1. (15 pts) Consider 42 measurements on air-pollution variables recorded at 12:00 noon in the Los Angeles area on different days. The data are in Table 1.5 of the textbook.

(a) (2 pts) A principal component analysis is performed on the observed data. Based on the output, what percentage of the total variation is explained by the first principal component?
Answer: 87.3%

(b) (4 pts) How many principal components are needed to obtain a good approximation of the data? Why?
Answer: 2 components. They explain about 95% of the total variation.

(c) (2 points) Provide an interpretation of the first principal component.
Answer: It is simply the “solar” radiation.

(d) (3 points) Now, a principal component analysis is performed using the correlation matrix. How many components are needed to obtain a good approximation of the data?
Answer: 3 components are needed to provide a reasonable approximation. It would require 5 components to achieve 90% variation.

(e) (4 points) Provide interpretations for the first two principal components.
Answer: The first component is basically a weighted average of the pollutants. The second component is essentially a contrast between (solar radiation + ozone concentration) and (NO + NO2 + HC).

2. (20 pts) Again, consider air-pollution variables of Problem 1. Questions 1 and 2 below are for the principal factor model, whereas questions 3 to 5 are for maximum-likelihood factor model.

(a) (5 points) Use the results of principal component analysis on the correlation matrix to construct a principal factor model analysis with 3 common factors. Write down the loading matrix, that is, the $L$ matrix in

$$ x = \mu + LF + \epsilon. $$

Answer: The loading matrix is obtained by multiplying the first three loading vectors by the corresponding square root of the eigenvalues. The result is given below:
(b) (5 points) What are the specific variances (or uniquenesses) of the factor model? 
Answer: The specific variances can be obtained by $1 - \text{diag}(LL')$. The results are 0.263, 0.456, 0.275, 0.205, 0.319, 0.278, and 0.278.

(c) (3 points) A factor analysis with 3 common factors is estimated by the maximum-likelihood method. Is the model adequate at the 1% level? Why? 
Answer: Yes, the large sample test has a $p$ value 0.0229 so that the model cannot be rejected at the 1% level.

(d) (2 points) Compute the communalities of the factor model. 
Answer: They are the diagonal elements of $LL'$. That is, 0.737, 0.544, 0.725, 0.795, 0.681, 0.722, and 0.722.

(e) (5 points) Based on the fitted factor model, what is the estimated covariance matrix $\hat{\Sigma} = \hat{L} \hat{L}' + \hat{\Psi}$ 
Answer: The covariance matrix is

```r
> S=LL+diag(psi)
> print(S,digits=3)
[1,] 1.0000 -0.1430 -0.3960 -0.3570 -0.0462 -0.2750  0.3610
[2,] -0.1431  1.0000  0.2390 -0.2460  0.1465  0.6130  0.0750
[3,] -0.3956  0.2390  1.0000  0.5380  0.6126  0.4010  0.3340
[4,] -0.3566 -0.2460  0.5380  1.0000  0.4635 -0.1340  0.2020
[5,] -0.0462  0.1470  0.6130  0.4630  1.0000  0.2590  0.5850
[6,] -0.2746  0.6130  0.4010 -0.1340  0.2585  1.0000  0.1040
[7,]  0.3613  0.0750  0.3340  0.2020  0.5850  0.1040  1.0000
```

3. (12 points) Consider the following four variables:

- $X_1$: 1973 nonprimary homicides
- $X_2$: 1973 primary homicides (homicides involving family or acquaintances)
- $Y_1$: 1970 severity of punishment (median months served)
• $Y_2$: 1970 certainty of punishment (number of admissions to prison divided by number of homicides)

The correlation matrix of $(X_1, X_2, Y_1, Y_2)'$ is

$$ R = \begin{bmatrix}
1.0 & 0.615 & -0.111 & -0.266 \\
0.615 & 1.0 & -0.195 & -0.085 \\
-0.111 & -0.195 & 1.0 & -0.269 \\
-0.266 & -0.085 & -0.269 & 1.0
\end{bmatrix}. $$

(a) (4 points) Find the sample canonical correlations between $X = (X_1, X_2)'$ and $Y = (Y_1, Y_2)'$.

Answer: The canonical correlations are $0.327$ and $0.171$.

(b) (4 points) Find the first canonical pair $\hat{U}_1$ and $\hat{V}_1$.

Answer: The vectors are $(1, -0.003)'$ and $(-0.524, -0.851)'$.

(c) (4 points) Find the second canonical pair $\hat{U}_2$ and $\hat{V}_2$.

Answer: The vectors are $(-0.523, 0.852)'$ and $(-0.923, -0.384)'$.

4. (16 pts) Consider the crude-oil samples of Problem 11.30 on page 661 of the textbook. We focus on two zones of sandstone, namely, Sub-Mulinia ($\pi_2$) and Upper ($\pi_3$) so that the sample sizes are 11 and 38, respectively. Assume that data are normally distributed with $N(\mu_i, \Sigma_i)$ for population $\pi_i$. Based on the R output, answer the following questions:

(a) (4 points) Test $H_0: \Sigma_2 = \Sigma_3$ vs. $H_a: \Sigma_2 \neq \Sigma_3$. What is the test statistic? Draw your conclusion using the 5% significance level.

Answer: Use the Box test. The test statistic is 18.77 with a $p$ value 0.22 so that we cannot reject the null hypothesis of equal covariance.

(b) (4 points) Test $H_0: \mu_2 = \mu_3$ vs. $H_a: \mu_2 \neq \mu_3$. What is the test statistic? Draw your conclusion using the 5% significance level.

Answer: Use the Behrens test. The test statistic is 106 with a $p$ value 0.00002. Thus, the null hypothesis of equal mean is rejected.

(c) (4 points) Construct Fisher’s (sample) linear discriminant function for the two populations.

Answer: The linear discriminant function is

$$ y = -0.711x_1 + 0.125x_2 - 9.978x_3 + 2.852x_4 + 0.006x_5 - 12.336. $$

(d) (4 points) Assume equal costs and equal prior probabilities. Assign the new observation $x_o = (5.0, 48.0, 0.1, 8.5, 4.0)'$ to either population.

Answer: For $x_o$, $y = 25.716 - 12.336 = 13.38 > 0$ so that the data point belongs to population $\pi_2$. 

3
5. (15 pts) We just experienced a severe financial crisis and like to examine the relationship between companies based the performance of their stocks. Consider the daily simple returns of the following nine companies: (1) AIG, (2) Bank of America (BAC), (3) Goldman Sachs (GS), (4) J.P. Morgan (JPM), (5) Morgan Stanley (MS), (6) Wells Fargo (WFC), (7) Boeing (BA), (8) Intel (INTC), and (9) Procter & Gamble (PG). The data span is from January 2006 to December 2009. The sample correlations of the stocks are given below.

```
> print(V,digits=2)
     aig  bac  gs  jpm  ms  wfc  ba  intc  pg
aig  1.00 0.51 0.37 0.45 0.41 0.42 0.34 0.31 0.28
bac  0.51 1.00 0.68 0.82 0.61 0.85 0.44 0.47 0.42
gs   0.37 0.68 1.00 0.72 0.81 0.66 0.49 0.51 0.44
jpm  0.45 0.82 0.72 1.00 0.84 0.46 0.52 0.48
ms   0.41 0.61 0.81 0.60 1.00 0.57 0.50 0.50 0.49
wfc  0.42 0.85 0.66 0.84 0.57 1.00 0.44 0.47 0.42
ba   0.34 0.44 0.49 0.46 0.50 0.44 1.00 0.50 0.51
intc 0.31 0.47 0.51 0.52 0.50 0.47 0.50 1.00 0.50
pg   0.28 0.42 0.44 0.48 0.49 0.42 0.51 0.50 1.00
```

Use the correlations to construct a distance measure \( d_{ij} = 1 - \rho_{ij} \), where \( \rho_{ij} \) denotes the correlation between stocks \( i \) and \( j \).

(a) (5 points) Use the distance and the hierarchical clustering method with single linkage to perform a classification. The first cluster is (BAC,WFC). Show the distance of \([\text{AIG},(\text{BAC},\text{WFC})]\). Draw the resulting dendrogram. Comment on the dendrogram.

Answer: The distance between \([\text{AIG},(\text{BAC},\text{WFC})]\) is 0.49. The dendrogram is in Figure 1. Except for AIG, the classification seems reasonable.

(b) (5 points) Use the distance and hierarchical clustering method with complete linkage to perform a classification. The first cluster is (BAC,WFC). Show the distance of \([\text{BA},(\text{BAC},\text{WFC})]\). Draw the resulting dendrogram. Comment on the dendrogram.

Answer: The distance between \([\text{BA},(\text{BAC},\text{WFC})]\) is 0.56. The dendrogram is in Figure 2. The classification matches with the common sense. In particular, AIG is the one companies that would fail without the government bailout.

(c) (5 points) Use the distance and the \( k \)-means method with \( k = 3 \) to perform a classification. What are the cluster sizes? List the members of each cluster. Comment on the clustering results.

Answer: The sizes are 4, 2, and 3. The members are

-  G1: AIG, BA, INTC, and PG.
6. (10 pts) Consider the quarterly U.S. real gross domestic product (GDP) and unemployment rate (UN) from 1948 to 2004 for 228 observations. Let $x_t = (GDP_t, UN_t)'$ and $y_t = (GDP_t - GDP_{t-1}, UN_t - UN_{t-1})'$. The sample canonical correlations between $x_{t-1}$ and $y_t$ are 0.20552 and 0.1285, respectively. Assume that $x_t$ follows a bivariate normal distribution with a positive definite covariance matrix. Let $\Sigma_{12} = \text{cov}(x_{t-1}, y_t)$. One way to test that $x_{t-1}$ and $y_t$ are independent is to test $H_0 : \Sigma_{12} = 0$ vs. $H_a : \Sigma_{12} \neq 0$. Perform the test by computing a proper test statistic and compute its $p$ value. Draw your conclusion using the 5% significant level.

Answer: Using the large sample inference of Section 10.6 with Bartlett correction, the...
Figure 2: Dendogram for complete linkage
A test statistic is

\[ T = -\left[228 - 1 - 0.5(2 + 2 + 1)\right] \sum_{i=1}^{2} \ln(1 - \rho_i^2) = 13.43. \]

Comparing with a chi-squared distribution with 4 degrees of freedom, the \( p \) value of the statistic is 0.0094. Therefore, the null hypothesis is rejected at the 5\% level. The null hypothesis would be marginally rejected at the 1\% level.

7. (12 pts) Answer the following questions briefly.

(a) (3 points) Consider the multivariate-multiple linear regression

\[ Y_{n \times m} = Z_{n \times (r+1)} \beta_{(r+1) \times m} + \epsilon_{n \times m}. \]

Write down the assumptions that the least-squares estimate is the best linear unbiased estimator of \( \beta \).

Answer: \( E(\epsilon) = 0 \). Let \( \epsilon_i \) denotes the \( i \)th row of \( \epsilon \), then \( \text{Var}(\epsilon_i) = \Sigma \), which is positive definite and \( \epsilon_i \) and \( \epsilon_j \) is uncorrelated for \( i \neq j \). In addition, \( \epsilon_i \) are normally distributed. Of course, \( Z \) is a full rank design matrix.

(b) (3 points) Consider a scalar response variable \( Y \) and a vector of explanatory variables \( X = (X_1, \ldots, X_p) \). The sliced inverse regression can be used to find a proper model for \( Y \). Describe the conditions under which the sliced inverse regression works well.

Answer: (1) The model is \( Y = g(\beta_1'X, \ldots, \beta_k'X, \epsilon) \), where \( \beta_i \) are \( p \)-dimensional real vectors, \( \epsilon \) denotes noises, and \( g(\cdot) \) is a well defined smooth function. (2) The linear design condition holds.

(c) (2 points) State the conditions under which the independent component analysis differs from the usual principal component analysis.

Answer: When the data are not normally distributed.

(d) (2 points) State the effects on the least-squares estimation in a multiple linear regression when the serial correlations of the residuals are overlooked.

Answer: (1) Variances of the least-squares estimates of the coefficients are improperly estimated. (2) May lead to inconsistent estimates in some situation.

(e) (2 points) Describe a method to detect outliers in a random sample of size \( n \) from a multivariate normal distribution.

Answer: Use the chi-squared QQ plot.