Biologia Quantitativa

Análise de Componentes Principais

2024/01

Depto. De Zoologia

20 de agosto de 2024

Referências básicas

- Pielou caps 3 e 4 (figs do ppt)
- Gotelli & Ellison. 2010. Princípios de Estatística em Ecologia. Editora Artmed
- Ayres et al. Bioestat 2.0. Cap 4.2

Princípios da PCA

- Combina visão geométrica e álgebra de matrizes
- Visão geométrica: rotação n-dimensional da nuvem de pontos na direção dos eixos ortogonais de maior dimensão
- Visão algébrica: multiplicação da matriz de dados por sua transposta para gerar matriz de correlação ou covariância
- Extração dos autovetores e autovalores da matriz
- Autovetores: eixos de maior variância
- Autovalores: percentagem da variância total explicada por cada eixo

Quatro tipos de PCA

- Matriz de correlação ou de covariância
- Projeção centrada ou não centrada
- Combinação dá quatro tipos de PCA

Projeção do sólido no espaço

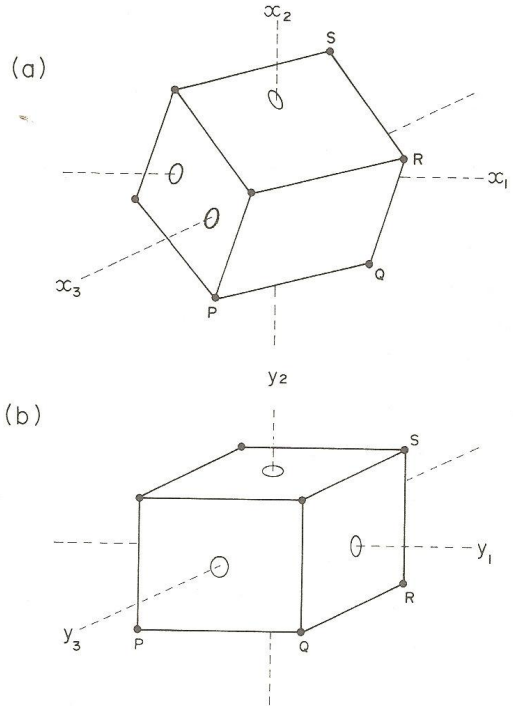


Figure 4.2. A box (cuboid) plotted in a three-dimensional coordinate frame in two different orientations. In (*a*) the box is oblique; in (*b*) it appears after rotation to an orientation that brings its edges parallel with the coordinate axes. The coordinates of the box's corners are denoted by *x*s in the upper graph and by *y*s in the lower graph. The width, height, and depth of the box are PQ, QR, and RS, respectively.

Comparação com Análise de Agrupamento
Também usa matriz de similaridade

TABLE 2.1.

A. DATA MATRIX #1. THE QUANTITIES OF 2 SPECIES IN 10 QUADRATS.

Quadrat	1	2	3	4	5	6	7	8	9	10
Species 1	12	20	28	11	22	8	13	20	39	16
Species 2	30	18	26	5	15	34	24	14	34	11

B. THE DISTANCE MATRIX (THE ROW AND COLUMN LABELS ARE THE QUADRAT NUMBERS)

	Quadrat									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	0	14.4	16.5	25.0	18.0	5.7	6.1	17.9	27.3	19.4
2		0	11.3	15.8	3.6	20.0	9.2	4.0	24.8	8.1
3			0	27.0	12.5	21.5	15.1	14.4	13.6	19.2
4				0	14.9	29.2	19.1	12.7	40.3	7.8
5					0	23.6	12.7	**2.2**	25.5	7.2
6						0	11.2	23.3	31.0	24.4
7							0	12.2	27.9	13.3
8								0	27.6	5.0
9									0	32.5
10										0

A análise de agrupamento cria árvore hierárquica a partir das distâncias

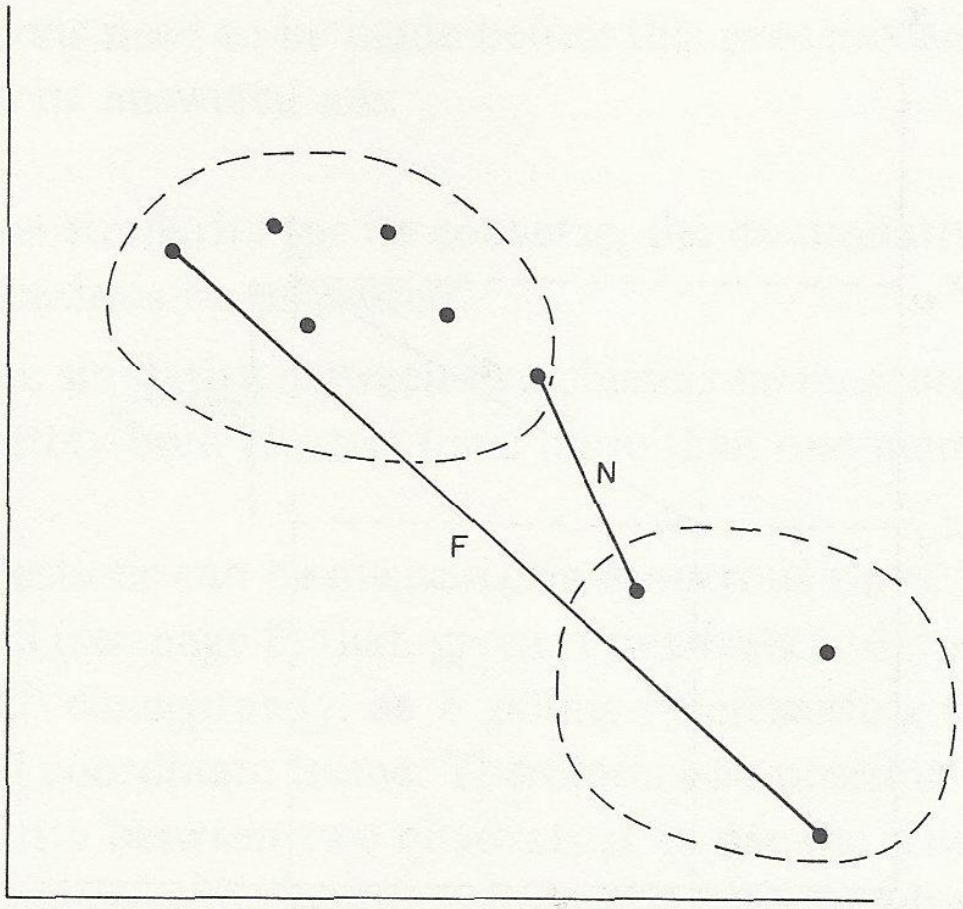
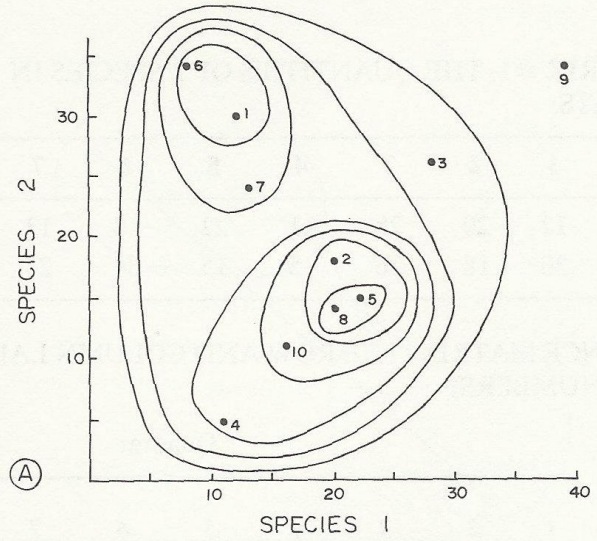


Figure 2.2. Two possible measures of the distance between two clusters. The nearest-neighbor distance N is the shortest distance, and the farthest-neighbor distance F is the longest distance between a member of one cluster and a member of the other.

Análise de Agrupamento



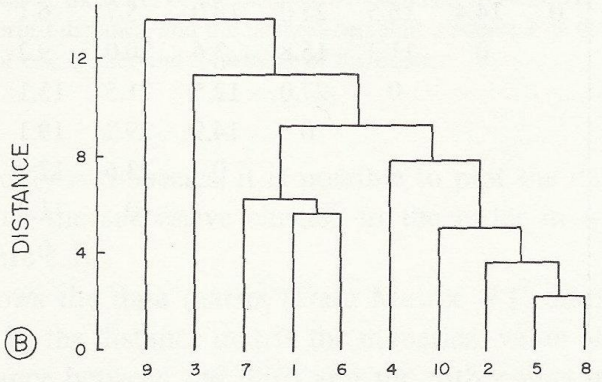


Figure 2.3. (*a*) The data points of Data Matrix #1 (see Table 2.1). The "contours" show the successive fusions with nearest-neighbor clustering except that, for clarity, the final contour enclosing all 10 points is omitted. (*b*) The corresponding dendrogram. Details are given in Table 2.3. The height of each node in the dendrogram is the distance between the pair of clusters whose fusion corresponds with the node.

A PCA faz a reprojeção dos dados criando novos eixos que são combinação de variáveis de acordo com a variância total explicada

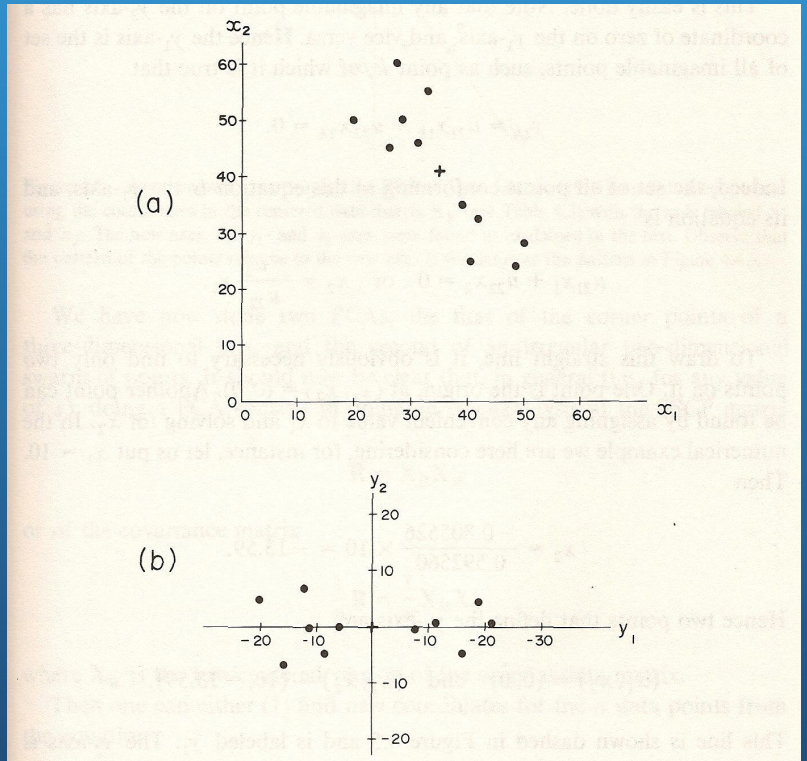


Figure 4.4. Two plots of Data Matrix #9. (*a*) The original untransformed data whose coordinates are given by **X** in Table 4.3. The cross marks the centroid of the swarm. (*b*) After PCA. The origin of the new coordinates is at the swarm's centroid. The swarm has been rotated. The coordinates of each point, measured along axes y_1 and y_2, are given by **Y** in Table 4.3.

Passos da PCA – Algebra Matricial

TABLE 4.3. THE STEPS IN A PRINCIPAL COMPONENTS ANALYSIS OF DATA MATRIX #9.

The 2×11 data matrix is

$$\mathbf{X} = \begin{pmatrix} 20 & 26 & 27 & 28 & 31 & 33 & 39 & 41 & 42 & 48 & 50 \\ 50 & 45 & 60 & 50 & 46 & 55 & 35 & 25 & 33 & 24 & 28 \end{pmatrix}.$$

The row-centered data matrix obtained by subtracting $\bar{x}_1 = 35$ and $\bar{x}_2 = 41$ from the first and second rows of $\mathbf{X}$, respectively, is

$$\mathbf{X}_R = \begin{pmatrix} -15 & -9 & -8 & -7 & -4 & -2 & 4 & 6 & 7 & 13 & 15 \\ 9 & 4 & 19 & 9 & 5 & 14 & -6 & -16 & -8 & -17 & -13 \end{pmatrix}.$$

The SSCP matrix is

$$\mathbf{R} = \begin{pmatrix} 934 & -1026 \\ -1026 & 1574 \end{pmatrix}.$$

The covariance matrix is

$$\frac{1}{n}\mathbf{R} = \begin{pmatrix} 84.9091 & -93.2727 \\ -93.2727 & 143.0909 \end{pmatrix}.$$

The matrix of eigenvectors is

$$\mathbf{U} = \begin{pmatrix} 0.592560 & -0.805526 \\ 0.805526 & 0.592560 \end{pmatrix}.$$

The eigenvalues of the covariance matrix are the nonzero elements of

$$\Lambda = \begin{pmatrix} 211.704 & 0 \\ 0 & 16.295 \end{pmatrix}.$$

The transformed data matrix (after rounding to one decimal place) is

$$\mathbf{Y} = \begin{pmatrix} -16.1 & -8.6 & -20.0 & -11.4 & -6.4 & -12.5 & 7.2 & 16.4 & 10.6 & 21.4 & 19.4 \\ -6.7 & -4.9 & 4.8 & -0.3 & -0.3 & 6.7 & -0.3 & -4.6 & 0.9 & 0.4 & 4.4 \end{pmatrix}$$

Centrando a fig anterior

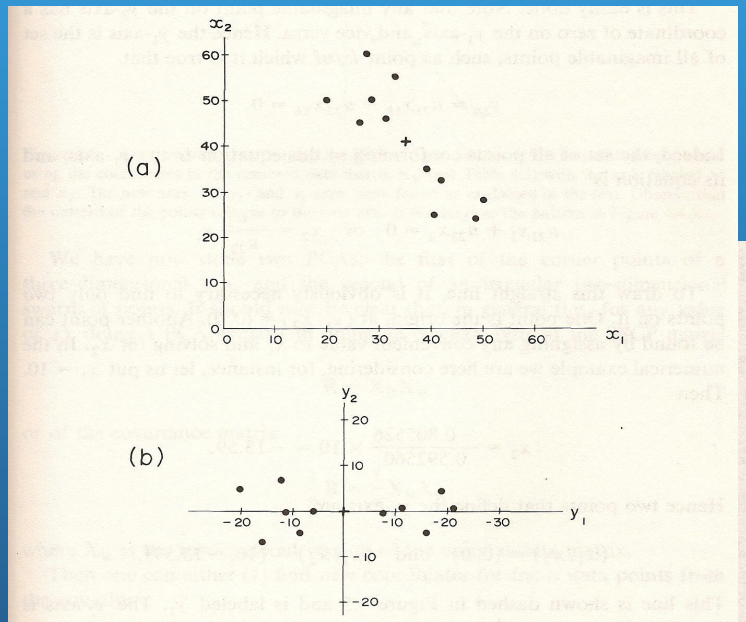


Figure 4.4. Two plots of Data Matrix #9. (*a*) The original untransformed data whose coordinates are given by **X** in Table 4.3. The cross marks the centroid of the swarm. (*b*) After PCA. The origin of the new coordinates is at the swarm's centroid. The swarm has been rotated. The coordinates of each point, measured along axes y_1 and y_2, are given by **Y** in Table 4.3.

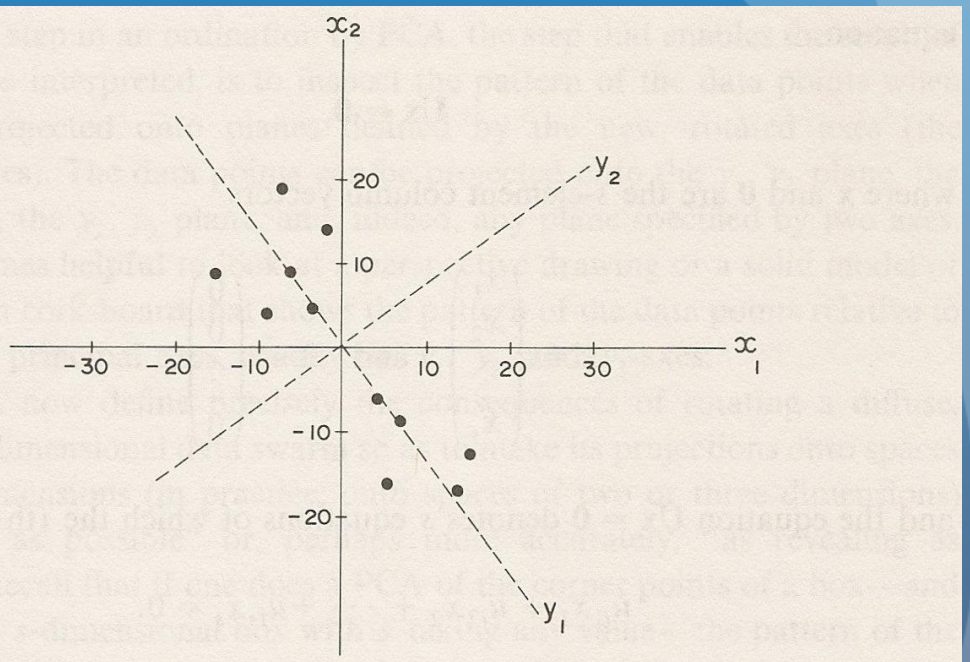


Figure 4.5. Another way of portraying the PCA of Data Matrix #9. The points were plotted using the coordinates in the centered data matrix $\mathbf{X_R}$ (see Table 4.3) with the axes labeled x_1 and x_2. The new axes, the y_1- and y_2-axes, were found as explained in the text. Observe that the pattern of the points relative to the new axes is the same as the pattern in Figure 4.4*b*.

Quatro formas de PCA

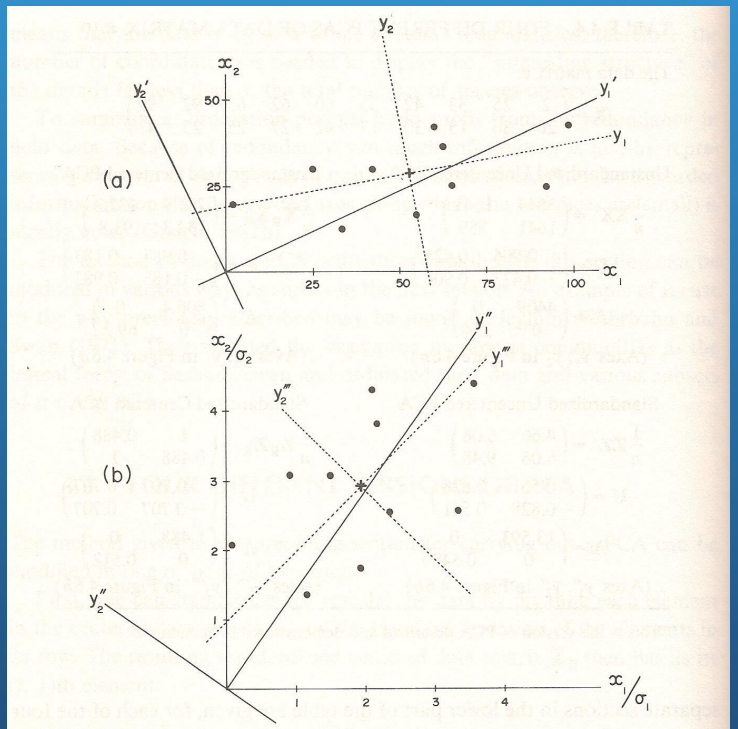


Figure 4.6. Four versions of PCA applied to Data Matrix #10 (see Table 4.4). (*a*) Unstandardized data. The raw, uncentered coordinates are measured along the x_1, x_2 axes. Uncentered PCA rotates the axes into the solid lines labeled y_1', y_2'. Centered PCA shifts the origin to the centroid of the swarm (marked +) and rotates the axes into the dotted lines y_1, y_2. (*b*) Standardized data. The uncentered but standardized data are measured along the $x_1/\sigma_1, x_2/\sigma_2$ axes. Uncentered PCA rotates the axes into the solid lines y_1'', y_2''. Centered PCA shifts the origin to the centroid and rotates the axes into the dotted lines y_1''', y_2'''.

Projeções centrada vs ñ centrada

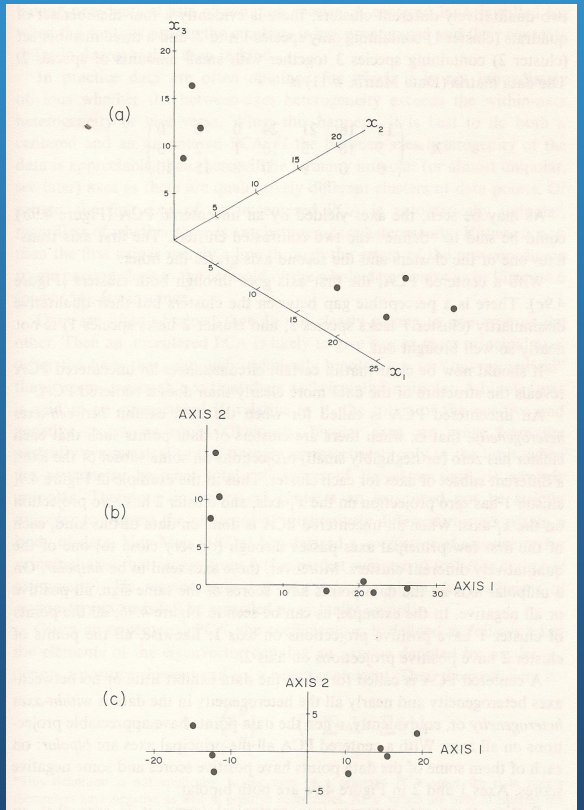


Figure 4.9. (*a*) Seven data points in three-space. They form two qualitatively different clusters. (*b*) A plot of the points in the coordinate frame formed by the first two principal axes resulting from an uncentered PCA. One cluster lies on axis 1 and the other very close to axis 2. (*c*) The corresponding plot after a centered PCA. Both clusters lie on axis 1.

Quatro tipos diferentes de PCA

TABLE 4.4. FOUR DIFFERENT PCAS OF DATA MATRIX #10.

The data matrix is

$$\mathbf{X} = \begin{pmatrix} 2 & 25 & 33 & 42 & 55 & 60 & 62 & 65 & 92 & 99 \\ 20 & 30 & 13 & 30 & 17 & 42 & 27 & 25 & 25 & 43 \end{pmatrix}.$$

Unstandardized Uncentered PCA

$$\frac{1}{n}\mathbf{X}\mathbf{X}' = \begin{pmatrix} 3644 & 1641 \\ 1641 & 889 \end{pmatrix}$$

$$\mathbf{U} = \begin{pmatrix} 0.906 & 0.423 \\ -0.423 & 0.906 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 4409 & 0 \\ 0 & 124 \end{pmatrix}$$

(Axes y_1', y_2' in Figure 4.6*a*)

Unstandardized Centered PCA[a]

$$\frac{1}{n}\mathbf{X}_\mathbf{R}\mathbf{X}_\mathbf{R}' = \begin{pmatrix} 781.9 & 132.3 \\ 132.3 & 93.8 \end{pmatrix}$$

$$\mathbf{U} = \begin{pmatrix} 0.983 & 0.183 \\ -0.183 & 0.983 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 806.4 & 0 \\ 0 & 69.2 \end{pmatrix}$$

(Axes y_1, y_2 in Figure 4.6*a*)

Standardized Uncentered PCA

$$\frac{1}{n}\mathbf{Z}\mathbf{Z}' = \begin{pmatrix} 4.66 & 6.06 \\ 6.06 & 9.48 \end{pmatrix}$$

$$\mathbf{U} = \begin{pmatrix} 0.561 & 0.828 \\ -0.828 & 0.561 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 13.593 & 0 \\ 0 & 0.548 \end{pmatrix}$$

(Axes y_1'', y_2'' in Figure 4.6*b*)

Standardized Centered PCA

$$\frac{1}{n}\mathbf{Z}_\mathbf{R}\mathbf{Z}_\mathbf{R}' = \begin{pmatrix} 1 & 0.488 \\ 0.488 & 1 \end{pmatrix}$$

$$\mathbf{U} = \begin{pmatrix} 0.707 & 0.707 \\ -0.707 & 0.707 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 1.488 & 0 \\ 0 & 0.512 \end{pmatrix}$$

(Axes y_1''', y_2''' in Figure 4.6*b*)

[a] This is the version of PCA described and demonstrated in Section 4.2.

Artigo Costa et al 2012

- Alometria de copaíba no cerrado de Minas Gerais
- PCA de correlação
- Variáveis: diametro tronco, larg copa, grau esbeltez, indice saliencia, grau deflexao copa
- Grafico mostra ordenacao, autovalores e autovetores

Forma da Copaíba por habitat

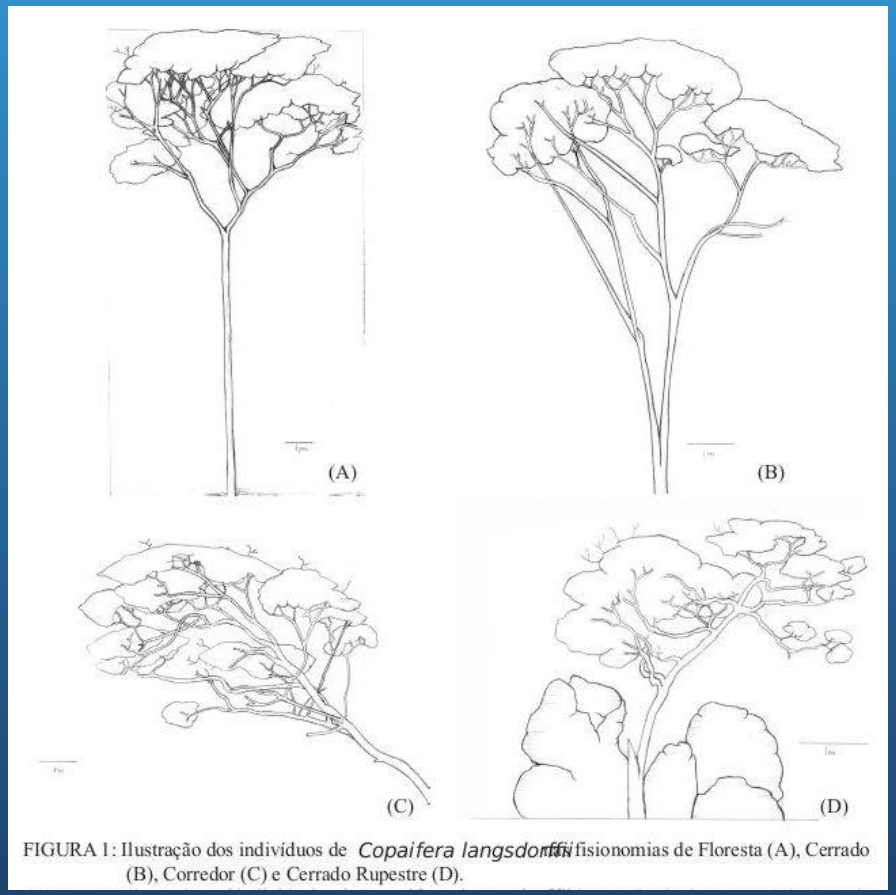


FIGURA 1: Ilustração dos indivíduos de *Copaifera langsdorffii* fisionomias de Floresta (A), Cerrado (B), Corredor (C) e Cerrado Rupestre (D).

Analise PCA – Copaiba cerrado

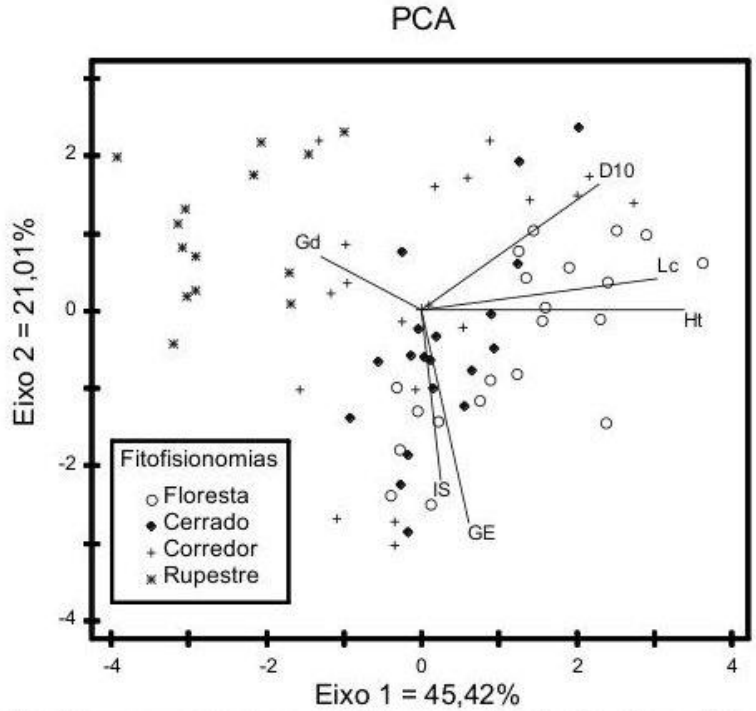


FIGURA 4: Análise de componentes principais (PCA) com variáveis dendrométricas Altura total (Ht); Diâmetro do tronco a uma altura de 10% da altura total (D10); Largura da copa (LC), morfométricas Grau de esbeltez (GE); Índice de Saliência (IS); e arquitetural Grau de deflexão da copa (Gd) das plantas nas fitofisionomias amostradas (Rupestre; Floresta; Cerrado; Corredor).

Dados e Análises

- 19 variáveis climáticas em 70 estações meteorológicas
- PCA de correlação centrada
- 2 primeiros eixos: 57,4% e 29,4% da variância (86,8%)
- Cinco principais variáveis eixo 1 temp: inv max diaria, inv min diaria, out min diaria, inv media diar, out media diaria
- Cinco principais variaveis eixo 2: prim max diaria, verao max diaria, verao media diaria, chuva media prim, chuva media verao

Interpretando o PCA – clima-veget

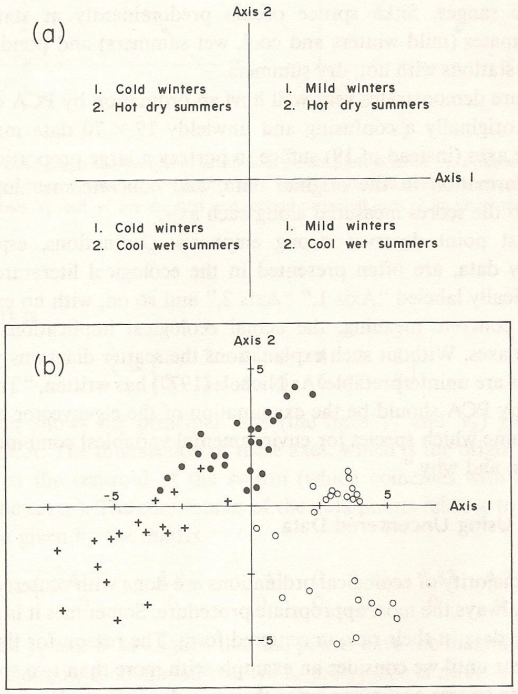


Figure 4.7. (*a*) A "qualitative" graph labeling the regions in the ordination of 70 British Columbia weather stations shown in (*b*). The first two axes obtained from a PCA of climatic data divide the coordinate frame into four regions. Axis 1 separates the stations into those with mild and those with cold winters. Axis 2 separates the stations into those with hot dry summers and those with cool wet summers. (*b*) A two-dimensional ordination, by PCA of the correlation matrix, of the 70 weather stations. The symbol ● denotes a station where ponderosa pine occurs; ○ denotes sitka spruce; +, neither species. The two species are never found together. (Adapted from Newnham, 1968.)

Testes de Hipóteses

- As análises de PCA podem também servir de base para extrair variáveis ou criar novas variáveis que por sua vez sejam usadas em testes univariados ou testes não paramétricos.
- Exemplo: págs 155-157 Pielou (The interpretation of ecological data