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

Solutions to Midterm

Part One. Basic concepts.

Problem A.

1. Bivariate normal as with mean $\boldsymbol{\mu}_z$ and covariance $\boldsymbol{\Sigma}_z$ given by
$$\boldsymbol{\mu}_z = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \quad \boldsymbol{\Sigma}_z = \begin{bmatrix} 6 & -2 \\ -2 & 4 \end{bmatrix}.$$
2. Univariate normal with mean and variance 17 and 21, respectively.
3. Bivariate normal with mean and covariance given by
$$\begin{bmatrix} 2.5 \\ 2.0 \end{bmatrix}, \quad \begin{bmatrix} 5 & 3 \\ 3 & 9 \end{bmatrix}.$$
4. Yes, because their covariance is zero. Let $\boldsymbol{c}_1 = (1, -1, 1)'$ and $\boldsymbol{c}_2 = (2, 1, -1)'$. Then, $\boldsymbol{c}_1'\boldsymbol{\Sigma}\boldsymbol{c}_2 = 0$.
5. χ^2_s.

Problem B.

1. 5-dimensional normal with mean zero and covariance matrix $\boldsymbol{\Sigma}$.
2. χ^2_5
3. χ^2_5
4. No longer χ^2_5 because $\text{cov}(\bar{\boldsymbol{X}})$ is not consistently estimated by $\boldsymbol{S}/n$.
5. $\frac{n-1}{n}\boldsymbol{\Sigma}$.

Problem C:

1. $\boldsymbol{\epsilon}_i$ are independent with mean zero and covariance matrix $\boldsymbol{\Sigma}$, which is time-invariant. Also, $E(\boldsymbol{\epsilon}_i|\boldsymbol{Z}_i) = \boldsymbol{0}$.
2. $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}(\boldsymbol{Z}'\boldsymbol{Y})$.

3. Because $\boldsymbol{Z}_i$ contains the constant 1 as one of its elements.

4. $\sqrt{n}[\text{vec}(\widehat{\boldsymbol{\beta}}) - \text{vec}(\boldsymbol{\beta})]$ is normally distributed with mean zero and covariance matrix $\boldsymbol{\Sigma} \bigotimes (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$.

5. $\text{cov}(\widehat{\boldsymbol{\epsilon}}_{(i)}) = \sigma_{ii}(\boldsymbol{I} - \boldsymbol{H})$, where σ_{ii} is the i-th diagonal element of $\boldsymbol{\Sigma}$ and $\boldsymbol{H} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$.

Problem D.

1. The data consist of a random sample from a multivariate distribution when the sample size is large. For small sample size, the population distribution is multivariate normal.

2. Yes, because the test statistic is both mean-adjusted and scale-adjusted. Simple derivation is available in the lecture note.

3. X_1 and X_2 are uncorrelated (or independent).

4. The data consist of random samples from the two populations. The sample size n_1 and n_2 are sufficiently large (compared with dimension). If the sample sizes are not sufficiently large, then one also assumes that the two populations have the same covariance matrix.

5. (1) Sample size is less than the number of explanatory variables. (2) The underlying model is sparse.

Problem E.

1. The EM algorithm and Markov chain Monte Carlos methods

2. The data are missing at random.

3. Transform the data into χ^2 random variates and compare the transformed data with quantiles of χ^2 distribution.

Part Two: Data analysis
Problem F. (21 points) Japanese Seishu wine.

1. The sample mean and sample variance of each explanatory variable are given below

```
> apply(X,2,mean)
        x1          x2          x3          x4          x5          x6
 4.2020000   1.5426667   0.8236667  -1.4066667   3.4883333   4.4040000
        x7          x8
15.9743333 121.7333333
> v1 <- cov(X)
> diag(v1)
```

```
          x1           x2           x3           x4           x5           x6
2.755448e-02 2.368920e-02 6.905851e-02 5.019264e+00 1.330833e-01 2.593559e-01
          x7           x8
3.512530e-01 1.541084e+03
```

2. The maximum Pearson correlation (in magnitude) between $\boldsymbol{Y}$ and $\boldsymbol{X}$ is 0.239 between "y2" and "x6".

3. Find the maximum Kendall's tau (in magnitude) between $\boldsymbol{Y}$ and $\boldsymbol{X}$ is 0.194 also between "y2" and "x6".

4. Use stepwise and AIC, the selected model is

$$y_{1i} = -7.48 + 1.29x_{1i} + 0.92x_{3i} + 0.08x_{4i} + 0.66x_{5i} - 0.005x_{8i} + \epsilon_t.$$

5. The $\widehat{\boldsymbol{\beta}}$ is

```
> m2 <- mmlr(y,X)
Beta-Hat matrix:
       y1     y2
   -4.140  4.935
x1  1.103 -0.955
x2  0.231 -0.222
x3  1.171  1.773
x4  0.111  0.048
x5  0.617 -0.058
x6  0.267  0.485
x7 -0.263 -0.209
x8 -0.004 -0.004
```

   Most of the estimates are not statistically significant, however.

6. The likelihood ratio test is 3.19 with p-value 0.53. Thus, the regressor $\boldsymbol{X}_3$ can be dropped.

7. The likelihood ratio test is 10.87 with p-value 0.028. Thus, at the 5% level, the regressor $\boldsymbol{X}_1$ contributes significantly to the regression.

Problem G. (10 points) This is a simulated example, the data generating model is

$$y_i = 5x_{30,i} - 5x_{50,i} + 2x_{1,i}x_{10,i} + \epsilon_i.$$

1. The LASSO results are shown in Figure 1 and 2. From the plots and cross-validation result, it seems that a fraction of $s \in [0.6, 1.0]$ would be fine. Thus, the solution is not unique. However, the dominating regressors are x_{30} and x_{50}. Since the interaction is not allowed in the design matrix, other variables might appear as a proxy.

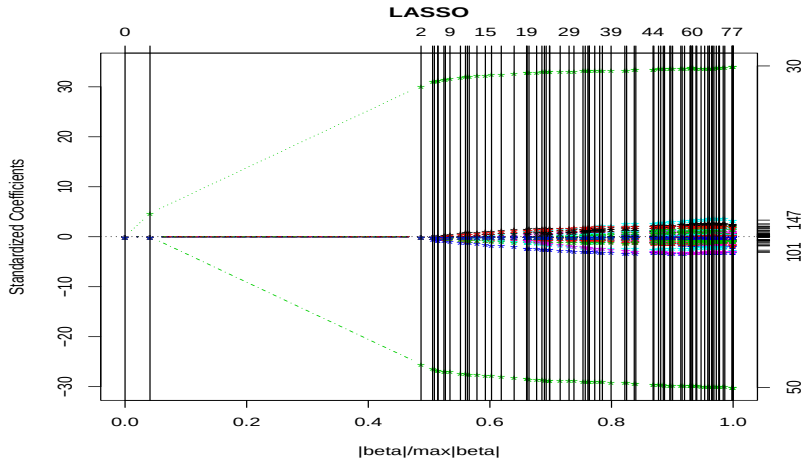

Figure 1: Output of LASSO for Problem G.

2. The `glmnet` results with $\alpha = 0.5$ are shown in Figures 3 and 4. Similar to part 1, the solution is not unique. For λ with $\ln(\lambda) \in [-2, 0]$ seems reasonable.

3. The LASSO results when interactions are allowed are shown in Figures 5 and 6. The data generating model should be included in the selected model. The fraction s seems to be between 0.8 and 1.0.

Problem H. Cell-phone data

1. The analysis is given below:

```
> fac1 <- factor(da[,1])
> fac2 <- factor(da[,2])
> fac3 <- factor(da[,3])
> Y <- as.matrix(da[,4:5])
> n1 <- manova(Y~fac1+fac2+fac3+fac1*fac2+fac1*fac3+fac2*fac3)
> summary(n1)
          Df  Pillai approx F num Df den Df    Pr(>F)
fac1       1 0.94651   70.774      2      8 8.189e-06 ***
fac2       1 0.98541  270.173      2      8 4.530e-08 ***
fac3       1 0.96983  128.573      2      8 8.287e-07 ***
fac1:fac2  1 0.72226   10.402      2      8  0.005951 **
fac1:fac3  1 0.18169    0.888      2      8  0.448407
fac2:fac3  1 0.22326    1.150      2      8  0.363995
Residuals  9
---
> summary(n1,test="Wilks")
```

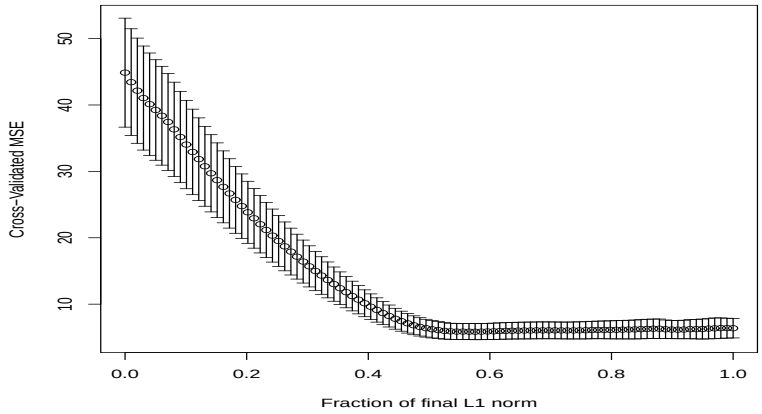

Figure 2: Cross-validation of LASSO for Problem G.

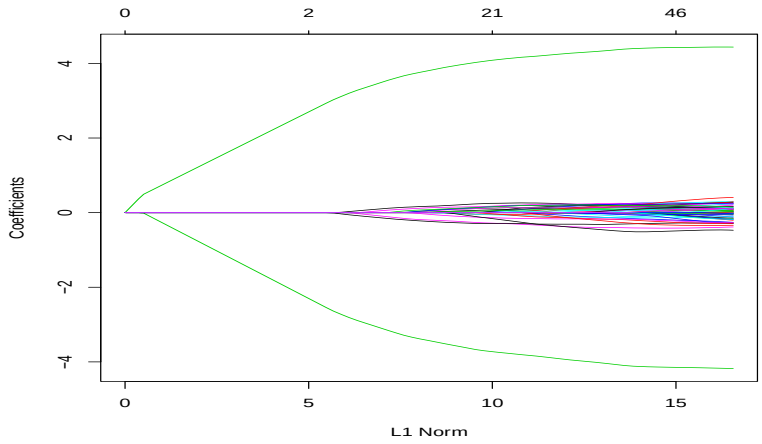

Figure 3: Output of `glmnet` for Problem G.

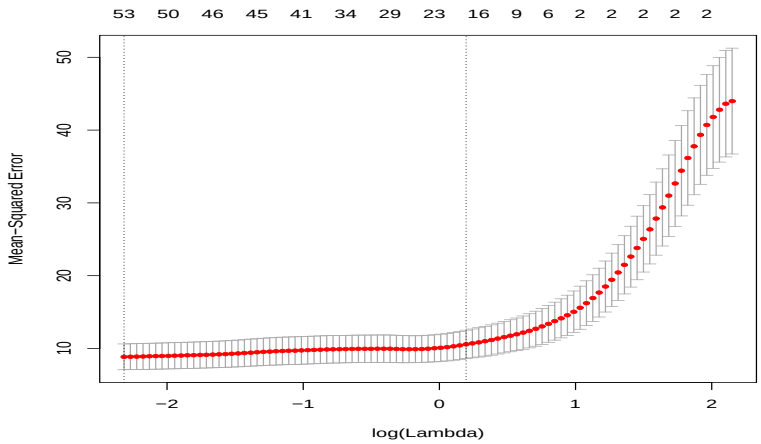

Figure 4: Cross-validation of `glmnet` for Problem G.

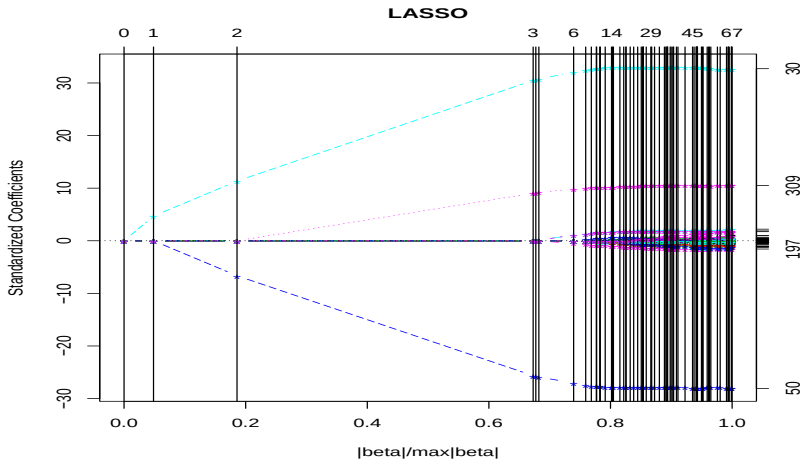

Figure 5: Output of LASSO for Problem G with interaction terms.

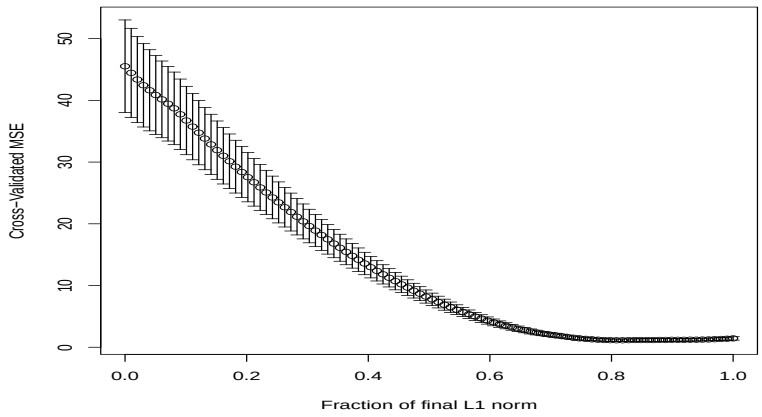

Figure 6: Cross-validation of LASSO for Problem G with interaction terms.

```
           Df   Wilks approx F num Df den Df    Pr(>F)
fac1        1 0.05349   70.774      2      8 8.189e-06 ***
fac2        1 0.01459  270.173      2      8 4.530e-08 ***
fac3        1 0.03017  128.573      2      8 8.287e-07 ***
fac1:fac2   1 0.27774   10.402      2      8  0.005951 **
fac1:fac3   1 0.81831    0.888      2      8  0.448407
fac2:fac3   1 0.77674    1.150      2      8  0.363995
Residuals   9
```

Based on Pillai or Wilks test, the interaction between severity and complexity and the three main effects are significant.

2. There are several ways to perform the analysis. I simply use the full model that includes main effects and all two-way interactions to form test. The result shows that the main effects are significant with likelihood ratio test 45.2 with p-value close to zero.

3. No, because the three possible dependent variables are linear combinations of two given. [You can perform the analysis to confirm it, but not required.]

Problem I. Monthly simple returns of IBM, MSFT (Microsoft), and the S&P composite index.

1. The sample means and covariance matrices are given below:

```
> x1 <- as.matrix(da[1:72,2:4])
> dim(da)
[1] 192   4
```

```
> x2 <- as.matrix(da[121:192,2:4])
> apply(x1,2,mean)
        msft          ibm           sp
-0.002916167  0.001442736 -0.001305014
> cov(x1)
            msft         ibm          sp
msft 0.012959379 0.007089793 0.002925037
ibm  0.007089793 0.009818525 0.003105217
sp   0.002925037 0.003105217 0.001924248
> apply(x2,2,mean)
       msft         ibm          sp
0.012797889 0.003497875 0.009153069
> cov(x2)
            msft          ibm           sp
msft 0.004580680 0.0010307783 0.0015974864
ibm  0.001030778 0.0020466945 0.0008785111
sp   0.001597486 0.0008785111 0.0014300346
```

2. Based on the Box-M test statistic, the null hypothesis of equal covariance matrices is rejected. The test statistic is 59.5 with p-value close to zero.

3. Based on Hotelling T^2 test (with un-equal covariances), one cannot reject the null hypothesis of equal mean vectors at the 5% level. The test statistic is 3.65 with p-value 0.31.