

MAE 5776

# ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

[pavan@ime.usp.br](mailto:pavan@ime.usp.br)

IME/USP/2020

# Dados Multivariados

Banco de Dados:

Unidades Amostras	Variáveis					
	1	2	...	j	...	p
1	$Y_{11}$	$Y_{12}$	...	$Y_{1j}$	...	$Y_{1p}$
2	$Y_{21}$	$Y_{22}$	...	$Y_{2j}$	...	$Y_{2p}$
...	...	...	...	...	...	...
i	$Y_{i1}$	$Y_{i2}$	...	$Y_{ij}$	...	$Y_{ip}$
...	...	...	...	...	...	...
n	$Y_{n1}$	$Y_{n2}$	...	$Y_{nj}$	...	$Y_{np}$



$Y_{n \times p} = (y_{ij})$ : Matriz de Dados



resposta do i-ésimo "indivíduo" na j-ésima variável

Espaço das unidades amostrais (indivíduos): Linhas de Y  
Espaço das variáveis: Colunas de Y

# MOTIVAÇÃO

Exemplos considerados em Manly, 2005.

Cães pré-históricos da Tailândia.

<b>Grupo</b>	<b>X1</b>	<b>X2</b>	<b>X3</b>	<b>X4</b>	<b>X5</b>	<b>X6</b>
G1	9.7	21.0	19.4	7.7	32.0	36.5
G2	8.1	16.7	18.3	7	30.3	32.9
G3	13.5	27.3	26.8	10.6	41.9	48.1
G4	11.5	24.3	24.5	9.3	40.0	44.6
G5	10.7	23.5	21.4	8.5	28.8	37.6
G6	9.6	22.6	21.1	8.3	34.4	43.1
Cão Pré-h	10.3	22.1	19.1	8.1	32.2	35.0

Discuta possíveis questões de interesse:

Qual grupo de animal (G1, G2, ..., G6) está mais “próximo” do cão pré-histórico? E qual está mais “distante” ?

# MOTIVAÇÃO

Dados apresentados em Mardia et al. (2003)

## MULTIVARIATE ANALYSIS

Table 1.4.1 Weights of cork deposits (in centigrams) for 28 trees in the four directions (after Rao, 1948)

N	E	S	W	N	E	S	W
72	66	76	77	91	79	100	75
60	53	66	63	56	68	47	50
56	57	64	58	79	65	70	61
41	29	36	38	81	80	68	58
32	32	35	36	78	55	67	60
30	35	34	26	46	38	37	38
39	39	31	27	39	35	34	37
42	43	31	25	32	30	30	32
37	40	31	25	60	50	67	54
33	29	27	36	35	37	48	39
32	30	34	28	39	36	39	31
63	45	74	63	50	34	37	40
54	46	60	52	43	37	39	50
47	51	52	43	48	54	57	43

Discuta possíveis questões de interesse: Em qual direção de crescimento das árvores há maior variabilidade no armazenamento de cortiça?

# MOTIVAÇÃO

Exemplos considerados em Manly, 2005.

Medidas biométricas (mm) de Pardais fêmea

(Hermon Bumps, 1898).

Variável que define grupos

Variáveis quantitativas avaliadas nas u.a.

Pardal	Sobrev.	X1	X2	X3	X4	X5
1	S	156	245	31.6	18.5	20.5
...	...					
21	S	159	236	31.5	18.0	21.5
22	N	155	240	31.4	18.0	20.7
...	...					
49	N	164	248	32.3	18.8	20.9

tamanhos amostrais

Análise dos dados  $\Rightarrow$  pode dar suporte à teoria da Seleção Natural de Darwin

Discuta possíveis questões de interesse.

# BANCO DE DADOS: HATCO

Unidades amostrais: Clientes da HATCO (Hair et al., 2005)

ID	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
1	4,1	0,6	6,9	4,7	2,4	2,3	5,2	0	32	4,2	1	0	1	1
2	1,8	3,0	6,3	6,6	2,5	4,0	8,4	1	43	4,3	0	1	0	1
3	3,4	5,2	5,7	6,0	4,3	2,7	8,2	1	48	5,2	0	1	1	2
4	2,7	1,0	7,1	5,9	1,8	2,3	7,8	1	32	3,9	0	1	1	1
5	6,0	0,9	9,6	7,8	3,4	4,6	4,5	0	58	6,8	1	0	1	3
6	1,9	3,3	7,9	4,8	2,6	1,9	9,7	1	45	4,4	0	1	1	2
...														
97	6,1	0,5	9,2	4,8	3,3	2,8	7,1	0	60	5,2	1	0	1	3
98	2,0	2,8	5,2	5,0	2,4	2,7	8,4	1	38	3,7	0	1	0	1
99	3,1	2,2	6,7	6,8	2,6	2,9	8,4	1	42	4,3	0	1	0	1
100	2,5	1,8	9,0	5,0	2,2	3,0	6,0	0	33	4,4	1	0	0	1

Var. da Percepção dos Clientes sobre o Fornecedor (HATCO)

Escala: [0,10]

Var. do Produto

Escala X9: [0,100]

Escala X10: [0,10]

Demais variáveis: Características do Cliente (empresa compradora)

# BANCO DE DADOS: Concentrações sanguíneas

Concentrações sanguíneas de açúcar, sódio e potássio em mulheres adultas

Indiv.	Açúcar	Sódio	Potássio
1	3,7	48,5	9,3
2	5,7	65,1	8
3	3,8	47,2	10,9
4	3,2	53,2	12
5	3,1	55,5	9,7
6	4,6	36,1	7,9
7	2,4	24,8	14
8	7,2	33,1	7,6
9	6,7	47,4	8,5
10	5,4	54,1	11,3
11	3,9	36,9	12,7
12	4,5	58,8	12,3
13	3,5	27,8	9,8
14	4,5	40,2	8,4
15	1,5	13,5	10,1
16	8,5	56,4	7,1
17	4,5	71,6	8,2
18	6,5	52,8	10,9
19	4,1	44,1	11,2
20	5,5	40,9	9,4

Como este grupo está caracterizado relativamente a essas concentrações sanguíneas?

Que valores são esperados?  
Há variabilidade entre as mulheres?  
Há correlação entre concentrações?

# Banco de Dados: Morfometria cefálica

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

Qual é a correlação entre a morfometria cefálica do primeiro e do segundo filho?

# Dados Multivariados

$Y_{n \times p} = (y_{ij})$ : Matriz de Dados

- Classificação das Variáveis:

Qualitativas {  
Nominal  
Ordinal

Quantitativas {  
Discreta  
Contínua

Análises  
clássicas

- $n > p$  (Hair et al., 2005:  $n > 100$ ;  $n > 5p$ ;  $|\rho| > 0,30$ )
- Observações (unidades amostrais) independentes:  $n$  é tamanho amostral
- Variáveis quantitativas:  $p$  é a dimensão dos dados multivariados

?



Mas, e no caso de:

- “Big data” com  $n \ll p$
- observações correlacionadas (Ex.: dados de famílias em que há parentesco entre os indivíduos)

# MOTIVAÇÃO

## Plataformas de Bancos de Dados

- Kaggle: <https://www.kaggle.com>

[ChestX-ray8 Hospital-Scale Chest CVPR 2017 paper.pdf](#) (Wang et al., 2016).

*This Dataset is comprised of 112,120 X-ray images with disease labels from 30,805 unique patients. There are 12 zip files in total and range from ~2gb to 4gb in size. Additionally, we took a random sample (5%) of these images containing 5,606 X-ray images and class labels...*

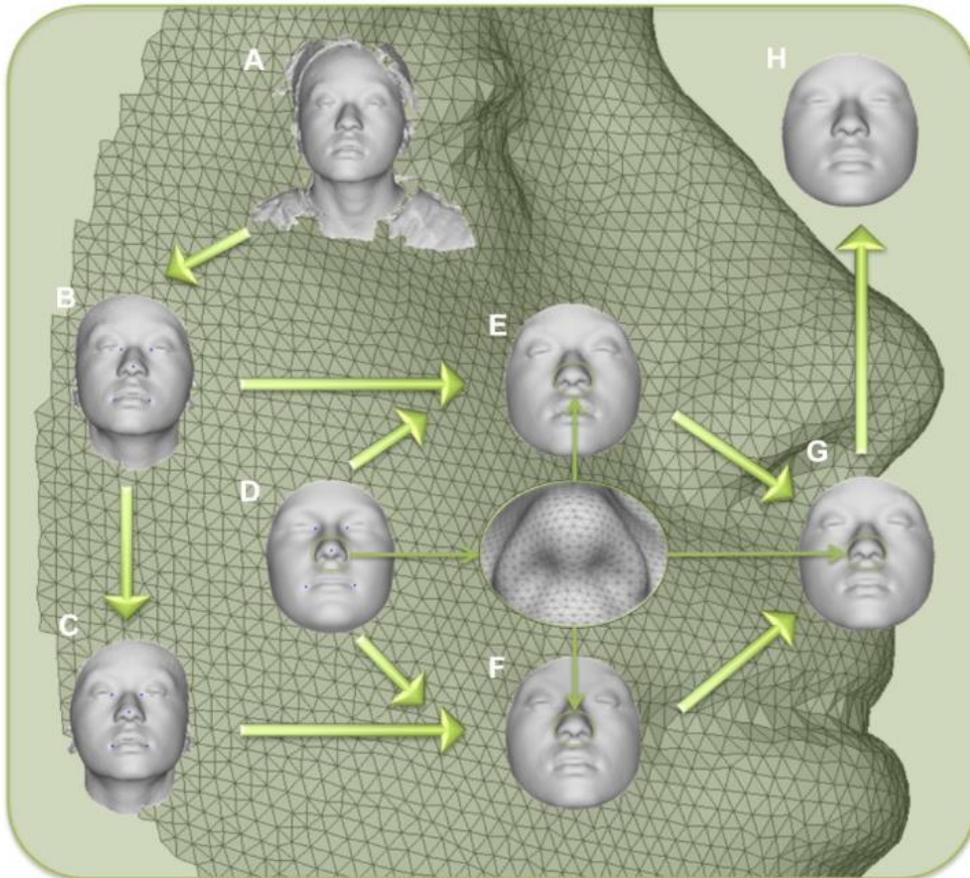
- 1000 Genomes: <https://www.internationalgenome.org/> Dados do genoma de populações mundiais

**EpiGen** (<http://epigen.grude.ufmg.br/>): mapeamento genético da asma, crescimento e longevidade. Inclui 6.600 indivíduos de Salvador (BA), Pelotas (RS) e Bambuí (MG). 2,5 milhões SNPs (amostra total), 5 milhões SNPs (subamostra) e o sequenciamento de 30 indivíduos.

**Projeto Corações de Baependi** (LBMGC-InCor/USP; Oliveira et al., 2008): estudo Longitudinal (2006, 2012, 2018), ≈2.000 indivíduos de 120 famílias, 1 milhão SNPs (amostra total), exoma e metaboloma (subamostra).

# Morfometria da Face Humana

⇒ A variabilidade da face humana tem sido largamente estudada cientificamente, mas sua predição é ainda objeto de grande desafio.



Máscara antropométrica para processamento e digitalização 3D da antropometria da face.

Claes, P. et al. (2014). Modeling 3D facial shape from DNA. Plos Genetics 10 (3):1-14.

# Transcriptoma

⇒ Em Transcriptômica, um grande desafio é entender as diferentes fontes de variação que influenciam a intensidade de expressão de genes. Por exemplo, a expressão diferencial de um gene é maior entre diferentes tecidos ou entre diferentes indivíduos (replicatas biológicas)?

Irizarry, R.A. and Love, M.I.

Data Analysis for the Life Sciences, 2015.

$$Y_{189 \times 22.215} = \begin{pmatrix} Y_1' \\ \dots \\ Y_{189}' \end{pmatrix} = (Y_{(1)}, \dots, Y_{(22.215)})$$

```
library(devtools)
install_github("genomicsclass/tissuesGeneExpression")

library(tissuesGeneExpression)
data(tissuesGeneExpression)
dim(e) ## e contains the expression data
## [1] 22215 189

table(tissue) ##tissue[i] tells us what tissue is represented by e[,i]
## tissue
## cerebellum colon endometrium hippocampus kidney liver placenta
## 38          34          15          31          39          26          6
```

# Matriz de Dados

$$Y_{n \times p} = (Y_{ij}) = \begin{pmatrix} Y_{1.}' \\ Y_{2.}' \\ \dots \\ Y_{n.}' \end{pmatrix} = (Y_{.1}, Y_{.2}, \dots, Y_{.p})$$

Espaço dos indivíduos:  $n$  vetores em um espaço  $p$ -dimensional ( $\mathfrak{R}^p$ )

$$Y_{i. (p \times 1)} = (Y_{i1}, Y_{i2}, \dots, Y_{ip})'; \quad i = 1, 2, \dots, n$$

Espaço das variáveis:  $p$  vetores em um espaço  $n$ -dimensional ( $\mathfrak{R}^n$ )

$$Y_{.j (n \times 1)} = (Y_{1j}, Y_{2j}, \dots, Y_{nj})'; \quad j = 1, 2, \dots, p$$



Explorar as propriedades geométricas de espaços vetoriais

# Estatísticas Descritivas Multivariadas

Dados Multivariados Quantitativos:

$Y_{n \times p}$

Vetor de médias:  $\bar{Y}_{p \times 1}$

Matriz de covariâncias:  $S_{p \times p} = (s_{j j'})$

Matriz de correlações:  $R_{p \times p} = (r_{j j'})$

Matriz de precisão:  $S^{-1}_{p \times p}$

Matriz de distâncias entre observações:  $D_{n \times n} = (d_{ii'}^2)$

Análise das variáveis!

Análise das observações!

# Estatísticas Descritivas Multivariadas

Dados Multivariados Quantitativos:

Uma amostra aleatória de  $n$  observações de uma população multivariada (em  $\mathbb{R}^p$ )

$Y_{n \times p}$

Vetor de médias:  $\bar{Y}_{p \times 1}$  ;  $\bar{Y}' = (\bar{Y}_1, \dots, \bar{Y}_p)$

Matriz de covariâncias:  $S_{p \times p} = (s_{jj'})$

Matriz de correlações:  $R_{p \times p} = (r_{jj'})$

$\bar{y}_j$  ?

$s_{jj'}$  ?

$r_{jj'}$  ?

Conceito de produto interno!

$$\bar{Y}_{p \times 1} = \begin{pmatrix} \bar{Y}_1 \\ \dots \\ \bar{Y}_p \end{pmatrix} = \frac{1}{n} \mathbf{1}'_n Y_{n \times p} = \frac{1}{n} Y' \mathbf{1}_n$$

**Centróide:** vetor de médias das p variáveis

**Matriz de covariâncias**

*Produto interno centrado*

$$S_{p \times p} = (s_{jj'}) = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})_{p \times 1} (Y_i - \bar{Y})'_{1 \times p} = \frac{1}{n} (Y'Y - n\bar{Y}\bar{Y}')$$

$$= \frac{1}{n} \left( Y'Y - \frac{1}{n} Y' \mathbf{1}_n \mathbf{1}_n' Y \right) = \frac{1}{n} \left[ Y' \left( I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' \right) Y \right] = \frac{1}{n} Y' H Y = (HY)' HY$$

*HY: centraliza as linhas de Y*

$$H_{n \times n} = I_n - \frac{1}{n} \mathbf{1} \mathbf{1}'$$

H é matriz simétrica e idempotente ( $H=H'$ ,  $H=H^2$ )

$$(Y'Y)_{p \times p} = \sum_{i=1}^n Y_i Y_i'$$

*Matriz de produto interno ordinário (matriz de soma de quadrados e produtos cruzados)*

# Estatísticas Descritivas Multivariadas

$$\bar{Y}_j = \frac{1}{n} \sum_{i=1}^n Y_{ij} = \frac{1}{n} Y_{.j}' \mathbf{1}_n$$

Média da variável j (escalar)

$$s_{jj'} = \frac{1}{n} \sum_{i=1}^n (Y_{ij} - \bar{Y}_j)(Y_{ij'} - \bar{Y}_{j'})$$

Covariância entre as variáveis j e j' (escalar)

*(Note: A diagram shows a red dashed box around the denominator 'n' and a blue arrow pointing to a red dashed box containing 'n-1', indicating the degrees of freedom for the variance calculation.)*

$$r_{jj'} = \frac{s_{jj'}}{\sqrt{s_{jj}} \sqrt{s_{j'j'}}} = \frac{\sum_{i=1}^n (Y_{ij} - \bar{Y}_j)(Y_{ij'} - \bar{Y}_{j'})}{\sqrt{\sum_{i=1}^n (Y_{ij} - \bar{Y}_j)^2} \sqrt{\sum_{i=1}^n (Y_{ij'} - \bar{Y}_{j'})^2}}$$

Note que a correlação entre as variáveis j e j' é igual à covariância entre as variáveis j e j' padronizadas

$$S_{p \times p} = \frac{1}{n} Y'HY = \begin{pmatrix} S_{11} & S_{12} & & S_{1p} \\ S_{21} & S_{22} & & S_{2p} \\ & & \dots & \\ S_{p1} & S_{p2} & & S_{pp} \end{pmatrix} ; H=H', H=H^2$$

- S é matriz positiva semidefinida (p.s.d)

$$a' Sa = \frac{1}{n} a' Y'HYa = \frac{1}{n} a' Y'H'HYa = \frac{1}{n} u'u \geq 0; \quad u = HYa$$

*n > p: S é p.d.*

- Matriz de correlações (matriz covariâncias das variáveis padronizadas)

$$R = D_{s_{jj}}^{-1/2} S D_{s_{jj}}^{-1/2} \quad S = D_{s_{jj}}^{1/2} R D_{s_{jj}}^{1/2}; \quad D_{s_{jj}}^{1/2} = \text{diag}(\sqrt{s_{jj}})$$

- Matriz de covariâncias com denominador (n-1)

*"no R": cov(Y)*

$$S_u = \frac{n}{n-1} S = \frac{1}{n-1} Y'HY$$

# Variabilidade Multivariada

- Variância total (traço da matriz de covariâncias)

$$trS = tr\Lambda = \sum_{j=1}^p \lambda_j$$

Autovalores da  
decomposição  
spectral de S

É importante para obter vetores  
reducionistas:  $\max(trS)$

- Variância generalizada (determinante da matriz de covariâncias)

$$|S| = |\Lambda| = \prod_{j=1}^p \lambda_j$$

$$|S| = (S_{11}S_{22}\dots S_{pp})|R|$$

É importante  
para obter  
EMVS na  $N_p$

# Matriz de Precisão - Correlação Parcial

- Matriz de Covariâncias e Correlação entre Variáveis

$$Y_{n \times 3} = (Y_{.1}, Y_{.2}, Y_{.3}) \quad S_{3 \times 3} = \begin{pmatrix} s_{11} & s_{12} & s_{13} \\ & s_{22} & s_{23} \\ & & s_{33} \end{pmatrix}$$



$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}} \sqrt{s_{kk}}}$$

Correlação (marginal) entre as variáveis  $Y_j$  e  $Y_k$

Grafos Não-direcionados

- Matriz de Precisão (Concentração) - Correlação Parcial entre Variáveis

$$S^{-1} = \begin{pmatrix} s_{11} & s_{12} & s_{13} \\ & s_{22} & s_{23} \\ & & s_{33} \end{pmatrix}^{-1} = \begin{pmatrix} \omega_{11} & \omega_{12} & \omega_{13} \\ & \omega_{22} & \omega_{23} \\ & & \omega_{33} \end{pmatrix}$$



$$r_{jk|l} = r(Y_j, Y_k | Y_l) = \frac{-\omega_{jk}}{\sqrt{\omega_{jj}} \sqrt{\omega_{kk}}};$$

Correlação parcial entre pares de variáveis dado as demais variáveis

$$r_{jk|l} = 0 \Leftrightarrow \omega_{jk} = 0$$

No caso da Normal Multivariada ( $N_p$ ): correlação parcial nula  $\Rightarrow$  independência aos pares entre variáveis  $\Rightarrow$  independência total

# Correlação Parcial – Aplicação em Aprendizado de Estruturas

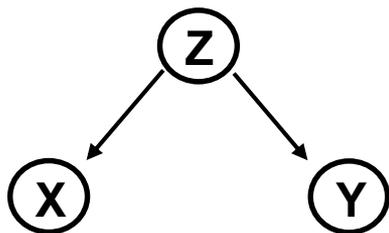
Ribeiro et al., 2016.

Padrões de dependência entre trios de variáveis

Grafos Direcionados

Caso 1

Z é causa comum  
Z é confundidor

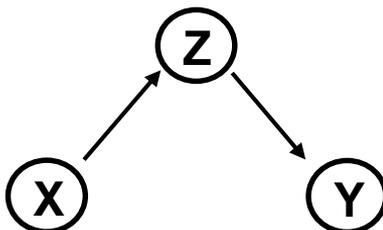


$(X \perp Y)$  correlacionados marginalmente

$(X \perp Y | Z)$  condicionalmente independentes

Caso 2

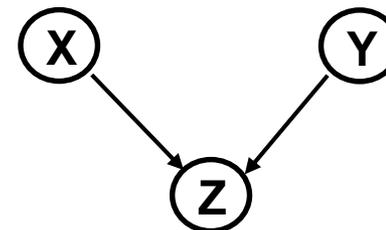
Z é efeito intermediário  
X tem efeito indireto sobre Y



Grafos diferentes mas com Modelos Probabilísticos Equivalentes

Caso 3

Z é efeito comum  
(colisão não conectada ou v-estrutura)



$(X \perp Y)$  independência marginal

$(X \perp Y | Z)$  dependência condicional

# Estatísticas Descritivas Multivariadas

Dados do depósito de cortiça (Tabela 1.4.1, Mardia et al., 2003)

Matriz de dados: n=28 observações avaliadas em p=4 variáveis.

$$Y_{28 \times 4} = (Y_{(N)}, Y_{(E)}, Y_{(S)}, Y_{(W)})$$

$$\bar{Y}_{4 \times 1} = \begin{matrix} & \text{N} & \text{E} & \text{S} & \text{W} \\ \text{N} & 50.53571 & 46.17857 & 49.67857 & 45.17857 \end{matrix}$$

$$S_{4 \times 4} = \begin{pmatrix} s_{11} & s_{12} & s_{13} & s_{14} \\ & s_{22} & s_{23} & s_{24} \\ & & s_{33} & s_{34} \\ & & & s_{44} \end{pmatrix} = \begin{matrix} & \text{N} & \text{E} & \text{S} & \text{W} \\ \text{N} & 290.41 & 223.75 & 288.44 & 226.27 \\ \text{E} & 223.75 & 219.93 & 229.06 & 171.37 \\ \text{S} & 288.44 & 229.06 & 350.00 & 259.54 \\ \text{W} & 226.27 & 171.37 & 259.54 & 226.00 \end{matrix}$$

divisor=n-1

Há indicação de heterocedasticidade?

$$R_{4 \times 4} = \left( r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}} \sqrt{s_{kk}}} \right) = \begin{matrix} & \text{N} & \text{E} & \text{S} & \text{W} \\ \text{N} & 1.00 & 0.89 & 0.90 & 0.88 \\ \text{E} & 0.89 & 1.00 & 0.83 & 0.77 \\ \text{S} & 0.90 & 0.83 & 1.00 & 0.92 \\ \text{W} & 0.88 & 0.77 & 0.92 & 1.00 \end{matrix}$$

A estrutura de correlação uniforme entre as variáveis parece ser adequada?

# Estruturas de Correlação

Principais estruturas de correlação	Definição
Independente	$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \quad R = I_p$
Permutável Equicorrelação, uniforme	$\begin{pmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{pmatrix} \quad R = (1 - \alpha)I_p + \alpha \mathbf{1}_p \mathbf{1}_p'$
Não estruturada	$\begin{pmatrix} 1 & \alpha_{1,2} & \cdots & \alpha_{1,t} \\ \alpha_{1,2} & 1 & \cdots & \alpha_{2,t} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{1,t} & \alpha_{2,t} & \cdots & 1 \end{pmatrix}$
Auto regressiva de ordem 1	$\begin{pmatrix} 1 & a & \cdots & a^{t-1} \\ a & 1 & \cdots & a^{t-2} \\ \vdots & \vdots & \ddots & \vdots \\ a^{t-1} & a^{t-2} & \cdots & 1 \end{pmatrix}$

# Correlação Parcial

Exemplo: Dados do peso de Cortiça - “Aprendizado” da relação entre variáveis

**S**

```
> round(cov(dcork), 2)
```

	N	E	S	W
N	290.41	223.75	288.44	226.27
E	223.75	219.93	229.06	171.37
S	288.44	229.06	350.00	259.54
W	226.27	171.37	259.54	226.00

**R**

```
> round(cor(dcork), 2)
```

	N	E	S	W
N	1.00	0.89	0.90	0.88
E	0.89	1.00	0.83	0.77
S	0.90	0.83	1.00	0.92
W	0.88	0.77	0.92	1.00

Correlação (marginal)  
alta entre todos os  
pares de variáveis

```
> library(ppcor)  
> pc<-pcor(x=dcork)  
> round(pc$estimate, 4)
```

	N	E	S	W
N	1.0000	0.5933	0.2683	0.3384
E	0.5933	1.0000	0.2067	<b>-0.1770</b>
S	0.2683	0.2067	1.0000	0.6330
W	0.3384	-0.1770	0.6330	1.0000

Matriz de Correlação Parcial  
entre pares de variáveis dado  
as demais (obtida de  $S^{-1}$ )

Testar a significância das  
correlações parciais!

# Correlação Parcial

Exemplo: Dados do peso de Cortiça - “Aprendizado” da relação entre variáveis

```
> library(ppcor)
> pc<-pcor(x=dcork)
> round(pc$estimate, 4)
```

	N	E	S	W
N	1.0000	0.5933	0.2683	0.3384
E	0.5933	1.0000	0.2067	-0.1770
S	0.2683	0.2067	1.0000	0.6330
W	0.3384	-0.1770	0.6330	1.0000

```
> library(ppcor)
> pc<-pcor(x=dcork)
> round(pc$p.value, 4)
```

	N	E	S	W
N	-	0.0014	0.1850	0.0909
E	0.0014	-	0.3111	0.3870
S	0.1850	0.3111	-	0.0005
W	0.0909	0.3870	0.0005	-

Teste de correlação parcial nula (Teste z de Fisher)

Níveis descritivos do teste de correlação parcial nula

Correlações Parciais:  $(N \perp E | S, W)$   $(S \perp W | N, E)$

Independência condicional:  $(N \perp S | N, E); (N \perp W | N, E); (S \perp E | N, W)$

Assim, as únicas correlações parciais significantes são entre os pares (N e E) e (S e W). Os demais pares de variáveis são condicionalmente independentes (independentes dados as demais).

Essa análise permite construir grafos não direcionados entre as variáveis.

Como fica o grafo se adotarmos um nível de significância individual de 10%?

# Transformações Lineares – Medidas de Distância

Matriz de distâncias

## ▪ Distância Euclidiana entre observações

$$Y_{i_{p \times 1}} \in \mathfrak{R}^p; \quad d_{E_{ik}}^2 = (Y_i - Y_k)' (Y_i - Y_k) \in \mathfrak{R} \Rightarrow D_{E_{n \times n}} = (d_{ik})$$

## ▪ Transformação de Escala (padronização) $\Rightarrow$ Distância de Pearson

$$Y_{i_{p \times 1}} \Rightarrow Y_{i_{p \times 1}}^* = D_{s_{jj}}^{-1/2} (Y_i - \bar{Y})$$

$$(Y_i^* - Y_k^*)' (Y_i^* - Y_k^*) = (Y_i - Y_k)' D^{-1} (Y_i - Y_k) = d_P^2(Y_i, Y_k)$$

Distância de Pearson (ao quadrado) entre pares de observações (é a distância Euclidiana das variáveis padronizadas)

$$Y_i^{*'} Y_i^* = (Y_i - \bar{Y})' D^{-1} (Y_i - \bar{Y}) = d_P^2(Y_i, C)$$

Distância de Pearson da observação ao centróide

# Transformações Lineares – Medidas de Distância

## ▪ Transformação de Mahalanobis $\Rightarrow$ Distância de Mahalanobis

$$Y_{i_{p \times 1}} \Rightarrow Z_{i_{p \times 1}} = S^{-1/2} (Y_i - \bar{Y}) \quad S_Z = I_p \quad \text{Variáveis independentes e variâncias unitárias}$$

$$(Z_i - Z_k)' (Z_i - Z_k) = (Y_i - Y_k)' S^{-1} (Y_i - Y_k) = d_M^2 (Y_i, Y_k) \quad \text{Distância de Mahalanobis (ao quadrado) entre pares de observações}$$

$$Z_i' Z_i = (Y_i - \bar{Y})' S^{-1} (Y_i - \bar{Y}) = d_M^2 (Y_i, C) \quad \text{Distância de Mahalanobis da observação ao centróide}$$

# Q-Espaço e R-Espaço

“Q-técnicas”

“R-técnicas”

- Distância entre Observações – Correlação entre Variáveis

$$Y_{i_{p \times 1}} \in \mathfrak{R}^p, \quad i = 1, \dots, n \quad \Rightarrow \quad d_{E_{ik}}^2 = (Y_i - Y_k)' (Y_i - Y_k)$$
$$d_{P_{ik}}^2 = (Y_i^* - Y_k^*)' (Y_i^* - Y_k^*) = (Y_i - Y_k)' D_{s_{jj}}^{-1} (Y_i - Y_k)$$

$$Y_{(j)_{n \times 1}} \in \mathfrak{R}^n, \quad j = 1, \dots, p \quad \Rightarrow \quad r_{jl} = \frac{s_{jl}}{\sqrt{s_{jj}} \sqrt{s_{ll}}}$$

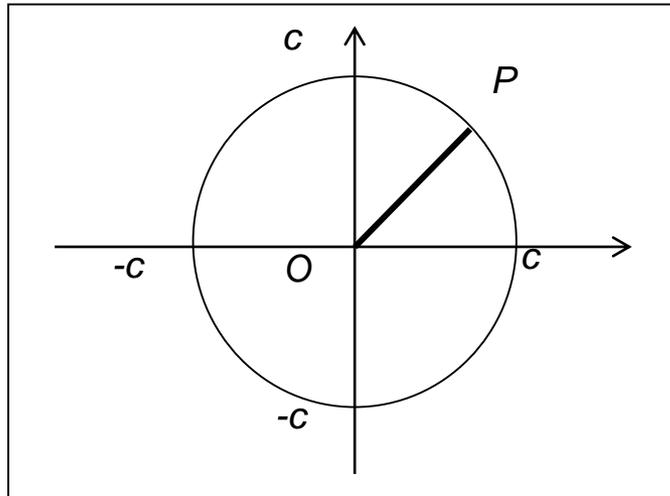


$$d_{P_{jl}}^2 = (Y_{(j)}^* - Y_{(l)}^*)' (Y_{(j)}^* - Y_{(l)}^*) = 2(1 - r_{jl})$$

A “distância” de Pearson entre variáveis é medida por sua correlação (coeficiente de correlação linear de Pearson).

# Distância entre Observações

Variáveis independentes e variâncias homogêneas  $\Rightarrow$  Distância Euclidiana



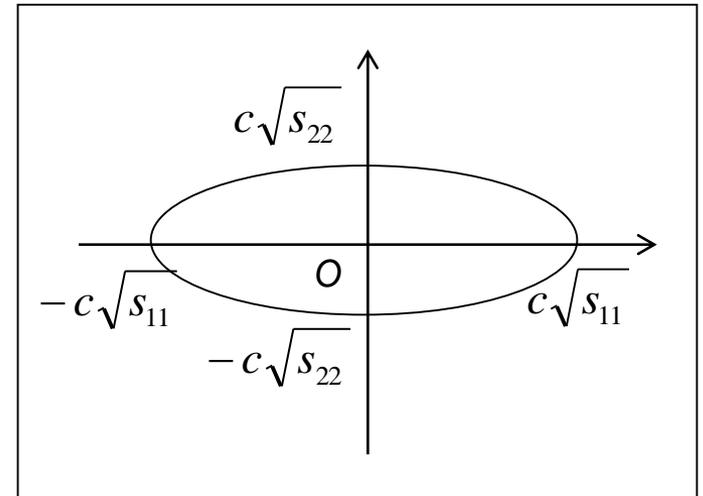
$$P = (y_1, y_2)$$

$$d_E^2(P, O) = y_1^2 + y_2^2 = c^2$$

Matricialmente e na dimensão p:

$$d_E^2(P, O) = (Y_i - \bar{Y})' (Y_i - \bar{Y})$$

Variáveis independentes e variâncias heterogêneas  $\Rightarrow$  Distância de Pearson



$$P = (y_1, y_2)$$

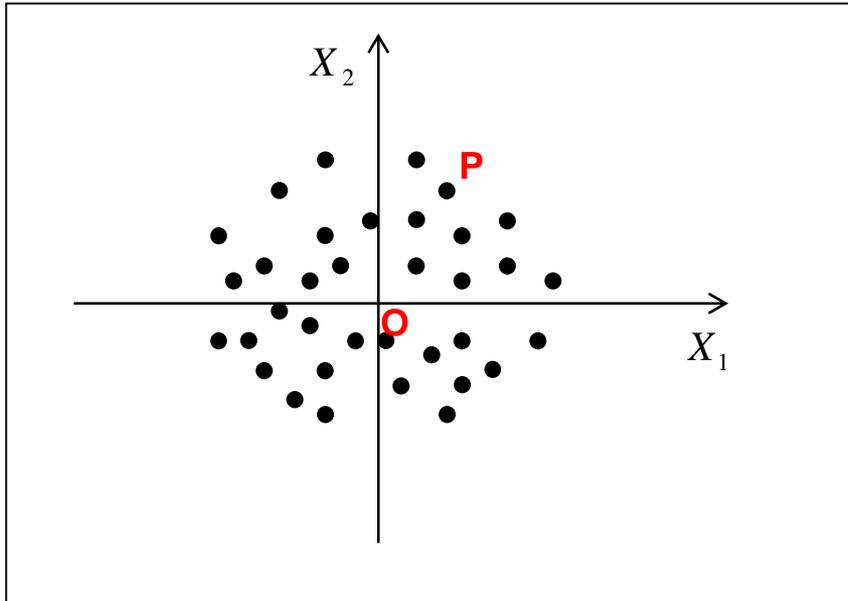
$$d_P^2(P, O) = \frac{y_1^2}{s_{11}} + \frac{y_2^2}{s_{22}} = c^2$$

Matricialmente e na dimensão p:

$$d_P^2(O, P) = (Y_i - \bar{Y})' D_{s_{jj}}^{-1} (Y_i - \bar{Y})$$

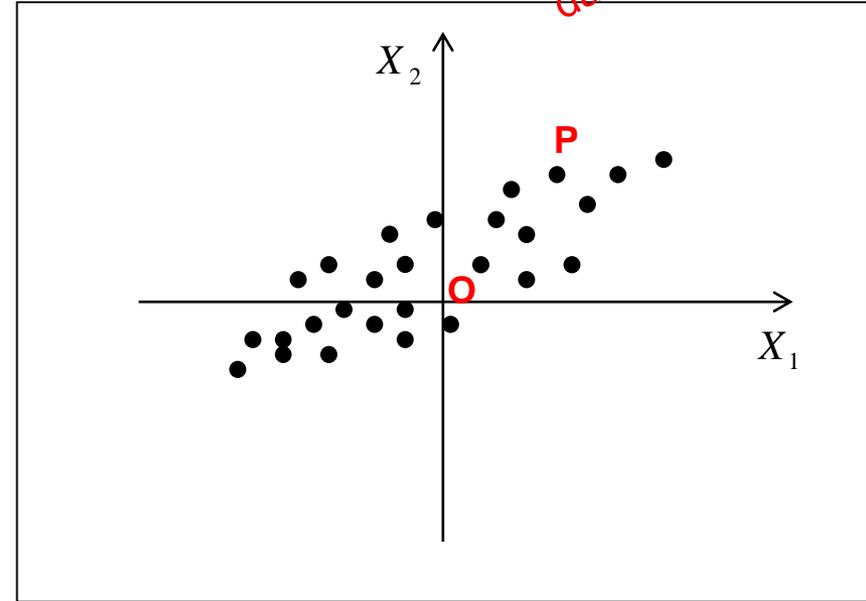
# Distância entre Observações

Considere dados bivariados



Variáveis independentes e variâncias homogêneas  $\Rightarrow$  Distância Euclidiana

$$d^2(P, O) = (Y_{p \times 1} - \bar{Y}_{p \times 1})' (Y_{p \times 1} - \bar{Y}_{p \times 1})$$



Variáveis correlacionadas e variâncias heterogêneas  $\Rightarrow$  Distância de Mahalanobis

$$d_M^2(P, O) = (Y_{p \times 1} - \bar{Y}_{p \times 1})' S^{-1} (Y_{p \times 1} - \bar{Y}_{p \times 1})$$

Variáveis independentes e variâncias heterogêneas  $\Rightarrow$  Distância de Pearson (Euclidiana Padronizada):

$$d^2(P, O) = (Y_{p \times 1} - \bar{Y}_{p \times 1})' D_{s_{jj}}^{-1} (Y_{p \times 1} - \bar{Y}_{p \times 1})$$



P e a origem O ( $=\bar{Y}$ ) são pontos p-dimensionais ( $\mathcal{R}^p$ )

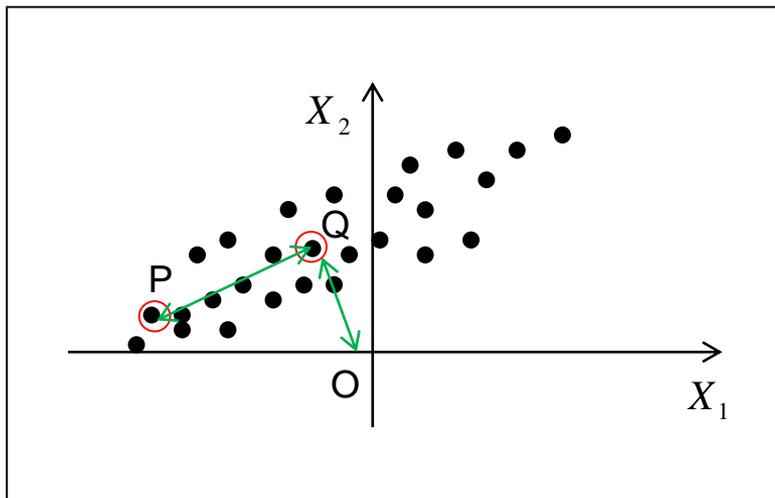
# Distância Estatística

⇒ Distância Euclidiana é apropriada para variáveis independentes e homocedásticas (variância homogênea).

⇒ Quando há heterocedasticidade uma alternativa é usar distância de Pearson, isto é, padronizar as variáveis.

⇒ No caso de variáveis correlacionadas e heterocedásticas uma alternativa é usar a distância de Mahalanobis.

## Motivação



Johnson and Wichern, 1992.

⇒ A distância Euclidiana de Q a P é maior do que a de Q à origem O.

⇒ Porém, note que O pode ser considerado ponto atípico (ocorre fora da nuvem de dispersão conjunta de  $X_1$  e  $X_2$ ), o que não ocorre com o ponto P.

⇒ Ao adotar medidas de distância estatística (Pearson, Mahalanobis), Q está mais próximo de P do que de O, o que pode ser mais razoável, considerando a dispersão dos pontos.

# Dados dos Cães Pré-históricos

```
> m<-round(colMeans(caes),2)
X1      X2      X3      X4      X5      X6
10.49 22.50 21.51  8.50 34.23 39.69
```

```
> v<-round(cov(caes),2)
      X1      X2      X3      X4      X5      X6
X1 2.88  5.25  4.85  1.93  6.53  7.74
X2 5.25 10.56  8.90  3.59 11.46 15.58
X3 4.85  8.90  9.61  3.51 13.43 16.31
X4 1.93  3.59  3.51  1.36  4.86  5.92
X5 6.53 11.46 13.43  4.86 24.36 24.68
X6 7.74 15.58 16.31  5.92 24.68 31.52
```

```
> r<-round(cor(caes),2)
      X1      X2      X3      X4      X5      X6
X1 1.00 0.95 0.92 0.98 0.78 0.81
X2 0.95 1.00 0.88 0.95 0.71 0.85
X3 0.92 0.88 1.00 0.97 0.88 0.94
X4 0.98 0.95 0.97 1.00 0.85 0.91
X5 0.78 0.71 0.88 0.85 1.00 0.89
X6 0.81 0.85 0.94 0.91 0.89 1.00
```

```
> trv<-sum(diag(v)) ## 80.29
> detv<-det(v) ## 0.114878
```

*divisor=(n-1)*

*Os cães são mais heterogêneos em qual variável?*

*A estrutura de correlação uniforme entre as variáveis parece ser adequada?*

```

> de<-round(dist(caes),2) ##dist Euclidiana
  1      2      3      4      5      6
2  6.21
3 18.70 24.34
4 13.13 18.55  5.99
5  4.83  9.44 18.38 13.64
6  7.43 12.94 12.50  7.26  7.98
7  2.03 6.62 19.20 13.78  5.09  8.67

```

Qual cão está mais próximo do pré-histórico? Calcule diferentes medidas de distância.

```

> ysd<-scale(caes, center = FALSE, scale = TRUE)
> dep<-round(dist(ysd),2) ##dist de Pearson
  1      2      3      4      5      6
2 0.26
3 0.72 0.96
4 0.45 0.68 0.29
5 0.20 0.43 0.61 0.38
6 0.20 0.42 0.57 0.30 0.22
7 0.09 0.33 0.68 0.43 0.17 0.22

```

```

> d2m<-round(mahalanobis(caes,m,v),2) ##dist Mahalanobis ao Centróide
3.06 4.97 5.08 5.42 5.21 5.00 4.74

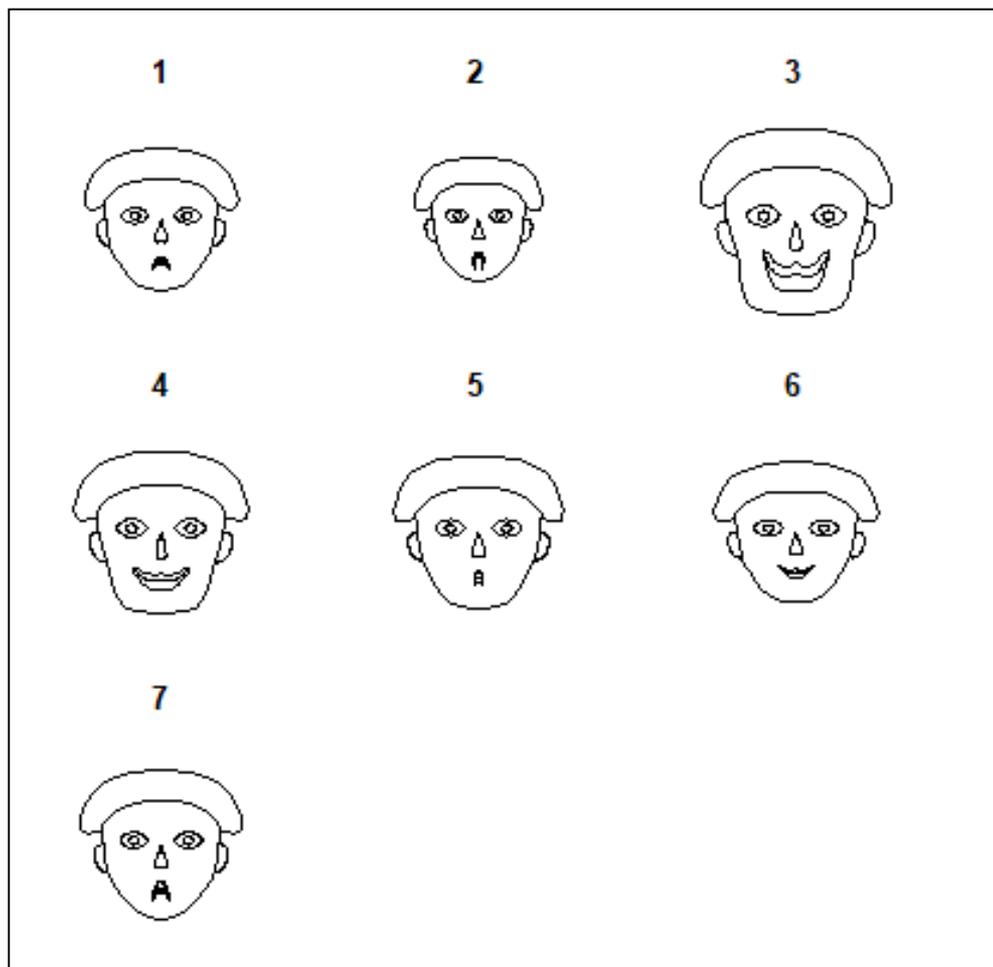
```

```

> dmlmy<-t(t(y[1,])-t(y[7,]))%*%solve(v)%*%(t(y[1,])-t(y[7,]))
  1      2      3      4      5      6      ##dist Mahalanobis ao Cão pré-h
7.68 11.84 11.53 11.75 11.76 11.89

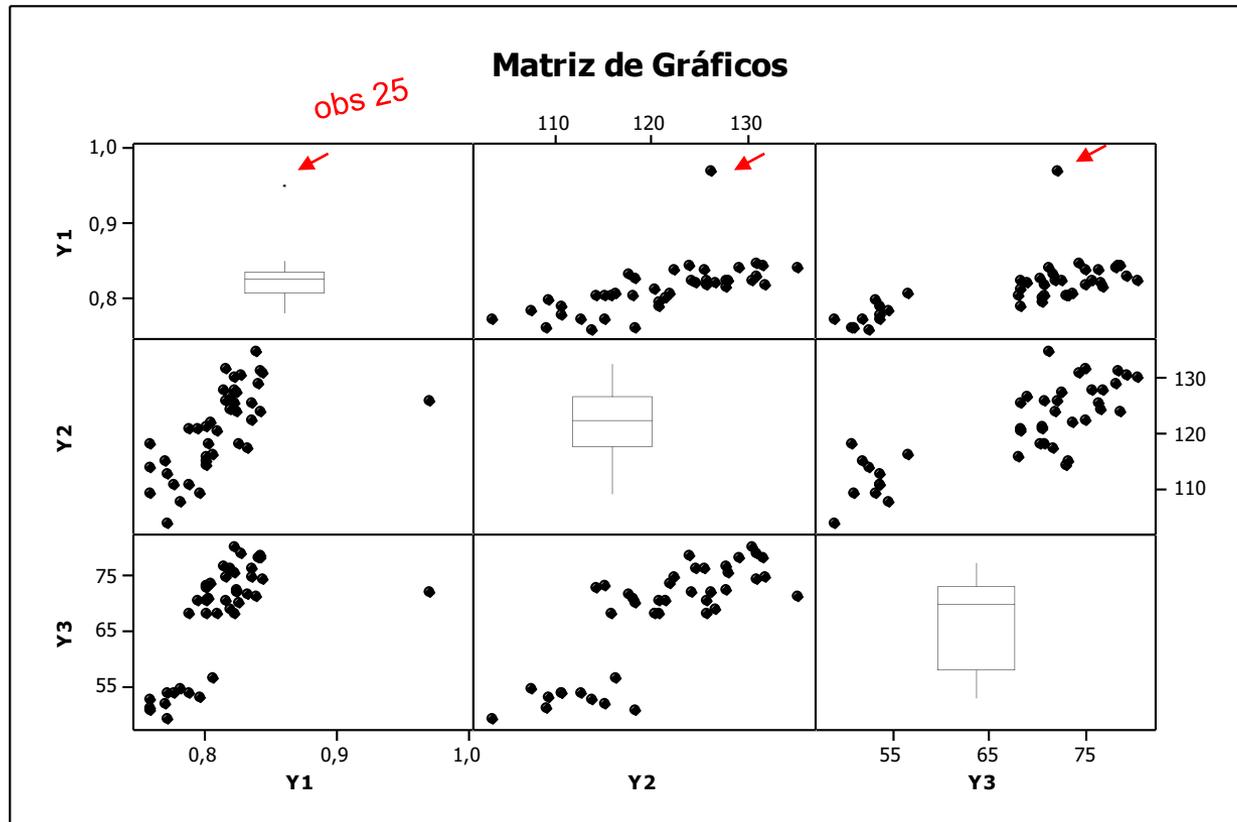
```

```
> library(TeachingDemos) ##faces de Chernoff  
> faces( as.matrix(y), fill=T) ##unidades amostrais
```



Qual cão está mais próximo do cão pré-histórico (obs 7)?  
Calcule diferentes medidas de distância!

# Observações Atípicas – Diagnóstico Gráfico



Johnson and Wichern, 2002

Causas para a ocorrência de observações atípicas:

- falta de controle amostral ou experimental
- amostragem de população heterogênea (estratificação)
- erros de mensuração
- erros no controle de qualidade dos dados
- variabilidade genuína

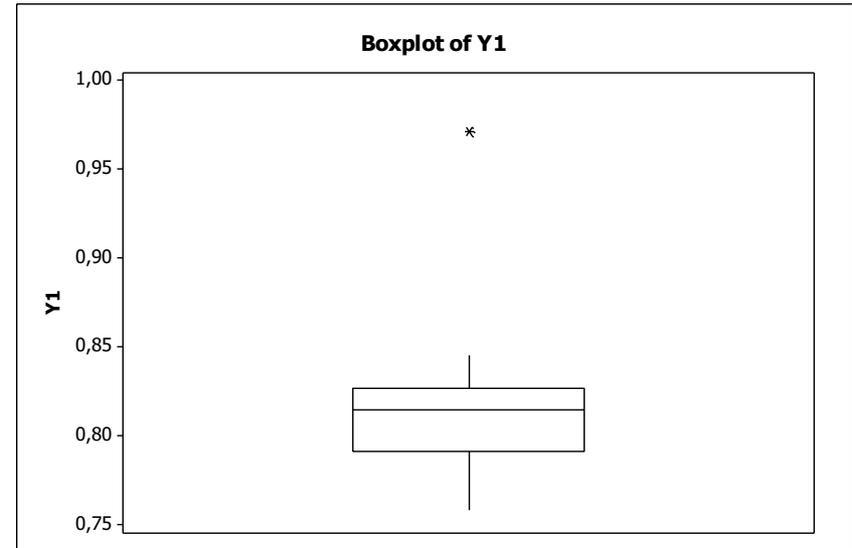
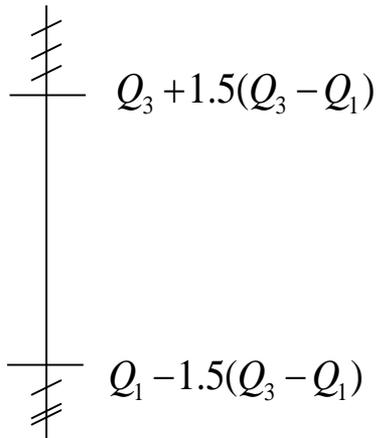
⇒ Obs 25 é um valor atípico (aberrante, *outlier*) unidimensional para Y1 (boxplot de Y1) mas não para Y2 ou Y3 (ver boxplot de Y2 e Y3)

⇒ Obs 25 é um valor atípico bidimensional para (Y1,Y2) e para (Y1,Y3), mas não para (Y2,Y3)

# Valor Atípico Unidimensional

$$Y = (Y_1, \dots, Y_n); \quad Y_i \in \mathfrak{R}, \quad i = 1, 2, \dots, n$$

- **Critério do Boxplot:**

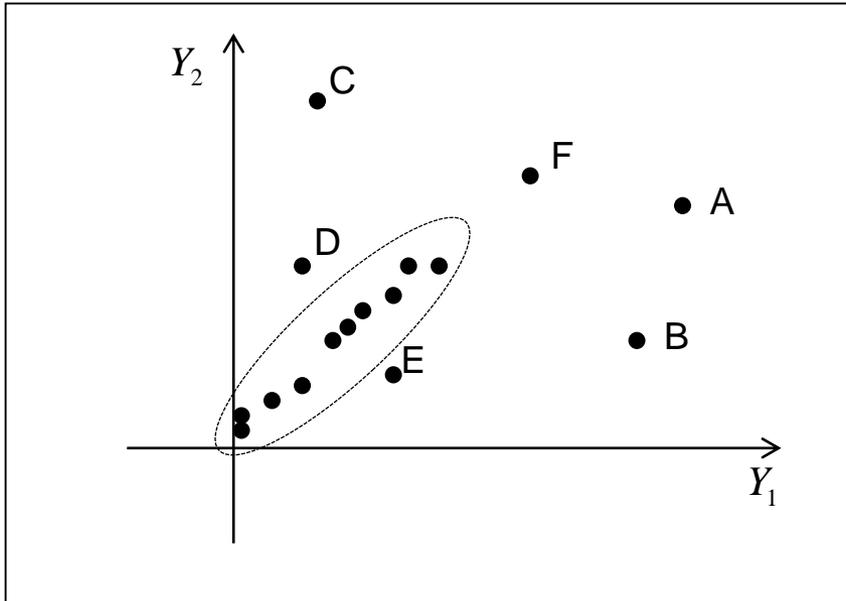


- Dados Padronizados:  $Z_i = \frac{Y_i - \bar{Y}}{s}$   $P(|Z| \geq 2,5) = 0,012$  (Hair et al., 1998)

a probabilidade de um ponto estar dentro do intervalo de concentração é  $(1-\alpha)$

- Medida de Distância:  $d^2 = \frac{(Y_i - \bar{Y})^2}{s^2} = Z_i^2 \stackrel{n \rightarrow \infty}{\sim} \chi_1^2 \Rightarrow P(d^2 \leq c^2) \leq (1-\alpha)$

# Valor Atípico Bidimensional



Qual é a influência destes pontos (A, B, C, D e F)

- na média e na variância de  $Y_1$  e  $Y_2$ ?
- na correlação entre  $Y_1$  e  $Y_2$ ?

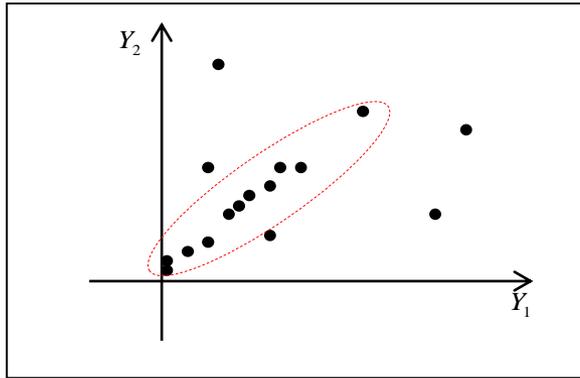
⇒ Ponto A: aberrante tanto para  $Y_1$  como para  $Y_2$

⇒ Pontos B (C) é aberrante para  $Y_1$  ( $Y_2$ ) mas não para  $Y_2$  ( $Y_1$ )

⇒ Pontos D e E: são aberrantes bidimensionais mas não unidimensionais

⇒ Ponto F: apesar de aberrante unidimensional (para  $Y_1$  e  $Y_2$ ) segue a tendência da nuvem de pontos amostrais

# Valor Atípico Bidimensional



Diagnóstico de valores atípicos bidimensionais (distantes da nuvem de dispersão conjunta dos pontos):

A distância Euclidiana de P ao centroide (C),

$$d_E^2(Y_i; C) = (Y_{i2 \times 1} - \bar{Y}_{2 \times 1})' (Y_{i2 \times 1} - \bar{Y}_{2 \times 1})$$

não leva em conta a correlação entre as variáveis. Além disso, há dificuldade em estabelecer um critério de diagnóstico com base nessa medida.

## Distância de Mahalanobis

$$d_M^2(Y_i; C) = (Y_{i2 \times 1} - \bar{Y}_{2 \times 1})' S^{-1} (Y_{i2 \times 1} - \bar{Y}_{2 \times 1})$$

S: matrix de covariância amostral

Diagnóstico:  $d_M^2 \stackrel{n \rightarrow \infty}{\sim} \chi_2^2 \Rightarrow P(d_M^2 \leq c^2) \leq (1 - \alpha)$

$$\left\{ (Y - \bar{Y})' S^{-1} (Y - \bar{Y}) = c^2 \right\}$$

**Elipse de concentração de pontos bidimensionais**, centrada na média. Os eixos principais da elipse (direções e comprimentos) podem ser obtidos da decomposição espectral de S.

# Valor Atípico Multidimensional

Identificação de valores atípicos (observações outliers)

Distância de Mahalanobis (Distância Generalizada) de pontos amostrais ao centróide:

$$d_M^2(P; C) = (Y_{p \times 1} - \bar{Y}_{p \times 1})' S^{-1} (Y_{p \times 1} - \bar{Y}_{p \times 1})$$

$$d_M^2 \stackrel{n \rightarrow \infty}{\sim} \chi_p^2 \Rightarrow P(d_M^2 \leq c^2) \leq (1 - \alpha)$$

$$\Rightarrow \frac{d_M^2}{p} \stackrel{(n-p) \rightarrow \infty}{\sim} t_p^2$$

critério  
assintótico e sob  
normalidade

$\left\{ (Y - \bar{Y})' S^{-1} (Y - \bar{Y}) = c^2 \right\}$  Estes **elipsóides** estão centrados na média, seus eixos estão na direção dos autovetores ( $V_j$ ) de  $S$  e os comprimentos desses eixos são proporcionais à raiz quadrada dos autovalores ( $\lambda_j$ ) de  $S$ .

Os elipsóides têm eixos  $\pm c \sqrt{\lambda_j} V_{kj}$ , em que  $|S - \lambda I_p| = 0$ ,  $SV_j = \lambda_j V_j$ ,  $j = 1, \dots, p$ .

No elipsóide os valores  $\mathbf{y} \in \mathbb{R}^p$ , satisfazendo a desigualdade

$$(\mathbf{y} - \bar{Y})' S^{-1} (\mathbf{y} - \bar{Y}) \leq \chi_p^2(\alpha), \text{ têm probabilidade } 1 - \alpha \text{ de ocorrerem.}$$

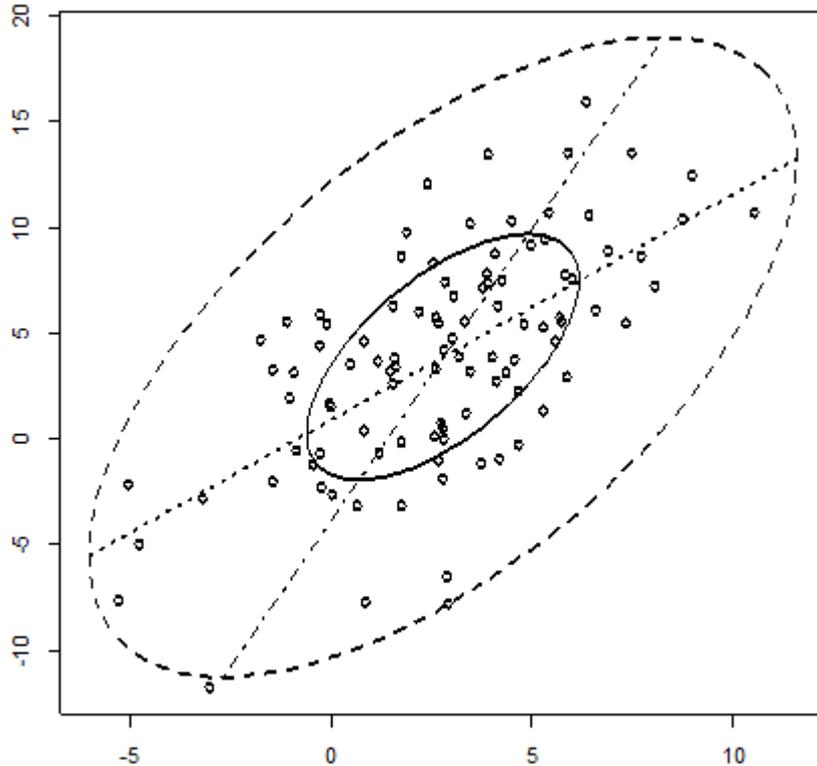
# BoxPlot Bivariado

## Elipse de Concentração de Observações

$$\mu' = (3, 4) \quad \Sigma = \begin{pmatrix} 9 & 10 \\ 10 & 25 \end{pmatrix}$$

**Boxplot Bivariado** (Everitt, 2007)

A elipse no centro inclui 50% dos dados.



A elipse maior fornece um critério (robusto) de diagnóstico de observações atípicas.

São apresentadas retas de regressão (de y vs. x e de x vs. y) com o estimador do centróide na intersecção.

A construção das retas de regressão pode ser por estimação robusta ou clássica. Quanto menor o ângulo entre as retas maior é o valor absoluto da correlação.

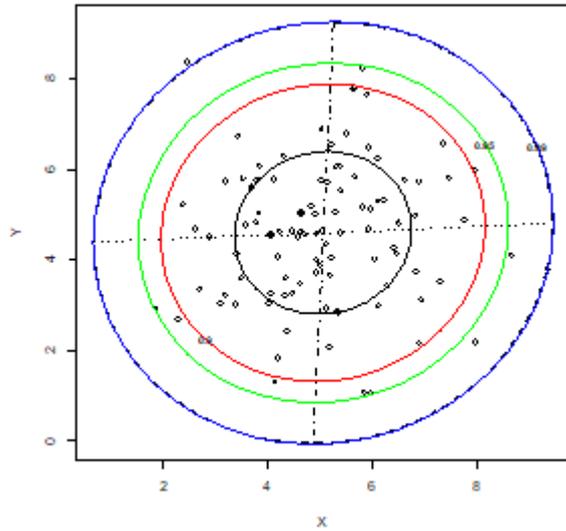
Note que na normal bivariada ambos, Y e X, são variáveis aleatórias. A construção das retas de regressão é um “abuso” justificado apenas para ilustrar as direções de eixos de variação dos dados.

$$d_M^2 = (Y_i - \bar{Y})' S^{-1} (Y_i - \bar{Y}) \leq c^2 = \chi_p^2(\alpha)$$

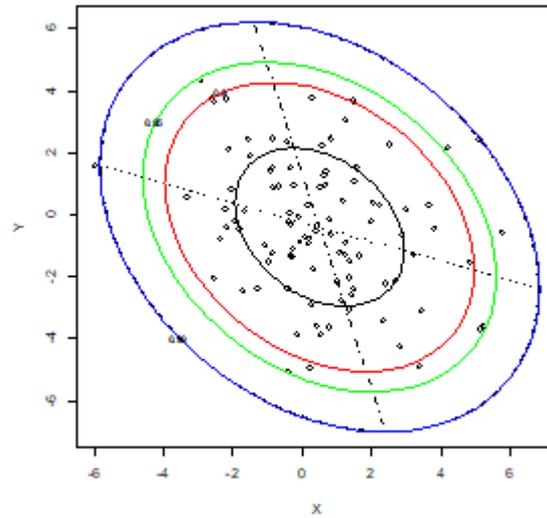


Em Everitt (2007) há outras alternativas de critérios robustos de diagnóstico de outliers bivariados.

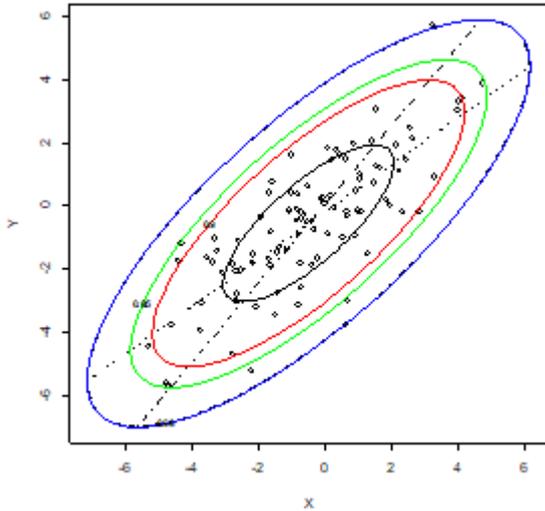
$$\mu' = (5,5) \quad \Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$



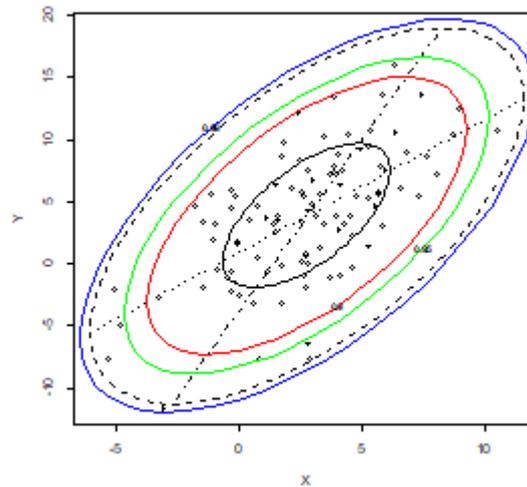
$$\mu' = (0,0) \quad \Sigma = \begin{pmatrix} 4 & -1 \\ -1 & 4 \end{pmatrix}$$



$$\mu' = (0,0) \quad \Sigma = \begin{pmatrix} 4 & 3 \\ 3 & 4 \end{pmatrix}$$



$$\mu' = (3,4) \quad \Sigma = \begin{pmatrix} 9 & 10 \\ 10 & 25 \end{pmatrix}$$



Função bivbox  
(Everitt, 2005)

# Aplicação: Banco de Dados HATCO

Unidades amostrais: Clientes da HATCO (Hair et al., 2005)

ID	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
1	4,1	0,6	6,9	4,7	2,4	2,3	5,2	0	32	4,2	1	0	1	1
2	1,8	3,0	6,3	6,6	2,5	4,0	8,4	1	43	4,3	0	1	0	1
3	3,4	5,2	5,7	6,0	4,3	2,7	8,2	1	48	5,2	0	1	1	2
4	2,7	1,0	7,1	5,9	1,8	2,3	7,8	1	32	3,9	0	1	1	1
5	6,0	0,9	9,6	7,8	3,4	4,6	4,5	0	58	6,8	1	0	1	3
6	1,9	3,3	7,9	4,8	2,6	1,9	9,7	1	45	4,4	0	1	1	2
...														
97	6,1	0,5	9,2	4,8	3,3	2,8	7,1	0	60	5,2	1	0	1	3
98	2,0	2,8	5,2	5,0	2,4	2,7	8,4	1	38	3,7	0	1	0	1
99	3,1	2,2	6,7	6,8	2,6	2,9	8,4	1	42	4,3	0	1	0	1
100	2,5	1,8	9,0	5,0	2,2	3,0	6,0	0	33	4,4	1	0	0	1

Var. da Percepção dos Clientes  
sobre o Fornecedor (HATCO)

Escala: [0,10]

Var. do Produto

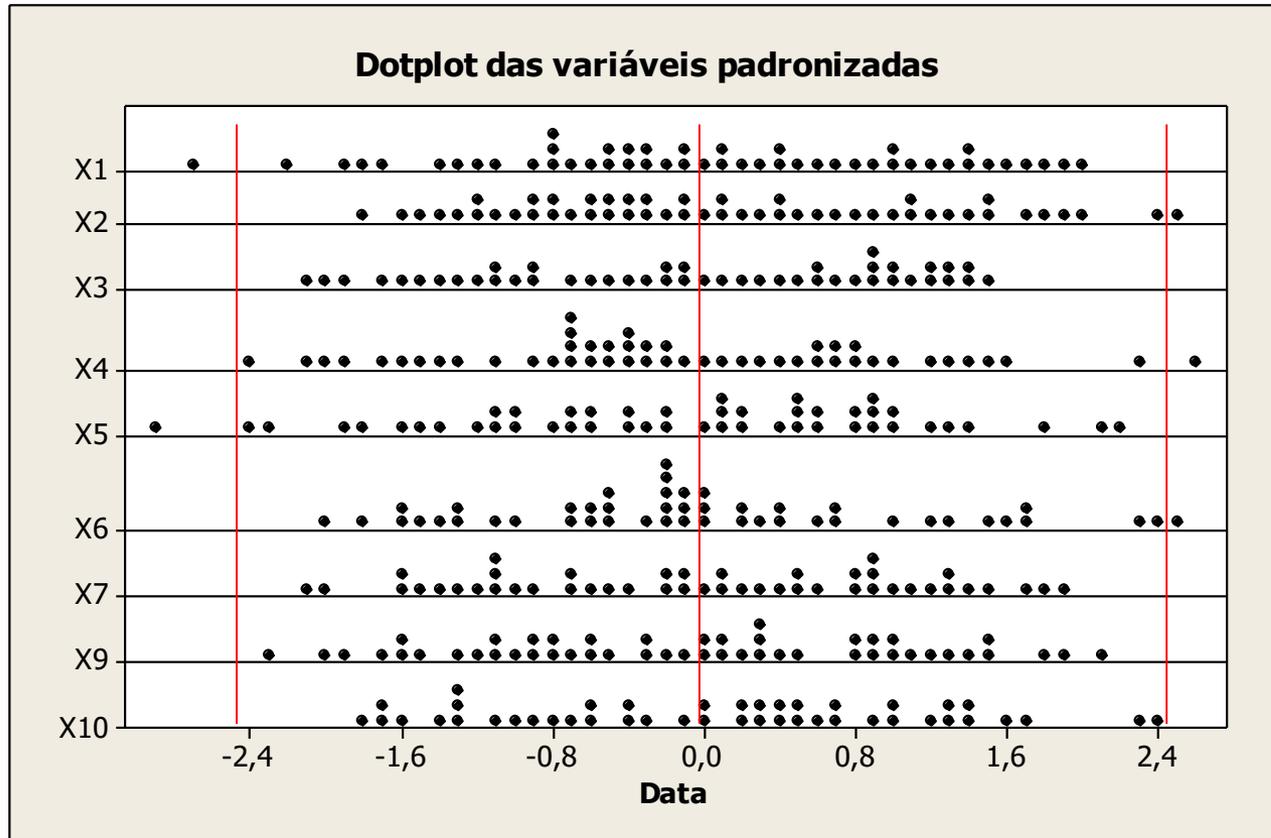
Escala X9: [0,100]

Escala X10: [0,10]

Demais variáveis:  
Características do  
Cliente (empresa  
compradora)

# Valor Atípico Unidimensional

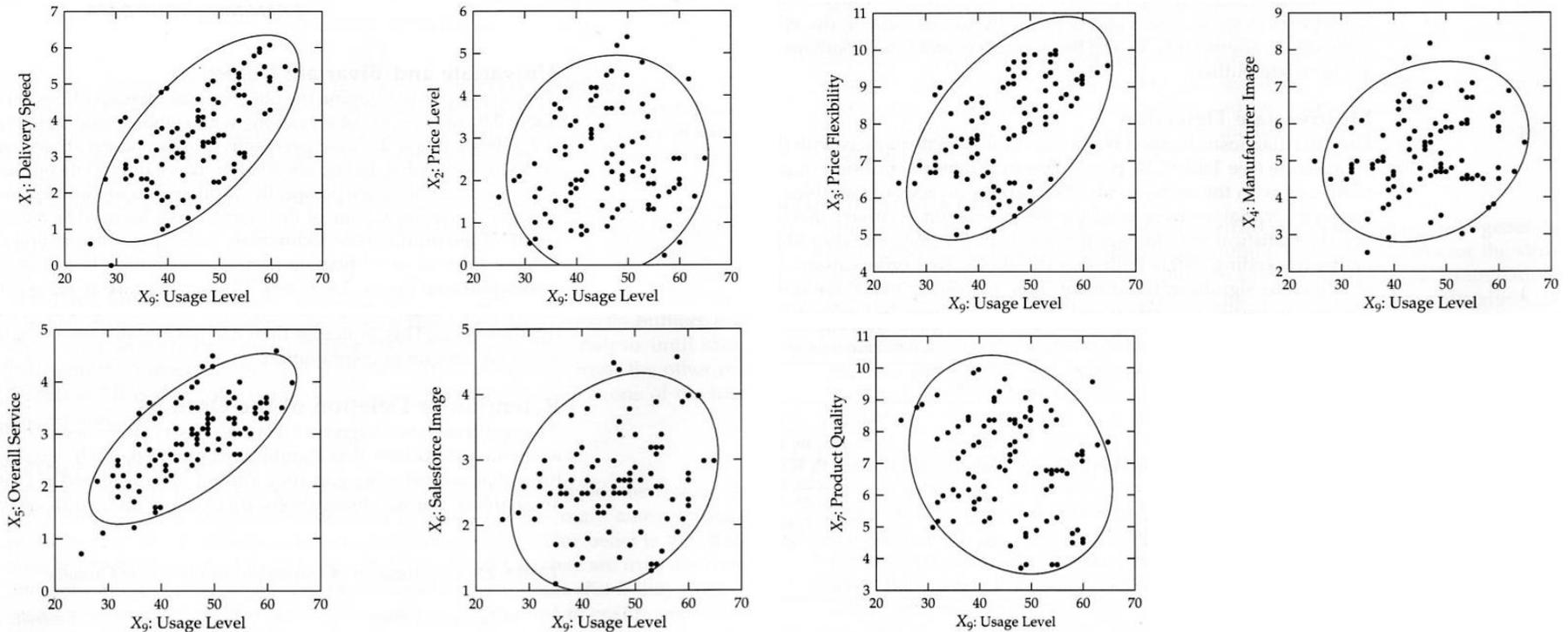
Variáveis padronizadas:  $Z_{i p \times 1} = D^{-1} \left( Y_{i p \times 1} - \bar{Y}_{p \times 1} \right)$



Casos atípicos: escores z excedendo  $\pm 2,5$

$$P\left(|Z_{ij}| \geq 2,5\right) = 0,012$$

# Valor Atípico Bidimensional



Elipse de concentração:  $(Y_i - \bar{Y})' S^{-1} (Y_i - \bar{Y}) \leq \chi_2^2(\alpha = 0,10) = 4,61$

A direção dos eixos da elipse é definida pelos autovetores e os raios são proporcionais aos autovalores .

# Valor Atípico Multidimensional

## Valores Aberrantes Univariados

Variável	Unidade Amostral
X1	39
X2	71
X3	-
X4	82
X5	96
X6	5, 42
X7	-
X9	-

$$Z_{i \times 1} = D_{s_{jj}}^{-1/2} (Y_{i \times 1} - \bar{Y}_{p \times 1})$$

$$P(|Z_{ij}| \geq 2,5) = 0,012$$

## Valores Aberrantes Bivariados

Variável	Unidade Amostral
X1	1, 3, 95, 96
X2	3, 49, 57, 71, 96, 97
X3	11, 57, 96, 100
X4	5, 22, 42, 50, 72, 82, 93, 96
X5	3, 22, 39, 57, 71, 96
X6	5, 7, 42, 82, 96
X7	57, 58, 95, 96

$$(Y_i - \bar{Y})' S^{-1} (Y_i - \bar{Y}) \leq \chi_2^2(\alpha = 0,10) = 4,61$$

# Valor Atípico Multidimensional

7 var. de percepção

$$\chi_7^2$$

$$(Y_i - \bar{Y})' S^{-1} (Y_i - \bar{Y}) \leq \chi_7^2(\alpha = 0,10) = 12,017$$

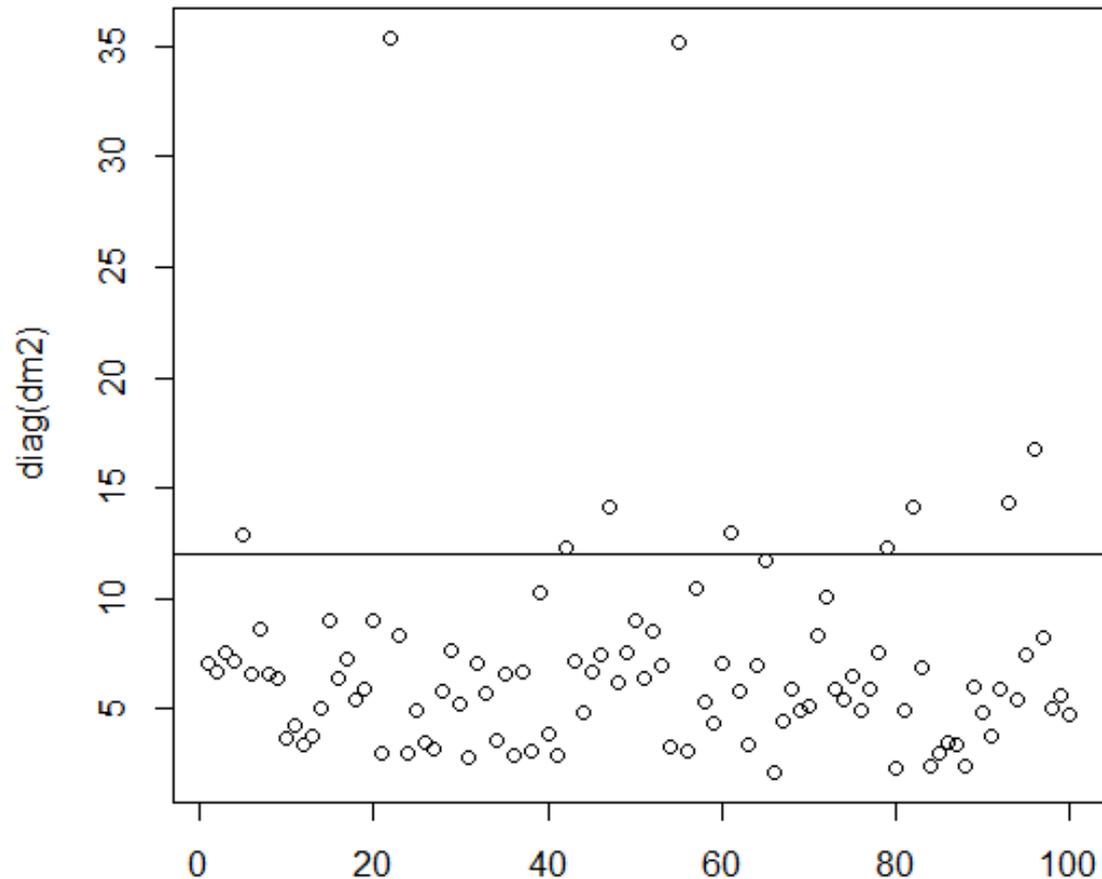
ID	DM2	Valor p
1	7,0313	0,425624
2	6,6914	0,461700
3	7,5671	0,372310
4	7,1030	0,418239
5	12,8700	0,075339
6	6,5166	0,480887
7	8,6345	0,279974
8	6,5626	0,475791
9	6,3753	0,496674
10	3,6258	0,821726
11	4,2366	0,752162
12	3,3892	0,846818
13	3,7682	0,806053
14	5,0297	0,656340
15	8,9619	0,255403
16	6,3977	0,494158
17	7,2120	0,407143
18	5,3501	0,617322
19	5,8987	0,551621
20	8,9619	0,255403
21	2,9782	0,887015
22	35,3897	0,000009
23	8,3329	0,304153
24	2,9743	0,887365
25	4,9087	0,671105

ID	DM2	Valor p
26	3,4629	0,839137
27	3,1712	0,868727
28	5,7654	0,567395
29	7,6008	0,369111
30	5,1885	0,636976
31	2,7506	0,907079
32	7,0239	0,426397
33	5,6784	0,577758
34	3,5292	0,832124
35	6,5393	0,478369
36	2,8996	0,894131
37	6,7041	0,460322
38	3,0296	0,882253
39	10,2131	0,176818
40	3,8267	0,799508
41	2,8975	0,894317
42	12,2819	0,091662
43	7,1289	0,415580
44	4,8189	0,682055
45	6,6704	0,463984
46	7,4751	0,381142
47	14,0940	0,049535
48	6,1518	0,522138
49	7,5605	0,372936
50	9,0287	0,250600

ID	DM2	Valor p
51	6,3615	0,498228
52	8,4673	0,293196
53	6,9126	0,438032
54	3,2441	0,861537
55	35,1975	0,000010
56	3,0818	0,877331
57	10,4877	0,162579
58	5,2655	0,627600
59	4,3477	0,738973
60	7,0123	0,427602
61	13,0008	0,072089
62	5,7983	0,563488
63	3,3222	0,853686
64	6,9256	0,436670
65	11,6830	0,111476
66	2,1085	0,953590
67	4,3815	0,734930
68	5,9247	0,548566
69	4,8776	0,674902
70	5,0570	0,653002
71	8,2943	0,307357
72	10,0950	0,183252
73	5,8874	0,552957
74	5,3634	0,615710
75	6,4714	0,485907

ID	DM2	Valor p
76	4,9253	0,669075
77	5,8474	0,557679
78	7,5221	0,376609
79	12,2787	0,091758
80	2,2699	0,943405
81	4,9429	0,666935
82	14,1176	0,049129
83	6,8374	0,446005
84	2,3659	0,936850
85	3,0164	0,883479
86	3,4932	0,835941
87	3,3539	0,850447
88	2,4175	0,933182
89	6,0111	0,538460
90	4,8599	0,677060
91	3,7626	0,806675
92	5,8410	0,558434
93	14,3281	0,045644
94	5,4068	0,610451
95	7,3910	0,389330
96	16,7080	0,019380
97	8,1952	0,315694
98	4,9898	0,661203
99	5,5865	0,588769
100	4,7045	0,695974

# Observação Atípica Multidimensional Diagnóstico – Distância de Mahalanobis



$$d_M^2 \leq \chi_7^2(\alpha = 0,10) = 12,017$$

# Tópicos Adicionais

Como medir “relação” entre variáveis ( $p$  vetores em  $\mathfrak{R}^n$ )?

Como medir “distância” entre observações ( $n$  vetores em  $\mathfrak{R}^p$ )?

- ✓ Correlação de Pearson: das variáveis originais ou padronizadas
- ✓ Distância Euclidiana, Distância de Pearson, Distância de Mahalanobis
- ✓ Matriz de Precisão
  
- Medidas de correlação robustas (cálculo de  $R$ ): Spearman, Kendall
- Correlações Parciais (Grafos Não direcionados)  $\Rightarrow$  Grafos Direcionados
- Matrizes de importância entre variáveis: AHP (Saaty, 1980)
- Dados composicionais: transformação de Aitchison (Aitchison, 2003)
- Matriz de dados heterogêneos (quantitativos e qualitativos): como combinar informação e propor medidas de distância entre observações? (medidas de distância ponderada – Johnson and Whichern, 2008)
- Dados incompletos: como calcular  $S$  e  $D$ ? (imputação, ponderação!)
- ...