

Modelos Matemáticos e Aplicações

Introdução à Estatística Multivariada

Jorge Cadima

Secção de Matemática (DCEB) - Instituto Superior de Agronomia (UL)

2021-22

Programa

- Alguns conceitos de teoria de matrizes e álgebra linear
- Análise em Componentes Principais (ACP - PCA)
- Análise Discriminante Linear (ADL - LDA)
- Análises Classificatórias (Cluster Analysis) (Prof. Pedro Silva)

Bibliografía

- Johnson, R.A. & Wichern, D.W. (2007) *Applied Multivariate Statistical Analysis*, 6th ed., Pearson Prentice Hall.
- Jolliffe, I.T. (2002) *Principal Component Analysis*, 2d. ed., Springer (Springer Series in Statistics)
- Krzanowski, W.J. (1998) *Principles of Multivariate Analysis: A User's Perspective*, Oxford Science Publications.
- Morrison, D.F. (1990) *Multivariate Statistical Methods*, 3rd.ed., McGraw-Hill.

Multivariada no :

- Everitt, B. & Hothorn, T. (2011) *An Introduction to Applied Multivariate Analysis with R*. Springer (Use R! Series).
- Zelterman, D. (2015). *Applied Multivariate Statistics with R*. Springer (Statistics for Biology and Health Series).

A matéria prima em ACP e ADL

Matriz de dados $\mathbf{X}_{n \times p}$, com conjuntos de observações:

- de p variáveis **quantitativas** (colunas);
- em n **indivíduos**, ou unidades experimentais (linhas).

Nota: Ao contrário do que sucede na modelação, aqui **todas as variáveis estão em pé de igualdade**.

Abordaremos a faceta **descritiva** (**geométrica**), quer da ACP, quer da ADL, embora em ambos os métodos se possam introduzir conceitos e abordagens probabilísticos.

Objectivo em ACP e ADL

Procuram-se **novas variáveis**, construídas a partir das p variáveis observadas **que salientem**:

- numa ACP: a **variabilidade** entre indivíduos;
- numa ADL: a **separação** entre subgrupos **conhecidos** dos indivíduos.

Em ambos os casos, as **novas variáveis** são **combinações lineares** das p **variáveis observadas**.

Motivação da ACP

Na **representação tradicional**, à **matriz de dados** $\mathbf{X}_{n \times p}$ corresponde uma **nuvem de n pontos em \mathbb{R}^p** :

$$\begin{array}{lll} p \text{ eixos} & \longleftrightarrow & p \text{ variáveis} \\ n \text{ pontos} & \longleftrightarrow & n \text{ indivíduos} \end{array}$$

Esta nuvem de pontos **não é visualizável para $p > 3$** .

A **ACP** pode ser vista como uma **técnica de redução “óptima” da dimensionalidade**: procuram-se os subespaços de dimensão $k < p$ onde a **projecção ortogonal** da nuvem de dados **preserva o máximo de variabilidade** (equivalentemente, **perde o mínimo de variabilidade**).

No caso duma redução para $k=2$ ou $k=3$ dimensões, ter-se-á uma aproximação visualizável da nuvem de pontos.

Um exemplo: os lavagantes de Somers

Dados: $p=13$ variáveis morfométricas sobre $n=63$ lavagantes

> lavagantes

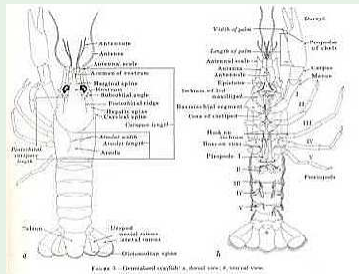
	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13
1	29.42	21.43	14.91	12.58	12.85	10.57	1.76	6.45	6.67	9.14	24.54	10.38	15.37
2	30.06	22.05	14.81	12.54	12.96	10.75	1.73	6.11	7.04	8.76	26.21	11.00	11.92
3	30.30	21.95	15.10	12.97	13.05	11.11	2.05	6.46	7.14	9.35	26.55	11.84	16.50
4	30.75	21.91	15.89	12.85	13.75	10.75	1.71	6.62	6.84	9.53	25.35	11.60	15.47
5	31.06	20.37	15.83	13.15	13.37	11.50	2.15	5.96	7.09	9.15	26.88	11.92	17.24
6	31.27	24.04	17.45	14.49	14.77	12.64	2.06	6.59	7.43	10.75	31.60	14.32	18.95
7	31.39	21.91	15.96	13.41	13.74	11.79	2.03	6.40	6.89	9.82	28.16	12.53	16.90
8	31.51	23.63	15.95	13.14	13.89	11.74	1.94	6.26	6.81	9.36	26.09	11.15	15.48
9	32.12	22.81	16.06	13.29	13.80	12.14	2.02	6.47	7.00	9.70	27.01	11.22	16.65
10	32.40	22.96	16.69	13.82	14.30	12.06	2.03	6.14	7.27	9.53	29.34	12.59	17.90
.....													
56	33.44	24.72	17.06	14.25	16.74	12.42	2.04	6.52	7.25	10.21	26.92	11.40	16.23
57	33.48	25.32	17.50	14.15	17.20	12.40	2.17	6.94	7.54	10.37	26.85	11.40	16.34
58	33.57	25.00	16.74	14.10	16.49	12.43	1.95	7.27	7.37	10.15	25.13	11.23	14.98
59	33.74	25.30	17.11	14.26	16.35	12.37	2.26	6.82	7.41	11.14	26.43	10.91	16.02
60	34.37	25.35	17.98	14.49	16.95	12.69	2.02	7.04	7.35	10.33	27.97	11.75	17.19
61	34.66	25.32	18.50	14.16	17.37	12.60	2.32	6.88	7.59	11.00	27.76	11.87	17.58
62	34.93	26.77	18.00	14.13	16.89	12.67	2.04	7.14	7.79	10.36	26.98	11.55	17.20
63	35.73	25.79	18.35	15.06	17.15	13.14	2.15	7.09	7.83	10.59	28.29	12.30	17.45



Estes $13 \times 63 = 819$ valores definem uma nuvem de 63 pontos em \mathbb{R}^{13} .

Os lavagantes de Somers (cont.)

Dados lavagantes: nomes completos das variáveis



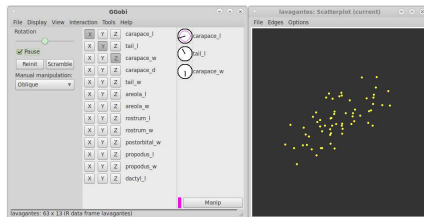
Thoma, Roger F., *A Field Guide to the Crayfishes of Obed Wild and Scenic River*, www.nps.gov.

x1	carapace_l	comprimento da carapaça
x2	tail_l	comprimento da cauda
x3	carapace_w	largura da carapaça
x4	carapace_d	profundidade da carapaça
x5	tail_w	largura da cauda
x6	areola_l	comprimento da aréola
x7	areola_w	largura da aréola
x8	rostrum_l	comprimento do rostro
x9	rostrum_w	largura do rostro
x10	postorbital_w	largura post-orbital
x11	propodus_l	comprimento da tenaz
x12	propodus_w	largura da tenaz
x13	dactyl_l	comprimento dátil

A representação gráfica de dados multivariados

Até $p=3$ é possível a representação usual dos dados (cada eixo é uma variável, cada indivíduo observado é um ponto nesse sistema de p eixos), por exemplo com o módulo (package) `rggobi`¹.

- > `library(rggobi)`
- > `ggobi(lavagantes)`



Continuamos apenas com visões parciais, resultantes de projectar ortogonalmente a nuvem de $n = 63$ pontos em \mathbb{R}^{13} sobre espaços tri-dimensionais.

¹O software `GGobi`, exterior ao R, é gratuito e de código aberto (www.ggobi.org)

Projecções ortogonais

Qualquer projecção empobrece a representação: ficamos com uma visão parcial. Distâncias podem ficar camufladas.

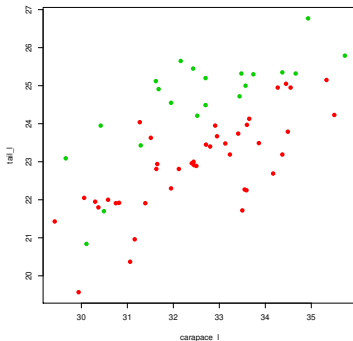
Mas,

- Porquê só projectar sobre (hiper)planos coordenados (definidos pelos eixos das variáveis)? Porque não outros (hiper)planos?
- Qual o (hiper)plano onde a projecção é mais fidedigna?
- Qual o critério de “projecção fidedigna”?

Ideia intuitiva: o subespaço onde se projecta deve preservar, tanto quanto possível, a variabilidade dos pontos. Esta é a abordagem que conduz à Análise em Componentes Principais.

Motivação: Análise Discriminante Linear (ADL)

A ACP trata todos os indivíduos por igual. Os 42 primeiros lavagantes são **machos** (21 reprodutores e 21 não reprodutores) e os 21 finais são **fêmeas**, e uma nuvem de pontos nas duas primeiras variáveis indicia que isso pode ser visível nas variáveis morfométricas.



ADL: identificar a combinação linear das variáveis que melhor separe os dois sexos.

Conceitos: Tipos de matrizes quadradas

Uma matriz diz-se **quadrada** se tem igual número de linhas e colunas.

Eis alguns tipos importantes de matrizes **quadradas**, $\mathbf{A}_{p \times p}$:

A Diagonal	$a_{ij} = 0$ se $i \neq j$ (e existe i tal que $a_{ii} \neq 0$)
\mathbf{I}_p Identidade	$\mathbf{A} = \mathbf{I}_p \iff a_{ij} = \begin{cases} 0 & \text{para } i \neq j \\ 1 & \text{para } i = j \end{cases}$
\mathbf{A}^{-1} Inversa de \mathbf{A}	$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_p$ (nem sempre existe, mas quando existe é única)
A Simétrica	$\mathbf{A}^t = \mathbf{A} \iff a_{ij} = a_{ji}, \forall i, j$
A Idempotente	$\mathbf{A}^2 = \mathbf{A}\mathbf{A} = \mathbf{A}$
A Ortogonal	$\mathbf{A}^{-1} = \mathbf{A}^t \iff \mathbf{A}^t\mathbf{A} = \mathbf{A}\mathbf{A}^t = \mathbf{I}_p$

Numa matriz ortogonal, quer as colunas, quer as linhas, formam conjuntos **ortonormados** (orthonormal) de vectores: vectores $\vec{\mathbf{a}}_i$ de norma um ($\|\vec{\mathbf{a}}_i\| = \sqrt{\vec{\mathbf{a}}_i^t \vec{\mathbf{a}}_i} = 1$) e ortogonais entre si ($\vec{\mathbf{a}}_i^t \vec{\mathbf{a}}_j = 0$, se $i \neq j$).

Matrizes de (co-)variâncias

As matrizes simétricas são importantes em Estatística porque as matrizes de (co)variâncias e de correlações são matrizes simétricas.

As matrizes de (co-)variâncias de conjuntos $n \times p$ de dados são da forma

$$\mathbf{S} = \frac{1}{n-1} \mathbf{X}^{\text{ct}} \mathbf{X}^{\text{c}},$$

onde \mathbf{X}^{c} é a matriz $n \times p$ cujas colunas são os p vectores centrados de observações, $\vec{\mathbf{x}}_j^{\text{c}}$, ou seja, a matriz de elemento genérico $x_{ij} - \bar{x}_{.j}$:

$$s_{jk} = \frac{1}{n-1} [\mathbf{X}^{\text{ct}} \mathbf{X}^{\text{c}}]_{jk} = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_{.j})(x_{ik} - \bar{x}_{.k}) = \text{Cov}_{\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_k}$$

Os valores próprios de matrizes de (co-)variâncias são sempre não negativos e, desde que não haja multicolinearidade nas variáveis subjacentes, são mesmo positivos.

Valores e vectores próprios (eigenvalues/vectors)

Definição: Valores e vectores próprios

Dada uma matriz real **quadrada** $\mathbf{A}_{p \times p}$, um **vector não-nulo** $\vec{x} \in \mathbb{C}^p$ diz-se um **vector próprio** de \mathbf{A} , e $\lambda \in \mathbb{C}$ diz-se o **valor próprio** correspondente se:

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

Valores e vectores próprios de matrizes simétricas

Se $\mathbf{A}_{p \times p}$ for uma matriz **simétrica**, os seus valores/vectores próprios têm boas propriedades:

- Os valores e vectores próprios são sempre **reais**.
- Vectores próprios associados a valores próprios diferentes são **ortogonais** entre si.
- Mesmo que haja valores próprios repetidos, é **possível determinar um conjunto ortonormado de p vectores próprios** (logo, p valores próprios).

A Decomposição Espectral duma matriz simétrica

Teorema da Decomposição Espectral

Seja $\mathbf{A}_{p \times p}$ uma matriz simétrica. Sejam:

- $\{\vec{\mathbf{v}}_i\}_{i=1}^p$ um conjunto ortonormado de vectores próprios; e
- $\{\lambda_i\}_{i=1}^p$ os seus p valores próprios correspondentes.

Definindo:

- a matriz diagonal $\mathbf{\Lambda}_{p \times p}$ cujos elementos diagonais são λ_i ; e
- a matriz (necessariamente ortogonal) $\mathbf{V}_{p \times p}$, cujas colunas são $\vec{\mathbf{v}}_i$;

tem-se:

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^t \quad \iff \quad \mathbf{A} = \sum_{i=1}^p \lambda_i \vec{\mathbf{v}}_i \vec{\mathbf{v}}_i^t .$$

Os valores e vectores próprios são a essência duma matriz simétrica \mathbf{A} .

Notas sobre as decomposições espectrais

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^t \iff \mathbf{A} = \sum_{i=1}^p \lambda_i \vec{\mathbf{v}}_i \vec{\mathbf{v}}_i^t .$$

- Se todos os **valores próprios** forem **diferentes**, os **vetores próprios** $\vec{\mathbf{v}}_i$ são **únicos**, a menos de troca de sinal (tanto se pode usar $\vec{\mathbf{v}}_i$ como $-\vec{\mathbf{v}}_i$).
- Ordenando os elementos diagonais de $\mathbf{\Lambda}$ ($\lambda_1 > \lambda_2 > \dots > \lambda_p$), a decomposição $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^t$ é **única** (a menos de trocas de sinais nas colunas de \mathbf{V}).
- Com **valores próprios iguais**, a decomposição **não é única**.

De facto, sejam $\vec{\mathbf{x}}_1$ e $\vec{\mathbf{x}}_2$ **vetores próprios** de \mathbf{A} com o mesmo valor próprio λ . Como $\mathbf{A}\vec{\mathbf{x}}_1 = \lambda\vec{\mathbf{x}}_1$ e $\mathbf{A}\vec{\mathbf{x}}_2 = \lambda\vec{\mathbf{x}}_2$, tem-se:

$$\mathbf{A}(\alpha\vec{\mathbf{x}}_1 + \beta\vec{\mathbf{x}}_2) = \alpha\mathbf{A}\vec{\mathbf{x}}_1 + \beta\mathbf{A}\vec{\mathbf{x}}_2 = \alpha \cdot \lambda\vec{\mathbf{x}}_1 + \beta \cdot \lambda\vec{\mathbf{x}}_2 = \lambda(\alpha\vec{\mathbf{x}}_1 + \beta\vec{\mathbf{x}}_2) ,$$

logo $\alpha\vec{\mathbf{x}}_1 + \beta\vec{\mathbf{x}}_2$ também é **vector próprio** de \mathbf{A} , associado ao mesmo valor próprio λ . **Todos os vetores do subespaço** gerado por $\vec{\mathbf{x}}_1$ e $\vec{\mathbf{x}}_2$ são **vetores próprios** associados a λ .

Traços (de matrizes quadradas)

Seja \mathbf{A} uma matriz quadrada:

- O traço de \mathbf{A} define-se como a soma dos elementos diagonais:

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^p a_{ii}.$$

- O traço é um operador linear, isto é,

$$\text{tr}(\alpha \mathbf{A} + \beta \mathbf{B}) = \alpha \text{tr}(\mathbf{A}) + \beta \text{tr}(\mathbf{B})$$

O produto interno de duas matrizes de igual dimensão, $\mathbf{A}_{n \times p}$ e $\mathbf{B}_{n \times p}$, costuma definir-se como:

$$\langle \mathbf{A}, \mathbf{B} \rangle = \text{tr}(\mathbf{A}^t \mathbf{B}) = \sum_{j=1}^p (\mathbf{A}^t \mathbf{B})_{jj} = \sum_{i=1}^n \sum_{j=1}^p a_{ij} b_{ij}.$$

Circularidade do traço

Produtos de duas matrizes: $\mathbf{A}_{n \times p}, \mathbf{B}_{p \times n} \implies \text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$.

(Mesmo que $\mathbf{AB} \neq \mathbf{BA}$: são ambos $\sum_{i=1}^n \sum_{j=1}^p a_{ij} b_{ji}$)

Produtos de 3 matrizes: $\mathbf{A}_{m \times k}, \mathbf{B}_{k \times p}, \mathbf{C}_{p \times m} \implies \text{tr}(\mathbf{ABC}) = \text{tr}(\mathbf{BCA})$.

(Aplicar o resultado anterior às duas matrizes \mathbf{A} e \mathbf{BC})

Produtos de n matrizes: Se $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \dots, \mathbf{A}_n$ matrizes de dimensões $(p_0 \times p_1), (p_1 \times p_2), (p_2 \times p_3), \dots, (p_{n-1} \times p_0)$, então,

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

(Aplicar resultado inicial às matrizes \mathbf{A}_1 e $\mathbf{A}_2 \mathbf{A}_3 \cdots \mathbf{A}_n$)

Pela Decomposição Espectral, é fácil ver que o traço duma matriz simétrica \mathbf{A} é também a soma dos seus valores próprios:

$$\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{V} \mathbf{\Lambda} \mathbf{V}^t) = \text{tr}(\underbrace{\mathbf{\Lambda} \mathbf{V}^t \mathbf{V}}_{=\mathbf{I}_p}) = \text{tr}(\mathbf{\Lambda}) = \sum_{i=1}^p \lambda_i.$$

ACP: Uma introdução estatística

Uma introdução frequente à ACP usa conceitos estatísticos.

Seja dada a matriz de dados $\mathbf{X}_{n \times p}$ (cada coluna corresponde a uma variável, e cada linha a um indivíduo observado).

Pretende-se determinar a combinação linear das p variáveis de variância máxima.

Isto é, pretende-se determinar o vector $\vec{\mathbf{v}} = (v_1, v_2, \dots, v_p) \in \mathbb{R}^p$ tal que

$$\mathbf{X}\vec{\mathbf{v}} = v_1 \vec{\mathbf{x}}_1 + v_2 \vec{\mathbf{x}}_2 + v_3 \vec{\mathbf{x}}_3 + \dots + v_p \vec{\mathbf{x}}_p$$

tenha variância máxima (sendo $\vec{\mathbf{x}}_j \in \mathbb{R}^n$ o vector das observações da variável j , ou seja, a j -ésima coluna de \mathbf{X}).

A variância de $\mathbf{X}\vec{\mathbf{v}}$ é dada por $\vec{\mathbf{v}}^t \mathbf{S} \vec{\mathbf{v}}$, sendo \mathbf{S} a matriz de (co)variâncias dos dados. Logo, quer-se o vector $\vec{\mathbf{v}}$ que maximize $\vec{\mathbf{v}}^t \mathbf{S} \vec{\mathbf{v}}$.

Variância de combinações lineares de variáveis

Seja \mathbf{S} a matriz de (co-)variâncias gerada pela matriz de dados \mathbf{X} .

A variância duma combinação linear das colunas de \mathbf{X} , $\vec{y} = \mathbf{X}\vec{a}$, é a forma quadrática de \mathbf{S} definida por \vec{a} :

$$\text{var}(\vec{y}) = \text{var}(\mathbf{X}\vec{a}) = \vec{a}^t \mathbf{S} \vec{a} .$$

De facto, $\vec{a}^t \mathbf{S} \vec{a} = \frac{1}{n-1} \vec{a}^t \mathbf{X}^c t \mathbf{X}^c \vec{a} = \frac{1}{n-1} \|\mathbf{X}^c \vec{a}\|^2$, e

$$\mathbf{X}^c \vec{a} = (\mathbf{I}_n - \mathbf{P}_{\vec{1}_n}) \mathbf{X} \vec{a} = \mathbf{X} \vec{a} - \mathbf{P}_{\vec{1}_n} \mathbf{X} \vec{a} = \vec{y} - (\bar{y}) \vec{1}_n$$

é o vector centrado da combinação linear $\vec{y} = \mathbf{X}\vec{a}$, de elemento genérico $y_i^c = y_i - \bar{y}$. Logo, $\vec{a}^t \mathbf{S} \vec{a}$ é a variância amostral de $\vec{y} = \mathbf{X}\vec{a}$:

$$\vec{a}^t \mathbf{S} \vec{a} = \frac{1}{n-1} \|\mathbf{X}^c \vec{a}\|^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = \text{var}(\vec{y}) .$$

A covariância entre diferentes combinações lineares, $\mathbf{X}\vec{a}$ e $\mathbf{X}\vec{b}$, é:

$$\text{Cov}[\mathbf{X}\vec{a}, \mathbf{X}\vec{b}] = \vec{a}^t \mathbf{S} \vec{b} .$$

Introdução estatística (cont.)

Sem outras restrições, o problema de maximizar $\vec{v}^t \mathbf{S} \vec{v}$ não tem solução, pois pode escolher-se \vec{v} de elementos arbitrariamente grandes.

Impõe-se a restrição de apenas considerar vectores de norma 1 (soma dos quadrados dos coeficientes 1), ou seja, vectores da forma $\frac{\vec{v}}{\|\vec{v}\|}$ (com $\vec{v} \neq \vec{0}$).

Assim, gera-se o problema de maximizar o chamado **quociente de Rayleigh-Ritz (Rayleigh-Ritz ratio)** de \mathbf{S} :

$$\max_{\vec{v} \in \mathbb{R}^p \setminus \{\vec{0}\}} \left[\frac{\vec{v}}{\|\vec{v}\|} \right]^t \mathbf{S} \frac{\vec{v}}{\|\vec{v}\|} = \max_{\vec{v} \in \mathbb{R}^p \setminus \{\vec{0}\}} \frac{\vec{v}^t \mathbf{S} \vec{v}}{\|\vec{v}\|^2} = \max_{\vec{v} \in \mathbb{R}^p \setminus \{\vec{0}\}} \frac{\vec{v}^t \mathbf{S} \vec{v}}{\vec{v}^t \vec{v}}$$

A solução é dada pelo **vector próprio \vec{v}_1** (de norma 1), associado ao maior valor próprio de \mathbf{S} , λ_1 .

Teorema de Rayleigh-Ritz

Seja $\mathbf{A}_{p \times p}$ matriz simétrica, com valores próprios por ordem decrescente:

$$\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{p-1} \geq \lambda_p = \lambda_{\min}.$$

- O maior valor próprio de \mathbf{A} verifica: $\lambda_{\max} = \max_{\vec{x} \neq \vec{0}} \frac{\vec{x}' \mathbf{A} \vec{x}}{\vec{x}' \vec{x}}$,

com $\vec{x} = \vec{v}_1$, o vector próprio correspondente a λ_{\max} .

- O menor valor próprio de \mathbf{A} verifica: $\lambda_{\min} = \min_{\vec{x} \neq \vec{0}} \frac{\vec{x}' \mathbf{A} \vec{x}}{\vec{x}' \vec{x}}$,

com $\vec{x} = \vec{v}_p$, o vector próprio correspondente a λ_{\min} .

- Os restantes valores (λ_j)/vectores (\vec{v}_j) próprios de \mathbf{A} também são caracterizáveis a partir do quociente de Rayleigh-Ritz de \mathbf{A} :

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

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

verificando-se as igualdades quando $\vec{x} = \vec{v}_j$.

A primeira Componente Principal

A primeira Componente Principal é a combinação linear $\vec{y}_1 = \mathbf{X}\vec{v}_1$, onde \vec{v}_1 é o vector próprio associado ao maior valor próprio de \mathbf{S} .

Nota: Se \vec{v} é vector próprio, $-\vec{v}$ também é. As soluções do problema definem direcções, mas não definem sentidos.

Tal como o vector próprio \vec{v}_1 , também a primeira CP, $\mathbf{X}\vec{v}_1$, está definida a menos duma multiplicação por -1 .

O vector de coeficientes \vec{v}_1 define a direcção no espaço \mathbb{R}^p de variância máxima da nuvem de n pontos definida pelos dados.

O valor próprio λ_1 é a variância desta primeira CP:

$$\text{var}(\vec{y}_1) = \text{var}(\mathbf{X}\vec{v}_1) = \vec{v}_1^t \mathbf{S} \vec{v}_1 = \vec{v}_1 \cdot \lambda_1 \vec{v}_1 = \lambda_1 \cdot \vec{v}_1^t \vec{v}_1 = \lambda_1 .$$

Quanto maior for λ_1 , mais alongada será a nuvem de pontos na direcção definida por esta primeira CP.

ACP: introdução estatística (cont.)

Fixada a 1a. CP, procura-se **nova combinação linear** $\vec{y} = \mathbf{X}\vec{v}$, com $\vec{v}^t\vec{v} = 1$, de **variância máxima**, **não-correlacionada com a anterior**.

Uma **correlação nula** equivale a uma **covariância nula**, e a **covariância** de duas **combinações lineares das colunas** duma matriz \mathbf{X} , no nosso caso $\mathbf{X}\vec{v}_1$ e $\mathbf{X}\vec{v}$, é dada por $\vec{v}^t\mathbf{S}\vec{v}_1$, sendo \mathbf{S} a matriz de covariâncias associada aos dados em \mathbf{X} .

Mas \vec{v}_1 é um vector próprio de \mathbf{S} , associado ao valor próprio λ_1 . Logo:

$$\text{cov}(\mathbf{X}\vec{v}, \mathbf{X}\vec{v}_1) = \vec{v}^t\mathbf{S}\vec{v}_1 = 0 \quad \Leftrightarrow \quad \lambda_1\vec{v}^t\vec{v}_1 = 0 \quad \Leftrightarrow \quad \vec{v} \perp \vec{v}_1 .$$

Logo, maximizar a variância de $\mathbf{X}\vec{v}$ sujeito à não correlação de $\mathbf{X}\vec{v}$ com $\mathbf{X}\vec{v}_1$ equivale a **maximizar** $\frac{\vec{v}^t\mathbf{S}\vec{v}}{\vec{v}^t\vec{v}}$, sujeito a \vec{v} e \vec{v}_1 serem ortogonais.

É de novo um problema associado aos quocientes de Rayleigh-Ritz.

ACP: introdução estatística (cont.)

Assim, maximizar a variância de $\mathbf{X}\vec{\mathbf{v}}$ sujeito à não correlação de $\mathbf{X}\vec{\mathbf{v}}$ com $\mathbf{X}\vec{\mathbf{v}}_1$ corresponde a tomar $\vec{\mathbf{v}} = \pm\vec{\mathbf{v}}_2$, o vector próprio de \mathbf{S} associado ao seu segundo maior valor próprio, λ_2 .

$\vec{\mathbf{y}}_2 = \pm\mathbf{X}\vec{\mathbf{v}}_2$ é a segunda componente principal, com variância λ_2 .

Sucessivas CPs são soluções do problema de determinar combinações lineares, não-correlacionadas entre si, de variância máxima.

A j -ésima componente principal é dada por $\vec{\mathbf{y}}_j = \pm\mathbf{X}\vec{\mathbf{v}}_j$, onde $\vec{\mathbf{v}}_j$ é o vector próprio de \mathbf{S} associado ao j -ésimo maior valor próprio λ_j .

A variância da j -ésima CP é o correspondente valor próprio:
 $var(\vec{\mathbf{y}}_j) = \lambda_j$.

O comando usual para efectuar uma ACP no R é o comando `prcomp`.

O comando `prcomp` tem um **único argumento obrigatório**: o nome do objecto contendo os dados, que deve ser da classe `data.frame` ou `matrix` (com **cada coluna associada a uma variável**).

Como noutros comandos R, o resultado é um objecto de classe `list`, contendo informação vária sobre o resultado da análise.

Nota: Existe também um comando `princomp`, mas por várias razões (incluindo a precisão numérica no caso de matrizes de (co-)variâncias quase singulares), **é preferível a utilização do comando `prcomp`**.

O comando `prcomp`

ACP dos lavagantes

```
> lav.acp <- prcomp(lavagantes)
> lav.acp
```

```
Standard deviations (1, ..., p=13):
```

```
[1] 4.4171243 2.1583124 0.9617894 0.7071970 0.6163559 0.4992560 0.4639879 <- desvios padrões
[8] 0.3848417 0.3362918 0.2500701 0.2060563 0.1770375 0.1405790 <- de cada CP
```

```
Rotation (n x k) = (13 x 13):
```

	PC1	PC2	PC3	PC4	PC5	
carapace_l	0.28762060	0.36935786	0.08475822	-0.31404094	-0.454639049	<- cada coluna é um
tail_l	0.10615292	0.61487598	-0.01728674	0.46421995	0.550775374	vector próprio v_j
carapace_w	0.19089393	0.22112280	0.09978650	-0.10987953	-0.186701149	da matriz de
carapace_d	0.13951311	0.14784642	0.13138041	0.01598041	0.105009202	(co)variâncias
tail_w	0.04682070	0.49290700	-0.05172379	0.06592005	-0.405755003	dos dados. Estes
areola_l	0.13858508	0.15588574	-0.03136931	-0.78849399	0.514893584	vectores contêm os
areola_w	0.02862658	0.02088959	-0.05104427	-0.01123927	-0.005062728	coeficientes das
rostrum_l	0.04321132	0.10238463	-0.00534869	0.10538116	-0.015312405	combinações lineares
rostrum_w	0.06381638	0.06445436	0.05636521	-0.02008425	-0.071806372	que definem as CPs.
postorbital_w	0.08947075	0.12850014	0.07576734	-0.01777992	0.021872310	
propodus_l	0.70705994	-0.28621233	0.04885310	0.16407517	0.077728529	
propodus_w	0.31334632	-0.14849063	0.69820134	0.07580938	0.026674997	
dactyl_l	0.46456390	-0.10926197	-0.67839228	0.05350023	-0.040805783	
[...]						<- omitidos, por razão de espaço, os restantes vectores de coeficientes.

Os coeficientes de cada combinação linear (colunas do objecto `Rotation`) são designados *loadings* em inglês.

Propriedades de CPs

- A soma das variâncias (inércia) das p componentes principais é igual à soma das variâncias das p variáveis originais:

$$\sum_{i=1}^p s_i^2 = \text{tr}(\mathbf{S}) = \text{tr}(\mathbf{V}\mathbf{\Lambda}\mathbf{V}^t) = \text{tr}(\mathbf{\Lambda}\mathbf{V}^t\mathbf{V}) = \text{tr}(\mathbf{\Lambda}) = \sum_{i=1}^p \lambda_i .$$

- Logo, pode afirmar-se que a j -ésima CP explica uma proporção da variabilidade (inércia) total igual a $\pi_j = \frac{\lambda_j}{\sum_{i=1}^p \lambda_i}$.
- Esta medida é extensível a subconjuntos de componentes principais. Às primeiras q CPs corresponde

$$\sum_{i=1}^q \pi_i \times 100\% = \frac{\sum_{i=1}^q \lambda_i}{\sum_{j=1}^p \lambda_j} \times 100\%$$

da variabilidade total (inércia) do conjunto de dados.

O comando `summary`

ACP dos lavagantes

```
> summary(lav.acp)
```

```
Importance of components:
```

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Std. Dev.	4.417	2.158	0.9618	0.7072	0.6164	0.49926	0.46399	0.38484	0.33629	0.25007	0.20606	0.17704	0.14058
Prop.Var.	0.727	0.173	0.0344	0.0186	0.0141	0.00928	0.00802	0.00551	0.00421	0.00233	0.00158	0.00117	0.00074
Cum.Prop.	0.727	0.900	0.9344	0.9530	0.9672	0.97645	0.98446	0.98998	0.99419	0.99652	0.99810	0.99926	1.00000

Na recta associada à primeira componente principal preservamos 72,7% da variabilidade total dos dados.

No plano associado às duas primeiras componentes principais preservamos 90,0% da variabilidade total dos dados.

No espaço a 3 dimensões definido pelas três primeiras CPs preservamos 93,4% da variabilidade total.

Apenas não visualizamos na representação tri-dimensional cerca de 6,6% da variabilidade total.

Vectores de *scores*

Por omissão, o comando `prcomp` não mostra os *scores* de cada individuo numa CP, ou seja, o valor de cada individuo na combinação linear $\vec{y}_j = \mathbf{X}\vec{v}_j$.

Os *scores* são guardados num objecto de nome `x`, na lista produzida pelo comando `prcomp`:

```
> names(lav.acp)
```

```
[1] "sdev"      "rotation" "center"   "scale"    "x"
```

```
> lav.acp$x
```

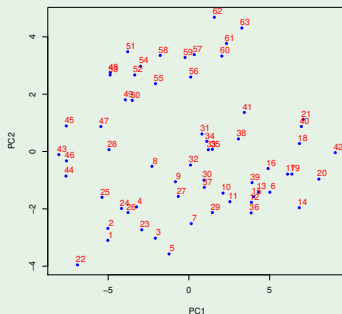
	PC1	PC2	PC3	PC4	PC5	PC6
1	-5.0216041	-3.09975004	-0.93638716	0.590170762	0.34242883	-0.311295721
2	-5.0199046	-2.68138921	1.93090666	0.652936303	0.71306147	2.411219117
3	-2.0772687	-3.02373521	-0.44934354	0.613510708	0.54941375	-0.365822245
[...]						
62	1.5767872	4.68339718	-0.49231884	0.246787192	-0.11313707	0.138658304
63	3.2782407	4.30830749	0.15373020	-0.562657698	-0.73379507	0.200035217
[...]						

São estas as coordenadas das representações gráficas a baixa dimensão que melhor preservam a variabilidade dos dados.

A melhor representação bidimensional

Primeiro plano principal lavagantes

```
> plot(lav.acp$x[,1:2],col="blue", pch=16, cex=0.8)  
> text(lav.acp$x[,1:2]+0.2, label=rownames(lavagantes), col="red")
```

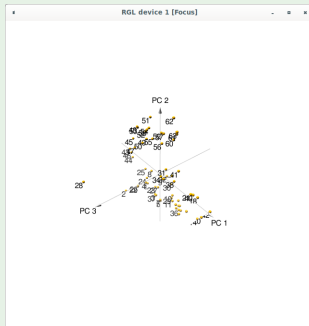


Os indivíduos 43 a 63 são fêmeas, o resto machos. A ACP não usou essa informação, mas apenas o seu reflexo na variabilidade das variáveis morfométricas.

A melhor representação tridimensional

O módulo (**package**) `pca3d` permite criar e rodar a nuvem tri-dimensional definida pelos **scores** nas 3 primeiras CPs:

```
> library(pca3d)
> pca3d(lav.acp, show.labels=TRUE)
```

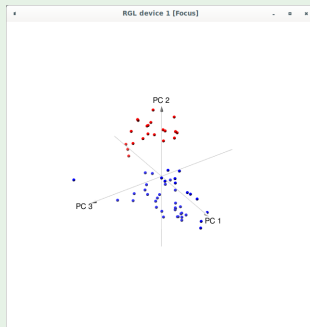


Verifica-se que a terceira CP separa a observação discrepante 28 das restantes. Havendo *outliers*, podem ser identificados por algumas CPs.

A melhor representação tridimensional

O módulo `pca3d` permite usar cores diferentes nos indivíduos:

```
> pca3d(lav.acp, col=rep(c("blue", "red"),c(42,21)))
```

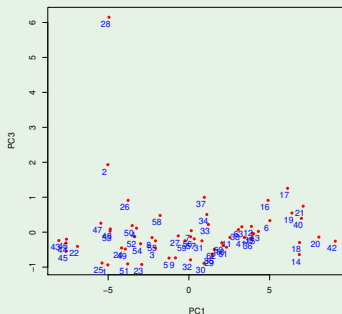


Verifica-se que as duas primeiras CPs separam machos (a azul) e fêmeas (a vermelho).

A observação 28 e a terceira CP

Observação atípica nos lavagantes

```
> plot(lav.acp$x[,c(1,3)], col="red", pch=16, cex=0.8)  
> text(lav.acp$x[,c(1,3)]-0.2, label=rownames(lavagantes), col="blue")
```

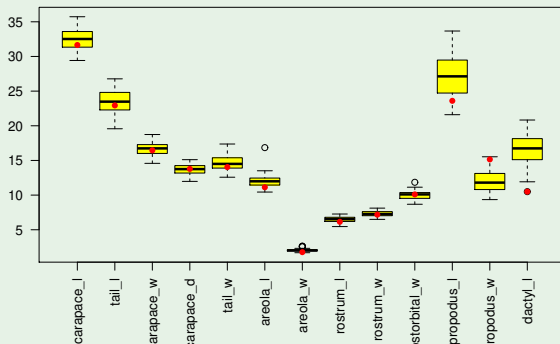


O individuo 28 contribui decisivamente para a terceira direcção ortogonal de maior variabilidade. Porquê? O que tem de diferente o individuo 28?

Revisitando o indivíduo 28

Observação atípica nos lavagantes

- > `boxplot(lavagantes, col="yellow", las=2)`
- > `points(1:13, lavagantes[28,], pch=16, col="red")`



O indivíduo 28 tem medições involgares na tenaz.

Decomposição em valores/vectores próprios

A informação gerada pelo comando `prcomp` poderia ser obtida através da **decomposição espectral** da matriz de (co-)variâncias dos dados, utilizando o comando `eigen`:

```
> eigen(var(lavagantes))
```

```
$values          <-- valores próprios
```

```
[1] 19.51098705  4.65831240  0.92503887  0.50012760  0.37989465  0.24925657  
[7]  0.21528474  0.14810313  0.11309220  0.06253506  0.04245919  0.03134228  
[13]  0.01976246
```

```
$vectors         <-- vectores próprios
```

```
          [,1]      [,2]      [,3]      [,4]      [,5]      [,6]  
[1,]  -0.28762060 -0.36935786 -0.08475822  0.31404094 -0.454639049  0.272071976  
[2,]  -0.10615292 -0.61487598  0.01728674 -0.46421995  0.550775374  0.088028646  
[3,]  -0.19089393 -0.22112280 -0.09978650  0.10987953 -0.186701149 -0.178125878  
[4,]  -0.13951311 -0.14784642 -0.13138041 -0.01598041  0.105009202 -0.171612241  
[5,]  -0.04682070 -0.49290700  0.05172379 -0.06592005 -0.405755003 -0.046182873  
[6,]  -0.13858508 -0.15588574  0.03136931  0.78849399  0.514893584 -0.004876079  
[7,]  -0.02862658 -0.02088959  0.05104427  0.01123927 -0.005062728  0.026873555  
[8,]  -0.04321132 -0.10238463  0.00534869 -0.10538116 -0.015312405 -0.029408152  
[9,]  -0.06381638 -0.06445436 -0.05636521  0.02008425 -0.071806372  0.007891374  
[10,] -0.08947075 -0.12850014 -0.07576734  0.01777992  0.021872310 -0.276900583  
[11,] -0.70705994  0.28621233 -0.04885310 -0.16407517  0.077728529  0.541197594  
[12,] -0.31334632  0.14849063 -0.69820134 -0.07580938  0.026674997 -0.476061633  
[13,] -0.46456390  0.10926197  0.67839228 -0.05350023 -0.040805783 -0.506989966  
[...]
```

Decomposição em valores/vectores próprios (cont.)

```
> sqrt(eigen(var(lavagantes))$val)
```

```
[1] 4.4171243 2.1583124 0.9617894 0.7071970 0.6163559 0.4992560 0.4639879  
[8] 0.3848417 0.3362918 0.2500701 0.2060563 0.1770375 0.1405790
```

```
> lav.acp$sdev
```

```
[1] 4.4171243 2.1583124 0.9617894 0.7071970 0.6163559 0.4992560 0.4639879  
[8] 0.3848417 0.3362918 0.2500701 0.2060563 0.1770375 0.1405790
```

```
> eigen(var(lavagantes))$vec
```

```
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]  
[1,] -0.28762060 -0.36935786 -0.08475822 -0.31404094 -0.454639049 -0.272071976  
[2,] -0.10615292 -0.61487598 0.01728674 0.46421995 0.550775374 -0.088028646  
[3,] -0.19089393 -0.22112280 -0.09978650 -0.10987953 -0.186701149 0.178125878  
[...]
```

```
> lav.acp$rot
```

```
      PC1      PC2      PC3      PC4      PC5  
carapace_l 0.28762060 0.36935786 0.08475822 -0.31404094 -0.454639049  
tail_l     0.10615292 0.61487598 -0.01728674 0.46421995 0.550775374  
carapace_w 0.19089393 0.22112280 0.09978650 -0.10987953 -0.186701149  
[...]
```

Nota: Repare-se como alguns vectores próprios diferem num factor de -1 .

Mais propriedades de CPs

Correlações entre CPs e variáveis

A correlação entre a i -ésima variável \vec{x}_i e a j -ésima CP $\mathbf{X}\vec{v}_j$ é:

$$\text{corr}(\vec{x}_i, \mathbf{X}\vec{v}_j) = \sqrt{\lambda_j} \cdot \frac{v_{ij}}{s_i}$$

- s_i — desvio padrão da variável \vec{x}_i
- $\frac{v_{ij}}{\sqrt{\lambda_j}}$ — coeficiente (loading) de \vec{x}_i na CP j
- $\sqrt{\lambda_j}$ — desvio padrão da j -ésima CP

Tem-se $\vec{x}_i = \mathbf{X}\vec{e}_i$, com \vec{e}_i o vector cujo único elemento não nulo é um 1 na posição i (i -ésimo vector da base canónica de \mathbb{R}^p).

A covariância entre as combinações lineares $\mathbf{X}\vec{v}_j$ e $\vec{x}_i = \mathbf{X}\vec{e}_i$ é dada por $\vec{e}_i^t \mathbf{S} \vec{v}_j$, onde \mathbf{S} é a matriz de (co-)variâncias dos dados. Logo,

$$\text{cov}(\vec{x}_i, \mathbf{X}\vec{v}_j) = \frac{\text{cov}(\mathbf{X}\vec{e}_i, \mathbf{X}\vec{v}_j)}{\sqrt{\text{var}(\vec{x}_i) \cdot \text{var}(\mathbf{X}\vec{v}_j)}} = \frac{\vec{e}_i^t \mathbf{S} \vec{v}_j}{s_i \cdot \sqrt{\lambda_j}} = \frac{\lambda_j \vec{e}_i^t \vec{v}_j}{s_i \cdot \sqrt{\lambda_j}} = \sqrt{\lambda_j} \frac{v_{ij}}{s_i} .$$

Interpretação de CPs

Correlações entre CPs e variáveis (lavagantes)

As correlações entre variáveis originais e CPs podem ser úteis para interpretar CPs. Pode-se usar a fórmula dada acima, ou o comando:

```
> round(cor(lavagantes, lav.acp$x), d=2)
```

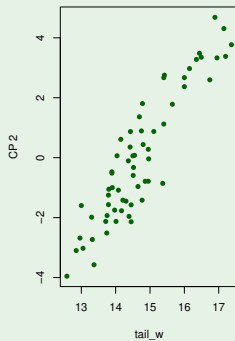
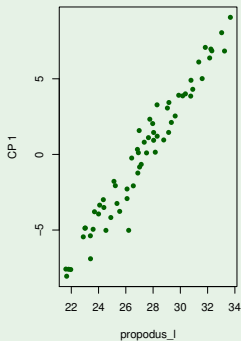
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
carapace_l	0.81	0.51	0.05	-0.14	-0.18	0.09	-0.15	0.06	-0.03	0.02	-0.02	0.01
tail_l	0.31	0.89	-0.01	0.22	0.23	0.03	-0.04	0.06	0.02	-0.01	-0.01	0.00
carapace_w	0.83	0.47	0.09	-0.08	-0.11	-0.09	0.02	-0.02	0.09	-0.14	0.12	0.05
carapace_d	0.78	0.41	0.16	0.01	0.08	-0.11	-0.06	-0.25	-0.33	-0.03	0.03	-0.01
tail_w	0.18	0.91	-0.04	0.04	-0.21	-0.02	0.28	-0.04	0.00	0.02	-0.04	-0.01
areola_l	0.64	0.35	-0.03	-0.58	0.33	0.00	0.11	0.01	0.01	0.03	0.00	0.00
areola_w	0.60	0.21	-0.23	-0.04	-0.01	0.06	0.10	0.09	-0.03	-0.14	0.08	-0.37
rostrum_l	0.50	0.58	-0.01	0.20	-0.02	-0.04	0.03	0.03	0.01	0.49	0.35	0.04
rostrum_w	0.76	0.38	0.15	-0.04	-0.12	0.01	-0.13	-0.03	0.12	-0.02	0.12	-0.40
postorbital_w	0.65	0.45	0.12	-0.02	0.02	-0.23	-0.23	-0.40	0.29	0.08	-0.09	0.01
propodus_l	0.98	-0.19	0.01	0.04	0.01	0.08	0.03	-0.03	0.01	0.00	0.00	0.00
propodus_w	0.87	-0.20	0.42	0.03	0.01	-0.15	0.03	0.08	0.00	0.01	-0.02	0.00
dactyl_l	0.94	-0.11	-0.30	0.02	-0.01	-0.12	-0.01	0.03	-0.01	0.00	-0.01	0.00

A primeira CP está muito fortemente correlacionada com as medições da **tenaz**, em particular **propodus_l**. A segunda CP está fortemente correlacionada com as medições da **cauda**, em particular **tail_w**.

Correlações entre CPs e variáveis (cont.)

Correlações CP/variáveis nos lavagantes

```
> par(mfrow=c(1,2))           <- cria uma "matriz 1x2 de gráficos"  
> plot(lavagantes[,11], lav.acp$x[,1], xlab="propodus_l", ylab="CP 1", pch=16, col="darkgreen")  
> plot(lavagantes[,5], lav.acp$x[,2], xlab="tail_w", ylab="CP 2", pch=16, col="darkgreen")  
> par(mfrow=c(1,1))         <- re-estabelece a janela gráfica original
```

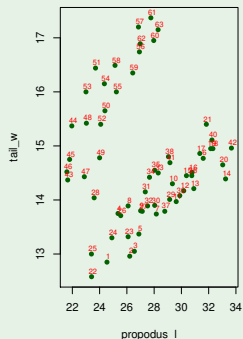
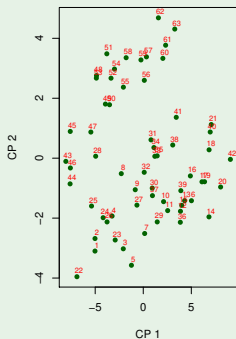


Correlações entre CPs e variáveis (cont.)

Ainda os lavagantes

As fortes correlações sugerem uma nuvem de pontos das duas variáveis originais:

```
> plot(lav.acp$x[,1:2], xlab="CP 1", ylab="CP 2", pch=16, col="darkgreen")  
> text(lav.acp$x[,1:2]+0.2, label=rownames(lavagantes), col="red", cex=0.7)  
> plot(lavagantes[,c(11,5)], xlab="propodus_l", ylab="tail_w", pch=16, col="darkgreen")  
> text(lavagantes[,c(11,5)]+0.1, label=rownames(lavagantes), col="red", cex=0.7)
```



ACP sobre a matriz de correlações

Uma característica pouco simpática da ACP (que a distingue, por exemplo, da regressão linear), é que os resultados duma ACP mudam se houver mudanças de escala diferenciadas nas várias variáveis.

Esta sensibilidade da ACP é natural, dadas as características do critério que se pretende otimizar: a variância.

Para tornar este problema, e sendo a generalidade das mudanças de escala transformações lineares, é hábito normalizar os dados antes de efectuar uma ACP:

$$x_{ij} \longrightarrow z_{ij} = \frac{x_{ij} - \bar{x}_{.j}}{s_j},$$

onde

- x_{ij} é a observação do individuo i na variável j ;
- $\bar{x}_{.j}$ é a média das n observações na variável j ;
- s_j é o desvio padrão das n observações na variável j ;
- z_{ij} é a observação normalizada do individuo i na variável j .

A centragem, na representação em \mathbb{R}^p

Qual o efeito de centrar a matriz \mathbf{X} na nuvem de pontos associada à representação tradicional dos dados, em \mathbb{R}^p ?

A transformação de \mathbf{X} em \mathbf{X}^c apenas altera a média de cada variável, que passa a ser zero. Geometricamente, o centro de gravidade da nuvem de n pontos em \mathbb{R}^p passa a ser a origem, ou seja, há uma translação do centro de gravidade:

$$(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p) \longrightarrow (0, 0, \dots, 0).$$

É habitual fazer a representação das CPs em \mathbb{R}^p com esta centragem (ou seja, a nuvem dos scores ter centro de gravidade na origem).

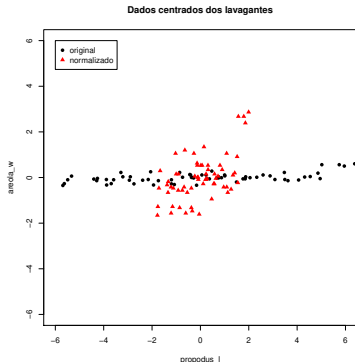
Corresponde a considerar a combinação linear das variáveis centradas (com os vectores de *loadings* usuais).

A normalização, na representação em \mathbb{R}^p

E qual o efeito de normalizar, ou seja (além de centrar), dividir cada variável pelo seu desvio padrão?

Todas as variáveis passam a ter igual variabilidade (variância 1). Logo, a nuvem de pontos em \mathbb{R}^p tende a ser mais esférica.

Ilustremos com os dados (centrados) dos lavagantes, e apenas as duas variáveis de maior, e menor, variância original:



$$s_7^2 = 0.04409$$

$$s_{11}^2 = 10.24217$$

Mudando a forma da nuvem, mudam também as direcções de variabilidade principal.

ACP sobre a matriz de correlações (cont.)

As matrizes de correlações são matrizes de covariâncias duma matriz \mathbf{Z} de dados centrados e reduzidos, de elemento genérico $z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$:

$$\mathbf{R} = \frac{1}{n-1} \mathbf{Z}^t \mathbf{Z} .$$

Assim, uma ACP sobre os dados normalizados é conhecida por ACP sobre a matriz de correlação.

Numa ACP sobre a matriz de correlação,

- as Componentes Principais são combinações lineares dos dados normalizados;
- Os vectores de coeficientes (*loadings*) das combinações lineares são os sucessivos vectores próprios da matriz de correlações \mathbf{R} ;
- As variâncias de sucessivas CPs são dadas pelos valores próprios de \mathbf{R} , cuja soma tem de ser $\text{tr}(\mathbf{R}) = p$.

Não há relação directa entre os resultados duma ou outra destas variantes de ACP.

ACP sobre a matriz de correlações no R

No R, há duas formas alternativas de efectuar uma ACP sobre **R**.

ACP sobre dados normalizados

```
> prcomp(scale(lavagantes)) % ou  
> prcomp(lavagantes, scale=TRUE)
```

Standard deviations:

```
[1] 2.8298571 1.4518966 0.8481395 0.7315674 0.6117634 0.5371346 0.5119344 0.4730480 0.4106900  
[10] 0.3761469 0.3016251 0.2178130 0.1793918
```

Rotation:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
carapace_l	0.3336487	-0.051654918	0.002147496	-0.05337901	0.05903158	-0.25593010	0.13991163
tail_l	0.2328489	-0.455025510	-0.004432513	0.02919494	-0.06389168	0.06642917	-0.32471231
carapace_w	0.3399357	-0.026168964	0.042817387	-0.05649310	0.11876996	-0.18817081	0.02954496
carapace_d	0.3161771	-0.001543245	0.174339992	-0.06927295	-0.01269919	0.02103474	-0.65346959
tail_w	0.1963703	-0.522307992	-0.097172600	0.02943249	0.06817824	-0.29897195	-0.06638706
areola_l	0.2625765	0.014998718	-0.203444780	-0.78727388	-0.41920392	0.00605338	0.19498049
areola_w	0.2320279	0.063340777	-0.813027317	0.19646231	0.26234962	0.17992496	-0.10423247
rostrum_l	0.2559610	-0.260192772	0.122258123	0.50436942	-0.58565962	0.13677260	0.30765231
rostrum_w	0.3122279	0.011301755	0.084409773	0.06116672	0.43328915	-0.24980467	0.49425052
postorbital_w	0.2883485	-0.080276403	0.361940139	-0.14548391	0.36223013	0.71927271	0.11234877
propodus_l	0.2741268	0.405235606	0.006549232	0.13377738	-0.13020525	-0.02606551	-0.05259459
propodus_w	0.2474141	0.398376708	0.281998129	0.09065523	0.00717611	-0.33417966	-0.19598386
dactyl_l	0.2740158	0.339649079	-0.152524450	0.15373369	-0.22361974	0.24824297	0.03271971
[...]							

As duas variantes de ACP

Os resultados duma e outra variante de ACP não são directamente comparáveis.

As duas variantes da ACP - dados lavagantes

```
> lav.acpR <- prcomp(lavagantes, scale=TRUE)
> summary(lav.acpR)
```

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Std.dev	2.830	1.4519	0.84814	0.73157	0.61176	0.53713	0.51193	0.47305	0.41069	0.37615	0.3016	0.21781	0.17939
Prp.Var	0.616	0.1621	0.05533	0.04117	0.02879	0.02219	0.02016	0.01721	0.01297	0.01088	0.0070	0.00365	0.00248
Cum.Prp	0.616	0.7782	0.83350	0.87466	0.90345	0.92565	0.94581	0.96302	0.97599	0.98688	0.9939	0.99752	1.00000

```
> summary(lav.acp)
```

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Std.dev	4.4171	2.1583	0.96179	0.70720	0.61636	0.49926	0.46399	0.38484	0.33629	0.25007	0.20606	0.17704	0.1406
Prp.Var	0.7265	0.1734	0.03444	0.01862	0.01415	0.00928	0.00802	0.00551	0.00421	0.00233	0.00158	0.00117	0.0007
Cum.Prp	0.7265	0.9000	0.93440	0.95302	0.96716	0.97645	0.98446	0.98998	0.99419	0.99652	0.99810	0.99926	1.0000

Em geral, numa ACP sobre a matriz de correlações são precisas mais CPs para alcançar uma mesma proporção da inércia total explicada.

As duas variantes de ACP (cont.)

Mudam também os vectores de *loadings* (vectores próprios de **S** e **R** são diferentes), bem como os vectores de *scores* a que dão origem.

Vejamos as *correlações* entre os dois tipos de CPs:

As duas variantes da ACP - dados lavagantes (cont.)

```
> round(cor(lav.acp$x, lav.acpR$x), d=2)
```

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
PC1	0.89	0.44	0.00	0.06	-0.07	-0.01	-0.02	-0.09	-0.04	0.04	0.01	0.04	0.03
PC2	0.44	-0.88	-0.03	-0.10	0.05	-0.05	-0.04	-0.01	-0.04	0.03	-0.05	-0.02	-0.04
PC3	0.05	0.05	0.53	-0.09	0.24	-0.42	-0.16	0.56	0.27	0.18	-0.13	0.07	0.07
PC4	-0.04	-0.10	0.19	0.79	0.09	0.11	-0.36	-0.05	-0.25	0.26	0.17	0.06	0.08
PC5	0.00	0.02	-0.03	-0.38	-0.37	0.34	-0.28	0.45	-0.42	0.31	0.21	-0.01	0.08
PC6	-0.05	-0.03	-0.21	0.02	-0.05	-0.26	0.11	0.01	-0.31	-0.05	-0.33	0.48	0.66
PC7	-0.02	-0.06	-0.26	-0.01	-0.26	-0.33	-0.12	-0.10	0.45	0.18	0.64	0.16	0.21
PC8	-0.05	0.04	-0.28	0.14	-0.27	-0.53	0.17	0.04	-0.22	0.45	-0.24	-0.15	-0.44
PC9	0.01	-0.02	0.10	-0.04	0.25	0.29	0.61	-0.08	0.10	0.64	0.08	-0.01	0.21
PC10	0.02	-0.10	0.22	0.27	-0.54	0.21	0.37	0.25	0.20	-0.17	-0.05	0.46	-0.24
PC11	0.05	-0.04	-0.04	0.27	-0.24	-0.05	0.26	0.32	0.04	-0.23	0.03	-0.68	0.40
PC12	-0.07	-0.03	0.33	-0.11	-0.47	0.07	-0.27	-0.45	0.26	0.23	-0.42	-0.18	0.21
PC13	-0.03	-0.02	0.56	-0.16	-0.13	-0.32	0.25	-0.30	-0.46	-0.15	0.38	-0.03	0.00

As duas variantes de ACP (cont.)

Correlações das CPs (dados normalizados) com as variáveis originais:

ACP sobre matriz de correlações - dados lavagantes

```
> round(cor(lavagantes, lav.acpR$x), d=2)
```

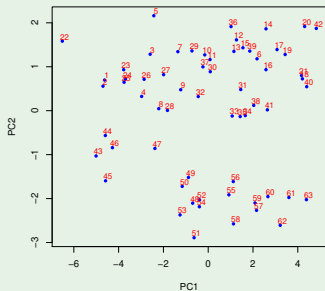
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
carapace_l	0.94	-0.07	0.00	-0.04	0.04	-0.14	0.07	-0.11	-0.03	-0.06	-0.24	0.05	-0.04
tail_l	0.66	-0.66	0.00	0.02	-0.04	0.04	-0.17	0.05	-0.24	0.19	-0.01	0.00	0.00
carapace_w	0.96	-0.04	0.04	-0.04	0.07	-0.10	0.02	-0.10	0.08	0.07	-0.03	-0.16	0.06
carapace_d	0.89	0.00	0.15	-0.05	-0.01	0.01	-0.33	0.07	-0.02	-0.24	0.02	-0.02	0.00
tail_w	0.56	-0.76	-0.08	0.02	0.04	-0.16	-0.03	-0.18	0.17	0.00	0.12	0.07	-0.01
areola_l	0.74	0.02	-0.17	-0.58	-0.26	0.00	0.10	0.09	0.02	0.01	0.04	0.01	0.00
areola_w	0.66	0.09	-0.69	0.14	0.16	0.10	-0.05	0.14	0.08	0.01	-0.02	0.00	0.00
rostrum_l	0.72	-0.38	0.10	0.37	-0.36	0.07	0.16	0.14	0.07	-0.05	-0.01	-0.01	0.01
rostrum_w	0.88	0.02	0.07	0.04	0.27	-0.13	0.25	0.12	-0.16	-0.09	0.10	0.00	-0.01
postorbital_w	0.82	-0.12	0.31	-0.11	0.22	0.39	0.06	-0.01	0.10	0.04	-0.01	0.03	0.00
propodus_l	0.78	0.59	0.01	0.10	-0.08	-0.01	-0.03	-0.08	-0.05	0.04	0.02	0.10	0.12
propodus_w	0.70	0.58	0.24	0.07	0.00	-0.18	-0.10	0.16	0.12	0.16	0.02	0.02	-0.07
dactyl_l	0.78	0.49	-0.13	0.11	-0.14	0.13	0.02	-0.26	-0.09	-0.01	0.07	-0.03	-0.08

- Em relação à ACP sobre dados originais, não só mudam as correlações entre CPs e variáveis, como também as interpretações possíveis.
- CP1 é agora essencialmente uma medida da **dimensão geral do animal**.
- CP2, mais difícil de interpretar, mas **contrasta dimensões de caudas e tenazes**.

Primeiro plano principal – dados normalizados

ACP sobre matriz de correlações - dados lavagantes (cont.)

- ```
> plot(lav.acpR$x[,1:2], col="blue", pch=16, cex=0.8)
> text(lav.acpR$x[,1:2]+0.1, label=rownames(lavagantes), col="red")
```

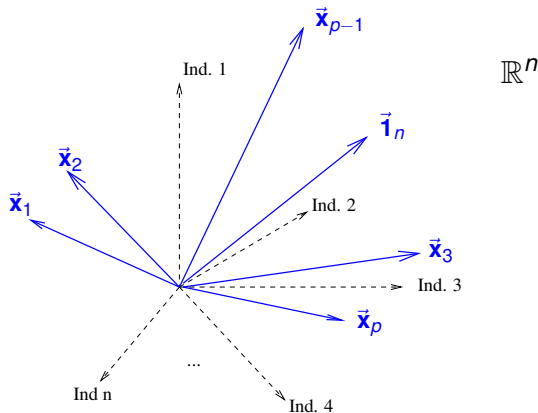


Simplificando: a CP 1 ordena por tamanho geral do organismo, e a CP 2 faz a separação de morfometria relacionada com o sexo.

# A representação em $\mathbb{R}^n$ , o espaço das variáveis

**Recordar:** Representação alternativa duma matriz de dados  $\mathbf{X}$ , no espaço das variáveis:

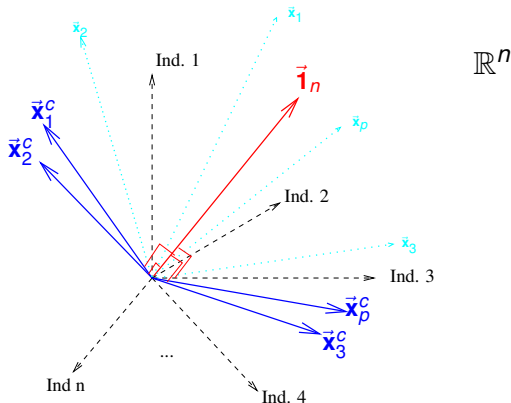
- cada eixo corresponde a um indivíduo observado;
- cada vector corresponde a uma variável.



# Variáveis centradas no espaço das variáveis

A representação mais interessante no espaço das variáveis é a das **variáveis centradas**, porque os conceitos geométricos introduzidos pelo habitual produto interno em  $\mathbb{R}^n$  têm interpretações estatísticas.

**Centrar** as colunas de  $\mathbf{X}$  corresponde a tornar os vectores que representam as variáveis ortogonais ao vector  $\vec{\mathbf{1}}_n$  dos  $n$  uns (a soma de qualquer coluna de  $\mathbf{X}^c$  é zero):



# Geometria e estatística no espaço das variáveis

O elemento genérico da matriz centrada dos dados,  $\mathbf{X}^c$ , é:

$$x_{ij}^c = x_{ij} - \bar{x}_{.j},$$

- $x_{ij}$  indica a observação do  $i$ -ésimo indivíduo na variável  $j$ ;
- $\bar{x}_{.j}$  indica a média das  $n$  observações na variável  $j$ .

Logo,

- a **norma usual** dum coluna  $\vec{x}_j^c$  de  $\mathbf{X}^c$  é proporcional ao **desvio padrão** dessa

variável:  $\|\vec{x}_j^c\| = \sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_{.j})^2} = \sqrt{n-1} s_j$ .

- o **produto interno** usual de duas diferentes colunas de  $\mathbf{X}^c$  é proporcional à **covariância** das respectivas variáveis:  $\langle \vec{x}_j^c, \vec{x}_k^c \rangle = (\vec{x}_j^c)^t \vec{x}_k^c = \sum_{i=1}^n (x_{ij} - \bar{x}_{.j})(x_{ik} - \bar{x}_{.k}) = (n-1) \text{cov}_{j,k}$ .
- o **coseno do ângulo** entre os vectores representando duas diferentes colunas de  $\mathbf{X}^c$  é o **coeficiente de correlação** das respectivas variáveis:

$$\cos \theta = \frac{\langle \vec{x}_j^c, \vec{x}_k^c \rangle}{\|\vec{x}_j^c\| \cdot \|\vec{x}_k^c\|} = \frac{(n-1) \cdot \text{cov}_{j,k}}{\sqrt{n-1} s_j \cdot \sqrt{n-1} s_k} = \frac{\text{cov}_{j,k}}{s_j \cdot s_k} = r_{j,k}$$

- Vectores centrados **ortogonais** correspondem a variáveis **não correlacionadas**.

# Intepretação da ACP no espaço das variáveis

A representação dos dados no espaço das variáveis ( $\mathbb{R}^n$ ) associa cada variável a um vector. Combinações lineares de variáveis são combinações lineares de vectores, logo novos vectores. As CPs também são representadas por vectores em  $\mathbb{R}^n$ .

Para vectores centrados, o quadrado do comprimento do vector é proporcional à variância da variável respectiva.

O critério da ACP (maximizar variância) corresponde a procurar combinações lineares dos vectores de comprimento máximo (com soma 1 de quadrados dos coeficientes).

É geometricamente intuitivo que variáveis com variância muito maior que as restantes tenham grande influência na definição da primeira CP (“dominam a primeira CP”).

# ACP é sensível a mudanças (diferenciadas) de escala

Qualquer transformação linear (afim) numa variável ( $x \rightarrow a + bx$ ), como são a maioria das mudanças de unidades de medida, **re-escala o vector centrado que a representa em  $\mathbb{R}^n$ , mantendo a direcção:**

- **constantes aditivas  $a$**  desaparecem na centragem, logo **não alteram o vector centrado correspondente em  $\mathbb{R}^n$ .**
- **constantes multiplicativas  $b$ :**
  - ▶ preservam a direcção do vector representativo ( $\vec{x}_j \rightarrow b\vec{x}_j$ );
  - ▶ mudam o sentido se  $b < 0$ ;
  - ▶ esticam o vector se  $|b| > 1$ , pois  $\|b\vec{x}_j\| = |b| \|\vec{x}_j\|$ ;
  - ▶ encolhem o vector se  $|b| < 1$ .

Logo, **o critério da ACP é sensível a mudanças de escalas diferenciadas nas  $p$  variáveis.**

## Interpretação de CPs em $\mathbb{R}^n$ (cont.)

E qual o efeito de **normalizar** as variáveis na representação em  $\mathbb{R}^n$ ?

A **normalização** dos dados (ACP sobre a matriz de correlações) **torna igual o tamanho de todos os vectores representativos das variáveis centradas**.

Logo,

- deixa de ser possível que haja vectores mais compridos do que outros, a condicionar as primeiras CPs;
- o que vai ser fundamental para determinar a direcção de maior comprimento é o **padrão de correlações** entre as variáveis, ou seja, a sua posição angular relativa;
- grupos de variáveis fortemente correlacionadas entre si tendem a “atrair” a primeira CP sobre os dados normalizados.



# Ainda a ACP sobre matriz de correlações

Em termos geométricos, a normalização das variáveis:

- Em  $\mathbb{R}^n$ , re-dimensiona cada um dos  $p$  vectores, que ficam com comprimento (norma) comum.
- Em  $\mathbb{R}^p$ , estica ou contrai cada eixo, com factores de alteração das escalas diferenciados para cada eixo. Muda a forma da nuvem de pontos.

Observações:

- A variabilidade total é  $\text{tr}(\mathbf{R}) = p$  (o número de variáveis).
- A correlação entre a variável  $\tilde{\mathbf{x}}_i$  e a  $j$ -ésima CP é agora  $\sqrt{\lambda_j^R} v_{ij}^R$ .
- Por vezes, os coeficientes das componentes numa ACP sobre a matriz de correlações são re-escalados de forma a que  $\tilde{\mathbf{v}}_j^t \tilde{\mathbf{v}}_j = \lambda_j$ . Nesse caso, os coeficientes da combinação linear são as correlações entre a variável e a CP.

# Advertências sobre ACPs (em geral)

- A redução da dimensionalidade associada à ACP não significa redução no número de *variáveis originais* com que se trabalha: cada CP é combinação linear de *todas* as variáveis observadas.
- É frequente procurar *interpretar* cada CP ignorando as variáveis cujos coeficientes (*loadings*) na combinação linear que define a CP são “próximos de zero”. Isto *pode induzir em erro*, e convém *utilizar informação complementar para validar as interpretações baseadas nos coeficientes*.
- Outra prática frequente, mas discutível, em ACP é a *rotação* das CPs: modificam-se os coeficientes da combinação linear para os aproximar de zero ou um, visando “simplificar a interpretação”. Mas esse objectivo pode ser *ilusório* (como vimos) e *sacrifica a optimalidade* das soluções.
- Alguns autores também chamam *componentes principais* aos vectores próprios de **S** ou **R** (vectores de *loadings*), gerando confusão.
- *Não* faz sentido que qualquer das variáveis originais seja uma variável *qualitativa (categórica)*.

## Outra introdução à ACP

A Análise em Componentes Principais pode ainda ser introduzida através do resultado fundamental da Teoria de Matrizes: a **Decomposição em Valores** (e Vectores) **Singulares**.

Tal como a Decomposição Espectral, a **DVS** envolve a **factorização numa matriz no produto de 3 matrizes**, das quais a central é uma matriz **diagonal** e as outras duas têm **colunas ortonormadas**. Mas:

- Enquanto a Decomposição Espectral apenas é válida para matrizes simétricas, **a DVS é válida para qualquer matriz**, incluindo matrizes rectangulares.
- **As três matrizes envolvidas na DVS são diferentes** e têm, em geral, dimensões diferentes.
- A DVS e a Decomposição Espectral **coincidem no caso de matrizes** (simétricas) **semi-definidas positivas**.

# Decomposição em Valores Singulares

## Decomposição em Valores Singulares (DVS)

Seja  $\mathbf{Y}_{n \times p}$  uma matriz genérica. É sempre possível factorizar  $\mathbf{Y}$  da seguinte forma:

$$\mathbf{Y} = \mathbf{W}\mathbf{\Delta}\mathbf{V}^t \iff \mathbf{Y} = \sum_{i=1}^p \delta_i \vec{\mathbf{w}}_i \vec{\mathbf{v}}_i^t,$$

onde

$\mathbf{\Delta}_{p \times p}$  matriz diagonal

$\mathbf{V}_{p \times p}$  matriz com colunas ortonormadas ( $\mathbf{V}^t\mathbf{V} = \mathbf{I}_p$ )

$\mathbf{W}_{n \times p}$  matriz com colunas ortonormadas ( $\mathbf{W}^t\mathbf{W} = \mathbf{I}_p$ )

$\delta_i$  elementos diagonais de  $\mathbf{\Delta}$  (valores singulares de  $\mathbf{Y}$ )

$\vec{\mathbf{w}}_i$  colunas de  $\mathbf{W}$  (vectores singulares esquerdos de  $\mathbf{Y}$ )

$\vec{\mathbf{v}}_i$  colunas de  $\mathbf{V}$  (vectores singulares direitos de  $\mathbf{Y}$ )

Admite-se que os valores singulares  $\delta_i$  estão por ordem decrescente.

## Observações sobre a DVS $\mathbf{Y} = \mathbf{W}\Delta\mathbf{V}^t$

- $\mathbf{Y}^t$  tem Decomposição em Valores Singulares  $\mathbf{Y}^t = \mathbf{V}\Delta\mathbf{W}^t$ .
- $\mathbf{Y}^t\mathbf{Y} = \mathbf{V}\Delta^2\mathbf{V}^t$  é uma Decomposição Espectral de  $\mathbf{Y}^t\mathbf{Y}$ . Logo,  $\mathbf{V}$  é matriz cujas colunas são **conjunto o.n. de vectores próprios de  $\mathbf{Y}^t\mathbf{Y}$** .
- $\mathbf{W}$  é a matriz análoga, de **vectores próprios de  $\mathbf{Y}\mathbf{Y}^t = \mathbf{W}\Delta^2\mathbf{W}^t$** .
- $\Delta$  é a matriz das **raízes quadradas dos valores próprios de  $\mathbf{Y}^t\mathbf{Y}$**  (iguais aos valores próprios não-nulos de  $\mathbf{Y}\mathbf{Y}^t$ ).
- A DVS duma matriz é sempre possível, mas não é única (pelo menos a troca de sinal nos pares de vectores).
- Se  $\mathbf{Y}$  é de **característica** (número máximo de colunas linearmente independentes)  $r < p$ , tem-se  $\delta_i = 0$  para  $i > r$ .
- Se  $\mathbf{Y}$  é de **característica**  $r < p$ , as  $p-r$  parcelas finais no somatório  $\mathbf{Y} = \sum_{i=1}^p \delta_i \vec{\mathbf{w}}_i \vec{\mathbf{v}}_i^t$  são matrizes nulas. Corresponde a ignorar as  $p-r$  colunas finais de  $\mathbf{V}$  e  $\mathbf{W}$ , e as  $p-r$  linhas e colunas finais de  $\Delta$ .  
A DVS assim definida é designada **Thin SVD** em inglês.

# A DVS e a ACP

A ACP é uma **Decomposição em Valores Singulares** da matriz centrada dos dados  $\mathbf{X}^c$ , a dividir por  $\sqrt{n-1}$ , (ou  $\sqrt{n}$ , consoante a convenção usada para criar a matriz de covariâncias):

$$\frac{1}{\sqrt{n-1}} \mathbf{X}^c = \mathbf{U} \mathbf{\Delta} \mathbf{V}^t,$$

porque:

$\mathbf{V}$  - matriz cujas **colunas** são os **vetores próprios** de  $\mathbf{S} = \frac{1}{n-1} \mathbf{X}^{cT} \mathbf{X}^c$ , ou seja, os **loadings** das CPs.

$\mathbf{\Delta}$  - matriz cujos **elementos diagonais** são as **raízes quadradas** dos **valores próprios** de  $\mathbf{S}$ , ou seja, os **desvios padrão das CPs**;

$\mathbf{X}^c \mathbf{V} = \sqrt{n-1} \mathbf{U} \mathbf{\Delta}$  - matriz cujas **colunas** são os **scores centrados** de cada indivíduo em cada CP.

$\mathbf{U} = \frac{1}{\sqrt{n-1}} \mathbf{X}^c \mathbf{V} \mathbf{\Delta}^{-1}$  - matriz dos **vetores singulares esquerdos**, que corresponde aos **vetores de scores normalizados**.

## DVS e ACP (cont.)

Confirmamos, efectuando a DVS da matriz  $\frac{1}{\sqrt{n-1}} \mathbf{X}^c$  no R, para os dados lavagantes.

A **centragem** dum matriz de dados pode fazer-se da seguinte forma:

```
> lav.centrado <- scale(lavagantes, scale=FALSE)
```

O comando `scale` pode fazer **simultaneamente** a **centragem** (subtracção da média) e **divisão pelo desvio padrão** das colunas dum matriz.

Cada uma destas operações é controlada por um argumento, respectivamente `center` e `scale`.

Por omissão, estes argumentos são TRUE. Qualquer das operações pode ser omitida dando ao correspondente argumento o valor **FALSE**.

No R uma Decomposição em Valores Singulares obtém-se com o comando `svd`.

# ACP e DVS (cont.)

## DVS com os lavagantes

```
> svd(lav.centrado/sqrt(62))
```

```
$d
[1] 4.4171243 2.1583124 0.9617894 0.7071970 0.6163559 0.4992560 0.4639879
[8] 0.3848417 0.3362918 0.2500701 0.2060563 0.1770375 0.1405790
```

```
$u
 [,1] [,2] [,3] [,4] [,5]
[1,] -0.144379990 -0.182396510 -0.123645871 0.1059842750 0.070557452
[2,] -0.144331125 -0.157779185 0.254967864 0.1172558607 0.146926297
[3,] -0.059725146 -0.177923620 -0.059333869 0.1101757182 0.113206688
[4,] -0.093246935 -0.113657051 0.014976742 0.0804924915 -0.069971697
[5,] -0.035380664 -0.210254166 -0.097758921 -0.1206499751 -0.146049537
[...]
```

```
$v
 [,1] [,2] [,3] [,4] [,5] [,6]
[1,] 0.28762060 0.36935786 0.08475822 -0.31404094 -0.454639049 0.272071976
[2,] 0.10615292 0.61487598 -0.01728674 0.46421995 0.550775374 0.088028646
[3,] 0.19089393 0.22112280 0.09978650 -0.10987953 -0.186701149 -0.178125878
[4,] 0.13951311 0.14784642 0.13138041 0.01598041 0.105009202 -0.171612241
[5,] 0.04682070 0.49290700 -0.05172379 0.06592005 -0.405755003 -0.046182873
[...]
```

**Atenção:** As componentes  $\$u$  e  $\$v$  são, respectivamente, as matrizes  $U$  e  $V$ .  
A componente  $\$d$  é um **vector**, com os **elementos diagonais** da matriz  $\Delta$ .



# ACP e DVS (cont.)

## SVD nos lavagantes (cont.)

```
> DVS <- svd(lav.centrado/sqrt(62))
> U <- DVS$u
> D <- diag(DVS$d) <- creates a diagonal matrix from vector DVS$d
> U %*% D * sqrt(62)
```

```
 [,1] [,2] [,3] [,4] [,5] [,6]
[1,] -5.0216041 -3.09975004 -0.93638716 0.590170762 0.34242883 -0.311295721
[2,] -5.0199046 -2.68138921 1.93090666 0.652936303 0.71306147 2.411219117
[3,] -2.0772687 -3.02373521 -0.44934354 0.613510708 0.54941375 -0.365822245
[...]
```

```
[62,] 1.5767872 4.68339718 -0.49231884 0.246787192 -0.11313707 0.138658304
[63,] 3.2782407 4.30830749 0.15373020 -0.562657698 -0.73379507 0.200035217
[...]
```

O comando `prcomp` usa a DVS:

```
> prcomp(lavagantes)$x
```

```
 PC1 PC2 PC3 PC4 PC5 PC6
1 -5.0216041 -3.09975004 -0.93638716 0.590170762 0.34242883 -0.311295721
2 -5.0199046 -2.68138921 1.93090666 0.652936303 0.71306147 2.411219117
3 -2.0772687 -3.02373521 -0.44934354 0.613510708 0.54941375 -0.365822245
[...]
```

```
62 1.5767872 4.68339718 -0.49231884 0.246787192 -0.11313707 0.138658304
63 3.2782407 4.30830749 0.15373020 -0.562657698 -0.73379507 0.200035217
[...]
```

# Relação ACP e DVS

Uma **matriz de dados**  $\mathbf{X}_{n \times p}$  é representada por uma nuvem de  $n$  pontos em  $\mathbb{R}^p$  ou, equivalentemente, um feixe de  $p$  vectores em  $\mathbb{R}^n$ .

Se  $\mathbf{Y}_{n \times p}$  fôr matriz de igual dimensão, mas **característica (rank)**  $r < p$ , a nuvem de  $n$  pontos correspondente estará num **subespaço de dimensão  $r$**  de  $\mathbb{R}^p$ . Analogamente, o seu feixe de  $p$  vectores gera um **subespaço de dimensão  $r$**  em  $\mathbb{R}^n$ .

## Problema geométrico

Identificar a matriz  $\mathbf{Y}_{n \times p}$ , de característica  $r$ , tal que os respectivos  $n$  pontos em  $\mathbb{R}^p$  minimizem a soma de quadrados das distâncias em relação aos  $n$  pontos associados à matriz de dados original  $\mathbf{X}_{n \times p}$ :

$$\sum_{i=1}^n \sum_{j=1}^p (x_{ij} - y_{ij})^2 .$$

O critério acima **minimiza igualmente** a soma de quadrados dos elementos das  $p$  colunas de  $\mathbf{X}$  e  $\mathbf{Y}$ , pelo que simultaneamente o feixe de  $p$  vectores de  $\mathbf{Y}$  estará “globalmente mais próximo” dos  $p$  vectores de  $\mathbf{X}$ .

# Teorema de Eckart-Young

## Teorema (Eckart-Young)

Seja  $\mathbf{X}_{n \times p}$  uma matriz de característica  $p$ . A matriz  $\mathbf{Y}_{n \times p}$  de característica  $r < p$  que minimiza a distância matricial usual  $\|\mathbf{X} - \mathbf{Y}\| = \sqrt{\sum_i \sum_j (x_{ij} - y_{ij})^2}$ , obtém-se da

seguinte forma:

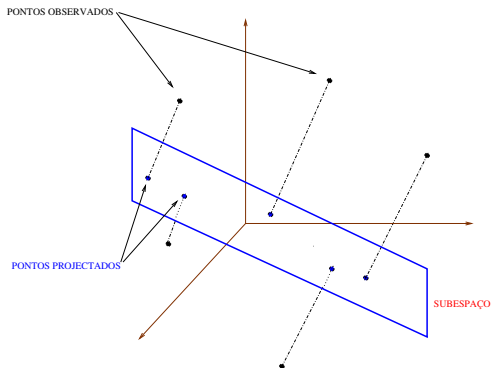
- Seja  $\mathbf{X} = \mathbf{W}\mathbf{\Delta}\mathbf{V}^t$  uma decomposição em valores singulares de  $\mathbf{X}$ .
- Sejam  $\mathbf{W}_r, \mathbf{V}_r$ , as matrizes constituídas pelas  $r$  colunas de  $\mathbf{W}$  e  $\mathbf{V}$ , respectivamente, associadas aos maiores valores singulares.
- Seja  $\mathbf{\Delta}_r$  a matriz diagonal de dimensão  $r \times r$  resultante de reter apenas os  $r$  maiores valores singulares em  $\mathbf{\Delta}$ .
- Então  $\mathbf{Y} = \mathbf{W}_r\mathbf{\Delta}_r\mathbf{V}_r^t$  (e esta é uma DVS de  $\mathbf{Y}$ ).

Nota 1: Como  $\mathbf{X} = \sum_{i=1}^p \delta_i \bar{\mathbf{w}}_i \bar{\mathbf{v}}_i^t$  é a DVS de  $\mathbf{X}$ ,  $\mathbf{Y}$  é a matriz que se obtém retendo apenas as  $r$  primeiras parcelas da DVS de  $\mathbf{X}$ :  $\mathbf{Y} = \sum_{i=1}^r \delta_i \bar{\mathbf{w}}_i \bar{\mathbf{v}}_i^t$ .

Nota 2: Assim, a ACP (a DVS de  $\mathbf{X}^c$ ) identifica, simultaneamente em  $\mathbb{R}^p$  e em  $\mathbb{R}^n$ , os subespaços de dimensão  $r$  onde a representação dos dados originais é o mais fidedigna possível, no sentido de ser globalmente mais próxima dos valores originais.

# As projecções em $\mathbb{R}^p$ e em $\mathbb{R}^n$

Para ambas as representações dos dados em  $\mathbf{X}^c$ , a ACP é a solução do problema de determinar o subespaço de dimensão  $r$  tal que a projecção ortogonal dos dados nesse subespaço minimiza a soma de quadrados das distâncias (na perpendicular) entre pontos originais e pontos projectados.



# Biplots

- Intimamente relacionada com a Decomposição em Valores Singulares numa matriz centrada de dados (logo, com uma ACP).
- Ideia fundamental do *biplot*: obter uma boa **representação simultânea (aproximada)** dos indivíduos e das variáveis (daí o prefixo *bi-*), **em baixa dimensão**.
- Nessa representação simultânea as **principais características estatísticas** numa ACP serão visíveis em termos geométricos.

## Biplots (cont.)

- Seja  $\mathbf{X}^c$  matriz centrada de dados, com SVD:  $\frac{1}{\sqrt{n-1}}\mathbf{X}^c = \mathbf{U}\mathbf{\Delta}\mathbf{V}^t$ .
- Definindo:

$$\begin{aligned}\mathbf{G} &= \mathbf{U} \\ \mathbf{H} &= \mathbf{V}\mathbf{\Delta}\end{aligned}$$

tem-se:  $\frac{1}{\sqrt{n-1}}\mathbf{X}^c = \mathbf{G}\mathbf{H}^t$ .

- Se  $\mathbf{X}^c$  é de característica (rank)  $p$ ,
  - ▶  $\mathbf{G}$  é  $n \times p$ , e existe correspondência entre linhas de  $\mathbf{G}$  e indivíduos.
  - ▶  $\mathbf{H}$  é  $p \times p$ , e existe correspondência entre linhas de  $\mathbf{H}$  e variáveis.
- As linhas de  $\mathbf{G}$  ( $g_{[i]}^t$ ) e de  $\mathbf{H}$  ( $h_{[j]}^t$ ) são marcadores de, respectivamente, indivíduos e variáveis, mas vivem no mesmo espaço ( $\mathbb{R}^p$ ) e podem ser representadas em simultâneo.
- O produto interno do marcador do indivíduo  $i$  e da variável  $j$  é o valor desse indivíduo nessa variável (centrada e a dividir por  $\sqrt{n-1}$ ):

$$g_{[i]}^t h_{[j]}^t = \frac{1}{\sqrt{n-1}} x_{ij}^c.$$

# Os marcadores de variáveis

Consideremos as **propriedades dos marcadores de variáveis**, que são **vetores de  $\mathbb{R}^p$** . Os produtos internos de marcadores de variáveis são:

$$\mathbf{H}\mathbf{H}^t = (\mathbf{V}\mathbf{\Delta})(\mathbf{V}\mathbf{\Delta})^t = \mathbf{V}\mathbf{\Delta}^2\mathbf{V}^t = \mathbf{S},$$

pois  $\mathbf{S} = \frac{1}{n-1}\mathbf{X}^c t \mathbf{X}^c = (\mathbf{U}\mathbf{\Delta}\mathbf{V}^t)^t (\mathbf{U}\mathbf{\Delta}\mathbf{V}^t) = \mathbf{V}\mathbf{\Delta}\mathbf{U}^t \mathbf{U}\mathbf{\Delta}\mathbf{V}^t = \mathbf{V}\mathbf{\Delta}^2\mathbf{V}^t$ .

Tendo em conta que **as linhas de  $\mathbf{H}$  são vetores de  $\mathbb{R}^p$** , **marcadores de variáveis**, tem-se:

- O produto interno entre cada par de marcadores de variáveis é a **covariância** entre as variáveis correspondentes.
- A **norma** de cada marcador de variável é o **desvio padrão** da variável correspondente.
- O **cosseno do ângulo** entre cada par de marcadores de variáveis é o **coeficiente de correlação** entre as variáveis correspondentes.

# Distâncias de Mahalanobis

Para compreender a leitura dos **marcadores de indivíduos**, é necessário o conceito de **distância (ao quadrado) de Mahalanobis**.

## Distâncias de Mahalanobis

Seja  $\mathbf{X}_{n \times p}$  uma matriz de dados, com matriz de covariâncias associada  $\mathbf{S}$ , linha genérica  $\vec{\mathbf{x}}_{[i]}$  e centro de gravidade  $\vec{\mathbf{m}}$ . Define-se:

- Distância de Mahalanobis do indivíduo  $i$  ao centro:

$$\|\vec{\mathbf{x}}_{[i]} - \vec{\mathbf{m}}\|_{\mathbf{S}^{-1}}^2 = (\vec{\mathbf{x}}_{[i]} - \vec{\mathbf{m}})^t \mathbf{S}^{-1} (\vec{\mathbf{x}}_{[i]} - \vec{\mathbf{m}}) .$$

- Distância de Mahalanobis entre os indivíduos  $i$  e  $j$ :

$$\|\vec{\mathbf{x}}_{[i]} - \vec{\mathbf{x}}_{[j]}\|_{\mathbf{S}^{-1}}^2 = (\vec{\mathbf{x}}_{[i]} - \vec{\mathbf{x}}_{[j]})^t \mathbf{S}^{-1} (\vec{\mathbf{x}}_{[i]} - \vec{\mathbf{x}}_{[j]}) .$$

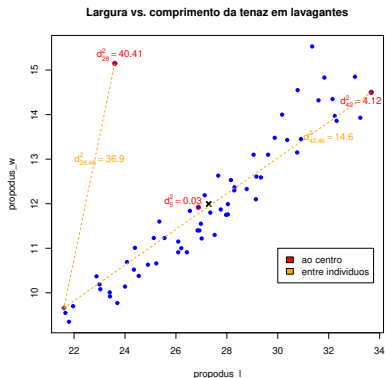
As distâncias Euclidianas usuais são dadas por expressões análogas, mas com a matriz identidade  $\mathbf{I}$  em vez da matriz  $\mathbf{S}^{-1}$ .



# Distâncias de Mahalanobis (cont.)

As distâncias (ao quadrado) de Mahalanobis levam em conta a forma da nuvem de pontos em  $\mathbb{R}^p$  (o padrão de covariâncias entre variáveis). São úteis para identificar *outliers* multivariados.

Eis a nuvem em  $\mathbb{R}^2$  apenas das variáveis `propodus_l` e `propodus_w`:



Os valores numéricos são distâncias de Mahalanobis.

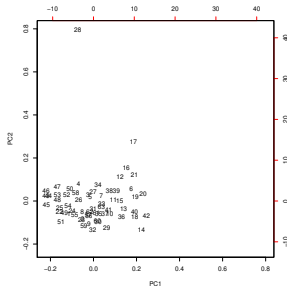
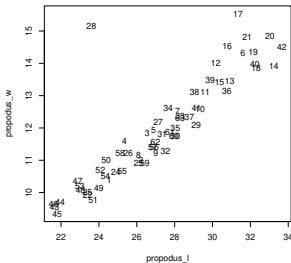
*Num biplot, a distância euclidiana entre marcadores de pontos é igual à distância de Mahalanobis entre indivíduos.*

# Os marcadores de indivíduos

- A distância euclidiana entre cada par de linhas de **G** é proporcional à distância de Mahalanobis entre os indivíduos correspondentes:

$$\|\vec{g}_{[i]} - \vec{g}_{[j]}\|^2 = (\vec{x}_{[i]} - \vec{x}_{[j]})^t \mathbf{S}^{-1} (\vec{x}_{[i]} - \vec{x}_{[j]}) = \|\vec{x}_{[i]} - \vec{x}_{[j]}\|_{\mathbf{S}^{-1}}^2.$$

Eis a nuvem só com as variáveis `propodus_l` e `propodus_w` e os seus marcadores de indivíduos no *biplot*:



No *biplot*, a distância euclidiana entre pontos é a distância de Mahalanobis entre indivíduos.

## Biplots (cont.)

A visualização dum *biplot* exige que se reduza a representação a um espaço a  $k = 2$  ou  $k = 3$  dimensões.

Tal é feito retendo apenas as coordenadas dos marcadores associadas às duas (ou três) primeiras dimensões:

- $\mathbf{G}^{(k)}$  submatriz  $n \times k$  com as  $k$  primeiras colunas de  $\mathbf{G}$ .
- $\mathbf{H}^{(k)}$  submatriz  $p \times k$  com as  $k$  primeiras colunas de  $\mathbf{H}$ .

As linhas de  $\mathbf{G}^{(k)}$  e  $\mathbf{H}^{(k)}$  são **marcadores de indivíduos e variáveis** e:

$$\frac{1}{\sqrt{n-1}} \tilde{\mathbf{X}}^c = \mathbf{G}^{(k)} \mathbf{H}^{(k)t}$$

é a melhor aproximação, de característica  $k$ , a  $\frac{1}{\sqrt{n-1}} \mathbf{X}^c$  (Teorema de Eckart-Young).

## Biplots (cont.)

Tomando  $k = 2$ , obtemos representação gráfica bidimensional, com

- marcadores de indivíduos representados por pontos; e
- marcadores de variáveis representados por vectores.

Tem-se, **aproximadamente**:

- o cosseno do ângulo entre marcadores de variáveis é o coeficiente de correlação entre as variáveis;
- o comprimento do marcador de cada variável é proporcional ao seu desvio padrão;
- a distância euclidiana entre marcadores de indivíduos é a distância de Mahalanobis entre esses indivíduos:

$$M_{ij} = (\vec{x}_{[i]} - \vec{x}_{[j]})^t \mathbf{S}^{-1} (\vec{x}_{[i]} - \vec{x}_{[j]}) ,$$

**A qualidade da aproximação mede-se como na ACP.**

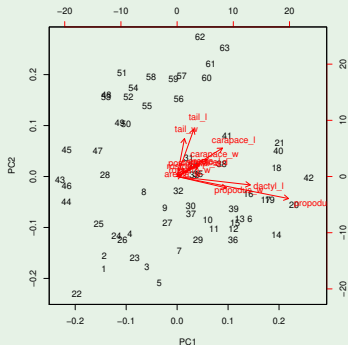
## Biplots (cont.)

Verificam-se ainda as seguintes propriedades **aproximadas** (por ser a  $k=2$  dimensões):

- o cosseno do ângulo entre cada vector e o eixo horizontal é aproximadamente o coeficiente de correlação entre a respectiva variável e a CP 1;
- o cosseno do ângulo entre cada vector e o eixo vertical é aproximadamente o coeficiente de correlação entre a respectiva variável e a CP 2;
- A projecção ortogonal de cada ponto sobre a direcção definida por um vector é aproximadamente o valor do respectivo individuo na variável correspondente.

## Função biplot nos lavagantes

```
> biplot(lav.acp)
```



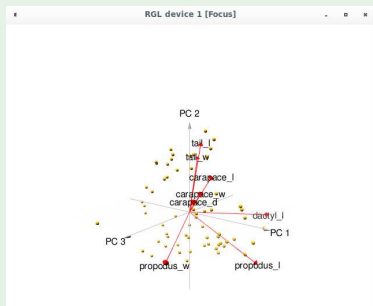
Os individuos do grupo 43-63 (fêmeas) têm tenazes mais pequenas e caudas maiores do que os machos.

## Biplots 3D com `pca3d`

### Biplots a três dimensões com o módulo `pca3d`

Acrescentando ao comando `pca3d` o argumento `biplot=TRUE`:

```
> library(pca3d)
> pca3d(lav.acp, biplot=TRUE)
```

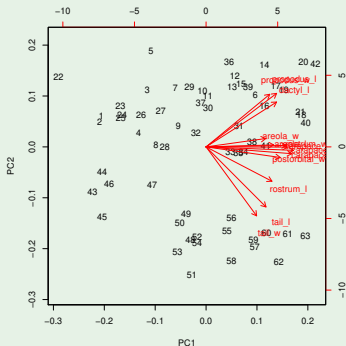


A imagem congela um momento da rotação. Por omissão a função não mostra todos os vectores marcadores de variáveis.

## Os *biplots* no (cont.)

### Função `biplot` nos lavagantes (dados normalizados)

```
> biplot(lav.acpR)
```



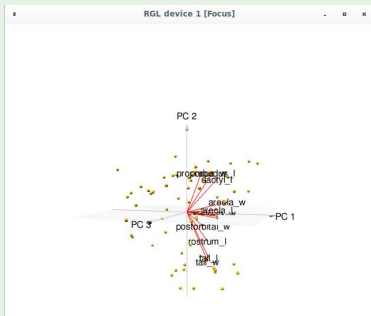
A separação machos/fêmeas continua a ser visível. A primeira CP aponta agora na direcção dum grupo de variáveis altamente correlacionadas entre si (tamanho?).



## Biplots 3D com `pca3d`

### Biplot a três dimensões para dados normalizados de lavagantes

```
> pca3d(lav.acpR, biplot=TRUE, biplot.vars=13)
```



O argumento `biplot.vars` permite controlar os marcadores de variáveis mostrados.

# Análise Discriminante Linear

# Análises Discriminantes

Conjunto de técnicas multivariadas em que:

- Parte-se do conhecimento de que  $n$  indivíduos observados pertencem a  $k$  subgrupos ou classes.
- Procura-se determinar funções das  $p$  variáveis observadas que melhor permitam distinguir ou discriminar esses subgrupos.

Análise Discriminante Linear (ou de Fisher):

Procuram-se combinações lineares das  $p$  variáveis observadas que melhor permitam distinguir ou discriminar esses subgrupos.

NOTA: Só trabalharemos num contexto descritivo, embora muitas vezes as Análises Discriminantes apareçam associadas a conceitos probabilísticos.

## Análise Discriminante Linear (cont.)

**Ponto de partida:** uma matriz de dados,  $\mathbf{X}_{n \times p}$ .

Os  $n$  indivíduos (linhas de  $\mathbf{X}$ ) definem uma partição em  $k$  subgrupos, que é conhecida. Podem ser vistos como  $k$  níveis dum factor.

**Objectivo informal:** determinar a melhor combinação linear  $\mathbf{X}\vec{a}$  das variáveis observadas, para assegurar que:

- indivíduos duma mesma classe tenham valores próximos, e
- indivíduos de classes diferentes tenham valores diferentes.

**Soluções:** combinações lineares  $\mathbf{X}\vec{a}$ , chamadas eixos discriminantes (ou variáveis canónicas).

A solução vai envolver projecções ortogonais sobre o subespaço de  $\mathbb{R}^n$  gerado pelas indicatrizes da constituição de cada subgrupo.

# A matriz da classificação

Matriz da classificação  $\mathbf{G}$ , cuja coluna  $i$  é indicatriz do subgrupo  $i$ :

$$\mathbf{G}_{n \times k} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 0 \\ \hline 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & \dots & 0 \\ \hline \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

Semelhante à matriz do modelo na ANOVA a um factor, mas na 1a. coluna indicatriz do nível 1 do factor. O subespaço de  $\mathbb{R}^n$  gerado pelas duas matrizes é igual.

Há relação estreita entre ADL e ANOVA a um factor, mas só usaremos conceitos descritivos na definição da ADL.

## A matriz da classificação (cont.)

Os vectores do espaço das colunas da matriz  $\mathbf{G}$  têm valor igual nos elementos de cada subgrupo, isto é,  $\vec{\mathbf{z}} \in \mathcal{C}(\mathbf{G})$  são da forma:

$$\vec{\mathbf{z}}^t = \left[ \underbrace{z_1 \ z_1 \ \dots \ z_1}_{n_1 \text{ vezes}} \mid \underbrace{z_2 \ z_2 \ \dots \ z_2}_{n_2 \text{ vezes}} \mid \dots \mid \underbrace{z_k \ z_k \ \dots \ z_k}_{n_k \text{ vezes}} \right]$$

Logo, **vectores em  $\mathcal{C}(\mathbf{G})$  são homogéneos no seio das classes.**

Mas não necessariamente heterogéneos entre classes:  $\mathcal{C}(\mathbf{G})$  inclui também o vector  $\vec{\mathbf{1}}_n$ , que não discrimina subgrupos.

Maximizar a heterogeneidade entre classes significa maximizar a variabilidade dos  $k$  valores  $\{z_j\}_{j=1}^k$ .

É desejável que a combinação linear esteja o mais longe possível de  $\mathcal{C}(\vec{\mathbf{1}}_n) \subset \mathcal{C}(\mathbf{G})$ , ou seja, **que seja ortogonal ao vector  $\vec{\mathbf{1}}_n$ .**

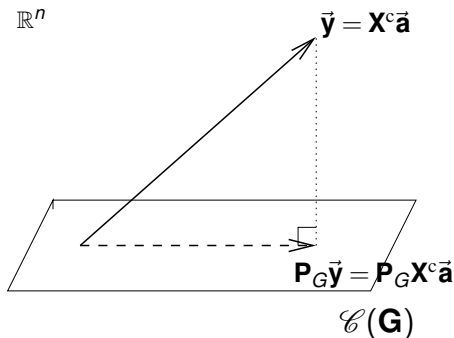
# Formulação do problema

Os vectores ortogonais ao vector  $\vec{1}_n$  são os vectores **centrados**.

Consideremos apenas as **combinações lineares centradas**:  $\mathbf{X}^c \vec{a}$

A projecção ortogonal de qualquer combinação linear centrada no espaço das colunas da matriz de classificação é dada por  $\mathbf{P}_G \mathbf{X}^c \vec{a}$ , onde  $\mathbf{P}_G = \mathbf{G}(\mathbf{G}^t \mathbf{G})^{-1} \mathbf{G}^t$ .

A projecção ortogonal cria um triângulo rectângulo:



## De novo Pitágoras

Pelo Teorema de Pitágoras, e como  $\mathbf{P}_G$  e  $\mathbf{I}_n$  são simétricas e idempotentes, tem-se:

$$\begin{aligned}\|\mathbf{X}^c \bar{\mathbf{a}}\|^2 &= \|\mathbf{P}_G \mathbf{X}^c \bar{\mathbf{a}}\|^2 + \|(\mathbf{I}_n - \mathbf{P}_G) \mathbf{X}^c \bar{\mathbf{a}}\|^2 \\ \Leftrightarrow \bar{\mathbf{a}}^t \mathbf{X}^{ct} \mathbf{X}^c \bar{\mathbf{a}} &= \bar{\mathbf{a}}^t \mathbf{X}^{ct} \mathbf{P}_G \mathbf{X}^c \bar{\mathbf{a}} + \bar{\mathbf{a}}^t \mathbf{X}^{ct} (\mathbf{I}_n - \mathbf{P}_G) \mathbf{X}^c \bar{\mathbf{a}}\end{aligned}$$

O membro esquerdo é  $\bar{\mathbf{a}}^t \mathbf{X}^{ct} \mathbf{X}^c \bar{\mathbf{a}} = (n-1) \cdot \bar{\mathbf{a}}^t \mathbf{S} \bar{\mathbf{a}}$ , ou seja,  $n-1$  vezes a variância da combinação linear  $\mathbf{X}^c \bar{\mathbf{a}}$ .

A combinação linear  $\mathbf{X}^c \bar{\mathbf{a}}$  desejável é a que, nesta decomposição, maximiza (em termos relativos) a primeira parcela do lado direito: como veremos, corresponde a maximizar a variabilidade dos valores  $z_j$  de cada grupo.





# Os vectores projectados $\mathbf{P}_G \vec{\mathbf{y}}$

Consideremos qualquer vector  $\vec{\mathbf{y}} \in \mathbb{R}^n$ , representado com dupla indexação  $i, j$ , sendo  $i$  subgrupo e  $j$  repetição. Considere-se também a sua projecção ortogonal sobre  $\mathcal{L}(G)$ :

$$\vec{\mathbf{y}} = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ \hline y_{21} \\ \vdots \\ y_{2n_2} \\ \hline \vdots \\ \hline y_{k1} \\ \vdots \\ y_{kn_k} \end{bmatrix} \qquad \mathbf{P}_G \vec{\mathbf{y}} = \begin{bmatrix} \bar{y}_{1.} \\ \vdots \\ \bar{y}_{1.} \\ \hline \bar{y}_{2.} \\ \vdots \\ \bar{y}_{2.} \\ \hline \vdots \\ \hline \bar{y}_{k.} \\ \vdots \\ \bar{y}_{k.} \end{bmatrix}$$

# Os vectores centrados projectados $\mathbf{P}_G \vec{y}^c$

Consideremos agora um vector previamente **centrado**:

$$\vec{y}^c = \begin{bmatrix} y_{11} - \bar{y}_{..} \\ \vdots \\ y_{1n_1} - \bar{y}_{..} \\ \hline y_{21} - \bar{y}_{..} \\ \vdots \\ y_{2n_2} - \bar{y}_{..} \\ \hline \vdots \\ \hline y_{k1} - \bar{y}_{..} \\ \vdots \\ y_{kn_k} - \bar{y}_{..} \end{bmatrix} \quad \mathbf{P}_G \vec{y}^c = \begin{bmatrix} \bar{y}_1 - \bar{y}_{..} \\ \vdots \\ \frac{\bar{y}_1 - \bar{y}_{..}}{\bar{y}_2 - \bar{y}_{..}} \\ \vdots \\ \frac{\bar{y}_2 - \bar{y}_{..}}{\bar{y}_2 - \bar{y}_{..}} \\ \vdots \\ \hline \bar{y}_k - \bar{y}_{..} \\ \vdots \\ \bar{y}_k - \bar{y}_{..} \end{bmatrix}$$

$\|\mathbf{P}_G \vec{y}^c\|^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y}_{..})^2$  mede a dispersão das médias de cada classe em torno da média geral  $\bar{y}_{..}$ . Corresponde a **SQF** na ANOVA de  $\vec{y}$  sobre o factor que define os grupos. É a **variabilidade inter-classes**, e convém ser grande: reflecte heterogeneidade entre classes.

# Os vectores $(\mathbf{I}_n - \mathbf{P}_G)\vec{y}$

Para qualquer vector  $\vec{y} \in \mathbb{R}^n$ , incluindo vectores **centrados**  $\vec{y}^c$ :

$$\vec{y} - \mathbf{P}_G \vec{y} = (\mathbf{I}_n - \mathbf{P}_G) \vec{y} = \begin{bmatrix} y_{11} - \bar{y}_1. \\ \vdots \\ y_{1n_1} - \bar{y}_1. \\ \hline y_{21} - \bar{y}_2. \\ \vdots \\ y_{2n_2} - \bar{y}_2. \\ \hline \vdots \\ \hline y_{k1} - \bar{y}_k. \\ \vdots \\ y_{kn_k} - \bar{y}_k. \end{bmatrix}$$

$$(\mathbf{I}_n - \mathbf{P}_G) \vec{y}^c = \begin{bmatrix} y_{11} - \bar{y}_1. \\ \vdots \\ y_{1n_1} - \bar{y}_1. \\ \hline y_{21} - \bar{y}_2. \\ \vdots \\ y_{2n_2} - \bar{y}_2. \\ \hline \vdots \\ \hline y_{k1} - \bar{y}_k. \\ \vdots \\ y_{kn_k} - \bar{y}_k. \end{bmatrix}$$

$\|(\mathbf{I}_n - \mathbf{P}_G)\vec{y}^c\|^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i.)^2$  mede a dispersão das observações individuais em torno da respectiva média de nível. Corresponde a **SQRE** na ANOVA de  $\vec{y}$  sobre o factor que define os grupos. É a **variabilidade intra-classes**, e convém ser pequena: mede homogeneidade das classes.

# De novo a equação resultante de Pitágoras

A equação do acetato 88 simplifica definindo as matrizes:

$$\begin{aligned} \mathbf{S} &= \frac{1}{n-1} \mathbf{X}^{ct} \mathbf{X}^c && \text{Matriz de variâncias-covariâncias de } \mathbf{X} \\ \mathbf{B} &= \frac{1}{n-1} \mathbf{X}^{ct} \mathbf{P}_G \mathbf{X}^c && \text{Matriz da variabilidade inter-classes} \\ \mathbf{W} &= \frac{1}{n-1} \mathbf{X}^{ct} (\mathbf{I}_n - \mathbf{P}_G) \mathbf{X}^c && \text{Matriz da variabilidade intra-classes} \end{aligned}$$

Tem-se:

$$\begin{aligned} \bar{\mathbf{a}}^t \underbrace{\mathbf{X}^{ct} \mathbf{X}^c}_{=(n-1) \cdot \mathbf{S}} \bar{\mathbf{a}} &= \bar{\mathbf{a}}^t \underbrace{\mathbf{X}^{ct} \mathbf{P}_G \mathbf{X}^c}_{=(n-1) \cdot \mathbf{B}} \bar{\mathbf{a}} + \bar{\mathbf{a}}^t \underbrace{\mathbf{X}^{ct} (\mathbf{I}_n - \mathbf{P}_G) \mathbf{X}^c}_{=(n-1) \cdot \mathbf{W}} \bar{\mathbf{a}} \end{aligned}$$

$$\iff \bar{\mathbf{a}}^t \mathbf{S} \bar{\mathbf{a}} = \bar{\mathbf{a}}^t \mathbf{B} \bar{\mathbf{a}} + \bar{\mathbf{a}}^t \mathbf{W} \bar{\mathbf{a}}$$

Uma formulação do problema (de Fisher): de entre as combinações lineares  $\mathbf{X}^c \bar{\mathbf{a}}$ , escolher a que maximiza:

$$\frac{\bar{\mathbf{a}}^t \mathbf{B} \bar{\mathbf{a}}}{\bar{\mathbf{a}}^t \mathbf{W} \bar{\mathbf{a}}}$$

Essa será a primeira função discriminante.

# Problema generalizado de valores próprios

## Teorema (Problema generalizado de valores próprios)

Sejam  $\mathbf{A}_{p \times p}$  e  $\mathbf{B}_{p \times p}$  matrizes simétricas ( $\mathbf{B}_{p \times p}$  só com valores próprios positivos).

- A maximização do quociente

$$\frac{\vec{\mathbf{x}}^t \mathbf{A} \vec{\mathbf{x}}}{\vec{\mathbf{x}}^t \mathbf{B} \vec{\mathbf{x}}}$$

está associada ao primeiro par próprio da matriz  $\mathbf{B}^{-1} \mathbf{A}$ ,  $(\lambda_1, \vec{\mathbf{x}}_1)$ .

- Sucessivos pares de valores/vectores próprios de  $\mathbf{B}^{-1} \mathbf{A}$  estão associados a sucessivos máximos do quociente  $\frac{\vec{\mathbf{x}}^t \mathbf{A} \vec{\mathbf{x}}}{\vec{\mathbf{x}}^t \mathbf{B} \vec{\mathbf{x}}}$ , sujeitos à exigência de  $\mathbf{B}$ -ortogonalidade de sucessivos vectores, isto é,  $\vec{\mathbf{x}}_i^t \mathbf{B} \vec{\mathbf{x}}_j = 0$ .

**Nota:** O produto de matrizes simétricas não é, em geral, simétrica, pelo que os valores/vectores próprios podem ser complexos. Mas os valores/vectores próprios de  $\mathbf{B}^{-1} \mathbf{A}$  são reais.

# Formulação de Fisher

**Solução:** Se  $\mathbf{W}$  for invertível, usar o problema de valores/vectores próprios generalizados (acetato 94): tomar  $\vec{\mathbf{a}} = \vec{\mathbf{a}}_1$ , o vector próprio do maior valor próprio de  $\mathbf{W}^{-1}\mathbf{B}$ .

O valor próprio  $\lambda_1 = \frac{\vec{\mathbf{a}}_1^t \mathbf{B} \vec{\mathbf{a}}_1}{\vec{\mathbf{a}}_1^t \mathbf{W} \vec{\mathbf{a}}_1}$  é a **capacidade discriminante** do eixo: razão entre as variabilidades entre- e intra-grupos nesse eixo (**ratio of between-group and within-group variability**).

Se o número de valores próprios não nulos de  $\mathbf{W}^{-1}\mathbf{B}$  for maior do que 1, podem procurar-se novas combinações lineares discriminantes.

Sucessivas soluções são as combinações lineares  $\mathbf{X}\vec{\mathbf{a}}_j$  com  $\vec{\mathbf{a}}_j$  dados por outros vectores próprios da matriz  $\mathbf{W}^{-1}\mathbf{B}$  associados a valores próprios não-nulos.

A **capacidade discriminante** destes novos eixos é dada pelo valor próprio  $\lambda_j = \frac{\vec{\mathbf{a}}_j^t \mathbf{B} \vec{\mathbf{a}}_j}{\vec{\mathbf{a}}_j^t \mathbf{W} \vec{\mathbf{a}}_j}$ .

# Observações

- Se  $k > n - p$ ,  $\mathbf{W}$  não é invertível. Em geral, se  $k \leq n - p$   $\mathbf{W}$  é invertível.
- As matrizes usadas na ADL verificam a relação  $\mathbf{S} = \mathbf{B} + \mathbf{W}$ .

De facto,

$$\mathbf{I}_n = \mathbf{P}_G + (\mathbf{I}_n - \mathbf{P}_G) \Rightarrow \mathbf{X}^{ct} \mathbf{I}_n \mathbf{X}^c = \mathbf{X}^{ct} \mathbf{P}_G \mathbf{X}^c + \mathbf{X}^{ct} (\mathbf{I}_n - \mathbf{P}_G) \mathbf{X}^c \Leftrightarrow \mathbf{S} = \mathbf{B} + \mathbf{W}$$

- Sucessivos eixos discriminantes são não correlacionados entre si.

Diferentes eixos discriminantes são da forma  $\mathbf{X}^c \bar{\mathbf{a}}_i$  e  $\mathbf{X}^c \bar{\mathbf{a}}_j$ , sendo  $\bar{\mathbf{a}}_i$  e  $\bar{\mathbf{a}}_j$  dois diferentes vectores próprios da matriz  $\mathbf{W}^{-1} \mathbf{B}$ . Sabemos ainda que  $\bar{\mathbf{a}}_i$  e  $\bar{\mathbf{a}}_j$  são  $\mathbf{W}$ -ortogonais. Logo, se  $i \neq j$ :

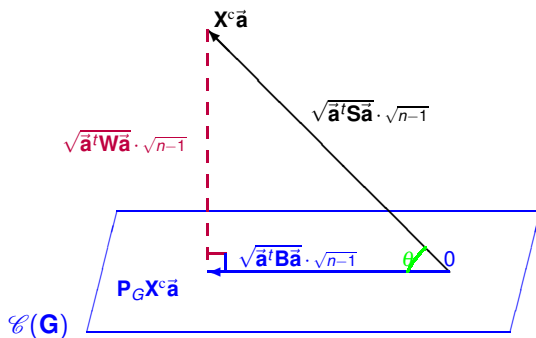
$$\begin{aligned} \mathbf{W}^{-1} \mathbf{B} \bar{\mathbf{a}}_j &= \lambda_j \bar{\mathbf{a}}_j \Rightarrow \mathbf{B} \bar{\mathbf{a}}_j = \lambda_j \mathbf{W} \bar{\mathbf{a}}_j \\ &\Rightarrow \mathbf{W} \bar{\mathbf{a}}_j + \mathbf{B} \bar{\mathbf{a}}_j = \mathbf{W} \bar{\mathbf{a}}_j + \lambda_j \mathbf{W} \bar{\mathbf{a}}_j \\ &\Rightarrow \mathbf{S} \bar{\mathbf{a}}_j = (1 + \lambda_j) \mathbf{W} \bar{\mathbf{a}}_j \\ &\Rightarrow \text{Cov}(\mathbf{X}^c \bar{\mathbf{a}}_i, \mathbf{X}^c \bar{\mathbf{a}}_j) = \bar{\mathbf{a}}_i^t \mathbf{S} \bar{\mathbf{a}}_j = (1 + \lambda_j) \bar{\mathbf{a}}_i^t \mathbf{W} \bar{\mathbf{a}}_j = 0 \end{aligned}$$



## Observações (cont.)

- Contrariamente à ACP, eixos discriminantes  $\mathbf{X}\vec{\mathbf{a}}_j$  não-correlacionados **não** significa vectores de coeficientes  $\vec{\mathbf{a}}_j$  ortogonais em  $\mathbb{R}^p$  (são **W**-ortogonais).
- **$\mathbf{W}^{-1}\mathbf{B}$  não pode ter mais de  $k-1$  valores próprios não-nulos.** Assim, não pode haver mais eixos discriminantes que o número de níveis do factor menos um.
- as soluções duma Análise Discriminante **são** invariantes a transformações lineares das escalas das variáveis.

# ADL - Resumo



Maximizar  $\frac{\vec{a}^t \mathbf{B} \vec{a}}{\vec{a}^t \mathbf{W} \vec{a}}$  é maximizar  $\text{ctg}^2(\theta)$ . Em cada eixo,  $\lambda_j = \text{ctg}^2(\theta_j)$ .

# Ainda a geometria numa ADL

Maximizar a cotangente do ângulo  $\theta$  quer dizer minimizar  $\theta$ .

Na ADL procuramos a combinação linear  $\mathbf{X}^c \vec{\mathbf{a}}$  das variáveis centradas (colunas de  $\mathbf{X}^c$ ) que forma o menor ângulo ( $\theta$ ) com o espaço gerado pelas indicatrizes de subgrupos (colunas de  $\mathbf{G}$ ).

Esse ângulo  $\theta$  é o menor ângulo entre dois subespaços de  $\mathbb{R}^n$ :

- o subespaço gerado pelas indicatrizes,  $\mathcal{L}(\mathbf{G})$ ; e
- o subespaço gerado pelas variáveis centradas,  $\mathcal{L}(\mathbf{X}^c)$ .

A capacidade discriminante das variáveis depende do menor ângulo entre  $\mathcal{L}(\mathbf{X}^c)$  e  $\mathcal{L}(\mathbf{G})$ , ou seja, da posição angular desses dois subespaços de  $\mathbb{R}^n$ .

# Formulações alternativas

Formulações alternativas que minimizam o ângulo  $\theta$ :

- 1 Minimizar o quadrado do seno de  $\theta$ .

i.e., minimizar a proporção da variabilidade total da combinação linear  $\mathbf{X}^c \vec{\mathbf{a}}$  que corresponde à variabilidade intra-classes

$$\frac{\vec{\mathbf{a}}^t \mathbf{W} \vec{\mathbf{a}}}{\vec{\mathbf{a}}^t \mathbf{S} \vec{\mathbf{a}}}$$

- 2 Maximizar o quadrado do cosseno do ângulo  $\theta$

ou seja, maximizar a proporção da variabilidade total da combinação linear  $\mathbf{X}^c \vec{\mathbf{a}}$  que corresponde à variabilidade inter-classes

$$\frac{\vec{\mathbf{a}}^t \mathbf{B} \vec{\mathbf{a}}}{\vec{\mathbf{a}}^t \mathbf{S} \vec{\mathbf{a}}}$$

## Relações entre formulações alternativas

Mas problema igual (minimizar  $\theta$ )  $\Rightarrow$  solução igual.

É fácil de verificar que são iguais:

- Vectores próprios de  $\mathbf{W}^{-1}\mathbf{B}$ ;
- Vectores próprios de  $\mathbf{S}^{-1}\mathbf{W}$ ;
- Vectores próprios de  $\mathbf{S}^{-1}\mathbf{B}$ ;

Logo, as combinações lineares  $\mathbf{X}^c\vec{\mathbf{a}}$  obtidas com as formulações alternativas são iguais.

Os correspondentes valores próprios **não** são iguais porque correspondem às diferentes funções trigonométricas. Mas, seja  $\vec{\mathbf{a}}$  vector próprio comum às três matrizes.

- Se  $\lambda$  é o valor próprio associado em  $\mathbf{W}^{-1}\mathbf{B}$ ;
- $\frac{1}{\lambda+1}$  é o valor próprio em  $\mathbf{S}^{-1}\mathbf{W}$  (queremos minimizar);
- $\frac{\lambda}{\lambda+1}$  é o valor próprio em  $\mathbf{S}^{-1}\mathbf{B}$  (queremos maximizar).

# ADL e ANOVA

Considere-se:

- uma ANOVA a um factor com as  $k$  classes;
- a **variável resposta**  $\vec{y} = \mathbf{X}^c \vec{a}$ .

O critério que define ADL equivale a pedir  $\vec{a}$  tal que a estatística do teste  $F$  da ANOVA aos efeitos do factor seja máxima:

$$F = \frac{QMF}{QMRE} = \frac{SQF}{SQRE} \cdot \frac{n-k}{k-1} = \frac{\|\mathbf{P}_G \vec{y}\|^2}{\|(\mathbf{I}_n - \mathbf{P}_G) \vec{y}\|^2} \cdot \frac{n-k}{k-1} = \frac{\vec{a}^t \mathbf{B} \vec{a}}{\vec{a}^t \mathbf{W} \vec{a}} \cdot \frac{n-k}{k-1}.$$

Os eixos discriminantes são as sucessivas combinações lineares, não correlacionadas, das  $p$  variáveis observadas que maximizam a separação entre as observações dos diversos níveis do factor.

## Classificação de novos indivíduos num eixo

Podemos classificar novos indivíduos, de “filiação” desconhecida.

Seja  $\vec{x}$  vector de observações de novo indivíduo nas  $p$  variáveis.

O respectivo valor (*score*) no eixo discriminante 1 é  $y^* = \vec{x}^t \vec{a}_1$ .

Comparando este valor com as  $k$  médias de classe nesse eixo,  $\bar{y}^{(1)}, \bar{y}^{(2)}, \dots, \bar{y}^{(k)}$ , podemos classificar na classe cujo centro de gravidade:

- Ihe está mais próxima, na habitual distância euclidiana:

$$\text{fica na classe } i \text{ se } |y^* - \bar{y}^{(i)}| < |y^* - \bar{y}^{(j)}|, \quad \forall j \neq i.$$

- Ihe está mais próxima, numa distância euclidiana inversamente ponderada pelo desvio padrão da classe:

$$\text{fica na classe } i \text{ se } \frac{|y^* - \bar{y}^{(i)}|}{s_y^{(i)}} < \frac{|y^* - \bar{y}^{(j)}|}{s_y^{(j)}}, \quad \forall j \neq i,$$

onde  $s_y^{(i)}$  indica o desvio padrão dos scores do grupo  $i$ .

## Classificação em $q$ eixos

Com  $q$  eixos discriminantes, um indivíduo tem vector de scores:  $\vec{y}^* = \vec{x}^t \mathbf{A}_q$ , com  $\mathbf{A}_q$  matriz  $p \times q$  cujas colunas são os vectores  $\vec{a}_1, \dots, \vec{a}_q$ .

Pode-se classificar o indivíduo na classe cujo centro de gravidade  $\vec{m}_{(i)}$ :

- esteja mais próximo de  $\vec{y}^*$ , na habitual distância euclidiana:  
*associar à classe  $i$  se  $\|\vec{y}^* - \vec{m}_{(i)}\| < \|\vec{y}^* - \vec{m}_{(j)}\|$ ,  $\forall j \neq i$*

- esteja mais próximo de  $\vec{y}^*$  na distância de Mahalanobis usual:  
*classe  $i$  se  $\|\vec{y}^* - \vec{m}_{(i)}\|_{\mathbf{S}^{-1}} < \|\vec{y}^* - \vec{m}_{(j)}\|_{\mathbf{S}^{-1}}$ ,  $\forall j \neq i$ ,*

onde  $\mathbf{S}$  é matriz de (co)variâncias dos scores das  $n$  observações.

- esteja mais próximo de  $\vec{y}^*$  na distância de Mahalanobis definida pela matriz de variâncias-covariâncias das observações dessa classe:

$$\textit{classe } i \textit{ se } \|\vec{y}^* - \vec{m}_{(i)}\|_{\mathbf{S}_i^{-1}} < \|\vec{y}^* - \vec{m}_{(j)}\|_{\mathbf{S}_j^{-1}}, \forall j \neq i,$$

onde  $\mathbf{S}_i$  é matriz de (co)variâncias dos scores do grupo  $i$ .



## ADL no R - a função `lda`

A função `lda`, do módulo `MASS`, fornece a informação básica para uma Análise Discriminante Linear (de Fisher).

A função `lda` foi concebida pensando num contexto inferencial (que não é o nosso e não é necessário para uma ADL). No entanto fornece a informação necessária para um contexto descritivo.

Exemplifiquemos com os dados dos lavagantes, em que as 21 primeiras observações são de machos reprodutores (grupo MR); as 21 seguintes de machos não-reprodutores (grupo MN); e as 21 observações finais são de fêmeas (grupo F).

É necessário construir o factor dos grupos e carregar o módulo `MASS`:

```
> lav.grupos <- factor(rep(c("MR", "MN", "F"), c(21, 21, 21)))
> library(MASS)
```

# O exemplo dos lavagantes

## ADL dos lavagantes com função `lda`

Na fórmula do comando `lda` o factor dos grupos é a variável resposta.

```
> lav.lda <- lda(lav.grupos ~ . , data=as.data.frame(lavagantes))
> lav.lda
```

Coefficients of linear discriminants:

|               | LD1           | LD2         |
|---------------|---------------|-------------|
| carapace_l    | -0.0005163473 | -1.19955746 |
| tail_l        | -0.1736612417 | 0.33191555  |
| carapace_w    | 0.1866238904  | -0.90101141 |
| carapace_d    | -0.3521185558 | -0.23124418 |
| tail_w        | -2.6055856004 | 1.28663805  |
| areola_l      | 0.3588957427  | -0.06043209 |
| areola_w      | -2.1123185437 | -0.03550332 |
| rostrum_l     | 1.2415578489  | 1.22874815  |
| rostrum_w     | -0.3314912527 | 1.39715849  |
| postorbital_w | 0.1940959791  | -1.59005854 |
| propodus_l    | 0.6321803333  | 0.17783018  |
| propodus_w    | 0.4297842346  | 0.71193763  |
| dactyl_l      | -0.0850563760 | 0.36615202  |

<- vectores de loadings

Proportion of trace:

| LD1    | LD2    |
|--------|--------|
| 0.9501 | 0.0499 |

<- proporção dos valores próprios não nulos de  $\text{inv}(\mathbf{W})\mathbf{B}$

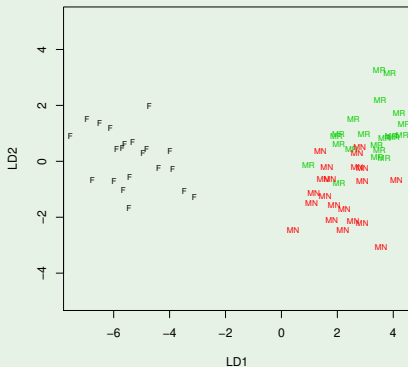
<- não são os valores do critério (razão de variâncias), definidos acima

## Exemplos dos lavagantes (cont.)

Existe um método `plot` para objectos da classe `lda`, produzidos pelo comando `lda`:

### ADL dos lavagantes

```
> plot(lav.lda, col=as.numeric(lav.grupos))
```



## Exemplos dos lavagantes (cont.)

Os **vectores de scores** usados para construir o gráfico podem ser obtidos com o comando `predict`, estando no argumento de saída `x`.

### ADL dos lavagantes

```
> predict(lav.lda)$x
```

|       | LD1        | LD2        |
|-------|------------|------------|
| 1     | 2.9590031  | 0.9654792  |
| 2     | 3.6848954  | 0.8131683  |
| 3     | 3.5259200  | 2.1811447  |
| 4     | 2.0462745  | 0.6083346  |
| [...] |            |            |
| 60    | -4.9547011 | 0.2934347  |
| 61    | -6.7592582 | -0.6571673 |
| 62    | -5.6927267 | 0.4566755  |
| 63    | -5.4276951 | -0.5692571 |

O comando `predict` pode ser usado para **determinar as coordenadas nos eixos discriminantes de um novo indivíduo**, de forma semelhante ao que foi visto para os modelos lineares.

# Exemplo dos lavagantes (cont.)

## ADL dos lavagantes

Definamos 3 novos indivíduos com valores iguais aos máximos de cada variável em cada subgrupo, e coloquemo-los nos novos eixos discriminantes:

```
> lxm1 <- apply(lavagantes[1:21,], 2, max)
> lxm2 <- apply(lavagantes[22:42,], 2, max)
> lxm3 <- apply(lavagantes[43:63,], 2, max)
> novos <- as.data.frame(rbind(lxm1, lxm2, lxm3))
> novos
```

```
 carapace_l tail_l carapace_w carapace_d tail_w areola_l areola_w rostrum_l
lxm1 35.33 25.15 18.36 14.57 15.40 13.26 2.60 7.06
lxm2 35.50 25.05 18.74 15.11 15.11 16.85 2.64 7.05
lxm3 35.73 26.77 18.50 15.06 17.37 13.14 2.32 7.27
 rostrum_w postorbital_w propodus_l propodus_w dactyl_l
lxm1 8.12 10.76 33.24 15.53 20.71
lxm2 7.74 11.85 33.67 15.15 20.83
lxm3 7.83 11.14 28.29 12.30 17.58
```

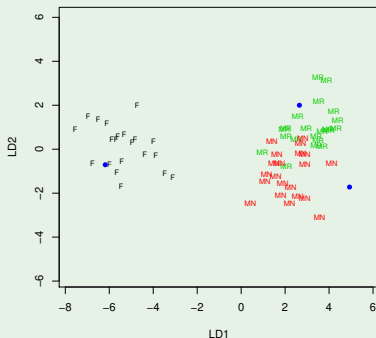
```
> predict(lav.lda, new=novos)$x
```

```
 LD1 LD2
lxm1 2.650187 1.9990716 <- coordenadas do primeiro individuo nos EDs
lxm2 4.931230 -1.7232999 <- coordenadas do segundo individuo nos EDs
lxm3 -6.183037 -0.7078646 <- coordenadas do terceiro individuo nos EDs
```

# Exemplos dos lavagantes (cont.)

## ADL dos lavagantes

```
> lxmp <- predict(lav.lda, new=novos)$x
> plot(lav.lda, col=as.numeric(lav.grupos), xlim=c(-8,6))
> points(lxmp, col="blue", pch=16)
```



# Observações sobre a função lda do MASS

**Atenção:** Na função lda,

- a matriz  $\mathbf{W}$  é definida como  $\mathbf{W} = \frac{1}{n-k} \mathbf{X}^{ct} (\mathbf{I}_n - \mathbf{P}_G) \mathbf{X}^c$ ;
- a matriz  $\mathbf{B}$  é definida como  $\mathbf{B} = \frac{1}{k-1} \mathbf{X}^{ct} \mathbf{P}_G \mathbf{X}^c$ ;
- deixa de ser válida a decomposição  $\mathbf{S} = \mathbf{W} + \mathbf{B}$ .
- os valores próprios de  $\mathbf{W}^{-1} \mathbf{B}$  com a função lda são  $\frac{n-k}{k-1}$  vezes os da nossa definição. É o valor da estatística do teste  $F$  na ANOVA de cada eixo discriminante sobre o factor;
- a componente **svd** dum objecto de classe lda dá as **raízes quadradas** dos valores próprios de  $\mathbf{W}^{-1} \mathbf{B}$  (definidos à moda de lda).

# Qualidade dos eixos discriminantes

## ADL dos lavagantes

```
> lav.lda$svd
[1] 21.345129 4.890076
> lav.lda$svd^2 <- Valores propios (e valores da estatística F)
[1] 455.61455 23.91285

> summary(aov(predict(lav.lda)$x[,1] ~ lav.grupos))
 Df Sum Sq Mean Sq F value Pr(>F)
lav.grupos 2 911.2 455.6 455.6 <2e-16
Residuals 60 60.0 1.0

> summary(aov(predict(lav.lda)$x[,2] ~ lav.grupos))
 Df Sum Sq Mean Sq F value Pr(>F)
lav.grupos 2 47.83 23.91 23.91 2.31e-08
Residuals 60 60.00 1.00
```



# Qualidade de eixos discriminantes

É possível obter os valores próprios da definição original de Fisher, multiplicando os valores próprios da função lda por  $\frac{k-1}{n-k}$ .

## ADL dos lavagantes

```
> lav.lda$svd^2*2/60 <- Valores próprios com W e B “à nossa moda”
[1] 15.1871516 0.7970949
```

A capacidade discriminante do primeiro eixo é 15.187, ou seja, a variabilidade entre os três grupos nesse eixo é 15.187 vezes maior que a variabilidade intra-grupo.

A capacidade discriminante do segundo eixo é fraca: 0.797, ou seja, a variabilidade entre os três grupos é, nesse eixo, **menor** que a variabilidade intra-grupo.

## Observações sobre a função `lda` (cont.)

- as proporções (dadas na listagem) de cada valor próprio não são afectadas pelas diferentes definições.

```
> val <- lav.lda$svd^2
> val/sum(val)
[1] 0.95013247 0.04986753
> val2 <- lav.lda$svd^2*2/60
> val2/sum(val2)
[1] 0.95013247 0.04986753
```

- a **W**-ortogonalidade dos *loadings* dados na listagem também é preservada (embora a norma ao quadrado dos vectores de *loadings* seja afectada, sendo  $\frac{n-k}{n-1}$  quando medida com a norma da matriz **W** definida no acetato 93).

# A classificação de novos indivíduos

O método `predict` da função `lda` classifica indivíduos nos grupos, com critérios baseados em conceitos inferenciais, mas próximos da classificação pelas distâncias de Mahalanobis. As classificações são guardadas no objecto `class`.

## Classificação dos lavagantes pela LDA

```
> predict(lav.lda)$class
```

```
[1] MR MR MR MR MR MR MR MN MN MR MR MR MR MR MR MR MR MR MR MR MN MR MN MN
[26] MN MR MN MN MN MN MN MN MN MN MN MN MR MN MN MN F F F F F F F F
[51] F F F F F F F F F F F F F
```

```
> predict(lav.lda, new=novos)$class
```

```
[1] MR MN F
```

# Tabelas de classificações

Tabelas de classificação podem ser criadas com a função `table`.

## Tabelas de classificação da LDA lavagantes

```
> lav.pred <- predict(lav.lda)$class
> table(lav.pred, lav.grupos)
```

```
 lav.grupos
lav.pred F MN MR
 F 21 0 0
 MN 0 18 2
 MR 0 3 19
```

- Todas as fêmeas foram bem classificadas.
- Três machos não reprodutores foram incorrectamente classificados como reprodutores.
- Dois machos reprodutores foram incorrectamente classificados como não reprodutores.

# Classificações erradas

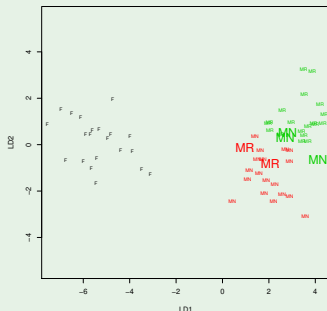
## Classificações erradas na LDA dos lavagantes

```
> (lav.grupos != predict(lav.lda)$class)
```

```
[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE
[17] FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE
[33] FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
> lav.mal <- (lav.grupos != predict(lav.lda)$class)
```

```
> plot(lav.lda, col=as.numeric(predict(lav.lda)$class), cex=0.7+lav.mal)
```



# O potencial do : uma outra função para ADL

## Exercício 17: a nossa função `adl`

```
> adl <- function(X,grupos){
 grupos <- as.factor(grupos)
 X <- as.matrix(X)
 k <- length(levels(grupos))
 n <- dim(X)[1]
 p <- dim(X)[2]
 Ind <- model.matrix(aov(X[,1] ~ -1 + grupos))
 PG <- Ind %>% solve(t(Ind)%>%Ind) %>% t(Ind)
 Xc <- scale(X, scale=F)
 B <- (t(Xc) %>% PG %>% Xc)/(n-1)
 W <- (t(Xc) %>% (diag(n)-PG) %>% Xc)/(n-1)
 valvec <- eigen(solve(W)%%B)
 val <- Re(valvec$val)[1:(k-1)]
 loadings <- Re(valvec$vec)[1:(k-1)]
 if (k>2) rownames(loadings) <- colnames(X)
 else if (k==2) names(loadings) <- colnames(X)
 rownames(B) <- colnames(X)
 colnames(B) <- colnames(X)
 rownames(W) <- colnames(X)
 colnames(W) <- colnames(X)
 if (k>2) colnames(loadings) <- paste("ED",1:(k-1),sep="")
 scores <- Xc %>% loadings
 rownames(scores) <- rownames(X)
 list(B=B,W=W,val=val,loadings=loadings,scores=scores)
}
```

`<- argumentos de entrada dados (X) e classes (grupos)`  
`<- garante que 'grupos' seja factor`  
`<- garante que 'X' seja matriz`  
`<- k: número de grupos`  
`<- n: número de indivíduos`  
`<- p: número de variáveis`  
`<- cria a matriz G indicada nos acetatos`  
`<- matriz de projecção P_G`  
`<- matriz centrada dos dados`  
`<- matriz B de variabilidade entre grupos`  
`<- matriz W de variabilidade intra-grupos`  
`<- valores e vectores próprios de inv(W)B`  
`<- valores próprios de inv(W)B`  
`<- vectores próprios de inv(W)B`  
`<- nomes de objectos de saída`  
`<- objecto (lista) de saída`

# A função adl em acção

```
> adl(lavagantes, lav.grupos)$val
[1] 15.1871516 0.7970949 <- comparar com valores anteriores

> adl(lavagantes, lav.grupos)$loadings <- de norma 1, W-ortogonais
```

|               | ED1          | ED2         |
|---------------|--------------|-------------|
| carapace_l    | 0.000138748  | -0.36607521 |
| tail_l        | 0.046664632  | 0.10129240  |
| carapace_w    | -0.050147834 | -0.27496636 |
| carapace_d    | 0.094618019  | -0.07056999 |
| tail_w        | 0.700148699  | 0.39265005  |
| areola_l      | -0.096439122 | -0.01844238 |
| areola_w      | 0.567602569  | -0.01083473 |
| rostrum_l     | -0.333619864 | 0.37498349  |
| rostrum_w     | 0.089075243  | 0.42637815  |
| postorbital_w | -0.052155664 | -0.48524647 |
| propodus_l    | -0.169873612 | 0.05426936  |
| propodus_w    | -0.115487617 | 0.21726572  |
| dactyl_l      | 0.022855557  | 0.11174052  |