다변량
8장 과제

과목	다변량데이터분석
담당교수	임태진 교수님
전공	산업정보시스템공학과
학번	20201368
이름	한채원
제출일	2022.11.30

유인물 CH-8b.pdf 를 참조하여 [유럽 산업 고용 데이터]에 대한 판별분석을 수행하시오.

1. 가정에 대한 검토

우선 고용패턴에 기초해서 국가 그룹들 간 구별을 한다.

```
> df = read.csv('euro.csv',row.names=1, strings=T)
> dfm = as.matrix(df[,-1])
> str(dfm)
 num [1:30, 1:9] 2.6 5.6 5.1 3.2 22.2 13.8 8.4 3.3 4.2 11.5 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:30] "Belgium" "Denmark" "France" "Germany" ...
  ..$ : chr [1:9] "AGR" "MIN" "MAN" "PS" ...
```

9개 비율 변수들 중 어느 하나는 100에서 나머지 변수들의 차로 표현될 수 있으므로 TC는 생략한다.

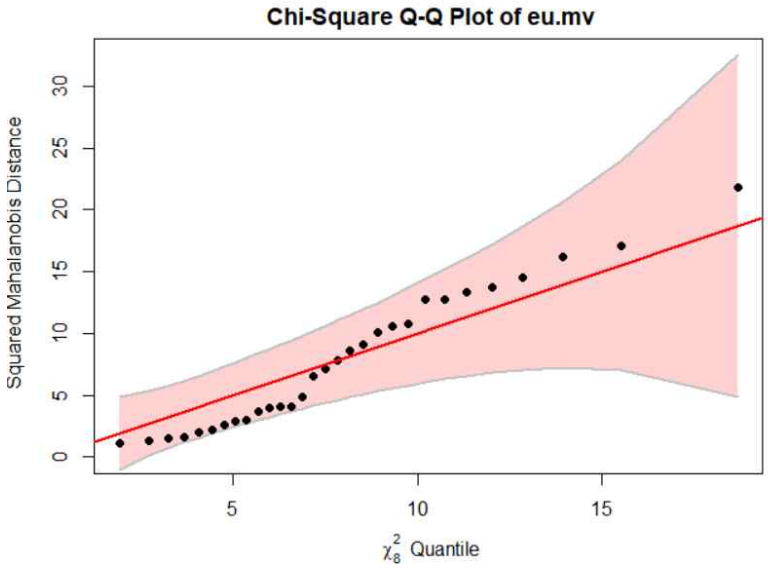
```
> all(round(rowSums(dfm))==100)
[1] TRUE
> dfm = dfm[,-9]
```

(1) MANOVA

```
> Group = df$Group
> eu.mv = manova(dfm~Group)
> (eusum = summary(eu.mv,test='Wilks'))
           Df     Wilks approx F num Df den Df     Pr(>F)
Group       3 0.083483     3.143     24 55.707 0.0002195 ***
Residuals 26
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

p값을 볼 때 이는 네 그룹들간의 유의미한 차이가 있다고 볼 수 있다.

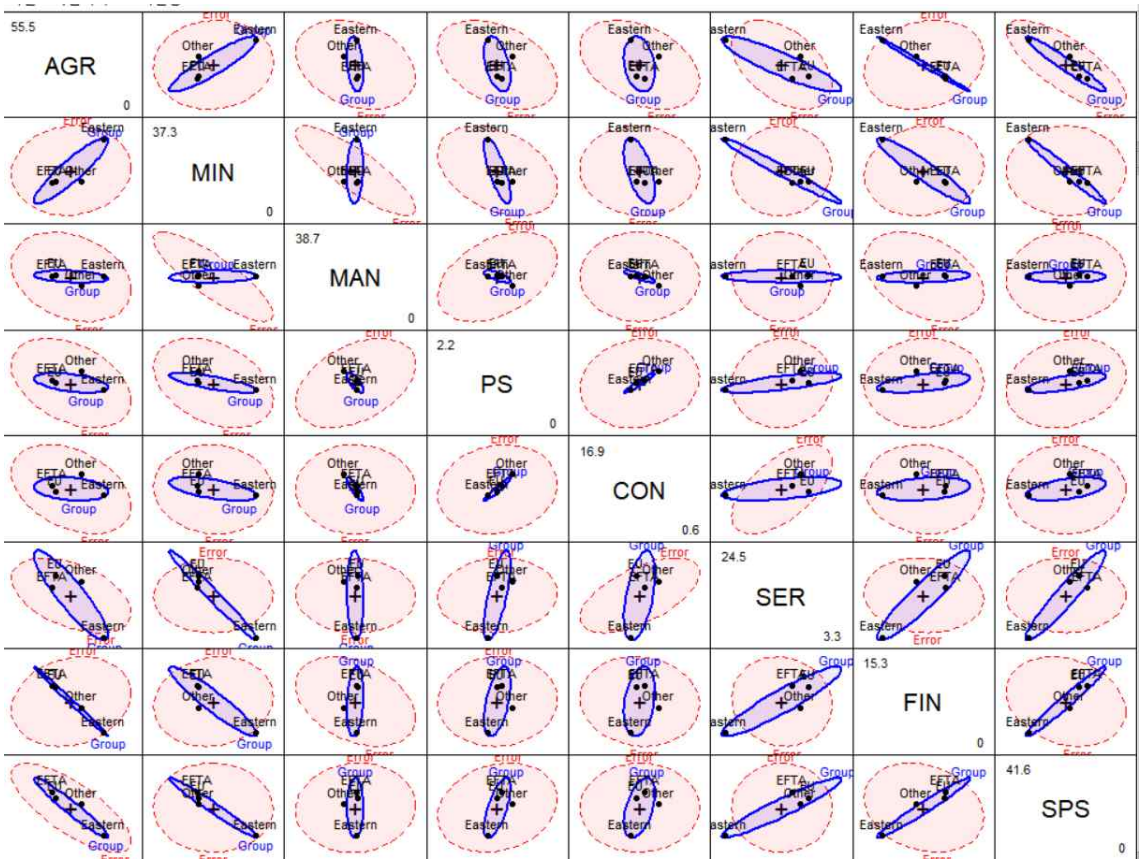
```
> heplots::cqplot(eu.mv)
```



그래프가 선형을 이루고 있으므로 이는 모집단이 정규성을 따름을 알 수 있다.

(2) Residuals

```
> MVN::mvn(eu.mv$resid,desc=F)
$multivariateNormality
             Test       HZ       p value MVN
1 Henze-Zirkler 1.191327 3.030909e-14  NO

$univariateNormality
              Test   Variable Statistic   p value Normality
1 Anderson-Darling      AGR      2.0077   <0.001      NO
2 Anderson-Darling      MIN      4.0158   <0.001      NO
3 Anderson-Darling      MAN      1.3809   0.0011      NO
4 Anderson-Darling      PS       0.6966   0.0618      YES
5 Anderson-Darling      CON      0.6885   0.0648      YES
6 Anderson-Darling      SER      0.1400   0.9703      YES
7 Anderson-Darling      FIN      1.1497   0.0044      NO
8 Anderson-Darling      SPS      0.3119   0.5319      YES

> win.graph(12,9)
> pairs(eu.mv,fill=T,fill.alpha = 0.1)
```



그래프를 살펴보면 이는 선형판별분석을 하는게 적절하다고 보여진다고 판단할 수 있다.

2. 선형 판별 분석

```
> eu.can = candisc(eu.mv)
> summary(eu.can)
```

Canonical Discriminant Analysis for Group:

	CanRsq	Eigenvalue	Difference	Percent	Cumulative
1	0.84249	5.34862	4.7788	87.3907	87.391
2	0.36297	0.56980	4.7788	9.3098	96.701
3	0.16801	0.20194	4.7788	3.2995	100.000

Class means:

	Can1	Can2	Can3
Eastern	3.53910	-0.025153	-0.090378
EFTA	-0.77816	0.669536	0.719944
EU	-1.55537	0.244758	-0.387236

```
Other    -1.24484 -1.688273  0.262547

 std coefficients:
         Can1     Can2      Can3
AGR  -4.77318  -7.5369  -8.17939
MIN  -2.25691  -4.4255  -6.79778
MAN  -3.55442  -5.4444  -8.64357
PS   -0.21713  -1.0109  -0.26303
CON  -0.62234  -1.9103  -1.46749
SER  -2.46115  -2.3550  -3.20356
FIN  -1.56365  -1.1749  -2.40713
SPS  -3.98354  -5.2866  -5.91895
```

정준변수의 계수들을 살펴본다.

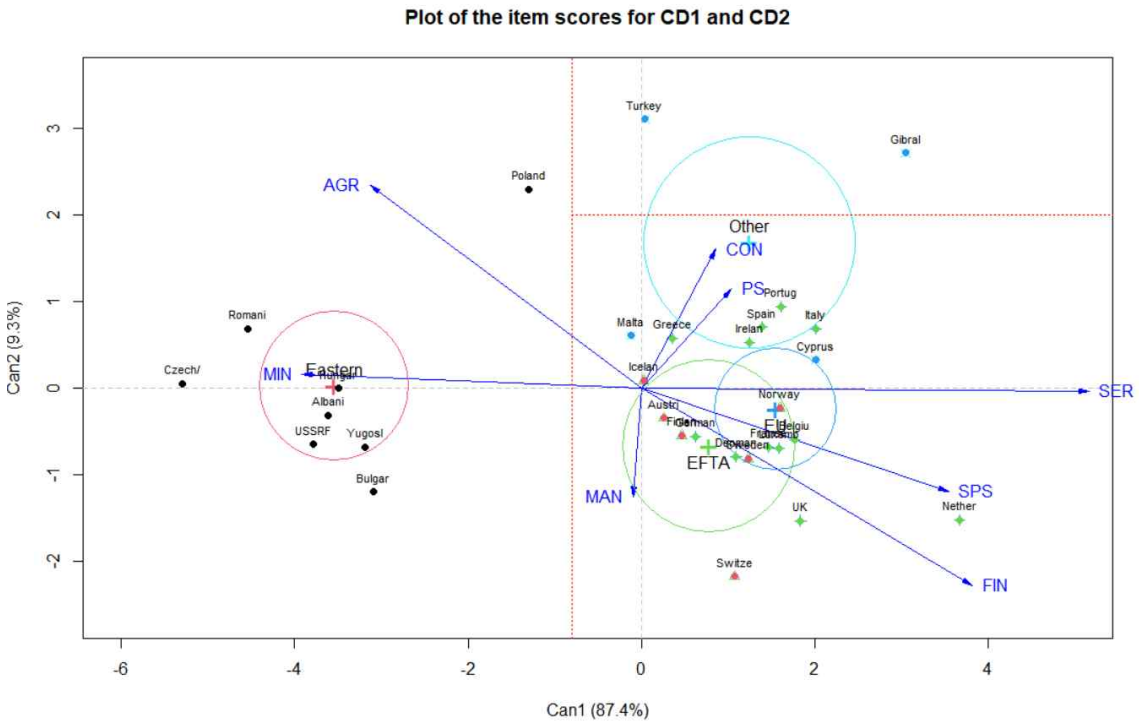
```
> round(eu.can$coeffs.raw,3)
      Can1   Can2   Can3
AGR  -0.427 -0.674 -0.732
MIN  -0.295 -0.579 -0.889
MAN  -0.359 -0.550 -0.873
PS   -0.339 -1.576 -0.410
CON  -0.222 -0.682 -0.524
SER  -0.688 -0.658 -0.895
FIN  -0.464 -0.349 -0.714
SPS  -0.514 -0.682 -0.764
```

원 변수와 정준변수와의 상관계수에 대해 알아본다. 세 정준변수에 해당하는 고유값들이 모두 양수이므로 상관관계를 고려해본다.

```
> eust1 = eu.can$structure;round(eust1,3)
      Can1   Can2   Can3
AGR   0.497 -0.372 -0.091
MIN   0.622 -0.025 -0.196
MAN   0.016  0.199 -0.117
PS   -0.166 -0.182  0.225
CON  -0.136 -0.255  0.339
SER  -0.821  0.006 -0.080
FIN  -0.607  0.362  0.089
SPS  -0.563  0.189  0.279

> eucs = eu.can$scores
> win.graph(12,8)
> plot(eu.can, xlim=c(-6,5), rev.axes=c(TRUE, TRUE))
```

Vector scale factor set to 6.295

```
> points(-eucs[,2], -eucs[,3], pch=19, col=Group)
> text(-eucs[,2], -eucs[,3], substr(rownames(df),1,6), cex=0.7, pos=3)
> title("Plot of the item scores for CD1 and CD2")
> abline(v=-0.8, lty=3, col="red")
> segments(-0.8,2, 6,2, lty=3, col="red")
```

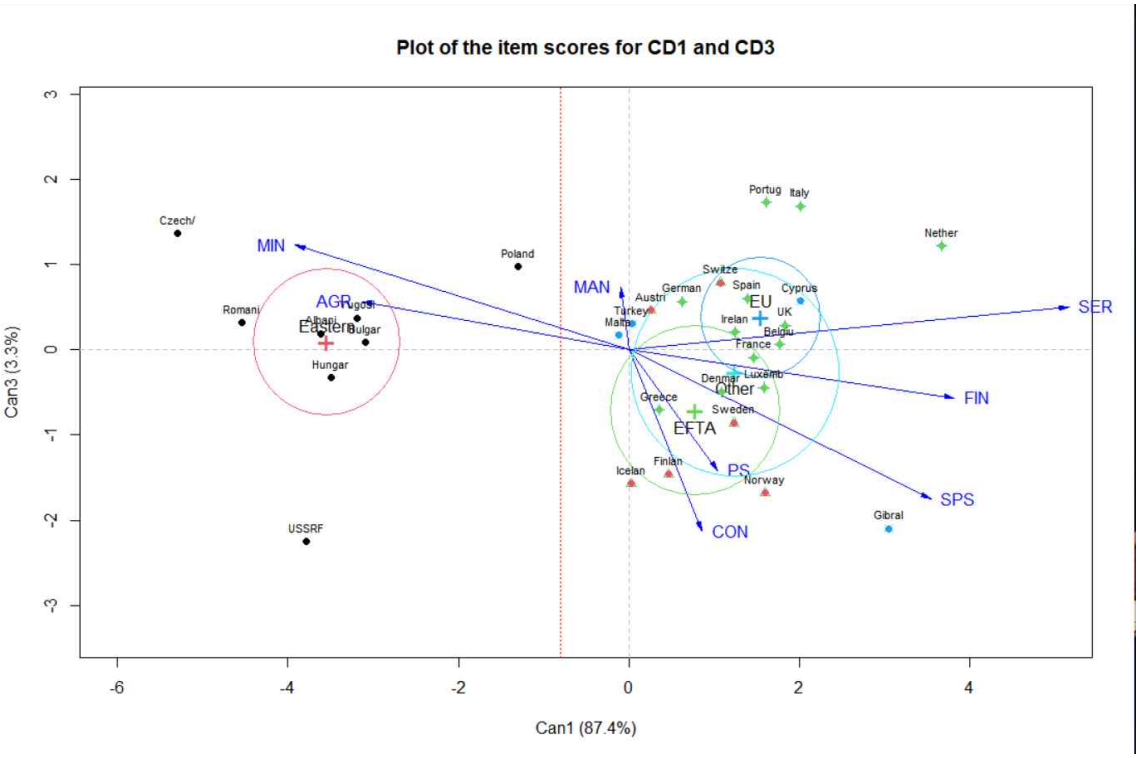


첫 번째 판별함수와 두 번째 판별함수간의 관계에 대해 살펴보면, 동유럽 국가에서는 서비스 산업보다 전통 산업이 중점인 것처럼 보인 반면에, 이를 제외한 다른 국가들은 반대의 성향을 보임을 파악할 수 있다. 또한 터키와 지브롤터는 농업과 건설이 중점적으로 보임을 확인할 수 있다.

```
> plot(eu.can, c(1,3), xlim=c(-6,5), rev.axes=c(TRUE, TRUE))
```

Vector scale factor set to 6.295

```
> points(-eucs[,2], -eucs[,4], pch=19, col=Group)
> text(-eucs[,2], -eucs[,4], substr(rownames(df),1,6), cex=0.7, pos=3)
> abline(v=-0.8, lty=3, col="red")
> title("Plot of the item scores for CD1 and CD3")
```



첫 번째 판별함수와 세 번째 판별함수간의 관계에 대해 살펴보면, 동유럽 국가에서는 서비스 산업보다 전통 산업이 중점인 것처럼 보인 반면에, 이를 제외한 다른 국가들은 반대의 성향을 보임을 파악할 수 있다.

위의 두 그래프에서 봤듯이 동부유럽 국가들은 그 외 다른 나라들과 구분하는데 성공적이지만, 다른 그룹들을 구분하는 것은 성공적이지 못하다.

3. 분류 예측 및 정확성 평가

```
> eu.lda = MASS::lda(Group~., data=df);eu.lda
Call:
lda(Group ~ ., data = df)

Prior probabilities of groups:
  Eastern      EFTA        EU     Other
0.2666667 0.2000000 0.4000000 0.1333333

Group means:
              AGR       MIN      MAN       PS      CON      SER
Eastern 21.462500 11.7875000 20.56250 0.6375000 6.912500  9.38750
EFTA     6.833333  0.3166667 20.53333 0.8666667 7.900000 16.75000
EU       7.666667  0.4500000 20.99167 0.7916667 7.283333 18.55833
Other   15.225000  0.4500000 17.25000 1.0500000 8.950000 17.70000
              FIN       SPS       TC
```

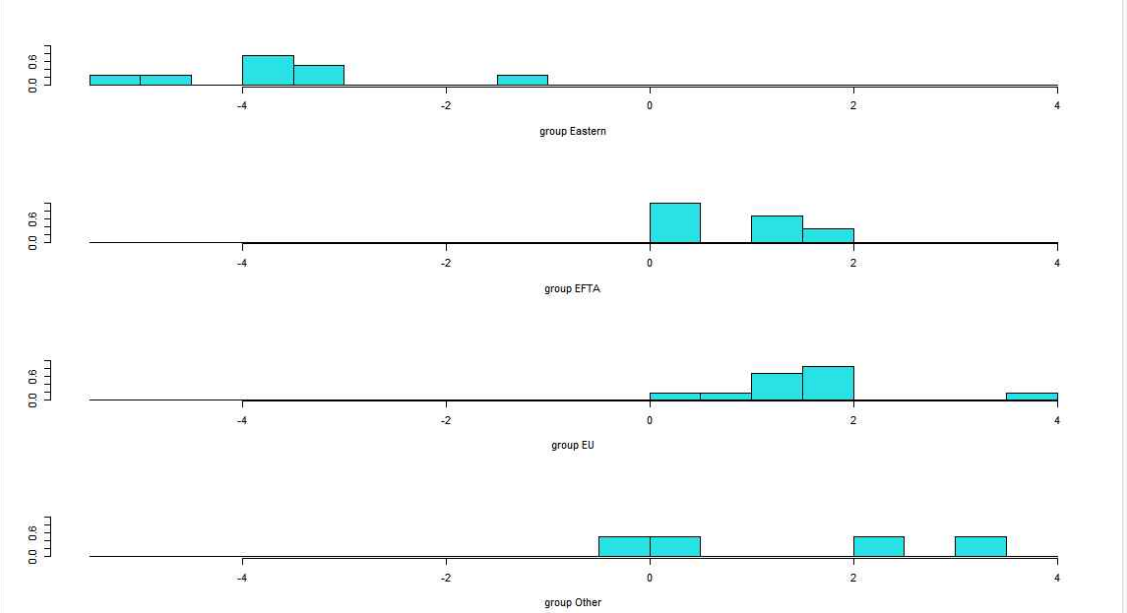
```
Eastern 3.000000 19.45000 6.800000
EFTA     8.500000 31.21667 7.016667
EU       8.391667 29.63333 6.208333
Other    5.950000 27.82500 5.650000

Coefficients of linear discriminants:
           LD1         LD2         LD3
AGR 0.9877410   9.623775 -3.445051
MIN 0.8544965   9.528676 -3.606094
MAN 0.9184923   9.506581 -3.595623
PS  0.9234748 10.748632 -3.116055
CON 0.7801160   9.577392 -3.227056
SER 1.2510719   9.653517 -3.618870
FIN 1.0285344   9.414589 -3.496837
SPS 1.0709938   9.571578 -3.456259
TC  0.5792659   9.336774 -2.927820
Proportion of trace:
   LD1    LD2    LD3
0.8470 0.1198 0.0332

> pred = predict(eu.lda); round(prec, 3)
> eupr = pred$class
> conf = table(list(Pred=eupr,Obs=df$Group))
> addmargins(conf)
          Obs
Pred       Eastern EFTA EU Other Sum
  Eastern        7    0  0     0   7
  EFTA           0    5  0     0   5
  EU             0    1 12     2  15
  Other          1    0  0     2   3
  Sum            8    6 12     4  30
```

다음과 같이 대각선의 값을 살펴보면, 예측력이 높다고 볼 수 있다.

```
> prec = diag(conf) / rowSums(conf); round(prec, 3)
Eastern    EFTA      EU   Other
  1.000   1.000   0.800   0.667
> sens = diag(conf) / colSums(conf); round(sens, 3)
Eastern    EFTA      EU   Other
  0.875   0.833   1.000   0.500
> caret::confusionMatrix(conf)
Overall Statistics
```

Accuracy : 0.8667
95% CI : (0.6928, 0.9624)
No Information Rate : 0.4
P-Value [Acc > NIR] : 1.769e-07
Kappa : 0.80710 ...

```
> ldahist(data = pred$x[,1], g=Group)
```



정밀도와 민감도와 그래프를 살펴보면 판별이 잘 된 케이스라고 볼 수 있지만 Eastern과 그 외 다른 국가들과의 구분은 뚜렷하지만 Eastern을 제외한 국가들끼리의 판별이 잘 된 케이스가 아니라고 볼 수 있다.