Implementação da Análise de Componentes Principal (PCA)

Murillo Freitas Bouzon

***Resumo*—O aumento da quantidade de informações trouxe novas possibilidades para a área de *machine learning*, podendo utilizar grandes bases de treinamento para o uso desses algoritmos. No entanto, a quantidade de características observadas também aumentou consequentemente, gerando casos onde a quantidade de características é maior do que o número de observações de uma amostra. Para resolver esse problema, pode-se utilizar o PCA, um método estatístico para redução de dimensionalidade ao encontrar um novo sistema de coordenadas para representar os dados originais de uma amostra. Sendo assim, neste trabalho foi feita a implementação do método PCA para visualizar os dados por novos eixos representados pelas componentes principais. Os resultados mostraram que foi possívei representar quase toda a variância dos dados com a primeira componente principal, facilitando a visualização ao rotacionar os dados para o novo sistema de coordenadas encontradas pelas componentes principais.**

***Index Terms*—Análise de Componentes Principais, Aprendizado Não-Supervisionado, Machine Learning, Inteligência Artificial**

I. INTRODUÇÃO

Com o aumento da quantidade de dados armazenados de fácil acesso, os algoritmos de aprendizado mostraram-se cada vez mais precisos em suas respostas, fazendo com que surgisse uma tendência em utilizá-los em conjunto com as grandes bases de dados disponíveis atualmente.

No entanto, para alguns casos, o tamanho da amostra acaba sendo menor do que a quantidade de características observadas de uma determinada base, denegrindo a qualidade do resultado final dos algoritmos pois acabam utilizando uma amostragem pequena para o treinamento em relação à quantidade de características presentes na base de dados, não sendo o suficiente para gerar resultados com uma boa precisão.

Uma das soluções para esse problema é utilizando a técnica Análise de Componentes Principais, também conhecida como PCA (Principal Components Analysis). Com o uso deste método, é possível reduzir a dimensionalidade dos seus dados baseando-se no cálculo da covariância entre eles para descartar características desnecessárias para representar os dados e então utilizar os novos dados reduzidos para realizar o treinamento do algoritmo de aprendizado, gerando melhores resultados graças a menor quantidade de características comparada com o tamanho da amostra.

O PCA é utilizado em diversos na literatura. Um exemplo da aplicação do PCA é o trabalho de [5], que propôs um método para construir e analisar diferentes características não vistas em imagens frontais 2D da face humana, tais como expressões faciais e traços de gênero. O PCA foi aplicado em conjunto com o método MLDA (Maximum Uncertainty Linear Discriminant Analysis) para reduzir a dimensionalidade dos dados e extrair informações discriminantes das fotos.

O trabalho de [1] também aplicou o PCA na área de reconhecimento facial, dividindo uma imagem em n sub-imagens e aplicando o PCA localmente para cada n região, extraindo características locais da face. No trabalho de [2] foi utilizado o PCA para compressão de imagens RGB e determinar a rotação de uma cena a partir de duas imagens. [3] foi outro trabalho que utilizou o PCA, em conjunto com entropia não-extensiva e redes neurais aplicados para classificação de iris.

Dada a importância deste método para a área de aprendizado de máquina, este trabalho tem como foco a implementação do PCA, sendo aplicado em bases de exemplo para encontrar as suas componentes principais e obter um melhor entendimento sobre o algoritmo.

II. CONCEITOS FUNDAMENTAIS

Nesta seção serão apresentadas as teorias relacionadas ao método desenvolvido neste trabalho

A. Covariância

Na estatística, a covariância é uma medida do grau de interdependência entre duas variáveis, sendo positiva em casos onde as duas variáveis possuem uma variação para o mesmo sentido, negativa em casos onde a variação entre elas possuem o sentido contrário e 0 se as duas variáveis forem independentes.

O cálculo da covariância entre duas variáveis x e y, dada uma amostra com N observações de uma população, é demonstrado pela Equação (1).

$$q_{x,y} = \frac{\sum_{i=1}^{N}(V_{i,x} - \bar{V_{i,x}})(V_{i,y} - \bar{V_{i,y}})}{N-1} \tag{1}$$

B. Autovalores e Autovetores

Dado um valor escalar λ multiplicado por um vetor x e uma matriz quadrada A, pode-se dizer que λ é autovalor de A caso exista um vetor $x! = 0$ tal que $Ax =$. O vetor x é chamado de autovetor. Ao aplicar uma matriz de transformação A em um autovetor x, apenas a sua magnitude e o seu sinal mudam, ou seja, a direção de Ax é a mesma de x.

C. Análise de Componentes Principais

A Análise de Componentes Principais, também conhecida como PCA (*Principal Component Analysis*) é uma técnica estatística proposta por [4]. Este método transforma um conjunto de dados aplicando uma transformação ortogonal entre eles,

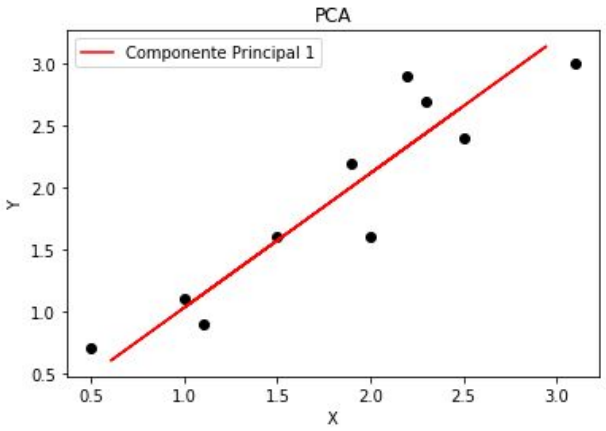


Figura 1: Representação gráfica da primeira componente principal sobre os dados originais.

convertendo-os para um novo conjunto de dados representado por novas variáveis linearmente independentes chamadas de componentes principais.

A primeira componente principal é calculada de forma que represente a maior variância dos dados originais. A segunda componente principal deve ser ortogonal a primeira, a terceira deve ser ortogonal a primeira e a segunda e assim sucessivamente. Para realizar o cálculo do PCA, calcula-se a matriz de covariância entre os dados e então é encontrado os autovalores e autovetores da matriz, onde os autovalores representa a variância que cada componente explica em relação aos dados e os autovetores são as próprias componentes principais. Com isso, é possível reduzir a dimensionalidade de um conjunto de observações, eliminando variáveis que explicam uma parte desprezível da variância entre os dados e facilitando tarefas que utilizam análise de dados e reconhecimento de padrões. A Figura 1 mostra um exemplo da aplicação do PCA sobre um conjunto de dados, apresentando a primeira componente principal que melhor explica a variância entre os dados.

III. METODOLOGIA

A metodologia deste trabalho é feita em 7 etapas, ilustrada pelo diagrama da Figura 2. A primeira etapa consiste na leitura da base de dados, que será descrita nas próximas seções. A seguir, para cada característica, é calculada a média e subtraída de cada observação. Na terceira etapa, calcula-se a matriz de covariância dos dados para então calcular os seus autovalores e autovetores nas etapas 4 e 5.

Nas duas últimas etapas, são selecionadas as componentes principais e então são aplicadas para transformar os dados em um novo conjunto representado pela componentes selecionadas, para assim serem visualizados em novos eixos.

A implementação foi feita na linguagem Python, sendo utilizado as bibliotecas *Numpy* para operação com matrizes e matplotlib para plotagem de gráficos.

IV. BASES DE DADOS

Para validar se as implementações foram realizadas corretamente, foram realizados experimentos em três bases de dados, sendo descritas a seguir:

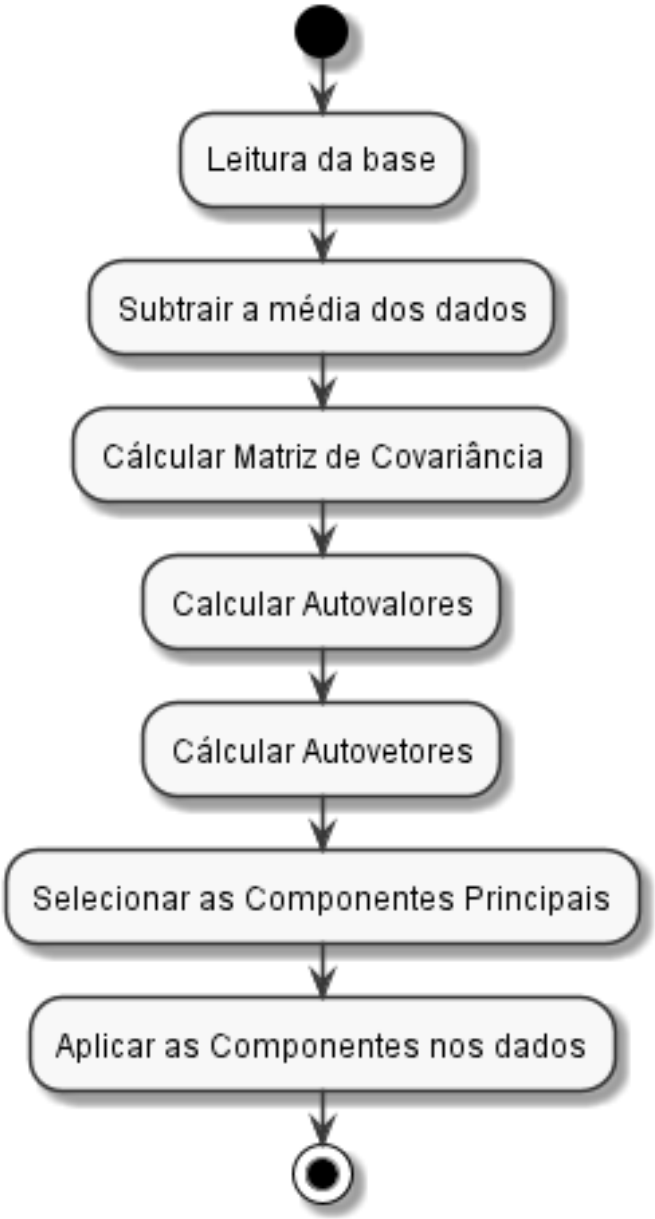


Figura 2: Metodologia do trabalho.

A. *Alps Water*

Esta base possui informações a respeito do ponto de ebulição da água em diferentes pressões atmosféricas, possuindo 17 observações e duas variáveis, sendo elas:

- Temperatura: A temperatura do ponto de ebulição da água medida em graus Fahrenheit (F).
- Pressão: Pressão atmosférica em polegadas de mercúrio (Hg).

A variável adotada como variável objetivo (y) foi a Pressão.

B. *Books X Grades*

Esta base possui informações a respeito do desempenho dos alunos em uma determinada matéria. Possui 3 variáveis e 40 observações e possui as seguinte variáveis:

- Livros: Quantidade de livros lidos pelo aluno.
- Presença: Quantidade de aulas que o aluno assistiu.
- Nota: A nota final do aluno na matéria.

Nesta base foi adotado a Nota como variável objetivo.

C. US Census Dataset

Esta base possui informações sobre o registro populacional dos Estados Unidos por década, possuindo 11 observações e 2 variáveis, sendo elas:

- Ano: O ano em que foi realizada a contagem.
- População: Número de habitantes registrados.

A variável adotada como objetivo foi a população.

V. Experimentos e Resultados

Para avaliar o método implementado, foram feitos experimentos nas três bases apresentadas na Seção IV. O experimento verificou duas coisas: A variância explicada por cada componente e a visualização dos dados representados pelas duas primeiras componentes principais.

Para encontrar a taxa da variância explicada por cada componente, encontrou-se os autovalores da matriz de covariância de cada uma das bases e foi feita a divisão deles pela soma total dos autovalores. A Figura 3 apresenta os resultados deste experimento, onde foi feito um gráfico de barras para cada uma das bases, sendo cada uma das barras a taxa da variância explicada por cada componente principal.

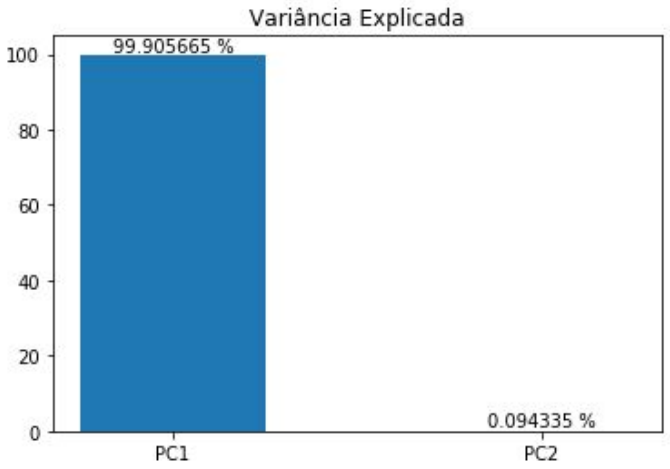
Os gráficos gerados mostram que para as três bases, a primeira componente principal representa a maior taxa de variância entre os dados, sendo quase 100% para as bases *Alpswater* e *US Census* e aproximadamente 95% para a base *Books X Grades*.

Em seguida, calculou-se os autovetores e projetou a primeira componente principal sobre os dados originais. A Figura 4 apresenta os gráficos com os dados originais de cada base e a primeira componente principal passando sobre eles.

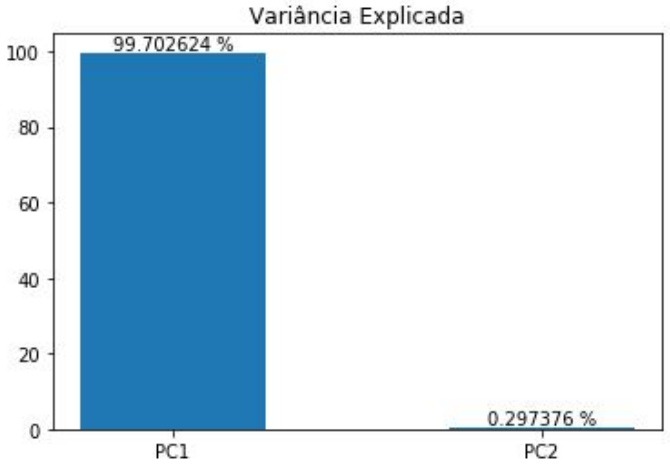
Os gráficos mostram que a primeira componente se aproxima muito dos dados originais. Além disso, é possível projetar os dados na primeira componente para poder visualizá-los.

Por fim, aplicou-se as duas primeiras componentes principais nos dados originais para visualizá-los em um novo sistema de coordenadas, com os eixos originais rotacionados. A Figura 5 apresenta os gráficos que representam o novo conjunto de dados após a rotação dos eixos de acordo com as duas primeiras componentes principais.

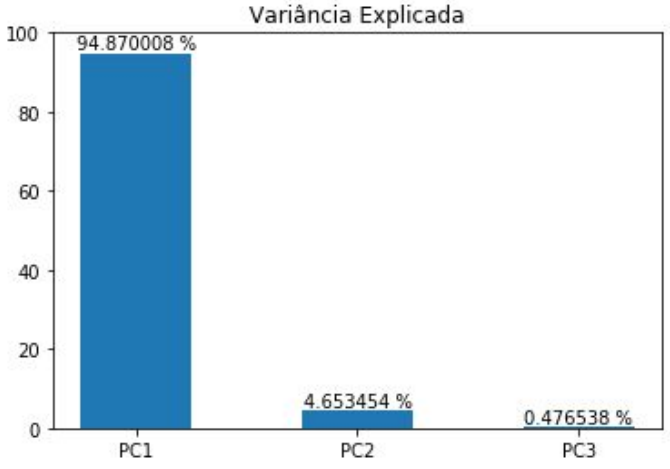
Pode-se dizer que é a visualização dos dados fica mais clara ao rotacionar os eixos, enxergando com mais facilidade em como os dados estão espalhados no espaço.



(a) Variância explicada por cada componente da base Alpswater.



(b) Variância explicada por cada componente da base US Census.



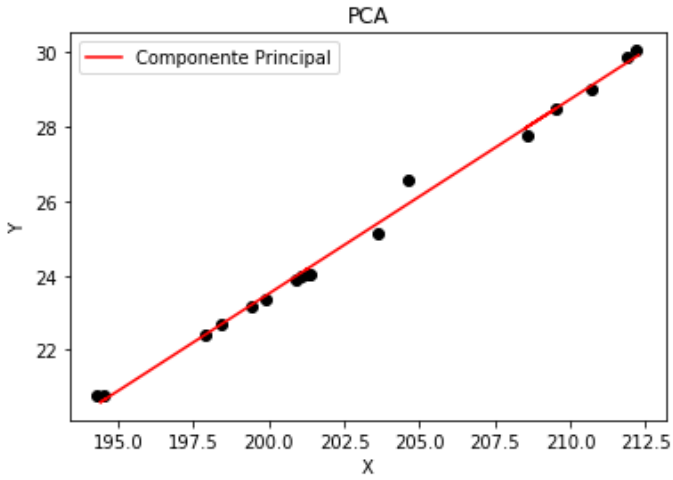
(c) Variância explicada por cada componente da base Books X Grades.

Figura 3: Gráficos com a taxa de variância explicada por cada componente das bases utilizadas no experimento.

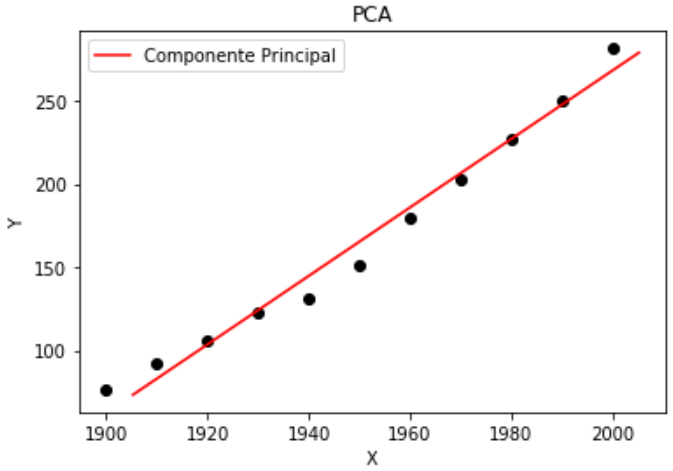
VI. Conclusão

Neste trabalho foi feita a implementação do método PCA aplicado para rotacionar os dados para melhorar a visualização e obter um melhor entendimento sobre as teorias que envolvem esse método e a sua importância na área de *machine learning*.

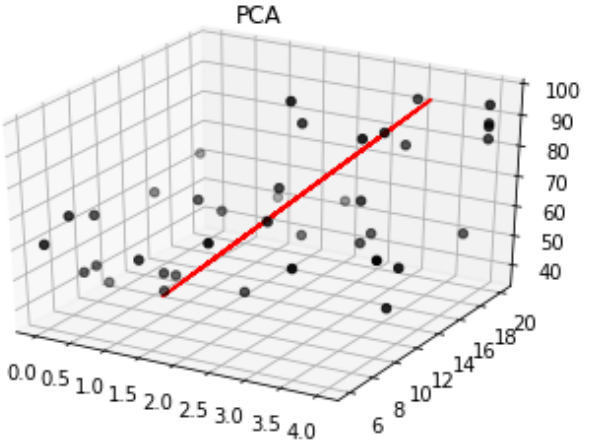
Foram feitos experimentos onde o método foi testado em três bases diferentes. Para cada uma das bases calculou-se os autovalores da matriz de covariância delas e foi verificada a taxa de variância explicada dada por cada uma das componentes principais. Foi observado que a primeira componente obteve a maior taxa de variância representando acima de 90% da variância total entre os dados.



(a) Gráfico com a primeira componente principal sobre a base Alpswater.



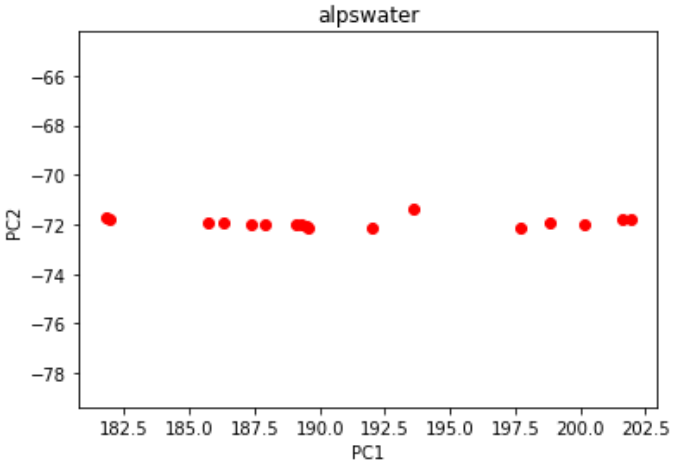
(b) Gráfico com a primeira componente principal sobre a base Census.



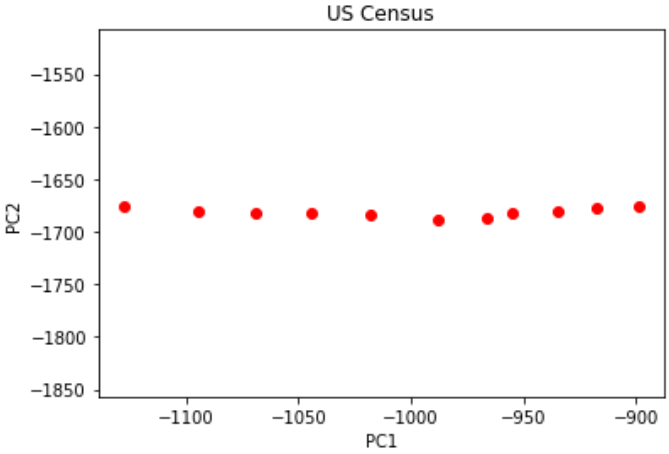
(c) Gráfico com a primeira componente principal sobre a base Books X Grades.

Figura 4: Gráficos com a primeira componente principal passando pelos dados originais de cada base utilizada no experimento.

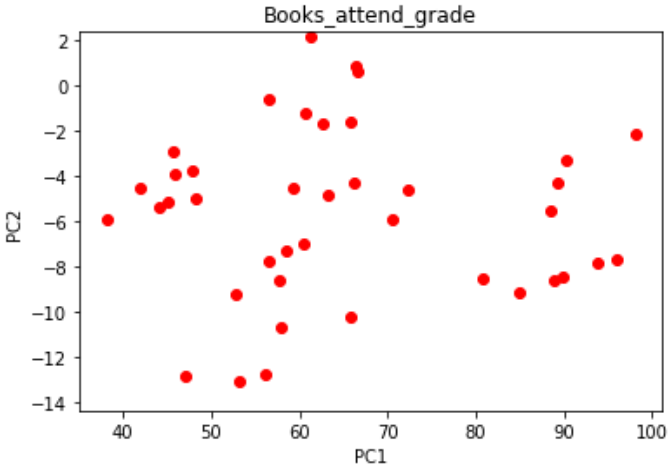
Também foi feita a rotação dos eixos de acordo com as componentes principais encontradas para facilitar a visualização



(a) Gráfico com o novo conjunto de dados após a aplicação das duas primeiras componentes principais da base Alpswater.



(b) Gráfico com o novo conjunto de dados após a aplicação das duas primeiras componentes principais da base Census.



(c) Gráfico com o novo conjunto de dados após a aplicação das duas primeiras componentes principais da Books X Grades.

Figura 5: Gráficos com os dados sendo representados pelas duas primeiras componentes principais para cada uma das bases utilizadas neste trabalho.

dos dados. Com isso, a visualização ficou mais claro, permitindo um melhor entendimento em como os dados encontram-se espalhados no espaço.

A técnica de PCA demonstra ser muito importante para a área de aprendizado, permitindo a redução da dimensionalidade dos dados ao eliminar componentes desnecessárias na representação dos dados, podendo ser utilizado em conjunto com algoritmos supervisionados para melhorar a precisão deles ao reduzir a quantidade de características em relação ao tamanho da amostra.

REFERÊNCIAS

[1] R. Gottumukkal and V. K. Asari. An improved face recognition technique based on modular pca approach. *Pattern Recognition Letters*, 25:429–436, 2004.

[2] M. Mudrová and A. Procházka. Principal component analysis in image processing. 2005.

[3] L. Nasseri, A. A. B. Shirazi, and N. Sadeghigol. Tsallis entropy, pca and neural network in novel algorithm of iris classification. *2011 World Congress on Information and Communication Technologies*, pages 385–390, 2011.

[4] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

[5] C. E. Thomaz, V. do Amaral, G. A. Giraldi, E. C. Kitani, J. R. Sato, and D. F. Gillies. A multi-linear discriminant analysis of 2d frontal face images. *2009 XXII Brazilian Symposium on Computer Graphics and Image Processing*, pages 216–223, 2009.