MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Sem/2020 - IME

Análise Multivariada

$$Y_{n\times p} = \left(Y_{ij}\right) \in \Re^{n\times p}$$

Já vimos ☺

- ✓ Estatísticas descritivas multivariadas, Episóides de Concentração, Boxplot Bivariado
- ✓ Distribuição N_p, Distribuições Amostrais (T^2 e W_p)
- ✓ $N_p(\mu_g;\Sigma_g)$: Inferências sobre μ_g (T^2, MANOVA, ICS, Correções para Múltiplos testes

Decomposições: SS_T e Y_{nxp}

Técnicas Multivariadas:

Já vimos ☺

- ✓ 1. Análise de Componentes Principais (CP)
- ✓ 2. Escalonamento Multidimensional (CoP)
- ✓ 3. Análise de Correspondência
- ✓ 4. Análise Fatorial

Quais são os critérios de otimização?

- Análise Discriminante (MANOVA)
- Análise de Agrupamento
- Análise de Correlação Canônica

Análise Discriminante

Técnicas de Redução de Dimensionalidade
Análise Discriminante

Análises Não-Supervisionadas

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

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

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

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

- Amostra Estratificada (G grupos): G=2 e G>2

  $\Rightarrow$ Solução de Fisher (Fatoração de matrizes: escores e cargas)

  $\Rightarrow$ 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

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

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}

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

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

Objetivos: ANÁLISE DISCRIMINANTE

Análise supervisionada,
Técnica de Aprendizado!

- 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

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

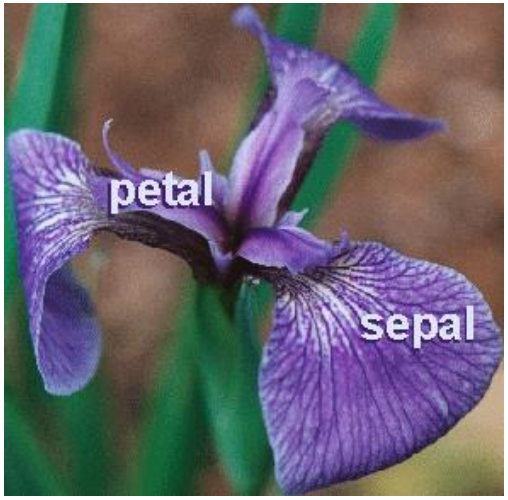
Como as características corporais dos pardais ($p=5$) podem ser usadas para predizer os grupos de pássaros Sobreviventes e Não-Sobreviventes?

CP x AD!!

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

Como caracterizar (predizer) as espécies de íris?

Análise Discriminante - Motivação

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

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

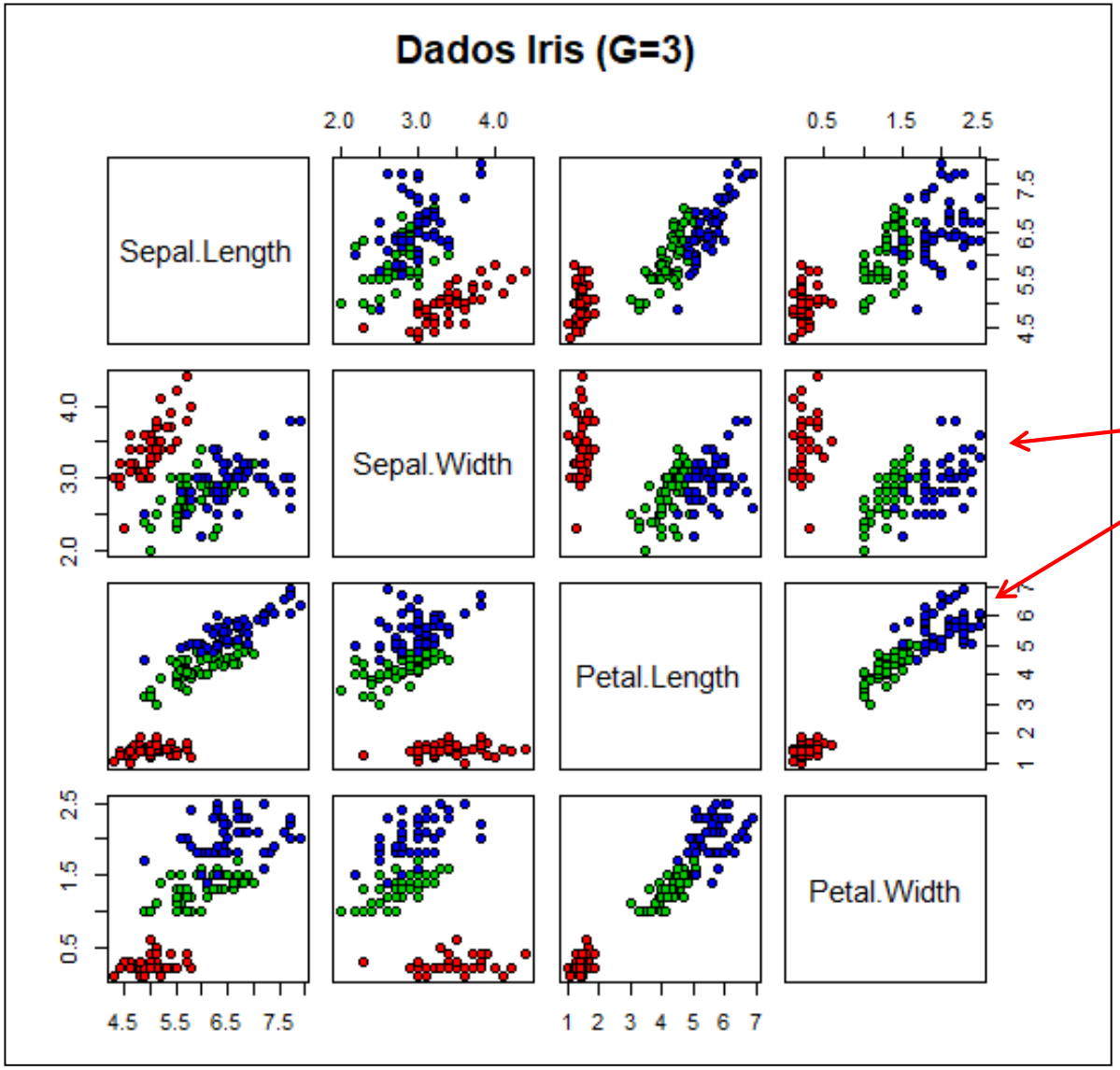
Irizarry, R.A. and Love, M.I.
Data Analysis for the Life Sciences, 2015.

```
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
G=7

Análise Discriminante



Considere a matriz de dispersão das variáveis, em particupar,

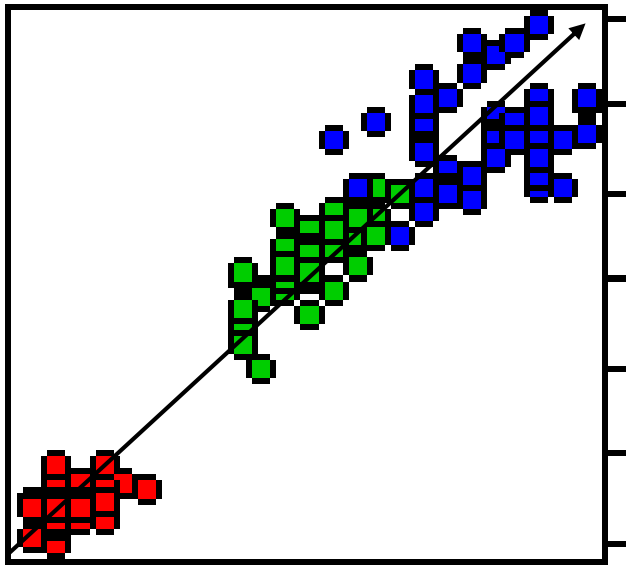
Sepal.Width x Petal.Width

Petal.Length x Petal.Width

Qual seria uma direção "ótima" (função linear das variáveis) 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

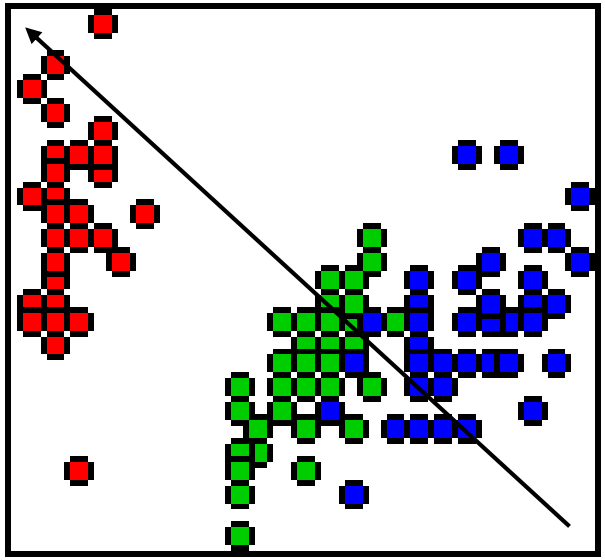


Gráfico de dispersão das observações (em $\Re^2$).

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

Que critério adotar na construção deste eixo de “máxima” discriminação?

Função Discriminante Linear de Fisher

Critério

G=2

O grupo ao qual a observação pertence é conhecido.

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

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

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

Matrizes de covariâncias homogêneas

Para G=2 a Proposta de Fisher é obter combinações lineares de $Y_i \in \Re^p$ tais que:

Vetor de cargas

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

Função Discriminante Linear de Fisher

Solução

$$Y_{n\times p}=\begin{bmatrix}Y_{1(n_1\times p)}\\ Y_{2(n_2\times p)}\end{bmatrix}\Rightarrow\begin{cases}Y_i\in\Re^p;\quad E(Y_i\mid\tau_1)=\mu_{1(p\times 1)}\\ E(Y_i\mid\tau_2)=\mu_{2(p\times 1)}\end{cases}\qquad \Sigma_1=\Sigma_2=\Sigma_{p\times p}$$

Matrizes de covariâncias homogêneas

Seja *X* uma combinação linear das variáveis multidimensionais $Y_{i\,(px1)}$. Então,

$$\Rightarrow\quad X_i=l'Y_i\qquad\begin{cases}\mu_{X_1}=E(X_i\mid\tau_1)=E(l'Y_i\mid\tau_1)=l'\mu_1\\ \mu_{X_2}=E(X_i\mid\tau_2)=E(l'Y_i\mid\tau_2)=l'\mu_2\end{cases}\qquad \sigma_X^2=Var(l'Y_i)=l'\Sigma l$$

Distância de Euclidiana Padronizada entre os centróides de X

Desigualdade de Cauchy-Schwarz

$$\max_{l;X=l'Y}\frac{(\mu_{X_1}-\mu_{X_2})^2}{\sigma_X^2};\quad \frac{l'(\mu_1-\mu_2)(\mu_1-\mu_2)'l}{l'\Sigma l}=\frac{l'\delta\,\delta'l}{l'\Sigma l}=\frac{(l'\delta)^2}{l'\Sigma l}\le(\delta'\Sigma^{-1}\delta)$$

$$(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)$$

Distância de Mahalanobis entre os centróides de Y

Vetor de cargas

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

Função Discriminante Linear de Fisher

Estimação

Suposição: independência, homocedasticidade (e prioris iguais)

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

Para dados amostrais: Adotar estimadores "apropriados" de $\mu_1,\ \mu_2,\ \Sigma$

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

Matriz de covariância comum

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

Matriz de covariância do grupo g

$$S_g = \left(n_g - 1\right)^{-1} \sum_{g=1}^{2} \sum_{i=1}^{n_g} \left(Y_{gi} - \bar{Y}_g\right)\left(Y_{gi} - \bar{Y}_g\right)'$$

Regra de Classificação Amostral: Alocar uma nova observação, Y_0 , aos Grupos

Y_0 ?

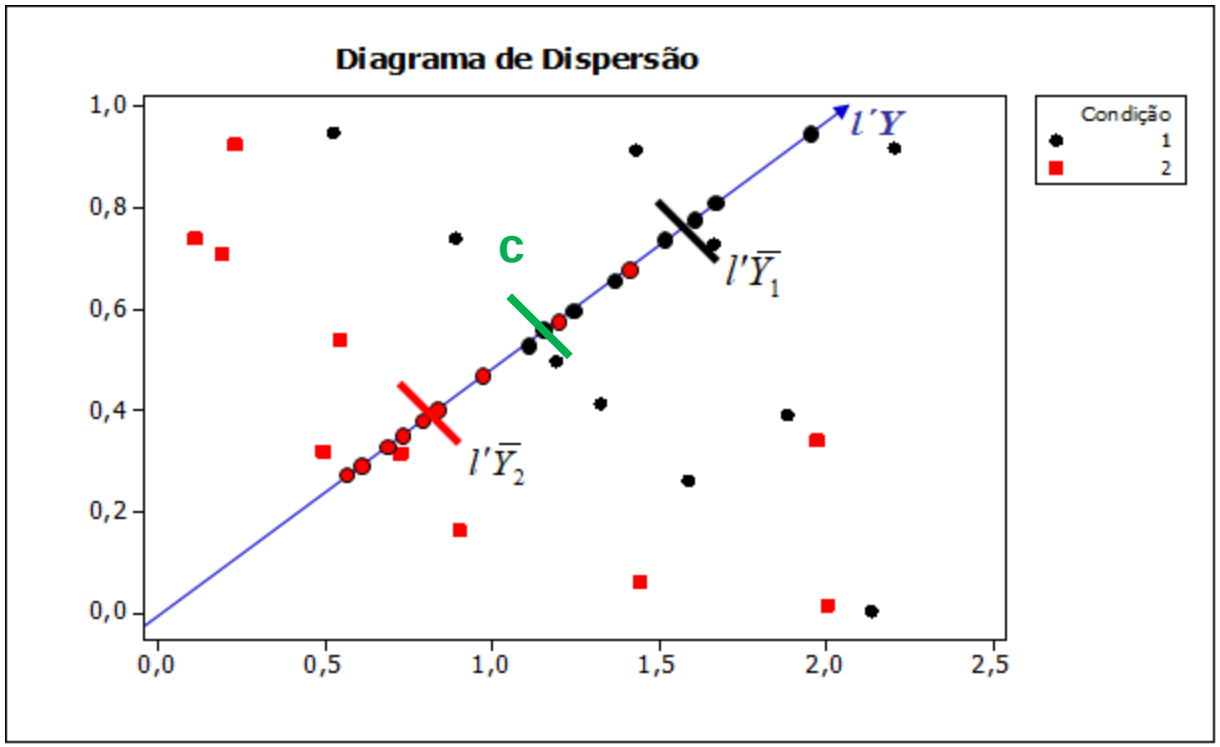
$$\begin{cases} X_0 = \hat{l}Y_0 \;\geq\; c \Rightarrow Y_0 \in \tau_1 \\ X_0 = \hat{l}Y_0 \;<\; c \Rightarrow Y_0 \in \tau_2 \end{cases}$$

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

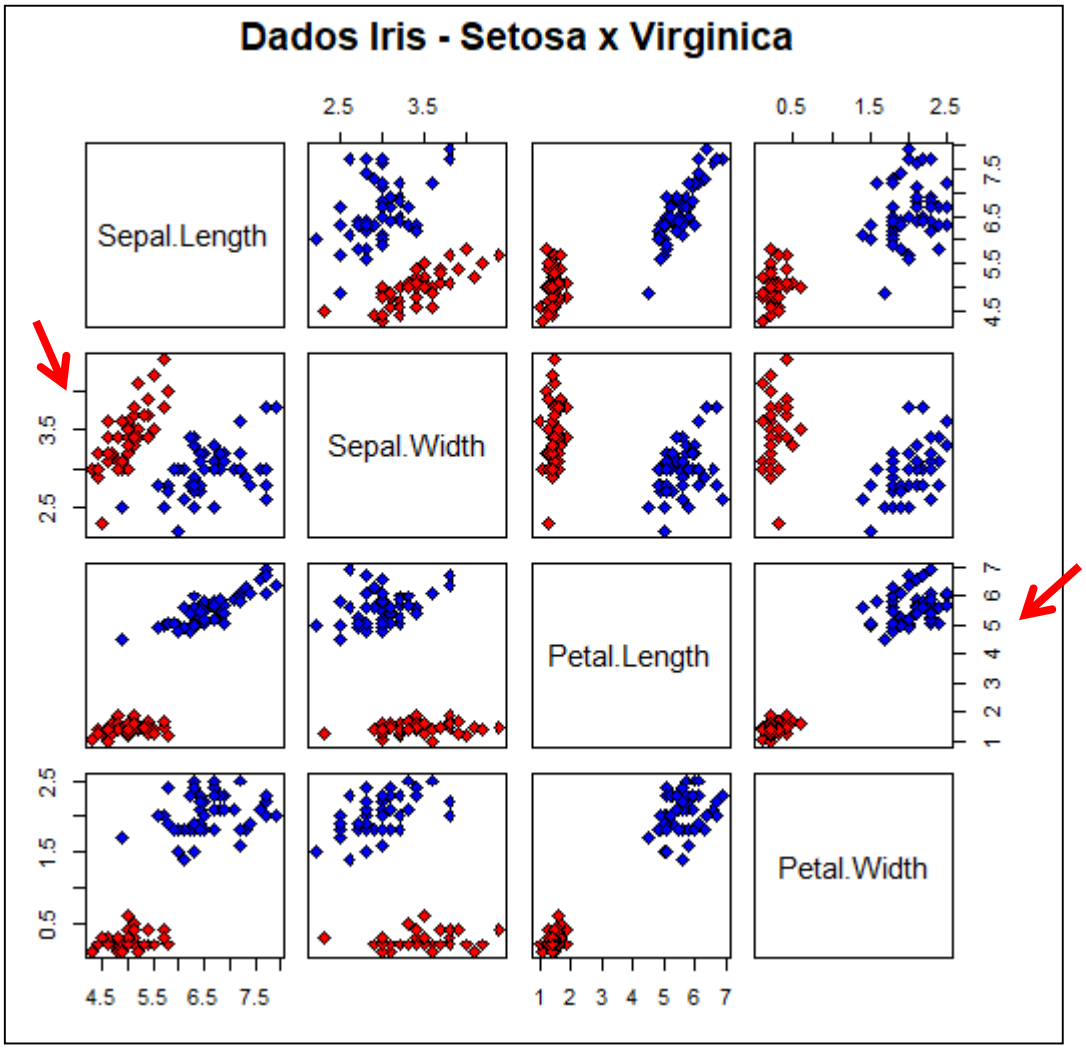
Função Discriminante Linear de Fisher

Critério

Dados hipotéticos: *p=2, G=2*



Qual é a direção da função discriminante?



Obter a Função Discriminante Linear de Fisher nos seguintes casos:

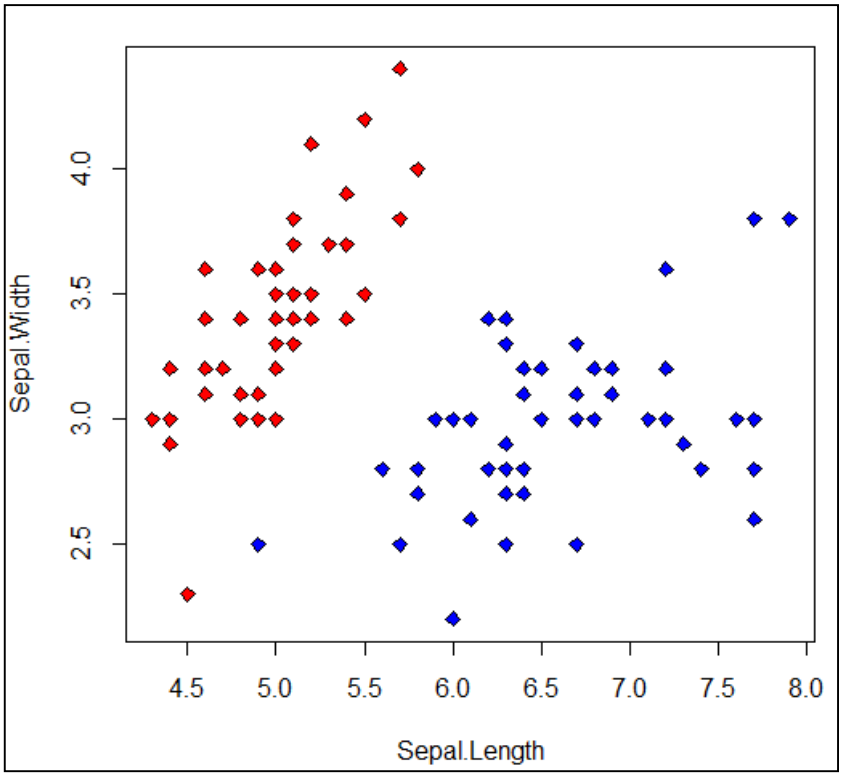
G=2: Setosa x Virginica

- ✓ 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

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

Avaliar a variável de maior Dif



Centróides por grupo

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

S.setosa

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

S.virginica

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

S.pooled

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

Avaliar a variável de menor Var

Cargas: $\hat{l}' = \left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1}$

Sepal.Length	Sepal.Width
-10.23536	11.64024

Critério:

c = -22.07399

Dados Iris, G=2(Setosa[1:50] x Virginica[101:150]) p=2

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

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

Setosa

	X_i		X_i
1	-11.459508	26	-16.256091
2	-15.232555	27	-11.599996
3	-10.857435	28	-12.483044
4	-10.997923	29	-13.647068
5	-9.271948	30	-10.857435
6	-9.874021	31	-13.044995
7	-7.505851	32	-15.694140
8	-11.599996	33	-5.498901
9	-11.278898	34	-7.405485
10	-14.068531	35	-14.068531
11	-12.202069	36	-13.928043
12	-9.552923	37	-15.553652
13	-14.209019	38	-8.248412
14	-9.091338	39	-10.114874
15	-12.804141	40	-12.623532
16	-7.124510	41	-10.435972
17	-9.874021	42	-19.286577
18	-11.459508	43	-7.786826
19	-14.108653	44	-10.435972
20	-7.967436	45	-7.967436
21	-15.694140	46	-14.209019
22	-9.131460	47	-7.967436
23	-5.177803	48	-9.833899
24	-13.787556	49	-11.178532
25	-9.552923	50	-12.764019

Virginica

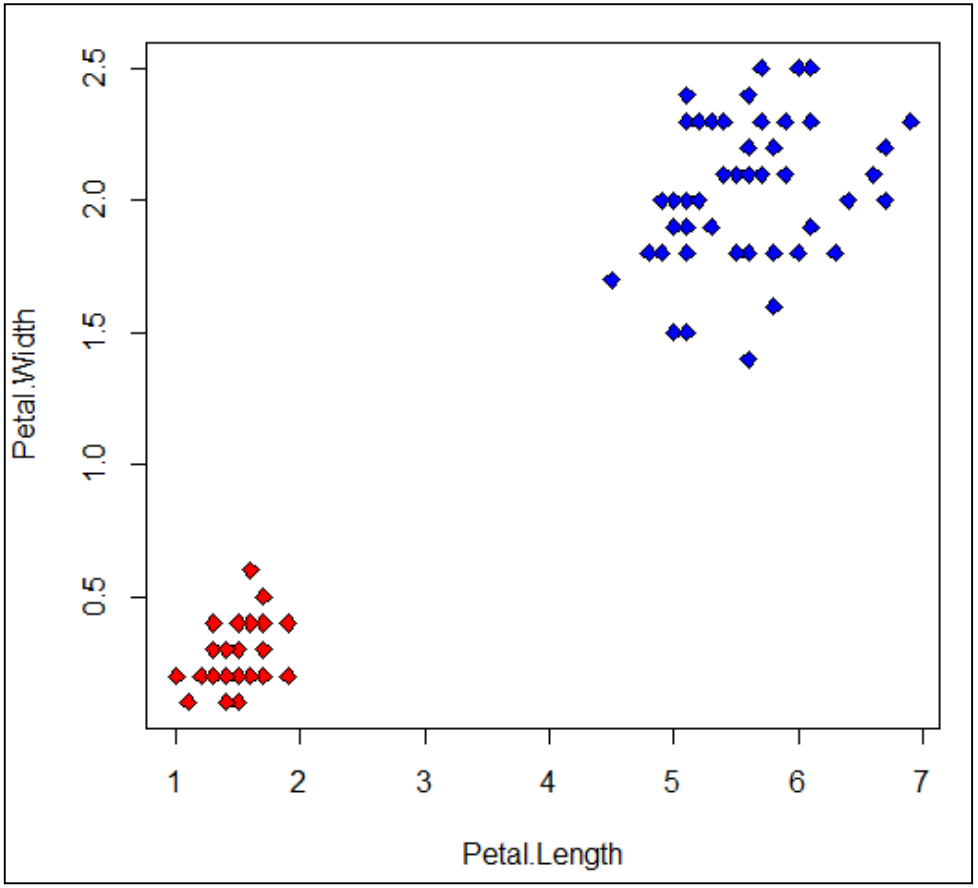
	X_i		X_i
101	-26.069989	126	-36.445839
102	-27.936452	127	-30.866573
103	-37.750350	128	-27.514989
104	-30.726085	129	-32.913645
105	-31.609133	130	-38.773886
106	-42.868031	131	-43.149006
107	-21.052674	132	-36.626448
108	-40.961446	133	-32.913645
109	-39.476325	134	-31.890109
110	-31.789743	135	-32.171084
111	-29.281086	136	-43.891567
112	-34.077669	137	-24.905966
113	-34.679742	138	-29.421573
114	-29.240964	139	-26.491453
115	-26.772428	140	-34.539254
116	-28.257549	141	-32.492182
117	-31.609133	142	-34.539254
118	-34.579376	143	-27.936452
119	-48.547663	144	-32.351694
120	-35.803644	145	-30.164134
121	-33.375230	146	-33.656206
122	-24.725356	147	-35.382180
123	-46.219615	148	-31.609133
124	-33.054133	149	-23.882429
125	-30.164134	150	-25.467916

	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

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



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:

	Petal.Length	Petal.Width
$\hat{l}'$	-19.74417	-28.61337

Critério:

c = -101.7476

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

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

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

Setosa

	X_i		X_i
1	-33.36452	26	-37.31335
2	-33.36452	27	-43.03602
3	-31.39010	28	-35.33893
4	-35.33893	29	-33.36452
5	-33.36452	30	-37.31335
6	-45.01044	31	-37.31335
7	-36.22585	32	-41.06161
8	-35.33893	33	-32.47760
9	-33.36452	34	-33.36452
10	-32.47760	35	-35.33893
11	-35.33893	36	-29.41568
12	-37.31335	37	-31.39010
13	-30.50318	38	-30.50318
14	-24.57993	39	-31.39010
15	-29.41568	40	-35.33893
16	-41.06161	41	-34.25143
17	-37.11277	42	-34.25143
18	-36.22585	43	-31.39010
19	-42.14910	44	-48.75870
20	-38.20027	45	-48.95927
21	-39.28777	46	-36.22585
22	-41.06161	47	-37.31335
23	-25.46685	48	-33.36452
24	-47.87178	49	-35.33893
25	-43.23660	50	-33.36452

Virginica

	X_i		X_i
101	-189.99845	126	-169.96910
102	-155.06068	127	-146.27609
103	-176.57869	128	-148.25051
104	-162.07143	129	-170.65544
105	-177.46561	130	-160.29759
106	-190.39961	131	-174.80485
107	-137.49150	132	-183.58944
108	-175.89235	133	-173.51677
109	-166.02026	134	-143.61533
110	-191.97287	135	-150.62608
111	-157.92202	136	-186.25020
112	-159.00951	137	-179.23945
113	-168.68102	138	-160.09701
114	-155.94760	139	-146.27609
115	-169.36736	140	-166.70660
116	-170.45486	141	-179.23945
117	-160.09701	142	-166.50603
118	-195.23536	143	-155.06068
119	-202.04554	144	-182.30136
120	-141.64091	145	-184.07520
121	-178.35253	146	-168.48044
122	-153.97318	147	-153.08626
123	-189.51269	148	-159.89643
124	-148.25051	149	-172.42928
125	-172.62986	150	-152.19934

Matriz de classificação (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

Coeficientes do discriminante linear

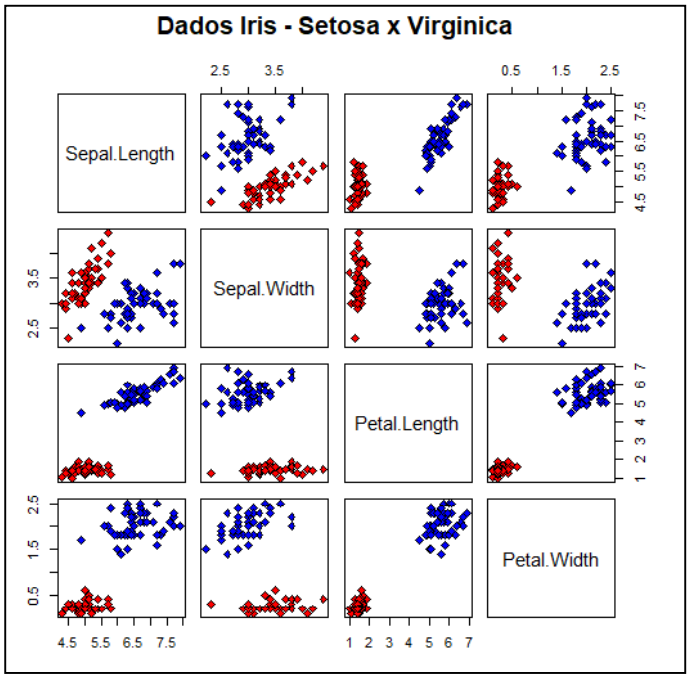
	Cargas
Sepal.L	-1.1338828
Sepal.W	-0.8603685
Petal.L	2.6138926
Petal.W	2.6310427

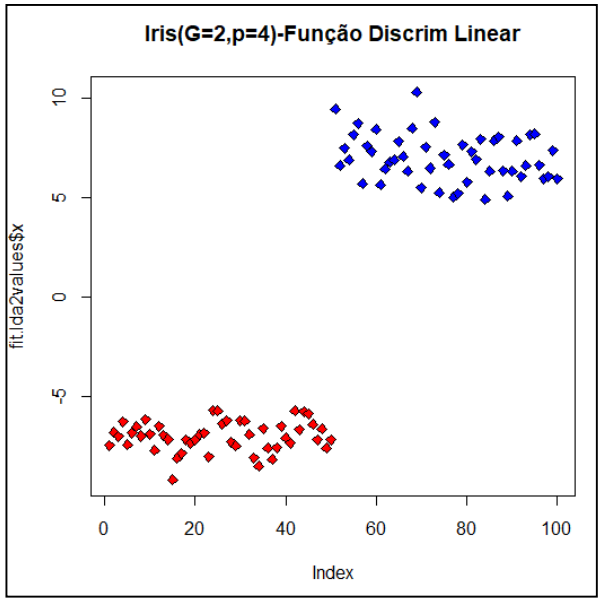
$$LD1_i = l'Y_i - c$$

LD1 ≤ 0 Setosa
> 0 Virginica

100% de classificação correta em ambos os grupos!

	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
(Suposição: único grupo)

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

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

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

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

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

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

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

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

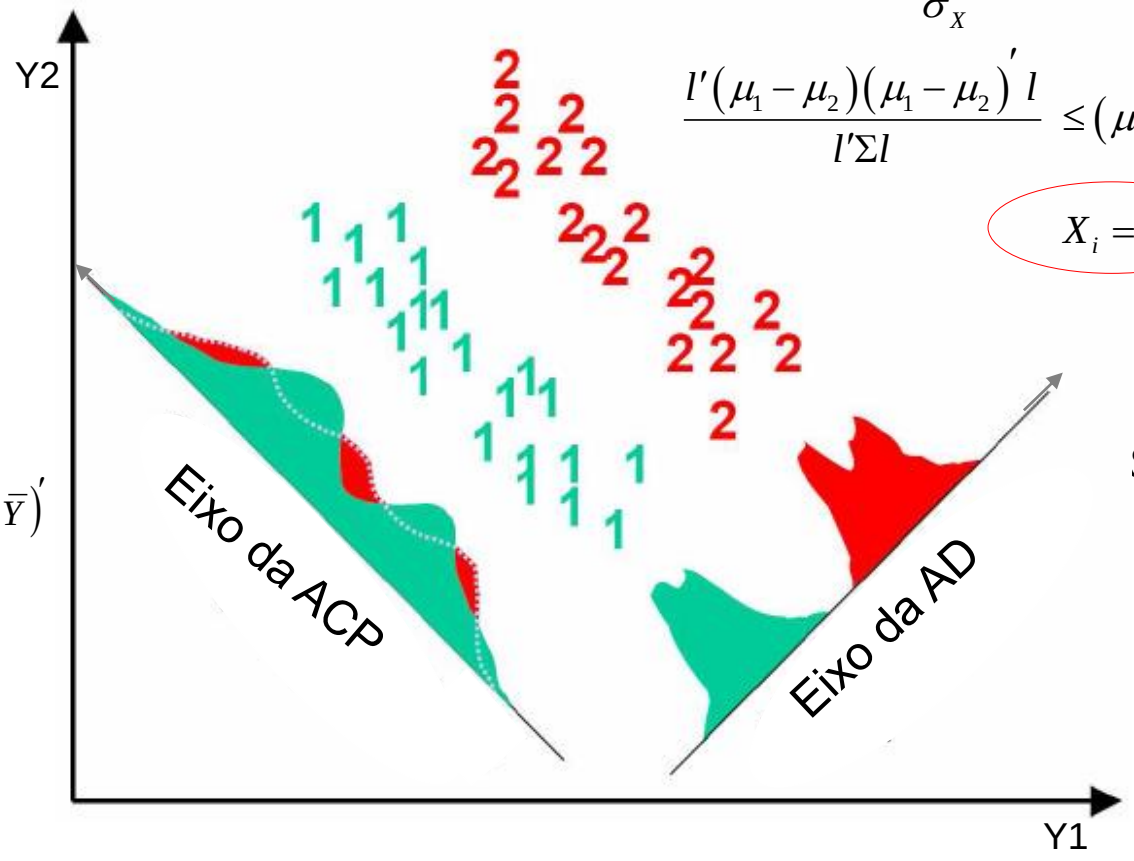
Análise Discriminante
G=2

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

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

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

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



$$S_T = S_B + S_W$$

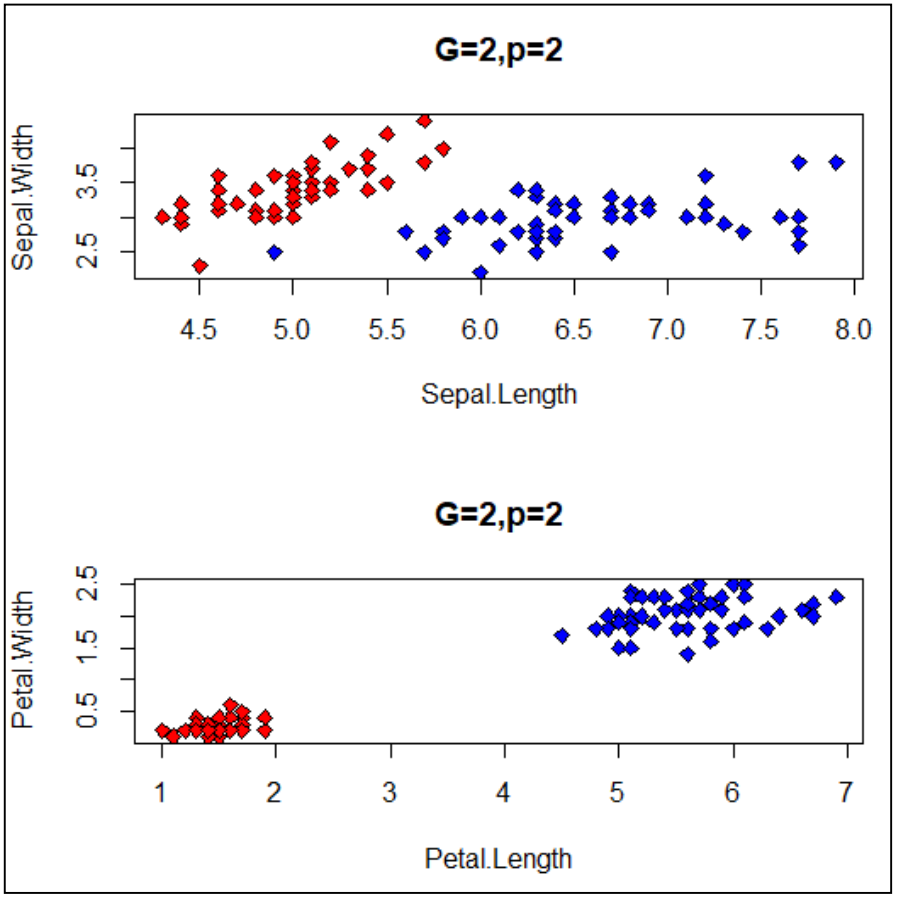
$$S_c = \frac{1}{n-G} S_W$$

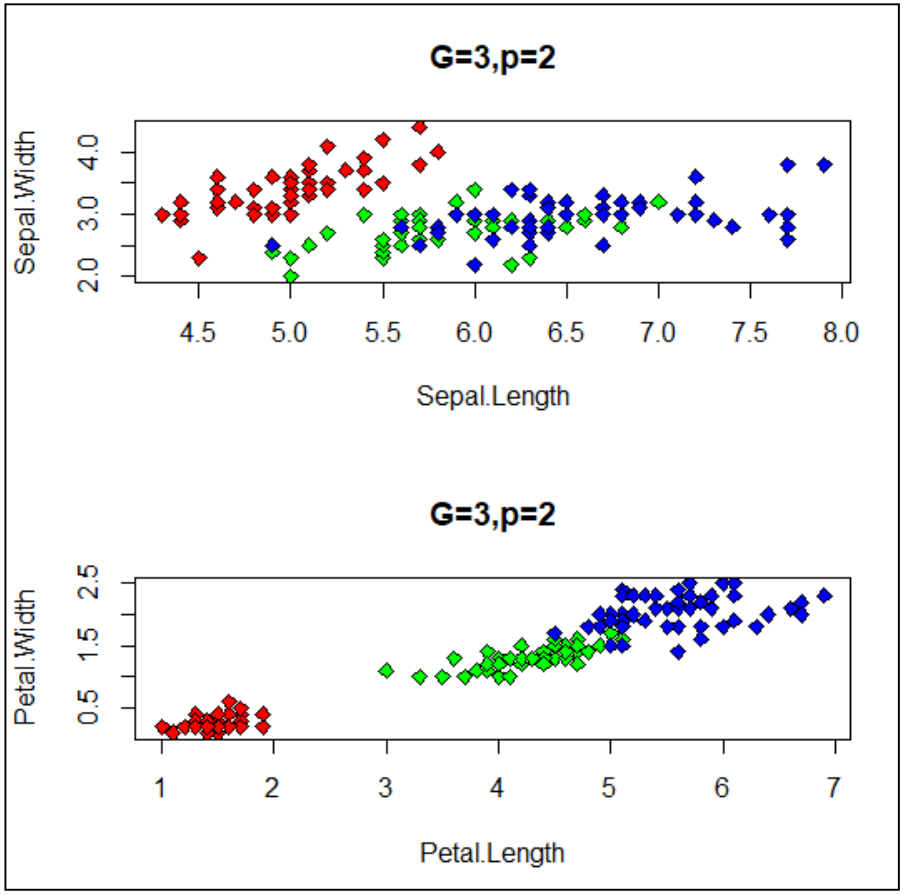
Funções Discriminantes de Fisher

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

G>2

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

Vetor de cargas

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

Σ: Matriz de covariância DENTRO de grupos

Matriz de cargas

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):

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

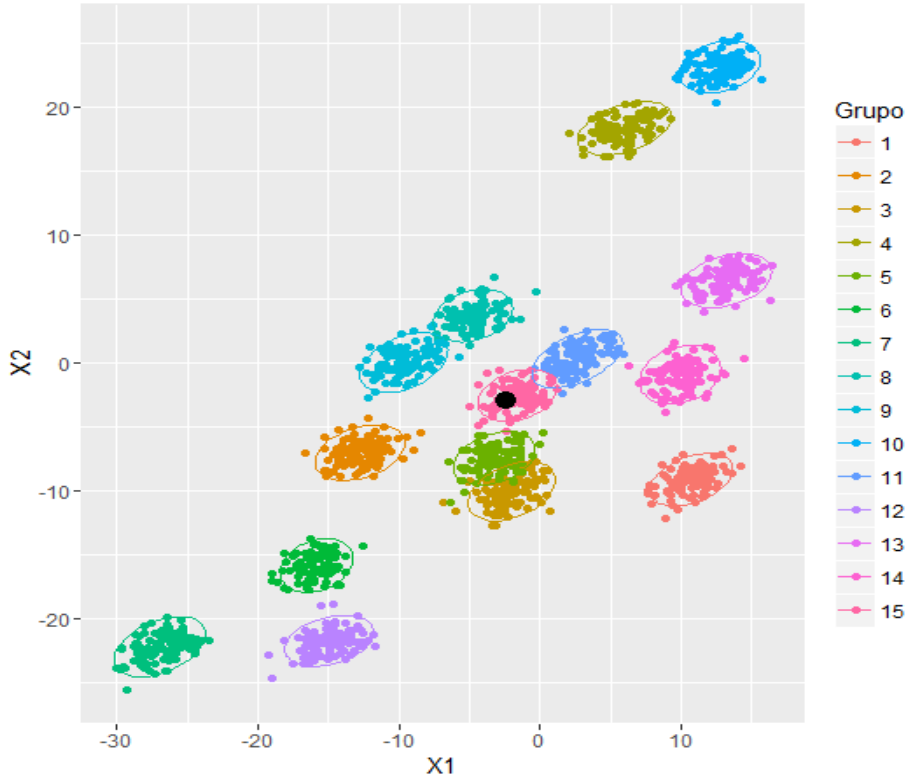
- Regra de Classificação Amostral:

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?

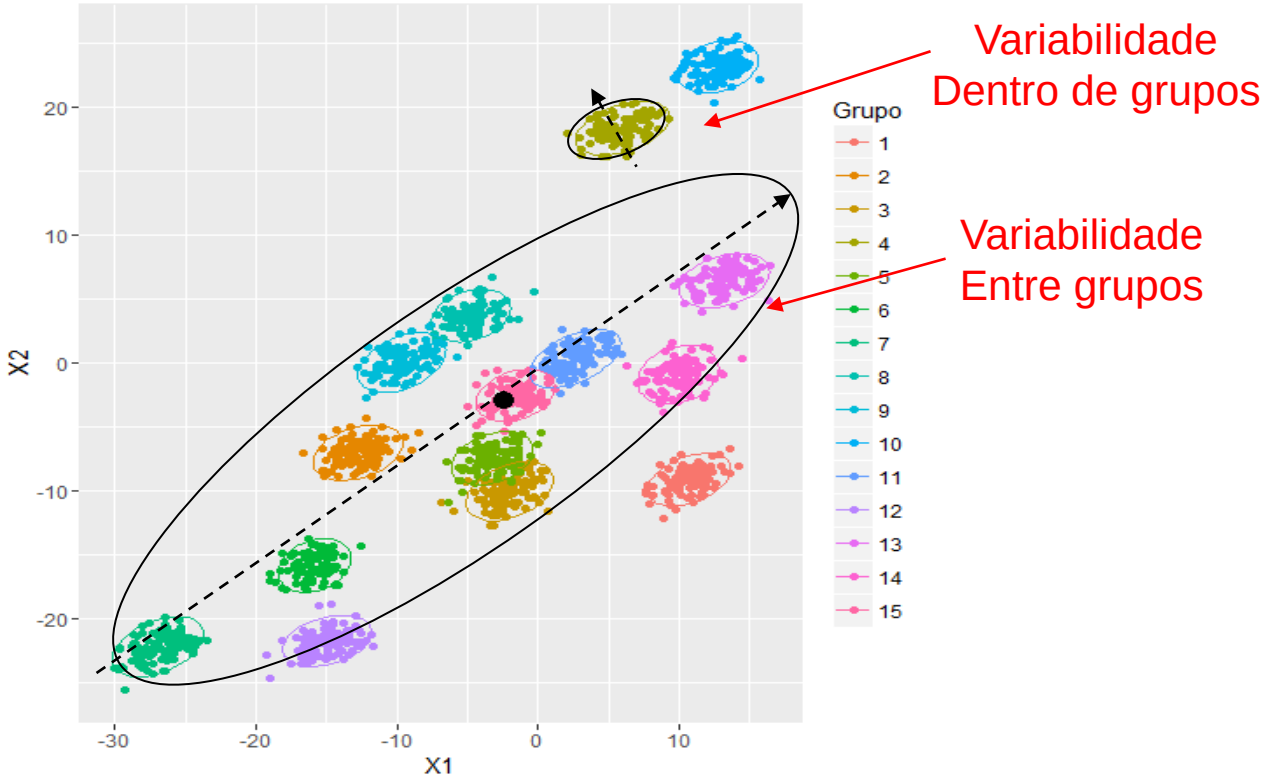
Função Discriminante de Fisher

Populações Estratificadas em Muitos Grupos (G>2)

Entendendo a direção discriminante

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

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



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

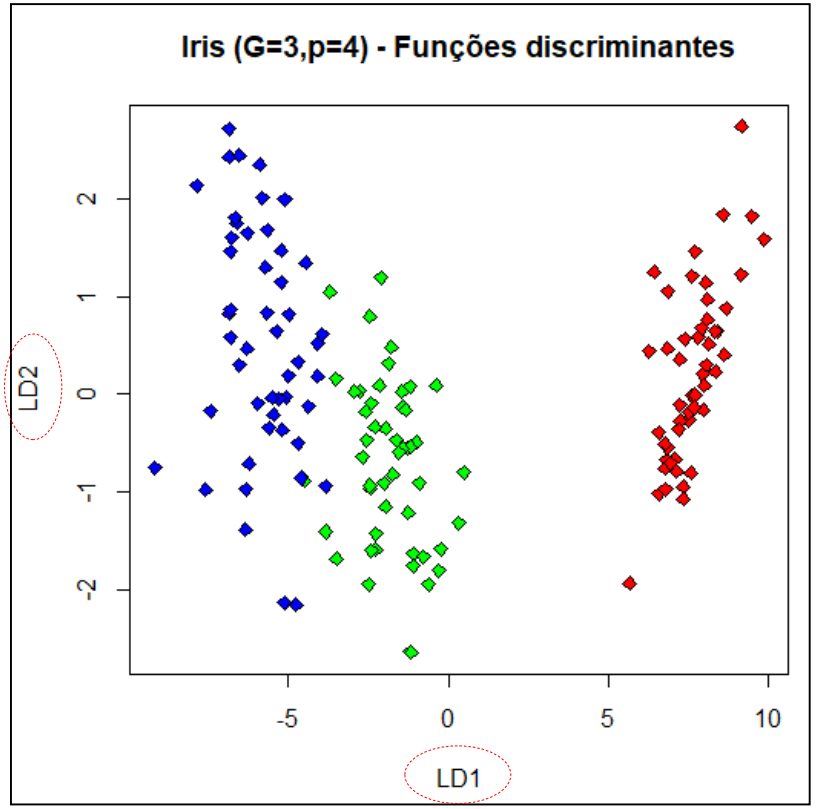
setosa	versicolor	virginica
1.00	0.96	0.98

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)



Análise Discriminante

- ✓ Regra Discriminante Linear de Fisher
- Métodos Probabilísticos: Regra de Classificação de Bayes
- Matriz de Classificação Correta (“Matriz de Confusão”):

  Método Empírico

  Validação Cruzada

- Métricas de Poder Preditivo
- Regressão Logística
- MANOVA