UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus Presidente Prudente

LARISSA VITÓRIA RIBEIRO DE ANDRADE

ESTUDO SOBRE PCA

PRESIDENTE PRUDENTE
2025

LARISSA VITÓRIA RIBEIRO DE ANDRADE

Estudo sobre PCA

Tarefa desenvolvido na disciplina de Álgebra Linear para Ciências de Dados no Programa de Pós-Graduação em Matemática Aplicada e Computacional da Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP-FCT.
Professor: Cássio Machiaveli Oishi

PRESIDENTE PRUDENTE
2025

Sumário

1 Decomposição SVD

A decomposição em valores singulares (SVD) de uma matriz

$$A \in \mathbb{C}^{m\times n}$$

é a fatoração

$$A = U\Sigma V^*$$

onde $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p) \in \mathbb{R}^{m\times n}$, com $p = \min\{m, n\}$, e $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$. Além disso, as matrizes quadradas

$$U = [u_1, u_2, \ldots, u_m] \in \mathbb{C}^{m\times m}$$

$$V = [v_1, v_2, \ldots, v_n] \in \mathbb{C}^{n\times n}$$

são unitárias.

Observação: As entradas da diagonal de Σ são chamadas de *valores singulares* de A, as colunas de U são chamadas de *vetores singulares à esquerda*, e as colunas de V são chamadas de *vetores singulares à direita* de A.

Definição (Forma reduzida): Seja $A \in \mathbb{C}^{m\times n}$ com posto $r \leq p = \min\{m, n\}$. A *forma reduzida da SVD* de A é a fatoração

$$A = U\Sigma V^*$$

com $\Sigma_r = \text{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}$, onde $\sigma_1 \geq \cdots \geq \sigma_r > 0$, $U = [u_1, \ldots, u_r] \in \mathbb{C}^{m\times r}$, $V = [v_1, \ldots, v_r] \in \mathbb{C}^{n\times r}$.

Neste caso, as colunas de U e V são ortonormais.

1.1 Propriedades:

a) Toda matriz $A \in \mathbb{C}^{m\times n}$ possui uma decomposição SVD na forma

$$A = U\Sigma V^*$$

b) Se $A \in \mathbb{R}^{m\times n}$, então U e V são matrizes reais.

c) Se $A \in \mathbb{C}^{m\times n}$ possui uma decomposição SVD, então para todo $j = 1, 2, \ldots, p = \min\{m, n\}$ temos:

$$Av_j = \sigma_j u_j, \quad A^*u_j = \sigma_j v_j, \quad u_j^* A v_j = \sigma_j$$

d) Se $U\Sigma V^*$ é uma decomposição de A, então $V\Sigma^T U^*$ é uma decomposição SVD de A^*.

e) Se $A \in \mathbb{C}^{m\times n}$ tem r valores singulares não nulos, então:

$$\begin{cases} \text{posto}(A) = r \\ A = \displaystyle\sum_{j=1}^{r} \sigma_j u_j v_j^* \\ \text{im}(A) = \text{gerado}\{u_1, \ldots, u_r\} \\ \text{espaço nulo}(A) = \text{gerado}\{v_{n+1}, \ldots, v_n\} \end{cases}$$

f) Para a forma reduzida da SVD, se $A \in \mathbb{R}^{m\times n}$, então $\hat{U}$ e $\hat{V}$ podem ser reais.

g) Se posto$(A) = r$, então A tem r valores singulares não nulos.

h)

$$\sigma_i(A) = \sqrt{\lambda_i(A^*A)} = \sqrt{\lambda_i(AA^*)}$$

i) Se $U\Sigma V^*$ é uma decomposição SVD de A, então as colunas de V são autovetores de A^*A, enquanto que as colunas de U são autovetores de AA^*.

j) Se $A \in \mathbb{C}^{n\times n}$ é Hermitiana com autovalores $\lambda_1, \lambda_2, \ldots, \lambda_n$, então seus valores singulares são $\sigma_1 = |\lambda_1|, \sigma_2 = |\lambda_2|, \ldots, \sigma_n = |\lambda_n|$.

Observação: Se $A \in \mathbb{R}^{n\times n}$, então as matrizes da decomposição SVD U, V e Σ também são matrizes reais de ordem $n \times n$, com

$$\Sigma = \begin{bmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \Sigma_n & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix}$$

Observação: Seja $r = n \leq m$. Então a matriz

$$C = A^TA \quad \text{é simétrica e positiva definida (S.P.D.)}$$

Denotamos os autovalores de C em ordem decrescente como λ_i, $i = 1, \ldots, n$. Note que:

$$C = A^TA = V\Sigma^TU^TU\Sigma V^T = V\Sigma^T\Sigma V^T$$

já que a matriz U é ortogonal.

Dessa forma, a matriz

$$\Sigma^T\Sigma \quad \text{é diagonal com elementos} \quad \sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2,$$

e V^T é uma matriz que garante uma "transformação semelhante" com C. Além disso,

$$\sigma_i = \sqrt{\lambda_i} \quad \text{e} \quad \|A\|_2 = \sigma_1$$

1.2 Teorema (Melhor aproximação de posto inferior de uma matriz)

A melhor aproximação de posto r denotada por A_r de uma matriz A, que admite a decomposição

$$A = U\Sigma V^T \quad \text{e a relação}$$

$$\|A - A_r\| \quad \text{é a matriz}$$

$$A_r = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

onde u_i e v_i são as colunas de U e V, respectivamente.

Observação: Esse teorema é aplicado para a construção de uma decomposição SVD "truncada", que faz a "redução dimensional" de um determinado problema.

2 Análise de Componentes Principais (PCA)

A Análise de Componentes Principais (PCA) é uma das aplicações centrais da SVD, fornecendo uma interpretação estatística do sistema de coordenadas hierárquico orientado por dados, usado para representar dados correlacionados de alta dimensão. Esse sistema de coordenadas envolve as matrizes de correlação. De forma importante, o PCA pré-processa os dados subtraindo a média e definindo a variância como unidade antes de realizar a SVD. A geometria do sistema de coordenadas resultante é determinada pelos componentes principais (PCs), que são não correlacionados (ortogonais) entre si, mas têm correlação máxima com as medições. Essa teoria foi desenvolvida em 1901 por Pearson [550] e, independentemente, por Hotelling na década de 1930 [338, 339]. Jolliffe [351] fornece um bom texto de referência.

Frequentemente, em estatística, um número de medições é coletado em um único experimento, e essas medições são tipicamente organizadas em um vetor linha. As medições podem ser características de um observável, como características demográficas de um indivíduo humano específico. Um número de experimentos é conduzido, e cada vetor de medições é disposto como uma linha em uma grande matriz $\mathbf{X}$, lembrando a estrutura de como os dados são registrados em uma planilha. No exemplo da demografia, a coleta dos experimentos pode ser feita por meio de pesquisas. Note que essa convenção para $\mathbf{X}$, consistindo de linhas de características, é diferente da convenção usada ao longo do restante deste capítulo, em que "instantâneos" de características individuais são dispostos como colunas. Contudo, optamos por ser consistentes

com a literatura de PCA nesta seção. Assim, a matriz ainda terá tamanho $n \times m$, embora possa ter mais linhas do que colunas, ou vice-versa.

2.1 Cálculo

Agora calculamos a linha média $\bar{x}$ (isto é, a média de todas as linhas) e a subtraímos de $\mathbf{X}$. A média $\bar{x}$ é dada por

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}, \tag{1.36}$$

e a matriz média é

$$\tilde{X} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \bar{x}. \tag{1.37}$$

Subtraindo $\tilde{X}$ de X resulta na matriz com a média subtraída B:

$$B = X - \tilde{X}. \tag{1.38}$$

A matriz de covariância de B é dada por

$$C = \frac{1}{n-1} B^* B. \tag{1.39}$$

Note que a covariância é normalizada por $n - 1$ em vez de n, mesmo havendo n pontos amostrais. Isso é conhecido como correção de Bessel, que compensa o fato de a variância amostral ser enviesada, pois não captura a variância da média amostral $\bar{x}$ em relação à verdadeira média. A matriz de covariância C é simétrica e definida seminegativa, tendo autovalores reais não-negativos. Cada entrada C_{ij} quantifica a correlação das características i e j em todos os experimentos.

Os componentes principais são os autovetores de C, e eles definem uma mudança de coordenadas na qual a matriz de covariância é diagonal:

$$CV = VD \implies C = VDV^* \implies D = V^*CV. \tag{1.40}$$

As colunas da matriz de autovetores V são os componentes principais, e os elementos da matriz diagonal D são as variâncias dos dados ao longo dessas direções. Essa transformação é garantida, pois C é Hermitiana e as colunas de V são ortonormais.

Nessas coordenadas dos componentes principais, todas as características são em grande parte não correlacionadas entre si.

A matriz de componentes principais V é também a matriz de vetores singulares à direita de B. Substituindo $B = U\Sigma V^*$ em (1.39) e comparando com (1.40) temos

$$C = \frac{1}{n-1} B^* B = \frac{1}{n-1} V \Sigma^2 V^* \quad \Longrightarrow \quad D = \frac{1}{n-1} \Sigma^2. \tag{1.41}$$

A variância dos dados nessas coordenadas, dada pelos elementos diagonais λ_k de D, está relacionada aos valores singulares como

$$\lambda_k = \frac{\sigma_k^2}{n-1}. \tag{1.42}$$

Assim, a SVD fornece uma abordagem numericamente robusta para calcular os componentes principais. Uma aproximação $\tilde{B}$ obtida mantendo apenas os primeiros r componentes principais terá uma variância faltante relacionada à norma de Frobenius ao quadrado mencionada em (1.7).

3 Resultados

3.1 Exercício

Resolver os exemplos *"Noisy Gaussian Data"* e *"Ovarian Cancer Data"*. Depois disso, use o dataset disponível em

`https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset`

e reproduza a análise para esse novo conjunto de dados.

Faça uma análise como na Seção D (*Processing the Data*) do artigo em

`https://ieeexplore.ieee.org/document/9399603`, explicando seus resultados. Você deve postar seu notebook, contendo o código, os resultados e as explicações.

3.2 Resolução

Inicialmente será apresentado os resultados a partir dos exemplos “Noisy Gaussian Data” e “Ovarian Cancer Data”, com isso, vamos observar que o PCA consegue reduzir a dimensionalidade dos dados mantendo as direções de maior variância.

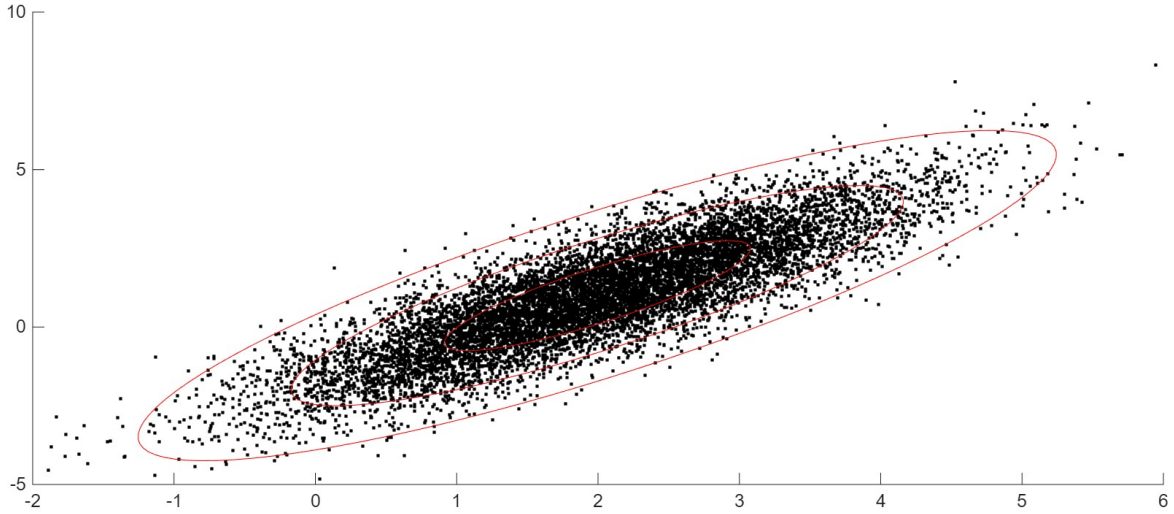

Figura 1

No caso do Noisy Gaussian Data, os primeiros componentes principais explicam quase toda a variabilidade, mostrando que mesmo em dados ruidosos a técnica consegue identificar a estrutura dominante e eliminar redundâncias. Isso confirma que podemos utilizar o PCA como método de filtragem e também para a redução de ruído.

- Esse conjunto de dados é formado por pontos que seguem uma distribuição gaussiana, mas com ruído adicionado.
- O PCA encontra a direção principal onde os dados mais variam (primeiro componente principal).
- Os primeiros componentes explicam quase toda a variabilidade dos dados, enquanto os outros capturam apenas o ruído.
- Com isso, o PCA atua como um filtro: mantém a estrutura dominante dos dados e elimina informações irrelevantes causadas pelo ruído, o que facilita a identificação dos padrões.

Agora, fazendo á análise para o Ovarian Cancer Data, os resultados mostram que os componentes principais separam bem os grupos, evidenciando padrões relevantes para classificação. Nesse caso, o PCA está atuando como uma etapa de pré-processamento fundamental para reduzir a dimensão do espaço e facilitar algoritmos de aprendizado supervisionado.

- O dataset contém informações biomédicas (perfis de expressão de proteínas de pacientes).
- Quando aplicamos o PCA nesse dataset, os primeiros componentes principais já conseguem fazer uma separação clara entre grupos (por exemplo, nesse caso, pacientes com câncer e pacientes saudáveis).
- Isso ocorre porque as direções de maior variância coincidem com as diferenças entre os grupos.
- Ao projetar os dados em apenas 2 ou 3 componentes principais, já é possível visualizar a separação natural das classes.
- O PCA, nesse caso, funciona como uma etapa de **pré-processamento**, reduzindo a dimensionalidade e facilitando a aplicação de algoritmos de classificação supervisionada.
- A imagem correspondente (gráfico de dispersão dos dois primeiros componentes principais) mostra pontos de diferentes cores/formatos se agrupando em regiões distintas, evidenciando o poder do PCA em revelar padrões relevantes.

Ou seja, no Ovarian Cancer Data, o PCA revela padrões de separação entre grupos, ajudando na classificação.

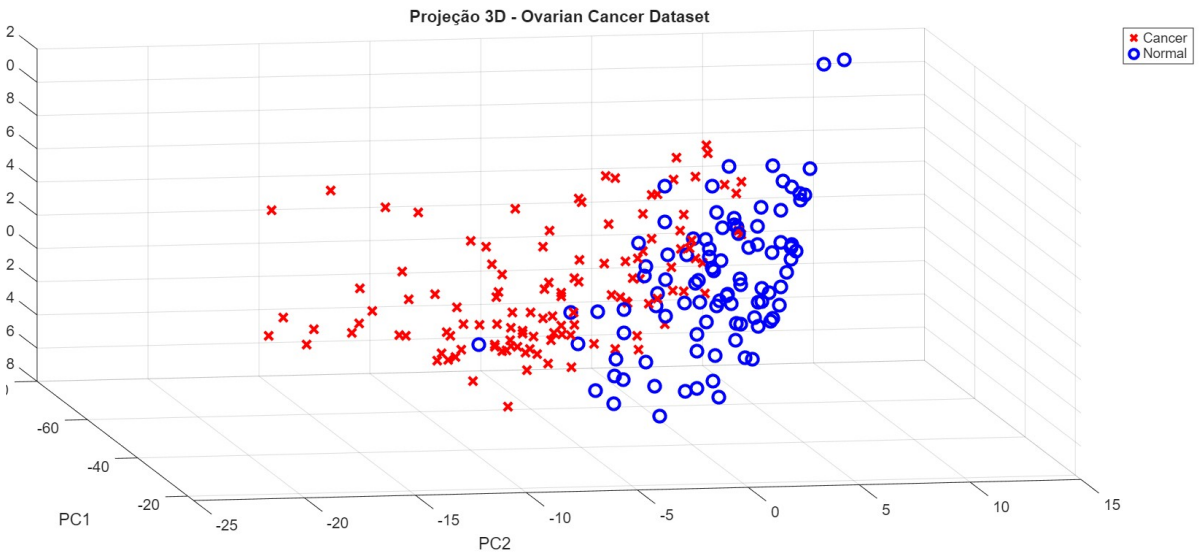

Figura 2

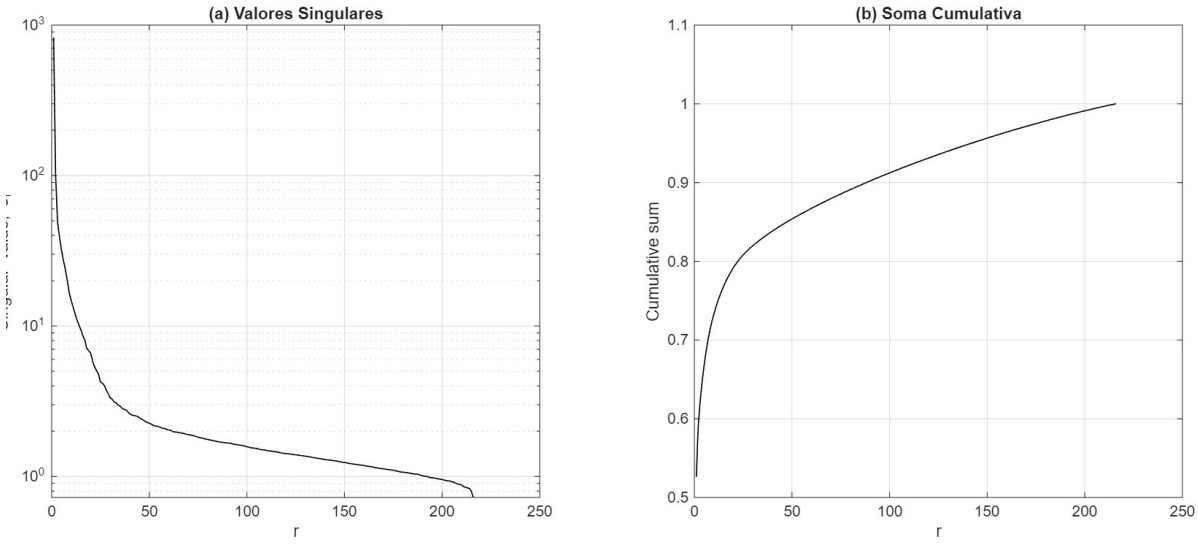
(a) Valores Singulares
10^3
10^2
10^1
10^0
0
50
100
150
200
250
r
(b) Soma Cumulativa
Cumulative sum
1.1
1
0.9
0.8
0.7
0.6
0.5
0
50
100
150
200
250
r

Figura 3

Veja que aplicando o mesmo procedimento ao Breast Cancer Dataset, os gráficos mostram que os primeiros componentes explicam a maior parte da variância dos dados, indicando que é possível representar as amostras em um espaço de baixa dimensão sem grande perda de informação. Isso é consistente com a Seção D (“Processing the Data”) do artigo de referência, em que o PCA é usado para extrair representações compactas dos dados antes de aplicar métodos de classificação.

- O dataset contém atributos de exames de pacientes (como medidas de células, raio, textura, concavidade, entre outros).
- Após aplicar o PCA, os primeiros componentes principais já se concetram na maior parte da variância dos dados.
- Isso mostra que é possível representar as amostras em um espaço de baixa dimensão (2D ou 3D) sem grande perda de informação relevante.
- Com isso, acontece a redução de dimensionalidade que acaba simplificando os dados pois diminui os números de variáveis, facilita a visualização e diminui os efeitos da “maldição da dimensionalidade”, que basicamente é tirar o excesso de informação não tão relevante que atrapalha enxergar os padrões.
- Esse resultado está de acordo com a Seção D (*Processing the Data*) do artigo de referência, onde o PCA é utilizado para extrair representações compactas antes de aplicar métodos de classificação.
- O gráfico resultante (dispersão dos dois primeiros componentes principais) evidencia que as classes (benigno vs. maligno) tendem a se separar em regiões distintas, confirmando a utilidade do PCA como etapa de pré-processamento.

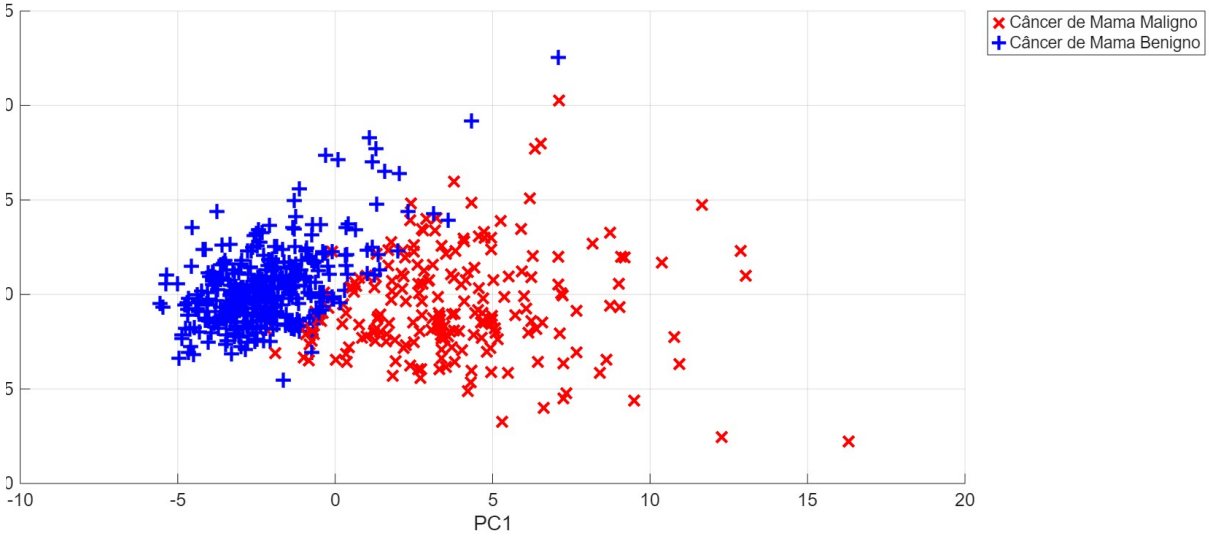

Figura 4

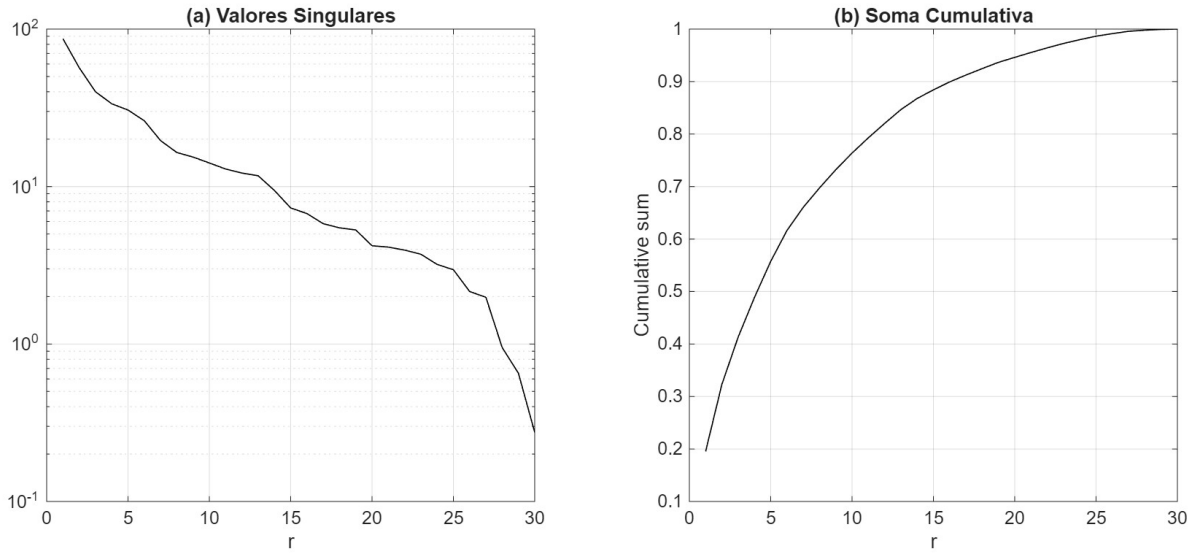

Figura 5

Assim, os resultados obtidos reforçam que o PCA não apenas reduz a dimensionalidade, mas também preserva a estrutura essencial dos dados, sendo uma ferramenta útil em contextos de análise exploratória e aprendizado de máquina.