Ramakrishna Mission Vivekananda University

Belur Math, Howrah, West Bengal

School of Mathematical Sciences, Department of Data Science

M.Sc. in Big Data Analytic 2016, Semester Exam — Date: 10 May 2017

Course : **DA310: Multivariate Statistics** — Time: 2 hrs

Instructor : Dr. Sudipta Das — *Max marks: 60*

Student signature and Id:

1. Consider the covariance matrix
$$\Sigma = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}.$$

   (a) Derive the first principal component.

   (b) Derive also, the first principal component form the correlation matrix corresponding to Σ.

   (c) When the principal components are preferred to be derived from the correlation matrix instead from the covariance matrix?

   [6+6+3=15]

2. Orthogonal factor model with p features and m common factors is described as follows:
$$X = \mu + LF + \epsilon.$$

   (a) Prove that
$$\Sigma = LL' + \Psi$$
   and state the assumptions needed to prove it. Is it, always, possible to get a consistent solution to the above equation?

   (b) The Σ and L matrices are given as
$$\Sigma = \begin{bmatrix} 19 & 30 & 2 & 12 \\ 30 & 57 & 5 & 23 \\ 2 & 5 & 38 & 47 \\ 12 & 23 & 47 & 68 \end{bmatrix} \text{ and } L = \begin{bmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{bmatrix},$$
   respectively. Find the Ψ matrix.

   (c) State two methods, used for estimation of factor loadings.

   (d) Why factor rotation is needed?

   (e) In which situation the varimax criterion is used?

   [(3+3+1)+2+2+2+2=15]

3. (a) To construct a procedure for detecting potential hemophilia A carriers, blood samples were assayed for two groups of women. The first group (normal) did not carry the

hemophilia gene. The second group (obligatory carrier) was selected from known hemophilia A carriers. Two variables were observed (AHF antihemophilic factor)

$$X_1 = \log_{10}(\text{AHF activity})$$

$$X_2 = \log_{10}(\text{AHF-like antigen}).$$

From the observations following informations are derived

$$\bar{x}_1 = \begin{bmatrix} -.0065 \\ -.0390 \end{bmatrix}, \bar{x}_2 = \begin{bmatrix} -.2483 \\ .0262 \end{bmatrix} \text{ and } S_{pooled}^{-1} = \begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix}.$$

Measurement of AHF activity and AHF-like antigen on a woman who may be a hemophilia A carrier give $x_1 = -.210$ and $x_2 = -.044$. Should this woman be classified as normal or obligatory carrier?

(b) Describe the procedure of estimating the logistic regression coefficients.

[9+6=15]

4. (a) Consider the distances between the pairs of five objects as follows:

$$D = \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix}.$$

Draw the complete linkage dendogram fir distance between five objects.

(b) Suppose we measure two variables X_1 and X_2 for each of four items A, B, C and D. The data are given in the following table

	Observations	
Item	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Consider the inital clusters (AB) and (CD). Find the final $K = 2$ clusters.

[8+7=15]

This exam has total 4 questions, for a total of 60 points and 0 bonus points.