Assignment 3

Deadline to hand in: Nov. 21 in class

Q. 1 (Do it by hands without using a computer.)

At the time of transplanting, four measurements of size of apples were made on each of 54 apple trees when they were planted. The four measurements are:

X_1 = tree weight in kg

X_2 = the square of the trunk circumference in cm^2

X_3 = the total length of the branches in cm

X_4 = a measure of the height of the tree in cm

The sample correlation matrix is:

$$R = \begin{bmatrix} 1 & & & \\ .7571 & 1 & & \\ .7792 & .6684 & 1 & \\ .5500 & .6017 & .3207 & 1 \end{bmatrix}$$

The eigenvalues and eigenvectors associated with R are

$$\text{eigenvalues}: \quad 2.855 \quad 0.721 \quad 0.253 \quad \hat{\lambda}_4$$

$$\text{eigenvectors}: \quad \begin{bmatrix} .55 \\ .54 \\ .49 \\ .42 \end{bmatrix} \quad \begin{bmatrix} -.16 \\ .06 \\ -.56 \\ .81 \end{bmatrix} \quad \begin{bmatrix} .35 \\ -.84 \\ .26 \\ .31 \end{bmatrix} \quad \begin{bmatrix} .74 \\ .02 \\ -.1 \\ -.28 \end{bmatrix}$$

(a) Find the value of $\hat{\lambda}_4$.

(b) Write down the formula for the first principal component.

(c) What is the percentage of total variance in the standardized variables accounted for by the first principal component?

(d) Give a one sentence interpretation of the first principal component.

(e) Give a one sentence interpretation of the second principal component.

(f) Estimate the correlation between the second principal component and the standardized total length of branches.

(g) Estimate the correlation between the scores for the first and second components.

Q. 2 The data, "spotowls.txt" is posted at Blackboard. It is obtained from the habitat association study. For each site, there is one line of data. The first column of the data indicates the nesting sites: "N" stands for "National Forest in Western Oregon" and "R" stands for random sites. The remaining seven columns contain percentages of mature forest within consecutive rings with outer diameters of 0.91, 1.18, 1.40, 1.60, 1.77, 2.41, 3.38 km, respectively.

(a) Obtain the sample correlation matrix for these data. Determine principal components for the percentages of mature forest for the seven rings. Write down the formulas for the first two principal components. Give an interpretation of each of these two components.

(b) Describe how the first two principal components correspond to patterns in the correlation matrix.

(c) What proportion of the total sample variance of the standardized variables is accounted by the two principal components described in part a?

(d) How many components are needed to adequately describe the variation in the standardized percentages of mature forest at these 60 sites. Explain.

(e) Plot scores for the first two principal components on a scatter plot. Use different symbols or colors for nesting sites (N) and randomly selected sites (R). What does this plot reveal?

(f) Test the null hypothesis that the vectors of mean percentage of mature forest are the same for nesting sites and random sites. Give a formula for your test statistic, degrees of freedom, and a p-value. State your conclusions.

Q. 3 A firm is attempting to evaluate the quality of its sales staff and is trying to find an examination or series of tests that may reveal the potential for good performance in sales. The firm has selected a random sample of 50 sales people and has evaluated each on 3 measures of performance: growth of sales, profitability of sales, and new-account sales. These measures have been converted to a scale, on which 100 indicates "average" performance. Each of the 50 individuals took each of 4 tests, which supported to measure creativity, mechanical reasoning, abstract reasoning, and mathematical ability, respectively. The $n = 50$ observations on $p = 7$ variables are listed in Table 9-12 on text book page 536. The data, "T9-12.DAT" is posted at Blackboard.

(a) Assume an orthogonal factor model for the standardized variable $Z_i = (X_i - \mu_i)/\sqrt{\sigma_{ii}},\ i = 1, 2, \ldots, 7$. Obtain the maximum likelihood solution for $m = 3$ common factors.

(b) Given your solution in (a), obtain the rotated loadings for $m = 3$. Interpret factor solutions.

(c) List the estimated communalities, specific variances, and $\hat{L}\hat{L}' + \hat{\Psi}$ for $m = 3$ solutions.

(d) Conduct a test of $H_0 : \Sigma = LL' + \Psi$ versus $H_1 : \Sigma \neq LL' + \Psi$ for $m = 3$ at $\alpha = 0.05$ level.

(e) Suppose a new salesperson, selected at random, obtains the test scores $\underset{\sim}{x}' = [x_1, x_2, \ldots, x_7] = [110, 98, 105, 15, 18, 12, 35]$. Calculate the salesperson's factor score using the regression model.

Note: the components of $\underset{\sim}{x}$ must be standardized using the sample mean and variance calculated from the original data.