STAT401 Homework #3
Spring 2022

Q1. Assume that data $X = (X_1, X_2)'$ follows a multivariate normal distribution with mean μ and the covariance matrix Σ given by

$$\mu = \begin{pmatrix} 2 \\ -2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

(a) Determine the population principal components Y_1 and Y_2.

(b) Calculate the proportion of the total population variance explained by the first principal component.

(c) For an observation $x = (3, -1)'$, calculate the principal component score.

(d) Convert Σ to a correlation matrix ρ. Determine the principal components Y_1 and Y_2 from ρ and calculate the proportion of total population variance explained by the first principal component.

(e) Compare the components calculated in (d) with those obtained in (a).

Q2. A study about grizzly bears was conducted to maintain a healthy population. The six variables, (1) weight (kg), (2) body length (cm), (3) neck (cm), (4) girth (cm), (5) head length (cm), and (6) head width (cm) were measured for 61 bears and the following shows summary statistics:

$$\bar{x} = \begin{pmatrix} 95.5 \\ 164.4 \\ 55.7 \\ 93.4 \\ 18.0 \\ 31.1 \end{pmatrix}, \quad S = \begin{pmatrix} 3266 & & & & & \\ 1344 & 722 & & & & \\ 732 & 324 & 179 & & & \\ 1176 & 537 & 281 & 475 & & \\ 163 & 80 & 39 & 64 & 10 & \\ 238 & 118 & 57 & 95 & 14 & 21 \end{pmatrix}.$$

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

(b) Perform a principal component analysis using the correlation matrix R. Can the data be effectively summarized in fewer than six dimension?

(c) Which analysis results would you recommend between (a) and (b)?

Q3. The weekly rates of return for five stocks listed on the New York Stock Exchange are given as follows (Data are available as an attached file, “STOCK.DAT”):

Week	JP Morgan	Citibank	Wells Fargo	Royal Dutch Shell	Exxon Mobil
1	0.0130338	-0.0078431	-0.0031889	-0.0447693	0.0052151
2	0.0084862	0.0166886	-0.0062100	0.0119560	0.0134890
⋮	⋮	⋮	⋮	⋮	⋮
103	-0.0127948	-0.0143678	-0.0187402	-0.0049759	-0.0163732

(a) Find the sample principal components using the sample correlation matrix *R*.

(b) Suggest an appropriate number of the principal components for this data set.

(c) Interpret chosen principal components in (b).

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

Q4. The radiotherapy data set discussed in Homework #1 (available as an attached file, “RADIOTHERAPY.DAT”) measures average ratings over the course of treatment for cancer patients undergoing radiotherapy. Among six variables, consider the following five variables.

X_1	Number of symptoms
X_2	Amount of activity (1–5 scale)
X_3	Amount of sleep (1–5 scale)
X_4	Amount of food consumed (1–3 scale)
X_5	Appetite (1–5 scale)

(a) Choose whether you would conduct principal component analysis using the sample variance-covariance matrix *S* or the sample correlation matrix *R*. Justify your choice.

(b) Find the sample principal components based on your choice of *S* or *R* in (a).

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

(d) For each observation, calculate the sum of $X_1 - X_5$. Compare this sum with the first principal component scores.

(e) Using the derived two principal components, produce a scatterplot of this data set. Discuss your findings.