UNIVERSITI SAINS MALAYSIA

First Semester Examination
2012/2013 Academic Session

January 2013

MSG 366 – Multivariate Analysis
[Analisis Multivariat]

Duration : 3 hours
[Masa : 3 jam]

Please check that this examination paper consists of TWENTY TWO pages of printed material before you begin the examination.

[Sila pastikan bahawa kertas peperiksaan ini mengandungi DUA PULUH DUA muka surat yang bercetak sebelum anda memulakan peperiksaan ini.]

Instructions: Answer **all ten** [10] questions.

*[**Arahan:** Jawab **semua sepuluh** [10] soalan.]*

In the event of any discrepancies, the English version shall be used.

[Sekiranya terdapat sebarang percanggahan pada soalan peperiksaan, versi Bahasa Inggeris hendaklah diguna pakai].

1. (a) Write down three properties that must be satisfied by a distance measure, $d(P, Q)$ between two points P and Q.

(b) For the two points $P = (2, 1, 4, 3, 1)$ and $Q = (3, 4, 1, 5, 7)$ in 5-dimensional space, determine
(i) the Euclidean distance
(ii) the city-block distance
(iii) the Canberra metric.

[15 marks]

1. (a) Tulis tiga sifat yang mesti dipenuhi oleh suatu ukuran jarak, d(P, Q) antara dua titik P dan Q.

(b) Bagi dua titik P = (2, 1, 4, 3, 1) dan Q = (3, 4, 1, 5, 7) dalam ruang 5-dimensi, tentukan
(i) jarak Euclidean
(ii) jarak blok-bandar
(iii) metrik Canberra.

[15 markah]

2. (a) Explain on how you would assess the assumption of normality for multivariate data.

(b) Suppose $\mathbf{X} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Based on the random sample

$$\mathbf{x} = \begin{bmatrix} 4 & 1 \\ 6 & 4 \\ 3 & 2 \\ 7 & 6 \\ 5 & 2 \end{bmatrix},$$

find the maximum likelihood estimates of the 2×1 mean vector $\boldsymbol{\mu}$ and the 2×2 covariance matrix $\boldsymbol{\Sigma}$.

[15 marks]

2. (a) Terangkan bagaimana anda akan menilai andaian kenormalan bagi data multivariat.

(b) Andaikan $\mathbf{X} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Berdasarkan sampel rawak

$$\mathbf{x} = \begin{bmatrix} 4 & 1 \\ 6 & 4 \\ 3 & 2 \\ 7 & 6 \\ 5 & 2 \end{bmatrix},$$

dapatkan anggaran kebolehjadian maksimum bagi vektor min 2×1 , $\boldsymbol{\mu}$ *dan anggaran kebolehjadian maksimum bagi matriks kovarians,* $\boldsymbol{\Sigma}$.

[15 markah]

3. Suppose the random vector $\mathbf{X} = (X_1, X_2, X_3)'$ has a multivariate normal distribution $\mathrm{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)'$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & r & r^2 \\ r & 1 & r \\ r^2 & r & 1 \end{pmatrix}.$$

Show that the conditional distribution of X_1 and X_3 given $X_2 = x_2$ is bivariate normal with mean = $\begin{pmatrix} \mu_1 + r(x_2 - \mu_2) \\ \mu_3 + r(x_2 - \mu_2) \end{pmatrix}$, and covariance = $\begin{pmatrix} 1-r^2 & 0 \\ 0 & 1-r^2 \end{pmatrix}$.

[20 marks]

3. Andaikan vektor rawak $\mathbf{X} = (X_1, X_2, X_3)'$ *mempunyai taburan normal multivariat* $\mathrm{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *yang mana* $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)'$ *dan*

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & r & r^2 \\ r & 1 & r \\ r^2 & r & 1 \end{pmatrix}.$$

Tunjukkan bahawa taburan bersyarat bagi X_1 dan X_3 *diberi* $X_2 = x_2$ *adalah normal bivariat dengan min* = $\begin{pmatrix} \mu_1 + r(x_2 - \mu_2) \\ \mu_3 + r(x_2 - \mu_2) \end{pmatrix}$, *dan kovarians* = $\begin{pmatrix} 1-r^2 & 0 \\ 0 & 1-r^2 \end{pmatrix}$.

[20 markah]

4. Let the random vector $\mathbf{X} = (X_1, X_2, X_3)'$ has a multivariate normal distribution $\mathrm{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = (3, 4, 1)' \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} 5 & 2 & 4 \\ 2 & 1 & 3 \\ 4 & 3 & 6 \end{pmatrix}.$$

Suppose $V_1 = aX_1 + bX_2$ and $V_2 = cX_2 + aX_3$, where *a, b* and *c* are constants.

(i) Show that the random variables V_1 and V_2 are independent if $b = c = -a$ $(a \neq 0)$.

(ii) Under the condition in (i), write down the joint probability density function of V_1 and V_2.

[20 marks]

4. *Biarkan vektor rawak* $\mathbf{X} = (X_1, X_2, X_3)'$ *mempunyai taburan normal multivariat* $\mathrm{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *yang mana*

$$\boldsymbol{\mu} = (3, 4, 1)' \text{ dan } \boldsymbol{\Sigma} = \begin{pmatrix} 5 & 2 & 4 \\ 2 & 1 & 3 \\ 4 & 3 & 6 \end{pmatrix}.$$

Andaikan $V_1 = aX_1 + bX_2$ *dan* $V_2 = cX_2 + aX_3$, *yang mana a, b dan c adalah pemalar-pemalar.*

(i) Tunjukkan bahawa pembolehubah-pembolehubah rawak V_1 dan V_2 *adalah tak bersandar jika* $b = c = -a \ (a \neq 0)$.

(ii) Bawah syarat (i), tulis taburan kebarangkalian tercantum bagi V_1 dan V_2.

.

[20 markah]

5. The scores of two subjects for a sample of 30 first year students at the School of Mathematical Sciences are obtained at the end of the semester. Let X_1 and X_2 denote the scores a student obtains for the subjects Linear Algebra and Elementary Statistics, respectively. The sample mean and the sample variance-covariance matrix obtained are as follows:

$$\bar{\mathbf{X}} = \begin{pmatrix} 60.2 \\ 66.5 \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 225 & 325 \\ 325 & 225 \end{pmatrix}.$$

(i) Use a multivariate test to test at the 5% level whether the mean scores on both subjects are equal to 65. Suppose 65 represent average score of the two subjects for students 5 years ago. Is there reason to believe that the scores now are different from the scores last five years? Explain. State any assumption you make before performing the multivariate test.

(ii) Obtain the 95% simultaneous confidence intervals for the two population means.

(iii) Obtain the 95% Bonferroni confidence intervals for the two population means and compare these intervals with the intervals in (ii) above.

[30 marks]

5. *Markah dua subjek bagi sampel 30 orang pelajar tahun pertama di Pusat Pengajian Sains Matematik diperolehi pada akhir semester. Biar* X_1 dan X_2 *masing-masing menunjukkan markah diperoleh pelajar untuk subjek Linear Algebra dan subjek Statistik Asas. Min sampel dan matriks varians-kovarians sampel yang diperolehi adalah seperti berikut:*

$$\bar{\mathbf{X}} = \begin{pmatrix} 60.2 \\ 66.5 \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 225 & 325 \\ 325 & 225 \end{pmatrix}.$$

(i) Gunakan suatu ujian multivariat untuk menguji pada tahap 5% sama ada min markah bagi kedua-dua subjek adalah sama dengan 65. Katakan 65 mewakili markah purata kedua-dua mata pelajaran untuk pelajar lima tahun yang lalu. Adakah terdapat sebab untuk mempercayai bahawa markah kini adalah berbeza daripada markah lima tahun lalu? Terangkan. Nyatakan sebarang andaian yang anda buat sebelum melaksanakan ujian multivariat tersebut.

(ii) Dapatkan selang keyakinan serentak 95% bagi dua min populasi tersebut.

(iii) Dapatkan selang keyakinan Bonferroni 95% bagi dua min populasi dan bandingkan selang-selang ini dengan selang-selang dalam (ii) di atas.

[30 markah]

6. The track records for men and women of a certain country are studied. 54 male and 54 female athletes of the country are randomly selected and the measurements on the set of two variables are compared:

X_1	Record time for 100 metres race (in seconds)
X_2	Record time for 200 metres race (in seconds)

The summary statistics of the observed record times for the two variables for men (M) and women (W) are as follows:

$$\bar{\mathbf{x}}_{\mathrm{M}} = \begin{pmatrix} 10.22 \\ 20.54 \end{pmatrix} \qquad \mathbf{S}_{\mathbf{M}} = \begin{pmatrix} 0.049 & 0.111 \\ 0.111 & 0.301 \end{pmatrix}$$

$$\bar{\mathbf{x}}_{\mathrm{W}} = \begin{pmatrix} 11.36 \\ 23.12 \end{pmatrix} \qquad \mathbf{S}_{\mathrm{W}} = \begin{pmatrix} 0.155 & 0.345 \\ 0.345 & 0.863 \end{pmatrix}$$

Use a multivariate test to test whether there is a difference in the population mean vectors between men and women. State any assumption you make.

[15 marks]

6. *Rekod trek bagi lelaki dan wanita sebuah negara dikaji. 54 atlit lelaki dan 54 atlit wanita negara tersebut dipilih secara rawak dan ukuran bagi set dua pembolehubah dibandingkan:*

X_1 *Rekod masa untuk 100 meter perlumbaan (dalam saat)*
X_2 *Rekod masa untuk 200 meter perlumbaan (dalam saat)*

Statistik-statistik ringkasan bagi cerapan rekod masa bagi dua pembolehubah untuk lelaki (M) dan wanita (W) adalah seperti berikut:

$$\bar{\mathbf{x}}_{\mathrm{M}} = \begin{pmatrix} 10.22 \\ 20.54 \end{pmatrix} \qquad \mathbf{S}_{\mathbf{M}} = \begin{pmatrix} 0.049 & 0.111 \\ 0.111 & 0.301 \end{pmatrix}$$

$$\bar{\mathbf{x}}_{\mathrm{W}} = \begin{pmatrix} 11.36 \\ 23.12 \end{pmatrix} \qquad \mathbf{S}_{\mathrm{W}} = \begin{pmatrix} 0.155 & 0.345 \\ 0.345 & 0.863 \end{pmatrix}$$

Guna suatu ujian multivariat untuk menguji sama ada terdapat perbezaan pada vektor-vektor min populasi antara lelaki dan perempuan. Nyatakan sebarang andaian yang dibuat.

[15 markah]

7. Observations on two responses are collected for three treatments. The observation vectors $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ are

Treatment 1: $\begin{bmatrix} 3 \\ 7 \end{bmatrix}, \begin{bmatrix} 2 \\ 5 \end{bmatrix}, \begin{bmatrix} 3 \\ 7 \end{bmatrix}, \begin{bmatrix} 4 \\ 9 \end{bmatrix}$

Treatment 2: $\begin{bmatrix} 6 \\ 7 \end{bmatrix}, \begin{bmatrix} 4 \\ 6 \end{bmatrix}, \begin{bmatrix} 5 \\ 8 \end{bmatrix}, \begin{bmatrix} 6 \\ 5 \end{bmatrix}, \begin{bmatrix} 4 \\ 4 \end{bmatrix}$

Treatment 3: $\begin{bmatrix} 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 4 \\ 6 \end{bmatrix}, \begin{bmatrix} 3 \\ 6 \end{bmatrix}$

Construct the one-way MANOVA table and test for treatment effects using $\alpha = 0.05$. Give your conclusion and state any assumption you make.

[20 marks]

7. *Cerapan untuk dua respon dikumpul bagi tiga rawatan. Vektor-vektor cerapan* $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ *adalah*

Rawatan 1: $\begin{bmatrix} 3 \\ 7 \end{bmatrix}, \begin{bmatrix} 2 \\ 5 \end{bmatrix}, \begin{bmatrix} 3 \\ 7 \end{bmatrix}, \begin{bmatrix} 4 \\ 9 \end{bmatrix}$

Rawatan 2: $\begin{bmatrix}6\\7\end{bmatrix}, \begin{bmatrix}4\\6\end{bmatrix}, \begin{bmatrix}5\\8\end{bmatrix}, \begin{bmatrix}6\\5\end{bmatrix}, \begin{bmatrix}4\\4\end{bmatrix}$

Rawatan 3: $\begin{bmatrix}2\\3\end{bmatrix}, \begin{bmatrix}4\\6\end{bmatrix}, \begin{bmatrix}3\\6\end{bmatrix}$

Bina jadual MANOVA satu-hala dan uji kesan rawatan menggunakan $\alpha = 0.05$. *Beri kesimpulan anda dan nyatakan sebarang andaian yang dibuat.*

[20 markah]

8. (a) What is the principal component analysis (PCA)? What are the objectives of PCA?

(b) Suppose $\mathbf{X} \sim N_2(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \begin{pmatrix}3 & 1\\1 & 2\end{pmatrix}$.

(i) Determine the principal components of $\boldsymbol{\Sigma}$.

(ii) Calculate the proportion of the total variance explained by the first principal component.

(iii) Give the joint distribution of the principal components, making sure you specify the mean vector and the variance-covariance matrix.

[25 marks]

8. (a) Apakah Analisis Komponen Prinsipal (AKP)? Apakah objektif-objektif AKP?

(b) Andaikan $\mathbf{X} \sim N_2(\mathbf{0}, \boldsymbol{\Sigma})$, *yang mana* $\boldsymbol{\Sigma} = \begin{pmatrix}3 & 1\\1 & 2\end{pmatrix}$.

(i) Tentukan komponen-komponen prinsipal bagi $\boldsymbol{\Sigma}$.

(ii) Hitung kadaran bagi jumlah varians yang diterangkan oleh komponen prinsipal pertama.

(iii) Beri taburan tercantum bagi komponen-komponen prinsipal, pastikan anda perincikan vektor min dan matriks varians-kovarians.

[25 markah]

9. Consider a dataset containing the national track records for men in 54 countries which have the measurements of eight variables:

X_1 Record time for 100 metres race (in seconds)
X_2 Record time for 200 metres race (in seconds)
X_3 Record time for 400 metres race (in seconds)

X_4 Record time for 800 metres race (in minutes)
X_5 Record time for 1500 metres race (in minutes)
X_6 Record time for 5000 metres race (in minutes)
X_7 Record time for 10,000 metres race (in minutes)
X_8 Record time for the marathon (in minutes)

Measurement on these eight variables provided the following summary statistics:

Variable	Mean	Variance
X1	10.217	0.0490
X2	20.541	0.301
X3	45.829	2.070
X4	1.7681	0.00275
X5	3.6533	0.0230
X6	13.618	0.579
X7	28.535	2.820
X8	133.48	80.14

A principal component analysis of the data has been carried out using Minitab yielding the following output:

Principal Component Analysis: X1, X2, X3, X4, X5, X6, X7, X8

Eigenanalysis of the xxxxxxxx Matrix

Eigenvalue	6.7033	0.6384	0.2275	0.2058	0.0976	0.0707	0.0469	0.0097
Proportion	0.838	0.080	0.028	0.026	0.012	0.009	0.006	0.001
Cumulative	0.838	0.918	0.946	0.972	0.984	0.993	0.999	1.000

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
X1	0.332	0.529	0.344	-0.381	0.300	-0.362	0.348	-0.066
X2	0.346	0.470	-0.004	-0.217	-0.541	0.349	-0.440	0.061
X3	0.339	0.345	-0.067	0.851	0.133	0.077	0.114	-0.003
X4	0.353	-0.089	-0.783	-0.134	-0.227	-0.341	0.259	-0.039
X5	0.366	-0.154	-0.244	-0.233	0.652	0.530	-0.147	-0.040
X6	0.370	-0.295	0.183	0.055	0.072	-0.359	-0.328	0.706
X7	0.366	-0.334	0.244	0.087	-0.061	-0.273	-0.351	-0.697
X8	0.354	-0.387	0.335	-0.018	-0.338	0.375	0.594	0.069

(i) Has the principal component analysis been performed on the covariance or the correlation matrix? Justify your answer and explain why this matrix has been used in this analysis.

(ii) Construct a scree plot using the information given in the output.

(iii) How many principal components do you think provide an adequate explanation of the data? Justify your answer.

(iv) How would you interpret the first two principal components?

[20 marks]

9. *Pertimbangkan suatu set data yang mengandungi rekod trek kebangsaan bagi lelaki dari 54 negara yang mempunyai ukuran-ukuran bagi lapan pembolehubah:*

X_1 *Rekod masa bagi 100 meter perlumbaan (dalam saat)*
X_2 *Rekod masa bagi 200 meter perlumbaan (dalam saat)*
X_3 *Rekod masa bagi 400 meter perlumbaan (dalam saat)*
X_4 *Rekod masa bagi 800 meter perlumbaan (dalam minit)*
X_5 *Rekod masa bagi 1500 meter perlumbaan (dalam minit)*
X_6 *Rekod masa bagi 5,000 meter perlumbaan (dalam minit)*
X_7 *Rekod masa bagi 10,000 meter perlumbaan (dalam minit)*
X_8 *Rekod masa bagi marathon (dalam minit)*

Ukuran pada lapan pembolehubah ini menghasilkan statistik ringkasan berikut:

P/ubah	Min	Varians
X1	10.217	0.0490
X2	20.541	0.301
X3	45.829	2.070
X4	1.7681	0.00275
X5	3.6533	0.0230
X6	13.618	0.579
X7	28.535	2.820
X8	133.48	80.14

Suatu analisis komponen prinsipal yang dilakukan menggunakan Minitab menghasilkan output berikut:

Principal Component Analysis: X1, X2, X3, X4, X5, X6, X7, X8

Eigenanalysis of the xxxxxxxx Matrix

Eigenvalue	6.7033	0.6384	0.2275	0.2058	0.0976	0.0707	0.0469	0.0097
Proportion	0.838	0.080	0.028	0.026	0.012	0.009	0.006	0.001
Cumulative	0.838	0.918	0.946	0.972	0.984	0.993	0.999	1.000

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
X1	0.332	0.529	0.344	-0.381	0.300	-0.362	0.348	-0.066
X2	0.346	0.470	-0.004	-0.217	-0.541	0.349	-0.440	0.061
X3	0.339	0.345	-0.067	0.851	0.133	0.077	0.114	-0.003
X4	0.353	-0.089	-0.783	-0.134	-0.227	-0.341	0.259	-0.039
X5	0.366	-0.154	-0.244	-0.233	0.652	0.530	-0.147	-0.040
X6	0.370	-0.295	0.183	0.055	0.072	-0.359	-0.328	0.706
X7	0.366	-0.334	0.244	0.087	-0.061	-0.273	-0.351	-0.697
X8	0.354	-0.387	0.335	-0.018	-0.338	0.375	0.594	0.069

(i) *Adakah analisis komponen prisipal dilakukan pada matriks kovarians atau matriks korelasi? Justifikasikan jawapan anda dan jelaskan mengapa matriks ini telah digunakan dalam analisis ini.*

(ii) *Bina plot scree meggunakan maklumat diberi dalam output.*

(iii) *Berapa komponen prinsipal yang anda fikir memberi penjelasan yang cukup untuk data tersebut? Justifikasikan jawapan anda.*

(iv) *Bagaimana anda akan mentafsir dua komponen prinsipal pertama?*

[20 markah]

10. The following data measure the amount of protein consumed for nine food groups in 25 European countries. The nine food groups are red meat (X1), white meat (X2), eggs (X3), milk (X4), fish (X5), cereal (X6), starch (X7), nuts (X8), and fruits and vegetables (X9).

County	X1	X2	X3	X4	X5	X6	X7	X8	X9
Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Austria	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3
Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0
Bulgaria	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Czechoslovakia	9.7	11.4	2.8	12.5	2.0	34.3	5.0	1.1	4.0
Denmark	10.6	10.8	3.7	25.0	9.9	21.9	4.8	0.7	2.4
E Germany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1.0	1.4
France	18.0	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Greece	10.2	3.0	2.8	17.6	5.9	41.7	2.2	7.8	6.5
Hungary	5.3	12.4	2.9	9.7	0.3	40.1	4.0	5.4	4.2
Ireland	13.9	10.0	4.7	25.8	2.2	24.0	6.2	1.6	2.9
Italy	9.0	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Netherlands	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Norway	9.4	4.7	2.7	23.3	9.7	23.0	4.6	1.6	2.7
Poland	6.9	10.2	2.7	19.3	3.0	36.1	5.9	2.0	6.6
Portugal	6.2	3.7	1.1	4.9	14.2	27.0	5.9	4.7	7.9
Romania	6.2	6.3	1.5	11.1	1.0	49.6	3.1	5.3	2.8
Spain	7.1	3.4	3.1	8.6	7.0	29.2	5.7	5.9	7.2
Sweden	9.9	7.8	3.5	4.7	7.5	19.5	3.7	1.4	2.0
Switzerland	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
UK	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
USSR	9.3	4.6	2.1	16.6	3.0	43.6	6.4	3.4	2.9
W Germany	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
Yugoslavia	4.4	5.0	1.2	9.5	0.6	55.9	3.0	5.7	3.2

Hierarchical Clustering analysis is performed using Minitab to determine if these 25 countries can be formed into groups suggested by the data. The results for two methods, single linkage and complete linkage are displayed as follows:

Cluster Analysis of Observations: C2, C3, C4, C5, C6, C7, C8, C9

Standardized Variables, Euclidean Distance, Single Linkage
Amalgamation Steps

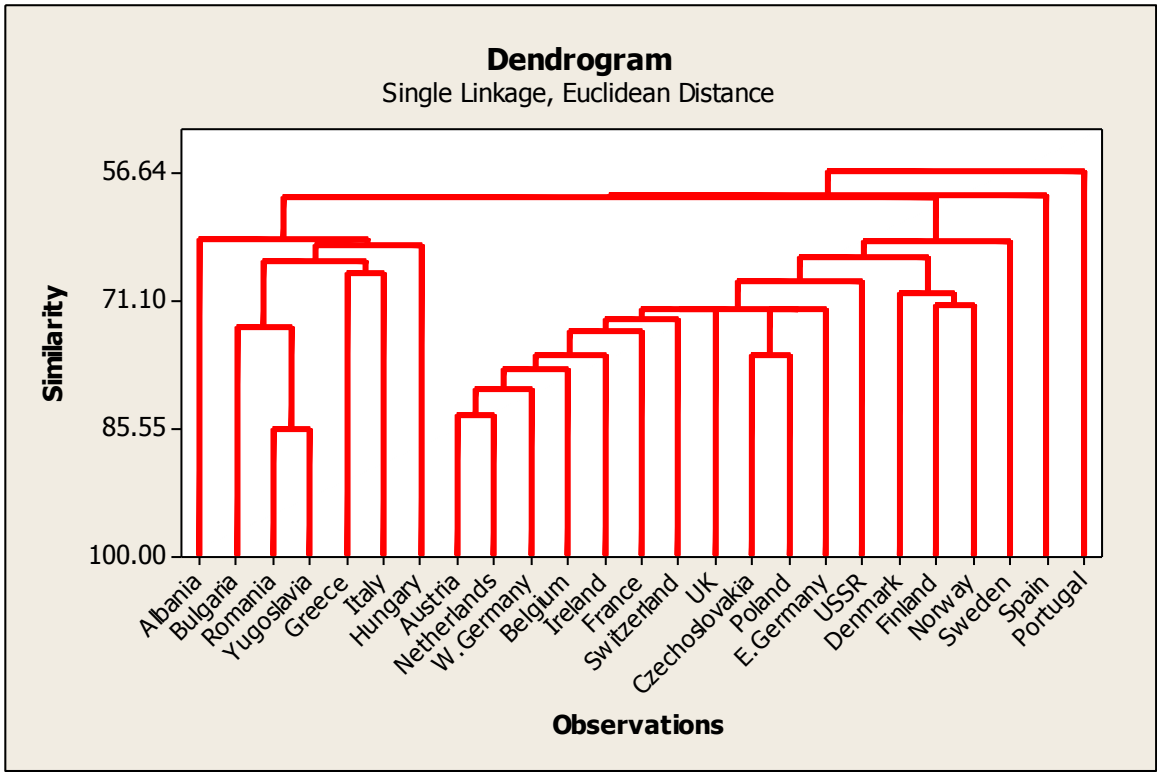
Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	24	85.6774	0.95934	18	25	18	2
2	23	84.0868	1.06587	2	14	2	2
3	22	81.1669	1.26145	2	24	2	3
4	21	78.9061	1.41288	2	3	2	4
5	20	77.4170	1.51262	2	12	2	5
6	19	77.3863	1.51467	5	16	5	2
7	18	74.6201	1.69996	2	9	2	6
8	17	74.3227	1.71988	4	18	4	3
9	16	73.4123	1.78085	2	21	2	7
10	15	72.2171	1.86091	2	22	2	8
11	14	72.1337	1.86650	5	7	5	3
12	13	72.0946	1.86912	2	5	2	11
13	12	71.6655	1.89786	8	15	8	2
14	11	70.3522	1.98582	6	8	6	3

...12/-

15	10	69.0642	2.07209	2	23	2	12
16	9	68.0072	2.14289	10	13	10	2
17	8	66.6629	2.23293	4	10	4	5
18	7	66.3732	2.25234	2	6	2	15
19	6	64.9799	2.34566	4	11	4	6
20	5	64.4307	2.38244	2	20	2	16
21	4	64.2885	2.39198	1	4	1	7
22	3	59.4603	2.71537	1	2	1	23
23	2	59.3892	2.72013	1	19	1	24
24	1	56.6429	2.90408	1	17	1	25

Final Partition
Number of clusters: 1

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	25	192	2.68208	4.23844



Cluster Analysis of Observations: C2, C3, C4, C5, C6, C7, C8, C9

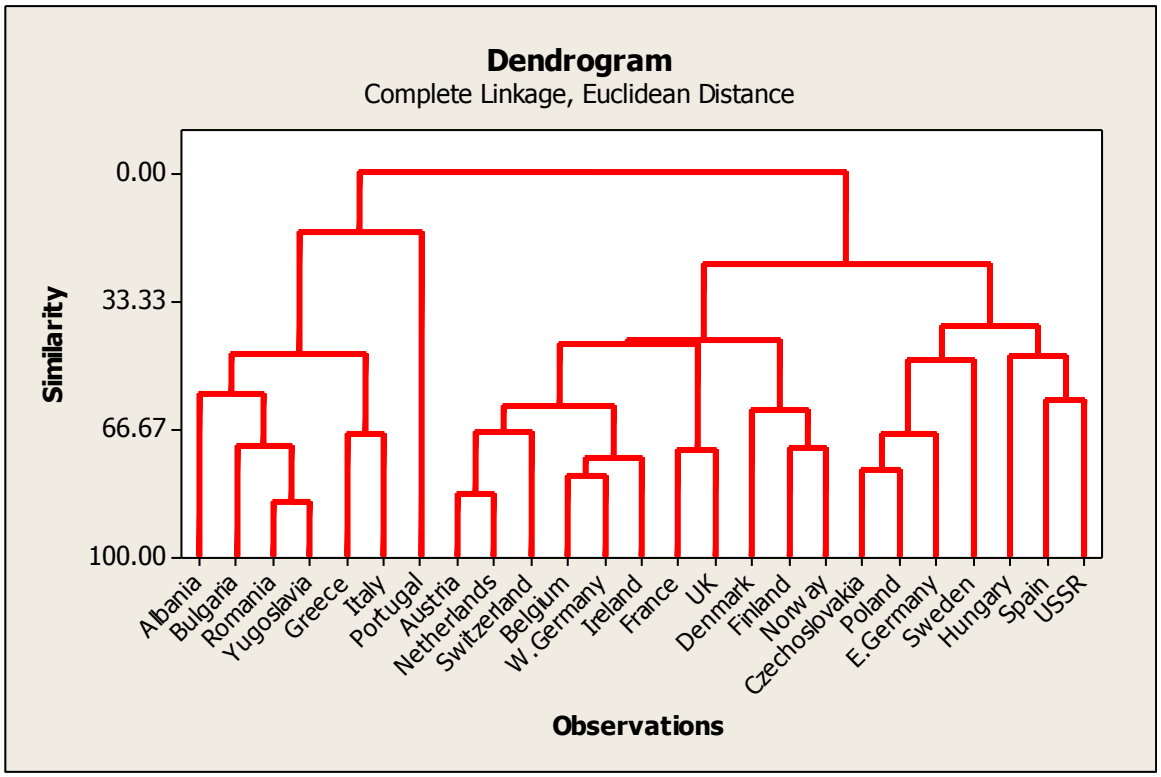
Standardized Variables, Euclidean Distance, Complete Linkage
Amalgamation Steps

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	24	85.6774	0.95934	18	25	18	2
2	23	84.0868	1.06587	2	14	2	2
3	22	78.9061	1.41288	3	24	3	2
4	21	77.3863	1.51467	5	16	5	2
5	20	74.3238	1.71980	3	12	3	3
6	19	72.2171	1.86091	9	22	9	2
7	18	71.6655	1.89786	8	15	8	2
8	17	71.4268	1.91385	4	18	4	3
9	16	68.3662	2.11885	5	7	5	3
10	15	68.0072	2.14289	10	13	10	2
11	14	67.5562	2.17310	2	21	2	3
12	13	61.8354	2.55628	6	8	6	3
13	12	60.8772	2.62046	2	3	2	6
14	11	59.3892	2.72013	19	23	19	2
15	10	57.8634	2.82233	1	4	1	4
16	9	49.0684	3.41142	5	20	5	4
17	8	47.6629	3.50556	11	19	11	3
18	7	47.3729	3.52499	1	10	1	6
19	6	44.6632	3.70648	2	9	2	8
20	5	43.7037	3.77075	2	6	2	11
21	4	39.7524	4.03541	5	11	5	7
22	3	23.8962	5.09746	2	5	2	18
23	2	15.6952	5.64677	1	17	1	7
24	1	0.0000	6.69805	1	2	1	25

Final Partition
Number of clusters: 1

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	25	192	2.68208	4.23844

...14/-



Interpret the results and compare the two methods.

[20 marks]

10. *Data berikut mengukur jumlah protein yang dimakan bagi sembilan kumpulan makanan di 25 buah negara Eropah. Sembilan kumpulan makanan tersebut adalah daging merah (X1), daging putih (X2), telur (X3), susu (X4), ikan (X5), bijirin (X6), kanji (X7), kekacang (X8), dan buah-buahan dan sayur-sayuran (X9).*

Negara	X1	X2	X3	X4	X5	X6	X7	X8	X9
Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Austria	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3
Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0
Bulgaria	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Czechoslovakia	9.7	11.4	2.8	12.5	2.0	34.3	5.0	1.1	4.0
Denmark	10.6	10.8	3.7	25.0	9.9	21.9	4.8	0.7	2.4
E Germany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1.0	1.4
France	18.0	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Greece	10.2	3.0	2.8	17.6	5.9	41.7	2.2	7.8	6.5
Hungary	5.3	12.4	2.9	9.7	0.3	40.1	4.0	5.4	4.2
Ireland	13.9	10.0	4.7	25.8	2.2	24.0	6.2	1.6	2.9
Italy	9.0	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Netherlands	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Norway	9.4	4.7	2.7	23.3	9.7	23.0	4.6	1.6	2.7
Poland	6.9	10.2	2.7	19.3	3.0	36.1	5.9	2.0	6.6

Portugal	*6.2*	*3.7*	*1.1*	*4.9*	*14.2*	*27.0*	*5.9*	*4.7*	*7.9*
Romania	*6.2*	*6.3*	*1.5*	*11.1*	*1.0*	*49.6*	*3.1*	*5.3*	*2.8*
Spain	*7.1*	*3.4*	*3.1*	*8.6*	*7.0*	*29.2*	*5.7*	*5.9*	*7.2*
Sweden	*9.9*	*7.8*	*3.5*	*4.7*	*7.5*	*19.5*	*3.7*	*1.4*	*2.0*
Switzerland	*13.1*	*10.1*	*3.1*	*23.8*	*2.3*	*25.6*	*2.8*	*2.4*	*4.9*
UK	*17.4*	*5.7*	*4.7*	*20.6*	*4.3*	*24.3*	*4.7*	*3.4*	*3.3*
USSR	*9.3*	*4.6*	*2.1*	*16.6*	*3.0*	*43.6*	*6.4*	*3.4*	*2.9*
W Germany	*11.4*	*12.5*	*4.1*	*18.8*	*3.4*	*18.6*	*5.2*	*1.5*	*3.8*
Yugoslavia	*4.4*	*5.0*	*1.2*	*9.5*	*0.6*	*55.9*	*3.0*	*5.7*	*3.2*

Analisis Pengelompokan Hierarki dilakukan menggunakan Minitab untuk menentukan jika 25 negara ini boleh dibentuk ke dalam kumpulan-kumpulan yang dicadangkan oleh data. Keputusan bagi dua kaedah, pautan tunggal dan pautan lengkap dipaparkan seperti berikut:

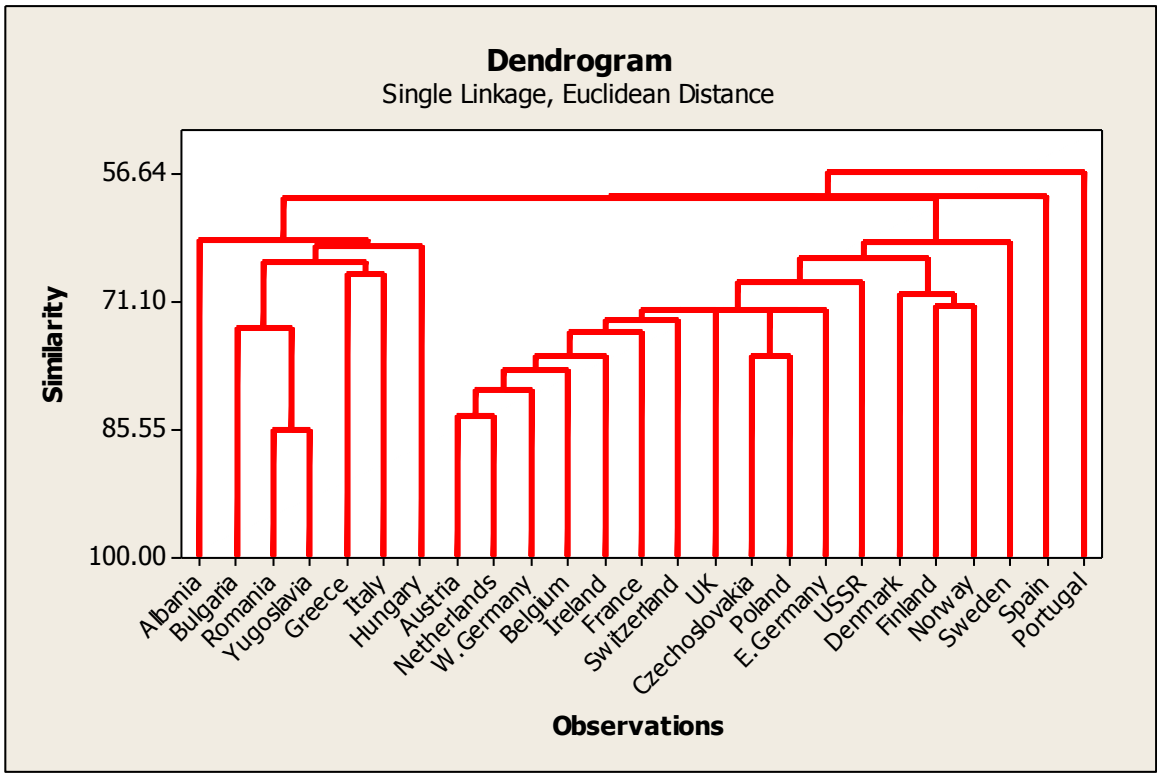
Cluster Analysis of Observations: C2, C3, C4, C5, C6, C7, C8, C9

Standardized Variables, Euclidean Distance, Single Linkage
Amalgamation Steps

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	24	85.6774	0.95934	18	25	18	2
2	23	84.0868	1.06587	2	14	2	2
3	22	81.1669	1.26145	2	24	2	3
4	21	78.9061	1.41288	2	3	2	4
5	20	77.4170	1.51262	2	12	2	5
6	19	77.3863	1.51467	5	16	5	2
7	18	74.6201	1.69996	2	9	2	6
8	17	74.3227	1.71988	4	18	4	3
9	16	73.4123	1.78085	2	21	2	7
10	15	72.2171	1.86091	2	22	2	8
11	14	72.1337	1.86650	5	7	5	3
12	13	72.0946	1.86912	2	5	2	11
13	12	71.6655	1.89786	8	15	8	2
14	11	70.3522	1.98582	6	8	6	3
15	10	69.0642	2.07209	2	23	2	12
16	9	68.0072	2.14289	10	13	10	2
17	8	66.6629	2.23293	4	10	4	5
18	7	66.3732	2.25234	2	6	2	15
19	6	64.9799	2.34566	4	11	4	6
20	5	64.4307	2.38244	2	20	2	16
21	4	64.2885	2.39198	1	4	1	7
22	3	59.4603	2.71537	1	2	1	23
23	2	59.3892	2.72013	1	19	1	24
24	1	56.6429	2.90408	1	17	1	25

Final Partition
Number of clusters: 1

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	25	192	2.68208	4.23844



Cluster Analysis of Observations: C2, C3, C4, C5, C6, C7, C8, C9

Standardized Variables, Euclidean Distance, Complete Linkage
Amalgamation Steps

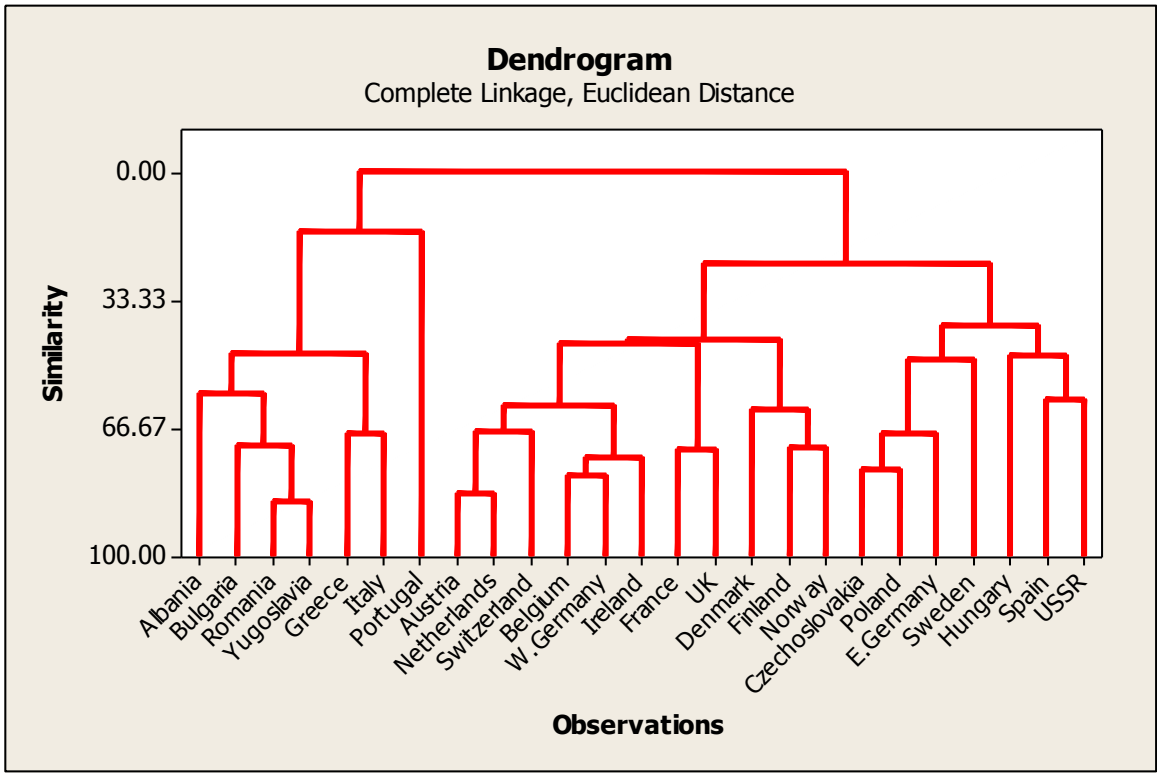
Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	24	85.6774	0.95934	18	25	18	2
2	23	84.0868	1.06587	2	14	2	2
3	22	78.9061	1.41288	3	24	3	2
4	21	77.3863	1.51467	5	16	5	2
5	20	74.3238	1.71980	3	12	3	3
6	19	72.2171	1.86091	9	22	9	2
7	18	71.6655	1.89786	8	15	8	2
8	17	71.4268	1.91385	4	18	4	3
9	16	68.3662	2.11885	5	7	5	3
10	15	68.0072	2.14289	10	13	10	2
11	14	67.5562	2.17310	2	21	2	3
12	13	61.8354	2.55628	6	8	6	3
13	12	60.8772	2.62046	2	3	2	6
14	11	59.3892	2.72013	19	23	19	2
15	10	57.8634	2.82233	1	4	1	4
16	9	49.0684	3.41142	5	20	5	4
17	8	47.6629	3.50556	11	19	11	3
18	7	47.3729	3.52499	1	10	1	6
19	6	44.6632	3.70648	2	9	2	8
20	5	43.7037	3.77075	2	6	2	11
21	4	39.7524	4.03541	5	11	5	7

...18/-

22	3	23.8962	5.09746	2	5	2	18
23	2	15.6952	5.64677	1	17	1	7
24	1	0.0000	6.69805	1	2	1	25

Final Partition
Number of clusters: 1

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	25	192	2.68208	4.23844



Tafsirkan keputusan dan bandingkan kedua-dua kaedah.

[20 markah]

APPENDIX/LAMPIRAN

MSG366: FORMULAE SHEET

1. Suppose $\mathbf{X}$ has $E(\mathbf{X}) = \boldsymbol{\mu}$ and $\text{Cov}(\mathbf{X}) = \boldsymbol{\Sigma}$. Thus $\mathbf{c}'\mathbf{X}$ has mean $\mathbf{c}'\boldsymbol{\mu}$ and variance $\mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}$.

2. Bivariate normal p.d.f.:

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}\ 1-\rho_{12}{}^2}} \exp\left\{ -\frac{1}{2\ 1-\rho_{12}{}^2} \left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}} \right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}} \right)^2 - 2\rho_{12} \left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}} \right) \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}} \right) \right] \right\}$$

3. Multivariate normal p.d.f.:

$$f(\mathbf{x}) = \frac{1}{2\pi^{\ p/2}\ |\boldsymbol{\Sigma}|^{1/2}} e^{-(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2}$$

4. If $\mathbf{X} \sim \text{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then
   (a) $\mathbf{a}'\mathbf{X} \sim N\ \mathbf{a}'\boldsymbol{\mu},\ \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$
   (b) $\mathbf{A}\mathbf{X} \sim \text{N}_q\ \mathbf{A}\boldsymbol{\mu},\ \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$
   (c) $\mathbf{X}+\mathbf{d} \sim \text{N}_p\ \boldsymbol{\mu}+\mathbf{d},\ \boldsymbol{\Sigma}$
   (d) $\mathbf{A}\mathbf{X}+\mathbf{d} \sim \text{N}_q\ \mathbf{A}\boldsymbol{\mu}+\mathbf{d},\ \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$
   (e) $\mathbf{X}-\boldsymbol{\mu}\ '\boldsymbol{\Sigma}^{-1}\ \mathbf{X}-\boldsymbol{\mu} \sim \chi_p^2$

5. Let $\mathbf{X}_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$, $j = 1, ..., n$ be mutually independent. Then $\mathbf{V}_1 = \sum_{j=1}^{n} c_j \mathbf{X}_j \sim N_p\left(\sum_{j=1}^{n} c_j \boldsymbol{\mu}_j, \left(\sum_{j=1}^{n} c_j^{\,2}\right)\boldsymbol{\Sigma}\right)$. Moreover, $\mathbf{V}_1$ and $\mathbf{V}_2 = \sum_{j=1}^{n} b_j \mathbf{X}_j$ are jointly multivariate normal with covariance matrix $\begin{pmatrix} \left(\sum_{j=1}^{n} c_j^{\,2}\right)\boldsymbol{\Sigma} & \mathbf{b}'\mathbf{c}\ \boldsymbol{\Sigma} \\ \mathbf{b}'\mathbf{c}\ \boldsymbol{\Sigma} & \left(\sum_{j=1}^{n} b_j^{\,2}\right)\boldsymbol{\Sigma} \end{pmatrix}$.

6. Let $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \hdashline \mathbf{X}_2 \end{bmatrix}$ be distributed as $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \hdashline \boldsymbol{\mu}_2 \end{bmatrix}$, $\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right]$ and $|\boldsymbol{\Sigma}_{22}| > 0$. Then the conditional distribution of $\mathbf{X}_1$, given that $\mathbf{X}_2 = \mathbf{x}_2$, is normal and has mean = $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$ and covariance = $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.

7. One-sample results:

(a) $T^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu})$

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{X}_j, \quad \mathbf{S} = \frac{1}{n-1}\sum_{j=1}^{n} (\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})'$$

$$T^2 \sim \frac{(n-1)p}{n-p} F_{p, n-p}$$

(b) $100(1-\alpha)\%$ simultaneous confidence intervals for $\mathbf{a}'\boldsymbol{\mu}$:

$$\mathbf{a}'\bar{\mathbf{X}} \pm \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p}(\alpha)\, \mathbf{a}'\mathbf{S}\mathbf{a}}$$

(c) $100(1-\alpha)\%$ Bonferroni confidence interval for μ_i:

$$\bar{x}_i \pm t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{ii}}{n}}$$

(d) $100(1-\alpha)\%$ large sample confidence interval for μ_i:

$$\bar{x}_i \pm \sqrt{\chi_p^2(\alpha)}\sqrt{\frac{s_{ii}}{n}}$$

8. Two-sample results (Paired comparisons):

(a) $T^2 = n\ \bar{\mathbf{D}} - \boldsymbol{\delta}\ '\boldsymbol{\delta}_d^{-1}\ \bar{\mathbf{D}} - \boldsymbol{\delta}$

$$\bar{\mathbf{D}} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{D}_j, \quad \mathbf{S}_d = \frac{1}{n-1}\sum_{j=1}^{n}\ \mathbf{D}_j - \bar{\mathbf{D}}\ \ \mathbf{D}_j - \bar{\mathbf{D}}\ '$$

$$T^2 \square \frac{n-1\ \ p}{n-p} F_{p,\,n-p}$$

(b) $100(1-\alpha)\%$ simultaneous confidence intervals for δ_i:

$$\bar{d}_i \quad \pm \quad \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p}(\alpha)}\ \sqrt{\frac{s_{d_i}^2}{n}}$$

9. Two-sample results (Independent samples):

(a) $$T^2 = \left[\ \mathbf{X}_1 - \mathbf{X}_2\ -\ \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\ \right]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_p\right]^{-1} \left[\ \mathbf{X}_1 - \mathbf{X}_2\ -\ \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\ \right]$$

$$T^2 \square \frac{n_1 + n_2 - 2\ \ p}{n_1 + n_2 - p - 1} F_{p,\,n_1+n_2-p-1}$$

$$\mathbf{S}_p = \frac{n_1 - 1\ \mathbf{S}_1 + \ \ n_2 - 1\ \mathbf{S}_2}{n_1 + n_2 - 2}$$

$$\mathbf{S}_i = \frac{\sum_{j=1}^{n_i}\ \mathbf{x}_{ij} - \bar{\mathbf{x}}_i\ \ \ \mathbf{x}_{ij} - \bar{\mathbf{x}}_i\ '}{n_i - 1}$$

(b) $100(1-\alpha)\%$ simultaneous confidence interval for $\mathbf{a}'\ \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$:

$$\mathbf{a}'\ \bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2\ \ \pm\ c\sqrt{\mathbf{a}'\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_p\mathbf{a}}$$

$$c^2 = \frac{n_1 + n_2 - 2\ \ p}{n_1 + n_2 - p - 1} F_{p,\,n_1+n_2-p-1}(\alpha)$$

(c) For large $n_1 - p$ and $n_2 - p$, an approximate $100(1-\alpha)\%$ simultaneous confidence interval for $\mathbf{a}'\ \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$:

$$\mathbf{a}'\ \bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 \ \pm\ c\sqrt{\mathbf{a}'\left(\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right)\mathbf{a}}$$

$$c^2 = \chi_p^2(\alpha)$$

10. One-way MANOVA

$$\mathbf{B} = \sum_{l=1}^{g} n_l\ \ \bar{\mathbf{x}}_l - \bar{\mathbf{x}}\ \ \ \bar{\mathbf{x}}_l - \bar{\mathbf{x}}\ '$$

$$\mathbf{W} = \sum_{l=1}^{g}\sum_{j=1}^{n_l}\ \mathbf{x}_{lj} - \bar{\mathbf{x}}_l\ \ \ \mathbf{x}_{lj} - \bar{\mathbf{x}}_l\ ' = (n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 + \ldots + (n_g - 1)\mathbf{S}_g$$

$$\Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|}$$

Distribution of Λ^* :

For $p = 1,\ g \geq 2$: $\left(\frac{n-g}{g-1}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{g-1,n-g}$

For $p = 2,\ g \geq 2$: $\left(\frac{n-g-1}{g-1}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2(g-1),2(n-g-1)}$

For $p \geq 1,\ g = 2$: $\left(\frac{n-p-1}{p}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{p,n-p-1}$

For $p \geq 1,\ g = 3$: $\left(\frac{n-p-2}{p}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2p,\,2(n-p-2)}$

$n = \sum n_l$

- ooo O ooo -