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
- ✓ 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 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 entre elas.

$\Rightarrow$ 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?

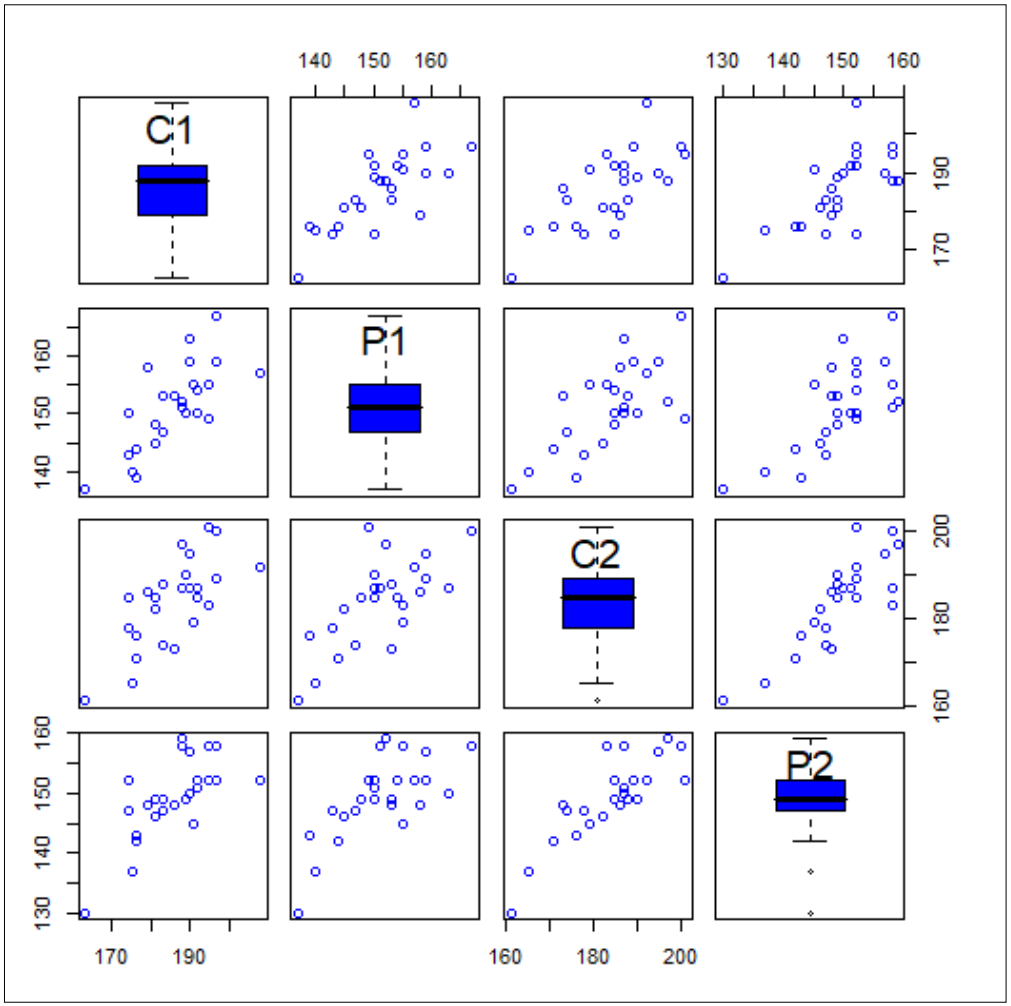
Discuta a estrutura dos dados.

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

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



Correlações:

	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 (1° e 2° filho)

.

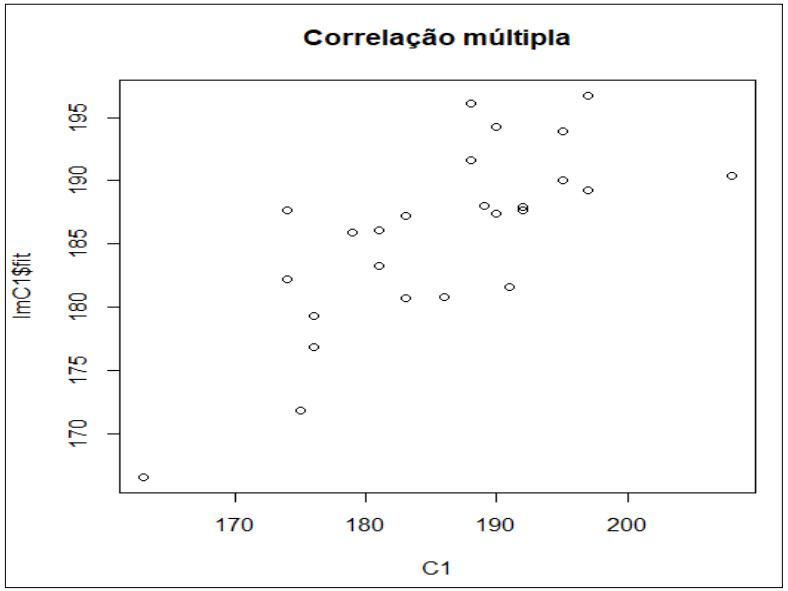
← Correlação ENTRE os grupos, para cada par de variável.

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).

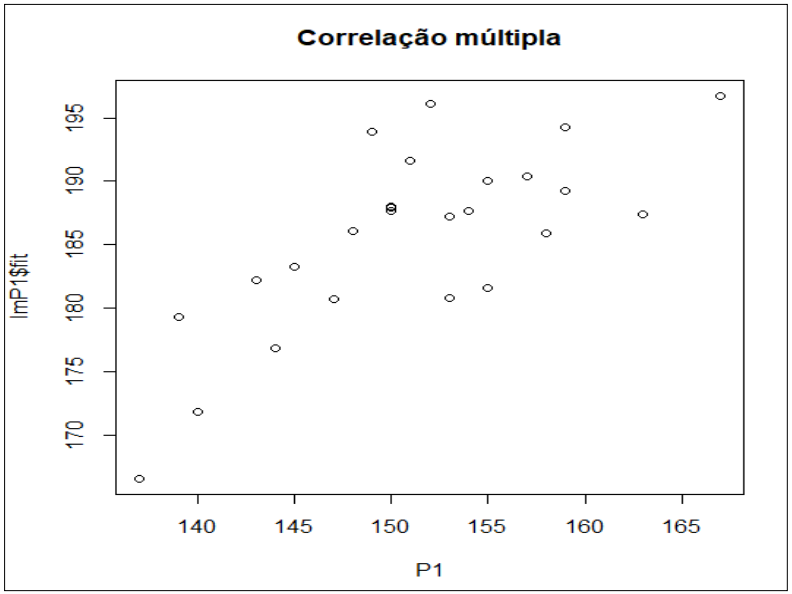
$$\rho_M\left[Y_{C1},\left(Y_{C2},Y_{P2}\right)\right]$$



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

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

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



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

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

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}).

	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

Outra medida de correlação:

Correlação Canônica - Exemplos

	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}		$Y_{n(p+q)}$

- 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!

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.

$$Y_{8\times(4+4)} = \left(Y1_{8\times4} \quad Y2_{8\times4}\right)$$

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

Correlação Canônica – Exemplos

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

		Expert 1			Expert 2				Expert 3		
Wine	Oak-type	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

$$Y_{6\times(3+4+3)} = \left(Y1_{6\times3} \quad Y2_{6\times4} \quad Y3_{6\times3}\right)$$ Correlação Canônica Múltipla

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

$$\left(Y1_{6\times3} \quad Y2_{6\times4}\right)$$

$$\left(Y1_{6\times3} \quad Y3_{6\times3}\right)$$

$$\left(Y2_{6\times4} \quad Y3_{6\times3}\right)$$

Correlação Canônica

Notação

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

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

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

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

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

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

Mede a covariância entre os dois conjuntos de variáveis

$$E\left(Y_i\right) = \mu_{(p+q)\times 1} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

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

Partição da matriz de covariância!

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

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

Problema de otimização bilinear decomposto em otilizações lineares!

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 (U_k,V_k), com k=1,2,…,min(n,p,q), representa o par de 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:

$$U^*_k = a^*_k{}' Y^*_1 = e^*_k{}'\, R_{11}^{-1/2}\, Y^*_1$$

$$V^*_k = b^*_k{}' Y^*_2 = f^*_k{}'\, R_{22}^{-1/2} Y^*_2$$

$$Corr\left(U^*_k, V^*_k\right) = \frac{a^*_k{}'\ \rho_{12}\ b^*_k}{\sqrt{a^*_k{}'\ \rho_{11}\ a^*_k}\sqrt{b^*_k{}'\ \rho_{22}\ b^*_k}} = \rho_{ck}$$

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

Resultado importante!

Interpretação Geométrica

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

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

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

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

Veja os Comandos R da aula!

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:

Matriz de Correlações

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

Número de variáveis canônicas $\leq \min(n,p\ q) = 2$

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

Redução para uma única dimensão deve bastar!

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

Autovetores:
Coeficientes (cargas) 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

Dados: Morfometria cefálica para os dois primeiros filhos de 25 famílias

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

$$U_1^* = 0,552\, Y^*_{C1} + 0,522\, Y^*_{P1} \qquad V_1^* = 0,505\, Y^*_{C2} + 0,538\, Y^*_{P2}$$

Estas são responsáveis pela maior correlação (r=0,789) entre as variáveis cefálicas dos dois primeiros filhos das famílias estudadas. As variáveis individuais contribuem com “pesos” muito próximos (≅média artimética).

A segunda variável canônica explica muito pouco (r=0,054) da correlação entre as variáveis dos dois primeiros filhos, sendo definida por:

$$U_2^* = 1,367\, Y^*_{C1} - 1,378\, Y^*_{P1} \qquad V_2^* = 1,767\, Y^*_{C\_C\_2} - 1,757\, Y^*_{P\_C\_2}$$

Correlação Canônica

Morfometria cefálica para os dois primeiros filhos de 25 famílias

Análise de Correlação Canônica das Variáveis Padronizadas:

$$U_1^* = 0,552\, Y^*_{C1} + 0,522\, Y^*_{P1}$$

$$V_1^* = 0,505\, Y^*_{C2} + 0,538\, Y^*_{P2}$$

$$\hat{\rho}_1^* = Corr\left(U_1^*, V_1^*\right) = 0,789$$

Análise de Correlação Canônica das Variáveis Originais:

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

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

$$U_1 = 0,057\, Y_{C1} + 0,071\, Y_{P1}$$

$$V_1 = 0,050\, Y_{C2} + 0,080\, Y_{P2}$$

$$\hat{\rho}_1 = Corr\left(U_1, V_1\right) = 0,789$$

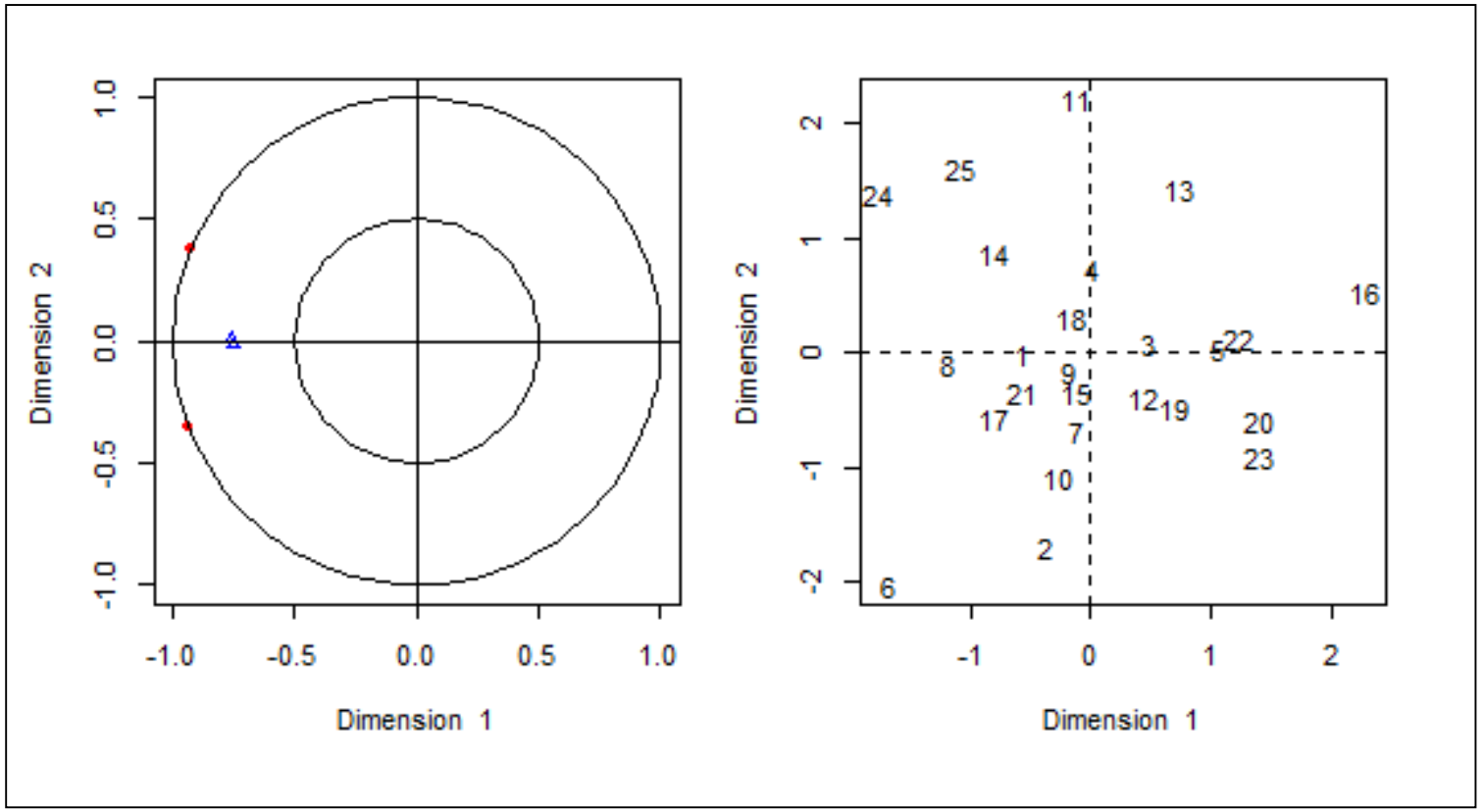
Correlação Canônica

Variáveis originais				Variáveis padronizadas				Variáveis canônicas			
Y_C1	**Y_P1**	**Y_C2**	**Y_P2**	**Z_C1**	**Z_P1**	**Z_C2**	**Z_P2**	**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

CCA-Filhos: Dados padronizados
Representação das cargas e dos escores canônicos



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

$$Y^*_{25\times(2+2)} = \left(Y^*_{1\ 25\times2} \quad Y^*_{2\ 25\times2} \right) \rightarrow \left(U^*_{25\times2} \quad V^*_{25\times2} \right)$$

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.

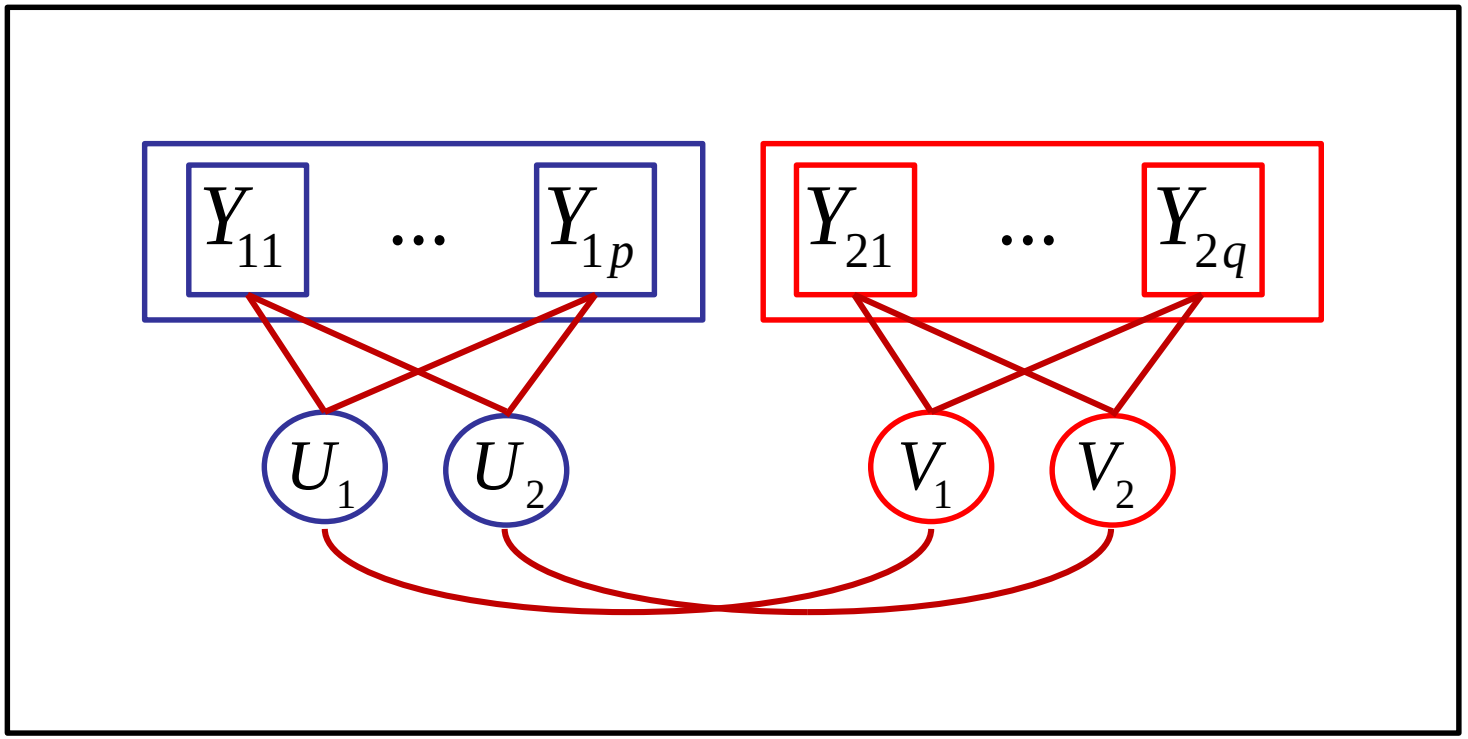
As correlações são invariantes por padronização!

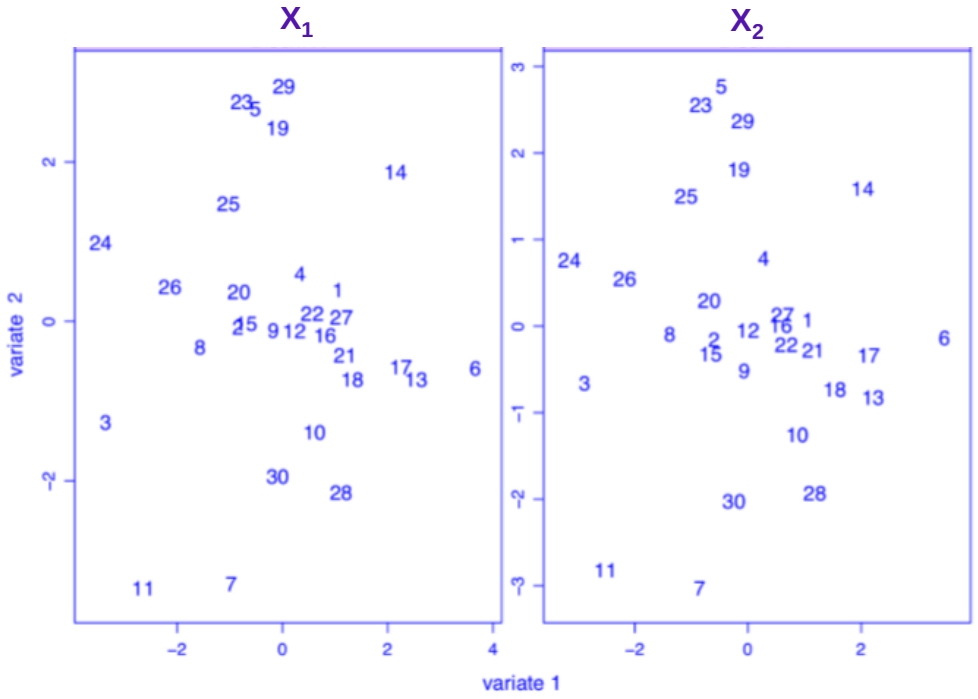
Correlação Canônica

Y_C1	Y_P1	Y_C2	Y_P2	Z_C1	Z_P1	Z_C2	Z_P2	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

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

Correlação Canônica
Integração de Bancos de Dados



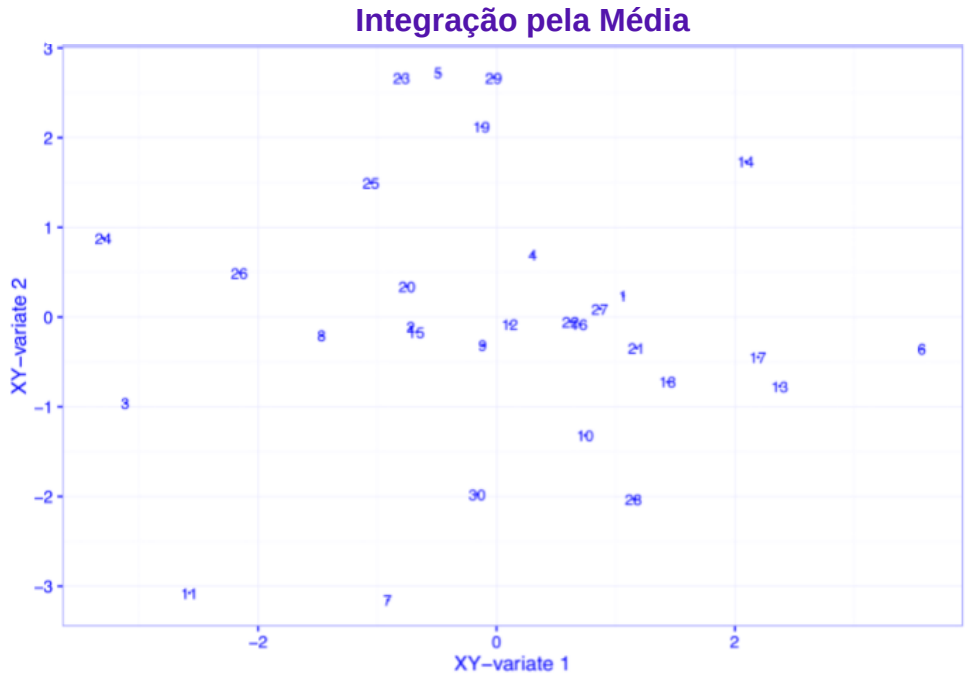


Análise de Correlação Canônica e Integração de Bancos de Dados:

$$\left[X_{1\,n\times 2} \quad X_{2\,n\times 2} \right]$$

	X_1		X_2	
	Var1	Var2	Var1	Var2
1				
2				
…				
30				

Integração de Bancos de Dados



Situação: mesmas variáveis (Var1 e Var2) avaliadas sob duas condições

Alternativa 1: obter a média das variáveis nos dois blocos.
Alternativa 2: obter a diferença entre as variáveis dos dois blocos.
Alternativa 3: obter as variáveis canônicas de cada bloco (Correlação canônica).

Diferentes critérios podem ser usados na integração de BD!