Universidade Federal de Uberlândia
Instituto de Matemática e Estatística
Curso de graduação em Estatística

Álgebra Linear
Prof. Dr. Augusto Duarte Pena

Análise de Componentes Principais

Daniel Barreto de Oliveira
Henrique Tomaz Gonzaga

04 de maio de 2025

1 Objetivo

A Análise de Componentes Principais (*Principal Component Analysis*; PCA) é uma técnica estatística que reduz a dimensionalidade de um conjunto de dados, transformando-o em um novo espaço de características definido por componentes principais. Esses componentes retêm a maior parte da variância dos dados originais [1]. Ao simplificar a estrutura de atributos problema, o PCA melhora a eficiência computacional e é amplamente utilizado, por exemplo, no pré-processamento de dados para algoritmos de aprendizado de máquina. Adicionalmente, o método ajuda a mitigar questões como multicolinearidade e sobreajuste [1].

A ideia central da PCA consiste em, dada uma nuvem de pontos no espaço original, encontrar uma projeção ortogonal em um hiperplano de dimensão reduzida que maximize a variância dos dados. Essa projeção preserva ao máximo a informação dos dados originais, capturando as direções de maior variância (componentes principais) [2]. Se tivermos um conjunto de dados de diversas espécies de plantas ou uma lista de parâmetros diagnósticos para uma certa doença, podemos aplicar PCA para modelar os dados e avaliar os possíveis agrupamentos formados.

2 Fundamentação Teórica

2.1 Centralização dos dados

Para encontrar a matriz de covariâncias é necessário primeiro centralizar os dados. Dado um conjunto de dados com m observações e n campos representados pela matriz

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix},$$

o vetor de médias de cada coluna dado por $\bar{x}$ é calculado da seguinte forma:

$$\bar{x} = \frac{1}{m} \cdot \begin{bmatrix} \sum_i x_{i1} \\ \sum_i x_{i2} \\ \vdots \\ \sum_i x_{in} \end{bmatrix}, x_{ij} \in X.$$

Dessa forma, os dados centralizados de X são encontrados subtraindo de cada coordenada a média de sua respectiva coluna, ou seja,

$$X_c = X - e\bar{x}^T,$$

sendo $e = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T$.

2.2 Matriz de Covariância

A covariância entre dois vetores $u = \begin{bmatrix} u_1 & \cdots & u_m \end{bmatrix}^T$ e $v = \begin{bmatrix} v_1 & \cdots & v_m \end{bmatrix}^T$ é dada por:

$$Cov(u, v) = E[(u - \bar{u}) \cdot (v - \bar{v})] = \sum_i \frac{(u_i - \bar{u}) \cdot (v_i - \bar{v})}{m - 1}.$$

Considerando $u_c = u - \bar{u}$ e $v_c = v - \bar{v}$, tem-se que

$$Cov(u, v) = E[u_c \cdot v_c] = \frac{1}{m-1} u_c^T v_c.$$

Tomando $x_1, \ldots x_n$ como os vetores colunas de X, a matriz de covariâncias de X é a seguinte:

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1}^2 & \sigma_{n2}^2 & \cdots & \sigma_{nn}^2 \end{bmatrix} = \frac{1}{m-1} X_c^T X_c,$$

com $\sigma_{ij}^2 = Cov(x_i, x_j)$ e $i, j = 1, \cdots n$. Uma vez que $Cov(x_i, x_j) = Cov(x_j, x_i)$, ou seja, $\sigma_{ij}^2 = \sigma_{ji}^2$, a matriz Σ é simétrica. Por conta da simetria, tem-se que a matriz Σ é diagonalizável e possui n autovetores linearmente independentes [2].

2.3 Autovalores e autovetores

Os autovalores, de modo geral, de uma matriz A são as raízes de um polinômio característico $p_A(\lambda)$. Para a matriz de covariância Σ, tem-se que os autovalores são as raízes λ_i (com $i = 1, \ldots n$) do polinômio

$$p_\Sigma(\lambda) = \det(\Sigma - I\lambda),$$

tais que

$$\lambda_1 \geq \cdots \geq \lambda_n,$$

com V sendo a matriz formada pelos autovetores $v_1, \ldots v_n$. Essa ordenação visa escolher os k maiores autovalores associados às maiores variâncias. Além disso, como Σ é simétrica, tem-se que seus autovetores podem serem escolhidos de modo que sejam ortonormais dois a dois, sendo assim, $V^{-1} = V^T$ [2].

2.4 Transformação dos dados em novo sistema de coordenadas

Escolhendo k componentes principais, obtém-se uma matriz V_k da forma

$$V_k = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{ik} \\ v_{21} & v_{22} & \cdots & v_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nk} \end{bmatrix},$$

formada pelos k primeiros autovetores. O novo sistema de coordenas para X é definida da seguinte maneira:

$$A = X_c V_k.$$

No qual a matriz A é uma projeção de X no hiperplano de k dimensões que possui a maior variação dos dados.

2.5 Padronização dos dados

Quando as variáveis possuem escalas diferentes, ao invés de apenas centralizar os dados, é necessário padronizá-los. Um vetor v_p é considerado padronizado quando:

$$v_p = \frac{v - \bar{v}}{\sigma} = \frac{v_c}{\sigma},$$

onde σ representa o desvio padrão do vetor, ou seja, $\sigma = \sqrt{\mathrm{Cov}(v, v)}$.

Definindo D como uma matriz diagonal cujos elementos são as raízes quadradas dos elementos da diagonal de Σ:

$$D = \begin{bmatrix} \sigma_{11} & & & \\ & \sigma_{22} & & \\ & & \ddots & \\ & & & \sigma_{nn} \end{bmatrix},$$

pode-se expressar o conjunto de dados padronizados como:

$$X_p = X_c D^{-1}.$$

Com os dados padronizados, a matriz de covariância é calculada de forma similar ao caso dos dados centralizados:

$$\Sigma_p = \frac{1}{m-1} X_p^T X_p.$$

Esta matriz Σ_p é denominada matriz de correlação e possui a seguinte estrutura:

$$\Sigma_p = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{bmatrix},$$

onde ρ_{ij} representa a correlação entre o i-ésimo e o j-ésimo vetores coluna de X. Assim como a matriz de covariância, a matriz de correlação Σ_p é simétrica, e seus autovalores e autovetores são determinados da mesma maneira que em Σ.

Ao realizar a projeção com base nos dados padronizados, obtém-se uma projeção em um hiperplano que desconsidera as escalas das variáveis do problema. A matriz de correlação é uma forma escalonada da matriz de covariância. [2].

3 Aplicação

Os códigos para a execução do PCA em bancos de dados e os resultados obtidos a seguir foram escritos na linguagem `R` utilizando as bibliotecas `ggplot2` e `dplyr`. O código completo está disponível em PCA.R. No relatório, optou-se por não incluir os códigos de visualização dos parâmetros que auxiliam na descrição dos resultados; mas que estão no mesmo link.

3.1 Problema dos valores ausentes

A função `impute_missing` trata os valores ausentes em um conjunto de dados. Ela passa por cada coluna e, para as colunas numéricas, substitui os valores ausentes (NA) pela média da coluna. A função `mean()`, com a opção `na.rm = TRUE`, ignora os valores ausentes ao calcular a média. Ao final, a função retorna com os dados completos para as análises que seguem.

```
impute_missing <- function(data){
  data_imputed <- data
  for (i in 1:ncol(data)){
    if (is.numeric(data[[i]])){
      data_imputed[[i]][is.na(data_imputed[[i]])] <- mean(
   data_imputed[[i]], na.rm = TRUE)
    }
  }
  return(data_imputed)
}
```

3.2 Transformando as coordenadas do conjunto de dados

A função `run_pca_manual` executa a PCA passo-a-passo. Primeiramente, trata os valores ausentes chamando a função `impute_missing`. Com o parâmetro `scale_data = TRUE`, os dados são padronizados com média zero e desvio padrão um. A função então

calcula a matriz de covariância dos dados e realiza a decomposição espectral dessa matriz, utilizando a função `eigen()`. Com isso, ela calcula os autovalores e autovetores, que representam a variância e as direções dos componentes principais. A função também calcula a projeção dos dados nos componentes principais, retornando uma lista com os autovalores, autovetores e pontuações dos componentes principais.

```
run_pca_manual <- function(data, scale_data = TRUE) {
  data <- impute_missing(data)

  if (scale_data) data <- scale(data, scale = TRUE)

  cov_matrix <- cov(data)

  eigens <- eigen(cov_matrix, symmetric = TRUE)

  scores <- data %*% eigens$vectors

  return(list(eigenvalues = eigens$values, eigenvectors =
   eigens$vectors, scores = scores))
}
```

3.3 Visualizando o plano com dois componentes principais

A função `plot_pca_static` gera os resultados da PCA em um gráfico de dispersão. Ela recebe as pontuações dos componentes principais (resultantes de `run_pca_manual`) e uma coluna de grupos para marcar os pontos no gráfico e diferenciar categorias. A função cria um `data.frame` com as pontuações dos primeiros componentes principais (PC1 e PC2) e utiliza o pacote `ggplot2` para gerar o gráfico. O eixo X representa o `PC1` e o eixo Y representa o `PC2`.

```
plot_pca_static <- function(scores, group_col = NULL) {
  pca_df <- data.frame(scores)
  colnames(pca_df) <- c("PC1", "PC2")

  if (!is.null(group_col)) pca_df$Group <- group_col

  ggplot(pca_df, aes(x = PC1, y = PC2, color = Group)) +
    geom_point(size = 3) +
    labs(title = "PCA - Plot Estático") +
    theme_minimal()
}
```

3.4 Exemplo de aplicação e interpretação

No exemplo de aplicação, a base de dados `iris` é utilizada para demonstrar o processo de PCA. O conjunto de dados consiste em 50 amostras das flores de três espécies de *Iris* (*I. setosa*, *I. virginica* e *I. versicolor*). Quatro variáveis foram medidas em cada amostra: o comprimento e a largura das sépalas e pétalas, em centímetros. Inicialmente, a coluna `Species`, que contém as classes das flores, é removida para que a análise seja realizada apenas nas variáveis numéricas. A função `run_pca_manual` é então aplicada

aos dados, com a opção `scale_data` ativada para padronizar as variáveis. Em seguida, a função `plot_pca_static` é utilizada para criar um gráfico de dispersão dos dois primeiros componentes principais (PC1 e PC2), com a variável `Species` para colorir os pontos. Assim, observamos como os diferentes grupos se separam na projeção dos componentes principais.

```
data(iris)

iris_data <- iris |>
  select(-Species)

pca_result <- run_pca_manual(iris_data, scale_data = TRUE)

plot_pca_static(pca_result$scores, group_col = iris$Species)
```

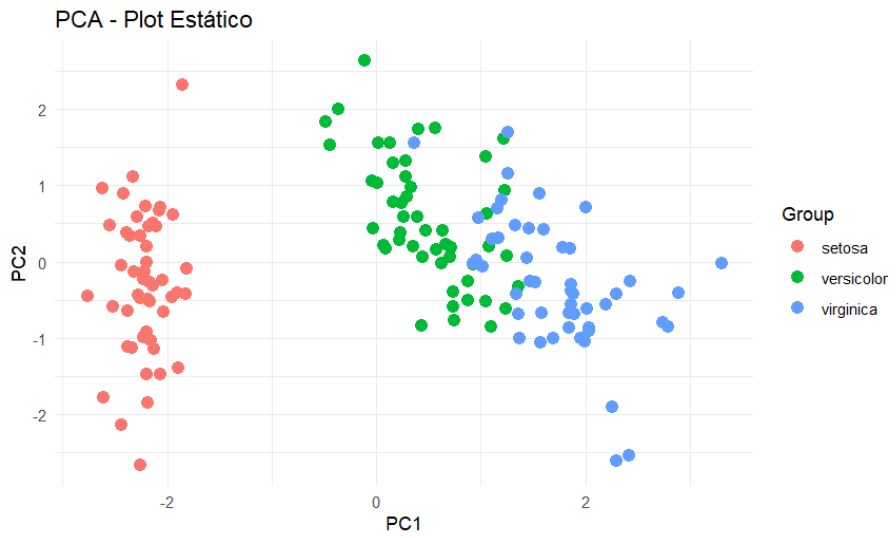

Figura 1: Gráfico gerado pela função `plot_pca_static`

O gráfico de PCA mostra a projeção dos dados nos dois primeiros componentes principais (PC1 e PC2), que explicam 95.81 % da variância total (PC1: 72.96 %, PC2: 22.85 %). Os autovetores revelam a contribuição de cada variável original na formação dos componentes.

- **PC1** possui coeficientes positivos e elevados para: comprimento da sépala e (0.521) e da pétala (0.580) e largura da pétala (0.565).

Flores com scores positivos em PC1 (à direita no gráfico) têm pétalas e sépalas maiores, típicas de *I. virginica*, enquanto scores negativos (à esquerda) correspondem a flores menores, como as de *I. setosa*.

- **PC2** é definido basicamente pela largura da sépala (-0.923), separando flores com sépalas mais largas (pontos inferiores no gráfico, como algumas *I. setosa*) daquelas com sépalas estreitas (pontos superiores, como *I. versicolor* e *I. virginica*).

A separação do *cluster* de *I. setosa* mostra sua diferença (valores baixos em PC1). A sobreposição parcial entre os *clusters* de *I. versicolor* e *I. virginica* (centro/direita)

ocorre porque PC2, ainda que útil, não captura totalmente variações sutis entre elas - exigindo componentes adicionais (ex.: PC3) para melhor separação/discriminação. PC3 pode revelar padrões ou agrupamentos que não são visíveis apenas nos dois primeiros componentes. Isso ajuda a entender que a variabilidade nos dados não é sempre linear ou facilmente capturada pelos dois primeiros componentes principais.

4 Conclusão

A Análise de Componentes Principais é uma ferramenta estatística fundamentada na álgebra linear, especialmente na decomposição espectral de matrizes de covariância ou correlação. Por meio da diagonalização dessas matrizes, a PCA identifica direções ortogonais os componentes principais que explicam a máxima variância possível dos dados. Essa propriedade a torna aplicável na redução de dimensionalidade, na visualização de estruturas em dados multivariados e na resolução de de redundâncias causadas por correlações entre variáveis. Assim, ao combinar a matemática estatística e a aplicabilidade prática, a PCA se estabelece como um método central na análise multivariada e exploratória de dados.

Referências

1 IBM. *What is principal component analysis (PCA)?* 2023. Https://www.ibm.com/think/topics/principal-component-analysis. Acesso em: 29 abr. 2025.

2 BENNETT, S. R. *Chapter 13: Principal Components Analysis*. 2021. <https://shainarace.github.io/LinearAlgebra/pca.html>. Acesso em: 28 abr. 2025.