과제 보고서

제목 : 9장, 10장, 13장 연습문제

과 목 명: 다변량 통계분석

제출일자: 2022.05.25

학 과: 정보통계학과

학 번: 2018015027

이 름: 김한탁

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

9장 주성분분석 연습문제

1. 다음 표는 11년 기간 동안($n=11$) 5개의 기상변수($p=5$)를 측정한 자료이다.

년	x_1	x_2	x_3	x_4	x_5
1920~1921	87.9	19.6	1.0	1661	28.37
1921~1922	89.9	15.2	90.1	968	23.77
1922~1923	153.0	19.7	56.6	1353	26.04
1923~1924	132.1	17.0	91.0	1293	25.74
1924~1925	88.8	18.3	93.7	1153	26.68
1925~1926	220.9	17.8	106.9	1286	24.29
1926~1927	117.7	17.8	65.5	1104	28.00
1927~1928	109.0	18.3	41.8	1574	28.37
1928~1929	156.1	17.8	57.4	1222	24.96
1929~1930	181.5	16.8	140.6	902	21.66
1930~1931	181.4	17.0	74.3	1150	24.37

x_1= 11월과 12월 강우량(단위: mm)

x_2= 7월 평균기온(단위: C)

x_3= 7월 강우량(단위: mm)

x_4= 7월 방사선량(단위: 알코올의 밀리미터)

x_5= 평균수확량(헥터당 퀸틀)

(a) 평균벡터, 공분산 및 상관행렬을 구하여라.

solve)

다음은 표의 데이터이다.

```
> x1<-c(87.9, 89.9, 153.0, 132.1, 88.8, 220.9, 117.7, 109.0, 156.1, 181.5, 181.4)
> x2<-c(19.6, 15.2, 19.7, 17.0, 18.3, 17.8, 17.8, 18.3, 17.8, 16.8, 17.0)
> x3<-c(1.0, 90.1, 56.6, 91.0, 93.7, 106.9, 65.5, 41.8, 57.4, 140.6, 74.3)
> x4<-c(1661, 968, 1353, 1293, 1153, 1286, 1104, 1574, 1222, 902, 1150)
> x5<-c(28.37, 23.77, 26.04, 25.74, 26.68, 24.29, 28.00, 28.37, 24.96, 21.66, 24.37)
> data<-cbind(x1, x2, x3, x4, x5)
```

a-1) 평균벡터 (R코드 및 결과)

```
> apply(data, 2, mean)
       x1         x2         x3         x4         x5
138.02727  17.75455  74.44545 1242.36364  25.65909
```

a-2) 공분산행렬 (R코드 및 결과)

```
> cov(data)
             x1         x2          x3          x4         x5
x1  1973.298182  -4.920636   799.56364 -2439.3509 -57.213873
x2    -4.920636   1.636727   -29.27873   217.1982   1.734655
x3   799.563636 -29.278727  1346.85873 -6822.7282 -62.080455
x4 -2439.350909 217.198182 -6822.72818 52914.6545 361.803364
x5   -57.213873   1.734655   -62.08045   361.8034   4.495809
```

a-3) 상관행렬 (R코드 및 결과)

```
> cor(data)
           x1          x2          x3          x4          x5
x1  1.00000000 -0.08658382  0.4904510 -0.2387206 -0.6074367
x2 -0.08658382  1.00000000 -0.6235956  0.7380402  0.6394711
x3  0.49045103 -0.62359556  1.0000000 -0.8081818 -0.7977925
x4 -0.23872062  0.73804024 -0.8081818  1.0000000  0.7417895
x5 -0.60743674  0.63947108 -0.7977925  0.7417895  1.0000000
```

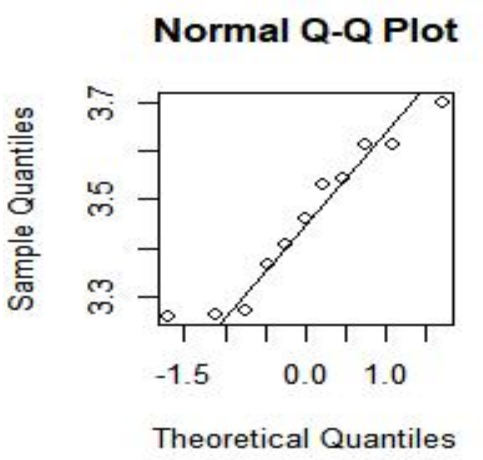
(b) 처음 4개 변수($x_1 \sim x_4$)의 공분산행렬을 이용하여 주성분분석을 실시하여라.

solve)

PCA 분석에 앞서, 변수의 왜도와 크기는 주성분에 영향을 미치므로, 변수에 대하여 왜도 변환, 중심화, 표준화를 수행하는 것이 바람직하다. 왜도 변환을 위해 정규화하는 방법은 로그 변환과 Box-Cox 변환이 존재한다. 문제에서는 두 방법 중 일반적인 Box-Cox변환을 이용하여 문제에 적용해보도록 하겠다. 그 후 promp() 함수를 통해 중심화(center=TRUE)와 표준화(scale=TRUE)를 수행할 수 있을 것이다.
다음은 변수 x_1에 대한 Box-Cox변환 과정이다.

변수x_1 Box-Cox변환 (R코드 및 결과)

```
> library(forecast)
> lambda1<-BoxCox.lambda(x1, method='loglik') #최적 lambda 선정
> lambda1
[1] -0.15
> x_1<-BoxCox(x1, lambda1)
> qqnorm(x_1)
> qqline(x_1)
```



위의 플롯은 변환된 변수 x_1의 정규확률그림으로, Box-Cox변환을 통해 정규분포에 가까운 형태로 바뀌었음을 알 수 있다. 마찬가지로 변수 $x_2 \sim x_4$에 대해서 Box-Cox변환을 수행한다.
다음은 모든 변수에 대하여 Box-Cox변환을 마친 뒤, PCA 분석을 수행하는 과정이다.

4개 변수($x_1 \sim x_4$)의 공분산행렬을 이용한 PCA 분석 (R코드 및 결과)

```
> data1<-cbind(x_1, x_2, x_3, x_4)
> pca1<-prcomp(data1, center=TRUE, scale=FALSE)
> print(pca1)
Standard deviations (1, .., p=4):
[1] 30.29910107 14.03202062  0.13007166  0.03199575

Rotation (n x k) = (4 x 4):
              PC1             PC2           PC3          PC4
x_1  0.001720989 -4.709608e-03 -0.998370209  0.056848782
x_2 -0.658104879 -7.529216e-01  0.002473118  0.000980029
x_3  0.752922184 -6.580934e-01  0.004308792 -0.001642290
x_4 -0.001786534  7.528136e-05 -0.056852830 -0.998380969
```

위의 prcomp() 함수를 이용하여 PCA 분석을 수행할 때, 중심화(center=TRUE)만 수행하였다. 이는 공분산 행렬을 이용하여 분석을 수행하는 조건으로 인해 표준화 과정이 생략된 것으로 설명될 수 있다.
print() 함수는 4개의 주성분 각각의 표준편차와 연속형 변수의 선형결합 계수를 나타내는 회전(또는 부하)을 제공하고 있다. 또한 처음 2개의 주성분의 표준편차가 나머지에 비해 매우 큰 것으로 보아, 처음 2개의 주성분이 자료 변동의 대부분을 설명할 것으로 예측된다. 각 주성분에 의해 설명되는 자료 분산의 자세한 비율은 summary() 함수를 통해 확인할 수 있다.

```
> summary(pca1)
Importance of components:
                            PC1      PC2      PC3   PC4
Standard deviation      30.2991 14.0320 0.13007 0.032
Proportion of Variance   0.8234  0.1766 0.00002 0.000
Cumulative Proportion    0.8234  1.0000 1.00000 1.000
```

위에서 print() 함수의 표준편차를 통해 예측하였듯이, 처음 2개의 주성분의 자료분산 비율이 0.99보다 크므로, 자료변동의 99% 이상을 설명함을 알 수 있다.
따라서 PCA 분석의 목적에 따라, 주성분 PC1과 PC2을 선택해 차원을 축소하는 것이 적절하다고 생각된다.

(c) 처음 4개 변수($x_1 \sim x_4$)의 상관행렬을 이용하여 주성분분석을 실시하여라.

solve)
4개 변수($x_1 \sim x_4$)의 상관행렬을 이용하여 PCA 분석을 수행하는 것 또한 (b)의 풀이와 비슷하게 진행된다.
4개의 변수를 Box-Cox변환을 통해, 정규화하여 왜도 변환을 수행한 뒤, promp() 함수를 통해 중심화와 표준화를 수행할 수 있을 것이다. 다만 공분산행렬이 아닌 상관행렬을 이용하기 때문에 (b)에서와 달리 중심화와 표준화가 동시에 진행되어야 할 것이다.
다음은 Box-Cox변환을 마친, 변수$x_1 \sim x_4$에 대하여 PCA 분석을 수행하는 과정이다.

4개 변수($x_1 \sim x_4$)의 상관행렬을 이용한 PCA 분석 (R코드 및 결과)

```
> data1<-cbind(x_1, x_2, x_3, x_4)
> pca2<-prcomp(data1, center=TRUE, scale=TRUE)
```

```
> print(pca2)
Standard deviations (1, .., p=4):
[1] 1.6094956 0.9896764 0.5315004 0.3841508

Rotation (n x k) = (4 x 4):
            PC1         PC2         PC3         PC4
x_1   0.2758358 -0.8875551 -0.2885809   0.2299600
x_2 -0.5176392 -0.3799705   0.7590974   0.1069734
x_3   0.5769123 -0.1356941   0.4221497 -0.6859657
x_4 -0.5684510 -0.2223858 -0.4028423 -0.6820016
```

print() 함수를 통해, 주성분 각각의 표준편차에 주목하여보자. 처음 2개의 주성분의 표준편차가 나머지에 비해 큰 편이므로, 자료변동의 많은 부분을 설명함을 예측할 수 있다. summary() 함수를 통해 각 주성분의 자료 분산 비율에 대해 자세히 알아보겠다.

```
> summary(pca2)
Importance of components:
                          PC1     PC2      PC3      PC4
Standard deviation     1.6095 0.9897 0.53150 0.38415
Proportion of Variance 0.6476 0.2449 0.07062 0.03689
Cumulative Proportion  0.6476 0.8925 0.96311 1.00000
```

위에서 print() 함수의 표준편차를 통해 예측하였듯이, 처음 2개의 주성분의 자료분산 비율이 0.89 정도인 것으로 보아, 자료변동의 대부분(89%)을 설명한다는 것을 알 수 있다. 하지만 PC3 또한 주성분으로 사용하게 된다면 자료변동을 96%까지 설명할 수 있게 되고, 이는 PC3를 포함할 때 더 많은 정보를 보존하는데 유의한 차이를 만들 것으로 생각된다.
따라서 주성분 PC1과 PC2 그리고 PC3을 선택해 차원을 축소하는 것이 적절하다고 생각된다.

(d) (b)와 (c)에서 구한 주성분을 해석하고, 어느 것이 더 큰 의미를 가지는가?

solve)
(b)에서 PCA 분석으로 4개의 주성분을 구할 수 있었다. 그 중 PC1과 PC2가 자료변동의 99%이상을 설명하기 때문에, 차원 축소를 위해 주성분을 PC1, PC2 2가지만 선택하는 것이 적절하다고 생각된다. 마찬가지로 (c)에서도 PCA 분석으로 4개의 주성분을 확인할 수 있다. 그 중 PC1과 PC2 그리고 PC3가 자료변동의 96%를 설명하므로, 차원 축소를 위한 주성분 선택은 PC1, PC2, PC3로 이루어질 수 있다.
PCA 분석은 목적은 데이터 셋의 차원을 축소하되 가능한 많은 정보(변동, variation)를 보존하는 것에 있다. 이러한 점을 고려하였을 때, 공분산행렬을 이용한 PCA 분석 결과(b)가 상관행렬을 이용한 PCA 분석 결과(c)보다 차원을 많이 줄이되, 더 많은 정보를 보존하여, PCA 분석의 목적에 더 적합하고, 더 큰 의미가 있을 것으로 보인다,
하지만 변수들의 척도(scale)가 다르다는 점(x_1, x_3: mm, x_2: C, x_4: 알코올의 밀리미터)을 고려하였을 때, 척도의 단위가 큰 특정 변수가 전체적인 경향을 좌우한다는 문제점을 가지고 있다. 이와 같은 문제점은 공분산행렬을 이용한 PCA 분석(b)에서 나타날 수 있으며, 4개의 주성분 중 2개 주성분 PC1, PC2만을 통해 자료변동 전부(99% 이상)를 설명한다는 점에서 그 문제점이 드러났을 것으로 생각된다.
따라서 위의 문제와 같이 변수들의 척도가 다른 경우, 변수 간의 관계에 중요도를 두어 모든 변수를 표준화하여 상대적인 변화에 초점을 둔 상관행렬을 이용한 PCA 분석(c)이 더 큰 의미를 가질 것으로 생각된다.

10장 인자분석 연습문제

1. 다음은 5개 교과목에 대한 성적 자료로부터 구해진 공분산행렬이다. 5개 교과목은 차례대로 Mechanics, Vector, Algebra, Analysis, Statistics이다. (9장 연습문제에서 다루어짐)

$$S = \begin{pmatrix} 302.3 & 125.8 & 100.4 & 105.1 & 116.1 \\ & 170.9 & 84.2 & 93.6 & 97.9 \\ & & 111.6 & 110.8 & 120.5 \\ & & & 217.9 & 153.8 \\ & & & & 294.4 \end{pmatrix}$$

위 변수 중 Mechanics(O)와 Vector(O) 과목은 open book 시험(O)의 결과이고, Algebra(C), Analysis(C), Statistics(C) 과목은 closed book 시험(C)의 결과라고 하자. 인자분석을 실시한 뒤 그 결과를 시각화하고 해석하여라.

Solve)
FA의 수행 절차는 다음과 같다.

[단계 1] 데이터의 수집과 탐색: 적절한 변수를 선택한다.
[단계 2] 인자의 수를 결정: PCA를 수행하여 정한다.
[단계 3] 미리 결정된 인자의 수를 이용하여 모형을 추정한다.
[단계 4] 회전(rotate)과 해석을 수행한다.
[단계 5] (a) 변경이 필요한지를 결정한다(예를 들어, 항목(들)을 제거하거나 포함.
(b) [단계 3]-[단계 4]를 반복한다.
[단계 6] 데이터에 대해 각 인자의 실현값을 구하고 추가분석에 이용한다.

위와 같은 FA의 수행 절차 중, 문제에서 데이터가 주어져 있고, open book 시험과 close book 시험으로 잠재인자가 2개일 것이라 가정하여, 단계 4부터 단계 6까지의 과정에 주목해 분석을 진행하도록 할 것이다.
다음은 FA를 수행하기 위해, 공분산행렬을 상관행렬로 변환하는 과정이다.
(상관행렬을 사용하는 이유는 동일한 리커트 척도에 의한 자료가 아닌 이상, 대부분(90% 이상)이 공분산행렬보다 상관행렬을 이용하기 때문이다.)

```
> S<-matrix(c(302.3, 125.8, 100.4, 105.1, 116.1,
+             125.8, 170.9, 84.2, 93.6, 97.9,
+             100.4,84.2, 111.6, 110.8, 120.5,
+             105.1, 93.6, 110.8, 217.9, 153.8,
+             116.1, 97.9, 120.5, 153.8, 294.4), 5, 5)
> S
      [,1]  [,2]  [,3]  [,4]  [,5]
[1,] 302.3 125.8 100.4 105.1 116.1
[2,] 125.8 170.9  84.2  93.6  97.9
[3,] 100.4  84.2 111.6 110.8 120.5
[4,] 105.1  93.6 110.8 217.9 153.8
[5,] 116.1  97.9 120.5 153.8 294.4
> r<-matrix(0, 5, 5)
> for(i in 1:5){
```

```
+    for(j in 1:5){
+      r[i,j]=S[i,j]/(sqrt(S[i, i])*sqrt(S[j, j]))
+    }
+ }
> r
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 1.0000000 0.5534655 0.5466163 0.4095011 0.3891744
[2,] 0.5534655 1.0000000 0.6096898 0.4850385 0.4364583
[3,] 0.5466163 0.6096898 1.0000000 0.7105245 0.6647924
[4,] 0.4095011 0.4850385 0.7105245 1.0000000 0.6072378
[5,] 0.3891744 0.4364583 0.6647924 0.6072378 1.0000000
```

FA를 수행하는데 fa() 함수를 이용할 수 있다. fa() 함수는 여러 가지 회전과 인자화 방법(common)을 제공한다. 문제에서의 두 인자 open book 시험과목과 close book 시험과목 간에 다소의 상관이 있을 것으로 생각하여 사교회전(rotate = "oblimin")을 사용하였다. 실제 인자 간의 상관 유무는 fa() 함수의 인자간 상관계수를 통해 판단할 수 있을 것이다. 다음은 fa() 함수를 이용한 FA를 수행하는 과정이다.

fa() 함수를 이용한 FA 수행 (R코드 및 결과)

```
> library(psych)
> library(GPArotation)
> FA<-fa(r=r, nfactors=2, rotate="oblimin", fm="pa")
> FA
Factor Analysis using method =  pa
Call: fa(r = r, nfactors = 2, rotate = "oblimin", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
     PA1   PA2   h2   u2 com
1  -0.02  0.74 0.52 0.48 1.0
2   0.07  0.71 0.59 0.41 1.0
3   0.72  0.22 0.81 0.19 1.2
4   0.84 -0.05 0.65 0.35 1.0
5   0.80 -0.06 0.57 0.43 1.0
```

위의 결과를 통해 인지부하를 살펴보면, 1, 2(Mechanics, Vector)는 모두 제2인자(PA2)에 0.7 근방의 매우 높은 인자부하를 갖고, 3, 4, 5(Algebra, Analysis, Statistics)는 제1인자(PA1)에 높은 인자부하를 가짐을 확인할 수 있다. 여기서 1, 2(Mechanics, Vector)는 모두 open book 시험의 결과이고, 3, 4, 5(Algebra, Analysis, Statistics)는 close book 시험의 결과라는 것이 문제에서 언급되어있다.

이를 통해 제1인자(PA1)는 close book 시험으로 이루어진 과목(Algebra, Analysis, Statistics)을 잘 나타내는 성격을 인자로 생각될 수 있고, 제2인자(PA2)는 open book 시험으로 이루어진 과목(Mechanics, Vector)을 잘 나타내는 성격의 인자로 생각될 수 있을 것이다.

다만 주목할 점은 3(Algebra)의 경우 4, 5(Analysis, Statistics)에 비해 PA1에 상대적으로 낮은 부하를 가지면서 PA2에도 약간의 부하를 가진다는 것이다.

```
                       PA1  PA2
SS loadings           1.95 1.19
Proportion Var        0.39 0.24
Cumulative Var        0.39 0.63
Proportion Explained  0.62 0.38
```

Cumulative Proportion 0.62 1.00

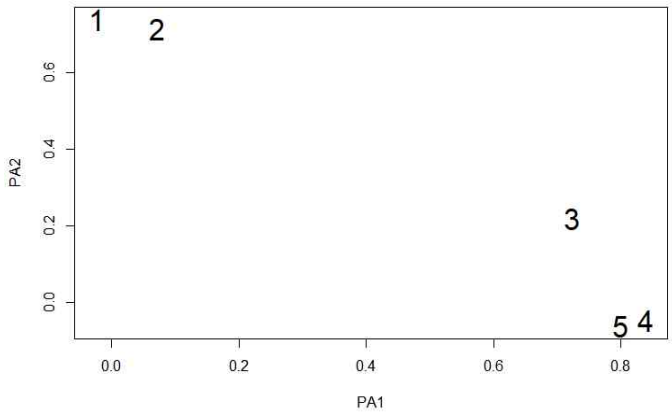
With factor correlations of

	PA1	PA2
PA1	1.00	0.76
PA2	0.76	1.00

두 인자는 각각 분산의 약 40%, 25% 정도를 설명하고 있다. 이는 두 개의 인자가 전체 분산의 63%를 설명한다는 의미이기도 하다. 또한 두 인자의 상관계수가 0.76인 것으로 보아, 상당 부분 상관되어 있으며, FA 수행 과정에서 인자 간 다소 상관되어 있을 것으로 예측하여 사교 회전을 선택한 것이 적절했다는 근거가 되기도 한다. 이어지는 과정들은 인자 부하의 시각화와 fa.diagram() 함수를 이용한 인자분석 결과이다.

인자 부하의 시각화 (R코드 및 결과)

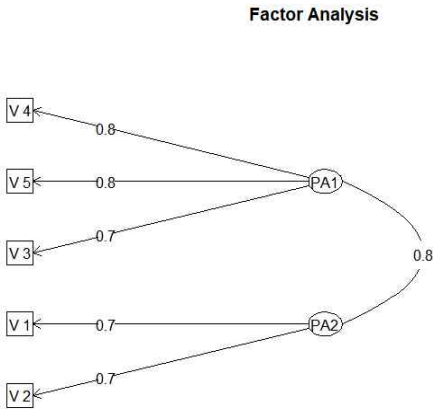
```
> load<-FA$loadings[,1:2]
> plot(load, type="n")
> text(load, cex=2)
```



위의 그래프를 보면, fa() 함수를 이용한 FA의 결과와 같이 1, 2(Mechanics, Vector)와 3, 4, 5(Algebra, Analysis, Statistics)가 두 개의 그룹으로 구분됨을 확인할 수 있다. FA의 과정에서 3(Algebra)의 인자부하에 대해 주목하였는데, 같은 이유로 3(Algebra)이 PA1에 높은 인지부하를 갖는 그룹에서 약간 떨어져있음을 알 수 있다.

fa.diagram() 함수를 이용한 인자분석 결과 (R코드 및 결과)

```
> fa.diagram(FA)
```



위 그림 또한 인자분석의 결과를 다른 방법으로 시각화하였을 뿐, fa() 함수를 이용한 FA의 결과와 같고, 인자부하량까지 거의 같다는 것을 알 수 있다.

13장 판별분석 연습문제

1. iris 자료에서 두 종류의 Species(setosa와 versicolor)만을 이용하여 선형판별분석을 수행하고자 한다. 단, 문제의 단순화를 위해 처음 두 개 변수 Sepal.Length와 Sepal.Width만 사용한다.

(a) 선형판별분석을 실시하고, 그 결과를 시각화하여라.

Solve)
다음은 lda() 함수를 이용한 선형판별분석(LDA) 수행과정이다.

lda() 함수를 아용한 선형판별분석(LDA) (R코드 및 결과)

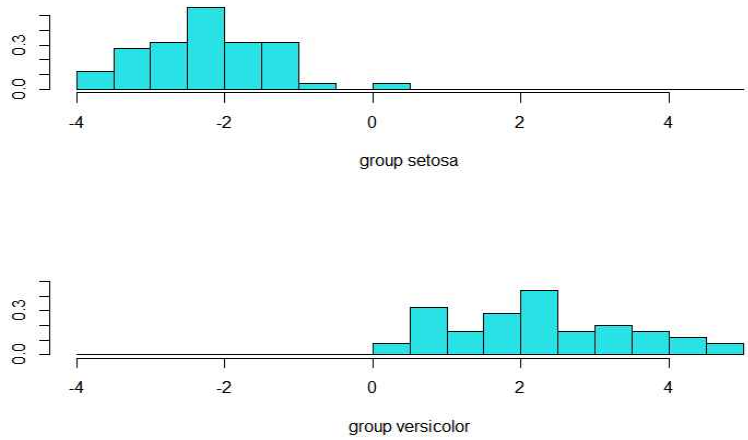
```
> library(MASS)
> data1<-subset(iris,Species=="setosa")[,c("Sepal.Length", "Sepal.Width", "Species")]
> data2<-subset(iris,Species=="versicolor")[,c("Sepal.Length", "Sepal.Width", "Species")]
> data<-rbind(data1, data2)
> r<-lda(formula = Species ~ Sepal.Length + Sepal.Width, data=data)
Warning message:
In lda.default(x, grouping, ...) : group virginica is empty
> r
Call:
lda(Species ~ Sepal.Length + Sepal.Width, data = data)

Prior probabilities of groups:
    setosa versicolor
       0.5        0.5

Group means:
           Sepal.Length Sepal.Width
setosa            5.006       3.428
versicolor        5.936       2.770

Coefficients of linear discriminants:
                   LD1
Sepal.Length  2.560968
Sepal.Width  -3.167079
> r$svd
[1] 22.32819
```

위의 결과에서 알 수 있듯이 lda() 함수는 각 군집의 사전확률과 빈도 및 각 공변량의 군집-특정적 평균을 제공한다. 문제에서는 prior= 옵션, 즉 군집의 사전확률에 대해 생략되었으므로 훈련용 자료의 군집비율인 0.5가 사용되었음을 알 수 있다. 추가로 1개의 선형판별식(2개의 군집을 분류하는 문제이기 때문일 것이다.)과 선형결합계수(scaling) 그리고 특잇값을 확인할 수 있다. 특잇값분해는 선형판별변수의 그룹-내와 그룹-간의 표준편차의 비를 나타낸다. 따라서 특잇값으로부터 선형판별식에 의해 설명되는 그룹-간 분산의 양을 계산할 수 있다.
다음은 특잇값으로부터 그룹-간 분산의 양을 구하는 과정이다.

```
> r$svd^2/sum(r$svd^2)
[1] 1
```

이 문제에서는 하나의 선형판별식이 존재하여, LD1이 자료의 그룹-간 분산 전체를 설명하는 것은 당연한 것으로 보인다.

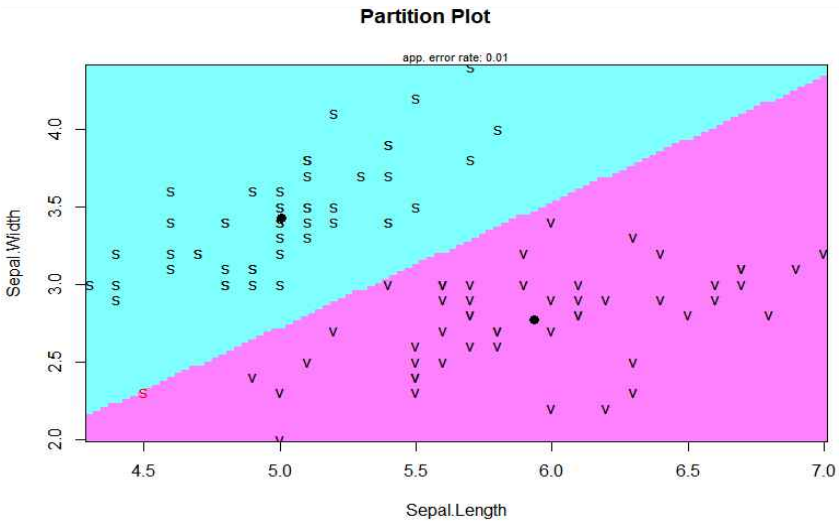
다음은 1개의 선형판별식에 따른 자료의 판별 차원상 분포이다.



추가로 {klaR} 패키지의 partimat() 함수를 이용하여, LDA의 결과를 시각화할 수 있다. 이 함수는 군집결과에 대한 이차원 분할 그림을 제공한다.

partimat() 함수를 이용한 LDA 결과 시각화(R코드 및 결과)

```
> library(klaR)
> partimat(Species ~ Sepal.Length + Sepal.Width, data=data, method="lda" )
Warning message:
In lda.default(x, grouping, ...) : group virginica is empty
```



위의 그림을 보면 lda() 함수를 이용한 LDA 결과에서 확인한 1개의 선형판별식이 대각선으로 나타나 있음을 확인할 수 있다. 또한 대각선을 기준으로 위쪽은 Species의 setosa의 군집으로, 아래로는 Species의 versicolor의 군집으로 명확하게 분류되었음을 보인다.

이처럼, versicolor 군집과 setosa 군집이 선형판별식으로 겹치는 부분 없이 명확하게 구분되는 것은 그룹-간 분산이 그룹-내 분산에 비해 크기 때문이라고 볼 수 있다,

(b) 다음 식을 이용하여 합동공분산행렬을 구하고, 이를 이용하여 아래 그림과 같이 등고선 그림(이변량 정규분포의 적합결과)를 시각화하여라.

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

여기서 S_1과 S_2는 두 집단의 표본 공분산행렬이다.

Solve)
다음은 문제의 식을 이용하여 합동공분산행렬을 구하는 과정이다.

합동공분산행렬 (R코드 및 결과)

```
> data1<-subset(iris,Species=="setosa")[,c("Sepal.Length", "Sepal.Width", "Species")]
> data2<-subset(iris,Species=="versicolor")[,c("Sepal.Length", "Sepal.Width", "Species")]
> s1<-cov(data1[, c("Sepal.Length", "Sepal.Width")])
> s2<-cov(data2[, c("Sepal.Length", "Sepal.Width")])

 > s1          #setosa의 (표본)공분산행렬
              Sepal.Length Sepal.Width
Sepal.Length   0.12424898  0.09921633
Sepal.Width    0.09921633  0.14368980

> s2       #versicolor의 (표본)공분산행렬
              Sepal.Length Sepal.Width
Sepal.Length   0.26643265  0.08518367
Sepal.Width    0.08518367  0.09846939

> s<-((s1*49)+(s2*49))/98
> s           #합동공분산행렬
              Sepal.Length Sepal.Width
Sepal.Length    0.1953408   0.0922000
Sepal.Width     0.0922000   0.1210796
```

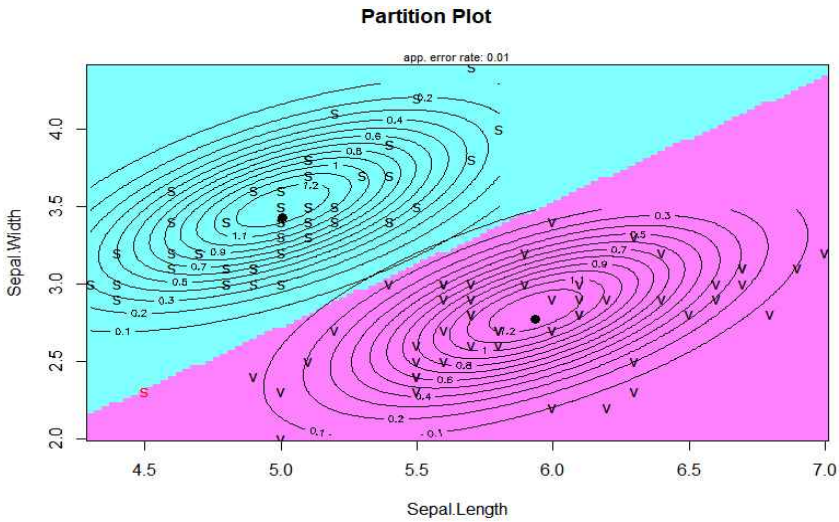
위의 코드에서 s1, s2는 각각 Species의 setosa와 versicolor의 표본공분산행렬이다.
자료를 통해, n_1, n_2이 50임을 알 수 있으므로, 합동공분산행렬(s)를 구할 수 있다.
또한 합동공분산행렬을 구하게 되면서, 등고선 그림(이변량 정규분포의 적합결과)를 시각화할 수 있다.

등고선 그림(이변량 정규분포의 적합결과) (R코드 및 결과)

```
> library(klaR)
> library(mvtnorm)
> partimat(Species~Sepal.Width+Sepal.Length,data=data,method="lda")
Warning message:
In lda.default(x, grouping, ...) : group virginica is empty
> x<-seq(min(data1$Sepal.Length),max(data1$Sepal.Length),0.01)
> y<-seq(min(data1$Sepal.Width),max(data1$Sepal.Width),0.01)
> f<-function(x,y){dmvnorm(cbind(x,y),mean=c(mean(data1$Sepal.Length),mean(data1$Sepal.Width))
```

,sigma=s)}

```
> contour(x,y,outer(x,y,f),add=TRUE)
> x<-seq(min(data2$Sepal.Length),max(data2$Sepal.Length),0.01)
> y<-seq(min(data2$Sepal.Width),max(data2$Sepal.Width),0.01)
> f<-function(x,y){dmvnorm(cbind(x,y),mean=c(mean(data2$Sepal.Length),mean(data2$Sepal.Width))
,sigma=s)}
> contour(x,y,outer(x,y,f),add=TRUE)
```



위의 결과는 (a)의 partimat() 함수를 이용하여, LDA의 결과를 시각화한 그림에 합동공분산행렬을 이용하여 등고선 그림을 추가한 것이다.