Chapter

10

Análise dos Componentes Principais (PCA)

Bernard Silva de Oliveira, Manoel Veríssimo dos Santos Neto, Rafael Rodrigues Silva e Wesley Modanez Freitas

Abstract

The main objective of the following study is to explore the Principal Component Analysis (PCA) technique and its practical applications. PCA is used to identify patterns in multivariate data, allowing these data to be expressed in a way that highlights their similarities and differences. This is achieved by transforming the original variables into a new set of variables, the principal components, which are orthogonal and ordered so that the first ones carry most of the variability in the data. The process includes standardizing the data to prevent undue influences from scale variations among attributes, calculating the covariance matrix to understand how variables are associated with each other, determining the eigenvalues and eigenvectors that identify the directions and magnitudes of the new features, and selecting the principal components based on explained variance. Furthermore, the study demonstrates the practical application of PCA in the selection of dimensionality in satellite images with pluviometric information, showing how this technique can be used to generate clusters in satellite data, identifying zones with homogeneous characteristics of rainfall regime.

Resumo

O objetivo principal do seguinte estudo é explorar a técnica de Análise de Componentes Principais (PCA) e suas aplicações práticas. A PCA é utilizada para identificar padrões em dados multivariados, permitindo a expressão desses dados de forma a destacar suas semelhanças e diferenças. Isso é alcançado transformando as variáveis originais em um novo conjunto de variáveis, as componentes principais, que são ortogonais e ordenadas de maneira que as primeiras carregam a maior parte da variabilidade nos dados. O processo inclui a padronização dos dados para evitar influências indevidas de variações de escala entre os atributos, o cálculo da matriz de covariância para entender como as variáveis estão associadas entre si, a determinação dos autovalores e

autovetores que identificam as direções e magnitudes das novas características, e a seleção dos componentes principais com base na variância explicada. Além disso, o estudo demonstra a aplicação prática da PCA na seleção de dimensionalidade em imagens de satélite com informações pluviométricas, mostrando como essa técnica pode ser utilizada para gerar agrupamentos em dados satelitais, identificando zonas com características homogêneas de regime de chuva.

10.1. Introdução

O objetivo principal do PCA é identificar padrões em dados e expressar esses dados de forma a destacar suas semelhanças e diferenças. Dessa maneira os dados podem ser reduzidos a uma forma que mantém as características mais importantes. Isso é realizado transformando as variáveis originais em um novo conjunto de variáveis, as componentes principais, que são ortogonais e ordenadas de forma que as primeiras carregam a maior parte da variabilidade nos dados.

10.2. Padronização dos dados

Os dados são normalmente padronizados, para garantir que o PCA não seja indevidamente influenciado por variações de escala entre os atributos. Isso envolve subtrair a média e dividir pelo desvio padrão de cada variável.

Definição: Para cada característica (variável) Xi,os dados são padronizados de acordo com a seguinte fórmula:

$Z_i = \frac{X_i - \mu_i}{\sigma_i}$

onde μ_i é a média e σ_i é o desvio padrão de X_i. Isso transforma cada característica para ter média zero e variância unitária.

Detalhes: Quando lidamos com uma matriz X de dimensão i × j, onde i é o número de observações e j é o número de características, a padronização é realizada para cada coluna (característica) da matriz.

1. Calcular a média da característica j: $\mu_j = \frac{1}{n}\sum_{i=j}^{n} X_{ij}$

X_{ij} representa o valor da característica j na observação i, e μ_j é a média de todos os valores da característica j.

2. Calcular o desvio padrão da característica j: $\sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \mu_j)^2}$

O desvio padrão σ_j mede a dispersão dos valores da característica j em torno da média μ_j.

3. Padronizar cada valor da característica j: $Z_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j}$

A matriz Z, resultante desse processo, terá as mesmas dimensões que a matriz original X, mas cada coluna j de Z terá média zero e desvio padrão unitário.

10.3. Matriz de Covariância

O cálculo da matriz de covariância ou correlação dos dados, ajudam a entender como as variáveis estão associadas entre si. A matriz de covariância é uma ferramenta estatística

que mede o grau de variação conjunta entre pares de características.

1. Definição da Matriz de Covariância: A matriz de covariância Σ de uma matriz de dados Z, onde cada linha representa uma observação e cada coluna uma característica padronizada, é definida como:

$\Sigma = \frac{1}{n-1} Z^T Z$

Z^T é a transposta de Z, e o produto $Z^T Z$ é uma matriz $p \times p$ (onde p é o número de características), cujos elementos representam as covariâncias entre todas as possíveis pares de características.

2. Elementos da Matriz de Covariância: Cada elemento σ_{jk} da matriz de covariância Σ pode ser interpretado como a covariância entre as características j e k e é calculado por:

$\sigma_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (Z_{ij} Z_{ik})$

onde Z_{ij} e Z_{ik} são os valores padronizados das características jj e kk, respectivamente, para a observação ii.

3. Propriedades da Matriz de Covariância: Simetria: A matriz de covariância Σ é sempre simétrica, ou seja, $\sigma_{jk} = \sigma_{kj}$. Diagonal: Os elementos na diagonal principal de Σ representam as variâncias de cada característica padronizada, pois a covariância de uma característica com ela mesma é a sua variância.

10.4. Cálculo dos Autovalores e Autovetores

Os autovalores e autovetores da matriz de covariância são calculados. Os autovetores determinam as direções das novas características (componentes principais), e os autovalores determinam a sua magnitude. Em outras palavras, os autovalores explicam a variância dos dados ao longo dos autovetores.

1. Equação Característica: A matriz de covariância Σ é uma matriz simétrica $p \times p$ (onde p é o número de características), e os autovalores λ e os autovetores v desta matriz são soluções da equação característica:

$\Sigma v = \lambda v$ ou, rearranjando termos: $(\Sigma - \lambda I)v = 0$

onde I é a matriz identidade de dimensão $p \times p$. Esta é uma equação de valor próprio, indicando que os autovalores λ são escalares e os autovetores v são vetores nãonulos.

2. Encontrando os Autovalores: Os autovalores são as soluções da equação determinante que resulta da equação característica:

$det(\Sigma - \lambda I) = 0$

Resolver essa equação polinomial para λ fornece os autovalores de Σ. Cada autovalor reflete a variância dos dados ao longo do eixo definido pelo seu autovetor correspondente.

3. Encontrando os Autovetores: Após determinar os autovalores, substituímos cada λ de volta na equação:

$$(\Sigma - \lambda I)v = 0$$

para encontrar o autovetor v correspondente a cada autovalor λ é necessário resolver o sistema de equações lineares.

4. Propriedades dos Autovetores: Ortogonalidade: Em matrizes simétricas como Σ, os autovetores associados a diferentes autovalores são ortogonais entre si. Isso implica que os componentes principais são independentes um do outro. Normalização: Os autovetores são normalmente normalizados para ter norma unitária, ou seja, o comprimento de cada autovetor é ajustado para ser igual a 1.

10.5. Seleção dos Componentes Principais

Os autovalores λ ordenados, da matriz de covariância Σ, são utilizados para selecionar os componentes principais. A ordenação reflete a importância de cada componente principal em termos da variância que ele captura dos dados.

1. Ordenação dos Autovalores: Depois de calcular os autovalores λ da matriz de covariância *Sigma*, o primeiro passo na seleção dos componentes principais é ordená-los em ordem decrescente. O autovalor mais alto corresponde ao componente principal que captura a maior parte da variância, e assim por diante.

2. Cálculo da Variância Explicada: Cada autovalor representa a variância ao longo da direção do seu autovetor correspondente. A soma de todos os autovalores dá a variância total. A fração da variância explicada por cada autovalor é calculada como:

$$VarianciaExplicada(\lambda_i) = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

onde λ_i é o i-ésimo autovalor e p é o número total de características (autovalores).

3. Decisão sobre o Número de Componentes: A seleção do número de componentes principais k a serem retidos depende de quantificar quanto da variância total dos dados desejamos capturar. Essa decisão pode ser tomada com o corte por Limiar de Variância: Escolher o menor número k de componentes principais tal que a soma de suas variâncias explicadas seja maior que um limiar predeterminado, como 85%, 90%, ou 95%, por exemplo.

10.6. Transformação dos Dados

Nessa etapa os dados originais são projetados nos novos eixos formados pelos componentes principais selecionados. Este processo reduz a dimensionalidade dos dados, mantendo ao mesmo tempo as características mais significativas em termos de variância.

1. Matriz de Projeção: Após selecionar os k primeiros autovetores da matriz de covariância Σ (os k autovetores correspondentes aos maiores autovalores), construímos a matriz de projeção V_k. Esta matriz $p \times k$ é formada pelos autovetores como colunas:

$$V_k = [v_1, v_2, ..., v_k]$$

Cada coluna v_i é um autovetor e representa um componente principal.

2. Projeção dos Dados: A matriz de dados padronizados Z é então projetada sobre os componentes principais utilizando a matriz de projeção V_k. A projeção é realizada

através da multiplicação da matriz Z pela matriz V_k:

$T_k = ZV_k$ onde:

Z é a matriz $n \times p$ dos dados, com n sendo o número de observações e p o número de características. T_k é a matriz resultante $n \times k$ dos dados transformados. Cada coluna de T_k corresponde a um dos componentes principais, e cada linha representa as coordenadas de uma observação no novo espaço de componentes principais.

10.7. Desafios e Limitações do PCA

Sensibilidade a Outliers O PCA minimiza a norma dos dados, o que dá um peso maior aos outliers. Outliers podem distorcer significativamente os componentes principais, afetando a interpretação dos dados.

Suposição de Linearidade O PCA assume que as relações entre as variáveis são lineares. Em dados com relações complexas e não lineares, o PCA pode falhar em capturar as variações significativas.

10.8. Prática 01 - Uso de Componentes principais (PCA) na seleção de dimensionalidade em imagens de satélite com informações pluviométricas

Na primeira prática, foi demonstrado o uso das componentes principais para seleção de dimensionalidade na geração de agrupamentos em dados satelitátios. Estes agrupamentos são de zonas com características homogêneas de regime de chuva no estado de Goiás no período de 20 anos, que corresponde do ano de 2001 a 2020.

Como referência bibliográfica, foi utilizado o estudo de Boschi et.al.(2011), onde o autor fez uso de técnicas de mineração de dados para análise da precipitação pluvial decenal no Rio Grande do Sul. Este, analisou o comportamento espaço-temporal do regime de chuva (precipitação) no Rio Grande do Sul, em dois períodos, o primeiro consiste de 1987 à 1996 e o segundo período de 1997 à 2006.

O modelo de mineração de dados empregado foi CRISP-DM (Cross Industry Standard Process for Data Mining) e o algoritmo para agrupamento, fez uso do KMeans. Os resultados do estudo de Boschi et.al (2011) foram que houve um incremento significativo de quantidade de chuva entre esses períodos (20 – 240 mm) e que entre 1997 e 2006, a apresentou picos mais intensos de chuva, tanto para mais (acima da média), quanto para menos (abaixo da média), sendo que os menores volumes foram observados especialmente nos últimos três anos da série (2004, 2005 e 2006).

Os dados utilizados na prática 01 são oriundos do sensor abordo do satélite GPM (Global Precipitation Measurement), é um projeto liderado pela NASA (The National Aeronautics and Space Administration) e JAXA (Japan Aerospace Exploration Agency) visando compreender os processos físico-meteorlógicos, com uma resolução espacial de 0,1° com informações de precipitação a cada 3 horas.

O ambiente de desenvolvimento para este exemplo foi o Google Earth Engine, na linguagem JavaScript. O Google Earth Engine combina um catálogo de vários petabytes de imagens de satélite e conjuntos de dados geoespaciais com recursos de análise e algo-

ritmos em escala planetária. O Earth Engine agora está disponível para uso comercial e permanece gratuito para uso acadêmico e de pesquisa.

Para execução dessa prática, foram executados os seguintes passos: **1º Passo - Dados:** Aquisição das imagens do satélite GPM por meio da plataforma Google Earth Engine(GEE) do satélite GPM entre 01 de janeiro de 2001 a 31 de dezembro 2020. A base do GPM começou a ser disponibilizada GEE a partir de junho de 2020, e encerrou setembro de 2021. A Informação das imagens do GPM são de precipitação mensal na unidade de mm/hr (milímetros por hora).

2º Passo - Conversão de unidade da precipitação: A conversão de unidade é feita para mudar a precipitação representação por mm/hr para mm/mensal, multiplicando por 720, pois o mês possui 720 horas.

3º Passo - Conversão da precipitação mensal para anual (acumulada): A conversão é feita através da soma de todos os valores de cada mês, gerando um único valor de precipitação. Exemplo do 3º Passo: No ano de 2021, eu tenho 12 imagens, onde cada imagem representa o mês, então, foi feito a soma de cada pixel em imagens diferentes, assim, tornando apenas uma imagem no ano de 2021 conforme a figura 01.

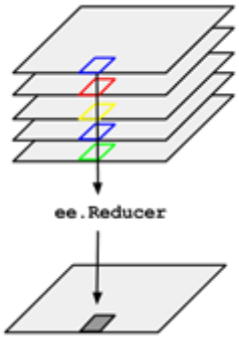

Figura 01 – Exemplo do modelo de redução das imagens aplicado.

4º Passo - Aplicação da técnica de componente principal: Nesta etapa foi calculado a média de cada imagem anual de precipitação, com o intuito de centralizar o valor original para o centro dos eixos; cálculo dos autovetores e autovalores do conjunto de dados; cálculo da variância de cada autovetores. Os dados originais não foram normalizados, pois a unidade de medida de todas as imagens era da mesma grandeza.

5º Passo - Seleção das componentes: Foram selecionadas as 5 primeiras componentes que representam 80% da variabilidade da base de dados. Utilizou o princípio de Pareto, a regra 80-20, onde 80% do resultado, provêm de 20% de ações.

6º Passo - Agrupamento de regiões na imagem: No agrupamento fez o uso do algoritmo Kmeans, com 5 agrupamentos.

Os resultados foram analisados usando: o gráfico da porcentagem dos autovalores em seus autovetores na figura 02; Acumulativo dos autovalores em cada componente (figura 03) e a distribuição espacial dos clusters gerados.

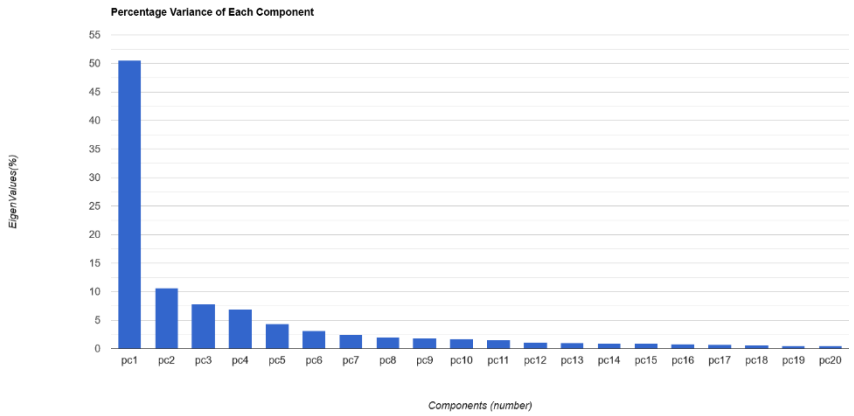

Figura 02 – Autovalores em porcentagem em cada componente.

Nota-se que a quantidade de componentes criadas foram a mesma quantidade de imagens e/ou atributos inseridos no algoritmo de cálculo das componentes principais. A primeira componente é responsável por 50% da variabilidade dos dados, seguindo pela 2ª componente com um pouco mais 10% e a 3ª componente entre 5 a 10% da variabilidade dos dados.

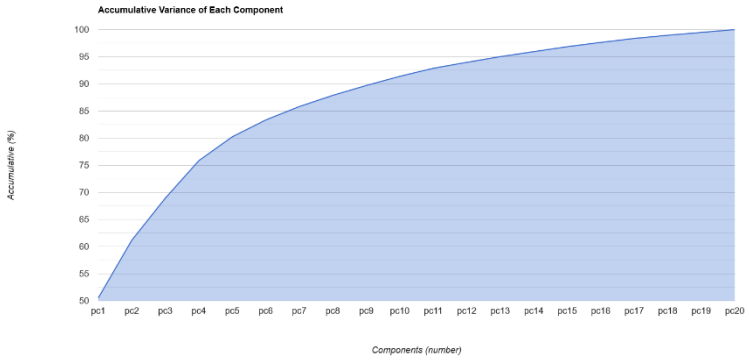

Figura 03 – Autovalores acumulados.

Conforme o gráfico de porcentagem de acumulação de variância das componentes principais geradas, foram selecionadas as 5 primeiras componentes, pois estas dão mais de 80% de variabilidade do conjunto de dados. Dessa forma, foi mantido o princípio de Pareto, onde 80% do trabalho, é causado por 20% das ações exercidas. Cinco componentes principais (aproximadamente 23%) equivale a 80% do conjunto de dados.

As figuras 04 e 05 mostram a dispersão entre a primeira e segunda componente principal e a distribuição geográfica do agrupamentos feito nos 5 primeiras componentes.

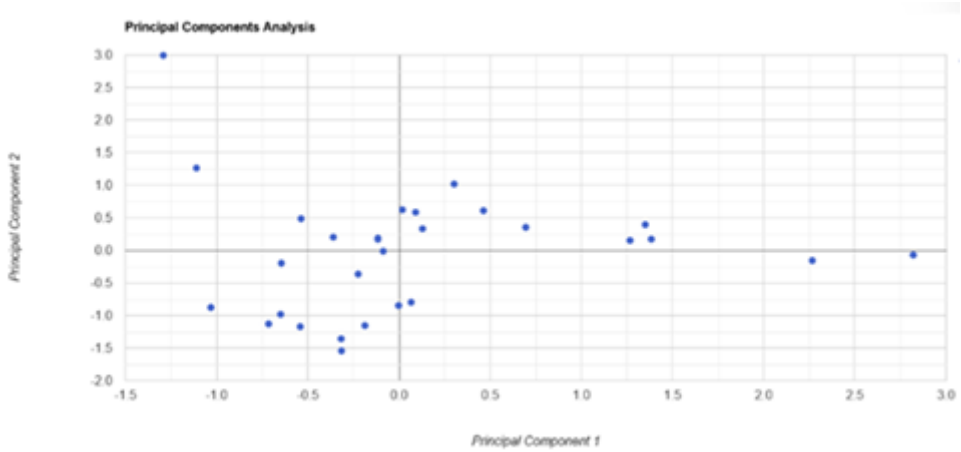

Figura 04 – Dispersão entre a primeira e a segunda componentes principais.

Color	Cluster	Precipitação Média Máxima	Precipitação Média Mínima
	1	1563.88	1416.12
	2	1473.78	**1229.73**
	3	**1838.23**	1446.86
	4	**1825.13**	1442.69
	5	**1638.30**	1577.39

Figura 05 – Agrupamentos gerados e seus valores mínimos e máximos de precipitação.

Conforme a figura 04, nota-se que os valores estão mais dispersos no eixo x (PC1) do que o eixo y (PC2), comprovando que a primeira componente tem variância maior que as outras componentes, conforme a figura 02, mais que 50%. Já figura 05, é visualizado a distribuição geográfica (georreferenciada) dos agrupamentos feitos pelo KMeans. Nota-se que o cluster 1 possui um regime de chuva(precipitação) mais estável, pois a variabilidade é pequena, ou seja, os valores máximos e mínimos são bem próximos em comparação aos outros clusters.

Já o cluster 2, possui os menores valores no seu intervalo de máximo e mínimo, que compreende entre 1416,12 e 1563,88 mm, sendo regiões com pouco presença de chuvas e regiões mais seca e não muito adaptada para culturas agrícolas que não necessitem de irrigação de forma natural.

10.9. Prática 02 - Uso de Componentes principais (PCA) para demonstração da influência das variáveis em um determinado grupo

Neste segundo caso prático o intuito é o de demonstrar que o PCA não apenas pode ser utilizado para redução de dimensionalidade mas também apresentar a influência das várias variáveis que compõem as dimensões de determinado estudo.

Os dados para este caso prático foram obtidos do estudo feito pelo Instituto Mauro

Borges do Estado de Goiás, instituto referência em pesquisas e estatísticas nas áreas de economia, geoprocessamento, geografia e ciências sociais, tendo como atribuições dentre outras a responsabilidade de elaboração de estudos que visem embasar a criação e aplicação de políticas públicas no Estado de Goiás.

Os dados extraídos são parte do Índice de Desempenho dos Municípios Goianos (IDM), tal estudo é feito desde o ano de 2010, tem um frequência bianual para sua apresentação e serve como base para apresentar um quadro qualitativo dos diversos aspectos que são utilizados como medida de como vem sendo aplicados os recursos públicos. Em essência traduzir o desempenho dos municípios goianos em um indicador sintético.

O IDM é uma medida sintética de parte do contexto socioeconômico dos municípios goianos em seis áreas de atuação: Economia, Educação, Infraestrutura, Saúde, Segurança e Trabalho. Cada dimensão contribui igualmente para a composição do índice final, ou seja, cada uma tem o mesmo peso no cálculo final.

São, ao todo, 37 variáveis selecionadas para conferir o desempenho dos municípios goianos, o que permite identificar a dinâmica temporal dos indicadores e a de casos de referências, dada a relativa comparabilidade entre os municípios, além de contribuir para o direcionamento de políticas públicas. A média geral do IDM, calculada com base na pontuação dos 246 municípios, ficou em 5,00 pontos, sendo o mesmo valor obtido pela mediana. A média geral do IDM não representa a nota do estado, mas a média dos 246 municípios analisados.

A Segurança apresenta a maior nota média (7,94) refletindo a situação da maioria dos municípios do estado, que possui menos de 100 mil habitantes e baixa ocorrência de crimes. A Saúde obteve a segunda melhor nota (7,72), pois a maioria dos municípios possui pelo menos um médico por mil habitantes e conta com 100% de cobertura da estratégia Saúde da Família. Por outro lado, a dimensão Economia apresenta a pior nota média (2,44), sinalizando que a maioria dos municípios tem economia pouco expressiva, quando comparado à Goiânia, primeira colocada.

Para execução dessa prática, foram executados os seguintes passos:

1º Passo - Aquisição dos dados: Os dados através do site https://goias.gov.br/imb/idm-indice-de-desempenho-dos-municipios-2020-2021/ Observação: Conforme a documentação do estudo os dados já estão normalizados o que dispensa esta etapa por nossa parte.

2º Passo - Configuração do ambiente com a instalação e utilização das bibliotecas: Nesta etapa, fez uso do sklearn para a aplicação do PCA, do matplotlib e seaborn para geração dos gráficos, pandas para armazenamento dos dados e o jupyter lab como o ambiente de desenvolvimento.

3º Passo: Geração dos autovalores e autovetores: Com o recurso da biblioteca sklearn.decomposition import PCA foi possível gerar os autovalores e autovetores, além da variância explicada e a razão da variância explicada.

4º Passo: Geração das matrizes de covariância e correlação: Os valores que servirão de base para a geração dos gráficos do 5º passo.

5º Passo: Visualização dos dados: Por meio de gráficos, demonstrou-se a variância entre cada uma das variáveis em cada dimensão.

Segue a tabela 01 e as figuras 06 e 07mostram informações da variância explicada e acumulada, correlação entre os IDM's as primeiras componentes, e gráfico de dispersão entre a primeira e segunda componente principal respectivamente.

Tabela 01 – Correlação entre os IDM's com as primeiras e segundas componentes principais.

	PC1	PC2
IDM	-0.302939	0.451487
IDM_Economia	-0.310779	-0.521476
IDM_Trabalho	-0.809690	0.124517
IDM_Educacao	-0.265914	0.457287
IDM_Seguranca	0.955214	-0.136754
IDM_Infraestrutura	-0.975208	0.072173
IDM_Saude	0.323612	0.857395

Tabela 02 – Os 10 municípios selecionados aleatoriamente.

Municipio	IDM	IDM_Economia	IDM_Trabalho	IDM_Educacao	IDM_Seguranca	IDM_Infraestrutura	IDM_Saude
Amaralina	4.59	2.23	2.34	4.87	9.63	1.21	7.28
Aporé	5.25	2.89	3.05	6.40	9.02	2.11	8.02
Buriti Alegre	4.97	2.59	3.18	5.86	6.07	4.19	7.91
Campos Belos	5.26	2.21	3.63	6.64	7.30	4.30	7.51
Damianópolis	4.97	0.73	2.97	6.34	8.60	2.36	8.82
Formosa	4.94	2.80	3.59	5.69	7.34	4.13	6.08
Goiatuba	5.28	3.41	3.36	5.60	7.57	2.88	8.84
Orizona	5.12	2.80	2.99	6.13	8.69	2.40	7.69
Porangatu	4.93	2.80	3.46	5.41	7.30	2.98	7.65
Santa Bárbara de Goiás	5.12	2.14	3.30	5.34	8.50	2.77	8.65

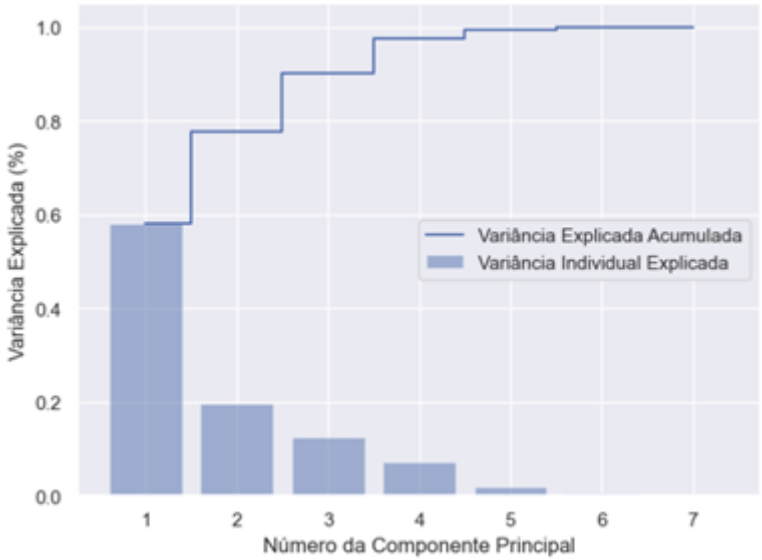

Figura 06 – Autovalores individuais e acumulados por componentes principais.

Figura 07 – Dispersão entre a primeira e a segunda componentes principais.

Para melhor visualização dos dados foi selecionada uma amostra aleatória dos dados de IDM e demais variáveis de 10 municípios e não todos os 246. Através do uso do PCA pôde se notar na figura 06 que 80% por cento dos dados estão nos componentes PC1 e PC2, diante disto na tabela 01 onde está a apresentada a correlação entre as variáveis, o valor mais expressivo para a PC1 é o do IDM da Infraestrutura com -0.975208, ao buscar na figura 07 quem seria o município de maior valor no eixo PC1 encontramos Amaralina, para provar que o dado foi plotado da forma correta observamos na tabela 02 o valor encontrado na coluna de Infraestrutura e registrado para Amaralina é de 1,21 o que faz todo sentido ao levar em consideração que na tabela 01 o valor do IDM da Infraestrutura é negativo demonstrando que quanto menor for o valor mais bem classificado está o município.

Outro fato a ser notado para comprovar que o uso do PCA foi feito com sucesso é analisar qual a variável que possui maior correlação com o IDM da Infraestrutura que nesse caso é o IDM da Segurança com 0,955214, note que o valor dessa variável é positivo então quanto maior for o valor melhor será a classificação do município, na tabela 02 Amaralina possui o valor de 9,63 o maior de todos os municípios analisados.

References

[Brunton 2020] Brunton, S. (2020). Principal component analysis (pca). https://www.youtube.com/watch?v=fkf4IBRSeEc. Acesso em: [02/05/2024].

[Brunton and Kutz 2019] Brunton, S. L. and Kutz, J. N. (2019). *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*, page 24–28. Cambridge University Press.

[Deisenroth et al. 2024] Deisenroth, M. P., Faisal, A. A., and Ong, C. S. (2024). *Mathematics for Machine Learning*, pages 317–347. Cambridge University Press.

[Du 2019] Du, T. (2019). Dimensionality reduction techniques for visualizing morphometric data: Comparing principal component analysis to nonlinear methods. *Evolutionary Biology*, 46.

[Neumayer et al. 2020] Neumayer, S., Nimmer, M., Setzer, S., and Steifdl, G. (2020). On the robust pca and weiszfeld's algorithm. *Applied Mathematics Optimization*, 82:1017–1048.

[Serrano 2019] Serrano, L. (2019). Principal component analysis (pca). https://www.youtube.com/watch?v=g-Hb26agBFg. Acesso em: [02/05/2024].