STAT 524
HW6

Satoshi Ido
34788706
11/07/2023

8.2. Convert the covariance matrix in Exercise 8.1 to a correlation matrix $\boldsymbol{\rho}$.

(a) Determine the principal components Y_1 and Y_2 from $\boldsymbol{\rho}$ and compute the proportion of total population variance explained by Y_1.

(b) Compare the components calculated in Part a with those obtained in Exercise 8.1. Are they the same? Should they be?

(c) Compute the correlations ρ_{Y_1,Z_1}, ρ_{Y_1,Z_2}, and ρ_{Y_2,Z_1}.

(a) $\Sigma = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$

$$\rho_{11} = \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} = \frac{5}{\sqrt{5\cdot 5}} = 1$$

$$\rho_{22} = \frac{\sigma_{22}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{22}}} = \frac{2}{\sqrt{2\cdot 2}} = 1$$

$$\rho_{12} = \rho_{21} = \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} = \frac{2}{\sqrt{10}} = 0.6325$$

$$\therefore \rho = \begin{bmatrix} 1 & 0.6325 \\ 0.6325 & 1 \end{bmatrix}$$

Let λ as eigen value and e as eigen vector of ρ

$$|\rho - \lambda I| = \begin{vmatrix} 1-\lambda & 0.6325 \\ 0.6325 & 1-\lambda \end{vmatrix} = 0$$

$$(1-\lambda)^2 - 0.6325^2 = 0$$

$$\lambda^2 - 2\lambda + 0.6 = 0$$

$$\lambda = 1 \pm 0.6325$$

$$\Rightarrow \lambda_1 = 1.6325, \quad \lambda_2 = 0.368$$

For $\lambda_1 = 1.6325$,

$$\begin{bmatrix} -0.6325 & 0.6325 \\ 0.6325 & -0.6325 \end{bmatrix} \begin{bmatrix} e_{11} \\ e_{21} \end{bmatrix} = 0$$

$$-0.6325\, e_{11} + 0.6325\, e_{21} = 0$$

This simplifies to $e_{11} = e_{21}$

Normalize these to $e_1 = \frac{1}{\sqrt{V_{11}^2 + V_{21}^2}} \begin{bmatrix} e_{11} \\ e_{21} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

For $\lambda_2 = 0.368$

$$\begin{bmatrix} 0.6325 & 0.6325 \\ 0.6325 & 0.6325 \end{bmatrix} \begin{bmatrix} e_{12} \\ e_{22} \end{bmatrix} = 0 \quad \Rightarrow \quad e_{12} = -e_{22}$$

Normalize these to $e_2 = \frac{1}{\sqrt{V_{12}^2 + V_{22}^2}} \begin{bmatrix} e_{12} \\ e_{22} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

∴ The principle components from ρ become

$$y_1 = e_1^T Z = 0.707\, Z_1 + 0.707\, Z_2$$

$$y_2 = e_2^T Z = 0.707\, Z_1 - 0.707\, Z_2 \quad //$$

∴ The proportion of total population variance by y_1

$$= \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{1.6328}{1.6328 + 0.368} = 0.8162 \quad \therefore \underline{81.62\%}$$

(d) No, they are not the same. The reason lies in the standardization of variables. The principal components from Σ are affected by the scale of the original values. Meanwhile, the correlation matrix is standardized version of the covariance matrix where each variable has a variance of 1. This way, all variables contribute equally to the components.

8.8. Use the results in Example 8.5.

(a) Compute the correlations r_{y_i, z_k} for $i = 1, 2$ and $k = 1, 2, \ldots, 5$. Do these correlations reinforce the interpretations given to the first two components? Explain.

(b) Test the hypothesis

$$H_0: \boldsymbol{\rho} = \boldsymbol{\rho}_0 = \begin{bmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{bmatrix}$$

versus

$$H_1: \boldsymbol{\rho} \neq \boldsymbol{\rho}_0$$

at the 5% level of significance. List any assumptions required in carrying out this test.

(a) Given $\hat{\lambda}^T = \begin{pmatrix} 2.437 \\ 1.407 \end{pmatrix}$ and $\hat{e}^T = \begin{pmatrix} 0.469, & 0.532, & 0.465, & 0.387, & 0.361 \\ -0.368, & -0.236, & -0.315, & 0.585, & 0.606 \end{pmatrix}$

$r_{\hat{y}_1, z_1} = 0.469 \cdot \sqrt{2.437} = 0.732$

$r_{\hat{y}_1, z_2} = 0.831$

$r_{\hat{y}_1, z_3} = 0.726$

$r_{\hat{y}_1, z_4} = 0.604$

$r_{\hat{y}_1, z_5} = 0.361$

$r_{\hat{y}_2, z_1} = -0.368 \cdot \sqrt{1.407} = -0.437$

$r_{\hat{y}_2, z_2} = -0.280$

$r_{\hat{y}_2, z_3} = -0.374$

$r_{\hat{y}_2, z_4} = 0.694$

$r_{\hat{y}_2, z_5} = 0.719$

The sample correlations of two principal components underpin the idea that the first component is roughly equally weighted sum of the five stocks, which we can call a general stock-market component. Meanwhile, the second component seems to represent a contrast between the banking stocks and the oil stocks, which we can call an industry component.

1e) The average of the off-diagonal elements in the k th cols become

$$\bar{r}_k = \frac{1}{p-1} \sum_{\substack{i=1 \\ i \neq k}}^{p} r_{ik} \qquad k = 1, 2, \cdots, p \quad , \quad p = 5$$

$$\bar{r}_1 = \frac{1}{4} (0.632 + 0.511 + 0.115 + 0.155) = 0.353$$

$$\bar{r}_2 = 0.435$$

$$\bar{r}_3 = 0.354$$

$$\bar{r}_4 = 0.326$$

$$\bar{r}_5 = 0.299$$

The overall average of the off-diagonal elements become

$$\bar{r} = \frac{2}{5(5-1)} \sum_{i<k}^{5} \sum^{5} r_{ik}$$

$$= \frac{1}{10} (0.632 + 0.511 + 0.115 + 0.155 + 0.574 + 0.322 + 0.213 + 0.183 + 0.146 + 0.683)$$

$$= 0.353$$

$$\hat{\gamma} = \frac{(5-1)^2 [1 - (1-0.353)^2]}{5 - (5-2)(1-0.353)^2} = 2.484$$

$$\therefore T = \frac{(103-1)}{(1-0.353)^2}\left[\sum\sum_{i<k}(r_{ik}-0.353)^2 - 2.464\sum_{k=1}^{5}(\bar{r}_k - 0.353)^2\right]$$

$$= \frac{102}{0.4186}\left(0.4472 - 2.484 \cdot 0.1037\right)$$

$$= 102.69$$

$$\chi^2_{(5+1)(5-2)/2}\ (\alpha = 0.05) = 16.92$$

$T > \chi^2_q(0.05) = 16.92$ Hence, we will reject H_0 at $\alpha = 0.05$ level of significant.

We have to assume a multivariate normal population, meaning a large sample test that all variables are independent are based of this test statistics. //

8.12. Consider the air-pollution data listed in Table 1.5. Your job is to summarize these data in fewer than $p = 7$ dimensions if possible. Conduct a principal component analysis of the data using both the covariance matrix **S** and the correlation matrix **R**. What have you learned? Does it make any difference which matrix is chosen for analysis? Can the data be summarized in three or fewer dimensions? Can you interpret the principal components?

Covariance matrix: S

```
> S
                      Wind Solar_radiation         CO         NO       NO_2
Wind             2.5000000      -2.7804878 -0.3780488 -0.4634146 -0.5853659
Solar_radiation -2.7804878     300.5156794  3.9094077 -1.3867596  6.7630662
CO              -0.3780488       3.9094077  1.5220674  0.6736353  2.3147503
NO              -0.4634146      -1.3867596  0.6736353  1.1823461  1.0882695
NO_2            -0.5853659       6.7630662  2.3147503  1.0882695 11.3635308
O3              -2.2317073      30.7909408  2.8217189 -0.8106852  3.1265970
HC               0.1707317       0.6236934  0.1416957  0.1765389  1.0441347
                        O3        HC
Wind            -2.2317073 0.1707317
Solar_radiation 30.7909408 0.6236934
CO               2.8217189 0.1416957
NO              -0.8106852 0.1765389
NO_2             3.1265970 1.0441347
O3              30.9785134 0.5946574
HC               0.5946574 0.4785134
```

Correlation matrix: R

```
> R
                      Wind Solar_radiation         CO          NO       NO_2
Wind             1.0000000     -0.10144191 -0.1938032 -0.26954261 -0.1098249
Solar_radiation -0.1014419      1.00000000  0.1827934 -0.07356907  0.1157320
CO              -0.1938032      0.18279338  1.0000000  0.50215246  0.5565838
NO              -0.2695426     -0.07356907  0.5021525  1.00000000  0.2968981
NO_2            -0.1098249      0.11573199  0.5565838  0.29689814  1.0000000
O3              -0.2535928      0.31912373  0.4109288 -0.13395214  0.1666422
HC               0.1560979      0.05201044  0.1660323  0.23470432  0.4477678
                        O3         HC
Wind            -0.2535928 0.15609793
Solar_radiation  0.3191237 0.05201044
CO               0.4109288 0.16603235
NO              -0.1339521 0.23470432
NO_2             0.1666422 0.44776780
O3               1.0000000 0.15445056
HC               0.1544506 1.00000000
```

Eigen system with S

```
> eigen_S
eigen() decomposition
$values
[1] 304.2578640  28.2761046  11.4644830   2.5243296   1.2795247   0.5287288
[7]   0.2096157

$vectors
             [,1]        [,2]        [,3]          [,4]          [,5]
[1,]  0.010039244  0.07622439  0.03087761  0.9203045748 -0.3423859285
[2,] -0.993199405  0.11615518  0.00659069 -0.0002118679 -0.0022391022
[3,] -0.014062314 -0.09956775 -0.18282641 -0.1382922410 -0.6500776063
[4,]  0.004710175  0.01320423 -0.13021553 -0.3277842624 -0.6431560485
[5,] -0.024255644 -0.15038113 -0.95526318  0.1023719020  0.2065840405
[6,] -0.112429558 -0.97335904  0.16981025  0.0632480276  0.0002935726
[7,] -0.002340785 -0.02382046 -0.08519558  0.1095073458 -0.0619613872
             [,6]         [,7]
[1,]  0.011779079 -0.169729925
[2,]  0.003353218 -0.001781987
[3,] -0.563893916  0.443577538
[4,]  0.497513370 -0.462855916
[5,] -0.009009299 -0.105029951
[6,]  0.051067254 -0.066992404
[7,]  0.657012233  0.738019426
```

Eigen system with R

```
> eigen_R
eigen() decomposition
$values
[1] 2.3367826 1.3860007 1.2040659 0.7270865 0.6534765 0.5366888 0.1558989

$vectors
           [,1]         [,2]       [,3]         [,4]        [,5]         [,6]
[1,]  0.2368211  0.278445138  0.6434744  0.172719491  0.56053441 -0.223579220
[2,] -0.2055665 -0.526613869  0.2244690  0.778136601 -0.15613432 -0.005700851
[3,] -0.5510839 -0.006819502 -0.1136089  0.005301798  0.57342221 -0.109538907
[4,] -0.3776151  0.434674253 -0.4070978  0.290503052 -0.05669070 -0.450234781
[5,] -0.4980161  0.199767367  0.1965567 -0.042428178  0.05021430  0.744968707
[6,] -0.3245506 -0.566973655  0.1598465 -0.507915905  0.08024349 -0.330583071
[7,] -0.3194032  0.307882771  0.5410484 -0.143082348 -0.56607057 -0.266469812
            [,7]
[1,]  0.24146701
[2,]  0.01126548
[3,] -0.58524622
[4,]  0.46088973
[5,]  0.33784371
[6,]  0.41707805
[7,] -0.31391372
```

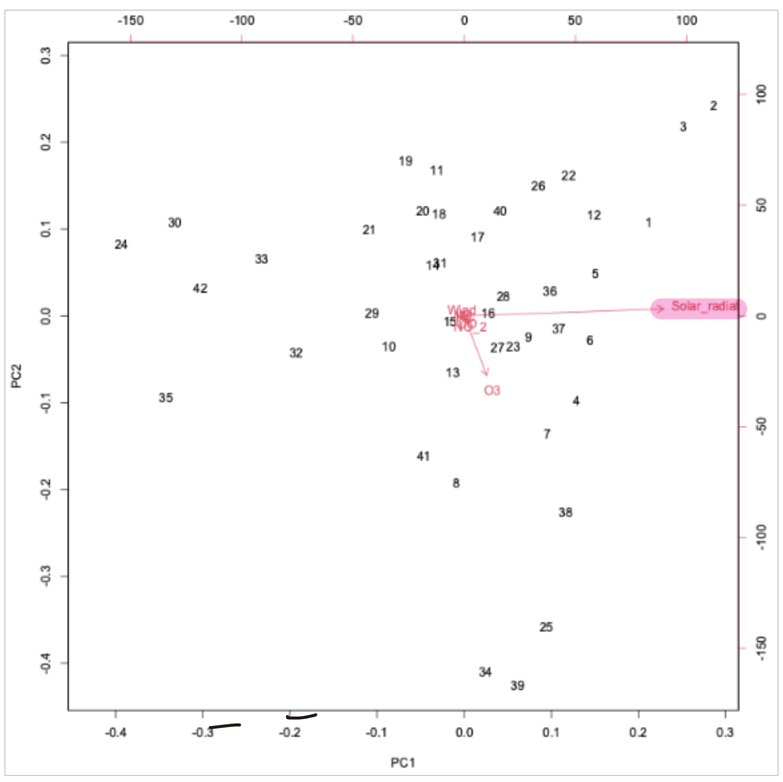
PCA with S

```
> summary_S
$std_dev
[1] 17.4429890  5.3175281  3.3859242  1.5888139  1.1311608  0.7271374  0.4578381

$prop_var
[1] 0.8729480010 0.0811271357 0.0328928146 0.0072425686 0.0036710916
[6] 0.0015169789 0.0006014096

$cum_prop_var
[1] 0.8729480 0.9540751 0.9869680 0.9942105 0.9978816 0.9993986 1.0000000

$loadings
            [,1]        [,2]        [,3]          [,4]          [,5]
[1,]   0.17511443  0.40532535  0.10454926  1.4621926872 -0.3872935261
[2,] -17.32436626  0.61765845  0.02231558 -0.0003366187 -0.0025327846
[3,]  -0.24528879 -0.52945432 -0.61903636 -0.2197206328 -0.7353422773
[4,]   0.08215953  0.07021387 -0.44089992 -0.5207881875 -0.7275128828
[5,]  -0.42309094 -0.79965586 -3.23444878  0.1626498993  0.2336797596
[6,]  -1.96110754 -5.17586403  0.57496463  0.1004893445  0.0003320778
[7,]  -0.04083029 -0.12666596 -0.28846577  0.1739867915 -0.0700882897
            [,6]         [,7]
[1,]  0.008565009 -0.0777088233
[2,]  0.002438250 -0.0008158614
[3,] -0.410028357  0.2030866895
[4,]  0.361760580 -0.2119130654
[5,] -0.006550999 -0.0480867113
[6,]  0.037132910 -0.0306716740
[7,]  0.477738168  0.3378933992
```

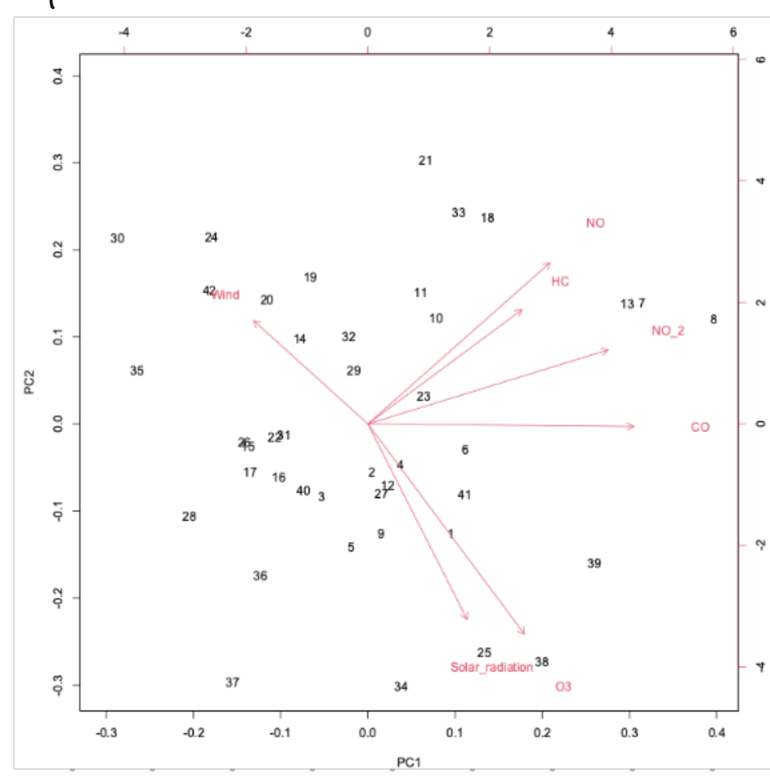
PCA with R

```
> summary_R
$std_dev
[1] 1.5286539 1.1772853 1.0972994 0.8526937 0.8083790 0.7325905 0.3948404

$prop_var
[1] 0.33382609 0.19800010 0.17200942 0.10386950 0.09335379 0.07666983 0.02227128

$cum_prop_var
[1] 0.3338261 0.5318262 0.7038356 0.8077051 0.9010589 0.9777287 1.0000000

$loadings
           [,1]         [,2]       [,3]         [,4]        [,5]         [,6]
[1,]  0.3620175  0.327809366  0.7060840  0.147276816  0.45312422 -0.163792005
[2,] -0.3142401 -0.619974764  0.2463097  0.663512149 -0.12621570 -0.004176389
[3,] -0.8424165 -0.008028499 -0.1246630  0.004520809  0.46354245 -0.080247159
[4,] -0.5772427  0.511735606 -0.4467081  0.247710112 -0.04582757 -0.329837708
[5,] -0.7612943  0.235183184  0.2156816 -0.036178238  0.04059219  0.545756973
[6,] -0.4961256 -0.667489747  0.1753995 -0.433096674  0.06486715 -0.242182006
[7,] -0.4882569  0.362465859  0.5936921 -0.122005412 -0.45759954 -0.195213244
             [,7]
[1,]  0.095340930
[2,]  0.004448066
[3,] -0.231078856
[4,]  0.181977886
[5,]  0.133394347
[6,]  0.164679265
[7,] -0.123945821
```

Biplot with S



Biplot with R



Given outputs from R, we know that we can make difference depending on types of matrices. For PCA with Covariance matrix, the first PC shows that 'Solar_radiation' contributes the most to the PC (-17.324), and the first PC accounts for 87%. The cumulative proportion of variance shows the first two PCs have a cumulative variance of over 95% in a dataset. Hence, the data can be summarized in fewer than three dimensions by PCA with covariance matrix.

For PCA with correlation matrix (= R), the first four PCs explain a variance of 80% in dataset. After the standardization of dataset, the size of variances are evenly distributed among PCs, with the first PC explaining only a third of the variance. In the case of PCA with the `R`, it is suggested not to be summarized in three or fewer dimensions to keep the original information.

9.2. Use the information in Exercise 9.1.
(a) Calculate communalities $h_i^2, i = 1, 2, 3$, and interpret these quantities.
(b) Calculate $\text{Corr}(Z_i, F_1)$ for $i = 1, 2, 3$. Which variable might carry the greatest weight in "naming" the common factor? Why?

(a) Given $m = 1$ factor model and

the loading matrix is $L = \begin{pmatrix} l_{11} \\ l_{21} \\ l_{31} \end{pmatrix} = \begin{pmatrix} 0.9 \\ 0.7 \\ 0.5 \end{pmatrix}$,

$h_1^2 = l_{11}^2 = 0.81$

$h_2^2 = l_{21}^2 = 0.49$

$h_3^2 = l_{31}^2 = 0.25$ //

$h_i^2, i = 1, 2, 3$ are coefficients that express the relationships between the observed variables (=3 obs) and the factors (= 1).

That is to say, h_1^2 displays the sum of squares of the loadings of the 1st variables. We can say similar things to h_2^2 and h_3^2 as well.

(d) $\text{Corr}(Z_i, F_1) = \text{Cov}(Z_i, F_1) = l_{i1}$ $\quad i=1,2,3$

$\therefore \text{Corr}(Z_1, F_1) = 0.9$, $\text{Corr}(Z_2, F_1) = 0.7$, $\text{Corr}(Z_3, F_1) = 0.5$

$\Rightarrow$ The first variable Z_1 has the largest correlation with the common factor F_1, meaning it might carry the greatest weight in naming the factor.

9.3. The eigenvalues and eigenvectors of the correlation matrix $\boldsymbol{\rho}$ in Exercise 9.1 are

$$\begin{aligned} \lambda_1 &= 1.96, & \mathbf{e}_1' &= [.625, .593, .507] \\ \lambda_2 &= .68, & \mathbf{e}_2' &= [-.219, -.491, .843] \\ \lambda_3 &= .36, & \mathbf{e}_3' &= [.749, -.638, -.177] \end{aligned}$$

(a) Assuming an $m = 1$ factor model, calculate the loading matrix $\mathbf{L}$ and matrix of specific variances $\boldsymbol{\Psi}$ using the principal component solution method. Compare the results with those in Exercise 9.1.

(b) What proportion of the total population variance is explained by the first common factor?

(a) $$L = \begin{pmatrix} l_{11} \\ l_{21} \\ l_{31} \end{pmatrix} = \sqrt{1.96} \cdot \begin{pmatrix} 0.625 \\ 0.593 \\ 0.507 \end{pmatrix} = \begin{pmatrix} 0.875 \\ 0.830 \\ 0.710 \end{pmatrix} //$$

Since $\rho = LL^T + \Psi \iff \Psi = \rho - LL^T$,

$$\Psi = \begin{bmatrix} 1.0 & 0.63 & 0.45 \\ 0.63 & 1.0 & 0.35 \\ 0.45 & 0.35 & 1.0 \end{bmatrix} - \begin{bmatrix} 0.875 & 0.830 & 0.710 \end{bmatrix} \begin{bmatrix} 0.875 \\ 0.83 \\ 0.710 \end{bmatrix}$$

$$\cong \begin{bmatrix} 0.234 \\ 0.311 \\ 0.5 \end{bmatrix} //$$

$\therefore$ It is slightly different from the results with those in Exercise 9.1.

(e) $\frac{\hat{\lambda}_1}{p} = \frac{1.96}{3} = 0.653$ //

9.10. The correlation matrix for chicken-bone measurements (see Example 9.14) is

$$\begin{bmatrix} 1.000 & & & & & \\ .505 & 1.000 & & & & \\ .569 & .422 & 1.000 & & & \\ .602 & .467 & .926 & 1.000 & & \\ .621 & .482 & .877 & .874 & 1.000 & \\ .603 & .450 & .878 & .894 & .937 & 1.000 \end{bmatrix}$$

The following estimated factor loadings were extracted by the maximum likelihood procedure:

Variable	Estimated factor loadings F_1	F_2	Varimax rotated estimated factor loadings F_1^*	F_2^*
1. Skull length	.602	.200	.484	.411
2. Skull breadth	.467	.154	.375	.319
3. Femur length	.926	.143	.603	.717
4. Tibia length	1.000	.000	.519	.855
5. Humerus length	.874	.476	.861	.499
6. Ulna length	.894	.327	.744	.594

Using the *unrotated* estimated factor loadings, obtain the maximum likelihood estimates of the following.

(a) The specific variances.

(b) The communalities.

(c) The proportion of variance explained by each factor.

(d) The residual matrix $\mathbf{R} - \hat{\mathbf{L}}_z\hat{\mathbf{L}}_z' - \hat{\mathbf{\Psi}}_z$.

(a) Given the R outputs below, we can see the specific variances $\psi^T = [0.596\ 0.758\ 0.122\ 0\ 0.0095\ 0.094]$ //

```
         [,1]     [,2]     [,3] [,4]     [,5]     [,6]
[1,] 0.597596 0.000000 0.000000    0 0.000000 0.000000
[2,] 0.000000 0.758195 0.000000    0 0.000000 0.000000
[3,] 0.000000 0.000000 0.122075    0 0.000000 0.000000
[4,] 0.000000 0.000000 0.000000    0 0.000000 0.000000
[5,] 0.000000 0.000000 0.000000    0 0.009548 0.000000
[6,] 0.000000 0.000000 0.000000    0 0.000000 0.093835
```

☆ $\psi = \Sigma - LL^T$

(b) The communalities $\hat{h}_j = \sum_{j=1}^{m} \hat{l}_{ij}$

$\hat{h}_1 = 0.402$

$\hat{h}_2 = 0.212$

$\hat{h}_3 = 0.878$

$\hat{h}_4 = 1$

$\hat{h}_5 = 0.991$

$\hat{h}_6 = 0.906$ //

(c) The proportion of variance explained by each factor $= \frac{1}{p}\sum_{j=1}^{p} l_{ij}^2$

For factor 1, 0.667

For factor 2, 0.067 //

(d) Given R output, the residual matrix, $R - \hat{L}_z \hat{L}_z^T - \hat{\Psi}_z$ be

=

```
> R - L %*% t(L) - P
          [,1]      [,2]      [,3] [,4]      [,5]      [,6]
[1,]  0.000000  0.193066 -0.017052    0 -0.000348 -0.000588
[2,]  0.193066  0.000000 -0.032464    0  0.000538 -0.017856
[3,] -0.017052 -0.032464  0.000000    0 -0.000392  0.003395
[4,]  0.000000  0.000000  0.000000    0  0.000000  0.000000
[5,] -0.000348  0.000538 -0.000392    0  0.000000 -0.000008
[6,] -0.000588 -0.017856  0.003395    0 -0.000008  0.000000
```