STA437 Assignment #3

Rui Qiu #999292509

2016-03-26

Problem 1

(a) Solution:

$$S = V\Lambda V^T = \begin{pmatrix} v_1 & \cdots & v_p \end{pmatrix} \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{pmatrix} \begin{pmatrix} v_1^T \\ \vdots \\ v_p^T \end{pmatrix}$$

$$= \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T + \cdots + \lambda_p v_p v_p^T$$

$$LL^T = \begin{pmatrix} \lambda_1^{\frac{1}{2}} v_1 & \cdots & \lambda_r^{\frac{1}{2}} v_r \end{pmatrix} \begin{pmatrix} \lambda_1^{\frac{1}{2}} v_1^T \\ \vdots \\ \lambda_r^{\frac{1}{2}} v_r^T \end{pmatrix}$$

$$= \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T + \cdots + \lambda_r v_r v_r^T$$

Note that $r < p$.
We claim that $\text{diag}(S) = \text{diag}(\Psi + LL^T)$, so

$$\Psi = \text{diag}(\Psi) = \text{diag}(S - LL^T) = \text{diag}(\lambda_{r+1} v_{r+1} v_{r+1}^T + \cdots + \lambda_p v_p v_p^T)$$

Observe that the product of column vector v_{r+1} and its transpose v_{r+1}^T generates a p-by-p matrix whose i-th diagonal line element is $v_{i(r+1)}^2$ and $v_{i(r+1)}$ is the i-th element in vector v_{r+1}

Therefore the i-th element in diagonal matrix Ψ, where i is any integer from 1 to p:

$$\psi_{ii} = \sum_{j=r+1}^{p} \lambda_j v_{ij}^2.$$

(b) Proof:
Since the diagonal elements of $D = \Psi + LL^T - S$ are 0 so we need only consider the off-diagonal elements of $LL^T - S$, which is equal to $-\lambda_{r+1} v_{r+1} v_{r+1}^T - \cdots - \lambda_p v_p v_p^T$.

We can see $LL^T - S$ as the sum of a bunch of matrices. And for each matrix, we take $-\lambda_{r+1}\boldsymbol{v_{r+1}}\boldsymbol{v_{r+1}^T}$ as example, we are interested in the sum of squares of its off-diagonal part, which can be expressed as:

$$\lambda_{r+1}^2 \cdot \sum_{i,j<r+1,i\neq j} v_{i(r+1)}^2 \, v_{j(r+1)}^2$$

And if we can show such part is no greater than λ_{r+1}^2, i.e., $\sum_{i,j<r+1,i\neq j} v_{i(r+1)}^2 \, v_{j(r+1)}^2 \leq 1$, then the desired inequation to prove is nothing but adding a bunch of similar inequations up.

Recall that V is an orthogonal matrix, so that any of its column $\boldsymbol{v_k}$ is an orthogonal unit vector. Therefore, $\sqrt{\sum_{i=1}^p v_{ik}^2} = 1$. And this is a sum of p non-negative terms which are all not greater than 1. Then for each term, we multiple it by another non-negative term no greater than 1. Consequently, the sum of these new p terms are no greater than 1 since each term becomes smaller (or stays equivalent).

So $\sum_{i=1,j=1}^p v_{i(r+1)}^2 \, v_{j(r+1)}^2 \leq 1$. And our previous version has already extracted the diagonal elements, so the sum should be even smaller with:

$$\sum_{i,j<r+1,i\neq j}^p v_{i(r+1)}^2 \, v_{j(r+1)}^2 \leq \sum_{i=1,j=1}^p v_{i(r+1)}^2 \, v_{j(r+1)}^2 \leq 1.$$

Both sides multiple by λ_{r+1}^2, then sum up from $r + 1$ to p, and finally we are done, have the following result:

$$\sum_{i=1}^p \sum_{j=1}^p d_{ij}^2 \leq \lambda_{r+1}^2 + \cdots + \lambda_p^2.$$

(c) Solution:
The LHS of result in part (b) is the total variations.

- We start from $t = p$, calculate $\sum_{i=1}^k \sum_{j=1}^k d_{ij}^2$.
- Calculate the ratio of variations that these t factors can explain, if this proportion is large enough, e.g. we set the critical level at 90% or 80%.
- Iterate on t, each time $t := t - 1$.
- If the proportion drops below the critical level, keep the previous t as that the number of factors.

Assignment #3-2

Problem 2

(a) Solution:

```
 1 > exam <- scan("marks.txt",what=list(0,0,0,0,0))
 2 Read 88 records
 3 > mec <- exam[[1]]
 4 > vec <- exam[[2]]
 5 > alg <- exam[[3]]
 6 > ana <- exam[[4]]
 7 > sta <- exam[[5]]
 8 > data <- cbind(mec,vec,alg,ana,sta)
 9 > factanal(data,factors=1)
10
11 Call:
12 factanal(x = data, factors = 1)
13
14 Uniquenesses:
15   mec   vec   alg   ana   sta
16 0.641 0.555 0.158 0.403 0.476
17
18 Loadings:
19     Factor1
20 mec 0.599
21 vec 0.667
22 alg 0.917
23 ana 0.772
24 sta 0.724
25
26               Factor1
27 SS loadings     2.766
28 Proportion Var  0.553
29
30 Test of the hypothesis that 1 factor is sufficient.
31 The chi square statistic is 8.65 on 5 degrees of freedom.
32 The p-value is 0.124
```
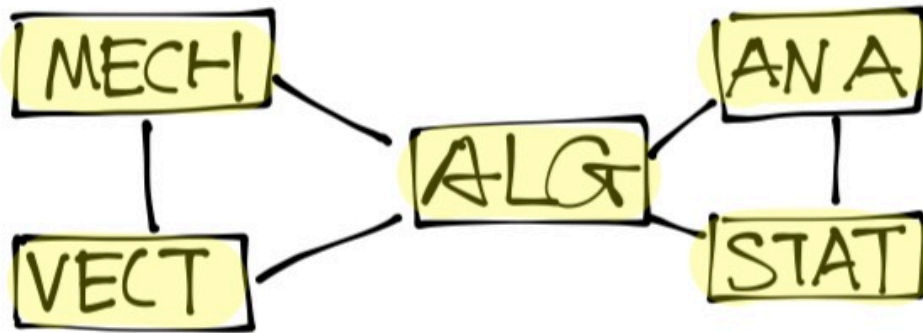
- The p-value of this model fitting the data is $0.124 > 0.05$, which is not very low, so we fail to reject. Hence **the single factor model fits the data somewhat adequately**.

(b) Solution:

- The **variable `alg` has the heaviest loading**, so it is the most important one.

Graphical Model (from simple analysis)

- As the graphical dependence model indicates derived directly from concentration matrix, the `alg` variable is relevant with all the other variables. In other words, variables `mec,vec,ana,sta` all depends on `alg`. Thus we conclude that **`alg` is the most important one.**

(c) Solution:

```
 1 > factanal(data,factors=2)
 2
 3 Call:
 4 factanal(x = data, factors = 2)
 5
 6 Uniquenesses:
 7   mec   vec   alg   ana   sta
 8 0.466 0.419 0.189 0.352 0.431
 9
10 Loadings:
11     Factor1 Factor2
12 mec 0.265   0.681
13 vec 0.356   0.674
14 alg 0.740   0.514
15 ana 0.738   0.322
16 sta 0.696   0.290
17
18               Factor1 Factor2
19 SS loadings     1.774   1.370
20 Proportion Var  0.355   0.274
21 Cumulative Var  0.355   0.629
22
23 Test of the hypothesis that 2 factors are sufficient.
24 The chi square statistic is 0.07 on 1 degree of freedom.
25 The p-value is 0.785
```

- The p-value of two factors model is higher than that of single factor model $(0.785 > 0.124)$, which is **an improvement.**
- The loadings of the default two factors model are not understandably easy to interpret. And in fact, the

`rotation` it used was `varimax`.

- So we do the same process but with different rotations `none` and `promax`.

```
 1 > factanal(data,factors=2,rotation="none")
 2
 3 Call:
 4 factanal(x = data, factors = 2, rotation = "none")
 5
 6 Uniquenesses:
 7    mec   vec   alg   ana   sta
 8 0.466 0.419 0.189 0.352 0.431
 9
10 Loadings:
11     Factor1 Factor2
12 mec  0.628   0.373
13 vec  0.695   0.312
14 alg  0.899
15 ana  0.780  -0.201
16 sta  0.727  -0.200
17
18               Factor1 Factor2
19 SS loadings     2.824   0.319
20 Proportion Var  0.565   0.064
21 Cumulative Var  0.565   0.629
22
23 Test of the hypothesis that 2 factors are sufficient.
24 The chi square statistic is 0.07 on 1 degree of freedom.
25 The p-value is 0.785
26 > factanal(data,factors=2,rotation="promax")
27
28 Call:
29 factanal(x = data, factors = 2, rotation = "promax")
30
31 Uniquenesses:
32    mec   vec   alg   ana   sta
33 0.466 0.419 0.189 0.352 0.431
34
35 Loadings:
36     Factor1 Factor2
37 mec          0.736
38 vec  0.113   0.680
39 alg  0.690   0.272
40 ana  0.788
41 sta  0.749
42
43               Factor1 Factor2
44 SS loadings     1.671   1.078
45 Proportion Var  0.334   0.216
46 Cumulative Var  0.334   0.550
47
48 Factor Correlations:
49          Factor1 Factor2
50 Factor1    1.000  -0.694
51 Factor2   -0.694   1.000
52
53 Test of the hypothesis that 2 factors are sufficient.
```

```
54 The chi square statistic is 0.07 on 1 degree of freedom.
55 The p-value is 0.785
```

- As the result shows, rotation `promax` seems to generate a possibly interpretable set of loadings with:
    - **loading 1: the marks of subjects Algebra, Analysis and Statistics.**
    - **loading 2: the marks of subjects Mechanics and Vectors.**
- `none` rotation has two sets of loadings, but it's not practically understandable. (Maybe loading 1 indicates the general level of all 5 subjects, loading 2 indicates the sum of Mechanics and Vectors minus the sum of Analysis and Statistics. And this does not make much sense in real life.)

Assignment #3-3

Problem 3

(a) Solution:

```
 1 > x <- scan("crabs.txt",skip=1,what=list("c","c",0,0,0,0,0,0))
 2 Read 200 records
 3 > colour <- x[[1]]
 4 > sex <- x[[2]]
 5 > FL <- x[[4]]
 6 > RW <- x[[5]]
 7 > CL <- x[[6]]
 8 > CW <- x[[7]]
 9 > BD <- x[[8]]
10 > group <- rep(0,200)
11 > group[sex=="M"&colour=="B"] <- 1
12 > group[sex=="F"&colour=="B"] <- 2
13 > group[sex=="M"&colour=="O"] <- 3
14 > group[sex=="F"&colour=="O"] <- 4
15 > group <- factor(group)
16 > library(MASS)
17 > r <- lda(group~FL+RW+CL+CW+BD,CV=T)
18 > sum(r$class!=group)/200
19 [1] 0.05
```

- `r$class` is the predicted classes of 200 crabs, `group` contains the real classes of 200 crabs. So **the estimated of misclassification rate is** 0.05.

(b) Solution:

```
 1 > r <- lda(group~FL+RW+CL+CW+BD)
 2 > r
 3 Call:
 4 lda(group ~ FL + RW + CL + CW + BD)
 5
 6 Prior probabilities of groups:
 7    1    2    3    4
 8 0.25 0.25 0.25 0.25
 9
10 Group means:
11       FL     RW     CL     CW     BD
12 1 14.842 11.718 32.014 36.810 13.350
13 2 13.270 12.138 28.102 32.624 11.816
14 3 16.626 12.262 33.688 37.188 15.324
15 4 17.594 14.836 34.618 39.036 15.632
16
17 Coefficients of linear discriminants:
18          LD1        LD2         LD3
```

```
19 FL -1.5543139 -0.1951885 -1.6667377
20 RW -0.6247546 -1.5394972  0.4558782
21 CL -0.1875489  1.0953923  0.6807588
22 CW  1.5156077 -0.6435178 -0.6548549
23 BD -1.3551090  0.5153193  1.2859743
24
25 Proportion of trace:
26    LD1    LD2    LD3
27 0.6861 0.2995 0.0144
28 > newdata <- data.frame(FL=18.7, RW=15.0, CL=35.0, CW=40.3, BD=16.6)
29 > predict(r,newdata)
30 $class
31 [1] 4
32 Levels: 1 2 3 4
33
34 $posterior
35            ⋮    ⋮    ⋮ 1            2            3            4
36 1 2.433387e-15 9.97961e-10 3.277912e-06 0.9999967
37
38 $x
39       ⋮      LD1         LD2         LD3
40 1 -4.393598 -2.095638 -1.433924
```

- This crab is very highly possible to be in **group 4**, so that **its sex is female and the colour is orange** (which agrees the result of Assignment 2).

(c) Solution:

```
1 > r1 <- qda(group~FL+RW+CL+CW+BD,CV=T)
2 > r2 <- qda(group~FL+RW+CL+CW+BD)
3 > predict(r2,newdata)
4 $class
5 [1] 4
6 Levels: 1 2 3 4
7
8 $posterior
9           ⋮    ⋮    1            2            3 4
10 1 1.81922e-32 1.460731e-35 1.179623e-14 1
11 > sum(r1$class!=group)/200
12 [1] 0.065
```

- The prediction agrees with part (b), which is female and orange.
- **The error rate of QDA** (0.065) is slightly higher than that of LDA (0.05) based on the leave-one-out cross validation.
- In this particular case, **LDA seems better with lower error rate under the same predicted result.**
- But we have notice that the LDA is based on an assumption of equal variance-covariance matrices between the input variables of the classes. And QDA should be preferred in more complex scenario where different variance-covariance matrices are introduced in different classes.