

**THE UNIVERSITY OF CHICAGO**

**Graduate School of Business**

Business 41912, Spring Quarter 2006, Mr. Ruey S. Tsay

**Solutions to Final Exam**

1. (20 pts) Consider the monthly excess stock returns of ten U.S. companies. The companies are (1) A.G. Edwards (AGE), (2) Citigroup (C), (3) Morgan Stanley (MWD), (4) Merrill Lynch (MER), (5) Dell, (6) Hewlett-Packard (HPQ), (7) IBM, (8) Alcoa (AA), (9) Caterpillar (CAT), and (10) Procter & Gamble (PG). These companies can roughly be classified into three industrial categories, namely “financial” (1-4), “high tech and computer” (5-7), and “others” (8-10). The sample period is from 1990 to 2003 for 168 observations. The sample correlation matrix of the returns is

	AGE	C	MWD	MER	Dell	HPQ	IBM	AA	CAT	PG
AGE	1.00	0.63	0.62	0.64	0.29	0.31	0.27	0.30	0.28	0.18
C	0.63	1.00	0.71	0.68	0.25	0.37	0.39	0.38	0.39	0.29
MWD	0.62	0.71	1.00	0.80	0.26	0.46	0.37	0.40	0.27	0.27
MER	0.64	0.68	0.80	1.00	0.23	0.47	0.31	0.37	0.28	0.27
Dell	0.29	0.25	0.26	0.23	1.00	0.45	0.36	0.33	0.11	0.10
HPQ	0.31	0.37	0.46	0.47	0.45	1.00	0.45	0.51	0.23	0.08
IBM	0.27	0.39	0.37	0.31	0.36	0.45	1.00	0.41	0.34	-0.01
AA	0.30	0.38	0.40	0.37	0.33	0.51	0.41	1.00	0.60	0.06
CAT	0.28	0.39	0.27	0.28	0.11	0.23	0.34	0.60	1.00	0.13
PG	0.18	0.30	0.27	0.27	0.10	0.08	-0.01	0.06	0.13	1.00

- (8 pts) Use the single linkage method to perform the clustering analysis. Show details of the first update of the distance matrix. Draw the dendrogram.

Answer: Since correlation measures similarity, we use negative correlation to represent distance. The minimum distance is  $-0.80$  between (3) and (4). Thus, we put MER and MWD into a group. The first update of the distance matrix is

	V1	V2 (V3V4)	V5	V6	V7	V8	V9	
V2	-0.63							
V3V4	-0.64	-0.71						
V5	-0.29	-0.25	-0.26					
V6	-0.31	-0.37	-0.47	-0.45				
V7	-0.27	-0.39	-0.37	-0.36	-0.45			
V8	-0.30	-0.38	-0.40	-0.33	-0.51	-0.41		
V9	-0.28	-0.39	-0.28	-0.11	-0.23	-0.34	-0.60	
V10	-0.18	-0.30	-0.27	-0.10	-0.08	0.01	-0.06	-0.13

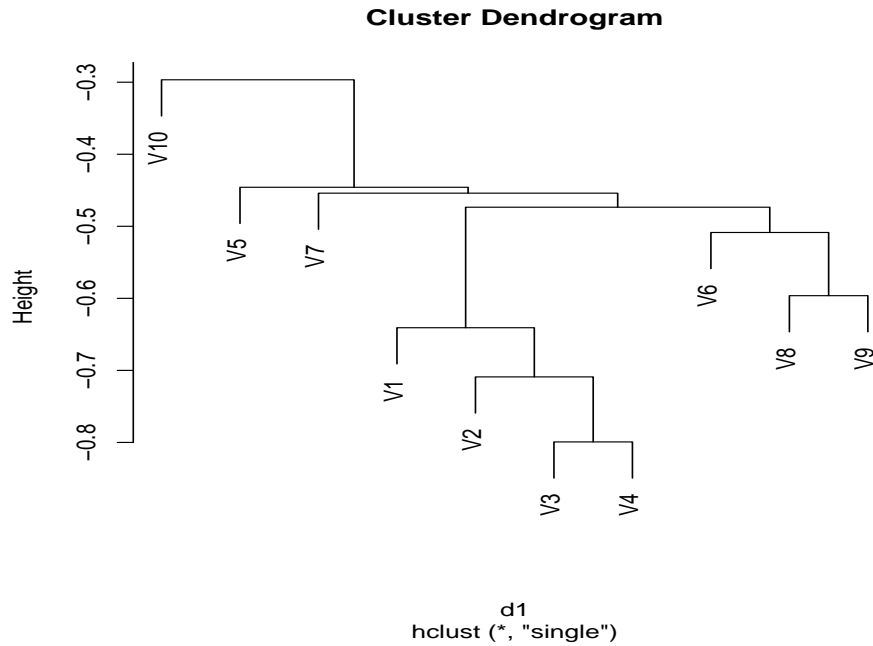


Figure 1: Dendrogram based on Single linkage

The dendrogram is shown in Figure 1.

- (8 pts) Use the complete linkage method to perform the clustering analysis. Show details of the first update of the distance matrix. Draw the dendrogram.

Answer: The first update of the distance matrix is

	V1	V2	V3V4	V5	V6	V7	V8	V9
V2	-0.63							
V3V4	-0.62	-0.68						
V5	-0.29	-0.25	-0.23					
V6	-0.31	-0.37	-0.46	-0.45				
V7	-0.27	-0.39	-0.31	-0.36	-0.45			
V8	-0.30	-0.38	-0.37	-0.33	-0.51	-0.41		
V9	-0.28	-0.39	-0.27	-0.11	-0.23	-0.34	-0.60	
V10	-0.18	-0.30	-0.27	-0.10	-0.08	0.01	-0.06	-0.13

The dendrogram is shown in Figure 2.

- Compare the two linkage methods. Is there any difference? Which method produces results that are close to the industrial categories?

Answer: Yes, the two dendrograms are different. The complete linkage provides results that are closer to the common industrial categories.

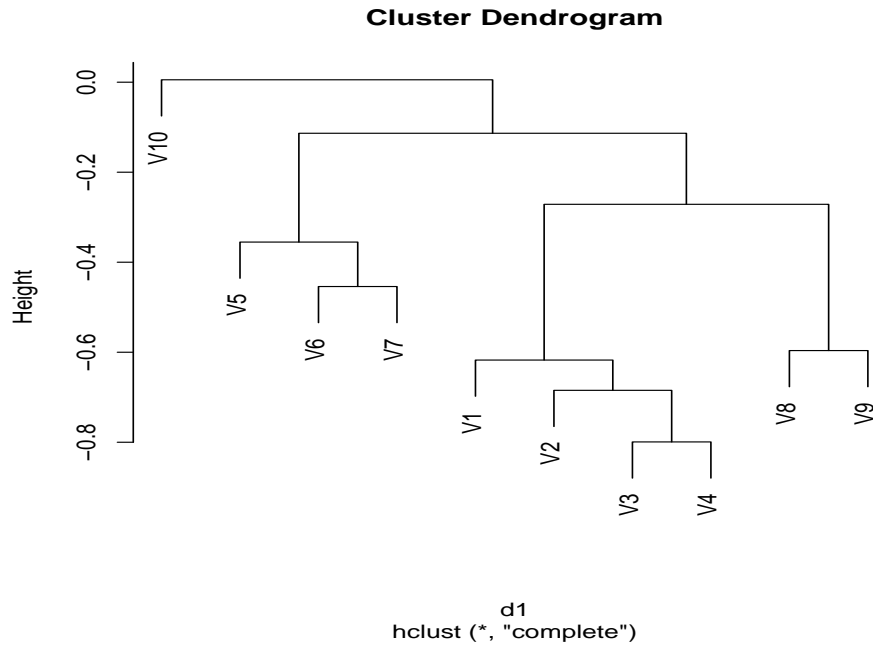


Figure 2: Dendrogram based on Complete Linkage

2. (15 pts) Again, consider the 10 monthly stock returns of Problem 1. The returns are standardized so that the variance of each return series is one. The attached output shows some selected results of factor analysis. First, the 2-factor model is rejected by the maximum likelihood method. The test statistic is 72.93 with p-value  $2.46 \times 10^{-6}$ . Consider, next, a 3-factor model. Use the output provided to answer the following questions:

- Obtain the *uniquenesses* for the fitted 3-factor model.

Answer: Using  $\mathbf{I} - \mathbf{L}\mathbf{L}' = \mathbf{\Psi}$ , we obtain the uniquenesses as 0.479

- Write down the three vectors of factor loadings.

Answer: The three loadings are

	Factor1	Factor2	Factor3
[1,]	0.639	0.297	-0.158
[2,]	0.686	0.406	-0.156
[3,]	0.838	0.292	-0.107
[4,]	0.824	0.296	-0.132
[5,]	0.344	0.123	0.420
[6,]	0.544	0.244	0.547
[7,]	0.352	0.352	0.338
[8,]	0.294	0.606	0.361

[9,]                    0.997  
 [10,]   0.250    0.137   -0.184

- Perform a maximum likelihood test to show that the three-factor model is not rejected at the 5% significance level. What is the test statistic? What is the reference distribution?

Answer: Using the large sample test (p. 499 of the text) with Bartlett's correction, we have the test statistic

$$T = [n - 1 - (2p + 4m + 5)/5] \ln \frac{|\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\Psi}|}{|\mathbf{S}_n|}.$$

In addition, using Eq. (9.41), we further obtain

$$T = [n - 1 - (2p + 4m + 5)/6] \ln \frac{|\hat{\mathbf{L}}_z\hat{\mathbf{L}}_z' + \hat{\Psi}_z|}{|\mathbf{R}|} \approx 22.72.$$

The reference distribution is  $\chi_{18}^2$ , and the p-value is 0.20. Thus, we cannot reject the null hypothesis at the 5% level.

Notice that, the returns are standardized and the determinant of  $\mathbf{R}$  is given. One needs only to compute the determinant of  $\hat{\mathbf{L}}_z\hat{\mathbf{L}}_z' + \hat{\Psi}_z$ .

3. (15 pts) Consider a study reported by C.R. Rao concerning head measurements of the first and second adult sons in a sample of 25 families. The measurements are

- $x_1$ : head length of the first son
- $x_2$ : head breadth of the first son
- $x_3$ : head length of the second son
- $x_4$ : head breadth of the second son.

The sample mean and covariance matrix are

$$\bar{\mathbf{x}} = \begin{bmatrix} 185.72 \\ 151.12 \\ 183.84 \\ 149.24 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} 1.0000 & 0.7346 & 0.7108 & 0.7040 \\ 0.7346 & 1.0000 & 0.6932 & 0.7086 \\ 0.7108 & 0.6932 & 1.0000 & 0.8392 \\ 0.7040 & 0.7086 & 0.8392 & 1.0000 \end{bmatrix}.$$

Answer the following questions:

- Obtain the canonical correlations and the associated canonical variates of head measurements of the two sons.

Answer: Details of the calculation is given below. The two canonical correlations are 0.789 and 0.054, respectively. The eigenvectors for the first son are

```

          [,1]      [,2]
[1,] 0.7149442 0.6991815
[2,] 0.6991815 -0.7149442

```

The eigenvectors for the second son are

```

          [,1]      [,2]
[1,] 0.5522325 0.03841978
[2,] 0.5628992 -0.03769174

```

- Assume normality. Test the hypothesis that the measurements between the two sons are uncorrelated at the 5% significance level. Show the test statistic and draw the conclusion.

Answer: The test statistic is  $T = -[24 - 0.5(2 + 2 + 1)] \ln[(1 - .622)(1 - .0029)] = 20.97$  with p-value 0.00032. Thus, we reject the null hypothesis. The measurements of the two sons are not independent.

- Let  $\rho_1^2 > \rho_2^2$  be the squared canonical correlations between the two sons. Test the hypothesis  $H_0 : \rho_2 = 0$  versus the alternative hypothesis  $H_a : \rho_2 \neq 0$  at the 5% significant level. Draw your conclusion.

Answer: The test statistic is  $T = -[24 - 0.5(2 + 2 + 1)] \ln(1 - 0.0029) = 0.0624$ . The p-value is 0.80. Thus, we cannot reject the null hypothesis of zero canonical correlation.

```

> r11=matrix(c(1,.7346, .7346, 1),2,2)
> r12=matrix(c(.7108,.6932,.7040,.7086),2,2)
> r22=matrix(c(1,.8392,.8392,1),2,2)
> m1=eigen(r11)
> m1
$values
[1] 1.7346 0.2654
$vectors
          [,1]      [,2]
[1,] 0.7071068 0.7071068
[2,] 0.7071068 -0.7071068

> x=m1$vectors
> d1=diag(sqrt(m1$values))
> a=x%%d1%%t(x)
> a      % square root matrix of r11
          [,1]      [,2]
[1,] 0.9161060 0.4009361
[2,] 0.4009361 0.9161060

```

```

> ainv=solve(a)
> r22inv=solve(r22)

> b=ainv%*%r12%*%r22inv%*%t(r12)%*%ainv
> m2=eigen(b)
> m2
$values
[1] 0.621816255 0.002896747
$vectors
      [,1]      [,2]
[1,] 0.7149442 0.6991815
[2,] 0.6991815 -0.7149442
> sqrt(m2$values)
[1] 0.78855327 0.05382144 % canonical correlations

> c1=m2$vectors % eigenvector of the first son
> m3=eigen(r22)
> g1=m3$vectors%*%diag(sqrt(m3$values))%*%t(m3$vectors)
> f=ginv%*%t(r12)%*%ainv%*%c1 % eigenvector of the second son
> f
      [,1]      [,2]
[1,] 0.5522325 0.03841978
[2,] 0.5628992 -0.03769174

*** Perform tests
> m2$values
[1] 0.621816255 0.002896747
>
> t1=(1-.621816)*(1-.002897)
> tt=-(24-0.5*(2+2+1))*log(t1)
> tt
[1] 20.96843
> p=pchisq(tt,4)
> p
[1] 0.9996787
> 1-p
[1] 0.0003212664
>
> t2=-(24-0.5*(2+2+1))*log(1-.002897)
> t2
[1] 0.0623759
> p1=pchisq(t2,1)

```

> p1  
 [1] 0.1972206  
 > 1-p1  
 [1] 0.8027794

4. (10 pts) Suppose that  $\mathbf{x}$  comes from one of two populations:

- $\pi_1$ : Normal with mean  $\boldsymbol{\mu}_1$ , covariance matrix  $\boldsymbol{\Sigma}_1$ , and probability density function  $f_1(\mathbf{x})$ .
- $\pi_2$ : Normal with mean  $\boldsymbol{\mu}_2$ , covariance matrix  $\boldsymbol{\Sigma}_2$ , and probability density function  $f_2(\mathbf{x})$ ,

where  $\boldsymbol{\Sigma}_i$  are positive definite. Assume further that the cost of misclassification is the same for both populations, and the prior probabilities are equal. Derive a discriminant equation for the two populations. Simplify the discriminant equation if  $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ .

Answer: Note that  $\ln \left[ \frac{c(1|2)}{c(2|1)} \times \frac{p_2}{p_1} \right] = 0$ . By plugging in the pdf of multivariate normal distribution, we obtain a quadratic discriminant equation

$$R_1 : -\frac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x} + (\boldsymbol{\mu}'_1\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}'_2\boldsymbol{\Sigma}_2^{-1})\mathbf{x} - k \geq 0,$$

where  $k = \frac{1}{2} \ln \left( \frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} \right) + \frac{1}{2}(\boldsymbol{\mu}'_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}'_2\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)$ .

If the covariance matrices are the same, then the discriminant equation becomes

$$R_1 : (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq 0.$$

5. (10 pts) Suppose that  $\mathbf{X}_1$ ,  $\mathbf{X}_2$  and  $\mathbf{X}_3$  are jointly multivariate normal with mean and covariance matrix given by

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_3 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{13} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \mathbf{0} \\ \boldsymbol{\Sigma}_{31} & \mathbf{0} & \boldsymbol{\Sigma}_{33} \end{bmatrix},$$

where  $\boldsymbol{\Sigma}_{ii}$  are positive definite and  $\mathbf{0}$  denotes a zero matrix.

- Derive the distribution of  $\mathbf{X}_1$  given  $\mathbf{X}_2 = \mathbf{x}_2$  and  $\mathbf{X}_3 = \mathbf{x}_3$ .

Answer: Since  $\mathbf{x}_2$  and  $\mathbf{x}_3$  are independent, we have  $(\mathbf{X}_1|\mathbf{x}_2, \mathbf{x}_3) = [(\mathbf{X}_1|\mathbf{x}_2)|\mathbf{x}_3]$ . Next,  $\mathbf{X}_1|\mathbf{x}_2$  is multivariate normal with mean and variance given by

$$\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \quad \boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

Applying the same result again, we obtain that  $\mathbf{X}_1|\mathbf{x}_2, \mathbf{x}_3$  is multivariate normal with mean and variance given by

$$\begin{aligned}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) + \boldsymbol{\Sigma}_{13}\boldsymbol{\Sigma}_{33}^{-1}(\mathbf{x}_3 - \boldsymbol{\mu}_3), \\ \boldsymbol{\Sigma}_{1|2,3} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} - \boldsymbol{\Sigma}_{13}\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\Sigma}_{31}.\end{aligned}$$

- Derive the distribution of  $\mathbf{X}_1$  given  $\mathbf{X}_2 + \mathbf{X}_3 = \mathbf{x}_0$ .

Answer: Let  $\mathbf{Y} = \mathbf{X}_2 + \mathbf{X}_3$  (assuming that  $\mathbf{X}_2$  and  $\mathbf{X}_3$  have the same dimension). From the independence,  $\mathbf{Y}$  is multivariate normal with mean  $\boldsymbol{\mu}_2 + \boldsymbol{\mu}_3$  and variance  $\boldsymbol{\Sigma}_{22} + \boldsymbol{\Sigma}_{33}$ . Furthermore, the joint distribution of  $\mathbf{X}_1$  and  $\mathbf{Y}$  is multivariate normal with mean  $\boldsymbol{\mu}_*$  and variance  $\boldsymbol{\Sigma}_*$ , where

$$\boldsymbol{\mu}_* = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 + \boldsymbol{\mu}_3 \end{bmatrix}, \quad \boldsymbol{\Sigma}_* = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} + \boldsymbol{\Sigma}_{13} \\ \boldsymbol{\Sigma}_{21} + \boldsymbol{\Sigma}_{31} & \boldsymbol{\Sigma}_{22} + \boldsymbol{\Sigma}_{33} \end{bmatrix}.$$

Consequently,  $\mathbf{X}_1|\mathbf{y} = \mathbf{x}_0$  is multivariate normal with mean and covariance matrix

$$\begin{aligned}\boldsymbol{\mu}_1 + (\boldsymbol{\Sigma}_{12} + \boldsymbol{\Sigma}_{13})(\boldsymbol{\Sigma}_{22} + \boldsymbol{\Sigma}_{33})^{-1}(\mathbf{x}_0 - \boldsymbol{\mu}_2 - \boldsymbol{\mu}_3), \\ \boldsymbol{\Sigma}_{11} - (\boldsymbol{\Sigma}_{12} + \boldsymbol{\Sigma}_{13})(\boldsymbol{\Sigma}_{22} + \boldsymbol{\Sigma}_{33})^{-1}(\boldsymbol{\Sigma}_{21} + \boldsymbol{\Sigma}_{31}).\end{aligned}$$

6. (15 pts) Consider the data on irises, page 657, of the textbook. Let  $\boldsymbol{\mu}_i$  be the mean vector of the population  $i$ . Assume that the data are from normal distributions.

- Construct a multivariate analysis of variance table to test the hypothesis  $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$  versus the alternative hypothesis  $\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j$  for some  $i$  and  $j$ . Use the 5% significance level to draw the conclusion.

Answer: The attached sample means of each species and the overall mean can be used to compute  $\mathbf{B}$  and the sample covariance matrix is used to compute  $\mathbf{B} + \mathbf{W}$  of the following MANOVA table:

Source	Matrix sum of squares	Degrees of freedom
Species	$\mathbf{B} = \sum_{i=1}^3 50(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'$	2
Residual	$\mathbf{W} = \sum_{i=1}^3 \sum_{j=1}^{50} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'$	147

The Wilks statistic is  $\frac{22096.88}{942754.6} = 0.023$ . Using the test statistics of Table 6.3 (p. 300), the F-ratio is 199.145, which compared with F distribution with degrees of freedom 8 and 288 has a p-value close to 0. Thus, we reject the null hypothesis of equal mean vector.

Using Bartlett's approximation,

$$-\left(150 - 1 - \frac{4 + 2}{2}\right) \ln(\Lambda^*) = 547.99.$$

Compared with  $\chi_8^2$ , this test statistic is highly significant.

- Focus on the population 1: Iris setosa. Construct a 95% confidence interval for each measurement using (1) one-at-a-time method, (2) simultaneous procedure, (3) the Bonferroni method.

Answer: The confidence intervals are given below

(1) "CR based on individual t"

	[,1]	[,2]
[1,]	4.9058235	5.1061765
[2,]	3.3202711	3.5357289
[3,]	1.4126452	1.5113548
[4,]	0.2160497	0.2759503

(2) "C.R. based on T^2"

	[,1]	[,2]
[1,]	4.8409111	5.1710889
[2,]	3.2504649	3.6055351
[3,]	1.3806643	1.5433357
[4,]	0.1966426	0.2953574

(3) "CR based on Bonferroni"

	[,1]	[,2]
[1,]	4.8767271	5.1352729
[2,]	3.2889811	3.5670189
[3,]	1.3983101	1.5256899
[4,]	0.2073506	0.2846494

7. (15 pts) Consider the quarterly U.S. real gross domestic product (gdp) and unemployment rate from 1948 to 2003. Let  $\mathbf{y}_t = (y_{1t}, y_{2t})'$ , where  $y_{1t}$  is the growth rate of gdp, and  $y_{2t}$  is the change in unemployment rate. Let  $\mathbf{x}_t = (1, y_{1,t-1}, y_{2,t-1}, y_{1,t-2}, y_{2,t-2})'$  be the vector of explanatory variables. That is, we use lag-1 and lag-2 of the dependent variables as the explanatory variable. Treating the problem as a multivariate linear regression one, we can estimate the model. Some R output is attached. Use the output to answer the questions:

- Write down the fitted model in the form

$$\mathbf{y}_t = \mathbf{x}_t \boldsymbol{\beta} + \boldsymbol{\epsilon}_t.$$

Answer: The fitted model is

$$\mathbf{y}_t = \mathbf{x}_t \boldsymbol{\beta} + \boldsymbol{\epsilon}_t,$$

where

$$\boldsymbol{\beta} = \begin{bmatrix} 0.0057 & 0.1901 \\ 0.1466 & -11.733 \\ -0.0087 & 0.4502 \\ 0.1730 & -10.098 \\ 0.0086 & -0.2991 \end{bmatrix}.$$

- Let  $\mathbf{x}_{1t} = (1, y_{1,t-1}, y_{2,t-1})'$  be a subset of  $\mathbf{x}_t$ . We also fit a multiple linear regression model using  $\mathbf{x}_{1t}$ . Perform the likelihood ratio test to hypothesis  $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$  versus the alternative hypothesis  $H_a : \boldsymbol{\beta}_2 \neq \mathbf{0}$ , where  $\boldsymbol{\beta}_2$  is defined as

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}.$$

What is the test statistic? Draw your conclusion at the 5% significance level.

Answer: The likelihood ratio test statistic is

$$\begin{aligned} & -[n - r - 1 - \frac{1}{2}(m - r + q + 1)] \ln \left( \frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_1|} \right) \\ & = -(225 - 4 - 1 - 0.5(2 - 4 + 2 + 1)) \ln(3.7951/4.2011) = 22.31. \end{aligned}$$

Compared with  $\chi_4^2$ , the p-value is 0.0002. Thus, reject the null hypothesis.

- Compute the  $t$ -ratios of elements of the  $\boldsymbol{\beta}_2$  estimate. Do these  $t$ -ratios provide the same conclusion as the likelihood ratio test?

Answer: The  $t$ -ratios are 1.92, 3.54,  $-3.59$ , and  $-4.04$ , respectively. Except for the first one, all  $t$ -ratios are significant at the 5% level. Thus,  $t$ -ratios provide the same conclusion as the likelihood ratio statistic.