Apontamentos de
ESTATÍSTICA MULTIVARIADA

Prof. Jorge Cadima
Mestrado em Matemática Aplicada às Ciências Biológicas

Departamento de Matemática
Instituto Superior de Agronomia
Universidade Técnica de Lisboa

Fevereiro 2010

Conteúdo

Introdução

A disciplina de Estatística Multivariada do Mestrado de Matemática Aplicada às Ciências Biológicas tem por finalidade familiarizar os alunos com os principais métodos utilizados na análise de dados multivariados.

Os métodos *descritivos* de análise multivariada constituem a primeira parte da disciplina. Do ponto de vista *descritivo*, ou de *análise exploratória dos dados*, o objecto de estudo na estatística multivariada é, frequentemente, uma *matriz de dados*, com n linhas, correspondentes a n *indivíduos* ou unidades estatísticas, e p colunas, correspondentes a p *variáveis* ou características observadas sobre cada um desses n individuos. Alternativamente, o ponto de partida poderá ser dado por uma *matriz* $n \times n$ *de semelhanças/dissemelhanças* entre n individuos, que pode, ou não, ser construída a partir duma matriz de dados do tipo acabado de referir. O estudo desses dados pode ser feito de várias maneiras, frequentemente recorrendo ao conceito de *espaço linear*. Nesse caso, o estudo desses dados será, em grande medida, uma aplicação da Álgebra Linear e Teoria das Matrizes.

Caso os dados disponíveis constituam uma amostra extraída aleatoriamente duma população (multivariada), será de todo o interesse proceder também a *inferência* estatística, procurando extrair conclusões relativas à população, a partir da amostra. Uma introdução aos métodos *inferenciais* em estatística multivariada constitui o objectivo da segunda parte da disciplina. Como seria de esperar, esta introdução exige alguns conceitos fundamentais em Distribuições de Probabilidades Multivariadas.

Estes apontamentos dizem respeito às técnicas *Descritivas* de Estatística Multivariada. Admitem que já são conhecidos resultados fundamentais da Álgebra Linear e Teoria de Matrizes, leccionados nas disciplinas de Complementos de Álgebra e Análise e Modelação Estatística deste Mestrado. A consulta dos respectivos apontamentos será indispensável para um aprofundamento desses resultados preliminares.

Na disciplina de Estatística Multivariada é utilizado o **programa informático R**. Trata-se de um programa baseado na linguagem computacional S, especialmente concebida para aplicações estatísticas, e exposta nos livros:

- Becker, R.A.; Chambers, J.M. & Wilks, A.R. (1988) *The S Language.* Wadsworth & Brooks/Cole
- Chambers, J.M. & Hastie, T. (1992) *Statistical Models in S.* Wadsworth & Brooks/Cole

A linguagem S conhece duas concretizações na forma de programas informáticos: uma comercial, e outra o programa (gratuito e de código público) R. Os dois programas diferem em vários aspectos de funcionalidade, compatibilidades, etc. Mas trata-se, no fundamental, de dois "dialectos" da linguagem S. John Chambers, co-autor dos dois livros acima referidos, integra o núcleo central de desenvolvimento do programa R.

O programa R pode ser descarregado gratuitamente através da Internet, a partir do endereço:

http://cran.r-project.org

ou em vários outros *sites* que reproduzem o mesmo conteúdo (*mirror sites*, cujos endereços estão indicados no *site* acima referido). Existem versões do programa R já compiladas para execução nos principais sistemas operativos (Linux, Macintosh, Windows).

Informação vária sobre o programa (Manuais, respostas a perguntas frequentes, páginas de Ajuda, Boletim informativo) podem ser também obtidos através da rede, a partir do endereço acima, ou em:

http://www.r-project.org

Capítulo 1

Noções de Álgebra Linear e Teoria de Matrizes

No início deste capítulo procede-se a uma revisão de conceitos fundamentais em Álgebra Linear e Teoria de Matrizes que serão necessários na discussão das técnicas descritivas de estatística multivariada. Muitos destes conceitos foram já estudados nas disciplinas de Complementos de Álgebra e Análise, no contexto do estudo dos espaços lineares mais comuns: os espaços euclideanos $\mathbb{R}^n$. A discussão aqui é feita de forma mais geral e são introduzidas algumas noções novas. Alguns conceitos aqui referidos são discutidos em mais pormenor nos apontamentos da disciplina de Modelação Estatística I. São ainda consideradas umas primeiras ferramentas elementares no estudo de dados multivariados. Finalmente, considera-se um resultado de importância capital em Teoria de Matrizes e nos métodos de Estatística Multivariada: a Decomposição em Valores Singulares.

1.1 Revisão de conceitos essenciais de matrizes

Uma matriz é uma colecção de números organizados em forma rectangular, $\mathbf{A}_{nxp} \equiv [a_{ij}]$, onde $i = 1, 2, ..., n$ representam as linhas e $j = 1, 2, ..., p$ representam as colunas.

Se $n = p$ a matriz diz-se **quadrada**.

A matriz **transposta** de $\mathbf{A}$, representada por $\mathbf{A}^t$ é uma matriz de tipo pxn, obtida escrevendo cada linha de $\mathbf{A}$ como a coluna correspondente de $\mathbf{A}^t$ (ou, análogamente, cada coluna de $\mathbf{A}$ como a linha correspondente de $\mathbf{A}^t$).

1.1.1 Operações sobre matrizes

- **SOMA:** $\mathbf{C}_{nxp} = \mathbf{A}_{nxp} + \mathbf{B}_{nxp} \Longleftrightarrow c_{ij} = a_{ij} + b_{ij}, \qquad \forall i \in \{1, 2, ..., n\}, j \in \{1, 2, ..., p\}.$

NOTA: As matrizes têm de ser do mesmo tipo, para que possam ser somadas.

- **MULTIPLICAÇÃO ESCALAR:** $\alpha\mathbf{A} \equiv [\alpha\ a_{ij}] \qquad \forall i,j.$
- **PRODUTO DE HADAMARD:** Não sendo este o conceito de produto de matrizes, é um conceito por vezes útil. Designa-se produto de Hadamard de duas matrizes do mesmo tipo, $\mathbf{A}_{n\times p}$, $\mathbf{B}_{n\times p}$, à matriz $\mathbf{C}_{n\times p}$ cujo elemento genérico é dado pelo produto dos correspondentes elementos de $\mathbf{A}$ e $\mathbf{B}$: $\mathbf{C}_{n\times p} = \mathbf{A}_{n\times p} \circ \mathbf{B}_{n\times p} \iff c_{ij} = a_{ij} \cdot b_{ij}, \qquad \forall i \in \{1,2,...,n\}, j \in \{1,2,...,p\}$.
- **PRODUTO MATRICIAL:** $\mathbf{C}_{nxk} = \mathbf{A}_{nxp}\mathbf{B}_{pxk} \iff c_{ij} =< \mathbf{a}_i^{\text{linha}}, \mathbf{b}_j^{\text{coluna}} >, \quad \forall i,j$, onde $\mathbf{a}_i^{\text{linha}}$ representa a i-ésima linha da matriz $\mathbf{A}$ e $\mathbf{b}_j^{\text{coluna}}$ representa a j-ésima coluna da matriz $\mathbf{B}$.

NOTAS:

1. A multiplicação de matrizes só é possível se as matrizes forem **compatíveis**, isto é, se o número de colunas de $\mathbf{A}$ fôr igual ao número de linhas de $\mathbf{B}$.
2. O produto matricial *não* é comutativo, *i.e.*, em geral $\mathbf{AB} \neq \mathbf{BA}$.
3. *Se a matriz* $\mathbf{B} = \mathbf{b}$ *é um vector coluna, então o produto* $\mathbf{Ab}$ *é uma combinação linear das colunas da matriz* $\mathbf{A}$*, em que os coeficientes da combinação linear são os elementos do vector* $\mathbf{b}$, *i.e.*, $\mathbf{Ab} = \sum_{i=1}^{p} b_i \mathbf{a}_i^{\text{coluna}}$.
4. Analogamente, se a matriz $\mathbf{A} = \mathbf{a}^t$ fôr um vector linha, então o produto $\mathbf{a}^t\mathbf{B}$ é uma combinação linear das linhas da matriz $\mathbf{B}$, em que que os coeficientes da combinação linear são os elementos do vector $\mathbf{a}$, *i.e.*, $\mathbf{a}^t\mathbf{B} = \sum_{i=1}^{n} a_i \mathbf{b}_i^{\text{linha}}$.
5. Sendo $\mathbf{B}$ uma matriz $n \times p$, $\mathbf{a} \in \mathbb{R}^n$ e $\mathbf{c} \in \mathbb{R}^p$, tem-se $\mathbf{a}^t\mathbf{Bc} = \sum_{i=1}^{n}\sum_{j=1}^{p} a_i b_{ij} c_j$, onde a_i e c_j são os elementos das posições i e j, respectivamente, dos vectores $\mathbf{a}$ e $\mathbf{c}$, e b_{ij} é o elemento da linha i e coluna j da matriz $\mathbf{B}$. (Verifique!)

1.1.2 Matrizes quadradas e traços de matrizes

$\mathbf{A}$ Matriz **Diagonal**	se $a_{ij} = 0$ quando $i \neq j$
$\mathbf{A}$ Matriz **Simétrica**	se $\mathbf{A}^t = \mathbf{A} \iff a_{ij} = a_{ji}, \forall i,j$
$\mathbf{I}_p$ Matriz **Identidade**	se é matriz diagonal com elementos diagonais todos iguais a 1 *i.e.*, $\mathbf{A} = \mathbf{I}_p \iff a_{ij} = \begin{cases} 0 & \text{para } i \neq j \\ 1 & \text{para } i = j \end{cases}$
$\mathbf{A}^{-1}$ Matriz **Inversa** de $\mathbf{A}$	se $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_p$ (nem sempre existe, mas quando existe é única)
$\mathbf{A}$ Matriz **Ortogonal**	se $\mathbf{A}^{-1} = \mathbf{A}^t \iff \mathbf{A}^t\mathbf{A} = \mathbf{A}\mathbf{A}^t = \mathbf{I}_p$
$\mathbf{A}$ Matriz **Idempotente**	se $\mathbf{A}^2 = \mathbf{AA} = \mathbf{A}$

Se $\mathbf{A}_{p\times p}$ é uma matriz *simétrica*, diz-se que $\mathbf{x}^t\mathbf{Ax}$ é uma **forma quadrática**, e tem-se:

A Matriz **Definida Positiva**	se $\forall \mathbf{x} \in \mathbb{R}^p - \{\mathbf{0}\}$, $\mathbf{x}^t\mathbf{A}\mathbf{x} > 0$
A Matriz **Semi-Definida Positiva**	se $\forall \mathbf{x} \in \mathbb{R}^p - \{\mathbf{0}\}$, $\mathbf{x}^t\mathbf{A}\mathbf{x} \geq 0$
A Matriz **Definida Negativa**	se $\forall \mathbf{x} \in \mathbb{R}^p - \{\mathbf{0}\}$, $\mathbf{x}^t\mathbf{A}\mathbf{x} < 0$
A Matriz **Semi-Definida Negativa**	se $\forall \mathbf{x} \in \mathbb{R}^p - \{\mathbf{0}\}$, $\mathbf{x}^t\mathbf{A}\mathbf{x} \leq 0$
A Matriz **Indefinida**	se $\mathbf{x}^t\mathbf{A}\mathbf{x}$ pode ter qualquer sinal.

Alguns factos adicionais, relacionados com matrizes, importantes para a matéria que se segue:

1. Se **D** é uma matriz *diagonal* e **A** é uma matriz compatível na multiplicação, então:
   (a) A matriz **AD** é a matriz que resulta de multiplicar cada *coluna* de **A** pelo correspondente elemento diagonal de **D**.
   (b) A matriz **DA** é a matriz que resulta de multiplicar cada *linha* de **A** pelo correspondente elemento diagonal de **D**.
2. O **traço** duma matriz quadrada é a soma dos seus elementos diagonais:
$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{p} a_{ii}.$$
3. O traço do produto matricial **AB**, em que $\mathbf{A} \in \mathbb{M}_{n\times p}$ e $\mathbf{B} \in \mathbb{M}_{p\times n}$, é dado por:
$$\text{tr}(\mathbf{AB}) = \sum_{\text{i=1}}^{\text{n}} < \mathbf{a}_{\text{i}}^{\text{linha}}, \mathbf{b}_{\text{i}}^{\text{coluna}} > = \sum_{\text{i=1}}^{\text{n}} \sum_{\text{j=1}}^{\text{p}} \text{a}_{\text{ij}}\text{b}_{\text{ji}},$$
   onde a_{ij} indica o elemento da linha i, coluna j da matriz **A**, e b_{ji} o elemento da linha j, coluna i da matriz **B**. (Verifique!).
4. **Circularidade do traço**.

Produtos de duas matrizes. Sejam $\mathbf{A} \in \mathbb{M}_{m\times k}$ e $\mathbf{B} \in \mathbb{M}_{k\times m}$. Então
$$\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA}) \ .$$

Demonstração. Tem-se $\text{tr}(\mathbf{AB}) = \sum_{\text{i=1}}^{\text{m}} < \mathbf{a}_{\text{i}}^{\text{linha}}, \mathbf{b}_{\text{i}}^{\text{coluna}} > = \sum_{\text{i=1}}^{\text{m}} (\sum_{\text{j=1}}^{\text{k}} \text{a}_{\text{ij}}\text{b}_{\text{ji}}) = \sum_{j=1}^{k} (\sum_{i=1}^{m} b_{ji}a_{ij}) = \sum_{j=1}^{k} < \mathbf{b}_{j}^{\text{linha}}, \mathbf{a}_{j}^{\text{coluna}} > = \text{tr}(\mathbf{BA})$. ∇

Observação: Este resultado é válido mesmo quando os produtos matriciais **AB** e **BA** sejam diferentes (como acontece em geral, visto o produto matricial não ser comutativo). Apenas é necessário exigir que seja possível construir os dois produtos (isto é, que as dimensões das matrizes **A** e **B** sejam compatíveis, quer para o produto matricial **AB**, quer para o produto matricial **BA**, para que se possa garantir que *os traços* das matrizes resultantes dos dois produtos sejam iguais.

Produtos de três matrizes. Sejam $\mathbf{A} \in \mathbb{M}_{m\times k}$, $\mathbf{B} \in \mathbb{M}_{k\times p}$ e $\mathbf{C} \in \mathbb{M}_{p\times m}$. Então

$$\mathrm{tr}(\mathbf{ABC}) = \mathrm{tr}(\mathbf{BCA}) \ .$$

Demonstração. Imediata a partir do produto de duas matrizes, tomando-se os produtos $\mathbf{AD}$ e $\mathbf{DA}$ com $\mathbf{D} = \mathbf{BC}$. ∇

Produtos de n matrizes. A circularidade do traço generaliza-se para produtos de qualquer número n de matrizes, de forma imediata. Sejam $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, ..., \mathbf{A}_n$ matrizes de dimensões $p_0 \times p_1$, $p_1 \times p_2$, $p_2 \times p_3$, ..., $p_{n-1} \times p_0$, respectivamente. Então,

$$\mathrm{tr}(\mathbf{A}_1\mathbf{A}_2\mathbf{A}_3\cdots\mathbf{A}_n) = \mathrm{tr}(\mathbf{A}_2\mathbf{A}_3\cdots\mathbf{A}_n\mathbf{A}_1) \ .$$

Observação: Atenção à exigência de que todos os produtos matricias indicados sejam possíveis, isto é, que as matrizes sejam compatíveis para os produtos indicados.

5. O traço é um **operador linear**, *i.e.*, $\mathrm{tr}(\alpha\mathbf{A} + \beta\mathbf{B}) = \alpha\mathrm{tr}(\mathbf{A}) + \beta\mathrm{tr}(\mathbf{B})$. (Verifique!).

6. O **determinante** duma matriz quadrada é a soma de $p!$ produtos de elementos da matriz, produtos da forma $a_{1,j_1}a_{2,j_2}a_{3,j_3}\cdots a_{p,j_p}$ onde os índices $(j_1, j_2, \ldots, j_p)$ correspondem a todas as $p!$ permutações dos inteiros de 1 a p, e onde metade das parcelas são multiplicadas por -1, segundo uma regra que não é relevante para os nossos propósitos[1]. Adiante veremos uma forma mais simples de calcular os determinantes das matrizes com que iremos trabalhar. O determinante de $\mathbf{A}$ costuma representar-se por $|\mathbf{A}|$.

1.2 Conceitos fundamentais de Álgebra Linear

1.2.1 Espaço linear, independência linear, base

Comecemos por relembrar a definição de *espaço linear*.

Definição 1.1 *Seja* $\mathbf{L}$ *um conjunto no qual se definem duas operações (fechadas em* $\mathbf{L}$*):*

(i) *Uma operação binária designada* soma *vectorial*

$$\mathbf{x}, \mathbf{y} \in \mathbf{L} \rightarrow \mathbf{x} + \mathbf{y} \in \mathbf{L}$$

[1] Vejam-se os apontamentos da disciplina de Complementos de Álgebra e Análise deste Mestrado, ou o livro *Horn, R. & Johnson, C.*, **Matrix Analysis**,Cambridge University Press, 1985, para mais pormenores

(ii) *Uma operação designada* multiplicação escalar *(real)*

$$\mathbf{x} \in \mathbf{L}, \alpha \in \mathbb{R} \rightarrow \alpha\mathbf{x} \in \mathbf{L}$$

O conjunto $\mathbf{L}$*, com estas duas operações designa-se um* ***espaço linear*** *(ou* ***vectorial****) se se verificarem as seguintes propriedades:*

(S) *A operação soma (vectorial) em* $\mathbf{L}$*:*

- **(S1)** *é* comutativa*, isto é,* $\mathbf{x}+\mathbf{y} = \mathbf{y}+\mathbf{x}$, $\forall \mathbf{x}, \mathbf{y} \in \mathbf{L}$.
- **(S2)** *é* associativa*, isto é,* $\mathbf{x}+(\mathbf{y}+\mathbf{z}) = (\mathbf{x}+\mathbf{y})+\mathbf{z}$, $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbf{L}$.
- **(S3)** *tem* elemento nulo*, isto é,* $\exists \mathbf{0} \in \mathbf{L}$ *tal que* $\mathbf{0}+\mathbf{x} = \mathbf{x}$, $\forall \mathbf{x} \in \mathbf{L}$.
- **(S4)** *admite* elementos inversos*, isto é,* $\forall \mathbf{x} \in \mathbf{L}$, $\exists -\mathbf{x} \in \mathbf{L}$ *tal que* $\mathbf{x}+(-\mathbf{x}) = \mathbf{0}$.

(ME) *A operação multiplicação escalar em* $\mathbf{L}$*:*

- **(ME1)** *é* quase-associativa*, isto é,* $\alpha(\beta\mathbf{x}) = (\alpha\beta)\mathbf{x}$, $\forall \alpha, \beta \in \mathbb{R}$, $\forall \mathbf{x} \in \mathbf{L}$.
- **(ME2)** *tem o número real 1 como* elemento identidade*, isto é,* $1\mathbf{x} = \mathbf{x}$, $\forall \mathbf{x} \in \mathbf{L}$.

E ainda:

- **(ME3)** A multiplicação escalar é distributiva em relação à soma vectorial*, isto é,*

$$\alpha(\mathbf{x}+\mathbf{y}) = \alpha\mathbf{x}+\alpha\mathbf{y}, \quad \forall \alpha \in \mathbb{R}, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbf{L}$$

- **(ME4)** A multiplicação escalar é distributiva em relação à soma de números reais*, isto é,*

$$(\alpha+\beta)\mathbf{x} = \alpha\mathbf{x}+\beta\mathbf{x}, \quad \forall \alpha, \beta \in \mathbb{R}, \quad \forall \mathbf{x} \in \mathbf{L}$$

Observações:

1. Os elementos de um espaço linear são designados **vectores**.
2. O inverso aditivo de um vector $\mathbf{x} \in \mathbf{L}$ resulta da sua multiplicação escalar pelo número real -1: $-\mathbf{x} = (-1)\mathbf{x}$, $\forall \mathbf{x} \in \mathbf{L}$.
3. A operação da subtracção está implicitamente definida em qualquer espaço linear: $\mathbf{x}-\mathbf{y} = \mathbf{x}+(-\mathbf{y})$, $\forall \mathbf{x}, \mathbf{y} \in \mathbf{L}$.
4. O elemento nulo da operação soma num espaço linear é *único*.
5. Cada vector de um espaço linear tem um inverso aditivo *único*.
6. A multiplicação escalar de qualquer vector $\mathbf{x} \in \mathbf{L}$ pelo número real zero resulta no elemento nulo da soma vectorial: $0\mathbf{x} = \mathbf{0}$, $\forall \mathbf{x} \in \mathbf{L}$.

Exemplos de espaços lineares:

1. $\mathbb{R}^n$ ($\forall n \in \mathbb{N}$), com as habituais operações.
2. $\mathbb{M}_{n\times p}$, o espaço de todas as matrizes reais de tipo $n\times p$, com a habitual operação de soma de matrizes e de produto de uma matriz por um número real.
3. $\mathbb{S}_p$, o espaço de todas as matrizes *simétricas* de tipo $p\times p$.
4. O conjunto de todos os polinómios de grau $\leq n$ (incluíndo o polinómio 0), com as habituais operações.
5. O conjunto das funções reais contínuas no intervalo [a,b], com as habituais operações: $h = f + g$ se $h(x) = f(x) + g(x)$, $\forall x \in$[a,b], e $h = \alpha f$, se $h(x) = \alpha f(x)$, $\forall x \in$[a,b].

Conjuntos com operações associadas que *não* são espaços lineares:

1. $\mathbb{R}_0^+$ com as habituais operações (pois, por exemplo, o conjunto não admite elementos inversos para a soma).
2. $\mathbb{Z}$ com as habituais operações (pois, por exemplo, o conjunto não é fechado para a multiplicação escalar).

Relembremos agora a definição de conceitos importantes associados ao conceito de espaços lineares.

Definição 1.2 *Seja* $\mathbf{L}$ *um espaço linear.*

1. *Sejam* $\mathbf{x}, \mathbf{y} \in \mathbf{L}$ *e* $\alpha, \beta \in \mathbb{R}$. *O vector* $\alpha\mathbf{x} + \beta\mathbf{y} \in \mathbf{L}$ *diz-se uma* ***combinação linear*** *dos vectores* $\mathbf{x}$ *e* $\mathbf{y}$.
2. *Um subconjunto* $\mathbf{M} \subseteq \mathbf{L}$ *diz-se um* ***conjunto gerador*** *de* $\mathbf{L}$ *se qualquer vector* $\mathbf{x} \in \mathbf{L}$ *se pode escrever como combinação linear de elementos de* $\mathbf{M}$.
3. *Um conjunto* $\{\mathbf{x}_i\}_{i=1}^n$ *de vectores de* $\mathbf{L}$ *diz-se* ***linearmente independente*** *se* $\sum_{i=1}^n \alpha_i \mathbf{x}_i = \mathbf{0} \Rightarrow \alpha_i = 0$, $\forall i = 1, ..., n$.
4. *Um conjunto linearmente independente e gerador de um espaço linear L diz-se uma* ***base*** *de L.*

Observações:

1. Quando um conjunto de vectores *não* é linearmente independente, diz-se *linearmente dependente* e, nesse caso, pelo menos um dos vectores do conjunto se pode escrever como combinação linear dos restantes.
2. Sejam M e N conjuntos de vectores no espaço linear L, tais que M⊂N. Então:

(a) M linearmente dependente ⇒ N linearmente dependente.

(b) N linearmente independente ⇒ M linearmente independente.

3. Os espaços lineares que possuam uma base com um número finito de vectores são particularmente "bem comportados". Todos os espaços lineares que nos interessam (no contexto descritivo em que nos situamos) estão neste caso.

Daqui em diante, quando se falar em espaços lineares admite-se sempre implicitamente que possuem uma base com um número finito de vectores.

Teorema 1.1 *Qualquer base de um espaço linear L tem o mesmo número de elementos.*

Definição 1.3 *O número de elementos de qualquer base de um espaço linear L designa-se a* ***dimensão*** *do espaço L e representa-se por dim(L).*

Teorema 1.2 *Seja L um espaço linear n-dimensional e* $\{\mathbf{x}_i\}_{i=1}^n$ *uma sua base. Então, qualquer vector* $\mathbf{x} \in L$ *se pode escrever de forma* **única** *como combinação linear dos vectores da base* $\{\mathbf{x}_i\}_{i=1}^n$.

Exemplos:

1. $\mathbb{R}^2$ é um espaço de dimensão 2. Uma base de $\mathbb{R}^2$ é constituída pelos vectores $\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ e $\mathbf{x}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Qualquer vector $\begin{bmatrix} a \\ b \end{bmatrix}$ se pode escrever como $(b-a)\mathbf{x}_1 + (2a-b)\mathbf{x}_2$.

2. $\mathbb{R}^n$ é um espaço n-dimensional. A base de $\mathbb{R}^n$ constituída pelos vectores $\{\mathbf{e}_i\}_{i=1}^n$, onde $\mathbf{e}_i$ é um vector com 1 na i-ésima posição e os restantes elementos iguais a zero, designa-se a **base canónica de** $\mathbb{R}^n$.

3. $\mathbb{M}_{n\times p}$ é um espaço np-dimensional. A *base canónica* deste espaço é constituída pelas matrizes $\mathbf{E}_{ij}$ (i=1,...,n ; j=1,...,p), que têm um 1 na i-ésima linha, j-ésima coluna, e zero nas restantes posições.

4. $\mathbb{S}_p$ é um espaço $p(p+1)/2$-dimensional. Uma base do espaço 6-dimensional $\mathbb{S}_3$ é dada por:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

5. O espaço linear dos polinómios de grau $\leq n$ é de dimensão $n+1$. Uma base deste espaço é constituída pelos polinómios $\{1, x, x^2, x^3, ..., x^n\}$.

Notas:

1. O espaço linear das funções contínuas em [a,b] é de dimensão infinita.

2. Num espaço linear de dimensão n, nenhum conjunto de menos de n vectores pode gerar o espaço e nenhum conjunto de mais de n vectores pode ser linearmente independente.

Definição 1.4 *Um subconjunto não vazio M de um espaço linear L diz-se um* ***subespaço linear*** *se, com as duas operações definidas no espaço linear L, satisfaz as propriedades indicadas na Definição 1.1 (pg. 6).*

Teorema 1.3 *Um subconjunto não vazio M dum espaço linear L é um* subespaço linear *se M fôr fechado para qualquer combinação linear dos seus elementos, i.e., se:*

$$\alpha\mathbf{x} + \beta\mathbf{y} \in \mathrm{M} \quad , \quad \forall \mathbf{x}, \mathbf{y} \in \mathrm{M}, \quad \alpha, \beta \in \mathbb{R}$$

Exercício 1.1 *Demonstre o Teorema anterior.*

Nota: Qualquer subespaço linear é ele próprio um espaço linear.

Exemplos:

1. $\mathbb{R}$ é um subespaço linear de $\mathbb{R}^2$.

2. $\mathbb{S}_p$ é um subespaço linear de $\mathbb{M}_{p\times p}$.

3. Para qualquer espaço linear L cujo elemento nulo da soma é $\mathbf{0}$, $\{\mathbf{0}\}$ é um subespaço linear de L.

4. Seja M um conjunto de elementos de L. O conjunto de todas as combinações lineares de elementos de M é um subespaço linear de L, designado o **subespaço gerado** pelo conjunto M.

Teorema 1.4 *Seja L um espaço linear e M, N dois seus subespaços lineares. Então $M \cap N$ também é um subespaço linear de L.*

Exercício 1.2 *Demonstre o Teorema anterior.*

Nota: M∪N *não* é, em geral, um subespaço linear.

Exercício 1.3 *Construa um exemplo em que M e N sejam subespaços, mas M∪N não seja um subespaço.*

1.2.2 Transformações Lineares

Relembremos ainda o conceito e algumas propriedades das transformações (aplicações) lineares.

Definição 1.5 *Sejam L,M espaços lineares. Uma* ***transformação (aplicação) linear*** $\mathbf{A}$ *de L em M é uma aplicação que associa a um vector* $\mathbf{x} \in L$*, outro vector* $\mathbf{A}(\mathbf{x}) \in M$*, tal que:*

$$\mathbf{A}(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha\mathbf{A}(\mathbf{x}) + \beta\mathbf{A}(\mathbf{y}) \quad , \quad \forall \mathbf{x}, \mathbf{y} \in \mathrm{L} \ , \ \forall \alpha, \beta \in \mathbb{R}$$

Observações:

1. É habitual escrever-se $\mathbf{Ax}$ em vez de $\mathbf{A}(\mathbf{x})$. Se M=L fala-se apenas numa aplicação linear em L.
2. Se L=$\mathbb{R}^p$ e M=$\mathbb{R}^n$, então as transformações lineares correspondem a matrizes de tipo $n \times p$.
3. Necessariamente, se $\mathbf{A}$ é uma transformação linear, a imagem do elemento nulo de L será o elemento nulo de M. De facto, $\mathbf{0}_L = \mathbf{x} - \mathbf{x} = \mathbf{x} + (-1)\mathbf{x}$ para qualquer elemento $\mathbf{x} \in$ L. Ora, pela definição de aplicação (transformação) linear, tem-se $\mathbf{A0}_L = \mathbf{A}(\mathbf{x} + (-1)\mathbf{x}) = \mathbf{Ax} + (-1)\mathbf{Ax} = \mathbf{Ax} - \mathbf{Ax} = \mathbf{0}_M$.

Definição 1.6 *Sejam L e M espaços lineares e* $\mathbf{A}$ *uma transformação linear de L em M. Considerem-se dois subconjuntos, definidos pela transformação linear* $\mathbf{A} : L \longrightarrow M$*:*

1. *O* ***conjunto imagem*** *de* $\mathbf{A}$*, representado por* $\mathcal{C}(\mathbf{A})$*, é o conjunto de elementos* ***de M*** *que são imagens da transformação* $\mathbf{A}$*, isto é, é o conjunto de elementos* $\mathbf{y} \in M$ *que se podem escrever na forma* $\mathbf{y} = \mathbf{Ax}$*, para algum elemento* $\mathbf{x} \in L$*.*
2. *O* ***núcleo*** *de* $\mathbf{A}$*, representado por* $\mathcal{N}(\mathbf{A})$*, é o conjunto de elementos* ***de L*** *cuja imagem pela aplicação* $\mathbf{A}$ *é o elemento nulo de L, isto é, é o conjunto dos vectores* $\mathbf{x} \in L$ *tais que* $\mathbf{Ax} = \mathbf{0} \in M$*.*

Teorema 1.5 *Sejam L e M espaços lineares e* $\mathbf{A}$ *uma transformação linear de L em M. Então, o núcleo de* $\mathbf{A}$*,* $\mathcal{N}(\mathbf{A})$ *é um* ***subespaço*** *de L, e o conjunto imagem,* $\mathcal{C}(\mathbf{A})$*, é um* ***subespaço*** *de M.*

Exercício 1.4 *Demonstre este Teorema.*

Definição 1.7 *Sejam L e M espaços lineares e* $\mathbf{A}$ *uma transformação linear de L em M. A dimensão do subespaço imagem* $\mathcal{C}(\mathbf{A})$ *diz-se a* ***característica*** *da transformação* $\mathbf{A}$ *e representa-se por* $car(\mathbf{A})$*. Assim,* $car(\mathbf{A}) = dim\,(\mathcal{C}(\mathbf{A}))$*.*

Generalizemos agora um resultado já estudado no contexto das transformações lineares entre espaços euclidianos, ou seja, no contexto de matrizes, e que relaciona a característica duma transformação linear com as dimensões do seu núcleo e do subespaço de partida.

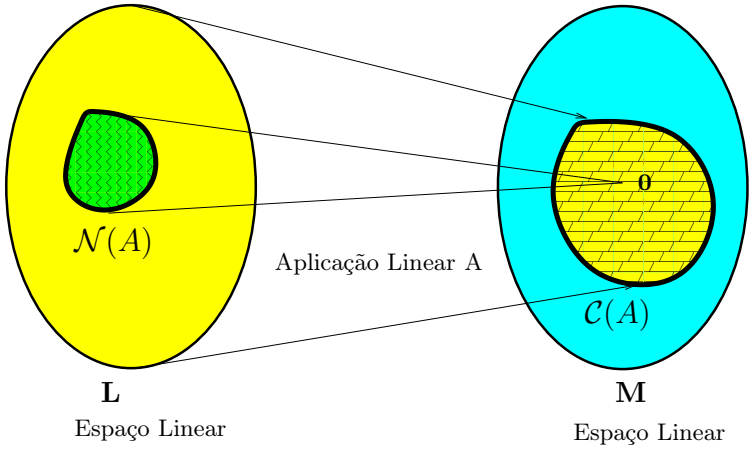


Figura 1.1: Os conjuntos Núcleo e Imagem, definidos por uma aplicação linear entre dois espaços lineares.

Teorema 1.6 *Seja* $\mathbf{A}$ *uma transformação linear entre os espaços lineares* L *e* M*. Então*

$$dim(L) \;=\; dim\left(\mathcal{N}(\mathbf{A})\right) \;+\; dim\left(\mathcal{C}(\mathbf{A})\right) \;. \tag{1.1}$$

Encerramos esta revisão com um resultado interessante: as transformações lineares entre espaços lineares formam, elas próprias, um espaço linear.

Teorema 1.7 *O conjunto* $\mathcal{T}(L, M)$ *das transformações lineares de* L *em* M *constitui um espaço linear com as operações* $(\mathbf{A}+\mathbf{B})\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{x}$ *e* $(\alpha\mathbf{A})\mathbf{x} = \alpha(\mathbf{A}\mathbf{x})$*.*

Exercício 1.5 *Demonstre este Teorema.*

Observação. Em particular, tem-se uma *transformação linear nula*, $\mathbf{0}$, que é elemento nulo para a operação soma em $\mathcal{T}(L, M)$, isto é, tal que para qualquer outra aplicação linear $\mathbf{A}$ se verifica $\mathbf{A}+\mathbf{0} = \mathbf{A}$. Essa transformação linear nula sobre L caracteriza-se pelo facto de $\mathbf{0x} = \mathbf{o}$, $\forall \mathbf{x} \in$L, onde $\mathbf{o}$ designa o elemento nulo do espaço linear M. É também consequência deste Teorema que a uma dada transformação linear corresponde sempre uma outra transformação linear que é o seu *inverso aditivo*. Ou seja, dada uma transformação linear de L em M, $\mathbf{A}$, existe sempre outra transformação linear de L em M, $-\mathbf{A}$, tal que $\mathbf{A} + (-\mathbf{A}) = \mathbf{0}$.

1.2.3 Produtos internos, Normas, Distâncias, Ângulos

Definição 1.8 *Um* ***produto interno*** *num espaço linear* L *é uma aplicação*

$$< \cdot, \cdot >: \mathrm{L} \times \mathrm{L} \longrightarrow \mathbb{R}$$

com as seguintes propriedades:

1. $< \mathbf{x}, \mathbf{y} >=< \mathbf{y}, \mathbf{x} > \quad \forall \mathbf{x}, \mathbf{y} \in \mathrm{L}$ [Simetria]
2. $< \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2, \mathbf{y} >= \alpha_1 < \mathbf{x}_1, \mathbf{y} > + \alpha_2 < \mathbf{x}_2, \mathbf{y} >, \quad \forall \mathbf{x}_1, \mathbf{x}_2, \mathbf{y} \in \mathrm{L}, \ \forall \alpha_1, \alpha_2 \in \mathbb{R}$ [Bilinearidade]
3. $< \mathbf{x}, \mathbf{x} > \geq 0, \quad \forall \mathbf{x} \in \mathrm{L}$, com a igualdade se e só se $\mathbf{x} = \mathbf{0}$ [Definida positiva]

Nota. Na disciplina de Complementos de Álgebra e Análise foi utilizada uma notação diferente para o produto interno entre vectores de $\mathcal{R}^n$: $\mathbf{x}|\mathbf{y}$ ou $\mathbf{x} \cdot \mathbf{y}$, em vez de $< \mathbf{x}, \mathbf{y} >$.

Nos espaços $\mathbb{R}^n$, **o produto interno mais frequente e mais habitual**, frequentemente designado **produto interno euclideano**, foi já introduzido na disciplina de Complementos de Álgebra e Análise:

$$< \mathbf{x}, \mathbf{y} >= \mathbf{x}^t \mathbf{y} = \sum_{i=1}^{n} x_i y_i \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \tag{1.2}$$

Este produto interno é um caso particular de produto interno. Outros produtos internos em $\mathbb{R}^n$ podem ser definidos com o auxílio de **matrizes definidas positivas**, conceito este que já foi introduzido na disciplina de Complementos de Álgebra e Análise. A noção de matriz definida positiva será retomada mais adiante nesta disciplina (ver página 5), mas será aqui recordado a fim de permitir discutir produtos internos em $\mathbb{R}^n$ alternativos.

Definição 1.9 *Uma matriz simétrica (logo quadrada)* $\mathbf{W} \in \mathbb{M}_{n \times n}$ *diz-se* ***definida positiva*** *se:*

$$\begin{aligned} \mathbf{x}^t \mathbf{W} \mathbf{x} \geq 0 \quad &, \quad \forall \mathbf{x} \in \mathbb{R}^n \\ \mathbf{x}^t \mathbf{W} \mathbf{x} = 0 \quad &\Longleftrightarrow \quad \mathbf{x} = \mathbf{0} \end{aligned}$$

Teorema 1.8 *Seja* $\mathbf{W} \in \mathbb{M}_{n \times n}$ *uma matriz definida positiva. A função* $< \cdot, \cdot >_{\mathbf{W}} : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$

$$< \mathbf{x}, \mathbf{y} >_{\mathbf{W}} = \mathbf{x}^t \mathbf{W} \mathbf{y} = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} x_i y_j \tag{1.3}$$

define um produto interno em $\mathbb{R}^n$.

Demonstração. A fim de provar que a função (1.3) define um produto interno, é preciso provar que verifica as condições da Definição 1.8. Ora, uma matriz definida positiva é, por definição, simétrica. Assim, $< \mathbf{y}, \mathbf{x} >_{\mathbf{W}} = \mathbf{y}^t \mathbf{W} \mathbf{x} = \mathbf{y}^t \mathbf{W}^t \mathbf{x} = (\mathbf{x}^t \mathbf{W} \mathbf{y})^t = \mathbf{x}^t \mathbf{W} \mathbf{y} =< \mathbf{x}, \mathbf{y} >_{\mathbf{W}}$. A penúltima passagem decorre da forma quadrática ser uma matriz 1×1. Logo, verifica-se a *simetria* da aplicação. A *bilinearidade* decorre directamente das propriedades dos produtos matriciais: $< \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2, \mathbf{y} >_{\mathbf{W}} = (\alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2)^t \mathbf{W} \mathbf{y} = \alpha_1 \mathbf{x}_1^t \mathbf{W} \mathbf{y} + \alpha_2 \mathbf{x}_2^t \mathbf{W} \mathbf{y} = \alpha_1 < \mathbf{x}_1, \mathbf{y} >_{\mathbf{W}} + \alpha_2 < \mathbf{x}_2, \mathbf{y} >_{\mathbf{W}}$. Finalmente, a terceira propriedade dos produtos internos decorre directamente do facto de $\mathbf{W}$ ser uma matriz definida positiva. ∇

Nota: O produto interno habitual corresponde a tomar $\mathbf{W} = \mathbf{I}$ na expressão 1.3.

Exemplo 1.1 *Num qualquer espaço linear de matrizes, $\mathbb{M}_{n\times p}$, o produto interno mais habitual é definido da seguinte forma:*

$$< \mathbf{A}, \mathbf{B} > \quad = \quad \mathrm{tr}(\mathbf{A}^t\mathbf{B}) \qquad , \qquad \forall \mathbf{A}, \mathbf{B} \in \mathbb{M}_{n\times p} \ ,$$

onde tr *indica o traço da matriz. É fácil de verificar que se trata dum produto interno bem definido. De facto, pela definição do produto matricial e de traço duma matriz, tem-se*

$$< \mathbf{A}, \mathbf{B} > \quad = \quad \mathrm{tr}(\mathbf{A}^t\mathbf{B}) \quad = \quad \sum_{i=1}^{n}\sum_{j=1}^{p} a_{ij}b_{ij} \ .$$

Esta última expressão corresponde a tomar o produto interno usual entre vectores, para dois vectores de $\mathbb{R}^{n\times p}$, definidos como tendo todos os elementos de cada matriz, "empilhados" coluna a coluna[2]*.*

Um resultado de grande importância associado a produtos internos é o seguinte Teorema.

Teorema 1.9 (Cauchy-Schwarz-Buniakovski) *Seja L um espaço linear com produto interno. Então:*

$$|< \mathbf{x},\mathbf{y} >| \quad \leq \quad \sqrt{< \mathbf{x},\mathbf{x} >} \cdot \sqrt{< \mathbf{y},\mathbf{y} >} \quad , \quad \forall \mathbf{x},\mathbf{y} \in \mathrm{L} \ , \tag{1.4}$$

tendo-se a igualdade se e só se um dos vectores fôr um múltiplo escalar do outro.

Demonstração. Para qualquer par de vectores $\mathbf{x}, \mathbf{y} \in L$, considere a combinação linear $\mathbf{x} - k\mathbf{y}$. Verifica-se, pela definição e propriedades do produto interno, que:

$$0 \quad \leq \quad < \mathbf{x} - k\mathbf{y} \ , \ \mathbf{x} - k\mathbf{y} > \quad = \quad < \mathbf{x}, \mathbf{x} > \ - 2k < \mathbf{x}, \mathbf{y} > \ + k^2 < \mathbf{y}, \mathbf{y} >$$

Mas a expressão final é um polinómio de segunda ordem em k, cujo gráfico é o de uma parábola com a concavidade voltada para cima. Uma vez que essa parabola só pode tomar valores não-negativos, ter-se-à uma de duas situações:

(i) o mínimo da parabola é positivo, pelo que as duas raízes do polinómio em k são complexas. Neste caso, o binómio discriminante na fórmula resolvente do polinómio em k (isto é, a parte "$b^2 - 4ac$"dessa fórmula resolvente) é negativa, isto é, no nosso caso:

$$4 < \mathbf{x}, \mathbf{y} >^2 \quad < \quad 4 < \mathbf{x}, \mathbf{x} > \cdot < \mathbf{y}, \mathbf{y} > \quad \Leftrightarrow \quad | < \mathbf{x}, \mathbf{y} > | \quad < \quad \sqrt{< \mathbf{x}, \mathbf{x} >} \cdot \sqrt{< \mathbf{y}, \mathbf{y} >} \ .$$

Este caso ocorre quando $< \mathbf{x} - k\mathbf{y}, \mathbf{x} - k\mathbf{y} > \quad > \quad 0$, isto é, quando $\mathbf{x} \neq k\mathbf{y}$;

(ii) o mínimo da parábola é zero, pelo que o polinómio tem uma dupla raíz real, e o binómio discriminante é zero. Neste caso, um raciocínio análogo ao caso anterior leva-nos à conclusão que

$$| < \mathbf{x}, \mathbf{y} > | \ = \ \sqrt{< \mathbf{x}, \mathbf{x} >} \cdot \sqrt{< \mathbf{y}, \mathbf{y} >} \ .$$

Este caso ocorre quando $< \mathbf{x} - k\mathbf{y}, \mathbf{x} - k\mathbf{y} >= 0$, isto é, quando $\mathbf{x} = k\mathbf{y}$. ∇

[2] *Esta operação que consiste em transformar uma matriz de tamanho $n \times p$ num vector de $\mathbb{R}^{n\times p}$, escrevendo os np elementos da matriz por ordem de colunas é por vezes designada a* vectorização *da matriz.*

A desigualdade de Cauchy-Schwarz-Buniakovski surge em numerosos contextos diferentes e é uma propriedade fundamental no estudo de produtos internos.

Normas

Vejamos agora o conceito de *norma* num espaço linear.

Definição 1.10 *Uma* ***norma*** *(comprimento) é uma função real* $\|\cdot\| : L \longrightarrow \mathbb{R}$*, que verifica as seguintes propriedades:*

1. $\|\mathbf{x}\| \geq 0, \qquad \forall \mathbf{x} \in L \quad e \quad \|\mathbf{x}\| = 0 \quad \Leftrightarrow \quad \mathbf{x} = \mathbf{0}$ [Positividade]
2. $\|c \cdot \mathbf{x}\| = |c| \cdot \|\mathbf{x}\|, \qquad \forall \mathbf{x} \in L, \qquad \forall c \in \mathbb{R}$ [Homogeneidade]
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|, \qquad \forall \mathbf{x}, \mathbf{y} \in L$ [Desigualdade Triangular]

Observações:

1. Um espaço linear com uma norma diz-se um **espaço normado**.
2. Um vector de norma 1 num espaço normado diz-se um **vector unitário**.

Nota: Em $\mathbb{R}^n$, a **norma habitual (euclideana)** é a norma definida por

$$\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{1.5}$$

Mas outras normas podem ser definidas nos espaços euclidianos $\mathbb{R}^n$, entre as quais se destacam a seguinte família de normas.

Teorema 1.10 *Seja* p *um número real tal que* $p \geq 1$*. A aplicação* $\|\cdot\|_p : \mathbb{R}^n \longrightarrow \mathbb{R}$ *definida como:*

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} \tag{1.6}$$

é uma norma em $\mathbb{R}^n$*, designada a* ***norma*** ℓ_p ***de Minkovski.***

Demonstração: O módulo na definição garante que todas as parcelas da soma são não-negativas, logo a primeira condição da Definição 1.10 verifica-se (não havendo parcelas negativas, todas têm de se anular para que a soma seja nula). A segunda condição é também imediata. Finalmente, a desigualdade triangular verifica-se neste caso (a famosa *desigualdade de Minkovski*), embora a demonstração não-trivial de tal facto seja aqui omitida. ∇

Nota: A habitual norma euclidiana é um caso particular desta família de normas, sendo a norma que resulta de tomar $p = 2$**, ou seja, é a norma** ℓ_2 **de Minkovski.**

Teorema 1.11 *A aplicação* $\|\cdot\|_\infty : \mathbb{R}^n \longrightarrow \mathbb{R}$ *definida como:*

$$\|\mathbf{x}\|_\infty = \max_i |x_i| \tag{1.7}$$

é uma norma em $\mathbb{R}^n$*, designada a* ***norma do máximo*** *ou* ***norma*** ℓ_∞*.*

Demonstração: É imediato que $\max_i |x_i| \geq 0$, e também que se o máximo desses valores absolutos é nulo, também todos os outros x_i terão de ser nulos, pelo que a positividade da aplicação $\|\cdot\|_\infty$ verifica-se. Por outro lado, para qualquer $i = 1, ..., n$, $|\alpha x_i| = |\alpha| \cdot |x_i|$. Logo, $\|\alpha\mathbf{x}\|_\infty = \max_i |\alpha| \cdot |x_i| = |\alpha| \cdot \max_i |x_i| = |\alpha| \cdot \|\mathbf{x}\|_\infty$, pelo que a aplicação também é homogénea. Finalmente, sabemos das propriedades dos módulos que, para quaisquer números reais x e y, $|x + y| \leq |x| + |y|$. Logo, $\|\mathbf{x} + \mathbf{y}\|_\infty = \max_i |x_i + y_i| \leq \max_i (|x_i| + |y_i|)$. Mas esta última expressão tem de ser menor ou igual que $\max_i |x_i| + \max_i |y_i| = \|\mathbf{x}\|_\infty + \|\mathbf{y}\|_\infty$, pelo que a desigualdade triangular também se verifica e a aplicação dada é uma norma. ∇

Nota: A norma do máximo[3] é designada também norma ℓ_∞ uma vez que é possível mostrar que surge como caso limite das normas de Minkovski, quando se toma $p \to \infty$.

Uma forma de ajudar a compreender o conceito de comprimento introduzido por cada uma destas normas consiste em ver quais são os vectores que, ao abrigo de cada norma, têm comprimento 1, ou seja, quais são os vectores unitários para cada norma. A esses vectores é hábito dar-se o nome de **bola unitária** para a norma em questão. Já sabemos que as bolas unitárias da habitual norma euclidiana estão na origem desta designação, uma vez que são, em $\mathbb{R}^2$ a circunferência de raio 1, centrada na origem; em $\mathbb{R}^3$ a esfera de raio 1, centrada na origem e em geral, em $\mathbb{R}^n$, a hiper-esfera de raio 1 e centro na origem. As bolas unitárias, em $\mathbb{R}^2$, para algumas das normas de Minkovski (e a norma do máximo) são dadas na Figura 1.2.

Teorema 1.12 *Se L é um espaço linear com a norma* $\|\cdot\|$*, verifica-se:*

1. $\|\mathbf{x} - \mathbf{y}\| \geq \left| \|\mathbf{x}\| - \|\mathbf{y}\| \right|$, $\quad \forall \mathbf{x}, \mathbf{y} \in L$
2. $\|\mathbf{x} - \mathbf{y}\| \leq \|\mathbf{x} - \mathbf{z}\| + \|\mathbf{z} - \mathbf{y}\|$, $\quad \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathrm{L}$

Exercício 1.6 *Demonstre o Teorema anterior.*

Em qualquer espaço linear L com produto interno $< \cdot, \cdot >$, pode sempre definir-se uma norma a partir do produto interno, como se verá no seguinte resultado.

[3] Em espaços lineares de dimensão infinita, é necessário substituir o *máximo* pelo *supremo* do elemento de L, sendo a norma nesse caso designada *norma do supremo*.

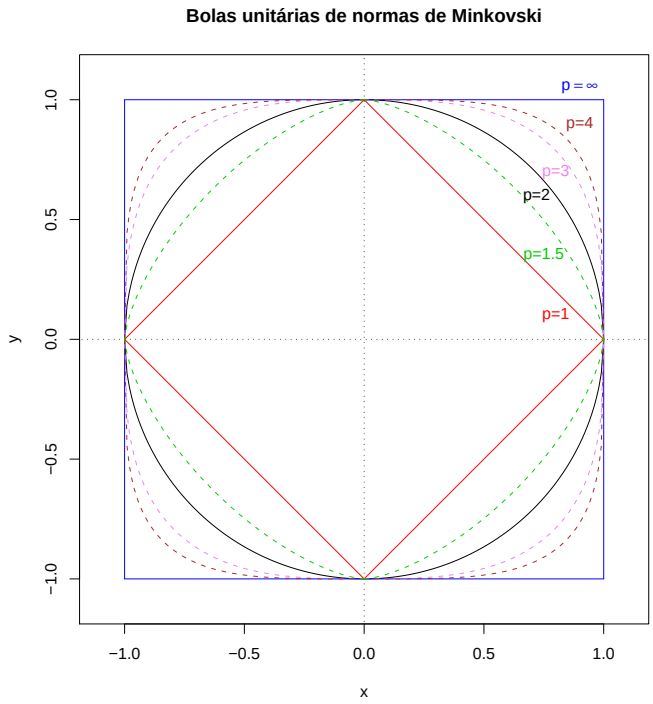


Figura 1.2: As bolas unitárias para algumas das normas de Minkovski: ℓ_1, $\ell_{3/2}$, ℓ_2 (a norma euclidiana), ℓ_3, ℓ_4 e ℓ_∞.

Teorema 1.13 *Seja L um espaço linear com um produto interno* $< \cdot, \cdot >$. *A aplicação* $\|\cdot\|_{<\cdot,\cdot>} : L \to \mathbf{R}$, *definida por*

$$\|\mathbf{x}\|_{<\cdot,\cdot>} = \sqrt{< \mathbf{x}, \mathbf{x} >}, \quad \forall \mathbf{x} \in \mathrm{L}$$

é uma norma em L, designada a ***norma induzida pelo produto interno****.*

Demonstração: Para verificar que a aplicação é, efectivamente, uma norma, será necessário confirmar se satisfaz as propriedades da Definição 1.10. A *positividade* verifica-se directamente a partir das propriedades do produto interno, uma vez que $< \mathbf{x}, \mathbf{x} >$ é não negativo, e é nulo se e só se $\mathbf{x} = \mathbf{0}$. A *homogeneidade* resulta de

$$\|\alpha\mathbf{x}\|_{<\cdot,\cdot>} = \sqrt{< \alpha\mathbf{x}, \alpha\mathbf{x} >} = \sqrt{\alpha^2 \ < \mathbf{x}, \mathbf{x} >} = |\alpha| \ \|\mathbf{x}\|_{<\cdot,\cdot>} .$$

Finalmente, a *desigualdade triangular* é consequência da desigualdade de Cauchy-Schwarz-Buniakovski, uma vez que

$$\begin{aligned} \|\mathbf{x}+\mathbf{y}\|^2_{<\cdot,\cdot>} &= <\mathbf{x}+\mathbf{y},\mathbf{x}+\mathbf{y}> \\ &= <\mathbf{x},\mathbf{x}> \ + 2<\mathbf{x},\mathbf{y}> \ + \ <\mathbf{y},\mathbf{y}> \\ &\leq \|\mathbf{x}\|^2_{<\cdot,\cdot>} \ + \ 2\,\|\mathbf{x}\|_{<\cdot,\cdot>}\cdot\|\mathbf{y}\|_{<\cdot,\cdot>} \ + \ \|\mathbf{y}\|^2_{<\cdot,\cdot>} \\ &= \left(\|\mathbf{x}\|_{<\cdot,\cdot>}+\|\mathbf{y}\|_{<\cdot,\cdot>}\right)^2 \\ \Longleftrightarrow \qquad \|\mathbf{x}+\mathbf{y}\|_{<\cdot,\cdot>} &\leq \|\mathbf{x}\|_{<\cdot,\cdot>}+\|\mathbf{y}\|_{<\cdot,\cdot>} \ . \end{aligned}$$

∇

Nota: Tendo em conta a noção de norma induzida, a desigualdade de Cauchy-Schwarz-Buniakovski pode ser expressa na seguinte forma mais mnemónica, que é *válida para quaisquer vectores* $\mathbf{x}$, $\mathbf{y}$, *em qualquer espaço linear* L *onde tenha sido definido um produto interno e a respectiva norma induzida*:

$$\left|<\mathbf{x},\mathbf{y}>\right| \quad \leq \quad \|\mathbf{x}\|\cdot\|\mathbf{y}\| \quad , \quad \forall \mathbf{x},\mathbf{y}\in \mathrm{L} \ , \tag{1.8}$$

Exemplo 1.2 *Em* $\mathbb{R}^n$, *a* ***norma habitual (euclidiana)*** *é a* ***norma induzida pelo produto interno usual****:*

$$\|\mathbf{x}\| = \sqrt{<\mathbf{x},\mathbf{x}>} = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{1.9}$$

Exemplo 1.3 *Considere-se o espaço euclidiano* $\mathbb{R}^n$. *Seja* $\mathbf{W}\in\mathbb{M}_{n\times n}$ *uma matriz definida positiva. A aplicação* $\|\cdot\|_{\mathbf{W}} : \mathbb{R}^n \longrightarrow \mathbb{R}$ *definida como:*

$$\|\mathbf{x}\|_{\mathbf{W}} = \sqrt{\mathbf{x}^t\mathbf{W}\mathbf{x}} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_ix_j} \tag{1.10}$$

é uma norma em $\mathbb{R}^n$.

Nota: A habitual norma euclidiana é um caso particular desta família de normas, sendo a norma que resulta de tomar $\mathbf{W}=\mathbf{I}$.

A **bola unitária** das normas induzidas por produtos internos $<\cdot,\cdot>_{\mathbf{W}}$ para matrizes definidas positivas **W são, em $\mathbb{R}^2$, elipses de centro na origem e cujos semi-eixos têm os seus comprimentos e direcções associadas aos valores e vectores próprios de W**. Esta afirmação á fácil de verificar para o caso de matrizes $\mathbf{W}$ que, além de definidas positivas, sejam diagonais. Nesse caso tem-se $\|\mathbf{x}\| = \mathbf{x}^t\mathbf{W}\mathbf{x} = w_{11}x_1^2 + w_{22}x_2^2$ (confirme!), pelo que a $\mathbf{W}$-norma unitária corresponde à equação

$$\left(\frac{x_1}{\frac{1}{\sqrt{w_1}}}\right)^2 + \left(\frac{x_2}{\frac{1}{\sqrt{w_2}}}\right)^2 = 1 \ ,$$

que é a equação duma elipse com semi-eixos na direcção dos eixos X_1 e X_2, de comprimento $\frac{1}{\sqrt{w_1}}$ e $\frac{1}{\sqrt{w_2}}$ respectivamente. **Para espaços $\mathbb{R}^n$ com $n \geq 3$ o resultado é análogo, sendo as bolas unitárias das normas-W dadas por elipsóides ou hiper-elipsóides**.

Exemplo 1.4 *Também em espaços de matrizes, a definição de um produto interno induz a definição duma norma matricial. Uma* ***norma*** *duma matriz $\mathbf{X} \in \mathbb{M}_{n\times p}$ é dada por:*

$$\|\mathbf{X}\| = \sqrt{\mathrm{tr}(\mathbf{X}^{\mathrm{t}}\mathbf{X})} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}^2} \tag{1.11}$$

Observação: Repare-se que **o quadrado da norma de uma matriz é a soma de quadrados de todos os elementos da matriz**. Este facto está de acordo com a noção de que uma norma é uma medida do *tamanho* dum elemento do espaço linear.

Distâncias

Consideremos agora o conceito de *distância* entre os elementos dum espaço linear L.

Definição 1.11 *Uma* ***distância*** *num espaço linear L é uma aplicação real $d : L\times L \to \mathbb{R}$, que verifica as seguintes propriedades:*

1. $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x}) \quad \forall \mathbf{x},\mathbf{y} \in \mathrm{L}$ *[Simétrica]*
2. $d(\mathbf{x},\mathbf{y}) \geq 0 \quad \forall \mathbf{x},\mathbf{y} \in \mathrm{L} \quad$ com $\; d(\mathbf{x},\mathbf{y}) = 0 \;$ sse $\; \mathbf{x} = \mathbf{y}$ *[Positiva]*
3. $d(\mathbf{x},\mathbf{y}) \leq d(\mathbf{x},\mathbf{z}) + d(\mathbf{y},\mathbf{z}) \quad \forall \mathbf{x},\mathbf{y},\mathbf{z} \in \mathrm{L}$ *[Desigualdade triangular]*

Teorema 1.14 *Dada uma norma $\|\cdot\|$ num espaço linear L, uma distância pode ser sempre definida à custa dessa norma:*

$$d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\| \quad , \quad \forall \mathbf{x},\mathbf{y} \in \mathrm{L}$$

Uma distância assim definida é designada uma ***distância induzida*** *pela norma $\|\cdot\|$.*

Exercício 1.7 *Demonstre o Teorema anterior.*

Exemplo 1.5 *Em $\mathbb{R}^n$, a habitual* ***distância (euclidiana)*** *é a* ***distância induzida pelo produto interno usual****:*

$$d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1.12}$$

Exercício 1.8 *Mostre que a habitual distância euclidiana verifica as propriedades de uma distância (Definição 1.11).*

Exemplo 1.6 *As* ***normas de Minkovski****, definidas nas equações (1.6) (pg. 15) e (1.7) (pg. 16) induzem distâncias em* $\mathbb{R}^n$*, designadas* ***distâncias de Minkovski****:*

$$d_p(\mathbf{x},\mathbf{y}) \;=\; \|\mathbf{x}-\mathbf{y}\|_p \;=\; \left(\sum_{i=1}^{n}|x_i-y_i|^p\right)^{\frac{1}{p}} \qquad ,\forall \mathbf{x},\mathbf{y}\in\mathbb{R}^n \;,\quad (p\geq 1)\;, \tag{1.13}$$

e

$$d_\infty(\mathbf{x},\mathbf{y}) \;=\; \|\mathbf{x}-\mathbf{y}\|_\infty \;=\; \max_i |x_i-y_i| \qquad ,\forall \mathbf{x},\mathbf{y}\in\mathbb{R}^n\;. \tag{1.14}$$

Exemplo 1.7 *Nos espaços de matrizes* $\mathbb{M}_{n\times p}$*, é possível definir uma* ***distância entre matrizes*** *a partir da norma induzida pelo habitual produto interno entre matrizes considerado no exemplo 1.1 (pg. 13):*

$$d(\mathbf{A},\mathbf{B}) \;=\; \|\mathbf{A}-\mathbf{B}\| \;=\; \sqrt{\mathrm{tr}\left((\mathbf{A}-\mathbf{B})^t(\mathbf{A}-\mathbf{B})\right)} \;=\; \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p}(a_{ij}-b_{ij})^2}\;. \tag{1.15}$$

Assim, o produto interno matricial usual considera que a distância entre duas matrizes é dada pela ***raíz quadrada da soma de quadrados das diferenças entre os seus elementos correspondentes****.*

Mas outras distâncias podem ser definidas nos espaços euclidianos, quer distâncias induzidas por outras normas, como as normas do Teorema 1.3, quer distâncias que não são induzidas por qualquer norma.

Uma das mais famosas distâncias entre vectores utilizada em estatística multivariada é a **distância de Mahalanobis**, que a seguir se define.

Definição 1.12 *Sejam* $\mathbf{x}$ *e* $\mathbf{y}$ *vectores de* $\mathbb{R}^p$ *correspondentes a observações de dois individuos em* p *variáveis. Seja* $\mathbf{\Sigma}$ *a matriz de variâncias-covariâncias associada às* p *variáveis observadas, que se admite ser uma matriz definida positiva. Designa-se* ***distância de Mahalanobis*** *entre os individuos associados a* $\mathbf{x}$ *e* $\mathbf{y}$ *à distância induzida pela norma* $\|\cdot\|_{\mathbf{\Sigma}^{-1}}$*:*

$$d_{\Sigma^{-1}}(\mathbf{x},\mathbf{y}) \;=\; \|\mathbf{x}-\mathbf{y}\|_{\mathbf{\Sigma}^{-1}} \;=\; \sqrt{(\mathbf{x}-\mathbf{y})^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{y})}\;. \tag{1.16}$$

Uma razão invocada para utilizar a distância de Mahalanobis reside no facto de permanecer invariante a transformações afins diferenciadas nas p variáveis, como no caso da conversão de unidades de medida do sistema métrico para unidades do sistema anglo-saxónico.

Ângulos e ortogonalidade

Já vimos como o conceito de produto interno permite definir comprimentos (normas) e distâncias em espaços lineares. Veremos agora que também o conceito de ângulo pode ser definido num espaço linear genérico a partir do conceito de produto interno.

Definição 1.13 *Seja L um espaço linear com produto interno $< \cdot, \cdot >$. Sejam $\mathbf{x}, \mathbf{y} \in L$, $\mathbf{x}, \mathbf{y} \neq \mathbf{0}$. O* ***ângulo*** *entre $\mathbf{x}$ e $\mathbf{y}$ define-se como $\angle(\mathbf{x}, \mathbf{y}) = \arccos\left(\frac{<\mathbf{x},\mathbf{y}>}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}\right)$.*

Observações:

1. Da definição resulta que o coseno do ângulo entre $\mathbf{x}$ e $\mathbf{y}$ $(\mathbf{x}, \mathbf{y} \neq \mathbf{0})$ é dado por $\cos(\mathbf{x}, \mathbf{y}) = \frac{<\mathbf{x},\mathbf{y}>}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}$
2. Quando $\mathbf{x} = \mathbf{0}$ ou $\mathbf{y} = \mathbf{0}$, o quociente que define o coseno resulta numa indeterminação, não estando nesse caso o ângulo bem definido.

Definição 1.14 *Seja L um espaço linear com produto interno. Dois vectores dizem-se* ***ortogonais*** *se $< \mathbf{x}, \mathbf{y} >= 0$. Nesse caso, escreve-se $\mathbf{x} \perp \mathbf{y}$.*

Observação: Da definição anterior resulta que, para $\mathbf{x}, \mathbf{y} \neq \mathbf{0}$, $\mathbf{x} \perp \mathbf{y} \Leftrightarrow \cos(\mathbf{x}, \mathbf{y}) = 0$. Repare-se que *a ortogonalidade depende do produto interno usado.*

Teorema 1.15 *Seja L um espaço linear, com produto interno. Um conjunto de n vectores não-nulos de L, ortogonais entre si dois a dois, é necessariamente um conjunto de vectores linearmente independente.*

Demonstração. Para que o conjunto de vectores $\{x_i\}_{i=1}^n$ seja linearmente independente é necessário que $\sum_{i=1}^n \alpha_i x_i = \mathbf{0}$ implique que $\alpha_i = 0$, $\forall\ i \in \{1, 2, ..., n\}$. Ora, se $\sum_{i=1}^n \alpha_i x_i = \mathbf{0}$, o produto interno desta combinação linear com *qualquer* vector x_j será nulo, pois $< \mathbf{0}, x_j >= 0$. Mas este produto interno também pode ser re-escrito, tendo em conta as propriedades do produto interno, como $\sum_{i=1}^n \alpha_i < x_i, x_j >$. Se os vectores são ortogonais dois a dois, o único produto interno que não se anula ocorre quando $j = i$, tendo-se então $\sum_{i=1}^n \alpha_i < x_i, x_j >= \alpha_j \|x_j\|^2$, que apenas se pode anular se $\alpha_j = 0$ (o vector x_j é por hipótese não nulo). Mas como o raciocínio é válido para qualquer $j \in \{1, 2, ..., n\}$, tem-se a condição para que os vectores formem um conjunto linearmente independente. ∇

Definição 1.15 *Seja L um espaço linear com produto interno e seja M um subespaço de L. O conjunto de vectores de L que são ortogonais a* todos *os vectores de M designa-se o* ***complemento ortogonal*** *de M em L, e representa-se por $M^{\perp}$.*

Teorema 1.16 *Seja L um espaço linear com produto interno e seja M um subespaço de L. O complemento ortogonal de M em L, $M^{\perp}$, é um* ***subespaço*** *linear de L.*

Demonstração. Para provar que $M^{\perp}$ é subespaço de L, basta provar que não é vazio e é fechado para as combinações lineares dos seus elementos, *i.e.*, que $\forall \mathbf{x}, \mathbf{y} \in M^{\perp} \Rightarrow \alpha\mathbf{x} + \beta\mathbf{y} \in M^{\perp}$, $\forall \alpha, \beta \in \mathbb{R}$. Ora, o elemento nulo de L tem de pertence a $M^{\perp}$, uma vez que o produto interno do elemento nulo com qualquer vector de L é zero, e portanto, em particular, $\mathbf{0}$ é ortogonal a todos os elementos de M. Por outro lado, a definição de $M^{\perp}$ leva a que:

$$< \alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z} > \quad = \quad \alpha < \mathbf{x}, \mathbf{z} > \; + \beta < \mathbf{y}, \mathbf{z} > \quad = \quad 0 \quad , \qquad \forall \mathbf{z} \in M \; .$$

Logo, $\alpha\mathbf{x} + \beta\mathbf{y} \in M^{\perp}$. ∇

Definição 1.16 *Seja L um espaço linear n-dimensional com produto interno. Uma base* $\{\mathbf{x}_i\}_{i=1}^{n}$ *de L diz-se uma* ***base ortonormada*** *se os vectores da base forem todos:*

1. *unitários, i.e., de norma um* ($\|\mathbf{x}_i\| = 1$, $\forall i$)*, na norma induzida pelo produto interno.*
2. *ortogonais entre si* ($< \mathbf{x}_i, \mathbf{x}_j >= 0$ *, se* $i \neq j$)*.*

Observações:

1. A base canónica de $\mathbb{R}^n$ é uma base ortonormada para o habitual produto interno em $\mathbb{R}^n$.
2. Qualquer espaço com produto interno possui uma base ortonormada. Recorde-se que o *processo de ortogonalização de Gram-Schmidt* permite transformar uma base genérica numa base ortonormada.

1.3 Projecções. O Teorema de Pitágoras

Definição 1.17 *Seja L um espaço linear e* L_1*,* L_2 *dois seus subespaços lineares.*

1. *O conjunto de elementos* $\mathbf{x} \in L$ *que se podem escrever como* $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$ *para algum vector* $\mathbf{x}_1 \in L_1$ *e algum vector* $\mathbf{x}_2 \in L_2$*, diz-se o* ***conjunto soma de*** L_1 ***e*** L_2 *e representa-se por* $L_1 + L_2$*.*
2. *Se cada vector* $\mathbf{x} \in L_1 + L_2$ *tem uma decomposição* única *como soma de uma parcela em* L_1 *e uma parcela em* L_2 *(i.e., uma decomposição única da forma* $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$ *com* $\mathbf{x}_1 \in L_1$ *e* $\mathbf{x}_2 \in L_2$*), diz-se que* L_1 *e* L_2 *definem uma* ***soma directa*** *do espaço* $L_1 + L_2$ *e escreve-se* $L_1 \oplus L_2$*.*

Teorema 1.17 *Seja L um espaço linear e* L_1*,* L_2 *dois seus subespaços lineares. A soma* $L_1 + L_2$ *é um subespaço de L.*

Teorema 1.18 *Seja L um espaço linear e M, N dois seus subespaços. Então* $L = M \oplus N$ *se e só se:*

1. $L = M + N$*.*

2. $M \cap N = \{\mathbf{0}\}$.

Teorema 1.19 *Seja $L = M \oplus N$. Então:*

1. *A reunião de uma base de M com uma base de N constitui uma base de L.*
2. *$dim(L) = dim(M) + dim(N)$*

Teorema 1.20 *Seja L um espaço linear com produto interno e M qualquer subespaço de L. Então,* $\mathrm{L} = \mathrm{M} \oplus \mathrm{M}^{\perp}$.

Observação: Isto significa que **qualquer vector de L se pode sempre escrever de forma *única* como a soma de um vector em M e de outro vector de $\mathrm{M}^{\perp}$, *i.e.*, ortogonal a M**.

Definição 1.18 *Seja $L = M \oplus N$. Uma aplicação $\mathbf{P}$ que associa a cada $\mathbf{z} \in L$ a sua componente única em M (i.e., tal que se $\mathbf{z} = \mathbf{x} + \mathbf{y}$, com $\mathbf{x} \in M$ e $\mathbf{y} \in N$, se tem $\mathbf{Pz} = \mathbf{x}$) diz-se uma **projecção de L sobre M, ao longo de N**. Se $N = M^{\perp}$, diz-se que $\mathbf{P}$ é a **projecção ortogonal de L sobre M**.*

É fácil de verificar que **as projecções são aplicações lineares**. Verifica-se então o seguinte resultado, que permite falar sempre em "o" projector sobre um subespaço, ao longo de outro.

Teorema 1.21 *Dado um espaço linear L e uma soma directa $L = M \oplus N$, o projector sobre M ao longo de N é único.*

Definição 1.19 *Uma aplicação linear $\mathbf{P}$ num espaço linear L diz-se:*

1. *uma aplicação **idempotente** se $\mathbf{P}^2 = \mathbf{P}$, onde por $\mathbf{P}^2$ se entende a aplicação $\mathbf{P}^2(\mathbf{x}) = \mathbf{P}(\mathbf{P}(\mathbf{x}))$.*
2. *uma aplicação **identidade** se $\mathbf{Px} = \mathbf{x}$, $\forall \mathbf{x} \in L$.*

Observação. É usual indicar uma aplicação identidade utilizando a letra $\mathbf{I}$.

Teorema 1.22 *Seja $\mathbf{P}$ uma aplicação linear no espaço linear L, e $\mathbf{I}$ a aplicação identidade. Então:*

1. *$\mathbf{P}$ é uma projecção em L se e só se $\mathbf{P}$ é idempotente.*
2. *Se $\mathbf{P}$ é idempotente, $\mathbf{P}$ projecta sobre o seu subespaço imagem, $\mathcal{C}(\mathbf{P})$, ao longo do seu núcleo, $\mathcal{N}(\mathbf{P})$.*
3. *Se $\mathbf{P}$ é idempotente, $\mathbf{I} - \mathbf{P}$ projecta sobre o núcleo de $\mathbf{P}$, $\mathcal{N}(\mathbf{P})$, ao longo da subespaço imagem de $\mathbf{P}$, $\mathcal{C}(\mathbf{P})$.*

Munindo o espaço linear L dum produto interno, e sendo M um subespaço de L, verifica-se ainda

4. *Se* $\mathbf{P}$ *é uma projecção ortogonal sobre um subespaço M de L, então* $\mathbf{I}-\mathbf{P}$ *é uma projecção ortogonal sobre* $M^{\perp}$.

Observações:

1. **Se P é uma aplicação idempotente, P projecta sobre o conjunto de vectores que permanecem invariantes sob o seu efeito** (isto é, o conjunto de vectores $\mathbf{x} \in$L tais que $\mathbf{Px} = \mathbf{x}$), **ao longo do núcleo de P** (isto é, ao longo do conjunto de vectores $\mathbf{x} \in$L tais que $\mathbf{Px} = \mathbf{0}$). Assim, **os vectores de um subespaço permanecem invariantes sob o efeito de um projector sobre esse subespaço**.

2. Em aplicações estatísticas, é frequente designar o vector $(\mathbf{I}-\mathbf{P})\mathbf{z}$ como o *vector residual* de $\mathbf{z}$ após a sua projecção ortogonal sobre M.

As projecções *ortogonais* desempenham um papel decisivo em muitos campos da Estatística Multivariada. A principal razão dessa importância reside no seguinte Teorema, de índole muito geral.

Teorema 1.23 *Seja L um espaço linear com produto interno e* $\|\cdot\|$ *a norma induzida pelo produto interno. Seja M um subespaço de L, e* $\mathbf{P}$ *o projector ortogonal sobre M. Dado qualquer vector (não-nulo)* $\mathbf{z} \in L$, *verifica-se:*

1. ***(Teorema de Pitágoras.)*** *O quadrado da norma de* $\mathbf{z}$ *é a soma dos quadrados das normas das suas componentes em M e em* $M^{\perp}$, *isto é:* $\|\mathbf{z}\|^2 = \|\mathbf{Pz}\|^2 + \|(\mathbf{I}-\mathbf{P})\mathbf{z}\|^2$.

2. *O cosseno do ângulo entre um vector* $\mathbf{z} \notin M^{\perp}$ *e a sua projecção ortogonal sobre M é dada por:*

$$\cos(\mathbf{z}, \mathbf{Pz}) = \frac{\|\mathbf{Pz}\|}{\|\mathbf{z}\|}$$

3. *O vector no subespaço M (que designaremos por* $\hat{\mathbf{z}}$*) mais próximo do vector* $\mathbf{z}$ *(isto é, o vector que minimiza a distância* $\|\mathbf{z}-\mathbf{y}\|$, *de entre todos os vectores* $\mathbf{y} \in L$*), é a projecção ortogonal de* $\mathbf{z}$ *sobre M, isto é,* $\hat{\mathbf{z}} = \mathbf{Pz}$.

4. *Os vectores no subespaço M que formam o mais pequeno ângulo com o vector* $\mathbf{z} \notin M^{\perp}$ *são os vectores que apontam no mesmo sentido que* $\mathbf{Pz}$, *ou seja, os vectores* $\mathbf{y} = \alpha\hat{\mathbf{z}}$, $\forall\, \alpha \in \mathbb{R}$.

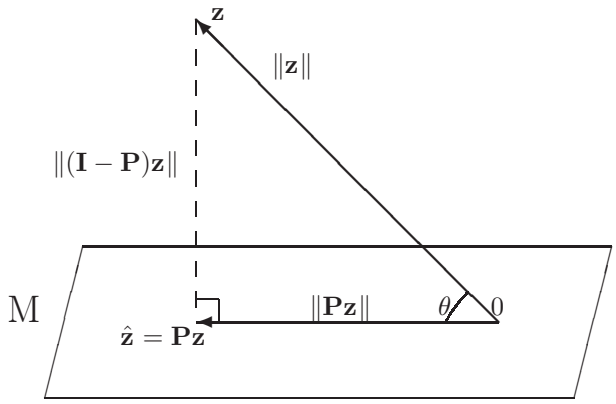


Figura 1.3: Ilustração do Teorema de Pitágoras.

Daqui em diante cingir-nos-emos apenas a projecções nos espaços $\mathbb{R}^k$.

Consideremos agora os espaços reais, $\mathbb{R}^k$, munidos do habitual produto interno Euclideano: $< \mathbf{x}, \mathbf{y} > = \mathbf{x}^t\mathbf{y}$. *As aplicações lineares de $\mathbb{R}^m$ em $\mathbb{R}^n$ correspondem às matrizes do tipo nxm*. As aplicações lineares em $\mathbb{R}^k$ correspondem às matrizes de tipo $k\times k$. Assim, a cada aplicação linear (e admitindo que se convenciona trabalhar apenas com as bases canónicas de $\mathbb{R}^k$) corresponde uma matriz $\mathbf{A} \in \mathbb{M}_{k\times k}$.

As matrizes de projecção ortogonal em $\mathbb{R}^k$ são as matrizes simétricas ($\mathbf{A}^t = \mathbf{A}$) e idempotentes ($\mathbf{A}^2 = \mathbf{A}$) de tipo $k\times k$, como prova o seguinte Teorema.

Teorema 1.24 *Seja $\mathbb{R}^k = M \oplus M^\perp$, em que M é um subespaço de $\mathbb{R}^k$. Considere o produto interno usual em $\mathbb{R}^k$. Seja* **B** *uma matriz* k×r *cujas r colunas formam uma qualquer base de M. A matriz* **P** *de projecção ortogonal sobre M é única e tem a forma:*

$$\mathbf{P} = \mathbf{B}(\mathbf{B}^t\mathbf{B})^{-1}\mathbf{B}^t$$

Teorema 1.25 *Seja* **P** *uma matriz de dimensão $n \times n$. Então,* **P** *é uma matriz de projecção ortogonal sobre algum subespaço de $\mathbb{R}^n$ se e só se* **P** *fôr simétrica e idempotente.*

As matrizes de projecção ortogonal em subespaços de $\mathbb{R}^n$ têm uma caracterização interessante dos seus valores e vectores próprios.

Teorema 1.26 *Seja M um subespaço r-dimensional de $\mathbb{R}^k$, e $\mathbf{P}_M$ a matriz de projecção ortogonal sobre M. Então:*

1. *Os valores próprios de $\mathbf{P}_M$ apenas tomam valor 0 ou 1, havendo precisamente $r = dim(M)$ valores próprios de valor 1 e $k - r = dim(M^\perp)$ valores próprios de valor 0.*
2. *Os vectores próprios associados a valores próprios 1 formam uma base ortonormada de M. Os vectores próprios associados a valores próprios 0 formam uma base ortonormada de $M^\perp$.*
3. *O traço da matriz de projecção $\mathbf{P}_M$ equivale à dimensão do subespaço M sobre o qual $\mathbf{P}_M$ projecta.*
4. *A matriz $\mathbf{P}_M$ é semi-definida positiva.*

Resumo:

Seja $\mathbf{y} \in \mathbb{R}^k$ um vector e M um subespaço linear r-dimensional de $\mathbb{R}^k$ com uma base constituída pelas colunas da matriz $\mathbf{B}$. A *projecção ortogonal de* $\mathbf{y}$ *sobre* M (com o produto interno usual) é o vector:

$$\hat{\mathbf{y}} = \mathbf{P}\mathbf{y} = \mathbf{B}(\mathbf{B}^t\mathbf{B})^{-1}\mathbf{B}^t\mathbf{y}$$

O vector (de tipo $r \times 1$):

$$(\mathbf{B}^t\mathbf{B})^{-1}\mathbf{B}^t\mathbf{y}$$

é o *vector dos r coeficientes* da combinação linear que define de forma única o *vector projectado* $\hat{\mathbf{y}} \in$M em termos dos vectores da base $\mathbf{B}$ de M.

Teorema 1.27 *Seja M um subespaço linear de $\mathbb{R}^k$ e N um subespaço próprio de M ($N \subset M \subset \mathbb{R}^k$). Sejam $\mathbf{P}_M$ e $\mathbf{P}_N$ as matrizes de projecção ortogonal sobre M e N, respectivamente. Sejam $\mathbf{P}_{M^\perp}$ e $\mathbf{P}_{N^\perp}$ as matrizes de projecção ortogonal sobre os complementos ortogonais de M e N. Então, tem-se:*

1. $\mathbf{P}_M\mathbf{P}_N = \mathbf{P}_N\mathbf{P}_M = \mathbf{P}_N$.
2. $\mathbf{P}_M\mathbf{P}_{N^\perp} = \mathbf{P}_{N^\perp}\mathbf{P}_M = \mathbf{P}_M - \mathbf{P}_N$.
3. $\mathbf{P}_N\mathbf{P}_{M^\perp} = \mathbf{P}_{M^\perp}\mathbf{P}_N = \mathbf{0}$.
4. $\mathbf{P}_{M^\perp}\mathbf{P}_{N^\perp} = \mathbf{P}_{N^\perp}\mathbf{P}_{M^\perp} = \mathbf{P}_{M^\perp}$.

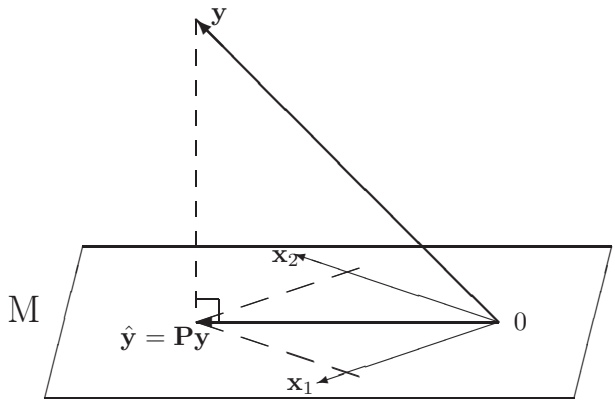


Figura 1.4: Projecção do vector $\mathbf{y}$ sobre o subespaço M, gerado pelos vectores $\mathbf{x}_1$ e $\mathbf{x}_2$. As coordenadas do vector projectado nos eixos $\mathbf{x}_1$ e $\mathbf{x}_2$ são dadas pelos elementos do vector $(\mathbf{B}^t\mathbf{B})^{-1}\mathbf{B}^t\mathbf{y}$, onde a matriz $\mathbf{B}$ é a matriz cujas duas colunas são os vectores da base, $\mathbf{x}_1$ e $\mathbf{x}_2$.

1.3.1 Mais propriedades de espaços imagem e núcleos

Sabemos (disciplina de Complementos de Álgebra e Análise) que se $\mathbf{A}$ é uma matriz $n \times p$ (ou seja, uma transformação linear de $\mathbb{R}^p$ em $\mathbb{R}^n$, então o complemento ortogonal do espaço coluna de qualquer matriz $\mathbf{A}$ é igual ao espaço nulo da transposta de $\mathbf{A}$, ou seja,

$$\mathcal{C}(\mathbf{A})^{\perp} = \mathcal{N}(\mathbf{A}^t) \tag{1.17}$$

Nesta Secção veremos alguns resultados adicionais relativos a espaços imagem e núcleos de matrizes.

O seguinte resultado auxiliar, relacionando o espaço imagem de um produto de matrizes com o espaço imagem do primeiro factor nesse produto, será de utilidade para obter resultados na Secção seguinte.

Teorema 1.28 *Seja $\mathbf{A} \in \mathbb{M}_{n\times p}$ e $\mathbf{B} \in \mathbb{M}_{p\times m}$. Então, verificam-se as seguintes relações:*

1. *entre os espaços imagem de $\mathbf{A}$ e de $\mathbf{AB}$:* $\mathcal{C}(\mathbf{AB}) \subset \mathcal{C}(\mathbf{A})$.
2. *entre os espaços nulos (núcleos) de $\mathbf{B}$ e $\mathbf{AB}$:* $\mathcal{N}(\mathbf{B}) \subset \mathcal{N}(\mathbf{AB})$.

Demonstração:

1. Tem-se

$$\begin{aligned} \mathbf{y} \in \mathcal{C}(\mathbf{AB}) &\iff \exists\, \mathbf{x} \text{ t.q. } \mathbf{y} = \mathbf{ABx} \\ &\implies \exists\, \mathbf{z} = \mathbf{Bx} \text{ t.q. } \mathbf{y} = \mathbf{Az} \\ &\iff \mathbf{y} \in \mathcal{C}(\mathbf{A}) \end{aligned}$$

2. Tem-se

$$\begin{aligned} \mathbf{x} \in \mathcal{N}(\mathbf{B}) &\iff \mathbf{Bx} = \mathbf{0}_p \\ &\implies \mathbf{ABx} = \mathbf{A0}_p = \mathbf{0}_n \\ &\iff \mathbf{x} \in \mathcal{N}(\mathbf{AB}) \end{aligned}$$

∇

No seguinte Teorema, toma-se $\mathbf{A} = \mathbf{B}^t$ (ou, o que é a mesma coisa, $\mathbf{A}^t = \mathbf{B}$) no Teorema anterior, para relacionar espaços imagem e núcleos de matrizes do tipo $\mathbf{XX}^t$, $\mathbf{X}^t\mathbf{X}$, $\mathbf{X}$ e $\mathbf{X}^t$.

Teorema 1.29 *Seja* $\mathbf{X}$ *uma matriz* $n \times p$. *Tem-se:*

1. $\mathcal{N}(\mathbf{XX}^t) = \mathcal{N}(\mathbf{X}^t)$.
2. $\mathcal{N}(\mathbf{X}^t\mathbf{X}) = \mathcal{N}(\mathbf{X})$.
3. $\mathcal{C}(\mathbf{XX}^t) = \mathcal{C}(\mathbf{X})$.
4. $\mathcal{C}(\mathbf{X}^t\mathbf{X}) = \mathcal{C}(\mathbf{X}^t)$.

Demonstração:

1. Que $\mathcal{N}(\mathbf{X}^t) \subset \mathcal{N}(\mathbf{XX}^t)$ é imediato a partir do Teorema 1.28, tomando $\mathbf{A} = \mathbf{X}$ e $\mathbf{B} = \mathbf{X}^t$. Falta provar a inclusão contrária: $\mathcal{N}(\mathbf{XX}^t) \subset \mathcal{N}(\mathbf{X}^t)$. Provar esta inclusão significa provar que se $\mathbf{y} \in \mathcal{N}(\mathbf{XX}^t)$, então $\mathbf{y} \in \mathcal{N}(\mathbf{X}^t)$. Por outras palavras, significa provar que se $\mathbf{XX}^t\mathbf{y} = \mathbf{0}_n$, então $\mathbf{X}^t\mathbf{y} = \mathbf{0}_p$. Mas

   $$\mathbf{XX}^t\mathbf{y} = \mathbf{0}_n \implies \mathbf{y}^t\mathbf{XX}^t\mathbf{y} = \mathbf{y}^t\mathbf{0}_n = 0 \iff \|\mathbf{X}^t\mathbf{y}\| = 0 \iff \mathbf{X}^t\mathbf{y} = \mathbf{0}_p\ ,$$

   pela definição de norma.

2. Análogo ao anterior, começando com $\mathbf{A} = \mathbf{X}^t$ e $\mathbf{B} = \mathbf{X}$.

3. Sabemos (equação 1.17) que $\mathcal{N}(\mathbf{X}^t) = \mathcal{C}(\mathbf{X})^\perp$ e $\mathcal{N}(\mathbf{XX}^t) = \mathcal{C}(\mathbf{XX}^t)^\perp$ ($\mathbf{XX}^t$ é simétrica). Logo, a partir do primeiro resultado, tem-se $\mathcal{C}(\mathbf{XX}^t)^\perp = \mathcal{C}(\mathbf{X})^\perp$. Como o complemento ortogonal dum complemento ortogonal é o subespaço inicial, tem-se $\mathcal{C}(\mathbf{XX}^t) = \mathcal{C}(\mathbf{X})$.

4. Análogo.

∇

1.4 Mais teoria de matrizes

1.4.1 Característica de matrizes

Definição 1.20 *A **característica duma matriz** $\mathbf{A}$ é a dimensão do subespaço imagem $\mathcal{C}(\mathbf{A})$ (o subespaço gerado pelas colunas da matriz $\mathbf{A}$). A característica de $\mathbf{A}$ costuma representar-se por $car(\mathbf{A})$*[4].

Algumas propriedades relevantes associadas à característica de matrizes, estudadas na disciplina de Complementos de Álgebra e Análise.

1. **A característica duma matriz é o número máximo de colunas linearmente independentes da matriz, que é necessariamente igual ao número máximo de linhas linearmente independentes de A** (mesmo que a matriz $\mathbf{A}$ seja rectangular).
2. **A característica duma matriz é igual à característica da sua transposta:** $car(\mathbf{A}) = car(\mathbf{A}^t)$ (o que é consequência directa da afirmação na alínea anterior).
3. **Uma matriz quadrada de tipo $p \times p$ é invertível se e só se tem característica** p.
4. Do ponto anterior resulta que, para qualquer matriz invertível $\mathbf{A}$, tem-se:

$$car(\mathbf{A}) = car(\mathbf{A}^{-1}) \tag{1.18}$$

Teorema 1.30 *Se $\mathbf{A}$ e $\mathbf{B}$ são matrizes compatíveis, a característica do seu produto não pode exceder a característica de qualquer das matrizes:*

$$car(\mathbf{AB}) \leq \min\{car(\mathbf{A}), car(\mathbf{B})\} \tag{1.19}$$

Demonstração: Por definição, a característica da matriz $\mathbf{AB}$ é a dimensão do subespaço imagem de $\mathbf{AB}$, isto é, a dimensão de $\mathcal{C}(\mathbf{AB})$. Ora, $\mathcal{C}(\mathbf{AB}) \subset \mathcal{C}(\mathbf{A})$ (Teorema 1.28). Logo, necessariamente $dim\,(\mathcal{C}(\mathbf{AB})) \leq dim\,(\mathcal{C}(\mathbf{A}))$. Por outro lado, e tendo em conta a alínea 2 das observações acima, $car(\mathbf{AB}) = car\,((\mathbf{AB})^t) = car(\mathbf{B}^t\mathbf{A}^t)$. Por um raciocínio análogo ao utilizado acima, tem-se $car(\mathbf{B}^t\mathbf{A}^t) \leq car(\mathbf{B}^t) = car(\mathbf{B})$. Logo, $car(\mathbf{AB}) \leq car(\mathbf{B})$. Se $car(\mathbf{AB})$ é majorada quer por $car(\mathbf{A})$, quer por $car(\mathbf{B})$, é majorada pela menor destas quantidades, c.q.d. ∇

O seguinte Teorema que afirma que se $\mathbf{A} = \mathbf{B}^t$ no teorema anterior, é possível garantir a igualdade das características.

[4] *Ou por $r(\mathbf{A})$ da palavra inglesa para característica de uma matriz, que é* rank.

Teorema 1.31 *Para qualquer matriz (mesmo rectangular)* $\mathbf{X}$*, tem-se:*

$$car(\mathbf{X}) = car(\mathbf{X}^t\mathbf{X}) = car(\mathbf{X}\mathbf{X}^t) \tag{1.20}$$

Demonstração: Sabemos do Teorema 1.29 que $\mathcal{C}(\mathbf{X}) = \mathcal{C}(\mathbf{X}\mathbf{X}^t)$ e $\mathcal{C}(\mathbf{X}^t) = \mathcal{C}(\mathbf{X}^t\mathbf{X})$. As dimensões de subespaços são também iguais. Mas pela alínea 2 acima, $dim(\mathcal{C}(\mathbf{X})) = dim(\mathcal{C}(\mathbf{X}^t))$. Logo, os quatro subespaços imagem referidos têm a mesma dimensão, isto é, as quatro matrizes $\mathbf{X}$, $\mathbf{X}^t$, $\mathbf{X}\mathbf{X}^t$ e $\mathbf{X}^t\mathbf{X}$ têm a mesma característica. ∇

1.4.2 Vectores e valores próprios

Definição 1.21 *Considere-se uma matriz quadrada* $\mathbf{A}_{pxp}$*. Um vector* $\mathbf{x} \in \mathbb{C}^p$ $(\mathbf{x} \neq \mathbf{0})$ *tal que:*

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

para algum escalar $\lambda \in \mathbb{C}$ *diz-se um* ***vector próprio*** *da matriz* $\mathbf{A}$*. O escalar* λ *diz-se o* ***valor próprio*** *associado ao vector próprio* $\mathbf{x}$*.*

Os valores próprios duma matriz de dimensão $p \times p$, $\mathbf{A}$, são as raízes do seu *polinómio característico*: $\det(\lambda\mathbf{I}_p - \mathbf{A})$=0. Este polinómio característico é um polinómio de ordem p em λ. Por isso, uma matriz $\mathbf{A}_{pxp}$ tem p valores próprios (reais ou complexos), embora alguns possam ser iguais (*i.e.*, o polinómio característico pode ter raízes repetidas).

O traço duma matriz é também a soma dos seus valores próprios: $\text{tr}(\mathbf{A}) = \sum_{i=1}^{p} \lambda_i$. O determinante duma matriz é também o produto dos seus valores próprios: $\det(\mathbf{A}) = \prod_{i=1}^{p} \lambda_i$.

Se $\mathbf{A}_{p\times p}$ fôr uma matriz real *simétrica*, os seus valores/vectores próprios são bem comportados:

- **Os valores próprios e os correspondentes vectores próprios são sempre *reais*.**
- **Vectores próprios associados a valores próprios diferentes são sempre ortogonais. E é sempre possível determinar um conjunto ortonormado de p vectores próprios, mesmo quando há valores próprios repetidos.**

Teorema 1.32 ***(Teorema da Decomposição Espectral).*** *Seja* $\mathbf{A} \in \mathbb{M}_{p\times p}$*. Então* $\mathbf{A}$ *é simétrica se e só se existir uma matriz ortogonal* $\mathbf{V} \in \mathbb{M}_{p\times p}$ *(com colunas* $\mathbf{v}_i$*) e uma matriz diagonal* $\mathbf{\Lambda} \in \mathbb{M}_{p\times p}$ *(com elementos diagonais* λ_i*), tal que:*

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^t \tag{1.21}$$

$$\iff \mathbf{A} = \sum_{i=1}^{p} \lambda_i \mathbf{v}_i \mathbf{v}_i^t \tag{1.22}$$

Observações:

1. Para uma discussão mais pormenorizada deste, e de anteriores resultados relativos a matrizes, vejam-se os apontamentos da disciplina de Complementos de Álgebra e Análise, ou, por exemplo, o livro *Matrix Analysis* de Horn, R. e Johnson, C. (Cambridge University Press,1985).

2. Os valores λ_i são valores próprios de $\mathbf{A}$ e os vectores $\mathbf{v}_i$ são vectores próprios de $\mathbf{A}$, constituíndo um conjunto ortonormado (pois $\mathbf{V}$ é matriz ortogonal).

3. Se todos os valores próprios forem diferentes, os vectores próprios $\mathbf{v}_i$ são únicos, a menos de troca de sinal (isto é, tanto se pode usar $\mathbf{v}_i$ como $-\mathbf{v}_i$). Nesse caso, uma ordenação dos valores próprios ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$) torna a decomposição única (a menos de trocas de sinais nas colunas de $\mathbf{V}$).

4. Se houver valores próprios iguais, a decomposição já não é única, nem com as ressalvas da observação anterior. De facto, seja $\mathbf{x}_i$ um vector próprio associado ao valor próprio λ, e seja $\mathbf{x}_j$ outro vector próprio (ortogonal a $\mathbf{x}_i$) associado ao mesmo valor próprio λ. Então, tem-se:

$$\begin{aligned} \mathbf{A}(\alpha\mathbf{x}_i + \beta\mathbf{x}_j) &= \alpha\mathbf{A}\mathbf{x}_i + \beta\mathbf{A}\mathbf{x}_j \\ &= \alpha\lambda\mathbf{x}_i + \beta\lambda\mathbf{x}_j \\ &= \lambda(\alpha\mathbf{x}_i + \beta\mathbf{x}_j) \quad \forall\alpha,\beta \in \mathbb{R} \end{aligned}$$

   Ou seja, qualquer combinação linear de $\mathbf{x}_i$ e $\mathbf{x}_j$ também será vector próprio de $\mathbf{A}$ associado ao valor próprio λ. Nesse caso, as colunas de $\mathbf{V}$ associadas a um mesmo valor próprio podem ser qualquer base ortonormada dum espaço cuja dimensão é dada pelo número de elementos diagonais de Λ (valores próprios) iguais.

5. Com base na Decomposição Espectral, é fácil demonstrar que o traço duma matriz simétrica $\mathbf{A}$, além de ser, por definição, a soma dos elementos diagonais de $\mathbf{A}$, é também a soma dos valores próprios de $\mathbf{A}$. (Demonstre!).

Na disciplina de Complementos de Álgebra e Análise foi já visto que existe uma forma alternativa de caracterizar as matrizes (semi-)definidas positivas, (semi-)definidas negativas e indefinidas em termos dos seus valores próprios, como se recorda no Teorema seguinte.

Teorema 1.33 *Seja* $\mathbf{A}$ *uma matriz simétrica, de tipo* $p \times p$*. Então:*

$\mathbf{A}$ *é definida positiva*	$\Longleftrightarrow$	*Todos os valores próprios de* $\mathbf{A}$ *são positivos.*
$\mathbf{A}$ *é semi-definida positiva*	$\Longleftrightarrow$	*Todos os valores próprios de* $\mathbf{A}$ *são não-negativos e pelo menos um é zero.*
$\mathbf{A}$ *é definida negativa*	$\Longleftrightarrow$	*Todos os valores próprios de* $\mathbf{A}$ *são negativos.*
$\mathbf{A}$ *é semi-definida negativa*	$\Longleftrightarrow$	*Todos os valores próprios de* $\mathbf{A}$ *são não-positivos e pelo menos um é zero.*
$\mathbf{A}$ *é indefinida*	$\Longleftrightarrow$	$\exists i : \lambda_i < 0, \quad \exists j : \lambda_j > 0$

Demonstração: Pelo Teorema da Decomposição Espectral (Teorema 1.32), a matriz simétrica $\mathbf{A}$ pode ser escrita como $\mathbf{A} = \sum_{i=1}^{p} \lambda_i \mathbf{v}_i \mathbf{v}_i^t$, para um conjunto ortonormado de p vectores próprios de $\mathbf{A}$, $\{\mathbf{v}_i\}_{i=1}^p$, e sendo λ_i o i-ésimo maior valor próprio de $\mathbf{A}$. Assim, para qualquer vector $\mathbf{x} \in \mathbb{R}^p$. tem-se:

$$\mathbf{x}^t \mathbf{A} \mathbf{x} \;=\; \mathbf{x}^t \left(\sum_{i=1}^{p} \lambda_i \mathbf{v}_i \mathbf{v}_i^t \right) \mathbf{x} \;=\; \sum_{i=1}^{p} \lambda_i (\mathbf{x}^t \mathbf{v}_i)(\mathbf{v}_i^t \mathbf{x}) \;=\; \sum_{i=1}^{p} \lambda_i (\mathbf{x}^t \mathbf{v}_i)^2 \; . \tag{1.23}$$

Ora, por definição (ver página 5), o vector nulo $\mathbf{x} = \mathbf{0}$ não é relevante para a classificação da matriz $\mathbf{A}$. Para vectores não-nulos, tem-se sempre $(\mathbf{x}^t \mathbf{v}_i)^2 \geq 0$ e para pelo menos um vector da base (isto é, pelo menos para um $\mathbf{v}_j$) tem de ter-se $(\mathbf{x}^t \mathbf{v}_j)^2 > 0$ (se $(\mathbf{x}^t \mathbf{v}_i)^2 = 0$ para todos os vectores da base de $\mathbb{R}^p$, então $\mathbf{x}$ é ortogonal a todos os vectores de $\mathbb{R}^p$, o que significa que está no complemento ortogonal de $\mathbb{R}^p$, ou seja no subespaço $\{\mathbf{0}\}$). Assim,

1. Se $\lambda_i > 0$, $\forall\, i$, então $\mathbf{x}^t \mathbf{A} \mathbf{x} = \sum_{i=1}^{p} \lambda_i (\mathbf{x}^t \mathbf{v}_i)^2 > 0$, $\forall \mathbf{x} \in \mathbb{R}^p \backslash \{\mathbf{0}\}$, pelo que $\mathbf{A}$ é definida positiva. Em sentido inverso, se $\mathbf{A}$ é definida positiva, tem-se $\mathbf{x}^t \mathbf{A} \mathbf{x} > 0$ para todos os vectores não nulos, e em particular, para $\mathbf{x} = \mathbf{v}_j$, $(\forall\, j)$. Logo $\mathbf{v}_j^t \mathbf{A} \mathbf{v}_j = \sum_{i=1}^{p} \lambda_i (\mathbf{v}_j^t \mathbf{v}_i)^2 = \lambda_j (\mathbf{v}_j^t \mathbf{v}_j)^2 = \lambda_j > 0$ $(\forall\, j)$.

2. Se $\lambda_j \geq 0$, $\forall\, j$, e $\lambda_j = 0$ para pelo menos um j, então $\mathbf{x}^t \mathbf{A} \mathbf{x} \geq 0$, $\forall \mathbf{x} \in \mathbb{R}^p \backslash \{\mathbf{0}\}$, mas existe pelo menos um vector não nulo (o vector próprio $\mathbf{v}_j$ associado a um valor próprio nulo λ_j) para o qual $\mathbf{v}_j^t \mathbf{A} \mathbf{v}_j = \lambda_j (\mathbf{v}_j^t \mathbf{v}_j)^2 = 0$. Assim, $\mathbf{A}$ é semi-definida positiva. Em sentido inverso, se $\mathbf{A}$ é semi-definida positiva, tem-se $\mathbf{x}^t \mathbf{A} \mathbf{x} \geq 0$ para todos os vectores não nulos, existindo algum $\mathbf{x} \neq \mathbf{0}$ tal que $\mathbf{x}^t \mathbf{A} \mathbf{x} = \sum_{i=1}^{p} \lambda_i (\mathbf{x}^t \mathbf{v}_i)^2 = 0$. Para que esta soma seja zero, das duas uma: ou existem parcelas com sinal diferente, ou todas as parcelas são nulas. A primeira possibilidade está excluída para matrizes semi-definidas positivas, uma vez que uma parcela ser negativa só pode ocorrer se o respectivo factor λ_j fosse negativo. Logo, todas as parcelas da soma têm de ser nulas. Mas uma parcela nula só pode ocorrer em duas circunstâncias: $\mathbf{x}^t \mathbf{v}_i = 0$ ou $\lambda_i = 0$. Nenhum vector não nulo $\mathbf{x}$ pode verificar a primeira condição para *todos* os vectores próprios $\{\mathbf{v}_i\}_{i=1}^p$. Logo, tem de haver pelo menos um valor próprio λ_i nulo.

3. Análogo à demonstração do primeiro ponto.

4. Análogo à demonstração do segundo ponto.

5. Se $\mathbf{A}$ tiver um valor próprio positivo, digamos λ_j e um valor próprio negativo, digamos λ_k, então para os respectivos vectores próprios $\mathbf{v}_j$ e $\mathbf{v}_k$ tem-se $\mathbf{v}_j^t \mathbf{A} \mathbf{v}_j = \lambda_j > 0$ e $\mathbf{v}_k^t \mathbf{A} \mathbf{v}_k = \lambda_k < 0$. Assim, $\mathbf{A}$ é indefinida. Partindo da hipótese de que $\mathbf{A}$ é indefinida, haverá pelo menos um vector $\mathbf{x}$ para o qual a forma quadrática $\mathbf{x}^t \mathbf{A} \mathbf{x}$ tem sinal positivo, e outro vector $\mathbf{y}$ para o qual $\mathbf{y}^t \mathbf{A} \mathbf{y}$ tem sinal negativo. Mas como se viu antes, se todos os valores próprios forem não-negativos, as formas quadráticas da equação (1.23) não podem tomar valor negativo, e se todos os valores próprios forem não positivos, as formas quadráticas não podem tomar valor positivo. Logo, havendo formas quadráticas com ambos os sinais, tem de haver pelo menos um valor próprios positivo e pelo menos um valor próprio negativo.

▽

O seguinte resultado é também de utilidade.

Teorema 1.34 *Seja $\mathbf{A}_{p\times p}$ uma matriz simétrica. Então:*

1. *A característica de $\mathbf{A}$, $k = car(\mathbf{A})$, é dada pelo número de valores próprios não nulos de $\mathbf{A}$.*
2. *Os vectores próprios de $\mathbf{A}$ associados aos valores próprios não nulos formam uma base ortonormada do subespaço imagem $\mathcal{C}(\mathbf{A})$.*

Demonstração: Considere a Decomposição Espectral de $\mathbf{A}$ dada na sua forma vectorial (equação (1.22), página 30). Qualquer vector no subespaço imagem $\mathcal{C}(\mathbf{A})$ é da forma $\mathbf{z} = \mathbf{A}\mathbf{x} = \sum_{i=1}^{p} \lambda_i \mathbf{v}_i \mathbf{v}_i^t \mathbf{x}$. Seja k o número de valores próprios não nulos de $\mathbf{A}$. Admita-se, sem perda de generalidade, que esses valores próprios não-nulos têm índices $i = 1 : k$. Então, tem-se $\mathbf{z} = \sum_{i=1}^{k} \mathbf{v}_i \, (\lambda_i \mathbf{v}_i^t \mathbf{x})$. Assim, qualquer vector do subespaço imagem $\mathcal{C}(\mathbf{A})$ se pode escrever como combinação linear dos k vectores próprios associados aos valores próprios não nulos, o que significa que esses vectores próprios são um *conjunto gerador* de $\mathcal{C}(\mathbf{A})$. Uma vez que os vectores próprios formam um conjunto ortogonal (e até ortonormado), são um conjunto linearmente independente (ver o Teorema 1.15, na página 21), pelo que se trata duma *base* de $\mathcal{C}(\mathbf{A})$ que, por conseguinte, tem dimensão k. ▽

1.4.3 Potências de matrizes simétricas

No caso de matrizes simétricas, tem-se:

$$\mathbf{A}^2 = \mathbf{A}\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^t \cdot \mathbf{V}\mathbf{\Lambda}\mathbf{V}^t = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^t,$$

onde $\mathbf{\Lambda}^2$ é a matriz diagonal cujo i-ésimo elemento é λ_i^2, sendo λ_i o i-ésimo elemento de $\mathbf{\Lambda}$. Em geral, para $k \in \mathbb{N}$:

$$\mathbf{A}^k = \mathbf{V}\mathbf{\Lambda}^k\mathbf{V}^t \quad \text{se} \quad \mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^t, \tag{1.24}$$

onde $\mathbf{\Lambda}^k$ é a matriz diagonal cujo i-ésimo elemento diagonal é λ_i^k.

A fórmula (1.24) tem ainda uma extensão natural para a potência -1, no caso de não existirem valores próprios nulos duma matriz simétrica $\mathbf{A}$. De facto, faz sentido escrever

$$\mathbf{A}^{-1} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^t$$

onde $\mathbf{\Lambda}^{-1}$ é a matriz diagonal dos recíprocos dos elementos diagonais de $\mathbf{\Lambda}$, que é a matriz inversa de $\mathbf{\Lambda}$, uma vez que esta fórmula nos dá a inversa da matriz $\mathbf{A}$ (que existe sempre para matrizes simétricas sem valores próprios nulos): $\mathbf{A}\mathbf{A}^{-1} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^t\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^t = \mathbf{V}\mathbf{\Lambda}\mathbf{\Lambda}^{-1}\mathbf{V}^t = \mathbf{V}\mathbf{V}^t = \mathbf{I}$ (verifique cada passagem!).

Esta notação é facilmente extensível a uma potência nula. De facto, e seguindo a convenção dos números reais de que uma potência com expoente nulo (e base não-nula) vale 1, podemos definir a potência nula duma matriz diagonal[5] como sendo a matriz identidade: $\mathbf{\Lambda}^0 = \mathbf{I}$. A partir desta definição, a potência nula

[5] Nesta definição, deve considerar-se a potência como aplicando-se apenas aos elementos diagonais da matriz.

duma matriz simétrica genérica vem também igual à identidade, de novo recorrendo à fórmula (1.24): $\mathbf{A}^0 = \mathbf{V}\mathbf{\Lambda}^0\mathbf{V}^t = \mathbf{V}\mathbf{I}\mathbf{V}^t = \mathbf{V}\mathbf{V}^t = \mathbf{I}$.

Igualmente, a não haver valores próprios negativos, a fórmula (1.24) pode abranger o caso de qualquer potência real k, sendo $\mathbf{\Lambda}^k$ a matriz diagonal dos λ_i^k. Assim, **para matrizes A definidas positivas, fica definida de forma única qualquer potência $\mathbf{A}^k$ ($k \in \mathbb{R}$) através da fórmula (1.24)**.

Nota: Note-se a relação imediata entre os valores/vectores próprios de $\mathbf{A}$ e os de $\mathbf{A}^k$: os vectores próprios são idênticos, enquanto que os valores próprios de $\mathbf{A}^k$ são dados pelas k-ésimas potências dos valores próprios de $\mathbf{A}$.

Estas definições dão-nos a possibilidade de trabalhar com potências de matrizes definidas positivas utilizando as mesmas regras algébricas que utilizamos para lidar com potências de números reais positivos. Por exemplo:

$$\mathbf{A}^k\mathbf{A}^m = \mathbf{A}^{k+m} \tag{1.25}$$

uma vez que $\mathbf{A}^k\mathbf{A}^m = \mathbf{V}\mathbf{\Lambda}^k\mathbf{V}^t\mathbf{V}\mathbf{\Lambda}^m\mathbf{V}^t = \mathbf{V}\mathbf{\Lambda}^{k+m}\mathbf{V}^t$, já que para as matrizes diagonais $\mathbf{\Lambda}$, os produtos de potências $\mathbf{\Lambda}^k\mathbf{\Lambda}^m$ obtêm-se mutliplicando os correspondentes elementos diagonais, obtendo-se a matriz diagonal $\mathbf{\Lambda}^{k+m}$ (confirme!).

Nota: Repare-se nas analogias com as regras para a definição de potências de números reais. No caso de a matriz $\mathbf{A}$ ser definida positiva, k, m podem ser quaisquer números reais (tomando-se $\mathbf{A}^0 = \mathbf{I}$). Se $\mathbf{A}$ fôr apenas semi-definida positiva, mas não definida positiva, as potências k, m não podem ser negativas, uma vez que λ_i^k não estará definida se $\lambda_i = 0$ e $k < 0$. Se $\mathbf{A}$ fôr apenas simétrica, mas não semi-definida positiva, k, m terão de ser números inteiros, uma vez que λ_i^k pode não estar definida para outros k, se $\lambda_i < 0$.

1.4.4 Mais resultados sobre matrizes

Vejamos agora mais resultados relativos a matrizes, que serão necessários para a compreensão das técnicas de análise multivariada que serão consideradas adiante.

Comecemos por um resultado que relaciona os valores e vectores próprios de uma matriz *simétrica* $\mathbf{A}$ com os valores da função $\frac{\mathbf{x}^t\mathbf{A}\mathbf{x}}{\mathbf{x}^t\mathbf{x}}$ ou, o que é equivalente, com os valores da forma quadrática $\mathbf{x}^t\mathbf{A}\mathbf{x}$ para vectores $\mathbf{x}$ de norma 1.

Teorema 1.35 ***(Teorema de Rayleigh-Ritz)*** *Seja $\mathbf{A}_{p\times p}$ uma matriz simétrica, cujos valores próprios são indexados por ordem descendente: $\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{p-1} \geq \lambda_p = \lambda_{\min}$.*

1. *O maior valor próprio de $\mathbf{A}$ é:*

$$\lambda_{\max} = \max_{\mathbf{x}\neq 0} \frac{\mathbf{x}^t\mathbf{A}\mathbf{x}}{\mathbf{x}^t\mathbf{x}}$$

O quociente $\frac{\mathbf{x}^t\mathbf{A}\mathbf{x}}{\mathbf{x}^t\mathbf{x}}$ toma esse valor quando $\mathbf{x} = \mathbf{v}_1$, o vector próprio associado ao maior valor próprio $\lambda_1 = \lambda_{\max}$.

2. *O menor valor próprio de* $\mathbf{A}$ *é:*

$$\lambda_{\min} = \min_{\mathbf{x}\neq 0} \frac{\mathbf{x}^t\mathbf{A}\mathbf{x}}{\mathbf{x}^t\mathbf{x}}$$

O quociente $\frac{\mathbf{x}^t\mathbf{A}\mathbf{x}}{\mathbf{x}^t\mathbf{x}}$ *toma esse valor quando* $\mathbf{x} = \mathbf{v}_p$*, o vector próprio associado ao menor valor próprio* $\lambda_p = \lambda_{\min}$.

3. *Os restantes valores/vectores próprios de* $\mathbf{A}$ *também são caracterizáveis a partir do* **quociente de Rayleigh-Ritz** *da matriz* $\mathbf{A}$*: seja* $\mathbf{v}_i$ *o vector próprio de* $\mathbf{A}$ *associado ao valor próprio* λ_i*. Então:*

$$\lambda_j = \max_{(\mathbf{x}\perp\mathbf{v}_1,\mathbf{v}_2,\ldots\mathbf{v}_{j-1})\wedge(\mathbf{x}\neq 0)} \frac{\mathbf{x}^t\mathbf{A}\mathbf{x}}{\mathbf{x}^t\mathbf{x}}$$

$$\lambda_j = \min_{(\mathbf{x}\perp\mathbf{v}_{j+1},\mathbf{v}_{j+2},\ldots\mathbf{v}_p)\wedge(\mathbf{x}\neq 0)} \frac{\mathbf{x}^t\mathbf{A}\mathbf{x}}{\mathbf{x}^t\mathbf{x}}$$

(i.e., os λ_j*'s são os sucessivos máximos - ou sucessivos mínimos - do quociente, sujeitos à exigência de que os vectores* $\mathbf{x}$ *considerados sejam ortogonais aos vectores próprios já determinados), verificando-se as igualdades quando* $\mathbf{x} = \mathbf{v}_j$.

Demonstração. Seja $\mathbf{A} \in \mathbb{S}_{p\times p}$, isto é, uma matriz simétrica de dimensão $p \times p$. Então, pelo Teorema da Decomposição Espectral (Teorema 1.32, p. 30) pode escrever-se $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^t = \sum_{i=1}^{p} \lambda_i \mathbf{v}_i \mathbf{v}_i^t$, onde $\mathbf{V}$ é uma matriz ortogonal $p \times p$, cujas colunas são vectores próprios $\{\mathbf{v}_i\}_{i=1}^p$ e $\mathbf{\Lambda}$ uma matriz diagonal, cujos elementos diagonais são os valores próprios $\{\lambda_i\}_{i=1}^p$. Nesse caso, para qualquer vector não-nulo de $\mathbb{R}^p$, $\mathbf{x}$, tem-se:

$$\mathbf{x}^t\mathbf{A}\mathbf{x} = \mathbf{x}^t\mathbf{V}\mathbf{\Lambda}\mathbf{V}^t\mathbf{x} = \mathbf{x}^t\left(\sum_{i=1}^{p}\lambda_i\mathbf{v}_i\mathbf{v}_i^t\right)\mathbf{x} = \sum_{i=1}^{p}\lambda_i(\mathbf{v}_i^t\mathbf{x})^2.$$

Então:

- Sendo $\lambda_{\max} = \lambda_1$ o maior dos valores próprios de $\mathbf{A}$, tem-se $\lambda_i \leq \lambda_{\max}$, para qualquer $i = 1, 2, ..., p$. Uma vez que $(\mathbf{v}_i^t\mathbf{x})^2 \geq 0$, tem-se $\lambda_i(\mathbf{v}_i^t\mathbf{x})^2 \leq \lambda_{\max}(\mathbf{v}_i^t\mathbf{x})^2$, $\forall i$, pelo que

$$\mathbf{x}^t\mathbf{A}\mathbf{x} = \sum_{i=1}^{p}\lambda_i(\mathbf{v}_i^t\mathbf{x})^2 \leq \lambda_{\max}\cdot\sum_{i=1}^{p}(\mathbf{v}_i^t\mathbf{x})^2 = \lambda_{\max}\cdot\mathbf{x}^t\mathbf{V}\mathbf{V}^t\mathbf{x} = \lambda_{\max}\cdot\mathbf{x}^t\mathbf{x}\ ,$$

  já que $\mathbf{V}$ é uma matriz ortogonal. Assim, $\frac{\mathbf{x}^t\mathbf{A}\mathbf{x}}{\mathbf{x}^t\mathbf{x}} \leq \lambda_{\max}$, $\forall \mathbf{x} \neq \mathbf{0}$. Falta apenas provar que existe um vector $\mathbf{x} \neq \mathbf{0}$ para o qual se verifica a igualdade. Mas tomando $\mathbf{x} = \mathbf{v}_1$ (o vector próprio associado ao maior valor próprio), o quociente $\frac{\mathbf{x}^t\mathbf{A}\mathbf{x}}{\mathbf{x}^t\mathbf{x}}$ toma o valor $\lambda_{\max}$, pelo que fica demonstrada a primeira alínea.

- Um raciocínio análogo, tomando o menor valor próprio e respectivo vector próprio, demonstra a segunda alínea.

- Consideremos os vectores $\mathbf{x} \in \mathbb{R}^p$ ortogonais a $\mathbf{v}_1$, o vector próprio a que corresponde o maior valor próprio de $\mathbf{A}$. Então, a forma quadrática $\mathbf{x}^t\mathbf{A}\mathbf{x}$ é igual a $\sum_{i=1}^{p}\lambda_i(\mathbf{v}_i^t\mathbf{x})^2$. Mas pela ortogonalidade de

$\mathbf{x}$ com $\mathbf{v}_1$, tem-se $\mathbf{x}^t\mathbf{A}\mathbf{x} = \sum_{i=2}^{p} \lambda_i(\mathbf{v}_i^t\mathbf{x})^2$. Um raciocínio análogo ao desenvolvido na demonstração da primeira alínea leva então à conclusão que, para vectores $\mathbf{x} \in \mathbb{R}^p$ ortogonais a $\mathbf{v}_1$, verifica-se

$$\mathbf{x}^t\mathbf{A}\mathbf{x} = \sum_{i=2}^{p} \lambda_i(\mathbf{v}_i^t\mathbf{x})^2 \leq \sum_{i=2}^{p} \lambda_2(\mathbf{v}_i^t\mathbf{x})^2 = \lambda_2 \sum_{i=2}^{p}(\mathbf{v}_i^t\mathbf{x})^2 = \lambda_2 \sum_{i=1}^{p}(\mathbf{v}_i^t\mathbf{x})^2 ,$$

uma vez que a nova parcela do último somatório é nula. Logo, e por analogia com o caso anterior, $\mathbf{x}^t\mathbf{A}\mathbf{x} \leq \lambda_2\mathbf{x}^t\mathbf{x}$, ou seja, $\frac{\mathbf{x}^t\mathbf{A}\mathbf{x}}{\mathbf{x}^t\mathbf{x}} \leq \lambda_2$. A igualdade verifica-se se $\mathbf{x} = \mathbf{v}_2$, o vector próprio associado a λ_2. Repetindo o raciocínio, mas agora para vectores $\mathbf{x}$ que sejam ortogonais, quer a $\mathbf{v}_1$, quer a $\mathbf{v}_2$ (isto é, que pertençam ao complemento ortogonal do subespaço gerado pelos vectores $\{\mathbf{v}_1, \mathbf{v}_2\}$), tem-se que $\frac{\mathbf{x}^t\mathbf{A}\mathbf{x}}{\mathbf{x}^t\mathbf{x}} \leq \lambda_3$, tendo-se a igualdade se $\mathbf{x} = \mathbf{v}_3$, e assim sucessivamente. Por outro lado, repetindo sucessivamente um raciocínio análogo ao utilizado na demonstração da segunda alínea deste Teorema, obtêm-se as caracterizações sucessivas dos valores próprios como *mínimos* dos quocientes de Rayleigh-Ritz, sujeitos às restrições indicadas no enunciado. ∇

Vejamos agora que as matrizes (semi-)definidas positivas podem ser sempre decompostas num produto da forma $\mathbf{X}^t\mathbf{X}$.

Teorema 1.36 *Seja $\mathbf{A} \in \mathbb{M}_{p\times p}$, simétrica. Então:*

1. *$\mathbf{A}$ é definida positiva se e só se existir uma matriz $\mathbf{X} \in \mathbb{M}_{m\times p}$ de característica p (isto é, com colunas linearmente independentes) tal que $\mathbf{A} = \mathbf{X}^t\mathbf{X}$.*
2. *$\mathbf{A}$ é semi-definida positiva de característica $k < p$ se e só se existir uma matriz $\mathbf{X} \in \mathbb{M}_{m\times p}$ de característica k tal que $\mathbf{A} = \mathbf{X}^t\mathbf{X}$.*

Nota: Neste enunciado, o número m de linhas da matriz $\mathbf{X}$ pode ser qualquer número natural não inferior à característica da matriz $\mathbf{A}$.

Demonstração. Temos:

1. ($\Longleftarrow$) Seja $\mathbf{A} = \mathbf{X}^t\mathbf{X}$, com $\mathbf{X} \in \mathbb{M}_{m\times p}$ de característica p (ou seja, com colunas linearmente independentes). Por definição, $\mathbf{A}$ é definida positiva quando $\mathbf{c}^t\mathbf{A}\mathbf{c} > 0, \quad \forall \mathbf{c} \neq 0$. Ora, $\mathbf{c}^t\mathbf{A}\mathbf{c} = \mathbf{c}^t\mathbf{X}^t\mathbf{X}\mathbf{c} = \|\mathbf{X}\mathbf{c}\|^2 \geq 0$ e:
$$\begin{aligned} \mathbf{c}^t\mathbf{A}\mathbf{c} = 0 \Longleftrightarrow \|\mathbf{X}\mathbf{c}\| = 0 &\iff \mathbf{X}\mathbf{c} = \mathbf{0} && \text{(propriedades das normas)} \\ &\iff \mathbf{c} = \mathbf{0} && (\mathbf{X} \text{ tem colunas linearmente independentes}) \end{aligned}$$
Logo, $\mathbf{A}$ é definida positiva.

   ($\Longrightarrow$) Se $\mathbf{A}$ é definida positiva, os seus p valores próprios são *positivos* (Teorema 1.33, página 31). Seja $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^t$ a Decomposição Espectral (Teorema 1.32, pg. 30) de $\mathbf{A}$. Pode-se definir $\mathbf{\Lambda}^{1/2}$ como a matriz diagonal das raízes quadradas dos valores próprios de $\mathbf{A}$ (como se viu na Subsecção 1.4.3, pg. 33). Seja $\mathbf{Q}_{m\times p}$ uma qualquer matriz de m ($m \geq p$) colunas ortonormadas, isto é, tal que $\mathbf{Q}^t\mathbf{Q} = \mathbf{I}_p$. Então, $\mathbf{X} = \mathbf{Q}\mathbf{\Lambda}^{1/2}\mathbf{V}^t$ é uma matriz tal que $\mathbf{X}^t\mathbf{X} = \mathbf{V}\mathbf{\Lambda}^{1/2}\mathbf{Q}^t\mathbf{Q}\mathbf{\Lambda}^{1/2}\mathbf{V}^t = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^t = \mathbf{A}$. Falta apenas provar que a matriz $\mathbf{X}$ é uma matriz de

característica p. Ora, se $\mathbf{A}$ é definida positiva, tem todos os seus p valores próprios não-nulos, e pelo primeiro ponto do Teorema 1.34, tem-se $car(\mathbf{A}) = p$. Por outro lado, o Teorema 1.31 garante que $car(\mathbf{A}) = car(\mathbf{X}^t\mathbf{X}) = car(\mathbf{X})$. Logo $car(\mathbf{X}) = p$.

2. ($\Longleftarrow$) Seja $\mathbf{A} = \mathbf{X}^t\mathbf{X}$, com $\mathbf{X}$ de característica $k < p$. Ora, $\forall\, \mathbf{c} \in \mathbb{R}^p$, $\mathbf{c}^t\mathbf{A}\mathbf{c} = \mathbf{c}^t\mathbf{X}^t\mathbf{X}\mathbf{c} = \|\mathbf{X}\mathbf{c}\|^2 \geq 0$ (pelas propriedades das normas). Logo, $\mathbf{A}$ apenas pode ser ou definida positiva, ou semi-definida positiva. Por outro lado, pelo Teorema 1.31, a característica de $\mathbf{A} = \mathbf{X}^t\mathbf{X}$ é igual à característica de $\mathbf{X}$, que é $k < p$. Mas pelo Teorema 1.34, esse é o número de valores próprios não nulos de $\mathbf{A}$, pelo que $\mathbf{A}$ tem $p - k$ valores próprios nulos, sendo apenas semi-definida positiva.

   ($\Longrightarrow$) Em relação à parte análoga do ponto anterior, falta apenas provar que a matriz $\mathbf{X} = \mathbf{Q}\mathbf{\Lambda}^{1/2}\mathbf{V}^t$ é uma matriz de característica $k < p$. Ora, se $\mathbf{A}$ é de característica $k < p$, o resultado (1.20) garante que $car(\mathbf{A}) = car(\mathbf{X}^t\mathbf{X}) = car(\mathbf{X})$. Logo, $car(\mathbf{X}) = k$. ∇

Antes de encerrar a discussão dos valores próprios, vejamos alguns resultados relativos às matrizes da forma $\mathbf{X}^t\mathbf{X}$ e $\mathbf{X}\mathbf{X}^t$, que possuem características importantes comuns.

Teorema 1.37 *Considere uma matriz $\mathbf{X} \in \mathbb{M}_{n\times p}$. verifica-se:*

1. *Os valores próprios não-nulos das matrizes $\mathbf{X}\mathbf{X}^t$ e $\mathbf{X}^t\mathbf{X}$ coincidem.*
2. *Seja $\mathbf{v}_i$ um vector próprio unitário da matriz $\mathbf{X}\mathbf{X}^t$ associado ao i-ésimo maior valor próprio não-nulo, λ_i. Seja $\mathbf{u}_i$ um vector próprio unitário da matriz $\mathbf{X}^t\mathbf{X}$ associado ao mesmo valor próprio λ_i. Pode-se sempre escolher esses vectores de forma a que:*
   - $\mathbf{u}_i = \frac{1}{\sqrt{\lambda_i}}\mathbf{X}^t\mathbf{v}_i$
   - $\mathbf{v}_i = \frac{1}{\sqrt{\lambda_i}}\mathbf{X}\mathbf{u}_i$

Notas:

1. A matriz $\mathbf{X}\mathbf{X}^t$ é de tipo $n \times n$ e, sendo simétrica, terá n valores próprios, associados a um conjunto ortogonal de vectores próprios. Por outro lado, a matriz $\mathbf{X}^t\mathbf{X}$ é de tipo $p \times p$, pelo que terá p valores próprios, também eles associados a vectores próprios ortogonais. Como veremos, os valores próprios não-nulos dessas matrizes têm de ser iguais. Admita-se que existem r desses valores próprios (tendo-se, necessariamente, $r \leq \min\{n, p\}$). Os restantes $p - r$ valores próprios de $\mathbf{X}^t\mathbf{X}$ têm de ser nulos, bem como os restantes $n - r$ valores próprios de $\mathbf{X}\mathbf{X}^t$.

2. A expressão cautelosa *"pode-se sempre escolher esses vectores próprios de forma a que"* é necessária porque, como se sabe, a escolha do conjunto de vectores próprios duma matriz simétrica não é única (mesmo admitindo que se trata de vectores próprios de norma 1), uma vez que todos os vectores próprios apenas estão definidos a menos de um sinal. Assim, seria sempre possível ter-se, por exemplo, $\mathbf{u}_i = -\frac{\mathbf{X}^t\mathbf{v}_i}{\sqrt{\lambda_i}}$. Além disso, caso haja valores próprios repetidos, os correspondentes vectores próprios podem ser escolhidos como sendo qualquer base ortonormada do subespaço por eles

gerado. Caso as matrizes $\mathbf{X}\mathbf{X}^t$ e $\mathbf{X}^t\mathbf{X}$ tenham todos os seus valores próprios não-nulos diferentes, poderia afirmar-se que a relação entre os respectivos vectores próprios é da forma:

- $\mathbf{u}_i = \pm\frac{1}{\sqrt{\lambda_i}}\mathbf{X}^t\mathbf{v}_i$
- $\mathbf{v}_i = \pm\frac{1}{\sqrt{\lambda_i}}\mathbf{X}\mathbf{u}_i$

Demonstração. Seja $\lambda_i,\ i = 1 : r$ um valor próprio não-nulo da matriz $\mathbf{X}\mathbf{X}^t$, associado ao vector próprio $\mathbf{v}_i$. Então, tem-se $\mathbf{X}\mathbf{X}^t\mathbf{v}_i = \lambda_i\mathbf{v}_i$. Multiplicando a equação à esquerda por $\mathbf{X}^t$, resulta a equação $\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{v}_i) = \lambda_i(\mathbf{X}^t\mathbf{v}_i)$, donde sai que λ_i é valor próprio de $\mathbf{X}^t\mathbf{X}$, associado ao vector próprio $\mathbf{X}^t\mathbf{v}_i$. Assim, todo o valor próprio não-nulo de $\mathbf{X}\mathbf{X}^t$ é também um valor próprio de $\mathbf{X}^t\mathbf{X}$. Analogamente, seja $\delta_i,\ i = 1 : s$ um valor próprio não-nulo de $\mathbf{X}^t\mathbf{X}$, associado ao vector próprio $\mathbf{u}_i$, tem-se $\mathbf{X}^t\mathbf{X}\mathbf{u}_i = \delta_i\mathbf{u}_i$. Multiplicando à esquerda por $\mathbf{X}$, vem $\mathbf{X}\mathbf{X}^t(\mathbf{X}\mathbf{u}_i) = \delta_i(\mathbf{X}\mathbf{u}_i)$, pelo que δ_i também é valor próprio de $\mathbf{X}\mathbf{X}^t$, associado ao vector próprio $\mathbf{X}\mathbf{u}_i$. Assim, todo o valor próprio não-nulo de $\mathbf{X}^t\mathbf{X}$ é também valor próprio de $\mathbf{X}\mathbf{X}^t$. Em conjunto com a conclusão anterior, tal significa que os valores próprios não-nulos das duas matrizes são iguais (tendo-se $r = s$). Além disso, os vectores próprios $\{\mathbf{X}^t\mathbf{v}_i\}_{i=1}^r$ formam um conjunto ortogonal, uma vez que $(\mathbf{X}^t\mathbf{v}_i)^t(\mathbf{X}^t\mathbf{v}_j) = \mathbf{v}_i^t\mathbf{X}\mathbf{X}^t\mathbf{v}_j = \lambda_j\mathbf{v}_i^t\mathbf{v}_j = 0$ se $i \neq j$, pois os vectores $\{\mathbf{v}_i\}_{i=1}^n$ são ortogonais. Para que os vectores $\{\mathbf{X}^t\mathbf{v}_i\}_{i=1}^r$ sejam vectores unitários, basta dividir esses vectores pela sua norma $\|\mathbf{X}^t\mathbf{v}_i\| = \sqrt{\mathbf{v}_i^t\mathbf{X}\mathbf{X}^t\mathbf{v}_i} = \sqrt{\lambda_i\mathbf{v}_i^t\mathbf{v}_i} = \sqrt{\lambda_i}$. Assim, os vectores $\{\frac{\mathbf{X}^t\mathbf{v}_i}{\sqrt{\lambda_i}}\}_{i=1}^r$ podem ser escolhidos como o conjunto ortonormado de vectores próprios $\{\mathbf{u}_i\}_{i=1}^r$ associados aos valores próprios não-nulos de $\mathbf{X}^t\mathbf{X}$. Um raciocínio análogo leva a demonstrar que se pode escrever $\mathbf{v}_i = \frac{\mathbf{X}\mathbf{u}_i}{\sqrt{\lambda_i}}$.

Nos métodos de estatística multivariada que serão estudados posteriormente, surge por vezes a necessidade de maximizar um quociente de formas quadráticas, $\frac{\mathbf{x}^t\mathbf{A}\mathbf{x}}{\mathbf{x}^t\mathbf{B}\mathbf{x}}$, onde $\mathbf{A}$ é uma matriz simétrica e $\mathbf{B}$ é uma matriz definida positiva. Os resultados seguintes dão-nos respostas a esta questão.

Teorema 1.38 *Seja $\mathbf{A}_{p\times p}$ uma matriz simétrica, e $\mathbf{B}_{p\times p}$ uma matriz definida positiva, com Decomposição Espectral $\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^t$. Então,*

1. *As soluções da equação*
$$\mathbf{A}\mathbf{x} \ = \ \lambda\mathbf{B}\mathbf{x}\ , \qquad (\mathbf{x} \in \mathbb{R}^p,\ \lambda \in \mathbb{R}) \tag{1.26}$$
são dadas pelos valores e vectores próprios da matriz $\mathbf{B}^{-1}\mathbf{A}$, existindo p pares $\{(\lambda_i, \mathbf{x}_i)\}_{i=1}^p$ distintos de soluções.

2. *Os valores próprios de $\mathbf{B}^{-1}\mathbf{A}$ são também valores próprios da matriz $\mathbf{C} = \mathbf{\Lambda}^{-1/2}\mathbf{V}^t\mathbf{A}\mathbf{V}\mathbf{\Lambda}^{-1/2}$.*

3. *Os p vectores próprios de $\mathbf{B}^{-1}\mathbf{A}$ não são ortogonais entre si, mas sim $\mathbf{B}$-ortogonais, isto é, $\mathbf{x}_i^t\mathbf{B}\mathbf{x}_j = 0$, se $i \neq j$.*

4. *Os vectores próprios $\{\mathbf{x}_i\}_{i=1}^p$ de $\mathbf{B}^{-1}\mathbf{A}$ estão directamente relacionados com os vectores próprios $\{\mathbf{y}_j\}_{j=1}^p$ da matriz $\mathbf{C} = \mathbf{\Lambda}^{-1/2}\mathbf{V}^t\mathbf{A}\mathbf{V}\mathbf{\Lambda}^{-1/2}$ através da relação $\mathbf{x} = \mathbf{V}\mathbf{\Lambda}^{-1/2}\mathbf{y}$. Os vectores próprios de $\mathbf{C}$ são ortogonais entre si.*

5. *A maximização do quociente*

$$\frac{\mathbf{x}^t\mathbf{A}\mathbf{x}}{\mathbf{x}^t\mathbf{B}\mathbf{x}} \tag{1.27}$$

está associada ao par $(\lambda_1, \mathbf{x}_1)$*, onde* λ_1 *é o maior valor próprio de* $\mathbf{B}^{-1}\mathbf{A}$ *e* $\mathbf{x}_1$ *é o correspondente vector próprio.*

6. *Os sucessivos pares de valores e vectores próprios de* $\mathbf{B}^{-1}\mathbf{A}$ *estão associados a sucessivos máximos do quociente* $\frac{\mathbf{x}^t\mathbf{A}\mathbf{x}}{\mathbf{x}^t\mathbf{B}\mathbf{x}}$*, sujeitos à exigência de que os vectores* $\mathbf{x}_i$ *considerados sejam* $\mathbf{B}$*-ortogonais às soluções* $\mathbf{x}_j$*,* $j = 1 : (i-1)$*, anteriormente obtidas.*

Demonstração.

Alíneas (1)-(4) Sendo a matriz $\mathbf{B}_{p\times p}$ definida positiva, os seus p valores próprios são não-nulos (são estritamente positivos) logo $\mathbf{B}$ é invertível. Assim,

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x} \iff \mathbf{B}^{-1}\mathbf{A}\mathbf{x} = \lambda\mathbf{x},$$

e é evidente que as soluções da equação $\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}$ são dadas por valores e vectores próprios da matriz $\mathbf{B}^{-1}\mathbf{A}$. No entanto, a matriz $\mathbf{B}^{-1}\mathbf{A}$ não é, em geral, simétrica (embora $\mathbf{A}$ e $\mathbf{B}^{-1}$ o sejam, o produto de matrizes simétricas não é, em geral, simétrica - veja-se o Exercíco 16, pg. 53). Para completar a demonstração da primeira afirmação, falta mostrar que existe um conjunto de precisamente p soluções distintas da equação $\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}$. Esta demonstração será agora feita, em simultâneo com as demonstrações dos pontos (2) a (4).

Tendo em conta a decomposição espectral da matriz $\mathbf{B}$ dada no enunciado (veja-se o Teorema 1.32, p. 30) tem-se,

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x} \iff \mathbf{A}\mathbf{x} = \lambda\mathbf{V}\mathbf{\Lambda}\mathbf{V}^t\mathbf{x}.$$

Ora, a matriz $\mathbf{V}$ é ortogonal, logo $\mathbf{V}^t$ é a sua inversa. Por outro lado, a matriz $\mathbf{\Lambda}$ é uma matriz diagonal que apenas tem elementos estritamente positivos na diagonal (pois $\mathbf{B}$ é definida positiva). Logo, é possível definir a sua raíz quadrada $\mathbf{\Lambda}^{1/2}$ (ver a secção 1.4.3), bem como a respectiva inversa $\mathbf{\Lambda}^{-1/2}$ (dada por uma matriz diagonal cujos elementos diagonais são os recíprocos das raízes quadradas dos elementos diagonais de $\mathbf{\Lambda}$). Defina-se então, para cada vector próprio $\mathbf{x}_i$ de $\mathbf{B}^{-1}\mathbf{A}$, um novo vector $\mathbf{y}_i = \mathbf{\Lambda}^{1/2}\mathbf{V}^t\mathbf{x}_i$. Dada a invertibilidade das matrizes, tem-se

$$\mathbf{y} = \mathbf{\Lambda}^{1/2}\mathbf{V}^t\mathbf{x} \iff \mathbf{x} = \mathbf{V}\mathbf{\Lambda}^{-1/2}\mathbf{y}. \tag{1.28}$$

Tem-se então

$$\begin{aligned}
& & \mathbf{B}^{-1}\mathbf{A}\mathbf{x} &= \lambda\mathbf{x} \\
\iff & \left(\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^t\right)\mathbf{A}\left(\mathbf{V}\mathbf{\Lambda}^{-1/2}\mathbf{y}\right) &= \lambda\mathbf{V}\mathbf{\Lambda}^{-1/2}\mathbf{y} \\
\iff & \mathbf{\Lambda}^{1/2}\mathbf{V}^t\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^t\mathbf{A}\mathbf{V}\mathbf{\Lambda}^{-1/2}\mathbf{y} &= \lambda\mathbf{y} \\
\iff & \mathbf{\Lambda}^{-1/2}\mathbf{V}^t\mathbf{A}\mathbf{V}\mathbf{\Lambda}^{-1/2}\mathbf{y} &= \lambda\mathbf{y}.
\end{aligned}$$

Assim, os valores próprios λ da matriz $\mathbf{B}^{-1}\mathbf{A}$ são também valores próprios, associados aos vectores próprios $\mathbf{y}$, da matriz $\mathbf{C} = \mathbf{\Lambda}^{-1/2}\mathbf{V}^t\mathbf{A}\mathbf{V}\mathbf{\Lambda}^{-1/2}$. Mas como esta matriz $\mathbf{C}$ é simétrica (confirme!),

podemos então garantir que existe um conjunto de p valores próprios $\{\lambda_i\}_{i=1}^p$, associados a p diferentes vectores próprios **ortogonais entre si**, $\{\mathbf{y}_i\}_{i=1}^p$. Estão assim provadas as afirmações dos pontos (1), (2) e (4). Quanto à $\mathbf{B}$-ortogonalidade dos vectores próprios $\mathbf{x}_i$ da matriz $\mathbf{B}^{-1}\mathbf{A}$, resulta de considerar, para $i \neq j$,

$$\mathbf{x}_i^t\mathbf{B}\mathbf{x}_j \;=\; \mathbf{x}_i^t\mathbf{V}\mathbf{\Lambda}\mathbf{V}^t\mathbf{x}_j \;=\; \mathbf{y}_i^t\mathbf{y}_j \;=\; 0,$$

utilizando a relação entre os vectores $\mathbf{x}$ e os vectores $\mathbf{y}$, e sabendo que os vectores $\mathbf{y}_i$ são ortogonais entre si.

Alínea (5) Determinar o vector $\mathbf{x}$ que maximiza o quociente $\frac{\mathbf{x}^t\mathbf{A}\mathbf{x}}{\mathbf{x}^t\mathbf{B}\mathbf{x}}$ equivale a determinar o vector $\mathbf{y}$ que maximiza o quociente equivalente

$$\frac{\mathbf{x}^t\mathbf{A}\mathbf{x}}{\mathbf{x}^t\mathbf{B}\mathbf{x}} \;=\; \frac{\mathbf{y}^t\left(\mathbf{\Lambda}^{-1/2}\mathbf{V}^t\mathbf{A}\mathbf{V}\mathbf{\Lambda}^{-1/2}\right)\mathbf{y}}{\mathbf{y}^t\mathbf{y}} \;=\; \frac{\mathbf{y}^t\mathbf{C}\mathbf{y}}{\mathbf{y}^t\mathbf{y}}. \tag{1.29}$$

Mas este novo quociente sabemos ser maximizado pelo vector próprio $\mathbf{y}_1$, associado ao maior valor proprio λ_1 da matriz $\mathbf{C}$ (ver o Teorema de Rayleigh-Ritz, na página 34). Assim, o correspondente vector $\mathbf{x}_1 = \mathbf{V}\mathbf{\Lambda}^{-1/2}\mathbf{y}_1$ maximiza o quociente inicial $\frac{\mathbf{x}^t\mathbf{A}\mathbf{x}}{\mathbf{x}^t\mathbf{B}\mathbf{x}}$, atribuindo-lhe o mesmo valor máximo λ_1.

Alínea (6) Sucessivos pares de valores e vectores próprios de $\mathbf{C}$ fornecem os sucessivos máximos do quociente $\frac{\mathbf{y}^t\mathbf{C}\mathbf{y}}{\mathbf{y}^t\mathbf{y}}$, sujeitos à ortogonalidade de sucessivas soluções $\mathbf{y}$ (Teorema 1.35). Já sabemos que, em termos dos vectores próprios $\mathbf{x}$ da matriz $\mathbf{B}^{-1}\mathbf{A}$, biunivocamente associados aos vectores próprios $\mathbf{y}$ de $\mathbf{C}$, esta exigência traduz-se na $\mathbf{B}$-ortogonalidade de diferentes $\mathbf{x}_i$.

∇

1.5 A Decomposição em Valores Singulares

Vejamos agora um dos mais importantes resultados em Teoria de Matrizes, a Decomposição em Valores Singulares duma matriz genérica. Este resultado permite factorizar qualquer matriz, mesmo uma matriz rectangular, de forma simultaneamente simples e poderosa. Como se verá seguidamente, as componentes desta factorização (valores e vectores singulares) desempenham um papel importante nas aplicações lineares definidas pela matriz que se decompõe e pela sua transposta. A Decomposição em Valores Singulares desempenha igualmente um papel crucial na Análise em Componentes Principais e outras técnicas de Estatística Multivariada.

Teorema 1.39 ***Decomposição em Valores Singulares (DVS).*** *Seja $\mathbf{X} \in \mathbb{M}_{n\times p}$ uma matriz de característica r. Então, pode-se escrever:*

$$\mathbf{X} \;=\; \mathbf{W}\mathbf{\Delta}\mathbf{V}^t \tag{1.30}$$

$$\iff \mathbf{X} \;=\; \sum_{i=1}^{r}\delta_i\mathbf{w}_i\mathbf{v}_i^t \tag{1.31}$$

onde

$\mathbf{\Delta}_{rxr}$ *é uma matriz diagonal com elementos diagonais positivos*

$\mathbf{V}_{pxr}$, $\mathbf{W}_{nxr}$ *são matrizes com colunas ortonormadas (isto e,* $\mathbf{V}^t\mathbf{V} = \mathbf{I}_r = \mathbf{W}^t\mathbf{W}$*)*

δ_i *são os elementos diagonais de* $\mathbf{\Delta}$ *(designados* ***valores singulares de*** $\mathbf{X}$*)*

$\mathbf{w}_i$ *são as colunas de* $\mathbf{W}$ *(designados* ***vectores singulares esquerdos de*** $\mathbf{X}$*)*

$\mathbf{v}_i$ *são as colunas de* $\mathbf{V}$ *(designados* ***vectores singulares direitos de*** $\mathbf{X}$*)*

Demonstração. A demonstração deste resultado será feita separando o caso mais simples, em que $\mathbf{X}$ é uma matriz com colunas linearmente independentes, isto é, de característica $r = p$ (Caso 1), do caso mais difícil em que as colunas de $\mathbf{X}$ não são linearmente independentes (Caso 2, com $r < p$).

Caso 1 Considere-se o produto $\mathbf{X}^t\mathbf{X}$ (que é uma matriz simétrica) e tome-se a sua Decomposição Espectral (Teorema 1.32, p.30):

$$\mathbf{X}^t\mathbf{X} \;=\; \mathbf{U}\Lambda\mathbf{U}^t.$$

Então, tomando

- $\mathbf{V} = \mathbf{U}$ (note-se que $\mathbf{U}$ é uma matriz ortogonal, isto é, $\mathbf{U}^t\mathbf{U} = \mathbf{U}\mathbf{U}^t = \mathbf{I}_p$),
- $\mathbf{\Delta} = \Lambda^{1/2}$, e
- $\mathbf{W} = \mathbf{X}\mathbf{U}\Lambda^{-1/2}$,

temos o resultado pretendido. De facto, ter-se-á $\mathbf{W}\mathbf{\Delta}\mathbf{V}^t = \mathbf{X}\mathbf{U}\Lambda^{-1/2}\Lambda^{1/2}\mathbf{U}^t = \mathbf{X}\mathbf{U}\mathbf{U}^t = \mathbf{X}$, uma vez que $\mathbf{U}$ é uma matriz ortogonal. Além disso, $\mathbf{\Delta}$ é, por definição, uma matriz diagonal com elementos diagonais positivos e $\mathbf{W}$ tem colunas ortonormadas, uma vez que $\mathbf{W}^t\mathbf{W} = \Lambda^{-1/2}\mathbf{U}^t\mathbf{X}^t\mathbf{X}\mathbf{U}\Lambda^{-1/2} = \Lambda^{-1/2}\mathbf{U}^t\mathbf{U}\Lambda\mathbf{U}^t\mathbf{U}\Lambda^{-1/2} = \mathbf{I}_p$, usando a decomposição espectral de $\mathbf{X}^t\mathbf{X}$ referida acima. Assim, a decomposição indicada no enunciado é possível.

Caso 2 Considere-se de novo a decomposição espectral da matriz (de tipo $p \times p$) $\mathbf{X}^t\mathbf{X}$, dada por $\mathbf{X}^t\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^t = \sum_{i=1}^{p} \lambda_i \mathbf{u}_i\mathbf{u}_i^t$. Uma vez que $\mathbf{X}$ é de característica $r < p$, e que a característica de $\mathbf{X}$ e de $\mathbf{X}^t\mathbf{X}$ coincidem (veja-se a equação (1.20), na página 30), apenas os r primeiros valores próprios de $\mathbf{X}^t\mathbf{X}$ são não-nulos, pelo que $\mathbf{X}^t\mathbf{X} = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i\mathbf{u}_i^t$. Do ponto de vista matricial, esta decomposição pode escrever-se na forma:

$$\mathbf{X}^t\mathbf{X} = \mathbf{U}_r\mathbf{\Lambda}_r\mathbf{U}_r^t,$$

onde $\mathbf{\Lambda}_r$ é a matriz $r \times r$ resultante de reter apenas as r linhas/colunas de $\mathbf{\Lambda}$ associadas aos valores próprios não-nulos, e $\mathbf{U}_r$ é a matriz $n \times r$ resultante de reter as r colunas de $\mathbf{U}$ associadas a esses valores próprios (confirme!). Note-se que $\mathbf{U}_r^t\mathbf{U}_r = \mathbf{I}_r$, mas não se pode verificar $\mathbf{U}_r\mathbf{U}_r^t = \mathbf{I}_p$ (porquê?). Ora, tomando

- $\mathbf{V} = \mathbf{U}_r$,
- $\mathbf{\Delta} = \mathbf{\Lambda}_r^{1/2}$, e

- $\mathbf{W} = \mathbf{X}\mathbf{U}_r\mathbf{\Lambda}_r^{-1/2}$,

tem-se a decomposição indicada no enunciado. De facto, $\mathbf{W}\mathbf{\Delta}\mathbf{V}^t = \mathbf{X}\mathbf{U}_r\mathbf{\Lambda}_r^{-1/2}\mathbf{\Lambda}_r^{1/2}\mathbf{U}_r^t = \mathbf{X}\mathbf{U}_r\mathbf{U}_r^t$. Pretende-se provar que este produto é igual a $\mathbf{X}$, ou seja, que $\mathbf{X}\mathbf{U}_r\mathbf{U}_r^t = \mathbf{X}$. Ora, o produto $\mathbf{U}_r\mathbf{U}_r^t$ é uma matriz $p \times p$, que projecta ortogonalmente sobre o subespaço de $\mathbb{R}^p$ gerado pelas colunas de $\mathbf{U}_r$ (é da forma $\mathbf{B}(\mathbf{B}^t\mathbf{B})^{-1}\mathbf{B}^t$, com $\mathbf{B} = \mathbf{U}_r$). Mas as colunas de $\mathbf{U}_r$ são vectores próprios unitários da matriz (simétrica) $\mathbf{X}^t\mathbf{X}$ associados a valores próprios não nulos, pelo que o Teorema 1.34 (p.33) garante que formam uma base ortonormada do espaço-coluna $\mathcal{C}(\mathbf{X}^t\mathbf{X})$. Por outro lado, o Teorema 1.28 (p.27) garante que $\mathcal{C}(\mathbf{X}^t\mathbf{X}) \subset \mathcal{C}(\mathbf{X}^t)$. Mas como o resultado (1.20) (p.30) garante que $car(\mathbf{X}) = car(\mathbf{X}^t\mathbf{X})$, temos que os subespaços encaixados, mas de igual dimensão, $\mathcal{C}(\mathbf{X}^t\mathbf{X})$ e $\mathcal{C}(\mathbf{X}^t)$, têm de ser idênticos. Resumindo, as colunas da matriz $\mathbf{U}_r$ formam uma base ortonormada do espaço $\mathcal{C}(\mathbf{X}^t)$, e a matriz $\mathbf{U}_r\mathbf{U}_r^t$ é uma matriz de projecção ortogonal sobre esse subespaço. Então, a projecção de vectores em $\mathcal{C}(\mathbf{X}^t)$ (como são as próprias colunas da matriz $\mathbf{X}^t$) sobre esse próprio subespaço, deixa-as invariantes. Por outras palavras, tem-se:

$$\begin{aligned} & \mathbf{U}_r\mathbf{U}_r^t\mathbf{X}^t &=& \mathbf{X}^t \\ \Longleftrightarrow \quad & \mathbf{X}\mathbf{U}_r\mathbf{U}_r^t &=& \mathbf{X}, \end{aligned}$$

como se queria mostrar. ∇

Observações.

1. A matriz $\mathbf{V}$ é uma matriz cujas colunas são um conjunto ortonormado de vectores próprios de $\mathbf{X}^t\mathbf{X}$, associados a valores próprios não-nulos. A matriz $\mathbf{\Delta}$ é a matriz das raízes quadradas dos valores próprios não-nulos de $\mathbf{X}^t\mathbf{X}$ (que são iguais aos de $\mathbf{X}\mathbf{X}^t$). Por definição, e tendo em conta o Teorema 1.37, $\mathbf{W}$ é uma matriz análoga a $\mathbf{V}$, para $\mathbf{X}\mathbf{X}^t$. **No que se segue, admite-se que os valores singulares δ_i estão ordenados por ordem decrescente**.

2. **A Decomposição em Valores Singulares (DVS) é válida para *qualquer* matriz** e não apenas para matrizes quadradas e simétricas, como a decomposição espectral.

3. A DVS duma matriz (simétrica) definida positiva coincide com a sua Decomposição Espectral (Exercício 22, página 53).

4. A DVS de uma matriz é sempre possível, mas apenas é única se não houver valores singulares repetidos, e mesmo assim, a menos de uma multiplicação escalar por -1 dos seus vectores singulares (aos pares, isto é, uma multiplicação dum vector singular esquerdo por -1 tem de estar acompanhada por análoga multiplicação do correspondente vector singular direito). Esta questão resulta directamente da discussão sobre existência e unicidade da decomposição espectral das matrizes $\mathbf{X}\mathbf{X}^t$ e $\mathbf{X}^t\mathbf{X}$.

5. **Se X tem Decomposição em Valores Singulares dada pela equação (1.31), então a transposta de X tem Decomposição em Valores Singulares dada por $\mathbf{X}^t = \mathbf{V}\mathbf{\Delta}\mathbf{W}^t$.**

6. Seria possível (e é frequente na literatura) definir a DVS de forma ligeiramente diferente, tomando a matriz diagonal $\mathbf{\Delta}$ de dimensões $p \times p$ (mesmo que $car(\mathbf{X}) < p$), a matriz $\mathbf{V}$ igualmente de dimensões $p \times p$ e a matriz $\mathbf{W}$ de dimensões $n \times p$. No caso de $r = car(\mathbf{X}) < p$, os últimos $p - r$ elementos diagonais de $\mathbf{\Delta}$ serão então nulos, e as últimas $p - r$ colunas de $\mathbf{V}$ serão qualquer base ortonormal de $\mathcal{N}(\mathbf{X})$. A matriz $\mathbf{V}$ será, neste caso, uma matriz ortogonal. As últimas $p - r$ colunas de $\mathbf{W}$ terão de ser vectores unitários de $\mathbb{R}^n$ ortogonais ao subespaço $\mathcal{C}(\mathbf{X})$. Para as nossas aplicações, esta versão da DVS não apresenta qualquer vantagem, embora seja mais conforme à definição usual da Decomposição Espectral de matrizes simétricas, onde é hábito manter as dimensões $p \times p$ das matrizes $\mathbf{P}$ e $\mathbf{\Lambda}$, mesmo que haja valores próprios nulos.

7. Sabemos que qualquer matriz de tipo $n \times p$ define uma aplicação linear de $\mathbb{R}^p$ em $\mathbb{R}^n$. **Os vectores e valores singulares duma matriz X desempenham um papel importante nas aplicações lineares definidas pelas matrizes X e $\mathbf{X}^t$. De facto, verifica-se (ver o Exercício 23, página 54):**

$$\mathbf{X}\mathbf{v}_i = \delta_i \mathbf{w}_i \qquad \text{e} \qquad \mathbf{X}^t \mathbf{w}_i = \delta_i \mathbf{v}_i \tag{1.32}$$

onde $\mathbf{v}_i$, $\mathbf{w}_i$ e δ_i são vectores e valores singulares de X. Veja-se a Figura 1.5.

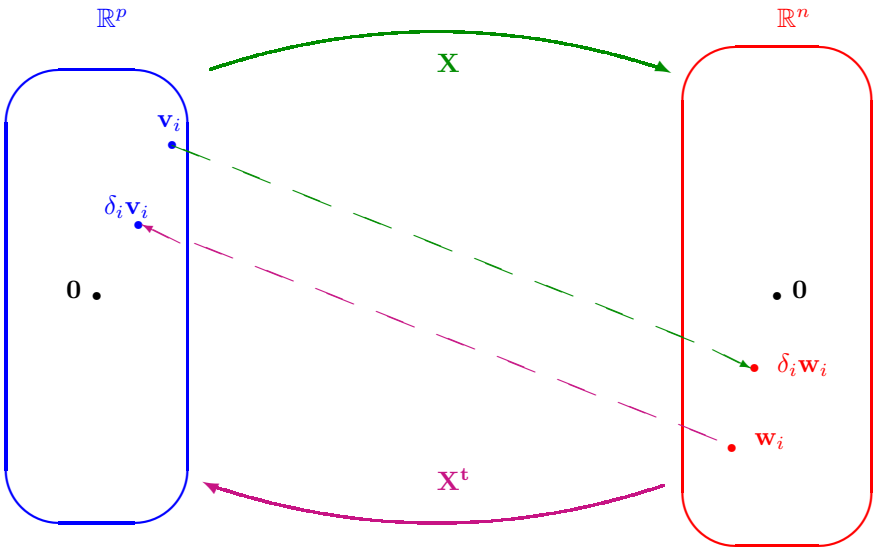


Figura 1.5: As aplicações definidas por uma matriz $\mathbf{X}_{n \times p}$ e a sua transposta. Os vectores $\mathbf{w}_i$ e $\mathbf{v}_i$ são vectores singulares de $\mathbf{X}$, associados aos valores singulares δ_i.

Teorema 1.40 *A norma duma matriz $\mathbf{X} \in \mathbb{M}_{n \times p}$ de característica r, dada na Definição 1.11 pode ser escrita em termos dos valores singulares de* $\mathbf{X}$*:*

$$\|\mathbf{X}\| = \sqrt{\sum_{i=1}^{r} \delta_i^2} = \|\underline{\delta}\|_2 \tag{1.33}$$

onde $\underline{\delta}$ designa o vector r-dimensional cujos elementos são os valores singulares de $\mathbf{X}$ e $\|\cdot\|_2$ a sua norma vectorial habitual.

Demonstração: Pela Definição 1.11, tem-se $\|\mathbf{X}\| = \sqrt{\text{tr}(\mathbf{X}^t\mathbf{X})}$. Seja $\mathbf{X} = \mathbf{W}\mathbf{\Delta}\mathbf{V}^t$ a Decomposição em Valores Singulares de $\mathbf{X}$. Então: $\|\mathbf{X}\| = \sqrt{\text{tr}(\mathbf{V}\mathbf{\Delta}\mathbf{W}^t\mathbf{W}\mathbf{\Delta}\mathbf{V}^t)} = \sqrt{\text{tr}(\mathbf{\Delta}^2\mathbf{V}^t\mathbf{V})}$ uma vez que as colunas de $\mathbf{W}$ são ortonormais e dada a circularidade do traço (página 5). Uma vez que as colunas de $\mathbf{V}$ também são ortonormais, tem-se: $\|\mathbf{X}\| = \sqrt{\text{tr}(\mathbf{\Delta}^2)} = \sqrt{\sum_{i=1}^{r}\delta_i^2}$, que é também a norma ℓ_2 do vector $\underline{\delta}$. ∇

Exercício 1.9 *Resolva os Exercícios 18, 22 e 23 (página 53).*

Seguem-se dois Teoremas importantes que ajudam a compreender a importância da DVS de uma matriz. A demonstração do segundo destes Teoremas é omitida.

Teorema 1.41 *Seja $\mathbf{X}_{n\times p}$ uma matriz de característica r, com Decomposição em Valores Singulares dada por $\mathbf{X} = \mathbf{W}\mathbf{\Delta}\mathbf{V}^t = \sum_{i=1}^{r}\delta_i\mathbf{w}_i\mathbf{v}_i^t$. Então as colunas $\{\mathbf{w}_i\}_{i=1}^r$ da matriz $\mathbf{W}$ formam uma base ortonormada do subespaço imagem da aplicação definida por $\mathbf{X}$, $\mathcal{C}(\mathbf{X}) \subset \mathbb{R}^n$, e as colunas $\{\mathbf{v}_i\}_{i=1}^r$ da matriz $\mathbf{V}$ formam uma base ortonormada do subespaço imagem da aplicação definida por $\mathbf{X}^t$, $\mathcal{C}(\mathbf{X}^t) \subset \mathbb{R}^p$.*

Demonstração. Como se viu acima, uma matriz $\mathbf{X}_{n\times p}$ define uma aplicação de $\mathbb{R}^p$ em $\mathbb{R}^n$. A matriz $\mathbf{X}$ ser de característica r significa que a dimensão do seu subespaço imagem $\mathcal{C}(\mathbf{X})$ é r. Para mostrar que as r colunas de $\mathbf{W}$ formam uma base do espaço $\mathcal{C}(\mathbf{X}) \subset \mathbb{R}^n$ basta provar que elas *geram* esse subespaço[6]. Provar que as colunas de $\mathbf{W}$ formam um conjunto gerador de $\mathcal{C}(\mathbf{X})$ equivale a provar que qualquer elemento desse conjunto imagem se pode escrever como combinação linear das colunas de $\mathbf{W}$. Os elementos $\mathbf{b} \in \mathcal{C}(\mathbf{X})$ são os elementos de $\mathbb{R}^n$ para os quais $\exists\ \mathbf{a} \in \mathbb{R}^p$ tal que $\mathbf{b} = \mathbf{X}\mathbf{a}$. Ora, $\mathbf{b} = \mathbf{X}\mathbf{a} = \mathbf{W}(\mathbf{\Delta}\mathbf{V}^t\mathbf{a})$, ou seja, $\mathbf{b}$ pode escrever-se como combinação linear das colunas de $\mathbf{W}$, sendo os coeficientes dessa combinação linear dados pelo vector $\mathbf{\Delta}\mathbf{V}^t\mathbf{a}$. Repetindo o raciocínio a partir da matriz transposta $\mathbf{X}^t = \mathbf{V}\mathbf{\Delta}\mathbf{W}^t$, obtemos o resultado análogo para as colunas da matriz $\mathbf{V}$. ∇

Teorema 1.42 *Seja $\mathbf{X}_{nxp}$ uma matriz de característica k. A matriz $\mathbf{Y}_{nxp}$ de característica $m < k$ que melhor aproxima $\mathbf{X}$, no sentido de minimizar a usual distância matricial $\|\mathbf{X}-\mathbf{Y}\| = \sqrt{\sum_i\sum_j (x_{ij}-y_{ij})^2}$, obtem-se da seguinte forma:*

- *Seja $\mathbf{X} = \mathbf{W}\mathbf{\Delta}\mathbf{V}^t$ a decomposição em valores singulares de $\mathbf{X}$.*
- *Sejam $\mathbf{W}_m$, $\mathbf{V}_m$, as matrizes constituídas pelas m colunas de $\mathbf{W}$ e $\mathbf{V}$, respectivamente, associadas aos maiores valores singulares. Seja $\mathbf{\Delta}_m$ a matriz diagonal de tipo $\text{m}\times\text{m}$ resultante de reter apenas as linhas e colunas de $\mathbf{\Delta}$ associadas com os m maiores valores singulares.*

[6] Como foi visto na disciplina de Complementos de Álgebra e Análise, qualquer conjunto gerador de um subespaço de dimensão r tem de ter pelo menos r vectores. Um conjunto gerador que tenha exactamente r vectores será necessariamente uma base desse subespaço (isto é, os vectores desse conjunto gerador terão de ser linearmente independentes caso sejam em igual número que a dimensão do subespaço).

- *Então* $\mathbf{Y} = \mathbf{W}_m \mathbf{\Delta}_m \mathbf{V}_m^t$.

Observações:

1. $\mathbf{Y} = \mathbf{W}_m \mathbf{\Delta}_m \mathbf{V}_m^t$ é uma DVS de $\mathbf{Y}$.
2. Usando a forma $\mathbf{X} = \sum_{i=1}^{k} \delta_i \mathbf{w}_i \mathbf{v}_i^t$ da DVS de $\mathbf{X}$, verificamos que **$\mathbf{Y}$ é a matriz que se obtem retendo apenas as m primeiras parcelas da DVS de $\mathbf{X}$.**

1.6 Primeiras Aplicações Estatísticas

1.6.1 As representações em $\mathbb{R}^p$ e em $\mathbb{R}^n$

Quando temos n observações de uma variável, podemos representá-las por um vector $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}^t = [x_1, x_2, x_3, ..., x_n]$. Em Estatística univariada ou bivariada, é habitual representar n observações de uma ou duas variáveis como n pontos sobre um eixo ou um plano definido por um par de eixos, eixos esses representativos da(s) variável(eis) observada(s). A esta representação chamaremos daqui em diante **representação em $\mathbb{R}^p$**. Os pontos aí representados correspondem a indivíduos. Mas é igualmente possível adoptar uma outra representação, no espaço $\mathbb{R}^n$, em que cada conjunto de n observações de uma variável é representada por um ponto/vector em $\mathbb{R}^n$ cujas coordenadas são as n observações. Esta representação, menos frequente quando se considerem apenas duas ou três variáveis, devido à óbvia perda de visibilidade que ela representa, é no entanto de grande utilidade quando se consideram várias variáveis. Na **representação em $\mathbb{R}^n$** cada ponto/vector aí representado corresponde a uma variável. Como veremos na secção seguinte, ela permite uma útil visualização geométrica de conceitos estatísticos.

1.6.2 Conceitos estatísticos em $\mathbb{R}^n$

Os indicadores estatísticos mais elementares têm interessantes significados geométricos quando se utiliza a representação dos dados no espaço dos indivíduos, *i.e.*, a representação em $\mathbb{R}^n$. Assim:

1. A **média** das n observações é o *coeficiente da projecção ortogonal do vector de observações* $\mathbf{x}$ *sobre o subespaço* $\mathcal{C}(\mathbf{1}_n)$ (onde ${\mathbf{1}_n}^t = [1\ 1\ 1\ ...1]$ é o vector dos n uns), *i.e.*, sobre a "bissectriz" do primeiro ortante de $\mathbb{R}^n$. De facto, a matriz de projecção ortogonal sobre esse subespaço é dada por:

$$\mathbf{P}_{\mathbf{1}_n} = \mathbf{1}_n ({\mathbf{1}_n}^t \mathbf{1}_n)^{-1} {\mathbf{1}_n}^t = \frac{1}{n} \mathbf{1}_n {\mathbf{1}_n}^t$$

   Logo, $\mathbf{P}_{\mathbf{1}_n} \mathbf{x} = \frac{1}{n} \mathbf{1}_n {\mathbf{1}_n}^t \mathbf{x} = \mathbf{1}_n \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) = \overline{x} \mathbf{1}_n$. (ver a Figura 1.6).

2. A **variável centrada em torno da sua média**, *i.e.*, o vector com componentes $x_i - \overline{x}$, é a *projecção ortogonal de* $\mathbf{x}$ *no subespaço* $\mathcal{C}(\mathbf{1}_n)^{\perp}$, *i.e.*, no complemento ortogonal do subespaço gerado pelo vector dos uns. Esse vector centrado é $\mathbf{x} - \overline{x}\mathbf{1}_n = (\mathbf{I} - \mathbf{P}_{\mathbf{1}_n})\mathbf{x}$, onde $\mathbf{I}$ é a matriz

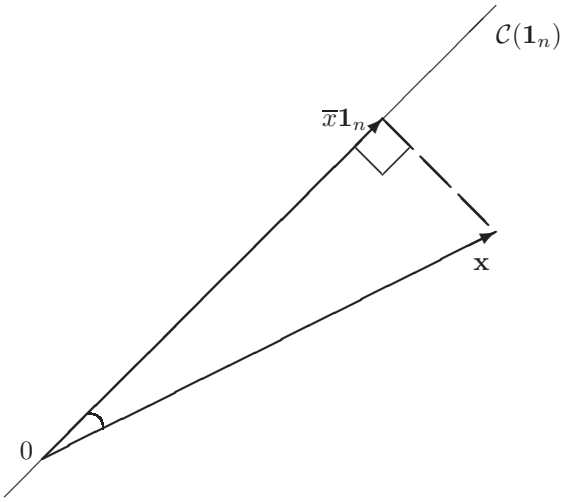


Figura 1.6: O significado geométrico de uma média no espaço de indivíduos.

identidade $n\times n$ que é a aplicação identidade em $\mathbb{R}^n$. O resto sai do Teorema 1.22 (página 23) relacionando os projectores $\mathbf{P}$ e $\mathbf{I}-\mathbf{P}$.

Note-se que é usual centrar as variáveis em torno da sua média em muitos indicadores estatísticos (variância, covariância, coeficiente de correlação). Essa centragem torna os resultados invariantes a translações da origem (*i.e.*, se $x_i \to x_i + a$, os valores $x_i - \overline{x}$ não sofrem alteração).

3. O **desvio padrão** das n observações é *proporcional à distância do vector* $\mathbf{x}$ *ao subespaço gerado pela coluna de uns,* $\mathcal{C}(\mathbf{1}_n)$. De facto, sabemos que a distância de um vector a um subespaço é dada pela menor distância entre o referido vector e qualquer vector do subespaço, e que essa menor distância atinge-se quando se considera a distância do vector à sua projecção ortogonal sobre o subespaço (Teorema 1.23, pg. 24). Assim, a distância em questão será o comprimento do vector centrado, $\mathbf{x}^c = (\mathbf{I} - \mathbf{P}_{\mathbf{1}_n})\mathbf{x}$, cujo quadrado é dado[7] por:

$$\|\mathbf{x}^c\|^2 = \|(\mathbf{I} - \mathbf{P}_{\mathbf{1}_n})\mathbf{x}\|^2 = \|\mathbf{x} - \overline{x}\mathbf{1}_n\|^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 = n \cdot \mathrm{var}(\mathbf{x}).$$

4. Por outro lado, tem-se, pelo Teorema de Pitágoras:

$$\begin{aligned} \mathrm{var}(\mathbf{x}) = \tfrac{1}{n}\|\mathbf{x} - \overline{x}\mathbf{1}_n\|^2 &= \tfrac{1}{n}\left(\|\mathbf{x}\|^2 - \|\overline{x}\mathbf{1}_n\|^2\right) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} x_i^2 - \overline{x}^2 \cdot \tfrac{1}{n}\|\mathbf{1}_n\|^2 \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} x_i^2 - \overline{x}^2 \end{aligned}$$

que é a *fórmula computacional da variância*. Assim, **a fórmula computacional da variância não é mais que uma aplicação do Teorema de Pitágoras (ver a Figura 1.7)**.

[7] Neste contexto descritivo definem-se variâncias e covariâncias de um conjunto de observações com denominador n.

5. O comprimento do vector $\mathbf{x}$ *não-centrado* é proporcional à raíz quadrada do *segundo momento não centrado* da variável, $m_2 = \frac{1}{n}\sum_{i=1}^{2} x_i^2 = \frac{1}{n}\|\mathbf{x}\|^2$.

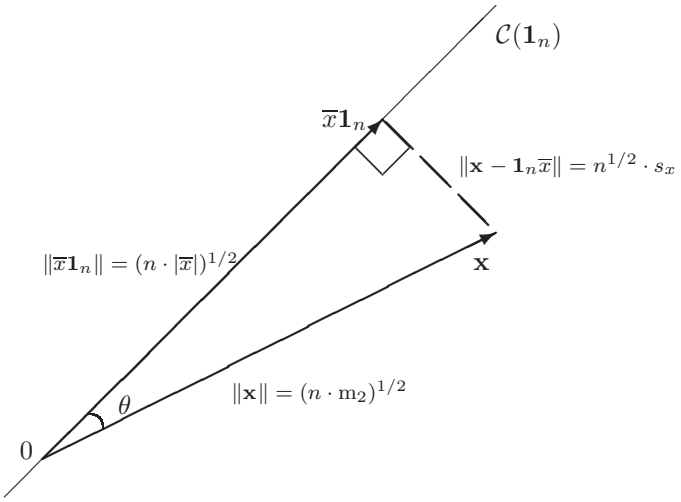


Figura 1.7: O significado geométrico do desvio padrão no espaço dos indivíduos.

Considerem-se agora n pares de observações sobre duas variáveis, $\{(x_i, y_i)\}_{i=1}^n$. Tem-se:

4. A **covariância** das observações de $\mathbf{x}$ e $\mathbf{y}$ é o *produto interno dos vectores projectados sobre* $\mathcal{C}(\mathbf{1}_n)^\perp$:

$$\mathrm{cov}(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n} < (\mathbf{I} - \mathbf{P}_{\mathbf{1}_n})\mathbf{x}, (\mathbf{I} - \mathbf{P}_{\mathbf{1}_n})\mathbf{y} >$$

$$\iff \mathrm{cov}(\mathbf{x},\mathbf{y}) = \frac{1}{n} < \mathbf{x}^c, \mathbf{y}^c > .$$

5. O **coeficiente de correlação** entre $\mathbf{x}$ e $\mathbf{y}$ é o *coseno do ângulo entre os vectores das variáveis centradas*. De facto:

$$r_{xy} = \frac{\mathrm{cov}(\mathbf{x},\mathbf{y})}{\sigma_x \cdot \sigma_y} = \frac{< (\mathbf{I} - \mathbf{P}_{\mathbf{1}_n})\mathbf{x}, (\mathbf{I} - \mathbf{P}_{\mathbf{1}_n})\mathbf{y} >}{\|(\mathbf{I} - \mathbf{P}_{\mathbf{1}_n})\mathbf{x}\| \cdot \|(\mathbf{I} - \mathbf{P}_{\mathbf{1}_n})\mathbf{y}\|} = \cos\left([\mathbf{I} - \mathbf{P}_{\mathbf{1}_n}]\mathbf{x}, [\mathbf{I} - \mathbf{P}_{\mathbf{1}_n}]\mathbf{y}\right)$$

$$\iff r_{xy} = \cos(\mathbf{x}^c, \mathbf{y}^c) .$$

1.6.3 Descrição Multivariada (p variáveis) - Primeiras ferramentas

Em modelos com várias variáveis preditoras, torna-se útil a representação matricial dos dados observados e de conceitos estatísticos associados. Designe-se por $\mathbf{X}$ a matriz cujas colunas representam as observações

de uma dada variável $\mathbf{x}_i$. Defina-se:

1. **Vector (px1) das médias**: $\overline{\mathbf{x}} = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_p \end{bmatrix} = \mathbf{X}^t \mathbf{1}_n (\mathbf{1}_n{}^t \mathbf{1}_n)^{-1}$

2. **Matriz ($p \times p$) das variâncias-covariâncias**:

$$\mathbf{\Sigma} = \begin{bmatrix} \text{var}_1 & \text{cov}_{1,2} & \text{cov}_{1,3} & \dots & \text{cov}_{1,p} \\ \text{cov}_{2,1} & \text{var}_2 & \text{cov}_{2,3} & \dots & \text{cov}_{2,p} \\ \text{cov}_{3,1} & \text{cov}_{3,2} & \text{var}_3 & \dots & \text{cov}_{3,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{cov}_{p,1} & \text{cov}_{p,2} & \text{cov}_{p,3} & \dots & \text{var}_p \end{bmatrix}$$

   Se $\mathbf{Y} = (\mathbf{I} - \mathbf{P}_{\mathbf{1}_n})\mathbf{X}$ é matriz de dados com colunas *centradas*, tem-se: $\mathbf{\Sigma} = \frac{1}{n}\mathbf{Y}^t\mathbf{Y} = \frac{1}{n}\mathbf{X}^t(\mathbf{I} - \mathbf{P}_{\mathbf{1}_n})\mathbf{X}$.

3. **Matriz ($p \times p$) das correlações**:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{1,2} & r_{1,3} & \dots & r_{1,p} \\ r_{2,1} & 1 & r_{2,3} & \dots & r_{2,p} \\ r_{3,1} & r_{3,2} & 1 & \dots & r_{3,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{p,1} & r_{p,2} & r_{p,3} & \dots & 1 \end{bmatrix}$$

   Se $\mathbf{Z}$ é matriz de dados com colunas *normalizadas*, tem-se: $\mathbf{R} = \frac{1}{n}\mathbf{Z}^t\mathbf{Z}$.

Notas:

1. $\mathbf{\Sigma} = \mathbf{DRD}$ onde $\mathbf{D}$ é a matriz diagonal ($p \times p$) dos desvios padrão.

2. $\mathbf{R} = \mathbf{D}^{-1}\mathbf{\Sigma}\mathbf{D}^{-1}$ onde $\mathbf{D}^{-1}$ é a inversa da matriz $\mathbf{D}$, isto é, a matriz (diagonal) dos recíprocos dos desvios padrão.

3. Já vimos que uma matriz de variâncias/covariâncias é da forma $\mathbf{\Sigma} = \frac{1}{n}\mathbf{Y}^t\mathbf{Y}$, e uma matriz de correlações é da forma $\mathbf{R} = \frac{1}{n}\mathbf{Z}^t\mathbf{Z}$. Assim, o Teorema 1.36 mostrou que **qualquer matriz de variâncias/covariâncias ou de correlações é sempre uma matriz definida positiva** se não houver multicolinearidade das variáveis que a definem, e que se houver multicolinearidade será apenas uma matriz semi-definida positiva. Em contrapartida, qualquer matriz definida positiva é uma matriz de variâncias/covariâncias (ou de correlações) para algum conjunto de dados (até mesmo para diferentes conjuntos de dados, inclusivamente com um número de indivíduos observados diferente).

Encerraremos esta secção de revisão de aplicações estatísticas lembrando que o Modelo Linear (Regressão Linear, Análise de Variância ou de Covariância de efeitos fixos), que foi já objecto de estudo na disciplina de Modelação Estatística deste Mestrado, tem uma forte componente geométrica, baseada nos conceitos de Álgebra Linear expostos até aqui. Trata-se de técnicas estatísticas que, sendo embora Multivariadas (uma vez que envolvem uma variável resposta e uma ou mais variáveis preditoras), são usualmente abordadas em disciplinas de Estatística introdutória.

1.7 Exercícios

Exercícios de Álgebra Linear e Teoria de Matrizes

1. Responda às seguintes questões:
   (a) Diga o que é um *espaço linear*.
   (b) Diga o que é uma *base* dum espaço linear.
   (c) Diga o que é uma *base ortonormada* dum espaço linear.
   (d) Diga o que é o *complemento ortogonal* dum subespaço linear.
   (e) Defina o conceito de *soma directa*.

2. Mostre que se $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n$ são vectores não-nulos, ortogonais dois a dois, têm de ser linearmente independentes.

3. Mostre que o conjunto de matrizes de tipo $n \times p$, associado às habituais operações de soma de matrizes e de multiplicação escalar, constitui um espaço linear.

4. Considere as seguintes matrizes no espaço linear de matrizes $\mathbb{M}_{3\times 2}$:
$$\mathbf{A} = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \quad \text{e} \quad \mathbf{B} = \begin{bmatrix} 2 & 8 \\ 4 & 10 \\ 6 & 12 \end{bmatrix}.$$
   (a) Determine o inverso aditivo de $\mathbf{A}$ no espaço linear $\mathbb{M}_{3\times 2}$.
   (b) As matrizes $\mathbf{A}$ e $\mathbf{B}$ são linearmente independentes? Justifique.
   (c) Calcule o produto matricial $\mathbf{A}^t\mathbf{B}$.
   (d) Se $< \cdot, \cdot >$ representa o produto interno usual no espaço de matrizes $\mathbb{M}_{n\times p}$, calcule $< \mathbf{A}, \mathbf{B} >$.

5. Considere o espaço $\mathbb{R}^n$ com o produto interno usual. Sejam $\mathbf{x}\,, \mathbf{y} \in \mathbb{R}^n$ vectores com as observações de duas variáveis X e Y nos mesmos n indivíduos. Sejam $\mathbf{x}^c \equiv [x_i - \overline{x}]$ e $\mathbf{y}^c \equiv [y_i - \overline{y}]$ os vectores centrados correspondentes.
   (a) Mostre que a covariância amostral de X e Y é dada por $cov_{xy} = \frac{1}{n-1} < \mathbf{x}^c, \mathbf{y}^c >$.
   (b) Mostre que a variância amostral de X é dada por $s_x^2 = \frac{1}{n-1} \|\mathbf{x}^c\|^2$.
   (c) Interprete a desigualdade de Cauchy-Schwarz-Buniakovski aplicada aos vectores $\mathbf{x}^c$ e $\mathbf{y}^c$, á luz das alíneas anteriores.
   (d) Interprete o coeficiente de correlação amostral entre X e Y, r_{xy}, à luz das alíneas anteriores.

6. Prove o Teorema 1.12, ou seja, mostre que, se L é um espaço linear com a norma $\|\cdot\|$, se verifica:
   (a) $\|\mathbf{x} - \mathbf{y}\| \geq \Big|\|\mathbf{x}\| - \|\mathbf{y}\|\Big|\,, \quad \forall \mathbf{x}, \mathbf{y} \in \mathrm{L}$
   (b) $\|\mathbf{x} - \mathbf{y}\| \leq \|\mathbf{x} - \mathbf{z}\| + \|\mathbf{z} - \mathbf{y}\|\,, \quad \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathrm{L}$

7. Mostre que, se L é um espaço linear com a norma $\|\cdot\|$, induzida pelo produto interno $<\cdot,\cdot>$, tem-se:
$$\|\mathbf{x}+\mathbf{y}\|^2 \;=\; \|\mathbf{x}\|^2 + 2<\mathbf{x},\mathbf{y}> + \|\mathbf{y}\|^2 \;, \quad \forall\, \mathbf{x},\mathbf{y} \in \mathrm{L}\;.$$

8. Mostre que as habituais operações elementares sobre as linhas/colunas de uma matriz $\mathbf{A}$, utilizadas no método de Gauss para resolver sistemas de equações, correspondem a pré-/pós-multiplicar a matriz $\mathbf{A}$ por matrizes convenientes. Em particular, mostre que:

   (a) Para trocar a linha i e a linha j da matriz $\mathbf{A}$ *pré*-multiplica-se a matriz $\mathbf{A}$ por uma matriz identidade com as linhas i e j trocadas. Por exemplo, se $\mathbf{A}$ tem quatro linhas e se pretende trocar as linhas 1 e 3, deve-se pré-multiplicar a matriz $\mathbf{A}$ pela matriz
$$\mathbf{E}_{1,3} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

   (b) Para trocar a coluna i e a coluna j da matriz $\mathbf{A}$ deve-se *pós*-multiplicar a matriz $\mathbf{A}$ por uma matriz identidade com as colunas i e j trocadas. Compare estas matrizes de permuta de colunas com as matrizes de permuta de linhas e comente.

   (c) Para que a linha i da matriz $\mathbf{A}$ passe a ser a soma de α vezes a actual linha i mais β vezes a actual linha j (com $\alpha, \beta \in \mathbb{R}$), deve-se *pré*-multiplicar a matriz $\mathbf{A}$ por uma matriz identidade, mas em que a linha i fica substituída por uma linha com α na posição i e β na posição j (sendo os restantes elementos dessa linha nulos). Por exemplo, se $\mathbf{A}$ tem quatro linhas e se pretende que a linha 3 passe a ser igual ao dobro da actual linha 3, menos o triplo da actual primeira linha, deve-se pré-multiplicar a matriz $\mathbf{A}$ pela matriz
$$\mathbf{E} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -3 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

   (d) Para que a coluna i da matriz $\mathbf{A}$ passe a ser a soma de α vezes a actual coluna i mais β vezes a actual coluna j (com $\alpha, \beta \in \mathbb{R}$), deve-se *pós*-multiplicar a matriz $\mathbf{A}$ por uma matriz identidade, mas em que a coluna i fica substituída por uma coluna com α na posição i e β na posição j (sendo os restantes elementos dessa coluna nulos). Compare estas matrizes de combinações lineares de colunas com as matrizes de combinações lineares de linhas e comente.

9. Considere os vectores de $\mathbb{R}^3$: $\mathbf{x} = (1,1,1)^t$ e $\mathbf{y} = (1,2,3)^t$.

   (a) Represente geometricamente $\mathbf{x}$ e $\mathbf{y}$ em $\mathbb{R}^3$.

   (b) Calcule, usando o produto interno usual em $\mathbb{R}^3$ e as respectivas norma e distância induzidas, o produto interno $<\mathbf{x},\mathbf{y}>$, as normas $\|\mathbf{x}\|$ e $\|\mathbf{y}\|$, a distância $d(\mathbf{x},\mathbf{y})$ e o cosseno do ângulo entre os vectores $\mathbf{x}$ e $\mathbf{y}$.

Considere agora a matriz $\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{9} \end{bmatrix}$.

(c) Mostre que $\mathbf{M}$ é uma matriz definida positiva.

(d) Com base no produto interno em $\mathbb{R}^3$ definido pela matriz $\mathbf{M}$ e as respectivas norma e distância induzidas, calcule o produto interno $< \mathbf{x}, \mathbf{y} >_M$, as normas $\|\mathbf{x}\|_M$ e $\|\mathbf{y}\|_M$, a distância $d_M(\mathbf{x}, \mathbf{y})$ e o cosseno do ângulo entre os vectores $\mathbf{x}$ e $\mathbf{y}$ definido pelo novo produto interno.

10. Defina-se a função $< \cdot, \cdot >_M : \mathbb{R}^n \mathrm{x} \mathbb{R}^n \longrightarrow \mathbb{R}_0^+$, dada por:

$$< \mathrm{x}, \mathrm{y} >_M = \mathrm{x}^{\mathrm{t}} \mathrm{M} \mathrm{y} \qquad \forall \quad \mathrm{x}, \mathrm{y} \in \mathbb{R}^n$$

(a) Mostre que esta função é um produto interno em $\mathbb{R}^n$ se e só se a matriz quadrada M fôr definida positiva.

(b) Seja $\mathcal{A}$ um subespaço de $\mathbb{R}^n$. Considere o subconjunto $\mathcal{A}^{\perp_M}$ de todos os vectores de $\mathbb{R}^n$ que são *M-ortogonais* com todos os vectores de $\mathcal{A}$, isto é:

$$\mathcal{A}^{\perp_M} = \{\mathrm{y} \in \mathbb{R}^n \ : < \mathrm{y}, \mathrm{x} >_M = 0, \forall \mathrm{x} \in \mathcal{A}\}$$

Mostre que $\mathcal{A}^{\perp_M}$ é um *subespaço* de $\mathbb{R}^n$.

(c) Mostre que $\mathbb{R}^n$ é uma soma directa dos subespaços $\mathcal{A}$ e $\mathcal{A}^{\perp_M}$, definidos na alínea anterior, isto é, mostre que $\mathbb{R}^n = \mathcal{A} \oplus \mathcal{A}^{\perp_M}$. (NOTA: Pode admitir que qualquer subespaço tem uma base M-ortonormada).

(d) Seja $\mathbf{A}$ uma matriz cujas colunas formam uma base qualquer do subespaço $\mathcal{A} \in \mathbb{R}^n$. Mostre que a matriz $P_{(A,M)} = A(A^t M A)^{-1} A^t M$ é a matriz de projecção M-ortogonal em $\mathcal{A}$, isto é, a matriz de projecção associada à soma directa da alínea anterior.

11. Seja $\mathbf{V}$ uma matriz ortogonal $p \times p$. Para qualquer vector $\mathbf{x} \in \mathbb{R}^p$, compare a norma euclidiana usual de $\mathbf{x}$ e de $\mathbf{V}\mathbf{x}$. Comente.

12. Considere a matriz $\mathbf{A} = \begin{bmatrix} 7 & \sqrt{6} \\ \sqrt{6} & 2 \end{bmatrix}$.

(a) Determine algebricamente os seus valores e vectores próprios, directamente a partir da definição.

(b) Determine os seus valores e vectores próprios utilizando o programa R.

(c) Confirme que se verifica a Decomposição Espectral de $\mathbf{A}$, na forma de produto matricial, utilizando os vectores e valores próprios obtidos nas alíneas anteriores.

(d) Confirme que se verifica a Decomposição Espectral de $\mathbf{A}$, agora na forma de somatório de matrizes de característica 1.

13. Seja $\mathbf{D}$ uma matriz diagonal. Determine uma Decomposição Espectral de $\mathbf{D}$.

14. Seja $\mathbf{V}$ uma matriz de dimensão 3×3, com colunas $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$. Seja $\mathbf{\Lambda}$ uma matriz 3×3 diagonal, com elementos diagonais $\lambda_1, \lambda_2, \lambda_3$. Mostre que se verifica a igualdade $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^t = \sum_{i=1}^{3} \lambda_i \mathbf{v}_i \mathbf{v}_i^t$.

15. Sejam $\mathbf{A}$ e $\mathbf{B}$ duas matrizes simétricas $p \times p$ que partilham um mesmo conjunto ortonormado de vectores próprios. Utilize a Decomposição Espectral de $\mathbf{A}$ e $\mathbf{B}$ para determinar o que pode afirmar sobre os valores próprios da matriz $\mathbf{A} + \mathbf{B}$.

16. Sejam $\mathbf{A}$ e $\mathbf{B}$ duas matrizes simétricas $p \times p$.

    (a) Verifique que o produto $\mathbf{AB}$ não é, em geral, simétrico.

    (b) Indique uma condição necessária e suficiente para que o produto $\mathbf{AB}$ seja simétrico.

17. Prove que, se $\mathbf{A}$ é uma matriz $n \times p$ e $\mathbf{B}$ uma matriz $p \times m$, se verificam as seguintes relações:

    (a) $\mathcal{C}(\mathbf{AB}) \subset \mathcal{C}(\mathbf{A})$.

    (b) $\mathcal{N}(\mathbf{B}) \subset \mathcal{N}(\mathbf{AB})$.

    (c) $car(\mathbf{AB}) \leq car(\mathbf{A})$.

    (d) $car(\mathbf{AB}) \leq car(\mathbf{B})$.

18. Considere as matrizes $\mathbf{X}^t\mathbf{X}$ e $\mathbf{X}\mathbf{X}^t$, onde $\mathbf{X}$ é uma matriz $n \times p$. Verifique que se λ_j é um valor próprio de $\mathbf{X}^t\mathbf{X}$, com vector próprio associado $\mathbf{c}_j$, então $\mathbf{X}\mathbf{c}_j$ é um vector próprio da matriz $\mathbf{X}\mathbf{X}^t$, para o mesmo valor próprio. Conversamente, se λ_j é um valor próprio de $\mathbf{X}\mathbf{X}^t$ com vector próprio associado $\mathbf{b}_j$, então $\mathbf{X}^t\mathbf{b}_j$ é um vector próprio de $\mathbf{X}^t\mathbf{X}$, com o mesmo valor próprio associado.

19. Seja $\mathbf{A}$ uma matriz simétrica $p \times p$ e $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^t$ a sua decomposição espectral. Mostre que se $\mathbf{A}$ tem característica r, é possível escrever $\mathbf{A} = \mathbf{V}_r\mathbf{\Lambda}_r\mathbf{V}_r^t$, onde $\mathbf{V}_r$ é a submatriz de tamanho $p \times r$ constituída pela r colunas de $\mathbf{V}$ associadas a valores próprios não-nulos, e $\mathbf{\Lambda}_r$ é a submatriz $r \times r$ de $\mathbf{\Lambda}$ cujos r elementos diagonais são os valores próprios não nulos de $\mathbf{A}$.

20. Considere uma matriz $\mathbf{X}_{n\times p}$, de característica p e com Decomposição em Valores Singulares $\mathbf{X} = \mathbf{W}\mathbf{\Delta}\mathbf{V}^t$. Sejam $\{\mathbf{v}_i\}_{i=1}^p$ os vectores singulares direitos, e $\{\mathbf{w}_i\}_{i=1}^p$ os vectores singulares esquerdos de $\mathbf{X}$. Considere a aplicação de $\mathbb{R}^p$ em $\mathbb{R}^n$ definida por $\mathbf{X}$.

    (a) Seja $\mathbf{z} = \sum_{i=1}^{p} \alpha_i \mathbf{v}_i$ um vector de $\mathbb{R}^p$, escrito como combinação linear dos p vectores singulares direitos de $\mathbf{X}$. Descreva a imagem de $\mathbf{z}$ através de $\mathbf{X}$ e interprete geometricamente.

    (b) Considere o conjunto de todos os vectores em $\mathbb{R}^p$ de norma euclidiana usual 1. Estes vectores formam uma hiperesfera em $\mathbb{R}^p$. Tendo em conta o resultado da alínea anterior, descreva a imagem desta hiperesfera quando submetida ao efeito da aplicação definida por $\mathbf{X}$.

21. Considere uma matriz $\mathbf{X}_{n\times p}$, de característica $r < p$ e com Decomposição em Valores Singulares $\mathbf{X} = \mathbf{W}\mathbf{\Delta}\mathbf{V}^t$. Indique o que sucede às imagens, através da aplicação definida pela matriz $\mathbf{X}$, dos vectores de $\mathbb{R}^p$ que pertencem ao complemento ortogonal do subespaço gerado pelas colunas de $\mathbf{V}$, $\mathcal{C}(\mathbf{V})^t$.

22. Verificar que a Decomposição em Valores Singulares duma matriz definida positiva (necessariamente quadrada e simétrica, pela definição de definida positiva) equivale à sua Decomposição Espectral (*i.e.*, à sua Decomposição em Valores e Vectores Próprios). Qual a relação entre as 2 decomposições no caso de uma matriz *semi*-definida positiva? E no caso de uma matriz definida *negativa*?

(**Observação:** Ver também o exercício 18).

23. Utilize a Decomposição em Valores Singulares duma matriz $\mathbf{X}$, na forma:

$$\mathbf{X} = \sum_{i=1}^{r} \delta_i \mathbf{w}_i \mathbf{v}_i^t$$

para mostrar que se $\mathbf{w}_i$ é um vector singular esquerdo associado ao valor singular δ_i, e $\mathbf{v}_i$ é um vector singular direito associado ao mesmo valor singular, então tem-se:

$$\mathbf{X}\mathbf{v}_i = \delta_i \mathbf{w}_i \qquad \text{e} \qquad \mathbf{X}^t \mathbf{w}_i = \delta_i \mathbf{v}_i$$

24. Considere uma matriz $\mathbf{B}$ e a matriz de projecção ortogonal sobre o subespaço gerado pelas colunas de $\mathbf{B}$, $\mathbf{P}_B = \mathbf{B}(\mathbf{B}^t\mathbf{B})^{-1}\mathbf{B}^t$. Utilizando a Decomposição em Valores Singulares da matriz $\mathbf{B}$, obtenha uma expressão alternativa para a matriz $\mathbf{P}_B$. Comente.

25. Considere o espaço linear das matrizes de tipo $n \times p$ (com as habituais operações de soma de matrizes e de produto duma matriz por um escalar). Mostre que a aplicação que a cada par de matrizes, $\mathbf{A}, \mathbf{B}$, nesse espaço associa o número real $\text{tr}(\mathbf{A}^t\mathbf{B})$ constitui um produto interno nesse espaço.

26. Considere o espaço linear $\mathbb{M}_n$ das matrizes quadradas de dimensão n (com as habituais operações de soma de matrizes e de produto duma matriz por um escalar).

    (a) Mostre que o conjunto das matrizes simétricas forma um subespaço linear de $\mathbb{M}_n$.

    (b) Uma matriz quadrada $\mathbf{A}$ diz-se *anti-simétrica* se $\mathbf{A}^t = -\mathbf{A}$. Mostre que o conjunto das matrizes anti-simétricas (de dimensão n) forma um subespaço linear de $\mathbb{M}_n$.

    (c) Mostre que o espaço linear $\mathbb{M}_n$ pode ser decomposto na soma directa do subespaço das matrizes simétricas com o subespaço das matrizes anti-simétricas. Caracterize as componentes (únicas) duma matriz quadrada genérica $\mathbf{A}$ nos subespaços das matrizes simétricas e anti-simétricas acima referidos.

Capítulo 2

Análise em Componentes Principais

A Análise em Componentes Principais (ACP) é, possivelmente, a técnica de Estatística Multivariada mais utilizada. Pode ser introduzida de várias formas alternativas, e os seus objectivos também podem ser expressos de diferentes formas. Na introdução à ACP feita nesta disciplina, privilegiar-se-ão os aspectos geométricos. Mais adiante será feita uma introdução alternativa, utilizando conceitos estatísticos.

2.1 Uma introdução geométrica

Seja dada uma **matriz de dados**, $\mathbf{Y} \in \mathbb{M}_{n\times p}$, correspondente a observações de p variáveis **quantitativas** em n individuos.

Do ponto de vista geométrico, estes dados admitem duas representações:

- **uma nuvem de n pontos em $\mathbb{R}^p$**: uma representação que associa um eixo a cada variável observada e onde às linhas da matriz $\mathbf{Y}$ (indivíduos) correspondem pontos nesse espaço $\mathbb{R}^p$.
- **um feixe de p vectores em $\mathbb{R}^n$**: uma representação onde a cada coluna da matriz de dados (variáveis) corresponde um ponto/vector em $\mathbb{R}^n$, estando os eixos do espaço associados aos individuos.

Qualquer destas representações não é visualizável pelo ser humano (excepto quando n ou p não excedem três). O **objectivo duma Análise em Componentes Principais** pode ser formulado do seguinte modo, ainda informal: aproximar a nuvem de pontos em $\mathbb{R}^p$ por outra nuvem de n pontos num subespaço de menor dimensão, da forma mais fidedigna possível.

As considerações já feitas sobre projecções sugerem que a solução do problema passe por projecções ortogonais sobre algum subespaço de dimensão inferior a p e n. Mas como determinar o melhor subespaço de dimensão q?

Na consideração deste problema é vantajoso começar por **centrar as colunas da matriz de dados**.

Assim, se $\mathbf{Y} \in \mathbb{M}_{n\times p}$ é uma matriz genérica de tipo $n \times p$, considere-se a matriz $\mathbf{X} = (\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{Y}$ obtida subtraíndo a cada coluna de $\mathbf{Y}$ a sua média ($\mathbf{P}_{\mathbf{1}_n}$ é a matriz de projecção ortogonal sobre o subespaço gerado pelo vector de n uns).

Nas representações geométricas já referidas, o efeito desta centragem das colunas da matriz de dados é a seguinte:

- **Na representação em** $\mathbb{R}^p$ efectua-se uma translação da nuvem de n pontos, de forma a fazer coincidir o seu centro de gravidade com a origem do sistema de eixos. De tal forma, permite-se que os subespaços a considerar para eventuais projecções estejam mais próximos da nuvem de pontos (recorde-se que qualquer subespaço tem de conter a origem - o elemento nulo do espaço).

- **Na representação em** $\mathbb{R}^n$ os vectores associados às variáveis centradas têm as propriedades geométricas com significado estatístico já discutidas anteriormente: o comprimento (norma) de qualquer vector é proporcional ao desvio padrão da variável a que está associado; e o cosseno do ângulo entre dois vectores é o coeficiente de correlação das variáveis correspondentes.

Foquemos a atenção na representação dos dados centrados no espaço $\mathbb{R}^p$. Já vimos que o objectivo da Análise em Componentes Principais (ACP) pode ser expresso como a procura dum subespaço de dimensão $q < \min\{p, n\}$ no qual a representação dos dados seja o mais "fiel" possível. Desde logo, coloca-se a questão de definir com rigor o critério de "fidelidade" que se pretende utilizar. O critério de qualidade da representação usado na ACP é o seguinte: escolher o **subespaço q-dimensional que minimize a soma dos quadrados das distâncias entre os pontos da nuvem original e as respectivas projecções ortogonais sobre o subespaço** (ver a Figura 2.1).

Considerem-se n pontos num subespaço q-dimensional de $\mathbb{R}^p$, e coloquem-se as respectivas coordenadas nas n linhas de uma matriz $\tilde{\mathbf{X}}_{n\times p}$. O facto de os pontos estarem num subespaço q-dimensional traduz-se no facto de a matriz $\tilde{\mathbf{X}}$ ser de característica q. Por outro lado, e como foi visto no Exemplo 1.7, a soma de quadrados das diferenças entre as coordenadas de cada linha de $\mathbf{X}$ e de a correspondente linha de $\tilde{\mathbf{X}}$ é dada por $\|\mathbf{X} - \tilde{\mathbf{X}}\|^2$. Assim, considerando o problema em termos do espaço $\mathbb{R}^p$, podemos dizer que:

através da ACP se procura a matriz $\tilde{\mathbf{X}}_{n\times p}$, *de característica q*, que minimiza $\|\mathbf{X} - \tilde{\mathbf{X}}\|$.

A solução do problema já foi encontrada no Teorema 1.42 (página 44), através da Decomposição em Valores Singulares, e baseia-se apenas nos q primeiros vectores singulares (esquerdos e direitos) e valores singulares. Ou seja,

a matriz $\tilde{\mathbf{X}}$ será dada por $\tilde{\mathbf{X}} = \mathbf{W}_q\mathbf{\Delta}_q\mathbf{V}_q^t$, em que as colunas de $\mathbf{W}_q$ e $\mathbf{V}_q$ são os q vectores singulares esquerdos e direitos da matriz $\mathbf{X}$ associados aos q maiores valores singulares, que constituem os elementos diagonais da matriz diagonal $\mathbf{\Delta}_q$.

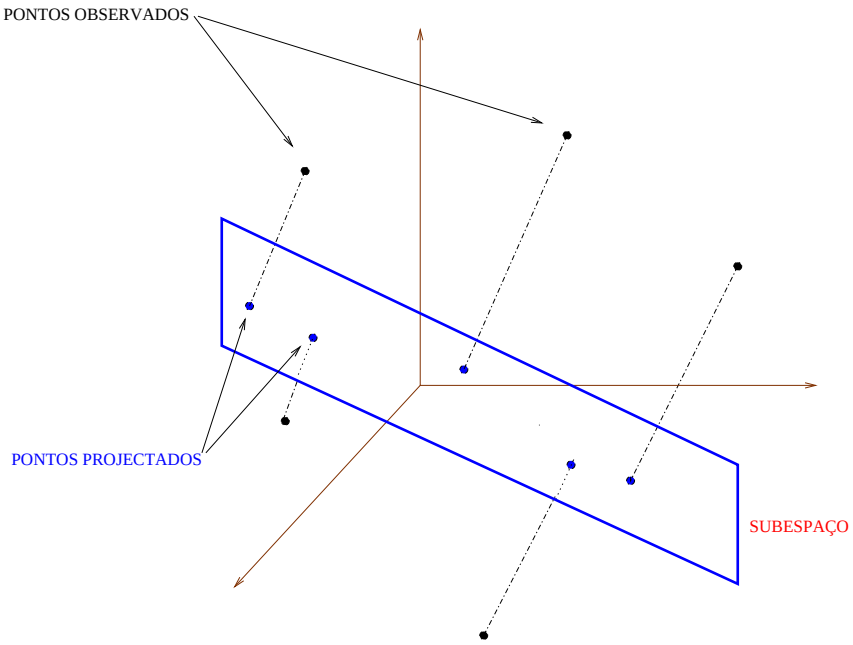


Figura 2.1: A projecção ortogonal de pontos sobre um subespaço. Nas projecções ortogonais minimiza-se a soma de quadrados das distâncias entre os pontos e as suas projecções.

É de assinalar que *esta solução para o problema que havia sido formulado considerando a configuração de n pontos que* $\mathbf{X}$ *define no espaço* $\mathbb{R}^p$*, é simultaneamente uma solução do problema análogo que se pode formular considerando o feixe de p vectores em* $\mathbb{R}^n$ *definidos pelas coordenadas das p colunas de* $\mathbf{X}$. De facto, considerem-se p vectores num subespaço q-dimensional de $\mathbb{R}^n$, e coloquem-se as respectivas coordenadas nas p colunas de uma matriz $\tilde{\mathbf{X}}_{n\times p}$. O facto de os vectores estarem num subespaço q-dimensional ir-se-á reflectir no facto de a matriz $\tilde{\mathbf{X}}$ ser de característica q. A soma de quadrados das diferenças entre as coordenadas de cada *coluna* de $\mathbf{X}$ e de a correspondente *coluna* de $\tilde{\mathbf{X}}$ é ainda dada por $\|\mathbf{X}-\tilde{\mathbf{X}}\|$. Assim, formulando o problema em termos do espaço $\mathbb{R}^n$, podemos dizer que **uma ACP procura a matriz $\tilde{\mathbf{X}}_{n\times p}$, *de característica q*, que minimiza** $\|\mathbf{X}-\tilde{\mathbf{X}}\|$. Como vemos, trata-se do mesmo problema algébrico que o resultante de formular o problema em termos da configuração de n pontos de $\mathbb{R}^p$ definida pelas linhas de $\mathbf{X}$ e a solução é, por conseguinte, idêntica.

Veremos agora que **a solução $\tilde{\mathbf{X}}$ do problema resulta de:**

Em $\mathbb{R}^p$ - Projectar ortogonalmente as n linhas de $\mathbf{X}$ sobre o subespaço gerado pelos q vectores próprios da matriz de variâncias/covariâncias $\mathbf{\Sigma}=\frac{1}{n}\mathbf{X}^t\mathbf{X}$, associados aos seus q maiores valores próprios. Representaremos esses vectores próprios por $\{\mathbf{v}_j\}_{j=1}^q$.

Em $\mathbb{R}^n$ - Projectar ortogonalmente as p colunas de $\mathbf{X}$ sobre o subespaço gerado pelas q combinações lineares das variáveis originais, dadas por $\{\mathbf{X}\mathbf{v}_j\}_{j=1}^q$ (espaço esse que é idêntico ao espaço gerado pelos q primeiros vectores próprios da matriz $\mathbf{X}\mathbf{X}^t$).

Designemos por $\mathbf{V}_q$ a matriz cujas colunas são os q vectores singulares direitos de $\mathbf{X}$ associados aos seus q maiores valores singulares (isto é, os q primeiros vectores próprios da matriz de variâncias-covariâncias definida por $\mathbf{X}$). A matriz de projecções ortogonais sobre o subespaço gerado pelas colunas de $\mathbf{V}_q$ é da forma $\mathbf{V}_q(\mathbf{V}_q^t\mathbf{V}_q)^{-1}\mathbf{V}_q^t$. Tendo em conta que as colunas de $\mathbf{V}_q$ formam um conjunto ortonormado, esta matriz de projecções reduz-se à matriz $\mathbf{V}_q\mathbf{V}_q^t$. Para projectar as *linhas* de $\mathbf{X}$, começamos por escrever essas linhas em colunas, *i.e.*, por considerar a matriz $\mathbf{X}^t$. Depois pré-multiplicamos essa matriz pela matriz de projecções acima referida. Finalmente, voltamos a escrever as colunas projectadas na forma de linhas, *i.e.*, transpomos o resultado. Daí resulta a matriz $(\mathbf{V}_q\mathbf{V}_q^t\mathbf{X}^t)^t = \mathbf{X}\mathbf{V}_q\mathbf{V}_q^t$. Tendo em conta a DVS de $\mathbf{X}$, temos então:

$$\begin{aligned}
\mathbf{X}\mathbf{V}_q\mathbf{V}_q^t &= (\mathbf{W}\boldsymbol{\Delta}\mathbf{V}^t)(\mathbf{V}_q\mathbf{V}_q^t) \\
&= (\textstyle\sum_{i=1}^p \delta_i\mathbf{w}_i\mathbf{v}_i^t)(\sum_{j=1}^q \mathbf{v}_j\mathbf{v}_j^t) \\
&= \textstyle\sum_{i=1}^q \delta_i\mathbf{w}_i(\mathbf{v}_i^t\mathbf{v}_i)\mathbf{v}_i^t && \text{pois } \mathbf{v}_i^t\mathbf{v}_j = 0 \text{ se } i \neq j \\
&= \textstyle\sum_{i=1}^q \delta_i\mathbf{w}_i\mathbf{v}_i^t && \text{pois } \mathbf{v}_i^t\mathbf{v}_i = 1 \\
&= \mathbf{W}_q\boldsymbol{\Delta}_q\mathbf{V}_q^t
\end{aligned}$$

Assim, as linhas de $\tilde{\mathbf{X}}$ resultam de projectar ortogonalmente as linhas de $\mathbf{X}$ sobre o subespaço (q-dimensional) de $\mathbb{R}^p$ gerado pelos q primeiros vectores próprios de $\boldsymbol{\Sigma}$.

Mas, ao mesmo tempo, as colunas de $\tilde{\mathbf{X}}$ resultam de projectar ortogonalmente as colunas de $\mathbf{X}$ sobre o subespaço (q-dimensional) de $\mathbb{R}^n$ gerado pelas q colunas de $\mathbf{X}\mathbf{V}_q$, isto é, pelas q colunas da matriz $\mathbf{W}_q\boldsymbol{\Delta}_q$. De facto, a matriz de projecção ortogonal sobre esse subespaço é a matriz:

$$(\mathbf{W}_q\boldsymbol{\Delta}_q)[(\boldsymbol{\Delta}_q\mathbf{W}_q^t)(\mathbf{W}_q\boldsymbol{\Delta}_q)]^{-1}(\boldsymbol{\Delta}_q\mathbf{W}_q^t).$$

Tendo em conta a ortonormalidade das colunas de $\mathbf{W}_q$ e a existência de inversa de $\boldsymbol{\Delta}_q$, ficamos simplesmente com a matriz $\mathbf{W}_q\mathbf{W}_q^t$. Projectando as colunas de $\mathbf{X}$ obtemos:

$$\mathbf{W}_q\mathbf{W}_q^t\mathbf{X} = \mathbf{W}_q\mathbf{W}_q^t\mathbf{W}\boldsymbol{\Delta}\mathbf{V}^t = \mathbf{W}_q\boldsymbol{\Delta}_q\mathbf{V}_q^t$$

(com passagens intermédias análogas às vistas acima para a discussão relativa à projecção das linhas).

As variáveis em $\mathbb{R}^n$, $\{\mathbf{X}\mathbf{v}_j\}_{j=1}^p$, obtidas efectuando combinações lineares das variáveis originais, utilizando como coeficientes nessas combinações lineares os elementos dos vectores próprios $\mathbf{v}_j$, designam-se as **componentes principais** das variáveis $\{\mathbf{x}_i\}_{i=1}^p$ e costumam ordenar-se pela ordem decrescente dos valores próprios de $\boldsymbol{\Sigma}$ associados aos vectores próprios $\mathbf{v}_j$.

Os vectores de coeficientes em $\mathbb{R}^p$, (*i.e.*, os vectores próprios $\mathbf{v}_j$ de $\boldsymbol{\Sigma}$) também são chamados componentes principais por alguns autores, mas é preferível distinguir os dois conjuntos de vectores, pelo que aqui falaremos em **eixos principais em $\mathbb{R}^p$**.

Como foi referido inicialmente, procurando directamente uma solução q-dimensional determinará o(s) *subespaço(s)* q-dimensional(is) sobre o(s) qual(is) projectar. Os vectores singulares de $\mathbf{X}$ constituem uma base desse(s) subespaço(s). Além disso, uma abordagem em que se considera primeiro o espaço unidimensional dará como eixo principal em $\mathbb{R}^p$ o vector próprio de $\boldsymbol{\Sigma}$ associado ao maior valor próprio, *i.e.*, $\mathbf{v}_1$, e como componente principal (em $\mathbb{R}^n$) o vector $\mathbf{X}\mathbf{v}_1 = \mathbf{w}_1\delta_1$. Subindo a dimensão do subespaço

alvo da projecção para dois, vemos pela solução acima obtida que esse subespaço em $\mathbb{R}^p$ é gerado pelos dois primeiros vectores próprios de $\mathbf{\Sigma}$, $\mathbf{v}_1$ e $\mathbf{v}_2$, pelo que poderemos dizer que o "novo" vector, $\mathbf{v}_2$ (que, note-se, é ortogonal a $\mathbf{v}_1$) representa o segundo eixo principal em $\mathbb{R}^p$. No espaço $\mathbb{R}^n$ teremos a solução bi-dimensional dada pela base constituída pelos dois vectores $\mathbf{X}\mathbf{v}_1$ (a solução uni-dimensional) e $\mathbf{X}\mathbf{v}_2$. Este "novo" vector designa-se a segunda componente principal. Note-se que, tal como os eixos principais em $\mathbb{R}^p$, é ortogonal à primeira componente principal, pois:

$$\begin{aligned} < \mathbf{X}\mathbf{v}_1, \mathbf{X}\mathbf{v}_2 > &= \mathbf{v}_1^t \mathbf{X}^t \mathbf{X} \mathbf{v}_2 \\ &= n \cdot \mathbf{v}_1^t \mathbf{\Sigma} \mathbf{v}_2 && \text{pois } \mathbf{\Sigma} = \tfrac{1}{n}\mathbf{X}^t\mathbf{X} \\ &= n \cdot \lambda_1 \mathbf{v}_1^t \mathbf{v}_2 && \text{pois } \mathbf{v}_1 \text{ é vector próprio de } \mathbf{\Sigma} \\ &= 0 && \text{pois } \mathbf{v}_2 \text{ também é vector próprio de } \mathbf{\Sigma}\text{, ortogonal a } \mathbf{v}_1 \end{aligned}$$

Sucessivos eixos principais em $\mathbb{R}^p$ e componentes principais em $\mathbb{R}^n$ resultam de forma análoga aos anteriores vectores próprios de $\mathbf{\Sigma}$ e os restantes vectores $\mathbf{X}\mathbf{v}_i$, mantendo a ortogonalidade aos eixos e componentes anteriores.

Alguns comentários resumindo os resultados obtidos até aqui:

1. **o subespaço q-dimensional de $\mathbb{R}^p$ gerado pelos q primeiros vectores próprios de $\mathbf{\Sigma}$ é o subespaço q-dimensional de $\mathbb{R}^p$ onde a nuvem de n pontos (associados às n linhas da matriz de dados X) fica melhor representada no sentido de minimizar a soma dos quadrados das distâncias entre os n pontos originais e os n pontos que resultam da sua projecção sobre o subespaço. Podemos designar este subespaço por o subespaço principal de dimensão q em $\mathbb{R}^p$.**

2. **Os q primeiros eixos principais em $\mathbb{R}^p$ (*i.e.*, os q primeiros vectores próprios de $\mathbf{\Sigma}$, dados pelas colunas da matriz $\mathbf{V}_q$) formam uma base ortonormada do subespaço principal em $\mathbb{R}^p$.**

3. **O subespaço q-dimensional de $\mathbb{R}^n$ gerado pelos q vectores $\mathbf{X}\mathbf{v}_i = \mathbf{w}_i\delta_i$ $(i = 1, \ldots, q)$ é o subespaço q-dimensional de $\mathbb{R}^n$ onde o feixe de p vectores (associados às p colunas da matriz de dados X, *i.e.*, às p variáveis) fica melhor representado no sentido de minimizar a soma dos quadrados das distâncias entre os p vectores originais e os p vectores que resultam da sua projecção sobre o subespaço. Podemos designar este subespaço por o subespaço principal de dimensão q em $\mathbb{R}^n$.**

4. **As q primeiras componentes principais (*i.e.*, os q vectores $\mathbf{X}\mathbf{v}_i = \mathbf{w}_i\delta_i$ acima referidos) formam uma *base ortogonal* do subespaço principal q-dimensional em $\mathbb{R}^n$**, pois como vimos anteriormente são vectores sucessivamente ortogonais. *Mas não formam uma base ortonormada*, pois não são vectores de norma 1 em $\mathbb{R}^n$, já que $\|\mathbf{X}\mathbf{v}_j\| = \sqrt{\mathbf{v}_j^t \mathbf{X}^t \mathbf{X} \mathbf{v}_j} = \sqrt{n \cdot \lambda_j}$.

5. **Uma base *ortonormada* do subespaço principal em $\mathbb{R}^n$ é dada pelos q primeiros vectores singulares esquerdos de X, isto é, pelas colunas da matriz $\mathbf{W}_q$.** De facto, trata-se de vectores *de norma 1* que apontam na mesma direcção que as componentes principais $\mathbf{W}_q\mathbf{\Delta}_q$, uma vez que

o efeito de multiplicar $\mathbf{W}_q$ por $\mathbf{\Delta}_q$ é apenas o de multiplicar cada coluna de $\mathbf{W}_q$ pelo elemento diagonal correspondente de $\mathbf{\Delta}_q$ (vejam-se os comentários na página 5).

6. Tendo em conta as relações entre conceitos estatísticos e conceitos geométricos em $\mathbb{R}^n$, discutidos na Secção 1.6 (página 45), as conclusões do ponto 4 dizem-nos que **as componentes principais**, que são novas variáveis obtidas como combinação linear das variáveis originais, **são variáveis não-correlacionadas entre si**. Também se verifica que **a variância duma componente principal é dada pelo valor próprio da matriz $\mathbf{\Sigma}$ que lhe está associado** (pois, como vimos na página 46, $\text{var}(\mathbf{X}\mathbf{v}_j) = \frac{1}{n}\|\mathbf{X}\mathbf{v}_j\|^2 = \lambda_j$).

7. As componentes principais são combinações lineares das p variáveis originais (são da forma $\mathbf{X}\mathbf{v}_j$), logo pertencem ao subespaço de $\mathbb{R}^n$ que é gerado pelas p variáveis originais (isto é, ao subespaço imagem de $\mathbf{X}$, $\mathcal{C}(\mathbf{X})$). A dimensão desse subespaço imagem é a característica da matriz $\mathbf{X}$, que admitimos ser p (o número de colunas de $\mathbf{X}$). O subespaço q-dimensional de $\mathbb{R}^n$ a que temos vindo a fazer referência é pois um subespaço q-dimensional deste subespaço p-dimensional de $\mathbb{R}^n$.

Mas as componentes principais (CPs) têm ainda outras propriedades, que por vezes são utilizadas para definir de forma alternativa o próprio conceito de CPs, como veremos de seguida.

2.2 Uma introdução estatística

A introdução à Análise de Componentes Principais (ACP) feita até aqui é uma abordagem essencialmente geométrica do problema. É, no entanto, frequente encontrar nos livros de Estatística Multivariada a ACP introduzida com critérios e considerações explicitamente estatísticos.

Seja dada a matriz de dados centrada $\mathbf{X}$. **Pretende-se determinar o vector $\mathbf{v} \in \mathbb{R}^p$ que *maximize a variância* da combinação linear $\mathbf{X}\mathbf{v}$ das variáveis observadas**, isto é, que maximize a forma quadrática $\mathbf{v}^t\mathbf{\Sigma}\mathbf{v}$ (tendo em atenção a discussão da relação entre norma e desvio padrão, feita na página 46). Sem outras restrições, o problema não tem solução, pois quanto maiores os coeficientes do vector $\mathbf{v}$, maior a variância. Vamos impôr a restrição de só considerar vectores de coeficientes $\mathbf{v}$ de tamanho fixo, e concretamente, de norma 1. Isto significa impor a restrição $\mathbf{v}^t\mathbf{v} = 1$.

Com esta restrição, maximizar a forma quadrática $\mathbf{v}^t\mathbf{\Sigma}\mathbf{v}$ equivale a maximizar o quociente $\frac{\mathbf{v}^t\mathbf{\Sigma}\mathbf{v}}{\mathbf{v}^t\mathbf{v}}$ (pois se $\mathbf{v}$ não é de norma 1, o vector normalizado $\frac{\mathbf{v}}{\|\mathbf{v}\|}$ já o será, e a forma quadrática por ele definida tem a forma do quociente de Rayleigh-Ritz). O Teorema de Rayleigh-Ritz (página 34) garante-nos que essa maximização ocorre para o vector próprio associado ao maior valor próprio de $\mathbf{\Sigma}$, *i.e.*, para $\mathbf{v}_1$.

Logo, **a combinação linear das variáveis originais procurada é a combinação $\mathbf{X}\mathbf{v}_1$, ou seja, o vector que já definimos como a primeira componente principal dos dados. O valor próprio λ_1 é assim a variância da primeira CP**.

Fixada a primeira CP, passamos a procurar uma nova combinação linear $\mathbf{X}\mathbf{v}$ (com $\mathbf{v}^t\mathbf{v} = 1$) de variância máxima, mas agora com a restrição adicional $r_{(\mathbf{X}\mathbf{v},\mathbf{X}\mathbf{v}_1)} = 0$, *i.e.*, com a restrição adicional de que a nova combinação linear seja não-correlacionada com a anterior. Esta exigência equivale a pedir que a

covariância entre $\mathbf{Xv}$ e a primeira componente $\mathbf{Xv}_1$ seja zero, e pelo que foi visto anteriormente (página 47) isso equivale a exigir que o produto interno entre os dois vectores em $\mathbb{R}^n$ seja nulo, *i.e.*, que: $\mathbf{v}^t\mathbf{X}^t\mathbf{Xv}_1 = n\cdot\mathbf{v}^t\mathbf{\Sigma v}_1 = n\cdot\lambda_1\mathbf{v}^t\mathbf{v}_1 = 0$. Ou seja, a exigência de não-correlação das duas combinações lineares equivale à exigência de ortogonalidade entre os vectores $\mathbf{v}$ e $\mathbf{v}_1$. Assim, pretendemos maximizar o quociente $\frac{\mathbf{v}^t\mathbf{\Sigma v}}{\mathbf{v}^t\mathbf{v}}$, sujeito à restrição $\mathbf{v}^t\mathbf{v}_1 = 0$. Mas esse é o problema que é resolvido tomando $\mathbf{v} = \mathbf{v}_2$, o segundo vector próprio de $\mathbf{\Sigma}$, associado ao valor próprio λ_2, como vimos na discussão do quociente de Rayleigh-Ritz (Teorema 1.35, página 34). Assim, esta segunda combinação linear é a segunda componente principal, com variância λ_2. **As restantes CPs surgem de forma análoga, como solução do problema de determinar sucessivas combinações lineares de variância máxima, não-correlacionadas entre si** (e com os vectores dos coeficientes das combinações lineares de norma 1). A j-ésima componente principal é dada por $\mathbf{Xv}_j$, onde $\mathbf{v}_j$ é o vector próprio de $\mathbf{\Sigma}$ associado ao j-ésimo maior valor próprio λ_j.

Uma **demonstração alternativa** da solução já obtida, será apresentada em seguida.

Para resolver o problema de *maximizar uma função de* $\mathbb{R}^n$ *em* $\mathbb{R}$ (a função $f(\mathbf{v}) = \mathbf{v}^t\mathbf{\Sigma v}$) *sujeita a uma restrição* (a restrição $\mathbf{v}^t\mathbf{v} = 1$), podemos usar o *Método dos Multiplicadores de Lagrange* (veja-se o Apêndice A para um breve resumo deste conceito). Será necessário criar a função auxiliar:

$$h(\mathbf{x}, \lambda) = \mathbf{x}^t\mathbf{\Sigma x} - \lambda(\mathbf{x}^t\mathbf{x} - 1)$$

As derivadas parciais desta função auxiliar geram o sistema:

$$\begin{array}{lll} \frac{\partial h}{\partial \lambda} = 0 & \iff & \mathbf{x}^t\mathbf{x} = 1 \qquad\qquad [\text{A restrição}] \\ \frac{\partial h}{\partial \mathbf{x}} = 0 & \iff & 2\mathbf{\Sigma x} - 2\lambda\mathbf{x} = 0 \\ & \iff & \mathbf{\Sigma x} = \lambda\mathbf{x} \end{array}$$

Nesse caso, os candidatos a extremos são os vectores próprios de $\mathbf{\Sigma}$. Uma vez que os valores da função nesses pontos são os valores próprios λ, podemos concluir imediatamente (sem analisar a matriz Hessiana associada à função) que o máximo da função vai ser o valor λ_1, alcançado em $\mathbf{v}_1$.

Para determinar a segunda solução do problema, há agora que maximizar a variância de $\mathbf{Xv}$ sujeitos a *duas* condições: $\mathbf{v}^t\mathbf{v} = 1$ e $\mathbf{v}^t\mathbf{v}_1 = 0$. A função auxiliar para o método dos multiplicadores de Lagrange será, neste caso:

$$h(\mathbf{x}, \lambda, \mu) = \mathbf{x}^t\mathbf{\Sigma x} - \lambda(\mathbf{x}^t\mathbf{x} - 1) - \mu(\mathbf{x}^t\mathbf{v}_1 - 0) \tag{2.1}$$

O sistema de derivadas parciais será agora:

$$\begin{array}{lll} \frac{\partial h}{\partial \lambda} = 0 & \iff & \mathbf{x}^t\mathbf{x} = 1 \\ \frac{\partial h}{\partial \mu} = 0 & \iff & \mathbf{x}^t\mathbf{v}_1 = 0 \\ \frac{\partial h}{\partial \mathbf{x}} = 0 & \iff & 2\mathbf{\Sigma x} - 2\lambda\mathbf{x} - \mu\mathbf{v}_1 = \mathbf{0} \end{array}$$

É possível verificar que este sistema de equações apenas pode ter solução se $\mu = 0$. De facto,

$$\begin{aligned} \frac{\partial h}{\partial \mathbf{x}} = 0 \quad &\Longrightarrow \quad 2\mathbf{v}_1^t\boldsymbol{\Sigma}\mathbf{x} - 2\lambda\mathbf{v}_1^t\mathbf{x} - \mu\mathbf{v}_1^t\mathbf{v}_1 = \mathbf{0} \\ &\Longleftrightarrow \quad 2\lambda_1\mathbf{v}_1^t\mathbf{x} - 2\lambda\mathbf{v}_1^t\mathbf{x} - \mu = 0 \qquad [\text{pois } \boldsymbol{\Sigma}\mathbf{v}_1 = \lambda\mathbf{v}_1] \\ &\Longleftrightarrow \quad \mu = 0 \qquad [\text{pois } \mathbf{v}_1^t\mathbf{x} = 0] \end{aligned}$$

Com este valor necessário de μ, a última equação do sistema de derivadas parciais vem: $\boldsymbol{\Sigma}\mathbf{x} = \lambda\mathbf{x}$. Ou seja, $\mathbf{x}$ é novamente um vector próprio de $\boldsymbol{\Sigma}$. A ortogonalidade ao primeiro vector próprio, e a condição de ser *máximo* da função com essa restrição implicam que se trata do vector próprio associado ao segundo maior valor próprio de $\boldsymbol{\Sigma}$. As restantes componentes obtêm-se de forma análoga.

2.3 Algumas propriedades e problemas numa ACP

2.3.1 Propriedades de CPs

Vejamos algumas propriedades adicionais das Componentes Principais de um conjunto de dados.

1. **A soma das variâncias das p componentes principais de p variáveis $\{\mathbf{x}_i\}_{i=1}^p$ é igual à soma das variâncias das p variáveis originais**. De facto, sabemos que a variância de cada CP é o valor próprio da matriz $\boldsymbol{\Sigma}$ que lhe está associado. Logo a soma das variâncias das p CPs é a soma dos seus p valores próprios. Mas isso é o *traço* da matriz $\boldsymbol{\Sigma}$, que é também a soma dos elementos diagonais de $\boldsymbol{\Sigma}$, *i.e.*, a soma das variâncias das p variáveis originais. Logo, a soma das variâncias das p variáveis originais, $\text{tr}(\boldsymbol{\Sigma})$, é a soma das variâncias das p CPs, $\text{tr}(\Lambda)$.

2. Devido à propriedade anterior, é frequente dizer que **a j-ésima CP explica uma proporção da variabilidade total** igual a:

$$\pi_j = \frac{\lambda_j}{\text{tr}(\boldsymbol{\Sigma})} \tag{2.2}$$

   Esta ideia é **extensível a subconjuntos de componentes principais, ou seja aos subespaços principais de dimensão q**. Assim, pode afirmar-se que **as primeiras q CPs "explicam $\frac{\sum_{j=1}^q \lambda_i}{\sum_{j=1}^p \lambda_j}$ x 100% da variabilidade total" do conjunto de dados**. É também frequente, em particular na literatura da escola francesa, designar a "variabilidade total" (*i.e.*, a soma das variâncias das p variáveis originais) como a **inércia** do conjunto de dados.

3. A **correlação entre a i-ésima variável $\mathbf{x}_i$ e a j-ésima CP $\mathbf{X}\mathbf{v}_j$** é dada por:

$$\boxed{\text{corr}(\mathbf{x}_i, \mathbf{X}\mathbf{v}_j) = \sqrt{\lambda_j} \cdot \frac{v_{ij}}{\sigma_i}} \tag{2.3}$$

   onde:

   - σ_i – é o desvio padrão da variável $\mathbf{x}_i$
   - v_{ij} – é o coeficiente de $\mathbf{x}_i$ na combinação linear que define a j-ésima CP, $\mathbf{X}\mathbf{v}_j$
   - λ_j – é a variância da componente $\mathbf{X}\mathbf{v}_j$ (j-ésimo valor próprio de $\boldsymbol{\Sigma}$)

De facto, tem-se:

$$\begin{aligned}\text{corr}(\mathbf{x}_i, \mathbf{X}\mathbf{v}_j) &= \frac{\frac{1}{n}<\mathbf{x}_i, \mathbf{X}\mathbf{v}_j>}{\frac{1}{\sqrt{n}}\|\mathbf{x}_i\| \cdot \frac{1}{\sqrt{n}}\|\mathbf{X}\mathbf{v}_j\|} \\ &= \frac{\frac{1}{n}\mathbf{x}_i^t \mathbf{X}\mathbf{v}_j}{\sqrt{\frac{1}{n}\mathbf{x}_i^t\mathbf{x}_i \cdot \frac{1}{n}\mathbf{v}_j^t\mathbf{X}^t\mathbf{X}\mathbf{v}_j}}\end{aligned}$$

Escrevendo $\mathbf{x}_i = \mathbf{X}\mathbf{e}_i$, onde $\mathbf{e}_i$ é o i-ésimo vector da base canónica de $\mathbb{R}^p$, vem:

$$\text{corr}(\mathbf{x}_i, \mathbf{X}\mathbf{v}_j) = \frac{\mathbf{e}_i^t \boldsymbol{\Sigma} \mathbf{v}_j}{\sqrt{\mathbf{e}_i^t \boldsymbol{\Sigma} \mathbf{e}_i \cdot \mathbf{v}_j^t \boldsymbol{\Sigma} \mathbf{v}_j}}$$

Como $\boldsymbol{\Sigma}\mathbf{v}_j = \lambda_j \mathbf{v}_j$ e $\mathbf{v}_j^t \mathbf{v}_j = 1$, temos:

$$\text{corr}(\mathbf{x}_i, \mathbf{X}\mathbf{v}_j) = \frac{\lambda_j \cdot \mathbf{e}_i^t \mathbf{v}_j}{\sqrt{\sigma_i^2 \cdot \lambda_j}} = \sqrt{\lambda_j} \cdot \frac{v_{ij}}{\sigma_i}$$

∇

Observação: *A covariância entre* $\mathbf{x}_i$ *e* $\mathbf{X}\mathbf{v}_j$ é apenas $\frac{1}{n} < \mathbf{x}_i, \mathbf{X}\mathbf{v}_j >= \lambda_j v_{ij}$.

4. **A representação do i-ésimo indivíduo (i-ésima linha de X) na componente $\mathbf{X}\mathbf{v}_j$ é o i-ésimo coeficiente desse vector $\mathbf{X}\mathbf{v}_j$. No espaço q-dimensional em $\mathbb{R}^n$, gerado pelas q primeiras CPs (colunas de $\mathbf{XV}_q$), o i-ésimo indivíduo é representado por um ponto cujas coordenadas (em relação à base representada pelas CPs) são dadas pela i-ésima linha de $\mathbf{XV}_q$.**

5. Na literatura de ACP em inglês é hábito designar os coeficientes das combinações lineares que definem as CPs (isto é, os coeficientes dos vectores $\mathbf{v}_j$) por *loadings*, e os coeficientes de cada indivíduo numa CP (*i.e.*, os coeficientes dos vectores $\mathbf{X}\mathbf{v}_j$) por *scores*.

2.3.2 ACP e Regressão Múltipla

A ACP é uma técnica Multivariada que, tal como uma Regressão Múltipla, envolve várias variáveis quantitativas. Importa salientar duas diferenças fundamentais da ACP em relação à Regressão Linear Múltipla:

- Ao contrário do que acontece na Regressão, na ACP não há uma variável "dependente" que se pretende aproximar por outras. Na ACP *todas* as p variáveis observadas têm à partida um papel análogo, e todas serão representadas no espaço q-dimensional.
- Na ACP o subespaço sobre o qual se projecta não está definido à partida, sendo a sua identificação a essência do problema, enquanto que na Regressão o subespaço sobre o qual se projecta é o subespaço gerado pelas $p+1$ variáveis $\mathbf{x}_i$ preditoras.

Quando existem apenas duas variáveis ($p = 2$), o primeiro eixo principal em $\mathbb{R}^p$ define a recta que *minimiza a soma dos quadrados das distâncias na perpendicular* entre os n pontos e a sua projecção na

recta (é esta a interpretação do critério geométrico utilizado para introduzir a ACP na Secção 2.1 quando $p = 2$). A recta de regressão linear, recorde-se, minimiza a soma dos quadrados das distâncias *verticais*, após definir uma das variáveis como a preditora e a outra (a colocar no eixo vertical) como variável resposta.

2.3.3 Problemas numa ACP

A Análise em Componentes Principais, introduzida através de duas abordagens diferentes (isto é, como solução de dois diferentes problemas) tem, além de numerosas virtudes e vantagens, alguns problemas. Chama-se desde já a atenção para *três problemas importantes* das Componentes Principais:

1. As Componentes Principais definem-se como combinação linear de p variáveis originais. Não faz muito sentido que qualquer dessas variáveis seja uma variável *qualitativa* (*categórica*), mesmo que as respectivas categorias tenham sido sujeitas a uma qualquer codificação.

2. Caso as p variáveis não tenham as mesmas unidades de medida, a combinação linear é insensata do ponto de vista "físico" (mistura "alhos com bugalhos").

3. Mais grave, **as CPs *não* são invariantes a mudanças multiplicativas diferenciadas nas escalas das variáveis**. De facto, seja $\mathbf{X} = \mathbf{W\Delta V}^t$ a DVS da matriz de dados $\mathbf{X}$. Já vimos que as componentes principais de $\mathbf{X}$ são as colunas de $\mathbf{W\Delta}$. Multiplicar cada coluna de $\mathbf{X}$ por um escalar corresponde a pós-multiplicar $\mathbf{X}$ pela matriz diagonal desses escalares: $\mathbf{X} \longrightarrow \mathbf{XD}$ (ver página 5). Mas $\mathbf{D}^t\mathbf{V}$ não é uma matriz de colunas ortonormadas (a menos que $\mathbf{D} = \alpha\mathbf{I}$ para $\alpha = \pm 1$), pelo que $\mathbf{XD} = \mathbf{W\Delta V}^t\mathbf{D}$ não é uma DVS de $\mathbf{X}$, ou seja, as colunas de $\mathbf{W\Delta}$ já não são as CPs de $\mathbf{XD}$. A implicação prática deste facto é que as componentes principais de dados medidos com o Sistema Métrico e as componentes principais dos mesmos dados expressos em, digamos, o Sistema Imperial britânico de unidades, não serão iguais. Tal facto levanta várias questões e é, talvez a faceta menos simpática da ACP. Mas é uma consequência natural e inevitável do seguinte facto: ao multiplicar-se diferentes variáveis por diferentes constantes de escala, a forma da nuvem de pontos em $\mathbb{R}^p$ (e o comprimento dos p vectores no feixe de vectores em $\mathbb{R}^n$) altera-se. Esta alteração tem de conduzir a diferentes resultados, tendo em vista os critérios subjacentes à ACP.

2.4 ACP sobre a Matriz de Correlações

Para tentar obviar aos problemas referidos na Secção 2.3, sugere-se frequentemente que (sobretudo quando as p variáveis não têm todas as mesmas unidades de medida) a ACP seja feita, não sobre as variáveis centradas, mas sobre as **variáveis centradas e reduzidas**, ou seja, que aos valores observados de cada variável seja subtraído o seu valor médio e que se divida pelo seu desvio padrão. Designando por y_{ij} o i-ésimo valor observado na j-ésima variável, esta abordagem consiste em trabalhar com a transformação:

$$\mathbf{y}_{ij} \longrightarrow \mathbf{z}_{ij} = \frac{\mathbf{y}_{ij} - \overline{y}_{\cdot j}}{\sigma_j}$$

Em termos geométricos, **na representação em $\mathbb{R}^n$, isto corresponde a dizer que cada um dos p vectores considerados até aqui será re-dimensionado de forma a ter o comprimento (norma) comum $\sqrt{n}$, *i.e.*, de forma a ter variância 1**. Em $\mathbb{R}^p$, a configuração da nuvem de n pontos sofre uma transformação muito mais complexa, sendo (além da translação do centro de gravidade da nuvem para a origem do referencial) cada eixo esticado (se o desvio padrão da variável correspondente fosse inferior a 1) ou contraído (se $\sigma_i > 1$), com factores de alteração das escalas diferenciados para cada eixo.

Em termos práticos, as consequências desta transformação prévia são duas:

1. **As componentes agora são combinações lineares das variáveis *reduzidas*, em que os coeficientes são dados pelos vectores próprios da matriz de correlações R** (que é a matriz de variâncias-covariâncias das variáveis reduzidas). Tal facto justifica que se designe uma ACP nestas condições como uma **ACP sobre a Matriz de Correlações**.

2. **As componentes resultantes optimizam os critérios anteriores para as variáveis reduzidas, mas já não para as variáveis originais**. Não existe relação directa entre as CPs duma ACP sobre a matriz das Covariâncias e as CPs duma ACP sobre a matriz das Correlações.

Observações:

1. No caso duma ACP sobre uma matriz de correlações $\mathbf{R}$, o denominador das proporções ou percentagens de "variabilidade explicada" será $\mathrm{tr}(\mathbf{R}) = \mathrm{p}$, *i.e.*, o número de variáveis analisadas.

2. **Numa ACP sobre uma matriz de correlações R, a correlação entre a variável $\mathbf{x}_i$ e a j-ésima CP, dada na equação (2.3), é dada apenas por $\sqrt{\lambda_j}\, v_{ij}$** (onde λ_j e v_{ij} indicam um valor próprio e a coordenada dum vector próprio de $\mathbf{R}$).

3. Devido à questão indicada na alínea anterior, *os coeficientes das componentes numa ACP sobre a matriz de correlações são por vezes re-escalados de forma a que $\mathbf{v}_j^t \mathbf{v}_j = \lambda_j$ (e não $\mathbf{v}_j^t \mathbf{v}_j = 1$). Nesse caso, os coeficientes da combinação linear são as correlações entre a variável e a CP em causa.*

Uma Análise em Componentes Principais baseada na matriz de Correlações dos dados é vivamente aconselhada quando as variáveis observadas estão expressas em diferentes unidades de medida (não compatíveis). Mas, como se viu, esta abordagem também não está isenta de problemas na sua interpretação e justificação. Felizmente, existe outra forma, menos conhecida, de justificar a utilização de uma ACP sobre a matriz de Correlações dos dados, que será discutida na Secção seguinte.

2.5 Outro critério para a ACP sobre a matriz de Correlações

As CPs baseadas numa matriz de correlações são também solução óptima de um outro problema – diferente do problema de maximizar variâncias que define a ACP.

Seja $\mathbf{Xv}$ uma qualquer combinação linear das colunas da matriz de dados centrada $\mathbf{X}$ (não necessariamente uma componente principal). A correlação entre $\mathbf{Xv}$ e a variável $\mathbf{x}_i$ é dada por:

$$\begin{aligned} \text{corr}(\mathbf{x}_i, \mathbf{Xv}) &= \frac{<\mathbf{x}_i, \mathbf{Xv}>}{\|\mathbf{x}_i\| \cdot \|\mathbf{Xv}\|} \\ &= \frac{\mathbf{x}_i^t \mathbf{Xv}}{\sqrt{\mathbf{x}_i^t \mathbf{x}_i} \cdot \sqrt{\mathbf{v}^t \mathbf{X}^t \mathbf{Xv}}} \end{aligned}$$

As p correlações entre as p variáveis $\mathbf{x}_i$ e a combinação linear $\mathbf{Xv}$ são os p elementos do vector:

$$\begin{aligned} r &= \frac{\frac{1}{\sqrt{n}} \mathbf{D}^{-1} \mathbf{X}^t \mathbf{Xv}}{\sqrt{\mathbf{v}^t \mathbf{X}^t \mathbf{Xv}}} \quad \text{onde } \mathbf{D}^{-1} \text{ é a matriz diagonal dos recíprocos} \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{dos desvios-padrão das } p \text{ variáveis} \\ &= \frac{\mathbf{D}^{-1} \mathbf{\Sigma v}}{\sqrt{\mathbf{v}^t \mathbf{\Sigma v}}} \end{aligned}$$

A soma dos quadrados destas p correlações é:

$$\|r\|^2 = \frac{\mathbf{v}^t \mathbf{\Sigma} \mathbf{D}^{-1} \cdot \mathbf{D}^{-1} \mathbf{\Sigma v}}{\mathbf{v}^t \mathbf{\Sigma v}}$$

Escrevendo, sem perda de generalidade, $\mathbf{v} = \mathbf{D}^{-1}\mathbf{b}$, tem-se:

$$\|r\|^2 = \frac{\mathbf{b}^t \mathbf{R}^2 \mathbf{b}}{\mathbf{b}^t \mathbf{R} \mathbf{b}}$$

onde $\mathbf{R} = \mathbf{D}^{-1}\mathbf{\Sigma}\mathbf{D}^{-1}$ é a matriz de correlações das p variáveis, como foi visto na página 48.

Vejamos agora qual a *combinação linear que maximiza a soma de quadrados das p correlações com cada variável original*, *i.e.*, qual a combinação linear $\mathbf{Xv}$ "globalmente mais correlacionada com as p variáveis", Trata-se dum critério diferente dos utilizados para definir Componentes Principais. O problema a resolver é o de determinar:

$$\max_{\mathbf{b} \in \mathbb{R}^p} \frac{\mathbf{b}^t \mathbf{R}^2 \mathbf{b}}{\mathbf{b}^t \mathbf{R} \mathbf{b}}. \tag{2.4}$$

Mas este novo problema é um caso particular do problema geral de maximização de um quociente de formas quadráticas, problema estudado no Teorema 1.38 (p.38). Sabemos assim que **a solução do problema consiste em considerar a combinação linear**

$$\mathbf{Xv} = \mathbf{X}\mathbf{D}^{-1}\mathbf{b}_1$$

onde $\mathbf{b}_1$ é o vector próprio associado ao maior valor próprio da matriz de correlações R, uma vez que a matriz $\mathbf{B}^{-1}\mathbf{A}$ referida no Teorema 1.38 é aqui dada por $\mathbf{R}^{-1}\mathbf{R}^2 = \mathbf{R}$.

Observação: A soma de quadrados das correlações entre esta nova variável e as p variáveis originais é o valor próprio λ_1 associado ao vector próprio $\mathbf{b}_1$.

Mas esta solução é a primeira componente principal das variáveis normalizadas z (pois $\mathbf{XD}^{-1}$ tem nas suas colunas as variáveis normalizadas), *i.e.*, é a primeira CP sobre a matriz de correlações.

Sabemos também, a partir do Teorema 1.38, que as restantes CPs da matriz de correlações são as combinações lineares que sucessivamente maximizam o quociente acima indicado, sujeitas às restrições de ortogonalidade com todas as soluções anteriores (note-se que, neste caso concreto, a matriz $\mathbf{B}^{-1}\mathbf{A}$ referida no Teorema 1.38 é a matriz de correlações $\mathbf{R}$, logo é uma matriz simétrica).

Este resultado não é apenas mais um critério optimizado por Componentes Principais. A importância deste resultado reside no facto do **critério não depender das unidades de medida das variáveis originais e ser portanto invariante face a mudanças lineares de escala nas variáveis**. Esta afirmação é facilmente confirmável se tivermos em conta que o critério é uma soma de quadrados de correlações entre cada combinação linear e as variáveis originais, e que correlações não se alteram (a não ser talvez no sinal) para transformações afins das variáveis.

Por outras palavras, **as combinações lineares das variáveis reduzidas $\mathbf{XD}^{-1}\mathbf{b}$ (onde b é vector próprio de R) só são componentes *principais* das variáveis reduzidas e não das variáveis nas suas unidades de medida originais. Mas são as combinações lineares sucessivamente "globalmente mais correlacionadas com as variáveis originais", independentemente das unidades de medida originais**.

O resultado que acabamos de referir tem uma **interpretação geométrica** intuitiva. Em $\mathbb{R}^n$, onde cada variável é representada por um vector, as variáveis do tipo $\mathbf{Xv}$ são os *vectores resultantes* das combinações lineares que os definem. Na ACP as componentes são determinadas pelo critério de maximizar a variância (proporcional ao quadrado da norma) da resultante. Este critério é, como vimos, sensível a mudanças de escala diferenciadas nas variáveis, pois o efeito dessas mudanças de escala é o de esticar/encolher os p vectores originais em $\mathbb{R}^n$, de forma diferenciada, o que transformará a própria direcção dos vectores resultantes.

Mas os *ângulos* entre vectores permanecem inalterados quando estes são apenas esticados/contraídos. Logo, **critérios baseados na dimensão de ângulos (cujos cossenos, recorde-se, são as correlações entre as variáveis) são insensíveis a mudanças de escala das variáveis**. É isso que acontece com o critério (2.4) da página 66.

2.6 Três advertências sobre ACP

Para completar esta introdução à Análise em Componentes Principais, refiram-se três aspectos associados à ACP onde, por vezes, surgem confusões e dúvidas.

1. **A redução da dimensionalidade introduzida pela ACP não significa redução no número de *variáveis originais* com que se trabalha**. Mesmo que venhamos a reter apenas q CPs, precisamos da totalidade das p variáveis originais para definir essas q CPs. Neste sentido, a "redução da dimensionalidade" é uma "falsa redução".

2. Por vezes procura-se **interpretar** cada CP em termos de algum subconjunto (de pequena dimensão) das p variáveis originais. Nesse sentido, é frequente ignorar as variáveis cujos coeficientes (*loadings*) na combinação linear que define a CP são "próximos de zero". *Tal prática pode induzir em erro*, pois a resultante duma combinação linear não depende apenas dos coeficientes, mas também do comprimento (variância) e da posição relativa (padrão de correlações) das variáveis que definem a combinação linear. Para um exemplo particularmente elucidativo de tais problemas, veja-se o Exercício 6 (página 81). Quando se efectuam tais interpretações, convém utilizar informação complementar para validar as interpretações baseadas nos coeficientes.

3. Outra prática frequente, mas discutível no contexto da ACP, é a da **rotação** das CPs. Tal prática altera os vectores de coeficientes das combinações lineares procurando tornar o maior número possível desses coeficientes próximos de zero para "simplificar a interpretação". Mas, como vimos na advertência anterior, esse objectivo pode ser ilusório. Além disso, as técnicas de rotação das Componentes Principais sacrificam a esse objectivo (potencialmente ilusório) de simplificar a interpretabilidade das CPs as características que nos levaram, desde logo, a procurar CPs: qualquer transformação das Componentes ou Eixos Principais já não resultará nos eixos que sucessivamente optimizam os problemas de representação mais fidedigna que estavam na origem do método.

2.7 Biplots

Intimamente relacionada com a Decomposição em Valores Singulares duma matriz centrada de dados (logo, com uma ACP), a técnica do *biplot* foi divulgada por Gabriel[1]. A ideia fundamental do *biplot* consiste em obter uma **representação simultânea dos indivíduos e das variáveis, num espaço de baixa dimensão, que forneça informação útil para a compreensão da relação entre indivíduos e variáveis, e entre estes e as Componentes Principais dos dados**.

Considere a DVS duma matriz centrada de dados, $\mathbf{X} = \mathbf{W}\mathbf{\Delta}\mathbf{V}^t$, que admitiremos ser de característica k. Considere as potências $\mathbf{\Delta}^\alpha$ da matriz diagonal $\mathbf{\Delta}$ (veja a discussão na página 33). Para qualquer valor $0 \leq \alpha \leq 1$, a decomposição DVS pode ser escrita na forma:

$$\mathbf{X} = \mathbf{W}\mathbf{\Delta}^\alpha\mathbf{\Delta}^{1-\alpha}\mathbf{V}^t \tag{2.5}$$

Definindo:

$$\mathbf{G} = \mathbf{W}\mathbf{\Delta}^\alpha \tag{2.6}$$

$$\mathbf{H} = \mathbf{V}\mathbf{\Delta}^{1-\alpha} \tag{2.7}$$

tem-se:

$$\mathbf{X} = \mathbf{G}\mathbf{H}^t \tag{2.8}$$

Repare-se que a matriz G é de tipo $n \times k$, ou seja existe uma correspondência entre as linhas de G e os indivíduos observados. Por seu turno, a matriz H é de tipo $p \times k$, e existe uma correspondência entre linhas de H e variáveis observadas.

[1] The biplot graphic display of matrices with application to principal component analysis, *Biometrika*, **58**, 453-467.

A equação (2.8) corresponde a dizer que **o elemento genérico (i, j) da matriz centrada de dados X é dada pelo produto interno da i-ésima linha de G com a j-ésima coluna de $\mathbf{H}^t$, isto é, a j-ésima linha de H**. Assim:

$$x_{ij} \quad = \quad < \mathbf{g}_i, \mathbf{h}_j > \quad = \quad \mathbf{g}_i^t \mathbf{h}_j \tag{2.9}$$

onde $\mathbf{g}_i$ designa a i-ésima linha da matriz $\mathbf{G}$ e $\mathbf{h}_j$ designa a j-ésima linha da matriz $\mathbf{H}$. Cada um dos vectores $\mathbf{g}_i$ ou $\mathbf{h}_j$ é um vector k-dimensional. Assim, no espaço $\mathbb{R}^k$ é possível representar cada indivíduo pela linha da matriz $\mathbf{G}$ que lhe está associada, e cada variável pela linha da matriz $\mathbf{H}$ correspondente, de tal forma que o produto interno entre estes representantes seja igual ao elemento x_{ij} da matriz $\mathbf{X}$.

Na prática, esta representação gráfica exacta não será possível, a não ser que $\mathbf{X}$ tenha característica $k = 2$ ou $k = 3$. No entanto, é possível **representar graficamente em $\mathbb{R}^2$ (ou até em $\mathbb{R}^3$) os indivíduos com marcadores cujas coordenadas são apenas os 2 (ou 3) primeiros elementos de cada vector $\mathbf{g}_i$, e *no mesmo sistema de eixos*, representar as variáveis com coordenadas dadas pelos primeiros elementos de $\mathbf{h}_j$**. É esta representação que se designa o **biplot**. **Os subvectores $\mathbf{g}_i^{(r)}$, $\mathbf{h}_j^{(r)}$, constituídos pelas r primeiras coordenadas dos vectores $\mathbf{g}_i$ e $\mathbf{h}_j$ designam-se marcadores no biplot, respectivamente dos indivíduos e das variáveis**. O *biplot* está associado à re-constituição aproximada, de característica $k = 2$ ou 3 da matriz $\mathbf{X}$, dada pela aproximação DVS já estudada no Teorema 1.42 (página 44). De facto, a matriz $\tilde{\mathbf{X}}$ que resulta de tomar o produto $\mathbf{G}^{(r)}\mathbf{H}^{(r)^t}$, onde as matrizes $\mathbf{G}^{(r)}$ e $\mathbf{H}^{(r)}$ resultam de reter apenas as r primeiras colunas das matrizes $\mathbf{G}$ e $\mathbf{H}$, é a melhor aproximação a $\mathbf{X}$, de caracteristica r, segundo foi visto nesse Teorema.

Para melhor compreensão do *biplot* e da sua utilidade, analisemos em mais pormenor o significado de um *biplot* correspondente a tomar $\alpha = 0$ na definição das matriz $\mathbf{G}$ e $\mathbf{H}$, que é a opção mais frequente.

Biplot com $\alpha = 0$

Se na definição das matrizes $\mathbf{G}$ e $\mathbf{H}$ (equações 2.6 e 2.7) se tomar $\alpha = 0$, fica-se com

$$\mathbf{G} \quad = \quad \mathbf{W} \quad \text{e} \quad \mathbf{H} \quad = \quad \mathbf{V\Delta} \tag{2.10}$$

Neste caso, verificam-se as seguintes particularidades:

1. **O produto interno entre cada par de linhas da matriz H é proporcional à covariância entre as variáveis correspondentes**.

2. **A norma de cada linha da matriz H é proporcional ao desvio padrão da variável correspondente**.

3. **O cosseno do ângulo entre cada par de linhas da matriz H é o coeficiente de correlação entre as variáveis correspondentes**.

4. **A distância euclidiana entre cada par de linhas da matriz G é proporcional à distância de Mahalanobis entre os indivíduos correspondentes**.

5. **A projecção ortogonal dos marcadores de indivíduos sobre o subespaço gerado pelo marcador da variável j é proporcional aos valores dos indivíduos na variável j (isto é, aos elementos x_{ij} da matriz original, X.**

De facto,

1. a matriz de covariâncias das variáveis observadas é dada por:
$$\frac{1}{n}\mathbf{X}^t\mathbf{X} \quad = \quad \frac{1}{n}\mathbf{H}\mathbf{G}^t\mathbf{G}\mathbf{H}^t \quad = \quad \frac{1}{n}\mathbf{H}\mathbf{H}^t \tag{2.11}$$
Assim, a covariância entre a variável i e a variável j é dada por $\frac{1}{n}$ vezes o produto interno da linha i e linha j da matriz $\mathbf{H}$ (coluna j de $\mathbf{H}^t$).

2. Imediato a partir do que se viu no ponto anterior, considerando os elementos diagonais da matriz.

3. Imediato, a partir da primeira alínea, tendo em atenção a definição de cosseno de ângulo e de coeficiente de correlação.

4. A distância de Mahalanobis entre o indivíduo i (o vector da linha i da matriz $\mathbf{X}$, que designaremos por $\mathbf{x}_{(i)}$ e que será tratado como um vector-coluna) e o indivíduo j (linha j da matriz $\mathbf{X}$, $\mathbf{x}_{(j)}$, na forma de vector-coluna) é dada por $(\mathbf{x}_{(i)} - \mathbf{x}_{(j)})^t\mathbf{\Sigma}^{-1}(\mathbf{x}_{(i)} - \mathbf{x}_{(j)})$, onde $\mathbf{\Sigma}$ indica a matriz de variâncias-covariâncias associada à matriz $\mathbf{X}$. Tendo em atenção que $\mathbf{X} = \mathbf{G}\mathbf{H}^t$, tem-se que a linha genérica i da matriz $\mathbf{X}$ é dada por $\mathbf{x}_{(i)}^t = \mathbf{g}_i^t\mathbf{H}^t$ (onde $\mathbf{g}_i$ é a linha i de $\mathbf{G}$). Assim, a distância de Mahalanobis entre $\mathbf{x}_{(i)}$ e $\mathbf{x}_{(j)}$ é dada por:
$$\begin{aligned}
d_M(\mathbf{x}_{(i)}, \mathbf{x}_{(j)}) &= (\mathbf{x}_{(i)} - \mathbf{x}_{(j)})^t\mathbf{\Sigma}^{-1}(\mathbf{x}_{(i)} - \mathbf{x}_{(j)}) \\
&= (\mathbf{g}_i - \mathbf{g}_j)^t\mathbf{H}^t\left(\frac{1}{n}\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{H}(\mathbf{g}_i - \mathbf{g}_j) \\
&= n\,(\mathbf{g}_i - \mathbf{g}_j)^t\mathbf{H}^t\left(\mathbf{H}\mathbf{H}^t\right)^{-1}\mathbf{H}(\mathbf{g}_i - \mathbf{g}_j) \\
&= n\,(\mathbf{g}_i - \mathbf{g}_j)^t\mathbf{\Delta}\mathbf{V}^t\left(\mathbf{V}\mathbf{\Delta}^2\mathbf{V}^t\right)^{-1}\mathbf{V}\mathbf{\Delta}(\mathbf{g}_i - \mathbf{g}_j) \\
&= n\,(\mathbf{g}_i - \mathbf{g}_j)^t\mathbf{\Delta}\mathbf{V}^t\left(\mathbf{V}\mathbf{\Delta}^{-2}\mathbf{V}^t\right)\mathbf{V}\mathbf{\Delta}(\mathbf{g}_i - \mathbf{g}_j) \\
&= n\,(\mathbf{g}_i - \mathbf{g}_j)^t\mathbf{\Delta}\mathbf{\Delta}^{-2}\mathbf{\Delta}(\mathbf{g}_i - \mathbf{g}_j) \\
&= n\,(\mathbf{g}_i - \mathbf{g}_j)^t(\mathbf{g}_i - \mathbf{g}_j)
\end{aligned} \tag{2.12}$$
A antepenúltima e penúltima passagens resultam da discussão de potências de matrizes simétricas (página 33) e da ortonormalidade das colunas da matriz $\mathbf{V}$, respectivamente.

5. A projecção ortogonal da linha $\mathbf{g}_i$ de $\mathbf{G}$ sobre a linha j de $\mathbf{H}$, $\mathbf{h}_j$, é dada por $\mathbf{h}_j(\mathbf{h}_j^t\mathbf{h}_j)^{-1}\mathbf{h}_j^t\mathbf{g}_i = \mathbf{h}_j(n\sigma_j^2)^{-1}\mathbf{h}_j^t\mathbf{g}_i$, uma vez que $\mathbf{h}_j^t\mathbf{h}_j$ é o j-ésimo elemento diagonal de $\mathbf{H}\mathbf{H}^t = \mathbf{X}^t\mathbf{X}$ (pela equação 2.11). Logo, o coeficiente de projecção do indivíduo i é $\frac{1}{n\sigma_j^2}\mathbf{h}_j^t\mathbf{g}_i$ que é, para cada variável j, proporcional a $\mathbf{h}_j^t\mathbf{g}_i$, ou seja, proprocional ao elemento (i, j) de $\mathbf{X}$, como se viu na equação (2.9).

Exercício 2.1 *Não seria em geral correcto partir, na derivação anterior, da expressão* $\mathbf{H}^t\left(\mathbf{H}\mathbf{H}^t\right)^{-1}\mathbf{H}$ *e escrever:* $\mathbf{H}^t\left(\mathbf{H}\mathbf{H}^t\right)^{-1}\mathbf{H} = \mathbf{H}^t(\mathbf{H}^t)^{-1}\mathbf{H}^{-1}\mathbf{H} = \mathbf{I}$ *(o que, aparentemente, produziria o resultado desejado). Diga porquê.*

Recorde-se que, no *biplot*, apenas se utilizam as 2 ou 3 primeiras coordenadas de cada vector $\mathbf{g}_i$ e $\mathbf{h}_j$. Assim, o resultado anterior indica-nos que, **num biplot com $\alpha = 0$,**

1. **o cosseno do ângulo entre cada par de marcadores de variáveis é uma aproximação do coeficiente de correlação entre essas variáveis**;
2. **o comprimento (euclidiano) desses mesmos marcadores é aproximadamente proporcional ao desvio padrão da correspondente variável**;
3. **a distância euclidiana entre cada par de marcadores de individuos é, aproximadamente, a distância de Mahalanobis entre os correspondentes individuos**;
4. **o produto interno entre um marcador de individuo e um marcador de variável, $\left(\mathbf{g}_i^{(r)}\right)^t \mathbf{h}_j^{(r)}$ é, aproximadamente, o valor que o individuo i toma na variável j (x_{ij}).**

Biplot com $\alpha = 1$

Se na definição das matrizes $\mathbf{G}$ e $\mathbf{H}$ (equações 2.6 e 2.7) se tomar $\alpha = 1$, fica-se com

$$\mathbf{G} = \mathbf{W\Delta} \quad \text{e} \quad \mathbf{H} = \mathbf{V} \tag{2.13}$$

Neste caso, **a linha genérica i da matriz G contêm os *scores* do individuo i nas Componentes Principais definidas pela matriz X, enquanto que a linha genérica j da matriz H é dada pelos coeficientes (*loadings*) da variável j em cada CP**. Neste caso, o *biplot* será constituido por:

1. **a nuvem de n pontos projectada no subespaço principal de dimensão r** (isto é, os marcadores de individuos corresponderão à representação r-dimensional óptima dos individuos, tal como fornecida pela representação gráfica da ACP anteriormente discutida);
2. **as distâncias euclidianas entre marcadores de individuos serão, aproximadamente, as distâncias euclidianas entre os individuos no espaço $\mathbb{R}^p$**;
3. **um feixe de vectores (marcadores de variáveis)** de interpretação menos útil quer no caso $\alpha = 0$;
4. o produto interno entre marcadores de individuos e de variáveis continua (como para qualquer escolha de α) a dar uma reconstituição aproximada de x_{ij}.

2.8 Um exemplo

Considere-se brevemente um exemplo, e o respectivo estudo no programa estatístico R. O principal comando para efectuar uma Análise em Componentes Principais no `R` é o comando `prcomp`, que permite efectuar uma Análise em Componentes Principais através da Decomposição em Valores Singulares duma matriz centrada de dados[2].

Consideremos agora um conjunto de dados relativos às medições de $p = 4$ variáveis sobre $n = 150$ lírios, sendo as variáveis o comprimento e largura das sépalas e das pétalas das flores[3]. Esses dados estão disponíveis nas distribuições padrão do `R`, numa *data frame* de nome `iris`.

O objecto `iris` é uma *data frame* com 150 linhas e cinco colunas, sendo a quinta e última coluna uma variável qualitativa (factor) indicando a espécie de lírio de cada flor individual. Foram consideradas 50 flores de cada uma de três espécies: *setosa*, *versicolor* e *virginica*.

Uma primeira visualização, que é útil para conjuntos de dados multivariados (desde que o número de variáveis observadas não seja excessivo), é dada pela matriz de gráficos de pontos para cada par de variáveis. Para *data frames*, é possível obtê-la através dum único comando, o comando `plot`. Aplicado às quatro variáveis numéricas do objecto `iris`, tem-se:

```
> plot(iris[,-5])
```

A fim de visualizar também as espécies de cada lírio, pode colorir-se de forma diferente os pontos de cada espécie, através do seguinte comando, obtendo-se o resultado da Figura 2.2.

```
> plot(iris[,-5],col=as.numeric(iris[,5]), pch=16)
```

Importa sublinhar que a representação gráfica da Figura 2.2 não é mais do que uma colecção de projecções ortogonais da nuvem de $n = 150$ pontos sobre os 6 planos coordenados definidos pelos possíveis pares de variáveis. A realidade da nuvem de $n = 150$ pontos em $p = 4$ variáveis definida pelas quatro primeiras colunas do objecto `iris` é mais complexa do que se pode inferir a partir dessa Figura.

Embora a estrutura dos gráficos da Figura 2.2 seja relativamente simples e de fácil interpretação (o que resulta, em parte, do número não muito elevado de variáveis observadas), ilustremos alguns aspectos duma ACP sobre estes dados.

Uma simples invocação do comando `prcomp` com a matriz de dados `iris[,-5]` produz os seguintes resultados:

```
> prcomp(iris[,-5])
```

[2]Para mais pormenores sobre este comando e as suas opções, dê, numa sessão de trabalho do R, o comando `help(prcomp)`. Veja também o comando `princomp`, que efectua uma ACP utilizando a decomposição espectral da matriz de covariâncias (ou correlações) dos dados.

[3]Estes dados já foram considerados na disciplina de Modelação Estatística I.

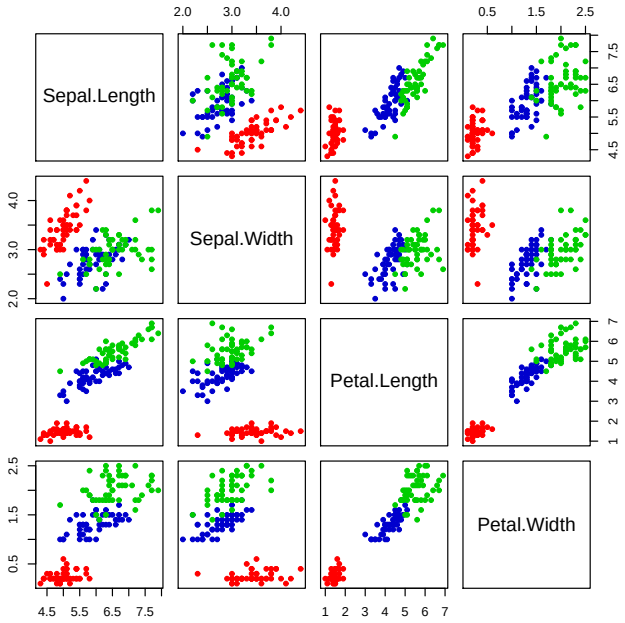


Figura 2.2: Quatro variáveis morfométricas observadas em 50 lírios de cada uma de três espécies.

```
Standard deviations:
[1] 2.0562689 0.4926162 0.2796596 0.1543862

Rotation:
                    PC1         PC2         PC3         PC4
Sepal.Length  0.36138659 -0.65658877  0.58202985  0.3154872
Sepal.Width  -0.08452251 -0.73016143 -0.59791083 -0.3197231
Petal.Length  0.85667061  0.17337266 -0.07623608 -0.4798390
Petal.Width   0.35828920  0.07548102 -0.54583143  0.7536574
```

São aqui visíveis os valores singulares (`Standard deviations`) e vectores singulares direitos (`Rotation`) da matriz. Recorde-se que os vectores singulares direitos (que são também os vectores próprios da matriz de variâncias/covariâncias dos dados) dão-nos os coeficientes da combinação linear das variáveis originais (centradas!) que definem cada CP. No nosso caso, a primeira CP dos dados dos lírios é definido pela seguinte combinação linear:

$$Z_1 = 0.36139 * Sepal.Length - 0.08452 * Sepal.Width + 0.85667 * Petal.Length + 0.35829 * Petal.Width$$

(De novo, recorde-se que se trata de combinações lineares das variáveis *centradas*, pelo que o valor associado a cada uma das $n = 150$ observações nesta CP é obtido substituíndo na equação os valores das variáveis *menos a média das observações dessa variável*, para cada observação). Os valores assim

obtidos, também conhecidos em inglês por *scores* de cada observação na CP, são calculados pelo comando `prcomp`, embora só sejam apresentados quando tal fôr explicitamente solicitado (devido ao número, em geral, elevado de indivíduos observados). Concretamente, o comando `prcomp` produz um objecto de saída cuja componente `x` contém esses valores dos *scores*. No caso dos lírios (e apresentando apenas parte dos resultados), tem-se:

```
> prcomp(iris[,-5])$x
..... (corte) ....
  146  1.944109795 -0.18753230 -0.177825091  0.4261959400
  147  1.527166615  0.37531698  0.121898172  0.2543674420
  148  1.764345717 -0.07885885 -0.130481631  0.1370012739
  149  1.900941614 -0.11662796 -0.723251563  0.0445953047
  150  1.390188862  0.28266094 -0.362909648 -0.1550386282
```

Um gráfico das coordenadas associadas a cada um dos $n = 150$ indivíduos nas duas primeiras componentes principais produz **a representação bi-dimensional mais fidedigna da nuvem de $n = 150$ pontos em $\mathbb{R}^4$** definida pela matriz `iris[,-5]`. Essa representação bidimensional é dada na Figura 2.3, que foi obtida com o comando:

```
> plot(prcomp(iris[,-5])$x[,1:2],cex=0.8,col=as.numeric(iris[,5]),pch=16)
```

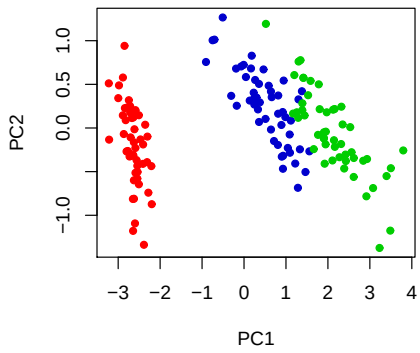


Figura 2.3: Primeiro plano principal definido pelas 150 observações dos lírios. O grupo de observações mais à esquerda é constituido pelas 50 observações da variedade *setosa*. Com alguma sobreposição, o grupo à direita engloba as observações da variedade *versicolor* (na zona mais central) e *virginica* (direita).

A qualidade desta representação bidimensional depende da proporção da inércia total dos dados que é preservada nesta projecção. Essa qualidade pode ser medida através do quociente da soma das variâncias das duas primeiras CPs (a soma dos quadrados dos dois maiores valores singulares da matriz de dados),

a dividir pela soma das variâncias da totalidade das CPs (o traço da matriz de variâncias associada aos dados). Embora se possam efectuar tais contas a partir dos valores acima indicados, o programa R permite obter desde logo esta medida da qualidade da projecção sobre o primeiro plano principal, da seguinte forma:

```
> summary(prcomp(iris[,-5]))
Importance of components:
                          PC1    PC2    PC3     PC4
Standard deviation     2.056 0.4926 0.2797 0.15439
Proportion of Variance 0.925 0.0531 0.0171 0.00521
Cumulative Proportion  0.925 0.9777 0.9948 1.00000
```

Assim, 97.7% da variabilidade total dos dados é preservada pela projecção da nuvem de pontos sobre o subespaço bi-dimensional de $\mathbb{R}^4$ que é gerado pelas duas primeiras CPs. Este resultado muito bom, que em parte é possível devido ao número reduzido de variáveis, significa que a imagem reproduzida na Figura 2.3 é uma imagem muito fidedigna da nuvem de pontos original.

Vale a pena fazer ainda alguns comentários relativos ao conjunto de dados analisado neste exemplo simples e introdutório à ACP.

As três espécies aparecem em grupos relativamente bem separados na projecção dos dados sobre o primeiro plano principal. Tal facto não é uma consequência obrigatória do método, uma vez que a ACP não utilizou em momento algum a informação sobre a espécie de proveniência de cada linha da matriz de dados. A ACP é frequentemente útil para identificar grupos de indivíduos observados que se distingam dos restantes, mas não é um método especificamente dirigido para esse fim. Esse tipo de objectivo seria melhor alcançado com métodos multivariados que se propõem distinguir grupos de indivíduos (como as Análises Classificatória e Discriminante que adiante se estudarão). A utilidade da ACP na separação de grupos de indivíduos será consequência de dois aspectos: (i) a separabilidade desses grupos nas variáveis observadas; e (ii) a coincidência das direcções em que essa separabilidade se possa evidenciar com as direcções de maior variabilidade dos dados, que são as identificadas por uma ACP.

A matriz de variâncias-covariâncias dos dados (cujos vectores próprios são os eixos principais em $\mathbb{R}^4$ acima indicados, e cujos valores próprios indicam a variância dos *scores* das observações quando projectadas sobre cada CP) é dada por:

```
> var(iris[,-5])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length   0.68569351 -0.04243400    1.2743154   0.5162707
Sepal.Width   -0.04243400  0.18997942   -0.3296564  -0.1216394
Petal.Length   1.27431544 -0.32965638    3.1162779   1.2956094
Petal.Width    0.51627069 -0.12163937    1.2956094   0.5810063
```

A observação das variâncias de cada variável observada (os elementos diagonais desta matriz) mostram que uma variável, a variável Comprimento das Pétalas, têm uma variabilidade bastante maior que as

restantes. É da natureza do método ACP que uma variável nessas condições esteja bastante associada à primeira Componente Principal (que identifica a direcção, em $\mathbb{R}^4$, de maior variabilidade dos dados). A confirmação desse facto pode ser obtida calculando os coeficientes de correlação da primeira CP com cada uma das $p = 4$ variáveis, dada pela fórmula (2.3):

```
> cor(iris[,-5],prcomp(iris[,-5])$x)
                    PC1        PC2         PC3         PC4
Sepal.Length  0.8974018 -0.3906044  0.19656672  0.05882002
Sepal.Width  -0.3987485 -0.8252287 -0.38363030 -0.11324764
Petal.Length  0.9978739  0.0483806 -0.01207737 -0.04196487
Petal.Width   0.9665475  0.0487816 -0.20026170  0.15264831
```

Repare-se na correlação quase máxima entre a primeira CP e a variável Comprimento das Pétalas, que confirma essa associação. A correlação também muito elevada da primeira CP com a variável Largura das Pétalas é consequência do facto de as duas medições das Pétalas estarem fortemente correlacionadas entre si (o ângulo entre os vectores que as representam em $\mathbb{R}^{150}$ é quase nulo), pelo que a fortíssima correlação da CP 1 com uma das variáveis implicaria sempre uma forte correlação com a outra. Veja-se a matriz de correlações entre as variáveis originais:

```
> cor(iris[,-5])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

Regressando à análise das correlações entre CPs e variáveis originais, saliente-se a correlação entre a segunda CP e a variável Largura das Sépalas. Esta variável, que está negativamente correlacionada com as outras três variáveis morfométricas (ou seja, o vector que a representa em $\mathbb{R}^{150}$ forma ângulos superiores a 90º com os vectores representando as restantes variáveis) tem uma influência importante na definição da segunda maior direcção de variabilidade dos dados (ortogonal à direcção definida pela CP 1).

Na Figura 2.4 vê-se um *biplot* obtido com o comando `biplot` do `R`, que na sua forma mais simples exige apenas como argumento o resultado duma ACP:

```
> biplot(prcomp(iris[,-5]))
```

Utilize este *biplot* (com[4] $\alpha = 0$) para visualizar as conclusões anteriores, relativas a este conjunto de dados. Em particular, repare-se na tradução gráfica da elevadíssima correlação entre as duas medições

[4] O comando `biplot` admite o parâmetro `scale` para fixar o valor de α. Assinale-se que o valor de `scale` corresponde a $1 - \alpha$. Assim, para obter um *biplot* com $\alpha = 1$ haverá que indicar `scale`$= 0$ no comando `biplot`. Por omissão, o comando toma o valor `scale`$= 1$, isto é, constrói um *biplot* com $\alpha = 0$.

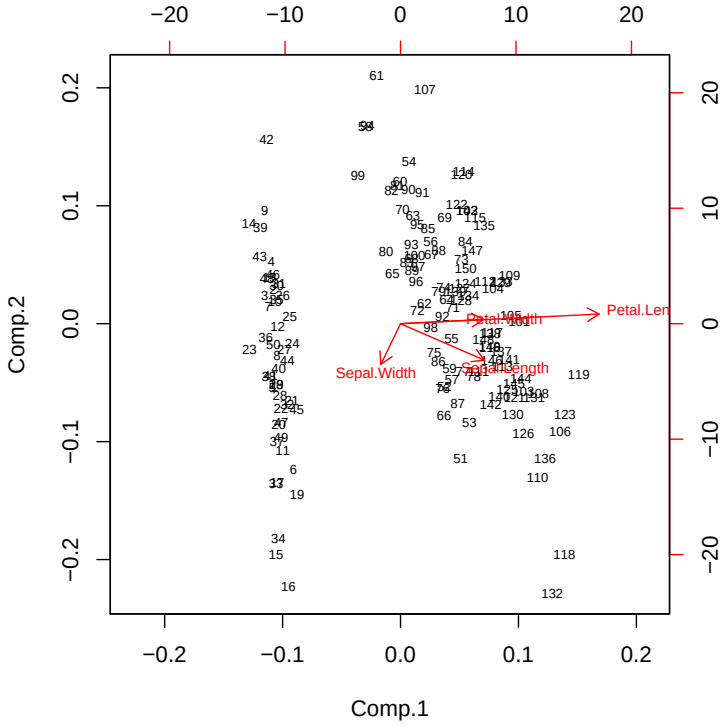


Figura 2.4: *Biplot* das 150 observações quadri-dimensionais dos lírios.

de pétalas (visível no facto de as setas associadas aos respectivos marcadores apontarem quase no mesmo sentido), assim como no facto de a variância da variável `Petal.Length` ser maior que as restantes (visível no maior comprimento do respectivo marcador). Assinale-se que este tipo de conclusão é extensível à relação entre marcadores de variáveis e eixos associados às Componentes Principais 1 (horizontal) e dois (vertical). Assim, o facto de as setas que servem de marcadores das variáveis correspondentes a pétalas serem quase horizontais reflecte a elevadíssima correlação dessas variáveis com a CP 1. A elevada correlação da CP 2 com a Largura das Sépalas é também ilustrada pelo facto de o marcador dessa variável ser aproximadamente vertical.

Também visível é o facto do grupo das *setosa* ter pétalas mais pequenas (a projecção ortogonal dos respectivos marcadores de indivíduos na direcção dos marcadores de Largura e Comprimento de Pétalas fica abaixo da média), da mesma forma que individuos como o 119, 123 e 106 parecem ter as maiores medições de pétalas.

Na leitura deste tipo de conclusões é sempre necessário ter a devida cautela: a representação bidimensional é apenas uma aproximação. A elevada proporção de variabilidade associada ás duas primeiras CPs neste caso (97,7%) é, neste sentido, muito tranquilizadora.

2.9 Exercícios

AVISO: Os conjuntos de dados de vários Exercícios deste Capítulo encontram-se disponíveis numa área de trabalho associada à disciplina, na máquina `prunus` do ISA. Para disponibilizar estes conjuntos de dados deve-se:

- Montar (*Map network drive*, no menu *Tools* do *My Computer*) a *drive*: \\`prunus`\`home`\`cadeiras`
- Abrir uma sessão do `R` (na directoria onde está a guardar o seu trabalho).
- A partir da sessão do `R`, seleccionar a opção *Files*, na barra de menus, e dentro da lista de opções disponibilizada, escolher *Load Workspace*.
- Na janela de diálogo que se abre, seleccionar a nova *drive* (que ficou associada ao `prunus`), depois a directoria `MMACB`, a seguir a directoria `em`, e finalmente o ficheiro `exerACP.RData`.

Se tudo correu bem, na sessão do `R` deverão estar agora disponíveis (confirme com o comando `ls()`) os objectos `lavagantes` (Exercíco 3), `framb` (Exercício 4), `trigo` (Exercício 5), `soil` (Exercício 6), `adelges` (Exercício 7).

1. Construa uma matrix de dimensões 4×3, com uma permutação aleatória dos inteiros de 1 a 12:

   ```
   > A <- matrix(nrow=4,ncol=3,sample(1:12,12))
   ```

   (a) Calcule, com o auxílio da função `svd` do `R`, a Decomposição em Valores Singulares de $\mathbf{A}$.

   (b) Calcule, com o auxílio da função `eigen` do `R`, a Decomposição Espectral das matrizes $\mathbf{A}^t\mathbf{A}$ e $\mathbf{A}\mathbf{A}^t$. Compare com os resultados da alínea anterior e comente.

   (c) Reconstrua, com o auxílio do `R` as três matrizes 4×3 dadas por $\mathbf{A}_i = \delta_i \, \mathbf{u}_i \, \mathbf{v}_i^t$, onde δ_i indica o i-ésimo maior valor singular de $\mathbf{A}$ e $\mathbf{u}_i$ e $\mathbf{v}_i$ são os correspondentes vectores singulares, respectivamente esquerdo e direito.

      i. Compare a matriz $\mathbf{A}_1$ com a matriz $\mathbf{A}$ e comente.

      ii. Compare as matrizes $\mathbf{A}_1 + \mathbf{A}_2$ e $\mathbf{A}_1 + \mathbf{A}_2 + \mathbf{A}_3$ com a matriz $\mathbf{A}$ e comente.

      iii. Reconstrua a matriz $\mathbf{A}_1 + \mathbf{A}_2$ com o auxílio do `R`, mas utilizando a versão matricial da Decomposição em Valores Singulares, ou seja, como

      ```
      >  U[,1:2]%*%D[1:2,1:2]%*%t(V[,1:2])
      ```

      sendo $\mathbf{U}$ e $\mathbf{V}$ as matrizes cujas colunas indicam os vectores singulares esquerdos e direitos, respectivamente, e $\mathbf{D}$ a matriz diagonal dos valores singulares de $\mathbf{A}$.

      iv. Calcule a norma matricial usual de $\mathbf{A}_1$ e compare com o maior valor singular de $\mathbf{A}$.

      v. Calcule a distância (induzida pela norma matricial usual) entre as matrizes $\mathbf{A}$ e $\mathbf{A}_1$. Qual a relação desta distância com os valores singulares de $\mathbf{A}$?

      vi. Calcule a norma matricial de $\mathbf{A}_1 + \mathbf{A}_2$ e relacione-a com os valores singulares de $\mathbf{A}$.

Se **A** é uma matriz de dados, com observações de 3 variáveis em 4 individuos, uma Análise em Componentes Principais está associada à DVS da matriz que se obtém *centrando as colunas de* **A**. Construa esta matriz centrada, **C**, através do seguinte comando do `R`:

```
C <- scale(A,scale=F)
```

(d) Compare a DVS da matriz **C** com a Decomposição Espectral da matriz de (co)variâncias de **A**. De seguida, compare a DVS da matriz $\frac{1}{\sqrt{3}}\mathbf{C}$ com a Decomposição Espectral da matriz de (co)variâncias de **A** (**Nota**: $\sqrt{3} = \sqrt{n-1}$).

(e) Compare uma Análise em Componentes Principais da matriz dos dados **A** com a DVS da matriz de dados centrados, **C**. Utilize o comando `prcomp` do `R` para efectuar a ACP.

2. Verifique que qualquer combinação linear de variáveis centradas (de média zero) é ainda uma variável centrada. Deduza daí que as Componentes Principais são variáveis centradas.

3. Foram observadas treze variáveis morfométricas em $n = 63$ lavagantes do lago Opeongo, no Canadá. As variáveis morfométricas observadas são:

carapace_l	comprimento da carapaça	**tail_l**	comprimento da cauda
carapace_w	largura da carapaça	**carapace_d**	profundidade da carapaça
tail_w	largura da cauda	**areola_l**	comprimento da areola
areola_w	largura da areola	**rostrum_l**	comprimento do rosto
rostrum_w	largura do rosto	**postorbital_w**	largura pós-orbital
propodus_l	comprimento da tenaz	**propodus_w**	largura da tenaz
dactyl_l	comprimento dáctil		

(a) Efectue uma Análise em Componentes Principais dos dados. Comente os resultados, e em particular, comente a qualidade da aproximação que se obtém quando se projecta a nuvem de $n = 63$ pontos em $\mathbb{R}^{13}$ sobre as duas principais direcções de variação em $\mathbb{R}^{13}$.

(b) Construa a nuvem de $n = 63$ pontos associada às duas primeiras componentes principais. Comente, tendo em conta que os 42 primeiros indivíduos são machos e os 21 restantes são fêmeas.

(c) Procure possíveis interpretações das duas primeiras CPs, analisando as respectivas correlações com cada uma das $p = 13$ variáveis originais.

(d) Inspeccione a matriz de (co)variâncias das treze variáveis originais. Comente os resultados da alínea anterior à luz desta matriz.

(e) Analise os gráficos dos $n = 63$ pontos definidos pelos vários pares de CPs que se obtêm a partir das 6 primeiras componentes principais. Identifique os individuos que aparecem destacados ao longo das CPs 3, 4, 5 e 6. Comente.

(f) Efectue agora uma ACP sobre os dados normalizados, isto é, uma ACP sobre a matriz de correlações. Comente as principais diferenças e semelhanças entre ambas as análises, e procure explicar esses resultados.

4. Num estudo sobre o cultivo de framboesas em estufa, observam-se 10 variáveis caracterizadoras de propriedades de frutos colhidos. Mais concretamente, recolhem-se framboesas em 14 plantas e determinam-se os valores médios desses 14 grupos de framboesas para as seguintes variáveis: diâmetro (x_1), altura (x_2), peso (x_3), brix (x_4), pH (x_5), uma outra medida de acidez, que será designada apenas por 'acidez' (x_6), açucar (x_7), e três variáveis indicadoras de propriedades cromáticas e designadas por 'L' (x_8), 'a' (x_9) e 'b' (x_{10}). Os valores assim obtidos foram:

Planta	Diâm.	Alt.	Peso	Brix	pH	Acidez	Açucar	L	a	b
1	2.0	2.1	3.71	8.4	2.78	1.39	5.12	23.0	17.0	4.2
2	2.1	2.0	3.79	8.4	2.84	1.49	5.40	26.3	17.8	5.6
3	2.0	1.7	3.65	8.7	2.89	1.51	5.38	21.3	13.1	4.2
4	2.0	1.8	3.83	8.6	2.91	1.44	5.23	22.5	13.9	4.3
5	1.8	1.8	3.95	8.0	2.84	1.62	3.44	26.2	15.9	6.1
6	2.0	1.9	4.18	8.2	3.00	1.74	3.42	26.4	16.9	6.6
7	2.1	2.2	4.37	8.1	3.00	1.68	3.48	29.2	18.1	7.1
8	1.8	1.9	3.97	8.0	2.96	1.57	3.34	26.0	22.5	7.8
9	1.8	1.8	3.43	8.2	2.75	1.46	2.02	27.4	16.4	6.1
10	1.9	1.9	3.78	8.0	2.75	1.54	2.14	29.0	15.0	5.8
11	1.9	1.9	3.42	8.0	2.73	1.26	2.06	27.2	17.1	6.4
12	2.0	1.9	3.60	8.1	2.71	1.18	2.02	27.5	16.6	6.5
13	1.9	1.7	2.87	8.4	2.94	1.32	3.86	29.9	19.1	6.6
14	2.1	1.9	3.74	8.8	3.20	1.46	3.89	26.6	17.1	4.6

   (a) Diga, justificando, se é adequado efectuar uma ACP sobre os dados originais (não normalizados), ou se considera preferível efectuar uma ACP sobre a matriz de correlações.

   (b) Diga, justificando, se uma Análise de Componentes Principais sobre a matriz das correlações permite representar de forma adequada o conjunto dos dados em apenas duas dimensões, sem grande perda de informação.

   (c) Entretanto, é obtida a informação adicional de que as 14 plantas não foram observadas todas nas mesmas datas, tendo os cortes sido efectuados em cinco datas diferentes, correspondendo os três primeiros grupos de quatro observações a três diferentes datas e as últimas duas observações a duas outras datas. Este facto é reflectido no primeiro plano principal resultante da análise anterior? Responda, identificando os pontos do gráfico da nuvem de pontos no primeiro plano principal.

   (d) *Caso a sua resposta á alínea anterior seja afirmativa*, diga, justificando, se é obrigatório que o primeiro plano principal reflicta esse tipo de estrutura dos dados em sub-grupos. *Caso a sua resposta á alínea anterior seja negativa*, diga como se pode explicar que essa estrutura não esteja reflectida no primeiro plano principal, dadas as propriedades optimizadas pelas primeiras componentes principais.

   (e) Admita agora que se procede a uma nova observação dos valores das 10 variáveis nas framboesas de uma nova planta e que se registaram os seguintes valores: 1.9, 2.0, 3.92, 8.1, 2.91, 1.48, 3.78, 27.2, 17.8, 5.8. Se pretendesse representar esta nova observação no primeiro plano principal, quais as coordenadas que deveria associar-lhe? Justifique a sua resposta e represente o ponto no gráfico dado acima. Seria capaz de indicar se a nova observação corresponde a alguma das

cinco datas de corte das anteriores observações? Responda, indicando quaisquer ressalvas que considere necessárias.

(f) Que significado atribui ao facto de haver observações (como as número 8, 13 ou 14, por exemplo) cujas coordenadas nas terceira e seguintes componentes principais são (em módulo) relativamente elevadas? Discuta brevemente as implicações desse facto numa análise dos resultados desta Análise de Componentes Principais.

5. Numa exploração agrícola da Bélgica registaram-se os valores de $p = 5$ variáveis meteorológicas ao longo de $n = 11$ anos agrícolas, na década de 1920. As cinco variáveis são:

$\mathbf{x}_1$	precipitação total em Novembro e Dezembro (mm)
$\mathbf{x}_2$	temperatura média em Julho (^{o}C)
$\mathbf{x}_3$	precipitação total em Julho (mm)
$\mathbf{x}_4$	radiação em Julho (mm de álcool)
$\mathbf{x}_5$	rendimento médio da colheita (quintais/ha)

Os valores observados foram:

Campanha	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
1920-21	87.9	19.6	1.0	1661	28.37
1921-22	89.9	15.2	90.1	968	23.77
1922-23	153.0	19.7	56.6	1353	26.04
1923-24	132.1	17.0	91.0	1293	25.74
1924-25	88.8	18.3	93.7	1153	26.68
1925-26	220.9	17.8	106.9	1286	24.29
1926-27	117.7	17.8	65.5	1104	28.00
1927-28	109.0	18.3	41.8	1574	28.37
1928-29	156.1	17.8	57.4	1222	24.96
1929-30	181.5	16.8	140.6	902	21.66
1930-31	181.4	17.0	74.3	1150	24.37

(a) Efectue uma Análise em Componentes Principais sobre a matriz de correlações destes dados, determinando as cinco Componentes Principais dos dados (justifique a opção por uma ACP sobre a matriz de correlações).

(b) Construa a melhor representação possível, a duas dimensões, da nuvem de $n = 11$ pontos em $\mathbb{R}^5$ (anos agrícolas) que os dados normalizados definem.

(c) Calcule os coeficientes de correlação entre a primeira Componente Principal e as cinco variáveis originais. Interprete os resultados obtidos.

(d) Construa um *biplot* dos dados normalizados e interprete-o. Indique uma medida da qualidade do *biplot*. Confirme as suas conclusões resultantes da leitura do *biplot*, analisando directamente os dados e os resultados duma ACP.

6. Considere o seguinte conjunto de dados, referido por Kendall (*Multivariate Analysis*, Charles Griffin & Co., 1980, pg. 20), e relativo a medições em 20 amostras de terras:

Amostra	Teor em limo (%)	Teor em argila (%)	Matéria orgânica (%)	Acidez (pH)
1	13.0	9.7	1.5	6.4
2	10.0	7.5	1.5	6.5
3	20.6	12.5	2.3	7.0
4	33.8	19.0	2.8	5.8
5	20.5	14.2	1.9	6.9
6	10.0	6.7	2.2	7.0
7	12.7	5.7	2.9	6.7
8	36.5	15.7	2.3	7.2
9	37.1	14.3	2.1	7.2
10	25.5	12.9	1.9	7.3
11	26.5	14.9	2.4	6.7
12	22.3	8.4	4.0	7.0
13	30.8	7.4	2.7	6.4
14	25.3	7.0	4.8	7.3
15	31.2	11.6	2.4	6.5
16	22.7	10.1	3.3	6.2
17	31.2	9.6	2.4	6.0
18	13.2	6.6	2.0	5.8
19	11.1	6.7	2.2	7.2
20	20.7	9.6	3.1	5.9

(a) Efectue uma Análise de Componentes Principais sobre a matriz de Covariâncias destes dados.

(b) Calcule o coeficiente de correlação entre cada Componente Principal e cada variável. Compare os valores obtidos com os coeficientes das variáveis nas combinações lineares que definem as CPs e veja como a tentativa de interpretar Componentes Principais apenas em função dos coeficientes (*loadings*) pode induzir em erro.

7. Num estudo dos afídios *Alate adelges* efectuaram-se medições de 19 variáveis sobre 40 indivíduos. As variáveis observadas, bem como as médias e variâncias dos valores observados, foram as seguintes:

Nome	Descrição	$\overline{x}$	s^2
COM	comprimento total do organismo	15.05	14.58
LAR	largura do corpo	7.14	4.05
CAA	comprimento da asa anterior	5.68	1.68
CAP	comprimento da asa posterior	3.45	0.83
E	número de espiráculos	4.88	0.11
AS1	comprimento do segmento de antena I	1.86	0.11
AS2	comprimento do segmento de antena II	1.69	0.11
AS3	comprimento do segmento de antena III	2.25	0.22
AS4	comprimento do segmento de antena IV	2.33	0.15
AS5	comprimento do segmento de antena V	2.73	0.15
S	número de sedas antenais	4.28	1.33
TAR	comprimento do tarso III	3.31	0.41
TIB	comprimento da tíbia III	3.38	0.58
FEM	comprimento do fémur III	2.57	0.34
ROS	rostrum	5.58	0.79
OVI	oviescapto	3.72	0.35
N	número de sedas do oviescapto	7.80	3.81
P	prega anal (var. qualitativa 0/1)	0.73	0.20
GAP	número de ganchos da asa posterior	2.38	0.25

As observações obtidas foram as seguintes:

COM	LAR	CAA	CAP	E	AS1	AS2	AS3	AS4	AS5	S	TAR	TIB	FEM	ROS	OVI	N	P	GAP
21.2	11.0	7.5	4.8	5	2.0	2.0	2.8	2.8	3.3	3	4.4	4.5	3.6	7.0	4.0	8	0	3
20.2	10.0	7.5	5.0	5	2.3	2.1	3.0	3.0	3.2	5	4.2	4.5	3.5	7.6	4.2	8	0	3
20.2	10.0	7.0	4.6	5	1.9	2.1	3.0	2.5	3.3	1	4.2	4.4	3.3	7.0	4.0	6	0	3
22.5	8.8	7.4	4.7	5	2.4	2.1	3.0	2.7	3.5	5	4.2	4.4	3.6	6.8	4.1	6	0	3
20.6	11.0	8.0	4.8	5	2.4	2.0	2.9	2.7	3.0	4	4.2	4.7	3.5	6.7	4.0	6	0	3
19.1	9.2	7.0	4.5	5	1.8	1.9	2.8	3.0	3.2	5	4.1	4.3	3.3	5.7	3.8	8	0	3.5
20.8	11.4	7.7	4.9	5	2.5	2.1	3.1	3.1	3.2	4	4.2	4.7	3.6	6.6	4.0	8	0	3
15.5	8.2	6.3	4.9	5	2.0	2.0	2.9	2.4	3.0	3	3.7	3.8	2.9	6.7	3.5	6	0	3.5
16.7	8.8	6.4	4.5	5	2.1	1.9	2.8	2.7	3.1	3	3.7	3.8	2.8	6.1	3.7	8	0	3
19.7	9.9	8.2	4.7	5	2.2	2.0	3.0	3.0	3.1	0	4.1	4.3	3.3	6.0	3.8	8	0	3
10.6	5.2	3.9	2.3	4	1.2	1.0	2.0	2.0	2.2	6	2.5	2.5	2.0	4.5	2.7	4	1	2
9.2	4.5	3.7	2.2	4	1.3	1.2	2.0	1.6	2.1	5	2.4	2.3	1.8	4.1	2.4	4	1	2
9.6	4.5	3.6	2.3	4	1.3	1.0	1.9	1.7	2.2	4	2.4	2.3	1.7	4.0	2.3	4	1	2
8.5	4.0	3.8	2.2	4	1.3	1.1	1.9	2.0	2.1	5	2.4	2.4	1.9	4.4	2.3	4	1	2
11.0	4.7	4.2	2.3	4	1.2	1.0	1.9	2.0	2.2	4	2.5	2.5	2.0	4.5	2.6	4	1	2
18.1	8.2	5.9	3.5	5	1.9	1.9	1.9	2.7	2.8	4	3.5	3.8	2.9	6.0	4.5	9	1	2
17.6	8.3	6.0	3.8	5	2.0	1.9	2.0	2.2	2.9	3	3.5	3.6	2.8	5.7	4.3	10	1	2
19.2	6.6	6.2	3.4	5	2.0	1.8	2.2	2.3	2.8	4	3.5	3.4	2.5	5.3	3.8	10	1	2
15.4	7.6	7.1	3.4	5	2.0	1.9	2.5	2.5	2.9	4	3.3	3.6	2.7	6.0	4.2	8	1	3
15.1	7.3	6.2	3.8	5	2.0	1.8	2.1	2.4	2.5	4	3.7	3.7	2.8	6.4	4.3	10	1	2.5
16.1	7.9	5.8	3.7	5	2.1	1.9	2.3	2.6	2.9	5	3.6	3.6	2.7	6.0	4.5	10	1	2
19.1	8.8	6.4	3.9	5	2.2	2.0	2.3	2.4	2.9	4	3.8	4.0	3.0	6.5	4.5	10	1	2.5
15.3	6.4	5.3	3.3	5	1.7	1.6	2.0	2.2	2.5	5	3.4	3.4	2.6	5.4	4.0	10	1	2
14.8	8.1	6.2	3.7	5	2.2	2.0	2.2	2.4	3.2	5	3.5	3.7	2.7	6.0	4.1	10	1	2
16.2	7.7	6.9	3.7	5	2.0	1.8	2.3	2.4	2.8	4	3.8	3.7	2.7	5.7	4.2	10	1	2.5
13.4	6.9	5.7	3.4	5	2.0	1.8	2.8	2.0	2.6	4	3.6	3.6	2.6	5.5	3.9	10	1	2
12.9	5.8	4.8	2.6	5	1.6	1.5	1.9	2.1	2.6	5	2.8	3.0	2.2	5.1	3.6	9	1	3
12.0	6.5	5.3	3.2	5	1.9	1.9	2.3	2.5	3.0	5	3.3	3.5	2.6	5.4	4.3	8	1	2
14.1	7.0	5.5	3.6	5	2.2	2.0	2.3	2.5	3.1	5	3.6	3.7	2.8	5.8	4.1	10	1	2
16.7	7.2	5.7	3.5	5	1.9	1.9	2.5	2.3	2.8	5	3.4	3.6	2.7	6.0	4.0	10	1	2.5
14.1	5.4	5.0	3.0	5	1.7	1.6	1.8	2.5	2.4	5	2.7	2.9	2.2	5.3	3.6	8	1	2
10.0	6.0	4.2	2.5	5	1.6	1.4	1.4	2.0	2.7	6	2.8	2.5	1.8	4.8	3.4	8	1	2
11.4	4.5	4.4	2.7	5	1.8	1.5	1.9	1.7	2.5	5	2.7	2.5	1.9	4.7	3.7	8	1	2
12.5	5.5	4.7	2.3	5	1.8	1.4	1.8	2.2	2.4	4	2.8	2.6	2.0	5.1	3.7	8	0	2
13.0	5.3	4.7	2.3	5	1.6	1.4	1.8	1.8	2.5	4	2.7	2.7	2.1	5.0	3.6	8	1	2
12.4	5.2	4.4	2.6	5	1.6	1.4	1.8	2.2	2.2	5	2.7	2.5	2.0	5.0	3.2	6	1	2
12.0	5.4	4.9	3.0	5	1.7	1.5	1.7	1.9	2.4	5	2.7	2.7	2.0	4.2	3.7	6	1	2
10.7	5.6	4.5	2.8	5	1.8	1.4	1.8	2.2	2.4	4	2.7	2.6	2.0	5.0	3.5	8	1	2
11.7	5.5	4.3	2.6	5	1.7	1.5	1.8	1.9	2.4	5	2.6	2.5	1.9	4.6	3.4	8	1	2
12.8	5.7	4.8	2.8	5	1.6	1.4	1.7	1.9	2.3	5	2.3	2.5	1.9	5.0	3.1	8	1	2

(a) Descreva sucintamente as principais características do feixe de vectores que representa as 19 variáveis (centradas,*mas não normalizadas*) no espaço $\mathbb{R}^{40}$.

(b) Efectue uma Análise de Componentes Principais sobre a matriz de correlações dos dados.

(c) Procure interpretar o significado das três primeiras componentes principais, à luz da informação disponível. Justifique.

(d) Construa a melhor representação bidimensional dos dados. Diga se a considera adequada. Identifique dois indivíduos cuja representação no primeiro plano principal seja menos fidedigna. Justifique.

(e) Na projecção da nuvem de pontos no plano definido pelos dois primeiros eixos principais aparece com alguma nitidez uma arrumação dos 40 indivíduos em grupos. Relacione essa arrumação com as variáveis originais e comente. (Caso faça sentido, sugira algum possível significado biológico para este facto).

(f) Não é muito frequente encontrar conjuntos de dados com 19 variáveis para os quais uma ACP sobre a matriz de correlações produza valores tão elevados dos primeiros 2 ou 3 valores próprios. O que pensa que pode justificar este facto, no nosso conjunto de dados?

(g) Avalie criticamente a utilização duma ACP neste conjunto de 19 variáveis, tendo em atenção a natureza de (algumas) delas. Sugira alternativas, no caso de considerar haver algum aspecto indesejável.

8. Num estudo sobre borboletas, foram observadas seis espécies em sete diferentes localidades das ilhas britânicas. Em cada localidade registou-se o número de observações de indivíduos da referida espécie por ano, ao longo de seis anos. Seguidamente, calculou-se para cada espécie e localidade a média das seis observações anuais, e efectuou-se uma transformação logarítmica dessas médias. As log-frequências médias por ano obtidas foram as seguintes:

Local	Espécie					
	1	2	3	4	5	6
A	4.4	3.9	4.3	1.9	7.0	0.1
C	3.7	3.8	3.7	3.7	3.7	0.8
E	5.1	3.5	4.5	6.0	2.7	2.3
G	2.9	3.6	3.8	1.5	0.6	3.4
I	2.1	3.5	2.3	3.4	1.0	0.2
K	5.4	0.0	0.0	0.0	2.6	0.0
M	4.9	2.4	0.0	0.3	1.2	0.2
Médias	4.071	2.957	2.657	2.400	2.686	1.000
Variâncias	1.502	1.943	3.790	4.480	???	1.763

A matriz de correlações das seis espécies, com base nestas observações é:

```
      1      2      3       4       5      6
1  1.000 -0.552 -0.302 -0.1792  0.3385 -0.216
2 -0.552  1.000  0.817  0.6107  0.2041  0.378
3 -0.302  0.817  1.000  0.7176  0.4111  0.549
4 -0.179  0.611  0.718  1.0000  0.0977  0.355
5  0.338  0.204  0.411  0.0977  1.0000 -0.368
6 -0.216  0.378  0.549  0.3552 -0.3679  1.000
```

A matriz das distâncias Euclideanas entre as sete localidades (com base nas observações acima referidas) é:

```
      A    C    E    G    I    K    M
A  0.00 3.93 6.39 7.39 6.91 7.59 7.56
C  3.93 0.00 3.35 4.68 3.51 6.82 5.94
E  6.39 3.35 0.00 5.59 5.28 8.60 7.79
G  7.39 4.68 5.59 0.00 4.11 7.17 5.65
I  6.91 3.51 5.28 4.11 0.00 6.53 4.90
K  7.59 6.82 8.60 7.17 6.53 0.00 2.85
M  7.56 5.94 7.79 5.65 4.90 2.85 0.00
```

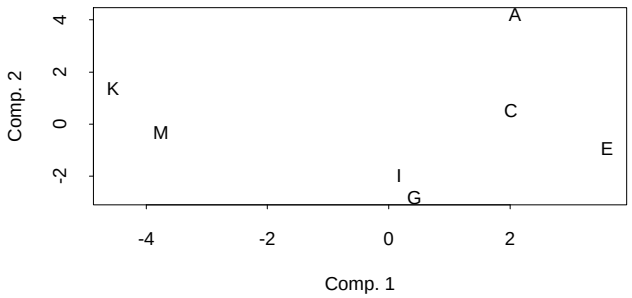
Uma Análise em Componentes Principais com base na matriz de covariâncias (onde se entendem localidades como indivíduos e espécies como variáveis) produziu os seguintes resultados, dos quais foram omitidos alguns valores:

```
Importance of components:
                       Comp. 1 Comp. 2 Comp. 3 Comp. 4 Comp. 5 Comp. 6
    Standard deviation   2.835   2.156   1.253   1.088   0.511   0.022
Proportion of Variance   0.512   0.296   0.100   0.075   0.017   0.000
 Cumulative Proportion   0.512   0.808   0.908   0.983   1.000   1.000

  Comp. 1 Comp. 2 Comp. 3 Comp. 4 Comp. 5 Comp. 6       Vectores de loadings
1  -0.113   0.276   0.529   0.520  -0.522  -0.297
2   0.382  -0.087  -0.372  -0.244  -0.800   0.089
3   0.615   0.015  -0.253   0.270   0.255  -0.647
4   0.588  -0.208     ???  -0.274   0.085   0.200
5   0.282   0.867  -0.092   0.030   0.117   0.382
6   0.192  -0.349  -0.143   0.721   0.038   0.547

  Comp. 1 Comp. 2 Comp. 3 Comp. 4 Comp. 5  Comp. 6      Correlações
1  -0.281  0.5244  0.5844  0.4989 -0.2352 -0.00563
2   0.840 -0.1458 -0.3615 -0.2056 -0.3171  0.00149
3   0.968  0.0176 -0.1762  0.1630  0.0725 -0.00773
4   0.851 -0.2287     ??? -0.1524  0.0221  0.00219
5   0.393  0.9172 -0.0564  0.0158  0.0293  0.00404
6   0.444 -0.6117 -0.1459  0.6384  0.0157  0.00957
```

O gráfico das sete localidades no primeiro plano principal é o seguinte:



(a) Comente a natureza das duas primeiras Componentes Principais e a sua importância relativa, com base na informação disponível.

(b) Comente se considera adequada a opção por uma Análise em Componentes Principais baseada na matriz de variâncias-covariâncias, ou se seria preferível ter feito uma ACP sobre a matriz de correlações.

(c) Como explica que a distância entre as localidades A e C, na matriz de distâncias Euclideanas, seja inferior à distância entre as localidades G e I, quando a localização desses pontos no primeiro plano principal nos dá a informação contrária? Comente.

(d) Preencha os valores omissos nas tabelas acima indicadas (a variância de variável 5; o coeficiente associado à variável 4, na componente 3; e a correlação entre a variável 4 e a componente 3).

9. Nas Estatísticas Agrícolas do INE (1973) indicam-se as seguintes produtividades (em t/ha) de 9 produções agrícolas, nos 20 concelhos do distrito de Santarém:

Concelho	Trigo	Milho	Centeio	Aveia	Cevada	Fava	Feijão	Grão	Batata
Abrantes	1.041	0.541	0.515	0.595	0.402	0.672	0.327	0.423	7.437
Alcanena	0.887	1.697	0.700	1.051	0.630	0.631	0.517	0.618	10.317
Almeirim	1.013	0.431	0.545	0.511	0.374	0.696	0.376	0.495	7.389
Alpiarça	1.293	1.803	0.891	0.413	1.094	0.591	0.518	0.500	17.678
Benavente	1.559	1.949	0.669	1.053	1.029	0.628	0.346	0.614	8.290
Cartaxo	0.925	1.600	0.544	0.696	0.460	0.657	0.352	0.469	9.071
Chamusca	1.103	3.144	0.379	0.321	0.423	0.542	0.543	0.442	17.199
Constância	1.516	0.524	0.321	0.562	0.571	0.474	0.381	0.485	11.271
Coruche	1.443	0.483	0.605	0.698	1.250	0.742	0.229	0.371	19.160
Entroncamento	1.023	4.120	0.716	0.621	0.707	1.057	0.533	0.700	20.600
F.do Zêzere	0.981	2.413	0.305	0.773	1.048	0.696	0.524	0.602	9.889
Golegã	1.223	3.777	0.646	0.330	0.763	0.763	0.672	0.311	8.113
Mação	0.839	0.772	0.306	0.362	0.260	0.600	0.293	0.420	8.468
Rio Maior	0.809	1.153	0.927	0.694	0.707	1.777	0.417	0.433	7.060
Salvaterra	1.509	1.100	1.034	0.697	1.582	1.138	0.636	0.516	10.791
Santarém	0.712	1.342	1.145	0.457	0.686	0.982	0.616	0.426	14.135
Sardoal	0.780	0.463	0.326	0.414	0.435	0.822	0.383	0.396	10.078
Tomar	1.000	1.928	0.430	0.863	1.080	0.913	0.404	0.687	9.320
Torres Novas	1.262	2.453	0.716	0.971	0.885	0.928	0.512	0.664	21.100
V.N.Barquinha	0.917	1.081	0.811	1.000	0.909	0.967	0.620	0.667	18.347

A matriz de variâncias-covariâncias resultante destes dados é a seguinte:

	Trigo	Milho	Centeio	Aveia	Cevada	Fava	Feijão	Grão	Batata
Trigo	1.302								
Milho	0.306	22.770							
Centeio	0.040	0.318	1.169						
Aveia	0.188	-0.112	0.203	1.091					
Cevada	0.958	0.768	0.734	0.655	2.222				
Fava	-0.400	0.175	0.783	0.228	0.492	1.600			
Feijão	-0.041	1.445	0.305	-0.013	0.239	0.166	0.299		
Grão	0.024	0.716	0.042	0.388	0.232	0.057	0.055	0.255	
Batata	4.493	32.965	5.853	2.006	8.927	-0.066	3.180	3.488	447.087

Uma Análise de Componentes Principais dos dados referidos acima (com base na matriz de (co)variâncias e utilzando o programa *Genstat*) produziu os seguintes resultados, para as 5 primeiras componentes:

***** Principal components analysis *****

*** Latent Roots ***

1	2	3	4	5
450.00	20.31	3.25	2.15	1.03

*** Percentage variation ***

1	2	3	4	5
94.18	4.25	0.68	0.45	0.22

*** Trace ***

477.8

*** Latent Vectors (Loadings) ***

	1	2	3	4	5
trigo	0.01008	0.00250	-0.25566	-0.60455	-0.27226
milho	0.07699	-0.99480	0.00908	-0.01042	0.00847
centeio	0.01309	0.00573	-0.38322	0.33513	-0.28307
aveia	0.00448	0.01329	-0.31187	-0.07449	0.82870
cevada	0.02006	-0.00454	-0.72296	-0.29216	-0.11798
fava	-0.00008	-0.00994	-0.39100	0.64940	-0.03922
feijao	0.00732	-0.05966	-0.08643	0.09531	-0.11025
grao	0.00787	-0.02201	-0.08260	-0.01772	0.36238
batata	0.99663	0.07739	0.02413	0.00822	0.00241

*** Principal Component Scores ***

	1	2	3	4	5
1	-4.928	0.728	0.284	-0.028	0.006
2	-1.958	-0.208	0.009	0.026	0.415
3	-4.984	0.828	0.304	0.033	-0.042
4	5.398	0.250	-0.101	-0.208	-0.356
5	-3.946	-0.606	-0.471	-0.545	0.210
6	-3.217	-0.200	0.278	0.022	0.134
7	5.001	-1.122	0.680	-0.108	-0.168
8	-1.102	1.038	0.285	-0.520	-0.082
9	6.773	1.698	-0.231	-0.362	-0.126
10	8.478	-1.836	0.142	0.295	0.064
11	-2.327	-0.962	-0.108	-0.232	0.217
12	-3.994	-2.462	-0.002	-0.114	-0.396
13	-3.892	0.577	0.649	0.039	-0.046
14	-5.246	0.075	-0.510	0.870	-0.047
15	-1.504	0.407	-1.050	-0.137	-0.329
16	1.821	0.429	-0.013	0.585	-0.250
17	-2.307	1.002	0.457	0.196	-0.039
18	-2.929	-0.519	-0.312	-0.089	0.278
19	8.856	-0.131	-0.105	0.009	0.250
20	6.005	1.014	-0.184	0.267	0.307

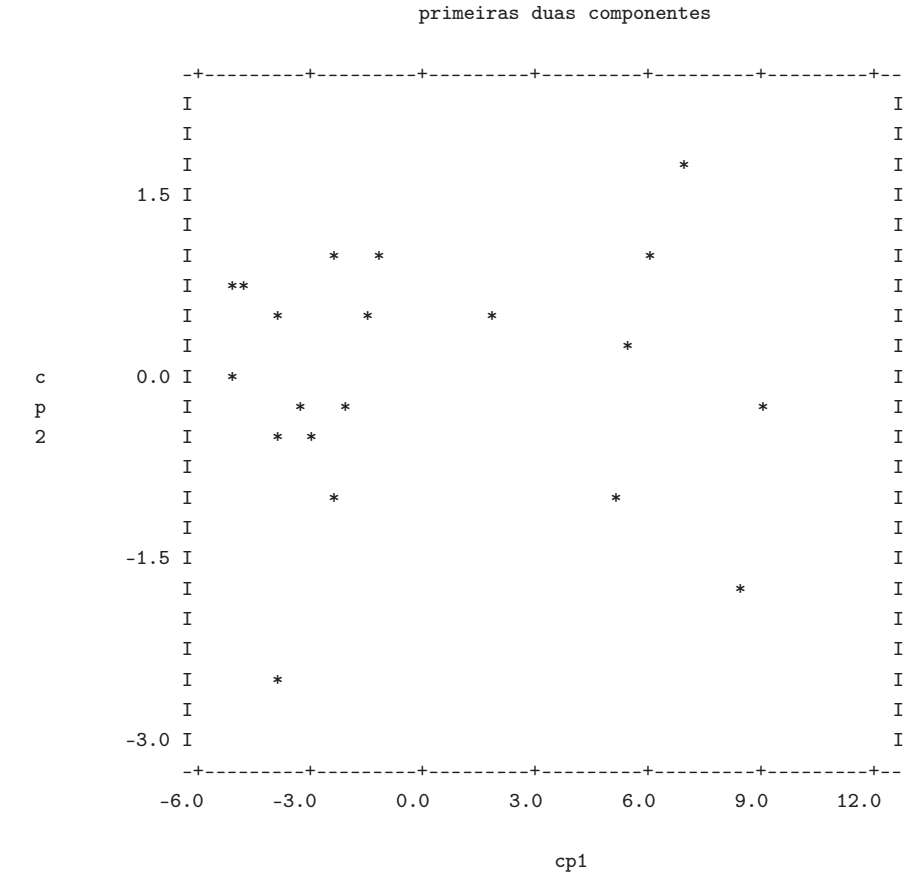
*** Sums of squares and products *** (Matriz de variancias-covariancias)

```
trigo     1.3023
milho     0.3061   22.7699
centeio   0.0398    0.3175    1.1687
aveia     0.1875   -0.1122    0.2035    1.0911
cevada    0.9580    0.7681    0.7341    0.6550    2.2225
fava     -0.4003    0.1748    0.7832    0.2284    0.4924    1.5995
feijao   -0.0413    1.4450    0.3054   -0.0130    0.2391    0.1657    0.2993
grao      0.0243    0.7156    0.0418    0.3878    0.2316    0.0573    0.0553    0.2545
batata    4.4929   32.9645    5.8525    2.0065    8.9272   -0.0664    3.1803    3.4878   447.0869
               1         2         3         4         5         6         7         8         9

*** Means ***
trigo     1.092       aveia     0.6541       feijao    0.4600
milho     1.639       cevada    0.7648       grao      0.5120
centeio   0.6266      fava      0.8138       batata   12.29
```

(a) Diga, justificando, se uma Análise de Componentes Principais sobre a matriz das covariâncias permite reduzir a dimensionalidade do conjunto dos dados sem grande perda de informação.

(b) A projecção da nuvem de 20 pontos (concelhos) sobre o plano definido pelos dois primeiros eixos principais é dada no gráfico abaixo. Identifique os 7 concelhos correspondentes aos pontos na metade direita do gráfico. Identifique ainda o ponto isolado no canto inferior esquerdo.



(c) Calcule os três coeficientes de correlação entre a primeira componente principal e as variáveis "batata","milho" e "feijão". Repita para a segunda componente principal.

(d) Procure interpretar a natureza das duas primeiras componentes principais, justificando as suas conclusões.

(e) Avalie criticamente a Análise de Componentes Principais (ACP) efectuada, indicando se se trata duma aplicação adequada desse método. Caso a sua resposta seja negativa, sugira aplicações alternativas do método (ACP) ao estudo das produtividades agrícolas no distrito de Santarém.

(f) Efectue agora uma Análise de Componentes Principais dos dados, mas baseada na matriz de correlações. Compare os resultados das duas ACPs e comente.

Capítulo 3

Análise Discriminante Linear

A expressão Análise Discriminante tem sido utilizada para identificar diversas técnicas multivariadas que, no entanto, têm um objectivo comum. Parte-se do conhecimento de que os n indivíduos observados pertencem a diversos subgrupos e procura-se determinar funções das p variáveis observadas que melhor permitam distinguir ou **discriminar** entre esses subgrupos ou classes.

3.1 Introdução

Como foi visto no Capítulo 2, Componentes Principais não são necessariamente boas soluções para efeitos de discriminação, pois as direcções de variabilidade principal não têm que coincidir com as direcções de melhor discriminação. Em Análise Discriminante coloca-se explicitamente o objectivo de separar subgrupos de individuos, subgrupos esses que são previamente conhecidos nos dados observados.

Neste Capítulo será abordada uma técnica discriminante, válida no contexto descritivo onde nos situamos, conhecida por Análise Discriminante Linear, ou de Fisher. Existem outras técnicas discriminantes, nomeadamente técnicas que se baseiam em modelos probabilisticos, que não serão abordadas aqui. A discriminação de Fisher tem a virtude de ser facilmente visualizável em termos geométricos. Além disso, não exige hipóteses adicionais (ao contrário das técnicas baseadas em modelos probabilisticos). Tem também a vantagem de permitir discriminar mais que dois diferentes sub-grupos (classes) sem grande complexidade, facto que nem sempre se verifica nos métodos baseados em considerações inferenciais.

Na Análise Discriminante de Fisher procuram-se as *combinações lineares* $\mathbf{Xa}$ das p variáveis observadas que melhor separem os subgrupos de individuos indicados, segundo um critério de separabilidade que adiante se discute em mais pormenor.

As soluções $\mathbf{Xa}$ obtidas designam-se **eixos discriminantes** ou também **variáveis canónicas**[1]. Podem

[1]Embora tal designação apareça também associada a um conceito completamente diferente, no âmbito duma técnica designada Análise das Correlações Canónicas.

ser utilizados para obter uma representação gráfica que saliente a distinção entre as classes. E podem também ser de utilidade para classificar futuros indivíduos (observados nas mesmas variáveis), do qual seja desconhecido à partida o subgrupo a que pertence.

Na Secção 3.2 descreve-se de forma mais pormenorizada o método.

3.2 O método em mais pormenor

O ponto de partida para uma Análise Discriminante é uma matriz $\mathbf{X}$ de dados observados, mas desta vez acompanhada pelo conhecimento de que os n indivíduos observados se distribuem por k classes (gerando uma **partição**, *i.e.*, **cada indivíduo pertence a uma e uma só classe**). Neste contexto (e ao contrário da notação usada no Capítulo 2 sobre Análise em Componentes Principais) **designamos por X a matriz de dados *sem* centragem prévia das colunas**.

O critério que preside à determinação de soluções na Análise Discriminante de Fisher baseia-se na seguinte ideia: **de entre as possíveis combinações lineares Xa das variáveis observadas, pretende-se escolher aquela em que os indivíduos de cada classe se tornam mais homogéneos, e as diversas classes se tornam mais heterogéneas entre si**; por outras palavras, pretendemos que os valores dos n_i indivíduos da i-ésima classe na variável $\mathbf{y} = \mathbf{Xa}$ sejam parecidos, e claramente distintos dos valores que os restantes indivíduos (não pertencentes à classe i) assumem, nessa combinação linear.

Ver-se-á em seguida que **a solução envolve uma projecção ortogonal da matriz dos dados centrados sobre o subespaço gerado pelas colunas indicatrizes da constituição de cada classe**. De facto, considere-se a matriz $\mathbf{C}$, cuja i-ésima coluna é uma coluna indicatriz de pertença ao i-ésimo subgrupo de indivíduos[2]. Admitindo (sem perda de generalidade) que os indivíduos duma mesma classe estão arrumados sequencialmente, a matriz $\mathbf{C}$ terá o aspecto indicado na equação (3.1).

[2]Esta matriz desempenha o papel que, no contexto da Análise de Variância é desempenhado pela *matriz do delineamento*. Nesse contexto, opta-se por construir uma matriz do delineamento com uma coluna de uns, e, para evitar os problemas de multicolinearidade, as restantes colunas eram dadas pelas variáveis indicatrizes de todos os níveis do Factor *menos um*. Esta opção é justificada, na disciplina de Modelação Estatística I, pelo facto de ser a que melhor se generaliza para ANOVAs com mais do que um Factor, e melhor permitir a integração da ANOVA no âmbito geral do Modelo Linear. Mas neste contexto, em que apenas existe um único Factor (os subgrupos) não existe a necessidade de assegurar uma solução que se possa generalizar para outras situações. Assim, é mais fácil expôr as ideas admitindo que a matriz do delineamento/classificação é constituída pelas k variáveis indicatrizes dos k subgrupos.

$$\mathbf{C} = \left[\begin{array}{ccccc}
1 & 0 & 0 & \cdots & 0 \\
1 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 0 & \cdots & 0 \\
-- & -- & -- & -- & -- \\
0 & 1 & 0 & \cdots & 0 \\
0 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 1 & 0 & \cdots & 0 \\
-- & -- & -- & -- & -- \\
0 & 0 & 1 & \cdots & 0 \\
0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 1 & \cdots & 0 \\
-- & -- & -- & -- & -- \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
-- & -- & -- & -- & -- \\
0 & 0 & 0 & \cdots & 1 \\
0 & 0 & 0 & \cdots & 1 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 1
\end{array}\right] \tag{3.1}$$

Observação: A matriz $\mathbf{C}$ designa-se a **matriz da classificação** uma vez que as suas colunas definem a classe a que cada indivíduo pertence. Note-se que *as colunas de* $\mathbf{C}$ *são sempre ortogonais entre si*, uma vez que nenhum indivíduo pode pertencer a mais do que uma classe. Note-se ainda que *a soma das* k *colunas da matriz* $\mathbf{C}$ *é o vector dos uns,* $\mathbf{1}_n$, uma vez que cada indivíduo pertence a uma (e uma só) classe. *O quadrado da norma da* j*-ésima coluna de* $\mathbf{C}$ *é* n_j*, o número de indivíduos que pertencem à* j*-ésima classe*.

Assinale-se que qualquer vector pertencente ao subespaço de $\mathbb{R}^n$ gerado pelas colunas da matriz $\mathbf{C}$ caracteriza-se por ter valor igual nos elementos associado às observações de cada subgrupo. Ou seja, os elementos $\mathbf{z} \in \mathcal{C}(\mathbf{C})$ são da forma:

$$\mathbf{z}^t = [\underbrace{z_1 \;\; z_1 \;\; \ldots \;\; z_1}_{n_1 \text{ vezes}} \mid \underbrace{z_2 \;\; z_2 \;\; \ldots \;\; z_2}_{n_2 \text{ vezes}} \mid \cdots \mid \underbrace{z_k \;\; z_k \;\; \ldots \;\; z_k}_{n_k \text{ vezes}}] \tag{3.2}$$

onde n_i $(i = 1 : k)$ indica o número de individuos associados à i-ésima classe.

A *matriz de projecções ortogonais sobre o subespaço (de* $\mathbb{R}^n$*) gerado pelas colunas de* $\mathbf{C}$ é:

$$\mathbf{P}_{\mathbf{C}} = \mathbf{C}(\mathbf{C}^t\mathbf{C})^{-1}\mathbf{C}^t \tag{3.3}$$

Regressemos ao problema de determinar uma "boa" combinação linear das colunas da matriz de dados, $\mathbf{Xa}$, para efeitos de separação de subgrupos.

Pelo que ficou dito, uma combinação linear $\mathbf{Xa}$ próxima do subespaço $\mathcal{C}(\mathbf{C})$ gerado pelas colunas da matriz $\mathbf{C}$ será uma nova variável na qual os valores de individuos associados a uma mesma classe serão semelhantes entre si. Mas existe ainda o outro aspecto do problema a considerar: desejamos que os valores associados a individuos de classes diferentes sejam tanto quanto possível diferentes. E a proximidade de $\mathbf{Xa}$ a $\mathcal{C}(\mathbf{C})$ apenas não garante essa condição. De facto, o subespaço $\mathcal{C}(\mathbf{C})$ também inclui os múltiplos escalares do vector $\mathbf{1}_n$ (confirme!), pelo que inclui vectores que em nada distinguem os individuos de classes diferentes. Assim, desejamos uma combinação linear $\mathbf{Xa}$ próxima de $\mathcal{C}(\mathbf{C})$, mas ao mesmo tempo o mais diferente possível dos vectores em $\mathcal{C}(\mathbf{1}_n)$, ou seja, desejamos uma combinação linear o mais próxima possível do subespaço $\mathcal{C}(\mathbf{C}) \cap \mathcal{C}(\mathbf{1}_n)^\perp$. A forma mais simples de garantir essa condição será **proceder à *centragem* de qualquer combinação linear $\mathbf{Xa}$**, uma vez que esses vectores, dados por $(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{Xa}$, pertencem necessariamente a $\mathcal{C}(\mathbf{1}_n)^\perp$. Assim, **o nosso objectivo será determinar a combinação linear centrada $(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{Xa}$ mais próxima possível do subespaço $\mathcal{C}(\mathbf{C})$**, o que sabemos resulta da projecção ortogonal dessa combinação linear centrada sobre o referido subespaço[3].

Explicitemos a operação de centragem das colunas de $\mathbf{X}$ (como foi feito na pg. 45). A matriz de dados centrados é $(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{X}$, e uma combinação linear das colunas desta matriz centrada é da forma:

$$\mathbf{z} = (\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{Xa}$$

Vamos agora escrever este vector como *a soma da sua componente no subespaço gerado pelas colunas de* $\mathbf{C}$ *e da sua componente no complemento ortogonal desse subespaço*, isto é, vamos recorrer à decomposição em soma directa

$$\mathbb{R}^n \quad = \quad \mathcal{C}(\mathbf{C}) \quad \oplus \quad \mathcal{C}(\mathbf{C})^\perp \tag{3.4}$$

obtendo-se então a seguinte decomposição do vector $\mathbf{z}$:

$$(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{Xa} \quad = \quad \mathbf{P}_{\mathbf{C}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{Xa} \; + \; (\mathbf{I}_n - \mathbf{P}_{\mathbf{C}})(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{Xa}$$

Repare-se ainda que, pelo Teorema 1.27 (p. 26):

$$\mathbf{P}_{\mathbf{C}}\mathbf{P}_{\mathbf{1}_n} = \mathbf{P}_{\mathbf{1}_n}$$

uma vez que o vector $\mathbf{1}_n$ pertence ao subespaço gerado pelas colunas de $\mathbf{C}$. Daí resulta que a decomposição acima referida se pode ainda escrever como:

$$(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{Xa} = (\mathbf{P}_{\mathbf{C}} - \mathbf{P}_{\mathbf{1}_n})\mathbf{Xa} + (\mathbf{I}_n - \mathbf{P}_{\mathbf{C}})\mathbf{Xa}$$

[3] De forma mais formal, e trabalhando com os conceitos de somas directas de mais do que dois subespaços estudados na disciplina de Modelação Estatística I, podemos dizer que consideramos o espaço $\mathbb{R}^n$ como soma directa de três seus subespaços: $\mathbb{R}^n = \mathcal{C}(\mathbf{1}_n) \oplus \left[\mathcal{C}(\mathbf{C}) \cap \mathcal{C}(\mathbf{1}_n)^\perp\right] \oplus \mathcal{C}(\mathbf{C})^\perp$. Procura-se a combinação linear $\mathbf{Xa}$ que esteja mais próxima do segundo desses subespaços.

Uma vez que se trata duma decomposição associada a uma projeção ortogonal, podemos aplicar o Teorema de Pitágoras (página 24) e concluir que:

$$\|(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{Xa}\|^2 = \|(\mathbf{P}_\mathbf{C} - \mathbf{P}_{\mathbf{1}_n})\mathbf{Xa}\|^2 + \|(\mathbf{I}_n - \mathbf{P}_\mathbf{C})\mathbf{Xa}\|^2 \tag{3.5}$$

A natureza das projecções efectuadas torna cada uma destas normas ao quadrado relevantes para o problema sob estudo.

Já sabemos que o membro esquerdo ($\|(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{Xa}\|^2$) é proporcional à variância dos valores dos indivíduos na variável $\mathbf{y} = \mathbf{Xa}$ (ver página 46), pelo que representa uma medida da *variabilidade total dos valores observados de* $\mathbf{y} = \mathbf{Xa}$. A fim de interpretar a natureza das parcelas do membro direito da igualdade, olhemos para a forma da matriz de projeção $\mathbf{P}_\mathbf{C}$.

Como vimos há pouco, a *matriz de projecções ortogonais sobre o subespaço (de* $\mathbb{R}^n$*) gerado pelas colunas de* $\mathbf{C}$ *é*: $\mathbf{P}_\mathbf{C} = \mathbf{C}(\mathbf{C}^t\mathbf{C})^{-1}\mathbf{C}^t$. Ora, a ortogonalidade das colunas de $\mathbf{C}$ implica que a matriz $\mathbf{C}^t\mathbf{C}$ é uma matriz diagonal, e que os seus k elementos diagonais são as dimensões de cada classe, $\{n_j\}_{j=1}^k$. Logo, a matriz inversa $(\mathbf{C}^t\mathbf{C})^{-1}$ é também uma matriz diagonal, cujos elementos diagonais são os recíprocos das dimensões das classes, $1/n_j$.

Do que acaba de ser dito resulta que a matriz de projeções ortogonais $\mathbf{P}_\mathbf{C}$ tem a forma:

$$\mathbf{P}_\mathbf{C} = \left[\begin{array}{cccc|c|c} \frac{1}{n_1} & \frac{1}{n_1} & \cdots & \frac{1}{n_1} & & \\ \vdots & \vdots & \ddots & \vdots & \mathbf{0}_{n_1 \mathrm{x} n_2} & \cdots \quad \mathbf{0}_{n_1 \mathrm{x} n_k} \\ \frac{1}{n_1} & \frac{1}{n_1} & \cdots & \frac{1}{n_1} & & \\ \hline & & & & \frac{1}{n_2} \; \frac{1}{n_2} \; \cdots \; \frac{1}{n_2} & \\ \mathbf{0}_{n_2 \mathrm{x} n_1} & & & & \vdots \quad \vdots \quad \ddots \quad \vdots & \cdots \quad \mathbf{0}_{n_2 \mathrm{x} n_k} \\ & & & & \frac{1}{n_2} \; \frac{1}{n_2} \; \cdots \; \frac{1}{n_2} & \\ \hline \multicolumn{4}{c|}{\vdots} & \multicolumn{2}{c}{\vdots \qquad\qquad \ddots \qquad\qquad \vdots} \\ \hline & & & & & \frac{1}{n_k} \; \frac{1}{n_k} \; \cdots \; \frac{1}{n_k} \\ \mathbf{0}_{n_k \mathrm{x} n_1} & & & & \mathbf{0}_{n_k \mathrm{x} n_2} & \cdots \quad \vdots \quad \vdots \quad \ddots \quad \vdots \\ & & & & & \frac{1}{n_k} \; \frac{1}{n_k} \; \cdots \; \frac{1}{n_k} \end{array}\right]$$

Exercício 3.1 *Confirme esta afirmação sobre a natureza da matriz* $\mathbf{P}_\mathbf{C}$*. Verifique que, se* $k = n$*, tem-se* $\mathbf{P}_\mathbf{C} = \mathbf{I}_n$*. Se* $k = 1$*, tem-se* $\mathbf{P}_\mathbf{C} = \mathbf{P}_{\mathbf{1}_n}$*. Veja as consequências desta forma da matriz* $\mathbf{P}_\mathbf{C}$ *nestes dois casos extremos.*

Assim sendo, o vector $\mathbf{P_C y} = \mathbf{P_C X a}$ será da forma:

$$\mathbf{P_C y} = \begin{bmatrix} \overline{y}^{(1)} \\ \vdots \\ \overline{y}^{(1)} \\ \hline \overline{y}^{(2)} \\ \vdots \\ \overline{y}^{(2)} \\ \hline \vdots \\ \hline \overline{y}^{(k)} \\ \vdots \\ \overline{y}^{(k)} \end{bmatrix}$$

Isto é, o vector $\mathbf{P_C y}$ é o vector n-dimensional cujas n_1 primeiros elementos são todos iguais à média dos valores de $\mathbf{y}$ na classe 1, os n_2 elementos seguintes são todos iguais à média dos valores de $\mathbf{y}$ para os indivíduos da segunda classe, e por aí fora.

Tem-se então:

$$(\mathbf{I}_n - \mathbf{P_C})\mathbf{y} = \begin{bmatrix} y_1^{(1)} - \overline{y}^{(1)} \\ y_2^{(1)} - \overline{y}^{(1)} \\ \vdots \\ y_{n_1}^{(1)} - \overline{y}^{(1)} \\ \hline y_1^{(2)} - \overline{y}^{(2)} \\ y_2^{(2)} - \overline{y}^{(2)} \\ \vdots \\ y_{n_2}^{(2)} - \overline{y}^{(2)} \\ \hline \vdots \\ \hline y_1^{(k)} - \overline{y}^{(k)} \\ y_2^{(k)} - \overline{y}^{(k)} \\ \vdots \\ y_{n_k}^{(k)} - \overline{y}^{(k)} \end{bmatrix}$$

onde $y_i^{(j)}$ é o valor de $\mathbf{y}$ para o i-ésimo elemento da j-ésima classe. Daí resulta que

$$\|(\mathbf{I}_n - \mathbf{P_C})\mathbf{y}\|^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_i^{(j)} - \overline{y}^{(j)})^2$$

ou seja, **$\|(\mathbf{I}_n - \mathbf{P_C})\mathbf{y}\|^2$ é a soma dos numeradores das variâncias de y em cada uma das k classes**. *Uma "boa" variável* $\mathbf{y} = (\mathbf{I} - \mathbf{P_{1_n}})\mathbf{X a}$ *será uma combinação linear para a qual esta parcela é "pequena", uma vez que esse facto reflectirá a existência de classes internamente homogéneas*. Designaremos esta parcela por **variabilidade intra-classes** dos dados.

Por outro lado, temos:

$$(\mathbf{P_C} - \mathbf{P_{1_n}})\mathbf{y} = \begin{bmatrix} \overline{y}^{(1)} - \overline{y} \\ \vdots \\ \overline{y}^{(1)} - \overline{y} \\ \hline \overline{y}^{(2)} - \overline{y} \\ \vdots \\ \overline{y}^{(2)} - \overline{y} \\ \hline \vdots \\ \hline \overline{y}^{(k)} - \overline{y} \\ \vdots \\ \overline{y}^{(k)} - \overline{y} \end{bmatrix}$$

A norma ao quadrado deste vector é, pois:

$$\|(\mathbf{P_C} - \mathbf{P_{1_n}})\mathbf{y}\|^2 = \sum_{j=1}^{k} n_j(\overline{y}^{(j)} - \overline{y})^2$$

Ou seja, *esta parcela mede a dispersão das médias de* $\mathbf{y}$ *de cada classe, em torno da média geral dos valores de* $\mathbf{y}$. *Uma "boa" combinação linear deverá produzir valores elevados desta parcela*, uma vez que tal facto reflectirá a **heterogeneidade entre classes** dessa variável $\mathbf{y}$. Designaremos esta parcela por **variabilidade inter-classes** dos dados.

Resumindo: A decomposição da combinação linear (centrada) $\mathbf{y} = (\mathbf{I} - \mathbf{P_{1_n}})\mathbf{Xa}$ na sua parcela projectada sobre o subespaço gerado pelas colunas da matriz da classificação e a sua parcela no respectivo complemento ortogonal gera uma aplicação do Teorema de Pitágoras que se resume na frase: **o numerador da variância dos indivíduos no eixo $\mathbf{y} = \mathbf{Xa}$ resulta da soma da variabilidade intra-classes com a variabilidade inter-classes**. Uma vez que a variabilidade total de $\mathbf{y}$ não depende da classificação definida pela matriz $\mathbf{C}$, tem-se que **uma combinação linear adequada para salientar a estrutura de subgrupos será um vector $\mathbf{Xa}$ que *minimize a variabilidade intra-classes* e, ao fazê-lo, estará simultaneamente a *maximizar a variabilidade inter-classes***.

Como determinar essa combinação linear, *i.e.*, como determinar o vector de coeficientes $\mathbf{a}$ na combinação $\mathbf{y} = \mathbf{Xa}$? A fim de facilitar a obtenção dessa solução, a expressão acima obtida será re-escrita em *notação matricial*. Assim:

$$\|(\mathbf{I}_n - \mathbf{P_{1_n}})\mathbf{Xa}\|^2 = \|(\mathbf{P_C} - \mathbf{P_{1_n}})\mathbf{Xa}\|^2 + \|(\mathbf{I}_n - \mathbf{P_C})\mathbf{Xa}\|^2$$

$$\iff \mathbf{a}^t\mathbf{X}^t(\mathbf{I}_n - \mathbf{P_{1_n}})\mathbf{Xa} = \mathbf{a}^t\mathbf{X}^t(\mathbf{P_C} - \mathbf{P_{1_n}})\mathbf{Xa} + \mathbf{a}^t\mathbf{X}^t(\mathbf{I}_n - \mathbf{P_C})\mathbf{Xa}$$

Designe-se:

$\mathbf{\Sigma} = \frac{1}{n}\mathbf{X}^t(\mathbf{I}_n - \mathbf{P_{1_n}})\mathbf{X}$	**Matriz de variâncias-covariâncias de X**
$\mathbf{H} = \frac{1}{n}\mathbf{X}^t(\mathbf{P_C} - \mathbf{P_{1_n}})\mathbf{X}$	**Matriz da variabilidade inter-classes**
$\mathbf{E} = \frac{1}{n}\mathbf{X}^t(\mathbf{I}_n - \mathbf{P_C})\mathbf{X}$	**Matriz da variabilidade intra-classes**

Tem-se:

$$\mathbf{\Sigma} = \mathbf{H} + \mathbf{E} \tag{3.6}$$

Exercício 3.2 *Demonstrar esta relação.*

A equação resultante do Teorema de Pitágoras pode agora re-escrever-se de forma simples como:

$$\mathbf{a}^t\mathbf{\Sigma}\mathbf{a} = \mathbf{a}^t\mathbf{H}\mathbf{a} + \mathbf{a}^t\mathbf{E}\mathbf{a} \tag{3.7}$$

Com base nestas novas designações, é possível re-formular o objectivo da Análise Discriminante que já havia sido enunciado na Secção 3.1: **de entre as combinações lineares Xa, escolher a que *maximiza* o quociente:**

$$\frac{\mathbf{a}^t\mathbf{H}\mathbf{a}}{\mathbf{a}^t\mathbf{E}\mathbf{a}} \tag{3.8}$$

Essa será a **primeira função discriminante**[4], ou **primeiro eixo discriminante em** $\mathbb{R}^n$.

Assim, o problema de identificar a *combinação linear que maximiza a discriminação*, é um caso particular do problema geral de maximização de um quociente de formas quadráticas, problema estudado no Teorema 1.38 (p.38). Sabemos então que, *se* $\mathbf{E}$ *fôr uma matriz definida positiva* **o vector de coeficientes a que se procura é o vector próprio da matriz** $\mathbf{E}^{-1}\mathbf{H}$ associado ao maior valor próprio de $\mathbf{E}^{-1}\mathbf{H}$, digamos o valor λ_1. Chegámos, pois, à primeira solução do nosso problema.

A existência da solução acima indicada depende da existência da inversa da matriz $\mathbf{E}$. Ora, $\mathbf{E}$ é uma matriz de tipo $p \times p$. Será invertível se fôr de característica plena p (ver o ponto 3, página 29). Uma vez que a característica de um produto de matrizes não pode exceder a menor das características dos factores nesse produto (ponto 1.19, página 29), tem-se:

$$\begin{aligned} car\left(\mathbf{X}^t(\mathbf{I}_n - \mathbf{P_C})\mathbf{X}\right) &\leq \min\{car(\mathbf{X}), car(\mathbf{I}_n - \mathbf{P_C}), car(\mathbf{X}^t)\} \\ &= \min\{p, n-k\} \qquad (\text{admitindo } car(\mathbf{X}) = p) \end{aligned}$$

Logo, se $k > n - p$, $\mathbf{E}$ não pode ser invertível. Em geral, para $k \leq n - p$ haverá invertibilidade.

A razão de ser do adjectivo "primeira" nas conclusões anteriores advém do facto de podermos estar interessados em determinar *novas combinações lineares discriminantes*, caso o primeiro eixo discriminante tenha uma fraca capacidade discriminante (e caso haja mais do que dois subgrupos de indivíduos, por razões que adiante se compreenderão). Tais novas combinações lineares deverão ser soluções dum problema análogo ($(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{X}\mathbf{a}$ mais próxima de $\mathcal{C}(\mathbf{C})$), mas agora sujeito à condição adicional de serem não-correlacionadas com a(s) solução(ões) anterior(es), isto é, de $\mathbf{a}^t\mathbf{\Sigma}\mathbf{a}_1 = 0$. Tendo em conta as propriedades de $\mathbf{a}_1$ (sabemos ser um vector próprio de $\mathbf{E}^{-1}\mathbf{H}$, associado ao valor próprio λ_1), podemos

[4] Assinale-se que, ao contrário do que acontece numa Análise em Componentes Principais, não é necessário impor qualquer exigência sobre a dimensão do vector de coeficientes $\mathbf{a}$. De facto, multiplicações do vector de coeficientes $\mathbf{a}$ por um escalar deixam invariante o quociente $\frac{\mathbf{a}^t\mathbf{H}\mathbf{a}}{\mathbf{a}^t\mathbf{E}\mathbf{a}}$, pelo que o critério (3.8) depende apenas da direcção do vector $\mathbf{a}$, e não da sua magnitude.

re-escrever a condição $\mathbf{a}^t\mathbf{\Sigma}\mathbf{a}_1 = 0$ sob a forma $\mathbf{a}^t\mathbf{E}\mathbf{a}_1 = 0$. De facto, se $\mathbf{E}^{-1}\mathbf{H}\mathbf{a}_1 = \lambda_1\mathbf{a}_1$, tem-se também $\mathbf{H}\mathbf{a}_1 = \lambda_1\mathbf{E}\mathbf{a}_1$. Logo, o produto $\mathbf{a}^t\mathbf{\Sigma}\mathbf{a}_1 = \mathbf{a}^t(\mathbf{E}+\mathbf{H})\mathbf{a}_1 = \mathbf{a}^t(\mathbf{E}+\lambda_1\mathbf{E})\mathbf{a}_1 = (1+\lambda_1)\mathbf{a}^t\mathbf{E}\mathbf{a}_1$. Assim, exigir a condição $\mathbf{a}^t\mathbf{\Sigma}\mathbf{a}_1 = 0$ equivale a exigir $\mathbf{a}^t\mathbf{E}\mathbf{a}_1 = 0$ (recorde-se que $\lambda_1 > 0$).

O teorema 1.38 (p.38) assegura-nos que **sucessivas soluções são combinações lineares $\mathbf{X}\mathbf{a}_j$ onde $\mathbf{a}_j$ são os restantes vectores próprios da matriz $\mathbf{E}^{-1}\mathbf{H}$ associados a valores próprios não-nulos. A *capacidade discriminante* destes novos eixos é medida pelo valor próprio λ_j associado.**

A combinação linear das variáveis (centradas) que maximiza a razão da variabilidade inter-classes para a variabilidade intra-classes é $\mathbf{X}\mathbf{a}_1$, sendo $\mathbf{a}_1$ o vector próprio associado ao maior valor próprio da matriz $\mathbf{E}^{-1}\mathbf{H}$. O valor λ_1 dessa razão é uma medida da *capacidade discriminante* de $\mathbf{X}\mathbf{a}_1$. Quanto maior λ_1, maior a capacidade discriminante de $\mathbf{X}\mathbf{a}_1$. Esta combinação linear designa-se a **primeira função** ou **eixo discriminante** dos dados. O vector de coeficientes dessa combinação linear, $(\mathbf{a}_1)$ designa-se o **primeiro eixo discriminante em $\mathbb{R}^p$.** Sucessivos eixos discriminantes $\mathbf{X}\mathbf{a}_j$, $(j = 2 : k-1)$, maximizam a capacidade discriminante de entre as combinações lineares das colunas de $\mathbf{X}$ não-correlacionadas com os eixos discriminantes obtidos anteriormente. São definidos pelos vectores próprios $\mathbf{a}_j$ da matriz $\mathbf{E}^{-1}\mathbf{H}$, associados aos restantes valores próprios não-nulos, λ_j $(j = 2 : k-1)$. Esses valores próprios, que representam igualmente a razão da variabilidade inter-classes para a variabilidade intra-classes asociada a esses eixos, medem a capacidade discriminante dos restantes eixos.

Observações:

1. **A matriz $\mathbf{E}^{-1}\mathbf{H}$ não pode ter mais de $k-1$ valores próprios não-nulos, se a classificação subjacente (determinada pelas colunas da matriz C) tiver k classes.** De facto, pelas propriedades das características de matrizes (pg. 29) e a definição da matriz $\mathbf{H}$ (pg. 96), tem-se:

$$car(\mathbf{E}^{-1}\mathbf{H}) \leq car(\mathbf{H}) = car(\mathbf{X}^t(\mathbf{P_C} - \mathbf{P_{1}}_n)\mathbf{X}) \leq car(\mathbf{P_C} - \mathbf{P_{1}}_n) = k-1.$$

   Assim, $\mathbf{E}^{-1}\mathbf{H}$ terá, quanto muito, $k-1$ valores próprios não-nulos, sendo portanto esse o número máximo de eixos discriminantes que se podem obter.

2. Contrariamente à ACP, o facto de as novas *funções discriminantes* $\mathbf{X}\mathbf{a}_j$ serem não-correlacionadas (*i.e.*, ortogonais em $\mathbb{R}^n$) *não* significa que os vectores de coeficientes $\mathbf{a}_j$ também sejam ortogonais em $\mathbb{R}^p$. Como vimos, esses vectores de coeficientes serão $\mathbf{E}$-ortogonais.

3. Também diversamente daquilo que acontece na ACP, **as soluções duma Análise Discriminante são invariantes a transformações lineares das escalas das variáveis.** Efectuar uma tal transformação linear corresponde a multiplicar a matriz de dados $\mathbf{X}$ por uma matriz diagonal $\mathbf{D}$:

$$\mathbf{X} \longrightarrow \mathbf{XD}$$

   As funções discriminantes dos dados transformados $\mathbf{XD}$ são combinações lineares da forma $\mathbf{XDb}$, onde $\mathbf{b}$ são vectores próprios da matriz análoga à matriz $\mathbf{E}^{-1}\mathbf{H}$, mas para os dados transformados,

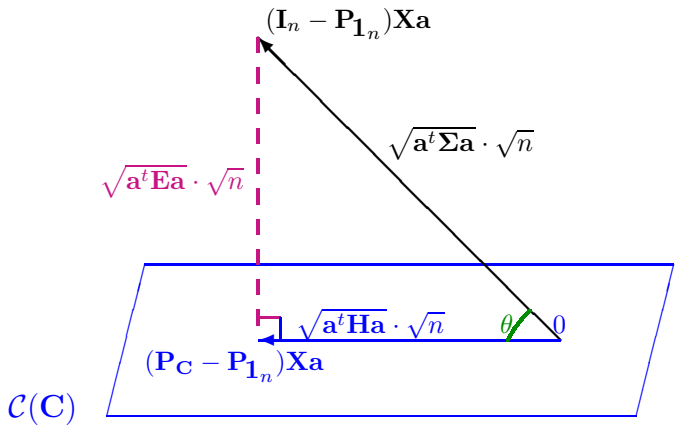
que facilmente se verifica ser a matriz $(\mathbf{DED})^{-1}(\mathbf{DHD})$. Assim, são vectores tais que:

$$(\mathbf{DED})^{-1}(\mathbf{DHD})\mathbf{b} = \mu\mathbf{b} \quad \Leftrightarrow \quad \mathbf{D}^{-1}\mathbf{E}^{-1}\mathbf{HDb} = \mu\mathbf{b} \quad \Leftrightarrow \quad \mathbf{E}^{-1}\mathbf{H}(\mathbf{Db}) = \mu(\mathbf{Db})$$

Logo, os vectores $\mathbf{Db}$ são vectores próprios da matriz $\mathbf{E}^{-1}\mathbf{H}$, pelo que as funções discriminantes dos dados originais são dadas por $\mathbf{X}(\mathbf{Db})$, ou seja, coincidem com as funções discriminantes dos dados transformados. (Embora os vectores de coeficientes em cada caso, $\mathbf{Db}$ e $\mathbf{b}$, não coincidam).

4. As funções discriminantes posteriores à primeira podem ser úteis (no caso de $k > 2$) para distinguir entre algumas classes cuja discriminação na primeira função discriminante seja fraca.

RESUMO



1. O Teorema de Pitágoras garante que $\mathbf{a}^t\mathbf{\Sigma a} = \mathbf{a}^t\mathbf{Ea} + \mathbf{a}^t\mathbf{Ha}$.
2. O critério de maximizar a razão da variabilidade inter-classes em relação à variabilidade intra-classes, *i.e.*, de maximizar o quociente $\frac{\mathbf{a}^t\mathbf{Ha}}{\mathbf{a}^t\mathbf{Ea}}$ corresponde a maximizar $\text{ctg}^2(\theta)$, *i.e.*, o quadrado da cotangente do ângulo entre o vector centrado $(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{Xa}$ e a sua projecção ortogonal sobre o subespaço gerado pelas colunas da matriz da classificação, $\mathbf{C}$. As funções discriminantes são os vectores desse subespaço que formam o menor ângulo com o vector original.
3. Os valores de $\text{ctg}^2(\theta)$, para cada função discriminante $\mathbf{Xa}_j$, são os valores próprios λ_j da matriz $\mathbf{E}^{-1}\mathbf{H}$ e representam a razão da variabilidade inter-classes sobre a variabilidade intra-classes, para os valores dessas combinação linear $\mathbf{Xa}_j$.
4. As funções discriminantes são invariantes a mudanças de escala nas variáveis.

3.3 A classificação de novos indivíduos

A utilização destes resultados da análise discriminante pode ser útil, como se referiu no início, quer para obter *representações gráficas* a baixa dimensão onde seja mais clara a distinção entre as classes, quer para obter *regras de classificação* de novos indivíduos nas classes. Assim, por exemplo, **observando um novo indivíduo nas p variáveis originais teremos um vector $p \times 1$ de observações que designaremos por x. O valor desse indivíduo na primeira função discriminante será** $y^* = \mathbf{x}^t \mathbf{a}_1$. Comparando este valor com as k médias de classe nessa função, $\overline{y}^{(1)}, \overline{y}^{(2)}, \ldots, \overline{y}^{(k)}$, poderemos:

1. **classificar o novo indivíduo na classe cujo centro de gravidade lhe está mais próxima, na habitual distância euclidiana**, isto é, associar o indivíduo à classe i se $|y^* - \overline{y}^{(i)}| < |y^* - \overline{y}^{(j)}|$, $\forall j \neq i$.

2. **classificar o novo indivíduo na classe cujo centro de gravidade lhe está mais próxima, ponderando a habitual distância euclidiana pelo inverso do desvio padrão das observações dessa classe**, isto é, associar o indivíduo à classe i se $\left|\frac{y^* - \overline{y}^{(i)}}{\sigma_y^{(i)}}\right| < \left|\frac{y^* - \overline{y}^{(j)}}{\sigma_y^{(j)}}\right|$, $\forall j \neq i$, onde $\sigma_y^{(i)}$ indica o desvio padrão das observações pertencentes ao grupo i, no eixo discriminante.

Esta segunda regra de classificação visa levar em consideração que a importância relativa de uma dada distância euclidiana pode ser diferente para classes muito homogéneas e para classes pouco homogéneas. Por exemplo, considere uma observação que se encontra à distância 1 dos centros de gravidade de duas diferentes classes. Se, no entanto, uma dessas classes tiver desvio padrão 0.1, enquanto que a outra tem desvio padrão 1.5, é bem mais provável que a observação à distância 1 pertença à segunda dessas classes.

Considere-se agora a situação mais geral, onde se definem q eixos discriminantes pelas combinações lineares das colunas de $\mathbf{X}$ cujos coeficientes são dados pelas colunas da matriz $\mathbf{A}_q$. Seja $\overline{\mathbf{y}}^{(i)}$ o vector das médias, nos q eixos discriminantes, das observações do i-ésimo grupo. Seja $\mathbf{x}$ uma nova observação nas p variáveis originais, que gera o ponto $\mathbf{y}^* = \mathbf{x}^t \mathbf{A}_q$ no espaço q-dimensional definido pelos q eixos discriminantes. Podem definir-se as seguintes regras de classificação de uma nova observação:

1. **classificar o novo indivíduo na classe cujo centro de gravidade lhe está mais próxima, na habitual distância euclidiana**, isto é, associar o indivíduo à classe i se $\|\mathbf{y}^* - \overline{\mathbf{y}}^{(i)}\| < \|\mathbf{y}^* - \overline{\mathbf{y}}^{(j)}\|$, $\forall j \neq i$, onde $\|\cdot\|$ representa a habitual distância euclidiana em $\mathbb{R}^q$.

2. **classificar o novo indivíduo na classe cujo centro de gravidade lhe está mais próxima na distância de Mahalanobis definida pela matriz de variâncias-covariâncias das observações dessa classe**, isto é, associar o indivíduo à classe i se $\|\mathbf{y}^* - \overline{\mathbf{y}}^{(i)}\|_{\mathbf{\Sigma}_i^{-1}} < \|\mathbf{y}^* - \overline{\mathbf{y}}^{(j)}\|_{\mathbf{\Sigma}_j^{-1}}$, $\forall j \neq i$, onde $\mathbf{\Sigma}_i$ indica a matriz de (co-)variâncias das observações do grupo i, e $\|\cdot\|_{\mathbf{\Sigma}_i^{-1}}$ indica a norma definida pela matriz definida positiva $\mathbf{\Sigma}_i^{-1}$.

Embora estejamos a fazer uma abordagem exclusivamente descritiva (não probabilística), refira-se que esta última regra, que se baseia nas distâncias de Mahalanobis, é equivalente a uma regra probabilística

em que se admite que o grupo i tem uma distribuição associada Multinormal, de vector médio $\boldsymbol{\mu}_i$ e matriz de (co)variâncias $\boldsymbol{\Sigma}_i$, e se associa uma nova observação ao grupo para o qual a nova observação é de *verosimilhança máxima*.

3.4 Formulações alternativas para o critério de discriminação

As funções discriminantes de Fisher foram obtidas através do critério de maximizar a razão da variabilidade inter-classes sobre a variabilidade intra-classes. Como se viu no resumo da página 99, este critério corresponde à minimização do ângulo θ entre o vector centrado da combinação linear das colunas de $\mathbf{X}$ e o subespaço gerado pelas colunas da matriz de classificação, $\mathbf{C}$ (isto é, o ângulo entre o vector centrado e a sua projecção ortogonal sobre o subespaço $\mathcal{C}(\mathbf{C})$). É geometricamente evidente que o critério de maximizar a função cotangente do ângulo θ pode ser substituído por um critério análogo que utilize qualquer outra função monótona do referido ângulo[5]. Em particular, poderiam formular-se os seguintes critérios:

1. ***Minimizar* o quadrado do *seno* do ângulo θ**, ou seja, **minimizar a proporção da variabilidade total da combinação linear Xa que corresponde à variabilidade intra-classes**

$$\frac{\mathbf{a}^t\mathbf{E}\mathbf{a}}{\mathbf{a}^t\boldsymbol{\Sigma}\mathbf{a}} \tag{3.9}$$

2. ***Maximizar* o quadrado do *cosseno* do ângulo θ** (que, recorde-se, é uma função *decrescente* do ângulo, no primeiro quadrante), ou seja, **maximizar a proporção da variabilidade total da combinação linear Xa que corresponde à variabilidade inter-classes**

$$\frac{\mathbf{a}^t\mathbf{H}\mathbf{a}}{\mathbf{a}^t\boldsymbol{\Sigma}\mathbf{a}} \tag{3.10}$$

Estas duas novas formulações, bem como a anterior, correspondem a um único problema geométrico: o de obter o vector $\mathbf{Xa}$ cuja projecção ortogonal (após centragem) sobre o subespaço $\mathcal{C}(\mathbf{C})$ minimize o ângulo θ. Sendo o problema o mesmo, as soluções também terão de ser as mesmas. Em termos algébricos, isso significa que os vectores próprios $\mathbf{a}$ da matriz $\mathbf{E}^{-1}\mathbf{H}$ (que sucessivamente maximizam o quociente $\frac{\mathbf{a}^t\mathbf{H}\mathbf{a}}{\mathbf{a}^t\mathbf{E}\mathbf{a}}$), terão de ser também vectores próprios das matrizes $\boldsymbol{\Sigma}^{-1}\mathbf{E}$ e $\boldsymbol{\Sigma}^{-1}\mathbf{H}$ associadas à optimização dos critérios (3.9) e (3.10), respectivamente. Confirmemos, algebricamente, que assim é.

Seja $(\lambda, \mathbf{a})$ um par de valor e vector próprio da matriz $\mathbf{E}^{-1}\mathbf{H}$, isto é, tal que:

$$\mathbf{E}^{-1}\mathbf{H}\mathbf{a} \;=\; \lambda\mathbf{a}.$$

[5] Monótona no primeiro quadrante, uma vez que os ângulos envolvidos em projecções ortogonais são sempre ângulos agudos.

Tendo em conta a relação (3.6) (p. 97), verifica-se:

$$\begin{aligned}
\mathbf{E}^{-1}\mathbf{H}\mathbf{a} = \lambda\mathbf{a} &\iff \mathbf{E}^{-1}(\boldsymbol{\Sigma}-\mathbf{E})\,\mathbf{a} = \lambda\mathbf{a} \\
&\iff \mathbf{E}^{-1}\boldsymbol{\Sigma}\mathbf{a}-\mathbf{a} = \lambda\mathbf{a} \\
&\iff \mathbf{a}-\boldsymbol{\Sigma}^{-1}\mathbf{E}\mathbf{a} = \lambda\boldsymbol{\Sigma}^{-1}\mathbf{E}\mathbf{a} \\
&\iff \mathbf{a} = (\lambda+1)\boldsymbol{\Sigma}^{-1}\mathbf{E}\mathbf{a} \\
&\iff \boldsymbol{\Sigma}^{-1}\mathbf{E}\mathbf{a} = \left(\frac{1}{\lambda+1}\right)\mathbf{a} \qquad (3.11)
\end{aligned}$$

Assim, **um vector próprio de $\mathbf{E}^{-1}\mathbf{H}$, com valor próprio associado λ, é também um vector próprio de $\boldsymbol{\Sigma}^{-1}\mathbf{E}$, com valor próprio associado $\frac{1}{\lambda+1}$**. Esta diferença nos valores próprios corresponde às diferentes funções trigonométricas do ângulo θ usadas em cada formulação. Assim, se λ é o quadrado da cotangente de θ (isto é, o valor do quociente $\frac{\mathbf{a}^t\mathbf{H}\mathbf{a}}{\mathbf{a}^t\mathbf{E}\mathbf{a}}$), $\frac{1}{\lambda+1}$ é o quadrado do seno de θ (isto é, o valor do quociente $\frac{\mathbf{a}^t\mathbf{E}\mathbf{a}}{\mathbf{a}^t\boldsymbol{\Sigma}\mathbf{a}}$)[6]. De forma análoga, a partir da relação (3.11) sai que:

$$\begin{aligned}
\mathbf{E}^{-1}\mathbf{H}\mathbf{a} = \lambda\mathbf{a} &\iff \boldsymbol{\Sigma}^{-1}\mathbf{E}\mathbf{a} = \left(\frac{1}{\lambda+1}\right)\mathbf{a} \\
&\iff \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-\mathbf{H})\mathbf{a} = \left(\frac{1}{\lambda+1}\right)\mathbf{a} \\
&\iff \mathbf{a}-\boldsymbol{\Sigma}^{-1}\mathbf{H}\mathbf{a} = \left(\frac{1}{\lambda+1}\right)\mathbf{a} \\
&\iff \boldsymbol{\Sigma}^{-1}\mathbf{H}\mathbf{a} = \mathbf{a}-\left(\frac{1}{\lambda+1}\right)\mathbf{a} \\
&\iff \boldsymbol{\Sigma}^{-1}\mathbf{H}\mathbf{a} = \left(1-\frac{1}{\lambda+1}\right)\mathbf{a} \\
&\iff \boldsymbol{\Sigma}^{-1}\mathbf{H}\mathbf{a} = \left(\frac{\lambda}{\lambda+1}\right)\mathbf{a} \qquad (3.12)
\end{aligned}$$

Assim, **um vector próprio de $\mathbf{E}^{-1}\mathbf{H}$, com valor próprio associado λ, é também um vector próprio de $\boldsymbol{\Sigma}^{-1}\mathbf{H}$, com valor próprio $\frac{\lambda}{\lambda+1}$**[7].

Observações:

1. Acabou-se de ver que maximizar a proporção de variabilidade inter-classes equivale a minimizar a proporção de variabilidade intra-classes, sendo ambas estas abordagens equivalentes a maximizar a razão entre as variabilidades inter- e intra-classes. Os valores destes quocientes (dados pelos valores próprios não-nulos das matrizes $\boldsymbol{\Sigma}^{-1}\mathbf{E}$, $\boldsymbol{\Sigma}^{-1}\mathbf{H}$ e $\mathbf{E}^{-1}\mathbf{H}$, respectivamente) diferem, mas os eixos discriminantes resultantes coincidem.

[6] Como se pode confirmar através de relações trigonométricas elementares: $\lambda = \mathrm{ctg}^2(\theta) = \frac{\cos^2(\theta)}{\sin^2(\theta)} = \frac{1-\sin^2(\theta)}{\sin^2(\theta)} = \frac{1}{\sin^2(\theta)} - 1 \Leftrightarrow \lambda+1 = \frac{1}{\sin^2(\theta)} \Leftrightarrow \frac{1}{\lambda+1} = \sin^2(\theta)$.

[7] Se λ é o quadrado da cotangente de θ, $\frac{\lambda}{\lambda+1}$ é o quadrado do cosseno de θ.

2. Caso procurassemos resolver directamente a nova formulação do problema de minimizar o quociente $\frac{\mathbf{a}^t\mathbf{E}\mathbf{a}}{\mathbf{a}^t\mathbf{\Sigma}\mathbf{a}}$, obteríamos, por analogia com o que foi feito antes, que os vectores $\mathbf{a}$ teriam de resolver a equação matricial $\mathbf{E}\mathbf{a} = \mu\mathbf{\Sigma}\mathbf{a}$, logo seriam vectores próprios da matriz $\mathbf{\Sigma}^{-1}\mathbf{E}$. Uma vez que agora o objectivo é o de *minimizar*, seria necessário tomar em primeiro lugar o vector próprio associado ao *menor* valor próprio (não-nulo) dessa matriz.

3. Os valores dos critérios nas duas novas formulações (3.9) e (3.10) não podem exceder 1. Este facto (que é evidente por se tratarem de quadrados de valores de funções seno e cosseno) é geometricamente evidente , uma vez que pela fórmula (3.7) (página 97), cujas parcelas são necessariamente não-negativas, os quocientes $\frac{\mathbf{a}^t\mathbf{E}\mathbf{a}}{\mathbf{a}^t\mathbf{\Sigma}\mathbf{a}}$ e $\frac{\mathbf{a}^t\mathbf{H}\mathbf{a}}{\mathbf{a}^t\mathbf{\Sigma}\mathbf{a}}$ não podem tomar valores superiores à unidade.

4. Os valores dos critérios nas formulações alternativas podem, também, ser considerados **medidas da capacidade discriminante** dos eixos. **Os valores próprios da matriz $\mathbf{\Sigma}^{-1}\mathbf{H}$** (ou seja, os valores do quociente $\frac{\mathbf{a}^t\mathbf{H}\mathbf{a}}{\mathbf{a}^t\mathbf{\Sigma}\mathbf{a}}$), em particular, **são de fácil interpretação: estando entre zero e um, valores mais próximos de um indicam uma maior capacidade discriminante**. Um valor 1 indicaria que, na combinação linear $\mathbf{X}\mathbf{a}$ das variáveis observadas, a totalidade da variabilidade dos dados corresponde a variabilidade *entre* classes. Necessariamente, esta situação estaria associada ao facto de todas as observações de um mesmo subgrupo terem, na referida combinação linear, o mesmo valor, ou seja, $\mathbf{X}\mathbf{a} \in \mathcal{C}(\mathbf{C})$. Atenção ao facto de que, no caso de ser usado o critério $\frac{\mathbf{a}^t\mathbf{E}\mathbf{a}}{\mathbf{a}^t\mathbf{\Sigma}\mathbf{a}}$, serão os *menores* valores a indicar uma melhor capacidade discriminante.

5. Diferentes textos e programas estatísticos podem utilizar uma ou outra destas três abordagens alternativas, sendo por isso da maior importância confirmar qual a definição exacta da medida de capacidade discriminante que nos é apresentada.

Exercício 3.3 *Resolver directamente os problemas colocados pelas formulações alternativas do critério e verificar os resultados que acabam de ser referidos.*

3.5 Uma abordagem no contexto da Análise de Variância

Embora não seja a forma mais habitual de introduzir a Análise Discriminante de Fisher, é possível relacionar o método utilizado para identificar as combinações lineares discriminantes com uma Análise de Variância a 1 Factor (totalmente casualizada e de efeitos fixos), em que os níveis do referido Factor são os subgrupos a que pertencem os n indivíduos observados. De facto, para qualquer combinação linear $\mathbf{X}\mathbf{a}$ das variáveis observadas, uma medida do maior ou menor grau de separação dos subgrupos pode ser dada pela estatística F do teste à existência de efeitos do factor, numa Análise de Variância a um factor, utilizando a combinação linear $\mathbf{X}\mathbf{a}$ como a variável resposta da análise. Um critério de escolha da combinação linear mais adequada para a separação dos subgrupos pode, assim, ser dado da seguinte forma: **escolher a combinação linear Xa que, usada como variável resposta numa ANOVA a 1 Factor, em que os subgrupos constituem os níveis do Factor, torne mais elevada a correspondente estatística F**. Na notação da disciplina de Modelação Estatística I, a estatística F

do teste ANOVA é dada por:

$$F = \frac{QMM}{QMRE} = \frac{SQM}{SQRE} \cdot \frac{n-(p+1)}{p} = \frac{\|(\mathbf{P_x} - \mathbf{P_{1_n}})\mathbf{y}\|^2}{\|(\mathbf{I}_n - \mathbf{P_x})\mathbf{y}\|^2} \cdot \frac{n-(p+1)}{p}$$

em que $p+1$ indica o número de níveis do Factor (subgrupos, no nosso contexto); n representa o número de indivíduos observados; $\mathbf{y}$ é a variável resposta; $\mathbf{P_x}$ representa a matriz de projecção ortogonal sobre o subespaço gerado pela matriz do delineamento (a matriz $\mathbf{C}$), sendo $\mathbf{I}_n - \mathbf{P_x}$ a matriz de projecção no complemento ortogonal desse espaço, e $\mathbf{P_x} - \mathbf{P_{1_n}}$ a matriz de projecção ortogonal sobre a intersecção desse subespaço com o subespaço $\mathcal{C}(\mathbf{1}_n)^\perp$, constituído pelos vectores centrados de $\mathbb{R}^n$. Importa relembrar que, *no nosso contexto, a variável resposta é dada por* $\mathbf{y} = \mathbf{Xa}$, *sendo* $\mathbf{X}$ *a matriz* $n \times p$ *dos dados observados, e tendo, pois, um significado diferente daquilo a que se chama a matriz* $\mathbf{X}$ *no contexto do Modelo Linear.* Essa matriz (que estava na origem da notação $\mathbf{P_x}$ usada mais acima) era a matriz do delineamento, isto é, a matriz cujas colunas estavam associadas às variáveis indicatrizes de pertença aos níveis do factor. No contexto da Análise Discriminante, chama-se por vezes **matriz da classificação** a essa matriz do delineamento, e é a matriz que designámos $\mathbf{C}$. Do conjunto destas adaptações de notação, e deixando cair o factor multiplicativo envolvendo n e p (que é sempre igual no nosso caso) sai a seguinte formulação do problema de determinar a combinação linear das variáveis observadas, $\mathbf{Xa}$, que melhor discrimina entre k grupos (classes) de indivíduos:

$$\text{determinar } \mathbf{a} \in \mathbb{R}^p \text{ que maximize } \frac{\|(\mathbf{P_C} - \mathbf{P_{1_n}})\mathbf{Xa}\|^2}{\|(\mathbf{I}_n - \mathbf{P_C})\mathbf{Xa}\|^2} \tag{3.13}$$

A abordagem da Análise Discriminante efectuada nas Secções anteriores foi feita sem recurso a resultados anteriores sobre ANOVA. Uma tal abordagem reforça a validade do método num contexto não-inferencial, não fazendo recurso às hipóteses de multinormalidade associadas à Análise de Variância clássica.

3.6 Um exemplo

O conjunto de dados analisado na Secção 2.8 serve como exemplo simples de aplicação de uma Análise Discriminante, uma vez que as $n = 150$ observações diziam respeito a lírios de três diferentes variedades havendo, pois, uma estrutura de subgrupos implícita nos dados. Essa estrutura (que não foi explicitamente utilizada na ACP) é o suporte para a seguinte pergunta: qual a combinação linear das $p = 4$ variáveis morfométricas que melhor salienta a separação entre as 3 espécies de lírios?

A fim de efectuar os cálculos, apoiemo-nos sobre a função `lda` (as iniciais de *Linear Discriminant Analysis*) que se encontra no módulo `MASS` do programa R. Mais pormenores sobre esta função podem ser encontrados na respectiva documentação[8]. O comando `lda` funciona com a seguinte informação mínima: onde se encontram as variáveis cujas combinações lineares se deseja estudar, e qual o critério de divisão em subgrupos dos dados (através do argumento `grouping`).

[8] Acessível com o comando `help(lda)`.

Estudemos os dados dos lírios, não esquecendo que pode ser necessário carregar os dados e o módulo `MASS`. Recorde-se também a estrutura do objecto `iris` que contém os referidos dados: uma `data.frame`, cujas quatro primeiras colunas contêm as variáveis morfométricas, e cuja quinta coluna é um factor indicativo da espécie de cada flôr observada.

```
> library(MASS)
> lda(iris[,-5],grouping=iris[,5])
Call:
lda.data.frame(iris[, -5], grouping = iris[, 5])

Prior probabilities of groups:
    setosa versicolor  virginica
 0.3333333  0.3333333  0.3333333

Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa            5.006       3.428        1.462       0.246
versicolor        5.936       2.770        4.260       1.326
virginica         6.588       2.974        5.552       2.026

Coefficients of linear discriminants:
                    LD1         LD2
Sepal.Length  0.8293776  0.02410215
Sepal.Width   1.5344731  2.16452123
Petal.Length -2.2012117 -0.93192121
Petal.Width  -2.8104603  2.83918785

Proportion of trace:
   LD1    LD2
0.9912 0.0088
```

A informação mais importante da listagem de resultados produzida é dada em `Coefficients of linear discriminants`: cada coluna de valores indica os coeficientes das variáveis observadas que definem cada eixo discriminante (ou seja, cada coluna contêm um vector próprio $\mathbf{a}_i$ da matriz $\mathbf{E}^{-1}\mathbf{H}$). O programa *R* normaliza estes vectores de coeficientes de forma a que verifiquem[9] a condição $\mathbf{a}^t\mathbf{E}\mathbf{a} = 1$. Assim, o primeiro eixo discriminante é dado pela combinação linear:

$$Y = 0.8294\, Sepal.Length + 1.5345\, Sepal.Width - 2.2012\, Petal.Length - 2.8105\, Petal.Width$$

[9] O programa *R* também define a matriz de variabilidades intra-classes $\mathbf{E}$ com a multiplicação pelo escalar $\frac{1}{n-k}$, em vez de $\frac{1}{n}$, como se admitiu nestas folhas. Essa definição tem em conta preocupações de natureza inferencial, que não são relevantes nesta discussão de natureza meramente descritiva.

Havendo $k = 3$ grupos de lírios, podem determinar-se $k - 1 = 2$ eixos discriminantes, como é patente nos resultados apresentados. Como medida da capacidade discriminante de cada eixo, a função `lda` devolve a proporção que cada valor próprio representa no traço da matriz $\mathbf{E}^{-1}\mathbf{H}$, isto é, a dimensão relativa de cada quociente $\frac{\mathbf{a}^t\mathbf{H}\mathbf{a}}{\mathbf{a}^t\mathbf{E}\mathbf{a}}$. Para se obter os quocientes das variabilidades inter- e intra-classes, referidas na discussão, pode pedir-se explicitamente as raízes quadradas dos valores próprios da matriz $\mathbf{E}^{-1}\mathbf{H}$, através do objecto de saída de nome `svd` (não confundir com a função do `R` de nome igual):

```
> lda(iris[,-5], grouping=iris[,5])$svd
[1] 48.642644  4.579983
> lda(iris[,-5], grouping=iris[,5])$svd^2
[1] 2366.10680   20.97624
```

Neste caso, tem-se $\frac{\mathbf{a}_1^t\mathbf{H}\mathbf{a}_1}{\mathbf{a}_1^t\mathbf{E}\mathbf{a}_1} = 2366.11$, o que corresponde a dizer que esse é a cotangente ao quadrado do ângulo entre o primeiro eixo discriminante $\mathbf{X}\mathbf{a}_1$ e o subespaço coluna da matriz da classificação, $\mathcal{C}(\mathbf{C})$. Trata-se dum ângulo de pouco superior a 1º. (Repare-se que a "Proporção do traço" indicada no resultado que, por omissão, o `R` produz quando se invoca o comando `lda` corresponde a ao quociente $\frac{2366.10680}{2366.10680+20.97624} = 0.9912126$).

No exemplo dos lírios, é evidente que o segundo eixo discriminante não acrescenta quase nada à discriminação que é possível efectuar com base no primeiro eixo. A Figura 3.6 ilustra esta situação, tendo sido obtida através do comando `plot`:

```
> plot(lda(iris[,-5], grouping=iris[,5]),abbrev=TRUE, col=as.numeric(iris[,5]))
```

Esta figura é possível porque o *package* `MASS` tem um método para o comando `plot` lidar com o resultado de Análises Lineares Discriminantes (efectuadas com o comando `lda`) que consiste em representar graficamente os individuos nos eixos discriminantes. Com esse método, cada ponto é automaticamente representado por uma legenda que indica o grupo a que pertence (legendas essas que são obtidas a partir do vector indicado no argumento `grouping`). O argumento `abbrev`, quando tomando o valor lógico "verdade" abrevia os nomes dos grupos (indicados na quinta coluna do objecto `iris`), de forma a não sobrecarregar o aspecto visual do gráfico. Registe-se que o *R* efectua a centragem dos eixos discriminantes (utilizando as médias gerais de cada variável, para as n observações), ou seja, trabalha com os eixos discriminantes $(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{X}\mathbf{a}$, e não $\mathbf{X}\mathbf{a}$.

```
> iris.lda <- lda(iris[,-5], grouping=iris[,5])
> iris.pred <- predict(iris.lda, new=data.frame(Sepal.Length=5, Sepal.Width=3,
+                                Petal.Length=1.5, Petal.Width=0.15))
> iris.pred

$class
[1] setosa
Levels: setosa versicolor virginica
```

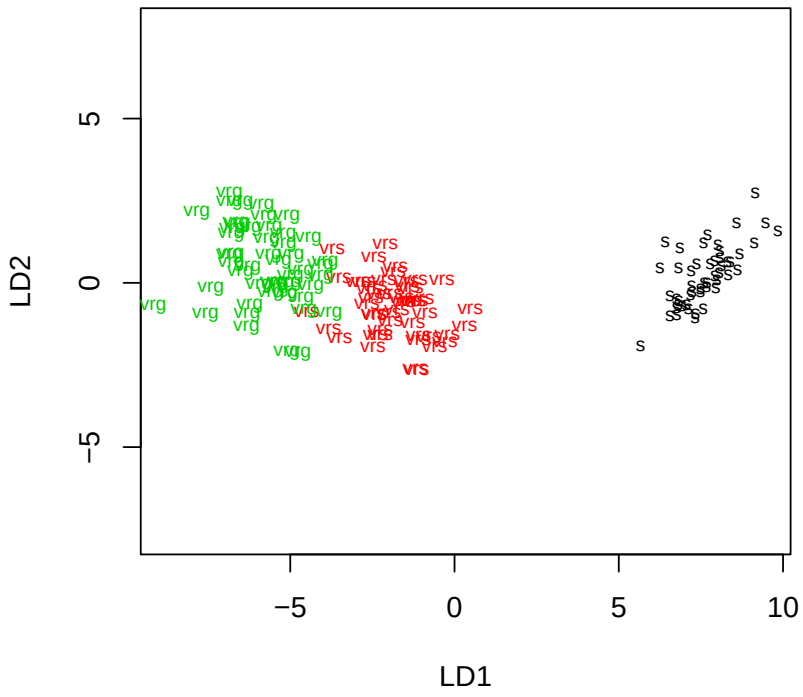


Figura 3.1: Representação dos $n = 150$ lírios nos dois Eixos Discriminantes. As coordenadas de cada ponto são dadas pelo valor do respectivo indivíduo nos dois eixos. Os símbolos que representam cada ponto são as abreviaturas das variedades de lírio correspondentes.

```
$posterior
     setosa   versicolor    virginica
[1,]      1 8.710725e-18 4.498939e-37
$x
          LD1       LD2
[1,] 7.132027 -1.019402
```

É possível colocar o ponto correspondente a esta nova observação em cima do gráfico, seleccionando a componente `x` do objecto de saída do comando anterior e utilizando o comando `points` para desenhar um ponto nas coordenadas correspondentes:

```
> points(iris.pred$x, pch=16, col="blue")
```

A fim de se poder classificar este novo ponto (o que já seria mais ou menos evidente a partir do posicionamento do ponto no gráfico), podemos calcular a média das observações de cada grupo nos eixos discriminantes. As coordenadas de cada um dos $n = 150$ pontos com que se fez o ajustamento, nos novos eixos discriminantes, podem ser pedidas usando de novo o comando `predict`, mas indicando como argumento `new` as próprias linhas da matriz de dados. Fazendo isso para as 50 linhas de cada espécie, e depois pedindo a média de cada coluna, teremos o valor médio por espécie em cada eixo discriminante:

```
> iris.pred.set <- predict(iris.lda, new=iris[iris[,5]=="setosa",-5])$x
> iris.pred.vrs <- predict(iris.lda, new=iris[iris[,5]=="versicolor",-5])$x
> iris.pred.vir <- predict(iris.lda, new=iris[iris[,5]=="virginica",-5])$x
> apply(iris.pred.set,2,"mean")
     LD1      LD2
7.607600 0.215133
> apply(iris.pred.vrs,2,"mean")
       LD1        LD2
-1.8250495 -0.7278996
> apply(iris.pred.vir,2,"mean")
       LD1        LD2
-5.7825504  0.5127666
```

Como foi visto, apenas será necessário considerar o primeiro eixo discriminante, no qual a nova observação tem coordenada 7.1320, muito próximo da média nesse eixo das observações do grupo *setosa* (7.6076) e muito longe das médias quer do grupo *versicolor* (-1.250), quer do grupo *virginica* (-5.7826). Assim, não parece haver grande dúvida que o novo indivíduo é da espécie *setosa*.

Repare-se que os desvios padrões de cada grupo no eixo discriminante 1 são todos muito próximos de 1 (o que se pode confirmar substituindo nos comandos acima o `mean` final por `sd`), pelo que a conclusão não seria alterada por utilizarmos a regra em que as distâncias são divididas pelos desvios padrões dos eixos:

Caso se utilize mais do que um eixo discriminante, podem ser calculadas as distâncias de Mahalanobis através dum comando do mesmo nome (ver `help(mahalanobis)`). Assim, e usando os dois eixos discriminantes do nosso exemplo, as distâncias de Mahalanobis do novo ponto (cujas coordenadas estão em `iris.pred$x`) até à média das observações de cada grupo, usando as matrizes de (co)variâncias de cada grupo podem ser obtidas da seguintes forma:

```
> mahalanobis(iris.pred$x, center=apply(iris.pred.set,2,"mean"), cov=var(iris.pred.set))
[1] 2.087987
> mahalanobis(iris.pred$x, center=apply(iris.pred.vrs,2,"mean"), cov=var(iris.pred.vrs))
[1] 78.96666
> mahalanobis(iris.pred$x, center=apply(iris.pred.vir,2,"mean"), cov=var(iris.pred.vir))
[1] 139.9377
```

3.7 Exercícios

AVISO: Alguns Exercícios deste Capítulo usam dados dos Exercícios do Capítulo de Análise em Componentes Principais. Alguns Exercícios precisam de novos conjuntos de dados que estão disponíveis na área de trabalho já referida no Capítulo sobre ACP. Para aceder a estes dados deve-se:

- Montar (*Map network drive*, no menu *Tools* do *My Computer*) a *drive*: \\`prunus`\`home`\`cadeiras`
- Abrir uma sessão do `R` (na directoria onde está a guardar o seu trabalho).
- A partir da sessão do `R`, seleccionar a opção *Files*, na barra de menus, e dentro da lista de opções disponibilizada, escolher *Load Workspace*.
- Na janela de diálogo que se abre, seleccionar a nova *drive* (que ficou associada ao `prunus`), depois a directoria `MMACB`, a seguir a directoria `em`, e finalmente o ficheiro `exerADL.RData`.

Na sessão do `R` deverão estar agora disponíveis os objectos `zebus` (Exercíco 2) e `videiras` (Exercício 4).

1. Considere de novo os dados morfométricos relativas a medições sobre lavagantes, considerados no Exercício 3 do Capítulo sobre Análise em Componentes Principais (pg. 79), e disponíveis na *data frame* `lavagantes`. Com o auxílio do `R` e da sua função `lda` (disponível no módulo `MASS`, como descrito na secção 3.6), responda às seguintes questões.

   (a) Efectue uma Análise Discriminante Linear, sabendo que os 42 primeiros individuos são machos e os 21 individuos restantes são fêmeas. Em particular,

      i. Justifique porque a função `lda` do `R` apenas produz um eixo discriminante.
      ii. Interprete a qualidade desse eixo discriminante na separação de machos e fêmeas.
      iii. Represente graficamente (com o auxílio das funções `predict` e `plot`) os 63 individuos, no eixo discriminante. Comente.
      iv. Calcule o coeficiente de correlação entre o eixo discriminante e as componentes principais do conjunto de dados. Comente.
      v. Descreva a matriz da classificação $\mathbf{C}$ associada a esta situação.
      vi. Determine o ângulo (em radianos e em graus) entre o eixo discriminante obtido e o espaço gerado pelas colunas da matriz da classificação $\mathbf{C}$.

   (b) Efectue uma nova Análise Discriminante Linear, sabendo que os 63 individuos se repartiam, na realidade, em três diferentes grupos de 21 individuos cada: machos reprodutores, machos não reprodutores e fêmeas. Em particular,

      i. Justifique a existência de dois eixos discriminantes.
      ii. Comente a qualidade dos eixos discriminantes obtidos, quer utilizando o critério clássico de Fisher, quer utilizando as formulações alternativas vistas nas aulas teóricas.

iii. Represente graficamente (com o auxílio das funções `predict` e `plot`) os 63 individuos, nos dois eixos discriminantes. Comente.

iv. Calcule o coeficiente de correlação entre os eixos discriminantes e as componentes principais do conjunto de dados. Comente.

v. Calcule o coeficiente de correlação entre os eixos discriminantes obtido considerando os três grupos e o único eixo discriminante obtido quando apenas se considerou a divisão entre machos e fêmeas. Comente.

2. Dez zebus e dez charolesas foram observados em três variáveis (v_1, v_2 e v_3). Os valores obtidos (dados de Diday *et. al.*, 1982) estão disponíveis na *data frame* `zebus` e são:

Zebus			Charolesas		
v_1	v_2	v_3	v_1	v_2	v_3
400	224	28.2	395	224	35.1
395	229	29.4	410	232	31.9
395	219	29.7	405	233	30.7
395	224	28.6	405	240	30.4
400	223	28.5	390	217	31.9
400	224	27.8	415	243	32.1
400	221	26.5	390	229	32.1
410	233	25.9	405	240	31.1
402	234	27.1	420	234	32.4
400	223	26.8	390	223	33.8

(a) Efectue uma Análise Discriminante dos dados e diga se considera que as 3 variáveis observadas permitem uma boa discriminação entre zebus e charolesas.

(b) Considere que chegaram os registos dos valores das três variáveis num novo animal: $v_1 = 403$, $v_2 = 231$, $v_3 = 31.1$. Calcule o *score* relativo a este individuo no eixo determinante. De que tipo de animal se trata?

(c) Calcule os coeficientes de correlação do eixo discriminante com cada uma das três variáveis. Diga se é possível interpretar a natureza do eixo discriminante, e o que isso sugere sobre a capacidade discriminante das variáveis originais.

3. No estudo sobre framboesas realizado na Secção de Horticultura do ISA e cujos dados foram utilizados no Exercício 4 de ACP (pg. 80), as framboesas foram colhidas em cinco datas diferentes. Tendo em conta esse facto, procurou-se verificar se seria possível discriminar as datas de corte através dos valores das variáveis observadas[10]. As datas de corte para cada planta foram:

[10] Naturalmente que as observações são insuficientes para uma análise deste tipo. No entanto, a pequena dimensão do conjunto de dados permite visualizar algumas questões importantes, que são referidas nas alíneas seguintes

Data de corte	Plantas
28 Novembro	1,2,3,4
13 Dezembro	5,6,7,8
16 Janeiro	9,10,11,12
20 Fevereiro	13
3 Abril	14

(a) Qual o número máximo de eixos discriminantes que será possível obter?

(b) Efectue uma Análise Discriminante Linear dos dados.

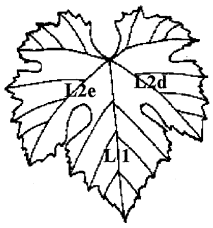
(c) Represente graficamente (com o auxílio das funções `predict` e `plot`) os 14 individuos, nos quatro eixos discriminantes. Comente.

(d) Qual o efeito, sobre as fórmulas da variabilidade intra- e inter-classes, resultantes do facto de dois dos grupos terem uma única observação?

(e) Repita a análise, mas utilizando agora *dados normalizados*. Qual espera que venham a ser os resultados? Confirme, vendo em particular os *scores* de cada indivíduo nos eixos dsicriminantes. Comente, em particular, os coeficientes de cada variável nos eixos discriminantes correspondentes, em cada caso. Qual tem de ser a relação entre estes coeficientes?

4. Na Secção de Horticultura do ISA foram seleccionadas 200 folhas de cada uma de três castas de videiras: Fernão Pires, Vital e Água Santa. Para cada folha obtiveram-se medições de:

   - áreas foliares **(Área)** (em cm^2);
   - o comprimento da nervura principal **(NP)** (em cm);
   - o comprimento da nervura lateral esquerda **(NLesq)** (em cm); e
   - o comprimento da nervura lateral direita **(NLdir)** (em cm).



   Os dados obtidos constam do objecto `videiras`. As suas 10 primeiras linhas são:

```
          Casta NLesq   NP NLdir Area
1  Fernao Pires  11.4 13.8  10.7  200
2  Fernao Pires   8.8  9.1   9.4  126
3  Fernao Pires  13.2 14.5  13.0  274
4  Fernao Pires  11.7 13.8  10.7  198
5  Fernao Pires   9.7 12.0  10.6  160
6  Fernao Pires  12.0 11.5  11.6  236
7  Fernao Pires  11.5 12.5  10.8  213
8  Fernao Pires   9.0  9.4   8.6  112
9  Fernao Pires  11.0 12.5  11.6  193
10 Fernao Pires  11.2 10.3  10.4  160
```

   (a) Efectue uma Análise Discriminante Linear que procure diferenciar as três castas a partir das variáveis observadas. Escreva a equação do primeiro eixo discriminante.

(b) Represente graficamente (com o auxílio das funções `predict` e `plot`) as 600 folhas, nos dois eixos discriminantes. Comente os resultados.

(c) Diga porque não é inteiramente correcta a seguinte afirmação: "não é possível distinguir as castas Vital e Água Santa a partir da forma das suas folhas".

5. Um estudo envolve nove tipos de medições sobre crânios de lobos *Canis lupus* L.. São efectuadas medições sobre 25 individuos, repartidos por 4 grupos: 6 machos do habitat 1, 3 fêmeas do habitat 1, 10 machos do habitat 2, 6 fêmeas do habitat 2. Os dados obtidos são (em mm.):

X1	X2	X3	X4	X5	X6	X7	X8	X9	Grupo
126	104	141	81.0	31.8	65.7	50.9	44.0	18.2	1
128	111	151	80.4	33.8	69.8	52.7	43.2	18.5	1
126	108	152	85.7	34.7	69.1	49.3	45.6	17.9	1
125	109	141	83.1	34.0	68.0	48.2	43.8	18.4	1
126	107	143	81.9	34.0	66.1	49.0	42.4	17.9	1
128	110	143	80.6	33.0	65.0	46.4	40.2	18.2	1
116	102	131	76.7	31.5	65.0	45.4	39.0	16.8	2
120	103	130	75.1	30.2	63.8	44.4	41.1	16.9	2
116	103	125	74.7	31.6	62.4	41.3	44.2	17.0	2
117	99	134	83.4	34.8	68.0	40.7	37.1	17.2	3
115	100	149	81.0	33.1	66.7	47.2	40.5	17.7	3
117	106	142	82.0	32.6	66.0	44.9	38.2	18.2	3
117	101	144	82.4	32.8	67.5	45.3	41.5	19.0	3
117	103	149	82.8	35.1	70.3	48.3	43.7	17.8	3
119	101	143	81.5	34.1	69.1	50.1	41.1	18.7	3
115	102	146	81.4	33.7	66.4	47.7	42.0	18.2	3
117	100	144	81.3	37.2	66.8	41.4	37.6	17.7	3
114	102	141	84.1	31.8	67.8	47.8	37.8	17.2	3
110	94	132	76.9	30.1	62.1	42.0	40.4	18.1	3
112	94	134	79.5	32.1	63.3	44.9	42.7	17.7	4
109	91	133	77.9	30.6	61.9	45.2	41.2	17.1	4
112	99	139	77.2	32.7	67.4	46.9	40.9	18.3	4
112	99	133	78.5	32.5	65.5	44.2	34.1	17.5	4
113	97	146	84.2	35.4	68.7	51.0	43.6	17.2	4
107	97	137	78.1	30.7	61.6	44.9	37.3	16.5	4

A matriz de correlações entre as 9 variáveis observadas (para a totalidade das 25 observações) é:

```
x[1]  1  1.000
x[2]  2  0.508  1.000
x[3]  3  0.404  0.491  1.000
x[4]  4  0.352  0.288  0.476  1.000
x[5]  5  0.569  0.349  0.505  0.563  1.000
x[6]  6  0.640  0.543  0.654  0.634  0.679  1.000
x[7]  7  0.302  0.285  0.691  0.349  0.116  0.595  1.000
x[8]  8  0.053 -0.183  0.380  0.205  0.145  0.224  0.443  1.000
x[9]  9  0.374  0.196  0.083 -0.278 -0.023  0.192  0.214  0.326  1.000
            1      2      3      4      5      6      7      8      9
```

Os resultados de uma Análise Discriminante num programa informático que opta pelo critério de maximizar a razão entre as variâncias inter-classes e as variâncias intra-classes, produz os seguintes resultados (que incluem uma nuvem de pontos nos dois primeiros eixos discriminantes):

```
*** Within-group means ***
                         1           2           3           4
x[1]            1        126.5       117.3       115.8       110.8
x[2]            2        108.2       102.7       100.8        96.2
x[3]            3        145.2       128.7       142.4       137.0
x[4]            4        82.12       75.50       81.68       79.23
x[5]            5        33.55       31.10       33.53       32.33
x[6]            6        67.28       63.73       67.07       64.73
x[7]            7        49.42       43.70       45.54       46.18
x[8]            8        43.20       41.43       39.99       39.97
x[9]            9        18.18       16.90       17.98       17.38
Constant       10        6.000       3.000      10.000       6.000

     *** Sums of squares and products ***
   1    100.60
   2     65.60    165.93
   3     95.63    149.30    557.90
   4     33.34     35.01    106.16     89.32
   5     42.64     33.58     89.13     39.80     55.91
   6     66.89     72.86    160.97     62.49     52.88    108.63
   7     38.21     46.28    206.12     41.58     10.92     78.31   159.30
   8      6.21    -27.75    105.57     22.81     12.80     27.46    65.81    138.49
   9      8.59      5.79      4.50     -6.02     -0.40      4.60     6.20      8.79      5.25
            1         2         3         4         5         6        7         8         9

*****  Canonical variate analysis  *****
***  Latent Roots  ***                              ***  Trace  ***
     l['Roots']                                       l['Trace']
              1           2           3
         20.516       6.326       0.749                   27.59

***  Percentage variation  ***
     l['Roots']
              1           2           3
          74.36       22.93        2.72

***  Latent Vectors (Loadings)  ***
             l['Vectors']
                       1           2           3
            1      0.5752      0.1764      0.0144
            2      0.2516     -0.1320      0.0511
            3     -0.1007      0.1810      0.1106
            4     -0.1994      0.5452     -0.0932
            5     -0.2143     -0.1745     -0.6717
            6     -0.1846     -0.3544      0.7015
            7      0.0369     -0.1041     -0.6510
            8      0.3322     -0.2247      0.1267
            9     -1.7654      1.8613     -0.2114
```

```
***  [Medias de grupo nas Variaveis Canonicas]  ***     ***  Inter-group distances  ***
                     1          2          3          1     0.000
            1       33.46      68.36       9.88          2     7.383     0.000
            2       32.43      61.17      11.18          3     8.578     9.778     0.000
            3       25.01      67.53      11.10          4    10.326     9.583     3.484     0.000
            4       23.77      64.82       9.29                   1         2         3         4

*** [Tres variaveis canonicas] ***
            1    33.72    68.02     8.50   |    14    26.08    66.38    11.78
            2    33.83    67.68    10.24   |    15    25.64    67.88     9.30
            3    32.44    69.58    11.22   |    16    24.99    67.67     9.80
            4    32.57    67.83    10.37   |    17    24.22    67.83    11.09
            5    33.48    67.95     8.69   |    18    24.90    66.97    10.85
            6    34.70    69.09    10.24   |    19    23.41    65.78    11.12
            7    30.13    61.76    10.50   |    20    24.76    65.57     9.11
            8    34.09    61.75    11.57   |    21    23.86    64.33     8.72
            9    33.07    59.99    11.47   |    22    23.51    64.32    10.95
           10    25.54    66.55    12.69   |    23    23.30    64.97    10.03
           11    24.70    67.72    10.92   |    24    23.65    65.83     7.99
           12    26.37    68.58    11.33   |    25    23.56    63.90     8.97
           13    24.21    69.96    12.16   |

              -+---------+---------+---------+---------+---------+---------+--
              I       *                                                       I
              I                                         *         *           I
              I               *                                               I
         68.0 I       * **  *                           *   *2                I
              I                                                               I
              I          * * *                                                I
              I    **                                                         I
              I         *                                                     I
              I   * *                                                         I
         64.0 I    2                                                          I
c             I                                                               I
v             I                                                               I
2             I                                  *              *             I
              I                                                               I
              I                                                               I
         60.0 I                                     *                        I
              I                                                               I
              I                                                               I
              I                                                               I
              I                                                               I
              I                                                               I
              I                                                               I
         56.0 I                                                               I
              -+---------+---------+---------+---------+---------+---------+--
             22.5       25.0       27.5       30.0       32.5       35.0       37.5

                                          cv1
```

(a) Qual é a primeira variável discriminante (canónica)? Qual a sua capacidade discriminante?

(b) Identifique as quatro classes de individuos na nuvem de pontos sobre as duas primeiras variáveis canónicas. Comente os resultados.

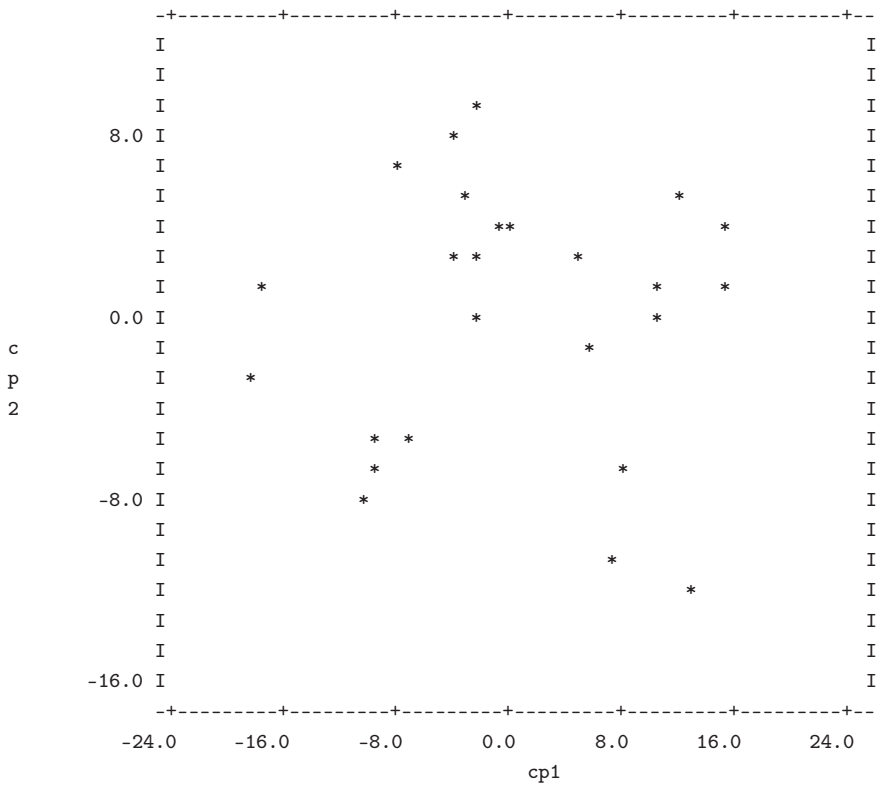
(c) A qual das 4 classes associaria um novo conjunto de observações, respeitantes a um lobo de sexo e habitat desconhecidos, com os seguintes valores para as 9 variáveis: 125, 104, 145, 81.1, 33.2, 68.2, 49.0, 43.3, 18.2? Justifique.

(d) Uma Análise de Componentes Principais sobre o conjunto das observações dos 25 individuos produz a seguinte aproximação óptima a 2 dimensões da nuvem de pontos dos 25 individuos no espaço 9-dimensional:

```
***  Principal Component Scores  ***

[NOTA: Este programa escreve estas Componentes de forma a que a soma de quadrados
de cada vector iguale o valor proprio associado a essa Componente.]
```

	1	2	3	4	5
1	-7.313	-5.264	-4.477	-0.936	-1.110
2	-18.793	-2.660	-1.629	3.222	-0.456
3	-17.691	1.128	-1.008	-1.956	1.650
4	-9.353	-6.250	-0.203	-1.637	-1.230
5	-9.654	-5.050	-0.673	-0.574	-0.066
6	-10.433	-8.234	2.395	1.297	1.897
7	8.001	-6.041	0.359	1.552	-1.975
8	6.842	-10.269	-1.380	1.066	0.207
9	12.567	-11.633	-2.065	-0.734	1.848
10	5.864	-1.639	6.059	-5.593	-1.396
11	-4.060	7.856	0.172	1.999	2.223
12	-2.348	-0.320	4.349	2.201	0.329
13	-2.489	3.328	0.764	-1.108	1.233
14	-8.319	6.328	-0.642	-0.545	0.390
15	-4.229	2.217	-1.325	-0.732	-2.765
16	-3.425	5.198	-0.595	1.481	1.205
17	-0.698	3.684	5.897	-2.267	3.391
18	0.177	3.544	2.662	0.986	-3.719
19	15.024	0.774	-1.478	-0.220	2.224
20	10.774	1.926	-3.511	-2.598	0.652
21	15.105	3.844	-4.195	-0.885	0.539
22	4.875	3.111	-1.415	1.941	-0.672
23	10.355	-0.440	4.642	1.792	-2.835
24	-2.498	9.939	-3.502	-2.153	-2.354
25	11.720	4.923	0.798	4.402	0.789



Compare com os resultados obtidos na alínea (5b) e comente. Como justifica as diferenças, sobretudo tendo em conta a palavra "óptima" na frase do enunciado que antecede o gráfico?

Capítulo 4

Análises Classificatórias (*Clustering*)

4.1 Introdução

Um problema que se coloca com alguma frequência é o de, dado um conjunto de n indivíduos, agrupar os n indivíduos em classes, ou subgrupos, de tal forma que cada subgrupo seja internamente homogéneo (*isto é*, constituído por indivíduos "semelhantes") e que os vários subgrupos sejam heterogéneos entre sí (*isto é*, os indivíduos de subgrupos diferentes sejam "dissemelhantes"). Em métodos de Análise Discriminante parte-se do pressuposto que uma tal subdivisão é conhecida num conjunto de dados que está disponível, e o objectivo é o de procurar direcções no espaço que evidenciem a separação desses subgrupos ou determinar uma regra para futuras classificações. Mas frequentemente *não* existe uma classificação desse tipo disponível, e o problema consiste em identificar quais (e quantas) são as diferentes classes de indivíduos existentes nos dados disponíveis.

Métodos que permitam determinar tais classes ou subgrupos designam-se métodos de **Análise Classificatória** (*Cluster Analysis*, em inglês). Nos dois extremos de possíveis classificações encontramos a classificação de *todos* os indivíduos numa única classe, e a classificação de *cada* indivíduo como uma classe separada. Existem dois grandes grupos de métodos de Análise Classificatória:

- **Métodos Hierárquicos -** O agrupamento em classes procede por etapas, em geral determinando-se a partir de n subgrupos (de um único indivíduo cada) sucessivas fusões de subgrupos considerados mais "semelhantes". Cada fusão reduz, em uma unidade, o número de subgrupos.

- **Métodos Não-Hierárquicos -** Fixa-se à partida o número k de classes que se pretende constituir e (regra geral) faz-se uma classificação inicial dos n indivíduos em k classes, ou determinam-se k "sementes" em torno das quais construir as classes. Através de transferências de indivíduos de uma classe para outra, ou de associações dos indivíduos às sementes das classes, procura-se determinar uma "boa" classificação, no sentido de tornar as classes mais internamente homogéneas e externamente heterogéneas.

4.2 Métodos Hierárquicos

Dado um conjunto de n indivíduos, o ponto de partida para os métodos de classificação hierárquicos será, em geral, uma matriz $n \times n$ cujo elemento genérico (i, j) é uma *medida de semelhança (ou dissemelhança)* entre o indivíduo i e o indivíduo j. Os critérios de semelhança (ou dissemelhança) utilizados podem ser diversos, havendo alguns critérios específicos para dados de diversos tipos, como adiante se verá. Com grande frequência, existe (como em métodos anteriores) uma matriz $\mathbf{X}_{n\times p}$ de observações multivariadas associadas aos indivíduos e que estão na origem da referida matriz de semelhanças/dissemelhanças. No entanto, pode não ser conhecida uma tal matriz, sendo apenas necessário conhecer a matriz $n \times n$ de semelhanças/dissemelhanças para que seja possível proceder a uma Análise de Classificação Hierárquica.

Como foi referido, parte-se duma **matriz $n \times n$ de semelhanças ou de dissemelhanças** entre os n indivíduos. Podem considerar-se inicialmente os n indivíduos como constituindo n classes diferentes. Procedendo por etapas, vai-se fundindo um par de classes em cada etapa. A fusão a efectuar numa dada etapa é a fusão dos dois subgrupos (classes) considerados mais "semelhantes". Este processo pode ser levado até às últimas consequências, *i.e.*, até à fusão de todos os indivíduos numa única classe.

A forma usual de representar graficamente as sucessivas fusões de subgrupos num método de classificação hierárquico é através dum **dendrograma**, *i.e.*, duma representação em forma de árvore do tipo indicado na Figura 4.1.

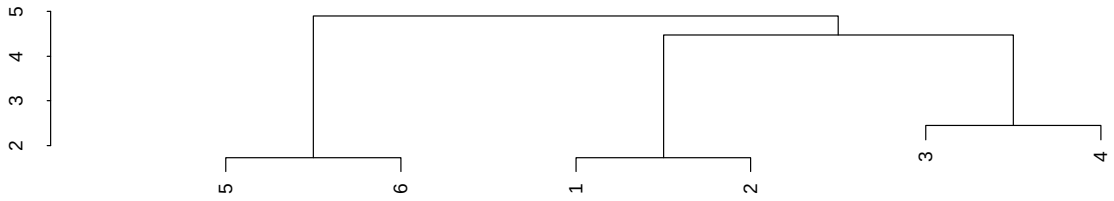


Figura 4.1: Dendrograma resultante duma Classificação Hierárquica

Um **corte** no dendrograma a qualquer nível de aglomeração produz uma *classificação em k subgrupos ($1 \leq k \leq n$).*

Note-se que um par de indivíduos que seja incluído numa mesma classe em qualquer etapa do processo não poderá mais ser separado em etapas posteriores, uma vez que estas consistem em *fusões* de classes já existentes.

Naturalmente, o processo apenas descrito constitui apenas *uma* das formas de efectuar uma Análise Classificatória Hierárquica, designada uma Análise Classificatória Hierárquica *Aglomeradora* ou *Ascendente*. É igualmente concebível proceder de forma análoga, mas em sentido inverso, *i.e.*, começar pela classe da

totalidade dos indivíduos e proceder à desagregação de classes anteriormente existentes pela separação de subgrupos considerados mais heterogéneos, efectuando assim uma Análise Classificatória Hierárquica *Desagregadora* ou *Descendente*. No entanto, a muito maior complexidade computacional deste procedimento torna menos habitual efectuar uma classificação hierárquica deste tipo.

O procedimento geral que acaba de ser descrito pode, no entanto dar origem a diferentes classificações, de acordo com duas questões-chave:

1. **o conceito de semelhança/dissemelhança entre 2 indivíduos**.
2. **o conceito de semelhança/dissemelhança entre 2 subgrupos**, também designado por **método aglomerador** ou **de fusão**.

Veremos seguidamente alguns destes conceitos de (dis)semelhanças entre indivíduos e entre classes.

4.3 (Dis)semelhanças entre indivíduos

Vamos admitir que o ponto de partida da classificação é uma matriz de dissemelhanças entre os n indivíduos. No caso de se tratar duma matriz de semelhanças, haverá ajustamentos a fazer, como se verá mais adiante. **A natureza da medida de dissemelhança irá condicionar a classificação que se seguirá e é, portanto, crucial que reflicta a natureza do problema sob estudo**.

4.3.1 Dissemelhanças e distâncias

De forma geral, **as dissemelhanças d_{ij} entre os indivíduos i e j** são medidas que reflectem as maiores ou menores *diferenças entre os valores que esses indivíduos registaram num conjunto de p variáveis*. No entanto, não é obrigatório que existam observações subjacentes de p variáveis, sendo até possível que as medidas de dissemelhança sejam atribuídas de forma subjectiva pelo investigador.

Uma medida de dissemelhança d_{ij} entre um individuo i e um individuo j deverá satisfazer algumas propriedades.

Quase sempre se exige a positividade, ou seja,

- $d_{ij} \geq 0, \quad \forall\, i, j = 1 : n;$
- $d_{ii} = 0, \quad \forall\, i = 1 : n.$

Na maioria das aplicações é também natural exigir a simetria, ou seja,

- $d_{ij} = d_{ji}, \quad \forall\, i, j = 1 : n.$

Fala-se em **distância**, em vez de apenas dissemelhança, no caso de as medidas de dissemelhança verificarem as condições indicadas na Definição 1.11 (p.19), ou seja (e utilizando a notação deste Capítulo), ou seja, no caso de além das anteriores conduições verificarem também a desigualdade triangular:

- $d_{ij} \leq d_{ik} + d_{jk}, \qquad \forall i, j, k$

4.3.2 Medidas de dissemelhança para dados quantitativos multivariados

Vejamos então algumas **medidas de distância d_{ij} entre indivíduos i e j**, construídas a partir de valores de p variáveis, traduzidos nos vectores $\mathbf{x}_{(i)}$ e $\mathbf{x}_{(j)}$. Estes vectores correspondem a linhas de uma matriz $\mathbf{X}_{n\times p}$ de dados, mas seguindo a convenção habitual para vectores, serão considerados vectores-coluna.

Distância euclidiana usual em $\mathbb{R}^p$. Tem-se:

$$\begin{aligned} d_{ij} &= \|\mathbf{x}_{(i)} - \mathbf{x}_{(j)}\| \\ &= \sqrt{(\mathbf{x}_{(i)} - \mathbf{x}_{(j)})^t(\mathbf{x}_{(i)} - \mathbf{x}_{(j)})} \\ &= \sqrt{\textstyle\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \\ \text{onde} \quad & \mathbf{x}_{(i)} \text{ é o vector do } i\text{-ésimo indivíduo} \\ & \mathbf{x}_{(j)} \text{ é o vector do } j\text{-ésimo indivíduo.} \end{aligned}$$

Trata-se duma distância no sentido da Definição 1.11 (p.19).

Distância Euclidiana Generalizada. Tem-se:

$$\begin{aligned} d_{ij} &= \|\mathbf{x}_{(i)} - \mathbf{x}_{(j)}\|_{\mathbf{W}} \\ &= \sqrt{(\mathbf{x}_{(i)} - \mathbf{x}_{(j)})^t\mathbf{W}(\mathbf{x}_{(i)} - \mathbf{x}_{(j)})} \\ &= \sqrt{\textstyle\sum_{k=1}^{p}\sum_{l=1}^{p} w_{kl}(x_{ik} - x_{jk})(x_{il} - x_{jl})} \\ \text{onde} \quad & \mathbf{W} \text{ é uma matriz definida positiva} \\ & \mathbf{x}_{(i)} \text{ é o vector do } i\text{-ésimo indivíduo} \\ & \mathbf{x}_{(j)} \text{ é o vector do } j\text{-ésimo indivíduo.} \end{aligned}$$

Este é o conceito de distância (definido na página 19) resultante dum produto interno definido no Teorema 1.3 (p.18). Alguns *casos especiais* resultam de escolher:

- $\mathbf{W} = \mathbf{D}^{-2}$, onde $\mathbf{D}^{-2}$ é a matriz diagonal dos recíprocos das variâncias, e que corresponde a tomar a distância euclidiana usual entre os dados *normalizados*.
- $\mathbf{W} = \Sigma^{-1}$, onde Σ é a matriz de variâncias-covariâncias das variáveis. Esta escolha gera a chamada **distância de Mahalanobis**, que é invariante a mudanças de escala nas variáveis[1].

[1] Veja-se, por exemplo, Mardia, Kent & Bibby (1979), *Multivariate Analysis*, Academic Press, para uma discussão mais pormenorizada desta distância.

Distâncias de Minkowski

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

com $\lambda = 1$ designa-se **métrica de Manhattan**, **distância ℓ_1** ou *city-block metric*.

com $\lambda = 2$ tem-se a **métrica euclidiana habitual** ou **distância ℓ_2** vista no ponto 4.3.2.

no limite $\lambda \to \infty$ obtem-se $d_{ij} = \max_k |x_{ik} - x_{jk}|$ que se designa a **métrica do máximo** ou **distância ℓ_∞**.

Em geral, quanto maior fôr o valor de λ, maior será o peso relativo de indivíduos muito dissemelhantes dos restantes.

Métrica de Canberra Para variáveis que apenas possam tomar valores não-negativos, pode definir-se a métrica:

$$d_{ij} = \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{(x_{ik} + x_{jk})}$$

em que o objectivo é obter uma medida de dissemelhança invariante a transformações de escala diferenciadas em cada variável. Este objectivo resulta de relativizar cada diferença através do denominador (cujas unidades de medida são iguais às do numerador). (NOTA: Este mesmo critério de dissemelhança, se usado com dados que possam tomar valores negativos, deixa de ter as propriedades de uma distância).

A escolha de qual o critério de distância entre indivíduos a utilizar não tem de se limitar às opções acima indicadas. Algumas destas opções são desaconselhadas para certos tipos de dados (por exemplo, as métricas de Minkowski podem não ser aconselháveis para o caso de as p variáveis terem diferentes unidades de medida).

Para certos tipos especiais de dados, existem outras medidas específicas de semelhança/dissemelhança. Refira-se, em particular, o caso de dados onde os indivíduos são apenas registados como estando ou não presentes numa série de locais (ou possuíndo ou não uma série de características dicotómicas).

4.3.3 Medidas de semelhança para dados binários

Ao contrário do que acontecia com técnicas anteriores, é possível efectuar Análises Classificatórias para observações de variáveis binárias e/ou qualitativas. Apenas será necessário poder construir uma matriz de dissemelhanças/semelhanças entre os indivíduos, considerada adequada pelo utilizador.

No caso de dados binários, isto é de observações sobre variáveis que apenas podem tomar dois diferentes valores (1/0, Presente/Ausente, Sim/Não, Macho/Fêmea, etc.) existem várias propostas de conceitos de *semelhança* na literatura[2].

[2]Para uma discussão de várias destas medidas veja-se, por exemplo, W.J.Krzanowski, *Principles of Multivariate Analysis*, Oxford Science Publications, 1988; P.Digby & R.Kempton, *Multivariate Analysis of Ecological Communities*, Chapman and Hall, 1987; ou ainda B.S.Everitt & S. Rabe-Hesketh, *The Analysis of Proximity Data*, Arnold, 1997.

Tal tipo de dados surge com frequência, em contextos biológicos. Por exemplo, ao **registar a "Presença" ou "Ausência" de várias espécies em p diferentes locais**, surge o problema de querer medir a (dis)semelhança entre um par de espécies, no que respeita à sua distribuição nos locais referidos. Trata-se dum caso particular do problema em que para dois indivíduos i e j se observam p variáveis binárias, e se pretende medir o grau de (dis)semelhança dos indivíduos. **As medidas de semelhança entre esses dois indivíduos baseiam-se, em geral, nas seguintes quatro quantidades**:

a – o número de variáveis para os quais ambos os indivíduos tomam o valor "Presente";

b – o número de variáveis para os quais o indivíduo i toma o valor "Presente" e o indivíduo j toma o valor "Ausente";

c – o número de variáveis para os quais o indivíduo i toma o valor "Ausente" e o indivíduo j toma o valor "Presente"; e

d – o número de variáveis em que ambos os indivíduos tomam o valor "Ausente"[3].

De entre os coeficientes de semelhança s_{ij} mais frequentes, destaquem-se:

Coeficiente de Concordância (*Matching Coefficient*) $s_{ij} = \frac{a+d}{a+b+c+d}$. Este coeficiente representa a proporção de variáveis em que há concordância nos valores dos indivíduos i e j. Toma valores entre 0 e 1.

Coeficiente de Jaccard $s_{ij} = \frac{a}{a+b+c}$. Este coeficiente representa o número de variáveis em que ambos os indivíduos têm valor "Presente", a dividir pelo número de variáveis em que pelo menos um dos indivíduos tem valor "Presente". Pode ser uma alternativa útil ao Coeficiente de Concordância caso os indivíduos registem valor "Ausente" na maioria das variáveis e não se considere que esse facto seja sinónimo de semelhança entre os indivíduos. Toma valores entre 0 e 1.

Coeficiente de Gower e Legendre $s_{ij} = \frac{(a+d)-(b+c)}{a+b+c+d}$. Este coeficiente toma a diferença entre concordâncias e discordâncias, relativamente ao número total de variáveis observadas. Ao contrário dos anteriores coeficientes, pode tomar valores negativos, situação que ocorre caso haja mais discordâncias do que concordâncias nos valores das variáveis para os indivíduos i e j. Toma valores entre -1 e 1.

4.3.4 Semelhanças e dissemelhanças entre individuos

Caso o ponto de partida seja um conjunto de medidas de *semelhança* entre cada par de indivíduos, pode proceder-se a trabalhar directamente sobre estas medidas de semelhança ou, alternativamente, converter as medidas de semelhança em medidas de dissemelhança. Caso sejam usadas medidas de semelhança são usuais as seguintes **formas de converter essas semelhanças em dissemelhanças**:

[3] Repare-se que, necessariamente, $p = a + b + c + d$.

1. $d_{ij} = 1 - s_{ij}$,
2. $d_{ij} = 1 - s_{ij}^2$,
3. $d_{ij} = \sqrt{1 - s_{ij}^2}$,
4. $d_{ij} = \sqrt{1 - s_{ij}}$

Note-se que, caso se utilize o Coeficiente de Concordância como medida de semelhança, a primeira destas regras de conversão equivale a tomar o coeficiente de dissemelhança $d_{ij} = \frac{b+c}{a+b+c+d}$, e a última regra de conversão produz a habitual distância euclidiana entre as linhas da matriz de zeros e uns associadas aos indivíduos em questão, caso se calculem distâncias relativas à distância máxima possível, que é $\sqrt{a+b+c+d}$.

4.4 Critérios de (des)agregação de classes

As medidas de dissemelhança entre indivíduos não representam a única opção a fazer numa Análise de Classificação. O investigador terá também de determinar o critério para medir a **dissemelhança entre classes** ou, como por vezes é dito, o **método de agregação de classes**. Dadas duas classes, G e H, como medir a semelhança/dissemelhança entre elas, de forma a proceder a fusões ou cisões? Alguns conceitos de dissemelhança habituais são:

Método do Vizinho Mais Próximo. (Em inglês, *nearest neighbour*, *single-link* ou *connected*). Consiste em considerar que a distância entre dois subgrupos é a *menor* distância entre um elemento dum subgrupo e um elemento do outro subgrupo:

$$D_{GH} = \min_{\substack{k \in G \\ l \in H}} d_{kl} \tag{4.1}$$

Método do Vizinho Mais Distante. (Em inglês, *furthest neighbour*, *complete-link* ou *compact*). Consiste em considerar que a distância entre dois subgrupos é a *maior* distância entre um elemento dum subgrupo e um elemento do outro subgrupo:

$$D_{GH} = \max_{\substack{k \in G \\ l \in H}} d_{kl} \tag{4.2}$$

Método das Distâncias Médias entre Grupos. (em inglês, *group average* ou *average link*). Consiste em considerar que a distância entre duas classes é a média de todas as distâncias entre pares de elementos (um em cada classe):

$$D_{GH} = \frac{1}{n_G \cdot n_H} \sum_{k=1}^{n_G} \sum_{l=1}^{n_H} d_{kl} \tag{4.3}$$

Consideremos ainda dois conceitos de dissemelhança entre classes válidos quando subjacente às dissemelhanças entre indivíduos existe, não apenas uma matriz nxn de dissemelhanças, mas uma matriz $\mathbf{X}_{n\times p}$ de dados multivariados.

Método da Inércia Mínima (Método de Ward). (Em inglês, *minimum variance method*). Considere-se a **inércia** duma classe G, *i.e.*, a soma de quadrados das diferenças entre cada indivíduo e o "indivíduo médio" dessa classe (dado pelo centro de gravidade da nuvem de n pontos em $\mathbb{R}^p$):

$$\mathbf{I}_G = \sum_{l=1}^{p} \left[\sum_{k\in G} (x_{kl} - \overline{x^G}_l)^2 \right] \tag{4.4}$$

onde $\overline{x^G}_l$ é a média dos valores da variável l para os indivíduos da classe G. Tome-se agora a distância entre as classes G e H como sendo o *aumento na soma total das inércias provocado pela fusão dos grupos G e H*. Por outras palavras, seja I_G a inércia da classe G, I_H a inércia da classe H e $I_{G\cup H}$ a inércia da classe resultante de fundir as classes G e H, então (uma vez que a fusão de G e H não afecta as inércias das restantes classes):

$$D_{GH} = I_{G\cup H} - (I_G + I_H) \tag{4.5}$$

É possível mostrar que se $\overline{\mathbf{x}_G}$ e $\overline{\mathbf{x}_H}$ são os vectores centro de gravidade das classes G e H, respectivamente, então:

$$D_{GH} = \frac{n_G n_H}{n_G + n_H} \|\overline{\mathbf{x}_G} - \overline{\mathbf{x}_H}\|^2 \tag{4.6}$$

onde $\|\cdot\|$ é a habitual norma euclidiana.

Exercício 4.1 *Demonstrar este resultado.*

Método dos Centróides. (Em inglês *Centroid method*). Neste caso toma-se a distância entre duas classes como sendo a *distância entre os centros de gravidade, ou outros pontos considerados 'representativos' (centróides), das classes*:

$$D_{GH} = \|\overline{\mathbf{x}_G} - \overline{\mathbf{x}_H}\| \tag{4.7}$$

Qualquer das definições de distâncias entre classes acima referidas pode ser vista como o caso particular duma única expressão para a distância entre a classe K e a fusão das classes G e H[4]. De facto, representando a distância entre duas classes K e G por D_{KG} e a distância entre qualquer classe K e a fusão das classes G e H por $D_{K\cdot GH}$, tem-se, na forma geral:

$$D_{K\cdot GH} = \alpha_G D_{KG} + \alpha_H D_{KH} + \beta D_{GH} + \gamma |D_{KG} - D_{KH}| \tag{4.8}$$

Nos casos particulares, temos:

[4]Este resultado deve-se a Lance e Williams, *A generalized sorting strategy for computerised classifications*, **Nature**, Vol. 212, pg. 218, 1966.

Vizinho Mais Próximo:

$$\alpha_G = \alpha_H = 1/2$$
$$\gamma = -1/2 \quad \beta = 0$$

Vizinho Mais Distante:

$$\alpha_G = \alpha_H = \gamma = 1/2$$
$$\beta = 0$$

Distâncias Médias entre Grupos:

$$\alpha_G = \frac{n_G}{n_G+n_H} \quad \alpha_H = \frac{n_H}{n_G+n_H}$$
$$\beta = \gamma = 0$$

Inércia Mínima:

$$\alpha_G = \frac{n_G+n_K}{n_G+n_H+n_K} \quad \alpha_H = \frac{n_H+n_K}{n_G+n_H+n_K}$$
$$\beta = \frac{-n_K}{n_G+n_H+n_K} \quad \gamma = 0$$

Centróides:

$$\alpha_G = \frac{n_G}{n_G+n_H} \quad \alpha_H = \frac{n_H}{n_G+n_H}$$
$$\beta = \frac{-n_G n_H}{(n_G+n_H)^2} \quad \gamma = 0$$

A escolha do método de agregação (*i.e.*, das distâncias D_{GH}) condicionará a classificação resultante. A opção por um outro método deve, pois, ser justificável com base na natureza dos dados e no objectivo da análise.

O facto de as classificações poderem diferir consoante os critérios de distâncias entre indivíduos e/ou entre classes, sugere que pode ser útil experimentar vários conceitos de distâncias, a fim de verificar se há robustez (consistência) na classificação geral. Quanto maior essa robustez, menos artificial será a classificação.

Observação: Não esquecer que qualquer método produzirá *sempre* uma classificação (em qualquer número de classes, consoante a altura a que optemos por cortar o dendrograma). Assim, a análise classificatória produz "classificações" mesmo onde elas possam não se justificar, pelo que será importante verificar a robustez dessas classificações.

Convém ter presente algumas características dos vários métodos de agregação:

- O método do Vizinho Mais Próximo tende a produzir classes "alongadas", com indivíduos que podem estar muito distantes entre si, mas pertencendo a uma mesma classe. Tal facto, conhecido pelo nome de **encadeamento**[5], resulta do facto de bastar que exista um elemento duma classe "próximo" de um único elemento doutra classe para que estas sejam atraídas, independentemente de haver outros elementos das classes que estejam muito distantes entre si. Do ponto de vista do dendrograma, o encadeamento tende a reflectir-se numa árvore com classes mal definidas, onde as fusões se sucedem a passo rápido.

[5] *chaining* em inglês.

- Os métodos do Vizinho Mais Próximo e do Vizinho Mais Distante são os únicos dos métodos acima referidos que são invariantes a transformações monótonas do conceito de dissemelhança (*i.e.*, produzem a mesma classificação antes e após uma transformação crescente ou decrescente das dissemelhanças).

- Os métodos do Vizinho Mais Distante, da Distância Média entre Grupos e dos Centróides têm tendência a produzir classes "esféricas" (caso os indivíduos sejam representáveis em $\mathbb{R}^p$), *i.e.*, classes onde não há grandes diferenças nas distâncias entre os pares de elementos mais distantes, ao longo de várias direcções.

- O método da Inércia Mínima (Ward) tem tendência a produzir classes com um número aproximadamente igual de indivíduos.

- O método do Vizinho Mais Próximo é o método mais "económico" do ponto de vista computacional.

- O método dos Centróides pode produzir as chamadas *inversões* no dendrograma, uma vez que não garante a monotonia nas dissemelhanças produzidas por sucessivas fusões.

Algumas palavras finais quanto à representação em árvore (dendrograma) dos resultados duma análise classificatória.

Em primeiro lugar refira-se que embora a classificação (para uma dada escolha de distâncias entre indivíduos e entre classes) seja única, **a representação em dendrograma *não* é única, uma vez que a ordem dos indivíduos (*i.e.*, das "folhas" da árvore) é arbitrária**. Por vezes, re-ordenações das folhas produzem árvores de aspecto diferente (embora a informação nelas contida seja idêntica). Em geral, os programas informáticos de Análise Classificatória tendem a escolher uma ordenação que evite que os "ramos" da árvore se cruzem.

Refira-se também que bons programas informáticos de Análise Classificatória produzirão uma representação gráfica em que é visível não apenas a hierarquia de agrupamento das várias classes, como também a *dissemelhança* entre as classes que se vão fundindo. Tal informação é reflectida na *altura* (admitindo que os indivíduos estão dispostos no eixo horizontal) das barras horizontais que unem as classes que se fundem.

Uma boa classificação deverá corresponder a um corte claro na árvore, *i.e.*, a um agrupamento que resulte de cortar o dendrograma numa zona onde as separações entre classes correspondam a grandes distâncias (o que se traduz em barras de junção de classes relativamente compridas). Tal facto reflectirá uma heterogeneidade entre classes, que como foi referido no início, é um dos objectivos da classificação. O outro objectivo, o da homogeneidade interna das classes, será tanto melhor conseguido quanto mais próximo das "folhas" (indivíduos) da árvore se fizer o corte.

4.5 Métodos Classificatórios Não-Hierárquicos

Numa **Classificação Não-Hierárquica**, é **frequente definir à partida o número k de classes que se pretende criar. O objectivo será determinar a classificação dos n indivíduos em k classes que optimize algum critério de homogeneidade interna e heterogeneidade externa**, como por exemplo um critério do tipo a **soma das inércias das k classes ser mínima**.

Determinar a solução deste problema, para um dado critério de classificação, exigiria uma pesquisa completa do valor do critério para *todas* as possíveis classificações de n indivíduos em k grupos. Tal pesquisa completa é inviável, excepto para pequenos conjuntos de indivíduos. (Note-se que, ainda por cima, as dimensões de cada classe não estão fixadas à partida, podendo tomar qualquer valor entre 1 e $n-(k-1)$). Torna-se assim necessário um algoritmo de pesquisa.

É possível começar por considerar uma *classificação inicial* dos n indivíduos em k classes. A escolha dessa classificação inicial pode corresponder a:

- Uma escolha subjectiva, resultante do estado do conhecimento dos dados no momento em que se inicia a Análise Classificatória.
- Uma escolha aleatória.
- Uma escolha determinística com algum critério simples (por exemplo, os primeiros n/k indivíduos constituem a primeira classe, os n/k seguintes a segunda classe, e por aí fora; escolher os indivíduos $i, \frac{n}{k}+i, 2\frac{n}{k}+i, \ldots, \frac{(k-1)n}{k}+i$ para constituir o i-ésimo grupo inicial; etc.).
- O resultado de cortar um dendrograma duma Classificação Hierárquica de forma a gerar k classes.
- Outros critérios.

Após esta escolha inicial, procede-se a uma pesquisa de classificações alternativas que gerem um valor do critério melhor. Para gerar classificações alternativas, pode admitir-se apenas o recurso a operações simples, como *transferências de um único indivíduo de uma para outra classe*; ou *permutas de dois indivíduos entre classes diferentes*. Desta forma, apenas se podem pesquisar classificações "próximas" duma classificação já alcançada. Uma classificação alternativa será aceite se produzir um valor do critério melhor que a classificação anterior.

Uma outra classe de métodos parte, não duma classificação inicial da totalidade dos n indivíduos, mas apenas de k "sementes" iniciais das classes (que poderão ser k indivíduos ou k "marcas" geradas por algum critério). Em seguida, vão-se agrupando os restantes indivíduos nas classes através de algum critério a optimizar, até se chegar a uma classificação da totalidade dos indivíduos em classes.

Exemplifiquemos através de um dos métodos de Classificação Não-Hierárquicos mais famosos, o **método das k-médias** (de Mc Queen, 1967):

1. Parta-se de k indivíduos iniciais (as *sementes*).

2. Associe-se os restantes $n - k$ indivíduos à classe cujo centro de gravidade lhe esteja mais próximo; (duas variantes do método resultam da decisão de re-calcular ou não o centro de gravidade duma classe após cada associação).

3. Após esta classificação resultante duma primeira associação de cada indivíduo, considerar o novo centro de gravidade de cada classe obtida como as novas "sementes" e fazer uma nova passagem pelos dados, procedendo a uma nova atribuição dos n indivíduos às classes. (Nesta passagem, é mais habitual deixar os centros de gravidade fixos).

Importa reter a ideia que **o desempenho de métodos deste tipo dependem de decisões** tais como:

- A escolha de critério a optimizar.
- A escolha da classificação inicial ou das sementes para o algoritmo.
- O critério de escolha de classificações alternativas (apenas se admitem transferências de um indivíduo? admitem-se ou não permutas?).
- A *ordem* pela qual se consideram os indivíduos aquando da decisão de os atribuir a uma ou outra classe (por exemplo, no método das k médias, a ordem pela qual se fará a associação de cada indivíduo às classes).

Destas decisões poderão resultar diferenças (maiores ou menores) nas classificações finais obtidas. Tal facto aconselha (como já foi referido aquando da discussão de Classificações Hierárquicas) que se procure alguma estabilidade global das classificações (para diferentes escolhas iniciais, diferentes ordens de consideração dos indivíduos, etc.) a fim de obter classificações mais "robustas".

4.6 A classificação de variáveis

Conceptualmente, nada impede que se considerem *as variáveis* observadas como os "indivíduos" a submeter ao agrupamento através duma Análise Classificatória. Nesse caso, o ponto de partida será uma matriz de dissemelhanças (ou semelhanças) entre variáveis.

Uma opção, consagrada em análises estatísticas multivariadas, para uma matriz de semelhanças entre variáveis é dada pela matriz de correlações entre variáveis, e será seguramente este o ponto de partida mais frequente (embora, naturalmente, não o único possível) para estudos de classificação de variáveis.

Menos clara será a questão de como se deve transformar uma matriz de correlações, que mede semelhanças entre variáveis, numa matriz de dissemelhanças. A natureza geométrica das correlações entre variáveis, discutidas na Secção 1.6 (p.45), que podem ser vistas como os cossenos dos ângulos entre os vectores representativos das variáveis centradas, sugere algumas opções para a conversão de medidas de semelhança em medidas de dissemelhança.

Uma possibilidade será a utilização da terceira transformação discutida na Subsecção 4.3.4 (p.122), ou seja, tomar $d_{ij} = \sqrt{1 - r_{ij}^2}$, onde r_{ij} indica a correlação entre as variáveis i e j. De facto, se r_{ij} é o cosseno do ângulo entre os vectores representativos das variáveis i e j, a dissemelhança d_{ij} agora definida será o módulo do *seno* desse mesmo ângulo[6]. Outra possibilidade consiste em utilizar *os ângulos* cujos cossenos são as correlações entre cada par de variáveis, ou seja, tomar $d_{ij} = \arccos(r_{ij})$.

Quaisquer que sejam as opções para critérios de semelhança e/ou dissemelhança entre variáveis, o procedimento de classificação que se segue será análogo aos procedimentos de classificação de indivíduos anteriormente considerados que utilizavam como ponto de partida matrizes de (dis)semelhanças.

4.7 A comparação de diferentes classificações

Nesta Secção aborda-se o problema de, dadas duas diferentes propostas de classificação de n indivíduos em k classes, quantificar o grau de semelhança entre essas classificações. Um índice deste tipo será de grande utilidade em pelo menos dois contextos:

- dada a grande diversidade de possíveis classificações de um mesmo grupo de indivíduos em k classes, torna-se útil poder avaliar em que medida os vários métodos de classificação tendem a produzir k subgrupos de constituição igual;

- no caso de se pretender validar uma classificação de um grupo de indivíduos que se sabe pertencerem a k diferentes subgrupos, será útil poder quantificar em que medida a classificação foi capaz de reproduzir essa estrutura.

O mais conhecido dos índices de comparação de duas classificações é o **Índice de Rand**. Existem duas expressões alternativas (mas equivalentes) para o índice de Rand associado a duas diferentes classificações dos mesmos n indivíduos em k classes.

A primeira expressão para o índice de Rand envolve a consideração de quantos, entre os $\binom{n}{2}$ pares de indíviduos diferentes, é que são classificados de forma análoga nas duas classificações, ou porque pertencem a uma mesma classe nas duas classificações, ou porque pertencem a classes diferentes nas duas classificações. Assim, tem-se:

$$R_k \quad = \quad \frac{A + B}{\binom{n}{2}} \ , \tag{4.9}$$

sendo A o número de pares de indivíduos que pertencem a uma mesma classe nas duas classificações, e B o número de pares de indivíduos que pertencem, nas duas classificações, a classes diferentes.

[6] Por outras palavras, será o seno do menor ângulo entre *as rectas* definidas por esses vectores, uma vez que ao elevar os coeficientes de correlação ao quadrado, se perde a informação sobre *o sentido* do vector, ficando apenas a informação sobre *a sua direcção*.

Exemplifiquemos com duas classificações de $n = 6$ indivíduos em $k = 3$ classes, indicadas pelos números 1, 2 e 3:

classif.1	1	1	2	2	3	3
classif.2	1	1	2	3	1	3

Neste caso, existe um único par de indivíduos que são sempre classificados numa mesma classe (os primeiro e segundo indivíduos, isto é, o par $(1, 2)$), pelo que $A = 1$. Por outro lado, dos $\binom{6}{2} = 15$ pares de seis indivíduos possíveis, existem $B = 9$ que, nas duas classificações, aparecem em classes diferentes (os pares $(1, 3)$, $(1, 4)$, $(1, 6)$, $(2, 3)$, $(2, 4)$, $(2, 6)$, $(3, 5)$, $(3, 6)$ e $(4, 5)$. Os restantes cinco pares – $(1, 5)$, $(2, 5)$, $(3, 4)$, $(4, 6)$ e $(5, 6)$ – surgem numa mesma classe numa das classificações, e em classes diferentes na outra classificação. Neste caso, o índice de Rand vale $R_3 = \frac{10}{15} = \frac{2}{3}$. Importa realçar que o valor do índice de Rand não depende dos valores numéricos usados para representar as classes, tomando o mesmo valor caso a primeira classificação tivesse sido representada pelo vector $(2, 2, 3, 3, 1, 1)$.

O índice de Rand não é mais do que o Coeficiente de Concordância (veja-se a página 122) para dados binários, aplicado aos dois vectores de zeros e uns que se obtém quando, em cada classificação, se associa a cada um dos $\binom{n}{2}$ pares possíveis de n individuos um 1, caso pertençam a uma mesma classe, ou um 0, caso pertençam a classes diferentes. É então fácil de verificar que, como qualquer Coeficiente de Concordância, **o índice de Rand toma valores no intervalo $[\mathbf{0}, \mathbf{1}]$. O valor máximo ($R_k = 1$) corresponderá a uma situação onde as duas classificações coincidam**, não havendo pares que esteja numa mesma classe num caso, e em classes diferentes no outro.

Uma expressão equivalente do índice de Rand é dada à custa do número m_{ij} de *individuos* que estão na classe i da primeira classificação, mas na classe j da segunda classificação. Designando por $m_{i\cdot} = \sum_{j=1}^{k} m_{ij}$ o número de indivíduos na classe i da primeira classificação (independentemente da classe a que pertençam na segunda classificação) e por por $m_{\cdot j} = \sum_{i=1}^{k} m_{ij}$ o número de indivíduos na classe j da segunda classificação, definam-se as quantidades:

$$T_k = \sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij}^2 - n \tag{4.10}$$

$$P_k = \sum_{i=1}^{k} m_{i\cdot}^2 - n \tag{4.11}$$

$$Q_k = \sum_{j=1}^{k} m_{\cdot j}^2 - n \ . \tag{4.12}$$

Tem-se então:

$$R_k = \frac{T_k - \frac{1}{2}(P_k + Q_k) + \binom{n}{2}}{\binom{n}{2}} \ .$$

Omite-se a demonstração da igualdade entre as duas expressões alternativas para o índice de Rand.

Outro índice proposto para efectuar a comparação de duas classificações é o **índice de Fowlkes e Mallows**, que também se baseia nas quantidades T_k, P_k e Q_k acima definidas[7]. Enquanto que no índice de Rand se privilegia uma comparação entre T_k e a média aritmética de P_k e Q_k, o índice de Fowlkes e Mallows compara T_k e a *média geométrica* de P_k e Q_k:

$$B_k \quad = \quad \frac{T_k}{\sqrt{P_k Q_k}} \; . \tag{4.13}$$

4.8 Exemplos

Os principais comandos do programa R, no contexto da **Análise Classificatória Hierárquica**, são:

`dist` **e** `as.dist` (para criar matrizes de distâncias entre indivíduos a partir de matrizes de dados);

`hclust` (para efectuar a Classificação);

`plot` **ou** `plclust` (para representar graficamente a classificação);

`cutree` (para cortar o dendrograma e produzir uma listagem dos indivíduos pertencentes a cada classe); e

`identify` **e** `rect.hclust` para separar graficamente as classes num dendrograma produzido pelo comando `plclust`.

No que respeita às Classificações Não-Hierárquicas, o programa informático R tem uma função que efectua uma variante do método das k-médias, designado as ***k*-médias de Hartigan**, em que o critério de atribuição dos indivíduos às classes (geradas por uma "semente" de k centros iniciais) é o de minimizar a soma das inércias das k classes, em relação aos seus centros. A função relevante do R é a função `kmeans`.

4.8.1 Classificando os lírios

Consideremos, mais uma vez, o conjunto de dados `iris`, já integrado no programa R, e procuremos classificar os $n = 150$ lírios, fingindo desconhecer os três grupos a que, na realidade, os lírios pertencem. Vejamos se, com base nas quatro variáveis morfométricas observadas, uma Análise Classificatória é capaz de reproduzir a existência de três diferentes grupos, associados às variedades de lírios.

Como foi visto nas Secções anteriores, haverá várias decisões a tomar a fim de efectuar uma Análise Classificatória, sendo a primeira de todas o critério de dissemelhança que deverá ser associado a cada par de indivíduos. Por omissão, o comando `dist` pega numa matriz ou *data frame* de dados e calcula uma matriz de distâncias euclidianas usuais entre cada par de indivíduos (linhas da matriz/*data frame* original). Através da opção `methods` do comando `dist`, poderão solicitar-se outros conceitos de distância.

[7] Aliás, é para salientar a relação entre os dois índices que as quantidades foram definidas da forma acima referida: no índice de Rand seria desnecessário incluir as parcelas "$-n$" no final de cada uma das quantidades, uma vez que se cancelam.

Actualmente, estão disponíveis no R as distâncias de Minkowski - incluíndo a métrica do máximo, com a designação de `maximum` - ou de Canberra, e ainda - para dados binários e com a designação de `binary` - a distância que resulta de associar a medida de semelhança Coeficiente de Jaccard (ver página 122) à conversão em dissemelhança dada pela relação $d_{ij} = 1 - s_{ij}$.

O comando `dist` produz um objecto da classe `dist`, ou seja, um objecto de tipo vector, contendo os elementos do triângulo inferior da matriz $n \times n$ de distâncias (aproveitando o facto de matrizes de distâncias serem simétricas e terem diagonais nulas para poupar na quantidade de informação a ser armazenada). Caso se disponha já de uma matriz de dissemelhanças completa ($n \times n$), criada por outra via, esta pode ser transformada numa estrutura de classe `dist` através do comando `as.dist`.

O resultado poderá então ser passado para o comando `hclust`, que efectua a classificação. Por omissão, este comando utiliza o Método do Vizinho Mais Distante para efectuar as sucessivas fusões de classes, podendo outros métodos de agregação ser solicitados através da opção `method`. Actualmente, os métodos de agregação disponíveis no R incluem, além do método do Vizinho Mais Distante (que é utilizado por omissão, e é indicando através da opção `complete` do argumento `method`), também o Método do Vizinho Mais Próximo (`method=''single''`), do Vizinho Médio (`method = ``average''`), dos Centróides (`method = ``average''`) ou de Ward (`method = ``ward''`), entre outros.

Os resultados do comando `hclust` são de difícil leitura (constituem um objecto de classe `hclust`). Serão de maior utilidade quando transformados num dendrograma, o que pode ser feito através do comando `plclust`, ou mais simplesmente solicitando um `plot` do resultado do comando `hclust`.

Na Figura 4.2 apresenta-se o dendrograma resultante de uma Análise Classificatória Hierárquica através do Método do Vizinho Mais Próximo, sobre a matriz de distâncias euclidianas entre as $n = 150$ linhas dos dados `iris`. Essa Figura foi obtida através do comando:

```
> plot(hclust(dist(iris[,-5]),method=''single''))
```

Como acontece sempre que o número de indivíduos é grande, a leitura das "folhas" do dendrograma não é fácil, sendo no entanto evidente a existência de dois grandes grupos. Para compreender quais os indivíduos que integram cada um desses grupos, pode utilizar-se o comando `cutree`, com o resultado indicado a seguir:

```
> cutree(hclust(dist(iris[,-5]),method="single"),k=2)
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
 21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
 41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
  1   1   1   1   1   1   1   1   1   1   2   2   2   2   2   2   2   2   2   2
 61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
 81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100
```

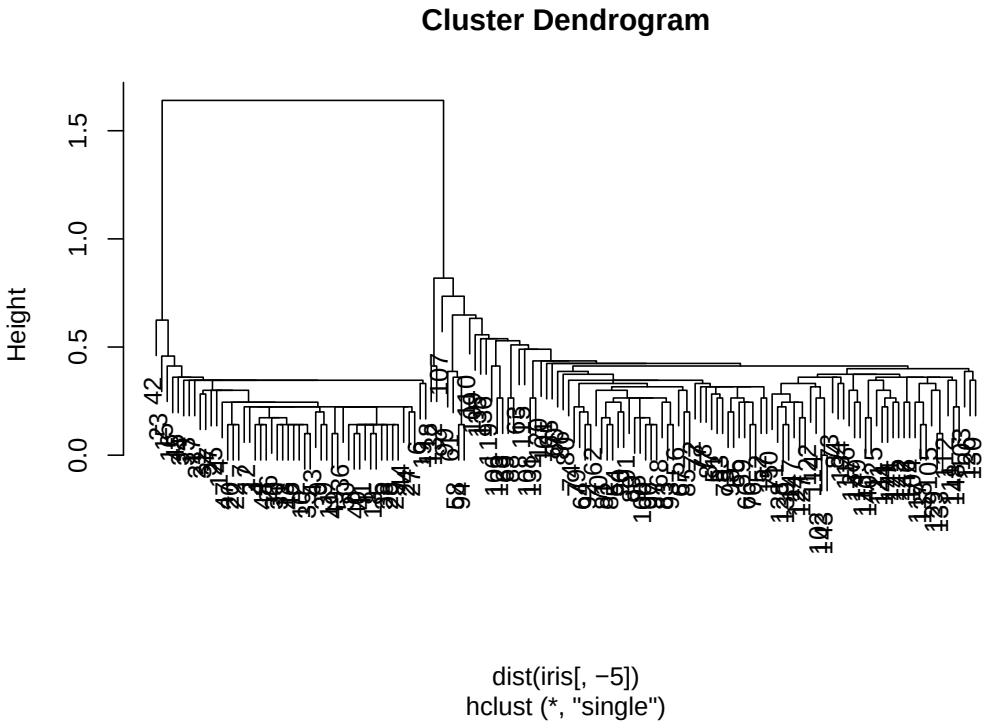


Figura 4.2: Dendrograma dos $n = 150$ lírios, classificados pelo Método do Vizinho Mais Próximo, com base numa matriz de distâncias euclidianas.

```
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
141 142 143 144 145 146 147 148 149 150
  2   2   2   2   2   2   2   2   2   2
```

Repare-se como o comando `cutree` exige a indicação (através do parâmetro `k`) do número de classes em que se deseja particionar os indivíduos. A composição das classes é o resultado de cortar o dendrograma a uma altura (que, no caso em estudo, pode ser a altura 1.0) na qual resultem dois grupos separados de "folhas". Da listagem produzida (e recordando que a *data frame* `iris` continha os lírios *setosa* nas 50 primeiras linhas) pode verificar-se que a separação deixa a totalidade dos lírios dessa variedade num grupo, juntando os 100 lírios das outras duas variedades no segundo grupo. Esta separação está de acordo com a visão dos dados quando projectados no primeiro Plano Principal (Figura 2.3, página 74). A leitura do dendrograma não justifica uma separação em mais do que 2 grupos, e a imposição de um terceiro grupo (que, recorde-se, é sempre possível, bastando cortar o dendrograma a uma altura adequada) não produz

a desejada classificação em separado das variedades *versicolor* e *virginica*. Para as variáveis observadas, critério de distância entre indivíduos e entre classes utilizadas, a separação dessas duas variedades não é possível.

Estes resultados podem, no entanto, modificar-se, caso se modifique alguma das opções acima referidas. Assim, por exemplo, uma classificação em três grupos, baseada ainda na matriz das distâncias euclidianas, mas utilizando o Método da Inércia Mínima (Método de Ward) para definir as semelhanças entre classes, produz uma classificação em três grupos muito próxima da classificação dos lírios pelas suas variedades:

```
> cutree(hclust(dist(iris[,-5]),method="ward"),k=3)
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
 21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
 41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
  1   1   1   1   1   1   1   1   1   1   2   2   2   2   2   2   2   2   2   2
 61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
 81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
  3   2   3   3   3   3   2   3   3   3   3   3   3   2   2   3   3   3   3   2
121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
  3   2   3   2   3   3   2   2   3   3   3   3   3   2   3   3   3   3   2   3
141 142 143 144 145 146 147 148 149 150
  3   3   2   3   3   3   2   3   3   2
```

As primeiras 100 observações (todas as *setosa* e *versicolor*) foram correctamente classificadas num mesmo grupo, e nenhuma flôr foi incorrectamente classificada no grupo das *setosa*. Já no que respeita aos 50 lírios *virginica*, foram globalmente classificados num terceiro grupo, mas em 14 casos foram incorrectamente associadas ao grupo das *versicolor*. No entanto, um olhar pelo dendrograma correspondente à classificação baseada neste método (Figura 4.3) revela que a existência de dois diferentes grupos é clara, mas já a subdivisão em três grupos é menos óbvia (embora seja claramente mais plausível que no caso anterior).

Considerações análogas emergem da utilização da técnica de Classificação Não-Hierárquica das k-médias. A exigência de $k = 2$ grupos finais com esse método produz os seguintes resultados (que, tratando-se de um método aleatório devido à ordem de escolha das tentativas de modificação dos subgrupos, poderão revelar algumas discrepâncias em futuras aplicações ao mesmo conjunto de dados):

```
> kmeans(iris[,-5],2)
K-means clustering with 2 clusters of sizes 97, 53
```

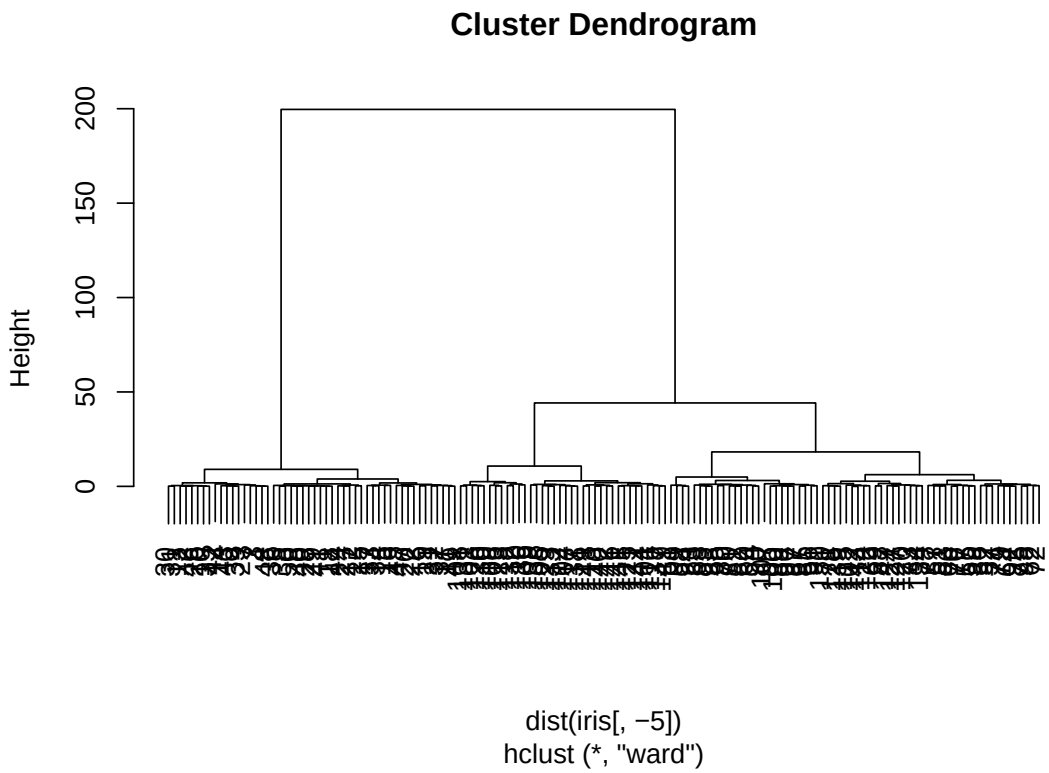


Figura 4.3: Dendrograma dos $n = 150$ lírios, classificados pelo Método da Inércia Mínima (Ward), com base numa matriz de distâncias euclidianas.

```
Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     6.301031    2.886598     4.958763   1.6958763
2     5.005660    3.369811     1.560377   0.2905660

Clustering vector:
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
 21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
 41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
  2   2   2   2   2   2   2   2   2   2   1   1   1   1   1   1   1   2   1   1
 61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
 81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100
  1   1   1   1   1   1   1   1   1   1   1   1   1   2   1   1   1   1   2   1
```

```
101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
141 142 143 144 145 146 147 148 149 150
  1   1   1   1   1   1   1   1   1   1

Within cluster sum of squares by cluster:
[1] 123.79588  28.55208

Available components:
[1] "cluster"  "centers"  "withinss" "size"
```

Além da classificação de cada indivíduo (`$cluster`), o comando devolve os centros de gravidade dos pontos em cada classe (`$centers`), a cardinalidade de cada classe (`$size`) e uma medida do grau de homogeneidade de cada classe (`$withinss`), dada pela inércia de cada classe (o critério de classificação visando minimizar a soma dessas inércias). No caso em questão é possível verificar que a classe 2 (que inclui as primeiras 50 observações) é mais homogénea que a segunda classe, mesmo levando em consideração que tem cerca de metade das observações, o que confirma a informação já obtida pelas análises anteriores.

4.8.2 A classificação das variáveis

Pode-se exemplificar o problema de classificar variáveis, já discutido na Secção 4.6 (página 128), definindo classes ou subgrupos das 4 variáveis observadas nos lírios: Comprimento e largura das Sépalas e das Pétalas. O ponto de partida para a nossa análise pode ser a matriz de correlações entre as variáveis, que é uma matriz de semelhanças entre variáveis, e que resulta ser:

```
> cor(iris[,-5])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

O pequeno número de variáveis observadas torna este exemplo quase desnecessário, sendo evidente que as variáveis associadas às medições das pétalas constituem um grupo relativamente coeso, ao qual se associa também o comprimento das sépalas, enquanto que a largura das sépalas constitui uma variável bastante diferente das restantes. Mas, se o exemplo é quase trivial, em contrapartida terá a vantagem de produzir dendrogramas legíveis!

A utilização do comando `hclust` do R exige a utilização de uma matriz de *dis*semelhanças, o que coloca o problema da conversão da matriz de correlações numa matriz de dissemelhanças. Utilizando como matriz de dissemelhanças a matriz dos *senos* dos ângulos em $\mathbb{R}^n$ representativos das variáveis (veja-se a discussão na Secção 4.6), o dendrograma resultante de utilizar os comandos `hclust` e `plclust` com o método de agregação do Vizinho Mais Distante (utilizado, como vimos antes, quando não se explicita outra opção) é dado na Figura 4.4.

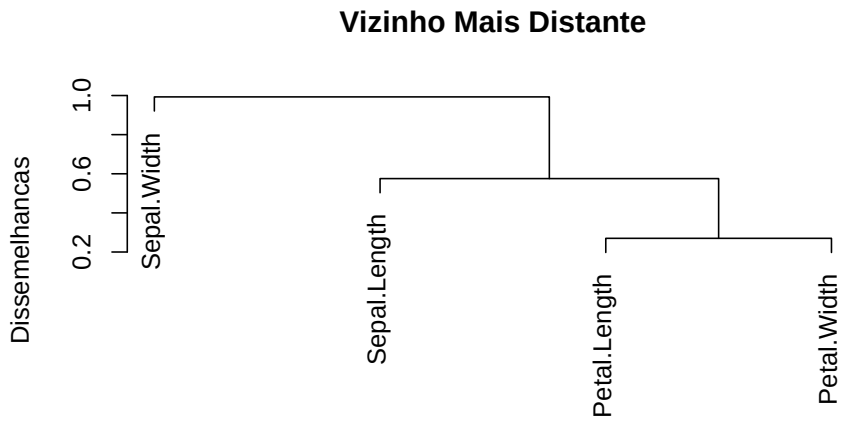


Figura 4.4: Dendrograma das $p = 4$ variáveis observadas no conjunto de dados dos lírios, classificados pelo Método do Vizinho Mais Distante, com medida de dissemelhanças entre variáveis dada por $d_{ij} = \sqrt{1 - r_{ij}^2}$.

Este dendrograma resultou de invocar o comando

```
> plot(hclust(as.dist(sqrt(1-cor(iris[,-5])^2))))
```

A opção `as.dist` é necessária para converter a matriz dos senos num objecto de tipo "`dist`", exigido pelo comando R `hclust` como sendo a natureza do objecto de entrada que contém as dissemelhanças entre (neste caso) as variáveis.

O dendrograma resultante de utilizar o critério de agregação do Vizinho Mais Próximo é praticamente idêntico.

Outra alternativa possível consiste em utilizar como medida de dissemelhança os arcos cujos cossenos são os coeficientes de correlação entre cada par de variáveis, o que representa os ângulos entre os vectores que, em $\mathbb{R}^n$ representam as variáveis centradas. Exprimindo os ângulos em graus, e através do comando

```
> plot(hclust(as.dist(180/pi*acos(cor(iris[,-5])))), main="",xlab="",
```

```
+                                                    sub="",ylab="Angulo (graus)")
```

obtém-se o dendrograma da Figura 4.5, que é semelhante ao anterior, mas separando mais claramente a variável largura das sépalas, das restantes.

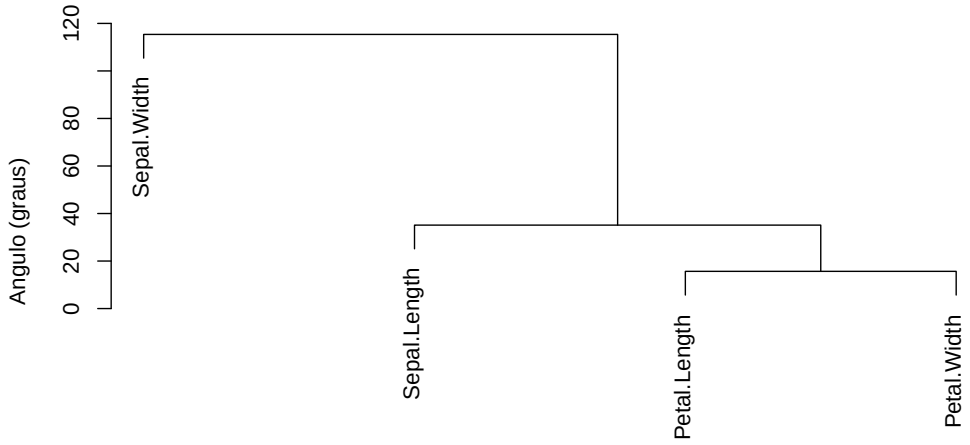


Figura 4.5: Dendrograma das $p = 4$ variáveis observadas no conjunto de dados dos lírios, classificados pelo Método do Vizinho Mais Distante, com medida de dissemelhanças entre variáveis dada por $d_{ij} = \arccos(r_{ij}) * \frac{180}{\pi}$.

4.8.3 Uma função na linguagem S para o índice de Rand

Uma das vantagens do programa R reside no facto de ser possível escrever código usando a linguagem de programação S para efectuar cálculos e operações, mesmo que estas não existam nos comandos já pre-disponibilizados no R[8]. Assim, é possível escrever uma pequena função para calcular o valor do índice de Rand associado a duas classificações que sejam disponibilizadas na forma de dois vectores numéricos de valores inteiros de 1 a k, indicando a classe a que cada um de n indivíduos pertence[9].

[8] Mas assinale-se a existência no *CRAN* de um grande, e sempre crescente, número de pacotes contribuídos gratuitamente por utilizadores do R, com código para efectuar muitas operações não existentes no programa base do R.

[9] Neste código simples não existem testes de validação do input para o caso de o utilizador fornecer valores sem sentido, como por exemplo, valores não inteiros. É boa prática de programação escrever código que contemple essas possibilidades. Mas neste caso optou-se por manter o código mais simples e legível.

Na nova função será necessário utilizar uma função pre-definida no R: a função `outer`. A função `outer` é um comando de grande utilidade que devolve um "produto externo" entre dois vectores (os dois primeiros argumentos), relativamente a uma função binária (o terceiro argumento). Por exemplo, o comando

```
> outer(1:10,1:10,"*")
```

cria uma tabuada (tabela de multiplicação) dos inteiros de 1 a 10. No nosso problema, admite-se que `vec` é um vector n-dimensional de números inteiros, indicando o subgrupo em que cada um de n individuos ficou classificado. Usar-se-á o comando `outer(vec,vec,"==")` para, a partir do referido vector, criar uma matriz de dimensão $n \times n$, cujo elemento (i, j) é dado pelo resultado (um valor lógico) de `vec[i] == vec[j]`. Assim, este comando cria uma matriz cujos elementos indicam se os elementos i e j do vector `vec` foram ou não classificados numa mesma classe. Ilustremos com base na classificação anteriormente discutida (subsecção 4.8.2) das quatro variáveis numéricas do conjunto `iris`:

```
> irisvar2 <- cutree(hclust(as.dist(sqrt(1-cor(iris[,-5]^2)))), k=2)
> irisvar2
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
           1            2            1            1

> outer(irisvar2,irisvar2,"==")
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length         TRUE       FALSE         TRUE        TRUE
Sepal.Width         FALSE        TRUE        FALSE       FALSE
Petal.Length         TRUE       FALSE         TRUE        TRUE
Petal.Width          TRUE       FALSE         TRUE        TRUE
```

Estamos agora em condições de escrever a função para calcular o índice de Rand, dadas duas classificações (`class1` e `class2`) de n indivíduos.

```
rand <- function(class1,class2){
         n <- length(class1)
         c <- as.dist(outer(class1,class1,"=="))
         d <- as.dist(outer(class2,class2,"=="))
       rand <- sum(c == d)/(n*(n-1)/2)
       rand
}
```

A função `rand` agora definida utiliza quatro comandos do R: `length`, `sum`, `outer` e `as.dist`. Os dois primeiros devolvem, respectivamente, o comprimento de um vector (isto é, o número dos seus elementos) e a soma dos elementos num vector ou matriz. O comando `outer` cria, como já se discutiu, uma matriz cujos elementos indicam se os elementos i e j do vector `class1` foram ou não classificados numa mesma

classe. Uma vez que nesta matriz estão incluídos os valores relativos à comparação de um elemento com ele próprio, e os pares de elementos diferentes são contemplados duas vezes, apenas interessa seleccionar o triângulo inferior da matriz, sendo isso que faz o comando `as.dist`. O resultado será um conjunto de $\binom{n}{2}$ valores lógicos, correspondentes aos $\binom{n}{2}$ pares de individuos que se podem formar a partir dos n indivíduos considerados, e em que o valor lógico `TRUE` corresponde a pares que pertencem a uma mesma classe, e os valores lógicos `FALSE` correspondem a pares que pertencem a classes diferentes. A instrução de código seguinte faz o mesmo para a segunda classificação a ser comparada. A penúltima instrução compara estes dois conjuntos de valores lógicos e calcula o coeficiente de Concordância entre eles, ou seja o índice de Rand.

O valor do índice de Rand obtido na comparação das classificações dos $n = 150$ lírios em $k = 3$ grupos obtidas com uma Classificação Hierárquica com base nos métodos do Vizinho Mais Próximo e de Ward (e da habitual distância Euclidiana) é 0.799.

```
> irisVMP <- hclust(dist(iris[,-5]),method="single")
> irisW <- hclust(dist(iris[,-5]),method="ward")
> rand(cutree(irisVMP,k=3),cutree(irisW,k=3))
[1] 0.7991946
```

Uma vez que neste caso é conhecida a classificação dos 150 lírios em 3 espécies, é possível comparar as classificações obtidas através das Análises Classificatórias com esta divisão por espécies. No caso da classificação resultante do Método do Vizinho Mais Próximo obtém-se um valor de 0.777 do índice de Rand, enquanto que o índice de Rand sobe para 0.892 na comparação entre a classificação por espécies e a classificação resultante de usar o Método de Ward. Os valores correspondentes para o índice de Fowlkes e Mallows são 0.790, 0.764 e 0.841.

4.9 Exercícios de Análises Classificatórias

1. Efectue uma Análise Classificatória Hierárquica dos dados do Exercício 7 do Capítulo sobre Análise em Componentes Principais (página 82), referentes aos afídios alados (*data frame* `adelges`). Utilize as distâncias Euclidianas entre os *indivíduos normalizados*. Considere três critérios de agregação de subgrupos:

   - Vizinho Mais Próximo;
   - Vizinho Mais Distante;
   - Vizinho Médio.

   (a) Comente as semelhanças e diferenças entre os três dendrogramas obtidos. Em particular,

      i. comente as diferenças do dendrograma resultante do método do vizinho mais próximo com os outros dois;

      ii. comente o que pode estar na origem dessas diferenças;

      iii. utilizando as funções `cutree` e `rect.hclust` do `R`, crie quatro classes em cada caso e compare-as.

   Sugira uma conclusão que lhe pareça adequada quanto à eventual existência de subgrupos entre os 40 afídios observados.

   (b) Compare os resultados desta Análise Classificatória com os resultados da Análise em Componentes Principais obtidos no Exercício 7 e comente.

   (c) Utilizando a função `rand` definida na Subsecção 4.8.3, calcule o índice de Rand comparando as classificações em quatro grupos obtidas pelos métodos do vizinho mais distante e do vizinho médio. Comente o valor obtido.

2. Efectue uma Análise Classificatória Hierárquica das 20 amostras de terra indicadas no Exercício 6 do Capítulo sobre ACP (página 81), a fim de analisar a eventual existência de diferentes classes de solos entre as observações. Utilize os seguintes critérios de dissemelhança:

   - distâncias de Canberra;
   - distâncias euclidianas sobre os individuos normalizados.

   Faça a análise utilizando os seguintes critérios de distâncias entre classes (*linkage method*):

   - vizinho mais distante.
   - distância média dos vizinhos.
   - distância de Ward.

   (a) Faça uma análise comparativa dos dendrogramas resultantes e comente.

   (b) Utilize as funções `identify` e `cutree` do `R` para criar, em cada caso, $k = 3$ classes.

   (c) Calcule o valor do índice de Rand para os pares das classificações em $k = 3$ grupos obtidos na alínea anterior. Comente.

3. Considere o conjunto de observações morfométricas sobre lavagantes discutido nos Capítulos anteriores.

   (a) Efectue análises classificatórias hierárquicas, com o método do Vizinho Mais Distante e:

   - a matriz das distâncias Euclidianas entre as linhas da matriz original de dados;
   - a matriz de distâncias Euclidianas entre as linhas da matriz *normalizada* dos dados;
   - a matriz das distâncias Euclidianas entre os *scores* das 63 observações nas duas primeiras Componentes Principais (sobre a matriz de dados não transformados).

   Comente os resultados, para cada uma destas análises, em termos da separação macho-fêmea que era visível com a projecção da nuvem de 63 indivíduos sobre o primeiro plano factorial, e ainda em termos da separação entre machos reprodutores e não reprodutores que sabemos existir entre os indivíduos observados.

   (b) Efectue uma classificação hierárquica usando as distâncias euclidianas sobre os dados normalizados e o critério do Vizinho Mais Próximo. Comente o dendrograma obtido.

   (c) Efectue agora uma Análise Classificatória Hierárquica visando agrupar *as variáveis* observadas. Utilize as correlações como medida de semelhança entre variáveis. Transforme estas semelhanças em dissemelhanças (veja as sugestões indicadas na Secção 4.6 (pg. 128)).

   (d) Efectue uma Classificação Não-Hierárquica em $k = 2$ grupos, utilizando a função `kmeans` do `R`. Em particular,

   i. Especifique apenas que deseja $k = 2$ classes;

   ii. Especifique o primeiro e o último individuos como sendo as sementes das 2 classes (veja a ajuda do comando `kmeans` do R e, em particular, o argumento `centers`).

   Comente os resultados.

4. Considere os dados relativos a medições de comprimentos e larguras de sepalas e petalas de 150 lírios, dadas no objecto `iris` do programa R, e já discutidos nos exemplos da Secção 4.8 (pg.131).

   (a) Efectue uma Análise Classificatória Hierárquica dos 150 lírios observados, com base na matriz de distâncias Euclidianas (tal como no primeiro exemplo da Secção referida), mas utilizando como método de agregação o Método de Ward.

   (b) Repita a alínea anterior, mas utilizando agora o Método de agregação do Vizinho Médio.

   (c) Efectue uma Análise Classificatória Hierárquica com os 150 lírios observados, utilizando as distâncias da métrica do máximo (métrica ℓ_∞), (também calculadas no R através do comando `dist`, com a opção `method=''maximum''`- veja a respectiva página de ajuda), e os Métodos de agregação do Vizinho Mais Distante, do Vizinho Mais Próximo, do Vizinho Médio e de Ward.

   (d) Repita a alínea anterior mas agora usando, como matriz de dissemelhanças, a matriz das distâncias de Manhattan (norma ℓ_1), que também é calculada pelo comando `dist`.

   (e) Repita a alínea anterior, mas utilizando agora a matriz de dissemelhanças de Canberra, definidas na Subsecção 4.3.2 (pg.120), e também calculadas pelo comando `dist` do R.

   (f) Compare as classificações obtidas em todos estes casos, utilizando o índice de Rand. Comente.

(g) Com base na *matriz de semelhanças entre classificações* obtida na alínea anterior, efectue uma Análise Classificatória (com critérios de distância à sua escolha) para classificar o resultado dos próprios métodos de Análise Classificatória usados! Parece-lhe possível agrupar os resultados em classes homogéneas? Comente.

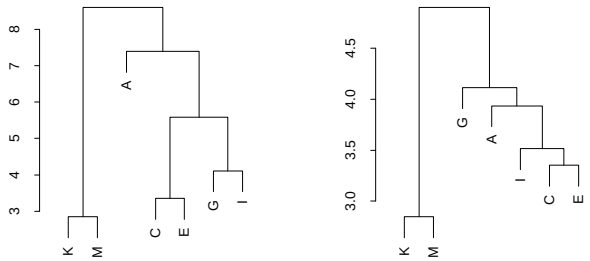
5. Ainda para os dados dos lírios, efectue uma Análise Classificatória Não-Hierárquica pelo método das k-médias, especificando:

   (a) a existência de três classes, deixando ao critério do programa R a escolha inicial de sementes para cada classe (tal como foi feito no exemplo da Subsecção 4.8.1, uma vez que é essa a opção utilizada caso não sejam explicitadas sementes).

   (b) a existência de três classes, mas agora escolhendo para sementes de cada classe os primeiros indivíduos de cada variedade, *i.e.*, os indivíduos 1, 51 e 101 (veja a ajuda do comando `kmeans` do R e, em particular, o argumento `centers`).

   Comente o efeito que a definição de sementes iniciais para cada classe parece ter tido na classificação dos lírios em $k = 3$ classes.

6. Efectuaram-se Análises Classificatórias sobre a matriz de distâncias do Exercício 8 do Capítulo sobre ACP (página 84), tendo sido considerados o método do Vizinho Mais Distante, e o Método do Vizinho Mais Próximo. Os dendrogramas resultantes foram, respectivamente: Como explicar



que no dendrograma da esquerda a localidade A seja a última a ser associada a alguma outra localidade, quando na matriz de distâncias subjacente à classificação lhe está associada a quarta menor distância (de entre as 21 distâncias entre diferentes pares das sete localidades)? Com base nos dois dendrogramas indicados e na ACP do Exercício 8, que classificação sugere para as sete localidades?

Capítulo 5

Representação Euclidiana de dissemelhanças (*MDS*)

5.1 Introdução

Nas Análises Classificatórias, o ponto de partida é uma matriz $n \times n$ de dissemelhanças d_{ij} entre n indivíduos. Vamos regressar a esse ponto de partida, mas agora com um objectivo diferente. Já não se trata de procurar classificar os indivíduos em subgrupos internamente homogéneos, mas sim de **procurar uma representação dos n indivíduos num espaço euclidiano, de tal forma que as distâncias euclidianas δ_{ij} entre cada par dos n pontos nessa representação sejam iguais, ou o mais próximas possível, às dissemelhanças d_{ij} da matriz inicial. Normalmente, procura-se também a melhor representação, a baixa dimensão (em geral $q = 2, 3$), nesse espaço euclidiano**.

Naturalmente que, se a matriz de dissemelhanças iniciais tiver sido calculada com base nas habituais distâncias euclidianas, o problema está resolvido se fôr conhecida a matriz de dados subjacente, e nesse caso, a representação óptima de baixa dimensão seria dada pela projecção da nuvem de pontos sobre o espaço definido pelas primeiras q Componentes Principais. A utilidade destes novos métodos reside na sua aplicabilidade: (i) no caso de não se conhecer a matriz de dados, e apenas se conhecer a matriz de distâncias euclidianas entre eles; (ii) no caso de dissemelhanças que não sejam representáveis por distâncias euclidianas (caso em que a Análise em Componentes Principais não é aplicável) e/ou (iii) no caso em que os indivíduos não são sequer representáveis por vectores de observações, mas sim por matrizes, funções, ou até apenas pelas dissemelhanças entre eles (caso, por exemplo, da matriz de distâncias rodoviárias entre localidades constante de muitos mapas).

No entanto, e como veremos em seguida, nem sempre será possível obter uma *representação euclidiana* dos indivíduos, estando essa possibilidade dependente da natureza e propriedades das medidas de dissemelhança utilizadas. Por vezes será possível obter representações aproximadas, mas casos haverá em que a utilidade destes métodos é questionável.

O conjunto de métodos cujo objectivo é o de procurar pontos num espaço euclidiano cujas coordenadas se situem a distâncias euclidianas δ_{ij} o mais fidedignas possíveis em relação às dissemelhanças d_{ij} dadas, é conhecido na literatura anglo-saxónica pela designação de *Multidimensional Scaling (MDS)*.

5.2 Matrizes Euclidianas

Comecemos por caracterizar as situações em que uma matriz de dissemelhanças é representável por distâncias euclidianas usuais num espaço a p dimensões.

Definição 5.1 *Seja* $\mathbf{D}$ *uma matriz* $n \times n$ *de dissemelhanças entre* n *individuos. A matriz* $\mathbf{D}$*, de elemento genérico* d_{ij}*, diz-se uma* ***matriz euclidiana*** *se existirem* n *pontos* $\{\mathbf{x}_{(i)}\}_{i=1}^{n} \in \mathbb{R}^p$ *tais que*

$$d_{ij}^2 \;=\; d^2(\mathbf{x}_{(i)}, \mathbf{x}_{(j)}) \;=\; \|\mathbf{x}_{(i)} - \mathbf{x}_{(j)}\|^2 \;=\; (\mathbf{x}_{(i)} - \mathbf{x}_{(j)})^t(\mathbf{x}_{(i)} - \mathbf{x}_{(j)}) \;. \tag{5.1}$$

Esta definição diz-nos que uma matriz é Euclidiana se fôr possível determinar n pontos em $\mathbb{R}^p$, tais que distância euclidiana usual entre qualquer par desses pontos seja precisamente a dissemelhança correspondente na matriz $\mathbf{D}$. **Sendo euclidiana, é possível determinar uma representação euclidiana exacta dos n pontos em p dimensões**. A notação $\mathbf{x}_{(i)}$ é sugerida pelo facto de que será conveniente considerar estes n pontos de $\mathbb{R}^p$ como as linhas[1] duma matriz $\mathbf{X}_{n\times p}$. Nesse caso, a i-ésima linha de $\mathbf{X}$ é o vector-linha $\mathbf{x}_{(i)}^t = \mathbf{e}_i^t\mathbf{X}$, onde $\mathbf{e}_i$ indica o i-ésimo vector da base canónica de $\mathbb{R}^n$. Analogamente, $\mathbf{x}_{(j)}^t = \mathbf{e}_j^t\mathbf{X}$, onde $\mathbf{e}_j$ indica o j-ésimo vector da base canónica de $\mathbb{R}^n$. Assim, a diferença entre os pontos que representam os individuos i e j é dada por:

$$(\mathbf{x}_{(i)} - \mathbf{x}_{(j)})^t \;=\; (\mathbf{e}_i - \mathbf{e}_j)^t\mathbf{X} \;. \tag{5.2}$$

e a dissemelhança ao quadrado será

$$d_{ij}^2 \;=\; (\mathbf{x}_{(i)} - \mathbf{x}_{(j)})^t(\mathbf{x}_{(i)} - \mathbf{x}_{(j)}) \;=\; (\mathbf{e}_i - \mathbf{e}_j)^t\mathbf{X}\mathbf{X}^t(\mathbf{e}_i - \mathbf{e}_j) \;. \tag{5.3}$$

Um facto digno de registo é que, caso uma matriz seja euclidiana, **os n vectores usados para representar as dissemelhanças em $\mathbb{R}^p$ não são únicos**, ou seja, a matriz $\mathbf{X}$ acima referida não é única. Tal facto é geometricamente intuitivo, uma vez que se uma dada configuração de n pontos em $\mathbb{R}^p$ satisfaz os requisitos de distâncias d_{ij}, **qualquer translacção, rotação ou reflexão** desses mesmos pontos também satisfaz as mesmas condições.

Em particular, será sempre possível considerar uma solução (linhas da matriz $\mathbf{X}$) cujo **centro de gravidade seja a origem do referencial em $\mathbb{R}^p$**. Isso corresponde a dizer que é sempre possível admitir que **a matriz tem as suas colunas centradas**, sendo por isso da forma $\mathbf{X} = (\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{X}$. Outra forma de argumentar a mesma ideia corresponde a notar que os vectores $\mathbf{e}_i - \mathbf{e}_j$ introduzidos na equação (5.2) pertencem ao complemento ortogonal do subespaço de $\mathbb{R}^n$ gerado pelo vector de n uns, uma vez

[1] Como sempre, consideramos que um vector é um vector-coluna, pelo que $\mathbf{x}_{(i)}$ indica um vector $p \times 1$ que, no entanto, será disposto ao longo duma linha da matriz $\mathbf{X}$.

que a soma das suas coordenadas é nula. Ou seja, $\mathbf{e}_i - \mathbf{e}_j \in \mathcal{C}(\mathbf{1}_n)^\perp$, $\forall i,j = 1:n$. Logo, esses vectores permanecem invariantes quando projectados sobre esse subespaço,

$$(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})(\mathbf{e}_i - \mathbf{e}_j) \;=\; \mathbf{e}_i - \mathbf{e}_j \;.$$

Deste modo, é possível re-escrever a equação (5.3) da seguinte forma:

$$d_{ij}^2 \;=\; (\mathbf{e}_i - \mathbf{e}_j)^t (\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n}) \mathbf{X}\mathbf{X}^t (\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})(\mathbf{e}_i - \mathbf{e}_j) \;, \tag{5.4}$$

sendo $(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{X}$ a matriz (de colunas centradas) cujas linhas contêm as coordenadas de cada um dos n pontos na configuração (em torno da origem) procurada.

Na próxima Secção abordam-se duas questões interligadas: (i) saber quando é que uma dada matriz $\mathbf{D}$ de dissemelhanças é euclidiana; e (ii) saber como determinar uma matriz $\mathbf{X}$ de coordenadas para os n pontos, com as características acima descritas.

5.3 A Análise em Coordenadas Principais

O mais tradicional método de *Scaling* é a **Análise em Coordenadas Principais** ou, na literatura em língua inglesa, *Principal Coordinate Analysis*, *Classical Scaling*, *Scaling* de Torgerson ou *Scaling* de Torgerson-Gower, . É esse o método que será agora desenvolvido em mais pormenor.

Para dar resposta às duas questões colocadas no final da Secção anterior (Secção 5.2), comecemos por representar a matriz dos produtos internos entre os vectores (centrados) $\mathbf{x}_{(i)}$:

$$\mathbf{Q} \;=\; (\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n}) \mathbf{X}\mathbf{X}^t (\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n}) \;. \tag{5.5}$$

Representando o elemento da linha i e coluna j de $\mathbf{Q}$ por q_{ij}, temos, a partir da equação (5.4) e da simetria de $\mathbf{Q}$:

$$\begin{aligned} d_{ij}^2 &= \mathbf{e}_i^t \mathbf{Q} \mathbf{e}_i - \mathbf{e}_i^t \mathbf{Q} \mathbf{e}_j - \mathbf{e}_j^t \mathbf{Q} \mathbf{e}_i + \mathbf{e}_j^t \mathbf{Q} \mathbf{e}_j \\ d_{ij}^2 &= q_{ii} - 2q_{ij} + q_{jj} \end{aligned} \tag{5.6}$$

Vejamos agora que *se apenas fôr conhecida a matriz de dissemelhanças* $\mathbf{D}$*, será possível recuperar a matriz* $\mathbf{Q}$ dos produtos internos. Uma vez que admitimos que as colunas de $\mathbf{X}$ estão centradas em torno da sua média, tem-se que terão de ser nulas:

- todas as somas das colunas de $\mathbf{Q}$;
- todas as somas das linhas de $\mathbf{Q}$;
- a soma de todos os elementos de $\mathbf{Q}$.

De facto,

$$\begin{array}{lcll} \mathbf{1}_n{}^t\mathbf{Q} = \mathbf{0} \in \mathbb{R}^n & \iff & \sum_{i=1}^n q_{ij} = 0 & ,\forall j \\ \mathbf{Q}\mathbf{1}_n = \mathbf{0} \in \mathbb{R}^n & \iff & \sum_{j=1}^n q_{ij} = 0 & ,\forall i \\ \mathbf{1}_n{}^t\mathbf{Q}\mathbf{1}_n = 0 & \iff & \sum_{i=1}^n \sum_{j=1}^n q_{ij} = 0 & \end{array}$$

Nesse caso, sempre a partir da equação (5.6), temos:

$$\left\{ \begin{array}{ll} i) & \sum_{i=1}^n d_{ij}^2 = \sum_{i=1}^n q_{ii} + nq_{jj} \\ ii) & \sum_{j=1}^n d_{ij}^2 = nq_{ii} + \sum_{j=1}^n q_{jj} \\ iii) & \sum_{i=1}^n \sum_{j=1}^n d_{ij}^2 = 2n \sum_{k=1}^n q_{kk} \end{array} \right.$$

O que equivale a:

$$\left\{ \begin{array}{rrll} iii') & \sum_{k=1}^n q_{kk} & = \frac{n}{2}(d^2)_{..} & \text{onde} \quad (d^2)_{..} = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n d_{ij}^2 \\ i') & q_{jj} & = (d^2)_{.j} - \frac{1}{2}(d^2)_{..} & \text{onde} \quad (d^2)_{.j} = \frac{1}{n}\sum_{i=1}^n d_{ij}^2 \\ ii') & q_{ii} & = (d^2)_{i.} - \frac{1}{2}(d^2)_{..} & \text{onde} \quad (d^2)_{i.} = \frac{1}{n}\sum_{j=1}^n d_{ij}^2 \end{array} \right.$$

Novamente a partir da equação (5.6), tem-se:

$$q_{ij} = -\frac{1}{2}(d_{ij}^2 - q_{ii} - q_{jj})$$

e substituíndo as expressões (i') e (ii'), vem:

$$q_{ij} = -\frac{1}{2}\left(d_{ij}^2 - (d^2)_{i.} - (d^2)_{.j} + (d^2)_{..}\right) \tag{5.7}$$

Assim, se apenas é conhecida a matriz de dissemelhanças D, é possível recuperar a matriz dos produtos internos entre indivíduos, Q. Com base nesta matriz $\mathbf{Q}$, é então possível criar uma matriz $n \times p$ cujas linhas correspondam à solução do problema. De facto, considere-se a Decomposição Espectral da matriz (simétrica) $\mathbf{Q}$:

$$\mathbf{Q} = (\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{X}\mathbf{X}^t(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n}) = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^t$$

Desde que a matriz Q seja semi-definida positiva, os elementos diagonais de $\mathbf{\Lambda}$ são não negativos e é possível definir a matriz $\mathbf{\Lambda}^{1/2}$. **A matriz $\mathbf{W}\mathbf{\Lambda}^{1/2}$ é uma matriz cujas n linhas podem representar os n indivíduos em $\mathbb{R}^p$, de forma a serem respeitados** os produtos internos, e logo **as distâncias euclidianas** d_{ij} entre eles.

Observações:

1. A existência desta solução depende de $\mathbf{Q}$ ser uma matriz semi-definida positiva. Na discussão acima, admitiu-se que $\mathbf{D}$ era euclidiana, e portanto $\mathbf{Q}$ terá de ser semi-definida positiva, porque a representação tem de ser possível. Repare-se que, dada a dupla centragem efectuada, $\mathbf{Q}$ não pode ser definida positiva, uma vez que tem pelo menos um valor próprio nulo. De facto, $\mathbf{Q}\mathbf{1}_n = (\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{X}\mathbf{X}^t(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{1}_n = (\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{X}\mathbf{X}^t(\mathbf{1}_n - \mathbf{1}_n) = \mathbf{0}$. Isto é, $\mathbf{Q}\mathbf{1}_n = 0 \cdot \mathbf{1}_n$.

2. Se na origem do problema esteve uma matriz de distâncias euclidianas entre n individuos observados em p variáveis, a solução agora produzida, ou seja, a matriz $\mathbf{W\Lambda}^{1/2}$ *não é a matriz de dados inicial*, mas sim a matriz dos *scores* de cada indivíduo nas p Componentes Principais definidas pelos dados.

3. Assinale-se que, para qualquer matriz de colunas ortogonais $\mathbf{R}$, ter-se-á: $(\mathbf{W\Lambda}^{1/2}\mathbf{R}^t)(\mathbf{W\Lambda}^{1/2}\mathbf{R}^t)^t = \mathbf{W\Lambda}^{1/2} \cdot (\mathbf{R}^t\mathbf{R}) \cdot \mathbf{\Lambda}^{1/2}\mathbf{W}^t = \mathbf{W\Lambda W}^t = \mathbf{Q}$. Esta indeterminação na reconstituição da matriz de dados original $\mathbf{X}$ a partir da matriz de distâncias entre indivíduos corresponde a dizer que **as distâncias fixam a configuração dos pontos, a menos de rotações e/ou reflexões em torno da origem**. $\mathbf{X}$.

4. A matriz $\mathbf{Q} = (\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{X}\mathbf{X}^t(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})$ tem de ter a mesma característica que a matriz $(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{X}$ (equação 1.20, pg.30), pelo que, admitindo que $n > p$ e que não há multicolinearidades entre as variáveis centradas (*i.e.*, que as p colunas de $(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{X}$ são linearmente independentes), teremos que $\mathbf{Q}$ não pode ter mais que $p-1$ valores próprios não-nulos (tendo em conta o valor próprio nulo acima referido). Pelo Teorema de Eckart-Yound, sabemos que a melhor representação q-dimensiona l $(q < p-1)$ dos n indivíduos será dada, pelas primeiras q colunas da matriz $\mathbf{W\Lambda}^{1/2}$, *i.e.*, pela matriz $\mathbf{W}_q\mathbf{\Lambda}_q^{1/2}$. A qualidade desta representação pode ser medida pelo quociente $(\sum_{i=1}^{q}\lambda_i)/(\sum_{j=1}^{p}\lambda_j)$, tal como em ACP.

5.3.1 Para uma matriz de dissemelhanças genérica

Vejamos agora em que consiste o método da Análise em Coordenadas Principais, num contexto em que a matriz de dissemelhanças **não** é euclidiana, ou em que **não se sabe** se é euclidiana. No que se segue, vamos apenas **admitir que a matriz das dissemelhanças, D, é simétrica e de elementos não negativos (com zeros na diagonal)**.

Dada uma matriz $\mathbf{D}$ de dissemelhanças (simétricas) d_{ij}, procede-se duma forma análoga à acima descrita para criar a matriz $\mathbf{Q}$. Concretamente, dão-se os seguintes passos:

1. **Cria-se, a partir da matriz de dissemelhanças D (de elemento genérico d_{ij}), uma nova matriz que se designará A, de elemento genérico:**

$$a_{ij} = -\frac{1}{2}d_{ij}^2$$

   Em termos matriciais, $\mathbf{A} = -\frac{1}{2}(\mathbf{D} \circ \mathbf{D})$, onde o símbolo "$\circ$" representa o **produto de Hadamard** entre duas matrizes (ver a página 4), definido como $\mathbf{B} \circ \mathbf{C} \equiv [b_{ij}c_{ij}]$, *i.e.*, como o produto de elementos correspondentes das duas matrizes (necessariamente de dimensões iguais).

2. **Constrói-se a matriz Q, obtida a partir da *dupla centragem* (centragem das linhas e das colunas) da matriz A, *i.e.*, a matriz Q terá elemento genérico:**

$$q_{ij} = a_{ij} - a_{i.} - a_{.j} + a_{..}$$

onde $a_{i.}, a_{.j}$ e $a_{..}$ são, respectivamente, as médias dos elementos da linha i, dos elementos da coluna j, e da totalidade dos elementos da matriz A. Em termos matriciais, tem-se:

$$\mathbf{Q} = (\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{A}(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n}) \tag{5.8}$$

3. A matriz $\mathbf{Q}$ será simétrica se a matriz $\mathbf{A}$ o fôr, e esta será simétrica se a matriz de dissemelhanças original, $\mathbf{D}$, fôr simétrica. **Sendo Q uma matriz simétrica, admite decomposição espectral:**

$$\mathbf{Q} = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^t$$

Como já se viu, *a dupla centragem* que faz parte da construção da matriz $\mathbf{Q}$ *gera um valor próprio nulo da matriz* $\mathbf{Q}$, associado ao vector próprio $\mathbf{1}_n$.

4. ***Caso a matriz Q seja semi-definida positiva*, procede-se como descrito atrás: determina-se a matriz**

$$\mathbf{Y} = \mathbf{W}\mathbf{\Lambda}^{1/2} \ ,$$

e as n linhas desta matriz correspondem às coordenadas que representam cada indivíduo num espaço euclidiano n-dimensional, $\mathbb{R}^n$.

5. **Projecta-se a nuvem de pontos assim obtida sobre $\mathbb{R}^q$ (em geral com $q = 2, 3$) retendo apenas as q primeiras colunas da matriz Y**, *i.e.*, através das linhas da matriz:

$$\mathbf{Y}_q = \mathbf{W}_q\mathbf{\Lambda}_q^{1/2}$$

onde $\mathbf{W}_q$ é a matriz $n \times q$ obtida retendo apenas as q primeiras colunas da matriz dos vectores próprios de $\mathbf{Q}$, e $\mathbf{\Lambda}_q^{1/2}$ é a matriz $q \times q$ obtida retendo apenas as raízes quadradas dos q primeiros valores próprios de $\mathbf{\Lambda}$, *i.e.*, retendo apenas as q primeiras linhas e colunas da matriz $\mathbf{\Lambda}^{1/2}$.

6. **Se Q não fôr semi-definida positiva**, alguns dos seus valores próprios serão negativos. Essa situação corresponde a dizer que **não existe representação exacta num espaço euclidiano real**, isto é, que não é possível garantir a representação dos indivíduos num espaço $\mathbb{R}^n$ de forma a respeitar as igualdades $d_{ij} = \delta_{ij}$ entre dissemelhanças iniciais e distâncias euclidianas no espaço $\mathbb{R}^n$. De facto, o Teorema 1.36 (pg.36) garante que, não sendo $\mathbf{Q}$ semi-definida positiva, nenhuma factorização da forma $\mathbf{Q} = \mathbf{Z}^t\mathbf{Z}$ é possível, não havendo, assim, marcadores (as eventuais linhas de $\mathbf{Z}$) que possam representar euclidianamente os indivíduos.

Portanto,

> **uma matriz de dissemelhanças D é euclidiana no espaço $\mathbb{R}^p$ se e só se a matriz $\mathbf{Q} = (\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{A}(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})$, com $\mathbf{A} = -\frac{1}{2}(\mathbf{D} \circ \mathbf{D})$, é semi-definida positiva de característica menor ou igual a p.**

5.3.2 Matrizes de dissemelhanças não-euclidianas

Como acaba de ser dito, nem todas as matrizes de dissemelhanças são euclidianas. **No caso de existirem valores próprios negativos da matriz Q**, duas alternativas podem ser consideradas, no âmbito da Análise em Coordenadas Principais:

1. Caso os valores próprios negativos de $\mathbf{Q}$ sejam de pequena magnitude (relativamente à soma dos valores próprios positivos), pode-se ignorá-los e trabalhar com uma configuração euclidiana aproximada, resultante de considerar apenas os vectores próprios associados aos valores próprios positivos de $\mathbf{Q}$. Em particular, pode-se **escolher o número máximo de eixos coordenados principais a reter de acordo com as seguintes regras**:

   Critério do traço : Reter eixos cuja soma de valores próprios associados seja aproximadamente igual ao traço da matriz $\mathbf{Q}$.

   Critério do valor absoluto : Reter eixos cujos valores próprios associados sejam maiores do que o módulo do menor valor próprio negativo.

   Na prática, porém, o critério mais frequente para a escolha de número de eixos coordenados principais a reter é o de escolher $q = 2$ ou 3, o que permite a tradução gráfica da representação euclidiana. Pode até ser possível interpretar os valores próprios de cada eixo como correspondendo à proporção de variabilidade total explicada pelo eixo (como faríamos no caso da representação euclidiana exacta ser possível). Dois critérios específicos, propostos na literatura **para medir a qualidade da representação obtida utilizando k eixos coordenados principais** (cujos valores próprios associados se admite serem todos positivos), são:

   (a) $P_1 = \dfrac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} |\lambda_i|}$;

   (b) $P_2 = \dfrac{\sum_{i=1}^{k} \lambda_i^2}{\sum_{i=1}^{n} \lambda_i^2}$.

2. Pode somar-se uma constante adequada, c^*, a todos os elementos não-diagonais da matriz de dissemelhanças, de tal forma a tornar a matriz $\mathbf{Q}$ uma matriz semi-definida postiva (*i.e.*, a tornar possível a representação euclidiana exacta). Tal opção, conhecida pelo nome de **problema da constante aditiva**, corresponde a inflacionar as dissemelhanças (entre indivíduos diferentes). A solução do problema[2], consiste em tomar c^* igual ao maior valor próprio da matriz:

$$\begin{bmatrix} \mathbf{0}_n & 2\mathbf{Q} \\ -\mathbf{I}_n & -4\mathbf{Q}(d_{ij}) \end{bmatrix}$$

   onde $\mathbf{Q}(d_{ij})$ corresponde à matriz obtida pela dupla centragem duma matriz análoga à matriz $\mathbf{A}$, mas com os elementos dados por $-\frac{1}{2}d_{ij}$, em vez de $-\frac{1}{2}d_{ij}^2$.

[2]Solução demonstrada por Cailliez, F. em *The analytical solution of the additive constant problem*, na revista *Psychometrika*, Vol. 48, No.2, pgs. 305-308 (1983).

5.4 Outras técnicas visando representações euclidianas

Dentro do princípio genérico de obter uma representação euclidiana a partir duma matriz de dissemelhanças **D** entre n indivíduos, pode-se formular o problema de forma diferente. Nesta formulação alternativa procuram-se pontos num espaço euclidiano de uma dada dimensão que, podendo não estar às distâncias euclidianas correspondentes às dissemelhanças d_{ij}, estejam a distâncias aproximadamente iguais a d_{ij} e de forma a **optimizar algum critério de qualidade de ajustamento**.

Estas técnicas de *Multidimensional Scaling (MDS)* podem ser sintetizada nos seguintes passos:

1. Fixa-se a dimensão q do espaço euclidiano onde se deseja a representação.
2. Parte-se duma configuração inicial de n indivíduos nesse espaço (*i.e.*, determina-se uma matriz $n \times q$ cujas linhas são as coordenadas de cada indivíduo). Uma escolha possível de configuração inicial pode ser a solução q-dimensional produzida pela Análise em Coordenadas Principais.
3. Calculam-se as distâncias euclidianas usuais entre os n pontos da configuração proposta.
4. Calcula-se o valor (para essa configuração) de algum critério de ajustamento que se deseja optimizar.
5. Efectuam-se alterações à configuração (de acordo com algum conjunto de regras pré-especificadas) e calcula-se o novo valor do critério.
6. Repete-se o passo anterior até alguma condição de paragem (usualmente ligada à ideia de que mexidas na configuração não estão a produzir melhorias no valor do critério).

Um critério possível de qualidade do ajustamento está associado ao nome de **Sammon** e é dado por:

$$Sam = \frac{1}{\sum_{i \neq j} d_{ij}} \sum_{i \neq j} \frac{(e_{ij} - d_{ij})^2}{d_{ij}} \tag{5.9}$$

onde d_{ij} representa a dissemelhança inicial entre os indivíduos i e j e e_{ij} representa a distância euclidiana habitual entre os representantes desses mesmos indivíduos na configuração que está sendo proposta. Como é evidente a partir desta definição, valores baixos do critério Sam estão associados a distâncias euclidianas globalmente próximas das dissemelhanças indicadas na matriz, pelo que o que se procura são configurações que minimizem Sam.

Seguramente o mais frequente critério de qualidade do ajustamento designa-se **STRESS** (que por vezes surge com algumas variantes). Foi proposto por Kruskal e Shepard, e é dado por:

$$STRESS = \sqrt{\frac{\sum_i \sum_{j<i} (e_{ij} - f(d_{ij}))^2}{\sum_i \sum_{j<i} e_{ij}^2}} \tag{5.10}$$

onde d_{ij} representa a dissemelhança inicial entre os indivíduos i e j, e_{ij} representa a distância euclidiana habitual entre os representantes desses mesmos indivíduos na configuração que está sendo proposta, e

f indica alguma função *crescente*[3] (em sentido lato, isto é, não decrescente). Escolhendo f como sendo a função identidade, vemos um critério cujo valor mínimo (zero) corresponderia a um conjunto de n pontos ideal, em que as distâncias euclidianas entre os pontos i e j são sempre iguais às dissemelhanças d_{ij} dadas na matriz $\mathbf{D}$. Permitir que f seja uma outra função crescente é uma opção indicada para casos em que os valores das dissemelhanças são algo subjectivos, sendo mais importantes **as ordens**[4] do que propriamente os valores dessas dissemelhanças. Mas pode também ser importante por permitir maior flexibilidade. Existe, aliás, um resultado teórico[5] que garante que, para matrizes de dissemelhanças simétricas, não negativas e com diagonal nula, é sempre possível encontrar uma solução com dissemelhanças assim alteradas, pelo menos num espaço de dimensão $n-2$. No caso de se admitirem as transformações crescentes associadas à função f, fala-se em **Scaling Não-Métrico** ou **Ordinal**. As transformações crescentes das dissemelhanças, $\hat{d}_{ij} = f(d_{ij})$ são por vezes chamadas as disparidades entre os individuos i e j. São obtidas através duma técnica de análise numérica designada *regressão isotónica*.

Algoritmos do tipo acima descrito exigem programas informáticos próprios, e são computacionalmente exigentes. Além disso, uma paragem do algoritmo pode não corresponder a uma solução globalmente óptima, mas apenas óptima numa vizinhança local das soluções que foram ensaiadas. Assim, é sugerido que se corra o algoritmo com diferentes configurações iniciais, a fim de testar a robustez da solução encontrada (e, no caso de surgirem diferentes soluções para diferentes configurações iniciais, deverá, naturalmente, optar-se pela que optimiza o valor do critério proposto).

5.5 Um exemplo

O exemplo paradigmático da Análise em Coordenadas Principais consiste em procurar construir o mapa de um país ou região, a partir duma matriz das distâncias rodoviárias ou em linha recta entre localidades desse país ou região. O Exercício 1 deste Capítulo (página 160) aborda essa questão no que respeita ao caso de Portugal. Nesta Secção será considerado o problema análogo, mas com base no objecto `eurodist`, já incluído no R, onde são dadas as distâncias rodoviárias (em km) entre 21 cidades europeias[6].

O comando do programa R para efectuar uma Análise em Coordenadas Principais é o comando `cmdscale`. Na sua forma mais simples, o comando exige um único argumento: a matriz de dissemelhanças, na forma dum objecto R de tipo `dist`, como é o `eurodist`. Por omissão, será procurada uma representação bi-dimensional, e serão devolvidas as coordenadas de cada ponto nos dois primeiros eixos coordenados principais. Esta dimensionalidade pode ser alterada através do argumento `k`. Os valores próprios associados podem ser solicitados através do argumento lógico `eig`. Caso se deseje utilizar a solução do problema da constante aditiva acima descrito, deve-se colocar o argumento lógico `add` como `TRUE`.

[3] A escolha duma função decrescente permitiria começar por uma medida de semelhança entre os individuos.

[4] Os *ranks*, em inglês.

[5] Lingoes, J.C. (1971), Some boundary conditions for a monotone analysis of symmetric matrices, *Psychometrika*, 36, 195-203.

[6] Informação retirada da *Cambridge Encyclopædia*.

No caso do nosso problema, é possível obter as coordenadas nos dois primeiros eixos, a partir da matriz de distâncias rodoviárias `eurodist`:

```
> cmdscale(eurodist)
                         [,1]        [,2]
Athens           2290.274680  1798.80293
Barcelona        -825.382790   546.81148
Brussels           59.183341  -367.08135
Calais            -82.845973  -429.91466
Cherbourg        -352.499435  -290.90843
Cologne           293.689633  -405.31194
Copenhagen        681.931545 -1108.64478
Geneva             -9.423364   240.40600
Gibralta        -2048.449113   642.45854
Hamburg           561.108970  -773.36929
Hook of Holland   164.921799  -549.36704
Lisbon          -1935.040811    49.12514
Lyons            -226.423236   187.08779
Madrid          -1423.353697   305.87513
Marseilles       -299.498710   388.80726
Milan             260.878046   416.67381
Munich            587.675679    81.18224
Paris            -156.836257  -211.13911
Rome              709.413282  1109.36665
Stockholm         839.445911 -1836.79055
Vienna            911.230500   205.93020
```

Uma representação gráfica destas coordenadas obtém-se imediatamente recorrendo ao comando `plot`, ou seja, apenas através do comando `plot(cmdscale(eurodist))`. A fim de incluir as legendas com os nomes das cidades junto aos pontos, podem dar-se os seguintes comandos, que produzem a Figura 5.1:

```
> plot(cmdscale(eurodist), type="n")
> text(cmdscale(eurodist), lab=labels(eurodist), cex=0.7, col=''red'')
```

Na disposição das localidades, Sul e Norte aparecem invertidos em relação às posições cartográficas convencionais. Como já foi referido, o método apenas é capaz de recuperar posições relativas, a menos de rotações rígidas, reflexões em torno dos eixos e translações da origem. Além disso, a disposição Sul/Norte habitual na cartografia actual não passe de uma convenção. Pode recuperar-se (de forma aproximada) essa disposição multiplicando as coordenadas do segundo eixo por -1. Introduzindo ainda uma ligeira rotação rígida, a fim de melhor adequar a posição das localidades aos mapas habituais, pode obter-se a representação indicada na Figura 5.2.

Embora a representação seja bastante adequada, permanece em aberto a questão da qualidade desta representação, do ponto de vista quantitativo. Solicitando a informação sobre a totalidade dos valores próprios da matriz $\mathbf{Q}$, obtem-se:

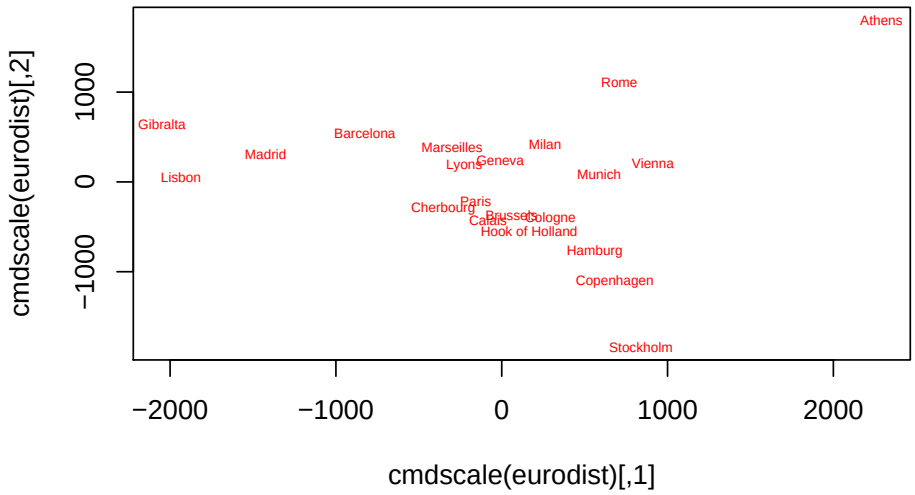


Figura 5.1: Representação gráfica de 21 localidades europeias nos dois primeiros eixos coordenados definidos pela matriz das respectivas distâncias rodoviárias.

```
> cmdscale(eurodist,k=20,eig=TRUE)$eig
 [1]  1.953838e+07  1.185656e+07  1.528844e+06  1.118742e+06  7.893472e+05
 [6]  5.816552e+05  2.623192e+05  1.925976e+05  1.450845e+05  1.079673e+05
[11]  5.139484e+04 -4.656613e-10 -9.496124e+03 -5.305820e+04 -1.322166e+05
[16] -2.573360e+05 -3.326719e+05 -5.162523e+05 -9.191491e+05 -1.006504e+06
Warning messages:
1: some of the first20eigenvalues are < 0 in: cmdscale(eurodist, k = 20, eig = TRUE)
2: NaNs produced in: sqrt(ev)
```

Como o próprio programa informático *R* previne, vários dos valores próprios de $\mathbf{Q}$ são negativos, e com valores absolutos bastante relevantes. Esta situação reflecte, como se viu anteriormente, a impossibilidade de forçar os 21 pontos a respeitarem, num espaço euclidiano de dimensão 20, as distâncias indicadas na matriz de dissemelhanças inicial. Tal facto não deve surpreender, uma vez que se trata de distâncias rodoviárias, e não de distâncias em linha recta. Pense-se, em particular, nas discrepâncias entre distâncias rodoviárias e distâncias em linha recta entre, por exemplo, Roma e Gibraltar, ou Atenas e Lisboa. Mas é de registar que o maior módulo dum valor próprio negativo é muito grande: cerca de um milhão, e mais de 5% do maior valor próprio positivo. Não obstante, o mapa produzido é reconhecível como uma aproximação das convencionais cartas bidimensionais da Europa.

É possível pedir ao programa *R* para resolver este problema através do método da constante aditiva, descrito na Secção anterior. Assim, colocando o argumento lógico `add` com o valor `TRUE`, obtêm-se os seguintes valores próprios da matriz modificada:

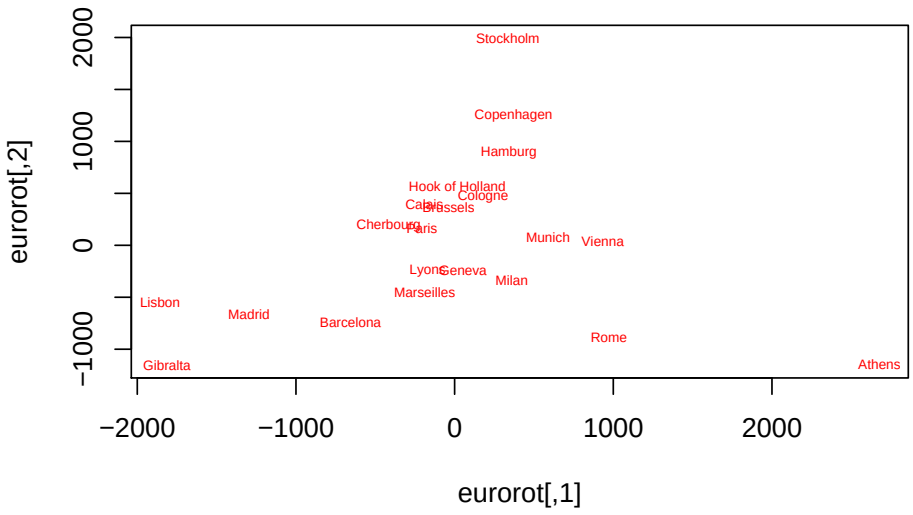


Figura 5.2: Representação gráfica de 21 localidades europeias nos dois primeiros eixos coordenados definidos pela matriz das respectivas distâncias rodoviárias, após uma rotação e reflexão conveniente.

```
> cmdscale(eurodist,k=20,eig=TRUE,add=TRUE)$eig
 [1] 42274973.8 31666186.4 16687778.0  8915634.6  6374452.9  5625824.9
 [7]  5414675.3  4318270.7  4039243.6  3690927.8  3501743.4  3433239.4
[13]  2838067.0  2689204.1  2621767.9  2262620.5  2031862.8  1690924.1
[19]  1202562.8    43505.2
```

A representação gráfica das 21 localidades nas coordenadas dos dois primeiros eixos obtidos após esta "correcção aditiva" da matriz de distâncias é dada na Figura 5.3 (já com simetrias em torno dos eixos e rotação adequadas).

Uma melhor compreensão dos efeitos da constante aditiva pode ser obtido sobrepondo neste gráfico as designações das localidades correspondentes à análise anterior, como se pode ver na Figura 5.4

No que respeita às técnicas algoritmicas de MDS, a distribuição base do `R` não disponibiliza funções para este fim. Mas no já referido módulo `MASS` encontram-se duas funções para correr outras tantas variantes destes métodos: as funções `sammon` e `isoMDS`.

A função `sammon` procura optimizar o critério (5.9). Tal como a função `cmdscale`, exige como argumento de entrada uma matriz de dissemelhanças, na forma duma estrutura `dist`. Assim, para correr o algoritmo de Sammon com os dados das distâncias entre cidades europeias, basta carregar o módulo `MASS` e depois invocar o comando `sammon`:

```
> library(MASS)
```

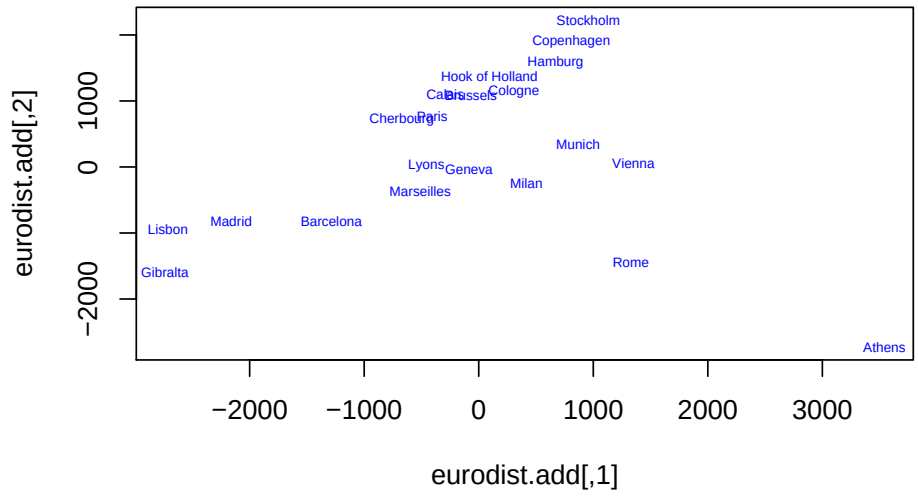


Figura 5.3: Representação gráfica de 21 localidades europeias nos dois primeiros eixos coordenados obtidos somando à matriz de distâncias rodoviárias uma constante aditiva que garanta uma representação euclidiana.

```
> sammon(eurodist)
Initial stress        : 0.01705
stress after  10 iters: 0.00951, magic = 0.500
stress after  20 iters: 0.00941, magic = 0.500
$points
                      [,1]        [,2]
Athens          1921.911103  1830.43091
Barcelona       -759.759842   606.08791
Brussels          80.198892  -443.36596
Calais          -106.206709  -512.10099
Cherbourg       -484.412859  -477.30458
Cologne          295.332393  -445.05488
Copenhagen       543.941038 -1091.58823
Geneva            -7.409617   269.68470
Gibralta        -1942.803908   727.82877
Hamburg           626.715320  -721.55069
Hook of Holland   185.861267  -658.28592
Lisbon          -1916.640601    83.58421
Lyons            -149.485025   217.70897
Madrid          -1372.706468   349.72549
Marseilles       -285.256846   514.72780
```

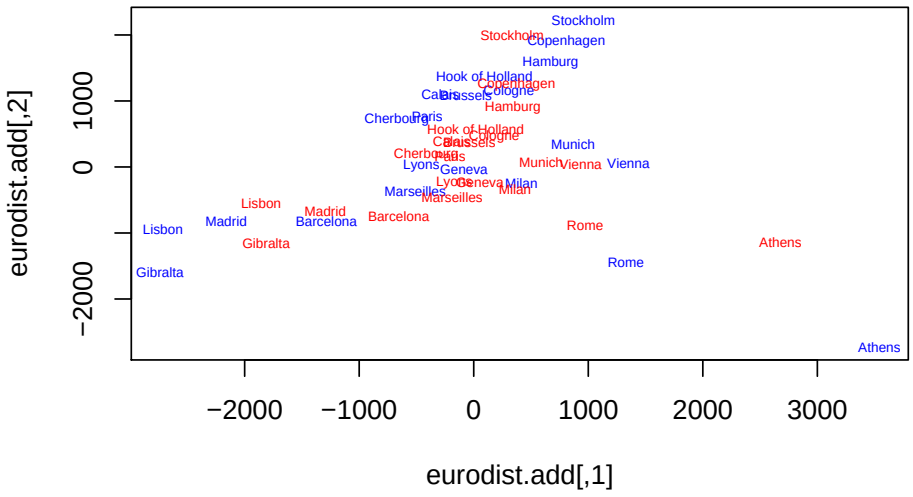


Figura 5.4: Sobreposição das representações gráficas de 21 localidades europeias dadas nas Figuras 5.2 (a azul, logo mais escuro) e 5.3 (a vermelho, mais claro).

```
Milan              273.108570   426.49832
Munich             569.246009   106.53333
Paris             -161.492199  -261.15118
Rome               698.672897  1023.66054
Stockholm          951.877637 -1716.50563
Vienna            1039.308949   170.43709
$stress
[1] 0.009413915
$call
sammon(d = eurodist)
```

Como se pode verificar, no objecto de saída encontram-se as coordenadas dos pontos (por omissão, o comando trabalha com `k`=2, mas pode ser alterado este valor do argumento), numa componente da lista com nome `points`. Também é devolvido o valor da função (5.9), no objecto de nome `stress` (não confundir com a função definida na equação 5.10). Este algoritmo necessita de uma configuração inicial que, por omissão, é a configuração (com a mesma dimensionalidade `k`) produzida pela Análise em Coordenadas Principais, ou seja, obtida com o comando `cmdscale`. Outra configuração inicial pode ser fornecida ao algoritmo através do argumento de nome `y`. A configuração de pontos final, cujas coordenadas são indicadas num objecto de nome `points`, é determinada por uma condição de paragem. Esta condição tanto pode ser o argumento `tol`, que especifica um limiar de variação no indicador *STRESS* abaixo da qual o algoritmo para, ou então um número máximo de iterações (argumento `niter`). Por omissão, a

função `sammon` trabalha com os valores `tol`=0.0001 e `maxit`=100.

Uma representação euclididana baseada em técnicas de *Scaling* não-métrico pode ser obtida, no programa R, utilizando o comando `isoMDS` do pacote `MASS`. A função `isoMDS` usa a regressão isotónica para determinar disparidades que substituem as dissemelhanças, de forma a minimizar o $SRESS$, mas preservando a ordem das dissemelhanças correspondentes. O comando `isoMDS` também exige como argumentos de entrada uma matriz de dissemelhanças (sob a forma dum objecto R de tipo `dist`) e, opcionalmente, uma configuração inicial de marcadores para cada indivíduo no espaço euclidiano de dimensão `k`. Por omissão, a função `isoMDS` considera `k`=2 e toma como configuração inicial a que resultou de efectuar uma Análise em Coordenadas Principais (através do comando `cmdscale`). Para os dados do exemplo desta Secção, bastaria dar o comando (caso já se tenha carregado o módulo `MASS`):

```
 > isoMDS(eurodist)
initial  value 7.505733
final  value 7.505688
converged
$points
                        [,1]        [,2]
Athens           2290.272645  1798.80178
Barcelona        -825.382640   546.81213
Brussels           59.183908  -367.08187
Calais            -82.845949  -429.91529
Cherbourg        -352.501462  -290.91022
Cologne           293.690514  -405.31357
Copenhagen        681.931345 -1108.64406
Geneva             -9.423649   240.40785
Gibralta        -2048.448589   642.45861
Hamburg           561.109087  -773.36918
Hook of Holland   164.922544  -549.36849
Lisbon          -1935.042299    49.12542
Lyons            -226.422997   187.08830
Madrid          -1423.353178   305.87512
Marseilles       -299.499291   388.80974
Milan             260.878614   416.67476
Munich            587.676381    81.18271
Paris            -156.836532  -211.13914
Rome              709.412629  1109.36503
Stockholm         839.446235 -1836.79051
Vienna            911.232682   205.93089
$stress
[1] 7.505688
> text(isoMDS(eurodist)$points, lab=labels(eurodist), cex=0.7, col="blue")
initial  value 7.505733
final  value 7.505688
converged
```

A condição de paragem determina-se de forma parecida com as da função `sammon`, mas neste caso o argumento que controla o número máximo de iterações chama-se `maxit` e, por omissão, tem valor 50, enquanto que o argumento `tol` tem, por omissão, valor `tol`=0.001. Neste caso, e uma vez que não foi especificada uma configuração inicial, é utilizada a configuração inicial resultante da Análise em Coordenadas Principais, verificando-se uma convergência rápida para uma configuração quase indistinguível da anterior, com um valor de *STRESS* de 7.505688 (comparado com o valor de *STRESS* de 7.505733 para a configuração produzida pela Análise em Coordenadas Principais). Assinale-se que, se tivesse sido pedida, como configuração inicial, o resultado do algoritmo da função `sammon`, através do comando

```
> isoMDS(eurodist,y=sammon(eurodist)$points)
```

a solução produzida já seria diferente, com um valor de $STRESS = 6.268603$. Este exemplo ilustra que pode ser conveniente ensaiar diferentes configurações iniciais. A Figura 5.5 foi criada com os comandos

```
> plot(isoMDS(eurodist)$points,type="n")
> text(isoMDS(eurodist)$points,labels=labels(eurodist),col="red",cex=0.6)
> text(sammon(eurodist)$points,labels=labels(eurodist),col="green",cex=0.6)
```

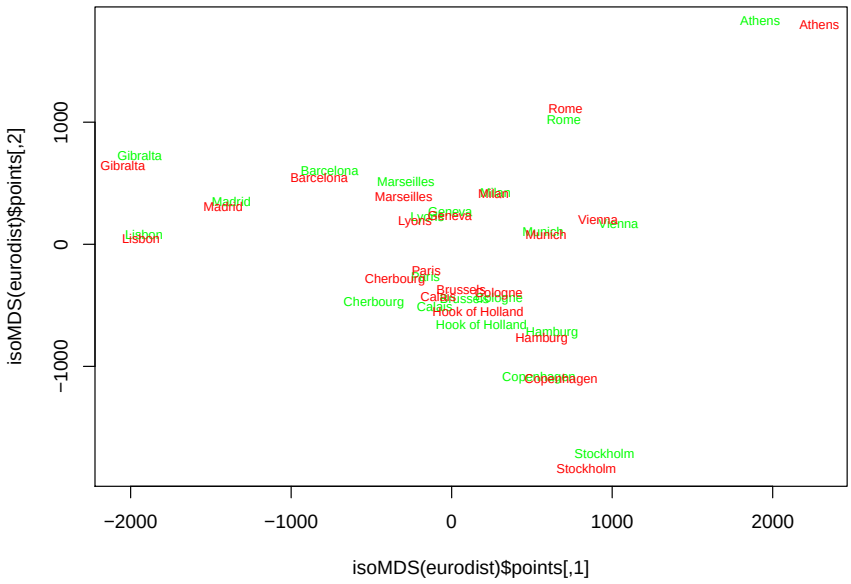


Figura 5.5: Sobreposição das representações gráficas de 21 localidades europeias através das funções `isoMDS` (a vermelho, logo mais escuro) e `sammon` (a verde, mais claro).

5.6 Exercícios

Na pasta da disciplina de Estatística Multivariada (`MMACB/em`) na área `//prunus/home/cadeiras`, está disponível uma ficheiro de nome `exerSca.RData`, que contem conjuntos de dados que serão utilizados nestes exercícios. Esse ficheiro deve ser lido para dentro duma sessão de trabalho do `R` (veja as instruções no início dos Exercícios do Capítulo 2 (página 78) para as instruções sobre como proceder).

1. O objecto `mapaacp` contem uma matriz simétrica (18×18) com as distâncias rodoviárias entre as 18 capitais de distrito, constantes do Mapa do Automóvel Clube de Portugal (em meados dos anos 90!). Com base nesses dados, responda às seguintes questões:

   (a) Efectue uma Análise em Coordenadas Principais, e produza a representação gráfica a duas dimensões correspondente. Coloque os nomes das colunas da matriz em cima das coordenadas. Procure identificar o mapa do território nacional. Comente.

   (b) Inspeccione os valores próprios da matriz (duplamente centrada) de produtos internos. Comente. Em particular, comente o significado da existência de valores próprios negativos.

   (c) Discuta as hipóteses de medir a qualidade da representação bidimensional que obteve na primeira alínea.

   (d) Caso optasse por utilizar o critério do valor absoluto para escolher quantas coordenadas principais reter, qual seria a sua resposta? Inspeccione as coordenadas que uma Análise em Coordenadas Principais associaria à(s) dimensão(ões) adicional(is). Comente. Em particular, diga quais as capitais de distrito que parecem sofrer mais com a representação bidimensional.

   (e) Utilize agora a função `sammon` do `R` para obter uma configuração bidimensional alternativa. Sobreponha a representação agora obtida à representação gerada pela Análise em Coordenadas Principais. Comente.

   (f) Qual foi a configuração inicial que a função `sammon` utilizou na alínea anterior?

   (g) Obtenha uma terceira configuração, desta vez utilizando a função `isoMDS`. Considere como configuração inicial a que obteve na primeira alínea. Comente.

   (h) Repita a alínea anterior, mas desta vez ponha o argumento `tol` com o valor 10^{-7}. Comente o significado desse argumento e comente as diferenças obtidas nos resultados.

   (i) Repita a alínea anterior, mas acrescentando o argumento `maxit` com valor 100. Experimente também o valor 200. Comente.

   (j) Utilize de novo a função `isoMDS`, mas com configuração inicial a configuração produzida na alínea 1e). Comente os resultados obtidos. Caso considere necessário, sobreponha esta configuração às anteriores.

2. No objecto `mapaacp42` está uma matriz de distâncias retirado novamente do Mapa do Automóvel Clube, mas desta vez com 42 localidades que incluem, além das 18 capitais de distrito, Fátima e mais 23 postos de fronteira. Repita o estudo do Exercício 1, agora para as 42 localidades.

3. Considere de novo os dados referentes aos $n = 63$ lavagantes, que já foram alvo da nossa atenção em Capítulos anteriores. Para esses $n = 63$ individuos, crie as seguintes matrizes de distâncias: (i) métrica Euclideana usual; (ii) métrica de Minkowski com $\lambda = 1$ ("Manhattan"); e (iii) métrica de Canberra.

   (a) Efectue uma Análise de Coordenadas Principais com base na matriz de distâncias Euclideanas habituais. Em particular:

      i. Compare numericamente os valores dos 63 individuos nos dois eixos coordenados produzidos pelo comando `cmdscale` com os *scores* dos 63 lavagantes nas duas primeiras componentes principais (sobre a matriz de covariâncias). Comente.

      ii. Pedindo (através do argumento `eig=TRUE` e com `k=62`) para se ver os valores próprios da matriz de produtos internos, surgem numerosos valores próprios negativos. Como se pode explicar esse facto, se a representação euclidiana exacta tem de ser possível (uma vez que conmeçámos com $n = 63$ vectores em $\mathbb{R}^{13}$ e trabalhámos com distãncias euclidianas)?

      iii. O grande número de valores próprios nulos da matriz de produtos internos é uma coincidência numérica ou é a consequência necessária de alguma característica dos dados? Justifique a sua resposta.

      iv. Compare de novo os 13 valores próprios não nulos da matriz de produtos internos com os treze valores próprios da matriz de variâncias dos dados originais. Porque não são iguais estes dois conjuntos de valores, uma vez que as configurações definidas são iguais?

   (b) Efectua-se uma nova Análise de Coordenadas Principais dos dados dos lavagantes, mas agora tendo por base a métrica de Canberra.

      i. Construa a representação gráfica bi-dimensional que melhor reflecte as distâncias de Canberra entre os lavagantes. Comente o resultado.

      ii. Volte a definir o gráfico resultante da Análise em Coordenadas Principais com base na matriz de distâncias euclidianas e sobreponha-lhe a configuração que agora obteve. Como explica a evidente diferença de escalas obtidas? É legítimo procurar sobrepôr as duas configurações? Que significado teria essa sobreposição? E como se poderia determinar o efeito de escala (factor da homotetia) que é necessário usar?

      iii. Discuta o facto de haver muito mais valores próprios não-nulos do que no caso da Análise feita com a métrica Euclideana. Em particular, comente o que pode estar na origem desse facto. O último valor próprio tem de ser exactamente zero?

   (c) Efectue a Análise de Coordenadas Principais com base na métrica de Minkowski com $\lambda = 1$.

      i. Compare os resultados com os resultados obtidos utilizando as distâncias Euclideanas e de Canberra. Comente.

      ii. Sobreponha a configuração agora obtida às configurações anteriores. Porque não surgem a maioria dos pontos no gráfico?

4. Num estudo meteorológico dispõe-se de 26 postos meteorológicos, em cada um dos quais foram medidas três variáveis: temperatura média, precipitação total e número de dias de chuva. Para cada posto, partiu-se duma matriz 4x3 com os valores médios de cada variável em cada uma das

quatro estações de um ano, e calcularam-se as respectivas 26 matrizes de correlações. Estas 26 matrizes foram então comparadas através duma medida de semelhança. A matriz de semelhanças resultante serviu, por sua vez, de base a uma Análise em Coordenadas Principais (*Classical Scaling*). Os 26 postos meteorológicos e a respectiva matriz de semelhanças foram:

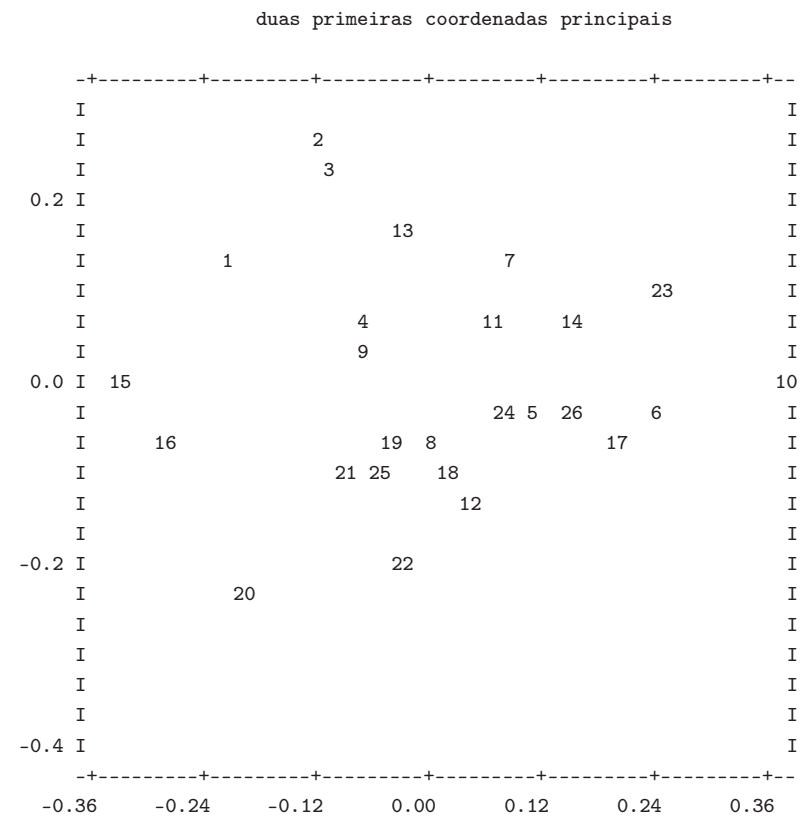
No.	Local	No.	Local	No.	Local	No.	Local
1	Viana do Castelo	8	Peso da Régua	14	Alcobaça	21	Santiago do Cacém
2	Braga	9	Viseu	15	Santarém	22	Beja
3	Santo Tirso	10	Guarda	16	Setúbal	23	Amareleja
4	Montalegre	11	Coimbra	17	Portalegre	24	S. Brás Alportel
5	Bragança	12	Castelo Branco	18	Elvas	25	Monchique
6	Pedras Salgadas	13	Pombal	19	Évora	26	Tavira
7	Mirandela			20	Alcácer do Sal		

```
 1  1.00
 2  0.98  1.00
 3  0.98  0.99  1.00
 4  0.96  0.96  0.97  1.00
 5  0.92  0.91  0.94  0.97  1.00
 6  0.88  0.88  0.90  0.93  0.98  1.00
 7  0.95  0.97  0.96  0.95  0.96  0.96  1.00
 8  0.95  0.92  0.94  0.97  0.98  0.97  0.96  1.00
 9  0.97  0.96  0.98  0.98  0.98  0.94  0.96  0.98  1.00
10  0.82  0.84  0.85  0.88  0.95  0.99  0.94  0.92  0.88  1.00
11  0.95  0.95  0.97  0.98  0.99  0.97  0.98  0.98  0.98  0.94  1.00
12  0.92  0.89  0.92  0.96  0.99  0.97  0.94  0.99  0.97  0.93  0.98  1.00
13  0.96  0.97  0.95  0.94  0.91  0.91  0.98  0.93  0.93  0.88  0.95  0.90  1.00
14  0.92  0.92  0.94  0.92  0.96  0.98  0.98  0.96  0.95  0.96  0.97  0.95  0.94  1.00
15  0.98  0.93  0.94  0.93  0.88  0.83  0.90  0.94  0.95  0.75  0.91  0.91  0.91  0.87  1.00
16  0.98  0.93  0.93  0.96  0.91  0.86  0.90  0.95  0.96  0.78  0.93  0.93  0.92  0.87  0.99  1.00
17  0.89  0.89  0.91  0.95  0.99  1.00  0.96  0.98  0.96  0.98  0.98  0.98  0.91  0.97  0.85  0.88
18  0.93  0.91  0.93  0.98  0.99  0.96  0.95  0.98  0.97  0.92  0.98  0.99  0.93  0.93  0.91  0.95
19  0.96  0.93  0.93  0.96  0.96  0.95  0.96  0.99  0.97  0.91  0.97  0.98  0.96  0.95  0.95  0.97
20  0.94  0.87  0.89  0.92  0.92  0.88  0.88  0.96  0.94  0.81  0.92  0.96  0.88  0.89  0.96  0.98
21  0.95  0.91  0.91  0.93  0.93  0.92  0.94  0.97  0.94  0.87  0.94  0.96  0.95  0.93  0.95  0.97
22  0.93  0.88  0.89  0.93  0.95  0.94  0.93  0.98  0.94  0.90  0.95  0.97  0.93  0.93  0.93  0.95
23  0.88  0.91  0.90  0.91  0.94  0.97  0.98  0.93  0.90  0.98  0.95  0.92  0.95  0.96  0.82  0.84
24  0.94  0.93  0.93  0.96  0.97  0.98  0.98  0.99  0.96  0.95  0.98  0.98  0.96  0.97  0.91  0.93
25  0.95  0.91  0.91  0.94  0.95  0.94  0.95  0.98  0.95  0.90  0.95  0.97  0.95  0.94  0.95  0.96
26  0.92  0.92  0.93  0.95  0.98  0.99  0.98  0.98  0.96  0.97  0.98  0.98  0.95  0.98  0.88  0.90
        1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16
17  1.00
18  0.97  1.00
19  0.96  0.98  1.00
20  0.90  0.95  0.97  1.00
21  0.93  0.95  0.99  0.97  1.00
22  0.95  0.97  0.99  0.98  0.99  1.00
23  0.96  0.92  0.94  0.84  0.92  0.92  1.00
24  0.98  0.98  0.99  0.94  0.98  0.98  0.97  1.00
25  0.95  0.97  1.00  0.97  1.00  1.00  0.93  0.99  1.00
26  0.99  0.97  0.98  0.92  0.96  0.97  0.98  0.99  0.97  1.00
      17    18    19    20    21    22    23    24    25    26
```

Foi efectuada uma Análise em Coordenadas Principais tendo por base esta matriz de semelhanças, sendo primeiro criada uma matriz de distâncias com base na relação $d_{ij} = 2(1 - s_{ij})$, onde d_{ij} é a distância entre os indivíduos i e j, e s_{ij} é a respectiva semelhança, obtida a partir da matriz acima. Os resultados desta Análise em Coordenadas Principais (incluíndo o gráfico dos 26 pontos nos dois primeiros eixos definidos pelo método) foram:

```
*****  Principal coordinates analysis  *****
***  Latent Roots  ***                                        ***  Trace  ***
    PCOlrv['Roots']                                           PCOlrv['Trace']
              1           2           3           4
         0.6922      0.3862      0.2276      0.0802                  1.428
***  Percentage variation  ***
    PCOlrv['Roots']
              1           2           3           4
          48.47       27.04       15.94        5.61
(NOTA: Nao houve valores proprios negativos)
```

```
***  Latent vectors (coordinates)  ***
             PCOlrv['Vectors']
                       1           2           3           4
           1    -0.21868     0.12749     0.01434    -0.02383
           2    -0.11719     0.27985     0.01375     0.02247
           3    -0.10625     0.22267    -0.10032    -0.06713
           4    -0.07027     0.07639    -0.14297     0.11702
           5     0.10327    -0.03700    -0.14909     0.02551
           6     0.23702    -0.04651    -0.03214    -0.01610
           7     0.08758     0.14944     0.06968    -0.00346
           8    -0.00268    -0.07880    -0.05761    -0.03926
           9    -0.07290     0.04783    -0.14725    -0.05563
          10     0.36670    -0.00689     0.02511    -0.00197
          11     0.05772     0.05338    -0.09652     0.02783
          12     0.04147    -0.14632    -0.09907    -0.01954
          13    -0.03531     0.17913     0.17870     0.06862
          14     0.14942     0.05949     0.02620    -0.15738
          15    -0.33270    -0.00104     0.01653    -0.06809
          16    -0.28681    -0.05791    -0.00579     0.04663
          17     0.19205    -0.05406    -0.07999    -0.00958
          18     0.01029    -0.08340    -0.09635     0.11896
          19    -0.04606    -0.07792     0.06328     0.02296
          20    -0.20601    -0.22108    -0.00213    -0.03130
          21    -0.10197    -0.11404     0.15419    -0.00047
          22    -0.03938    -0.18600     0.09439     0.00550
          23     0.23578     0.10041     0.14019     0.03848
          24     0.06922    -0.04332     0.06335     0.01801
          25    -0.05970    -0.11616     0.11771    -0.00076
          26     0.14539    -0.02564     0.03181    -0.01750
```



(a) Discuta o significado do gráfico acima reproduzido. Discuta também a qualidade da informação que pode extrair desse gráfico.

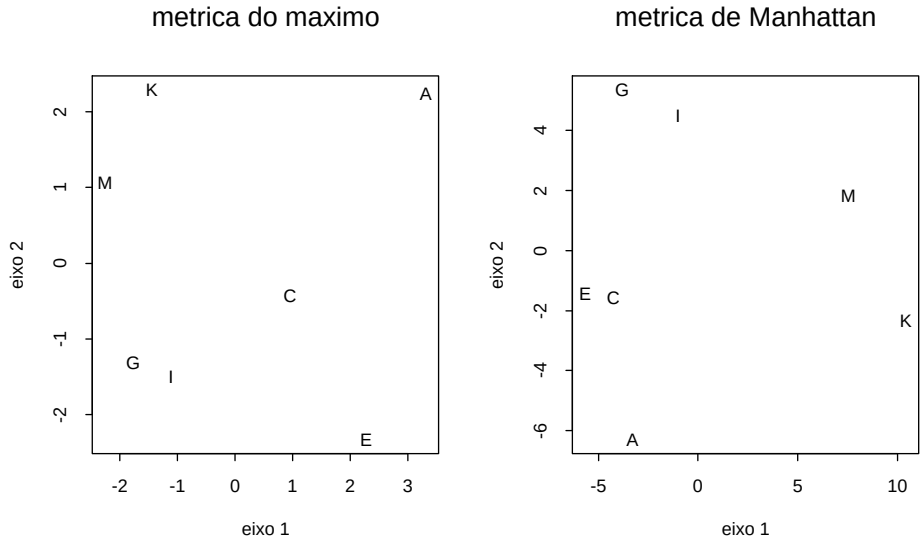
(b) Independentemente da sua resposta à alínea anterior, e admitindo que o gráfico é uma representação adequada, procure interpretar os eixos e a arrumação dos pontos com base nos seus conhecimentos geográficos e meteorológicos.

(c) Que significado pode associar ao facto de não ter havido valores próprios negativos nos resultados desta Análise?

5. Procurou-se representar num espaço Euclideano bi-dimensional as distâncias entre as sete localidades do estudo referido no enunciado do Exercício 8 do Capítulo 2 sobre Análise em Componentes Principais, no caso de essas distâncias serem medidas a partir das observações de log-frequências médias, mas com base nas métricas *do máximo* e *de Manhattan*. Seguem-se as representações obtidas e os sete valores próprios de cada matriz de distâncias.

```
> cmdscale(dist(moth.sub,metric="max"),eig=T,k=7)$eig
[1]  2.852e+001  2.099e+001  6.226e+000  2.505e+000  4.917e-007 -4.126e-001 -4.384e+000

> cmdscale(dist(moth.sub,metric="man"),eig=T,k=7)$eig
[1]  2.417e+002  1.018e+002  6.559e+001  3.396e+001  3.728e+000 -1.323e-005 -9.750e+000
```



(a) Comente a qualidade da representação das matrizes de distâncias das métricas do máximo e de Manhattan.

(b) Compare com a qualidade da representação que se obtem utilizando a matriz das distâncias Euclideanas. Comente.

Capítulo 6

Análise em Correlações Canónicas

Neste Capítulo analisaremos a Análise em Correlações Canónicas, uma técnica que parte da existência de **dois conjuntos de variáveis**, e procura sucessivas **combinações lineares de cada conjunto que sejam o mais correlacionadas possível**.

Num contexto descritivo, a existência de dois conjuntos de variáveis significa a existência de duas matrizes de dados, $\mathbf{X}_{n\times p}$ e $\mathbf{Y}_{n\times q}$, cujas linhas correspondem a observações, para um mesmo individuo, dos respectivos valores nos dois conjuntos de variáveis. A cardinalidade de cada conjunto de variáveis não tem de ser igual (isto é, tem-se, em geral, $p \neq q$), nem a sua natureza tem de ser semelhante. Procurar combinações lineares de cada conjunto de variáveis significa, neste contexto descritivo, procurar vectores de $\mathcal{C}(\mathbf{X})$ e $\mathcal{C}(\mathbf{Y})$, e o objectivo desta Análise é procurar vectores desse tipo o mais possível correlacionados, com sucessivas soluções ortogonais às soluções anteriores. Assim, a atenção da Análise em Correlações Canónicas centra-se na relação entre os subespaços gerados pelas colunas de $\mathbf{X}$ e pelas colunas de $\mathbf{Y}$. É desde já possível antever que, no caso da intersecção $\mathcal{C}(\mathbf{X}) \cap \mathcal{C}(\mathbf{Y})$ incluir mais do que a origem de $\mathbb{R}^n$ (que pertence, obrigatoriamente, a qualquer subespaço), terá de haver uma combinação linear comum a $\mathcal{C}(\mathbf{X})$ e $\mathcal{C}(\mathbf{Y})$, pelo que haverá pelo menos uma correlação canónica igual à identidade.

No que se segue **admite-se que as colunas, quer de X, quer de Y formam conjuntos linearmente independentes**, isto é, que não há multicolinearidades em qualquer dos grupos de variáveis. Admitimos também que **as colunas destas matrizes foram centradas**.

6.1 O método

Combinações lineares das colunas de $\mathbf{X}$ são da forma $\mathbf{Xa}$, para algum vector $\mathbf{a} \in \mathbb{R}^p$. Combinações lineares das colunas de $\mathbf{Y}$ são da forma $\mathbf{Yb}$, para vectores $\mathbf{b} \in \mathbb{R}^q$. **Procuram-se os vectores** $\mathbf{a} \in \mathbb{R}^p$

e $\mathbf{b} \in \mathbb{R}^q$ que maximizem a correlação[1] entre Xa e Yb:

$$\max_{\mathbf{a}\in\mathbb{R}^p,\mathbf{b}\in\mathbb{R}^q} \frac{\mathbf{a}^t\boldsymbol{\Sigma}_{X,Y}\mathbf{b}}{\sqrt{\mathbf{a}^t\boldsymbol{\Sigma}_{X,X}\mathbf{a}\cdot\mathbf{b}^t\boldsymbol{\Sigma}_{Y,Y}\mathbf{b}}} \tag{6.1}$$

onde $\boldsymbol{\Sigma}_{X,X} = \frac{1}{n}\mathbf{X}^t\mathbf{X}$ indica a matriz $p\times p$ de (co-)variâncias entre as variáveis do primeiro grupo, $\boldsymbol{\Sigma}_{Y,Y} = \frac{1}{n}\mathbf{Y}^t\mathbf{Y}$ indica a matriz $q\times q$ de (co-)variâncias entre as variáveis do segundo grupo, e $\boldsymbol{\Sigma}_{X,Y} = \frac{1}{n}\mathbf{X}^t\mathbf{Y}$ indica a matriz $p\times q$ de covariâncias cruzadas entre variáveis do primeiro grupo e variáveis do segundo grupo. Assinale-se que as matrizes de covariâncias e covariâncias cruzadas acima indicadas podem ser vistas como submatrizes da matriz de (co-)variâncias da totalidade das variáveis, vistas como integrando um único grupo. Assim, sendo $\boldsymbol{\Sigma}$ a matriz $(p+q)\times(p+q)$ de variâncias-covariâncias das $p+q$ variáveis que constituem as colunas das matrizes $\mathbf{X}$ e $\mathbf{Y}$, tem-se:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{X,X} & \vdots & \boldsymbol{\Sigma}_{X,Y} \\ \cdots\cdots & \cdots & \cdots\cdots \\ \boldsymbol{\Sigma}_{Y,X} & \vdots & \boldsymbol{\Sigma}_{Y,Y} \end{bmatrix} \tag{6.2}$$

Exercício 6.1 *Confirme a estrutura da matriz Σ indicada na equação (6.2).*

O primeiro par de soluções

No entanto, não será necessário trabalhar directamente com a expressão (6.1). De facto, quaisquer que venham a ser as combinações lineares $\mathbf{Xa}$ de colunas de $\mathbf{X}$ e $\mathbf{Yb}$ de colunas de $\mathbf{Y}$ que resolvam o problema, sabemos que a combinação linear de colunas de $\mathbf{Y}$ mais correlacionada com $\mathbf{Xa}$ será um vector colinear com a projecção ortogonal de $\mathbf{Xa}$ sobre $\mathcal{C}(\mathbf{Y})$ (Teorema de Pitágoras, página 24, último ponto). Por outras palavras, tem de ter-se $\mathbf{Yb} = \alpha\mathbf{P}_Y\mathbf{Xa}$, para algum escalar (não nulo) α, sendo $\mathbf{P}_Y$ a matriz de projecção ortogonal sobre o subespaço gerado pelas colunas da matriz $\mathbf{Y}$. De forma análoga, qualquer que venha a ser a combinação linear $\mathbf{Yb}$ de colunas de $\mathbf{Y}$ que resolva o problema, a combinação linear de colunas de $\mathbf{X}$ mais correlacionada com $\mathbf{Yb}$ tem de ser colinear com a projecção ortogonal de $\mathbf{Yb}$ sobre $\mathcal{C}(\mathbf{X})$. Ou seja, tem de ter-se $\mathbf{Xa} = \beta\mathbf{P}_X\mathbf{Yb}$, para algum escalar β, sendo $\mathbf{P}_X$ a matriz de projecção ortogonal sobre o subespaço gerado pelas colunas da matriz $\mathbf{X}$. Assim, os vectores $\mathbf{a}, \mathbf{b}$ procurados serão soluções do sistema:

$$\begin{cases} \mathbf{Xa} &= \beta\ \mathbf{P}_X\mathbf{Yb} \\ \mathbf{Yb} &= \alpha\ \mathbf{P}_Y\mathbf{Xa} \end{cases} \tag{6.3}$$

Substituindo em cada um dos membros direitos a expressão dada pela outra equação, resulta o seguinte sistema:

$$\begin{cases} \mathbf{Xa} &= \alpha\beta\ \mathbf{P}_X\mathbf{P}_Y\mathbf{Xa} \\ \mathbf{Yb} &= \alpha\beta\ \mathbf{P}_Y\mathbf{P}_X\mathbf{Yb} \end{cases} \tag{6.4}$$

[1] **No caso de haver uma correlação negativa entre Xa e Yb, haverá uma correlação positiva, de igual magnitude, entre $-\mathbf{Xa} = \mathbf{X}(-\mathbf{a})$ e Yb, pelo que não é limitativo falar apenas na maior correlação, abstraíndo do sinal dessa correlação.**

Por outras palavras, **a combinação linear Xa é um vector próprio da matriz $\mathbf{P}_X\mathbf{P}_Y$, e a combinação linear Yb é um vector próprio da matriz $\mathbf{P}_Y\mathbf{P}_X$**, que é a transposta da matriz anterior. Ambos estes vectores próprios estão associados a um mesmo valor próprio: $\lambda = \frac{1}{\alpha\beta}$. Ora, estas matrizes não são simétricas, pois $(\mathbf{P}_X\mathbf{P}_Y)^t = \mathbf{P}_Y\mathbf{P}_X \neq \mathbf{P}_X\mathbf{P}_Y$, em geral. Mas facilmente se mostra que **estes vectores próprios são igualmente vectores próprios de matrizes simétricas, com os mesmos valores próprios associados. Assim, é sempre possível obter um conjunto ortonormado de tais vectores e quer os vectores, quer os valores próprios, são reais**. De facto, uma combinação linear $\mathbf{Xa}$ das colunas de $\mathbf{X}$ pertence ao subespaço gerado por estas, $\mathcal{C}(\mathbf{X})$. Logo, permanece invariante quando projectada ortogonalmente sobre $\mathcal{C}(\mathbf{X})$. Assim, $\mathbf{P}_X\mathbf{Xa} = \mathbf{Xa}$. Este facto permite re-escrever a primeira equação do sistema 6.4 como: $\mathbf{P}_X\mathbf{P}_Y\mathbf{P}_X\boldsymbol{\eta} = \lambda\boldsymbol{\eta}$, onde $\boldsymbol{\eta} = \mathbf{Xa}$. Procedendo de forma análoga para a combinação linear no outro espaço, $\mathbf{Yb}$, resulta que as soluções $\boldsymbol{\eta} = \mathbf{Xa}$ e $\boldsymbol{\nu} = \mathbf{Yb}$ do sistema (6.4) são simultaneamente soluções do sistema:

$$\begin{cases} \mathbf{P}_X\mathbf{P}_Y\mathbf{P}_X\boldsymbol{\eta} = \lambda\boldsymbol{\eta} \\ \mathbf{P}_Y\mathbf{P}_X\mathbf{P}_Y\boldsymbol{\nu} = \lambda\boldsymbol{\nu} \end{cases} \tag{6.5}$$

Assim, as soluções do sistema (6.5) são dadas por um conjunto ortonormado de vectores $\{\eta_j\}_{j=1}^n$, vectores próprios de $\mathbf{P}_X\mathbf{P}_Y\mathbf{P}_X$, e $\{\nu_j\}_{j=1}^n$, vectores próprios de $\mathbf{P}_Y\mathbf{P}_X\mathbf{P}_Y$, que partilham um valor próprio comum às duas matrizes.

Ora, **os valores próprios não-nulos destas duas matrizes têm de ser iguais**, tendo em conta o Teorema 1.37, uma vez que tomando $\mathbf{A} = \mathbf{P}_X\mathbf{P}_Y$, tem-se $\mathbf{A}^t = \mathbf{P}_Y\mathbf{P}_X$, e as duas matrizes são da forma $\mathbf{A}\mathbf{A}^t = \mathbf{P}_X\mathbf{P}_Y\mathbf{P}_X$ e $\mathbf{A}^t\mathbf{A} = \mathbf{P}_Y\mathbf{P}_X\mathbf{P}_Y$. Assim, todos os vectores próprios associados a valores próprios não-nulos das duas matrizes surgem aos pares, sendo potenciais soluções do problema.

Para compreender qual o par destas combinações lineares que maximiza o critério (6.1), vejamos o significado dos valores próprios (necessariamente reais) $\{\lambda_j\}_{j=1}^n$ comuns a cada par de vectores próprios. **O valor próprio λ_j comum a $\boldsymbol{\eta}_j = \mathbf{Xa}_j$ e a $\boldsymbol{\nu}_j = \mathbf{Yb}_j$ representa o quadrado da correlação entre $\boldsymbol{\eta}_j$ e $\boldsymbol{\nu}_j$**. De facto, e como vimos acima, o vector $\boldsymbol{\nu}_j$ terá de ser proporcional à projecção ortogonal do vector $\boldsymbol{\eta}_j$ sobre $\mathcal{C}(\mathbf{Y})$. Logo, e tendo em conta que os vectores $\{\boldsymbol{\eta}_j\}_{j=1}^n$ formam um conjunto ortonormado,

$$\begin{aligned} \operatorname{cor}(\underline{\eta_j}, \boldsymbol{\nu}_j) &= \operatorname{cor}(\boldsymbol{\eta}_j, \mathbf{P}_Y\boldsymbol{\eta}_j) = \frac{\boldsymbol{\eta}_j^t\mathbf{P}_Y\boldsymbol{\eta}_j}{\sqrt{\boldsymbol{\eta}_j^t\boldsymbol{\eta}_j}\cdot\sqrt{\boldsymbol{\eta}_j^t\mathbf{P}_Y\boldsymbol{\eta}_j}} = \sqrt{\frac{\boldsymbol{\eta}_j^t\mathbf{P}_Y\boldsymbol{\eta}_j}{1}} \\ &= \sqrt{\boldsymbol{\eta}_j^t\mathbf{P}_X\mathbf{P}_Y\mathbf{P}_X\boldsymbol{\eta}_j} = \sqrt{\lambda_j\boldsymbol{\eta}_j^t\boldsymbol{\eta}_j} = \sqrt{\lambda_j} \end{aligned} \tag{6.6}$$

Assim sendo, é evidente que **as combinações lineares de cada grupo de variáveis que maximizam o critério (6.1) são os vectores $\boldsymbol{\eta}_1$ e $\boldsymbol{\nu}_1$, vectores próprios associados ao *maior* valor próprio, λ_1, comum a $\mathbf{P}_X\mathbf{P}_Y\mathbf{P}_X$ e $\mathbf{P}_Y\mathbf{P}_X\mathbf{P}_Y$**.

Outros pares de variáveis canónicas

Os restantes vectores e valores próprios das matrizes $\mathbf{P}_X\mathbf{P}_Y\mathbf{P}_X$ e $\mathbf{P}_Y\mathbf{P}_X\mathbf{P}_Y$ também são úteis. O par de vectores próprios $\boldsymbol{\eta}_2$ e $\boldsymbol{\nu}_2$ maximizam, *de entre as combinações lineares* $\boldsymbol{\eta} = \mathbf{Xa}$ *ortogonais a* $\boldsymbol{\eta}_1$ *e*

as combinações lineares $\boldsymbol{\nu} = \mathbf{Yb}$ *ortogonais a* $\boldsymbol{\nu}_1$, o critério (6.1), sendo λ_2 o valor do quadrado desse máximo restrito do critério. Sucessivos pares de vectores próprios $\boldsymbol{\eta}_j$, $\boldsymbol{\nu}_j$ são as sucessivas soluções de problemas análogos.

As combinações lineares $\boldsymbol{\eta}_j = \mathbf{Xa}_j$ e $\boldsymbol{\nu}_j = \mathbf{Yb}_j$ designam-se **variáveis de correlação canónica**[2], enquanto que os coeficientes de correlações entre os pares de variáveis de correlação canónica (isto é, os valores $\sqrt{\lambda_j}$) se designam as **correlações canónicas** dos dois conjuntos de variáveis.

Os coeficientes das variáveis de correlação canónica

Embora estejam identificadas as combinações lineares que são sucessivas soluções do problema, falta indicar explicitamente quais os *vectores de coeficientes que definem essas variáveis de correlação canónica*, isto é, quais os vectores $\mathbf{a}_j$ e $\mathbf{b}_j$. Esses vectores de coeficientes, designados **vectores (de coeficientes) de correlação canónica** são dados por:

$$\begin{cases} \mathbf{a}_j &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\eta}_j \\ \mathbf{b}_j &= (\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t\boldsymbol{\nu}_j \end{cases} \tag{6.7}$$

De facto, já vimos que $\boldsymbol{\eta}_j = \mathbf{Xa}_j = \mathbf{P}_X\mathbf{Xa}_j = \mathbf{P}_X\boldsymbol{\eta}_j$. Uma vez que se admitiu que as colunas de $\mathbf{X}$ eram linearmente independentes, cada sua combinação linear tem coeficientes únicos. Assim, o vector de coeficientes $\mathbf{a}_j$ tem de ser igual ao vector $(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\eta}_j$ que multiplica à direita $\mathbf{X}$ na última expressão. O caso dos coeficientes $\mathbf{b}_j$ sai por analogia.

Uma expressão alternativa mais comum, para estes vectores de coeficientes resulta de, pelas equações do sistema (6.4), se ter:

$$\begin{cases} \boldsymbol{\eta}_j &= \frac{1}{\lambda_j}\mathbf{P}_X\mathbf{P}_Y\mathbf{Xa}_j \\ \boldsymbol{\nu}_j &= \frac{1}{\lambda_j}\mathbf{P}_Y\mathbf{P}_X\mathbf{Yb}_j \end{cases} \tag{6.8}$$

Substituindo estas expressões no sistema (6.7), obtem-se:

$$\begin{cases} \lambda_j\mathbf{a}_j &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}(\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t\mathbf{Xa}_j \\ \lambda_j\mathbf{b}_j &= (\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t\mathbf{Y}(\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Yb}_j \end{cases}$$

que, simplificando, e tendo em atenção que $\Sigma_{X,X} = \frac{1}{n}\mathbf{X}^t\mathbf{X}$, $\Sigma_{X,Y} = \frac{1}{n}\mathbf{X}^t\mathbf{Y}$, $\Sigma_{Y,X} = \frac{1}{n}\mathbf{Y}^t\mathbf{X}$ e $\Sigma_{Y,Y} = \frac{1}{n}\mathbf{Y}^t\mathbf{Y}$, resulta no sistema:

$$\begin{cases} \lambda_j\mathbf{a}_j &= (\boldsymbol{\Sigma}_{X,X})^{-1}\boldsymbol{\Sigma}_{X,Y}(\boldsymbol{\Sigma}_{Y,Y})^{-1}\boldsymbol{\Sigma}_{Y,X}\mathbf{a}_j \\ \lambda_j\mathbf{b}_j &= (\boldsymbol{\Sigma}_{Y,Y})^{-1}\boldsymbol{\Sigma}_{Y,X}(\boldsymbol{\Sigma}_{X,X})^{-1}\boldsymbol{\Sigma}_{X,Y}\mathbf{b}_j \end{cases} \tag{6.9}$$

Por outras palavras, **os vectores de coeficientes das variáveis de correlações canónicas são os vectores próprios das matrizes** $(\Sigma_{X,X})^{-1}\Sigma_{X,Y}(\Sigma_{Y,Y})^{-1}\Sigma_{Y,X}$ **e** $(\Sigma_{Y,Y})^{-1}\Sigma_{Y,X}(\Sigma_{X,X})^{-1}\Sigma_{X,Y}$**, associados ao valor próprio comum** λ_j.

Observações:

[2] Ou até, simplesmente, **variáveis canónicas**. Atenção ao facto de, por vezes, se utilizar uma designação análoga para os eixos obtidos através duma Análise Discriminante Linear.

1. **Não podem existir mais do que $\min\{p, q\}$ pares de variáveis de correlações canónicas**. Tal facto é evidente do ponto de vista geométrico: não podem existir mais de k vectores ortogonais num espaço de dimensão k, e sendo as variáveis de correlações canónicas ortogonais entre si, o seu número não pode exceder a menor das dimensões dos espaços $\mathcal{C}(\mathbf{X})$ e $\mathcal{C}(\mathbf{Y})$. Mas também do ponto de vista dos conceitos de Teoria de Matrizes. Seguindo um raciocínio análogo ao usado aquando da discussão do número de eixos discriminantes (página 98), e tendo em conta que o número de pares de variáveis de correlações canónicas corresponderá ao número de valores próprios não nulos de $\mathbf{P}_X\mathbf{P}_Y$ (ou $\mathbf{P}_Y\mathbf{P}_X$), tem-se:

$$car(\mathbf{P}_X\mathbf{P}_Y) \leq \min\{car(\mathbf{P}_X), car(\mathbf{P}_Y)\} = \min\{p, q\} \tag{6.10}$$

   uma vez que as características de matrizes de projecção ortogonal são iguais aos seus traços, e estes têm valor igual às dimensões dos subespaços sobre os quais projectam (Teorema 1.26, página 26).

2. O papel do grupo de variáveis que constituem as colunas da matriz $\mathbf{X}$ e do grupo de variáveis que constituem as colunas da matriz $\mathbf{Y}$ é intercambiável, como pode verificar-se a partir da observação do sistema (6.5). Nesse sentido, o método é **simétrico em X e Y**.

3. A Análise das Correlações Canónicas é **invariante a transformações lineares (diferenciadas) nas escalas das variáveis de qualquer dos grupos**. De facto, e como foi visto aquando do estudo de métodos anteriores, esse tipo de transformações correspondem a pós-multiplicar as matrizes de colunas centradas $\mathbf{X}$ e $\mathbf{Y}$ por matrizes diagonais. Mas esse tipo de transformação não altera as matrizes de projecção ortogonal $\mathbf{P}_X$ e $\mathbf{P}_Y$, logo não altera os vectores/valores próprios de $\mathbf{P}_X\mathbf{P}_Y$ e $\mathbf{P}_Y\mathbf{P}_X$. No entanto, **os coeficientes das combinações lineares são alterados, caso se proceda a efectuar essas mudanças**, da forma necessária para manter a resultante da combinação linear inalterada. Assim, por exemplo, admita-se que se procede à normalização das colunas da matriz $\mathbf{X}$ (o que, tendo em conta que as variáveis foram centradas, significa apenas dividir cada coluna pelo respectivo desvio padrão). Esta operação corresponde a multiplicar $\mathbf{X}$, à direita, pela matriz diagonal $\mathbf{D}^{-1}$, cujo i-ésimo elemento diagonal é o recíproco do desvio padrão da variável i. Nesse caso, o vector de coeficientes teria de ser da forma $\mathbf{D}\mathbf{a}_i$, onde $\mathbf{a}_i$ é o vector dos coeficientes dos dados não-normalizados $\mathbf{X}$. É desta forma que a combinação linear fica igual: $\mathbf{X}\mathbf{D}^{-1}\mathbf{D}\mathbf{a}_i = \mathbf{X}\mathbf{a}_i = \boldsymbol{\eta}_i$.

6.2 A ACC como generalização de métodos anteriores

Alguns dos métodos de Estatística Multivariada anteriormente estudados podem ser vistos como casos particulares da Análise em Correlações Canónicas. Assim:

1. Se $q = 1$, e a matriz $\mathbf{Y}$ é constituída por um único vector-coluna $\mathbf{y}$, e nesse caso procurar a combinação linear das colunas de $\mathbf{X}$ mais correlacionada com $\mathbf{y}$ é aquilo que se fazia no estudo do **Modelo Linear**, sendo $\mathbf{y}$ a variável resposta e $\mathbf{X}$ a matriz das variáveis preditoras. Uma vez que $q = 1$, apenas haverá *um* par de *variáveis de correlação canónica*, que serão $\mathbf{y}$ e a sua projecção

sobre $\mathcal{C}(\mathbf{X})$, $\hat{\mathbf{y}}$. O *coeficiente de correlação canónica* será a correlação entre estas duas variáveis, ou seja, a raíz quadrada do Coeficiente de Determinação do Modelo. Recorde-se que no Modelo Linear, embora a variável $\mathbf{y}$ tenha de ser quantitativa, as colunas da matriz $\mathbf{X}$ podiam ser todas quantitativas (caso da Regressão Linear), todas indicatrizes (Análise de Variância) ou uma mistura destes dois tipos (Análise de Covariância). Qualquer destas situações pode ser encarada, do ponto de vista descritivo, como um caso particular duma Análise em Correlações Canónicas.

2. Uma **Análise Discriminante** também pode ser encarada como um caso particular duma Análise em Correlações Canónicas, em que, para além das variáveis observadas que constituem as colunas da matriz $\mathbf{X}$, a matriz $\mathbf{Y}$ será a matriz da classificação (a matriz $\mathbf{C}$, definida na página 92). Relembre-se que, na Análise Discriminante procuravam-se combinações lineares das colunas de $\mathbf{X}$, não correlacionadas entre si, cujas projecções sobre o subespaço $\mathcal{C}(\mathbf{C})$ gerado pelas colunas da matriz de classificação formassem o menor ângulo possível, isto é fossem o mais correlacionadas possível com alguma combinação linear das variáveis indicatrizes que geram $\mathcal{C}(\mathbf{C})$. Os cossenos ao quadrado desses sucessivos ângulos seriam os sucessivos máximos do critério $\frac{\mathbf{a}^t\mathbf{H}\mathbf{a}}{\mathbf{a}^t\Sigma\mathbf{a}}$ (ver equação 3.10, página 101). Esses cossenos ao quadrado são as **correlações canónicas** entre as variáveis observadas e as variáveis indicatrizes da classificação. As **variáveis de correlação canónica do espaço $\mathcal{C}(\mathbf{X})$** são os **eixos discriminantes**. As **variáveis de correlação canónica do espaço $\mathcal{C}(\mathbf{C})$** são os vectores $\mathbf{P}_C\mathbf{P}_X\mathbf{a} = \mathbf{P}_C\mathbf{y}$ definidos na página 95, ou seja, os vectores das médias de cada grupo no respectivo eixo. Fazendo a ponte com a nota anterior, pode confirmar-se a já referida relação entre uma Análise Discriminante Linear e uma ANOVA a um factor: o primeiro eixo discriminante é a combinação linear das variáveis observadas que maximiza a correlação com o espaço gerado pelas variáveis indicatrizes de nível, logo que torna máxima a estatística F do teste ANOVA correspodente.

6.3 Um exemplo

Consideremos a utilização de uma Análise de Correlações Canónicas numa situação em que existem dois conjuntos distintos de várias variáveis quantitativas.

Uma Análise de Correlações Canónicas pode ser efectuada no R recorrendo ao comando `cancor`.

Um exemplo muito simples, de novo com os dados referentes às medições de $p = 4$ variáveis em $n = 150$ lírios, pode consistir em procurar a combinação linear das duas medições sobre as sépalas (largura e comprimento) mais correlacionada com uma qualquer combinação linear das duas medições das pétalas (largura e comprimento). Trata-se de um exemplo que vale mais para ilustrar o método do que pelo seu valor intrínseco. Com base no conjunto de dados `iris` (já utilizado nos exemplos da Capítulos anteriores), o comando

```
> cancor(iris[,1:2],iris[,3:4])
```

produz os seguintes resultados:

```
$cor
[1] 0.9409690 0.1239369
$xcoef
                    [,1]       [,2]
Sepal.Length -0.08757435 0.04749411
Sepal.Width   0.07004363 0.17582970
$ycoef
                    [,1]       [,2]
Petal.Length -0.06956302 -0.1571867
Petal.Width   0.05683849  0.3940121
$xcenter
Sepal.Length  Sepal.Width
    5.843333     3.057333
$ycenter
Petal.Length  Petal.Width
    3.758000     1.199333
```

Cada coluna das matrizes `xcoef` e `ycoef` contém os coeficientes que definem combinações lineares de cada grupo das variáveis originais. A primeira coluna de cada matriz produz o par de combinações lineares de cada grupo de variáveis que está mais correlacionado, sendo o valor do respectivo coeficiente de correlação dado no primeiro objecto de saída do comando, o objecto `cor`. Assim, no nosso caso, temos que a combinação linear $-0.0876 * Sepal.Length + 0.0700 * Sepal.Width$ e $-0.0696 * Petal.Length + 0.0568 * Petal.Width$ têm uma correlação 0.941. Trata-se da maior correlação possível entre combinações lineares das variáveis associadas às sépalas e combinações lineares das variáveis associadas às pétalas. A variável resultante da combinação linear de medições de sépalas acima referida pode ser obtida pelo comando:

```
> iris.cancor <- cancor(iris[,1:2], iris[,3:4])
> as.matrix(iris[,1:2])%*%iris.cancor$xcoef[,1]
```

(assinale-se a necessidade de converter a `data.frame iris` numa matriz, a fim de se poder efectuar a multiplicação matricial das duas primeiras colunas de `iris` pelo vector dos coeficientes da respectiva combinação linear). A correspondente combinação linear das medições das pétalas é, naturalmente, dada de forma análoga:

```
> as.matrix(iris[,3:4])%*%iris.cancor$ycoef[,1]
```

Uma representação gráfica da relação entre estas duas variáveis é dada na Figura 6.1. É visível a boa correlação linear entre o primeiro par de variáveis de correlação canónica, traduzida no coeficiente de correlação 0.94.

A seguir, indicam-se as correlações entre cada uma das variáveis originais e cada uma das variáveis canónicas agora definidas. Por razões de espaço, os nomes das variáveis foram encurtados para a colagem de "S" (sépalas) ou "P" (pétalas) e "l" (comprimento, ou *length*) e "w" (largura, ou *width*); os nomes das variáveis de correlação canónica que resultam de combinações lineares das medições das sépalas para

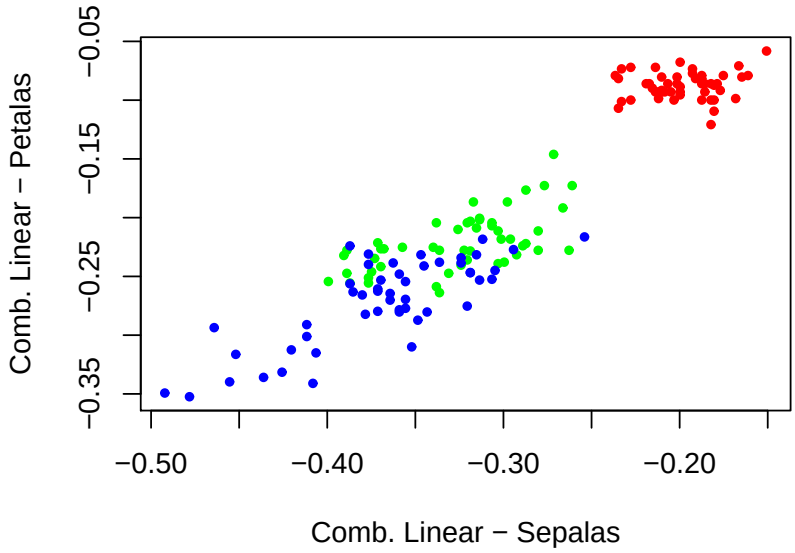


Figura 6.1: Representação gráfica dos $n = 150$ lírios no par de primeiras variáveis de correlação canónicas. As flores de cada variedade foram indicadas em cores diferentes: a azul (os pontos mais escuros, numa impressão a preto e branco) as *Virginica*; a verde (os pontos mais claros) as *Versicolor*, a vermelho (os pontos de tonalidade cinzenta intermédia) as *Setosa*.

"vcS"; e os das respectivas parceiras, que resultam das combinações lineares das medições de pétalas, para "vcP".

```
         Sl     Sw     Pl     Pw   vcS1  vcS2   vcP1  vcP2
Sl    1.000 -0.118  0.872  0.818 -0.929 0.370 -0.874 0.046
Sw   -0.118  1.000 -0.428 -0.366  0.477 0.879  0.449 0.109
Pl    0.872 -0.428  1.000  0.963 -0.931 0.018 -0.990 0.143
Pw    0.818 -0.366  0.963  1.000 -0.860 0.050 -0.914 0.405
vcS1 -0.929  0.477 -0.931 -0.860  1.000 0.000  0.941 0.000
vcS2  0.370  0.879  0.018  0.050  0.000 1.000  0.000 0.124
vcP1 -0.874  0.449 -0.990 -0.914  0.941 0.000  1.000 0.000
vcP2  0.046  0.109  0.143  0.405  0.000 0.124  0.000 1.000
```

O bloco (de dimensão 4×4) do canto superior esquerdo é a matriz de correlações entre as quatro variáveis observadas. O bloco (de dimensão 4×4) do canto inferior direito confirma que as sucessivas variáveis de correlação canónica, de cada grupo, são não correlacionadas entre si, enquanto que as correlações entre cada par de variáveis canónicas são, respectivamente, 0.941 e 0.124. Os restantes blocos da matriz acima indicada contêm as correlações entre variáveis originais e variáveis canónicas (o bloco do canto

superior direito sendo a transposta do bloco do canto inferior esquerdo). Assim, é possível constatar que a primeira variável canónica resultante das medições de sépalas está bastante correlacionada com o comprimento das sépalas, e a segunda variável canónica com a respectiva largura, enquanto que a primeira variável canónica resultante das medições das pétalas está fortemente correlacionada com as duas medições das pétalas (que já estavam fortemente correlacionadas entre si), sendo a segunda variável canónica das pétalas necessariamente mal correlacionada com ambas as medições originais das pétalas.

6.4 Exercícios

1. Em relação aos dados de medições morfométricas sobre lavagantes, já considerados em Capítulos anteriores:

   (a) Efectue uma análise de correlações canónicas entre, por um lado, as três medições relativas à carapaça, e por outro, as duas medições relativas à cauda. Comente os valores obtidos para as correlações canónicas.

   (b) Repita a alínea anterior, mas partindo dos dados *normalizados* dos lavagantes. Compare os resultados com a alínea anterior e comente.

   (c) Efectue uma análise de correlações canónicas entre, por um lado, as três medições relativas à carapaça, e por outro, as duas medições relativas à tenaz, mais o comprimento do dedo da tenaz (*propodus_w*, *propodus_l* e *dactyl_l*). Comente os valores obtidos para as correlações canónicas. Porque é que o *número* de correlações canónicas aqui indicado é diferente do indicado na alínea anterior (independentemente do seu valor)?

   (d) Uma análise em correlações canónicas, efectuada no programa `R`, para comparar as três variáveis relativas à carapaça com as duas variáveis relativas ao *rostrum* produziu os seguintes resultados:

```
 > cancor(lavagantes[,c(1,3,4)],lavagantes[,c(8,9)])
$cor
[1] 0.88961798 0.09519876
$xcoef
                  [,1]        [,2]       [,3]
carapace_l -0.04056729  0.07813553  0.2073225
carapace_w -0.05462914 -0.29617100 -0.2081173
carapace_d -0.01489208  0.25194699 -0.1669188
$ycoef
                 [,1]       [,2]
rostrum_l -0.09576172  0.4080075
rostrum_w -0.27565079 -0.3298531
$xcenter
carapace_l carapace_w carapace_d
  32.53079   16.67540   13.67016
$ycenter
rostrum_l rostrum_w
 6.507460  7.286349
```

   Comente a grande diferença entre as duas correlações canónicas obtidas, procurando os significados algébrico e biológico dessa diferença.

   (e) Uma análise em correlações canónicas, efectuada no programa `R`, para comparar as três variáveis relativas à carapaça com a variável largura do *rostrum* produziu os seguintes resultados:

```
> cancor(lavagantes[,c(1,3,4)],lavagantes[,c(9)])
$cor
[1] 0.866356
$xcoef
                  [,1]          [,2]        [,3]
carapace_l 0.038592666 -0.2164290816  0.04901566
carapace_w 0.062048219  0.3365496391  0.12999529
carapace_d 0.008561461 -0.0008288925 -0.30246809
$ycoef
          [,1]
[1,] 0.3437287
$xcenter
carapace_l carapace_w carapace_d
  32.53079   16.67540   13.67016
$ycenter
[1] 7.286349
```

Efectue uma Regressão Linear da variável largura do *rostrum* sobre as três variáveis de medições sobre a carapaça.

i. Indique qual a relação entre o coeficiente de correlação canónica acima indicado e o coeficiente de determinação (R^2) da regressão linear agora ajustada.

ii. Divida os coeficientes do (único) eixo de correlação canónica pelos correspondentes coeficientes da superfície ajustada pela regressão linear. Comente.

2. Considere uma Análise em Correlações Canónicas entre dois grupos de variáveis: um grupo X constituído por p variáveis, e um grupo Y constituído por q variáveis. Sejam $(\eta_i, \nu_i)_{i=1}^{\min(p,q)}$ os pares de eixos canónicos de cada grupo de variáveis. Mostre que $cor(\eta_i, \nu_j) = 0$ se $i \neq j$.

3. Considere os dados relativos a framboesas, já discutidos nos Exercícios dos Capítulos reltivos à Análise em Componentes Principais e à Análise Discriminante Linear (e disponíveis na *data frame* de nome `framb`). Das dez variáveis, as sete primeiras dizem respeito a características físicas e químicas gerais, enquanto que as três últimas reflectem características cromáticas dos frutos. Efectue uma Análise em Correlações Canónicas, considerando estes dois grupos de variáveis.

   (a) Discuta o significado algébrico de apenas existirem três correlações canónicas.

   (b) Discuta o significado algébrico de existirem duas correlações canónicas muito elevadas. Indique uma possível causa para este facto.

   (c) Calcule os coeficientes de correlação entre o conjunto das 16 variáveis envolvidas na discussão: as dez variáveis originais e os três eixos canónicos associados a cada um do dois grupos de variáveis. Procure interpretações para os eixos canónicos de cada grupo de variáveis, com base nessas correlações.

Capítulo 7

Análise de Correspondências

Um outro conjunto de técnicas de análise de dados tem por objectivo estudar **tabelas de dupla entrada**, isto é, quadros a duas dimensões, de valores não-negativos, em que cada dimensão está associada aos níveis de um factor (critério de classificação).

Um exemplo frequente deste tipo de dados em aplicações biológicas é o de tabelas de frequências de observações, em determinado conjunto de *locais* (um dos factores), de um dado conjunto de *espécies* (o outro factor). Um dos principais objectivos de interesse no estudo deste tipo de dados será o de analisar eventuais preferências (ou aversões) de determinadas espécies por alguns dos locais analisados. De utilidade será também uma visualização gráfica, de baixa dimensão ($m = 2$ ou $m = 3$) com marcadores de linhas (locais) e colunas (espécies) que reflicta a informação sobre as associações sugerideas na tabela. Em particular, será de interesse analisar eventuais desvios à hipótese de *independência* na distribuição das observações pelas várias combinações de níveis de um e outro factor.

Existem várias técnicas aparentadas com este objectivo genérico. Entre elas, a mais famosa é a Análise Factorial de Correspondências, da escola francesa de análise de dados multivariados[1]. Mas técnicas aparentadas encontram-se também sob as designações de *Reciprocal averaging*, *dual scaling* ou apenas *correspondence analysis*. As diferenças entre estas variantes dizem respeito à forma de introdução dos conceitos básicos e à forma de apresentação dos resultados.

[1]Para uma abordagem muito completa do método sob esta perspectiva, pode consultar a obra de M. Greenacre referida na bibliografia desta disciplina: *Theory and Application of Correspondance Analysis*, da Academic Press (1984). No `R` existe um módulo bastante completo de análises multivariadas na perspectiva da escola francesa, designado `ade4`.

7.1 Tabelas de contingência e outras tabelas de dupla entrada

A matéria prima deste conjunto de técnicas são, como se disse, matrizes de dupla entrada[2], cujos valores são não-negativos.

Quando os valores da tabela são **frequências** de observação de cada uma das possíveis combinações de níveis dos dois factores de classificação, os quadros designam-se **tabela de contingências**.

Exemplo 7.1 *Como se assinalou mais acima, um exemplo frequente de tabelas de contingência em contextos biológicos diz respeito a tabelas cujas margens correspondem a diferentes locais (sites em inglês) e espécies (species). Um exemplo duma tal tabela encontra-se no módulo* `MASS` *do* `R`*, numa data frame de nome* `waders`*. trata-se duma tabela de dimensão* 15×19*, onde as linhas correspondem a* 15 *locais na costa ocidental de África, as colunas correspondem a* 19 *espécies de aves limícolas (para mais pormenores, consultar* `help(waders)`*, após ter carregado o módelo). Eis a tabela:*

```
> library(MASS)
> waders
   S1   S2   S3  S4   S5   S6   S7  S8  S9 S10 S11  S12  S13   S14  S15   S16  S17  S18  S19
A  12 2027    0   0 2070   39  219 153   0  15  51 8336 2031 14941   19  3566    0    5    0
B  99 2112    9  87 3481  470 2063  28  17 145  31 1515 1917 17321 3378 20164  177 1759   53
C 197  160    0   4  126   17    1  32   0   2   9  477    1   548   13   273    0    0    0
D   0   17    0   3   50    6    4   7   0   1   2   16    0     0    3    69    1    0    0
E  77 1948    0  19  310    1    1  64   0  22  81 2792  221  7422   10  4519   12    0    0
F  19  203   48  45   20  433    0   0  11 167  12    1    0    26 1790  2916  473  658   55
G 1023 2655   0  18  320   49    8 121   9  82  48 3411   14  9101   43  3230  587   10    5
H  87  745 1447 125 4330  789  228 529 289 904  34 1710 7869  2247 4558 40880 7166 1632  498
I 788 2174    0  19  224  178    1 423   0 195 162 2161   25  1784    3  1254    0    0    0
J  82  350  760 197  858  962   10 511 251 987 191   34   87   417 4496 15835 5327 1312 1020
K 474  930    0  10  316  161    0  90   0  39  48 1183  166  4626   65   127    4    0    0
L  77  249  160 136  999  645   15 851 101 723 266  495   83  1253 1864  4107 1939  623  527
M  22  144    0   4    1    1    0  10   0   2   9  125    5   411    0     3    0    0    0
N   0  791    0   0    4   38    1  56   1  30  54   95    0  1726    0     0    0    0    0
O   0  360  128  43  364 1628   63 287 328 641 850   83   67    48 6499  9094 5647 1333  582
```

Este tipo de tabela pode ser útil no estudo de eventuais preferências, ou rejeições, de determinados locais, por parte de algumas espécies.

Na tabela acima referida, os valores das somas de linha e de coluna não estão predeterminados.

Uma variante da tabela acima indicada consistiria em substituir as frequências absolutas pelas frequências relativas (relativas ao total de observações). Assim, se $\mathbf{T}$ indicasse a tabela de frequências absolutas, e

[2]Não serão consideradas nesta disciplina generalizações da Análise de Correspondências para situações onde se têm tabels de três ou mais dimensões.

n o número total de observações associado à tabela, a tabela $\mathbf{F} = \mathbf{T}/n$ fornece as frequências relativas de cada combinação (i, j) de níveis dos factores A e B. Este tipo de tabela designa-se uma **tabela de correspondências**. A soma dos elementos duma matriz $\mathbf{F}$ de correspondências é igual a 1, e a tabela pode ser vista como uma **estimativa de distribuição de probabilidades bivariada**.

Outra variante da tabela acima indicada seria dada no caso de se assinalar, não a frequência de observações, mas apenas a **presença ou ausência** das espécies nos locais. Nesse caso, os elementos não-nulos da tabela seriam todos iguais a 1 (indicando a presença). Neste caso podemos falar numa **matriz de incidências** ou **matriz indicatriz**.

O conceito de tabela de contingências (ou de matriz de incidências) não se restringe ao caso de tabelas de locais × espécies.

Exemplo 7.2 *Ainda no módelo* `MASS` *encontra-se uma outra tabela de contingências, na data frame* `caith`. *Trata-se da classificação de* 5387 *habitantes de Caithness, na Escócia, de acordo com dois factores: côr dos olhos (com* 4 *níveis: blue, light, medium, dark) e côr do cabelo (com* 5 *níveis: fair, red, medium, dark, black). A tabela é a seguinte:*

```
> caith
       fair red medium dark black
blue    326  38    241  110     3
light   688 116    584  188     4
medium  343  84    909  412    26
dark     98  48    403  681    85
```

Mais uma vez, uma tabela deste tipo sugere a procura de associações preferenciais (ou raras) entre côr de olhos e côr de cabelo. Análises de Correspondências visam dar uma resposta a esta questão, acompanhando com frequência os resultados duma representação visual a baixa dimensão que ajude na compreensão das relações.

7.2 Alguns conceitos e notação

Como foi indicado acima, o ponto de partida para uma Análise de Correspondências é uma tabela de dupla entrada, que admitiremos na discussão ser uma tabela de contingências.

Designemos os dois factores de classificação por A, com a níveis, e B, com b níveis. Admitimos que o factor A está associado às *linhas* da tabela, e o factor B às suas *colunas*.

A **tabela de contingências** será, assim, da forma

$$\mathbf{T} \quad = \quad \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1j} & \cdots & n_{1,b-1} & n_{1,b} \\ n_{21} & n_{22} & \cdots & n_{2j} & \cdots & n_{2,b-1} & n_{2,b} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ n_{i1} & n_{i2} & \cdots & n_{ij} & \cdots & n_{i,b-1} & n_{i,b} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ n_{a-1,1} & n_{a-1,2} & \cdots & n_{a-1,j} & \cdots & n_{a-1,b-1} & n_{a-1,b} \\ n_{a,1} & n_{a,2} & \cdots & n_{a,j} & \cdots & n_{a,b-1} & n_{a,b} \end{bmatrix} \tag{7.1}$$

cujo (i, j)-ésimo elemento n_{ij} indica o número de observações (frequência absoluta) efectuadas na combinação do nível i do factor A com o nível j do factor B.

A soma das frequências na linha i da tabela de contingência $\mathbf{T}$,

$$n_{i.} \quad = \quad \sum_{j=1}^{b} n_{ij} \qquad (i = 1 : a) \ ,$$

indica o número total de observações (frequência absoluta) associado ao nível i do primeiro factor de classificação. Assim, caso se trate duma matriz em que as linhas correspondem a locais e as colunas a espécies, **a soma de cada linha indica a frequência absoluta de observações em cada local**. Como sabemos, o vector das somas de linhas é dado pelo produto matricial $\mathbf{T}\mathbf{1}_b$.

Analogamente, a soma das frequências na coluna j da tabela,

$$n_{.j} \quad = \quad \sum_{i=1}^{a} n_{ij} \qquad (j = 1 : b) \ ,$$

indica o número total de observações associado a esse nível do segundo factor de classificação (o número de observações por espécie, no caso duma tabela locais $\times$ espécies). Sabemos que o vector-linha das somas de coluna é dado pelo produto matricial $\mathbf{1}_a^t\mathbf{T}$.

O número total de observações (em qualquer combinação de níveis dos dois factores) é dado por:

$$n_{..} \quad = \quad \sum_{i=1}^{a} n_{i.} \quad = \quad \sum_{j=1}^{b} n_{.j} \quad = \quad \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}$$

Este valor é obtido matricialmente pelo produto

$$n_{..} = \mathbf{1}_a^t\mathbf{T}\mathbf{1}_b \ .$$

A *frequência relativa* da linha i da tabela (nível i do factor A, independentemente de quais os níveis do factor B correspondentes) é dada por:

$$r_i \quad = \quad \frac{n_{i.}}{n_{..}} \qquad (i = 1 : a) \ . \tag{7.2}$$

O vector **r** destas frequências relativas de linha[3] indica a proporção de observações em cada nível do primeiro factor (cada linha). Assim, **numa tabela do tipo locais × espécies, o vector r fornece a proporção de observações em cada um dos a locais**. Do ponto de vista matricial, o vector das frequências relativas de linhas calcula-se como:

$$\mathbf{r} \quad = \quad \frac{\mathbf{T}\mathbf{1}_b}{\mathbf{1}_a^t \mathbf{T}\mathbf{1}_b} \quad = \quad \frac{1}{n_{..}} \mathbf{T}\mathbf{1}_b \ .$$

De forma análoga, a *frequência relativa* da coluna j da tabela é dada por:

$$c_j \quad = \quad \frac{n_{.j}}{n_{..}} \qquad (j = 1 : b) \ , \tag{7.3}$$

e define-se o vector **c** cujos b elementos são as frequências relativas associadas a cada coluna. Numa tabela do tipo locais × espécies, o vector **c** fornece a proporção de observações de cada uma das b espécies. Do ponto de vista matricial, o vector (coluna) das frequências relativas de colunas calcula-se como:

$$\mathbf{c} \quad = \quad \frac{\mathbf{T}^t\mathbf{1}_a}{\mathbf{1}_a^t \mathbf{T}\mathbf{1}_b} \quad = \quad \frac{1}{n_{..}} \mathbf{T}^t\mathbf{1}_a.$$

A **tabela das correspondências**, ou tabela das frequências relativas (relativas ao número total de observações) é dado por

$$\mathbf{F} \quad = \quad \frac{\mathbf{T}}{n_{..}} \quad = \quad \begin{bmatrix} \frac{n_{11}}{n_{..}} & \frac{n_{12}}{n_{..}} & \dots & \frac{n_{1,j}}{n_{..}} & \dots & \frac{n_{1,b-1}}{n_{..}} & \frac{n_{1,b}}{n_{..}} \\ \frac{n_{21}}{n_{..}} & \frac{n_{22}}{n_{..}} & \dots & \frac{n_{2,j}}{n_{..}} & \dots & \frac{n_{2,b-1}}{n_{..}} & \frac{n_{2,b}}{n_{..}} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ \frac{n_{i,1}}{n_{..}} & \frac{n_{i,2}}{n_{..}} & \dots & \frac{n_{i,j}}{n_{..}} & \dots & \frac{n_{i,b-1}}{n_{..}} & \frac{n_{i,b}}{n_{..}} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ \frac{n_{a-1,1}}{n_{..}} & \frac{n_{a-1,2}}{n_{..}} & \dots & \frac{n_{a-1,j}}{n_{..}} & \dots & \frac{n_{a-1,b-1}}{n_{..}} & \frac{n_{a-1,b}}{n_{..}} \\ \frac{n_{a,1}}{n_{..}} & \frac{n_{a,2}}{n_{..}} & \dots & \frac{n_{a,j}}{n_{..}} & \dots & \frac{n_{a,b-1}}{n_{..}} & \frac{n_{a,b}}{n_{..}} \end{bmatrix} \tag{7.4}$$

Utilizando esta tabela de frequências relativas **F**, obtêm-se fórmulas mais simples para os vectores de frequências relativas de linha e de coluna:

$$\mathbf{r} \ = \ \mathbf{F}\mathbf{1_b} \tag{7.5}$$

e

$$\mathbf{c} \ = \ \mathbf{F}^t\mathbf{1_a} \tag{7.6}$$

Já se viu que, se **T** é uma tabela de contingências, a matriz **F** pode ser vista como uma estimativa da distribuição de probabilidades bivariada associada ao problema sob estudo. De igual forma, **os vectores r e c são estimativas das distribuições de probabilidades marginais**, associadas, respectivamente, ao factor A e ao factor B.

[3] É hábito usar as iniciais das palavras inglesas com que se designam as linhas e as colunas de uma matriz – *rows* e *columns* – para referenciar conceitos associados ao primeiro e segundo factores, respectivamente.

Perfis de linha e perfis de coluna

Por **perfil da linha** i entende-se o conjunto das frequências observadas para cada elemento dessa linha, **relativas ao total de observações nessa linha**. Assim, o perfil da linha i é dado pelos b valores:

$$pl_j^{(i)} \quad = \quad \frac{n_{ij}}{n_{i.}} \qquad (j = 1 : b) \; . \tag{7.7}$$

No caso de uma tabela do tipo locais × espécies, um perfil de linha corresponderá à distribuição, por espécie, das observações numa dada localidade, ou seja, ao **perfil da localidade**.

Do ponto de vista matricial, a matriz $\mathbf{P_L}$ dos perfis de linha calcula-se através do produto

$$\mathbf{P_L} \quad = \quad \mathbf{D_r^{-1}F} \; , \tag{7.8}$$

onde $\mathbf{D_r^{-1}}$ é a matriz diagonal $(a \times a)$ cuja diagonal é dada pelos recíprocos do vector de frequências relativas de linha, ou seja,

$$\mathbf{D_r^{-1}} \; = \; \begin{bmatrix} \frac{1}{r_1} & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{r_2} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{r_{a-1}} & 0 \\ 0 & 0 & \cdots & 0 & \frac{1}{r_a} \end{bmatrix} \; .$$

De forma análoga, por **perfil da coluna** j entende-se o conjunto das frequências observadas para cada elemento dessa coluna, **relativas ao total de observações nessa coluna**. Assim, o perfil da coluna j é dado pelos a valores:

$$pc_i^{(j)} \quad = \quad \frac{n_{ij}}{n_{.j}} \qquad (i = 1 : a) \; . \tag{7.9}$$

No caso de uma tabela do tipo locais × espécies, um perfil de coluna corresponderá à distribuição das observações de uma dada espécie, por localidade, ou seja, ao **perfil da espécie**.

A matriz $\mathbf{P_C}$ dos perfis de coluna calcula-se através do produto

$$\mathbf{P_C} \quad = \quad \mathbf{FD_c^{-1}} \; , \tag{7.10}$$

onde $\mathbf{D_c^{-1}}$ é a matriz diagonal $(b \times b)$ cuja diagonal é dada pelos recíprocos do vector de frequências relativas de coluna.

Exemplo 7.3 *Consideremos um pequeno exemplo, para assentar ideias. Admita-se que se tem uma tabela das frequências absolutas de observações, do tipo locais × espécies, com $a = 3$ locais e $b = 5$ espécies,*

$$\mathbf{T} \quad = \quad \begin{bmatrix} 12 & 3 & 15 & 10 & 20 \\ 9 & 6 & 12 & 0 & 18 \\ 21 & 24 & 6 & 12 & 32 \end{bmatrix} \; .$$

A matriz de correspondências associada é dada por

$$\mathbf{F} \quad = \quad \frac{\mathbf{T}}{200} \quad = \quad \begin{bmatrix} 0.060 & 0.015 & 0.075 & 0.050 & 0.100 \\ 0.045 & 0.030 & 0.060 & 0.000 & 0.090 \\ 0.105 & 0.120 & 0.030 & 0.060 & 0.160 \end{bmatrix} \; .$$

O vector $\mathbf{r}$ *das frequências relativas de linhas é dado por*

$$\mathbf{r} \quad = \quad (0.300, 0.225, 0.475) \ .$$

O vector $\mathbf{c}$ *das frequências relativas de colunas é dado por*

$$\mathbf{c} \quad = \quad (0.210, 0.165, 0.165, 0.110, 0.350) \ .$$

O perfil da primeira linha é, tendo em conta que $n_{1.} = 60$*, o vector*

$$\mathbf{pl}^{(1)} \quad = \quad \left(\frac{12}{60} \ , \ \frac{3}{60} \ , \ \frac{15}{60} \ , \ \frac{10}{60} \ , \ \frac{20}{60} \right) \quad = \quad (\ 0.200 \ , \ 0.050 \ , \ 0.250 \ , \ 0.167 \ , \ 0.333 \ ) \, .$$

Os restantes perfis de linha são dados pelo produto matricial (tendo em conta que o valor $n_{..} = 200$ *cancela em todos os produtos):*

$$\begin{aligned} \mathbf{P_L} \quad = \quad \mathbf{D_r^{-1}F} \quad &= \quad \begin{bmatrix} \frac{1}{60} & 0 & 0 \\ 0 & \frac{1}{45} & 0 \\ 0 & 0 & \frac{1}{95} \end{bmatrix} \cdot \begin{bmatrix} 12 & 3 & 15 & 10 & 20 \\ 9 & 6 & 12 & 0 & 18 \\ 21 & 24 & 6 & 12 & 32 \end{bmatrix} \\ &= \quad \begin{bmatrix} 0.2000000 & 0.0500000 & 0.2500000 & 0.1666667 & 0.3333333 \\ 0.2000000 & 0.1333333 & 0.2666667 & 0.0000000 & 0.4000000 \\ 0.2210526 & 0.2526316 & 0.0631579 & 0.1263158 & 0.3368421 \end{bmatrix} \ . \end{aligned}$$

A interpretação dos valores nesta matriz de perfis de linha é evidente: na primeira localidade, 20% *das observações correspondem à primeira espécie,* 5% *à segunda espécie,* 25% *à terceira espécie,* $16,7\%$ *à quarta espécie, e o restante terço de observações à quinta e última espécie. Estas proporções variam nos restantes locais.*

Por outro lado, o perfil da primeira coluna é, tendo em conta que $n_{.1} = 42$*, o vector*

$$\mathbf{pc}^{(1)} \quad = \quad \left(\frac{12}{42} \ , \ \frac{9}{42} \ , \ \frac{21}{42} \right) \quad = \quad (\ 0.2857 \ , \ 0.2143 \ , \ 0.5000 \ ) \, .$$

Os restantes perfis de coluna são dados pelo produto matricial

$$\begin{aligned} \mathbf{P_C} \quad = \quad \mathbf{FD_c^{-1}} \quad &= \quad \begin{bmatrix} 12 & 3 & 15 & 10 & 20 \\ 9 & 6 & 12 & 0 & 18 \\ 21 & 24 & 6 & 12 & 32 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{42} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{33} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{33} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{22} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{70} \end{bmatrix} \\ &= \quad \begin{bmatrix} 0.2857143 & 0.0909091 & 0.4545455 & 0.4545455 & 0.2857143 \\ 0.2142857 & 0.1818182 & 0.3636364 & 0.0000000 & 0.2571429 \\ 0.5000000 & 0.7272727 & 0.1818182 & 0.5454545 & 0.4571429 \end{bmatrix} \ . \end{aligned}$$

As proporções devem agora ser lidas por coluna: assim, das observações da primeira espécie, cerca de $28,5\%$ *foram efectuadas no primeiro local, cerca de* $21,4\%$ *no segundo local, e metade foram efectuadas no terceiro local.*

Assinale-se que a soma das colunas da matriz de perfil de linhas é um vector de uns, $\mathbf{1}_3$*, facto que é matricialmente fácil de verificar, tendo em conta que essa soma de linhas é dada pelo produto matricial*

$$\mathbf{P_L 1_5} \;=\; \mathbf{D_r^{-1} F 1_5} \;=\; \mathbf{1}_3 \;.$$

De facto, o produto $\mathbf{F1_5}$ *é igual ao vector da soma de linhas de* $\mathbf{F}$*, cujos recíprocos se encontram nas posições diagonais da matriz* $\mathbf{D_r^{-1}}$*. Analogamente, a soma das linhas da matriz de perfil de colunas é também um vector de uns:*

$$\mathbf{P}_C^t \mathbf{1}_3 \;=\; \mathbf{D_c^{-1} F}^t \mathbf{1}_3 \;=\; \mathbf{1}_5 \;.$$

7.3 A hipótese de independência

No caso de existir independência entre os factores de classificação, a probabilidade p_{ij} de ter uma observação na célula (i,j) da tabela de contingências será dada pelo produto das respectivas probabilidades marginais, ou seja, $p_{ij} = p_{i.} \times p_{.j}$, na notação convencional. Nesse caso, ter-se-á como valor esperado para o número de observações que recaem na célula (i,j) (de entre um total de $n_{..}$ observações), o produto

$$E_{ij} \quad = \quad n_{..} \times p_{i.} \times p_{.j} \;.$$

Ora, as probabilidades marginais podem ser estimadas pelas frequências relativas marginais da tabela, ou seja, $\hat{p}_{i.} = \frac{n_{i.}}{n_{..}} = r_i$ e $\hat{p}_{.j} = \frac{n_{.j}}{n_{..}} = c_j$. Logo, o valor esperado estimado será, dada a hipótese de independência,

$$\hat{E}_{ij} \quad = \quad n_{..} \times \hat{p}_{i.} \times \hat{p}_{.j} \;=\; n_{..} \frac{n_{i.}}{n_{..}} \times \frac{n_{.j}}{n_{..}} = n_{..} r_i c_j \;.$$

Assim, a hipótese de independência daria à tabela de contingências uma estrutura simples, aproximada por:

$$\begin{aligned} & \mathbf{T} \;=\; n_{..} \mathbf{r c}^t \\ \Leftrightarrow \quad & \mathbf{F} \;=\; \mathbf{r c}^t \;. \end{aligned}$$

A matriz dada pela diferença

$$\mathbf{F} \;-\; \mathbf{r c}^t$$

fornece informação sobre desvios à independência na tabela de contingências. Quanto mais próximo de zero estiver a generalidade dos valores dessa matriz, mais plausível será a hipótese de independência entre os factores de classificação. Assim, *a norma da matriz* $\|\mathbf{F} - \mathbf{r c}^t\|$ *pode ser vista como um índice global de desvio à hipótese de independência*. O quadrado dessa norma, como sabemos da Definição (1.4), na página 19, é dada pela soma dos quadrados dos seus elementos, isto é, por

$$\|\mathbf{F} - \mathbf{r c}^t\|^2 \quad = \quad \sum_{i=1}^{r} \sum_{j=1}^{c} \left(\frac{n_{ij}}{n_{..}} - \frac{n_{i.}}{n_{..}} \frac{n_{.j}}{n_{..}} \right)^2 \quad = \quad \sum_{i=1}^{r} \sum_{j=1}^{c} \left(\frac{O_{ij} - \hat{E}_{ij}}{n_{..}} \right)^2 \;, \tag{7.11}$$

usando a habitual notação dos testes χ^2 de independência, onde $O_{ij} = n_{ij}$ designa o número de observações na célula (i,j) e $\hat{E}_{ij} = n_{..} \frac{n_{i.}}{n_{..}} \frac{n_{.j}}{n_{..}} = \frac{n_{i.} n_{.j}}{n_{..}}$ o número esperado estimado de observações nessa mesma

célula. *A semelhança da expressão (7.11) com a estatística do teste χ^2 à independência duma tabela de contingências sugere que, em vez de se analisar a matriz $\mathbf{F}-\mathbf{r}\mathbf{c}^t$, se analise antes a matriz que se obtém dividindo cada uma das suas linhas pela raíz quadrada do produto das frequências relativas de linha, e cada coluna pela raíz quadrada do produto das frequências relativas de coluna.* De facto, seja $\mathbf{D}_r^{-\frac{1}{2}}$ a matriz diagonal dada por

$$\mathbf{D}_r^{-\frac{1}{2}} = \begin{bmatrix} \frac{1}{\sqrt{r_1}} & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{\sqrt{r_2}} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{r_{a-1}}} & 0 \\ 0 & 0 & \cdots & 0 & \frac{1}{\sqrt{r_a}} \end{bmatrix}, \tag{7.12}$$

e $\mathbf{D}_c^{-\frac{1}{2}}$ a matriz diagonal dada por

$$\mathbf{D}_c^{-\frac{1}{2}} = \begin{bmatrix} \frac{1}{\sqrt{c_1}} & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{\sqrt{c_2}} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{c_{b-1}}} & 0 \\ 0 & 0 & \cdots & 0 & \frac{1}{\sqrt{c_b}} \end{bmatrix}. \tag{7.13}$$

Então, o elemento genérico da matriz

$$\mathbf{D}_r^{-\frac{1}{2}} \left(\mathbf{F}-\mathbf{r}\mathbf{c}^t\right) \mathbf{D}_c^{-\frac{1}{2}} \tag{7.14}$$

é dado por $\frac{\frac{n_{ij}}{n_{..}}-r_i c_j}{\sqrt{r_i \cdot c_j}}$, e a soma dos quadrados dos elementos dessa matriz será assim dada por

$$\|\mathbf{D}_r^{-\frac{1}{2}} \left(\mathbf{F}-\mathbf{r}\mathbf{c}^t\right) \mathbf{D}_c^{-\frac{1}{2}}\|^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{\left(\frac{n_{ij}}{n_{..}} - \frac{n_{i.}}{n_{..}}\frac{n_{.j}}{n_{..}}\right)^2}{\frac{n_{i.}}{n_{..}}\frac{n_{.j}}{n_{..}}} = \frac{1}{n_{..}} \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{\left(O_{ij}-\hat{E}_{ij}\right)^2}{\hat{E}_{ij}}. \tag{7.15}$$

Por outras palavras, **o quadrado da norma da matriz (7.14) é a estatística do teste χ^2 à independência dos factores, a dividir pelo número total de observações**. Valores "grandes" desta norma ao quadrado indiciam violação da hipótese de independência, pelo que a dimensão da generalidade dos valores da matriz $\mathbf{D}_r^{-\frac{1}{2}} \left(\mathbf{F}-\mathbf{r}\mathbf{c}^t\right) \mathbf{D}_c^{-\frac{1}{2}}$ é indiciadora de falta de independência. Mais, cada parcela da estatística do teste χ^2 está associada a um elemento da matriz (7.14), pelo que será possível analisar nos elementos dessa matriz quais as combinações de níveis de um e outro factor que mais contribuem para o valor final da estatística do χ^2 e que, em caso de rejeiçao da hipótese de independência seriam as mais "responsáveis" por essa ausência de independência.

O estudo da matriz $\mathbf{D}_r^{-\frac{1}{2}} \left(\mathbf{F}-\mathbf{r}\mathbf{c}^t\right) \mathbf{D}_c^{-\frac{1}{2}}$ será assim de grande interesse na consideração da hipótese de independência entre os factores de classificação. Nesta disciplina, esta matriz será designada a **matriz normalizada dos desvios à independência**.

7.4 As nuvens de perfis

Cada perfil de linha, ou seja, cada linha da matriz de perfis de linha, $\mathbf{P_L}$, define um ponto no espaço a b dimensões, $\mathbb{R}^b$. A nuvem de a pontos em $\mathbb{R}^b$ assim resultante pode ser designada a **nuvem de perfis de linha**, e representada por **N(A)**. Infelizmente, e a menos que $b \leq 3$ desta nuvem não é visualizável.

Uma representação análoga dos perfis de coluna (colunas de $\mathbf{P_C}$) é possível no espaço $\mathbb{R}^a$, onde "vivem" as colunas da matriz. A nuvem de b pontos em $\mathbb{R}^a$ assim resultante pode ser designada a **nuvem de perfis de coluna**, e representada por **N(B)**. Também aqui, e a menos que $a \leq 3$, esta nuvem não é visualizável.

Embora estas duas nuvens de pontos vivam em espaços diferentes ($N(A) \subset \mathbb{R}^b$ e $N(B) \subset \mathbb{R}^a$), há **várias relações entre estas duas nuvens de pontos**.

Uma **primeira relação** diz respeito à dimensão dos subespaços onde cada nuvem está efectivamente contida. A dimensão do subespaço de $\mathbb{R}^b$ onde está contida a nuvem dos perfis de linha, $N(A)$, é dada pela característica da matriz $\mathbf{P_L}$. Mas

$$car(\mathbf{P_L}) \;=\; car(\mathbf{D_r^{-1}F}) \;\leq\; \min\left\{car(\mathbf{D_r^{-1}}), car(\mathbf{F})\right\} \;\leq\; \min\{a, \min(a,b))\;=\;\min(a,b)\;.$$

Em geral, quando não há dependências lineares nas linhas da tabela $\mathbf{F}$, esta característica é precisamente a menor das dimensões da tabela. De forma análoga, a nuvem de perfis de coluna (colunas de $\mathbf{P_C}$) está num subespaço de dimensão

$$car(\mathbf{P_C}) \;=\; car(\mathbf{FD_c^{-1}}) \;\leq\; \min\left\{car(\mathbf{D_c^{-1}}), car(\mathbf{F})\right\} \;\leq\; \min\{b, \min(a,b))\;=\;\min(a,b)\;.$$

Também aqui, se não há dependências lineares nas colunas da tabela $\mathbf{F}$, esta característica é precisamente $\min(a,b)$, e portanto ambas as nuvens vivem em subespaços de igual dimensão (embora em espaços diferentes).

Uma **segunda relação diz respeito aos centros de gravidade (ponderados pela frequência de cada ponto) das duas nuvens de pontos**. De facto,

- a **média ponderada das coordenadas dos a perfis de linha** (sendo as ponderações dadas pelas frequências relativas de linha, r_i) é dada pelo vector

$$(\mathbf{rP_L})^t \;=\; \mathbf{P_L}^t\mathbf{r} \;=\; \mathbf{F}^t\mathbf{D_r^{-1}r} \;=\; \mathbf{F}^t\mathbf{1}_a \;=\; \mathbf{c}\;, \tag{7.16}$$

  tendo em conta as equações (7.8) e (7.6).

- a **média ponderada das coordenadas dos b perfis de coluna** (sendo as ponderações dadas pelas frequências relativas de coluna, c_j) é dada pelo vector

$$\mathbf{P_C c} \;=\; \mathbf{FD_c^{-1}c} \;=\; \mathbf{F1}_b \;=\; \mathbf{r}\;, \tag{7.17}$$

  tendo em conta as equações (7.10) e (7.5).

Assim, a nuvem $N(A)$ de perfis de linha (que são pontos em $\mathbb{R}^b$) tem centro de gravidade (ponderado) dada pelo vector das frequências relativas das colunas (igualmente um ponto $\mathbf{c} \in \mathbb{R}^b$). Por seu lado, a a nuvem $N(B)$ de perfis de coluna (que são pontos em $\mathbb{R}^a$) tem centro de gravidade (ponderado) dada pelo vector das frequências relativas das linhas (igualmente um ponto $\mathbf{r} \in \mathbb{R}^a$).

Façamos agora a **centragem** de cada nuvem de pontos, ou seja, desloquemos a origem de $\mathbb{R}^b$ para o centro de gravidade da nuvem $N(A)$ e a origem de $\mathbb{R}^a$ para o centro de gravidade da nuvem $N(B)$.

A **matriz $a \times b$ dos perfis centrados de linha** corresponde à matriz que se obtem subtraíndo de cada linha de $\mathbf{P_L}$ o vector $\mathbf{c}$. Em notação matricial, serão as linhas da matriz

$$\mathbf{P_L} \; - \; \mathbf{1}_a \mathbf{c}^t \; . \tag{7.18}$$

De forma análoga, a **matriz $a \times b$ dos perfis centrados de coluna** corresponde à matriz que se obtem subtraíndo de cada coluna de $\mathbf{P_C}$ o vector $\mathbf{r}$. Em notação matricial, serão as colunas da matriz

$$\mathbf{P_C} \; - \; \mathbf{r}\mathbf{1}_b^t \; . \tag{7.19}$$

Uma forma de medir a dispersão de cada nuvem de pontos é medir a sua **inércia**, ou seja, a soma de quadrados das distâncias de cada ponto, perfil de linha, em relação ao seu centro de gravidade. Caso trabalhassemos com a inércia simples (sem ponderações), a inércia da nuvem de perfis de linha seria a soma de quadrados de todos os elementos da matriz (7.18), ou seja, a norma (usual) ao quadrado dessa matriz:

$$\sum_{i=1}^{a} \sum_{j=1}^{b} (pl_j^{(i)} - c_j)^2 \; .$$

De forma análoga, a inércia simples da nuvem de perfis de coluna seria a soma de quadrados dos elementos da matriz (7.19), isto é a norma (usual) ao quadrado dessa matriz.

Mas há argumentos que justificam utilizar inércias ponderadas, com métricas diferentes da euclidiana usual. De facto, o peso relativo de cada linha (ou coluna) não é igual. Assim, faz sentido que no cálculo da inércia dos perfis de linha se dê ponderações proporcionais à frequência relativa de cada linha, ou seja, a contribuição de cada linha seja multiplicada por r_i. Se se optar também por ponderar cada parcela de forma *inversa* à frequência relativa da coluna a que diz respeito, teremos a seguinte expressão para a **inércia (ponderada) da nuvem de perfis de linha, $N(A)$**:

$$\sum_{i=1}^{a} \sum_{j=1}^{b} r_i \, \frac{(pl_j^{(i)} - c_j)^2}{c_j} \; . \tag{7.20}$$

Em notação matricial, trata-se da soma de quadrados (ou seja, norma usual ao quadrado) da matriz:

$$\mathbf{D_r}^{1/2} \left(\mathbf{P_L} - \mathbf{1}_a \mathbf{c}^t \right) \mathbf{D_c}^{-1/2} \; = \; \mathbf{D_r}^{1/2} \left(\mathbf{D_r}^{-1} \mathbf{F} - \mathbf{D_r}^{-1} \mathbf{r}\mathbf{c}^t \right) \mathbf{D_c}^{-1/2} \; = \; \mathbf{D_r}^{-1/2} \left(\mathbf{F} - \mathbf{r}\mathbf{c}^t \right) \mathbf{D_c}^{-1/2} \; , \tag{7.21}$$

que não é mais do que a matriz dos desvios normalizados à hipótese de independência (equação 7.14). Assim, **a inércia (ponderada) da nuvem $N(A)$ dos perfis de linha tem de ser $\frac{1}{n_{\cdot\cdot}} \chi^2$**, onde χ^2 indica o valor da estatística no teste de igual nome á independência dos factores.

Um raciocínio análogo conduz à conclusão que a **inércia (ponderada) da nuvem de perfis de coluna,** $N(B)$:

$$\sum_{i=1}^{a}\sum_{j=1}^{b} c_j \frac{(pc_i^{(j)} - r_i)^2}{r_i} \ . \tag{7.22}$$

Em notação matricial, trata-se da soma de quadrados (ou seja, norma usual ao quadrado) da matriz:

$$\mathbf{D_r}^{-1/2}\left(\mathbf{P_C} - \mathbf{r}\mathbf{1}_b^t\right)\mathbf{D_c}^{1/2} \ = \ \mathbf{D_r}^{-1/2}\left(\mathbf{F}\mathbf{D_c}^{-1} - \mathbf{r}\mathbf{c}^t\mathbf{D_c}^{-1}\right)\mathbf{D_c}^{1/2} \ = \ \mathbf{D_r}^{-1/2}\left(\mathbf{F} - \mathbf{r}\mathbf{c}^t\right)\mathbf{D_c}^{-1/2} \ , \tag{7.23}$$

que é, de novo, a matriz dos desvios normalizados à hipótese de independência (equação 7.14). Assim, **a inércia (ponderada) da nuvem $N(B)$ dos perfis de coluna é igual à inércia (ponderada) da nuvem $N(A)$:** $\frac{1}{n_{\cdot\cdot}}\chi^2$, onde χ^2 indica o valor da estatística no teste de igual nome á independência dos factores. Vimos pois uma **terceira relação** importante entre as duas nuvens de perfis: partilham a mesma inércia ponderada.

Permanece o problema da visualização das nuvens $N(A)$ e $N(B)$. tratando-se de nuvens de pontos em subespaços (diferentes) de dimensão menor ou igual a $\min(a, b)$, põe-se o problema de obter uma aproximação a baixa dimensão. Mas, uma vez que os perfis normalizados são, em ambos os casos, dados pela matriz (7.14) dos desvios normalizados à independência. O problema da representação a baixa dimensão será agora discutido.

7.5 A análise factorial da matriz normalizada dos desvios

A fim de se poder proceder a representações gráfica das correspondências entre níveis de um e outro factor que possam ser evidenciadas pela matriz $\mathbf{D}_r^{-\frac{1}{2}}\left(\mathbf{F} - \mathbf{r}\mathbf{c}^t\right)\mathbf{D}_c^{-\frac{1}{2}}$, convém efectuar uma aproximação de baixa dimensionalidade a essa matriz[4].

Tratando-se de uma matriz que, em geral, nem sequer é quadrada[5], recorre-se à **Decomposição em Valores Singulares da matriz normalizada dos desvios à independência**. Recorde-se (matéria da Secção 1.5, a partir da página 40) que a DVS garante que é possível escrever qualquer matriz $\mathbf{X}_{a\times b}$ de característica k na forma

$$\mathbf{X} \ = \ \mathbf{W}\Delta\mathbf{V}^t \ = \ \sum_{i=1}^{k}\delta_i\mathbf{w}_i\mathbf{v}_i^t \ ,$$

sendo os vectores $\mathbf{w}$ (as colunas da matriz $\mathbf{W}$) um conjunto ortonormado de vectores em $\mathbb{R}^a$, os vectores $\mathbf{v}$ (as colunas da matriz $\mathbf{V}$) um conjunto ortonormado de vectores em $\mathbb{R}^b$, e os δs (os elementos diagonais da matriz diagonal Δ) um conjunto de escalares não negativos, sendo ainda o escalar k a característica da matriz $\mathbf{X}$. Estas quantidades são designadas vectores singulares (respectivamente esquerdos e direitos) e valores singulares da matriz $\mathbf{X}$.

[4] Ou seja, convém efectuar aquilo a que os franceses designam a *analyse factorielle* da matriz das correspondências.

[5] Apenas por coincidência haverá igualdade entre o número de níveis, a e b, dos dois factores.

Como foi visto na Secção relativa à DVS, a melhor aproximação à matriz $\mathbf{X}$ de cardinalidade $m < k$ é dada retendo as colunas das matrizes $\mathbf{W}$ e $\mathbf{V}$ e linhas/colunas da matriz Δ associadas aos maiores valores singulares. Na forma da DVS dada pelo somatório de matrizes da forma $\delta_i \mathbf{w}_i \mathbf{v}_i^t$, isso significa reter apenas as m parcelas associadas aos m maiores valores singulares:

$$\mathbf{X}_{[m]} \;=\; \mathbf{W}_{[m]} \Delta_{[m]} \mathbf{V}_{[m]}^t \;=\; \sum_{i=1}^{m} \delta_i \mathbf{w}_i \mathbf{v}_i^t \;.$$

Esta aproximação de baixa cardinalidade será agora representada graficamente num espaço de dimensão m, por marcadores de linhas e marcadores de colunas, de tal forma a permitir a visualização das correspondências entre linhas e colunas da matriz $\mathbf{X}$, que no nosso caso é a matriz de correspondências normalizada. A ideia geral aqui utilizada é análoga à ideia que presidiu à técnica do *biplot* já estudada no capítulo de Análise em Componentes Principais (Capítulo 2, pg. 55).

É possível utilizar vários critérios de escolha para marcadores de perfis de linhas e marcadores de perfis de colunas, e muitas das diferenças entre as várias variantes de análises de correspondência têm a que ver com estas diversas opções.

Uma hipótese possível consiste em escolher para marcadores de linha

$$\mathbf{R}_{[m]} \quad = \quad \mathbf{W}_{[m]} \Delta_{[m]}^{1/2} \tag{7.24}$$

e os marcadores de coluna serão dados por

$$\mathbf{C}_{[m]} \quad = \quad \mathbf{V}_{[m]} \Delta_{[m]}^{1/2} \tag{7.25}$$

onde $\mathbf{W}_{[m]}$, $\mathbf{V}_{[m]}$ e $\Delta_{[m]}$ são as matrizes associadas à aproximação de característica m da matriz de correspondências $\mathbf{D}_r^{-\frac{1}{2}} (\mathbf{F} - \mathbf{r}\mathbf{c}^t) \mathbf{D}_c^{-\frac{1}{2}}$. Tem-se que essa matriz aproximada é dada por $\mathbf{R}_{[m]} \mathbf{C}_{[m]}^t$, ou seja, cada produto interno entre marcador de linha e marcador de coluna reproduz o elemento correspondente na matriz de correspondências. **Em geral, considera-se** $m = 2$, a fim de permitir uma rápida representação visual, embora a qualidade desta aproximação no plano depende do tamanho relativo dos valores singulares δ_1 e δ_2.

A matriz $\mathbf{R}_{[m]}$ dos marcadores de linha é uma matriz com a linhas (tantas quantas as linhas da matriz de correspondências) e m colunas (tantas quantas a cardinalidade da aproximação, que é também a dimensão do espaço onde se fará a representação gráfica). Por seu lado, a matriz $\mathbf{C}_{[m]}$ dos marcadores de colunas tem b linhas (tantas quantas as colunas da matriz de correspondências) e m colunas. As linhas de cada uma destas matrizes são usadas como os marcadores de linhas e colunas associadas à matriz de correspondências. **Marcadores de linha próximos indicam que os respectivos perfis de linha são semelhantes**. Da mesma forma, **Marcadores de coluna próximos indicam que os respectivos perfis de coluna são semelhantes**.

Uma vez que, para quaisquer vectores $\mathbf{x}$, $\mathbf{y}$, dum espaço real, o produto interno usual é dado por

$$< \mathbf{x}, \mathbf{y} > \quad = \quad \mathbf{x}^t \mathbf{y} \quad = \quad \|\mathbf{x}\| \cdot \|\mathbf{y}\| \cdot \cos(\theta) \;,$$

onde θ é o ângulo entre os vectores $\mathbf{x}$ e $\mathbf{y}$, temos as seguintes relações:

- **marcadores de linha (ou coluna) próximos da origem** (ou seja, com norma aproximadamente nula) indicam que os elementos na matriz de correspondências (aproximada) associados a essa linha (ou coluna) são igualmente próximos de zero, ou seja, pouco contribuem para o valor da estatística χ^2 e portanto **estão próximos dos valores que seria de esperar sob a hipótese de independência**;

- mesmo que um par marcador de linha/marcador de coluna esteja distantes da origem (normas afastadas de zero), **caso o ângulo formado por esse par de marcadores de linha e coluna seja aproximadamente recto (ou seja,** $\cos(\theta) \approx 0$**)**, então essa combinação de linha/coluna tem igualmente valor próximo de zero na matriz de correspondências aproximada, pelo que **se trata duma célula onde a frequência observada está próxima do que seria de esperar sob a hipótese de independência**;

- pelo contrário, **se um par marcador de linha/marcador de coluna estiver distante da origem e formar um ângulo quase nulo ou quase raso (**$\cos(\theta) \approx 1$ **ou** $\cos(\theta) \approx -1$**), então a célula associada a essa linha/coluna tem frequência observada muito diferente do que seria de esperar sob a hipótese de independência, para mais no caso de** $\cos(\theta) \approx 1$**, para menos no caso de** $\cos(\theta) \approx -1$.

A representação gráfica destes marcadores permite assim visualizar, de forma aproximada, quais as combinações de linhas/colunas onde existem desvios apreciáveis à hipótese de independência, quer por haver fenómenos de agregação, quer por haver fenómenos de aversão.

7.6 Relação com uma Análise em Correlações Canónicas

A Análise Factorial de Correspondências tem uma ponte com o método das Correlações Canónicas estudado no Capítulo anterior. De facto, qualquer tabela de contingências com um total de n observações pode ser visto como uma síntese da informação contida numa matriz $n \times (a + b)$ em que a cada linha corresponde um dos indivíduos observados, e as $a + b$ colunas são colunas indicatrizes dos a níveis do factor A e dos b níveis do factor B. Nesse caso, a tabela de contingências resultaria do produto cruzado das a indicatrizes associadas ao factor A com as b indicatrizes associadas ao factor B.

Caso se procedesse a efectuar uma Análise em Correlações Canónicas, em que as indicatrizes associadas a cada factor formam cada um dos grupos de variáveis, **mas omitindo a centragem prévia das "variáveis" indicatrizes**, os coeficientes de correlação canónica que se obtêm são os valores singulares da matriz normalizada dos desvios à independência, estando os respectivos eixos de correlação canónica associados aos vectores singulares.

7.7 Um exemplo

Consideremos a tabela de contingência referida no exemplo 7.2, e relativa à classificação de $n_{..} = 5387$ habitantes de Caithness de acordo com a sua côr de olhos (factor A) e de cabelo (factor B)[6].

A matriz de correspondências associada a esta tabela de contingências é:

```
> caith.F <- as.matrix(caith/sum(caith))
> caith.F
             fair         red     medium       dark        black
blue   0.06051606 0.007054019 0.04473733 0.02041953 0.0005568962
light  0.12771487 0.021533321 0.10840913 0.03489883 0.0007425283
medium 0.06367180 0.015593094 0.16873956 0.07648042 0.0048264340
dark   0.01819194 0.008910340 0.07480973 0.12641544 0.0157787266
```

A função `as.matrix`, não sendo imprescindível neste momento, é útil para transformar a *data frame* numa matriz e permitir efectuar operações algébricas simples no `R`.

Os vectores de frequências relativas de linha e de coluna são, respectivamente:

```
> caith.r <- apply(caith,1,sum)/sum(caith)
> caith.r
     blue     light    medium      dark
0.1332838 0.2932987 0.3293113 0.2441062
> caith.c <- apply(caith,2,sum)/sum(caith)
> caith.c
      fair        red     medium       dark      black
0.27009467 0.05309077 0.39669575 0.25821422 0.02190459
```

Nestes vectores lemos as proporções de cada categoria de côr de olhos e de côr de cabelo na população de Caithness como um todo. Mas está ausente a informação sobre associações de cores de olhos e cabelo.

No caso (implausível) de as cores de olhos e cabelo serem independentes, e tomando as estimativas de probabilidades marginais dadas pelos vectores $\mathbf{r}$ e $\mathbf{c}$ acima calculados, esperar-se-ia uma matriz $\mathbf{F}$ próxima de $\mathbf{r}\mathbf{c}^t$ que, no nosso caso, é a matriz

```
> caith.r %*% t(caith.c)
           fair         red     medium       dark       black
[1,] 0.03599925 0.007076142 0.05287313 0.03441578 0.002919527
[2,] 0.07921841 0.015571454 0.11635034 0.07573389 0.006424586
```

[6]Recorde-se que, para disponibilizar esta tabela de contingência, é necessário carregar o módulo `MASS` para a sessão de trabalho.

```
[3,] 0.08894523 0.017483392 0.13063639 0.08503286 0.007213428
[4,] 0.06593178 0.012959786 0.09683588 0.06303169 0.005347045
```

A fim de voltar a colocar os nomes de linhas nesta nova matriz (perdidos durante os cálculos), pode proceder-se como indicado seguidamente.

```
> Fcar1 <- caith.r %*% t(caith.c)
> rownames(Fcar1) <- rownames(caith)
> Fcar1
             fair         red     medium       dark       black
blue   0.03599925 0.007076142 0.05287313 0.03441578 0.002919527
light  0.07921841 0.015571454 0.11635034 0.07573389 0.006424586
medium 0.08894523 0.017483392 0.13063639 0.08503286 0.007213428
dark   0.06593178 0.012959786 0.09683588 0.06303169 0.005347045
```

Constata-se que algumas associações são bastante mais frequentes do que a hipótese de independência deixaria supôr. Por exemplo, há cerca de três vezes mais pessoas de cabelo escuro e olhos pretos do que seria previsível pela hipótese de independência, e cerca do dobro da associação olhos e cabelo "escuros". Por outro lado, algumas associações, como cabelo claro e olhos escuros, são muito menos frequentes na realidade do que a hipótese de independência faria prever. Efectivamente, um teste χ^2 de Pearson à independência dos dois factores (disponível no R através do comando `chisq.test`) produz uma rejeição muito clara da hipótese nula de independência:

```
> chisq.test(caith)

Pearson's Chi-squared test

data:  caith
X-squared = 1240.039, df = 12, p-value < 2.2e-16
```

A fim de efectuar uma Análise de Correspondências no R, pode utilizar-se a função `corresp`, disponível no módulo `MASS`. Para utilizar a função é obrigatório especificar a tabela sobre a qual se deseja trabalhar. É também conveniente indicar a dimensão da resposta que se pretende, uma vez que, por omissão, a função produz um único par de coeficientes de linhas e colunas, e respectivo valor singular.

Alternativamente, pode utilizar-se a função `dudi.coa`, disponível no módulo `ade4`. Este último módulo tem numerosas outras funções de utilidade em análises multivariadas.

7.8 Exercícios de Análise de Correspondências

1. Considere a tabela de tipo locais × espécies, relativa a aves limícolas, apresentada no Exemplo 7.1. A *data frame* com estes dados chama-se `waders` e encontra-se no módulo `MASS`.

   (a) Construa os vectores **r** e **c** de frequências relativas de linhas e de colunas, respectivamente. Interprete os resultados.

   (b) Construa a matriz dos desvios (não-normalizados) à independência, dados por $\mathbf{F} - \mathbf{r}\mathbf{c}^t$. Comente as associações que mais se destacam do que seria de esperar, sob a hipótese de independência.

   (c) Utilize a função `corresp` (também do módulo `MASS`) para efectuar uma Análise de Correspondências desta tabela. Em particular, peça uma aproximação a mais do que um dimensão (através do argumento `nf`) e represente graficamente, a duas dimensões, a tabela (através da função `biplot`). Comente a qualidade deste gráfico.

   (d) Repita a alínea anterior, mas utilizando apenas uma dimensão (`nf`=1). Utilize a função `plot` para produzir o respectivo gráfico. Interprete esse gráfico.

   (e) Utilizando o `R`, construa a matriz normalizada dos desvios à independência, dada na equação (7.14). Calcule a sua Decompoisção em Valores Singulares (utilizando a função `svd`). Compare com os resultados obtidos através do comando `corresp` e comente. Procure explicar as diferenças nos valores dos coeficientes.

Capítulo 8

Inferência Multivariada

Página em construção

8.1 Exercícios de Inferência Multivariada

1. Considere a matriz 2×2 dada por $\mathbf{A} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$.

   (a) Considere o vector $\mathbf{x} = (r, 0)$ (para qualquer $r > 0$) e o vector $\mathbf{Ax}$. Desenhe esses pontos num referencial $x0y$. Comente sobre qual o efeito da matriz $\mathbf{A}$ sobre os vectores do eixo dos xx.

   (b) Considere o vector $\mathbf{y} = (0, r)$ (para qualquer $r > 0$) e o vector $\mathbf{Ay}$. Desenhe esses pontos num referencial $x0y$. Comente sobre qual o efeito da matriz $\mathbf{A}$ sobre os vectores do eixo dos yy.

   (c) Diga qual o efeito de pré-multiplicar um vector genérico (x, y) pela matriz $\mathbf{A}$.

   (d) Mostre que a matriz $\mathbf{A}$ é uma matriz ortogonal.

   (e) Mostre que qualquer matriz ortogonal 2×2 tem de ter a estrutura da matriz $\mathbf{A}$, para algum $\theta \in [0, 2\pi[$.

2. Considere os dados dos lírios, na *data frame* `iris`.

   (a) Para o par de variáveis constituído pelas duas medições sobre as pétalas,

      i. Calcule o vector médio amostral e a matriz de variâncias amostral.

      ii. Efectue um teste de hipóteses ao nível de significância $\alpha = 0.05$ para saber se é admissível que o vector das médias populacionais de comprimento e largura (respectivamente) das pétalas seja o vector $\boldsymbol{\mu} = (3.8, 1.1)$.

      iii. Construa os intervalos de confiança a 95% para a média populacional de cada variável separadamente, utilizando os tradicionais resultados das disciplinas introdutórias de Estatística.

      iv. Com o auxílio do `R`, construa a elipse a 95% de confiança para o vector do par de médias populacionais dessas variáveis, admitindo que os dados são uma amostra aleatória duma única população multinormal. Compare o resultado com os intervalos de confiança obtidos separadamente para a média populacional de cada variável e comente. (**Nota:** Utilize a função experimental `elipse.mu` que está disponível no objecto `elipsemu.RData`, que pode ser descarregado a partir da área da disciplina de Estatística Multivariada em `\\prunus\home\cadeiras`).

      v. Calcule os valores e vectores próprios da matriz de variâncias amostral das duas medições das pétalas. Interprete o seu significado.

   (b) Repita a alínea anterior, mas agora usando as duas larguras (sépalas e pétalas). Em particular,

      i. Comente a forma diferente da elipse agora obtida. A que se deve essa forma diferente? Que ilações pode tirar, no que respeita à utilidade das regiões de confiança conjuntas para pares de médias?

      ii. Trace a recta de regressão linear de largura da pétaa sobre largura da sépala, por cima do gráfico. Porque não coincide a recta com a direcção de variabilidade principal?

   (c) Repita as alínea anteriores, mas agora usando as duas medidas de sépalas. Comente.

3. Considere os dados relativos às medições morfométricas sobre lavagantes, já estudados em Capítulos anteriores e disponíveis na *data frame* `lavagantes`. Recorde-se que os $n = 63$ individuos se subdividiam em $k = 3$ grupos de igual dimensão, constituídos por machos reprodutores, machos não reprodutores e fêmeas (organizados na *data frame* por linhas contíguas).

   (a) Calcule as médias amostrais de cada variável, para os $n_c = 21$ lavagantes de cada subgrupo. **Nota:** explore o comando `by` do `R`.

   (b) Utilize uma Análise de Variância Multivariada (com a estatística de Wilks) para verificar se o vector das $p = 13$ médias de nível populacionais se pode considerar igual nos três grupos. **Nota:** Explore o comando `manova` do `R`. Quais os pressupostos que tem de admitir para que esta análise seja válida?

   (c) Considere agora apenas as $p = 7$ variáveis relativas à carapaça (variáveis 1, 3 e 4), areola (variáveis 6 e 7) e *rostrum* (variáveis 8 e 9). Efectue uma MANOVA para determinar se pode admitir a igualdade dos três vectores de médias populacionais, utilizando:

   - a estatística de Wilks;
   - a estatística de Bartlett-Pillai;
   - a estatística de Hotelling-Lawley;
   - a estatística de Roy.

   Comente os resultados obtidos.

   (d) Construa as nuvens de pontos para todos os pares de variáveis referidos na alínea anterior, utilizando cores diferentes para os individuos de cada grupo. Comente o resultado à luz dos resultados da MANOVA da alínea anterior.

   (e) Considere agora as seis variáveis não consideradas na alínea 3c): medições das caudas (variáveis 2 e 5), largura pós-orbital (variável 10) e medições da tenaz (variáveis 11, 12 e 13).

      i. Efectue uma MANOVA (com base numa estatística à sua escolha) para analisar a hipótese da igualdade dos $k = 3$ vectores de médias populacionais dos machos dos dois tipos e as fêmeas. Comente os resultados.

      ii. Compare as conclusões da subalínea anterior com as conclusões da Análise em Componentes Principais destes dados, efectuada no Capítulo 2.

      iii. Construa as nuvens de pontos dos pares de variáveis agora considerados, distinguindo os individuos de cada grupo. Comente os resultados, à luz das conclusões a que chegou.

4. Considere os dados relativos às medições sobre $n = 600$ folhas de videira (*data frame* `videiras`) de três diferentes castas.

   (a) Construa as nuvens de pontos para cada par de variáveis, utlizando cores diferentes para as folhas de cada casta. Comente.

   (b) Calcule as médias amostrais de cada variável, em cada casta. Comente.

   (c) Efectue uma MANOVA para estudar se é possível considerar iguais os vectores médios populacionais das quatro variáveis, em cada casta. Comente os resultados, à luz das alíneas anteriores.

Apêndice A

Funções de $\mathbb{R}^n$ – revisão

Lembrem-se algumas noções básicas sobre funções em $\mathbb{R}^n$, e nomeadamente o cálculo de extremos locais de tais funções, estudadas na disciplina de Complementos de Álgebra e Análise. Para uma discussão mais pormenorizada deste tema, consulte-se a Bibliografia dessa disciplina ou, por exemplo, *Calculus in Vector Spaces* (2a. edição), Corbin, L.J. & Szczaba, R.H., Marcel Dekker, 1995.

Como em muitas outras situações, a utilização de notação matricial simplifica notavelmente a discussão.

Seja:

$$\begin{array}{rl} f: & \mathbb{R}^n \longrightarrow \mathbb{R} \\ & \mathbf{x} \longrightarrow f(\mathbf{x}) \end{array}$$

Admita-se que as funções às quais se faz referência são sempre diferenciáveis.

Represente-se o **operador** que a cada função (do tipo acima referido) faz corresponder o vector das suas derivadas parciais por $\frac{\partial}{\partial \mathbf{x}}$, isto é:

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

e seja $\left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\right)_{(\mathbf{x}=\mathbf{x}_0)} \in \mathbb{R}^n$ o valor desse vector de derivadas parciais no ponto $\mathbf{x} = \mathbf{x}_0$, ou **gradiente** de f no ponto $\mathbf{x} = \mathbf{x}_0$. Os gradientes de alguns tipos de funções são particularmente simples de calcular. De facto, é fácil verificar que:

1. $f(\mathbf{x}) = c \quad , \quad \forall \mathbf{x} \in \mathbb{R}^n \quad \Longrightarrow \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{0} \quad , \forall \mathbf{x} \in \mathbb{R}^n$
2. $f(\mathbf{x}) = \mathbf{a}^t\mathbf{x} = \sum_{i=1}^n a_i x_i \quad \Longrightarrow \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{a}$, para qualquer vector de coeficientes $\mathbf{a} \in \mathbb{R}^n$.
3. **(Forma quadrática)**. $f(\mathbf{x}) = \mathbf{x}^t\mathbf{A}\mathbf{x} = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j \Longrightarrow \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$, para qualquer matriz simétrica $\mathbf{A}_{nxn}$.

4. **(Forma quadrática generalizada)**. $f(\mathbf{x}) = (\mathbf{a} - \mathbf{C}\mathbf{x})^t \mathbf{A}(\mathbf{a} - \mathbf{C}\mathbf{x}) \Longrightarrow \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = -2\mathbf{C}^t\mathbf{A}(\mathbf{a} - \mathbf{C}\mathbf{x})$, onde $\mathbf{A}_{pxp}$ é qualquer matriz simétrica, $\mathbf{C}_{pxn}$ é uma matriz de coeficientes e $\mathbf{a}_{px1}$ é um vector de coeficientes.

Por sua vez, a matriz das segundas derivadas parciais, isto é, a matriz $\mathbf{H}$ cujo elemento genérico da linha i, coluna j é dado por:

$$(\mathbf{H})_{(i,j)} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}$$

designa-se a **matriz Hessiana** de f. Tem-se:

1. $f(\mathbf{x}) = \mathbf{x}^t\mathbf{A}\mathbf{x} \quad \Longrightarrow \quad \mathbf{H} = 2\mathbf{A}$
2. $f(\mathbf{x}) = (\mathbf{a} - \mathbf{C}\mathbf{x})^t\mathbf{A}(\mathbf{a} - \mathbf{C}\mathbf{x}) \quad \Longrightarrow \quad \mathbf{H} = 2\mathbf{C}^t\mathbf{A}\mathbf{C}$

Para calcular extremos locais de funções em $\mathbb{R}^n$ (que admitem segunda derivada), tem-se:

1. **Condição necessária (pontos de estacionaridade):**

$$\left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\right)_{(\mathbf{x}=\mathbf{x}_0)} = \mathbf{0}$$

2. **Condição suficiente (se $\mathbf{x}_0$ é ponto de estacionaridade)**:

$\mathbf{H}_{(\mathbf{x}=\mathbf{x}_0)}$	definida positiva	$\Longrightarrow \mathbf{x}_0$ é mínimo.
$\mathbf{H}_{(\mathbf{x}=\mathbf{x}_0)}$	definida negativa	$\Longrightarrow \mathbf{x}_0$ é máximo.
$\mathbf{H}_{(\mathbf{x}=\mathbf{x}_0)}$	indefinida	$\Longrightarrow \mathbf{x}_0$ não é máximo nem mínimo

Se $\mathbf{H}_{(\mathbf{x}=\mathbf{x}_0)}$ fôr semi-definida (positiva ou negativa), $\mathbf{x}_0$ poderá ou não ser extremo (mínimo ou máximo, respectivamente), mas não é possível garantir sem uma análise ulterior. Não sendo essencial para os nossos propósitos, não abordaremos ulteriormente esta questão.

O Método dos Multiplicadores de Lagrange

Seja $f : \mathbb{R}^n \longrightarrow \mathbb{R}$. Se se pretende determinar um extremo (máximo ou mínimo) de f, sujeito a restrições adicionais da forma $g(\mathbf{x}) = \mathbf{c}$, para alguma função $g : \mathbb{R}^n \longrightarrow \mathbb{R}$, então constrói-se a função auxiliar $h : \mathbb{R}^{n+1} \longrightarrow \mathbb{R}$:

$$h(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda g(\mathbf{x} - \mathbf{c})$$

A constante λ associada à segunda parcela designa-se um *multiplicador de Lagrange*. Os pontos de estacionaridade (pontos críticos) da função auxiliar h são os pontos que satisfazem as $(n+1)$ condições:

$$\begin{array}{lll} \frac{\partial h}{\partial \lambda} = 0 & \Longleftrightarrow & g(\mathbf{x}) = \mathbf{c} \qquad \text{[A restrição]} \\ \frac{\partial h}{\partial \mathbf{x}} = 0 & \Longleftrightarrow & \frac{\partial f}{\partial \mathbf{x}} - \lambda \cdot \frac{\partial g}{\partial \mathbf{x}} = \mathbf{0} \end{array}$$

Assim, os pontos críticos de h satisfazem sempre a restrição. E note-se que se $\mathbf{x}$ é um vector que satisfaz a condição, as funções h e f tomam o mesmo valor nesse ponto, pois a parcela que as distingue anula-se. Esses pontos críticos são, pois, os candidatos a extremos da função f, sujeitos à restrição. O método generaliza-se de forma óbvia quando existem mais do que uma restrição, havendo tantas parcelas com multiplicadores de Lagrange quantas as restrições adicionais, na função auxiliar.