STA2005: Applied Multivariate Stat Assignment #3

Due on March 30, 2016

K. Knight, W1-2pm, F1-3pm

Yi Li

Problem 1

(a) Since we know that the Ψ is a diagonal matrix with the diagonal elements of $\Psi + LL^T$ equal to S. The diagonal elements of LL^T are $LL^T_{ii} = \sum_{j=1}^{r} l_{ij}$. Thus the i-th diagonal element of Ψ is $\Psi_{ii} = S_{ii} - \sum_{j=1}^{r} l_{ij} = \sum_{j=r+1}^{p} \lambda_j V_{ij}^2$

(b) Let's assume $R = LL^T - S$ and $\{R_{ij}\}$ are the elements of R, so D is equal to R with all diagonal elements in D are 0. Thus we will have $\sum_{i=1}^{p}\sum_{j=1}^{p} d_{ij}^2 \leq \sum_{i=1}^{p}\sum_{j=1}^{p} R_{ij}^2$. We also know $R = LL^T - S = \sum_{k=r+1}^{p} \lambda_k V_k V_k^T$, so $R_{ij} = \sum_{k=r+1}^{p} \lambda_k V_{ik} V_{jk}$, then $R_{ij}^2 = (\sum_{k=r+1}^{p} \lambda_k V_{ik} V_{jk})^2 = \lambda_{(r+1)}^2 V_{i(r+1)}^2 V_{j(r+1)}^2 + \lambda_{(r+2)}^2 V_{i(r+2)}^2 V_{j(r+2)}^2 + ... + \lambda_{(p)}^2 V_{i(p)}^2 V_{j(p)}^2 + 2\lambda_{r+1}\lambda_{r+2} V_{i(r+1)} V_{j(r+1)} V_{i(r+2)} V_{j(r+2)} + 2\lambda_{r+1}\lambda_{r+2} V_{i(r+1)} V_{j(r+1)} V_{i(r+2)} V_{j(r+2)} + .. + 2\lambda_{p-1}\lambda_p V_{i(p-1)} V_{j(p-1)} V_{i(p)} V_{j(p)}$ (some square terms and some non-square terms).
So far, R_{ij}^2 are split by square terms, like $\lambda_{(r+1)}^2 V_{i(r+1)}^2 V_{j(r+1)}^2$ and non-square terms, like $2\lambda_{r+1}\lambda_{r+2} V_{i(r+1)} V_{j(r+1)} V_{i(r+2)} V_{j(r+2)}$ for the non-square terms, we can sum up the whole column, which will be zero. Here take one of those terms as an example, $\sum_{i=1}^{p} 2\lambda_{r+1}\lambda_{r+2} V_{i(r+1)} V_{j(r+1)} V_{i(r+2)} V_{j(r+2)} = 2\lambda_{r+1}\lambda_{r+2} V_{j(r+1)} V_{j(r+2)} (\sum_{i=1}^{p} V_{i(r+1)} V_{i(r+2)}) = 0$
(here we use the truth that $V_{(r+1)}$ and $V_{(r+2)}$ are orthogonal so that $\sum_{i=1}^{p} V_{i(r+1)} V_{i(r+2)} = 0$).
For those square terms, we can group them based on $(r+1), (r+2), ..., (p)$ and sum up them along both row and column. Here we use those terms have $(r+1)$ (where coms from $\lambda_{(r+1)} V_{(r+1)} V_{(r+1)}^T$) as an example. $\sum_{i=1}^{p}\sum_{j=1}^{p} \lambda_{(r+1)}^2 V_{i(r+1)}^2 V_{j(r+1)}^2 = \lambda_{(r+1)}^2 \sum_{i=1}^{p} V_{i(r+1)}^2 (\sum_{j=1}^{p} V_{j(r+1)}^2) = \lambda_{(r+1)}^2$ (here we use the truth that $V_{(r+1)}$ is a unit vector so that $\sum_{i=1}^{p} V_{i(r+1)}^2 = \sum_{j=1}^{p} V_{j(r+1)}^2 = 1$). Thus $\sum_{j=1}^{p} R_{ij}^2 = \lambda_{(r+1)}^2 + \lambda_{(r+2)}^2 + ... + \lambda_{(p)}^2$
Through above derivation, we prove that $\sum_{i=1}^{p}\sum_{j=1}^{p} d_{ij}^2 \leq \sum_{i=1}^{p}\sum_{j=1}^{p} R_{ij}^2 = \lambda_{(r+1)}^2 + \lambda_{(r+2)}^2 + ... + \lambda_{(p)}^2$

(c) Since $D = \Psi + LL^T - S$, $\sum_{i=1}^{p}\sum_{j=1}^{p} d_{ij}^2$ is the total variance or the L2 norm residues of the model fitting. Thus $\sum_{i=1}^{p}\sum_{j=1}^{p} d_{ij}^2$ should follow $\chi^2(\nu)$, where $\nu = 1/2(p-r)^2 - 1/2(p+r)$. So we choose the number of factors based on the following steps. First step: we starts from r=1, we use $\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} R_{ij}^2 = \lambda_{(r+1)}^2 + \lambda_{(r+2)}^2 + ... + \lambda_{(p)}^2 = \lambda_{(2)}^2 + \lambda_{(3)}^2 + ... + \lambda_{(p)}^2$ to estimate the total variance $\sum_{i=1}^{p}\sum_{j=1}^{p} d_{ij}^2$. Second step: compute $\nu = 1/2(p-r)^2 - 1/2(p+r)$ and the probability $P(\chi(\nu)^2 > \chi^2)$. If this probability is too small, we need to increase r and repeat these two steps until we find appropriate r as well as a larger p-value.

Problem 2

(a) The result of the single factor analysis is show in Figure. 1. The p-value of model fitting is 0.124, which is not too small. So we can say this model fits the data adequately in some degrees.

```
Call:
factanal(x = examdata, factors = 1)

Uniquenesses:
  mec   vec   alg   ana   sta
0.641 0.555 0.158 0.403 0.476

Loadings:
    Factor1
mec 0.599
vec 0.667
alg 0.917
ana 0.772
sta 0.724

                Factor1
SS loadings       2.766
Proportion Var    0.553

Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 8.65 on 5 degrees of freedom.
The p-value is 0.124
```

Figure 1: Factor analysis using a single factor model.

(b) Based on Figure.1, we find the variable ALG has the heaviest loading, so we can conclude that it seems to be most important. Recall the graphic model shown in Figure. 2 based on the concentration matrix, the variable ALG is at center and connected with other variables, which indicates that all the other variables are dependent on the variable ALG. Thus this variable ALG is the most important one.

(c) The result of the factor analysis using a two factor model is show in Figure. 3. The p-value of model fitting is 0.785, which is significantly larger than the single factor model, so the model seems to be an improvement over the single model. Here we use the 'varimax' rotation as default. The loadings for the two factors here are essentially mirror images of each other, the first giving higher loadings to ALG, ANA and STA with the second giving higher loadings to MEC, VEC and ALG.
To better interpret the factor loadings, we can try different rotations, Figure. 4 and 5 shows the two factor model analysis with none and 'promax' rotations respectively. For the none rotation, the loadings are similar to the loadings for the first two principal components. But for the 'promax' rotation, we find the first factor loading is 0 for MEC and second factor loadings are 0 for ANA and STA. Then we can interpret the factor loadings as the dependence structure shown as Figure. 6 (Note that because the 'promax' rotation is not an orthogonal rotation, the two factors are correlated).

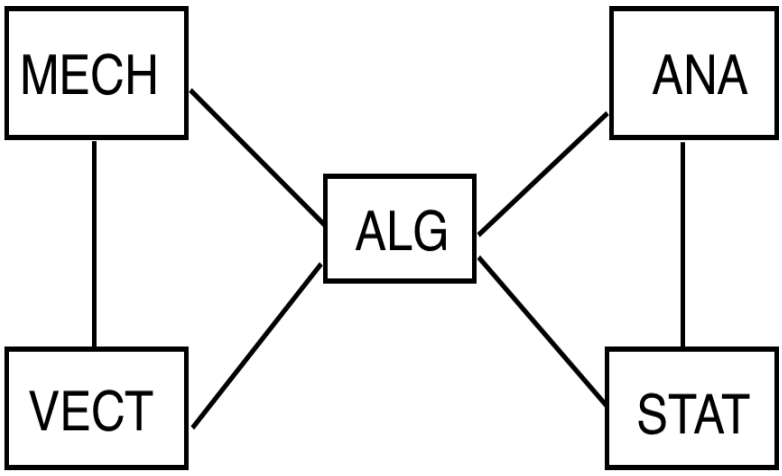

Figure 2: Graphic model from concentration matix.

```
Call:
factanal(x = examdata, factors = 2)

Uniquenesses:
  mec   vec   alg   ana   sta
0.466 0.419 0.189 0.352 0.431

Loadings:
    Factor1 Factor2
mec 0.265   0.681
vec 0.356   0.674
alg 0.740   0.514
ana 0.738   0.322
sta 0.696   0.290

               Factor1 Factor2
SS loadings      1.774   1.370
Proportion Var   0.355   0.274
Cumulative Var   0.355   0.629

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 0.07 on 1 degree of freedom.
The p-value is 0.785
```

Figure 3: Factor analysis using a two factor model with default rotation.

Listing 1: R code for Problem 2

```
#set working directory
setwd("/Users/Harryliyi/Documents/Course/STA437/HW3")
getwd()
```

```
Call:
factanal(x = examdata, factors = 2, rotation = "none")

Uniquenesses:
  mec   vec   alg   ana   sta
0.466 0.419 0.189 0.352 0.431

Loadings:
    Factor1 Factor2
mec  0.628   0.373
vec  0.695   0.312
alg  0.899
ana  0.780  -0.201
sta  0.727  -0.200

                Factor1 Factor2
SS loadings       2.824   0.319
Proportion Var    0.565   0.064
Cumulative Var    0.565   0.629

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 0.07 on 1 degree of freedom.
The p-value is 0.785
```

Figure 4: Factor analysis using a two factor model with none rotation.

```
Call:
factanal(x = examdata, factors = 2, rotation = "promax")

Uniquenesses:
  mec   vec   alg   ana   sta
0.466 0.419 0.189 0.352 0.431

Loadings:
    Factor1 Factor2
mec           0.736
vec  0.113    0.680
alg  0.690    0.272
ana  0.788
sta  0.749

                Factor1 Factor2
SS loadings       1.671   1.078
Proportion Var    0.334   0.216
Cumulative Var    0.334   0.550

Factor Correlations:
         Factor1 Factor2
Factor1   1.000  -0.694
Factor2  -0.694   1.000

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 0.07 on 1 degree of freedom.
The p-value is 0.785
```

Figure 5: Factor analysis using a two factor model with 'promax' rotation.

```
exam <- scan("marks.txt",what=list(0,0,0,0,0))
mec <- exam[[1]]
vec <- exam[[2]]
alg <- exam[[3]]
```

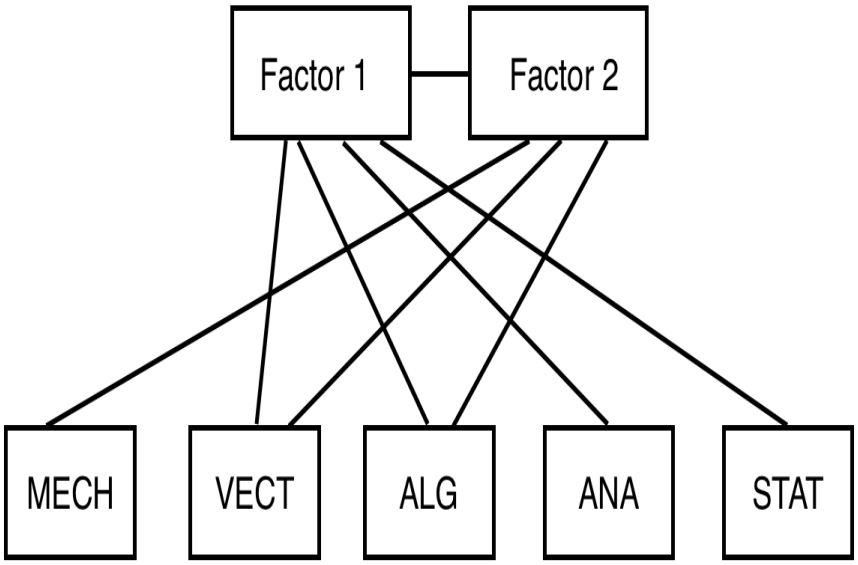

Figure 6: Dependence structure using two factor model with 'promax' rotation.

```
ana <- exam[[4]]
sta <- exam[[5]]

examdata <- cbind(mec,vec,alg,ana,sta)
fa1 <-  factanal(examdata,factors=1)
print(fa1)

fa2 <-  factanal(examdata,factors=2)
print(fa2)

fa3 <-  factanal(examdata,factors=2,rotation="promax")
print(fa3)

fa4 <-  factanal(examdata,factors=2,rotation="none")
print(fa4)
```

Problem 3

(a) We can firstly do a linear discriminant analysis with leave-one-out cross validation for each observation. Then we can compare the analysis result with the real group for each observation to estimate the misclassification rate as 0.05 (See the code in List. 2).

(b) Given the following measurements for the 5 variables: FL = 18.7, RW = 15.0, CL = 35.0, CW = 40.3, BD = 16.6, the prediction based on the linear discriminant analysis is class 4, so sex is female and color is orange (Also see the code in List. 2).

(c) We can repeat above analysis based on quadratic discriminant analysis. It will give us the same prediction for the crab as female sex and orange color, But the misclassification rate of qda is 0.065. So we can conclude that LDA seems to be better than QDA in this particular case since the error rate of LDA (0.05) is slightly small than QDA (0.065) based on leave-one-out cross validation. However QDA probably does better in more complex data since it allows different covariance matrices among different class.

Listing 2: R code for Problem 3

```
#set working directory
setwd("/Users/Harryliyi/Documents/Course/STA437/HW3")
getwd()

x <- scan("crabs.txt",skip=1,what=list("c","c",0,0,0,0,0,0))
colour <- x[[1]]
sex <- x[[2]]
FL <- x[[4]]
RW <- x[[5]]
CL <- x[[6]]
CW <- x[[7]]
BD <- x[[8]]

group <- rep(0,200)
group[sex=="M"&colour=="B"] <- 1
group[sex=="F"&colour=="B"] <- 2
group[sex=="M"&colour=="O"] <- 3
group[sex=="F"&colour=="O"] <- 4
group <- factor(group)

library(MASS) # library with lda
r1 <- lda(group~FL+RW+CL+CW+BD, CV=T)

ldaer <- sum(r1$class!=group)/200
print(ldaer)

r2 <- lda(group~FL+RW+CL+CW+BD)
print(r2)
newdata <- data.frame(FL=18.7,RW=15.0,CL=35.0,CW=40.3,BD=16.6)
prel <- predict(r2,newdata)
print(prel$class)

r3 <- qda(group~FL+RW+CL+CW+BD, CV=T)
qdaer <- sum(r3$class!=group)/200
print(qdaer)

r4 <- qda(group~FL+RW+CL+CW+BD)
print(r4)
preq <- predict(r4,newdata)
print(preq$class)
```

Problem 4

(a)