Exercises

8.1. Determine the population principal components Y_1 and Y_2 for the covariance matrix

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

Also, calculate the proportion of the total population variance explained by the first principal component.

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}.

8.3. Let

$$\boldsymbol{\Sigma} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

Determine the principal components Y_1, Y_2, and Y_3. What can you say about the eigenvectors (and principal components) associated with eigenvalues that are not distinct?

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

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

8.5. (a) Find the eigenvalues of the correlation matrix

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

Are your results consistent with (8-16) and (8-17)?

(b) Verify the eigenvalue–eigenvector pairs for the $p \times p$ matrix $\boldsymbol{\rho}$ given in (8-15).

8.6. Data on x_1 = sales and x_2 = profits for the 10 largest companies in the world were listed in Exercise 1.4 of Chapter 1.
From Example 4.12

$$\bar{\mathbf{x}} = \begin{bmatrix} 155.60 \\ 14.70 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 7476.45 & 303.62 \\ 303.62 & 26.19 \end{bmatrix}$$

(a) Determine the sample principal components and their variances for these data. (You may need the quadratic formula to solve for the eigenvalues of $\mathbf{S}$.)

(b) Find the proportion of the total sample variance explained by $\hat{y}_1$.

(c) Sketch the constant density ellipse $(\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) = 1.4$, and indicate the principal components $\hat{y}_1$ and $\hat{y}_2$ on your graph.

(d) Compute the correlation coefficients $r_{\hat{y}_1,x_k}$, $k = 1, 2$. What interpretation, if any, can you give to the first principal component?

8.7. Convert the covariance matrix $\mathbf{S}$ in Exercise 8.6 to a sample correlation matrix $\mathbf{R}$.

(a) Find the sample principal components $\hat{y}_1$, $\hat{y}_2$ and their variances.

(b) Compute the proportion of the total sample variance explained by $\hat{y}_1$.

(c) Compute the correlation coefficients $r_{\hat{y}_1,z_k}$, $k = 1, 2$. Interpret $\hat{y}_1$.

(d) Compare the components obtained in Part a with those obtained in Exercise 8.6(a). Given the original data displayed in Exercise 1.4, do you feel that it is better to determine principal components from the sample covariance matrix or sample correlation matrix? Explain.

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.

8.9. *(A test that all variables are independent.)*

(a) Consider that the normal theory likelihood ratio test of H_0: $\boldsymbol{\Sigma}$ is the diagonal matrix

$$\begin{bmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{bmatrix}, \quad \sigma_{ii} > 0$$

Show that the test is as follows: Reject H_0 if

$$\Lambda = \frac{|\mathbf{S}|^{n/2}}{\prod_{i=1}^{p} s_{ii}^{n/2}} = |\mathbf{R}|^{n/2} < c$$

For a large sample size, $-2 \ln \Lambda$ is approximately $\chi^2_{p(p-1)/2}$. Bartlett [3] suggests that the test statistic $-2[1 - (2p + 11)/6n] \ln \Lambda$ be used in place of $-2 \ln \Lambda$. This results in an improved chi-square approximation. The large sample α critical point is $\chi^2_{p(p-1)/2}(\alpha)$. Note that testing $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$ is the same as testing $\boldsymbol{\rho} = \mathbf{I}$.

(b) Show that the likelihood ratio test of H_0: $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ rejects H_0 if

$$\Lambda = \frac{|\mathbf{S}|^{n/2}}{(\text{tr}(\mathbf{S})/p)^{np/2}} = \left[\frac{\prod_{i=1}^{p} \hat{\lambda}_i}{\left(\frac{1}{p} \sum_{i=1}^{p} \hat{\lambda}_i \right)^p} \right]^{n/2} = \left[\frac{\text{geometric mean } \hat{\lambda}_i}{\text{arithmetic mean } \hat{\lambda}_i} \right]^{np/2} < c$$

For a large sample size, Bartlett [3] suggests that

$$-2[1 - (2p^2 + p + 2)/6pn] \ln \Lambda$$

is approximately $\chi^2_{(p+2)(p-1)/2}$. Thus, the large sample α critical point is $\chi^2_{(p+2)(p-1)/2}(\alpha)$. This test is called a *sphericity test*, because the constant density contours are spheres when $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$.

Hint:

(a) $\max_{\mu,\Sigma} L(\mu, \Sigma)$ is given by (5-10), and $\max L(\mu, \Sigma_0)$ is the product of the univariate likelihoods, $\max_{\mu_i \sigma_{ii}} (2\pi)^{-n/2} \sigma_{ii}^{-n/2} \exp\left[-\sum_{j=1}^{n} (x_{ji} - \mu_i)^2 / 2\sigma_{ii}\right]$. Hence $\hat{\mu}_i = n^{-1} \sum_{j=1}^{n} x_{ji}$ and $\hat{\sigma}_{ii} = (1/n) \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2$. The divisor n cancels in Λ, so **S** may be used.

(b) Verify $\hat{\sigma}^2 = \left[\sum_{j=1}^{n} (x_{j1} - \bar{x}_1)^2 + \cdots + \sum_{j=1}^{n} (x_{jp} - \bar{x}_p)^2\right] \Big/ np$ under H_0. Again, the divisors n cancel in the statistic, so **S** may be used. Use Result 5.2 to calculate the chi-square degrees of freedom.

The following exercises require the use of a computer.

8.10. The weekly rates of return for five stocks listed on the New York Stock Exchange are given in Table 8.4. (See the stock-price data on the following website: www.prenhall.com/statistics.)

(a) Construct the sample covariance matrix **S**, and find the sample principal components in (8-20). (Note that the sample mean vector $\bar{\mathbf{x}}$ is displayed in Example 8.5.)

(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.

Table 8.4 Stock-Price Data (Weekly Rate Of Return)

Week	J P Morgan	Citibank	Wells Fargo	Royal Dutch Shell	Exxon Mobil
1	0.01303	−0.00784	−0.00319	−0.04477	0.00522
2	0.00849	0.01669	−0.00621	0.01196	0.01349
3	−0.01792	−0.00864	0.01004	0	−0.00614
4	0.02156	−0.00349	0.01744	−0.02859	−0.00695
5	0.01082	0.00372	−0.01013	0.02919	0.04098
6	0.01017	−0.01220	−0.00838	0.01371	0.00299
7	0.01113	0.02800	0.00807	0.03054	0.00323
8	0.04848	−0.00515	0.01825	0.00633	0.00768
9	−0.03449	−0.01380	−0.00805	−0.02990	−0.01081
10	−0.00466	0.02099	−0.00608	−0.02039	−0.01267
⋮	⋮	⋮	⋮	⋮	⋮
94	0.03732	0.03593	0.02528	0.05819	0.01697
95	0.02380	0.00311	−0.00688	0.01225	0.02817
96	0.02568	0.05253	0.04070	−0.03166	−0.01885
97	−0.00606	0.00863	0.00584	0.04456	0.03059
98	0.02174	0.02296	0.02920	0.00844	0.03193
99	0.00337	−0.01531	−0.02382	−0.00167	−0.01723
100	0.00336	0.00290	−0.00305	−0.00122	−0.00970
101	0.01701	0.00951	0.01820	−0.01618	−0.00756
102	0.01039	−0.00266	0.00443	−0.00248	−0.01645
103	−0.01279	−0.01437	−0.01874	−0.00498	−0.01637

8.11. Consider the census-tract data listed in Table 8.5. Suppose the observations on X_5 = median value home were recorded in ten thousands, rather than hundred thousands, of dollars; that is, multiply all the numbers listed in the sixth column of the table by 10.

(a) Construct the sample covariance matrix **S** for the census-tract data when X_5 = median value home is recorded in ten thousands of dollars. (Note that this covariance matrix can be obtained from the covariance matrix given in Example 8.3 by multiplying the off-diagonal elements in the fifth column and row by 10 and the diagonal element s_{55} by 100. Why?)

(b) Obtain the eigenvalue–eigenvector pairs and the first two sample principal components for the covariance matrix in Part a.

(c) Compute the proportion of total variance explained by the first two principal components obtained in Part b. Calculate the correlation coefficients, r_{y_i, x_k}, and interpret these components if possible. Compare your results with the results in Example 8.3. What can you say about the effects of this change in scale on the principal components?

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?

Table 8.5 Census-tract Data

Tract	Total population (thousands)	Professional degree (percent)	Employed age over 16 (percent)	Government employment (percent)	Median home value ($100,000)
1	2.67	5.71	69.02	30.3	1.48
2	2.25	4.37	72.98	43.3	1.44
3	3.12	10.27	64.94	32.0	2.11
4	5.14	7.44	71.29	24.5	1.85
5	5.54	9.25	74.94	31.0	2.23
6	5.04	4.84	53.61	48.2	1.60
7	3.14	4.82	67.00	37.6	1.52
8	2.43	2.40	67.20	36.8	1.40
9	5.38	4.30	83.03	19.7	2.07
10	7.34	2.73	72.60	24.5	1.42
	⋮	⋮	⋮	⋮	⋮
52	7.25	1.16	78.52	23.6	1.50
53	5.44	2.93	73.59	22.3	1.65
54	5.83	4.47	77.33	26.2	2.16
55	3.74	2.26	79.70	20.2	1.58
56	9.21	2.36	74.58	21.8	1.72
57	2.14	6.30	86.54	17.4	2.80
58	6.62	4.79	78.84	20.0	2.33
59	4.24	5.82	71.39	27.1	1.69
60	4.72	4.71	78.01	20.6	1.55
61	6.48	4.93	74.23	20.9	1.98

Note: Observations from adjacent census tracts are likely to be correlated. That is, these 61 observations may not constitute a random sample. Complete data set available at www.prenhall.com/statistics.

8.13. In the radiotherapy data listed in Table 1.7 (see also the radiotherapy data on the website www.prenhall.com/statistics), the $n = 98$ observations on $p = 6$ variables represent patients' reactions to radiotherapy.

(a) Obtain the covariance and correlation matrices **S** and **R** for these data.

(b) Pick one of the matrices **S** or **R** (justify your choice), and determine the eigenvalues and eigenvectors. Prepare a table showing, in decreasing order of size, the percent that each eigenvalue contributes to the total sample variance.

(c) Given the results in Part b, decide on the number of important sample principal components. Is it possible to summarize the radiotherapy data with a single reaction-index component? Explain.

(d) Prepare a table of the correlation coefficients between each principal component you decide to retain and the original variables. If possible, interpret the components.

8.14. Perform a principal component analysis using the sample covariance matrix of the sweat data given in Example 5.2. Construct a Q–Q plot for each of the important principal components. Are there any suspect observations? Explain.

8.15. The four sample standard deviations for the postbirth weights discussed in Example 8.6 are

$$\sqrt{s_{11}} = 32.9909, \quad \sqrt{s_{22}} = 33.5918, \quad \sqrt{s_{33}} = 36.5534, \quad \text{and} \quad \sqrt{s_{44}} = 37.3517$$

Use these and the correlations given in Example 8.6 to construct the sample covariance matrix **S**. Perform a principal component analysis using **S**.

8.16. Over a period of five years in the 1990s, yearly samples of fishermen on 28 lakes in Wisconsin were asked to report the time they spent fishing and how many of each type of game fish they caught. Their responses were then converted to a catch rate per hour for

x_1 = Bluegill $\quad x_2$ = Black crappie $\quad x_3$ = Smallmouth bass

x_4 = Largemouth bass $\quad x_5$ = Walleye $\quad x_6$ = Northern pike

The estimated correlation matrix (courtesy of Jodi Barnet)

$$\mathbf{R} = \begin{bmatrix} 1 & .4919 & .2636 & .4653 & -.2277 & .0652 \\ .4919 & 1 & .3127 & .3506 & -.1917 & .2045 \\ .2635 & .3127 & 1 & .4108 & .0647 & .2493 \\ .4653 & .3506 & .4108 & 1 & -.2249 & .2293 \\ -.2277 & -.1917 & .0647 & -.2249 & 1 & -.2144 \\ .0652 & .2045 & .2493 & .2293 & -.2144 & 1 \end{bmatrix}$$

is based on a sample of about 120. (There were a few missing values.)

Fish caught by the same fisherman live alongside of each other, so the data should provide some evidence on how the fish group. The first four fish belong to the centrarchids, the most plentiful family. The walleye is the most popular fish to eat.

(a) Comment on the pattern of correlation within the centrarchid family x_1 through x_4. Does the walleye appear to group with the other fish?

(b) Perform a principal component analysis using only x_1 through x_4. Interpret your results.

(c) Perform a principal component analysis using all six variables. Interpret your results.

8.17. Using the data on bone mineral content in Table 1.8, perform a principal component analysis of **S**.

8.18. The data on national track records for women are listed in Table 1.9.

(a) Obtain the sample correlation matrix **R** for these data, and determine its eigenvalues and eigenvectors.

(b) Determine the first two principal components for the standardized variables. Prepare a table showing the correlations of the standardized variables with the components, and the cumulative percentage of the total (standardized) sample variance explained by the two components.

(c) Interpret the two principal components obtained in Part b. (Note that the first component is essentially a normalized unit vector and might measure the athletic excellence of a given nation. The second component might measure the relative strength of a nation at the various running distances.)

(d) Rank the nations based on their score on the first principal component. Does this ranking correspond with your intuitive notion of athletic excellence for the various countries?

8.19. Refer to Exercise 8.18. Convert the national track records for women in Table 1.9 to speeds measured in meters per second. Notice that the records for 800 m, 1500 m, 3000 m, and the marathon are given in minutes. The marathon is 26.2 miles, or 42,195 meters, long. Perform a principal components analysis using the covariance matrix **S** of the speed data. Compare the results with the results in Exercise 8.18. Do your interpretations of the components differ? If the nations are ranked on the basis of their score on the first principal component, does the subsequent ranking differ from that in Exercise 8.18? Which analysis do you prefer? Why?

8.20. The data on national track records for men are listed in Table 8.6. (See also the data on national track records for men on the website www.prenhall.com/statistics) Repeat the principal component analysis outlined in Exercise 8.18 for the men. Are the results consistent with those obtained from the women's data?

8.21. Refer to Exercise 8.20. Convert the national track records for men in Table 8.6 to speeds measured in meters per second. Notice that the records for 800 m, 1500 m, 5000 m, 10,000 m and the marathon are given in minutes. The marathon is 26.2 miles, or 42,195 meters, long. Perform a principal component analysis using the covariance matrix **S** of the speed data. Compare the results with the results in Exercise 8.20. Which analysis do you prefer? Why?

8.22. Consider the data on bulls in Table 1.10. Utilizing the seven variables YrHgt, FtFrBody, PrctFFB, Frame, BkFat, SaleHt, and SaleWt, perform a principal component analysis using the covariance matrix **S** and the correlation matrix **R**. Your analysis should include the following:

(a) Determine the appropriate number of components to effectively summarize the sample variability. Construct a scree plot to aid your determination.

(b) Interpret the sample principal components.

(c) Do you think it is possible to develop a "body size" or "body configuration" index from the data on the seven variables above? Explain.

(d) Using the values for the first two principal components, plot the data in a two-dimensional space with $\hat{y}_1$ along the vertical axis and $\hat{y}_2$ along the horizontal axis. Can you distinguish groups representing the three breeds of cattle? Are there any outliers?

(e) Construct a Q–Q plot using the first principal component. Interpret the plot.

Table 8.6 National Track Records for Men

Country	100 m (s)	200 m (s)	400 m (s)	800 m (min)	1500 m (min)	5000 m (min)	10,000 m (min)	Marathon (min)
Argentina	10.23	20.37	46.18	1.77	3.68	13.33	27.65	129.57
Australia	9.93	20.06	44.38	1.74	3.53	12.93	27.53	127.51
Austria	10.15	20.45	45.80	1.77	3.58	13.26	27.72	132.22
Belgium	10.14	20.19	45.02	1.73	3.57	12.83	26.87	127.20
Bermuda	10.27	20.30	45.26	1.79	3.70	14.64	30.49	146.37
Brazil	10.00	19.89	44.29	1.70	3.57	13.48	28.13	126.05
Canada	9.84	20.17	44.72	1.75	3.53	13.23	27.60	130.09
Chile	10.10	20.15	45.92	1.76	3.65	13.39	28.09	132.19
China	10.17	20.42	45.25	1.77	3.61	13.42	28.17	129.18
Columbia	10.29	20.85	45.84	1.80	3.72	13.49	27.88	131.17
Cook Islands	10.97	22.46	51.40	1.94	4.24	16.70	35.38	171.26
Costa Rica	10.32	20.96	46.42	1.87	3.84	13.75	28.81	133.23
Czech Republic	10.24	20.61	45.77	1.75	3.58	13.42	27.80	131.57
Denmark	10.29	20.52	45.89	1.69	3.52	13.42	27.91	129.43
DominicanRepublic	10.16	20.65	44.90	1.81	3.73	14.31	30.43	146.00
Finland	10.21	20.47	45.49	1.74	3.61	13.27	27.52	131.15
France	10.02	20.16	44.64	1.72	3.48	12.98	27.38	126.36
Germany	10.06	20.23	44.33	1.73	3.53	12.91	27.36	128.47
Great Britain	9.87	19.94	44.36	1.70	3.49	13.01	27.30	127.13
Greece	10.11	19.85	45.57	1.75	3.61	13.48	28.12	132.04
Guatemala	10.32	21.09	48.44	1.82	3.74	13.98	29.34	132.53
Hungary	10.08	20.11	45.43	1.76	3.59	13.45	28.03	132.10
India	10.33	20.73	45.48	1.76	3.63	13.50	28.81	132.00
Indonesia	10.20	20.93	46.37	1.83	3.77	14.21	29.65	139.18
Ireland	10.35	20.54	45.58	1.75	3.56	13.07	27.78	129.15
Israel	10.20	20.89	46.59	1.80	3.70	13.66	28.72	134.21
Italy	10.01	19.72	45.26	1.73	3.35	13.09	27.28	127.29
Japan	10.00	20.03	44.78	1.77	3.62	13.22	27.58	126.16
Kenya	10.28	20.43	44.18	1.70	3.44	12.66	26.46	124.55
Korea, South	10.34	20.41	45.37	1.74	3.64	13.84	28.51	127.20
Korea, North	10.60	21.23	46.95	1.82	3.77	13.90	28.45	129.26
Luxembourg	10.41	20.77	47.90	1.76	3.67	13.64	28.77	134.03
Malaysia	10.30	20.92	46.41	1.79	3.76	14.11	29.50	149.27
Mauritius	10.13	20.06	44.69	1.80	3.83	14.15	29.84	143.07
Mexico	10.21	20.40	44.31	1.78	3.63	13.13	27.14	127.19
Myanmar(Burma)	10.64	21.52	48.63	1.80	3.80	14.19	29.62	139.57
Netherlands	10.19	20.19	45.68	1.73	3.55	13.22	27.44	128.31
New Zealand	10.11	20.42	46.09	1.74	3.54	13.21	27.70	128.59
Norway	10.08	20.17	46.11	1.71	3.62	13.11	27.54	130.17
Papua New Guinea	10.40	21.18	46.77	1.80	4.00	14.72	31.36	148.13
Philippines	10.57	21.43	45.57	1.80	3.82	13.97	29.04	138.44
Poland	10.00	19.98	44.62	1.72	3.59	13.29	27.89	129.23
Portugal	9.86	20.12	46.11	1.75	3.50	13.05	27.21	126.36
Romania	10.21	20.75	45.77	1.76	3.57	13.25	27.67	132.30
Russia	10.11	20.23	44.60	1.71	3.54	13.20	27.90	129.16
Samoa	10.78	21.86	49.98	1.94	4.01	16.28	34.71	161.50
Singapore	10.37	21.14	47.60	1.84	3.86	14.96	31.32	144.22
Spain	10.17	20.59	44.96	1.73	3.48	13.04	27.24	127.23
Sweden	10.18	20.43	45.54	1.76	3.61	13.29	27.93	130.38
Switzerland	10.16	20.41	44.99	1.71	3.53	13.13	27.90	129.56
Taiwan	10.36	20.81	46.72	1.79	3.77	13.91	29.20	134.35
Thailand	10.23	20.69	46.05	1.81	3.77	14.25	29.67	139.33
Turkey	10.38	21.04	46.63	1.78	3.59	13.45	28.33	130.25
U.S.A.	9.78	19.32	43.18	1.71	3.46	12.97	27.23	125.38

Source: *IAAF/ATES Track and Field Statistics Handbook* for the Helsinki 2005 Olympics. Courtesy of Ottavio Castellini.

8.23. A naturalist for the Alaska Fish and Game Department studies grizzly bears with the goal of maintaining a healthy population. Measurements on $n = 61$ bears provided the following summary statistics:

Variable	Weight (kg)	Body length (cm)	Neck (cm)	Girth (cm)	Head length (cm)	Head width (cm)
Sample mean $\bar{x}$	95.52	164.38	55.69	93.39	17.98	31.13

Covariance matrix

$$\mathbf{S} = \begin{bmatrix} 3266.46 & 1343.97 & 731.54 & 1175.50 & 162.68 & 238.37 \\ 1343.97 & 721.91 & 324.25 & 537.35 & 80.17 & 117.73 \\ 731.54 & 324.25 & 179.28 & 281.17 & 39.15 & 56.80 \\ 1175.50 & 537.35 & 281.17 & 474.98 & 63.73 & 94.85 \\ 162.68 & 80.17 & 39.15 & 63.73 & 9.95 & 13.88 \\ 238.37 & 117.73 & 56.80 & 94.85 & 13.88 & 21.26 \end{bmatrix}$$

(a) Perform a principal component analysis using the covariance matrix. Can the data be effectively summarized in fewer than six dimensions?

(b) Perform a principal component analysis using the correlation matrix.

(c) Comment on the similarities and differences between the two analyses.

8.24. Refer to Example 8.10 and the data in Table 5.8, page 240. Add the variable x_6 = regular overtime hours whose values are (read across)

6187	7336	6988	6964	8425	6778	5922	7307
7679	8259	10954	9353	6291	4969	4825	6019

and redo Example 8.10.

8.25. Refer to the police overtime hours data in Example 8.10. Construct an alternate control chart, based on the sum of squares d_{Uj}^2, to monitor the unexplained variation in the original observations summarized by the additional principal components.

8.26. Consider the psychological profile data in Table 4.6. Using the five variables, Indep, Supp, Benev, Conform and Leader, performs a principal component analysis using the covariance matrix **S** and the correlation matrix **R**. Your analysis should include the following:

(a) Determine the appropriate number of components to effectively summarize the variability. Construct a scree plot to aid in your determination.

(b) Interpret the sample principal components.

(c) Using the values for the first two principal components, plot the data in a two-dimensional space with $\hat{y}_1$ along the vertical axis and $\hat{y}_2$ along the horizontal axis. Can you distinguish groups representing the two socioeconomic levels and/or the two genders? Are there any outliers?

(d) Construct a 95% confidence interval for λ_1, the variance of the first population principal component from the covariance matrix.

8.27. The pulp and paper properties data is given in Table 7.7. Using the four paper variables, BL (breaking length), EM (elastic modulus), SF (Stress at failure) and BS (burst strength), perform a principal component analysis using the covariance matrix **S** and the correlation matrix **R**. Your analysis should include the following:

(a) Determine the appropriate number of components to effectively summarize the variability. Construct a scree plot to aid in your determination.

(b) Interpret the sample principal components.

(c) Do you think it it is possible to develop a "paper strength" index that effectively contains the information in the four paper variables? Explain.

(d) Using the values for the first two principal components, plot the data in a two-dimensional space with $\hat{y}_1$ along the vertical axis and $\hat{y}_2$ along the horizontal axis. Identify any outliers in this data set.

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 and on the website www.prenhall.com/statistics

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

Table 8.7 Mali Family Farm Data

Family	DistRD	Cotton	Maize	Sorg	Millet	Bull	Cattle	Goats
12	80	1.5	1.00	3.0	.25	2	0	1
54	8	6.0	4.00	0	1.00	6	32	5
11	13	.5	1.00	0	0	0	0	0
21	13	2.0	2.50	1.0	0	1	0	5
61	30	3.0	5.00	0	0	4	21	0
20	70	0	2.00	3.0	0	2	0	3
29	35	1.5	2.00	0	0	0	0	0
29	35	2.0	3.00	2.0	0	0	0	0
57	9	5.0	5.00	0	0	4	5	2
23	33	2.0	2.00	1.0	0	2	1	7
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
20	0	1.5	1.00	3.0	0	1	6	0
27	41	1.1	.25	1.5	1.50	0	3	1
18	500	2.0	1.00	1.5	.50	1	0	0
30	19	2.0	2.00	4.0	1.00	2	0	5
77	18	8.0	4.00	6.0	4.00	6	8	6
21	500	5.0	1.00	3.0	4.00	1	0	5
13	100	.5	.50	0	1.00	0	0	4
24	100	2.0	3.00	0	.50	3	14	10
29	90	2.0	1.50	1.5	1.50	2	0	2
57	90	10.0	7.00	0	1.50	7	8	7

Source: Data courtesy of Jay Angerer.

(b) Perform a principal component analysis using the correlation matrix **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?

8.29. Refer to Exercise 5.28. Using the covariance matrix **S** for the first 30 cases of car body assembly data, obtain the sample principal components.

(a) Construct a 95% ellipse format chart using the first two principal components $\hat{y}_1$ and $\hat{y}_2$. Identify the car locations that appear to be out of control.

(b) Construct an alternative control chart, based on the sum of squares $d^2_{U\,j}$, to monitor the variation in the original observations summarized by the remaining four principal components. Interpret this chart.

References

1. Anderson, T. W. *An Introduction to Multivariate Statistical Analysis* (3rd ed.). New York: John Wiley, 2003.
2. Anderson, T. W. "Asymptotic Theory for Principal Components Analysis." *Annals of Mathematical Statistics*, **34** (1963), 122–148.
3. Bartlett, M. S. "A Note on Multiplying Factors for Various Chi-Squared Approximations." *Journal of the Royal Statistical Society (B)*, **16** (1954), 296–298.
4. Dawkins, B. "Multivariate Analysis of National Track Records." *The American Statistician*, **43** (1989), 110–115.
5. Girschick, M. A. "On the Sampling Theory of Roots of Determinantal Equations." *Annals of Mathematical Statistics*, **10** (1939), 203–224.
6. Hotelling, H. "Analysis of a Complex of Statistical Variables into Principal Components." *Journal of Educational Psychology*, **24** (1933). 417–441, 498–520.
7. Hotelling, H. "The Most Predictable Criterion." *Journal of Educational Psychology*, **26** (1935), 139–142.
8. Hotelling, H. "Simplified Calculation of Principal Components." *Psychometrika*, **1** (1936), 27–35.
9. Hotelling, H. "Relations between Two Sets of Variates." *Biometrika*, **28** (1936), 321–377.
10. Jolicoeur, P. "The Multivariate Generalization of the Allometry Equation." *Biometrics*, **19** (1963), 497–499.
11. Jolicoeur, P., and J. E. Mosimann. "Size and Shape Variation in the Painted Turtle: A Principal Component Analysis." *Growth*, **24** (1960), 339–354.
12. King, B. "Market and Industry Factors in Stock Price Behavior." *Journal of Business*, **39** (1966), 139–190.
13. Kourti, T., and J. McGregor, "Multivariate SPC Methods for Process and Product Monitoring," *Journal of Quality Technology*, **28** (1996), 409–428.
14. Lawley, D. N. "On Testing a Set of Correlation Coefficients for Equality." *Annals of Mathematical Statistics*, **34** (1963), 149–151.
15. Rao, C. R. *Linear Statistical Inference and Its Applications* (2nd ed.). New York: Wiley-Interscience, 2002.
16. Rencher, A. C. "Interpretation of Canonical Discriminant Functions, Canonical Variates and Principal Components." *The American Statistician*, **46** (1992), 217–225.