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, modelos MANOVA, Correlação Canônica

3. Simulação de Monte Carlo, Intervalos de Confiança Bootstrap

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 supervisionada

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

Análise Discriminante

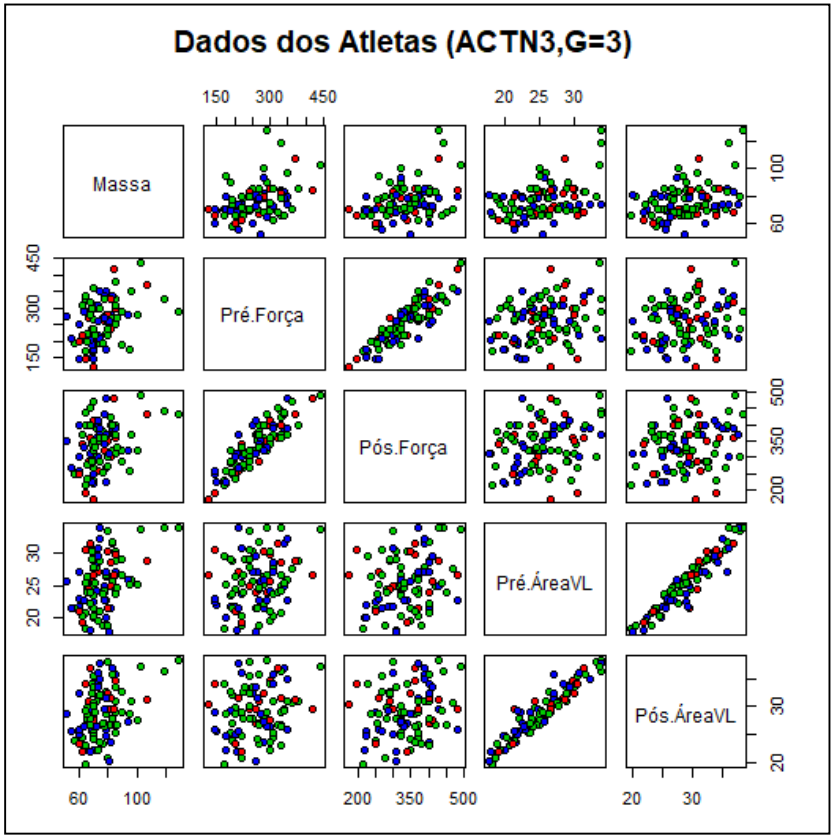
Revisando!

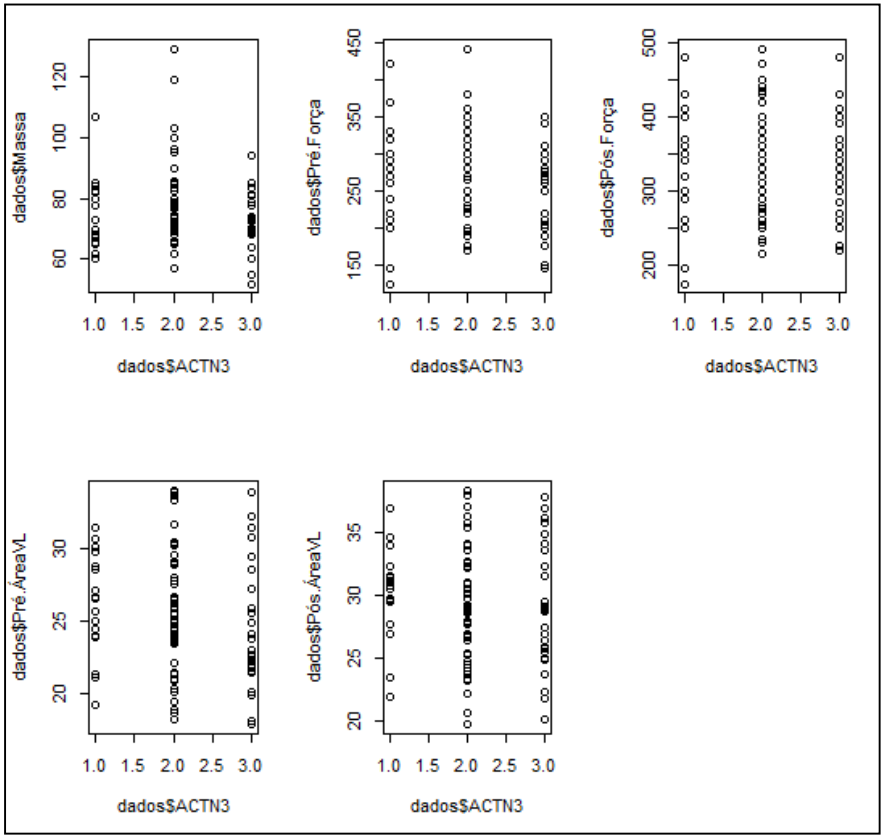
Projeto ACTN3 e Força (dados de atletas)

	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

Obtenha as variáveis discriminantes. Quais variáveis mais contribuem para a discriminação dos grupos genotípicos?

ACTN3		
1 (red)	**2** (green)	**3** (blue)
17	50	28





Análise Discriminante

Médias das variáveis por grupo genotípico (centróide):

	Massa	Pré.Força	Pós.Força	Pré.ÁreaVL	Pós.ÁreaVL	
1	75.24706	267.0588	334.7059	26.11594	29.93076	$\bar{Y}_1$
2	78.25400	267.1000	339.5000	25.75210	29.03864	$\bar{Y}_2$
3	72.31786	249.1071	335.8929	24.62186	28.79032	$\bar{Y}_3$

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

B: Matriz de SQPC devido ao efeito de grupo (ENTRE)

	Massa	Pré.Força	Pós.Força	Pré.ÁreaVL	Pós.ÁreaVL
Massa	643.18	1838.60	427.11	111.01	14.46
Pré.Força	1838.60	6385.67	851.46	433.95	168.09
Pós.Força	427.11	851.46	404.45	35.59	-31.85
Pré.ÁreaVL	111.01	433.95	35.59	31.19	15.58
Pós.ÁreaVL	14.46	168.09	-31.85	15.58	14.55

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

Sc: matriz de covariância comum aos grupos genotípicos

	Massa	Pré.Força	Pós.Força	Pré.ÁreaVL	Pós.ÁreaVL
Massa	151.63	304.45	305.42	20.05	19.63
Pré.Força	304.45	4051.20	3888.70	67.52	58.87
Pós.Força	305.42	3888.70	4736.83	72.21	67.51
Pré.ÁreaVL	20.05	67.52	72.21	16.41	17.30
Pós.ÁreaVL	19.63	58.87	67.51	17.30	20.13

Função discriminante:

$$l_k' Y_{i_{p\times 1}}$$

$$= l_{k1}Y_{i1} + \ldots + l_{k5}Y_{i5}$$

Solução:

$$\frac{l_k' B l_k}{l_k' S_c l_k} = \lambda_k$$

l_k : autovetor

λ_k : autovalor

Análise Discriminante -Tabela de MANOVA

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_G = \mu$$

F.V.	g.l.	Matriz de SQPC
Grupo	**G-1**	$B_{p\times p} = \sum_{g=1}^{G} n_g \left(\bar{y}_g - \bar{y}\right)\left(\bar{y}_g - \bar{y}\right)'$
Resíduo	**n-G**	$W_{p\times p} = \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(y_{ig} - \bar{y}_g\right)\left(y_{ig} - \bar{y}_g\right)'$ $\quad S_c = W_{p\times p}/(n-G)$
TOTAL	**n-1**	$B + W = \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(y_{ig} - \bar{y}\right)\left(y_{ig} - \bar{y}\right)'$

Função discriminante de Fisher: $l_k' Y_{i_{p\times 1}} = l_{k1}Y_{i1} + l_{k2}Y_{i2} + \ldots + l_{kp}Y_{ip}$ escore

Obtida da decomposição spectral: $\dfrac{l_k' B l_k}{l_k' S_c l_k} = \lambda_k$

l_k : autovetor (carga)

λ_k : autovalor

Análise Discriminante -Tabela de MANOVA

$$H_0: \mu_1 = \mu_2 = ... = \mu_G = \mu$$

F.V.	g.l.	Matriz de SQPC

Grupo **G-1**

B: Matriz de SQPC devido ao efeito de grupo (ENTRE)

	Massa	Pré.Força	Pós.Força	Pré.ÁreaVL	Pós.ÁreaVL
Massa	643.18	1838.60	427.11	111.01	14.46
Pré.Força	1838.60	6385.67	851.46	433.95	168.09
Pós.Força	427.11	851.46	404.45	35.59	-31.85
Pré.ÁreaVL	111.01	433.95	35.59	31.19	15.58
Pós.ÁreaVL	14.46	168.09	-31.85	15.58	14.55

Resíduo **n-G**

W: Matriz de SQPC devido ao efeito residual (DENTRO)

$S_c = W/(n-G)$

	Massa	Pré.Força	Pós.Força	Pré.ÁreaVL	Pós.ÁreaVL
Massa	13949.61	28009.63	28098.64	1844.58	1806.00
Pré.Força	28009.63	372710.12	357760.12	6211.88	5416.23
Pós.Força	28098.64	357760.12	435788.71	6643.46	6210.76
Pré.ÁreaVL	1844.58	6211.88	6643.46	1510.02	1591.72
Pós.ÁreaVL	1806.00	5416.23	6210.76	1591.72	1851.51

MANOVA: Testar H_0 (Inexistência de efeito de Grupo considerando "*p*" variáveis)

Diferentes estatísticas podem ser adotadas: Pillai, Wilks, Hotelling-Lawley, Roy

ANOVA: Testar H_0 (para cada uma das variáveis)

Análise Discriminante -Tabela de MANOVA

MANOVA - Teste de H_0 (Igualdade das médias dos grupos: p=5)

Método	Df	Estat	approxF	numDf	denDf	Pr(>F)
Pillai	2	0.16743	1.6263	10	178	0.1023
Wilks	2	0.83646	1.6437	10	176	0.0977
Hotelling-Lawley	2	0.19085	1.6604	10	174	0.0935
Roy	2	0.16209	2.8853	5	89	**0.0184**

Há pelo menos uma diferença entre os grupos para alguma combinação linear entre as variáveis

ANOVA - Teste de H_0 (Igualdade das médias dos grupos para cada variável: p=1)

Variável	SQGrupo	SQRes	F	valor-p
Massa	643.2	13949.6	2.121	0.1257
Pré.Força	6385.7	372710.1	0.7881	0.4577
Pós.Força	404.4	435788.7	0.0427	0.9582
Pré.ÁreaVL	31.2	1510.0	0.9503	0.3904
Pós.ÁreaVL	14.5	1851.5	0.3614	0.6977
Grau Liberd	2	92		

Individualmente, para cada variável, não há evidência de efeito de Grupo.

Análise Discriminante

Probabilidades a Priori dos Grupos genotípicos:

```
        1         2         3
0.1789474 0.5263158 0.2947368
```

Médias das variáveis por grupo:

```
       Massa  Pré.Força  Pós.Força Pré.ÁreaVL Pós.ÁreaVL
1   75.24706   267.0588   334.7059   26.11594   29.93076
2   78.25400   267.1000   339.5000   25.75210   29.03864
3   72.31786   249.1071   335.8929   24.62186   28.79032
```

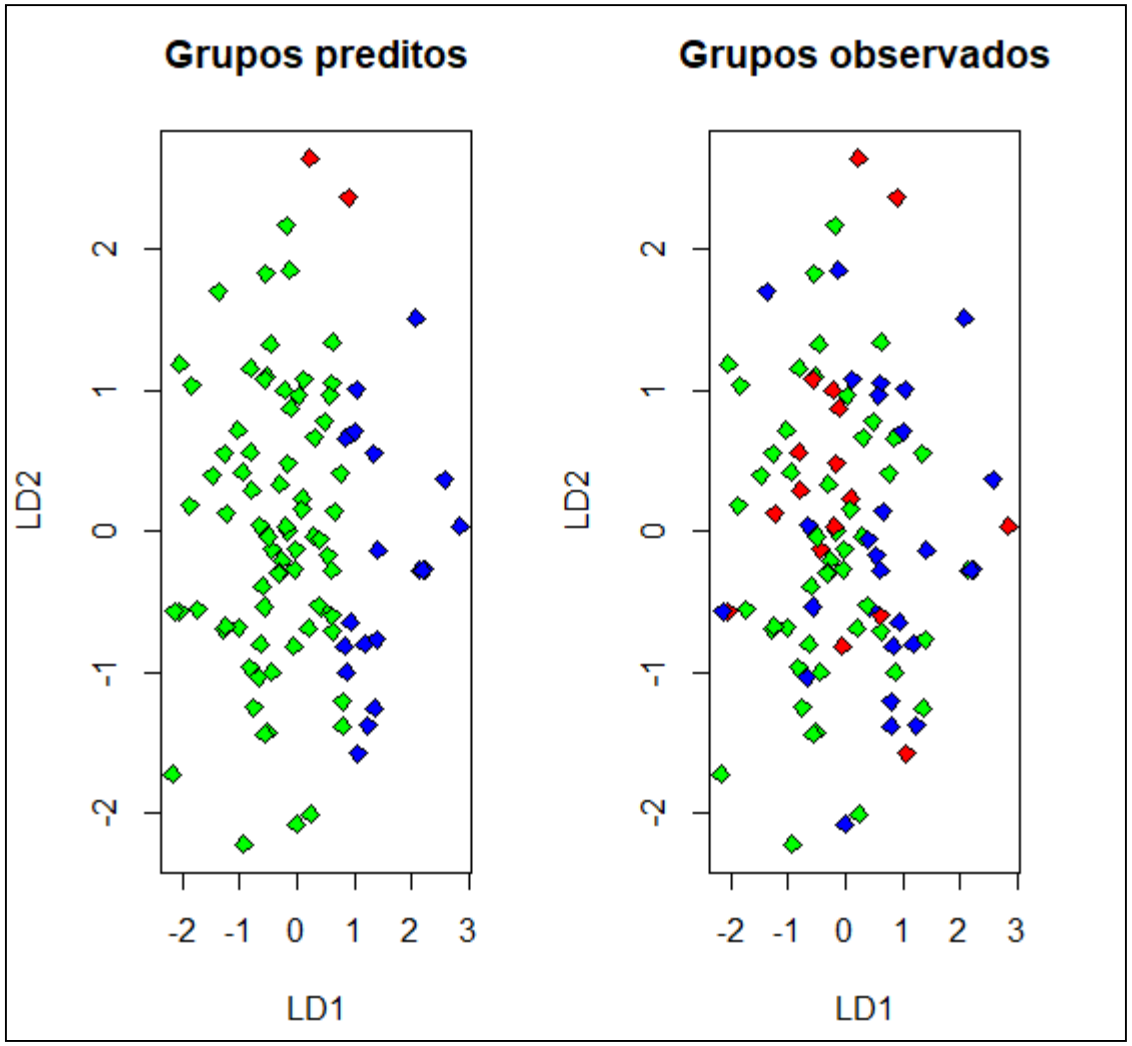
Coeficientes discriminantes (cargas): #FD ≤ min(n,p,G-1)=2

```
                    LD1         LD2
Massa       -0.03251980 -0.04252504
Pré.Força   -0.01439875  0.02519128
Pós.Força    0.01468766 -0.02127497
Pré.ÁreaVL -0.55904290 -0.20243173
Pós.ÁreaVL  0.48865084  0.32290125
```

Proporção Explicada pelas funções discriminantes:

```
   LD1    LD2
0.8493 0.1507
```

Análise Discriminante



Red=XX Green=RX Blue=RR

ACTN3

	Predito		
	1	**2**	**3**
1	2	13	2
2	0	44	6
3	0	17	11

%ClassCorreta

1	2	3	
11.8	88.0	39.3	60.0

Validação Cruzada

	1	2	3
1	0	14	3
2	1	41	8
3	0	18	10

%ClassCorreta-CV

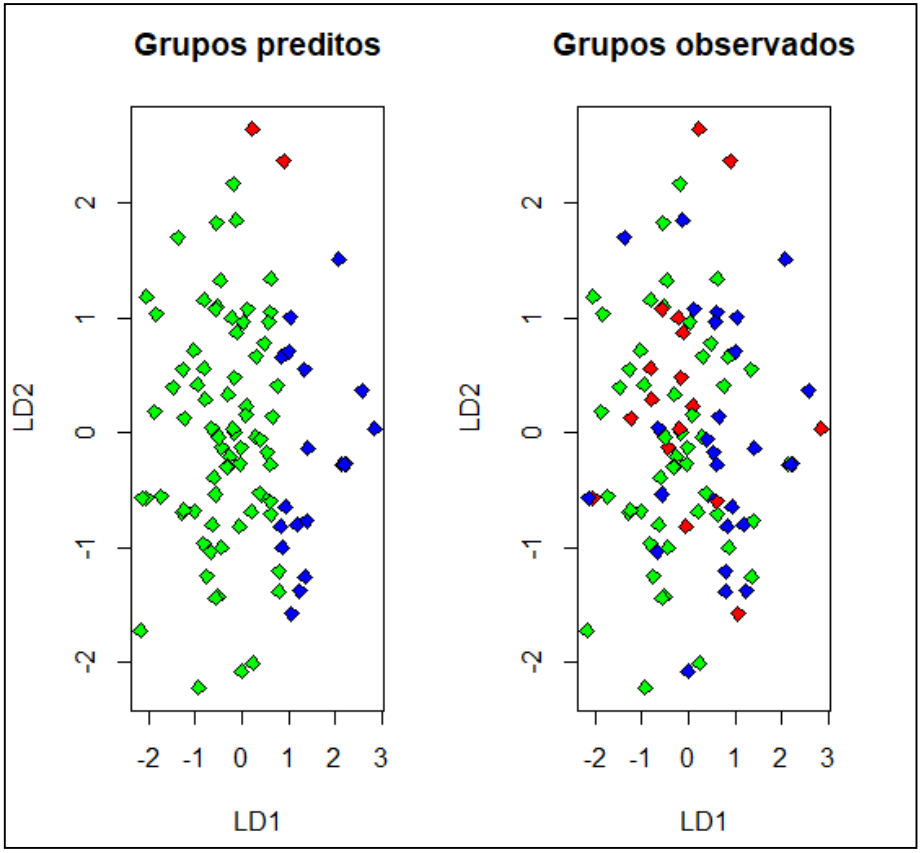
1	2	3	
0.0	82.0	35.7	53.7

Análise Discriminante

Revisando!

Supondo Prioris Amostrais (%)

```
   1    2    3
17.9 52.6 29.5
```

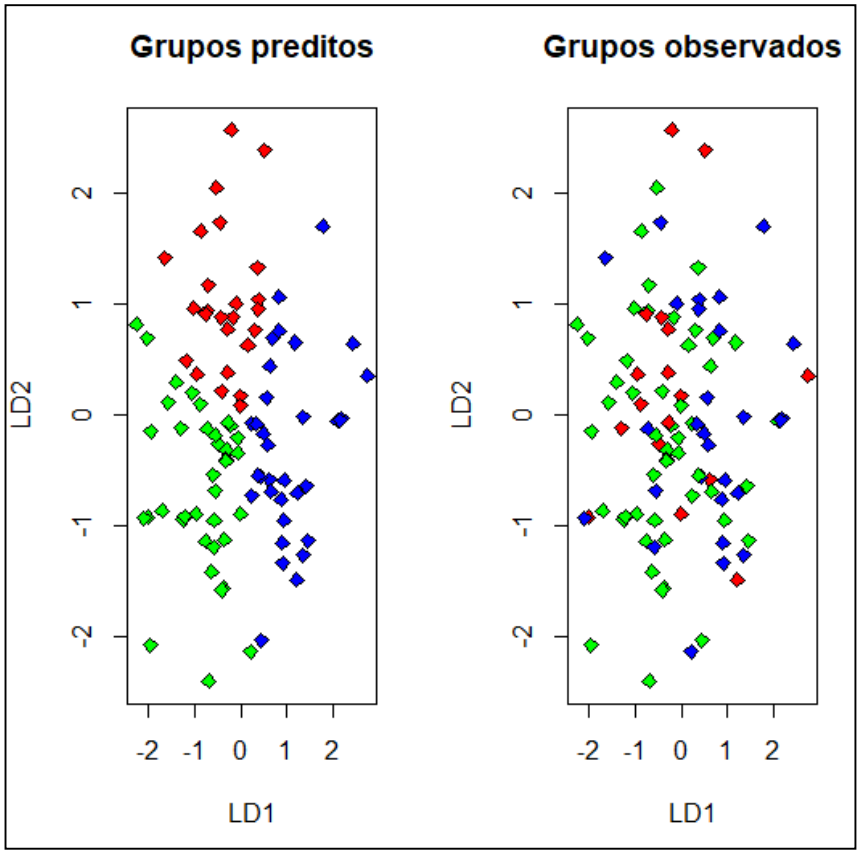


ACTN3
%ClassCorreta-CV

```
   1    2    3
 0.0 82.0 35.7 53.7
```

Supondo Prioris Iguais (%)

```
   1    2    3
33.3 33.3 33.3
```



ACTN3
%ClassCorreta-CV

```
   1    2    3
29.4 46.0 64.3 48.4
```

Análise de Correlação Canônica

Análise de Correlação Canônica

Análise Não-Supervisionada

	Variáveis						
Unidades Amostrais	Y1	Y2	…	Yp			Y(p+q)
1	Y_{11}	Y_{12}		Y_{1p}			$Y_{1(p+q)}$
2	Y_{21}	Y_{22}		Y_{2p}			$Y_{2(p+q)}$
…	…	…	…	…		…	…
n	Y_{n1}	Y_{n2}		Y_{np}			$Y_{n(p+q)}$

Objetivo:

- Estudar o relacionamento (integração) ENTRE dois “conjuntos de variáveis” (p+q)

ANÁLISE DE “CORRELAÇÃO CANÔNICA”

$\Rightarrow$ Obter Variáveis Canônicas (escores, var. latentes, vetores reducionistas) de cada subconjunto das variáveis originais, com máxima correlação.

$\Rightarrow$ Realizar a integração de dois bancos de dados.

Correlação entre Conjuntos de Variáveis
Motivação

Morfometria cefálica para os dois primeiros filhos de 25 famílias (Everitt, 2007)

	1° Filho		2° Filho	
Família	Comprimento	Perímetro	Comprimento	Perímetro
1	191	155	179	145
2	195	149	201	152
3	181	148	185	149
4	183	153	188	149
5	176	144	171	142
6	208	157	192	152
7	189	150	190	149
8	197	159	189	152
9	188	152	197	159
10	192	150	187	151
11	179	158	186	148
12	183	147	174	147
13	174	150	185	152
14	190	159	195	157
15	188	151	187	158
16	163	137	161	130
17	195	155	183	158
18	186	153	173	148
19	181	145	182	146
20	175	140	165	137
21	192	154	185	152
22	174	143	178	147
23	176	139	176	143
24	197	167	200	158
25	190	163	187	150

Como relacionar os irmãos com base em ambas medidas cefálicas?

Como definir uma medida de correlação (escalar) para o caso multidimensional?

Discuta a estrutura dos dados.

Neste caso, tem-se as mesmas variáveis (comprimento e perímetro) avaliadas em cada nível de um fator de estratificação (1° e 2° filhos). As famílias definem o pareamento ou dependência entre os dois conjuntos.

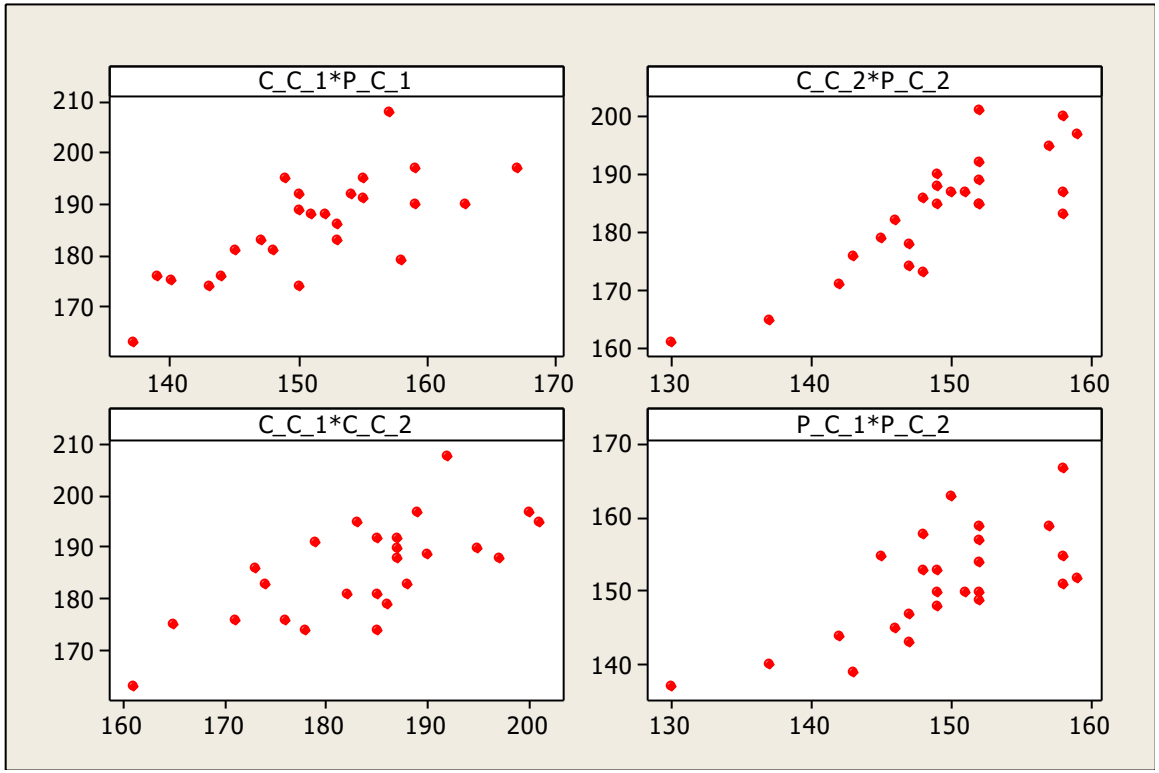
A análise se estende para situações de dois conjuntos de variáveis diferentes!

Diferentes Medidas de Correlação

Coeficientes de Correlação Linear de Pearson para os dados de morfometria cefálica:

r	C_C_1	P_C_1	C_C_2
P_C_1	0,735		
C_C_2	0,711	0,693	
P_C_2	0,704	0,709	0,839

Correlações de menor interesse.



Correlação entre as variáveis DENTRO do grupo.

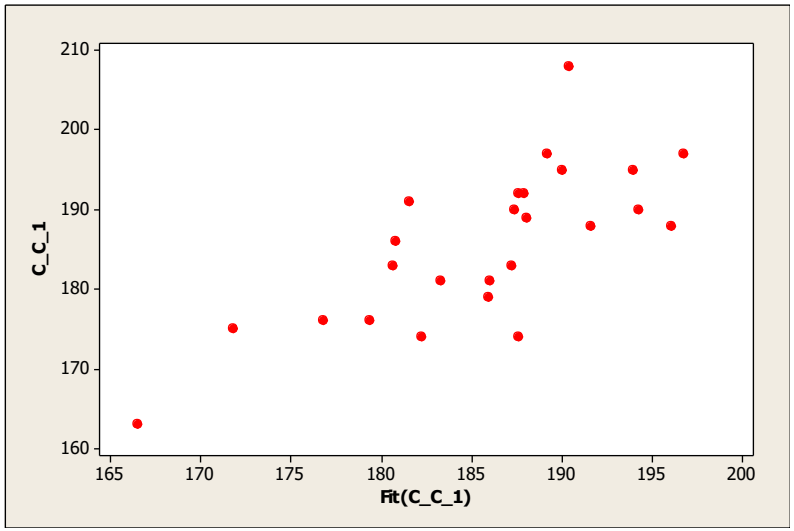
Correlação ENTRE os grupos, para cada variável.

Diferentes Medidas de Correlação

Coeficiente de Correlação Múltipla

$\Rightarrow$ É a correlação linear de Pearson entre cada variável de um conjunto e seu preditor linear (função das variáveis do outro conjunto).

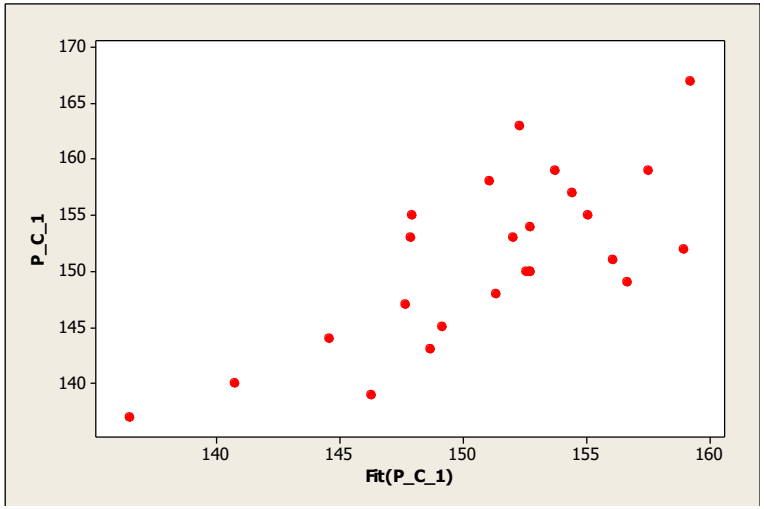
$$\rho_M\left[Y_{C\_C\_1}, \left(Y_{C\_C\_2}, Y_{P\_C\_2}\right)\right]$$



$$\rho_P\left(Y_{C\_C\_1}, \hat{Y}_{C\_C\_1}\right) = 0{,}738$$

$$Y_{C\_C\_1} = \beta_0 + \beta_1 Y_{C\_C\_2} + \beta_2 Y_{P\_C\_2} + e$$

$$\rho_M\left[Y_{P\_C\_1}, \left(Y_{C\_C\_2}, Y_{P\_C\_2}\right)\right]$$



$$\rho_P\left(Y_{P\_C\_1}, \hat{Y}_{P\_C\_1}\right) = 0{,}731$$

$$Y_{P\_C\_1} = \beta_0 + \beta_1 Y_{C\_C\_2} + \beta_2 Y_{P\_C\_2} + e$$

Diferentes Medidas de Correlação

Coeficiente de Correlação Parcial

$\Rightarrow$ Considere a distribuição condicional de vetores de variáveis aleatórias

$$Y_{1p\times 1};\quad E\left(Y_{1p\times 1}\right)=\mu_1 \quad Cov\left(Y_{1p\times 1}\right)=\Sigma_{11p\times p} \qquad Y_{2q\times 1};\quad E\left(Y_{2q\times 1}\right)=\mu_2 \quad Cov\left(Y_{2q\times 1}\right)=\Sigma_{22q\times q}$$

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix};\quad E\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}=\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix};\quad Cov\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}=\Sigma_{(p+q)\times(p+q)}=\begin{bmatrix} \Sigma_{11p\times p} & \Sigma_{12p\times q} \\ \Sigma_{21q\times p} & \Sigma_{22q\times q} \end{bmatrix}$$

$$E\left(Y_2 \mid Y_1\right)=\mu_2-\Sigma_{21}\Sigma_{11}^{-1}\left(Y_1-\mu_1\right) \qquad Cov\left(Y_2 \mid Y_1\right)=\Sigma_{22.1}=\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

Correlação entre Y_{2j} e Y_{2k}, eliminando o efeito das variáveis $Y_1=(Y_{11}, \ldots, Y_{1q})$:

$$\rho\left(Y_{2j}, Y_{2k} \mid Y_1\right)=\frac{\sigma_{jk.1}}{\sqrt{\sigma_{jj.1}}\sqrt{\sigma_{kk.1}}};$$ $\sigma_{jk.1}$ é a casela jk da matriz $\Sigma_{22.1}$

Pode ser obtido de matrizes de precisão (Σ^{-1}).

Outra medida de correlação:

Correlação Canônica - Exemplos

- Relacionar variáveis da mãe com variáveis do recém-nascido.
- Relacionar variáveis do sedimento com variáveis da coluna de água de um rio, considerando vários pontos de coleta.
- Relacionar variáveis clínicas com variáveis do genoma.
- Relacionar variáveis do Genoma com variáveis do Transcriptoma
- …

Integração de bancos de dados!

Correlação Canônica

Aplicação

Projeto ACTN3 e Força (dados de atletas)

Atleta	ACTN3	Massa	Pré.Força	Pré.ÁreaVL	Pós.Força	Pós.ÁreaVL
1	1	73.0	210	27.133	260	31.398
2	1	78.0	260	23.841	320	26.950
3	3	70.0	220	23.755	320	28.937
4	2	78.5	280	26.241	360	28.794
...						
94	2	95.0	175	25.120	260	28.605
95	2	71.0	220	25.452	330	29.029

$$Y_{95\times(3+2)} = \left(Y_{1\,95\times3} \quad Y_{2\,95\times2} \right)$$

Obtenha as variáveis canônicas relacionando as variáveis dos dois conjuntos de dados (Pré e Pós).

Correlação Canônica

Notação

Dados de um vetor de variáveis aleatórias particionado em Dois Conjuntos de Variáveis:

$$Y_{n\times(p+q)} = \left(Y_{1n\times p} \quad Y_{2n\times q}\right); \qquad Y_{i\,(p+q)\times 1} \overset{iid}{\sim} (\mu;\Sigma)$$

$$Y_{i\,(p+q)\times 1} = \begin{bmatrix} Y_{1i\,p\times 1} \\ Y_{2i\,q\times 1} \end{bmatrix}$$

$$E\left(Y_{1i\,p\times 1}\right) = \mu_1 \quad Cov\left(Y_{1i\,p\times 1}\right) = \Sigma_{11\,p\times p}$$

$$E\left(Y_{2i\,q\times 1}\right) = \mu_2 \quad Cov\left(Y_{2i\,q\times 1}\right) = \Sigma_{22\,q\times q}$$

$$Cov\left(Y_{1i\,p\times 1}, Y_{2i\,q\times 1}\right) = \Sigma_{12\,p\times q} = \Sigma'_{21q\times p}$$

Mede a covariância entre os dois conjuntos de variáveis

$$E\left(Y_i\right) = \mu_{(p+q)\times 1} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad Cov\left(Y_i\right) = \Sigma_{(p+q)\times(p+q)} = \begin{bmatrix} \Sigma_{11\,p\times p} & \Sigma_{12\,p\times q} \\ \Sigma_{21q\times p} & \Sigma_{22\,q\times q} \end{bmatrix}$$

Correlação Canônica

Como Resumir "Correlações" entre Dois Conjuntos de Variáveis?

Obter combinações lineares de cada conjunto!

$$U_i = a' Y_{1i} \qquad V_i = b' Y_{2i}$$

$$\begin{cases} Var(U_i) = a' \Sigma_{11} a \qquad Var(V_i) = b' \Sigma_{22} b \\ Cov(U_i, V_i) = a' \Sigma_{12} b \end{cases}$$

Obter vetores $\boldsymbol{a} \in \Re^p$ e $\boldsymbol{b} \in \Re^q$, tal que (independentemente, de i):

$$Corr(U,V) = \frac{Cov(U,V)}{\sqrt{Var(U)}\sqrt{Var(V)}} = \frac{a' \Sigma_{12} b}{\sqrt{a' \Sigma_{11} a}\sqrt{b' \Sigma_{22} b}}$$ seja máxima.

$\Rightarrow$ Encontrar o primeiro par de combinações lineares, U_1 e V_1, padronizadas (variâncias unitárias), que maximizam a correlação canônica definida acima.

$\Rightarrow$ Caso seja de interesse, encontrar o segundo par de variáveis padronizadas, U_2 e V_2, que maximizem a correlação canônica entre todas as escolhas não correlacionadas com o primeiro par $\Rightarrow$ e assim por diante até *m=min(n,p,q)*.

Correlação Canônica

$$\left.\begin{array}{l} U = a' Y_1 \\ V = b' Y_2 \end{array}\right\} \quad \max_{a,b} Corr(U,V) = \max_{a,b} \frac{a' \Sigma_{12} b}{\sqrt{a' \Sigma_{11} a}\sqrt{b' \Sigma_{22} b}}$$

equivale a maximizar:

$$\Rightarrow \max_{a \in \Re^p} \frac{a' \Sigma_{12}\Sigma_{22}^{-1} \Sigma_{21} a}{a' \Sigma_{11} a} \qquad \Rightarrow \max_{b \in \Re^q} \frac{b' \Sigma_{21}\Sigma_{11}^{-1} \Sigma_{12} b}{b' \Sigma_{22} b}$$

Solução: O $\max_{a,b} Corr(U,V) = \rho_{c1}$ é atingido pelo primeiro par de combinações lineares, dado por: (Mardia, 1979)

$$U_1 = \underbrace{e_1' \Sigma_{11}^{-1/2}}_{a_1'} Y_1 \qquad V_1 = \underbrace{f_1' \Sigma_{22}^{-1/2}}_{b_1'} Y_2$$

Os escores U e V são obtidos a partir de projeções que compartilham os mesmos autovalores. "a" e "b" são os coeficientes (cargas) da variável canônica.

$[Corr(U,V)]^2$

$\Rightarrow \rho_{c1}^2$ e e_1 são o maior autovalor e o autovetor de $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$

$\Rightarrow \rho_{c1}^2$ e f_1 são o maior autovalor e o autovetor de $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$

Correlação Canônica

$$\max_{a,b} Corr\left(U,V\right) = \rho_{c1} \quad \Rightarrow \quad \begin{aligned} U_1 &= a_1' Y_1 = e_1' \, \Sigma_{11}^{-1/2} \, Y_1 \\ V_1 &= b_1' Y_2 = f_1' \, \Sigma_{22}^{-1/2} \, Y_2 \end{aligned}$$

O k-ésimo par de variáveis canônicas (com k=1,2,…,min(n,p,q)) representam as combinações lineares de cada conjunto com máxima correlação e independente das demais:.

$$U_k = e_k' \, \Sigma_{11}^{-1/2} \, Y_1, \qquad V_k = f_k' \, \Sigma_{22}^{-1/2} \, Y_2; \qquad Corr\left(U_k, V_k\right) = \rho_{ck}$$

k-ésimo coeficiente de correlação canônico.

$$\Re^{(p+q)} \rightarrow \Re^{(m+m)}; m \leq \min(n, p, q)$$

Critério de redução de dimensionalidade com o compromisso de maximizar a correlação entre os conjuntos de dados.

Correlação Canônica

Solução: $\max_{a,b} Corr\left(U_1, V_1\right) = \rho_{c1}$ é atingido pelo primeiro par de variáveis canônicas, dado por

$$U_1 = a_1' Y_1 = e_1' \, \Sigma_{11}^{-1/2} \, Y_1 \qquad V_1 = b_1' Y_2 = f_1' \, \Sigma_{22}^{-1/2} \, Y_2$$

$\Rightarrow \lambda_1$ e e_1 são o maior autovalor e seu autovetor de $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$

$\Rightarrow \lambda_1$ e f_1 são o maior autovalor e seu autovetor de $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$

As demais variáveis canônicas $(U_2, V_2), \ldots, (U_k, V_k), \ldots, (U_m, V_m)$ satisfazem:

$$\begin{cases} Var(U_k) = Var(V_k) = 1 \\ Cov(U_k, U_l) = Corr(U_k, U_l) = 0 \quad k \neq l \\ Cov(V_k, V_l) = Corr(V_k, V_l) = 0 \quad k \neq l \\ Cov(U_k, V_l) = Corr(U_k, V_l) = 0 \quad k \neq l \end{cases}$$

$$\Rightarrow \quad Cov(U, V) = \begin{pmatrix} I_m & \Lambda^{1/2} \\ \Lambda^{1/2} & I_m \end{pmatrix};$$

$$\Lambda^{1/2} = \left(\sqrt{\lambda_j} = \rho_{cj}\right)$$

Correlação Canônica

Considere as variáveis padronizadas:

$$Y_i = \begin{bmatrix} Y_{1i\,(p\times 1)} \\ Y_{2i\,(q\times 1)} \end{bmatrix} \Rightarrow Y^*_{i\ (p+q)\times 1} = \begin{bmatrix} Y^*_{1i\ p\times 1} \\ Y^*_{2i\,q\times 1} \end{bmatrix} = \begin{bmatrix} D_{11}^{-1/2}\left(Y_{1i} - \mu_1\right) \\ D_{22}^{-1/2}\left(Y_{2i} - \mu_2\right) \end{bmatrix}$$

$\Rightarrow$ As variáveis canônicas são da forma:

$$\left.\begin{array}{l} U_k^* = a_k^{*\prime} Y_1^* = e_k^{*\prime}\ R_{11}^{-1/2}\ Y_1^* \\ V_k^* = b_k^{*\prime} Y_2^* = f_k^{*\prime}\ R_{22}^{-1/2} Y_2^* \end{array}\right\} \quad Corr\left(U_k^*, V_k^*\right) = \frac{a_k^{*\prime}\ \rho_{12}\ b_k^*}{\sqrt{a_k^{*\prime}\ \rho_{11}\ a_k^*}\sqrt{b_k^{*\prime}\ \rho_{22}\ b_k^*}} = \rho_{ck}$$

As correlações canônicas são invariantes por padronização dos dados!

$$\rho_{ck} = \sqrt{\lambda_k} = \sqrt{\lambda_k^*}$$

$\Rightarrow$ λ_k^* , e_k^* : k-ésimo autovalor e autovetor de $R_{11}^{-1/2} R_{12} R_{22}^{-1} R_{21} R_{11}^{-1/2}$

$\Rightarrow$ λ_k^* , f_k^* : k-ésimo autovalor e autovetor de $R_{22}^{-1/2} R_{21} R_{11}^{-1} R_{12} R_{22}^{-1/2}$

Correlação Canônica

Relação entre as Variáveis Canônicas obtidas das Variáveis Originais e das Variáveis Padronizadas

Variáveis Originais

$$Y_{(p+q)\times 1} = \begin{bmatrix} Y_{1\,p\times 1} \\ Y_{2\,q\times 1} \end{bmatrix}$$

$$U_k = a_k' Y_1 = e_k' \; \Sigma_{11}^{-1/2} \; Y_1$$

$$V_k = b_k' Y_2 = f_k' \; \Sigma_{22}^{-1/2} \; Y_2$$

$\Rightarrow$

Variáveis Padronizadas

$$Y_{i\ (p+q)\times 1}^{*} = \begin{bmatrix} Y_{1i\ p\times 1}^{*} \\ Y_{2i\,q\times 1}^{*} \end{bmatrix} = \begin{bmatrix} D_{11}^{-1/2}\left(Y_{1i} - \mu_1\right) \\ D_{22}^{-1/2}\left(Y_{2i} - \mu_2\right) \end{bmatrix}$$

$$U_k^{*} = a_k^{*}{}' Y_1^{*} = e_k^{*}{}' \; R_{11}^{-1/2} \; Y_1^{*}$$

$$V_k^{*} = b_k^{*}{}' Y_2^{*} = f_k^{*}{}' \; R_{22}^{-1/2} Y_2^{*}$$

$$a_k' Y_1 = a_k' \left(Y_1 - \mu_1\right) = a_{k1}\left(Y_{11} - \mu_{11}\right) + \ldots + a_{kp}\left(Y_{1p} - \mu_{1p}\right)$$

$$= a_{k1}\sqrt{\sigma_{11}}\,\frac{\left(Y_{11} - \mu_{11}\right)}{\sqrt{\sigma_{11}}} + \ldots + a_{kp}\sqrt{\sigma_{pp}}\,\frac{\left(Y_{1p} - \mu_{1p}\right)}{\sqrt{\sigma_{pp}}}$$

$$= a_{k1}^{*} Y_{11}^{*} + \ldots + a_{kp}^{*} Y_{11}^{*} = a_k^{*}{}' Y_1^{*}$$

$\Rightarrow$

$$a_k^{*}{}' = a_k{}' D_{11}^{1/2}$$

$\Rightarrow$

$$b_k^{*}{}' = b_k{}' D_{22}^{1/2}$$

$$Y_{(p+q)\times 1} = \begin{bmatrix} Y_{1\,p\times 1} \\ Y_{2\,q\times 1} \end{bmatrix} \quad \Rightarrow \quad Y^*_{i\ (p+q)\times 1} = \begin{bmatrix} Y^*_{1i\ p\times 1} \\ Y^*_{2i\,q\times 1} \end{bmatrix} = \begin{bmatrix} D_{11}^{-1/2}\left(Y_{1i} - \mu_1\right) \\ D_{22}^{-1/2}\left(Y_{2i} - \mu_2\right) \end{bmatrix}$$

$$U_k = a_k' Y_1 \qquad U^*_k = a_k^{*\prime} Y_1^* = a_k' D_{11}^{1/2} Y_1^*$$

$$V_k = b_k' Y_2 \qquad V_k^* = b_k^{*\prime} Y_2^* = b_k' D_{22}^{1/2} Y_2^*$$

$$\rho_c\left(U_k^*, V_k^*\right) = \frac{a_k^{*\prime} R_{12} b_k^*}{\sqrt{a_k^{*\prime} R_{11} a_k^*}\sqrt{b_k^{*\prime} R_{22} b_k^*}} = a_k^{*\prime} R_{12} b_k^* = a_k' D_{11}^{1/2} Corr\left(Y_1^*, Y_2^*\right) D_{22}^{1/2} b_k$$

$$= a_k' D_{11}^{1/2} Corr\left(D_{11}^{-1/2}\left(Y_1 - \mu_1\right), D_{22}^{-1/2}\left(Y_2 - \mu_2\right)\right) D_{22}^{1/2} b_k$$

$$= a_k' Corr\left(\left(Y_1 - \mu_1\right), \left(Y_2 - \mu_2\right)\right) b_k = a_k' Corr\left(Y_1, Y_2\right) b_k = \rho_c\left(U_k, V_k\right)$$

A correlação canônica é invariante por padronização

- Os coeficientes canônicos das variáveis padronizadas podem ser obtidos diretamente dos coeficientes das variáveis originais
- O coeficiente de correlação canônico das variáveis originais e das variáveis padronizadas é o mesmo (invariantes por padronização dos dados)

Correlação Canônica

Interpretação Geométrica

$$\max_{a,b} Corr(U,V) = \rho_{c1} \quad \Rightarrow \quad U_1 = a_1'Y_1 = e_1' \, \Sigma_{11}^{-1/2} \, Y_1$$

$$V_1 = b_1'Y_2 = f_1' \, \Sigma_{22}^{-1/2} \, Y_2$$

$$U_1 = a_1'Y_1 = e_1' \underbrace{\Sigma_{11}^{-1/2}}_{P_1 \, \Lambda^{-1/2} \, P_1'} Y_1 = e_1' \, P_1 \, \Lambda^{-1/2} \, P_1' \, Y_1$$

Decomposição spectral de Σ_{11}

Componente Principal de Y_1

Fator Comum de Y_1 (CP padronizado)

A variável canônica U_1 resulta de uma rotação orthogonal (via P_1 e determinada por Σ_{11}) do CP padronizado seguida por outra rotação orthogonal (via e_1 e determinada por $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$)

Correlação Canônica

Morfometria cefálica para os dois primeiros filhos de 25 famílias

	1° Filho		2° Filho	
Família	Comprimento	Perímetro	Comprimento	Perímetro
1	191	155	179	145
2	195	149	201	152
3	181	148	185	149
4	183	153	188	149
5	176	144	171	142
6	208	157	192	152
7	189	150	190	149
8	197	159	189	152
9	188	152	197	159
10	192	150	187	151
11	179	158	186	148
12	183	147	174	147
13	174	150	185	152
14	190	159	195	157
15	188	151	187	158
16	163	137	161	130
17	195	155	183	158
18	186	153	173	148
19	181	145	182	146
20	175	140	165	137
21	192	154	185	152
22	174	143	178	147
23	176	139	176	143
24	197	167	200	158
25	190	163	187	150
Média	185,72	151,12	183,84	149,24
Var.	95,29	54,36	100,81	45,02

Obtenha as variáveis canônicas das variáveis padronizadas.

Interprete os resultados.

Correlação Canônica

Morfometria cefálica para os dois primeiros filhos de 25 famílias

Considere a análise de Correlação Canônica das Variáveis Padronizadas:

$$R_{11} = \begin{pmatrix} 1 & 0{,}73456 \\ 0{,}73456 & 1 \end{pmatrix} \qquad R_{22} = \begin{pmatrix} 1 & 0{,}83925 \\ 0{,}83925 & 1 \end{pmatrix}$$

$$R_{12} = \begin{pmatrix} 0{,}7108 & 0{,}704 \\ 0{,}6932 & 0{,}7086 \end{pmatrix}$$

Todas as correlações são altas $\Rightarrow \lambda_2 \cong 0$

Autovalores: 0,6218 0,0029 $\Rightarrow$ $\hat{\rho}^*_{c1} = \sqrt{0,6218} = 0,7886$ $\hat{\rho}^*_{c2} = 0,0539$

Coeficientes das Variáveis canônicas:

$$A^*_{2\times 2} = \begin{pmatrix} a^*_1 & a^*_2 \end{pmatrix} \quad a^*_1 = \begin{pmatrix} 0{,}552 \\ 0{,}522 \end{pmatrix} \quad a^*_2 = \begin{pmatrix} 1{,}367 \\ -1{,}378 \end{pmatrix}$$

$$B^*_{2\times 2} = \begin{pmatrix} b^*_1 & b^*_2 \end{pmatrix} \quad b^*_1 = \begin{pmatrix} 0{,}505 \\ 0{,}538 \end{pmatrix} \quad b^*_2 = \begin{pmatrix} 1{,}767 \\ -1{,}757 \end{pmatrix}$$

Correlação Canônica

Morfometria cefálica para os dois primeiros filhos de 25 famílias

Se somente a primeira variável canônica (das variáveis padronizadas) é usada, temos:

$$U_1^* = 0{,}552\, Y^*_{C\_C\_1} + 0{,}522\, Y^*_{P\_C\_1} \qquad V_1^* = 0{,}505\, Y^*_{C\_C\_2} + 0{,}538\, Y^*_{P\_C\_2}$$

Estas são responsáveis pela maior correlação (r=0,79) entre as variáveis cefálicas dos dois primeiros filhos das famílias estudadas. As variáveis individuais contribuem com “pesos” muito próximos.

A segunda variável canônica explica muito pouco (r=0,05) da correlação entre as variáveis dos dois primeiros filhos, sendo definida por:

$$U_2^* = 1{,}367\, Y^*_{C\_C\_1} - 1{,}378\, Y^*_{P\_C\_1} \qquad V_2^* = 1{,}767\, Y^*_{C\_C\_2} - 1{,}757\, Y^*_{P\_C\_2}$$

Correlação Canônica

Morfometria cefálica para os dois primeiros filhos de 25 famílias

Análise de Correlação Canônica das Variáveis Padronizadas:

$$U_1^* = 0{,}552\, Y^*_{C\_C\_1} + 0{,}522\, Y^*_{P\_C\_1}$$

$$V_1^* = 0{,}505\, Y^*_{C\_C\_2} + 0{,}538\, Y^*_{P\_C\_2}$$

$$\hat{\rho}_1^* = Corr\left(U_1^*, V_1^*\right) = 0{,}79$$

Análise de Correlação Canônica das Variáveis Originais:

$$\Rightarrow a_1 = a_1^{*\prime} D_{11}^{-1/2} = \begin{pmatrix} 0{,}552 & 0{,}522 \end{pmatrix} \begin{pmatrix} 1/\sqrt{95{,}29} & 0 \\ 0 & 1/\sqrt{54{,}36} \end{pmatrix} = \begin{pmatrix} 0{,}057 & 0{,}071 \end{pmatrix}$$

$$\Rightarrow b_1 = b_1^{*\prime} D_{22}^{-1/2} = \begin{pmatrix} 0{,}505 & 0{,}538 \end{pmatrix} \begin{pmatrix} 1/\sqrt{100{,}81} & 0 \\ 0 & 1/\sqrt{45{,}02} \end{pmatrix} = \begin{pmatrix} 0{,}050 & 0{,}080 \end{pmatrix}$$

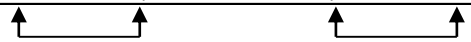
$$U_1 = 0{,}057\, Y_{C\_C\_1} + 0{,}071\, Y_{P\_C\_1}$$

$$V_1 = 0{,}050\, Y_{C\_C\_2} + 0{,}080\, Y_{P\_C\_2}$$

$$\hat{\rho}_1 = Corr\left(U_1, V_1\right) = 0{,}79$$

Correlação Canônica

Variáveis originais				Variáveis padronizadas				Variáveis canônicas			
Y_CC1	Y_PC1	Y_CC2	Y_PC2	Y*_CC1	Y*_PC1	Y*_CC2	Y*_PC2	U*1	V*1	U1	V1
191	155	179	145	0,541	0,526	-0,482	-0,632	0,573	-0,583	21,892	20,550
195	149	201	152	0,951	-0,288	1,709	0,411	0,375	1,084	21,694	22,210
181	148	185	149	-0,484	-0,423	0,116	-0,036	-0,488	0,039	20,825	21,170
183	153	188	149	-0,279	0,255	0,414	-0,036	-0,021	0,190	21,294	21,320
176	144	171	142	-0,996	-0,966	-1,279	-1,079	-1,054	-1,226	20,256	19,910
208	157	192	152	2,282	0,798	0,813	0,411	1,676	0,632	23,003	21,760
189	150	190	149	0,336	-0,152	0,614	-0,036	0,106	0,291	21,423	21,420
197	159	189	152	1,156	1,069	0,514	0,411	1,196	0,481	22,518	21,610
188	152	197	159	0,234	0,119	1,311	1,455	0,191	1,444	21,508	22,570
192	150	187	151	0,643	-0,152	0,315	0,262	0,276	0,300	21,594	21,430
179	158	186	148	-0,688	0,933	0,215	-0,185	0,107	0,009	21,421	21,140
183	147	174	147	-0,279	-0,559	-0,980	-0,334	-0,446	-0,675	20,868	20,460
174	150	185	152	-1,201	-0,152	0,116	0,411	-0,742	0,280	20,568	21,410
190	159	195	157	0,438	1,069	1,112	1,156	0,800	1,184	22,119	22,310
188	151	187	158	0,234	-0,016	0,315	1,306	0,120	0,861	21,437	21,990
163	137	161	130	-2,327	-1,915	-2,275	-2,867	-2,284	-2,691	19,018	18,450
195	155	183	158	0,951	0,526	-0,084	1,306	0,799	0,660	22,120	21,790
186	153	173	148	0,029	0,255	-1,080	-0,185	0,149	-0,645	21,465	20,490
181	145	182	146	-0,484	-0,830	-0,183	-0,483	-0,700	-0,352	20,612	20,780
175	140	165	137	-1,098	-1,508	-1,876	-1,824	-1,393	-1,929	19,915	19,210
192	154	185	152	0,643	0,391	0,116	0,411	0,559	0,280	21,878	21,410
174	143	178	147	-1,201	-1,101	-0,582	-0,334	-1,238	-0,473	20,071	20,660
176	139	176	143	-0,996	-1,644	-0,781	-0,930	-1,408	-0,895	19,901	20,240
197	167	200	158	1,156	2,154	1,610	1,306	1,762	1,515	23,086	22,640
190	163	187	150	0,438	1,611	0,315	0,113	1,083	0,220	22,403	21,350



r(U*1,V*1) = 0,789 r(U1,V1) = 0,789

Correlação Canônica

Propriedades das Variáveis Canônicas (*min(n,p,q)*)

- Variâncias Unitárias: $Var(U_k) = Var(V_k) = 1$
- Não Correlacionadas (Entre pares): $Corr(U_k, U_l) = Corr(V_k, V_l) = Corr(U_k, V_l) = 0$
- Correlação Máxima (Dentro do par): $Corr(U_k, V_k) = \rho_{ck} = \sqrt{\lambda_k}$

- Correlação entre as Variáveis Canônicas e as Variáveis Originais: $(A_{p\times m}; B_{q\times m})$

$$U_{i\,m\times 1} = A'Y_{1i}$$

$$V_{i\,m\times 1} = B'Y_{2i}$$

$$Corr(U;Y_1) = A'\Sigma_{11}D_{11}^{-1/2} = A^{*\prime}R_{11} = Corr(U^*, Y_1^*)$$

$$Corr(U;Y_2) = A'\Sigma_{12}D_{22}^{-1/2} = A^{*\prime}R_{12} = Corr(U^*, Y_2^*)$$

$$Corr(V;Y_1) = B'\Sigma_{21}D_{11}^{-1/2} = B^{*\prime}R_{21} = Corr(V^*, Y_1^*)$$

$$Corr(V;Y_2) = B'\Sigma_{22}D_{22}^{-1/2} = B^{*\prime}R_{22} = Corr(V^*, Y_2^*)$$

Na prática, calcular a correlação de Pearson entre essas variáveis!

Correlação Canônica

Morfometria cefálica para os dois primeiros filhos de 25 famílias

$$A^* = \begin{pmatrix} 0,552 & 0,522 \\ 1,367 & -1,378 \end{pmatrix}' \qquad B^* = \begin{pmatrix} 0,505 & 0,538 \\ 1,767 & -1,757 \end{pmatrix}'$$

$$Corr(U^*, Y_1^*) = A^{*\prime} R_{11} = \begin{pmatrix} 0,9354 & 0,9275 \\ 0,3548 & -0,3737 \end{pmatrix} \begin{matrix} \leftarrow \rho(U_1^*, Y_1^*) \\ \leftarrow \rho(U_2^*, Y_1^*) \end{matrix}$$

Correlações das variáveis canônicas com as variáveis do primeiro filho

$$Corr(U^*, Y_2^*) = A^{*\prime} R_{12} = \begin{pmatrix} 0,7542 & 0,7585 \\ 0,0164 & -0,0141 \end{pmatrix}$$

$$Corr(V^*, Y_1^*) = B^{*\prime} R_{21} = \begin{pmatrix} 0,7377 & 0,7313 \\ 0,0191 & -0,0201 \end{pmatrix}$$

Note que as primeiras variáveis canônicas, U_1 e V_1, têm as maiores correlações com as variáveis originais.

$$Corr(V^*, Y_2^*) = B^{*\prime} R_{22} = \begin{pmatrix} 0,9565 & 0,9618 \\ 0,2924 & -0,2740 \end{pmatrix} \begin{matrix} \leftarrow \rho(V_1^*, Y_2^*) \\ \leftarrow \rho(V_2^*, Y_2^*) \end{matrix}$$

Correlação Canônica

Y_CC1	Y_PC1	Y_CC2	Y_PC2	Y*CC1	Y*_PC1	Y*_CC2	Y*_PC2	U*1	V*1	U1	V1
191	155	179	145	0,541	0,526	-0,482	-0,632	0,573	-0,583	21,892	20,550
195	149	201	152	0,951	-0,288	1,709	0,411	0,375	1,084	21,694	22,210
181	148	185	149	-0,484	-0,423	0,116	-0,036	-0,488	0,039	20,825	21,170
183	153	188	149	-0,279	0,255	0,414	-0,036	-0,021	0,190	21,294	21,320
176	144	171	142	-0,996	-0,966	-1,279	-1,079	-1,054	-1,226	20,256	19,910
208	157	192	152	2,282	0,798	0,813	0,411	1,676	0,632	23,003	21,760
189	150	190	149	0,336	-0,152	0,614	-0,036	0,106	0,291	21,423	21,420
197	159	189	152	1,156	1,069	0,514	0,411	1,196	0,481	22,518	21,610
188	152	197	159	0,234	0,119	1,311	1,455	0,191	1,444	21,508	22,570
192	150	187	151	0,643	-0,152	0,315	0,262	0,276	0,300	21,594	21,430
179	158	186	148	-0,688	0,933	0,215	-0,185	0,107	0,009	21,421	21,140
183	147	174	147	-0,279	-0,559	-0,980	-0,334	-0,446	-0,675	20,868	20,460
174	150	185	152	-1,201	-0,152	0,116	0,411	-0,742	0,280	20,568	21,410
190	159	195	157	0,438	1,069	1,112	1,156	0,800	1,184	22,119	22,310
188	151	187	158	0,234	-0,016	0,315	1,306	0,120	0,861	21,437	21,990
163	137	161	130	-2,327	-1,915	-2,275	-2,867	-2,284	-2,691	19,018	18,450
195	155	183	158	0,951	0,526	-0,084	1,306	0,799	0,660	22,120	21,790
186	153	173	148	0,029	0,255	-1,080	-0,185	0,149	-0,645	21,465	20,490
181	145	182	146	-0,484	-0,830	-0,183	-0,483	-0,700	-0,352	20,612	20,780
175	140	165	137	-1,098	-1,508	-1,876	-1,824	-1,393	-1,929	19,915	19,210
192	154	185	152	0,643	0,391	0,116	0,411	0,559	0,280	21,878	21,410
174	143	178	147	-1,201	-1,101	-0,582	-0,334	-1,238	-0,473	20,071	20,660
176	139	176	143	-0,996	-1,644	-0,781	-0,930	-1,408	-0,895	19,901	20,240
197	167	200	158	1,156	2,154	1,610	1,306	1,762	1,515	23,086	22,640
190	163	187	150	0,438	1,611	0,315	0,113	1,083	0,220	22,403	21,350

Calcular as correlações entre as variáveis duas a duas de interesse!
(mesmos resultados, menos fórmulas, mais intuição)

Correlação Canônica

Aplicação

Projeto ACTN3 e Força (dados de atletas)

Obtenha as variáveis canônicas relacionando as variáveis Pré e Pós.

Atleta	ACTN3	Massa	Pré.Força	Pré.ÁreaVL	Pós.Força	Pós.ÁreaVL
1	1	73.0	210	27.133	260	31.398
2	1	78.0	260	23.841	320	26.950
3	3	70.0	220	23.755	320	28.937
4	2	78.5	280	26.241	360	28.794
...						
94	2	95.0	175	25.120	260	28.605
95	2	71.0	220	25.452	330	29.029

Matrizes de Correlação (R1, R2 e R12): Dados Pré e Pós

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1.0000000	0.4013053	0.4123615	0.3575432	0.3488574
[2,]	0.4013053	1.0000000	0.2749434	0.8818818	0.2099585
[3,]	0.4123615	0.2749434	1.0000000	0.2575988	0.9477690
[4,]	0.3575432	0.8818818	0.2575988	1.0000000	0.2165755
[5,]	0.3488574	0.2099585	0.9477690	0.2165755	1.0000000

$$\begin{matrix} Y_{195\times3} \rightarrow U_1 \\ Y_{295\times2} \rightarrow V_1 \end{matrix} \Rightarrow \rho_c\left(U_1, V_1\right)$$

$$\begin{matrix} Y^*_{1\ 95\times3} \rightarrow U^*_1 \\ Y^*_{2\ 95\times2} \rightarrow V^*_1 \end{matrix} \Rightarrow \rho_c\left(U_1, V_1\right) = \rho_c\left(U^*_1, V^*_1\right)$$

Correlação Canônica

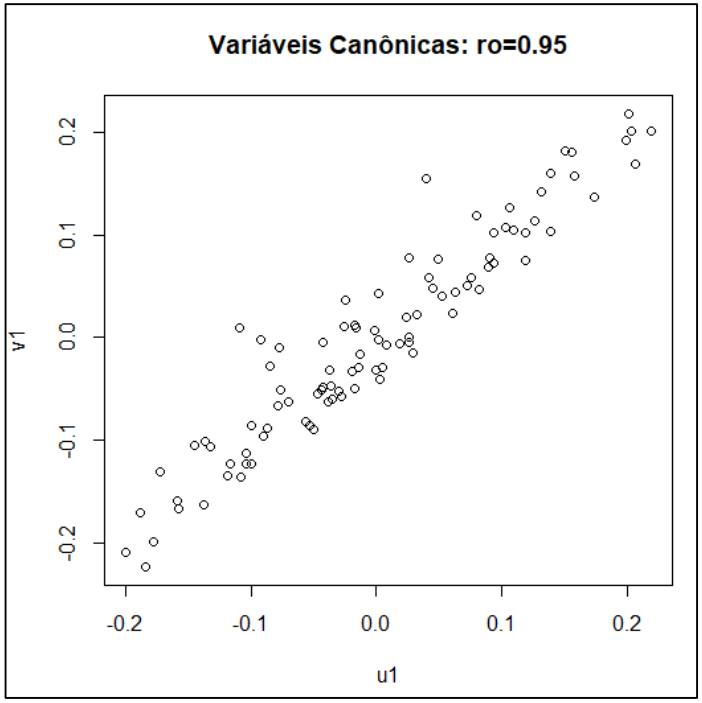
Aplicação

Projeto ACTN3 e Força (dados de atletas)

Obtenha as variáveis canônicas relacionando as variáveis Pré e Pós.

Variáveis originais

```
Cargas canônicas (var. "Pré")
                [,1]           [,2]           [,3]
[1,] -3.064514e-04  6.846222e-05 -0.0096161836
[2,] -2.146354e-05  1.684536e-03  0.0006009558
[3,]  2.593360e-02 -6.738612e-03  0.0088035299
Cargas canônicas (var. "Pós")
               [,1]          [,2]
[1,] 5.317268e-05  0.001550019
[2,] 2.295958e-02 -0.005926085
```



Correlação Canônica

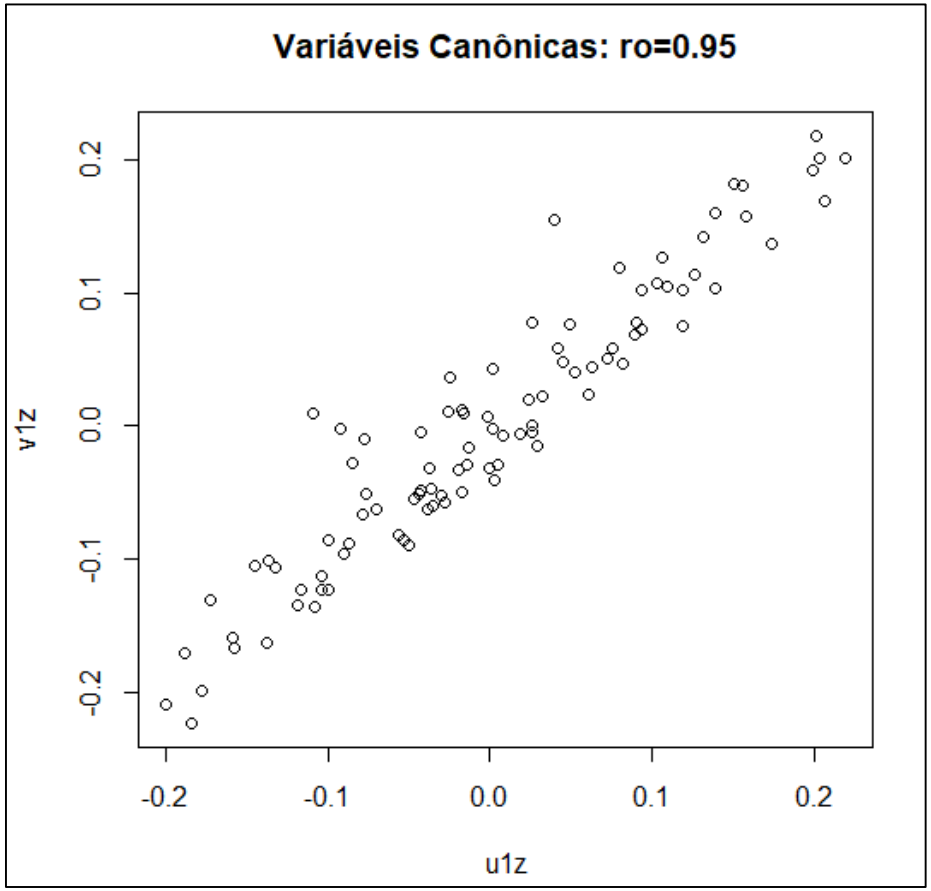
Aplicação

Projeto ACTN3 e Força (dados de atletas)

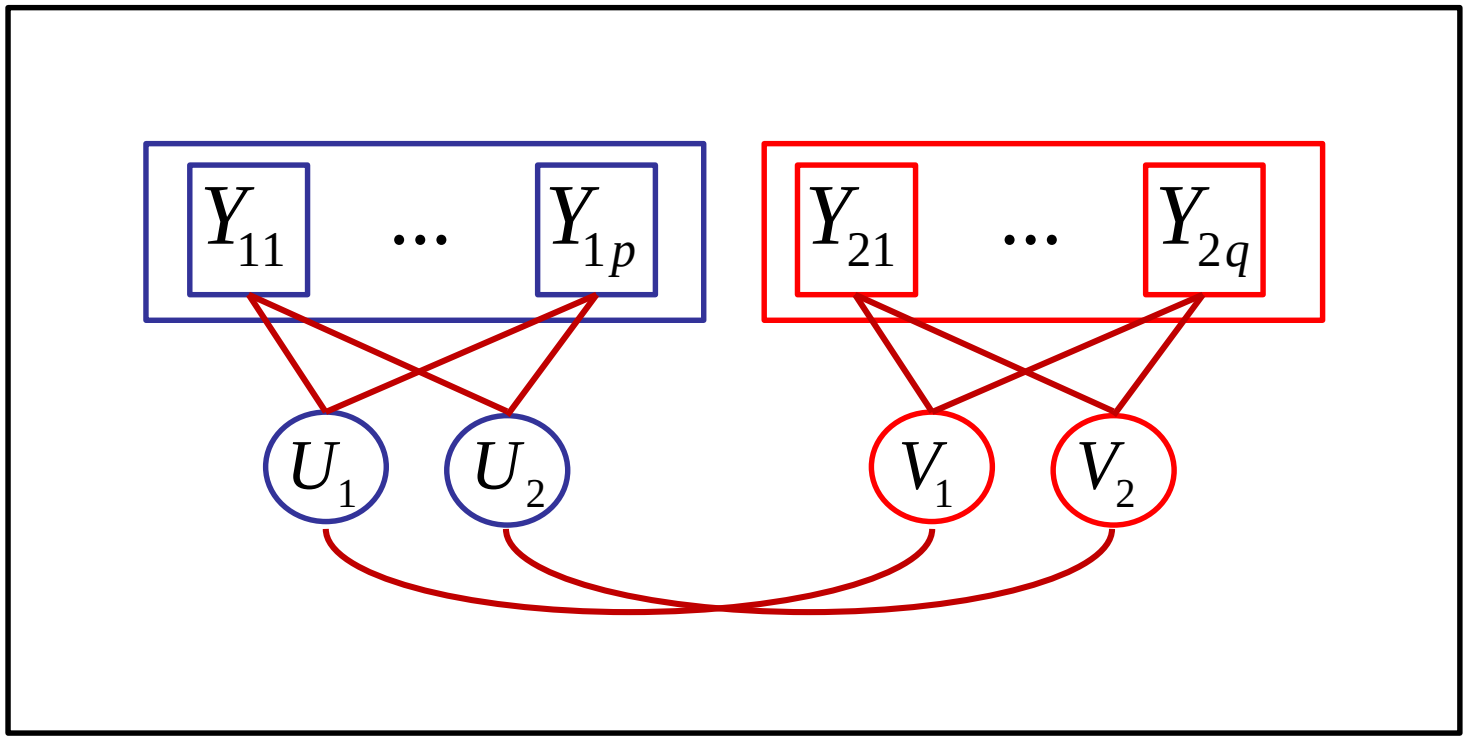
Obtenha as variáveis canônicas relacionando as variáveis Pré e Pós.

Variáveis padronizadas

```
Cargas canônicas (var. "Pré")
              [,1]           [,2]
[1,] -0.003818272  0.0008530142
[2,] -0.001363051  0.1069771302
[3,]  0.105009851 -0.0272858627
Cargas canônicas (var. "Pós")
            [,1]         [,2]
[1,] 0.00362213  0.10558752
[2,] 0.10229702 -0.02640383
```



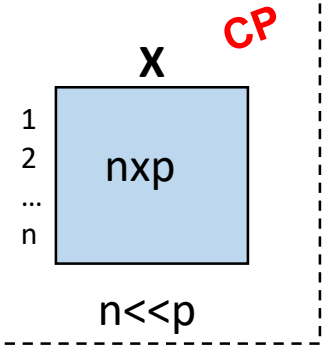
Correlação Canônica
Integração de Bancos de Dados



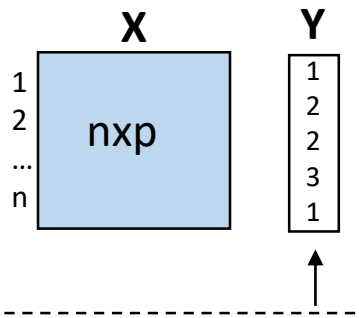
N-Integração de Bancos de Dados

Um Único BD

Análise Não-Supervisionada CP



Análise Supervisionada AD



N-Integração entre múltiplos níveis de informação avaliados nas mesmas unidades amostrais!

⇒ **Aplicação em Dados MultiOmics**

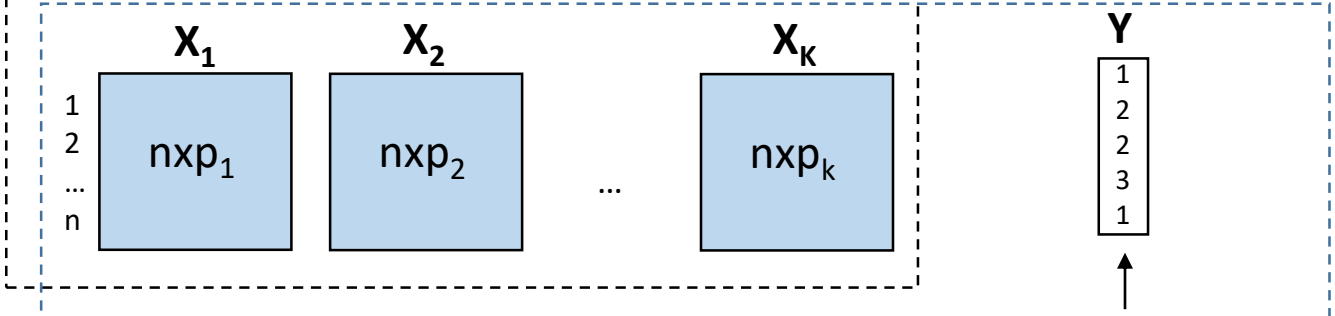
X: Matriz(es) de Dados

Y: Resposta de interesse (em geral, Classes)

Múltiplos BD (Multimodais, Multivisão)

Análise Não-Supervisionada CCg

An. Supervisionada ADg



P-Integração: mesmas variáveis em diferentes unidades amostrais (estudos multicentros, Meta-análise)

Redução de Dimensionalidade
Obtenção de Vetores Reducionistas

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

Revisando

- Componentes Principais: $Y_i \sim \left(\mu_{p\times 1}; \Sigma_{p\times p}\right) \rightarrow \quad Z_{ki} = V_k' Y_i, \; k = 1, 2, ..., m$

$$\Sigma = V_k \Lambda V_k' \quad \Rightarrow \quad f\left(\Sigma; V_k\right) = \frac{V_k' \Sigma V_k}{V_k' V_k} = \lambda_k, \quad V_k' V_k = 1$$

- Análise Discriminante (Linear de Fisher): $f\left(\Sigma_w^{-1}\Sigma_b; a\right) = \dfrac{a'\Sigma_b a}{a'\Sigma_w a}, \quad a'\Sigma_w a = 1$ $\qquad \Sigma = \Sigma_b + \Sigma_w$

$$Y_{n\times p}; n = \sum n_g; \quad Y_i \rightarrow a'Y_i$$

- Análise de Correlação Canônica:

$$Y_{i(p+q)\times 1} = \begin{pmatrix} Y_{1i\,p\times 1} \\ Y_{2i\;q\times 1} \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

$$f_1\left(\Sigma_{11}^{-1}\boldsymbol{\Sigma}_{12}\Sigma_{22}^{-1}\,\Sigma_{21}; a\right) = \frac{a'\boldsymbol{\Sigma}_{12}\Sigma_{22}^{-1}\,\Sigma_{21}a}{a'\Sigma_{11}a}, \quad a'\Sigma_{11}a = 1$$

$$f_2\left(\Sigma_{22}^{-1}\boldsymbol{\Sigma}_{21}\Sigma_{11}^{-1}\,\Sigma_{12}; b\right) = \frac{b'\boldsymbol{\Sigma}_{21}\Sigma_{11}^{-1}\,\Sigma_{12}b}{b'\Sigma_{22}b}, \quad b'\Sigma_{22}b = 1$$

Redução de Dimensionalidade - Apoio do R

- eigen(S) : Decomposição espectral. Recebe uma matriz da forma quadrática a ser analisada ($\Re^{pxp}$ ou $\Re^{nxn}$)

- princomp(Y): CP. Recebe Y_{nxp} e realiza a decomposição espectral de R ou S (com divisor n)

- prcomp(Y) : CP. Recebe Y_{nxp} e realiza a decomposição espectral de R ou S (com divisor n-1)
  $\rightarrow$ suporta n<p

- svd(Y): rDecomposição em valores singulares. Recebe Y_{nxp} (n<p, n>p) e realiza a decomposição em valores singulares de $\Re^{pxp}$ e $\Re^{nxn}$.
  Para comparar com *eigen* é preciso "padronizar" autovalores: $\lambda_{eigen} = \left(\lambda_{svd} / \sqrt{n-1} \right)^2$

- cmdscale: recebe a matriz de distâncias D ou Similaridade entre observações e realiza a Análise de Escalonamento Multidimensional (Análise de Coordenadas Principais)
  Ver também os pacotes do R: "sammon" e "isoMDS"

- ca: realiza a Análise de Correspondência.

- lda: recebe as (p+1)-variáveis e realiza a Análise Discriminante (solução geral)

- cc(Y_1,Y_2) da biblioteca CCA: realiza a Análise de Correlação Canônica

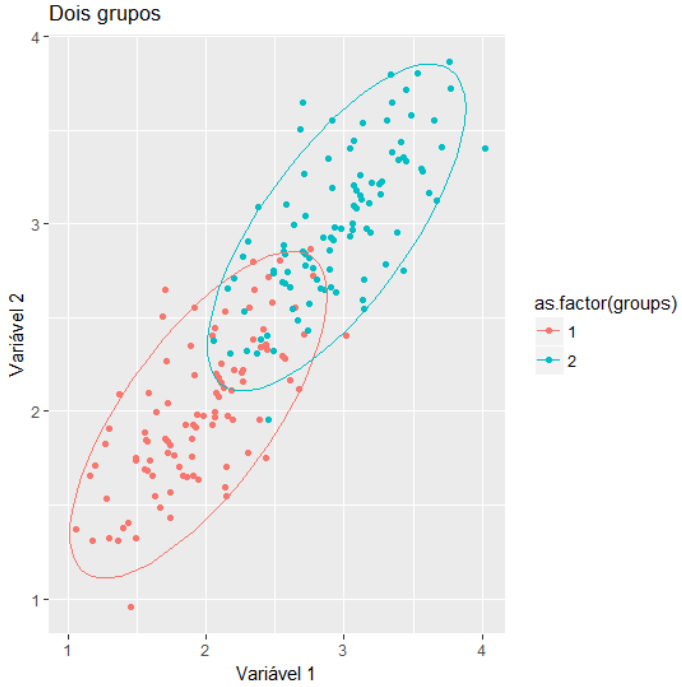
Onde estão os Vetores Reducionistas?

Um gráfico pode valer mais que mil palavras mas pode exigir milhares de palavras para construí-lo. Tukey

Obter a direção do CP e do Eixo Discriminante.

Observações independentes. Indicação da elipse de concentração (95%).

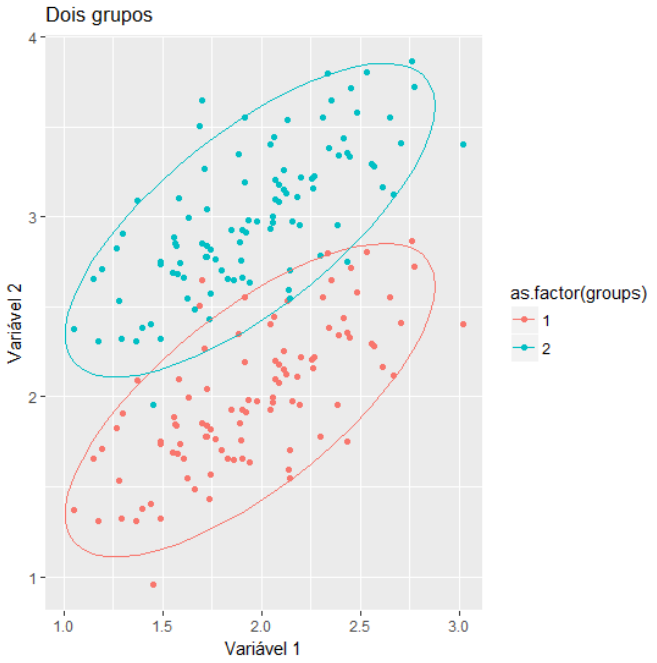
Exemplo 1



$$n = 200,\ p = 2,\ \mu_1 = (2,2),\ \mu_2 = (3,3)$$

$$R_{2\times 2} = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}; \sigma = (0.4 \quad 0.4)$$

Exemplo 2



$$n = 200,\ p = 2,\ \mu_1 = (2,2),\ \mu_2 = (2,3)$$

$$R_{2\times 2} = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}; \sigma = (0.4 \quad 0.4)$$

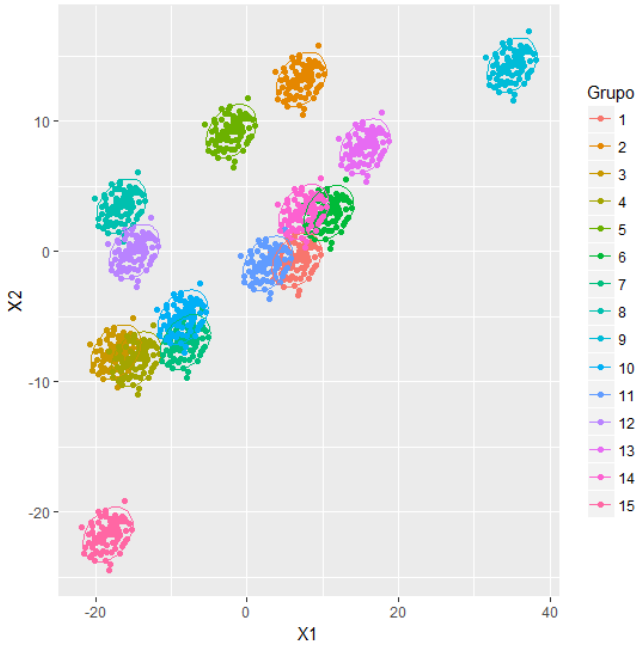
Onde estão os Vetores Reducionistas?

Obter a direção do Eixo Discriminante.

Observações independentes ENTRE e DENTRO de grupos.

Exemplo 3: “Sinais Iguais”

$$T = B + W$$

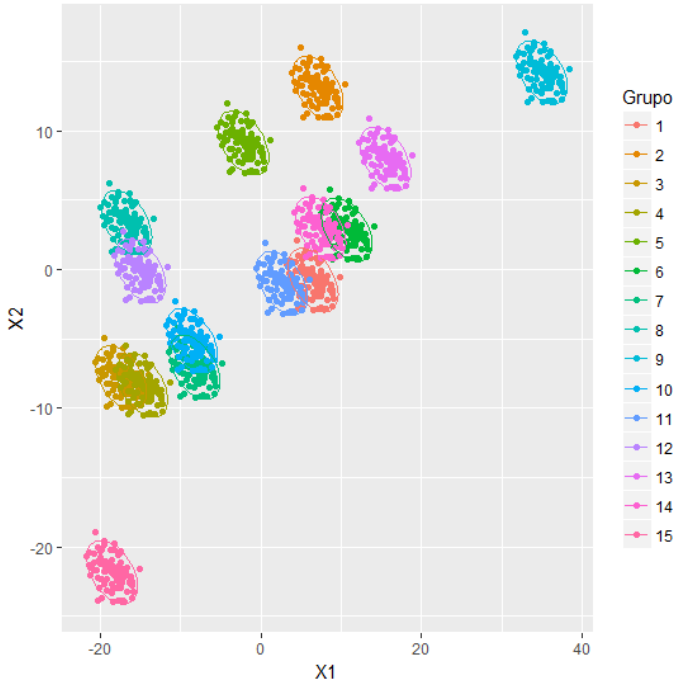


$$G = 15,\ n_g = 100,\ \mu_{\cdot} = (0,0)$$

$$S_b = \begin{pmatrix} 150 & 100 \\ 100 & 150 \end{pmatrix},\ S_w = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

Exemplo 4: “Sinais Opostos”

$$T = B + W$$



$$G = 15,\ n_g = 100,\ \mu_{\cdot} = (0,0)$$

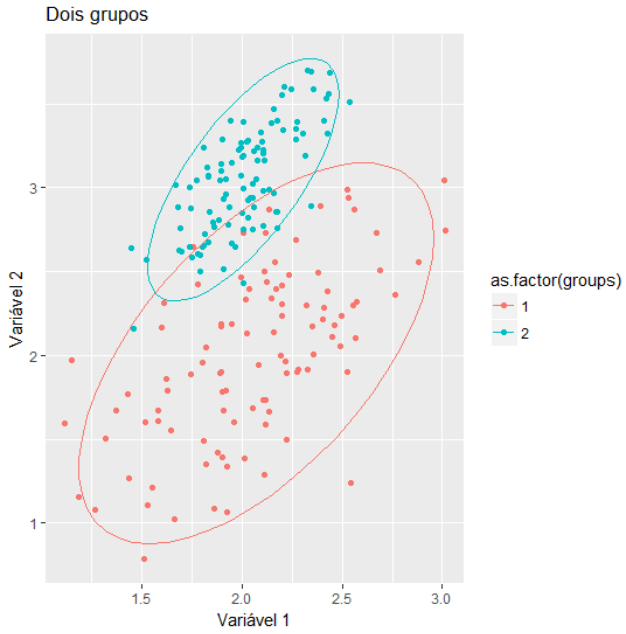
$$S_b = \begin{pmatrix} 150 & 100 \\ 100 & 150 \end{pmatrix},\ S_w = \begin{pmatrix} 2 & -0.5 \\ -0.5 & 1 \end{pmatrix}$$

Onde estão os Vetores Reducionistas?

Obter a direção da Variável Canônica.

Observações independentes avaliadas em $\Re^{p+q}$.

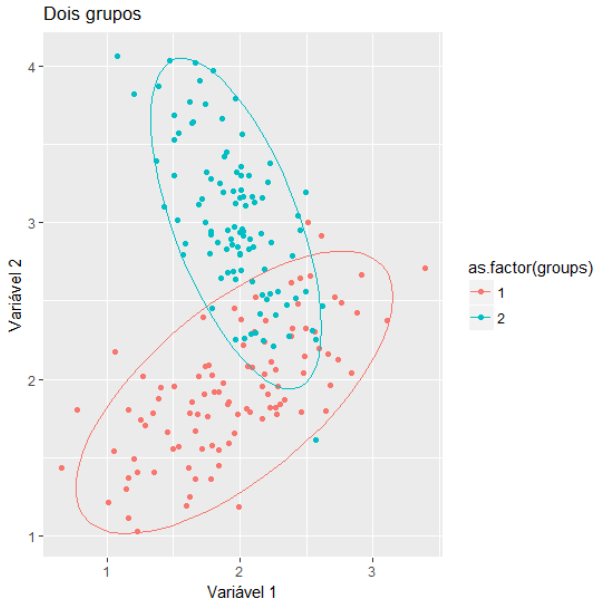
Exemplo 5: Correlações de mesmo sinal



$$R = \begin{pmatrix} 1 & 0.65 & 0.2 & 0.5 \\ 0.65 & 1 & 0.1 & 0.4 \\ 0.2 & 0.1 & 1 & 0.7 \\ 0.5 & 0.4 & 0.7 & 1 \end{pmatrix}$$

$$\sigma_1 = (0.4 \quad 0.5); \sigma_2 = (0.2 \quad 0.3); \quad \mu_1 = (2,2); \mu_2 = (3,3)$$

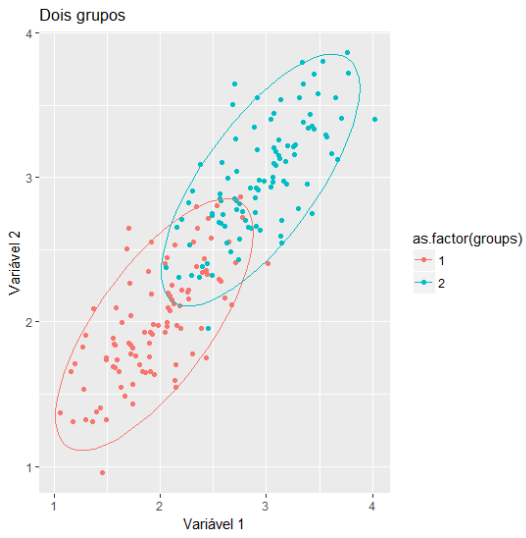
Exemplo 6: Correlações de sinal diferentes



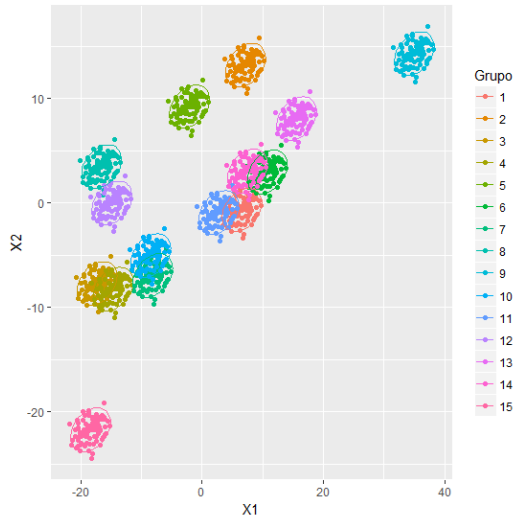
$$R = \begin{pmatrix} 1 & 0.7 & 0.2 & 0.5 \\ 0.7 & 1 & 0.3 & 0.4 \\ 0.2 & 0.3 & 1 & -0.7 \\ 0.5 & 0.4 & -0.7 & 1 \end{pmatrix}$$

$$\sigma_1 = (0.5 \quad 0.4); \sigma_2 = (0.3 \quad 0.5); \quad \mu_1 = (2,2); \mu_2 = (3,3)$$

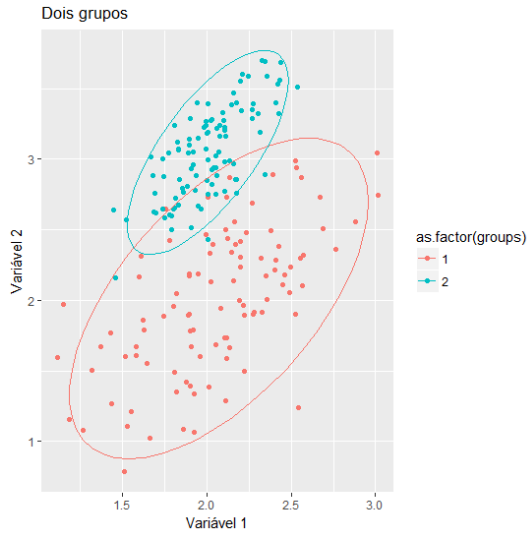
Exemplo 1



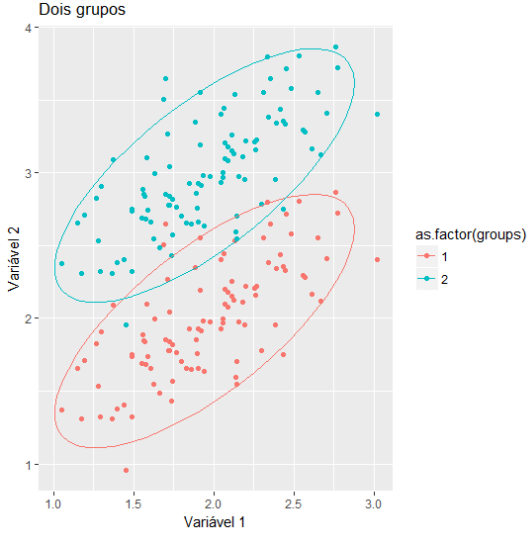
Exemplo 3



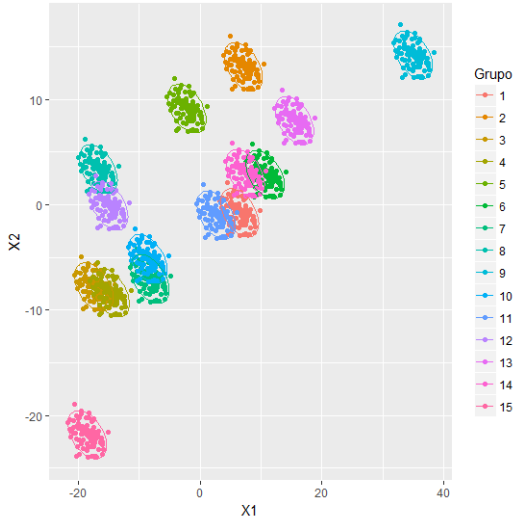
Exemplo 5



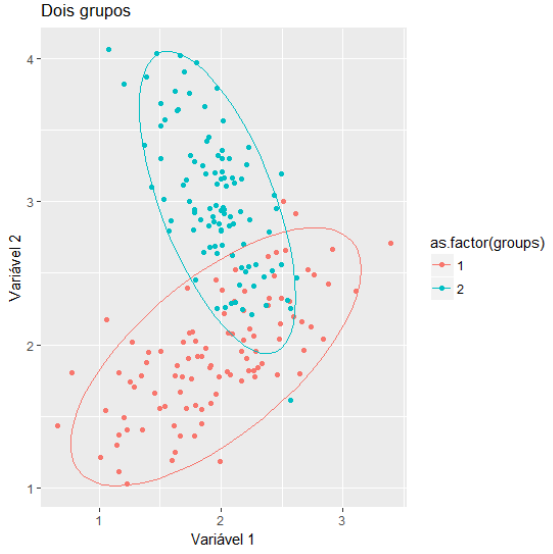
Exemplo 2



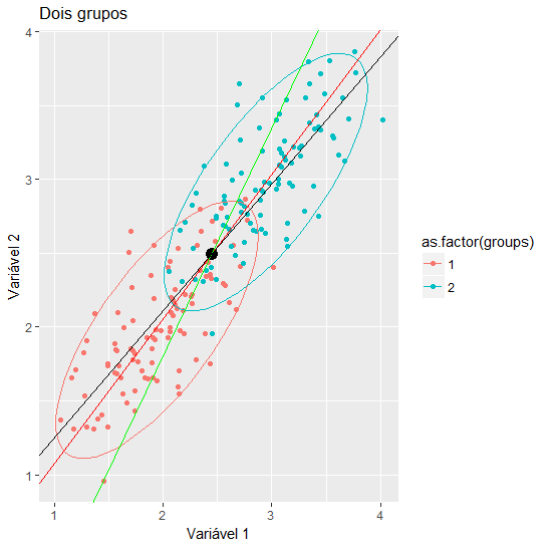
Exemplo 4



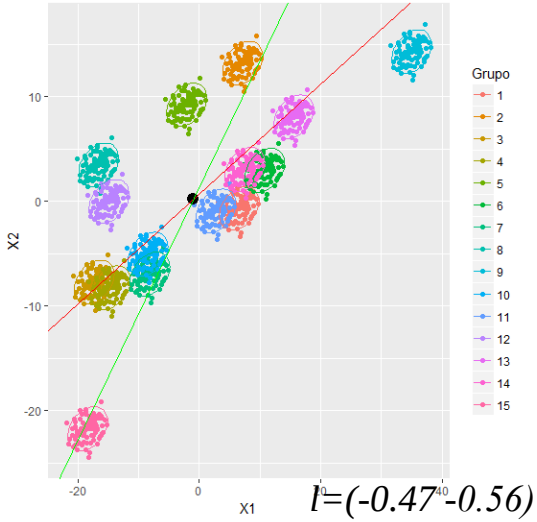
Exemplo 6



Exemplo 1

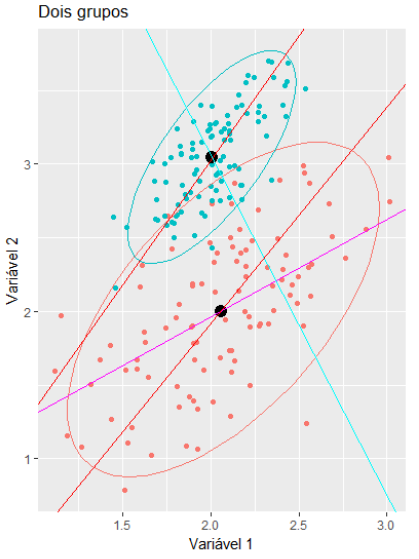


Exemplo 3



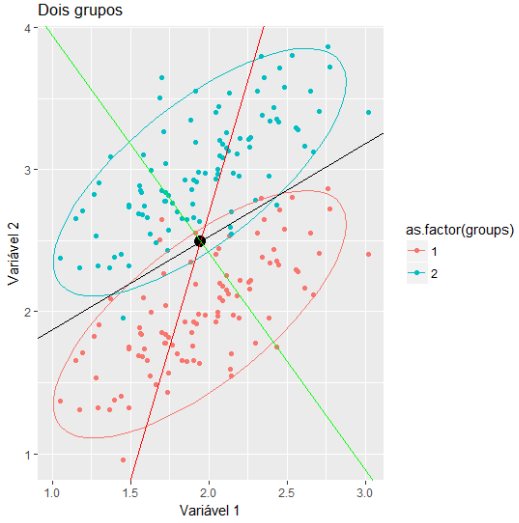
l=(-0.47 -0.56)'
%Expl=0.78

Exemplo 5

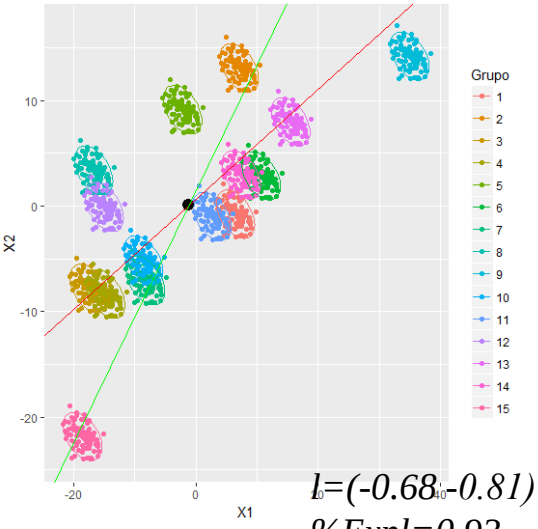


ρ_c=0.56

Exemplo 2

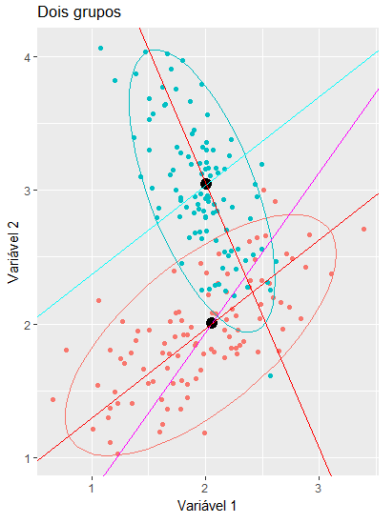


Exemplo 4



l=(-0.68 -0.81)'
%Expl=0.93

Exemplo 6



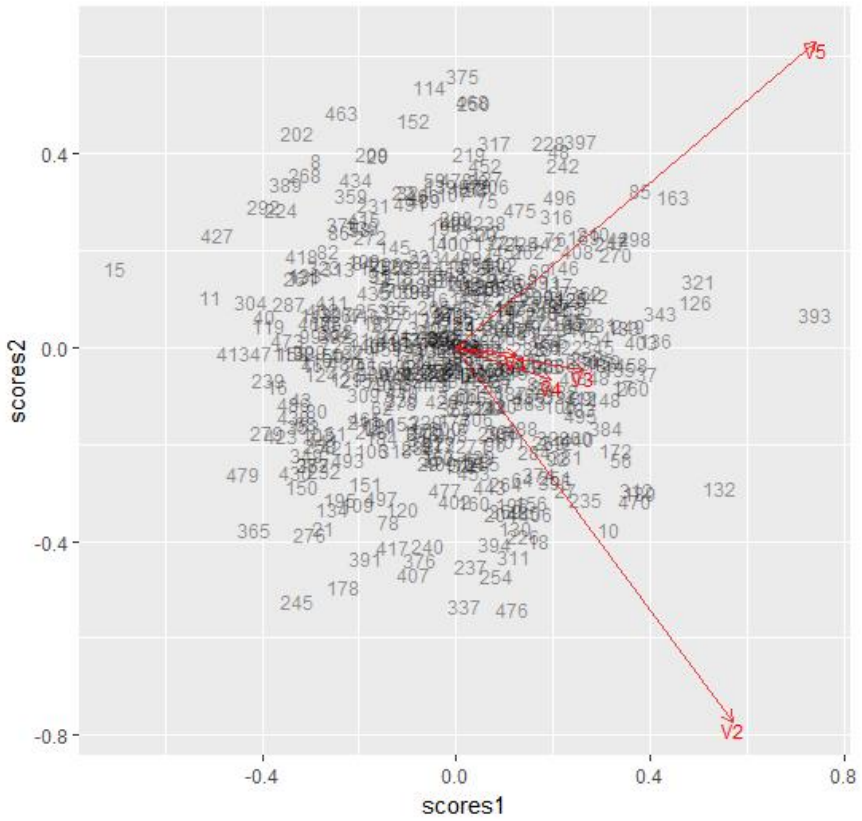
ρ_c=0.98

Preto:reta de MQ Vermelho:vetor de CP Verde:vetor discriminante Azul e rosa:variáveis canônicas

Dados Multivariados: $Y_{n\times p}$
Problemas em Alta Dimensionalidade

Big-n (n >> p)

Biplot: n=500 p=5

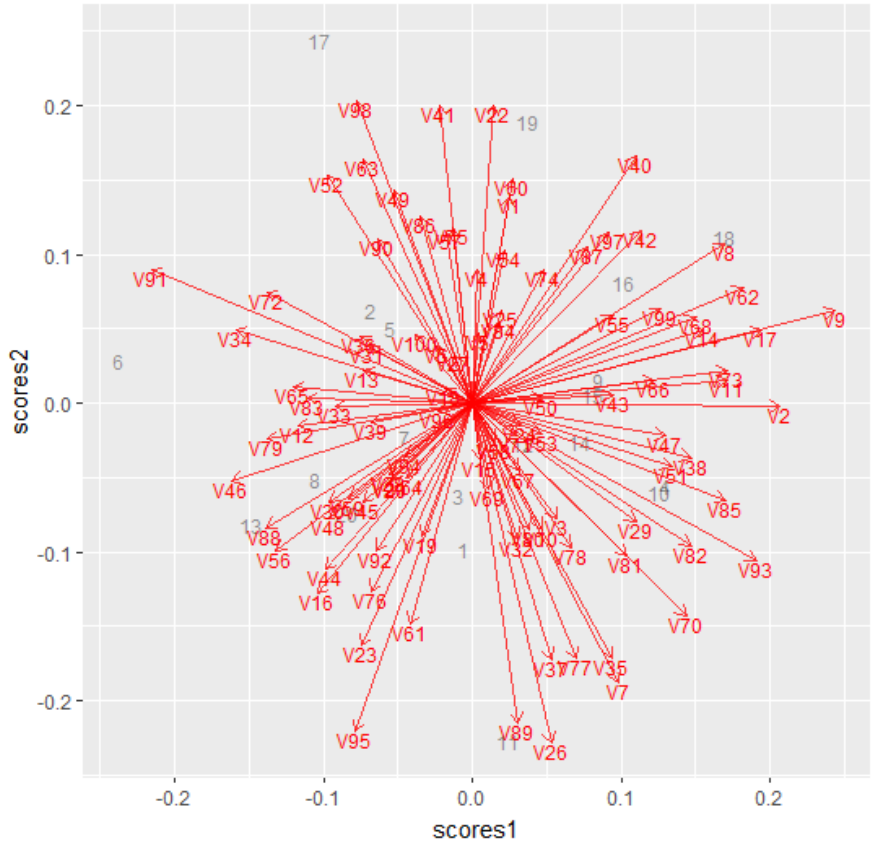


Sumarização e visualização ?

(Black Screen Problem - BSP: *R_alpha blending*)

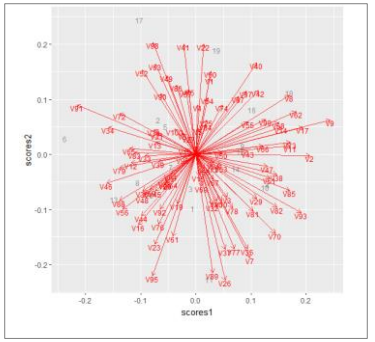
Big-p (n << p)

Biplot: n=20 p=100



Soluções regularizadas e penalizadas!

Dados Multivariados – *Big-p*



- *Big-p*: Redução de dimensionalidade em espaços com n<<p

Componente Principal Regularizado e Penalizado (Elastic Net:Zou, Hastie, Tibishirani, 2006)

Passo 1: Obter as **Coordenadas Principais** a partir de D_{nxn}

$$Y_{n\times p} = U\Lambda^{1/2}V' \overset{n<<p}{\Rightarrow} Z_j = U_j\lambda_j^{1/2}$$

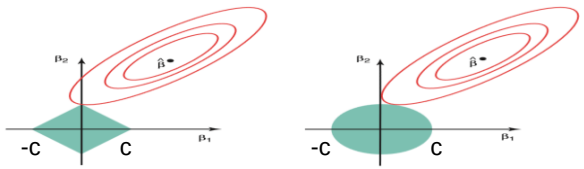
Desvantagem: não atribui pesos nulos às variáveis!

Passo 2: Obter os Componentes Principais Regularizados e Penalizados a partir das Coordenadas Principais

$$\hat{\beta} = \arg\min_{\beta} \left\{ \left\| Z_j - Y\beta \right\|_2^2 + \lambda_1 \left\| \beta \right\|_2^2 + \lambda_2 \left\| \beta \right\|_1 \right\}; \quad \hat{v}_j = \frac{\hat{\beta}}{\left\| \hat{\beta} \right\|_2}; \quad \hat{Z}_j = Y\hat{v}_j$$

obter o vetor β (de pesos) | regularização | penalização

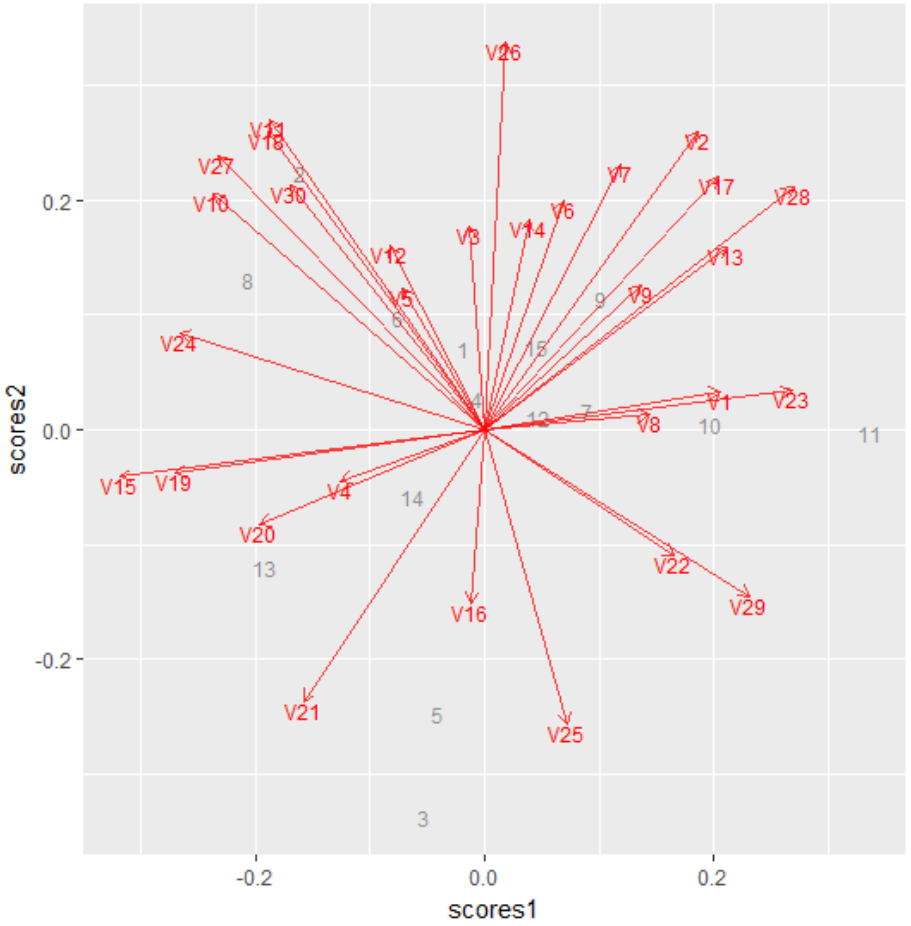
Soluções sujeitas a restrições



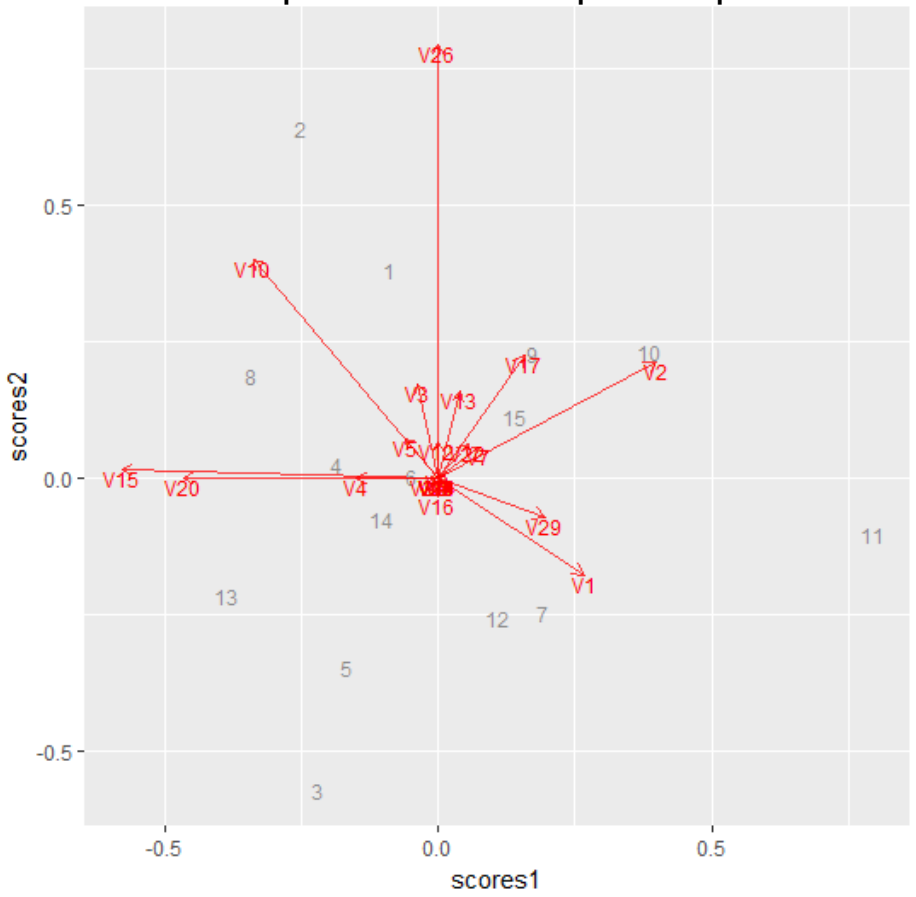
Componentes Principais – n<<p

Representação Biplot: n=15 p=30

R-prcomp: Coordenadas Principais



R-SPCA do pacote ElasticNet:
Componentes Principais Esparsos



Redução de Dimensionalidade em Espaços com n<<p - Apoio do R

- Componentes Principais: *SPCA* (Sparse principal Component Analysis)
- Análise Discriminante: *sparseLDA* (Sparse Linear Discriminant Analysis)
- Análise de Correlação Canônica: *PMA* (Penalized Matrix Analysis) – CCA
  rcc (Correlação Canônica Regularizada)

- Veja também outras análises de redução de dimensionalidade:

  PLS (Partial Least Square) (Tibshirani et al., 2015; Abdi, 2010)