Redução da complexidade de dados financeiros: aplicação da Análise dos Componentes Principais (PCA) em conjuntos de dados de ações negociadas na Bolsa de Valores do Brasil (B3) no período de 2018 a 2023

Flavia do Valle
Thalita Routh

Abstract

A análise minuciosa de investimentos desempenha um papel fundamental na tomada de decisões no mercado de ações. Com os avanços da inteligência artificial e das ferramentas de análise de dados, o uso de tecnologia avançada na análise de papéis torna-se cada vez mais relevante no mercado financeiro. Neste estudo, utilizou-se o método de Análise dos Componentes Principais (PCA) em Python para reduzir a dimensionalidade das variáveis em um conjunto de dados, transformando-as em componentes principais. Os resultados obtidos demonstraram uma redução no tempo de processamento e uma otimização dos modelos preditivos, evidenciando o potencial da metodologia proposta no suporte a pesquisas financeiras e análises de investimentos, bem como na seleção criteriosa de ativos diversificados que contribuam para a mitigação de riscos.

Palavras-chave: PCA; ações; variância; covariância; vetor.

]

Reducing the complexity of financial data: Application of Principal Component Analysis (PCA) on datasets of stocks traded on the Brazilian Stock Exchange (B3) from 2018 to 2023.

Abstract

Thorough investment analysis plays a crucial role in decision-making in the stock market. With advancements in artificial intelligence and data analysis tools, the use of advanced technology in analyzing stocks becomes increasingly relevant in the financial market. In this study, the Principal Component Analysis (PCA) method was employed in Python to reduce the dimensionality of variables in a dataset, transforming them into principal components. The obtained results demonstrated a reduction in processing time and optimization of predictive models, highlighting the potential of the proposed methodology in supporting financial research and investment analysis, as well as in the careful selection of diversified assets that contribute to risk mitigation.

Keywords: PCA; stocks; variance; covariance; vector.

1 Introdução

O mercado de ações representa uma das formas mais relevantes de investimento adotadas em escala global. No entanto, sua análise é considerada um desafio constante no universo financeiro, especialmente devido à imensa quantidade de dados disponíveis e à velocidade crescente das informações que, por sua vez, demanda criteriosa avaliação de diversos fatores, como conjuntura econômica, histórico do setor e das empresas pautadas, margens líquidas e brutas, liquidez média diária das ações, entre outros importantes indicadores de potencial retorno de investimentos.

Com o intuito de alcançar uma gestão de riscos eficiente, especialistas no âmbito financeiro enfatizam fortemente a adoção de uma estratégia que envolva a composição de uma carteira de investimentos diversificada. Contudo, é imperativo salientar que a efetividade dessa estratégia pode ser comprometida caso sejam empregados métodos de análise desatualizados, os quais acarretem tempos de processamento prolongados. Nesse contexto, à medida que os avanços nas tecnologias de inteligência artificial progridem, constata-se um crescimento notável no emprego de abordagens metodológicas sofisticadas para a análise de dados financeiros. Essa prática vem se tornando cada vez mais comum no contexto do processo decisório de investimentos, uma vez que possibilita uma análise mais precisa e ágil. (HEUVEL; KAPTEIN, 2022)

Devido à sua capacidade de reduzir a dimensionalidade dos dados, identificar padrões e estruturas subjacentes, bem como investigar as relações entre as variáveis de um conjunto de dados, a análise dos componentes principais (PCA) tem sido amplamente adotada como uma valiosa ferramenta de estatística multivariada na ciência de dados, abrangendo diversos campos de estudo, incluindo finanças. A relevância da utilização da análise dos componentes principais (PCA) na análise de investimentos reside no fato de que essa medida estatística facilita a compreensão das complexas relações existentes entre os dados financeiros, contribuindo de maneira significativa para a tomada de decisões no mercado de investimentos. (SPECTOR et al., 2023).

Na esfera financeira, a aplicação desta abordagem possibilita a seleção de diversos critérios visando a avaliação das correlações e comportamento de ativos com base em suas particularidades, tais como histórico de preços, volatilidade e outros elementos de relevância pertinente. Por exemplo, é factível compreender a correlação e a variância do desempenho de títulos em relação a um período temporal específico ou por meio de um conjunto de métricas financeiras. Ademais, essa estratégia pode revelar-se especialmente benéfica em contextos de mercados mais flutuantes, possibilitando que o investidor diversifique de maneira mais eficaz sua carteira ao encurtar o tempo de processamento dos dados e ao obter acesso às informações tempestivas.

Para realizar a análise dos componentes principais (PCA), foi necessário pré-processar os dados das ações, normalizando-os e construindo um modelo vetorial que representasse adequadamente as características específicas das ações em análise. Nesse sentido, optou-se por utilizar a linguagem de programação Python, cujas funcionalidades permitiram a conversão dos dados das ações em vetores, possibilitando o uso da função PCA. Assim, tornou-se possível empregar modelos vetoriais que proporcionaram a análise de padrões, bem como a exploração de correlações entre diferentes ativos e a identificação de títulos com potencial de desempenho favorável no mercado de ações brasileiro.

Inicialmente, serão expostos os princípios teóricos subjacentes à análise estatística multivariada, enfocando especificamente o método dos componentes principais (PCA). Na sequência, serão delineados os procedimentos metodológicos adotados, os quais englobam a utilização de modelos vetoriais e ferramentas avançadas de análise de dados, com o propósito de identificar padrões, tendências e correlações entre as ações. Os resultados obtidos serão exibidos, enfatizando descobertas relevantes e insights significativos que possam ser empregados em estudos financeiros.

2 Revisão da Literatura

A análise multivariada é uma metodologia estatística que permite examinar de forma conjunta a relação entre múltiplas variáveis. A aplicação desse conceito é valiosa para identificar padrões complexos e explorar as interações não lineares. A análise de componentes principais (PCA), por exemplo, desempenha um papel significativo neste campo estatístico e da ciência de dados, tal qual área multidisciplinar integra conhecimentos e técnicas do ramo da estatística, matemática, ciência da computação e outras especialidades, com o propósito de obter estimativas e identificar tendências em uma base de dados.

Essa abordagem utiliza métodos, como inferência estatística, mineração de dados, a aprendizagem de máquina e os princípios da álgebra linear, para coletar, organizar, analisar e interpretar grandes volumes de dados, visando solucionar problemas complexos e embasar o processo de tomada de decisão em evidências. (HEUVEL; KAPTEIN, 2022, SPECTOR et al., 2023).

No contexto financeiro, a técnica de Análise de Componentes Principais (PCA) desempenha um papel fundamental ao permitir uma síntese do conteúdo informativo presente em conjuntos de dados de alta dimensionalidade, por meio da extração de um conjunto reduzido de variáveis não correlacionadas, denominadas componentes principais, visando simplificar sua representação sem comprometer substancialmente a quantidade de informações contidas nos dados. Tais componentes são obtidas como combinações lineares dos fatores originais, onde o primeiro componente é selecionado com base na sua capacidade de explicar a maior variabilidade presente entre as variáveis.

Os componentes principais subsequentes são calculados de forma a explicar a dispersão restante de maneira ordenada e consecutiva. Dessa forma, o PCA busca capturar gradativamente as informações mais relevantes contidas no conjunto de dados ao preservar a maior parte da variabilidade original. (RAMOVIC; AKERMAN, 2021).

Diante do exposto, quando se trata da aplicação de processos estatísticos baseados em dados no contexto financeiro, destaca-se o seu uso no mercado de ações, que é caracterizado por uma quantidade massiva de informações e flutuações constantes. Nesse contexto, uma das principais vantagens da análise de dados é a capacidade de identificar padrões, tendências e correlações ocultas nos dados financeiros. (ROBERT, 2021)

Com esse propósito, pode-se destacar primeiramente a exploração da variância presente nos dados como uma das abordagens fundamentais, cujo parâmetro é uma métrica de dispersão que quantifica o afastamento dos valores de um conjunto de dados em relação à sua média. No âmbito financeiro, a variância é frequentemente utilizada para avaliar a volatilidade de uma ação ou de uma carteira de investimentos. Uma variância alta indica uma maior flutuação nos retornos, o que implica em um maior risco (ZAGIDULLINA, 2021).

Outra métrica relevante é a covariância, que representa o grau de associação linear entre duas variáveis. A covariância mede a tendência de duas variáveis de se moverem juntas em termos de sua relação linear. No campo financeiro, a covariância entre os retornos de duas ações pode fornecer informações sobre como elas se movem em relação uma à outra. No entanto, a covariância não é facilmente interpretável, pois seu valor depende das unidades das variáveis.

Para contornar esse problema, é comum utilizar a matriz de correlação, que é uma medida padronizada da relação linear entre as variáveis. A matriz de correlação varia de -1 a 1, onde -1 indica uma relação negativa perfeita, 1 indica uma relação positiva perfeita e 0 indica ausência de correlação (ZAGIDULLINA, 2021).

Ao condensar as informações contidas nos conjuntos de dados, tais técnicas contribuem para a tomada de decisões embasadas, viabilizando uma análise mais abrangente dos padrões e tendências do mercado financeiro. A adoção do PCA e a redução de dimensionalidade, em conjunto com outras metodologias estatísticas, conferem aos investidores uma posição competitiva, ao possibilitarem uma análise mais ágil e precisa de grandes volumes de dados.

3 Procedimentos Metodológicos

A metodologia adotada para a realização da presente pesquisa foi baseada na concepção e desenvolvimento de um algoritmo projetado para efetuar o processamento e a estruturação dos dados brutos correspondente ao mercado de ações brasileiro (B3) de 5 anos, abrangendo o período de 2018 a 2022, que totalizou 300 mil empresas. Toda análise pode ser vista no Github, através do link:

https://github.com/thalitarouth

A fim de possibilitar uma análise minuciosa e uma manipulação precisa dessas informações, por meio de uma sequência estruturada de etapas, buscou-se alcançar resultados consistentes e confiáveis, permitindo explorar de maneira profunda os insights contidos nos dados e extrair conhecimentos relevantes para a compreensão do cenário financeiro. Desse modo, os seguintes passos adotados foram os seguintes:

3.1 Leitura do arquivo de dados brutos

O algoritmo realizou a leitura do 2 arquivo fornecido contendo os dados brutos da B3.

Atributo	Descrição
TIPREG	Tipo de registro
DATPRE	Data de pregão
CODNEG	Código de negociação do ativo
NOMRES	Nome resumido da empresa emissora do ativo
ESPECI	Especificação do tipo de ação ou ativo
PREABE	Preço de abertura
PREMAX	Preço máximo
PREMIN	Preço mínimo
PREMED	Preço médio
PREULT	Preço de fechamento
PREOFC	Preço da melhor oferta de compra
PREOFV	Preço da melhor oferta de venda
QUATOT	Quantidade total de títulos negociados
VOLTOT	Volume total de títulos negociados
CODISI	Código ISIN do ativo
DISMES	Número de dias úteis do mês

(1)

3.2 Extração dos dados

Para cada linha do arquivo, o algoritmo extraiu as informações relevantes de cada campo.

```
            # Adiciona os dados à lista
            ordem.append([tipreg, datpre, codneg, nomres, especi, preabe, premax, premin,
                                  premed, preult, preofc, preofv, quatot, voltot, codisi, di
smes])

            # Atualiza a barra de carregamento
            pbar.update(1)

    # Criação do DataFrame com os dados organizados
    df = pd.DataFrame(ordem, columns=['TIPREG', 'DATPRE', 'CODNEG', 'NOMRES', 'ESPECI',
                                                  'PREABE', 'PREMAX', 'PREMIN', 'PREMED', 'PREU
LT',
                                                  'PREOFC', 'PREOFV', 'QUATOT', 'VOLTOT', 'CODI
SI', 'DISMES'])
```

(2)

3.3 Criação do DataFrame

Após o tratamento e organização dos dados, um DataFrame foi criado utilizando a biblioteca pandas do Python. O DataFrame foi estruturado de forma tabular, com as colunas corretamente nomeadas.

```
    df_dict = {
        'TIPREG': 'Tipo de registro',
        'DATPRE': 'Data de pregão',
        'CODNEG': 'Código de negociação do ativo',
        'NOMRES': 'Nome resumido da empresa emissora do ativo',
        'ESPECI': 'Especificação do tipo de ação ou ativo',
        'PREABE': 'Preço de abertura',
        'PREMAX': 'Preço máximo',
        'PREMIN': 'Preço mínimo',
        'PREMED': 'Preço médio',
        'PREULT': 'Preço de fechamento',
        'PREOFC': 'Preço da melhor oferta de compra',
        'PREOFV': 'Preço da melhor oferta de venda',
        'QUATOT': 'Quantidade total de títulos negociados',
        'VOLTOT': 'Volume total de títulos negociados',
        'CODISI': 'Código ISIN do ativo',
        'DISMES': 'Número de dias úteis do mês'
    }

    df.rename(columns=df_dict)
```

(3)

Com o processamento de todas as linhas, o algoritmo utiliza a lista 'ordem' para criar um DataFrame com as colunas corretamente nomeadas. As colunas são renomeadas utilizando um dicionário de mapeamento, o que facilita a compreensão dos dados. O DataFrame resultante é retornado como saída da função.

3.4 Disponibilização do código-fonte

O código-fonte do algoritmo desenvolvido está disponível em um repositório online, permitindo o acesso e utilização por outros pesquisadores interessados na análise de dados da B3.

```
display(md('#### Início do Tratamento...'))
df18 = tratamento('/kaggle/input/b3-data-analyctis/data/COTAHIST_A2018.TXT')
df19 = tratamento('/kaggle/input/b3-data-analyctis/data/COTAHIST_A2019.TXT')
df20 = tratamento('/kaggle/input/b3-data-analyctis/data/COTAHIST_A2020.TXT')
df21 = tratamento('/kaggle/input/b3-data-analyctis/data/COTAHIST_A2021.TXT')
df22 = tratamento('/kaggle/input/b3-data-analyctis/data/COTAHIST_A2022.TXT')

df = pd.concat([df18, df19, df20, df21, df22], axis=0).reset_index(drop=True)

display(md('#### Exportando Dados para Backup...'))
df18.to_csv('COTAHIST_T2018.csv', sep =';')
df19.to_csv('COTAHIST_T2019.csv', sep =';')
df20.to_csv('COTAHIST_T2020.csv', sep =';')
df21.to_csv('COTAHIST_T2021.csv', sep =';')
df22.to_csv('COTAHIST_T2022.csv', sep =';')

df.to_csv('COTAHIST_T1822.csv', sep =';')
display(md('#### Dados Exportados e Prontos para a Análise `df`'))
```

(4)

3.5 Transformação dos dados

As transformações e formatações necessárias incluíram a conversão de formatos de data, tratamento de valores ausentes e ajustes de escala.

```
numeric = ['PREABE', 'PREMAX', 'PREMIN', 'PREMED', 'PREULT', 'PREOFC', 'QUATOT', 'VOLTOT']

dfn18 = df18[numeric]
dfn19 = df19[numeric]
dfn20 = df20[numeric]
dfn21 = df21[numeric]
dfn22 = df22[numeric]

dfn = df[numeric]
```

(5)

3.6 Realização das Análises dos Componentes

A análise e os resultados obtidos neste estudo visaram identificar a presença de correlações lineares próximas entre as ações baseada na aplicação de uma diversificada gama de componentes principais.

Criação da Matriz de Covariância, para isso é necessário a padronização das variáveis para que sejam comparáveis entre si.

```
# Pegando as médias das colunas
mean = np.mean(dfn18, axis=0)

# Padronizando os dados | A matriz subtraido pelo vetor de médias, m x - média
P = dfn18 - np.tile(mean, (dfn18.shape[0], 1))

# Dividindo pelo desvio padrão
M = P / dfn18.std()
M
```

(6)

3.7 Seleção de Colunas Númericas

Para a realização da análise, só é possivel realizar em somente colunas numéricas.

	PREABE	PREMAX	PREMIN	PREMED	PREULT	PREOFC	QUATOT	VOLTOT
0	-0.053389	-0.053906	-0.052894	-0.053416	-0.053416	-0.062176	-0.014712	-0.067713
1	-0.053483	-0.054049	-0.052838	-0.053403	-0.053473	-0.062293	-0.017335	-0.092685
2	-0.027228	-0.027902	-0.026884	-0.027460	-0.027544	0.174361	-0.017317	-0.091840
3	-0.027228	-0.022789	-0.026734	-0.025429	-0.027394	0.171952	-0.017340	-0.092680
4	-0.060156	-0.060748	-0.059515	-0.060101	-0.060095	-0.124586	-0.017308	-0.092656
...	...	...	...	...	...	...	...	...
580150	0.007568	0.006867	0.005753	0.005144	0.007569	0.497873	-0.017211	-0.083678
580151	-0.000243	-0.000750	-0.000180	-0.000267	-0.000060	0.438812	-0.016860	-0.061889
580152	0.004694	0.004058	0.003878	0.004269	0.004632	0.482887	-0.016783	-0.054377
580153	0.001638	0.000942	0.002315	0.001670	0.001633	0.456442	-0.017328	-0.091837
580154	-0.051783	-0.052195	-0.051149	-0.051697	-0.051536	-0.045956	-0.017336	-0.092675

580155 rows × 8 columns

(7)

Foram escolhidas 5 anos de dados da Bolsa de Valores do Brasil (B3), nas quais serão análisadas o comportamente ao longo desses período e após isso a mesma análise será aplicada sobre os anos individualmente.

Na ausência de tal adequação e organização dos dados brutos, conforme evidenciado nos quadros apresentados, a viabilidade e replicabilidade da pesquisa seria comprometida, dificultando a obtenção de resultados confiáveis. Dessa forma,a implementação desse algoritmo foi essencial para possibilitar a aplicação das metodologias mencionadas neste estudo, as quais sao fundamentais para a análise e manipulação de dados.

4 Realização das Análises dos Componentes

A análise e os resultados obtidos neste estudo visaram identificar a presença de correlações lineares próximas entre as ações por meio da utilização de uma abordagem abrangente baseada na aplicação de uma diversificada gama de componentes principais.

4.1 Matriz de Correlação

Através da análise da matriz de correlação foi possível identificar o grau de interdependência linear entre as ações e determinar se existem padrões de comportamento semelhantes ou opostos entre elas, conforme exibido a seguir:

```
# Matriz de Correlação
Cor = M.T.dot(M) / (dfn18.shape[0] - 1)
Cor
```

	PREABE	PREMAX	PREMIN	PREMED	PREULT	PREOFC	QUATOT	VOLTOT
PREABE	1.000000	0.999920	0.999940	0.999960	0.999888	0.102142	-0.000996	0.140601
PREMAX	0.999920	1.000000	0.999850	0.999945	0.999934	0.103195	-0.001006	0.140480
PREMIN	0.999940	0.999850	1.000000	0.999962	0.999931	0.101052	-0.000987	0.140706
PREMED	0.999960	0.999945	0.999962	1.000000	0.999961	0.102036	-0.000996	0.140606
PREULT	0.999888	0.999934	0.999931	0.999961	1.000000	0.102214	-0.000996	0.140592
PREOFC	0.102142	0.103195	0.101052	0.102036	0.102214	1.000000	-0.001797	0.003360
QUATOT	-0.000996	-0.001006	-0.000987	-0.000996	-0.000996	-0.001797	1.000000	0.046614
VOLTOT	0.140601	0.140480	0.140706	0.140606	0.140592	0.003360	0.046614	1.000000

(8)

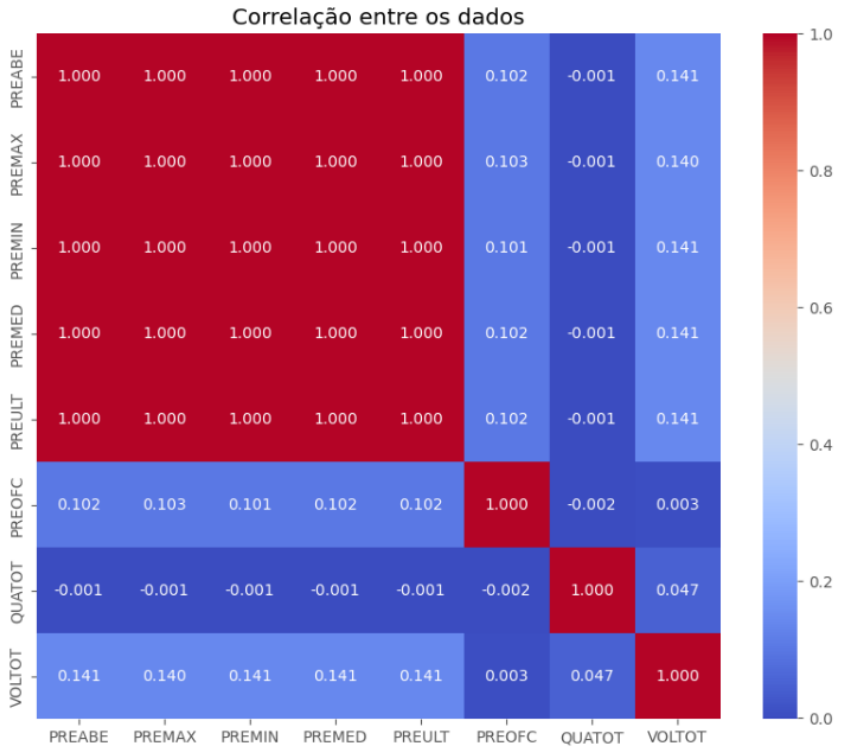
Ao realizar a análise da matriz de correlação entre as ações selecionadas, observou-se que algumas apresentaram correlações positivas significativas, indicando uma tendência de movimento conjunto. Isso sugere que essas ações podem responder de maneira similar a eventos e influências externas, como mudanças no mercado financeiro ou notícias econômicas.

Por outro lado, também foram identificadas correlações negativas entre algumas ações, o que implica em um comportamento oposto entre elas. Isso pode ser resultado de fatores específicos que afetam negativamente uma ação enquanto beneficiam outra, como concorrência direta ou características setoriais distintas.

A análise da matriz de correlação entre as ações permite ainda identificar a presença de ações com correlação próxima a zero, apresentadas em azul na figura abaixo, o que indica baixa interdependência linear entre elas. Essas ações podem ser consideradas como potenciais ativos de diversificação, pois seus movimentos de preço e retorno tendem a ser menos afetados por fatores comuns.

```
plt.figure(figsize=(10, 8))
sns.heatmap(Cor, annot=True, cmap='coolwarm', fmt=".3f", square=True)
plt.title('Correlação entre os dados')
plt.show()
```

(9)



(10)

Tais ações podem ser consideradas como potenciais ativos de diversificação, pois seus movimentos de preço e retorno tendem a ser menos afetados por fatores comuns. Essa análise proporciona uma visão abrangente das relações entre as ações estudadas, auxiliando os investidores a selecionarem ativos com baixa correlação entre si para reduzir o risco e maximizar o potencial de retorno.

É importante ressaltar que a análise da matriz de correlação deve ser complementada por outras técnicas estatísticas e de modelagem para uma compreensão mais completa do comportamento dos ativos financeiros.

4.2 Matriz de Covariância

A análise da matriz de covariância permitiu compreender as relações de variabilidade conjunta entre ações, tornando possivel avaliar não apenas a direção das relações entre os ativos financeiros em questão, mas também a intensidade dessas relações, demonstradas na Tabela a seguir:

```
# Matriz de Covariância | Matriz Transposta
Cov = P.T.dot(P) / (dfn18.shape[0] - 1)
Cov
```

	PREABE	PREMAX	PREMIN	PREMED	PREULT	PREOFC	QUATOT
PREABE	2.561531e+06	2.563285e+06	2.559636e+06	2.561338e+06	2.561370e+06	2.781785e+04	-5.730980e+07
PREMAX	2.563285e+06	2.565450e+06	2.561362e+06	2.563257e+06	2.563447e+06	2.812602e+04	-5.789413e+07
PREMIN	2.559636e+06	2.561362e+06	2.558046e+06	2.559599e+06	2.559737e+06	2.750221e+04	-5.674623e+07
PREMED	2.561338e+06	2.563257e+06	2.559599e+06	2.561348e+06	2.561466e+06	2.778782e+04	-5.728212e+07
PREULT	2.561370e+06	2.563447e+06	2.559737e+06	2.561466e+06	2.561781e+06	2.783873e+04	-5.731649e+07
PREOFC	2.781785e+04	2.812602e+04	2.750221e+04	2.778782e+04	2.783873e+04	2.895573e+04	-1.098817e+07
QUATOT	-5.730980e+07	-5.789413e+07	-5.674623e+07	-5.728212e+07	-5.731649e+07	-1.098817e+07	1.291656e+15
VOLTOT	1.261221e+10	1.261104e+10	1.261305e+10	1.261221e+10	1.261206e+10	3.204118e+07	9.389557e+13

(11)

Ao examinar a matriz de covariância, observa-se que os valores diagonais representam as variâncias individuais de cada ação, indicando a medida de dispersão dos retornos de cada ativo isoladamente. Quanto maior o valor na diagonal principal, maior é a variabilidade dos retornos de uma ação em particular.

Além disso, os valores fora da diagonal principal da matriz de covariância representam as covariâncias entre pares de ações, sendo possível identificar os pares de ações que apresentam maior covariância. Esses valores indicam o grau de associação linear entre os retornos das ações, permitindo avaliar se elas tendem a se mover em conjunto, variando de -1 a 1. O coeficiente de correlação próximo de 1 indica uma forte correlação positiva, enquanto um valor próximo de -1 indica uma forte correlação negativa. A covariância próxima de zero indica uma relação fraca ou inexistente entre os retornos das ações.

A partir dos valores da matriz de covariância, é possível calcular a volatilidade da carteira e, por meio de técnicas de otimização, encontrar a combinação ideal de ativos que maximiza o retorno esperado para um determinado nível de risco. No entanto, por assumir uma relação linear entre os retornos das ações, a análise da matriz de covariância pode não ser válida em todos os casos, sendo necessario combinar essa análise com outras técnicas estatiticas para uma compreensão mais abrangente.

4.3 Busca dos Autovalores e Vetores da Matriz

A busca dos autovetores e autovalores da matriz de covariância foi realizada, permitindo compreender a variabilidade dos dados e sua distribuição acumulada. O calculo da variância explicada de cada autovalor é dada por:

```
autovalores, autovetores = np.linalg.eig(Cov)
display(md(f'### · 8 Autovalores\n> {autovalores}\n\n### · Autovetores\n> {autovetores}\n'))

auto = [(np.abs(autovalores[i]), autovetores[:, i]) for i in range(len(autovalores))]
auto.sort()
auto.reverse()

# Organizando os Autovalores e os Autovetores do menor para o maior
pd.DataFrame(auto, columns=['Autovalor', 'Autovetor'])
```

(12)

	Autovalor	Autovetor
0	3.146032e+15	[4.0028735775089105e-06, 4.002494546244071e-06...
1	1.286901e+15	[-5.400825846056778e-07, -5.4049030344817e-07,...
2	1.255382e+07	[-0.44719839264737804, -0.44754277393959097, -...
3	2.865544e+04	[-0.0020915672132915857, 0.008077178935381934,...
4	3.884625e+02	[-0.21064737787115395, 0.7097710538311673, -0....
5	2.773475e+02	[-0.6833018362336096, -0.20543387270716235, 0....
6	3.882083e+01	[-0.320365387438038, -0.0745767189795413, -0.1...
7	1.395408e+01	[0.4313932414231744, -0.4980978547324639, -0.5...

(13)

Os autovalores foram ordenados, indicando a quantidade de variabilidade presente nos dados, ou seja, a proporção de variância que a primeira variável acumula, cujos dados estão ordenados pelos autovalores.

A análise da soma dos autovalores e da porcentagem de variância acumulada da primeira variável fornece informações essenciais sobre a estrutura de variabilidade das ações, assim como a visualização da porcentagem de variância acumulada da é um indicador importante para compreender o quão representativa essa variável é em relação à variabilidade total dos dados.

```
# Soma  dos Autovalores
total_sum = sum(autovalores)
display(md(f'### Soma dos Autovalores\n> {total_sum}'))

# Visualização da porcentagem de variância dos dados totais que a primeira variavel acumula
var = [(i / total_sum) * 100 for i in sorted(autovalores, reverse=True)]
display(md(f'### Visualização da porcentagem de variância dos dados totais que a primeira variavel a
cumula\n> {[round(vars, 2) for vars in var]}')) # Pequena Amostra
```

(14)

No caso apresentado, a vizualização da variancia dos dados totais que a primeira variavel acumula é representada por [72.09, 27.91, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0], isto é, 72.09% da variância total dos dados, é indicativo de que essa variável é responsável por capturar uma parcela significativa da variabilidade presente nos dados analisados.

No que se refere a soma dos autovalores, neste caso, igual a 6661507990521419.0, representa a quantidade de variabilidade capturada pelos componentes principais utilizados na análise em relação a primeira variável.

A partir dessas informações, foi possível identificar quais componentes são mais relevantes na análise e descartar aqueles que possuem uma contribuição insignificante para a variabilidade total.

4.4 Aplicação da Análise dos Componentes Principais (PCA)

Em seguida, os valores projetados nos componentes principais são adicionados ao DataFrame original, enriquecendo-o com as informações relevantes para análise posterior.

```
# Soma acumulativa
acum_var = np.cumsum(var)

# Criando um DataFrame para a visualização
comp = ['PCA %s' %i for i in range(1, len(autovalores) +1)]

# Criação do DataFrame para facilitar a Visualização
pca = pd.DataFrame({'Componente':comp, 'Autovalor (λ)':autovalores, 'Variância % (fi)':var, 'Freq. A
cumulada (Fi)':acum_var}).set_index('Componente')
display(pca)

# Apartir da utilização de 4 componentes, não é possivel a visualização, pois é o Plano R4,
# Então, no minimo é possivel a visualização no Plano R3 com 3 componentes, 93%

fig = plt.figure(figsize=(20,7))
specs = gridspec.GridSpec( nrows=1, ncols=2, figure=fig)

ax1 = fig.add_subplot( specs[ 0 , 0 ] )
ax2 = fig.add_subplot( specs[ 0 , 1 ] )

ax1.semilogy(pca['Variância % (fi)'], '-o',  color='mediumseagreen')
ax1.set_title('Valores Singulares')
ax1.set_xlabel('PCA 1 até PCA 8')
ax1.set_xticks([])

ax2.plot(np.cumsum(pca['Variância % (fi)']) / np.sum(pca['Variância % (fi)']), '-o', color='steelblu
e')
ax2.set_title('Valores singulares: Soma Acumulativa (%)');
ax2.set_xlabel('PCA 1 até PCA 8')
ax2.set_xticks([]);
plt.savefig('plot01.png', dpi=120)
```

(15)

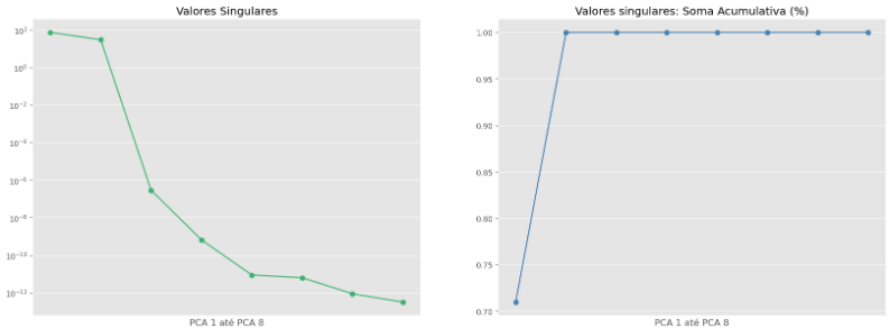
O código apresentado aplica o PCA para reduzir a dimensionalidade dos dados, utilizando autovetores para identificar as direções de maior variabilidade. Por fim, a Tabela resultante, é apresentada de forma visualmente organizada, permitindo uma análise mais clara dos resultados.

	Autovalor (λ)	Variância % (fi)	Freq. Acumulada (Fi)
Componente			
PCA 1	3.146032e+15	7.096954e+01	70.969536
PCA 2	1.286901e+15	2.903046e+01	100.000000
PCA 3	1.255382e+07	2.831944e-07	100.000000
PCA 4	2.865544e+04	6.464216e-10	100.000000
PCA 5	3.884625e+02	8.763101e-12	100.000000
PCA 6	2.773475e+02	6.256524e-12	100.000000
PCA 7	1.395408e+01	8.757368e-13	100.000000
PCA 8	3.882083e+01	3.147821e-13	100.000000

(16)

Em seguida, foram extraídos os autovetores correspondentes aos componentes principais. Esses autovetores são armazenados na matriz A para uso posterior. Daí, a matriz de dados original é multiplicada pela matriz A contendo os autovetores selecionados, cuja operação resultou em uma nova matriz, denominada X. Essa matriz representa os valores projetados nos componentes principais, ou seja, é uma representação reduzida dos dados originais em um espaço de menor dimensionalidade. Para facilitar a interpretação e análise dos resultados, os nomes das colunas da matriz X foram atribuídos às variáveis PCA 1 a PCA 8.

Adicionalmente, gráficos foram gerados para auxiliar na visualização dos resultados.



(17)

O primeiro gráfico exibe a porcentagem de variância explicada por cada componente, permitindo a identificação dos componentes com maior contribuição para a explicação da variabilidade dos dados. Já o segundo gráfico apresenta a soma acumulativa da porcentagem de variância explicada, fornecendo uma perspectiva sobre a quantidade de variância que é explicada à medida que se considera um número crescente de componentes.

Em um primeiro momento, o número de componentes principais é determinado a partir das dimensões da matriz resultante. Isso permitiu compreender a quantidade de informações que será preservada durante o processo de redução dimensional.

Com base na análise realizada, constatou-se que, a partir do quarto componente, não é possível visualizar os dados em um espaço bidimensional ou tridimensional. Isso indica que, para obter uma representação visual adequada dos dados, é necessário considerar no mínimo três componentes, os quais explicam aproximadamente 93% da variância total. Assim, a matriz de dados original é enriquecida com as colunas correspondentes aos componentes principais. Essa etapa é realizada para facilitar a análise posterior dos dados.

Após a criação da matriz dfpc, são selecionadas apenas as observações que possuem códigos de negociação específicos, como 'PETR4', 'BRF SA', 'VALE3', 'BOVB11F' e 'AAPL34'. Essa filtragem permite focar a análise nos ativos financeiros de interesse.

```
n_comp = pca.shape[0]
autovetores = [i[1] for i in auto]
A = autovetores[:3]

target = 'NOMRES'
# Multiplicar a matriz pela matriz com 3 autovetores
X_ = np.dot(df18[[target, 'PREABE', 'PREMAX', 'PREMIN', 'PREMED', 'PREULT', 'PREOFC', 'QUATOT', 'VOL
TOT']].set_index(target), np.array(A).T)
new = pd.DataFrame(X_,columns = ['PCA 1', 'PCA 2', 'PCA 3'])
new[target] = df18[target]
new
```

(18)

	PCA 1	PCA 2	PCA 3	NOMRES
0	1.405975e+06	23430.144548	-20.847485	ALLIAR
1	3.350940e+03	55.614828	-33.212340	ALLIAR
2	5.072758e+04	-1667.416276	-126.196925	APPLE
3	3.636132e+03	-123.035894	-131.951598	APPLE
4	5.031316e+03	946.778832	-9.260591	ABC BRASIL
...	...	...	...	...
580150	5.077801e+05	-21024.225772	-243.978624	FII XP INDL
580151	1.728071e+06	-70167.934685	-208.323105	FII XP LOG
580152	2.148653e+06	-88703.395742	-221.057819	FII XP MALLS
580153	5.084588e+04	-2079.926137	-230.372527	FIP XP OMEGA
580154	3.912988e+03	24.151918	-39.580404	FII TRXE COR

(19)

```
new1 = new.loc[(new['NOMRES'] == 'APPLE') | (new['NOMRES'] == 'BRF SA'), :]
#new1 = new
fig = px.scatter_3d(new1, x='PCA 1', y='PCA 2', z='PCA 3', color='NOMRES',
                    title='Visualização 3D dos PCAs', size_max=2, opacity=0.5)
fig.update_layout(scene=dict(
    xaxis_title='PCA 1',
    yaxis_title='PCA 2',
    zaxis_title='PCA 3'
))
fig.show()
```

(20)

```
sns.set(rc={'figure.figsize':(21,9)})
sns.pairplot( new, vars=['PCA 1', 'PCA 2', 'PCA 3'], hue='CODNEG', height=3, aspect=10/5 )._legend.r
emove()
plt.savefig('plot02.png', dpi=120)
plt.show()
```

(21)

Em resumo, o código aplicou a técnica de componentes principais (PCA) aos dados, projetando as observações nos componentes principais selecionados. Isso resulta em uma representação de menor dimensionalidade, facilitando a análise posterior.

4.5 Criar um dicionário de mapeamento de cores para CODNEG

Para a visualização dos resultados, foi criado um dicionário de mapeamento de cores para as diferentes ações. Em seguida, um gráfico de dispersão foi gerado, destacando as duas primeiras componentes principais (PCA 1 e PCA 2), sendo as cores dos pontos correspondentes às ações mapeadas pelo dicionário.

$\text{codneg}_colors = 'PETR4' :' red','VALE3' :' blue','ITUB4' :' green',$
Adicione mais mapeamentos de CODNEG para cores, se necessário
plt.figure(figsize=(10, 6))
Plot das PCAs com destaque de cor pelo CODNEG plt.scatter(dfpc['PCA 1'], dfpc['PCA 2'], c=dfpc['CODNEG'].map(codneg_colors), $label =' PCA1e2'$)
Configuração do eixo x e y plt.xlabel('PCA 1') plt.ylabel('PCA 2')
Configuração do título e legenda plt.title('Análises dos 2 Maiores Componentes') plt.legend()
Configuração dos ticks do eixo x plt.xticks(rotation=45)
Exibição do gráfico plt.show()

4.6 Distribuição dos Valores Projetados nos Componentes

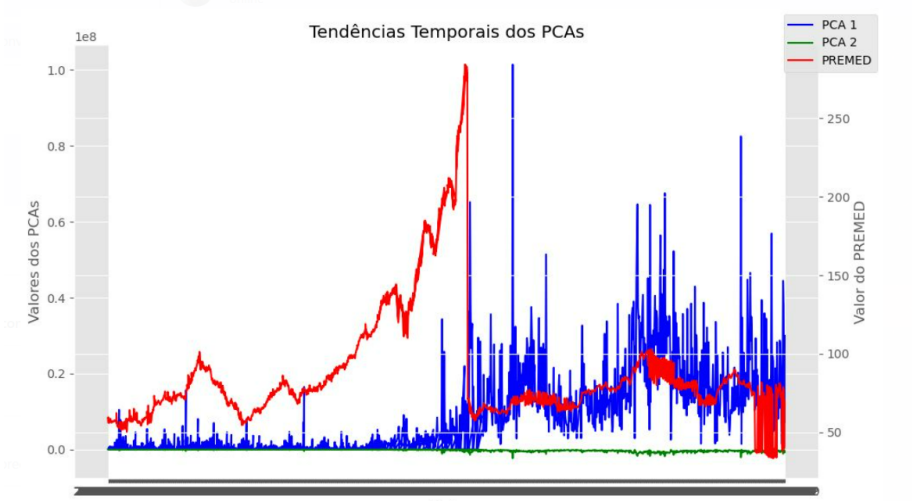
No contexto da análise de componentes principais (PCA), uma dessas análises consiste na identificação dos pontos extremos, ou seja, observações atípicas ou casos de destaque que se destacam em relação aos demais. Outra análise importante é o agrupamento de observações, que busca identificar grupos semelhantes entre os dados. Por meio da PCA, é possível visualizar a proximidade ou semelhança entre as observações, permitindo a identificação de grupos distintos e a compreensão das características que os diferenciam.

```python
import plotly.graph_objects as go

# Seleciona os 3 maiores componentes principais
pca_3d = autovetores[:, :3]

# Cria o gráfico de dispersão tridimensional
fig = go.Figure(data=go.Scatter3d(
    x=pca_3d[:, 0],
    y=pca_3d[:, 1],
    z=pca_3d[:, 2],
    mode='markers',
    marker=dict(
        size=5,
        color=pca_3d[:, 2],  # Colorir com base no componente 3
        colorscale='Viridis',  # Escolher uma escala de cores
        opacity=0.8
    )
))

# Configurações dos eixos
fig.update_layout(
    scene=dict(
        xaxis_title='PCA 1',
        yaxis_title='PCA 2',
        zaxis_title='PCA 3'
    )
)

# Exibe o gráfico 3D
fig.show()
```

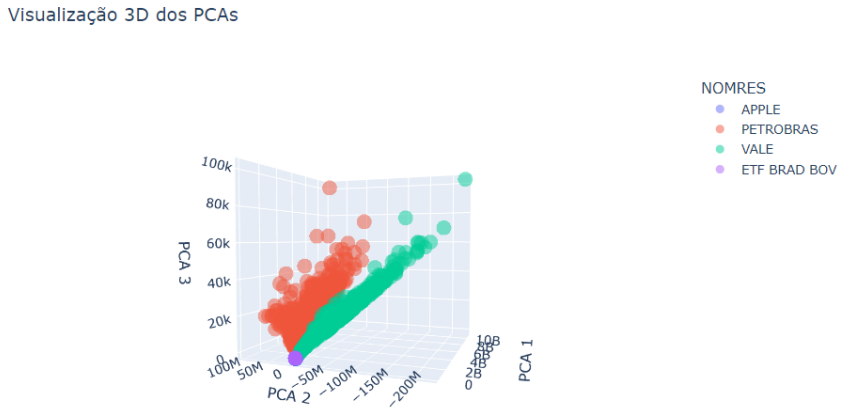
(22)

4.7 Exibição dos gráficos

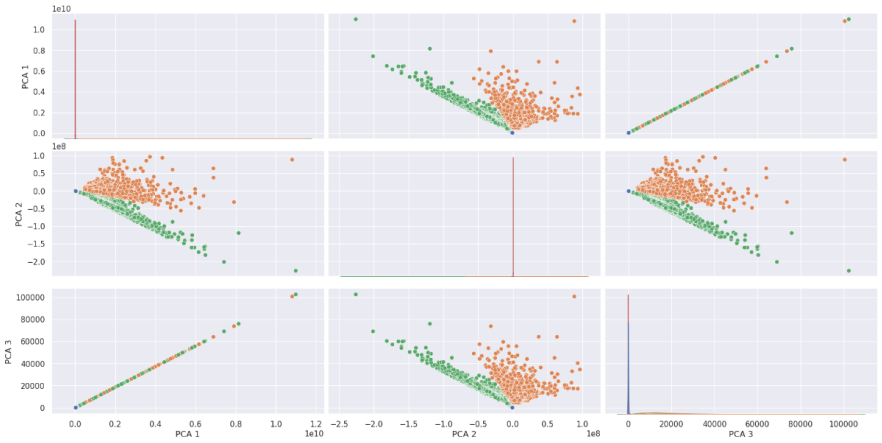
A seguir a demonstração da tendência temporal dos PCAs e a visualização em 3D.



(23)



(24)



(25)

Tabelas demonstrativas dos PCAs:

	Unnamed: 0	TIPREG	DATPRE	CODNEG	NOMRES	ESPECI	PREABE	PREMAX	PREMIN	PREMED	PREULT	PREOFC
0	0	1	2018-01-02	AALR3	ALLIAR	ON NM	14.94	15.16	14.70	14.84	14.89	14.75
1	1	1	2018-01-02	AALR3F	ALLIAR	ON NM	14.79	14.93	14.79	14.86	14.80	14.73
2	2	1	2018-01-02	AAPL34	APPLE	DRN	56.81	56.81	56.30	56.38	56.30	55.00
3	3	1	2018-01-02	AAPL34F	APPLE	DRN	56.81	65.00	56.54	59.63	56.54	54.59
4	4	1	2018-01-02	ABCB2	ABC BRASIL	DIR PRE N2	4.11	4.20	4.11	4.14	4.20	4.13
...	...	...	...	...	...	...	...	...	...	...	...	...
6560327	6560327	1	2022-12-05	BRFS3T	BRF SA	ON NM	8.56	8.57	8.25	8.40	8.26	0.00
6560328	6560328	1	2022-12-06	BRFS3T	BRF SA	ON NM	8.11	8.12	7.99	8.10	8.00	0.00
6560329	6560329	1	2022-12-12	BRFS3T	BRF SA	ON NM	7.48	7.49	7.04	7.16	7.05	0.00
6560330	6560330	1	2022-12-19	BRFS3T	BRF SA	ON NM	6.80	6.95	6.80	6.89	6.95	0.00
6560331	6560331	1	2022-12-21	BRFS3T	BRF SA	ON NM	7.12	7.18	7.04	7.16	7.18	0.00

(26)

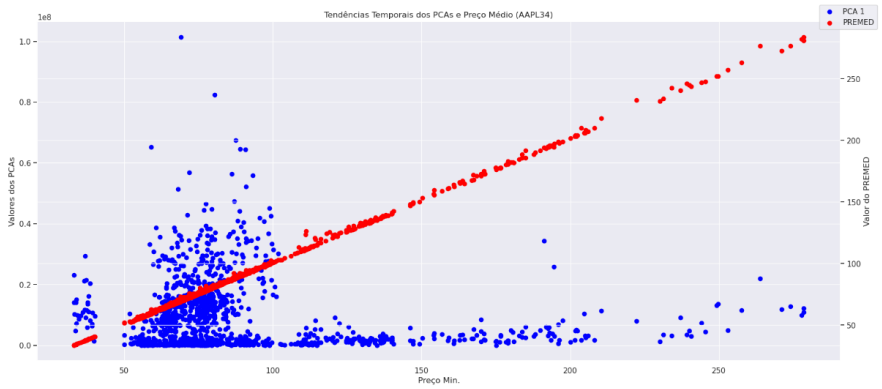
PREOFV	QUATOT	VOLTOT	CODISI	DISMES	PCA 1	PCA 2	PCA 3	PCA 4	PCA 5
14.89	94500	1402991.00	NaN	0	1.405533e+06	42309.362092	-20.347294	-14.634561	0.157752
15.16	225	3343.84	NaN	0	3.349891e+03	100.610773	-33.205931	-14.572894	0.387943
56.30	900	50747.00	NaN	0	5.074540e+04	-986.027286	-126.159872	-54.401617	1.521007
62.00	61	3637.70	NaN	0	3.637456e+03	-74.193887	-131.934878	-54.137362	-4.117890
6.00	1200	4977.00	NaN	0	5.018148e+03	1014.260769	-9.257258	-4.087424	0.063632
...	...	...	...	...	...	...	...	...	...
0.00	20000	168151.75	NaN	0	1.687787e+05	13738.841459	-17.241833	0.076090	-0.172628
0.00	12100	98010.91	NaN	0	9.839279e+04	8450.241811	-17.123079	0.078121	-0.060572
0.00	30300	217028.59	NaN	0	2.180045e+05	22215.806560	-14.188662	0.060200	-0.253197
0.00	26700	183980.88	NaN	0	1.848458e+05	19846.115248	-13.676659	0.062738	-0.107336
0.00	19500	139767.46	NaN	0	1.403954e+05	14293.749595	-14.662332	0.066914	-0.088452

(27)

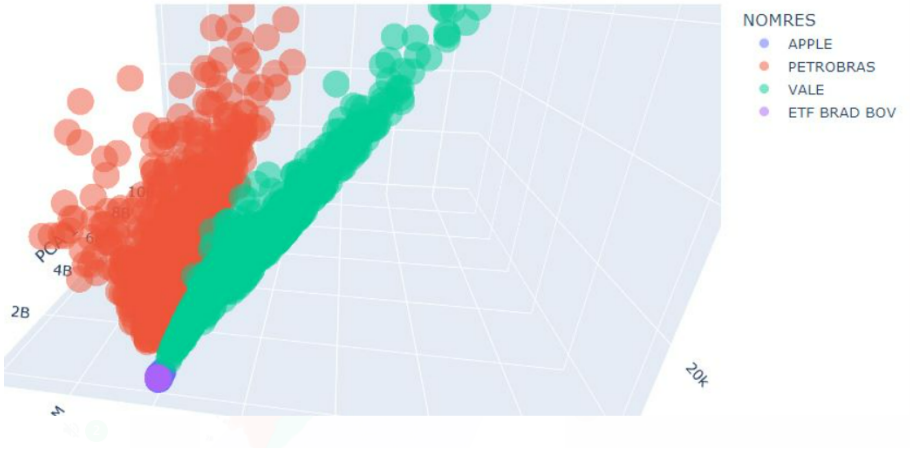
PCA 1	PCA 2	PCA 3	PCA 4	PCA 5	PCA 6	PCA 7	PCA 8
1.405533e+06	42309.362092	-20.347294	-14.634561	0.157752	-0.053222	-0.053989	-0.041167
3.349891e+03	100.610773	-33.205931	-14.572894	0.387943	0.026432	0.057119	-0.061316
5.074540e+04	-986.027286	-126.159872	-54.401617	1.521007	-0.284217	-0.112210	-0.082788
3.637456e+03	-74.193887	-131.934878	-54.137362	-4.117890	-0.979753	2.127213	-3.559108
5.018148e+03	1014.260769	-9.257258	-4.087424	0.063632	0.062467	-0.010138	-0.005511
...	...	...	...	...	...	...	...
1.687787e+05	13738.841459	-17.241833	0.076090	-0.172628	-0.243303	-0.005660	0.000403
9.839279e+04	8450.241811	-17.123079	0.078121	-0.060572	-0.089249	0.039273	0.013651
2.180045e+05	22215.806560	-14.188662	0.060200	-0.253197	-0.349315	-0.085148	-0.024531
1.848458e+05	19846.115248	-13.676659	0.062738	-0.107336	0.089569	0.013002	0.006634
1.403954e+05	14293.749595	-14.662332	0.066914	-0.088452	0.027861	0.013584	0.048320

(28)

Tendência temporal dos PCAs e preço médio



(29)



(30)

5 Conclusão

A análise utilizando componentes principais proporciona uma visão abrangente e multivariada de um conjunto de dados. Cada componente principal é uma combinação linear das variáveis originais. O primeiro componente principal busca explicar a maior variabilidade presente nos dados, enquanto o segundo componente busca explicar a variabilidade restante, e assim por diante até o último componente.

É possível escolher a quantidade de componentes que, quando somados, representam uma porcentagem próxima a cem porcento da variabilidade total dos dados. Através da utilização desses componentes, podemos extrair informações importantes e produzir visualizações de dados em diversos planos. No entanto, é importante destacar que só podemos visualizar até o terceiro plano. Por essa razão, a maioria das análises costuma se limitar a utilizar até quatro componentes.

O algoritmo apresentado neste artigo é uma solução eficiente e organizada para o processamento de dados brutos da bolsa de valores. Ele permite extrair informações relevantes dos campos do arquivo de dados, armazená-las em um DataFrame e renomear as colunas para facilitar a interpretação dos dados. Com o uso de barras de progresso, o algoritmo fornece feedback visual sobre o progresso do processamento dos dados. Essa solução pode ser facilmente aplicada em projetos de análise de dados financeiros.

6 Referências

ROBERT, Tonny. Análise multivariada de dados. Brazil, Editora Senac São Paulo, 2021.

SPECTOR, Alfred Z. et al. Data Science in Context: Foundations, Challenges, Opportunities. 2023.

ZAGIDULLINA, Aygul. High-Dimensional Covariance Matrix Estimation: An Introduction to Random Matrix Theory. Springer Nature, 2021.