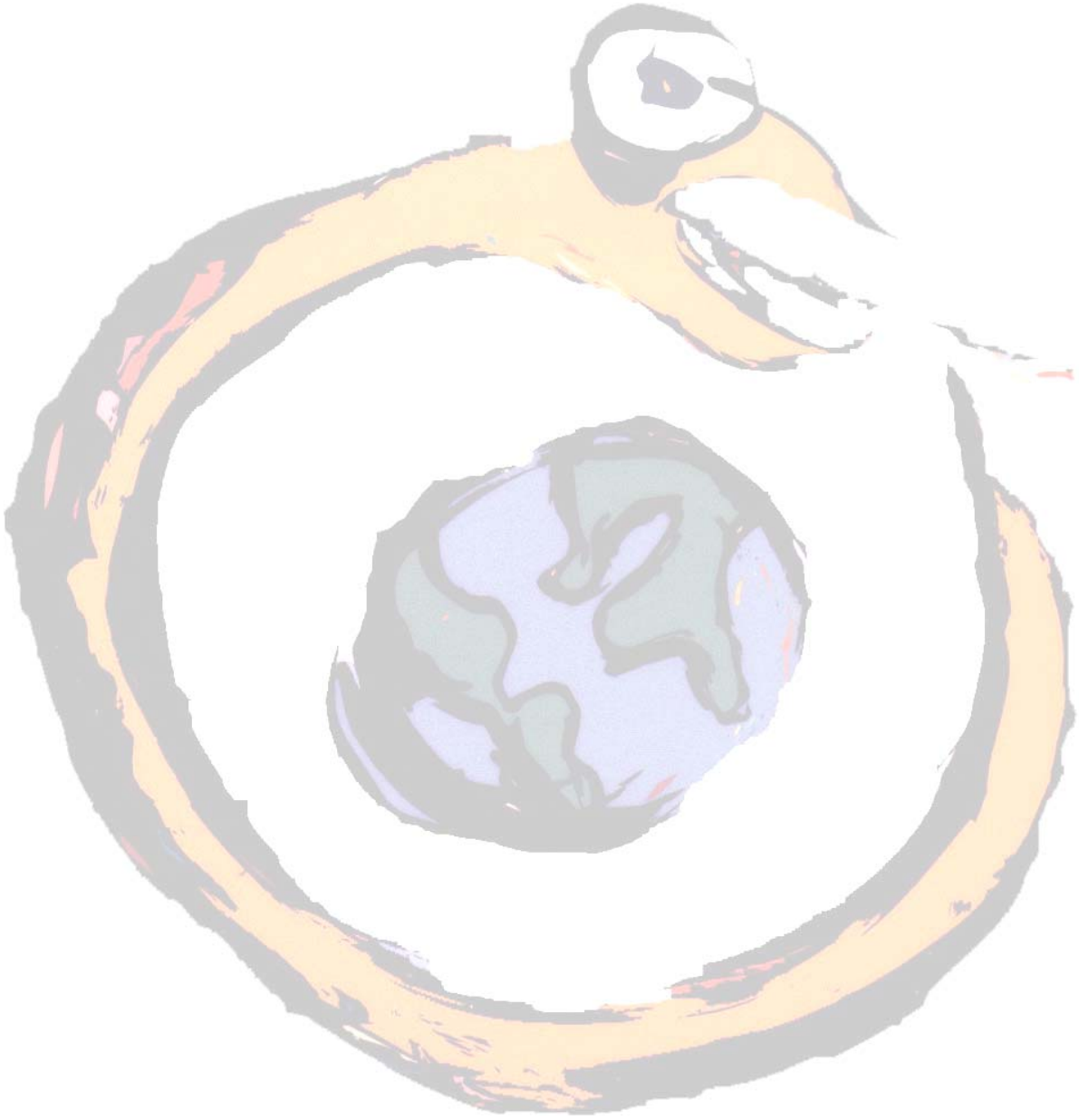
Faculdade de Economia da Universidade de Coimbra

Estatística Multivariada Aplicada

Pedro Lopes Ferreira

2000



Sumário

1

Introdução à análise multivariada

1.1 A organização dos dados

Sendo este um curso de estatística multivariada, iremos analisar medições feitas em várias variáveis ou características. Estas medições (dados) são normalmente apresentadas quer graficamente, quer sob a forma matricial.

Assim, se considerarmos n medições em p variáveis, x_{ij} representará a medição da variável j no item i. A sua representação matricial será

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} \\ \cdot & \cdot & \dots & \cdot & \dots & \cdot \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} \\ \cdot & \cdot & \dots & \cdot & \dots & \cdot \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix}$$

Esta matriz X contém os dados de todas as observações em todas as variáveis.

Exemplo 1.1: Pretende-se estudar as vendas dos livros de uma livraria e, para isso, recolheu-se uma amostra de 4 recibos, indicando cada um deles o número de livros vendidos e o total gasto (em centenas de escudos). Numa forma tabular temos os seguintes dados:

Variável	Nome				
1	Total	42	52	48	58
2	No. livros	4	5	4	3

Representando numa forma matricial obtemos a matriz **X** com duas linhas (variáveis) e quatro colunas (itens):

$$X = \begin{bmatrix} 42 & 52 & 48 & 58 \\ 4 & 5 & 4 & 3 \end{bmatrix}'$$

o

1.2 Estatísticas descritivas

Se considerarmos $x_{1j}, x_{2j}, \ldots, x_{ij}, \ldots, x_{nj}$ como representando as n medições feitas na variável j (coluna j da matriz X), podemos denominar por $\bar{x}_j$ a média amostral da variável j

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad j = 1,2,\ldots,p$$

Do mesmo modo, a medida de dispersão variância amostral da variável i é dada por

$$s_i^2 = s_{ii} = \frac{1}{n}\sum_{i=1}^{n} (x_{ij} - xj)^2 \qquad i = 1,2,\ldots,p$$

A raiz quadrada da variância amostral, $\sqrt{s_{jj}}$ é denominada desvio padrão amostral.

Podemos também estar interessados em determinar o grau de associação linear entre duas variáveis j e k. Isto consegue-se através da covariância amostral representada pela média dos produtos dos desvios em relação às respectivas médias

$$s_{ik} = s_{ki} = \frac{1}{n}\sum_{i=1}^{n} (x_{ij} - xj)\ (x_{ik} - x_k) \qquad i = 1,2,\ldots,p \ ;\ k = 1,2,\ldots,p$$

Se valores altos de uma variável foram observados conjuntamente com valores altos de outra variável, e valores pequenos também ocorrerem conjuntamente, s_{jk} será positiva. Se valores altos de uma variável ocorrerem com valores pequenos de outra variável, s_{jk} será negativa. Caso não exista associação entre os valores de ambas as varáveis, s_{jk} será aproximadamente nula.

Finalmente, consideremos o coeficiente de correlação amostral de Pearson, uma medida de associação linear entre duas variáveis, independente das unidades de medida e com valores entre -1 e +1.

$$r_{jk} = r_{kj} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_i)^2}\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}}$$

para $i = 1,2,\ldots,p$ e $k = 1,2,\ldots,p$.

Esta última medida constitui, como facilmente se pode observar, uma versão estandardizada da covariância amostral.

De notar que, se substituirmos os valores originais x_{ij} e x_{ik} pelos correspondentes valores estandardizados $(x_{ij} - \bar{x}_j) / \sqrt{s_{jj}}$ e $(x_{ik} - \bar{x}_k) / \sqrt{s_{kk}}$, o coeficiente de correlação amostral r_{jk} pode ser visto como a covariância amostral. Após a estandardização, ambas as variáveis podem ser comparadas, pois passam a estar nas mesmas unidades.

Voltando, de novo, à apresentação matricial, baseando-nos na matriz X com n medições (linhas) em p variáveis (colunas), as médias amostrais são representadas por

$$\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \cdots \\ \bar{x}_p \end{bmatrix}$$

as variâncias e covariâncias amostrais por

$$S_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ . & . & \cdots & . \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

e as correlações amostrais por

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ . & . & \cdots & . \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

Reparar que as matrizes S_n e R são matrizes simétricas de ordem p.

<u>Exemplo 1.1 (cont):</u>

Pegando de novo na matriz

$$X' = \begin{bmatrix} 42 & 52 & 48 & 58 \\ 4 & 5 & 4 & 3 \end{bmatrix}'$$

podemos determinar o vector $\bar{x}$ e as matrizes S_n e R. Assim,

$$\bar{x_1} = \frac{1}{4}\sum_{i=1}^{4} x_{i1} = \frac{1}{4}(42+52+48+58) = 50$$

$\bar{x}_2 = \frac{1}{4}\sum_{i=1}^{4} x_{i2} = \frac{1}{4}(4+5+4+3) = 4$

e então, $\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} = \begin{bmatrix} 50 \\ 4 \end{bmatrix}$

Do mesmo modo,

$s_{11} = \frac{1}{4}\sum_{i=1}^{4}(x_{i1}-\bar{x}_1)^2 = \frac{1}{4}\left[(42-50)^2+(52-50)^2+(48-50)^2+(58-50)^2\right] = 34$

$s_{22} = \frac{1}{4}\sum_{i=1}^{4}(x_{i2}-\bar{x}_2)^2 = \frac{1}{4}\left[(4-4)^2+(5-4)^2+(4-4)^2+(3-4)^2\right] = .5$

$s_{12} = \frac{1}{4}\sum_{i=1}^{4}(x_{i1}-\bar{x}_1)\ (x_{i2}-\bar{x}_2) =$

$= \frac{1}{4}\left[(42-50)(4-4)+(52-50)(5-4)+(48-50)(4-4)+(58-50)(3-4)\right] = -1.5$

$S_n = \begin{bmatrix} 34 & -1.5 \\ -1.5 & .5 \end{bmatrix}$

Finalmente, a correlação amostral é dada por

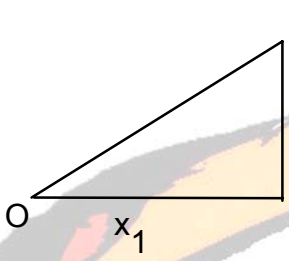
$r_{12} = r_{21} = \frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} = \frac{-1.5}{\sqrt{34}\sqrt{.5}} = -.36$

$R = \begin{bmatrix} 1 & -.36 \\ -.36 & 1 \end{bmatrix}$

o

1.3 Distâncias

A maioria das técnicas multivariadas são baseadas no conceito simples de distância. Se considerarmos um plano e um ponto P = (x_1,x_2) nesse plano, a distância d(O,P) entre a origem e esse ponto é dada por



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

Figura 1.1 – Teorema de Pitágoras

Num caso mais geral, se os pontos tiverem p coordenadas, então P = (x_1,x_2, …,x_p), O=(0,0,…,0) e $d(O,P) = \sqrt{x_1^2 + x_2^2 + \ldots + x_p^2}$

Desta última equação, e elevando ao quadrado ambos os termos, podemos dizer que todos os pontos (x_1,x_2, …,x_p) que estejam a uma mesma distância quadrada da origem, satisfazem a equação

$$d^2(O,P) = x_1^2 + x_2^2 + \ldots + x_p^2$$

Se se tratar de um espaço onde p=2 , esta equação não é mais do que a equação de uma circunferência de centro (0,0) e raio d(0,P).

A distância em linha recta entre dois pontos quaisquer P e Q com coordenadas P=(x_1,x_2, …,x_p) e Q=(y_1,y_2, …,y_p) é dada por

$$d(P,Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_p - y_p)^2}$$

Ora também aqui se faz sentir o eventual problema das várias dimensões terem unidades de medida distintas. Mais ainda, as medições das diversas coordenadas podem estar sujeitas a variações aleatórias com intensidades diferentes. Por isso, uma distância baseada numa linha recta, ou euclideana, não é a mais apropriada. Necessitamos então de um outro tipo de medição de distâncias e, porque este novo tipo de distância vai ter em conta as diferenças de variação, denomina-la-emos distância estatística.

Para ilustrar o conceito de distância estatística, suponhamos que temos n pares de medições em duas variáveis independentes x_1 e x_2. Além disso, suponhamos também que a variação das medições da variável x_1 é maior do que a das medições em x_2.

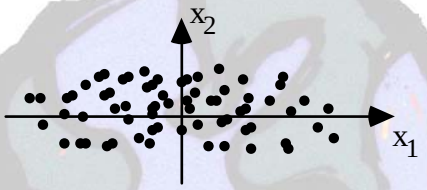


Figura 1.2 – Diagrama de pontos

Neste caso, a solução passa, de novo, pela estandardização das coordenadas, dividindo cada uma delas pelo respectivo desvio padrão amostral. Assim, uma distância estatística do ponto P=(x_1,x_2) à origem O=(0,0) é dada por

$$d(O,P) = \sqrt{\left(\frac{x_1}{\sqrt{s_{11}}}\right)^2 + \left(\frac{x_2}{\sqrt{s_{22}}}\right)^2} = \sqrt{\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}}}$$

Se compararmos esta equação com a anteriormente apresentada, podemos concluir que a diferença reside na aplicação de pesos $k_1 = 1/s_{11}$ e $k_2 = 1/s_{22}$,

respectivamente, a x_1^2 e x_2^2. Também aqui todos os pontos de coordenadas (x_1,x_2) a uma distância quadrada constante c^2 da origem devem satisfazer a

$$\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}} = c^2$$

Esta última equação não é mais do que a equação de uma elipse centrada na origem com os eixos principais a coincidirem com os eixos do sistema de coordenadas.

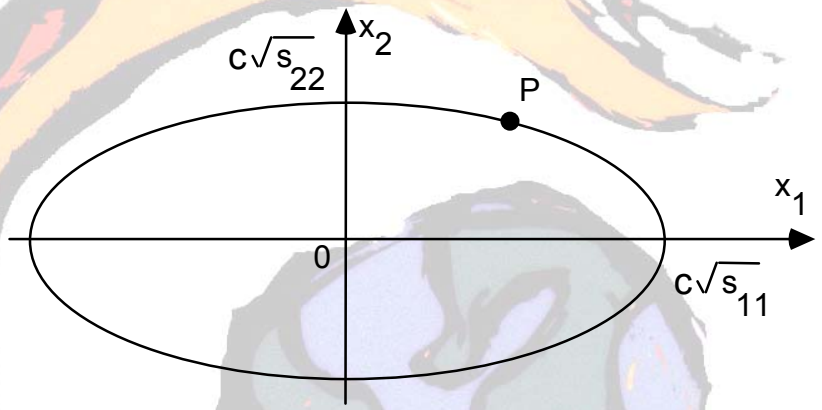


Figura 1.3 – Elipse centrada na origem

Exemplo 1.2: Suponhamos que temos duas variáveis independentes com médias $\bar{x}_1 = \bar{x}_2 = 0$ e com variâncias $s_{11} = 4$ e $s_{22} = 1$.

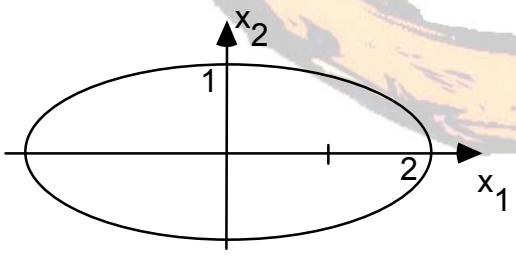


Figura 1.4 – Elipse

A distância de um qualquer ponto $P=(x_1,x_2)$ à origem $O=(0,0)$ é dada, neste caso por

$$d^2(O,P) = \frac{x_1^2}{4} + \frac{x_2^2}{1}$$

Todos os pontos (x_1,x_2) que estão a uma distância constante 1 da origem satisfazem a equação

$$\frac{x_1^2}{4} + \frac{x_2^2}{1} = 1$$

correspondendo à equação da elipse centrada em (0,0), com os eixos principais segundo os eixos x_1 e x_2 e com meias distâncias iguais a $\sqrt{4} = 2$ e $\sqrt{1} = 1$, respectivamente.

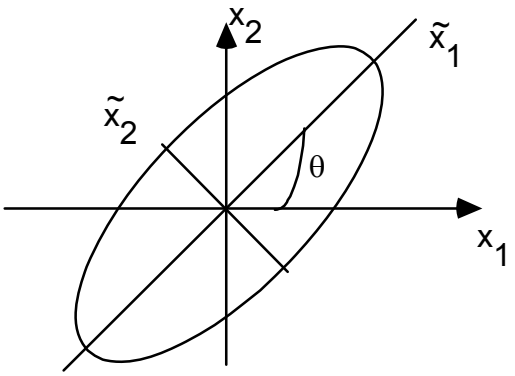
o

Generalizando para p variáveis, podemos determinar a distância estatística entre dois pontos $P=(x_1,x_2, \ldots,x_p)$ e $Q=(y_1,y_2, \ldots,y_p)$ através da equação

$$d(P,Q) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \frac{(x_2 - y_2)^2}{s_{22}} + \ldots + \frac{(x_p - y_p)^2}{s_{pp}}}$$

com $s_{11}, s_{22}, \ldots, s_{pp}$ as variâncias construídas a partir das n medições nas variáveis $x_1,x_2, \ldots,x_p$, respectivamente. Todos os pontos P a uma distância quadrada de Q estão colocados num hiperelipsóide centrado em Q com os eixos principais paralelos aos eixos do sistema de coordenadas. Obviamente, se todas as variâncias fossem iguais, encontramos a distância euclideana já atrás apresentada.

Temos até agora analisado a situação em que os eixos da elipse dos dados coincidem com os eixos do sistema de coordenadas. Ora, há situações onde isto não acontece, isto é, em que a variável x_1 não varia independentemente da variável x_2 e, neste caso, o coeficiente de correlação amostral não é nulo.



Da figura ao lado vemos que basta rodarmos o sistema original de eixos de um ângulo θ para termos uma situação semelhante às anteriores.

Figura 1.5 – Elipse com ângulo θ

Isto corresponde a passarmos a usar as novas variáveis

$\tilde{x}_1 = x_1 \cos(\theta) + x_2 \sin(\theta)$ $\qquad$ $\tilde{x}_2 = - x_1 \sin(\theta) + x_2 \cos(\theta)$

A distância entre o ponto P=($\tilde{x}_1$,$\tilde{x}_2$) e a origem O=(0,0) é então definida como

$$d(O,P) = \sqrt{\frac{\tilde{x}_1^2}{\tilde{s}_{11}} + \frac{\tilde{x}_2^2}{\tilde{s}_{22}}} = \sqrt{a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2}$$

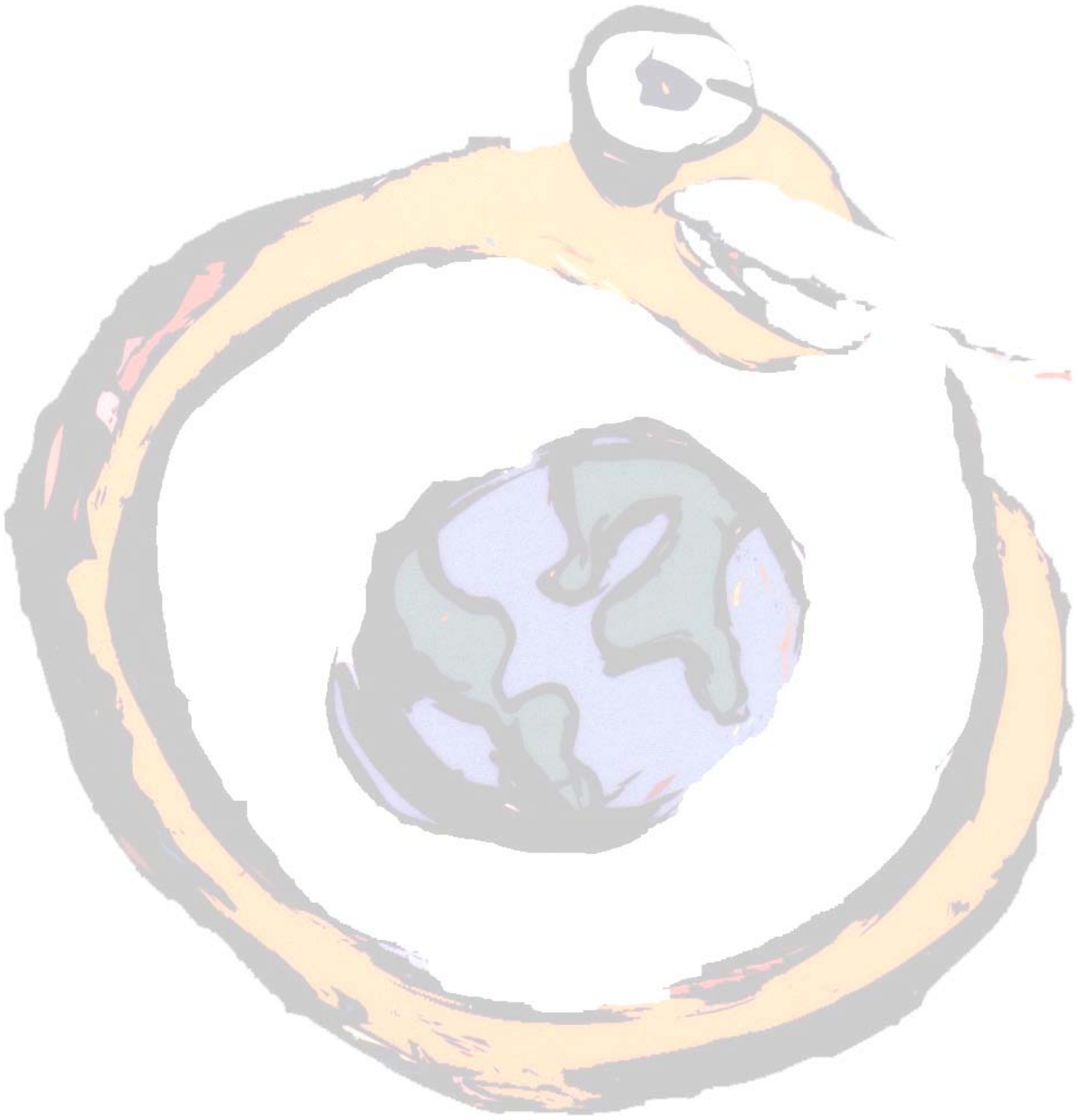
Nesta fase não é vital sabermos como determinar os valores destes a´s. O que é importante é vermos que existe um termo de produto cruzado indicador da correlação r_{12} não nula. Mais ainda, quando olhamos para a equação correspondente às duas variáveis independentes, vemos que

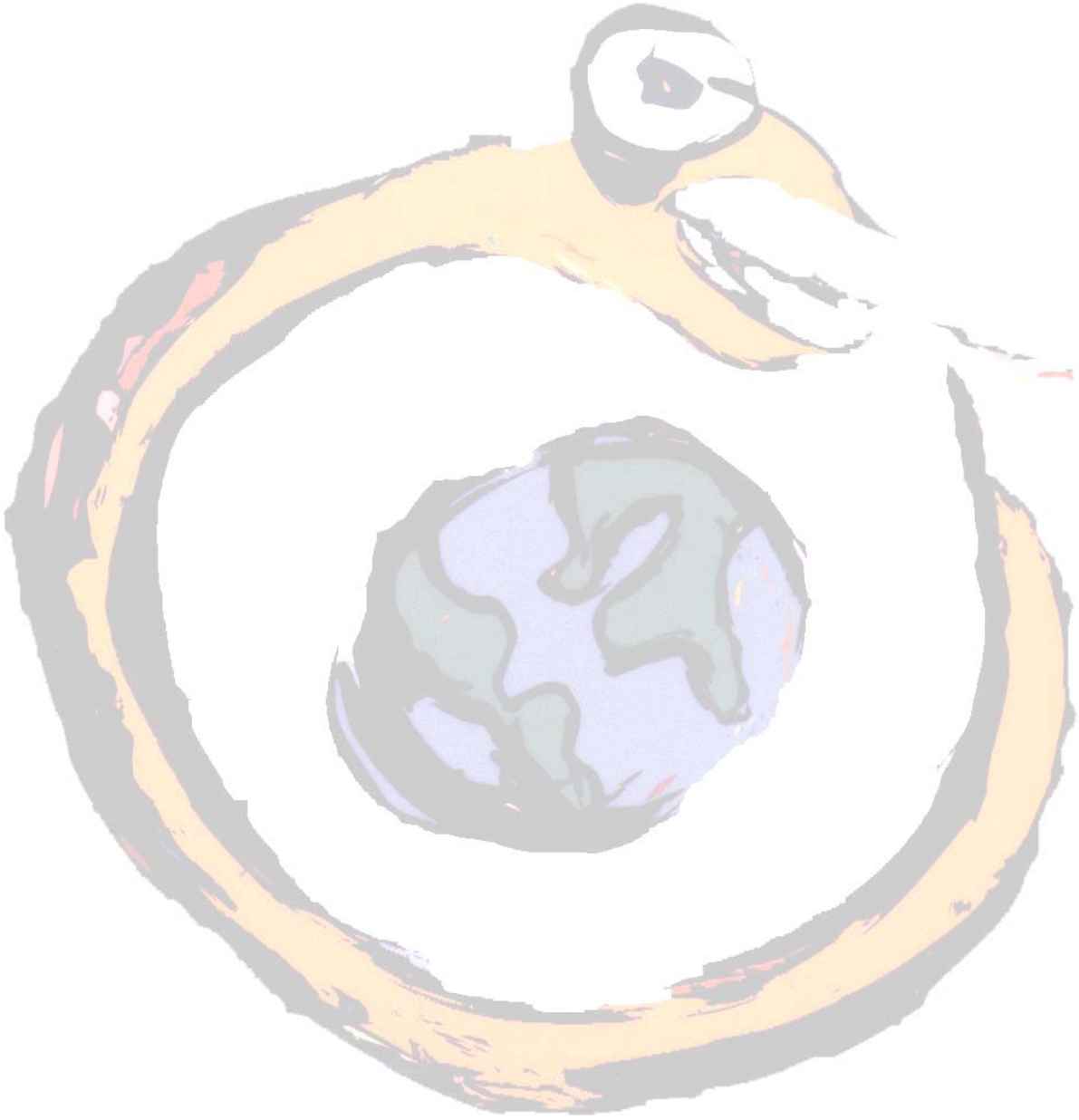
$$a_{11} = \frac{1}{s_{11}} \qquad a_{22} = \frac{1}{s_{22}} \qquad a_{12} = 0$$

De uma maneira geral, a distância estatística do ponto P=(x_1,x_2) ao ponto fixo Q=(y_1,y_2) para variáveis correlacionadas é dada por

$$d(P,Q) = \sqrt{a_{11}(x_1 - y_1)^2 + 2a_{12}(x_1 - y_1)(x_2 - y_2) + a_{22}(x_2 - y_2)^2}$$

As coordenadas de todos os pontos $P=(x_1,x_2)$ que estejam a uma distância quadrada constante c^2 de Q, definem uma elipse centrada em Q. A generalização das fórmulas anteriores para p dimensões é imediata.





2

Álgebra matricial e vectores aleatórios

2.1 Alguns conceitos básicos

Vejamos alguns conceitos que nos irão ser úteis mais tarde.

Sendo dado um vector $\mathbf{x}' = [\, x_1, x_2, \ldots, x_n \,]$ com n componentes, definimos comprimento deste vector como sendo o valor dado por

$$L_{\mathbf{x}} = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$$

Assim, pré-multiplicando **x** pelo inverso do seu comprimento, $L_{\mathbf{x}}^{-1}\mathbf{x}$, obtém-se o vector unitário (com comprimento 1) e com a mesma direcção de **x**.

Um outro conceito também importante é o de ângulo. Se tivermos dois vectores num plano com um ângulo θ entre eles, podemos considerar que $\theta = \theta_2 - \theta_1$, sendo θ_1 e θ_2 os ângulos que, respectivamente, **x** e **y** fazem com a primeira coordenada (ver Figura 2.1).

Assim, sabendo que

$$\cos(\theta_1) = \frac{x_1}{L_{\mathbf{x}}} \qquad\qquad \cos(\theta_2) = \frac{y_1}{L_{\mathbf{y}}}$$

$$\sin(\theta_1) = \frac{x_2}{L_x} \qquad \sin(\theta_2) = \frac{y_2}{L_y}$$

e que $\cos(\theta) = \cos(\theta_2 - \theta_1) = \cos(\theta_2)\cos(\theta_1) + \sin(\theta_2)\sin(\theta_1)$

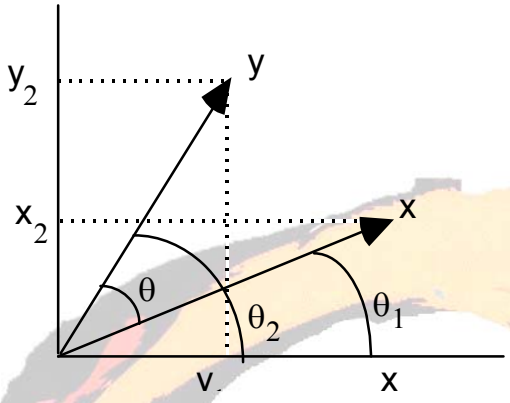


Figura 2.1 – Diferença entre ângulos

obtemos

$$\cos(\theta) = \cos(\theta_1 - \theta_2) = \left(\frac{y_1}{L_y}\right)\left(\frac{x_1}{L_x}\right) + \left(\frac{y_2}{L_y}\right)\left(\frac{x_2}{L_x}\right) = \frac{x_1y_1 + x_2y_2}{L_xL_y}$$

Como o produto interno de dois vectores é dado por $\mathbf{x'y} = x_1y_1 + x_2y_2$ podemos re-escrever as equações referentes a L_x e a $\cos(\theta)$ da seguinte maneira:

$$L_x = \sqrt{\mathbf{x'x}} \qquad \text{e} \qquad \cos(\theta) = \frac{\mathbf{x'y}}{L_xL_y} = \frac{\mathbf{x'y}}{\sqrt{\mathbf{x'x}}\sqrt{\mathbf{y'y}}}$$

Deste modo, dizemos que **x** e **y** são perpendiculares quando $\mathbf{x'y} = \mathbf{0}$.

Exemplo 2.1: Sendo dados os vectores $\mathbf{x}' = [\,1, 3, 2\,]$ e $\mathbf{y}' = [\,-2, 1, -1\,]$, determinar o valor do comprimento de **x** e de **y** e o ângulo que eles fazem entre si.

Como $\mathbf{x}'\mathbf{x} = 1^2 + 3^2 + 2^2 = 14$

$\mathbf{y}'\mathbf{y} = (-2)^2 + 1^2 + (-1)^2 = 6$

$\mathbf{x}'\mathbf{y} = 1(-2) + 3(1) + 2(-1) = -1$

então $L_{\mathbf{x}} = \sqrt{\mathbf{x}'\mathbf{x}} = \sqrt{14} = 3.74$

$L_{\mathbf{y}} = \sqrt{\mathbf{y}'\mathbf{y}} = \sqrt{6} = 2.45$

$\cos(\theta) = \frac{\mathbf{x}'\mathbf{y}}{L_x L_y} = \frac{-1}{(3.74)(2.45)} = -.109$, donde, $\theta = 96.3^\circ$

o

Diz-se que um conjunto de vectores $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_k}$ é linearmente dependente se existirem as constantes $c_1, c_2, \ldots, c_k$, não todas nulas, tal que

$$c_1 \mathbf{x_1} + c_2 \mathbf{x_2} + \ldots + c_k \mathbf{x_k} = \mathbf{0}$$

Exemplo 2.2: Determinar a dependência linear dos vectores $\mathbf{x_1}' = [\,1, 2, 1\,]$, $\mathbf{x_2}' = [\,1, 0, -1\,]$ e $\mathbf{x_3}' = [\,1, -2, 1\,]$.

A equação $c_1 \mathbf{x_1} + c_2 \mathbf{x_2} + c_3 \mathbf{x_3} = \mathbf{0}$ implica o sistema

$$\begin{cases} c_1 + c_2 + c_3 = 0 \\ 2c_1 \quad - 2c_3 = 0 \\ c_1 - c_2 + c_3 = 0 \end{cases}$$

que possui uma única solução $c_1 = c_2 = c_3 = 0$.

Neste caso, dizemos que os vectores $\mathbf{x_1}$, $\mathbf{x_2}$ e $\mathbf{x_3}$ são linearmente independentes.

o

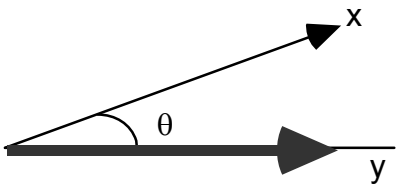


Figura 2.2 – Projecção de **x** em **y**

A projecção (ou sombra) de um vector **x** num vector **y** é dada por

$$\frac{\mathbf{x}'\mathbf{y}}{\mathbf{y}'\mathbf{y}}\mathbf{y} = \frac{\mathbf{x}'\mathbf{y}}{L_y}\frac{1}{L_y}\mathbf{y}$$

tendo $L_y^{-1}\mathbf{y}$, o comprimento unitário. O comprimento desta projecção é

$$\frac{|\mathbf{x}'\mathbf{y}|}{L_y} = L_{\mathbf{x}}\left|\frac{\mathbf{x}'\mathbf{y}}{L_x L_y}\right| = L_{\mathbf{x}}\,|\cos(\theta)|$$

O último conceito muito usado na estatística multivariada é o de valor próprio e vector próprio. Uma matriz quadrada A tem um valor próprio λ com o correspondente vector próprio $\mathbf{x} \neq \mathbf{0}$ se

$$A\ \mathbf{x} = \lambda\ \mathbf{x}$$

Isto é, os valores próprios são as raízes da equação característica $|A - \lambda I| = 0$.

Exemplo 2.3: Determinar os valores e vectores próprios da matriz $\mathbf{A} = \begin{bmatrix} 1 & -5 \\ -5 & 1 \end{bmatrix}$

$$|\mathbf{A} - \lambda\,\mathbf{I}| = \mathbf{0} \Rightarrow \begin{vmatrix} 1-\lambda & -5 \\ -5 & 1-\lambda \end{vmatrix} = \mathbf{0} \Rightarrow (1-\lambda)^2 - 25 = 0 \Rightarrow \lambda_1 = 6 \text{ ou } \lambda_2 = -4$$

Para λ_1=6, $\mathbf{A}\,\mathbf{e} = \lambda_1\,\mathbf{e} \Rightarrow \begin{bmatrix} 1 & -5 \\ -5 & 1 \end{bmatrix}\begin{bmatrix} e_{11} \\ e_{21} \end{bmatrix} = 6 \begin{bmatrix} e_{11} \\ e_{21} \end{bmatrix}$

$$\begin{cases} e_{11} - 5e_{21} = 6e_{11} \\ -5e_{11} + e_{21} = 6e_{21} \end{cases} \Rightarrow \begin{cases} e_{11} = \dfrac{1}{\sqrt{2}} \\ e_{21} = \dfrac{-1}{\sqrt{2}} \end{cases}$$

$\mathbf{e}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix}$ é um vector próprio normalizado correspondente ao valor próprio λ_1=6.

De modo idêntico se encontra $\mathbf{e}_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$ como sendo o vector próprio correspondente a λ_2 = -4.

o

2.2 Matrizes definidas positivas

Dois dos pilares fundamentais da estatística multivariada são o conceito de distância e o pressuposto de que os dados estão distribuídos segundo uma distribuição normal multivariada. Os produtos de matrizes resultantes da combinação destes conceitos são denominados formas quadráticas. Assim, neste capítulo iremos falar em particular sobre as formas quadráticas não negativas e as matrizes definidas positivas associadas.

Muitas vezes, também, os resultados que envolvem formas quadráticas e matrizes simétricas são consequência directa do que se denomina decomposição espectral definida numa matriz simétrica $A_{k\times k}$ definida como

$$\underset{(k\times k)}{\mathbf{A}} = \underset{(k\times 1)\,(1\times k)}{\lambda_1\ \mathbf{e}_1\ \mathbf{e}_1'} + \underset{(k\times 1)(1\times k)}{\lambda_2\ \mathbf{e}_2\ \mathbf{e}_2'} + \ldots + \underset{(k\times 1)(1\times k)}{\lambda_k\ \mathbf{e}_k\ \mathbf{e}_k'}$$

onde $\lambda_1, \lambda_2, \ldots, \lambda_k$ são os valores próprios de A e $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k$ os correspondentes vectores próprios normalizados, isto é, $\mathbf{e}_i'\ \mathbf{e}_i = 1$ (i = 1, 2, …, k) e $\mathbf{e}_i'\ \mathbf{e}_j = 0$ (i ≠ j).

Exemplo 2.4: Sendo dada a matriz $\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$, obtêm-se os valores próprios $\lambda_1 = 4$ e $\lambda_2 = 2$. O vector próprio correspondente ao primeiro valor próprio é $\mathbf{e}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

Tornamo-lo único, normalizando-o (comprimento igual à unidade), isto é, dividindo cada elemento do vector por $\sqrt{e_{11}^2 + e_{21}^2} = \sqrt{1^2 + 1^2} = \sqrt{2}$

Encontra-se $\mathbf{e}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$. Do mesmo modo se obtinha $\mathbf{e}_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix}$.

Reparar que $\mathbf{e}_1 \perp \mathbf{e}_2$, isto é, $\mathbf{e}_1'\ \mathbf{e}_2 = 0$.

Verificando a decomposição espectral,

$$\begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} = 4 \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} + 2 \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix} =$$

$$= 4\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} + 2\begin{bmatrix} \frac{1}{2} & \frac{-1}{2} \\ \frac{-1}{2} & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} + \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$$

o

Sempre que a matriz **A** (k×k) simétrica seja tal que **x´A x** seja sempre maior ou igual a zero, qualquer que seja o vector $\mathbf{x}' = [x_1 \quad x_2 \quad \dots \quad x_n] \neq [0 \quad 0 \quad \dots \quad 0]$, denominamo-la definida não-negativa ou semi-definida positiva. **A** é chamada definida positiva se **x´A x** > 0 para todo o vector **x** ≠ **0**. À componente **x´A x** damos o nome de forma quadrática.

Para k = 2,

$$\mathbf{x'A\,x} = [x_1 \quad x_2]\begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = [x_1 \quad x_2]\begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{12}x_1 + a_{22}x_2 \end{bmatrix}$$

$$= a_{11}x_1^2 + a_{12}x_1x_2 + a_{12}x_1x_2 + a_{22}x_2^2 = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2$$

$$= d^2(0,\mathbf{x}) = c^2$$

Pela decomposição espectral; $\mathbf{A} = \lambda_1 \mathbf{e}_1 \mathbf{e}_1' + \lambda_2 \mathbf{e}_2 \mathbf{e}_2'$

e então $\mathbf{x'A\,x} = \lambda_1 (\mathbf{x'e}_1)^2 + \lambda_2 (\mathbf{x'e}_2)^2$.

Assim; $c^2 = \lambda_1 y_1^2 + \lambda_2 y_2^2$ é uma elipse em $y_1 = \mathbf{x'e}_1$ e $y_2 = \mathbf{x'e}_2$

Facilmente se verifica que $\mathbf{x} = c\,\lambda_1^{-1/2}\mathbf{e}_1$ satisfaz $\mathbf{x'A\,x} = \lambda_1 (c\,\lambda_1^{-1/2}\mathbf{e}_1'\mathbf{e}_1)^2 = c^2$

e $\mathbf{x} = c\,\lambda_2^{-1/2}\,\mathbf{e}_2$ nos dá a distância na direcção $\mathbf{e}_2$

Deste modo os pontos situados a uma distância c fazem parte de uma elipse cujos eixos são dados pelos vectores próprios de **A** com comprimentos proporcionais aos inversos das raízes quadradas dos valores próprios. A constante de proporcionalidade é c.

Esta conclusão é ilustrada na figura abaixo.

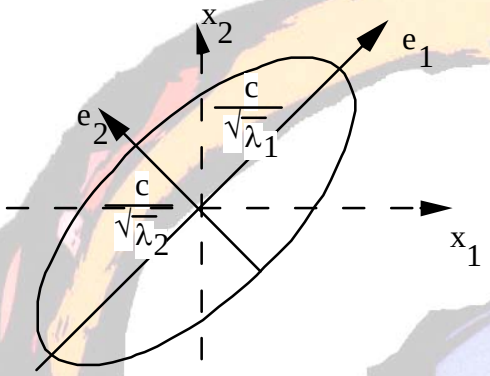


Figura 2.3 – Elipse de distância constante

Com p > 2, os pontos $\mathbf{x}' = [x_1 \quad x_2 \quad \ldots \quad x_p]$ a uma distância constante

$c = \sqrt{\mathbf{x}'\mathbf{A}\mathbf{x}}$ da origem encontram-se no elipsóide

$$c^2 = \lambda_1 (\mathbf{x}'\mathbf{e}_1)^2 + \ldots + \lambda_p (\mathbf{x}'\mathbf{e}_p)^2$$

cujos eixos são dados pelos vectores próprios de **A** . A meia distância na direcção de $\mathbf{e}_i$ é igual a $\frac{c}{\sqrt{\lambda_i}}$, i = 1, 2, …, p, onde $\lambda_1, \lambda_2, \ldots, \lambda_p$, são os valores próprios de **A**.

2.3 Médias e covariâncias de combinações lineares

Um vector aleatório é um vector cujos elementos são variáveis aleatórias. Do mesmo modo, uma matriz aleatória é uma matriz cujos elementos são variáveis aleatórias.

A combinação linear $\mathbf{c}'\mathbf{X} = c_1X_1 + \ldots + c_pX_p$ tem

média $E(\mathbf{c}'\mathbf{X}) = \mathbf{c}'\boldsymbol{\mu}$

e variância $Var(\mathbf{c}'\mathbf{X}) = \mathbf{c}'\sum\mathbf{c}$

onde $\boldsymbol{\mu} = E(\mathbf{X})$ e $\sum = Cov(\mathbf{X}) = E\left[(X-\mu)(X-\mu)'\right]$

Exemplo 2.5: Consideremos a matriz $\mathbf{X'} = \begin{bmatrix} 2 & 3 & 1 \\ -2 & 5 & 0 \end{bmatrix}'$

A média desta matriz é $\boldsymbol{\mu} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$

e a matriz das covariâncias é $\sum = \begin{bmatrix} 2/3 & 2/3 \\ 2/3 & 26/3 \end{bmatrix}$

Assim, a combinação linear $Y = 3X_1 + 2X_2$, isto é, $\begin{bmatrix} 3 & 2 \end{bmatrix}\begin{bmatrix} 2 & 3 & 1 \\ -2 & 5 & 0 \end{bmatrix}$,

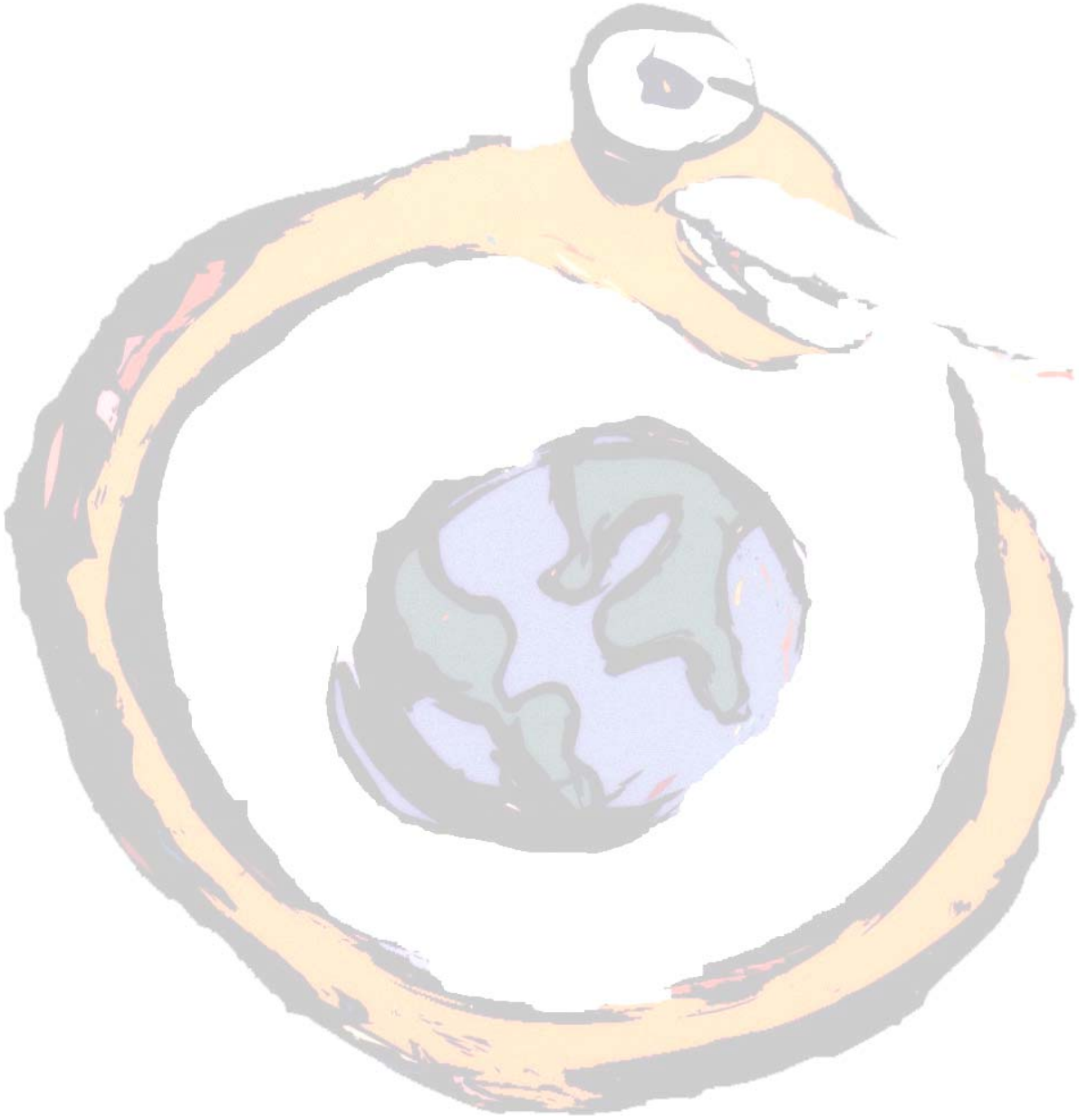
terá a média $E(\mathbf{Y}'\mathbf{X}) = \begin{bmatrix} 3 & 2 \end{bmatrix}\begin{bmatrix} 2 \\ 1 \end{bmatrix} = 8$

e a variância $Var(\mathbf{Y}'\mathbf{X}) = \begin{bmatrix} 3 & 2 \end{bmatrix}\begin{bmatrix} 2/3 & 2/3 \\ 2/3 & 26/3 \end{bmatrix}\begin{bmatrix} 3 \\ 2 \end{bmatrix} = 48.67$

o

Além dos resultados anteriores podemos também afirmar que, sendo dado duas combinações lineares **a´X** e **b´X**, a covariância entre elas é dada por

$$\text{Cov}(\mathbf{a}'\mathbf{X},\mathbf{b}'\mathbf{X}) = \mathbf{a}' \sum \mathbf{b}$$



3

Geometria amostral e amostragem aleatória

Neste capítulo iremos analisar as interpretações geométricas das estatísticas descritivas amostrais $\bar{x}$, S_n e R. Será também introduzido o conceito de variância generalizada para descrever a variabilidade.

3.1 Geometria da amostra

Tal como já atrás vimos, as n observações em p variáveis podem ser dispostas numa matriz n×p

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} x'_1 \\ x'_2 \\ \cdot \\ \cdot \\ \cdot \\ x'_n \end{bmatrix}$$

onde cada linha representa uma observação multivariada (vector x_i , i= 1, n).

Assim, a variabilidade ocorre em várias direcções e é quantificada através da matriz S_n das variâncias. Um valor numérico desta variabilidade é dado pelo determinante de S_n.

Exemplo 3.1: Determinar o vector média $\bar{\mathbf{x}}$ da matriz $X' = \begin{bmatrix} 4 & -1 & 3 \\ 1 & 3 & 5 \end{bmatrix}'$, apresente os n = 3 pontos num espaço a p = 2 dimensões e localize $\bar{\mathbf{x}}$.

$$\bar{\mathbf{x}} = \begin{bmatrix} \frac{4-1+3}{3} \\ \frac{1+3+5}{3} \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

O gráfico de pontos correspondente será,

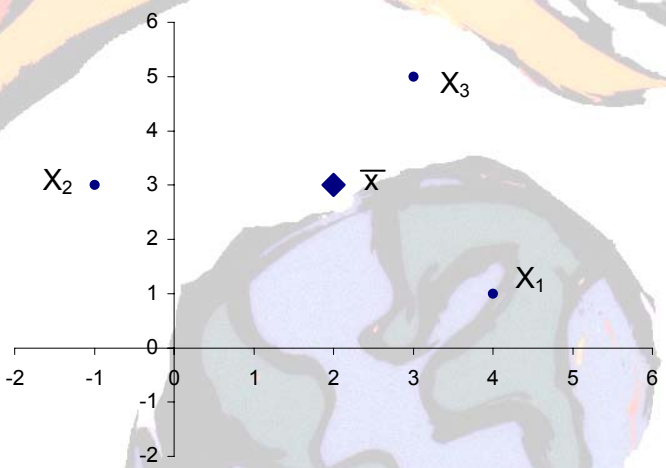


Figura 3.1 – Representação dos pontos x_1, x_2, x_3 e médio

o

Em alternativa a esta interpretação geométrica, podemos considerar os dados como sendo p pontos num espaço a n dimensões.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = [y_1 \quad y_2 \quad \cdots \quad y_p]$$

Nesta nova interpretação, as coordenadas do i-ésimo ponto $y_i = [x_{1i}, x_{2i}, \ldots, x_{ni}]$ são as n medições da i-ésima variável.

Exemplo 3.2: Usando a mesma matriz do exemplo anterior, representar o vectores y'_1 e y'_2.

$$y'_1 = [4 \quad -1 \quad 3] \qquad y'_2 = [1 \quad 3 \quad 5]$$

O gráfico de pontos correspondente será,

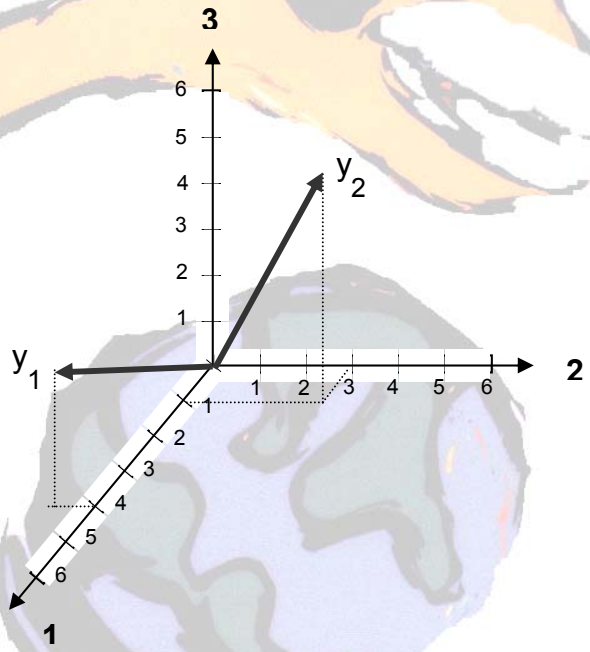


Figura 3.2 – Representação dos vectores y_1 e y_2

o

Também é possível dar-se uma interpretação geométrica ao processo de determinação da média amostral. Para isso começamos por definir o vector $n \times 1$

$$\mathbf{1}'_n = \mathbf{1}' = [1 \quad 1 \quad \ldots \quad 1]$$

que, por definição, forma ângulos iguais com cada uma das n coordenadas.

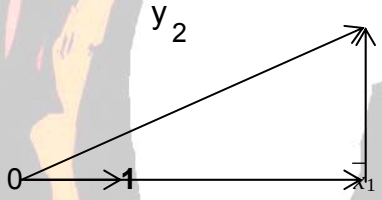
Deste modo, $\frac{1}{\sqrt{n}}\mathbf{1}$ tem comprimento unitário e direcção do ângulo igualitário.

A projecção de $\mathbf{y}_i$ no vector unitário é dada por

$$\mathbf{y}'_i \left(\frac{1}{\sqrt{n}}\mathbf{1}\right)\frac{1}{\sqrt{n}}\mathbf{1} = \frac{x_{i1}+x_{i2}+\ldots+x_{in}}{n}\mathbf{1} = \bar{x}_i \mathbf{1}$$

isto é, a média amostral $\bar{x}_i = \mathbf{y}'_i \mathbf{1}/n$ corresponde ao múltiplo de **1** necessário para obter a projecção de $\mathbf{y}_i$ na linha determinada por **1**.

Além disso, para cada $\mathbf{y}_j$ podemos determinar o vector desvio $\mathbf{d}_j$, desvio entre $\mathbf{y}_j$ e $\bar{x}_i \mathbf{1}$.



$$\mathbf{d}_j = \mathbf{y}_j - \bar{x}_j \mathbf{1} = \begin{bmatrix} x_{1i}-\bar{x}_i \\ x_{2i}-\bar{x}_i \\ \cdots \\ x_{ni}-\bar{x}_i \end{bmatrix}$$

Figura 3.3 – Diferença entre vectores

Exemplo 3.3: Ainda com a mesma matriz **X**,

$$\bar{x}_1 \mathbf{1} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} \qquad \bar{x}_2 \mathbf{1} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix}$$

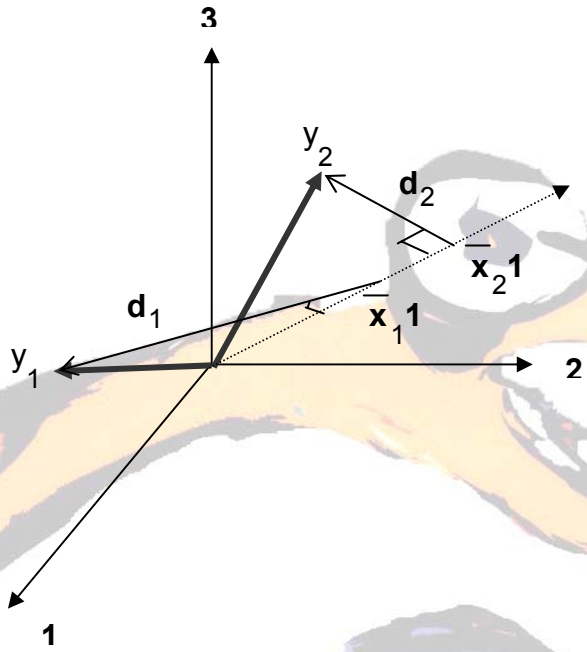
Consequentemente,

$$\mathbf{d}_1 = \mathbf{y}_1 - \bar{x}_1 \mathbf{1} = \begin{bmatrix} 4 \\ -1 \\ 3 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix}$$

$$\mathbf{d}_2 = \mathbf{y}_2 - \bar{x}_2 \mathbf{1} = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} - \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} = \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix}$$

Figura 3.4 – Vectores desvios

o



É fácil ver que

$$L_{d_i}^2 = \mathbf{d}_i' \mathbf{d}_i = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

isto é, o quadrado do comprimento do vector desvio é igual à soma dos quadrados dos desvios.

Do mesmo modo,

$$\mathbf{d}_i' \mathbf{d}_k = \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k) = L_{d_i} L_{d_k} \cos(\theta_{ik})$$

e então,

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \cos(\theta_{ik})$$

O coseno do ângulo é o coeficiente de correlação amostral. Assim, se dois vectores tiverem aproximadamente a mesma orientação, a correlação amostral será próxima da unidade. Se estes dois vectores forem quase perpendiculares, a

correlação amostral é quase nula. Se os dois vectores estiverem orientados aproximadamente em direcções opostas, a correlação amostral será próxima de -1.

Exemplo 3.4: Com os resultados dos exemplos anteriores,

$$\mathbf{d}_1'\,\mathbf{d}_1 = \begin{bmatrix} 2 & -3 & 1 \end{bmatrix}\begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix} = 14 = 3\,s_{11}$$

$$\mathbf{d}_2'\,\mathbf{d}_2 = \begin{bmatrix} -2 & 0 & 2 \end{bmatrix}\begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix} = 8 = 3\,s_{22}$$

$$\mathbf{d}_1'\,\mathbf{d}_2 = \begin{bmatrix} 2 & -3 & 1 \end{bmatrix}\begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix} = -2 = 3\,s_{12}$$

$$S_n = \begin{bmatrix} \frac{14}{3} & \frac{-2}{3} \\ \frac{-2}{3} & \frac{8}{3} \end{bmatrix}$$

$$r_{12} = \frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} = \frac{\frac{-2}{3}}{\sqrt{\frac{14}{3}}\sqrt{\frac{8}{3}}} = -.189$$

$$R = \begin{bmatrix} 1 & -.189 \\ -.189 & 1 \end{bmatrix}$$

□

3.2 Amostragem aleatória

Para estudarmos a variabilidade amostral de $\bar{x}$ e S_n e para podermos inferir os resultados para toda a população, temos de estabelecer alguns pressupostos relativamente às variáveis que constituem o conjunto das observações.

Dada a matriz

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} x'_1 \\ x'_2 \\ . \\ . \\ . \\ x'_n \end{bmatrix}$$

dizemos que $\mathbf{x}_1$, $\mathbf{x}_2$, … , $\mathbf{x}_n$ formam uma amostra aleatória se constituírem observações independentes, possuindo uma distribuição conjunta $f(\mathbf{x}) = f(\mathbf{x}_1)\, f(\mathbf{x}_2) \dots f(\mathbf{x}_n)$.

Se $\boldsymbol{\mu}$ e $\sum$ representarem, respectivamente, o vector média e a matriz de variâncias da amostra aleatória $\mathbf{x}_1$, $\mathbf{x}_2$, … , $\mathbf{x}_n$, então $\bar{\mathbf{x}}$ é um estimador não enviesado de $\boldsymbol{\mu}$ $[E(\bar{\mathbf{x}}) = \boldsymbol{\mu}]$ e $S = \frac{n}{n-1} S_n$ é um estimador não enviesado de $\sum$, isto é, $E(\frac{n}{n-1} S_n) = \sum$.

A matriz amostral não enviesada das variâncias é

$$S = \frac{n}{n-1} S_n = \frac{1}{n-1} \sum_{j=1}^{n} (\mathbf{x}_j - \bar{\mathbf{x}})\ (\mathbf{x}_j - \bar{\mathbf{x}})'$$

3.3 Variância generalizada

A variância é normalmente descrita pela matriz das variâncias

$$S = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} = \left\{ s_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{\mathbf{x}}_j)\ (x_{ik} - \bar{\mathbf{x}}_k) \right\}$$

Um único valor numérico que representa toda a variação expressa em S é a variância amostral generalizada dada pelo determinante de S.

Variância amostral generalizada = | S |

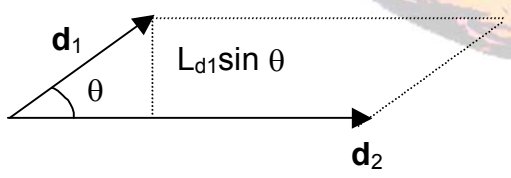
Exemplo 3.5: Consideremos a matriz $S = \begin{bmatrix} 14808 & 14213 \\ 14213 & 15538 \end{bmatrix}$

A variância generalizada é dada por

| S | = (14808) (15538) - (14213) (14213) = 28.08×10^6.

□

Vejamos de seguida uma interpretação geométrica para |S|. Consideremos então a área gerada pelos dois vectores desvio $\mathbf{d}_1 = \mathbf{y}_1 - \bar{x}_1\ \mathbf{1}$ e $\mathbf{d}_2 = \mathbf{y}_2 - \bar{x}_2\ \mathbf{1}$



Área = $\left[L_{d_1} \sin(\theta)\right] L_{d_2}$

= $L_{d_1} L_{d_2} \sqrt{1 - \cos^2 \theta}$

= (n - 1) $\sqrt{s_{11} s_{22} (1 - r_{12}^2)}$

Figura 3.5 – Área gerada pelos desvios

Por outro lado,

$$|S| = \begin{vmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{vmatrix} = \begin{vmatrix} s_{11} & \sqrt{s_{11}}\sqrt{s_{22}}\, r_{12} \\ \sqrt{s_{11}}\sqrt{s_{22}}\, r_{12} & s_{22} \end{vmatrix}$$

$$= s_{11}\, s_{22} - s_{11}\, s_{22}\, r_{12}^2 = s_{11}\, s_{22}\, (1 - r_{12}^2)$$

Destes dois últimos resultados, podemos concluir que

$$|S| = \frac{\text{área}^2}{(n-1)^2} = (n-1)^{-2}\ \text{área}^2$$

Generalizando para um p-espaço obtemos

$$\text{Variância amostral generalizada} = |S| = (n-1)^{-p}\ (\text{volume})^2$$

isto é, para um determinado conjunto de dados, a variância amostral generalizada é proporcional ao quadrado do volume gerado pelos p vectores desvio.

As duas figuras abaixo representam, respectivamente, uma grande e uma pequena variância amostral generalizada para p = 3 no espaço das observações.

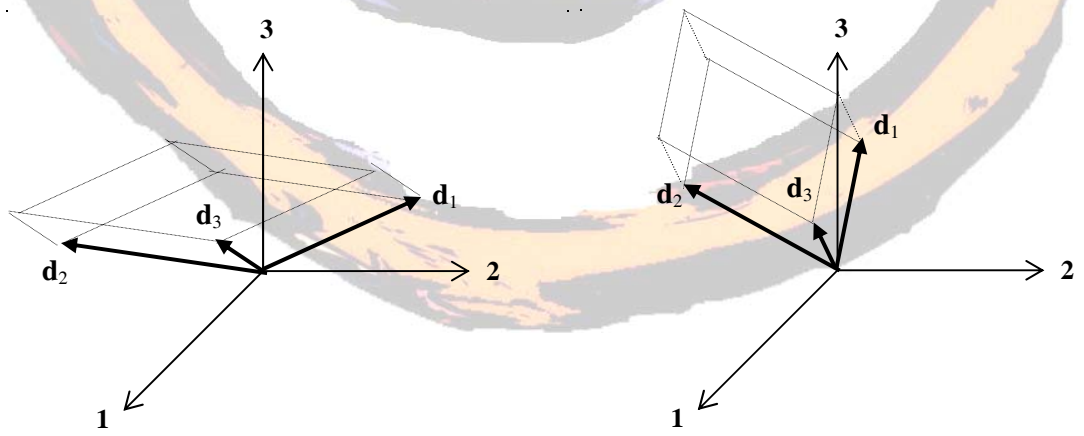


Figura 3.6 - Representação geométrica da variância generalizada

A variância generalizada tem também interpretação no gráfico de pontos num p-espaço. Consideremos, para isso, a média amostral $\bar{\mathbf{x}}' = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p]$.

As coordenadas $x' = [x_1, x_2; \ldots, x_p]$ dos pontos a uma distância constante c de $\bar{\mathbf{x}}$ satisfazem

$$(\mathbf{x} - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}}) = c^2$$

que define uma elipse (p = 2) centrada em $\bar{\mathbf{x}}$.

Usando o cálculo integral, podemos verificar que o volume do hiper-elipsóide está relacionado com o valor de | S |

$$\text{Volume de } \{x : (\mathbf{x} - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}}) = c^2\} = k_p \, |S|^{1/2} \, c^p$$

ou

(volume do elipsóide)^2 = (constante) (variância amostral generalizada)

Apesar da sua interpretação geométrica, a variância amostral generalizada é limitada como indicador descritivo de uma matriz amostral de variâncias. Para ilustrar isto vejamos o exemplo que se segue.

Exemplo 3.6: Consideremos as matrizes

$$S = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix} \qquad S = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix} \qquad S = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$

todas elas com a mesma variância generalizada | S | = 9 mas com distintos coeficientes de correlação, respectivamente, .8, -.8 e 0.

o

Ora, prova-se que o determinante de uma qualquer matriz **A** p×p pode ser escrito como o produto dos seus valores próprios $\lambda_1, \lambda_1, \ldots, \lambda_p$, isto é, $|\mathbf{A}| = \prod_{i=1}^{p} \lambda_i$.

Assim, os valores próprios podem dar-nos informação referente à variabilidade em todas as direcções numa representação p-espacial e, por isso, é útil não só analisarmos os valores individuais assim como o seu produto.

A variância generalizada é nula quando e apenas quando pelo menos um vector desvio estiver no hiperplano formado por todas as combinações lineares dos outros, isto é, quando as linhas de uma matriz de desvios forem linearmente dependentes.

Exemplo 3.7: Dada a matriz $X = \begin{bmatrix} 1 & 4 & 4 \\ 2 & 1 & 0 \\ 5 & 6 & 4 \end{bmatrix}$,

a matriz das médias é $\bar{\mathbf{x}}' = [\,3\,,\,1\,,\,5\,]$ e então $X - \bar{x}\ \mathbf{1}' = \begin{bmatrix} -2 & 1 & 1 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}$.

Os desvios residuais são $\mathbf{d}_1 = \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}$, $\mathbf{d}_2 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$ e $\mathbf{d}_3 = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$.

Como $\mathbf{d}_3 = \mathbf{d}_1 + 2\,\mathbf{d}_2$, há degenerescência nas linhas e |S| = 0, pois o volume a três dimensões formado pelos três vectores é nulo.

o

$|S| = 0$ significa, em termos matemáticos, que as medições em algumas variáveis podem ser retiradas do estudo. Por outro lado $|S|$ também será nulo se o tamanho da amostra for menor ou igual ao número de variáveis, isto é, $n \leq p$.

Se estivermos a trabalhar com variáveis estandardizadas, podemos dizer que a variância amostral generalizada é dada pelo determinante de R:

$$\left(\begin{array}{c} \text{Variância amostral generalizada} \\ \text{das variáveis estandardizadas} \end{array} \right) = |R| = (n-1)^{-p} \ (\text{volume})^2$$

Como $|S|$ e $|R|$ estão relacionadas por $|S| = (s_{11}\, s_{22} \ldots s_{pp})\ |R|$, podemos escrever

$$(n-1)^p\ |S| = (n-1)^p\ (s_{11}\, s_{22} \ldots s_{pp})\ |R|$$

Exemplo 3.8: Sendo dada a matriz $S = \begin{bmatrix} 4 & 3 & 1 \\ 3 & 9 & 2 \\ 1 & 2 & 1 \end{bmatrix}$, $s_{11} = 4$; $s_{22} = 9$ e $s_{33} = 1$.

Além disso, $R = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{2}{3} \\ \frac{1}{2} & \frac{2}{3} & 1 \end{bmatrix}$. Como $|S| = 14$ e $|R| = \frac{7}{18}$, confirma-se que

$$14 = |S| = s_{11}\ s_{22}\ s_{33}\ |R| = (4)\,(9)\,(1)\left(\frac{7}{18}\right) = 14$$

o

Concluímos esta discussão apresentando o conceito de variância amostral total cujo valor corresponde ao valor do traço da matriz S, isto é, à soma dos elementos da sua diagonal.

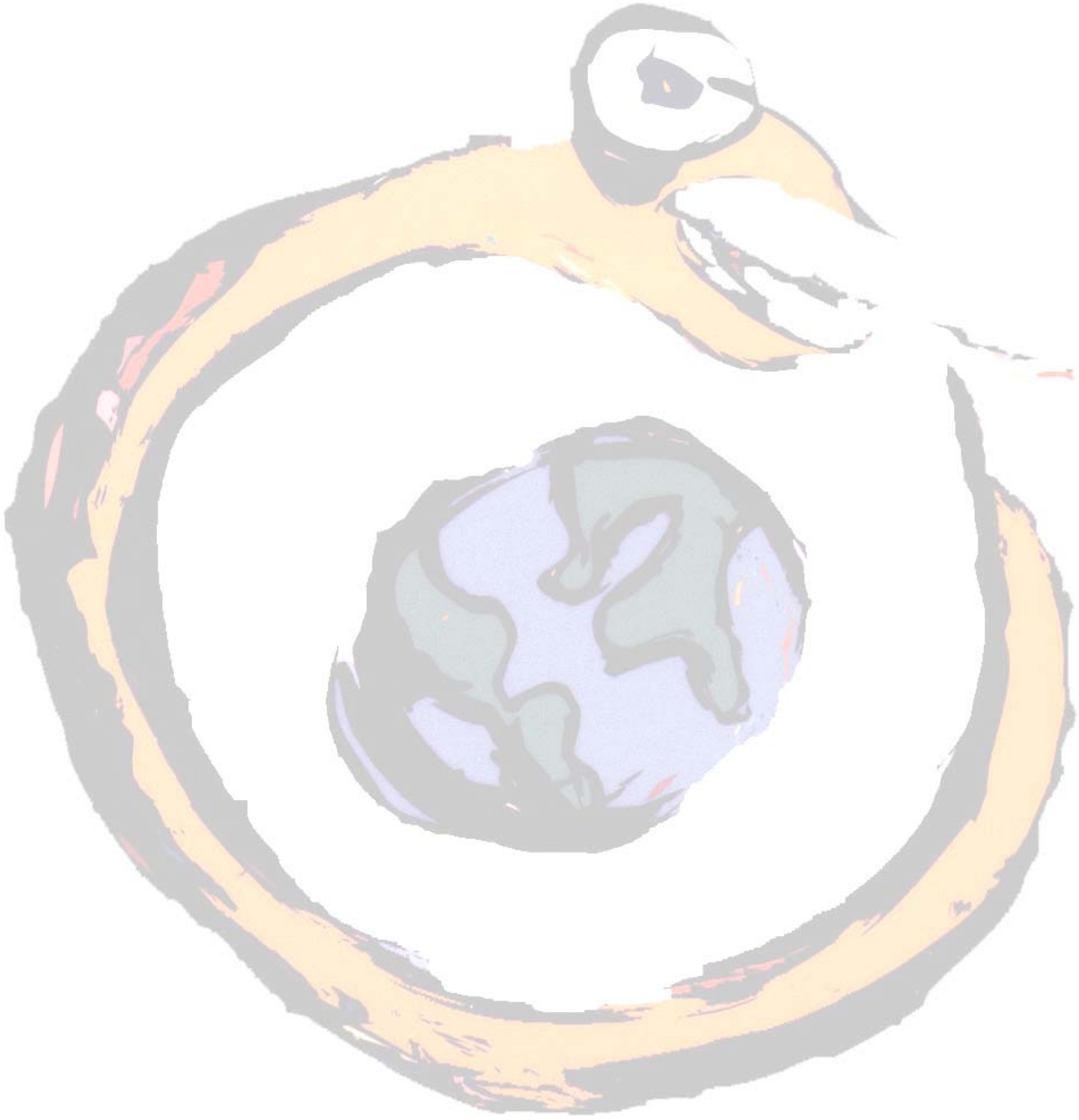
Variância amostral total = $s_{11} + s_{22} + \ldots + s_{pp}$

Exemplo 3.9: A variância amostral total da matriz $S = \begin{bmatrix} 14808 & 14213 \\ 14213 & 15538 \end{bmatrix}$ é

$$s_{11} + s_{22} = 14808 + 15538 = 30346.$$

A variância amostral total da matriz $S = \begin{bmatrix} 3 & -\frac{3}{2} & 0 \\ -\frac{3}{2} & 1 & \frac{1}{2} \\ 0 & \frac{1}{2} & 1 \end{bmatrix}$ é $s_{11} + s_{22} + s_{33} = 3 + 1 + 1 = 5.$ o

Geometricamente, a variância amostral total corresponde à soma dos quadrados dos comprimentos dos p vectores residuais $\mathbf{d}_1 = \mathbf{y}_1 - \bar{x}_1 \mathbf{1}, \ldots, \mathbf{d}_p = \mathbf{y}_p - \bar{x}_p \mathbf{1}$ dividida por n - 1.



4

Distribuição normal multivariada

A generalização da tão conhecida curva normal para várias dimensões desempenha um papel fundamental na análise multivariada.

4.1 A densidade normal multivariada

A densidade normal multivariada consiste numa generalização, para p ≥ 2, da densidade da curva normal

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(x-\mu)/\sigma]^2/2} \qquad -\infty < x < \infty$$

O termo $\left(\frac{x-\mu}{\sigma}\right)^2 = (x - \mu)(\sigma^2)^{-1}(x - \mu)$ no expoente da função densidade não é mais do que a distância quadrada de x a μ em unidades estandardizadas de desvio. Generalizando para um vector **x** de dimensão p×1, podemos escrever

$$(\mathbf{x} - \boldsymbol{\mu})' \sum^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

onde o vector **μ** representa o valor esperado do vector aleatório **x** e a matriz ∑ p×p é a matriz da variâncias.

A função densidade normal p-dimensional $N_p(\boldsymbol{\mu}, \Sigma)$ para o vector aleatório $\mathbf{x} = [X_1, X_2, \ldots, X_p]'$ é

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-(1/2)(\mathbf{x}-\boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

onde $-\infty < x_i < \infty$, $i = 1, 2, \ldots, p$.

Exemplo 4.1: Consideremos o espaço p = 2.

Neste espaço $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ e $\Sigma = \begin{bmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{bmatrix}$

Calculando a inversa da matriz de variâncias, obtemos

$$\Sigma^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{bmatrix} \sigma_{22} & \sigma_{12} \\ \sigma_{12} & \sigma_{11} \end{bmatrix}$$

Assim, a distância quadrada $(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$ fica igual a

$$= \begin{bmatrix} x_1 - \mu_1 & x_1 - \mu_1 \end{bmatrix} \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{bmatrix} \sigma_{22} & -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} \\ -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} & \sigma_{11} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}$$

$$= \frac{\sigma_{22}(x_1 - \mu_1)^2 + \sigma_{11}(x_2 - \mu_2)^2 - 2\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}(x_1 - \mu_1)(x2 - \mu2)}{\sigma_{11}\sigma_{22}(1 - \rho_{12}^2)}$$

$$= \frac{1}{1-\rho_{12}^2} \left[\left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}} \right)^2 + \left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}} \right)^2 - 2\rho_{12} \left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}} \right) \left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}} \right) \right]$$

Deste modo,

$$f(x_1,x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}}$$

$$\times \exp\left\{\frac{-1}{2(1-\rho_{12}^2)}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2+\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2-2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\right]\right\}$$

Olhando para esta última equação, podemos dizer que se $\rho_{12} = 0$, a densidade conjunta pode ser escrita como um produto de duas densidades normais univariadas, isto é, se X_1 e X_2 não estão correlacionadas, $f(x_1,x_2) = f(x_1)\, f(x_2)$, isto é, X_1 e X_2 são independentes.

o

Do que atrás ficou dito, podemos concluir que a densidade normal multivariada é constante nas superfícies onde a distância quadrada $(\mathbf{x} - \boldsymbol{\mu})' \sum^{-1} (\mathbf{x} - \boldsymbol{\mu})$ for constante. Os eixos de cada elipsóide de constante densidade têm a direcção dos vectores próprios de $\sum^{-1}$ e os comprimentos proporcionais aos inversos das raízes quadradas dos valores próprios de $\sum$.

Uma vez que $\sum \mathbf{e} = \lambda \mathbf{e} \Rightarrow \sum^{-1} \mathbf{e} = \frac{1}{\lambda}\mathbf{e}$, os valores próprios de $\sum^{-1}$ podem ser determinados através dos valores próprios de $\sum$.

Deste modo, podemos afirmar que os contornos de densidade constante da distribuição normal p-dimensional constituem elipsóides definidos por $\mathbf{x}$ tal que $(\mathbf{x} - \boldsymbol{\mu})' \sum^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2$. Estes elipsóides são centrados em $\boldsymbol{\mu}$ e possuem eixos com comprimento $\pm c \sqrt{\lambda i}\ \mathbf{e}_i$, onde $\sum \mathbf{e}_i = \lambda_i \mathbf{e}_i$, i=1, …, p.

Exemplo 4.2: Consideremos o caso em que $\sigma_{11} = \sigma_{22}$.

$$|\Sigma - \lambda \mathrm{I}| = 0 \quad \Rightarrow \quad \begin{vmatrix} \sigma_{11} - \lambda & \sigma_{12} \\ \sigma_{12} & \sigma_{11} - \lambda \end{vmatrix} = 0$$

$$\Rightarrow \quad (\lambda - \sigma_{11} - \sigma_{12})\,(\lambda - \sigma_{11} + \sigma_{12}) = 0$$

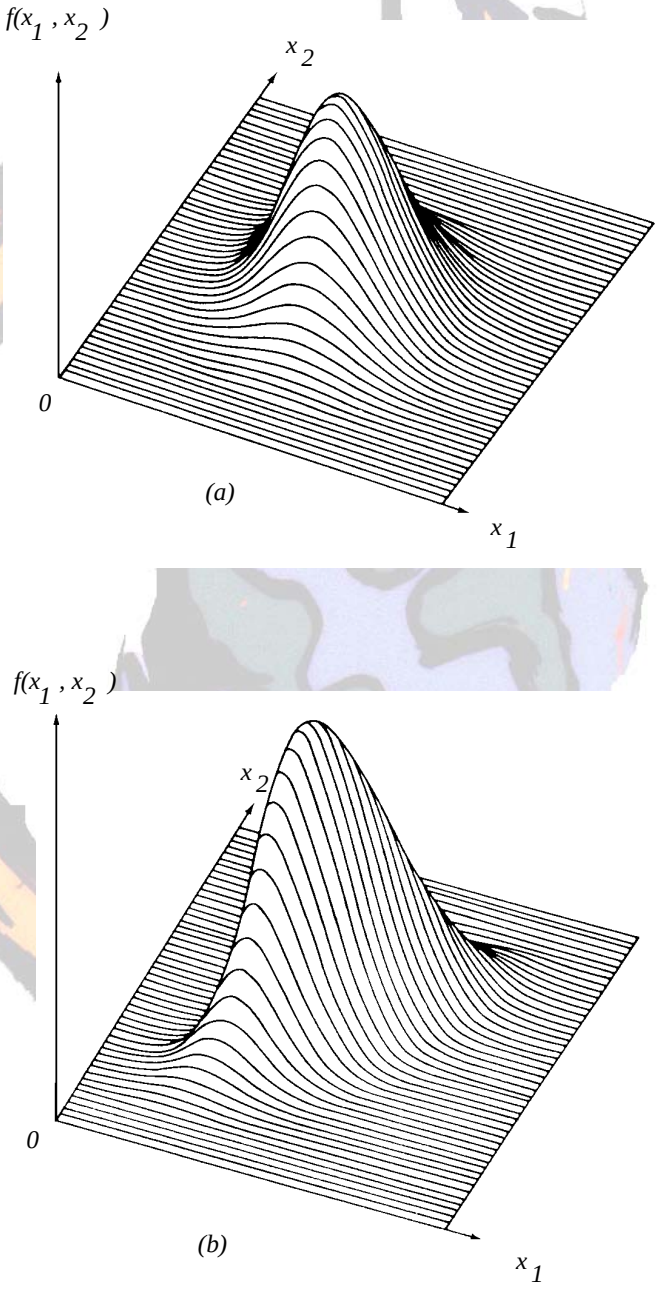


Figura 4.1 — Duas distribuições normais bivariadas

(a) $\sigma_{11} = \sigma_{22}$ e $\rho_{12} = 0$ (b) $\sigma_{11} = \sigma_{22}$ e $\rho_{12} = .75$

Então, os valores próprios são $\lambda_1 = \sigma_{11} + \sigma_{12}$ e $\lambda_2 = \sigma_{11} - \sigma_{12}$. O vector próprio $\mathbf{e}_1$ correspondente ao valor próprio λ_1 é dado por

$$\begin{bmatrix} s_{11} & s_{12} \\ s_{12} & s_{11} \end{bmatrix} \begin{bmatrix} e_{11} \\ e_{21} \end{bmatrix} = (\sigma_{11} + \sigma_{12}) \begin{bmatrix} e_{11} \\ e_{21} \end{bmatrix}$$

$$\Rightarrow \quad \mathbf{e}_1 = \begin{bmatrix} e_{11} \\ e_{21} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

De modo idêntico $\mathbf{e}_2 = \begin{bmatrix} e_{12} \\ e_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix}$

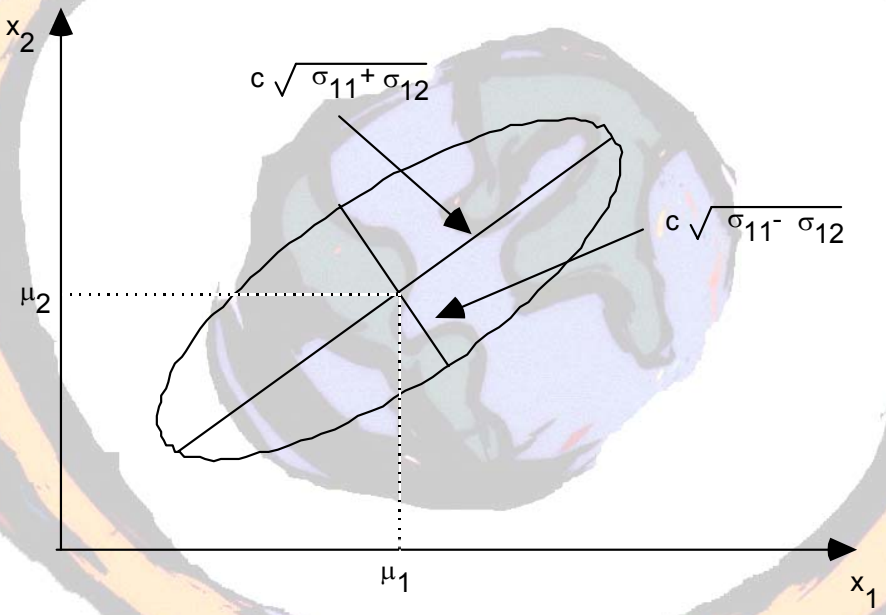


Figura 4.2 - Contorno de densidade constante para uma distribuição normal bivariada com $\sigma_{11} = \sigma_{22}$ e $\sigma_{12} > 0$ (ou $\rho_{12} > 0$)

Quando $\sigma_{12} > 0$, $\lambda_1 = \sigma_{11} + \sigma_{12}$ é o maior valor próprio e o correspondente vector próprio $e_1' = \left[1/\sqrt{2}, 1/\sqrt{2} \right]$] situa-se na recta a 45º que passa por $\boldsymbol{\mu}' = [\mu_1, \mu_2]$. Como os eixos das elipses de densidade constante são dados por $\pm c\sqrt{\lambda_1 e_1}$ e

$\pm c\sqrt{\lambda_2 e_2}$, com cada vector próprio de comprimento unitário, o maior eixo está associado ao maior dos valores próprios.

o

A densidade normal p-variada

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-(1/2)(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

tem um valor máximo quando a distância quadrada $(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})$ for nula, isto é, quando $\mathbf{x}=\boldsymbol{\mu}$. Deste modo, $\boldsymbol{\mu}$ é o ponto de densidade máxima, ou moda, ao mesmo tempo que constitui o valor esperado de **X**, ou média.

4.2 Propriedades da distribuição normal

Vejamos, de seguida, algumas propriedades da distribuição normal. Assim, sendo dado o vector aleatório **x** com uma distribuição normal multivariada, $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$,

- Combinações lineares das componentes de **X** são normalmente distribuídas.

$$\mathbf{a}'\mathbf{X} = a_1\mathbf{X}_1 + a_2\mathbf{X}_2 + \ldots + a_p\mathbf{X}_p \sim N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\Sigma\mathbf{a})$$

$$\underset{(q\times p)}{A}\ \underset{(p\times 1)}{X} = \begin{bmatrix} a_{11}X_1 + \ldots + a_{1p}Xp \\ a_{21}X_1 + \ldots + a_{2p}Xp \\ \ldots \\ a_{q1}X_1 + \ldots + a_{qp}Xp \end{bmatrix} \sim N_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\Sigma\mathbf{A}')$$

$$\underset{(p\times 1)}{X} + \underset{(p\times 1)}{d} \sim N_p(\boldsymbol{\mu}, \mathbf{d}\sum)$$

- Todos os subconjuntos das componentes de X seguem uma distribuição normal multivariada. Se dividirmos **X**, **μ** e ∑

$$\underset{(p\times 1)}{X} = \begin{bmatrix} \underset{(q\times 1)}{X_1} \\ \hline \underset{((p-q)\times 1)}{X_2} \end{bmatrix} \qquad \underset{(p\times 1)}{\mu} = \begin{bmatrix} \underset{(q\times 1)}{\mu_1} \\ \hline \underset{((p-q)\times 1)}{\mu_2} \end{bmatrix}$$

$$\underset{(p\times p)}{\Sigma} = \left[\begin{array}{c|c} \underset{(q\times q)}{\Sigma_{11}} & \underset{(q\times(p-q))}{\Sigma_{12}} \\ \hline \underset{((p-q)\times q)}{\Sigma_{21}} & \underset{((p-q)\times(p-q))}{\Sigma_{22}} \end{array}\right]$$

então, por exemplo, $\mathbf{X}_1 \sim N_q(\boldsymbol{\mu}_1, \sum_{11})$.

- Se $\mathbf{X}_1$ $(q_1\times 1)$ e $\mathbf{X}_2$ $(q_2\times 1)$ forem independentes, então $Cov(\mathbf{X}_1,\mathbf{X}_2) = \mathbf{0}$, sendo **0** uma matriz $(q_1\times q_2)$ de zeros.

- As distribuições condicionais das componentes são normais multivariadas.

Se $X = \left[\begin{array}{c} \mathbf{X}_1 \\ \hline \mathbf{X}_2 \end{array}\right] \sim N_p(\boldsymbol{\mu}, \sum)$ com $\boldsymbol{\mu} = \left[\begin{array}{c} \boldsymbol{\mu}_1 \\ \hline \boldsymbol{\mu}_2 \end{array}\right]$, $\sum = \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right]$

e $|\sum_{22}| > 0$, então a distribuição condicional de $\mathbf{X}_1$ dado $\mathbf{X}_2 = \mathbf{x}_2$ é normal com a média = $\boldsymbol{\mu}_1 + \sum_{12} \Sigma_{22}^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2)$ e covariância = $\sum_{11} - \sum_{12} \Sigma_{22}^{-1} \sum_{21}$.

Notar que a covariância não depende do valor de $\mathbf{x}_2$ da variável condicionante.

- Se $|\sum| > 0$, então $(\mathbf{x} - \boldsymbol{\mu})' \sum^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2_p$, uma distribuição de qui-quadrado com p graus de liberdade.

- A distribuição $N_p(\boldsymbol{\mu},\sum)$ atribui uma probabilidade 1- α ao elipsóide

$$\left\{x : (\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \chi^2_p(\alpha)\right\}$$

sendo $\chi^2_p(\alpha)$ o percentil de ordem (100α) da distribuição χ^2_p.

4.3 A forma amostral da distribuição normal multivariada

Sendo dado $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ uma amostra aleatória de uma população normal com média $\boldsymbol{\mu}$ e covariância $\sum$, os estimadores de máxima verosimilhança para $\boldsymbol{\mu}$ e $\sum$ são dados, respectivamente, por

$$\hat{\boldsymbol{\mu}} = \overline{\mathbf{X}}$$

$$\hat{\sum} = \frac{1}{n}\sum_{j=1}^{n}(\mathbf{X}_j - \overline{\mathbf{X}})(\mathbf{X}_j - \overline{\mathbf{X}})' = \frac{n-1}{n}\mathrm{S}$$

Notar que o estimador $\overline{\mathbf{X}}$ é um vector aleatório e que o estimador $\hat{\sum}$ é uma matriz aleatória.

Estes estimadores de máxima verosimilhança possuem a propriedade da invariância. Isto significa, por exemplo, que o estimador de máxima verosimilhança de $\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}$ é $\hat{\boldsymbol{\mu}}'\hat{\Sigma}^{-1}\hat{\boldsymbol{\mu}}$ e que o estimador de máxima verosimilhança de $\sqrt{\sigma_{jj}}$ é $\sqrt{\hat{\sigma}_{jj}}$,

com $\hat{\sigma}_{jj} = \frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \bar{X}j)^2$ como sendo o estimador de máxima verosimilhança de σ_{jj} = Var($\mathbf{X}_j$).

Tratando-se de populações normais, toda a informação amostral da matriz de dados **X** está contida em $\bar{\mathbf{X}}$ e **S**; qualquer que seja o tamanho n da amostra. Como esta afirmação não é necessariamente verdadeira para populações não normais, é sempre conveniente testar os pressupostos da normal multivariada.

4.4 Distribuição amostral de $\bar{\mathbf{X}}$ e S

No caso univariado (p = 1) sabemos que $\bar{X}$ segue uma distribuição normal com média μ e variância $\frac{1}{n}\sigma^2$. O resultado para o caso multivariado (p ≥ 2) é idêntico. $\bar{\mathbf{X}}$ segue uma distribuição normal com média **μ** e matriz de covariância $\frac{1}{n}$ ∑.

Ora, como ∑ é desconhecida, a distribuição de $\bar{\mathbf{X}}$ não pode ser usada directamente para inferir acerca de **μ**. Contudo, S independente de **μ** fornece-nos informação suficiente acerca de ∑. À medida que o tamanho da amostra cresce, $\bar{\mathbf{X}}$ e S são regidos por algumas propriedades independentemente das características da população-pai. O único requisito que existe é que esta população-pai, qualquer que seja a sua forma, tenha uma média **μ** e uma covariância finita ∑.

Pela Lei dos Grandes Números e sempre que o tamanho da amostra seja grande, existe uma grande probabilidade de que $\bar{\mathbf{X}}$ se aproxime de **μ** e que S se

aproxime de $\sum$. Precisando um pouco mais (Teorema do Limite Central), sejam $\mathbf{X}_1$, $\mathbf{X}_2$, …, $\mathbf{X}_n$ uma observação independente de uma qualquer população com média $\boldsymbol{\mu}$ e covariância finita $\sum$. Então, para amostras grandes (n deve ser grande relativamente a p), $\sqrt{n}\,(\overline{\mathbf{X}} - \boldsymbol{\mu})$ aproximadamente segue uma distribuição $N_p(\mathbf{0}, \sum)$.

Quando $\overline{\mathbf{X}} \sim N_p(\boldsymbol{\mu}, \frac{1}{n} \sum)$ ou seja, quando $\sqrt{n}\,(\overline{\mathbf{X}} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}; \sum)$, pode também demonstrar-se que $n\,(\overline{\mathbf{X}} - \boldsymbol{\mu})' \sum^{-1} (\overline{\mathbf{X}} - \boldsymbol{\mu}) \sim \chi^2_p$.

Reparar, finalmente, que, para n grande e muito maior do que p, substituir Σ^{-1} por S^{-1} não afecta seriamente a aproximação.

5

Inferência acerca do vector média

Nos capítulos anteriores apresentaram-se os conceitos básicos para uma melhor compreensão da estatística multivariada. Neste capítulo iremos analisar a inferência (testes e regiões de confiança) referentes ao vector média de uma população normal.

5.1 T^2 de Hotelling

Uma generalização natural da distância quadrada

$$t^2 = \frac{(X-\mu_o)^2}{s^2/n} = n\,(\overline{X} - \mu_o)\,(s^2)^{-1}\,(\overline{X} - \mu_o)$$

é a correspondente multivariada

$$T^2 = (\overline{\mathbf{X}} - \boldsymbol{\mu}_\mathbf{o})' \left(\frac{1}{n}S\right)^{-1} (\overline{\mathbf{X}} - \boldsymbol{\mu}_\mathbf{o}) = n\,(\overline{\mathbf{X}} - \boldsymbol{\mu}_\mathbf{o})'\,\mathbf{S}^{-1}\,(\overline{\mathbf{X}} - \boldsymbol{\mu}_\mathbf{o})$$

onde $\underset{(p\times 1)}{\overline{X}} = \frac{1}{n}\sum_{j=1}^{n} X_j$

$$\underset{(p\times p)}{S} = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)\left(X_j - \overline{X}\right)' \qquad \underset{(p\times 1)}{\mu_0} = \begin{bmatrix} \mu_{10} \\ \mu_{20} \\ \vdots \\ \mu_{p0} \end{bmatrix}$$

e $\frac{1}{n}$ S representa a matriz estimada das covariâncias de $\overline{\mathbf{X}}$.

A estatística T^2 é denominada T^2 de Hotelling em homenagem a Harold Hotelling, pioneiro da estatística multivariada. Se a distância generalizada observada T^2 for grande, isto é; se $\bar{\mathbf{x}}$ estiver muito longe de $\boldsymbol{\mu}_0$, a hipótese H_0: $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ será rejeitada. Ora, para podermos ter uma ideia da grandeza da distância T^2, utilizamos o conhecimento que temos da sua distribuição. De facto,

$$T^2 \sim \frac{(n-1)p}{(n-p)} F_{p,\, n-p}$$

onde $F_{p,n-p}$ indica uma variável aleatória com uma distribuição F com p e n-p graus de liberdade.

Considerando então a amostra aleatória $\mathbf{X}_1$, $\mathbf{X}_2$, …, $\mathbf{X}_n$ de uma população $N_p(\boldsymbol{\mu}, \Sigma)$,

$$\alpha = P\left[T^2 > \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)\right] = \left[n(\bar{\mathbf{X}}-\boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu}) > \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)\right]$$

quaisquer que sejam os valores verdadeiros de $\boldsymbol{\mu}$ e Σ, com $F_{p,n-p}(\alpha)$ a representar o percentil de ordem (100α) da distribuição $F_{p,n-p}$.

O que já foi dito é suficiente para testar H_0: $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ contra H_1: $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. A um nível de significância α, rejeitamos H_0 em favor de H_1 se

$$T^2 = n\,(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'\,\mathbf{S}^{-1}\,(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \quad > \quad \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)$$

Exemplo 5.1: Analisou-se a transpiração de 20 mulheres saudáveis, tendo sido usadas as três variáveis X_1 = taxa de transpiração, X_2 = conteúdo de sódio e

X_3 = conteúdo de potássio. Os valores encontrados levaram aos seguintes resultados:

$$\overline{\mathbf{X}} = \begin{bmatrix} 4.640 \\ 45.400 \\ 9.965 \end{bmatrix} \quad S = \begin{bmatrix} 2.879 & 10.002 & -1.810 \\ 10.002 & 199.798 & -5.627 \\ -1.810 & -5.627 & 3.628 \end{bmatrix} \quad \text{e} \quad S^{-1} = \begin{bmatrix} .586 & -.022 & .258 \\ -.022 & .006 & -.002 \\ .258 & -.002 & .402 \end{bmatrix}$$

Testar a hipótese H_0: $\boldsymbol{\mu}' = [\ 4\ ,\ 50,\ 10\ ]$ contra H_1: $\boldsymbol{\mu}' \neq [\ 4\ ,\ 50,\ 10\ ]$ a um nível de confiança de $\alpha = .10$.

Ora $T^2 = n\,(\overline{\mathbf{X}} - \boldsymbol{\mu}_0)'\,\mathbf{S}^{-1}\,(\overline{\mathbf{X}} - \boldsymbol{\mu}_0)$

$$= 20\ [4.640 - 4\ ;\ 45.400 - 50\ ;\ 9.965 - 10] \begin{bmatrix} .586 & -.022 & .258 \\ -.022 & .006 & -.002 \\ .258 & -.002 & .402 \end{bmatrix} \begin{bmatrix} 4.640 - 4 \\ 45.400 - 50 \\ 9.965 - 10 \end{bmatrix}$$

$$= 20\ [.640\ ;\ -4.600\ ;\ -.035\ ] \begin{bmatrix} .467 \\ -.042 \\ .160 \end{bmatrix} = 9{,}74$$

Comparando o valor observado T^2 com o valor crítico

$$\frac{(n-1)p}{(n-p)} F_{p,n-p}(.10) = \frac{19(3)}{17} F_{3,17}(.10) = (3.353)\ (2.44) = 8{,}18$$

podemos concluir que $T^2 = 9.74 > 8.18$ e, portanto, rejeitamos H_0 ao nível de confiança de 90%.

o

5.2 Regiões de confiança

Seja $\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2 \quad \ldots \quad \mathbf{X}_n]$ a matriz de dados e θ um vector de parâmetros desconhecidos de uma população. A região R($\mathbf{X}$) é chamada região 100(1-α)% confiança se, antes da amostra ser seleccionada,

$$P[R(X) \text{ incluir o verdadeiro valor para } \theta] = 1 - \alpha$$

Adaptando este conceito à média $\boldsymbol{\mu}$, obtemos

$$P\left[n(\overline{\mathbf{X}} - \boldsymbol{\mu})' \mathbf{S}^{-1} (\overline{\mathbf{X}} - \boldsymbol{\mu}) \leq \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha) \right] = 1 - \alpha$$

Por outras palavras, $\overline{\mathbf{X}}$ estará a uma distância $\sqrt{\frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)}$ de $\boldsymbol{\mu}$, com probabilidade 1 - α, desde que a distância seja definida em termos de $\left(\frac{1}{n} S\right)^{-1}$.

Para se saber se um qualquer valor $\boldsymbol{\mu}_0$ pertence à região de confiança, necessitamos de determinar o valor da distância quadrada generalizada

$$n\,(\overline{\mathbf{x}} - \boldsymbol{\mu}_0)'\, S^{-1}\, (\overline{\mathbf{x}} - \boldsymbol{\mu}_0)$$

e compará-la com o valor de $\frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)$. Caso a distância seja maior do que este último valor, $\boldsymbol{\mu}_0$ não pertencerá à região de confiança.

Os eixos do elipsóide de confiança e os seus respectivos comprimentos podem ser determinados através dos próprios valores próprios λ_i e dos vectores próprios $\mathbf{e}_i$ de S.

Centrado em $\bar{\mathbf{x}}$, os eixos do elipsóide

$$n\,(\bar{\mathbf{x}} - \boldsymbol{\mu})' \mathbf{S}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \leq c^2 = \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)$$

são $\pm \sqrt{\lambda_i}\sqrt{\frac{(n-1)p}{n(n-p)} F_{p,n-p}(\alpha)}\,\mathbf{e}_i$; onde $\mathbf{S}\mathbf{e}_i = \lambda_i \mathbf{e}_i$, i = 1, 2, …, p.

Exemplo 5.2: Num estudo de 42 aparelhos de microondas, foram medidas as radiações emitidas pelos aparelhos, respectivamente, com as portas fechadas (X_1) e com as portas abertas (X_2). Para os 42 pares de observações , encontrou-se

$$\bar{X} = \begin{bmatrix} .564 \\ .603 \end{bmatrix} \quad S = \begin{bmatrix} .0144 & .0117 \\ .0117 & .0146 \end{bmatrix} \quad \text{e} \quad S^{-1} = \begin{bmatrix} 203.018 & -163.391 \\ -163.391 & 200.228 \end{bmatrix}$$

Os pares de valores próprios e vectores próprios para **S** são

$\lambda_1 = .026$ $\qquad$ $\mathbf{e}_1' = [\ .704,\ .710\ ]$

$\lambda_2 = .002$ $\qquad$ $\mathbf{e}_2' = [\ -.710,\ .704\ ]$

A elipse a 95% de confiança para $\boldsymbol{\mu}$ consiste em todos os valores (μ_1 , μ_2) que satisfazem a inequação

$$42\,[.564 - \mu_1\ ;\ .603 - \mu_2] \begin{bmatrix} 203.018 & -163.391 \\ -163.391 & 200.228 \end{bmatrix} \begin{bmatrix} .564 - \mu_1 \\ .603 - \mu_2 \end{bmatrix} \leq \frac{2(41)}{40} F_{2,40}(.05)$$

Como $F_{2,40}(.05) = 3.23$, obtém-se,

$42(203.018)(.564-\mu_1)^2 + 42(200.228)(.603-\mu_2)^2 - 84(163.391)(.564-\mu_1)(.603-\mu_2) \leq 6.62$

Para determinar se $\boldsymbol{\mu}' = [\ .562\ ,\ .589\ ]$ pertence à região de confiança, calculamos a expressão anterior para $\mu_1 = .562$ e $\mu_2 = .589$, encontrando-se o valor $1.30 \leq 6.62$. Concluímos então que se situa na região de confiança.

Do mesmo modo, um teste de H_0: $\boldsymbol{\mu} = \begin{bmatrix} .562 \\ .589 \end{bmatrix}$ não será rejeitado em favor de H_1: $\boldsymbol{\mu} \neq \begin{bmatrix} .562 \\ .589 \end{bmatrix}$ a um nível de significância $\alpha = .05$.

O elipsóide de confiança conjunta está centrado em $\overline{\mathbf{X}} = \begin{bmatrix} .564 \\ .603 \end{bmatrix}$ e, respectivamente, com metades dos eixos maior e menor iguais a

$$\sqrt{\lambda_1}\sqrt{\frac{(n-1)p}{n(n-p)}F_{p,n-p}(\alpha)} = \sqrt{.026}\sqrt{\frac{2(41)}{42(40)}(3.23)} = .064$$

e

$$\sqrt{\lambda_2}\sqrt{\frac{(n-1)p}{n(n-p)}F_{p,n-p}(\alpha)} = \sqrt{.002}\sqrt{\frac{2(41)}{42(40)}(3.23)} = .018$$

Estes eixos encontram-se segundo $\mathbf{e}_1' = [\ .704,\ .710\ ]$ e $\mathbf{e}_2' = [\ -.710,\ .704\ ]$.

Pode-se facilmente ver que o eixo maior é cerca de 3.6 vezes maior do que o eixo menor.

o

Consideremos agora $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \sum)$ e a combinação linear $\mathbf{Z} = \mathbf{c}'\mathbf{X} = c_1\mathbf{X}_1 + c_2\mathbf{X}_2 + \ldots + c_p\mathbf{X}_p$. Então, para $\mathbf{c}$ fixo e σ_z^2 desconhecido, um intervalo de confiança a $100(1 - \alpha)\%$ para $\boldsymbol{\mu}_Z = \mathbf{c}'\boldsymbol{\mu}$ é dado por

$$\mathbf{c}'\overline{\mathbf{x}} - t_{n-1}(\alpha/2)\frac{\sqrt{c'\mathbf{S}c}}{\sqrt{n}} \leq \mathbf{c}'\boldsymbol{\mu} \leq \mathbf{c}'\overline{\mathbf{x}} + t_{n-1}(\alpha/2)\frac{\sqrt{c'\mathbf{S}c}}{\sqrt{n}}$$

onde $t_{n-1}(\alpha/2)$ é o percentil superior de ordem $100(\alpha/2)$ de uma distribuição t com n-1 graus de liberdade.

Esta desigualdade pode ser interpretada como uma afirmação em relação às componentes do vector média **μ**. Por exemplo, com **c´**= [1, 0, …, 0], **c´μ** = μ_1 torna-se no intervalo de confiança já por nós conhecido para a média de uma população normal, sendo **c´**S**c** = s_{11}.

Podemos deste modo construir vários intervalos de confiança para os componentes de **μ**, cada um deles associado a um coeficiente de confiança de 1-α. Basta para isso escolher os vectores **c** apropriados. Contudo, a confiança associada a todos os intervalos quando tomados em conjunto não é igual a 1-α.

Sendo dada a amostra aleatória $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ de uma população $N_p(\boldsymbol{\mu}, \Sigma)$, com ∑ definida positiva, para todos os **c** simultaneamente, o intervalo

$$\left(c'\mathbf{X} - \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p}(\alpha) c' S c} \quad ; \quad c'\mathbf{X} + \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p}(\alpha) c' S c} \right)$$

contém **c´μ** com probabilidade 1-α.

Estes intervalos simultâneos são, por vezes, denominados, intervalos T^2 pois a probabilidade de cobertura é determinada pela distribuição de T^2. As escolhas **c´**= [1, 0, …, 0], **c´**= [0, 1, …, 0], …, **c´**= [0, 0, …, 1] permitem-nos concluir que todos os intervalos

$$\overline{x_1} - \sqrt{\frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha)} \sqrt{\frac{s_{11}}{n}} \leq \mu_1 \leq \overline{x_1} + \sqrt{\frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha)} \sqrt{\frac{s_{11}}{n}}$$

$$\overline{x_2} - \sqrt{\frac{p(n-1)}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{22}}{n}} \le \mu_2 \le \overline{x_2} + \sqrt{\frac{p(n-1)}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{22}}{n}}$$

…

$$\overline{x_p} - \sqrt{\frac{p(n-1)}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{pp}}{n}} \le \mu_p \le \overline{x_p} + \sqrt{\frac{p(n-1)}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{pp}}{n}}$$

se verificam com um coeficiente de confiança de 1-α.

Reparar que, por exemplo, para se obter um intervalo de confiança para $\mu_i - \mu_k$ basta usar-se $c_i = c_k = 1$ no vector $\mathbf{c}' = [\, 0, \ldots, c_i, 0, \ldots, -c_k, \ldots, 0\,]$ a que corresponde $\mathbf{c}'\mathbf{S}\mathbf{c} = s_{ii} - 2s_{ik} + s_{kk}$, obtendo-se o intervalo

$$\overline{x_i} - \overline{x_k} \pm \sqrt{\frac{p(n-1)}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{ii} - 2s_{ik} + s_{kk}}{n}}$$

Exemplo 5.3: 87 alunos de um liceu obtiveram classificações em três exames especiais: X_1 = ciências sociais, X_2 = verbal e X_3 = ciências exactas. Os resultados obtidos foram:

$$\overline{\mathbf{X}} = \begin{bmatrix} 527.74 \\ 54.69 \\ 25.13 \end{bmatrix} \text{ e } \qquad \mathbf{S} = \begin{bmatrix} 5691.34 & 600.51 & 217.25 \\ 600.51 & 126.05 & 23.37 \\ 217.25 & 23.37 & 23.11 \end{bmatrix}$$

Para encontrar os intervalos simultâneos de confiança a 95% para μ_1, μ_2 e μ_3 necessitamos calcular o valor

$$\frac{p(n-1)}{(n-p)}F_{p,n-p}(\alpha) = \frac{3(87-1)}{(87-3)}F_{3,84}(.05) = \frac{3(86)}{84}(2.7) = 8.29$$

obtendo assim os intervalos

$$527.74 - \sqrt{8.29}\sqrt{\frac{5691.34}{87}} \leq \mu_1 \leq 527.74 + \sqrt{8.29}\sqrt{\frac{5691.34}{87}} \qquad 504.45 \leq \mu_1 \leq 551.03$$

$$54.69 - \sqrt{8.29}\sqrt{\frac{126.05}{87}} \leq \mu_2 \leq 54.69 + \sqrt{8.29}\sqrt{\frac{126.05}{87}} \qquad 51.22 \leq \mu_2 \leq 58.16$$

$$25.13 - \sqrt{8.29}\sqrt{\frac{23.11}{87}} \leq \mu_3 \leq 25.13 + \sqrt{8.29}\sqrt{\frac{23.11}{87}} \qquad 23.65 \leq \mu_3 \leq 26.61$$

o

Se o número m de médias μ_i ou de combinações lineares $\mathbf{c}'\boldsymbol{\mu} = c_1\mu_1 + c_2\mu_2 + \ldots + c_p\mu_p$ for pequeno, os intervalos de confiança simultâneos podem ser obtidos de uma forma mais precisa. Tais intervalos de confiança, denominados de Bonferroni, são baseados nos intervalos t individuais

$$\bar{x}_i \pm t_{n-1}\left(\frac{\alpha_i}{2}\right)\sqrt{\frac{s_{ii}}{n}} \qquad i = 1, 2, \ldots, m$$

com $\alpha_i = \alpha/m$. Assim, para um nível de confiança global maior ou igual a $1 - \alpha$, podemos obter m = p intervalos:

$$\bar{x}_1 - t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{11}}{n}} \leq \mu_1 \leq \bar{x}_1 + t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{11}}{n}}$$

$$\bar{x}_2 - t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{22}}{n}} \leq \mu_2 \leq \bar{x}_2 + t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{22}}{n}}$$

$$\bar{x}_p - t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{pp}}{n}} \leq \mu_p \leq \bar{x}_p + t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{pp}}{n}}$$

Exemplo 5.4: Voltando aos dados da transpiração, podemos obter os intervalos de confiança de Bonferroni a 95% para μ_1, μ_2 e μ_3 correspondentes à escolha de α_i = .05/3, i=1, 2, 3.

Como n = 20 e $t_{19}(.05/2(3)) = t_{19}(.0083) = 2.625$, temos

$$\bar{x}_1 \pm t_{19}(.0083)\sqrt{\frac{s_{11}}{n}} = 4.64 \pm 2.625\sqrt{\frac{2.879}{20}} \qquad 3.64 \le \mu_1 \le 5.64$$

$$\bar{x}_2 \pm t_{19}(.0083)\sqrt{\frac{s_{22}}{n}} = 45.4 \pm 2.625\sqrt{\frac{199.798}{20}} \qquad 37.10 \le \mu_2 \le 53.70$$

$$\bar{x}_3 \pm t_{19}(.0083)\sqrt{\frac{s_{33}}{n}} = 9.965 \pm 2.625\sqrt{\frac{3.628}{20}} \qquad 8.85 \le \mu_3 \le 11.08$$

o

5.3 Inferências para grandes amostras

Quando o tamanho da amostra é grande, os testes de hipóteses e as regiões de confiança para **μ** podem ser construídos sem o pressuposto da existência de uma população normal, mesmo tratando-se de distribuições discretas. Todas as inferências de amostras grandes são baseadas na distribuição χ^2.

$(\bar{\mathbf{X}} - \boldsymbol{\mu})' \left(\frac{1}{n}S\right)^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}) = n(\bar{\mathbf{X}} - \boldsymbol{\mu})' S^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu})$ é aproximadamente χ^2

com p graus de liberdade e, então,

$$P\left[n(\bar{\mathbf{X}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \le \chi_p^2(\alpha)\right] = 1 - \alpha$$

onde $\chi^2_p(\alpha)$ é o percentil superior de ordem (100α) da distribuição χ^2_p.

Seja $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ uma amostra aleatória de uma população com média $\boldsymbol{\mu}$ e matriz de covariância definida positiva Σ. Quando n - p for grande,

- a hipótese H_0: $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ é rejeitada em favor de H_1: $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, a um nível de significância aproximadamente α se

$$n\,(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'\,\mathbf{S}^{-1}\,(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) > \chi^2_p(\alpha)$$

- $\mathbf{c}'\bar{\mathbf{X}} \pm \sqrt{\chi^2_p(\alpha)}\sqrt{\dfrac{\mathbf{c}'\mathbf{S}\mathbf{c}}{n}}$ contém $\mathbf{c}'\boldsymbol{\mu}$, para todo $\mathbf{c}$, com probabilidade aproximadamente 1-α. Consequentemente, os intervalos de confiança simultâneos a 100(1-α)%

$$\bar{x}_1 \pm \sqrt{\chi^2_p(\alpha)}\sqrt{\frac{s_{11}}{n}} \text{ contém } \mu_1$$

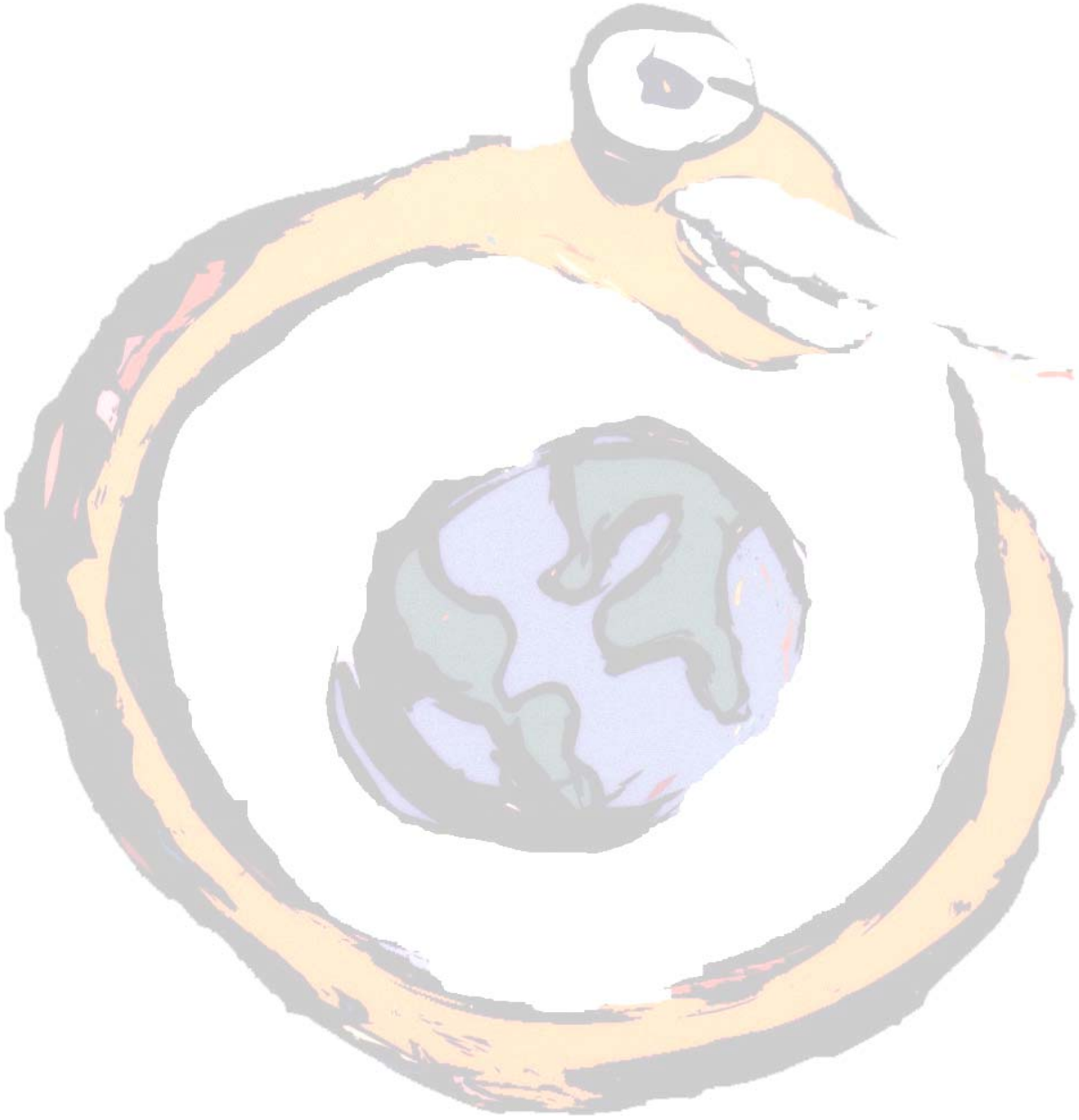
$$\bar{x}_2 \pm \sqrt{\chi^2_p(\alpha)}\sqrt{\frac{s_{22}}{n}} \text{ contém } \mu_2$$

$$\ldots$$

$$\bar{x}_p \pm \sqrt{\chi^2_p(\alpha)}\sqrt{\frac{s_{pp}}{n}} \text{ contém } \mu_p$$

Além disso, para todos os pares (μ_i, μ_k), i, k = 1, 2, …, p, as elipses amostrais centradas na média

$$n\,[\bar{x}_i - \mu_i\ ;\ \bar{x}_k - \mu_k] \begin{bmatrix} s_{ii} & s_{ik} \\ s_{ik} & s_{kk} \end{bmatrix}^{-1} \begin{bmatrix} x_{-i} - \mu_i \\ x_{-k} - \mu_k \end{bmatrix} \leq \chi^2_p(\alpha) \quad \text{contém} \quad (\mu_i, \mu_k)$$



6

Comparação entre duas médias multivariadas

Neste capítulo iremos estender o conhecimento à comparação entre dois vectores média. Mais uma vez iremos partir de procedimentos univariados e generalizaremos para o caso multivariado.

6.1 Comparações emparelhadas

Por vezes, as medições são feitas em condições experimentais diversas, com o objectivo de testar se as respostas diferem significativamente. É o caso, por exemplo, de um teste de eficácia de um novo medicamento que requer que haja medições antes e após um determinado tratamento. As respostas emparelhadas podem ser comparadas analisando as respectivas diferenças.

No caso univariado, e considerando X_{1j} e X_{2j}, respectivamente, as medições "antes" e "após", os efeitos são representados pelas diferenças $d_j = x_{1j} - x_{2j}$, j= 1, 2, …, n. Partindo do pressuposto de que as diferenças D_j representam observações independentes de uma distribuição $N(\delta, \sigma_d^2)$, a variável

$$t = \frac{(\overline{D} - \delta)}{s_d / \sqrt{n}};$$

onde $\overline{D} = \frac{1}{n}\sum_{j=1}^{n} D_j$ e $s_d^2 = \frac{1}{n-1}\sum_{j=1}^{n}(D_j - \overline{D})^2$, segue uma distribuição t com n-1 graus de liberdade.

Consequentemente, a um nível α, o teste $H_0: \delta = 0$ contra $H_0: \delta \neq 0$ pode ser conduzido comparando $|t|$ com $t_{n-1}(\alpha/2)$. Do mesmo modo, um intervalo de confiança a 100(1-α)% para a diferença média $\delta = E(X_{1j} - X_{2j})$ pode ser obtido pela expressão

$$\overline{d} - t_{n-1}(\alpha/2)\frac{s_d}{\sqrt{n}} \leq \delta \leq \overline{d} + t_{n-1}(\alpha/2)\frac{s_d}{\sqrt{n}}$$

Ao generalizar para o caso multivariado, vamos necessitar de distinguir entre p respostas, 2 tratamentos e n unidades experimentais. Obtemos assim as p variáveis aleatórias de diferenças

$$D_{1j} = X_{11j} - X_{21j}$$

$$D_{2j} = X_{12j} - X_{22j}$$

$$\dots$$

$$D_{pj} = X_{1pj} - X_{2pj}$$

ou, em forma matricial,

$$\begin{bmatrix} X_{111} & X_{112} & \dots & X_{11n} \\ . & . & \dots & . \\ X_{1p1} & X_{1p2} & \dots & X_{1pn} \end{bmatrix} - \begin{bmatrix} X_{211} & X_{212} & \dots & X_{21n} \\ . & . & \dots & . \\ X_{2p1} & X_{2p2} & \dots & X_{2pn} \end{bmatrix} = \begin{bmatrix} D_{11} & D_{12} & \dots & D_{1n} \\ . & . & \dots & . \\ D_{p1} & D_{p2} & \dots & D_{pn} \end{bmatrix}$$

Considerando $\mathbf{D}_j' = \lfloor D_{1j} D_{2j} \dots D_{pj} \rfloor \quad (j = 1,2,\dots,n)$,

$E(\mathbf{D}_j) = \delta = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \cdots \\ \delta_p \end{bmatrix}$ e $\text{cov}(\mathbf{D}_j) = \sum_d$.

Se, além disso, $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_n$ forem vectores aleatórios independentes $N_p(\delta, \sum_d)$, então

$$T^2 = n(\overline{\mathbf{D}} - \delta)' \mathbf{S}_d^{-1} (\overline{\mathbf{D}} - \delta)$$

onde $\overline{\mathbf{D}} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{D}_j$ e $\mathbf{S}_d = \frac{1}{n-1}\sum_{j=1}^{n} (\mathbf{D}_j - \overline{\mathbf{D}})'(\mathbf{D}_j - \overline{\mathbf{D}})$ é distribuído como uma variável aleatória $\frac{(n-1)p}{(n-p)} F_{p,n-p}$.

Se ambos n e n-p forem grandes, T^2 é aproximadamente distribuída como χ_p^2, independentemente da forma da população subjacente das diferenças.

Sendo observadas as diferenças $\mathbf{d}_j = \lfloor d_{1j} d_{2j} \ldots d_{pj} \rfloor \, (j = 1,2,\ldots,n)$, rejeitamos H_0: $\delta = 0$ contra H_1: $\delta \neq 0$ a um nível α para uma população $N_p(\delta, \sum_d)$ se o valor observado

$$T^2 = n \overline{\mathbf{d}}' \mathbf{S}_d^{-1} \overline{\mathbf{d}} > \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)$$

onde $F_{p;n-p}(\alpha)$ é o valor do percentil de ordem 100α de uma distribuição F com p e n-p graus de liberdade.

Uma região de confiança a 100(1-α)% para δ é formado por todos os δ tal que

$$(\overline{\mathbf{d}} - \delta)\, \mathbf{S}_d^{-1}\, (\overline{\mathbf{d}} - \mathbf{d}) \leq \frac{(n-1)p}{n(n-p)} F_{p,n-p}(\alpha)$$

Os intervalos simultâneos de confiança a (1-α)% para δ_i são dados por

$$\delta_i: \quad \overline{d_i} \pm \sqrt{\frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)} \sqrt{\frac{s_{d_i}^2}{n}}$$

onde $\overline{d_i}$ é o elemento de ordem i de $\overline{\mathbf{d}}$ e $s_{d_i}^2$ é o i-ésimo elemento da diagonal de $\mathbf{S}_d$.

Para n-p grande; $\frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)$ aproxima-se da distribuição $\chi_p^2(\alpha)$, e a normalidade não é mais necessária.

Os intervalos de confiança simultâneos de Bonferroni a 100(1-α)% para as médias individuais de diferença, δ_i, são

$$\delta_i: \quad \overline{d_i} \pm t_{n-p}\left(\frac{\alpha}{2p}\right) \sqrt{\frac{s_{d_i}^2}{n}}$$

onde $t_{n-p}\left(\frac{\alpha}{2p}\right)$ é o percentil de ordem 100(α/2p) de uma distribuição t com n-p graus de liberdade.

Exemplo 6.1: Um conjunto de 11 amostras de água foi enviado a dois laboratórios para a análise da necessidade oxigénio bioquímico (NOB) e de sólidos suspensos (SS). Os dados são apresentados a seguir:

	Laboratório 1		Laboratório 2	
Amostra j	x_{11j} (NOB)	x_{12j} (SS)	x_{21j} (NOB)	x_{22j} (SS)
1	6	27	25	15
2	6	23	28	33
3	18	64	36	22
4	8	44	35	29
5	11	30	15	31
6	34	75	44	64
7	28	26	42	30
8	71	124	54	64
9	43	54	34	56
10	33	30	29	20
11	20	14	39	21

Será que os resultados provenientes dos dois laboratórios coincidem? Se existir diferença, de que tipo é?

A estatística T^2 para o teste H_0: $\delta' = [\ \delta_1 , \delta_2\ ] = [\ 0 , 0\ ]$ contra H_0: $\delta \neq \mathbf{0}$ é construída a partir das observações de diferenças:

$d_{1j} = x_{11j} - x_{21j}$	-19	-22	-18	-27	-4	-10	-14	17	9	4	-19
$d_{2j} = x_{12j} - x_{22j}$	12	10	42	15	-1	11	- 4	60	-2	10	-7

Então,

$\overline{\mathbf{d}} = \begin{bmatrix} \bar{d}_1 \\ \bar{d}_1 \end{bmatrix} = \begin{bmatrix} -9.36 \\ 13.27 \end{bmatrix}$; $\mathbf{S}_d = \begin{bmatrix} 199.26 & 88.38 \\ 88.38 & 418.61 \end{bmatrix}$ e

$T^2 = 11\ [-9.36;\ 13.27] \begin{bmatrix} .0055 & -.0012 \\ -.0012 & .0026 \end{bmatrix} \begin{bmatrix} -9.36 \\ 13.27 \end{bmatrix} = 13.6$

Com $\alpha = .05$; encontramos $\frac{(n-1)p}{(n-p)} F_{p;n-p}(.05) = \frac{2\,(10)}{9} F_{2;9}(.05) = 9.47$

Como $T^2 = 13.6 > 9.47$, rejeitamos H_0 e concluímos que existe uma diferença média não nula entre as medições dos dois laboratórios. Dos dados parece evidente que o primeiro laboratório tende a produzir medições mais baixas para NOB e mais altas para SS do que o segundo laboratório.

Os intervalos de confiança simultâneos a 95% para as médias das diferenças $\delta_1 e \delta_2$ são, respectivamente,

$\overline{d_1} \pm \sqrt{\frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)} \sqrt{\frac{s_{d_1}^2}{n}} = -9.36 \pm \sqrt{9.47} \sqrt{\frac{199.26}{11}}$ ou (-22.46 ; 3.74)

$\overline{d_2} \pm \sqrt{\frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)} \sqrt{\frac{s_{d_1}^2}{n}} = 13.27 \pm \sqrt{9.47} \sqrt{\frac{418.61}{11}}$ ou (-5.71 ; 32.25)

O intervalo de confiança simultâneo a 95% inclui o valor zero e, no entanto, como vimos, a hipótese H_0: $\delta = \mathbf{0}$ foi rejeitada.

De facto, o ponto $\delta = \mathbf{0}$ encontra-se fora da região de confiança a 95%, o que é consistente com o teste T^2. Os intervalos de confiança simultâneos dizem respeito ao conjunto de todos o conjunto de intervalos que podem ser construídos a partir das possíveis combinações $c_1\delta_1 + c_2\delta_2$, de que os intervalos calculados

correspondem às escolhas ($c_1 = 1$, $c_2 = 0$) e ($c_1 = 0$, $c_2 = 1$). Estes intervalos contêm o valor zero; no entanto, outras escolhas para c_1e c_2 produzem intervalos simultâneos que não contêm zero. Sabemos, sim , que se a hipótese H_0: $\delta = \mathbf{0}$ não tivesse sido rejeitada, todos os intervalos simultâneos incluiriam zero. Os intervalos de Bonferroni também cobrem o valor zero.

o

6.2 Comparações em desenhos de medidas repetidas

Outra generalização da estatística t univariada consiste no caso de q tratamentos serem comparados relativamente a uma única variável de resposta. Cada indivíduo ou unidade experimental recebe o tratamento uma vez em vários períodos de tempo. A observação de ordem j é

$$\mathbf{X}_j = \begin{bmatrix} X_{1j} \\ X_{2j} \\ . \\ . \\ . \\ X_{qj} \end{bmatrix} \qquad j = 1, 2, \ldots, n$$

onde X_{ij} corresponde ao tratamento de ordem i no indivíduo ou unidade experimental j.

Representando por **C** a matriz de contraste onde as q-1 linhas são linearmente independentes, podemos formular a hipótese de que não há diferenças nos tratamentos (igualdade das médias dos tratamentos) fazendo $\mathbf{C\mu = 0}$, qualquer que seja a escolha da matriz de contraste **C**.

Considerando uma população $N_p(\boldsymbol{\mu}, \sum)$, uma matriz de contraste **C** e um nível α, a hipótese H_0: $\mathbf{C\mu = 0}$ é rejeitada em relação à hipótese H_1: $\mathbf{C\mu \neq 0}$ se

$$T^2 = n\,(\mathbf{C}\bar{\mathbf{x}})'\,(\mathbf{CSC}')^{-1}\,\mathbf{C}\bar{\mathbf{x}} > \frac{(n-1)(q-1)}{(n-q+1)} F_{q-1,n-q+1}(\alpha)$$

onde $F_{q-1,n-q+1}(\alpha)$ é o percentil de ordem 100α de uma distribuição F, com q-1 e n-q+1 graus de liberdade.

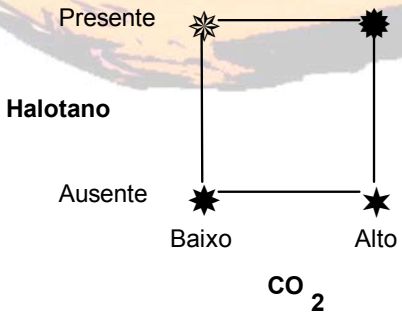
A região de confiança para os contrastes **Cμ** é determinada pelo conjunto de todos os **Cμ** tal que

$$n\,(\mathbf{C}\bar{\mathbf{x}} - \mathbf{C}\boldsymbol{\mu})'\,(\mathbf{CSC}')^{-1}\,(\mathbf{C}\bar{\mathbf{x}} - \mathbf{C}\boldsymbol{\mu}) \leq \frac{(n-1)(q-1)}{(n-q+1)} F_{q-1,n-q+1}(\alpha)$$

Consequentemente, os intervalos simultâneos de confiança a 100(1-α)% para um único contraste **c´μ** é dado por

$$\mathbf{c}'\boldsymbol{\mu} : \mathbf{c}'\bar{\mathbf{x}} \pm \sqrt{\frac{(n-1)(q-1)}{(n-q+1)} F_{q-1,n-q+1}(\alpha)} \sqrt{\frac{c'Sc}{n}}$$

Exemplo 6.2: Num teste de eficácia de um novo anestésico, foi escolhida uma amostra de 19 cães aos quais foi administrado dióxido de carbono (CO_2) a dois níveis de pressão (alto e baixo), seguido da adição de halotano (H) e da repetição de dióxido de carbono.



tratamento 1 = CO_2 alto sem H

tratamento 2 = CO_2 baixo sem H

tratamento 3 = CO_2 alto com H

tratamento 4 = CO_2 baixo com H

Os dados referentes aos milisegundos entre batidas do coração estão apresentados a seguir:

	Tratamento			
Cão	1	2	3	4
1	426	609	556	600
2	253	236	392	395
3	359	433	349	357
4	432	431	522	600
5	405	426	513	513
6	324	438	507	539
7	310	312	410	456
8	326	326	350	504
9	375	447	547	548
10	256	286	403	422
11	349	382	473	497
12	429	410	488	547
13	348	377	447	514
14	412	473	472	446
15	347	326	455	468
16	434	458	637	524
17	364	367	432	469
18	420	395	508	531
19	397	556	645	625

Com base neste desenho de medidas repetidas, vamos analisar os efeitos anestésicos da pressão de CO_2 e do halotano. Representando por μ_1, μ_2, μ_3, e μ_4, respectivamente, as respostas médias nos tratamentos 1, 2, 3 e 4, estamos interessados nos seguintes três contrastes de tratamento:

$(\mu_3 + \mu_4) - (\mu_1 + \mu_2)$ contraste halotano, representando a diferença entre a presença e a ausência do halotano

$(\mu_1 + \mu_3) - (\mu_2 + \mu_4)$ contraste CO_2, representando a diferença entre as pressões baixa e alta de CO_2

$(\mu_1 + \mu_4) - (\mu_2 + \mu_3)$ contraste interacção, representando a influência do halotano nas diferenças de pressão de CO_2

Com $\boldsymbol{\mu}' = [\mu_1 \ \mu_2 \ \mu_3 \ \mu_4]$, a matriz de contraste é $\mathbf{C} = \begin{bmatrix} -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$

Dos dados acima, $\bar{\mathbf{x}} = \begin{bmatrix} 368.21 \\ 404.63 \\ 479.26 \\ 502.89 \end{bmatrix}$ e $\mathbf{S} = \begin{bmatrix} 2819.19 & . & . & . \\ 3568.42 & 7963.14 & . & . \\ 2943.49 & 5303.98 & 6851.32 & . \\ 2295.35 & 4065.44 & 4499.63 & 4878.99 \end{bmatrix}$

Então; $\mathbf{C}\bar{\mathbf{x}} = \begin{bmatrix} 209.31 \\ -60.05 \\ -12.79 \end{bmatrix}$, $\mathbf{CSC}' = \begin{bmatrix} 9432.32 & 1098.92 & 927.62 \\ 1098.92 & 5195.84 & 914.54 \\ 927.62 & 914.54 & 7557.44 \end{bmatrix}$

e $T^2 = n(\mathbf{C}\bar{\mathbf{x}})'(\mathbf{CSC}')^{-1}(\mathbf{C}\bar{\mathbf{x}}) = 19\,(6.11) = 116.$

Com $\alpha = .05$, $\frac{(n-1)(q-1)}{(n-q+1)} F_{q-1;n-q+1}(\alpha) = \frac{18(3)}{16} F_{3;16}(5) = \frac{18(3)}{16}(3.24) = 10.94.$

Como $T^2 = 116 > 10.94$, rejeitamos H_0: $\mathbf{C\mu} = \mathbf{0}$ (não há efeitos do tratamento).

Para detectarmos quais os contrastes responsáveis pela rejeição de H_0, construímos os intervalos simultâneos de confiança a 95% para estes contrastes. Assim, a influência de halotano é estimada pelo intervalo

- $(\overline{x_3} + \overline{x_4}) - (\overline{x_1} + \overline{x_2}) \pm \sqrt{\frac{18\,(3)}{16} F_{3,16}(.05)} \sqrt{\frac{c_1' S c_1}{19}}$

$$= 209.31 \pm \sqrt{10.94}\sqrt{\frac{9432.32}{19}} = 209.31 \pm 73.70$$

Do mesmo modo, os contrastes restantes são estimados por

- influência da pressão $CO_2 = (\mu_1 + \mu_3) - (\mu_2 + \mu_4)$:

$$= -60.05 \pm \sqrt{10.94}\sqrt{\frac{5195.84}{19}} = -60.05 \pm 54.70$$

- interacção H - $CO_2 = (\mu_1 + \mu_4) - (\mu_2 + \mu 3_4)$:

$$= -12.79 \pm \sqrt{10.94}\sqrt{\frac{7557.44}{19}} = -12.79 \pm 65.97$$

Podemos ver, do primeiro intervalo, que existe um efeito do halotano. A presença do halotano produz tempos mais longos entre batidas do coração, o que acontece a ambos os níveis de pressão de CO_2 (pois o contraste de interacção não é significativamente diferente de zero). O segundo intervalo de confiança também indica que há um efeito devido à pressão de CO_2, provocando as baixas pressões maiores tempos entre batidas.

Há, no entanto, que referir que estes resultados devem ser encarados com algum cuidado, uma vez que as experiências com halotano tem necessariamente de ser realizadas após as experiências sem halotano. Assim, o efeito encontrado derivado à presença do halotano pode também ser derivado ao factor tempo.

o

6.3 Comparações entre duas populações

É também possível compararmos as respostas obtidas em duas populações. Consideremos uma amostra aleatória de tamanho n_1 de uma população 1 e uma amostra de tamanho n_2 de uma população 2. As observações em p variáveis são tais que:

Amostra	Estatísticas
População 1 $\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1n_1}$	$\overline{\mathbf{x}_1} = \frac{1}{n_1}\sum_{j=1}^{n_1}\mathbf{x}_{1j}$ $\quad \mathbf{S}_1 = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(\mathbf{x}_{1j}-\overline{\mathbf{x}_1})(\mathbf{x}_{1j}-\overline{\mathbf{x}_1})'$
População 2 $\mathbf{x}_{21}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{2n_2}$	$\overline{\mathbf{x}_2} = \frac{1}{n_2}\sum_{j=1}^{n_2}\mathbf{x}_{2j}$ $\quad \mathbf{S}_2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(\mathbf{x}_{2j}-\overline{\mathbf{x}_2})(\mathbf{x}_{2j}-\overline{\mathbf{x}_2})'$

Pretendemos inferir acerca da diferença entre os vectores média de ambas as populações ($\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$). Será que $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ (isto é, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$)? E se $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \mathbf{0}$, que médias são diferentes?

Para se responder a estas questões, há que se partir de alguns pressupostos. Assim,

- A amostra $\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1}$ é aleatória de comprimento n_1 de uma população p-variada com vector média $\boldsymbol{\mu}_1$ e matriz de covariância $\sum_1$.
- A amostra $\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2}$ é aleatória de comprimento n_2 de uma população p-variada com vector média $\boldsymbol{\mu}_2$ e matriz de covariância $\sum_2$.
- $\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1}$ são independentes de $\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2}$.

Além disto, quando n_1 e n_2 são pequenos,

- Ambas as populações são normais multivariadas.
- Igual matriz de covariância ($\Sigma_1 = \Sigma_2 = \Sigma$).

Neste último caso há, portanto necessidade de estimar a covariância comum Σ, fazendo

$$\mathbf{S}_{comum} = \frac{\sum_{j=1}^{n_1}(\mathbf{x}_{1j}-\bar{\mathbf{x}}_1)(\mathbf{x}_{1j}-\bar{\mathbf{x}}_1)' + \sum_{j=1}^{n_2}(\mathbf{x}_{2j}-\bar{\mathbf{x}}_2)(\mathbf{x}_{2j}-\bar{\mathbf{x}}_2)'}{n_1+n_2-2}$$

$$= \frac{(n_1-1)\mathbf{S}_1+(n_2-1)\mathbf{S}_2}{n_1+n_2-2}$$

Como $\mathbf{S}_{comum}$ estima Σ, podemos afirmar que $\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\mathbf{S}_{comum}$ é um estimador de $Cov(\bar{X}_1 - \bar{X}_2)$.

Sendo dado o teste $H_0: \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \delta_0$ contra $H_1: \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \delta_0$; rejeitamos H_0 se

$$T^2 = (\bar{x}_1 - \bar{x}_2 - \delta_0)' \left[\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\mathbf{S}_{comum}\right]^{-1} (\bar{x}_1 - \bar{x}_2 - \delta_0) > c^2$$

onde $c^2 = \frac{(n_1+n_2-2)p}{(n_1+n_2-p-1)} F_{p,n_1+n_2-p-1}(\alpha)$.

Exemplo 6.3: 50 barras de sabão são fabricadas de cada um de dois processos e duas características X_1 = espuma e X_2 = suavidade são medidas. Foram obtidas as seguintes estatísticas:

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 8.3 \\ 4.1 \end{bmatrix} \qquad \mathbf{S}_1 = \begin{bmatrix} 2 & 1 \\ 1 & 6 \end{bmatrix}$$

$$\bar{\mathbf{x}}_2 = \begin{bmatrix} 10.2 \\ 3.9 \end{bmatrix} \qquad \mathbf{S}_1 = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$$

Obter uma região de confiança a 95% para $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

Como S_1 e S_2 são aproximadamente iguais, faz sentido encontrar-se uma matriz comum de covariâncias:

$$S_{comum} = \frac{(50-1)\mathbf{S}_1 + (50-1)\mathbf{S}_2}{50+50-2} = \begin{bmatrix} 2 & 1 \\ 1 & 5 \end{bmatrix}$$

Como $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 = \begin{bmatrix} -1.9 \\ .2 \end{bmatrix}$, a elipse de confiança está centrada em [-1.9; .2]´, sendo os valores e vectores próprios de $\mathbf{S}_{comum}$ obtidos através da equação

$$0 = |\mathbf{S}_{comum} - \lambda \mathbf{I}| = \begin{bmatrix} 2-l & 1 \\ 1 & 5-l \end{bmatrix} = \lambda^2 - 7\lambda + 9.$$

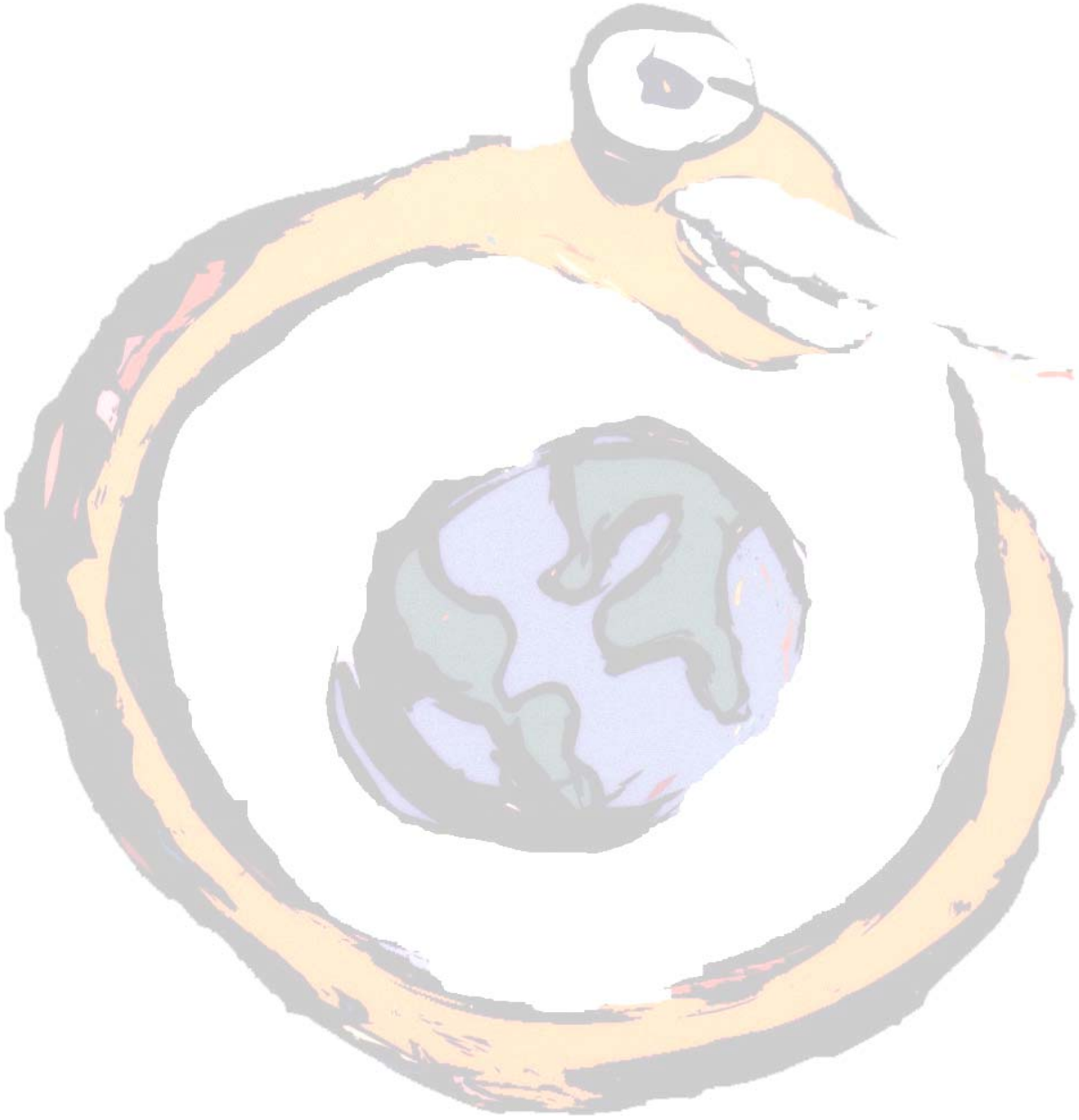
Deste modo; $\lambda_1 = 5.303$ $\mathbf{e}_1' = [\ .290;\ .957\ ]$

$\lambda_2 = 1.697$ $\mathbf{e}_2' = [\ .957;\ -.290\ ]$

Além disso; $\left(\frac{1}{n_1} + \frac{1}{n_2}\right) c^2 = \left(\frac{1}{50} + \frac{1}{50}\right) \frac{(98)(2)}{(97)} F_{2,97}(.05) = .25$

A elipse de confiança estende-se $\sqrt{\lambda_i}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)c^2} = \sqrt{\lambda_i}\sqrt{.25}$ unidades segundo o vector próprio $\mathbf{e}_i$; isto é; 1.15 unidades na direcção de $\mathbf{e}_1$ e .65 unidades na direcção de $\mathbf{e}_2$. É óbvio que $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$ não pertence à elipse sendo, portanto, possível concluirmos que os dois métodos de fabricação de sabão produzem resultados diferentes. Parece que os dois tipos de sabão têm a mesma suavidade, produzindo o segundo maior espuma.

o



7

Análise de componentes principais e análise factorial

7.1 Introdução

Os nomes que compõem o título deste capítulo são frequentemente usados de uma maneira menos precisa, chegando mesmo a acontecer que investigadores afirmem que estão a levar a cabo uma análise factorial quando, de facto, estão a proceder a uma análise de componentes principais.

Consideremos as variáveis 'temperatura' e 'horas de sol' numa determinada região. O valor 0.9 de coeficiente de correlação entre ambas as variáveis pode ser representado pelo ângulo entre estas variáveis, quando representadas vectorialmente. A questão que a análise factorial pretende responder é a seguinte

> Podem estes dois vectores ser substituídos por um único vector de referência, denominado factor, de tal modo que retenha a maior parte da informação respeitante à correlação existente entre as variáveis originais?

Intuitivamente parece que o melhor vector de referência é o que divide ao meio o ângulo de 25° entre os dois vectores. Na Figura 7.1. a variável 'temperatura' é representada por **T**, as 'horas de sol' por **S** e o vector de referência por $\mathbf{F}_1$. Este vector faz um ângulo de 12.5° com **T** e com **S**. O coseno de 12.5°, igual a 0.976, representa a correlação entre **T** e $\mathbf{F}_1$ e entre **S** e $\mathbf{F}_1$. Na linguagem da análise factorial, a correlação entre uma variável e um factor é denominada peso (*loading*) da variável no factor.

Também já vimos que o quadrado do coeficiente de correlação, R^2, representa a quantidade da variância partilhada por ambas as variáveis. No nosso caso, a quantidade de variância partilhada por **T** e $\mathbf{F}_1$ é $(0.976)^2 = 0.95$, também chamada variância do factor comum.

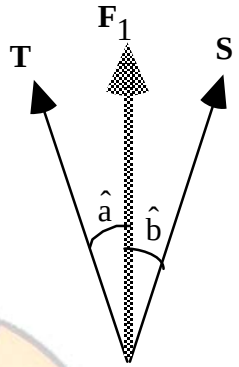


Figura 7.1 — Diagrama vectorial representando o primeiro vector de referência F_1 ($\hat{a} = \hat{b} = 12.5^{\circ}$)

A variância explicada pelo factor $\mathbf{F}_1$ através de **T** e de **S** é obtida pela soma dos quadrados dos pesos de **T** e de **S** em $\mathbf{F}_1$, isto é, $(0.9762)^2+(0.9762)^2=1.9$.

Como a variância total de cada uma das variáveis **T** e **S** é 1, a variância máxima que pode ser extraída por $\mathbf{F}_1$ é igual a 1 + 1 = 2 e, portanto, a percentagem da variância extraída por $\mathbf{F}_1$ é $\frac{1.9}{2}$ x100 = 95. Isto já nos dá 95% da representação da relação entre ambas. No entanto, para obter a imagem completa, temos de desenhar o outro vector $\mathbf{F}_2$, fazendo um ângulo recto (ou ortogonal) com $\mathbf{F}_1$.

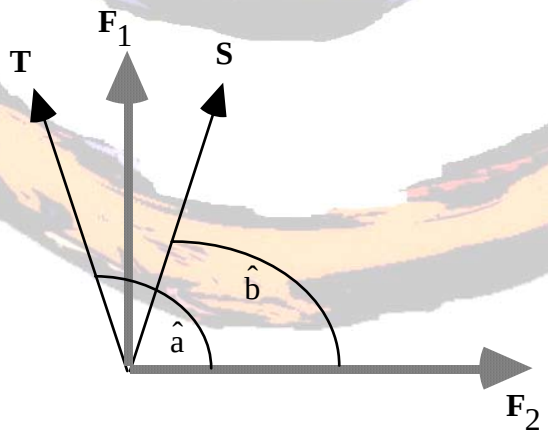


Figura 7.2 — Diagrama vectorial representando dois vectores de referência $\mathbf{F}_1$ e $\mathbf{F}_2$

$(\hat{a} = 102.5^{\circ}\ ;\ \hat{b} = 77.5^{\circ})$

Os ângulos formados por **T** e **S** com $\mathbf{F}_2$ são, respectivamente, 102.5° e 77.5°, correspondendo aos pesos $\cos(102.5^{\circ}) = -0.216$ e $\cos(77.5^{\circ}) = 0.216$. A variância extraída por $\mathbf{F}_2$ é $(-0.216)^2 + (0.216)^2 = 0.1$ e a percentagem de variância extraída é 5%.

Estes resultados podem ser resumidos na seguinte tabela:

Variáveis	Factores		Comunalidade
	1	2	
T	0.976	-0.216	1.0
S	0.976	0.216	1.0
Variância extraída	1.9	0.1	2.0
Percentagem da variância	95	5	100

A última coluna, a comunalidade, é encontrada pela soma das variâncias do factor comum. Assim, por exemplo para T, temos $(0.976)^2 + (-0.216)^2 = 1.0$ que corresponde à quantidade de variância que é partilhada com as outras variáveis.

7.2 Componentes principais

Com a análise das componentes principais pretende-se explicar a estrutura das variâncias-covariâncias através de algumas combinações lineares das variáveis originais. Embora as p componentes sejam necessárias para reproduzir toda a variabilidade do sistema, normalmente grande parte desta variabilidade pode ser

atribuída a um número menor k de componentes principais. Existirá, assim, quase tanta informação quanta a existente com as p variáveis originais. As k componentes principais podem substituir as p variáveis e o conjunto inicial de dados, com n medições em p variáveis, pode então ser reduzido num conjunto de n medições em k variáveis.

A análise das componentes principais é utilizada mais como um meio do que como um fim, constituindo um passo intermédio para investigações mais extensas, como por exemplo, as baseadas em regressões ou análises de agrupamentos (*clusters)*.

Algebricamente, as componentes principais são combinações lineares das p variáveis aleatórias X_1, X_2, …, X_p e correspondem geometricamente à selecção de um novo sistema de coordenadas. Sendo apenas dependentes da matriz $\sum$ de covariâncias (ou da matriz ρ de correlações) as componentes principais não necessitam, para a sua construção, do pressuposto da normalidade multivariada.

Sendo dada a matriz $\sum$ de covariâncias associada ao vector aleatório $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ e os pares de valores-vectores próprios $(\lambda_1, \mathbf{e}_1)$, $(\lambda_2, \mathbf{e}_2)$, …, $(\lambda_p, \mathbf{e}_p)$, onde $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ são todos não nulos, a componente principal de ordem i é dada por

$$Y_i = \mathbf{e}_i,' \mathbf{X} = \mathbf{e}_{1i} X_1 + \mathbf{e}_{2i} X_2 + \ldots + \mathbf{e}_{pi} X_p \qquad i = 1, 2, \ldots, p$$

As componentes principais são não correlacionadas [$Cor(Y_i, Y_k) = \mathbf{e}_i' \sum \mathbf{e}_k = 0 \ (i \neq k)$] e têm variâncias iguais aos valores próprios de $\sum$ [$Var(Y_i) = \mathbf{e}_i' \sum \mathbf{e}_i = \lambda_i \ (i = 1, 2, \ldots, p)$].

Além disso, se $Y_1 = \mathbf{e}_1' \mathbf{X}$, $Y_2 = \mathbf{e}_2' \mathbf{X}$, …, $Y_p = \mathbf{e}_p' \mathbf{X}$ forem as componentes principais,

- $\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} = \sum_{j=1}^{p} Var(X_i) = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \sum_{j=1}^{p} Var(Y_i)$

- Variância total da população $= \sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} = \lambda_1 + \lambda_2 + \ldots + \lambda_p$

- $\left(\begin{array}{c} \text{Proporção da variância} \\ \text{total da população} \\ \text{devida à componente} \\ \text{principal de ordem k} \end{array} \right) = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}, k = 1,2,\ldots, p$

- Os coeficientes de correlação entre as componentes Y_i e as variáveis X_k (i, k = 1, 2, …, p) são dados por $\rho_{Y_i, X_k} = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$

Exemplo 7.1: Suponhamos que as variáveis X_1, X_2 e X_3 possuem a seguinte matriz de covariâncias:

$$\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Pode ser verificado que os pares valores-vectores próprios são:

$\lambda_1 = 5.83$ $\quad \mathbf{e}_1' = [\ .383; -.924; 0\ ]$

$\lambda_2 = 2.00$ $\quad \mathbf{e}_2' = [\ 0; 0; 1\ ]$

$\lambda_3 = 0.17$ $\quad \mathbf{e}_3' = [\ .924; .383; 0\ ]$

As componentes principais são então,

$$Y_1 = \mathbf{e}_1' \mathbf{X} = .383\, X_1 - .924\, X_2$$

$$Y_2 = \mathbf{e}_2' \mathbf{X} = X_3$$

$$Y_3 = \mathbf{e}_3' \mathbf{X} = .924\, X_1 - .383\, X_2$$

Facilmente se vê, por exemplo, que

$$\begin{aligned} \text{Var}(Y_1) &= \text{Var}(.383\, X_1 - .924\, X_2) \\ &= (.383)^2 \text{Var}(X_1) + (-.924)^2 \text{Var}(X_2) - 2(.383)(-.924) \text{Cov}(X_1, X_2) \\ &= 5.83 = \lambda_1 \end{aligned}$$

$$\begin{aligned} \text{Cov}(Y_1, Y_2) &= \text{Cov}(.383\, X_1 - .924\, X_2, X_3) \\ &= .383\, \text{Cov}(X_1, X_3) - .924\, \text{Cov}(X_1, X_3) \\ &= 0 \end{aligned}$$

Verifica-se também que

$$\sigma_{11} + \sigma_{22} + \sigma_{33} = 1 + 5 + 2 = \lambda_1 + \lambda_2 + \lambda_3 = 5.83 + 2.00 + .17 = 8$$

A proporção da variância total devida à primeira componente principal é

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{5.83}{8} = .73$$

e as primeiras duas componentes principais são responsáveis por $\frac{5{,}83+2}{8} = 98\%$ da variância da população. Neste caso as componentes Y_1 e Y_2 podem substituir as três variáveis originais com pouca perda de informação.

Finalmente, como $\rho_{Y_1, X_1} = \frac{e_{11}\sqrt{l_1}}{\sqrt{\sigma_{11}}} = \frac{.383\sqrt{5.83}}{\sqrt{1}} = .925$

$$\rho_{Y_1, X_2} = \frac{e_{21}\sqrt{l_1}}{\sqrt{\sigma_{22}}} = \frac{-.924\sqrt{5.83}}{\sqrt{5}} = -.998$$

podemos concluir que X_1 e X_2 são, cada um, igualmente importantes para a primeira componente principal. Além disto,

$$\rho_{Y_2, X_1} = \rho_{Y_2, X_2} = 0 \quad \text{e} \quad \rho_{Y_2, X_3} = \frac{e_{32}\sqrt{l_2}}{\sqrt{\sigma_{33}}} = \frac{\sqrt{2}}{\sqrt{2}} = 1$$

As restantes correlações podem ser desprezadas uma vez que a terceira componente não é importante.

o

As componentes principais $y_1 = \mathbf{e}_1' \mathbf{x}$, $y_2 = \mathbf{e}_2' \mathbf{x}$, …, $y_p = \mathbf{e}_p' \mathbf{x}$ posicionam-se nas direcções dos eixos do elipsóide de densidade constante. Assim, qualquer ponto no eixo de ordem i do elipsóide tem **x** coordenadas proporcionais a $\mathbf{e}_i' \mathbf{x} = [\, e_{1i}, e_{2i}, \ldots, e_{pi} \,]$ e, necessariamente, coordenadas das componentes principais da forma [0, …, 0, y_i, 0, …, 0].

A Figura 7.3 é uma elipse de densidade constante e as componentes principais para um vector aleatório normal bivariado com $\mu = 0$ e $\rho = .75$. Podemos ver que as

componentes principais são obtidas rodando o sistema inicial de coordenadas de um ângulo θ até coincidir com os eixos da elipse de densidade constante. O mesmo é válido para p > 2.

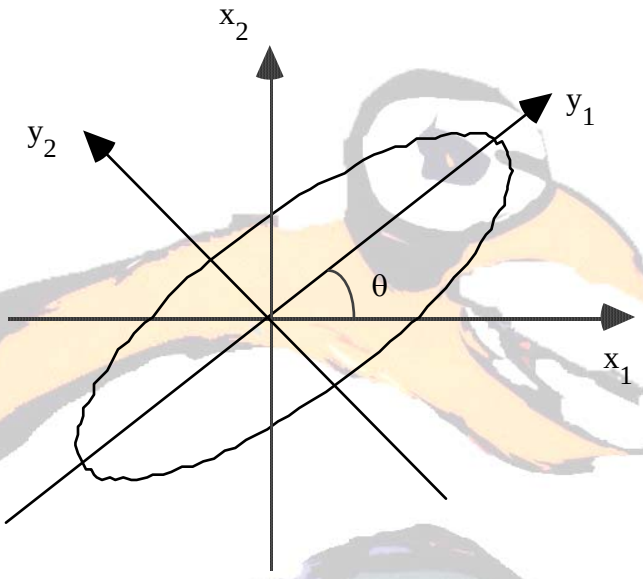


Figura 7.3 - Elipse de densidade constante e as componentes principais y_1 e y_2

Embora não necessariamente iguais às obtidas anteriormente, podemos também encontrar as componentes principais para as variáveis estandardizadas. A componente principal de ordem i das variáveis estandardizadas $\mathbf{Z}' = [Z_1, Z_2, \ldots, Z_p]$ com Cov ($\mathbf{Z}$) = ρ é dada por

$$Y_i = \mathbf{e}_i' \mathbf{Z} = \mathbf{e}_i' (\mathbf{V}^{1/2})^{-1} (\mathbf{X} - \boldsymbol{\mu}), \quad i = 1, 2, \ldots, p$$

Além disto, sendo $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ os pares valores-vectores próprios de ρ com $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$,

- $\sum_{j=1}^{p} Var(Y_i) = \sum_{j=1}^{p} Var(Z_i) = p$

- $\rho_{Y_i; Z_k} = e_{ki} \sqrt{\lambda_i} \qquad (i, k = 1,2,\ldots, p)$

- $\left(\begin{array}{c}\text{Proporção da variância total}\\ \text{da população estandardizada}\\ \text{devida à componente}\\ \text{principal de ordem k}\end{array}\right) = \frac{\lambda_k}{p}, \qquad k = 1,2,\ldots,p$

Exemplo 7.2: Consideremos a matriz de covariâncias $\Sigma = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}$ e a correspondente matriz de correlações $\rho = \begin{bmatrix} 1 & .4 \\ .4 & 1 \end{bmatrix}$

Os pares valores-vectores próprios de Σ são $\lambda_1 = 100.16$ $\mathbf{e}'_1 = [\ .040;\ .999\ ]$

$\lambda_2 = .84$ $\mathbf{e}'_2 = [\ .999.\ -.040\ ]$

e, para ρ, $\lambda_1 = 1 + \rho = 1.4$ $\mathbf{e}'_1 = [\ .707.\ .707\ ]$

$\lambda_2 = 1 - \rho = .6$ $\mathbf{e}'_2 = [\ .707;\ -.707\ ]$

As correspondentes componentes principais são então, para Σ:

$$Y_1 = .040\ X_1 + .999\ X_2$$

$$Y_2 = .999\ X_1 - .040\ X_2$$

e para ρ:

$$Y_1 = .707\ Z_1 + .707\ Z_2 = .707 \left(\frac{X_1 - \mu_1}{1}\right) + .707 \left(\frac{X_2 - \mu_2}{10}\right) = .707\ (X_1 - \mu_1) + .0707\ (X_2 - \mu_2)$$

$$Y_2 = .707\ Z_1 - .707\ Z_2 = .707 \left(\frac{X_1 - \mu_1}{1}\right) - .707 \left(\frac{X_2 - \mu_2}{10}\right) = .707\ (X_1 - \mu_1) - .0707\ (X_2 - \mu_2)$$

Devido à sua maior variância, X_2 domina completamente a primeira componente principal obtida a partir de $\sum$. Esta primeira componente principal explica $\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{100.16}{101} = .992$ da variância total da população.

Contudo, quando as variáveis X_1 e X_2 são estandardizadas, as variáveis resultantes contribuem de modo idêntico para as componentes principais obtidas de ρ. Assim, como

$\rho_{Y_1, Z_1} = e_{11}\sqrt{\lambda_1} = .707\sqrt{1.4} = .837$ e $\rho_{Y_1, Z_2} = e_{21}\sqrt{\lambda_1} = .707\sqrt{1.4} = .837$

a primeira componente principal explica $\frac{\lambda_1}{p} = \frac{1.4}{2} = .7$ da variância total da população estandardizada.

o

Do exemplo anterior pode concluir-se que as componentes principais obtidas de $\sum$ são diferentes das obtidas de ρ. Além disso, um conjunto de componentes principais não é uma função simples do outro, dando, portanto valor à estandardização.

Exemplo 7.3: Sejam x_1, x_2, x_3, x_4 e x_5 observações semanais das taxas de retorno das acções de cinco empresas (Allied Chemical, DuPont, Union Carbide, Exxon e Texaco). Após 100 semanas consecutivas, obteve-se

$$\bar{\mathbf{x}}' = [\ .0054;\ .0048;\ .0057;\ .0063;\ .0037\ ]$$

e

$$R = \begin{bmatrix} 1.000 & .577 & .509 & .387 & .462 \\ .577 & 1.000 & .599 & .389 & .322 \\ .509 & .599 & 1.000 & .436 & .426 \\ .387 & .389 & .436 & 1.000 & .523 \\ .462 & .322 & .426 & .523 & 1.000 \end{bmatrix}$$

Os valores próprios e os correspondentes vectores próprios normalizados de R são

$\lambda_1 = 2.857$ $\quad \mathbf{e}'_1 = [\ .464, .457, .470, .421, .421\ ]$

$\lambda_2 = .809$ $\quad \mathbf{e}'_2 = [\ .240, .509, .260, -.526, -.582\ ]$

$\lambda_3 = .540$ $\quad \mathbf{e}'_3 = [\ -.612, .178, .335, .541, -.435\ ]$

$\lambda_4 = .452$ $\quad \mathbf{e}'_4 = [\ .387, .206, -.662, .472, -.382\ ]$

$\lambda_5 = .343$ $\quad \mathbf{e}'_5 = [\ -.451, .676, -.400, -.176, .385\ ]$

Usando as variáveis estandardizadas, obtermos as primeiras duas componentes principais

$$y_1 = \mathbf{e}'_1\ \mathbf{z} = .464\, z_1 + .457\, z_2 + .470\, z_3 + .421\, z_4 + .421\, z_5$$

$$Y_2 = \mathbf{e}'_2\ \mathbf{z} = .240\, z_1 + .509\, z_2 + .260\, z_3 - .526\, z_4 + .582\, z_5$$

Estas componentes, que explicam $\left(\frac{\lambda_1 + \lambda_2}{p}\right)100\% = \left(\frac{2.857 + .809}{5}\right)100\% = 73\%$ têm uma interpretação interessante. A primeira componente consiste num índice das cinco acções e pode ser chamada 'componente de mercado'. A segunda componente representa um contraste entre as acções de empresas químicas (Allied Chemical, DuPont e Union Carbide) e as acções das empresas petrolíferas (Exxon e Texaco) podendo ser denominado componente industrial.

As restantes componentes, de difícil interpretação, representam no seu conjunto a variação provavelmente específica de cada acção.

o

7.3 Análise factorial

O objectivo essencial da análise factorial é descrever, se possível, as relações de covariância entre as várias variáveis em termos de um número reduzido de quantidades aleatórias subjacentes, mas não observáveis, chamadas factores.

A análise factorial pode ser vista como uma extensão da análise das componentes principais, uma vez que ambas podem ser encaradas como aproximações à matriz das covariâncias. Contudo, a aproximação feita pelo modelo da análise factorial é mais elaborada e centra-se na análise da consistência dos dados com uma estrutura pré-definida.

Considerando o vector aleatório **X** de dados observados, com p componentes, média **μ** e matriz de covariâncias ∑, o modelo factorial parte do conceito de que **X** é linearmente dependente de algumas variáveis não observáveis F_1, F_2, …, F_m, chamados factores comuns, e p fontes de variação ε_1, ε_2, …, ε_m, chamados erros ou factores específicos.

Numa forma matricial, o modelo de análise factorial é

$$\underset{(p\times 1)}{X-\mu} = \underset{(p\times m)}{L}\ \underset{(m\times 1)}{F} + \underset{(p\times 1)}{\varepsilon}$$

ou seja,

$$\mathbf{X}_1 - \boldsymbol{\mu}_1 = \ell_{11}\,\mathbf{F}_1 + \ell_{12}\,\mathbf{F}_2 + \ldots + \ell_{1m}\,\mathbf{F}_m + \varepsilon_1$$

$$\mathbf{X}_2 - \boldsymbol{\mu}_2 = \ell_{21}\,\mathbf{F}_1 + \ell_{22}\,\mathbf{F}_2 + \ldots + \ell_{2m}\,\mathbf{F}_m + \varepsilon_2$$

…

$$\mathbf{X}_p - \boldsymbol{\mu}_p = \ell_{p1}\,\mathbf{F}_1 + \ell_{p2}\,\mathbf{F}_2 + \ldots + \ell_{pm}\,\mathbf{F}_m + \varepsilon_p$$

onde μ_i representa a média da variável i, ε_i o factor específico de ordem i, F_i o factor comum de ordem i e ℓ_{ij} o peso (*loadings*) da variável i no factor j.

Além disso, as variáveis aleatórias $F_1, F_2, \ldots, F_m$, assim como os erros $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m$ não são observáveis, o que permite distinguir este modelo da representação linear onde os **X** independentes podem ser observados.

Para este modelo partimos do pressuposto que

- $E(\mathbf{F}) = \underset{(m\times 1)}{0}$; $\quad Cov(\mathbf{F}) = E[\mathbf{FF'}] = \underset{(m\times m)}{I}$
- $E(\boldsymbol{\varepsilon}) = \underset{(p\times 1)}{0}$; $\quad Cov(\boldsymbol{\varepsilon}) = E[\boldsymbol{\varepsilon\varepsilon'}] = \underset{(p\times p)}{\Psi} = \begin{bmatrix} \Psi_1 & 0 & \ldots & 0 \\ 0 & \Psi_2 & \ldots & 0 \\ . & . & \ldots & . \\ 0 & 0 & \ldots & \Psi_p \end{bmatrix}$
- **F** e $\boldsymbol{\varepsilon}$ são independentes; isto é; $Cov(\boldsymbol{\varepsilon}; \mathbf{F}) = E(\boldsymbol{\varepsilon}\,\mathbf{F'}) = \underset{(p\times m)}{0}$

Como já atrás vimos, comunalidade representa a parte da variância da variável i devida aos m factores comuns. Deste modo, a variância de X_i pode ser dada por

$$Var(X_i) = \text{comunalidade } h_i^2 + \text{variância específica } \psi_i$$

$$\sigma_{ii} = \left[\ell_{i1}^2 + \ell_{i1}^2 + \ldots + \ell_{i1}^2\right] + \psi_i$$

Exemplo 7.4: Consideremos a matriz de covariâncias

$$\Sigma = \begin{bmatrix} 19 & 30 & 2 & 12 \\ 30 & 57 & 5 & 23 \\ 2 & 5 & 38 & 47 \\ 12 & 23 & 47 & 68 \end{bmatrix}$$

A igualdade $\Sigma = \mathbf{L}\,\mathbf{L}' + \Psi$, ou seja,

$$\begin{bmatrix} 19 & 30 & 2 & 12 \\ 30 & 57 & 5 & 23 \\ 2 & 5 & 38 & 47 \\ 12 & 23 & 47 & 68 \end{bmatrix} = \begin{bmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} 4 & 7 & -1 & 1 \\ 1 & 2 & 6 & 8 \end{bmatrix} + \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

pode ser verificada pela álgebra matricial. Deste modo, Σ tem a estrutura produzida por um modelo factorial ortogonal com m=2.

Sendo $\mathbf{L} = \begin{bmatrix} \ell_{11} & \ell_{12} \\ \ell_{21} & \ell_{22} \\ \ell_{31} & \ell_{32} \\ \ell_{41} & \ell_{42} \end{bmatrix} = \begin{bmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{bmatrix}$ e $\Psi = \begin{bmatrix} \Psi_1 & 0 & 0 & 0 \\ 0 & \Psi_2 & 0 & 0 \\ 0 & 0 & \Psi_3 & 0 \\ 0 & 0 & 0 & \Psi_4 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$

a comunalidade de X_1 é

$$h_1^2 = \ell_{11}^2 + \ell_{12}^2 = 4^2 + 1^2 = 17$$

e a variância de X_1 pode ser decomposta da seguinte maneira

$$\sigma_{11} = h_1^2 + \psi_1 = 17 + 2 = 19$$

As restantes variáveis podem ser decompostas de maneira análoga.

o

Infelizmente, quando o número m de factores é muito menor do que o número p de variáveis, a maioria das matrizes de covariância não podem ser factorizadas da forma $\mathbf{L}\,\mathbf{L}' + \Psi$. Há, neste caso, necessidade de se utilizarem métodos de estimação apropriados para $\mathbf{L}$ e Ψ e se proceder previamente a algumas transformações ortogonais, sabendo nós de antemão que, quer os pesos, quer as comunalidades, não são alterados por qualquer transformação ortogonal.

Comecemos pela estimação. A matriz de covariância amostral S é um estimador da matriz $\sum$ de covariância desconhecida da população. Se os elementos fora da diagonal de S são pequenos (ou os correspondentes valores em R essencialmente nulos), as variáveis não estão relacionadas e a análise factorial não se apresenta de muita utilidade, uma vez que, neste caso, os factores específicos desempenham um papel fundamental, não fazendo sentido a construção dos factores comuns.

Se $\sum$ se desvia significativamente de uma matriz diagonal, então faz sentido usar-se um modelo factorial, sendo primeiramente necessário estimar-se os pesos ℓ_{11} e as variâncias específicas ψ_1. De entre os métodos existentes para a estimação destes parâmetros, usaremos, neste capítulo, apenas o método das componentes principais, que passaremos a expor. As soluções encontradas poderão então ser rodadas (através de transformações) com vista a uma melhor interpretação.

A análise factorial de componentes principais da matriz amostral S de covariâncias, usando uma decomposição espectral, é especificada em termos dos

seus pares de valores-vectores próprios estimados ($\hat{\ell}_1$, $\hat{\mathbf{e}}_1$); ($\hat{\ell}_2$, $\hat{\mathbf{e}}_2$), …, ($\hat{\ell}_p$, $\hat{\mathbf{e}}_p$) onde $\hat{\ell}_1 \geq \hat{\ell}_2 \geq \ldots \geq \hat{\ell}_p$. Sendo m < p o número dos factores comuns; a matriz dos pesos factoriais estimados $\{\tilde{\ell}_{ij}\}$ é dada por

$$\tilde{\mathbf{L}} = \left[\sqrt{\hat{\ell}_1}\hat{\mathbf{e}}_1 \mid \sqrt{\hat{\ell}_2}\hat{\mathbf{e}}_2 \mid \ldots \mid \sqrt{\hat{\ell}_m}\hat{\mathbf{e}}_m\right]$$

As variâncias específicas estimadas são fornecidas pelos elementos da diagonal da matriz $S - \tilde{L}\tilde{L}'$;

$$\Psi = \begin{bmatrix} \tilde{\psi}_1 & 0 & \ldots & 0 \\ 0 & \tilde{\psi}_2 & \ldots & 0 \\ . & . & \ldots & . \\ 0 & 0 & \ldots & \tilde{\psi}_p \end{bmatrix} \qquad \text{com} \qquad \tilde{\psi}_i = s_{ii} - \sum_{j=1}^{m} \tilde{\ell}_{ij}^2$$

e as comunalidades são estimadas da forma que se segue

$$\tilde{h}_i^2 = \tilde{\ell}_{i1}^2 + \tilde{\ell}_{i2}^2 + \ldots + \tilde{\ell}_{im}^2$$

Há que notar que a análise factorial das componentes principais da matriz amostral de correlações é obtida de maneira idêntica, começando pela matriz R em vez de S. Além disso, os pesos factoriais estimados para um determinado factor não são alterados quando o número de factores aumenta. A proporção da variância amostral total devida ao factor j é dada por:

- para uma análise factorial de S:

$$\left(\begin{array}{c}\text{Proporção da variância amostral} \\ \text{total devida ao factor j}\end{array}\right) = \frac{\hat{\lambda}_j}{s_{11} + s_{22} + \ldots + s_{pp}}$$

- para uma análise factorial de R:

$$\left(\begin{array}{c}\text{Proporção da variância amostral}\\ \text{total devida ao factor j}\end{array}\right) = \frac{\hat{\lambda}_j}{p}$$

Exemplo 7.5: Voltando ao Exemplo 8.3 referente às n=100 observações semanais das taxas de retorno das acções de p=5 empresas químicas e onde se encontraram as primeiras duas componentes principais obtidas a partir de R, é fácil determinar as soluções das componentes principais para o modelo ortogonal com m=1 e m=2. Assim, para encontrar os pesos factoriais estimados basta multiplicar os coeficientes das componentes principais amostrais (vectores próprios de R) pela raízes quadradas dos correspondentes valores próprios.

A seguir, são apresentados os pesos factoriais estimados F, as comunalidades, as variâncias específicas e a proporção da variância total (estandardizada) amostral explicada por cada factor, para as soluções com m=1 e com m=2.

	Solução 1 factor		Solução 2 factores		
	Pesos factoriais estimados	Variâncias específicas	Pesos factoriais estimados		Variâncias específicas
Variável	F_1	$\tilde{\Psi}_i = 1 - \tilde{h}_i^2$	F_1	F_2	$\tilde{\Psi}_i = 1 - \tilde{h}_i^2$
1. Allied Chemical	.783	.39	.783	-.217	.34
2. DuPont	.773	.40	.773	-.458	.19
3. Union Carbide	.794	.37	.794	-.234	.31
4. Exxon	.713	.49	.713	.472	.27
5. Texaco	.712	.49	.712	.524	.22
Proporção da variância total (estandardizada) amostral explicada	.571		.571	.733	

A matriz residual correspondente à solução para m=2 factores é

$$\mathbf{R} - \hat{\mathbf{L}}\hat{\mathbf{L}}' - \hat{\Psi} = \begin{bmatrix} 0 & -.127 & -.164 & -.069 & .017 \\ -.127 & 0 & -.122 & .055 & .012 \\ -.164 & -.122 & 0 & -.019 & -.017 \\ -.069 & .055 & -.019 & 0 & -.232 \\ .017 & 012 & -.017 & -.232 & 0 \end{bmatrix}$$

A proporção da variância total explicada pela solução com dois factores é apreciavelmente maior do que a correspondente à solução com apenas um factor. Mais uma vez se vê que o primeiro factor F_1 representa condições económicas gerais e pode ser chamado factor de mercado, todas as acções têm um peso alto e todos os pesos são mais ou menos iguais. O segundo factor, que permite a separação das empresas químicas com acções de petróleo das empresas químicas sem acções de petróleo, pode ser denominado factor de indústria.

o

Como já vimos anteriormente, todos os pesos factoriais obtidos pelos pesos iniciais através de uma transformação ortogonal têm idêntica capacidade de produzir a matriz de covariância (ou de correlação). Ora, pela álgebra matricial sabemos que uma transformação ortogonal corresponde a uma rotação rígida dos eixos coordenados. Por esta razão, a uma transformação ortogonal dos pesos factoriais damos o nome de rotação factorial.

Se $\hat{\mathbf{L}}$ é uma matriz p×m de pesos factoriais estimados obtidos por um qualquer método, então $\hat{\mathbf{L}}^* = \hat{\mathbf{L}}\,\mathbf{T}$ (onde $\mathbf{TT}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$) é a matriz p×m de pesos após rotação. Como consequência directa da rotação, a matriz residual mantêm-se inalterada, assim como as variâncias específicas $\hat{\Psi}_i$ ¡ e as comunalidades $\hat{h}_i^2$. Isto significa que, sob o ponto de vista matemático, é imaterial usarmos $\hat{\mathbf{L}}$ ou $\hat{\mathbf{L}}^*$.

Exemplo 7.6: Consideremos a seguinte matriz de correlações referentes às notas em p=6 áreas de n=220 alunos de uma escola:

$$\mathbf{R} = \begin{array}{c} \begin{array}{cccccc} \text{Port} & \text{Franc} & \text{Hist} & \text{Aritm} & \text{Álgeb} & \text{Geomet} \end{array} \\ \begin{bmatrix} 1.0 & .439 & .410 & .288 & .329 & .248 \\ & 1.0 & .351 & .354 & .320 & .329 \\ & & 1.0 & .164 & .190 & .181 \\ & & & 1.0 & .595 & .470 \\ & & & & 1.0 & .464 \\ & & & & & 1.0 \end{bmatrix} \end{array}$$

A solução para m=2 factores comuns é apresentada a seguir:

	Factores rodados		Comunalidades
	F1	F2	$\hat{h}_i^2$
1. Português	.553	.429	.490
2. Francês	.568	.288	.406
3. História	.392	.450	.356
4. Aritmética	.740	-.273	.623
5. Álgebra	.724	-.211	.569
6. Geometria	.595	-.132	.372

Todas as variáveis têm pesos positivos no primeiro factor, factor geral de inteligência. No entanto, em relação ao segundo factor, há bipolarização entre as disciplinas matemáticas e as não-matemáticas. Os pares $(\tilde{\ell}_{i1}^2; \tilde{\ell}_{i2}^2)$ de pesos factorias estão apresentados na Figura 7.4.

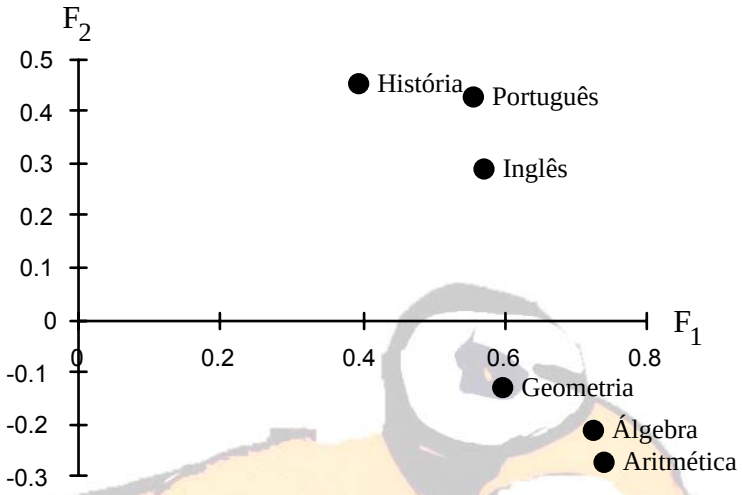


Figura 7.4 — Pesos factoriais

Rodando o sistema de eixos de θ= 20º, fazemos com que o transformado do eixo F_1 passe pelo ponto (γ^2_{41} ; γ^2_{42}), como o representado na Figura 7.5.

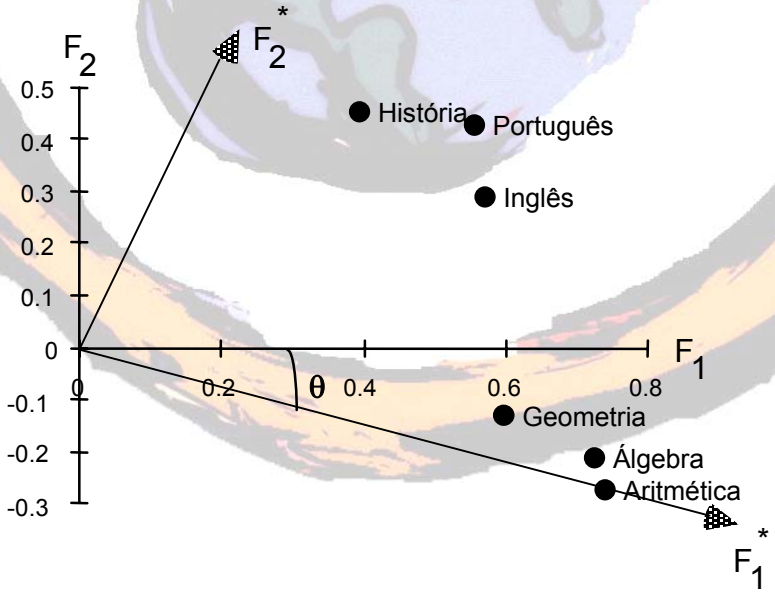


Figura 7.5 — Rotação factorial

Quando isto é feito todos os pontos se encontram no primeiro quadrante (todos os pesos factoriais são positivos) e os dois grupos de variáveis são evidenciados. Isto corresponde à seguinte tabela de pesos estimados após rotação.

Variável	Pesos factoriais estimados após rotação F_1^*	F_2^*	Comunalidades $\hat{h}^{*2}_i = \hat{h}^2_i$
1. Português	.369	.594	.490
2. Francês	.433	.467	.406
3. História	.211	.558	.356
4. Aritmética	.789	.001	.623
5. Álgebra	.752	.054	.569
6. Geometria	.604	.083	.372

Como se pode verificar; as comunalidades não se alteraram.

o

Ora, esta rotação pode ser conseguida analiticamente, por exemplo através do critério varimax. Considerando $\tau^*_{ij} = \frac{\hat{\tau}^*_{ij}}{\hat{h}^*_i}$, o procedimento varimax selecciona a transformação ortogonal **T** tal que maximiza

$$V = \frac{1}{p} \sum_{j=1}^{m} \left[\sum_{i=1}^{p} \tau^{*4}_{ij} - \frac{\left(\sum_{i=1}^{p} \tau^{*2}_{ij} \right)^2}{p} \right]$$

É importante salientar-se que as rotações ortogonais são apropriadas para modelos factoriais onde se pressupõe que os factores são independentes. Caso isso se não verifique existem rotações oblíquas (não ortogonais), mas que não se regem pelo modelo atrás indicado.

For fim, falta ainda debruçarmo-nos um pouco sobre um problema prático referente ao número de factores a escolher e a utilizar numa análise específica. Um dos critérios mais vulgares é reter apenas factores com valores próprios maiores do que 1, quando usada a matriz de correlações. Outra alternativa é analisar o gráfico dos valores próprios e parar a análise no ponto onde a linha deste gráfico começa a ser quase paralela com o eixo horizontal. Este última alternativa; denominada teste de base de montanha (*scree test*) está ilustrada na Figura 7.6.

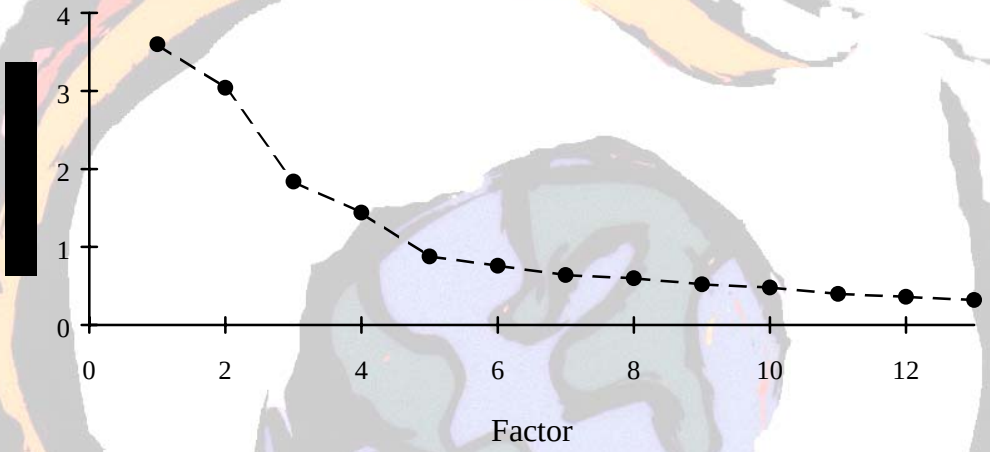
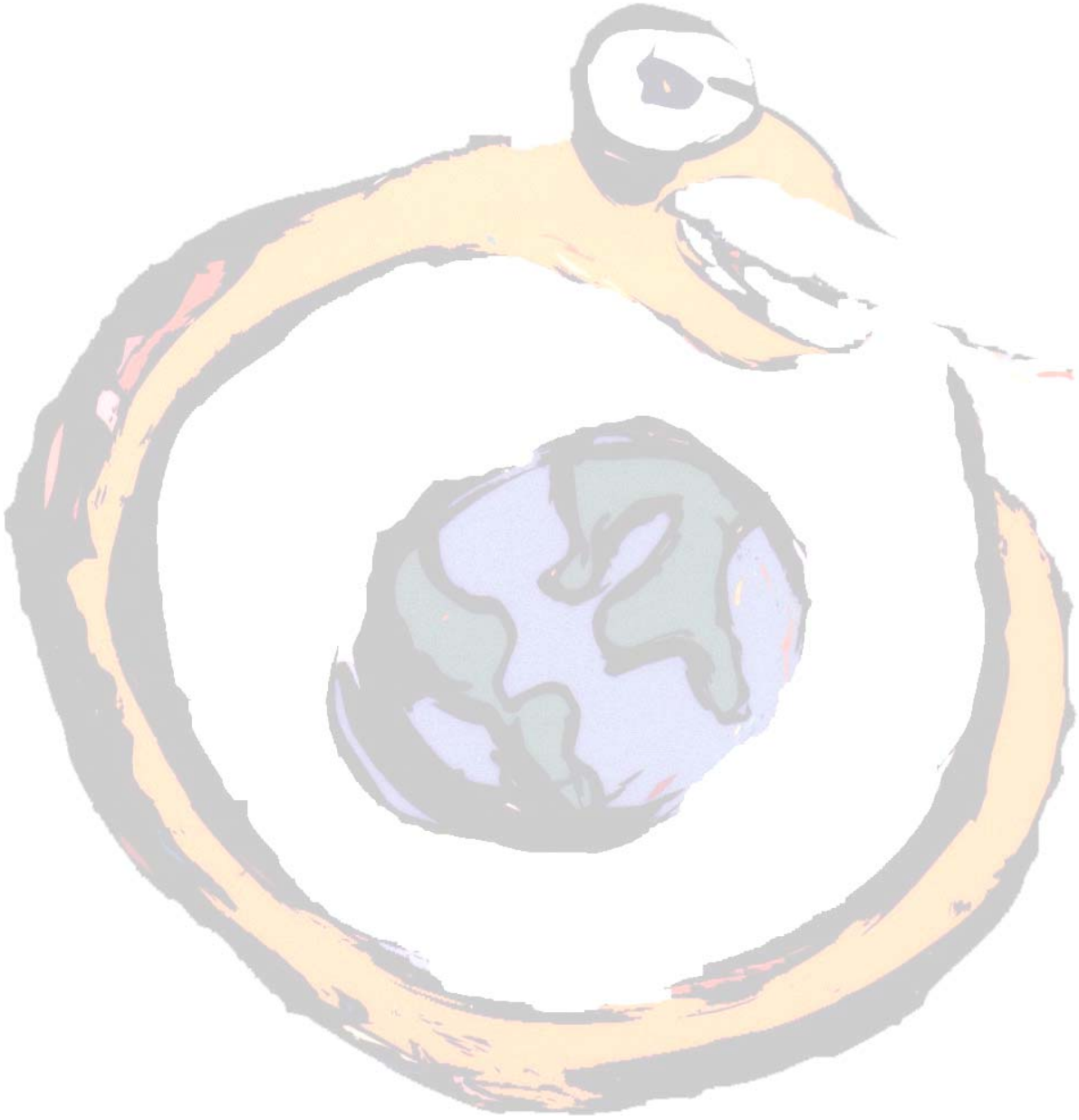


Figura 7.6 — Exemplo de um teste de base de montanha

Segundo este gráfico o investigador concluiria que não deveriam ser extraídos mais de cinco factores.



8

Análise de agrupamentos (*clusters*)

8.1 Introdução

Uma outra técnica exploratória de dados é aquela que pesquisa a existência de grupos naturais de indivíduos ou de variáveis. A aplicação desta técnica não pressupõe qualquer característica da estrutura do agrupamento; apenas se baseia em medidas de semelhança ou de distância entre objectos e na escolha de critérios de agregação.

De uma maneira geral, esta análise passa pelas seguintes fases:

1. Selecção da amostra de indivíduos a agrupar;
2. Definição de variáveis para permitir o agrupamento dos indivíduos;
3. Definição de uma medida de semelhança ou de distância;
4. Escolha de um critério de agregação ou desagregação
5. Validação dos resultados encontrados.

8.2 Medidas de semelhança

A escolha das medidas de semelhança envolve sempre uma grande componente de subjectividade para além das características das variáveis e das escalas usadas para a medição. Normalmente os indivíduos são agrupados à custa

de distâncias. As variáveis podem, por exemplo, ser agrupadas com base no coeficiente de correlação.

8.2.1 Medidas de distância

De entre as várias medidas normalmente utilizadas para determinar a distância entre elementos de uma matriz de dados, destacam-se as seguintes:

1. Distância Euclideana – a distância entre dois indivíduos i e j é a raiz quadrada do somatório dos quadrados das diferenças entre os valores de i e j para todas as variáveis:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} .(x_{ik} - x_{jk})^2}$$

2. Quadrado da distância Euclideana – a distância entre dois indivíduos i e j é o somatório dos quadrados das diferenças entre os valores de i e j para todas as variáveis:

$$d_{ij}^2 = \sum_{k=1}^{p} .(x_{ik} - x_{jk})^2$$

3. Distância absoluta (*city block*) – a distância entre dois indivíduos i e j é o somatório dos valores absolutos das diferenças entre os valores de i e j para todas as variáveis:

$$d_{ij} = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right| .$$

4. Distância de Minkowski – generalização da distância absoluta (para m=1) e da distância Euclideana (para m=2):

$$d_{ij} = \left(\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^{m} \right)^{1/m}$$

5. Distância generalizada (de Mahalanobis) – medida que utiliza a matriz das variâncias:

$$d_{ij} = \left(x_i - x_j\right)' \sum^{-1} \left(x_i - x_j\right)$$

Para exemplificar a utilização destas medidas consideremos a seguinte matriz de 5 observações em 3 variáveis:

	X_1	X_2	X_3
1	1,06	9,2	151
2	1,10	9,2	245
3	1,34	13,0	168
4	1,43	15,4	113
5	1,16	11,7	104

As matrizes a seguir apresentam as medições das distâncias para a matriz de dados.

Distância Euclideana

	1	2	3	4
2	94,0			
3	17,4	77,1		
4	38,5	132,1	55,0	
5	47,1	141,0	64,0	9,7

Quadrado da distância Euclideana

	1	2	3	4
2	8836,0			
3	303,5	5943,5		
4	1482,6	17462,5	3030,8	
5	2215,3	19887,3	4097,7	94,8

Distância absoluta (*city block*)

	1	2	3	4
2	94,0			
3	21,1	81,0		
4	44,6	138,5	57,5	

5	49,6	143,6	65,5	13,0

Distância generalizada (de Mahalanobis)

	1	**2**	**3**	**4**
2	36,6			
3	21,4	34,0		
4	40,0	35,5	19,0	
5	21,2	33,8	40,0	18,8

Como se pode ver, principalmente quando a distância generalizada é comparada com as outras, as variáveis que apresentam variações e unidades de medidas elevadas tendem a anular o efeito das outras variáveis.

8.2.2 Medidas de associação

Por outro lado, todas estas variáveis são quantitativas. No entanto, também as variáveis qualitativas podem ser introduzidas neste tipo de análise à custa da sua transformação em variáveis binárias, com o valor 1 no casos da presença de uma determinada característica de interesse e 0 nos casos contrários.

Considerando os indivíduos i e j, medidos através de p variáveis binárias, constrói-se a tabela seguinte

	Indivíduo j		**Totais**
Indivíduo i	**1**	**0**	
1	a	b	a + b
0	c	d	c + d
Totais	a + c	b + d	

onde ***a*** corresponde ao número de características existentes (valor 1) em ambos os indivíduos, ***d*** ao número de características ausentes (valor 0) em ambos os indivíduos,

b ao número de características presentes em i e ausentes em j, e ***c*** ao número de características ausentes em i e presentes em j.

Alguns dos coeficientes de emparelhamento e de semelhança são os apresentados a seguir:

1	$\frac{a+d}{a+b+c+d}$	Igual peso às as presenças e as ausências simultâneas;
2	$\frac{2(a+d)}{2(a+d)+b+c}$	Peso duplo às presenças e ausências simultâneas;
3	$\frac{a+d}{a+d+2(b+c)}$	Peso duplo às situações discordantes; inclusão das ausências simultâneas
4	$\frac{2a}{2a+b+c}$	Peso duplo às presenças ausências simultâneas; exclusão das ausências simultâneas.
5	$\frac{a}{a+2(b+c)}$	Peso duplo as situações discordantes; exclusão das ausências simultâneas.
6	$\frac{a}{b+c}$	Quociente entre presenças simultâneas e situações discordantes; exclusão das ausências simultâneas.

Suponhamos agora outros cinco indivíduos com as seguintes características:

Indivíduo	Altura (cm)	Peso (Kg)	Olhos	Cabelo	Canhoto	Sexo
1	173	64	Verdes	Louros	Não	Fem
2	185	84	Castanhos	Castanhos	Não	Masc
3	170	75	Azuis	Louros	Não	Masc
4	163	54	Castanhos	Castanhos	Não	Fem
5	193	95	Castanhos	Castanhos	Sim	Masc

Definamos as seis variáveis binárias X_1, X_2, X_3, X_4, X_5, e X_6 do seguinte modo:

$$X_1 = \begin{cases} 1 & altura \geq 183cm \\ 0 & altura < 183cm \end{cases} \qquad X_4 = \begin{cases} 1 & cabelo\,louro \\ 0 & cabelo\,nâo\,louro \end{cases}$$

$$X_2 = \begin{cases} 1 & peso \geq 68\,Kg \\ 0 & peso < 68\,Kg \end{cases} \qquad X_5 = \begin{cases} 1 & não\,canhoto \\ 0 & canhoto \end{cases}$$

$$X_3 = \begin{cases} 1 & cabelos\,cas\tanh os \\ 0 & cabelos\,não\,cas\tanh os \end{cases} \qquad X_6 = \begin{cases} 1 & sexo\ fe\min ino \\ 0 & sexo\,masculino \end{cases}$$

As pontuações para os indivíduos 1 e 2 para as 6 variáveis são

Indivíduo	X_1	X_2	X_3	X_4	X_5	X_6
1	0	0	0	1	1	1
2	1	1	1	0	1	0

E o número de coincidências são indicadas pela tabela de duas entradas:

	Indivíduo 2		Totais
Indivíduo 1	**1**	**0**	
1	1	2	3
0	3	0	3
Totais	4	2	6

Utilizando o primeiro coeficiente de semelhança, obtemos $\frac{a+d}{a+b+c+d} = \frac{1+0}{6} = \frac{1}{6}$ e, continuando, a seguinte matriz:

	1	2	3	4	5
1	1				
2	$\frac{1}{6}$	1			
3	$\frac{4}{6}$	$\frac{3}{6}$	1		
4	$\frac{4}{6}$	$\frac{3}{6}$	$\frac{2}{6}$	1	
5	0	$\frac{5}{6}$	$\frac{2}{6}$	$\frac{2}{6}$	1

o que demonstra que os indivíduos 2 e 5 são mais semelhantes entre si e que os indivíduos 1 e 5 são menos semelhantes entre si. Os dois subgrupos que se podiam criar eram (1 3 5) e (2 5).

Todos estes coeficientes de semelhança s_{ij} podem variar entre 0 e 1 e a sua relação com as distâncias d_{ij} permite a sua construção através da fórmula:

$$s_{ij} = \frac{1}{1 + dij}$$

8.3 Critérios de agregação e desagregação

No processo de agrupamento há necessidade de estimar as distâncias entre os grupos já formados e outros grupos ou indivíduos. Também aqui não existe o melhor método de desagregação, tendo o investigador que utilizar vários critérios e comparar os resultados.

De entre os critério de agregação mais utilizados podemos citar o critério do vizinho mais próximo (*single linkage*), o critério do vizinho mais afastado (*complete linkage*), o critério da média dos grupos (average linkage), o critério do centróide e o critério de Ward

8.3.1 Critério do vizinho mais próximo (*single linkage*)

Dados dois grupos (i,j) e (k), a distancia entre eles é igual à menor distância entre os elementos dos dois grupos, isto é,

$$d_{(i,j)k} = \min\{d_{ik}; d_{jk}\}$$

Com este critério, cada indivíduo terá mais tendência para se agrupar a um grupo já definido do que para formar o núcleo de um novo grupo. Isto constitui uma desvantagem, principal responsável pela fraca utilização deste critério.

8.3.2 Critério do vizinho mais afastado (*complete linkage*)

Dados dois grupos (i,j) e (k), a distancia entre eles é igual à maior distância entre os elementos dos dois grupos, isto é,

$$d_{(i,j)k} = \max\{d_{ik}; d_{jk}\}$$

Com este critério, cada grupo passa a ser definido como o conjunto dos indivíduos em que cada um é mais semelhante a todos os outros do grupo do que a qualquer outro elemento. Os grupos assim criados são mais compactos

8.3.3 Critério da média dos grupos (*average linkage*)

Dados dois grupos (i,j) e (k), a distancia entre eles é a média entre todos os pares de indivíduos constituídos por todos os elementos dos dois grupos.

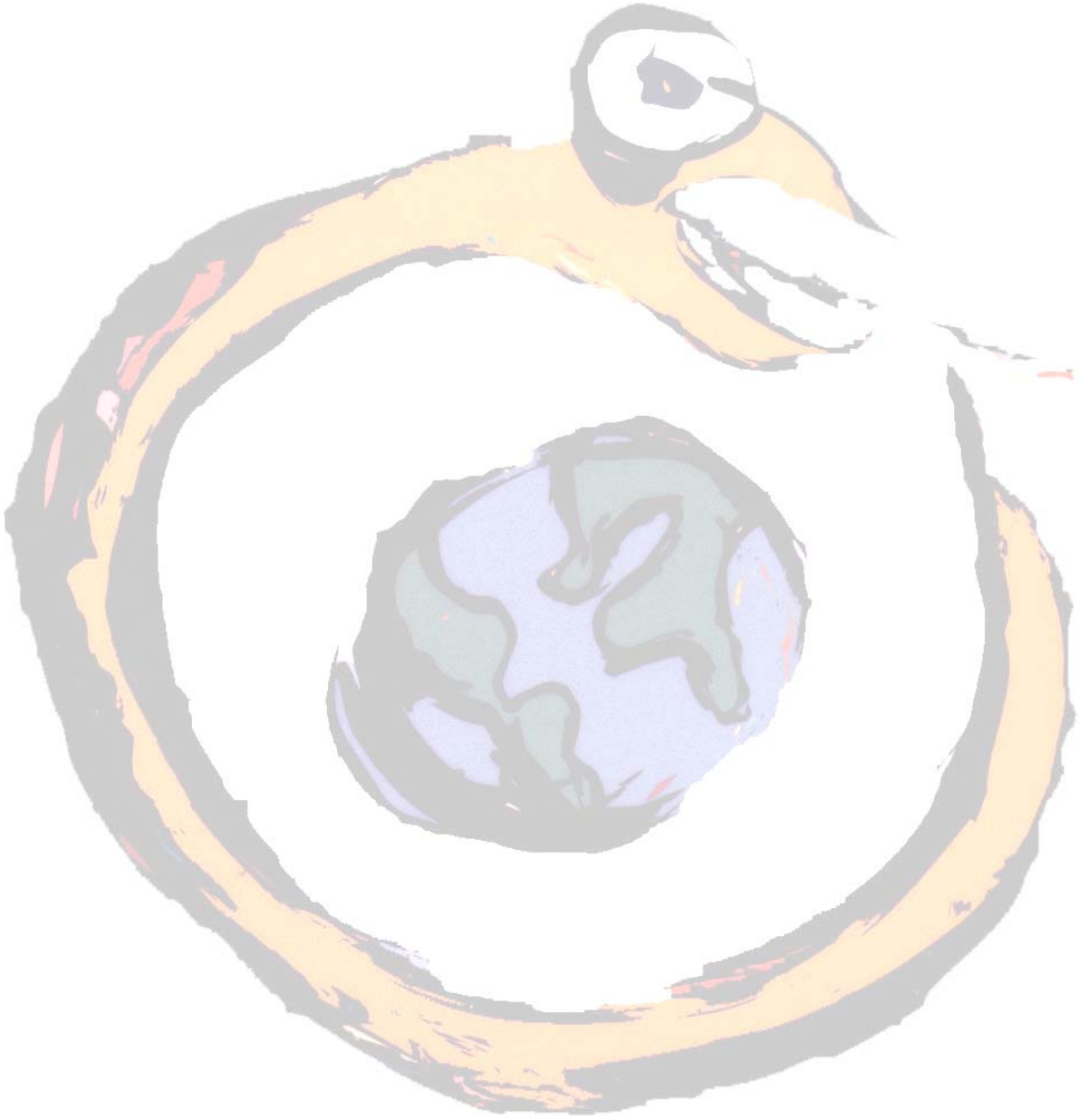
Constitui-se como uma estratégia intermédia das restantes.

8.3.4 Critério do centróide

Dados dois grupos, a distancia entre eles é a distância entre os respectivos centróides, média das variáveis caracterizadoras dos indivíduos de cada grupo.

8.3.5 Critério de Ward

Este critério baseia-se na comparação entre a aplicação da medida da soma dos quadrados dos desvios das observações em relação à média dos grupos. Primeiro são calculadas as médias das variáveis de cada grupo; em seguida, é calculado o quadrado da distância Euclideana entre essas médias e os valores das variáveis para cada indivíduo. Por fim, somam-se as distâncias para todos os indivíduos e optimiza-se a variância mínima dentro dos grupos.



Referências bibliográficas

Aldenferfer MS, Blashfield RK. *Cluster analysis*. Sage university paper series on quantitative applications in the social sciences, 07-044. Beverly Hills: Sage, 1984.

Alt M. *Exploring hyperspace . A non-mathematical explanation of multivariate analysis*. London: McGraw-Hill, 1990.

Bryman A, Cramer D. An*álise de dados em ciências sociais: introdução às técnicas utilizando o SPSS*. Oeiras: Celta Editora, 1992.

Dunteman GH. *Principal componentes analysis*. Sage university paper series on quantitative applications in the social sciences, 07-069. Beverly Hills: Sage, 1989.

Hair JF, Anderson RE, Tatham RL, Black WC. *Multivariate data analysis*. Englewood Clifs: Prentice-Hall, 1995.

Jobson JD. *Applied multivariate analysis. Volume II: Categorical and multivariate methods*. New York: Springer-Verlag, 1992.

Johnson RA, Wichern D. *applied multivariate analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

Kim J-O, Mueller C. *Introduction to factor analysis*. Sage university paper series on quantitative applications in the social sciences, 07-013. Beverly Hills: Sage, 1978.

Kim J-O, Mueller C. *Factor analysis. Statistical methods and practical issues*. Sage university paper series on quantitative applications in the social sciences, 07-014. Beverly Hills: Sage, 1978.