**THE UNIVERSITY OF CHICAGO**
**Graduate School of Business**
Business 41912, Spring Quarter 2006, Mr. Ruey S. Tsay

# Final Exam

**Notes**:

1. Open book and notes. The exam time is 3 hours.
2. Write your solutions in the answer book. Mark the solution clearly.
3. You are required to pledge to uphold the honor code of GSB in the 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 | 0.1 8 |
| 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.
- (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.
- Compare the two linkage methods. Is there any difference? Which method produces results that are close to the industrial categories?

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.
- Write down the three vectors of factor loadings.
- 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?

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

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

Answer the following questions:

- Obtain the canonical correlations and the associated canonical variates of head measurements of the two sons.
- Assume normaility. Test the hypothesis that the measurements between the two sons are ucorrelated at the 5% significance level. Show the test statistic and draw the conclusion.
- Let $\rho_1^2 > \rho_2^2$ be the squared canonical correlations between the two sons. Test the hypithesis $H_0 : \rho_2 = 0$ versus the alternative hypothesis $H_a : \rho_2 \neq 0$ at the 5% significant level. Draw your conclusion.

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

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

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

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

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

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

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

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% significane level to draw the conclusion.
   - Forcus 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 Berferroni method.

7. (15 pts) Consider the quarterly U.S. real gross demonstic product (gdp) and unemployment rate from 1948 to 2003. Let $\boldsymbol{y}_t = (y_{1t}, y_{2t})'$, where $y_{1t}$ is the growth rate of gdp, and $y_{2t}$ is the change in unemployment rate. Let $\boldsymbol{x}_t = (1, y_{1,t-1}, y_{2,t-1}, y_{1,t-2}, y_{2,t-2})'$ 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

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

- Let $\boldsymbol{x}_{1t} = (1, y_{1,t-1}, y_{2,t-1})'$ be a subset of $\boldsymbol{x}_t$. We also fit a multiple linear regression model using $\boldsymbol{x}_{1t}$. 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.
- 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?

```
** Output for Problem 2 **
> dim(da)
[1] 168  10
> s=cov(da)
> d1=sqrt(diag(s))
> s1=diag(d1)
> s2=solve(s1)
> y=as.matrix(da)%*%s2   % Standardize the data
> y1=as.matrix(y)

> m1=factanal(y1,2)   % 2 factor model
> m1
Call:
factanal(x = y1, factors = 2)

Uniquenesses:
 [1] 0.484 0.361 0.205 0.214 0.836 0.611 0.714 0.191 0.597 0.900
Loadings:
      Factor1 Factor2
 [1,] 0.671   0.256
 [2,] 0.719   0.349
 ....
[10,] 0.315

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 72.93 on 26 degrees of freedom.
The p-value is 2.46e-06

> m2=factanal(y1,3,rotation="none",scores="regression")
> m2
Call:
factanal(x=y1,factors=3,scores="regression",rotation="none")

Uniquenesses:
 [1] 0.4xx xxx xxxx (output deleted.)

Loadings:
      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

               Factor1 Factor2 Factor3
SS loadings      2.947   2.003   0.832
Proportion Var   0.295   0.200   0.083
Cumulative Var   0.295   0.495   0.578

> s=cov(y)
> g2=det(s)
> g2
[1] 0.01143887

** Problem 6 **
> da=read.table("T11-5.DAT")
> dim(da)
[1] 150   5
> y=as.matrix(da[,1:4])
> y1=y[1:50,]
> y2=y[51:100,]
> y3=y[101:150,]

> apply(y1,2,mean)
   V1    V2    V3    V4
5.006 3.428 1.462 0.246
> apply(y2,2,mean)
   V1    V2    V3    V4
5.936 2.770 4.260 1.326
> apply(y3,2,mean)
   V1    V2    V3    V4
6.588 2.974 5.552 2.026

> apply(y,2,mean)
      V1       V2       V3       V4
5.843333 3.057333 3.758000 1.199333

> s1=cov(y1)
> s1
```

```
           V1         V2          V3          V4
V1 0.12424898 0.09921633 0.016355102 0.010330612
V2 0.09921633 0.14368980 0.011697959 0.009297959
V3 0.01635510 0.01169796 0.030159184 0.006069388
V4 0.01033061 0.00929796 0.006069388 0.011106122

> s2=cov(y2)
> s2
           V1         V2         V3         V4
V1 0.26643265 0.08518367 0.18289796 0.05577959
V2 0.08518367 0.09846939 0.08265306 0.04120408
V3 0.18289796 0.08265306 0.22081633 0.07310204
V4 0.05577959 0.04120408 0.07310204 0.03910612

> s3=cov(y3)
> s3
           V1         V2         V3         V4
V1 0.40434286 0.09376327 0.30328980 0.04909388
V2 0.09376327 0.10400408 0.07137959 0.04762857
V3 0.30328980 0.07137959 0.30458776 0.04882449
V4 0.04909388 0.04762857 0.04882449 0.07543265
>
> sp=(49*s1+49*s2+49*s3)/147
> sp
           V1         V2         V3         V4
V1 0.26500816 0.09272109 0.16751429 0.03840136
V2 0.09272109 0.11538776 0.05524354 0.03271020
V3 0.16751429 0.05524354 0.18518776 0.04266531
V4 0.03840136 0.03271020 0.04266531 0.04188163

> v1=cov(y)
> v1
            V1          V2         V3         V4
V1  0.68569351 -0.04243400  1.2743154  0.5162707
V2 -0.04243400  0.18997942 -0.3296564 -0.1216394
V3  1.27431544 -0.32965638  3.1162779  1.2956094
V4  0.51627069 -0.12163937  1.2956094  0.5810063

> v2=v1*149
> v2
           V1         V2       V3        V4
V1 102.168333  -6.322667 189.8730  76.92433
```

```
V2  -6.322667  28.306933 -49.1188 -18.12427
V3 189.873000 -49.118800 464.3254 193.04580
V4  76.924333 -18.124267 193.0458  86.56993

> sp*147
        V1      V2      V3     V4
V1 38.9562 13.6300 24.6246 5.6450
V2 13.6300 16.9620  8.1208 4.8084
V3 24.6246  8.1208 27.2226 6.2718
V4  5.6450  4.8084  6.2718 6.1566

> det(sp*147)
[1] 22096.88

> det(v1*149)
[1] 942754.6

** Problem 7 **
> da=read.table("q-gdpun.txt")
> dim(da)
[1] 228   5
> gdp=diff(da[,4])   % Obtain growth rates of GDP
> un=diff(da[,5])    % Obtain changes in unemployment rate.

> xx=cbind(gdp,un)
> dim(xx)
[1] 227   2
>
> y=xx[3:227,]    % Construct y-vector
> x=cbind(rep(1,225),xx[2:226,],xx[1:225,]) %
> dim(x)
[1] 225   5
>
> xtx=t(x)%*%x
> xty=t(x)%*%y
> xtxin=solve(xtx)
> b=xtxin%*%xty
> b
              gdp          un
      0.005725224   0.1900745
gdp1  0.146636138 -11.7330084
un1  -0.008719171   0.4502194
```

```
gdp2  0.173036572 -10.0980491
un2   0.008555776  -0.2990811

> resi=y-x%*%b
> sig=(t(resi)%*%resi)/(225-5)
> sig
              gdp           un
gdp  8.392534e-05 -0.001705331
un  -1.705331e-03  0.081950267

> xtxin
                      1gdp         un1       gdp2         un2
      0.01928975 -0.8468530 -0.02330566 -0.8861283 -0.00882072
gdp1 -0.84685299 92.3106663  1.94734535  6.9275714 -0.30084285
un1  -0.02330566  1.9473453  0.09189436  0.7638365 -0.02709015
gdp2 -0.88612832  6.9275714  0.76383648 96.1690244  1.30640987
un2  -0.00882072 -0.3008428 -0.02709015  1.3064099  0.06629015
>
> x1=x[,1:3]          % Construct subset of x-vector
> ztz=t(x1)%*%x1
> ztzin=solve(ztz)
> zty=t(x1)%*%y
> b1=ztzin%*%zty
> b1
             gdp          un
      0.006908423   0.1077575
gdp1  0.184656938 -12.3228530
un1  -0.005304290   0.4054643

> resi1=y-x1%*%b1
> sig1=t(resi1)%*%resi1/(225-3)
> sig1
             gdp1          un1
gdp1  8.814463e-05 -0.001865036
un1  -1.865036e-03  0.088420616
>
> det(sig*220/225)
[1] 3.795085e-06

> det(sig1*222/225)
[1] 4.201132e-06
```