Multivariate Analysis - Homework 5

1. For multivariate regression models, using the same notations as slides, derive

   (a) the distribution of predicted vector $\hat{\mathbf{y}}_0$ of a new response $\mathbf{y}_0$, with observed $\mathbf{x}_0$;

   (b) the $(1-\alpha)\%$ confidence region/ellipse of $E(\mathbf{y}_0)$;

   (c) the $(1-\alpha)\%$ simultaneous confidence intervals of all the p components of $E(\mathbf{y}_0)$ (for the derivation of simultaneous confidence interval, please refer to Result 5.3 in P225 of book "*Applied Multivariate Statistical Analysis*");

   (d) the $(1-\alpha)\%$ simultaneous prediction intervals of all the p components of $\mathbf{y}_0$.

2. Use (1) inequality, (2) Lagrange multiplier, (3) SVD to obtain principal components.

3. For the covariance matrix of $\mathbf{y} = (Y_1, Y_2)'$,

$$\mathbf{\Sigma} = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix},$$

   (a) Determine the population principal components $\mathbf{z} = (Z_1, Z_2)'$ by hand.

   (b) Compute the proportion of total population variance explained by the first principal component Z_1.

   (c) Suppose the original variables follows bivariate normal distribution with mean vector $(1, 2)'$. Sketch the constant density ellipse and indicate the principal components on your graph.

   (d) Convert the covariance matrix to a correlation matrix $\mathbf{P}$. Determine the principal components $\mathbf{v} = (V_1, V_2)'$ from $\mathbf{P}$, and compute the proportion of total population variance explained by the first principal component V_1.

   (e) Compare the components calculated in (d) with those obtained in (b). Are they the same? Should they be?

   (f) Find the correlation matrix $CORR(\mathbf{z}, \mathbf{y})$, $CORR(\mathbf{v}, \mathbf{y})$ and $CORR(\mathbf{z}, \mathbf{v})$, respectively. Comment.

4. We present the results of several factor analyses on bone and skull measurements of white leghorn fowl. The full data set consists of $n = 276$ measurements on bone dimensions:

$$\text{Head:} \quad \begin{cases} X_1 = \text{skull length} \\ X_2 = \text{skull breadth} \end{cases}$$

$$\text{Leg:} \quad \begin{cases} X_3 = \text{femur length} \\ X_4 = \text{tibia length} \end{cases}$$

$$\text{Wing:} \quad \begin{cases} X_5 = \text{humerus length} \\ X_6 = \text{ulna length} \end{cases}$$

The sample correlation matrix is

$$\mathbf{R} = \begin{bmatrix} 1.000 & .505 & .569 & .602 & .621 & .603 \\ .505 & 1.000 & .422 & .467 & .482 & .450 \\ .569 & .422 & 1.000 & .926 & .877 & .878 \\ .602 & .467 & .926 & 1.000 & .874 & .894 \\ .621 & .482 & .877 & .874 & 1.000 & .937 \\ .603 & .450 & .878 & .894 & .937 & 1.000 \end{bmatrix}$$

(a) Obtain the estimated factor loadings by the principal component method, suppose 3 factors are used.

(b) Estimate the specific variances.

(c) Estimate the communalities.

(d) Estimate the proportion of variance explained by each factor.

(e) Compute the residual matrix and comment the result.

(f) Suppose the rotated factor loadings are as follows, give the interpretation of the factors.

Rotated estimated loadings

F_1^*	F_2^*	F_3^*
.355	.244	.902
.235	.949	.211
.921	.164	.218
.904	.212	.252
.888	.228	.283
.908	.192	.264

5. (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.