Applied Multivariate Statistical Analysis: Homework 5

Homework format: all homework must be written in latex. You must turn in both your tex and pdf files. Attach your code and computer output if there is any programming.

1. Suppose that $n_1 = 10$ and $n_2 = 12$ observations are made on two random vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, where $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are both assumed to have bivariate normal distributions, denoted by D_1 and D_2, with a common covariance matrix Σ, but possibly different mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. The sample mean vectors and pooled covariance matrix are $\bar{\boldsymbol{x}}_1 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$, $\bar{\boldsymbol{x}}_2 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ and

   $$S_{\text{pooled}} = \begin{pmatrix} 7.3 & -1.1 \\ -1.1 & 4.8 \end{pmatrix}.$$

   (a) Test for the difference in population mean vectors with the significance level $\alpha = 0.1$.

   (b) Construct a (sample) linear discriminant function (also called Fisher's linear discriminant function).

   (c) Assign a new observation $\boldsymbol{x}_0' = (0, 1)$ to either population D_1 or D_2, assuming equal prior probabilities.

2. Perform a discriminant/classification analysis for the data on hemophilia carriers, hemophilia.DAT, where the 1st column contains the true group membership (1=Noncarriers, 2=Obligatory carriers), the 2nd and 3rd columns contain the measurements $\log_{10}$(AHF activity) and $\log_{10}$(AHF antigen) for each case.

   (a) Obtain the sample linear discriminant function assuming equal prior probabilities.

   (b) Evaluate this linear discriminant rule by (i) directly classifying all cases using the above sample linear discriminant function; (ii)Cross-Validation procedure. What do you find from the confusion matrices obtained from (i) and (ii)? Which evaluation method is more appropriate? Explain.

(c) Classify the 10 new cases in hemophiliaNew.DAT using the discriminant function in (a).

3. Let $\Sigma = \begin{pmatrix} 5 & 2 \\ 2 & 4 \end{pmatrix}$ be the covariance matrix of a bivariate random vector $\boldsymbol{X}$ with $E\boldsymbol{X} = \mathbf{0}$.

(a) Determine the population principal components Y_1 and Y_2 from Σ. Also calculate the proportions of the total population variance explained by Y_1 and Y_2.

(b) Convert the covariance matrix Σ to a correlation matrix ρ. Determine the principal components Y_1^* and Y_2^* from ρ in terms of the standardized random vector $\boldsymbol{Z} = (Z_1, Z_2)'$, where $Z_1 = X_1/\sqrt{\sigma_{11}}$ and $Z_2 = X_2/\sqrt{\sigma_{22}}$. Also calculate the proportions of the total variance explained by Y_1^* and Y_2^*.

(c) Compare the components obtained in (a) and (b) in an appropriate way. Are they the same?

(d) Compute the correlations ρ_{Y_i,X_j} and $\rho_{Y_i^*,Z_j}$ for $i, j = 1, 2$.

4. The covariance matrix $\Sigma = \begin{pmatrix} 1.0 & .63 & .45 \\ .63 & 1.0 & .35 \\ .45 & .35 & 1.0 \end{pmatrix}$ is for the $p = 3$ standardized random variables Z_1, Z_2 and Z_3.

(a) The above Σ can be generated by the $m = 1$ factor model $Z_1 = .9F_1 + \epsilon_1$, $Z_2 = .7F_1 + \epsilon_2$, $Z_3 = .5F_1 + \epsilon_3$, where $\text{var}(F_1) = 1$, $\text{cov}(F_1, \boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \Psi = \text{diag}(\Psi_1, \Psi_2, \Psi_3)$. Find Ψ.

(b) Calculate the communalities h_i^2, $i = 1, 2, 3$, and interpret these quantities.

(c) Calculate $\text{corr}(Z_i, F_1)$ for $i = 1, 2, 3$. Which variable might carry the greatest weight in "naming" the common factor?

(d) Find the eigenvalues and normalized eigenvectors for the covariance matrix Σ. Still assuming an $m = 1$ factor model, use the principal component method to compute the loading matrix L^* and the matrix of specific variances Ψ^*. Compare the results with those in (a). Are they the same?