STA 4702/5701 – Exam 3 Practice Problems

Q.1. A study considered a model involving subway stations in Tehran, Iran. The authors had 2 sets of variables, each measured for each of the n = 22 subway stations.

$X^{(1)}$ = Population, Number of Workers in a particular economic sector, Degree of Functional Mix, Place-to-Movement, and Place-Through-Movement (p=5)

$X^{(2)}$ = Frequency of Train Services, Number of Stations w/in 45 minutes travel time, Passenger Frequency, Proximity to Central Business District, Node-to-Movement, Node-Through-Movement (q=6)

p.1.a. The eigenvalues of $R_{11}^{-1/2} R_{12} R_{22}^{-1} R_{21} R_{11}^{-1/2}$ are:

0.78420220 0.43837667 0.22056279 0.08968695 0.02091465

What is the correlation between the first canonical variate for $X^{(1)}$ and the first canonical variate for $X^{(2)}$?

What is the correlation between the second canonical variate for $X^{(1)}$ and the second canonical variate for $X^{(2)}$?

What is the correlation between the first canonical variate for $X^{(1)}$ and the second canonical variate for $X^{(1)}$?

p.1.b. Test H_0: $\boldsymbol{\Sigma}_{12} = \mathbf{0}$ p.1.c. Test H_0: $\rho_3 = \rho_4 = \rho_5 = 0$

Q.2. A multivariate multiple regresiion model was fit, relating m=3 texture scores to r=5 physiochemical predictors.

Y_1 = Hardness, Y_2 = Gumminess, Y_3 = Chewiness

Z_1 = Moisture, Z_2 = Amylase, Z_3 = Water Absorption, Z_4 = Swelling, Z_5 = Solids Content

Two models were fit:

Model 1: $E\{Y_k\} = \beta_{0k} + \beta_{1k} Z_1 + \beta_{2k} Z_2 + \beta_{3k} Z_3 + \beta_{4k} Z_4 + \beta_{5k} Z_5 \quad k = 1,2,3$

Model 2: $E\{Y_k\} = \beta_{0k} + \beta_{1k} Z_1 + \beta_{2k} Z_2 \quad k = 1,2,3$

Results for Model 1 are given below.

```
Response Y1 :
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -10495.060   4415.928  -2.377   0.0258 *  
Z1             402.449    235.203   1.711   0.1000 .  
Z2             227.385     38.285   5.939 3.96e-06 ***
Z3              11.408      9.367   1.218   0.2351    
Z4              -1.996      7.945  -0.251   0.8038    
Z5              15.393     24.226   0.635   0.5312    
Residual standard error: 424.7 on 24 degrees of freedom
Multiple R-squared:  0.6532,    Adjusted R-squared:  0.581 
F-statistic: 9.042 on 5 and 24 DF,  p-value: 6.151e-05

Response Y2 :
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -8534.728   2595.555  -3.288   0.0031 ** 
Z1            324.992    138.245   2.351   0.0273 *  
Z2            141.141     22.503   6.272 1.75e-06 ***
Z3              5.280      5.505   0.959   0.3471    
Z4              1.671      4.670   0.358   0.7236    
Z5             15.271     14.240   1.072   0.2942    
Residual standard error: 249.6 on 24 degrees of freedom
Multiple R-squared:  0.6516,    Adjusted R-squared:  0.5791 
F-statistic: 8.979 on 5 and 24 DF,  p-value: 6.475e-05

Response Y3 :
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7092.0228  2566.9409  -2.763 0.010822 *  
Z1            325.9379   136.7212   2.384 0.025385 *  
Z2             96.1867    22.2550   4.322 0.000233 ***
Z3              3.3346     5.4448   0.612 0.545998    
Z4             -0.0744     4.6183  -0.016 0.987280    
Z5             17.2643    14.0826   1.226 0.232118    
Residual standard error: 246.9 on 24 degrees of freedom
Multiple R-squared:  0.4692,    Adjusted R-squared:  0.3586 
F-statistic: 4.243 on 5 and 24 DF,  p-value: 0.006619
```

p.2.a. Give the predicted value for each response when Z_1=15, Z_2=24, Z_3=230, Z_4=235, Z_5=10

p.2.b. The ML estimates of $\Sigma = V\{\mathbf{Y}\}$ for models 1 and 2 are given below: Test
H_0: $\beta_{31}=\beta_{32}=\beta_{33}=\beta_{41}=\beta_{42}=\beta_{43}=\beta_{51}=\beta_{52}=\beta_{53}=0$

```
> Y <- cbind(Y1,Y2,Y3)
> n <- nrow(Y)
> Z1 <- cbind(rep(1,n),X1,X2,X3,X4,X5)
> Z2 <- cbind(rep(1,n),X1,X2)
> beta.hat1 <- solve(t(Z1)%*%Z1) %*% t(Z1) %*% Y
> beta.hat2 <- solve(t(Z2)%*%Z2) %*% t(Z2) %*% Y
> 
> E1 <- Y - Z1 %*% beta.hat1
> E2 <- Y - Z2 %*% beta.hat2
> 
> (Sigma.hat <- (1/n) * (t(E1) %*% E1))
          Y1       Y2       Y3
Y1 144267.51 79026.95 62432.82
Y2  79026.95 49840.82 42928.50
Y3  62432.82 42928.50 48747.95
> (Sigma.hat1 <- (1/n) * (t(E2) %*% E2))
          Y1       Y2       Y3
Y1 155187.15 85622.10 67251.98
Y2  85622.10 54898.31 46664.51
Y3  67251.98 46664.51 52318.60
> det(Sigma.hat); det(Sigma.hat1)
[1] 9.54529e+12
[1] 1.33586e+13
```

Q.3. A study compared n = 40 lager beers in terms of Total phenolic content, melanoidin content, and p = 5 measures of antioxidant activity. Consider a principal component analysis of the 5 antioxidant activity variables (dsa, asa, orac, rp, and mca) based on the Correlation matrix.

```
> X <- cbind(dsa,asa,orac,rp,mca)
>
> (R <- cor(X))
           dsa       asa      orac        rp       mca
dsa  1.0000000 0.4551698 0.5360284 0.6132432 0.5406189
asa  0.4551698 1.0000000 0.2003063 0.6613946 0.3522524
orac 0.5360284 0.2003063 1.0000000 0.3189525 0.1791062
rp   0.6132432 0.6613946 0.3189525 1.0000000 0.3743024
mca  0.5406189 0.3522524 0.1791062 0.3743024 1.0000000
>
> eigen(R)$val
[1] 2.7416852 0.9031943 0.7426515 0.3568033 0.2556657
> eigen(R)$vec
           [,1]       [,2]       [,3]       [,4]       [,5]
[1,] -0.5224466 -0.2277687  0.1623841 -0.4372943  0.6764437
[2,] -0.4468155  0.4591393 -0.3951008  0.5923713  0.2872952
[3,] -0.3447843 -0.8047989 -0.1794726  0.3682451 -0.2561401
[4,] -0.5018368  0.2335563 -0.3396039 -0.5144648 -0.5600058
[5,] -0.3958397  0.1872506  0.8185264  0.2399821 -0.2840268
```

p.3.a. Give the first principal component of the standardized variables. How would you interpret it?

p.3.b. What proporion of the standardized sample variance is due to the first principal component?

p.3.c. Give the cumulative proportion of variation due to components 1:5.

p.3.d. Compute the correlation between orac and the 2nd principal component.

p.3.e. Compute a 95% Confidence Interval for λ_1.

Q.4. A study considered agricultural production for n = 22 countries in the 1950s. The variables were: Agricultural output ($1million), population active in agriculture (1000s), arables land equivalent (1000s of acres), and productive livestock (1000s of animals). The correlation matrix, its eigenvalues and eigenvectors are given below.

```
> (R <- cor(X))
          x1        x2        x3        x5
x1 1.0000000 0.4737335 0.9635610 0.8761381
x2 0.4737335 1.0000000 0.5720992 0.6960911
x3 0.9635610 0.5720992 1.0000000 0.9449781
x5 0.8761381 0.6960911 0.9449781 1.0000000
> round(eigen(R)$val,4)
[1] 3.2981 0.6057 0.0782 0.0180
> round(eigen(R)$vec,4)
        [,1]    [,2]    [,3]    [,4]
[1,] -0.5124 -0.4121 -0.5900  0.4685
[2,] -0.4018  0.8729 -0.2761 -0.0193
[3,] -0.5359 -0.2612  0.0103 -0.8028
[4,] -0.5374  0.0007  0.7587  0.3682
```

p.4.a. For the factor analytic model with m = 2, compute estimates for L and Ψ

p.4.b. What proportion of the standardized sample variance is due to the first factor?

p.4.c. Maximum likelihood estimates of L_Z and Ψ are given below along with the determinants of Sigma-hat under the m=1 model, and R. Test whether m = 1.

```
> (Sigma.hat <- mlfa$loadings %*% t(mlfa$loadings) + diag(mlfa$uniquenesses))
          x1        x2        x3        x5
x1 1.0000003 0.5540950 0.9616386 0.9114152
x2 0.5540950 0.9999992 0.5734351 0.5434864
x3 0.9616386 0.5734351 1.0002037 0.9432272
x5 0.9114152 0.5434864 0.9432272 0.9999995
> det(Sigma.hat)
[1] 0.005581524
> det(R)
[1] 0.002811384

Call:
factanal(x = X, factors = 1)

Uniquenesses:
   x1    x2    x3    x5
0.071 0.670 0.005 0.106

Loadings:
   Factor1
x1 0.964
x2 0.575
x3 0.998
x5 0.945

               Factor1
SS loadings      3.149
Proportion Var   0.787
```

Q.5. A discriminant analysis is conducted to classify NHL and EPL players by Height and Weight. Random samples of $n_{NHL} = n_{EPL} = 100$ players to generate Fisher's discriminant function to classify players by league. The results for the 2 samples are given below.

```
      Xbar1   Xbar2    Diff Spooled          INV(Sp)
   73.1862  71.168  2.0182  6.4397  29.4987  0.3307 -0.0383
  202.2000 166.260 35.9400 29.4987 254.7638 -0.0383  0.0084
```

p.5.a. Compute $\hat{\mathbf{a}}' = \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)' S_{\text{pooled}}^{-1}$ and $\hat{m} = \frac{1}{2}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)' S_{\text{pooled}}^{-1}\left(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2\right)$

p.5.b. The confusion matrix for the holdout samples (617 NHL players and 426 EPL players) is given below, based on the function generated for the training sample. Compute the estimate of the Expected actual error rate.

```
> (classtab <- table(league,classify))
      classify
league   1   2
     1 517 100
     2  95 331
```

Q.6. There are 2 populations of individuals: π_1, π_2. Two variables are measured on each individual, both of which range between 0 and 1. The prior probabilities are $p_1 = 0.25$, $p_2 = 0.75$ and the cost of misclassification is twice as high for individuals from population 1 than for individuals from population 2.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad f_1(\mathbf{x}) = f_1(x_1, x_2) = x_2^{\frac{1-x_1}{x_1}} \quad 0 < x_1, x_2 < 1 \qquad f_2(\mathbf{x}) = f_2(x_1, x_2) = 2x_2^{\frac{2-x_1}{x_1}} \quad 0 < x_1, x_2 < 1$$

Which population would the following points **x** be allocated to: (.10,.10), (.10,.90), (.90,.10), (.9,.9), (.5,.5)

Q.7. The market capitalizations (in $100B), gross profits (in $100B), and the revenues (in $100B) for Facebook, Apple, Amazon, Netflix, and Google (aka Alphabet) as of 8:00AM, 4/29/2019 are given in the following table.

Company	MktCap	Profits	Revenues
Facebook	5.47	0.47	0.59
Apple	9.63	1.02	2.62
Amazon	9.60	0.94	2.42
Netflix	1.64	0.06	0.17
Google (Alphabet)	8.86	0.77	1.37

p.7.a. Compute the matrix of distances among the 5 firms.

p.7.b. Cluster the 5 firms by single linkage, complete linkage, and average linkage. Draw a dendogram based on average linkage.

Q.8. The following table gives the Height (inches), Number of Instagram followers (millions), net worth ($1M), and age (years) of the 5 Kardashian/Jenner sisters.

Sister	Height	InstaFollow	NetWorth	Age
Kim	62	127	350	40
Kourtney	60	73.5	35	38
Khloe	70	86.9	40	34
Kendall	70	104	30	23
Kylie	66	127	1000	21

p.8.a. Obtain the correlation matrix for the 4 variables.

p.8.b. Obtain a cluster analysis of the 4 variables by single linkage, complete linkage, and average linkage. Draw a dendogram based on average linkage.