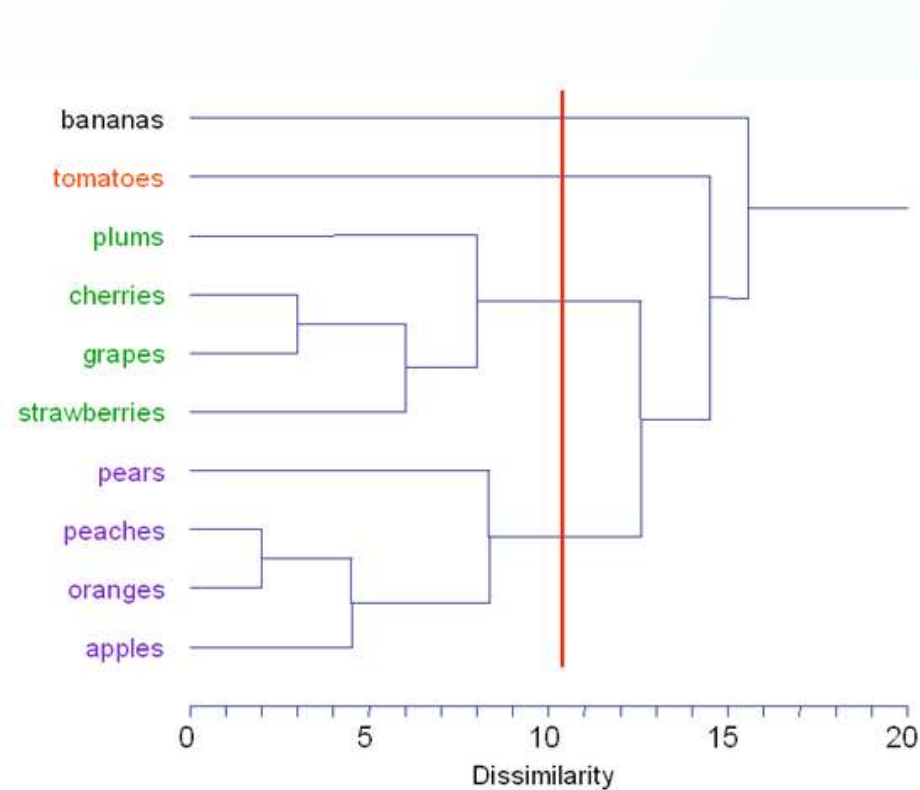
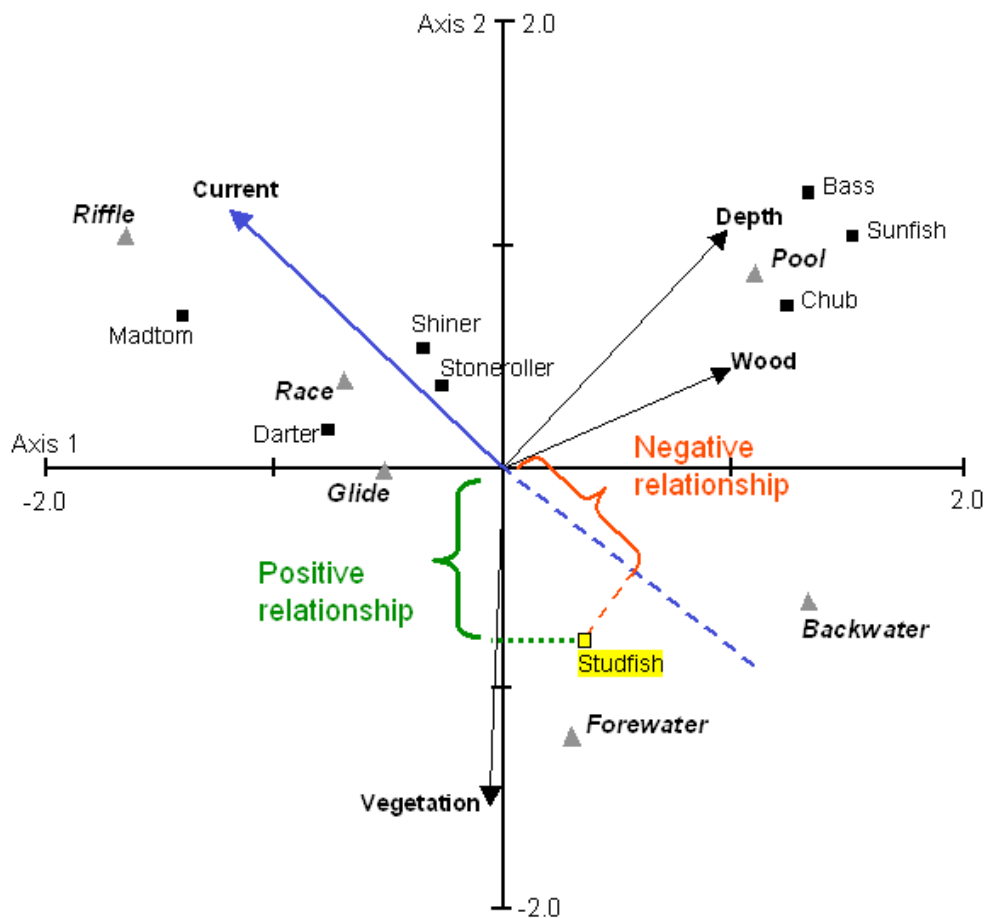
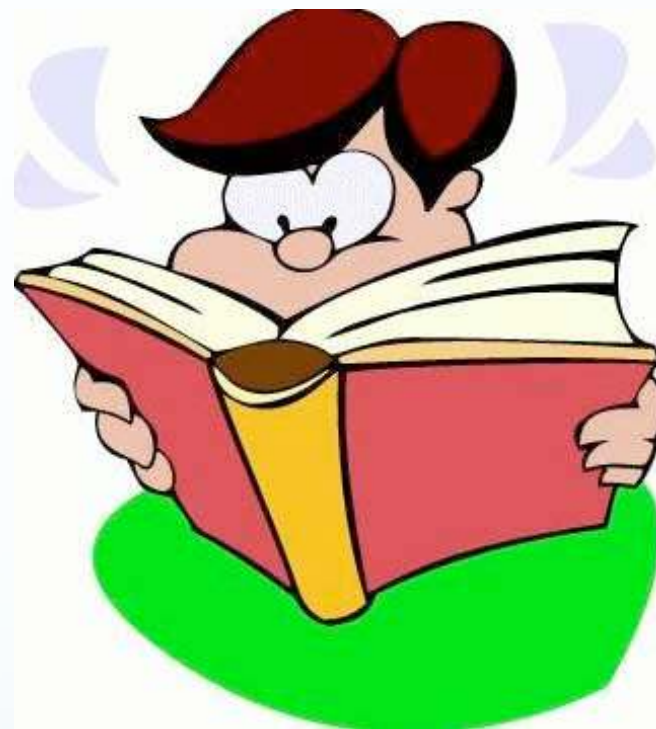


INTRODUÇÃO À ESTATÍSTICA MULTIVARIADA





Informação

vs.

Formação

- **Objetivos Principais**

- Principal direção de variação dos dados, correlação de matrizes, diferenças entre grupos

- **Análises Exploratórias vs. Testes**

- Hipóteses (expectativas)

MULTIVARIATE ANALYSIS IN ECOLOGY AND SYSTEMATICS: PANACEA OR PANDORA'S BOX?

Frances C. James

Department of Biological Science, Florida State University, Tallahassee, Florida
32306

Charles E. McCulloch

Biometrics Unit, Cornell University, Ithaca, New York 14853

Numerical Ecology

SECOND ENGLISH EDITION

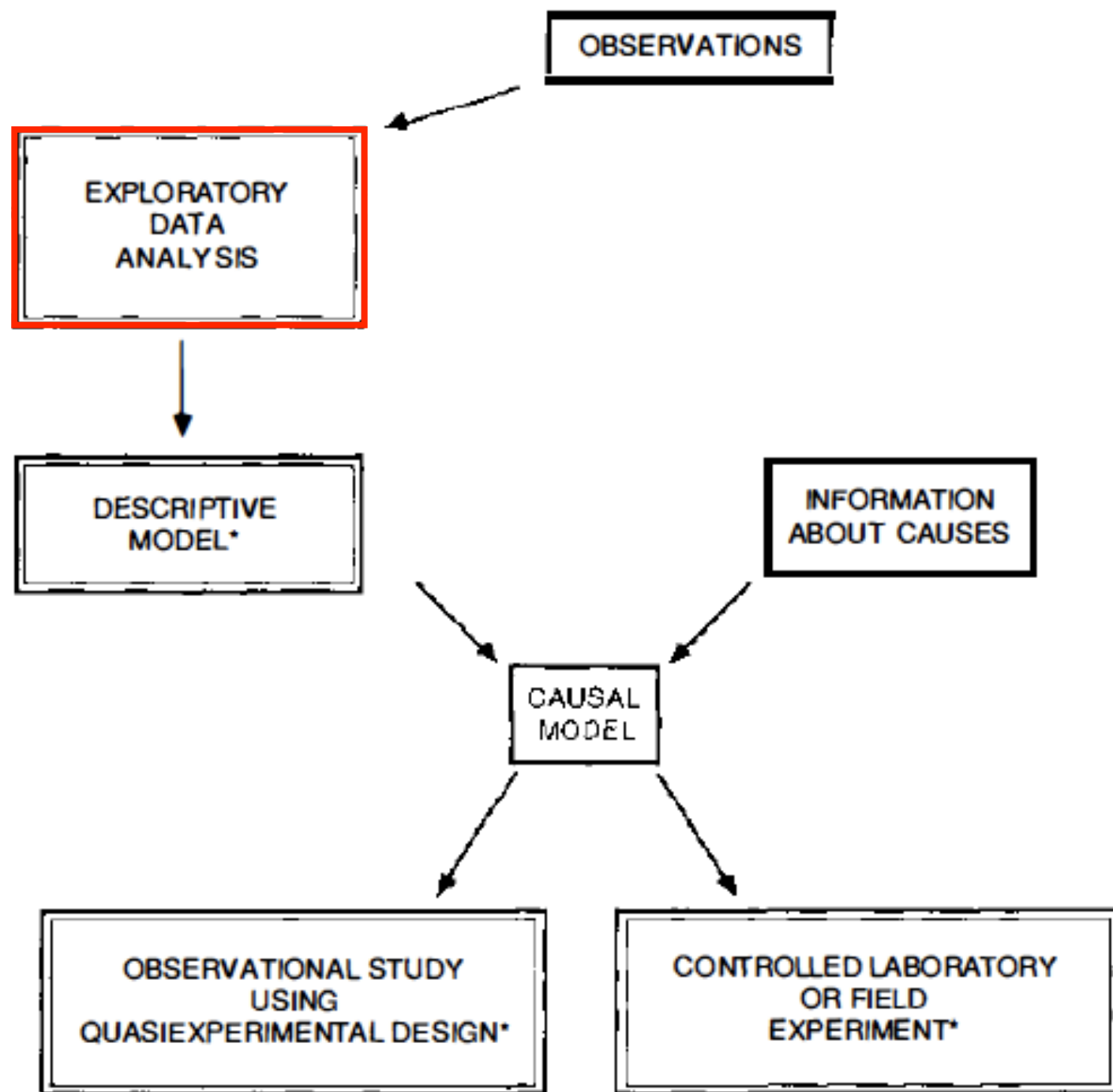
Pierre LEGENDRE, Professor

*Département de sciences biologiques, Université de Montréal,
Montréal (Québec) H3C 3J7, Canada*

and

Louis LEGENDRE, Professor

*Département de biologie, Université Laval,
Québec (Québec) G1K 7P4, Canada*

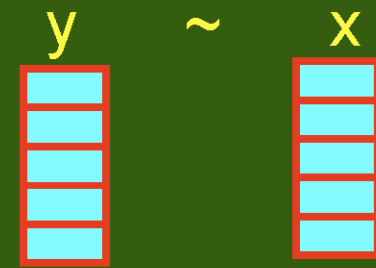


- Organização de dados
 - Variáveis na coluna e unidades amostrais nas linhas
- Tipos de testes

Análises Multivariadas

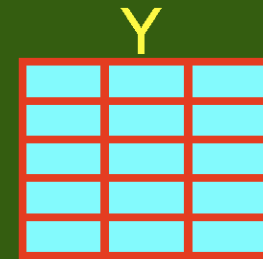
Univariadas $\rightarrow y \sim x$

Regressão, Anova, Reg. Logística



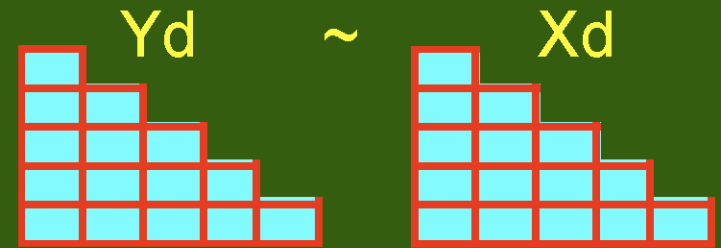
Multivariadas exploratórias $\rightarrow Y$

Ordenação e Classificação



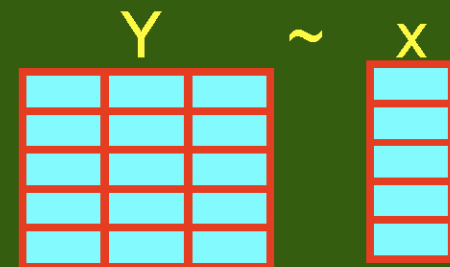
Multivariadas (testes) $\rightarrow Y_d \sim X_d$

Teste de Mantel



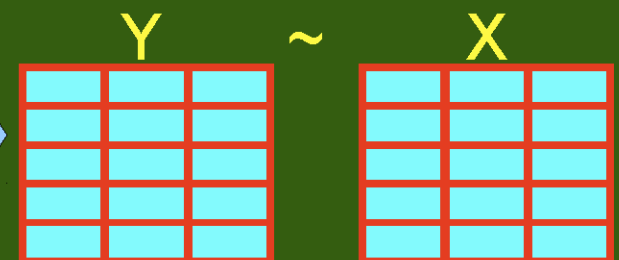
Multivariadas (testes) $\rightarrow Y \sim x$

Manova, db-Manova



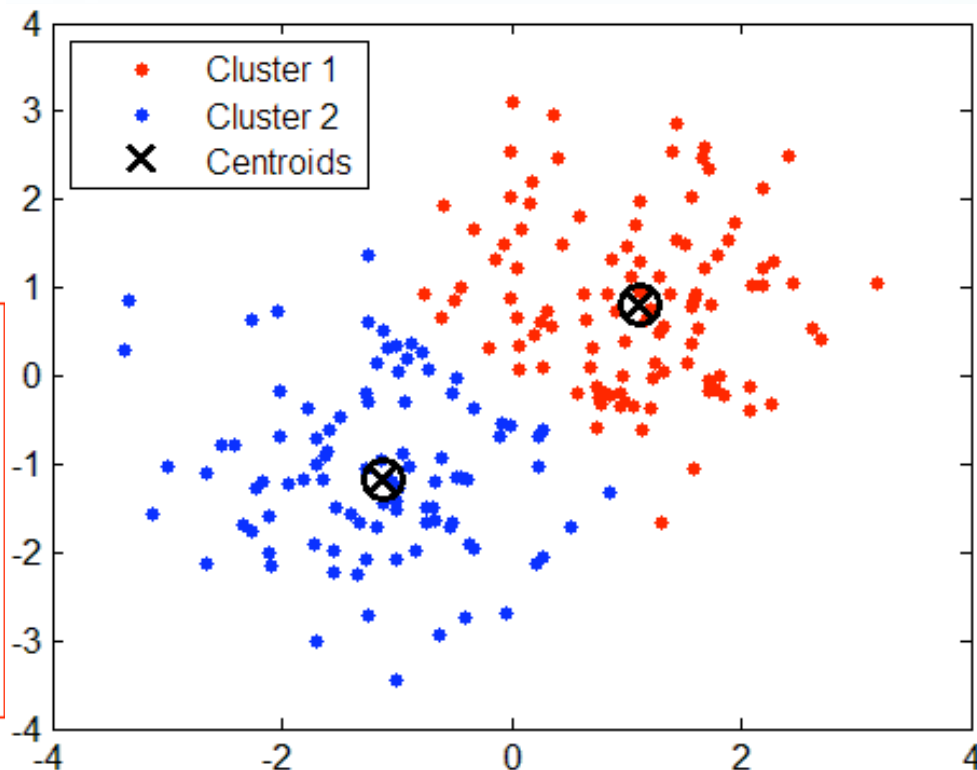
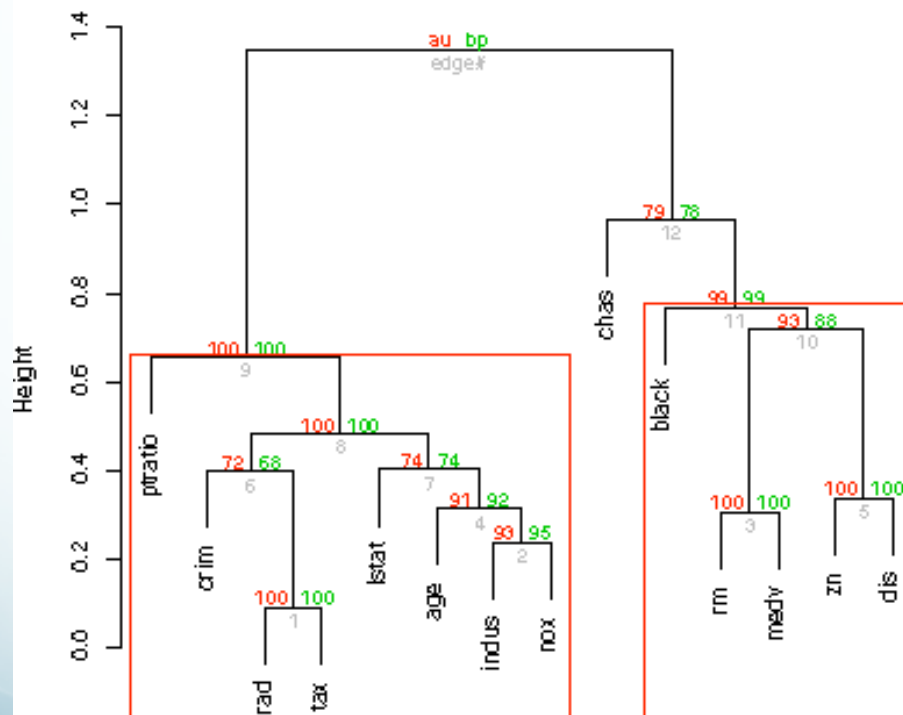
Multivariada (ordenações restritas) $\rightarrow Y \sim X$

CCA, RDA, CapScale



Classificação

Cluster dendrogram with AU/BP values (%)

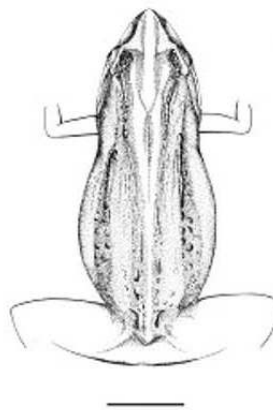
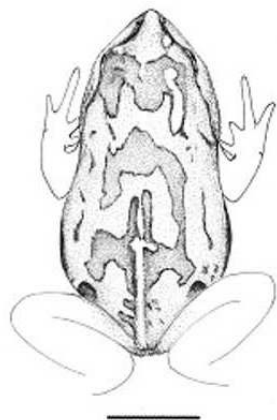
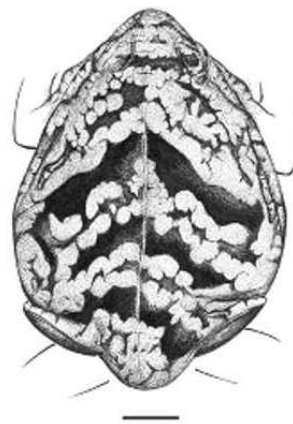
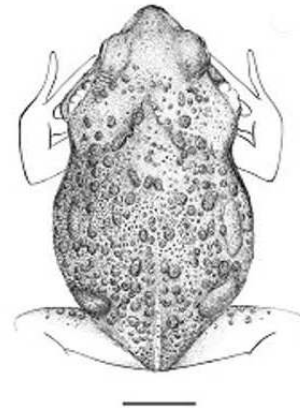
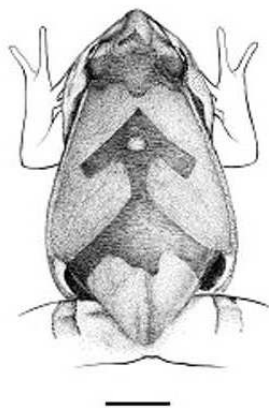


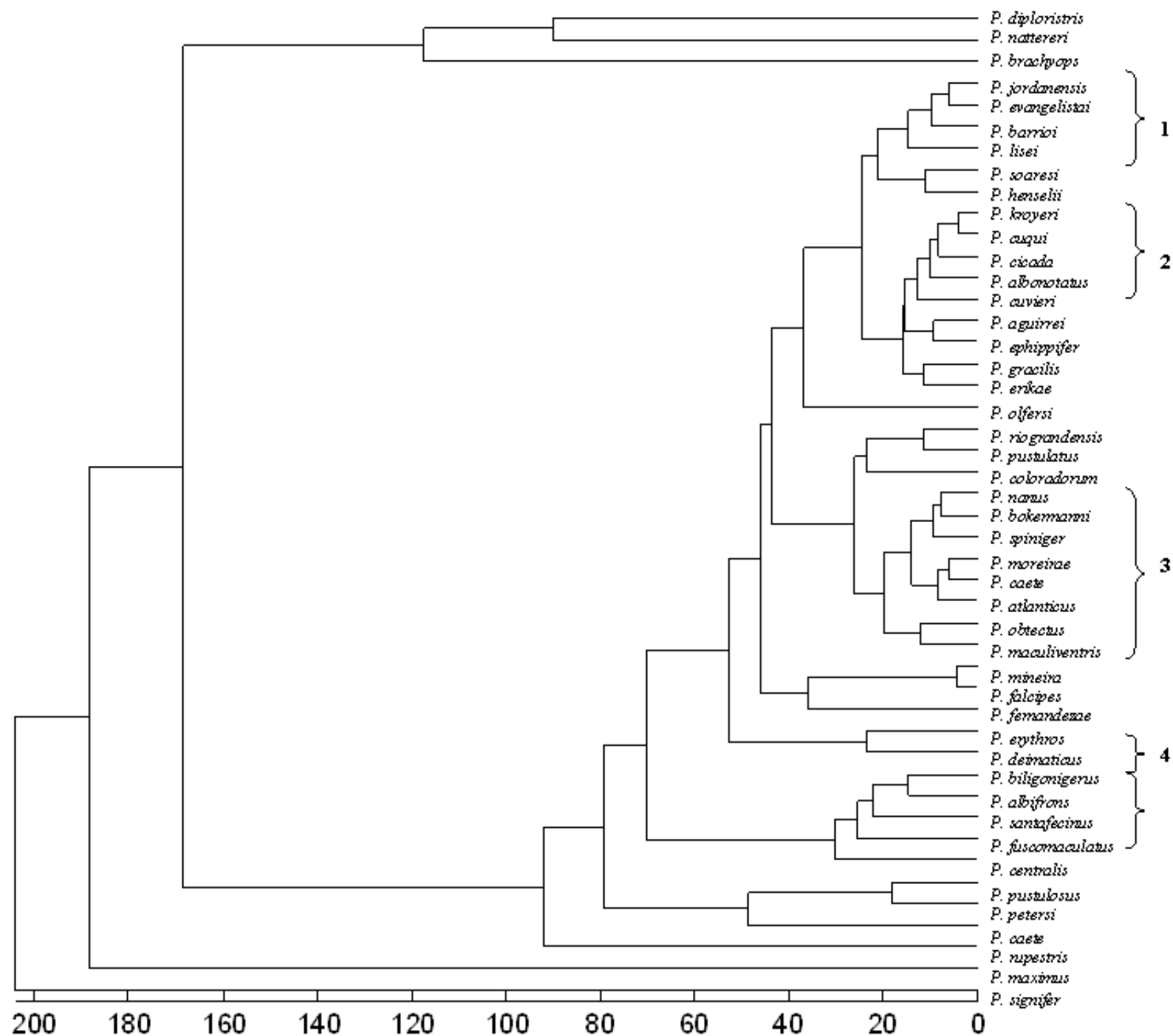
Agrupamento

- Análise de Agrupamento Hierárquico
- Análise de Espécie Indicadora (IndVal)

Análise de Agrupamento

- Resumir um grande volume de informação
- Agrupa objetos por grau de similaridade
- Busca por descontinuidades no conjunto de dados para formar grupos
- Descontinuidade X Gradiente
- Usos em Sistemática => Escola Fenética





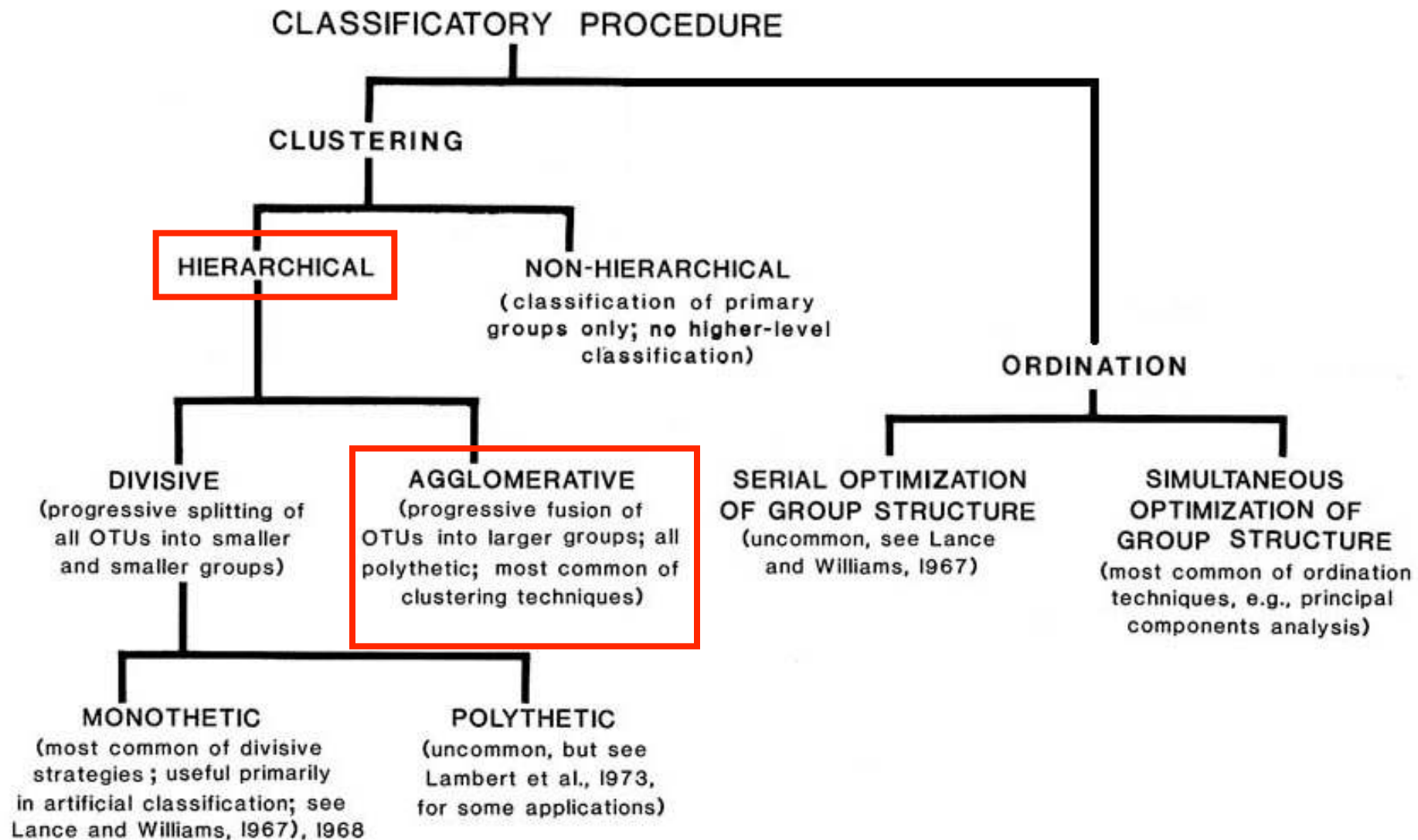


FIGURE 7.2 Outline of major phenetic classificatory algorithms. (Based on Williams 1971:306; Clifford and Stephenson 1975:29; and R. Sokal *in litt.*)

Exemplo da lógica do agrupamento hierárquico com *single linkage*

- **Métodos de agrupamento**

- UPGMA e Distância mínima de Ward

- **Índices de Similaridade/Distância**

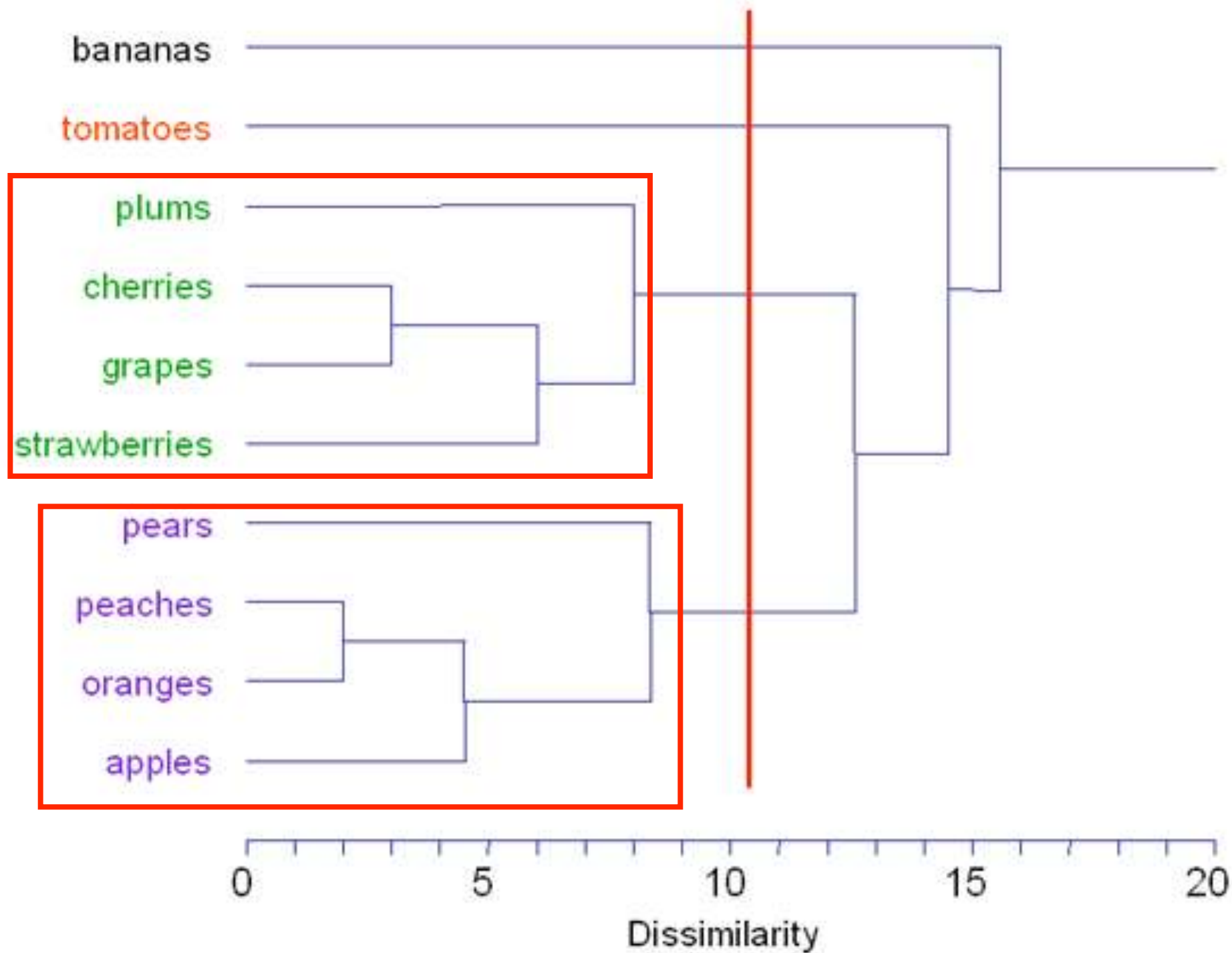
- Índices binários assimétricos: Jaccard, Sørensen
- Índices quantitativos assimétricos: Bray-Curtis, Morisita-Horn, Gower2
- Índices métricos: Distância Euclidiana, Manhattan, Canberra

Representatividade do dendrograma

- Coeficiente de Correlação Cofenética
- Correlação de Pearson entre a matriz de similaridade e a matriz cofenética.
- Regra do 0,8 .

Definindo os grupos: Nível de corte

- Vários critérios
 - Bootstrap dos nós,
 - 50% similaridade,
 - inspeção visual etc.
 - Pacote *pvclust*



The Annals of Statistics

2004, Vol. 32, No. 6, 2616–2641

DOI 10.1214/009053604000000823

© Institute of Mathematical Statistics, 2004

APPROXIMATELY UNBIASED TESTS OF REGIONS USING MULTISTEP-MULTISCALE BOOTSTRAP RESAMPLING¹

BY HIDETOSHI SHIMODAIRA

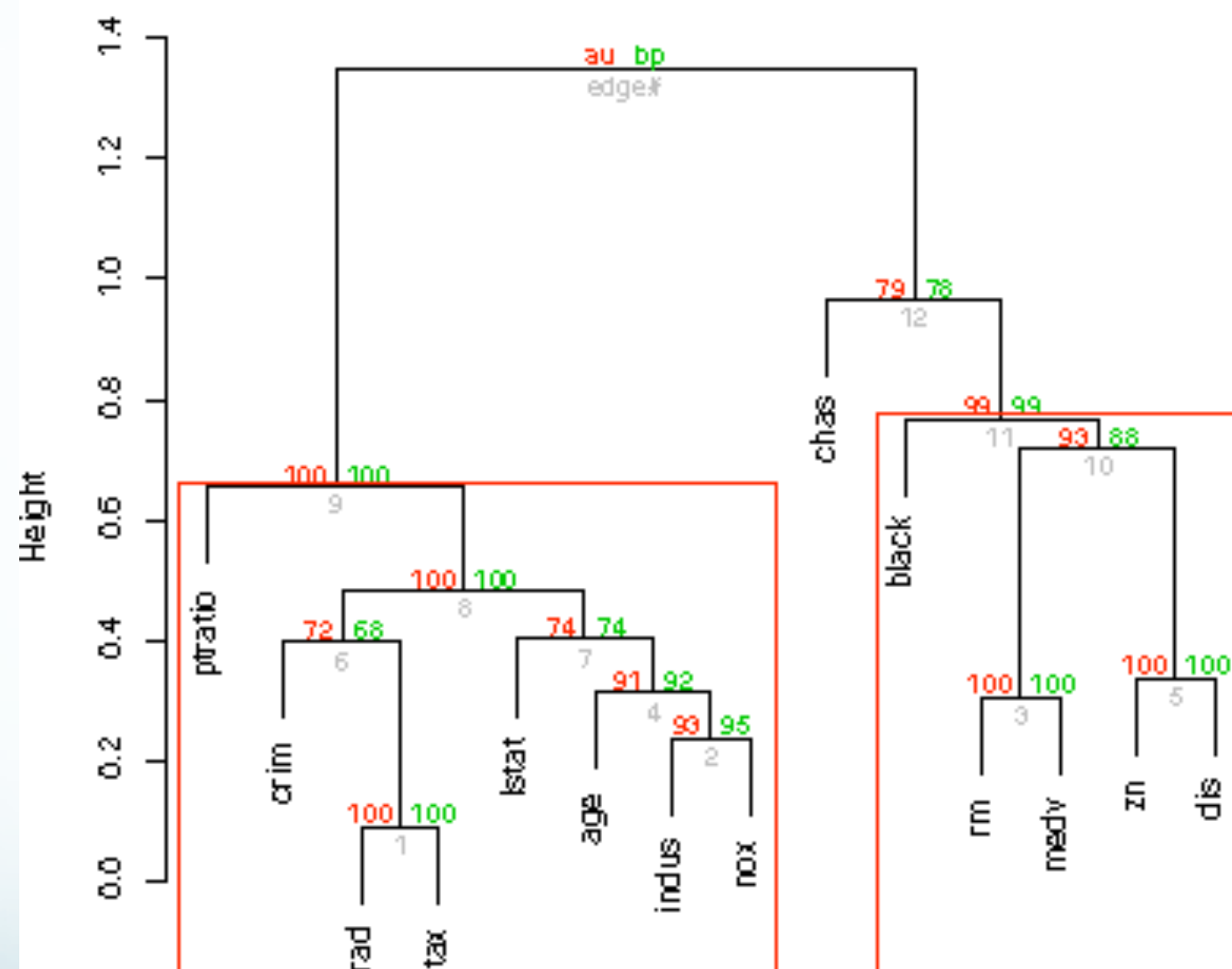
PVCLUST

- Reamostragem utilizando bootstrap em multiescala.
 - Utiliza 2 estimadores para calcular o valor de p: **AU** e **BP**
 - Verifica se os agrupamentos (clusters) se manteriam se o tamanho amostral fosse aumentado.
 - função pvrect circula grupos com **AU** maiores que 95%

```
> result <- pvclust(lung, method.dist="cor", method.hclust="average", nboot=1000)
Bootstrap (r = 0.5)... Done.
Bootstrap (r = 0.6)... Done.
Bootstrap (r = 0.7)... Done.
Bootstrap (r = 0.8)... Done.
Bootstrap (r = 0.9)... Done.
Bootstrap (r = 1.0)... Done.
Bootstrap (r = 1.1)... Done.
Bootstrap (r = 1.2)... Done.
Bootstrap (r = 1.3)... Done.
Bootstrap (r = 1.4)... Done.
```

?pvclust

Cluster dendrogram with AU/BP values (%)



Exemplo e Exercícios

Análise de espécie indicadora (IndVal)

Ecological Monographs, 67(3), 1997, pp. 345–366
© 1997 by the Ecological Society of America

SPECIES ASSEMBLAGES AND INDICATOR SPECIES: THE NEED FOR A FLEXIBLE ASYMMETRICAL APPROACH

MARC DUFRÊNE¹ AND PIERRE LEGENDRE²

IndVal

- Identificar espécies indicadoras de grupos pré-estabelecidos
- Conceitos
 - Especificidade => só em 1 grupo
 - Fidelidade => todas as u.a. do grupo
 - Espécie indicadora ótima => ocorre em todas as u.a. da 1 grupo e somente naquele grupo
 - Indicadora assimétrica (rara): contribuem para especificidade do hábitat mas não para a formação de grupos

IndVal

- Circularidade da análise original
 - Utilização de uma informação ambiental

Procedimentos da Análise

- 1) Matriz de distância e classificação dos agrupamentos
- 2) Classificação dos grupos baseada em variáveis ambientais
- 3) Cálculo da especificidade e fidelidade
- 4) Comparação dos resultados com os dados originais (indicadoras assimétricas e simétricas)

Interpretação dos Resultados

X12	1	0.7500	0.028
X19	1	0.7500	0.029
X14	1	0.4455	0.038
X21	2	0.9706	0.004
X5	4	0.9556	0.003

Sum of probabilities = 10.111

Sum of Indicator Values = 11.66

Sum of Significant Indicator Values = 6.64

Number of Significant Indicators = 8

Significant Indicator Distribution

1 2 4

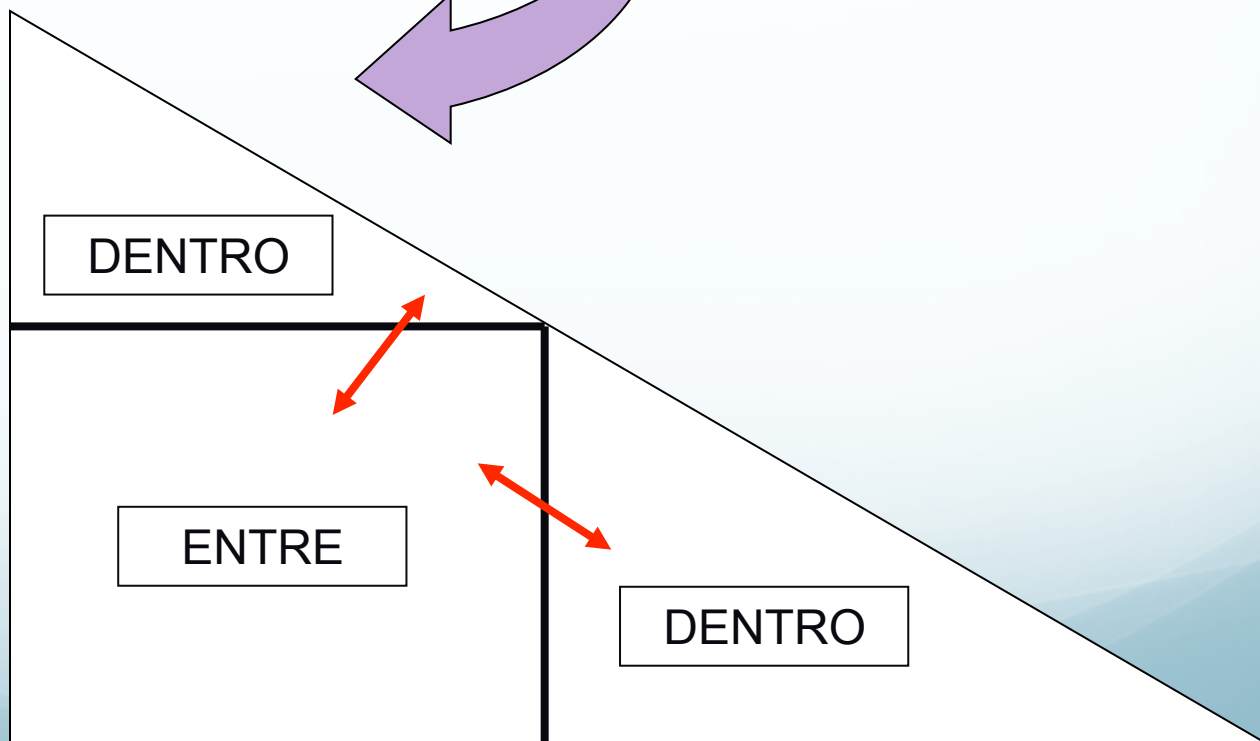
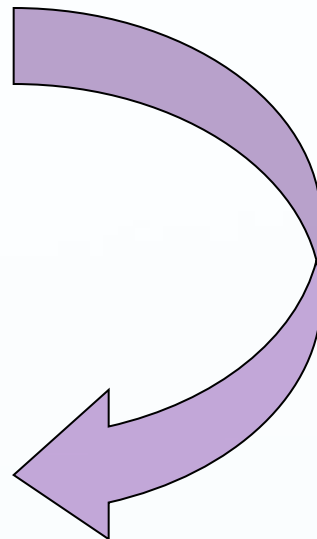
Exemplo no R

Comparação de médias entre grupos

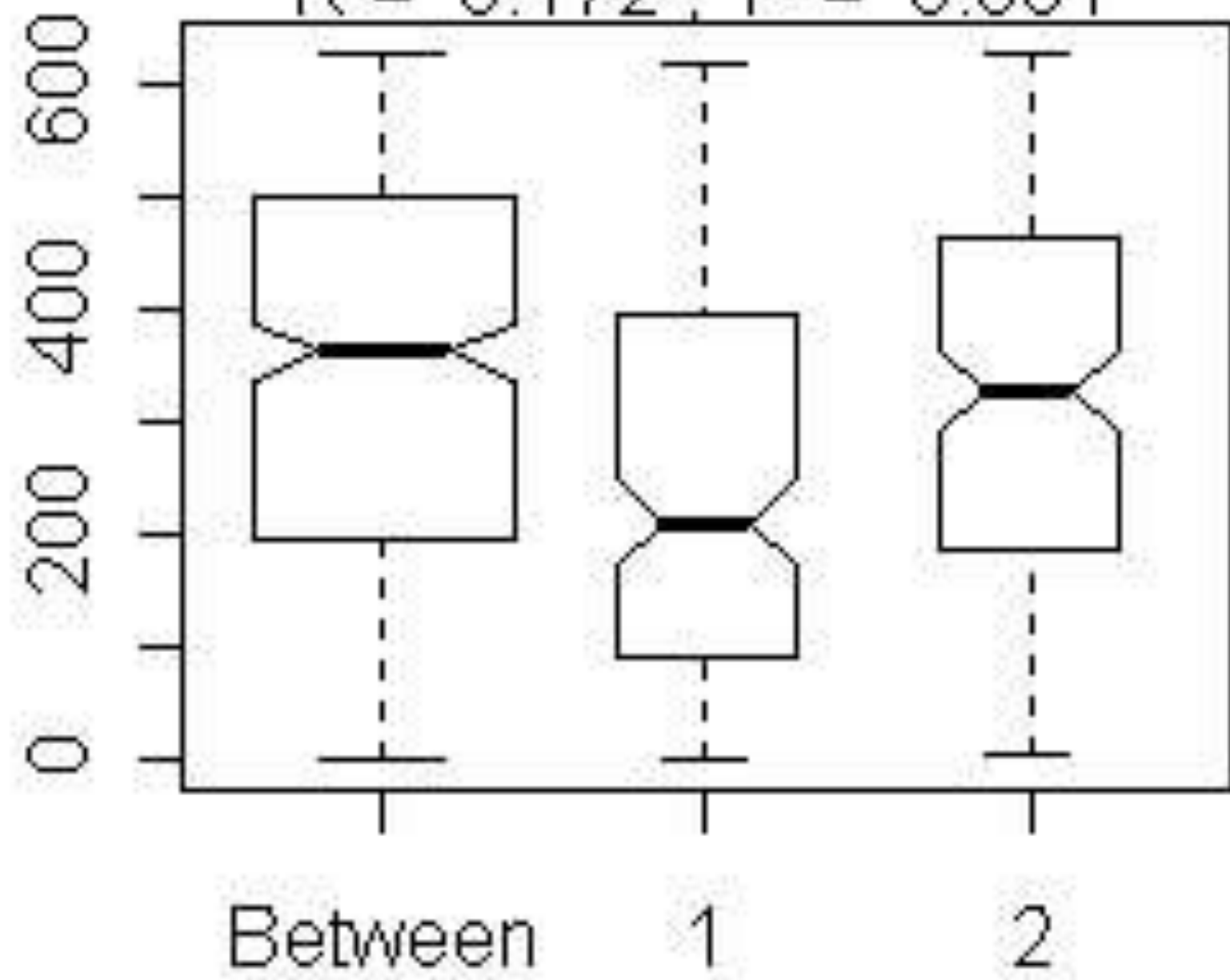
ANOSIM

- Tipo de MANOVA
- Compara se a similaridade é menor *dentro* que *entre* grupos
- Ranqueia as similaridades: maior similaridade = 1
- Estatística R (-1 a 1), significância avaliada por permutação
- Recomendado para 2 grupos

1.	ANN	DEAL (5)	NELDA (20)	JUDY(12)
2.	BUD	JUSTIN (15)	NORRIS (19)	JOHN (10)
3.	DIANE	DONNA (6)	LAURA (16)	JUDY (12)
4.	DORIS	JUDY (12)	LAURA (16)	JUNE (13)
5.	DEAL	JERRI (14)	ANN (1)	JUNE (13)
6	DONNA	DIANE (3)	NELDA (20)	JUNE (13)
7.	DICK	SOL (23)	JIM (11)	BUD (2)
8.	GARY	JUNE (13)	DORIS (4)	NORRIS (19)
9.	HARRY	JIM (11)	DICK (7)	BUD (2)
10.	JOHN	JUSTIN (15)	BUD (2)	NORRIS (19)
11.	JIM	MIKE (17)	JUSTIN (15)	BUD (2)
12.	JUDY	NELDA (20)	LAURA (16)	ANN (1)
13.	JUNE	DORIS (4)	LAURA (16)	JUDY (12)
14.	JERRI	DEAL (5)	NELDA (20)	DICK (7)
15.	JUSTIN	SOL (23)	DICK (7)	NORRIS (19)
16.	LAURA	JUDY (12)	NELDA (20)	DORIS (4)
17.	MIKE	JIM (11)	SOL (23)	NORRIS (19)
18.	MILLARD	SAM (22)	VICTOR (24)	JIM (11)
19.	NORRIS	JUSTIN (15)	NELDA (20)	JIM (11)
20.	NELDA	JUDY (12)	NORRIS (19)	LAURA (16)
21.	PRUDENCE	NELDA (20)	DORIS (4)	JUDY (12)
22.	SAM	VICTOR (24)	MIKE (17)	JIM (11)
23.	SOL	NORRIS (19)	DICK (7)	JIM (11)
24.	VICTOR	SAM (22)	MIKE (17)	JIM (11)



$R = 0.172$, $P = 0.001$



MRPP

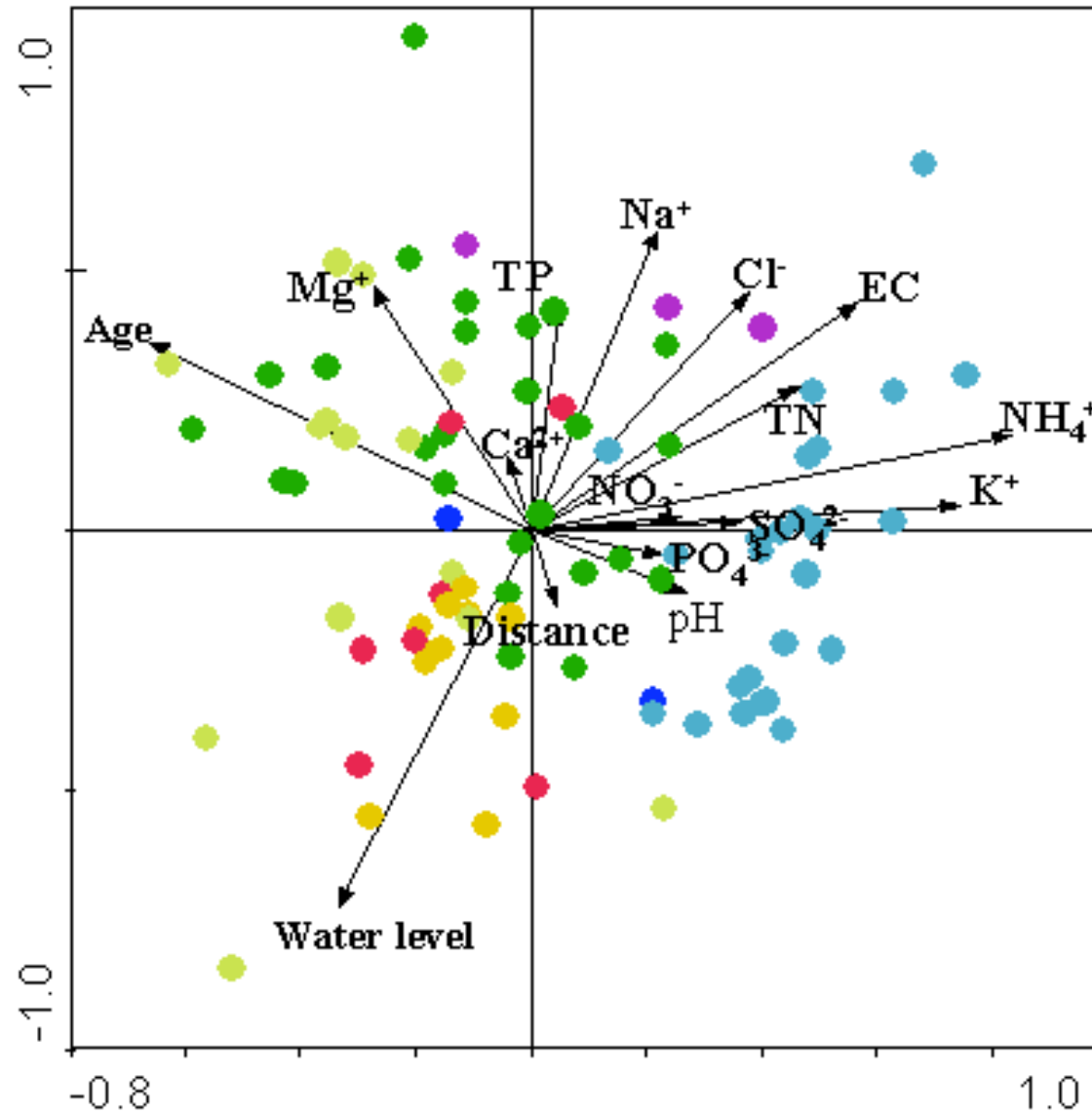
- Lógica semelhante ao ANOSIM
- Recomendado para mais de 2 grupos
- Estatística δ representa a média ponderada geral das médias dentro dos grupos dos pares de dissimilaridade.

IMPORTANTE!!!

- A diferença que o δ mede pode ser de localização (diferenças na média) ou de dispersão (diferenças nas distâncias dentro dos grupos). Ou seja, o teste pode encontrar uma diferença significativa entre 2 grupos simplesmente porque um destes grupos tem uma maior dissimilaridade entre as u.a.

Exemplo no R e Exercícios

Ordenação



Introdução

- Concebidas para análises psicométricas e sociológicas.
- Introduzidas na ecologia inicialmente na classificação de comunidades vegetais.

GRADIENT ANALYSIS OF VEGETATION*

By

R. H. WHITTAKER

Não se perca na sopa de
letrinhas

Ordenação irrestrita

- PCA = Análise de Componentes Principais
- PCoA= Análise de Coordenadas Principais
- nMDS = Escalonamento multidimensional não-métrico

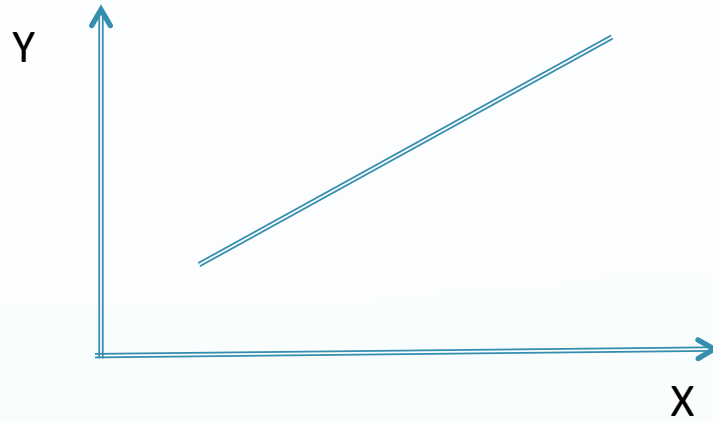
Ordenação irrestrita

- Como os objetos se distribuem ao longo de gradientes => situação mais comum na ecologia

PCA

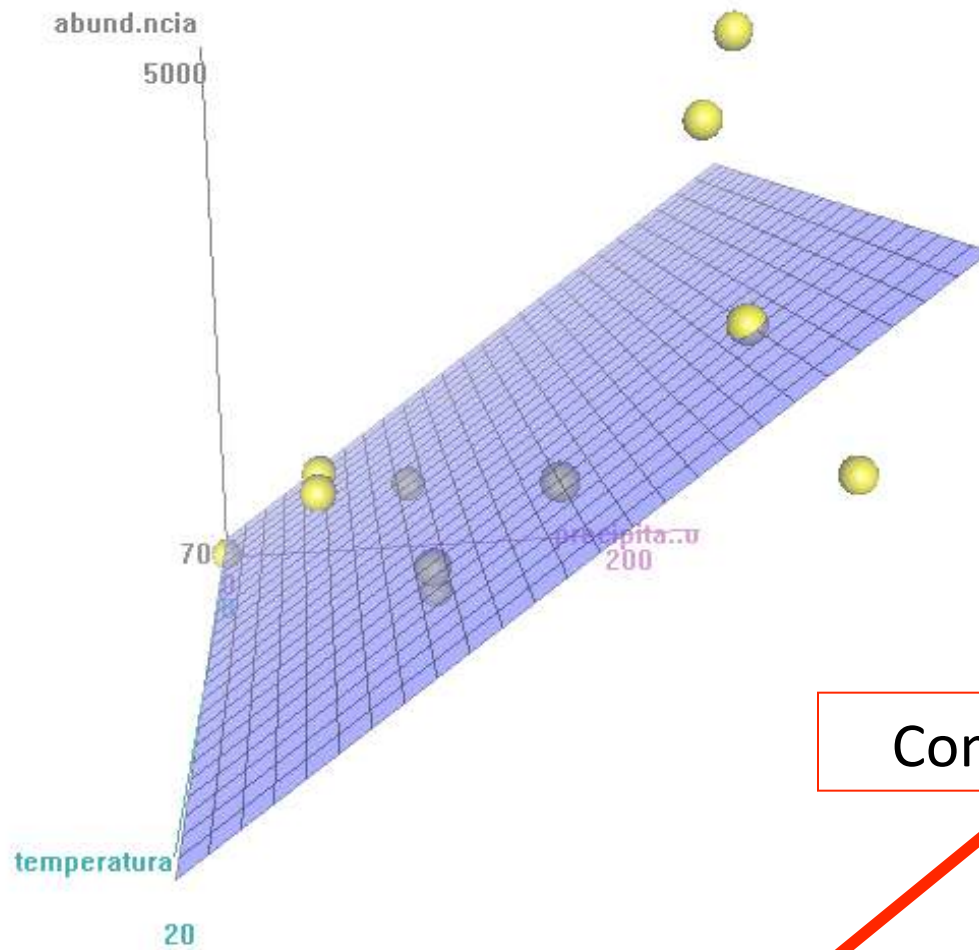
- Usada para reduzir a dimensionalidade dos dados
- Relação entre as amostras
- Maximiza a variância entre os objetos

PCA



$$Y=a+bX$$

PCA



Combinação Linear

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

Quadro

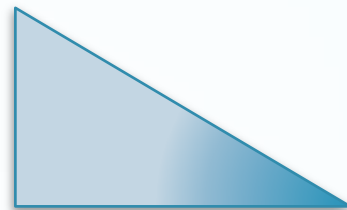
1	ACV	1440.0	1814.00	1090.0
2	DE	1000.0	1000.00	1000.0
3	DAW	1000.0	1000.00	1000.0
4	DAW	1000.0	1000.00	1000.0
5	DAW	1000.0	1000.00	1000.0
6	DAW	1000.0	1000.00	1000.0
7	DAW	1000.0	1000.00	1000.0
8	DAW	1000.0	1000.00	1000.0
9	DAW	1000.0	1000.00	1000.0
10	DAW	1000.0	1000.00	1000.0
11	DAW	1000.0	1000.00	1000.0
12	DAW	1000.0	1000.00	1000.0
13	DAW	1000.0	1000.00	1000.0
14	DAW	1000.0	1000.00	1000.0
15	DAW	1000.0	1000.00	1000.0
16	DAW	1000.0	1000.00	1000.0
17	DAW	1000.0	1000.00	1000.0
18	DAW	1000.0	1000.00	1000.0
19	DAW	1000.0	1000.00	1000.0
20	DAW	1000.0	1000.00	1000.0
21	DAW	1000.0	1000.00	1000.0
22	DAW	1000.0	1000.00	1000.0
23	DAW	1000.0	1000.00	1000.0
24	DAW	1000.0	1000.00	1000.0

Transformação



1	ACV	1440.0	1814.00	1090.0
2	DE	1000.0	1000.00	1000.0
3	DAW	1000.0	1000.00	1000.0
4	DAW	1000.0	1000.00	1000.0
5	DAW	1000.0	1000.00	1000.0
6	DAW	1000.0	1000.00	1000.0
7	DAW	1000.0	1000.00	1000.0
8	DAW	1000.0	1000.00	1000.0
9	DAW	1000.0	1000.00	1000.0
10	DAW	1000.0	1000.00	1000.0
11	DAW	1000.0	1000.00	1000.0
12	DAW	1000.0	1000.00	1000.0
13	DAW	1000.0	1000.00	1000.0
14	DAW	1000.0	1000.00	1000.0
15	DAW	1000.0	1000.00	1000.0
16	DAW	1000.0	1000.00	1000.0
17	DAW	1000.0	1000.00	1000.0
18	DAW	1000.0	1000.00	1000.0
19	DAW	1000.0	1000.00	1000.0
20	DAW	1000.0	1000.00	1000.0
21	DAW	1000.0	1000.00	1000.0
22	DAW	1000.0	1000.00	1000.0
23	DAW	1000.0	1000.00	1000.0
24	DAW	1000.0	1000.00	1000.0

Correlação de Pearson

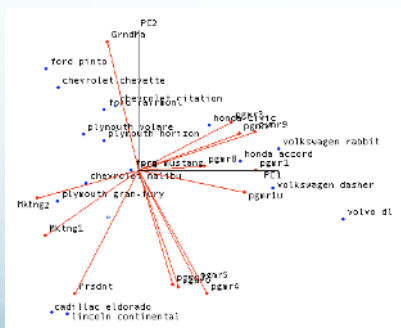


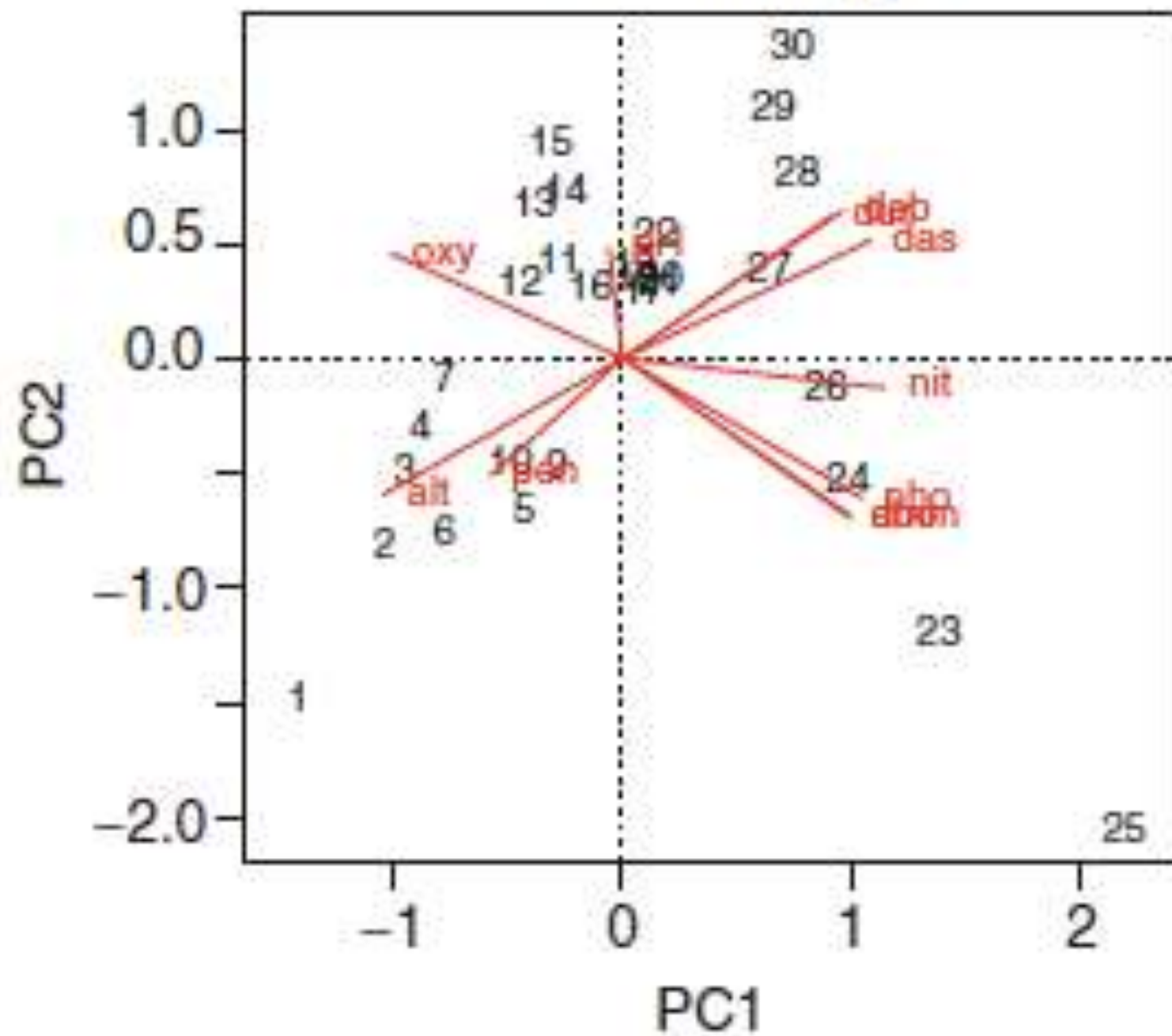
Autovetores
e
Autovalores

Autovetores
e
Autovalores



Combinações lineares





Vantagens PCA

- Retira multicolinearidade: cria um novo conjunto de variáveis independentes.
- Combina variáveis únicas originais em um novo conjunto de variáveis
- Matriz de correlação: quando os dados *não* estão na mesma unidade de medida
- Matriz de covariância: quando os dados estão na mesma unidade

Limitações da PCA

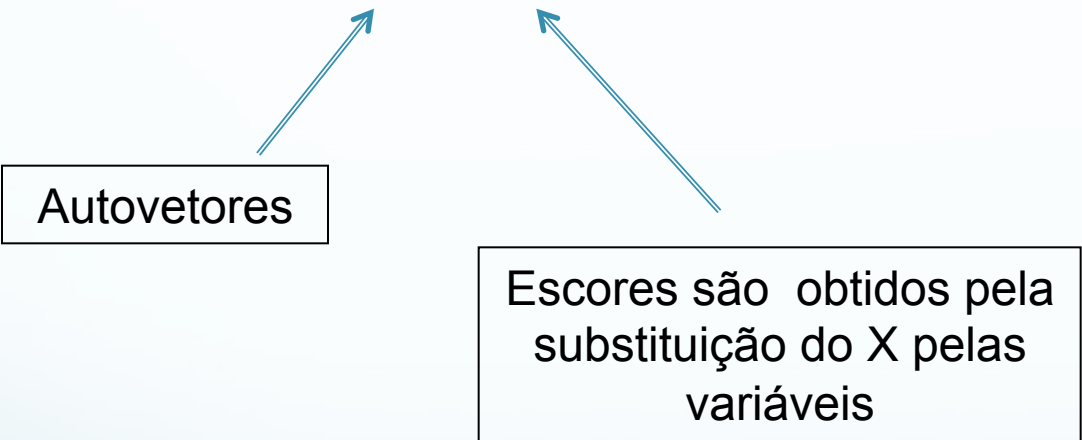
- Coeficiente de distância
 - Distância Euclidiana
 - Variáveis ambientais e medidas (morfológicas, distância etc.)
- Sensível a *outliers*
- Duplas ausências (muitos zeros) e dados ausentes (NA's)
- Mais variáveis que unidades amostrais
- Se a % de explicação dos eixos muito similar => sem estrutura clara nos dados

Conceitos

- **Autovalores** (eigenvalues): % de explicação de cada eixo, % variação nos dados que o eixo explica.
- **Autovetores** (eigenvectors)/Loading: Importância das variáveis para formação dos eixos
- **Escores**: posição das unidades amostrais no eixo de ordenação e das variáveis (setas) no espaço de ordenação

$$Z = a_1 X_1 + a_2 X_2 + a_3 X_3 \dots a_n X_n$$

Autovetores



Escores são obtidos pela
substituição do X pelas
variáveis

Critérios de parada

- Kaiser-Guttman: médias dos autovetores
- Regra do dedão: 75%
- Esferidade de Barlet
- Diagrama
- Broken-Stick: eixos maiores que o valor preditor pelo modelo

A análise é satisfatória?

- Os 3 primeiros eixos explicam em torno de 70% da variação dos dados
- É possível extrair os autovetores para utilizar em uma análise de regressão

PCoA

- Semelhante à PCA
- Centralização dupla: manter a relação euclidiana entre as u.a.
- Permite uso de diferentes índices de similaridade
- Ordenar dados quando só a matrix de similaridade está disponível
- Robusta para NA's, duplas ausências, *dummy* e quando $P > n$

Desvantagens da PCoA

- Não informa quais variáveis influenciam a distribuição dos objetos
- Não informa relação das variáveis e os eixos
- Não permite interpretar os eixos com base na projeção dos descritores

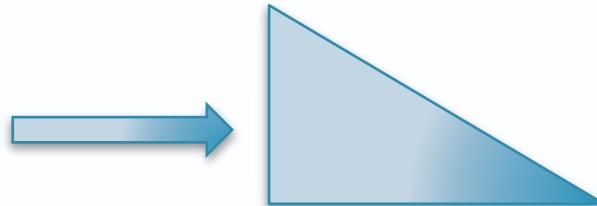
A análise é adequada?

- Autovalores altos e negativos: matriz de distância inadequada

nMDS

- Também permite utilizar qualquer coeficiente de distância.
- Robusto para duplas ausências, NA's e dados de incidência
- STRESS (STandard Residuals Sum of Squares): quanto a posição dos obj. se desviam da distâncias originais após os escalonamento.
- Não é possível extrair os eixos para utilizar em outras análises (não-métrico).

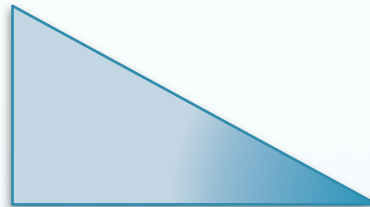
Atribuição de
Escores aleatórios
aos eixos



Aleatorização do
STRESS

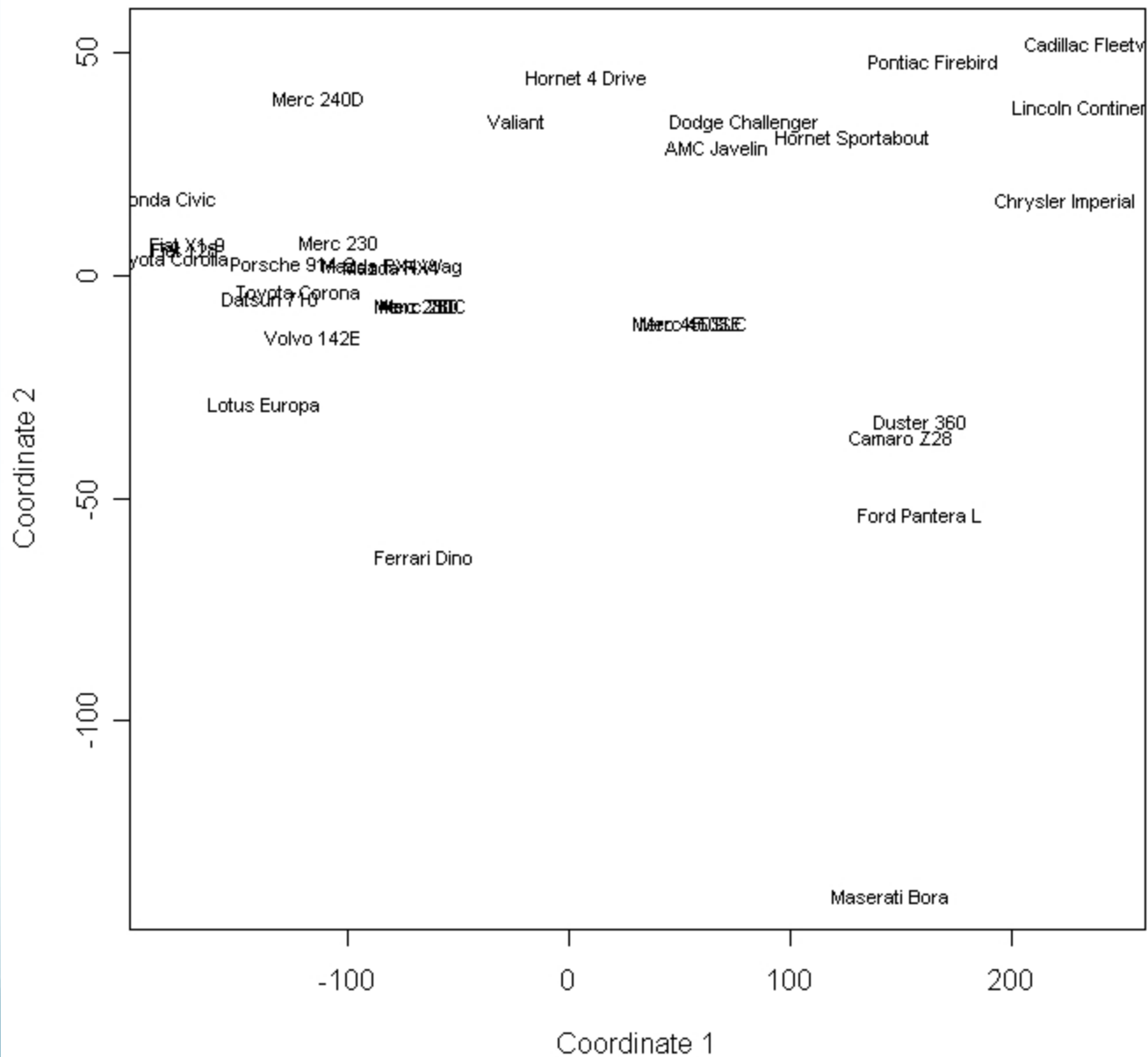
ρ

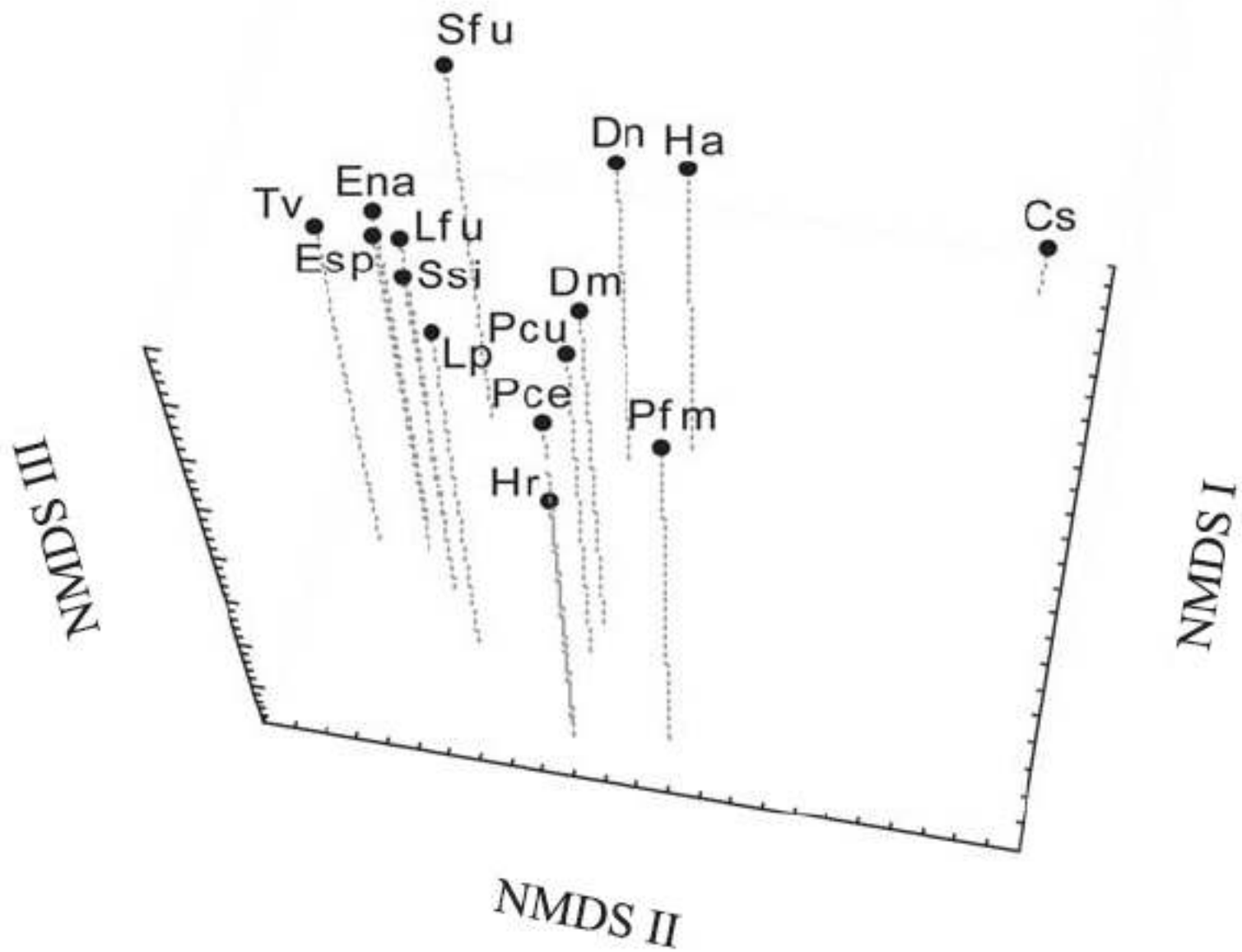
STRESS



1	ADN	2446.30	PRIMA (0)	PRIMA (0)
2	QDE	PRIMA (0)	PRIMA (0)	PRIMA (0)
3	QDE	2446.30	PRIMA (0)	PRIMA (0)
4	QDE	PRIMA (0)	PRIMA (0)	PRIMA (0)
5	QDE	2446.30	PRIMA (0)	PRIMA (0)
6	QDE	2446.30	PRIMA (0)	PRIMA (0)
7	QDE	2446.30	PRIMA (0)	PRIMA (0)
8	QDE	2446.30	PRIMA (0)	PRIMA (0)
9	QDE	2446.30	PRIMA (0)	PRIMA (0)
10	QDE	2446.30	PRIMA (0)	PRIMA (0)
11	QDE	2446.30	PRIMA (0)	PRIMA (0)
12	QDE	2446.30	PRIMA (0)	PRIMA (0)
13	QDE	2446.30	PRIMA (0)	PRIMA (0)
14	QDE	2446.30	PRIMA (0)	PRIMA (0)
15	QDE	2446.30	PRIMA (0)	PRIMA (0)
16	QDE	2446.30	PRIMA (0)	PRIMA (0)
17	QDE	2446.30	PRIMA (0)	PRIMA (0)
18	QDE	2446.30	PRIMA (0)	PRIMA (0)
19	QDE	2446.30	PRIMA (0)	PRIMA (0)
20	QDE	2446.30	PRIMA (0)	PRIMA (0)
21	QDE	2446.30	PRIMA (0)	PRIMA (0)
22	QDE	2446.30	PRIMA (0)	PRIMA (0)
23	QDE	2446.30	PRIMA (0)	PRIMA (0)
24	QDE	2446.30	PRIMA (0)	PRIMA (0)
25	QDE	2446.30	PRIMA (0)	PRIMA (0)
26	QDE	2446.30	PRIMA (0)	PRIMA (0)
27	QDE	2446.30	PRIMA (0)	PRIMA (0)
28	QDE	2446.30	PRIMA (0)	PRIMA (0)
29	QDE	2446.30	PRIMA (0)	PRIMA (0)
30	QDE	2446.30	PRIMA (0)	PRIMA (0)

Nonmetric MDS





STRESS

- Regra de dedão para interpretação (Clarke, 1993):
- $\text{STRESS} < 0.05 \Rightarrow$ Excelente
- $\text{STRESS} < 0.1 \Rightarrow$ Bom
- $\text{STRESS} < 0.2 \Rightarrow$ Razoável
- $\text{STRESS} > 0.2 \Rightarrow$ Interpretação inviável
- ??? Tente utilizar outro coeficiente de distância

Ordenação restrita

- CCA= Análise de Correlação Canônica
- RDA= Análise de Redundância

CCA

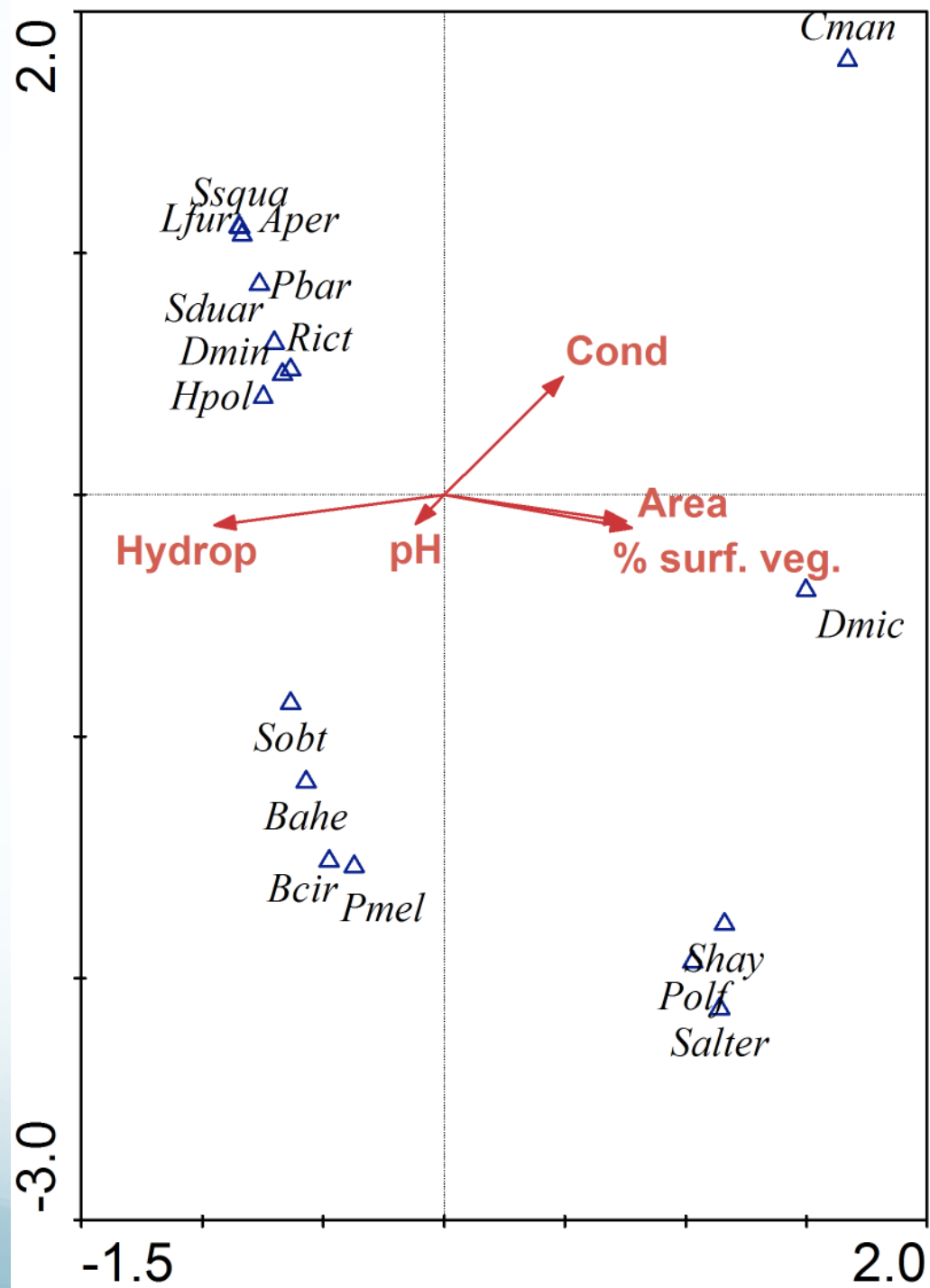
- Identificar a influência das variáveis ambientais nos padrões de composição e abundância
- Estrutura de correlação dentro da matriz de espécies e entre a matriz ambiental e a de espécies
- Resposta unimodal aos gradientes
- Gradientes incompletos
- Maximiza a separação dos nichos das espécies

CCA

- Ideal quando há mais unidades amostrais que descritores.
- É possível realizar a análise somente com dados de incidência
- Baseada na distância de $\chi^2 \Rightarrow$ dá sobre peso para espécies raras (!), considere retirar as sp. raras se não foram adequadamente amostradas

RDA

- Resposta linear das espécies em relação aos gradientes
- Regressão múltipla multivariada seguida de PCA com valores ajustados



Interpretação do biplot

- Ótimo da sp => projeção perpendicular na seta
- Posição da amostra no eixo => projeção perpendicular no eixo
- Quanto maior a seta, mais importante a variável para explicar a distribuição das espécies
- Quadrante da seta => correlação positiva
- Quanto menor o ângulo entre a seta e o eixo, maior a correlação
- Problema da seleção de variáveis

Respostas lineares ou unimodais?

- Procedimento de aleatorização de Monte Carlo baseado na estatística F para avaliar a significância dos autovalores dos eixos