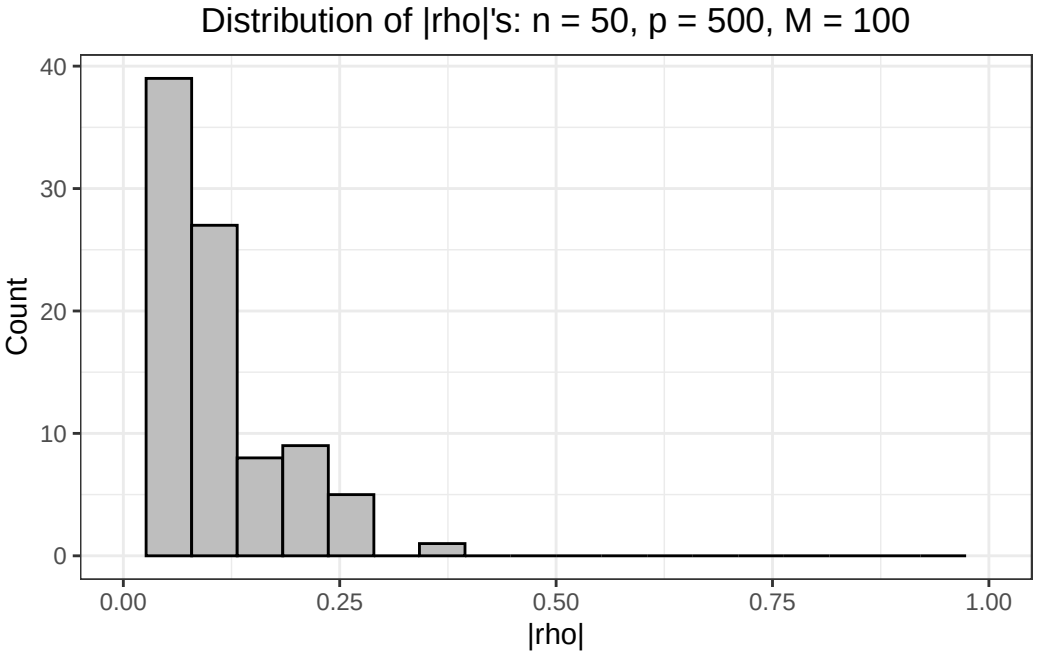
Below, we generate $\{\rho_1, \dots, \rho_{100}\}$ for $n = 50$, $p = 100$ and plot the $|\rho_i|$'s in a histogram. This time, notice that most of the mass of the distribution is spread across the interval $[0, 0.5]$, and there are *no* indices i for which $|\rho_i| > 0.8$, let alone $|\rho_i| \approx 1$. So, not only does this histogram deviate from the one we saw in Part (b), but it runs totally counter to our hypothesis in Part (a)! This is most likely because $p = 100$, the dimension of our random vector X, now exceeds $n = 50$, the size of our sample. Now that $\rho = \text{Corr}(\hat{e}_1, v_1) \not\approx \pm 1$, the first sample principal component does *not* closely coincide with an eigenvector corresponding to the largest eigenvalue of Σ. As such, the first sample principal component no longer does a very good job of capturing the direction of largest variance in the data.



3.4 Part (d)

Try larger p in Part (c) without breaking your laptop (keep $n - 50$). How far could you go? Plot the histogram(s) and comment.

Below, we manage to generate $\{\rho_1, \dots, \rho_{100}\}$ for $n = 50,\ p = 500$ and plot the $|\rho_i|$'s in a histogram. This time, the results are even worse than in Part (c): most of the mass of the distribution is around 0, and there are no indices i for which $|\rho_i| > 0.5$! Now that the dimension $p = 500$ of our random vector X is even larger than our sample size $n = 50$ than before, the first sample principal component of the data is even less codirectional with an eigenvector corresponding to the largest eigenvalue of Σ. Now, the first sample principal component is almost surely useless for visualizing the direction of largest variation in the data.



3.5 Part (e)

A researcher acquired data of gene expression levels of 4,000 genes for each of 100 patients. To find which genes are important, the researcher conducted PCA on the data and obtained a nice "L" shape in the scree plot. Subsequently the genes with larger loadings on the first couple of principal components were reported as important. Why is the researcher's conclusion problematic?

This situation bears a strong resemblance to our simulation in Part (d). The researcher is studying a random vector of gene expression levels with dimension $p = 4,000$, but they only have a sample of $n = 100$ patients, so that $n \ll p$. We saw in the simulations above that when there are far fewer observations in a sample than there are dimensions of a random vector, PCA yields sample principal components that are weakly correlated with the directions of greatest variation they are intended to identify. So, in this case, the first couple of sample principal components (corresponding to the largest eigenvalues of the sample covariance matrix) most likely do *not* correspond to the directions of greatest variation in gene expression. The researcher's claims that the genes with large loadings on these PCs are "imporant" reveal a failure to acknowledge this fact. In fact, in this case, the researcher's PCA gives little to no evidence that these genes are highly relevant to their work.

4 Exercise 5: Factor Analysis Using PC and ML Methods

Utilize the same data `Harman5.txt` as in Exercise 2 in this assignment.

4.1 Part (a)

Obtain the principal component solution to the factor model $X = \mu + LF + \varepsilon$ with the number of factors $m = 2$,

4.1.1 Part (i)

Using original data.

The loadings on the $m = 2$ common factors (i.e., the $m = 2$ columns of the L matrix) are as follows:

```
Q4PCcov = princomp(harman5, cor = FALSE)
rtev_PCcov = Q4PCcov$sdev

Q4L_PCcov = cbind(rtev_PCcov[1] * Q4PCcov$loading[,1],
                  rtev_PCcov[2] * Q4PCcov$loading[,2])
Q4L_PCcov
```

```
                  [,1]          [,2]
population    128.294381 3290.11053671
schooling       1.476329   -0.02958778
employment    164.228267 1155.73195839
professional   86.337287   45.09404891
housevalue   6095.608334 -101.02351948
```

4.1.2 Part (ii)

Using normalized (variance $= 1$) data.

The loadings on the $m = 2$ common factors (i.e., the $m = 2$ columns of the L matrix) are as follows:

```
Q4PCcor = princomp(harman5, cor = TRUE)
rtev_PCcor = Q4PCcor$sdev

Q4L_PCcor = cbind(rtev_PCcor[1] * Q4PCcor$loading[,1],
                  rtev_PCcor[2] * Q4PCcor$loading[,2])
Q4L_PCcor
```

```
                 [,1]       [,2]
population   0.5809571  0.8064212
schooling    0.7670373 -0.5447561
employment   0.6724314  0.7260453
professional 0.9323926 -0.1043054
housevalue   0.7911612 -0.5581795
```

4.2 Part (b)

Find the maximum likelihood estimates of L and Ψ for $m = 2$. What happens if you try $m = 3$?

4.2.1 Part (i): $m = 2$ Factors

For $m = 2$ factors, the maximum likelihood estimate of the L matrix is as follows:

```
Q4_ML2 = factanal(harman5, 2, rotation = "none")
Q4L_ML2 = Q4_ML2$loadings[1:5,1:2]
Q4L_ML2
```

```
                Factor1    Factor2
population   -0.02594191 0.99716856
schooling     0.89737779 0.03766743
employment    0.09128508 0.97751671
professional  0.77697154 0.45959388
housevalue    0.96121991 0.04611779
```

and the maximum likelihood estimate of the Ψ matrix is as follows:

```
Q4Psi_ML2 = Q4_ML2$uniq %>% round(3) %>% diag()
Q4Psi_ML2
```

```
      [,1]  [,2]  [,3]  [,4]  [,5]
[1,] 0.005 0.000 0.000 0.000 0.000
[2,] 0.000 0.193 0.000 0.000 0.000
[3,] 0.000 0.000 0.036 0.000 0.000
[4,] 0.000 0.000 0.000 0.185 0.000
[5,] 0.000 0.000 0.000 0.000 0.074
```

4.2.2 Part (ii): $m = 3$ Factors

When we attempt to estimate the factor model using the maximum likelihood procedure for $m = 3$ factors, we receive the following error message.

```
Q4_ML3 = factanal(harman5, 3, rotation = "none")
```

```
Error in factanal(harman5, 3, rotation = "none"):
3 factors are too many for 5 variables
```

Recall that the degrees of freedom associated with the maximum likelihood procedure is $df = \frac{(p-m)^2-p-m}{2}$. Here, if we tried to use three factors, $p = 5$ and $m = 3$, so the degrees of freedom would be $df = \frac{(5-3)^2-5-3}{2} = -2 < 0$. As such, the model cannot be estimated using the ML procedure for $m = 3$ factors.

4.3 Part (c)

Compare the factors obtained by principal component methods and by maximum likelihood, especially on their estimates of the covariance or correlation matrix. Compare the entries in the residual matrices. Which method is better in estimating the correlation matrix?

Using the principal component method with $m = 2$ factors on the correlation matrix (i.e., on data with standardized variances), the residual matrix is as follows:

```
resid_PC = cor(harman5) - Q4L_PCcor%*%t(Q4L_PCcor) -
           diag(rep(1,5)) - diag(Q4L_PCcor%*%t(Q4L_PCcor))
resid_PC %>% round(3)
```

```
             population schooling employment professional housevalue
population        -1.976    -0.984     -0.992       -1.007     -0.975
schooling         -0.882    -1.770     -0.851       -0.966     -0.933
employment        -0.983    -0.945     -1.959       -1.016     -0.984
professional      -0.899    -0.961     -0.917       -1.760     -0.898
housevalue        -0.925    -0.985     -0.942       -0.956     -1.875
```

Under the principal component approach, the sum of absolute residuals and the sum of squared residuals are as follows:

```
c(sum(abs(resid_PC)),
  sum(resid_PC^2))
```

```
[1] 28.33994 35.57468
```

Meanwhile, using the maximum likelihood method with $m = 2$ factors (which is necessarily on the correlation matrix), the residual matrix is:

```
resid_ML = cor(harman5) - Q4L_ML2%*%t(Q4L_ML2) - Q4Psi_ML2
resid_ML %>% round(3)
```

```
             population schooling employment professional housevalue
population         0.000    -0.005      0.000        0.001      0.001
schooling         -0.005     0.000      0.036       -0.023     -0.001
employment         0.000     0.036      0.000       -0.005     -0.011
professional       0.001    -0.023     -0.005        0.000      0.010
housevalue         0.001    -0.001     -0.011        0.010      0.000
```

Under the maximum likelihood approach, the sum of absolute residuals and the sum of squared residuals are now:

```
c(sum(abs(resid_ML)),
  sum(resid_ML^2))
```

```
[1] 0.185819784 0.004129502
```

Notice that the sum of absolute residuals and the sum of squared residuals are each significantly smaller using the ML method compared to the PC method. As such, the ML method is better at estimating the correlation matrix for these data.

6. Factor model w/ $m=1$ factor. Covariance matrix:

$$\Sigma = \begin{bmatrix} 5 & 2 & 3 \\ 2 & 6 & 6 \\ 3 & 6 & 10 \end{bmatrix} = \begin{bmatrix} \ell_1 \\ \ell_2 \\ \ell_3 \end{bmatrix} \begin{bmatrix} \ell_1 & \ell_2 & \ell_3 \end{bmatrix} + \begin{bmatrix} \Psi_1 & 0 & 0 \\ 0 & \Psi_2 & 0 \\ 0 & 0 & \Psi_3 \end{bmatrix}$$

6a. How many variables are there in this study? Write the population factor model in vector-matrix form & indicate the dimensions of terms.

- Since Σ is a (3×3) matrix, there are three variables in this study.
- The population factor model is therefore

$$\underset{(3\times 1)}{X} = \underset{(3\times 1)}{\mu} + \underset{(3\times 1)(1\times 1)}{LF} + \underset{(3\times 1)}{\varepsilon}$$

6b. Set up & solve a sys. of eqns for ℓ_i, Ψ_i, $i=1,2,3$.

$$\cdot\ \Sigma = \begin{bmatrix} 5 & 2 & 3 \\ 2 & 6 & 6 \\ 3 & 6 & 10 \end{bmatrix} = \begin{bmatrix} \ell_1 \\ \ell_2 \\ \ell_3 \end{bmatrix} \begin{bmatrix} \ell_1 & \ell_2 & \ell_3 \end{bmatrix} + \begin{bmatrix} \Psi_1 & 0 & 0 \\ 0 & \Psi_2 & 0 \\ 0 & 0 & \Psi_3 \end{bmatrix}$$

$$= \begin{bmatrix} \ell_1^2 & \ell_1\ell_2 & \ell_1\ell_3 \\ \ell_1\ell_2 & \ell_2^2 & \ell_2\ell_3 \\ \ell_1\ell_3 & \ell_2\ell_3 & \ell_3^2 \end{bmatrix} + \begin{bmatrix} \Psi_1 & 0 & 0 \\ 0 & \Psi_2 & 0 \\ 0 & 0 & \Psi_3 \end{bmatrix}$$

$$= \begin{bmatrix} \ell_1^2+\Psi_1 & \ell_1\ell_2 & \ell_1\ell_3 \\ \ell_1\ell_2 & \ell_2^2+\Psi_2 & \ell_2\ell_3 \\ \ell_1\ell_3 & \ell_2\ell_3 & \ell_3^2+\Psi_3 \end{bmatrix}$$

- So,

$$\begin{cases} (1)\ \ell_1^2 + \Psi_1 = 5 \\ (2)\ \ell_2^2 + \Psi_2 = 6 \\ (3)\ \ell_3^2 + \Psi_3 = 10 \\ (4)\ \ell_1\ell_2 = 2 \\ (5)\ \ell_1\ell_3 = 3 \\ (6)\ \ell_2\ell_3 = 6 \end{cases}$$

- ④ $\Rightarrow \ell_1 = \frac{2}{\ell_2}$
- ⑤ $\Rightarrow \ell_1 = \frac{3}{\ell_3}$
- ④, ⑤ $\Rightarrow \frac{2}{\ell_2} = \frac{3}{\ell_3} \Rightarrow 3\ell_2 = 2\ell_3 \Rightarrow \ell_2 = \frac{2}{3}\ell_3$
- ④, ⑤, ⑥ $\Rightarrow \ell_2\ell_3 = 6 \Rightarrow (\frac{2}{3}\ell_3)\ell_3 = 6 \Rightarrow \ell_3^2 = 9 \Rightarrow \boxed{\ell_3 = 3}$
- ⑤ $\Rightarrow \ell_1\ell_3 = 3 \Rightarrow 3\ell_1 = 3 \Rightarrow \boxed{\ell_1 = 1}$
- ④ $\Rightarrow \ell_1\ell_2 = 2 \Rightarrow \boxed{\ell_2 = 2}$
- ① $\Rightarrow \ell_1^2 + \Psi_1 = 5 \Rightarrow 1 + \Psi_1 = 5 \Rightarrow \boxed{\Psi_1 = 4}$
- ② $\Rightarrow \ell_2^2 + \Psi_2 = 6 \Rightarrow 4 + \Psi_2 = 6 \Rightarrow \boxed{\Psi_2 = 2}$
- ③ $\Rightarrow \ell_3^2 + \Psi_3 = 10 \Rightarrow 9 + \Psi_3 = 10 \Rightarrow \boxed{\Psi_3 = 1}$

- So, and $L = (\ell_1, \ell_2, \ell_3) = (1, 2, 3)$, $\Psi = \text{diag}(\Psi_1, \Psi_2, \Psi_3) = \text{diag}(4, 2, 1)$

6c. For each variable X_i, calc. the % of $\text{Var}(X_i)$ explained by the common factor.

- $h_1^2 = \ell_1^2 = 1$; $\text{Var}(X_1) = 5$
  So, $\frac{h_1^2}{\text{Var}(X_1)} = \frac{1}{5}$ or 20% of the variance of X_1 is explained by the common factor.
- $h_2^2 = \ell_2^2 = 4$; $\text{Var}(X_2) = 6$
  So, $\frac{h_2^2}{\text{Var}(X_2)} = \frac{4}{6} = \frac{2}{3}$ or ~66.7% of the variance of X_2 is explained by the common factor.
- $h_3^2 = \ell_3^2 = 9$; $\text{Var}(X_3) = 10$
  So, $\frac{h_3^2}{\text{Var}(X_3)} = \frac{9}{10}$ or 90% of the variance of X_3 is explained by the common factor.

6d. Now $\Sigma = \begin{bmatrix} 5 & 2 & 3 \\ 2 & 6 & 6 \\ 3 & 6 & 8 \end{bmatrix}$.

Repeat Parts (a)–(c). Comment on results for Part (c). Is the factor model reasonable?

- Note that just Σ_{33} has changed.
- Σ is still (3×3), so these are still 3 variables.
- The population factor model is still

$$\underset{(3\times1)}{X} = \underset{(3\times1)}{\mu} + \underset{(3\times1)(1\times1)}{LF} + \underset{(3\times1)}{\varepsilon}$$

- Since only Σ_{33} has changed, the system is

$$\begin{aligned} \ell_1^2 + \Psi_1 &= 5 \\ \ell_2^2 + \Psi_2 &= 6 \\ \ell_3^2 + \Psi_3 &= 8 \\ \ell_1\ell_2 &= 2 \\ \ell_1\ell_3 &= 3 \\ \ell_2\ell_3 &= 6 \end{aligned}$$

- So still $\ell_1 = 1$, $\ell_2 = 2$, $\ell_3 = 3$, $\Psi_1 = 4$, $\Psi_2 = 2$. Now, $\ell_3^2 + \Psi_3 = 8 \Rightarrow 9 + \Psi_3 = 8 \Rightarrow \Psi_3 = -1$.
- So, $L = (\ell_1, \ell_2, \ell_3) = (1, 2, 3)$, and $\Psi = \text{diag}(\Psi_1, \Psi_2, \Psi_3) = (4, 2, -1)$.

- $h_1^2 = \ell_1^2 = 1$; $\text{Var}(X_1) = 5$. So, still $\frac{h_1^2}{\text{Var}(X_1)} = \frac{1}{5}$ or 20% of the variance of X_1 is explained by the common factor.

- $h_2^2 = \ell_2^2 = 4$; $\text{Var}(X_2) = 6$. So, still $\frac{h_2^2}{\text{Var}(X_2)} = \frac{4}{6} = \frac{2}{3}$ or ~66.7% of the variance of X_2 is explained by the common factor.

- But now $h_3^2 = l_3^2 = 9$; $Var(X_3 = 8)$.

So, $\frac{h_3^2}{Var(X_3)} = \frac{9}{8} > 1$ or 112.5% > 100% of the variance of X_3 is purportedly explained by the common factor. This is obviously nonsensical: it makes no sense for the common factor F to explain more than 100% of the variance of X_3!