Multivariate Analysis

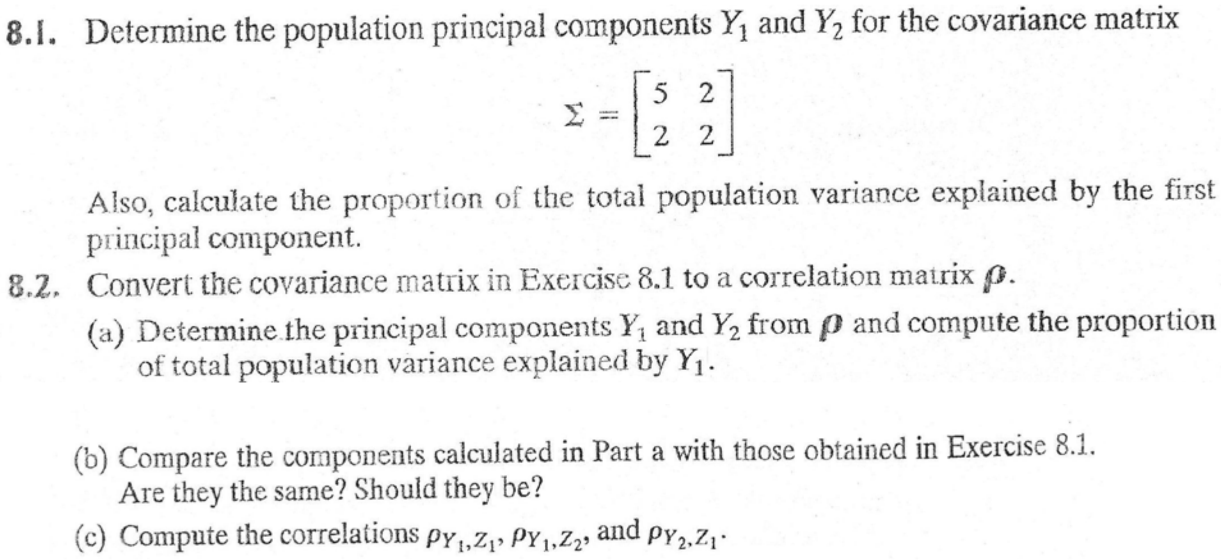
Assignment #2

Due: 4/10/2018

I. Perform the calculations in calculator or Matlab.

1. (10%) Textbook 8.1

2. (15%) Textbook 8.2



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 $\boldsymbol{\rho}$.

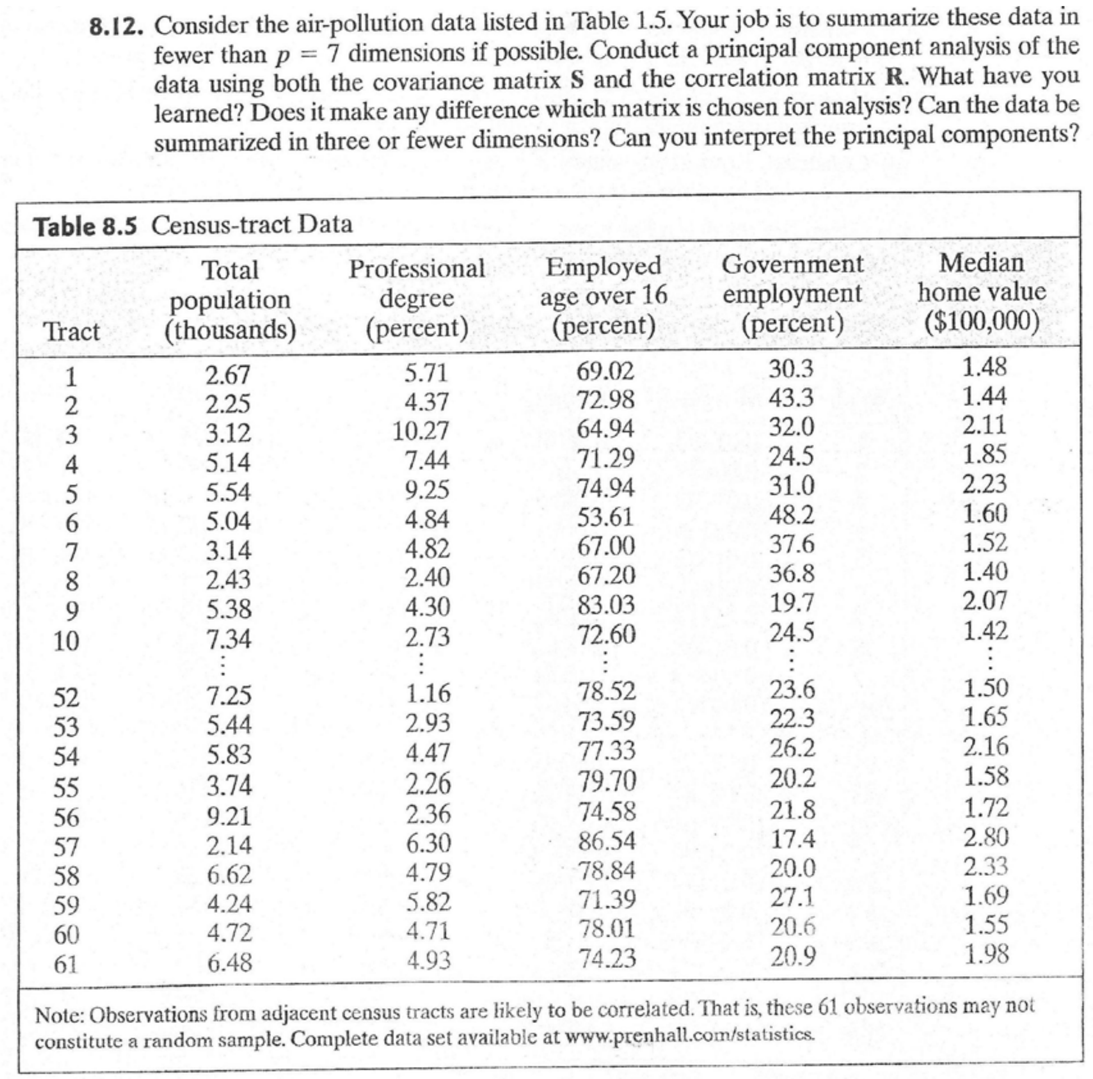
(a) Determine the principal components Y_1 and Y_2 from $\boldsymbol{\rho}$ and compute the proportion of total population variance explained by Y_1.

(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}.

II.Perform the calculations in SAS.

3. (20%) Textbook 8.12



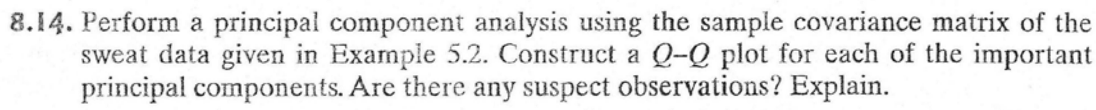
8.12. Consider the air-pollution data listed in Table 1.5. Your job is to summarize these data in fewer than $p = 7$ dimensions if possible. Conduct a principal component analysis of the data using both the covariance matrix **S** and the correlation matrix **R**. What have you learned? Does it make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

4. (10%) Textbook 8.14



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.

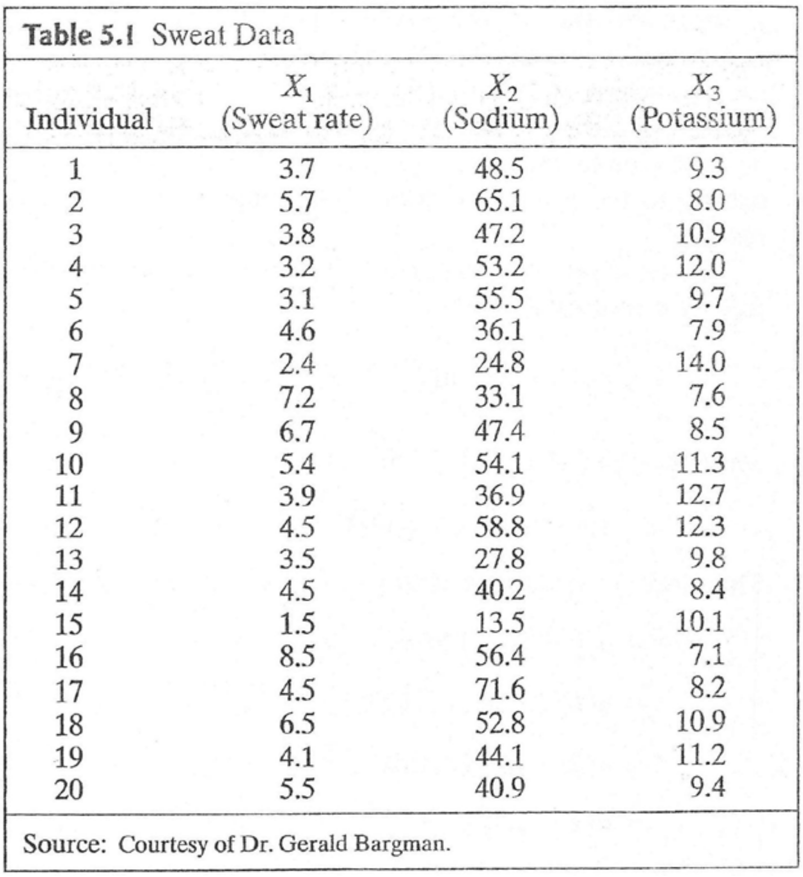


Table 5.1 Sweat Data

Individual	X_1 (Sweat rate)	X_2 (Sodium)	X_3 (Potassium)
1	3.7	48.5	9.3
2	5.7	65.1	8.0
3	3.8	47.2	10.9
4	3.2	53.2	12.0
5	3.1	55.5	9.7
6	4.6	36.1	7.9
7	2.4	24.8	14.0
8	7.2	33.1	7.6
9	6.7	47.4	8.5
10	5.4	54.1	11.3
11	3.9	36.9	12.7
12	4.5	58.8	12.3
13	3.5	27.8	9.8
14	4.5	40.2	8.4
15	1.5	13.5	10.1
16	8.5	56.4	7.1
17	4.5	71.6	8.2
18	6.5	52.8	10.9
19	4.1	44.1	11.2
20	5.5	40.9	9.4

Source: Courtesy of Dr. Gerald Bargman.

5. (30%)

File FOODP.DAT gives the average price in cents per pound of five food items in 24 U.S. cities. Use principal components analysis to analyze the data (correlation matrix) and determine the appropriate number of principal components.

(a) (10%) Show Scree plot, principal components and interpret their meanings

(b) (5%) Show the first two principal components scores

(c) (10%) Plot the first two components scores in a scatter plot. How many “clusters” can be identified visually?

(d) (5%) Which city is the least expensive? Which city is the most expensive?

6. (15%)

Following the 1973-74 Arab oil embargo and the subsequent dramatic increase in oil prices, a study was conducted in three cities to estimate the potential demand for mass transportation. Five hundred and ninety-seven (597) responded to the twenty statements on a five-point scale

(1=disagree strongly to 5=agree strongly). Use principal components analysis to analyze the data (correlation matrix) and identify the key perceptions about the energy crisis.

The data collected are given in file MASST.DAT. The twenty statements are given below:

X1 If the energy shortage gets any worse, the country will be in bad shape.
X2 The worst of the energy crisis has passed.
X3 Science and technology will be able to resolve the energy crisis without conservation.
X4 Saving energy requires you to make major sacrifices.
X5 The energy crisis is for real.
X6 Utility companies should be allowed to burn cheaper fuel even though this would cause more pollution.
X7 The petroleum companies have not done all they can to solve the energy problem.
X8 Congress has done all it can to solve the energy problem.
X9 Rationing of energy resources will be necessary for at least the next five years.
X10 Conserving electricity will save me money in the long run.
X11 The natural gas companies have done all they can to solve the energy
problem.
X12 My electricity bill would be the same no matter what I did.
X13 There is not much an average citizen can do to save electricity.
X14 President Carter has done all he can to solve the energy problem.
X15 At this point in time our traditional energy resources (coal, oil, natural gas, etc.) are insufficient to continue energy consumption at the present rate.
X16 It would be easy for me to cut down on the use of electricity in my home.
X17 We should forget about reducing pollution until our energy problems are solved.
X18 My personal conservation efforts have little impact on total consumption of energy.
X19 Because of the abundance of coal, industries should be encouraged to switch to coal as a fuel despite the air pollution it causes.

X20 The electric companies have not done all they can to solve the energy problem.

(a) (10%) Show the first five sample principal components loadings. What is the percentage of total variance they account for?
(b) (5%) Interpret the first five sample principal components.