Modelos Matemáticos e Aplicações
Estatística Multivariada

Jorge Cadima

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

2017-18

Programa

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

Bibliografia

- 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).

Conceitos introdutórios em ACP e ADL

Matéria prima: Matriz de dados $\mathbf{Y}_{n\times p}$: conjuntos de observações de p variáveis (quantitativas) em n individuos, ou unidades experimentais.

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

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

Quer numa Análise em Componentes Principais (ACP), quer numa Análise Discriminante Linear (ADL) procuram-se novas variáveis, construídas a partir das p variáveis observadas que salientem:

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

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

A representação usual dos dados

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

$$\begin{array}{ccc} p \text{ eixos} & \longleftrightarrow & p \text{ variáveis} \\ n \text{ pontos} & \longleftrightarrow & n \text{ individuos} \end{array}$$

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

A ACP pode ser vista como uma técnica de redução "óptima" da dimensionalidade. 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

A representação gráfica de dados multivariados

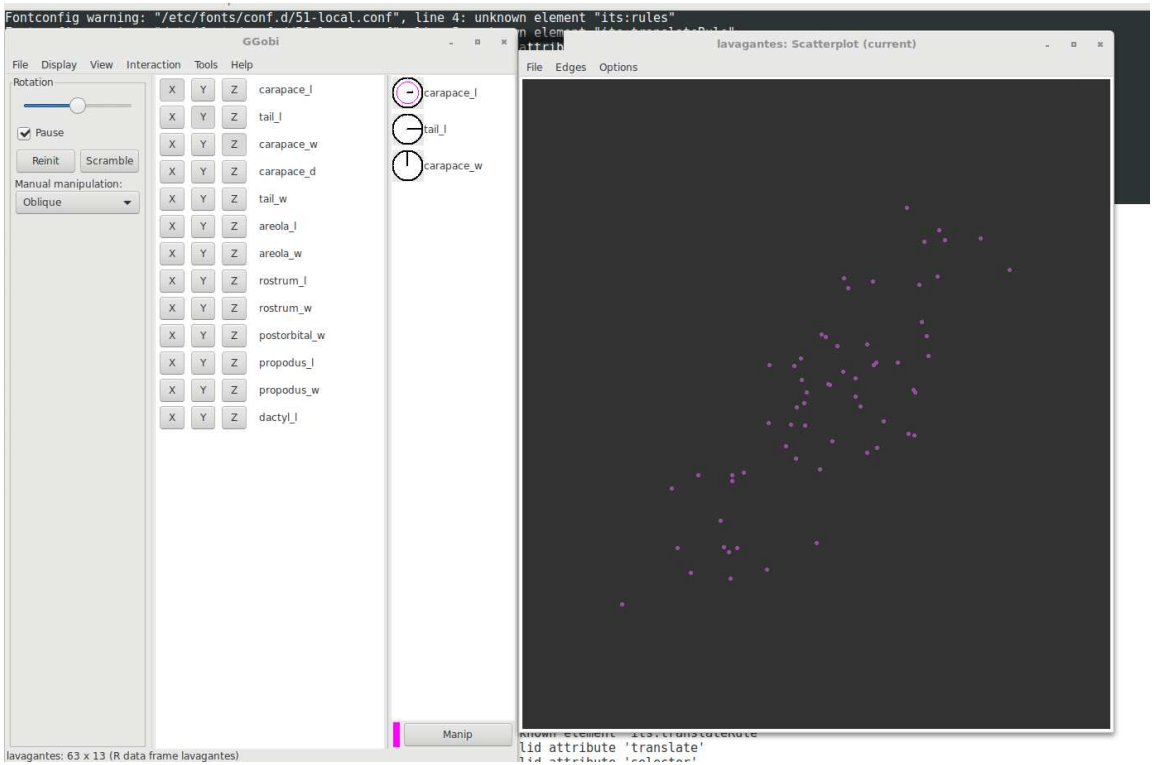
Até $p = 3$ é possível a representação usual dos dados: cada variável associada a um eixo, e cada indivíduo observado é dado por um ponto nesse sistema de p eixos.

Uma ferramenta de visualização tri-dimensional encontra-se no módulo `rggobi`. Este módulo necessita que esteja previamente instalado no sistema (exterior ao `R`) o *software* `GGobi` (gratuito e de código aberto, `www.ggobi.org`).

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

A representação gráfica (cont.)

O módulo `rggobi` abre uma janela de controlo e uma janela gráfica. Para visualizar a 3 dimensões, seleccione no menu `View` a opção `Rotation` (pare a rotação seleccionando a opção `Pause`). Pode seleccionar as três variáveis que quer visualizar.



No menu `Interaction` seleccione a opção `Identify` e, com o rato, passeie pela nuvem de pontos, identificando observações individuais.

Projecções ortogonais

Mas continuamos a ter visões parciais, resultantes de projectar ortogonalmente a nuvem de $n = 63$ pontos em $\mathbb{R}^{13}$ sobre espaços (neste caso tri-dimensionais) coordenados.

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

Mas,

- Porquê só (hiper)planos coordenados? Porque não outros (hiper)planos?
- Qual o plano onde a projecção é mais fidedigna?
- Qual o critério de "projecção fidedigna"?

Ideia intuitiva: estar globalmente mais perto da generalidade dos pontos. Esta é a abordagem que conduz à Análise em Componentes Principais.

Conceitos de Teoria de Matrizes e Álgebra Linear

A Estatística Multivariada, sobretudo descritiva, precisa de:

- Conceitos de Álgebra Linear, como:
  - Espaços e subespaços lineares (vectoriais);
  - Combinações lineares e independência linear;
  - Bases e dimensões de espaços lineares;
  - Produtos internos e conceitos geométricos associados;
  - Projecções (sobretudo ortogonais) sobre subespaços.
- Conceitos de Teoria de Matrizes, como:
  - Operações sobre matrizes e tipos de matrizes;
  - Transformações lineares e matrizes;
  - Valores e vectores próprios;
  - Decomposição espectral de matrizes simétricas;
  - Decomposição em Valores Singulares duma matriz genérica;
  - Problema generalizado de valores próprios.

Matrizes quadradas

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

$\mathbf{A}$ Diagonal	$a_{ij} = 0$ se $i \neq j$ (e existe i tal que $a_{ii} \neq 0$)
$\mathbf{A}$ Simétrica	$\mathbf{A}^t = \mathbf{A} \iff a_{ij} = a_{ji}, \quad \forall i, j$
$\mathbf{I}_p$ 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)
$\mathbf{A}$ Ortogonal	$\mathbf{A}^{-1} = \mathbf{A}^t \iff \mathbf{A}^t\mathbf{A} = \mathbf{A}\mathbf{A}^t = \mathbf{I}_p$
$\mathbf{A}$ Idempotente	$\mathbf{A}^2 = \mathbf{A}\mathbf{A} = \mathbf{A}$

Matrizes simétricas

Seja $\mathbf{A}_{p\times p}$ uma matriz simétrica (necessariamente quadrada).

Para qualquer vector $\vec{\mathbf{x}} \in \mathbb{R}^p$, diz-se que $\vec{\mathbf{x}}^t\mathbf{A}\vec{\mathbf{x}}$ é uma forma quadrática, e tem-se:

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

Valores e vectores próprios

Definição

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

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

Nota: Se $\mathbf{A}_{p\times p}$ fôr 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 sempre ortogonais.
- Mesmo que haja valores próprios repetidos, é possível determinar um conjunto ortonormado de p vectores próprios.

Teorema da Decomposição Espectral

Teorema (Decomposição espectral)

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

- uma matriz ortogonal $\mathbf{V}_{p\times p}$ (com colunas $\vec{\mathbf{v}}_i$); e
- uma matriz diagonal $\mathbf{\Lambda}_{p\times p}$ (com elementos diagonais λ_i),

tais que:

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

Notas:

- Os λ_i são valores próprios e os vectores $\vec{\mathbf{v}}_i$ são um conjunto ortonormado de (correspondentes) vectores próprios de $\mathbf{A}$.
- Se todos os valores próprios forem diferentes, os vectores 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 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, a decomposição é única (a menos de trocas de sinais nos vectores).

A Decomposição Espectral (cont.)

Notas:

- Com valores próprios iguais, a decomposição não é única. Qualquer combinação linear de vectores próprios associados ao valor próprio λ também será vector próprio de $\mathbf{A}$ associado a λ.
- Os valores próprios duma matriz simétrica determinam o sinal das suas formas quadráticas. De facto,
  $\vec{\mathbf{x}}^t\mathbf{A}\vec{\mathbf{x}} = \vec{\mathbf{x}}^t\left(\sum_{i=1}^{p}\lambda_i\vec{\mathbf{v}}_i\vec{\mathbf{v}}_i^t\right)\vec{\mathbf{x}} = \sum_{i=1}^{p}\lambda_i(\vec{\mathbf{x}}^t\vec{\mathbf{v}}_i)(\vec{\mathbf{v}}_i^t\vec{\mathbf{x}}) = \sum_{i=1}^{p}\lambda_i(\vec{\mathbf{x}}^t\vec{\mathbf{v}}_i)^2$. Logo,

Teorema

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

$\mathbf{A}$ definida positiva	$\iff$	$\lambda_i > 0, \forall i$
$\mathbf{A}$ semi-definida positiva	$\iff$	$\lambda_i \geq 0, \forall i$ (pelo menos um zero)
$\mathbf{A}$ definida negativa	$\iff$	$\lambda_i < 0, \forall i$
$\mathbf{A}$ semi-definida negativa	$\iff$	$\lambda_i \leq 0, \forall i$ (pelo menos um zero)
$\mathbf{A}$ é indefinida	$\iff$	$\exists i : \lambda_i < 0, \quad \exists j : \lambda_j > 0.$

Potências de matrizes simétricas

Seja $\mathbf{A}$ uma matriz simétrica, $p \times p$, e seja $\mathbf{A}_{p\times p} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^t$ a sua decomposição espectral.

- Para qualquer $k \in \mathbb{N}$, tem-se $\mathbf{A}^k = \mathbf{V}\boldsymbol{\Lambda}^k\mathbf{V}^t$, com $\boldsymbol{\Lambda}^k \equiv \left[\lambda_i^k\right]$.
- Se $\mathbf{A}$ invertível,
  - define-se $\mathbf{A}^0 = \mathbf{V}\boldsymbol{\Lambda}^0\mathbf{V}^t = \mathbf{V}\mathbf{V}^t = \mathbf{I}_{p\times p}$;
  - tem-se, para a inversa, que $\mathbf{A}^{-1} = \mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}^t$, com $\boldsymbol{\Lambda}^{-1} \equiv \left[\frac{1}{\lambda_i}\right]$;
  - para $k \in \mathbb{N}$, $\mathbf{A}^{-k} = \mathbf{V}\boldsymbol{\Lambda}^{-k}\mathbf{V}^t = (\mathbf{A}^k)^{-1} = (\mathbf{A}^{-1})^k$
- Se $\mathbf{A}$ definida positiva, define-se $\mathbf{A}^k = \mathbf{V}\boldsymbol{\Lambda}^k\mathbf{V}^t$ para qualquer $k \in \mathbb{R}$.

Para os expoentes que faz sentido em cada caso, definiu-se uma álgebra de potências análoga à dos números reais:

$$\mathbf{A}^k\mathbf{A}^m \quad = \quad \mathbf{A}^{k+m}\,.$$

Traços (de matrizes quadradas)

Pela Decomposição Espectral, é fácil ver que o traço duma matriz simétrica **A** é também a soma dos seus valores próprios.

- O traço é a soma dos elementos diagonais:

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

- O traço é um operador linear, isto é,

$$\mathrm{tr}(\alpha\,\mathbf{A} + \beta\,\mathbf{B}) \quad = \quad \alpha\,\mathrm{tr}(\mathbf{A}) + \beta\,\mathrm{tr}(\mathbf{B})$$

.

- O traço do produto matricial $\mathbf{A}_{n\times p}\mathbf{B}_{p\times n}$ é dado por:

$$\mathrm{tr}(\mathbf{AB}) \;=\; \sum_{i=1}^{n}(\mathbf{AB})_{ii} \;=\; \sum_{i=1}^{n}\sum_{j=1}^{p} a_{ij}b_{ji}\;.$$

Circularidade do traço

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

(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: $\mathbf{A}_{m\times k}, \mathbf{B}_{k\times p}, \mathbf{C}_{p\times m} \Rightarrow \operatorname{tr}(\mathbf{ABC}) = \operatorname{tr}(\mathbf{BCA})$.

(Aplicar o resultado anterior às duas matrizes **A** e **BC**)

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

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

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

Matrizes de (co-)variâncias e correlações

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

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

onde $\mathbf{X}^{c}$ é a matriz $n \times p$ cujas colunas são os p vectores centrados de observações.

As matrizes de correlações são análogas, mas usando matrizes $\mathbf{Z}$ de dados centrados e reduzidos:

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

Matrizes de (co-)variâncias e correlações (cont.)

Mais do que simétricas, as matrizes de covariâncias e de correlações são necessariamente semi-definidas positivas:

$$\vec{\mathbf{a}}^t \mathbf{S} \vec{\mathbf{a}} = \tfrac{1}{n-1} \vec{\mathbf{a}}^t \mathbf{X}^{c\,t} \mathbf{X}^c \vec{\mathbf{a}} = \tfrac{1}{n-1} \left\| \mathbf{X}^c \vec{\mathbf{a}} \right\|^2 \geq 0 \,.$$

São definidas positivas se e só se as p colunas da matriz $\mathbf{X}^c$ (ou $\mathbf{Z}$) são linearmente independentes (i.e., sse não há multicolinearidade), pois nesse caso:

$$\vec{\mathbf{a}}^t \mathbf{S} \vec{\mathbf{a}} = 0 \quad \Leftrightarrow \quad \left\| \mathbf{X}^c \vec{\mathbf{a}} \right\|^2 = 0 \quad \Leftrightarrow \quad \mathbf{X}^c \vec{\mathbf{a}} = \vec{\mathbf{0}}_n \quad \Leftrightarrow \quad \vec{\mathbf{a}} = \vec{\mathbf{0}}_p \,.$$

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

A centragem da matriz de dados

Seja

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

a matriz $(n \times n)$ de projecção ortogonal sobre o subespaço (de $\mathbb{R}^n$) gerado pelo vector $\vec{1}_n$ dos n uns.

Para qualquer vector $\vec{\mathbf{a}}$, o produto $\mathbf{P}_{\vec{1}_n}\vec{\mathbf{a}}$ é o vector que repete n vezes a média dos elementos de $\vec{\mathbf{a}}$.

Para qualquer matriz $\mathbf{A}_{n\times p}$, o produto $\mathbf{P}_{\vec{1}_n}\mathbf{A}$ é a matriz $n \times p$ que, na coluna j, repete n vezes a média dos valores da coluna j de $\mathbf{A}$.

Seja $\mathbf{X}$ uma matriz $n \times p$ de dados. A correspondente matriz centrada de dados, $\mathbf{X}^c$, obtém-se através do produto:

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

Variância de combinações lineares de variáveis

Seja **S** a matriz de (co-)variâncias gerada pela matriz de dados **X**.

A variância duma combinação linear das colunas de **X**, $\mathbf{X}\vec{\mathbf{a}}$, é a forma quadrática:

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

De facto,

$$\vec{\mathbf{a}}^t\mathbf{S}\vec{\mathbf{a}} = \tfrac{1}{n-1}\vec{\mathbf{a}}^t\mathbf{X}^{c\,t}\mathbf{X}^c\vec{\mathbf{a}} = \tfrac{1}{n-1}\left\|\mathbf{X}^c\vec{\mathbf{a}}\right\|^2 \ .$$

Mas $\mathbf{X}^c\vec{\mathbf{a}} = (\mathbf{I}_n - \mathbf{P}_{\vec{\mathbf{1}}_n})\mathbf{X}\vec{\mathbf{a}}$ é o vector centrado da combinação linear $\vec{\mathbf{y}} = \mathbf{X}\vec{\mathbf{a}}$, de elemento genérico $y_i^c = y_i - \overline{y}$. Logo,

$$\tfrac{1}{n-1}\left\|\mathbf{X}^c\vec{\mathbf{a}}\right\|^2 = \tfrac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 \ ,$$

que é a variância amostral da combinação linear $\vec{\mathbf{y}} = \mathbf{X}\vec{\mathbf{a}}$.

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

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

Relação entre **S** e **R**

Se **D** diagonal e **A** é uma matriz compatível na multiplicação, então:

- **AD** é a matriz que resulta de multiplicar cada *coluna* de **A** pelo correspondente elemento diagonal de **D**.
- **DA** é a matriz que resulta de multiplicar cada *linha* de **A** pelo correspondente elemento diagonal de **D**.

Seja $\mathbf{X}^c{}_{n\times p}$ uma matriz de dados centrada. Seja **D** uma matriz diagonal cujos elementos são os recíprocos do desvios-padrão de cada variável ($d_{ii} = \frac{1}{s_i}$). Então, a matriz de dados normalizados **Z** surge do produto:

$$\mathbf{Z} = \mathbf{X}^c\mathbf{D}.$$

Assim (e como para qualquer matriz diagonal, $\mathbf{D}^t = \mathbf{D}$), tem-se:

$$\mathbf{R} = \tfrac{1}{n-1}\mathbf{Z}^t\mathbf{Z} = \tfrac{1}{n-1}\mathbf{D}\mathbf{X}^{c\,t}\mathbf{X}^c\mathbf{D} = \mathbf{DSD}\ .$$

ACP: Uma introdução estatística

Uma introdução frequente à ACP usa conceitos estatísticos.

Seja dada a matriz de dados $n \times p$, $\mathbf{X}$.

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

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

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

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 $\vec{\mathbf{v}}^t \mathbf{S} \vec{\mathbf{v}}$.

Introdução estatística (cont.)

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

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

Assim, gera-se o problema de maximizar o chamado quociente de Rayleigh-Ritz de $\mathbf{S}$:

$$\max_{\vec{\mathbf{v}} \in \mathbb{R}^p \setminus \{\vec{\mathbf{0}}\}} \frac{\vec{\mathbf{v}}^t \mathbf{S} \vec{\mathbf{v}}}{\vec{\mathbf{v}}^t \vec{\mathbf{v}}}$$

A solução é dada pelo vector próprio $\vec{\mathbf{v}}_1$, 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: $\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{p-1} \geq \lambda_p = \lambda_{\min}$.

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

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

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

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

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

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

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

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

A primeira Componente Principal

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

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

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

É a direcção (não o sentido) 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. Quanto maior fôr, 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 $\vec{\mathbf{y}} = \mathbf{X}\vec{\mathbf{v}}$, com $\vec{\mathbf{v}}^t\vec{\mathbf{v}} = 1$, de variância máxima, não-correlacionada com a anterior.

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

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

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

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

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: $var(\vec{\mathbf{y}}_j) = \lambda_j$.

A ACP no R

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 vária razões (incluindo a precisão numérica no caso de matrizes de (co-)variâncias quase singulares), é preferível a utilização do comando `prcomp`.

O comando prcomp

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

```
Standard deviations:
[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:
                   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`

> summary(lav.acp)

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12    PC13
Standard deviation     4.4170 2.1580 0.9618 0.70720 0.61640 0.49926 0.46399 0.38484 0.33629 0.25007 0.20606 0.17704 0.14058
Proportion of Variance 0.7270 0.1734 0.0344 0.00186 0.00141 0.00928 0.00802 0.00551 0.00421 0.00233 0.00158 0.00117 0.00074
Cumulative Proportion  0.7270 0.9004 0.9344 0.95300 0.96720 0.97645 0.98446 0.98998 0.99419 0.99652 0.99810 0.99926 1.00000
```

Na recta definida pela primeira componente principal preservamos $72,7\%$ da variabilidade total dos dados.

No plano definido pelas 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 que cada individuo toma na combinação linear $\vec{\mathbf{y}}_j = \mathbf{X}\vec{\mathbf{v}}_j$.

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

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

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

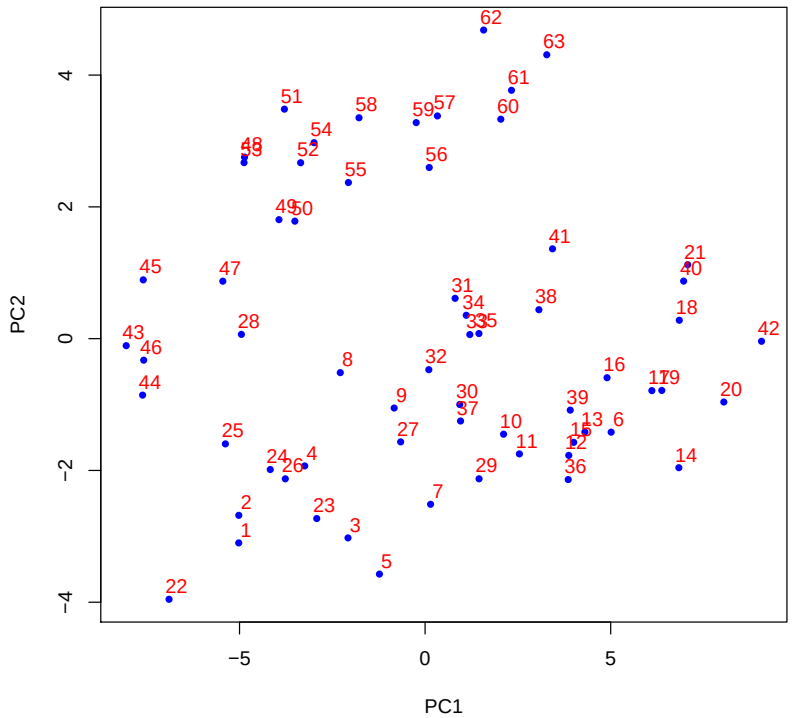
```
> lav.acp$x
```

```
          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

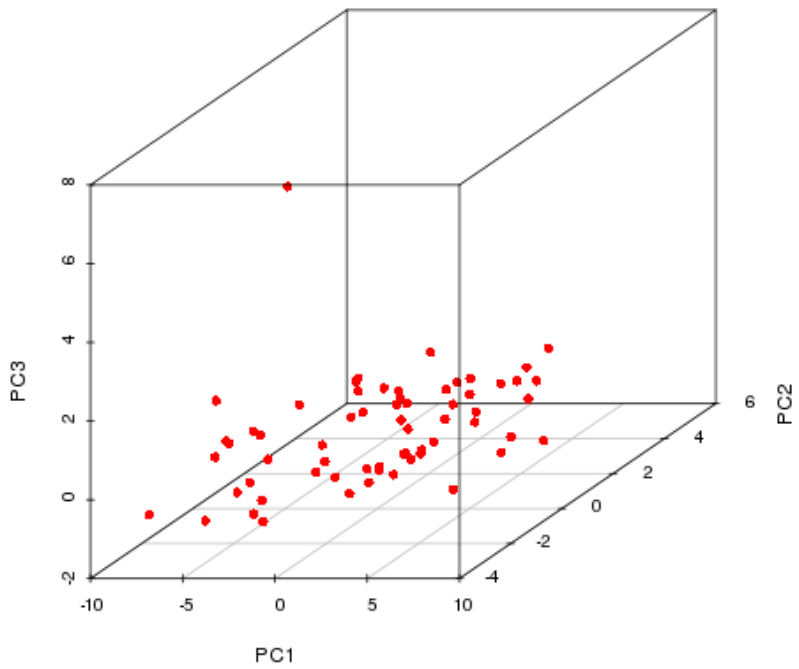
```
> plot(lav.acp$x[,1:2],col="blue", pch=16, cex=0.8)
> text(lav.acp$x[,1:2]+0.2, label=rownames(lavagantes), col="red")
```



Os individuos 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

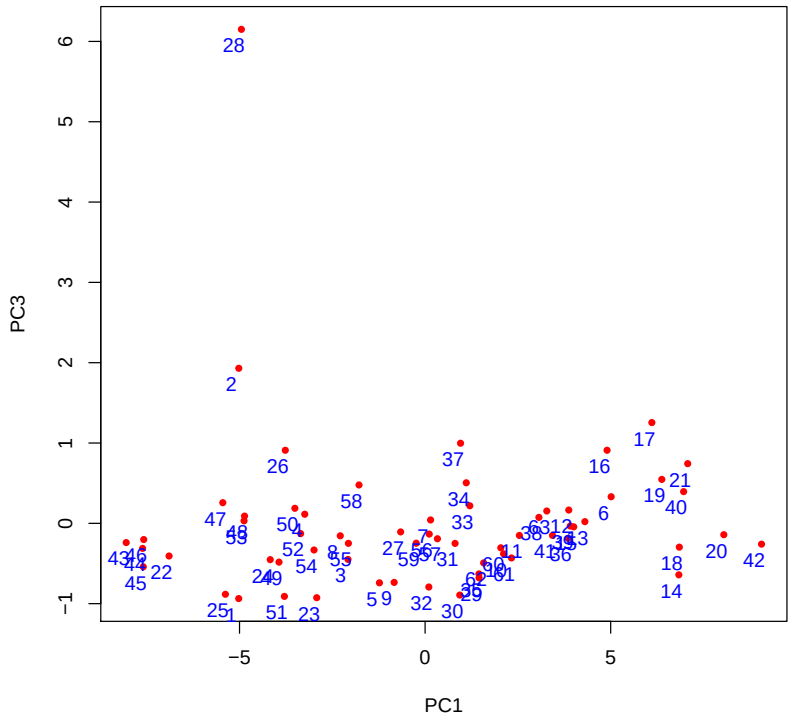
Pode ser obtida usando o módulo `rggobi` (embora o gráfico abaixo seja obtido doutra forma):



No gráfico verifica-se que a terceira CP separa uma observação discrepante (*outlier*) das restantes. Acontece com alguma frequência em CPs posteriores.

A observação 28 e a terceira CP

```
> 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")
```

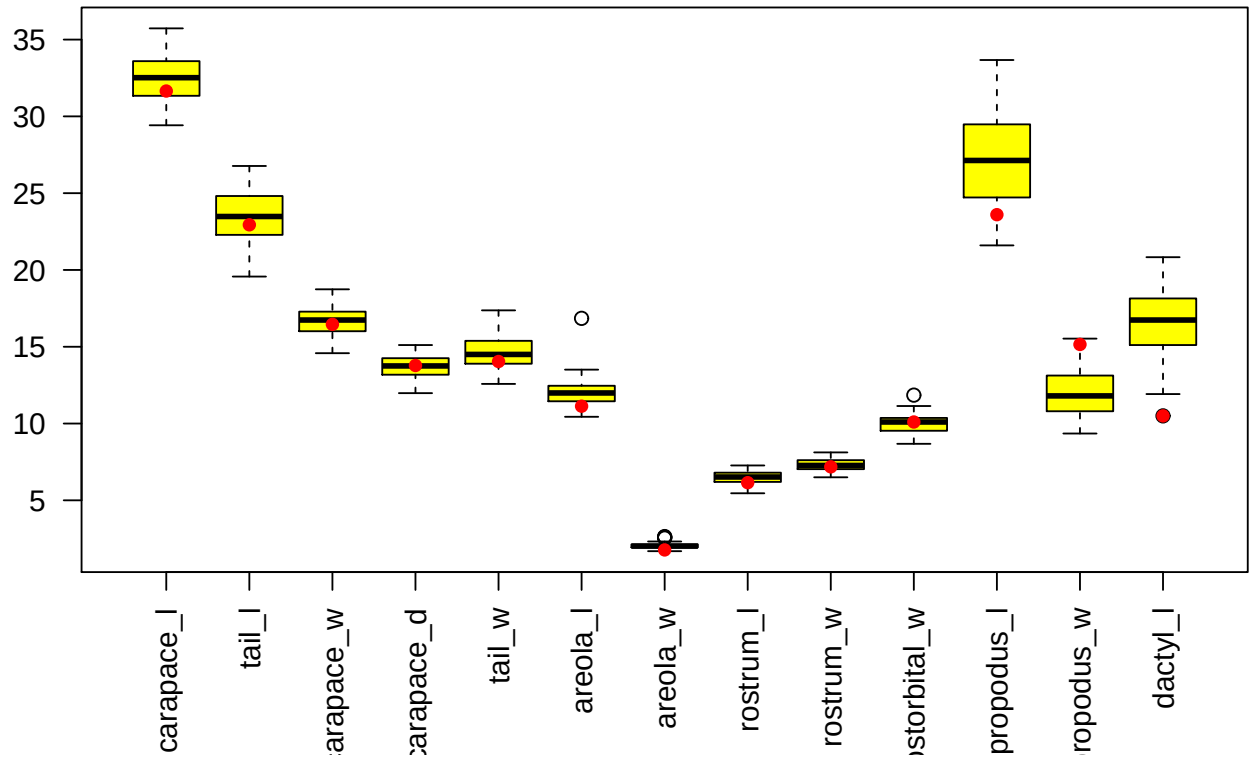


O individuo 28 contribui decisivamente para a terceira direcção ortogonal de maior variabilidade.

Revisitando o individuo 28

O que tem de diferente o individuo 28?

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



Decomposição em valores/vectores próprios

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

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

```
$values
[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
             [,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
[4,] -0.13951311 -0.14784642 -0.13138041  0.01598041  0.105009202  0.171612241
[...]

> 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
carapace_d  0.13951311  0.14784642  0.13138041  0.01598041  0.105009202
[...]
```

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

Mais propriedades de CPs

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

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

$$\begin{array}{rcl} s_i & - & \text{desvio padrão da variável } \vec{\mathbf{x}}_i \\ v_{ij} & - & \text{coeficiente de } \vec{\mathbf{x}}_i \text{ na CP } j \\ \lambda_j & - & \text{variância da } j\text{-ésima CP} \end{array}$$

Nota: A covariância entre duas combinações lineares das colunas de $\mathbf{X}$, como o são $\mathbf{X}\vec{\mathbf{v}}_j$ e $\vec{\mathbf{x}}_i = \mathbf{X}\vec{\mathbf{e}}_i$, (com $\vec{\mathbf{e}}_i$ o i-ésimo vector da base canónica de $\mathbb{R}^p$) é dada por $\vec{\mathbf{e}}_i^t \mathbf{S} \vec{\mathbf{v}}_j$, onde $\mathbf{S}$ é a matriz de (co-)variâncias dos dados. Logo,

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

Interpretação de CPs

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

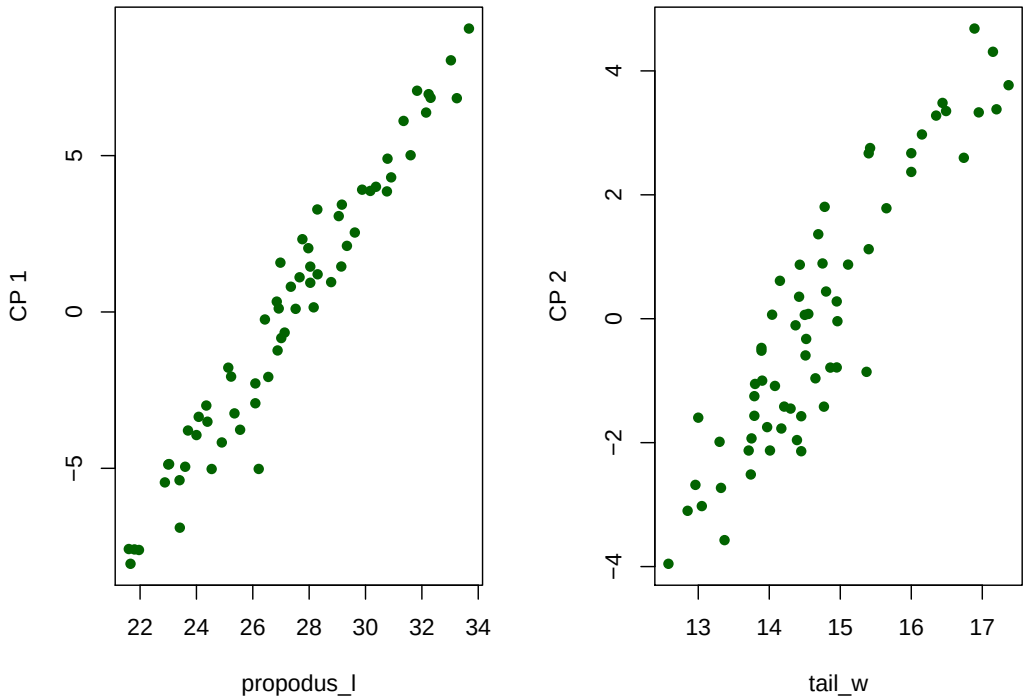
```
> round(cor(lavagantes, lav.acp$x),d=2)
```

	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

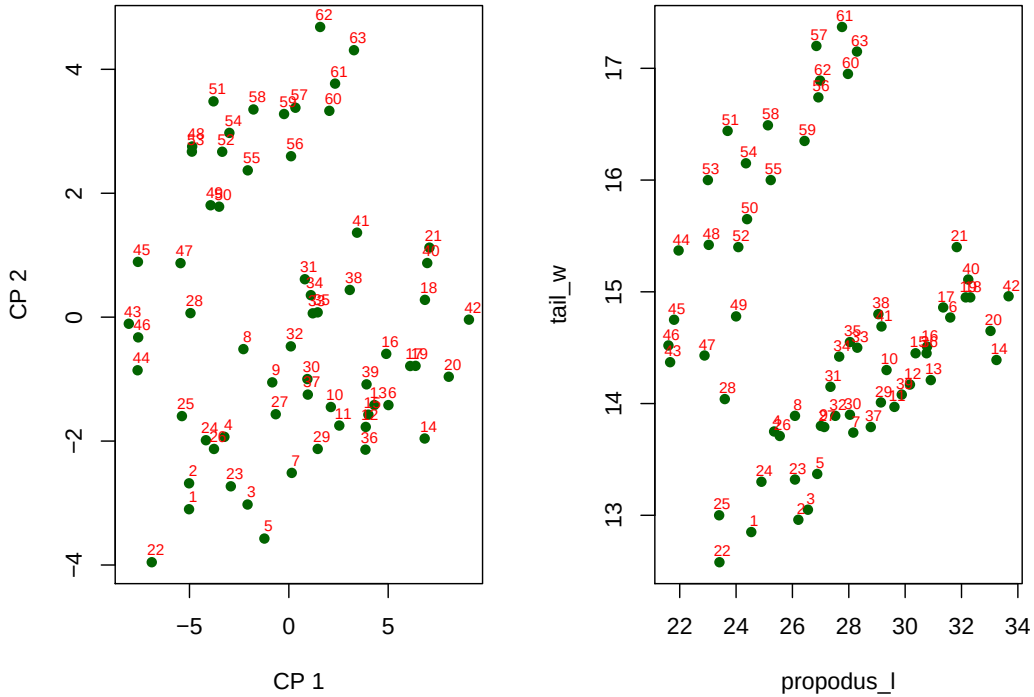
```
> par(mfrow=c(1,2))            <- cria uma ‘‘matriz 1x2  de gráficos’’
> plot(lavagantes[,11], lav.acp$x[,1], xlab="propodus_l", ylab="CP 1", pch=16, col="darkgreen")
> plot(lavagantes[,5], lav.acp$x[,2], xlab="tail_w", ylab="CP 2", pch=16, col="darkgreen")
> par(mfrow=c(1,1))            <- re-estabelece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



Correlações entre CPs e variáveis (cont.)

As fortes correlações sugerem uma análise da nuvem de pontos definida pelas 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 optimizar: a variância.

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

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

onde

- x_{ij} é a observação do individuo i na variável j;
- $\overline{x}_{.j}$ 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.

ACP sobre a matriz de correlações (cont.)

A matriz de (co)variâncias dos dados normalizados é a matriz de correlações **R** dos dados originais (ou normalizados). 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 **R**;
- As variâncias de sucessivas CPs são dadas pelos valores próprios de **R**, cuja soma tem de ser $\mathrm{tr}(\mathbf{R}) = p$.

ACP sobre a matriz de correlações no R

No R, há duas formas alternativas de efectuar esta Análise:

```
> prcomp(scale(lavagantes))
> 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 ACP não são directamente comparáveis.

```
> 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 0.14058
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 0.00074
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 1.00000
```

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:

```
> 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.)

Vejamos as correlações entre as CPs sobre os dados normalizados e as variáveis originais:

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

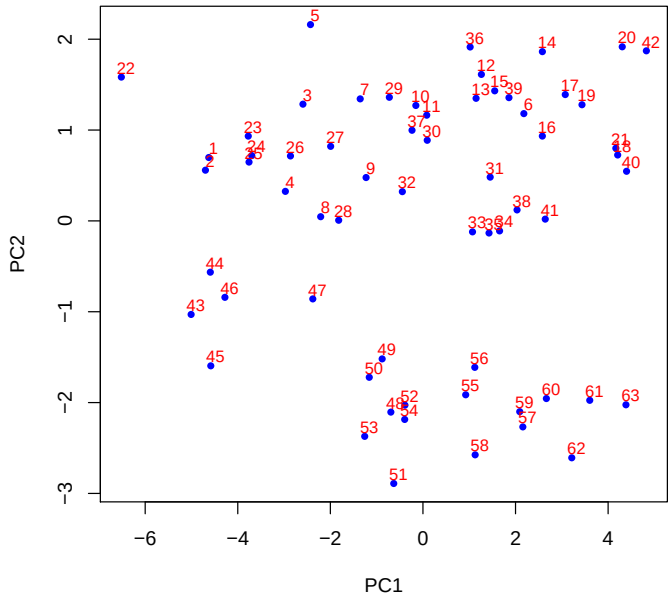
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
carapace_l	0.94	-0.07	0.00	-0.04	0.04	-0.14	0.07	-0.11	-0.03	-0.06	-0.24	0.05	-0.04
tail_l	0.66	-0.66	0.00	0.02	-0.04	0.04	-0.17	0.05	-0.24	0.19	-0.01	0.00	0.00
carapace_w	0.96	-0.04	0.04	-0.04	0.07	-0.10	0.02	-0.10	0.08	0.07	-0.03	-0.16	0.06
carapace_d	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

Notas:

- 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

```
> 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")
```

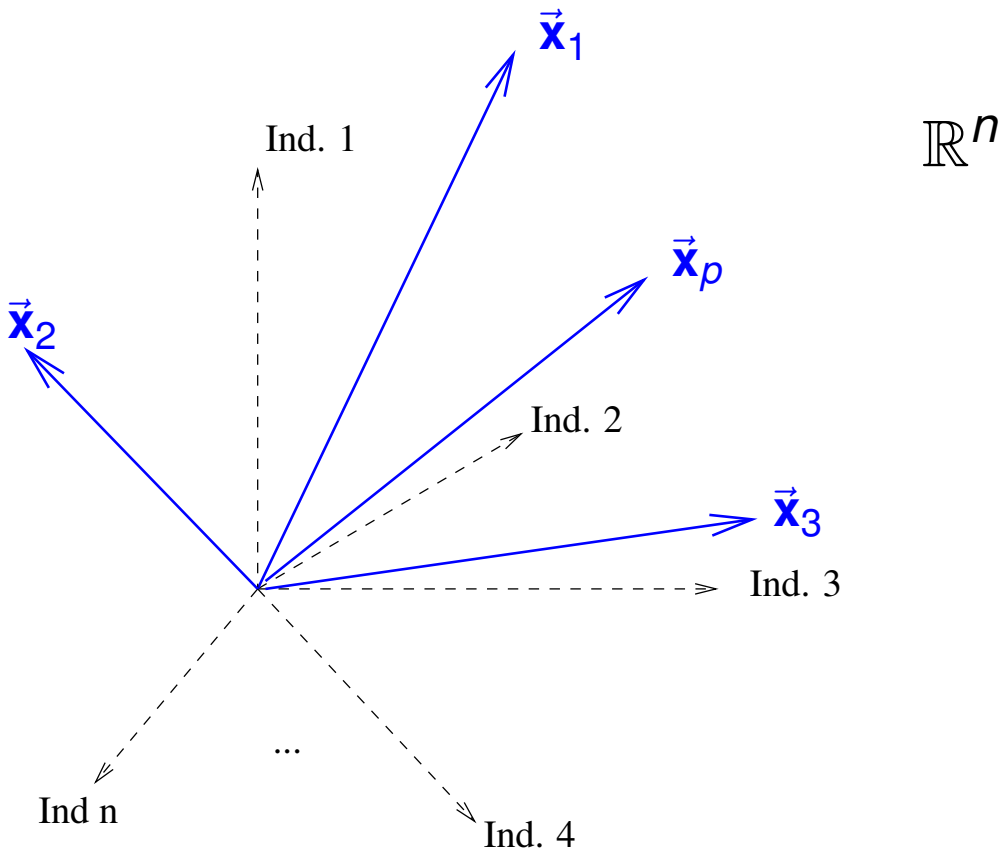


Em síntese: a CP 1 ordena por tamanho geral do organismo, e a CP 2 faz a separação de sexos.

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

Representação alternativa duma matriz de dados **X**, no espaço das variáveis:

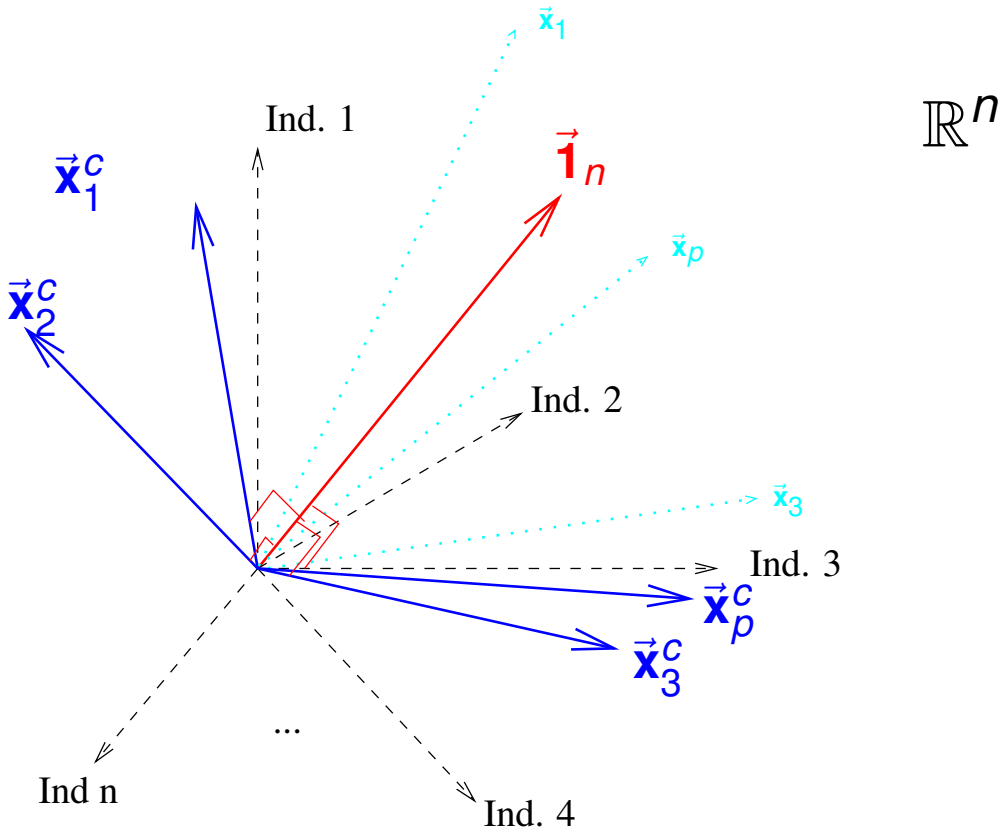
- cada eixo corresponde a um indivíduo observado;
- cada vector corresponde a uma variável.



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:



Geometria e estatística no espaço das variáveis

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

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

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

Logo,

- a norma usual duma coluna de $\mathbf{X}^c$ é proporcional ao desvio padrão dessa variável: $\|\vec{\mathbf{x}}_j^c\| = \sqrt{n-1}\, s_j$.
- o produto interno usual de duas diferentes colunas de $\mathbf{X}^c$ é proporcional à covariância das respectivas variáveis: $< \vec{\mathbf{x}}_j^c, \vec{\mathbf{x}}_k^c >= (n-1)\,\mathrm{cov}_{i,j}$.
- o cosseno do ângulo entre os vectores representando duas diferentes colunas de $\mathbf{X}^c$ é o coeficiente de correlação das respectivas variáveis: $\cos\theta = \frac{<\vec{\mathbf{x}}_j^c, \vec{\mathbf{x}}_k^c>}{\|\vec{\mathbf{x}}_i^c\| \cdot \|\vec{\mathbf{x}}_j^c\|} = \frac{\mathrm{cov}_{i,j}}{s_i \cdot s_j} = r_{i,j}$
- Vectores centrados ortogonais correspondem a variáveis não correlacionadas.

Intepretação de CPs 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.

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

O critério da ACP 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").

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

Transformações lineares (afins) nas variáveis correspondem a re-escalar os vectores que as representam (mantendo a direcção):

- constantes aditivas desaparecem na centragem, logo não alteram o vector correspondente em $\mathbb{R}^n$.
- constantes multiplicativas c correspondem a multiplicar o vector correspondente por c, logo:
  - preservam a direcção do vector representativo;
  - mudam o sentido se $c < 0$;
  - esticam o vector se $|c| > 1$;
  - encolhem o vector se $|c| < 1$.

A normalização dos dados associada à ACP sobre a matriz de correlações corresponde a tornar todos os vectores, representativos das variáveis centradas, do mesmo tamanho.

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

A matriz centrada, $\mathbf{X}^c$, na representação em $\mathbb{R}^p$

Qual o efeito de centrar a matriz $\mathbf{X}$ na nuvem de pontos associada aos 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:

$$(\overline{x_1}, \overline{x_2}, ..., \overline{x_p}) \quad \longrightarrow \quad (0, 0, ..., 0) \ .$$

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

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 $\mathbf{S}$ (vectores de *loadings*), gerando confusão.
- Não faz sentido que qualquer das variáveis originais seja uma variável *qualitativa* (*categórica*).

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 $\mathrm{tr}(\mathbf{R}) = p$ (o número de variáveis).
- A correlação entre a variável $\vec{\mathbf{x}}_j$ 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 $\vec{\mathbf{v}}_j^t \vec{\mathbf{v}}_j = \lambda_j$. Nesse caso, os coeficientes da combinação linear são as correlações entre a variável e a CP.

Outro critério

As CPs da matriz de correlações são também solução óptima de outro problema: determinar a combinação linear que maximiza a soma de quadrados das p correlações com cada variável original.

O vector de correlações entre cada variável $\vec{\mathbf{x}}_j = \mathbf{X}\vec{\mathbf{e}}_j$ e a combinação linear $\mathbf{X}\vec{\mathbf{v}}$ é:

$$\vec{\mathbf{r}} \equiv \left[\frac{\vec{\mathbf{e}}_j^t \mathbf{S}\vec{\mathbf{v}}}{s_j \cdot \sqrt{\vec{\mathbf{v}}^t \mathbf{S}\vec{\mathbf{v}}}}\right] \quad \Leftrightarrow \quad \vec{\mathbf{r}} = \frac{1}{\sqrt{\vec{\mathbf{v}}^t \mathbf{S}\vec{\mathbf{v}}}} \mathbf{D}\mathbf{S}\vec{\mathbf{v}} \ ,$$

onde $\mathbf{D} \equiv diag[\frac{1}{s_j}]$.

A soma de quadrados destas p correlações é:

$$\|\vec{\mathbf{r}}\|^2 = \vec{\mathbf{r}}^t\vec{\mathbf{r}} = \frac{\vec{\mathbf{v}}^t \mathbf{S}\mathbf{D} \cdot \mathbf{D}\mathbf{S}\vec{\mathbf{v}}}{\vec{\mathbf{v}}^t \mathbf{S}\vec{\mathbf{v}}}$$

Ora, $\mathbf{R} = \mathbf{D}\mathbf{S}\mathbf{D}$, pelo que tomando $\vec{\mathbf{v}} = \mathbf{D}\vec{\mathbf{b}}$, tem-se:

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

Problema generalizado de valores próprios

Teorema

Seja $\mathbf{A}_{p\times p}$ uma matriz simétrica, e $\mathbf{B}_{p\times p}$ uma matriz definida positiva.

- 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, $\vec{\mathbf{x}}_i^t\mathbf{B}\vec{\mathbf{x}}_j = 0$, se $i \neq j$ e $\vec{\mathbf{x}}_i^t\mathbf{B}\vec{\mathbf{x}}_j = 1$ se $i = j$.

De novo o critério alternativo

Vimos que a combinação linear $\mathbf{X}\vec{\mathbf{v}}$ "globalmente mais correlacionada com as p variáveis originais" maximiza o quociente $\|\vec{\mathbf{r}}\|^2 = \frac{\vec{\mathbf{b}}^t\mathbf{R}^2\vec{\mathbf{b}}}{\vec{\mathbf{b}}^t\mathbf{R}\vec{\mathbf{b}}}$ com $\vec{\mathbf{v}} = \mathbf{D}\vec{\mathbf{b}}$ (sendo $\mathbf{R} = \mathbf{DSD}$ a matriz de correlações das p variáveis).

A solução é dada pelo primeiro par próprio de $\mathbf{R}^{-1}\mathbf{R}^2 = \mathbf{R}$:
$\vec{\mathbf{y}}_1 = \mathbf{X}\vec{\mathbf{v}} = \mathbf{XD}\vec{\mathbf{b}}_1$, a primeira CP sobre a matriz de correlações de $\mathbf{X}$, $\mathbf{R}$.

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

As restantes CPs da matriz de correlações são as combinações lineares que sucessivamente maximizam o quociente, sujeitas às restrições de ortogonalidade (usual) com todas as soluções anteriores.

O critério alternativo (cont.)

O critério é invariante a mudanças lineares de escala nas variáveis, porque não depende das unidades de medida das variáveis originais: envolve a soma de quadrados de correlações entre as combinações lineares e as variáveis originais.

Interpretação geométrica do novo critério, em $\mathbb{R}^n$: procuram-se as combinações lineares das variáveis que maximizam a soma dos quadrados dos cossenos dos ângulos com os vectores das variáveis originais. Como transformações lineares (afins) das variáveis não alteram esses ângulos, o critério fica invariante.

As combinações lineares centradas das variáveis reduzidas $\mathbf{X}^c\mathbf{D}\vec{\mathbf{b}}$ (onde $\vec{\mathbf{b}}$ é vector próprio de $\mathbf{R}$):

- só são componentes principais das variáveis reduzidas;
- mas são as combinações lineares sucessivamente "globalmente mais correlacionadas com as variáveis originais", independentemente das unidades de medida originais.

Outra introdução à ACP

Decomposição em Valores Singulares

Seja $\mathbf{Y}_{n\times p}$ $(n \geq p)$ uma matriz genérica, de característica r. É sempre possível factorizar $\mathbf{Y}$ da seguinte forma:

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

onde

$\mathbf{\Delta}_{r\text{x}r}$ matriz diagonal

$\mathbf{V}_{p\text{x}r}$ matriz com colunas ortonormadas $(\mathbf{V}^t\mathbf{V} = \mathbf{I}_r)$

$\mathbf{W}_{n\text{x}r}$ matriz com colunas ortonormadas $(\mathbf{W}^t\mathbf{W} = \mathbf{I}_r)$

δ_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 $\mathbf{Y}$)

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

- $\mathbf{Y}^t$ tem Decomposição em Valores Singulares $\mathbf{Y}^t = \mathbf{V}\boldsymbol{\Delta}\mathbf{W}^t$.
- $\mathbf{Y}^t\mathbf{Y} = \mathbf{V}\boldsymbol{\Delta}^2\mathbf{V}^t$ 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 $\mathbf{Y}^t\mathbf{Y}$, associados a valores próprios não-nulos.
- $\mathbf{W}$ é a matriz análoga, de vectores próprios de $\mathbf{Y}\mathbf{Y}^t = \mathbf{W}\boldsymbol{\Delta}^2\mathbf{W}^t$.
- $\boldsymbol{\Delta}$ é a matriz das raízes quadradas dos valores próprios não-nulos de $\mathbf{Y}^t\mathbf{Y}$ (iguais aos de $\mathbf{Y}\mathbf{Y}^t$). Admite-se que os valores singulares δ_i estão ordenados por ordem decrescente.
- A Decomposição em Valores Singulares (DVS) é válida para qualquer matriz.
- A DVS duma matriz é sempre possível, mas não é única (pelo menos a troca de sinal nos pares de vectores).

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):

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

onde

- $\mathbf{V}$ - matriz cujas colunas são os vectores próprios de $\mathbf{S} = \frac{1}{n-1}\mathbf{X}^{c\,t}\mathbf{X}^c$, 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 individuo em cada CP.
- $\mathbf{U} = \frac{1}{\sqrt{n-1}}\mathbf{X}^c\mathbf{V}\Delta^{-1}$ - matriz dos vectores singulares esquerdos, que corresponde aos vectores de *scores* normalizados.

DVS e ACP (cont.)

Iremos confirmar, efectuando a DVS da matriz $\frac{1}{\sqrt{n-1}}\mathbf{X}^c$ no R.

A centragem duma 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 duma 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`.

ACP e DVS (cont.)

No R uma Decomposição em Valores Singulares faz-se através do comando `svd`. Vejamos uma ACP feita com este comando:

```
> 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** referidas acima. A componente `$d` é um vector, com os elementos diagonais da matriz **Δ**.

ACP e DVS (cont.)

O produto das matrizes **U** e **Δ**:

```
> DVS <- svd(lav.centrado/sqrt(62))
> U <- DVS$u
> D <- diag(DVS$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
[...]
```

```
> 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
[...]
```

Teorema de Eckart-Young

Teorema

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

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

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

Nota 2: O critério usado minimiza, simultaneamente, a distância entre colunas correspondentes, e entre linhas correspondentes, das matrizes $\mathbf{X}$ e $\mathbf{Y}$.

Teorema de Eckart-Young (cont.)

Aplicando o Teorema de Eckart-Young às matrizes de dados centradas $\mathbf{X}^c$:

- para qualquer m $(1 \leq m \leq car(\mathbf{X}^c))$, a matriz $\mathbf{X}^c_{(m)}$ que resulta de reter apenas as parcelas na DVS de $\mathbf{X}^c$ com os m maiores valores singulares, é a matriz $n \times p$ de característica m, globalmente mais próxima de $\mathbf{X}$.
- O significado geométrico de $\mathbf{X}^c_{(m)}$ ter característica m é que as suas colunas estão num subespaço de dimensão m de $\mathbb{R}^n$, e as suas linhas estão num subespaço de dimensão m de $\mathbb{R}^p$.
- Estes subespaços são gerados pelos m vectores singulares esquerdos (em $\mathbb{R}^n$) e direitos (em $\mathbb{R}^p$), associados aos m maiores valores singulares. Ou seja, são gerados pelas m primeiras CPs (em $\mathbb{R}^n$) e respectivos vectores de *loadings* (em $\mathbb{R}^p$).
- Assim, a ACP (a DVS de $\mathbf{X}^c$) identifica, simultaneamente em $\mathbb{R}^p$ e em $\mathbb{R}^n$, os subespaços de dimensão m onde a representação dos dados originais é o mais fidedigna possível, no sentido de ser globalmente mais próxima dos valores originais.

Projecções sobre subespaços

No Módulo II (Modelo Linear) falámos de projecções ortogonais sobre subespaços de espaços vectoriais $\mathbb{R}^k$.

Em particular, viu-se que se $\mathbf{B}$ é uma matriz cujas m colunas formam uma base dum subespaço m-dimensional $\mathscr{B}$ de $\mathbb{R}^k$, a matriz $\mathbf{P} = \mathbf{B}(\mathbf{B}^t\mathbf{B})^{-1}\mathbf{B}^t$ projecta os vectores de $\mathbb{R}^k$ ortogonalmente sobre $\mathscr{B}$.

Se os vectores da base formarem um conjunto ortonormado, tem-se $\mathbf{B}^t\mathbf{B} = \mathbf{I}_m$, pelo que a matriz de projecção ortogonal fica apenas:

$$\mathbf{P} = \mathbf{B}\mathbf{B}^t = \sum_{j=1}^{m} \vec{\mathbf{b}}_j \vec{\mathbf{b}}_j^t \, .$$

A matriz $\mathbf{X}^{\mathrm{c}}_{(\mathrm{m})}$ dos acetatos anteriores é a matriz que se obtém:

- projectando ortogonalmente as colunas de $\mathbf{X}^{\mathrm{c}}$ sobre o subespaço gerado pelas m primeiras CPs;
- projectando ortogonalmente as linhas de $\mathbf{X}^{\mathrm{c}}$ sobre o subespaço gerado pelos m primeiros vectores próprios de $\mathbf{S}$ (ou $\mathbf{R}$).

Projecções em $\mathbb{R}^n$

Cada CP (vector de *scores*) é um múltiplo escalar do correspondente vector singular esquerdo, $\vec{\mathbf{u}}_i$. O conjunto desses vectores $\vec{\mathbf{u}}_i$ é um conjunto ortonormado de vectores.

Logo, a matriz de projecção ortogonal sobre o subespaço gerado pelas m primeiras CPs é $\mathbf{P}_u = \mathbf{U}_m \mathbf{U}_m^t = \sum_{j=1}^{m} \vec{\mathbf{u}}_j \vec{\mathbf{u}}_j^t$.

A projecção das colunas de $\frac{1}{\sqrt{n-1}}\mathbf{X}^c$ é dada por:

$$\begin{aligned} \mathbf{P}_u \tfrac{1}{\sqrt{n-1}}\mathbf{X}^c &= \left(\sum_{j=1}^{m} \vec{\mathbf{u}}_j \vec{\mathbf{u}}_j^t\right)\left(\sum_{i=1}^{r} \delta_i \vec{\mathbf{u}}_i \vec{\mathbf{v}}_i^t\right) = \sum_{j=1}^{m}\sum_{i=1}^{r} \delta_i \vec{\mathbf{u}}_j \vec{\mathbf{u}}_j^t \vec{\mathbf{u}}_i \vec{\mathbf{v}}_i^t \\ &= \sum_{j=1}^{m} \delta_j \vec{\mathbf{u}}_j \vec{\mathbf{v}}_j^t = \mathbf{X}^c{}_{(m)} \ , \end{aligned}$$

já que $\vec{\mathbf{u}}_i^t \vec{\mathbf{u}}_j = 1$ se $i = j \leq m$ e 0 noutros casos.

Projecções em $\mathbb{R}^p$

Os vectores próprios de **S** (ou **R**), $\vec{\mathbf{v}}_j$, formam um conjunto o.n. de vectores. Logo, a matriz de projecção ortogonal sobre o subespaço gerado pelos m primeiros vectores próprios de **S** (ou **R**) é $\mathbf{P}_v = \mathbf{V}_m \mathbf{V}_m^t$.

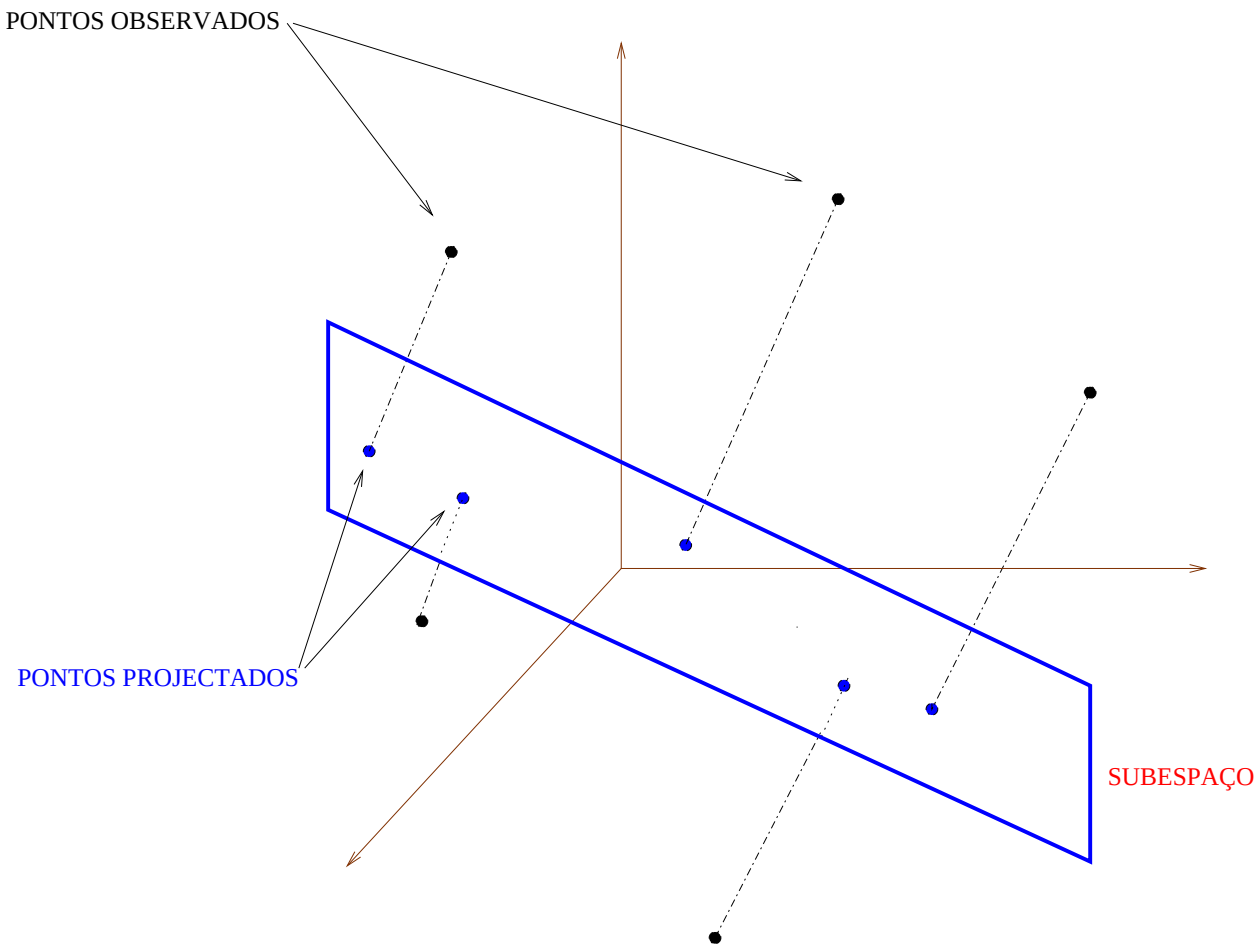
A projecção das linhas de $\frac{1}{\sqrt{n-1}}\mathbf{X}^c$ equivale às projecção das colunas da sua transposta, cuja SVD é $\sum_{i=1}^{r} \delta_i \vec{\mathbf{v}}_i \vec{\mathbf{u}}_i^t$. Logo,

$$\begin{aligned}
\mathbf{P}_v \left(\tfrac{1}{\sqrt{n-1}}\mathbf{X}^c\right)^t &= \left(\sum_{j=1}^{m} \vec{\mathbf{v}}_j \vec{\mathbf{v}}_j^t\right)\left(\sum_{i=1}^{r} \delta_i \vec{\mathbf{v}}_i \vec{\mathbf{u}}_i^t\right) = \sum_{j=1}^{m}\sum_{i=1}^{r} \delta_i \vec{\mathbf{v}}_j \vec{\mathbf{v}}_j^t \vec{\mathbf{v}}_i \vec{\mathbf{u}}_i^t \\
&= \sum_{j=1}^{m} \delta_j \vec{\mathbf{v}}_j \vec{\mathbf{u}}_j^t = {\mathbf{X}^c_{(m)}}^t ,
\end{aligned}$$

As linhas de $\mathbf{X}^c_{(m)}$ correspondem à projecção das linhas de $\mathbf{X}^c$ sobre o subespaço de $\mathbb{R}^p$ gerado pelos m primeiros vectores próprios de **S** (**R**).

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 m 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 duma matriz centrada de dados (logo, com uma ACP).
- Ideia fundamental do *biplot*: obter uma boa representação simultânea (aproximada) dos indivíduos e das variáveis, em baixa dimensão.
- Nessa representação simultânea as principais características estatísticas duma ACP serão visíveis em termos geométricos.

Biplots (cont.)

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

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

  tem-se: $\frac{1}{\sqrt{n-1}}\mathbf{X}^c = \mathbf{G}\mathbf{H}^t$.
- Se $\mathbf{X}^c$ é de característica r,
  - $\mathbf{G}$ é $n \times r$, e existe correspondência entre linhas de $\mathbf{G}$ e indivíduos.
  - $\mathbf{H}$ é $p \times r$, e existe correspondência entre linhas de $\mathbf{H}$ e variáveis.
- As linhas de $\mathbf{G}$ e $\mathbf{H}$ são marcadores de, respectivamente, indivíduos e variáveis, mas vivem no mesmo espaço ($\mathbb{R}^r$) e podem ser representadas em simultâneo.

Os marcadores de variáveis

Consideremos as propriedades dos marcadores de variáveis, que são vectores de $\mathbb{R}^r$. Tem-se:

$$\mathbf{H}\mathbf{H}^t = \mathbf{S} \quad (\text{ou } \mathbf{R}) .$$

Logo,

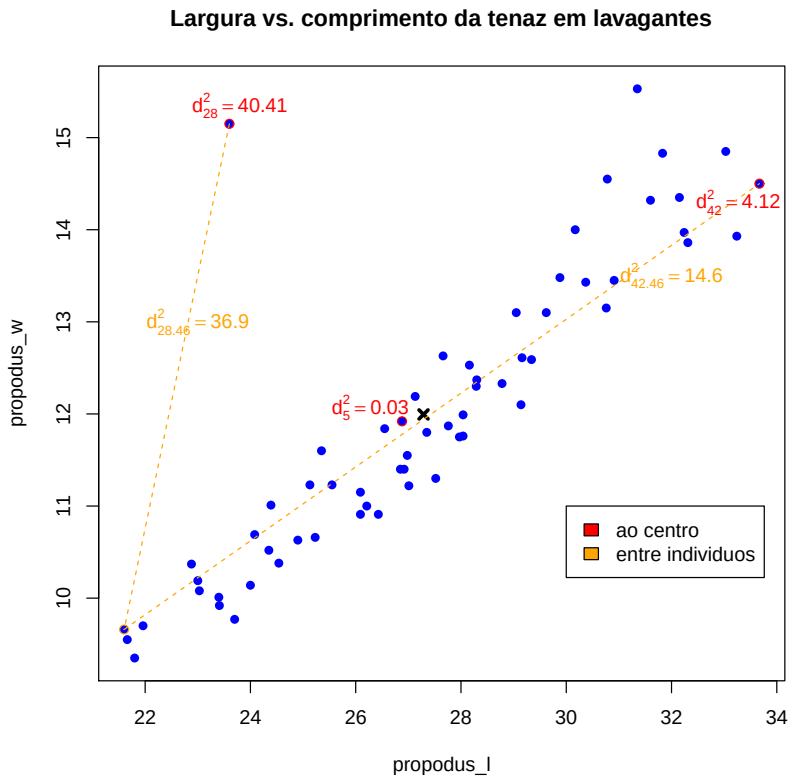
- O produto interno entre cada par de linhas da matriz **H** é a covariância entre as variáveis correspondentes.
- A norma de cada linha da matriz **H** é o desvio padrão da variável correspondente.
- O cosseno do ângulo entre cada par de linhas da matriz **H** é o coeficiente de correlação entre as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

Os marcadores de indivíduos

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

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

As distâncias de Mahalanobis levam em conta o padrão de covariâncias entre variáveis e são úteis para identificar *outliers* multivariados:



À esquerda, as distâncias no gráfico são euclidianas. Os valores numéricos são distâncias de Mahalanobis. Num *biplot, a distância euclidiana entre pontos é igual à 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 individuos e variáveis e:

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

é a melhor aproximação, de caracteristica 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 individuos 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 individuos é a distância de Mahalanobis entre esses individuos:

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

A qualidade da aproximação mede-se como na ACP.

Biplots (cont.)

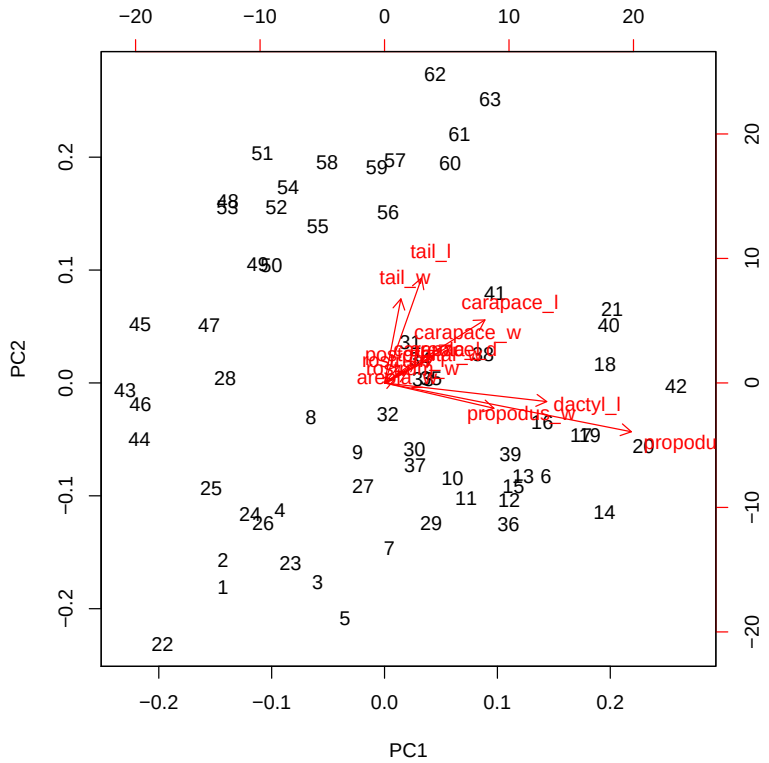
Verificam-se ainda as seguintes propriedades aproximadas (por ser a $k = 2$ dimensões):

- 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.

Os *biplots* no R

Função `biplot`, aplicada ao resultado duma ACP:

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

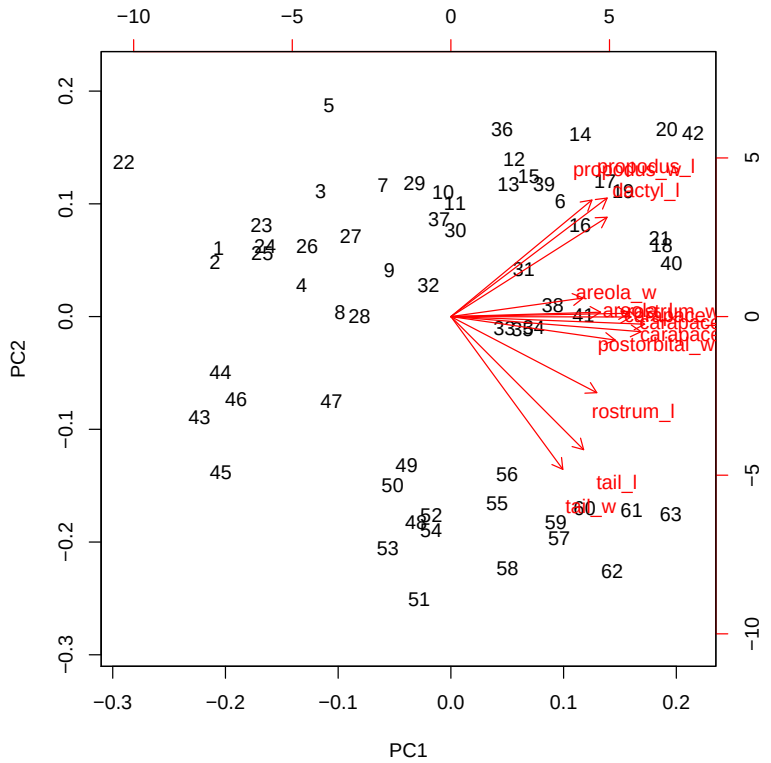


Os individuos do grupo 43-63 (fêmeas) têm tenazes mais pequenas e caudas maiores do que os machos.

Os *biplots* no R (cont.)

Eis o *biplot* dos 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?).

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.

Análise Discriminante Linear (cont.)

Ponto de partida: uma matriz de dados, $\mathbf{X}_{n\times p}$.
Os n individuos (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 $\mathbf{X}\vec{\mathbf{a}}$ das variáveis observadas, para que:

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

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

A solução vai envolver projecções ortogonais sobre o subespaço gerado pelas colunas indicatrizes da constituição de cada subgrupo.

A matriz da classificação

Matriz da classificação **G**, cuja coluna *i* é indicatriz do subgrupo *i*:

$$\mathbf{G} = \left[\begin{array}{ccccc}
1 & 0 & 0 & \cdots & 0 \\
1 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 0 & \cdots & 0 \\
\hline
0 & 1 & 0 & \cdots & 0 \\
0 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 1 & 0 & \cdots & 0 \\
\hline
0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 1 & \cdots & 0 \\
\hline
\vdots & \vdots & \vdots & \ddots & \vdots \\
\hline
0 & 0 & 0 & \cdots & 1 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 1
\end{array}\right]$$

Nota: 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.

A matriz da classificação (cont.)

- As colunas de $\mathbf{G}$ são ortogonais entre si.
- A soma das k colunas de $\mathbf{G}$ é o vector $\vec{\mathbf{1}}_n$, logo $\vec{\mathbf{1}}_n \in \mathscr{C}(\mathbf{G})$.
- O quadrado da norma da coluna j é n_j (no. indivíduos da classe j).
- Os vectores do espaço das colunas da matriz $\mathbf{G}$ têm valor igual nos elementos de cada subgrupo, isto é, $\vec{\mathbf{z}} \in \mathscr{C}(\mathbf{G})$ são da forma:

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

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

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

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

Formulação do problema

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

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

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

A projecção ortogonal cria um triângulo rectângulo. Pelo Teorema de Pitágoras, tem-se:

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

Vamos ver que a combinação linear $\mathbf{X}^c\vec{\mathbf{a}}$ desejável é a que, nesta decomposição, maximiza (em termos relativos) a primeira parcela do lado direito.

De novo a equação resultante de Pitágoras

A equação final do acetato 90 simplifica definindo as matrizes:

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

Tem-se:

$$\vec{\mathbf{a}}^t \underbrace{\mathbf{X}^{c\,t}\mathbf{X}^{c}}_{=(n-1)\cdot\mathbf{S}} \vec{\mathbf{a}} = \vec{\mathbf{a}}^t \underbrace{\mathbf{X}^{c\,t}\mathbf{P}_G\mathbf{X}^{c}}_{=(n-1)\cdot\mathbf{B}} \vec{\mathbf{a}} + \vec{\mathbf{a}}^t \underbrace{\mathbf{X}^{c\,t}(\mathbf{I}_n - \mathbf{P}_G)\mathbf{X}^{c}}_{=(n-1)\cdot\mathbf{W}} \vec{\mathbf{a}}$$

$$\Longleftrightarrow \qquad \vec{\mathbf{a}}^t\mathbf{S}\vec{\mathbf{a}} = \vec{\mathbf{a}}^t\mathbf{B}\vec{\mathbf{a}} + \vec{\mathbf{a}}^t\mathbf{W}\vec{\mathbf{a}}$$

Quer-se a maximização relativa da primeira parcela do lado direito.

A matriz de projecções ortogonais $\mathbf{P_G}$

$$\mathbf{P}_G = \mathbf{G}(\mathbf{G}^t\mathbf{G})^{-1}\mathbf{G}^t = \left[\begin{array}{c|c|c|c} \begin{matrix} \frac{1}{n_1} & \frac{1}{n_1} & \cdots & \frac{1}{n_1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n_1} & \frac{1}{n_1} & \cdots & \frac{1}{n_1} \end{matrix} & \mathbf{0}_{n_1 \text{x} n_2} & \cdots & \mathbf{0}_{n_1 \text{x} n_k} \\ \hline \mathbf{0}_{n_2 \text{x} n_1} & \begin{matrix} \frac{1}{n_2} & \frac{1}{n_2} & \cdots & \frac{1}{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n_2} & \frac{1}{n_2} & \cdots & \frac{1}{n_2} \end{matrix} & \cdots & \mathbf{0}_{n_2 \text{x} n_k} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline \mathbf{0}_{n_k \text{x} n_1} & \mathbf{0}_{n_k \text{x} n_2} & \cdots & \begin{matrix} \frac{1}{n_k} & \frac{1}{n_k} & \cdots & \frac{1}{n_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n_k} & \frac{1}{n_k} & \cdots & \frac{1}{n_k} \end{matrix} \end{array}\right]$$

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

Consideremos vectores com dupla indexação i,j, sendo i subgrupo e j repetição.

Para qualquer vector $\vec{\mathbf{y}} \in \mathbb{R}^n$:

$$\mathbf{P}_G\vec{\mathbf{y}} = \begin{bmatrix} \overline{y}_{1.} \\ \vdots \\ \overline{y}_{1.} \\ \hline \overline{y}_{2.} \\ \vdots \\ \overline{y}_{2.} \\ \hline \vdots \\ \hline \overline{y}_{k.} \\ \vdots \\ \overline{y}_{k.} \end{bmatrix}$$

Se $\vec{\mathbf{y}}^c$ é um vector centrado ($\vec{\mathbf{y}} \perp \vec{\mathbf{1}}_n$):

$$\mathbf{P}_G\vec{\mathbf{y}}^c = \begin{bmatrix} \overline{y}_{1.} - \overline{y}_{..} \\ \vdots \\ \overline{y}_{1.} - \overline{y}_{..} \\ \hline \overline{y}_{2.} - \overline{y}_{..} \\ \vdots \\ \overline{y}_{2.} - \overline{y}_{..} \\ \hline \vdots \\ \hline \overline{y}_{k.} - \overline{y}_{..} \\ \vdots \\ \overline{y}_{k.} - \overline{y}_{..} \end{bmatrix}$$

Para um vector centrado $\vec{\mathbf{y}}^c$, $\|\mathbf{P}_G\vec{\mathbf{y}}^c\|^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{y}_{i.} - \overline{y}_{..})^2$ mede dispersão das médias de cada classe em torno da média geral de $\vec{\mathbf{y}}$. É a variabilidade inter-classes, e convém ser grande: reflecte heterogeneidade entre classes.

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

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

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

$\|(\mathbf{I}_n - \mathbf{P}_G)\vec{\mathbf{y}}^c\|^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i.})^2$ mede a dispersão das observações individuais em torno da respectiva média de nível. É a variabilidade intra-classes, e convém ser pequena: mede homogeneidade das classes.

Formulação de Fisher

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

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

Essa será a primeira função discriminante.

Solução: Se $\mathbf{W}$ fôr definida positiva, usar o resultado associado ao problema de valores/vectores próprios generalizados (acetato 62):

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.

Formulação de Fisher (cont.)

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

Pelo Teorema do problema de valores próprios generalizados (acetato 62), sucessivas soluções são as combinações lineares $\mathbf{X}\vec{\mathbf{a}}_j$ com $\vec{\mathbf{a}}_j$ os restantes 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 λ_j associado.

Existência de $\mathbf{W}^{-1}$

$\mathbf{W}_{p\times p}$ é invertível se fôr de característica plena p.

Resultado geral em matrizes: $car(\mathbf{AB}) \leq \min\{car(\mathbf{A}), car(\mathbf{B})\}$.

Como $\mathbf{W} = \frac{1}{n-1}\mathbf{X}^{c\,t}(\mathbf{I}_n - \mathbf{P}_G)\mathbf{X}^c$, tem-se
$car(\mathbf{W}) \leq \min\{car(\mathbf{X}^c), car(\mathbf{I}_n - \mathbf{P}_G)\}$ (já que $car(\mathbf{X}^c) = car(\mathbf{X}^{c\,t})$).

Admitindo que não há dependências lineares nas colunas de $\mathbf{X}^c$ (multicolinearidade), e que $n > p$, tem-se $car(\mathbf{X}^c) = p$. A característica duma matriz de projecção ortogonal (como $\mathbf{P}_G$) é a dimensão do subespaço sobre o qual projecta, logo $car(\mathbf{P}_G) = k$ e $car(\mathbf{I}_n - \mathbf{P}_G) = n - k$.

Assim, tem-se $car(\mathbf{W}) \leq \min\{p, n-k\}$.

Se $k > n - p$, $\mathbf{W}$ não pode ser invertível.
Em geral, para $k \leq n - p$ haverá invertibilidade.

Observações

- As matrizes usadas na ADL verificam a relação $\mathbf{S} = \mathbf{B} + \mathbf{W}$.

  De facto,

$$\begin{aligned} \mathbf{I}_n = \mathbf{P}_G + (\mathbf{I}_n - \mathbf{P}_G) \quad &\Rightarrow \quad \mathbf{X}^{c\,t}\mathbf{I}_n\mathbf{X}^c = \mathbf{X}^{c\,t}\mathbf{P}_G\mathbf{X}^c + \mathbf{X}^{c\,t}(\mathbf{I}_n - \mathbf{P}_G)\mathbf{X}^c \\ &\Leftrightarrow \quad \mathbf{S} = \mathbf{B} + \mathbf{W} \end{aligned}$$

- Sucessivos eixos discriminantes são não correlacionados entre si.

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

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

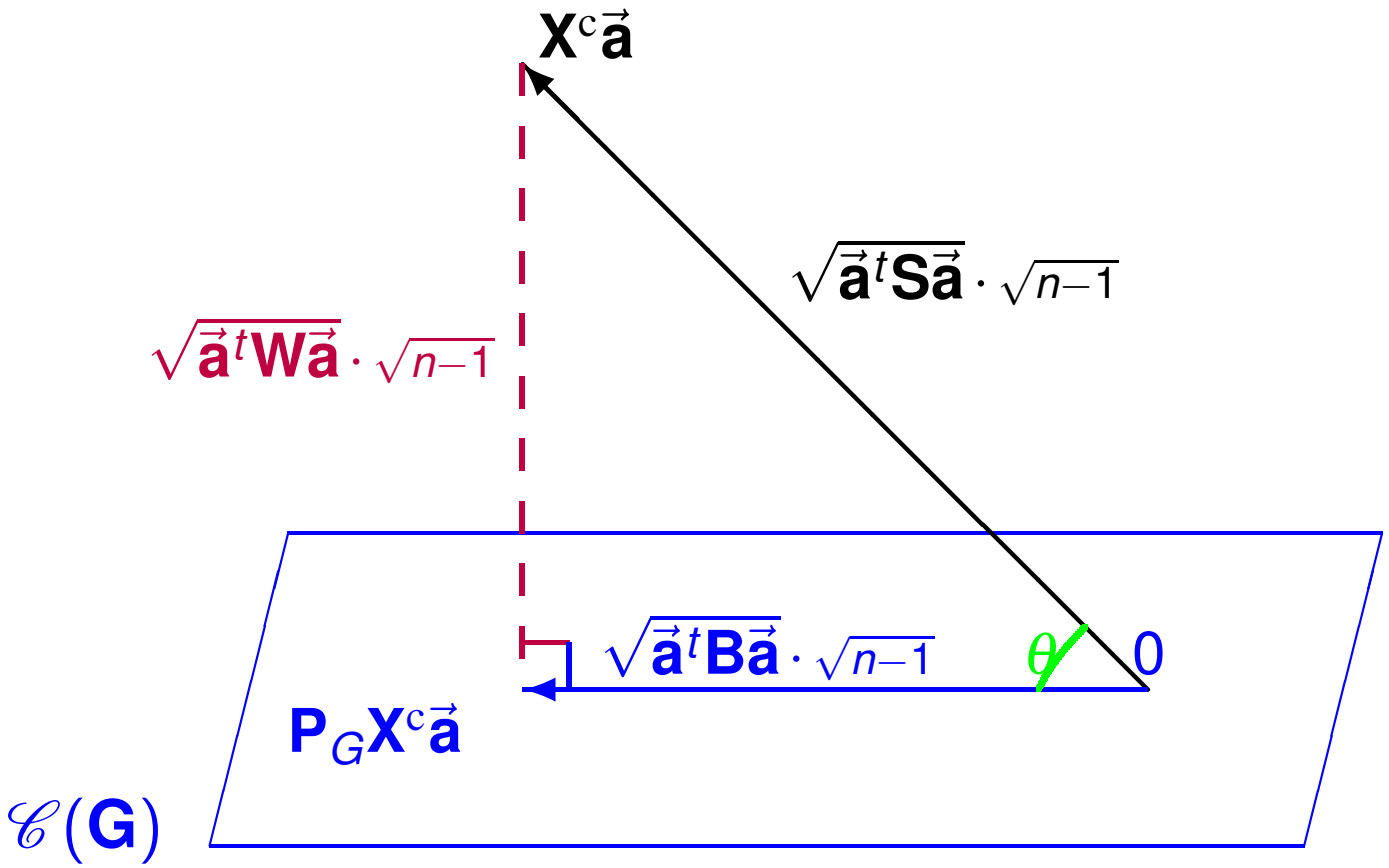
Observações (cont.)

- 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 $\mathbf{W}$-ortogonais).
- $\mathbf{W}^{-1}\mathbf{B}$ não pode ter mais de $k-1$ valores próprios não-nulos:

$$car(\mathbf{W}^{-1}\mathbf{B}) \leq car(\mathbf{B}) = car(\mathbf{X}^{c\,t}\mathbf{P}_G\mathbf{X}^c) \leq car(\mathbf{P}_G) = k\ .$$

  Mas como o vector projectado é centrado (ortogonal a $\vec{\mathbf{1}}_n \in \mathscr{C}(\mathbf{G})$), perde-se mais uma dimensão: $k-1$ é o número máximo de eixos discriminantes.
- as soluções duma Análise Discriminante são invariantes a transformações lineares das escalas das variáveis.

ADL - Resumo



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

Formulações alternativas

Maximizar a cotangente do ângulo θ quer dizer minimizar θ.
Formulações alternativas que minimizam o ângulo:

1. Minimizar o quadrado do seno de θ.
   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 θ
   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 θ) $\Rightarrow$ 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 λ 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 $\mathbf{S}^{-1}\mathbf{B}$;

ADL e ANOVA

Considere-se:

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

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

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

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 individuos num eixo

Podemos classificar novos indivíduos, de "filiação" desconhecida.

Seja $\vec{\mathbf{x}}$ vector de observações de novo individuo nas p variáveis.
O respectivo valor (*score*) no eixo discriminante 1 é $y^* = \vec{\mathbf{x}}^t \vec{\mathbf{a}}_1$.
Comparando este valor com as k médias de classe nesse eixo, $\overline{y}^{(1)}, \overline{y}^{(2)}, \ldots, \overline{y}^{(k)}$, podemos classificar na classe cujo centro de gravidade:

- lhe está mais próxima, na habitual distância euclidiana:

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

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

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

  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*: $\vec{\mathbf{y}}^* = \vec{\mathbf{x}}^t \mathbf{A}_q$, com $\mathbf{A}_q$ matriz $p \times q$ cujas colunas são os vectores $\vec{\mathbf{a}}_1, ..., \vec{\mathbf{a}}_q$.

Pode-se classificar o indívíduo na classe cujo centro de gravidade $\vec{\bar{\mathbf{y}}}_{(i)}$:

- esteja mais próximo de $\vec{\mathbf{y}}^*$, na habitual distância euclidiana:

  $$\textit{associar à classe } i \textit{ se } \|\vec{\mathbf{y}}^* - \vec{\bar{\mathbf{y}}}_{(i)}\| < \|\vec{\mathbf{y}}^* - \vec{\bar{\mathbf{y}}}_{(j)}\|, \forall j \neq i$$

- esteja mais próximo de $\vec{\mathbf{y}}^*$ na distância de Mahalanobis usual:

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

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

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

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

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

ADL no R

A função `lda`, 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

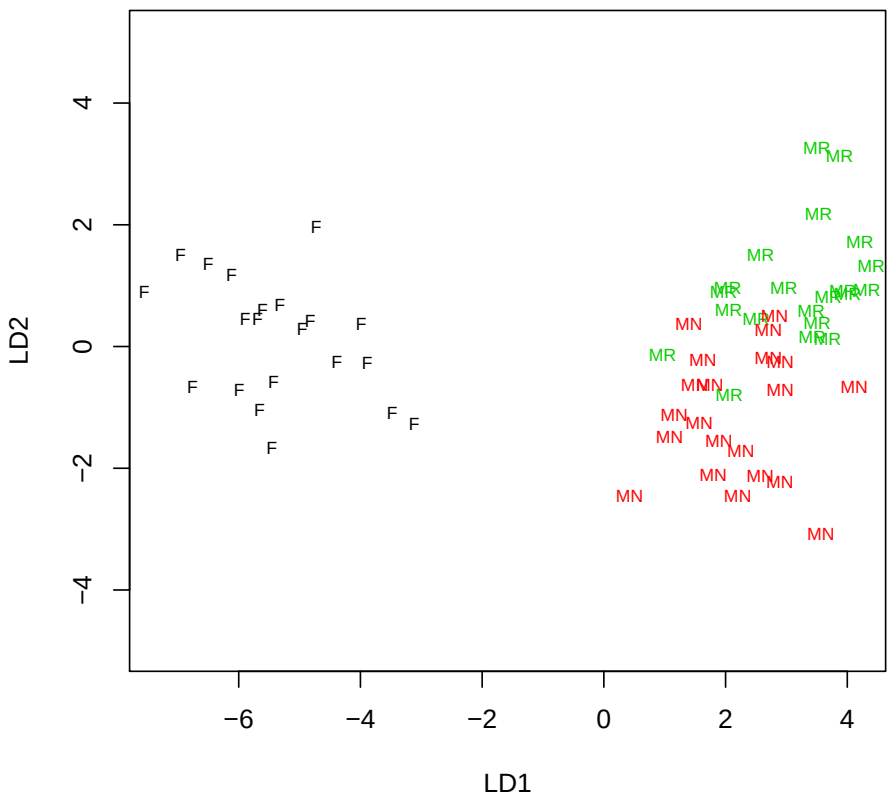
Eis o comando, e principais resultados para a ADL dos lavagantes:

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

[...]
Coefficients of linear discriminants:                    <- vectores de loadings
                        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

Proportion of trace:                    <- proporção dos valores próprios não nulos de W^(-1)B
   LD1    LD2
0.9501 0.0499
```

Exemplos dos lavagantes (cont.)

Existe um método `plot` para objectos da classe `lda`, como os que são produzidos pelos comando `lda`:

```
> 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`:

```
> 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.)

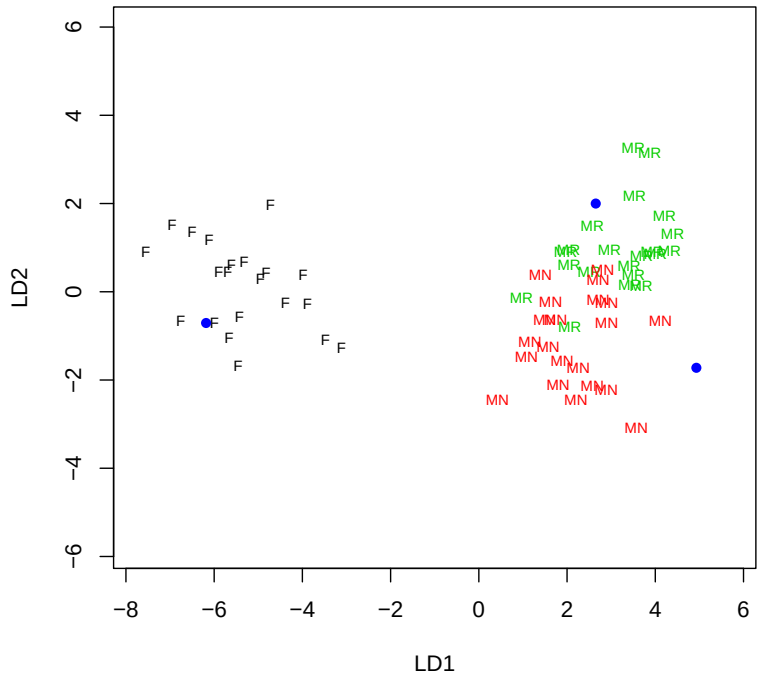
Definamos três novos indivíduos com todas as observações iguais aos máximos de cada variável em cada subgrupo, e coloque-mo-los nos novos eixos discriminantes:

```
> 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
lxm2  4.931230 -1.7232999
lxm3 -6.183037 -0.7078646
```

Exemplos dos lavagantes (cont.)

Existe um método `plot` para objectos da classe `lda`, como os que são produzidos pelos comando `lda`:

```
> 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`

Na função `lda`:

- a matriz $\mathbf{W}$ é definida como $\mathbf{W} = \frac{1}{n-k}\mathbf{X}^{c\,t}(\mathbf{I}_n - \mathbf{P}_G)\mathbf{X}^c$;
- a matriz $\mathbf{B}$ é definida como $\mathbf{B} = \frac{1}{k-1}\mathbf{X}^{c\,t}\mathbf{P}_G\mathbf{X}^c$;
- logo, cada valor próprio de $\mathbf{W}^{-1}\mathbf{B}$ dado pela função `lda` é o valor da estatística do teste F na ANOVA de cada eixo discriminante sobre o factor. Mas deixa de ser válida a decomposição $\mathbf{S} = \mathbf{W} + \mathbf{B}$.
- a função `lda` devolve as raízes quadradas dos valores próprios de $\mathbf{W}^{-1}\mathbf{B}$ na componente `svd`:

```
> lav.lda$svd
[1] 21.345129  4.890076
> lav.lda$svd^2  <- Valores proprios (e valores da estatistica F)
[1] 455.61455  23.91285
> lav.lda$svd^2*2/60  <- Valores próprios com W e B originais
[1] 15.1871516  0.7970949
```

Observações sobre a função `lda` (cont.)

- as proporções representadas por cada valor próprio (dadas na listagem) não são afectadas por esta diferente definição.

```
> 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 91).

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 baseada em distâncias de Mahalanobis. Essas classificações são guardadas no objecto `class` da função `predict`:

```
> 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 MR MN MR MN MN
[26] MN MR MN MN 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 classificação podem ser criadas com a função `table`:

```
> 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
```