Multivariate Analysis Homework 3

A49109720 Yi-Chen Zhang

April 13, 2018

8.4. Find the principal components and the proportion of the total population variance explained by each when the covariance matrix is

\[ \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1^2 \rho & 0 \\ \sigma_1^2 \rho & \sigma_2^2 & \sigma_2^2 \rho \\ 0 & \sigma_2^2 \rho & \sigma_2^2 \end{pmatrix}, \quad -\frac{1}{\sqrt{2}} < \rho < \frac{1}{\sqrt{2}} \]

Sol. To find the eigenvalues, we let \( \det(\Sigma - \lambda I) = 0 \), i.e.,

\[
\begin{vmatrix} \sigma_1^2 - \lambda & \sigma_1^2 \rho & 0 \\ \sigma_1^2 \rho & \sigma_2^2 - \lambda & \sigma_2^2 \rho \\ 0 & \sigma_2^2 \rho & \sigma_2^2 - \lambda \end{vmatrix} = 0.
\]

By solving the system, we obtain the characteristic polynomial in terms of \( \lambda \) as:

\[(\sigma_1^2 - \lambda)(\sigma_2^4 - 2\sigma_1^2\rho^2 - 2\lambda\sigma_2^2 + \lambda^2) = 0\]

and we get \( \lambda_1 = \sigma_1^2(1 + \sqrt{2}\rho) \), \( \lambda_2 = \sigma_2^2 \), and \( \lambda_3 = \sigma_2^2(1 - \sqrt{2}\rho) \). To solve the eigenvector, we need to solve \( \Sigma e_i = \lambda_i e_i \), for \( i = 1, 2, 3 \). We found that

\[
e_1 = \begin{pmatrix} 1/2 \\ \sqrt{2}/2 \\ 1/2 \end{pmatrix}, \quad e_2 = \begin{pmatrix} \sqrt{2}/2 \\ 0 \\ -\sqrt{2}/2 \end{pmatrix}, \quad \text{and} \quad e_3 = \begin{pmatrix} 1/2 \\ -\sqrt{2}/2 \\ 1/2 \end{pmatrix}
\]

Therefore, the principal components become

\[
Y_1 = e_1^T X = \frac{1}{2}X_1 + \frac{\sqrt{2}}{2}X_2 + \frac{1}{2}X_3
\]

\[
Y_2 = e_2^T X = \frac{\sqrt{2}}{2}X_1 - \frac{\sqrt{2}}{2}X_3
\]

\[
Y_3 = e_3^T X = \frac{1}{2}X_1 - \frac{\sqrt{2}}{2}X_2 + \frac{1}{2}X_3
\]

The total population variance is

\[
\sum_{i=1}^{3} \text{Var}(Y_i) = \sum_{i=1}^{3} \lambda_i = \sigma_1^2(1 + \sqrt{2}\rho) + \sigma_2^2 + \sigma_2^2(1 - \sqrt{2}\rho) = 3\sigma^2
\]

and the proportion of total population variance explained by each principal components is: \( \frac{1}{3}(1 + \sqrt{2}\rho) \), \( \frac{1}{3} \), and \( \frac{1}{3}(1 - \sqrt{2}\rho) \), for \( Y_1 \), \( Y_2 \), and \( Y_3 \), respectively.

8.10. The weekly rates of return for five stocks listed on the New York Stock Exchange are given in Table 8.4.
(a) Construct the sample covariance matrix \( S \), and find the sample principal components in (8-20).

(b) Determine the proportion of the total sample variance explained by the first three principal components. Interpret these components.

(c) Construct Bonferroni simultaneous 90% confidence intervals for the variances \( \lambda_1, \lambda_2, \) and \( \lambda_3 \) of the first three population components \( Y_1, Y_2, \) and \( Y_3 \).

(d) Given the results in Parts (a)-(c), do you feel that the stock rates-of-return data can be summarized in fewer than five dimensions? Explain.

Sol. (a) The sample covariance matrix \( S \) is shown below:

\[
\begin{array}{cccccc}
\text{JPMorgan} & \text{CitiBank} & \text{WellsFargo} & \text{RoyDutShell} & \text{ExxonMobil} \\
0.00043327 & 0.00027567 & 0.00015903 & 0.00006412 & 0.00008897 \\
0.00027567 & 0.00043872 & 0.00017997 & 0.00018145 & 0.00012326 \\
0.00015903 & 0.00017997 & 0.00022397 & 0.00007341 & 0.00006055 \\
0.00006412 & 0.00018145 & 0.00007341 & 0.00072250 & 0.00050828 \\
0.00008897 & 0.00012326 & 0.00006055 & 0.00050828 & 0.00076567 \\
\end{array}
\]

The sample principle components are:

\[
\text{Standard deviations (1, \ldots, p=5):} \\
[1] 0.03698213 \quad 0.02647942 \quad 0.01593118 \quad 0.01194163 \quad 0.01090352 \\
\]

\[
\text{Rotation (n x k) = (5 x 5):} \\
\begin{array}{ccccc}
\text{PC1} & \text{PC2} & \text{PC3} & \text{PC4} & \text{PC5} \\
\text{JPMorgan} & -0.2228228 & 0.6252260 & -0.32611218 & 0.6627590 & -0.11765952 \\
\text{CitiBank} & -0.3072900 & 0.5703900 & 0.24959014 & -0.4140935 & 0.58860803 \\
\text{WellsFargo} & -0.1548103 & 0.3445049 & 0.03763929 & -0.4970499 & -0.78030428 \\
\text{RoyDutShell} & -0.6389680 & -0.2479475 & 0.64249741 & 0.3088689 & -0.14845546 \\
\text{ExxonMobil} & -0.6509044 & -0.3218478 & -0.64586064 & -0.2163758 & 0.09371777 \\
\end{array}
\]

(b) From part (a),

\[
\hat{\lambda}_1 = 0.00137, \quad \hat{\lambda}_2 = 0.00070, \quad \hat{\lambda}_3 = 0.00025, \quad \hat{\lambda}_4 = 0.00014, \quad \hat{\lambda}_5 = 0.00012,
\]

so the total sample variance is \( \sum_{i=1}^{5} \hat{\lambda}_i = 0.00258 \) and the proportion of total variance explained by the first three component is \( \frac{\sum_{i=1}^{3} \hat{\lambda}_i}{\sum_{i=1}^{5} \hat{\lambda}_i} = 0.8988 \). The first component might be interpreted as a “market” component with the greatest weight on Royal Dutch Shell and Exxon Mobil, the second component as an “industry” component that separates bank and gas companies, and the third component is a contrast between these five stocks which is difficult to interpret.

(c) The Bonferroni simultaneous 100(1 - \( \alpha \))% confidence interval for \( \lambda_i \) can be constructed by

\[
\frac{\hat{\lambda}_i}{1 + z \left( \frac{\alpha}{2m} \right) \sqrt{\frac{2}{n}}} \leq \lambda_i \leq \frac{\hat{\lambda}_i}{1 - z \left( \frac{\alpha}{2m} \right) \sqrt{\frac{2}{n}}}
\]

Thus the 90% confidence intervals for the three variance of the population components are:

\[
\lambda_1 : (0.001055, 0.001944)
\]
(d) Stock returns are probably best summarized in two dimensions with 80% of the total variation accounted for by a “market” component and an “industry” component without much loss of information.

8.28. Survey data were collected as part of a study to assess options for enhancing food security through the sustainable use of natural resources in the Sikasso region of Mali (West Africa). A total of \( n = 76 \) farmers were surveyed and observations on the nine variables

\[
\begin{align*}
  x_1 &= \text{Family (total number of individuals in household)} \\
  x_2 &= \text{DistRd (distance in kilometers to nearest passable road)} \\
  x_3 &= \text{Cotton (hectares of cotton planted in year 2000)} \\
  x_4 &= \text{Maize (hectares of maize planted in year 2000)} \\
  x_5 &= \text{Sorg (hectares of sorghum planted in year 2000)} \\
  x_6 &= \text{Millet (hectares of millet planted in year 2000)} \\
  x_7 &= \text{Bull (total number of bullocks or draft animals)} \\
  x_8 &= \text{Cattle (total); } x_9 = \text{Goats (total)}
\end{align*}
\]

were recorded. The data are listed in Table 8.7.

(a) Construct two-dimensional scatterplots of Family versus DistRd, and DistRd versus Cattle. Remove any obvious outliers from the data set.

(b) Perform a principal component analysis using the correlation matrix \( \mathbf{R} \). Determine the number of components to effectively summarize the variability. Use the proportion of variation explained and a scree plot to aid in your determination.

(c) Interpret the first five principal components. Can you identify, for example, a “farm size” component? A, perhaps, “goats and distance to road” component?

Sol. (a) Scatterplots of the two pairs of specified variables are shown in Figure 1.

![Scatterplots](image-url)

Figure 1: Scatterplots of Family versus DistRd (left) and DistRd versus Cattle (right).
Based on these scatterplots, we removed the four outliers (observations 25, 34, 69, 72) from the dataset. The scatterplots with outlier removed are plotted in Figure 2.

(b) The principal component analysis of $R$ follows. Removing the outliers has some but relatively little effect on the analysis.

Standard deviations (1, ..., p=9):

[1] 2.0457593 1.1992026 1.0413933 0.8898414 0.7773833 0.6050916
[7] 0.4899220 0.4145180 0.3437368

Rotation (n x k) = (9 x 9):

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family</td>
<td>0.433842713</td>
<td>-0.065088959</td>
<td>0.09840025</td>
<td>-0.17120143</td>
<td>0.01132705</td>
</tr>
<tr>
<td>DistRD</td>
<td>0.007587031</td>
<td>0.496670914</td>
<td>-0.56856059</td>
<td>-0.49561039</td>
<td>-0.3776811</td>
</tr>
<tr>
<td>Cotton</td>
<td>0.446140316</td>
<td>0.008917253</td>
<td>0.13211700</td>
<td>0.02733684</td>
<td>-0.21870789</td>
</tr>
<tr>
<td>Maize</td>
<td>0.352228405</td>
<td>0.35271495</td>
<td>0.38820350</td>
<td>-0.24020492</td>
<td>-0.07920345</td>
</tr>
<tr>
<td>Sorg</td>
<td>0.203622111</td>
<td>-0.60367416</td>
<td>-0.11149246</td>
<td>0.05854254</td>
<td>-0.64457383</td>
</tr>
<tr>
<td>Millet</td>
<td>0.240361102</td>
<td>-0.415159516</td>
<td>-0.1159577</td>
<td>-0.61632679</td>
<td>0.52696668</td>
</tr>
<tr>
<td>Bull</td>
<td>0.445273680</td>
<td>0.068042477</td>
<td>-0.03038787</td>
<td>0.14559178</td>
<td>-0.0282987</td>
</tr>
<tr>
<td>Cattle</td>
<td>0.355411543</td>
<td>0.284473439</td>
<td>0.01382636</td>
<td>0.37293370</td>
<td>0.21753184</td>
</tr>
<tr>
<td>Goats</td>
<td>0.254549533</td>
<td>-0.048668251</td>
<td>-0.68695528</td>
<td>0.35078804</td>
<td>0.24867109</td>
</tr>
<tr>
<td>PC6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PC7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PC8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PC9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>PC6</th>
<th>PC7</th>
<th>PC8</th>
<th>PC9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family</td>
<td>-0.03997862</td>
<td>-0.79746017</td>
<td>-0.26281017</td>
<td>-0.24862206</td>
</tr>
<tr>
<td>DistRD</td>
<td>0.18658220</td>
<td>0.02106965</td>
<td>-0.04790053</td>
<td>-0.06469259</td>
</tr>
<tr>
<td>Cotton</td>
<td>-0.19968612</td>
<td>0.36124785</td>
<td>0.32948454</td>
<td>-0.67521059</td>
</tr>
<tr>
<td>Maize</td>
<td>-0.27321206</td>
<td>-0.02328279</td>
<td>0.36297395</td>
<td>0.57444950</td>
</tr>
<tr>
<td>Sorg</td>
<td>0.24598733</td>
<td>-0.02061874</td>
<td>0.12556392</td>
<td>0.29340194</td>
</tr>
<tr>
<td>Millet</td>
<td>0.18077867</td>
<td>0.24070610</td>
<td>0.07713302</td>
<td>0.04795829</td>
</tr>
<tr>
<td>Bull</td>
<td>-0.13405394</td>
<td>0.39621919</td>
<td>-0.75050803</td>
<td>0.18962561</td>
</tr>
<tr>
<td>Cattle</td>
<td>0.75905049</td>
<td>-0.01063587</td>
<td>0.16866186</td>
<td>0.03806691</td>
</tr>
<tr>
<td>Goats</td>
<td>-0.40218231</td>
<td>-0.13068360</td>
<td>0.27368097</td>
<td>0.14936105</td>
</tr>
</tbody>
</table>
The proportion of variance explained by each component are: 46.50%, 15.98%, 12.05%, 8.80%, 6.71%, 4.07%, 2.67%, 1.91%, and 1.31%. Based on the screeplot and cumulative proportion of variance plot in Figure 3, we would like to choose the first five principal components to summarize this dataset. The first five components explain about 90% of the total variability in the data set and seems a reasonable number given the screeplot.

![Screeplot and cumulative proportion of variance plot](image)

Figure 3: Screeplot (left) and cumulative proportion of variance (right).

(c) All the variables (all crops, all livestock, family) except for distance to road (RistRd) load about equally on the first component. This component might be called a farm size component. Millet and sorghum load negative and distance to road and maize load positively on the second component. Without additional subject matter knowledge, this component is difficult to interpret. The third component is essentially a distance to the road and goats component. This component might represent subsistence farms. The fourth component appears to be a contrast between distance to road and millet versus cattle and goats. Again, this component is difficult to interpret. The fifth component appears to contrast sorghum with millet.

10.2. The $(2 \times 1)$ random vectors $X^{(1)}$ and $X^{(2)}$ have the joint mean vector and joint covariance matrix

$$
\mu = \begin{pmatrix} 
\mu^{(1)} \\
\mu^{(2)}
\end{pmatrix} = \begin{pmatrix} 
-3 \\
2 \\
0 \\
1
\end{pmatrix};
$$

$$
\Sigma = \begin{pmatrix} 
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{pmatrix} = \begin{pmatrix} 
8 & 2 & 3 & 1 \\
2 & 5 & -1 & 3 \\
3 & -1 & 6 & -2 \\
1 & 3 & -2 & 7
\end{pmatrix}
$$

(a) Calculate the canonical correlations $\rho_1$, $\rho_2$.

(b) Determine the canonical variate pairs $(U_1, V_1)$ and $(U_2, V_2)$.

(c) Let $U = (U_1, U_2)^T$ and $V = (V_1, V_2)^T$. From the first principles, evaluate

$$
E \left( \frac{U}{V} \right) \quad \text{and} \quad Cov \left( \frac{U}{V} \right) = \begin{pmatrix} 
\Sigma_{UU} & \Sigma_{UV} \\
\Sigma_{VU} & \Sigma_{VV}
\end{pmatrix}
$$
Compare your results with the properties in Result 10.1.

**Sol.** (a) The inverse and square root of the inverse of $\Sigma_{11}$ and $\Sigma_{22}$ are calculated by R compiled in the Appendix. We have

$$
\Sigma_{11}^{-\frac{1}{2}} = \begin{pmatrix} 0.1389 & -0.0556 \\ -0.0556 & 0.2222 \end{pmatrix}, \quad \Sigma_{12}^{-\frac{1}{2}} = \begin{pmatrix} 0.3667 & -0.0667 \\ -0.0667 & 0.4667 \end{pmatrix},
$$

$$
\Sigma_{22}^{-\frac{1}{2}} = \begin{pmatrix} 0.1842 & 0.0526 \\ 0.0526 & 0.1579 \end{pmatrix}, \quad \Sigma_{22}^{-\frac{1}{2}} = \begin{pmatrix} 0.4243 & 0.0645 \\ 0.0645 & 0.3921 \end{pmatrix}.
$$

Since $\rho^2 = (\rho_1^2, \rho_2^2)$ are the eigenvalues of the matrix $\Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}} \Sigma_{11}^{-\frac{1}{2}}$ with corresponding $(2 \times 1)$ eigenvectors $h_1, h_2$. (The quantities $\rho^2$ are also the eigenvalues of the matrix $\Sigma_{22}^{-\frac{1}{2}} \Sigma_{21} \Sigma_{11}^{-\frac{1}{2}} \Sigma_{22}^{-\frac{1}{2}}$ with corresponding $(2 \times 1)$ eigenvectors $f_1, f_2.$)

$$
\Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}} \Sigma_{21} \Sigma_{11}^{-\frac{1}{2}} = \begin{pmatrix} 0.2756 & -0.0322 \\ -0.0322 & 0.2690 \end{pmatrix}
$$

and

$$
\Sigma_{22}^{-\frac{1}{2}} \Sigma_{21} \Sigma_{11}^{-\frac{1}{2}} \Sigma_{22}^{-\frac{1}{2}} = \begin{pmatrix} 0.2946 & -0.0234 \\ -0.0234 & 0.2500 \end{pmatrix}
$$

The eigenvalues are $(\rho_1^2, \rho_2^2) = (0.3046, 0.2400)$ with the corresponding eigenvectors $H = (h_1, h_2)$ and $Q = (f_1, f_2)$, respectively. Here

$$
h_1 = \begin{pmatrix} -0.7422 \\ 0.6702 \end{pmatrix}, \quad h_2 = \begin{pmatrix} -0.6702 \\ -0.7422 \end{pmatrix}, \quad f_1 = \begin{pmatrix} -0.9194 \\ 0.3936 \end{pmatrix}, \quad \text{and} \quad f_2 = \begin{pmatrix} -0.3936 \\ -0.9193 \end{pmatrix}.
$$

So the canonical correlations $(\rho_1, \rho_2) = (0.5519, 0.4899)$.

(b) The canonical variate pairs:

$$
U_1 = h_1^T \Sigma_{11}^{-\frac{1}{2}} X^{(1)} = -0.3168 X_1^{(1)} + 0.3622 X_2^{(1)},
$$

$$
V_1 = f_1^T \Sigma_{22}^{-\frac{1}{2}} X^{(2)} = -0.3647 X_1^{(2)} + 0.0951 X_2^{(2)},
$$

$$
U_2 = h_2^T \Sigma_{11}^{-\frac{1}{2}} X^{(1)} = -0.1962 X_1^{(1)} - 0.3017 X_2^{(1)},
$$

$$
V_2 = f_1^T \Sigma_{22}^{-\frac{1}{2}} X^{(2)} = -0.2263 X_1^{(2)} - 0.3858 X_2^{(2)}.
$$

(c) Since $U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = H^T \Sigma_{11}^{-\frac{1}{2}} X^{(1)}$ and $V = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} = Q^T \Sigma_{22}^{-\frac{1}{2}} X^{(2)}$

$$
E \left( \begin{pmatrix} U \\ V \end{pmatrix} \right) = \begin{pmatrix} H^T \Sigma_{11}^{-\frac{1}{2}} \mu^{(1)} \\ Q^T \Sigma_{22}^{-\frac{1}{2}} \mu^{(2)} \end{pmatrix} = \begin{pmatrix} 1.6749 \\ -0.0146 \\ 0.9091 \end{pmatrix},
$$

$$
\text{Cov} \left( \begin{pmatrix} U \\ V \end{pmatrix} \right) = \begin{pmatrix} H^T \Sigma_{11}^{-\frac{1}{2}} X^{(1)} \\ Q^T \Sigma_{22}^{-\frac{1}{2}} X^{(2)} \end{pmatrix} = \begin{pmatrix} H^T \Sigma_{11}^{-\frac{1}{2}} \Sigma_{11}^{-\frac{1}{2}} H & H^T \Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}} Q \\ Q^T \Sigma_{22}^{-\frac{1}{2}} \Sigma_{21} \Sigma_{11}^{-\frac{1}{2}} H & Q^T \Sigma_{22}^{-\frac{1}{2}} \Sigma_{22}^{-\frac{1}{2}} Q \end{pmatrix}
$$

$$
= \begin{pmatrix} 1 & 0 & 0.5519 & 0 \\ 0 & 1 & 0 & 0.4899 \\ 0.5519 & 0 & 1 & 0 \\ 0 & 0.4899 & 0 & 1 \end{pmatrix}
$$
The above result shows that $\text{Corr}(U_k, V_k) = \rho_k$ and 

\[
\text{Var}(U_k) = \text{Var}(V_k) = 1 \\
\text{Cov}(U_k, U_l) = \text{Corr}(U_k, U_l) = 0 \quad k \neq l \\
\text{Cov}(V_k, V_l) = \text{Corr}(V_k, V_l) = 0 \quad k \neq l \\
\text{Cov}(U_k, V_l) = \text{Corr}(U_k, V_l) = 0 \quad k \neq l
\]

for $k, l = 1, 2$. This result coincide with the properties in Result 10.1.

10.10. In a study of poverty, crime, and deterrence, Parker and Smith [10] report certain summary crime statistics in various states for the years 1970 and 1973. A portion of their sample correlation matrix is

\[
R = \begin{pmatrix}
R_{11} & R_{12} \\
R_{21} & R_{22}
\end{pmatrix} = \begin{pmatrix}
1.000 & 0.615 & -0.111 & -0.266 \\
0.615 & 1.000 & -0.195 & -0.085 \\
-0.111 & -0.195 & 1.000 & -0.269 \\
-0.266 & -0.085 & -0.269 & 1.000
\end{pmatrix}
\]

The variables are

- $X_1^{(1)} = 1973$ nonprimary homicides
- $X_2^{(1)} = 1973$ primary homicides (homicides involving family or acquaintances)
- $X_1^{(2)} = 1970$ severity of punishment (median months served)
- $X_2^{(2)} = 1970$ certainty of punishment (number of admissions to prison divided by number of homicides)

(a) Find the sample canonical correlations.

(b) Determine the first canonical pair $\hat{U}_1$, $\hat{V}_1$ and interpret these quantities.

\textbf{Sol.} (a) The inverse and square root of the inverse of $R_{11}$ and $R_{22}$ are calculated by $R$ compiled in the Appendix. We have

\[
R^{-1}_{11} = \begin{pmatrix}
1.6083 & -0.9891 \\
-0.9891 & 1.6083
\end{pmatrix}, \quad R^{-\frac{1}{2}}_{11} = \begin{pmatrix}
1.1993 & -0.4124 \\
-0.4124 & 1.1993
\end{pmatrix}
\]

\[
R^{-1}_{22} = \begin{pmatrix}
1.0780 & 0.2900 \\
0.2900 & 1.0780
\end{pmatrix}, \quad R^{-\frac{1}{2}}_{22} = \begin{pmatrix}
1.0287 & 0.1410 \\
0.1410 & 1.0287
\end{pmatrix}
\]

\[
R^{-\frac{1}{2}}_{11} R_{12} R^{-\frac{1}{2}}_{22} R_{21} R^{-\frac{1}{2}}_{11} = \begin{pmatrix}
0.0986 & 0.0237 \\
0.0237 & 0.0374
\end{pmatrix}
\]

and

\[
R^{-\frac{1}{2}}_{22} R_{21} R^{-\frac{1}{2}}_{11} R_{12} R^{-\frac{1}{2}}_{22} = \begin{pmatrix}
0.0459 & 0.0318 \\
0.0318 & 0.0900
\end{pmatrix}
\]

The eigenvalues are $(\rho_1^2, \rho_2^2) = (0.1067, 0.0293)$ with the corresponding eigenvectors $H = (h_1, h_2)$ and $Q = (f_1, f_2)$, respectively. Here

\[
h_1 = \begin{pmatrix}
-0.9463 \\
-0.3232
\end{pmatrix}, \quad h_2 = \begin{pmatrix}
0.3232 \\
-0.9463
\end{pmatrix}, \quad f_1 = \begin{pmatrix}
0.4634 \\
0.8861
\end{pmatrix}, \quad \text{and} \quad f_2 = \begin{pmatrix}
-0.8861 \\
0.4634
\end{pmatrix}.
\]

So the canonical correlations $(\rho_1, \rho_2) = (0.3266, 0.1711)$. 
(b) The first canonical variate pairs:

\[
\hat{U}_1 = h^T_1 \Sigma_{11}^{-\frac{1}{2}} Z^{(1)} = -1.0016 Z^{(1)}_1 + 0.0026 Z^{(1)}_2 \approx -Z^{(1)}_1
\]

\[
\hat{V}_1 = f^T_1 \Sigma_{22}^{-\frac{1}{2}} Z^{(2)} = 0.6016 Z^{(2)}_1 + 0.9769 Z^{(2)}_2 \approx \frac{3}{5} Z^{(2)}_1 + Z^{(2)}_2
\]

Since \(\hat{U}_1\) approximately equals \(-Z^{(1)}_1\), we can interpret the canonical variate \(\hat{U}_1\) as the standardized \(X^{(1)}_1 = 1973\) nonprimary homicides. On the other hand, \(\hat{V}_1\) approximately equals \(\frac{3}{5} Z^{(2)}_1 + Z^{(2)}_2\), we can interpret the canonical variate \(\hat{V}_1\) as a punishment index. Punishment appears to be correlated with nonprimary homicides but not primary homicides.

10.13. Waugh [12] provides information about \(n = 138\) samples of Canadian hard red spring wheat and the flour made from the samples. The \(p = 5\) wheat measurements (in standardized form) were

- \(z^{(1)}_1\) = kernel texture
- \(z^{(1)}_2\) = test weight
- \(z^{(1)}_3\) = damaged kernels
- \(z^{(1)}_4\) = foreign material
- \(z^{(1)}_5\) = crude protein in the wheat

The \(q = 4\) (standardized) flour measurements were

- \(z^{(2)}_1\) = wheat per barrel of flour
- \(z^{(2)}_2\) = ash in flour
- \(z^{(2)}_3\) = crude protein in flour
- \(z^{(2)}_4\) = gluten quality index

The sample correlation matrix was

\[
R = \begin{pmatrix}
1.000 & & \\
0.754 & 1.000 & \\
-0.690 & -0.712 & 1.000 \\
-0.446 & -0.515 & 0.323 & 1.000 \\
0.692 & 0.412 & -0.444 & -0.334 & 1.000 \\
-0.605 & -0.772 & 0.737 & 0.527 & -0.383 & 1.000 \\
-0.479 & -0.419 & 0.361 & 0.461 & -0.505 & 0.251 & 1.000 \\
0.780 & 0.542 & -0.546 & -0.393 & 0.737 & -0.490 & -0.434 & 1.000 \\
-0.152 & -0.102 & 0.172 & -0.019 & -0.148 & 0.250 & -0.079 & -0.163 & 1.000
\end{pmatrix}
\]

(a) Find the sample canonical variates corresponding to significant (at the \(\alpha = 0.01\) level) canonical correlations.

(b) Interpret the first sample canonical variates \(\hat{U}_1, \hat{V}_1\). Do they in some sense represent the overall quality of the wheat and flour, respectively?
(c) What proportion of the total sample variance of the first set $Z^{(1)}$ is explained by the canonical variate $\hat{U}_1$? What proportion of the total sample variance of the $Z^{(2)}$ set is explained by the canonical variate $\hat{V}_1$? Discuss your answers.

Sol. (a) We calculate the canonical correlation by $R$ compiled in Appendix. The canonical correlations are: $(\hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3, \hat{\rho}_4) = (0.9158, 0.6706, 0.2544, 0.0940)$. We then run the hypothesis testing and summarize the results in the following table:

<table>
<thead>
<tr>
<th>Null hypothesis</th>
<th>Test Statistic</th>
<th>Df</th>
<th>$\chi^2$-value</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0: \Sigma_{12} = \rho_{12} = 0$</td>
<td>329.6947</td>
<td>20</td>
<td>37.5662</td>
<td>Reject $H_0$</td>
</tr>
<tr>
<td>$H_0: \rho_1 \neq 0$, $\rho_2 = \rho_3 = \rho_4 = 0$</td>
<td>88.8550</td>
<td>12</td>
<td>26.2170</td>
<td>Reject $H_0$</td>
</tr>
<tr>
<td>$H_0: \rho_1 \neq 0$, $\rho_2 \neq 0$, $\rho_3 = \rho_4 = 0$</td>
<td>10.0012</td>
<td>6</td>
<td>16.8119</td>
<td>Do not reject $H_0$</td>
</tr>
</tbody>
</table>

$\hat{U}_1 = -0.1176z_{(1)}^1 - 0.3004z_{(1)}^2 + 0.3160z_{(1)}^3 + 0.2509z_{(1)}^4 - 0.2937z_{(1)}^5$

$\hat{V}_1 = 0.5930z_{(2)}^1 + 0.2856z_{(2)}^2 - 0.4017z_{(2)}^3 - 0.0366z_{(2)}^4$

$\hat{U}_2 = -1.0204z_{(1)}^1 + 0.7809z_{(1)}^2 - 0.5102z_{(1)}^3 - 0.2463z_{(1)}^4 - 0.5000z_{(1)}^5$

$\hat{V}_2 = 0.9809z_{(2)}^1 - 0.0020z_{(2)}^2 + 0.9956z_{(2)}^3 - 0.1821z_{(2)}^4$

(b) $\hat{U}_1$ appears to measure quality of wheat as a contrast between negative aspects $z_{(1)}^1$, $z_{(1)}^2$, and $z_{(1)}^5$ and positive aspects $z_{(1)}^3$ and $z_{(1)}^4$. $\hat{V}_1$ appears to measure the quality of the flour as represented by $z_{(2)}^1$, $z_{(2)}^2$, and $z_{(2)}^3$.

(c) We find that the proportion of the total sample variance of the first set $Z^{(1)}$ explained by the canonical variate $\hat{U}_1$ is $\rho_{1}^{(1)} = \frac{1}{p}(a_{(1)}^T a_{(1)})^T = 0.6297$ and the proportion of the total sample variance of the $Z^{(2)}$ set explained by the canonical variate $\hat{V}_1$ is $\rho_{2}^{(2)} = \frac{1}{q}(b_{(2)}^T b_{(2)})^T = 0.4453$.

Appendix

R code for Problem 8.10.

```r
> stocks <- read.table('./T8-4.DAT', col.names = c("JPMorgan", "CitiBank", "WellsFargo", "RoyDutShell", "ExxonMobil"))
> # (a)
> S <- cov(stocks)
> pca <- prcomp(stocks)
> # (b)
> lambda <- pca$sdev^2
> cumsum(lambda/sum(lambda))
> # (c)
> n <- nrow(stocks)
> alpha <- 0.1
> m <- 3
> CI.LB <- lambda[1:m]/(1+qnorm(1-alpha/(2*m))*sqrt(2/n))
> CI.UB <- lambda[1:m]/(1-qnorm(1-alpha/(2*m))*sqrt(2/n))

9
R code for Problem 8.28.

```r
>
> # (a) scatterplots of Family versus DistRD
> plot(farm$Family, farm$DistRD, xlab="Family", ylab="DistRD")
> plot(farm$DistRD, farm$Cattle, xlab="DistRD", ylab="Cattle")
>
> farm1 <- farm[-c(25,34,69,72),]
> plot(farm1$Family, farm1$DistRD, xlab="Family", ylab="DistRD")
> plot(farm1$DistRD, farm1$Cattle, xlab="DistRD", ylab="Cattle")
>
> # (b) PCA on correlation matrix R
> pca <- prcomp(farm1, center = TRUE, scale = TRUE)
> # screeplot
> plot(1:length(pca$sdev), pca$sdev^2, type="b", +
      xlab="Number of PCs", ylab="Variance explained")
>
> # proportion of variance explained
> plot(1:length(pca$sdev), cumsum(pca$sdev^2)/sum(pca$sdev^2), type="b", +
      xlab="number of PCs", ylab="Cumulative proportion")
>
> # proportion of variance explained by each component
> pca$sdev^2/sum(pca$sdev^2)*100
```

R code for Problem 10.2.

```r
> mu1 <- c(-3,2)
> mu2 <- c(0,1)
> S11 <- matrix(c(8,2,2,5), nrow=2, ncol=2)
> S12 <- matrix(c(3,-1,1,3), nrow=2, ncol=2)
> S21 <- t(S12)
> S22 <- matrix(c(6,-2,-2,7), nrow=2, ncol=2)
>
> eig11 <- eigen(S11)
> S11inv <- solve(S11)
> S11invsq <- eig11$vectors %*% diag(sqrt(eig11$values)^(-1)) %*% t(eig11$vectors)
>
> eig22 <- eigen(S22)
> S22inv <- solve(S22)
> S22invsq <- eig22$vectors %*% diag(sqrt(eig22$values)^(-1)) %*% t(eig22$vectors)
>
> # (a)
> rho <- sqrt(eigen(S11invsq %*% S12 %*% S22inv %*% S21 %*% S11invsq)$values)
>
> # (b)
> H <- eigen(S11invsq %*% S12 %*% S22inv %*% S21 %*% S11invsq)$vectors
> Q <- eigen(S22invsq %*% S21 %*% S11inv %*% S12 %*% S22invsq)$vectors
>
> t(H) %*% S11invsq
> t(Q) %*% S22invsq
```
\[ \begin{align*}
EU & = t(H) \%*% S11invsq \%*% mu1 \\
EV & = t(Q) \%*% S22invsq \%*% mu2 \\
SUU & = t(H) \%*% S11invsq \%*% S11invsq \%*% H \\
SUV & = t(H) \%*% S11invsq \%*% S12 \%*% S22invsq \%*% Q \\
SVU & = t(Q) \%*% S22invsq \%*% S21 \%*% S11invsq \%*% H \\
SVV & = t(Q) \%*% S22invsq \%*% S22 \%*% S22invsq \%*% Q
\end{align*} \]

R code for Problem 10.10.

\[
\begin{align*}
R11 & = \text{matrix}(c(1,0.615,0.615,1), \text{nrow}=2, \text{ncol}=2) \\
R12 & = \text{matrix}(c(-0.111,-0.195,-0.266,-0.085), \text{nrow}=2, \text{ncol}=2) \\
R21 & = t(R12) \\
R22 & = \text{matrix}(c(1,-0.269,-0.269,1), \text{nrow}=2, \text{ncol}=2)
\end{align*}
\]

\[# (a)\]

eig11 <- eigen(R11)

R11inv <- solve(R11)

R11invsq <- eig11$vectors \%*% diag(sqrt(eig11$values)^(-1)) \%*% t(eig11$vectors)

\[
\begin{align*}
eig22 & = \text{eigen}(R22) \\
R22inv & = \text{solve}(R22) \\
R22invsq & = eig22$vectors \%*% diag(sqrt(eig22$values)^(-1)) \%*% t(eig22$vectors)
\end{align*}
\]

\[
\text{rho} = \text{sqrt(eigen}(R11invsq \%*% R12 \%*% R22inv \%*% R21 \%*% R11invsq)$values)
\]

\[# (b)\]

H & = \text{eigen}(R11invsq \%*% R12 \%*% R22inv \%*% R21 \%*% R11invsq)$vectors

Q & = \text{eigen}(R22invsq \%*% R21 \%*% R11inv \%*% R12 \%*% R22invsq)$vectors

\[
\begin{align*}
t(H[,1]) & \%*% R11invsq \\
t(Q[,1]) & \%*% R22invsq
\end{align*}
\]

R code for Problem 10.13.

\[
\begin{align*}
R11 & = \text{matrix}(c(1.000,0.754,-0.690,-0.446,0.692, \\
+ 0.754,1.000,-0.712,-0.515,0.412, \\
+ -0.690,-0.712,1.000,0.323,-0.444, \\
+ -0.446,-0.515,0.323,1.000,-0.334, \\
+ 0.692,0.412,-0.444,-0.334,1.000), \text{nrow}=5, \text{ncol}=5, \text{byrow}=\text{TRUE}) \\
R21 & = \text{matrix}(c(-0.605,-0.772,0.737,0.527,-0.383, \\
+ -0.479,-0.419,0.361,0.461,-0.505, \\
+ 0.780,0.542,-0.546,-0.393,0.737, \\
+ -0.152,-0.102,0.172,-0.019,-0.148), \text{nrow}=4, \text{ncol}=5, \text{byrow}=\text{TRUE}) \\
R22 & = \text{matrix}(c(1.000,0.251,-0.490,0.250, \\
+ 0.251,1.000,-0.434,-0.079, \\
+ -0.490,-0.434,1.000,-0.163, \\
+ 0.250,-0.079,-0.163,1.000), \text{nrow}=4, \text{ncol}=4, \text{byrow}=\text{TRUE})
\end{align*}
\]
> n <- 138
> p <- 5
> q <- 4
> d <- min(p,q)
> alpha <- 0.01
>
> eig11 <- eigen(R11)
> R11inv <- solve(R11)
> R11invvsq <- eig11$vectors %*% diag(sqrt(eig11$values)^(-1)) %*% t(eig11$vectors)
>
> eig22 <- eigen(R22)
> R22inv <- solve(R22)
> R22invvsq <- eig22$vectors %*% diag(sqrt(eig22$values)^(-1)) %*% t(eig22$vectors)
>
> rho2 <- eigen(R11invvsq %*% R12 %*% R22inv %*% R21 %*% R11invvsq)$values
> rho2[p] <- 0
> rho <- sqrt(rho2)
>
> H <- eigen(R11invvsq %*% R12 %*% R22inv %*% R21 %*% R11invvsq)$vectors
> Q <- eigen(R22invvsq %*% R21 %*% R11inv %*% R12 %*% R22invvsq)$vectors
>
> # (a) Sequential test
> for ( i in 1:d){
+   TS <- -(n-1-1/2*(p+q+1))*log(prod(1-rho2[i:d]))
+   dfs <- (p-1)*(q-1)
+   pval <- qchisq(1-alpha,df=dfs)
+   print(c(TS, dfs, pval))
+ }
>
> U <- t(H[,1:2]) %*% R11invvsq
> V <- t(Q[,1:2]) %*% R22invvsq
>
> # (c)
> Azinv <- solve(t(H) %*% R11invvsq)
> Bzinv <- solve(t(Q) %*% R22invvsq)
>
> rho1exp <- crossprod(Azinv[,1])/p
> rho2exp <- crossprod(Bzinv[,1])/q