MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Sem/2022 - IME

Análise Multivariada

Já vimos ☺

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

- Revisão de Metodologias Clássicas de obtenção de vetores reducionistas:

$$\Re^{p} \rightarrow \Re^{m}$$

Variáveis latentes: escores e cargas
(combinações lineares de Y)

n > p
Obs iid
Dados quantitativos

- ✓ Componentes Principais (*m≤min(n,p)*)
- ✓ Escalonamento Multidimensional – CoP (*m≤posto*(D_{nxn}))
- ✓ Análise de Correspondência (Tabelas de contingência IxJ, m<min(I-1,J-1))
- ✓ Análise Fatorial Exploratória (via CP, *m≤min(n,p);* via SEM)
- ✓ Análise de Agrupamento (das observações, das variáveis, heatmap)
- ✓ Análise Discriminante (Solução de Fisher, *m≤min(n,p, G-1); Regra de Bayes*)
- ✓ Correlação Canônica (*m≤min(n,p,q)*)

⇒ Solução em Espaços Duais ($\Re^{nxp}$, $\Re^{pxp}$, $\Re^{nxn}$)
⇒ Visualização dos Dados: Representações Bi-Plot

Revisando ☺

Componentes Principais (Pearson, 1901)

$$Y_{i\,p\times 1} \overset{iid}{\sim} (\mu\,;\Sigma\,)$$

Suposições: AASn de uma única população com matriz de cov Σ, obs independentes, n>p

Unidades Amostrais	Variáveis 1	2	…	j	…	p
1	Y_{11}	Y_{12}		Y_{1j}		Y_{1p}
…	…	…	…	…	…	…
n	Y_{n1}	Y_{n2}		Y_{nj}		Y_{np}

Redução de dimensionalidade:

$$\boxed{\Re^p \to \Re^m; \quad m \le \min(n, p)}$$

$$Y_{n\times p} \to Z_{n\times m}; \; Z_{ki} = V_k{}'Y_i$$

escore

carga

$$V_k\;; \quad \underset{\|a_k\|=1}{\arg\max} \frac{a_k{}'\Sigma a_k}{a_k{}'a_k}$$

$$= \underset{\|a_k\|=1}{\arg\max} \sum_{k=1}^{m} Var\left(Z_{ki}\right)$$

tr(Σ)

Solução: Decomposição Espectral em $\Re^{pxp}$

$$\Sigma = V\Lambda V'; \quad \Lambda = D_{\lambda_j}; VV' = V'V = I_p; \quad \left|\Sigma - \lambda I_p\right| = 0; \quad \Sigma V_k = \lambda_k V_k$$

$$\Sigma_Y = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22}^2 & \dots & \sigma_{2p} \\ & & \dots & \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp}^2 \end{pmatrix} \boxed{= \lambda_1 V_1 V_1' + \dots + \lambda_m V_m V_m'} + \dots + \lambda_p V_p V_p'; \quad \lambda_1 \ge \dots \ge \lambda_m \ge \dots \ge \lambda_p$$

Componentes Principais – Coordenadas Principais

Equivalência das Soluções em Espaços Duais

$Y^*_{n\times p}$ **:** Matriz de dados ("**padronizados**") de posto r=min(n,p)

$$Y^*_{n\times p} = U^*_{n\times n} \begin{pmatrix} \Lambda_r^{*1/2} & 0 \\ 0 & 0 \end{pmatrix} V^*_{p\times p}{}'$$

Análise em $\Re^{nxp}$ (Decomposição em valores singulares)

Análise em $\Re^{nxn}$ (Decomposição espectral de YY'))

Análise em $\Re^{pxp}$ Decomposição espectral de Y'Y)

Análise em $\Re^{nxn}$

Análise em $\Re^{pxp}$

*Dnxn ↔ Y*Y*'*

Matriz de similaridade entre indivíduos

$$Y^*Y^{*\prime} = U^*\Lambda^*U^{*\prime}$$

Y'Y*=(n-1)R_u*

Matriz de similaridade entre variáveis

$$Y^{*\prime}Y^* = V^*\Lambda^{*1/2}V^{*\prime}$$

$$\boxed{U^*_{n\times n}\Lambda_m^{*1/2}} = \boxed{Y^*_{n\times p}V^*_{p\times m}} \qquad m \le r$$

Coordenadas Principais

Componentes Principais

Componentes Principais – Coordenadas Principais

Equivalência das Soluções em Espaços Duais

$Y_{n\times p}$ **:** Matriz de dados ("**originais**") de posto r=min(n,p)

$H = I_n - n^{-1}1_n1_n'$

$HH' = H'H = H^2 = H \quad (HY)_{n\times p} = U_{n\times n}\begin{pmatrix}\Lambda_r^{1/2} & 0 \\ 0 & 0\end{pmatrix}V'_{p\times p}$

HY: são as linhas de Y centradas na média das p variáveis

Análise em $\Re^{nxp}$ (Decomposição em valores singulares)

Análise em $\Re^{nxn}$ Análise em $\Re^{pxp}$

Dnxn ↔ (HY)(HY)'

$HYY'H = U\Lambda U'$ $\quad Y'HY = V\Lambda V'$

(HY)'(HY)=(n-1)S_u

$\boxed{U_{n\times n}\Lambda_m^{1/2}} = \boxed{(HY)_{n\times p}V_{p\times m}}$ $\quad m \leq r$

Coordenadas Principais

Componentes Principais

Biplots: representação gráfica simultânea de *n* observações e *p* variáveis em $\Re^2$

Sem perder generalidade, considere os dados padronizados Y:

$$Y_{n\times p} = \boxed{U_{n\times n}\ \Lambda^{1/2}}\ \boxed{V'_{p\times p}}$$

Matriz de "escores dos CP"

Matriz de "cargas"

$$\begin{cases} YY' = U\Lambda U' & \text{Análise em } \Re^{nxn} \\ Y'Y = V\Lambda V' & \text{Análise em } \Re^{pxp} \end{cases}$$

$\Re^p \to \Re^m$

m=2

$$Y_{n\times p};\quad Y \approx [U_1\ U_2]_{n\times 2}\,\Lambda_2^{1/2}\,[V_1\ V_2]'_{2\times p} = [U_1\ U_2]\,\Lambda_2^{1/2-c/2+c/2}\,[V_1\ V_2]'$$

$$Y \approx \left(U_1\lambda_1^{1/2-c/2}\quad U_2\lambda_2^{1/2-c/2}\right)\left(\lambda_1^{c/2}V_1\quad \lambda_2^{c/2}V_2\right)'$$

Análises sob os mesmos autovalores

$U_1\lambda_1^{1/2-c/2} \quad\times\quad U_2\lambda_2^{1/2-c/2}$ *n pontos*

$\lambda_1^{c/2}V_1 \quad\times\quad \lambda_2^{c/2}V_2$ *p pontos*

c=0: linhas em coordenadas principais e colunas em coordenadas padronizadas (mapa assimétrico)
c=1: linhas em coordenadas padronizadas e colunas em coordenadas principais (mapa assimétrico)
c=1/2: representação sob diferentes "*zoons*"

Análise Fatorial Exploratória (Spearman, 1904)

Redução de dimensionalidade $\Rightarrow$ Modelagem estrutural da dependência de "p" variáveis por meio de "m" fatores comuns além de termos específicos:

$\Re^p \to \Re^m; \quad m \le \min(n, p)$ (Solução via CP)

$$Y_{n\times p};\begin{cases} Y_1^i - \mu_1 = \phi_{11}F_{i1} + \phi_{12}F_{i2} + \ldots + \phi_{1m}F_{im} + e_{i1} \\ Y_2^i - \mu_2 = \phi_{21}F_{i1} + \phi_{22}F_{i2} + \ldots + \phi_{2m}F_{im} + e_{i2} \\ \quad\ldots \\ Y_p^i - \mu_p = \phi_{p1}F_{i1} + \phi_{p2}F_{i2} + \ldots + \phi_{pm}F_{im} + e_{ip} \end{cases}$$

Suposição:

$$Y_i - \mu = \Phi\,\mathbf{f}_i + e_i;$$

$$\mathbf{f}_{i\,m\times 1} \overset{iid}{\sim} (0; I_m) \perp e_{i\,p\times 1} \overset{iid}{\sim} (0; \Psi)$$

$$\Leftrightarrow \boxed{\Sigma_{p\times p} \cong \underbrace{\Phi_{p\times m}\Phi'_{m\times p}}_{\text{Comunalidade}} + \underbrace{\Psi_{p\times p}}_{\text{Especificidade}}}$$

$\Phi = (\phi_{ij})$: Matriz de cargas fatoriais

Ψ : Matriz (diagonal) de fatores específicos

$\mathbf{f} = (F_1, \ldots, F_m)'$: Vetor de fatores comuns (escores, variáveis latentes)

Solução via Componentes Principais

$$\Sigma = V\Lambda V' = (V\Lambda^{1/2})(V\Lambda^{1/2})'$$

$$\Rightarrow Z_{ip\times 1} = V' Y_i$$

$$\Rightarrow \boxed{Y_i = \underbrace{(V\Lambda^{1/2})}_{\text{Cargas fatoriais}}\underbrace{(\Lambda^{-1/2}Z_i)}_{\text{Fatores comuns (CP padronizados)}} + e_i}$$

$\to$ Solução via MVS

Análise Discriminante Linear (Fisher, 1938)

Análise supervisionada

$$\Re^{(p+1)} \to \Re^{m};\ m \le \min(n, p; G-1)$$

Obter combinações lineares para a máxima separação dos grupos:

	Unidades Amostrais	Variáveis 1	2	…	j	…	p
Grupo1	1	Y_{111}	Y_{112}		Y_{11j}		Y_{11p}
	2	Y_{121}	Y_{122}		Y_{12j}		Y_{12p}
	…	…	…	…	…		…
	n_1	Y_{1n11}	Y_{1n12}		Y_{1n1j}		Y_{1n1p}
Grupo2	1	Y_{211}	Y_{212}		Y_{21j}		Y_{21p}
	2	Y_{221}	Y_{222}		Y_{22j}		Y_{22p}
	…	…	…	…	…		…
	n_2	Y_{2n21}	Y_{2n22}		Y_{2n2j}		Y_{2n2p}

$Y_{n\times(p+1)}$

p variáveis mais grupo

... G grupos

$$Y_{i\ p\times 1} \mid \tau_g \overset{iid}{\sim} (\mu_g; \Sigma_g) \quad \Rightarrow X_i = l' Y_i \overset{iid}{\sim} (l'\mu_g; l'\Sigma l)$$

Suposição $\Rightarrow$ $\Sigma_g = \Sigma$

Solução (linear) de Fisher:

$$\frac{l' \sum_{g=1}^{G} (\bar{Y}_g - \bar{Y}_.)(\bar{Y}_g - \bar{Y}_.)' l}{l' S_W l} = \boxed{\frac{l' SS_B\, l}{l' S_W l}}$$

Situação ideal para discriminação: variáveis com covariâncias ENTRE e DENTRO de sinais contrários!

$$SS_T = SS_B + SS_W; \quad S_W = SS_W / (n-G) \leftarrow \text{MANOVA}$$

Redução de Dimensionalidade
Obtenção de Vetores Reducionistas

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

- Componentes Principais: $f\left(\Sigma;a\right) = \boxed{\dfrac{a'\Sigma a}{a'a}, \quad a'a = 1} \Rightarrow Z_{ki} = a_k{}'Y_i \qquad Cov(Y) = \Sigma$

- Análise Fatorial Exploratória (via CP): $\Rightarrow F_{ki} = \lambda^{-1/2} Z_{ki} \qquad \Sigma = \Phi\Phi' + Diag\left(\Psi_{jj}\right)$

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

  $Y_{n\times p}; n = \sum n_g$

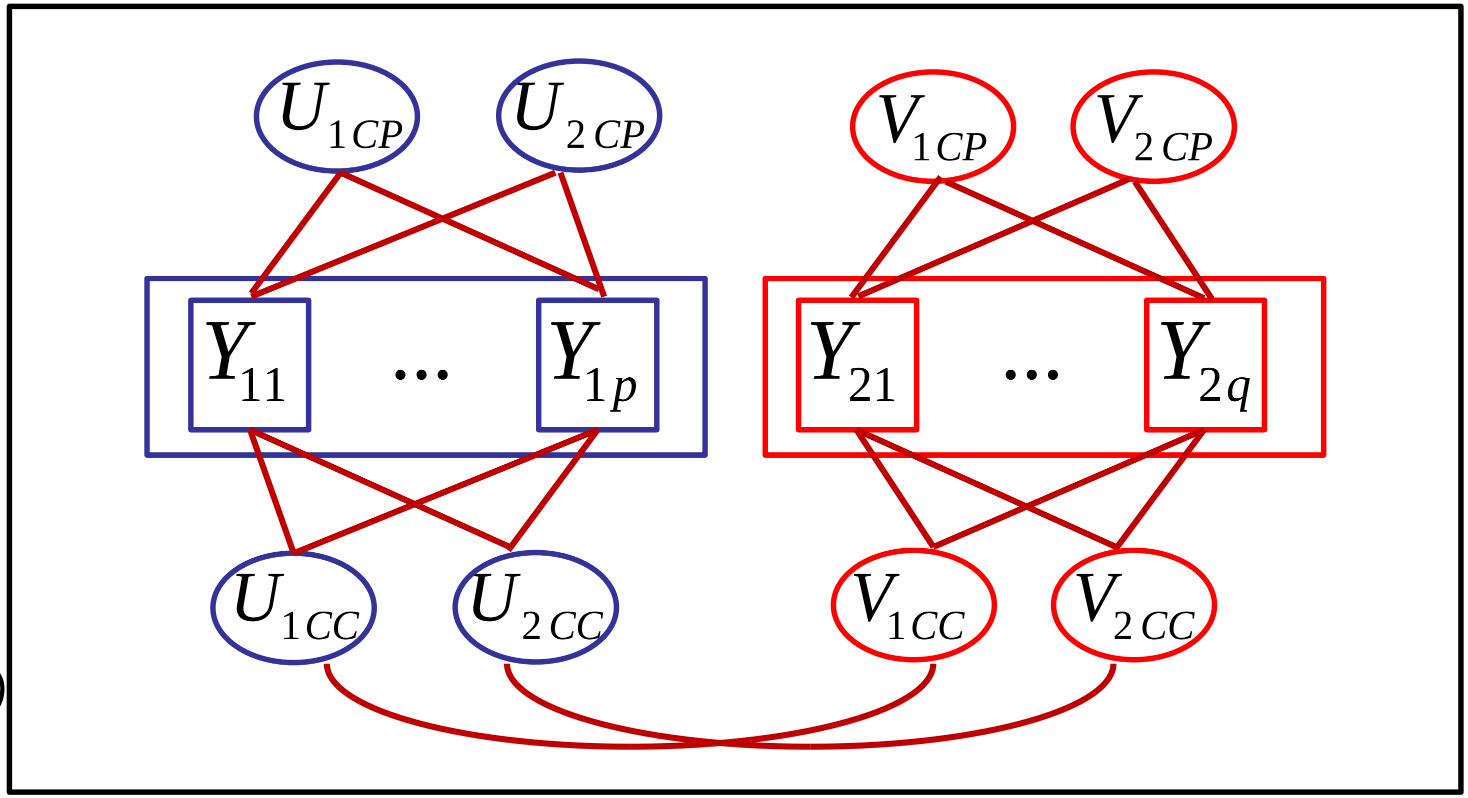
- Análise de Correlação Canônica:

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

$$\begin{cases} f_1\left(\Sigma_{11}^{-1}\mathbf{\Sigma}_{12}\Sigma_{22}^{-1}\,\Sigma_{21};a\right) = \boxed{\dfrac{a'\mathbf{\Sigma}_{12}\Sigma_{22}^{-1}\,\Sigma_{21}a}{a'\Sigma_{11}a}, \quad a'\Sigma_{11}a = 1} \\[2em] f_2\left(\Sigma_{22}^{-1}\mathbf{\Sigma}_{21}\Sigma_{11}^{-1}\,\Sigma_{12};b\right) = \boxed{\dfrac{b'\mathbf{\Sigma}_{21}\Sigma_{11}^{-1}\,\Sigma_{12}b}{b'\Sigma_{22}b}, \quad b'\Sigma_{22}b = 1} \end{cases}$$

Redução de Dimensionalidade

Componente Principal
m≤(n,"p")



Correlação
Canônica
m≤min(n,p,q)

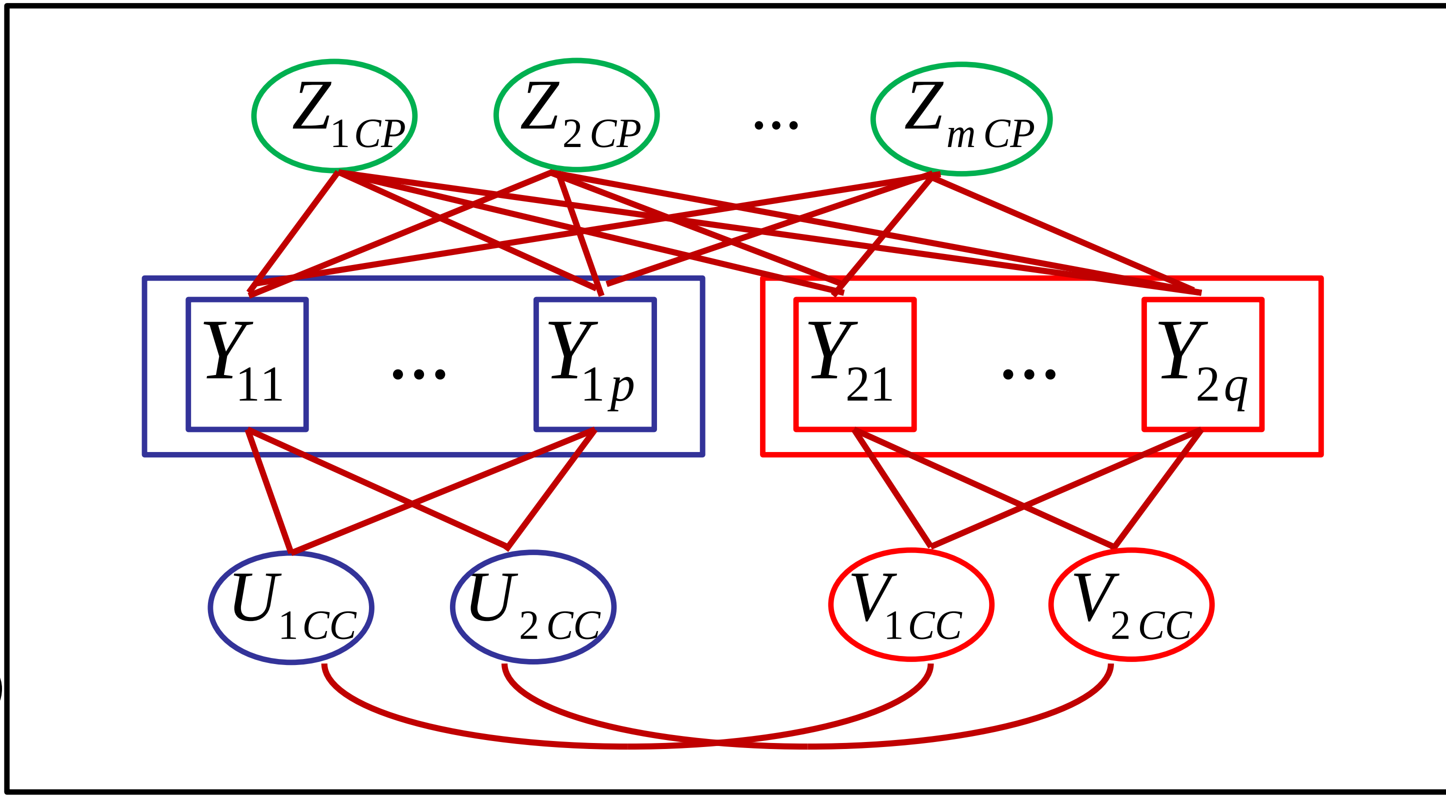
$$U_1 = a_1' Y_1 = e_1' \underbrace{\Sigma_{11}^{-1/2}}_{P_1 \Lambda^{-1/2} P_1'} Y_1 = e_1' \underbrace{P_1 \Lambda^{-1/2} \underbrace{P_1' Y_1}_{\text{CP de } Y_1}}_{\text{CP padronizado de } Y_1 \text{ (Fator Comum)}}$$

Decomposição spectral de Σ_{11}

Redução de Dimensionalidade

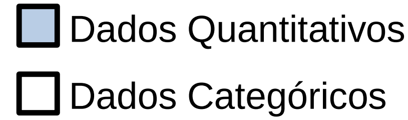
Diferentes alternativas de análise para obter os vetores reducionistas de um conjunto de dados multivariados

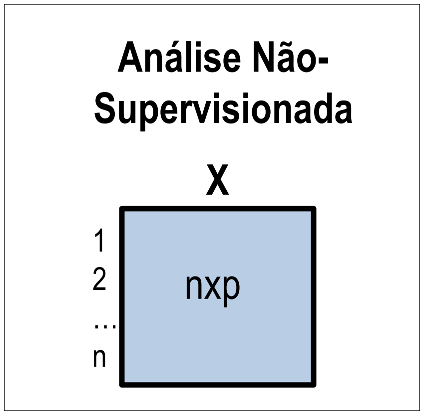
Componente Principal
m≤(n,"p+q")



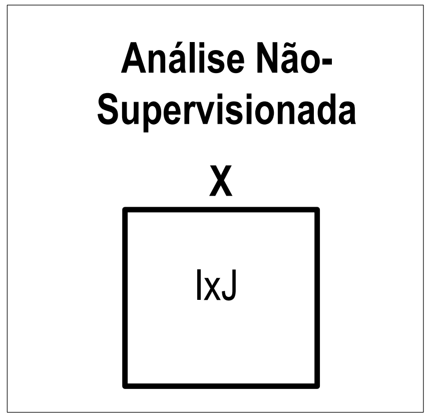
Correlação Canônica
m≤min(n,p,q)

Análises Multivariadas

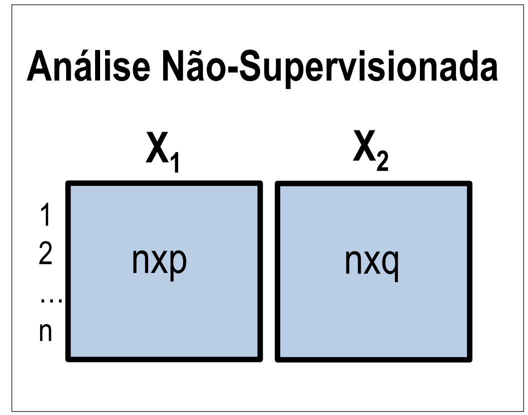




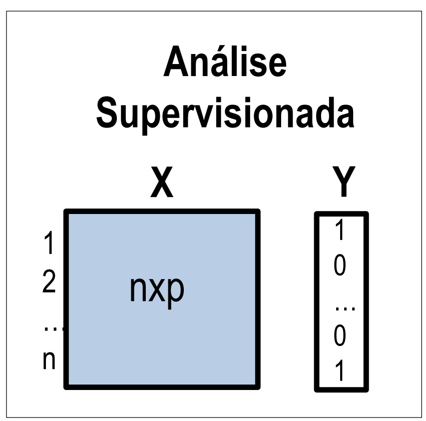
CP, CoP



An Correspond.



ACC



AD

Redução de Dimensionalidade - Apoio do R

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

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

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

- svd(Y): 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: $\lambda_{eigen} = \left(\lambda_{svd} \big/ \sqrt{n-1} \right)^2$

- cmdscale: recebe a matriz de distâncias D e realiza a An. de Escalonamento Multidimensional (An. de Coordenadas Principais) – Soluções não-métricas: "sammon" e "isoMDS"

- ca: realiza a Análise de Correspondência.

- factanal: recebe Y_{nxp} ou R e realiza a Análise Fatorial Exploratória, solução por MVS.
  $\rightarrow$ fa (AF Exploratória, library(psych)) $\rightarrow$ cfa (AF Confirmatória, library(lavaan))

- lda: recebe as (p+1)-variáveis e realiza a Análise Discriminante (solução geral)
  $\rightarrow$ linda (solução linear de Fisher, library(DiscriMiner))

- cancor(Y_1,Y_2): realiza a Análise de Correlação Canônica
  $\rightarrow$ cc (library(CCA))

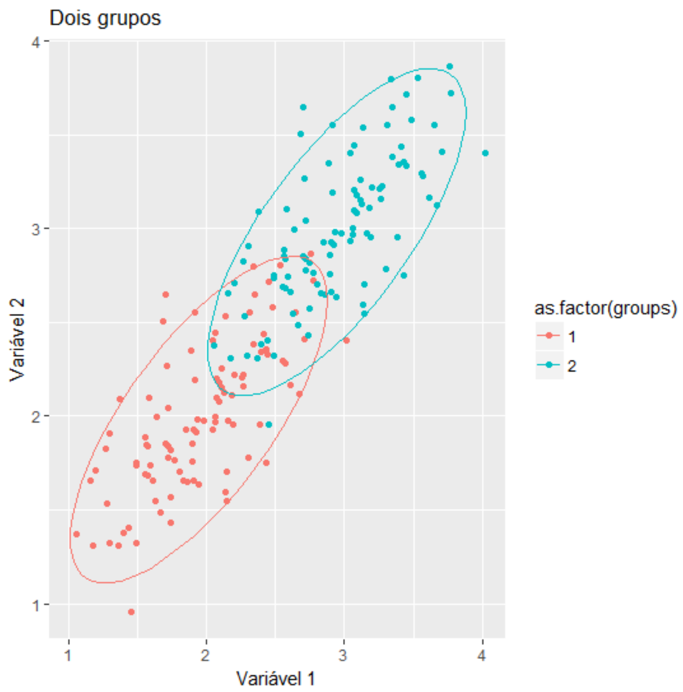
Onde estão os Vetores Reducionistas?

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%).

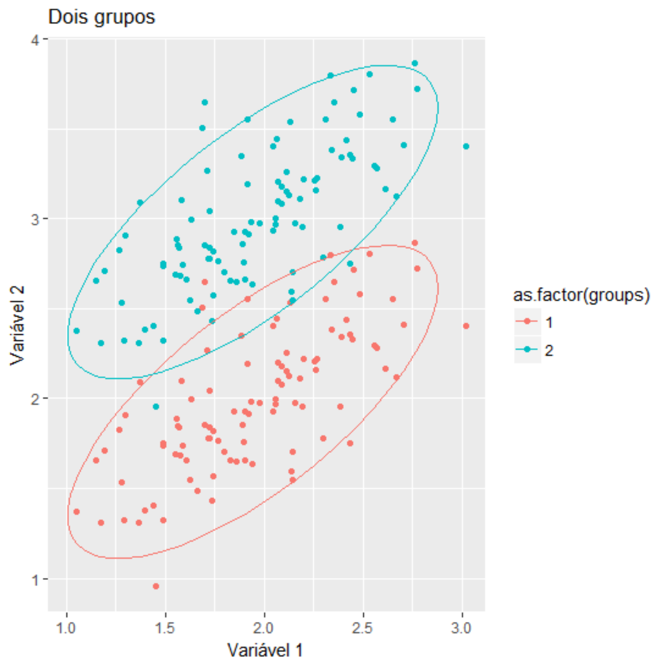
Exemplo 1



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

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

Exemplo 2



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

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

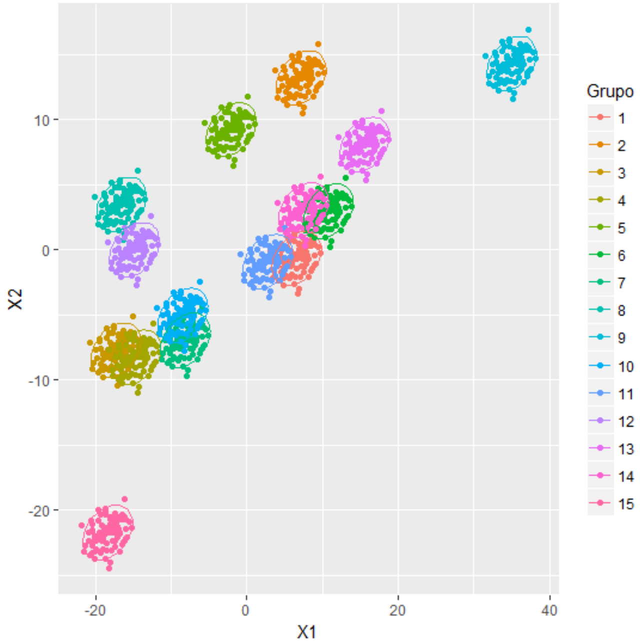
Onde estão os Vetores Reducionistas?

Obter a direção do Eixo Discriminante.

Observações independentes ENTRE e DENTRO de grupos.

Exemplo 3: "Sinais Iguais"

$$T = B + W$$

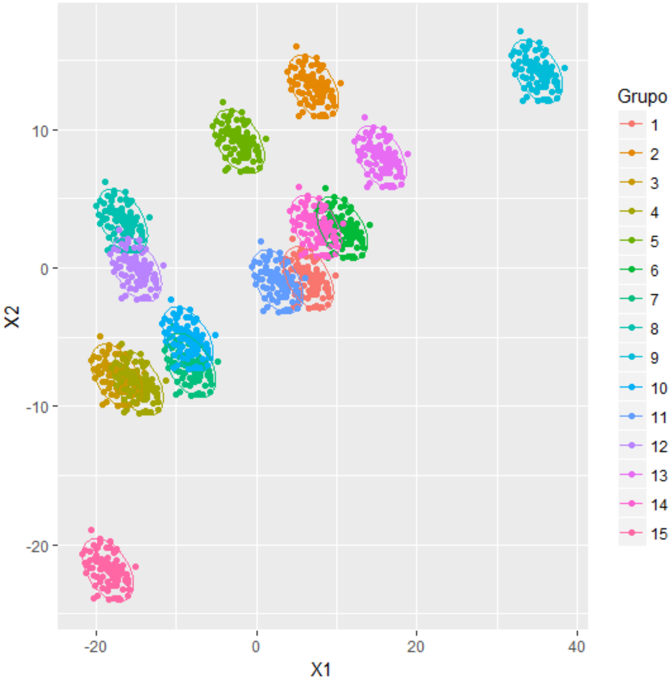


$$G = 15, n_g = 100, \mu_{.} = (0,0)$$

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

Exemplo 4: "Sinais Opostos"

$$T = B + W$$



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

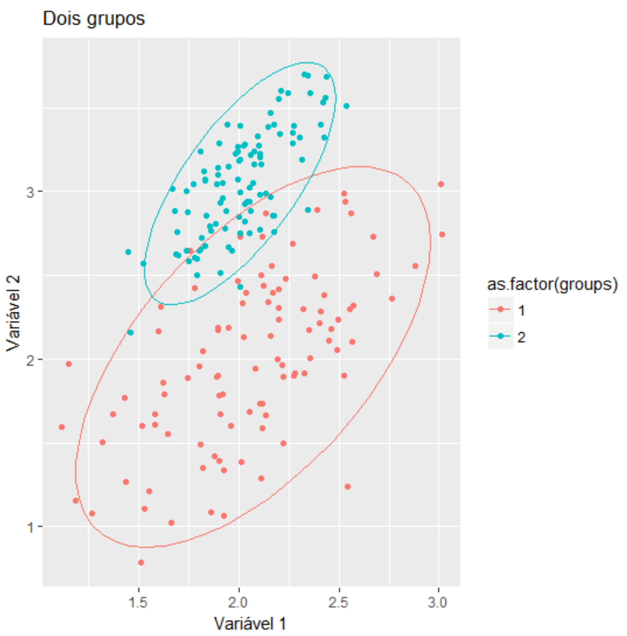
$$S_b = \begin{pmatrix} 150 & 100 \\ 100 & 150 \end{pmatrix}, S_w = \begin{pmatrix} 2 & -0.5 \\ -0.5 & 1 \end{pmatrix}$$

Onde estão os Vetores Reducionistas?

Obter a direção da Variável Canônica.

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

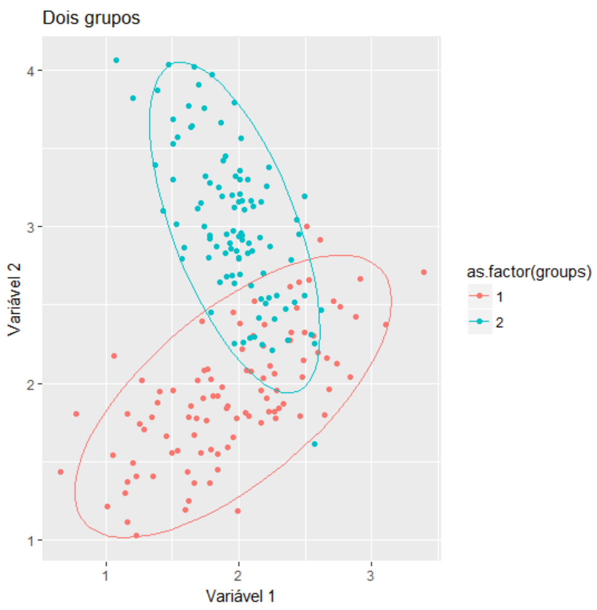
Exemplo 5: Correlações de mesmo sinal



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

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

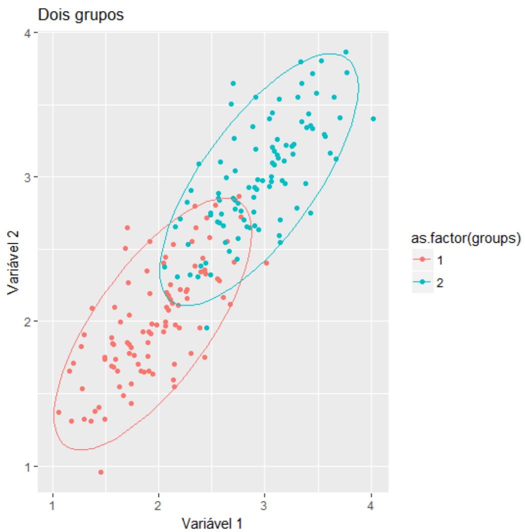
Exemplo 6: Correlações de sinal diferentes



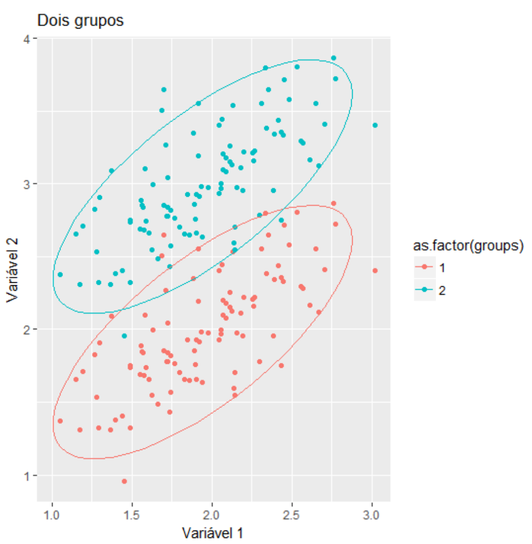
$$R = \begin{pmatrix} 1 & 0.7 & 0.2 & 0.5 \\ 0.7 & 1 & 0.3 & 0.4 \\ 0.2 & 0.3 & 1 & -0.7 \\ 0.5 & 0.4 & -0.7 & 1 \end{pmatrix}$$

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

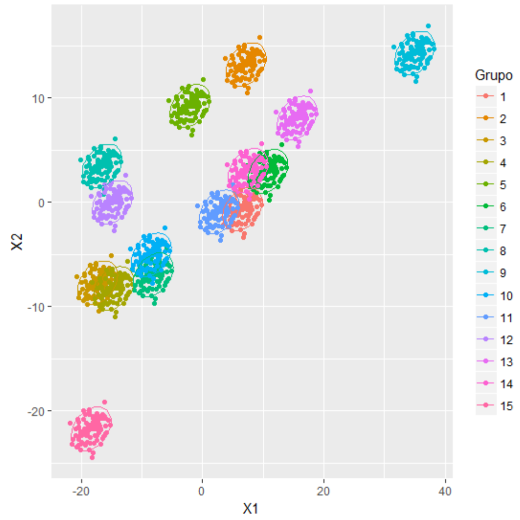
Exemplo 1



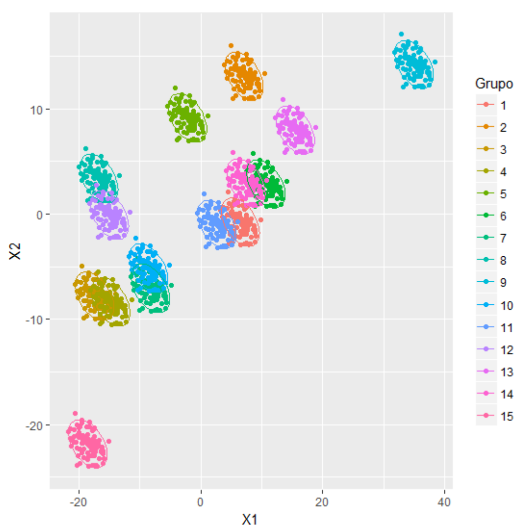
Exemplo 2



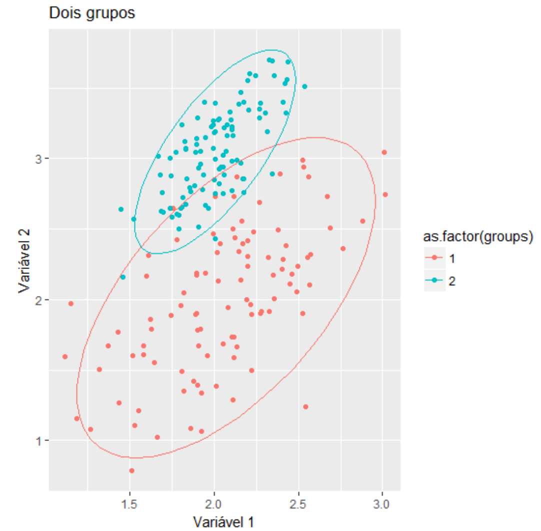
Exemplo 3



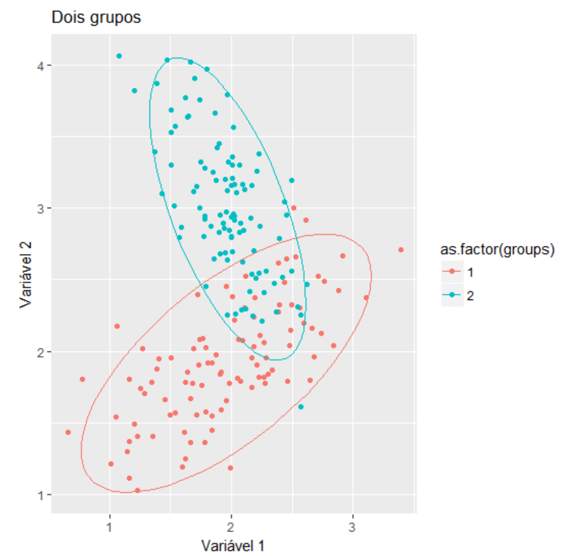
Exemplo 4



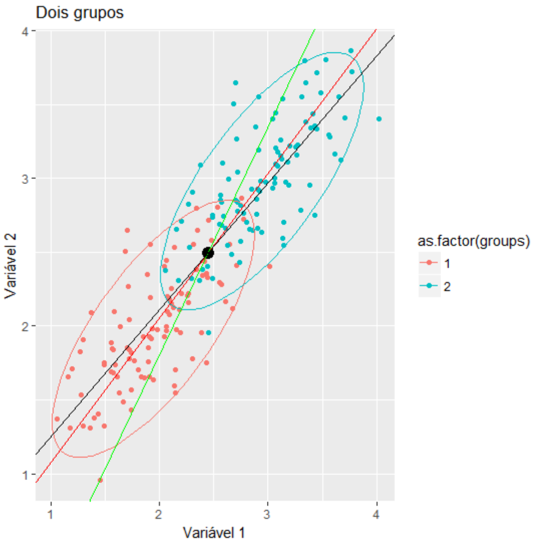
Exemplo 5



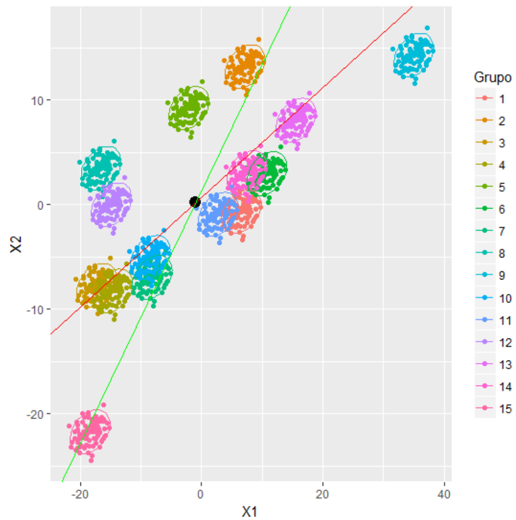
Exemplo 6



Exemplo 1

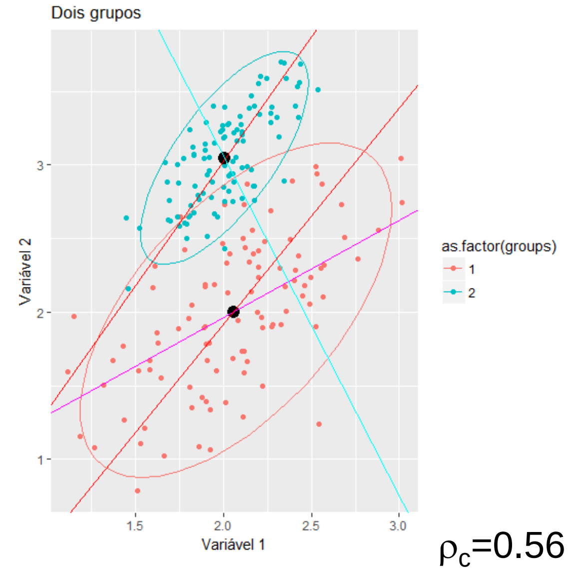


Exemplo 3



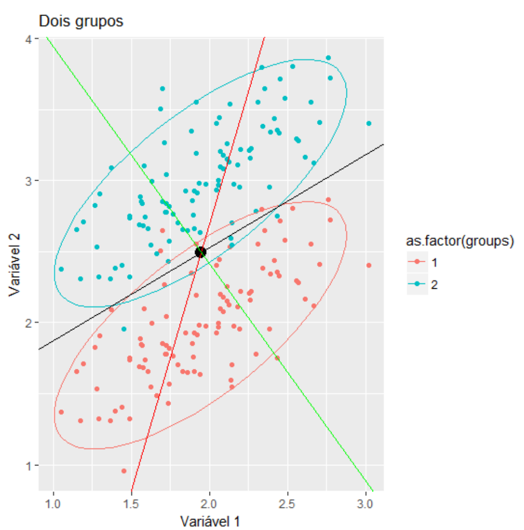
l=(-0.47 -0.56)'%Expl=0.78

Exemplo 5

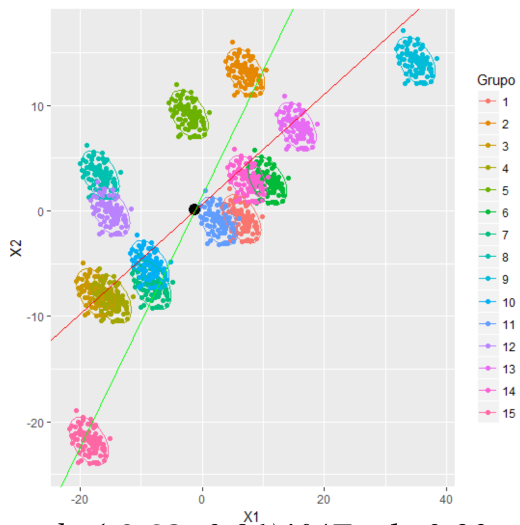


ρ_c=0.56

Exemplo 2

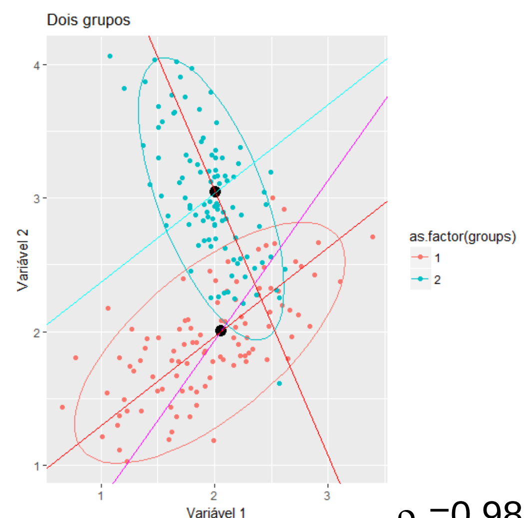


Exemplo 4



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

Análises em Espaços Duais

$$Y_{n\times p} = \left(Y_{ij}\right) \in \Re^{n\times p}; \quad \boxed{Y_{n\times p} = X_{n\times q}\,\beta_{q\times p} + E_{n\times p}}$$

$Y_1 \quad Y_2 \ldots \quad Y_p$	$X_1 \quad X_2 \ldots \quad X_q$
$Y_{n\times p}$	$X_{n\times q}$

Y: Matriz de Dados

X: Matriz de Planejamento, Regressores ou Covariáveis

$$\hat{Y}_{n\times p} = X\,\hat{\beta} = X\left(X'X\right)^{-1}X'Y = PY \Rightarrow S_{X\ p\times p} = Y'\left[P - \frac{1_n 1_n'}{n}\right]Y$$

$$\hat{e}_{n\times p} = Y - \hat{Y} = \left(I_n - P\right)Y \Rightarrow S_{E\ p\times p} = Y'\left[I_n - P\right]Y$$

$$S_X + S_E = S_{p\times p}$$

$$S_{p\times p} = Y'\,H\,Y$$

$$H = \left(I_n - \frac{1_n 1_n'}{n}\right)$$

Matriz de Gower (1966, 1999)

$$D_{n\times n} = \left(d_{ik}\right) \rightarrow A_{n\times n} = \left(-\frac{1}{2}d_{ik}^2\right) \rightarrow \boxed{G_{n\times n} = \left(I_n - \frac{1_n 1_n'}{n}\right)A\left(I_n - \frac{1_n 1_n'}{n}\right) = HYY'H}$$

Análises em Espaços Duais

$$Y_{n\times p};\quad Y_{n\times p} = X_{n\times q}\,\beta_{q\times p} + E_{n\times p};\quad \hat{Y} = X\hat{\beta} = X\left(X'X\right)^{-1}X'Y = PY$$

		ASCA Partição de Y	MANOVA Partição de S	AOD Partição de D
FV	#g.l.	$\mathfrak{R}^{nxp}$	$\mathfrak{R}^{pxp}$	$\mathfrak{R}^{nxn}$
"X"	q-*1*	$\hat{Y}_{n\times p}$	$S_X = Y'\left[P - \frac{1_n1_n'}{n}\right]Y$	$\left[P - \frac{1_n1_n'}{n}\right]A\left[P - \frac{1_n1_n'}{n}\right] = PGP$
Res	*n-q*	$Y_{n\times p} - \hat{Y}_{n\times p}$	$S_E = Y'\left[I_n - P\right]Y$	$\left[I_n - P\right]A\left[I_n - P\right] = \left[I_n - P\right]G\left[I_n - P\right]$
Total	*n-1*	$Y_{n\times p}$	$S = Y'HY = Y'\left[I_n - \frac{1_n1_n'}{n}\right]Y$ CP	$HYY'H = \left[I_n - \frac{1_n1_n'}{n}\right]A\left[I_n - \frac{1_n1_n'}{n}\right] = G$ CoP

Estatísticas de Teste de

$$H_0 : \notin Efeito\ de\ X$$

$$F_{Wilks} = \frac{\left|S_E\right|}{\left|S_X + S_E\right|}$$

$$F_{Pillai} = tr\left(\frac{S_X/(q-1)}{S_E/(n-q)}\right)$$

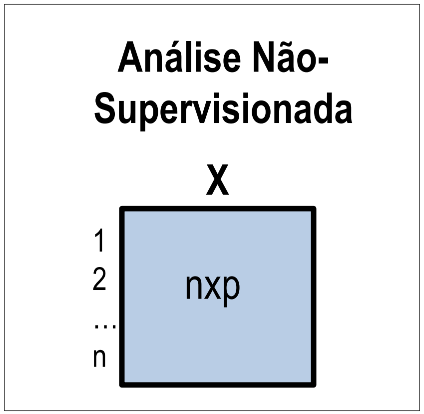
$$T_{Mantel} = \frac{tr\left(PGP\right)/\left(q-1\right)}{tr\left[\left(I_n - P\right)G\left(I_n - P\right)\right]/\left(n-q\right)}$$

Significância avaliada via testes de permutação.

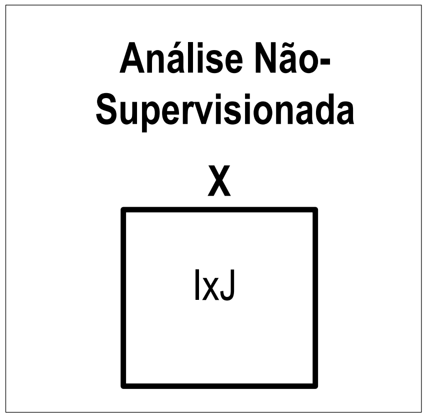
Análises Multivariadas

Dados Quantitativos

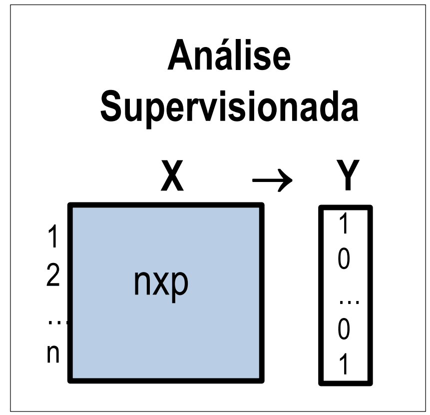
Dados Categóricos



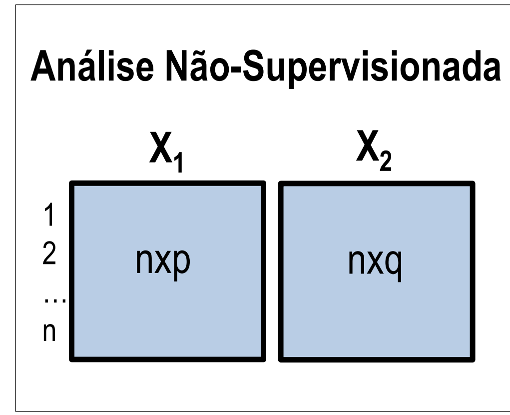
CP, CoP



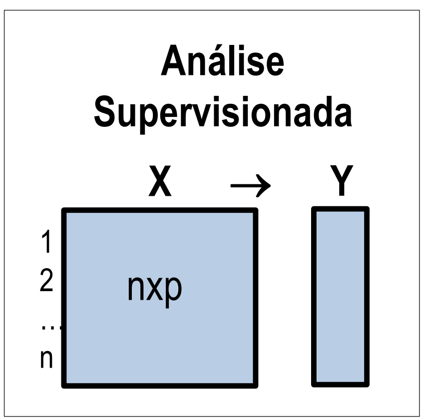
An Correspond.



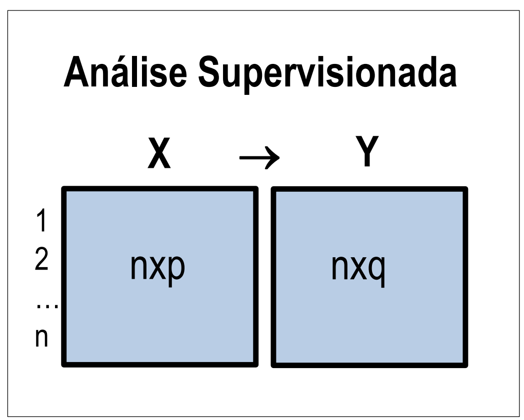
AD



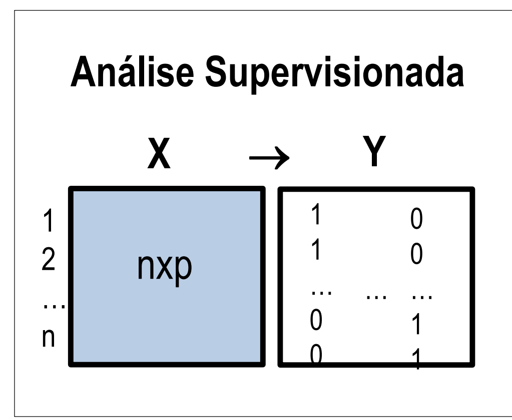
ACC



PLS (Partial Least Square)



PLS para múltiplas respostas



ACC_AD