MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Sem/2022 - 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)

Decomposições: Y_{nxp} , $SS_{T\ pxp}$,e D_{nxn}

- $N_p(\mu_g;\Sigma_g)$: Inferências sobre μ_g (T^2, MANOVA, ICS, Correções para Múltiplos testes

Técnicas Multivariadas:

✓ Análise de Componentes Principais

- Escalonamento Multidimensional
- Análise de Correspondência

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

Revisando ☺ Componentes Principais

Dados dos cães

	Y1	Y2	Y3	Y4	Y5	Y6
[1,]	9.7	21.0	19.4	7.7	32.0	36.5
[2,]	8.1	16.7	18.3	7.0	30.3	32.9
[3,]	13.5	27.3	26.8	10.6	41.9	48.1
[4,]	11.5	24.3	24.5	9.3	40.0	44.6
[5,]	10.7	23.5	21.4	8.5	28.8	37.6
[6,]	9.6	22.6	21.1	8.3	34.4	43.1
[7,]	10.3	22.1	19.1	8.1	32.2	35.0

$$Y_{n\times p} \rightarrow Z_{n\times p} = YV; \quad Z_{ik} = V_k{}'Y_i$$

$$\Sigma_{p\times p} = V\Lambda V' \rightarrow Cov(Z) = \Lambda = Diag(\lambda_j)$$

$$tr(\Sigma) = tr(\Lambda) = \sum_{j=1}^{p} \lambda_j$$

Matriz de Covariância S:

	Y1	Y2	Y3	Y4	Y5	Y6
Y1	**2.88**	5.25	4.85	1.93	6.53	7.74
Y2	5.25	**10.56**	8.90	3.59	11.46	15.58
Y3	4.85	8.90	**9.61**	3.51	13.43	16.31
Y4	1.93	3.59	3.51	**1.36**	4.86	5.92
Y5	6.53	11.46	13.43	4.86	**24.36**	24.68
Y6	7.74	15.58	16.31	5.92	24.68	**31.52**

Autovalores de S (λ_j):

72.61 4.86 2.13 0.66 0.02 0.01

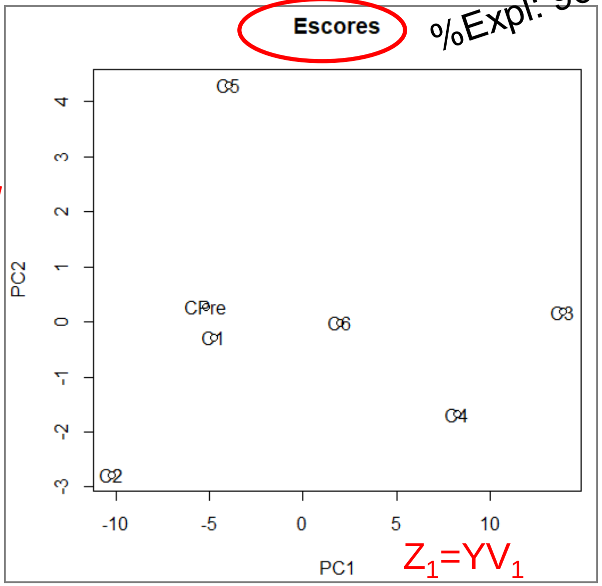
Autovetores de S (V_j):

	V1	V2	V3	V4	V5	V6
Y1	0.18	0.23	-0.41	0.10	0.65	-0.56
Y2	0.34	0.64	-0.33	-0.47	-0.37	0.09
Y3	0.35	0.15	-0.15	0.84	-0.36	0.01
Y4	0.13	0.11	-0.15	0.11	0.52	0.82
Y5	0.55	-0.70	-0.39	-0.21	-0.09	0.03
Y6	0.65	0.10	0.72	-0.08	0.18	-0.09

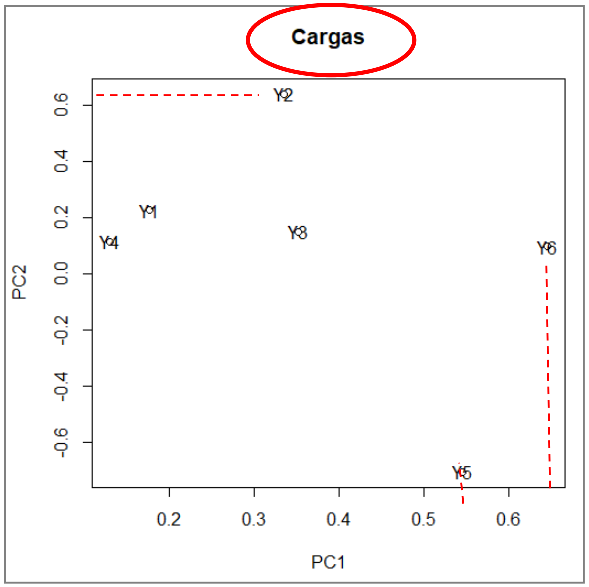
Cargas

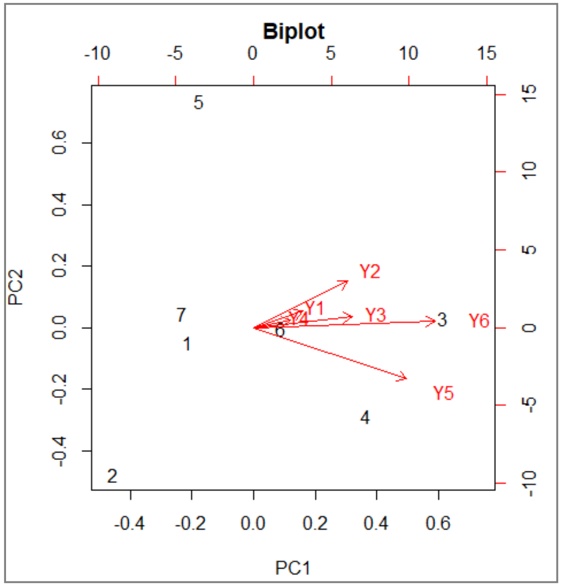
%Expl: 96%

$Z_2=YV_2$



$Z_1=YV_1$





Escalonamento Multidimensional ou

Análise de Coordenadas Principais

Dados Multivariados

Y_{nxp}: não é conhecida
Somente D (ou C) é conhecida

$$D = \begin{pmatrix} 0 & & & \\ d_{21} & 0 & & \\ \dots & \dots & \dots & \\ d_{n1} & d_{n2} & \dots & 0 \end{pmatrix}_{n \times n} \qquad C = \begin{pmatrix} 1 & & & \\ r_{21} & 1 & & \\ \dots & \dots & \dots & \\ r_{n1} & r_{n2} & \dots & 1 \end{pmatrix}_{n \times n}$$

D: Matriz de Distâncias entre indivíduos

C: Matriz de Similaridades entre indivíduos

<u>Objetivos:</u>

- A partir de matrizes de distância D (ou similaridade C) entre *n* objetos obter uma representação das correspondentes observações Y_{nxp} que geraram D (ou C);
- A partir de D (ou C) obter Eixos Principais (Coordenadas Principais) $\Rightarrow$ Identificar a dimensão das observações multivariadas que geraram D

Escalonamento Multidimensional

Análise baseada no espaço linha da matriz de dados

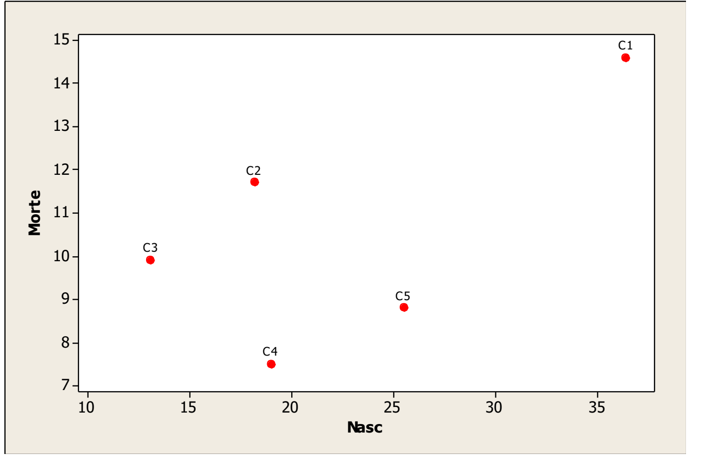
Escalonamento Multidimensional

Motivação: Representação gráfica de observações

$Y_{5\times 2}$: Matriz conhecida

Cidade	Nascimento	Morte
C1	36,4	14,6
C2	18,2	11,7
C3	13,1	9,9
C4	19	7,5
C5	25,5	8,8

Representação das cidades



$$d_{12}; \quad d_{12}^2 = (36{,}4 - 18{,}2)^2 + (14{,}6 - 11{,}7)^2$$

	C1	C2	C3	C4	C5
C1	0				
C2	18,43	0			
C3	23,76	5,41	0		
C4	18,79	4,28	6,37	0	
C5	12,34	7,85	12,45	6,63	0

Suponha que o único dado disponível corresponde à Matriz de Distância Euclidiana D entre as observações.

Como representar as observações (obter Y)?

Escalonamento Multidimensional

Matriz de Distância Euclidiana entre os 7 cães (p=6 variáveis)

Matriz de Distância Euclidiana

	1	2	3	4	5	6	7
1	0						
2	6.21	0					
3	18.70	24.34	0				
4	13.13	18.55	5.99	0			
5	4.83	9.44	18.38	13.64	0		
6	7.43	12.94	12.50	7.26	7.98	0	
7	2.03	6.62	19.20	13.78	5.09	8.67	0

Como representar os 7 pontos em um gráfico?

$$D_{7\times7} \xrightarrow{?} Y_{7\times m}; \quad m = 2$$

Escalonamento Multidimensional

Matriz de Distância* ("postos") entre 6 Docerias

	A	B	C	D	E	F
A	-					
B	2	-				
C	13	12	-			
D	4	6	9	-		
E	3	5	10	1	-	
F	8	7	11	14	15	-

D não é matriz de distância Euclidiana mas uma distância empírica!

*1: é o par mais similar 15: é o par menos similar (6(6-1)/2)

Como representar os 6 pontos em um gráfico?

$$D_{6\times 6} \xrightarrow{?} Y_{6\times m}; \quad m = 2$$

Escalonamento Multidimensional

Distâncias (em km) entre 12 cidades $\Rightarrow$ matriz de "distância" empírica

	1	2	3	4	5	6	7	8	9	10	11	12
1	0											
2	244	0										
3	218	350	0									
4	284	77	369	0								
5	197	167	347	242	0							
6	312	444	94	463	441	0						
7	215	221	150	236	279	245	0					
8	469	583	251	598	598	169	380	0				
9	166	242	116	257	269	210	55	349	0			
10	212	53	298	72	170	392	168	531	190	0		
11	253	325	57	340	359	143	117	264	91	273	0	
12	270	168	284	164	277	378	143	514	174	111	256	0

D não é matriz de distância Euclidiana mas uma distância empírica!

Como representar os 12 pontos em um gráfico?

$$D_{12\times 12} \xrightarrow{?} Y_{12\times m};\quad m = 2$$

Escalonamento Multidimensional

Notação:

Dada uma matriz de distâncias D,

$$D = \left(d_{i\,j}\right)_{n\times n}$$

O objetivo do Escalonamento Multidimensional é predizer pontos, $\hat{Y}_i \in \Re^m$

i=1,…,n, tal que, se $\hat{d}_{i\,k}$ é a distância Euclidiana entre $\hat{Y}_i$ e $\hat{Y}_k$, então

$\hat{D} = \left(\hat{d}_{ik}\right)$ é uma "boa aproximação" para D .

Solução:

- Métodos métricos $\Rightarrow$ baseados na teoria da decomposição espectral
- Métodos não métricos $\Rightarrow$ baseados na minimização de funções objetivo como o "stress".

Escalonamento Multidimensional

<u>Solução Clássica (Métrica) na dimensão ***m***</u>

(Mardia, 1979)

Dado *D*, matriz de distância Euclidiana $\Leftrightarrow$ Existe $Y_{n\times p}$ matriz de dados tal que:

$$D = \left(d_{ik}\right)_{n\times n} \; ; \quad d_{ik}^2 = \sum_{j=1}^{p}\left(y_{ij} - y_{kj}\right)^2$$

d_{ik} : conhecido

y_{ik} : desconhecido

Logo, existe: $B_{n\times n} = Y_{n\times p}\, Y'_{p\times n} \Longrightarrow b_{ik} = \sum_{j=1}^{p} y_{ij}\, y_{kj}$

Matriz de Produto Interno entre as linhas de Y

$$\begin{cases} d_{ik}^2 = b_{ii} + b_{kk} - 2b_{ik} \\ b_{ik} = -\dfrac{1}{2}\left(d_{ik}^2 - d_{i.}^2 - d_{.k}^2 + d_{..}^2\right) \end{cases}$$

B: calculada de ***D***

Escalonamento Multidimensional

Solução Clássica na dimensão ***m***

$$D=\left(d_{ik}\right)_{n\times n} \quad \Leftrightarrow \quad Y_{n\times p}\ ? \qquad d_{ik}^2=\sum_{j=1}^{p}\left(y_{ij}-y_{kj}\right)^2$$

$$B=YY' \quad \Rightarrow \quad B=\left(b_{ik}=-\frac{1}{2}\left(d_{ik}^2-d_{i.}^2-d_{.k}^2+d_{..}^2\right)\right)$$

Logo, temos:

$B_{n\times n}=Y\,Y'$ Matriz p.s.d. (sob n>p) e sua Decomposição Espectral é:

$$=U\,\Lambda\,U'=U\Lambda^{1/2}\,\Lambda^{1/2}U'=\left(U\Lambda^{1/2}\right)\left(U\Lambda^{1/2}\right)' \qquad \boxed{\Rightarrow \hat{Y}=\left(U\Lambda^{1/2}\right)}$$

- Quando n>p, o posto de D é p. Logo, há (n-p) autovalores nulos.
- Podemos escolher uma representação para Y em uma dimensão m (m<p) .

$$\hat{Y}=U\Lambda^{1/2}=\begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ u_{21} & u_{22} & \dots & u_{2n} \\ & & \dots & \\ u_{n1} & u_{n1} & \dots & u_{nn} \end{pmatrix}\begin{pmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 \\ & & \dots & \\ 0 & 0 & \dots & \sqrt{\lambda_n} \end{pmatrix}$$

$$n>p\Rightarrow \lambda_{(p+1)}=\dots=\lambda_n=0$$

Escalonamento Multidimensional

Solução Clássica na dimensão ***m***

$$D = \left(d_{ik}\right)_{n\times n} \quad \left\{ \begin{array}{l} \Leftrightarrow \qquad Y_{n\times p} \; ? \\ B_{n\times n} = \left(b_{ik} = -\dfrac{1}{2}\left(d_{ik}^2 - d_{i.}^2 - d_{.k}^2 + d_{..}^2\right) \right) \end{array} \right.$$

Matriz de distâncias

Obter os "*m*" primeiros componentes da decomposição espectral de ***B***:

Autovalores: $\boxed{\lambda_1 > \lambda_2 > \ldots > \lambda_m} > \lambda_{m+1} > \ldots > \lambda_n > 0$

Autovetores normalizados: $U = \left(\boxed{U_1, U_2, \ldots, U_m}, U_{m+1}, \ldots, U_p, U_{p+1}, \ldots, U_n\right)$

$\Rightarrow$ As coordenadas do vetor $\boldsymbol{Y_i}$ são obtidas a partir da *i*-ésima linha da matriz $U=(u_{ij})$

$$\hat{Y}_i = \left(U_1 \ldots U_m\right)_i \, \Lambda_m^{1/2} = \left(u_{i1}\sqrt{\lambda_1}, u_{i2}\sqrt{\lambda_2}, \ldots, u_{im}\sqrt{\lambda_m}\right)$$

Escalonamento Multidimensional

Matriz de Distância Euclidiana entre os 7 cães (p=6)

Matriz de distância Euclidiana entre as observações

	1	2	3	4	5	6	7
1	0.00	6.21	18.70	13.13	4.83	7.43	2.03
2	6.21	0.00	24.34	18.55	9.44	12.94	6.62
3	18.70	24.34	0.00	5.99	18.38	12.50	19.20
4	13.13	18.55	5.99	0.00	13.64	7.26	13.78
5	4.83	9.44	18.38	13.64	0.00	7.98	5.09
6	7.43	12.94	12.50	7.26	7.98	0.00	8.67
7	2.03	6.62	19.20	13.78	5.09	8.67	0.00

Matriz B

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]
[1,]	23.09	48.94	-66.33	-38.97	17.32	-9.68	25.62
[2,]	48.94	113.39	-142.41	-79.68	29.54	-20.68	50.89
[3,]	-66.33	-142.41	194.13	114.78	-54.35	25.24	-71.07
[4,]	-38.97	-79.68	114.78	71.28	-39.90	15.65	-43.17
[5,]	17.32	29.54	-54.35	-39.90	34.88	-8.09	20.62
[6,]	-9.68	-20.68	25.24	15.65	-8.09	12.69	-15.14
[7,]	25.62	50.89	-71.07	-43.17	20.62	-15.14	32.25

Escalonamento Multidimensional

Dados dos Cães (n=7; p=6)

Matriz B

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]
[1,]	23.09	48.94	-66.33	-38.97	17.32	-9.68	25.62
[2,]	48.94	113.39	-142.41	-79.68	29.54	-20.68	50.89
[3,]	-66.33	-142.41	194.13	114.78	-54.35	25.24	-71.07
[4,]	-38.97	-79.68	114.78	71.28	-39.90	15.65	-43.17
[5,]	17.32	29.54	-54.35	-39.90	34.88	-8.09	20.62
[6,]	-9.68	-20.68	25.24	15.65	-8.09	12.69	-15.14
[7,]	25.62	50.89	-71.07	-43.17	20.62	-15.14	32.25

Decomposição Espectral de B

Autovalores

435.64 29.17 12.78 3.95 0.15 0.03 0.00

Dimensão: m=2 explica 96,49%

$$\boxed{\Rightarrow \hat{Y} = \left(U\Lambda^{1/2}\right)}$$

Autovetores

	U1	U2	U3	U4	U5	U6	U7
[1,]	-0.23	-0.05	0.05	0.25	-0.02	0.86	-0.38
[2,]	-0.49	-0.52	-0.06	-0.52	-0.22	-0.18	-0.38
[3,]	0.67	0.03	0.23	-0.19	-0.57	0.04	-0.38
[4,]	0.40	-0.31	0.07	-0.11	0.77	-0.03	-0.38
[5,]	-0.19	0.80	-0.05	-0.39	0.18	-0.05	-0.38
[6,]	0.10	0.00	-0.79	0.40	-0.11	-0.21	-0.38
[7,]	-0.25	0.05	0.55	0.55	-0.03	-0.43	-0.38

Coordenadas Principais:

	$\hat{Y}_1$	$\hat{Y}_2$
1	-4.76	-0.28
2	-10.23	-2.78
3	13.90	0.18
4	8.26	-1.68
5	-3.98	4.29
6	2.01	0.00
7	-5.21	0.26

$\sqrt{435.64}\,(-0.23)$

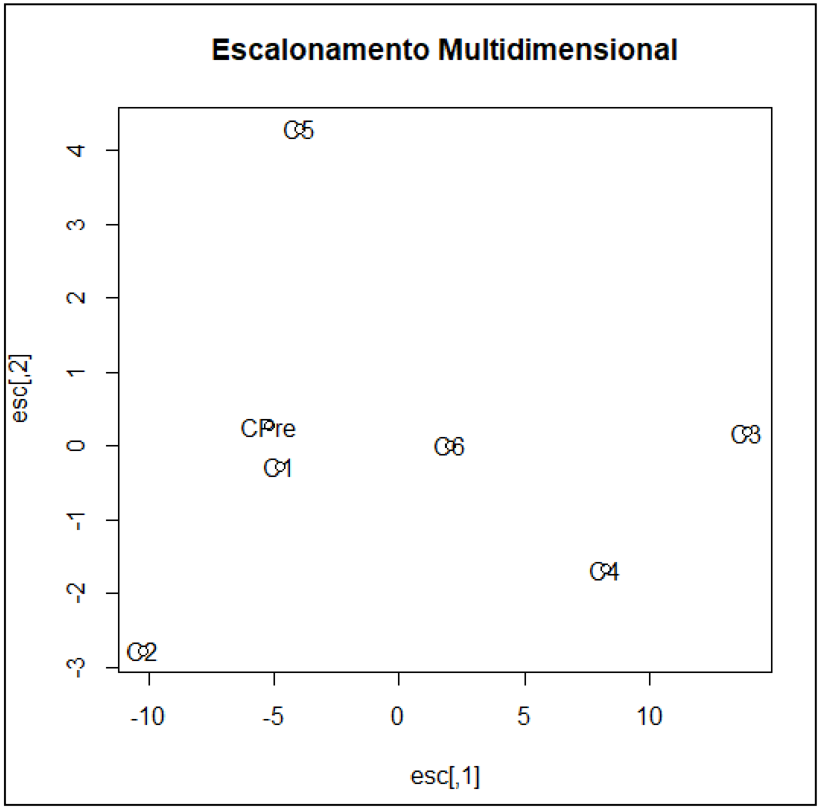
$\sqrt{29.17}\,(-0.05)$

Escalonamento Multidimensional

Representação dos 7 cães obtida da matriz de distância Euclidiana (D)

Coordenadas Principais:

	$\hat{Y}_1$	$\hat{Y}_2$
1	-4.76	-0.28
2	-10.23	-2.78
3	13.90	0.18
4	8.26	-1.68
5	-3.98	4.29
6	2.01	0.00
7	-5.21	0.26



Escalonamento Multidimensional

Critério de otimalidade da solução Métrica: $\hat{D} \cong D$

Matrizes de Distância Euclidiana (observada e predita) para os dados dos cães (n=7).

Triangular superior: Matriz de distância Euclidiana observada (D)
Triangular inferior: Matriz de distância Euclidiana predita das Coordenadas Principais ($\hat{D}$)

	1	2	3	4	5	6	7
1	0	6.21	18.70	13.13	4.83	7.43	2.03
2	6.01	0	24.34	18.55	9.44	12.94	6.62
3	18.67	24.31	0	5.99	18.38	12.50	19.20
4	13.10	18.52	5.94	0	13.64	7.26	13.78
5	4.64	9.44	18.35	13.62	0	7.98	5.09
6	6.78	12.55	11.89	6.48	7.36	0	8.67
7	0.70	5.87	19.11	13.61	4.21	7.22	0

$= D$

$= \hat{D}$

Usar os desvios para avaliar a qualidade da representação das observações em $\Re^2$

Se D é matriz de distância Euclidiana a solução métrica é ótima no sentido de minimizar a SQ dos desvios.

Análise de Componentes Principais e Coordenadas Principais

Resultado importante!

Análise de CP
Decomposição espectral de Σ

$$Y_{n\times p};\ \boxed{\Sigma_{p\times p} = V_{p\times p}\Lambda_p V'_{p\times p}} \Rightarrow \boxed{Z_{n\times p} = YV}$$

Decomposição de Y em valores singulares

$$\boxed{Y = U_{n\times n}\Lambda_n^{1/2}V'_{p\times p};}\quad n \geq p \Rightarrow \lambda_{p+1} = \ldots = \lambda_n = 0$$

$$\boxed{YV} = U\ \Lambda^{1/2}V'\,V = \boxed{U\Lambda^{1/2}}$$

Componentes principais | Coordenadas principais

Análise de EM
Decomposição espectral de B

$$Y_{n\times p}; D_{n\times n} \rightarrow \boxed{B_{n\times n} = U_{n\times n}\Lambda_n U'_{n\times n}} \Rightarrow \boxed{Y_{n\times p} = U_{n\times p}\Lambda^{1/2}}$$

- Os ***m*** primeiros Componentes Principais são "ótimos" ⇒ a soma de suas variâncias é maior que qualquer outro conjunto de ***m*** combinações lineares não correlacionadas.
- as ***m*** primeiras Coordenadas Principais são "ótimas" ⇒ a projeção de Y no sub-espaço de dimensão ***m*** de $\Re^p$ é mais próxima (em distância Euclidiana) da configuração original do que qualquer outra ($\hat{D} \cong D$)

Componentes Principais – Coordenadas Principais
Solução via Espaços Duais

Resultado importante!

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

Análise no espaço das variáveis: $\Re^{pxp}$

$$Y'Y=\Sigma_{p\times p}=V_{p\times p}\begin{pmatrix}\Lambda_r & 0\\ 0 & 0\end{pmatrix}V'_{p\times p}\quad\Rightarrow\quad \overset{\approx Y}{Y_{n\times p}V_{p\times r}}$$

r Componentes Principais

Análise no espaço dos indivíduos: $\Re^{nxn}$

$$D_{n\times n}\Rightarrow B_{n\times n}=YY'=U_{n\times n}\begin{pmatrix}\Lambda_r & 0\\ 0 & 0\end{pmatrix}U'_{n\times n}\Rightarrow \overset{\approx Y}{U\Lambda_r^{1/2}}$$

Escalonamento Multidimensional:
r Coordenadas Principais
obtidas da Matriz de Distâncias

Análise no espaço $\Re^{nxp}$

$$Y_{n\times p}=U_{n\times n}\begin{pmatrix}\Lambda_r^{1/2} & 0\\ 0 & 0\end{pmatrix}V'_{p\times p}\quad\Rightarrow\quad \boxed{Y_{n\times p}V_{p\times r}=U_{n\times n}\Lambda_r^{1/2}}$$

Equivalência entre os Componentes Principais e as Cordenadas Principais

n<<p : Componentes Principais de Y podem ser obtidos da decomposição espectral da matriz de distâncias D (nxn), de dimensão muito menor que Σ (pxp)

Componentes Principais – Coordenadas Principais

Equivalência das Soluções em Espaços Duais

Resultado importante!

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

HY: linhas de Y centradas

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

B=HYY'H

(n-1)S_u=Y'HY

Análise em $\Re^{nxn}$

Análise em $\Re^{pxp}$

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

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

Coordenadas Principais

Componentes Principais (das linhas de HY)

Coordenadas Principais

Dados Cães: n=7; p=6 variáveis

	X1	X2	X3	X4	X5	X6
[1,]	9.7	21.0	19.4	7.7	32.0	36.5
[2,]	8.1	16.7	18.3	7.0	30.3	32.9
[3,]	13.5	27.3	26.8	10.6	41.9	48.1
[4,]	11.5	24.3	24.5	9.3	40.0	44.6
[5,]	10.7	23.5	21.4	8.5	28.8	37.6
[6,]	9.6	22.6	21.1	8.3	34.4	43.1
[7,]	10.3	22.1	19.1	8.1	32.2	35.0

Matriz de Distância Euclidiana (D)

	1	2	3	4	5	6	7
1	0.00	6.21	18.70	13.13	4.83	7.43	2.03
2	6.21	0.00	24.34	18.55	9.44	12.94	6.62
3	18.70	24.34	0.00	5.99	18.38	12.50	19.20
4	13.13	18.55	5.99	0.00	13.64	7.26	13.78
5	4.83	9.44	18.38	13.64	0.00	7.98	5.09
6	7.43	12.94	12.50	7.26	7.98	0.00	8.67
7	2.03	6.62	19.20	13.78	5.09	8.67	0.00

Escalonamento Multidimensional

Coordenadas Principais

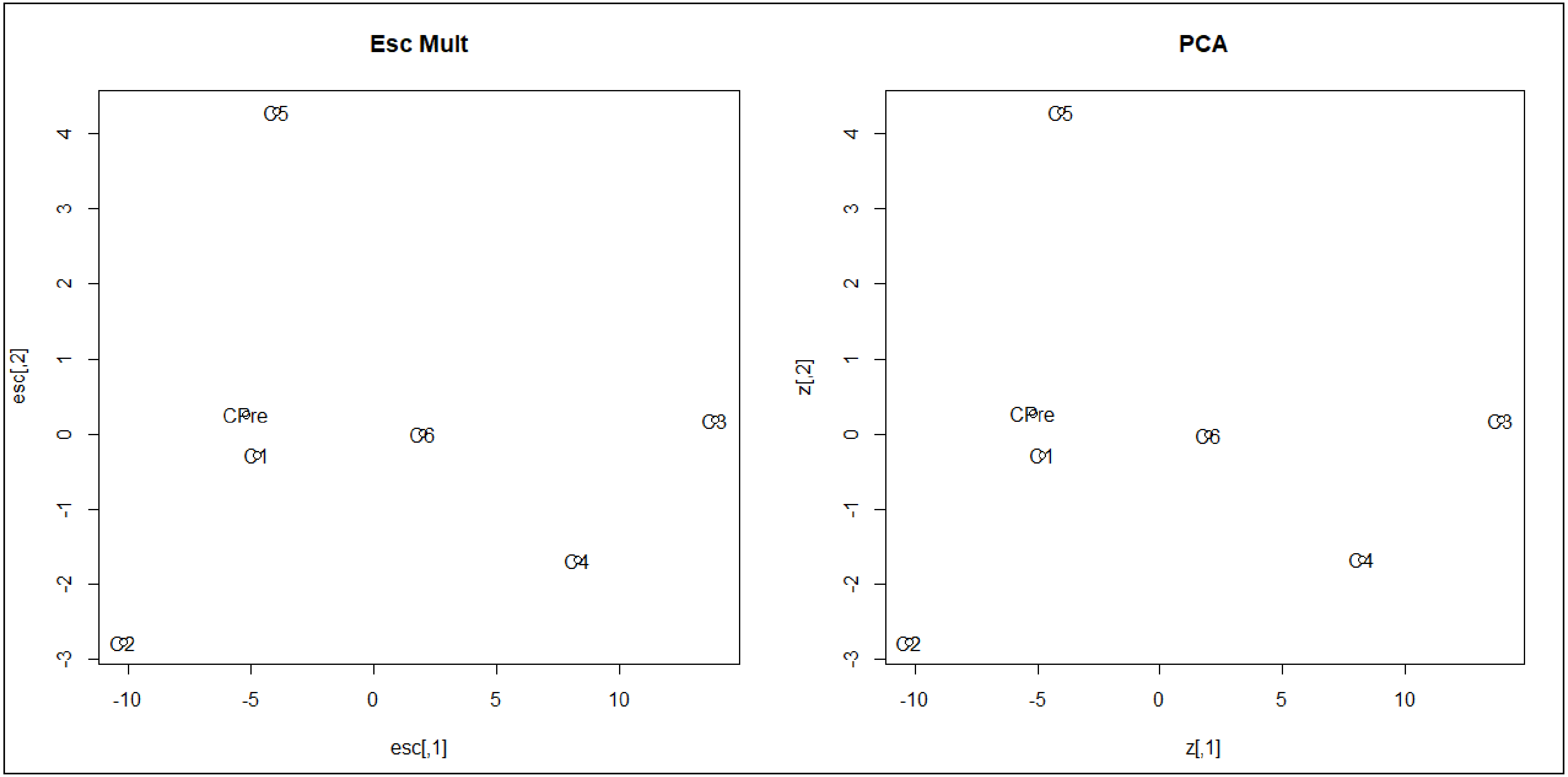
	$\hat{Y}_1$	$\hat{Y}_2$
1	-4.76	-0.28
2	-10.23	-2.78
3	13.90	0.18
4	8.26	-1.68
5	-3.98	4.29
6	2.01	0.00
7	-5.21	0.26

As Coordenadas Principais representam uma escala (dados) construída a partir da informação das distâncias (D)

Sendo D diatância Euclidiana as CoP são os Escores dos Componentes Principais

Componentes Principais e Coordenadas Principais Equivalência

Coordenadas Principais obtidas de D_{nxn} $\Rightarrow$ Representação (em $\Re^2$) equivalente aos Componentes Principais obtidos de S_{pxp}



Escalonamento Multidimensional

A análise de Coordenadas Principais (CoP ou Escalonamento Multidimensional) é baseada em uma matriz de Distâncias (nxn) entre observações enquanto a análise de Componentes Principais (CP) é baseada em uma matriz de covariâncias (pxp) entre variáveis.

Equivalências entre CoP e CP:

1. A análise de Coordenadas Principais da matriz de distâncias Euclidianas é equivalente à análise de Componentes Principais da matriz de covariâncias (S).
2. A análise de Coordenadas Principais da matriz de distâncias de Penrose (ou Pearson) é equivalente à análise de Componentes Principais da matriz de correlação (R).

Na redução de dimensionalidade, a análise de Coordenadas Principais pode ser aplicada de maneira **mais geral**, para diferentes escolhas de matriz de distâncias entre observações (Manhattan, ou até mesmo distâncias empíricas). Neste caso, quando a matriz D não Euclidiana, NÃO está garantida a equivalência entre as duas analyses, CoP e CP.

Escalonamento Multidimensional

Métodos Não-Métricos

- **D** é considerada uma matriz de “dissimilaridade” geral (não precisa ser de distância Euclidiana)
- Os elementos de **D** podem ser ordenados

$$d_{ik}^{(1)} \leq d_{ik}^{(2)} \leq \ldots \leq d_{ik}^{(q)}; \quad q = n\left(n-1\right)/2$$

- Seja $\hat{D}$, tal que os elementos $\hat{d}_{ik}$ estão monotonicamente relacionados aos elementos d_{ik}

$$d_{ik} < d_{rs} \quad \Rightarrow \quad \hat{d}_{ik} \leq \hat{d}_{rs} \quad ; \; \mathrm{i} < k, \; \mathrm{r} < \mathrm{s}$$

- Seja Y_{nxm} uma configuração em $\Re^m$ com distâncias $\hat{d}_{ik}$. Y_{nxm} é ótima no sentido de minimizar seu “stress”, definido como:

$$S^2\left(Y\right) = \frac{\sum_{i<k}\left(d_{ik} - \hat{d}_{ik}\right)^2}{\sum_{i<k}\left(d_{ik} - \bar{d}\right)^2}$$

Medida de stress de Y: mede quanto da variância de d_{ik} NÃO é explicada pelas m coordenadas principais

Escalonamento Multidimensional

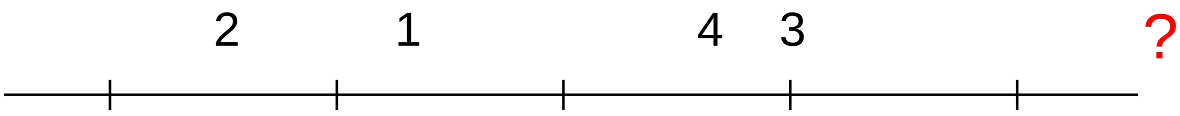
Distância Euclidiana entre os 7 cães (considerando as p=6 variáveis)

Matriz de Distância Euclidiana

	1	2	3	4	5	6	7
1	0						
2	6.21	0					
3	18.70	24.34	0				
4	13.13	18.55	5.99	0			
5	4.83	9.44	18.38	13.64	0		
6	7.43	12.94	12.50	7.26	7.98	0	
7	2.03	6.62	19.20	13.78	5.09	8.67	0

$$d_{43} < d_{21} < d_{41} < d_{42} < d_{31} < d_{32}$$

Tente localizar os cães 1, 2, 3 e 4 em uma única dimensão (m=1):



Solução: obter uma escala (*Y*) que minimize o *stress*

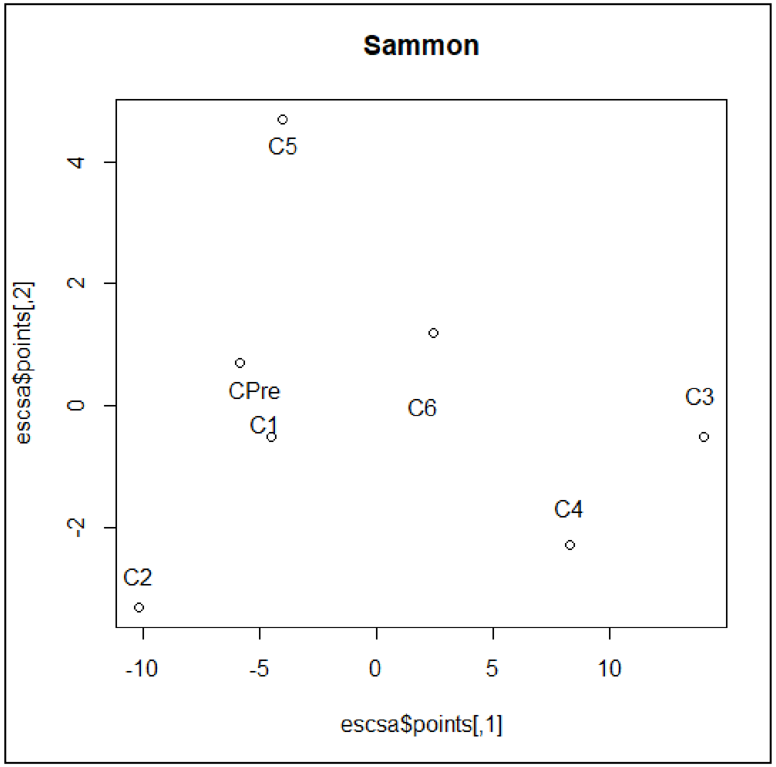
Escalonamento Multidimensional

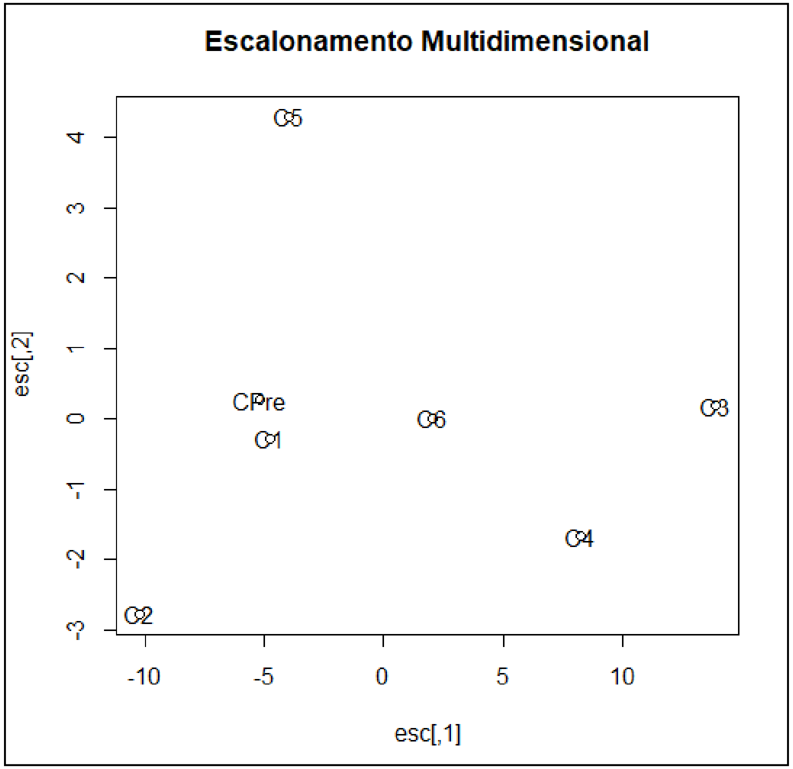
Solução Não Métrica: Sammon

	[,1]	[,2]
1	-4.53	-0.51
2	-10.20	-3.31
3	14.02	-0.51
4	8.27	-2.28
5	-4.07	4.70
6	2.40	1.20
7	-5.90	0.71

Solução Métrica

	[,1]	[,2]
1	-4.76	-0.28
2	-10.23	-2.78
3	13.90	0.18
4	8.26	-1.68
5	-3.98	4.29
6	2.01	0.00
7	-5.21	0.26





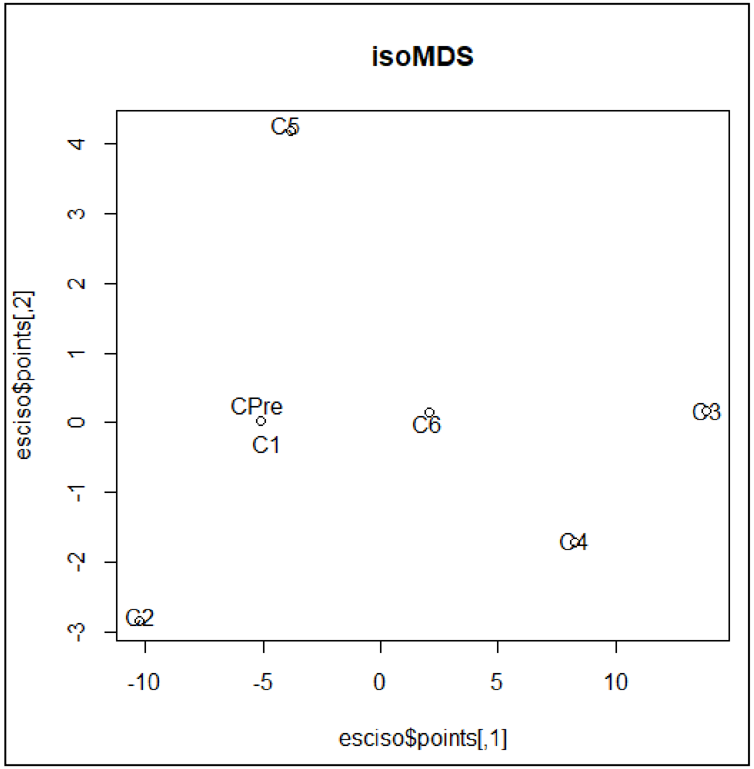
Escalonamento Multidimensional

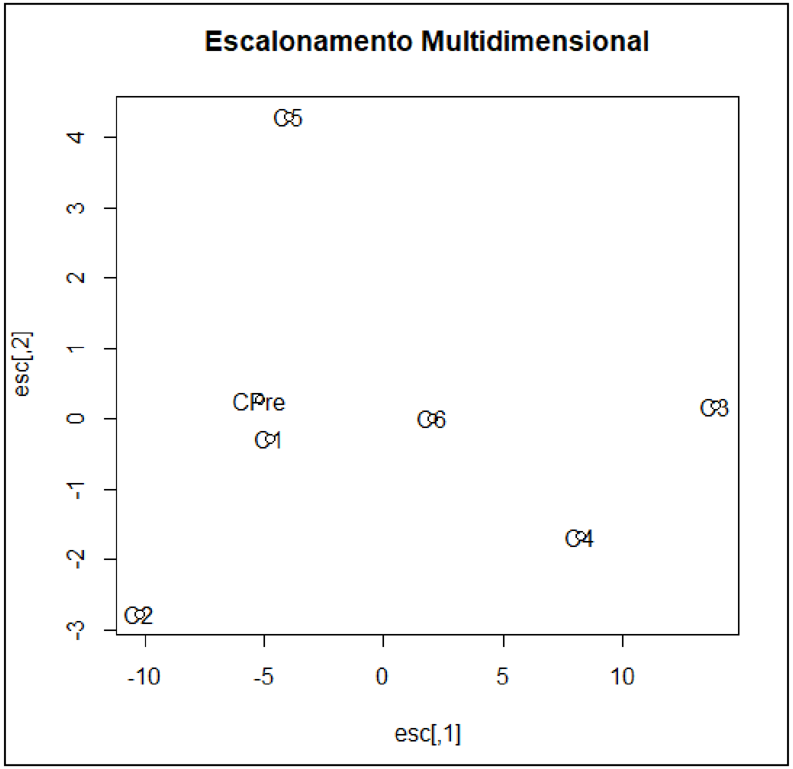
Solução Não Métrica: IsoMDS

	[,1]	[,2]
1	-5.08	0.03
2	-10.26	-2.84
3	13.88	0.17
4	8.30	-1.72
5	-3.82	4.19
6	2.06	0.15
7	-5.08	0.02

Solução Métrica

	[,1]	[,2]
1	-4.76	-0.28
2	-10.23	-2.78
3	13.90	0.18
4	8.26	-1.68
5	-3.98	4.29
6	2.01	0.00
7	-5.21	0.26





Espaços Duais

Dados dos cães: Ynxp

	Y1	Y2	Y3	Y4	Y5	Y6
[1,]	9.7	21.0	19.4	7.7	32.0	36.5
[2,]	8.1	16.7	18.3	7.0	30.3	32.9
[3,]	13.5	27.3	26.8	10.6	41.9	48.1
[4,]	11.5	24.3	24.5	9.3	40.0	44.6
[5,]	10.7	23.5	21.4	8.5	28.8	37.6
[6,]	9.6	22.6	21.1	8.3	34.4	43.1
[7,]	10.3	22.1	19.1	8.1	32.2	35.0

→ CP

Matriz de Covariância: Spxp

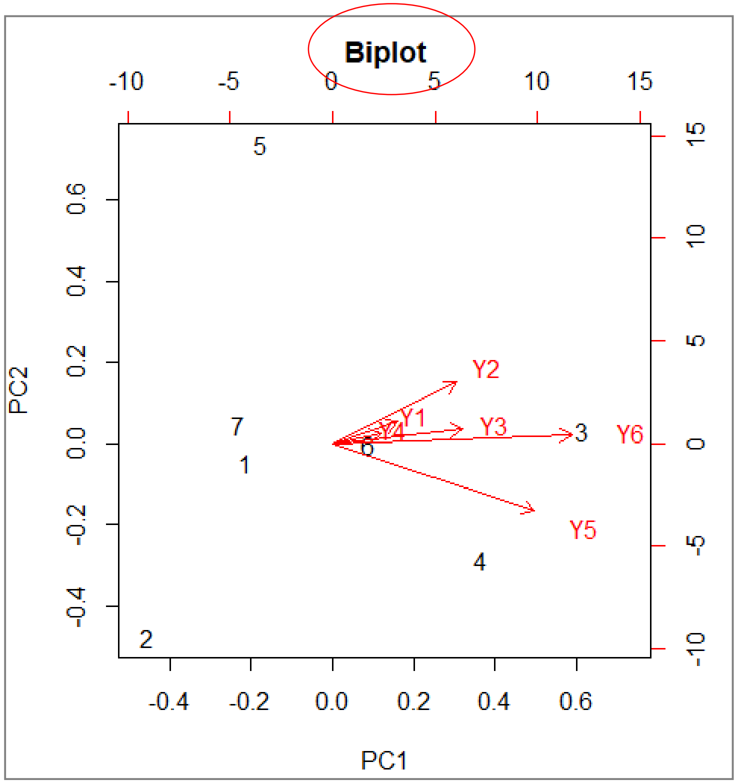
	Y1	Y2	Y3	Y4	Y5	Y6
Y1	**2.88**	5.25	4.85	1.93	6.53	7.74
Y2	5.25	**10.56**	8.90	3.59	11.46	15.58
Y3	4.85	8.90	**9.61**	3.51	13.43	16.31
Y4	1.93	3.59	3.51	**1.36**	4.86	5.92
Y5	6.53	11.46	13.43	4.86	**24.36**	24.68
Y6	7.74	15.58	16.31	5.92	24.68	**31.52**

→ CoP

Matriz de Distância Euclidiana: Dnxn

	1	2	3	4	5	6	7
1	0.00	6.21	18.70	13.13	4.83	7.43	2.03
2	6.21	0.00	24.34	18.55	9.44	12.94	6.62
3	18.70	24.34	0.00	5.99	18.38	12.50	19.20
4	13.13	18.55	5.99	0.00	13.64	7.26	13.78
5	4.83	9.44	18.38	13.64	0.00	7.98	5.09
6	7.43	12.94	12.50	7.26	7.98	0.00	8.67
7	2.03	6.62	19.20	13.78	5.09	8.67	0.00

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



Biplots

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

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

Matriz de "escores dos CP"

Matriz de "cargas"

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

$\Re^p \rightarrow \Re^m$

m=2

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

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

Análises das Linhas e das colunas sob os "mesmos" autovalores

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

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

c=0: linhas em coordenadas principais e colunas em coordenadas padronizadas
c=1: linhas em coordenadas padronizadas e colunas em coordenadas principais
c=1/2: representação mais geral

Pesquisar a diferença na aplicação das técnicas **PCA, PCoA e t-SNE**:
PCA = Principal Component Analysis (Análise de Componentes Principais)
PCoA = Principal Coordinate Analysis (Análise de Coordenadas Principais)
t-SNE = Distributed Stochastic Neighbor Embedding (Mapeamento por Vizinhança Estocástica t-Distribuída)

Lembre que:
1. $Y=UD^{1/2}V'$; sendo U e V matrizes de autovetores e D matriz diagonal com os autovalores.
2. $R=VDV'$, que é a base para a PCA.
3. $B=UDU'$, a base para o Escalonamento Multidimensional (Sol. Métrica) - PCoA.

Neste caso, podemos ter uma redução de dimensionalidade obtendo os m=2 primeiros PCs, bem como as duas primeiras Pco, dados por:
De 2: PCA1=YV1 e PCA2=YV2, V1 e V2 os dois primeiros autovetores das colunas de Y
De 3: $PCoA1=U1D^{1/2}$ e $PCoA2=U2D^{1\backslash 2}$, com U1 e U2 os dois primeiros autovetores das linhas de Y

Ainda, pode ser mostrado que: $YV= (UD^{1/2}V')V= UD^{1/2}$. Assim, pode ser estabelecida equivalência entre PCA e PCoA. Contudo, as PCo's podem ser formuladas por meio de soluções mais gerais do que esta solução via a decomposição espectral da matriz B. O Escalonamento Multidimensional pode ser entendido como uma técnica multivariada de dados que, mais que realizar uma redução de dimensionalidade, permite construir escalas, eixos de representação de observações a partir de uma matriz de distância (ou de similaridade) qualquer (não, necessariamente, Euclidiana). As PCo's obtidas de B são soluções métricas, mas a teoria também engloba soluções não-métricas (muitas vezes chamadas de soluções computacionais ou não-lineares). A t-SNE pode ser entendida como uma solução algorítmica nesta classe geral de soluções não-métricas que resolve o problema de construir uma escala para representar em uma baixa dimensão as unidades amostrais de um espaço de alta dimensão (big-p). A superioridade de PCA (ou PCoA) ou de t-SNE (bem como de seus derivados, como t-ETE), depende da estrutura dos dados e do problema de otimização imposto pelo objetivo da pesquisa, por exemplo, a solução PCA é ótima para dados bem representados por elipsoides de concentração.