ANÁLISE DE COMPONENTES PRINCIPAIS - PCA

A ideia básica da PCA

- Se forem coletados dados que formem uma matriz de *n* amostras/parcelas por *m* variáveis/espécies, haverá uma grande quantidade de correlações na variabilidade das variáveis das amostras.
- Por exemplo, uma matriz de 100 amostras e 50 variáveis pode ser reduzida a 100 amostras em 5 ou menos componentes.
- Esses componentes podem ser considerados como “super-variáveis” feitas de combinações altamente correlacionadas das 50 variáveis iniciais.

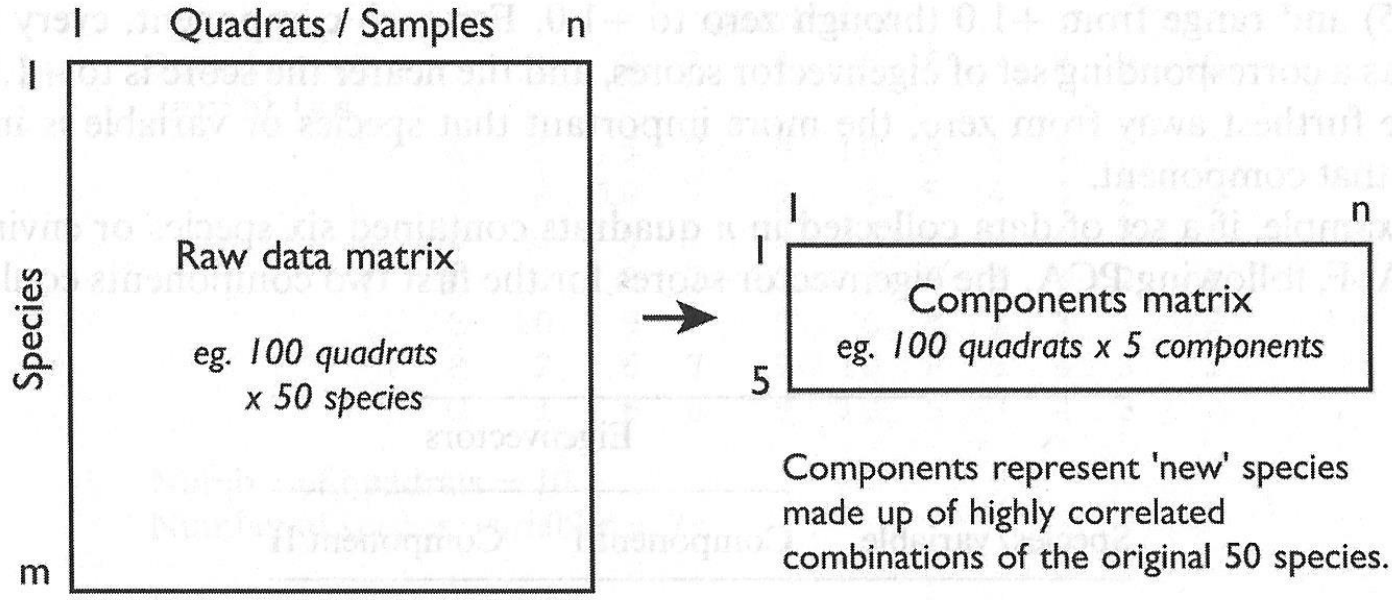


Figure 6.16 Reduction of many species or environmental/biotic variables into a few components.

Características

- Uma das mais importantes características dos componentes é que, como todas as variáveis são altamente relacionadas entre si, os novos componentes são completamente não correlacionados, ou seja, são ortogonais.
- Como na matriz original cada amostra apresenta um valor para cada variável, na nova matriz componente, cada amostra tem um valor para cada componente, que tomados juntos são conhecidos como valores dos componentes.
- O núcleo de qualquer PCA são autovetores e autovalores.

Autovetores

- São conjuntos de valores que representam o peso de cada variável original sobre cada componente.
- Os autovetores são escalados como coeficientes de correlação e variam de +1,0 a -1,0 (passando pelo zero).
- Para cada componente, todas as variáveis têm um conjunto de autovetores correspondentes, e quanto mais próximo de +1,0 ou -1,0 está o autovetor, mais importante é a variável para o componente.

Exemplo

As an example, if a set of data collected in *n* quadrats contained six species or environmental variables A–F, following PCA, the eigenvector scores for the first two components could be:

	Eigenvectors	
Species/variable	Component I	Component II
A	0.91	0.21
B	0.79	0.19
C	0.75	0.29
D	0.01	−0.90
E	−0.12	−0.86
F	−0.01	−0.93

Autovalores

- São valores que representam a contribuição relativa de cada componente na explicação da variação total dos dados.
- Existe um autovalor para cada componente, e o tamanho do autovalor para o componente é uma indicação direta da importância do componente na explicação da variação total dentro do conjunto de dados. Ou seja, o autovalor explica a importância do componente sobre a variação total dos dados.

Explicação geométrica para a PCA

- O ponto inicial de uma PCA é uma matriz com amostras e variáveis:

Species or variables	Quadrats 1	2	3	4	5	6	7	8	9	10
A	1	5	7	9	10	8	6	4	3	2
B	8	10	7	9	6	5	4	3	1	2
C	3	6	7	9	10	8	5	4	2	1
D	4	8	7	10	9	6	5	3	2	1
E	10	9	7	8	6	5	4	3	1	2
F	2	6	7	9	10	8	5	4	3	1
G	1	6	8	9	10	5	7	4	3	2

Number of quadrats = 10
Number of species/variables = 7

Análise em modo R – ordenar amostras

- Primeiramente os dados devem ser padronizados, expressando cada variável em unidades de desvio padrão em relação a uma média igual a zero.
- Se os dados não são padronizados a análise será tendenciosa, favorável às variáveis com mais alta variância.
- Uma alternativa é usar a matriz de correlações como medida de similaridade.

Matriz de correlações – *Pearson*

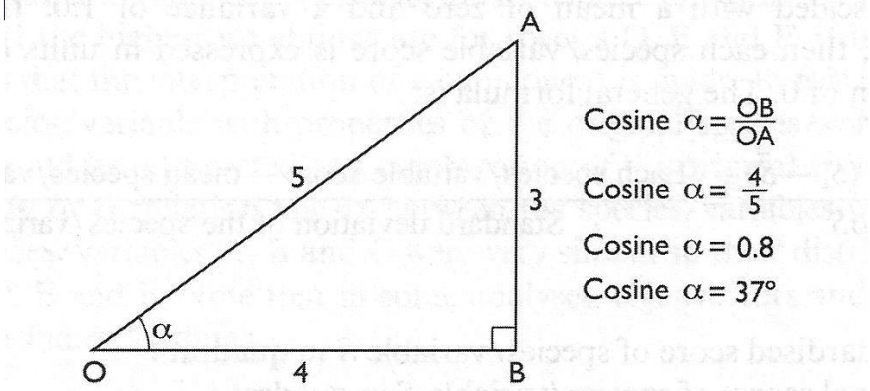
Species/variable

A	1.00						
B	0.35	1.00					
C	0.95	0.58	1.00				
D	0.83	0.79	0.93	1.00			
E	0.20	0.96	0.47	0.67	1.00		
F	0.98	0.49	0.99	0.90	0.36	1.00	
G	0.93	0.42	0.88	0.85	0.26	0.90	1.00
	A	B	C	D	E	F	G

Species/variable

O coeficiente de correlação é dado pelo cosseno do ângulo em um triângulo retângulo

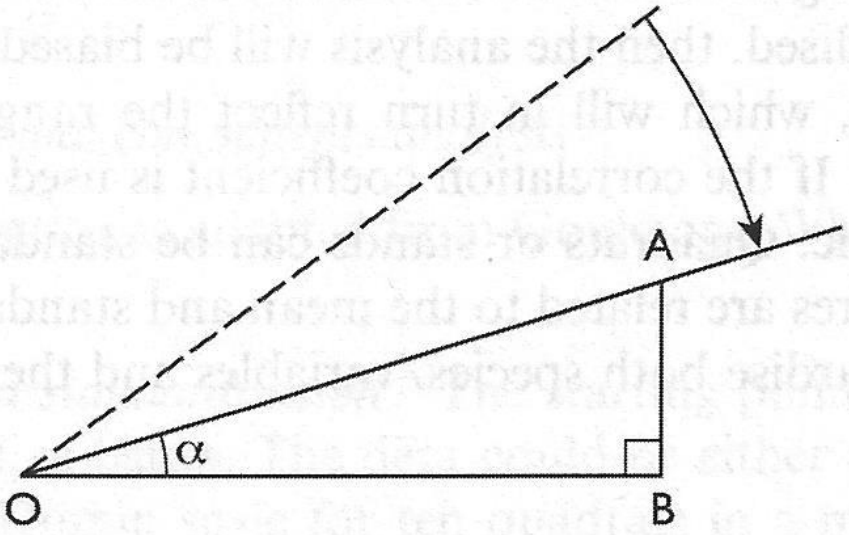
- Fundamental para o entendimento geométrico da PCA é a ideia de que o coeficiente de correlação pode ser expresso como o cosse



- Em um triângulo retângulo, o cosseno de um ângulo é definido como a razão entre o comprimento do lado adjacente ao ângulo e o comprimento da hipotenusa.

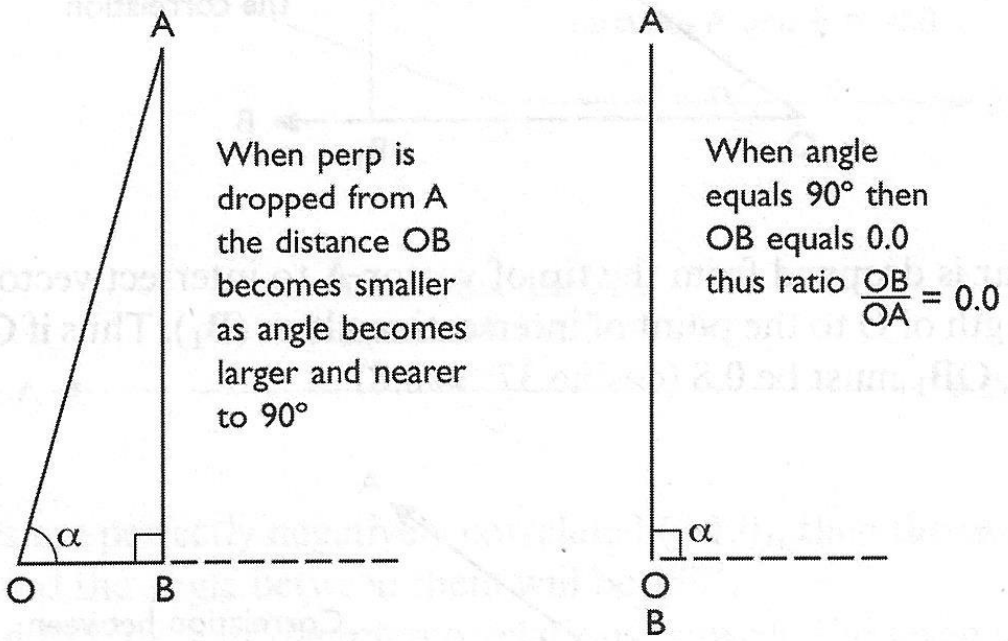
O próximo passo

- Movendo-se AO para baixo diminui-se o ângulo alfa e os dois lados AO e OB tornam-se mais similares até serem idênticos (alfa = zero).



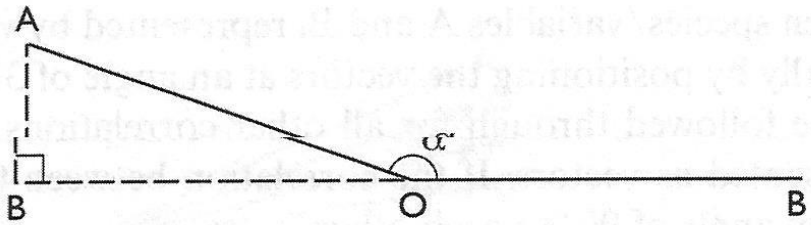
Outro passo

- Mover AO a partir de OB até que AO fique quase na vertical. Se alfa igual a 90º a distância OB será zero.



Passo 4

- Se AO é movimentado além do ponto O, o ângulo alfa torna-se maior do que 90º e o cosseno torna-se negativo, aproximando-se de -1,0 (cosseno 180º = -1,0).



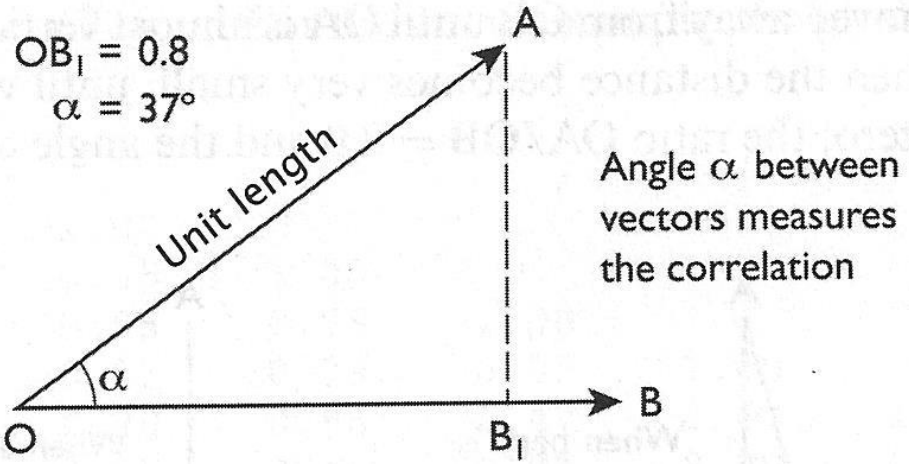
O vetor

- De acordo com o exposto, o cosseno, expresso como uma razão entre dois lados de um triângulo retângulo, varia de -1,0 a +1,0, como o coeficiente de correlação. Isso torna possível representar correlações entre variáveis geometricamente.
- Cada variável na análise pode ser considerada um vetor.
- Um vetor é uma linha que possui propriedades de comprimento e direção.

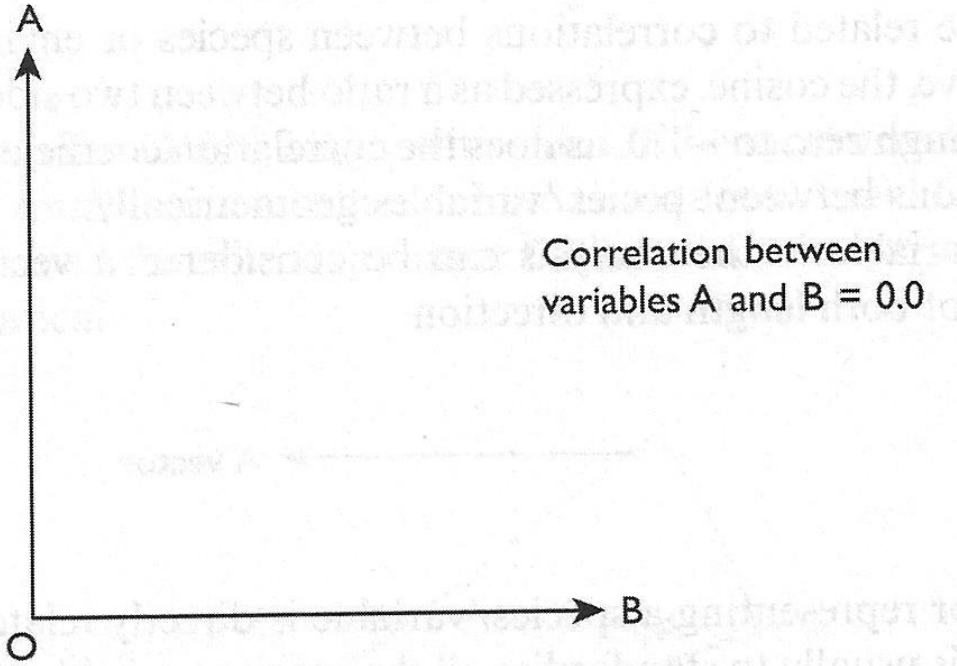
O vetor

O vetor

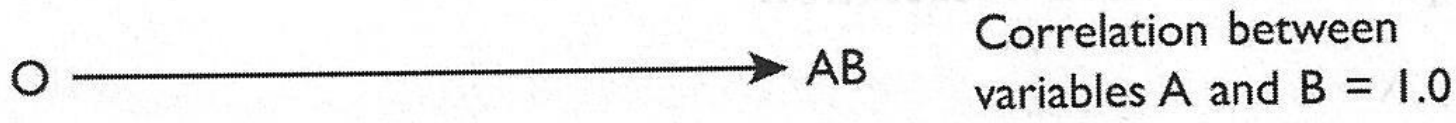
- O comprimento de um vetor representando uma variável está diretamente relacionado à sua variância, mas como o primeiro passo na PCA geralmente é padronizar as variáveis à variância 1,0, os vetores representando todas as variáveis na análise possuirão o mesmo comprimento: 1,0.



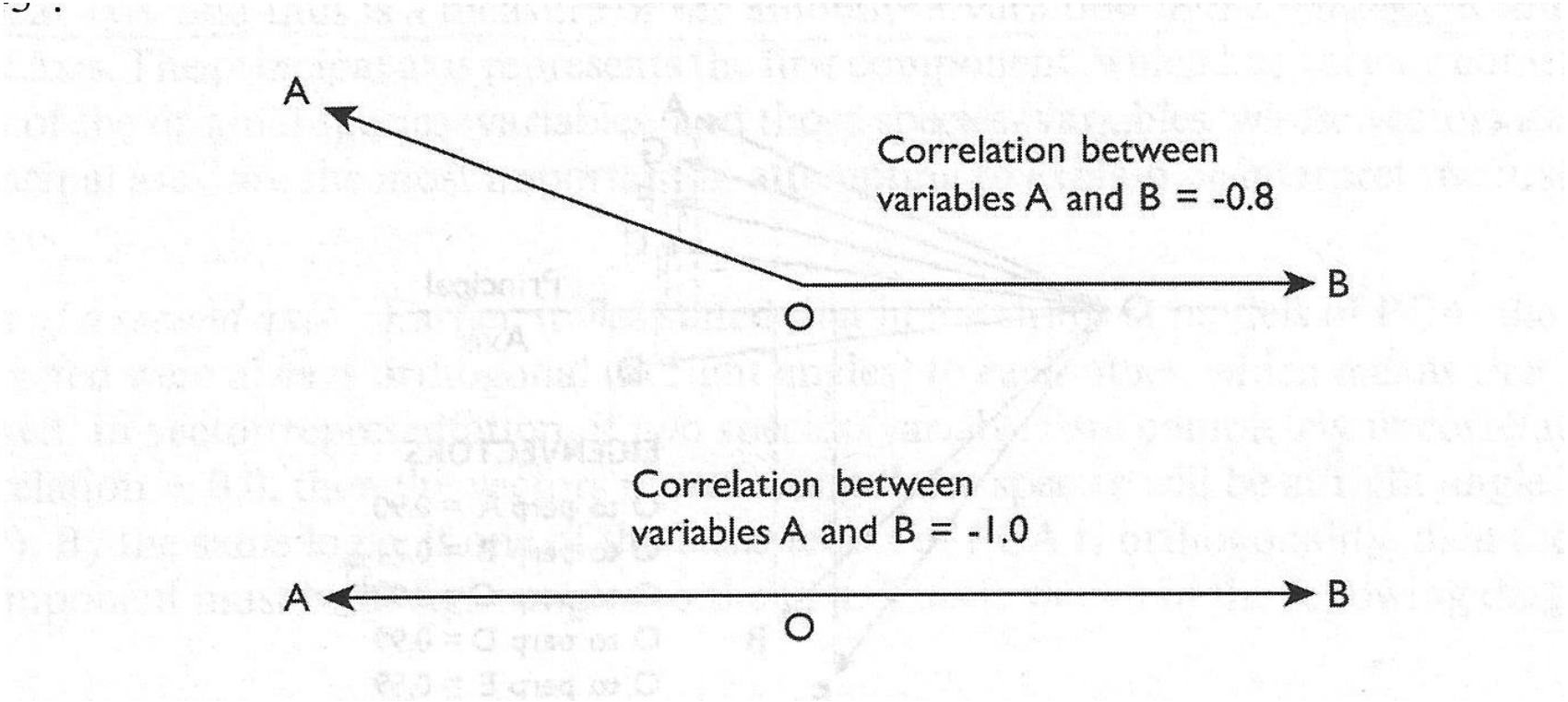
Correlação entre duas variáveis A e B



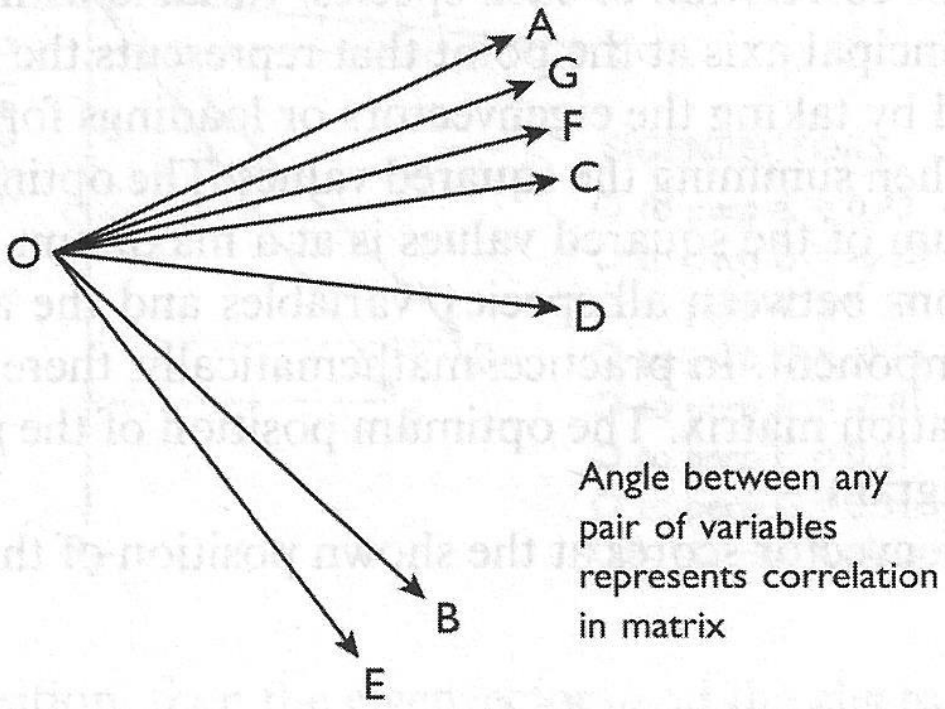
Correlação entre duas variáveis A e B



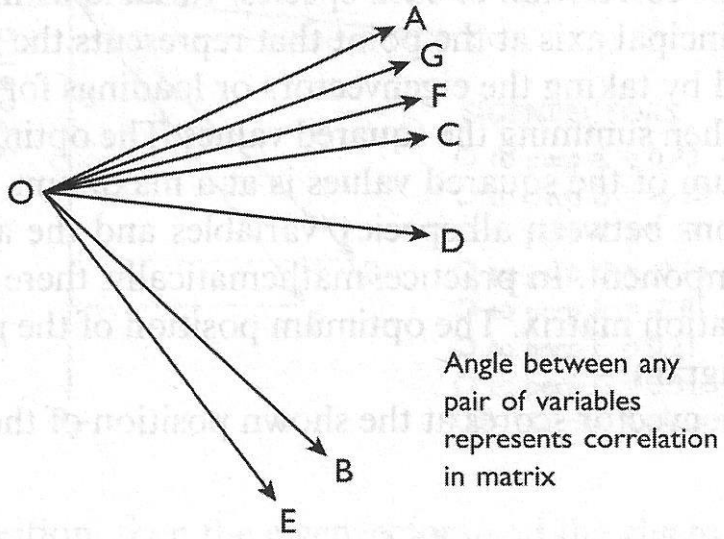
Correlação entre duas variáveis A e B



Intercorrelações entre muitas variáveis



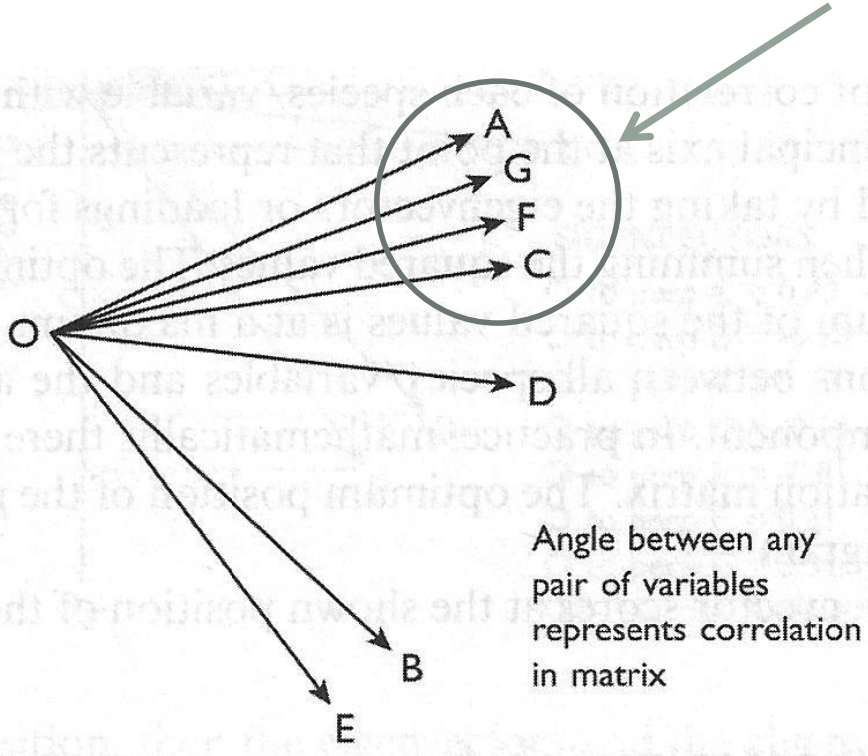
Intercorrelações entre muitas variáveis



- Note que as correlações deveriam estar representadas em sete dimensões – uma para cada variável – mas, como é impossível visualizar mais do que três dimensões, para propósitos de explicação, as correlações foram cuidadosamente escolhidas de modo a serem representadas em duas dimensões.

O eixo principal

Variáveis altamente correlacionadas
tendem a ficar juntas e na mesma direção



Species/variable

A	1.00						
B	0.35	1.00					
C	0.95	0.58	1.00				
D	0.83	0.79	0.93	1.00			
E	0.20	0.96	0.47	0.67	1.00		
F	0.98	0.49	0.99	0.90	0.36	1.00	
G	0.93	0.42	0.88	0.85	0.26	0.90	1.00
	A	B	C	D	E	F	G

Species/variable

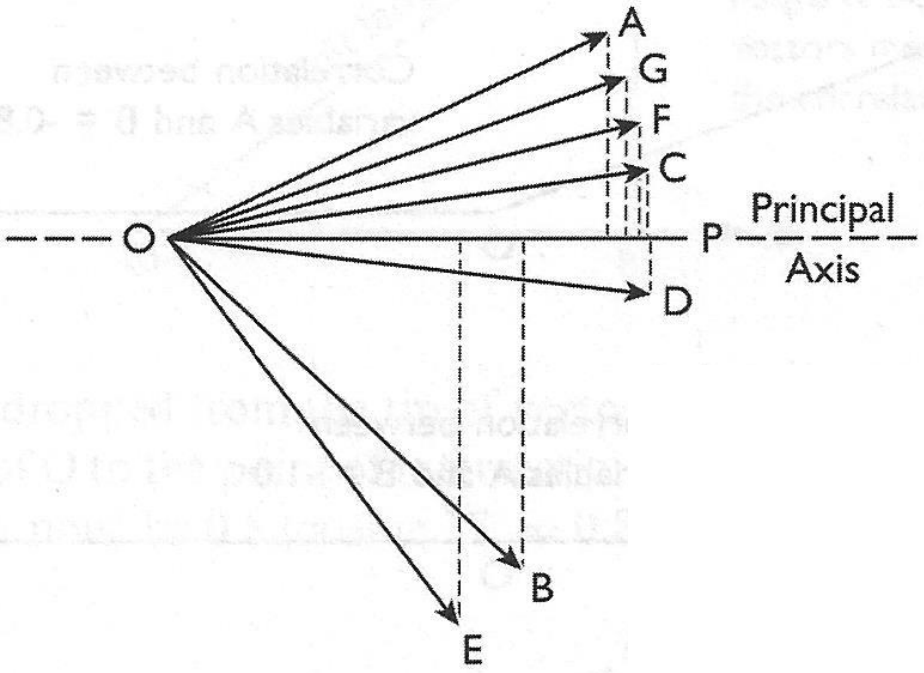
O eixo principal

- O objetivo da análise em componentes é identificar o sentido geral dos vetores, passando um eixo através da origem em comum, de forma que cada vetor que representa uma variável apresente um ângulo reto com o eixo.

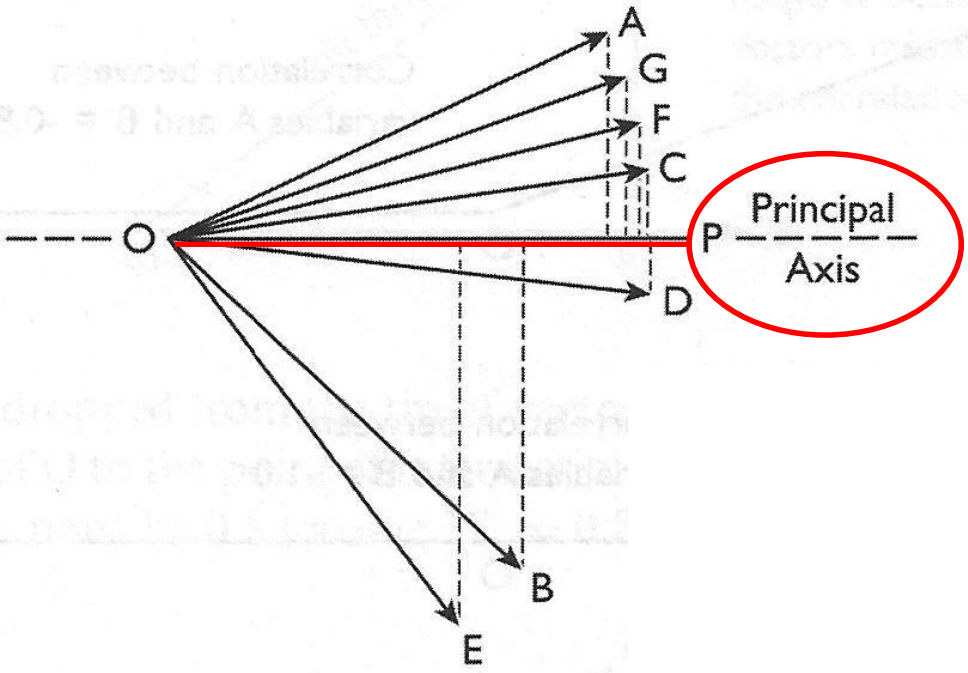
O eixo principal

- O objetivo da análise em componentes é identificar o sentido geral dos vetores, passando um eixo através da origem em comum, de forma que cada vetor que representa uma variável apresente um ângulo reto com o eixo.
- Como os vetores têm comprimento unitário, a projeção sobre o eixo não será maior do que o cosseno do ângulo entre os vetores e o eixo principal

O eixo principal



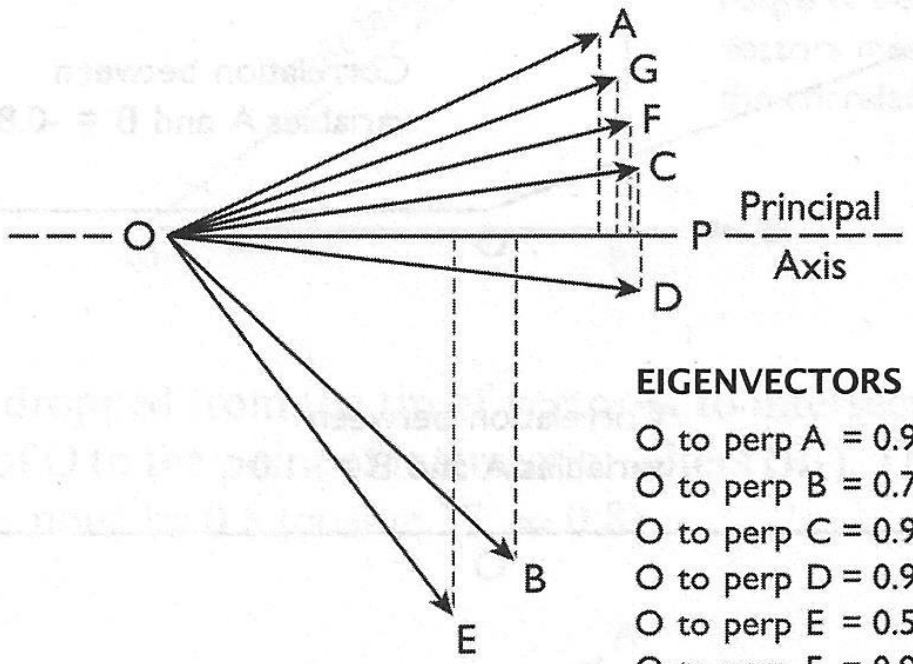
O eixo principal



Como traçar o eixo principal e como saber quando ele mostra a direção que está fortemente relacionada a todos os vetores?

- O eixo principal é oscilado lentamente ao redor da origem (O) e a cada instante, o comprimento das projeções perpendiculares de todos os vetores sobre o eixo são medidas.
- Essas projeções são os autovetores, ou as cargas de todas as variáveis sobre o eixo, e são vistas como equivalentes ao nível de correlação de cada variável com o eixo principal.

Como traçar o eixo principal e como saber quando ele mostra a direção mais fortemente relacionada a todos os vetores?



EIGENVECTORS

O to perp A = 0.90
O to perp B = 0.71
O to perp C = 0.97
O to perp D = 0.99
O to perp E = 0.59
O to perp F = 0.95
O to perp G = 0.92
OP (Principal Axis) = 1.0

A posição ótima do eixo

- A posição que representa a máxima direção no sentido de todos os vetores é encontrada somando-se o quadrado dos valores.
- A posição ótima do eixo principal será a que maximiza a soma dos quadrados dos autovetores.
- Esse valor representa a soma de quadrados das correlações entre todas as variáveis e o eixo. E é o autovalor do eixo principal, o primeiro componente.

O autovalor

Species/variable	Eigenvector
A	0.90
B	0.71
C	0.97
D	0.99
E	0.59
F	0.95
G	0.92

The eigenvalue is thus: $(0.90)^2 + (0.71)^2 + (0.97)^2 + (0.99)^2 + (0.59)^2 + (0.95)^2 + (0.92)^2 = 5.33$.

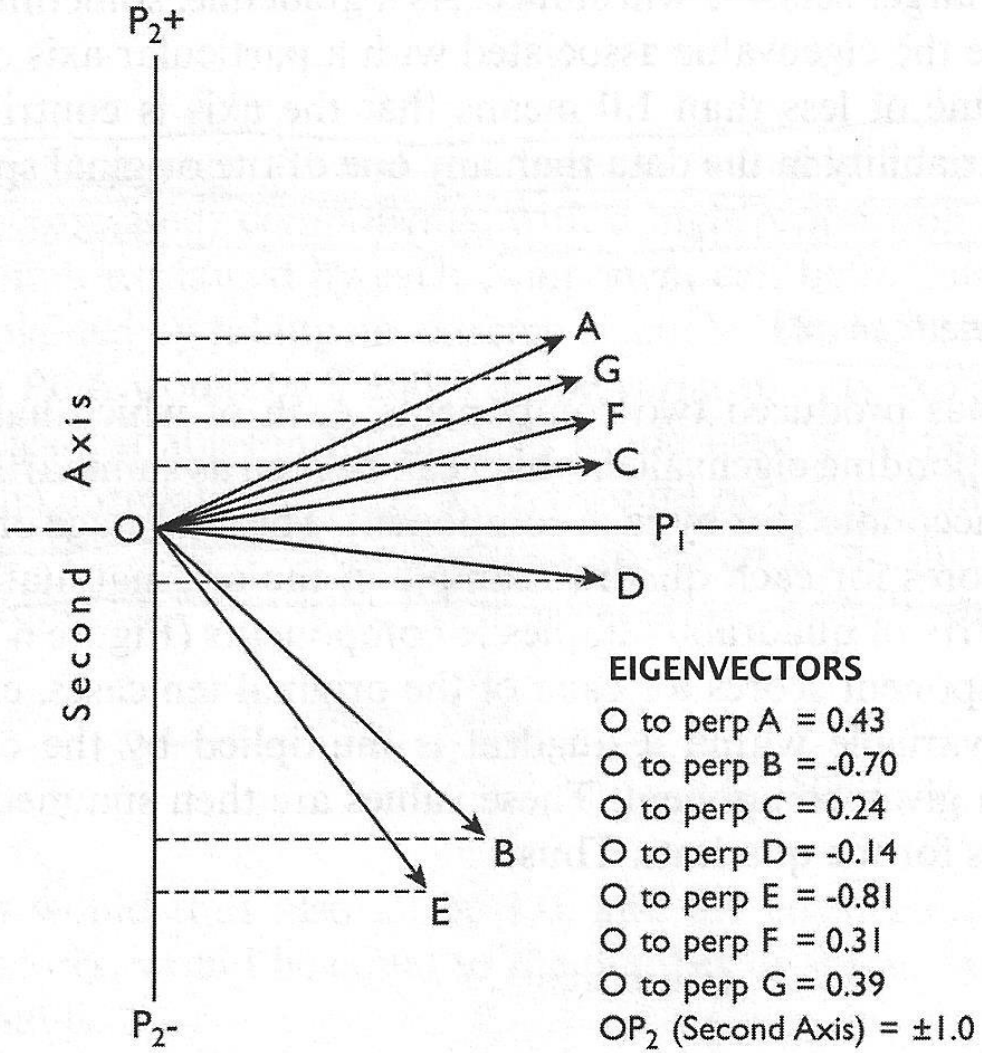
- O autovalor representa o nível mais alto possível de correlação de todas as variáveis com o eixo principal, e portanto, é uma medida da variação total do conjunto de dados representada pelo eixo.

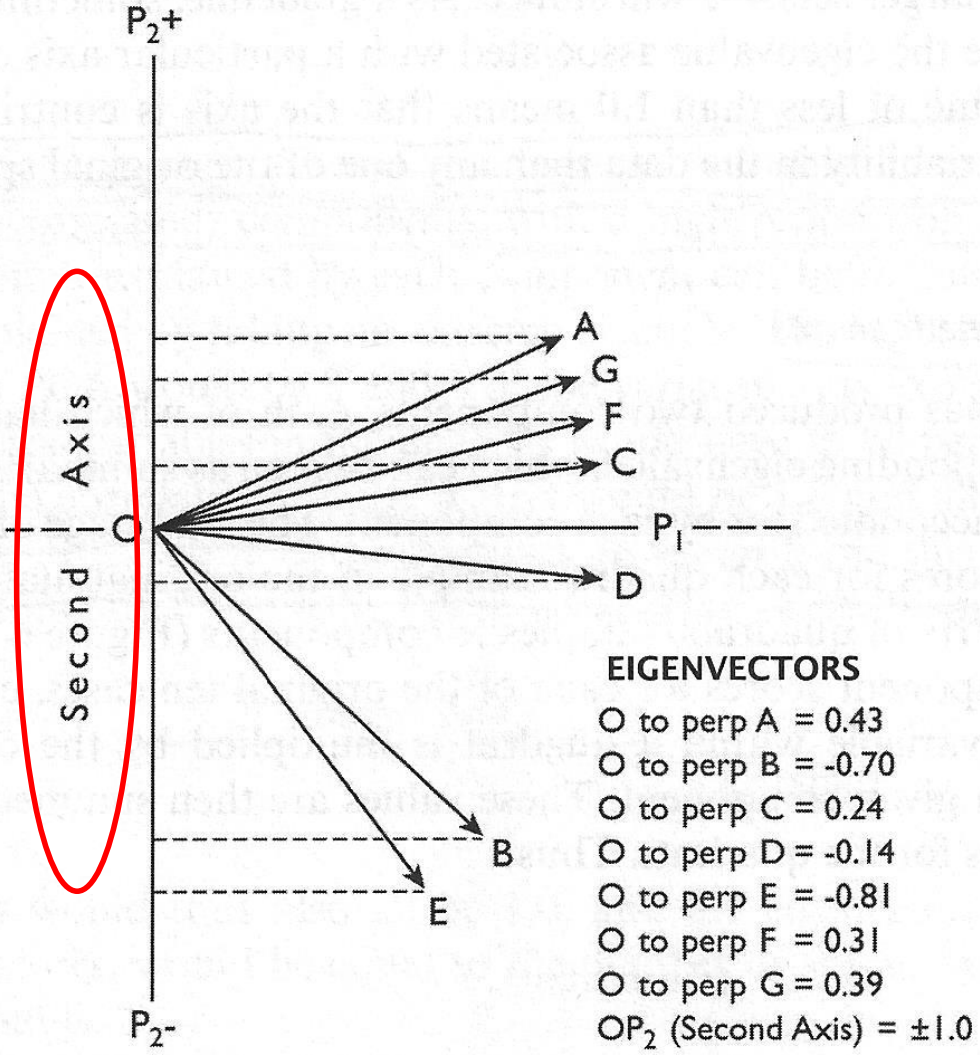
O autovalor

- O eixo principal representa o primeiro componente, que tem diferentes contribuições de cada uma das variáveis originais.
- As variáveis, cujos vetores estão mais próximos do eixo principal, são as mais importantes para tentar explicar ou interpretar o primeiro eixo.

O segundo eixo

- Na representação vetorial, se duas variáveis não estão correlacionadas entre si, a correlação é zero e os vetores devem ficar posicionados a um ângulo de 90º entre si.
- Dessa forma, o segundo eixo deve ser ortogonal ao primeiro, ou seja, não correlacionado com este.


P2+
A
G
F
C
O
P1
D
Second Axis
B
E
P2-
EIGENVECTORS
O to perp A = 0.43
O to perp B = -0.70
O to perp C = 0.24
O to perp D = -0.14
O to perp E = -0.81
O to perp F = 0.31
O to perp G = 0.39
OP2 (Second Axis) = ±1.0


P2+
Second Axis
O
A
G
F
C
P1
D
B
E
P2-
EIGENVECTORS
O to perp A = 0.43
O to perp B = -0.70
O to perp C = 0.24
O to perp D = -0.14
O to perp E = -0.81
O to perp F = 0.31
O to perp G = 0.39
OP2 (Second Axis) = ±1.0

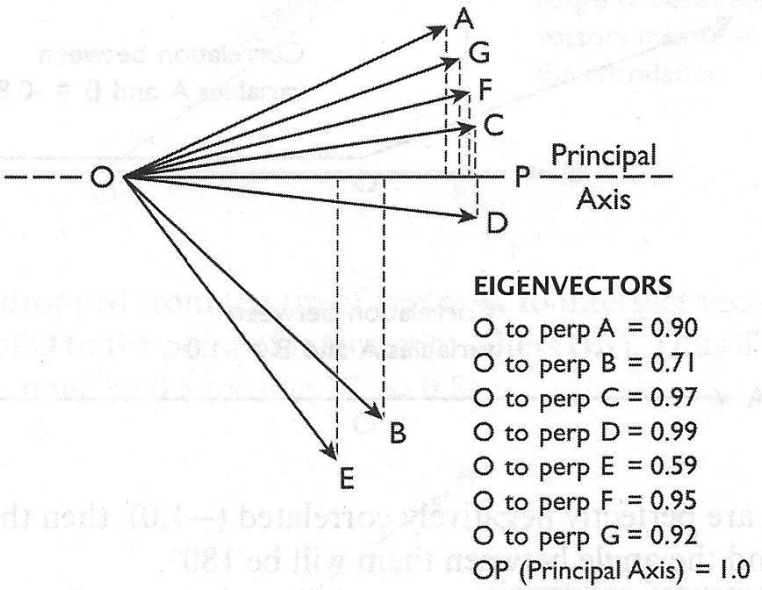
Species/variable	Eigenvector
A	0.43
B	−0.70
C	0.24
D	−0.14
E	−0.81
F	0.31
G	0.39

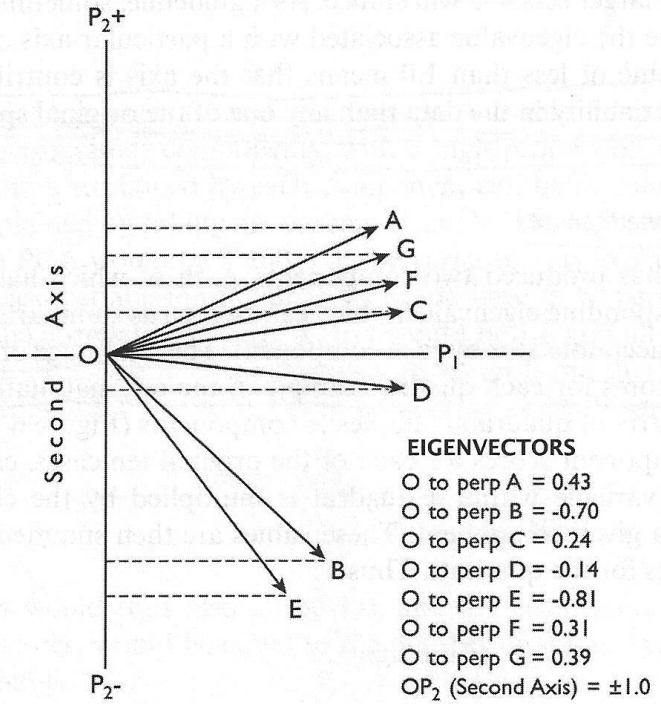
Autovalor = 1,66

- Notar que o segundo autovalor é muito mais baixo que o primeiro eixo.
- Uma das características da PCA é que os eixos são extraídos em ordem descendente de importância em termos de sua contribuição à variação total nos dados.

Características

- Notar que as variáveis que tinham autovetores altos sobre o primeiro eixo tendem a ter menor autovetor no segundo eixo.





Mais eixos

- Outros eixos podem ser extraídos exatamente da mesma maneira, sempre em ângulo reto com os demais.
- Matematicamente o número de eixo pode ser igual ao número de variáveis, mas eixos mais altos contribuem pouco com a explicação.
- Como regra geral, eixos são extraídos até que o eixo associado apresente autovalor ≥ 1,0.

Cálculo do valor do componente

- A análise produziu dois componentes, cada um com um conjunto de autovetores e um autovalor.
- A etapa final da PCA é calcular o valor do componente para cada amostra na matriz original.
- É com esses valores que se constrói a matriz reduzida de amostras e componentes.

Cálculo do valor do componente

- A análise produziu dois componentes, cada um com um conjunto de autovetores e um autovalor.
- A etapa final da PCA é calcular o valor do componente para cada amostra na matriz original.
- É com esses valores que se constrói a matriz reduzida de amostras e componentes.

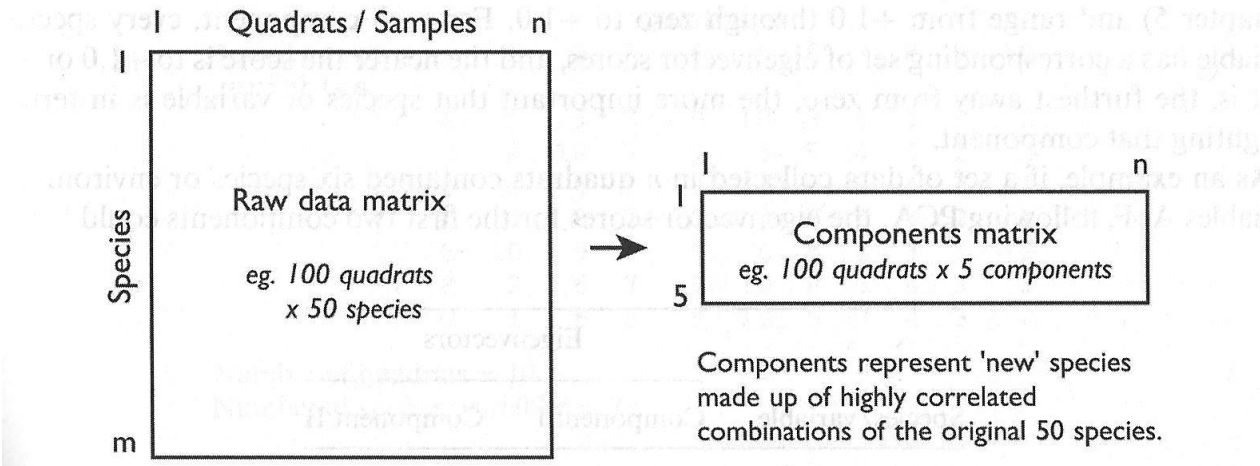


Figure 6.16 Reduction of many species or environmental/biotic variables into a few components.

A matriz reduzida

- Para obter o valor componente para cada uma das dez amostras originais, cada valor original, de cada variável, dentro de uma amostra, é multiplicado pelo autovetor dessa variável sobre um dado componente.
- A soma desses valores em cada amostra corresponde ao valor do componente.

	Quadrats										Eigenvectors	
Species or variables	1	2	3	4	5	6	7	8	9	10	Component I	Component II
A	1	5	7	9	10	8	6	4	3	2	0.90	0.43
B	8	10	7	9	6	5	4	3	1	2	0.71	−0.70
C	3	6	7	9	10	8	5	4	2	1	0.97	0.24
D	4	8	7	10	9	6	5	3	2	1	0.99	−0.14
E	10	9	7	8	6	5	4	3	1	2	0.59	−0.81
F	2	6	7	9	10	8	5	4	3	1	0.95	0.31
G	1	6	8	9	10	5	7	4	3	2	0.92	0.39

To obtain the component score for quadrat 1 on component I:

$$\text{Score} = (1 \times 0.90) + (8 \times 0.71) + (3 \times 0.97) + (4 \times 0.99) + (10 \times 0.59) + (2 \times 0.95) + (1 \times 0.92) = 22.17.$$

Matriz de componentes e amostras

Components matrix

	Quadrats									
Components	1	2	3	4	5	6	7	8	9	10
I	22.17									
II										

Matriz de componentes e amostras

Components matrix

	Quadrats									
Components	1	2	3	4	5	6	7	8	9	10
I	22.17	41.80	42.90	54.40	53.80	39.50	31.40	21.70	13.40	9.10
II	−12.10	−7.60	−2.00	−1.80	−3.40	1.40	1.30	0.50	2.10	−0.97

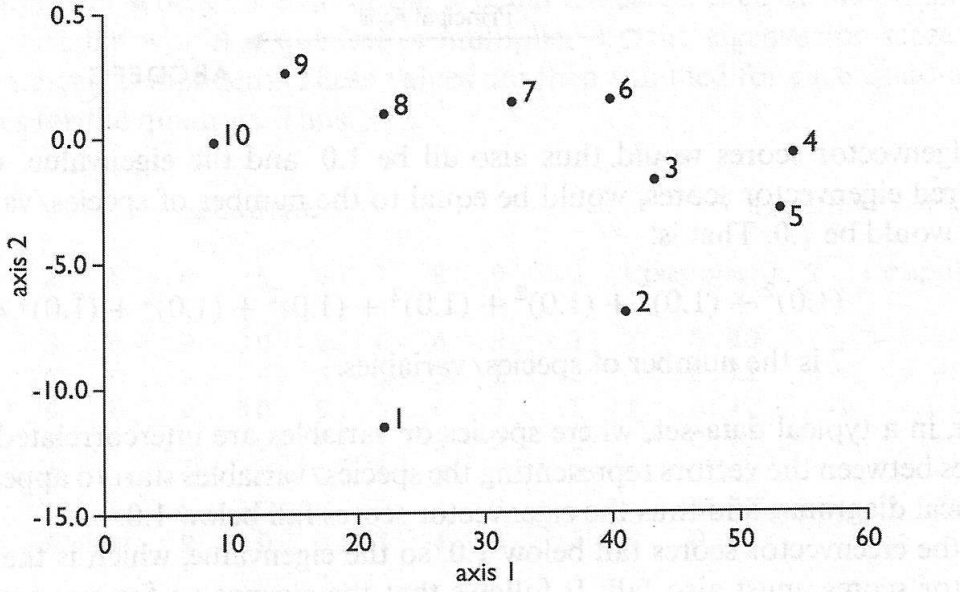
- A matriz de 10 amostras e sete variáveis foi reduzida a dez amostras com dois componentes ortogonais, com uma alta proporção da variação total dos dados explicada por esses componentes.
- A proporção exata explicada por cada componente pode ser calculada da seguinte forma:

O diagrama de ordenação

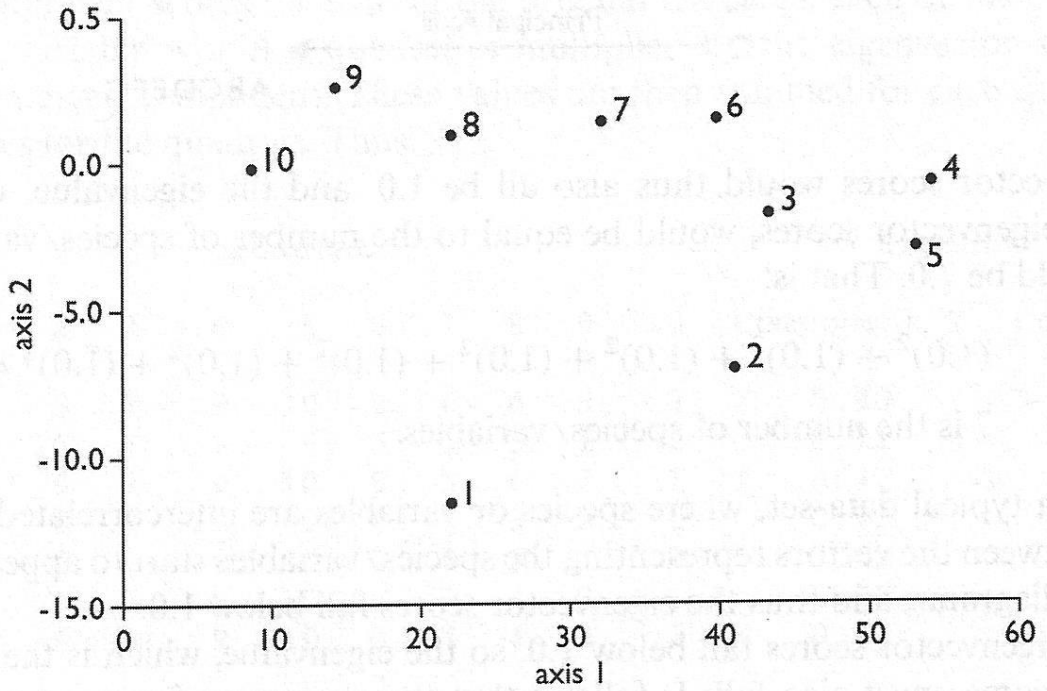
- O diagrama de ordenação é construído plotando-se as amostras em função da matriz reduzida em dois componentes.

Components matrix

	Quadrats									
Components	1	2	3	4	5	6	7	8	9	10
I	22.17	41.80	42.90	54.40	53.80	39.50	31.40	21.70	13.40	9.10
II	−12.10	−7.60	−2.00	−1.80	−3.40	1.40	1.30	0.50	2.10	−0.97



A interpretação



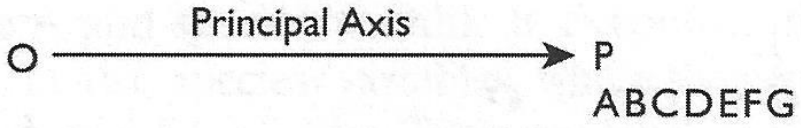
- A distância entre dois pontos quaisquer representando amostras é uma aproximação da sua similaridade em relação às variáveis.
- As amostras 4 e 5 são similares entre si, assim como as amostras 9 e 10, mas as amostras 9 e 10 são ambas diferentes das amostras 4 e 5.

Exemplo

- A mais alta explicação possível de qualquer componente da PCA seria 100% da variação sendo explicada por um único eixo.
- Para isso ocorrer todas as variáveis na análise teriam que ser perfeitamente correlacionadas entre si., isto é, todas as correlações na matriz seriam iguais a 1,0 e todas as variáveis seriam identicamente distribuídas nas amostras.

Exemplo

- A mais alta explicação possível de qualquer componente da PCA seria 100% da variação sendo explicada por um único eixo.
- Para isso ocorrer todas as variáveis na análise teriam que ser perfeitamente correlacionadas entre si. Isto é, todas as correlações na matriz seriam iguais a 1,0 e todas as variáveis seriam identicamente distribuídas nas amostras.



- Todos os autovetores serão iguais a 1,0 e o autovalor será:

$$1^2+1^2+1^2+1^2+1^2+1^2+1^2 = 7$$

- 7 = número de variáveis

De volta à matriz reduzida

- Como em uma matriz de dados onde as variáveis estão intercorrelacionadas a vários níveis, os ângulos entre os vetores produz autovetores menores do que 1,0.
- O autovalor para qualquer componente é uma medida exata da proporção da variação total nos dados explicada pelo componente, e pode ser expresso como uma fração do autovalor sobre o número de variáveis envolvidas na análise.

De volta à matriz reduzida

- Como em uma matriz de dados onde as variáveis estão intercorrelacionadas a vários níveis, os ângulos entre os vetores produz autovetores menores do que 1,0.
- O autovalor para qualquer componente é uma medida exata da proporção da variação total nos dados explicada pelo componente, e pode ser expresso como uma fração do autovalor sobre o número de variáveis envolvidas na análise.

$$\frac{5.33 \text{ (the eigenvalue)}}{7.00 \text{ (the number of species/variables)}} \times 100 = 76.20\%$$

De volta à matriz reduzida

$$\frac{5.33\ \text{(the eigenvalue)}}{7.00\ \text{(the number of species/variables)}} \times 100 = 76.20\%$$

The percentage explanation of the second axis is:

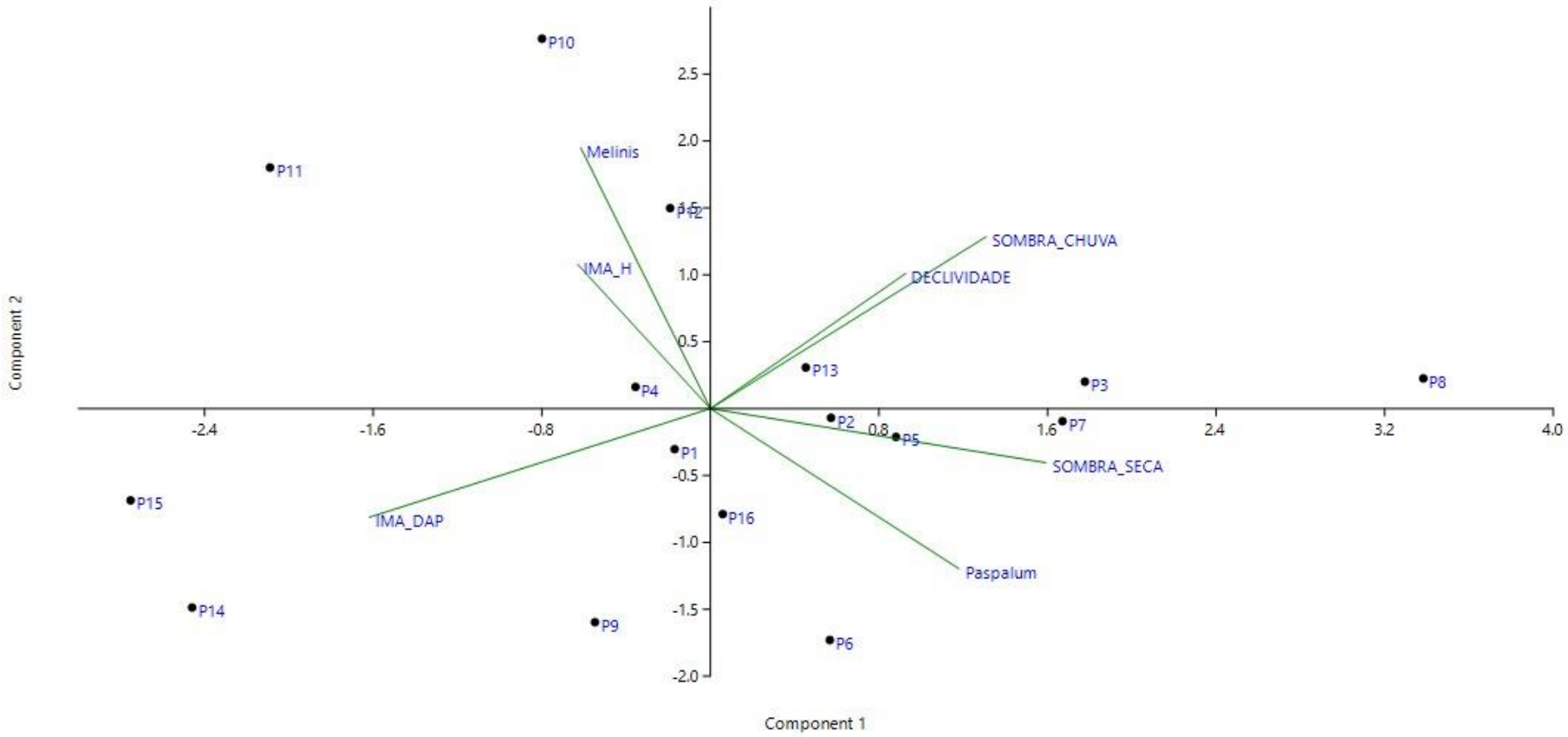
$$\frac{1.66\ \text{(the eigenvalue)}}{7.00\ \text{(the number of species/variables)}} \times 100 = 23.65\%$$

The cumulative explanation of the two components is:

$$76.20 + 23.65 = 99.85\%$$

O método Biplot

- Ambos os valores das espécies e amostras são plotados no mesmo gráfico, mas em escalas diferentes.
- A direção da seta indica a direção na qual a abundância de uma espécie aumenta mais rapidamente.
- O comprimento da seta indica a taxa de mudança na abundância nesta direção.
- Portanto, uma seta comprida indica taxa de mudança gradual na abundância enquanto que uma seta curta representa uma mudança muito rápida.


P10
P11
P12
P13
P4
P3
P8
P2
P7
P5
P1
P15
P16
P14
P9
P6
Melinis
IMA_H
IMA_DAP
SOMBRA_CHUVA
DECLIVIDADE
SOMBRA_SECA
Paspalum
2.5
2.0
1.0
0.5
-0.5
-1.0
-1.5
-2.0
-2.4
-1.6
-0.8
0.8
1.6
2.4
3.2
4.0
Component 2
Component 1