Universidade de Brasília

DEPARTAMENTO DE ESTATÍSTICA

16 julho 2023

# Lista 10 - Análise de Discriminantes e Classificação

Prof. Dr. George von Borries

Análise Multivariada 1

Aluno: Bruno Gondim Toledo | Matrícula: 15/0167636

## 85. Johnson e Wichern - Exercício 11.1.

### a)

A função de discriminante linear é dada por:

$\hat{y} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}^{-1}_{\mathbf{pooled}} \mathbf{x} = \hat{\mathbf{a}}' \mathbf{x}$.

Como

$$\mathbf{S}^{-1}_{\mathbf{pooled}} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$$

Então a função de discriminante linear será igual a $-2x_1$.

### b)

$\hat{m} = \frac{1}{2}(\hat{y}_1 + \hat{y}_2) = \frac{1}{2}(\hat{\mathbf{a}}'\bar{\mathbf{x}}_1 + \hat{\mathbf{a}}'\bar{\mathbf{x}}_2)$ =-8.

Portanto, devemos colocar $x_0'$ na população $\pi_1$ se $\hat{y}_0 = [2 \;\; 7] x_0 \geq \hat{m}$ =-8. Caso contrário, devemos atribuir $x_0$ para a população $\pi_2$. Neste caso, $x_0$ calculado = -4; portanto, atribuímos este à população $\pi_1$

## 86. Johnson e Wichern - Exercício 11.2.

### a)

```
dados <- read_table("dados/tabela11.1.txt",
                    col_types = cols(X5 = col_skip()))

lcf <- function(x){
  X1 <- as.matrix(dados[,1:2])
  X2 <- as.matrix(dados[,3:4])
  colnames(X1) <- NULL
  colnames(X2) <- NULL
  al <- t(t(t(colMeans(X1)))-t(t(colMeans(X2)))) %*% solve(cov_pooled(X1,X2))
  limite <- .5*(al%*%colMeans(X1)+al%*%colMeans(X2))
  fronteira <- al%*%t(x)
  if(fronteira<=limite){
    return("2")
  }else{
    return("1")
  }
}

lcf(x10)
```

```
## [1] "2"
```

```
lcf(x10*100)
```

```
## [1] "1"
```

```
lcf(t(X1[2,]))
```

```
## [1] "2"
```

Como podemos perceber, a função construída avalia as observações e classifica segundo o modelo se devem ser agrupadas na população 1 ou 2, retornando no *output* somente o valor do grupo ao qual deve ser classificado o novo elemento ("1" ou "2").

### b)

A matriz de confusão será da forma:

| | pop1 | pop2 |
|---|---|---|
| pop1 | 11 | 1 |
| pop2 | 2 | 10 |

Desta forma, podemos como em uma tabela de contingência ver diretamente quais valores foram corretamente classificados, e quais não foram. Isso é essencial para conjuntos grandes, onde começa a ficar difícil contar pontinhos no gráfico..

### c)

A taxa de erro aparente é, em suma, a razão dos valores classificados equivocadamente, pelo total. Neste caso, será 0.125. Ou seja, estamos errando 12,5% das classificações com este algoritmo.

### d)

Os pressupostos deste modelo são que as observações contidas todas as populações seguem distribuição normal multivariada, com matrizes de covariâncias iguais.

## 87. Johnson e Wichern - Exercício 11.4.

### a)

A observação x será classificada como pertencente à população $\pi_1$ se $\frac{f_1(x)}{f_2(x)} \geq (\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1}) = (\frac{100}{50})(\frac{0,2}{0,8}) = 0,5$. Caso contrário, deve ser classificada como pertencente à população $\pi_2$.

### b)

Neste caso, como $f_1(x) = 0,3$ e $f_2(x) = 0,5$, então: $\frac{f_1(x)}{f_2(x)} = 0,6 \geq 0,5$. Portanto, devemos classificar $x$ como pertencente à população $\pi_1$

## 88. Johnson e Wichern - Exercício 11.10.

### a)

As hipóteses do teste serão:

$$\begin{cases} H_0)\ \boldsymbol{\mu_1} = \boldsymbol{\mu_2} \\ H_1)\ \boldsymbol{\mu_1} \neq \boldsymbol{\mu_2} \end{cases}$$

A estatística $T^2$ de Hotelling's para duas amostras é dada por:

$$T^2 = (\bar{x}_1 - \bar{x}_2)' \left[ \ (\tfrac{1}{n_1} + \tfrac{1}{n_2}\ S_{pooled}) \ \right]^{-1} (\bar{x}_1 - \bar{x}_2)$$

Que para este conjunto, será $T^2 = 14.5217134$.

Sob $H_0$);

$T^2 \sim \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{p,n_1+n_2-p-1}$

Então; $T^2 = 14.5217134 \geq \frac{(11+12-2)2}{11+12-2-1} F_{2,20} \approx 5.4374336$ para o nível de significância $\alpha = 0,1$. Portanto, rejeitamos a hipótese nula $H_0$); ou seja, temos evidências estatísticas para acreditar que as médias dos grupos diferem, confirmando a suspeita do enunciado.

### b)

O discriminante linear de Fisher será:

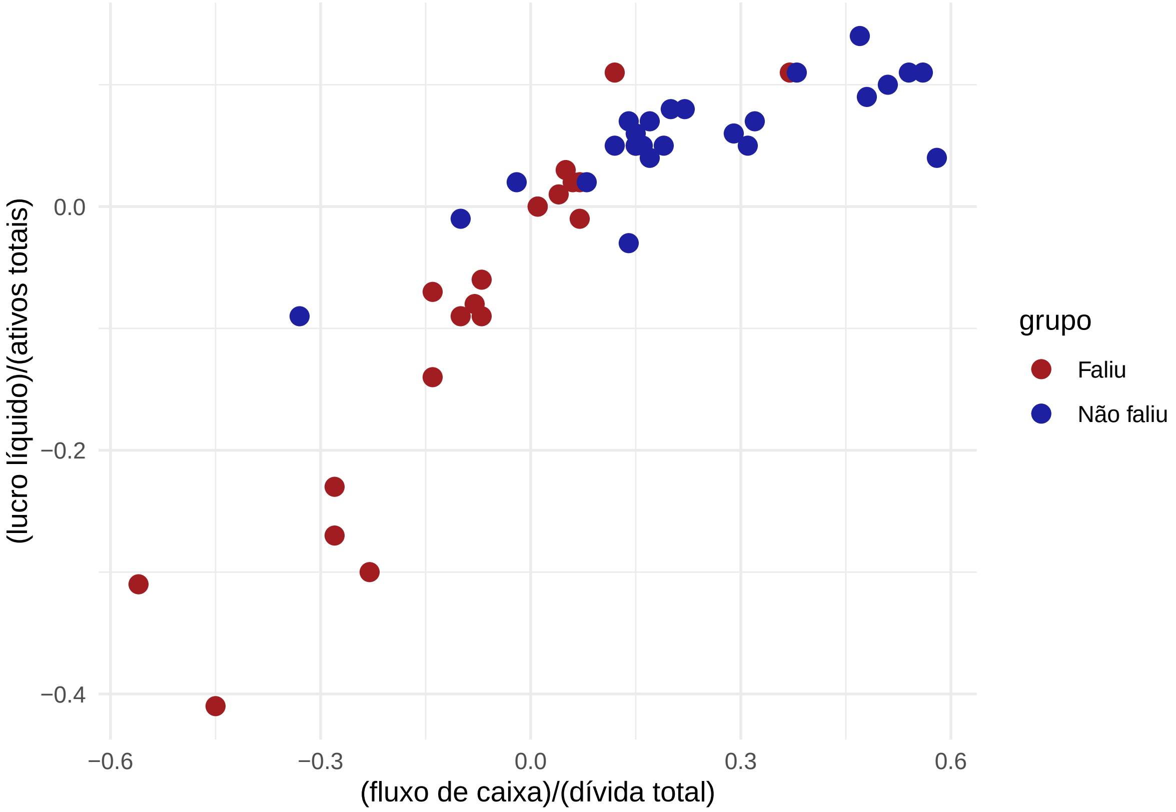
$\hat{y}_0 = \hat{\mathbf{a}}' x_0 =$ -0.4906887 $x_1$ -0.5291162 $x_2$

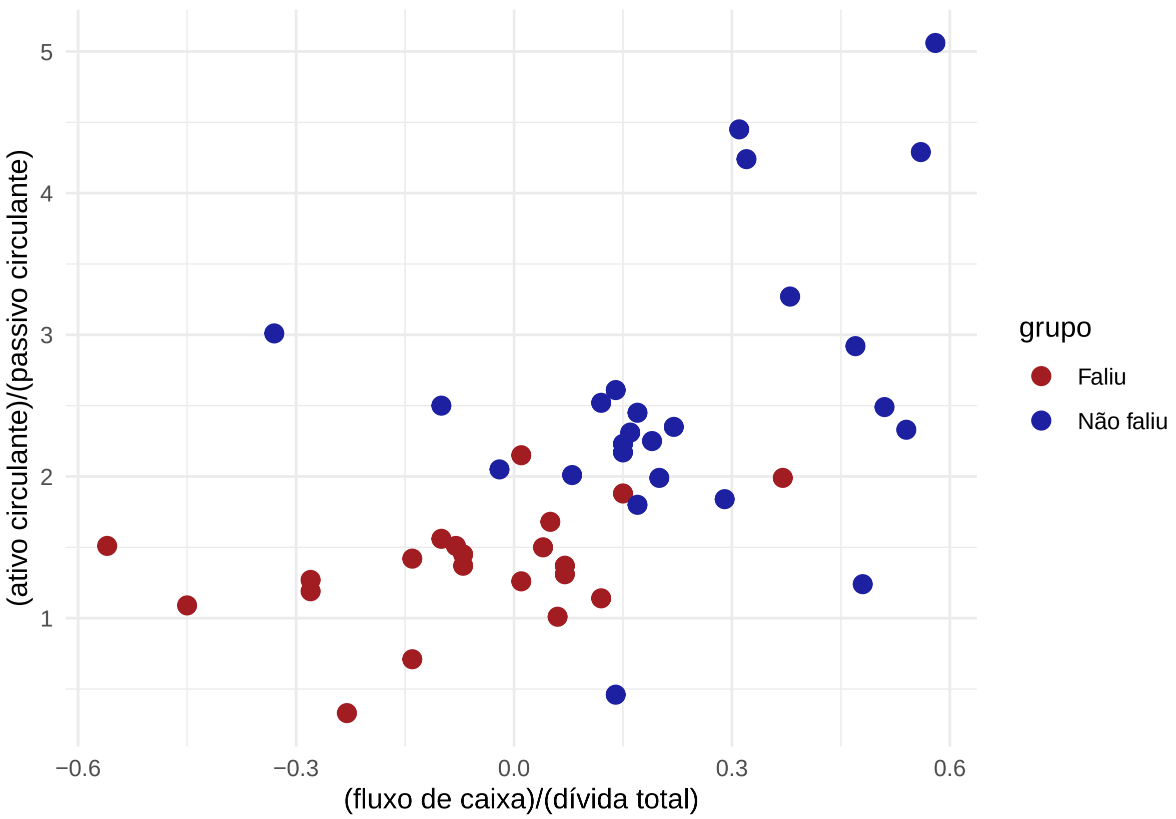
## c)

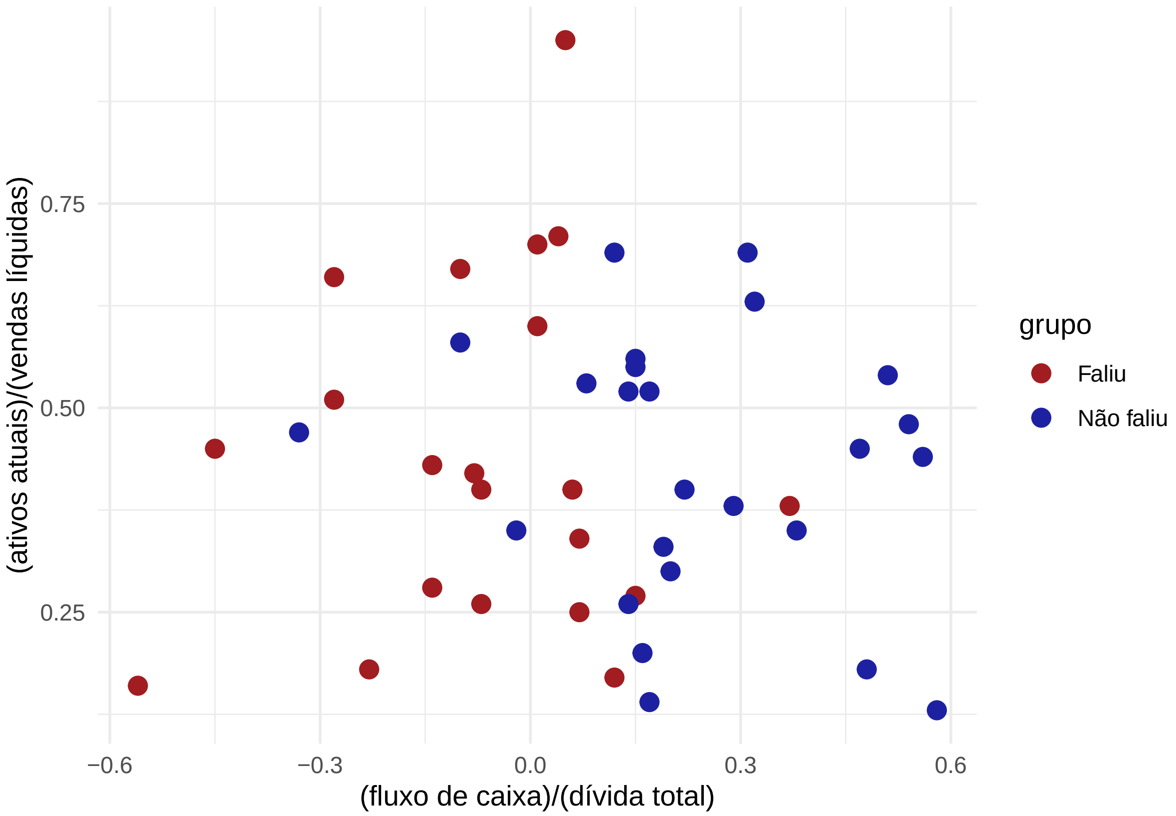
Neste caso, $\hat{m}$ = -0.2453444. Para $x_0' = [0 \;\; 1]$, o discriminante linear será $\hat{y}$ = -0.4906887 (0) -0.5291162 (1) = -0.5291162 $<$ -0.2453444. Portanto, devemos classificar $x_0'$ como pertencente à população $\pi_2$.

### 89. Johnson e Wichern - Exercício 11.24.

**a)**






Em todos os gráficos, os pontos lembram a forma de elipsoides. Portanto, graficamente, não é possível rejeitar a normalidade bivariada dos dados.

## b)

Considerando 1 como o grupo de empresas que faliram (falidos) e 2 como o grupo de empresas que não faliram ainda (ativos), temos os vetores de média $\mu_1', \mu_2'$ dados respectivamente por: [-0.0690476, -0.0814286],[0.2352, 0.0556], e matrizes de covariância $S_1 =$

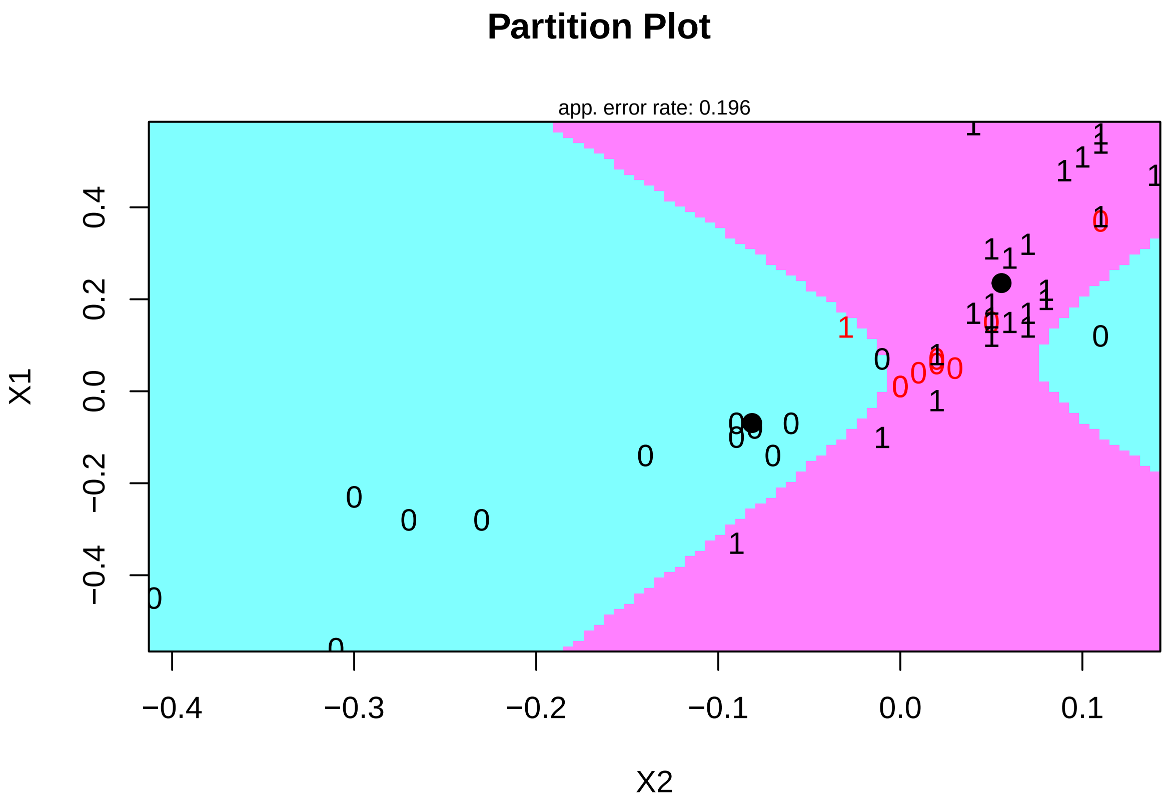
| | X1 | X2 |
|---|---|---|
| X1 | 0.0441290 | 0.0284764 |
| X2 | 0.0284764 | 0.0210029 |

e $S_2 =$

| | X1 | X2 |
|---|---|---|
| X1 | 0.0470510 | 0.0085072 |
| X2 | 0.0085072 | 0.0023757 |

## c)

Como para este conjunto não rejeitamos a hipótese de normalidade multivariada (apesar de termos feito apenas análise gráfica), e, apesar de não termos testado a igualdade das variâncias, elas aparentam ser diferentes; portanto a abordagem mais adequada para este caso é a análise discriminante quadrática abaixo. No caso, foram definido custos e prioris iguais para ambos os grupos.




Matriz de confusão:

| | 0 | 1 |
|---|---|---|
| 0 | 13 | 8 |
| 1 | 1 | 24 |

Proporção de classificações corretas em cada grupo:

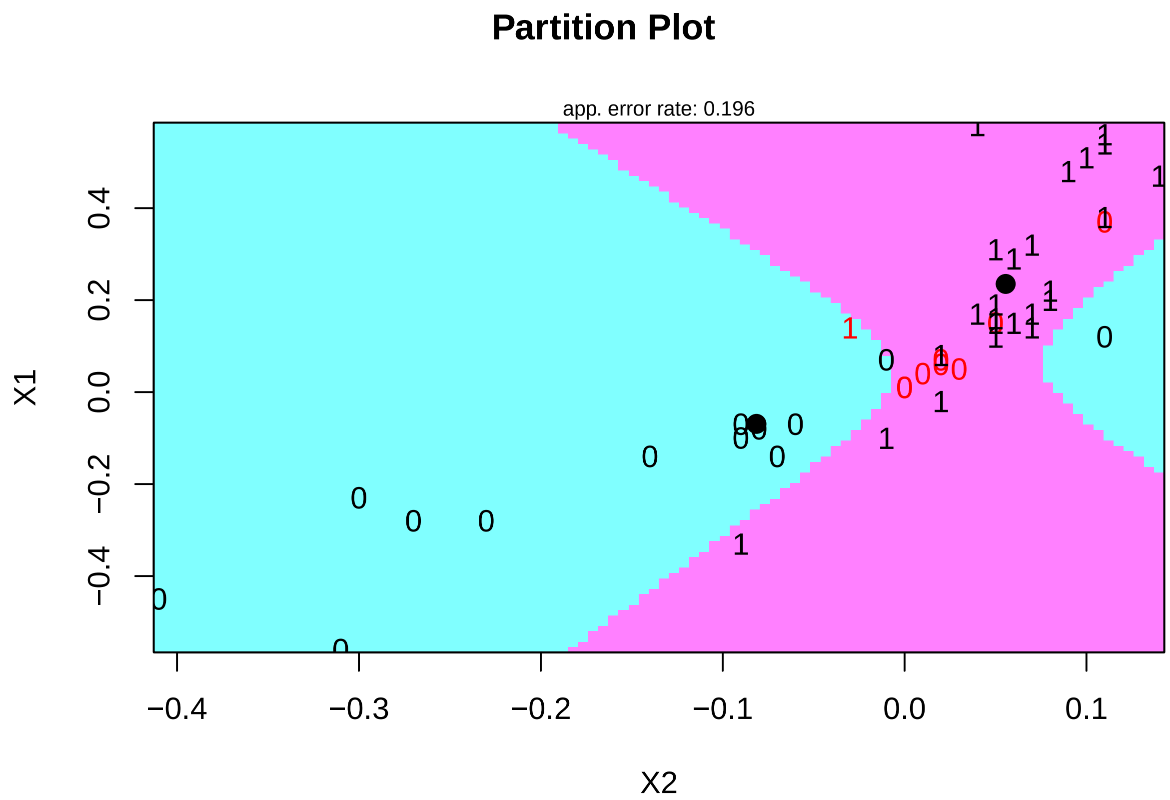
| | x |
|---|---|
| 0 | 0.6190476 |
| 1 | 0.9600000 |

Proporção total de classificação correta: 0.8043478

## d)

O erro aparente (APER) deste conjunto foi calculado como sendo 0.1956522; enquanto que a estimação da taxa de erro aparente ($\hat{E}(AER)$) foi calculada como 0.2173913. Notamos que apesar de o erro estimado via validação cruzada Jackknife ter sido maior que o erro aparente, esta é uma estimativa mais robusta em comparação com o resultado sem validação cruzada.

## e)




Matriz de confusão:

| | 0 | 1 |
|---|---|---|
| 0 | 9 | 12 |
| 1 | 0 | 25 |

Proporção de classificações corretas em cada grupo:

| | x |
|---|---|
| 0 | 0.4285714 |
| 1 | 1.0000000 |

Proporção total de classificação correta: 0.7391304

O erro aparente (APER):0.2608696

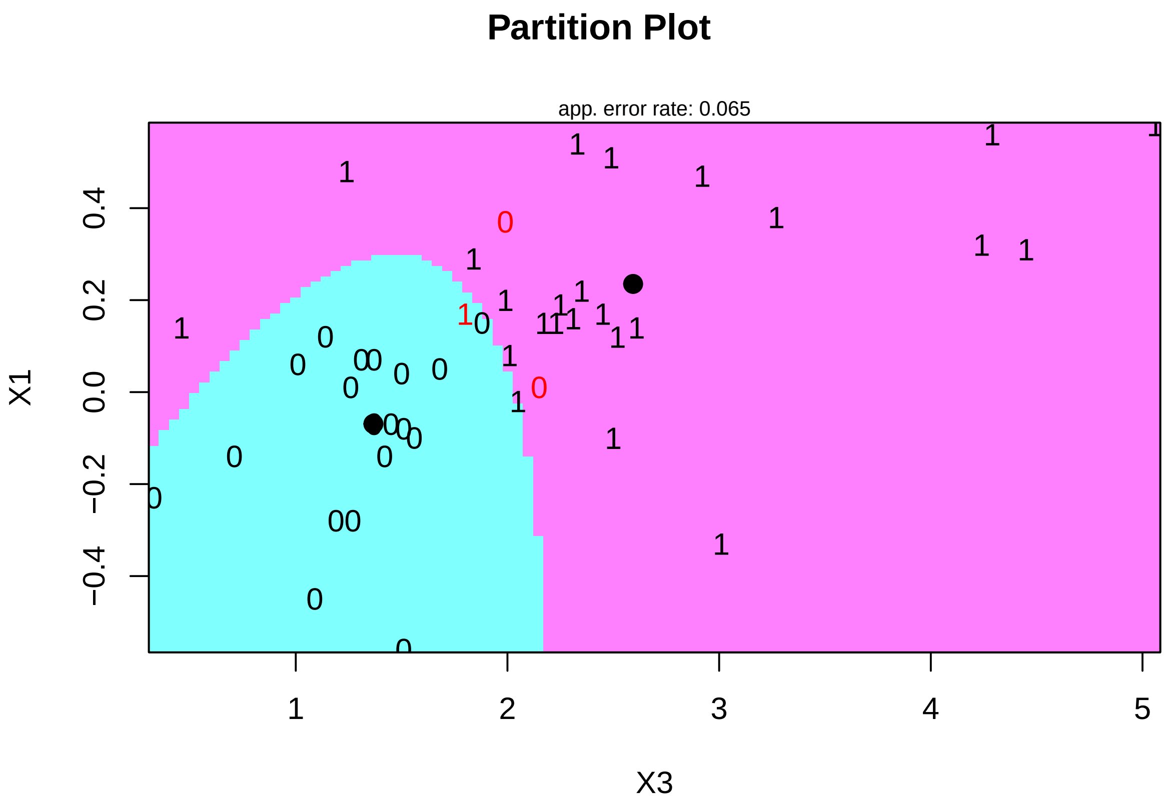
Estimativa da taxa de erro aparente ($\hat{E}(AER)$): 0.2608696

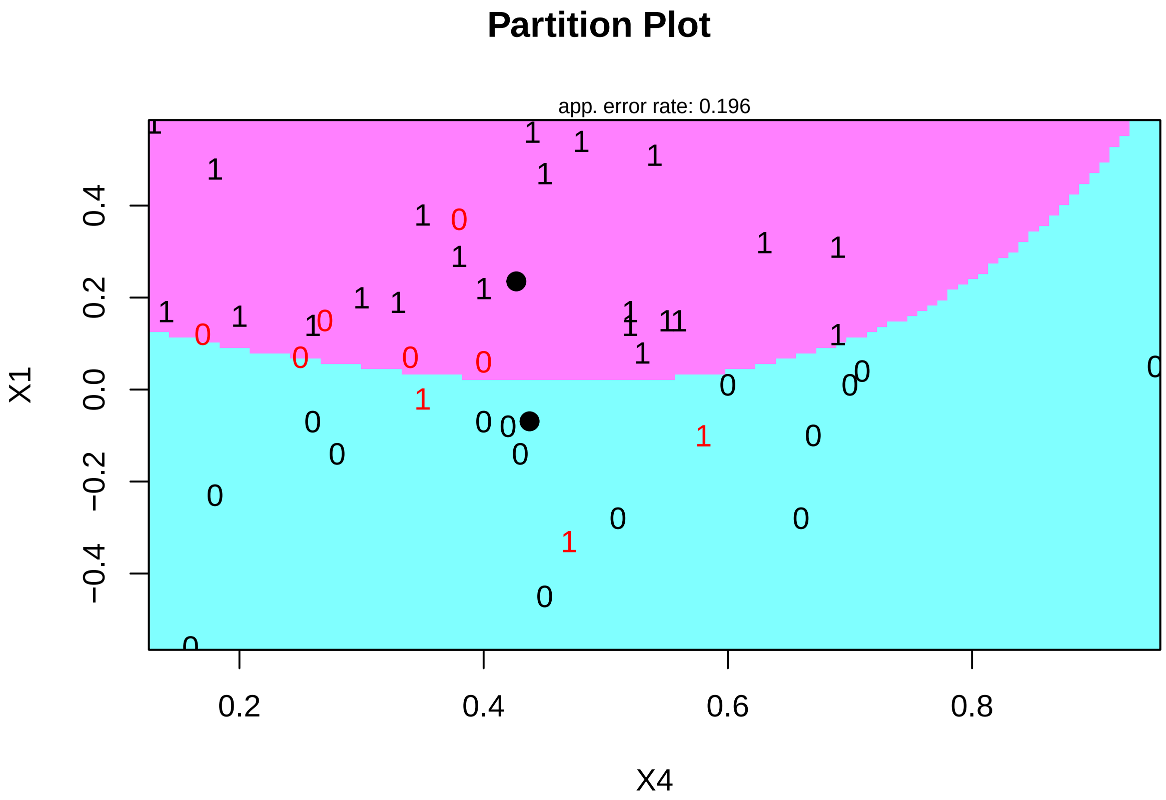
Analisando os APER e $\hat{E}(AER)$, concluímos que as prioris iguais ($p_1 = 0,5; p_2 = 0,5$) tem um erro de classificação inferior se comparado as prioris desiguais ($p_1 = 0,05; p_2 = 0,95$). Neste caso, notamos que tanto o APER quanto o $\hat{E}(AER)$ deram resultados idênticos.

## f)

Como as matrizes $S_1$ e $S_2$ aparentam ser diferentes, esta técnica não aparenta ser a mais adequada. Entretanto, tomando como base apenas a performance do APER = 0.173913, até que a classificação por discriminantes lineares não ficou ruim, com resultados até melhores do que os obtido pelos discriminantes quadráticos.

g)




**Vetores de média e matrizes de covariância para as variáveis (x1,x3):**

Vetor média $\mu_1' =$ -0.0690476, 1.3666667

Vetor Média $\mu_3' = 0.2352, 2.5936$

Matriz de covariância $S_1 =$

| | X1 | X3 |
|---|---|---|
| X1 | 0.0441290 | 0.0344933 |
| X3 | 0.0344933 | 0.1643033 |

Matriz de covariância $S_3 =$

| | X1 | X3 |
|---|---|---|
| X1 | 0.0470510 | 0.0749305 |
| X3 | 0.0749305 | 1.0467740 |

**Análise discriminante quadrática, com prioris = (0,5;0,5), utilizando as variáveis (x1,x3):**

Matriz de confusão:

| | 0 | 1 |
|---|---|---|
| 0 | 19 | 2 |
| 1 | 3 | 22 |

Proporção de classificações corretas em cada grupo:

| | x |
|---|---|
| 0 | 0.9047619 |
| 1 | 0.8800000 |

Proporção total de classificação correta: 0.8913043

Erro aparente (APER):0.1086957

Estimativa da taxa de erro aparente ($\hat{E}(AER)$): 0.1304348

**Análise discriminante quadrática, com prioris = (0,05;0,95), utilizando as variáveis (x1,x3):**

Matriz de confusão:

| | 0 | 1 |
|---|---|---|
| 0 | 4 | 17 |
| 1 | 0 | 25 |

Proporção de classificações corretas em cada grupo:

| | x |
|---|---|
| 0 | 0.1904762 |
| 1 | 1.0000000 |

Proporção total de classificação correta: 0.6304348

Erro aparente (APER):0.3695652

Estimativa da taxa de erro aparente ($\hat{E}(AER)$): 0.3913043

**Vetores de média e matrizes de covariância para as variáveis (x1,x4):**

Vetor média $\mu_1' =$ -0.0690476, 0.437619

Vetor Média $\mu_3' =$ 0.2352, 0.4268

Matriz de covariância $S_1 =$

| | X1 | X4 |
|---|---|---|
| X1 | 0.0441290 | 0.0041474 |
| X4 | 0.0041474 | 0.0445790 |

Matriz de covariância $S_3 =$

| | X1 | X4 |
|---|---|---|
| X1 | 0.0470510 | -0.0067035 |
| X4 | -0.0067035 | 0.0263810 |

**Análise discriminante quadrática, com prioris = (0,5;0,5), utilizando as variáveis (x1,x4):**

Matriz de confusão:

| | 0 | 1 |
|---|---|---|
| 0 | 17 | 4 |
| 1 | 4 | 21 |

Proporção de classificações corretas em cada grupo:

| | x |
|---|---|
| 0 | 0.8095238 |
| 1 | 0.8400000 |

Proporção total de classificação correta: 0.826087

Erro aparente (APER):0.173913

Estimativa da taxa de erro aparente ($\hat{E}(AER)$): 0.2173913

**Análise discriminante quadrática, com prioris = (0,05;0,95), utilizando as variáveis (x1,x4):**

Matriz de confusão:

| | 0 | 1 |
|---|---|---|
| 0 | 3 | 18 |
| 1 | 0 | 25 |

Proporção de classificações corretas em cada grupo:

| | x |
|---|---|
| 0 | 0.1428571 |
| 1 | 1.0000000 |

Proporção total de classificação correta: 0.6086957

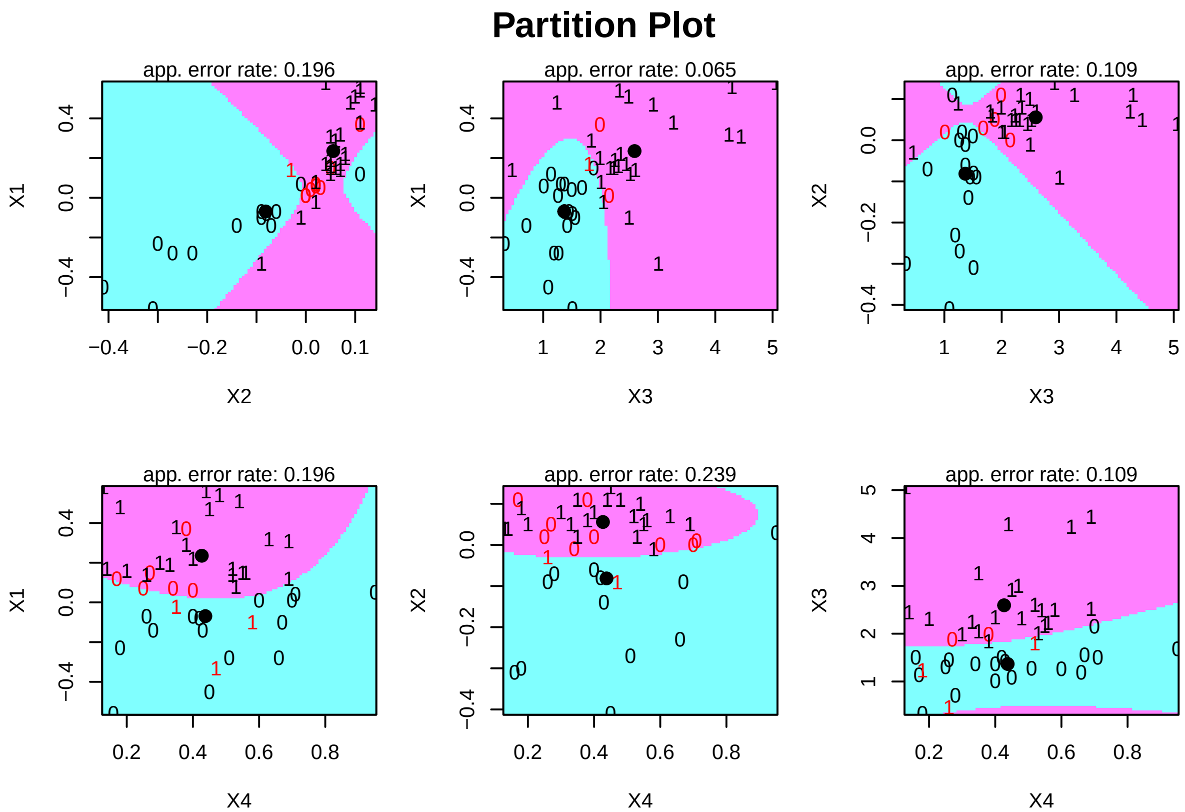
Erro aparente (APER):0.3913043

Estimativa da taxa de erro aparente ($\hat{E}(AER)$): 0.4565217

**Conclusões:**

De fato, os resultados encontrados foram bastante distintos para cada caso. Analisando somente os APER e $\hat{E}(AER)$, notamos que a análise em que foi observado o menor valor de ambos foi a análise executada utilizando as variáveis $(x_1, x_3)$, com prioris iguais $(0,5; 0,5)$, enquanto que os maiores valores foram observados para o modelo em que utilizei as variáveis $(x_1, x_4)$ com prioris desiguais $(0,05; 0,95)$. O modelo que menos variou estas duas estatísticas para ambas as prioris testadas $(0,5; 0,5)$ e $(0,05; 0,95)$ foi o modelo inicialmente testado com as variáveis $(x_1, x_2)$. Com base nisso, podemos concluir que tanto a escolha das variáveis quanto a escolha das prioris, influenciam bastante na qualidade do modelo final.

**h)**




**Vetores de média e matrizes de covariância para as variáveis (x1,x2,x3,x4):**

Vetor média $\mu_1' =$ -0.0690476, -0.0814286, 1.3666667, 0.437619

Vetor Média $\mu_3' =$ 0.2352, 0.0556, 2.5936, 0.4268

Matriz de covariância $S_1 =$

| | X1 | X2 | X3 | X4 |
|---|---|---|---|---|
| X1 | 0.0441290 | 0.0284764 | 0.0344933 | 0.0041474 |
| X2 | 0.0284764 | 0.0210029 | 0.0260200 | 0.0034414 |
| X3 | 0.0344933 | 0.0260200 | 0.1643033 | 0.0327817 |
| X4 | 0.0041474 | 0.0034414 | 0.0327817 | 0.0445790 |

Matriz de covariância $S_3 =$

| | X1 | X2 | X3 | X4 |
|---|---|---|---|---|
| X1 | 0.0470510 | 0.0085072 | 0.0749305 | -0.0067035 |

| | X1 | X2 | X3 | X4 |
|---|---|---|---|---|
| X2 | 0.0085072 | 0.0023757 | 0.0085832 | 0.0001853 |
| X3 | 0.0749305 | 0.0085832 | 1.0467740 | 0.0326328 |
| X4 | -0.0067035 | 0.0001853 | 0.0326328 | 0.0263810 |

**Análise discriminante quadrática, com prioris = (0,5;0,5), utilizando as variáveis (x1,x2,x3,x4):**

Matriz de confusão:

| | 0 | 1 |
|---|---|---|
| 0 | 19 | 2 |
| 1 | 1 | 24 |

Proporção de classificações corretas em cada grupo:

| | x |
|---|---|
| 0 | 0.9047619 |
| 1 | 0.9600000 |

Proporção total de classificação correta: 0.9347826

Erro aparente (APER):0.0652174

Estimativa da taxa de erro aparente ($\hat{E}(AER)$): 0.1086957

**Análise discriminante quadrática, com prioris = (0,05;0,95), utilizando as variáveis (x1,x2,x3,x4):**

Matriz de confusão:

| | 0 | 1 |
|---|---|---|
| 0 | 12 | 9 |
| 1 | 0 | 25 |

Proporção de classificações corretas em cada grupo:

| | x |
|---|---|
| 0 | 0.5714286 |
| 1 | 1.0000000 |

Proporção total de classificação correta: 0.8043478

Erro aparente (APER):0.1956522

Estimativa da taxa de erro aparente ($\hat{E}(AER)$): 0.2391304

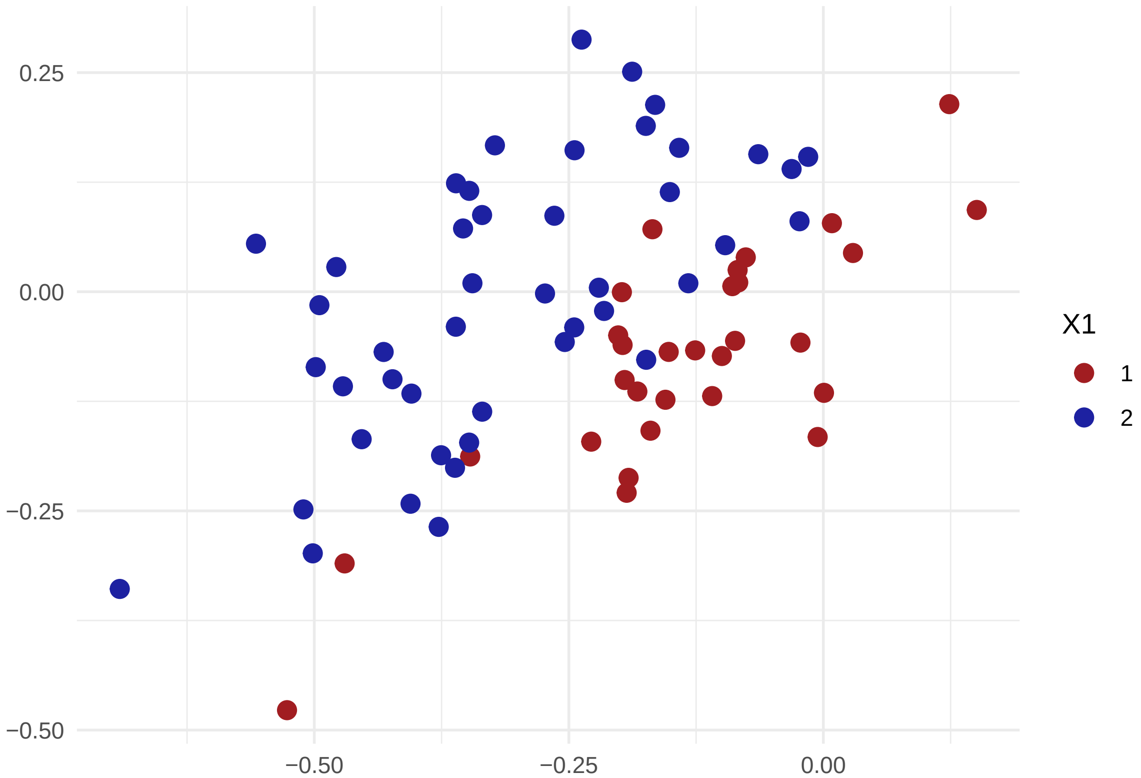
No caso da inclusão de todas as 4 variáveis, o classificador com prioris iguais produziu as melhores classificações (menores APER e $\hat{E}(AER)$). Também neste caso, o classificador com prioris $(0,05; 0,95)$ produziu um APER significativamente maior que o mesmo modelo com prioris iguais, porém foram os menores valores se comparados com os valores observados nos demais modelos com prioris $(0,05; 0,95)$.

**Conclusões:**

Isto nos leva a acreditar que a inclusão de mais variáveis foi bom para o modelo, produzindo os menores erros aparentes. Entretando, a diferença não foi tão substantiva assim, então, deve-se considerar questões como verba para coleta de tantas variáveis, complexidade da análise e viabilidade de novas coletas caso deseje-se seguir com o modelo mais preciso.

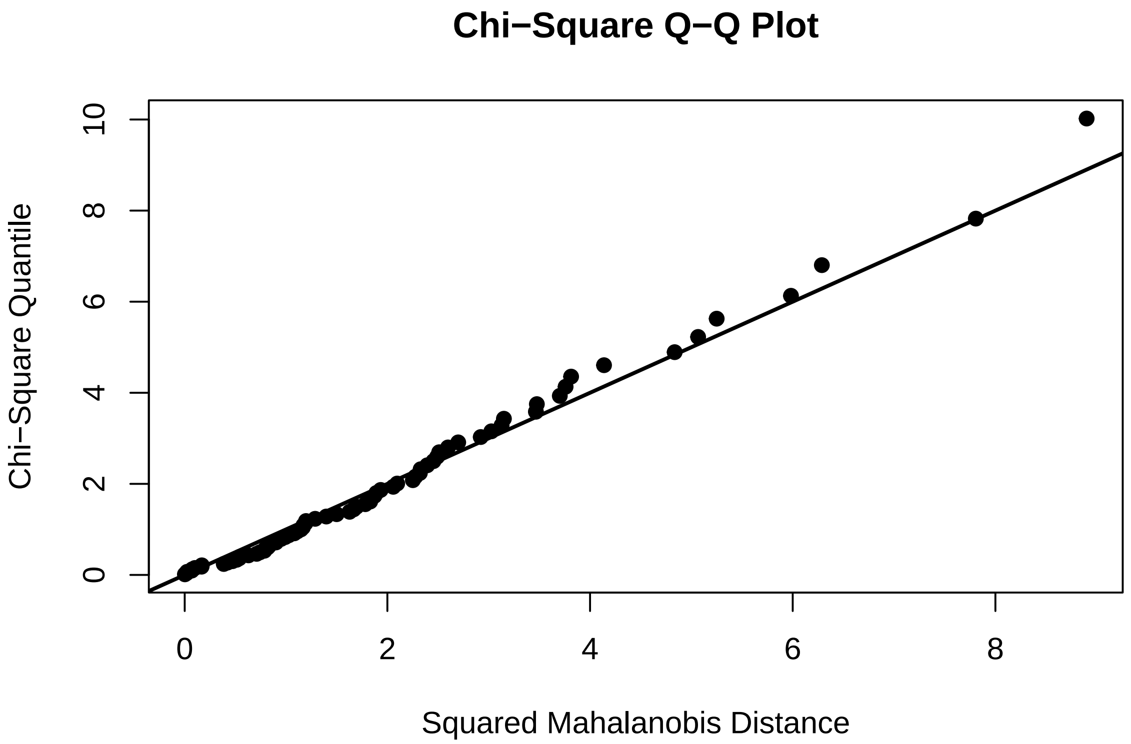
## 90. Johnson e Wichern - Exercício 11.32.

### a)




```
##
##  Shapiro-Wilk normality test
##
## data:  dados$X2
## W = 0.98496, p-value = 0.5185
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dados$X3
## W = 0.99255, p-value = 0.9428
```




```
##             Anderson-Darling test for Multivariate Normality
##
##   data : dados[, 2:3]
##
##   AD              : 0.7583493
##   p-value         : 0.239676
##
##   Result  : Data are multivariate normal (sig.level = 0.05)
```

Através da análise visual, não é possível rejeitar a normalidade bivariada, visto que os pontos aparentam formar uma elipsoide. Foi realizado ainda testes de Shapiro-Wilk nas duas marginais, que também não rejeitaram a normalidade; univariada, neste caso. Foi ainda utilizado o teste de Anderson-Darling para normalidade multivariada do pacote `mvnTest`, que também não rejeitou a normalidade multivariada. Portanto, não temos evidências para rejeitar a hipótese de normalidade multivariada dos dados.

### b)

Matriz de confusão:

| FALSE | TRUE |
|---|---|
| 20 | 1 |
| 12 | 19 |

Proporção de classificações corretas em cada grupo:

| x |
|---|
| 0.9523810 |
| 0.6129032 |

Proporção total de classificação correta: 0.75

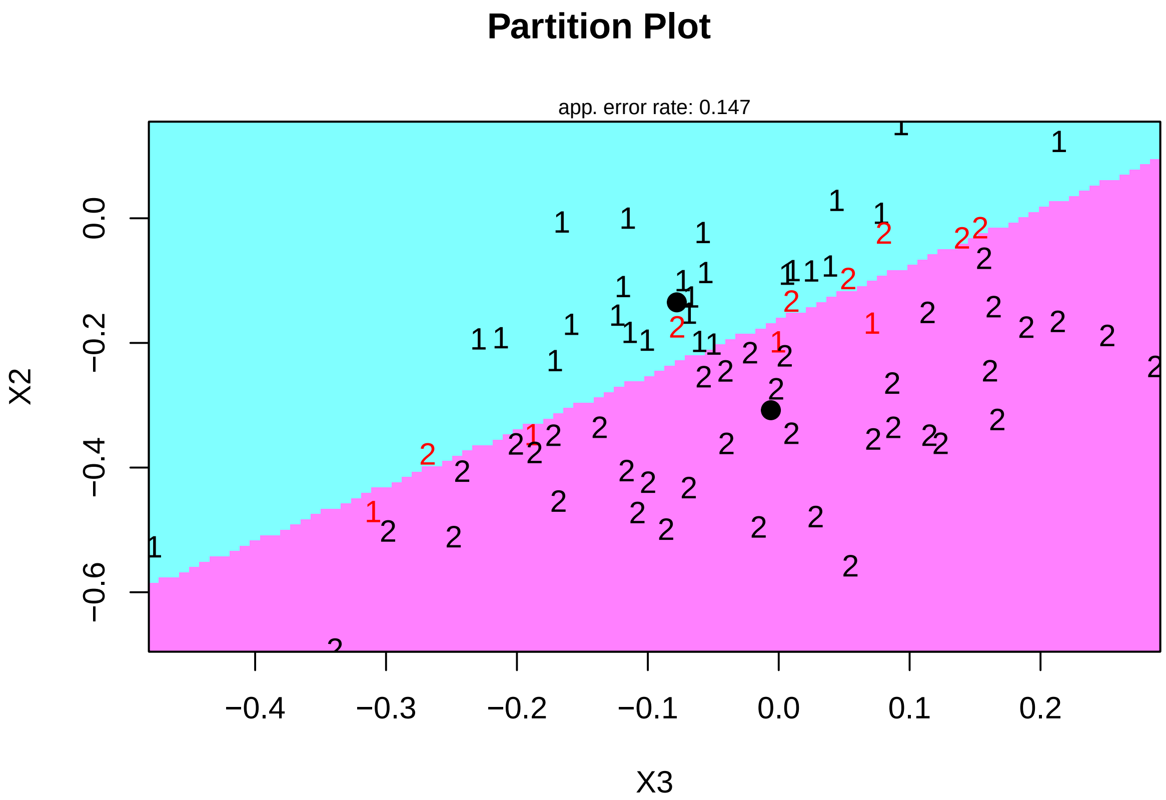
Com isso, temos que a taxa de erro do modelo pontual é de 25%. Esta é relativamente maior do que a encontrada pelos outros métodos de validação utilizados até agora.

**c)**

| pop3 | Freq |
|---|---|
| p1 | 10 |

Todas as 10 novas observações foram classificadas como percentence à população $\pi_1$

**d)**




Matriz de confusão:

| FALSE | TRUE |
|---|---|
| 19 | 2 |
| 5 | 26 |

Proporção de classificações corretas em cada grupo:

| x |
|---|
| 0.9047619 |
| 0.8387097 |

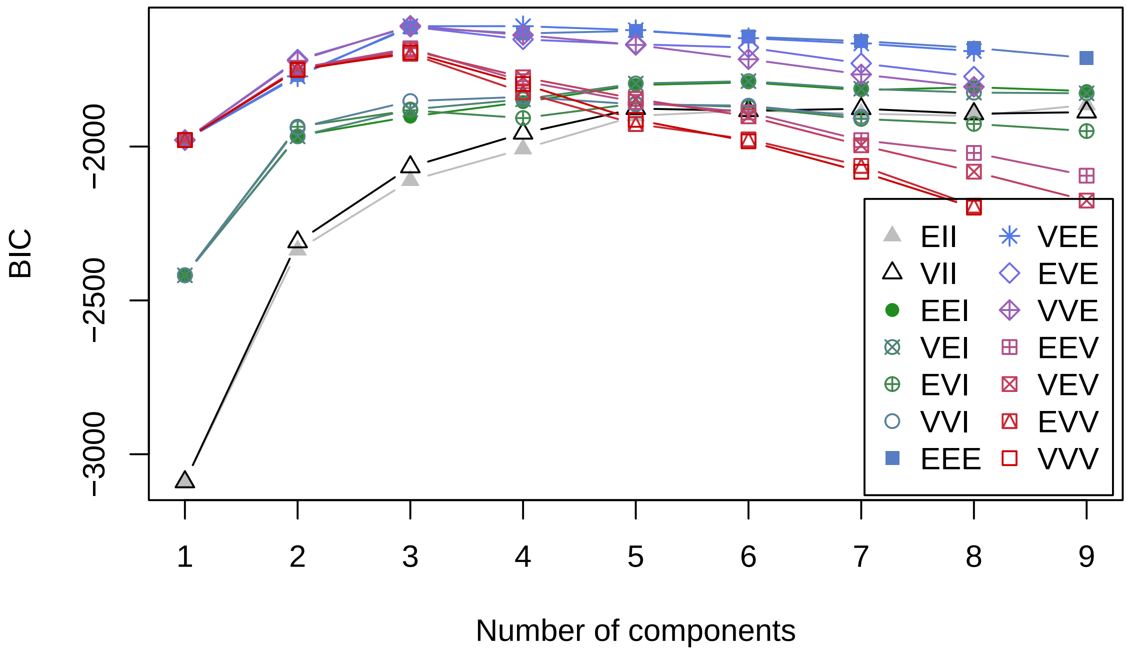
Proporção total de classificação correta: 0.8653846

Com isso, temos que a taxa de erro do modelo pontual é de 13.4615385%. Percebemos que a taxa de erro caiu consideravelmente ao escolher esta outra priori.

Além disso, todas as 10 novas observações foram novamente classificadas como pertencente à população $\pi_1$. Este é um resultado que não impressiona, visto que já haviam sido classificados assim com a priori igual, então era de se esperar que confirmasse este resultado com uma priori maior para a população 1.

## 91.

```
## Best BIC values:
##              VVE,3        VEE,4        VEE,3
## BIC      -1607.574 -1608.767736 -1608.793746
## BIC diff     0.000    -1.194096    -1.220106
```




```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VVE (ellipsoidal, equal orientation) model with 3 components:
##
##  log-likelihood   n df       BIC      ICL
##       -663.3814 200 53 -1607.574 -1607.71
##
## Clustering table:
##  1  2  3
## 18 98 84
```

```
##
## Class         1  2  3
##   Genuína     2 98  0
##   Falsificada 16  0 84
```

A mistura de normais não operou tão bem quanto os discriminantes lineares e quadráticos. Enquanto nesses dois, 199 das 200 notas foram classificadas corretamente, o algorítmo de mistura de normais encontrou m=3 como o número ideal de clusters (sendo que neste caso sabemos que há apenas dois:

genuínas e falsificadas). Com isso, classificou corretamente 182 das 200 notas. Interessante notar que não houve classificação de notas falsas como notas genuínas ou vice-versa; e sim algumas notas desses dois grupos foram classificadas em outro cluster, que seria talvez um cluster de "confusão", ou seja, notas em que não estava claro o suficiente se eram genuínas ou classificadas.

Erro aparente (APER) do modelo de discriminante linear: 0.005

Erro aparente (APER) do modelo de discriminante quadrático: 0.005

Erro aparente (APER) do modelo de mistura de normais: 0.09

Índice de Rand ajustado do modelo de mistura de normais: 0.8418856