**# 1 (Exercise 14.6)**

Swiss Bank Note 데이터에는 100개의 진짜 bank note와 100개의 가짜 bank note에 대한 각각의 측정값이 기록되어 있다. 이 여섯 종류의 측정값으로 진짜 bank note와 가짜 bank note를 구별하는 것이 목표이다. 이때 판별분석이 이용된다. 판별분석이란 우리가 관심 있는 개체가 어떤 집단에 속하는지 구분하는 함수에 대한 분석 기법이다. 판별분석을 통해 판별함수를 만들 수 있다. 판별함수는 집단을 판별해내는 함수로 선형판별함수와 비선형판별함수가 있다. 선형판별함수는 독립변수들의 선형결합으로 이루어져 있다. 선형판별함수는 비선형판별함수와 달리 첫 번째 그룹 관측치의 공분산 행렬과 두 번째 그룹 관측치의 공분산 행렬이 같다고 가정한다. 이렇게 판별함수를 만들어낸 이후에는 새로운 개체가 주어졌을 때 이 개체가 어떤 집단에 속해 있는지 알 수 있다. 그리고 판별함수의 계수를 통해 집단을 판별할 때 가장 공헌이 큰 변수를 찾아낼 수 있다. 따라서 판별분석을 통해 진짜 bank note와 가짜 bank note를 구별해내는 판별함수를 만들기로 한다.

진짜 bank note와 가짜 bank note에 대한 측정값 각각의 공분산 행렬을 비교해보면 차이가 작다. 따라서 선형판별함수로 두 집단을 판별한다. 먼저 진짜 bank note 관측치 공분산 행렬과 가짜 bank note 관측치 공분산 행렬을 통해서 $\boldsymbol{W}$ 행렬을 구한다.

$$\boldsymbol{W} = \mathbf{100}(\boldsymbol{S_f} + \boldsymbol{S_g})$$

$\boldsymbol{W}$ 행렬을 통해서 벡터 $\boldsymbol{a}$를 구한다. 이 벡터는 집단 간 분산과 집단 내 분산의 비율을 최대화한다.

$$\boldsymbol{a} = \boldsymbol{W^{-1}}(\overline{\boldsymbol{x_g}} - \overline{\boldsymbol{x_f}}) = (\mathbf{0.0002}, \mathbf{0.0289}, -\mathbf{0.0295}, -\mathbf{0.0388}, -\mathbf{0.0409}, \mathbf{0.0541})^{\boldsymbol{T}}$$

그리고 판별함수가 만들어진다. 진짜 bank note의 경우에는 다음 식이 성립하면 진짜 bank note로 판별한다.

$$\boldsymbol{a^T}(\boldsymbol{x_g} - \overline{\boldsymbol{x}}) \geq \mathbf{0}$$

모든 관측치에 대해 판별함수를 적용하여 classification을 한다. 그리고 misclassification된 것이 몇 개인지 구한다. [그림1]은 confusion matrix이다. 가짜 bank note인데 진짜 bank note로 분류된 경우는 없다. 하지만 진짜 bank note인데 가짜 bank note라고 판별한 경우는 1건 존재한다. 따라서 misclassification된 경우는 총 1번 밖에 없으며 전체 error rate는 0.005로 매우 낮다.

[그림2]는 진짜 bank note 관측치와 가짜 bank note 관측치를 projection 했을 때의 그래프이다. 각각 평균이 서로 멀리 떨어져 있어 겹치는 부분이 적기 때문에 잘 분류된 것으로 해석된다. 이로써 진짜 bank note와 가짜 bank note는 형태가 다르며 그 형태로 둘을 구분할 수 있다는 것을 알 수 있다.

**# 2 (Exercise 14.7)**

WAIS 데이터는 두 그룹의 사람들을 대상으로 한 4번의 테스트(information, similarities, arithmetic, picture completion) 결과를 보여주고 있다. 사람들은 12명의 노년층(Group2)과 37명의 노년층이 아닌 사람들(Group1)로 나누어져 있다. 해당 테스트의 결과로 판별분석을 하여 노년층과 노년층이 아닌 사람으로 분류한다. 만약 판별함수가 노년층과 노년층이 아닌 사람으로 잘 분류된다면 노년층과 노년층이 아닌 사람은 해당 능력에 차이가 있음을 알 수 있을 것이다.

위와 비슷한 방법으로 판별함수를 만든다. 집단 간 분산과 집단 내 분산의 비율을 최대화하는 벡터 $\boldsymbol{a}$는 다음과 같다.

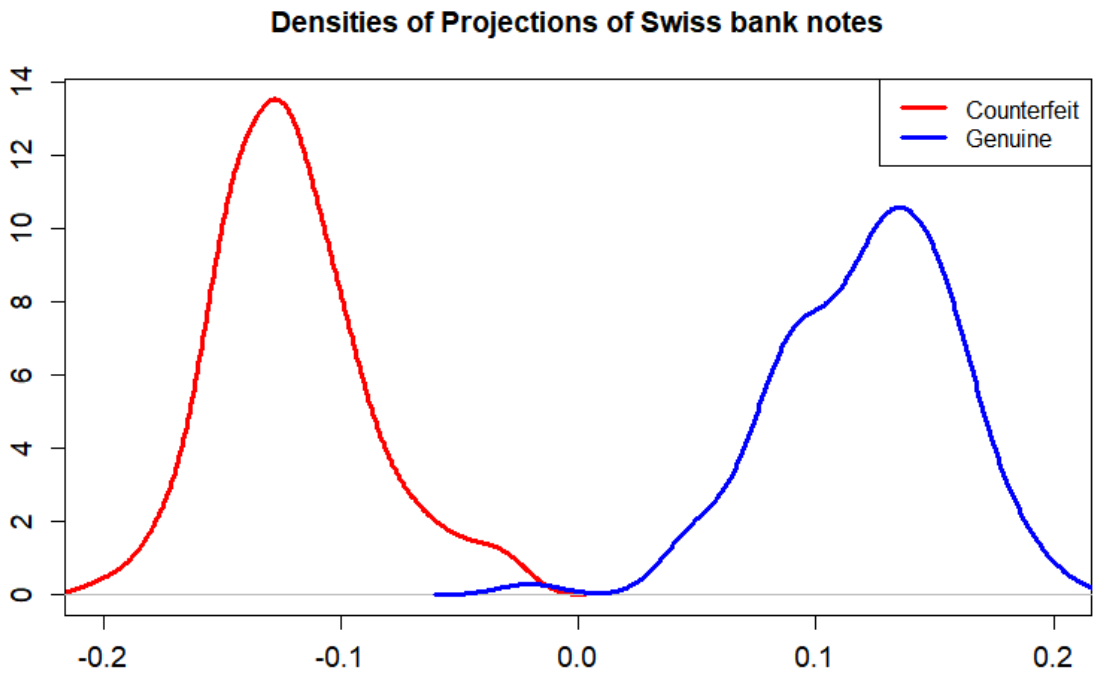
$$\boldsymbol{a} = (\mathbf{0.0007}, \mathbf{0.0006}, \mathbf{0.00001}, \mathbf{0.0012})^{\boldsymbol{T}}$$

그리고 그 classification의 결과를 나타내는 confusion matrix는 [그림3]이다. Group1에 해당하는데 Group2라고 분류한 경우는 10건이다. 그리고 Group2에 해당하는데 Group1이라고 분류한 경우는 3건이다. 총 12건이 misclassification되었다. 총 error rate는 0.265이다. 위의 결과보다는 error rate가 높다.

[그림4]의 그래프는 Group1과 Group2의 테스트 성적을 projection 시킨 것을 나타내는 그래프이다. Group1과 Group2의 평균이 멀리 떨어져 있고 Group2의 분산이 커서 [그림2]보다 겹치는 부분이 많다. 따라서 위의 결과보다 error rate가 높게 나온 것이다. Error rate가 다소 높기 때문에 이 판별함수는 두 그룹으로 잘 분류한다고 할 수는 없다.

| | Genuine | Counterfeit |
|---|---|---|
| **Predicted Genuine** | 99 | 0 |
| **Predicted Counterfeit** | 1 | 100 |

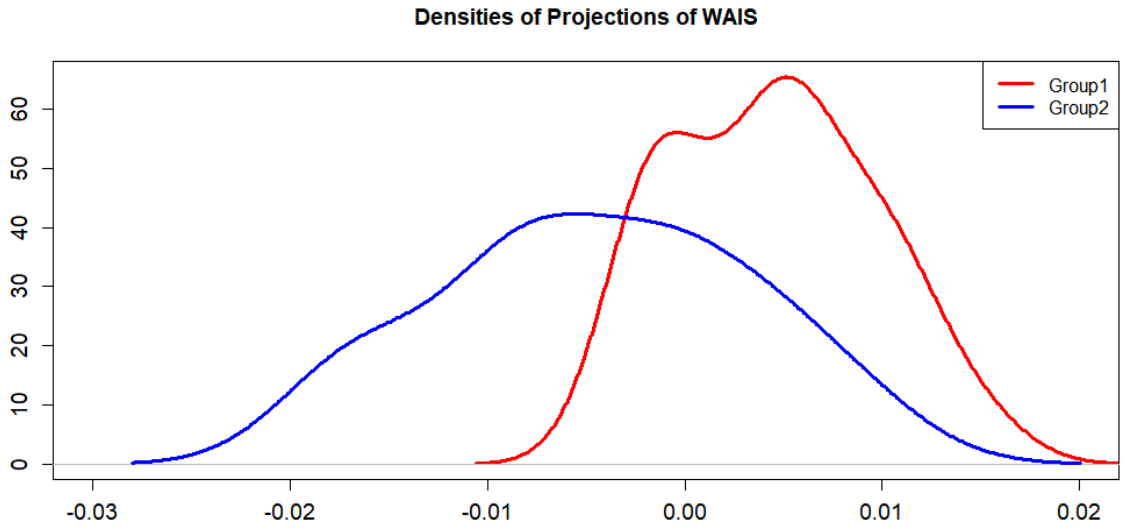
[그림1]




[그림2]

| | Group1 | Group2 |
|---|---|---|
| **Predicted Group1** | 27 | 3 |
| **Predicted Group2** | 10 | 9 |

[그림3]


Densities of Projections of WAIS
Group1
Group2
0
10
20
30
40
50
60
-0.03
-0.02
-0.01
0.00
0.01
0.02

**[그림4]**