

Análise de Componentes Principais

Jhessica Letícia Kirch
Universidade de São Paulo

Simpósio de Microbiologia Agrícola
11 de abril de 2023



DEFINIÇÃO DE COMPONENTES PRINCIPAIS

- Descrita por Karl Pearson (1901);
- É um dos métodos multivariados mais utilizados;
- Objetivo: Tomar p variáveis X_1, X_2, \dots, X_p e encontrar combinações lineares destas para produzir índices Z_1, Z_2, \dots, Z_p que sejam não correlacionados na ordem de sua importância e que descreva toda a variação dos dados;
- A falta de correlação significa que os índices estão medindo diferentes **dimensões** dos dados.

DEFINIÇÃO DE COMPONENTES PRINCIPAIS

- A ordem é tal que $Var(Z_1) \geq Var(Z_2) \geq \dots \geq Var(Z_p)$;
- Os índices Z_i são também variáveis e são os Componentes Principais (CP);
- Na ACP, espera-se que a maioria das últimas variâncias **sejam baixas**, de modo que grande parte da explicação de variabilidade das variáveis originais se concentre em **poucos componentes** Z_i , resumindo assim o espaço dimensão variável.

DEFINIÇÃO DE COMPONENTES PRINCIPAIS

Desvantagens:

- Ao reduzir o número de variáveis, há perda da informação de variabilidade das variáveis originais.
- A ACP nem sempre funciona! (às vezes, mesmo com a redução ainda continua grande). É o caso de variáveis originais pouco correlacionadas.

OBJETIVO: redução dimensional.

DEFINIÇÃO DE COMPONENTES PRINCIPAIS

- Seja X uma matriz de n observações e p variáveis. O primeiro componente principal é uma combinação linear tal que:

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$

em que a_{ij} é o coeficiente associado à importância da j -ésima variável resposta em Z_i .

- A determinação desses coeficientes é feita por meio da técnica de autovalores e autovetores da matriz de covariância C ou correlação R .

AUTOVALORES E AUTOVETORES

- A decomposição espectral, técnica matemática que consiste na determinação de autovalores e autovetores, ocupa um lugar central na análise multivariada.
- Considere a matriz quadrada A ($p \times p$). Se existe um escalar λ e um vetor v não nulo tal que

$$Av = \lambda v$$

então o λ é denominado autovalor e v é seu autovetor associado.

AUTOVALORES E AUTOVETORES

- O autovalor é obtido por meio de:

$$\det(A - \lambda I)$$

- O autovetor associado é obtido por meio de:

$$(A - \lambda_i I)x = \mathbf{0}$$

AUTOVALORES E AUTOVETORES

▪ Exemplo:

$$A = \begin{bmatrix} 1 & -1 \\ -4 & 1 \end{bmatrix}$$

- Primeiro autovalor e autovetor associado:

$$\lambda = 3 \quad \text{e} \quad \mathbf{v} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

- Verificando que $A\mathbf{v} = \lambda\mathbf{v}$

AUTOVALORES E AUTOVETORES

- Verificando que $A\mathbf{v} = \lambda\mathbf{v}$

$$A = \begin{bmatrix} 1 & -1 \\ -4 & 1 \end{bmatrix}, \quad \lambda = 3 \quad \text{e} \quad \mathbf{v} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

$$A\mathbf{v} = \begin{bmatrix} 1 & -1 \\ -4 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \end{bmatrix} = \begin{bmatrix} 3 \\ -6 \end{bmatrix}$$

$$\lambda\mathbf{v} = 3 \begin{bmatrix} 1 \\ -2 \end{bmatrix} = \begin{bmatrix} 3 \\ -6 \end{bmatrix}$$

AUTOVALORES E AUTOVETORES

- Segundo autovalor e autovetor associado:

$$\lambda = -1 \quad \text{e} \quad \mathbf{v} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

- Verificando que $A\mathbf{v} = \lambda\mathbf{v}$

$$A\mathbf{v} = \begin{bmatrix} 1 & -1 \\ -4 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ -2 \end{bmatrix}$$

$$\lambda\mathbf{v} = -1 \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ -2 \end{bmatrix}$$

AUTOVALORES E AUTOVETORES NA ACP

- As variâncias dos componentes principais são os autovalores da matriz C .
- Isto é $Var(Z_i) = \lambda_i$
- Os coeficientes $a_{i1}, a_{i2}, \dots, a_{ip}$ são os elementos do autovetor associado a λ_i , escalonados de forma que

$$a_{i1}^2 + a_{i2}^2 + \dots + a_{ip}^2 = 1$$

DEFINIÇÃO DE COMPONENTES PRINCIPAIS

Exemplo: Pardais sobreviventes da tempestade

Tabela 1. Correlações entre as cinco medidas do corpo das pardocas.

	X_1	X_2	X_3	X_4	X_5
X_1	1,000				
X_2	0,735	1,000			
X_3	0,662	0,674	1,000		
X_4	0,645	0,769	0,763	1,000	
X_5	0,605	0,529	0,526	0,607	1,000

DEFINIÇÃO DE COMPONENTES PRINCIPAIS

- X_1 = comprimento total;
- X_2 = extensão alar;
- X_3 = comprimento do bico e cabeça;
- X_4 = comprimento do úmero;
- X_5 = comprimento da quilha do esterno.

DEFINIÇÃO DE COMPONENTES PRINCIPAIS

Exemplo da Tabela 1 em $p = 5$ medidas altamente correlacionadas do corpo de $n = 49$ pardocas temos:

$$var(Z_1) = 3,62$$

$$[var(Z_2) = 0,53, \quad var(Z_3) = 0,39, \\ var(Z_4) = 0,30, \quad var(Z_5) = 0,16]$$

O primeiro componente é visivelmente o mais importante para representar a variação total das 49 pardocas!

DEFINIÇÃO DE COMPONENTES PRINCIPAIS

Com as variáveis X_i padronizadas, temos:

$$Z_1 = 0,45X_1 + 0,46X_2 + 0,45X_3 + 0,47X_4 + 0,40X_5$$

expressando um índice de tamanho! (Z_1 consiste em uma média das medidas).

- X_1 = comprimento total;
 - X_2 = extensão alar;
- X_3 = comprimento do bico e cabeça;
 - X_4 = comprimento do úmero;
- X_5 = comprimento da quilha do esterno.

PROCEDIMENTO PARA UMA ANÁLISE DE C.P.

1. Parte-se de um conjunto de n indivíduos e p variáveis.
2. O primeiro CP é:

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$

sob à condição:

$$a_{11}^2 + a_{12}^2 + \cdots + a_{1p}^2 = 1$$

PROCEDIMENTO PARA UMA ANÁLISE DE C.P.

3. O segundo CP é:

$$Z_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$

sob à condição:

$$a_{21}^2 + a_{22}^2 + \cdots + a_{2p}^2 = 1$$

com a condição de que a **correlação** entre Z_1 e Z_2 seja zero!

PROCEDIMENTO PARA UMA ANÁLISE DE C.P.

4. O terceiro CP é:

$$Z_3 = a_{31}X_1 + a_{32}X_2 + \cdots + a_{3p}X_p$$

sob à condição:

$$a_{31}^2 + a_{32}^2 + \cdots + a_{3p}^2 = 1$$

com a condição de que a correlação entre $(Z_1$ e $Z_3)$ e $(Z_2$ e $Z_3)$ sejam **nulas**!

E assim por diante até o máximo de p componentes.

PROCEDIMENTO PARA UMA ANÁLISE DE C.P.

Procedimento de cálculo:

1. Parte-se da matriz de covariâncias C ou R .
2. Calcula-se os p autovalores $(\lambda_1, \lambda_2, \dots, \lambda_p)$ e os p autovetores (a_1, a_2, \dots, a_p) de C ou R .
3. Tem-se então que

$$Z_1 = Xa_1$$

são os valores (escores) do primeiro componente.

PROCEDIMENTO PARA UMA ANÁLISE DE C.P.

4. Descarte os componentes que expliquem pouco da variação total dos dados.

Dizemos que a ACP é bem sucedida quando há uma significativa redução dimensional com o mínimo de perda de informação.

CRITÉRIOS

- Percentual de explicação maior que 80% (sugestão);
- Número de autovalores maiores que 1.

PROCEDIMENTO PARA UMA ANÁLISE DE C.P.

INTERPRETAÇÃO:

- A interpretação dos componentes deve ser feita em termos das magnitudes dos coeficientes associados às variáveis originais
- Portanto, os coeficientes indicam um “**peso**” da variável original.

EXEMPLO

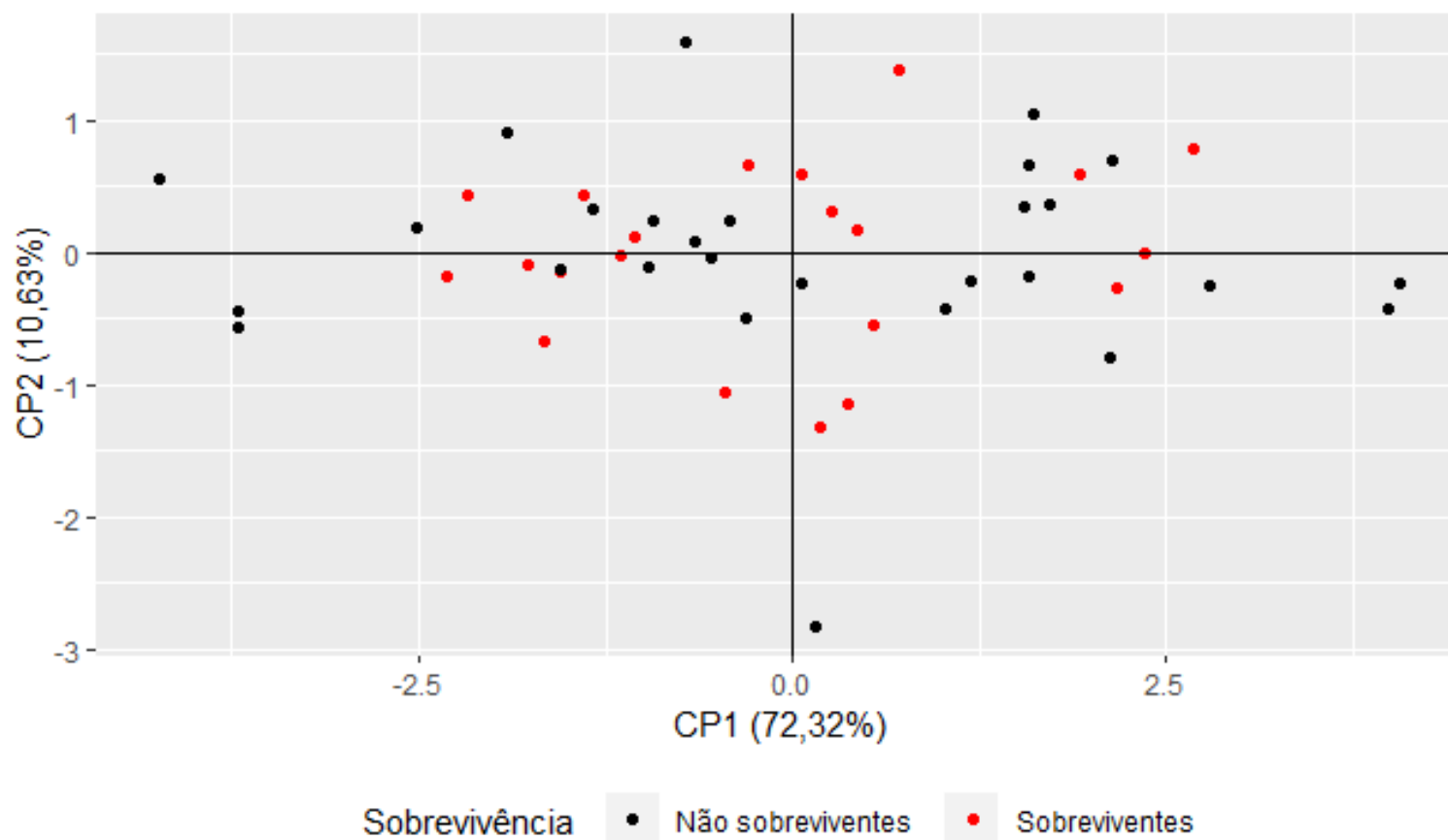
Medidas do corpo de pardocas

Interpretação dos CP:

- CP1, (Z_1): índice dos tamanhos das pardocas e explica 72% da variância total.
- CP2, (Z_2): representa uma diferença de forma entre as pardocas e explica 10,6% da variância total.

EXEMPLO

Figura 1 Representação de 49 pardocas contra valores para os dois primeiros componentes principais, CP1 e CP2.



EXEMPLO

Obs.:

- Nota-se que os pássaros com valores extremos para o 1º CP não sobreviveram. Isso é sugestivo também para o 2º CP.
- Os valores dos autovetores podem sair com sinais trocados em alguns pacotes computacionais. Isso não é um erro! Ele continua medindo exatamente o mesmo aspecto dos dados, mas na direção oposta. Continua sendo uma base do espaço de vetores.

EXEMPLO 2

Emprego nos países europeus

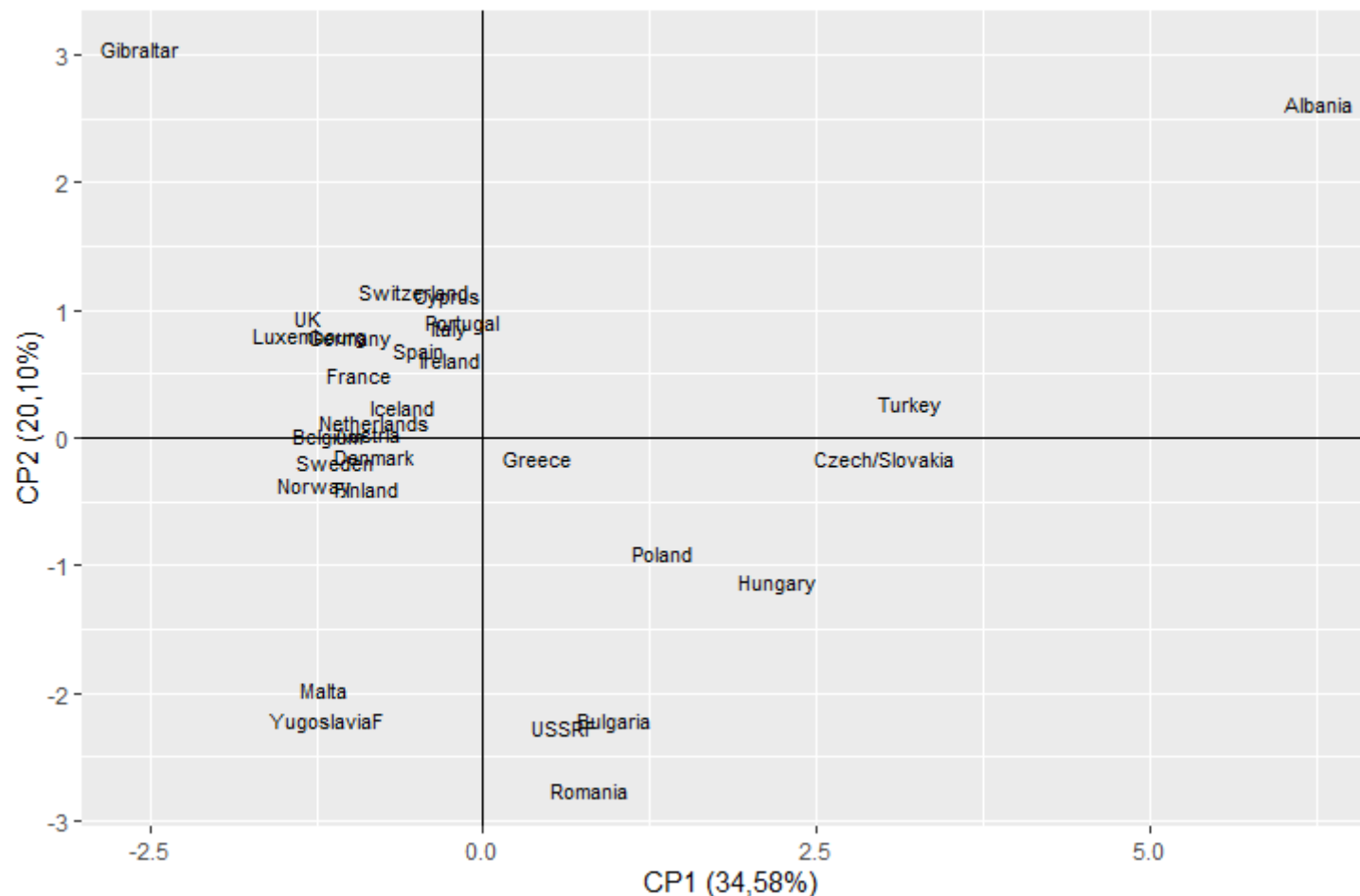
Interpretação dos CP:

Z_1 é um contraste entre os números engajados em AGR (agricultura, florestal e pesca) MIN (mineração e exploração de pedreiras) versus os números engajados em outras ocupações.

Z_2 é o contraste entre os números para MAN (fabricação) e TC (transporte e comunicação) com os números em CON (construção), SER (indústrias e serviços) e FIN (finança)

EXEMPLO 2

Figura 2 Países europeus representados contra os primeiros dois componentes principais para variáveis de emprego.



ACP COM ANÁLISE DE AGRUPAMENTOS

- Alguns algoritmos de análise de agrupamentos começam fazendo uma ACP para reduzir o número de variáveis originais.
- Pode mudar drasticamente os resultados obtidos.
- É uma opção de melhorar a representação gráfica quando os dois primeiros CP's contam por uma alta porcentagem de variação dos dados.