THE UNIVERSITY OF CHICAGO
Booth School of Business
Business 41912, Spring Quarter 2018, Mr. Ruey S. Tsay

Solutions to Midterm

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 $\boldsymbol{Z} = (X_1, X_2)'$.

$$\boldsymbol{Z} \sim N\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 4 & -3 \\ -3 & 6 \end{bmatrix} \right)$$

2. Are X_1 and X_3 independent? Why?

   Yes, X_1 and X_3 are independent, because their covariance is zero.

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

   Yes, because (X_1, X_2) is independent of X_3.

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

   No, because X_1 and X_2 are dependent, even though X_3 and X_2 are independent.

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

   By direct calculation, the conditional distribution is $N(1.5, 2.5)$.

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 $\boldsymbol{X}_1 - \boldsymbol{X}_2$?

   Mean is zero and covariance matrix is $2\boldsymbol{\Sigma}$.

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

   Proof: $E(\bar{\boldsymbol{X}}) = \frac{1}{n}\sum_{i=1}^{n} E(\boldsymbol{X}_i) = \boldsymbol{\mu}$.

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

   Proof: $\text{var}(\bar{\boldsymbol{X}}) = \frac{1}{n^2}\sum_{i=1}^{n} \text{var}(\boldsymbol{X}_i) = \frac{1}{n}\boldsymbol{\Sigma}$.

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

   Proof: Note that $(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})' = \boldsymbol{X}_i\boldsymbol{X}_i' - \boldsymbol{X}_i\bar{\boldsymbol{X}}' - \bar{\boldsymbol{X}}\boldsymbol{X}_i' + \bar{\boldsymbol{X}}\bar{\boldsymbol{X}}'$ and without loss of generality, we may assume $\boldsymbol{\mu} = \boldsymbol{0}$. Then, taking the summation and using result of question 3, we have

$$E(\boldsymbol{S}) = \frac{1}{n-1}\left(n\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\right) = \boldsymbol{\Sigma}.$$

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 $n \to \infty$?

   Answer: χ^2_p.

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 $\boldsymbol{u}'\widehat{\boldsymbol{\epsilon}} = \boldsymbol{0}$.

   Proof: Since the first column of $\boldsymbol{Z}$ is $\boldsymbol{u}$, and we have $\boldsymbol{Z}'\widehat{\boldsymbol{\epsilon}} = \boldsymbol{0}$, we have $\boldsymbol{u}\widehat{\boldsymbol{\epsilon}} = \boldsymbol{0}$.

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

   Proof: $\sum_{i=1}^{n} h_i i = tr(\boldsymbol{H}) = tr[(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Z}] = tr(\boldsymbol{I}_{p+1}) = p + 1$.

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

   True, because $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ is positive definite.

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

   False, should be $\boldsymbol{\Sigma}$.

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

   False, they are correlated (sum to zero).

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

   False, the denominator is $n - p - 1$, not n.

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 $\boldsymbol{x}$.

   Answer: The linear combination $\boldsymbol{a}'\boldsymbol{x}$ such that $\text{var}(\boldsymbol{a}'\boldsymbol{x})$ is maximized subject to the constraint $\boldsymbol{a}'\boldsymbol{a} = 1$.

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 $\boldsymbol{x}$?

   Answer: $y = \boldsymbol{e}_2'\boldsymbol{x}$ is the 2nd principal component.

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?

   Answer: True, eigenvalues are consistent estimates.

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 $\boldsymbol{x}$?

   Answer: The column means of $\boldsymbol{X}$ are zero.

5. Describe a weakness of principal component analysis.

   Answer: Results depend on the scales of the components of $\boldsymbol{x}$. Or PCA is sensitive to outliers.

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

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

   Answer: Any two of (a) forward selection, (b) backward elimination, (c) stepwise regression, (d) C_p criterion or other information criteria.

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

   Answer: The samples are indepdendent across the population, or the sample sizes are sufficiently large, or the covariance matrices are the same across the populations, or the samples for each population are random samples.

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 $\mathbf{0}$ and positive-definite covariance matrix $\boldsymbol{\Sigma}$. Let $\bar{\boldsymbol{x}}$ be the sample mean of the data. What is $E(\bar{\boldsymbol{x}})$?

   Answer: $E(\bar{\boldsymbol{x}}) = \boldsymbol{\mu}$, which is the mean of each $\boldsymbol{x}_i$.

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

   Answer: Slight modification of the VAR(1) case shown in the lecture. From the model, $\mathrm{var}(\boldsymbol{x}_t) = \boldsymbol{\Sigma} + \boldsymbol{\theta}\boldsymbol{\Sigma}\boldsymbol{\theta}'$ and $\mathrm{cov}(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}) = -\boldsymbol{\theta}\boldsymbol{\Sigma}$. Other covariances are zero.

$$\begin{aligned}
\mathrm{Cov}(\bar{\boldsymbol{x}}) &= \frac{1}{n^2}\mathrm{Cov}(\sum_{i=1}^{n}\boldsymbol{x}_i) \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{n}\mathrm{var}(\boldsymbol{x}_i) + 2\sum_{i=2}^{n}\mathrm{cov}(\boldsymbol{x}_i, \boldsymbol{x}_{i-1})\right] \\
&= \frac{1}{n^2}\left[n(\boldsymbol{\Sigma} + \boldsymbol{\theta}\boldsymbol{\Sigma}\boldsymbol{\theta}') - 2(n-1)\boldsymbol{\theta}\boldsymbol{\Sigma}\right] \\
&= \frac{1}{n}[\boldsymbol{\Sigma} + \boldsymbol{\theta}\boldsymbol{\Sigma}\boldsymbol{\theta}' - 2\frac{n-1}{n}\boldsymbol{\theta}\boldsymbol{\Sigma}].
\end{aligned}$$

Part Two: Data analysis
Please see the R output for further details.

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.

   Answer: 0.346 and -0.137.

2. Find the maximum and minimum of Kendall's tau between y and x.

   Answer: 0.230 and -0.091.

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.

   Answer: The Box M test is 110.83 with p-value close to zero so that the two covariance matrices are different at the 5% level.

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.

   Answer: The Behrens test is 159.1 with p-value close to 0. The null hypothesis is rejected at the 5% level. The same conclusion is reched if Hotelling T^2 test was used.

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 $\boldsymbol{X} = (1, x_1, x_2, x_3)'$.

   Answer: The results are given below:

```
Beta-Hat matrix:
       y1     y2     y3
   54.870 65.679 58.106
y1  0.054 -0.048  0.018
y2 -0.024  0.163  0.012
y3  0.107 -0.036  0.125
LS residual covariance matrix:
       y1     y2     y3
y1 91.689 20.214  3.136
y2 20.214 67.549 14.509
y3  3.136 14.509 70.508
```

6. Is the overall regression statistically significant? Why?

   Answer: Yes, the test statistic is 18.56 with p-value 0.029.

7. Test the significance of x_1 given x_2 and x_3.

   Answer: The test statistic is 2.66 with p-value 0.45 so that x_1 is not statistically significant given x_2 and x_3.

8. Test the significance of x_3 given x_1 and x_2.

   Answer: The test statistic is 8.33 with p-value 0.04 sothat x_3 is statistically significant 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.]

   Answer: The results are (based on a simple R code I wrote)

```
Point prediction:
    y1     y2     y3
68.587 73.665 73.626
Simultaneous C.I. with prob 0.95
        [,1]    [,2]
[1,] 63.9120 73.2616
[2,] 69.6522 77.6772
[3,] 69.5261 77.7249
```

10. (3 points) Calculate simultaneous 95% prediction intervals for the elements of $E(\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}_0)$.

    Answer:

```
Simultaneous P.I. with prob 0.95
        [,1]    [,2]
[1,] 38.5499 98.6237
[2,] 47.8833 99.4461
[3,] 47.2855 99.9656
```

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?

   Answer: No, the interaction is not statistically significant with test statistic 0.93 and p-value 0.95.

2. Is the effect of rotational velocity significant? Why?

   Answer: No, the velocity is not statistically significant. The test statistic is 0.692 with p-value 0.18.

3. Is the effect of lubricants significant? Why?

   Answer: Yes, the lubricant is statistically significant. The test statistic is 0.47 with p-value 1.87×10^{-4}.

4. List the assumptions used in the analysis.

   Answer: The observations are random sample and normally distributed.

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.

   Answer: The penalty parameter is 0.0892 and important predictors are variables 6, 15, 21, 55, and 66.

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.

   Answer: The penalty parameter is 0.141 and the selected predictors are variables 6, 15, 21, 55, and 66.

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.

   Answer: The predictors are variables 6, 15, 21, 55, 66 and the interaction $x_5 x_6$.

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.

   Answer: For the 1st PCA, the maximum loading is company 26. For the 2nd PCA, the maximum loading is company 30.

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

   Answer: 0.67 and 0.16, respectively.

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

   Answer: 61.4%

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?

   Answer: The first two components explain about 18.2% and 10.4%, respectively.

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

   Answer: The loading shows that the 2nd sparse principal component mainly consists of the 30th company. It indicates that the company appears to be an outlying company.