MTH443A: Mid semester examination
Full Marks: 60

[1] Let $\underline{X} = (X_1, X_2)^T$ be a random vector such that $\underline{X} \sim N_2(\underline{0}, \Sigma)$ $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$.

(a) Find the two principal components Y_1 and Y_2 derived from Σ.

(b) Find the proportion of total variation in $\underline{X}$ explained by the first principal component Y_1.

(c) Verify whether or not the two principal components are independent.

(d) Find $Correl(X_1, Y_2)$.

(e) Let $\underline{Y} = (Y_1, Y_2)^T$ and $\underline{Z} = \begin{pmatrix} \Sigma^{-\frac{1}{2}} \underline{X} \\ \underline{Y} \end{pmatrix}$. Prove or disprove "total variation of $\underline{Z} = 2$".

16 (4+2+3+3+4) 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

$$D = \begin{pmatrix} 0 & 12 & 10 & 9 & 16 & 7 \\ 12 & 0 & 3 & 6 & 11 & 13 \\ 10 & 3 & 0 & 5 & 4 & 9 \\ 9 & 6 & 5 & 0 & 3 & 2 \\ 16 & 11 & 4 & 3 & 0 & 14 \\ 7 & 13 & 9 & 2 & 14 & 0 \end{pmatrix}$$

(a) Construct the dendogram tree corresponding to an agglomerative complete linkage hierarchical clustering algorithm. Identify the clusters at merger level 10 from the dendogram.

(b) Suppose $\mathcal{C}_1 = \{C_1, C_2, C_3, C_4\}$ and $\mathcal{C}_2 = \{C_5, C_6\}$. Find average linkage distance between $\mathcal{C}_1$ and $\mathcal{C}_2$.

14 (12+2) marks

[3] Let $\mathcal{X} = \{20,10,16,2,3,4,8,4,1,12,11,19,18,21,5,11,12,19,2,11\}$ be an observed sample of size 20 from a population with unknown probability density function $f(x)$.

(a) Compute kernel density estimate at the points 7 and 24 using the rectangular kernel

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

with kernel bandwidth, h, equal to 4.

(b) Find density estimate at the points 3 and 17 using a 4-nearest neighbor approach.

(c) Compute kernel density estimate at the point 24 using the triangular kernel

$$K(z) = \begin{cases} 1 - |z|, & \text{if } |z| \le 1 \\ 0, & \text{otherwise} \end{cases}'$$

with kernel bandwidth, h, equal to 4.

17 (6+6+5) marks

[4] Consider the Bhattacharya distance between 2 p-dimensional populations, π_1 and π_2,

$$J_B = -\log_e\left(\int \{p(\underline{x}|\pi_1)\, p(\underline{x}|\pi_2)\}^{\frac{1}{2}}\, d\underline{x}\right)$$

$p(\underline{x}|\pi_i)$ denotes the joint probability density function under π_i, $i = 1,2$; $\underline{x} = (x_1, \dots, x_p)^T$.
Prove or disprove "$J_B = \sum_{i=1}^{p} J_B^i$, where J_B^i is the Bhattacharya distance corresponding to the i^{th} dimension of the 2 populations.

7 marks

[5] Let $\underline{x}_1 = \begin{pmatrix}1\\2\end{pmatrix}$, $\underline{x}_2 = \begin{pmatrix}3\\2\end{pmatrix}$, $\underline{x}_3 = \begin{pmatrix}1\\6\end{pmatrix}$ and $\underline{x}_4 = \begin{pmatrix}5\\4\end{pmatrix}$ be observed feature vectors of 4 cases, C_1, C_2, C_3, C_4, respectively. 2 different clustering algorithms (Algorithm A and Algorithm B) yield the following partitions:

Algorithm A partition: $\{C_1,\ C_3\},\ \{C_2,\ C_4\}$
Algorithm B partition: $\{C_1,\ C_4\},\ \{C_2,\ C_3\}$

Let $S_B = \frac{1}{n}\sum_{j=1}^{g}\sum_{i=1}^{n} Z_{ji}\left(\underline{m}_j - \overline{m}\right)\left(\underline{m}_j - \overline{m}\right)^T$ be the between cluster sum of squares and cross product scatter matrix for a fixed number, g, of clusters obtained from n cases. $Z_{ji} = 1$, if $\underset{\sim}{x}_i \in$ cluster j; 0, otherwise. $\underset{\sim}{m}_j$ is the mean of cluster j and $\overline{m}$ is the overall sample mean vector. Which of the above partitions would you prefer if clustering criterion based on $trace(S_B)$ is to be used?

6 marks

(1)

(a) $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ $\rho = \frac{1}{2}$

$|\Sigma - \lambda I| = 0 \Rightarrow (1-\lambda)^2 - \rho^2 = 0$

$\lambda = 1-\rho, 1+\rho$

$\lambda_1 = \frac{3}{2}, \quad \lambda_2 = \frac{1}{2}$

$\Sigma \underset{\sim}{x} = \lambda \underset{\sim}{x}$

for $\lambda_1 = \frac{3}{2}$, $\underset{\sim}{e}_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$

& for $\lambda_2 = \frac{1}{2}$, $\underline{e}_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$

PCs $Y_1 = \frac{1}{\sqrt{2}} X_1 + \frac{1}{\sqrt{2}} X_2$

$Y_2 = \frac{1}{\sqrt{2}} X_1 - \frac{1}{\sqrt{2}} X_2$ 4

(b) Proportion of total variation in $\underset{\sim}{x}$ explained by Y_1

$= \frac{3/2}{2} = 0.75$ 2

(c) $\text{Cov}(Y_1, Y_2) = 0$

$\underset{\sim}{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \underset{\sim}{X} = A \underset{\sim}{X} \sim N_2\left(\underset{\sim}{0}, \begin{pmatrix} 3/2 & 0 \\ 0 & 1/2 \end{pmatrix}\right)$

Since $\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N_2$ and are uncorrelated, Y_1 & Y_2 are indep 3

(d) $\text{Corr}^n(X_1, Y_2) = \frac{1}{\sqrt{2}} \sqrt{\frac{1/2}{1}} = \frac{1}{2}$ 3

(e) $\underset{\sim}{Z} = \begin{pmatrix} \Sigma^{-1/2} \underline{X} \\ \underline{Y} \end{pmatrix}$

$\text{Cov}(\underset{\sim}{Z}) = \begin{pmatrix} \text{Cov}(\Sigma^{-1/2}\underline{X}) & \text{Cov}(\Sigma^{-1/2}\underline{X}, \underline{Y}) \\ \text{Cov}(\underline{Y}, \Sigma^{-1/2}\underline{X}) & \text{Cov}(\underline{Y}) \end{pmatrix}$

$$\Rightarrow \operatorname{Cov}(\underline{z}) = \begin{pmatrix} I_2 & \operatorname{Cov}(\Sigma^{-1/2}\underline{x}, \underline{y}) \\ \operatorname{Cov}(\underline{y}, \Sigma^{-1/2}\underline{x}) & D_\lambda \end{pmatrix} \quad D_\lambda = \begin{pmatrix} 3/2 & 0 \\ 0 & 1/2 \end{pmatrix}$$ — 2

total variation in $\underline{z} = \operatorname{tr}(\operatorname{Cov}\underline{z}) = 4$ — 2

$\neq 2$ disproved.

(2)

$$D = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{pmatrix} 0 \\ 12 & 0 \\ 10 & 3 & 0 \\ 9 & 6 & 5 & 0 \\ 16 & 11 & 4 & 3 & 0 \\ 7 & 13 & 9 & \boxed{2} & 14 & 0 \end{pmatrix}$$

$(4,6) \to$ at level 2

$$D_1 = \begin{array}{c} 1 \\ 2 \\ 3 \\ 5 \\ (4,6) \end{array} \begin{pmatrix} 0 \\ 12 & 0 \\ 10 & \boxed{3} & 0 \\ 16 & 11 & 4 & 0 \\ 9 & 13 & 9 & 14 & 0 \end{pmatrix}$$ 2

$(2,3) \to$ at level 3

$$D_2 = \begin{array}{c} 1 \\ 5 \\ (4,6) \\ (2,3) \end{array} \begin{pmatrix} 0 \\ 16 & 0 \\ \boxed{9} & 14 & 0 \\ 12 & 11 & 13 & 0 \end{pmatrix}$$ 2

$(1,(4,6)) \to$ at level 9

$$D_3 = \begin{array}{c} 5 \\ (2,3) \\ (1,4,6) \end{array} \begin{pmatrix} 0 \\ \boxed{11} & 0 \\ 16 & 13 & 0 \end{pmatrix}$$ 2

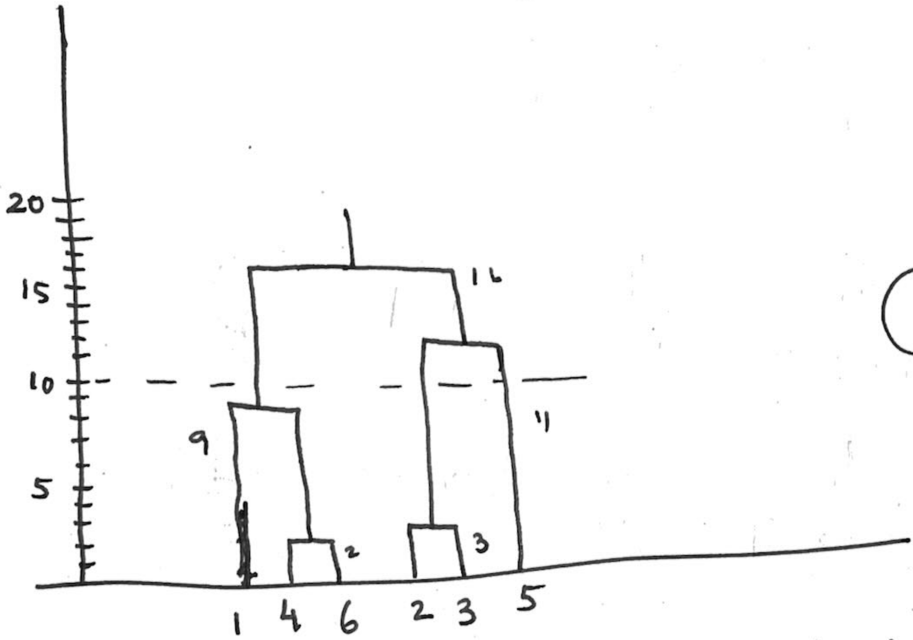
$(5,(2,3)) \to$ at level 11

$$D_5 = \begin{array}{c} (5,2,3) \\ (1,4,6) \end{array}\begin{pmatrix} 0 & \\ 16 & 0 \end{pmatrix}$$

2

$(1, 2, 3, 4, 5, 6) \rightarrow$ at level 16

Dendogram



2

Cluster at level 10 : $(c_1, c_4, c_6), (c_2, c_3), c_5$ — 2

(b) $\ell_1 = (c_1, c_2, c_3, c_4)$ & $\ell_2 = (c_5, c_6)$

Avg linkage dist betn ℓ_1 & ℓ_2

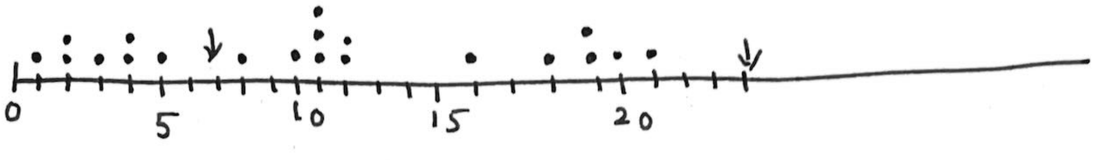
$$= \frac{1}{n_{\ell_1} n_{\ell_2}} \sum_{i \in \ell_1} \sum_{j \in \ell_2} d_{ij}$$

$$= \frac{1}{8}\left((16+7) + (11+13) + (4+9) + (3+2)\right)$$

$$= \frac{65}{8} = 8.125$$

2

(3)



(a) $$f_1^R(7) = \frac{1}{20 \times 4}\left(\frac{1}{2}(9)\right) = \frac{9}{160}$$ (3)

$$f_1^R(24) = \frac{1}{20 \times 4}\left(\frac{1}{2}(2)\right) = \frac{2}{160}$$ (3)

(b) $$f^{4NN}(3) = \frac{4}{20 \times 2} = \frac{4}{40} \quad ; r=1$$ (3)

$$f^{4NN}(17) = \frac{4}{20 \times 4} = \frac{4}{80} \quad ; r=2$$ (3)

(c) $$f^{TK}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$$

$$K(z) = \begin{cases} 1-|z|, & |z| \le 1 \\ 0, & o/w \end{cases}$$

i.e. $$f^{TK}(x) = \frac{1}{20 \times 4}\left(\sum_{i=1}^{20}\left(1-\left|\frac{x-x_i}{4}\right|\right) I_{(|x-x_i| \le 4)}\right)$$

$$\Rightarrow f^{TK}(24) = \frac{1}{80}\left(\left(1-\left|\frac{20-24}{4}\right|\right)+\left(1-\left|\frac{21-24}{4}\right|\right)\right)$$

$$= \frac{1}{80}\left(1-\frac{3}{4}\right) = \frac{1}{320}$$

(5)

(4)
$$J_B = -\log_e \left(\int \{ p(\underline{x}|\pi_1)\, p(\underline{x}|\pi_2) \}^{1/2} d\underline{x} \right)$$

If $x_1, \ldots x_p$ are indep then

$$p(\underline{x}|\pi_i) = \prod_{j=1}^{p} p(x_j|\pi_i)$$

and
$$\int \{ p(\underset{\sim}{x}|\pi_1)\, p(\underline{x}|\pi_2) \}^{1/2} d\underline{x}$$
$$= \prod_{j=1}^{p} \int (p(x_j|\pi_1)\, p(x_j|\pi_2))^{1/2} dx_j$$

& $$J_B = -\sum_{j=1}^{p} \log_e \left\{ \int (p(x_j|\pi_1)\, p(x_j|\pi_2))^{1/2} dx_j \right\}$$
$$= + \sum_{j=1}^{p} J_B^i$$

Unless the components of $\underset{\sim}{X}$ are indep $J_B \neq \sum_{j=1}^{p} J_B^i$ (*)

(7)

<u>Counter example</u>

Suppose π_1 is $N_p(\underset{\sim}{\mu}_1, \Sigma)$
& π_2 is $N_p(\underline{\mu}_2, \Sigma)$ $\Sigma \neq$ diagonal p.d. matrix

Then $$J_B = \frac{1}{8} \underbrace{(\underline{\mu}_2 - \underline{\mu}_1)' \Sigma^{-1} (\underline{\mu}_2 - \underline{\mu}_1)}_{\text{Sq of Mahalanobis distance}}$$

$$\neq \sum_{j=1}^{p} J_B^i .$$

$$\left(J_B^i = \frac{1}{8} \{ (\underset{\sim}{\mu}_{2i} - \mu_{1i})^2 / \sigma_{ii} \} \right)$$

(Give full marks if (*) is ~~correctly~~ concluded (without any counter example.))

(5) $S_B = \frac{1}{4} \sum_{j=1}^{2} \sum_{i=1}^{4} z_{ji} (\underset{\sim}{m}_j - \bar{\underset{\sim}{m}})(\underset{\sim}{m}_j - \bar{\underset{\sim}{m}})'$

$S_W = \frac{1}{4} \sum_{j=1}^{2} \sum_{i=1}^{4} z_{ji} (\underset{\sim}{x}_i - \underline{m}_j)(\underline{x}_i - \underline{m}_j)'$

maximization of $tr(S_B) \Leftrightarrow$ minimization of $tr(S_W)$

One can calculate either $tr(S_W)$ or $tr(S_B)$

e.g. use $tr(S_W)$

Algorithm A partition $c_1 \Rightarrow \underset{\sim}{x}_1 = \begin{pmatrix}1\\2\end{pmatrix}$; $c_2: \underline{x}_2 = \begin{pmatrix}3\\2\end{pmatrix}$; $c_3: \underline{x}_3 = \begin{pmatrix}1\\6\end{pmatrix}$
$c_4: \underline{x}_4 = \begin{pmatrix}5\\4\end{pmatrix}$

$(c_1, c_3) \rightarrow \underset{\sim}{m}_1 = \begin{pmatrix}1\\4\end{pmatrix}$

$(c_2, c_4) \rightarrow \underline{m}_2 = \begin{pmatrix}4\\3\end{pmatrix}$

$tr\, S_W^A = \frac{1}{4}\left(\sum_{i=1}^{4} z_{1i} |\underline{x}_i - \underset{\sim}{m}_1|^2 + \sum_{i=1}^{4} z_{2i} |\underline{x}_i - \underset{\sim}{m}_2|^2\right)$

$= \frac{1}{4}\left(\{4+4\} + \{2+2\}\right) = 3.$ 3

Algorithm B partition

$(c_1, c_4) \rightarrow \underset{\sim}{m}_1 = \begin{pmatrix}3\\3\end{pmatrix}$

$(c_2, c_3) \rightarrow \underline{m}_2 = \begin{pmatrix}2\\4\end{pmatrix}$

$tr\, S_W^B = \frac{1}{4}\left(\{5+5\} + \{5+5\}\right) = 5$

Since $tr\, S_W^A < tr\, S_W^B$ $(\Leftrightarrow tr\, S_B^A > tr\, S_B^B)$

Preferred partition is A 3

((*) Give full marks if the $tr(S_B)$ is calculated)