THE UNIVERSITY OF CHICAGO
Booth School of Business
Business 41912, Spring Quarter 2018, 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. The exam has two parts. 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 $\boldsymbol{A}'$. Also, the (i, j)th element of the matrix $\boldsymbol{A}$ is a_{ij}.

Part One. Basic concepts.

Problem A. (10 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

$$\boldsymbol{\mu} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 4 & -3 & 0 \\ -3 & 6 & 0 \\ 0 & 0 & 5 \end{bmatrix}.$$

Answer the following questions:

1. Find the joint distribution of $\boldsymbol{Z} = (X_1, X_2)'$.

2. Are X_1 and X_3 independent? Why?

3. Are $3X_1 - X_2$ and X_3 independent? Why?

4. Are (X_1, X_3) and X_2 independent? Why?

5. What is the conditional distribution of $X_1|X_2 = 1$?

Problem B. (10 points) Suppose that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are random samples from a p-dimensional multivariate distribution with mean $\boldsymbol{\mu}$ and positive-definite covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'$ be the sample covariance matrix and $\bar{\boldsymbol{X}}$ be the sample mean. Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ be another random sample from the same distribution. Denote the sample mean and sample covariance of this 2nd sample by $\bar{\boldsymbol{Y}}$ and $\boldsymbol{S}_2$, respectively. The two random samples are independent.

1. Obtain the mean and covariance matrix of $\boldsymbol{X}_1 - \boldsymbol{X}_2$?

2. Show that $E(\bar{\boldsymbol{X}}) = \boldsymbol{\mu}$.

3. Show that $\text{var}(\bar{\boldsymbol{X}}) = \frac{1}{n}\boldsymbol{\Sigma}$.

4. Show that $E(\boldsymbol{S}) = \boldsymbol{\Sigma}$.

5. What is the limiting distribution of $n(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})'(\boldsymbol{S} + \boldsymbol{S}_2)^{-1}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})$ as $n \to \infty$?

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 an 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 an m-dimensional random noise with mean zero and positive-definite covariance matrix $\boldsymbol{\Sigma}$. 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 and we assume that $\boldsymbol{\epsilon}_i$ are uncorrelated. 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 $\boldsymbol{\beta}$. Also, let $\boldsymbol{H} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$ be the hat matrix of the regression.

1. Let $\boldsymbol{u}$ be a n-dimensional vector of 1. Show that $\boldsymbol{u}'\widehat{\boldsymbol{\epsilon}} = \boldsymbol{0}$.

2. Show that $\sum_{i=1}^{n} h_{ii} = r + 1$.

3. (True or False) $h_{ii} > 0$ for all i.

4. (True or False) $\text{Cov}(\text{vec}(\widehat{\boldsymbol{\beta}})) = \boldsymbol{\Sigma}^{-1} \otimes (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$.

5. (True or False) Residuals $\widehat{\boldsymbol{\epsilon}}_i$ and $\widehat{\boldsymbol{\epsilon}}_j$ are uncorrelated if $i \neq j$.

6. (True or False) $E(\frac{1}{n}\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}) = \boldsymbol{\Sigma}$.

Problem D. (10 points) Let $\boldsymbol{x} = (x_1, \ldots, x_p)$ be a p-dimensional random variable that has a continuous distribution with mean zero and positive-definite covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be a random sample from $\boldsymbol{x}$ and denote the data matrix by $\boldsymbol{X}$, which is a $n \times p$ matrix.

1. Define the first principal component of $\boldsymbol{x}$.

2. Let $(\lambda_i, \boldsymbol{e}_i)$ be the ith eigenvalue-eigenvector pair of $\boldsymbol{\Sigma}$, where $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ and $\boldsymbol{e}_i'\boldsymbol{e}_i = 1$. What is the second principal component of $\boldsymbol{x}$?

3. Let $(\hat{\lambda}_i, \widehat{\boldsymbol{e}}_i)$ be the ith eigenvalue-eigenvector pair of $\boldsymbol{S}$, where $\hat{\lambda}_1 > \hat{\lambda}_2 > \cdots > \hat{\lambda}_p$ and $\boldsymbol{S}$ is the sample covariance matrix. True or False that $\hat{\lambda}_i$ is a consistent estimate of λ_i?

4. Under what condition on the data matrix $\boldsymbol{X}$ that the singular value decomposition of $\boldsymbol{X}$ provides a quick principal component analysis for $\boldsymbol{x}$?

5. Describe a weakness of principal component analysis.

Problem E. (8 points) Answer briefly the following questions.

1. Describe two methods for model selection in the linear regression analysis.

2. One can use multivariate analysis of variance to compare the mean vectors of several populations. Describe two assumptions used in such an analysis.

3. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be a realization of the p-dimensional time series model $\boldsymbol{x}_t = \boldsymbol{\mu} + \boldsymbol{a}_t - \boldsymbol{\theta}\boldsymbol{a}_{t-1}$, where $\{\boldsymbol{a}_t\}$ is a sequence of independent and identically distributed Gaussian random vector with mean $\boldsymbol{0}$ and positive-definite covariance matrix $\boldsymbol{\Sigma}$. Let $\bar{\boldsymbol{x}}$ be the sample mean of the data. What is $E(\bar{\boldsymbol{x}})$?

4. What is $\text{Cov}(\bar{\boldsymbol{x}})$, where $\bar{\boldsymbol{x}}$ is defined in the prior question.

Part Two: Data analysis

Problem F. (22 points) Consider the measurements of blood glucose levels on three occasions for 50 women. The dependent variables y's represent fasting glucose measurements on the three occasions and the predictors x's are glucose measurements 1 hour after sugar intake. The data are available on the course web: `Glucose.DAT`.

1. Find the maximum and minimum of Spearman's rho between y and x.
2. Find the maximum and minimum of Kendall's tau between y and x.
3. Let $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ be the covariance matrices of y's and x's, respectively. Test the hypothesis $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ versus $H_a : \boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. Draw a conclusion.
4. Let $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ be the mean vectors of y's and x's, respectively. Test the hypothesis $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $H_a : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. Draw a conclusion.
5. Obtain the least squares estimate of $\boldsymbol{\beta}$ and the residual covariance matrix of the linear regression
$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i,$$
where $\boldsymbol{Y} = (y_1, y_2, y_3)'$ and $\boldsymbol{X} = (1, x_1, x_2, x_3)'$.
6. Is the overall regression statistically significant? Why?
7. Test the significance of x_1 given x_2 and x_3.
8. Test the significance of x_3 given x_1 and x_2.
9. (3 points) Let $\boldsymbol{x}_0 = (1, 100, 100, 100)'$. Compute simultaneous 95% confidence intervals for the elements of $E(\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}_0)$. [You should update the "ama.R" file. The new version of `mmlr` provides the $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ in the output.]
10. (3 points) Calculate simultaneous 95% prediction intervals for the elements of $E(\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}_0)$.

Problem G: (8 points) An experiment involved a 2×4 design with 4 replications was conducted to study the properties of bar steel. The two factors are rotational velocity [A_1 (fast) and A_2 (slow)] and lubricants [B_i for $i = 1, 2, 3, 4$]. The measurements are y_1 = ultimate torque and y_2 = ultimate strain. The data are available on course web `Barsteel.DAT`. Use the Wilk's statistic to answer the following three questions.

1. Is there any interaction between the two factors? Why?
2. Is the effect of rotational velocity significant? Why?
3. Is the effect of lubricants significant? Why?

4. List the assumptions used in the analysis.

Problem H: (10 points) Consider the data set in `ProblemH.txt`. It consists of two scalar dependent variables [y (column 1) and $y1$ (column 2)] and 300 regressors $x_1, \ldots, x_{300}$ (columns 3 to 302). You may use the following command to create column names for the data and design matrix for Question 3 below.

```
da <- read.table(``ProblemH.txt'')
y <- da[,1]; y1 <- da[,2]
x <- da[,3:302]
c1 <- paste(``x'',1:300,sep=``'')
colnames(x) <- c1
X <- model.matrix(y1~.+(x1+x2+x3+x4+x5+x6+x7+x8+x9+x10)^2,data=data.frame(x))
X <- X[,-1]  ## remove the intercept from the design matrix.
```

1. Apply LASSO to regress y on x and use 10-fold cross-validation to select the penalty parameter. What is the selected penalty parameter? With this choice of λ, identify the predictors of the LASSO regression with absolute coefficient greater than 0.5.

2. Apply elastic-net regress y on x with $\alpha = 0.5$ (in package **glmnet**). Again, use 10-fold cross-validation to select the penalty parameter. What is the selected penalty parameter λ? Identify the predictors of the resulting regression with absolute coefficient greater than 0.5.

3. (4 points) In addition to the individual predictors, we add all pairwise interactions of $\{x_1, \ldots, x_{10}\}$. Apply elastic-net to regress $y1$ on X with $\alpha = 0.8$. Identify the predictors with absolute coefficient greater than 0.5. Note that this problem concerns new dependent variable and new predictors.

Problem I. (10 points) The file `m-apca40stocks.txt` contains monthly returns of 40 stocks for two years. The first row of the file contains the codes for companies, and should be removed. Here we have $p = 40$ and $n = 24$.

1. Perform the traditional principal component analysis. Identify the variable that has the maximum loading (in absolute value) of the first two principal components.

2. What are the variances of the first two principal components?

3. What percentage of total variability is explained by the first two principal components?

4. Perform a sparse principal component analysis. For instance, you may use the package **elasticnet** with command `spca` and subcommand `para=c(2,1)`. What percentages of total variability are explained by the first two sparse principal components?

5. Identify the second sparse principal component. What does it mean?