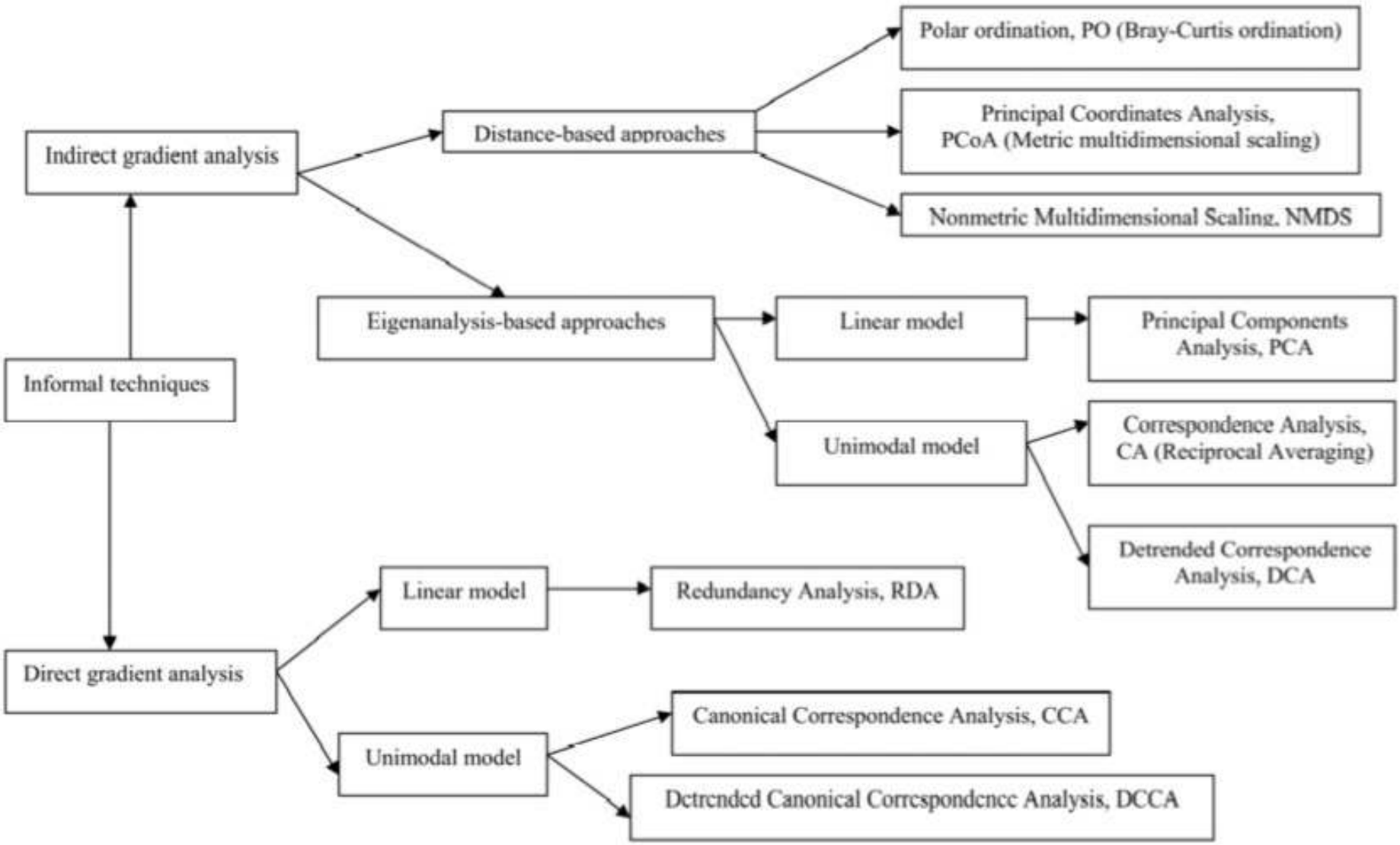
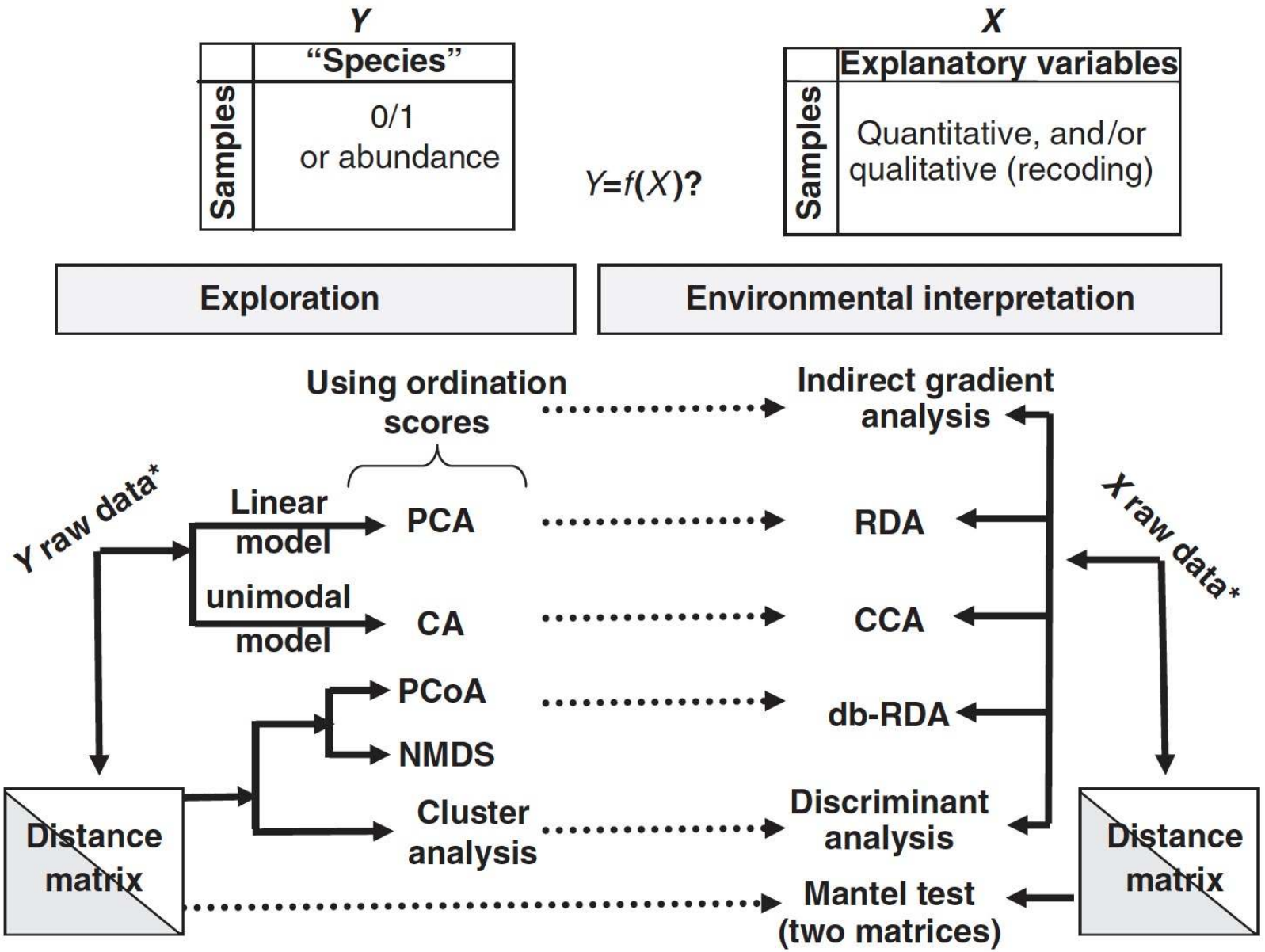
"Você já pensou em desistir da disciplina ?"

Aula 4: Ordenações irrestritas

Cap. 9 Legendre & Legendre

Árvore de decisão de análises de ordenação





Ramette 2007

Quais as análises que veremos hoje?

- PCA
- PCoA
- CA
  - DCA
  - Não falaremos de MCA
- nMDS

Table 9.1 Domains of application of the ordination methods presented in this chapter.

Method	Distance preserved	Variables
Principal component analysis (PCA)	Euclidean distance	Quantitative data, linear relationships (beware of double-zeros)
Correspondence analysis (CA)	χ^2 distance	Non-negative, dimensionally homogeneous quantitative or binary data; species frequencies or presence/absence data
Principal coordinate analysis (PCoA), metric (multidimensional) scaling, classical scaling	Any distance measure	Quantitative, semiquantitative, qualitative, or mixed
Nonmetric multidimensional scaling (nMDS)	Any distance measure	Quantitative, semiquantitative, qualitative, or mixed

Análise de Componentes Principais (PCA)

Densidade de macroinvertebrados

Dossel

Riqueza de plantas aquáticas

pH

Temperatura

Condutividade

Composição de peixes

Vegetação interna

Volume

Velocidade da correnteza

Densidade de algas

Turbidez

Riqueza de macroinvertebrados

Slide Thiago Gonçalves Souza

Matrizes de dados ecológicos

Espécies

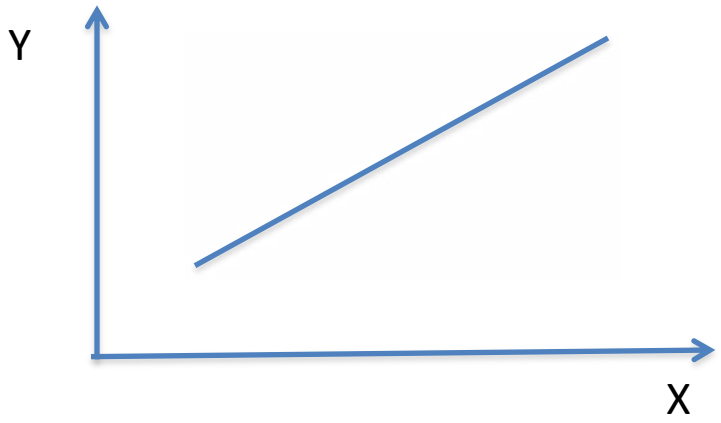
	Sp1	Sp2	Sp3	Sp4	Spn
Área 1	2	12	14	0	2
Área 2	10	0	11	0	15
Área 3	18	19	10	1	19
Área 4	6	21	9	0	8
Área 5	0	0	1	18	0
Área 6	0	0	2	9	0
Área 7	0	0	0	19	0
Área 8	1	0	1	21	0

Variáveis ambientais

Temp	pH	O_2	CO_2
2	12	14	0
10	0	11	0
18	19	10	1
6	21	9	0
0	0	1	18
0	0	2	9
0	0	0	19
1	0	1	21

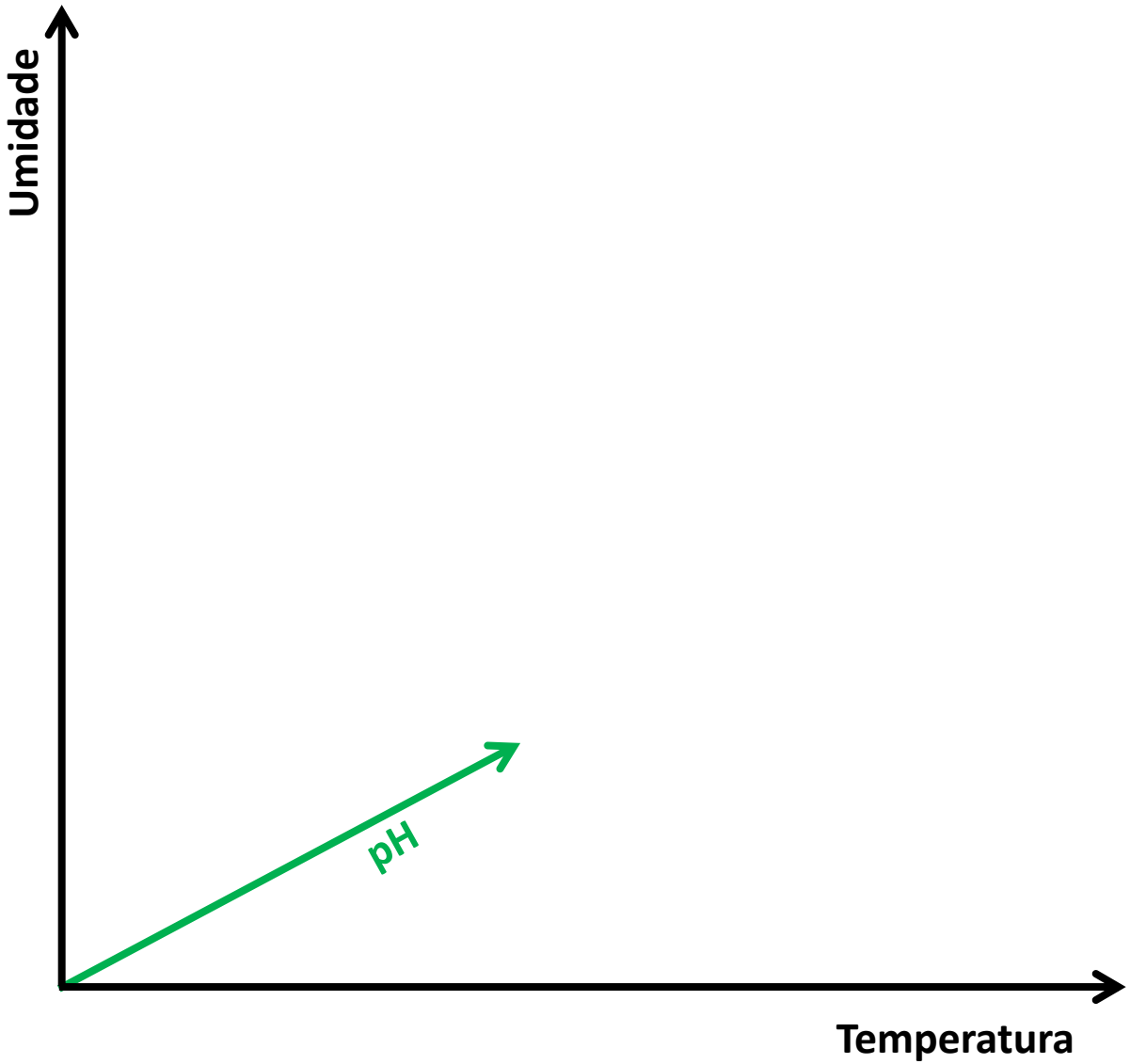
Slide Thiago Gonçalves Souza

PCA



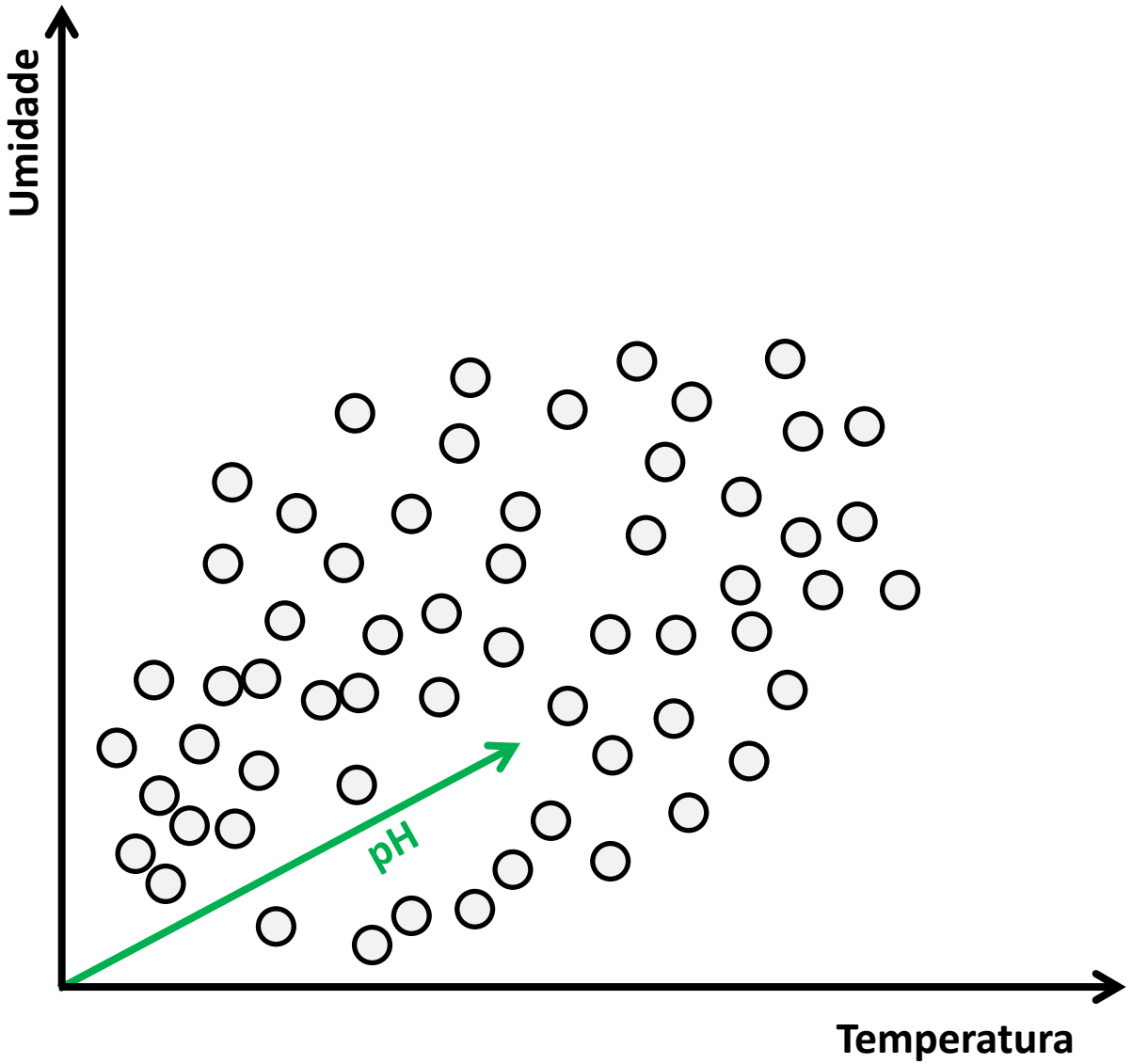
$Y=a+bX$

Dados multivariados



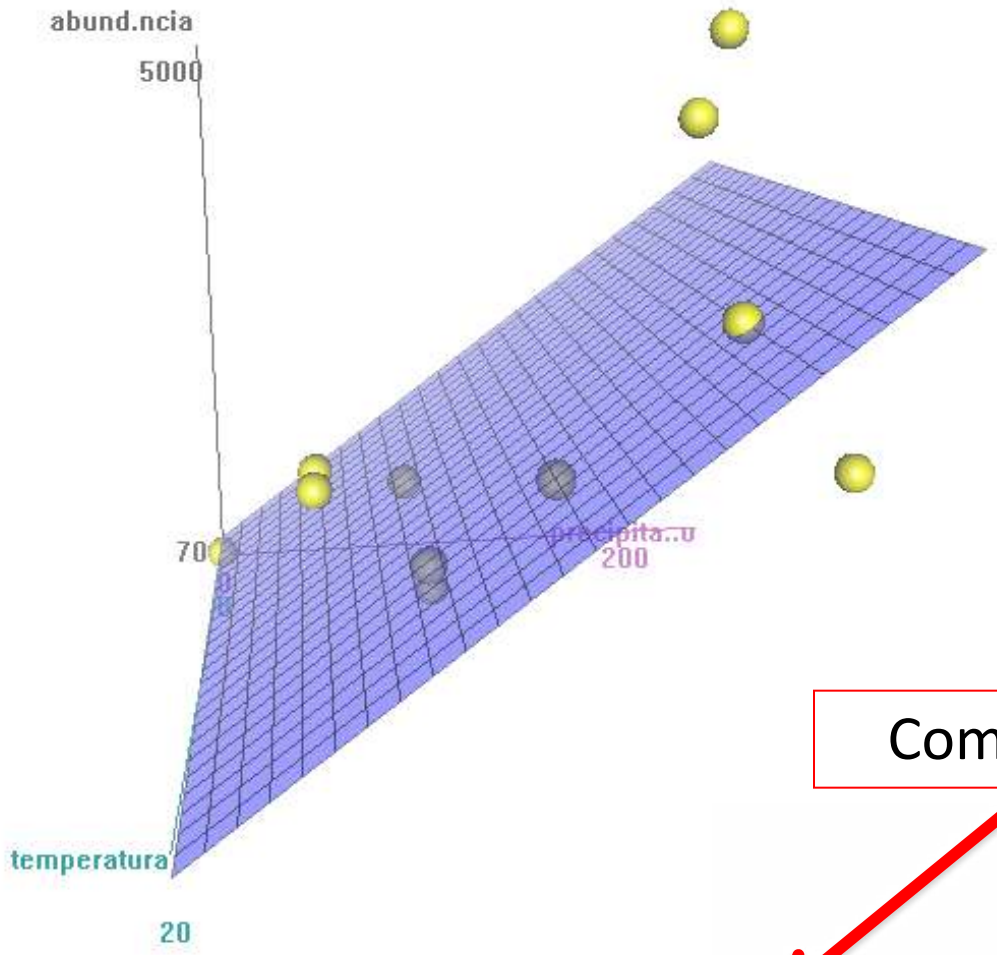
Slide Thiago Gonçalves Souza

Dados multivariados



Slide Thiago Gonçalves Souza

PCA



Combinação Linear

$$Y=a+bX_1+bX_2+\ldots+bX_n$$

Principal Component Analysis
PCA

1. Reduz a dimensionalidade de dados multivariados

Slide Thiago Gonçalves Souza

Principal Component Analysis
PCA

1. Reduz a dimensionalidade de dados multivariados

2. A **PCA** se aplica à tabela de dados em que as linhas são os **indivíduos** e as colunas **variáveis quantitativas**. Portanto, **preserva a distância Euclidiana**

Slide Thiago Gonçalves Souza

Principal Component Analysis
PCA

1. Reduz a dimensionalidade de dados multivariados

2. A **PCA** se aplica à tabela de dados em que as linhas são os **indivíduos** e as colunas **variáveis quantitativas**. Portanto, **preserva a distância Euclidiana**

3. OU se aplica à tabela de dados em que as linhas são as **espécies** e as colunas **locais de coleta, *desde que previamente transformadas (veja aula 2)*.**

Principal Component Analysis
PCA

1. Reduz a dimensionalidade de dados multivariados

2. A **PCA** se aplica a tabela de dados em que as linhas são os **indivíduos** e as colunas **variáveis quantitativas**. Portanto, **preserva a distância Euclidiana**

3. OU se aplica a tabela de dados em que as linhas são as **espécies** e as colunas **locais de coleta, *desde que previamente transformadas (veja aula 2)*.**

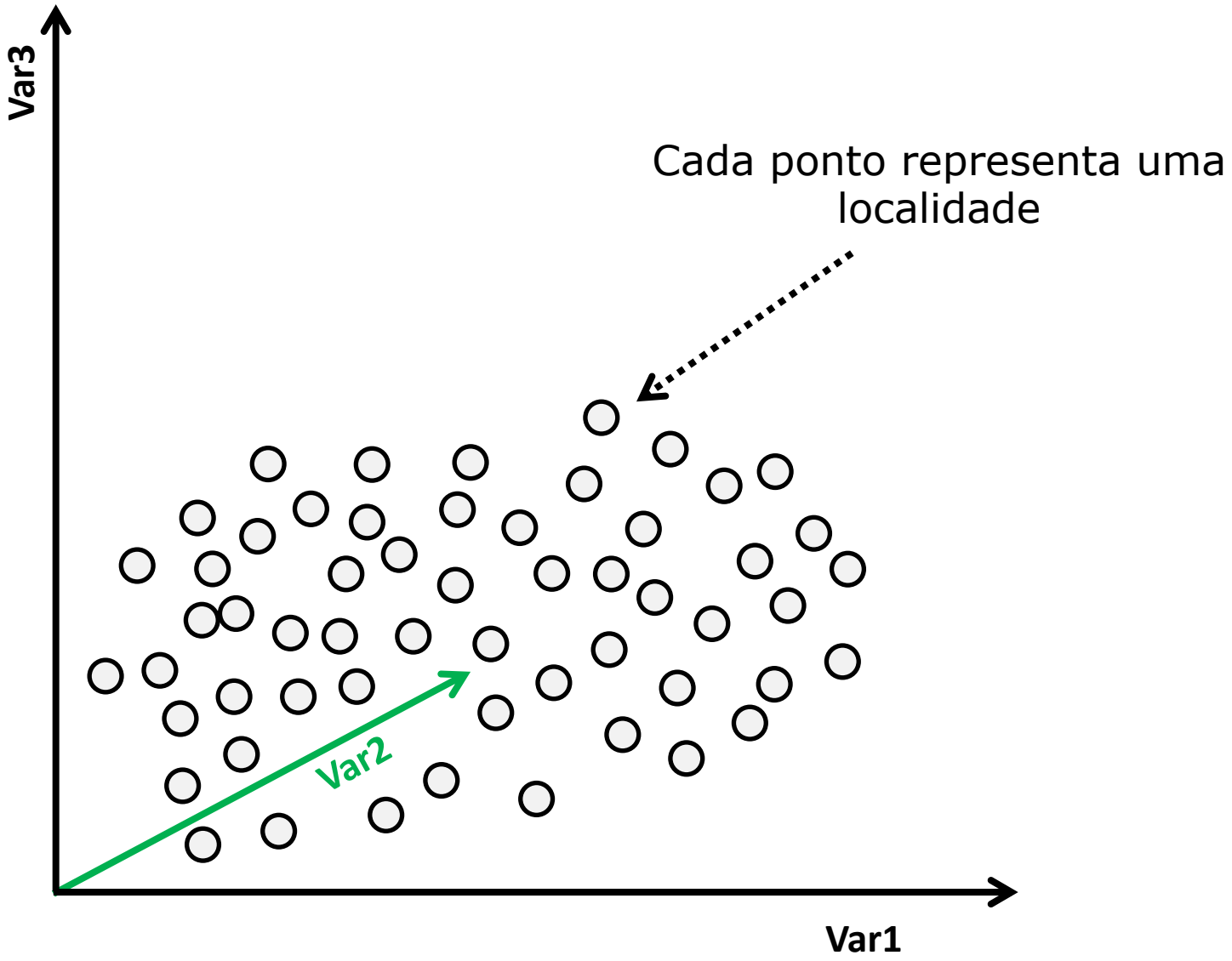
4. Dimensionalidade efetiva = requer um menor número de dimensões que representem a maior parte da variação nos dados

PCA será feita pela autoanálise da seguinte matriz:

$$\mathbf{S} = (n-1)^{-1}\mathbf{Y}_c{}'\mathbf{Y}_c, \text{ where } \mathbf{Y}_c \text{ is matrix } \mathbf{Y} \text{ column-centred.}$$

Vejamos agora o passo-a-passo

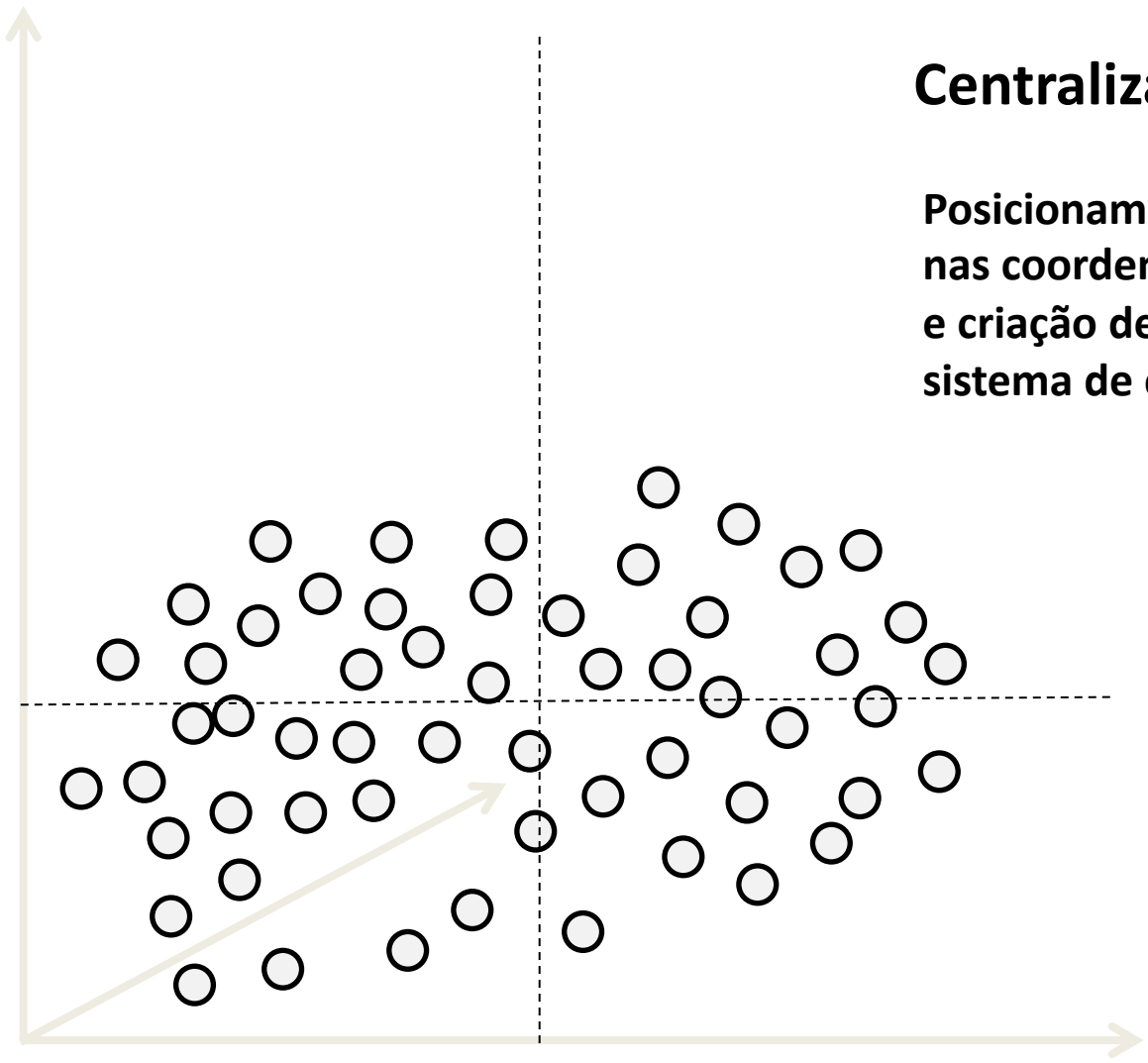
PCA



Slide Thiago Gonçalves Souza

Centralização

Posicionamento dos eixos nas coordenadas 0,0 e criação de um novo sistema de coordenadas



$$\mathbf{Y} = \begin{bmatrix} 2 & 1 \\ 3 & 4 \\ 5 & 0 \\ 7 & 6 \\ 9 & 2 \end{bmatrix} \qquad \mathbf{Y}_c = [y - \bar{y}] \qquad \mathbf{Y}_c \begin{bmatrix} -3.2 & -1.6 \\ -2.2 & 1.4 \\ -0.2 & -2.6 \\ 1.8 & 3.4 \\ 3.8 & -0.6 \end{bmatrix}$$

$$\bar{y}_1 \quad \bar{y}_2$$

O que significa isso na prática?

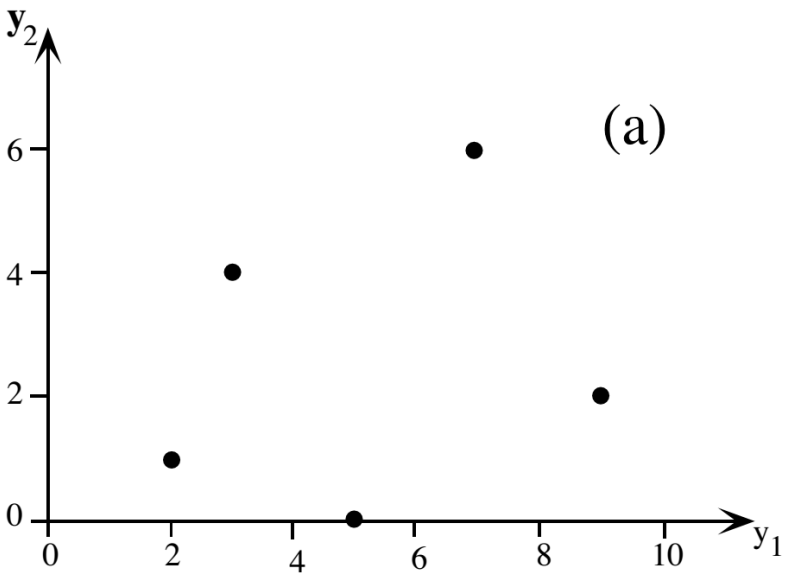
$$\mathbf{Y} = \begin{bmatrix} 2 & 1 \\ 3 & 4 \\ 5 & 0 \\ 7 & 6 \\ 9 & 2 \end{bmatrix}$$

Brutos

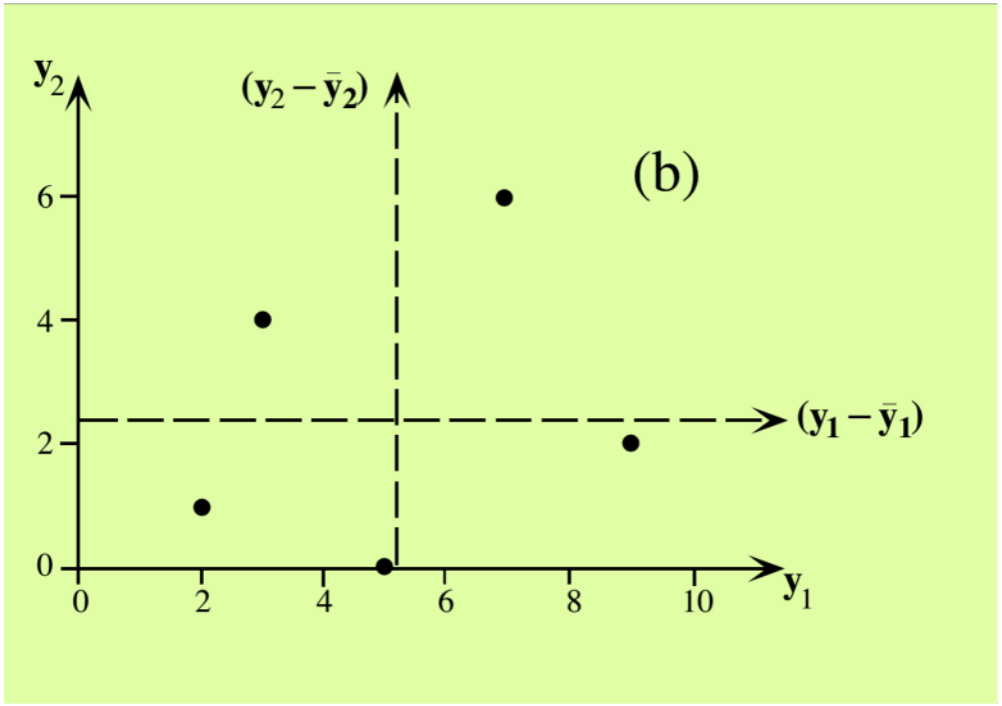
$$\mathbf{Y}_c = \begin{bmatrix} -3.2 & -1.6 \\ -2.2 & 1.4 \\ -0.2 & -2.6 \\ 1.8 & 3.4 \\ 3.8 & -0.6 \end{bmatrix}$$

Centralizados

$\overline{Y} = 5.2 \quad 2.6$



Brutos



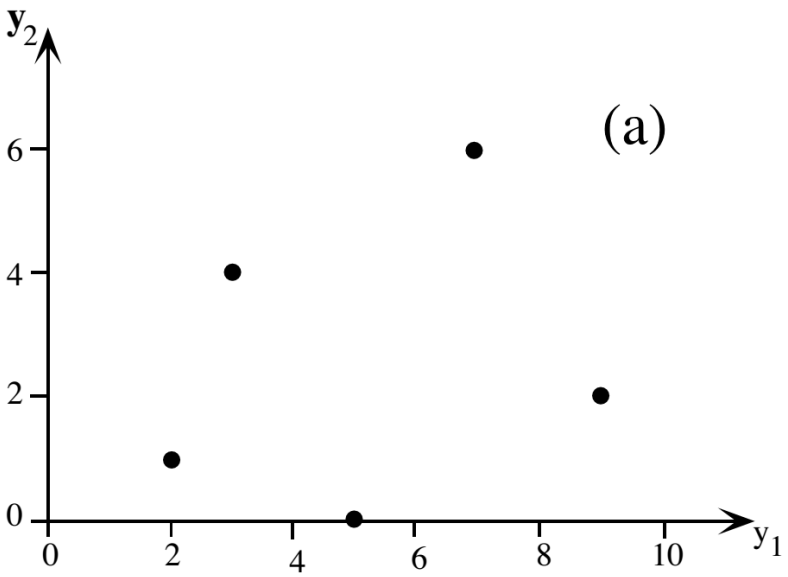
Centralizados

$$\mathbf{Y} = \begin{bmatrix} 2 & 1 \\ 3 & 4 \\ 5 & 0 \\ 7 & 6 \\ 9 & 2 \end{bmatrix} - \quad \mathbf{Y}_c = \begin{bmatrix} -3.2 & -1.6 \\ -2.2 & 1.4 \\ -0.2 & -2.6 \\ 1.8 & 3.4 \\ 3.8 & -0.6 \end{bmatrix}$$

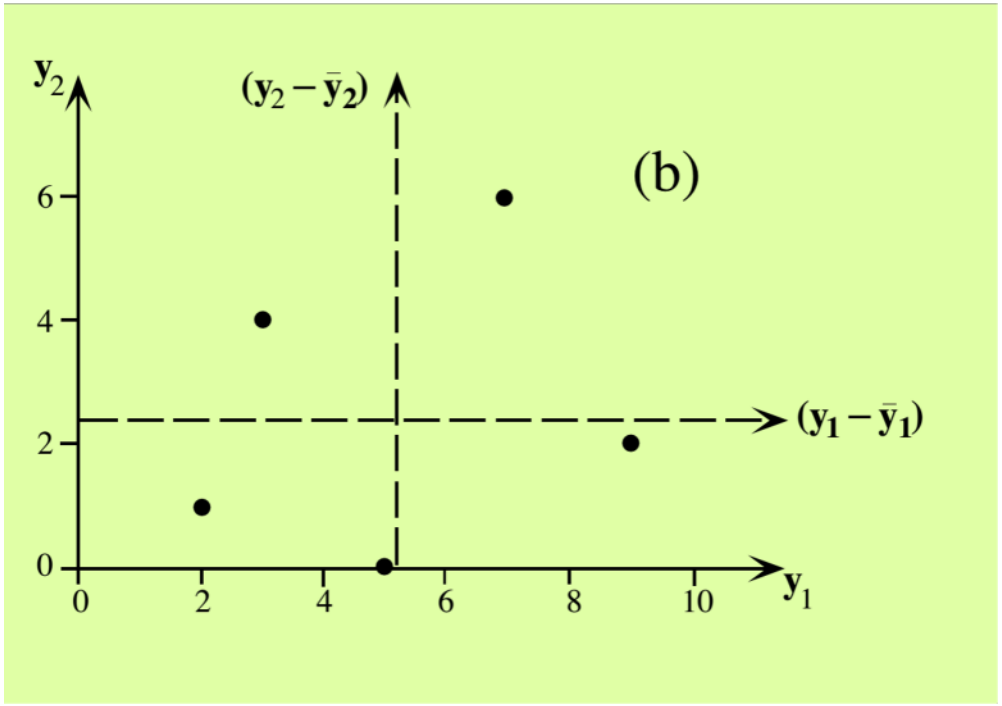
Brutos

Centralizados

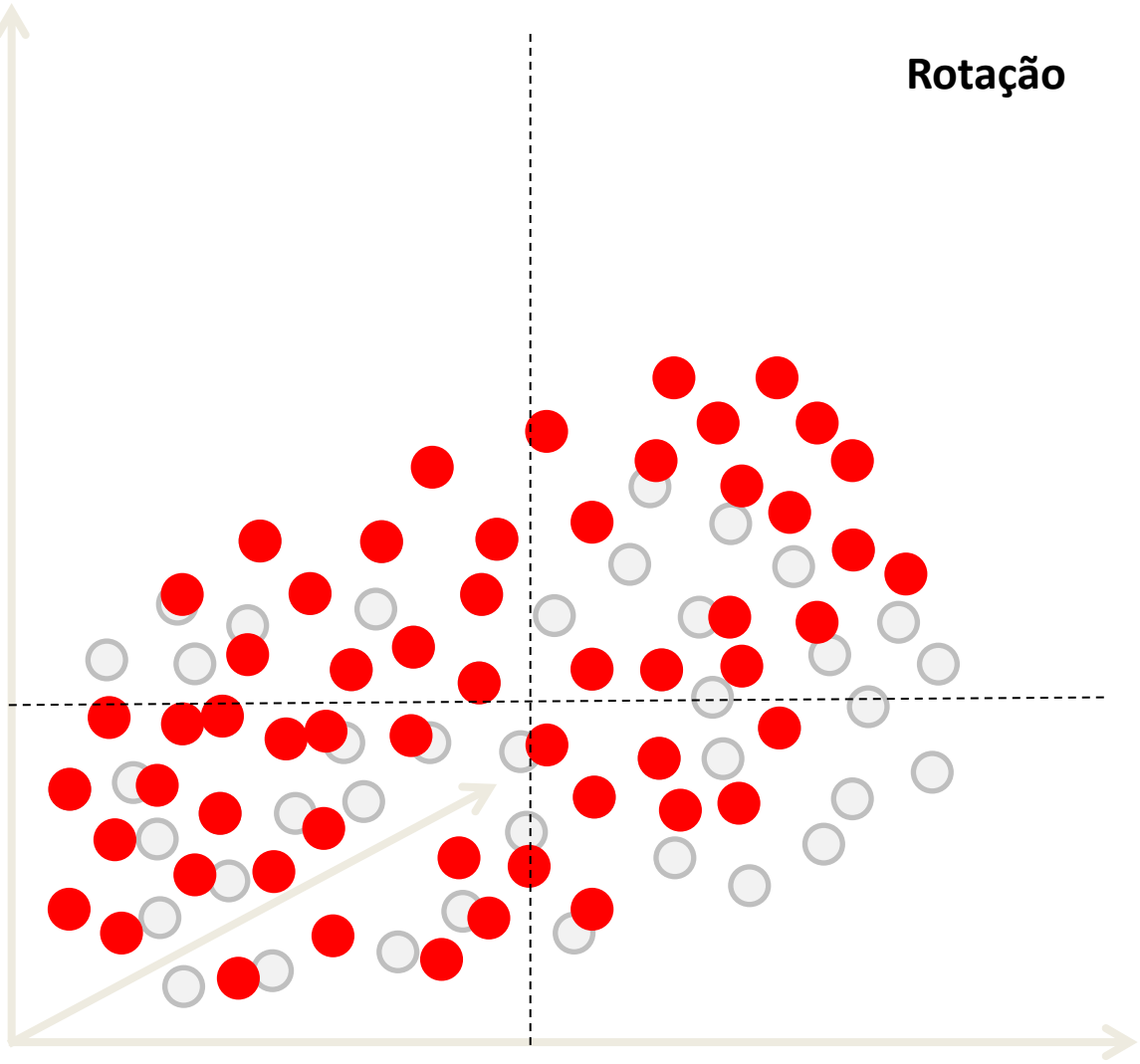
$\overline{Y} = 5.2 \quad 2.6$



Brutos

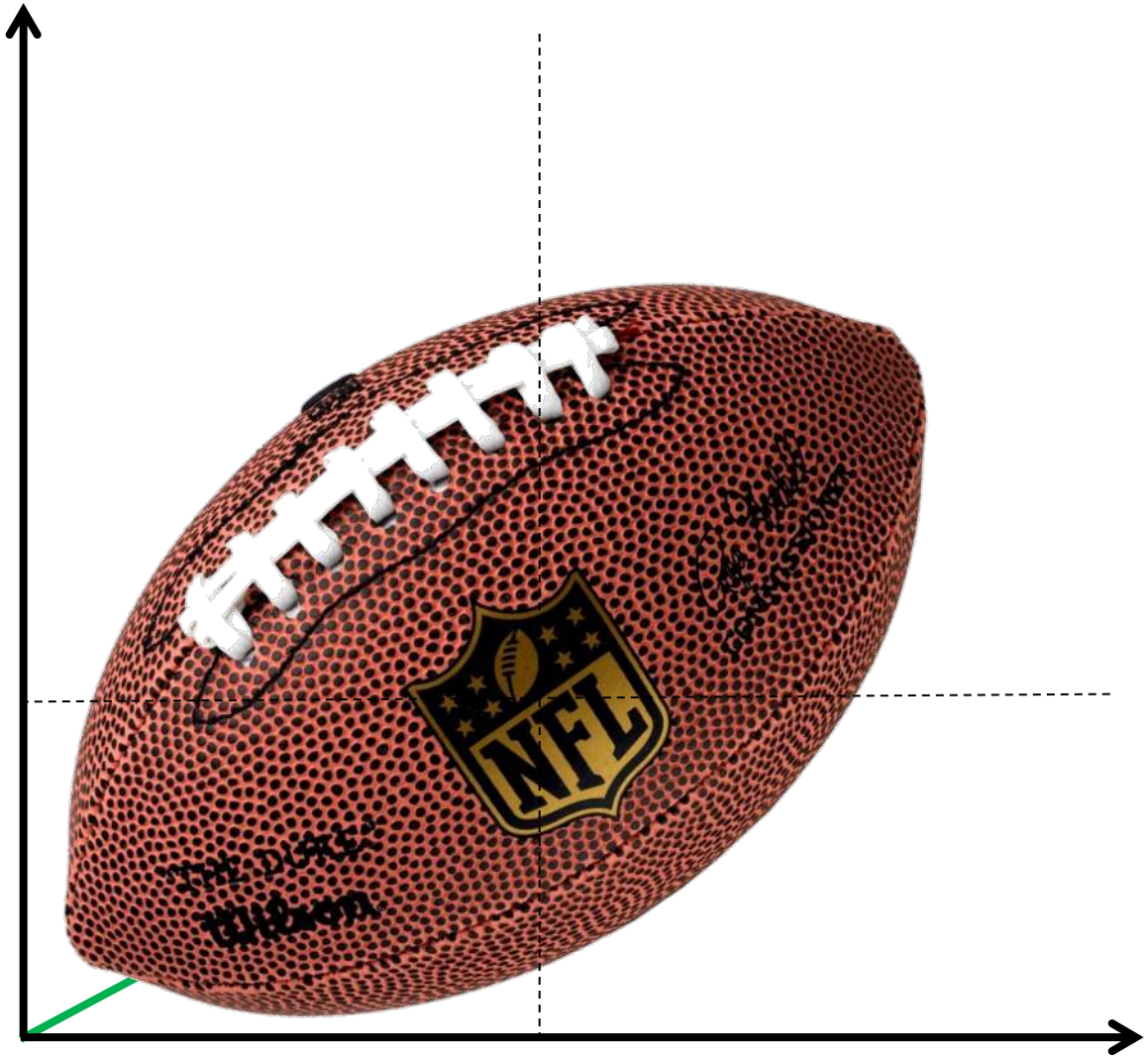


Centralizados



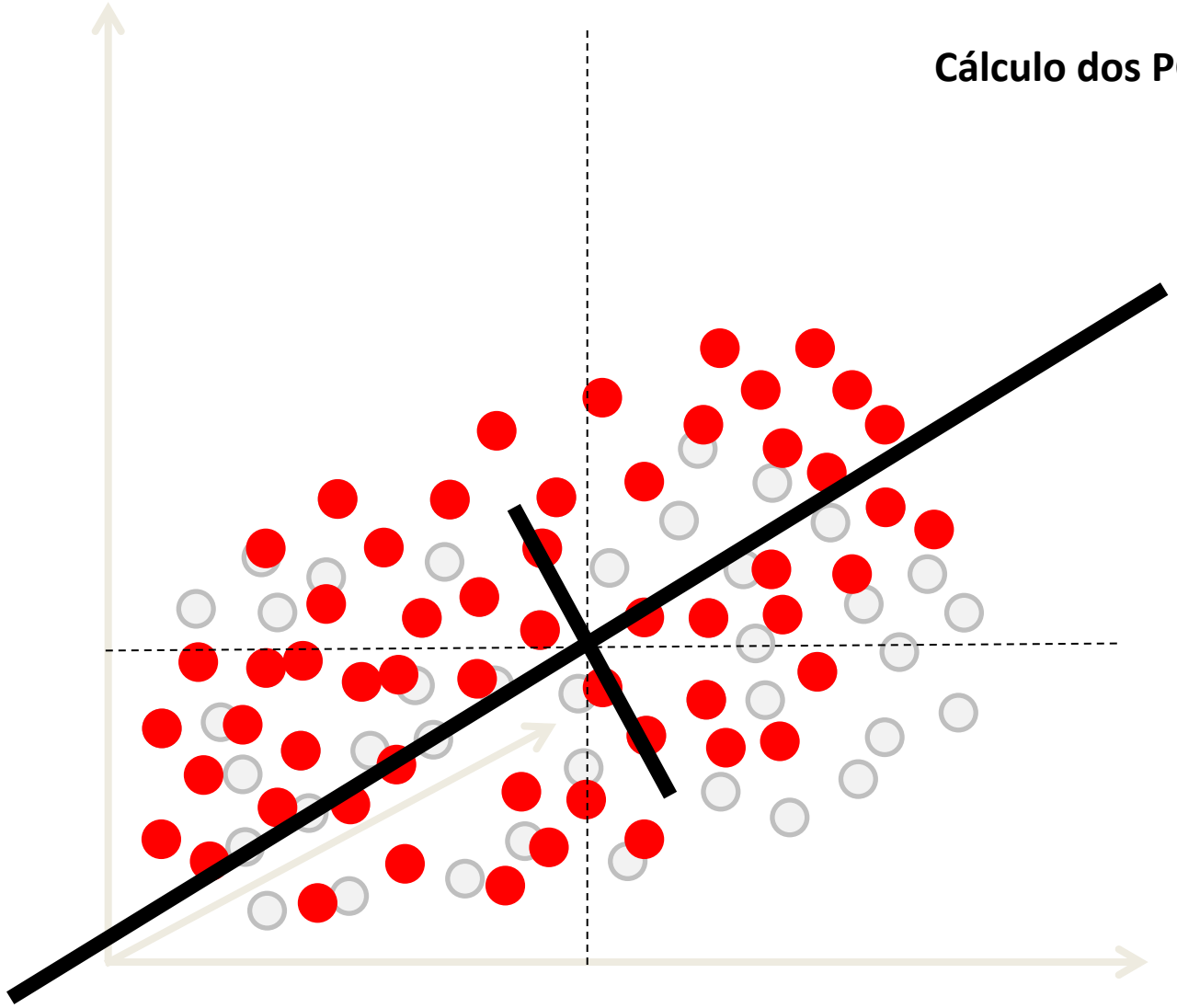
Slide Thiago Gonçalves Souza

Rotaciona os dados para encontrar dimensões com maior variação



Slide Thiago Gonçalves Souza

Cálculo dos PCs



Slide Thiago Gonçalves Souza

Componentes Principais

- Sinônimo de autovetores normalizados, numa PCA
- Eixos da PCA, correspondendo a um novo sistema de coordenadas, criados representar o eixo de maior variância dos dados
- **Número de eixos = número de variáveis**
- Calculados segundo a equação:

autovetores

$$(\mathbf{S} - \lambda_k \mathbf{I})\,\mathbf{u}_k = \mathbf{0}$$

(eq. 4.23) whose characteristic equation

autovalores

$$\left|\mathbf{S} - \lambda_k \mathbf{I}\right| = 0$$

Usada para cálculo dos autovalores

Calculando os autovetores e autovalores de uma matriz

1. Matriz de dispersão $$\mathbf{S} = \frac{1}{n-1}\mathbf{Y}_c{}'\,\mathbf{Y}_c = \begin{bmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{bmatrix}$$

2. Input na Equação característica $$\left|\mathbf{S} - \lambda_k \mathbf{I}\right| = \left| \begin{bmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{bmatrix} - \begin{bmatrix} \lambda_k & 0 \\ 0 & \lambda_k \end{bmatrix} \right| = 0$$

3. Cálculo dos autovetores e autovalores $$(\mathbf{S} - \lambda_k \mathbf{I})\,\mathbf{u}_k = \mathbf{0}$$

4. Depois de normalizados (padronizados para = 1) tornam-se os componentes principais (eixos) $$\mathbf{U} = \begin{bmatrix} 0.8944 & -0.4472 \\ 0.4472 & 0.8944 \end{bmatrix}$$

Calculando os autovetores e autovalores de uma matriz

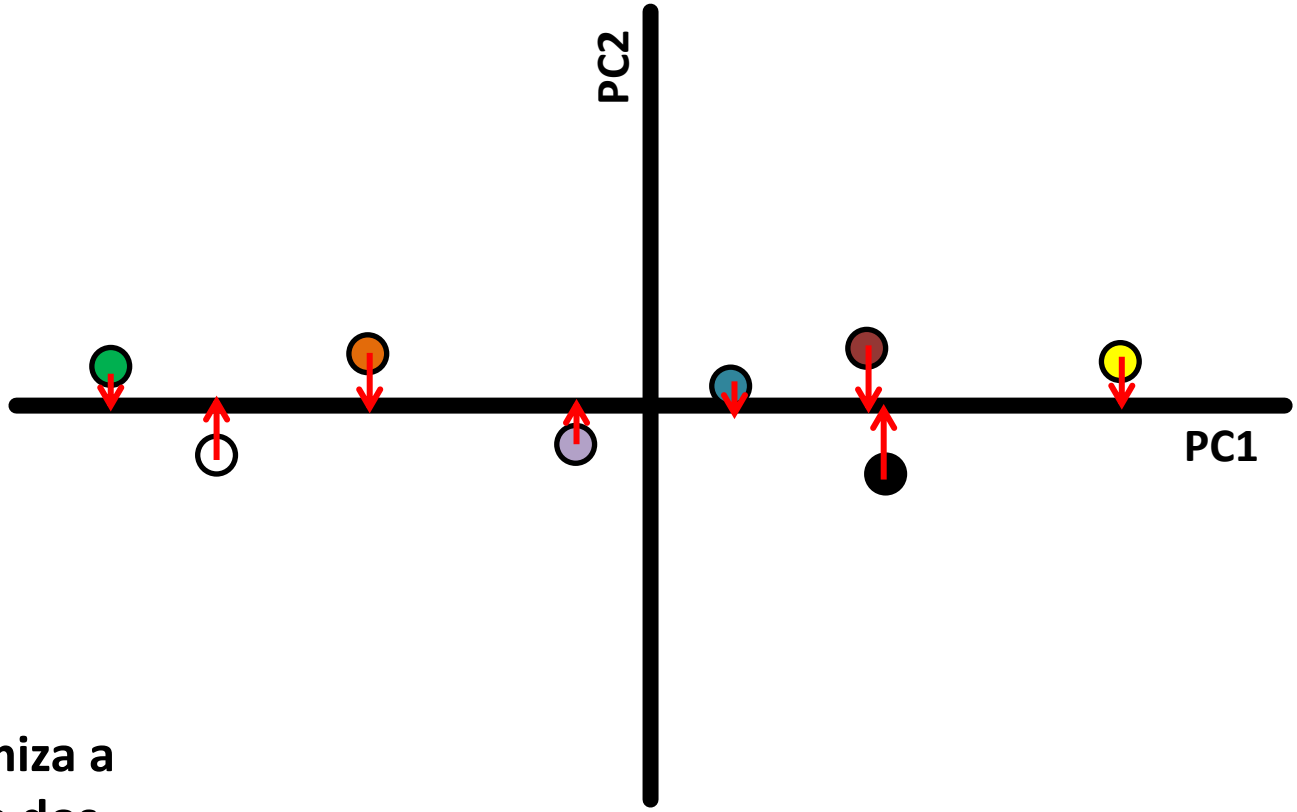
Matriz de dispersão $$\mathbf{S} = \frac{1}{n-1}\mathbf{Y}_c{}' \mathbf{Y}_c = \begin{bmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{bmatrix}$$

Input na Equação característica $$\left|\mathbf{S} - \lambda_k \mathbf{I}\right| = \left| \begin{bmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{bmatrix} - \begin{bmatrix} \lambda_k & 0 \\ 0 & \lambda_k \end{bmatrix} \right| = 0$$

8.2 (64.3%)+5.8 (35.7%)=14

Cálculo dos autovetores e autovalores $$(\mathbf{S} - \lambda_k \mathbf{I})\,\mathbf{u}_k = \mathbf{0}$$

Depois de normalizados (padronizados para = 1) tornam-se os componentes principais (eixos) $$\mathbf{U} = \begin{bmatrix} 0.8944 & -0.4472 \\ 0.4472 & 0.8944 \end{bmatrix}$$

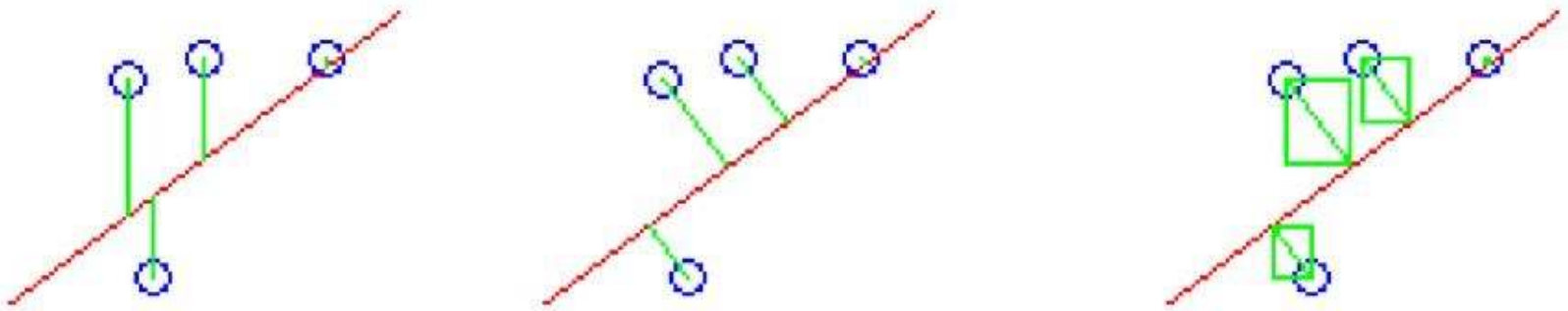


Minimiza a soma dos quadrados dos resíduos

Slide Thiago Gonçalves Souza

Three Ways of Calculating a Regression

Least Squares | **Major Axis** | **Reduced Major Axis**



Least squares regression is appropriate when there is uncertainty only regarding the *y*-variable. If both variables are subject to sampling and measurement error, major axis or reduced major axis regression is recommended. In the first two cases, the sum of the squared distances indicated by the green lines is minimized. In the final case, it is the areas of the triangles bounded by the horizontal and vertical green lines that are summed and minimized.

http://www.blc.arizona.edu/courses/schaffer/182/Giraffe/Regression.htm

A posição de um objeto (eg., pontos de amostragem) ao longo dos eixos da PCA (autovetores) é dada pela seguinte *combinação linear*:

$$f_{i1} = (y_{i1} - \bar{y}_1)\, u_{11} + \ldots + (y_{ip} - \bar{y}_p)\, u_{p1} = [y - \bar{y}]_i \mathbf{u}_1$$

Autovetores são ortogonais, por definição (Veja aula 1)

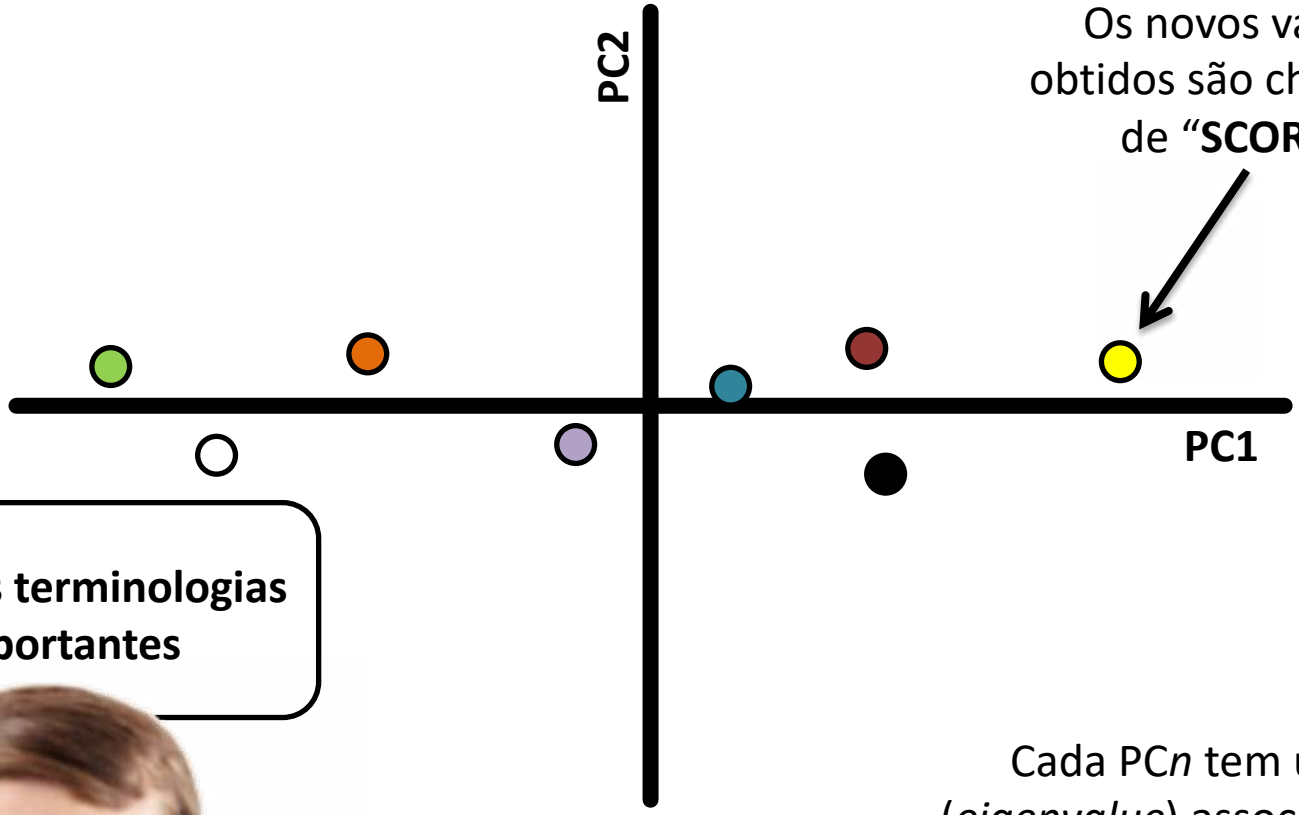
Ortho-gonality

It is easy to check the orthogonality of the two eigenvectors: their cross-product $\mathbf{u}'_1\mathbf{u}_2 = (0.8944 \times (-0.4472)) + (0.4472 \times 0.8944) = 0$. Moreover, Section 4.4 has

p. 432 L&L

Slide Thiago Gonçalves Souza

Olha a combinação linear aí, gente! → $PC1 = 4.2\ Y_1 - 3.8\ Y_2 ... + a_{in} Y_n$



Os novos valores obtidos são chamados de "**SCORES**"

Algumas terminologias importantes

Cada PC*n* tem um **autovalor** (*eigenvalue*) associado. A soma dos autovalores de todos os eixos representa a **porcentagem total de variação.**

O procedimento anterior implica em que a posição relativa dos objetos no espaço multimensional rotacionado (componentes principais) é preservada, ou seja, equivale à distância Euclidiana entre eles

Como essas informações aparecem no R?

```
Partitioning of variance:
               Inertia Proportion
Total             35.4          1
Unconstrained     35.4          1

Eigenvalues, and their contribution to the variance

Importance of components:
                          PC1    PC2    PC3    PC4
Eigenvalue              4.625  3.492  2.444  2.297
Proportion Explained    0.130  0.098  0.069  0.064
Cumulative Proportion   0.130  0.229  0.298  0.363
```

total variance of the dataset

eigenvalue of the first unconstrained axis

variance represented by the first unconstrained axis = eig_1/total_variance

cumulative variance represented by the first two unconstrained axes = (eig_1+eig_2)/total_variance

Slide Thiago Gonçalves Souza

Como interpretar os eixos?

- **Loadings**: correlação de Pearson entre PCs e as variáveis originais
- Não se pode testar significância da correlação, já que os eixos são combinações lineares
- Intervalo de confiança baseado num Bootstrap modificado (Peres-Neto et al. 2003)

Ecology, 84(9), 2003, pp. 2347–2363

GIVING MEANINGFUL INTERPRETATION TO ORDINATION AXES: ASSESSING LOADING SIGNIFICANCE IN PRINCIPAL COMPONENT ANALYSIS

PEDRO R. PERES-NETO,[1] DONALD A. JACKSON, AND KEITH M. SOMERS

Variables	PC1	PC2	PCA3
T	0.009	−0.138	0.670*
pH	0.496	−0.514*	−0.216
EC	0.974*	0.069	0.012
DO	0.576*	−0.267	0.270
NH_3^-	0.124	−0.916*	0.033
NO_2^-	0.963*	0.169	0.029
NO_3^-	0.116	0.097	0.739*
$PO_4^{\equiv}$	0.967*	0.082	−0.041
Cl^-	0.977*	0.146	−0.027
$SO_4^=$	0.949*	0.219	−0.080
Na^+	0.913*	0.208	−0.081
Ca^{++}	0.904*	−0.308	0.040
Mg^{++}	0.067	−0.877*	−0.069
Variance explained by components	6.93	2.22	1.14
Percent of total variance explained	53.31	17.07	8.76

https://www.researchgate.net/publication/26312587_Assessment_of_surface_water_quality_of_the_Ceyhan_River_basin_Turkey

Quantos eixos interpretar?

Available online at www.sciencedirect.com
SCIENCE DIRECT
ELSEVIER
Computational Statistics & Data Analysis 49 (2005) 974–997
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

How many principal components? stopping rules for determining the number of non-trivial axes revisited

Pedro R. Peres-Neto*, Donald A. Jackson, Keith M. Somers

- Critérios de seleção de eixos (existem > 20)
  - Broken stick
  - Scree plot
  - Kaiser-Guttman
    - Interpretar os eixos com $\lambda > 1$

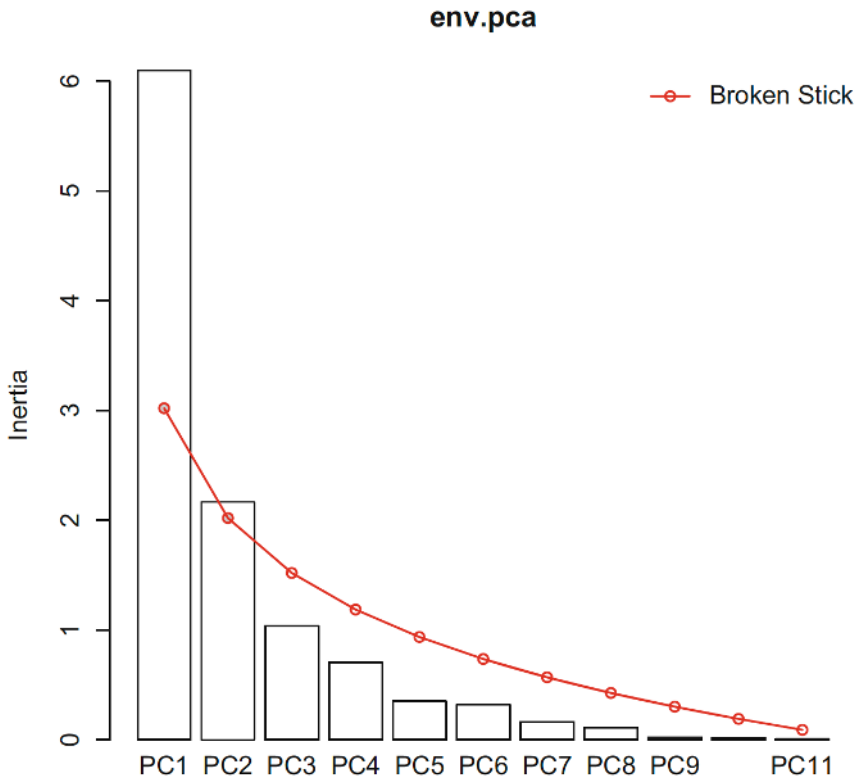
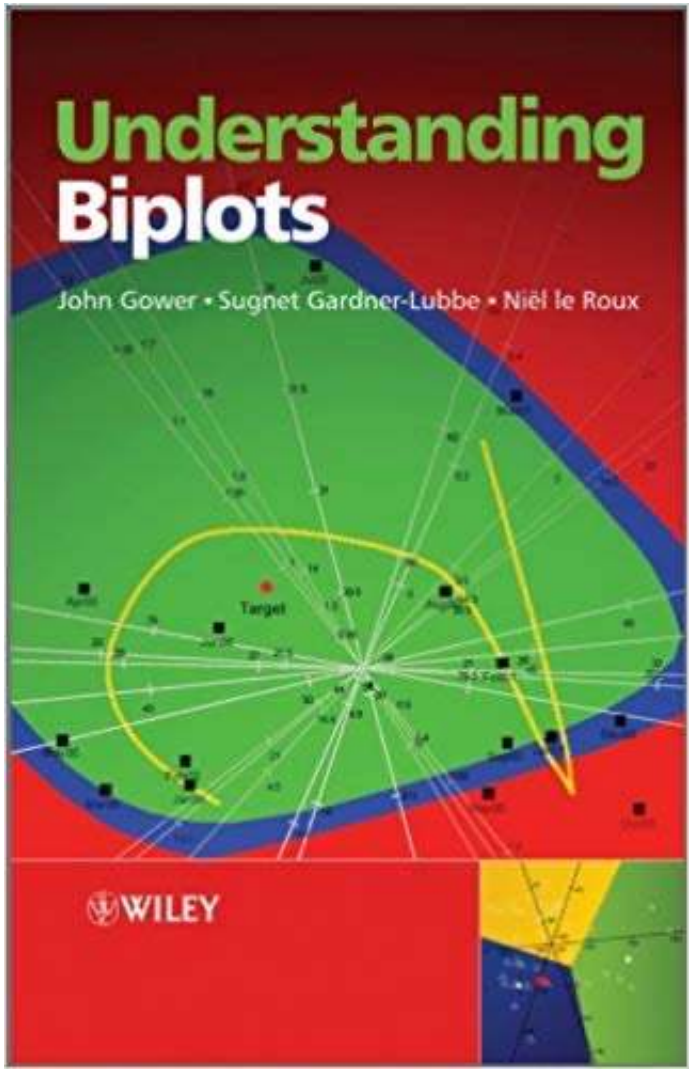


Fig. 5.1 Scree plot and broken stick model to help assess the number of interpretable axes in PCA. Application to the Doubs environmental data

Veja Peres-Neto et al. 2005 para mais detalhes

p. 448-9 L&L

Como interpretar um biplot?



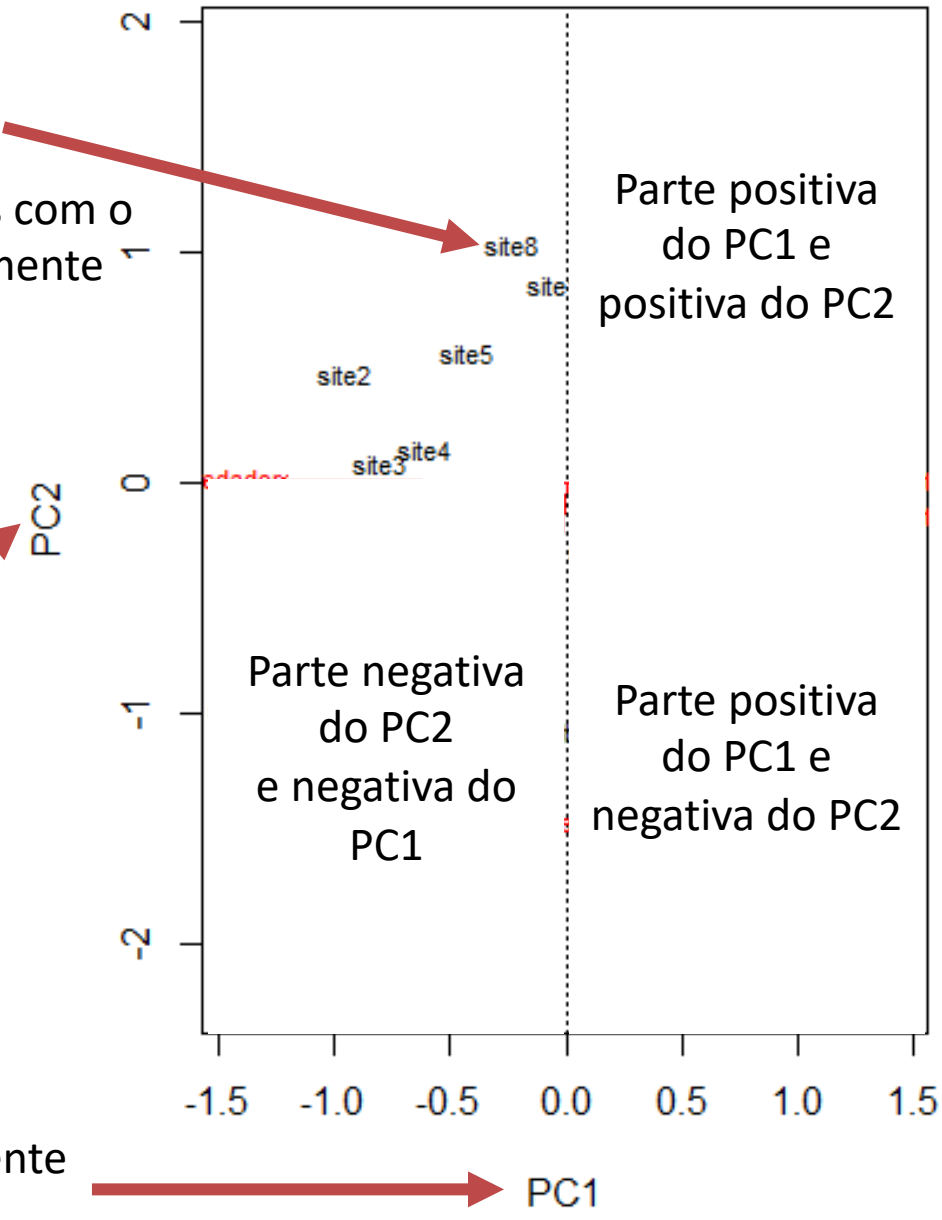
O que é um biplot?

- Diagrama de ordenação que representa dois tipos de informação
  - Posição dos descritores (e.g., variáveis ambientais)
    - Geralmente mostrados como setas
  - Posição dos objetos (e.g., locais)
- Dois tipos:
  - Scaling 1 (enfatiza distância)
    - Bom para mostrar relação entre variáveis
  - Scaling 2 (enfatiza correlação)
    - Bom para mostrar relações entre objetos

p. 443-5 L&L

No scaling 1

Posição (score) dos objetos Positivamente correlacionados com o PC2 e negativamente com o PC1



Distância entre objetos é a sua distância Euclidiana. O mesmo não ocorre no scaling 2

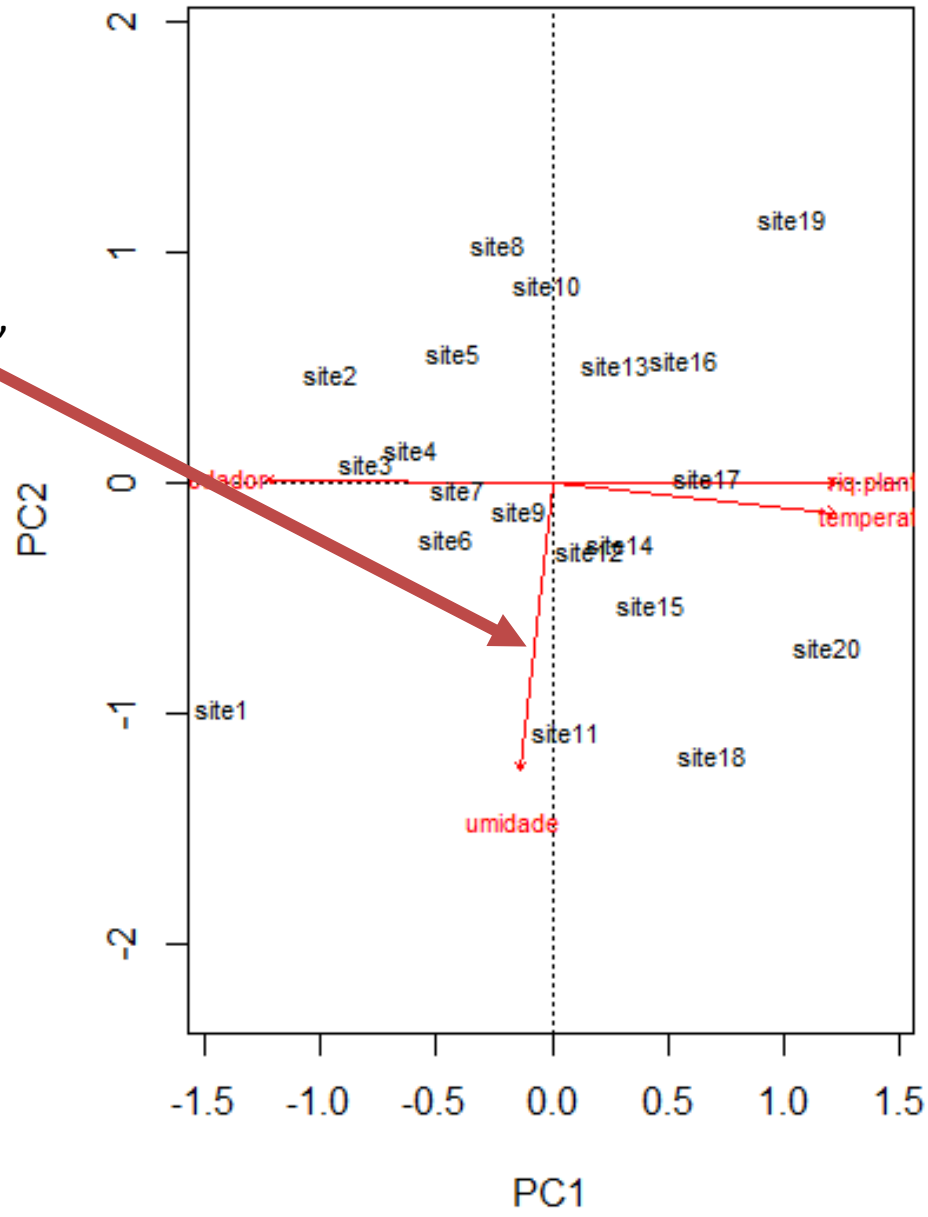
2º componente principal

Dividido em 4 "quadrantes"

1º componente principal

No scaling 1

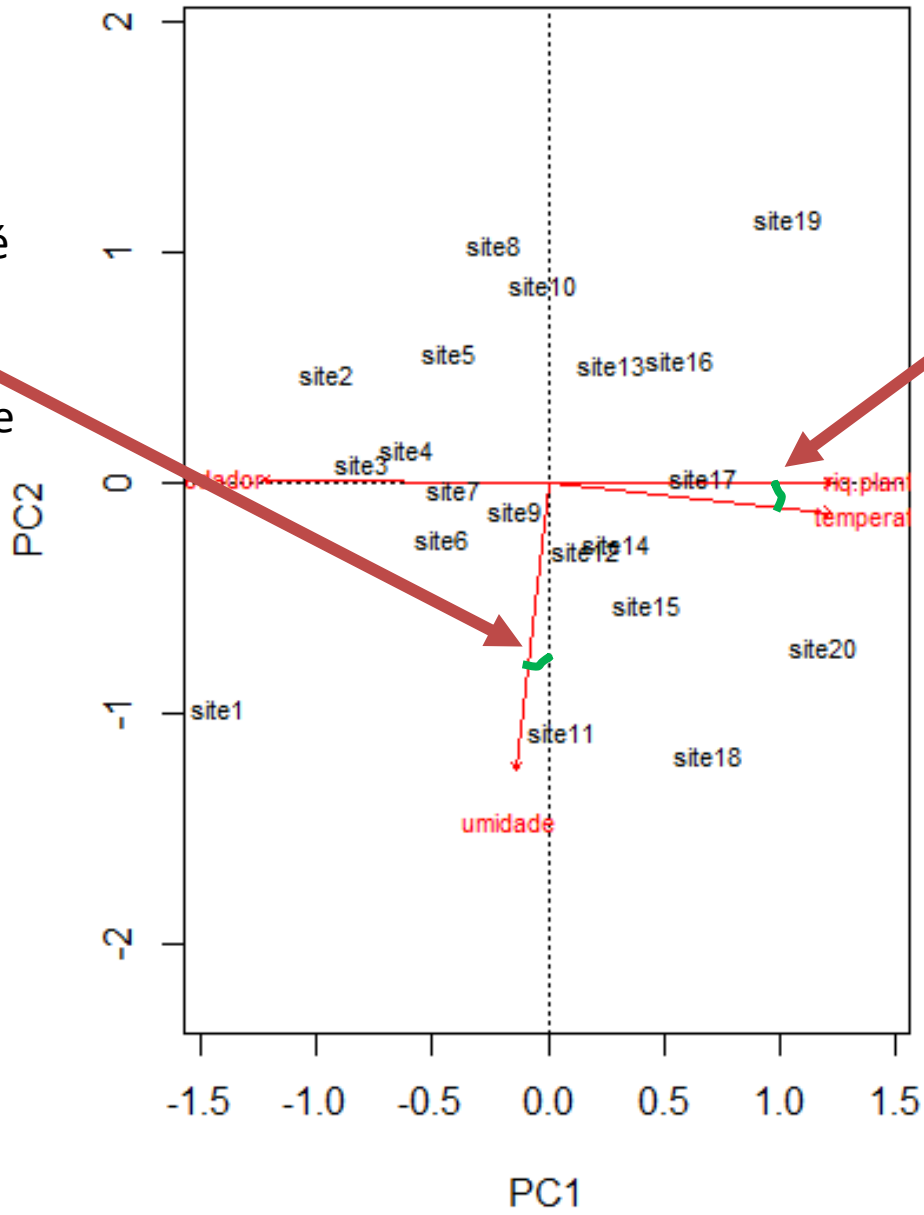
Quanto maior a seta, maior a correlação do descritor com o eixo



No scaling 2

Ângulo* de curvatura da seta indica o quanto um descritor é correlacionado com o PC

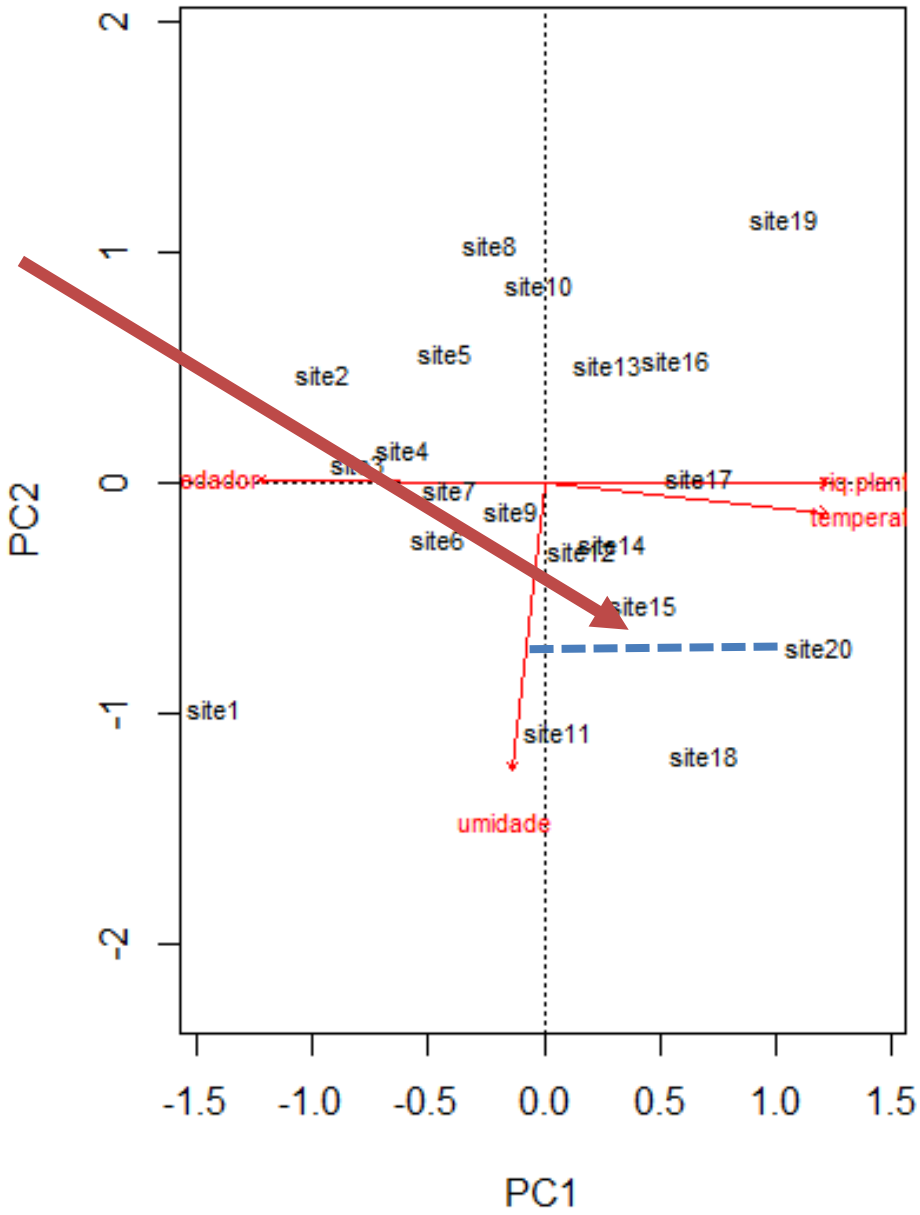
Neste caso, a umidade é mais correlacionada com o PC2, do que PC1



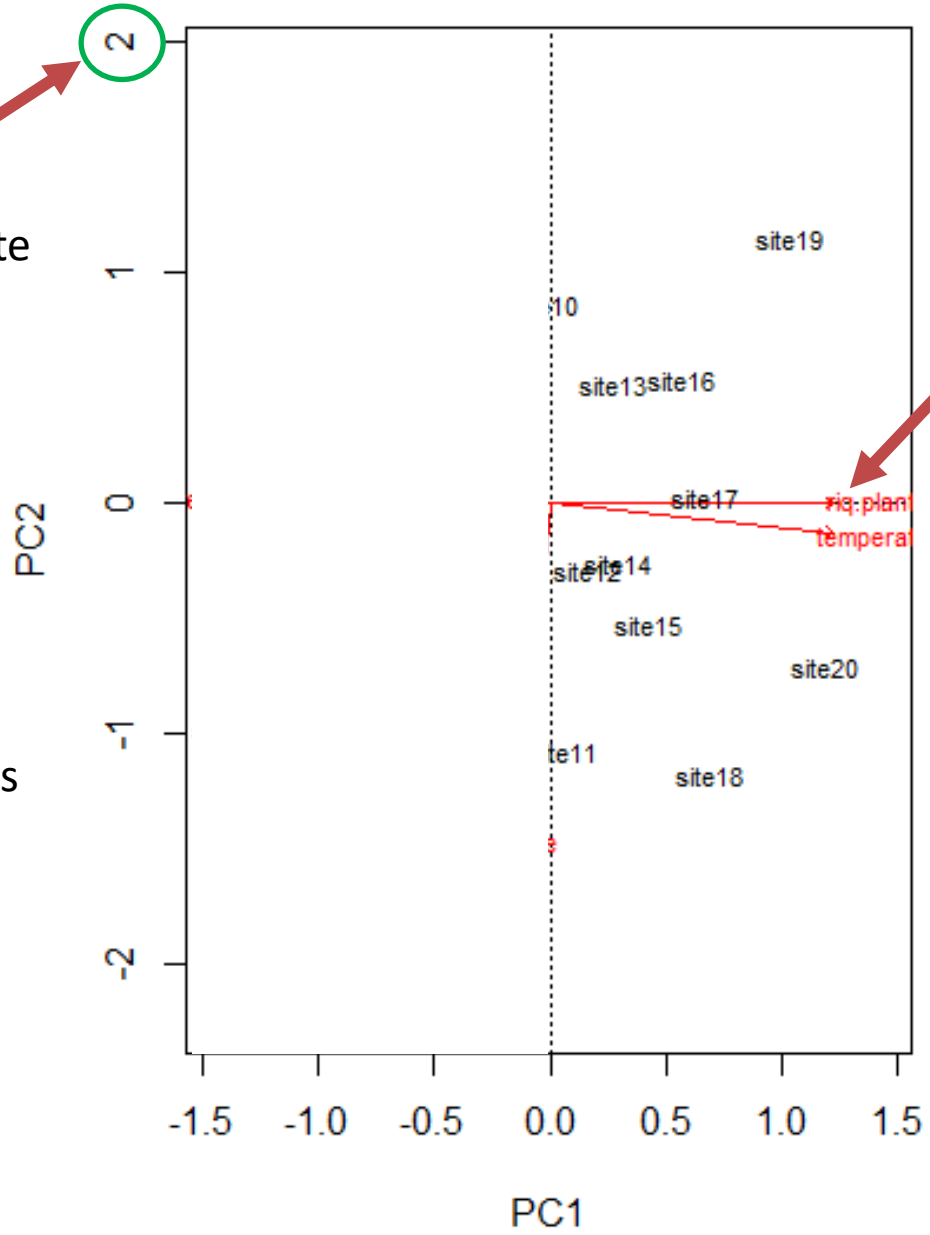
Ângulo entre os descritores indica as suas covariâncias

No scaling 2

Ao projetar em 90º um objeto em direção ao descritor, obtém-se a posição do objeto ao longo do descritor



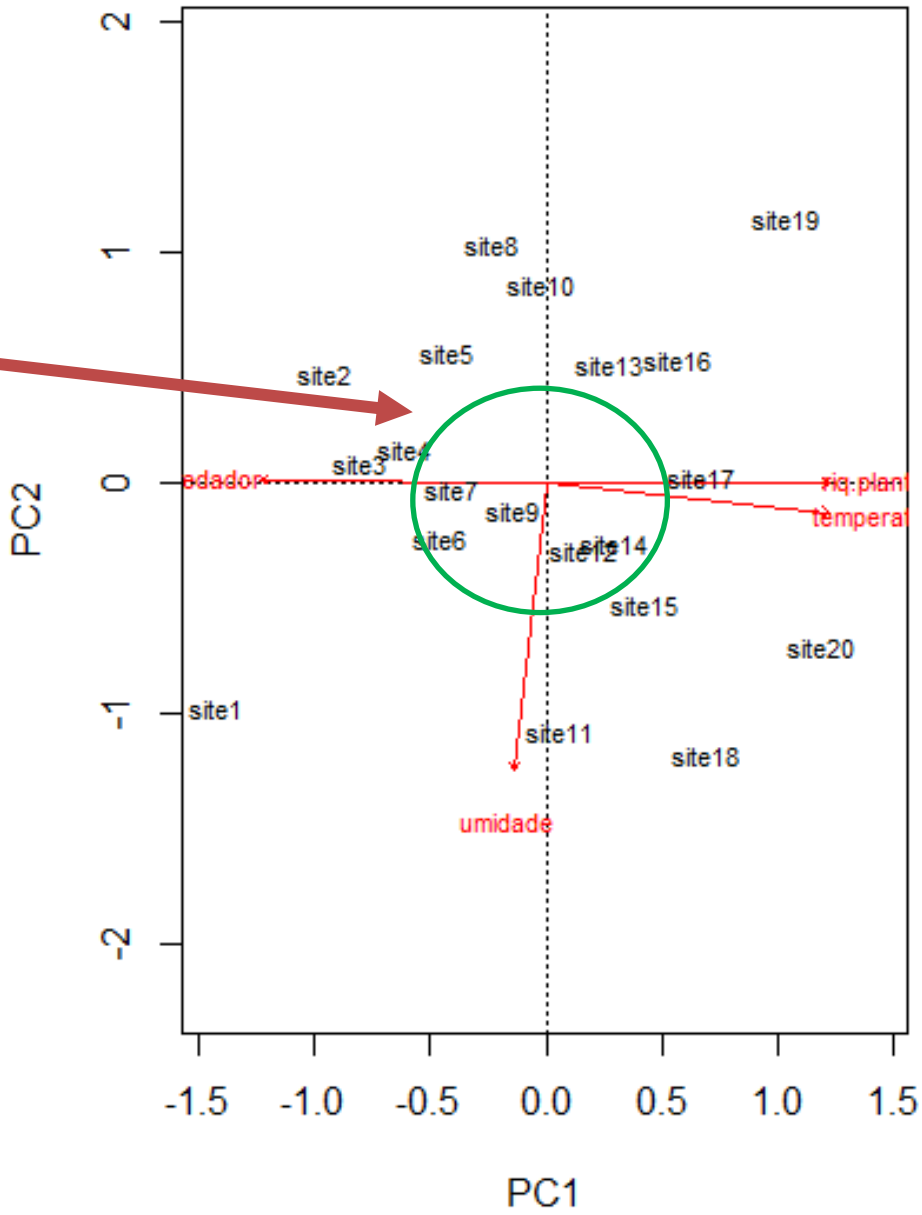
Elementos do Autovetor (componente principal)



Posição (score) dos descritores Positivamente correlacionados com o PC1 e negativamente com o PC2

Descritores são mostrados como seta para diferenciá-los dos objetos, que são mostrados como pontos

Descritores e objetos que estão no centro do diagrama de ordenação indica que contribuíram pouco pra formação tanto com o PC1 e PC2



Círculo de equilíbrio

- Maneira visual

e rápida de se estimar

a relevância de descritores

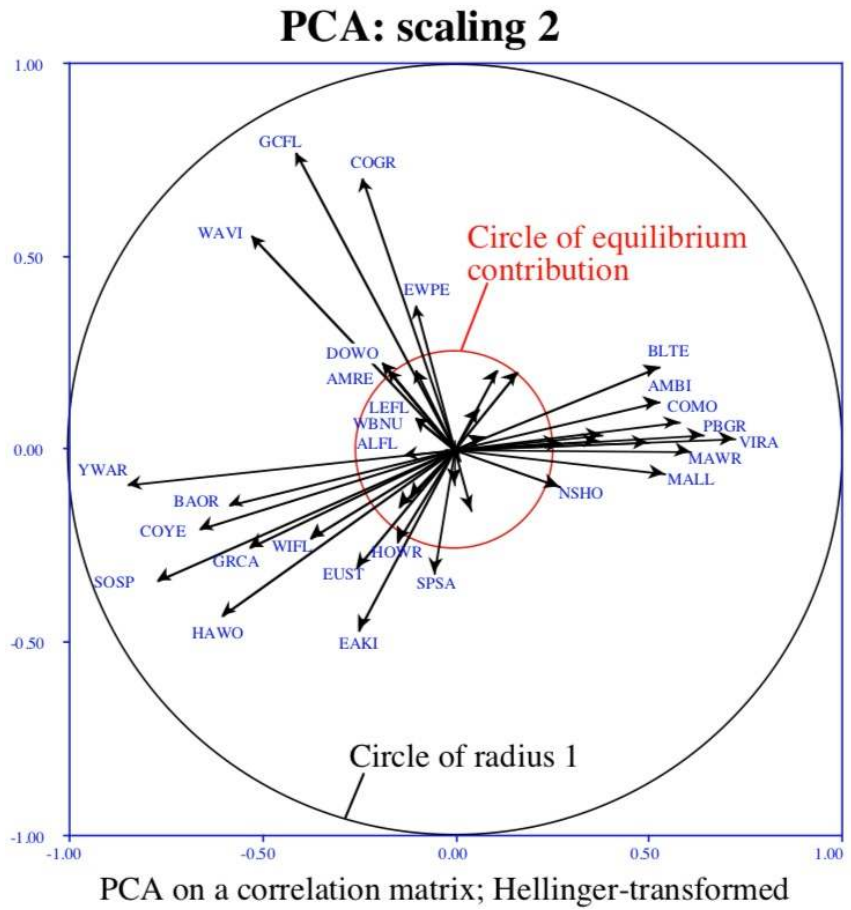
para a construção dos

componentes principais

- As setas que estiverem

dentro do círculo têm

pouca relevância



p. 437-9 L&L

Table 9.4 Questions that can be addressed in the course of a principal component analysis and the answers found in Section 9.1.

Before starting a principal component analysis	*Pages*
1) Are the descriptors appropriate?	
⇒ Quantitative descriptors; multinormality; not too many zeros.	450-452
2) Are the descriptors dimensionally homogeneous?	
⇒ If YES, conduct the analysis on the dispersion matrix	442, 446
⇒ If NO, conduct the analysis on the correlation matrix	445-448
3) Purpose of the ordination in reduced space:	
⇒ To preserve and display the relative positions of the objects: scale the eigenvectors to unit lengths to obtain matrix **U**	432
Draw a distance biplot (descriptors: **U**; objects: **F** = **YU**)	444
⇒ To display the correlations among descriptors: scale the eigenvectors to $\sqrt{\lambda}$ to obtain matrix $\mathbf{U}_{sc2}$	436
Draw a correlation biplot (descriptors: $\mathbf{U}_{sc2}$; objects: $\mathbf{G} = \mathbf{F}\boldsymbol{\Lambda}^{-1/2}$) (beware: Euclidean distances among objects are not preserved)	444
While examining the results of a principal component analysis	
1) How informative is a representation of the objects in an m-dimensional reduced space?	
⇒ Compute eq. 9.5	433
2) Are the distances among objects well preserved in the reduced space?	
⇒ Compare Euclidean distances using a Shepard diagram	427-428
3) Which eigenvalues are important?	
⇒ Is λ_k larger than the mean of the λ's?	448
⇒ Is the percentage of the variance corresponding to λ_k larger than the corresponding value in the broken stick model?	449
4) What are the descriptors that contribute the most to the formation of the reduced space?	
⇒ Compute the equilibrium contribution of descriptors and, when appropriate, draw the circle	437, 439, 442, 448
⇒ Compute correlations between descriptors and principal axes	440, 442, 448
⇒ Compute the table of "Cumulative fit per descriptor"	441-442
5) How to represent the objects in the reduced space?	
⇒ Scaling 1: $\mathbf{F} = \mathbf{Y}_c\mathbf{U}$; scaling 2: $\mathbf{G} = \mathbf{F}\boldsymbol{\Lambda}^{-1/2}$	433-435, 443-444
⇒ Compute the table of "Cumulative percent fit of the objects"	443

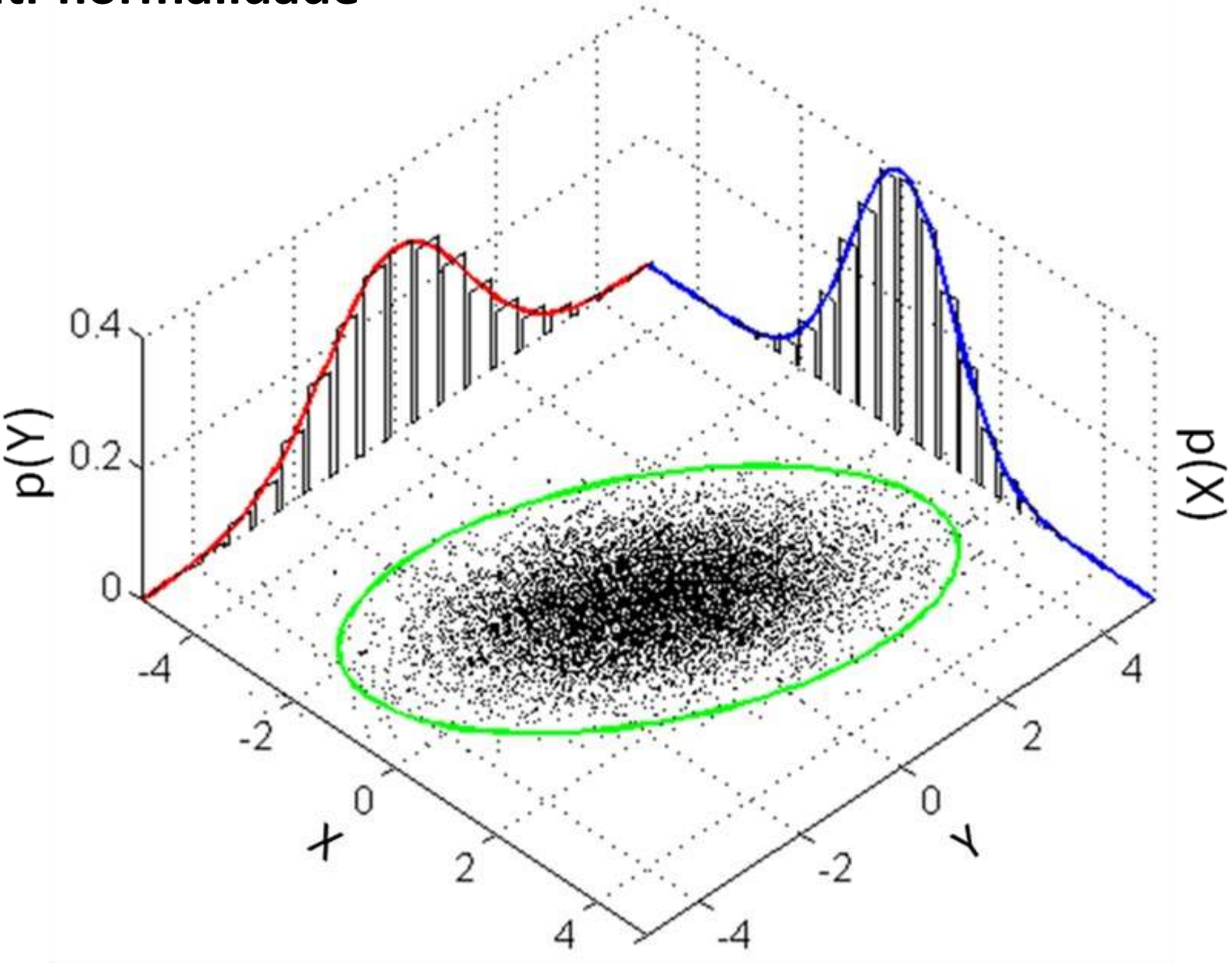
- 1) *Independência* dos componentes entre si, de modo a não invalidar a ortogonalidade. Útil para gerar variáveis combinadas.

- 2) *Aditividade*: os componentes devem somar suas contribuições, sem interação. A soma dos eigenvalues (valores próprios) dos componentes (eixos) dá a variância total dos pontos em relação a todos os eixos (*SS* das correlações).

- 3) *Linearidade* das relações entre descritores e entre descritores e objetos (gradiente) é assumida. Se espera relação unimodal, faça uma CA (daqui a alguns slides)

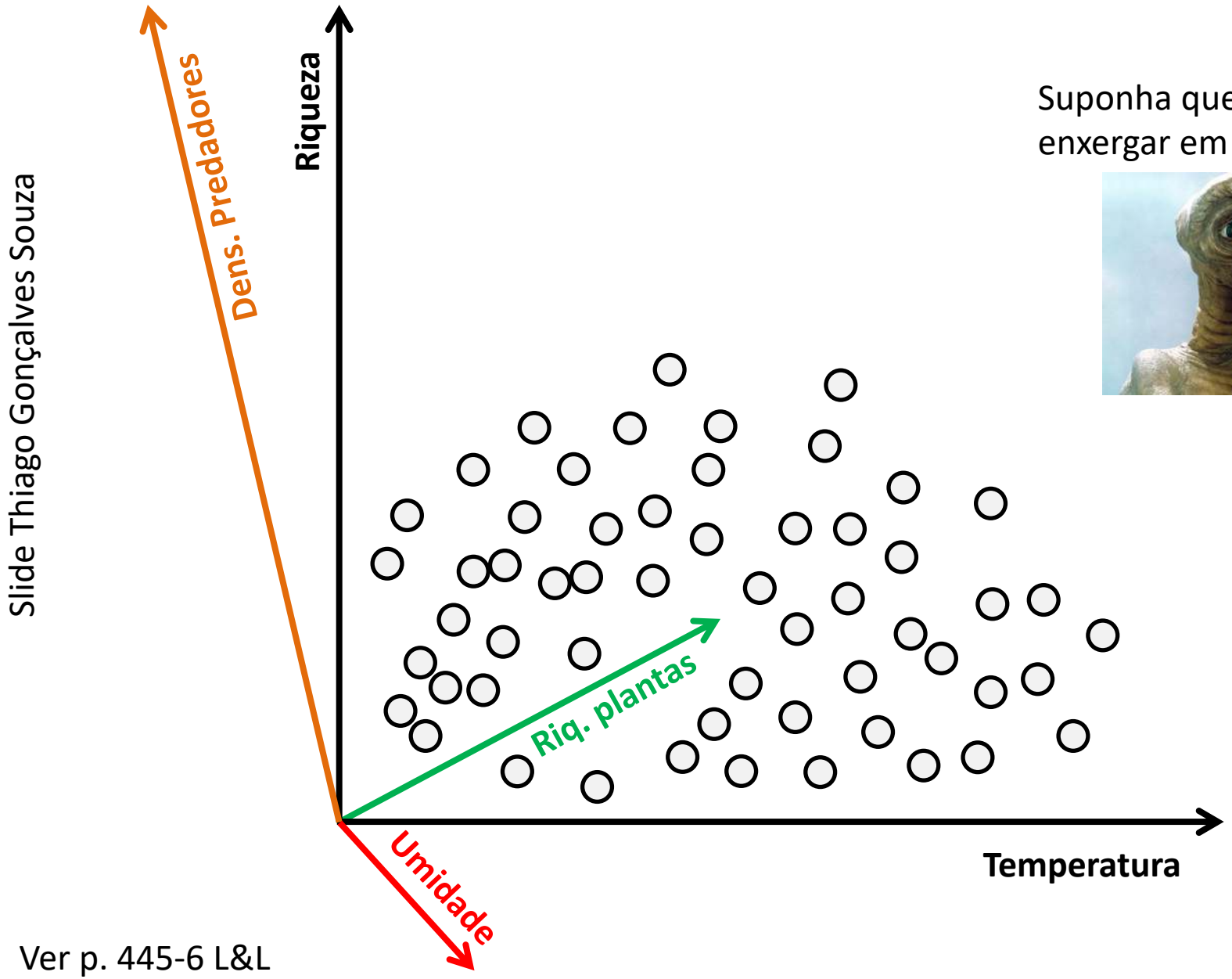
- 4) *Multi-normalidade* dos dados, especialmente para dar sentido estatístico aos resultados. Logo, não é possível calcular uma PCA de dados ordinais, categóricos ou multiestado. (mas veja Hill-Smith 1976, ade4::dudi.hillsmith)

Multi-normalidade



Slide Thiago Gonçalves Souza

PROBLEMA: utilização de variáveis não-padronizadas



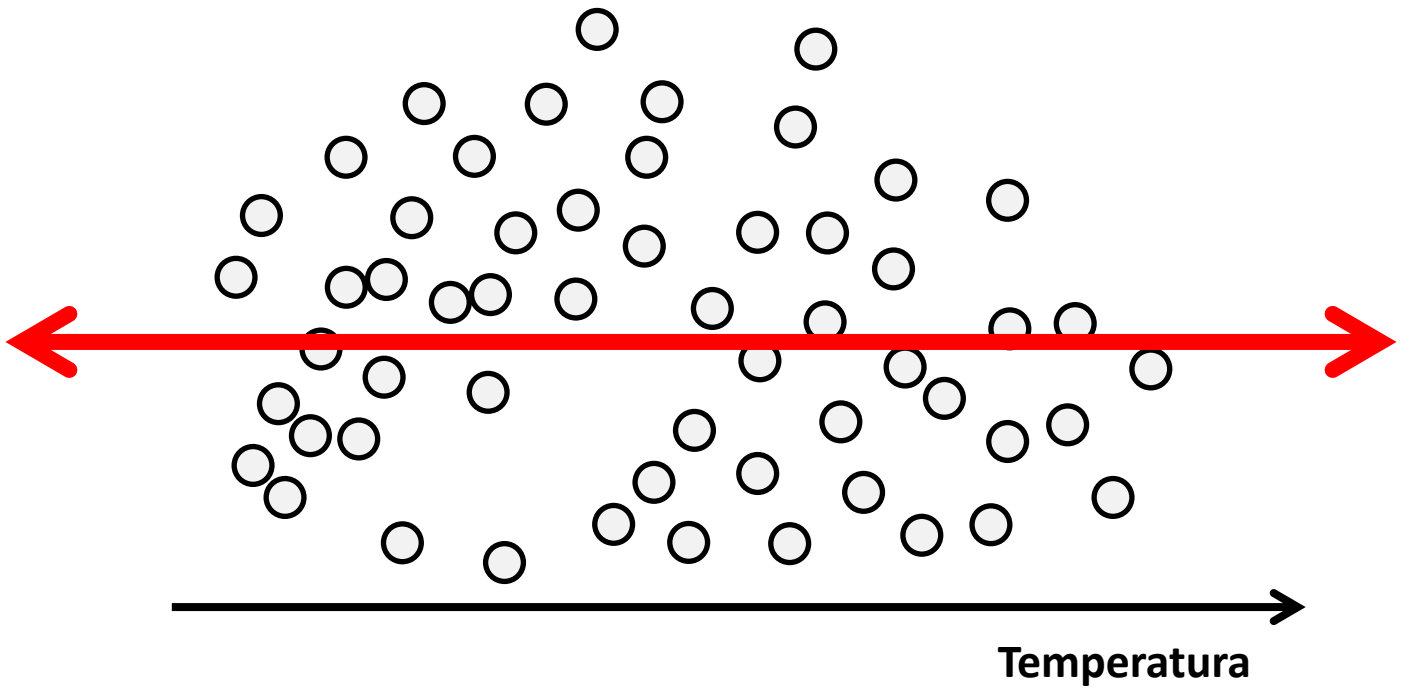
Suponha que você consiga enxergar em 5 dimensões

Slide Thiago Gonçalves Souza

Ver p. 445-6 L&L

PROBLEMA 1: utilização de variáveis não-padronizadas

As variáveis com maiores unidades de medida podem dominar a ordenação



Slide Thiago Gonçalves Souza

Ver p. 445-6 L&L

PCA de matriz de correlação ou covariância?

- PCA sobre variáveis brutas (centralizadas) (= PCA usando **matriz covariância**) só é apropriado quando as variáveis **são dimensionalmente homogêneas**.
- PCA usando **matriz de correlação**: quando as variáveis são **dimensionalmente heterogêneas** ou **quando se quer dar peso igual para todas as variáveis**

Ver p. 445-6 L&L

Só tem uma última regra...

p. 462-4
p. 450-1 L&L

(1) Unconstrained ordination analysis

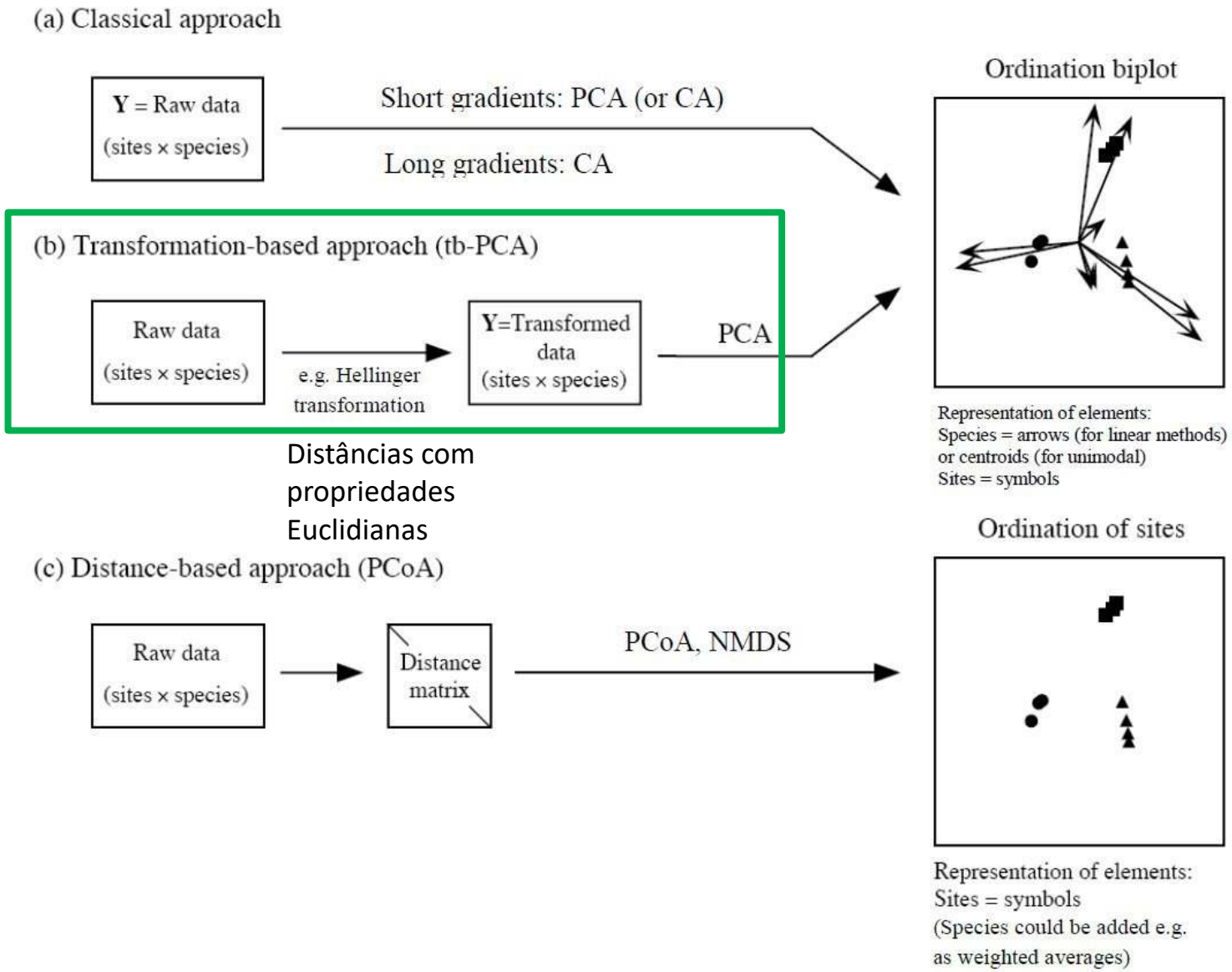


Fig. 9.8, p. 462-4 L&L

Análise de Correspondência (CA)

p. 464-82 L&L

O que é? Como funciona?

- Equivalente a uma PCA feita numa matriz de correlação que foi transformada com chi-quadrado
- Melhor desempenho do que PCA quando se tem muitos zeros na matriz
  - Ideal para análise de *gradientes longos*
- Aplicável a uma **matriz de espécies por locais** *contendo tanto incidência quanto abundância*
- *Relação unimodal* dos objetos (espécies) e descritores (locais)
- Sinônimo de Reciprocal Averaging (RA)

Como interpretar o biplot de uma CA?

- Distância entre objetos no biplot reflete a distância de chi-quadrado entre eles
- Locais mais próximos entre si no espaço reduzido possuem frequência relativa de espécies similar
- Espécies próximas de locais (setas) indicam que aquele local tem alta abundância (ou presença) da espécie

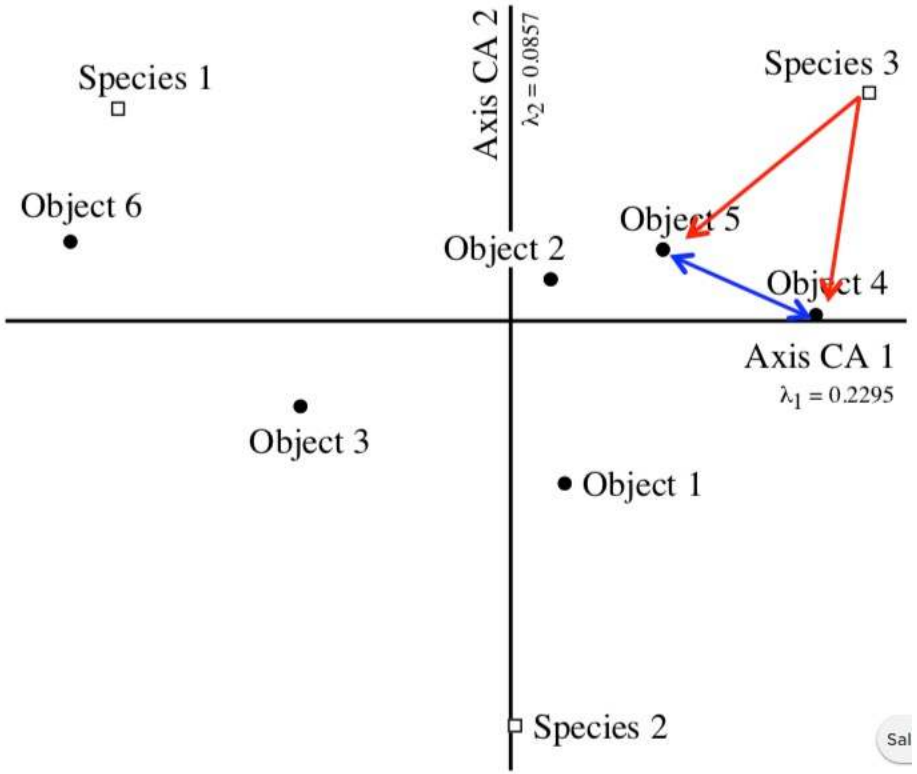
Ordenação de locais

4. Correspondence analysis (CA)

Scalings:

CA scaling type 1

	Spec.1	Spec.2	Spec.3
Obj.1	1	5	2
Obj.2	4	4	6
Obj.3	3	3	0
Obj.4	1	3	5
Obj.5	2	2	4
Obj.6	4	1	0



Slide P. Legendre

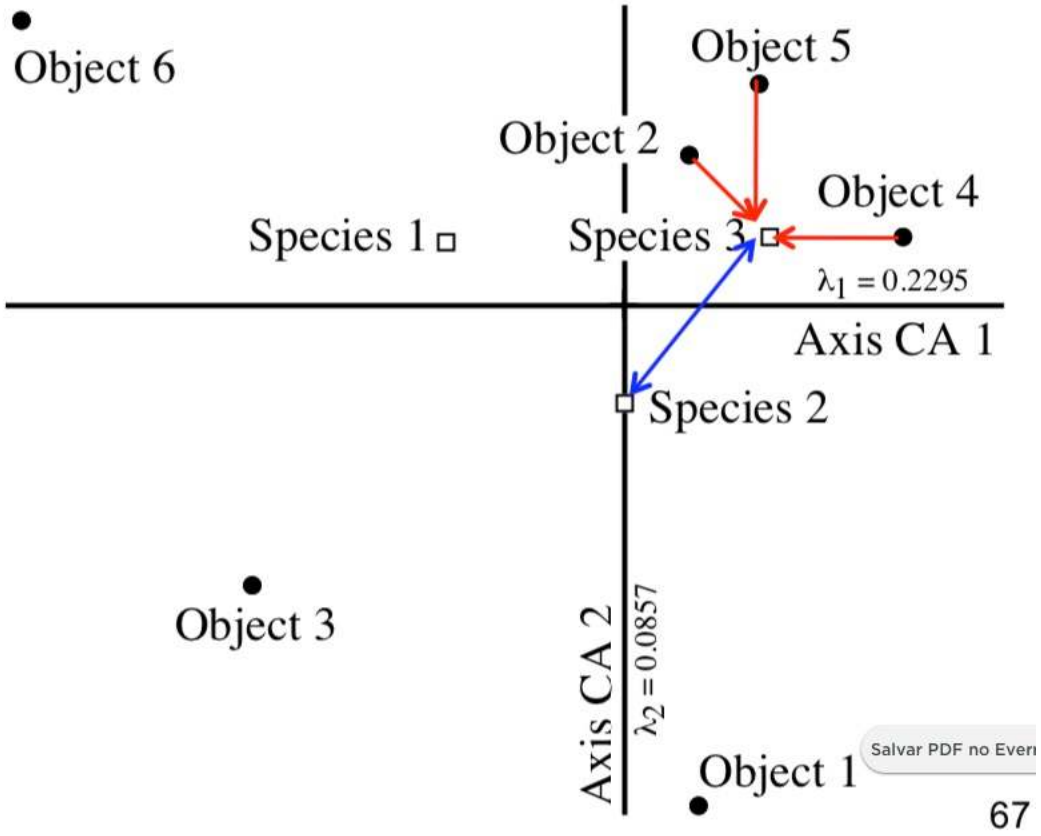
Ordenação de espécies

4. Correspondence analysis (CA)

Scalings:

CA scaling type 2

	Spec.1	Spec.2	Spec.3
Obj.1	1	5	2
Obj.2	4	4	6
Obj.3	3	3	0
Obj.4	1	3	5
Obj.5	2	2	4
Obj.6	4	1	0

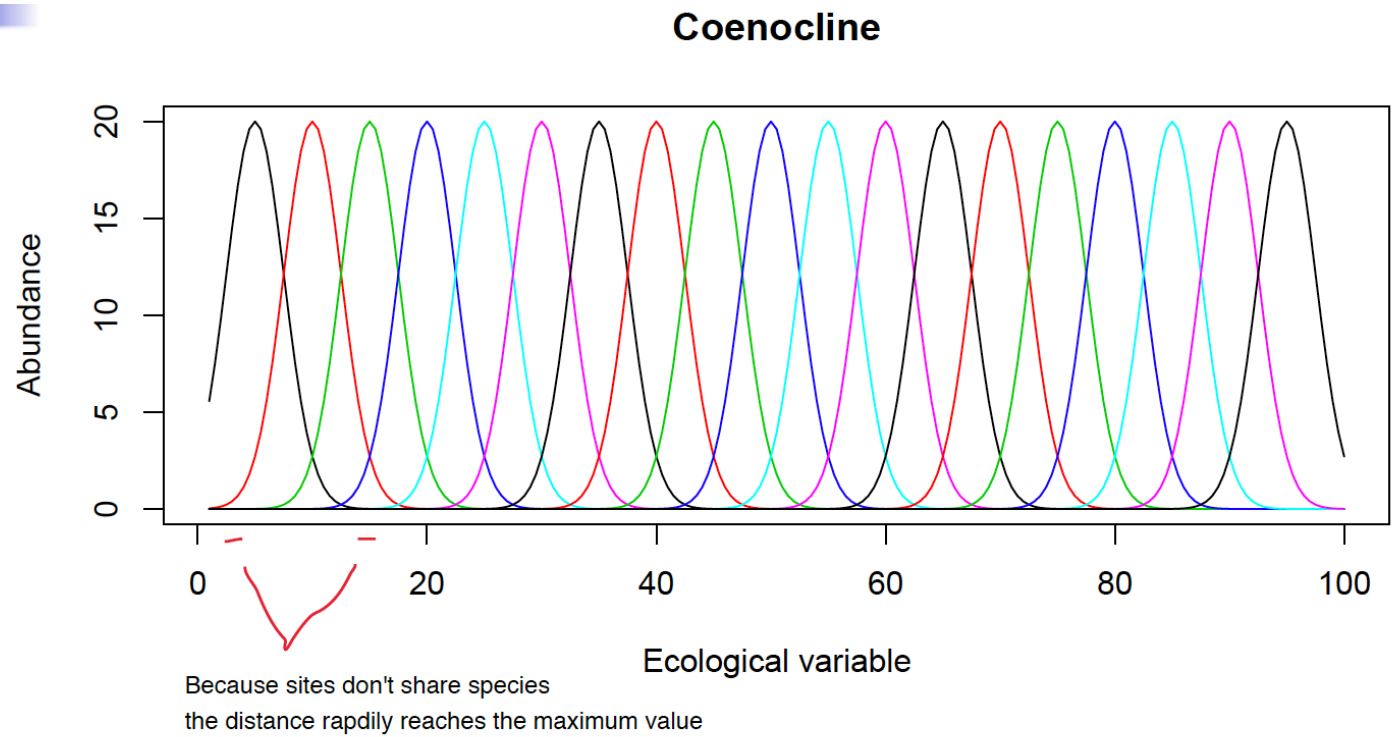


Slide P. Legendre

Detalhes da CA

- Problema com espécies raras ou pouco frequentes
  - Muitos zeros
  - Distância de chi-quadrado dá muito peso para espécies raras. Coeficiente assimétrico (exclui duplos zeros)
  - Aparecem muito deslocadas no diagrama de ordenação
- Efeito do arco
  - Ocorre em gradientes fortes (alto turnover)
  - Nos extremos do gradiente série de turnover é distorcida
  - Distância de chi-quadrado tem um limite superior

Simulação de distribuições de espécies



There are 69% zeros in the data file (99 × 19).

CA is an appropriate method for ordination of this type of data because it approximates a Gaussian ordination.

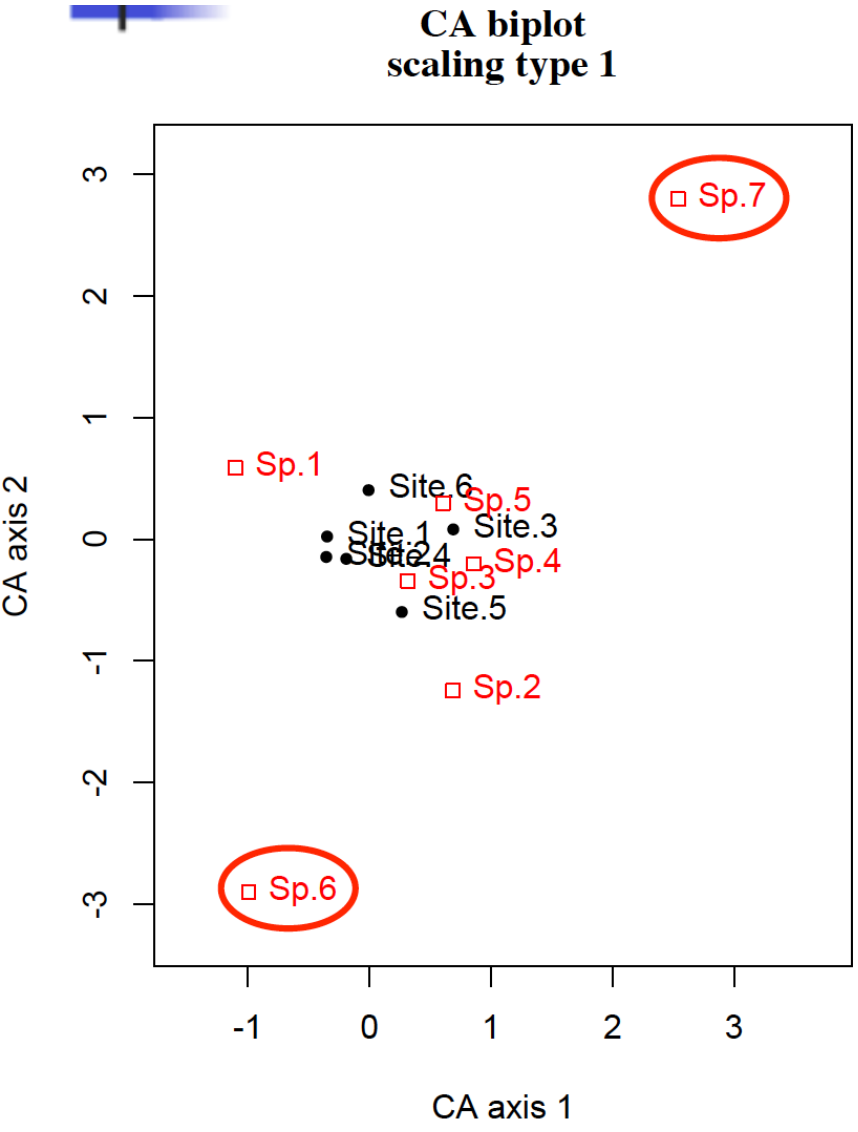
Slide P. Legendre

2. In scaling 1 biplots, rare species with few occurrences may take extreme values, meaning that they may be located far from the origin. First example (artificial data):

	Sp1	Sp2	Sp3	Sp4	Sp5	Sp6	Sp7
Site.1	45	10	15	10	3	2	0
Site.2	25	8	10	3	2	3	0
Site.3	7	15	20	12	5	0	10
Site.4	25	10	20	3	3	2	0
Site.5	7	15	10	10	2	3	0
Site.6	45	8	15	12	5	0	10
Sp.sums	154	66	90	50	20	10	20
Occur.	6	6	6	6	6	4	2

Species 6 and 7 have the smallest numbers of occurrences.

Species 5 has small total abundance.



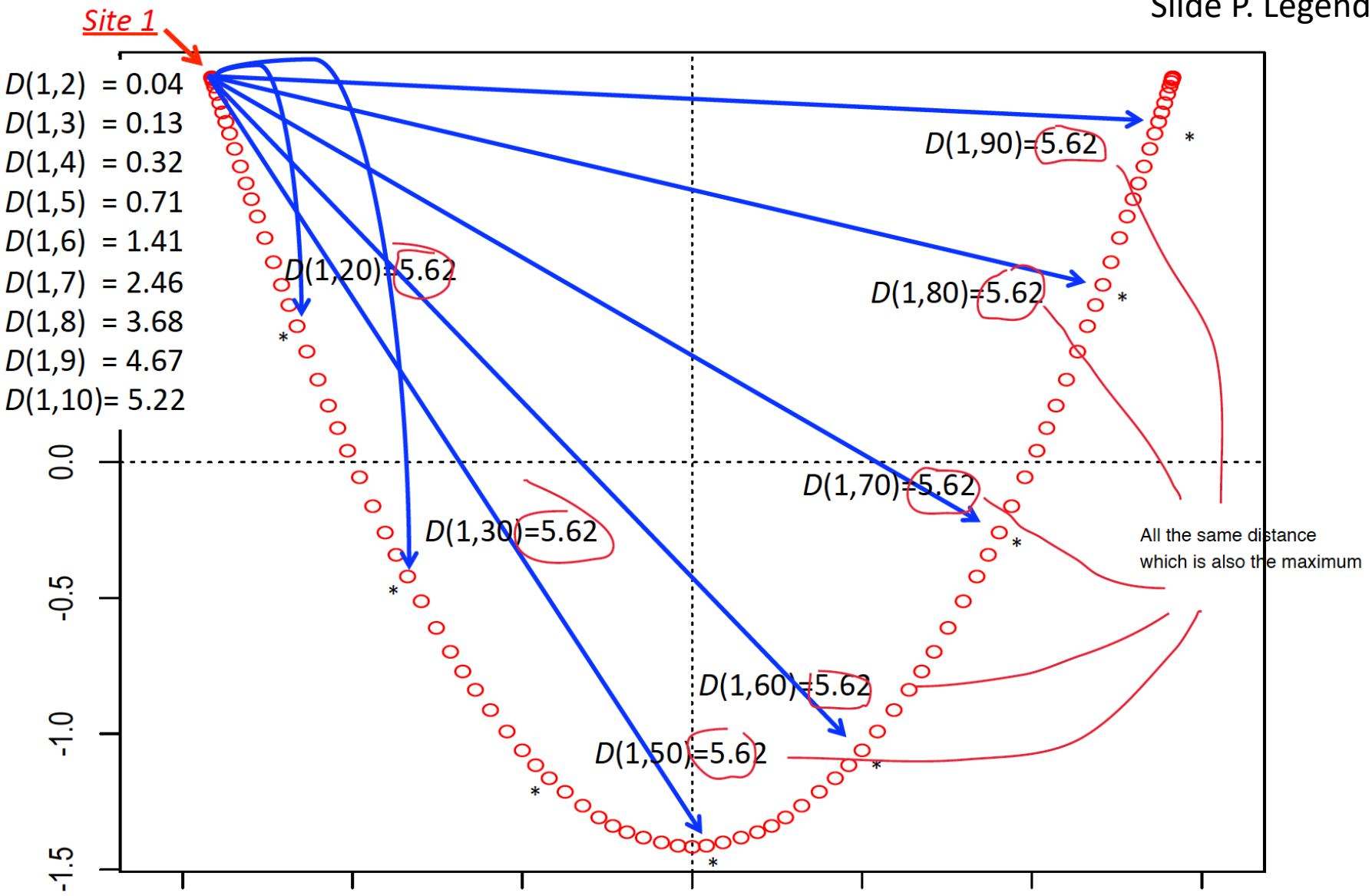

```
# Matrix V, species in scaling 1 biplot

       Axis1  Axis2  Axis3  Axis4
Sp.1 -1.102  0.587 -0.201 -0.043
Sp.2  0.687 -1.243 -0.263 -0.303
Sp.3  0.316 -0.344  1.534  0.621
Sp.4  0.857 -0.203 -1.880  1.267
Sp.5  0.607  0.296  0.537 -0.311
Sp.6 -0.993 -2.902 -0.700 -4.468
Sp.7  2.540  2.795  0.026 -2.088
```

Rare species #6 and 7 (low occurrences) are located far from the center of the plot.

- Species 6 is found in sites $\{1, 2, 4, 5\}$; it pulls these sites to the lower-left corner.
- Species 7 is found in sites $\{3, 6\}$; it pulls these sites towards the upper-right corner of the plot.

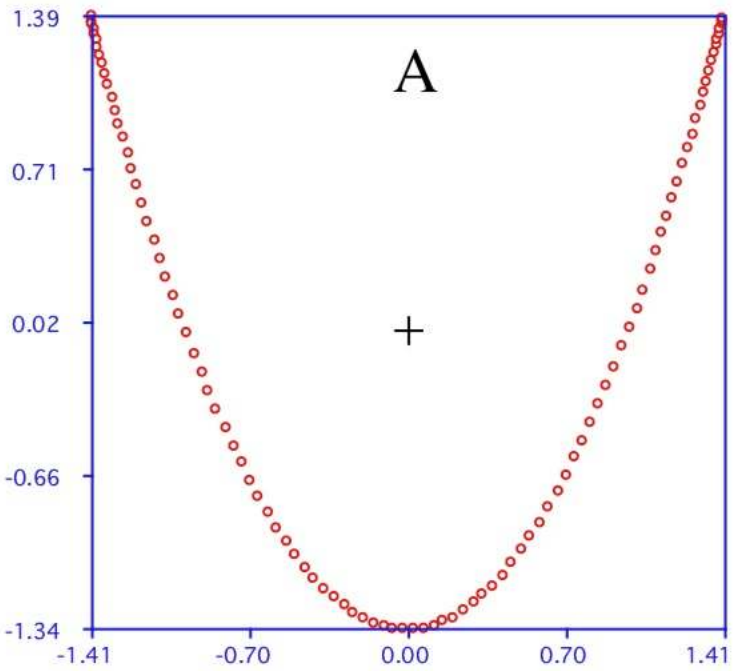
Slide P. Legendre

Slide P. Legendre

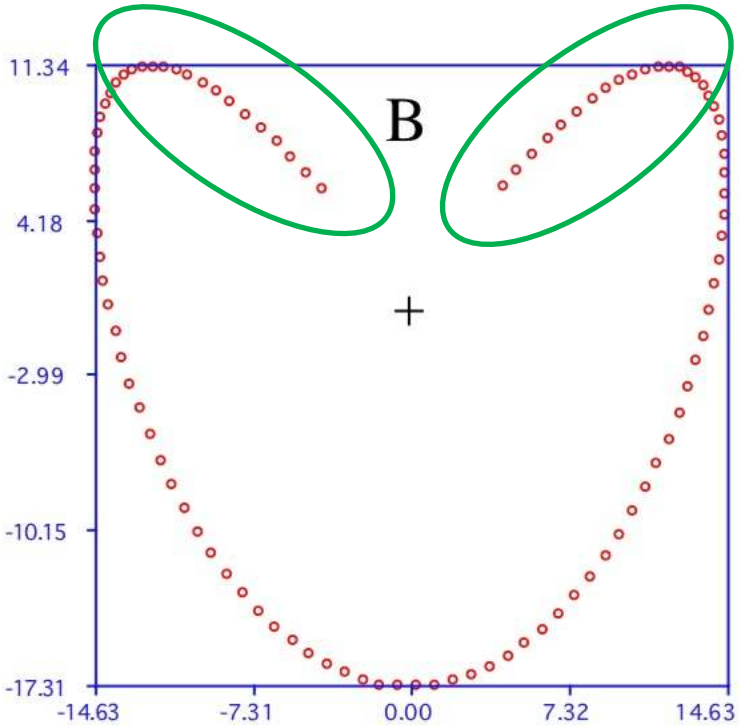


Chi-square distances between *Site 1* and selected sites along the coenocline. The distance reaches a maximum after 13 steps and does not increase thereafter. The plot reacts by forming an arch.

Arch and horseshoe effects



CA: arch effect

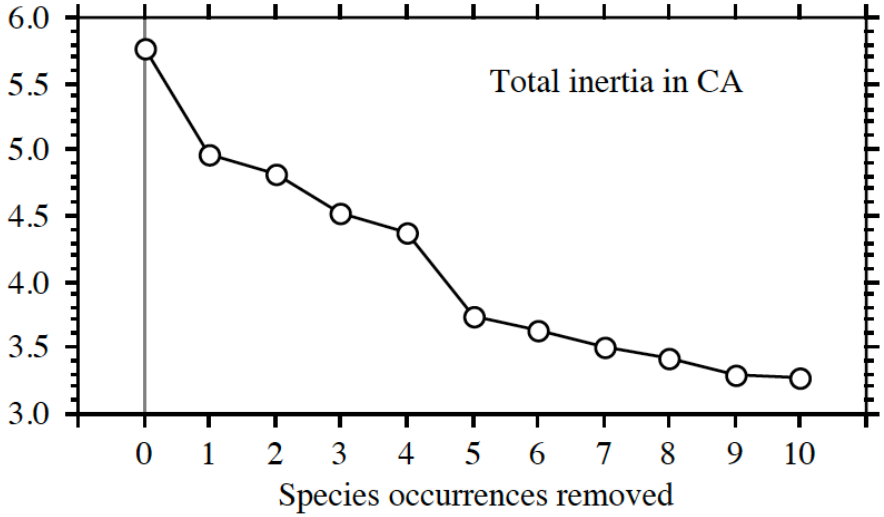


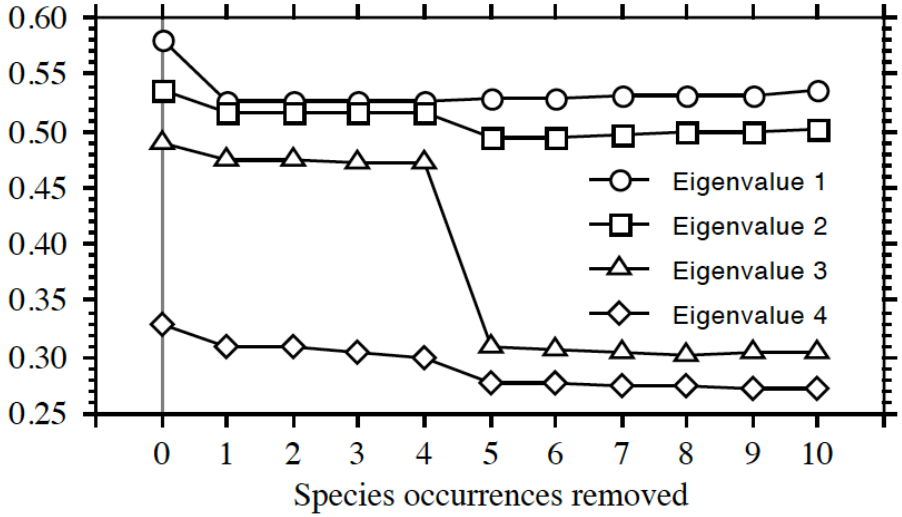
PCA: horseshoe effect

Slide P. Legendre

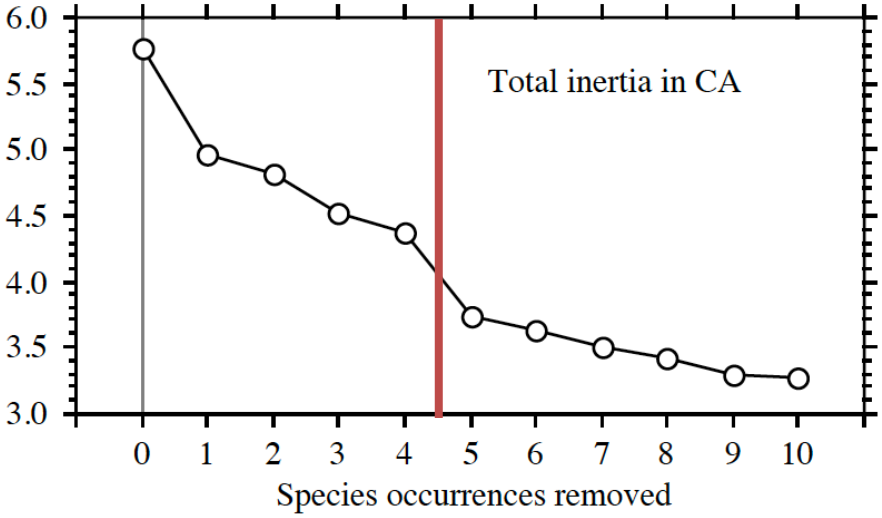
O que fazer nesses casos?

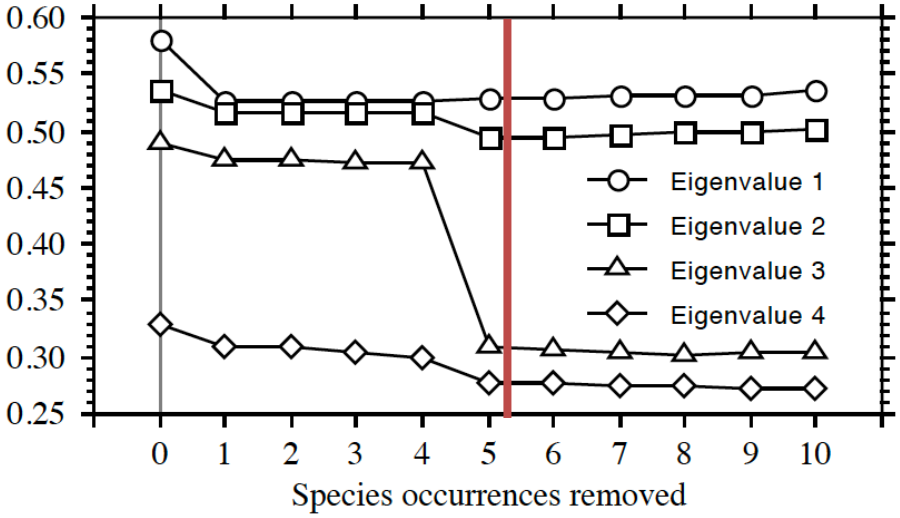
- O 2º eixo é uma transformação quadrática do 1º
  - 2º eixo é inútil
- Espécies raras são “exceções”
- O que fazer com elas depende do objetivo do estudo e do tipo de organismo
- Considere exclui-las da análise
  - Repetir a CA eliminando espécies com 1 indivíduo, depois dois etc
  - Note se a inercia (variação total) aumenta ao se eliminar essas espécies
  - Plote os autovalores e procure uma quebra


6.0
5.5
5.0
4.5
4.0
3.5
3.0
Total inertia in CA
0 1 2 3 4 5 6 7 8 9 10
Species occurrences removed


0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
Eigenvalue 1
Eigenvalue 2
Eigenvalue 3
Eigenvalue 4
0 1 2 3 4 5 6 7 8 9 10
Species occurrences removed

Espécies com até 4 indivíduos
podem ser removidas, pois não
afetam a análise

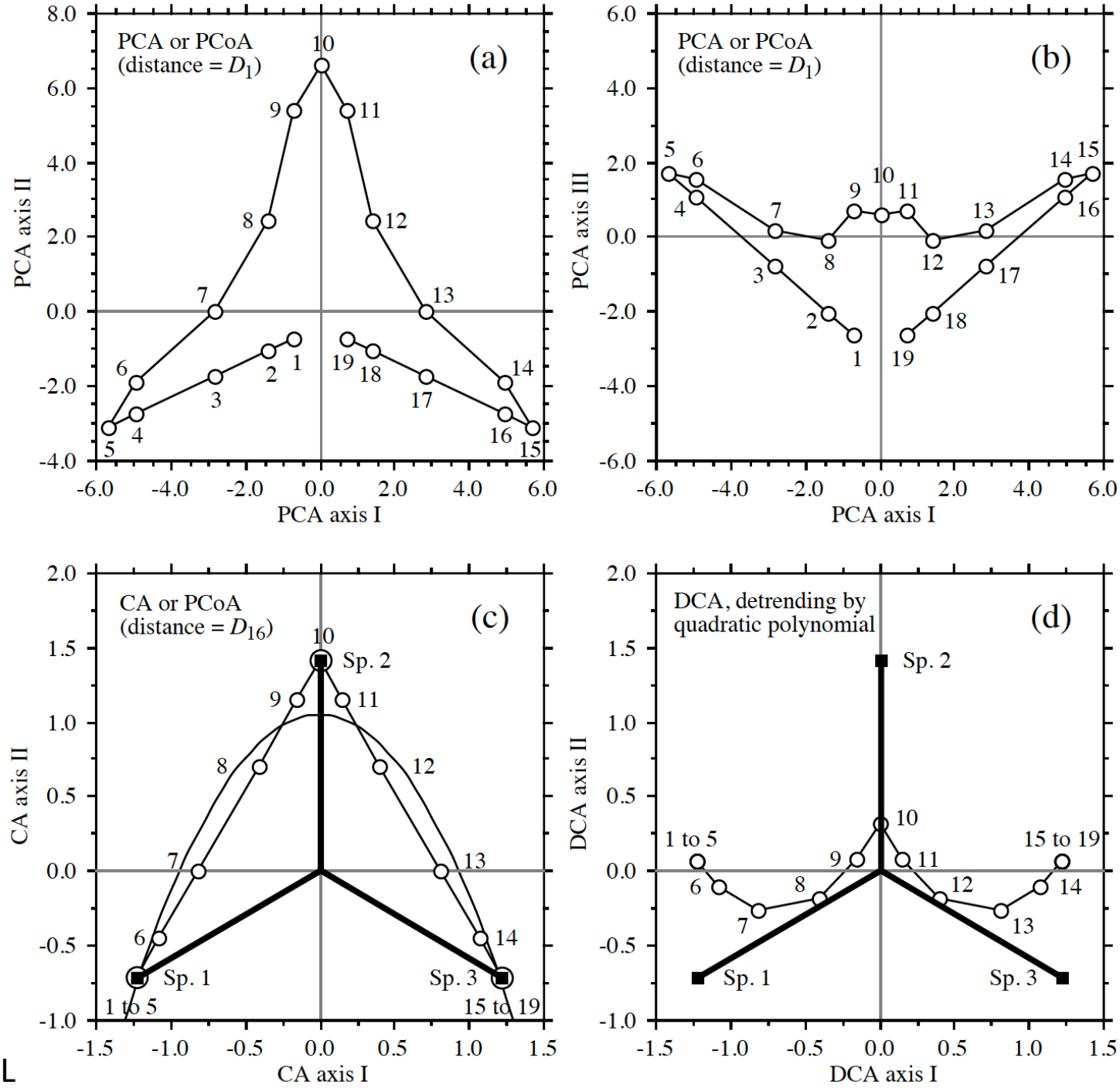




Detrended Correspondence Analysis (DCA)

- Arcos em diagramas de ordenação são indicativos de não-linearidade
- DCA destendencia o arco dividindo-o em segmentos (10-46)
- O comprimento do DCA1 é muito útil para determinar o comprimento do gradiente ambiental
  - Gradiente unimodal => 4 SP
  - Gradiente linear => 1-3 SD

p. 483-7 L&L


PCA or PCoA
(distance = D_1)
(a)
PCA axis II
PCA axis I
PCA or PCoA
(distance = D_1)
(b)
PCA axis III
PCA axis I
CA or PCoA
(distance = D_{16})
(c)
Sp. 1
Sp. 2
Sp. 3
1 to 5
15 to 19
CA axis II
CA axis I
DCA, detrending by
quadratic polynomial
(d)
Sp. 1
Sp. 2
Sp. 3
1 to 5
15 to 19
DCA axis II
DCA axis I

p. 484 L&L

Análise de Coordenadas Principais (PCoA) ou Escalonamento multidimensional *métrico*

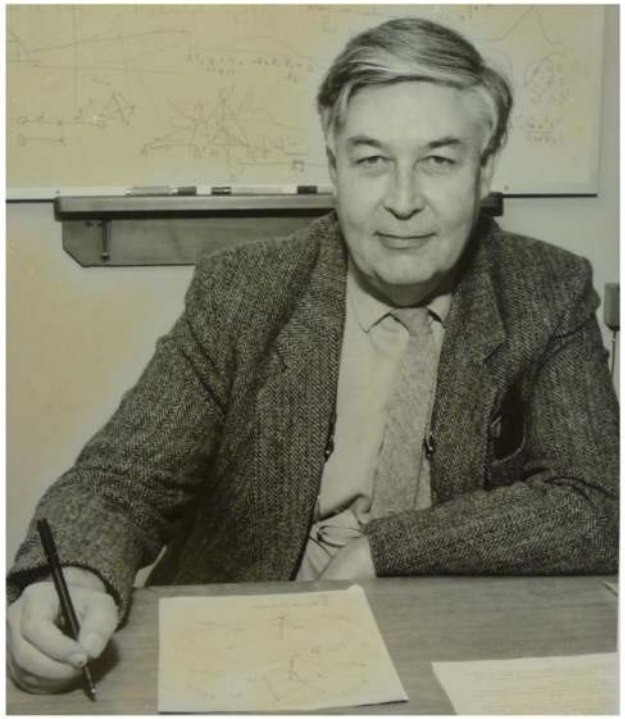


Figure 1. *John Gower, circa 1985.*

p. 492-512 L&L

Principal Coordinate Analysis
PCoA

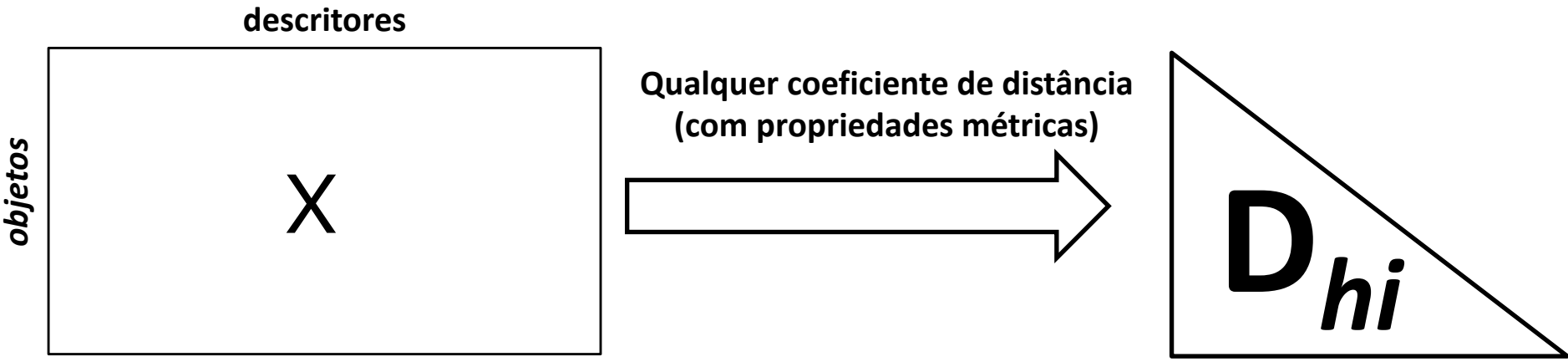
1. Reduz a dimensionalidade de dados multivariados

2. A **PCoA** se aplica a tabela de dados em que as linhas são os **indivíduos** e as colunas **variáveis (quant. ou qualitativas)**.

→ **Aceita qualquer tipo de dados**

→ **Binário, quant e quali**

Slide Thiago Gonçalves



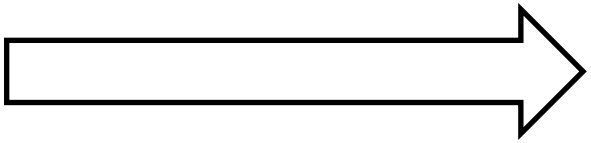
Slide Thiago Gonçalves

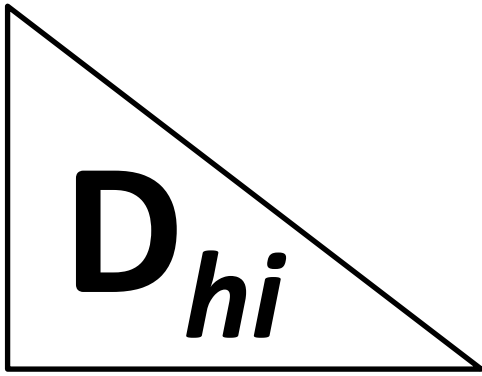
descritores

objetos

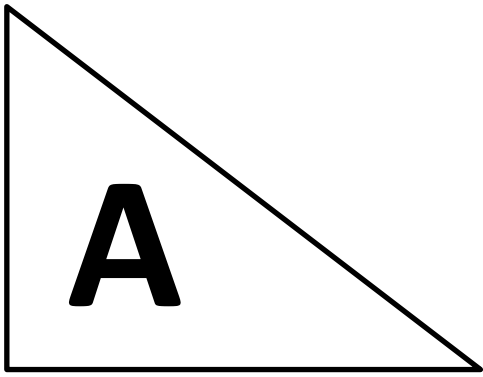
X

Qualquer coeficiente de distância (com propriedades métricas)





$$A_{hi} = -\frac{1}{2} D^2_{hi}$$



Slide Thiago Gonçalves

descritores

objetos

X

Qualquer coeficiente de distância (com propriedades métricas)

D_{hi}

$$A_{hi} = -\frac{1}{2} D^2_{hi}$$

	Eigenvalues			
	λ_1	λ_2	...	λ_c
Objects	**Eigenvectors**			
$\mathbf{x}_1$	u_{11}	u_{12}	...	u_{1c}
$\mathbf{x}_2$	u_{21}	u_{22}	...	u_{2c}
⋮	⋮			⋮
$\mathbf{x}_h$	u_{h1}	u_{h2}	...	u_{hc}
⋮	⋮			⋮
$\mathbf{x}_i$	u_{i1}	u_{i2}	...	u_{ic}
⋮	⋮			⋮
$\mathbf{x}_n$	u_{n1}	u_{n2}	...	u_{nc}
Lengths: $\sqrt{\sum_i u_{ik}^2}$ =	$\sqrt{\lambda_1}$	$\sqrt{\lambda_2}$	...	$\sqrt{\lambda_c}$
Centroid: $[\bar{u}_k]$ =	0	0	...	0

Produz mais que um autovalor = 0

Centralização

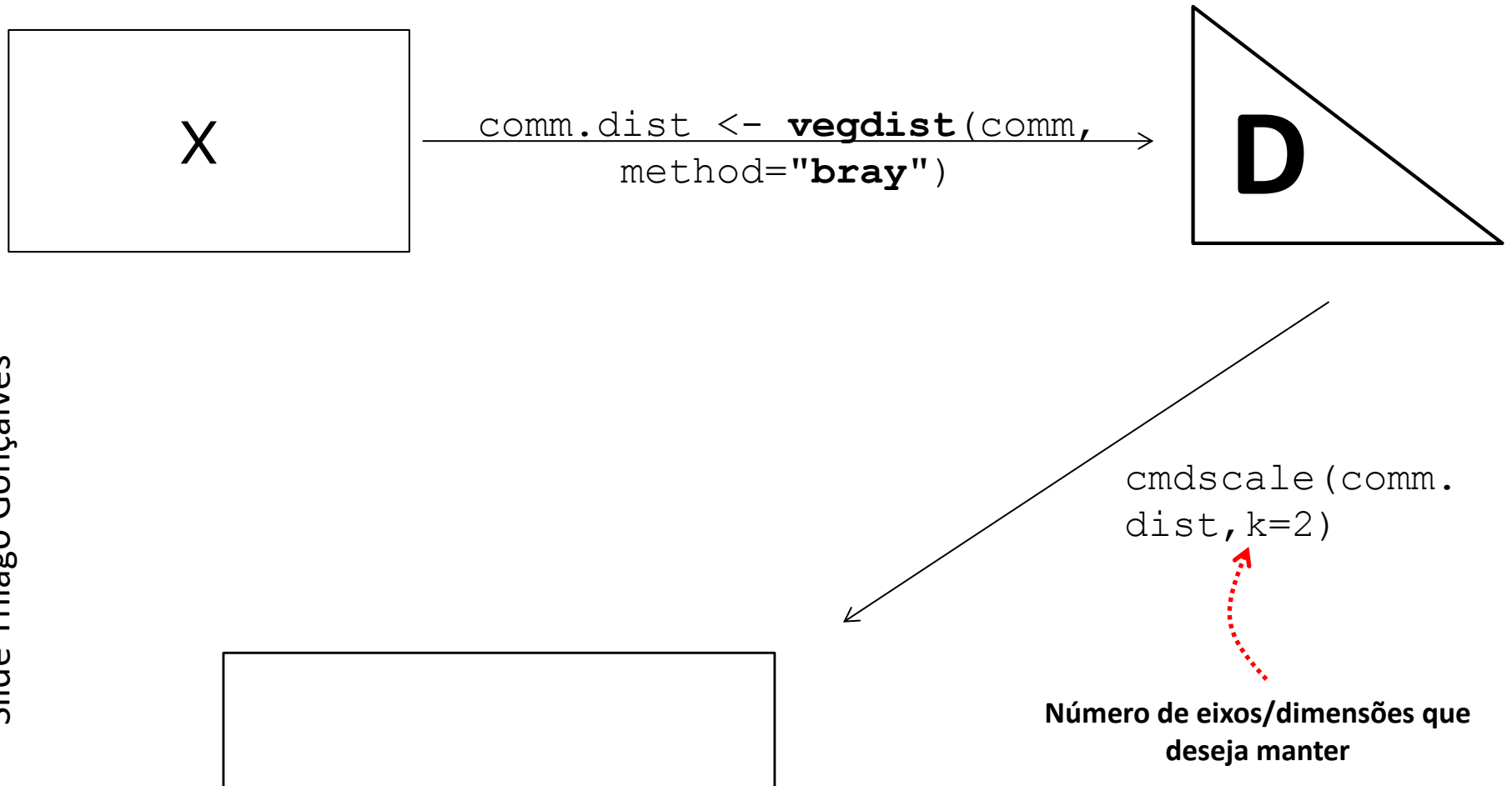
A

$$\delta_{hi} = a_{hi} - \bar{a}_h - \bar{a}_i + \bar{a}$$

	Eigenvalues			
	λ_1	λ_2	$\cdots$	λ_c
Objects	**Eigenvectors**			
$\mathbf{x}_1$	u_{11}	u_{12}	$\cdots$	u_{1c}
$\mathbf{x}_2$	u_{21}	u_{22}	$\cdots$	u_{2c}
$\vdots$	$\vdots$			$\vdots$
$\mathbf{x}_h$	u_{h1}	u_{h2}	$\cdots$	u_{hc}
$\vdots$	$\vdots$			$\vdots$
$\mathbf{x}_i$	u_{i1}	u_{i2}	$\cdots$	u_{ic}
$\vdots$	$\vdots$			$\vdots$
$\mathbf{x}_n$	u_{n1}	u_{n2}	$\cdots$	u_{nc}
Lengths: $\sqrt{\sum_i u_{ik}^2}$ =	$\sqrt{\lambda_1}$	$\sqrt{\lambda_2}$	$\cdots$	$\sqrt{\lambda_c}$
Centroid: $[\bar{u}_k]$ =	0	0	$\cdots$	0

Escalonamento dos autovetores e autovalores

Passo-a-passo no R



Slide Thiago Gonçalves

Propriedades da PCoA

- Também produz eixos em função das variáveis originais
  - Coordenadas principais são autovetores escalonados pela raiz quadrada do comprimento dos autovalores
  - Somente a parte Euclidiana dos coeficientes de distância é representada no biplot
- **Número de autovetores = número de variáveis – 1**
  - Devido aos procedimentos de centralização e representação no espaço Euclidiano, o que causa perda de informações (veja slides anteriores)

Propriedades da PCoA

- PCoA conduzida numa matriz de distância Euclidiana fornece os mesmos resultados que uma PCA
- Sinais dos objetos ao longo dos eixos podem inverter, já que o objetivo é mostrar a distância entre eles no espaço reduzido (posição relativa) e não a sua posição absoluta
- Qualidade da representação no espaço reduzido pode ser acessada com um diagrama de Shepard

Diagrama de Shepard

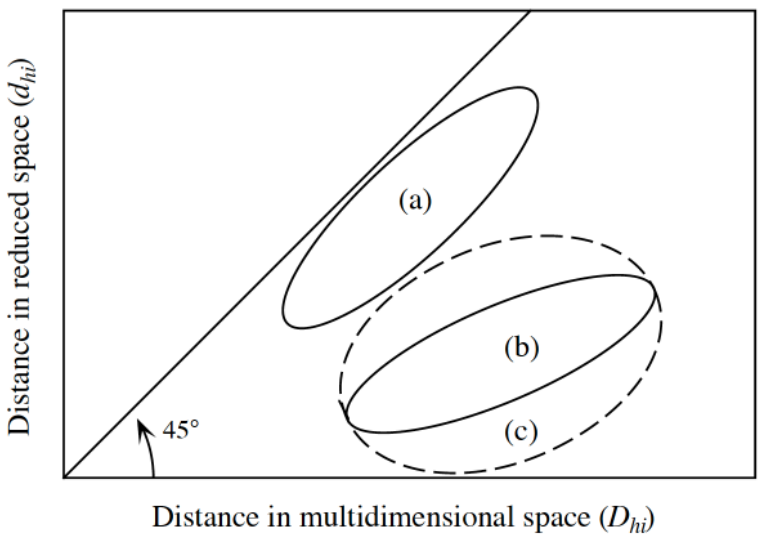
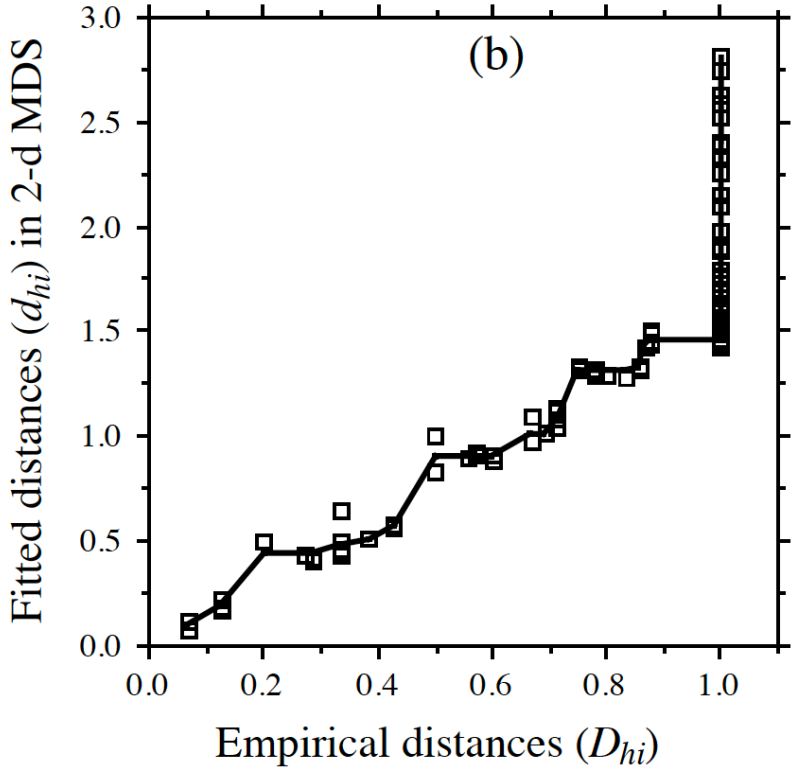


Figure 9.1 Shepard diagram. Three situations encountered when comparing distances among objects, in the p-dimensional space of the p original descriptors (abscissa) *versus* the d-dimensional reduced space (ordinate). The figure only shows the contours of the scatters of points. (a) The projection in reduced space accounts for a high fraction of the variance; the relative positions of objects in the d-dimensional reduced space are similar to those in the p-dimensional space. (b) The projection accounts for a small fraction of the variance, but the relative positions of the objects are similar in the two spaces. (c) Same as (b), but the relative positions of the objects differ in the two spaces. Adapted from Rohlf (1972). Compare to Fig. 8.24.



Quanto mais pontos em cima da linha melhor a representação gráfica no espaço reduzido

Propriedades da PCoA

- PCoA só pode representar as relações entre objetos (Q mode) ou descritores (R mode), mas não os dois ao mesmo tempo
  - Isso tem implicações para como o biplot é produzido
- Solução para isso é calcular os scores das variáveis depois de uma PCoA em Q mode usando correlações ou médias ponderadas
  - Implementado em ape::biplot.pcoa

Autovetores com autovalores negativos

- São produzidos quando um coeficiente não-métrico ou semi-métrico é usado

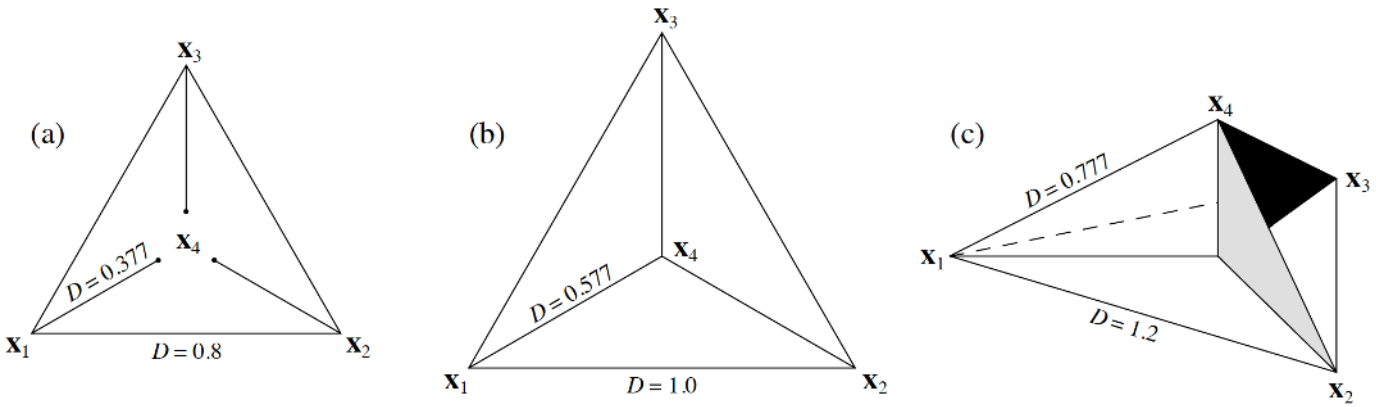


Figure 9.17 (a) Distances among four points constructed in such a way that the system cannot be represented in Euclidean space because the three lines going towards point $\mathbf{x}_4$ do not meet. (b) By adding a constant to all distances ($c_2 = 0.2$ in the present case), correction method 2 makes the system Euclidean; in this example, the distances can be associated with a representation of the points in two-dimensional space. (c) When increasing the distances further (adding again 0.2 to each distance in the present case), the system remains Euclidean but requires more dimensions for representation (three dimensions in this example).

Correções para autovalores negativos

- Duas correções
  - Lingoes

$$\hat{D}_{hi} = \sqrt{D_{hi}^2 + 2c_1} \qquad \text{for } h \neq i$$

  - Cailliez

$$\hat{D}_{hi} = D_{hi} + c_2 \qquad \text{for } h \neq i$$

- Transformam distâncias pequenas em grandes

Non-metric Multidimensional Scaling (nMDS)

William Kruskal

p. 512-9 L&L

O que é um nMDS?

- Procedimento iterativo de ordenação que representa objetos no espaço reduzido, cujo número de eixos é escolhido *a priori* pelo usuário
  - Geralmente 2 ou 3 dimensões
- Também utiliza matrizes de distâncias construídas com qualquer coeficiente, métrico ou não-métrico
- Transforma as distâncias originais em ranques
- Os ranques de distância Euclidiana entre pontos na ordenação têm uma relação monotônica não-métrica em relação a qualquer dissimilaridade original
- A função de transferência das dissimilaridades observadas para as distâncias na ordenação é não-métrica, mas a configuração do resultado da ordenação é métrico

O que é um nMDS?

- nMDS comprime as distâncias entre objetos em 2 ou 3D de forma mais eficiente do que PCoA
- Sempre obtém uma representação Euclidiana, mesmo para distâncias não-Euclidiana
- Mas faz isso de forma não-linear
- Requer mais tempo de computação do que PCoA, portanto não é recomendado quando se tem grandes conjuntos de dados

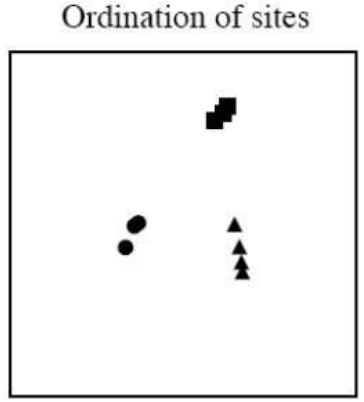
descritores

objetos

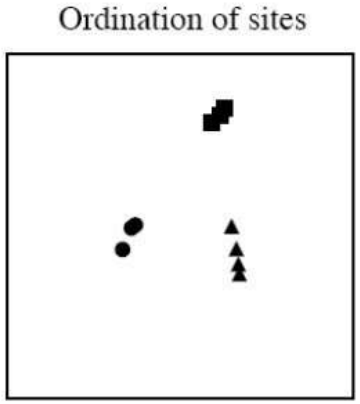
Qualquer coeficiente de distância
(métrico ou semi-métrico)

D_{hi}

Especifique o número de
dimensões desejado
p.ex. k=2



Cálcule STRESS
usando uma das 3
fórmulas



Modifique o arranjo
dos objetos e construa um novo
espaço reduzido

Repita os dois
últimos passos
até encontrar a solução
que minimize
o STRESS

Como o nMDS funciona

- Não é um método baseado em autoanálise de uma matriz
  - Eixos são arbritários
- Portanto, não podemos usar posteriormente os eixos/autovetores em análises
  - Muito bom para visualização de dados
  - Permite missing data
- Cálculo do STRESS
  - STandard REsiduals Sum of Squares
  - 3 tipos (fórmulas)
  - Mede a qualidade da representação (diagrama), distorção das distâncias originais

McCune & Grace 2002

Table 16.4. Rules of thumb for interpreting final stress from nonmetric multidimensional scaling, using Kruskal's stress formula 1 multiplied by 100.

Stress	
	Kruskal's rules of thumb
2.5	Excellent
5	Good
10	Fair
20	Poor
	Clarke's rules of thumb
< 5	An excellent representation with no prospect of misinterpretation. This is, however, rarely achieved.
5-10	A good ordination with no real risk of drawing false inferences
10-20	Can still correspond to a usable picture, although values at the upper end suggest a potential to mislead. Too much reliance should not be placed on the details of the plot.
> 20	Likely to yield a plot that is relatively dangerous to interpret. By the time stress is 35-40 the samples are placed essentially at random, with little relation to the original ranked distances.

Quanto mais baixo melhor!

Mas quão baixo é baixo o suficiente?

Algumas regras de dedão para ajudar a definir

nMDS

- Vantagens
  - Não assume que variáveis estão linearmente relacionadas
  - O uso de ranks de distância tende a linearizar relações entre distâncias de espécies e variáveis.
  - Tende a diminuir o problema do zero-truncation
  - Permite que se use qualquer coeficiente de distância

nMDS

- Desvantagens
  - Falha em encontrar a melhor solução (arranjo dos objetos no espaço reduzido que distorça o mínimo possível as distâncias originais na matriz)
    - vegan::metaMDS resolve isso calculando várias vezes o nMDS começando de pontos aleatórios diferentes
  - Cálculo lento pra conjuntos de dados grandes

Diferenças entre PCoA e nMDS

	nMDS	PCoA
Solução	Algoritmo iterativo de aproximação	Autoanálise da matriz de distância
Estabilidade da solução	Pode variar a cada vez que o algoritmo é usado, dependendo do ponto de início (aleatório)	Única. Pode ser usada como "semente" para outras técnicas
Tratamento das dissimilaridades	Distorcidas durante o cálculo, fazendo com que não reflitam as distâncias originais	Não são distorcidas
Construção da solução final ótima	Solução pode depender do critério (tipo de STRESS) usado; impossível saber a priori qual usar	Os primeiros eixos maximizam a variância das observações
Qualidade do espaço reduzido	Difícil escolher os parâmetros a priori. Solução é contingente ao no. de eixos escolhidos e muda se no. ≠ for escolhido	Permite encontrar todos os eixos correspondendo a uma matriz de dissimilaridade, permitindo a reconstrução exata das distâncias originais

p. 194-5 Borcard et al. 2018

Diferenças entre PCoA e nMDS

	nMDS	PCoA
Medida de ajuste	STRESS não indica a proporção de variância representada no diagrama de ordenação	O pseudo-R^2 (soma dos autovalores dos primeiros eixos/soma de todos os autovalores) indica a fração da variância dos dados representada na ordenação. Útil como medida da qualidade da ordenação (espaço reduzido).

nMDS é útil, na maioria das vezes, somente para representar graficamente distâncias entre locais em poucas dimensões

p. 194-5 Borcard et al. 2018


That's all Folks!
For today