ANÁLISE DE COMPONENTES PRINCIPAIS (A.C.P.)

$(Principal\ Component\ Analysis - PCA)$

Material preparado pelo Prof. Dr. César Gonçalves de Lima – FZEA/USP

ANÁLISE DE COMPONENTES PRINCIPAIS (A.C.P.)

Objetivo: A partir de um número grande de p variáveis $X_1, X_2, \cdots, X_p$ medidas em n indivíduos, encontrar algumas combinações lineares dessas variáveis para produzir *novas* variáveis

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p$$

que não sejam correlacionadas entre si e que descrevam a maior parte da variabilidade dos dados originais.

- As novas variáveis (ou índices) são chamados **componentes principais** (CP).
- A falta de correlação entre os CP's significa que eles medem diferentes **dimensões** dos dados e quando percebemos o significado dessas dimensões, podemos dar nomes aos CP's.

Os CP's são obtidos de tal forma que:

$$var(Z_1) \geq var(Z_2) \geq \cdots \geq var(Z_p)$$

Em que $var(Z_i)$ denota a variância amostral do componente Z_i.

- **Espera-se** que poucos componentes possam explicar a maior parte da variabilidade dos dados iniciais resultantes da avaliação de um número grande de variáveis e que os últimos componentes tenham variâncias desprezíveis (quase nulas).
- A **ACP nem sempre funciona**, ou seja, nem sempre conseguimos que um grande número de variáveis iniciais seja reduzido a um pequeno número de componentes principais.
- Os **melhores resultados** serão obtidos quando as variáveis originais forem altamente correlacionadas, positiva ou negativamente.

PROCEDIMENTO DE ANÁLISE

O primeiro componente principal ($CP1$) é uma combinação linear das variáveis:

$$CP1 =, a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$

tal que: $var(Z_1)$ é a maior possível e $a_{11}^2 + a_{12}^2 + \cdots + a_{1p}^2 = 1$

O segundo componente principal (CP2) é outra combinação

$$CP2 =, a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$$

Que é obtido de tal forma que: $var(CP_2) \leq var(CP_1)$, $\sum_{j=1}^{p} a_{2j}^2 = 1$ e $corr(CP_1, CP_2) = 0$.

Os outros componentes $CP_3, \ldots, CP_p$, são obtidos de forma análoga.

- Se existem p variáveis originais existirão, no máximo, p componentes principais.

- O cálculo dos coeficientes a_{ij} envolve a obtenção dos autovalores e autovetores de uma **matriz de covariâncias** (dados originais) ou **matriz de correlações** (dados padronizados), que é uma operação algébrica conhecida como **decomposição espectral** da matriz de variâncias e covariâncias ou da matriz de correlações.

- As variâncias dos CP´s são os autovalores (raízes características) desta matriz de covariâncias ou da matriz de correlações.

- Autovalores negativos não são possíveis para uma matriz de covariâncias ou de correlações.

- A cada **autovalor**, λ_i, está associado um **autovetor** de componentes $\{a_{i1}, a_{i2}, \cdots, a_{ip}\}$, que são os **coeficientes** do i-ésimo componente principal:

$$CP_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p.$$

PROPRIEDADES IMPORTANTES:

- $\sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} var(X_i)$, no caso de se usar a matriz de variâncias e covariâncias das variáveis originais.

- $\sum_{i=1}^{p} \lambda_i = p$ (número de variáveis originais), no caso da ACP basear-se nas variáveis padronizadas (matriz de correlações).

Exemplo 1. Consideremos as cinco medidas feitas nos corpos das 49 pardocas (planilha **Pardocas**)

Variável	X_1	X_2	X_3	X_4	X_5	**Total**
Média	157.98	241.33	31.46	18.47	20.83	–
Variância	13.35	25.68	0.632	0.318	0.982	**40.969**

Em que 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.

Note que as variáveis apresentam variâncias bem diferentes e as correlações entre as variáveis são todas positivas e superiores a 0,52, indicando que a ACP terá êxito.

Calculando a correlação linear de Pearson entre as variáveis, obtemos:

Correlação: X1; X2; X3; X4; X5

	X1	X2	X3	X4
X2	0.735			
X3	0.662	0.674		
X4	0.645	0.769	0.763	
X5	0.608	0.530	0.526	0.606

Conteúdo da Célula: Correlação de Pearson

Importante:

- Quando as variâncias das variáveis forem muito diferentes é indicado realizar a ACP com os dados padronizados.
- A padronização dos dados das variáveis não altera as correlações entre elas e será utilizada neste exemplo.

Para verificar fazemos **Calc > Padronização**, em **Colunas de entrada** escolher as variáveis X1, ..., X5,em **Armazenar resultados em:** digitar Z1, Z2, Z3, Z4, Z5 e marcar a opção **Subtrair média e dividir pelo desvio padrão**.

Calculando a correlação entre as variáveis padronizadas, podemos verificar que são iguais às calculadas com as variáveis originais.

Correlação: Z1; Z2; Z3; Z4; Z5

	Z1	Z2	Z3	Z4
Z2	0.735			
Z3	0.662	0.674		
Z4	0.645	0.769	0.763	
Z5	0.608	0.530	0.526	0.606

Conteúdo da Célula: Correlação de Pearson

Para realizar a ACP dos dados da planilha **Pardocas**, entramos em **Estat > Multivaridado > Componentes Principais**, em **Variáveis** selecionamos **X1-X5** e em **Tipo de Matriz** marcamos **Correlação** se quisermos trabalhar com os <u>dados padronizados</u>.

Os resultados serão idênticos se usarmos as **Variáveis** Z1-Z5 e em **Tipo de Matriz** marcarmos qualquer uma das opções.

Análise de Componentes Principais: X1; X2; X3; X4; X5

Autoanálise (Autovalores e Autovetores) da Matriz de Correlação

Autovalor	3.6173	0.5313	0.3856	0.3015	0.1644
Proporção	0.723	0.106	0.077	0.060	0.033
Acumulado	0.723	0.830	0.907	0.967	1.000

Neste quadro aparecem os autovalores da matriz de correlação e as proporções da variabilidade total dos dados que é explicada pelos $CP's$.

Note que somente o primeiro autovalor é superior a um.

A soma dos autovalores é igual a 5, que é o número de variáveis medidas nas pardocas, porque usamos os dados padronizados.

Os coeficientes dos componentes principais também são listados:

Variável	CP1	CP2	CP3	CP4	CP5
X1	0.452	0.062	0.687	-0.422	-0.376
X2	0.462	-0.297	0.348	0.545	0.530
X3	0.450	-0.329	-0.451	-0.606	0.343
X4	0.471	-0.191	-0.409	0.390	-0.650
X5	0.398	0.873	-0.190	0.072	0.194

O primeiro componente principal é uma combinação linear das variáveis padronizadas:

$$CP1 = 0{,}452X_1 + 0{,}462X_2 + 0{,}450X_3 + 0{,}471X_4 + 0{,}398X_5$$

e explica 72,3% da variabilidade total dos dados padronizados e os outros quatro CP's explicam, juntos, somente 27,7%.

Autovetores

Variável	CP1	CP2	CP3	CP4	CP5
X1	0.452	0.062	0.687	-0.422	-0.376
X2	0.462	-0.297	0.348	0.545	0.530
X3	0.450	-0.329	-0.451	-0.606	0.343
X4	0.471	-0.191	-0.409	0.390	-0.650
X5	0.398	0.873	-0.190	0.072	0.194

Como os coeficientes do CP1 são aproximadamente iguais, podemos interpretar Z_1 como um **índice de tamanho das pardocas**, porque pondera todas as medidas.

Portanto, 72,3% da variabilidade nos dados padronizados está relacionada com as diferenças de tamanho das pardocas.

O $CP2$ é um contraste entre as variáveis X_2 (extensão alar), X_3 (comprimento do bico e cabeça) e X_4 (comprimento do úmero) de um lado e X_5 (comprimento da quilha do esterno) do outro.

$$CP2 = 0{,}062X_1 - 0{,}2597X_2 - 0{,}329X_3 - 0{,}191X_4 + 0{,}873X_5$$

Representa um **índice de diferença de forma** entre pardocas e assumirá valores altos para valores altos de X_5 e baixos de X_2, X_3 e X_4 e assumirá valores baixos e positivos para valores altos de X_2, X_3 e X_4 e valores baixos de X_5.

Juntos, os dois primeiros componentes principais explicam 83% da variabilidade inicial das variáveis padronizadas.

Podemos calcular os **escores** das 49 pardocas relativos a cada um dos CP's, substituindo, por exemplo, na expressão do CP1:

$$CP1 = 0.452X_1 + 0.462X_2 + 0.451X_3 + 0.471X_4 + 0.398X_5$$

os <u>valores padronizados</u> das variáveis $X_1, X_2, \ldots, X_5$ de cada uma das pardocas.

O Minitab pode calcular <u>automaticamente</u> os escores de cada observação para os dois primeiros CP's.

Basta indicar em **Número de componentes para calcular** o número 2, clicar em **Armazenamento** e em **Escores** escrever os nomes **Escore1 Escore2** das colunas onde os escores serão armazenados.

Com os escores calculados, podemos construir um gráfico de dispersão para estudar a distribuição das aves sobreviventes ou não, com base nos índices de tamanho ($CP1$) e de diferença de formas ($CP2$).

Em **Gráficos > Gráfico de dispersão** marcamos a opção **Com grupos**, em **Variáveis Y** escolhemos **Escore2** e em **Variáveis X** escolhemos **Escore1**. Em **Variáveis categóricas para o Agrupamento** escolher a coluna **Sobreviveu** e em **Múltiplos Gráficos** marcar a opção **Sobrepostas no mesmo gráfico**.

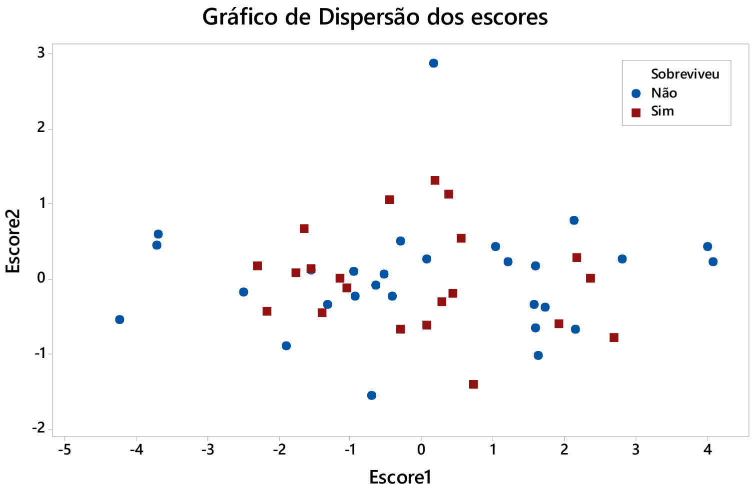


Gráfico de Dispersão dos escores
Sobreviveu
Não
Sim
Escore2
Escore1
3
2
1
0
-1
-2
-5
-4
-3
-2
-1
0
1
2
3
4

A partir da distribuição dos escores dos dois primeiros CP's, Bryan Manly comentou que:

- As pardocas com escores extremos (muito altos ou muito baixos) de $CP1$ não sobreviveram, ou seja, a seleção estabilizadora pode ter agido contra os pássaros muito grandes ou muito pequenos.
- Com um pouco de "esforço" e "boa vontade" nós podemos perceber que o mesmo acontece em relação ao $CP2$, diferença de forma.

UMA APLICAÇÃO

Como o CP1 explica 72,3% da variância total dos dados originais, podemos comparar as médias dos escores do CP1 dos grupos de aves sobreviventes e de aves não sobreviventes usando um teste t para duas amostras.

Antes de comparar as médias dos dois grupos, precisamos saber se as suas variâncias podem ser consideradas iguais ou não, testando

$$H_0: \sigma^2_{Não} = \sigma^2_{Sim} \ versus \ H_a: \sigma^2_{Não} \neq \sigma^2_{Sim}$$

Em S**tat > Estatísticas básicas > Teste para 2 variâncias** marcamos a opção: **As duas amostras estão em uma coluna**; em **Amostras** selecionamos a coluna **Escore1** e em **Identificação de amostra** selecionamos a coluna **Sobreviveu**.

Em **Opções > Razão** escolhemos **(variância da amostra1)/ (variân-cia da amostra 2)** e marcamos a opção **Usar o teste e intervalos de confiança com base na distribuição normal**. Em **Resultados** marcamos somente a opção **Teste**.

Teste

Hipótese nula	H_0: σ_1^2 / σ_2^2 = 1
Hipótese alternativa	H_1: σ_1^2 / σ_2^2 ≠ 1
Nível de significância	α = 0.05

Método	Estatística de teste	GL1	GL2	Valor-p
F	2.09	27	20	0.093

O valor-p $= 0{,}093 > 0{,}05$ indica que a hipótese de igualdade das variâncias (H_0: $\sigma_{Não}^2 = \sigma_{Sim}^2$) deve ser aceita, o que permite concluir que as variâncias de Escore1 dos dois grupos de pardocas devem ser consideradas iguais.

Para comparar as médias: em **Estat > Estatísticas básicas > Teste t para duas amostras**, marcamos a opção **As duas amostras estão em uma coluna**; em **Amostras** selecionamos a coluna **Escore1,** em **Identificação da amostra** selecionamos a coluna **Sobreviveu** e em **Opções** marcamos **Assumir variâncias iguais.**

Método

μ_1: média de Escore1 quando Sobreviveu = Não
μ_2: média de Escore1 quando Sobreviveu = Sim
Diferença: $\mu_1 - \mu_2$

Assumiu-se igualdade de variâncias para esta análise.

Teste

Hipótese nula	$H_0: \mu_1 - \mu_2 = 0$
Hipótese alternativa	$H_1: \mu_1 - \mu_2 \neq 0$

Valor-T	GL	Valor-p
0.33	47	0.746

A comparação das duas médias pelo teste t resultou não significativa (valor-p $= 0{,}746$) indicando que as médias dos escores de CP1 dos dois grupos de pardocas devem ser consideradas iguais, ou seja, os dois grupos apresentam o mesmo índice médio de tamanho.

CRITÉRIOS PARA ESCOLHA DO NÚMERO IDEAL DE CP´s

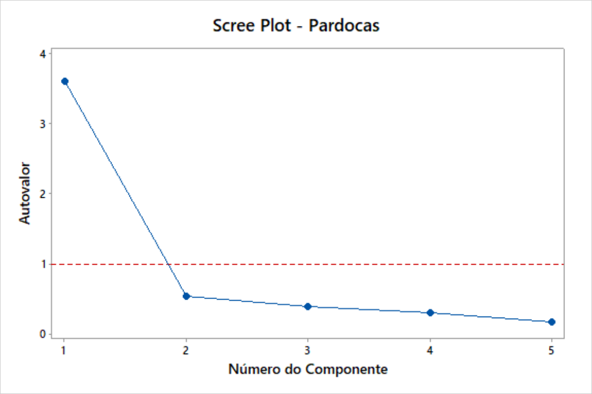
- Escolher um número de CP´s que explique uma porcentagem da variação total das variáveis superior a 90% ou 95%.

- Se o estudo estiver baseado na matriz de correlações (dados padronizados), escolher os CP's relativos a autovalores superiores a um ($\lambda_i \geq 1$), ou seja, mantem-se no estudo os CP's que conseguem explicar ao menos a quantidade de variabilidade de uma variável padronizada.

- Escolher o ponto (cotovelo) do *Scree-plot* em que os valores de λ_i tendem a se estabilizar.

Autoanálise (Autovalores e Autovetores) da Matriz de Correlação

Autovalor	3.6173	0.5313	0.3856	0.3015	0.1644
Proporção	0.723	0.106	0.077	0.060	0.033
Acumulado	0.723	0.830	0.907	0.967	1.000



No exemplo das Pardocas temos um único autovalor superior a 1, que explica 72,3% da variação dos dados originais.

O cotovelo indica a escolha por 2 CP's, se quisermos explicar mais de 80% da variação, 3 CP´s.

COMENTÁRIOS GERAIS

- A ACP não pressupõe nenhuma distribuição de probabilidades para as variáveis em estudo.

- Como o objetivo da ACP é <u>resumir</u> as informações das p variáveis originais, sugere-se a escolha de um número pequeno de CP´s.

- Geralmente, quando os CP´s são extraídas da matriz de correlações necessita-se de um número maior de componentes para explicar uma boa parte da variabilidade dos dados.

- CP´s resultantes de autovalores próximos a zero podem ser eliminados do estudo.

Exemplo 2. Considere os dados de poluição atmosférica apresentados na planilha **ACP - Air-pollution.MTW**, de $p = 7$ variáveis registradas ao meio dia em Los Angeles: Velocidade do vento, Radiação solar, monóxido de carbono (CO), monóxido de nitrogênio ou óxido nítrico (NO), dióxido de nitrogênio (NO2), ozônio (O3) e hidrocarboneto (HC). (Fonte: Johnson & Wichern, 2007).

Estatísticas Descritivas: Vento; Radiação; CO; NO; NO2; O3; HC

Variável	Média	Variância
Vento	7.500	2.500
Radiação	73.860	300.520
CO	4.548	1.522
NO	2.190	1.182
NO2	10.048	11.364
O3	9.405	30.979
HC	3.095	0.479

A variância dos dados de radiação solar (X2) é bem maior que a das outras variáveis, principalmente dos hidrocarbonetos.

Matriz CORR1

1.00000	-0.10144	-0.19380	-0.26954	-0.10982	-0.25359	0.15610
-0.10144	1.00000	0.18279	-0.07357	0.11573	0.31912	0.05201
-0.19380	0.18279	1.00000	0.50215	0.55658	0.41093	0.16603
-0.26954	-0.07357	0.50215	1.00000	0.29690	-0.13395	0.23470
-0.10982	0.11573	0.55658	0.29690	1.00000	0.16664	0.44777
-0.25359	0.31912	0.41093	-0.13395	0.16664	1.00000	0.15445
0.15610	0.05201	0.16603	0.23470	0.44777	0.15445	1.00000

As correlações entre as variáveis não são altas, sendo algumas positivas e outras negativas. Esses aspectos podem tornar os resultados da ACP, pouco proveitosos.

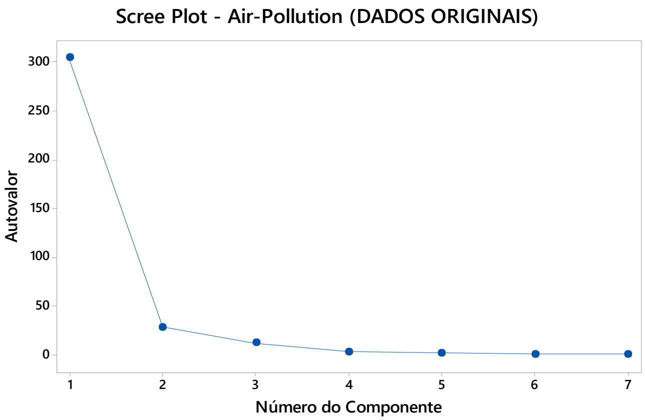
Desconsiderando o comentário sobre as diferenças nas variâncias das variáveis, vamos iniciar o estudo realizando uma ACP com os dados originais (não padronizados).

Em **Estat > Multivariada > Componentes principais > Variáveis** selecionar as variáveis **C1-C7**, em **Tipo de matriz** marcar **Covariância** e em **Gráficos** marcar **Scree-plot - gráfico de perfis de autovalores.**

Autoanálise (Autovalores e Autovetores) da Matriz de Covariância

Autovalor	304.26	28.28	11.46	2.52	1.28	0.53	0.21
Proporção	0.873	0.081	0.033	0.007	0.004	0.002	0.001
Acumulado	0.873	0.954	0.987	0.994	0.998	0.999	1.000

Avaliando o Scree-plot e as porcentagens acumuladas percebe-se que com dois CP´s conseguimos explicar 95,4% da variabilidade total dos dados originais.


Scree Plot - Air-Pollution (DADOS ORIGINAIS)
Autovalor
300
250
200
150
100
50
0
1
2
3
4
5
6
7
Número do Componente

Note que os maiores coeficientes dos 4 primeiros CP´s estão associados às 4 variáveis com maiores variâncias, quais sejam Radiação, O3, NO2 e Vento.

Variável	**CP1**	**CP2**	**CP3**	**CP4**	**CP5**	**CP6**	**CP7**
Vento	-0.010	-0.076	0.031	**-0.920**	0.342	-0.012	-0.170
Radiação	**0.993**	-0.116	0.007	0.000	0.002	-0.003	-0.002
CO	0.014	0.100	-0.183	0.138	0.650	0.564	0.444
NO	-0.005	-0.013	-0.130	0.328	0.643	-0.498	-0.463
NO2	0.024	0.150	**-0.955**	-0.102	-0.207	0.009	-0.105
O3	0.112	**0.973**	0.170	-0.063	-0.000	-0.051	-0.067
HC	0.002	0.024	-0.085	-0.110	0.062	-0.657	0.738

Se refizermos a ACP utilizando a matriz de correlações, ou seja, padronizando as variáveis, obteremos os seguintes resultados:

Autoanálise (Autovalores e Autovetores) da Matriz de Correlação

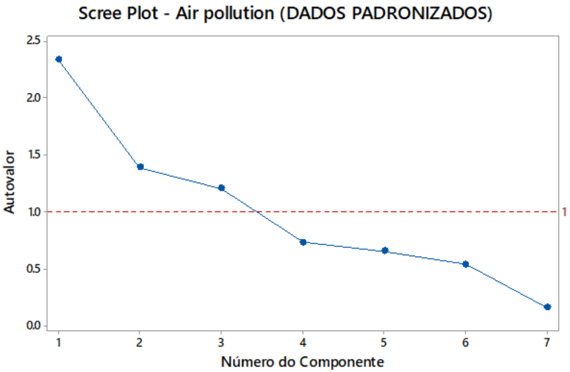
Autovalor	2.3368	1.3860	1.2041	0.7271	0.6535	0.5367	0.1559
Proporção	0.334	0.198	0.172	0.104	0.093	0.077	0.022
Acumulado	0.334	0.532	0.704	0.808	0.901	0.978	1.000

Neste caso, para explicar mais de 90% da variabilidade dos dados padronizados precisaremos de cinco CP's!

Esse resultado **desanimador** (resumir um conjunto de 7 variáveis por 5 CP's) pode ser explicado pelo fato de as variáveis originais apresentarem baixos valores dos coeficientes de correlação.

$\therefore$ A ACP não vai auxiliar na diminuição da dimensão deste problema.

Mesmo assim vamos continuar a análise com os dados padronizados e com três primeiros CP's que são superiores a um e juntos explicam 70,4% da variabilidade das variáveis padronizadas.



Com os dados padronizados, o $scree - plot$ é pouco informativo, não mostrando claramente a localização do cotovelo.

Temos somente três CP's com autovalores maiores que um.

Note que os coeficientes dos 3 CP´s, calculados com as variáveis padronizadas, não evidenciam uma única variável como a mais importante naquele CP.

Autovetores

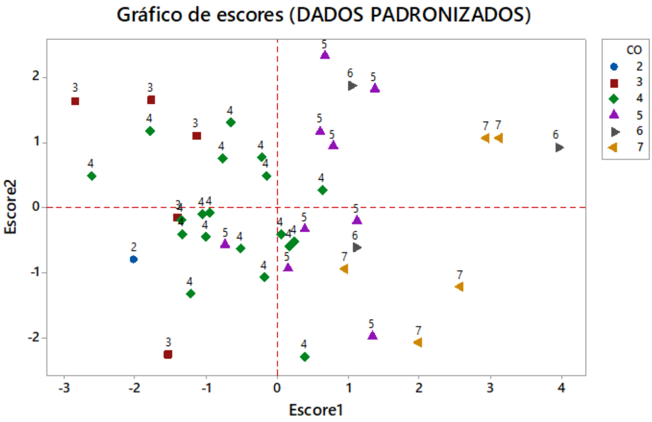
Variável	CP1	CP2	CP3
Vento	-0.237	0.278	-0.643
Radiação	0.206	-0.527	-0.224
CO	0.551	-0.007	0.114
NO	0.378	0.435	0.407
NO2	0.498	0.200	-0.197
O3	0.325	-0.567	-0.160
HC	0.319	0.308	-0.541

Problema: Como dar nomes aos componentes principais?

Exemplo: O primeiro CP é um contraste entre a velocidade do vento e as demais medidas(?)

Valores altos e positivos de CP1 indicam altos níveis de poluição.

Podemos visualizar a distribuição dos escores dos dois primeiros CP's e identificar os valores (altos ou baixos) de CO.



No gráfico dos escores percebe-se a presença de altos valores de CO (5, 6 e 7) nos quadrantes 1 e 4 e baixos valores (2, 3 e 4) nos quadrantes 2 e 3.

A planilha **Exemplos.xlsx** apresenta outros conjuntos de dados sobre o assunto.

Utilize-os para exercitar a ACP, analisando a correlação linear entre as variáveis utilizadas no estudo (o que pode garantir sucesso na análise); olhando para as variâncias das variáveis, decida sobre a necessidade (ou não) de utilizar as variáveis padronizadas e sobre o número de CP´s; tente dar nomes aos CP´s e busque agrupamentos interessantes entre os indivíduos com base nos seus escores.