MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Semestre IME/2022

MAE 5776 – Análise Multivariada

MAE5776: ANÁLISE MULTIVARIADA – 1º Sem/2022 – IME/USP

PROGRAMA

Conteúdo (geral):

1. Introdução: estrutura de dados, medidas resumo multivariadas, propriedades em espaços duais, elipses de concentração de dados, *outliers* multivariados.
2. Distribuição Normal Multivariada: propriedades, estimação, distribuições amostrais, testes de hipóteses para vetores de médias e matrizes de covariância. Regiões (elipsoides) de confiança para vetores de Médias.
3. Técnicas (clássicas) de redução da dimensionalidade (n>p e observações independentes). Teoria de Fatoração de Matrizes na redução de dimensionalidade e integração de bancos de dados.
4. Técnicas de redução da dimensionalidade em espaços mais gerais (n<<p, n muito grande, observações não independentes): soluções em espaços duais, soluções regularizadas e penalizadas, reamostragem.
5. Temas adicionais: Modelos de Equações Estruturais; Teoria de Grafos Probabilísticos; Análise de Dados heterogêneos.

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

Estatísticas Descritivas Multivariadas

Caracterização nutricional de 27 produtos alimentícios (Everitt, 2007)

	energia	proteina	gordura	calcio	ferro
[1,]	340	20	28	9	2.6
[2,]	245	21	17	9	2.7
[3,]	420	15	39	7	2.0
[4,]	375	19	32	9	2.5
[5,]	180	22	10	17	3.7
[6,]	115	20	3	8	1.4
[7,]	170	25	7	12	1.5
[8,]	160	26	5	14	5.9
[9,]	265	20	20	9	2.6
[10,]	300	18	25	9	2.3
[11,]	340	20	28	9	2.5
[12,]	340	19	29	9	2.5
[13,]	355	19	30	9	2.4
[14,]	205	18	14	7	2.5
[15,]	185	23	9	9	2.7
[16,]	135	22	4	25	0.6
[17,]	70	11	1	82	6.0
[18,]	45	7	1	74	5.4
[19,]	90	14	2	38	0.8
[20,]	135	16	5	15	0.5
[21,]	200	19	13	5	1.0
[22,]	155	16	9	157	1.8
[23,]	195	16	11	14	1.3
[24,]	120	17	5	159	0.7
[25,]	180	22	9	367	2.5
[26,]	170	25	7	7	1.2
[27,]	110	23	1	98	2.6

$Y_{27\times 5}$: matriz de dados

n=27 unidades amostrais independentes (corresponde a uma amostra aleatória de alguma população de interesse)

p=5 variáveis (quantitativas)

$$Y_{12\times 3} = 29$$

$\Rightarrow$ 27 vetores (linha) na dimensão 5 ($\Re^5$)
$\Rightarrow$ 5 vetores (coluna) na dimensão 27 ($\Re^{27}$)

Pense em medidas descritivas úteis para resumir informações sobre os “indivíduos” e sobre as “variáveis” !

Matriz de Dados (Quantitativos)

$$Y_{n\times p}=\begin{pmatrix} Y_{1.}\,' \\ Y_{2.}\,' \\ \cdots \\ Y_{n.}\,' \end{pmatrix}=\left(Y_{.1},Y_{.2},\ldots,Y_{.p}\right)$$

“Q-espaço”

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

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

“R-espaço”

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

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

Explorar as propriedades geométricas de espaços vetoriais

Matriz de Dados (Quantitativos)

$$Y_{i.} \in \Re^p$$
$$\left(\mu\,;\Sigma\right)$$

População de interesse

Amostra aleatória de "n" observações

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

$$Y_{i.\;p\times 1} = \left(Y_{ij}\right) \overset{iid}{\sim} \left(\mu_{p\times 1};\Sigma_{p\times p}\right)$$

Considere, inicialmente, que uma amostra aleatória de "*n*" vetores de dimensão "*p*" é extraída de uma população de interesse, com vetor de medias (centróide) μ e matriz de covariância Σ.

Estatísticas Descritivas Multivariadas

$Y_{n \times p}$

Vetor de Médias amostral: $\bar{Y}_{p\times 1};\quad \bar{Y} = \left(\bar{Y}_1, \ldots, \bar{Y}_p\right)$

Estatísticas das variáveis

"R-técnicas"

Matriz de Covariâncias amostral: $S_{p\times p} = \left(S_{j\,j'}\right)$

"R-técnicas"

Matriz de correlações amostral: $R_{p\times p} = \left(R_{j\,j'}\right)$

"Q-técnicas"

Matriz de distâncias (Euclidiana) amostral: $D_{n\times n} = \left(d^2_{E\,i\,i'}\right)$

Estatísticas das unidades amostrais

$\bar{Y}_j$? $\quad S_{j\,j'}$? $\quad R_{j\,j'}$? $\quad d^2_{E\,i\,i'}$?

Estatísticas Descritivas Univariadas

$$\bar{Y}_j = \frac{1}{n}\sum_{i=1}^{n} Y_{ij}$$

Média amostral da variável j (escalar)

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

$$\boxed{n-1} \quad S_{u\ jj'}$$

Covariância entre as variáveis j e j' (escalar)

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

Correlação entre as variáveis j e j' . É a Covariância entre as variáveis j e j' padronizadas

$$d^2_{E_{ik}} = d^2_{ik} = \left(Y_{i.} - Y_{k.}\right)'\left(Y_{i.} - Y_{k.}\right) = \sum_{j=1}^{p}(Y_{ij} - Y_{kj})^2;$$

$$Y_{i._{p\times 1}} \in \Re^p \Rightarrow d_{E\,ik} \in \Re;$$

Distância entre os indivíduos i e k (escalar)

Estatísticas Descritivas Multivariadas

Notação Matricial

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

Centróide amostral: vetor de médias para as p variáveis

$$1_{n\ n\times 1} = \begin{pmatrix} 1 \\ 1 \\ \dots \\ 1 \end{pmatrix}; \quad 1_{n\ 1\times n}' = \begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix}$$

Estatísticas Descritivas Multivariadas
Notação Matricial

Matriz de covariâncias

$$S_{p\times p} = \left(s_{jj}\right) = \begin{pmatrix} s_{11} & s_{12} & & s_{1p} \\ s_{21} & s_{22} & & s_{2p} \\ & & \cdots & \\ s_{p1} & s_{p2} & & s_{pp} \end{pmatrix}$$

Matriz simétrica
Diagonal com as variâncias
Nas triangulares (superior e inferior) as covariâncias

Matriz de correlações

$$R_{p\times p} = \left(r_{jj}\right) = \begin{pmatrix} 1 & r_{12} & & r_{1p} \\ r_{21} & 1 & & r_{2p} \\ & & \cdots & \\ r_{p1} & r_{p2} & & 1 \end{pmatrix}$$

Matriz simétrica
Diagonal com valores 1
Nas triangulares (superior e inferior) as correlações

Estatísticas Descritivas Multivariadas
Notação Matricial

Matriz de covariâncias

Produto interno centrado

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

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

$$= (HY)'\,HY$$

$$H = I_n - \frac{1}{n}1_n1_n{}'$$

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

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

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.), tal que, para $a \in \Re^p$:

n>p: S é p.d.

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

- Matriz de correlações

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

- Matriz de covariâncias com denominador (n-1)

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

"no pacote R":
S_u=cov(Y)

Medidas de Variabilidade Multivariada

Critério de otimalidade: obter vetores reducionistas que maximizam o tr(S)

- Variância total (traço da matriz de covariâncias)

$$trS = \sum_{j=1}^{p} s_{jj} = \sum_{j=1}^{p} \lambda_j$$

λ_j : Matriz diagonal com os autovalores de S (obtidos da decomposição spectral de S)

- Variância generalizada (determinante da matriz de covariâncias)

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

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}; \quad tr\left(S\right) = S_{11} + S_{22}; \quad |S| = S_{11} S_{22} - S_{12} S_{21}$$

Já comece a revisar estes conceitos!

Decomposição Espectral de Matrizes (Quadradas)

$$S_{p\times p} = V\,\Lambda\,V'; \quad VV' = V'V = I$$

V_{pxp}: Matriz de autovetores (ortonormal)

$\Lambda_{pxp} = (\lambda_j)$: Matriz diagonal de autovalores

$$\left|S - \lambda I_p\right| = 0$$

Obter os autovalores de S (raízes características)

$$SV_j = \lambda_j V_j$$

Obter os autovetores de S

Revise também:
Decomposição em Valores Singulares de Matrizes Retangulares

Exemplo:

$$S_{2\times 2} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}; \quad S_{11} = S_{22}$$

- Autovalores:

$$\left|S - \lambda I_2\right| = 0 \quad \Rightarrow \quad \lambda_1 = S_{11} + S_{12} \quad \lambda_2 = S_{11} - S_{12}$$

- Autovetores: $\Sigma P_j = \lambda_j P_j$

$$\Rightarrow \quad P_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \quad P_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}; \quad P_j{}'P_j = 1 \quad P_j{}'P_{j'} = 0$$

Padronização de Variáveis

Variável original $\Rightarrow$ Variável padronizada

$$Y_{ij} \Rightarrow Y_{ij}^{*} = \frac{Y_{ij} - \bar{Y}_j}{\sqrt{S_{jj}}}; \quad i = 1, \ldots, n; \; j = 1, \ldots, p \quad \Rightarrow Y_{i_{p\times 1}}^{*} = D_{s_{jj}}^{-1/2}\left(Y_i - \bar{Y}\right)$$

$$Y_{n\times p} \quad \Rightarrow \quad S_{p\times p} \quad R_{p\times p} \quad tr(S) \quad |S|$$

Matriz de covariâncias e de Correlação das variáveis originais

Variância total e Variância generalizada das variáveis originais

$$Y_{n\times p}^{*} \quad \Rightarrow \quad S_{p\times p}^{*} = R_{p\times p} \quad tr(R) \quad |R|$$

Matriz de covariâncias das variáveis padronizadas = Matriz de Correlação das variáveis originais

Variância total das variáveis padronizadas é igual a p!

Estatísticas Descritivas Multivariadas

Caracterização nutricional de 27 produtos alimentícios (Everitt, 2007)

	energia	proteina	gordura	calcio	ferro
[1,]	340	20	28	9	2.6
[2,]	245	21	17	9	2.7
[3,]	420	15	39	7	2.0
[4,]	375	19	32	9	2.5
[5,]	180	22	10	17	3.7
[6,]	115	20	3	8	1.4
[7,]	170	25	7	12	1.5
[8,]	160	26	5	14	5.9
[9,]	265	20	20	9	2.6
[10,]	300	18	25	9	2.3
[11,]	340	20	28	9	2.5
[12,]	340	19	29	9	2.5
[13,]	355	19	30	9	2.4
[14,]	205	18	14	7	2.5
[15,]	185	23	9	9	2.7
[16,]	135	22	4	25	0.6
[17,]	70	11	1	82	6.0
[18,]	45	7	1	74	5.4
[19,]	90	14	2	38	0.8
[20,]	135	16	5	15	0.5
[21,]	200	19	13	5	1.0
[22,]	155	16	9	157	1.8
[23,]	195	16	11	14	1.3
[24,]	120	17	5	159	0.7
[25,]	180	22	9	367	2.5
[26,]	170	25	7	7	1.2
[27,]	110	23	1	98	2.6

$Y_{27\times5}$: matriz de dados

Algumas questões:

Qual é o estado nutricional médio destes produtos alimentícios?

Os produtos alimentícios são mais diferentes relativamente a qual variável nutricional?

O perfil nutricional destes produtos pode ser considerado uniforme ou há uma compensação nas composições deles?

Estatísticas Descritivas Multivariadas

Matriz de dados: Avaliações nutricionais (p=5) de n=27 produtos alimentícios.

$$Y_{27\times5} = \left(Y_{1\,27\times1}, Y_{2\,27\times1}, Y_{3\,27\times1}, Y_{4\,27\times1}, Y_{5\,27\times1}\right)$$

$\bar{Y}_{4\times1} =$

energia	proteína	gordura	cálcio	ferro
207.41	19.00	13.48	43.96	2.38

Há indicação de heterocedasticidade? (Note que as var. podem ter escalas diferentes)

$S_{5\times5} =$

	energia	proteina	gordura	calcio	ferro
energia	**10243.02**	74.81	1124.57	-2530.29	-14.75
proteina	74.81	18.08	1.19	-28.23	-1.08
gordura	1124.57	1.19	126.72	-270.67	-1.00
calcio	-2530.29	-28.23	-270.67	6089.34	5.05
ferro	-14.75	-1.08	-1.00	5.05	2.13

?

Variância total: tr(S) = 16479.3 Variância generalizada: det(S)= 841117207

$R_{5\times5} =$

	energia	proteina	gordura	calcio	ferro
energia	1.00	0.17	0.99	-0.32	-0.10
proteina	0.17	1.00	0.02	-0.09	-0.17
gordura	0.99	0.02	1.00	-0.31	-0.06
calcio	-0.32	-0.09	-0.31	1.00	0.04
ferro	-0.10	-0.17	-0.06	0.04	1.00

A estrutura de correlação sugere ser "uniforme" entre as variáveis?

	Dados originais					Dados padronizados				
	energia	proteina	gordura	calcio	ferro	energia	proteina	gordura	calcio	ferro
[1,]	340	20	28	9	2.6	1.31	0.24	1.29	-0.45	0.15
[2,]	245	21	17	9	2.7	0.37	0.47	0.31	-0.45	0.22
[3,]	420	15	39	7	2.0	2.10	-0.94	2.27	-0.47	-0.26
[4,]	375	19	32	9	2.5	1.66	0.00	1.65	-0.45	0.08
[5,]	180	22	10	17	3.7	-0.27	0.71	-0.31	-0.35	0.91
[6,]	115	20	3	8	1.4	-0.91	0.24	-0.93	-0.46	-0.67
[7,]	170	25	7	12	1.5	-0.37	1.41	-0.58	-0.41	-0.60
[8,]	160	26	5	14	5.9	-0.47	1.65	-0.75	-0.38	2.41
[9,]	265	20	20	9	2.6	0.57	0.24	0.58	-0.45	0.15
[10,]	300	18	25	9	2.3	0.91	-0.24	1.02	-0.45	-0.05
[11,]	340	20	28	9	2.5	1.31	0.24	1.29	-0.45	0.08
[12,]	340	19	29	9	2.5	1.31	0.00	1.38	-0.45	0.08
[13,]	355	19	30	9	2.4	1.46	0.00	1.47	-0.45	0.02
[14,]	205	18	14	7	2.5	-0.02	-0.24	0.05	-0.47	0.08
[15,]	185	23	9	9	2.7	-0.22	0.94	-0.40	-0.45	0.22
[16,]	135	22	4	25	0.6	-0.72	0.71	-0.84	-0.24	-1.22
[17,]	70	11	1	82	6.0	-1.36	-1.88	-1.11	0.49	2.48
[18,]	45	7	1	74	5.4	-1.60	-2.82	-1.11	0.38	2.07
[19,]	90	14	2	38	0.8	-1.16	-1.18	-1.02	-0.08	-1.08
[20,]	135	16	5	15	0.5	-0.72	-0.71	-0.75	-0.37	-1.29
[21,]	200	19	13	5	1.0	-0.07	0.00	-0.04	-0.50	-0.94
[22,]	155	16	9	157	1.8	-0.52	-0.71	-0.40	1.45	-0.40
[23,]	195	16	11	14	1.3	-0.12	-0.71	-0.22	-0.38	-0.74
[24,]	120	17	5	159	0.7	-0.86	-0.47	-0.75	1.47	-1.15
[25,]	180	22	9	367	2.5	-0.27	0.71	-0.40	4.14	0.08
[26,]	170	25	7	7	1.2	-0.37	1.41	-0.58	-0.47	-0.81
[27,]	110	23	1	98	2.6	-0.96	0.94	-1.11	0.69	0.15

Centróide de Y^*? Matriz de Covariância de Y^*?

Estatísticas Descritivas Multivariadas

Dados Padronizados: Avaliações nutricionais

Aproximadamente "0"

$$\bar{Y}^*_{4\times1} =$$

energia	proteina	gordura	calcio	ferro
-1.227156e-16	-6.167906e-18	5.962309e-17	2.569961e-17	4.465307e-17

$$S^*_{5\times5} = R =$$

	energia	proteina	gordura	calcio	ferro
energia	1.00	0.17	0.99	-0.32	-0.10
proteina	0.17	1.00	0.02	-0.09	-0.17
gordura	0.99	0.02	1.00	-0.31	-0.06
calcio	-0.32	-0.09	-0.31	1.00	0.04
ferro	-0.10	-0.17	-0.06	0.04	1.00

Cov(Y*) = Cor(Y)

Variância total das var padronizadas: tr(R) = 5 = "p"

Variância generalizada das var padronizadas: det(R)= 0.002758483

Estruturas de Correlação entre Variáveis

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 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	$\begin{pmatrix} 1 & a & \cdots & a^{t-1} \\ a & 1 & \cdots & a^{t-2} \\ \vdots & \vdots & \ddots & \vdots \\ a^{t-1} & a^{t-2} & \cdots & 1 \end{pmatrix}$	

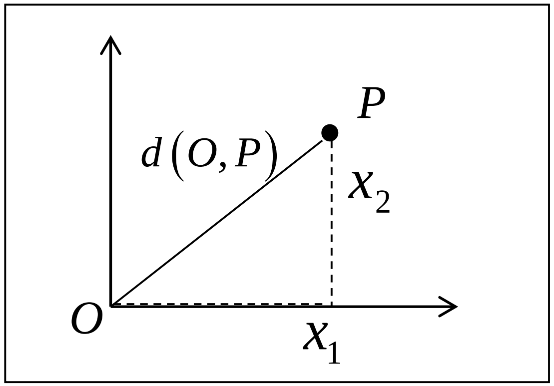
Casos Mais Gerais

Como medir **“dependência” entre variáveis** (espaço das colunas de Y_{nxp})?

- Correlação de Pearson ($\rho_{Y1,Y2}$)
- Correlação de Spearman: análises mais robustas
- Correlações Parciais ($\rho_{Y1,Y2|X}$): S-1 (“aprendizado de estruturas”)
- Matrizes de importância entre variáveis (Saaty, 1980): a matriz não é simétrica e sim recíproca
- Modelagem por cópulas (construção de distribuições conjuntas a partir de distribuições marginais)

Como medir “dependência” entre observações (n vetores em $\Re^p$)?
Vamos adotar Medidas de Distância.

Medidas de Distância entre Observações



Ponto genérico em $\Re^2$

$$P = (x_1, x_2)'$$

Origem

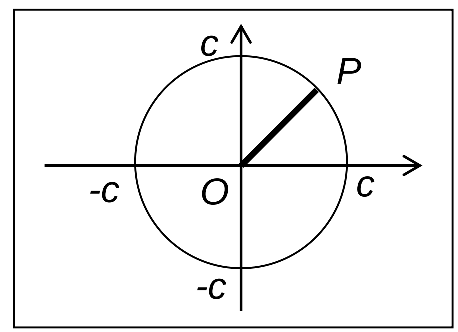
$$O = (0,0)$$

Distância Euclidiana de *P* a *O* (Teorema de Pitágoras)

$$d^2(O,P) = x_1^2 + x_2^2 \;\Rightarrow\; d(O,P) = \sqrt{x_1^2 + x_2^2}$$

Pontos a uma distância c^2 da origem satisfazem à equação:

$$d^2(O,P) = x_1^2 + x_2^2 = c^2$$

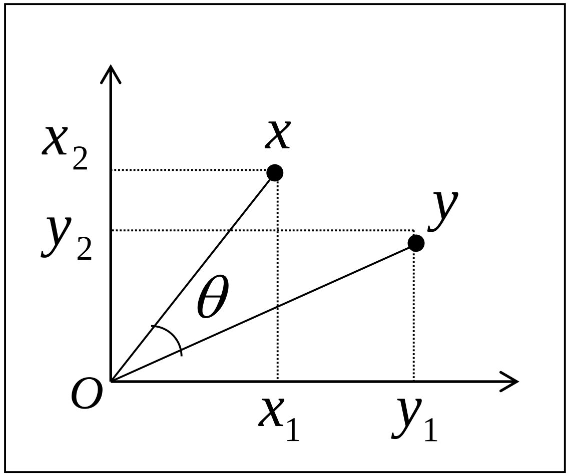


Generalizando para pontos p-dimensionais (em $\Re^p$):

$$d^2(O,P) = x_1^2 + x_2^2 + \ldots + x_p^2 \quad \Rightarrow \quad d^2(O,P) = (x_1, \ldots, x_p)\begin{pmatrix} x_1 \\ \ldots \\ x_p \end{pmatrix} = x'x$$

Produto interno

Medidas de Distância entre Observações



$$x = (x_1, x_2)' \qquad y = (y_1, y_2)'$$

Distância Euclidiana entre os pontos x e y :

$$d^2(x,y) = (y_1 - x_1)^2 + (y_2 - x_2)^2$$
$$= (x-y)'(x-y)$$

Generalizando para pontos p-dimensionais (em $\Re^p$):

$$d^2(x,y) = (x_1 - y_1)^2 + \ldots + (x_p - y_p)^2 \quad = \sum_{j=1}^{p} (x_j - y_j)^2$$

Distância Euclidiana

```
Matriz de distância Euclidiana
(para as 6 primeiras observações dos dados nutricionais)

       1      2      3      4      5      6
1   0.00  95.64  80.93  35.24 161.22 226.39
2  95.64   0.00 176.49 130.88  65.88 130.77
3  80.93 176.49   0.00  45.76 242.06 307.16
4  35.24 130.88  45.76   0.00 196.43 261.62
5 161.22  65.88 242.06 196.43   0.00  66.06
6 226.39 130.77 307.16 261.62  66.06   0.00

> dat[1,]
 energia proteina  gordura   calcio    ferro
   340.0     20.0     28.0      9.0      2.6
> dat[4,]
 energia proteina  gordura   calcio    ferro
   375.0     19.0     32.0      9.0      2.5
> dat[3,]
 energia proteina  gordura   calcio    ferro
     420       15       39        7        2
> dat[6,]
 energia proteina  gordura   calcio    ferro
   115.0     20.0      3.0      8.0      1.4
```

Quais produtos alimentícios (dentre os 6 apresentados) são mais “parecidos” nutricionalmente?

Quais são mais “diferentes”?

Distância Euclidiana Padronizada (Distância de Pearson)

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

```
Matriz de distância Euclidiana Padronizada (Pearson)
(para as 6 primeiras observações dos dados nutricionais)
     1    2    3    4    5    6
1 0.00 1.38 1.77 0.55 2.42 3.25
2 1.38 0.00 3.01 1.91 1.15 2.01
3 1.77 3.01 0.00 1.26 4.04 4.57
4 0.55 1.91 1.26 0.00 2.95 3.72
5 2.42 1.15 4.04 2.95 0.00 1.87
6 3.25 2.01 4.57 3.72 1.87 0.00

Dados Padronizados das observações 1 e 4
Calcule a distância de Pearson:

> datp[1,]
 energia proteina  gordura   calcio    ferro
3.359425 4.704005 2.487334 0.115334 1.779777
> datp[4,]
 energia proteina  gordura   calcio    ferro
3.705248 4.468804 2.842667 0.115334 1.711324
```

É a distância Euclidiana entre as observações padronizadas!

Quais produtos são mais “parecidos” nutricionalmente?

Quais são mais “diferentes”?

Transformações Lineares – Medidas de Distância

- Transformação de Escala (padronização)

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

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

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

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

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

Distância de Pearson ao quadrado (Euclidiana padronizada)

- Transformação de Mahalanobis

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

$$S_Z = I_p$$

Variáveis independentes e variâncias unitárias

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

Distância de Mahalanobis ao quadrado

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

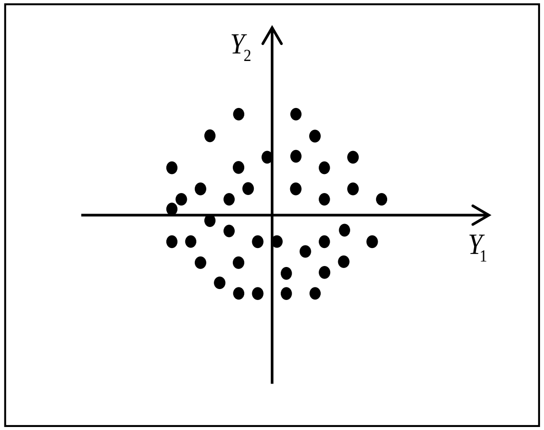
Veremos como obter Coordenadas Principais neste caso (Ex. 1.6.1 de Mardia et al., 2003)

Distâncias Estatísticas

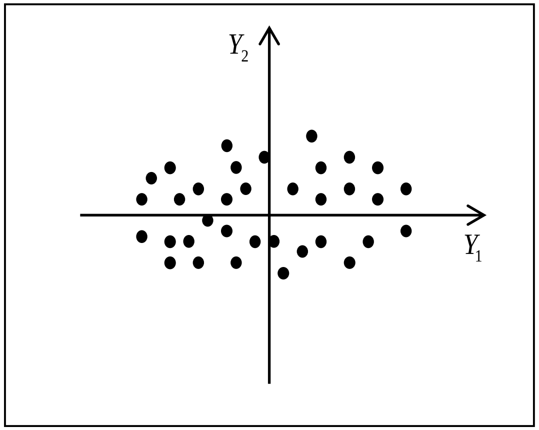
⇒ Para definir Medidas de Distância entre observações é necessário considerar diferenças na variação dos pontos nos eixos e o grau de correlação entre eles.

A dispersão dos pontos é maior na direção Y1 ou na direção Y2 ?

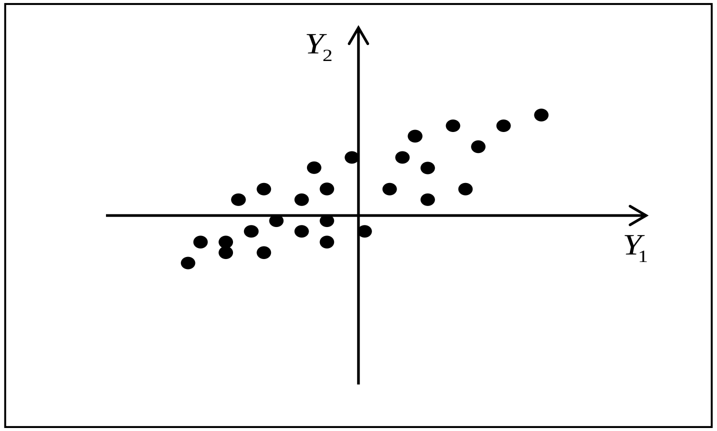
Como é a correlação entre as variáveis?



Variâncias homogêneas
Observações não correlacionadas



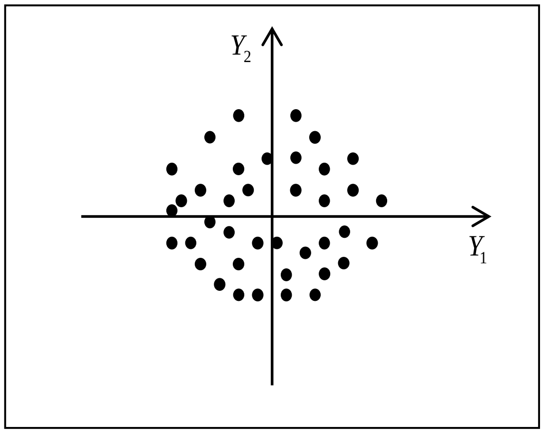
Variâncias heterogêneas
Observações não correlacionadas



Variâncias heterogêneas
Observações correlacionadas

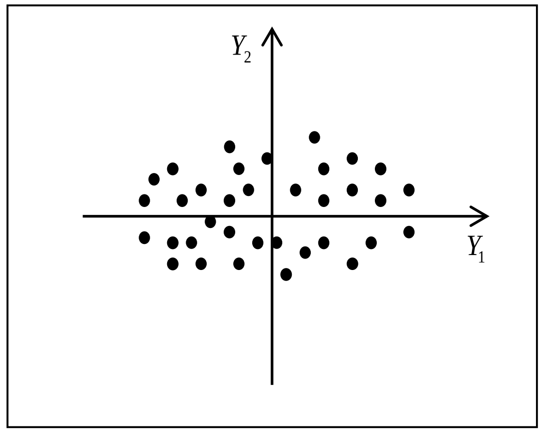
Distâncias Estatísticas

⇒ Para definir Medidas de Distância entre observações é necessário considerar diferenças na variação dos pontos nos eixos e o grau de correlação entre eles.



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

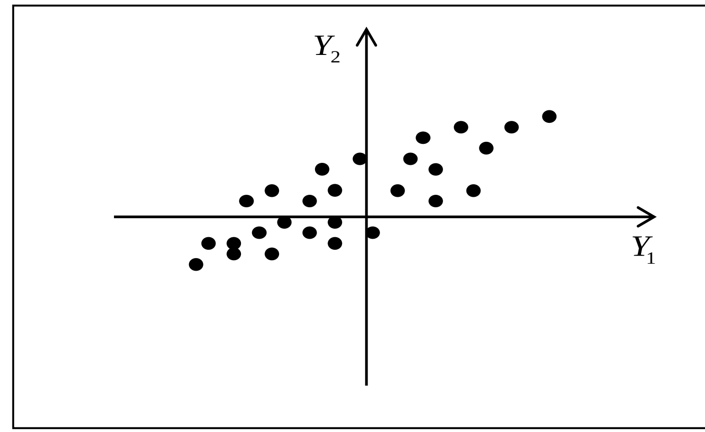
Coordenadas originais



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

Coordenadas padronizadas

(corrige pela heterocedasticidade)



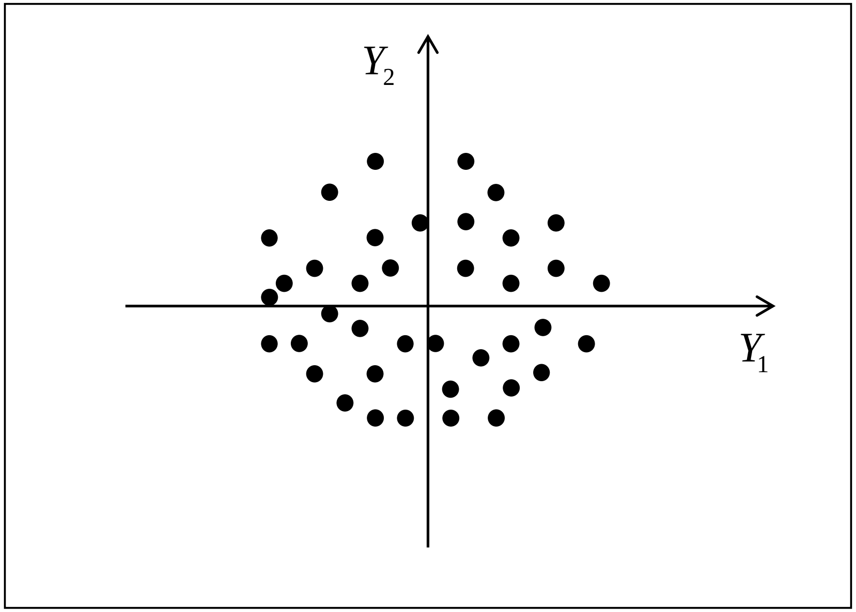
$$d^2(O,P) = a_{11}y_1^2 + \underbrace{a_{12}y_1y_2}_{} + a_{22}y_2^2$$

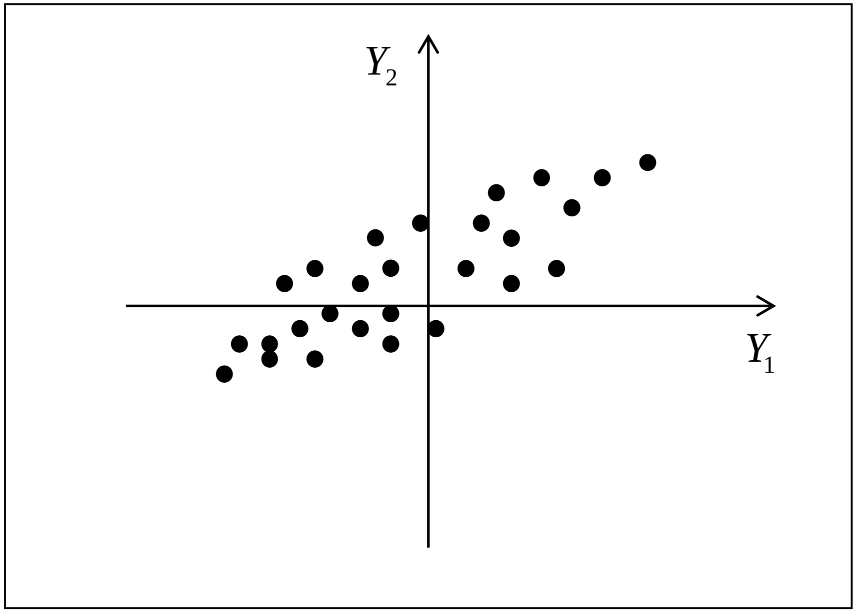
Coordenadas rotacionadas padronizadas

(corrige pela heterocedasticidade e correlação)

Distâncias Estatísticas

Expressão matricial





Variáveis independentes e homocedásticas $\Rightarrow$ **distância Euclidiana** (ordinária)

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

Variáveis correlacionadas e heterocedásticas $\Rightarrow$ **distância de Mahalanobis**

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

Variáveis independentes e heterocedásticas

$\Rightarrow$ **distância de Pearson** (Euclidiana Padronizada)

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

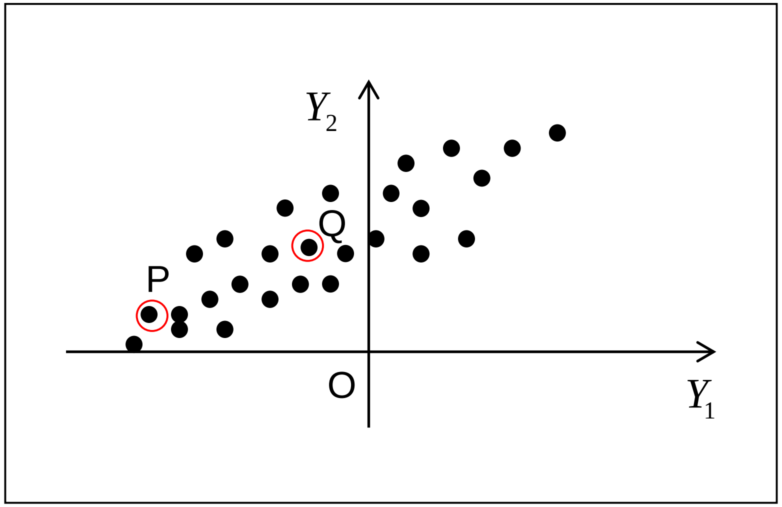
Distância Estatística

⇒ 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 heterocedasticidade uma alternativa é usar a distância de Pearson nas variáveis rotacionadas (Mahalanobis).

Motivação



Johnson and Wichern, 2006.

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

⇒ Porém, note que O é ponto aberrante, ocorre fora da nuvem de dispersão conjunta de Y1 e Y2, o que não ocorre com o ponto P.

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

Distâncias Estatísticas

Matriz de distância Euclidiana
(entre observações)

	1	2	3	4	5	6
1	0.00	95.64	80.93	35.24	161.22	226.39
2	95.64	0.00	176.49	130.88	65.88	130.77
3	80.93	176.49	0.00	**45.76**	242.06	307.16
4	35.24	130.88	45.76	0.00	196.43	261.62
5	161.22	65.88	242.06	196.43	0.00	66.06
6	226.39	130.77	307.16	261.62	66.06	0.00

Dados Nutricionais (distâncias para as 6 primeiras observações)

Matriz de distância de Pearson
(entre observações)

	1	2	3	4	5	6
1	0.00	1.38	1.77	0.55	2.42	3.25
2	1.38	0.00	3.01	1.91	**1.15**	2.01
3	1.77	3.01	0.00	1.26	4.04	4.57
4	0.55	1.91	1.26	0.00	2.95	3.72
5	2.42	1.15	4.04	2.95	0.00	1.87
6	3.25	2.01	4.57	3.72	1.87	0.00

Distância de Mahalanobis é usada como critério de diagnóstico de observações *outliers* multivariadas

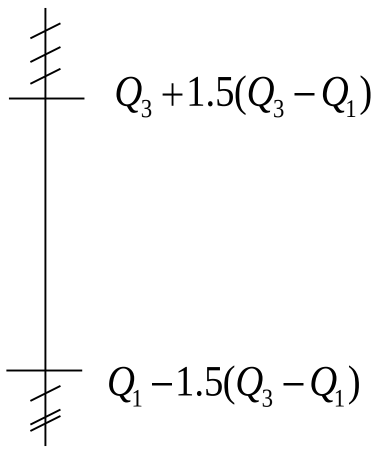
Distância de Mahalanobis
(das observações ao centróide)

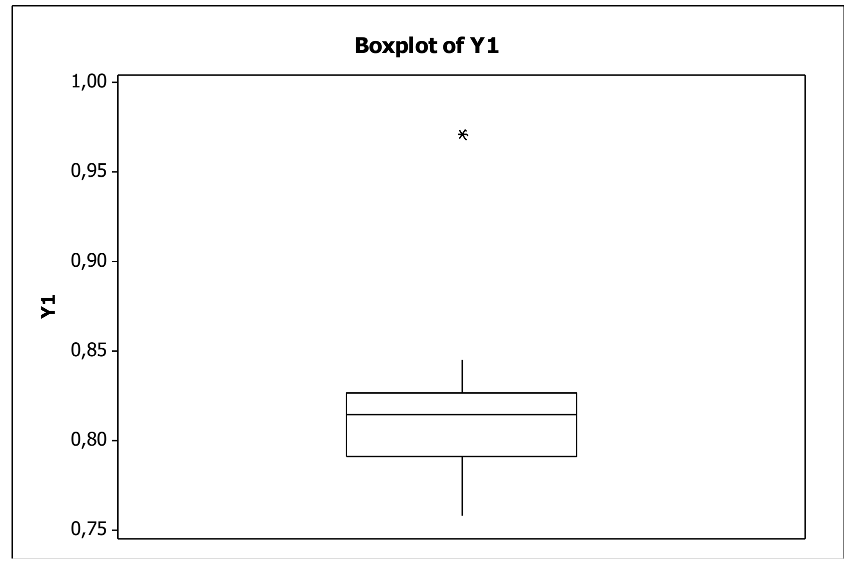
```
 [1]  1.81  0.53  6.27  3.27  2.94  2.66  2.92 12.17  0.86  2.18  1.77
[12]  2.52  2.25  0.72  1.81  2.84  9.81 13.09  4.47  7.24  1.91  2.92
[23] 12.29  5.44 19.60  3.25  2.45
```

Valor Atípico Unidimensional

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

- **Critério do Boxplot**:



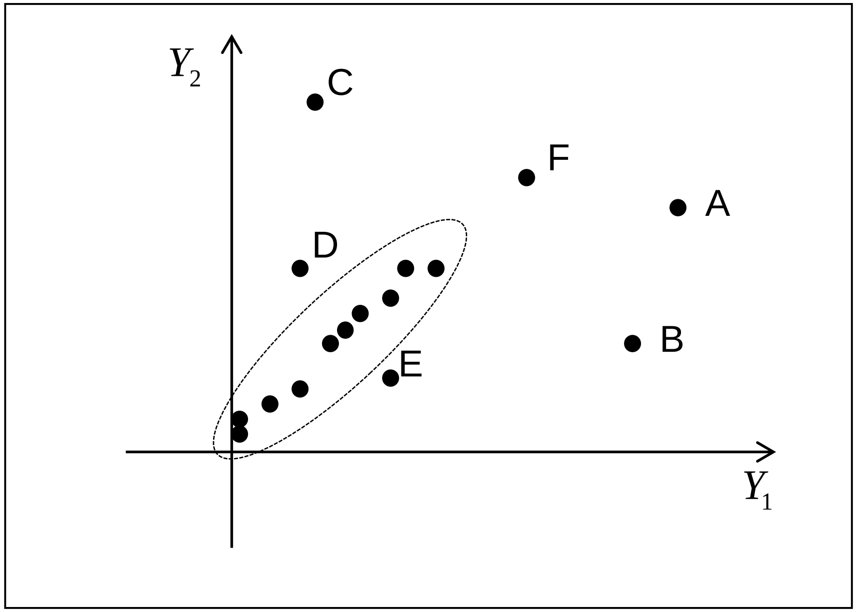


- Dados Padronizados: $Z_i = \dfrac{Y_i - \bar{Y}}{s}$ $\quad P(|Z| \geq 2{,}5) = 0{,}012$ (Hair et al., 1998)

a probabilidade de um ponto estar dentro do intervalo de concentração é (1-α)

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

Valor Atípico Bidimensional



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

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

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

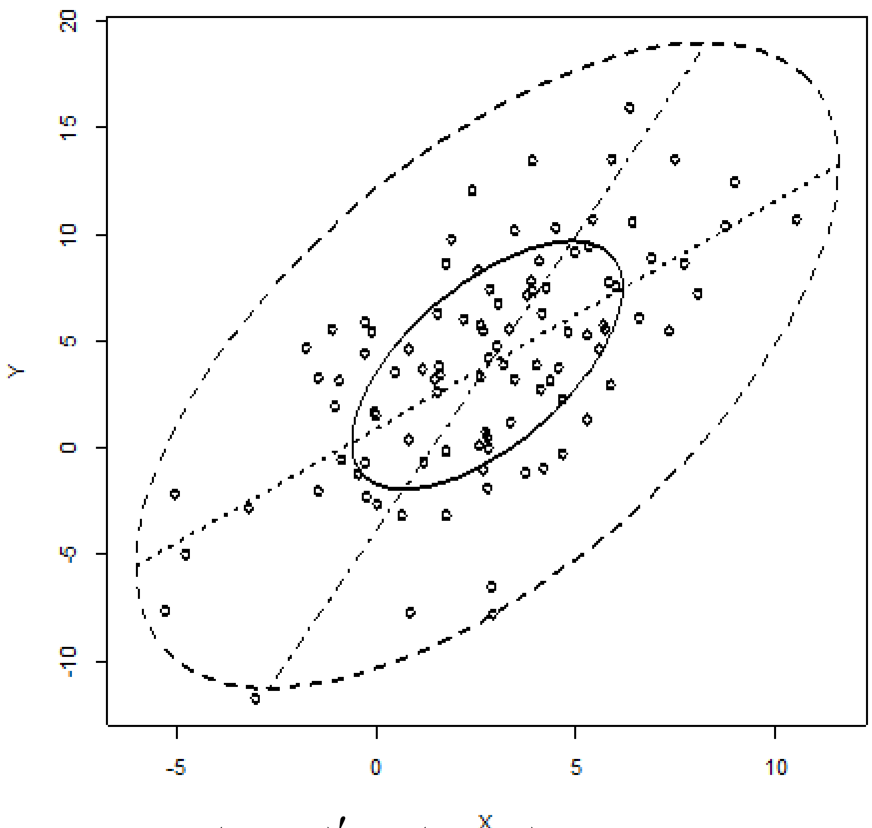
$\Rightarrow$ Pontos B (C) é aberrante para Y1 (Y2) mas não para Y2 (Y1)

$\Rightarrow$ 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

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.

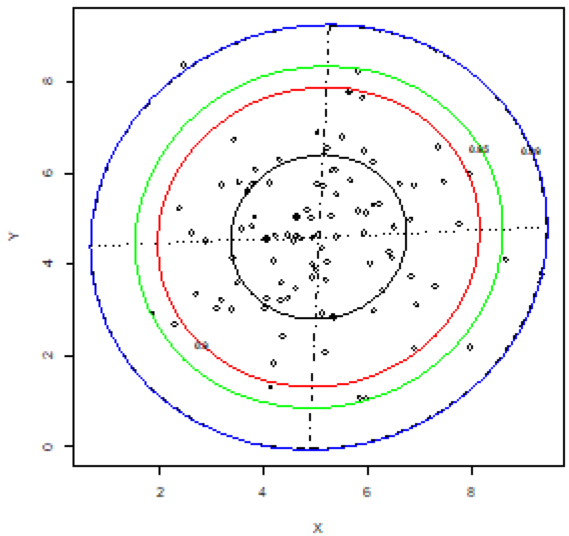
A elipse maior fornece um critério (robusto) de diagnóstico de observações atípicas.

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

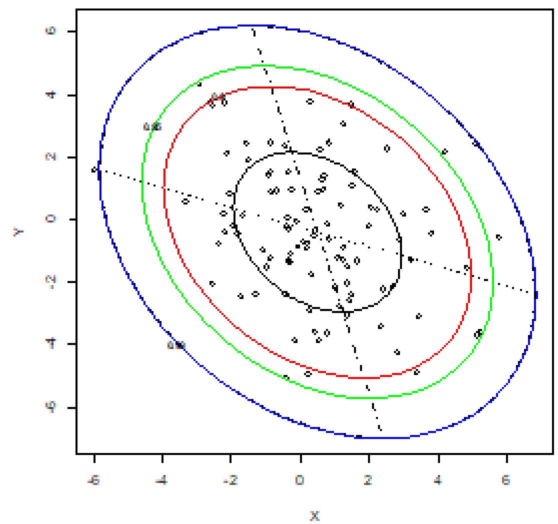
A construção das retas de regressão pode ser por estimacão robusta ou clássica. Quanto menor o ângulo entre as retas maior é o valor absoluto da correlação.

Função bivbox (Everitt, 2005)

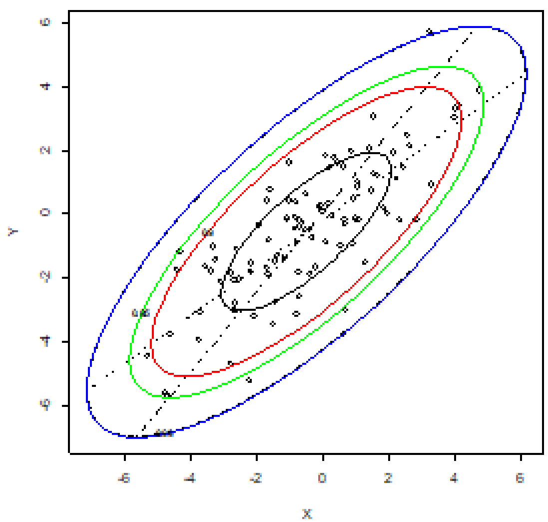
$$\mu' = (5,5) \quad \Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$



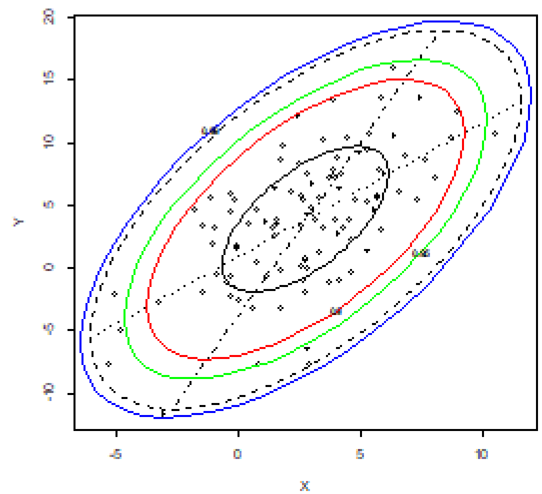
$$\mu' = (0,0) \quad \Sigma = \begin{pmatrix} 4 & -1 \\ -1 & 4 \end{pmatrix}$$



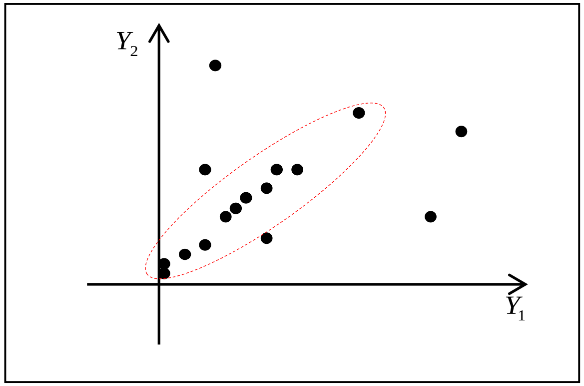
$$\mu' = (0,0) \quad \Sigma = \begin{pmatrix} 4 & 3 \\ 3 & 4 \end{pmatrix}$$



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



Valor Atípico Multidimensional



Diagnóstico de observações atípicas bidimensionais:

Diagnosticar observações distantes da nuvem de dispersão conjunta dos pontos.

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

Elipse de concentração de pontos bidimensionais, centrada na média.

Distância de Mahalanobis (das observações ao centroide)

S: matrix de covariância

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

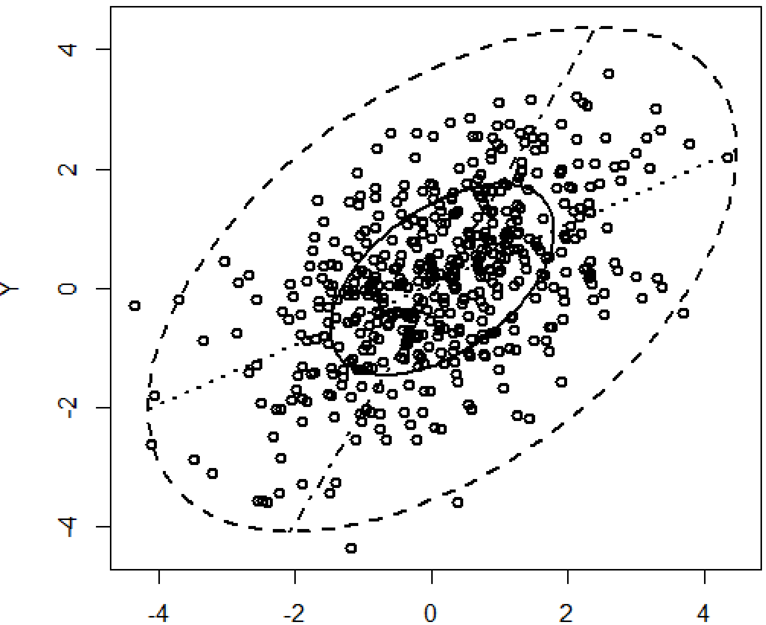
$Y_{i \in \Re^2}$

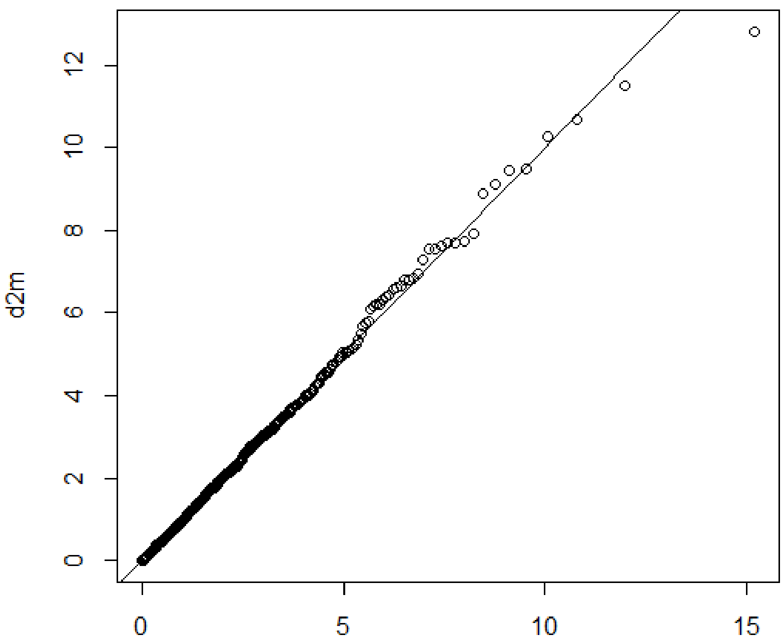
Diagnóstico (p=2): $d_M^2 \overset{n\to\infty}{\sim} \chi_2^2 \Rightarrow P\left(d_M^2 \leq c^2\right) \leq (1-\alpha)$

Dimensão geral

Diagnóstico ($Y_i \in \Re^p$): $d_M^2 \overset{n\to\infty}{\sim} \chi_p^2 \Rightarrow P\left(d_M^2 \leq c^2\right) \leq (1-\alpha)$

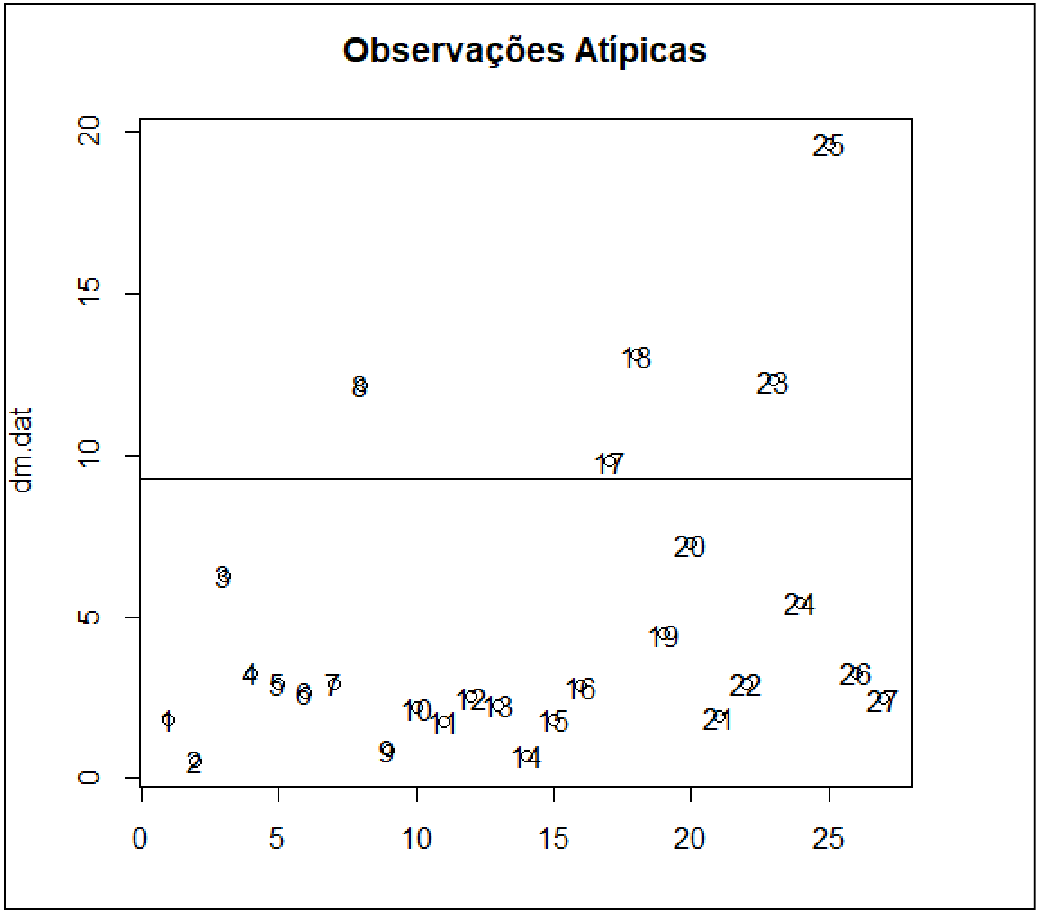
Distância de Mahalanobis Diagnóstico de *Outliers* (via a distribuição Qui-Quadrado)





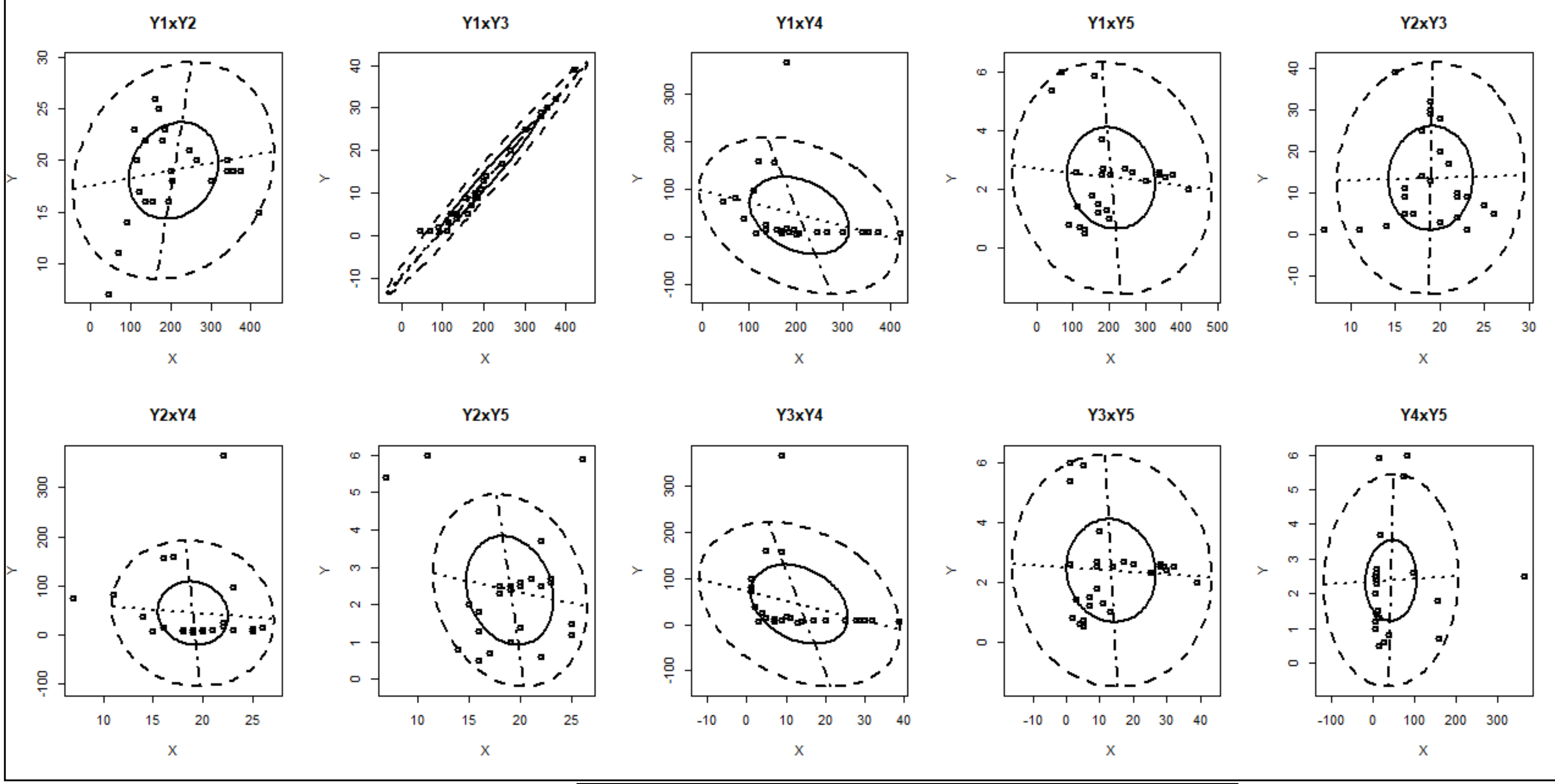

```
library(MASS)
mu<-c(0,0)
sigma<-matrix(c(2,1,1,2),ncol=2)
n<-500
y<-mvrnorm(n,mu,sigma)
mi<-colMeans(y)
s<-cov(y)
par(mfrow=c(1,2))
bivbox(y, method ="O")
# Copy Everitt's bivbox function
d2m<-mahalanobis(y,mi,s)
quantis <- qchisq(ppoints(length(y)),df=2)
qqplot(quantis, d2m)
abline(0,1)
```

Observação Atípica Multidimensional
Diagnóstico – Distância de Mahalanobis



Diagnóstico de *outliers* para os dados nutricionais:

$$d_M^2 \leq \chi_5^2(\alpha = 0,10) = 9.236$$

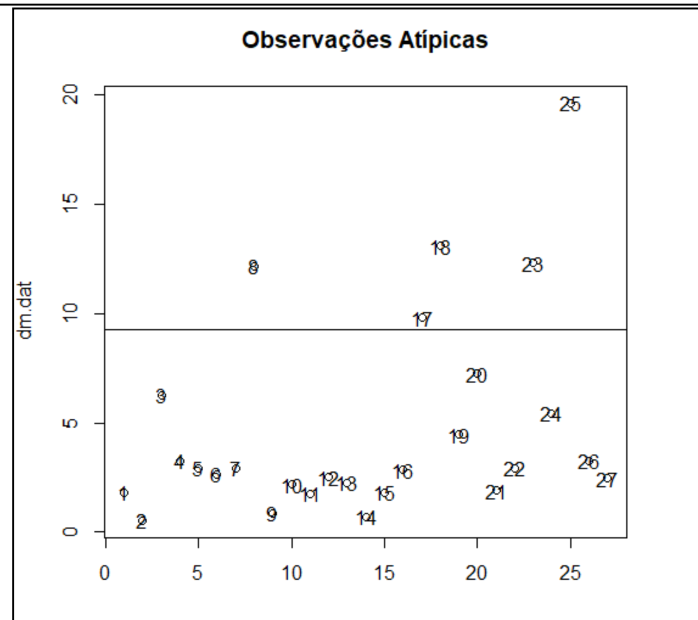


Dados Nutricionais de 27 produtos alimentícios.

Diagnóstico de observações atípicas:

Boxplot bivariado (p=2)

Distância de Mahalanobis (p=5)



$$d_M^2 \leq \chi_5^2(\alpha = 0{,}10) = 9.236$$

Intervalos de Concentração de Observações
Regiões de Concentração de Observações

Caso univariado

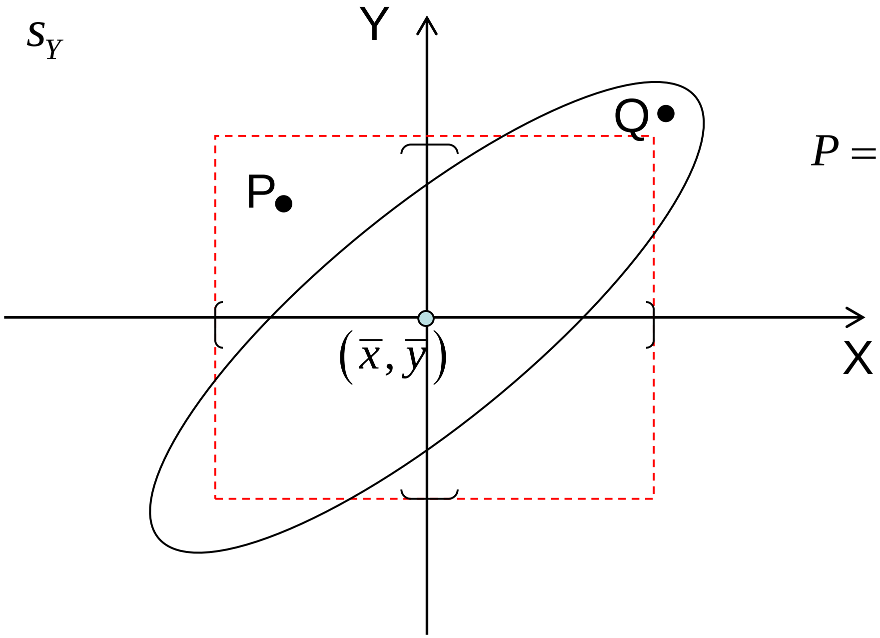
$$z_X = \frac{x - \bar{x}}{s_X}, \quad z_Y = \frac{y - \bar{y}}{s_Y}$$

$$z^2 \sim \chi^2_1$$

Caso multivariado (p=2)

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

$$P = (x_1, y_1); \quad Q = (x_2, y_2)$$



Avaliando os intervalos de concentração dos dados bem como a elipse de concentração:

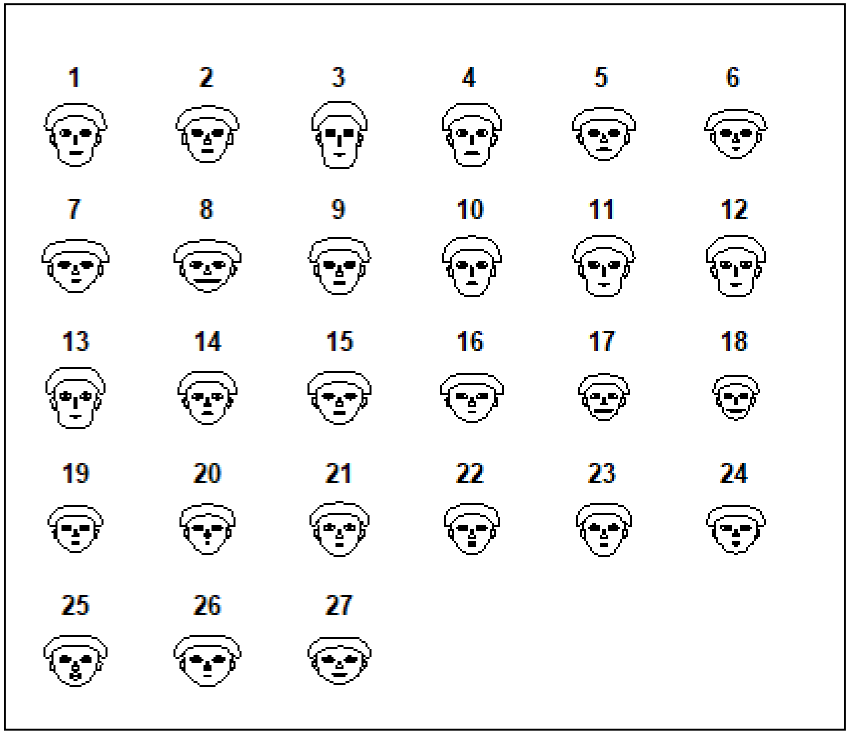
⇒ P não é *outlier* univariado (nem na variável X e nem na variável Y), mas é *outilier* bivariado.

⇒ Q é *outlier* tanto na direção do eixo X como de Y, mas não é outlier bivariado.

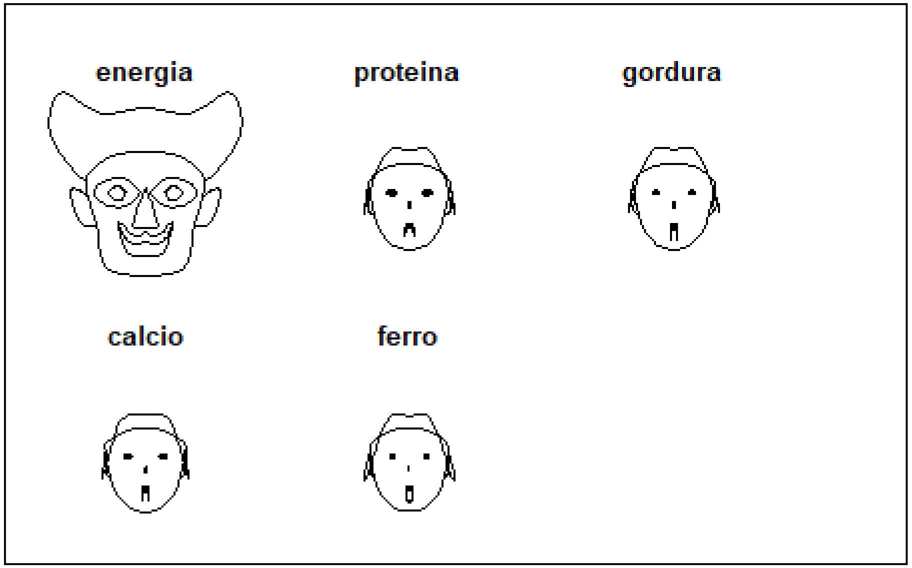
Dados Multivariados
Representação Gráfica (Pictorial)

Faces de Chernoff

Qual produto (n=27) é mais diferente dos demais?



Qual componente nutricional (p=5) mais diferencia os produtos alimentícios?



Avaliar a contribuição das variáveis ao perfil multivariado $\Rightarrow$ **Análise de Componentes Principais**

Conteúdo Extra a Pesquisar

Implementar no R!

Gerar dados da Normal Multivariada

$$Y_i \in \mathbb{R}^p \overset{iid}{\sim} N\left(\mu_{p\times 1}; \Sigma_{p\times p}\right) \quad i = 1, 2, \ldots, n$$

Decomposição Espectral da Matriz de Covariâncias

$$\Sigma_{p\times p} = V\, D_{\lambda_j} V'$$

$V_{p\times p}$: Matriz de "p" Autovetores coluna de Y

D_{λ_j} : Matriz Diagonal de "p" Autovalores

Decomposição em Valores Singulares da Matriz de Dados

$$Y_{n\times p} = U\ D_{\lambda_j} V' \quad \Rightarrow \quad \begin{cases} \left(Y'Y\right)_{p\times p} = VDV' \\ \left(YY'\right)_{n\times n} = UDU' \end{cases}$$

$U_{n\times n}$: Matriz de "n" Autovetores linha de Y

$V_{p\times p}$: Matriz de "p" Autovetores coluna de Y

D_{λ_j} : Matriz com os "p" Autovalores