MTH443A: Midsem Examination
Full Marks: 50

[1] Let the covariance matrix of the random vector $\underline{X} = (X_1, \dots, X_p)^T$ $(p > 1)$ be $\Sigma = (\sigma_{ij})$; where $\sigma_{ii} = 1$ for all $i = 1, \dots, p$ and $\sigma_{ij} = \rho$ for all $i \neq j$ and $i, j = 1, \dots, p$.

(a) Prove or disprove the statement: For every ρ such that $|\rho| < 1$, elements of $\underline{X}$ are **not** "linearly related with probability 1". $-1 < \rho < 1$

(b) Suppose $p = 5$ and $\rho = -\frac{1}{5}$, find the proportion of total variation in $\underline{X}$ explained by the first principal component derived from the covariance matrix of $\underline{X}$.

10 marks

[2] The distance matrix corresponding to 6 multidimensional cases $C_1, C_2, C_3, C_4, C_5, C_6$ is given by

$$D = \begin{bmatrix} 0 & 23 & 14 & 24 & 13 & 7 \\ 23 & 0 & 10 & 21 & 12 & 11 \\ 14 & 10 & 0 & 9 & 3 & 7 \\ 24 & 21 & 9 & 0 & 6 & 17 \\ 13 & 12 & 3 & 6 & 0 & 8 \\ 7 & 11 & 7 & 17 & 8 & 0 \end{bmatrix}$$

(a) Construct the dendogram tree corresponding to an agglomerative complete linkage hierarchical clustering algorithm.
(b) Identify the clusters of cases at a merger level 15.
(c) Find the level at which we get 3 clusters and list the objects in the clusters.

10 Marks

[3] Consider the following dataset with 15 observations of a bivariate random vector $\underline{X} = (X_1, X_2)^T$:

$$\mathcal{X} = \left\{ \begin{matrix} (12,10), (5,10), (6,3), (8,9), (1,21), (9,8), (8,1), (7,10), \\ (20,1), (2,4), (1,10), (5,12), (5,1), (6,7), (21,4) \end{matrix} \right\},$$

where, for the i^{th} observation pair $(x_{i1}, x_{i2}) \in \mathcal{X}$, x_{i1} denotes the i^{th} observed value of X_1 and x_{i2} denotes the i^{th} observed value of X_2.

(a) Compute kernel density estimate of the variable X_1 at the points 10 and 15 using the rectangular kernel

$$K(z) = \begin{cases} \frac{1}{2}, & \text{if } |z| \leq 1 \\ 0, & \text{otherwise} \end{cases},$$

and with kernel bandwidth $h = 4$.

(b) Compute the density estimate of the variable X_2 at the points 9 and 21 using a 4-nearest neighbor approach.

(c) Compute an estimate of the bivariate joint density of (X_1, X_2) at the point $(9,9)$, assuming independence of the 2 components and using rectangular kernel (with kernel bandwidth $h = 4$) based kernel density estimates of the two components.

15 Marks

[4] Let $\underline{x}_1 = (0,2)^T$, $\underline{x}_2 = (8,2)^T$, $\underline{x}_3 = (4,6)^T$, $\underline{x}_4 = (6,4)^T$, $\underline{x}_5 = (4,4)^T$ and $\underline{x}_6 = (6,2)^T$ be observed feature vectors of 6 cases $C_1, C_2, C_3, C_4, C_5, C_6$, respectively.

(a) 2 different clustering algorithms gave the following final cluster partitions:

Algorithm I Partition: $(C_1, C_2), (C_3, C_4, C_5, C_6)$

Algorithm II Partition: $(C_1, C_3), (C_2, C_4, C_5, C_6)$

Let

$$S_W = \frac{1}{n} \sum_{j=1}^{g} \sum_{i=1}^{n} Z_{ji} (\underline{x}_i - \underline{m}_j)(\underline{x}_i - \underline{m}_j)^T$$

be the pooled within cluster sum of squares and cross product scatter matrix for a fixed number, g, of clusters obtained from n cases.

$$Z_{ji} = \begin{cases} 1, & \text{if } \underline{x}_i \in \text{cluster } j \\ 0, & \text{otherwise} \end{cases}$$

and $\underline{m}_j$ is the mean of cluster j.

Which of the above partitions would you prefer if clustering criterion based on $trace(S_W)$ is to be used?

(b) Staring from the random partition $(C_3, C_2), (C_1, C_4, C_5, C_6)$, obtain k-means clustering of the cases, with $k = 2$.

15 marks

(1)
(a) $$\underset{p\times p}{\Sigma} = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ & 1 & \cdots & \rho \\ & & \ddots & \\ & & & 1 \end{pmatrix}$$

Σ is cov matrix $\Leftrightarrow$ Σ is p.s.d.

$$|\Sigma| = (1-\rho)^{p-1}(1+(p-1)\rho)$$

For $p=3$, if $\rho = -\frac{1}{p-1} = -\frac{1}{2} > -1$, then $|\Sigma| = 0$ and the elements of $\underset{\sim}{x}$ are linearly related w.p.1

Hence, the statement is disproved. 5

For all $\rho \ni -\frac{1}{p-1} < \rho < 1$, $\Sigma > 0$ and hence elements of $\underset{\sim}{x}$ are not linearly related.

(b) For $p = 5$, $\rho = -\frac{1}{5}$

$|\Sigma - \lambda I| = 0$ gives eigen values as

$(1-\rho), (1-\rho), (1-\rho), (1-\rho), (1+(p-1)\rho)$

i.e. $\frac{6}{5}, \frac{6}{5}, \frac{6}{5}, \frac{6}{5}, \frac{1}{5}$

$\Rightarrow$ proportion of total variation in $\underline{x}$ explained by 1^{st} PC

is $\frac{6}{25}$ 5

(2)

$$D = \begin{array}{c} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \end{array} \begin{pmatrix} 0 & & & & & \\ 23 & 0 & & & & \\ 14 & 10 & 0 & & & \\ 24 & 21 & 9 & 0 & & \\ 13 & 12 & \boxed{3} & 6 & 0 & \\ 7 & 11 & 7 & 17 & 8 & 0 \end{pmatrix}$$

1st merger (C_3, C_5) — at level 3 $\quad 1\frac{1}{2}$

Updated distance matrix

$$D_2 = \begin{array}{c} C_1 \\ C_2 \\ C_4 \\ C_6 \\ (C_3, C_5) \end{array} \begin{pmatrix} 0 & & & & \\ 23 & 0 & & & \\ 24 & 21 & 0 & & \\ \boxed{7} & 11 & 17 & 0 & \\ 14 & 12 & 9 & 8 & 0 \end{pmatrix}$$

$d_{(1,(3,5))} = \max(14, 13) = 14$

$d_{(2,(3,5))} = 12$, $d_{(4,(3,5))} = 9$, $d_{(6,(3,5))} =$

2nd merger (C_1, C_6) — at level 7 $\quad 1\frac{1}{2}$

D_3

$$D_3 = \begin{array}{c} 2 \\ 4 \\ (C_3, C_5) \\ (C_1, C_6) \end{array} \begin{pmatrix} 0 & & & \\ 21 & 0 & & \\ 12 & \boxed{9} & 0 & \\ 23 & 24 & 14 & 0 \end{pmatrix}$$

3rd merger $(C_4, (C_3, C_5))$ — at level 9 $\quad 1\frac{1}{2}$

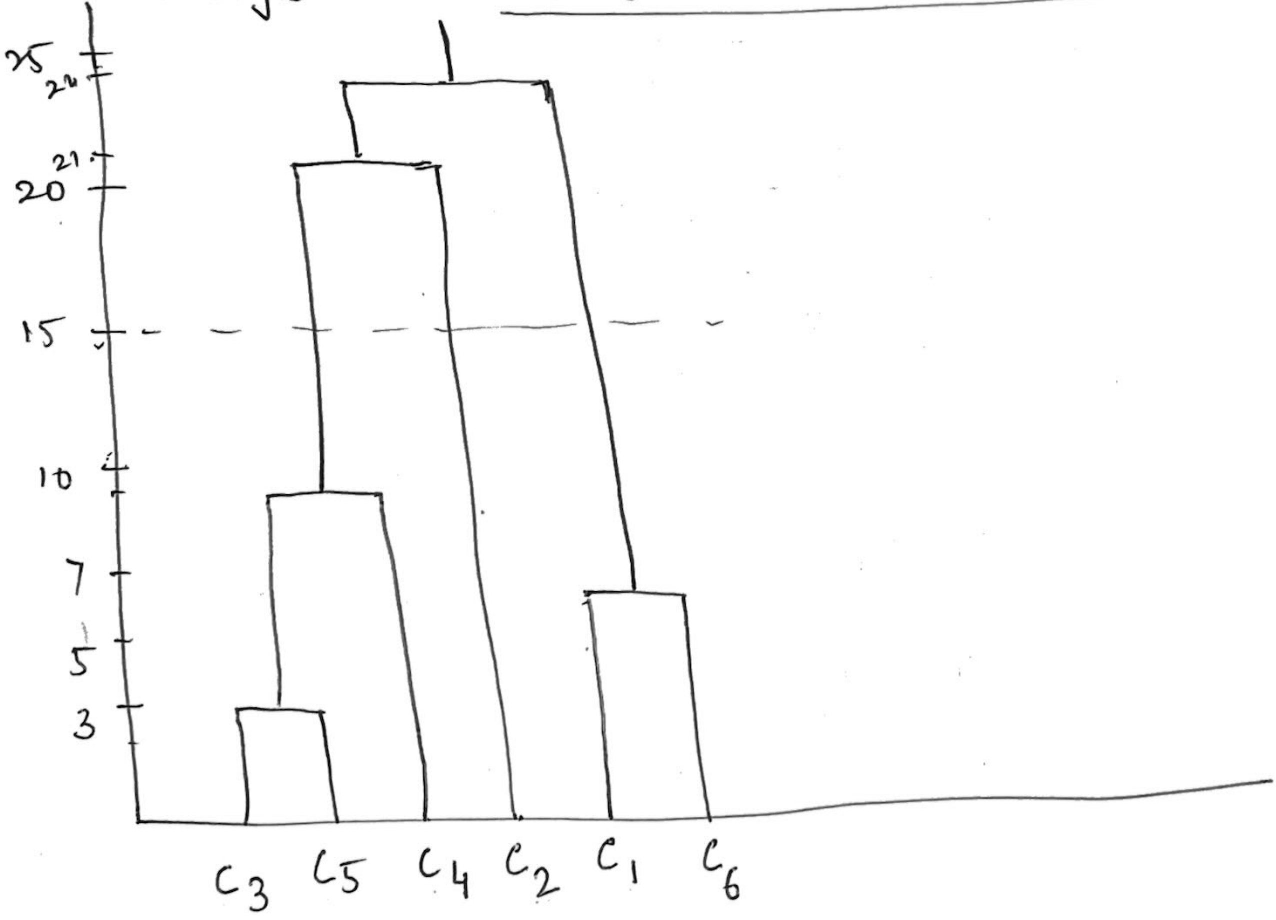
$$D_4 = \begin{array}{l} C_2 \\ (C_1, C_6) \\ (C_4, (C_3, C_5)) \end{array} \begin{pmatrix} 0 & & \\ 23 & 0 & \\ \boxed{21} & 24 & 0 \end{pmatrix}$$

4^{th} merger $(C_2, (C_4, C_3, C_5))$ — at level 21 $1\frac{1}{2}$

$$D_5 = \begin{array}{l} (C_1, C_6) \\ (C_2, C_4, C_3, C_5) \end{array} \begin{pmatrix} 0 & \\ 24 & 0 \end{pmatrix}$$

5^{th} merger $(C_1, C_6, C_2, C_4, C_3, C_5)$ - at level 24

$1\frac{1}{2}$



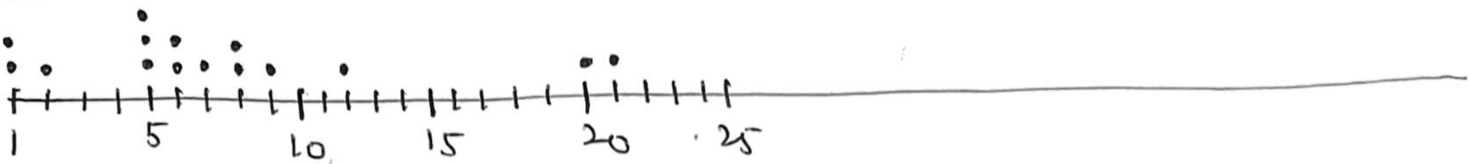
Dendogram

(b) Cluster at level 15 — (C_3, C_4, C_5), (C_2), (C_1, C_6) $1\frac{1}{2}$

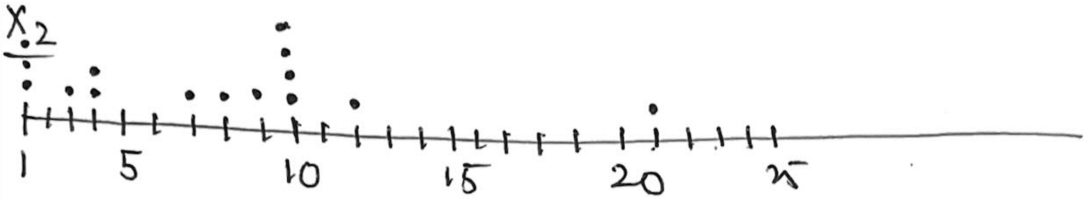
(c) At merger level 9, we get 3 clusters
(C_3, C_4, C_5), (C_2), (C_1, C_6) 1

(3)

$\underline{X_1}$



$\underline{X_2}$



(a) KDE

$$f_1(10) = \frac{1}{15 \times 4} \sum_{i=1}^{15} \frac{1}{2} I_{(|10 - x_i| \leq 4)}$$

$$= \frac{1}{60} \times \frac{1}{2}(7) = \frac{7}{120} \quad - 3$$

$$f_1(15) = \frac{1}{15 \times 4} \sum_{i=1}^{15} \frac{1}{2} I_{(|15 - x_i| \leq 4)}$$

$$= \frac{1}{120}(1) = \frac{1}{120} \quad - 3.$$

(b) K-nn

$$f_2(9) = \frac{4}{15}\left(V_4(9)\right)^{-1} = \frac{4}{15} \times \frac{1}{2} = \frac{4}{30} = \frac{2}{15} \quad 3$$

$$f_2(21) = \frac{4}{15}\left(V_4(21)\right)^{-1} = \frac{4}{15} \times \frac{1}{22} = \frac{2}{165} \quad 3$$

(c) $f_{1,2}(9,9) = f_1(9)\, f_2(9)$

$f_1(9) = \frac{10}{120}$ $\qquad f_2(9) = \frac{8}{120}$

$\Rightarrow f_{1,2}(9,9) = \frac{1}{12 \times 15}$ $\quad 3$

(4) (a)

$$\operatorname{tr}(S_W) = \frac{1}{n}\sum_{j=1}^{g}\sum_{i=1}^{n} z_{ji}\,\|\underline{x}_i - \underline{m}_j\|^2$$

i.e. $$= \frac{1}{6}\sum_{i=1}^{6}\sum_{j=1}^{2} z_{ji}\,\|\underline{x}_i - \underline{m}_j\|^2$$

Algorithm I partition

$(c_1, c_2) \rightarrow \underset{\sim}{m}_1 = (4, 2)$

$(c_3, c_4, c_5, c_6) \rightarrow \underset{\sim}{m}_2 = (5, 4)$

$$(\operatorname{tr} S_W)^{I} = \frac{1}{6}\left(\left\{\|\underset{\sim}{x}_1 - \underset{\sim}{m}_1\|^2 + \|\underline{x}_2 - \underset{\sim}{m}_1\|^2\right\}\right.$$
$$\left. + \left\{\|\underline{x}_3 - \underline{m}_2\|^2 + \|\underline{x}_4 - \underline{m}_2\|^2 + \|\underline{x}_5 - m_3\|^2 + \|\underline{x}_6 - \underline{m}_3\|^2\right\}\right)$$
$$= \frac{1}{6}\left(\{32\} + \{12\}\right) = \frac{44}{6}$$

$3\frac{1}{2}$

Algorithm II partition

$(c_1, c_3) \rightarrow \underset{\sim}{m}_1 = (2, 4)$

$(c_2, c_4, c_5, c_6) \rightarrow \underset{\sim}{m}_2 = (6, 3)$

$$(\operatorname{tr} S_W)^{II} = \frac{1}{6}\left(\{16\} + \{12\}\right) = \frac{28}{6}$$
$$< (\operatorname{tr} S_W)^{I}$$

$3\frac{1}{2}$

$\Rightarrow$ Algorithm II partition is the preferred one

(b) Initial partition

I: (C_2, C_3) II: (C_1, C_4, C_5, C_6)

Centr: $(6,4)$ $(4,3)$

C_4 is closer to (C_2, C_3) centroid than it's initial centroid

$\Rightarrow$ relocation of C_4 to I

New partition

I: (C_2, C_3, C_4) II: (C_1, C_5, C_6)

Centr: $(6,3)$ $\left(\frac{10}{3}, \frac{8}{3}\right)$

C_3 is closer to II centroid than I centroid

$\Rightarrow$ relocation of C_3 to II

New partition

I: (C_2, C_4) II: (C_1, C_3, C_5, C_6)

Centr: $(7,3)$ $\left(\frac{14}{4}, \frac{14}{4}\right)$

C_6 is closer to I centroid than II centroid

$\Rightarrow$ relocation of C_6 to I

8

New partition

I: (C_2, C_4, C_6) II: (C_1, C_3, C_5)

Centr: $\left(\frac{20}{3}, \frac{8}{3}\right)$ $\left(\frac{8}{3}, 3\right)$

No relocation reqd

Final partition: (C_2, C_4, C_6), (C_1, C_3, C_5)