Análise da Redução de Dimensionalidade com PCA

João Vitor Ferreira Lima Luigi Leone Duarte Yamamoto Leite

Manuela Silva de Andrade

1 Introdução

.

A Análise de Componentes Principais (PCA) é uma técnica essencial na análise de dados, visando a redução de dimensionalidade. Nesta seção, destacaremos a importância do PCA na simplificação de conjuntos de dados complexos, fornecendo uma visão geral dos objetivos e relevância desta abordagem em diversas disciplinas.

Análise de Componentes Principais (PCA) desempenha um papel essencial tanto na estatística multivariada quanto no aprendizado de máquina, sendo crucial para a redução de dimensionalidade e a identificação de padrões significativos em conjuntos de dados complexos. Essa técnica tem aplicações abrangentes, estendendo-se desde a análise de dados financeiros até o processamento de imagens e o reconhecimento de padrões.

No âmago da PCA está a transformação linear dos dados originais, resultando em um novo sistema de coordenadas no qual as variáveis são reorganizadas com base na variabilidade dos dados. Essa reconfiguração possibilita a identificação e o destaque das principais direções ao longo das quais os dados variam, revelando estruturas subjacentes e simplificando a interpretação dos conjuntos de dados.

Este trabalho se propõe a aprofundar nos princípios teóricos da Análise de Componentes Principais, abordando sua aplicabilidade em diversos contextos, destacando suas vantagens, reconhecendo suas limitações e explorando as implicações práticas de sua implementação. Além disso, serão apresentados exemplos concretos para ilustrar como a PCA tem desempenhado um papel fundamental em avanços notáveis em diversas disciplinas, evidenciando sua versatilidade e importância na análise de conjuntos de dados complexos e dimensionalmente extensos.

2 Processo de extração das componentes principais

As etapas para extrair as componentes principais envolvem uma série de procedimentos, incluindo:

- Coleta de M amostras de vetores de dimensão n para formar o conjunto de dados.
- Cálculo do vetor médio, obtendo a média aritmética das amostras coletadas.
- Centralização dos dados, onde cada elemento do conjunto é subtraído pela média calculada na etapa anterior.
- Cálculo da matriz de covariância, que expressa a variabilidade conjunta dos dados através do produto das diferenças obtidas na centralização.
- Identificação dos autovalores e autovetores da matriz de covariância, os quais indicam a direção e a magnitude das componentes principais.
- O autovetor associado ao maior autovalor corresponde à primeira componente principal, representando o relacionamento mais significativo entre as dimensões dos dados. Este processo se repete para as demais componentes principais, seguindo a ordem decrescente dos autovalores.

Dessa forma, as componentes principais são extraídas de maneira sistemática, proporcionando uma representação eficiente das relações entre as diferentes dimensões do conjunto de dados.

3 Descrição do Conjunto de Dados: Exemplo de uso do PCA

.

Antes de aplicar o PCA, é crucial compreender o conjunto de dados em questão. Abordaremos a natureza e o contexto do conjunto de dados escolhido, contextualizando sua relevância na aplicação da análise de componentes principais.

O Conjunto de Dados Iris é uma base de dados clássica amplamente utilizada em estudos de aprendizado de máquina e análise estatística. Ele contém informações sobre três espécies de flores de íris: setosa, versicolor e virginica. As medidas fornecidas incluem o comprimento e a largura das sépalas e pétalas, características botânicas que são frequentemente utilizadas para classificar espécies de flores.

Em termos mais específicos, para cada amostra de íris no conjunto de dados, são registrados quatro atributos: comprimento da sépala, largura da sépala, comprimento da pétala e largura da pétala. Essas medidas são expressas em centímetros.

O objetivo típico ao utilizar esse conjunto de dados é explorar métodos de classificação ou técnicas de redução de dimensionalidade, como a Análise de Componentes Principais (PCA). A distinção clara entre as espécies com base nas características fornecidas torna o conjunto de dados Iris uma excelente escolha para demonstrar técnicas de machine learning e análise exploratória de dados.

3.1 Dimensão dos Dados:

Número de Variáveis: 4

Número de Observações: 150

O conjunto de dados possui quatro variáveis, representando o comprimento e a largura das sépalas e pétalas de 150 flores de íris. Cada observação está associada a uma espécie de íris.

3.2 Visualização das Primeiras Linhas do Conjunto de Dados:

As primeiras linhas mostram as medidas das sépalas e pétalas, bem como a espécie de íris associada (0, 1 ou 2).

O conjunto de dados compreende 50 amostras de cada uma das três variedades de Iris (Iris Setosa, Iris virginica e Iris versicolor). Foram registradas quatro características para cada amostra, incluindo as medidas de comprimento e largura das sépalas e pétalas, expressas em centímetros. Todos os dados foram introduzidos na tabela do software estatístico "PAST".

	Color	Symbol	Name	sepal_length	sepal_width	petal_length	petal_width
Type	Darkviolet	Dot	-	-	-	-	-
Name	Darkviolet	Dot	species	sepal_length	sepal_width	petal_length	petal_width
Iris-setosa	Darkviolet	Dot	Iris-setosa	48,18104687	32,80019083	12,2445546	0,642562397
Iris-setosa	Darkviolet	Dot	Iris-setosa	46,18104687	0,800190826	12,2445546	0,642562397
Iris-setosa	Darkviolet	Dot	Iris-setosa	44,18104687	29,80019083	11,2445546	0,642562397
Iris-setosa	Darkviolet	Dot	Iris-setosa	43,18104687	28,80019083	13,2445546	0,642562397
Iris-setosa	Darkviolet	Dot	Iris-setosa	2,181046871	33,80019083	12,2445546	0,642562397
Iris-setosa	Darkviolet	Dot	Iris-setosa	51,18104687	36,80019083	15,2445546	2,642562397
Iris-setosa	Darkviolet	Dot	Iris-setosa	43,18104687	31,80019083	12,2445546	1,642562397
Iris-setosa	Darkviolet	Dot	Iris-setosa	2,181046871	31,80019083	13,2445546	0,642562397
Iris-setosa	Darkviolet	Dot	Iris-setosa	41,18104687	26,80019083	12,2445546	0,642562397
Iris-setosa	Darkviolet	Dot	Iris-setosa	46,18104687	28,80019083	13,2445546	-0,357437603
Iris-setosa	Darkviolet	Dot	Iris-setosa	51,18104687	34,80019083	13,2445546	0,642562397
Iris-setosa	Darkviolet	Dot	Iris-setosa	45,18104687	31,80019083	14,2445546	0,642562397
Iris-setosa	Darkviolet	Dot	Iris-setosa	45,18104687	0,800190826	12,2445546	-0,357437603
Iris-setosa	Darkviolet	Dot	Iris-setosa	40,18104687	0,800190826	9,244554604	-0,357437603
Iris-setosa	Darkviolet	Dot	Iris-setosa	55,18104687	1,800190826	10,2445546	0,642562397
Iris-setosa	Darkviolet	Dot	Iris-setosa	54,18104687	41,80019083	13,2445546	2,642562397
Iris-setosa	Darkviolet	Dot	Iris-setosa	51,18104687	36,80019083	11,2445546	2,642562397
Iris-setosa	Darkviolet	Dot	Iris-setosa	48,18104687	32,80019083	12,2445546	1,642562397
Iris-setosa	Darkviolet	Dot	Iris-setosa	54,18104687	35,80019083	15,2445546	1,642562397
Iris-setosa	Darkviolet	Dot	Iris-setosa	48,18104687	35,80019083	13,2445546	1,642562397
Iris-setosa	Darkviolet	Dot	Iris-setosa	51,18104687	31,80019083	15,2445546	0,642562397
Iris-setosa	Darkviolet	Dot	Iris-setosa	48,18104687	34,80019083	13,2445546	2,642562397
Iris-setosa	Darkviolet	Dot	Iris-setosa	43,18104687	33,80019083	-0,755445396	0,642562397

Figure 1: Conjunto de dados na área tabular do software "PAST".

3.3 Matriz de Covariância:

.

Os dados deste estudo estão normalizados e suas escalas são equilibradas. Portanto, não há impedimento para a utilização da matriz de covariância no cálculo das Componentes Principais (PCAs).

PC	Eigenvalue	% variance
1	2,08886	52,356
2	0,970712	24,33
3	0,606774	15,208
4	0,323362	8,1049

Figure 2: Ferramenta de PCA

A escolha pela aplicação do método de Análise de Componentes Principais (PCA) foi feita a partir da ferramenta de ordenação multivariada. Duas opções para o cálculo desse método são disponibilizadas pelo software: utilizando uma matriz de covariância ou de correlação.

Em conjuntos de dados com disparidades significativas nas escalas e unidades de medida das variáveis (ou seja, dados não normalizados), ocorre uma disparidade acentuada entre as variâncias, especificamente no contexto da matriz de covariância. Isso acontece porque as componentes são influenciadas pelas variáveis de maior variância, resultando em componentes principais predominantemente controladas por uma única variável, o que reduz sua utilidade.

3.4 Autovalores e Autovetores:

Os autovalores e autovetores são fundamentais para o PCA. Os autovalores indicam a quantidade de variância explicada por cada componente principal, enquanto os autovetores apontam as direções dessas componentes.

```
Autovalores:
[4.22824171 0.24267075 0.0782095  0.02383509]

Autovetores:
[[ 0.36138659 -0.65658877 -0.58202985  0.31548719]
 [-0.08452251 -0.73016143  0.59791083 -0.3197231 ]
 [ 0.85667061  0.17337266  0.07623608 -0.47983899]
 [ 0.3582892   0.07548102  0.54583143  0.75365743]]
```

Figure 3: Autovalores e Autovetores

3.5 Maiores Autovalores e Autovetores:

Foram focados os dois maiores autovalores e seus autovetores correspondentes, pois eles capturam a maior parte da variância nos dados.

```
Autovalor 1: 4.228241706034869
Autovetor 1: [0.36138659 -0.08452251  0.85667061  0.3582892 ]

Autovalor 2: 0.24267074792863338
Autovetor 2: [-0.65658877 -0.73016143  0.17337266  0.07548102]
```

Figure 4: Maiores Autovalores e Autovetores(Top 2)

Dessa forma, a opção foi pela utilização da matriz de variância-covariância para conduzir a análise. A análise da figura revela o cálculo de 4 componentes principais, seus autovalores e a proporção de sua contribuição para os dados originais.
Diante desse contexto, a escolha é utilizar apenas os dois primeiros componentes principais, visto que o terceiro e o quarto componentes representam uma quantidade insignificante de informações no conjunto de dados. Em termos simples, a soma dos PC1 e PC2 equivale a cerca de 76 (porcento) da massa de dados.

3.6 Plotagem (Imagem):

,

A plotagem bidimensional usando PCA destaca a distribuição das amostras nas duas primeiras componentes principais, permitindo uma visualização eficaz dos padrões existentes nos dados.

A guia seguinte, denominada Scatter plot, exibe um gráfico de dispersão. Inicialmente, utilizando apenas o PCA1 nos dois eixos, representando os valores dispersos em uma reta, podemos notar uma tendência evidente pois a maioria das Irís-virginica que estão representadas pelos pontos vermelho estão posicionada em valores positivos da reta, ao passo que as Irís-setosa que estão representadas pelos pontos roxo ocupam principalmente os valores negativos.

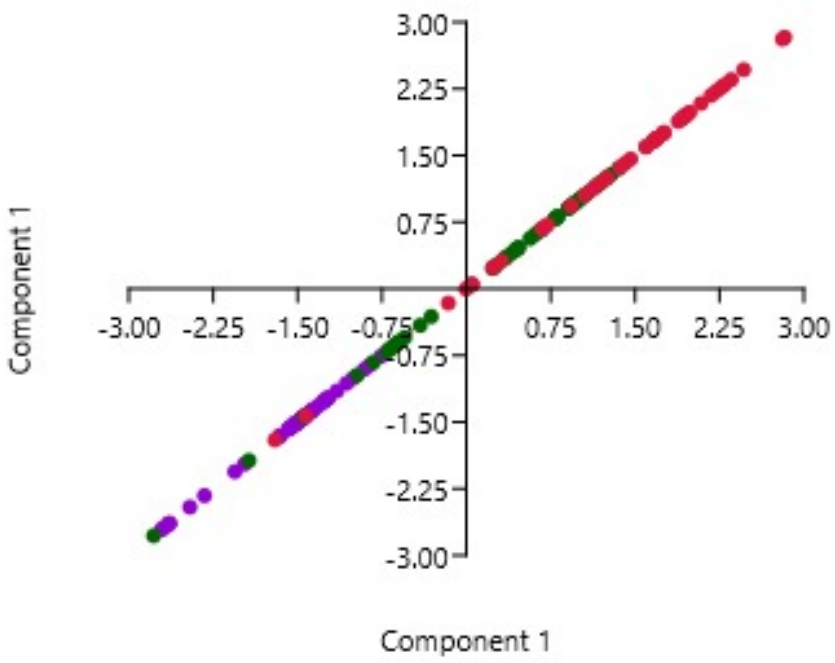

Figure 5: Scatter Plot

A fim de aprimorar a compreensão dos resultados obtidos, optamos por designar o eixo Y como PCA2. Isso nos permite entender como a dispersão desses indivíduos ocorre no gráfico em uma direção não ortogonal ao PCA1.

Inicialmente, é possível observar que as Íris setosas estão predominantemente posicionadas nos valores negativos do PC1, diferentemente das Irís Virgínicas e Irís Versicolor. Além disso, nota-se que a Iris setosa apresenta uma largura de sépala significativamente maior quando averiguadas as variáveis originais.

Essa disparidade ocorre devido à relação entre as medidas das sépalas e das pétalas, que evidenciam diferenças ao analisarmos as diversas espécies de Íris. Em geral, as Irís virgínicas e Irís versicolor tendem a ter uma média de largura de sépala menor em relação à Iris setosa.

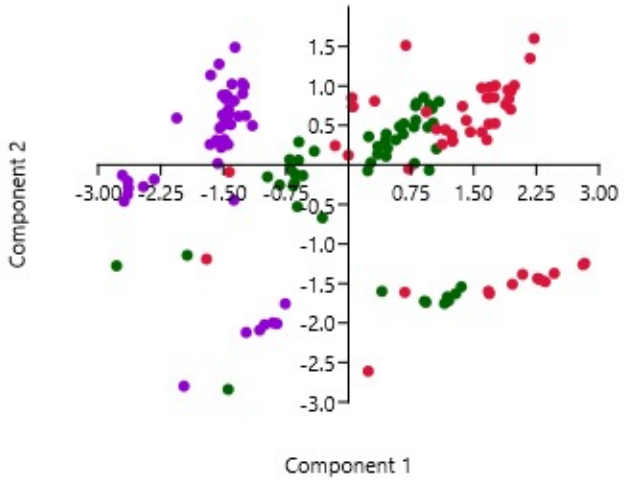

Figure 6: Gráfico de dispersão de dois eixos

4 Análise dos Resultados:

.

Iris setosa está principalmente posicionada em valores negativos do PCA1, enquanto Iris Virginica e Iris Versicolor ocupam valores positivos. Ao designar o eixo Y como PCA2, observamos a dispersão em uma direção não ortogonal ao PCA1, revelando diferenças nas medidas de sépalas e pétalas.

A Iris setosa destaca-se por características distintas, como largura de sépala significativamente maior.

As Iris Virginica e Versicolor exibem diferenças mais sutis, mas ainda discerníveis, nas medidas de sépalas e pétalas.

A escolha de focar nos dois primeiros componentes principais permite uma redução significativa na dimensionalidade dos dados, mantendo a informação essencial.

5 Conclusão:

.

É possível identificar que a aplicação da Análise de Componentes Principais (PCA) ao Conjunto de Dados Iris revelou uma estratégia eficaz na redução de dimensionalidade e na identificação de padrões subjacentes nas características florais. Através da transformação dos dados originais em um novo conjunto de componentes principais, pudemos condensar a informação mantendo as principais variações nos dados.

Os resultados visuais apresentados na análise bidimensional mostraram uma clara separação entre as diferentes espécies de íris, indicando que as características escolhidas para o estudo são distintivas o suficiente para permitir uma identificação eficaz. A eficácia do PCA em simplificar a representação dos dados ressalta sua utilidade na análise exploratória e na visualização de padrões complexos.

E possível explorar variações no número de componentes principais retidos para entender como isso afeta a representação dos dados e a capacidade de distinguir entre as espécies. No contexto de aplicações práticas contínuas, a compreensão aprimorada das características florais através do PCA pode ser utilizada em projetos de classificação automática de espécies de íris.

A normalização dos dados permitiu a utilização da matriz de covariância no cálculo das Componentes Principais (PCAs) com escalas equilibradas, utilizando a ferramenta PAST. A escolha do método PCA baseou-se na ordenação multivariada, optando pela matriz de covariância em situações de disparidades significativas nas escalas das variáveis para evitar distorções nos resultados.

Os autovalores e autovetores foram fundamentais para compreender a quantidade de variância explicada por cada componente principal e as direções dessas componentes. Focamos nos dois maiores autovalores e seus autovetores correspondentes, representando aproximadamente 76 (porcento) da

variância total. A decisão de utilizar apenas os dois primeiros componentes principais foi tomada devido à insignificância do terceiro e quarto componentes.

A plotagem bidimensional usando PCA proporcionou uma visualização eficaz dos padrões nos dados. A análise do scatter plot revelou tendências distintas na distribuição das espécies, destacando a disparidade nas medidas de sépalas e pétalas entre Iris setosa e as outras variedades, com Iris setosa apresentando uma largura de sépala significativamente maior.

A utilização do PCA facilitou a compreensão das relações complexas entre as variáveis do conjunto de dados da íris. A escolha dos componentes principais e a visualização bidimensional permitiram a identificação de padrões distintos, contribuindo para uma compreensão mais profunda da variabilidade nas medidas da íris e evidenciando a utilidade do PCA na simplificação e interpretação de conjuntos de dados multidimensionais.

Em resumo, a aplicação do PCA à base de dados Iris forneceu insights e representou um primeiro passo essencial na exploração da capacidade dessa técnica em revelar padrões e simplificar a representação de dados complexos.

6 Referências

.

COSTA, Nilson Santos. PCA no software PAST. [Vídeo online]. YouTube. Publicado em 24 de maio de 2023. Disponível em: https://www.youtube.com/watch?v=Hj5S5KE80S4. Acesso em: 25/11/2023.

CONCI, Alberto. Análise de Componentes Principais (PCA) - Apostila. [Documento online]. Universidade Federal Fluminense. Disponível em: http://www2.ic.uff.br/ aconci/PCA-ACP.pdf. Acesso em: 25/11/2023.

ARSHID, A. Iris Flower Dataset. [Conjunto de dados online]. Kaggle. Disponível em: https://www.kaggle.com/datasets/arshid/iris-flower-dataset. Acesso em: 25/11/2023.

Link do GitHUB: https://github.com/manuandradex/Trabalho$_PCA$