

MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler
pavan@ime.usp.br

1º Sem/2022 - IME

Análise Multivariada

$$Y_{n \times p} = (Y_{ij}) \in \mathbb{R}^{n \times p}$$

Já vimos 😊

- ✓ Estatísticas descritivas multivariadas, Episódios 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 $Y_{n \times p}$



Técnicas Multivariadas:

Já vimos 😊

- ✓ 1. Análise de Componentes Principais (CP)
- ✓ 2. Escalonamento Multidimensional (CoP)
- ✓ 3. Análise de Correspondência
- ✓ 4. Análise Fatorial
- ✓ 5. Análise Discriminante (MANOVA)
- ✓ 6. Análise de Agrupamento

- Análise de Correlação Canônica 

Análise de Correlação Canônica

Análise de Correlação Canônica

Análise Não-Supervisionada

	Variáveis					
Unidades Amostras	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”

⇒ Obter Variáveis Canônicas (escores, var. latentes, vetores reducionistas) de cada subconjunto das variáveis originais, com máxima correlação entre elas.

⇒ 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)

Família	1° Filho		2° Filho	
	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?

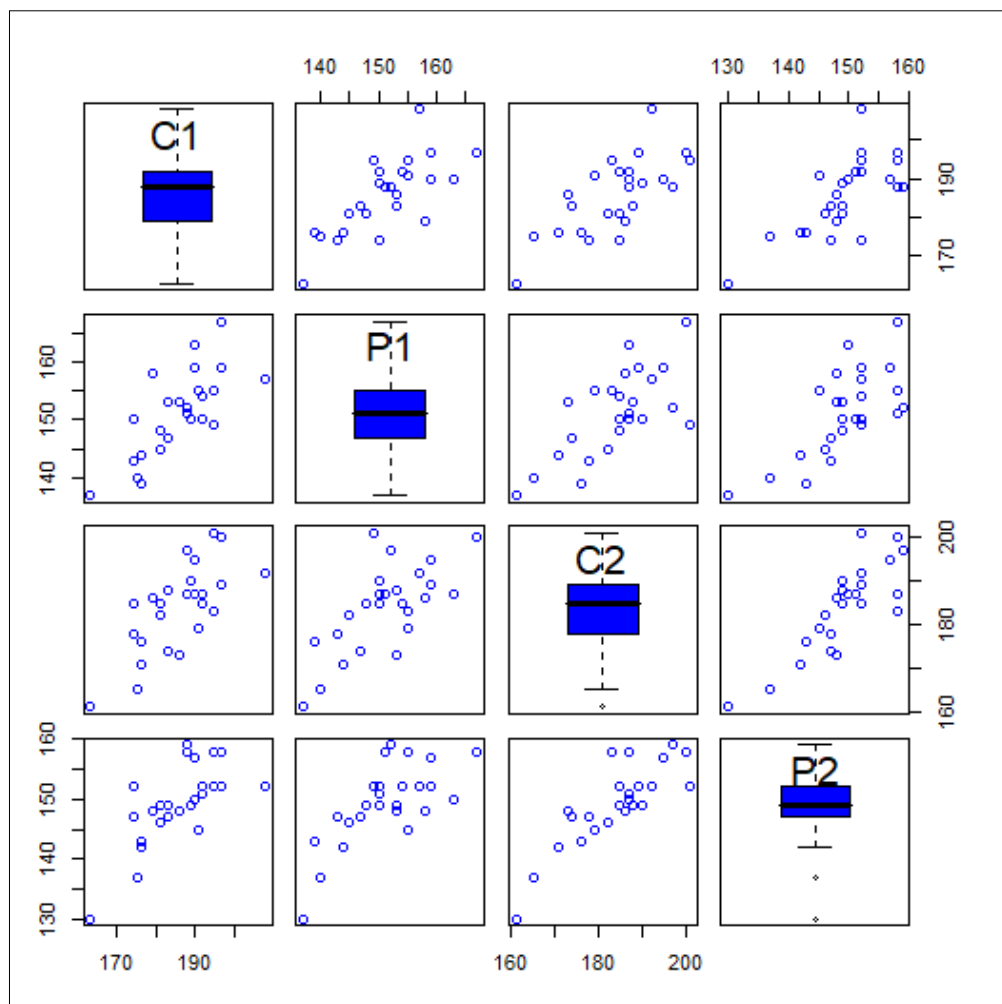
Discuta a estrutura dos dados.

Neste caso, tem-se as mesmas variáveis (comprimento e perímetro) são 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 de CC se estende para situações de dois conjuntos de variáveis diferentes!

Diferentes Medidas de Correlação

Coeficiente de Correlação Linear de Pearson
para Pares de variáveis - Dados de Morfometria Cefálica



Correlações (marginais):

	C1	P1	C2	P2
C1	1.00	0.73	0.71	0.70
P1		1.00	0.69	0.71
C2			1.00	0.84

← Correlação entre as variáveis DENTRO de cada grupo (1º e 2º filho)

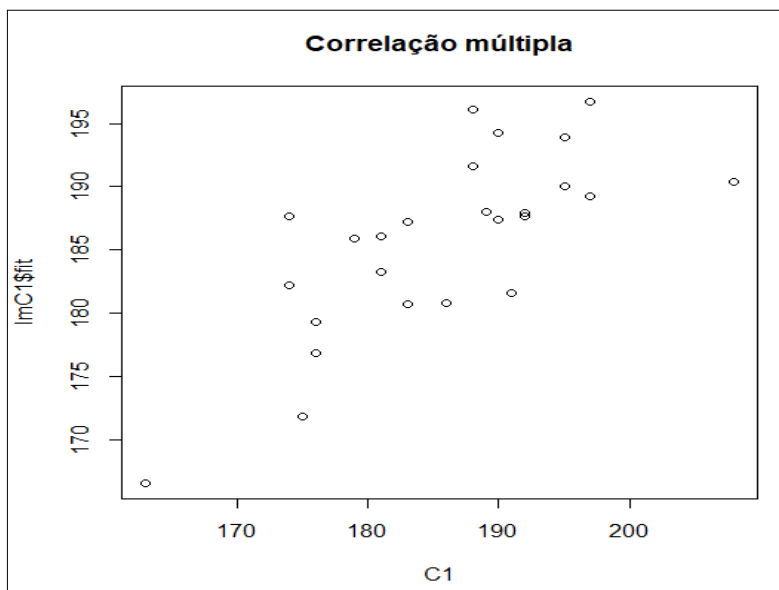
← Correlação ENTRE os grupos, para cada par de variável.

Diferentes Medidas de Correlação

Coeficiente de Correlação Múltipla

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

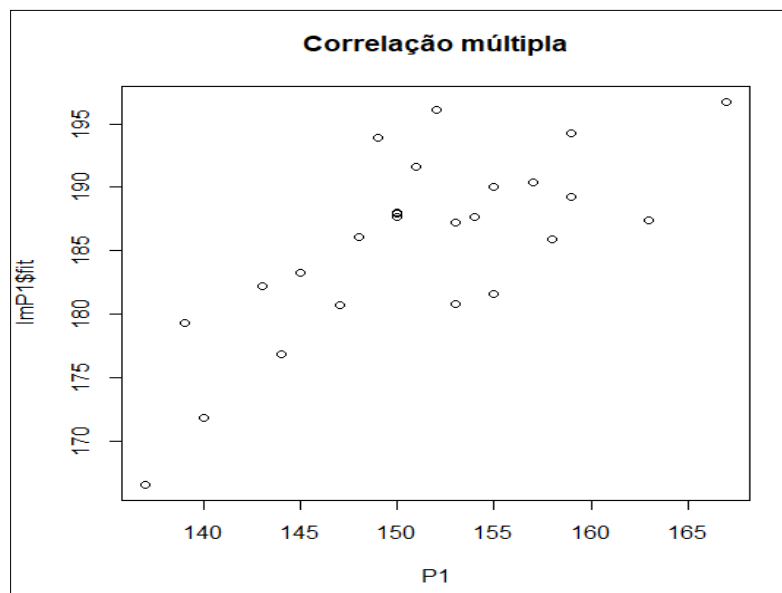
$$\rho_M [Y_{C1}, (Y_{C2}, Y_{P2})]$$



$$\rho_P (Y_{C1}, \hat{Y}_{C1|C2,P2}) = 0,738$$

$$Y_{C1} = \beta_0 + \beta_1 Y_{C2} + \beta_2 Y_{P2} + e$$

$$\rho_M [Y_{P1}, (Y_{C2}, Y_{P2})]$$



$$\rho_P (Y_{P1}, \hat{Y}_{P1|C2,P2}) = 0,731$$

$$Y_{P1} = \beta_0 + \beta_1 Y_{C2} + \beta_2 Y_{P2} + e$$

Diferentes Medidas de Correlação

Coeficiente de Correlação Parcial – Útil para Inferência Causal

⇒ Considere a distribuição condicional das variáveis do Filho 2 dado as do Filho 1!

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

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

$$E(Y_2 | Y_1) = \mu_2 - \Sigma_{21} \Sigma_{11}^{-1} (Y_1 - \mu_1) \quad Cov(Y_2 | Y_1) = \Sigma_{22.1} = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$$

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

$$\rho(Y_{2j}, Y_{2k} | Y_1) = \frac{\sigma_{jk.1}}{\sqrt{\sigma_{jj.1}} \sqrt{\sigma_{kk.1}}};$$

$\sigma_{jk.1}$ é a casela jk da matriz $\Sigma_{22.1}$

Σ^{-1}	C1	P1	C2	P2
C1	1.000	0.425	0.223	0.152
P1		1.000	0.132	0.225
C2			1.000	0.626

A correlação parcial (de pares de variáveis dado as demais) pode ser obtida da matriz $\Sigma_{22.1}$ ou diretamente da matriz de precisão Σ^{-1} .

Outra medida de
correlação:

Correlação Canônica - Exemplos

Unidades Amostrais	Variáveis					
	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)}

- 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 de pacientes.
- Relacionar variáveis da folha com variáveis do tronco de plantas.
- ...



Integração de
bancos de dados!

Análise de Correlação Canônica - Motivação

Dados “Iris” do R Medidas do comprimento e largura da **pétala e sépala** de 50 flores de iris de de três espécies (setosa, versicolor e virginica).

$$Y_{150 \times 4} \Leftrightarrow \begin{pmatrix} Y_{G=1 \ 50 \times 4} \\ Y_{G=2 \ 50 \times 4} \\ Y_{G=3 \ 50 \times 4} \end{pmatrix} \Leftrightarrow Y_{150 \times (2+2)} = \begin{pmatrix} Y_{150 \times 2} & Y_{150 \times 2} \end{pmatrix}$$

↑ ↑
CP **AD** **Correlação Canônica**



	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

Reduzir a dimensionalidade dos dados por obter Escores das variáveis da Sépala mais correlacionados com as variáveis da Pétala.

Correlação Canônica – Exemplos

TABLE 1.
Wine tasting data from Abdi and Valentin (2007).

Wine	Oak-type	Expert 1			Expert 2				Expert 3		
		Fruity	Woody	Coffee	Red fruit	Roasted	Vanillin	Woody	Fruity	Butter	Woody
1	1	1	6	7	2	5	7	6	3	6	7
2	2	5	3	2	4	4	4	2	4	4	3
3	2	6	1	1	5	2	1	1	7	1	1
4	2	7	1	2	7	2	1	2	2	2	2
5	1	2	5	4	3	5	6	5	2	6	6
6	1	3	4	4	3	5	4	5	1	7	5

$$Y_{6 \times (3+4+3)} = (Y1_{6 \times 3} \quad Y2_{6 \times 4} \quad Y3_{6 \times 3}) \quad \text{Correlação Canônica Múltipla}$$

→ Correlação Canônica entre Dois Grupos de Variáveis
(Pares de Bancos de Dados)

$$(Y1_{6 \times 3} \quad Y2_{6 \times 4})$$

$$(Y1_{6 \times 3} \quad Y3_{6 \times 3})$$

$$(Y2_{6 \times 4} \quad Y3_{6 \times 3})$$

Correlação Canônica – Exemplos

TABLE 1. Male and female views of working wives in eight countries (International Social Survey Programme, 1989)

		Should wife stay at home...? (response percentage)			
	Country	... before first child (1)	... after first child (2)	... when first child is at school (3)	... when all children are at school (4)
Male	D	6.3	78.3	51.4	14.6
	GB	3.0	74.7	15.3	4.0
	US	7.6	61.1	16.2	7.1
	A	5.1	75.4	45.7	12.2
	H	18.9	58.4	22.1	8.7
	NL	3.0	60.0	17.3	3.6
	I	11.1	49.6	23.6	21.7
	IR	7.0	56.4	33.5	9.2
Female	D	6.1	73.9	47.7	14.5
	GB	2.4	66.6	10.0	1.9
	US	4.0	50.0	10.3	3.8
	A	2.9	69.4	40.5	7.3
	H	7.2	46.5	14.9	3.4
	NL	1.5	52.2	10.0	2.3
	I	3.8	38.3	12.0	10.0
	IR	5.8	54.6	20.7	5.9

Each value is the percentage of respondents who are in favour of the wife staying at home in the following four periods: (1) before the first child is born; (2) after the birth of the first child; (3) after the first child has gone to school; and (4) after all children are at school. The countries surveyed are: D—Germany, GB—Great Britain, US—United States of America, A—Austria, H—Hungary, NL—Netherlands, I—Italy, IR—Republic of Ireland.

$$Y_{8 \times (4+4)} = (Y1_{8 \times 4} \quad Y2_{8 \times 4})$$

O pareamento
das observações
é dado pelo país!

Greenacre, M (2003).
SVD of matched
matrices.

Correlação Canônica

Notação

A matriz de dados multivariados está particionada em Dois Conjuntos de Variáveis:

$$Y_{n \times (p+q)} = \begin{pmatrix} Y_{1n \times p} & Y_{2n \times q} \end{pmatrix}; \quad Y_{i(p+q) \times 1} \stackrel{iid}{\sim} \left(\mu_{(p+q) \times 1}; \Sigma_{(p+q) \times (p+q)} \right)$$

$$Y_{i(p+q) \times 1} = \begin{bmatrix} Y_{1i p \times 1} \\ Y_{2i q \times 1} \end{bmatrix} \left\{ \begin{array}{ll} E(Y_{1i p \times 1}) = \mu_1 & Cov(Y_{1i p \times 1}) = \Sigma_{11 p \times p} \\ E(Y_{2i q \times 1}) = \mu_2 & Cov(Y_{2i q \times 1}) = \Sigma_{22 q \times q} \\ Cov(Y_{1i p \times 1}, Y_{2i q \times 1}) = \Sigma_{12 p \times q} = \Sigma'_{21 q \times p} \end{array} \right.$$

Mede a covariância entre os dois conjuntos de variáveis

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

Partição da matriz de covariância!

Análise de Correlação Canônica

Notação

	Variáveis					
Unidades Amostras	Y1	Y2	...	Y _p		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)}

$\mu_{1 \times p}$

$\Sigma_{11 \times p \times p}$

$\mu_{2 \times q}$

$\Sigma_{22 \times q \times q}$

$\Sigma_{12 \times p \times q}$

mede a covariância entre os dois conjuntos de variáveis



$$\Sigma_{(p+q) \times (p+q)} = \begin{pmatrix} \Sigma_{11 \times p \times p} & \Sigma_{12 \times p \times q} \\ \Sigma_{21 \times q \times p} & \Sigma_{22 \times q \times q} \end{pmatrix}$$

Matriz de covariância para as (p+q) variáveis

Correlação Canônica

Como Resumir “Correlações” entre Dois Conjuntos de Variáveis?

	Y_1			Y_2		
Indiv	Y_{11}	...	Y_{1p}	Y_{21}	...	Y_{2q}
1	Y_{111}	...	Y_{1p1}	Y_{211}	...	Y_{2p1}
2	Y_{112}	...	Y_{1p2}	Y_{212}	...	Y_{2p2}
...
n	Y_{11n}	...	Y_{1pn}	Y_{21n}	...	Y_{2pn}

Obter combinações lineares de cada conjunto!

$$U_i = a' Y_{1i}$$

$$a_1 Y_{11i} + a_2 Y_{12i} + \dots + a_p Y_{1pi}$$

$$V_i = b' Y_{2i}$$

$$b_1 Y_{21i} + b_2 Y_{22i} + \dots + b_q Y_{2qi}$$

$$Var(U_i) = a' \Sigma_{11} a$$

$$Var(V_i) = b' \Sigma_{22} b$$

$$Cov(U, V) = a' \Sigma_{12} b$$

Tal que: U e V tenham correlação máxima!

Correlação Canônica

Obter U de Y_1 e V de Y_2 , tal que, a correlação entre U e V seja máxima

$$\begin{array}{l} U_i = a' Y_{1i} \\ V_i = b' Y_{2i} \end{array} \quad \left\{ \begin{array}{l} Var(U_i) = a' \Sigma_{11} a \\ Var(V_i) = b' \Sigma_{22} b \end{array} \right. \quad Cov(U_i, V_i) = a' \Sigma_{12} b$$

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

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

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

⇒ Caso seja de interesse, encontrar um segundo par de variáveis padronizadas, U_2 e V_2 , com correlação canônica máxima e que não sejam correlacionadas com o primeiro par ⇒ e assim por diante até $m = \min(n, p, q)$

Correlação Canônica

$$\max_{a,b} \text{Corr}(U,V) = \max_{a,b} \frac{a' \Sigma_{12} b}{\sqrt{a' \Sigma_{11} a} \sqrt{b' \Sigma_{22} b}}$$

equivale a maximizar:



$$\Rightarrow \max_{a \in \mathbb{R}^p} \frac{a' \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} a}{a' \Sigma_{11} a}$$

$$\Rightarrow \max_{b \in \mathbb{R}^q} \frac{b' \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} b}{b' \Sigma_{22} b}$$

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

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

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

Os **escores** U e V são projeções dos dados que compartilham os mesmos autovalores. Os vetores “a” e “b” contêm as **cargas** atribuídas às variáveis na variável canônica.

$[\text{Corr}(U,V)]^2$

$\Rightarrow \rho_{c1}^2$ e e_1 são o maior autovalor e o autovetor de

$\Rightarrow \rho_{c1}^2$ e f_1 são o maior autovalor e o autovetor de

$$\Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1/2}$$

$$\Sigma_{22}^{-1/2} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1/2}$$

Correlação Canônica

$$\max_{a,b} \text{Corr}(U,V) = \rho_{c1} \quad \Rightarrow \quad \begin{aligned} U_1 &= a'_1 Y_1 = e'_1 \Sigma_{11}^{-1/2} Y_1 \\ V_1 &= b'_1 Y_2 = f'_1 \Sigma_{22}^{-1/2} Y_2 \end{aligned}$$

O **k-ésimo par de variáveis canônicas** (U_k e V_k , com $k \leq \min(n,p,q)$) representam as combinações lineares de cada conjunto de variáveis, com máxima correlação e independente das demais:

$$U_k = e'_k \Sigma_{11}^{-1/2} Y_1, \quad V_k = f'_k \Sigma_{22}^{-1/2} Y_2; \quad \text{Corr}(U_k, V_k) = \rho_{ck}$$

k-ésimo coeficiente de correlação canônico.

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

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

Resumindo:

Correlação Canônica

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

$$U_1 = a'_1 Y_1 = e'_1 \Sigma_{11}^{-1/2} Y_1 \quad V_1 = b'_1 Y_2 = f'_1 \Sigma_{22}^{-1/2} Y_2$$

$\Rightarrow \lambda_1$ 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 $(U_2, V_2), \dots, (U_k, V_k), \dots, (U_m, V_m)$ satisfazem:

$$\left\{ \begin{array}{l} \text{Var}(U_k) = \text{Var}(V_k) = 1 \\ \text{Cov}(U_k, U_l) = \text{Corr}(U_k, U_l) = 0 \quad k \neq l \\ \text{Cov}(V_k, V_l) = \text{Corr}(V_k, V_l) = 0 \quad k \neq l \\ \text{Cov}(U_k, V_l) = \text{Corr}(U_k, V_l) = 0 \quad k \neq l \end{array} \right.$$

\Rightarrow

$$\text{Cov}(U, V) = \begin{pmatrix} I_m & \Lambda^{1/2} \\ \Lambda^{1/2} & I_m \end{pmatrix};$$
$$\Lambda^{1/2} = (\sqrt{\lambda_j} = \rho_{cj})$$

Correlação Canônica

Considere as variáveis, Y1 e Y2, padronizadas:

$$Y_i = \begin{bmatrix} Y_{1i(p \times 1)} \\ Y_{2i(q \times 1)} \end{bmatrix} \Rightarrow Y_{i(p+q) \times 1}^* = \begin{bmatrix} Y_{1i p \times 1}^* \\ Y_{2i q \times 1}^* \end{bmatrix} = \begin{bmatrix} Y_{1ji}^* = (Y_{1ji} - \mu_{1j}) / \sigma_{1j} \\ Y_{2ki}^* = (Y_{2ki} - \mu_{2k}) / \sigma_{2k} \end{bmatrix} \quad \begin{matrix} i=1, \dots, n \\ j=1, \dots, p \quad k=1, \dots, q \end{matrix}$$

\Rightarrow As variáveis canônicas (dos dados padronizados) são da forma:

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

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

As correlações canônicas
são invariantes por
padronização dos dados!

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

$\Rightarrow \lambda_k^*, f_k^*$: 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

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

$$U_k = a'_k Y_1 = e'_k \Sigma_{11}^{-1/2} Y_1$$

$$V_k = b'_k Y_2 = f'_k \Sigma_{22}^{-1/2} Y_2$$

Variáveis Padronizadas

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

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

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

$$\left. \begin{aligned} a'_k Y_1 &= a'_k (Y_1 - \mu_1) = a_{k1} (Y_{11} - \mu_{11}) + \dots + a_{kp} (Y_{1p} - \mu_{1p}) \\ &= \boxed{a_{k1} \sqrt{\sigma_{11}}} \frac{(Y_{11} - \mu_{11})}{\sqrt{\sigma_{11}}} + \dots + \boxed{a_{kp} \sqrt{\sigma_{pp}}} \frac{(Y_{1p} - \mu_{1p})}{\sqrt{\sigma_{pp}}} \\ &= a_{k1}^* Y_{11}^* + \dots + a_{kp}^* Y_{1p}^* = a_k^{*'} Y_1^* \end{aligned} \right\} \Rightarrow \begin{aligned} a_k^{*'} &= a_k' D_{11}^{1/2} \\ b_k^{*'} &= b_k' D_{22}^{1/2} \end{aligned}$$

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

$$U_k = a'_k Y_1$$

$$V_k = b'_k Y_2$$

\Rightarrow

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

$$U_k^* = a_k'^* Y_1^* = a_k' D_{11}^{1/2} Y_1^*$$

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

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

A correlação canônica é invariante por padronização

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

$$= a_k' \text{Corr}((Y_1 - \mu_1), (Y_2 - \mu_2)) b_k = a_k' \text{Corr}(Y_1, Y_2) b_k = \rho_c(U_k, V_k)$$

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

Correlação Canônica

Morfometria cefálica para os dois primeiros filhos de 25 famílias

Família	1° Filho		2° Filho	
	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:

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

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

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

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

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

$$\begin{array}{l} \text{Coeficientes das} \\ \text{Variáveis canônicas:} \end{array} \left\{ \begin{array}{ll} A_{2 \times 2}^* = \begin{pmatrix} a_1^* & a_2^* \end{pmatrix} & a_1^* = \begin{pmatrix} 0,552 \\ 0,522 \end{pmatrix} & a_2^* = \begin{pmatrix} 1,367 \\ -1,378 \end{pmatrix} \\ B_{2 \times 2}^* = \begin{pmatrix} b_1^* & b_2^* \end{pmatrix} & b_1^* = \begin{pmatrix} 0,505 \\ 0,538 \end{pmatrix} & b_2^* = \begin{pmatrix} 1,767 \\ -1,757 \end{pmatrix} \end{array} \right.$$

Correlação Canônica

Morfometria cefálica para os dois primeiros filhos de 25 famílias

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

$$U_1^* = 0,552 Y_{C_C_1}^* + 0,522 Y_{P_C_1}^*$$

$$V_1^* = 0,505 Y_{C_C_2}^* + 0,538 Y_{P_C_2}^*$$

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:

$$U_2^* = 1,367 Y_{C_C_1}^* - 1,378 Y_{P_C_1}^*$$

$$V_2^* = 1,767 Y_{C_C_2}^* - 1,757 Y_{P_C_2}^*$$

Correlação Canônica

Morfometria cefálica para os dois primeiros filhos de 25 famílias

Análise de Correlação Canônica das Variáveis Padronizadas:

$$U_1^* = 0,552 Y_{C_C_1}^* + 0,522 Y_{P_C_1}^*$$

$$V_1^* = 0,505 Y_{C_C_2}^* + 0,538 Y_{P_C_2}^*$$

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



Análise de Correlação Canônica das Variáveis Originais:

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

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

$$U_1 = 0,057 Y_{C_C_1} + 0,071 Y_{P_C_1}$$

$$V_1 = 0,050 Y_{C_C_2} + 0,080 Y_{P_C_2}$$

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

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

$$r(U^*1, V^*1) = 0,789 \quad r(U1, V1) = 0,789$$

Há interesse em calcular as correlações entre as variáveis canônicas e cada uma das variáveis originais (ou padronizadas) \Rightarrow calcular as correspondentes correlações de Pearson

Correlação Canônica

Morfometria cefálica para os dois primeiros filhos de 25 famílias

Interpretação das variáveis canônicas das variáveis padronizadas



Cargas:

$$A^* = \begin{pmatrix} 0,552 & 0,522 \\ 1,367 & -1,378 \end{pmatrix}'$$

$$B^* = \begin{pmatrix} 0,505 & 0,538 \\ 1,767 & -1,757 \end{pmatrix}'$$

Correlações:

$$\text{Corr}(U_1^*, Y_{11}^*) = 0,935 \quad \text{Corr}(U_1^*, Y_{12}^*) = 0,927$$

$$\text{Corr}(U_2^*, Y_{11}^*) = 0,354 \quad \text{Corr}(U_2^*, Y_{12}^*) = -0,373$$

$$\text{Corr}(V_1^*, Y_{11}^*) = 0,737 \quad \text{Corr}(V_1^*, Y_{12}^*) = 0,731$$

$$\text{Corr}(V_2^*, Y_{11}^*) = 0,019 \quad \text{Corr}(V_2^*, Y_{12}^*) = 0,0191$$

$$\text{Corr}(U_1^*, Y_{21}^*) = 0,754 \quad \text{Corr}(U_1^*, Y_{22}^*) = 0,758$$

$$\text{Corr}(U_2^*, Y_{21}^*) = 0,016 \quad \text{Corr}(U_2^*, Y_{22}^*) = -0,014$$


$$\text{Corr}(V_1^*, Y_{21}^*) = 0,956 \quad \text{Corr}(V_1^*, Y_{22}^*) = 0,961$$

$$\text{Corr}(V_2^*, Y_{21}^*) = 0,292 \quad \text{Corr}(V_2^*, Y_{22}^*) = -0,274$$

Note que o primeiro par de variáveis canônicas, (U_1^*, V_1^*) , tem as maiores correlações com as correspondentes variáveis padronizadas.

Correlação Canônica

Morfometria cefálica para os dois primeiros filhos de 25 famílias


$$Y_{25 \times (2+2)}^* = \begin{pmatrix} Y_{1 \ 25 \times 2}^* & Y_{2 \ 25 \times 2}^* \end{pmatrix} \rightarrow \begin{pmatrix} U_{25 \times 2}^* & V_{25 \times 2}^* \end{pmatrix}$$

Correlação (Y1*,U*)

	U_{1}^*	U_{2}^*
C1	0.9352877	-0.3538884
P1	0.9271512	0.3746875

Correlação (Y2*,U*)

	U_{1}^*	U_{2}^*
C2	0.7539771	-0.01572908
P2	0.7582663	0.01474027

Correlação (Y1*,V*)

	V_{1}^*	V_{2}^*
C1	0.7374817	-0.01901786
P1	0.7310660	0.02013559

Correlação (Y2*,V*)

	V_{1}^*	V_{2}^*
C2	0.9562074	-0.2926900
P2	0.9616470	0.2742901

As primeiras variáveis canônicas, U_{1}^* e V_{1}^* , têm as maiores correlações com as variáveis padronizadas.

As correlações são invariantes por padronização!

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: $(A_{p \times m}; B_{q \times m})$

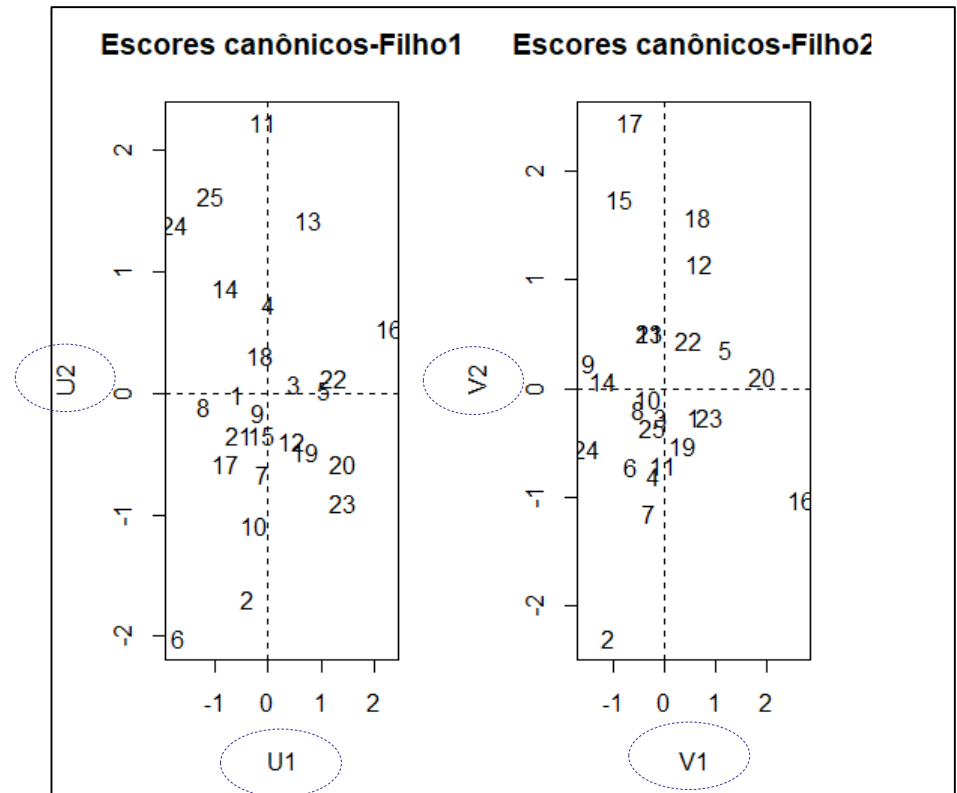
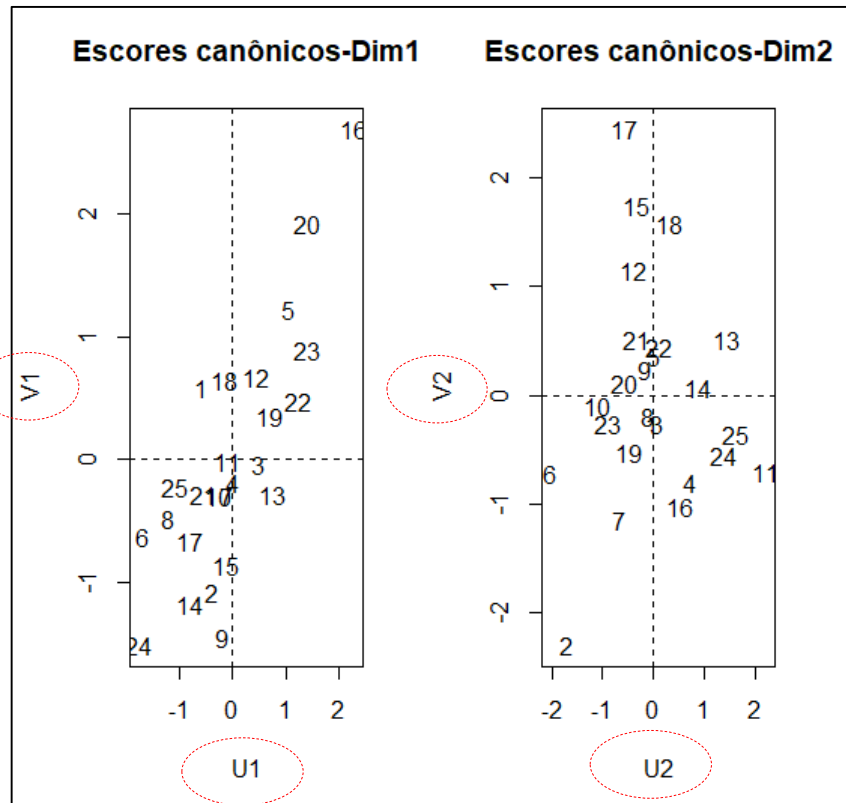
$$\begin{cases} U_{i \times m \times 1} = A' Y_{1i} \\ V_{i \times m \times 1} = B' Y_{2i} \end{cases} \left\{ \begin{array}{l} Corr(U; Y_1) = A' \Sigma_{11} D_{11}^{-1/2} = A'^* R_{11} = Corr(U^*, Y_1^*) \\ Corr(U; Y_2) = A' \Sigma_{12} D_{22}^{-1/2} = A'^* R_{12} = Corr(U^*, Y_2^*) \\ Corr(V; Y_1) = B' \Sigma_{21} D_{11}^{-1/2} = B'^* R_{21} = Corr(V^*, Y_1^*) \\ Corr(V; Y_2) = B' \Sigma_{22} D_{22}^{-1/2} = B'^* R_{22} = Corr(V^*, Y_2^*) \end{array} \right.$$

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

Correlação Canônica

CCA: Dados dos Filhos padronizados

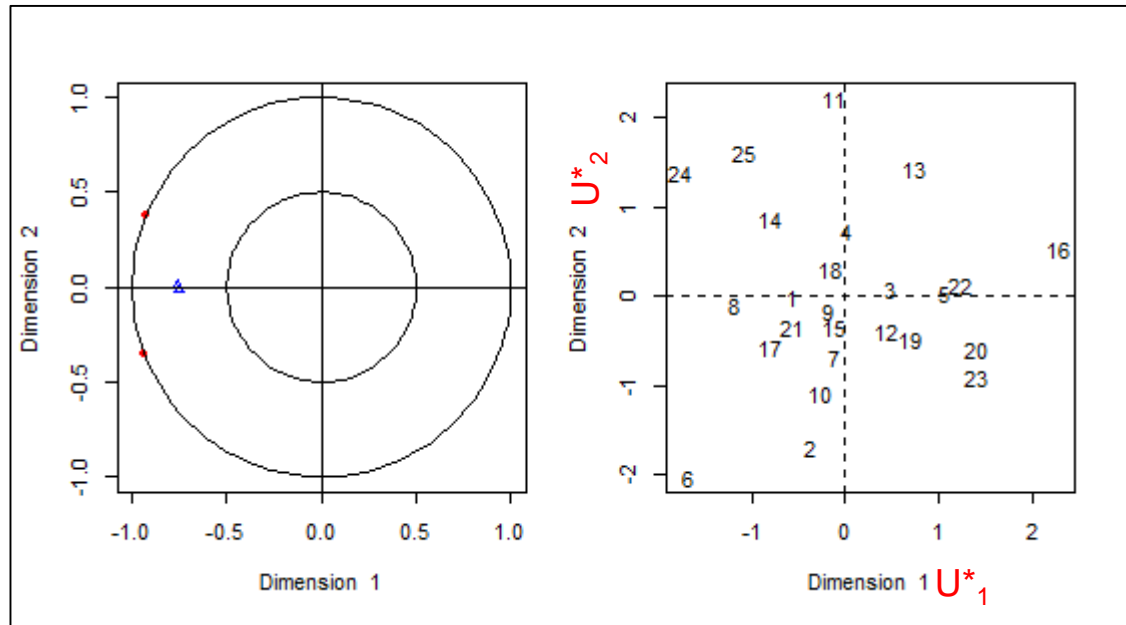
Representação dos escores canônicos



Correlação Canônica

CCA: Dados dos Filhos padronizados

Representação das cargas e dos escores canônicos



Correlação ($Y1^*, U^*$)

	U^*_1	U^*_2
C1	0.9352877	-0.3538884
P1	0.9271512	0.3746875

Correlação ($Y2^*, U^*$)

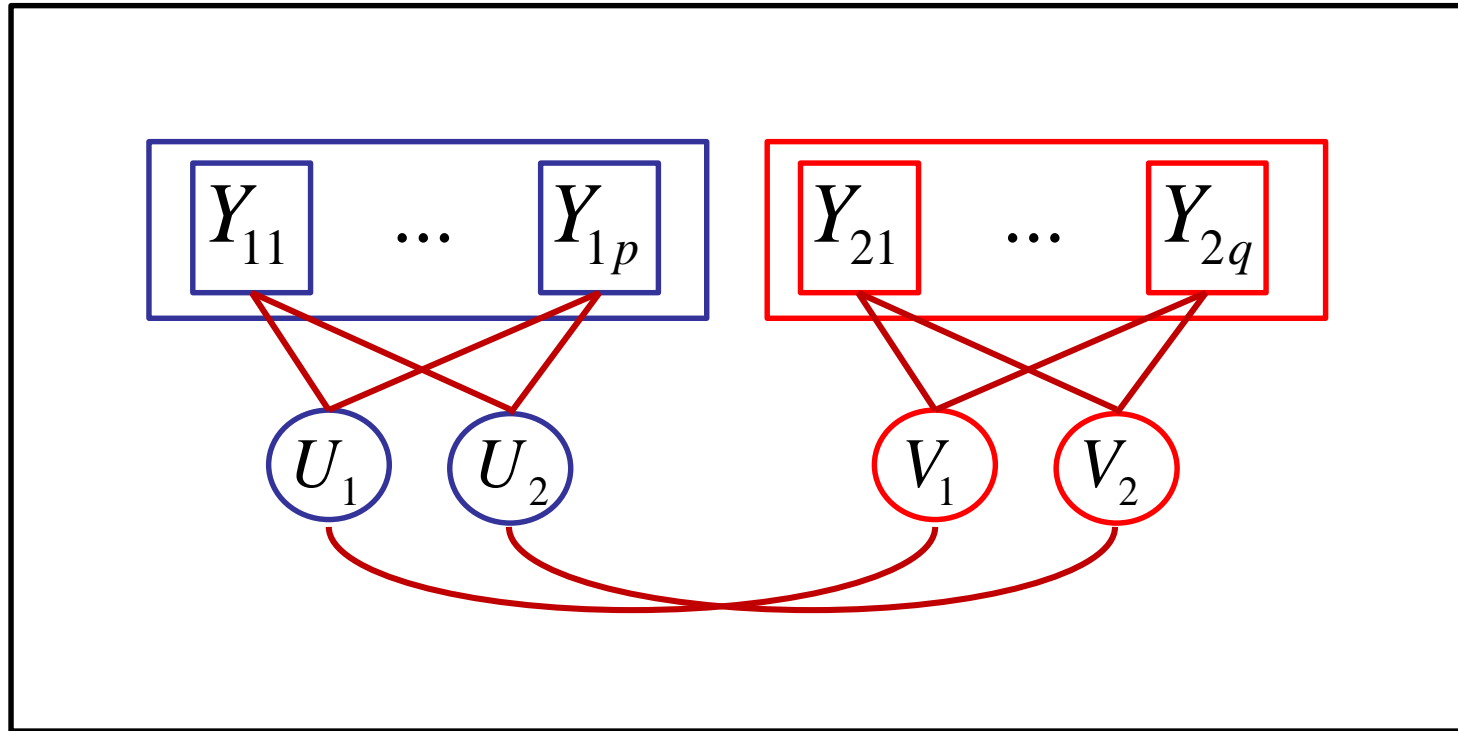
	U^*_1	U^*_2
C2	0.7539771	-0.01572908
P2	0.7582663	0.01474027

Correlações dos escores,
 U^*_1 e U^*_2 , com as
variáveis

Escores (na dimensão 1 e 2)
dos indivíduos para o primeiro
grupo de variáveis (Filho 1)

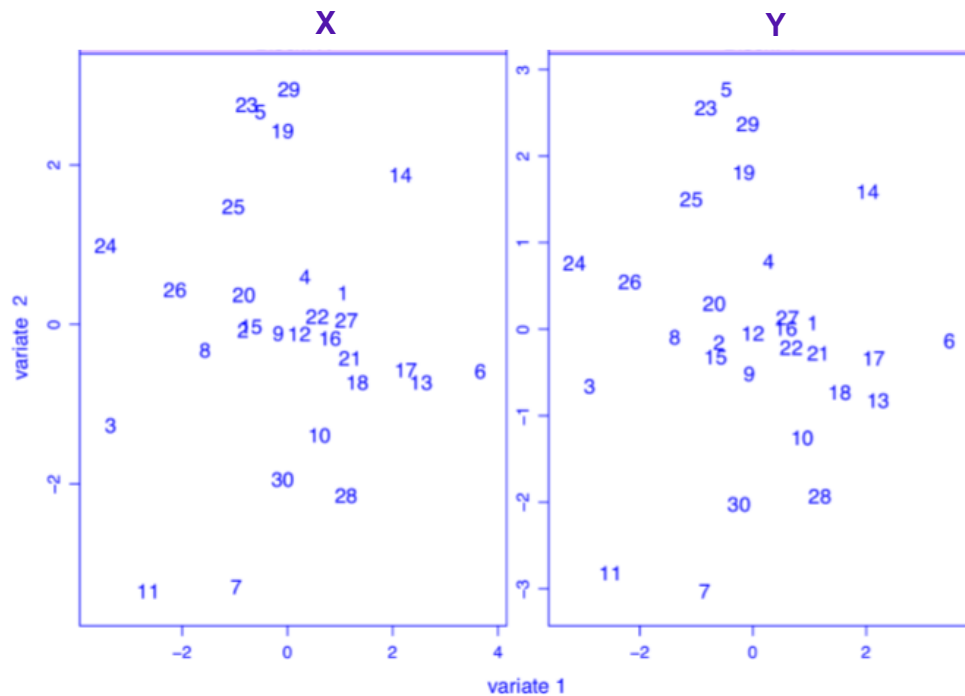
Correlação Canônica

Obtenção de Escores para a Integração de Bancos de Dados



U_1 e V_1 são escores, obtidos da redução de dimensionalidade dos dados $Y_{n \times (p+q)}$. É o par de variáveis latentes, de cada conjunto de dados, com maior correlação. Gráfico $U_1 \times V_1$ pode ser usado para representar o padrão de dispersão das observações nesta redução de dimensionalidade!

Análise de Correlação Canônica e Integração de Bancos de Dados:

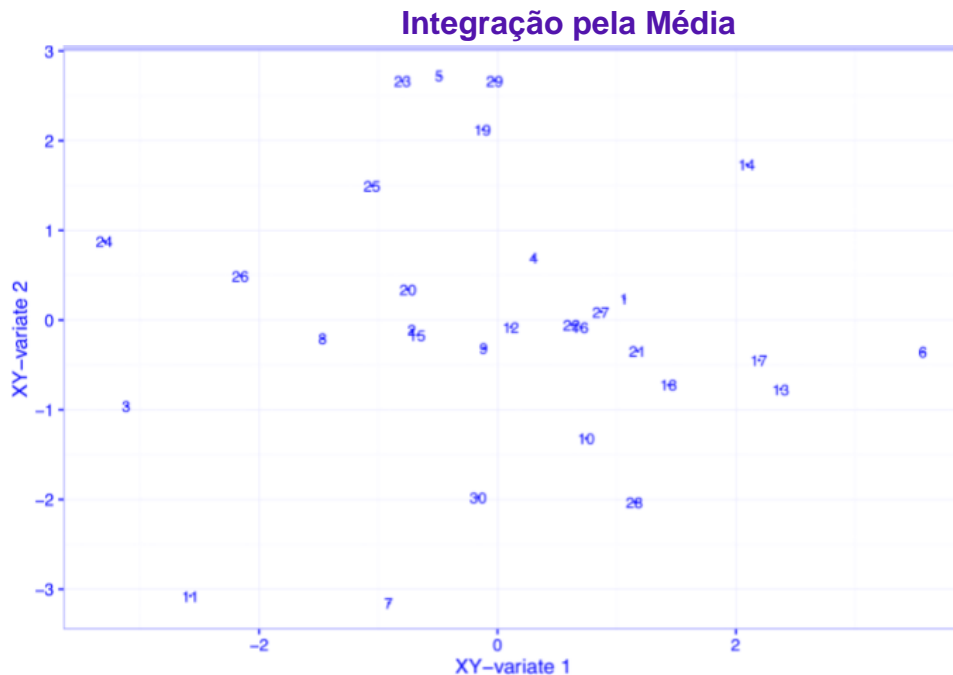


$$\begin{bmatrix} X_{30 \times 2} & Y_{30 \times 2} \end{bmatrix}$$

	X		Y	
	Var1	Var2	Var1	Var2
1				
2				
...				
30				

Integração de Bancos de Dados

Dados: mesmas variáveis (Var1 e Var2) avaliadas sob as condições X e Y



Alternativa 1: Integração pela Média (obter a média de cada variável)

Alternativa 2: Integração pela Diferença (obter a diferença entre X e Y em cada var)

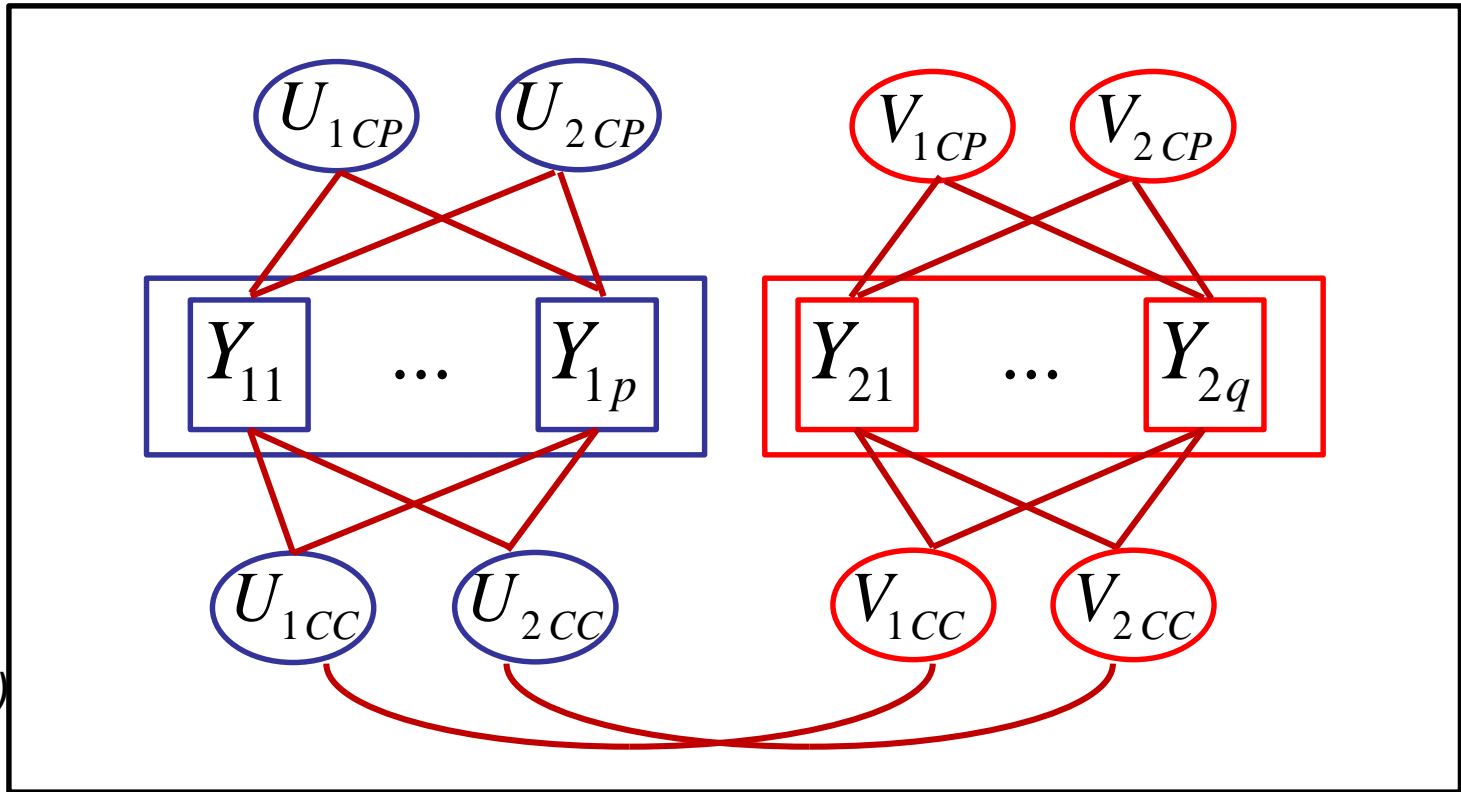
Alternativa 3: obter as variáveis canônicas U1 e V1 (Correlação canônica).

Diferentes critérios podem ser usados na Integração de BD!

Redução de Dimensionalidade

Componente
Principal
 $m \leq (n, "p")$
 $m \leq (n, "q")$

Correlação
Canônica
 $m \leq \min(n, p, q)$



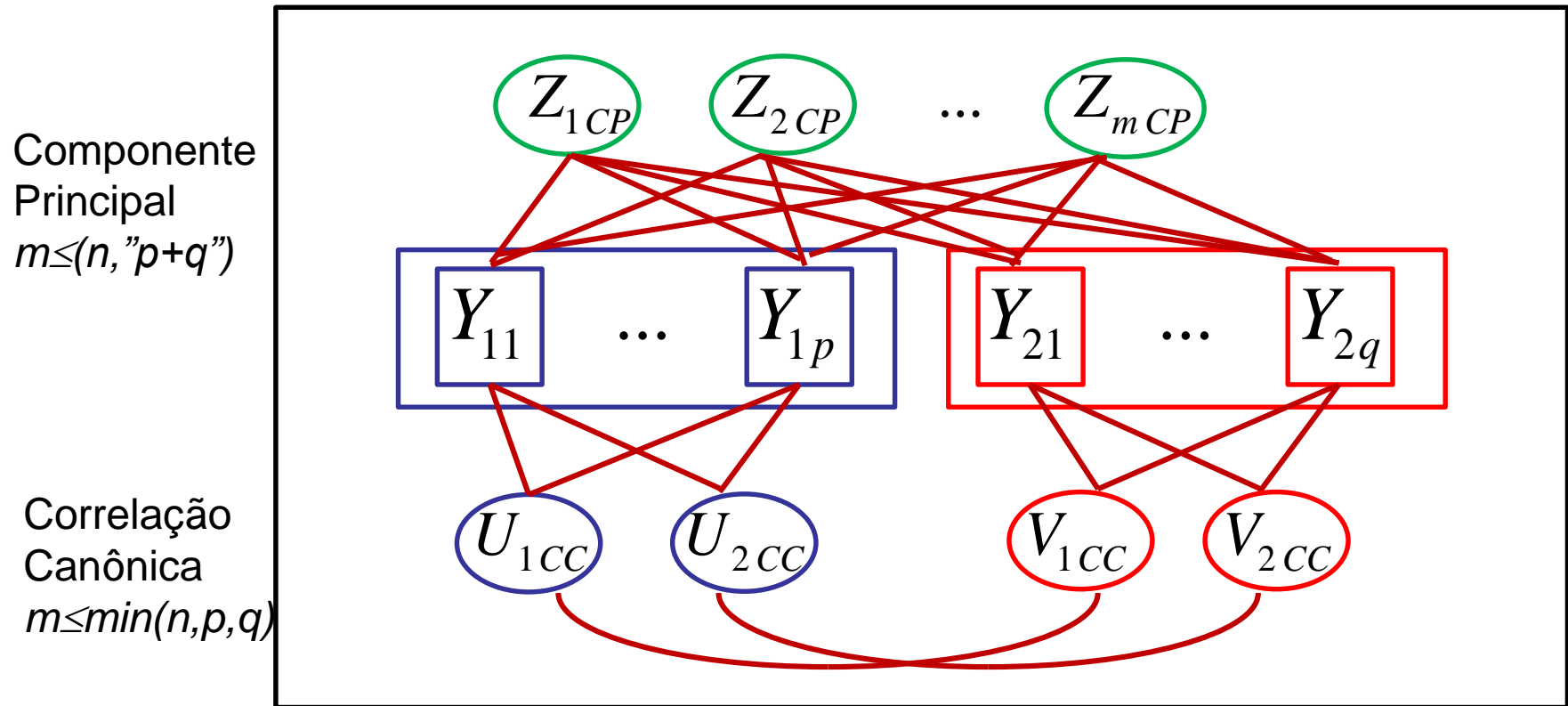
$$U_1 = a'_1 Y_1 = e'_1 \underbrace{\Sigma_{11}^{-1/2}}_{P_1 \Lambda^{-1/2} P_1'} Y_1 = e'_1 P_1 \underbrace{\Lambda^{-1/2} P_1' Y_1}_{\text{CP de } Y_1}$$

Decomposição
spectral de Σ_{11}

CP padronizado de Y_1
(Fator Comum)

Redução de Dimensionalidade

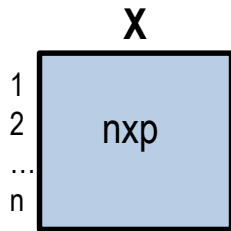
Diferentes alternativas de análise para obter os vetores reducionistas de um conjunto de dados multivariados



Análises Multivariadas

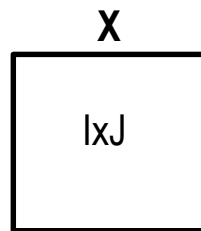
■ Dados Quantitativos
□ Dados Categóricos

Análise Não-Supervisionada



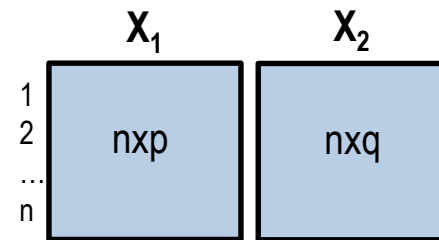
CP, CoP

Análise Não-Supervisionada



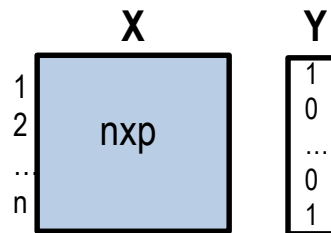
An Correspond.

Análise Não-Supervisionada



ACC

Análise Supervisionada

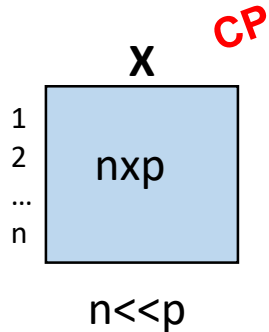


AD

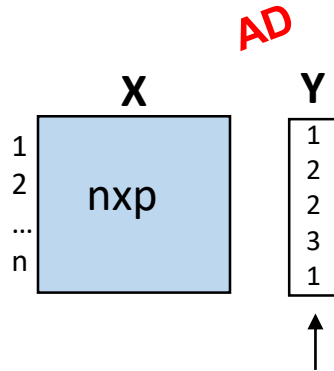
N-Integração de Bancos de Dados

Um Único BD

Análise Não-Supervisionada



Análise Supervisionada



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

P-Integração entre múltiplos níveis de informação avaliados nas mesmas variáveis (metanálise)!

X: Matriz(es) de Dados

Y: Resposta de interesse
(em geral, Classes)

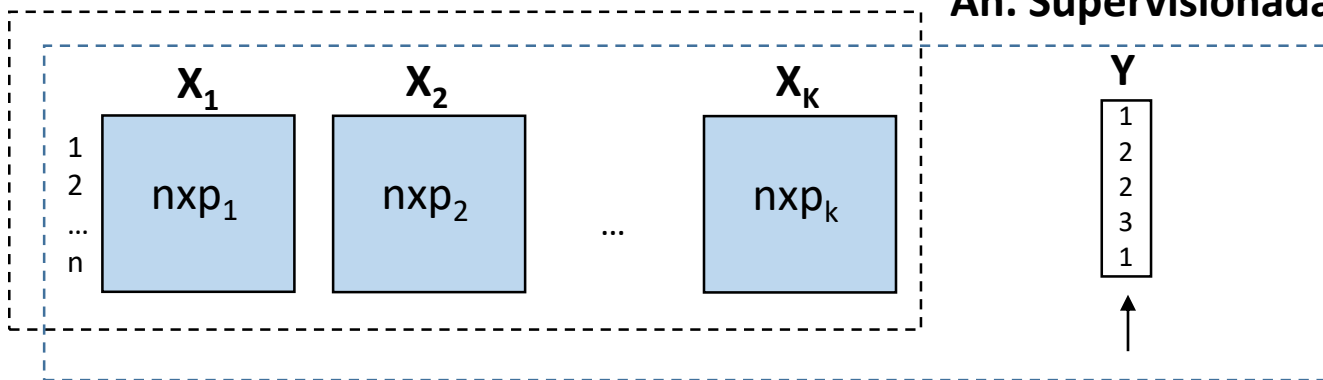
Múltiplos BD (Multimodais, Multivisão)

Análise Não-Supervisionada

CCg

An. Supervisionada

ADg



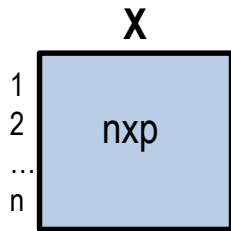
⇒ **Multionics
(R)**

**Análises
Generalizadas**

Análises Multivariadas

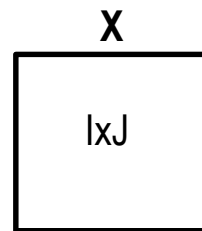
■ Dados Quantitativos
□ Dados Categóricos

Análise Não-Supervisionada



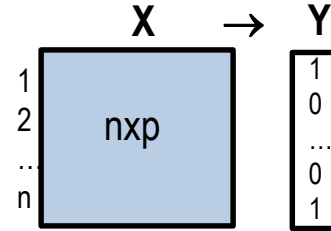
CP, CoP

Análise Não-Supervisionada



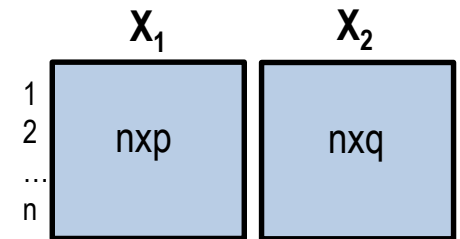
An Correspond.

Análise Supervisionada



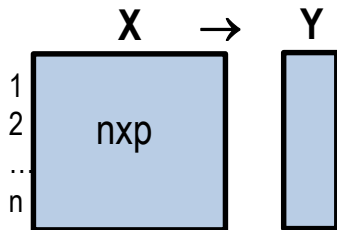
AD

Análise Não-Supervisionada



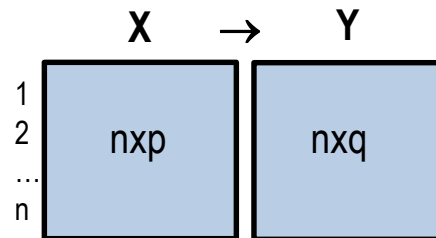
ACC

Análise Supervisionada



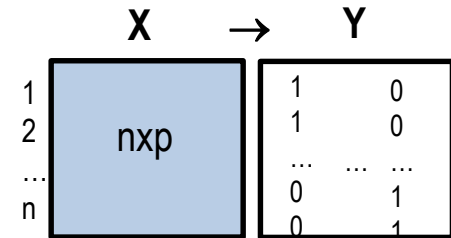
PLS (Partial Least Square)

Análise Supervisionada



PLS para múltiplas respostas

Análise Supervisionada



ACC_AD