THE UNIVERSITY OF CHICAGO
Booth School of Business
Business 41912, Spring Quarter 2016, Mr. Ruey S. Tsay

Midterm

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

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

Notes:

1. There are two parts in the exam. The first part consists of some simple questions whereas the second part focuses on data analysis. The total time is 180 minutes with the first 60 minutes (maximum) allocated to the first part. You can immediately start the second part of the exam once you hand in the solutions of the first part.
2. Turn off cell phones. No communication is allowed during the exam. Except for downloading data of part 2, no Internet connection is allowed.
3. You may consult textbooks during the second part of the exam.
4. Write your answers in a bluebook. Mark the solution clearly.
5. All tests are based on the 5% significance level.
6. The transpose of the matrix $\boldsymbol{A}$ is $\boldsymbol{A}'$.

Part One. Basic concepts.

Problem A. (12 points) Suppose that $\boldsymbol{X} = (X_1, X_2, X_3)'$ follows a 3-dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ given by

$$\boldsymbol{\mu} = \begin{bmatrix} 3 \\ 1 \\ 4 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 6 & 1 & -2 \\ 1 & 13 & 4 \\ -2 & 4 & 4 \end{bmatrix}.$$

Answer the following questions:

1. Find the joint distribution of $\boldsymbol{Z} = (X_1, X_3)'$.

2. Find the distribution of $Z = 2X_1 - X_2 + 3X_3$.

3. (4 points) Find the conditional joint distribution of $Z = (X_1, X_2)'$ given $X_3 = 5$.

4. Are the random variables $Z_1 = X_1 - X_2 + X_3$ and $Z_2 = 2X_1 + X_2 - X_3$ independent? Why?

5. What is the distribution of $(\boldsymbol{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$?

Problem B. (10 points) Suppose that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are random samples from a 5-dimensional multivariate Non-Gaussian distribution with mean $\boldsymbol{\mu}$ and positive-definite covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{S}$ be the sample covariance matrix and $\bar{\boldsymbol{X}}$ be the sample mean. What are the limiting distributions of the following random vectors as $n \to \infty$?

1. $\sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$.

2. $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$.

3. $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$.

4. Describe briefly the impact of serial dependence of $\boldsymbol{X}_i$ on the limiting distribution of question (3).

5. Let $\boldsymbol{S}_n = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'$. What is $E(\boldsymbol{S}_n)$?

Problem C: (12 points) Consider the multivariate multiple linear regression (mmlr) model,

$$\boldsymbol{Y}_i' = \boldsymbol{Z}_i'\boldsymbol{\beta} + \boldsymbol{\epsilon}_i', \quad i = 1, \ldots, n,$$

where $\boldsymbol{Y}_i$ is a m-dimensional dependent vector, $\boldsymbol{Z}_i = (1, Z_{1i}, \ldots, Z_{p,i})'$ is a $(p+1)$ dimensional vector and $\boldsymbol{\epsilon}_i = (\epsilon_{1i}, \ldots, \epsilon_{mi})'$ is a m-dimensional random noise. In matrix form, the model can be written as

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

where $\boldsymbol{Z}$ is a full rank design matrix. Assume that $n > (p+1) + m$. Let $\widehat{\boldsymbol{Y}} = \boldsymbol{Z}\widehat{\boldsymbol{\beta}}$ be the fitted value of the mmlr model and $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ be the residual matrix, where $\widehat{\boldsymbol{\beta}}$ is the ordinary least squares estimate of $\boldsymbol{\beta}$. Answer the following questions:

1. (4 pints) State the three common assumptions used for the mmlr model.

2. Write down $\widehat{\boldsymbol{\beta}}$.

3. Why is $\boldsymbol{Z}'\widehat{\boldsymbol{\epsilon}} = \boldsymbol{0}$?

4. What is asymptotic distribution of $\widehat{\boldsymbol{\beta}}$?

5. Focus on the ith dependent variable, i.e. the i-column of the matrix $\boldsymbol{Y}$. Denote it by $\boldsymbol{Y}_{(i)}$. Let $\widehat{\boldsymbol{\epsilon}}_{(i)} = \boldsymbol{Y}_i - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{(i)}$ be the residual vector of $\boldsymbol{Y}_{(i)}$. What is $\text{cov}(\widehat{\boldsymbol{\epsilon}}_{(i)})$?

Problem D. (10 points) Briefly answer the following questions:

1. What are the assumptions used to justify the use of Hotelling's T^2 statistic?
2. True or false that the Hotelling's T^2 statistic is location and scale invariant? Why?
3. Suppose that X_i is normally distributed with mean μ_i and variance σ_i^2 for $i = 1$ and 2. Let $\boldsymbol{X} = (X_1, X_2)'$. Under what condition that $\boldsymbol{X}$ follows a bivariate normal distribution?
4. What are the assumptions used to compare the mean vectors of two populations?
5. Describe two situations under which the LASSO regression are useful.

Problem E. (5 points) Answer briefly the following questions.

1. Describe two methods for handle missing values in a multivariate data set?
2. (1 point) State the key assumption used in your answers of question (1).
3. Describe a method for detecting outliers in multivariate data analysis.

Part Two: Data analysis

Problem F. (21 points) To study the relationship between two measurements (y_1 = taste and y_2 = odor) and eight variables (x_1 = pH, x_2 = acidity 1, x_3 = acidity 2, x_4 = sake meter, x_5 = direct reducing sugar, x_6 = total sugar, x_7 = alcohol, and x_8 = formyl-nitrogen), Siotani et al. (1963, Processing of ISM) collected data from 30 brands of Japanese Seishu wine. The data are available from the course web: `seishu.txt`.

1. Compute the sample mean and covariance matrix of $\boldsymbol{X} = (x_1, \ldots, x_8)'$.
2. Find the maximum Pearson correlation (in magnitude) between $\boldsymbol{Y}$ and $\boldsymbol{X}$ and identify the corresponding pair of variables.
3. Find the maximum Kendall's tau (in magnitude) between $\boldsymbol{Y}$ and $\boldsymbol{X}$ and identify the corresponding pair of variables.
4. Focus on y_1 only. Find the linear regression model selected by the stepwise procedure. Write down the fitted model.
5. Let $\boldsymbol{Y} = (y_1, y_2)'$. Obtain the multivariate multiple linear regression between $\boldsymbol{Y}$ and the regressors $\boldsymbol{X}$.

6. Partition $\boldsymbol{X}$ into three categories, namely $\boldsymbol{X}_1 = (x_1, x_3)'$, $\boldsymbol{X}_2 = (x_2, x_4, x_5, x_6)'$, and $\boldsymbol{X}_3 = (x_7, x_8)'$. Test the significant of $\boldsymbol{X}_3$ given that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are in the model. Draw the conclusion.

7. Test the significance of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ are in the model. Draw the conclusion.

Problem G. (10 points) Consider the data set in `ProblemG.txt`. It consists of a scalar dependent variable y (column 1) and 300 regressors $x_1, \ldots, x_{300}$ (columns 2 to 301). You may use the following command to create column names for the data and design matrix for question 3 below.

```
da <- read.table(''ProblemG.txt'')
y <- da[,1]
x <- da[,2:301]
c1 <- paste(''x'',1:300,sep='''')
colnames(x) <- c1
X <- model.matrix(y~.+(x1+x2+x3+x4+x5+x6+x7+x8+x9+x10)^2,data=data.frame(x))
```

1. Obtain the LASSO regression for the data. Then, use 10-fold cross-validation to select a sparse model. Write down the fitted sparse model.

2. Obtain an elastic-net regression for the data with $\alpha = 0.5$. Then, use 10-fold cross-validation to select a sparse model. Write down the fitted sparse model.

3. (4 points) In addition to individual regressors, we add all pairwise interactions of $\{x_1, \ldots, x_{10}\}$. Obtain a LASSO regression for the resulting data set. Use 10-fold cross-validation to select a sparse model. Write down the fitted sparse model.

Problem H. (12 points) Consider Problem 6.41 of the textbook. It is concerned with fixing problems of cell phone relay towers. A problem was initially classified as low or high severity, simple or complex in complexity. The engineer assigned was rated as relatively new (novice) or expert (guru). Two time measurements were taken. There are the time to assess the problem and plan an attach and the time to implement the solution; both measured in hours. The data are given in `cellphone.txt` (you may ignore the last column, which is the sum of two time measurements).

1. Perform a multivariate analysis of variance analysis to check the effect of each factor. Perform statistical tests to confirm the findings.

2. Use multivariate multiple linear regression to test the main effect of each factor. Draw the conclusion.

3. (2 points) Will there be any difference in the conclusion if one uses implementation time and total resolution time as dependent variables? Why?

Problem I. (8 points) Consider the monthly simple returns of IBM, MSFT (Microsoft), and the S&P composite index from January 2000 to December 2015. The data are in the file `m-ibmmsftsp0015.txt`. The goal is to compare performance of the three assets before and after the sub-prime financial crisis. To this end, consider the first 72 data points (from 2000 to 2005) as the first sample and the last 72 data points (from 2010 to 2015) as the second sample. Thus, the observations from Row 73 to Row 120 should be excluded in the analysis. Answer the following questions:

1. (2 points) Compute the sample mean and sample covariance matrix of each sample.

2. Are the two samples have the same covariance matrix? Perform a test and draw the conclusion.

3. Are the mean returns of the two samples equal? Perform a proper test and draw the conclusion.