Introdução a Métodos Estatísticos para a Bioinformática

Profa. Júlia Maria Pavan Soler

pavan@ime.usp.br

IBI 5086 – Bioinformática - IME/USP

2º Sem/2020

Programa

- ➢ Álgebra linear básica: cálculo matricial, determinantes, sistemas lineares, produto interno, norma, ortogonalidade, autovalores e autovetores
- ✓ Estrutura de Dados: variáveis (resposta, explicativa), unidades amostrais e experimentais

- ✓ 1.1. Comparação de 2 ou mais grupos: Testes Clássicos (teste t, Wilcoxon, ANOVA), Testes de Aleatorização, Comparações Múltiplas, Efeitos Genéticos
- ✓ 1.2. Análise de Tabelas de Contingência: Testes Qui-Quadrado, Regressão Logística.

2. **Análise Multivariada de Dados**: Componentes Principais, Análise Discriminante e Classificação, Correlação Canônica, modelos MANOVA

3. Simulação de Monte Carlo, Intervalos de Confiança Bootstrap

Redução de Dimensionalidade em $\Re^p$

	Variáveis					
Unidades Amostrais	1	2	…	j	…	p
1	Y_{11}	Y_{12}		Y_{1j}		Y_{1p}
2	Y_{21}	Y_{22}		Y_{2j}		Y_{2p}
…	…	…	…	…		…
i	Y_{i1}	Y_{i2}		Y_{ij}		Y_{ip}
…	…	…	…	…	…	…
n	Y_{n1}	Y_{n2}		Y_{nj}		Y_{np}

$$Y_{n\times p};\quad n > p \qquad \Re^p \rightarrow \Re^m, m < p$$

Redução de Dimensionalidade $\Rightarrow$ obter ***m*** vetores (m<p) que são combinações lineares das ***p*** variáveis originais e atendem a critérios de otimalidade

$Y_{n\times p}$ ***p*** vetores das respostas para ***n*** indivíduos $\Rightarrow$ Vetores de **Escores** para os n indivíduos (CP)

Vetores de **Cargas** (pesos) às ***p*** variáveis

Ténicas Multivariadas de Redução de Dimensionalidade

Como obter vetores reducionistas de dados?

Depende:
- Estrutura dos Dados
- Objetivo da análise

- ✓ Análise de Componentes Principais: $Y_{n\times p} \Rightarrow \Re^{p\times p}$
- ✓ Escalonamento Multidimensional: $Y_{n\times p} \Rightarrow D^{n\times n}$
- ✓ Análise de Correspondência: $Y_{n\times p} \Rightarrow [0,1]^{I\times J}$

Análises não supervisionadas

- Análise Discriminante $Y_{n\times(p+1)} \Rightarrow \Re^{p\times p}$ $\Rightarrow$ MANOVA
- Análise de Agrupamento: $Y_{n\times p} \Rightarrow D^{n\times n}$
- Análise de Correlação Canônica: $Y_{n\times(p+q)} \Rightarrow \Re^{p\times q}(\Re^{p\times p}, \Re^{q\times q})$

Análise supervisionada

Análise Multivariada - Estatísticas Descritivas

Projeto ACTN3 e Força (dados de atletas)

n = 95 p= (5+1)

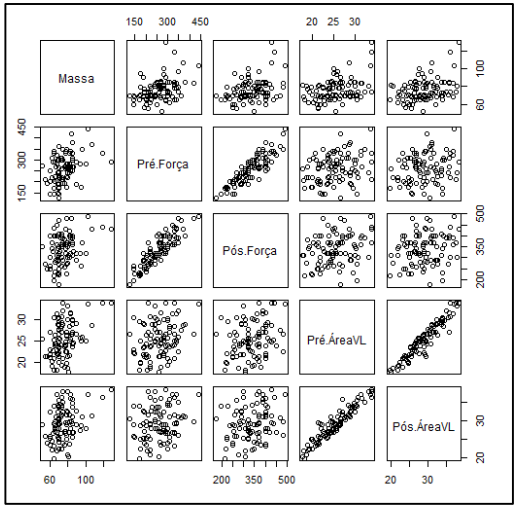
	Massa	ACTN3	Pré.Força	Pós.Força	Pré.ÁreaVL	Pós.ÁreaVL
1	73.0	1	210	260	27.133	31.398
2	78.0	1	260	320	23.841	26.950
3	70.0	3	220	320	23.755	28.937
...						
94	95.0	2	175	260	25.120	28.605
95	71.0	2	220	330	25.452	29.029

Vetor centróide

Massa	Pré.Força	Pós.Força	Pré.ÁreaVL	Pós.ÁreaVL
75.97	261.79	337.58	25.48	29.13

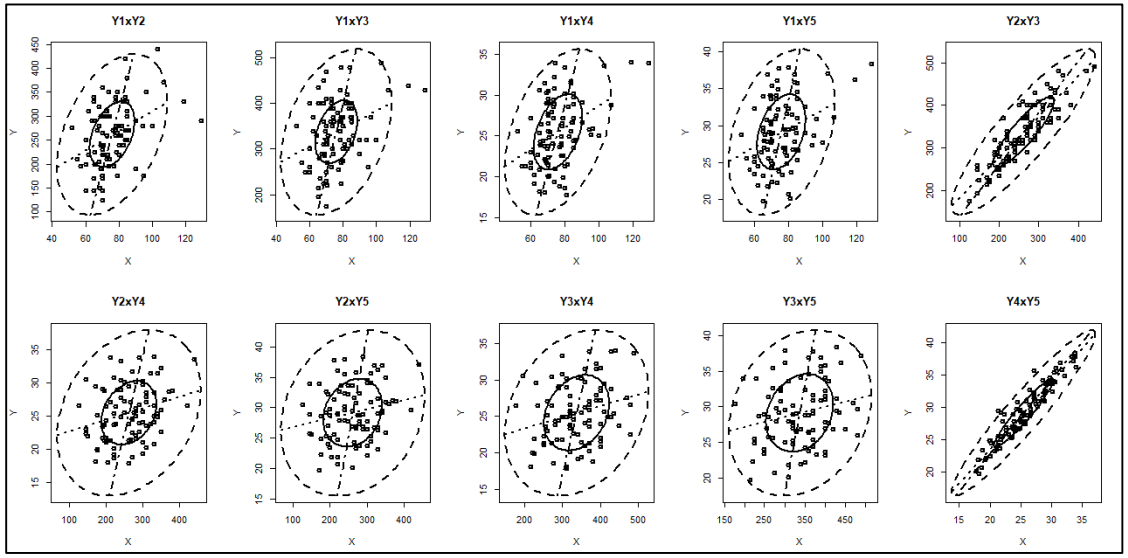
Matriz de Covariância (superior) e **Correlação** (inferior)

	Massa	Pré.Força	Pós.Força	Pré.ÁreaVL	Pós.ÁreaVL
Massa	**155.24**	317.53	303.47	20.80	19.37
Pré.Força	0.40	**4032.93**	3815.02	70.70	59.41
Pós.Força	0.36	0.88	**4640.35**	71.05	65.73
Pré.ÁreaVL	0.41	0.27	0.26	**16.40**	17.10
Pós.ÁreaVL	0.35	0.21	0.22	0.95	**19.85**

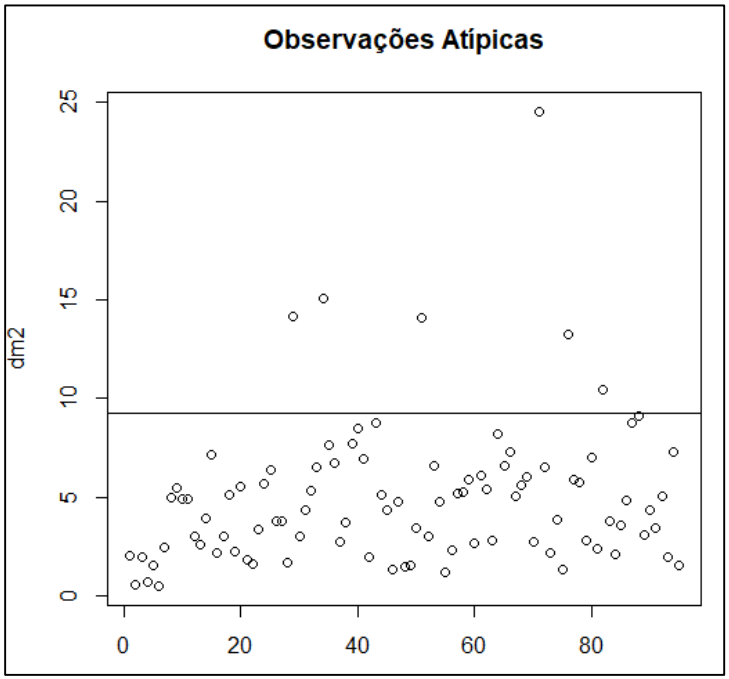


Análise Multivariada – Observações Atípicas

Box-plot bivariado: observações atípicas (p=2)



Distância de Manhattan (p=5)



$\chi^2(p=5\ ,\ 90\%) = 9.24$

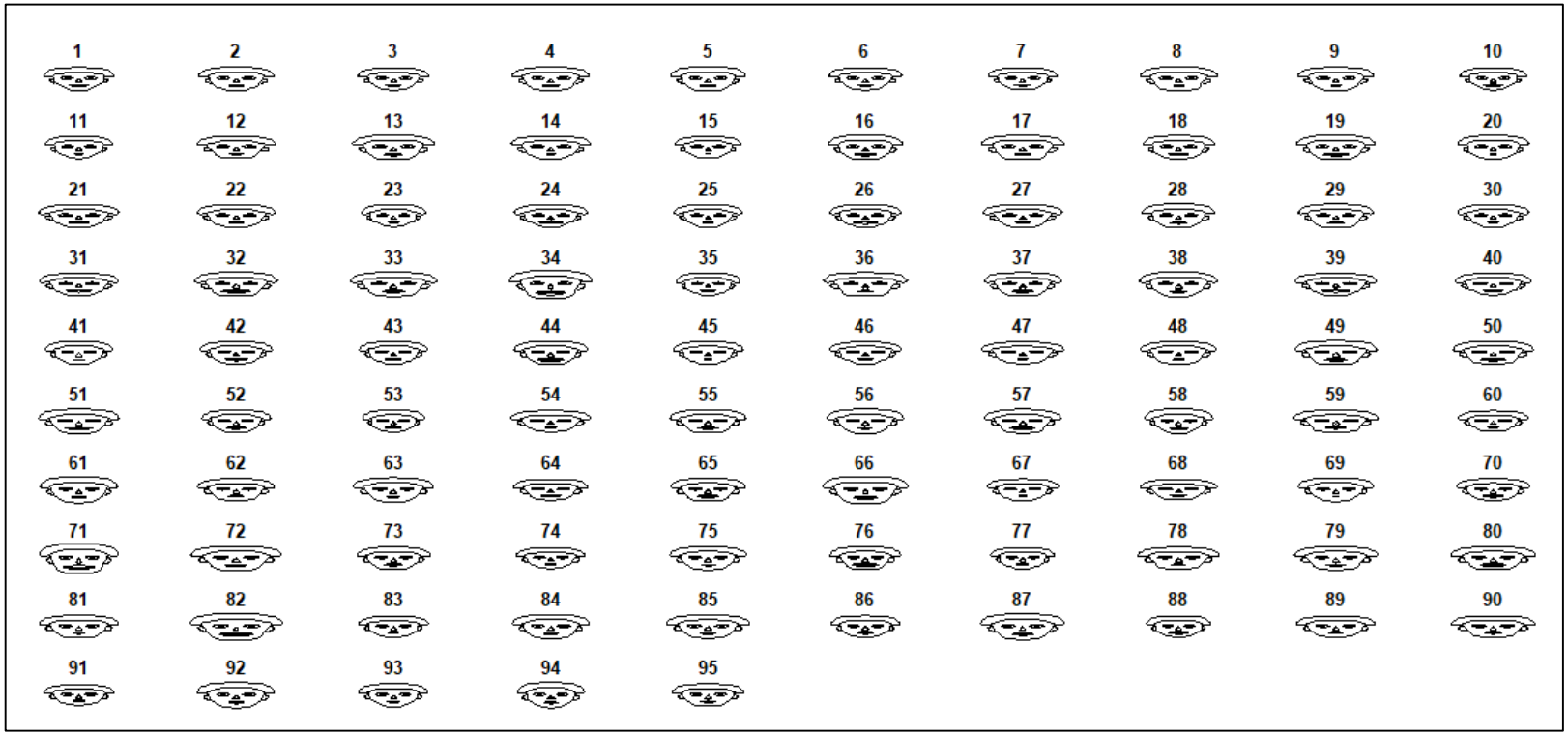
Escore Z (Atletas atípicos)

	dm2	Massa	Pré.Força	Pós.Força	Pré.ÁreaVL	Pós.ÁreaVL
82	10.42	2.17	2.81	2.24	2.01	1.80
76	13.25	-0.14	-0.82	0.48	2.08	1.95
51	14.06	0.73	0.76	0.77	0.41	1.48
29	14.16	0.32	-0.03	0.48	-1.03	0.07
34	15.06	3.45	1.07	1.50	2.11	1.60
71	24.54	4.26	0.44	1.36	2.07	2.08

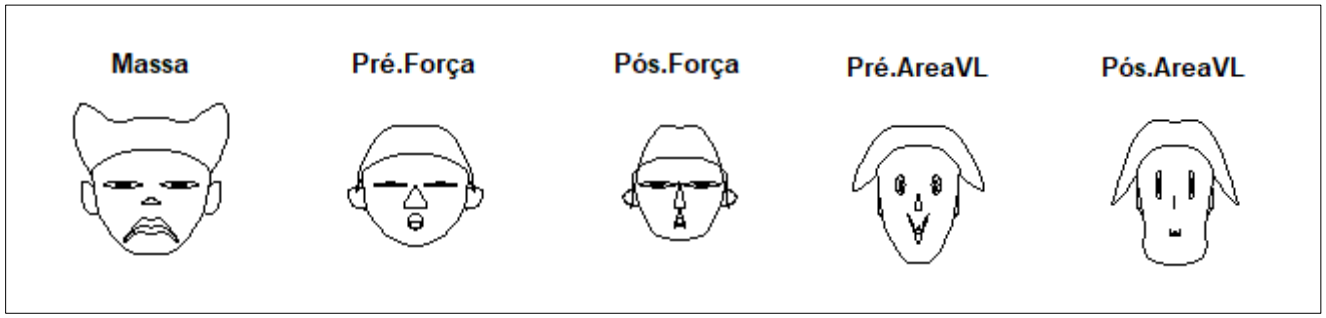
Análise Multivariada – Observações e Variáveis

Projeto ACTN3 e Força

Atletas



Variáveis



Componentes Principais – Coordenadas Principais

S_{pxp} D_{nxn}

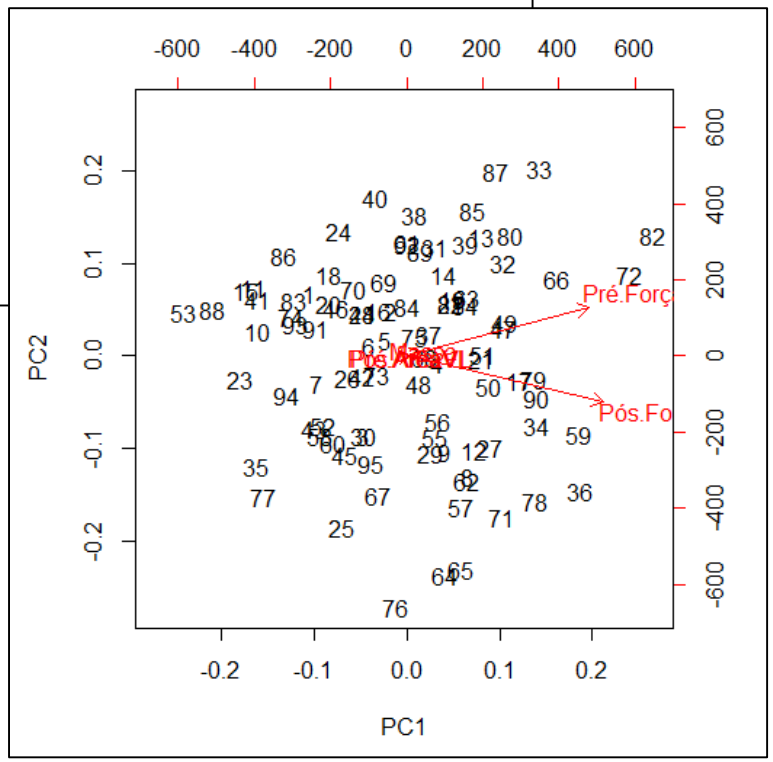
Projeto ACTN3 e Força: Análise dos Dados originais (S)

Autovalores (importância dos CP):

	PC1	PC2	PC3	PC4	PC5
Standard deviation	90.4980	22.61788	11.56229	5.36349	0.9289
Proportion of Variance	0.9239	0.05771	0.01508	0.00325	0.0001
Cumulative Proportion	0.9239	**0.98158**	0.99666	0.99990	1.0000

Autovetores (cargas)

	PC1	**PC2**	PC3	PC4	PC5
Massa	0.05	0.07	0.97	-0.20	0.01
Pré.Força	0.68	0.73	-0.09	0.00	0.01
Pós.Força	0.73	-0.68	0.01	-0.01	0.00
Pré.ÁreaVL	0.01	0.01	0.14	0.64	-0.75
Pós.ÁreaVL	0.01	0.00	0.14	0.74	0.66



Biplot: representação simultânea dos atletas e das variáveis

As variáveis de Força são as que mais contribuem para a redução de dimensionalidade.

Componentes Principais – Coordenadas Principais

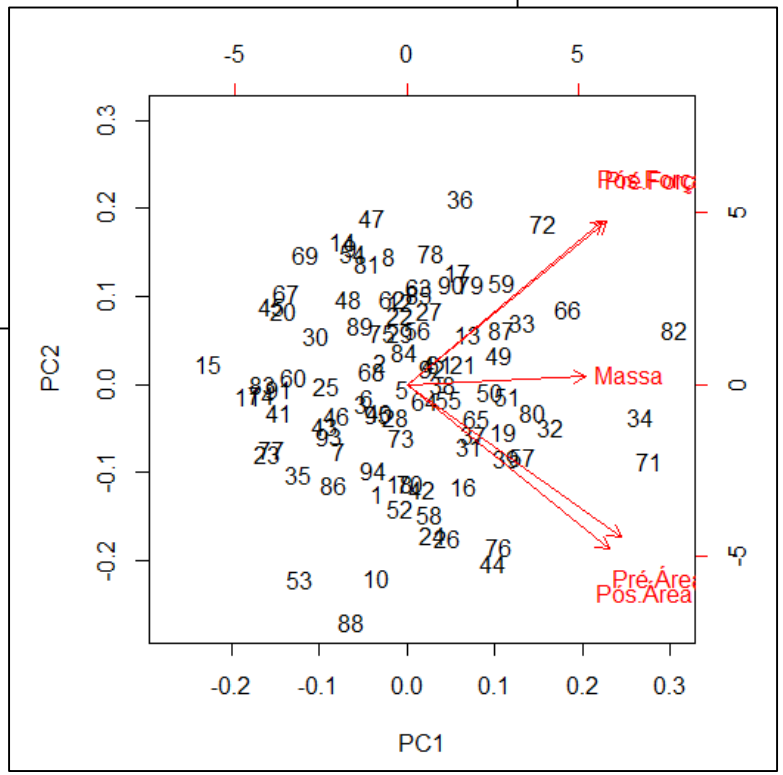
Projeto ACTN3 e Força: Análise dos Dados Normalizados (R)

Autovalores (importância dos CP):

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.6524	1.1983	0.8173	0.34393	0.21785
Proportion of Variance	0.5461	0.2872	0.1336	0.02366	0.00949
Cumulative Proportion	0.5461	**0.8333**	0.9668	0.99051	1.00000

Autovetores (cargas)

	PC1	**PC2**	PC3	PC4	PC5
Massa	0.40	0.02	0.91	-0.05	0.04
Pré.Força	0.45	0.51	-0.18	0.70	0.12
Pós.Força	0.44	0.51	-0.25	-0.69	-0.09
Pré.ÁreaVL	0.48	-0.47	-0.16	0.11	-0.71
Pós.ÁreaVL	0.46	-0.51	-0.22	-0.10	0.69



Biplot: representação dos atletas e das variáveis

Todas as variáveis contribuem para a redução de dimensionalidade

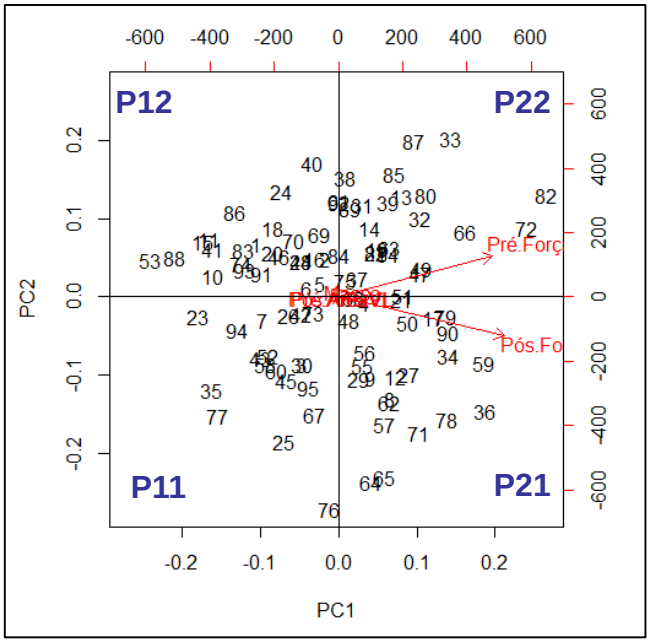
Análise de Correspondência

Representação gráfica das variáveis em uma tabela de contingência

Projeto ACTN3 e Força

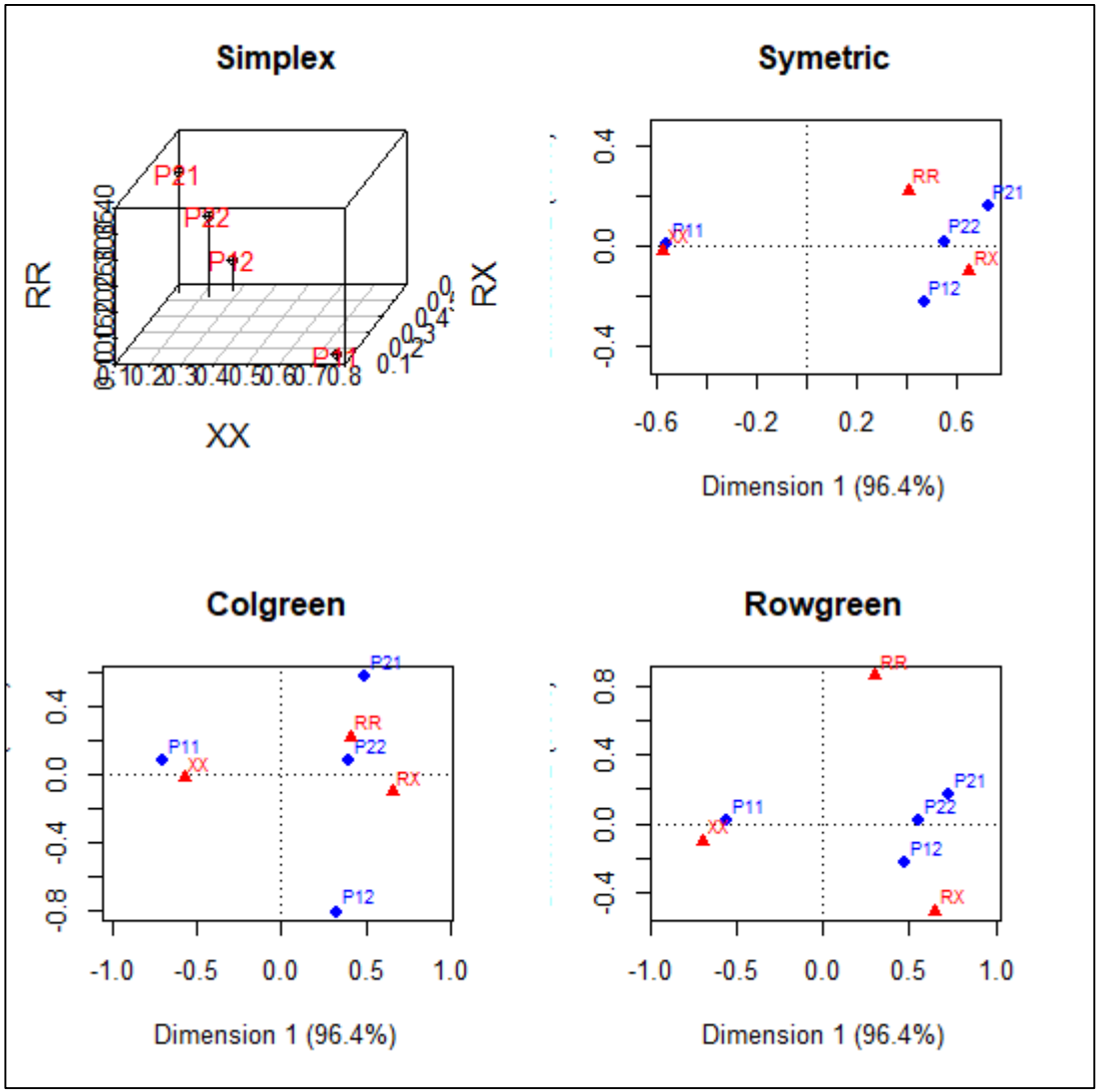
Biplot – S

Categorização das Variáveis



Distribuição dos atletas

	XX	RX	RR
P11	1	9	9
P12	7	14	4
P21	3	13	8
P22	6	14	7



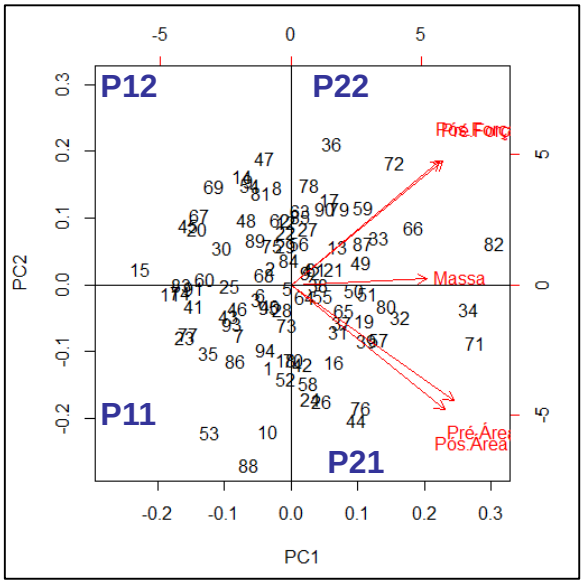
X-squared=53.04, df=6, p-value=1.151e-09

Análise de Correspondência

Representação gráfica das variáveis em uma tabela de contingência

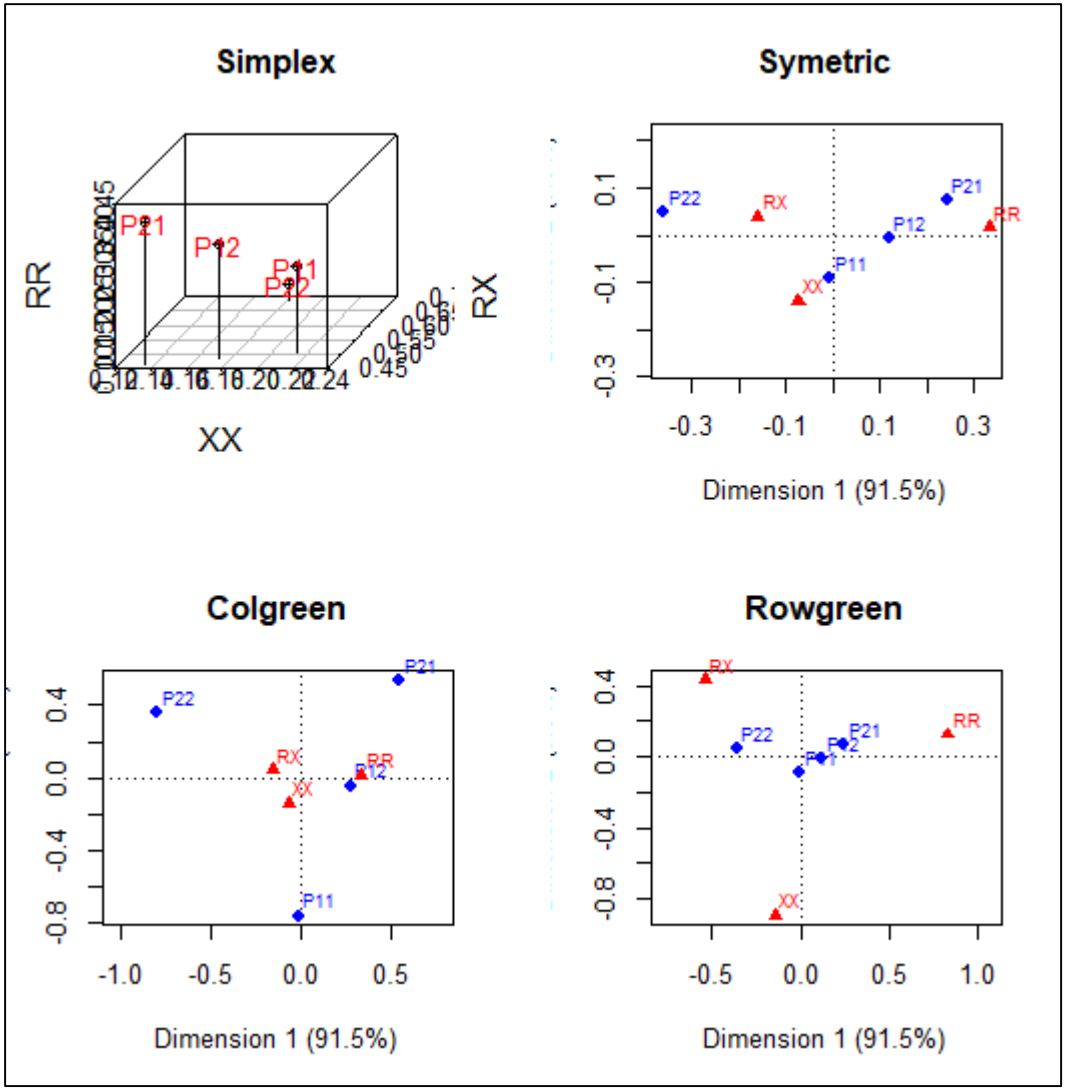
Projeto ACTN3 e Força

Biplot – R
Categorização



Distribuição dos atletas

	XX	RX	RR
P11	6	14	8
P12	4	11	8
P21	3	10	9
P22	4	15	3



X-squared=4.86, df=6, p-value=0.5616

Análise Discriminante

Análise Clássica

n > p

Observações iid ***agrupadas***

(respostas quantitativas)

Técnicas de Redução de Dimensionalidade
Análise Discriminante

Análises Não-Supervisionadas

- ✓ Componentes Principais
- ✓ Coordenadas Principais
- ✓ An. de Correspondência

$$Y_{n\times p};\quad Y_{i_{p\times 1}} \overset{iid}{\sim} (\mu;\Sigma);\quad i=1,...,n$$

$$\Re^p \rightarrow \Re^m;\quad m<\min(n,p);\quad n>p$$

Análise Discriminante: $Y_{n\times p};\quad n=\sum_{g=1}^{G} n_g;\quad Y_{gi_{p\times 1}} \overset{iid}{\sim} (\mu_g;\Sigma_g),\ g=1,2,...,G$

$$\Re^p \rightarrow \Re^m;\quad m<p$$

Análise Supervisionada

- Populações Estratificadas (G grupos): G=2 e G>2

  $\Rightarrow$ Solução de Fisher

  - Solução Probabilística (Regra Discriminante de Bayes)

População Estratificada

τ_1 τ_2 ... τ_G

$$\Rightarrow Y_i \mid \tau_g = \left(Y_{i1}, Y_{i2}, \ldots Y_{ip}\right)' \in \Re^p; \quad g = 1, \ldots, G$$

$$E\left(Y_i \mid \tau_1\right) = \mu_{1(p\times 1)} \qquad E\left(Y_i \mid \tau_2\right) = \mu_{2(p\times 1)} \qquad \ldots \qquad E\left(Y_i \mid \tau_G\right) = \mu_{G(p\times 1)}$$

$$Cov\left(Y_i \mid \tau_1\right) = \Sigma_{1(p\times p)} \qquad Cov\left(Y_i \mid \tau_2\right) = \Sigma_{2(p\times p)} \qquad \ldots \qquad Cov\left(Y_i \mid \tau_G\right) = \Sigma_{G(p\times p)}$$

n_1 n_2 n_G

Amostra

Grupos	Unidades amostrais	Variáveis					
		1	2	…	j	…	p
1	1	Y_{11}	Y_{12}		Y_{1j}		Y_{1p}
	2	Y_{21}	Y_{22}		Y_{2j}		Y_{2p}
	…	…	…	…	…		…
…	i	Y_{i1}	Y_{i2}		Y_{ij}		Y_{ip}
	…	…	…	…	…	…	…
G	n	Y_{n1}	Y_{n2}		Y_{nj}		Y_{np}

Matriz de Dados

$$Y_{n\times p} = \begin{pmatrix} Y_1 \\ \ldots \\ Y_G \end{pmatrix}$$

$$Y_{g(n_g \times p)}$$

$$n = \sum_{g=1}^{G} n_g$$

Objetivos da ANÁLISE DISCRIMINANTE

- Obter Funções Discriminantes das “p” variáveis

  - Redução de dimensionalidade: $\Re^p \rightarrow \Re^m$; $m \leq \min[n, p, (G-1)]$

  - Classificação de “novas” observações (predição de grupos)

Análise Discriminante - Motivação

Projeto ACTN3 e Força (dados de atletas)

n = 95 p= (5+1)

	Massa	ACTN3	Pré.Força	Pós.Força	Pré.ÁreaVL	Pós.ÁreaVL
1	73.0	1	210	260	27.133	31.398
2	78.0	1	260	320	23.841	26.950
3	70.0	3	220	320	23.755	28.937
...						
94	95.0	2	175	260	25.120	28.605
95	71.0	2	220	330	25.452	29.029

Atletas agrupados segundo
o genótipo em ACTN3

Obter uma função das variáveis que melhor faça a predição dos Grupos $\Rightarrow$ seleção de variáveis mais discriminatórias.

Análise Discriminante - Motivação

1. Medidas biométricas (mm) de Pardais fêmea

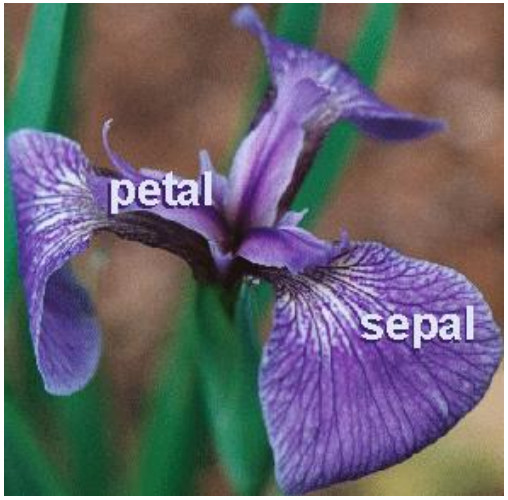
(Manly, 2005; Hermon Bumps, 1898).

Pardal	Sobrev.	X1	X2	X3	X4	X5
1	S	156	245	31.6	18.5	20.5
…	…					
21	S	159	236	31.5	18.0	21.5
22	N	155	240	31.4	18.0	20.7
…	…					
49	N	164	248	32.3	18.8	20.9

Como as características corporais dos pardais pode ser usada para predizer os grupos de Sobreviventes e Não-Sobreviventes?

Análise Discriminante - Motivação

Dados “Iris” do R (Fisher, RA, 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, Part II: 179–188)



Medidas do comprimento e largura da pétala e sépala de 50 flores de íris de cada uma de três espécies (setosa, versicolor e virginica).

$$Y_{150\times4} = \begin{pmatrix} Y_{G=1\ 50\times4} \\ Y_{G=2\ 50\times4} \\ Y_{G=3\ 50\times4} \end{pmatrix}$$

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
...					
50	5.0	3.3	1.4	0.2	setosa
51	7.0	3.2	4.7	1.4	versicolor
...					
100	5.7	2.8	4.1	1.3	versicolor
101	6.3	3.3	6.0	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica
...					
150	5.9	3.0	5.1	1.8	virginica

Análise Discriminante - Motivação

⇒ Dados do Transcriptoma: A expressão de “genes” pode ser usada para predizer (caracterizar) diferentes tecidos tumorais (cancer)?

Irizarry, R.A. and Love, M.I.
Data Analysis for the Life Sciences, 2015.

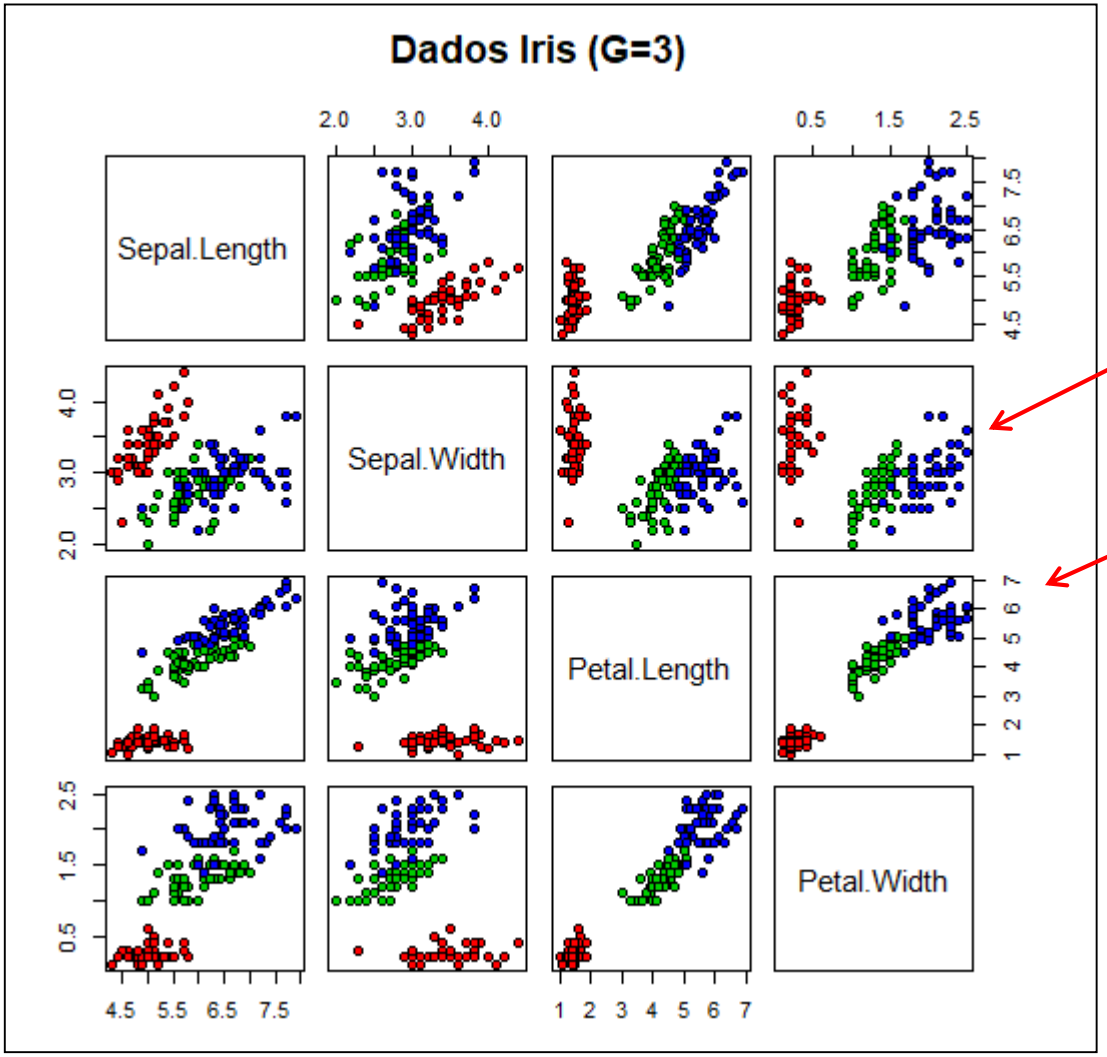
$$Y_{189\times 22.215} = \begin{pmatrix} Y_1' \\ \dots \\ Y_{189}' \end{pmatrix} = \left(Y_{(1)},\dots,Y_{(22.215)}\right)$$

```
library(devtools)
install_github("genomicsclass/tissuesGeneExpression")

library(tissuesGeneExpression)
data(tissuesGeneExpression)
dim(e)  ## e contains the expression data
## [1] 22215   189

table(tissue) ##tissue[i] tells us what tissue is represented by e[,i]
## tissue
## cerebellum colon endometrium hippocampus kidney liver placenta
##         38    34         15          31     39    26         6
```

p=22.215

n=189

Análise Discriminante



Considerando a dispersão das variáveis:

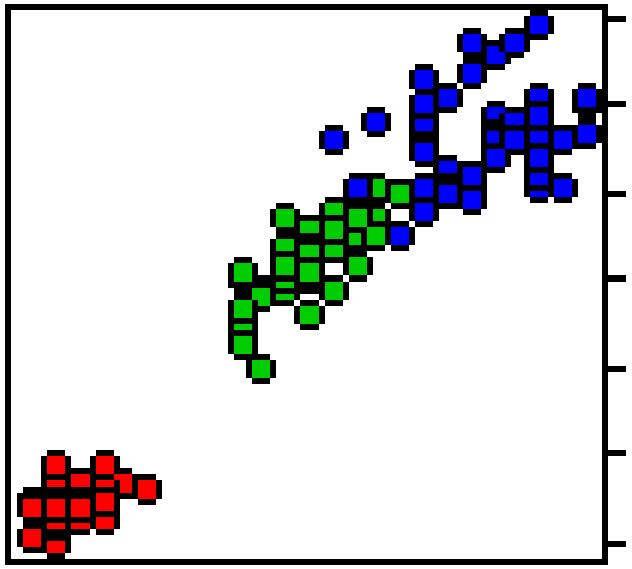
Sepal.Width x Petal.Width

Petal.Length x Petal.Width

Qual seria uma direção “otima” (função linear) para a discriminação das espécies?

Análise Discriminante

Dados Iris – G=3
Pepal.Length x Petal.Width



Dados Iris – G=3
Sepal.Width x Petal.Width

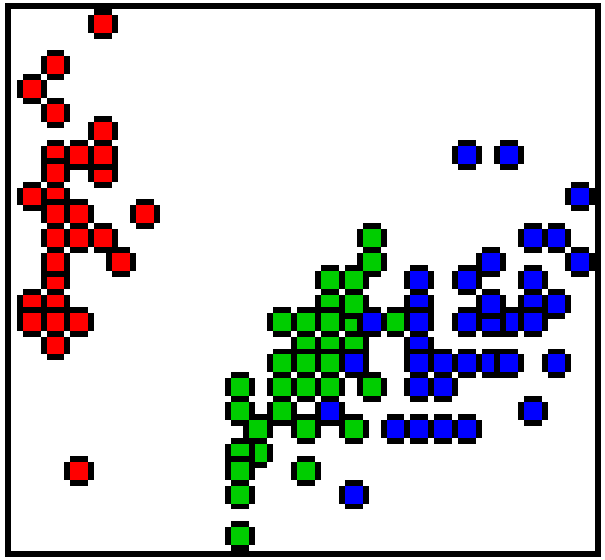
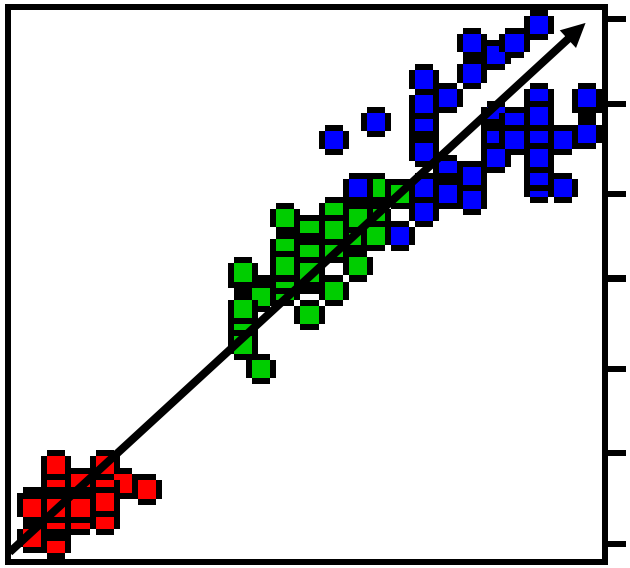


Gráfico de dispersão das observações (em $\Re^2$).

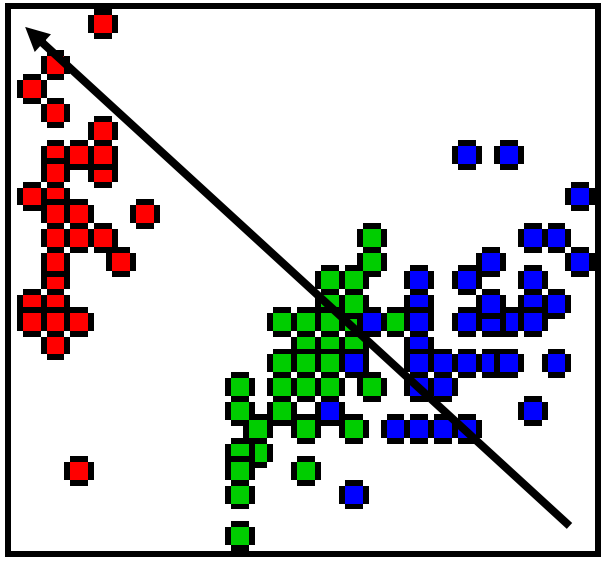
Indique uma direção (terceiro eixo, $l'Y$) que defina uma função discriminante linear de separação dos grupos!

Análise Discriminante

Dados Iris – G=3
Pepal.Length x Petal.Width



Dados Iris – G=3
Sepal.Width x Petal.Width



Indicação de um terceiro eixo, $l'Y$ (em preto), que define uma função discriminante linear de separação entre os grupos.

Como obter essa direção discriminante?

$$Y_i \in \Re^p; \quad Y_i = \left(Y_{i1}, Y_{i2}, ..., Y_{ip}\right) \quad \rightarrow \quad X = l' Y_i = l_1 Y_{i1} + l_2 Y_{i2} + ... + l_p Y_{ip}$$

obter Cargas e os correspondentes Escores

Função Discriminante Linear de Fisher

Formulação de Fisher – Caso de 2 Populações

Considere uma População constituída por observações multivariadas (quantitativas) e estratificada em dois subgrupos, tal que:

O grupo ao qual a observação pertence é conhecido.

$$Y_{n\times p} = \begin{bmatrix} Y_{1(n_1\times p)} \\ Y_{2(n_2\times p)} \end{bmatrix} \Rightarrow \begin{cases} Y_i \in \Re^p; \quad E(Y_i \mid \tau_1) = \mu_{1(p\times 1)} \quad Cov(Y_i \mid \tau_1) = \Sigma_{1(p\times p)} \\ \qquad\qquad E(Y_i \mid \tau_2) = \mu_{2(p\times 1)} \quad Cov(Y_i \mid \tau_2) = \Sigma_{2(p\times p)} \end{cases}$$

Suposição $\Rightarrow \quad \Sigma_1 = \Sigma_2 = \Sigma$

Matrizes de covariâncias homogêneas

Para G=2: Proposta de Fisher é obter combinações lineares das "p" variáveis que maximizem a distância entre os centróides dos grupos na função discriminante :

$$Y_i \in \Re^p \rightarrow X = l'\,Y_i; \qquad l = \arg\max_{l;=l'Y} \frac{(\mu_{X1} - \mu_{X2})^2}{\sigma_X^2} = \arg\max_l \frac{(l'\mu_1 - l'\mu_2)^2}{l'\Sigma l}$$

Solução ao problema de otimização:

$$l_{p\times 1} = \Sigma^{-1}(\mu_1 - \mu_2); \qquad X_i = l'\,Y_i = (\mu_1 - \mu_2)'\,\Sigma^{-1}\,Y_i$$

Cargas **Escores**

Função Discriminante Linear de Fisher
Estimação

Suposição: Obs. independentes, Matriz de covariâncias homogêneas, prioris iguais.

$$Y_{n\times p} = \begin{bmatrix} Y_{1(n_1\times p)} \\ Y_{2(n_2\times p)} \end{bmatrix} \Rightarrow \qquad Y_i \in \Re^p \rightarrow X_i = l'\,Y_i = \left(\mu_1 - \mu_2\right)' \Sigma^{-1} Y_i$$

Para dados amostrais: Adotar estimadores "apropriados" de $\hat{\mu}_1,\ \hat{\mu}_2,\ \hat{\Sigma}$

$$X_i = l'Y_i = \left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1} Y_i$$

Função discriminante

$\bar{Y}_1 \quad \bar{Y}_2 \quad S_c$

$$S_{c_{p\times p}} = \frac{\left(n_1 - 1\right)S_1 + \left(n_2 - 1\right)S_2}{n_1 + n_2 - 2}$$

Matriz de covariância comum

Regra de Classificação Amostral: Alocação de uma nova observação aos Grupos

$$Y_0 = \left(Y_{01}, Y_{02}, \ldots Y_{0p}\right)? \rightarrow \qquad X_0 = l'Y_0 \qquad \begin{cases} X_0 \geq c \Rightarrow Y_0 \in \tau_1 \\ X_0 < c \Rightarrow Y_0 \in \tau_2 \end{cases}$$

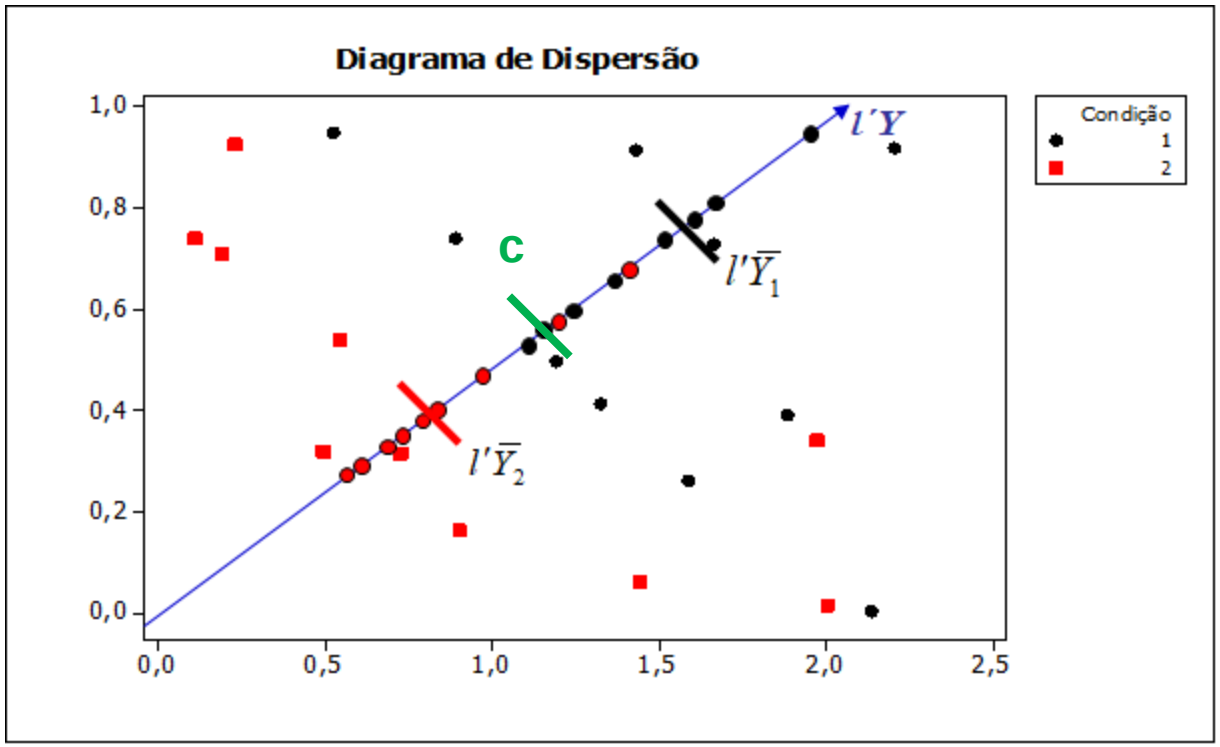
c ?

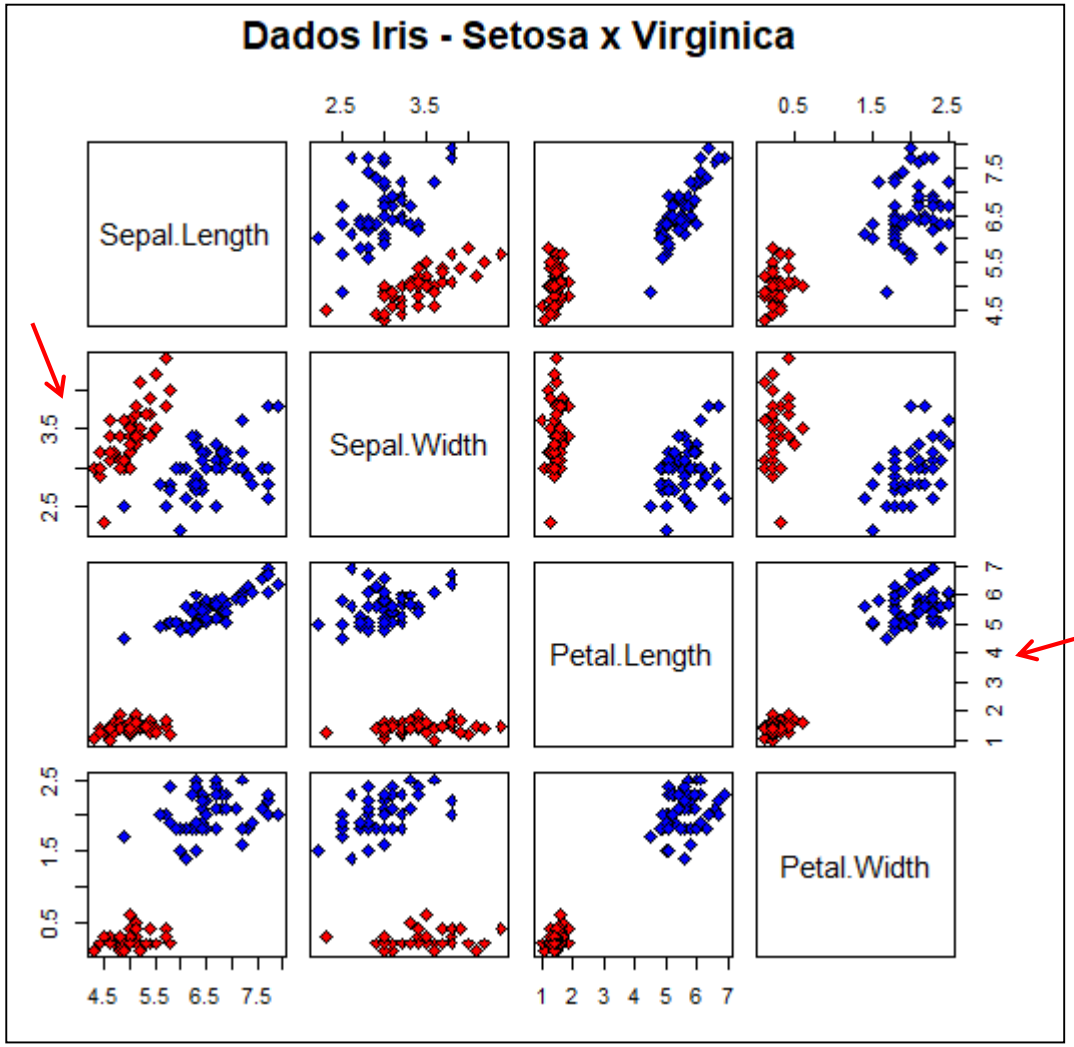
$$c = \bar{X} = \frac{1}{2}\left(\bar{X}_1 + \bar{X}_2\right) = \frac{1}{2}\left(l'\bar{Y}_1 + l'\bar{Y}_2\right) = \frac{1}{2}l'\left(\bar{Y}_1 + \bar{Y}_2\right) = \frac{1}{2}\left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1}\left(\bar{Y}_1 + \bar{Y}_2\right)$$

Função Discriminante Linear de Fisher

Critério

Dados hipotéticos: *p=2, G=2*





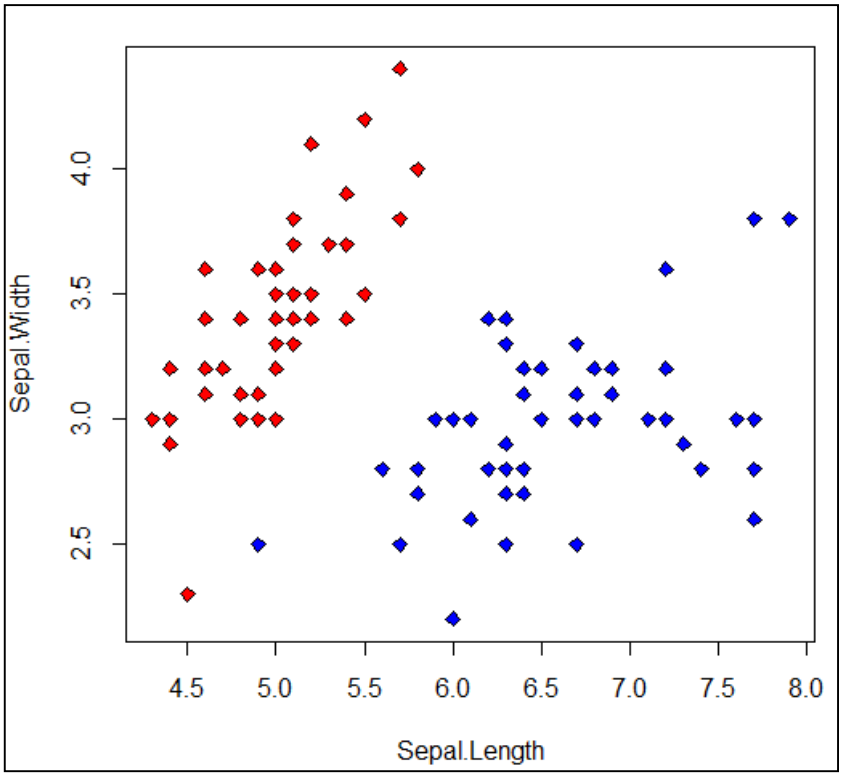
Obter a Função Discriminante Linear de Fisher nos seguintes casos:

G=2: Setosa x Virginica

- ✓ p=2: Sepal.Length
  Sepal.Width
- ✓ p=2: Petal.Length
  Petal.Width
- ✓ p=4: Sepal.Length
  Sepal.Width
  Petal.Length
  Petal.Width

Dados Iris, G=2 (Setosa x Virginica) p=2

$$X_i = l'Y_i = \left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1} Y_i$$



Centróides por grupo

	Sepal.Length	Sepal.Width
setosa	5.006	3.428
virginica	6.588	2.974
Dif	**-1.582**	**0.454**

Matriz de Covariância: S.setosa

	Sepal.Length	Sepal.Width
Sepal.Length	0.12424898	0.09921633
Sepal.Width	0.09921633	0.14368980

Matriz de Covariância: S.virginica

	Sepal.Length	Sepal.Width
Sepal.Length	0.40434286	0.09376327
Sepal.Width	0.09376327	0.10400408

Matriz de Covariância: S.pooled

	Sepal.Length	Sepal.Width
Sepal.Length	**0.2642959**	0.0964898
Sepal.Width	0.0964898	**0.1238469**

Cargas da Função discriminante

	Sepal.Length	Sepal.Width
l'	-10.23536	11.64024

c = -22.07399

Dados Iris, G=2 (Setosa x Virginica) p=2, n=150

$$X_i = l'Y_i = \left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1} Y_i$$

$$X_i = -10.23536 * Sepal.Length_i + 11.640 * Sepal.Width_i \begin{cases} \geq -22.07399 \Rightarrow Setosa \\ < -22.07399 \Rightarrow Virginica \end{cases}$$

	X_i
1	-11.459508
2	-15.232555
3	-10.857435
4	-10.997923
5	-9.271948
6	-9.874021
7	-7.505851
8	-11.599996
9	-11.278898
10	-14.068531
11	-12.202069
12	-9.552923
13	-14.209019
14	-9.091338
15	-12.804141
16	-7.124510
17	-9.874021
18	-11.459508
19	-14.108653
20	-7.967436
21	-15.694140
22	-9.131460
23	-5.177803
24	-13.787556
25	-9.552923

	X_i
26	-16.256091
27	-11.599996
28	-12.483044
29	-13.647068
30	-10.857435
31	-13.044995
32	-15.694140
33	-5.498901
34	-7.405485
35	-14.068531
36	-13.928043
37	-15.553652
38	-8.248412
39	-10.114874
40	-12.623532
41	-10.435972
42	-19.286577
43	-7.786826
44	-10.435972
45	-7.967436
46	-14.209019
47	-7.967436
48	-9.833899
49	-11.178532
50	-12.764019

	X_i
101	-26.069989
102	-27.936452
103	-37.750350
104	-30.726085
105	-31.609133
106	-42.868031
107	-21.052674
108	-40.961446
109	-39.476325
110	-31.789743
111	-29.281086
112	-34.077669
113	-34.679742
114	-29.240964
115	-26.772428
116	-28.257549
117	-31.609133
118	-34.579376
119	-48.547663
120	-35.803644
121	-33.375230
122	-24.725356
123	-46.219615
124	-33.054133
125	-30.164134

	X_i
126	-36.445839
127	-30.866573
128	-27.514989
129	-32.913645
130	-38.773886
131	-43.149006
132	-36.626448
133	-32.913645
134	-31.890109
135	-32.171084
136	-43.891567
137	-24.905966
138	-29.421573
139	-26.491453
140	-34.539254
141	-32.492182
142	-34.539254
143	-27.936452
144	-32.351694
145	-30.164134
146	-33.656206
147	-35.382180
148	-31.609133
149	-23.882429
150	-25.467916

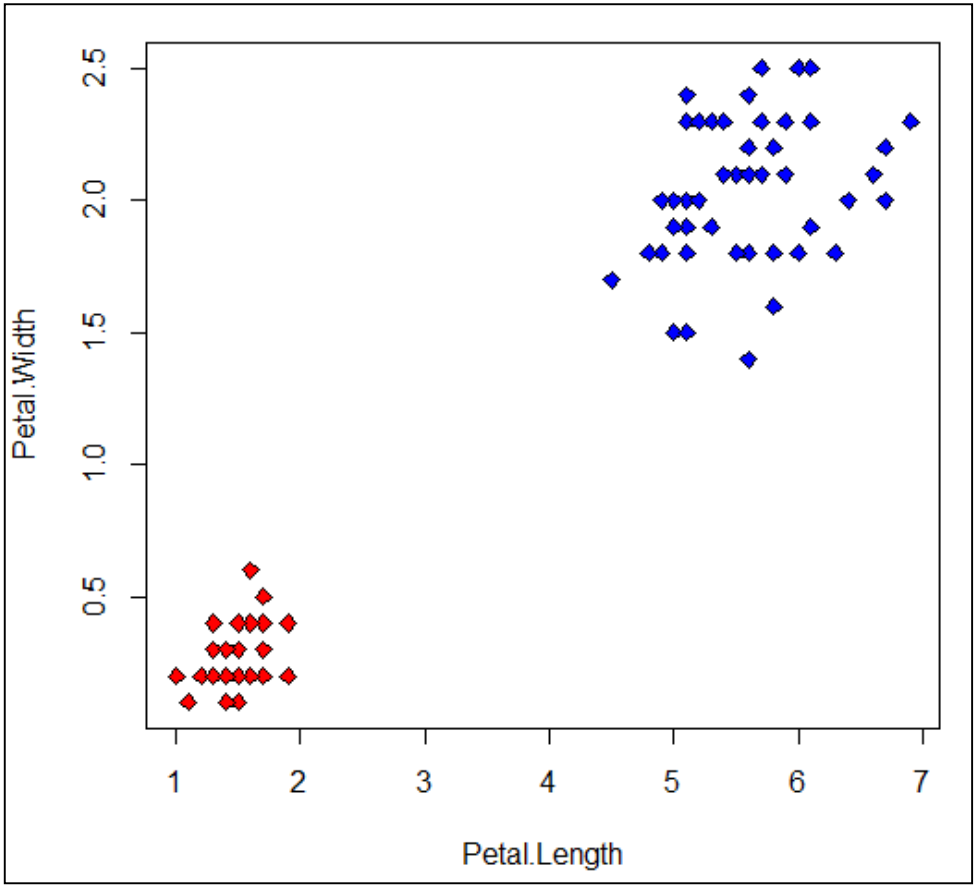
Matriz de classificação (ou confusão)

	Predito	
	setosa	virginica
setosa	50	0
virginica	1	49

% de classificação correta: 100 Setosa
98 Virginica

Dados Iris, G=2(Setosa x Virginica) p=2, n=150

$$X_i = l'Y_i = \left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1} Y_i$$



Centróides por grupos

	Petal.Length	Petal.Width
setosa	1.462	0.246
virginica	5.552	2.026
Dif	**-4.090**	**-1.78**

S.setosa

	Petal.Length	Petal.Width
Petal.Length	0.030159184	0.006069388
Petal.Width	0.006069388	0.011106122

S.virginica

	Petal.Length	Petal.Width
Petal.Length	0.30458776	0.04882449
Petal.Width	0.04882449	0.07543265

S.pooled

	Petal.Length	Petal.Width
Petal.Length	**0.16737347**	0.02744694
Petal.Width	0.02744694	**0.04326939**

Cargas da Função discriminante

	Petal.Length	Petal.Width
l'	-19.74417	-28.61337

c = -101.7476

Dados Iris, G=2 (Setosa x Virginica) p=2

$$X_i = l'Y_i = \left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1} Y_i$$

$$X_i = -19.74417 * Petal.Length_i - 28.6133711.640 * Petal.Width_i \begin{cases} \geq -101.7476 \Rightarrow Setosa \\ < -101.7476 \Rightarrow Virginica \end{cases}$$

	X_i
1	-33.36452
2	-33.36452
3	-31.39010
4	-35.33893
5	-33.36452
6	-45.01044
7	-36.22585
8	-35.33893
9	-33.36452
10	-32.47760
11	-35.33893
12	-37.31335
13	-30.50318
14	-24.57993
15	-29.41568
16	-41.06161
17	-37.11277
18	-36.22585
19	-42.14910
20	-38.20027
21	-39.28777
22	-41.06161
23	-25.46685
24	-47.87178
25	-43.23660

	X_i
26	-37.31335
27	-43.03602
28	-35.33893
29	-33.36452
30	-37.31335
31	-37.31335
32	-41.06161
33	-32.47760
34	-33.36452
35	-35.33893
36	-29.41568
37	-31.39010
38	-30.50318
39	-31.39010
40	-35.33893
41	-34.25143
42	-34.25143
43	-31.39010
44	-48.75870
45	-48.95927
46	-36.22585
47	-37.31335
48	-33.36452
49	-35.33893
50	-33.36452

	X_i
101	-189.99845
102	-155.06068
103	-176.57869
104	-162.07143
105	-177.46561
106	-190.39961
107	-137.49150
108	-175.89235
109	-166.02026
110	-191.97287
111	-157.92202
112	-159.00951
113	-168.68102
114	-155.94760
115	-169.36736
116	-170.45486
117	-160.09701
118	-195.23536
119	-202.04554
120	-141.64091
121	-178.35253
122	-153.97318
123	-189.51269
124	-148.25051
125	-172.62986

	X_i
126	-169.96910
127	-146.27609
128	-148.25051
129	-170.65544
130	-160.29759
131	-174.80485
132	-183.58944
133	-173.51677
134	-143.61533
135	-150.62608
136	-186.25020
137	-179.23945
138	-160.09701
139	-146.27609
140	-166.70660
141	-179.23945
142	-166.50603
143	-155.06068
144	-182.30136
145	-184.07520
146	-168.48044
147	-153.08626
148	-159.89643
149	-172.42928
150	-152.19934

Matriz de classificação (ou confusão)

	Predito	
	setosa	virginica
setosa	50	0
virginica	0	50

100% de classificação correta!

Dados Iris, G=2 (Setosa x Virginica) p=4

Centróide dos grupos:

	Sepal.L	Sepal.W	Petal.L	Petal.W
setosa	5.006	3.428	1.462	0.246
virginica	6.588	2.974	5.552	2.026

S.comum

0.264	0.096	0.160	0.030
0.096	0.124	0.042	0.028
0.160	0.042	0.167	0.027
0.030	0.028	0.027	0.043

Cargas do discriminante linear

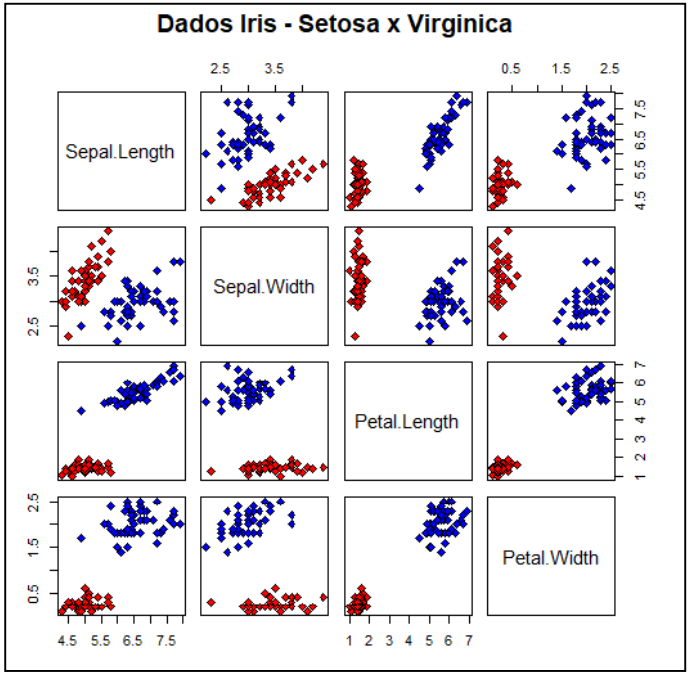
	LD1
Sepal.L	-1.1338828
Sepal.W	-0.8603685
Petal.L	2.6138926
Petal.W	2.6310427

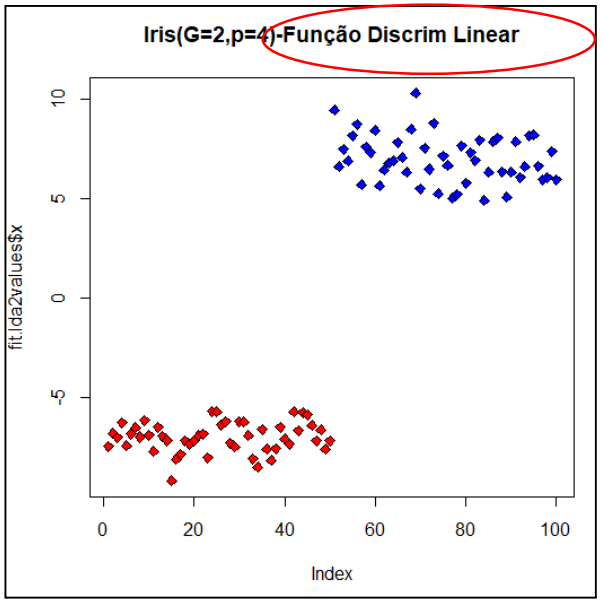
LD1 ≤ 0 Setosa
> 0 Virginica

100% de classificação correta em ambos os grupos!

Escore da função discriminante

	LD1
1	-7.437061
2	-6.780101
3	-6.986787
4	-6.264583
5	-7.409710
6	-6.810997
…	
47	-7.172393
48	-6.612009
49	-7.574522
50	-7.151599
101	9.449657
102	6.601690
103	7.486855
104	6.906517
105	8.168899
106	8.749638
107	5.699715
108	7.602359
…	
148	6.074355
149	7.382464
150	5.967087





Análise de Componentes Principais (Único grupo)

$$Y_{n\times p} \Rightarrow Z_{n\times m}$$

$$\Sigma = V\Lambda V'$$

$$\Sigma\, V_j = \lambda_j V_j$$

$$Z_{ji} = V_j' Y_i$$

escore

$$V_j = a;\ \max_{\|a\|=1} \frac{a'\Sigma a}{a'a}$$

$$\hat{\Sigma} = \frac{1}{n-1} S_T$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\left(Y_i - \bar{Y}\right)'$$

$$Z_{n\times m} = Y_{n\times p} V_{p\times m}$$

Análise Discriminante G=2 (ou mais grupos)

$$Y_{n\times p};\ n = \sum_{g=1}^{G} n_g;\quad Y_{i\ p\times 1} \Rightarrow X_i = l'Y_i$$

$$\max_{l;X=l'Y} \frac{\left(\mu_{X_1} - \mu_{X_2}\right)^2}{\sigma_X^2};$$

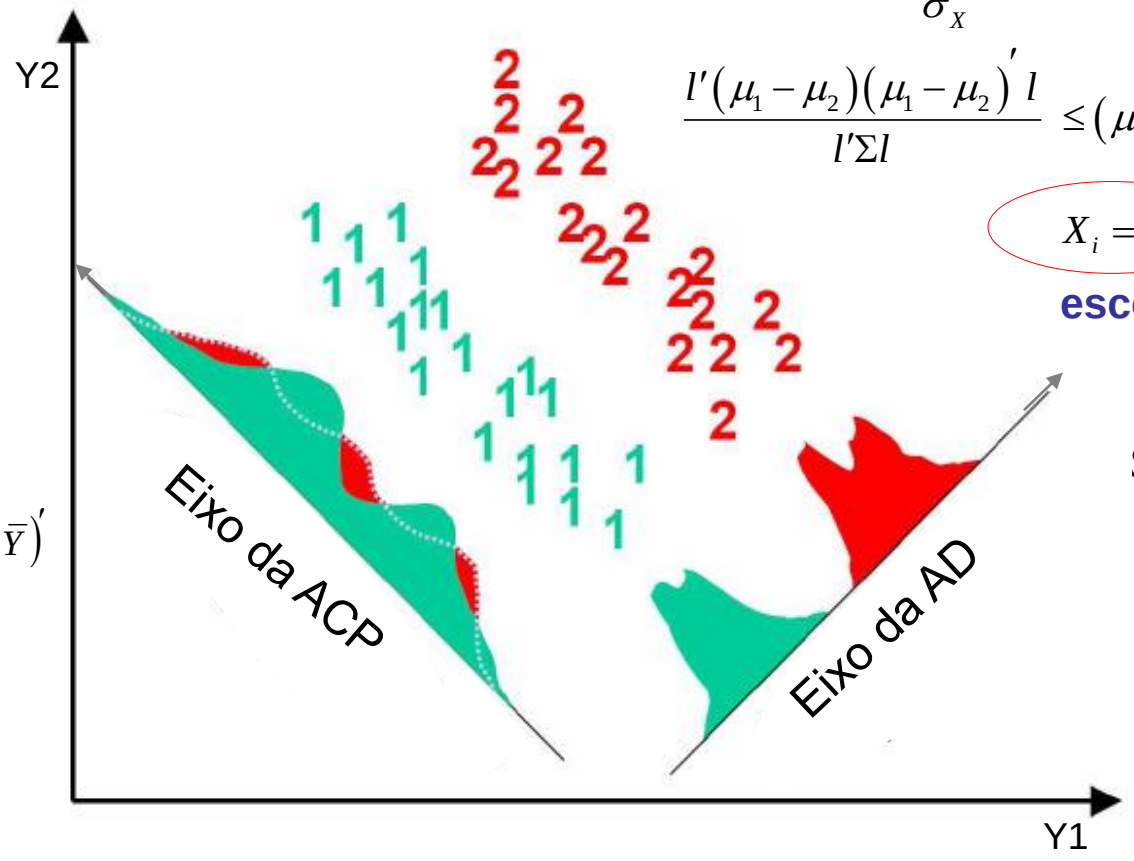
$$\frac{l'\left(\mu_1 - \mu_2\right)\left(\mu_1 - \mu_2\right)' l}{l'\Sigma l} \le \left(\mu_1 - \mu_2\right)' \Sigma^{-1} \left(\mu_1 - \mu_2\right)$$

$$X_i = l'Y_i = \left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1} Y_i$$

escore

$$S_T = S_B + S_W$$

$$S_c = \frac{1}{n-G} S_W$$

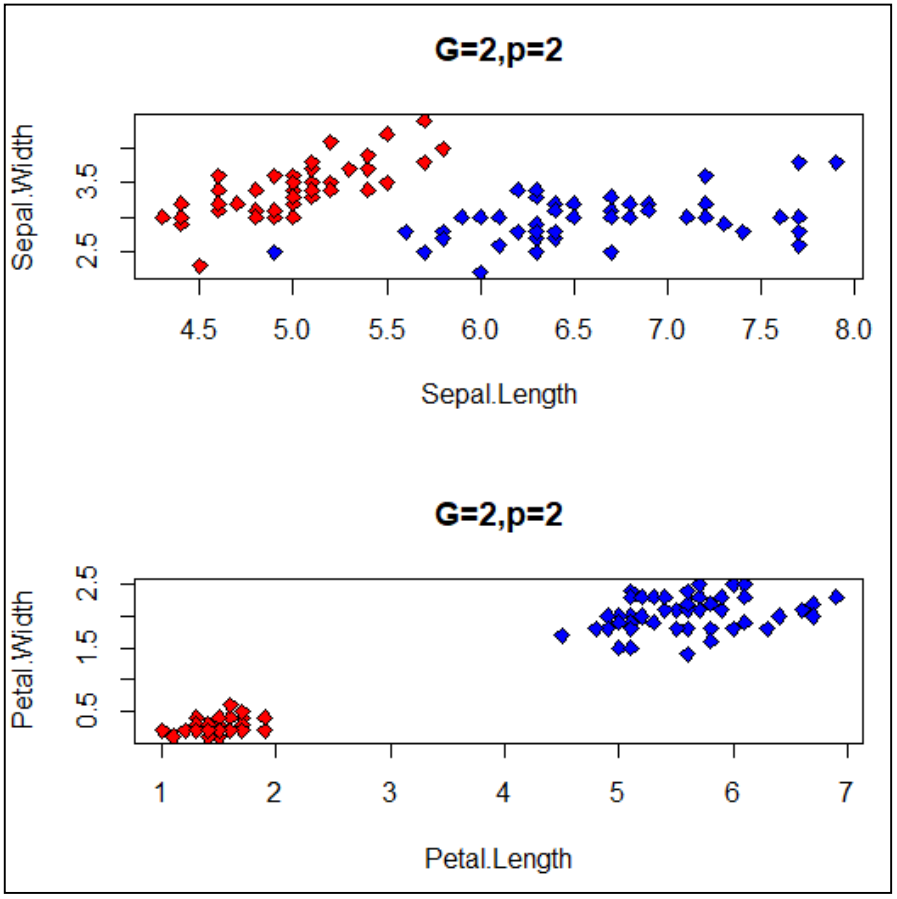


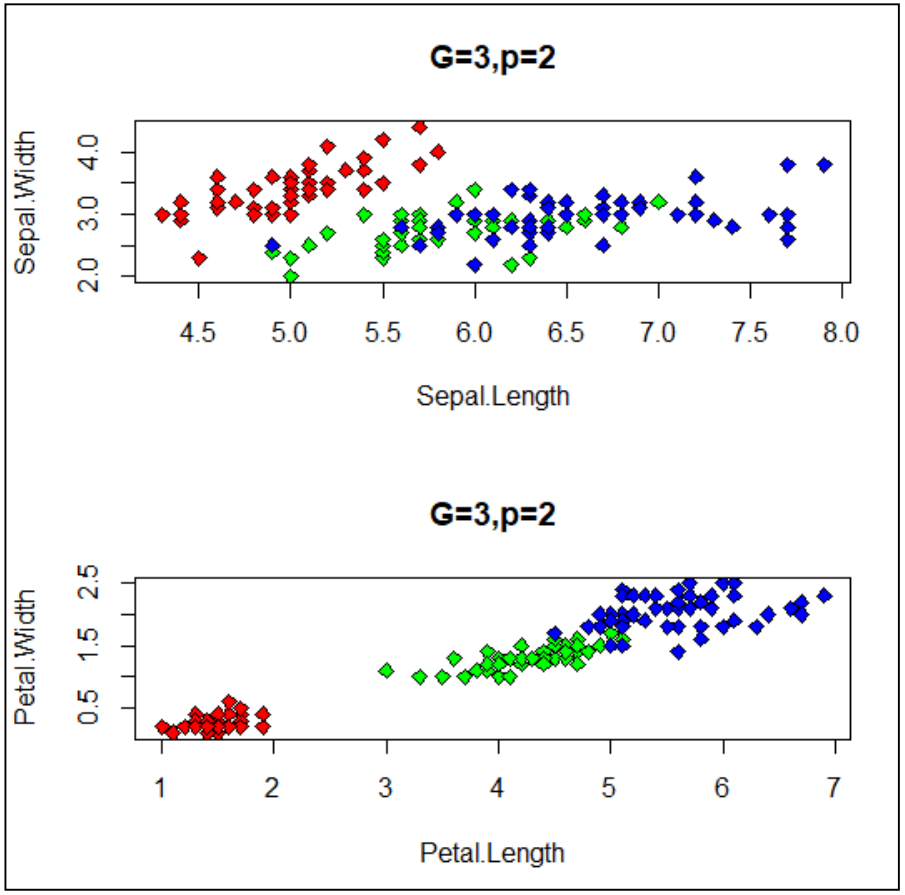
Funções Discriminantes de Fisher

Solução de Fisher: G=2 $\rightarrow$ G>2

Número de funções discriminantes: $m \leq \min(n, p, G-1)$

Como obter as funções discriminantes?





Funções Discriminantes de Fisher

Solução de Fisher para Muitas Populações

$$Y_{n\times p} = \begin{bmatrix} Y_{1(n_1\times p)} \\ \dots \\ Y_{G(n_G\times p)} \end{bmatrix} \Rightarrow \begin{cases} Y_i \in \Re^p; \quad E(Y_i \mid \tau_1) = \mu_{1(p\times 1)} \quad \dots \quad E(Y_i \mid \tau_G) = \mu_{G(p\times 1)} \\ Cov(Y_i \mid \tau_g) = \Sigma_g = \Sigma_{(p\times p)}, \quad g = 1,2,\dots,G \end{cases}$$

Matriz de covariância homogênea

$$Y_i \in \Re^p \rightarrow X_i = l' Y_i;$$

B: Matriz de covariância ENTRE grupos

$$l = \arg\max_{l;X=l'Y} \frac{\sum_{g=1}^{G} \left(\mu_{X_g} - \bar{\mu}_X\right)^2}{\sigma_X^2} \Rightarrow \frac{\sum_{g=1}^{G} \left(l'\mu_g - l'\bar{\mu}\right)^2}{\sigma_X^2} = \frac{l'\sum_{g=1}^{G}\left(\mu_g - \bar{\mu}\right)\left(\mu_g - \bar{\mu}\right)' l}{l'\Sigma l} = \frac{l'\, B\, l}{l'\Sigma l}$$

Σ : Matriz de covariância DENTRO de grupos

As funções discriminantes, $L_{p\times m} = (l_1,\dots,l_m)$, são obtidas a partir dos autovetores da matriz $\Sigma^{-1} B$, restritos a $L' \Sigma L = I$.

$$\Sigma^{-1/2}\, B\, \Sigma^{-1/2} = P\Lambda P'; \quad L = \Sigma^{-1/2} P; \quad m \le \min(n, p, G-1)$$

Matriz simétrica

Funções Discriminantes de Fisher

Método de Fisher para Muitas Populações

- Dados Amostrais: maximizar a função em termos de estimadores apropriados

$$\frac{l' B\, l}{l'\Sigma l} \quad \Rightarrow \quad \hat{l} = \arg\max_l \frac{l' \hat{B}\, l}{l' \hat{\Sigma}\, l} \qquad \Rightarrow \hat{L}_{p\times m} = \left(\hat{l}_1,...,\hat{l}_m\right)$$

Matriz de "Soma de Quadrados e Produtos Cruzados Entre grupos" (SQPC da MANOVA):

$$\hat{B}_{p\times p} = S_B = \sum_{g=1}^{G} n_g \left(\bar{Y}_g - \bar{Y}\right)\left(\bar{Y}_g - \bar{Y}\right)'$$

Matriz de "Quadrado Médio e Produto Cruzado Dentro de grupos" (QMPC da MANOVA):

$$\hat{\Sigma} = S_{c_{p\times p}} = \frac{\left(n_1 - 1\right)S_1 + ... + \left(n_G - 1\right)S_G}{n_1 + ... + n_G - G} = \frac{1}{n-G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(Y_{gi} - \bar{Y}_g\right)\left(Y_{gi} - \bar{Y}_g\right)' = \frac{1}{n-G} S_W$$

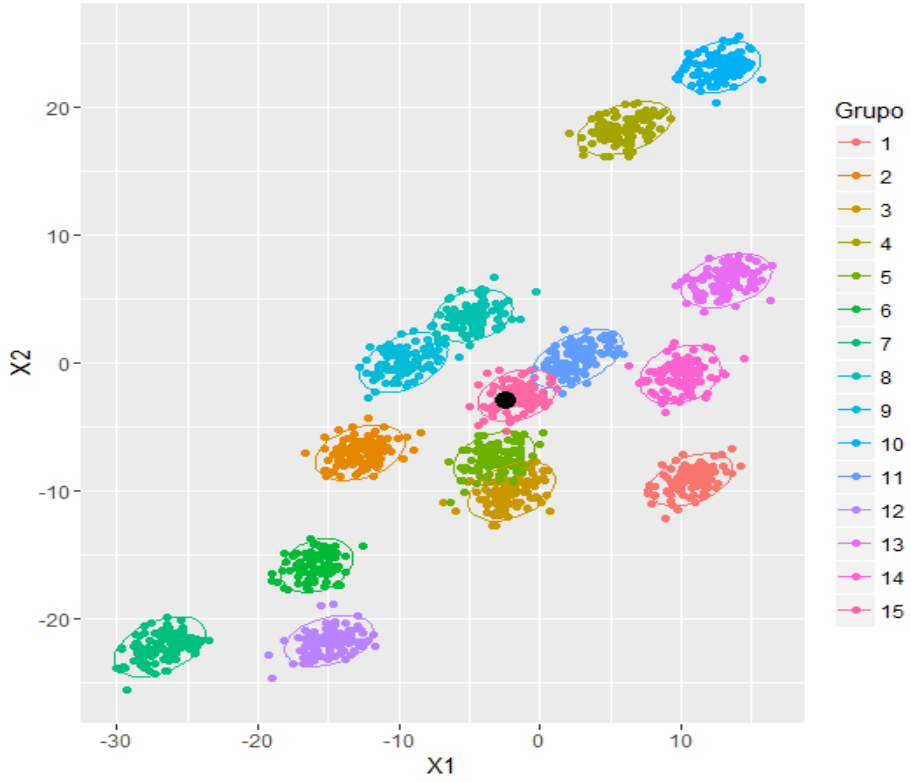
- Regra de Classificação Amostral:

Alocar a observação Y_0 ($\in \Re^p$) à população τ_k em que o valor da função discriminante X_0 ($\in \Re^m$) está mais "próxima" de seu centróide

Distância Euclidiana Padronizada

Funções Discriminantes de Fisher

Populações Estratificadas em Muitos Grupos (G>2)



$$m \leq \min\left(n, p, G-1\right)$$

Como realizar a redução dos dados (p=2, G=15)? Uma única dimensão ($X = l'Y$) é suficiente para uma boa discriminação dos grupos?

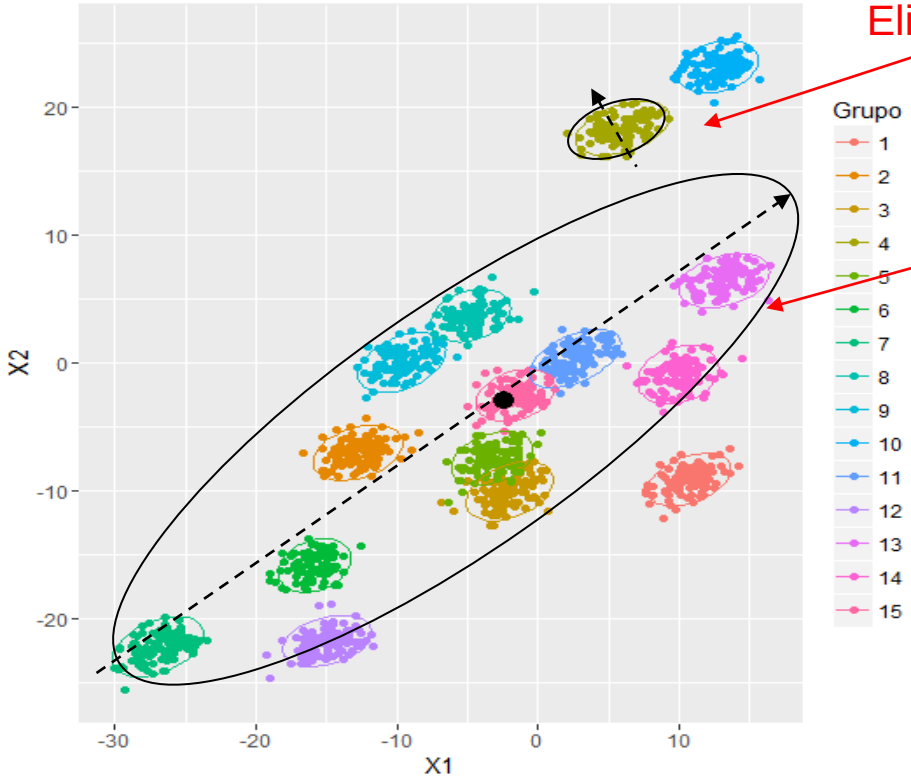
Função Discriminante de Fisher

Populações Estratificadas em Muitos Grupos (G>2)

Entendendo a direção discriminante

$$\max_l \frac{l' \hat{B} l}{l' \hat{\Sigma} l}$$

$$m \leq \min(n, p, G-1)$$



Elipse menor: Variabilidade Dentro de grupos

Elipse maior: Variabilidade Entre grupos

A direção discriminante ótima é aquela que maximiza B (eixo de variação ENTRE grupos) relativamente a Σ (eixo de variação DENTRO de grupos).

Alocar a observação Y_0 ($\in \Re^p$) à população τ_k em que o valor da função discriminante X_0 ($\in \Re^m$) está mais "próxima" de seu centróide

Dados Iris: G=3 e p=4

Probabilidades a priori dos grupos

setosa	versicolor	virginica
0.3333333	0.3333333	0.3333333

Centróides dos grupos

	Sepal.L	Sepal.W	Petal.L	Petal.W
setosa	5.006	3.428	1.462	0.246
versicolor	5.936	2.770	4.260	1.326
virginica	6.588	2.974	5.552	2.026

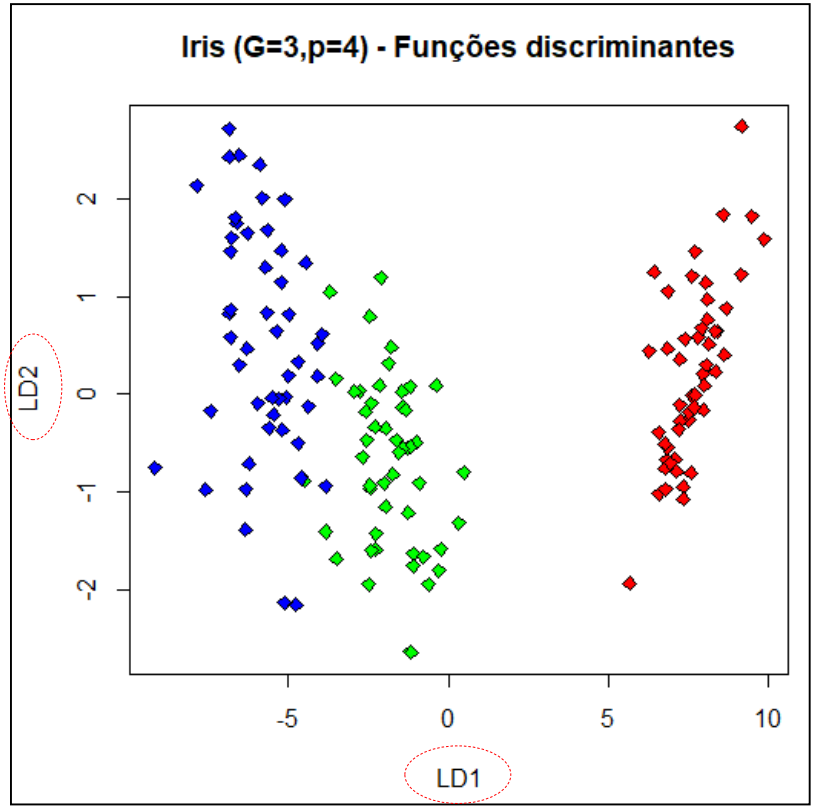
Cargas das Funções discriminantes

	LD1	LD2
Sepal.L	0.8293776	0.02410215
Sepal.W	1.5344731	2.16452123
Petal.L	-2.2012117	-0.93192121
Petal.W	-2.8104603	2.83918785

Matriz de Classificação

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	2
virginica	0	1	49

setosa	versicolor	virginica
1.00	0.96	0.98



Redução de dimensionalidade em Análise discriminante: m=min(n,p,G-1)=2

Centróide do espaço discriminante

		LD1	LD2
1	setosa	7.607600	0.2151330
2	versicolor	-1.825049	-0.7278996
3	virginica	-5.782550	0.5127666

X_0 é classificada no grupo ao qual possui menor distância (Euclidiana Padronizada) ao centróide

Análise Discriminante

Validação Empírica de um Algoritmo de Classificação Amostral

Métricas de validação via a Matriz de Classificação

Matriz de Classificação (ou de Confusão)

	Predito		
Verdade	τ_1 +	τ_2 -	
τ_1 +	n_{1c} V+	n_{1M} F-	n_1
τ_2 -	n_{2M} F+	n_{2c} V-	n_2

- Taxa de Erro Aparente (proporção de itens mal classificados):

$$TxErro = \frac{n_{1M} + n_{2M}}{n_1 + n_2} = \frac{F_+ + F_-}{n}$$ Estima Pr(classificação errada)

- Acurácia: $$Acurácia = \frac{n_{1C} + n_{2C}}{n_1 + n_2} = \frac{V_+ + V_-}{n}$$ Estima Pr(classificação correta)

Métricas de Validação via a Matriz de Classificação

Matriz de Classificação (ou de Confusão)

	Predito		
Verdade	τ_1 +	τ_2 -	
τ_1 +	n_{1c} V+	n_{1M} F-	n_1
τ_2 -	n_{2M} F+	n_{2c} V-	n_2

- Sensibilidade = $\frac{V_+}{V_+ + F_-}$ = Pr(classificação + | +)
- Especificidade = $\frac{V_-}{F_+ + V_-}$ = Pr(classificação - | -)

Poder Preditivo via a Curva ROC: Sensibilidade x (1-Especificidade)

- Preditivo Positivo = $\frac{V_+}{F_+ + V_+}$ Precisão do classificador
- Preditivo Negativo = $\frac{V_-}{F_- + V_-}$
- Escore F1 = $2\frac{\mathrm{Pr}\,ecisão * Sensibilidade}{\mathrm{Pr}\,ecisão + Sensibilidade}$

Média harmônica da precisão e sensibilidade

Análise Discriminante

Validação de um Algoritmo de Classificação

Matriz de Classificação (ou de Confusão)

	Predito		
Verdade	τ_1	τ_2	
τ_1	n_{1c} **V+**	n_{1M} **F+**	n_1
τ_2	n_{2M} **F-**	n_{2c} **V-**	n_2

TxErro subestima a Probabilidade de erro de classificação (populacional): os mesmos dados são usados para Treinamento e Teste do algoritmo

Alternativas

- Método de Particionamento (*Data Split*): particiona os dados em Amostra de Treinamento e Amostra de Validação (Teste)
- Método de "Validação Cruzada" (Cross-validation)

Análise Discriminante

Validação de um Algoritmo de Classificação Amostral

Validação Cruzada pelo método Leave-One-Out (*Fold*=N)

1. Inicie com as observações de τ_1 . Omita uma obs deste grupo e obtenha a função de classificação baseada nos remanescentes N-1=(n1-1)+n2 observações (supondo G=2)
2. Classifique a obs omitida usando a função calculada no passo 1
3. Repetir os passos 1 e 2 até que todas as obs de τ_1 tenham sido classificadas. Calcule o número de erros de classificação neste grupo
4. Repita os passos de 1 a 3 para as observações do grupo 2.

Taxa de Erro de Classificação esperada é dada por:

$$TxErro = \frac{n_{1M}^{Cross} + n_{2M}^{Cross}}{n_1 + n_2}$$

Algoritmos de CV podem usar *Fold*=k

Análise Discriminante
Normalização de Variáveis

	Unidades Amostrais	Variáveis 1	2	…	j	…	p
G1	1	Y_{111}	Y_{112}		Y_{11j}		Y_{11p}
	2	Y_{121}	Y_{122}		Y_{12j}		Y_{12p}
	…	…	…	…	…		…
	n_1	Y_{1n11}	Y_{1n12}		Y_{1n1j}		Y_{1n1p}
G2	1	Y_{211}	Y_{212}		Y_{21j}		Y_{21p}
	2	Y_{221}	Y_{222}		Y_{22j}		Y_{22p}
	…	…	…	…	…		…
	n_2	Y_{2n21}	Y_{2n22}		Y_{2n2j}		Y_{2n2p}

$\bar{Y}_{1\,p\times1}$ $S_{1\,p\times p}$

$\bar{Y}_{2\,p\times1}$ $S_{2\,p\times p}$

$\bar{Y}_{p\times1}$ $S_{c\,p\times p}$

Na AD a normalização das variáveis é usada com a finalidade de facilitar a interpretação das cargas das variáveis na função discriminante e no cálculo de "*c*". O comando "lda" do R adota a "normalização" das variáveis para calcular as funções discriminantes, mas o "linda" não. A normalização da variável j avaliada no indivíduo i do grupo g é dada por:

$$Y^*_{gij} = \left(\frac{Y_{gij} - \bar{Y}_j}{s_{gj}} \right)$$

Média: para cada j independente de grupo

Variância: para cada grupo g e variável j

$$\bar{Y}_j = \frac{1}{n_1 + n_2} \sum_{g=1}^{2} \sum_{i=1}^{n_g} Y_{gij}$$

Média da variável j (j=1,..,p), independente de grupo

$$S_{gj} = \frac{1}{n_g - 1} \sum_{g=1}^{2} \left(Y_{gij} - \bar{Y}_{gj} \right)^2$$

Variância da variável j no grupo g

Análise Discriminante

- ✓ Regra Discriminante Linear de Fisher

- ▪ Métodos Probabilísticos de Análise Discriminante:

  Regra de Classificação de Bayes

  Regressão Logística

- ▪ MANOVA: estimar as matrizes de covariância ENTRE (S_B) e DENTRO (S_W) de grupos sob diferentes modelos (ajuste por covariáveis)