Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Dissertação de Mestrado

Análises Multivariadas Aplicadas à Gestão de Investimentos e Alocação de Recursos no Contexto de Entidades Fechadas de Previdência Complementar

por

Louise Barbosa dos Santos Machado

Brasília, 26 de fevereiro de 2026

Análises Multivariadas Aplicadas à Gestão de Investimentos e Alocação de Recursos no Contexto de Entidades Fechadas de Previdência Complementar

por

Louise Barbosa dos Santos Machado

Dissertação apresentada ao Departamento de Estatística da Universidade de Brasília, como requisito parcial para obtenção do título de Mestre em Estatística.

Orientador: Prof. Dr. Gladston Luiz da Silva

Brasília, 26 de fevereiro de 2026

Dissertação submetida ao Programa de Pós-Graduação em Estatística do Departamento de Estatística da Universidade de Brasília como parte dos requisitos para a obtenção do título de Mestre em Estatística.

Texto aprovado por:

Prof. Gladston Luiz da Silva

Orientador, PGEST/UnB

Prof. Pedro Luiz Pizzigatti Corrêa

Poli/USP

Prof. Guilherme Souza Rodrigues

EST/UnB

Prof. Felipe Sousa Quintino

Suplente, EST/UnB

"Busquem, pois, em primeiro lugar o Reino de Deus e a Sua justiça, e todas essas coisas serão acrescentadas a vocês.".

(Mateus 6.33)

Para meu marido, meu filho, meus pais e meu irmão.

Agradecimentos

Agradeço, em primeiro lugar, a Deus por me capacitar e me dar sabedoria para enfrentar todos os desafios.

Ao meu marido por estar ao meu lado na conclusão de mais uma etapa. Por apoiar meus sonhos e desafios, minha carreira profissional e acadêmica, sempre acreditando na minha capacidade e me incentivando a ir mais longe.

Aos meus pais por serem meus exemplos de persistência, estudo e dedicação. Por terem incentivado meus estudos, me proporcionado as melhores condições que poderiam me dar, por me ensinarem o valor do estudo e a batalhar pelo que desejo.

Ao meu irmão e melhor amigo que sempre incentivou minha profissão e meus estudos, me mostrando novas perspectivas e possibilidades de áreas que poderia seguir. Por estar comigo em todos os momentos e ter dividido e comemorado cada etapa e conquista que tive desde nossa infância até o momento.

Ao prof. dr. Gladston Luiz da Silva pela orientação desta dissertação e a todos os professores do departamento de estatística que fizeram parte da minha formação acadêmica na graduação e no mestrado.

Por fim, aos meus amigos que apoiaram toda essa nova jornada e acreditaram na minha capacidade para concluí-la.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

Ao realizar uma gestão de ativos financeiros, deseja-se alcançar o maior retorno possível para um determinado nível de risco. Isso pode ser obtido por meio da diversificação da carteira, buscando alocar os recursos financeiros em classes de ativos que sejam pouco correlacionadas, não correlacionadas ou correlacionadas negativamente. Para auxiliar este objetivo, técnicas estatísticas multivariadas podem ser utilizadas para realizar o agrupamento dos ativos financeiros com base nos seus dados de retorno histórico, identificando padrões de variação comuns entre eles e criando grupos que sejam heterogêneos entre si, mas homogêneos internamente. A Análise de Componentes Principais Funcionais (FPCA) considera que os dados possuem relação temporal e são tratados como funções ao invés de observações pontuais. Dessa forma, essa técnica permite extrair informações a respeito da tendência das séries temporais em estudo. Além disso, técnicas de agrupamento, como método de Ward, *K-means* e modelo de misturas finitas de normais, podem ser utilizados juntamente com a FPCA para criação de grupos de ativos com características semelhantes. A aplicação dessas técnicas foi feita em dados de retornos diários de 2022 a 2025 de ativos de renda fixa, renda variável e fundos imobiliários. Os resultados mostraram que os métodos de agrupamento conseguiram separar os ativos em três grupos, em geral, semelhantes aos seus segmentos, sendo o resultado do modelo de misturas finitas mais condizente com essa divisão. As componentes principais extraíram tendência decrescente no período estudado e um padrão de parábola.

Palavras-chave: técnicas estatísticas multivariadas, dados funcionais, componentes principais, técnicas de agrupamento, gestão de investimento.

Abstract

MULTIVARIATE ANALYSIS APPLIED TO ASSET MANAGEMENT AND RESOURCE ALLOCATION IN THE CONTEXT OF PENSION FUNDS

In financial assets management, the goal is to achieve the highest possible return for a given level of risk. This can be accomplished through portfolio diversification, by allocating financial resources in asset classes that are weakly correlated, uncorrelated or negatively correlated. Multivariate statistical techniques can be employed to group financial assets based on historical return data, by identifying groups that are heterogenous among themselves, but homogenous within themselves. The Functional Principal Components Analysis (FPCA) assumes that the data has temporal correlation and is treated as functions instead of single observations, which allows the extraction of information regarding the tendency of the time series being studied. Furthermore, grouping techniques such as Ward's method, K-means and the finite mixture model of Normal distributions can be used along with FPCA to create asset groups with similar characteristics. These techniques were applied on daily returns from 2022 to 2025 of fixed income, equities and real estate investimento funds assets. The results showed that the clustering methods were generally able to separate the assets into three groups, broadly corresponding to their segments, with the finite mixture model yielding results that were more consistent with this aggregation. The principal components extracted revealed a decreasing trend over the period of study and a parabolic pattern.

Keywords: multivariate statistical methods, functional data, principal components, clustering methods, asset management.

Sumário

Lista de Tabelas

Lista de Figuras

Abreviações e Siglas

AIC	*Akaike Information Criterion*
ALM	*Asset Liability Management*
ANBIMA	Associação Brasileira das Entidades dos Mercados Financeiro e de Capitais
BIC	*Bayesian Information Criterion*
Bovespa	Bolsa de Valores de São Paulo
CMN	Conselho Monetário Nacional
CRI	Certificados de Recebíveis Imobiliários
EFPC	Entidades Fechadas de Previdência Complementar
EM	*Expectation-Maximization*
ETF	*Exchange Traded Funds*
FII	Fundos de Investimento Imobiliário
FPCA	*Functional Principal Components Analysis*
IBOV	Índice Bovespa
ICL	*Integrated Completed Likelihood*
IDA	Índice de Debêntures ANBIMA
IDkA	Índice de Duração Constante
IFIX	Índice de Fundos de Investimentos Imobiliários
IGP-M	Índice Geral de Preços - Mercado
IHFA	Índice de *Hedge Funds* ANBIMA

IMA	Índice de Mercado ANBIMA
IPCA	Índice Nacional de Preços ao Consumidor
LFT	Letras Financeiras do Tesouro ou Tesouro Selic
LTN	Letras do Tesouro Nacional
NTN-B	Notas do Tesouro Nacional - Série B
NTN-C	Notas do Tesouro Nacional - Série C
NTN-F	Notas do Tesouro Nacional - Série F
PCA	*Principal Components Analysis*
SVD	*Single Value Decomposition*
VSO	*Volume-Shape-Orientation*

Capítulo 1

Introdução

1.1 Considerações iniciais

A gestão de ativos é uma área do mercado financeiro voltada para o investimento em diferentes tipos de ativos, como ações, títulos públicos e fundos imobiliários (Union Bancaire Privée, 2025). As entidades fechadas de previdência complementar (EFPC) atuam como gestoras de ativos financeiros dos planos de benefícios e, dessa forma, seguem certas restrições regulamentares e contratuais. Em 2025, a Resolução CMN 5.202/2025, que modificou a Resolução CMN 4.994/2022, tratou, entre outros assuntos, sobre a política de investimentos das EFPC. Foi estabelecido, nessas resoluções, que este documento deveria conter a separação dos investimentos de acordo com os seguintes segmentos de aplicação: renda fixa, renda variável, estruturado, imobiliário, operações com participantes e investimentos no exterior (Conselho Monetário Nacional, 2025).

Esses segmentos são grandes grupos formados por ativos que possuem características semelhantes, porém dentro de cada um ainda pode haver diversos aspectos que podem divergir. De forma geral, além da separação por segmento, as EFPC utilizam a divisão dos ativos em classes, que consistem em grupos menores de instrumentos financeiros que seguem as mesmas normas e regulamentações, respondem de maneira semelhante às condições de mercado e, assim como

os segmentos, possuem características em comum (Union Bancaire Privée, 2025). Essas classes são formadas empiricamente considerando, além dos aspectos citados, estratégias da instituição que facilitem a organização e gestão de seus investimentos.

Assim, as EFPC conseguem diversificar seus investimentos por meio da alocação em diferentes classes de ativos. De acordo com Markowitz (1952), a diversificação do portfólio é uma estratégia fundamental para gerenciar o risco e o retorno dos investimentos. Com ela, é possível reduzir o risco total da carteira sem obrigatoriamente diminuir o retorno esperado, o que é de suma importância para os planos de benefícios das previdências complementares que buscam um portfólio eficiente que consiste em maximizar o retorno dos investimentos condicionado a um determinado nível de risco. Essa redução ocorre devido à correlação entre os ativos que pode ser baixa ou até negativa, implicando que, quando um ativo está em período de valorização, o outro está em desvalorização. Logo, uma carteira diversificada, com recursos em classes de ativos não-correlacionadas, apresenta um retorno esperado maior para um determinado risco do que de uma carteira com menor diversificação (Union Bancaire Privée, 2025).

Como mencionado, uma das maneiras que as EFPC utilizam para gerenciar seus investimentos é por meio das classes de ativos. Isso é chamado comumente de macroalocação e a distribuição dos recursos é feita por meio da ferramenta ALM (*Asset Liability Management*) que, ao final, sugere uma alocação ótima para a carteira de cada plano das entidades (Marques, 2011). Assim, é importante que as classes sejam pouco relacionadas entre si, em geral, e os ativos dentro delas tenham alta correlação e características semelhantes, para que a escolha de um portfólio eficiente satisfaça a teoria de Markowitz (1952), aumentando as perspectivas de rentabilidade dos planos de benefícios sujeito a um nível mais baixo de risco, ao considerar os objetivos e restrições de investimentos de cada um.

Para isso, técnicas estatísticas, como clusterização, podem ser utilizadas para realizar a separação dos ativos financeiros em classes, uma vez que o objetivo dessa ferramenta é criar grupos relevantes. Segundo Bouveyron et al. (2019), os métodos de agrupamento têm se fortalecido para resolver situações em que os conjuntos de dados possuem uma estrutura de *cluster*, mas

que não é identificada pelos humanos de forma natural. Ademais, como o volume de dados está cada vez maior, a identificação de características que auxiliem nesta separação se torna mais desafiadora e algoritmos automatizados podem ser utilizados para substituir a ação humana nesta tarefa.

A metodologia de agrupamento possui diversas aplicações e perspectivas de ampliação de seu uso nos problemas de mercado. Entre elas, pode-se destacar a segmentação de clientes de acordo com seus comportamentos, preferências e necessidades financeiras para que as instituições possam oferecer soluções personalizadas, a detecção de fraude identificando padrões anormais em transações financeiras (LI, 2024) e melhoria na pontuação de avaliação de crédito ao observar características relevantes que influenciam na solvência financeira (Tang, Tian e Wu, 2022). Em relação ao tema de investimentos, pode-se citar como exemplo o trabalho de Duarte e Castro (2020) que trata a respeito do uso do algoritmo k-medoids para separar os ativos correlacionados da Bolsa de Valores (B3) e melhorar a otimização da carteira de ativos.

Juntamente com as técnicas de clusterização, podem ser utilizadas as análises de componentes principais (PCA) que proporcionam, entre outros aspectos, a redução de dimensionalidade e facilidade na interpretação (Johnson e Wichern, 2007). De maneira geral, o método de PCA auxilia na formação de grupos porque, por meio das componentes que são combinações lineares das variáveis, é possível identificar certas relações nos dados que não haviam sido observadas. Além disso, os autores Johnson e Wichern (2007) também destacam que esta técnica é usualmente utilizada em conjunto com as análises de *clusters* como uma ferramenta para se auxiliar no objetivo final da formação dos grupos.

1.2 Justificativa

A segregação dos ativos em classes é feita, geralmente, pelos próprios gestores financeiros de forma empírica, considerando, principalmente, estratégias de investimentos e características dos ativos. Isso pode gerar certas classes de ativos altamente correlacionadas e redundantes,

podendo prejudicar a diversificação do portfólio que é importante para se obter maiores rentabilidades com um menor nível de risco do que uma carteira menos diversificada. Portanto, deseja-se avaliar como técnicas estatísticas multivariadas podem contribuir para a gestão de investimentos por meio da formação de classes de ativos financeiros.

Para estruturar essa avaliação, o presente trabalho utiliza a abordagem da metodologia CRISP-DM (Chapman et al., 1999) que é utilizada em projetos de modelagem e mineração de dados. De maneira geral, esse processo organiza o estudo em fases de entendimento do problema, compreensão e preparação dos dados, modelagem, avaliação e implantação dos resultados, descrevendo as principais tarefas de cada uma dessas etapas. Assim, é possível construir uma sequência clara para o desenvolvimento de modelos estatísticos - neste trabalho, a aplicação de técnicas multivariadas à formação de classes de ativos financeiros - auxiliando no processo de tomada de decisões ao longo das análises e aderência das soluções aos objetivos de negócio.

1.3 Objetivos

1.3.1 Objetivo Geral

O objetivo geral desta dissertação é propor, por meio de análises multivariadas, classes de ativos financeiros que sejam heterogêneas entre si e homogêneas internamente, com o intuito de auxiliar na gestão de investimentos e alocação de recursos no contexto de entidades fechadas de previdência complementar. Para isso, foram aplicados os métodos estatísticos de componentes principais e clusterização, a um conjunto de dados que contém a informação a respeito dos retornos dos ativos financeiros.

1.3.2 Objetivos Específicos

Para atingir o objetivo geral, são definidos os seguintes objetivos específicos:

1. Definir características dos ativos a serem consideradas nas análises.

2. Identificar e selecionar as variáveis para a formação das classes de ativos.

3. Separar os ativos financeiros em classes.

4. Avaliar se as classes propostas são adequadas.

1.4 Estrutura da Dissertação

Esta dissertação está estruturada em quatro capítulos, sendo este o Capítulo 1 referente à introdução. Ele contém a introdução, a justificativa e os objetivos geral e específicos e tem a finalidade de contextualizar a gestão de investimentos em entidades fechadas de previdência complementar, os normativos associados a esta gestão e como os ativos financeiros são segmentados. Além disso, também são introduzidas as técnicas estatísticas que foram utilizadas.

O Capítulo 2 apresenta o referencial teórico relativo à Análise de Componentes Principais, análise de agrupamentos e trabalhos relacionados ao tema. Nele, é apresentada uma revisão de literatura a respeito de cada tema, trazendo referências a artigos e livros consolidados sobre as técnicas, além de artigos que façam a aplicação destas.

O Capítulo 3 descreve o estudo de caso, em que são apresentados o entendimento do negócio, a preparação dos dados e a modelagem proposta. O Capítulo 4 é composto pelos resultados obtidos por meio das técnicas aplicadas, apresentando uma discussão e avaliação delas. Por fim, o Capítulo 5 apresenta as conclusões e proposta de trabalhos futuros.

Capítulo 2

Revisão de Literatura

2.1 Gestão de Investimento

A Gestão de Investimentos é uma área das finanças voltada para a administração de ativos financeiros. De forma geral, ela consiste em realizar investimentos em diferentes tipos de ativos, como ações, fundos imobiliários e títulos públicos, e os gestores atuam segundo restrições regulamentares e contratuais que irão variar conforme a área de atuação da empresa (Union Bancaire Privée, 2025). Além disso, tem como base a Teoria Moderna de Portfólio, proposta por Markowitz (1952), na qual é estabelecido que o risco do portfólio depende das covariâncias entre os ativos.

Em sua teoria, Markowitz (1952) tem como princípio central a diversificação que consiste em investir em diferentes ativos de forma que o risco da carteira seja minimizado, para um determinado nível de retorno esperado, ou o retorno esperado seja maximizado para um mesmo nível de risco. Esse equilíbrio entre risco e retorno deu origem ao conceito de fronteira eficiente (pontos em azul da Figura 2.1 a seguir), na qual o portfólio eficiente é aquele ponto em que uma linha reta tangencia a curva da fronteira eficiente.

Com isso, ele entende que o risco do portfólio não consiste em uma média ponderada dos ativos ou apenas na soma dos riscos individuais, uma vez que a relação entre os ativos desem-

penha um papel fundamental na determinação do risco geral da carteira. A relação entre os ativos pode ser expressa em termos da matriz de covariâncias ou de correlação e essa medida traz benefícios para a diversificação, uma vez que, quanto menor a correlação entre os ativos, menor será o risco.

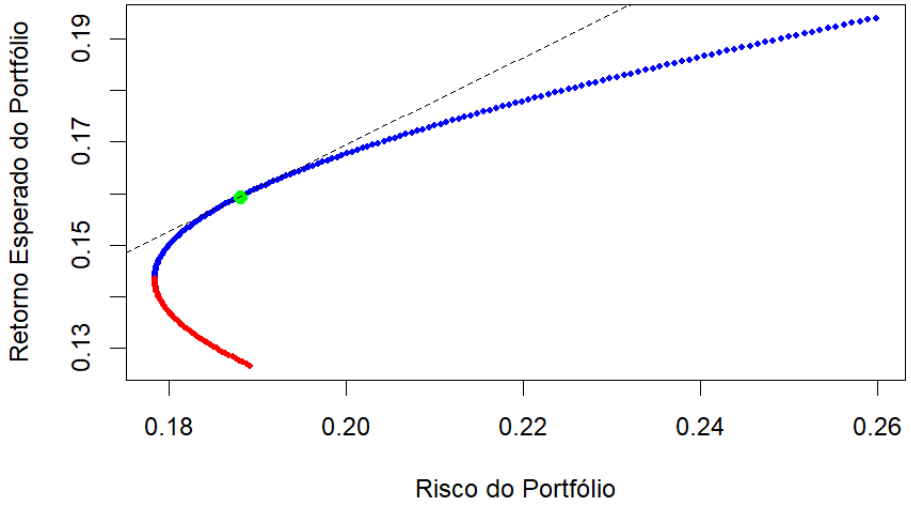

Figura 2.1: Fronteira Eficiente de Markowitz (1952)

Assim, a variância de um portfólio com n ativos pode ser expressa por meio da Equação (2.1) a seguir (Markowitz, 1952):

$$\sigma_p^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \sigma_{ij} \ , \tag{2.1}$$

na qual:

- σ_p^2 é a variância do portfólio e sua raiz quadrada, σ_p, representa o risco - a volatilidade total do portfólio;
- w_i e w_j são os pesos dos ativos na carteira;

- σ_{ij} é a covariância entre os retornos dos ativos i e j.

Como $\sigma_{ij} = Cov(i,j) = \rho_{ij}\sigma_i\sigma_j$, sendo σ_i e σ_j o desvio padrão (volatilidade) do ativo i e j, respectivamente, é possível reescrever a Equação (2.1) em termos da correlação entre dois ativos ρ_{ij}:

$$\sigma_p^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \rho_{ij}\sigma_i\sigma_j \tag{2.2}$$

Com o uso da correlação, pode-se observar de maneira mais simples como esta impacta na diminuição do risco geral da carteira. Bodie, Kane e Marcus (2014) destacam que, quando $\rho_{ij} = 1$, o cálculo da variância do portfólio se resume à média ponderada das variâncias dos ativos. Logo, quanto menor a correlação, menor será a variância e, consequentemente, menor a volatilidade da carteira, mostrando o benefício da diversificação.

É possível olhar para uma determinada carteira por meio de classes de ativos, ao invés de um conjunto de vários ativos individualmente. As classes de ativos representam grupos de ativos financeiros que têm características semelhantes de risco, retorno e comportamento em diferentes condições de mercado, como ações, títulos públicos e *commodities*, e a formação dessas classes é feita pelos gestores de investimentos que se baseiam em critérios empíricos, como natureza dos instrumentos e setor de atividade econômica do emissor (Bodie, Kane e Marcus, 2014).

No contexto das Entidades Fechadas de Previdência Complementar (EFPC), que são operadoras de planos de benefícios complementares à previdência pública (Superintendência Nacional de Previdência Complementar (PREVIC), 2022), existe um conjunto de regulamentos específicos que abordam os tipos de investimentos que podem ser realizados nessas fundações, como a Resolução CMN 5.202/2025 do Conselho Monetário Nacional (2025). Um dos assuntos tratados nessa resolução é a respeito do volume máximo de alocação em cada tipo de segmento de aplicação (renda variável, renda fixa, estruturados, imobiliário, investimento no exterior e operações com participantes), trazendo, assim, restrições que devem ser consideradas na gestão de investimentos de planos de benefícios.

Além disso, a Resolução CMN 5.202/2025 também apresentou algumas mudanças em relação aos tipos de ativos financeiros que as EFPC poderiam investir. A partir desta resolução, fica proibido, de maneira explícita, o investimento em criptoativos e passa a ser autorizado o investimento em debêntures de infraestrutura, FIAGROS e créditos de carbono (Conselho Monetário Nacional, 2025), refletindo a busca por maior diversificação e alinhamento com critérios de sustentabilidade que devem ser seguidos pelas entidades.

Logo, a segmentação de ativos em classes adequadas é relevante para o contexto das EFPC que demandam estratégias de diversificação para o cumprimento dos compromissos atuariais de longo prazo. É preciso, então, evitar classes de ativos altamente correlacionadas entre si para que os princípios estabelecidos por Markowitz (1952) não sejam comprometidos e para evitar concentração de risco em determinados tipos de ativos, o que viola os pressupostos da teoria de otimização de carteiras.

2.2 Análise de Componentes Principais

Existem pesquisas em que a quantidade de variáveis a serem estudadas é grande, o que pode dificultar as análises individuais. Uma alternativa que torna isso viável é utilizar técnicas de redução de dimensionalidade, como as componentes principais, em que são obtidas combinações lineares das variáveis e armazena-se, para serem analisadas, apenas as que possuem variâncias maiores, eliminando as demais (Anderson, 2003). Além disso, essa abordagem também auxilia na interpretação dos resultados e, frequentemente, mostra relações nos dados que não foram identificadas anteriormente (Johnson e Wichern, 2007).

Os primeiros estudos relacionados ao que seria conhecido, então, como Análise de Componentes Principais (do inglês, *Principal Components Analysis* - PCA) foram feitos por Pearson (1901). Em seu artigo, o autor teve como objetivo representar um sistema de pontos em uma linha ou plano que "melhor se ajuste" aos dados, de forma que minimize o erro dessa projeção. Para isso, ele propôs a minimização da soma dos quadrados das distâncias perpendiculares dos

pontos à linha ou ao plano de ajuste, o que era diferente dos métodos de mínimos quadrados até então utilizados. Este critério teve como prerrogativa considerar que todas as variáveis, sejam elas dependentes, sejam independentes, estão sujeitas a um erro experimental.

Além disso, Pearson (1901) introduziu os conceitos de elipsoide de resíduos e de correlação e a relação geométrica entre eles que permite interpretar problemas de ajuste em termos da variação mínima e máxima dos dados - conceito este que foi desenvolvido posteriormente na técnica de PCA. O elipsoide de resíduos representa a dispersão dos desvios perpendiculares ao plano (ou linha) de melhor ajuste e está relacionado à minimização da soma dos quadrados das distâncias perpendiculares. Já o elipsoide de correlação aborda a estrutura de correlação entre as variáveis e seus eixos principais correspondem às combinações lineares de variáveis não-correlacionadas. Logo, Pearson (1901) demonstra que esses elipsoides são geometricamente perpendiculares entre si, o que fundamentou a ortogonalidade das componentes principais introduzidas por Hotelling (1933).

Com base na relação geométrica entre os elipsoides, é possível identificar a linha e o plano de melhor ajuste. A linha de melhor ajuste irá coincidir, em direção, ao maior eixo do elipsoide de correlação e, por isso, corresponde à direção de "variação não-correlacionada" (Pearson, 1901). Essa definição foi formalizada por Hotelling (1933) como a primeira componente principal, representando a direção de maximização da variância explicada dos dados, enquanto a última componente seria o menor eixo do elipsoide de correlação. Por sua vez, o plano de melhor ajuste será perpendicular ao menor eixo do elipsoide de correlação, buscando minimizar a dispersão dos resíduos, o que permite reduzir a dimensionalidade dos dados preservando a maior quantidade de informação (Pearson, 1901).

Hotelling (1933) ampliou o estudo de Pearson (1901) para o caso multivariado e introduziu os conceitos da técnica de Análise de Componentes Principais. O autor tem como objetivo obter variáveis independentes que expliquem grande parte da variabilidade dos dados multivariados. Para isso, ele utiliza como exemplo testes educacionais que mediram a velocidade e poder da leitura, velocidade aritmética e poder aritmético dos estudantes. Com o uso das componentes

principais, foi possível obter duas componentes que explicaram, juntas, 82,5% e foram interpretadas como uma componente de "habilidade geral" e uma que contrasta as habilidades verbais e aritméticas dos estudantes. Assim, o autor demonstra que é possível reduzir as dimensões das análises de quatro variáveis para apenas duas que contemplem a maior variância dos dados.

Essa redução de dimensionalidade é feita por meio da rotação do eixo de coordenadas original para um novo eixo, no qual a primeira componente será o eixo com maior variabilidade e a segunda será o eixo perpendicular à primeira e que contenha a segunda maior variabilidade dos dados (Anderson, 2003). Portanto, esses novos eixos irão representar a direção de maior variabilidade e promovem uma descrição mais parcimoniosa da estrutura da matriz de covariância. Para obter a variabilidade total dos dados, seriam necessárias p componentes, em que p representa o número total de variáveis presentes no conjunto de dados (Johnson e Wichern, 2007). Porém, como citado anteriormente, o método consiste em obter um conjunto de menor dimensão que seja capaz de conter a maior parte da variância total dos dados.

A rotação dos eixos de coordenadas é uma representação geométrica para as combinações lineares de p variáveis $\mathbf{X} = [X_1, ..., X_p]$ (Johnson e Wichern, 2007), sendo essas combinações as componentes principais mencionadas. Como elas dependem da matriz de covariâncias $\boldsymbol{\Sigma}$ ou da matriz de correlação $\boldsymbol{\rho}$, seu desenvolvimento não necessita da suposição de normalidade multivariada dos dados, mas, para os casos em que se deseja fazer inferências com as componentes, é preciso seguir essa suposição.

Assim, as componentes principais são as combinações lineares não-correlacionadas que maximizam a variância $Var(Y_i) = \mathbf{a}_i' \boldsymbol{\Sigma} \mathbf{a}_i$, com $i = 1, 2, \cdots, p$ e $\boldsymbol{\Sigma}$ a matriz de covariâncias de $\mathbf{X}$, e podem ser representadas conforme a Equação (2.3), utilizando a notação de Johnson e Wichern (2007):

$$
\begin{aligned}
Y_1 &= \mathbf{a_1'X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\
Y_2 &= \mathbf{a_2'X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\
&\vdots \\
Y_p &= \mathbf{a_p'X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \,,
\end{aligned} \tag{2.3}
$$

em que:

- $\mathbf{X}$ é o vetor de variáveis originais;
- $\mathbf{a_i}$ é o vetor de coeficientes da i-ésima componente principal.

O processo de obtenção dessas componentes principais segue um procedimento, introduzido por Hotelling (1933), de maximização sequencial da variância sob restrições de normalização e ortogonalidade. A primeira componente (Y_1), por exemplo, é determinada pelo vetor de coeficientes $\mathbf{a_1}$ que maximiza a variância $Var(Y_1) = \mathbf{a}_1'\mathbf{\Sigma a}_1$, sujeita à restrição de normalização $\mathbf{a}_1'\mathbf{a}_1 = 1$ (Johnson e Wichern, 2007). As próximas componentes, $Y_2, \cdots, Y_p$, são obtidas pelo mesmo princípio de maximização da variância, porém sujeito também à restrição de ortogonalidade em que cada componente seguinte Y_j deve ser não-correlacionada com as componentes anteriores Y_i, $\forall i < j$. Isso significa que $Cov(Y_i, Y_j) = \mathbf{a}_i'\mathbf{\Sigma a}_j = 0$, $\forall i \neq j, i < j$ (Johnson e Wichern, 2007).

A restrição de normalização, $\mathbf{a}_i'\mathbf{a}_i = 1$, evita que a variância possa ser aumentada por meio do escalonamento arbitrário dos coeficientes ao multiplicá-los por uma constante $k > 1$, resultando em uma variância k^2 vezes maior. Para eliminar o problema da indeterminação (Johnson e Wichern, 2007), a restrição normaliza o vetor $\mathbf{a_i}$, garantindo que se busque a direção de máxima variabilidade, ao invés de aumento da magnitude dos coeficientes, e mantendo, assim, a interpretação geométrica das componentes no espaço transformado (Anderson, 2003).

Essa solução de maximização da variância, sujeita a restrições, é obtida por meio do método dos multiplicadores de Lagrange, abordado por Hotelling (1933), Anderson (2003) e Jolliffe (2002). Segundo Jolliffe (2002), ele consiste em maximizar:

$$\phi = \mathbf{a}_1' \mathbf{\Sigma} \mathbf{a}_1 - \lambda(\mathbf{a}_1' \mathbf{a}_1 - 1) \,, \tag{2.4}$$

no qual λ corresponde ao multiplicador de Lagrange.

Para maximizar a Equação (2.4), é preciso derivar com relação a $\mathbf{a_1}$ e igualar a zero, conforme apresentado a seguir (Equação 2.5).

$$\frac{\partial \phi}{\partial \mathbf{a}_1} = \mathbf{\Sigma} \mathbf{a}_1 - \lambda \mathbf{a}_1 = 0 \quad \Rightarrow \quad \mathbf{\Sigma} \mathbf{a}_1 = \lambda \mathbf{a}_1 \,. \tag{2.5}$$

A Equação (2.5) mostra que λ é o autovalor da matriz de covariâncias $\mathbf{\Sigma}$ e $\mathbf{a}_1$ é o autovetor associado ao respectivo autovalor. Logo, o maior autovalor λ_1 fornece a máxima variância e o autovetor associado a ele define a direção da primeira componente principal. Além disso, com base no Resultado 8.1 de Johnson e Wichern (2007), a Equação (2.3) pode ser reescrita como:

$$Y_i = \mathbf{e_i'} \mathbf{X} = e_{i1} X_1 + e_{i2} X_2 + \cdots + e_{ip} X_p \,, \tag{2.6}$$

em que $Var(Y_i) = \mathbf{e_i'} \mathbf{\Sigma} \mathbf{e_i} = \lambda_i$, $i = 1, \cdots, p$, e $Cov(Y_i, Y_j) = \mathbf{e_i'} \mathbf{\Sigma} \mathbf{e_j} = 0$, para $i \neq j$.

Algumas propriedades das componentes principais são derivadas da Equação (2.6) e são importantes para a interpretação dos resultados obtidos com esta técnica. A primeira consiste na preservação da variância total dos dados: a soma das variâncias das variáveis originais, σ_{ii}, é igual à soma das variâncias das componentes principais, λ_i, como apresentado na Equação (2.7). Essa igualdade garante o princípio da técnica de PCA citado no início da seção em que a variabilidade total dos dados é mantida durante a transformação e rotação dos eixos (Johnson e Wichern, 2007).

$$\sum_{i=1}^{p} Var(X_i) = \sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} \lambda_i \, . \tag{2.7}$$

A preservação da variância (Equação (2.7)) é fundamentada na decomposição espectral da matriz de covariância que é consequência da Equação (2.8).

$$\mathbf{\Sigma} = \mathbf{A\Lambda A}' \, , \tag{2.8}$$

em que $\mathbf{\Lambda}$ é a matriz diagonal de autovalores, $\lambda_1, \cdots, \lambda_p$, e $\mathbf{A}$ é a matriz ortogonal de autovetores (Hotelling, 1933; Jolliffe, 2002). Essa equação mostra como cada autovalor λ_i quantifica a variância explicada pela i-ésima componente principal, estabelecendo a conexão direta entre a álgebra matricial e a interpretação estatística das componentes (Johnson e Wichern, 2007).

Como consequência da Equação (2.8), Jolliffe (2002) estabelece a decomposição espectral, apresentado na Equação (2.9). Esse resultado permite separar as variâncias combinadas dos elementos de $\mathbf{X}$ em contribuições decrescentes de cada componente principal, além de decompor toda a matriz de covariância. O autor também destaca que a decomposição espectral ajuda a explicar os elementos fora da diagonal principal de $\mathbf{\Sigma}$.

$$\mathbf{\Sigma} = \lambda_1 \mathbf{a_1 a_1'} + \lambda_2 \mathbf{a_2 a_2'} + \cdots + \lambda_p \mathbf{a_p a_p'} \, . \tag{2.9}$$

Uma alternativa à decomposição espectral da matriz de covariâncias é a decomposição em valores singulares (do inglês, *singular value decomposition* - SVD). De acordo com Jolliffe (2002), uma matriz arbitrária $\mathbf{X}$ de dimensão $(n \times p)$, sendo n o número de observações e p a quantidade de variáveis medidas em relação a suas médias, pode ser escrita como (Equação (2.10)):

$$\mathbf{X} = \mathbf{ULA}' \, , \tag{2.10}$$

na qual:

- $\mathbf{U}$, $\mathbf{A}$ são as matrizes ortogonais de dimensão $(n \times r)$ e $(p \times r)$, respectivamente;
- $\mathbf{L}$ é a matriz diagonal de dimensão $(r \times r)$ com os valores singulares;
- r é o *rank* de $\mathbf{X}$.

Destaca-se que os autovetores da matriz de covariância são equivalentes às colunas de $\mathbf{A}$ e os quadrados dos valores singulares (elementos da diagonal de $\mathbf{L}$), divididos por $(n-1)$, fornecem as variâncias das componentes principais. Essa abordagem de decomposição em valores singulares é importante por ser um método computacionalmente eficiente para obtenção das componentes principais e é útil para conjuntos de dados de alta dimensão.

Por meio da Equação (2.7) e sabendo, como consequência da Equação (2.6), que os autovalores de $\mathbf{\Sigma}$ representam a variância de cada componente principal Y_i, $Var(Y_i) = \lambda_i$, pode-se calcular a proporção da variância explicada por cada componente (Equação (2.11)). Com isso, obtém-se também a variância acumulada pelas primeiras q componentes que é um critério fundamental para determinar quantas componentes reter na análise (Jolliffe, 2002).

$$\text{Proporção da variância total da } k\text{-ésima componente principal} = \frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i} \,. \tag{2.11}$$

A decisão de quantas componentes reter na análise deve considerar um equilíbrio entre usar poucas componentes e preservar a variabilidade dos dados originais (Jolliffe, 2002). A proporção acumulada da variância total, representada pela soma das proporções da Equação (2.11), é uma das maneiras utilizadas, na qual se estabelece um limiar mínimo que as primeiras q componentes devem atingir conjuntamente (Jolliffe e Cadima, 2016).

Ademais, pode-se utilizar o critério de Kaiser, conforme apresentado por Jolliffe (2002), especialmente nos casos de obtenção das componentes principais por meio da matriz de correlação. Esse critério sugere reter as componentes cujos autovalores/variâncias são maiores que 1, pois parte da suposição de que, se todos os elementos de $\mathbf{X}$ são independentes, todas as com-

ponentes são iguais às variáveis originais e possuem variâncias iguais a 1. O autor destaca que este critério pode ser conservador em algumas situações ao sugerir poucas componentes quando seriam necessárias mais.

De forma complementar, Johnson e Wichern (2007) apresentam o gráfico *Scree Plot* como uma alternativa visual para a seleção do número de componentes. Ele irá representar os autovalores em ordem decrescente no eixo vertical e o número da respectiva componente no eixo horizontal, conforme Figura 2.2, e a decisão de quantas componentes reter no estudo é feita observando os autovalores que não possuem diferença relevante. No caso apresentado a seguir, considerando apenas o resultado gráfico, a sugestão seria de duas componentes. Os autores recomendam a aplicação conjunta desses critérios de seleção devido às perspectivas complementares que cada um oferece sobre a estrutura dos dados.

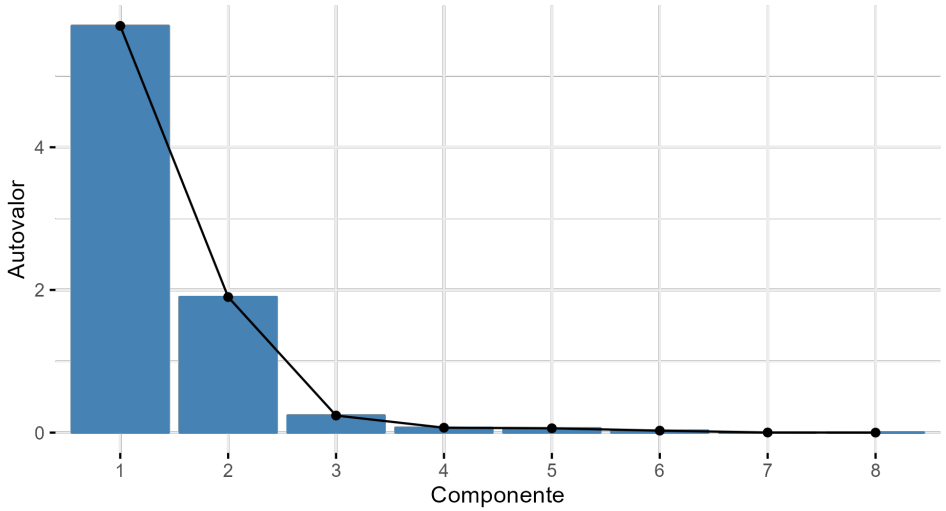

Figura 2.2: *Scree Plot* para seleção do número de componentes

Além desses critérios descritivos apresentados, a literatura também apresenta testes estatísticos que auxiliam na seleção de componentes. Um exemplo é o teste de significância de autovalores que, assim como demais técnicas inferenciais relacionadas à adequação dos dados para análise de componentes principais, não será abordado neste trabalho.

Outro aspecto importante a respeito das componentes principais, está relacionado com os

autovetores $\mathbf{e_i'} = [e_{i1}, e_{i2}, \cdots, e_{ip}]$. De acordo com Johnson e Wichern (2007), cada elemento desse vetor é proporcional ao coeficiente de correlação entre Y_i e X_k (Equação (2.12)) e mede a importância da k-ésima variável na i-ésima componente principal. O coeficiente de correlação, por sua vez, ajuda a interpretar as componentes, porém não indicam a importância da variável X_k na componente Y_i com a presença das demais variáveis; ele irá medir a contribuição individual. Dessa forma, os autores recomendam uma análise conjunta dos coeficientes dos autovetores e das correlações para ajudar na interpretação das componentes principais.

$$\rho_{Y_i,X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}, \tag{2.12}$$

em que $i, k = 1, 2, \cdots, p$.

Ajuste por meio da matriz de correlação

Os resultados apresentados anteriormente consideraram, em geral, a formação das componentes principais por meio da matriz de covariâncias. Porém, devem ser avaliados alguns aspectos do conjunto de dados em análise para escolher entre o uso da matriz de covariância ou de correlação. Johnson e Wichern (2007) e Jolliffe (2002) alertam que utilizar a matriz de covariâncias na PCA é apropriado quando todas as variáveis possuem a mesma unidade de medida e quando as variâncias das variáveis não diferem muito. Além disso, Jolliffe (2002) destaca também que uma vantagem da escolha da matriz de covariâncias é relacionada à inferência das componentes principais populacionais que é mais fácil utilizando a matriz de covariâncias ao invés da matriz de correlação. Como, em geral, a PCA é utilizada de maneira descritiva, essa vantagem não é relevante.

Nesse caso, recomenda-se o uso de variáveis padronizadas e da matriz de correlação e a i-ésima componente é dada conforme Equação (2.13) a seguir:

$$Y_i = \mathbf{e'}_i\mathbf{Z}\,, \tag{2.13}$$

em que $\mathbf{Z}$ é o vetor de variáveis padronizadas obtidas por (Equação (2.14)) e $\mathbf{V}^{\frac{1}{2}}$ é a matriz diagonal dos desvios padrões:

$$\mathbf{Z} = (\mathbf{V}^{\frac{1}{2}})^{-1}(\mathbf{X} - \mu)\,. \tag{2.14}$$

Apesar da diferença na composição das componentes principais, as equações apresentadas anteriormente são semelhantes para o caso das componentes obtidas por meio das variáveis padronizadas. Existe uma ressalva de que a variância de Z_i é unitária, logo:

$$\sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} Var(Z_i) = p \tag{2.15}$$

e

$$\rho_{Y_i,Z_k} = e_{ik}\sqrt{\lambda_i}\,, \tag{2.16}$$

com $i, k = 1, \cdots, p$.

Além disso, como $\sum_{i=1}^{p} Var(Y_i) = p$, então, a Equação (2.11) se resuma a $\frac{\lambda_k}{p}$, sendo λ_k's os autovalores de $\boldsymbol{\rho}$.

2.2.1 Análise de Componentes Principais Amostrais

A aplicação da Análise de Componentes Principais requer, de maneira geral, a estimação por meio de dados amostrais. Na maioria dos problemas estatísticos, os parâmetros populacionais, como a matriz de covariâncias $\boldsymbol{\Sigma}$ e seus respectivos autovalores e autovetores, são desconhecidos e devem ser estimados por meio das estatísticas amostrais correspondentes (Jolliffe e Cadima, 2016). O objetivo, portanto, é construir combinações lineares não-correlacionadas das características observadas que representem a maior parte da variância amostral; essas combinações serão as componentes principais amostrais (Johnson e Wichern, 2007).

A obtenção das componentes amostrais segue o mesmo princípio estabelecido anteriormente

na seção 2.2, porém com o uso das estimativas amostrais. Considera-se uma amostra composta por n observações independentes, $\mathbf{x_1}, \cdots, \mathbf{x_n}$, em que $\mathbf{x}$ representa o vetor de cada observação medida em p variáveis. Esses dados possuem um vetor de médias $\overline{\mathbf{x}}$, matriz de covariâncias $\mathbf{S}$ e matriz de correlação $\mathbf{R}$ (Johnson e Wichern, 2007). Logo, é possível obter uma combinação linear qualquer da seguinte forma (Equação (2.17)):

$$\mathbf{a}_1'\mathbf{x} = a_{11}x_{j1} + a_{12}x_{j2} + \cdots + a_{1p}x_{jp}\,, \tag{2.17}$$

que possui média $\mathbf{a}_1'\overline{\mathbf{x}}$, variância amostral $\mathbf{a}_1'\mathbf{S}\mathbf{a}_1$ e covariância amostral entre duas combinações lineares igual a $\mathbf{a}_1'\mathbf{S}\mathbf{a}_2$, $j = 1, \cdots, n$.

Assim como no caso populacional, as componentes principais amostrais serão aquelas combinações lineares que irão maximizar a variância amostral sujeitas às restrições de normalidade e ortogonalidade. Portanto, a primeira componente será a combinação $\mathbf{a}_1'\mathbf{x}_j$ que maximiza a variância amostral $\mathbf{a}_1'\mathbf{S}\mathbf{a}_1$, sujeita a $\mathbf{a}_1'\mathbf{a}_1 = 1$ (restrição de normalidade). A próxima i-ésima componente será a combinação linear que, além de maximizar a variância amostral $\mathbf{a}_i'\mathbf{S}\mathbf{a}_i$ e estar sujeita à restrição de normalidade $\mathbf{a}_i'\mathbf{a}_i = 1$, deve satisfazer a restrição de ortogonalidade, em que a covariância amostral entre duas componentes deve ser igual a zero: $Cov(\mathbf{a_i'x_j}, \mathbf{a_k'x_j}) = \mathbf{a}_i'\mathbf{S}\mathbf{a}_k = 0$ (Johnson e Wichern, 2007).

Vale ressaltar que, para maximizar a variância amostral, a matriz $\mathbf{S}$ deve ser positiva definida, conforme apresentado por Johnson e Wichern (2007). Com isso, o valor máximo é o maior autovalor $\hat{\lambda}_1$ para a escolha de $\mathbf{a}_1' = \hat{\mathbf{e}}_1$, em que $\hat{\mathbf{e}}_1$ é o autovetor de $\mathbf{S}$. Ainda, os resultados apresentados no caso populacional podem ser obtidos de maneira semelhante para o caso amostral. Sendo assim, tem-se que, se $\mathbf{S}$ é a matriz de covariâncias amostrais com autovalores $\hat{\lambda}_1, \cdots, \hat{\lambda}_p$ e autovetores $\hat{\mathbf{e}}_1, \cdots, \hat{\mathbf{e}}_p$ que fornecem as direções de máxima variabilidade na amostra, a i-ésima componente principal amostral pode ser escrita como (Equação (2.18)):

$$\hat{y}_i = \hat{\mathbf{e}}_i\mathbf{x} = \hat{e}_{i1}x_1 + \hat{e}_{i2}x_2 + \cdots + \hat{e}_{ip}x_p\,, \tag{2.18}$$

com $i = 1, \cdots, p$, $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$, $\hat{\text{Var}}(\hat{y}_k) = \hat{\lambda}_k$, $k = 1, \cdots, p$, e $\hat{\text{Cov}}(\hat{y}_i, \hat{y}_k) = 0, \forall i \neq k$.

Espera-se que os autovalores sejam maiores do que zero para indicar certa importância da componente correspondente. Johnson e Wichern (2007) destacam que, caso uma componente apresente um autovalor associado próximo a zero, existe uma indicação de dependência linear nos dados que não foi percebida antes da análise de componentes principais, mostrando possível redundância das variáveis que devem ser avaliadas para retirada da análise.

Semelhantemente às Equações (2.7) e (2.12), tem-se os seguintes resultados para as componentes principais amostrais:

$$\sum_{i=1}^{p} \hat{\text{Var}}(\hat{y}_i) = \sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p \tag{2.19}$$

e

$$r_{\hat{y}_i, x_k} = \frac{\hat{e}_{ik}\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}, \quad i, k = 1, \cdots p \tag{2.20}$$

Interpretação das Componentes Principais Amostrais

Uma etapa importante da Análise de Componentes Principais (PCA) é a interpretação de seus resultados para atribuir significado à análise. Conforme destacado por Hotelling (1933) e reforçado por Anderson (2003), esse processo envolve a análise conjunta de diversos aspectos, como *loadings*, correlações entre componentes e variáveis, *scores* das observações, proporção explicada pela variância e representações gráficas por meio do *biplot*.

Os *loadings*, ou cargas, são os valores e_{ij} do autovetor $\mathbf{e}_i$ correspondente à i-ésima componente (Equação (2.18)). Eles medem a contribuição relativa de cada variável original para a formação das componentes principais, como já mencionado anteriormente, em que valores absolutos elevados indicam forte influência da variável j na componente i (Jolliffe, 2002) e, ao considerar o sinal da carga, é possível identificar as que possuem associação direta com a componente (sinal positivo) e quais tem relação inversa (sinal negativo) (Johnson e Wichern,

2007). Dessa forma, variáveis que possuem *loadings* próximos a zero não são consideradas importantes para a determinada componente (Jolliffe, 2002).

É importante, juntamente com a análise dos *loadings*, verificar a correlação entre a componente principal e a variável, como apresentado na Equação (2.20). Essa correlação quantifica a associação linear entre cada variável e componente, independentemente das unidades de medida e desconsiderando, também, a presença das demais variáveis (Johnson e Wichern, 2007). Por ser invariante à dimensão das variáveis, ela permite comparar os resultados de duas correlações de variáveis com unidades de medida distintas com a mesma componente principal, podendo indicar, assim, qual está mais relacionada linearmente com a componente. Os autores também destacam que, nos casos em que a correlação é forte, mas os *loadings* são mais baixos, existe a possibilidade de outras variáveis estarem influenciando a variável em questão.

Além disso, essas interpretações numéricas podem ser complementadas por uma representação gráfica por meio do gráfico *biplot*. Proposto por Gabriel (1971), esse gráfico consiste em mostrar as projeções das observações e as direções das variáveis, relacionadas aos *loadings*, nos eixos das duas primeiras componentes principais. Nesta representação, o comprimento do vetor de cada variável será proporcional à contribuição desta para as componentes, refletindo a magnitude do *loading*, o ângulo entre o vetor e o eixo da componente mostra a correlação entre a variável e a componente em questão e o ângulo entre vetores de duas variáveis representa a correlação entre elas (Jolliffe, 2002; Jolliffe e Cadima, 2016).

2.3 Análise de Componentes Principais Funcionais

A Análise de Componentes Principais Funcionais (FPCA) é uma alternativa à forma clássica de PCA para redução de dados aplicada a dados funcionais, em que cada observação é uma função, como curva de crescimento e séries temporais. A Análise de Componentes Principais "clássica" não é apropriada para dados que tenham relação temporal. Porém, com o advento da tecnologia e as grandes bases de dados, as informações têm sido cada vez mais armazenadas

considerando um intervalo de tempo (Wang, Chiou e Müller, 2016). Com isso, surgiu-se a necessidade de adequação das técnicas clássicas de PCA.

2.3.1 Dados Funcionais

Para compreender a Análise de Componentes Principais Funcionais, é necessário entender a definição de dados funcionais. Os dados funcionais são conjuntos de dados em que cada observação corresponde a uma função ao longo de um domínio contínuo. Eles surgem em diversas áreas do conhecimento, como medicina para avaliação de curvas de crescimento, meteorologia para estimar temperaturas ao longo do ano e finanças ao verificar curvas de retornos ou volatilidade de ativos ao longo do tempo. Por exemplo, para analisar a altura de indivíduos, a análise funcional considera toda a curva de crescimento do indivíduo como uma função suave do tempo, ao invés de avaliar a altura apenas em idades fixas (Ramsay e Silverman, 1997).

Nesses conjuntos de dados, cada unidade amostral é representada por uma função contínua $x_i(t)$, na qual $t \in \mathcal{T} \subseteq \mathbb{R}$ é um intervalo contínuo e i indica o indivíduo ou objeto de estudo. Essas funções representam um produto interno $\langle ., . \rangle$ que permite medir distâncias e ângulos entre funções. Na prática, as funções $x_i(t)$ não são observadas em todos os pontos de $\mathcal{T}$, mas em valores específicos t_j, em que $j = 1, \cdots, m_i$ identifica a ordem desses pontos no domínio (Equação (2.21)) (Ramsay e Silverman, 1997).

$$y_{ij} = x_i(t_j) + \epsilon_{ij}, \qquad j = 1, \cdots, m_i\,, \tag{2.21}$$

na qual:

- y_{ij} é o valor observado para o indivíduo/objeto i;
- ϵ_{ij} é o erro de medição (ruído);
- m é o número de observações que pode variar para cada indivíduo/objeto i.

Apesar das funções serem observadas em pontos discretos, como medições diárias ou mensais, pressupõe-se que existe uma estrutura subjacente que pode ser recuperada por técnicas de suavização ou interpolação. Assim, o objetivo inicial da análise de dados funcionais é reconstruir essas funções $x_i(t)$ a partir dos dados observados, utilizando expansão em bases (Ramsay e Silverman, 1997; Jolliffe e Cadima, 2016) que assume que qualquer função $x_i(t)$ pode ser aproximada por uma combinação linear finita de funções base conhecidas (Ramsay e Silverman, 1997):

$$x_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t) \, , \tag{2.22}$$

em que:

- $\phi_k(t)$, $k = 1, \ldots, K$, representa o conjunto de funções base pré-escolhidas;
- c_{ik} são os coeficientes específicos para cada observação i e base k;
- K é o número de funções base (dimensão da aproximação).

A escolha da base impacta diretamente a flexibilidade da representação. As bases de Fourier, $\phi_k(t) = \exp^{i\omega_k t}$, são ideais para dados periódicos, enquanto as B-splines permitem um controle local por meio de nós, adequando-se a tendências não periódicas.

A reconstrução das funções deve preservar a suavidade pressuposta do processo e permitir a avaliação da função em qualquer ponto t do domínio. Para isso, deve-se estimar os coeficientes c_{ik} de forma que a função representada na Equação (2.22) seja contínua, suave e possível de ser avaliada em qualquer t. A estimação dos c_{ik} é feita pela minimização do erro quadrático entre os valores observados e os preditos por meio da expansão da Equação (2.23) a seguir (Ramsay e Silverman, 1997):

$$\min_{\{c_{ik}\}} \sum_{j=1}^{m_i} \left(y_{ij} - \sum_{k=1}^{K} c_{ij}\phi_k(t_j) \right)^2 . \tag{2.23}$$

Após o ajuste dos dados funcionais como funções suaves, é necessário avaliar a variabilidade dos dados. Isso pode ser feito por meio da estrutura de covariância funcional que captura dependências entre os valores da função em diferentes pontos do domínio. Portanto, a função de covariância para dados funcionais é definida como (Jolliffe e Cadima, 2016):

$$S(s,t) = Cov[x(s), x(t)] = \frac{1}{n-1}\sum_{i=1}^{n}[x_i(s) - \overline{x}(s)][x_i(t) - \overline{x}(t)] = \frac{1}{n-1}\sum_{i=1}^{n} x_i^*(s)x_i^*(t), \tag{2.24}$$

em que $\overline{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t)$. Essa função é simétrica, $S(s,t) = S(t,s)$ e não-negativa definida. Os elementos da diagonal principal, $S(t,t)$, correspondem à função de variância que descreve a variabilidade pontual dos dados e os elementos fora da diagonal, $S(s,t)$, revelam os padrões de dependência temporal ou espacial.

Pode-se escrever a Equação (2.22) em termos matriciais, resultando em $\mathbf{x} = \mathbf{C}\phi$. Assim, a função de variância-covariância da Equação (2.24) é equivalente a (Ramsay e Silverman, 1997):

$$v(s,t) = \frac{1}{n}\phi'(s)\mathbf{C}'\mathbf{C}\phi(t)\,, \tag{2.25}$$

com ϕ' sendo o vetor transposto de ϕ. Ramsay e Silverman (1997) destacam que realizar o cálculo da função de variância-covariância utilizando como denominador n ou $n-1$ não tem diferenças essenciais para o contexto de análise de componentes principais.

2.3.2 Componentes Principais Funcionais

Semelhantemente ao caso clássico da PCA, a Análise de Componentes Principais Funcionais também está relacionada à decomposição da estrutura de covariâncias. A FPCA busca as combinações lineares funcionais que capturem os padrões dominantes de variação ao longo do domínio contínuo $\mathcal{T}$ (Ramsay e Silverman, 1997; Wang, Chiou e Müller, 2016). Logo, as componentes principais clássicas, obtidas por meio do somatório das variáveis originais (Equação

(2.3)), são reformuladas no formato de integral por serem, na FPCA, combinações de funções (Equação (2.26)).

$$f_i = \int \beta(s) x_i^*(s) ds = \langle \beta, x_i^* \rangle. \tag{2.26}$$

Para a aplicação da FPCA, considera-se que os dados são centrados na média funcional (Ramsay e Silverman, 1997). Assim, $x_i^*(t)$ é equivalente a $x_i(t) - \overline{x}(t)$, em que $\overline{x}(t)$ é a média funcional calculada por meio da Equação (2.27) a seguir.

$$\overline{x}(t) = \frac{1}{n} \sum_{i=1}^{n} x_i(t) \, . \tag{2.27}$$

A função de covariâncias $S(s,t)$ da Equação (2.24) substitui a matriz de covariâncias presente na PCA clássica. Dessa forma, a equação análoga à Equação (2.5), apresentada a seguir, é uma transformação por integral que tem solução analítica por meio das funções próprias $a(s)$ e que reflete a natureza funcional de $S(s,t)$ e os produtos internos (Jolliffe e Cadima, 2016).

$$\int S(s,t) a(t) dt = \lambda a(s) \, . \tag{2.28}$$

Jolliffe e Cadima (2016) destacam que as funções próprias, $a(t)$, não podem ser determinadas, mas que a solução da Equação (2.22) é uma alternativa destacada pelos autores Ramsay e Silverman (1997). Essa abordagem supõe que qualquer função $x_i(t)$ pode ser escrita como combinações lineares finitas de um conjunto de B funções bases conhecidas, $\phi_1(t), \cdots, \phi_B(t)$, e é vantajosa porque permite incorporar suavização e controlar a complexidade do modelo por meio da escolha de B.

Além da definição das funções base, também é importante considerar a quantidade de componentes principais a serem retidas. Técnicas como proporção da variância explicada e o gráfico *Scree Plot* (Jolliffe, 2002) podem ser utilizadas para auxiliar na definição da quantidade de componentes principais funcionais. Porém, Li, Wang e Carroll (2013) propõem uma aborda-

gem com base em modelos, utilizando critérios de informação, como AIC e BIC, adaptados ao contexto de análise de dados funcionais.

Em relação à interpretação das componentes principais funcionais, Ramsay e Silverman (1997) recomendam a visualização das componentes como perturbações da média funcional. Isso pode ser feito por meio de somas e subtrações de múltiplos da função da componente principal em questão à média e permite identificar padrões de variação associados a cada componente, refletindo como cada uma altera a forma média das funções. De maneira geral, esse método consiste em apresentar em um gráfico (Figura 2.3) a curva média e as curvas das perturbações obtidas pela Equação (2.29).

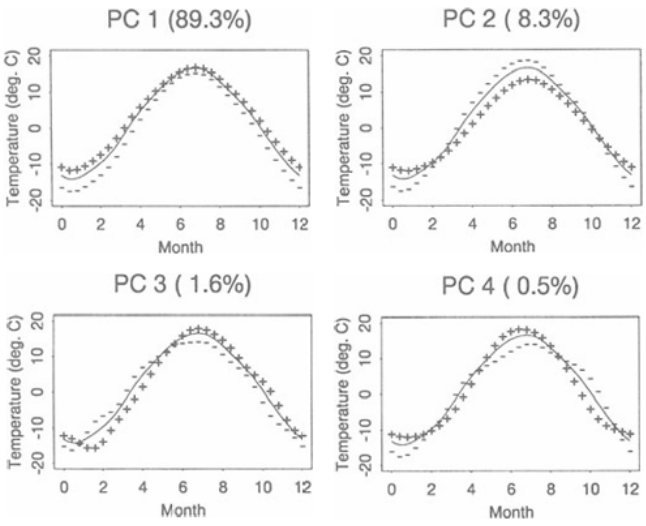

Figura 2.3: Curvas de temperatura média e os efeitos das perturbações para cada componente principal (Ramsay e Silverman, 1997).

$$\overline{x}(t) \pm C\hat{\gamma}_j, \quad \hat{\gamma}_j = j\text{-ésima componente principal.} \tag{2.29}$$

A constante C é obtida por meio da raiz quadrada da Equação (2.30). Ela é definida como a raiz do erro quadrático médio entre a média funcional $\hat{\mu} = \overline{x}(t)$ e a média global $\overline{\mu}$ (Ramsay e Silverman, 1997).

$$C^2 = T^{-1}||\hat{\mu} - \overline{\mu}||^2 = \frac{1}{T}\int(\overline{x}(t) - \overline{\mu})^2 dt, \tag{2.30}$$

com T sendo o comprimento do domínio (intervalo de tempo considerado) e $\overline{\mu} = T^{-1}\int \hat{\mu}(t)dt$ é o valor médio das médias funcionais ao longo de todo o intervalo de tempo. Assim, os autores Ramsay e Silverman (1997) recomendam representar de forma gráfica as médias $\hat{\mu} = \overline{x}(t)$ e $\hat{\mu} \pm 0.2C\hat{\gamma}_j$, em que o valor de 0.2 foi escolhido para facilitar a interpretação dos resultados.

Ademais, a análise dos *scores* funcionais das observações também é uma forma importante de se avaliar os resultados das componentes principais. Os *scores*, representados pela Equação (2.26), indicam quanto da variação associada de cada componente está presente nas funções (observações no formato de dados funcionais). Essa análise é feita por meio de um gráfico (Figura 2.4) composto pelos *scores* das observações nas duas primeiras componentes e, com ele, é possível realizar a identificação de agrupamentos ou tendências específicas ao longo do domínio $\mathcal{T}$.

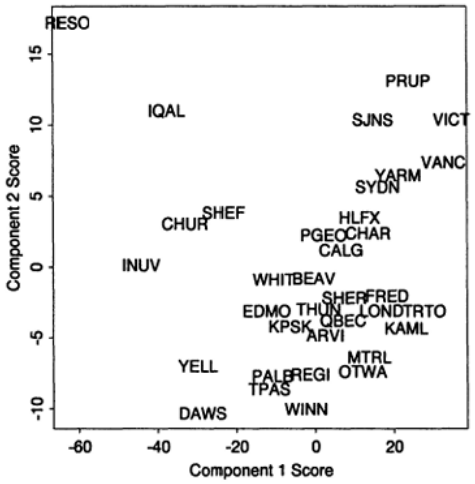

Figura 2.4: *Scores* do clima de cada estação nas duas primeiras componentes principais (Ramsay e Silverman, 1997).

2.4 Análise de Agrupamento

As técnicas de agrupamento têm se fortalecido como uma forma de explorar conjuntos de dados que possuem uma estrutura de grupos e que não são identificadas facilmente (Bouveyron et al., 2019). Com elas, é possível identificar a presença de valores discrepantes nos dados, avaliar a dimensionalidade e identificar possíveis relações que sejam interessantes para o estudo. Portanto, essas técnicas têm como objetivo dividir um conjunto de observações em grupos homogêneos, nos quais as observações dentro do mesmo grupo são semelhantes entre si e distintas das demais de outros grupos (Johnson e Wichern, 2007).

Diferentemente dos métodos de classificação, o agrupamento não possui nenhuma suposição sobre a quantidade ou estrutura dos grupos, mas utiliza medidas de similaridade ou dissimilaridade para formação destes. A medida de dissimilaridade mais comum é a distância euclidiana e, geralmente, é a escolhida para os problemas de agrupamento. Considerando as observações p-dimensionais $\mathbf{x}' = [x_1, x_2, \cdots, x_p]$ e $\mathbf{y}' = [y_1, y_2, \cdots, y_p]$, a distância euclidiana é obtida por meio da Equação (2.31) (Johnson e Wichern, 2007):

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2} = \sqrt{(\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y})}\,. \qquad (2.31)$$

Existem três principais métodos de agrupamento de dados: métodos hierárquicos, não hierárquicos e baseados em modelos estatísticos. Os hierárquicos consistem em agrupar ou separar as observações sucessivamente e podem ser visualizados por meio de dendrogramas. Os métodos não hierárquicos, por sua vez, buscam juntar as observações em uma quantidade pré-definida de grupos, sem que seja necessário determinar a matriz de similaridades, tornando estes algoritmos mais apropriados para grandes conjuntos de dados do que os algoritmos hierárquicos. Por fim, os métodos de agrupamento baseados em modelos incluem modelos estatísticos para identificar grupos por meio de misturas de distribuições (Johnson e Wichern, 2007).

2.4.1 Métodos Hierárquicos

Os métodos hierárquicos são divididos de duas formas: algoritmos aglomerativos e algoritmos divisivos. Os aglomerativos iniciam com cada observação em um grupo distinto e, a cada passo, as observações mais semelhantes são agrupadas, de acordo com um critério de similaridade, até que todas estejam em um único grupo. Já os algoritmos divisivos funcionam de maneira contrária aos aglomerativos: consideram que todas as observações são parte de um único grupo e são divididas em mais grupos utilizando medidas de dissimilaridade, até que cada observação seja um grupo distinto (Johnson e Wichern, 2007).

Os autores destacam os métodos aglomerativos, especificamente aqueles que utilizam critérios de ligação, pois são adequados para agrupamento de observações e de variáveis, o que não é válido para os demais métodos aglomerativos. Para realizar o agrupamento de n observações por meio dos métodos de ligação, é necessário, primeiro, iniciar com uma observação em cada grupo, como qualquer outro método aglomerativo, contendo uma matriz $\mathbf{D}$ de distâncias ou similaridades $n \times n$. Em seguida, são comparadas as medidas para os pares de grupos a fim de identificar quais são as medidas de distâncias ou similaridades mais semelhantes e, então, unir esses grupos. Dessa forma, haverão novas medidas e a matriz de distâncias deve ser atualizada para o novo formato. Esse processo deve ser repetido $n - 1$ vezes a fim de que, ao final, todas as observações sejam parte de um único grupo.

Existem diferentes métodos hierárquicos aglomerativos que utilizam critérios de ligação. Entre eles, pode-se destacar o *Single Linkage*, *Complete Linkage* e *Average Linkage*. Cada um deles utiliza uma medida de comparação dos grupos, sendo, respectivamente, a distância mínima ou "vizinho mais próximo", máxima distância ou "vizinho mais distante" e distância média (Johnson e Wichern, 2007). Além disso, um outro método aglomerativo que pode ser destacado é o Método de Ward que se baseia na minimização da "perda de informação" e será o foco desta dissertação.

Método de Ward

Entre os algoritmos hierárquicos aglomerativos, pode-se destacar o método de Ward, introduzido em 1963. Para Ward (1963, *apud* Johnson e Wichern (2007)), os agrupamentos hierárquicos têm como base a minimização da perda de informação ao juntar dois grupos e, em geral, essa perda é considerada como um aumento na soma de quadrados dos erros (ESS). Para um dado grupo k, ESS_k representa a soma dos quadrados dos desvios de cada observação em relação à média do grupo e, para uma determinada quantidade de grupos K, ESS é definido como a soma dos ESS_k, $k = 1, \cdots, K$, conforme Equação (2.32):

$$ESS = \sum_{k=1}^{K} ESS_k \, . \tag{2.32}$$

Assim, a cada etapa da análise, é considerada a junção de todos os possíveis pares de grupos e aqueles cuja união resultar em um menor aumento da ESS serão agrupados. Inicialmente, por ser um algoritmo aglomerativo em que cada observação faz parte de um grupo, $ESS_k = 0$ e, consequentemente, $ESS = 0$. Porém, ao final do processo, quando todas as N observações estiverem em um único grupo, ESS será dado por (Johnson e Wichern, 2007):

$$ESS = \sum_{j=1}^{N} = (\mathbf{x}_j - \overline{\mathbf{x}})'(\mathbf{x}_j - \overline{\mathbf{x}}) \, , \tag{2.33}$$

em que $\mathbf{x}_j$ é a medida multivariada associada à j-ésima observação e $\overline{\mathbf{x}}$ é a média de todas as observações.

Além disso, para realizar o agrupamento das observações, o método de Ward utiliza um critério de dissimilaridade que tem como base a distância euclidiana entre dois centroides (Kaufman e Rousseeuw, 1990). A dissimilaridade entre dois grupos G_i e G_j é dada pela raiz quadrada da Equação (2.34) a seguir:

$$d^2(G_i, G_j) = \frac{|G_i| \cdot |G_j|}{|G_i| + |G_j|} ||\mu_i - \mu_j||^2 \, . \tag{2.34}$$

2.4.2 Métodos Não Hierárquicos

Diferentemente dos métodos hierárquicos apresentados anteriormente, os não hierárquicos são voltados para agrupamento de observações e não são utilizados para agrupamento de variáveis. Eles também não precisam da construção da matriz de distâncias ou similaridades e, por isso, podem ser aplicados em conjuntos de dados maiores do que no caso das técnicas hierárquicas. Para iniciar o agrupamento, as técnicas não hierárquicas consideram uma divisão inicial das observações em grupos ou um conjunto inicial de "sementes" que formarão o núcleo dos grupos (Johnson e Wichern, 2007).

Seguindo este princípio, pode-se destacar o algoritmo *K-means* (MacQueen, 1967) e a ordenação do centróide mais próximo (Anderson, 1973 *apud* SAS Institute Inc. (2014)) como métodos não hierárquicos. Ambos possuem o mesmo objetivo e consistem em agrupar as observações nos grupos cujas médias sejam mais próximas aos valores observados. No método de ordenação do centróide mais próximo, são selecionadas certas "sementes" dos grupos para representarem as médias de cada grupo e, em seguida, são comparados os valores de cada observação com essas médias. Aqueles valores que forem mais próximos de uma determinada média serão agrupados. O processo termina quando não há mais mudanças nos grupos. A seguir, será detalhado o método *K-means* que é muito conhecido e utilizado.

K-means

Ele foi proposto por MacQueen (1967) e consiste, de forma geral, em agrupar as observações cujos valores sejam mais próximos ao centróide (média) do grupo. O algoritmo inicia selecionando k pontos aleatórios, sendo cada um deles um grupo. Cada novo ponto tem seu valor comparado à média do grupo e, caso o valor seja próximo, ele é agrupado e uma nova média para o grupo é calculada a fim de contemplar a nova informação que foi agregada.

O objetivo principal do algoritmo é minimizar a variância dentro dos grupos, definida pela soma dos quadrados das distâncias entre cada observação e o centróide do grupo que pode ser

visto, também, como minimizar a distância euclidiana (Equação (2.31)) ao quadrado (Johnson e Wichern, 2007). Assim, deseja-se minimizar o resultado da Equação (2.35) a seguir e, para determinar o ponto de parada do algoritmo, é medida a diferença entre a soma dos quadrados de duas iterações consecutivas; caso essa diferença seja menor do que um valor pré-estabelecido, considera-se que não há mais mudanças relevantes na atribuição das observações aos grupos.

$$E = \sum \mathbf{d}_{i,C_i}^2 = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \,, \tag{2.35}$$

em que μ_i é a média do i-ésimo grupo C_i.

O K*-means* se destaca pela simplicidade conceitual, eficiência computacional e capacidade de lidar com grandes conjuntos de dados (Johnson e Wichern, 2007). Além disso, autores como Peng e Müller (2008, *apud* Jolliffe e Cadima (2016)) e Chiou e Li (2007, *apud* Jolliffe e Cadima (2016)) utilizam essa algoritmo juntamente com a FPCA para agrupamento de dados funcionais. Por outro lado, essa técnica apresenta algumas limitações, como a sensibilidade na escolha dos centróides iniciais que pode levar a convergências em ótimos locais e a necessidade de se especificar inicialmente o número de grupos k, pois diferentes escolhas podem levar a resultados muito distintos (Johnson e Wichern, 2007).

2.4.3 Métodos Baseados em Modelos Estatísticos

Métodos hierárquicos e não hierárquicos agrupam objetos por meio de semelhanças entre eles, utilizando, por exemplo, a distância euclidiana como parâmetro para esta união. Com isso, podem ser considerados procedimentos intuitivos (Johnson e Wichern, 2007), porém surgem algumas questões decorrentes da ausência de uma medida de probabilidade associada a eles. Entre elas, pode-se destacar as perguntas apresentadas por Bouveyron et al. (2019), como qual o grau de certeza incerteza associado ao agrupamento proposto e como tratar os valores atípicos (*outliers*).

Nesse contexto, surgiu a alternativa de agrupamento baseado em modelos estatísticos. Esse

método tem como principal forma de modelagem a mistura finita de distribuições e considera que cada grupo é formado por uma distribuição de probabilidade (Bouveyron et al., 2019). Logo, é possível obter inferências associadas à modelagem, verificar o grau de incerteza do agrupamento e obter critérios quantitativos para auxiliar na escolha do número de grupos por meio, por exemplo, de medidas de comparação de modelos, como o BIC. Dessa forma, é possível avaliar, considerando o contexto de cada estudo, quando é mais apropriado utilizar um modelo mais simples, porém com mais grupos, ou um modelo mais complexo com uma quantidade menor de grupos.

A abordagem de modelos de misturas finitas considera que cada observação do conjunto de dados - $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{id})$, $i = 1, \ldots, n$ - é um vetor em $\mathbb{R}^d$, em que d é o número de dimensões (variáveis) presentes no estudo. Elas são formadas por meio de uma mistura finita de distribuições de probabilidades, chamadas de componentes de mistura, sendo sua densidade representada pela Equação (2.36) a seguir (Bouveyron et al., 2019):

$$p(\boldsymbol{y}_i) = \sum_{k=1}^{K} \tau_k f_k(\boldsymbol{y}_i|\boldsymbol{\theta}_k), \tag{2.36}$$

em que:

- K é o número de componentes/grupos;
- τ_k a probabilidade da observação ter sido gerada pelo k-ésimo componente, com $\tau_k \geq 0$ e $\sum_{k=1}^{K} \tau_k = 1$;
- $\boldsymbol{\theta}_k$ o vetor de parâmetros;
- $f_k(\boldsymbol{y}_i|\boldsymbol{\theta}_k)$ é a densidade do k-ésimo componente, dado $\boldsymbol{\theta}_k$.

Assim, cada componente corresponde a um grupo latente na população, transformando o problema de agrupamento em uma inferência sobre os parâmetros $\{\tau_k, \boldsymbol{\theta}_k\}$ e sobre a alocação das observações aos componentes da mistura.

Modelo de Misturas Finitas de Normais

No contexto de dados contínuos, o modelo de misturas mais usual é o de misturas finitas de normais multivariadas (Johnson e Wichern, 2007). Nesse caso, assume-se que o componente k da mistura segue uma distribuição $\mathcal{N}_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ e, portanto, a densidade condicional $f_k(\boldsymbol{y}_i|\boldsymbol{\theta}_k)$ da Equação (2.36) pode ser reescrita conforme Equação (2.37), parametrizada pelo vetor de médias $\boldsymbol{\mu}_g$ e a matriz de covariâncias $\boldsymbol{\Sigma}_g$ (Bouveyron et al., 2019):

$$\phi_k(\boldsymbol{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = |2\pi\boldsymbol{\Sigma}_k|^{\frac{-1}{2}} exp\left\{\frac{-1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu}_k)\right\}. \tag{2.37}$$

A relação entre o modelo de misturas e o agrupamento pode ser apresentada por meio de variáveis latentes $\boldsymbol{z}_i = (z_{i,1}, \ldots, z_{i,K})$ independentes e identicamente distribuídas, em que $z_{i,k} = 1$ se a observação i pertence ao grupo k e $z_{i,k} = 0$ caso contrário. Dessa forma, assume-se que $\boldsymbol{z}_i$ segue uma distribuição multinomial de K categorias com probabilidades $\tau_1, \ldots, \tau_K$ e que $\boldsymbol{y}_i$ é gerada por $\phi_k(.|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ se $z_{ik} = 1$. A partir disso, tem-se a probabilidade condicional estimada, apresentada na Equação (2.38), de que a observação i pertence ao grupo k, quando a verossimilhança observada (Equação (2.39)) é máxima.

$$\hat{z}_{i,k} = \frac{\hat{\tau}_k f_k(y_i|\hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{h=1}^{K} \hat{\tau}_h f_h(y_i|\hat{\mu}_h, \hat{\Sigma}_h)}\ . \tag{2.38}$$

$$\mathcal{L}_O(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n}\sum_{k=1}^{K} \tau_k\ \phi_k(y_i|\mu_k, \Sigma_k). \tag{2.39}$$

Esse resultado pode ser utilizado para classificar as observações em grupos, atribuindo-as ao grupo com maior $\hat{z}_{i,k}$, e para avaliar a incerteza de classificação, utilizando a Equação (2.40) a seguir. Com isso, obtém-se uma forma de quantificar a incerteza associada a cada agrupamento e o modelo apresenta uma maneira de identificar observações com classificação ambígua que podem ser interpretadas como *outliers* ou pontos de fronteira entre grupos.

$$\text{Uncer}_i = 1 - \max_{k=1,\ldots,K} \hat{z}_{i,k} \, . \tag{2.40}$$

Em um modelo de misturas finitas de normais multivariadas, utiliza-se o algoritmo EM (*Expectation-Maximization*) para a estimação dos parâmetros por meio de estimadores de máxima verossimilhança. Ele é indicado para situações em que há a presença de variáveis latentes, como no caso dos indicadores de grupo $z_{i,k}$ que são considerados como dados que "completam" o conjunto de dados e tornam a log-verossimilhança mais simples. Logo, tem-se que a log-verossimilhança dos dados completos é:

$$l_C(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{i,k} \, log[\tau_k \, \phi_k(y_i|\mu_k, \Sigma_k)]. \tag{2.41}$$

O algoritmo EM alterna entre dois passos até a convergência, sendo o passo E de esperança (*expectation*) e o M de maximização. Em problemas de misturas finitas de normais, o passo E calcula as probabilidades a posteriori de pertencimento das observações aos componentes (Equação (2.38)), enquanto o passo M maximiza a log-verossimilhança dos dados completos, apresentada anteriormente, em termos de τ_k e θ_k, considerando $z_{i,k}$ fixo - valor obtido no passo E. Além disso, é possível obter estimativas de probabilidades e médias por meio de fórmulas fechadas utilizando os resultados do passo E. Assim, tem-se que, na iteração s, o passo E do algoritmo para o caso de misturas finitas de normais multivariadas é dado pela Equação (2.42):

$$\hat{z}_{i,k}^{(s)} = \frac{\hat{\tau}_k^{(s-1)} f_k(y_i|\hat{\mu}_k^{(s-1)}, \hat{\Sigma}_k^{(s-1)})}{\sum_{h=1}^{K} \hat{\tau}_h^{(s-1)} f_h(y_i|\hat{\mu}_h^{(s-1)}, \hat{\Sigma}_h^{(s-1)})} \tag{2.42}$$

e as estimativas do passo M são obtidas por meio da Equação (2.43):

$$\hat{\tau}_k^{(s)} = \frac{\hat{n}_k^{(s-1)}}{n} \;\; ; \;\; \hat{\mu}_k^{(s)} = \frac{\sum_{i=1}^{n} \hat{z}_{i,k}^{(s-1)} y_i}{\hat{n}_k^{(s-1)}} \;\; ; \;\; \hat{n}_k^{(s-1)} = \sum_{i=1}^{n} \hat{z}_{i,k}^{(s-1)}. \tag{2.43}$$

As etapas E e M são repetidas até que um critério de convergência seja satisfeito, como a estabilização da log-verossimilhança ou a variação relativa dos parâmetros abaixo de um limiar

predefinido. O algoritmo EM não assegura a convergência para o máximo global, tornando a escolha de valores iniciais um aspecto relevante do procedimento de estimação (Bouveyron et al., 2019).

O processo de estimação do modelo de misturas finitas de normais multivariadas pode passar por certas dificuldades, como problemas na precisão e dificuldade de interpretação dos resultados, devido à quantidade de parâmetros, obtida por meio do resultado $(K-1)+Kd+K[d(d+1)/2]$ (Bouveyron et al., 2019). Com o intuito de corrigir esse problema, utiliza-se a parametrização da matriz de covariâncias $\boldsymbol{\Sigma}_k$ por meio da decomposição espectral, também conhecida como decomposição geométrica ou VSO (*Volume-Shape-Orientation*), apresentada a seguir (Equação (2.44)):

$$\boldsymbol{\Sigma}_k = \boldsymbol{\lambda}_k \boldsymbol{D}_k \boldsymbol{A}_k \boldsymbol{D}_k^T. \tag{2.44}$$

Cada elemento da decomposição da Equação (2.44) representa um aspecto geométrico da matriz de covariâncias, sendo $\boldsymbol{\lambda}_k$ o volume, $\boldsymbol{A}_k$ uma matriz diagonal que controla a forma e $\boldsymbol{D}_k$ a matriz de autovetores que determina a orientação do grupo. Esses elementos possuem restrições que levam a duas opções de modelos univariados e 14 multivariados, permitindo explorar a geometria dos grupos ao equilibrar capacidade de ajuste e complexidade do modelo. Bouveyron et al. (2019) destacam as seguintes restrições para cada componente geométrico de $\boldsymbol{\Sigma}_k$:

- Para o volume ($\boldsymbol{\lambda}_k$): tem-se a alternativa de igualdade (“E") quando os volumes dos grupos são considerados iguais ou variável (“V") caso não haja essa limitação.
- Para o formato ($\boldsymbol{A}_k$): pode-se ter a restrição de igualdade (“E") nos casos em que $\boldsymbol{A}_k \equiv \boldsymbol{A}$ para $k = 1, \ldots, K$, variável (“V") se não forem restritos ou “I"se os grupos forem considerados esféricos - nesse caso, considera-se que $\boldsymbol{A}_k = \boldsymbol{I}$, para $k = 1, \ldots, K$.
- Para a orientação ($\boldsymbol{D}_k$): como no caso do formato, as orientações podem ser consideradas

iguais (“E"), variáveis (“V") ou esféricas (“I"), no qual $\boldsymbol{D}_k = \boldsymbol{I}$, $k = 1, \ldots, K$.

Por fim, assim como em outros casos de modelagem estatística, a escolha do número de componentes K e da parametrização de $\boldsymbol{\Sigma}_k$ para agrupamento das observações também são tratadas como um problema de seleção de modelos e podem utilizar critérios para avaliação e comparação. Os autores Bouveyron et al. (2019) destacam o Critério de Informação Bayesiano (BIC) e o *Integrated Completed Likelihood* (ICL) como métricas apropriadas para este contexto. O BIC de um modelo M_j, apresentado na Equação (2.45), é mais apropriado para os casos em que se deseja estimar o número de componentes do modelo de mistura, em vez do número de grupos existentes no conjunto de dados. Por outro lado, utiliza-se o ICL, da Equação (2.46), para casos em que o objetivo do estudo é o agrupamento em si, tendo um melhor desempenho na escolha do número de grupos, sem ter como objetivo encontrar o melhor modelo de mistura que se ajuste aos dados.

$$BIC_{M_k} = 2 \cdot log\, p(D|\hat{\theta}_{M_j}, M_j) - v_{M_j} log(n), \tag{2.45}$$

com $p(D|\hat{\theta}_{M_j}, M_j)$ sendo a verossimilhança do modelo e v_{M_j} representando o número de parâmetros independentes a serem estimados.

$$ICL = 2 \cdot log\, p(y, z^*|\hat{\theta}_{M_j}, M_j) - v_{M_j} log(n), \tag{2.46}$$

no qual z^* é a estimação do máximo a posteriori (MAP) do agrupamento z que satisfaz: $z^*_{i,k} = 1$ se $\hat{z}_{i,k} = \arg\max_h \hat{z}_{i,h}$ e $z^*_{i,k} = 0$ caso contrário.

Como o objetivo principal do uso do ICL é obter o número de grupos presentes no conjunto de dados, ele é mais indicado em casos em que os grupos estão bem definidos e tende a escolher quantidade de grupos iguais ou menores que o BIC. Isso ocorre devido ao termo de entropia esperada da classificação (Equação (2.48)) presente no cálculo desta métrica que penaliza o BIC e está associado à incerteza da classificação (Equação 2.47). A entropia é mais elevada

quando existe uma maior incerteza da classificação das observações nos grupos e $\hat{z}_{i,k} = \frac{1}{K_{M_j}}$, $\forall k$, e é menor quando $\hat{z}_{i,k}$ é igual a 0 ou 1.

$$ICL = BIC - E(M_j), \tag{2.47}$$

em que:

$$E(M_j) = -\sum_{i=1}^{n}\sum_{k=1}^{K_{M_j}} \hat{z}_{i,k} log(\hat{z}_{i,k}) \tag{2.48}$$

2.5 Trabalhos Correlatos

As técnicas apresentadas anteriormente podem ser utilizadas em diversas áreas de aplicação. A Análise de Componentes Principais Funcionais é utilizada em conjunto com a análise de agrupamento no trabalho de Lin et al. (2015), abordando uma extensão para a FPCA bidimensional e um método de agrupamento com base na seleção aleatória de características para o algoritmo *k-means*. Esse artigo foi aplicado dentro do contexto de imagens médicas para diagnóstico de câncer e busca utilizar essas técnicas como um apoio ao diagnóstico médico para identificação de padrões nos tumores, subtipos de doença e de possíveis respostas diferentes a um tratamento.

Outro artigo que utiliza a abordagem de componentes principais, porém com a extensão FPCA multinível, é de Zablocki et al. (2024). Neste estudo, a FPCA multinível é aplicada ao contexto de avaliação do comportamento sedentário, identificado como um fator de risco para diversas doenças crônicas, de mulheres após período de menopausa. Foram obtidas componentes que indicavam padrões de flutuação no movimento ao sentar (forma como pessoas se movem enquanto estão sentadas) e variação no movimento durante diferentes durações de episódios de sentar (menos de 30 minutos sentado, entre 30 e 39 minutos e acima de 39 minutos). Dessa forma, o estudo mostrou que esses padrões estão associados à saúde cardiovascular das mulheres em análise e observou o sedentarismo como fator de risco cardiovascular.

O método de Ward, juntamente com outras técnicas de agrupamento hierárquicas, foi explorado por Hidayatullah e Sofro (2024) no contexto de gestão de resíduos na Indonésia. Neste trabalho, os autores exploram que políticas únicas de gestão de resíduos não são eficientes e que perfis específicos de resíduos podem orientar intervenções mais direcionadas. Para isso, foram utilizados os métodos hierárquicos de ligação simples, completa e média, e o método de Ward com variáveis que indicam o volume de resíduos de cada tipo, sendo de domicílios, escritórios, mercados, comércio, instalações públicas, áreas regionais e outras fontes. O artigo mostrou, ao final, que o método de Ward foi o mais adequado e gerou 14 grupos que permitiram a identificação de perfis de resíduos em cada região da Indonésia, colaborando para a hipótese inicial de que políticas únicas são ineficientes.

Com o objetivo de comparar o resultado de duas abordagens de pré-processamento para agrupamento de séries temporais, o artigo de Lee, Lin e Stolz (2024) explora o método *K-means* juntamente com a técnica NP-*Free* e de normalização Z (do inglês, *Z-Normalization*). Os autores destacam que o *K-means* é amplamente utilizado devido à simplicidade, porém requer pré-processamentos para transformar escalas dos dados de maneira que se tornem semelhantes. O artigo mostra que as duas abordagens possuem resultados diferentes, sendo a primeira mais demorada, mas preserva as nuances específicas de cada série, trazendo uma melhor precisão na separação dos grupos, enquanto a segunda é mais rápida computacionalmente, porém une séries temporais que são semelhantes superficialmente.

O artigo de Tang, Tian e Wu (2022) apresenta diferentes métodos de agrupamento aplicados ao setor financeiro. Foram abordadas as técnicas de K-*means*, *minimum spanning tree* e agrupamentos hierárquicos, dentro do contexto de análise de crédito, mercado de ações, estratégias de negociação e seleção de portfólio. O artigo destaca, por exemplo, que o agrupamento auxilia na seleção de variáveis para previsão de inadimplência e na tomada de decisão de crédito, ao ser aplicado em um cenário de análise de crédito. Já no contexto das estratégias de negociação, são formados grupos de padrões de preços e volumes que podem revelar ineficiências de mercado e identificar estratégias de negociação mais eficazes.

Em relação à aplicação relacionada à seleção de portfólio, pode-se destacar o trabalho de Duarte e Castro (2020) que propõe uma maneira de criar grupos de ativos por meio da técnica de *k-medoids*. O objetivo é segmentar os ativos, especificamente ações da Bolsa de Valores Brasileira (B3), em grupos de ativos correlacionados e, em seguida, alocar recursos para cada grupo. Dessa forma, o artigo mostra que técnicas de agrupamento podem contribuir na diversificação e gestão de risco por agrupar ativos com padrões de comportamento semelhantes.

Outro artigo que pode ser citado em relação ao tema de análise de portfólio é o de Burca et al. (2021). Ele propõe uma abordagem alternativa ao modelo clássico de Markowitz (1952) por meio do uso de técnicas de mineração de dados. O foco do artigo está no mercado de capitais romeno e são utilizadas técnicas de Análise de Componentes Principais nos retornos semanais e betas para reduzir a colinearidade entre as ações da bolsa e técnicas de agrupamento para seleção de ativos. O resultado final mostra que a metodologia proposta melhora tanto o retorno quanto o perfil de risco do portfólio final comparado aos métodos tradicionais.

Na busca da literatura a respeito de fundos de pensão, não foram encontrados estudos que abordam diretamente o tema de formação de classes de ativos financeiros por meio de técnicas multivariadas. Porém, algumas referências auxiliam no entendimento de negócio ao mostrar como fundos de pensão, em diversos países, distribuem recursos entre classes, como o documento de OECD (2025). Além disso, o relatório de PwC e AMAFORE (2016) apresenta as melhores práticas de investimento em fundos de pensão, incluindo a divisão entre classes e critérios de diversificação e gestão de risco.

Capítulo 3

Estudo de Caso

Esse estudo é voltado para o contexto de entidades fechadas de previdência complementar (EFPC), como introduzido no Capítulo 1, que atuam como gestoras financeiras de planos de benefícios. Como gestoras, as EFPC buscam diversificar o portfólio dos planos, equilibrando níveis de risco e retorno e investindo em diferentes segmentos de aplicação. Este trabalho se restringiu a três segmentos em que as EFPC têm investimentos, em geral - renda fixa, renda variável e segmento imobiliário.

Dessa forma, o presente capítulo tem como objetivo mostrar as características do estudo de caso realizado nesta dissertação, relatando o entendimento do negócio, preparação dos dados e método proposto, conforme a metodologia CRISP-DM (Chapman et al., 1999). Foram utilizados dados referentes aos retornos de ativos financeiros do mercado brasileiro a partir de 2022. O objetivo do estudo foi realizar, utilizando o *software* R, o agrupamento dos ativos por meio de técnicas como método de Ward, K*-means* e misturas finitas de normais, após ajuste das variáveis pelas componentes principais funcionais.

3.1 Entendimento do Negócio

O mercado financeiro brasileiro é formado por diversos ativos que podem ser identificados por meio de segmentos de aplicação, como renda fixa, renda variável e imobiliário. Cada um desses segmentos tem características particulares quanto ao nível de risco de mercado e tipo de aplicação que os compõem. Em geral, dentro do contexto das EFPC, cada segmento pode ser composto pelos seguintes ativos (Conselho Monetário Nacional, 2025):

- Renda fixa: títulos de dívida pública, cotas de classes de ETF de renda fixa, ativos financeiros de renda fixa de emissão com obrigação ou coobrigação de instituições financeiras autorizadas a funcionar pelo Banco Central do Brasil, de emissão de sociedade por ações de capital aberto, obrigações de organismos multilaterais emitidas no País, debêntures incentivada e de infraestrutura, cotas de classes de fundo de investimento em direitos creditórios (FIDC), classes de investimento em cotas de FIDC, cédulas de crédito bancário (CCB), certificados de céludas de crédito bancário (CCCB), cédulas de produto rural (CPR), certificados de direitos creditórios do agronegócio (CDCA), certificados de recebíveis do agronegócio (CRA) e *warrant* agropecuário (WA).
- Renda variável: ações, certificados de depósito de valores mobiliários, cotas de fundos de índice referenciados em ações, *Brazilian Depositary Receipts* (BDR), certificados representativos de ouro físico.
- Imobiliário: cotas de fundos de investimento imobiliário (FII), certificados de recebíveis imobiliários (CRI) e cédulas de crédito imobiliário (CCI).

Os ativos de renda fixa possuem uma particularidade: não estão sempre presentes nas negociações do mercado, pois cada um pode ser inserido no mercado em determinado momento e tem data de vencimento (momento em que o ativo não será mais negociado). Dessa forma, foram considerados índices ANBIMA que retratam o comportamento de cada tipo de ativo de renda fixa para as análises. São eles (ANBIMA, s.d.):

- IDA (Índice de Debêntures ANBIMA): reflete o comportamento dos ativos de uma carteira de dívida privada, principalmente debêntures, que possuem séries com prazo superior a um mês. Ele é composto pelos seguintes subíndices:
  - IDA-DI: composto por debêntures remuneradas pela variação da taxa DI.
  - IDA-IPCA: composto por debêntures indexadas ao Índice Nacional de Preços ao Consumidos Amplo (IPCA).
  - IDA-IPCA Infraestrutura: além de ser composta por debêntures indexadas ao IPCA, também são debêntures "incentivadas" que consistem nas que oferecem benefícios fiscais aos investidores.
  - IDA-IPCA ex-Infraestrutura: esse índice é formado pelas debêntures indexadas ao IPCA que não possuem benefícios fiscais.
- IDA LIQ (Índice de Debêntures ANBIMA Liquidez): índice semelhante ao IDA, porém acrescenta critérios de liquidez.
- IDkA (Índice de Duração Constante ANBIMA): representa o desempenho de investimentos em títulos públicos com prazos fixos, mantendo o mesmo período até o vencimento. Ele se divide em:
  - Prefixado: contém a informação sobre o desempenho da curva de juros nominais dos títulos públicos prefixados, como as LTNs e NTN-F.
  - IPCA: neste caso, a curva de juros será atrelada aos títulos públicos indexados à inflação, como as NTN-Bs e o Tesouro IPCA+. Ela é calculada de acordo com os prazos de vencimentos: dois, três, cinco, dez, quinze, vinte e trinta anos.
- IHFA (Índice de *Hedge Funds* ANBIMA): representa o desempenho ao longo do tempo dos fundos multimercado com gestão ativa.

- IMA (Índice de Mercado ANBIMA): apresenta o desempenho da carteira de títulos públicos ao longo do temo e serve como referência para investimentos de renda fixa. Como existe uma variedade de títulos, o IMA é dividido nos seguintes subíndices:
  - IDA-Geral: composto por todos os títulos da dívida pública.
  - IDA-Geral ex-C: composto pela maioria dos títulos da dívida pública. São excluídos apenas os papéis indexados ao IGP-M, como as NTN-Cs e o Tesouro IGPM+ com juros semestrais.
  - IRF-M: composto por títulos públicos prefixados - LTN e NTF com juros semestrais. Ele pode ser subdividido nos índices IRF-M 1, IRF-M 1+ que variam a composição pelo prazo de vencimento, sendo até um ano ou acima de um ano, respectivamente, e IRF-M P2 que foi criado para contemplar os ETFs.
  - IMA-B: composto por títulos públicos indexados ao IPCA com juros semestrais que são as NTN-Bs. Assim como o IRF-M, possui uma divisão de acordo com o prazo de vencimento por meio dos índices IMA-B 5, IMA-B 5+ que são os títulos com vencimento de até cinco anos e igual ou superior a cinco anos, respectivamente, e IMA-B 5 P2 para incluir os ETFs.
  - IMA-S: composto por títulos pós-fixados vinculados à taxa básica de juros (Selic), especificamente as LFTs.

3.2 Preparação dos Dados

Os dados foram coletados por meio da plataforma da Economática no dia 25/08/2025, considerando os retornos diários de cada ativo financeiro até o dia 22/08/2025. Como mencionado na seção 3.1, os ativos de renda fixa foram representados por índices ANBIMA, sendo considerados os seguintes índices: IDA-DI, IDA-IPCA, IDA LIQ DI, IDA LIQ IPCA, IDA LIQ IPCA INF, IDkA IPCA 2A, IDkA IPCA 3A, IDkA IPCA 5A, IDkA IPCA 10A, IDkA IPCA 15A,

IDkA IPCA 20A, IDkA IPCA 30A, IDkA PRE 1A, IDkA PRE 2A, IDkA PRE 3A, IDkA PRE 3M, IDkA PRE 5A, IHFA, IRF-M 1, IRF M 1+, IMA-B 5, IMA-B 5+, IMA-S. Para o caso da renda variável e mercado imobiliário, foram utilizados os ativos financeiros que compõem índices da Bolsa de Valores do Brasil (B3), especificamente IBOV (Índice Bovespa) e IFIX (Índice de Fundos de Investimentos Imobiliários), selecionando todos os ativos disponíveis na plataforma, com base nas informações de B3 S.A. – Brasil (s.d.) referentes à composição dos índices no segundo quadrimestre de 2025.

Tabela 3.1: Quantidade de ativos financeiros por segmento obtidos da coleta da plataforma Economática

Segmento	**Qtde. de ativos**
Fundos Imobiliários	111
Renda Fixa	23
Renda Variável	82
Total	**216**

A Tabela 3.1 acima apresenta a separação dos ativos obtidos por meio da coleta dos dados utilizando a plataforma da Economática. Apesar de ter uma grande quantidade de fundos imobiliários e ações a mercado, foi observado que muitos desses ativos não possui um histórico longo de retornos - muitos deles não apresentavam informação entre 2018 e março de 2022. Por essa razão, foram excluídos das análises todos os ativos financeiros sem essas informações, garantindo, assim, que os restantes para as análises tivessem informações mais completas dentro do período a ser analisado. Dessa forma, restaram apenas 33 fundos imobiliários e 71 ações a mercado.

Optou-se por realizar as análises a partir do primeiro dia útil do ano de 2022 (03/01/2022), priorizando os anos que sucederam a pandemia da COVID-19. Esse recorte tem como objetivo concentrar as análises em um mercado mais homogêneo, evitando a mistura com o período pré-pandêmico e o choque inicial associado à COVID-19 que foi caracterizado por um aumento da volatilidade e quebras estruturais nas séries de retornos (Zeng et al., 2024). No contexto do mercado financeiro brasileiro, os autores Ashikawa e Marçal (2025) mostram que a pandemia

gerou um ambiente de elevada incerteza e mudanças na dinâmica da inflação brasileira, exigindo o uso de preditores robustos por conta da presença dessas quebras estruturais e da dificuldade de projeção em um cenário de choques intensos e persistentes. Além disso, Pereira e Arevalo (2024) destacam que a COVID-19 esteve associada a quebras na série da taxa Selic e em preços de ações de empresas listadas no Ibovespa, indicando alterações relevantes na relação entre política monetária e mercado acionário no período de 2015 a 2022, o que reforça a interpretação de que o choque pandêmico introduziu um novo regime no mercado financeiro brasileiro.

3.3 Método

O trabalho foi dividido em duas etapas principais: Análise de Componentes Principais Funcionais e Análise de Agrupamento. Para realizar a FPCA, é recomendado que as séries a serem analisadas não tenham valores ausentes para evitar que sejam gerados coeficientes sem informação e erros em decomposição espectral. Assim, foi realizada uma interpolação linear com limite de 10 dias de lacuna, a fim de garantir que as funções estejam definidas de forma contínua ao longo do domínio temporal em estudo (Ramsay e Silverman, 1997). Após isso, é preciso que as funções estejam centradas em suas respectivas médias funcionais. Portanto, foi aplicada a Equação (3.1) a cada observação dos dados que, neste momento, como visto na seção 2.3.1, estão apresentados de maneira discreta, em pontos específicos de $\mathcal{T}$.

$$x_i^*(t) = x_i(t) - \overline{x}(t)\ . \tag{3.1}$$

Uma vez que as observações foram centralizadas em suas respectivas médias, foi definida qual técnica de suavização utilizar: bases de Fourier ou *B-splines*. Elas são aplicadas para corrigir o formato "discreto" dos dados coletados, reconstruindo a estrutura das curvas de $x_i^*(t)$, e assumem que qualquer função pode ser aproximada por uma combinação linear de funções base conhecidas, como apresentado na Equação (2.17). A escolha da base depende das características dos dados, sendo as bases de Fourier indicadas para funções periódicas e as *B-splines*

para dados não periódicos.

Devido à natureza não periódica dos retornos diários de ativos financeiros analisados - característica reconhecida na literatura de análise funcional, optou-se pela escolha das bases *B-splines* (Ramsay e Silverman, 1997). Essas séries podem ser influenciadas por choques econômicos e eventos imprevisíveis, não exibindo, consequentemente, padrões cíclicos fixos que colaborem para o uso de bases de Fourier, sendo as *B-splines* mais apropriadas para dados financeiros por conta de seu controle local de suavização e eficiência computacional. Após a suavização, observou-se que dois ativos - CRFB3 e JBSS3 - estavam gerando coeficientes com valores ausentes e, portanto, foram retirados do estudo.

É importante destacar que, para aplicação da técnica de suavização, as datas foram transformadas em valores numéricos, sendo o primeiro dia da série igual a zero, e normalizadas. Essa prática é recomendada pelos autores Ramsay e Silverman (1997) pois garante que todas as séries estejam no mesmo domínio temporal e permite que as técnicas de análise funcional e decomposição em autovalores fiquem mais estáveis por se tratar de um domínio limitado. Portanto, utilizou-se a normalização apresentada na Equação (3.2).

$$\text{Tempo normalizado} = \frac{\text{Tempo } - \min(\text{Tempo})}{\max(\text{Tempo}) - \min(\text{Tempo})} \,. \tag{3.2}$$

Com o ajuste dos dados para um formato de dados funcionais, definiu-se a função de covariância, conforme Equação (2.24), e a Análise de Componentes Principais Funcionais pôde ser ajustada. Nesta etapa, foram calculados os autovalores funcionais que correspondem aos coeficientes das componentes principais funcionais e os autovalores representam a variância explicada por cada componente. Além disso, calculou-se, para cada função observada, os *scores* associados a cada componente que correspondem a quanto cada modo de variação contribui para a estrutura de cada função individual do conjunto.

Espera-se, por meio da FPCA, obter padrões dominantes de variação entre os ativos dos conjuntos de dados que contribuam para a formação de classes heterogêneas entre si. Assim,

foi selecionada a quantidade de componentes principais que devem ser retidas no estudo pela análise da proporção da variância explicada e pelo gráfico *Scree Plot*. Ao determinar o número de componentes, segue-se para a interpretação desses resultados a fim de identificar quais características foram destacadas por cada componente.

Como apresentado na seção 2.3, a interpretação e visualização dos resultados pode ser feita de algumas maneiras. Entre elas, há a visualização das componentes como perturbações da média funcional que facilitam a compreensão de variação predominante nos dados. Ademais, outra opção é a análise dos *scores* que podem auxiliar na identificação de agrupamentos, tendências e pontos discrepantes entre as funções observadas.

Logo, utilizando o resultado dos *scores* funcionais das observações, foi realizado o agrupamento dos ativos financeiros em classes, por meio do uso das técnicas do método de Ward, o algoritmo K*-means* e o modelo de misturas finitas de normais multivariadas. Esses métodos foram comparados utilizando como critérios, por exemplo, a quantidade de classes resultantes, quantos ativos financeiros foram agrupados em cada uma e a característica de cada grupo.

Por fim, foi realizada uma análise a respeito da formação de cada classe de ativos. Para isso, verificou-se a composição de cada classe, elencando seus principais ativos e quais foram os segmentos contemplados por cada uma. Além disso, foram verificados quais padrões de variação cada classe possui a fim de sugerir uma nomenclatura para idenficação de cada uma.

Capítulo 4

Resultados

O presente capítulo apresenta os resultados obtidos por meio das análises descritivas e multivariadas mencionadas anteriormente no Capítulo 2, aplicadas aos retornos diários de ativos financeiros.

4.1 Análise Descritiva

Conforme explicitado no Capítulo 3, foram considerados ativos dos segmentos renda fixa, renda variável e fundos imobiliários, totalizando, respectivamente, em 23, 69 e 33 ativos de cada segmento, como mostrado na Tabela 4.1 a seguir. Essa quantidade final de ativos é consequência dos passos apresentados na seção 3.2.

Tabela 4.1: Quantidade de ativos financeiros utilizados no estudo por segmento

Segmento	Qtde. de ativos
Fundos Imobiliários	33
Renda Fixa	23
Renda Variável	69
Total	**125**

Para exemplificar a variação do retorno dos ativos em estudo, foram selecionados três ativos de cada segmento, considerando as ações e os fundos imobiliários com maior participação nos

índices IBOV e IFIX, respectivamente, e, no caso dos índices de renda fixa, foram escolhidos um de cada tipo, conforme apresentado na seção 3.1. As Figuras 4.1, 4.2 e 4.3 apresentam os retornos mensais entre 2022 e julho de 2025, em porcentagem, em torno de uma linha pontilhada que representa o retorno nulo - igual a zero.

Observa-se, na Figura 4.1, que todas as séries apresentam elevada volatilidade, com meses de ganhos expressivos (acima de 30% em alguns pontos, no caso de PETR3 e VALE3, e acima de 20% no caso de ITUB4) alternados com quedas acentuadas (abaixo de -20%). Esse é um comportamento esperado de ativos de renda variável que são marcados pela alta variação do retorno ao longo do tempo. Pode-se destacar, ao avaliar o gráfico, que as ações PETR3 e VALE3 tendem a registrar variações extremas mais frequentes, indicando maior risco, enquanto ITUB4 exibe oscilações relativamente mais moderadas, quando comparada com as demais ações, ainda que também sujeita a meses negativos relevantes. Dessa forma, a ausência de um padrão claro de tendência comum indica que, embora haja momentos em que as três ações se movimentam na mesma direção, a correlação entre elas não é perfeita, sugerindo benefícios potenciais de diversificação ao combiná-las em uma carteira.

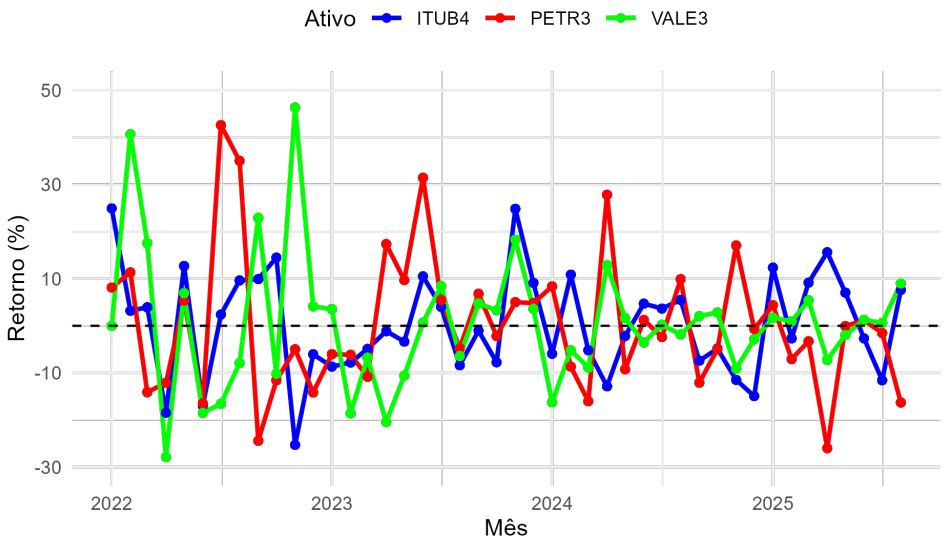

Figura 4.1: Retorno mensal de três ativos de renda variável de janeiro de 2022 a julho de 2025

Ao avaliar os fundos imobiliários (Figura 4.2), nota-se que as oscilações desses ativos são menores do que a observada nas ações a mercado, apresentando, em geral, retornos entre -10% e 10%. Os três fundos exibem trajetória relativamente estável, com alternância de meses positivos e negativos, porém sem mudanças bruscas ou tendências prolongadas de queda, o que é coerente com a natureza mais defensiva dos FIIs de crédito (KNCR11, KNIP11) e dos fundos imobiliários de shoppings (XPML11), quando comparados a ações individuais. Além disso, observa-se que a volatilidade dos FIIs é claramente menor, tanto em amplitude dos retornos mensais quanto na frequência de grandes oscilações, sugerindo um perfil de risco mais moderado e mais adequado para investidores que buscam suavizar a variabilidade da carteira ao longo do tempo.

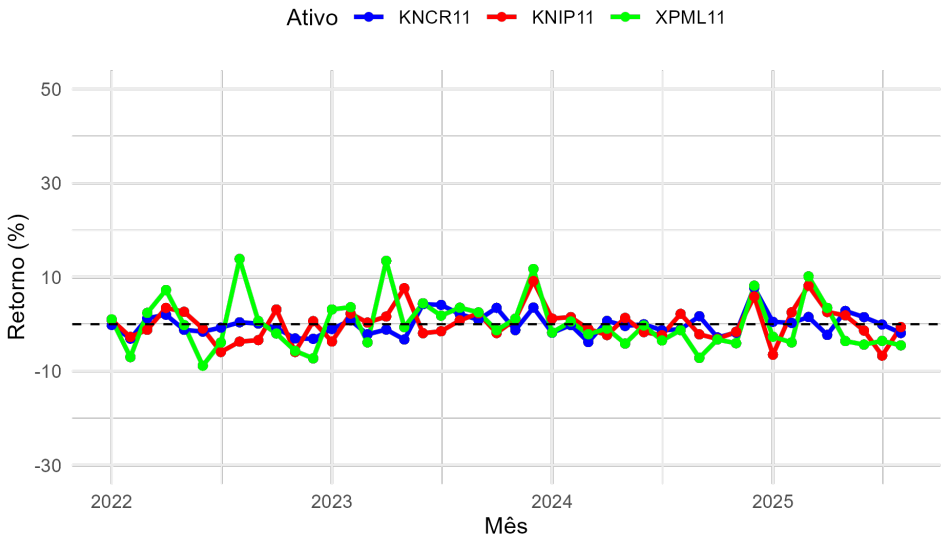

Figura 4.2: Retorno mensal de três fundos imobiliários de janeiro de 2022 a julho de 2025

Por fim, a Figura 4.3 apresenta o comportamento de três índices de renda fixa. Diferentemente dos gráficos anteriores, as oscilações são extremamente baixas, com a maioria dos pontos concentrados em um intervalo de -2% a 2%, aproximadamente, refletindo a baixa volatilidade inerente aos ativos de renda fixa indexados à inflação e taxa de juros. Os fundos seguem trajetórias próximas, especialmente em 2023 e 2024, quando todos registram retornos positivos, o

que indica alta correlação entre eles, provavelmente impulsionada por fatores macroeconômicos comuns como Selic e IPCA. Ademais, os índices de renda fixa exibem volatilidade inferior às ações (que apresentaram picos de -30% e 50%) e aos FIIs (até 15%), posicionando-se como o grupo mais estável, ideal para preservação de capital em períodos de incerteza.

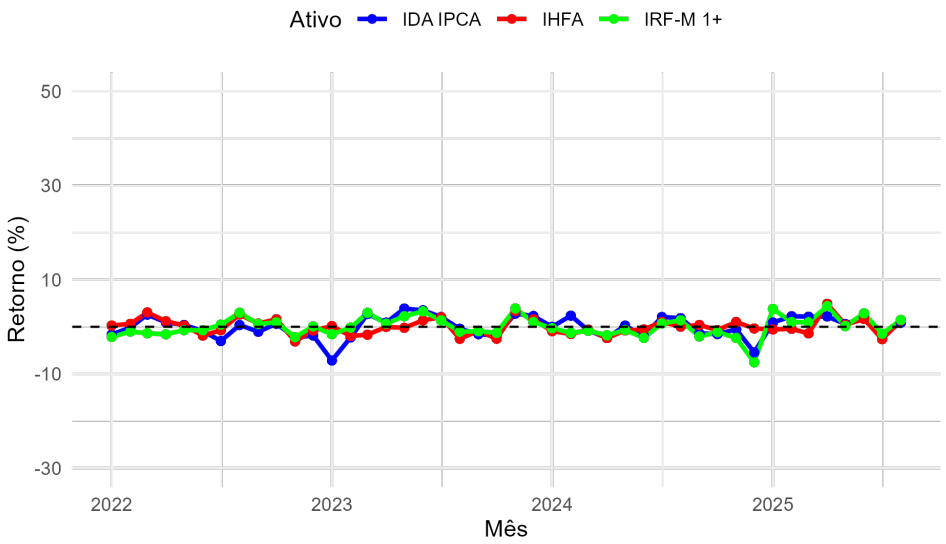

Figura 4.3: Retorno mensal de três índices de renda fixa de janeiro de 2022 a julho de 2025

4.2 Análise de Componentes Principais Funcionais

Como exposto no Capítulo 3, a FPCA foi ajustada após o ajuste dos dados para um formato funcional e a definição da função de covariância. Para avaliar a quantidade de componentes apropriadas para seguir com as próximas etapas do estudo, foi utilizado o gráfico *Scree Plot* (Figura 4.4), que retrata a variância (autovalor) e o número da respectiva componente, juntamente com a proporção da variância total explicada (Tabela 4.2). Por meio do gráfico apresentado a seguir, nota-se que a indicação é de que o uso de duas componentes é suficiente para resumir a variabilidade presente nos dados.

Além disso, observando a Tabela 4.2, nota-se que a proporção da variância de cada componente também indica que o uso de duas componentes é apropriado. As duas primeiras com-

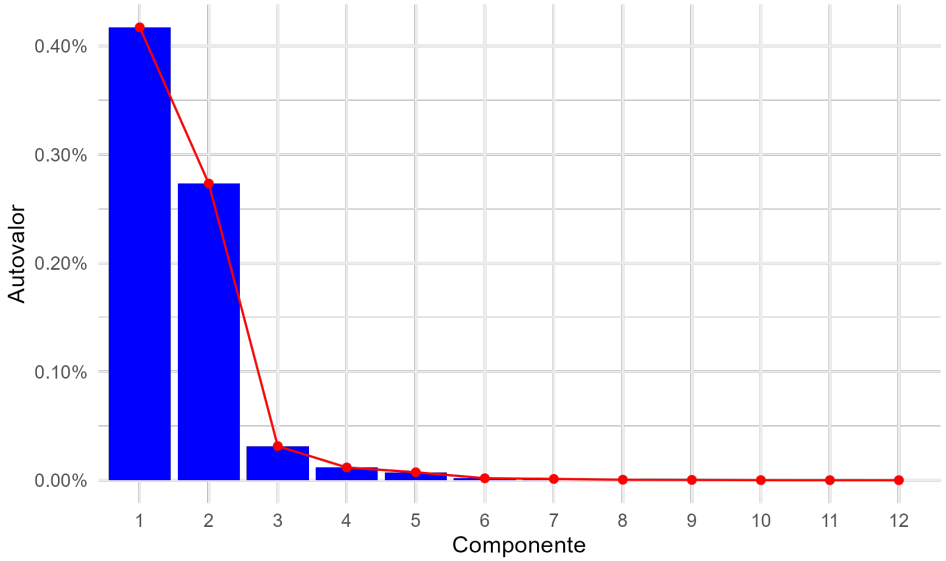

Figura 4.4: *Scree Plot* para seleção do número de componentes da FPCA

ponentes apresentaram, respectivamente, uma proporção da variância igual a 56% e 36,69%, enquanto a terceira já apresenta uma grande queda para apenas 4,22%. Com isso, resulta-se que o valor da variância total acumulada, para duas componentes, é de 92,69%, aproximadamente, o que mostra que mais de 90% da variabilidade dos dados pode ser explicada apenas pelas duas primeiras componentes. Isso traz vantagens para a visualização das observações em grupos, posteriormente, uma vez que é possível representa-las em duas dimensões.

Tabela 4.2: Proporção da variância total e variância total acumulada por número de componente principal

Componente	Proporção da Variância Total	Variância Acumulada
1	56,00%	56,00%
2	36,69%	92,69%
3	4,22%	96,91%
4	1,58%	98,49%
5	0,98%	99,47%
6	0,24%	99,72%

As duas primeiras componentes principais funcionais, também chamadas de harmônicas funcionais, sintetizam os principais padrões de variação presentes nas séries dos ativos finan-

ceiros analisados. De forma geral, a primeira harmônica captura a variação média global dos retornos ao longo do tempo, enquanto a segunda descreve padrões de variação opostos em diferentes instantes, evidenciando mudanças de comportamento em períodos específicos da trajetória temporal.

O gráfico apresentado na Figura 4.5 mostra as tendências extraídas por cada harmônica e mostra que a PC1 possui um padrão geral de "queda", capturando a tendência de longo prazo dos retornos dos ativos. Ativos financeiros com *scores* positivos para esta harmônica tendem a apresentar retornos mais altos no início da série e mais baixos ao final. Em relação à segunda harmônica, o formato de parábola representa o comportamento de ativos que possuem desempenho diferenciado no meio da série em relação às extremidades do período, indicando que esta componente pode capturar movimentos temporários de valorização ou desvalorização conjunta de ativos. Logo, ativos com *score* positivo nesta componente têm retornos com comportamento mais favorável no meio do período.

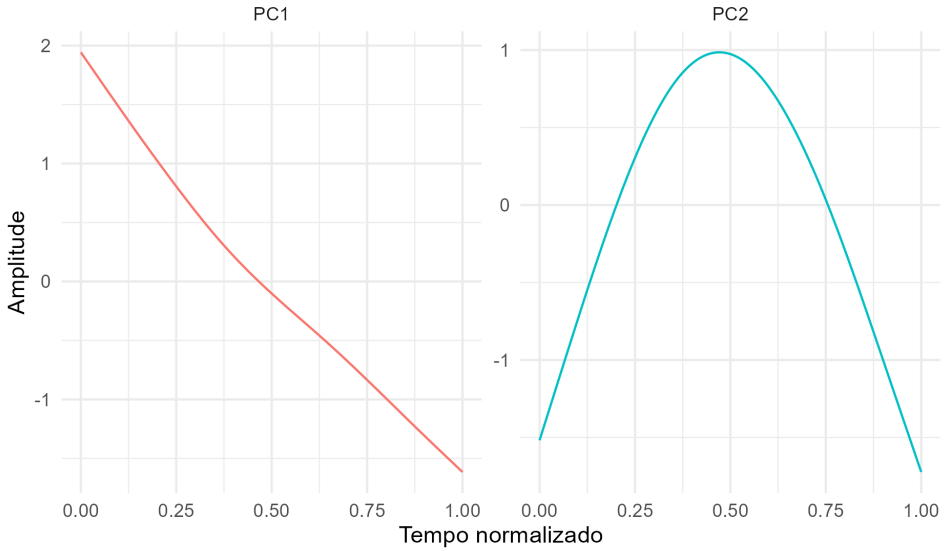

Figura 4.5: Duas primeiras harmônicas resultantes da FPCA

Além disso, conforme recomendado por Ramsay e Silverman (1997), pode-se utilizar a visualização das componentes como perturbações da média funcional, contribuindo para a iden-

tificação dos padrões de variação associados a cada componente ao refletir como cada uma altera a forma média das funções. Assim como a Figura 4.5, os gráficos da Figura 4.6 também apresentam a tendência linear decrescente da PC1 e o formato de parábola da PC2. Isso é evidenciado pelas linhas tracejadas vermelhas que representam a perturbação positiva (conforme Equação (2.29)) de cada componente, enquanto a azul retrata a perturbação negativa, com o comportamento oposto ao apresentado pelas harmônicas.

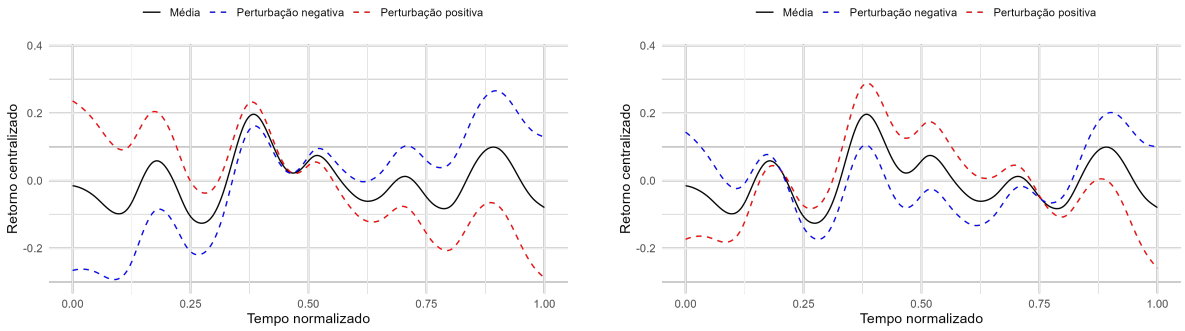

Figura 4.6: Curvas de retorno médio centrados em zero e os efeitos das perturbações das duas primeiras componentes principais funcionais (PC1 e PC2, respectivamente)

As Figuras 4.7, 4.8 e 4.9 mostram a reconstrução em bases *B-splines*, por meio de duas componentes, da curva dos retornos centralizados em zero do índice de renda fixa IRF-M 1+, da ação da Petrobras (PETR3) e do fundo imobiliário de shopping XPML11. Por meio desses gráficos, é possível notar que a suavização extraiu o comportamento geral de variação ao longo do tempo, retirando a flutuação diária que sugere a presença de ruídos na série. Assim, as duas primeiras componentes funcionais capturaram a dinâmica média e padrões gerais de variação do índice no período analisado.

Dessa forma, pode-se avaliar a dispersão dos *scores* dos ativos financeiros nas componentes funcionais a fim de identificar possíveis agrupamentos ou tendências específicas ao longo do domínio $\mathcal{T}$. Por meio da Figura 4.10, é possível observar que ativos de um mesmo segmento estão, em geral, mais próximos, principalmente os fundos imobiliários e os índices de renda fixa. Ativos de renda variável estão mais dispersos do que os dos demais segmentos e apresen-

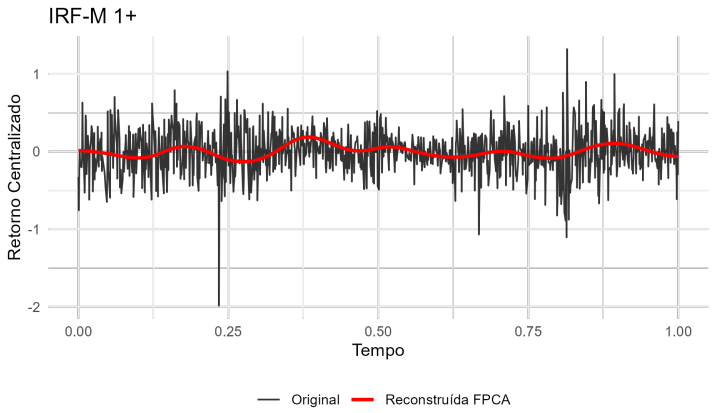

Figura 4.7: Representação original e reconstruída pela FPCA dos retornos diários do índice IRF-M 1+

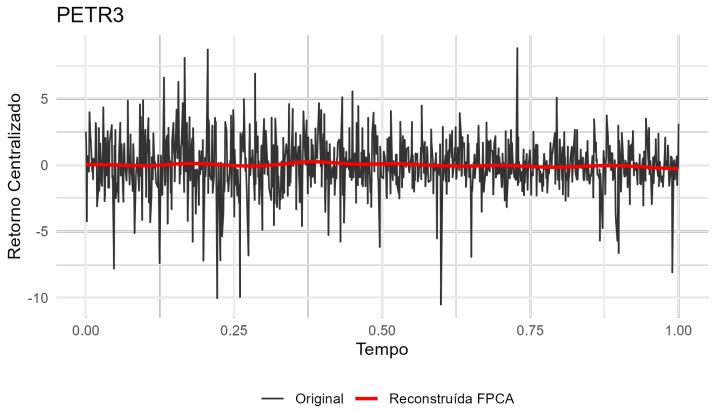

Figura 4.8: Representação original e reconstruída pela FPCA dos retornos diários da ação da Petrobras (PETR3)

tam comportamentos mais acentuados - *scores* em cada componente mais elevados do que os demais. Nota-se, portanto, que o gráfico não expressa uma divisão clara sobre o agrupamento desses ativos, mas traz a indicação de que podem haver interseções entre os segmentos, devido à proximidade dos fundos imobiliários e os índices de renda fixa e a um ativo de fundo imobiliário que está próximo a outros ativos de renda variável.

Vale destacar também que há um ativo que pode ser um indicativo de *outlier* por ser o único com *score* positivo e acima de 0,2 em ambas as componentes. Esse ativo possui uma tendência decrescente, conforme indicado pela primeira harmônica, porém apresenta o desempenho

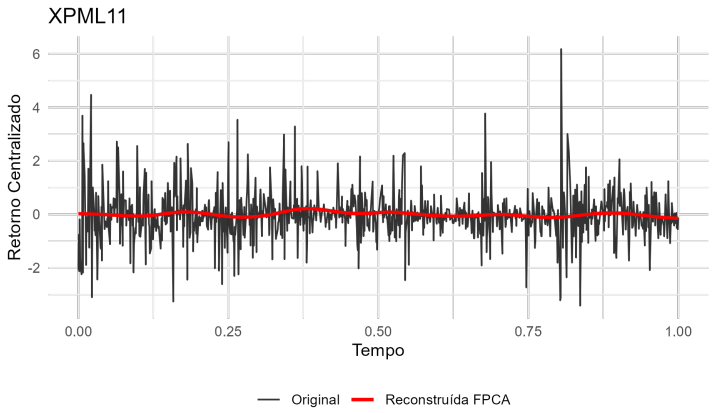

Figura 4.9: Representação original e reconstruída pela FPCA dos retornos diários do fundo imobiliário de shopping XPML11

diferenciado no meio da série, como expressa o padrão da segunda componente (Figura 4.5).

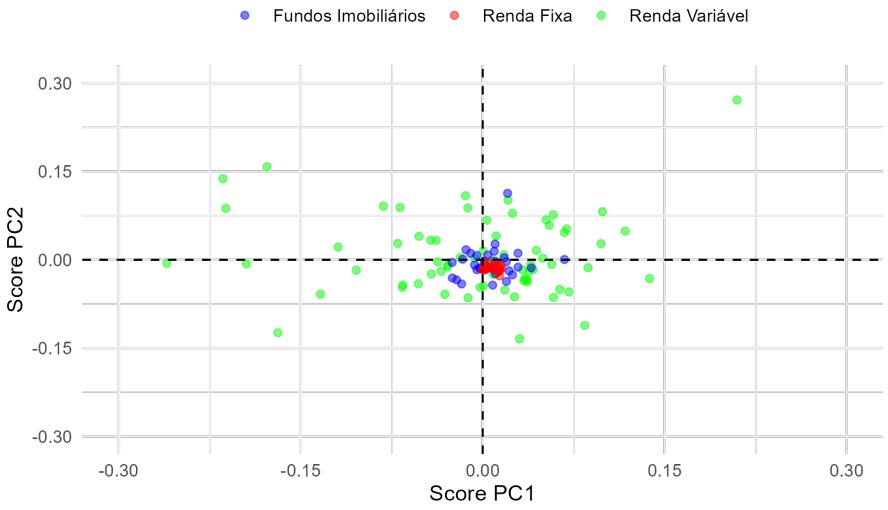

Figura 4.10: *Scores* dos ativos financeiros nas duas primeiras componentes principais

4.3 Agrupamentos

Com a obtenção dos *scores* de cada ativo nas componentes principais, foram feitos os agrupamentos utilizando o método de Ward, *K-means* e misturas finitas de normais.

4.3.1 Método de Ward

O método de Ward foi utilizado, inicialmente, como uma forma de visualizar cada etapa do agrupamento, uma vez que é possível representá-lo por meio de um dendrograma. A Figura 4.11 mostra a separação sugerida pelo método e, utilizando o corte da distância igual a $0,4$, nota-se que foram formados três grupos: o primeiro (da esquerda para a direita) é menor e é formado apenas por ativos de renda variável, o segundo é o maior deles e une ativos dos três diferentes segmentos, e o terceiro é composto majoritariamente por ativos de renda variável, porém com um fundo imobiliário correspondente ao ativo HTMX11 (fundo imobiliário de hotel).

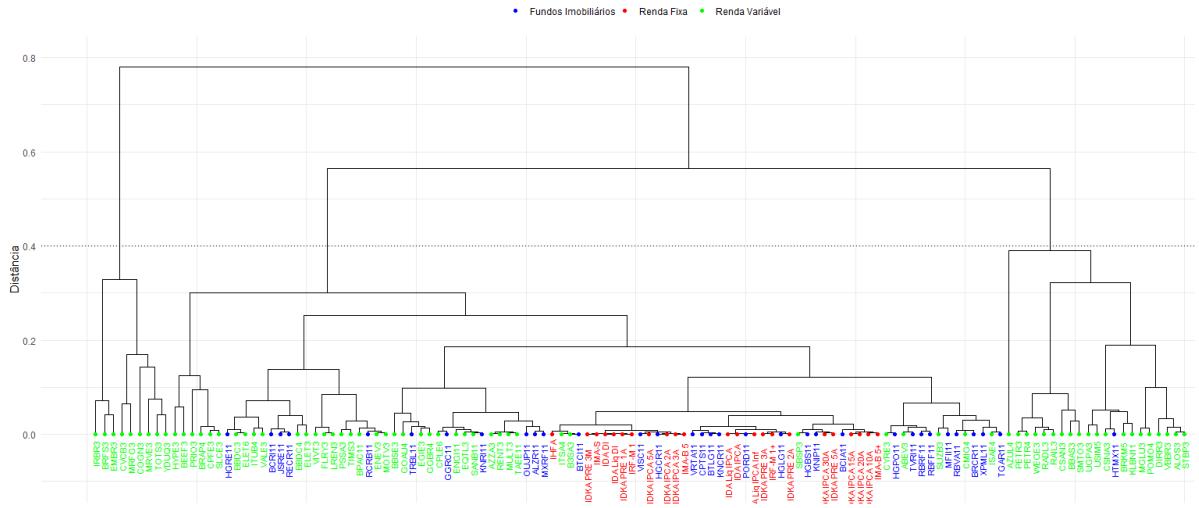

Figura 4.11: Dendrograma do agrupamento de ativos financeiros por meio do método de Ward

Com auxílio da Tabela 4.3 a seguir, é possível observar a quantidade de ativos em cada grupo e a divisão por segmento. O primeiro grupo, formado apenas por renda variável, contempla 9 ativos que são voltados para o setor financeiro/seguros, alimentos de proteína animal, indústria/aeroespacial, educação, turismo, construção civil e tecnologia. Além disso, também é observado um outro grupo - denominado como "Grupo 3" - que têm, além do fundo imobiliário de hotel, maioria de renda variável, porém voltadas para os setores de energia/combustíveis, indústria de bens de capital, siderurgia, logística/transporte, financeiro, varejo, construção/imobiliário e aviação. Vale destacar também que o grupo 2 é o mais diverso dos três, sendo

composto por 95 ativos, sendo 32 fundos imobiliários, 23 de renda fixa e 40 de renda variável.

Tabela 4.3: Quantidade de ativos por grupo e por segmento após agrupamento por meio do método de Ward

Segmento	Grupo 1	Grupo 2	Grupo 3	Total
Fundos Imobiliários	-	32	1	**33**
Renda Fixa	-	23	-	**23**
Renda Variável	9	40	20	**69**
Total	**9**	**95**	**21**	**125**

O gráfico da Figura 4.12 apresenta da divisão dos grupos por meio da dispersão dos *scores* dos ativos financeiros nas duas primeiras componentes principais funcionais. Nota-se que o grupo 1 é composto por ativos que possuem valores negativos maiores, em módulo, do que $-0,1$ na primeira componente, expressando que possuem comportamento diferente do expresso por meio da primeira harmônica, ilustrada na Figura 4.5. Já o grupo 3 é formado por ativos financeiros com *score* positivo na segunda componente principal, indicando que seu retorno possui comportamento semelhante a uma parábola no período analisado.

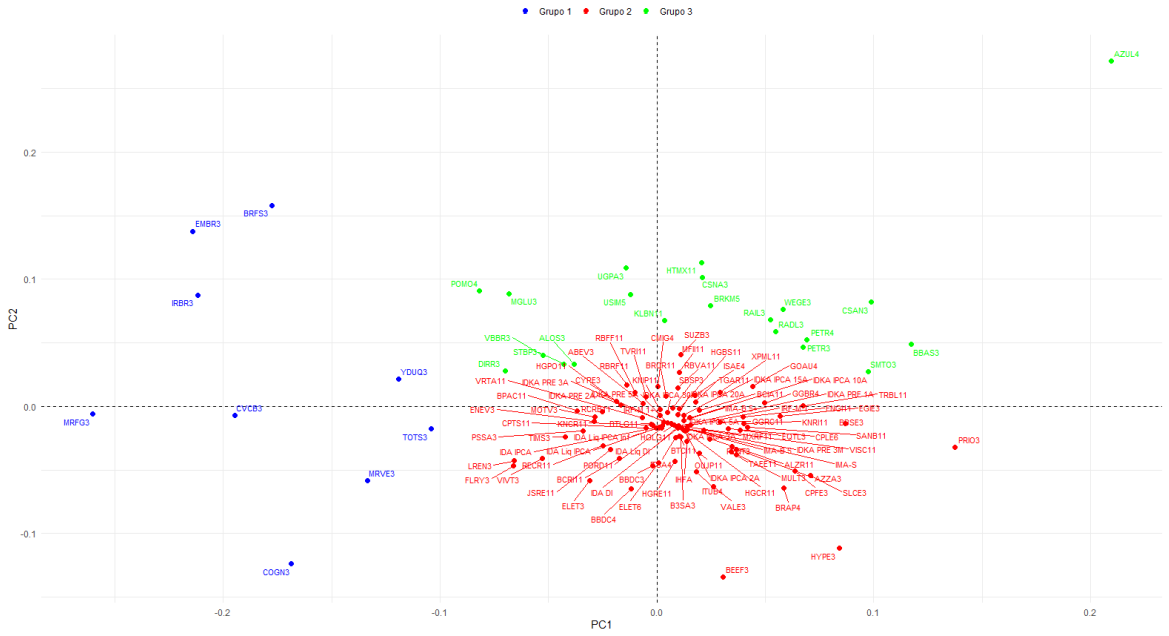

Figura 4.12: *Scores* dos ativos financeiros nas duas primeiras componentes principais funcionais divididos pelo método de Ward

4.3.2 *K-means*

Semelhantemente ao agrupamento realizado pelo método de Ward, foi aplicado o método *K-means* a fim de verificar se o resultado seria mais adequado do que o observado na seção anterior. A escolha do número de grupos foi feita com auxílio do dendograma (Figura 4.11)) e com a avaliação do gráfico a seguir (Figura 4.13) que ilustra a soma de quadrados ao considerar cada quantidade de grupos. É apropriado escolher a quantidade de grupos em que se observa uma quebra na diminuição do valor da soma de quadrados. Com isso, optou-se por utilizar três grupos.

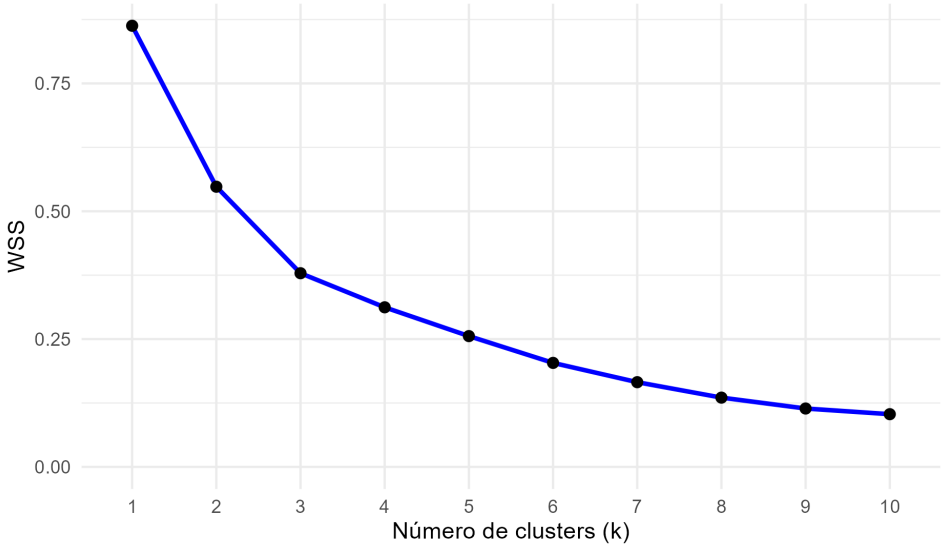

Figura 4.13: Gráfico de *Elbow* para seleção do número de grupos com *K-means*

Observa-se que o algoritmo *K-means* apresentou resultado semelhante ao método de Ward, com a diferença de que o maior grupo teve três ações de renda variável a mais em relação ao agrupamento da Tabela 4.3. Além diso, o Grupo 3 - que era composto por 20 ações - também foi reduzido, passando para 14 ações (Tabela 4.4) dos setores de energia/combustíveis, indústria/bens de capital, siderurgia, logística/transporte, financeiro, varejo e aviação, e permanecendo com o fundo imobiliário HTMX11. O Grupo 1 passou a ser composto por 12 ações, sendo elas dos setores de alimentos de proteína animal, educação, turismo, construção civil, indústria/ae-

roespacial, seguros, varejo, bens de capital/transporte e tecnologia.

Tabela 4.4: Quantidade de ativos por grupo e por segmento após agrupamento por meio de *K-means*

Segmento	Grupo 1	Grupo 2	Grupo 3	Total
Fundos Imobiliários	-	32	1	33
Renda Fixa	-	23	-	23
Renda Variável	12	43	14	69
Total	**12**	**98**	**15**	**125**

A seguir, observa-se o gráfico de dispersão dos *scores* dos ativos financeiros com a indicação do respectivo grupo formado por meio do algoritmo *K-means* (Figura 4.14). Em relação ao resultado obtido pelo método de Ward (Figura 4.12), destaca-se a presença de três ativos dos setores de bens de capital, varejo e construção civil que passaram a integrar o Grupo 1, sendo eles, respectivamente, POMO4, MGLU3 e DIRR3. Além disso, nota-se que o terceiro grupo retirou três ações que estavam mais próximas ao restante do grande grupo 1 - VBBR3, STBP3 e ALOS3 - sendo elas dos setores de energia, logística/transporte e imobiliário, respectivamente.

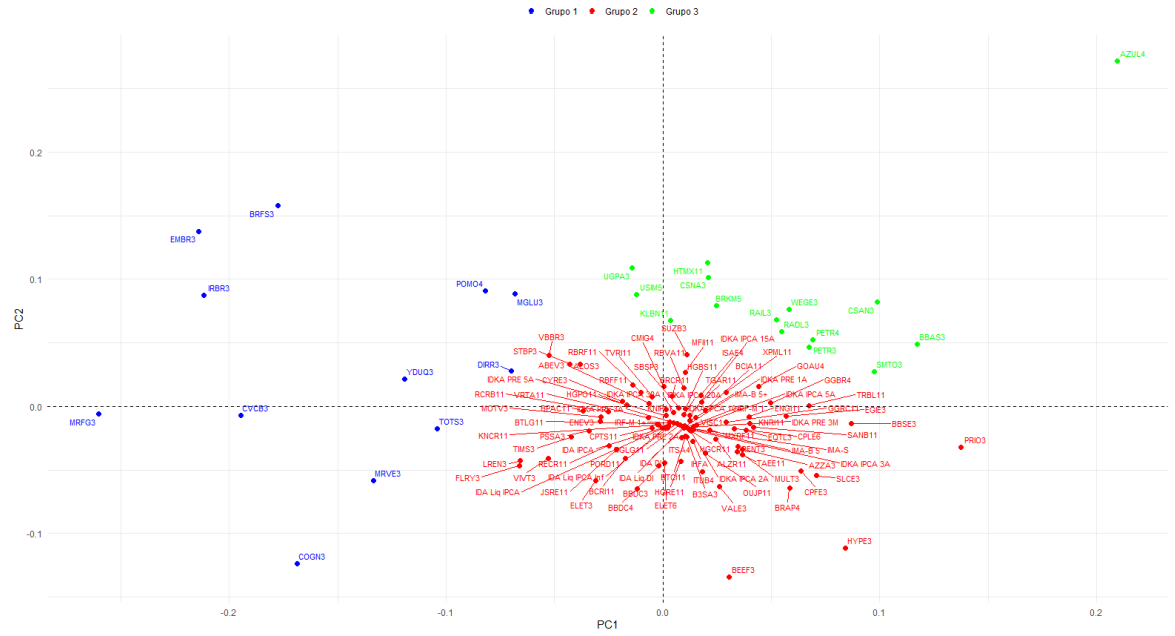

Figura 4.14: *Scores* dos ativos financeiros nas duas primeiras componentes principais funcionais divididos pelo algoritmo *K-means*

4.3.3 Modelo de Misturas Finitas de Normais

Diferentemente dos métodos anteriores, a escolha do número de grupos para o modelo de misturas finitas é feita por meio de medidas de comparação de modelos, como o BIC apresentado na seção 2.4.3. Logo, utilizando o BIC para a escolha do modelo, foi obtido o resultado apresentado na Tabela 4.5 que mostra que o melhor modelo é o VEI - volume variado, forma igual, orientação esférica - com três grupos.

Tabela 4.5: Medidas de comparação de modelos resultante do modelo de misturas finitas VEI ajustado para o agrupamento

Medida	Valor
Log-verossimilhança	451,55
BIC	845,15
ICL	802.38

A separação dos ativos financeiros em estudo por meio do método de misturas finitas de normais apresentou um resultado diferente do que foi observado utilizando os métodos apresentados anteriormente. Pode-se perceber, pela Tabela 4.6 a seguir, que este algoritmo distribuiu as ações (renda variável) em dois grupos, principalmente, e alocou os ativos de renda fixa em um grupo separado de grande parte dos ativos de renda variável e fundos imobiliários (FII). O primeiro grupo está semelhante ao que foi observado nos grupos 3 do método de Ward e algoritmo *K-means* (Tabela 4.3 e 4.4), tendo um único FII junto com ações.

Tabela 4.6: Quantidade de ativos por grupo e por segmento após agrupamento por meio de misturas finitas de normais

Segmento	Grupo 1	Grupo 2	Grupo 3	Total
Fundos Imobiliários	1	24	8	**33**
Renda Fixa	-	2	21	**23**
Renda Variável	35	32	2	**69**
Total	**36**	**58**	**31**	**125**

Com auxílio do gráfico apresentado na Figura 4.15, pode-se identificar quais são os ativos presentes em cada grupo e verificar as características de cada grupo formado. É possível obser-

var que o Grupo 1 é formado pelos ativos de renda variável que estão mais distantes do centro (valor zero para ambas as componentes), incluindo o FII de hotel HTMX11, como nos demais agrupamentos nos quais ele também se mostrou semelhante às ações. O Grupo 2, por sua vez, é formado por diversos fundos imobiliários e ações e tem como destaque dois índices de renda fixa: IDkA IPCA 30 anos que representa o desempenho de investimentos em títulos públicos, com prazos fixos, indexados à inflação, cujo prazo de vencimento é de 30 anos; e IHFA que corresponde ao desempenho ao longo do tempo de fundos multimercado com gestão ativa. Por fim, o Grupo 3 é composto por ativos com valores de *scores* muito próximos em ambas as componentes e majoritariamente formado por índices de renda fixa. Os oito fundos imobiliários que compõem o grupo são cinco FII de papel (investimentos voltados para títulos de dívida ligados à imóveis, como Certificado de Recebíveis Imobiliários (CRI) e letras financeiras) e três de tijolo de logística e shopping/varejo. Já os dois ativos de renda variável consistem em ações da bolsa de valores brasileira (B3) e do Itaú S.A (*holding* brasileira de investimentos).

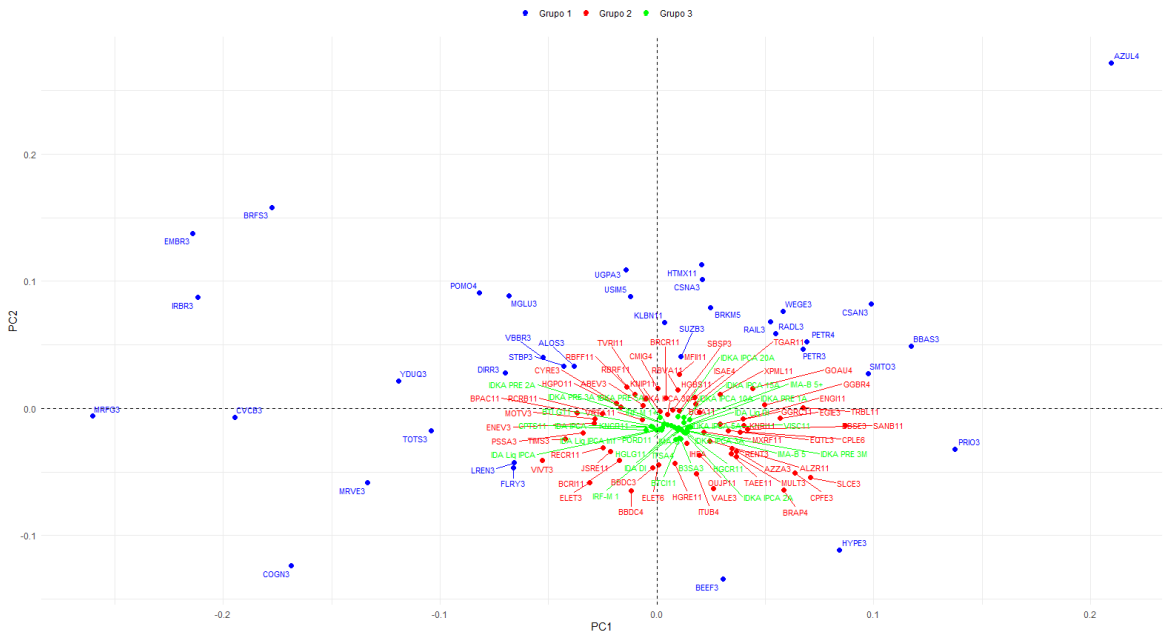

Figura 4.15: *Scores* dos ativos financeiros nas duas primeiras componentes principais funcionais divididos pelo método de misturas finitas de normais

Capítulo 5

Considerações Finais

Um dos principais desafios na gestão de investimentos é a diversificação da carteira por meio da alocação dos recursos em diferentes classes de ativos. No caso das EFPC, essa prática é importante para garantir que os recursos dos planos de benefícios estejam distribuídos de tal forma que a expectativa de retorno seja alta para um baixo nível de risco. Portanto, deve-se buscar ativos que sejam, no mínimo, pouco correlacionados entre si, podendo apresentar correlação nula ou negativa. Porém, essas instituições constroem as classes por meio de avaliações e características empíricas relacionadas ao negócio, por vezes sem considerar que certas classes podem estar altamente correlacionadas.

Neste contexto, foi proposto o uso de técnicas estatísticas multivariadas que fossem voltadas para a construção de grupos, capazes de extrair a correlação entre os ativos financeiros e o comportamento de seus respectivos retornos ao longo dos anos. A Análise de Componentes Principais Funcionais, além de auxiliar na redução de dimensionalidade - como a técnica de PCA clássica, se mostrou eficiente para extrair a tendência dos retornos dos ativos financeiros em estudo, sendo cada componente responsável por representar um tipo de comportamento. Foi visto, neste estudo, que a primeira componente extraiu um padrão geral de "queda", capturando uma tendência decrescente em todo o período analisado, enquanto a segunda componente extraiu um formato de parábola, representando retornos mais elevados no meio do período.

Com os resultados obtidos pela FPCA, foram aplicadas três técnicas de agrupamento - método de Ward, *K-means* e Modelo de Misturas Finitas de Normais - a fim de extrair grupos e verificar qual dos agrupamentos foi mais condizente com as características dos ativos. Cada um faz parte de um tipo de método de agrupamento, sendo, respectivamente, hierárquicos, não hierárquicos e baseados em modelos estatísticos, e podem ser utilizados em diferentes contextos de acordo com o objetivo de cada estudo. Além disso, possuem suas próprias vantagens e desvantagens que devem ser avaliadas dentro de cada estudo.

O método de Ward é eficiente para estudos iniciais e exploratórios, devido a sua característica de hierarquia. Além disso, a visualização dos resultados ocorre por meio de um dendrograma, o que permite inspecionar a estrutura dos grupos em vários níveis de corte, sem precisar fixar o número de grupos anteriormente. Por outro lado, é um algoritmo pouco prático para grandes amostras, devido à complexidade computacional, é sensível à escala das variáveis, pois utiliza a distância euclidiana para realizar os agrupamentos e tem melhor desempenho quando os grupos são aproximadamente esféricos e com tamanhos semelhantes.

O *K-means*, por sua vez, é mais adequado para grandes bases de dados, sendo um dos métodos mais simples e conhecidos atualmente. Ele utiliza o critério de minimização da soma de quadrados intra-grupo e produz, em geral, grupos relativamente compactos e sua interpretação é feita por meio dos centróides que correspondem às médias dos grupos. Apesar disso, possui a desvantagem de ser um algoritmo sensível à escolha inicial dos centróides, podendo convergir para mínimos locais, além de ser necessário escolher o número de grupos antes de realizar o agrupamento, utilizando critérios como o gráfico de *elbow*. Assim como o método de Ward, também é mais adequado nos casos em que os grupos são esféricos e de tamanhos parecidos, apresentando dificuldades, também, em casos em que os grupos se sobrepõem.

Diferentemente dos métodos mencionados acima, o Modelo de Misturas Finitas de Normais utiliza a distribuição dos dados como uma mistura de distribuições normais, permitindo que cada grupo seja interpretado como um componente de um modelo estatístico com parâmetros. Além disso, por ser um método com base probabilística, é possível mensurar a incerteza

associada ao agrupamento obtido e possibilita o uso de métricas como BIC e ICL para a escolha do modelo. Uma outra vantagem em relação aos demais métodos é que permite covariâncias gerais, capturando grupos com volume, forma e orientações distintos. Uma das principais desvantagens deste método é o uso do algoritmo EM para estimação da máxima verossimilhança, sendo sensível a valores iniciais e podendo convergir para máximos locais.

Ao aplicar cada um desses métodos de agrupamento nos *scores* resultantes da FPCA, foram obtidos três diferentes resultados, sendo os obtidos pelo método de Ward e *K-means* semelhantes. Estes algoritmos dividiram os grupos, em geral, pelo quadrante em que se encontravam no gráfico - pela distância entre as posições de cada ativo financeiro. Pode-se notar, pelos gráficos das Figuras 4.12 e 4.14, que um grupo está mais localizado no canto superior esquerdo do gráfico, outro no canto superior direito, com algumas exceções, e o último mais ao centro. O modelo de misturas finitas de normais, por sua vez, agrupou os ativos utilizando, de maneira geral, os *scores*, em módulo, mais elevados em cada componente, sendo possível ver que os ativos mais distantes do centro foram colocados em um grupo e capturando a diferença da variabilidade dos ativos.

Por conseguinte, as técnicas aplicadas foram eficientes para alcançar o objetivo inicial de separar os ativos em grupos e identificar características dos grupos e dos retornos dos ativos financeiros em estudo. Apesar da divergência encontrada entre as aplicações de cada método de agrupamento, foi possível notar que os algoritmos identificaram semelhanças entre alguns ativos de diferentes segmentos, gerando grupos, por exemplo, mais voltados para investimentos em renda fixa, mas que é composto, além de índices ANBIMA, por fundos imobiliários que investem em CRI e letras financeiras. Isso mostra que técnicas estatísticas multivariadas podem auxiliar na gestão de investimentos ao apresentar resultados que, por vezes, podem não ser considerados no momento da separação dos ativos em classes.

Como descrito anteriormente, os métodos aplicados apresentaram divergências nos resultados e a principal característica associada a eles é a natureza dos dados que foram utilizados neste estudo. Os retornos dos ativos financeiros foram centralizados, porém não foram normalizados

a fim de preservar a variabilidade de cada série que é uma característica importante na separação de ativos em classes e que foi considerada, de forma eficiente, no modelo de misturas finitas de normais. Porém, o método de Ward e o algoritmo *K-means* apresentaram resultados limitados, uma vez que são voltados para separação pela diferença de posição e não conseguem capturar diferentes variâncias. Neste caso, para uma melhor eficiência dessas técnicas, recomenda-se um estudo futuro utilizando o agrupamento dos retornos financeiros normalizados para que seja retirada a influência da variância da série dos ativos financeiros.

Sugere-se, além disso, que novos estudos sejam feitos utilizando ativos financeiros de outros segmentos e comparando com outras técnicas de redução de dimensionalidade e agrupamento que sejam mais robustas computacionalmente, como o UMAP e o HDBSCAN. Em 2020, foi revisado um artigo, publicado pela primeira vez em 2018, sobre o algoritmo UMAP (McInnes, Healy e Melville, 2020) que consiste em uma técnica de redução de dimensionalidade projetada para preservar a estrutura local e global dos dados, além de ser útil para visualização e pré-processamento de dados de alta dimensão. Ele se destaca, em relação a outros métodos, por sua velocidade e por conseguir lidar com grandes volumes de dados. O HDBSCAN (McInnes, Healy, Astels et al., 2017) pode ser utilizado juntamente com o UMAP para realizar o agrupamento dos ativos, uma vez que é um algoritmo eficiente para estudos em que não é conhecido o número de grupos e em que pode haver presença de ruídos ou *outliers*.

Referências Bibliográficas

ANBIMA (s.d.). *Índices*. Disponível em: https://www.anbima.com.br/pt_br/informar/precos-e-indices/indices/indices.htm. Acesso em: 29 de junho de 2025.

Anderson, Theodore Wilbur (2003). *An Introduction to Multivariate Statistical Analysis*. 3rd. John Wiley & Sons.

Ashikawa, Rodrigo e Marçal, Emerson Fernandes (2025). "Do robust predictors improve the accuracy of inflation forecasts in moments of structural break?" *Revista Brasileira de Economia* 79.2.

B3 S.A. – Brasil Bolsa, Balcão (s.d.). *Market Data e Índices*. Disponível em: https://www.b3.com.br/pt_br/market-data-e-indices/. Acesso em: 15 jul. 2025.

Bodie, Zvi, Kane, Alex e Marcus, Alan J. (2014). *Investments*. 10th (Global Edition). The McGraw-Hill/Irwin series in finance, insurance, and real estate. New York, NY: McGraw-Hill Education.

Bouveyron, Charles et al. (2019). *Model-based clustering and classification for data science: with applications in R*. Cambridge University Press.

Burca, Valentin et al. (2021). "Portfolio Optimization Using an Alternative Approach. Towards Data Mining Techniques Way". *Economic Alternatives* 1, pp. 113–133.

Chapman, Pete et al. (1999). "The CRISP-DM Process Model". *no. C*.

Conselho Monetário Nacional (2025). *Resolução nº 5.202, de 2025*. Diário Oficial da União.

Duarte, Flavio Gabriel e Castro, Leandro Nunes De (2020). "A framework to perform asset allocation based on partitional clustering". *IEEE access* 8, pp. 110775–110788.

Gabriel, K. R. (1971). "The biplot graphic display of matrices with application to principal component analysis". *Biometrika* 58.3, pp. 453–467.

Hidayatullah, Syarif e Sofro, A'yunin (2024). "Hierarchical Cluster Analysis Based on Waste Sources in Indonesia in 2022". *ComTech: Computer, Mathematics and Engineering Applications* 15.2, pp. 93–99.

Hotelling, Harold (1933). "Analysis of a complex of statistical variables into principal components". *PsycARTICLES* 24.6, pp. 417–441.

Johnson, Richard A. e Wichern, Dean W. (2007). *Applied Multivariate Statistical Analysis*. 6th. Pearson Prentice Hall.

Jolliffe, Ian T. (2002). *Principal Component Analysis*. 2nd. Springer.

Jolliffe, Ian T. e Cadima, Jorge (2016). "Principal component analysis: a review and recent developments". *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* 374.2065.

Kaufman, Leonard e Rousseeuw, Peter J. (1990). *Finding Groups in Data*. John Wiley & Sons.

Lee, Ming-Chang, Lin, Jia-Chun e Stolz, Volker (2024). "Evaluation of k-means time series clustering based on z-normalization and NP-Free". *arXiv preprint arXiv:2401.15773*.

LI, Hongzong (2024). *Clustering Techniques in FinTech: Applications and Prospect*. Disponível em: `https://hkaift.com/clustering-techniques-in-fintech-applications-and-prospect/`. Acesso em: 13 mar. 2025.

Li, Yehua, Wang, Naisyin e Carroll, Raymond J (2013). "Selecting the number of principal components in functional data". *Journal of the American Statistical Association* 108.504, pp. 1284–1294.

Lin, Nan et al. (2015). "Functional principal component analysis and randomized sparse clustering algorithm for medical image analysis". *PLoS One* 10.7.

MacQueen, James (1967). "Some methods for classification and analysis of multivariate observations". *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Vol. 5. University of California press, pp. 281–298.

Markowitz, Harry (1952). "Modern portfolio theory". *Journal of Finance* 7, pp. 77–91.

Marques, Demósthenes (2011). "Asset and Liability Management (ALM) para entidades fechadas de previdência complementar no Brasil: Validação de um modelo de otimização com a aplicação a um caso prático". Diss. de mestr. Universidade de Brasília.

McInnes, Leland, Healy, John, Astels, Steve et al. (2017). "hdbscan: Hierarchical density based clustering." *J. Open Source Softw.* 2.11, p. 205.

McInnes, Leland, Healy, John e Melville, James (2020). "Umap: Uniform manifold approximation and projection for dimension reduction". *arXiv preprint arXiv:1802.03426*.

OECD (2025). *Pensions at a Glance 2025: OECD and G20 Indicators*. OECD Publishing, Paris. DOI: https://doi.org/10.1787/e40274c1-en.

Pearson, Karl (1901). "On lines and planes of closest fit to systems of points in space". *Philosophical Magazine* 2, pp. 559–572.

Pereira, Denis Derkian Martins e Arevalo, Jorge Luis Sanchez (2024). "Análise de quebras estruturais na taxa básica de juros brasileira e no mercado de ações sob o advento da COVID-19". *International Journal of Professional Business Review* 9.4, pp. 1–16.

PwC e AMAFORE (2016). *Global Pension Funds: Best practices in the pension funds investment process*. PricewaterhouseCoopers.

Ramsay, J. O. e Silverman, B. W. (1997). *Functional Data Analysis*. 1st. Springer Series in Statistics.

SAS Institute Inc. (2014). *SAS/STAT® 13.2 User's Guide: The FASTCLUS Procedure*. Capítulo 38 do SAS/STAT® 13.2 User's Guide. SAS Institute Inc. Cary, NC, USA.

Superintendência Nacional de Previdência Complementar (PREVIC) (2022). *Entidades Fechadas de Previdência Complementar (EFPC)*. Disponível em: https://www.gov.br/previc/pt-br/previdencia-complementar-fechada/entidades-

fechadas-de-previdencia-complementar-efpc. Atualizado em 19/07/2022. Acesso em 24/07/2025.

Tang, Geyang, Tian, Rujian e Wu, Bingdi (2022). "An overview of clustering methods in the financial world". Em: *2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022)*. Atlantis Press International B.V., pp. 524–529.

Union Bancaire Privée (2025). *Glossário*. Disponível em: https://www.ubp.com/pt/glossario. Acesso em: 10 de março de 2025.

Wang, Jane-Ling, Chiou, Jeng-Min e Müller, Hans-Georg (2016). "Functional data analysis". *Annual Review of Statistics and its application* 3.1, pp. 257–295.

Zablocki, Rong W et al. (2024). "Using functional principal component analysis (FPCA) to quantify sitting patterns derived from wearable sensors". *International Journal of Behavioral Nutrition and Physical Activity* 21.1, p. 48.

Zeng, Jin et al. (2024). "The effect of the Covid pandemic on stock market volatility: Separating initial impact from time-to-recovery". *Data Science in Finance and Economics*.