Aplicações de Álgebra Linear na Probabilidade e Estatística

Palavras-Chave: Álgebra Linear, Cadeias de Markov, Análise Multivariada

Aluno e Orientador:

André Lírio Nunes Santos – IFGW - UNICAMP

Prof. Dr. Alex Rodrigo dos Santos Sousa, IMECC - UNICAMP

INTRODUÇÃO:

Este projeto de iniciação científica teve como objetivo realizar um estudo teórico das aplicações da álgebra linear em temas estatísticos, com foco especial na análise de componentes principais e técnicas de clustering. Para isso, foram aplicados dois exemplos práticos utilizando o software RStudio. O primeiro exemplo consistiu na aplicação da análise de componentes principais em uma planilha de dados fornecida pelo IBGE, com o intuito de explorar a estrutura e as relações presentes nos dados. O segundo exemplo abordou o problema de Spike Sorting, uma técnica utilizada no processamento de sinais neuronais, onde a análise de componentes principais foi aplicada para identificar e separar diferentes padrões de atividade neuronal. Esses exemplos ilustram a relevância e a utilidade da álgebra linear na área estatística, fornecendo insights valiosos para a interpretação e análise de dados complexos. Os resultados obtidos demonstraram a eficácia dessas aplicações e abriram possibilidades para futuras investigações nessa interseção entre álgebra linear e estatística.

METODOLOGIA:

Revisão conceitual para devida implementação dos códigos no RStudio e Excel inicialmente para realizar as análises de Análise de Componentes Principais (PCA). Esses códigos foram desenvolvidos com o objetivo de receber os dados brutos como input e produzir como output os autovalores e autovetores resultantes da PCA, juntamente com a porcentagem de representatividade dos valores obtidos e gráficos para demonstrar tais efeitos; com estes resultados, criar um segundo input para gerar um gráfico usando como coordenadas os dois autovetores mais relevantes para aplicação do método de clustering.

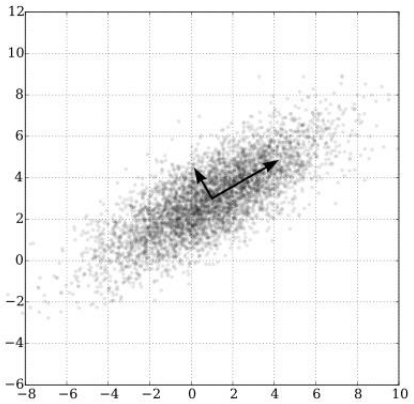

Fig 1: Imagem ilustrativa de PCA

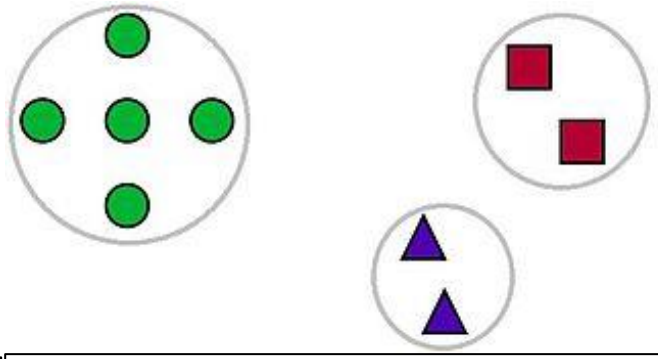

Figura 2: Figura ilustrativa de Clustering

Essa abordagem permitiu uma análise detalhada das principais componentes presentes nos dados, fornecendo informações essenciais para a compreensão das estruturas subjacentes e das relações entre as variáveis estudadas.

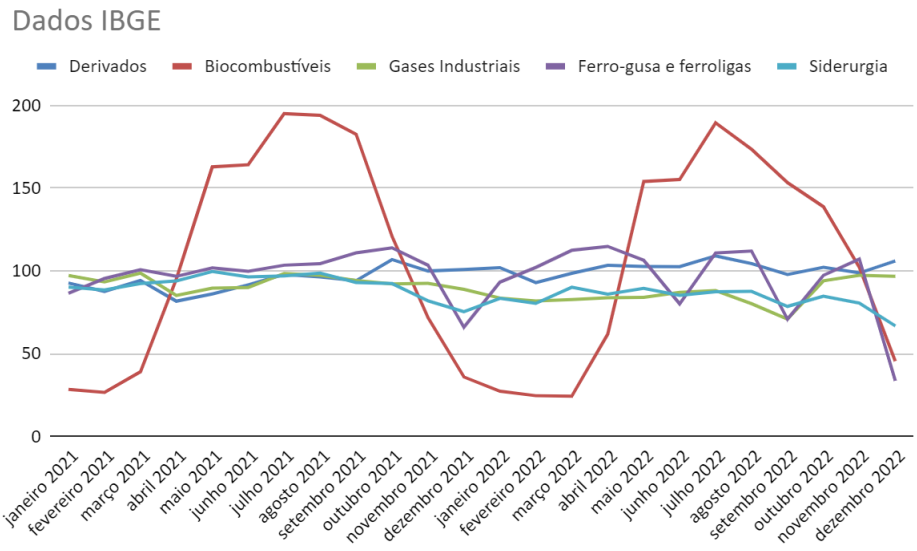

Figura 3: Dados brutos IBGE

Os dados utilizados foram a partir de dois conjuntos distintos: Sidra IBGE a respeito de produção física industrial (Relatório 7511), no que se refere a produção de: derivados de petróleo, biocombustíveis, gases industriais, ferro-gusa e ferroligas e siderurgia; e pelo data sharing da CRCNS (Collaborative Research in Computational Neuroscience) sobre eletrodos no hipocampo para buscar por sinais neuronais distintos, processo conhecido como spike sorting, uma vez concluídas as aplicações foram tomadas as devidas conclusões sobre os conjuntos, vale ressaltar que inicialmente os dados da CRCNS tiveram um desempenho insatisfatório para obter qualquer conclusão, em função disso foi aplicado um filtro de threshold para melhor visualização dos elementos.

A técnica utilizada pode ser descrita como segue: inicialmente, com os dados originais, é gerada uma matriz de correlação para obter os autovalores e autovetores associados a essa matriz, os autovetores são perpendiculares entre si. Em seguida, selecionam-se os dois autovetores com os maiores autovalores associados, e então é realizada a multiplicação desses autovetores pelos dados brutos. O resultado dessa operação é a imagem como apresentada nos resultados do estudo. Por fim, é aplicada a técnica de clustering por centroides para identificar padrões aceitáveis entre os valores. É importante ressaltar que a representatividade dos resultados obtidos está diretamente relacionada à relação entre os dois maiores autovalores, representada por $\frac{(\lambda_1+\lambda_2)}{\Sigma\,\lambda_i}$ onde λ_i representa os autovalores em ordem decrescente. Essa relação desempenha um papel crucial na interpretação das componentes principais e na avaliação da representatividade dos resultados em relação aos dados originais.

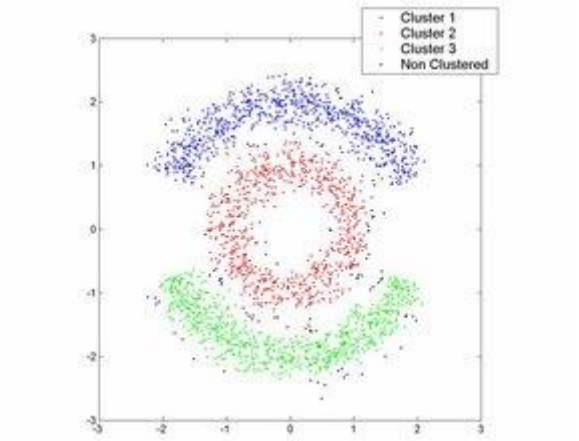

Figura 4: Imagem ilustrativa de clustering superparamagnético de spike sorting

RESULTADOS E DISCUSSÃO:

Os resultados da técnica de PCA e clustering podem ser visualizados nos gráficos apresentados. Os dados do IBGE alcançaram um índice de representatividade de 99,2%, o que indica que a maior parte das relações de produção dos objetos de interesse foi capturada pela análise. Além disso, a representação gráfica assemelha-se a uma reta, o que sugere uma relação linear entre as variáveis estudadas. Essa tendência linear é bastante significativa e abrange quase todas as relações de produção.

Os clusters obtidos refletem dois períodos oscilantes na produção: um primeiro período, abrangendo de dezembro até abril, e um segundo período, de março até outubro. Dezembro até abril representa um período de crescimento na produção de biocombustíveis, enquanto março até outubro representa um período de declínio na produção de siderurgia. O mês de novembro atua como um mês de transição, sendo representado pelos três pontos centrais no

gráfico. Essa divisão em clusters destaca a sazonalidade e a flutuação da produção ao longo do ano, revelando padrões importantes que podem ser úteis para a compreensão e tomada de decisões em contextos específicos.

Em relação aos dados de Spike sorting, foi aplicada a mesma técnica, mas os resultados apresentaram uma representatividade de apenas 90%. Os dados se mostraram extremamente espalhados e dispersos, tornando difícil aplicar o clustering de forma adequada. Diante dessa dificuldade, uma tentativa foi feita para contornar o problema, utilizando um filtro de thresholding nos dados. Embora essa abordagem tenha melhorado os resultados, ainda não foi suficiente para obter informações significativas sobre os dados.

Essa análise ressalta os desafios específicos associados ao problema de Spike sorting e a complexidade inerente ao tratamento desses dados. Apesar das tentativas de melhorar os resultados, a natureza peculiar dos dados e a alta dispersão tornam a análise mais desafiadora. Essa experiência aponta para a necessidade de buscar abordagens mais avançadas e estratégias adicionais para lidar com esses tipos de dados complexos, o que pode requerer a incorporação de outros métodos e técnicas de processamento e análise.

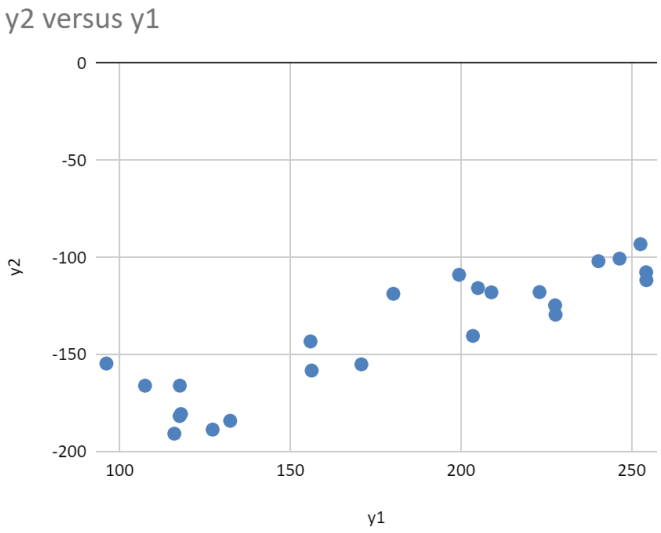

Figura 5: Resultados Obtidos dos dados IBGE

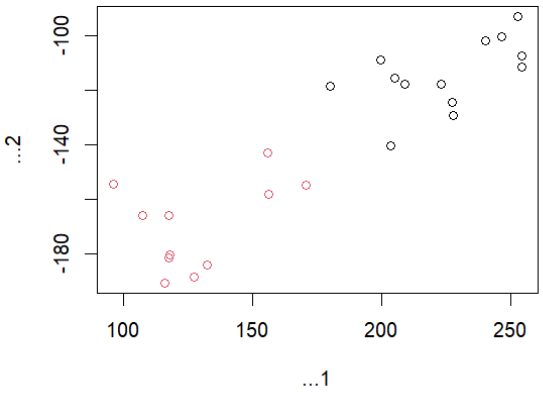

Figura 6: Dados IBGE com a divisão de clustering

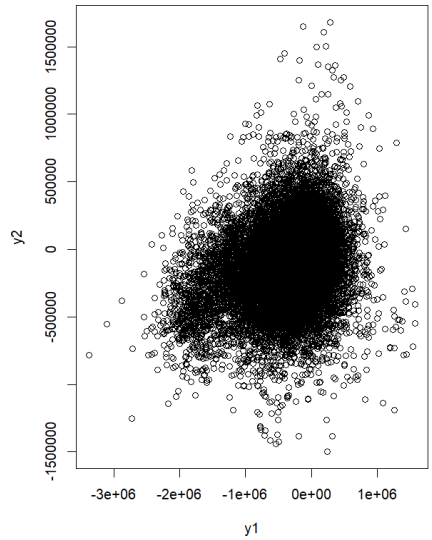

Figura 7: Dados Spike Sorting sem filtragem

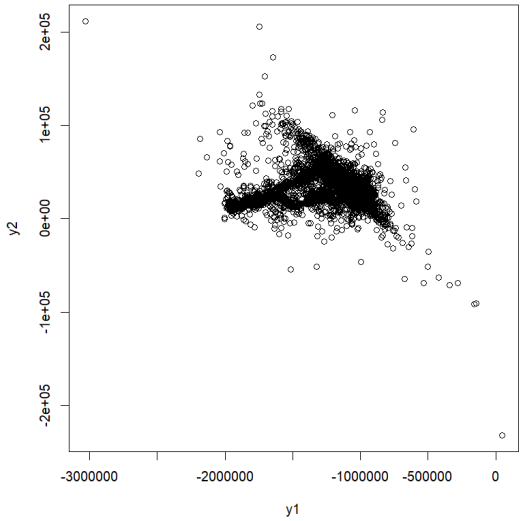

Figura 8: Dados Spike Sorting com Filtragem

CONCLUSÕES:

A pesquisa alcançou com sucesso seus objetivos ao explorar as relações entre álgebra linear e estatística: envolveu uma sólida base teórica, além da implementação prática dos conceitos através da codificação em RStudio. A abordagem permitiu a exploração de aplicações em pesquisas de vanguarda, que se mostraram desafiadoras devido à natureza complexa dos problemas investigados.

BIBLIOGRAFIA

[1]Multivariate Statistical Analysis, Sexta Edição, Prentice-Hall, Nova Jersey, 1998.

[2]COELHO, Flávio; LOURENÇO, Mary. Um Curso de Álgebra Linear. 2. ed. rev. e aum. São Paulo: Editora da Universidade de São Paulo, 2005. 261 p. ISBN 978-85-314-0594-5.

[3]UNSUPERVISED Spike Detection and Sorting with Wavelets and Superparamagnetic Clustering. Neural Computation, Massachusetts Institute of Technology, p. 1661-1687, 30 jan. 2004.

[4]https://sidra.ibge.gov.br/home/ipca/brasil Acesso em: julho de 2019.