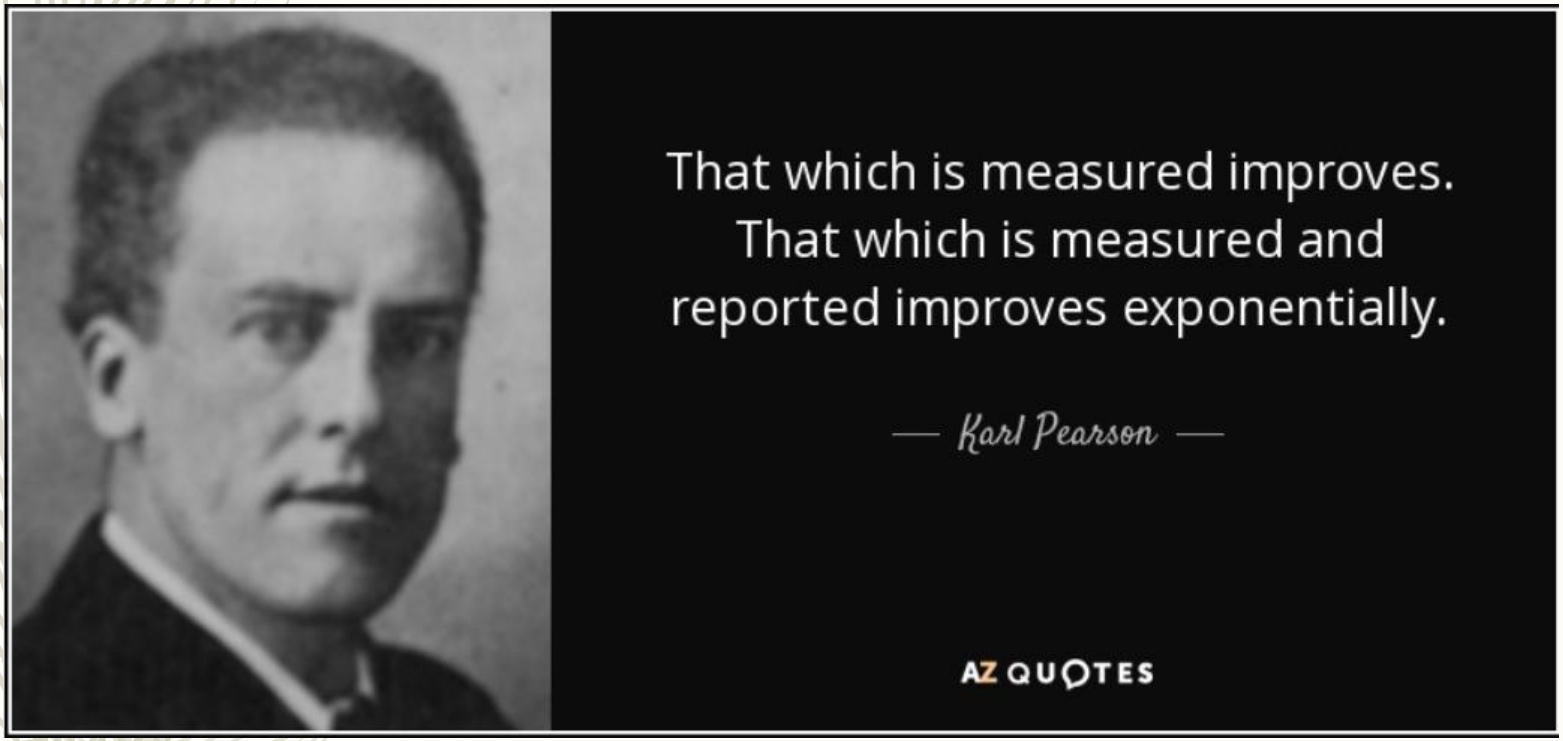
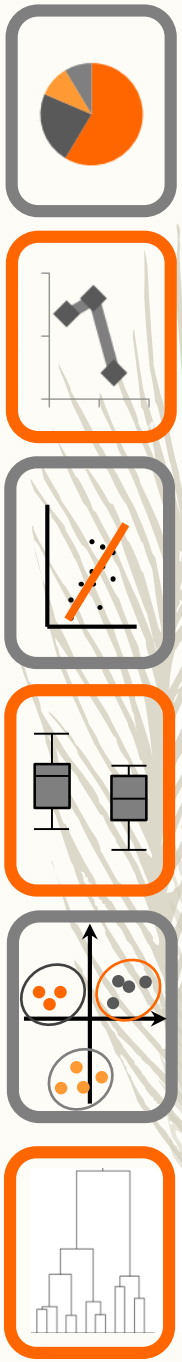
Aula 24 Goodies*

* Goodies related to animals, plants and numbers…

Ecologia Numérica - Aula Teórica 24 – 09-12-2018



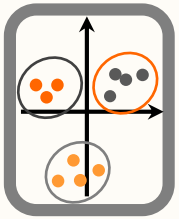
https://www.azquotes.com/quote/727622



ecologia numérica

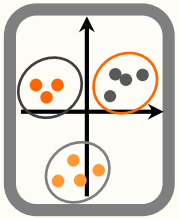
Introdução à análise multivariada

ordenação



ordenação

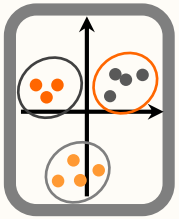
- Quais as técnicas de ordenação a utilizar quando temos várias variáveis de interesse?
- Quais as suas diferenças e âmbito de aplicação?
- Como interpretar os resultados?
- Que técnica escolher?



Ordenação

Âmbito:

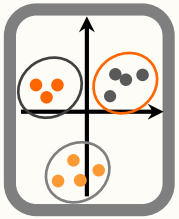
Ordenar entidades ou objectos num espaço de dimensão reduzida, com base nas suas características.



Ordenação

Principais objectivos:

- Redução de dimensionalidade;
- Evidenciar padrões nos dados;
- Evidenciar as relações entre variáveis e variáveis e observações.



Ordenação: técnicas mais utilizadas

- Análise de componentes principais (PCA);
- Análise de correspondências (CA);
- Análise canónica de correspondências (CCA);
- Escalamento multidimensional (MDS);
- Análise discriminante (DA)

Estas classificações dependem dos autores, mas para mim a DA é uma análise de classificação, não de ordenação. O objectivo é classificar objectos (e.g. locais, espécies, filmes, pessoas, marcas, etc) em grupos em função das suas caracteristicas, e não ordenar objectos num espaço multidimensional (apesar de ser isso que a DA faz, ver onde é que as observações ficam colocadas num espaço multidimensional para as separar por grupos nesse espaço)

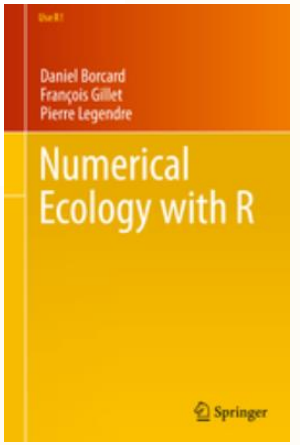
The methods that are presented in this chapter are:

- *Principal component analysis (PCA)*: the main eigenvector-based method. Works on raw, quantitative data. Preserves the Euclidean distance among sites.
- *Correspondence analysis (CA)*: works on data that must be frequencies or frequency-like, dimensionally homogeneous, and non-negative. Preserves the χ^2 distance among rows or columns. Mainly used in ecology to analyse species data tables.

[redacted]

- *Nonmetric multidimensional scaling (NMDS)*: unlike the three others, this is not an eigenvector-based method. NMDS tries to represent the set of objects along a predetermined number of axes while preserving the ordering relationships among them.

[redacted] NMDS can produce ordinations from any square distance matrix.



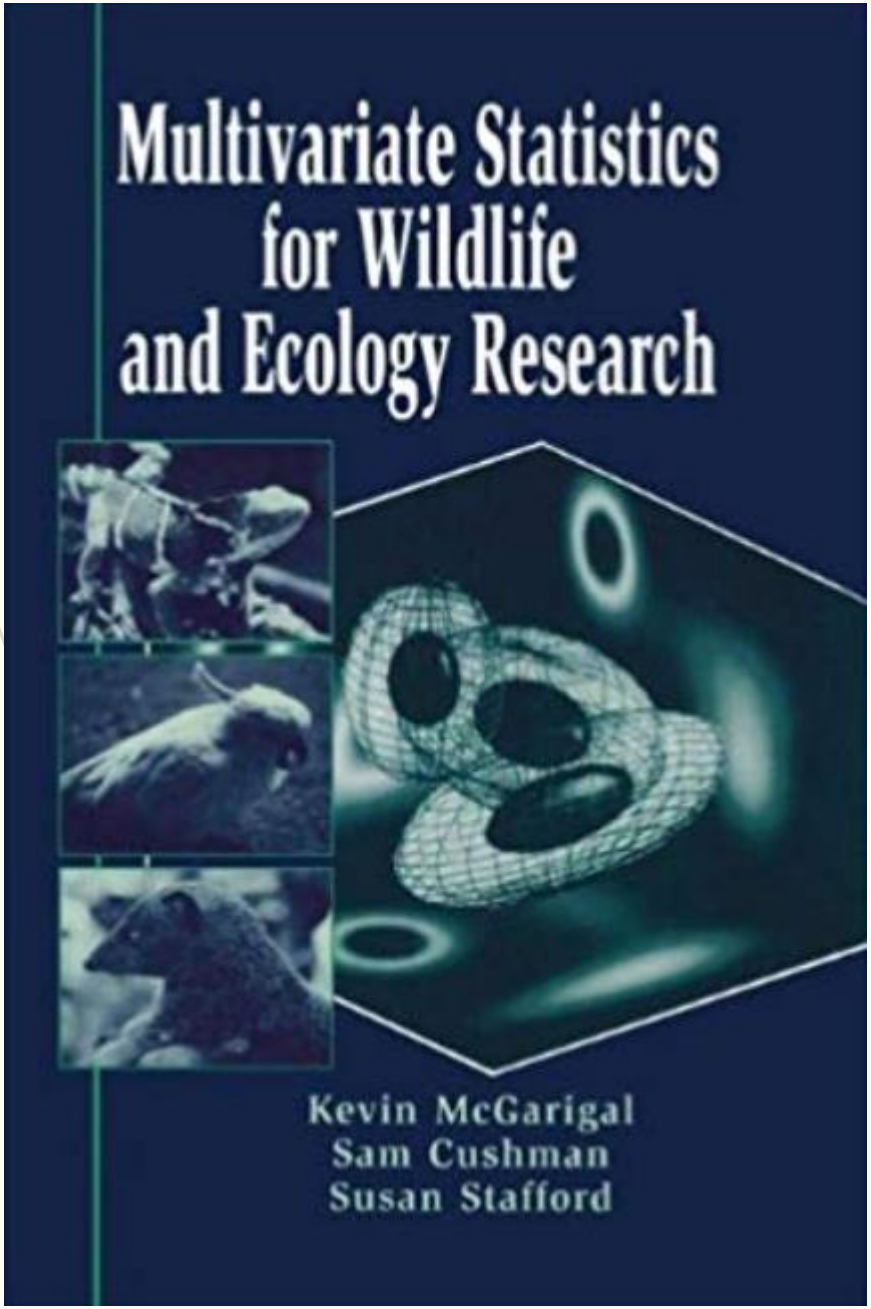


TABLE 1.2 Alternative terminology for multivariate techniques. The labels used in this book are given in the left-hand column.

Technique	Alternative Names
Ordination	
Principal components analysis	Factor analysis
Polar ordination	Bray and Curtis ordination
Factor analysis	None
Nonmetric multidimensional scaling	None
Reciprocal averaging	Correspondence analysis
Detrended correspondence analysis	None
Canonical correspondence analysis	None
Cluster analysis	Botryology
	Classification
	Clumping
	Grouping
	Morphometrics
	Nosography
	Nosology
	Numerical taxonomy
	Partitioning
	Q-analysis
	Segmentation analysis
	Systematics
	Taximetrics
	Taxonomics
	Typology
	Unsupervised pattern recognition
Discriminant analysis	
Canonical analysis of discriminance	Discriminant analysis
	Discriminant function analysis
	Multiple discriminant analysis
	Descriptive discriminant analysis
	Canonical variates analysis
	Fisher's linear discriminant function analysis
Classification	Discriminant analysis
	Discriminant function analysis
	Multiple discriminant analysis
	Predictive discriminant analysis
	Fisher's linear discriminant function analysis
Canonical correlation analysis	Canonical analysis

with predictive discrimination. Similarly, the terms *factor analysis* and *principal components analysis* are sometimes used interchangeably, when in fact they are distinctly different techniques (albeit techniques that accomplish a similar end).

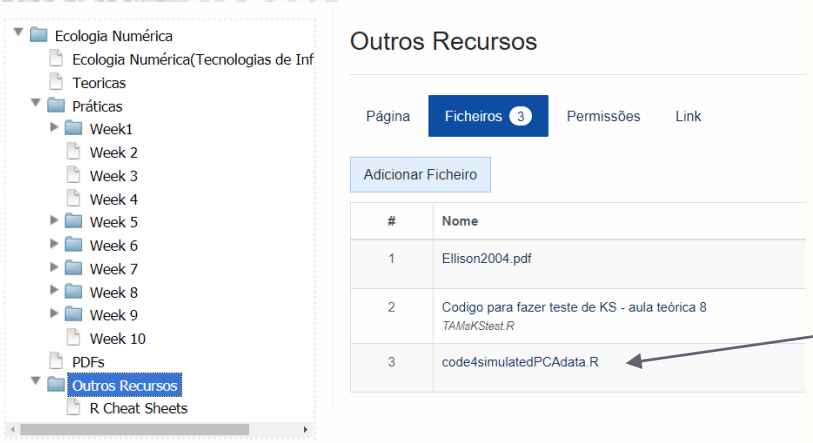
Equally confusing is the terminology used to define or classify variables. In this book, variable labels are used repeatedly to describe techniques and, more importantly, to compare and contrast various techniques. Consequently, it is very important that you fully understand the meaning of the various labels. Variables are labeled

A simulated example*, but a useful example to understand what is at stake in a multivariate analysis (here, a PCA)!

Informação:

1. taxa por locais
2. variáveis ambientais por locais

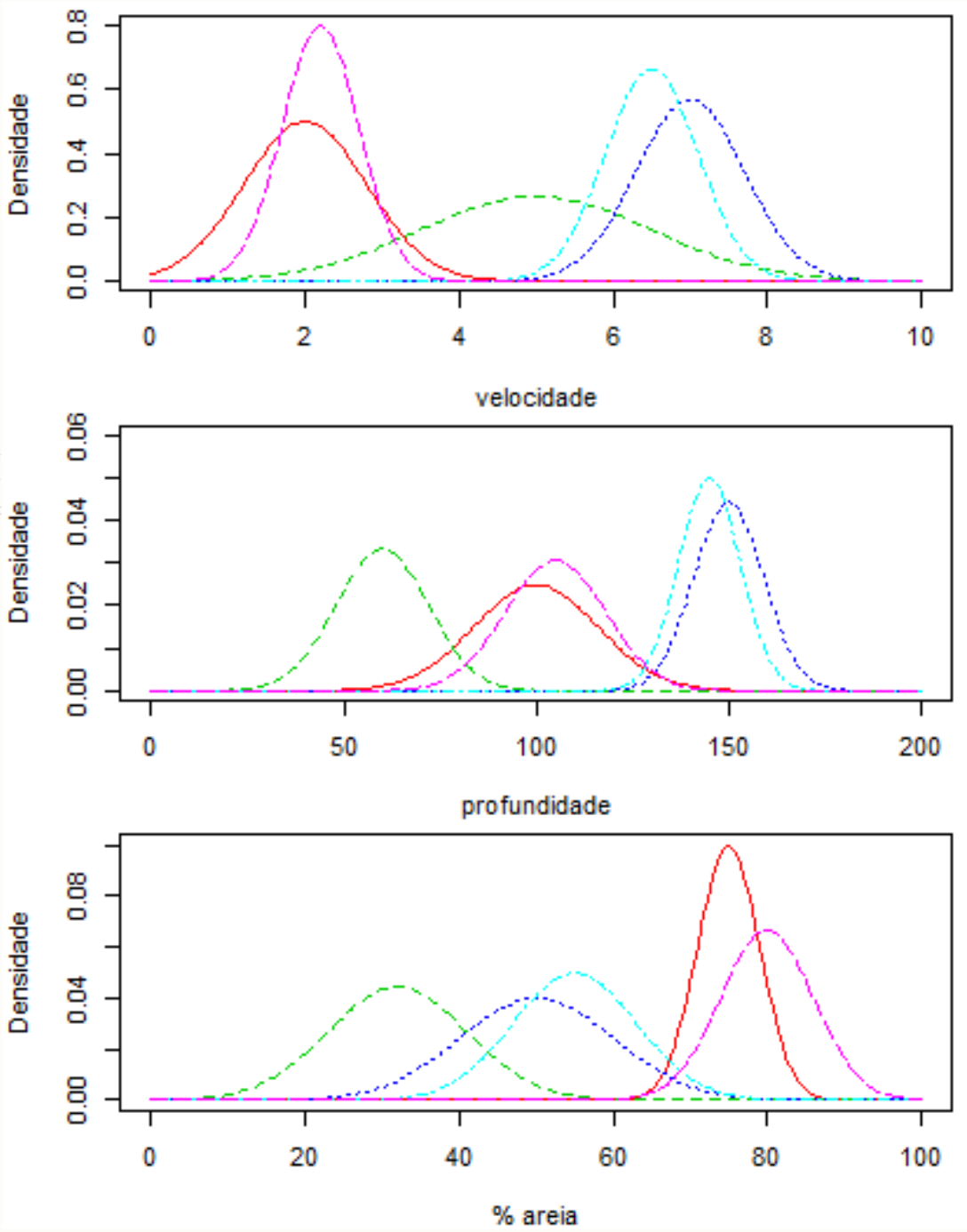
- Note-se que as espécies reagem às variáveis ambientais, mas numa análise indireta as variáveis ambientais não são consideradas.



* O Código para simular estes dados está no FENIX

- 5 espécies
- 20 locais
- Espécies azuis e espécies rosa-vermelha com requisitos ecológicos muito semelhantes, espécie verde diferente das outras

```
> round(ambvars,1)
   velocidade profundidade areia
1         1.1          63.3  55.3
2         6.2          60.5  64.6
3         6.1          31.8  31.2
4         6.2           8.0  62.2
5         8.6          43.8  33.0
6         6.4         162.1  50.2
7         0.1         105.1  67.7
8         2.3         182.9  48.5
9         6.7         166.3  24.4
10        5.1           9.2  76.5
11        6.9          91.2   7.4
12        5.4          53.0  31.0
13        2.8          60.9  71.7
14        9.2         101.5  50.5
15        2.9          36.2  15.3
16        8.4         151.9  50.4
17        2.9          40.2  49.4
18        2.7          51.8  75.1
19        1.9         198.4  17.5
20        2.3         161.5  84.8
```



- 3 variáveis ambientais (velocidade, profundidade, % areia)
- 5 espécies
- 20 locais
- Espécies azuis e espécies rosa-vermelha com requisitos ecológicos muito semelhantes, espécie verde diferente das outras

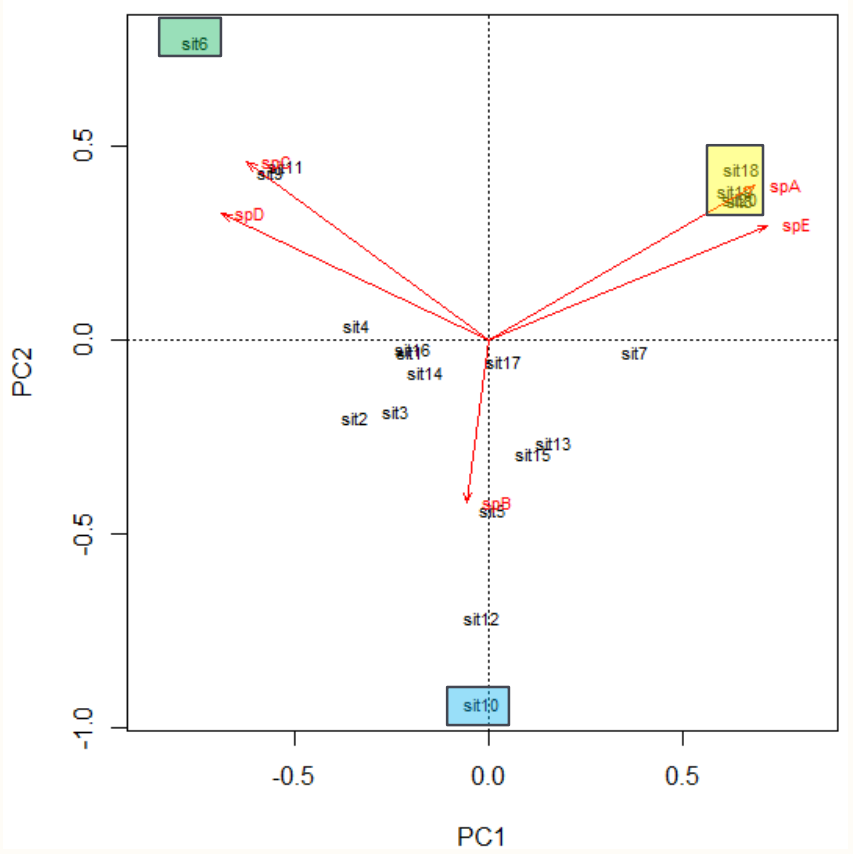
- Cada local encontra-se num espaço multivariado (de dimensão 3) no que diz respeito às variáveis ambientais que o caracterizam

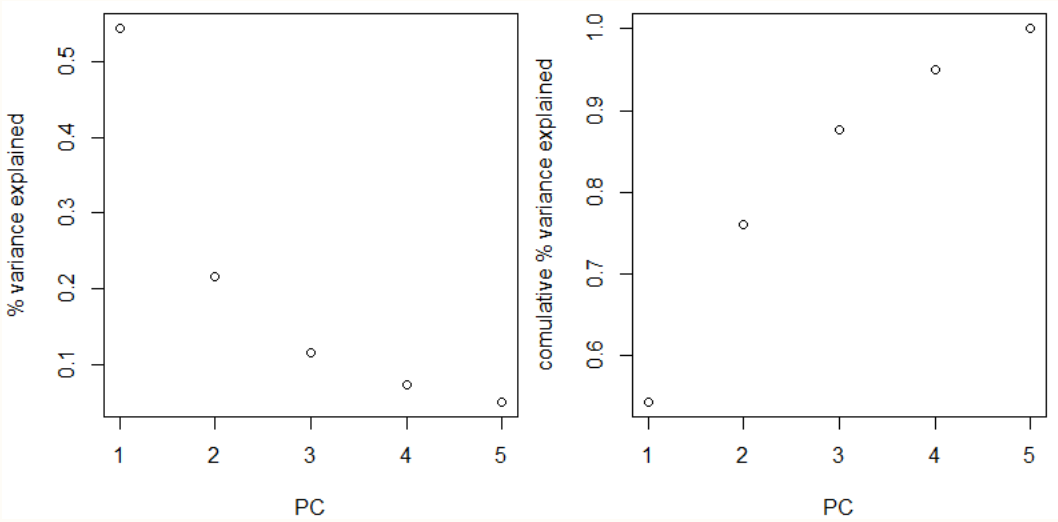
- Cada local encontra-se num espaço multivariado (de dimensão 5 em relação às espécies que lá estão presentes)

- O objectivo de uma PCA é representar os locais (e as variáveis) num espaço dimensional reduzido (tipicamente, 2 dimensões, ou seja, um biplot)

Densidades por local

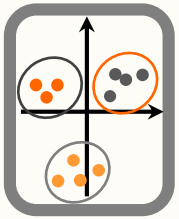
```
> round(specbyloc,2)
    spA  spB  spC  spD  spE
1  0.10 0.00 0.00 0.70 0.03
2  0.00 0.41 0.15 0.78 0.00
3  0.00 0.15 0.08 0.59 0.00
4  0.00 0.01 0.16 0.80 0.00
5  0.00 0.05 0.00 0.00 0.00
6  0.00 0.00 1.21 0.97 0.00
7  0.00 0.00 0.00 0.00 0.94
8  0.72 0.00 0.00 0.00 0.95
9  0.00 0.00 0.70 0.92 0.00
10 0.00 0.94 0.00 0.00 0.00
11 0.00 0.00 0.98 0.55 0.00
12 0.00 0.57 0.00 0.03 0.00
13 0.12 0.07 0.00 0.00 0.29
14 0.00 0.00 0.52 0.00 0.00
15 0.07 0.00 0.00 0.00 0.19
16 0.00 0.00 0.62 0.00 0.00
17 0.10 0.00 0.32 0.00 0.25
18 1.20 0.00 0.00 0.00 0.50
19 0.94 0.00 0.00 0.00 0.71
20 0.72 0.00 0.00 0.00 0.95
```





- **Relações:**
  1. Entre locais, em função das espécies neles presentes
  2. Entre locais, em função das variáveis ambientais que apresentam
  3. Entre espécies, em função das variáveis ambientais que preferem
  4. Entre espécies, em função dos locais em que se encontram simultaneamente presentes ou ausentes

  Objectivos:
  1. compreender e descrever as relações de proximidade entre os locais para explicar quais os fatores determinantes na organização (na ordenação) das comunidades
  2. perceber como é que as variáveis ambientais contribuem para estruturas os gradientes observados nas comunidades animais



Análise de Componentes Principais

Transforma um conjunto de dados
em combinações lineares, facilitando a interpretação

Método

A partir de k variáveis originais : x1,x2,...,xk:
Produzem-se k novas variáveis: y1,y2,...,yk:

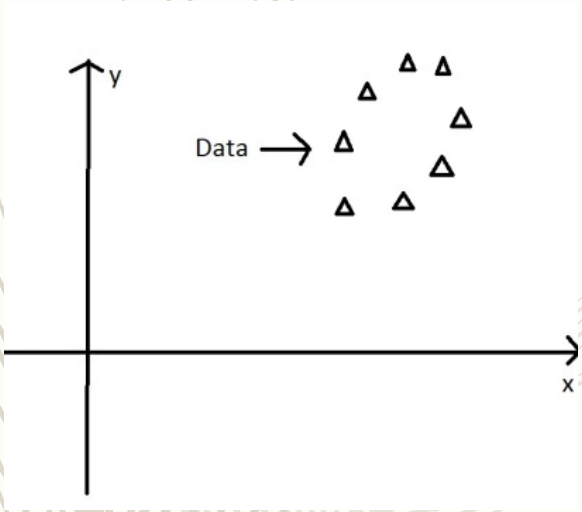
$$y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1k}x_k$$
$$y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2k}x_k$$
$$\ldots$$
$$y_k = a_{k1}x_1 + a_{k2}x_2 + \ldots + a_{kk}x_k$$

→ Componentes Principais

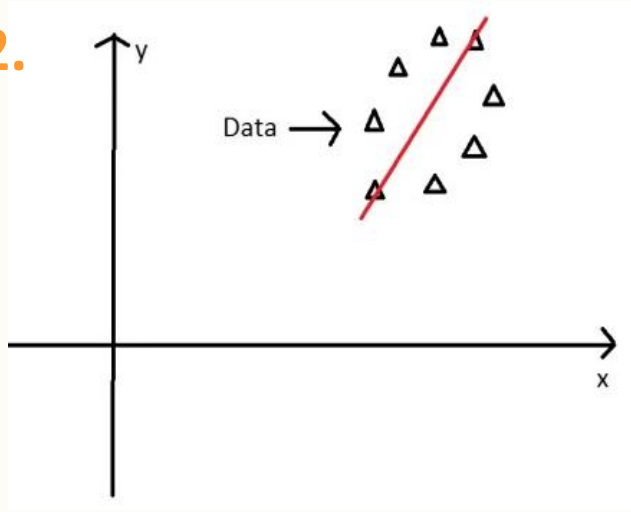
Os y's são combinações lineares dos x's (i.e. das abundancies das espécies)

Análise de Componentes Principais

1.

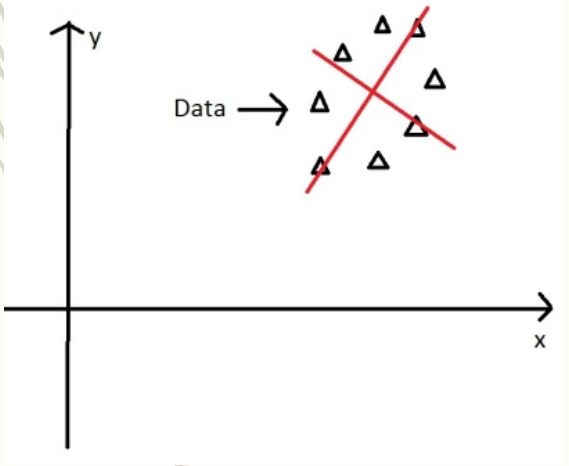


2.



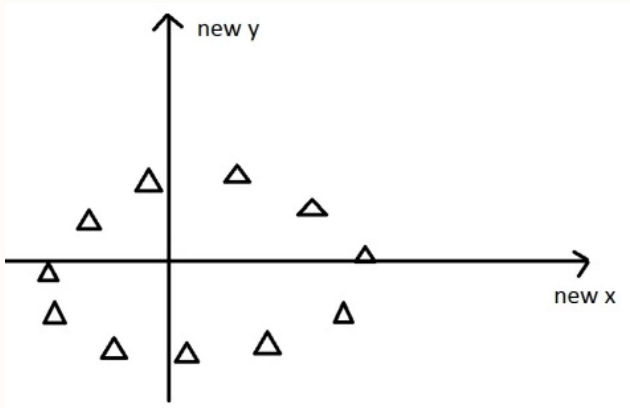
1º CP: onde a variância é maior

3.



2º CP: ortogonal ao 1º

4.



Novos eixos = 1º e 2º CP's

Análise de Componentes Principais

Dados Originais

$$X = \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{bmatrix}$$

Matriz Variância – Covariância

$$\sum = \begin{bmatrix} Var_{(x_1)} & Cov_{(x_1x_2)} & Cov_{(x_1x_3)} \\ Cov_{(x_2x_1)} & Var_{(x_2)} & Cov_{(x_2x_3)} \\ Cov_{(x_3x_1)} & Cov_{(x_3x_2)} & Var_{(x_3)} \end{bmatrix}$$

$$\det\left[\sum - \lambda I\right] = 0$$

Componentes Principais

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$

$$Y_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$

...

$$Y_p = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p$$

$$\lambda_1 > \lambda_2 > \lambda_3 > \ldots > \lambda_p \qquad \tilde{a} = \begin{bmatrix} a_{i1} \\ \vdots \\ a_{ip} \end{bmatrix}$$

Análise de Componentes Principais

Vectores próprios: "peso" de cada variável nas componentes principais

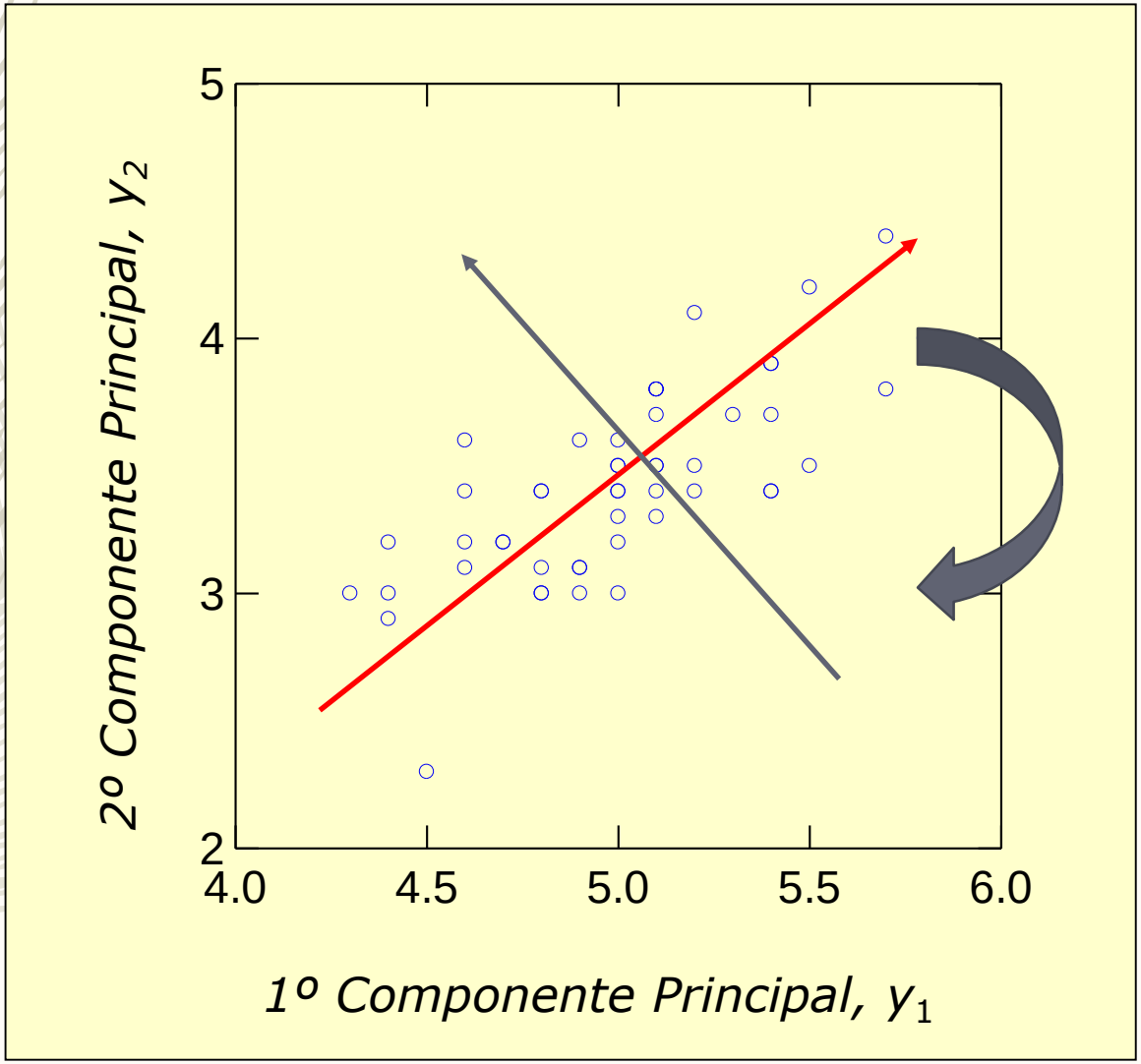
Valores próprios: informação que cada componente principal "carrega"

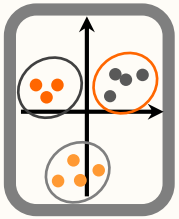
Componentes Principais (yk's):

- são ortogonais, i.e. não correlacionados
- y1 explica a maior quantidade possível da variância
- y2 explica a maior quantidade possível da variância remanescente

… etc.

Análise de Componentes Principais

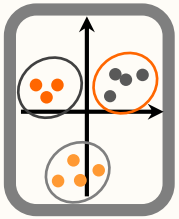




Análise de Componentes Principais

A extração dos CP's pode ser baseada na:

- Matriz de variância-covariância (argumento de `princomp cor=FALSE`, quando as variáveis são medidas nas mesmas unidades), ou queremos que as abundâncias influenciem a ordenação
- Matriz de correlação (argumento de `princomp cor=TRUE`, ou `scale(dados)`, se estamos mais interessados nas abundancias relativas e não na abundancia total

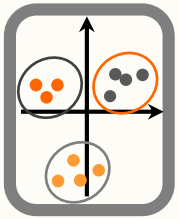


Análise de Componentes Principais

Então, sendo

$$y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1k}x_k$$
$$y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2k}x_k$$
$$\ldots$$
$$y_k = a_{k1}x_1 + a_{k2}x_2 + \ldots + a_{kk}x_k$$

xj's são padronizados (redução e centragem) se for utilizada a matriz de correlação (no R, `scale(dados)`)



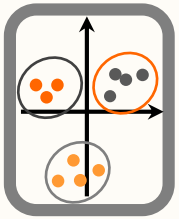
Análise de Componentes Principais

{a11,a12,…,a1k} é o 1º vector próprio (eigenvector) da matriz de correlação/covariância, o qual inclui os coeficientes da 1ª componente principal;

{a21,a22,…,a2k} é o 2º vector próprio (eigenvector) da matriz de correlação/covariância, o qual inclui os coeficientes da 2ª componente principal;

…

{ak1,ak2,…,akk} é o k-ésimo vector próprio (eigenvector) da matriz de correlação/covariância, o qual inclui os coeficientes da k-ésima componente principal.



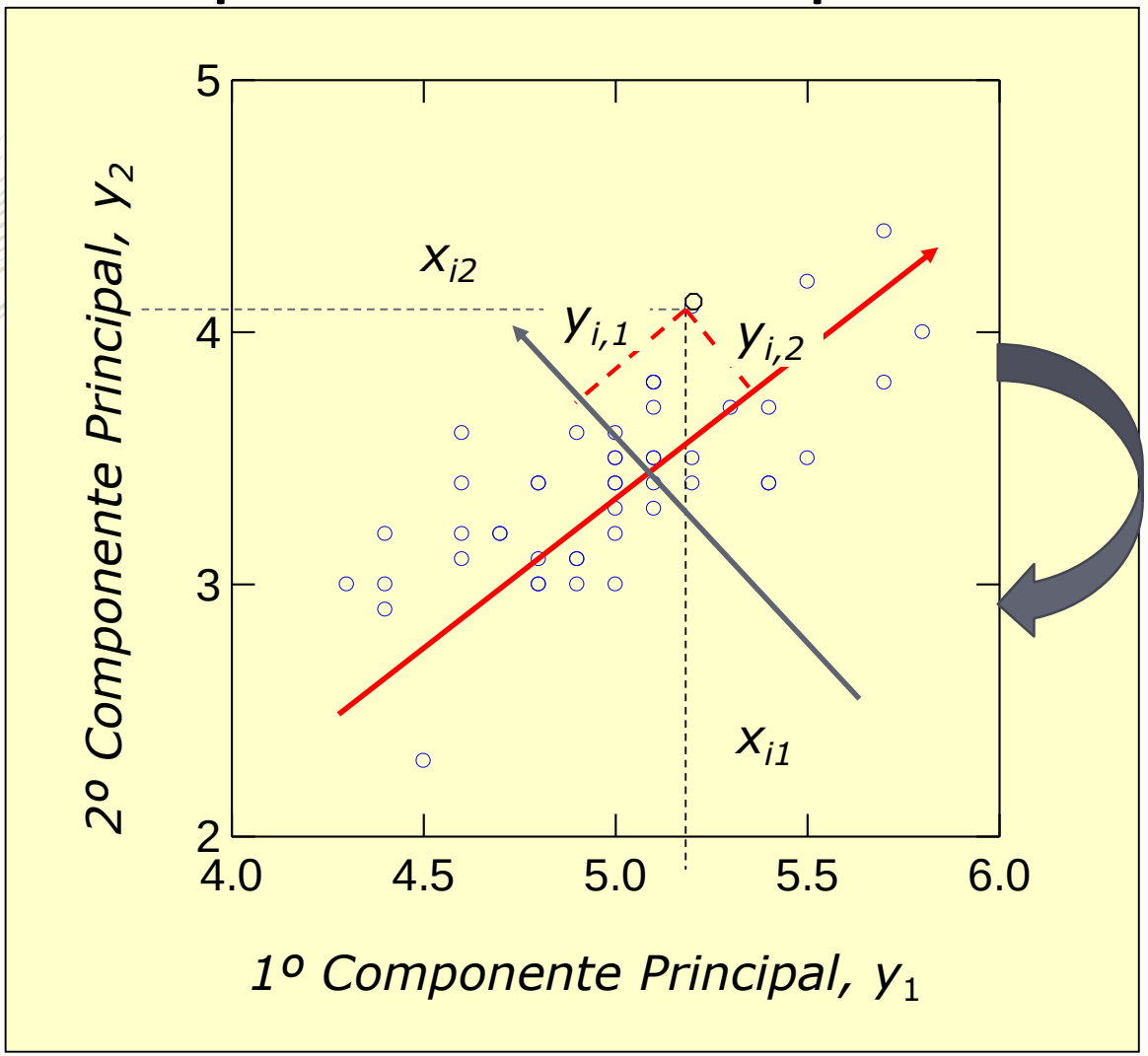
Análise de Componentes Principais

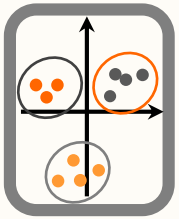
O score (coordenada) da i-ésima observação no j-ésimo componente principal é

$$y_{i,j} = a_{j1}x_{i1} + a_{j2}x_{i2} + \ldots + a_{jk}x_{ik}$$

Análise de Componentes Principais

Scores (coordenadas)





Análise de Componentes Principais

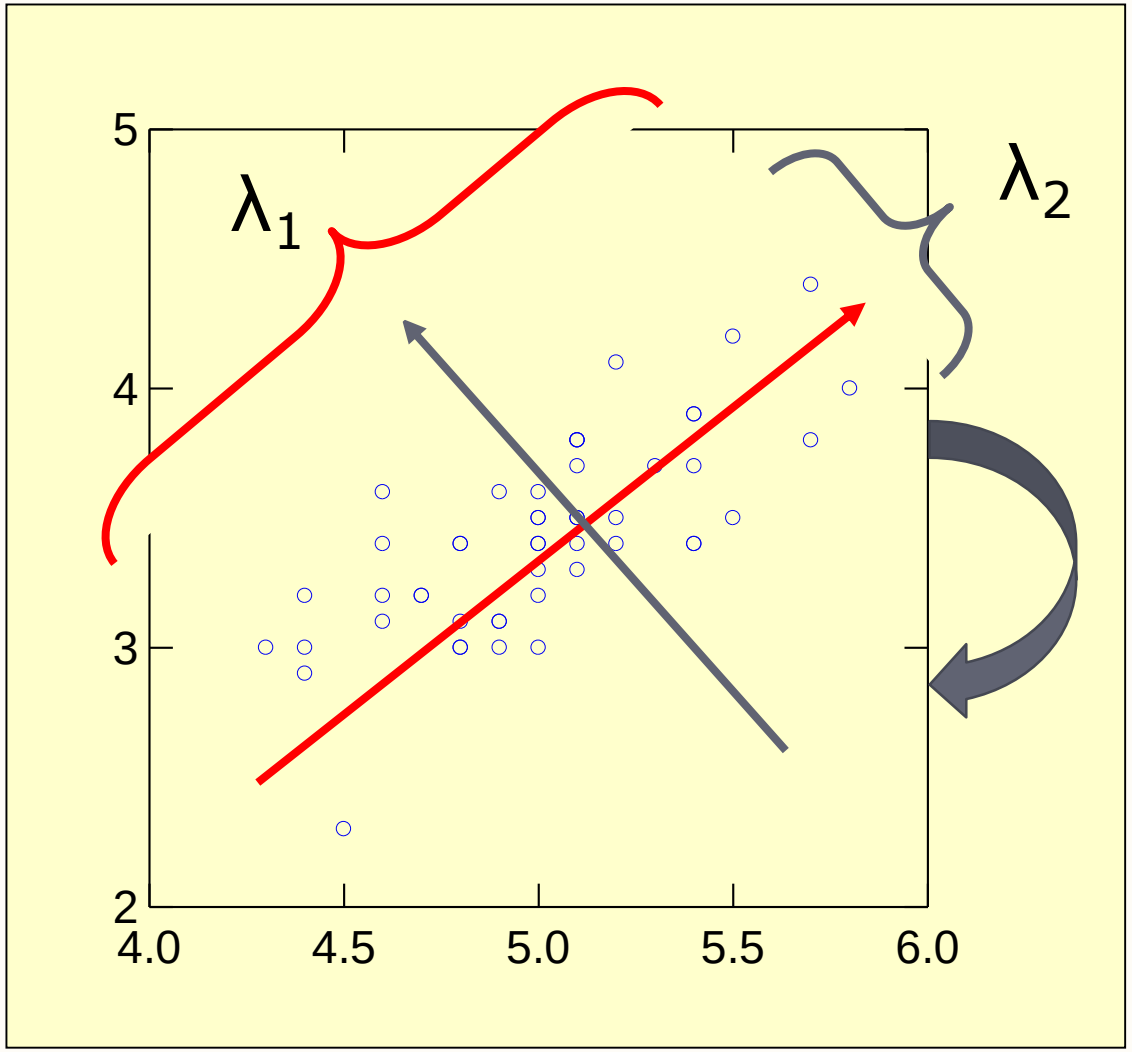
A quantidade de variância explicada pelo:

- 1º componente principal, λ1, é o 1º valor próprio (eigenvalue)
- 2º componente principal, λ2, é o 2º valor próprio (eigenvalue)
- …

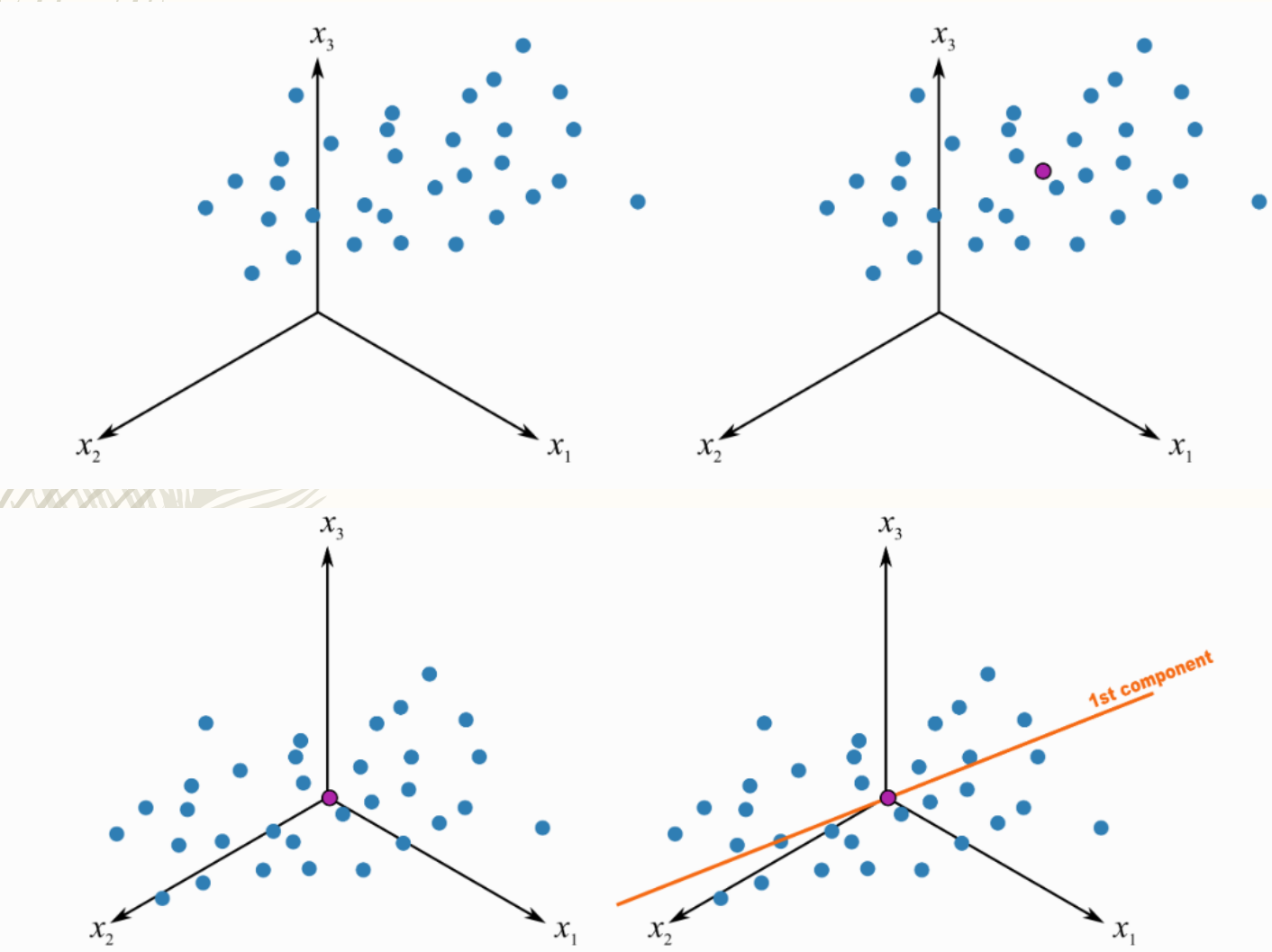
λ1 > λ2 > λ3 > λ4 > …

Análise de Componentes Principais

Valores próprios
(eigenvalues)

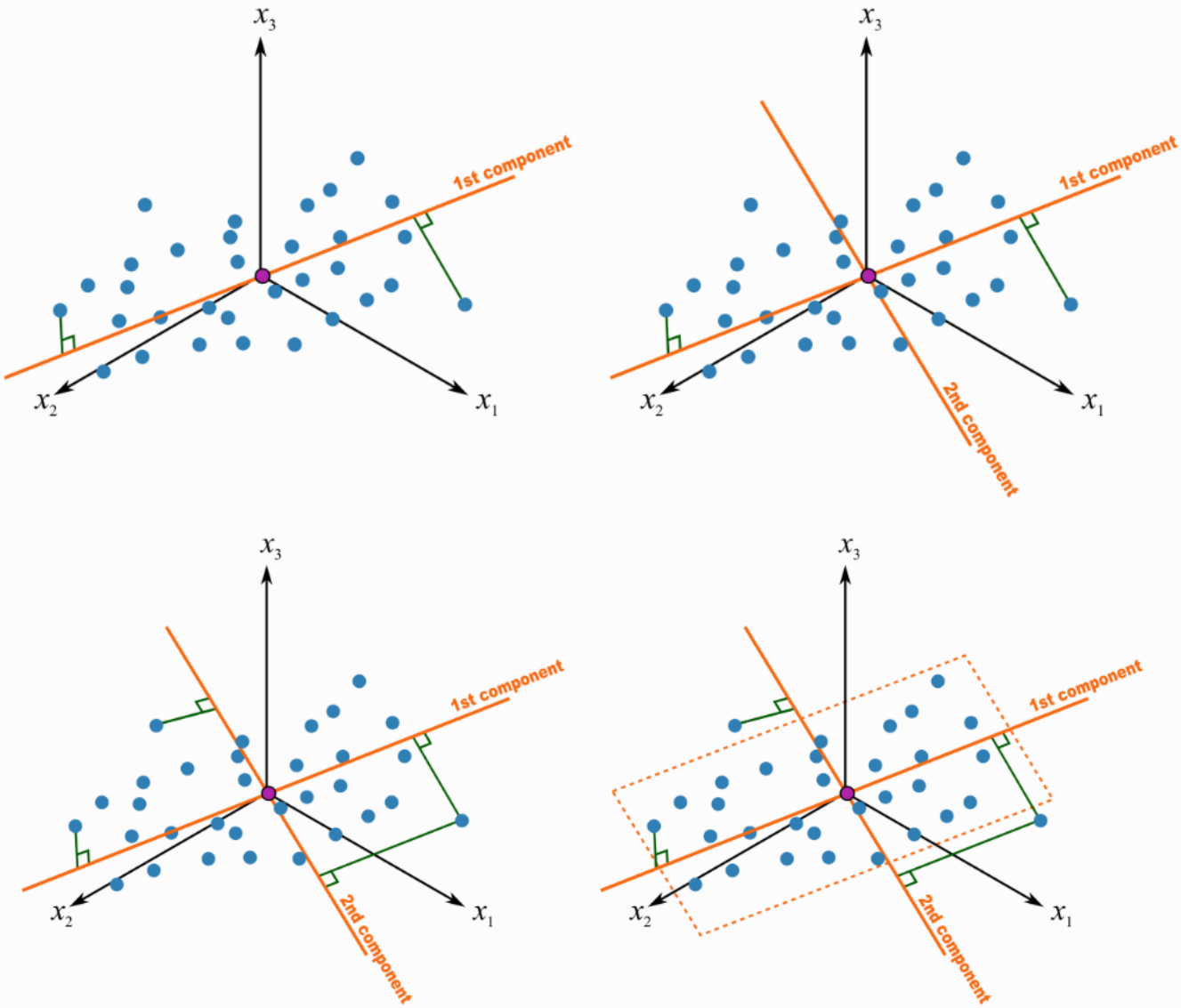


Interpretação Geométrica de uma ACP



https://learnche.org/pid/latent-variable-modelling/principal-component-analysis/geometric-explanation-of-pca

Interpretação Geométrica de uma ACP

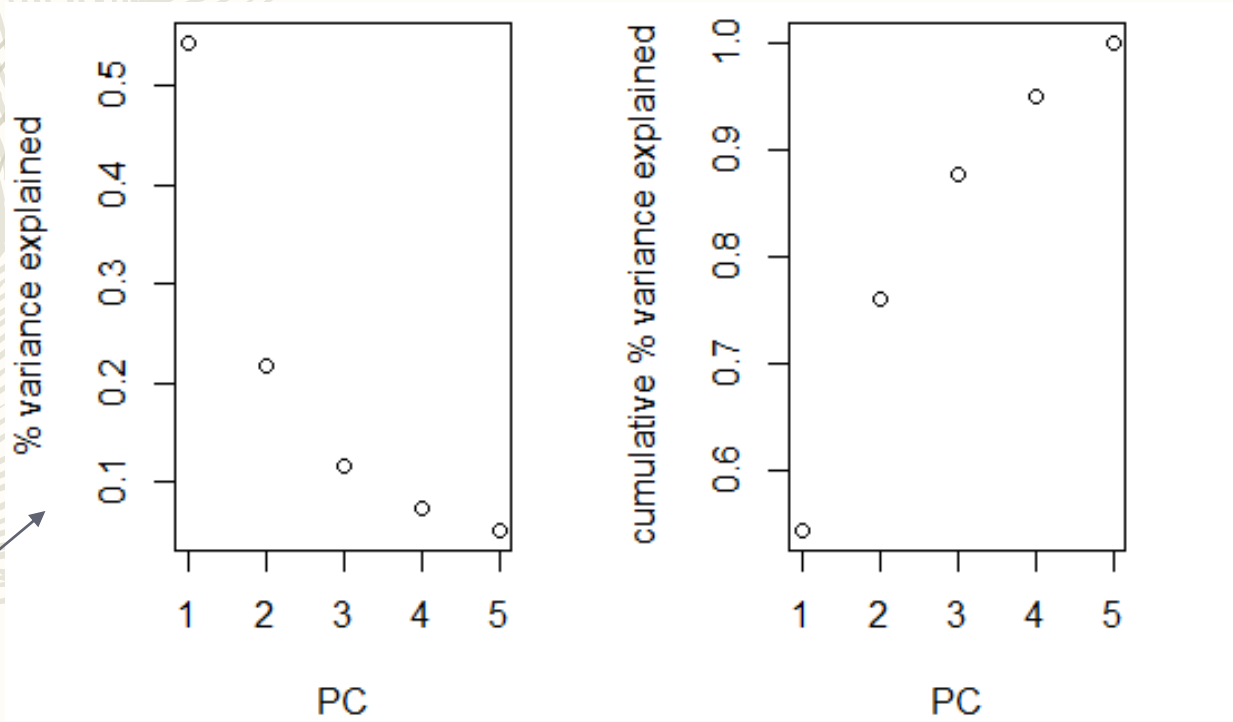


https://learnche.org/pid/latent-variable-modelling/principal-component-analysis/geometric-explanation-of-pca

Implementação no R

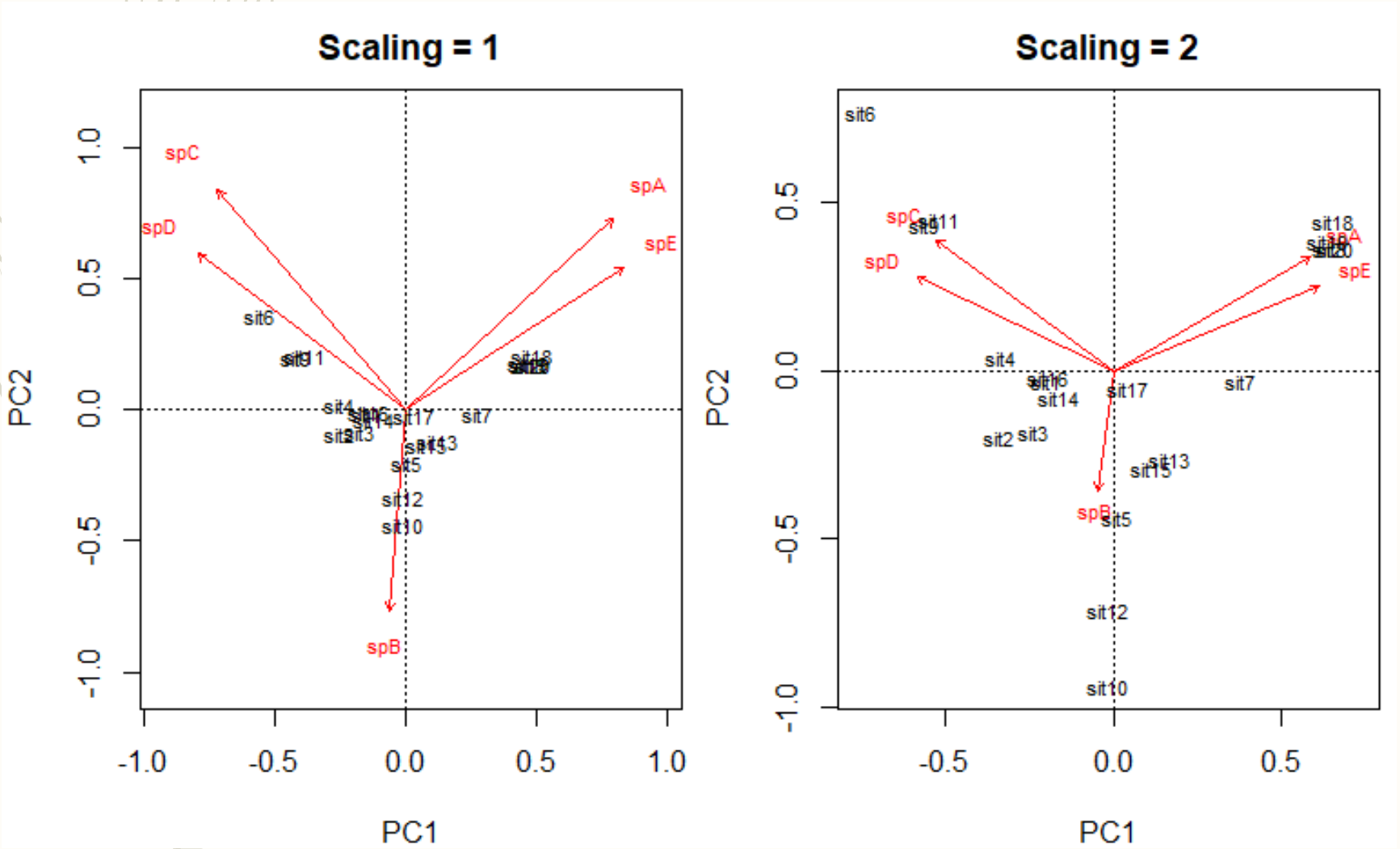
```
option 1
library(vegan)
myPCA=rda(specbyloc)
#option 2
myPCA2=princomp(specbyloc)
```

```
par(mfrow=c(1,2),mar=c(4,4,0.5,2.5))
#variance is the square of the standard deviations
eigval=myPCA2$sdev^2
plot(1:5,eigval/sum(eigval),ylab="% variance explained",xlab="PC")
plot(1:5,cumsum(eigval/sum(eigval)),
ylab="cumulative % variance explained",xlab="PC")
```



Scree diagram

```
par(mfrow=c(1,2),mar=c(4,4,2.5,0.5))
biplot(myPCA,scaling=1,main="Scaling = 1")
biplot(myPCA,scaling=2,main="Scaling = 2")
```



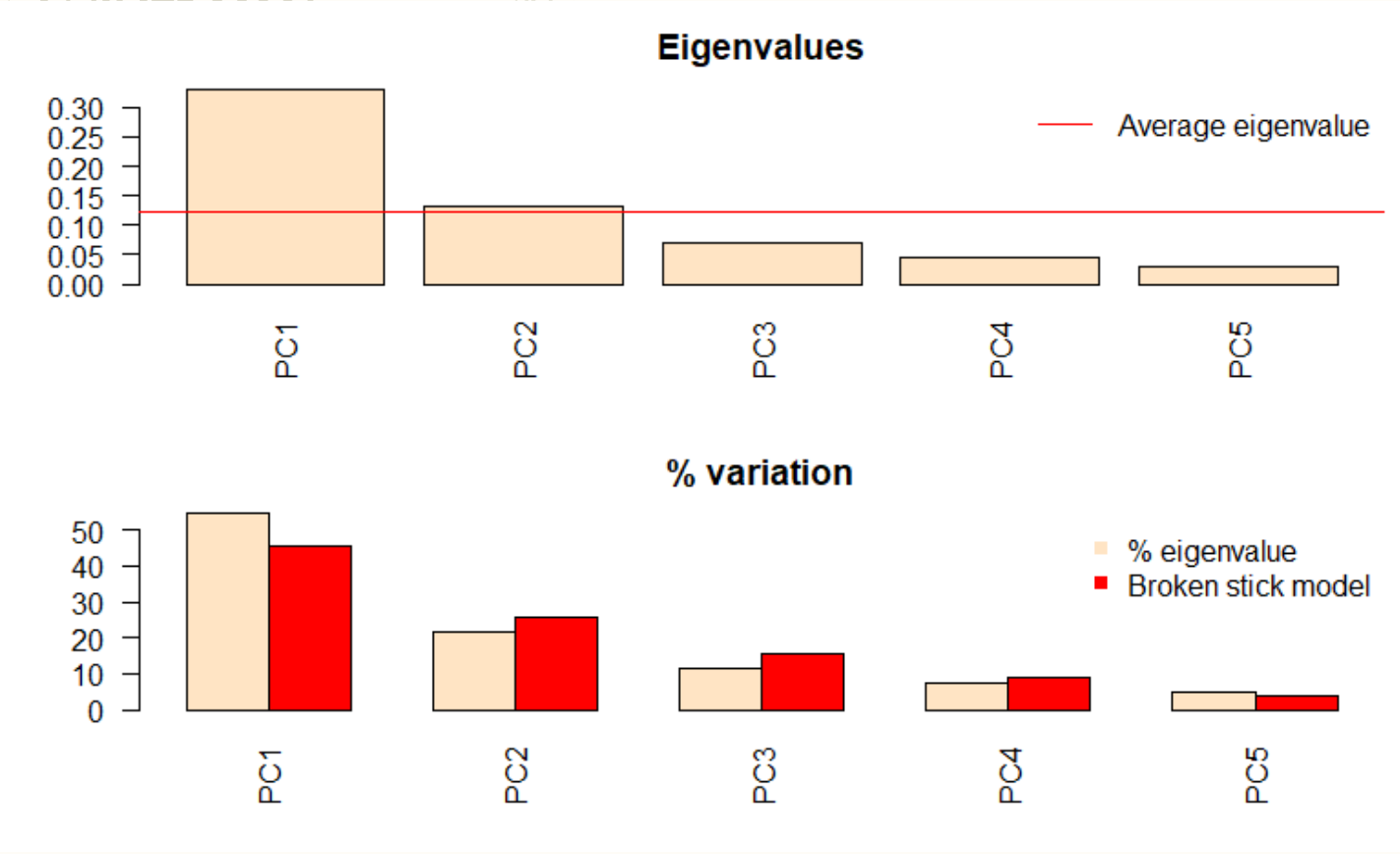
Eigenvectors scaled to unit length. Distances between objects are approximations of their Euclidean distances in multidimensional space. Angles among descriptors are useless.

A different scaling such that angles among descriptors represent correlation between descriptors.

A nice function in brocardfunctions.R – `evplot` – find it under FENIX aula 23.

How many PC's to interpret? (well, pragmatically 2 is all we can deal with really but…)

```
evplot(myPCA$CA$eig)
```



criteria

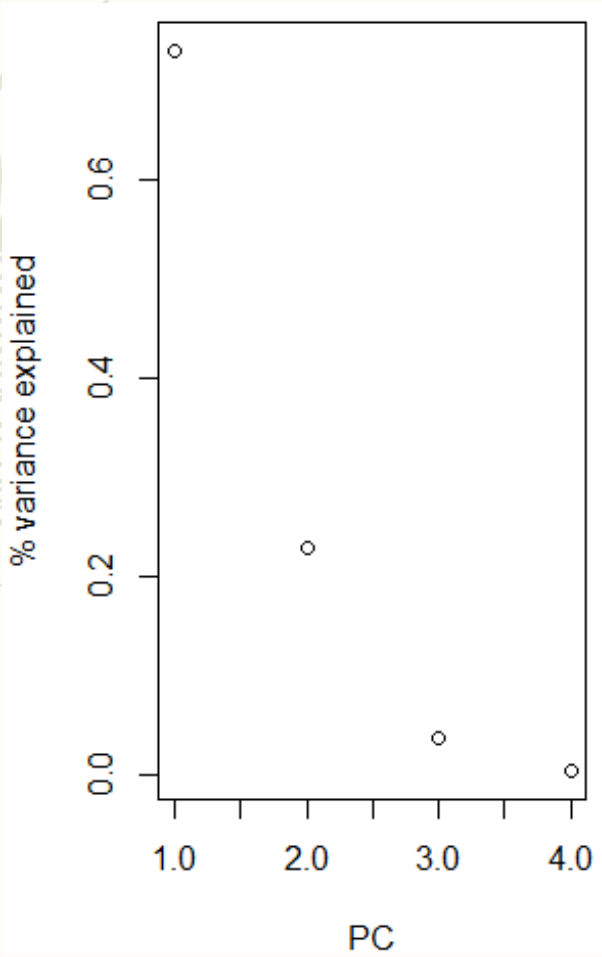
Kaiser-Guttman

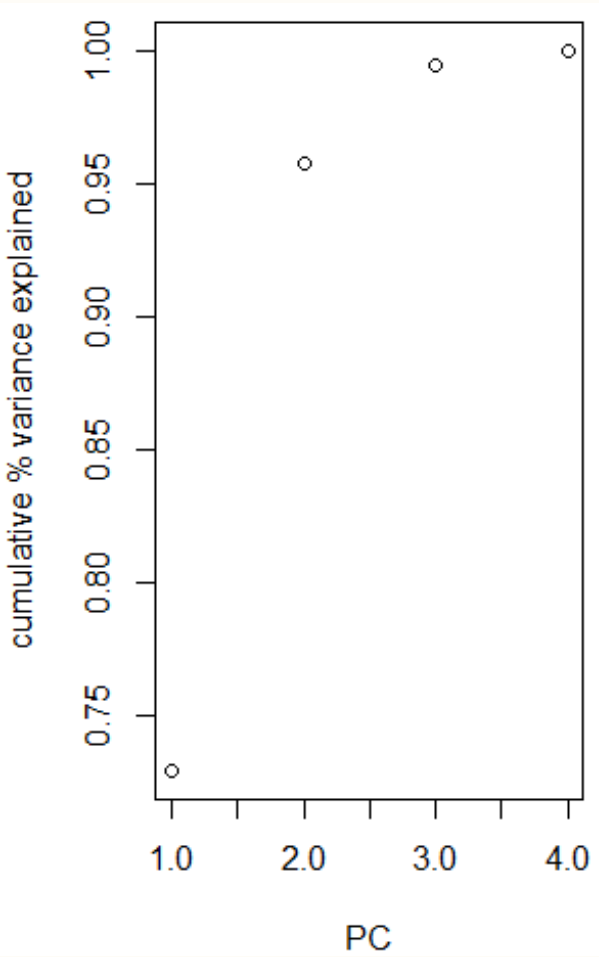
2 PCs

broken-stick

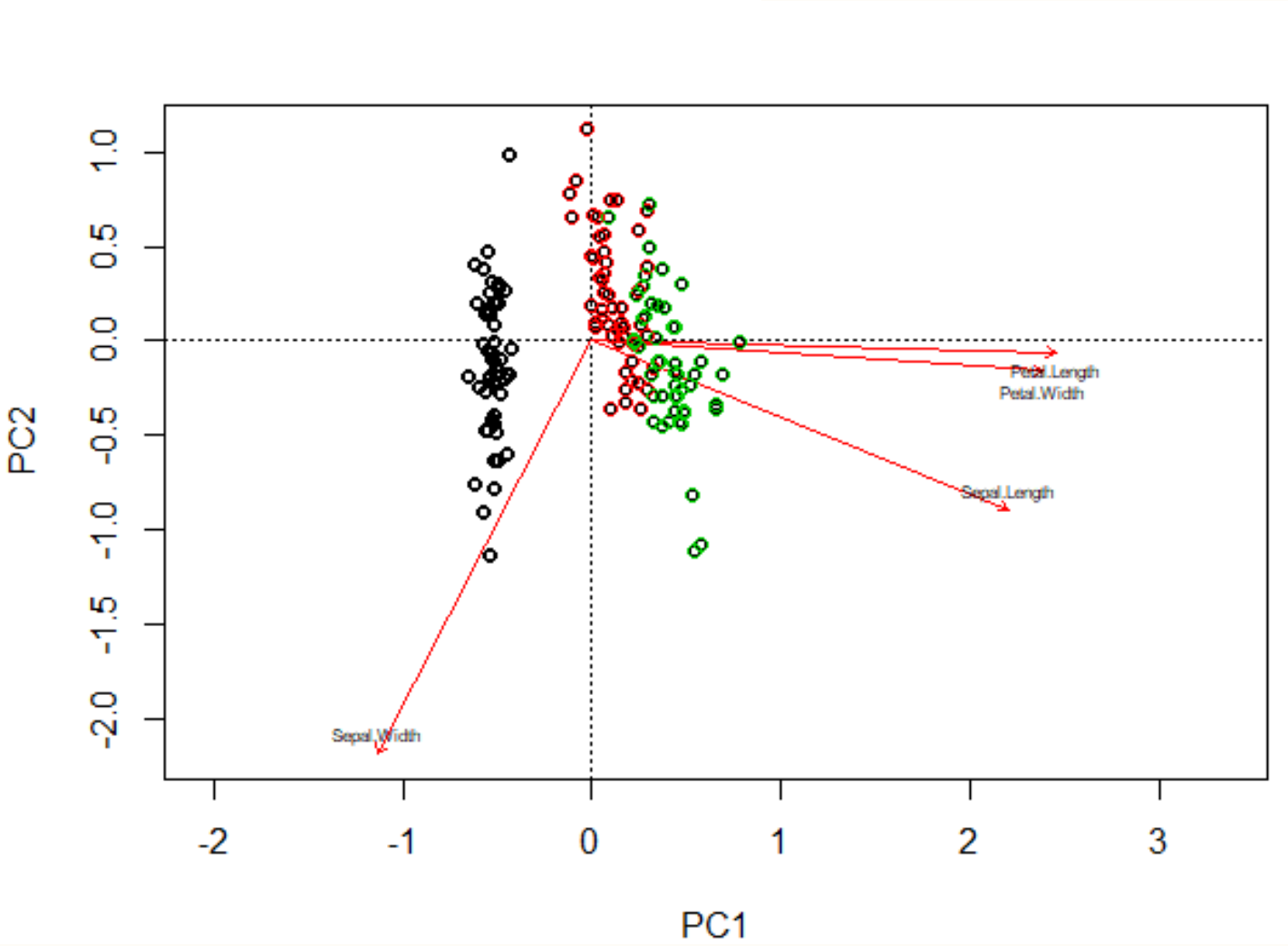
1 PC

```
#another example, now with data iris
data(iris)
myPCAiris=rda(scale(iris[,1:4]))
par(mfrow=c(1,2),mar=c(4,4,0.5,2.5))
eigvali=myPCAiris$CA$eig
plot(1:4,eigvali/sum(eigvali),ylab="% variance explained",xlab="PC")
plot(1:4,cumsum(eigvali/sum(eigvali)),
ylab="cumulative % variance explained",xlab="PC")
```



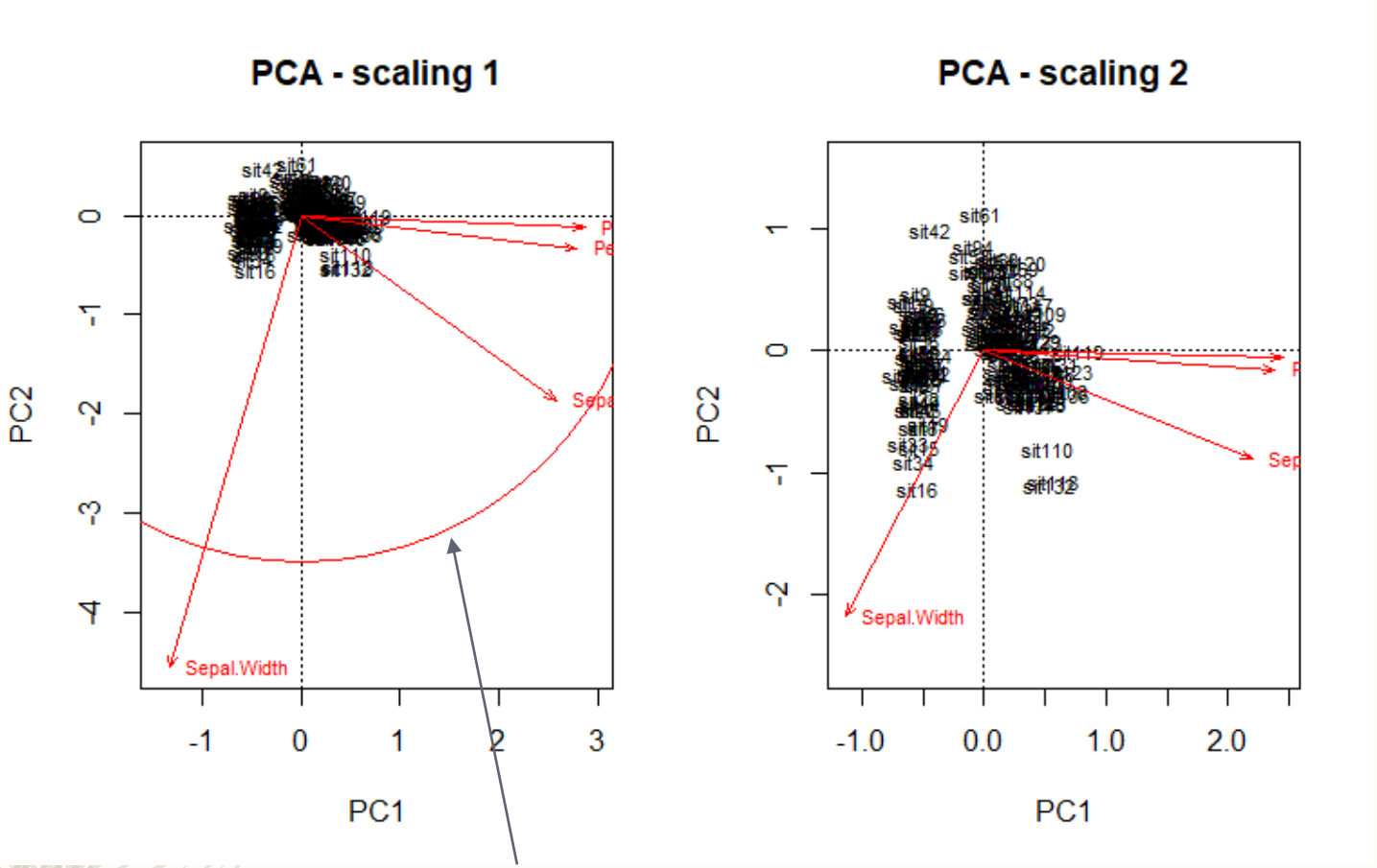


```
par(mfrow=c(1,1),mar=c(4,4,2.5,0.5))
biplot(myPCAiris)
#add colors to points
points(summary(myPCAiris)$sites[,1:2],col=iris[,5])
#for some reason vars names not appearing
text(summary(myPCAiris)$species[,1],
summary(myPCAiris)$species[,2]+c(0.1,0.1,-0.1,-0.1),
names(iris[,1:4]),cex=0.5)
```



Function in file `brocardfunctions.R`

```
cleanplot.pca(myPCAiris)
```



First, the *scaling 1 biplot* displays a feature that must be explained. The circle is called a *circle of equilibrium contribution*. Its radius is equal to $\sqrt{d / p}$, where d is the number of axes represented in the biplot (usually $d=2$) and p is the number of dimensions of the PCA space (i.e. usually the number of variables of the data matrix).[2] The radius of this circle represents the length of the vector representing a variable that would contribute equally to all the dimensions of the PCA space. Therefore, for any given pair of axes, the variables that have vectors longer than this radius make a higher contribution than average and can be interpreted with confidence.

`Sepal.Width` seems the most important variable in this case