THE UNIVERSITY OF CHICAGO
Graduate School of Business
Business 41912, Spring Quarter 2008, Mr. Ruey S. Tsay

Final Exam

GSB Honor Code:
I pledge my honor that I have not violated the Honor Code during this examination.

Signature: **Name**: **ID**:

Notes:

1. Open book and notes. The exam time is 3 hours.
2. Write your answers in a bluebook. Mark the solution clearly.
3. All tests are based on the 5% significance level.
4. R output is attached for some problems.
5. Round your answer to 2 digits.

1. (13 pts) Consider the monthly log returns, in percentages, of five U.S. stocks from 1990 to 1999 for 120 observations. The stocks are (1) IBM, (2) Hewlett-Packard (HPQ), (3) Intel (INTC), (4) Merill Lynch (MER), and (5) Morgan Stanley Dean Witter (MWD). Let $\boldsymbol{X}_t = (IBM_t, HPQ_t, INTC_t)'$ and $\boldsymbol{Y}_t = (MER_t, MWD_t)'$ so that $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ represent returns of high-tech and financial stocks, respectively. Based on the attached output, answer the following questions:
   (a) (2 pts) What are the canonical correlation coefficients between $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$?
   (b) (4 pts) Obtain the first two canonical variates of $\boldsymbol{X}_t$. Show the steps taken to obtain your answer.
   (c) (3 points) Let $\rho_1^* > \rho_2^*$ be the two canonical correlations between $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$. Test the null hypothesis $H_o : \rho_2^* = 0$ versus $H_a : \rho_2^* \neq 0$. What is the statistic? Draw your conclusion.

(d) (4 points) Setup the hypotheses to test the null that $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ are uncorrelated. Calculate the test statistic? Draw your conclusion.

2. (10 pts) Again, consider the five monthly stock returns of Problem 1. The correlation matrix of the stocks is

	IBM	HPQ	INTC	MER	MWD
IBM	1.00	0.42	0.30	0.18	0.18
HPQ	0.42	1.00	0.45	0.36	0.38
INTC	0.30	0.45	1.00	0.25	0.24
MER	0.18	0.36	0.25	1.00	0.79
MWD	0.18	0.38	0.24	0.79	1.00

Answer the following questions:

(a) (2 points) Use the correlations to construct a distance measure between the five stocks. Write down the distance.

(b) (2 points) Perform the hierarchical cluster analysis using the single linkage method. Show details of the first updating of the distance and draw the dendrogram.

(c) (2 points) Perform the hierarchical cluster analysis using the complete linkage method. Show details of the first updating of the distance and draw the dendrogram.

(d) (2 points) Perform the hierarchical cluster analysis using the average linkage method. Show details of the first updating of the distance and draw the dendrogram.

(e) (2 points) Compare and comment on the three linkage methods. In particular, is there any difference among them in this particular instance?

3. (15 points) Consider again the monthly log returns of five stocks in Problem 1. The eigenvalues and eigenvectors of the sample correlation matrix are given in the output. Answer the following questions.

(a) (3 points) Assume one common factor only. Obtain an orthogonal factor model using the principal component analysis method. Write down the factor loadings and the specific variances.

(b) (4 points) Assume that there are two common factors. Obtain an orthogonal factor model using the principal component analysis method. Write down the factor loadings and the specific variance.

(c) (2 points) What is the proportion of total variance explained by the prior two-factor model?

(d) (2 points) The factor analysis using the maximum likelihood method is given in the output. Write down the fitted orthogonal factor model, including factor loadings and the specific variances.

(e) (4 points) For the maximum likelihood method, what is the large sample test statistic for testing $m = 2$ factors when the Bartlett's correction is used? What is the p-value? Draw your conclusion.

4. (12 pts) Assume that $n_1 = 11$ and $n_2 = 12$ observations were randomly collected from Populations 1 and 2. Assume also that the observations are bivariate and follow multivariate normal distributions $N_2(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ for $i = 1$ and 2. Suppose that the summary statistics of the samples are

$$\bar{\boldsymbol{x}}_1 = \begin{bmatrix} -1 \\ -2 \end{bmatrix}, \quad \bar{\boldsymbol{x}}_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad \boldsymbol{S}_{pooled} = \begin{bmatrix} 4.0 & -1.0 \\ -1.0 & 5.0 \end{bmatrix}.$$

Answer the following questions:

(a) (4 points) Test $H_o : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $H_a : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ using the Hotelling's two-sample T^2-statistic. Draw the conclusion.

(b) (4 points) Construct Fisher's (sample) linear discrimiant function for the two populations.

(c) (4 points) Assume equal costs and equal prior probabilities. Assign the new observation $\boldsymbol{x}_o = (0, 1)'$ to either population.

5. (10 pts) Consider the multiple linear regression model

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{Z}_{n\times(r+1)}\boldsymbol{\beta}_{(r+1)\times 1} + \boldsymbol{\epsilon}_{n\times 1},$$

where $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$ with $\boldsymbol{I}_n$ being the $n \times n$ identity matrix, and $\boldsymbol{Z}$ is of full rank $(r + 1)$ with the first column consisting of 1. Let $\boldsymbol{H} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$ be the hat-matrix, $\widehat{\boldsymbol{\beta}}$ be the least squares estimate of $\boldsymbol{\beta}$, and $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}$ be the residual vectors. Prove the following statements:

(a) (2 points) Let $\hat{\epsilon}_i$ be the ith element of $\widehat{\boldsymbol{\epsilon}}$. Then, $\sum_{i=1}^{n} \hat{\epsilon}_i = 0$.

(b) (3 points) $E(\widehat{\boldsymbol{\epsilon}}\widehat{\boldsymbol{\epsilon}}') = (\boldsymbol{I} - \boldsymbol{H})\sigma^2$.

(c) (5 points) If $r = 1$, i.e. simple linear regression, then

$$h_{jj} = \frac{1}{n} + \frac{(z_j - \bar{z})^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2},$$

where h_{jj} is the (j, j)th element of $\boldsymbol{H}$ and z_j is the jth element of the 2nd column of $\boldsymbol{Z}$.

6. (10 pts) Assume that $\boldsymbol{X}$ is a p-dimensional normal random vector with mean zero and covariance matrix $\boldsymbol{\Sigma}$. Suppose that a radnom sample $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ of n observations is available. Let $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}$, where $\boldsymbol{X}' = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$. Prove the following statements:

(a) (4 points) $\widehat{\boldsymbol{\Sigma}}$ is the maximum likelihood estimate of $\boldsymbol{\Sigma}$.

(b) (2 points) $E(\widehat{\boldsymbol{\Sigma}}) = \boldsymbol{\Sigma}$.

(c) (4 points) When $n < p$, $\widehat{\boldsymbol{\Sigma}}$ is singular. Show that the non-zero eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$ are the same as those of $\boldsymbol{X}\boldsymbol{X}'$, which is a $n \times n$ matrix. *Hint: use the identity*

$$|\boldsymbol{I}_p - \boldsymbol{X}'\boldsymbol{X}| = |\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{X}'|.$$

7. (15 pts) Anacondas are some of the largest snakes in the world. Jesus Ravis and his associates capture a snake and measure its (i) snout vent length (cm) or the length from the snout of the snake to its vent where it evaculates waste and (ii) weights (kilograms). A sample of these measurements is shown in Table 6.19 of the textbook (p. 357). Some summary statistics of the data are given in the output. Use the information to asnwer the following questions:

   (a) Test the equality of the two covariance matrices between male and female snakes.

   (b) Test the equality of the means between male and female snakes based on the result of part 1 (i.e. to pool or not to pool the covariances).

   (c) Construct the 95% Bonferroni confidence intervals for the mean differences between males and females on both length and weight.

8. (15 pts) Consider the monthly 30-year fixed mortgage rates (M_t) and the 2-year Treasury Constant Maturity interest rates (I_t) of the U.S. from June 1976 to May 2008 for 384 observations. Let $\boldsymbol{Y}_t = (M_t, I_t)'$ be the 2-dimensional dependent variable and $\boldsymbol{X}_t = (1, \boldsymbol{Y}'_{t-1}, \boldsymbol{Y}'_{t-2}, \boldsymbol{Y}'_{t-3}, \boldsymbol{Y}'_{t-4})'$ be an 9-dimensional independent variable, including the first element being 1. This is equivalent to employing a VAR(4) model for $\boldsymbol{Y}_t$. Let $\boldsymbol{Z}_t = (1, \boldsymbol{Y}'_{t-1}, \boldsymbol{Y}'_{t-2}, \boldsymbol{Y}'_{t-3})'$ be a 7-dimensional independent variable. Thus, using $\boldsymbol{Z}_t$ is equivalent to fitting a VAR(3) model to $\boldsymbol{Y}_t$. More presicely, we write the VAR(3) model as

$$\boldsymbol{Y}'_t = \boldsymbol{Z}'_t\boldsymbol{\beta}_3 + \boldsymbol{\epsilon}'_t$$

and the VAR(4) model as

$$\boldsymbol{Y}'_t = \boldsymbol{X}'_t\boldsymbol{\beta}_4 + \boldsymbol{e}'_t,$$

where

$$\boldsymbol{\beta}_4 = \begin{bmatrix} \boldsymbol{\beta}_3 \\ \boldsymbol{\beta}_* \end{bmatrix} \quad \text{with} \quad \boldsymbol{\beta}_* \quad \text{being a } 2 \times 2 \text{ matrix.}$$

Some R output is attached. Use the information to answer the following questions.

   (a) (4 points) Test the hypothesis $H_o : \boldsymbol{\beta}_* = \mathbf{0}$ versus $H_a : \boldsymbol{\beta}_* \neq \mathbf{0}$. Under the normality assumption, what is the maximum likelihood ratio test statistic? Draw your conclusion.

   (b) (2 points) Write down the fitted VAR(3) model, including covariance matrix of the residuals.

(c) (3 points) Focus on the VAR(3) fit. Let β_{ij} be the (i, j)th element of $\boldsymbol{\beta}_3$. Test the hypothesis $H_o : \beta_{11} = 0$ versus $H_a : \beta_{11} \neq 0$. What is the t-ratio? What is the associated p-value? Draw your conclusion.

(d) (3 points) Again, focus on the VAR(3) fit. Test the hypothesis $H_o : \beta_{21} = 0$ versus $H_a : \beta_{21} \neq 0$. What is the t-ratio? What is the associated p-value? Draw your conclusion.

(e) (3 points) Again, focus on the VAR(3) fit. What is the estimated covariance between β_{21} and β_{22}?

Problem 1

```
> x=read.table("m-5cln.txt")
> dim(x)
[1] 120   5
> apply(x,2,mean)
      V1       V2       V3       V4       V5
1.465728 1.974066 3.050518 2.294586 2.364150
> apply(x,2,var)
       V1        V2        V3        V4        V5
 73.10415 103.60482 113.96073 105.55809 109.91059

> colnames(x) <- c("IBM","HPQ","INTC","MER","MWD")
> X1=x[,1:3]
> X2=x[,4:5]
> S11=cor(X1)
> S12=cor(X1,X2)
> S22=cor(X2)
> S11inv=solve(S11)
> S22inv=solve(S22)
> print(S11,digits=3)
       IBM   HPQ  INTC
IBM  1.000 0.419 0.297
HPQ  0.419 1.000 0.450
INTC 0.297 0.450 1.000
> print(S22,digits=3)
      MER   MWD
MER 1.000 0.794
MWD 0.794 1.000
> print(S12,digits=3)
       MER   MWD
IBM  0.183 0.182
HPQ  0.359 0.382
INTC 0.255 0.240
> print(S11inv,digits=3)
        IBM    HPQ   INTC
IBM   1.235 -0.442 -0.168
HPQ  -0.442  1.412 -0.504
INTC -0.168 -0.504  1.276
> print(S22inv,digits=3)
      MER   MWD
MER  2.70 -2.14
MWD -2.14  2.70
```

```
> TT=S22inv%*%t(S12)%*%S11inv%*%S12
> m1=eigen(TT)
> m1
$values
[1] 0.163101518 0.002671486

$vectors
           [,1]       [,2]
[1,] -0.5808415 -0.7205799
[2,] -0.8140167  0.6933719

*** Problem 2 ****
> setwd("C:/teaching/ama")
> x=read.table("m-5cln.txt")
> R=cor(x)
> print(R,digits=2)
     V1   V2   V3   V4   V5
V1 1.00 0.42 0.30 0.18 0.18
V2 0.42 1.00 0.45 0.36 0.38
V3 0.30 0.45 1.00 0.25 0.24
V4 0.18 0.36 0.25 1.00 0.79
V5 0.18 0.38 0.24 0.79 1.00

** Problem 3 ***
> mm=eigen(R)
> print(mm$values,digits=3)
[1] 2.456 1.145 0.699 0.495 0.205
> print(mm$vectors,digits=3)
       [,1]   [,2]    [,3]    [,4]    [,5]
[1,] -0.342  0.525  0.6906  0.3618  0.0118
[2,] -0.474  0.314 -0.0425 -0.8199 -0.0496
[3,] -0.387  0.405 -0.7166  0.4140  0.0339
[4,] -0.503 -0.481  0.0517  0.1467 -0.7008
[5,] -0.505 -0.481  0.0708  0.0617  0.7107
> f1=factanal(x,2,cor=T,scores=c("regression"))
> f1

Call:
factanal(x = x, factors = 2, scores = c("regression"), cor = T)

Uniquenesses:
```

```
   V1    V2    V3    V4    V5
0.718 0.371 0.676 0.360 0.005

Loadings:
   Factor1 Factor2
V1         0.523
V2 0.254   0.751
V3 0.146   0.550
V4 0.768   0.226
V5 0.982   0.176

               Factor1 Factor2
SS loadings      1.647   1.224
Proportion Var   0.329   0.245
Cumulative Var   0.329   0.574
>
> names(f1)
 [1] "converged"    "loadings"     "uniquenesses" "correlation"  "criteria"
 [6] "factors"      "dof"          "method"       "scores"       "STATISTIC"
[11] "PVAL"         "n.obs"        "call"

> L=f1$loadings
> Psi=f1$uniquenesses
> U=as.matrix(L)%*%t(as.matrix(L))+diag(Psi)
> det(U)
[1] 0.2000050
> det(R)
[1] 0.1996378

*** Snake example ****
> x=read.table("T6-19.DAT")
> dim(x)
[1] 56  3

> F=x[1:28,1:2]
> M=x[29:56,1:2]
> mu1=apply(F,2,mean)
> mu1
       V1        V2
348.27500  37.26357
> mu2=apply(M,2,mean)
> mu2
```

```
        V1          V2
228.753571   7.290357
> S1=var(F)
> S2=var(M)
> Sp=0.5*(S1+S2)
> deter=c(det(S1),det(S2),det(Sp))
> deter
[1] 221168.3129    699.5423  86170.3848
> S1inv=solve(S1)
> S2inv=solve(S2)
> Spinv=solve(Sp)
> print(S1inv,digits=3)
         V1       V2
V1  0.00183 -0.00586
V2 -0.00586  0.02129
> print(S2inv,digits=3)
         V1      V2
V1  0.00633 -0.0559
V2 -0.05593  0.7197
> print(Spinv,digits=3)
         V1       V2
V1  0.00237 -0.00775
V2 -0.00775  0.03025
> print(S1)
         V1        V2
V1 4709.298 1296.7605
V2 1296.760  404.0424
> print(S2)
          V1        V2
V1 503.47888 39.126536
V2  39.12654  4.430033
> print(Sp)
          V1       V2
V1 2606.3882 667.9435
V2  667.9435 204.2362
>
*** Mortgage Rate and GS2 interest Rate
> mort=read.table("MORTG.txt")
> dim(mort)
[1] 384   4
> gs2=read.table("GS2.txt")
> dim(gs2)
```

```
[1] 384   4
> mort=mort[,4]
> gs2=gs2[,4]
> Y=cbind(mort[5:384],gs2[5:384])
*** VAR(4) model
> X=cbind(rep(1,380),mort[4:383],gs2[4:383],mort[3:382],gs2[3:382],mort[2:381],
gs2[2:381],mort[1:380],gs2[1:380])
> XPX=t(X)%*%X
> XPY=t(X)%*%Y
> XPXinv=solve(XPX)
> beta=XPXinv%*%XPY
> print(beta,digits=3)
          [,1]    [,2]
 [1,]  0.15059  0.0915
 [2,]  1.02206 -0.5609
 [3,]  0.47910  1.6273
 [4,] -0.20197  0.5669
 [5,] -0.47236 -0.6990
 [6,]  0.16201  0.0234
 [7,]  0.04987  0.0883
 [8,] -0.03540 -0.0455
 [9,] -0.00654 -0.0093
> resi=Y-X%*%beta
> Sig4=t(resi)%*%resi/380

> det(Sig4)
[1] 0.005236972

*** VAR(3) model
> Z=X[,1:7]
> ZPZ=t(Z)%*%Z
> ZPY=t(Z)%*%Y
> ZPZinv=solve(ZPZ)
> beta3=ZPZinv%*%ZPY
> print(beta3,digits=3)
        [,1]    [,2]
[1,]  0.1561  0.0988
[2,]  1.0113 -0.5749
[3,]  0.4825  1.6316
[4,] -0.1610  0.6206
[5,] -0.4840 -0.7140
[6,]  0.0937 -0.0653
```

```
[7,]  0.0544  0.0934
> resi3=Y-Z%*%beta3
> Sig3=t(resi3)%*%resi3/380    % MLE of covariance matrix.
> det(Sig3)
[1] 0.005248375

> Sig=t(resi3)%*%resi3/(380-7)  % LSE of covariance matrix
> print(Sig3,digits=3)
       [,1]   [,2]
[1,] 0.0507 0.0544
[2,] 0.0544 0.1657

> print(ZPZinv,digits=2)
         [,1]    [,2]    [,3]    [,4]    [,5]     [,6]    [,7]
[1,]  5.8e-02 -0.0086  0.0015 -0.0053  0.0041  7.3e-05  0.0051
[2,] -8.6e-03  0.0786 -0.0259 -0.0868  0.0012  1.1e-02  0.0214
[3,]  1.5e-03 -0.0259  0.0246  0.0387 -0.0265 -1.4e-02  0.0028
[4,] -5.3e-03 -0.0868  0.0387  0.1516 -0.0379 -6.3e-02 -0.0028
[5,]  4.1e-03  0.0012 -0.0265 -0.0379  0.0630  3.5e-02 -0.0349
[6,]  7.3e-05  0.0114 -0.0136 -0.0628  0.0350  5.1e-02 -0.0203
[7,]  5.1e-03  0.0214  0.0028 -0.0028 -0.0349 -2.0e-02  0.0336
```