Multivariate Analysis - Homework 6

Please upload your homework on SPOC before 8:00pm, May 18, including all details needed. For R exercises, R markdown is highly encouraged; for other parts, try to use LaTex.

1. Let

$$\mathbf{\Sigma} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}$$

   Determine the principal components. What can you say about the eigenvectors and principal components associated with eigenvalues that are not distinct?

2. For two sets of variables $\mathbf{x}$ and $\mathbf{y}$, the covariance matrices are

$$\mathbf{\Sigma}_{XX} = \mathbf{\Sigma}_{YY} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \text{ and } \mathbf{\Sigma}_{XY} = \begin{pmatrix} \rho & \rho \\ \rho & \rho \end{pmatrix}.$$

   Find the canonical correlations and the canonical variates. Generalize your results to general dimension of $\mathbf{x}$ and $\mathbf{y}$.

3. In the principal comoponent method for the factor analysis, prove that

$$\text{Sum of squared entries of } \mathbf{S} - (\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}}) \leq \textstyle\sum_{k=m+1}^{p} \hat{\lambda}_k^2.$$

   (Hint: 1. sum of squared entries of $\mathbf{S} - (\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}})$ can be controlled by that of $\mathbf{S} - \hat{\mathbf{L}}\hat{\mathbf{L}}'$; 2. sum of squared entries of a matrix $\mathbf{A}$ equals to $tr(\mathbf{A}\mathbf{A}')$.)

4. (R exercise.) Generate a random sample of size $n = 100$ from a three-dimensional Gaussian (i.e. normal) distribution, where one of the variables has very high variance (relative to the other two). Carry out PCA on these data using the covariance matrix and the correlation matrix. In each case, find the eigenvalues and eigenvectors, draw the scree plot, compute the PC scores, and plot all pairwise PC scores in a matrix plot. Compare results. Code by yourself. DO NOT USE R package.

5. (R exercise.) Carry out a PCA of Fishers iris data in R. These data consist of 50 observations on each of three species of iris: Iris setosa, Iris versicolor, and Iris virginica. The four quantitative variables are sepal length, sepal width, petal length, and petal width. Compute the PC scores and plot all pairwise sets of PC scores using the four variables in a matrix plot. Explain your results, taking into consideration the species labels.

6. (R exercise.) Consider the air-pollution data (attached in a separate .dat file.)

   (a) Conduct a principal component analysis of the data using both the covariance matrix and correlation matrix. What have you learned? Give the detail of the analysis, your conclusion remarks and the interpretations.

   (b) Refer to the variables Y_1, Y_2, Y_5, Y_6. Obtain the principal component solution to a factor model with $m = 1$ and $m = 2$.

   (c) Find the maximum likelihood estimates of $\mathbf{L}$ and $\mathbf{\Psi}$ for $m = 1$ and $m = 2$.

   (d) Compare the factorization obtained by the principal component and maximum likelihood methods.

   (e) Rotate the factors using varimax methods based on the principal component method with $m = 2$. Interpret the result.

   (f) Plot the rotated factor scores associated with (d) and analyze the typical values.