INSTITUTO DE INFORMÁTICA
Universidade Federal de Goiás

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
TÓPICOS ESPECIAIS EM FUNDAMENTOS DE COMPUTAÇÃO – MATEMÁTICA E ESTATÍSTICA PARA CIÊNCIA DE DADOS
PROF. DR. ROMMEL MELGAÇO BARBOSA

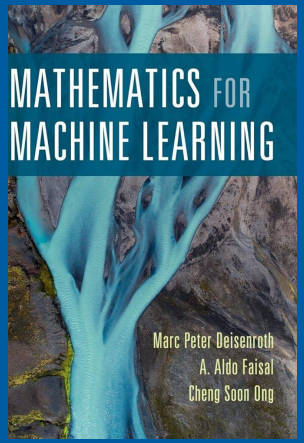
SEMINÁRIOS

PCA

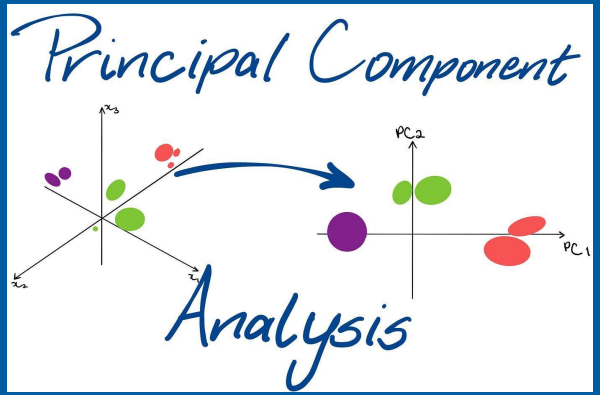
Bernard Silva de Oliveira
Manoel Veríssimo dos Santos Neto
Rafael Rodrigues Silva
Wesley Modanez Freitas

Abril/2024

PPGCC PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
INF INSTITUTO DE INFORMÁTICA
UFG UNIVERSIDADE FEDERAL DE GOIÁS



Capítulo 10
PCA



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.

Como o PCA Funciona?

- Padronização dos Dados
- Covariância / Matriz de Correlação
- Cálculo dos Autovalores e Autovetores
- Seleção de Componentes Principais
- Transformação dos Dados

Padronização dos Dados

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.

Padronização dos Dados

Definição: Para cada característica (variável) X_i, 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.

Padronização dos Dados

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*:

$$\mu_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}$$

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

Padronização dos Dados

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.

Padronização dos Dados

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.

Covariância / Matriz de Correlação

Covariância / Matriz de Correlação

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.

Covariância / Matriz de Correlação

1. Definição da Matriz de Covariância: ΣA 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:

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

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

Covariância / Matriz de Correlação

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**.

Covariância / Matriz de Correlação

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.

Cálculo dos Autovalores e Autovetores

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.

Cálculo dos Autovalores e Autovetores

1. Equação Característica: A matriz de covariância **Σ** é uma matriz simétrica **p × 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 × p**. Esta é uma equação de valor próprio, indicando que os autovalores **λ** são escalares e os autovetores **v** são vetores não-nulos.

Cálculo dos Autovalores e Autovetores

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

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

Polinômio Característico

Para $\lambda \in \mathbb{R}$ e uma matriz quadrada $\boldsymbol{A} \in \mathbb{R}^{n \times n}$

$$p\boldsymbol{A}(\lambda) := det(\boldsymbol{A} - \lambda \boldsymbol{I})$$
$$= c_0 + c_1\lambda + c_2\lambda^2 + \cdots + c_{n-1}\lambda^{n-1} + (-1)^n\lambda^n ,$$

$c_0, \ldots, c_{n-1} \in \mathbb{R}$.

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.

Cálculo dos Autovalores e Autovetores

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.

Cálculo dos Autovalores e Autovetores

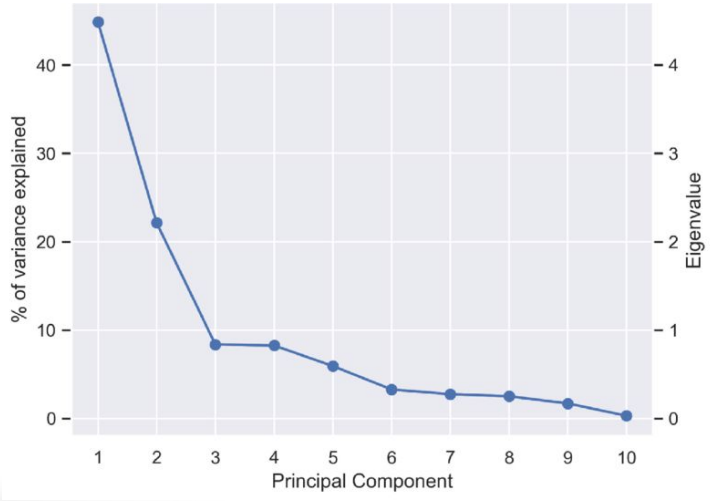
4. Propriedades dos Autovetores:

- **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.

Seleção dos Componentes Principais

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.



Seleção dos Componentes Principais

1. Ordenação dos Autovalores: Depois de calcular os autovalores **λ** da matriz de covariância **Σ**, 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:

$$\text{Variância Explicada}(\lambda_i) = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

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

Seleção dos Componentes Principais

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.

Transformação dos Dados

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.

Transformação dos Dados

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 $\mathbf{V_k}$. Esta matriz **p × k** é formada pelos autovetores como colunas:

$$V_k = [v_1, v_2, \ldots, v_k]$$

Cada coluna $\mathbf{v_i}$ é um autovetor e representa um componente principal.

Transformação dos Dados

2. Projeção dos Dados: A matriz de dados padronizados **Z** é então projetada sobre os componentes principais utilizando a matriz de projeção $\mathbf{V_k}$. A projeção é realizada através da multiplicação da matriz **Z** pela matriz $\mathbf{V_k}$:

$$T_k = ZV_k$$ Onde:

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

Ilustração Prática do PCA

Ilustração Prática do PCA

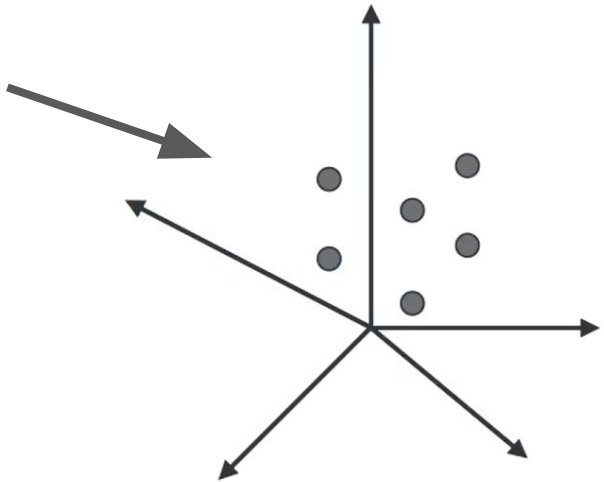
Large Table

X_1	X_2	X_3	X_4	X_5
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*

Ilustração Prática do PCA

Large Table

X1	X2	X3	X4	X5
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*



5D Plot

Ilustração Prática do PCA

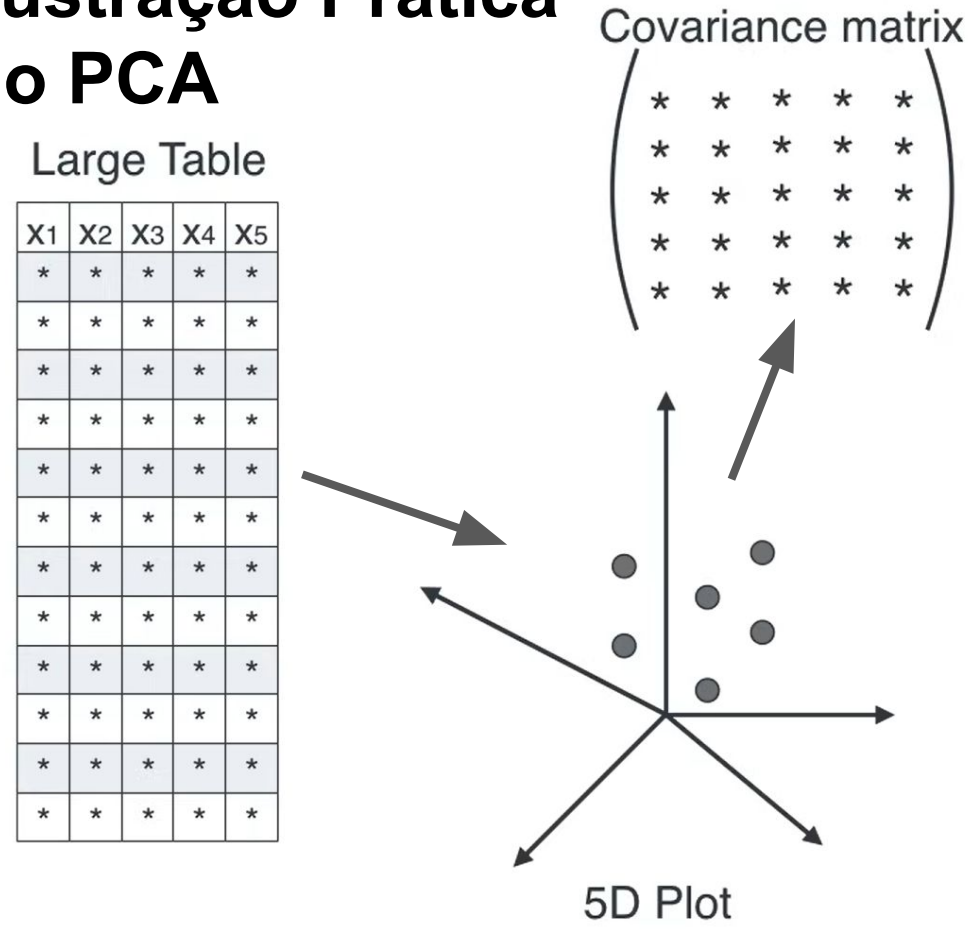


Ilustração Prática do PCA

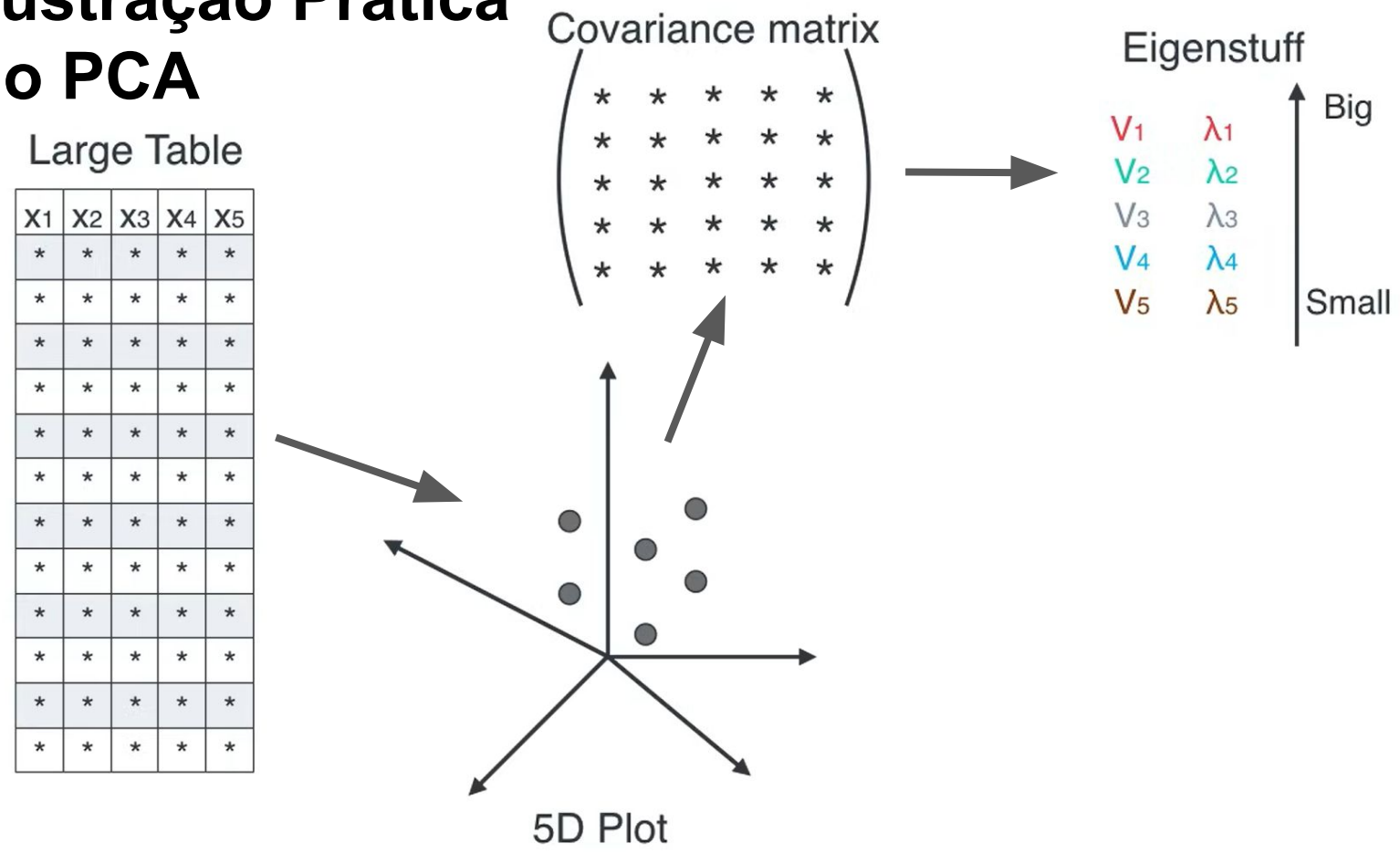


Ilustração Prática do PCA

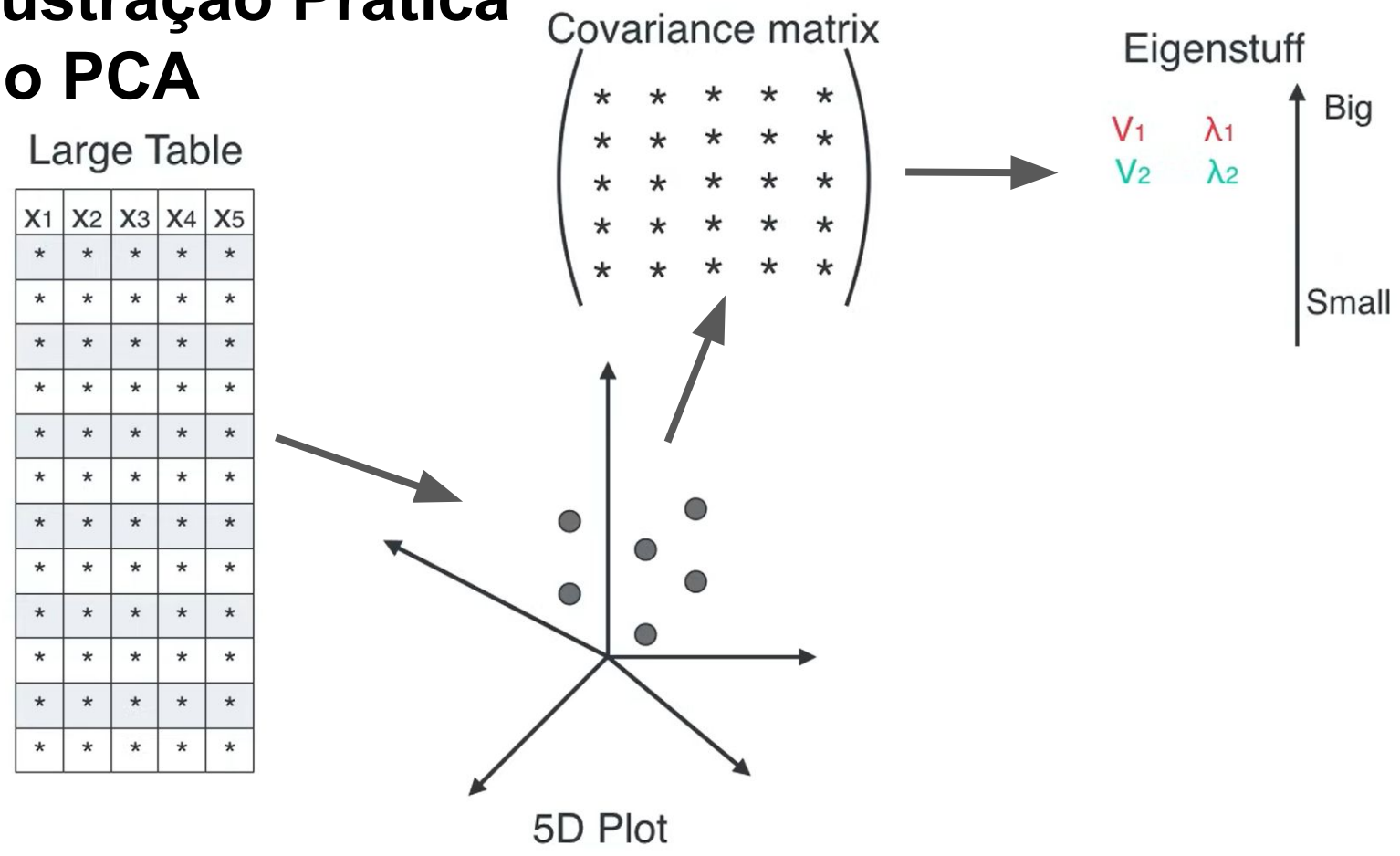


Ilustração Prática do PCA

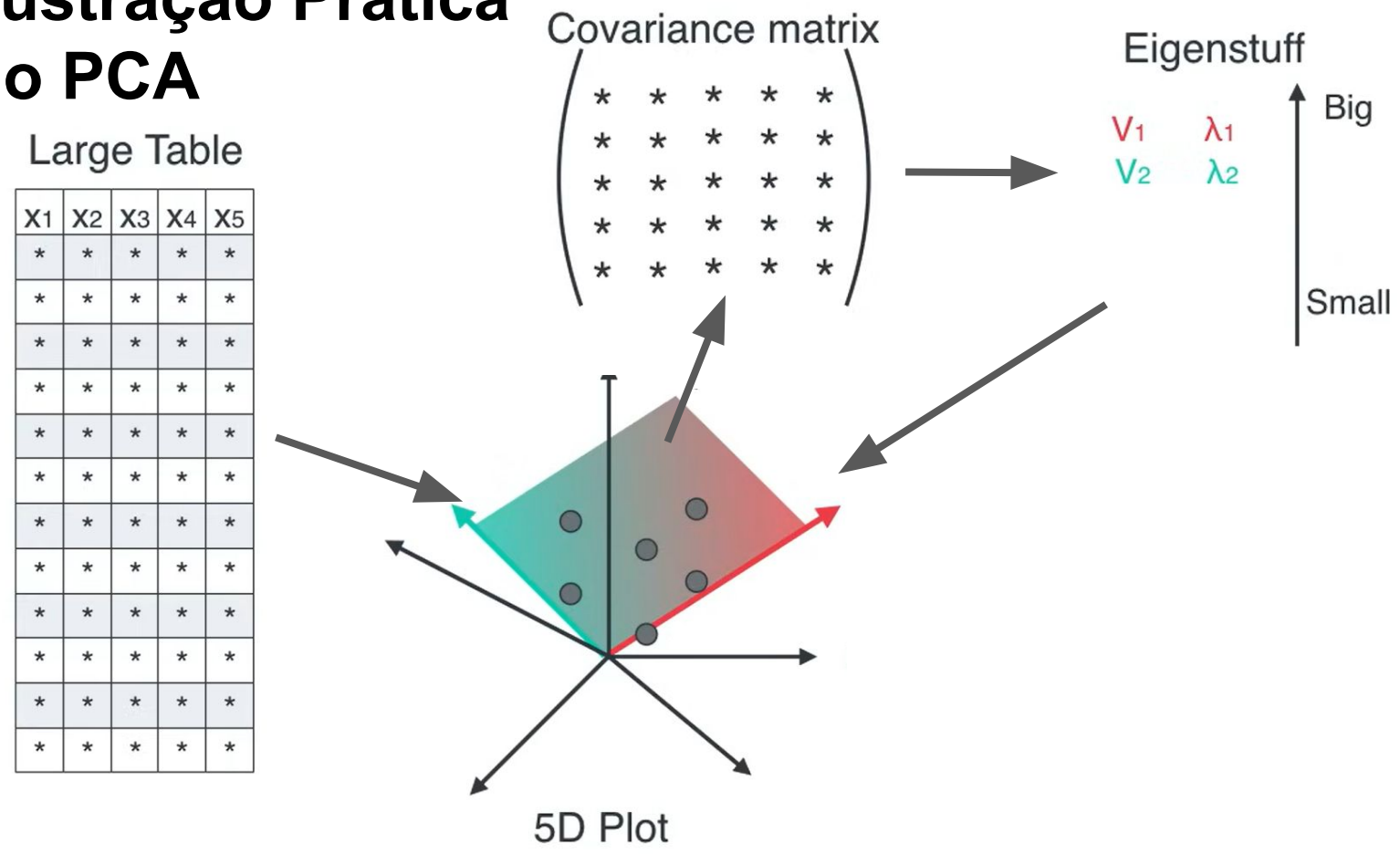


Ilustração Prática do PCA

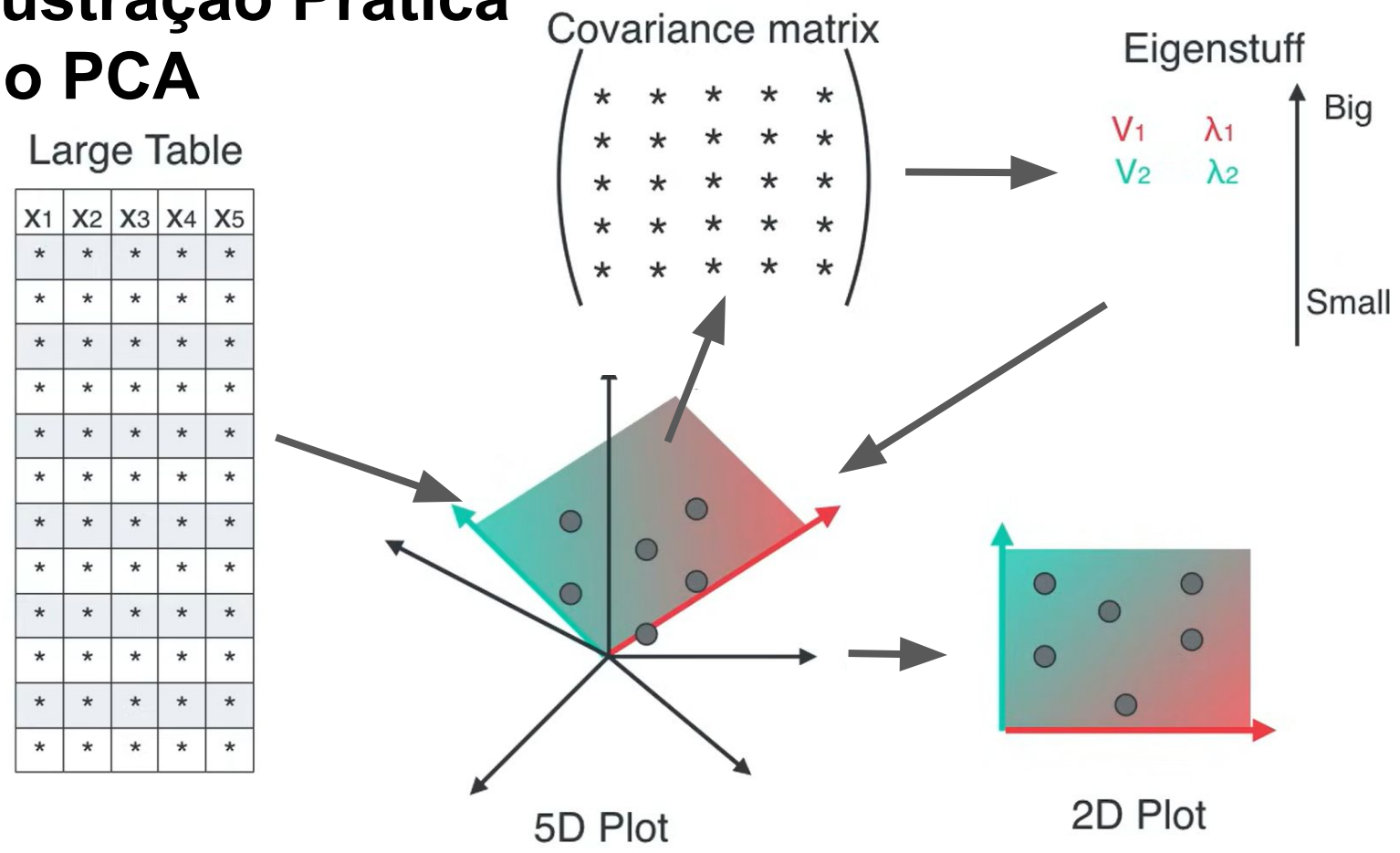
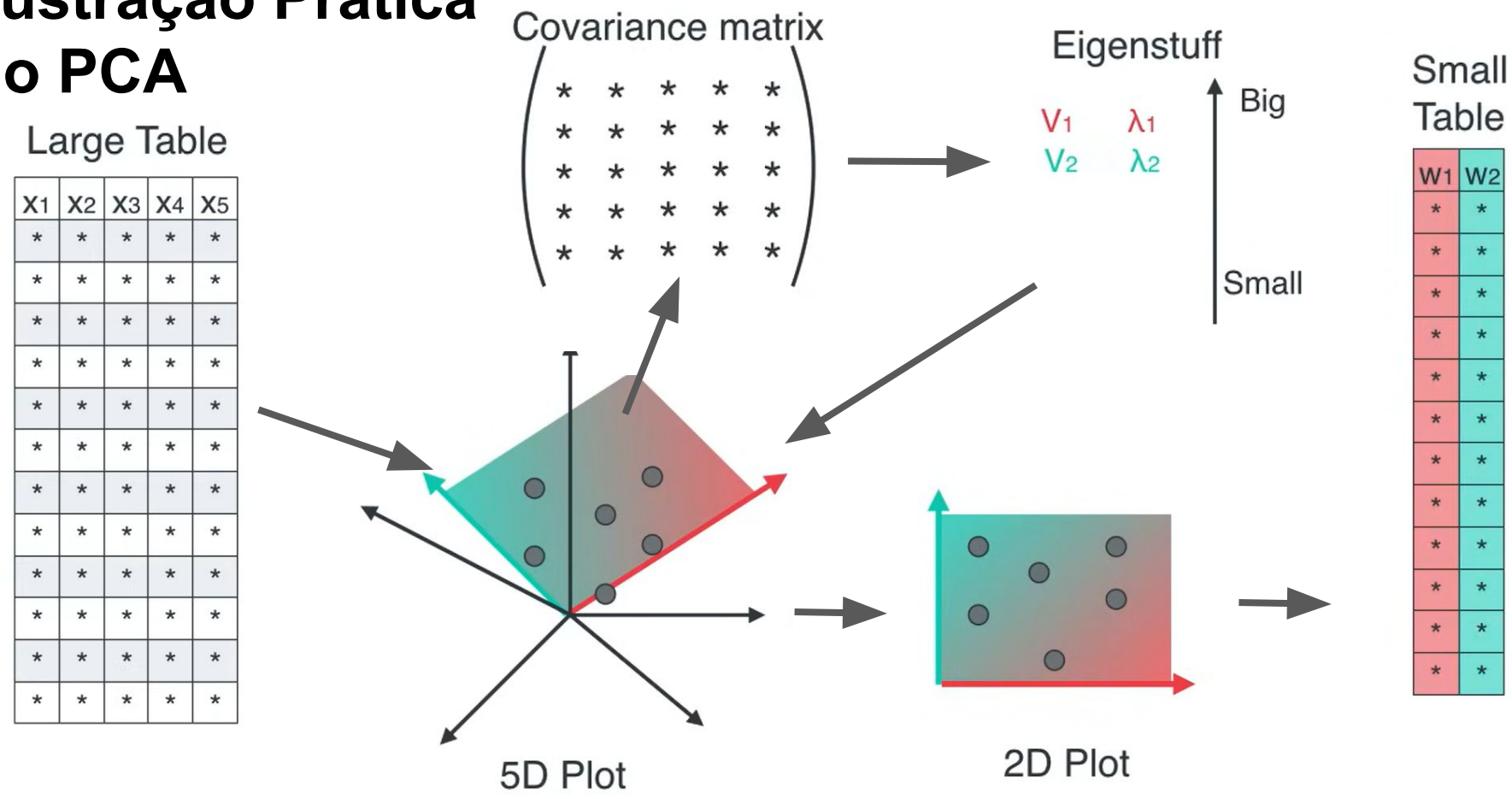
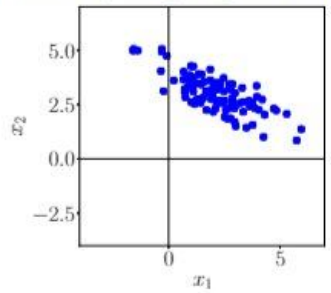
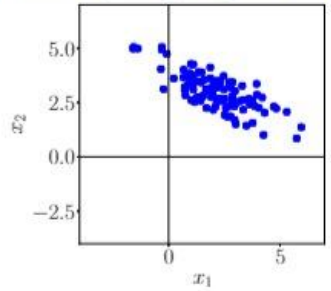


Ilustração Prática do PCA

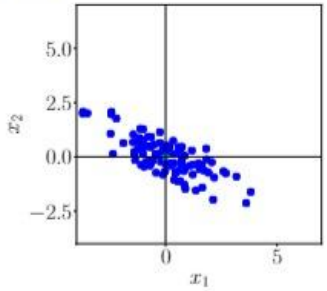




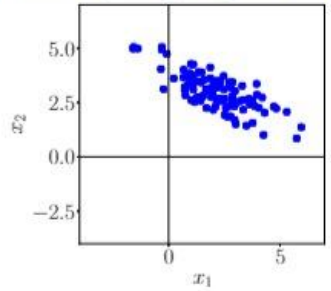
(a) Original dataset.



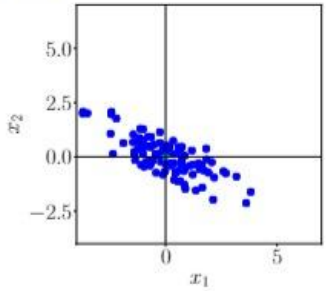
(a) Original dataset.



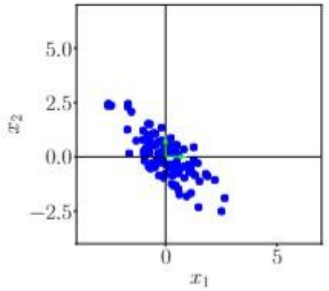
(b) Step 1: Centering by subtracting the mean from each data point.



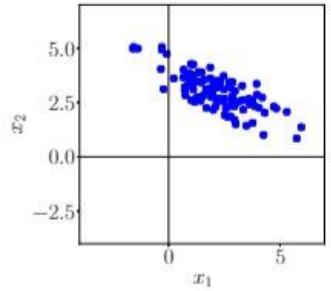
(a) Original dataset.



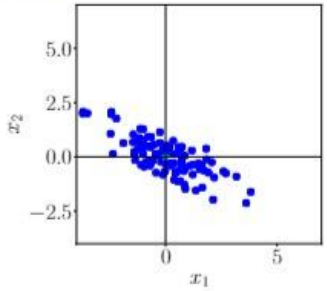
(b) Step 1: Centering by subtracting the mean from each data point.



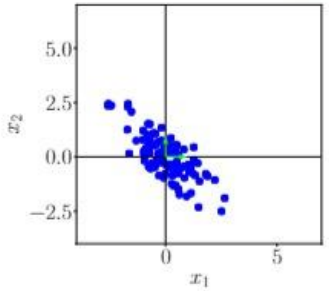
(c) Step 2: Dividing by the standard deviation to make the data unit free. Data has variance 1 along each axis.



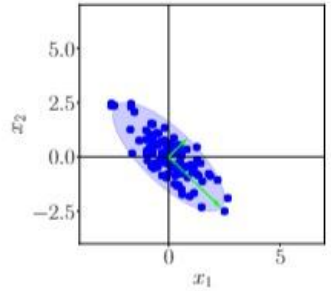
(a) Original dataset.



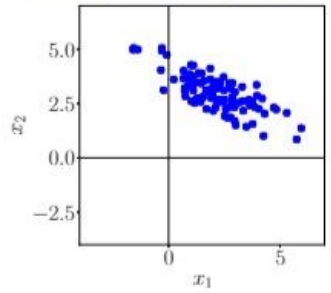
(b) Step 1: Centering by subtracting the mean from each data point.



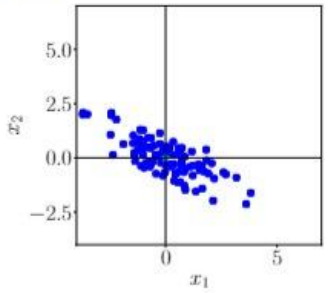
(c) Step 2: Dividing by the standard deviation to make the data unit free. Data has variance 1 along each axis.



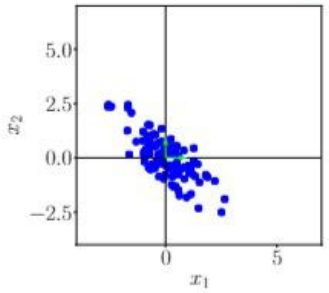
(d) Step 3: Compute eigenvalues and eigenvectors (arrows) of the data covariance matrix (ellipse).



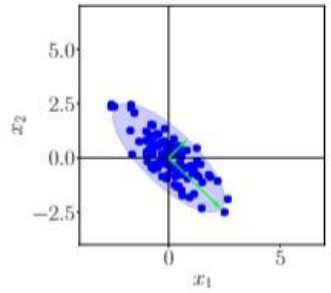
(a) Original dataset.



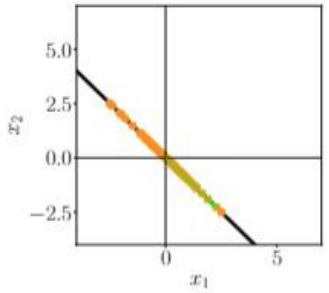
(b) Step 1: Centering by subtracting the mean from each data point.



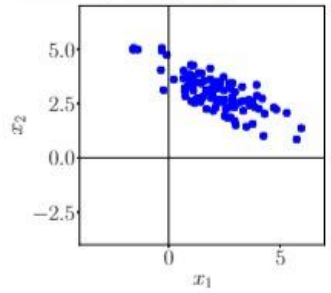
(c) Step 2: Dividing by the standard deviation to make the data unit free. Data has variance 1 along each axis.



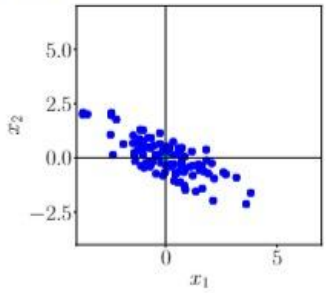
(d) Step 3: Compute eigenvalues and eigenvectors (arrows) of the data covariance matrix (ellipse).



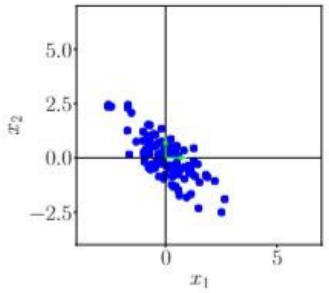
(e) Step 4: Project data onto the principal subspace.



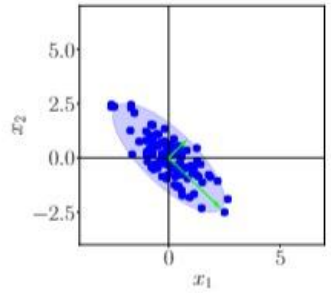
(a) Original dataset.



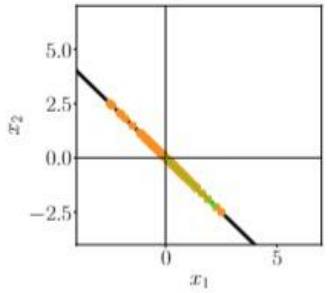
(b) Step 1: Centering by subtracting the mean from each data point.



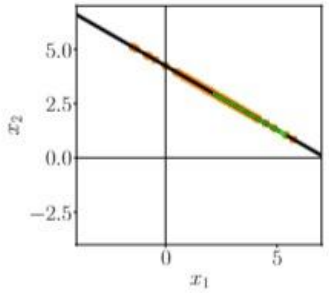
(c) Step 2: Dividing by the standard deviation to make the data unit free. Data has variance 1 along each axis.



(d) Step 3: Compute eigenvalues and eigenvectors (arrows) of the data covariance matrix (ellipse).



(e) Step 4: Project data onto the principal subspace.

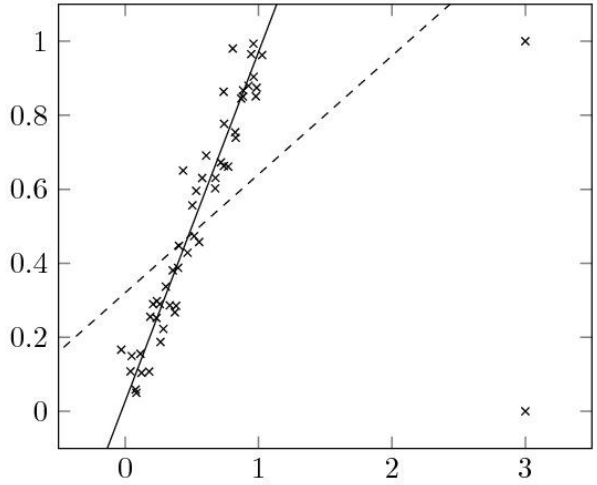


(f) Undo the standardization and move projected data back into the original data space from (a).

Desafios e Limitações do PCA

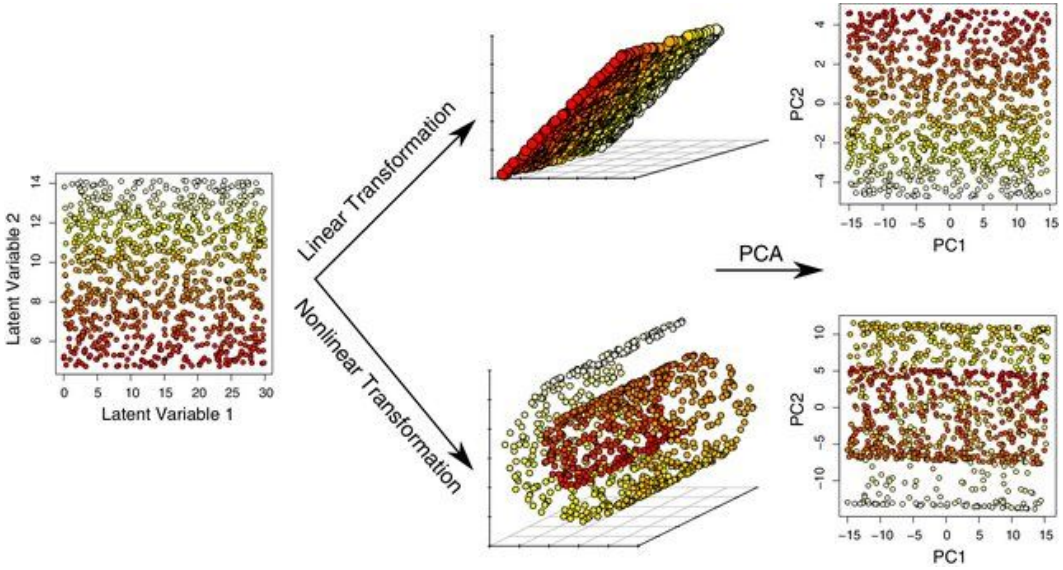
Sensibilidade a Outliers

- 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.



Exemplo Prático 01

Exemplo Prático 01

- Neste exemplo iremos demonstrar o uso da técnica de componentes principais para seleção de dimensionalidade na geração de agrupamentos em dados satelitários.

- Agrupamentos de zonas homogêneas do regime pluviométrico no estado de Goiás.

Exemplo Prático 01

- A base de dados utilizadas neste exemplo são imagens adquiridas por sensores a bordo do satélites artificial com informações mensais de precipitação .

- Área de estudo escolhido foi o estado de Goiás, no período de 20 anos, de 2001 à 2021.

Exemplo Prático 01
Referência Bibliográfica

Técnicas de mineração de dados para análise da precipitação pluvial decenal no Rio Grande do Sul

Data mining techniques for decennial analysis of rainfall in Rio Grande do Sul

AUTORIA | SCIMAGO INSTITUTIONS RANKINGS

» Resumos
» Text
» Datas de Publicação
» Histórico

Resumos

O objetivo deste trabalho foi analisar o comportamento espaçotemporal da precipitação pluvial no Estado do Rio Grande do Sul, entre os decênios de 1987-1996 e 1997-2006, por meio de técnicas de mineração de dados. As séries históricas foram adquiridas no sistema de informações hidrológicas Hidroweb. A metodologia utilizada teve como base o modelo CRISP-DM (Cross Industry Standard Process for Data Mining). Foram definidas áreas pluviometricamente homogêneas para os decênios de 1987-1996 e 1997-2006. Em seguida, pela sobreposição dos agrupamentos obtidos para os dois períodos, encontraram-se seis zonas comuns aos dois decênios (A a F). As alterações ocorridas foram avaliadas nas seguintes escalas temporais: anual, sazonal e mensalmente. Os resultados indicaram incrementos significativos (20 a 240 mm) na precipitação anual em todas as zonas, exceto na zona A. Na análise sazonal, as variações foram aleatórias, sendo que, na primavera, todas as zonas apresentaram incremento significativo (44 a 142 mm). Na análise mensal, destaca-se a redução ocorrida no mês de janeiro em todas as zonas, exceto na E. Nos demais meses, as variações foram aleatórias. Os resultados mostram que, entre os decênios, houve uma alteração no volume da precipitação pluvial em todas as escalas temporais analisadas.

Exemplo Prático 01
Referência Bibliográfica

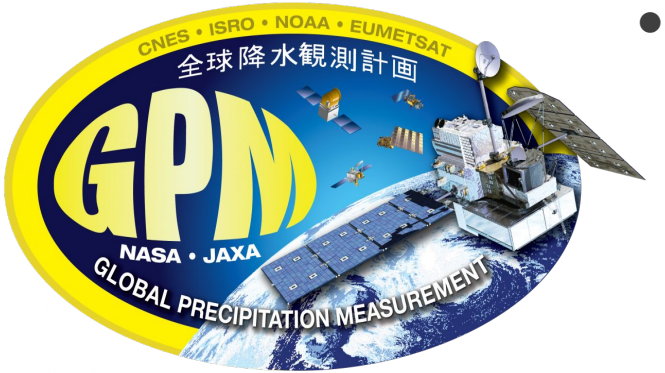
- Analisar o comportamento espaço-temporal da precipitação no estado do Rio Grande do Sul.
- Fez uso do modelo CRISP-DM.
- Algoritmo para clusterização: KMeans.

Exemplo Prático 01

- Dados utilizados neste trabalho é oriundo do sensor a bordo do satélite utilizado GPM (Global Precipitation Measurement Mission).

- Ambiente de desenvolvimento para este exemplo foi o Google Earth Engine, na linguagem JavaScript.

Exemplo Prático 01
Metodologia

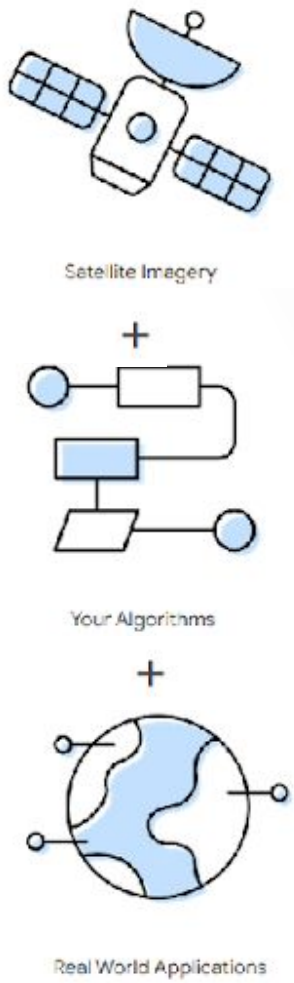


- Projeto liderado pela NASA e Jaxa, visando compreender os processos físico-meteorológicos envolvidos na precipitação com uma resolução espacial cerca de 0,1º à cada 3 horas.

Exemplo Prático 01
Metodologia

- O Google Earth Engine combina um catálogo de vários petabytes de imagens de satélite e conjuntos de dados geoespaciais com recursos de análise e algoritmos em escala planetária.

- O Earth Engine agora está disponível para uso comercial e permanece gratuito para uso acadêmico e de pesquisa.



Exemplo Prático 01
Metodologia - Procedimento

- 1º Passo: 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).

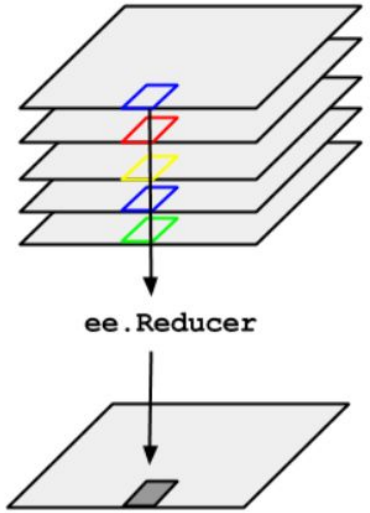
Exemplo Prático 01
Metodologia - Procedimento

- 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 Prático 01
Metodologia - Procedimento

- 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.



Exemplo Prático 01
Metodologia - Procedimento

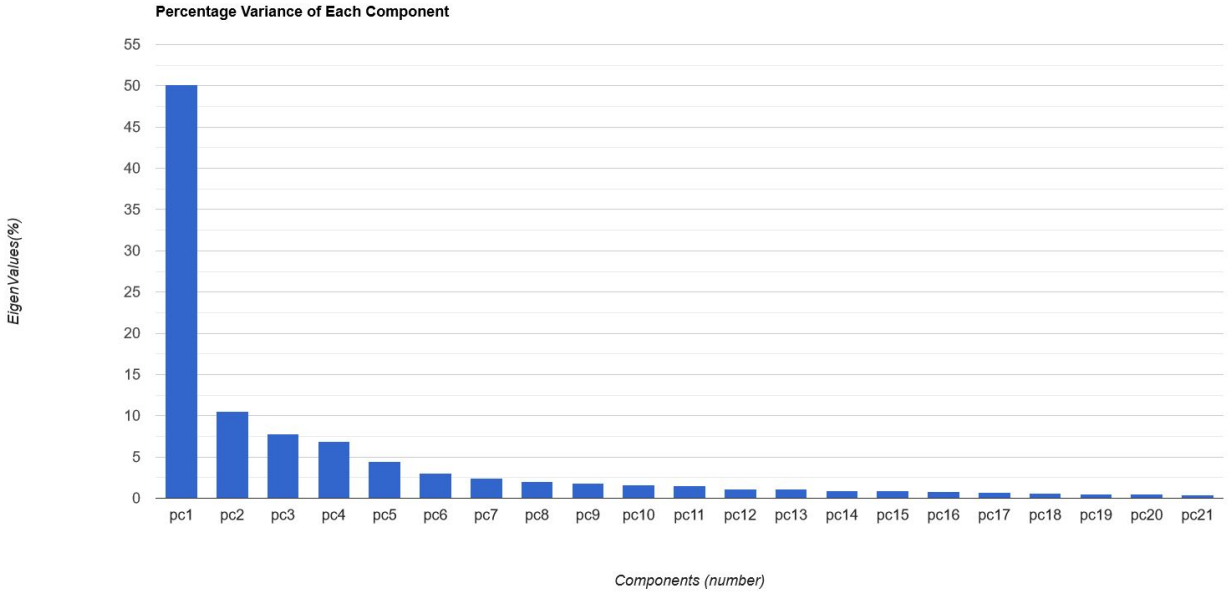
- 4º Passo: Aplicação da técnica de componente principal: nesta etapa foi calculado a média de cada imagem anual de precipitação, com o intuito de centralizar o valor original para o centro dos eixos; cálculo dos autovetores e autovalores do conjunto de dados; cálculo da variância de cada autovetores.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

Exemplo Prático 01
Metodologia - Procedimento

- 5º Passo: Seleção de 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.

Exemplo Prático 01 - Resultados

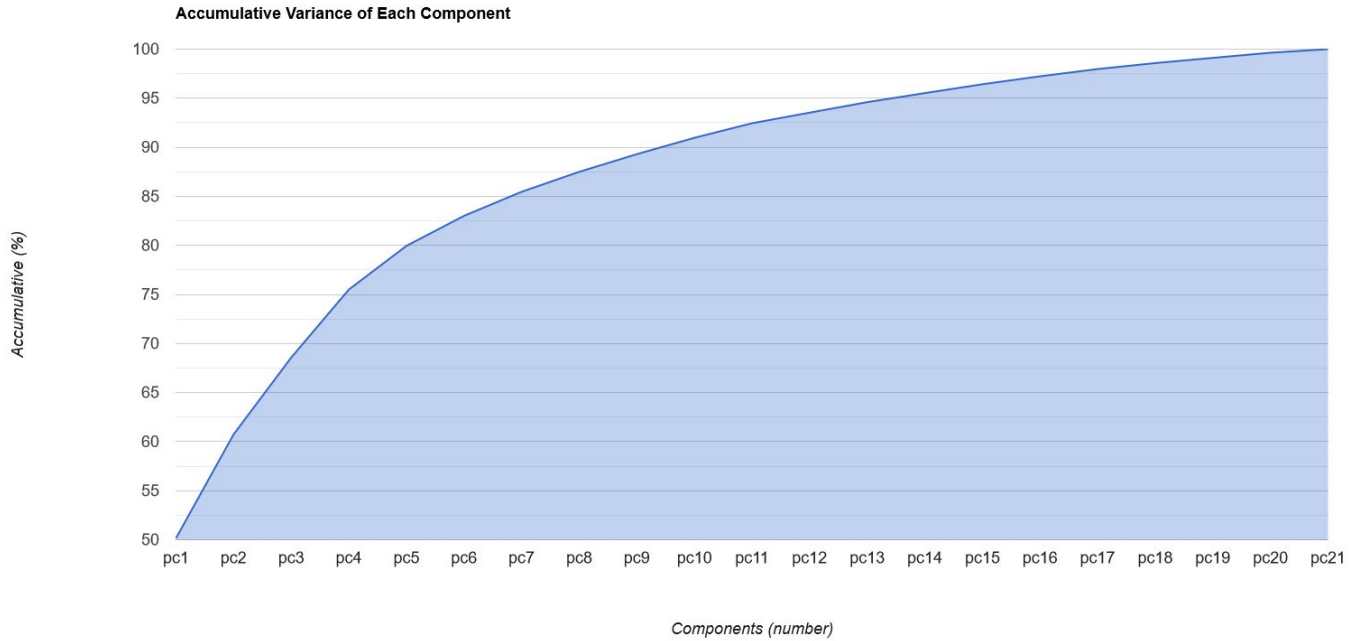


Exemplo Prático 01 - Resultados

- 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.

Exemplo Prático 01 - Resultados

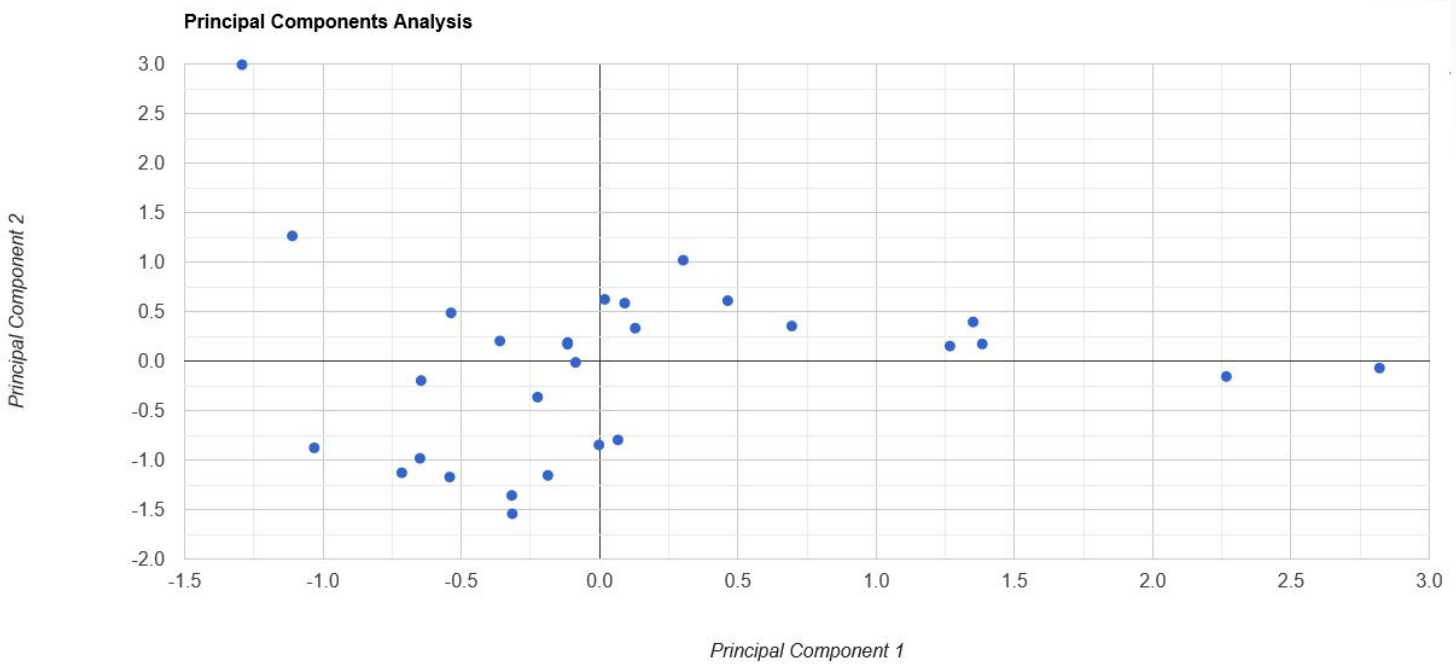


Exemplo Prático 01 - Resultados

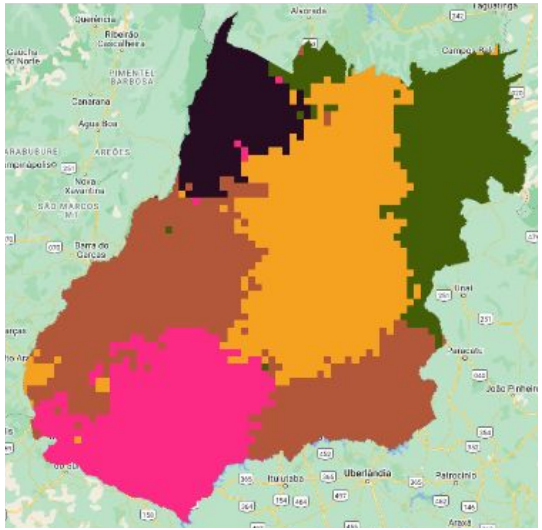
- Conforme o gráfico de porcentagem de acumulação de variância das componentes principais geradas, foram selecionados as 5 primeiras componentes, pois estas dão mais de 80% de variabilidade do conjunto de dados.

- Dessa forma, foi mantido o princípio de Pareto, onde 80% do trabalho, é causado por 20% das ações exercidas. Cinco componentes principais (~ 23%) equivale a 80% do conjunto de dados.

Exemplo Prático 01 - Resultados



Exemplo Prático 01 - Resultados



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

Exemplo Prático 02

Exemplo Prático 02
Objetivo

- Demonstrar com o uso da técnica de componentes principais para agrupamento (PCA), os dados do Índice de Desempenho dos Municípios Goianos (IDM), criado pelo Instituto Mauro Borges do Estado de Goiás (IMB) e como cada uma de suas variáveis influenciam nas respectivas áreas de estudo.

Exemplo Prático 02
O que é o IDM?

- O IMB pública bianualmente o Índice de Desempenho dos Municípios – IDM – desde 2010.

- O índice é uma medida sintética de parte do contexto socioeconômico dos municípios goianos em seis áreas de atuação: Economia, Educação, Infraestrutura, Saúde, Segurança e Trabalho.

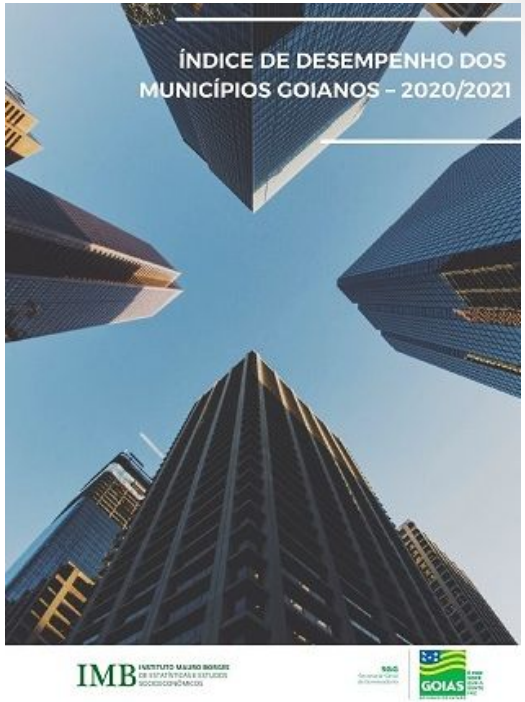
- Cada dimensão contribui igualmente para a composição do índice final, ou seja, cada uma tem o mesmo peso no cálculo final.

Exemplo Prático 02
O que é o IDM?

- 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.

Exemplo Prático 02
Origem dos dados

A base de dados utilizadas neste exemplo são os dados produzidos no estudo feito pelo IMB para os anos de 2020 e 2021. (https://goias.gov.br/imb/idm-indice-de-desempenho-dos-municipios-2020-2021/)



Exemplo Prático 02
Os dados

Dimensão	Quantidade de Variáveis	Origem dos dados das variáveis	Ano Base
Economia	7	IBGE/IMB/SICONFI	2016/2018/2019
Trabalho	4	RAIS	2019
Educação	9	INEP/IMB	2019
Segurança Pública	5	SSP-GO	2019
Infraestrutura	4	SNIS/IMB/ENEL/CHESP/ANATEL	2019
Saúde	8	SES-GO	2019

Exemplo Prático 02
Metodologia - Procedimento

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

Exemplo Prático 02
Metodologia - Procedimento

- 1º Passo: Aquisição dos dados através do site https://goias.gov.br/imb/idm-indice-de-desempenho-dos-municipios-2020-2021/

- 2º Passo: Instalação e utilização das bibliotecas: Scikit-Learn para aplicação do PCA, Matplotlib e Seaborn para geração dos gráficos, Pandas para armazenamento dos dados e JupyterLab como IDE.

- 3º Passo: Geração dos autovalores e autovetores.

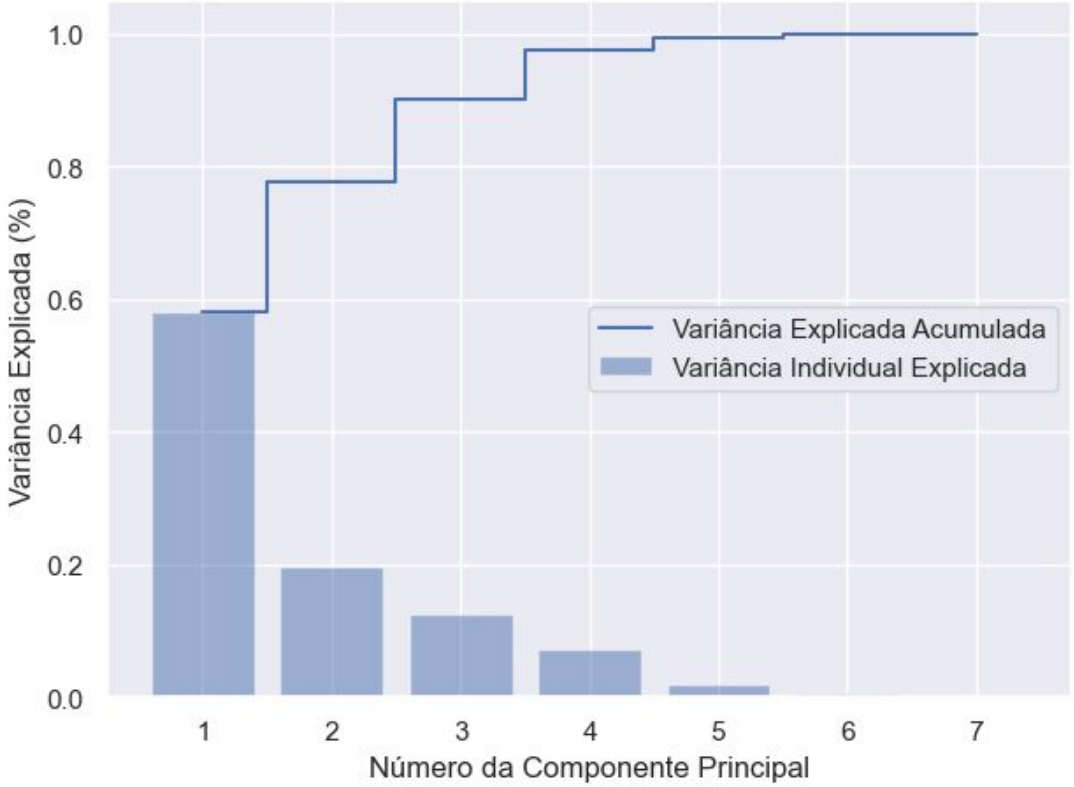
- 4º Passo: Geração das matrizes de covariância e correlação.

Exemplo Prático 02
Metodologia - Procedimento

- 5º Passo: Apresentação do gráfico demonstrando a variância entre cada uma das variáveis.

- 6º Passo: Análise dos resultados e conclusões finais.

Exemplo Prático 01 - Resultados



Exemplo Prático 01 - Resultados

	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

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

Referências

Livros

- DEISENROTH, Marc Peter; FAISAL, A. Aldo; ONG, Cheng Soon. Mathematics for Machine Learning. 2024.
- BRUNTON, Steven L.; KUTZ, J. Nathan. Data Driven Science & Engineering. Seattle: University of Washington, 2017.

Vídeos do YouTube

- Principal Component Analysis (PCA), Serrano.Academy. Disponível em: https://www.youtube.com/watch?v=g-Hb26agBFg. Acesso em: 02/05/2024.
- Principal Component Analysis (PCA), Steve Brunton. Disponível em: https://www.youtube.com/watch?v=fkf4IBRSeEc. Acesso em: 02/05/2024.

Artigos

- Neumayer, Sebastian & Nimmer, Max & Setzer, Simon & Steifdl, Gabriele. (2020). On the Robust PCA and Weiszfeld's Algorithm. Applied Mathematics & Optimization. 82. 1017-1048. 10.1007/s00245-019-09566-1.
- Du, Trina. (2019). Dimensionality Reduction Techniques for Visualizing Morphometric Data: Comparing Principal Component Analysis to Nonlinear Methods. Evolutionary Biology. 46. 10.1007/s11692-018-9464-9.
- Boschi, Raquel S., Stanley R. de M. Oliveira, and Eduardo D. Assad. "Técnicas de mineração de dados para análise da precipitação pluvial decenal no Rio Grande do Sul." Engenharia Agrícola 31 (2011): 1189-1201.