COLORADOSCHOOLOFMINES

MATH437/537 **Fall, 2022**

Homework 5. Due November 4

1. Let $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_3$ and $\boldsymbol{v}$ be the following 14×1 vectors:

$$\boldsymbol{\mu}_1 = \mathbf{0}, \quad \boldsymbol{\mu}_2 = (3, \ldots, 3)^\top, \quad \boldsymbol{\mu}_3 = (-3, \ldots, -3)^\top, \quad \boldsymbol{v}_1 = (1, -1, 1, -1, \ldots, 1, -1)^\top/\sqrt{14},$$

$$\boldsymbol{v}_2 = (-2, 1, -2, 1, \ldots, -2, 1)^\top/6.$$

   (a) Draw a sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{32}$ where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{16}$ are iid $N_{14}(\boldsymbol{\mu}_1, \boldsymbol{I})$, $\boldsymbol{X}_{17}, \ldots, \boldsymbol{X}_{24}$ are iid $N_{14}(\boldsymbol{\mu}_2, \boldsymbol{v}_1\boldsymbol{v}_1^\top)$ and $\boldsymbol{X}_{25}, \ldots, \boldsymbol{X}_{32}$ are iid $N_{14}(\boldsymbol{\mu}_3, \boldsymbol{v}_1\boldsymbol{v}_1^\top + \boldsymbol{v}_2\boldsymbol{v}_2^\top)$. These samples give you a data matrix $\boldsymbol{X}$ of size 32×14.

   (b) Make a plot of $(\lambda_1 + \cdots + \lambda_k)/(\lambda_1 + \cdots + \lambda_{14})$ as a function of k, where λ_i are the eigenvalues of the sample covariance matrix. Interpret the results.

   (c) Make a plot of the data in $\boldsymbol{X}$ in the first two sample PCs.

   (d) Repeat (a)-(c) twice using two new sets of data of the same sample size. You will see the sample variability inherent in sample principal components.

2. Measurements of four variables were made for each of a random sample of 500 sharks. The eigenvalues of the sample covariance matrix $\boldsymbol{S}$ are 14.1, 4.3, 1.2 and 0.4. The eigenvectors corresponding to the two largest eigenvalues are

$$\boldsymbol{u}_1 = (0.39,\ 0.42,\ 0.44,\ 0.69)^\top \quad \text{and} \quad \boldsymbol{u}_2 = (0.40,\ 0.39,\ 0.42,\ -0.72)^\top.$$

   (a) What is the sample variance of the first sample principal component?

   (b) How much of the total sample variance is a accounted for by the first two sample principal components?

   (c) What are the first two sample principal components corresponding to an arbitrary data point $(x_1, \ldots, x_4)$?

   (d) What are the coordinates of $(x_1, \ldots, x_4)$ you put in a biplot of the data?

3. For this question you need the data in Table 11.1 (p.288) in Zelterman of the numbers of Canadian health care workers per 100,000 residents in 2006, listed by province or territory (it is also available in CANVAS, file name `Canmed.txt`). Before doing anything, omit the row 'Canada total'. Also, Zelterman suggests scaling the data (use function `scale` in `R`).

   (a) Draw a dendogram of the data using complete linkage (`R` comands given in homework 4).

(b) Draw a biplot. The basic commands in R are:

```
pc = princomp(scale(data))
biplot(pc, col = c(1, 2), cex = c(.75, .75),xlim = c( -.45, .45),
    xlab = "PC1",ylab = "PC2", main = "title")
```

(c) Can you conclude any useful information about the health care in Canada from the dendogram and biplot?

4. Suppose $\boldsymbol{X} \sim N_3(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1/4 & 0 \\ 1/4 & 1 & 1/4 \\ 0 & 1/4 & 1 \end{pmatrix}.$$

(a) Find the principal components Y_1, Y_2 and Y_3 of $\boldsymbol{X}$.

(b) Find the Karhunen-Loève expansion of $\boldsymbol{X}$ (explained in class).

(c) Find the best rank-one approximation of $\boldsymbol{\Sigma}$ and determine the distribution of the first term in the Karhunen-Loève expansion of $\boldsymbol{X}$ (this answers a question asked in class).