Unidade III – Análise de Componentes Principais

3.1 Introdução

A análise de componentes principais (ACP) é uma técnica de estatística multivariada que consiste em transformar, utilizando álgebra linear, um conjunto de variáveis em outro conjunto de variáveis denominadas componentes principais, que são combinações lineares das variáveis originais.
A ideia da técnica de ACP foi originalmente desenvolvida por Karl Pearson em 1901, com um trabalho sobre o ajuste de um sistema de pontos em um espaço multivariado a uma linha ou a um plano. Esse trabalho foi retomado por Hottelling em 1933 que utilizou a técnica para analisar estruturas de correlação e definiu a técnica como é conhecida atualmente.
Os objetivos principais da ACP são: redução da dimensionalidade de dados multivariados, identificação de padrões ocultos em um conjunto de dados, identificação de variáveis correlacionadas e identificação de valores aberrantes. Essa técnica é vista como exploratória ou intermediária e apresenta melhores resultados quando existem variáveis correlacionadas no conjunto de dados.
Os componentes principais gerados são independentes entre si e estimados com o propósito de reter, em ordem de estimação, o máximo de informação, em termos da variação total contida nos dados.
O desenvolvimento da ACP não requer pressuposições de normalidade multivariada, entretanto, se a normalidade existir, podem ser feitas interpretações úteis utilizando a elipsoide de densidade da Normal.

3.2 Dados

Considere a matriz de dados multivariados X, em que observamos p variáveis de n indivíduos extraídos de uma população π, que possui vetor de médias μ e matriz de covariâncias, ∑. X é representada por:

$$\underset{n \times p}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

A estrutura de interdependência entre as variáveis da matriz de dados é representada pela matriz de covariância **S** ou pela matriz de correlação **R**:

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} = \left\{ s_{ik} = \frac{1}{n-1} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k) \right\} \quad \mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}, \quad r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}}$$

Entretanto, não é simples o entendimento dessa estrutura através das variáveis X_1, X_2, X_3,.., X_p. Assim, a análise de componentes principais transforma essa estrutura complicada, representada pelas variáveis X_1, X_2, X_3,...,X_p, em uma outra estrutura representada pelas variáveis Y_1,Y_2,Y_3,...,Y_p os componentes principais, que são não correlacionados e definidas em ordem decrescente da quantidade de variância que são capazes de explicar, para que seja possível comparar os indivíduos usando apenas as variáveis Y_j que apresentam maior variância. A solução é dada a partir da matriz de covariância S ou da matriz de correlação R.
Como as variáveis são geralmente mensuradas em unidades de medidas diferentes entre si, é conveniente padronizar as variáveis. A padronização pode ser feita com média zero e variância 1, ou com variância 1 e média qualquer:
Padronização com média zero e variância 1

$$z_{jk} = \frac{x_{jk} - \bar{x}_k}{s(x_k)}, \qquad j = 1,2,...,n \quad \text{e} \quad k = 1,2,...,p.$$

Padronização com variância 1 e média qualquer

$$z_{jk} = \frac{x_{jk}}{s(x_k)}, \qquad j = 1,2,...,n \quad \text{e} \quad k = 1,2,...,p.$$

Em que $\bar{x}_k$ e $s(x_k)$ são, respectivamente, a estimativa da média e o desvio padrão da variável k.

3.3 Componentes principais (CP)

Os componentes principais são as combinações lineares das variáveis mensuradas X que maximizam a variação total da amostra. Os componentes principais são mutuamente ortogonais e são determinados resolvendo-se a equação característica da matriz S ou R, isto é:

$$|\mathbf{S} - \lambda\mathbf{I}| = 0$$

Se a matriz S for de posto completo (posto = p), isto é, não apresentar nenhuma coluna que seja combinação linear de outra, a equação característica terá 'p' raízes, que são os autovalores ou raízes características da matriz S.

Sejam $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_p$ as raízes da equação característica da matriz R ou S, então $\lambda_1 > \lambda_2 > \lambda_3 > \ldots > \lambda_p$.

Sabemos que para cada autovalor λ_i existe um autovetor normalizado e_i, que são ortogonais entre si.

$$\mathbf{e}_i = \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{ip} \end{bmatrix}$$

Os autovetores e_i apresentam as seguintes propriedades:

$$\sum_{j=1}^{p} e_{ij}^2 = 1 \quad \text{e} \quad \sum_{j=1}^{p} e_{ij}.e_{kj} = 0 \qquad (\mathbf{e_i}.\mathbf{e}_k = 0 \; para \; i \neq k)$$

Sendo e_i o autovetor correspondente ao autovalor λ_i , e $\mathbf{X}' = [X_1, X_2, X_3, .., X_p]$, então os componentes principais são dados por:

$$\begin{aligned} Y_1 &= \mathbf{e_1'}\boldsymbol{X} = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\ Y_2 &= \mathbf{e_2'}\boldsymbol{X} = e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\ &\vdots \\ Y_p &= \mathbf{e_p'}\boldsymbol{X} = e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p \end{aligned}$$

Os componentes principais apresentam as seguintes propriedades:

1) A variância do componente principal Y_i é igual ao valor do autovalor λ_i.
$$\widehat{Var}(Y_k) = \lambda_k, \qquad k = 1,2,\ldots,p$$
2) O primeiro componente é o que apresenta maior variância e assim por diante:
$$\widehat{Var}(Y_1) > \widehat{Var}(Y_2) > \cdots > \widehat{Var}(Y_p)$$
3) O total de variância das variáveis originais é igual ao somatório dos autovalores que é igual ao total de variância dos componentes principais, que é igual ao traço da matriz de covariâncias:
$$\sum \widehat{Var}(X_k) = \sum \lambda_k = \sum \widehat{Var}(Y_k) = \text{traço}(\mathbf{S})$$
4) Os componentes principais não são correlacionados entre si:
$$Cov(Y_i, Y_k) = 0, \quad \text{para } i \neq k$$

3.4 Contribuição de cada componente principal

A contribuição C_k de cada componente principal $\mathbf{Y}_k$ é expressa em porcentagem. É calculada dividindo-se a variância de $\mathbf{Y}_k$ pela variância total. Representa a proporção de variância total explicada pelo componente principal $\mathbf{Y}_k$ e representa a sua importância na análise.

$$C_k = \frac{\widehat{Var}(Y_k)}{\sum \widehat{Var}(Y_k)}.100 = \frac{\lambda_k}{\sum \lambda_k}.100 = \frac{\lambda_k}{traço(\mathbf{S})}.100$$

Pode-se também avaliar a proporção da variação total explicada pelos primeiros k componentes principais ($PV_{k's}$):

$$PV_{k's} = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_k + \cdots + \lambda_p}$$

A proporção da variação total explicada pelos primeiros CP's é uma medida da quantidade de informação retida pela redução de p para k dimensão.

Em certos estudos é desejável que a variância acumulada nos dois primeiros CP's exceda 70-80%. Nesta condição, a distorção das coordenadas no gráfico de dispersão, cujos eixos são os componentes principais, será considerada aceitável e as inferências no estudo satisfatório.

3.5 Interpretação de cada CP

A interpretação do CP é dada pelo grau de influência que cada variável $\mathbf{X}_k$ tem sobre o componente $\mathbf{Y}_j$. O grau de influência é mensurado pela correlação entre cada $\mathbf{X}_k$ e o componente Y_j que está sendo interpretado. Por exemplo a correlação entre X_k e Y_1 é:

$$Cor(X_k, Y_1) = \frac{Cov(X_k, Y_1)}{\sqrt{Var(X_k)}\sqrt{Var(Y_1)}}$$

Mas, lembrando que $Y_i = \mathbf{e}_i'\boldsymbol{X}$ e reescrevendo X_k como $X_k = \boldsymbol{a}_k'\boldsymbol{X}$ com $\boldsymbol{a}_k' = [0, \cdots, 0,1,0, \cdots, 0]$, tem-se:

$$Cov(X_k, Y_i,) = Cov(\boldsymbol{a}_k'\boldsymbol{X}, \mathbf{e}_i'\boldsymbol{X}) = \boldsymbol{a}_k' Cov(\boldsymbol{X})\boldsymbol{e}_i = \boldsymbol{a}_k'\boldsymbol{S}\boldsymbol{e}_i$$

mas, pela equação linear dos autovalores e autovetores, se **A** é uma matriz quadrada, existe a relação: **Ax** = λ**x**, para algum autovetor x e um autovalor λ. Aplicando ao presente caso, a matriz quadrada é **S**, o autovetor é $\mathbf{e}_i$ e os autovalores são os λ's, tem-se então:

$$Cov(X_k, Y_i,) = \boldsymbol{a}_k'\boldsymbol{S}\boldsymbol{e}_j = \boldsymbol{a}_k'\lambda_j\boldsymbol{e}_j = \lambda_j\boldsymbol{a}_k'\boldsymbol{e}_j = \lambda_j e_{jk}$$

Voltando à correlação e substituindo a covariância e as variâncias, tem-se:

$$Cor(X_k, Y_1) = \frac{\lambda_1 e_{1k}}{\sqrt{\widehat{Var}(X_k)}\sqrt{\lambda_1}} = \frac{e_{1k}\sqrt{\lambda_1}}{\sqrt{\widehat{Var}(X_k)}}$$

Para comparar a influência de X_1, X_2, ..., X_p sobre Y_1, deve ser analisado o peso ou loading de cada variável sobre o componente Y_1. O peso de cada variável sobre um determinado componente é dado por:

$$w_1 = \frac{e_{11}}{\sqrt{\widehat{Var}(X_1)}}, \qquad w_2 = \frac{e_{12}}{\sqrt{\widehat{Var}(X_2)}}, \cdots, w_p = \frac{e_{1p}}{\sqrt{\widehat{Var}(X_p)}}$$

Se o objetivo da análise for a obtenção de índices, prática muito comum em Economia, a análise geralmente termina aqui. Mas se o objetivo da análise for comparar ou agrupar indivíduos, a análise continua e é necessário calcular os escores para cada CP que será utilizado na análise.

3.6 Escores dos CP's

Os escores são os valores dos componentes principais. Após a redução da dimensão de p para k, os k componentes principais serão os novos indivíduos e toda análise é feita utilizando-se os escores desses componentes. No Quadro 1 é exemplificado a organização de um conjunto de dados composto por n tratamentos, p variáveis e k componentes principais.

Quadro 1. Conjunto de dados com *n* indivíduos, *p* variáveis e *k* CP's

Indivíduos	Variáveis				Escores dos componentes principais			
	X_1	X_2	⋯	X_p	Y_1	Y_2	⋯	Y_k
1	X_{11}	X_{12}	⋯	X_{1p}	Y_{11}	Y_{12}	⋯	Y_{1k}
2	X_{21}	X_{22}	⋯	X_{2p}	Y_{21}	Y_{22}	⋯	Y_{2k}
⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮
n	X_{n1}	X_{n2}	⋯	X_{np}	Y_{n1}	Y_{n2}	⋯	Y_{nk}

Os escores do primeiro CP para os n tratamentos são:

$$\begin{aligned} Y_{11} &= e_{11}X_{11} + e_{12}X_{12} + \cdots + e_{1p}X_{1p} \\ Y_{21} &= e_{11}X_{21} + e_{12}X_{22} + \cdots + e_{1p}X_{2p} \\ &\vdots \\ Y_{n1} &= e_{11}X_{n1} + e_{12}X_{n2} + \cdots + e_{1p}X_{np} \end{aligned}$$

Exemplo 1. A matriz **X** representa uma amostra de uma vegetação constituída de duas espécies e cinco parcelas (indivíduos):

$$\mathbf{X} = \begin{bmatrix} 2 & 0 \\ 5 & 1 \\ 2 & 4 \\ 1 & 3 \\ 0 & 1 \end{bmatrix}, \quad \bar{\mathbf{X}} = \begin{bmatrix} 2.0 \\ 1.8 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 3.5 & -0.5 \\ -0.5 & 2.7 \end{bmatrix}$$

Para calcular os autovalores:

$$|\mathbf{S} - \lambda \boldsymbol{I}| = 0 \Rightarrow \begin{vmatrix} 3.5 - \lambda & -0.5 \\ -0.5 & 2.7 - \lambda \end{vmatrix} = 9.45 - 3.5\lambda - 2.7\lambda + \lambda^2 - 0.25 = \lambda^2 - 6.2\lambda + 9.2 = 0$$

Cujas raízes são $\lambda_1 = 3.74$ e $\lambda_2 = 2.46$.

Encontrar os autovetores:

para $\lambda = 3.74$, resolve-se a equação:

$$(\mathbf{S} - 3.74\boldsymbol{I})\boldsymbol{y} = \mathbf{0} \Rightarrow \begin{bmatrix} 3.5 - 3.74 & -0.5 \\ -0.5 & 2.7 - 3.74 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Que pode ser escrito como:

$$\begin{aligned} -0.24y_1 - 0.50y_2 &= 0 \\ -0.50y_1 - 1.04y_2 &= 0 \end{aligned}$$

Um vetor solução pode ser escrito com $y_1 = c$ como uma constante arbitrária.

$$\boldsymbol{v}_1 = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} -2.08y_2 \\ y_2 \end{bmatrix} = y_2 \begin{bmatrix} -2.08 \\ 1 \end{bmatrix} = c \begin{bmatrix} -2.08 \\ 1 \end{bmatrix}$$

para $\lambda = 2.46$, resolve-se a equação:

$$(\mathbf{S} - 2.46\boldsymbol{I})\boldsymbol{y} = \mathbf{0} \Rightarrow \begin{bmatrix} 3.5 - 2.46 & -0.5 \\ -0.5 & 2.7 - 2.46 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Que pode ser escrito como:

$$\begin{aligned} 1.04y_1 - 0.50y_2 &= 0 \\ -0.50y_1 + 0.24y_2 &= 0 \end{aligned}$$

Um vetor solução pode ser escrito com $y_1 = c$ como uma constante arbitrária.

$$\boldsymbol{v}_2 = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0.48y_2 \\ y_2 \end{bmatrix} = y_2 \begin{bmatrix} 0.48 \\ 1 \end{bmatrix} = c \begin{bmatrix} 0.48 \\ 1 \end{bmatrix}$$

Para normalizar os autovetores, basta encontrar a norma (comprimento) de cada autovetor e dividir os elementos do vetor por esse valor:

Para $\boldsymbol{v}_1$, a norma será: $\|\boldsymbol{v}_1\| = \sqrt{(-2.08)^2 + 1^2} = 2.3$, logo o autovetor normalizado será: $e_1 =: \frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|} = \begin{bmatrix} -0.90 \\ 0.43 \end{bmatrix}$

Para $\boldsymbol{v}_2$, a norma será: $\|\boldsymbol{v}_2\| = \sqrt{(0.48)^2 + 1^2} = 1.1$, logo o autovetor normalizado será: $e_2 =: \frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_2\|} = \begin{bmatrix} 0.43 \\ 0.90 \end{bmatrix}$.

Pode ser verificado que a soma dos autovalores é igual a soma das variâncias das espécies:

$S_{11} + S_{22} = \lambda_1 + \lambda_2 = 3{,}5 + 2{,}7 = 3{,}74 + 2{,}46 = 6{,}2$

A contribuição de cada PC é dada por:

$$C_k = \frac{\widehat{Var}(Y_k)}{\sum \widehat{Var}(Y_k)} . 100 = \frac{\lambda_k}{\sum \lambda_k} . 100 = \frac{\lambda_k}{traço(\mathbf{S})} . 100, \text{ logo:}$$

$$C_1 = \frac{3.74}{6.20} . 100 = 60.33\% \quad \text{e} \quad C_2 = \frac{2.46}{6.20} . 100 = 39.67\%$$

Verifica-se, neste caso, que 60,37% da variação total está concentrada em Y_1, ou seja, Y_1 explica 60,33% da variação total. O segundo componente principal (Y_2) explica 39,67% da variação total.

Vamos verificar o grau de influência de cada CP:

Os coeficientes de correlação entre Y_1 e as variáveis X_1 e X_2 são:

$$Cor(X_1, Y_1) = \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{\widehat{Var}(X_1)}} = \frac{-0.90\sqrt{3.74}}{\sqrt{3.50}} = -0.93 \quad \text{e} \quad Cor(X_2, Y_1) = \frac{e_{21}\sqrt{\lambda_1}}{\sqrt{\widehat{Var}(X_2)}} = \frac{0.43\sqrt{3.74}}{\sqrt{2.70}} = 0.51$$

Verifica-se que existe uma grande correlação entre Y_1 e X_1, mostrando que X_1 é de grande importância para o primeiro CP.
O coeficiente de correlação entre Y_2 e as variáveis X_1 e X_2 são:

$$Cor(X_1, Y_2) = \frac{e_{12}\sqrt{\lambda_2}}{\sqrt{\widehat{Var}(X_1)}} = \frac{0.43\sqrt{2.46}}{\sqrt{3.50}} = 0.36 \quad \text{e} \quad Cor(X_2, Y_2) = \frac{e_{22}\sqrt{\lambda_2}}{\sqrt{\widehat{Var}(X_2)}} = \frac{0.90\sqrt{2.46}}{\sqrt{2.70}} = 0.86$$

Verifica-se que a variável X_2 é a de maior importância para o segundo CP (Y_2).

Em resumo, tem-se:

Componente Principal	Variância		CPA (Autovetores)	
	Autovalor	(%)	X_1	X_2
Y_1	3,74	60,33	-0,90	0,43
Y_2	2,46	39,67	0,43	0,90

CPA = Coeficiente de ponderação associado.

Escores dos CP's:

$$\mathbf{X} = \begin{bmatrix} 2 & 0 \\ 5 & 1 \\ 2 & 4 \\ 1 & 3 \\ 0 & 1 \end{bmatrix}$$

Y_{11} = -0,90x2 + 0,43x0 = -1,80
Y_{12} = -0,90x5 + 0,43x1 = -4,07
Y_{13} = -0,90x2 + 0,43x4 = -0,08
⋮
Y_{25} = 0,43x0 + 0,90x1 = 0,901

Que pode ser organizado da seguinte maneira:

Parcelas	Componentes	
	Y_1	Y_2
1	-1,80	0,86
2	-4,07	3,05
3	-0,08	4,46
4	0,39	3,13
5	0,43	0,90
Variância	3,740	2,460

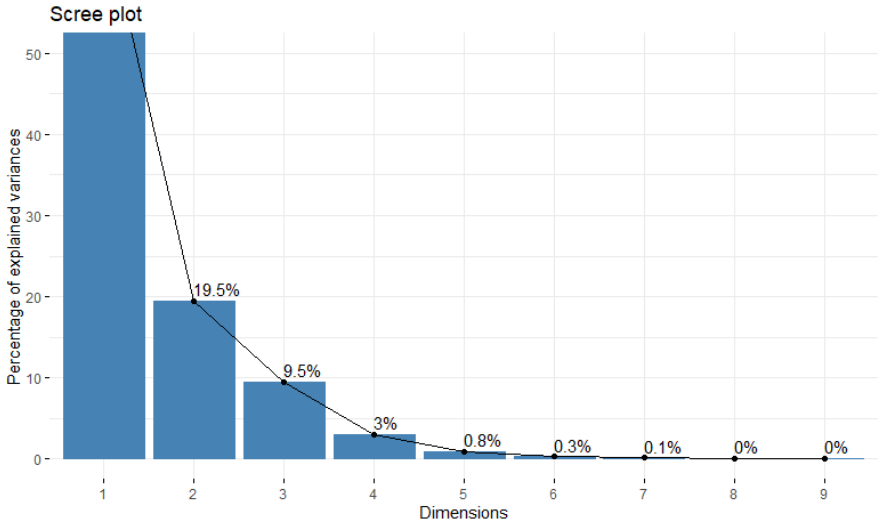
Os escores dos componentes são coordenadas retangulares da ordenação e podem ser plotados produzindo diagramas que mostram a distribuição agrupada dos CP's.

3.7. Gráficos

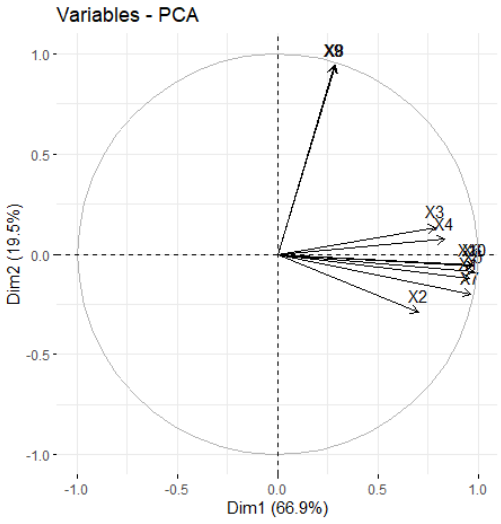
Usualmente o uso de softwares nos permitem a exibição de gráficos de apoio que permitem uma melhor compreensão na ACP.

Para esse exemplo apresentaremos alguns gráficos que são usuais:

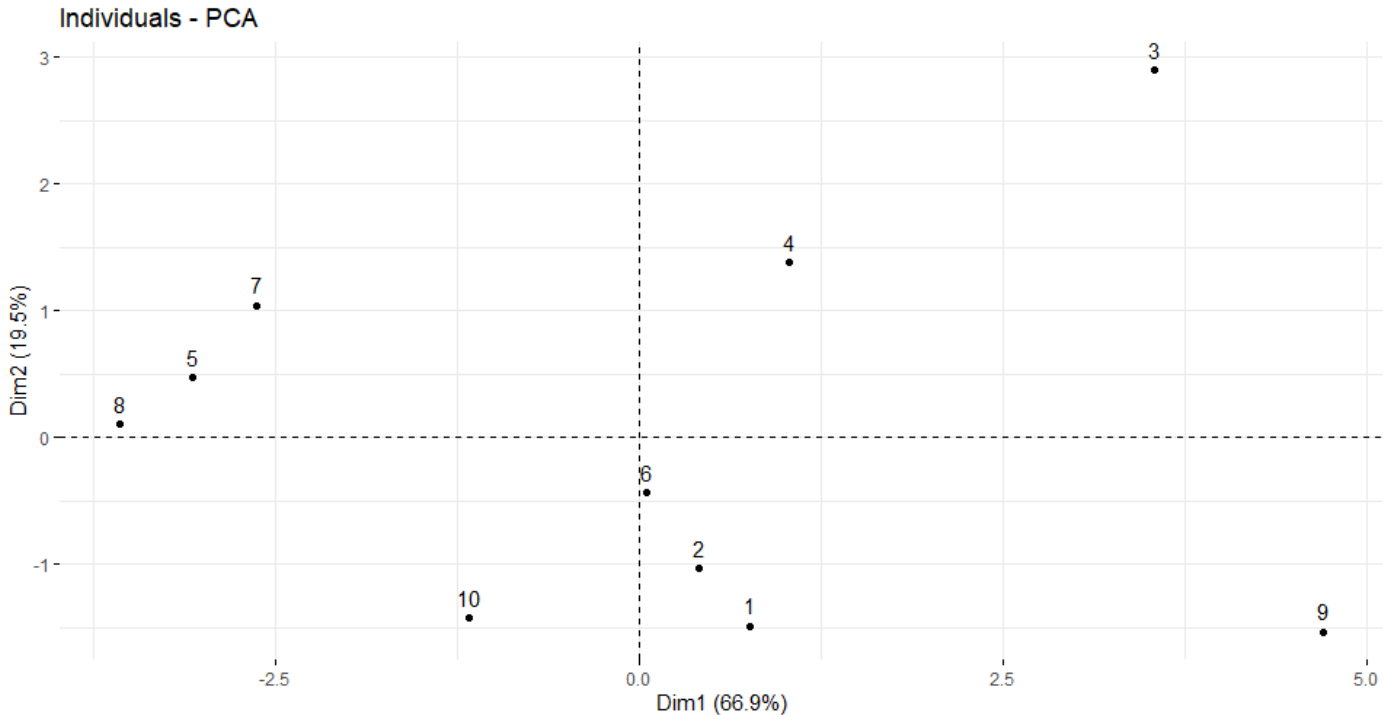
1) **Gráfico de autovalores** (Scree plot): fornece um método para determinar o número de componentes principais. O scree plot exibe o número do componente versus o autovalor correspondente. Os autovalores da matriz de correlação são iguais às variâncias dos componentes principais; portanto, escolha o número de componentes com base no tamanho dos autovalores.
   O padrão ideal é uma curva acentuada, seguida por uma curva e depois por uma linha reta. Escolha o número de componentes que estão na curva antes do primeiro ponto que inicia a tendência da linha. Na prática, você pode ter dificuldade em interpretar um plano de scree. Use seu conhecimento dos dados e os resultados dos outros métodos de seleção de componentes para ajudar a decidir o número de componentes a serem retidos. Nesse exemplo, o gráfico indica que podem ser escolhidos dois ou três componentes.



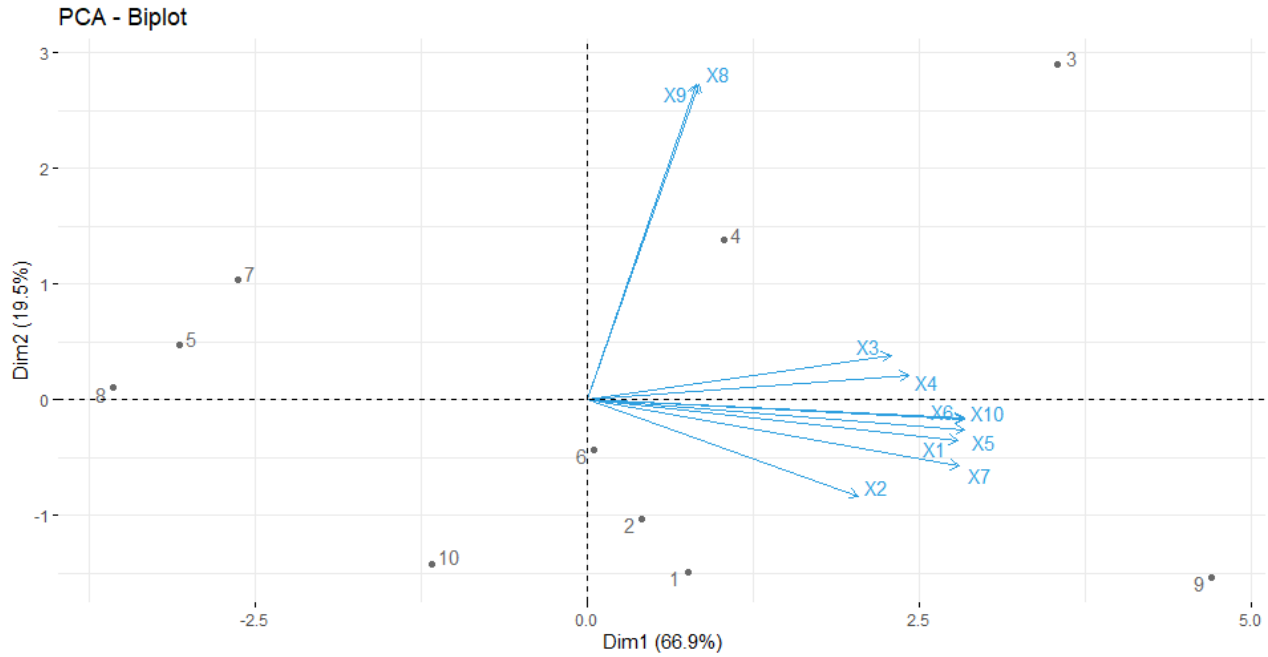
2) **Gráfico da influência das variáveis** (gráfico das cargas fatoriais)
   Os gráficos de cargas fatoriais representam os coeficientes de cada variável para o primeiro componente versus os coeficientes para o segundo componente. Usado para identificar quais variáveis têm o maior efeito em cada componente. As cargas fatoriais podem variar de -1 a 1. As cargas fatoriais próximas de -1 ou 1 indicam que a variável influencia fortemente o componente. As cargas fatoriais próximas de 0 indicam que a variável tem uma influência fraca no componente. Avaliar as cargas fatoriais também pode ajudá-lo a caracterizar cada componente em termos das variáveis. No exemplo a seguir verifica-se um grupo de variáveis bastante correlacionadas e duas variáveis separadas, que não são muito correlacionadas com as demais, mas que são fortemente correlacionadas entre si ficam sobrepostas no gráfico). Essas variáveis (X8 e X9) estão representando o CP 2, enquanto o restante das variáveis representam o CP 1.



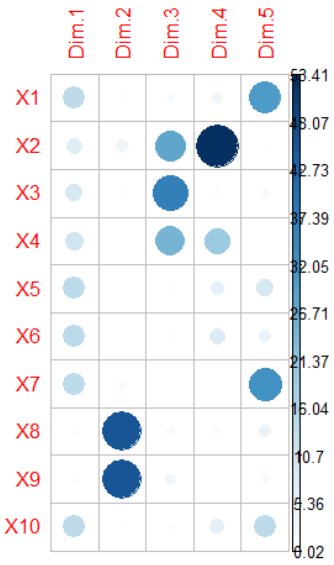
3) **Gráfico individual**: O gráfico de indivíduos representa graficamente cada observação (indivíduo) do segundo componente principal versus os escores do primeiro componente principal. Se os dois primeiros componentes forem responsáveis pela maior parte da variação nos dados, você poderá usar o gráfico para avaliar a estrutura de dados e detectar clusters, valores discrepantes e tendências. O gráfico pode revelar agrupamentos de pontos, que podem indicar duas ou mais distribuições separadas nos dados.



4) **Biplot**: O biplot sobrepõe o gráfico de indivíduos e o gráfico de cargas fatoriais. Usado para avaliar a estrutura dos dados e as cargas fatoriais dos primeiros dois componentes em um gráfico. São representados os escores do CP1 versus os escores do CP2, bem como as cargas fatoriais de ambos os componentes.



5) **Gráfico de contribuições das variáveis nos PC's**
As contribuições das variáveis na contabilização da variabilidade em um determinado CP são apresentadas nesse gráfico. As variáveis que são mais correlacionadas com o PC1 e PC2 são as mais importantes para explicar a variabilidade dos dados. As variáveis que não são correlacionadas com nenhum PC ou correlacionadas com as últimas dimensões (últimos PC's) são variáveis com baixa contribuições e candidatas a serem removidas para simplificar a análise.



Exemplo 2. Dados de um teste de progênies de ***Eucalyptus* sp.**, em que foram avaliadas 10 características (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9 e X_{10}) em 10 progênies, num delineamento em blocos ao acaso com quatro repetições e seis plantas por parcela, realizou-se a análise por componentes principais. A seguir são apresentados as matrizes de médias, variância, covariância e de correlações.

Quadro 1: Médias das 10 Progênies em Relação a 10 características (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9 e X_{10})

Prog.	Características									
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	10.7542	0.6708	16.4708	12.8417	0.0750	0.0575	0.0175	0.4786	0.3659	0.1559
2	10.3417	0.6000	17.0833	13.0708	0.0731	0.0556	0.0175	0.4791	0.3647	0.1513
3	11.2625	0.6750	17.0250	13.2875	0.0832	0.0649	0.0184	0.5509	0.4274	0.1842
4	10.3583	0.6083	16.7542	13.1375	0.0768	0.0587	0.0181	0.5230	0.3975	0.1475
5	9.8083	0.5542	15.9250	11.6000	0.0616	0.0480	0.0136	0.4943	0.3846	0.1244
6	10.2292	0.6833	16.6208	13.0708	0.0691	0.0525	0.0167	0.4953	0.3750	0.1402
7	9.6042	0.6500	15.7333	11.5958	0.0621	0.0479	0.0142	0.5147	0.3939	0.1201
8	9.5208	0.5833	15.8167	11.6208	0.0579	0.0439	0.0140	0.4950	0.3758	0.1169
9	11.6333	0.7458	16.6833	12.9125	0.0954	0.0736	0.0218	0.4924	0.3769	0.1979
10	10.4292	0.6792	15.7208	11.7958	0.0687	0.0527	0.0161	0.4803	0.3674	0.1422

Quadro 2 – Médias Padronizadas das 10 Progênies em Relação a 10 Características (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9 e X_{10})

Prog.	Características									
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	15.78	11.66	30.73	17.47	06.69	06.54	07.04	20.82	19.03	05.90
2	15.17	10.43	31.87	17.78	06.52	06.32	07.04	20.84	18.97	05.73
3	16.52	11.73	31.76	18.08	07.42	07.38	07.41	23.97	22.23	06.98
4	15.19	10.57	31.26	17.88	06.85	06.68	07.28	22.75	20.68	05.59
5	14.39	9.63	29.71	15.78	5.50	5.36	5.55	21.47	19.23	4.71
6	15.01	11.87	31.01	17.78	06.17	05.97	06.72	21.55	19.51	05.31
7	14.09	11.29	29.35	15.78	05.54	05.45	05.71	22.39	20.49	04.55
8	13.97	10.14	29.51	15.81	05.17	04.99	05.63	21.54	19.55	04.43
9	17.07	12.96	31.12	17.57	08.51	08.37	08.77	21.42	19.60	07.49
10	15.30	11.80	29.33	16.05	06.13	05.99	06.48	20.89	19.11	05.38

* Padronização : $z_{jk} = \frac{x_{jk}}{s(x_k)}$

Quadro 3 – Matriz de Variâncias e Covariâncias Entre as Variáveis originais

0.4646	0.0291	0.2361	0.3507	0.0074	0.0058	0.0016	0.0026	0.0025	0.0178
	0.0033	0.0730	0.0171	0.0005	0.0004	0.0001	0.00004	0.000009	0.0011
		0.2872	0.3772	0.0142	0.0032	0.0009	0.0034	0.0025	0.0099
			0.5401	0.0061	0.0046	0.0014	0.0044	0.0030	0.0141
				0.0001	0.0001	0.00003	0.00005	0.00005	0.0003
					0.00008	0.00002	0.00005	0.00004	0.00006
						0.000006	0.000006	0.000004	0.000006
							0.00053	0.0004	0.0001
								0.0004	0.0001
									0.0007

Quadro 4 – Matriz de Correlação entre Variáveis Originais

1,0	0,7419	0,6462	0,7000	0,9626	0,9663	0,9263	0,1668	0,1932	0,9885
	1,0	0,2391	0,4050	0,7043	0,6992	0,7097	0,0295	0,0081	0,7032
		1,0	0,9577	0,6977	0,6835	0,7294	0,2726	0,2418	0,7035
			1,0	0,7365	0,7176	0,7860	0,2619	0,2112	0,7263
				1,0	0,9983	0,9785	0,2060	0,2087	0,9768
					1,0	0,9647	0,2352	0,2457	0,9803
						1,0	0,1012	0,0768	0,9395
							1,0	0,9857	0,2279
								1,0	0,2485
									1,0

Baseado na teoria descrita anteriormente sobre componentes principais, os autovalores e autovetores associados são apresentados a seguir (Quadro 5). Estes foram obtidos a partir da matriz de correlação entre as características originais (R) (ou matriz de covariâncias entre as características padronizadas).

A obtenção destes autovalores e autovetores associados por um processo manual é impraticável. Desta forma, utilizou-se a plataforma R, mas pode ser usado qualquer outro software.

Quadro 5 – Componentes Principais Obtidos da Análise de 10 Características (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9 e X_{10}) (variáveis padronizadas)

Componente Principal	Variância		Coeficiente de Ponderação Associado (Autovetores)									
	Autovalor	Acumul. (%)	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Y_1	6.6879	66.88	0.3713	0.2715	0.3041	0.3218	0.3788	0.3778	0.3728	0.1117	0.1090	0.3778
Y_2	1.9454	86.33	-0.0862	-0.2067	0.0954	0.0537	-0.0639	-0.0406	-0.1411	0.6752	0.6744	-0.0398
Y_3	0.9508	95.79	-0.1511	-0.5276	0.6051	0.4980	-0.0863	-0.1151	0.0137	-0.1405	-0.1838	-0.1012
Y_4	0.2950	98.84	-0.1973	0.7308	0.1132	0.4413	-0.2375	-0.2799	-0.0621	0.1362	-0.0554	-0.2432
Y_5	0.0849	99.64	-0.5509	-0.0580	-0.1532	-0.0238	0.3027	0.2162	0.5664	0.2195	-0.1465	-0.3746
Y_6	0.0255	99.87	-0.4299	0.2392	0.6111	-0.5532	0.0263	0.0539	-0.0943	-0.0505	0.0272	0.2501
Y_7	0.0099	99.99	0.1078	-0.0855	-0.0135	-0.1496	-0.2984	-0.5076	0.4868	0.3609	-0.2933	0.3981
Y_8	0.0003	99.99	0.4077	0.0557	0.2788	-0.2880	-01295	-0.1633	0.4069	-0.2360	0.3195	-0.5507
Y_9	0.00009	99.99	-0.3560	-0.0092	-0.2113	0.1968	-0.1537	-0.1377	0.3015	-0.5032	0.5305	0.3488
Y_{10}	0.00002	100.00	-0.0029	-0.0047	0.0002	0.0008	-0.7529	0.6411	0.1175	0.0589	-00686	0.0057

No Quadro 3, pode-se constatar numericamente que:

$$\sum \lambda_k = \sum \widehat{Var}(Y_k) = \text{traço}(\mathbf{S}) = 10$$

Os resultados apresentados no Quadro 5 evidenciam que o primeiro componente principal (Y_1) explica 66,88% da variação total disponível. Os dois primeiros componentes principais (Y_1 e Y_2) explicam 86,33% e os três primeiros (Y_1, Y_2 e Y_3) explicam 95,84% da variância total disponível. Portanto, para o presente exemplo, a técnica de componentes principais sumariza muito bem a variação total disponível dos dados amostrais pelo três primeiros componentes principais.

Assim, a utilização destes componentes no estudo de divergência genética por meio da dispersão dos escores em gráficos cujos eixos são os referidos componentes (Y_1 e Y_2), apresentará resultados satisfatórios.

Em estudos que utilizam a técnica dos componentes principais como meio de descartes de variáveis com a finalidade de redução de mão-de-obra, tempo e custo despendido na análise e interpretação dos dados experimentais, a importância relativa das características pode ser avaliada pela magnitude do coeficiente de ponderação destas. Assim, com base em MARDIA et al. (1978) e CRUZ e REGAZZI (1994), para o presente exemplo, identifica-se, em ordem crescente, os caracteres X_5, X_{10}, X_6, X_3, X_7 e X_2, com maiores pesos em Y_{10} (-0,7529), Y_9 (0,5305), Y_8 (-0,5507), Y_7 (-0,5076), Y_6 (0,6111), Y_5 (0,5664) e Y_4 (0,7308), respectivamente, como os de menores importância no estudo realizado, são possíveis de descarte.

No exemplo em consideração, o descarte de X_2, X_3, X_5, X_6, X_7 e X_{10} é minimizado pela presença de X_1 e X_4, cujas correlações entre estas são altas (ver matriz de correlações entre variáveis originais). O descarte da variável X_9 é minimizado pela presença de X_8, cuja correlação com X_9 é de 0,9857.

Os escores relativos a cada progênie, em cada componente, é estimado com base nas informações do Quadro 2 (médias padronizadas das 10 progênies em relação as 10 características X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9 e X_{10}) e do Quadro 5 (componentes principais obtidos da análise de 10 características X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9 e X_{10}). Assim, tem-se:

$$Y_{11} = 0,3713\ (15,78) + 0,2715\ (11,66) + 0,3041\ (30,73)$$
$$+ 0,3218\ (17,47) + 0,3788\ (6,69) + 0,3778\ (6,54) +$$
$$+ 0,3728\ (7,04) + 0,1117\ (20,82) + 0,1090\ (19,03) +$$
$$+ 0,3778\ (5,90)$$
$$Y_{11} = 38,2770$$

Os demais escores encontram-se no Quadro 6.

Com base na Figura 2, observa-se que, em relação aos caracteres considerados, as progênies 1, 2, 6 e 10 e as progênies 5, 7 e 8 são as mais similares, havendo, entretanto, considerável divergência entre as progênies 3, 4 e 9.

As distâncias gráficas podem se estimadas pelas distâncias Euclidianas:

$$\mathbf{dcp_{ii}} = [(Y_{i1} - Y'_{i'1})^2 + (Y_{i2} - Y'_{i'2})^2]^{½}$$

Quadro 6 – Escores relativos a 10 progênies, obtidos em relação aos dois primeiros CP's

Genótipos	Y_1'	Y_2
1	38.2570	25.0736
2	37.9302	25.5050
3	40.8988	29.2319
4	38.5141	27.7958
5	34.6232	26.9303
6	37.5891	26.9303
7	35.0461	27.4731
8	34.1564	26.5848
9	41.9986	25.0295
10	36.4273	25.1353

Por esta expressão são obtidas as medidas de dissimilaridade, que são apresentadas no Quadro 7. Como ilustração é obtida a estimativa de $dcp_{1,2}$:

$$dcp_{1,2} = [\ (38,2570 - 37,9302)^2 + (25,0736 - 25,5050)^2]^{½}$$
$$dcp_{1,2} = 0,5412$$

Quadro 7 – Dissimilaridade entre Genótipos, com Base na Distância Euclidiana, Obtida de Escores dos Dois Primeiros Componentes Principais

-	0.5412	4.9265	2.7343	4.0807	1.2030	4.0083	4.3702	3.7419	1.8307
	-	4.7647	2.3640	3.6011	0.6635	3.4915	3.9252	4.0961	1.5477
		-	2.7837	6.6843	4.5745	6.1111	7.2434	4.3439	6.0644
			-	3.9860	1.9544	3.4828	4.5228	4.4491	3.3813
				-	3.0870	0.6882	0.5808	7.6164	2.5450
					-	2.9022	3.4705	4.5315	1.4937
						-	1.2574	7.3692	2.7152
							-	7.9949	2.6941
								-	5.5723
									-

Quantos Componentes usar na análise?

Um autovalor > 1 indica que os PCs são responsáveis por mais variância do que por uma das variáveis originais em dados padronizados. Isso é comumente usado como um ponto de corte para o qual os PCs são retidos. Isso é verdadeiro apenas quando os dados são padronizados. Você também pode limitar o número de componentes àquele número que representa uma certa fração da variância total. Por exemplo, se você estiver satisfeito com 70% da variância total explicada em seguida, use o número de componentes para conseguir isso.