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

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & \sigma^2\rho & 0 \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ 0 & \sigma^2\rho & \sigma^2 \end{pmatrix}, \quad -\frac{1}{\sqrt{2}} < \rho < \frac{1}{\sqrt{2}}$$

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

$$\begin{vmatrix} \sigma^2-\lambda & \sigma^2\rho & 0 \\ \sigma^2\rho & \sigma^2-\lambda & \sigma^2\rho \\ 0 & \sigma^2\rho & \sigma^2-\lambda \end{vmatrix} = 0.$$

By solving the system, we obtain the characteristic polynomial in terms of λ as:

$$(\sigma^2 - \lambda)(\sigma^4 - 2\sigma^4\rho^2 - 2\lambda\sigma^2 + \lambda^2) = 0$$

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

$$\boldsymbol{e}_1 = \begin{pmatrix} 1/2 \\ \sqrt{2}/2 \\ 1/2 \end{pmatrix}, \quad \boldsymbol{e}_2 = \begin{pmatrix} \sqrt{2}/2 \\ 0 \\ -\sqrt{2}/2 \end{pmatrix}, \quad \text{and} \quad \boldsymbol{e}_3 = \begin{pmatrix} 1/2 \\ -\sqrt{2}/2 \\ 1/2 \end{pmatrix}$$

Therefore, the principal components become

$$\begin{aligned} Y_1 = \boldsymbol{e}_1^T\boldsymbol{X} &= \frac{1}{2}X_1 + \frac{\sqrt{2}}{2}X_2 + \frac{1}{2}X_3 \\ Y_2 = \boldsymbol{e}_2^T\boldsymbol{X} &= \frac{\sqrt{2}}{2}X_1 - \frac{\sqrt{2}}{2}X_3 \\ Y_3 = \boldsymbol{e}_3^T\boldsymbol{X} &= \frac{1}{2}X_1 - \frac{\sqrt{2}}{2}X_2 + \frac{1}{2}X_3 \end{aligned}$$

The total population variance is

$$\sum_{i=1}^{3} Var(Y_i) = \sum_{i=1}^{3} \lambda_i = \sigma^2(1+\sqrt{2}\rho) + \sigma^2 + \sigma^2(1-\sqrt{2}\rho) = 3\sigma^2$$

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 $\boldsymbol{S}$, 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 λ_1, λ_2, and λ_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 $\boldsymbol{S}$ is shown below:

```
              JPMorgan   CitiBank WellsFargo RoyDutShell ExxonMobil
JPMorgan    0.00043327 0.00027567 0.00015903  0.00006412 0.00008897
CitiBank    0.00027567 0.00043872 0.00017997  0.00018145 0.00012326
WellsFargo  0.00015903 0.00017997 0.00022397  0.00007341 0.00006055
RoyDutShell 0.00006412 0.00018145 0.00007341  0.00072250 0.00050828
ExxonMobil  0.00008897 0.00012326 0.00006055  0.00050828 0.00076567
```

The sample principle components are:

```
Standard deviations (1, .., p=5):
[1] 0.03698213 0.02647942 0.01593118 0.01194163 0.01090352

Rotation (n x k) = (5 x 5):
                   PC1        PC2         PC3        PC4         PC5
JPMorgan    -0.2228228  0.6252260 -0.32611218  0.6627590 -0.11765952
CitiBank    -0.3072900  0.5703900  0.24959014 -0.4140935  0.58860803
WellsFargo  -0.1548103  0.3445049  0.03763929 -0.4970499 -0.78030428
RoyDutShell -0.6389680 -0.2479475  0.64249741  0.3088689 -0.14845546
ExxonMobil  -0.6509044 -0.3218478 -0.64586064 -0.2163758  0.09371777
```

(b) From part (a),

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

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 $\sum_{i=1}^{3} \hat{\lambda}_i / \sum_{i=1}^{5} \hat{\lambda}_i = 0.8988$. The first component might be interpreted as a "market" component with the greast 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 λ_i can be constructed by

$$\frac{\hat{\lambda}_i}{1 + z\left(\frac{\alpha}{2m}\right)\sqrt{\frac{2}{n}}} \le \lambda_i \le \frac{\hat{\lambda}_i}{1 - z\left(\frac{\alpha}{2m}\right)\sqrt{\frac{2}{n}}}.$$

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

$$\lambda_1 : (0.001055, 0.001944)$$

$$\lambda_2 : (0.000541, 0.000997)$$
$$\lambda_3 : (0.000196, 0.000361)$$

(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

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

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 $\boldsymbol{R}$. 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.

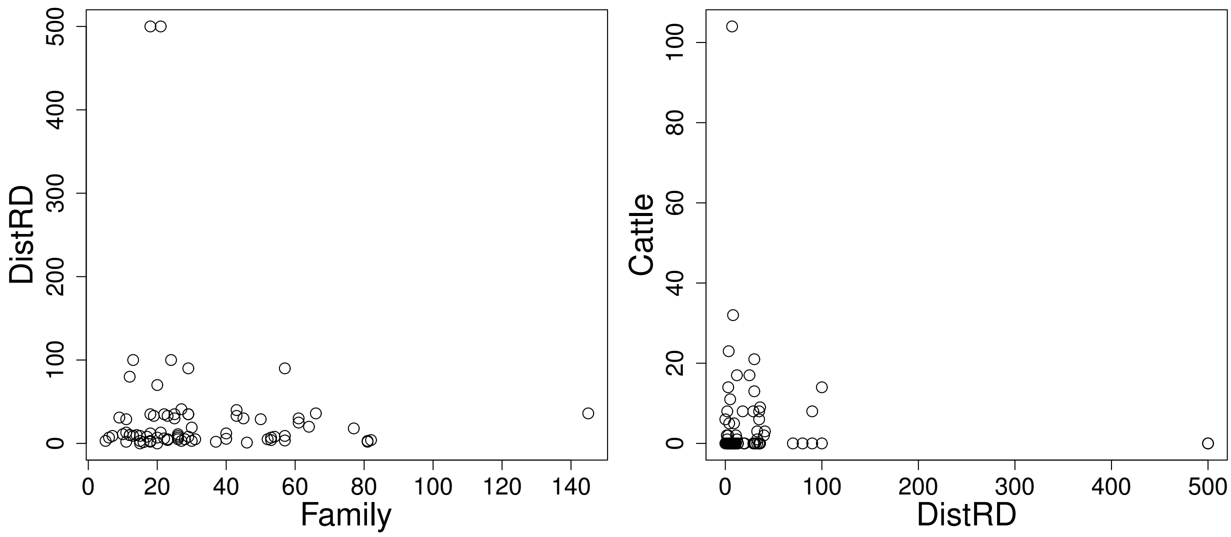

Figure 1: Scatterplots of Family versue DistRd (left) and DistRd versus Cattle (right).

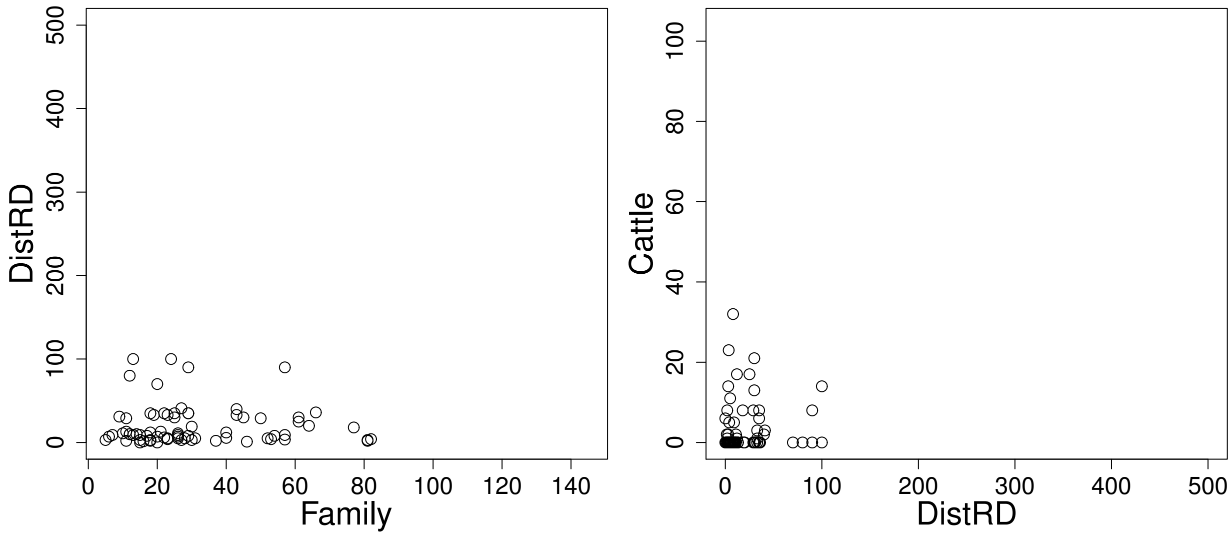

Figure 2: Scatterplots of Family versue DistRd (left) and DistRd versus Cattle (right). The outliers are removed from the dataset.

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 $\boldsymbol{R}$ 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):
              PC1          PC2         PC3         PC4         PC5
Family 0.433842713 -0.065088695  0.09840025 -0.17120143  0.01132705
DistRD 0.007587031  0.496670914 -0.56856059 -0.49561039 -0.37766811
Cotton 0.446140316  0.008917253  0.13211700  0.02733684 -0.21870789
Maize  0.352228405  0.352571495  0.38820350 -0.24020492 -0.07920345
Sorg   0.203622111 -0.603667416 -0.11149246  0.05854254 -0.64457738
Millet 0.240361102 -0.415159516 -0.11595977 -0.61632679  0.52696668
Bull   0.445273680  0.068042477 -0.03038787  0.14559178 -0.02829987
Cattle 0.355411548  0.284473439  0.01382636  0.37293370  0.21753184
Goats  0.254549533 -0.048668251 -0.68695528  0.35078804  0.24867109
              PC6         PC7         PC8         PC9
Family -0.03997862 -0.79746017 -0.26281017 -0.24862206
DistRD  0.18658220  0.02106965 -0.04790053 -0.06469259
Cotton -0.19968612  0.36124785  0.32948454 -0.67521059
Maize  -0.27321206 -0.02382879  0.36297395  0.57444950
Sorg    0.24598733 -0.02061874  0.12556392  0.29340194
Millet  0.18077867  0.24070610  0.07713302  0.04795829
Bull   -0.13405398  0.39621919 -0.75050803  0.18962561
Cattle  0.75905049 -0.01063587  0.16866186  0.03806691
Goats  -0.40218231 -0.13068360  0.27368097  0.14936105
```

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.

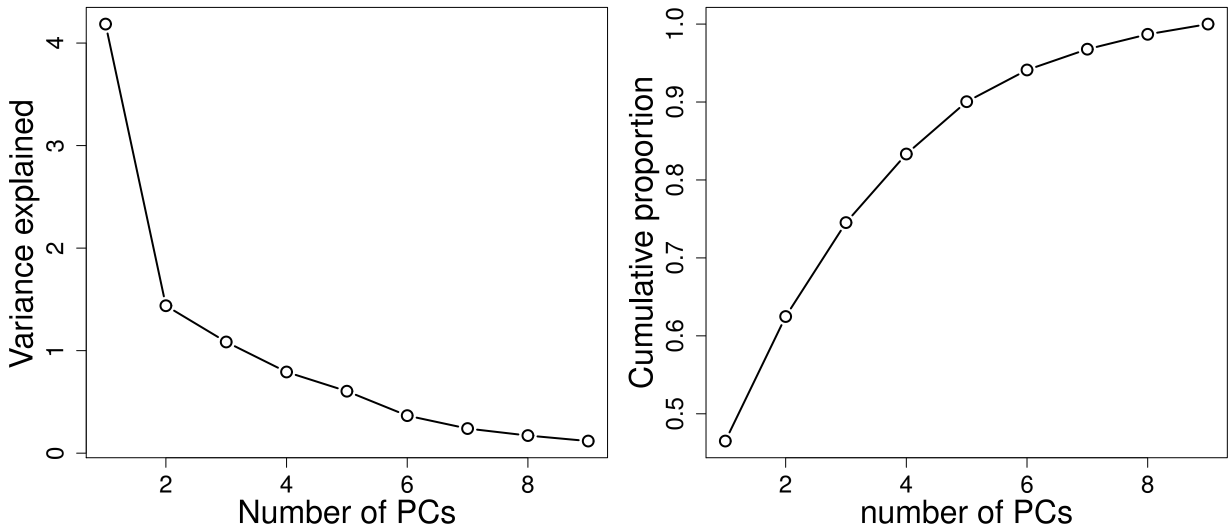

Figure 3: Screeplot (left) and cumulative proportion of variance (right).

(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 (2×1) random vectors $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ have the joint mean vector and joint covariance matrix

$$\boldsymbol{\mu} = \left(\frac{\boldsymbol{\mu}^{(1)}}{\boldsymbol{\mu}^{(2)}}\right) = \left(\begin{array}{c} -3 \\ 2 \\ \hline 0 \\ 1 \end{array}\right);$$

$$\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right) = \left(\begin{array}{rr|rr} 8 & 2 & 3 & 1 \\ 2 & 5 & -1 & 3 \\ \hline 3 & -1 & 6 & -2 \\ 1 & 3 & -2 & 7 \end{array}\right)$$

(a) Calculate the canonical correlations ρ_1, ρ_2.

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

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

$$E\left(\frac{\boldsymbol{U}}{\boldsymbol{V}}\right) \quad \text{and} \quad Cov\left(\frac{\boldsymbol{U}}{\boldsymbol{V}}\right) = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{UU} & \boldsymbol{\Sigma}_{UV} \\ \hline \boldsymbol{\Sigma}_{VU} & \boldsymbol{\Sigma}_{VV} \end{array}\right)$$

Compare your results with the properties in Result 10.1.

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

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

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

Since $\boldsymbol{\rho}^2 = (\rho_1^2, \rho_2^2)$ are the eigenvalues of the matrix $\mathbf{\Sigma}_{11}^{-\frac{1}{2}}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-\frac{1}{2}}$ with corresponding (2×1) eigenvectors $\boldsymbol{h}_1, \boldsymbol{h}_2$. (The quantities $\boldsymbol{\rho}^2$ are also the eigenvalues of the matrix $\mathbf{\Sigma}_{22}^{-\frac{1}{2}}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-\frac{1}{2}}$ with corresponding (2×1) eigenvectors $\boldsymbol{f}_1, \boldsymbol{f}_2$.)

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

and

$$\mathbf{\Sigma}_{22}^{-\frac{1}{2}}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-\frac{1}{2}} = \begin{pmatrix} 0.2946 & -0.0234 \\ -0.0234 & 0.2500 \end{pmatrix}$$

The eigenvalues are $(\rho_1^2, \rho_2^2) = (0.3046, 0.2400)$ with the corresponding eigenvectors $\boldsymbol{H} = (\boldsymbol{h}_1, \boldsymbol{h}_2)$ and $\boldsymbol{Q} = (\boldsymbol{f}_1, \boldsymbol{f}_2)$, respectively. Here

$$\boldsymbol{h}_1 = \begin{pmatrix} -0.7422 \\ 0.6702 \end{pmatrix}, \quad \boldsymbol{h}_2 = \begin{pmatrix} -0.6702 \\ -0.7422 \end{pmatrix}, \quad \boldsymbol{f}_1 = \begin{pmatrix} -0.9194 \\ 0.3936 \end{pmatrix}, \quad \text{and } \boldsymbol{f}_2 = \begin{pmatrix} -0.3936 \\ -0.9193 \end{pmatrix}.$$

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

(b) The canonical variate pairs:

$$\begin{aligned} U_1 &= \boldsymbol{h}_1^T\mathbf{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{X}^{(1)} = -0.3168X_1^{(1)} + 0.3622X_2^{(1)} \\ V_1 &= \boldsymbol{f}_1^T\mathbf{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{X}^{(2)} = -0.3647X_1^{(2)} + 0.0951X_2^{(2)} \\ U_2 &= \boldsymbol{h}_2^T\mathbf{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{X}^{(1)} = -0.1962X_1^{(1)} - 0.3017X_2^{(1)} \\ V_2 &= \boldsymbol{f}_1^T\mathbf{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{X}^{(2)} = -0.2263X_1^{(2)} - 0.3858X_2^{(2)} \end{aligned}$$

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

$$E\left(\frac{\boldsymbol{U}}{\boldsymbol{V}}\right) = \left(\frac{\boldsymbol{H}^T\mathbf{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{\mu}^{(1)}}{\boldsymbol{Q}^T\mathbf{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{\mu}^{(2)}}\right) = \left(\begin{array}{c} 1.6749 \\ -0.0146 \\ \hline 0.0951 \\ -0.3858 \end{array}\right)$$

$$\begin{aligned} Cov\left(\frac{\boldsymbol{U}}{\boldsymbol{V}}\right) &= Cov\left(\frac{\boldsymbol{H}^T\mathbf{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{X}^{(1)}}{\boldsymbol{Q}^T\mathbf{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{X}^{(2)}}\right) \\ &= \left(\begin{array}{c|c} \boldsymbol{H}^T\mathbf{\Sigma}_{11}^{-\frac{1}{2}}\mathbf{\Sigma}_{11}\mathbf{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{H} & \boldsymbol{H}^T\mathbf{\Sigma}_{11}^{-\frac{1}{2}}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{Q} \\ \hline \boldsymbol{Q}^T\mathbf{\Sigma}_{22}^{-\frac{1}{2}}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{H} & \boldsymbol{Q}^T\mathbf{\Sigma}_{22}^{-\frac{1}{2}}\mathbf{\Sigma}_{22}\mathbf{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{Q} \end{array}\right) \\ &= \left(\begin{array}{cc|cc} 1 & 0 & 0.5519 & 0 \\ 0 & 1 & 0 & 0.4899 \\ \hline 0.5519 & 0 & 1 & 0 \\ 0 & 0.4899 & 0 & 1 \end{array}\right) \end{aligned}$$

The above result shows that $Corr(U_k, V_k) = \rho_k$ and

$$\begin{aligned} Var(U_k) &= Var(V_k) = 1 \\ Cov(U_k, U_l) &= Corr(U_k, U_l) = 0 \quad k \neq l \\ Cov(V_k, V_l) &= Corr(V_k, V_l) = 0 \quad k \neq l \\ Cov(U_k, V_l) &= Corr(U_k, V_l) = 0 \quad k \neq l \end{aligned}$$

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

$$\boldsymbol{R} = \left(\begin{array}{c|c} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \hline \boldsymbol{R}_{21} & \boldsymbol{R}_{22} \end{array}\right) = \left(\begin{array}{rr|rr} 1.000 & 0.615 & -0.111 & -0.266 \\ 0.615 & 1.000 & -0.195 & -0.085 \\ \hline -0.111 & -0.195 & 1.000 & -0.269 \\ -0.266 & -0.085 & -0.269 & 1.000 \end{array}\right)$$

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.

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

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

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

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

and

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

The eigenvalues are $(\rho_1^2, \rho_2^2) = (0.1067, 0.0293)$ with the corresponding eigenvectors $H = (\boldsymbol{h}_1, \boldsymbol{h}_2)$ and $\boldsymbol{Q} = (\boldsymbol{f}_1, \boldsymbol{f}_2)$, respectively. Here

$$\boldsymbol{h}_1 = \begin{pmatrix} -0.9463 \\ -0.3232 \end{pmatrix}, \ \boldsymbol{h}_2 = \begin{pmatrix} 0.3232 \\ -0.9463 \end{pmatrix}, \ \boldsymbol{f}_1 = \begin{pmatrix} 0.4634 \\ 0.8861 \end{pmatrix}, \ \text{and } \boldsymbol{f}_2 = \begin{pmatrix} -0.8861 \\ 0.4634 \end{pmatrix}.$$

So the canonical correlations $(\rho_1, \rho_2) = (0.3266, 0.1711)$.

(b) The first canonical variate pairs:

$$\hat{U}_1 = \boldsymbol{h}_1^T \boldsymbol{\Sigma}_{11}^{-\frac{1}{2}} \boldsymbol{Z}^{(1)} = -1.0016 Z_1^{(1)} + 0.0026 Z_2^{(1)} \approx -Z_1^{(1)}$$
$$\hat{V}_1 = \boldsymbol{f}_1^T \boldsymbol{\Sigma}_{22}^{-\frac{1}{2}} \boldsymbol{Z}^{(2)} = 0.6016 Z_1^{(2)} + 0.9769 Z_2^{(2)} \approx \frac{3}{5} Z_1^{(2)} + Z_2^{(2)}$$

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 $\frac{3}{5} Z_1^{(2)} + Z_2^{(2)}$, 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

$$\begin{aligned} z_1^{(1)} &= \text{kernel texture} \\ z_2^{(1)} &= \text{test weight} \\ z_3^{(1)} &= \text{damaged kernels} \\ z_4^{(1)} &= \text{foreign material} \\ z_5^{(1)} &= \text{crude protein in the wheat} \end{aligned}$$

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

$$\begin{aligned} z_1^{(2)} &= \text{wheat per barrel of flour} \\ z_2^{(2)} &= \text{ash in flour} \\ z_3^{(2)} &= \text{crude protein in flour} \\ z_4^{(2)} &= \text{gluten quality index} \end{aligned}$$

The sample correlation matrix was

$$\boldsymbol{R} = \left(\begin{array}{c|c} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \hline \boldsymbol{R}_{21} & \boldsymbol{R}_{22} \end{array}\right)$$
$$= \left(\begin{array}{rrrrr|rrrr} 1.000 & & & & & & & & \\ 0.754 & 1.000 & & & & & & & \\ -0.690 & -0.712 & 1.000 & & & & & & \\ -0.446 & -0.515 & 0.323 & 1.000 & & & & & \\ 0.692 & 0.412 & -0.444 & -0.334 & 1.000 & & & & \\ \hline -0.605 & -0.772 & 0.737 & 0.527 & -0.383 & 1.000 & & & \\ -0.479 & -0.419 & 0.361 & 0.461 & -0.505 & 0.251 & 1.000 & & \\ 0.780 & 0.542 & -0.546 & -0.393 & 0.737 & -0.490 & -0.434 & 1.000 & \\ -0.152 & -0.102 & 0.172 & -0.019 & -0.148 & 0.250 & -0.079 & -0.163 & 1.000 \end{array}\right)$$

(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 $\boldsymbol{Z}^{(1)}$ is explained by the canonical variate $\hat{U}_1$? What proportion of the total sample variance of the $\boldsymbol{Z}^{(2)}$ 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:

Null hypothesis	Test Statisitc	Df	χ^2-value	Conclusion
$H_0: \boldsymbol{\Sigma}_{12} = \rho_{12} = 0$	329.6947	20	37.5662	Reject H_0
$H_0: \rho_1 \neq 0, \rho_2 = \rho_3 = \rho_4 = 0$	88.8550	12	26.2170	Reject H_0
$H_0: \rho_1 \neq 0, \rho_2 \neq 0, \rho_3 = \rho_4 = 0$	10.0012	6	16.8119	Do not reject H_0

$$\hat{U}_1 = -0.1176 z_1^{(1)} - 0.3004 z_2^{(1)} + 0.3160 z_3^{(1)} + 0.2509 z_4^{(1)} - 0.2937 z_5^{(1)}$$
$$\hat{V}_1 = 0.5930 z_1^{(2)} + 0.2856 z_2^{(2)} - 0.4017 z_3^{(2)} - 0.0366 z_4^{(2)}$$
$$\hat{U}_2 = -1.0204 z_1^{(1)} + 0.7809 z_2^{(1)} - 0.5102 z_3^{(1)} - 0.2463 z_4^{(1)} - 0.5000 z_5^{(1)}$$
$$\hat{V}_2 = 0.9809 z_1^{(2)} - 0.0020 z_2^{(2)} + 0.9956 z_3^{(2)} - 0.1821 z_4^{(2)}$$

(b) $\hat{U}_1$ appears to measure quality of wheat as a contrast between negative aspects $z_1^{(1)}$, $z_2^{(1)}$, and $z_5^{(1)}$ and positive aspects $z_3^{(1)}$ and $z_4^{(1)}$. $\hat{V}_1$ appears to measure the quality of the flour as represented by $z_1^{(2)}$, $z_2^{(2)}$, and $z_3^{(2)}$.

(c) We find that the proportion of the total sample variance of the first set $\boldsymbol{Z}^{(1)}$ explained by the canonical variate $\hat{U}_1$ is $\rho_1^{(1)} = \frac{1}{p}(\boldsymbol{a}_z^{(1)})^T \boldsymbol{a}_z^{(1)} = 0.6297$ and the proportion of the total sample variance of the $\boldsymbol{Z}^{(2)}$ set explained by the canonical variate $\hat{V}_1$ is $\rho_1^{(2)} = \frac{1}{q}(\boldsymbol{b}_z^{(1)})^T \boldsymbol{b}_z^{(1)} = 0.4453$.

Appendix

R code for Problem 8.10.

```
> 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))
```

R code for Problem 8.28.

```
> farm <- read.table('./T8-7.DAT', col.names = c("Family", "DistRD", "Cotton",
+                     "Maize", "Sorg", "Millet", "Bull", "Cattle", "Goats"))
>
> # (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")
>
> # porportion 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.

```
> 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
```

```
>
> # (c)
> EU <- t(H)%*%S11invsq%*%mu1
> EV <- t(Q)%*%S22invsq%*%mu2
>
> SUU <- t(H)%*% S11invsq %*% S11 %*% S11invsq %*% H
> SUV <- t(H)%*% S11invsq %*% S12 %*% S22invsq %*% Q
> SVU <- t(Q)%*% S22invsq %*% S21 %*% S11invsq %*% H
> SVV <- t(Q)%*% S22invsq %*% S22 %*% S22invsq %*% Q
```

R code for Problem 10.10.

```
> R11 <- matrix(c(1,0.615,0.615,1), nrow=2, ncol=2)
> R12 <- matrix(c(-0.111,-0.195,-0.266,-0.085), nrow=2, ncol=2)
> R21 <- t(R12)
> R22 <- matrix(c(1,-0.269,-0.269,1), nrow=2, ncol=2)
>
> # (a)
> eig11 <- eigen(R11)
> R11inv <- solve(R11)
> R11invsq <- eig11$vectors %*% diag(sqrt(eig11$values)^(-1)) %*% t(eig11$vectors)
>
> eig22 <- eigen(R22)
> R22inv <- solve(R22)
> R22invsq <- eig22$vectors %*% diag(sqrt(eig22$values)^(-1)) %*% t(eig22$vectors)
>
> rho <- sqrt(eigen(R11invsq %*% R12 %*% R22inv %*% R21 %*% R11invsq)$values)
>
> # (b)
> H <- eigen(R11invsq %*% R12 %*% R22inv %*% R21 %*% R11invsq)$vectors
> Q <- eigen(R22invsq %*% R21 %*% R11inv %*% R12 %*% R22invsq)$vectors
>
> t(H[,1]) %*% R11invsq
> t(Q[,1]) %*% R22invsq
```

R code for Problem 10.13.

```
> R11 <- 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), nrow=5, ncol=5, byrow=TRUE)
> R21 <- 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), nrow=4, ncol=5, byrow=TRUE)
> R12 <- t(R21)
> R22 <- 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), nrow=4, ncol=4, byrow=TRUE)
>
```

```
> n <- 138
> p <- 5
> q <- 4
> d <- min(p,q)
> alpha <- 0.01
>
> eig11 <- eigen(R11)
> R11inv <- solve(R11)
> R11invsq <- eig11$vectors %*% diag(sqrt(eig11$values)^(-1)) %*% t(eig11$vectors)
>
> eig22 <- eigen(R22)
> R22inv <- solve(R22)
> R22invsq <- eig22$vectors %*% diag(sqrt(eig22$values)^(-1)) %*% t(eig22$vectors)
>
> rho2 <- eigen(R11invsq %*% R12 %*% R22inv %*% R21 %*% R11invsq)$values
> rho2[p] <- 0
> rho <- sqrt(rho2)
>
> H <- eigen(R11invsq %*% R12 %*% R22inv %*% R21 %*% R11invsq)$vectors
> Q <- eigen(R22invsq %*% R21 %*% R11inv %*% R12 %*% R22invsq)$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-(i-1))*(q-(i-1))
+   pval <- qchisq(1-alpha,df=dfs)
+   print(c(TS, dfs, pval))
+ }
>
> U <- t(H[,1:2]) %*% R11invsq
> V <- t(Q[,1:2]) %*% R22invsq
>
> # (c)
> Azinv <- solve(t(H) %*% R11invsq)
> Bzinv <- solve(t(Q) %*% R22invsq)
>
> rho1exp <- crossprod(Azinv[,1])/p
> rho2exp <- crossprod(Bzinv[,1])/q
```