# MAE 5776

# ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Semestre IME/2019

# Dados Multivariados

Banco de Dados:

| | Variáveis | | | | | |
|---|---|---|---|---|---|---|
| Unidades Amostrais | 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 | $Y_{n1}$ | $Y_{n2}$ | | $Y_{nj}$ | | $Y_{np}$ |

$Y_{n\times p} = \left( y_{ij} \right)$: 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.

Medidas biométricas (mm) de Pardais fêmea (Hermon Bumps, 1898).

Variável que define grupos

Variáveis quantitativas avaliadas nas u.a.

tamanhos amostrais

| 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 |

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

Discuta possíveis questões de interesse.

# MOTIVAÇÃO

Exemplos considerados em Manly, 2005.

Cães pré-históricos da Tailândia.

| Grupo | X1 | X2 | X3 | X4 | X5 | X6 |
|---|---|---|---|---|---|---|
| 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 está mais “próximo” do cão pré-histórico? E qual está mais “distante” ?

# MOTIVAÇÃO

Exemplos considerados em Manly, 2005.

Força de trabalho em países da Europa.

| País | Região | Agr | Min | Man | PS | Cons | Ser | Fin | SPS | TC |
|---|---|---|---|---|---|---|---|---|---|---|
| P1 | U.E. | 2.6 | 0.2 | 20.8 | 0.8 | 6.3 | 16.9 | 8.7 | 36.9 | 6.8 |
| … | … | | | | | | | | | |
| P11 | U.E. | | | | | | | | | |
| P12 | EFTA | 7.4 | 0.3 | 26.9 | 1.2 | 8.5 | 19.1 | 6.7 | 23.3 | 6.4 |
| … | … | | | | | | | | | |
| P17 | EFTA | | | | | | | | | |
| P18 | Leste | 55.5 | 19.4 | 0.0 | 0.0 | 3.4 | 3.3 | 15.3 | 0.0 | 3.0 |
| … | … | | | | | | | | | |
| P25 | Leste | | | | | | | | | |
| P26 | Outro | 13.5 | 0.3 | 19.0 | 0.5 | 9.1 | 23.7 | 6.7 | 21.2 | 6.0 |
| … | … | | | | | | | | | |
| P30 | Outro | 44.8 | 0.9 | 15.3 | 0.2 | 5.2 | 12.4 | 2.4 | 14.5 | 4.4 |

Discuta possíveis questões de interesse:

Como “discriminar” as regiões? Com base na força de trabalho de um novo país, como “classificá-lo”?

# MOTIVAÇÃO

Dados apresentados em Mardia et al. (2003)



*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?

# 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)

# Dados Multivariados

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

- Classificação das Variáveis:

| | | | |
|---|---|---|---|
| Qualitativas | Nominal | Quantitativas | Discreta |
| | Ordinal | | Contínua |

Análises clássicas

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

**?** 

"Big data": n << p

Observações correlacionadas (*clusters* – região, grau de parentesco): como calcular o tamanho amostral "efetivo"

# 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…*

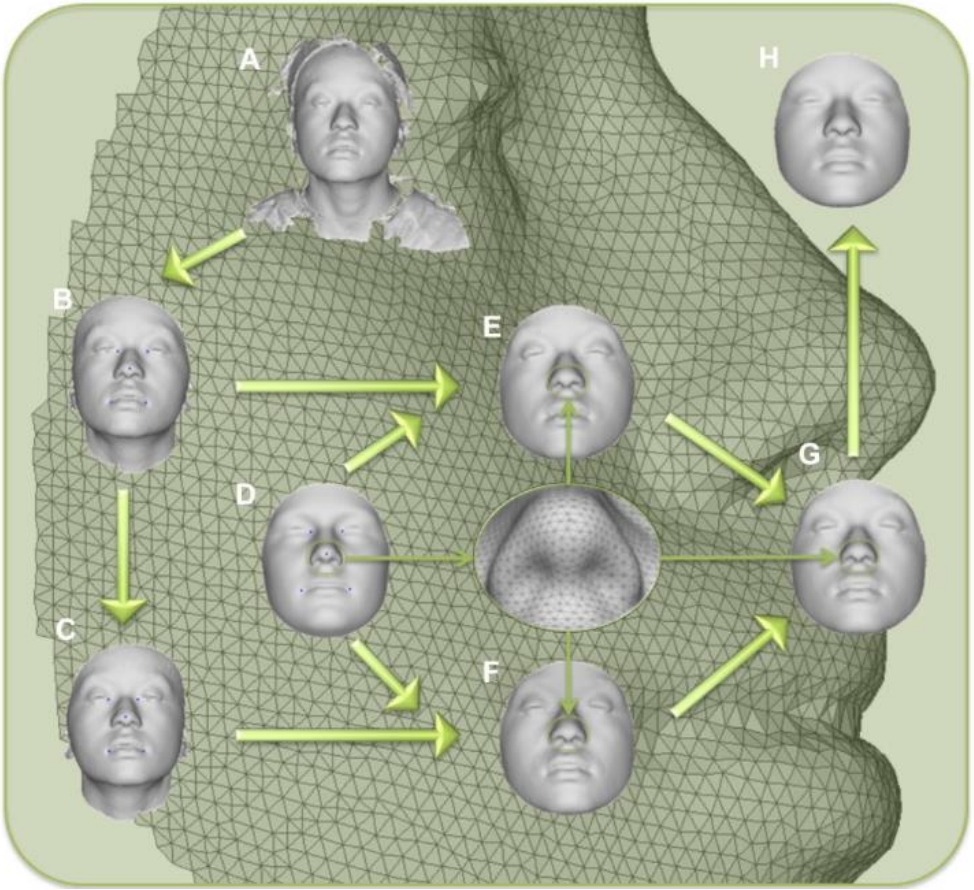
- Estudos do genoma de populações mundiais

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 (fontes dessa variabilidade) 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.

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

```
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

$$Y_{n\times p} = \begin{pmatrix} Y_1' \\ Y_2' \\ \dots \\ Y_n' \end{pmatrix} = \left(Y_{(1)}, Y_{(2)}, \dots, Y_{(p)}\right)$$

"Q-espaço"

Espaço dos indivíduos: n vetores em um espaço p-dimensional ($\Re^p$)

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

"R-espaço"

Espaço das variáveis: p vetores em um espaço n-dimensional ($\Re^n$)

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

Explorar as propriedades geométricas de espaços vetoriais

# Estatísticas Descritivas

Dados Multivariados Quantitativos:

Uma amostra aleatória de n observações de uma população multivariada ($\Re^p$)

$$Y_{n\times p}$$

Vetor de medias: $\bar{y}_{p\times 1} \quad ; \bar{y}' = \left(\bar{y}_1, \ldots, \bar{y}_p\right)$

Matriz de covariâncias: $S_{p\times p} = \left(s_{j\,j'}\right)$ "R-técnicas"

Matriz de correlações: $R_{p\times p} = \left(r_{j\,j'}\right)$ "R-técnicas"

$\bar{y}_j$ ? $\quad s_{j\,j'}$ ? $\quad r_{j\,j'}$ ? Conceito de produto interno!

# Estatísticas Descritivas

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

Média da variável j (escalar)

$1$ : vetor (nx1) de valores 1

$1'$ : vetor transposto

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

Covariância entre as variáveis j e j' (escalar)

*n-1*

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

Correlação entre as variáveis j e j' = Covariância entre as variáveis j e j' padronizadas

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

Centróide: vetor de médias das p variáveis

Matriz de covariâncias

Produto interno centrado

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

$$= \frac{1}{n}\left(Y'Y - \frac{1}{n}Y'1\,1'Y\right) = \frac{1}{n}\left[Y'\left(I_n - \frac{1}{n}11'\right)Y\right] = \frac{1}{n}Y'HY$$

HY: centralizar as linhas de Y

$$= \left(HY\right)' HY$$

$$H = I_n - \frac{1}{n}11' = I_n - n^{-1}JJ'$$

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

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

Produto interno ordinário
(soma de quadrados e produtos cruzados)

$$S_{p\times p} = \frac{1}{n} Y'HY = \begin{pmatrix} s_{11} & s_{12} & & s_{1p} \\ s_{21} & s_{22} & & s_{2p} \\ & & \cdots & \\ s_{p1} & s_{p2} & & s_{pp} \end{pmatrix}$$ ; H=H', H=H²

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

n>p: S é p.d.

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

- Matriz de correlações

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

- Matriz de covariâncias com denominador (n-1)

"no R": cov(Y)

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

- Decomposição Espectral (de matriz quadrada)

$$S = V\Lambda V'; \quad VV' = V'V = I$$

V: Matriz de autovetores (ortonormal)
$\Lambda$: Matriz diagonal de autovalores

$$\left|S - \lambda I_p\right| = 0$$ Obter os autovalores (raízes características)

$$SV_j = \lambda_j V_j$$ Obter os autovetores

Obter vetores reducionistas: max(trS)

- Variância total (traço da matriz de covariâncias)

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

- Variância generalizada (determinante da matriz de covariâncias)

E.M.V.S.

$$\left|S\right| = \left|\Lambda\right| = \prod_{j=1}^{p} \lambda_j \qquad \left|S\right| = \left(S_{11}S_{22}...S_{pp}\right)\left|R\right|$$

# Matriz de Precisão - Correlação Parcial

- Matriz de Covariâncias e Correlação entre Variáveis

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

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

Correlação (marginal) entre as variáveis Yj e Yk

Grafos Não-direcionados

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

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

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

Correlação parcial entre pares de variáveis dado as demais variáveis

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

No caso da $N_p$: correlação parcial nula $\Rightarrow$ independência aos pares $\Rightarrow$ independência total

# Correlação Parcial - Aprendizado de Estruturas

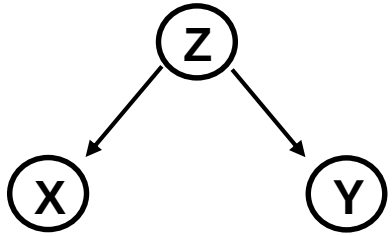
Ribeiro et al., 2016.

Padrões de dependência entre trios de variáveis

Grafos Direcionados

## Caso 1

Z é causa comum
Z é confundidor

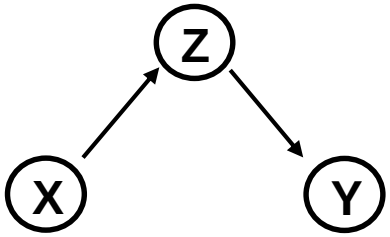



$(X \not\perp Y)$ correlacionados marginalmente

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

## Caso 2

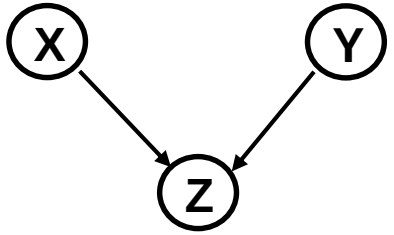
Z é efeito intermediário
X tem efeito indireto sobre Y




Grafos diferentes mas com Modelos Probabilísticos Equivalentes

## Caso 3

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




$(X \perp Y)$ independência marginal

$(X \not\perp Y \mid Z)$ dependência condicional

# Transformações Lineares de Variáveis

- Combinações Lineares

i-ésimo indivíduo

Combinação linear ou um contraste de interesse

- $Y_{i_{p\times 1}} \in \Re^p \Rightarrow X_{i_{1\times 1}} \in \Re;$ $\quad X_i = a'Y_i = \sum_{j=1}^{p} a_j Y_{ij}; \quad a \in \Re^p$

para n indivíduos

$$X_{n\times 1} = Y_{n\times p} a_{p\times 1}$$

$$\bar{x}_{1\times 1} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n} a'Y_i = a'\bar{Y}$$

$$s_{X_{1\times 1}} = \frac{1}{n}\sum a'\left(Y_i - \bar{Y}\right)\left(Y_i - \bar{Y}\right)' a = a' S_Y a$$

i-ésimo indivíduo

- $Y_{i_{p\times 1}} \in \Re^p \Rightarrow X_{i_{q\times 1}} \in \Re^q;$ $\quad X_i = A_{q\times p} Y_i + b_{q\times 1}; \quad A \in \Re^{q\times p}, b \in \Re^p$

para n indivíduos

$$X_{n\times q} = YA' + 1_n b'$$

$$\bar{X}_{q\times 1} = \bar{Y}A' + b$$

$$S_{X_{q\times q}} = AS_Y A';$$

# Transformações Lineares – Medidas de Distância

"Q-técnicas"

⇒ **Distância Euclidiana entre observações**

Matriz de distâncias

$$Y_{i_{p\times 1}} \in \Re^p \Rightarrow d_E \in \Re; \quad d^2_{E_{ik}} = d^2_{ik} = (Y_i - Y_k)'(Y_i - Y_k) \Rightarrow D_{E\,n\times n} = (d_{ik})$$

- Seja: $B_{n\times n} = (b_{ik}); \quad b_{ik} = (Y_i - \bar{Y})'(Y_k - \bar{Y});$

$$B_{n\times n} = HY(HY)' = HYY'H = H\,M\,H; \quad H = I_n - n^{-1}1_n1_n', \; M = \left(-\frac{1}{2}d^2_{ik}\right);$$

Então: $d^2_{ik} = b_{ii} + b_{kk} - 2b_{ik}$

Obter Coordenadas Principais (HY) a partir de distâncias entre objetos

$$b_{ik} = (Y_i - \bar{Y})'(Y_k - \bar{Y}) = -\frac{1}{2}\left(d^2_{ik} - d^2_{i.} + d^2_{.k} - 2d^2_{..}\right)$$

$$d^2_{i.} = \frac{1}{n}\sum_{j=1}^{n} d^2_{ij}; \quad d^2_{.j} = \frac{1}{n}\sum_{i=1}^{n} d^2_{ij}; \quad d^2_{..} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d^2_{ij}$$

# Transformações Lineares – Medidas de Distância

- Transformação de Escala (padronização)

"Q-espaço"

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

$$S_{Y^*} = \frac{1}{n}\sum_{i=1}^{n} Y_i^* Y_i^{*\prime} = \frac{1}{n} Y^{*\prime} Y^* = R_{Y^*} = R_Y$$

Matriz de covariâncias (ou de correlações) das variáveis padronizadas é a matriz de correlação das variáveis originais

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

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

**Distância de Pearson** ao quadrado (Euclidiana padronizada)

"Q-espaço"

- Transformação de Mahalanobis

$$Y_{i_{p\times 1}} \Rightarrow Z_{i_{p\times 1}} = S^{-1/2}\left(Y_i - \bar{Y}\right)$$

$$S_Z = I_p$$

Variáveis independentes e variâncias unitárias

$$Z_i^{\prime} Z_i = \left(Y_i - \bar{Y}\right)' S^{-1}\left(Y_i - \bar{Y}\right) = d_M^2(Y_i, C)$$

**Distância de Mahalanobis** ao quadrado

$$\left(Z_i - Z_k\right)'\left(Z_i - Z_k\right) = \left(Y_i - Y_k\right)' S^{-1}\left(Y_i - Y_k\right) = d_M^2(Y_i, Y_k)$$

Como obter Coordenadas Principais neste caso?
Ex. 1.6.1 (Mardia et al., 2003)

# Q-Espaço e R-Espaço

"Q-técnicas"

"R-técnicas"

- Distância entre Observações – Correlação entre Variáveis

$$Y_{i_{p\times 1}} \in \Re^p, \quad i = 1,\ldots,n \quad \Rightarrow \quad d^2_{E_{ik}} = \left(Y_i - Y_k\right)' \left(Y_i - Y_k\right)$$

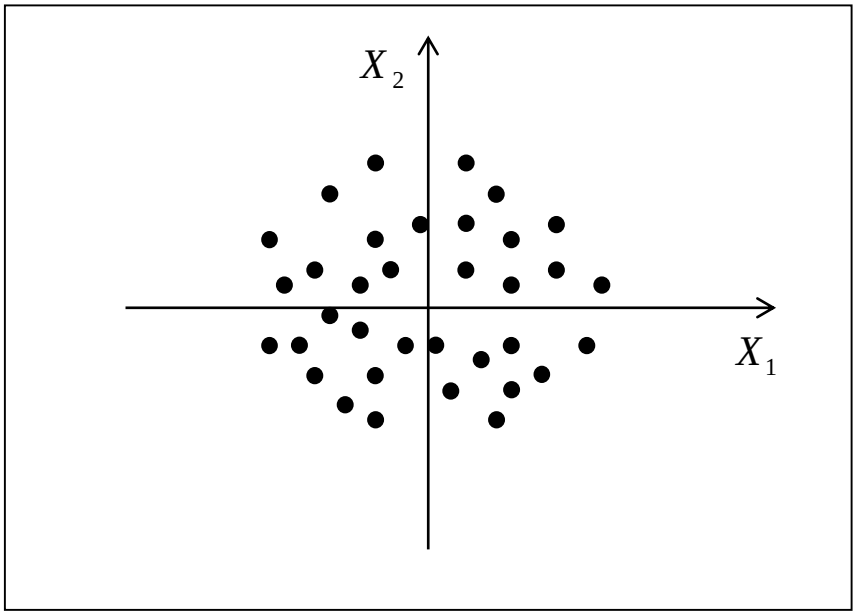
$$d^2_{P_{ik}} = \left(Y^*_i - Y^*_k\right)' \left(Y^*_i - Y^*_k\right) = \left(Y_i - Y_k\right)' D^{-1}_{s_{jj}} \left(Y_i - Y_k\right)$$

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

$$d^2_{P_{jl}} = \left(Y^*_{(j)} - Y^*_{(l)}\right)' \left(Y^*_{(j)} - Y^*_{(l)}\right) = 2\left(1 - r_{jl}\right)$$

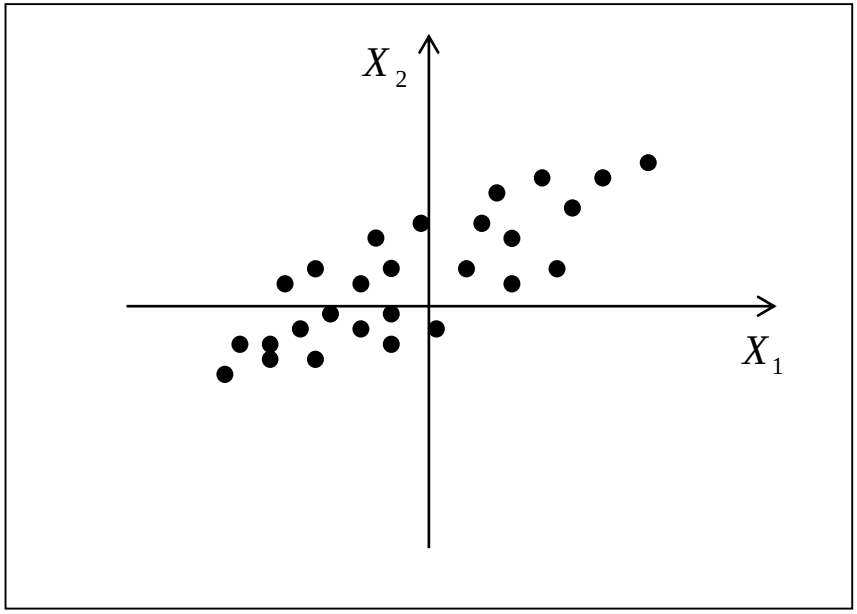
A "distância" de Pearson entre variáveis é medida por sua correlação (coeficiente de correlação linear de Pearson).

# Distância entre Observações




Variáveis independentes e variâncias homogêneas $\Rightarrow$ Distância Euclidiana

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




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

$$d_M^2(P,\mu) = \left(Y_{p\times 1} - \mu_{p\times 1}\right)' S^{-1} \left(Y_{p\times 1} - \mu_{p\times 1}\right)$$

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

$$d^2(P,\mu) = \left(Y_{p\times 1} - \mu_{p\times 1}\right)' D_{s_{jj}}^{-1} \left(Y_{p\times 1} - \mu_{p\times 1}\right)$$

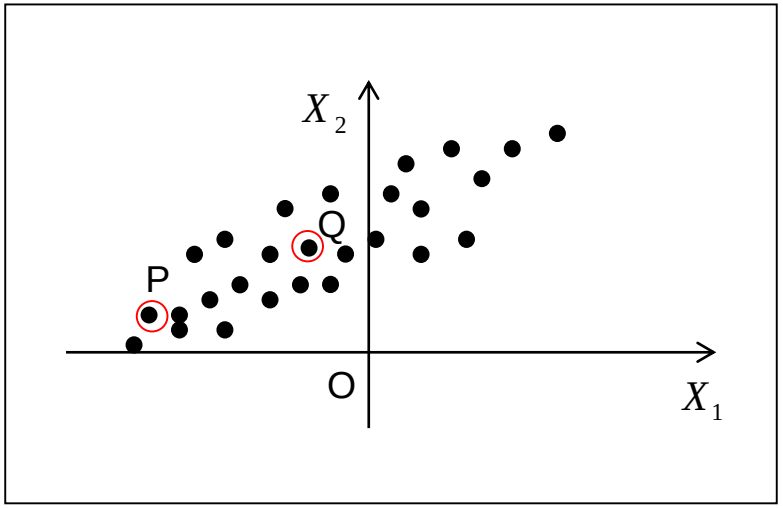
# Distância Estatística

⇒ 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 X1 e X2), 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.

# 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.

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

$$\bar{Y}_{4\times 1} =$$

| N | E | S | W |
|---|---|---|---|
| 50.53571 | 46.17857 | 49.67857 | 45.17857 |

divisor=n-1

$$S_{4\times 4} = \begin{pmatrix} s_{11} & s_{12} & s_{13} & s_{14} \\ & s_{22} & s_{23} & s_{24} \\ & & s_{33} & s_{34} \\ & & & s_{44} \end{pmatrix} =$$

| | 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 |

Há indicação de heterocedasticidade?

$$R_{4\times 4} = \left(r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}\right) =$$

| | 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 |

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

# Estruturas de Correlação

| Principais estruturas de correlação | Definição | |
|---|---|---|
| Independente | $\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$ | $R = I_p$ |
| Permutável<br>Equicorrelação, uniforme | $\begin{pmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{pmatrix}$ | $R = (1-\alpha)I_p + \alpha 1_p 1_p'$ |
| 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 | $\begin{pmatrix} 1 & a & \cdots & a^{t-1} \\ a & 1 & \cdots & a^{t-2} \\ \vdots & \vdots & \ddots & \vdots \\ a^{t-1} & a^{t-2} & \cdots & 1 \end{pmatrix}$ | |

# Estatísticas Descritivas Multivariadas

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

$$Y_{28\times 4} = \left(Y_{(N)}, Y_{(E)}, Y_{(S)}, Y_{(W)}\right) \qquad Y \in \mathbb{R}^4 \rightarrow X \in \mathbb{R}^3$$

$$X = YA'; \qquad A = \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}$$

Matriz de contrastes

$$X_1 = (N+S) - (E+W)$$
$$X_2 = N - S$$
$$X_3 = E - W$$

$$\bar{X}_{3\times 1} = \left(8{,}857 \quad 0{,}857 \quad 1{,}000\right)' \qquad S_{3\times 3} = \begin{pmatrix} 124{,}12 & -20{,}27 & -25{,}96 \\ & 61{,}27 & 26{,}96 \\ & & 99{,}50 \end{pmatrix}$$

Padronize as variáveis e *recalcule* o centróide e a matriz de covariâncias de X. Na transformação considere a soma de quadrados dos coeficientes igual a 1.

# 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)*

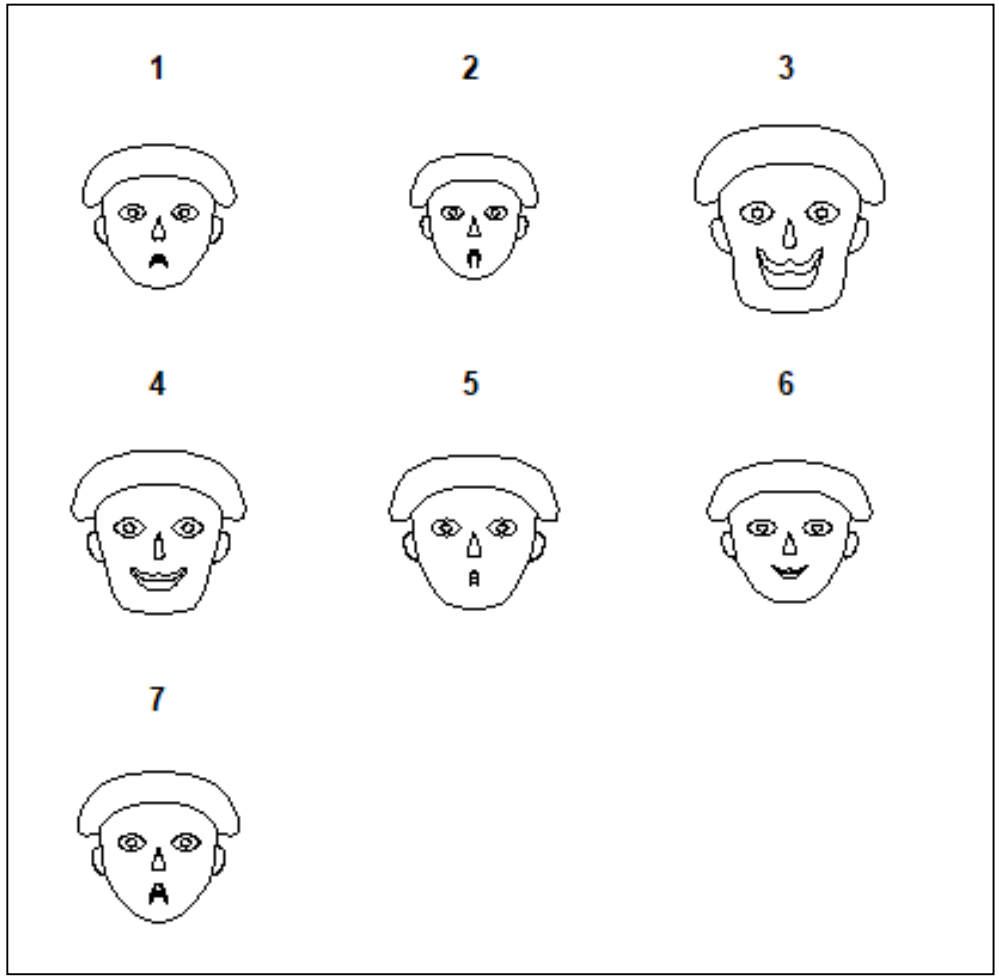
Os cães são mais heterogêneos relativamente a qual variável?

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

> 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

> dm1my<-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
```

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

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




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

# Correlação Parcial

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

```
> 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
```

```
> 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 -0.1770
S 0.2683  0.2067 1.0000  0.6330
W 0.3384 -0.1770 0.6330  1.0000
```

Correlação parcial entre pares de variáveis dado as demais.

Testar a significância das *pcor*!

# 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 0.0000 0.0014 0.1850 0.0909
E 0.0014 0.0000 0.3111 0.3870
S 0.1850 0.3111 0.0000 0.0005
W 0.0909 0.3870 0.0005 0.0000
```

Teste de correlação parcial nula (Teste z de Fisher)

Correlações Parciais:

dependências condicionais

S — W — N — E

# Correlação Parcial

Exemplo: Dados dos Cães - “Aprendizado” da relação entre variáveis

```
> 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
```

Correlação (marginal) entre pares de variáveis

```
> 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
```

```
> library(ppcor)
> pc<-pcor(x=caes)
> round(pc$estimate,4)
       X1      X2      X3      X4      X5      X6
X1  1.0000  0.8102  0.4772  0.7061  0.7124 -0.9128
X2  0.8102  1.0000 -0.7213 -0.1992 -0.8415  0.8920
X3  0.4772 -0.7213  1.0000  0.1973 -0.5309  0.6757
X4  0.7061 -0.1992  0.1973  1.0000 -0.2226  0.4754
X5  0.7124 -0.8415 -0.5309 -0.2226  1.0000  0.8364
X6 -0.9128  0.8920  0.6757  0.4754  0.8364  1.0000
```

Correlação parcial entre pares de variáveis dado as demais.

Testar a significância das *pcor*!

# Correlação Parcial

Exemplo: Dados dos Cães Pré-históricos - “Aprendizado” da relação entre variáveis

```
> pc<-pcor(x=caes)
> round(pc$estimate,2)
      X1    X2    X3    X4    X5    X6
X1  1.00  0.81  0.48  0.71  0.71 -0.91
X2  0.81  1.00 -0.72 -0.20 -0.84  0.89
X3  0.48 -0.72  1.00  0.20 -0.53  0.68
X4  0.71 -0.20  0.20  1.00 -0.22  0.48
X5  0.71 -0.84 -0.53 -0.22  1.00  0.84
X6 -0.91  0.89  0.68  0.48  0.84  1.00
```

```
> pc<-pcor(x=caes)
> round(pc$p.value,3)
      X1    X2    X3    X4    X5    X6
X1 0.000 0.399 0.683 0.501 0.495 0.268
X2 0.399 0.000 0.487 0.872 0.363 0.299
X3 0.683 0.487 0.000 0.874 0.644 0.528
X4 0.501 0.872 0.874 0.000 0.857 0.685
X5 0.495 0.363 0.644 0.857 0.000 0.369
X6 0.268 0.299 0.528 0.685 0.369 0.000
```

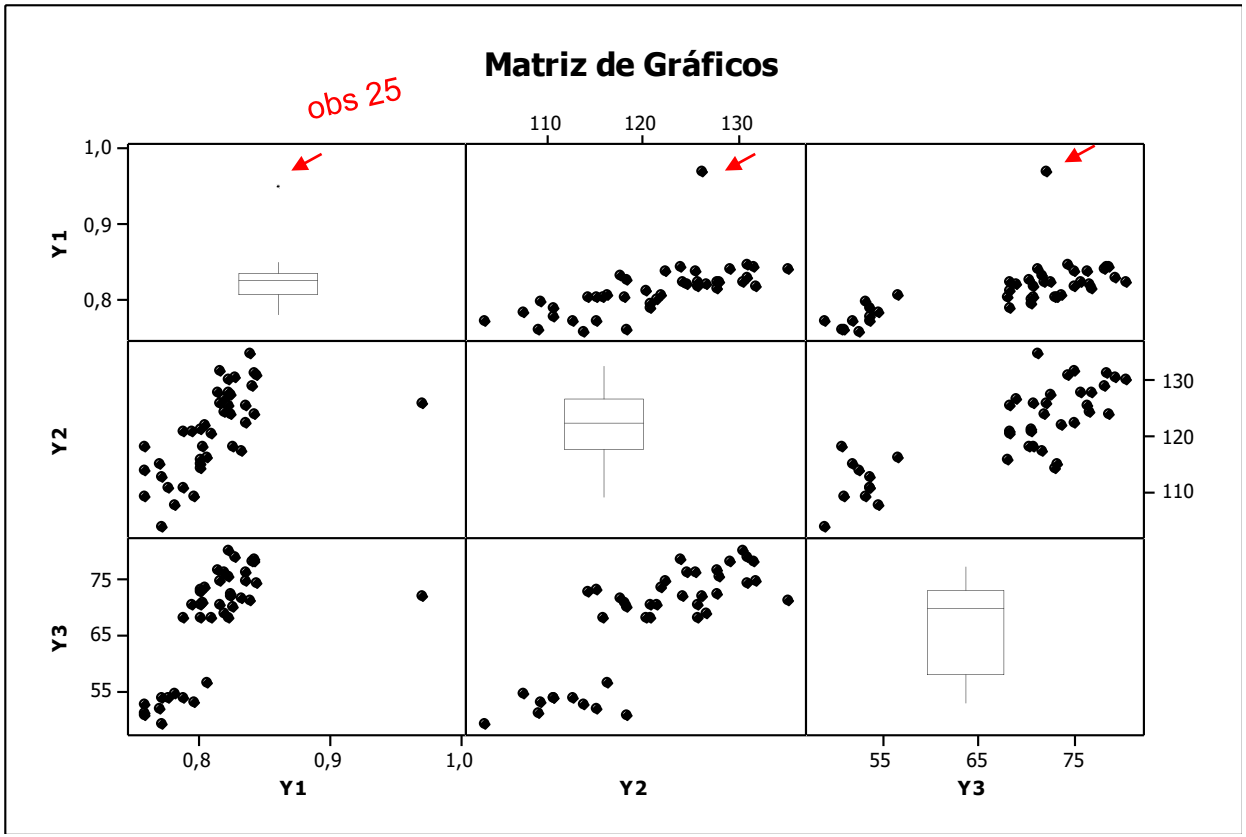
Independência condicional

Correlações Parciais:

X1 X2 X3 X4 X5 X6

Caso 1 ou 2: variáveis correlacionadas marginalmente MAS independentes condicionalmente

# Observações Atípicas – Diagnóstico Gráfico




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

Johnson and Wichern, 2002

⇒ 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)

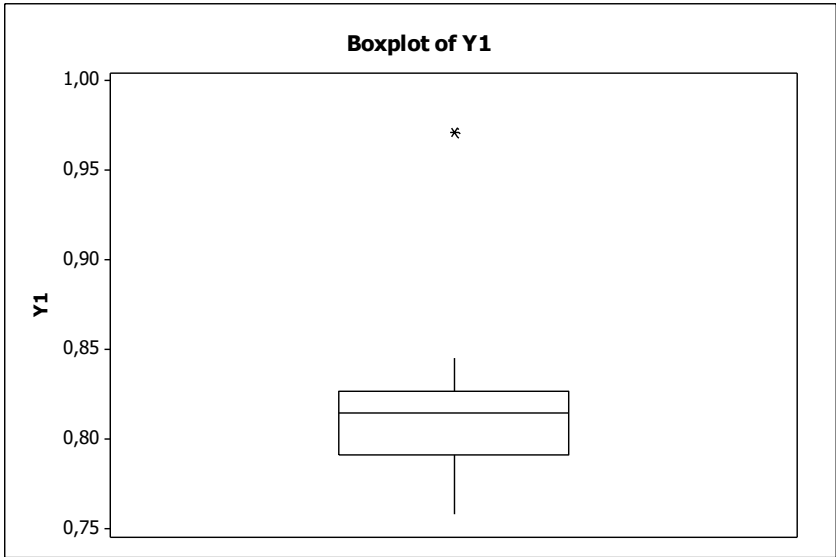
# Valor Atípico Unidimensional

$$Y = (Y1,\ldots,Y_n); \quad Y_i \in \Re, \quad i = 1,2,\ldots,n$$

- **Critério do Boxplot**:

$Q_3 + 1.5(Q_3 - Q_1)$

$Q_1 - 1.5(Q_3 - Q_1)$

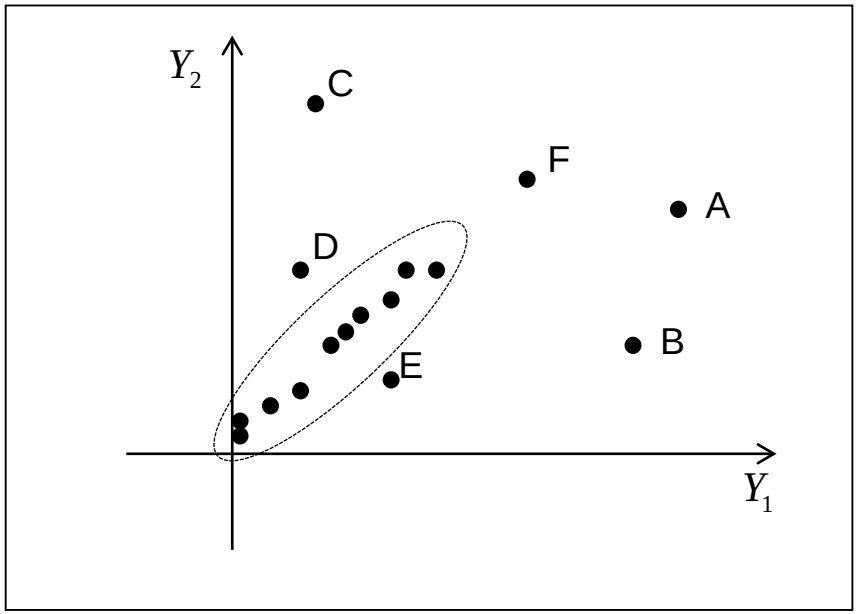



- Dados Padronizados: $Z_i = \dfrac{Y_i - \bar{Y}}{s}$ $P(|Z| \geq 2{,}5) = 0{,}012$ (Hair et al., 1998)

- Medida de Distância: $d^2 = \dfrac{(Y_i - \bar{Y})^2}{s^2} \overset{(n-p)\to\infty}{\sim} \chi^2_1 \Rightarrow P(d^2 \leq c^2) \leq (1-\alpha)$

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

# Valor Atípico Bidimensional




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

- na média e na variância de Y1 e Y2?
- na correlação entre Y1 e Y2?

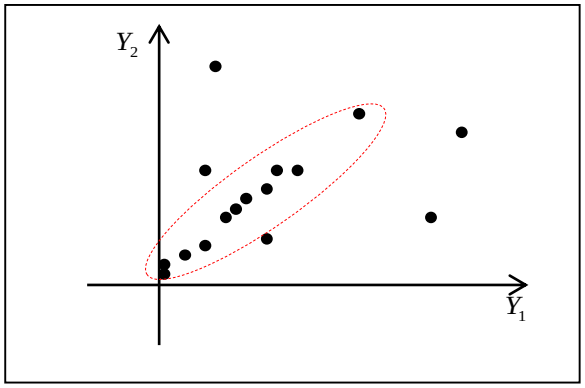
$\Rightarrow$ Ponto A: aberrante tanto para Y1 como para Y2

$\Rightarrow$ Pontos B (C) é aberrante para Y1 (Y2) mas não para Y2 (Y1)

$\Rightarrow$ Pontos D e E: são aberrantes bidimensionais mas não unidimensionais

$\Rightarrow$ Ponto F: apesar de aberrante unidimensional (para Y1 e Y2) 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 centro,

$$d_E^2\left(Y_i;C\right)=\left(Y_{i\,2\times1}-\overline{Y}_{2\times1}\right)'\left(Y_{i\,2\times1}-\overline{Y}_{2\times1}\right)$$

não leva em conta a correlação e há dificuldade em estabelecer um critério de diagnóstico.

## Distância de Mahalanobis

S: matrix de covariância

$$d_M^2\left(Y_i;C\right)=\left(Y_{i\,2\times1}-\overline{Y}_{2\times1}\right)'S^{-1}\left(Y_{i\,2\times1}-\overline{Y}_{2\times1}\right)$$

Diagnóstico: $d_M^2 \overset{n\to\infty}{\sim} \chi_2^2 \Rightarrow P\left(d_M^2 \le c^2\right)\le(1-\alpha)$

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

Elipse de concentração de pontos bidimensionais, centrada na média. Os eixos principais (direção e comprimento) 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:

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

$$d_M^2 \overset{n\to\infty}{\sim} \chi_p^2 \quad \Rightarrow \quad P\left(d_M^2 \le c^2\right) \le (1-\alpha)$$

$$\Rightarrow \quad \frac{d_M^2}{p} \overset{(n-p)\to\infty}{\sim} t_p^2$$

critério assintótico e sob normalidade

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

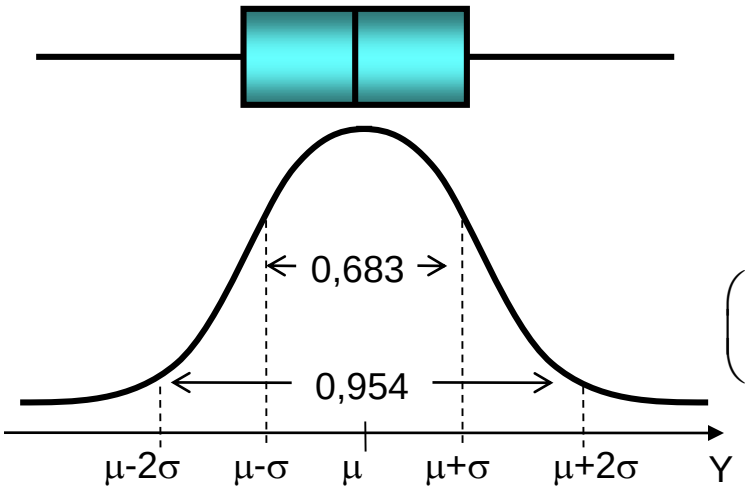
Os elipsóides têm eixos $\pm c\sqrt{\lambda_j}\, V_{kj}$ , em que $|S - \lambda I_p| = 0, \;\; SV_j = \lambda_j V_j, \;\; j = 1,\ldots p.$

No elipsóide sólido, os valores $\mathbf{y} \in \Re^p$, satisfazendo

$$\left(\mathbf{y} - \bar{Y}\right)' S^{-1} \left(\mathbf{y} - \bar{Y}\right) \le \chi_p^2(\alpha)$$

têm probabilidade 1-$\alpha$ de ocorrerem.

# Distribuição Normal Multivariada




$$Y_i \sim N_1\left(\mu, \sigma^2\right)$$ Normal Univariada

$$\Rightarrow f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left[(y_i-\mu)/\sigma\right]^2/2} \quad I_{(-\infty,\infty)}(y_i)$$

$\left(\frac{y_i-\mu}{\sigma}\right)^2 = (y_i-\mu)\sigma^{-2}(y_i-\mu)$ mede a distância ao quadrado de y a μ em unidades de desvios padrão

Normal p-Variada

$$Y_{i\ p\times 1} \sim N_p\left(\mu_{p\times 1}; \Sigma_{p\times p}\right)$$

**Σ** : matriz simétrica positiva-definida

$$\Rightarrow f(y_i) = \frac{1}{(2\pi)^{p/2}\left|\Sigma\right|^{1/2}} e^{-(y_i-\mu)'\Sigma^{-1}(y_i-\mu)/2} \quad I_{(-\infty,\infty)}(y_{ij});\ j = 1,\ldots,p$$

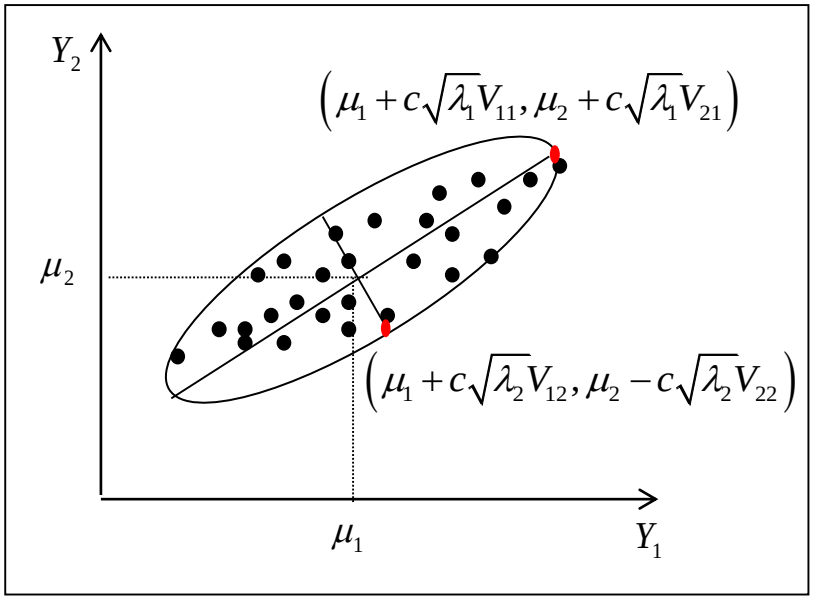
$\left\{d_M^2(Y_i;\mu) = (Y_i-\mu)'\Sigma^{-1}(Y_i-\mu) = c^2\right\}$ : a densidade é constante onde a distância ao quadrado é constante ⇒ elipsóide centrado em **μ**

# Elipse de Concentração de Observações

Exemplo da Normal bivariada:

$$\mathbf{Y}_{2\times1} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N_2\left(\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \boldsymbol{\Sigma}_{2\times2} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}\right); \qquad \sigma_{11} = \sigma_{22}$$

Elipse de Concentração de observações




$$d_M^2 \sim \chi_2^2 \quad \Rightarrow \quad P\left(d_M^2 \le c^2\right) \le (1-\alpha)$$

Obter *c* para a inclusão de 90%, 95% e 98% dos pontos amostrais.

Os eixos da elipse de concentração são calculados pela decomposição espectral de $\Sigma$:

$$\left|\Sigma - \lambda I_2\right| = 0 \qquad \text{autovalores}$$

$$\Rightarrow \quad \lambda_1 = \sigma_{11} + \sigma_{12} \qquad \lambda_2 = \sigma_{11} - \sigma_{12}$$

$$\Sigma V_j = \lambda_j V_j \qquad \text{autovetores}$$

$$\Rightarrow \quad V_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \qquad V_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$$

$$\Rightarrow \quad V_j' V_j = 1 \qquad V_j' V_{j'} = 0$$

# Elipsoide de Concentração de Observações

Distribuição Normal Multivariada:

$$Y_{i\,p\times 1} \sim N_p\left(\mu_{p\times 1};\Sigma_{p\times p}\right); \qquad f_{Y_i}(y)=\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}e^{-(y-\mu)'\Sigma^{-1}(y-\mu)/2} \qquad |\Sigma|>0,\ y\in\Re^p\ ,i=1,\ldots,n$$

$$d_M^2=(Y_i-\mu)'\Sigma^{-1}(Y_i-\mu)=c^2$$

$\Rightarrow$ elipsoide centrado em $\mu$: a densidade é constante onde a distância ao quadrado é constante

$$d_M^2=(Y_i-\mu)'\Sigma^{-1}(Y_i-\mu)=(Y_i-\mu)'\left[\sum_{j=1}^{p}\lambda_j^{-1}V_jV_j'\right](Y_i-\mu)=\sum_{j=1}^{p}\left(\frac{V_j'(Y_i-\mu)}{\lambda_j}\right)^2\sim\chi_p^2$$

$$\Sigma=V\Lambda V'=\sum_{j=1}^{p}\lambda_jV_jV_j'; \quad \Sigma^{-1}=V\Lambda^{-1}V'=\sum_{j=1}^{p}\lambda_j^{-1}V_jV_j' \qquad iid\ N_1(0;1)$$

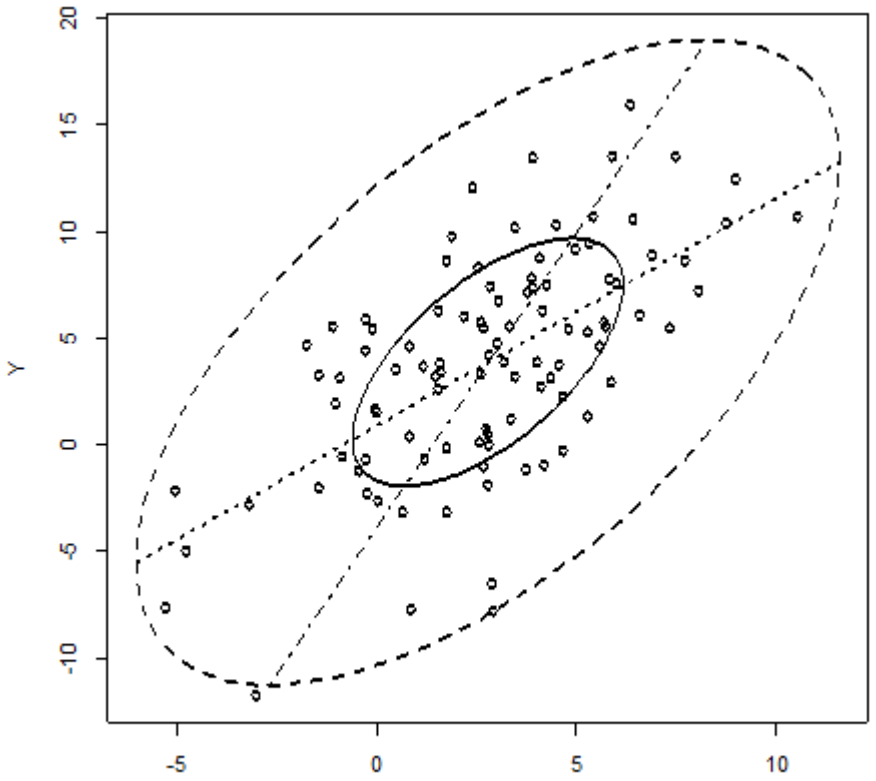
$$P\left(Y_i\in\Re^p;\ d_M^2(Y_i;\mu)\le c^2\right)=1-\alpha; \qquad c^2=\chi_p^2(\alpha)$$

Percentil (1-$\alpha$)100% da Qui-Quadrado

Elipsoide de Concentração para os dados $Y_i$ $\Rightarrow$ pode ser usado como um critério de diagnóstico de pontos *outliers*

# BoxPlot Bivariado
# Elipse de Concentração de Observações

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




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

Alternativas de critérios robustos de diagnóstico (Everitt, 2007)

Boxplot Bivariado (Everitt, 2007)
A elipse no centro inclui 50% dos dados.

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

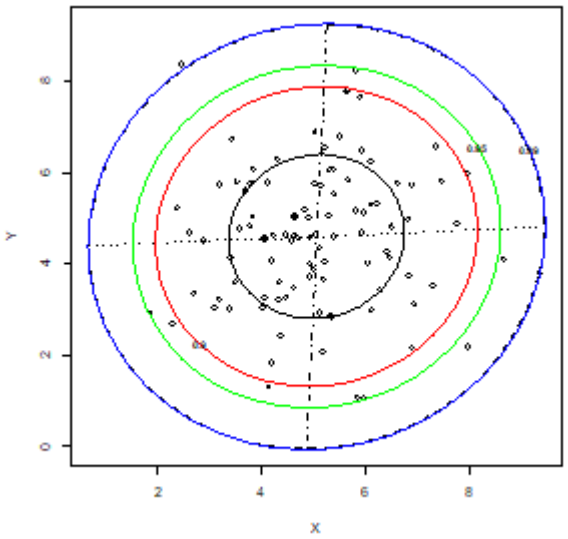
São apresentadas retas de regressão (de y vs. x e de x vs. y) com o estimador do centróide na intersecção.

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

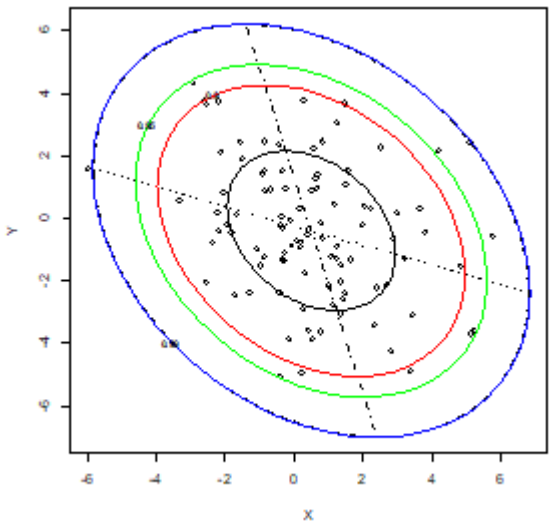
Note que na normal bivariada ambos, Y e X, são variáveis aleatórias. A construção das retas de regressão é um “abuso” justificado pela intenção de ilustrar as direções de eixos de variação dos dados.

Função bivbox (Everitt, 2005)

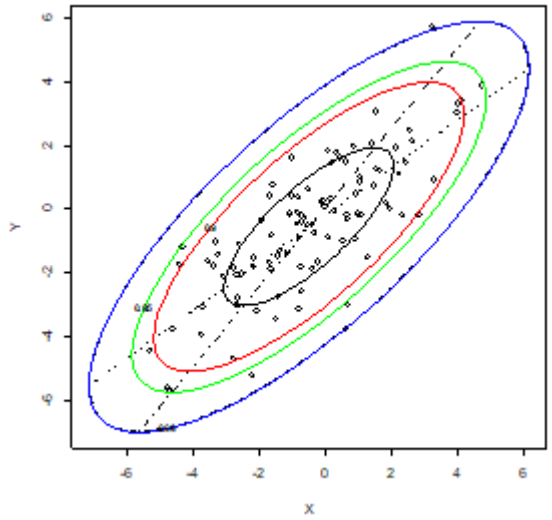
$$\mu' = (5,5) \quad \Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$



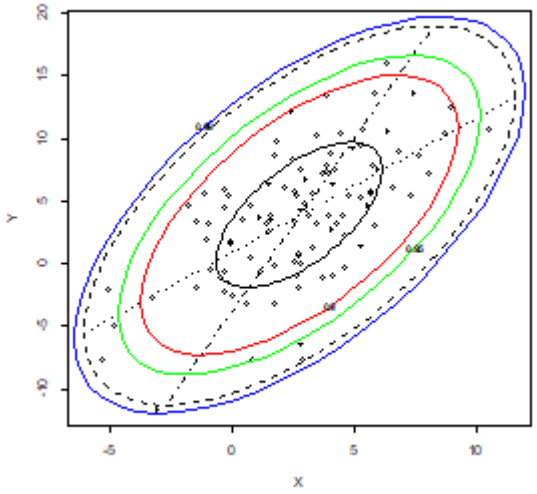

$$\mu' = (0,0) \quad \Sigma = \begin{pmatrix} 4 & -1 \\ -1 & 4 \end{pmatrix}$$




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




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



# 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

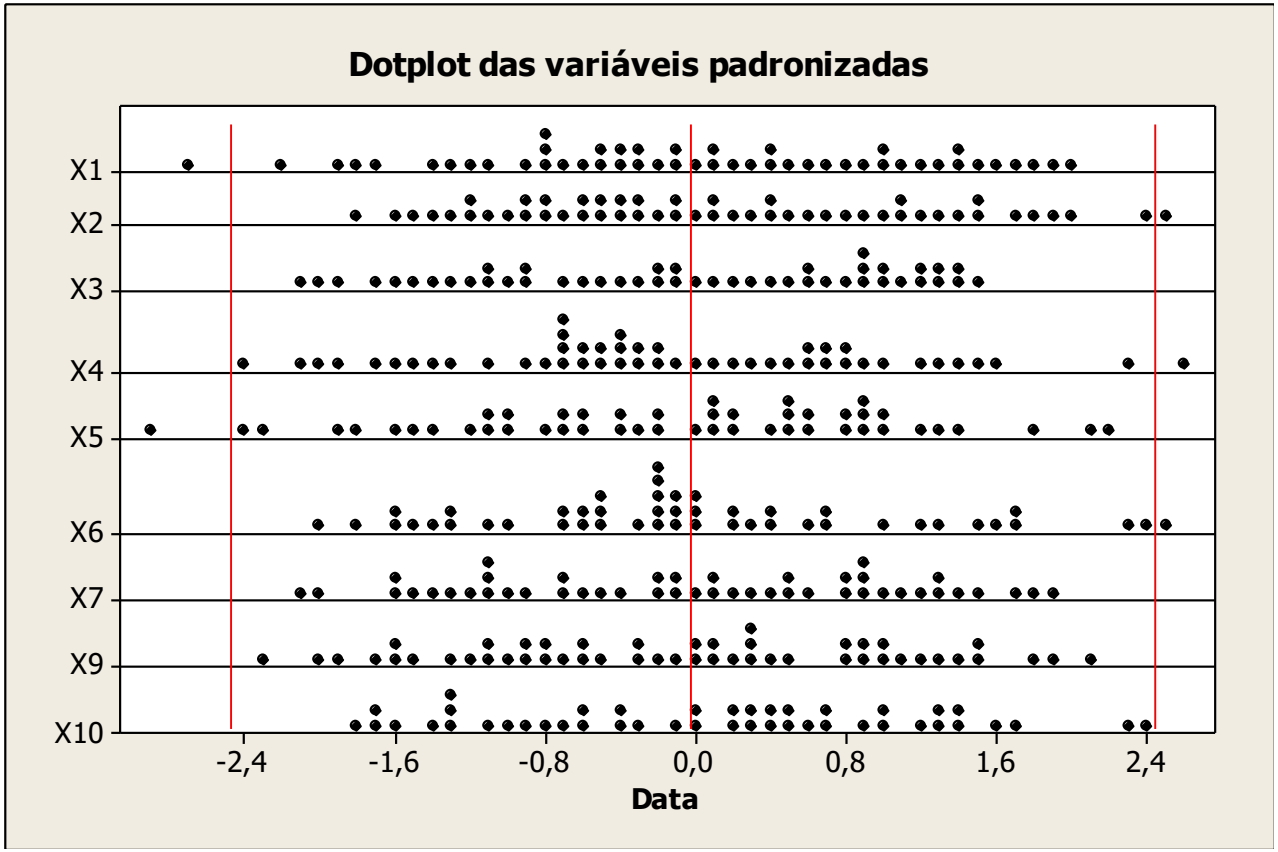
Escala X9: [0,100]

Escala X10: [0,10]

Demais variáveis: Características do Cliente (empresa compradora)

# Valor Atípico Unidimensional

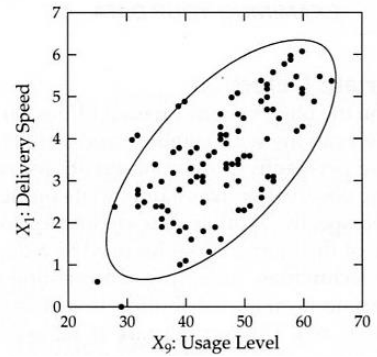
Variáveis padronizadas: $Z_{i\,p\times 1} = D^{-1}_{\sqrt{s_{jj}}}\left(Y_{i\,p\times 1} - \bar{Y}_{p\times 1}\right)$

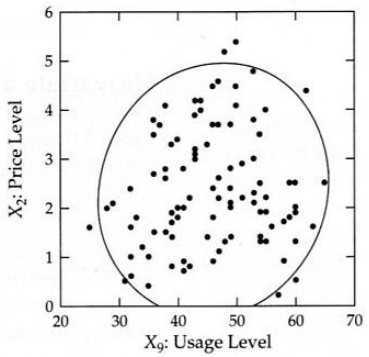


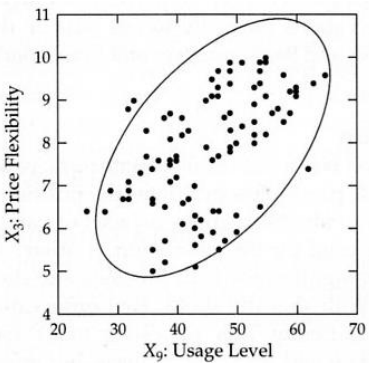

Casos atípicos: escores z excedendo $\pm$ 2,5

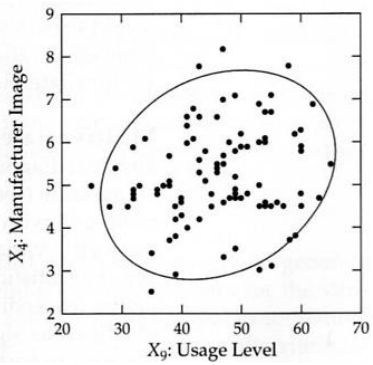
$$P\left(\left|Z_{ij}\right| \geq 2{,}5\right) = 0{,}012$$

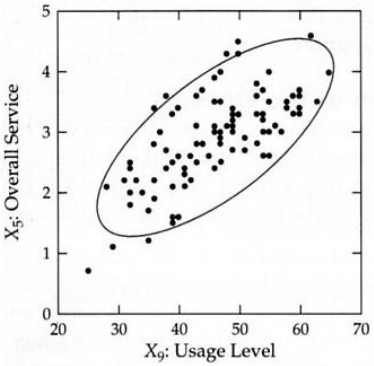
# Valor Atípico Bidimensional

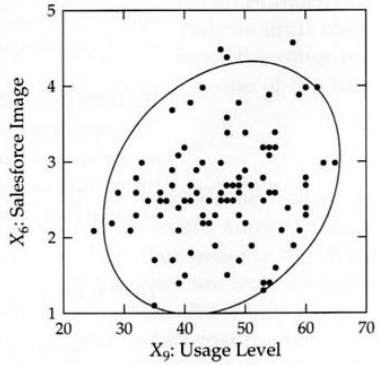


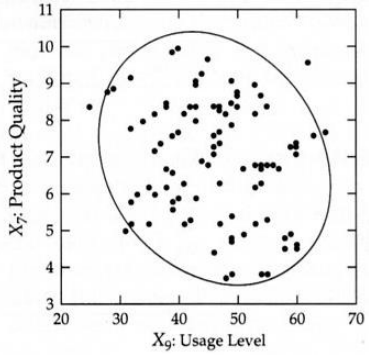



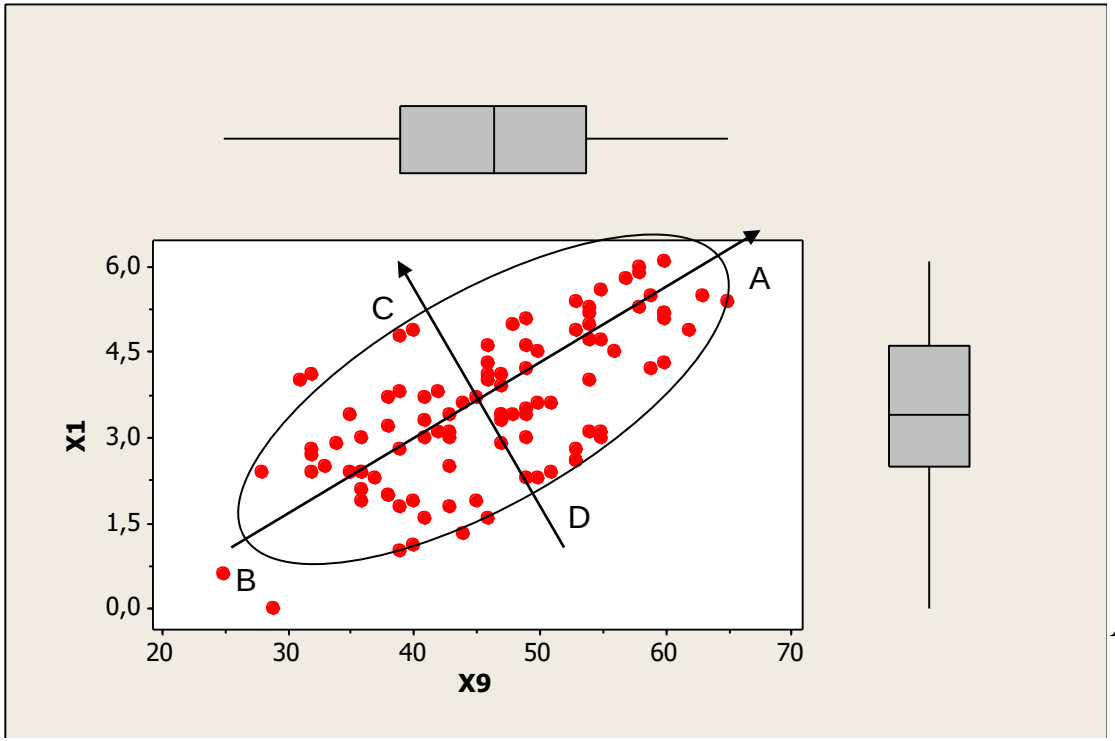




Elipse de concentração: $\left(Y_i - \bar{Y}\right)' S^{-1}\left(Y_i - \bar{Y}\right) \leq \chi_2^2(\alpha = 0{,}10) = 4{,}61$

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

# Elipse de Concentração de Pontos




$$\bar{Y}_{2\times1} = \begin{pmatrix} \bar{Y}_9 = 46{,}1 \\ \bar{Y}_1 = 3{,}5 \end{pmatrix}$$

$$\mathbf{S}_{2\times2} = \begin{pmatrix} s_9^2 = 80{,}80 & s_{91} = 8{,}08 \\ s_{91} = 8{,}08 & s_1^2 = 1{,}74 \end{pmatrix}$$

Decomposição Espectral de S:

$$V'SV = \Lambda$$

$$|S - \lambda I| = 0 \quad \Rightarrow \quad SV_j = \lambda_j V_j$$

$$\Lambda = \begin{pmatrix} 81{,}61 & 0 \\ 0 & 0{,}94 \end{pmatrix} \quad V = \begin{pmatrix} 0{,}99 & -0{,}10 \\ 0{,}10 & 0{,}99 \end{pmatrix}$$

$$A = \left(\bar{Y}_9 + \sqrt{4{,}6}\sqrt{81{,}6}\ \ 0{,}99\ ;\ \bar{Y}_1 + \sqrt{4{,}6}\sqrt{81{,}6}\ \ 0{,}10\right)$$

$$B = \left(\bar{Y}_9 - \sqrt{4{,}6}\sqrt{81{,}6}\ \ 0{,}99\ ;\ \bar{Y}_1 - \sqrt{4{,}6}\sqrt{81{,}6}\ \ 0{,}10\right)$$

$$C = \left(\bar{Y}_9 + \sqrt{4{,}6}\sqrt{0{,}94}\ \ (-0{,}10)\ ;\ \bar{Y}_1 + \sqrt{4{,}6}\sqrt{0{,}94}\ \ 0{,}99\right)$$

$$D = \left(\bar{Y}_9 - \sqrt{4{,}6}\sqrt{0{,}94}\ \ (-0{,}10)\ ;\ \bar{Y}_1 - \sqrt{4{,}6}\sqrt{0{,}94}\ \ 0{,}99\right)$$

# Valor Atípico Multidimensional

Valores Aberrantes Univariados

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

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 |

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

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

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

# Valor Atípico Multidimensional

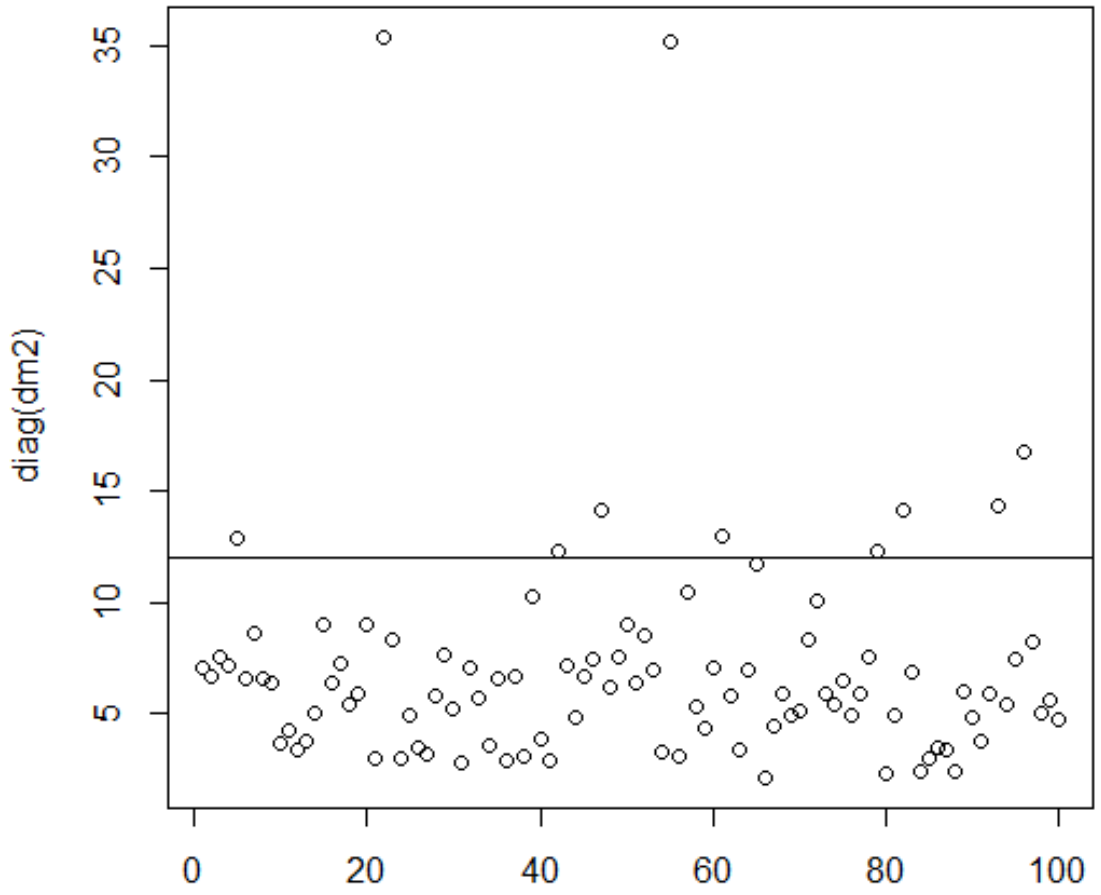
7 var. de percepção

$\chi^2_7$

$$\left(Y_i - \bar{Y}\right)' S^{-1} \left(Y_i - \bar{Y}\right) \leq \chi^2_7 \left(\alpha = 0{,}10\right) = 12{,}017$$

| 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 |
| 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 |
| 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 |
| 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 $\Re^n$)?

Como medir “distância” entre observações (n vetores em $\Re^p$)?

- ✓ 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 quantitativos e qualitativos: como combinar informação? (medidas de distância ponderada – Johnson and Whichern, 2008)
- Dados incompletos: imputação
- …