MAE 5905: Introdução à Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística
Universidade de São Paulo
pam@ime.usp.br
http://www.ime.usp.br/∼ pam

Aula 13

23 de maio de 2023

Sumário

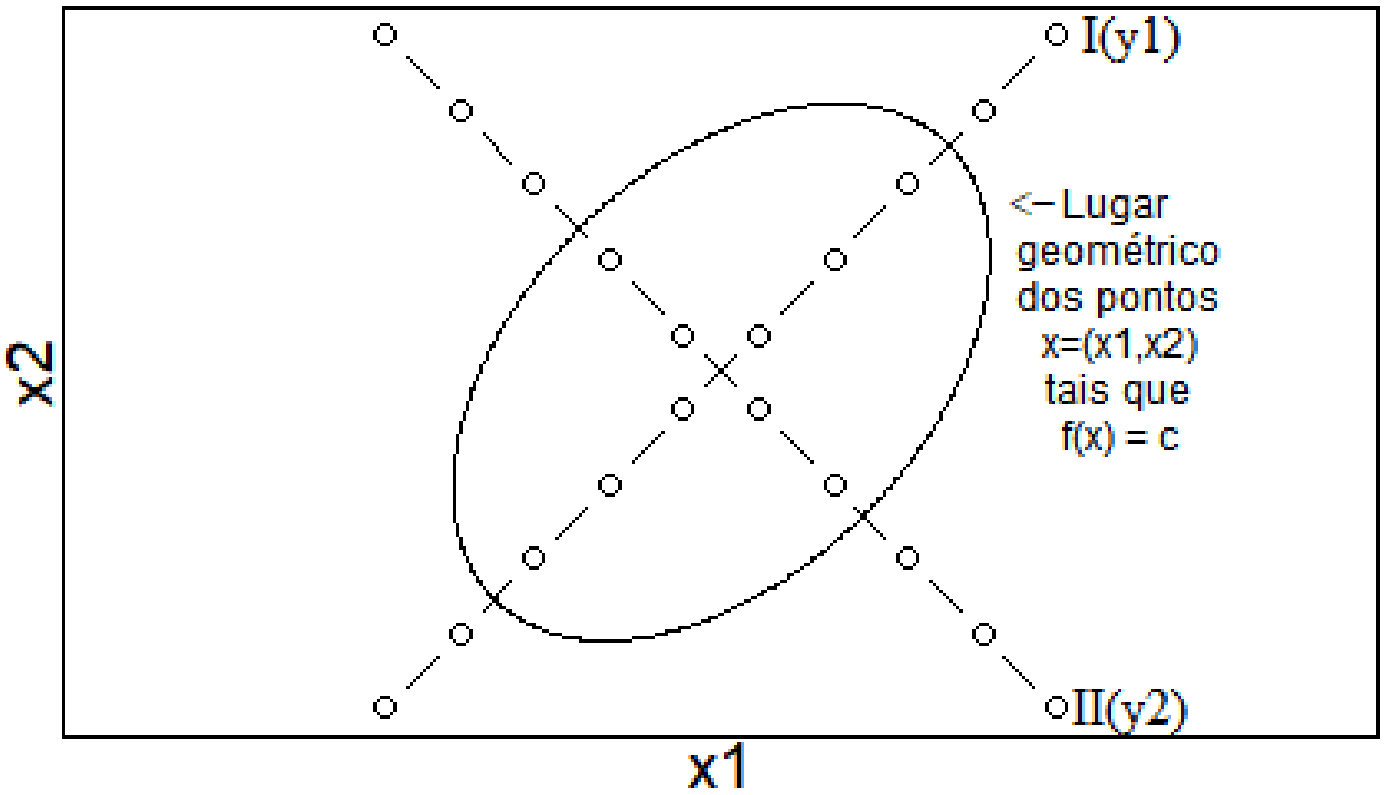
Preliminares

- As técnicas de Análise de Componentes Principais (ACP), Análise de Fatorial (AF) e Análise de Componentes Independentes (ACI) têm como objetivo reduzir a dimensionalidade de observações multivariadas com base em sua estrutura de dependência.
- A ideia que as permeia é a obtenção de poucos fatores, obtidos como funções das características observadas, que conservem, pelo menos aproximadamente, a estrutura de covariância das variáveis originais.
- Esses poucos fatores vão substituir as variáveis originais em análises subsequentes, servindo, por exemplo, como variáveis explicativas em modelos de regressão. Por esse motivo, a interpretação dessas novas variáveis é muito importante.
- Dessas técnicas, ACP e AF são bastante conhecidas há muito tempo. A ACI (ou ICA, em Inglês) é mais recente, da década de 1990–2000, e não muito contemplada em textos de Estatística.

ACP – introdução

- A técnica de Componentes Principais consiste numa transformação ortogonal dos eixos de coordenadas de um sistema multivariado.
- A orientação dos novos eixos é determinada por meio da partição sequencial da variância total das observações em porções cada vez menores de modo que, ao primeiro eixo transformado, corresponda o maior componente da partição da variância; ao segundo eixo transformado, a parcela seguinte e assim por diante.
- Se os primeiros eixos forem tais que uma grande parcela da variância seja explicada por eles, poderemos desprezar os demais e trabalhar apenas com os primeiros em análises subsequentes.
- Consideremos duas variáveis X_1 e X_2 com distribuição normal bivariada com média μ e matriz de covariâncias $\mathbf{\Sigma}$.
- O gráfico correspondente aos pontos em que a função densidade de probabilidade é constante é uma elipse; um exemplo está apresentado na Figura 1. Admitimos dados com distribuição normal apenas para finalidade didática. Em geral, essa suposição não é necessária.

ACP – introdução



ACP – metodologia

- À medida em que a correlação entre X_1 e X_2 aumenta, o comprimento do eixo maior da elipse também aumenta e o do eixo menor diminui até que a elipse se degenera em um segmento de reta no caso limite em que as variáveis são perfeitamente correlacionadas, *i.e.*, em que o correspondente coeficiente de correlação linear é igual a 1.
- Na Figura 1, o eixo I corresponde ao eixo maior e o eixo II, ao menor. O eixo I pode ser expresso por intermédio de uma combinação linear de X_1 e X_2, ou seja

$$Y_1 = \beta_1 X_1 + \beta_2 X_2 \tag{1}$$

- No caso extremo, em que X_1 e X_2 são perfeitamente correlacionadas, toda a variabilidade pode ser explicada por meio de Y_1.
- Quando a correlação entre X_1 e X_2 não é perfeita, Y_1 explica apenas uma parcela de sua variabilidade. A outra parcela é explicada por meio de um segundo eixo, a saber

$$Y_2 = \gamma_1 X_1 + \gamma_2 X_2. \tag{2}$$

ACP – metodologia

Na Figura 2 apresentamos um gráfico de dispersão correspondente a n observações (X_{1i}, X_{2i}) do par (X_1, X_2).

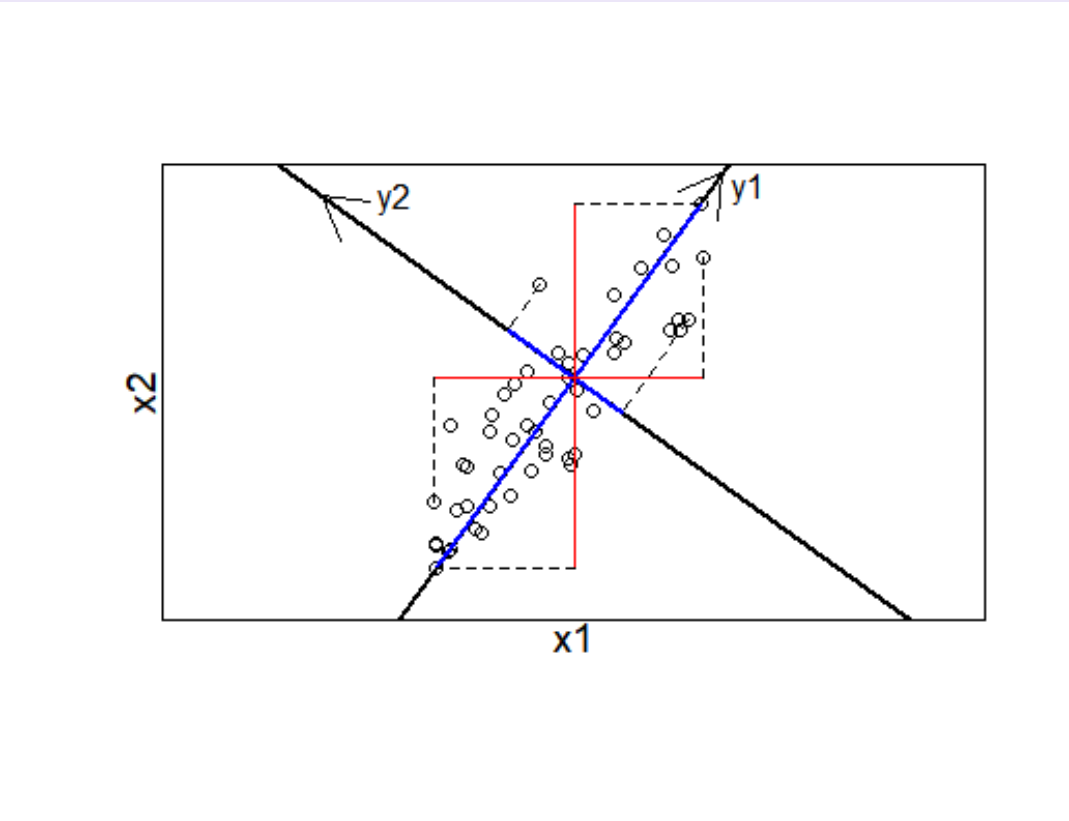


Figura: Gráfico de dispersão de n observações do par (X_1, X_2).

ACP – metodologia

- A variabilidade no sistema de eixos correspondente a (X_1, X_2) pode ser expressa como

$$\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)^2 + \sum_{i=1}^{n}(X_{2i} - \overline{X}_2)^2$$

em que $\overline{X}_1$ e $\overline{X}_2$ são, respectivamente, as médias dos n valores de X_{1i} e X_{2i}.

- No sistema de eixos correspondente a (Y_1, Y_2), a variabilidade é expressa como

$$\sum_{i=1}^{n}(Y_{1i} - \overline{Y}_1)^2 + \sum_{i=1}^{n}(Y_{2i} - \overline{Y}_2)^2$$

em que $\overline{Y}_1$ e $\overline{Y}_2$ têm interpretações similares a $\overline{X}_1$ e $\overline{X}_2$.

ACP – metodologia

- Se a correlação entre X_1 e X_2 for "grande", é possível obter valores de β_1, β_2, γ_1, γ_2 de tal forma que Y_1 e Y_2 em (1) e (2) sejam tais que

$$\sum_{i=1}^{n}(Y_{1i} - \overline{Y}_1)^2 \gg \sum_{i=1}^{n}(Y_{2i} - \overline{Y}_2)^2.$$

- Nesse caso, podemos utilizar apenas Y_1 como variável para explicar a variabilidade de X_1 e X_2.
- Para descrever o processo de obtenção das componentes principais, consideremos o caso geral em que $\mathbf{x}_1, \ldots, \mathbf{x}_n$ corresponde a uma amostra aleatória de uma variável $\mathbf{X} = (X_1, \ldots, X_p)^\top$ com p componentes e para a qual o vetor de médias é $\boldsymbol{\mu}$ e a matriz de covariâncias é $\boldsymbol{\Sigma}$.

ACP – metodologia

- Sejam $\bar{\mathbf{x}} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i$ e $\mathbf{S} = (n-1)^{-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$, respectivamente, o vetor de médias amostrais e a matriz de covariâncias amostral.
- A técnica consiste em procurar sequencialmente p combinações lineares (denominadas componentes principais) de $X_1, \ldots, X_p$ tais que à primeira corresponda a maior parcela de sua variabilidade, à segunda, a segunda maior parcela e assim por diante e que, além disso, sejam não correlacionadas entre si.
- A primeira componente principal é a combinação linear

$$Y_1 = \boldsymbol{\beta}_1^\top \mathbf{X} = \beta_{11} X_1 + \ldots + \beta_{1p} X_p$$

com $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1p})^\top$, para a qual a variância $Var(Y_1) = \boldsymbol{\beta}_1^\top \boldsymbol{\Sigma} \boldsymbol{\beta}_1$ é máxima.

ACP – metodologia

- Uma estimativa da primeira componente principal calculada com base na amostra é a combinação linear $\widehat{Y}_1 = \widehat{\boldsymbol{\beta}}_1^\top \mathbf{x}$ para a qual $\widehat{Var(Y_1)} = \widehat{\boldsymbol{\beta}}_1^\top \mathbf{S} \widehat{\boldsymbol{\beta}}_1$ é máxima.
- Este problema não tem solução sem uma restrição adicional, pois se tomarmos $\widehat{\boldsymbol{\beta}}_1^* = c\widehat{\boldsymbol{\beta}}_1$ com c denotando uma constante arbitrária, podemos tornar a variância $\widehat{Var(Y_1)} = Var(\widehat{\boldsymbol{\beta}}_1^* \mathbf{x})$ arbitrariamente grande, tomando c arbitrariamente grande.
- A restrição adicional mais usada consiste em padronizar $\boldsymbol{\beta}_1$ por meio de $\boldsymbol{\beta}_1^\top \boldsymbol{\beta}_1 = 1$.
- Consequentemente, o problema de determinação da primeira componente principal se resume a obter $\widehat{\boldsymbol{\beta}}_1$ tal que

$$\text{Maximize} \quad \boldsymbol{\beta}_1^\top \boldsymbol{\Sigma} \boldsymbol{\beta}_1, \quad \text{sujeito a } \boldsymbol{\beta}_1^\top \boldsymbol{\beta}_1 = 1.$$

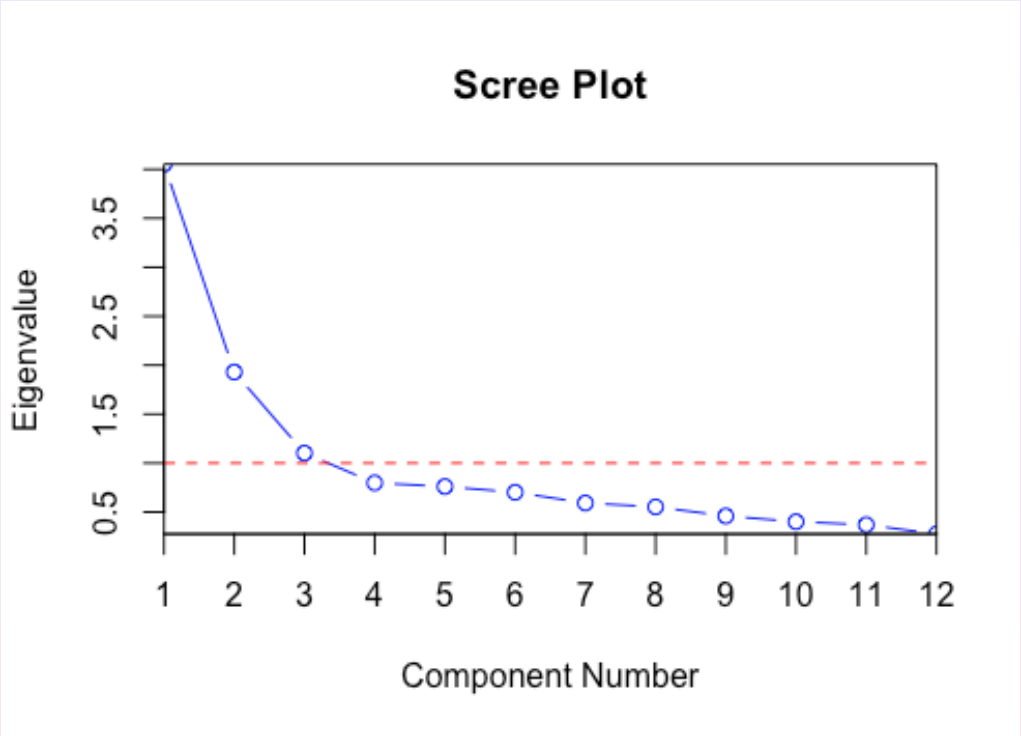
ACP – metodologia

- A solução desse problema pode ser encontrada por meio da aplicação de multiplicadores de Lagrange.
- Dada a primeira componente principal, obtém-se a segunda, $Y_2 = \boldsymbol{\beta}_2^\top \mathbf{X}$, por meio da maximização de $\boldsymbol{\beta}_2^\top \boldsymbol{\Sigma} \boldsymbol{\beta}_2$ sujeito a $\boldsymbol{\beta}_2^\top \boldsymbol{\beta}_2 = 1$ e $\boldsymbol{\beta}_1^\top \boldsymbol{\beta}_2 = 0$ (para garantir a ortogonalidade).
- Note que a ortogonalidade das componentes principais implica que a soma de suas variâncias seja igual à variância total do sistema de variáveis. Esse procedimento é repetido até a determinação da p-ésima componente principal.
- Os coeficientes das componentes principais estimadas são os autovetores $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ da matriz $\mathbf{S}$ e suas variâncias são os autovalores $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \ldots \geq \widehat{\lambda}_p$ correspondentes.

ACP – metodologia

- Como a variância total do sistema é $\text{tr}(\mathbf{S}) = \sum_{i=1}^{p} \widehat{\lambda}_i$, a contribuição da i-ésima componente principal é $\widehat{\lambda}_i / \text{tr}(\mathbf{S})$. Lembrando que $\mathbf{S} = \sum_{i=1}^{p} \widehat{\lambda}_i \widehat{\boldsymbol{\beta}}_i \widehat{\boldsymbol{\beta}}_i^{\top}$ podemos verificar quão bem ela pode ser aproximada com um menor número de componentes principais. Detalhes podem ser obtidos na Nota de Capítulo 2.
- Na prática, a determinação do número de componentes principais a reter como novo conjunto de variáveis para futuras análises pode ser realizada por meio do elbow plot, também conhecido como scree plot, que consiste num gráfico cartesiano com os autovalores no eixo vertical e os índices correspondentes às suas magnitudes (em ordem decrescente) no eixo das abscissas.
- Um exemplo está apresentado na Figura 3

ACP – metodologia



ACP – metodologia

- A ideia é acrescentar componentes principais até que sua contribuição para a explicação da variância do sistema não apresente contribuições "relevantes". Com base na Figura 3, apenas as duas primeiras componentes principais poderiam ser suficientes, pois a contribuição das seguintes é apenas marginal.
- Suponhamos que r componentes principais, $Y_1, \ldots, Y_r$ explicam uma parcela "substancial"da variabilidade do sistema multivariado. Então para a k-ésima unidade amostral, podemos substituir os valores das variáveis originais $\mathbf{x}_k = (x_{1k}, \ldots, x_{pk})^\top$ pelos correspondentes **escores** associados às componentes principais, nomeadamente, $\widehat{Y}_{1k}, \ldots, \widehat{Y}_{rk}$ com $\widehat{Y}_{ik} = \widehat{\boldsymbol{\beta}}_i^\top \mathbf{x}_k$.
- Infelizmente, nem a matriz de covariâncias nem os correspondentes autovalores são invariantes relativamente a mudanças de escala. Em outras palavras, mudanças nas unidades de medida das características $X_1, \ldots, X_p$ podem acarretar mudanças na forma e na posição dos elipsoides correspondentes a pontos em que a função densidade é constante.

ACP – metodologia

- É difícil interpretar combinações lineares de características com unidades de medida diferentes e uma possível solução é padronizá-las por meio de transformações do tipo $Z_{ij} = (X_{ij} - \overline{X}_i)/S_i$ em que $\overline{X}_i$ e S_i representam, respectivamente a média e o desvio padrão de X_{ij} antes da obtenção das componentes principais.
- A utilização da matriz de correlações **R** obtida por meio dessa transformação pode ser utilizada na determinação das componentes principais; no entanto, os resultados são, em geral, diferentes e não é possível passar de uma solução a outra por meio de uma mudança de escala dos coeficientes.
- Se as características de interesse foram medidas com as mesmas unidades, é preferível extrair as componentes principais utilizando a matriz de covariâncias amostrais **S**.

ACP – metodologia

- Lembrando que a i-ésima componente principal é $Y_i = \beta_{i1}X_1 + \ldots + \beta_{ip}X_p$ do sistema multivariado, se as variáveis $X_1, \ldots, X_p$ tiverem variâncias similares ou forem variáveis padronizadas, os coeficientes β_{ij} indicam a importância e a direção da j-ésima variável relativamente à i-ésima componente principal.
- Nos casos em que as variâncias das variáveis originais são diferentes, convém avaliar sua importância relativa na definição das componentes principais por meio dos correspondentes coeficientes de correlação. O vetor de covariâncias entre as variáveis originais e a i-ésima componente principal é

$$Cov(\mathbf{I}_p\mathbf{X}, \boldsymbol{\beta}_i\mathbf{X}) = \mathbf{I}_p\boldsymbol{\Sigma}\boldsymbol{\beta}_i = \boldsymbol{\Sigma}\boldsymbol{\beta}_i = \lambda_i\boldsymbol{\beta}_i$$

pois $(\boldsymbol{\Sigma} - \lambda_i\mathbf{I}_p)\boldsymbol{\beta}_i = \mathbf{0}$ (ver Nota de Capítulo 1).

ACP – metodologia

- Uma estimativa desse vetor de covariâncias é $\widehat{\lambda}_i \widehat{\boldsymbol{\beta}}_i$.
- Consequentemente, uma estimativa do coeficiente de correlação entre a j-ésima variável original e a i-ésima componente principal é

$$\widehat{Corr}(X_j, \beta_{ij}) = \frac{\widehat{\lambda}_i \widehat{\beta}_{ij}}{\sqrt{\widehat{\lambda}_i} s_j} = \frac{\sqrt{\widehat{\lambda}_i} \widehat{\beta}_{ij}}{s_j}$$

  em que s_j é o desvio padrão amostral de X_j.
- As componentes principais podem ser encaradas como um conjunto de variáveis latentes (ou fatores) não correlacionados que servem para descrever o sistema multivariado original sem as dificuldades relacionadas com sua estrutura de correlação. Os coeficientes associados a cada componente principal servem para descrevê-las em termos das variáveis originais.
- A comparação entre unidades realizada por meio da componente principal Y_i é independente da comparação baseada na componente Y_j. No entanto, essa comparação pode ser ilusória se essas componentes principais não tiverem uma interpretação simples.

ACP – metodologia

- Como, em geral, isso não é a regra, costuma-se utilizar essa técnica como um passo intermediário para a obtenção de um dos possíveis conjuntos de combinações lineares ortogonais das variáveis originais passíveis de interpretação como variáveis latentes.
- Esses conjuntos estão relacionados entre si por meio de rotações rígidas (transformações ortogonais) e são equivalentes no sentido de aproximar as correlações entre as variáveis originais. Apesar de essas rotações rígidas implicarem uma perda da característica de ordenação das componentes principais em termos de porcentagem de explicação da variabilidade do sistema multivariado, muitas vezes produzem ganhos interpretativos.
- Há muitas opções computacionais para a análise de componentes principais no sistema `R`, dentre as quais destacamos (funções e pacotes entre parênteses) prcomp (`stats`), princomp (`stats`) e pca (`FactoMineR`). Como há vários métodos tanto para a extração quanto para a rotação das componentes principais, nem sempre os resultados coincidem. A análise deve ser realizada por tentativa e erro tendo como objetivo um sistema com interpretação adequada.

ACP – Exemplo 1

- **Exemplo 1.** Num estudo em que se pretendia avaliar o efeito de variáveis climáticas na ocorrência de suicídios por enforcamento na cidade de São Paulo foram observadas X_1 = temperatura máxima, X_2 = temperatura mínima, X_3 = temperatura média, X_4 = precipitação e X_5 = nebulosidade diárias para o período de 01/07/2006 e 31/07/2006. Os dados estão disponíveis em `http://www.ime.usp.br/~jmsinger/MorettinSinger/suicidios.xls` e detalhes em Zerbini et al. (2018)],
- Para reduzir o número de variáveis a serem utilizadas como variáveis explicativas numa regressão logística tenso como variável resposta a ocorrência de suicídios por enforcamento nesse período consideramos uma análise de componentes principais.
- Os coeficientes das cinco componentes bem como as porcentagem da variância total do sistema explicada por cada uma delas (além da porcentagem acumulada correspondente) estão indicados na Tabela 1.

ACP – Exemplo 1

Tabela: Coeficientes das componentes principais e porcentagens da variância explicada: Exemplo 1

Variável	Componentes principais				
	CP1	CP2	CP3	CP4	CP5
Temperatura máxima	0,93	-0,25	0,05	0,26	0,06
Temperatura mínima	0,91	0,29	-0,18	-0,24	0,06
Temperatura média	0,99	-0,02	-0,03	0,00	-0,11
Precipitação	0,12	0,75	0,66	0,02	0,00
Nebulosidade	-0,12	0,85	-0,50	0,14	-0,01
% Variância	54	28	14	3	0
% Acumulada	54	82	97	100	100

ACP – Exemplo 1

- Uma análise do gráfico da escarpa sedimentar correspondente representado na Figura 4, conjuntamente com um exame da variância acumulada na Tabela 1 sugere que apenas duas componentes principais podem ser empregadas como resumo, retendo 82% da variância total.
- A primeira componente principal pode ser interpretada como percepção térmica e segunda, percepção de acinzentamento. Valores dessas novas variáveis para cada unidade amostral são calculadas como $CP_1 = 0,93 * tempmax + 0,91 * tempmin + 0,99 * tempmed$ e $CP_2 = 0,75 * precip + 0,85 * nebul$, com as variáveis originais devidamente padronizadas.
- O gráfico *biplot* disposto na Figura 5 é conveniente para representar a relação entre as duas componentes principais e as variáveis originais.

ACP – Exemplo 1

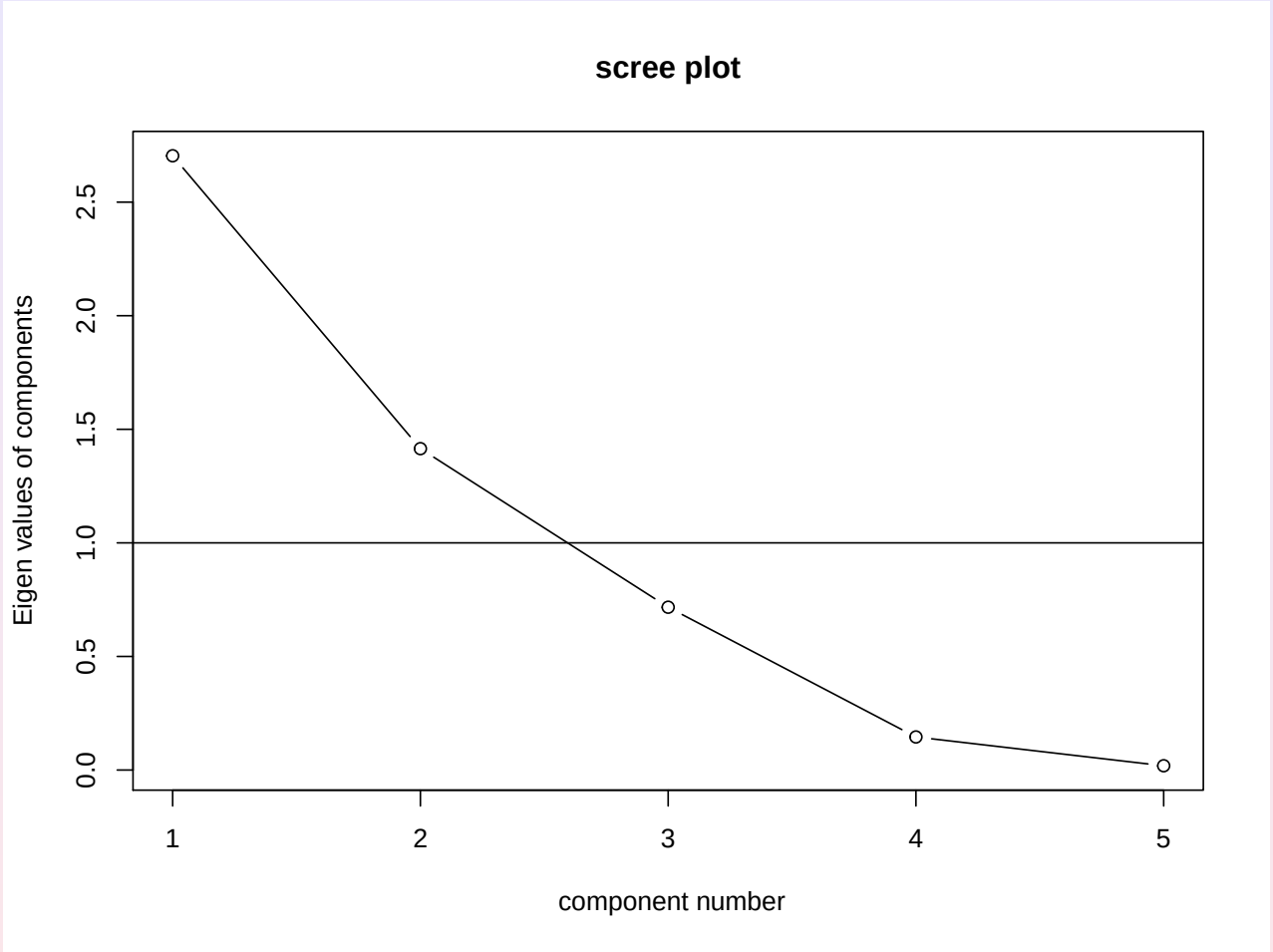


Figura: Gráfico da escarpa sedimentar (ou do cotovelo) para os dados do Exemplo 1.

ACP – Exemplo 1

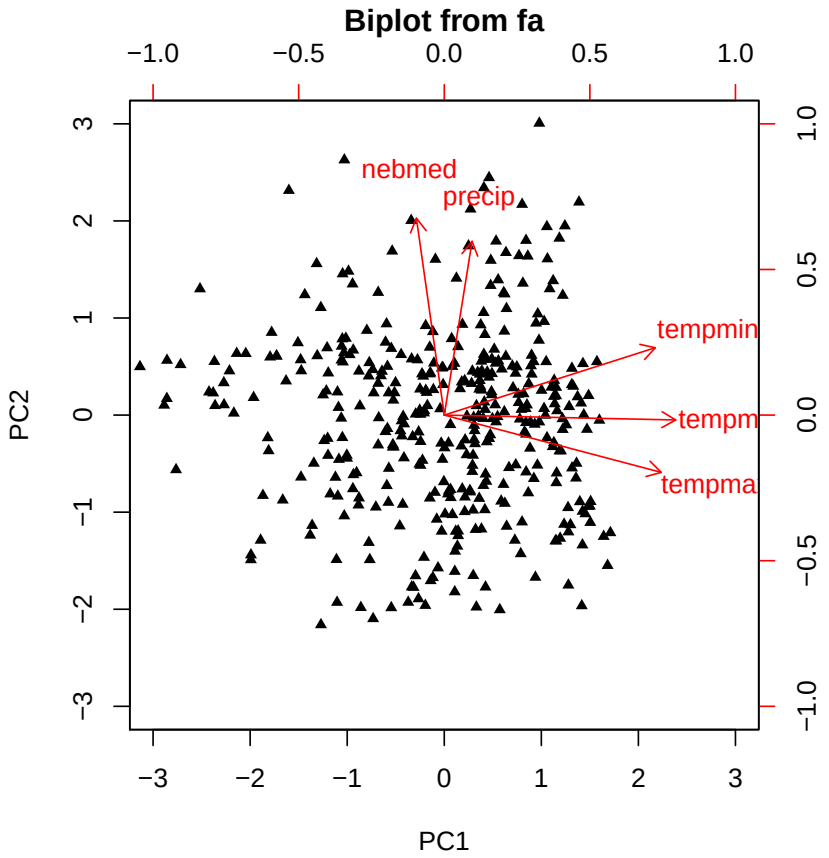


Figura: Gráfico *biplot* para os dados do Exemplo 1.

Referências

Hastie, T., Tibshirani, R. and Friedman, J. (2017). *The Elements of Statistical Learning*, 2nd Edition, Springer.

Härdle, W.K. and Simar, L. (2015). *Applied Multivariate Statistical Analysis.* Springer.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). *Introduction to Statistical Learning*. Springer.

Morettin, P. A. e Singer, J. M. (2022). *Estatística e Ciência de Dados*. LTC: Rio de Janeiro.