UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

Arthur Marchito Leite

Comparação de Técnicas de Agrupamento: K-means, Método Hierárquico de Ward e Misturas de Gaussianas Aplicadas a Dados Reais

Juiz de Fora

2025

Arthur Marchito Leite

Comparação de Técnicas de Agrupamento: K-means, Método Hierárquico de Ward e Misturas de Gaussianas Aplicadas a Dados Reais

Trabalho de conclusão de curso ao Bacharelado em Estatística da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Bacharel em Estatística.

Orientadora: Prof.ª Dr.ª Camila Borelli Zeller

Juiz de Fora

2025

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Leite Marchito, Arthur.
Comparação de Técnicas de Agrupamento: K-means, Método Hierárquico de Ward e Misturas de Gaussianas Aplicadas a Dados Reais / Arthur Marchito Leite. – 2025.
108 f. : il.

Orientadora: Prof.ª Dr.ª Camila Borelli Zeller
Dissertação (graduação) – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Departamento de Estatística, 2025.

1. Agrupamento. 2. Aprendizado não-supervisionado. 3. Cluster. I. Zeller, Camila Borelli, orient. II. Comparação de Técnicas de Agrupamento: K-means, Método de Ward e Misturas de Gaussianas com Aplicações em Dados Reais.

Arthur Marchito Leite

Comparação de Técnicas de Agrupamento: K-means, Método Hierárquico de Ward e Misturas de Gaussianas Aplicadas a Dados Reais

Trabalho de conclusão de curso ao Bacharelado em Estatística da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Bacharel em Estatística.

Aprovada em 18 de Dezembro de 2026

BANCA EXAMINADORA

Prof.ª Dr.ª Camila Borelli Zeller - Orientador
Universidade Federal de Juiz de Fora

Prof. Dr. Lupércio França Bessegato
Universidade Federal de Juiz de Fora

Profª. Drª. Bárbara da Costa Campos Dias
Universidade Federal de Juiz de Fora

Dedico este trabalho à minha avó Lenita.

AGRADECIMENTOS

Agradeço, primeiramente, a Deus, por toda proteção, força e cuidado ao longo da minha vida e desta jornada acadêmica. Expresso minha profunda gratidão à minha família, em especial à minha avó Lenita, à minha mãe Telma e à minha tia, pelo amor incondicional, pelo apoio constante e por estarem ao meu lado em todos os momentos.

Registro também meus sinceros agradecimentos ao professor Lupercio, pelas conversas, orientações e conselhos, que certamente tiveram grande impacto na minha trajetória. À minha orientadora, professora Camila, agradeço por acreditar em mim, pela dedicação, paciência e por todo acompanhamento ao longo do desenvolvimento deste trabalho. Agradeço igualmente à professora Bárbara por ter aceitado participar da banca avaliadora e contribuir com sua experiência.

Sou grato, ainda, aos amigos que encontrei ao longo da jornada universitária. À minha primeira amiga do curso, Catharina, agradeço pelo companheirismo e pela parceria desde o início. Aos amigos Camila, Daniel, Joysce, Gustavo, Natasha e Pedro, deixo meu muito obrigado por compartilharem comigo momentos de aprendizado, apoio e muita risada. A presença de vocês tornou essa caminhada mais leve e inesquecível.

A todos que, de alguma forma, contribuíram para que este trabalho se concretizasse, deixo meu sincero agradecimento.

“Deem-me um ponto de fé, e eu moverei o mundo.”

— Fiódor Dostoiévski

RESUMO

No aprendizado não supervisionado, os métodos de agrupamento são fundamentais para a análise de dados, organizando elementos em grupos (ou clusters) com características semelhantes. Este estudo explora e compara três técnicas amplamente usadas: K-means, método de Ward e misturas de gaussianas. Apesar de compartilharem o objetivo de formar clusters, esses métodos diferem em suas abordagens e pressupostos. O K-means busca minimizar a variabilidade interna dos clusters, enquanto o método de Ward constrói um dendrograma, possibilitando a escolha flexível do número de grupos. As misturas de gaussianas oferecem uma abordagem probabilística, atribuindo probabilidades de pertencimento a cada cluster. Utilizando dados reais, avaliamos o desempenho de cada técnica, concluindo que a escolha do método depende das propriedades dos dados e do objetivo da análise. Este estudo fornece diretrizes práticas para a aplicação dos métodos em diferentes contextos analíticos.

Palavras-chave: Agrupamento; K-means; Método de Ward; Misturas de gaussianas; Dados reais.

ABSTRACT

In the context of unsupervised learning, clustering methods are essential for data analysis, allowing the organization of elements into groups (or clusters) based on similar characteristics. This study aims to explore and compare three widely used clustering methods in the statistical literature: K-means, Ward's method, and Gaussian mixtures. Although all methods share the goal of cluster formation, they differ significantly in their approaches and underlying assumptions. Using real datasets, we applied each of the aforementioned methods, evaluating their performance and effectiveness in group formation. We concluded that the choice of clustering method should consider the intrinsic properties of the data and the analysis objectives. This study contributes to the understanding of the differences and similarities between these methods, providing practical guidelines for applications in various data analysis scenarios.

Keywords: clustering; K-means; Ward's method; Gaussian Mixture Models; Real Data.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

No campo da ciência de dados e do aprendizado de máquina, o aprendizado não supervisionado desempenha um papel crucial na descoberta de padrões ocultos em conjuntos de dados que não possuem rótulos predefinidos (Murphy, 2013). Diferentemente do aprendizado supervisionado, que depende de dados rotulados para treinar modelos, o não supervisionado busca identificar estruturas latentes nos dados (James et al., 2013). Entre suas diversas abordagens, destacam-se os algoritmos de agrupamento (clustering), amplamente utilizados para segmentar dados em grupos (ou clusters) homogêneos, nos quais os elementos apresentam maior similaridade entre si do que em relação aos demais (Hastie et al., 2009).

A análise de agrupamento, também conhecida como análise de clusters, tem como principal objetivo revelar estruturas subjacentes e padrões de similaridade entre objetos, identificando agrupamentos naturais dentro de um conjunto de dados (Jain, 2010). Desde os trabalhos pioneiros de Sokal e Sneath (1963), essa área tem despertado crescente interesse devido à sua ampla aplicação em domínios como biologia, medicina, arqueologia, serviço social, pesquisa de mercado e educação (Hartigan, 1975; Everitt, 1993).

Por exemplo, Morettin e Singer (2022) destacam a aplicação dessa técnica em áreas como segmentação de imagens, bioinformática, reconhecimento de padrões e compressão de grandes volumes de dados. Na bioinformática, é especialmente utilizada para analisar expressão gênica com base em microarrays ou sequenciamento de DNA. Além disso, Izbicki e Santos (2020) e Arabie e Hubert (1994) enfatizam seu uso no marketing, onde empresas segmentam clientes para personalizar estratégias comerciais, como o envio de propagandas direcionadas. No setor leiteiro, os algoritmos de agrupamento permitem classificar produtores com características semelhantes, otimizando a gestão e o uso eficiente de recursos (Seidel et al., 2008).

Diversas abordagens estão disponíveis para a realização de agrupamentos, cada uma com características próprias e aplicabilidades específicas. Este trabalho concentra-se na comparação de três métodos amplamente utilizados: K-means, o método de Ward e as misturas de Gaussianas. O K-means, descrito por Hartigan e Wong (1979), destaca-se pela simplicidade e eficiência computacional. O método de Ward, introduzido por Ward (1963), adota uma abordagem hierárquica com foco na minimização da variabilidade interna dos grupos. Já as misturas de Gaussianas, conforme discutido por McLachlan e Peel (2000), fornecem uma modelagem probabilística mais flexível, capaz de identificar clusters com diferentes formas e tamanhos. Embora todos esses métodos compartilhem o objetivo de particionar os dados em subconjuntos semelhantes, diferem quanto aos pressupostos, algoritmos e formas de interpretação.

A literatura sobre agrupamento é vasta e abrange décadas de pesquisa. Os trabalhos

clássicos de Johnson (1967), Lance e Williams (1968), Jardine e Sibson (1968), Fisher e Ness (1971), Milligan (1979), Dubien e Warde (1979), Wong (1982) e Wong e Lane (1983) estabeleceram os fundamentos teóricos para diferentes abordagens. Avanços posteriores, como os de Jain e Flynn 1996), Frigui e Krishnapuram (1997), Shi e Malik (2000), Iwayama e Tokunaga (1995) e Sahami (1999), ampliaram essas contribuições, com destaque para aplicações em aprendizado de máquina e visão computacional. Mais recentemente, estudos como os de Gan, Ma e Wu (2007) e Marquez (2019) vêm consolidando esses conhecimentos e explorando o uso do agrupamento em cenários de big data. Revisões atuais, como as de de Zhou et al. (2024), Ren et al., (2024) e Wang et al., (2024), destacam o crescimento das abordagens baseadas em aprendizado profundo e a evolução de algoritmos mais sofisticados, capazes de lidar com a complexidade dos dados contemporâneos. Esses trabalhos refletem o crescente interesse da comunidade científica em aprimorar algoritmos de agrupamento para lidar com a complexidade e heterogeneidade dos dados.

1.1 OBJETIVOS

Este trabalho tem como objetivo realizar um estudo aprofundado e comparativo dos algoritmos de agrupamento K-means, método de Ward e misturas de Gaussianas. A pesquisa busca explorar as relações, semelhanças e diferenças entre esses métodos, fornecendo uma análise detalhada de seu desempenho em diferentes cenários de dados. Pretende-se avaliar suas vantagens, limitações e condições de aplicação, investigando como cada um se comporta em termos de eficiência, precisão e robustez em distintos contextos. Além disso, o trabalho visa oferecer diretrizes práticas sobre a escolha do método mais adequado para problemas reais de clustering, com base nas características específicas dos dados e nos objetivos da análise.

1.2 APRESENTAÇÃO DOS CAPÍTULOS E AMBIENTE DE DESENVOLVIMENTO

No Capítulo 2, são apresentados os conceitos fundamentais relacionados à análise de agrupamentos, incluindo as etapas do processo, as medidas de similaridade e uma introdução aos métodos de clusterização abordados. O objetivo é fornecer o embasamento teórico necessário para a compreensão dos algoritmos analisados.

Nos Capítulos 3, 4 e 5, é detalhada a metodologia adotada, com a apresentação dos aspectos teóricos dos métodos hierárquicos aglomerativos, com ênfase no método de Ward, do algoritmo K-means e das Misturas Finitas de Gaussianas, respectivamente.

O Capítulo 6 descreve os métodos utilizados para a escolha do número de grupos, e o Capítulo 7 apresenta as estratégias de validação dos agrupamentos.

No Capítulo 8, é introduzida uma aplicação complementar baseada na segmentação de imagens, em que as técnicas de agrupamento são utilizadas para identificar regiões

homogêneas de cor, oferecendo uma perspectiva alternativa da utilização dos métodos estudados.

O Capítulo 9 apresenta a aplicação prática principal da metodologia. Nesse capítulo, são analisados e discutidos os resultados obtidos pelos diferentes algoritmos. Por fim, o Capítulo 10 reúne as conclusões do estudo, com a discussão dos principais achados, suas implicações e possíveis direções para pesquisas futuras.

Para a execução da parte prática deste trabalho, foram utilizados os seguintes equipamentos, softwares e ferramentas. Os scripts desenvolvidos encontram-se disponíveis no link do GitHub indicado ao final da lista.

- **Sistema Operacional**: Windows 11;
- **Especificações do dispositivo**:
  - **Modelo**: Samsung Book 2;
  - **Processador (CPU)**: Intel Core i5-1235U;
  - **Memória RAM**: 8 GB;
  - **Placa Gráfica (GPU)**: Intel Iris Xe Graphics.
- **Linguagem de Programação**: R, versão 4.3.1 (R Core Team, 2023);
- **Ambiente de Desenvolvimento Integrado**: RStudio, versão 2023.6.2.561 (Posit Team, 2023);
- **Principais pacotes utilizados**:
  - `stats` ;
  - `cluster`;
  - `mclust`;
  - `factoextra`;
  - `ggplot2`;
  - `jpeg`
  - `tidyverse`.
- **Repositório GitHub**: *(https://github.com/ArthurMarchito/TCC).*

2 CONCEITOS INICIAIS

Neste capítulo, iremos introduzir alguns conceitos importantes para contextualizar o tema tratado neste trabalho.

2.1 TRATAMENTO E PADRONIZAÇÃO DOS DADOS

O tratamento e a padronização dos dados são etapas cruciais no processo de análise de agrupamentos. Antes de aplicar algoritmos de clustering, é recomendável estudar o comportamento dos dados e garantir que variáveis e observações estejam em condições apropriadas para esse tipo de análise (Fávero e Belfiore, 2017). Esse preparo envolve tarefas que podem influenciar significativamente a formação e a interpretação dos clusters resultantes, como a análise de outliers, a normalização de variáveis e a eliminação de dados inconsistentes ou irrelevantes.

2.1.1 Tratamento dos Dados

O tratamento de dados visa preparar os dados para assegurar a sua qualidade e eficiência no processo de agrupamento. As etapas essenciais, segundo Doni (2004), incluem:

- **Eliminação de Dados Duplicados ou Corrompidos**: Dados duplicados ou corrompidos são removidos.
- **Tratamento de Outliers**: Outliers são dados que apresentam valores significativamente fora do esperado para uma variável. A decisão de manter ou excluir outliers depende dos objetivos da análise. A exclusão pode ser recomendada quando esses valores distorcem a estrutura de agrupamento de forma não representativa (Fávero e Belfiore, 2017). Uma alternativa recomendada também é conduzir a análise de agrupamento tanto com quanto sem os outliers e, posteriormente, verificar se as conclusões se modificam na presença ou na ausência dessas observações atípicas.
- **Valores Faltantes ou Inválidos**: Dados com valores ausentes ou inválidos devem ser corrigidos ou removidos para minimizar o impacto de inconsistências.
- **Transformação dos Dados**: A transformação visa adequar os atributos para o processo de agrupamento:
  - **Normalização**: Trata dados com atributos em diferentes escalas, especialmente quando se pretende que todas as variáveis tenham influência balanceada no agrupamento.
  - **Tratamento de Atributos**: Assegura que tipos de dados diferentes sejam representados adequadamente no processo de clustering.

2.1.2 Normalização dos Dados por Z-scores

A normalização ajusta variáveis que estão em escalas diferentes para uma mesma escala, permitindo que todas as variáveis tenham a mesma influência no cálculo das medidas de distância entre observações. Quando as variáveis possuem unidades distintas, como peso (em quilogramas) e altura (em centímetros), uma delas pode dominar o cálculo dessas medidas devido à sua magnitude numérica. A padronização, como o procedimento de Z-scores, é uma técnica comum para eliminar essa influência arbitrária. Ela transforma cada valor x_i de uma variável X para uma nova escala com média zero e desvio padrão 1. Esse processo garante que todas as variáveis padronizadas tenham a mesma escala e impacto na análise. A fórmula de Z-score é dada por:

$$z_i = \frac{x_i - \bar{x}}{s_x}, \tag{2.1}$$

em que z_i é o valor padronizado da observação i, x_i é o valor original da variável, $\bar{x}$ corresponde à média amostral e s_x ao desvio padrão da variável x. Por meio dessa transformação, assegura-se que a estrutura dos agrupamentos reflita as relações intrínsecas entre as observações, sem que as diferentes unidades de medida das variáveis exerçam influência indevida no cálculo das distâncias (Fávero e Belfiore, 2017).

2.2 ANÁLISES DE AGRUPAMENTOS

A análise de agrupamentos, também conhecida como clustering, é uma técnica amplamente utilizada para identificar grupos homogêneos, nos quais os elementos de um mesmo grupo apresentem alta similaridade entre si e, simultaneamente, sejam significativamente diferentes dos elementos pertencentes a outros grupos. Johnson, Wichern et al., (2002) destacam que essa coesão interna e diferenciação externa são características fundamentais para o sucesso da análise de agrupamento. De modo geral, o objetivo do clustering é classificar n objetos avaliados por p variáveis, de forma que aqueles que compartilhem características semelhantes sejam agrupados (Faria, 2012).

Segundo Izbicki e Santos (2020), um método de agrupamento, na prática, busca particionar os elementos de uma amostra em subconjuntos distintos $C_1, C_2, \ldots, C_K$, atendendo a duas condições fundamentais. Primeiro, cada elemento deve pertencer a um, e apenas um, desses subconjuntos, o que matematicamente se representa como $C_1 \cup C_2 \cup \cdots \cup C_K = \{1, 2, \ldots, n\}$. Em segundo lugar, os subconjuntos precisam ser mutuamente exclusivos, significando que não há sobreposição entre eles, ou seja, $C_i \cap C_j = \emptyset \quad \forall\, i \neq j$.

No contexto do aprendizado não supervisionado, Morettin e Singer (2022) aponta que a análise de agrupamento é frequentemente empregada para organizar dados em grupos com base em uma medida de similaridade ou dissimilaridade entre observações, também chamadas de medidas de parecença, de forma que elementos de um mesmo grupo

apresentem maior proximidade entre si. Essa premissa fundamenta a alocação conjunta dos itens em clusters, reforçando a homogeneidade desejada.

2.3 PROCESSO DE ANÁLISE DE AGRUPAMENTO

Segundo Artes e Barroso (2023), a análise de agrupamento consiste em etapas fundamentais que orientam o processo de formação e avaliação de clusters:

1. **Escolha do critério de similaridade**: A definição do critério de similaridade é um passo inicial crucial. Nessa etapa, decide-se se as variáveis devem ou não ser transformadas (como a padronização), e qual medida de similaridade será utilizada para a formação dos grupos. Um exemplo comum é a utilização da distância euclidiana para medir a proximidade entre os pontos.

2. **Definição do número de grupos**: Definir a quantidade de clusters é uma etapa que pode ser orientada por conhecimento prévio dos dados, como no caso de conjuntos em que se sabe de antemão o número de classes (por exemplo, espécies em estudos biológicos). Em análises exploratórias ou aplicadas, como a segmentação de mercado, pode ser conveniente selecionar um número menor de grupos para facilitar a interpretação, como no caso de apenas dois clusters. Além disso, existem métodos estatísticos para auxiliar essa escolha, como o método do cotovelo, a silhueta média e o BIC, que permitem avaliar de forma quantitativa o número ideal de grupos. No entanto, o número de clusters também pode ser ajustado após uma análise preliminar dos resultados.

3. **Formação dos grupos**: Nessa etapa, define-se qual algoritmo de agrupamento será utilizado, como K-means, método de Ward ou misturas de Gaussianas, entre outros, com o objetivo de identificar os clusters de acordo com os critérios estabelecidos.

4. **Validação do agrupamento**: Após a formação dos clusters, é necessário verificar se os grupos realmente refletem padrões diferenciados entre as variáveis. Esse processo de validação é fundamental para confirmar a robustez do agrupamento. Existem métodos de validação interna, como o índice de silhueta, que avalia a coesão e separação dos grupos, além de validações externas, que podem incluir métricas como a acurácia média da matriz de confusão e o índice de Rand Ajustado. Essas métricas serão abordadas com maior profundidade em uma seção específica, onde os métodos de avaliação aplicados serão descritos em detalhe.

5. **Interpretação dos grupos**: Em muitas aplicações, após a formação dos grupos, é necessário interpretá-los. A utilização de estatísticas descritivas é recomendada para caracterizar cada grupo, destacando as características que os distinguem.

2.4 MEDIDAS DE PARECENÇA

Para realizar uma análise de agrupamento, a primeira etapa consiste em selecionar uma medida de dissimilaridade ou similaridade entre observações, a qual servirá de base para a formação dos grupos (Fávero e Belfiore, 2017). Existem dois tipos principais de medidas de parecença: as de similaridade, onde valores mais altos indicam maior semelhança, e as de dissimilaridade, nas quais valores mais altos indicam menor semelhança (Artes e Barroso, 2023).

Bussab, Miazaki e Andrade (1990) mencionam que as medidas de similaridade podem ser convertidas em medidas de dissimilaridade, o que permite maior flexibilidade no uso das métricas. Como a maioria dos algoritmos de agrupamento opera com dissimilaridades, é comum realizar essa conversão. Por exemplo, uma matriz de dissimilaridade pode ser obtida a partir da correlação entre os objetos usando a fórmula $d(.) = 1 - \text{corr}(.)$. Nesse caso, a dissimilaridade é definida como a diferença entre 1 e o coeficiente de correlação, ou seja, quanto menor a correlação, maior a dissimilaridade. Quando a correlação positiva e negativa têm significados equivalentes, é possível utilizar $d(.) = 1 - |\text{corr}(.)|$ para a conversão.

A correlação de Pearson, usada para calcular $\text{corr}(X, Y)$, é dada pela expressão:

$$\text{corr}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{2.2}$$

onde, x_i e y_i representam os valores observados das variáveis X e Y para o elemento i-iésimo da amostra, $\bar{x}$ e $\bar{y}$ correspondem às médias amostrais das variáveis X e Y, respectivamente. O índice n indica o número total de observações na amostra. Essa medida expressa o grau de associação linear entre duas variáveis, resultando em um coeficiente que varia de -1 a 1. Valores próximos de 1 indicam uma forte correlação positiva; próximos de -1, uma forte correlação negativa; e valores próximos de 0 sugerem fraca ou nenhuma associação linear.

Segundo Linden (2009), uma função de dissimilaridade é uma função $d : \Gamma \times \Gamma \to \mathbb{R}$, definida sobre o conjunto de objetos Γ que deve satisfazer certas propriedades fundamentais:

- **Não-negatividade:** $d(x, y) \geq 0$ para todos os $x, y \in \Gamma$, sendo que $d(x, y) = 0$ se, e somente se, $x = y$.
- **Simetria:** $d(x, y) = d(y, x)$ para todos os $x, y \in \Gamma$.
- **Desigualdade triangular:** $d(x, z) \leq d(x, y) + d(y, z)$ para todos os $x, y, z \in \Gamma$.

Para que uma medida de dissimilaridade possa ser interpretada como uma distância, ela deve ser não negativa, simétrica e satisfazer a desigualdade triangular. O atendimento dessas propriedades garante um comportamento matematicamente consistente, permitindo

que a dissimilaridade seja tratada formalmente como uma distância entre elementos. Como observado por Izbicki e Santos (2020), existem diversas formas de mensurar dissimilaridade, cada uma mais adequada a diferentes contextos analíticos. A seguir, discutem-se algumas dessas métricas, especialmente aquelas aplicáveis a dados numéricos, com foco em como se adaptam aos objetivos específicos dos algoritmos de agrupamento.

2.4.1 Distância de Minkowski

A distância de Minkowski é uma métrica geral que abrange diversas outras medidas de dissimilaridade como casos particulares, permitindo maior flexibilidade na escolha do método de comparação entre observações. Para duas observações $\mathbf{X} = (x_1, x_2, \ldots, x_p)$ e $\mathbf{Y} = (y_1, y_2, \ldots, y_p)$, a distância de Minkowski é dada por:

$$d(\mathbf{X}, \mathbf{Y}) = \left(\sum_{i=1}^{p} |x_i - y_i|^m \right)^{\frac{1}{m}}, \tag{2.3}$$

onde, x_i e y_i são os valores das variáveis correspondentes às observações $\mathbf{X}$ e $\mathbf{Y}$ em cada uma das p dimensões. O parâmetro m determina o tipo de distância calculada e, ao ser ajustado, permite que a fórmula se adapte a diferentes contextos e características dos dados.

2.4.2 Distância de Manhattan

Com $m = 1$, a medida de Minkowski resulta na distância de Manhattan, ou distância em "blocos", onde a distância entre dois pontos é a soma das diferenças absolutas entre as coordenadas:

$$d(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{p} |x_i - y_i|, \tag{2.4}$$

De acordo com Doni (2004), a distância de Manhattan frequentemente gera resultados similares aos da distância euclidiana. No entanto, sua principal distinção é que, na distância de Manhattan, as diferenças entre as coordenadas não são elevadas ao quadrado, o que reduz o impacto de grandes discrepâncias em uma única dimensão.

2.4.3 Distância Euclidiana

A distância euclidiana também é um caso particular da distância de Minkowski obtido quando se fixa $m = 2$. Essa métrica é uma das mais utilizadas em tarefas de agrupamento, classificação e visualização de dados, por representar geometricamente à distância em linha reta entre dois pontos em um espaço multidimensional. Sua fórmula é dada por:

$$d(\mathbf{X}, \mathbf{Y}) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}, \tag{2.5}$$

Nesse contexto, quanto menor for o valor da distância euclidiana entre dois pontos, maior será a semelhança entre as observações. Por essa razão, ela é frequentemente aplicada quando se deseja medir a proximidade direta entre elementos em um espaço vetorial, assumindo que todas as variáveis estejam na mesma escala ou tenham sido previamente padronizadas.

2.4.4 Distância de Chebyshev

Quando m tende ao infinito, a distância de Minkowski torna-se a distância de Chebyshev, onde apenas a maior diferença entre as variáveis é considerada:

$$d(\mathbf{X}, \mathbf{Y}) = \max_i |x_i - y_i|, \quad i = 1, \ldots, p. \tag{2.6}$$

Esta medida é útil em cenários onde a maior discrepância entre as variáveis é mais relevante do que a soma total das diferenças.

2.4.5 Distância Generalizada ou Ponderada

A distância generalizada, também conhecida como distância euclidiana ponderada, é uma extensão da distância euclidiana tradicional que permite incorporar pesos para cada dimensão, tornando-a útil em casos onde diferentes variáveis têm importâncias distintas. A fórmula para essa distância entre $\mathbf{X} = (x_1, x_2, \ldots, x_p)$ e $\mathbf{Y} = (y_1, y_2, \ldots, y_p)$ é:

$$d(X, Y) = \sqrt{\sum_{i=1}^{p} w_i(x_i - y_i)^2}, \tag{2.7}$$

onde, w_i representa o peso atribuído à i-ésima variável, ajustando a influência de cada diferença $(x_i - y_i)$ no valor total da distância. Quando todos os pesos são iguais a 1, a fórmula se reduz à distância euclidiana tradicional. No entanto, ao modificar esses pesos, é possível enfatizar ou reduzir o impacto de determinadas variáveis, de acordo com seu papel na análise ou com o conhecimento prévio do problema.

2.4.6 Distância de Mahalanobis

A distância de Mahalanobis é uma generalização da distância euclidiana ponderada que considera a estrutura de correlação entre as variáveis. Ela é obtida quando os pesos w_i, utilizados na expressão (2.7), correspondem às entradas da matriz de covariância inversa das variáveis. a distância entre dois vetores $\mathbf{X}$ e $\mathbf{Y}$ é definida como:

$$d(\mathbf{X}, \mathbf{Y}) = \sqrt{(\mathbf{X} - \mathbf{Y})^T \mathbf{\Sigma}^{-1} (\mathbf{X} - \mathbf{Y})}, \tag{2.8}$$

onde, $\mathbf{\Sigma}^{-1}$ representa a matriz de covariância inversa das variáveis. O termo $(\mathbf{X} - \mathbf{Y})^T$ indica a transposição do vetor de diferenças entre as observações $(\mathbf{X} - \mathbf{Y})$. Essa métrica é particularmente útil em contextos nos quais as variáveis estão correlacionadas entre si, pois ela ajusta a influência de cada componente da diferença $X - Y$ com base não apenas na variabilidade individual das variáveis, mas também em sua variabilidade conjunta (Johnson, Wichern et al., 2002).

2.5 CATEGORIAS DOS ALGORITMOS DE AGRUPAMENTOS

Os algoritmos de agrupamento podem ser classificados em diferentes categorias e métodos, conforme suas abordagens e características. Segundo Hastie, Tibshirani e Friedman (2009), esses algoritmos são organizados em três categorias principais:

- **Algoritmos baseados em abordagens combinatórias:** incluem métodos como algoritmos hierárquicos e de partição.
- **Algoritmos baseados em modelos:** Assumem que os dados seguem uma combinação de distribuições, cada uma representando um grupo.
- **Algoritmos de procura por modas:** Algoritmos que identificam as modas das distribuições para definir grupos, agrupando observações próximas de cada moda.

Neste trabalho, conforme discutido na seção de objetivos, o foco será nas duas primeiras categorias: abordagens combinatórias, com ênfase nos algoritmos hierárquicos aglomerativos (Ward) e de partição (K-means), e abordagens baseadas em modelos, utilizando Misturas de Gaussianas. Essas técnicas serão detalhadas nas seções seguintes para proporcionar uma compreensão aprofundada de suas características e aplicabilidades.

3 MÉTODOS HIERÁRQUICOS AGLOMERATIVOS

Os métodos hierárquicos aglomerativos compõem um dos principais subgrupos da abordagem combinatória, pois essa classe de algoritmos trabalha diretamente com os dados, sem fazer suposições sobre a distribuição subjacente. Esses métodos utilizam uma matriz que expressa as similaridades ou dissimilaridades entre os objetos (Hastie et al., 2009).

De acordo com Anderberg (1973), os passos gerais de um método hierárquico aglomerativo são:

- **Inicialização:** Cada elemento da amostra é inicialmente tratado como um grupo independente, resultando em n grupos, $C_1, C_2, \ldots, C_n$, onde C_i representa o grupo contendo o i-ésimo objeto.
- **Cálculo de Similaridade ou Dissimilaridade:** Uma matriz de dissimilaridade é construída para medir as distâncias entre todos os pares de grupos. Para dois grupos C_i e C_j, a dissimilaridade é denotada por d_{ij}, que pode ser calculada com diversas métricas, como a distância euclidiana, por exemplo ou outra apropriada ao contexto.
- **Combinação dos Grupos Mais Próximos:** Identificam-se os dois grupos com menor dissimilaridade (ou maior similaridade) na matriz, representados por C_i e C_j. Esses grupos são então combinados em um novo grupo, denotado por $C_l = C_i \cup C_j$.
- **Atualização da Matriz de Dissimilaridade:** Após a combinação, a matriz de dissimilaridade é atualizada para refletir as novas distâncias entre o grupo C_l e todos os demais grupos restantes na estrutura hierárquica.
- **Iteração do Processo:** Os passos de combinação e atualização são repetidos sucessivamente até que se obtenha o número desejado de grupos, ou até que todos os objetos estejam reunidos em um único grupo.

O processo de fusão nos métodos aglomerativos é realizado de forma iterativa, resultando em uma hierarquia de agrupamentos que pode ser visualizada por meio de um dendrograma. Esse gráfico constitui uma representação bidimensional das etapas de fusão, na qual cada nível indica como o conjunto de dados é particionado ao longo do procedimento. Dessa forma, o dendrograma facilita tanto a análise visual quanto a interpretação das relações entre os grupos formados (Pereira, 1993). Além disso, sua estrutura permite acompanhar a organização progressiva dos agrupamentos durante a execução do método, tornando o processo mais claro e intuitivo (Everitt et al., 2011). A Figura 1 apresenta um exemplo ilustrativo.

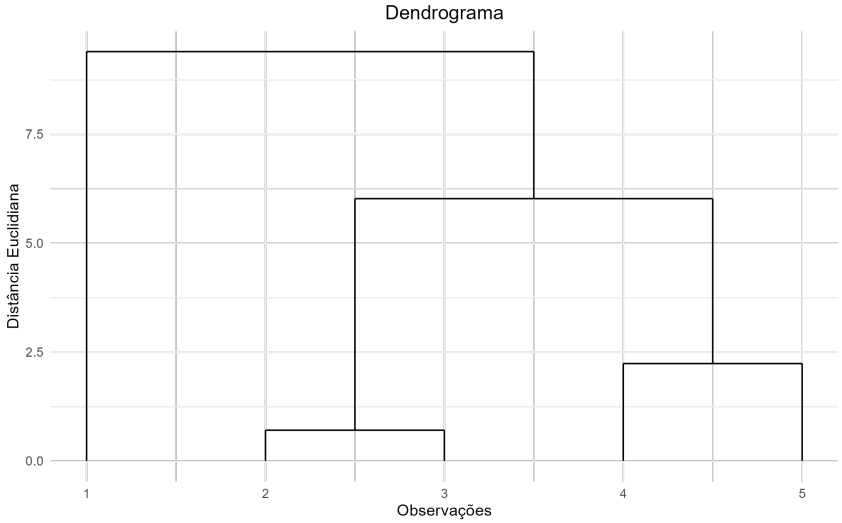

Figura 1 – Dendrograma didático com cinco observações.

Na Figura 1, as cinco observações são representadas pelos rótulos 1, 2, 3, 4 e 5, e a medida de dissimilaridade utilizada é a distância euclidiana. A construção de um dendrograma em métodos hierárquicos aglomerativos inicia-se com o cálculo das distâncias entre todos os pares de observações. Em seguida, identifica-se o par mais semelhante, que é unido para formar o primeiro cluster. Após essa fusão, as distâncias entre o novo grupo e os demais são recalculadas conforme o critério de ligação adotado.

Esse procedimento é repetido de forma iterativa, sempre unindo os grupos ou observações mais próximos, até que todas as unidades pertençam a um único agrupamento. O dendrograma registra graficamente essas etapas de fusão, e a altura em que cada união ocorre representa o custo associado àquela junção: no método de Ward, esse custo reflete o aumento da variabilidade interna (inércia); nos demais métodos, corresponde diretamente à distância entre os grupos envolvidos. No início, cada observação constitui um cluster isolado, e as fusões seguem das combinações mais semelhantes para as menos semelhantes, permitindo visualizar claramente a estrutura de agrupamento.

Além do aspecto visual, é importante destacar uma característica fundamental dos métodos hierárquicos aglomerativos: sua natureza estritamente hierárquica. Uma vez que duas observações são unidas em um estágio inicial, elas permanecem juntas até o final do processo, sem possibilidade de realocação. Essa rigidez contrasta com métodos de particionamento, que permitem que as observações mudem de grupo durante o processo iterativo, o que pode resultar em melhor coesão interna e separação entre clusters (Kaufman e Rousseeuw, 1990).

Outro fator central na formação da hierarquia é a escolha do critério de ligação, mencionado anteriormente, que é o responsável por atualizar a matriz de distâncias a cada fusão. Essa escolha influencia diretamente a forma final dos clusters e a interpretação do dendrograma (Everitt et al., 2011). Conforme discutido por Anderberg (1973), entre

os critérios mais utilizados estão a ligação simples, ligação completa, ligação média e o método de Ward, cada um com vantagens e limitações dependendo da estrutura dos dados e dos objetivos da análise.

A escolha do critério de ligação é formalizada pela fórmula de recorrência proposta por Lance e Williams (1968), que fornece uma abordagem geral para atualizar as distâncias entre grupos em cada etapa do algoritmo (Pereira, 1993). Essa fórmula é suficientemente flexível para incorporar diferentes critérios de ligação, por meio do ajuste de parâmetros específicos. Ela define a distância d_{kl}, onde C_k representa um grupo qualquer e C_l é um grupo formado pela união de dois grupos anteriores $C_l = C_i \cup C_j$, de acordo com a seguinte equação:

$$d_{kl} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} + \gamma |d_{ik} - d_{jk}|, \tag{3.1}$$

onde, os coeficientes α_i, α_j, β e γ determinam a forma como a distância entre o novo grupo C_l e qualquer outro grupo C_k será calculada com base nas distâncias anteriores entre os grupos C_i, C_j e C_k. Cada critério de ligação define um conjunto específico desses parâmetros, o que altera a forma de construção da hierarquia de agrupamento.

No caso da **ligação simples** (*single linkage*), a distância entre dois grupos é definida como a menor distância entre quaisquer dois elementos pertencentes a esses grupos. Na fórmula definida em (3.1), isso é obtido com os seguintes parâmetros: $\alpha_i = \alpha_j = 0{,}5$, $\beta = 0$ e $\gamma = -0{,}5$. Esse critério tende a formar clusters alongados, pois é sensível a cadeias de proximidade.

Na **ligação completa** (*complete linkage*), a distância entre grupos é definida como a maior distância entre quaisquer dois elementos dos grupos. Os parâmetros nesse caso são: $\alpha_i = \alpha_j = 0{,}5$, $\beta = 0$ e $\gamma = 0{,}5$. Esse método favorece a formação de grupos mais compactos e esféricos, sendo menos sensível a outliers do que o *single linkage*.

Já o **método de Ward** busca minimizar a variância total dentro dos grupos, ou seja, forma os agrupamentos de modo a minimizar o aumento do erro quadrático total (soma dos quadrados dos desvios intra-grupo) a cada fusão. Na formulação de LANCE; WILLIAMS, esse método utiliza os seguintes parâmetros:

$$\alpha_i = \frac{n_i + n_k}{n_i + n_j + n_k}, \quad \alpha_j = \frac{n_j + n_k}{n_i + n_j + n_k}, \quad \beta = -\frac{n_k}{n_i + n_j + n_k}, \quad \gamma = 0, \tag{3.2}$$

onde n_i, n_j e n_k representam os tamanhos dos respectivos grupos. Esse método tende a formar grupos de tamanhos similares e com variância interna homogênea. Assim, ela fornece uma estrutura unificada para a implementação de diferentes estratégias de ligação.

Na subseção seguinte, discutiremos o método de Ward com maior profundidade. A escolha desse método se justifica por sua proximidade conceitual com o K-means, um algo-

ritmo de partição amplamente utilizado que também busca minimizar a variância interna dos grupos e que será detalhado no capítulo posterior. Isso possibilita uma comparação direta entre os dois métodos, facilitando a avaliação de suas capacidades de agrupamento. Além disso, incluímos no estudo as misturas finitas de distribuições, uma abordagem probabilística para formação de clusters. Comparar as misturas finitas com os métodos baseados em distância possibilita contrastar a flexibilidade de um modelo probabilístico com a estrutura determinística dos métodos hierárquicos, permitindo uma análise mais abrangente da adequação de cada técnica a diferentes cenários. Esses cenários incluem, por exemplo, dados com grupos bem separados ou sobrepostos, distribuições aproximadamente esféricas ou assimétricas, além de diferentes escalas e níveis de variabilidade.

3.1 MÉTODO HIERÁRQUICO DE WARD

O método de Ward, proposto inicialmente por Ward (1963), é uma técnica de agrupamento hierárquico aglomerativo também conhecida como método da mínima variância (Mingoti, 2005). Seu principal objetivo é formar grupos que apresentem alta homogeneidade interna, ou seja, cujos elementos sejam, semelhantes entre si. Para isso, o método busca minimizar o aumento da variabilidade dentro dos grupos a cada etapa de fusão.

A característica distintiva do método de Ward é o uso da soma dos quadrados dos desvios internos como critério de agrupamento. Em cada iteração do algoritmo, são combinados os dois grupos cuja união resulta no menor acréscimo da soma de quadrados intra-grupo, promovendo a formação de clusters mais compactos e homogêneos ao longo de todo o processo (Hair et al., 2005).

Conforme definido por Ward (1963), a função objetivo do método é a minimização da soma de quadrados SQ dos desvios dos objetos em relação ao centroide do grupo. O centróide, nesse contexto, corresponde ao vetor formado pelas médias de cada variável dentro do grupo, representando assim o "ponto médio" multivariado que sintetiza as características gerais de seus elementos. A distância entre uma observação e o centróide indica o quão diferente essa observação é do perfil médio do grupo, e a soma dessas distâncias ao quadrado mede a variabilidade interna do cluster. A soma de quadrados SQ dentro de um grupo C_k é expressa como:

$$SQ_k = \sum_{i \in C_k} \|\mathbf{x}_i - \bar{\mathbf{x}}_k\|^2, \tag{3.3}$$

onde x_i representa a observação i pertencente ao grupo C_k, e $\bar{x}$ é o centroide desse grupo. A norma $||.||$ denota a distância euclidiana.

Durante o processo de aglomeração, a variabilidade total dos agrupamentos, ou seja, a soma de quadrados entre todos os grupos formados em um determinado estágio, é

dada por:

$$SQ_{total} = \sum_{k=1}^{K} SQ_k, \tag{3.4}$$

onde K é o número de grupos em cada estágio. No estágio inicial, quando cada grupo é composto por um único objeto, a SQ_{total} é zero, pois não há variabilidade interna. À medida que o processo de fusão avança, a união de dois grupos $C_l = C_i \cup C_j$ gera um incremento na SQ_{total}, calculado por:

$$\Delta SQ_{total} = SQ_l - SQ_i - SQ_j. \tag{3.5}$$

Esse incremento ΔSQ_{total} representa o acréscimo de variabilidade resultante da fusão, e o algoritmo do método de Ward seleciona, em cada etapa, o par de grupos cuja união produz o menor valor para esse aumento. Dessa forma, preserva-se a maior homogeneidade possível entre os elementos de cada cluster ao longo do processo (Pereira, 1993).

Como exemplo didático, aplicamos o Método de Ward a três observações em um espaço bidimensional, dadas por $\mathbf{x}_1 = (1, 2)$, $\mathbf{x}_2 = (2, 2)$ e $\mathbf{x}_3 = (8, 8)$. No início do processo cada ponto constitui o próprio grupo, o que implica $SQ_{total} = 0$. Em seguida avaliamos quanto a soma de quadrados intra-grupo aumenta quando possíveis fusões são realizadas. Ao considerar a união entre $\mathbf{x}_1$ e $\mathbf{x}_2$, o centróide do grupo formado pelos dois pontos torna-se

$$\bar{\mathbf{x}}_{12} = \left(\frac{1+2}{2}, \frac{2+2}{2}\right) = (1.5,\ 2).$$

Com esse centróide, a soma de quadrados interna passa a

$$SQ_{12} = \|(1, 2) - (1.5, 2)\|^2 + \|(2, 2) - (1.5, 2)\|^2 = (0.5)^2 + (-0.5)^2 = 0.25 + 0.25 = 0.5.$$

Como os grupos originais tinham variabilidade nula, o incremento provocado pela fusão é $\Delta SQ_{total} = 0.5 - 0 - 0 = 0.5$. Ao analisar a fusão entre $\mathbf{x}_1$ e $\mathbf{x}_3$, o centróide seria

$$\bar{\mathbf{x}}_{13} = \left(\frac{1+8}{2}, \frac{2+8}{2}\right) = (4.5,\ 5).$$

A soma de quadrados interna correspondente torna-se, então,

$$SQ_{13} = \|(1, 2) - (4.5, 5)\|^2 + \|(8, 8) - (4.5, 5)\|^2 = 20.25 + 20.25 = 40.5.$$

A diferença entre os valores de incremento evidencia que a união de $\mathbf{x}_1$ e $\mathbf{x}_2$ produz uma perda de homogeneidade muito menor do que a fusão envolvendo $\mathbf{x}_3$, que está mais distante no espaço. Assim, segundo o critério de Ward, a primeira fusão realizada pelo algoritmo ocorreria entre $\mathbf{x}_1$ e $\mathbf{x}_2$, pois essa junção ocasiona o menor aumento da variabilidade interna e, consequentemente, preserva melhor a compacidade dos grupos formados.

Dando continuidade, conforme demonstrado por Wishart (1969), o incremento na soma de quadrados total decorrente da fusão de dois grupos pode ser reformulado da seguinte maneira:

$$\Delta SQ_{total} = \frac{n_i n_j}{n_i + n_j} \left\| \bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j \right\|^2 . \tag{3.6}$$

Essa expressão mostra que a variação total ΔSQ_{total} depende diretamente da distância entre os centróides dos grupos C_i e C_j, denotados por $\bar{\mathbf{x}}_i$ e $\bar{\mathbf{x}}_j$, respectivamente. Quanto menor essa distância, menor será o aumento da variabilidade total. Dessa forma, o método de Ward favorece a união de grupos mais próximos entre si, promovendo uma estrutura de clusters mais coesa. Os termos n_i e n_j representam o número de elementos nos grupos C_i e C_j, respectivamente. Eles introduzem um fator de ponderação que ajusta a importância do incremento de variabilidade de acordo com o tamanho dos grupos envolvidos, penalizando fusões desbalanceadas. Assim, a expressão (3.6) favorece a união de grupos com tamanhos similares e centróides próximos, reduzindo a variabilidade interna ao longo do processo.

Com base nisso, o critério de dissimilaridade utilizado no método de Ward pode ser definido como:

$$d_{ij} = \frac{n_i n_j}{n_i + n_j} \left\| \bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j \right\|^2 . \tag{3.7}$$

Essa métrica de dissimilaridade mostra que a fusão entre dois grupos é considerada mais apropriada quando a distância entre seus centróides é pequena e seus tamanhos são similares. Assim, a cada etapa do algoritmo, o método de Ward seleciona o par de clusters cuja fusão acarreta o menor aumento na variabilidade interna, promovendo a formação de grupos internamente homogêneos.

Além de sua formulação direta, o método de Ward também pode ser adaptado à estrutura recursiva proposta por Lance e Williams (1968), conforme observado na expressão (3.1). Nesse contexto, a distância entre um grupo C_k e o grupo resultante da união $C_l = C_i \cup C_j$ é dada por:

$$d_{kl} = \frac{n_i + n_k}{n_i + n_j + n_k} d_{ik} + \frac{n_j + n_k}{n_i + n_j + n_k} d_{jk} - \frac{n_k}{n_i + n_j + n_k} d_{ij}. \tag{3.8}$$

essa expressão recursiva permite atualizar eficientemente a matriz de dissimilaridades ao longo do processo hierárquico, mantendo a consistência com o critério de mínima variância adotado por Ward.

O método de Ward também apresenta propriedades matemáticas relevantes, como a monotonicidade das dissimilaridades ao longo das iterações, o que assegura que a variabilidade interna total dos grupos não diminui durante o processo de fusão (Pereira, 1993). Embora possa ser aplicado com diferentes medidas de dissimilaridade, seu uso

com distâncias alternativas pode dificultar a interpretação dos agrupamentos (Anderberg, 1973). Uma característica marcante do método é sua tendência a formar grupos de tamanhos semelhantes, favorecendo a agregação de objetos a clusters menores em casos de equidistância (Kalkstein, 1987). No entanto, sua sensibilidade a outliers exige atenção prévia ao tratamento dos dados, pois a presença de valores atípicos pode comprometer a qualidade das soluções obtidas (Milligan, 1979). Dessa forma, o método se destaca como uma abordagem eficaz para a formação de clusters compactos, onde se deseja maximizar a homogeneidade interna dos grupos.

.

3.1.1 Exemplificação do procedimento

O funcionamento do método Ward pode ser descrito nas seguintes etapas principais:

1. **Inicialização dos Grupos:**
   - Tratar cada objeto individualmente como um grupo, resultando em N grupos no início.
2. **Cálculo da Dissimilaridade:**
   - Construir uma matriz de dissimilaridade, com a distância entre cada par de grupos calculada. A dissimilaridade entre grupos é medida pelo aumento da soma dos quadrados entre grupos (Ward minimiza a variância dentro dos clusters).
3. **Combinação de Grupos:**
   - Em cada etapa, dois grupos são unidos. A combinação ocorre entre aqueles que resultam no menor aumento na soma de quadrados total (SQ_{total}).
   - O critério de fusão é baseado no aumento mínimo de SQ_{total} dado por ΔSQ_{total}, que é proporcional à distância euclidiana entre os centróides dos grupos a serem combinados.
4. **Atualização das Distâncias:**
   - Após a fusão de dois grupos, as distâncias entre o novo grupo formado e os demais grupos são recalculadas.
   - A fórmula utilizada para essa atualização segue a equação de dissimilaridade (3.7), ponderando as distâncias com base no tamanho dos grupos.
5. **Repetição do Processo:**
   - As etapas de fusão e atualização das distâncias são repetidas iterativamente.

- Em cada etapa, o número de grupos é reduzido em 1, até que reste apenas um único grupo contendo todas as observações.

4 MÉTODO K-MEANS

Os métodos de partição, como descrito por Artes e Barroso (2023), têm como objetivo dividir um conjunto de dados em grupos distintos, de forma que cada ponto de dado seja alocado a um único grupo, resultando em uma segmentação clara e mutuamente exclusiva dos dados. Essa abordagem é especialmente útil quando o número de grupos desejado é previamente definido e as observações precisam ser distribuídas de maneira a minimizar a variabilidade interna de cada grupo, ao mesmo tempo em que maximizam a variabilidade entre os grupos (Bussab et al., 1990).

O K-means, foco deste capítulo, é um dos métodos de partição mais amplamente utilizados em análise de agrupamento, sendo frequentemente empregado como ponto de partida para explorar a estrutura dos dados, devido à sua simplicidade, eficiência e fácil implementação (Jain, 2010). Seu objetivo é particionar um conjunto de n observações b-dimensionais, representadas por $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^b\}, i = 1, \ldots, n$, em K clusters $C = \{C_k\}, k = 1, \ldots, K$, de modo a minimizar o erro quadrado total entre os pontos de cada cluster e seu centróide. O erro quadrado dentro de um cluster C_k é expresso pela seguinte equação:

$$SQ_k = \sum_{i \in C_k} \|\mathbf{x}_i - \bar{\mathbf{x}}_k\|^2, \tag{4.1}$$

onde $\|\mathbf{x}_i - \bar{\mathbf{x}}_k\|$ representa a distância Euclidiana entre um ponto x_i e o centróide $\bar{\mathbf{x}}_k$. O objetivo global do algoritmo é minimizar a soma total dos erros quadrados para todos os K clusters, conforme a equação:

$$SQ_{\text{total}} = \sum_{k=1}^{K} SQ_k \tag{4.2}$$

Minimizar a função objetivo do algoritmo K-means é um problema classificado como *NP-difícil* (NP-hard), o que significa que, com os recursos computacionais atuais, não se conhece um algoritmo capaz de resolvê-lo de forma exata, em tempo polinomial, para todos os casos possíveis (Drineas et al., 1999). Conforme definido por por Cormen et al., (2009), um problema é considerado NP-hard quando é, no mínimo, tão difícil quanto qualquer problema da classe NP, no sentido de que qualquer problema em NP pode ser reduzido a ele em tempo polinomial. Diferentemente dos problemas NP-completos que além de serem NP-hard, pertencem à classe NP (ou seja, têm soluções verificáveis em tempo polinomial), os problemas NP-hard não precisam necessariamente ser problemas de decisão, podendo inclusive estar fora de NP.

Em termos simples, problemas NP-difíceis são aqueles cuja complexidade cresce de forma tão acentuada com o aumento do tamanho da entrada que se tornam intratáveis

para grandes volumes de dados. Embora seja possível verificar rapidamente se uma solução proposta é válida, encontrar essa solução pode exigir tempo exponencial no pior caso.

Um exemplo clássico de problema NP-difícil é o *Problema do Caixeiro Viajante* (Travelling Salesman Problem — TSP), no qual se busca a menor rota possível que permita a um vendedor visitar uma série de cidades apenas uma vez e retornar ao ponto de origem (Cormen et al., 2009). À medida que o número de cidades aumenta, o número de combinações possíveis cresce de forma exponencial, tornando a busca pela solução ótima extremamente custosa em termos computacionais.

Para lidar com essa dificuldade no K-means, o algoritmo emprega estratégias heurísticas e iterativas para encontrar soluções viáveis em tempo aceitável, movendo-se em direção a um mínimo local da função objetivo. Isso significa que, embora o algoritmo não garanta encontrar o mínimo global, ele converge para uma solução satisfatória em tempo razoável. Estudos mostram que, em muitos casos, especialmente quando os clusters estão bem separados, o K-means tem alta probabilidade de alcançar a solução ótima global (Meila, 2006).

O funcionamento do algoritmo K-means ocorre de forma iterativa, alternando entre duas etapas fundamentais:

- **Etapa 1 – Atribuição:** Para cada ponto $\mathbf{x}_i$, determinar o centróide mais próximo e atribuí-lo ao cluster correspondente. Isso é feito com base na distância euclidiana, conforme:

$$c_i^{(t)} = \arg\min_k \|\mathbf{x}_i - \bar{\mathbf{x}}_k^{(t-1)}\|^2, \tag{4.3}$$

  onde $c_i^{(t)}$ representa o rótulo do cluster atribuído ao ponto $\mathbf{x}_i$ na iteração t, e $\bar{\mathbf{x}}_k^{(t-1)}$ é o centróide do cluster C_k obtido na iteração anterior.

- **Etapa 2 – Atualização dos centróides:** Após a etapa de atribuição, cada cluster $C_k^{(t)}$ possui um conjunto de observações associadas. O centróide do cluster C_k na iteração t é então recalculado como a média aritmética de todos os vetores pertencentes ao cluster. Como cada observação $\mathbf{x}_i$ é um vetor em $\mathbb{R}^b$, a soma $\sum_{\mathbf{x}_i \in C_k^{(t)}} \mathbf{x}_i$ corresponde à soma componente a componente desses vetores. Assim, o novo centróide é dado por:

$$\bar{\mathbf{x}}_k^{(t)} = \frac{1}{|C_k^{(t)}|} \sum_{\mathbf{x}_i \in C_k^{(t)}} \mathbf{x}_i, \tag{4.4}$$

  Em termos explícitos, se cada vetor é escrito como

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip}),$$

então o centróide recalculado é o vetor

$$\bar{\mathbf{x}}_k^{(t)} = \left(\frac{1}{|C_k^{(t)}|} \sum_{\mathbf{x}_i \in C_k^{(t)}} x_{i1},\ \frac{1}{|C_k^{(t)}|} \sum_{\mathbf{x}_i \in C_k^{(t)}} x_{i2},\ \ldots,\ \frac{1}{|C_k^{(t)}|} \sum_{\mathbf{x}_i \in C_k^{(t)}} x_{ip} \right).$$

onde $|C_k^{(t)}|$ representa o número de observações atribuídas ao cluster C_k na iteração t. Dessa forma, cada coordenada do centróide é a média das coordenadas correspondentes das observações do cluster.

Essas duas etapas são repetidas sucessivamente até que o algoritmo atinja convergência, o que ocorre quando as atribuições deixam de mudar ou quando a variação dos centróides entre iterações se torna inferior a um limiar pré-estabelecido. Nesse estágio, considera-se que o procedimento estabilizou em um mínimo local da função objetivo, isto é, a soma de quadrados total intra-clusters definida em (4.2), e o processo iterativo é então encerrado.

4.0.1 Critério de Qualidade da Partição

Os critérios de qualidade em algoritmos de partição, conforme elucidado por Artes e Barroso (2023), buscam identificar uma configuração de agrupamento que apresente alta homogeneidade interna (isto é, elementos semelhantes dentro de um mesmo grupo) e elevada heterogeneidade externa (ou seja, grupos bem distintos entre si). No caso do algoritmo *K-means*, esse critério é representado pela minimização da Soma de Quadrados Dentro dos Grupos SQ_k, conforme definida em (4.1), a qual quantifica a variabilidade interna de cada grupo, medindo o quão distantes as observações estão de seus respectivos centróides. A função objetivo do *K-means*, portanto, consiste em minimizar essa soma total, conforme apresentado na equação (4.2).

Esse critério é inspirado na decomposição da variabilidade utilizada na Análise de Variância (ANOVA), em que a soma total dos quadrados (SQ_{Total}) é decomposta em dois componentes: a variabilidade entre os grupos e a variabilidade dentro dos grupos. No contexto do *K-means*, o foco está na minimização da variabilidade intra-grupo, isto é, na redução das distâncias entre os pontos e seus centróides, promovendo assim a formação de grupos mais coesos.

O método de Ward, utilizado em técnicas hierárquicas, adota um critério semelhante: a cada etapa do processo de fusão de *clusters*, seleciona-se a união entre os grupos cuja junção produza o menor incremento em ΔSQ_{total}. Dessa forma, tanto o *K-means* quanto o método de Ward compartilham o objetivo de otimizar a homogeneidade interna dos agrupamentos, ainda que empreguem estratégias distintas para a construção dos *clusters*.

4.0.2 Exemplificação do procedimento

O funcionamento do método K-means pode ser descrito nas quatros etapas principais (Giolo, 2017):

- **Inicialização:** Escolha K centróides a priori por meio de algum método escolhido e mais adequado.
- **Atribuição:** Cada ponto de dados é atribuído ao centróide mais próximo, utilizando alguma medida de distância, geralmente a euclidiana. Isso resulta na formação de K grupos. Cada ponto pertence exclusivamente a um único grupo.
- **Atualização:** Recalcule os centróides como a média das observações atribuídas a cada grupo.
- **Iteração:** Repita os passos de atribuição e atualização até que nenhuma observação seja mais realocada ou até que se atinja um número máximo de iterações, ou que algum critério de qualidade seja atingido.

A Figura 2 apresenta uma ilustração do funcionamento iterativo do algoritmo K-means aplicada ao conjunto de dados *Old Faithful*, extraída de Bishop e Nasrabadi (2006). Esse conjunto é composto por medições da duração das erupções e do intervalo de tempo entre eventos sucessivos do gêiser localizado no Parque Nacional de Yellowstone. Os pontos no plano bidimensional representam as observações, enquanto as cores indicam a partição final obtida pelo algoritmo após a convergência. Os símbolos em formato de "X" correspondem aos centróides estimados para cada agrupamento, os quais são atualizados iterativamente ao longo do processo de otimização.

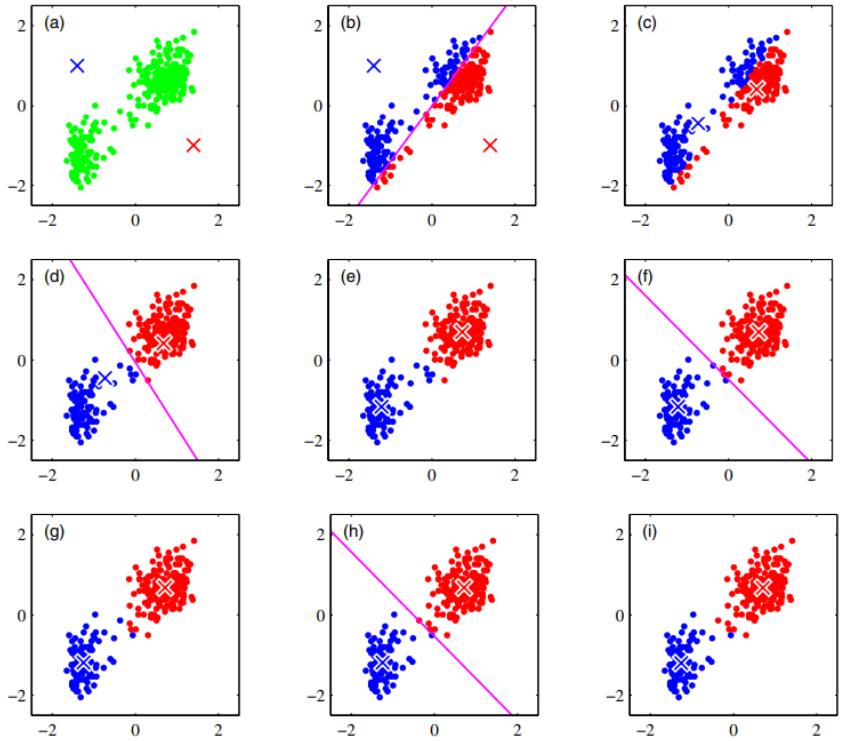

Figura 2 – Ilustração do algoritmo K-means utilizando o conjunto de dados Old Faithful.
Fonte: Adaptado de Bishop (2006).

A imagem evidencia o caráter iterativo do algoritmo: à medida que os centróides são atualizados, os pontos são continuamente realocados para o cluster cujo centróide está mais próximo. Esse mecanismo reforça a busca pela minimização da variabilidade interna, permitindo que o método identifique regiões naturais de concentração no espaço de atributos. Assim, ao final do procedimento, conforme observado no quadro (i), formam-se grupos mais compactos e bem separados.

5 MISTURA FINITAS DE DENSIDADES

Os métodos de agrupamento baseados em modelos são uma abordagem poderosa para identificar padrões em dados heterogêneos. Diferente de técnicas tradicionais, esses métodos atribuem probabilidades a cada observação, refletindo a incerteza na alocação dos grupos (Banfield e Raftery, 1993).

Uma das principais estratégias dentro dessa abordagem é o modelo de misturas finitas de densidades, no qual a distribuição dos dados é representada como uma combinação de múltiplas distribuições componentes. Isso permite capturar padrões complexos, como multimodalidade e assimetria, sendo útil quando uma única distribuição não descreve bem a estrutura dos dados (McLachlan e Peel, 2000). Apesar de suas vantagens, a aplicação de modelos de mistura envolve desafios relevantes, como a definição do número ótimo de componentes e a interpretação dos grupos formados, especialmente quando há sobreposição entre as distribuições.

Neste capítulo, apresentamos o conceito de misturas finitas de densidades, exploramos suas principais características e discutimos a estimação de parâmetros via algoritmo Expectation-Maximization (EM). Por fim, abordamos critérios para seleção de modelos e aplicações na classificação não supervisionada. Diferentemente dos demais métodos apresentados ao longo deste trabalho, em que adotamos K para representar o número de grupos, no contexto de misturas utilizaremos a notação G, mais tradicional na literatura da área.

5.1 MISTURA FINITAS DE DENSIDADES

Para mais informações detalhadas sobre misturas finitas de densidades, ver Everitt e Hand (1981), Titterington (1985), McLachlan e Basford (1988), McLachlan e Peel (2000) e Frühwirth-Schnatter (2006), Dávila, Cabral e Zeller (2018).

5.1.1 Definição

Considere um vetor aleatório $\mathbf{Y} \in \mathbb{R}^p$, cuja função densidade de probabilidade é expressa como:

$$f(\mathbf{y}) = \sum_{j=1}^{G} p_j \psi_j(\mathbf{y}), \quad p_j \geq 0 \quad \text{e} \quad \sum_{j=1}^{G} p_j = 1, \tag{5.1}$$

Nesse modelo, $f(\cdot)$ é denominada uma mistura finita de densidades e é composta por G componentes. Os parâmetros $p_1, \ldots, p_G$ representam as proporções associadas a cada componente, enquanto $\psi_1, \ldots, \psi_G$ correspondem às funções de densidade que constituem a mistura. Essas funções de densidade $\psi_j(\cdot)$ podem ser escritas de forma mais específicas considerando que pertencem a uma família paramétrica, resultando na seguinte forma:

$$f(\mathbf{y};\boldsymbol{\theta}) = \sum_{j=1}^{G} p_j \psi_j(\mathbf{y};\boldsymbol{\theta_j}), \tag{5.2}$$

onde, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_G^\top)^\top$, sendo $\boldsymbol{\theta}_j$ os parâmetros que definem cada componente ψ_j. É importante destacar que esses parâmetros não precisam, necessariamente, estar definidos no mesmo espaço paramétrico. No entanto, em muitas aplicações práticas, incluindo as abordadas neste trabalho, como as componentes ψ_j pertencem a uma mesma família paramétrica, os parâmetros $\boldsymbol{\theta_j}$ também estão definidos no mesmo espaço paramétrico.

5.1.2 Conceitos Fundamentais

Dois conceitos centrais na formulação e aplicação de modelos de misturas finitas de distribuições são a *identificabilidade* e a *distribuição marginal*.

A **identificabilidade** é uma propriedade essencial para assegurar a consistência estatística do modelo. Em termos práticos, um modelo é identificável quando diferentes conjuntos de parâmetros não produzem a mesma distribuição de probabilidade. Caso contrário, a ausência de identificabilidade pode resultar em estimadores inconsistentes e, consequentemente, comprometer a validade das inferências estatísticas. Além disso, a não identificabilidade introduz ambiguidade na interpretação dos resultados, uma vez que diferentes combinações de parâmetros podem gerar a mesma distribuição observada.

A **distribuição marginal**, por outro lado, diz respeito à preservação da estrutura da mistura quando se consideram subconjuntos das variáveis originalmente modeladas. Em outras palavras, mesmo ao reduzir o número de variáveis em análise, a mistura mantém suas G componentes, o que garante a coerência estrutural do modelo frente à projeção de variáveis. Esses conceitos irão ser explorados mais a fundo a seguir.

5.1.2.1 Distribuição Marginal

Conforme discutido por Basso (2009), o resultado a seguir mostra que a marginal de uma mistura também é uma mistura.

Teorema 5.1. *Seja* $\mathbf{Y}$ *um vetor aleatório em* $\mathbb{R}^b$*, cuja distribuição é uma mistura finita de densidades com* G *componentes. Então, a distribuição de qualquer vetor aleatório* $\mathbf{X}$*, formado por um subconjunto das variáveis de* $\mathbf{Y}$*, pertencente a* $\mathbb{R}^q$ *com* $q < b$*, também é uma mistura finita de densidades com* G *componentes.*

Demonstração: Seja $\mathbf{X}$ um vetor aleatório cujas variáveis componentes formam um subconjunto das variáveis de $\mathbf{Y}$. Então, a densidade de $\mathbf{X}$ é dada por

$$f_{\mathbf{X}}(\mathbf{x}) = \int f_{\mathbf{Y}}(\mathbf{y})\, d\mathbf{y}_{(\mathbf{X})}, \tag{5.3}$$

onde a integral é tomada nas componentes de $\mathbf{Y}$ que não pertencem a $\mathbf{X}$. Substituindo $f_{\mathbf{Y}}(\mathbf{y})$ pela representação da mistura, temos:

$$f_{\mathbf{X}}(\mathbf{x}) = \int \sum_{j=1}^{G} p_j \psi_j(\mathbf{y}) \, d\mathbf{y}_{(\mathbf{X})}. \tag{5.4}$$

Trocando a ordem da soma e da integral:

$$f_{\mathbf{X}}(\mathbf{x}) = \sum_{j=1}^{G} p_j \int \psi_j(\mathbf{y}) \, d\mathbf{y}_{(\mathbf{X})}. \tag{5.5}$$

Como $\int \psi_j(\mathbf{y}) \, d\mathbf{y}_{(\mathbf{X})}$ é a densidade marginal de ψ_j em $\mathbf{x}$, denotada por $\psi_j(\mathbf{x})$, temos:

$$f_{\mathbf{X}}(\mathbf{x}) = \sum_{j=1}^{G} p_j \psi_j(\mathbf{x}). \tag{5.6}$$

Portanto, $f_{\mathbf{X}}(\mathbf{x})$ também é uma mistura finita de densidades com G componentes, o que conclui a demonstração. □

5.1.2.2 Identificabilidade

A identificabilidade é uma propriedade fundamental na modelagem estatística, pois garante que diferentes configurações dos parâmetros resultem em distribuições distintas. Em uma família paramétrica $\mathcal{F}$ de funções densidade de probabilidade $\psi(\cdot; \boldsymbol{\theta})$, essa propriedade é satisfeita quando valores distintos de $\boldsymbol{\theta}$ correspondem a membros diferentes da família de distribuições (McLachlan e Peel, 2000). Ou seja, se $\psi(\cdot; \boldsymbol{\theta}_1) = \psi(\cdot; \boldsymbol{\theta}_2)$, então necessariamente $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

Entretanto, quando lidamos com modelos de mistura finita de distribuições, essa definição clássica precisa ser expandida. Isso se deve à possibilidade de se obter a mesma função densidade de mistura por meio de diferentes permutações dos componentes, como discutido em Titterington, Smith e Makov (1985). A permutabilidade dos índices dos componentes implica que duas representações diferentes da mistura podem ser estatisticamente indistinguíveis, apesar de utilizarem conjuntos distintos de parâmetros.

Para ilustrar essa ambiguidade, Duda e Hart (1988) apresentam o exemplo de uma mistura de duas distribuições normais com variância unitária, cuja função densidade é expressa como:

$$f(y; \boldsymbol{\theta}) = p_1 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y - \mu_1)^2\right) + p_2 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y - \mu_2)^2\right). \tag{5.7}$$

Neste caso, a troca de índices entre os componentes, isto é, permutar $\mu_1 \leftrightarrow \mu_2$ e $p_1 \leftrightarrow p_2$ não altera a função densidade resultante. Portanto, o modelo permanece estatisticamente idêntico, ainda que os parâmetros tenham sido modificados. Isso evidencia

que, em misturas finitas, a identificabilidade deve ser definida *até uma permutação* dos componentes.

Com o objetivo de lidar com esse desafio, a identificabilidade em modelos de mistura é formalmente definida a seguir.

Seja $\mathcal{F} = \{\psi(\mathbf{y};\boldsymbol{\theta}) : \mathbf{y} \in \mathbb{R}^p, \boldsymbol{\theta} \in \Omega\}$ uma família paramétrica de densidades, e

$$\mathcal{P} = \left\{ f(\mathbf{y};\boldsymbol{\theta}) : f(\mathbf{y};\boldsymbol{\theta}) = \sum_{j=1}^{N} p_j \psi(\mathbf{y};\boldsymbol{\theta}_j),\ p_j \geq 0,\ \sum_{j=1}^{N} p_j = 1,\ \psi(\mathbf{y};\boldsymbol{\theta}) \in \mathcal{F},\ \boldsymbol{\theta} = (\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta}_N) \right\}$$

uma classe de misturas finitas de densidades. Diz-se que $\mathcal{P}$ é identificável se, para quaisquer dois membros

$$f(\mathbf{y};\boldsymbol{\theta}) = \sum_{j=1}^{N} p_j \psi(\mathbf{y};\boldsymbol{\theta}_j) \quad \text{e} \quad f(\mathbf{y};\boldsymbol{\theta}') = \sum_{j=1}^{N'} p'_j \psi(\mathbf{y};\boldsymbol{\theta}'_j),$$

tem-se que $f(\mathbf{y};\boldsymbol{\theta}) = f(\mathbf{y};\boldsymbol{\theta}')$ se, e somente se, $N = N'$ e existe uma permutação dos índices tal que $p_j = p'_j$ e $\psi(\mathbf{y};\boldsymbol{\theta}_j) = \psi(\mathbf{y};\boldsymbol{\theta}'_j)$ para todo $j = 1, \ldots, N$.

Conforme discutido por McLachlan e Basford (1988), dificuldades adicionais surgem quando todas as componentes pertencem à mesma família de distribuições. Nesses casos, a densidade da mistura pode permanecer inalterada mesmo sob diversas permutações dos componentes, o que compromete a unicidade dos parâmetros estimados. Assim, embora a mistura como um todo seja identificável no sentido estatístico, o vetor de parâmetros $\boldsymbol{\theta}_j$ não é único, tornando a densidade invariável sob permutações dos índices dos componentes.

A ausência de identificabilidade pode, portanto, acarretar sérios desafios na prática estatística. A função de verossimilhança, por exemplo, pode apresentar múltiplos máximos locais associados a diferentes vetores $\boldsymbol{\theta}_j$, o que dificulta uma estimação confiável. De acordo com McLachlan e Krishnan (2007), esse problema torna-se especialmente relevante quando o interesse está na obtenção de estimativas específicas para cada parâmetro do modelo. Nesse contexto, torna-se necessário impor condições adicionais, como a distinção entre os parâmetros das componentes ou a adoção de restrições estruturais, a fim de assegurar a identificabilidade e garantir a validade inferencial dos modelos de mistura ajustados.

5.2 A ESTRUTURA LATENTE DOS MODELOS DE MISTURAS DE DENSIDADES

Em muitas situações práticas, os dados disponíveis para análise são intrinsecamente incompletos, significando que informações cruciais não estão diretamente observáveis. Nos modelos de mistura de densidades, essa incompletude reside na ausência de informação de agrupamento, ou seja, não se sabe a qual dos G grupos ou componentes subjacentes cada observação $\mathbf{y}_i$ pertence. Por exemplo, em um estudo sobre o desempenho de estudantes

em uma prova, as notas $\mathbf{y}_i$ são observadas, mas a classificação de cada estudante em níveis de proficiência (baixa, média, alta) é desconhecida. Essa lacuna torna o conjunto de dados observados $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ incompleto.

Para contornar essa limitação e viabilizar a estimação, introduz-se a estrutura de dados completos, que engloba tanto os dados observados $\mathbf{Y}$ quanto os dados latentes não observáveis $\mathbf{Z}$. O conjunto de dados completos é, portanto, $\mathbf{Y}_c = (\mathbf{Y}^\top, \mathbf{Z}^\top)^\top$. O elemento central dessa formulação é a variável latente Z_i, que identifica o componente da mistura responsável pela geração da observação $\mathbf{Y}_i$. Essa variável assume valores no conjunto $\{1, 2, \ldots, G\}$, com probabilidades associadas $p_1, p_2, \ldots, p_G$, satisfazendo $\sum_{j=1}^{G} p_j = 1$. Além disso, assume-se que a densidade de $\mathbf{Y}_i$ condicional a $Z_i = j$ é dada por $\psi_j(\mathbf{y}_i)$, para $j = 1, \ldots, G$, conforme apresentado na expressão (5.1). Assim, Z_i desempenha o papel de uma variável latente que indica de qual componente da mistura a observação $\mathbf{Y}_i$ foi gerada. Para simplificar a notação e permitir formulações algébricas mais convenientes, define-se um vetor indicador binário de dimensão G, denotado por $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, \ldots, Z_{iG})^T$, com $\mathbf{Z}_i \sim \text{Multinomial}(1; p_1, p_2, \ldots, p_G)$, tal que,

$$Z_{ij} = \begin{cases} 1, & \text{se } \mathbf{Y}_i \text{ pertence ao componente } j \\ 0, & \text{caso contrário} \end{cases} \tag{5.8}$$

A distribuição de $\mathbf{Z}_i$ é dada por,

$$\Pr(Z_{i1} = z_{i1}, \ldots, Z_{iG} = z_{iG}) = \prod_{j=1}^{G} p_j^{z_{ij}}, \qquad \text{com} \quad \sum_{j=1}^{G} z_{ij} = 1. \tag{5.9}$$

Condicionalmente ao evento $Z_{ij} = 1$, tem-se $Y_i \mid (Z_{ij} = 1) \sim \psi_j(\cdot; \theta_j)$. Assim, a densidade condicional de Y_i dado Z_i pode ser expressa como,

$$f_{Y_i|Z_i}(y_i \mid z_i) = \prod_{j=1}^{G} \left[\psi_j(y_i; \theta_j)\right]^{z_{ij}}. \tag{5.10}$$

Assim, a densidade conjunta de Y_i dado Z_i pode ser escrita como,

$$f_{Y_i,Z_i}(y_i, z_i) = \Pr(Z_i = z_i)\, f_{Y_i|Z_i}(y_i \mid z_i). \tag{5.11}$$

Substituindo as expressões anteriores, obtém-se,

$$f_{Y_i,Z_i}(y_i, z_i) = \left(\prod_{j=1}^{G} p_j^{z_{ij}}\right) \left(\prod_{j=1}^{G} \psi_j(y_i; \theta_j)^{z_{ij}}\right).$$

Como ambos os produtos percorrem o mesmo índice j, segue que,

$$f_{Y_i,Z_i}(y_i, z_i) = \prod_{j=1}^{G} \left[p_j\, \psi_j(y_i; \theta_j)\right]^{z_{ij}}. \tag{5.12}$$

Assumindo independência entre os pares (Y_i, Z_i), a verossimilhança completa é dada por,

$$L_c(\theta) = \prod_{i=1}^{n} f_{Y_i,Z_i}(y_i, z_i) = \prod_{i=1}^{n}\prod_{j=1}^{G} \left[p_j\, \psi_j(y_i;\theta_j)\right]^{z_{ij}} . \tag{5.13}$$

onde $\theta = (p_1, \ldots, p_G, \theta_1, \ldots, \theta_G)$ representa o vetor completo de parâmetros do modelo. Aplicando o logaritmo natural à expressão (5.13), obtém-se a log-verossimilhança completa,

$$\ell_c(\theta) = \log L_c(\theta) = \sum_{i=1}^{n}\sum_{j=1}^{G} \log\left(\left[p_j\, \psi_j(y_i;\theta_j)\right]^{z_{ij}}\right) . \tag{5.14}$$

Utilizando as propriedades $\log(a^b) = b\log(a)$ e $\log(ab) = \log a + \log b$, chega-se a

$$\ell_c(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{G} z_{ij}\left[\log p_j + \log \psi_j(\mathbf{y}_i;\boldsymbol{\theta}_j)\right] . \tag{5.15}$$

Essa formulação aditiva, separando explicitamente os termos associados aos pesos de mistura p_j e aos parâmetros específicos de cada componente $\boldsymbol{\theta}_j$, é fundamental para a etapa de Maximização (M) do algoritmo EM. Essa estutura será discutida na próxima seção.

5.3 O ALGORITMO EM PARA MODELOS DE MISTURAS

Conforme apresentado por McLachlan e Peel (2000), o algoritmo EM é um método iterativo para estimar parâmetros em modelos com variáveis latentes, como os modelos de mistura de distribuições. Ele é composto por duas etapas fundamentais: na etapa E, calcula-se a esperança da log-verossimilhança completa, condicionada aos dados observados e aos valores atuais dos parâmetros; na etapa M, maximiza-se essa esperança em relação aos parâmetros. Esse ciclo E-M é então iterado sucessivamente até que os valores estimados atinjam convergência.

5.3.1 Etapa E (Esperança)

Nesta etapa, os dados não observáveis, representados pela variável Z_{ij}, são tratados por meio do cálculo da esperança da log-verossimilhança, $\ell_c(\boldsymbol{\theta})$, condicionada aos valores observados $\mathbf{y}$ e aos parâmetros estimados na iteração anterior, $\boldsymbol{\theta}^{(k)}$. Dessa forma, obtém-se a função Q dada por:

$$Q(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(k)}) = \mathbb{E}\left[\ell_c(\boldsymbol{\theta}) \mid \mathbf{y}, \widehat{\boldsymbol{\theta}}^{(k)}\right] \tag{5.16}$$

$$= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{G} Z_{ij}\left(\log(p_j) + \log(\psi_j(\mathbf{y}_i;\boldsymbol{\theta}_j))\right) \middle| \mathbf{y}, \widehat{\boldsymbol{\theta}}^{(k)}\right] \tag{5.17}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{G} \mathbb{E}\left[Z_{ij} \mid \mathbf{y}; \widehat{\boldsymbol{\theta}}^{(k)} \right] \left(\log(p_j) + \log(\psi_j(\mathbf{y}_i; \boldsymbol{\theta}_j)) \right). \tag{5.18}$$

Como a log-verossimilhança completa é linear em relação às variáveis latentes Z_{ij}, a etapa E se reduz ao cálculo da esperança condicional desses indicadores, isto é, das probabilidades a posteriori de que cada observação $\mathbf{y}_i$ pertença a j-ésima componente da mistura. Assim definimos em (5.19),

$$\mathbb{E}\left[Z_{ij} \mid \mathbf{y}; \widehat{\boldsymbol{\theta}}^{(k)} \right] = P(Z_{ij} = 1 \mid \mathbf{y}; \widehat{\boldsymbol{\theta}}^{(k)}) = \widehat{z}_{ij}. \tag{5.19}$$

onde,

$$\hat{z}_{ij} = \frac{\hat{p}_j^{(k)} \psi(\mathbf{y}_i \mid \hat{\boldsymbol{\theta}}_j^{(k)})}{\sum_{l=1}^{G} \hat{p}_l^{(k)} \psi(\mathbf{y}_i \mid \hat{\boldsymbol{\theta}}_l^{(k)})}, \tag{5.20}$$

com $i = 1, \ldots, n$ e $j = 1, \ldots, G$. Dessa forma, a expressão da função $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ pode ser escrita como:

$$Q(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(k)}) = \sum_{i=1}^{n} \sum_{j=1}^{N} \hat{z}_{ij} \left[\log p_j^{(k)} + \log \psi(y_i \mid \boldsymbol{\theta}_j^{(k)}) \right]. \tag{5.21}$$

5.3.2 Etapa M (Maximização)

Na etapa M, busca-se maximizar a função $Q(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(k)})$ em relação ao vetor de parâmetros $\boldsymbol{\theta}$, de modo a obter as novas estimativas para a iteração seguinte, isto é, $\hat{\boldsymbol{\theta}}^{(k+1)}$. Para isso, definimos $n_j = \sum_{i=1}^{n} z_{ij}$, o que permite reescrever a expressão da log-verossimilhança completa em (5.15) como

$$\ell_c(\boldsymbol{\theta}) = \sum_{j=1}^{G} n_j \log p_j + \sum_{j=1}^{G} \sum_{i=1}^{n} z_{ij} \log \psi_j(\mathbf{y}_i; \boldsymbol{\theta}_j). \tag{5.22}$$

Essa decomposição mostra que os pesos de mistura que maximizam a log-verossimilhança completa são dados por

$$\hat{p}_j = \frac{n_j}{n}. \tag{5.23}$$

Contudo, como os indicadores latentes Z_{ij} não são observáveis, o algoritmo EM substitui seus valores pelas esperanças condicionais $\hat{z}_{ij}$, obtidas na etapa E da iteração k. Assim, a atualização das proporções de mistura na iteração $k+1$ é dada por

$$\hat{p}_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} \hat{z}_{ij}. \tag{5.24}$$

Essa estimativa de p_j é obtida por meio da maximização da função (5.18), sob a restrição de que as proporções de mistura devem somar 1, isto é, $\sum_{j=1}^{G} p_j = 1$. Para lidar com essa

restrição, utiliza-se o método dos multiplicadores de Lagrange, que permite incorporar a condição de soma unitária diretamente no processo de otimização. Dessa forma, encontra-se a expressão analítica apresentada em (5.24) para a estimação de p_j.

O algoritmo EM descrito nesta seção corresponde à sua formulação geral, na qual se consideram como dados aumentados apenas as variáveis latentes Z_{ij}. É importante destacar que, por se tratar de um procedimento iterativo, o algoritmo repete suas etapas até que seja atingido um critério de convergência previamente estipulado. Nesse trabalho, iremos considerar que a convergência é alcançada quando a variação relativa da log-verossimilhança dos dados observados entre duas iterações consecutivas torna-se suficientemente pequena, isto é:

$$\left\| \frac{\ell(\widehat{\boldsymbol{\theta}}^{(k)} \mid \mathbf{y}) - \ell(\widehat{\boldsymbol{\theta}}^{(k+1)} \mid \mathbf{y})}{\ell(\widehat{\boldsymbol{\theta}}^{(k+1)} \mid \mathbf{y})} \right\| < \epsilon, \tag{5.25}$$

em que ϵ é um valor arbitrariamente pequeno, indicando a estabilização da sequência de valores de log-verossimilhança. Adicionalmente, Dempster, Laird e Rubin (1977) demonstraram que a função de verossimilhança dos dados observados não decresce a cada iteração do algoritmo EM, ou seja,

$$\left\| \ell(\boldsymbol{\theta}^{(k+1)} \mid \mathbf{y}) \geq \ell(\boldsymbol{\theta}^{(k)} \mid \mathbf{y}) \right\|, \tag{5.26}$$

para $k = 0, 1, 2, \ldots$. Esse resultado garante que a sequência de valores da verossimilhança observada gerada pelo algoritmo EM é monotonicamente não decrescente e converge para um ponto estacionário da função de verossimilhança. Esse ponto estacionário pode corresponder a um máximo local, a um ponto de sela ou, eventualmente, ao máximo global. Em modelos de mistura, a verossimilhança é geralmente multimodal, de modo que o algoritmo não assegura a obtenção do máximo global, embora a sequência seja limitada superiormente nas aplicações práticas.

5.4 MISTURAS FINITAS DE MODELOS NORMAIS MULTIVARIADOS PARCIMONIOSOS

No presente trabalho, optou-se pela utilização dos Modelos Gaussianos Parcimoniosos. Contudo, antes de apresentá-los, introduzimos inicialmente o modelo geral de misturas de distribuições normais. Nesse contexto, assume-se que a densidade dos dados observados pode ser representada como uma combinação ponderada de G distribuições normais multivariadas. Para ilustrar essa formulação, consideremos que cada componente segue uma distribuição normal q-variada, de modo que o j-ésimo componente possui densidade $N_q(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, com peso p_j, em que $\boldsymbol{\mu}_j$ e $\boldsymbol{\Sigma}_j$ representam, respectivamente, o vetor de médias e a matriz de covariância, para $j = 1, \ldots, G$.

Sob essa suposição, a etapa M do algoritmo EM consiste em maximizar a função

$$\sum_{i=1}^{n}\sum_{j=1}^{G}\hat{z}_{ij}^{(k+1)}\log\psi_q(\mathbf{y}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \tag{5.27}$$

em relação a $\boldsymbol{\mu}_j$ e $\boldsymbol{\Sigma}_j$, onde $\psi_q(\mathbf{y}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ denota a densidade normal multivariada $N_q(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$.

Fixado um componente j, essa maximização reduz-se à log-verossimilhança ponderada:

$$\sum_{i=1}^{n}\hat{z}_{ij}^{(k+1)}\log\psi_q(\mathbf{y}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j). \tag{5.28}$$

A solução desse problema é conhecida e admite forma fechada. As atualizações para o vetor de médias do componente j são dadas por:

$$\hat{\boldsymbol{\mu}}_j^{(k+1)} = \frac{\sum_{i=1}^{n}\hat{z}_{ij}^{(k+1)}\mathbf{y}_i}{\sum_{i=1}^{n}\hat{z}_{ij}^{(k+1)}}. \tag{5.29}$$

A matriz de covariância é atualizada conforme:

$$\hat{\boldsymbol{\Sigma}}_j^{(k+1)} = \frac{\sum_{i=1}^{n}\hat{z}_{ij}^{(k+1)}\left(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_j^{(k+1)}\right)\left(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_j^{(k+1)}\right)^{\top}}{\sum_{i=1}^{n}\hat{z}_{ij}^{(k+1)}}, \qquad j = 1, \ldots, G. \tag{5.30}$$

Dando continuidade à discussão, ao introduzirmos os Modelos Gaussianos Parcimoniosos (GPCM) no contexto de misturas de distribuições normais multivariadas, é importante destacar que cada componente da mistura possui sua própria matriz de covariância. Nesse sentido, Celeux e Govaert (1995) propuseram uma decomposição útil dessa matriz, que permite interpretar separadamente três aspectos geométricos fundamentais de cada grupo: volume, forma e orientação. Essa decomposição é dada por:

$$\boldsymbol{\Sigma}_g = \lambda_g, \mathbf{D}_g, \mathbf{A}_g, \mathbf{D}_g^{\top}, \tag{5.31}$$

em que $\lambda_g > 0$ é o fator de volume, $\mathbf{D}_g$ é uma matriz ortogonal que define a orientação, e $\mathbf{A}_g$ é uma matriz diagonal de forma, sujeita à restrição $\det(\mathbf{A}_g) = 1$. Essa parametrização constitui a base dos 14 modelos gaussianos parcimoniosos.

O escalar λ_g representa o volume do agrupamento.Ele corresponde à média geométrica dos autovalores originais $\alpha_{g1}, \ldots, \alpha_{gp}$ da matriz de covariância:

$$\lambda_g = \left(\prod_{j=1}^{p}\alpha_{gj}\right)^{1/p}. \tag{5.32}$$

Assim, λ_g atua como um fator global de escala: valores mais altos indicam grupos mais dispersos, enquanto valores menores correspondem a grupos mais compactos.

A matriz $\mathbf{D}_g$ é uma matriz ortogonal ($\mathbf{D}_g^\top \mathbf{D}_g = \mathbf{I}$) composta pelos autovetores de $\mathbf{\Sigma}_g$. Ela determina a orientação dos eixos principais do elipsoide de densidade associado ao componente g.Em termos geométricos, $\mathbf{D}_g$ realiza uma rotação no espaço, alinhando o elipsoide às direções de maior variabilidade dos dados.

Por fim, a matriz $\mathbf{A}_g$ é diagonal e contém os autovalores normalizados de $\mathbf{\Sigma}_g$:

$$\mathbf{A}g = \frac{\text{diag}(\alpha g1, \ldots, \alpha_{gp})}{\left(\prod_{j=1}^{p} \alpha_{gj}\right)^{1/p}}. \tag{5.33}$$

A normalização garante $\det(\mathbf{A}_g) = 1$. Dessa forma, $\mathbf{A}_g$ controla exclusivamente a forma relativa do elipsoide,isto é, como ele se alonga ou se achata em cada direçãos,em influenciar o volume global. Quando todos os elementos diagonais são iguais, obtém-se um grupo esférico, para o qual $\mathbf{A}_g = \mathbf{I}$.

A decomposição da matriz de covariância $\mathbf{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g^\top$ pode ser entendida como uma transformação progressiva da bola unitária, onde:

1. **Aplicação de $\mathbf{A}_g$:** Transforma a esfera em um elipsoide, definindo a forma do agrupamento.
2. **Aplicação $\mathbf{D}_g$:** Rotaciona o elipsoide no espaço, definindo a orientação.
3. **Aplicação λ_g:** Expande ou contrai o elipsoide, definindo o volume.

Essa separação geométrica torna possível a construção dos modelos gaussianos parcimoniosos. A parcimônia decorre da possibilidade de restringir (E = equal) ou permitir variar (V = variable) os elementos λ_g, $\mathbf{A}_g$ e $\mathbf{D}_g$ entre os G componentes da mistura, resultando nos 14 modelos apresentados na Tabela 1. Esses modelos constituem a base conceitual das implementações do pacote `Mclust` (Scrucca et al., 2016).

Tabela 1 – Modelos Gaussianos Parcimoniosos (GPCM) disponíveis no pacote `Mclust`.

Modelo	$\boldsymbol{\Sigma}_g$	Distribuição	Volume	Forma	Orientação
EII	$\lambda\mathbf{I}$	Esférica	Igual	Igual	—
VII	$\lambda_g\mathbf{I}$	Esférica	Variável	Igual	—
EEI	$\lambda\mathbf{A}$	Diagonal	Igual	Igual	Eixos coordenados
VEI	$\lambda_g\mathbf{A}$	Diagonal	Variável	Igual	Eixos coordenados
EVI	$\lambda\mathbf{A}_g$	Diagonal	Igual	Variável	Eixos coordenados
VVI	$\lambda_g\mathbf{A}_g$	Diagonal	Variável	Variável	Eixos coordenados
EEE	$\lambda\mathbf{D}\mathbf{A}\mathbf{D}^\top$	Elipsoidal	Igual	Igual	Igual
EVE	$\lambda\mathbf{D}_g\mathbf{A}\mathbf{D}_g^\top$	Elipsoidal	Igual	Variável	Igual
VEE	$\lambda_g\mathbf{D}\mathbf{A}\mathbf{D}^\top$	Elipsoidal	Variável	Igual	Igual
VVE	$\lambda_g\mathbf{D}_g\mathbf{A}\mathbf{D}_g^\top$	Elipsoidal	Variável	Variável	Igual
EEV	$\lambda\mathbf{D}\mathbf{A}_g\mathbf{D}^\top$	Elipsoidal	Igual	Igual	Variável
VEV	$\lambda_g\mathbf{D}\mathbf{A}_g\mathbf{D}^\top$	Elipsoidal	Variável	Igual	Variável
EVV	$\lambda\mathbf{D}_g\mathbf{A}_g\mathbf{D}_g^\top$	Elipsoidal	Igual	Variável	Variável
VVV	$\lambda_g\mathbf{D}_g\mathbf{A}_g\mathbf{D}_g^\top$	Elipsoidal	Variável	Variável	Variável

Fonte: Adaptado de Celeux e Govaert (1995) e Scrucca et al. (2016).

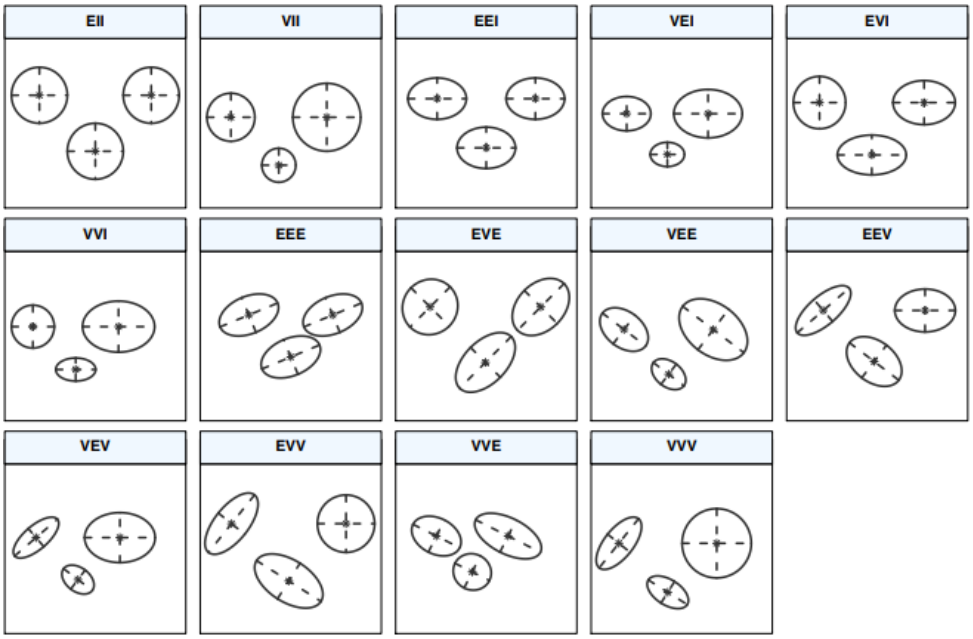

Figura 3 – Elipses de isodensidade para cada um dos 14 modelos Gaussianos obtidos por decomposição em autovalores, no caso de três grupos em duas dimensões.
Fonte: Scrucca et al. (2016).

A decomposição da expressão (5.31) estabelece, portanto, uma base estatística robusta e interpretável, separando claramente os três aspectos fundamentais da distribuição multivariada: volume, forma e orientação. Essa estrutura é essencial para a seleção do modelo de mistura mais adequado aos dados, otimizando o equilíbrio entre parcimônia e flexibilidade. Por fim, vale destacar que utilizaremos o critério BIC para selecionar

automaticamente o melhor entre os 14 modelos, considerando tanto o ajuste quanto a complexidade do modelo. O BIC será discutido com mais detalhes no Capítulo 6.

5.5 IMPLEMENTAÇÃO VIA PACOTE MCLUST

No presente trabalho, a estimação dos modelos gaussianos parcimoniosos foi realizada por meio do pacote `mclust`. Esse pacote é amplamente utilizado na linguagem R para tarefas de agrupamento, classificação e estimação de densidades baseadas em modelos de misturas finitas de distribuições gaussianas. Desenvolvido originalmente por Fraley e Raftery (2002), o `mclust` consolidou-se como uma das ferramentas mais completas para análise de dados multivariados sob a abordagem de modelos baseados em mistura, oferecendo algoritmos eficientes, critérios de seleção de modelos e uma implementação abrangente dos 14 modelos gaussianos parcimoniosos.

Além disso, o pacote incorpora metodologias modernas de inicialização, diagnóstico e validação, o que o torna particularmente adequado para aplicações práticas em estatística multivariada e aprendizado de máquina. O `mclust` está disponível gratuitamente no *Comprehensive R Archive Network* (CRAN), onde é continuamente atualizado e mantido.

O pacote mclust ajusta modelos de misturas gaussianas por meio do algoritmo EM (Expectation-Maximization). Como o EM é sensível ao ponto inicial, o desempenho e a convergência do ajuste dependem crucialmente da escolha de valores iniciais apropriados para as proporções de mistura, médias e matrizes de covariância. Por padrão, o mclust emprega o procedimento Model-Based Hierarchical Agglomerative Clustering (MBHAC), proposto por Banfield e Raftery (1993), para determinação das classificações iniciais. A seguir descrevemos em detalhes como esse procedimento funciona.

O MBHAC é um método hierárquico aglomerativo baseado em modelos, construído a partir de uma formulação probabilística de misturas gaussianas. Seu objetivo é gerar uma sequência hierárquica de fusões entre grupos, iniciando com cada observação representada como um cluster individual e, em seguida, combinando iterativamente os dois grupos cuja fusão provoca a menor redução na verossimilhança de classificação, conceito que será detalhado adiante.

Dado um conjunto de dados $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, o procedimento do *MBHAC* pode ser descrito da seguinte forma:

1. Cada observação $\mathbf{y}_i$ inicia como um *cluster* próprio.

2. Para cada par de *clusters* (a, b), calcula-se a diminuição na verossimilhança de classificação (ℓ_{class}) caso esses dois grupos fossem fundidos.

3. Funde-se o par cuja fusão resulta na menor queda de ℓ_{class}.

4. O processo é repetido até restar apenas um *cluster*, originando uma árvore hierárquica orientada por modelo (*model-based*).

Para um número desejado de componentes G, o mclust realiza o corte dessa árvore no nível apropriado e extrai uma classificação inicial dura (*hard*), na qual $z_{ij} \in \{0, 1\}$ indica a afiliação da observação i ao cluster j.

A partir dessa classificação, o pacote computa as estimativas iniciais dos parâmetros do modelo ($\boldsymbol{\theta}^{(0)}$):

$$\hat{p}_j^{(0)} = \frac{n_j}{n}, \tag{5.34}$$

$$\hat{\boldsymbol{\mu}}_j^{(0)} = \frac{1}{n_j} \sum_{i:z_{ij}=1} \mathbf{y}_i, \tag{5.35}$$

$$\hat{\boldsymbol{\Sigma}}_j^{(0)} = \mathrm{cov}\{\mathbf{x}_i : z_{ij} = 1\}, \tag{5.36}$$

onde $n_j = \sum_{i=1}^{n} z_{ij}$ representa o tamanho do *cluster* j. Posteriormente, a matriz de covariância $\hat{\boldsymbol{\Sigma}}_j^{(0)}$ é ajustada de acordo com a parametrização de parcimônia considerada (por exemplo, modelos EEE, VVV etc.). Esses valores constituem o ponto de partida do algoritmo EM.

A verossimilhança de classificação é uma versão modificada da verossimilhança usual de misturas, construída sob a suposição de que as classificações são conhecidas (classificação dura). Para

$$Z_{ij} = \begin{cases} 1, & \text{se } \mathbf{y}_i \text{ pertence ao cluster } j, \\ 0, & \text{caso contrário,} \end{cases}$$

define-se:

$$L_{\text{class}} = \prod_{i=1}^{n} \prod_{j=1}^{G} \left[p_j \, \psi(\mathbf{y}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \right]^{z_{ij}}, \tag{5.37}$$

onde $\psi(\cdot)$ denota a densidade gaussiana do componente j.

A log-verossimilhança correspondente é:

$$\ell_{\text{class}} = \sum_{i=1}^{n} \sum_{j=1}^{G} z_{ij} \left(\log p_j + \log \psi(\mathbf{y}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \right). \tag{5.38}$$

Essa forma de verossimilhança é particularmente conveniente no MBHAC porque é separável por clusters, o que permite avaliar de maneira eficiente o impacto de uma fusão. Em cada etapa, o par escolhido é aquele cuja união causa a menor redução em ℓ_{class}, produzindo uma hierarquia coerente com a estrutura probabilística do modelo de mistura.

Após a etapa de inicialização via MBHAC, o algoritmo EM é utilizado para refinar as estimativas do modelo. Diferentemente do MBHAC, o EM trabalha com classificações suaves (*soft*), expressas pelas responsabilidades:

$$\hat{z}_{ij} = \frac{\hat{p}_j \psi(\mathbf{y}_i \mid \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)}{\sum_{l=1}^{G} \hat{p}_l \psi(\mathbf{y}_i \mid \hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}_l)}, \tag{5.39}$$

as quais satisfazem $0 < \hat{z_{ij}} < 1$ e $\sum_{j=1}^{G} z_{ij} = 1$.

Ao término da estimação, o `mclust` retorna tanto essas probabilidades a posteriori quanto uma classificação dura obtida pelo critério MAP:

$$\hat{c}(i) = \arg \max_{j \in \{1,\ldots,G\}} \hat{z}_{ij}.$$

onde $\hat{c}_i$ representa a afiliação final da observação ao componente cuja probabilidade a posteriori é máxima. Assim, o vetor de classificações finais é dado por $\hat{c} = (\hat{c}_1, \ldots, \hat{c}_n)$, com $\hat{c}_i \in \{1, \ldots, G\}$.

A classificação MAP pode ser interpretada como o estimador de Bayes sob perda 0–1, pois minimiza o risco esperado quando o objetivo é selecionar o componente mais provável para cada observação.

É importante destacar que a classificação final não faz parte da maximização da verossimilhança: ela é uma etapa posterior ao ajuste do modelo. O EM maximiza a verossimilhança marginal do modelo de mistura, produzindo responsabilidades probabilísticas z_{ij} (*soft*) que são consistentes com o ajuste de Máxima Verossimilhança. A classificação MAP é apenas uma representação **discreta** (*hard*) baseada na regra de decisão:

$$\text{Classificar } \mathbf{y}_i \text{ no componente } j \quad \Longleftrightarrow \quad \mathbb{P}(j \mid \mathbf{y}_i) \text{ é máximo.}$$

Dessa forma, o MBHAC fornece a inicialização por meio de classificações duras hierárquicas, enquanto o EM realiza o ajuste final com base em classificações *soft*. A classificação dura devolvida pelo `mclust` resulta da aplicação da regra MAP às responsabilidades estimadas, oferecendo uma interpretação clara e coerente da estrutura de grupos inferida pelo modelo, sem alterar os parâmetros ajustados.

5.5.1 Exemplificação do procedimento

O funcionamento dos modelos de misturas de distribuições pode ser estruturado nas seguintes etapas principais:

- **Inicialização (MBHAC):** O `mclust` inicia o processo utilizando o *Model-Based Hierarchical Agglomerative Clustering* (MBHAC). Esse procedimento constrói uma hierarquia de fusões entre grupos, escolhendo sempre a combinação que menos deteriora a verossimilhança. A partir dessa árvore hierárquica, o algoritmo extrai classificações iniciais para vários valores possíveis de G, fornecendo um ponto de partida consistente para o EM.

- **Etapa E (Esperança):** Com os parâmetros iniciais, o algoritmo calcula, para cada observação, a probabilidade de pertencer a cada componente. Essas probabilidades são chamadas de responsabilidades e representam uma classificação *soft*, em que cada ponto tem graus de pertencimento aos grupos.

- **Etapa M (Maximização):** Usando as responsabilidades, o algoritmo atualiza os parâmetros do modelo (proporções dos grupos, médias e matrizes de covariância). No `mclust`, essa atualização respeita as estruturas de covariância definidas pelos modelos parsimoniosos (como EII, VII, VVV etc.), ajustando apenas os parâmetros permitidos em cada modelo.

- **Iteração EM:** As etapas E e M são repetidas até que os parâmetros parem de mudar de forma significativa, indicando convergência. Esse procedimento é realizado para diferentes valores de G e para todas as estruturas de covariância possíveis.

- **Escolha do modelo (BIC):** Para cada combinação de G e estrutura de covariância, o `mclust` calcula o valor do BIC. O modelo escolhido é aquele que maximiza o BIC, equilibrando qualidade de ajuste e complexidade.

6 ESCOLHA DE NÚMERO DE GRUPOS

A definição do número de grupos constitui uma etapa fundamental em qualquer análise de agrupamento, pois influencia diretamente tanto a qualidade das partições obtidas quanto a interpretação dos resultados. Determinar a quantidade ideal de clusters é um dos desafios mais recorrentes na área, como destacado por Sarle (1983). A maioria dos algoritmos de agrupamento amplamente utilizados requer que o número de grupos seja especificado previamente, assim como a métrica de distância a ser empregada no processo de particionamento.

Essa decisão exerce impacto direto sobre o desempenho e a representatividade do modelo, uma vez que afeta a capacidade do método de capturar a estrutura intrínseca dos dados. Uma escolha inadequada pode gerar agrupamentos pouco informativos ou induzir interpretações equivocadas. Por exemplo, um número reduzido de clusters pode mascarar subestruturas relevantes, enquanto um número excessivo pode resultar em divisões artificiais e de difícil interpretação (Hastie et al., 2009; Kaufman e Rousseeuw, 1990).

Diversas abordagens foram propostas na literatura para auxiliar na definição do número ótimo de clusters, cada uma oferecendo uma perspectiva distinta sobre o equilíbrio entre ajuste e complexidade do modelo. Nesta seção, são discutidos os principais critérios utilizados para essa finalidade, considerando os métodos empregados neste trabalho: Ward, K-means e Misturas Finitas de Gaussianas.

No método de Ward, a escolha do número de grupos é feita ao final do processo hierárquico de fusão, sendo o dendrograma a principal ferramenta visual para essa decisão. O corte do dendrograma em um determinado nível de distância permite identificar a configuração de clusters que melhor representa a estrutura dos dados. Critérios baseados no comportamento dos níveis de fusão ou de similaridade, como os propostos por Giolo (2017), podem auxiliar na determinação do ponto de corte mais adequado.

No caso do K-means, o número de grupos deve ser definido antes da execução do algoritmo, o que torna essa escolha uma etapa crucial. Entre os critérios mais utilizados estão o método do cotovelo e a análise do coeficiente de silhueta, que buscam equilibrar a variabilidade explicada e a simplicidade da partição, evitando tanto a subsegmentação quanto o excesso de grupos artificiais.

Por fim, para os modelos de Misturas Finitas de Gaussianas, a determinação do número de componentes foi realizada com base no Critério de Informação Bayesiano (BIC). Esse critério penaliza modelos excessivamente complexos, promovendo a parcimônia e prevenindo o sobreajuste, ao mesmo tempo em que preserva a qualidade do ajuste aos dados.

Assim, cada técnica apresenta uma abordagem específica para determinar o número

ideal de grupos, variando entre métodos visuais, empíricos e baseados em critérios de informação. As subseções seguintes detalham os procedimentos adotados em cada caso.

6.1 ANÁLISE DO COMPORTAMENTO DO NÍVEL DE FUSÃO

Conforme o algoritmo de agrupamento hierárquico avança de um estágio K para o estágio $K+1$, observa-se uma diminuição progressiva na similaridade entre os grupos que estão sendo unidos. Consequentemente, a distância entre os grupos que se fundem em cada etapa do processo aumenta. Ao representar graficamente o passo do agrupamento (ou o número de grupos) em função do nível de distância (ou fusão) de cada etapa, é possível identificar pontos em que ocorrem aumentos significativos nas distâncias, conhecidos como *saltos*. Esses *pontos de salto* indicam potenciais momentos de interrupção do processo de agrupamento, sugerindo o número ideal de grupos K e a composição final dos clusters. Quando a função de distância exibe vários desses pontos de salto, é recomendável analisar uma faixa de valores para K, considerando que esses pontos delimitam uma região de valores promissores para o número de grupos. A visualização desses pontos de aumento abrupto nas distâncias, bem como a perda acentuada de similaridade, é essencial para uma escolha fundamentada do número de grupos finais.

6.2 ANÁLISE DO COMPORTAMENTO DO NÍVEL DE SIMILARIDADE

A análise do nível de similaridade em cada etapa do processo de agrupamento é uma técnica útil para determinar o ponto ideal de interrupção do algoritmo. Essa abordagem busca identificar saltos acentuados na similaridade entre os grupos que estão sendo unidos. Esses saltos indicam que a similaridade entre os grupos restantes está diminuindo significativamente, sugerindo que o processo de fusão deve ser interrompido nesse ponto, conforme discutido por Giolo (2017).

Sejam C_i e C_l os grupos que se unem em uma etapa específica do algoritmo. O nível de similaridade entre esses grupos é definido pela seguinte equação:

$$S_{il} = 1 - \frac{d_{il}}{\max\{d_{jr} : j, r = 1, 2, \ldots, n\}}, \tag{6.1}$$

onde d_{il} representa a distância entre os grupos C_i e C_l, e $\max\{d_{jr} : j, r = 1, 2, \ldots, n\}$ corresponde à maior distância encontrada na matriz de distância $D_{n\times n}$, calculada no início do processo de agrupamento. Em outras palavras, a similaridade S_{il} mede o quão próximos estão os grupos C_i e C_l em relação à maior distância inicial, padronizando o valor entre 0 e 1.

O número final de grupos, K, é determinado pelo estágio em que o algoritmo é interrompido, ou seja, no ponto em que ocorre uma queda acentuada na similaridade entre os grupos que estão se unindo. Para definir o número adequado de grupos, o

usuário deve estabelecer um intervalo de similaridade apropriado, assegurando que o agrupamento resultante seja consistente com os objetivos da análise. Dessa forma, a análise de similaridade fornece uma referência prática para identificar a melhor partição dos dados.

6.3 MÉTODO DO COTOVELO

Para o algoritmo K-means, um dos métodos mais comuns para determinar o número adequado de clusters é o critério do cotovelo. Esse método observa a variação explicada pelo agrupamento em função do número de clusters, buscando identificar um ponto em que a adição de novos clusters não resulta em uma redução significativa da variabilidade residual. Esse ponto, onde ocorre uma mudança abrupta na curva, é conhecido como o *cotovelo* da curva (Bholowalia e Kumar, 2014).

A lógica por trás do método do cotovelo está relacionada ao conceito de soma de quadrados dentro dos clusters (SQ_k), que mede a compactação dos clusters formados. A soma de quadrados total dentro dos grupos (SQ_{total}) é calculada da seguinte forma:

$$SQ_{total} = \sum_{k=1}^{K} SQ_k = \sum_{k=1}^{K} \sum_{x \in C_k} \|x - \bar{x}_k\|^2 \tag{6.2}$$

onde K é o número total de clusters, C_k representa o conjunto de observações pertencentes ao cluster k, x_i é um ponto de dado pertencente a C_k, e $\bar{x}_k$ é o centroide do cluster C_k. Um valor menor de SQ_{total} indica maior compactação dos pontos dentro de cada cluster, refletindo uma boa coesão do agrupamento.

No gráfico do método do cotovelo, SQ_{total} é traçado em função de K. À medida que K aumenta, SQ_{total} tende a diminuir, pois os clusters se tornam menores e mais específicos. O objetivo do método do cotovelo é encontrar o valor de K após o qual a diminuição em SQ_{total} se torna marginal, indicando que clusters adicionais não trazem uma melhora significativa na qualidade do agrupamento. Este ponto de inflexão na curva é o *cotovelo* e sugere o número ideal de clusters (Hastie et al., 2009). Em resumo, o método do cotovelo ajuda a evitar a criação de clusters artificiais e a supersegmentação, permitindo identificar um valor de K que maximiza a variabilidade explicada entre os clusters sem incluir clusters desnecessários.

6.4 CRITÉRIO DA SILHUETA MÉDIA

A silhueta média é uma das métricas mais utilizadas para determinar o número ideal de clusters no K-means, por fornecer uma medida conjunta de coesão e separação entre grupos (Artes e Barroso, 2023). Para cada possível número de clusters K, calcula-se a silhueta média $\bar{s}(K)$, que indica o quão bem definidos estão os agrupamentos formados

(Kaufman e Rousseeuw, 1990). O cálculo da silhueta baseia-se em duas quantidades fundamentais associadas a cada ponto i:

- $a(i)$: distância média entre o ponto i e os demais elementos do seu próprio cluster (coesão interna);
- $b(i)$: menor distância média entre o ponto i e os elementos de qualquer outro cluster, diferente daquele ao qual pertence (separação externa).

Considere que o ponto i pertence ao cluster C_k. A distância média $a(i)$ é calculada como a média das distâncias entre i e todos os outros pontos do mesmo grupo, excluindo o próprio ponto. Assim,

$$a(i) = \frac{\sum_{j \in C_k,\, j \neq i} d_{ij}}{n_{C_k} - 1}, \tag{6.3}$$

onde n_{C_k} é o número de elementos do cluster C_k e d_{ij} representa a distância entre os pontos i e j.

Para calcular $b(i)$, identifica-se o cluster C_l ($l \neq k$) que minimiza a distância média entre o ponto i e seus elementos. Para esse cluster mais próximo, define-se:

$$b(i) = \frac{\sum_{j \in C_l\, j \neq i} d_{ij}}{n_{C_l}}, \tag{6.4}$$

onde C_l é o cluster vizinho mais próximo de C_k para o ponto i, e n_{C_l} é sua cardinalidade.

Com estas quantidades, o coeficiente de silhueta para o ponto i é dado por:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{6.5}$$

que assume valores no intervalo $[-1, 1]$. Valores próximos de 1 indicam que o ponto está bem alocado, com forte coesão interna e boa separação entre clusters. Valores próximos de zero sugerem que o ponto está na fronteira entre dois grupos, enquanto valores negativos indicam possível alocação inadequada.

A qualidade de uma partição com K grupos é medida pela silhueta média $\bar{s}(K)$. O número de clusters ideal é aquele que maximiza essa quantidade:

$$K^* = \arg\max_k \; \bar{s}(K), \tag{6.6}$$

ou seja, o valor de K cuja partição produz a maior silhueta média. Esse procedimento permite identificar o particionamento que melhor reflete a estrutura inerente aos dados, evitando tanto a supersegmentação quanto a formação de agrupamentos artificiais.

6.5 CRITÉRIO DE INFORMAÇÃO BAYESIANO (BIC)

No contexto do *model-based clustering*, o número de componentes G da mistura geralmente é desconhecido e precisa ser determinado. Para isso, ajustam-se modelos para diferentes valores de G e, posteriormente, seleciona-se aquele que apresenta o melhor desempenho segundo um critério de escolha (Benites et al., 2025). Neste trabalho, utilizou-se o Critério de Informação Bayesiano (BIC), proposto por Schwarz (1978), que busca equilibrar qualidade de ajuste e complexidade do modelo, penalizando configurações com número excessivo de parâmetros. Para um modelo ψ com j componentes, o BIC é definido como:

$$BIC = 2\ell_c(\hat{\boldsymbol{\theta}}) - m_{\psi,j} \log n, \tag{6.7}$$

em que $\ell_c(\hat{\boldsymbol{\theta}})$ representa a log-verossimilhança maximizada calculada nos parâmetros estimados, $m_{\psi,j}$ é o número de parâmetros independentes do modelo, e j indica o número de componentes da mistura.

O BIC depende da partição inicial, pois a log-verossimilhança é obtida a partir dos estimadores de máxima verossimilhança penalizados pela complexidade do modelo. Esse critério é amplamente adotado no pacote `mclust`, que seleciona automaticamente o modelo e o número de componentes com base no maior valor de BIC, interpretado como maior evidência a favor da estrutura escolhida (Scrucca e Raftery ,2015). Resultados empíricos apresentados por Zeller et al., (2018) e Mirfarah, Naderi e Chen (2021) reforçam a robustez do BIC na identificação do número adequado de componentes em modelos de misturas, justificando sua popularidade em aplicações práticas.

7 VALIDAÇÃO DOS AGRUPAMENTOS

A validação de agrupamentos é um passo essencial para avaliar a qualidade dos resultados obtidos em uma análise de cluster. Este processo utiliza métodos formais, quantitativos e objetivos para assegurar que os grupos formados são consistentes e representam de maneira significativa a estrutura subjacente dos dados, evitando interpretações que possam ser influenciadas por ruído ou acaso (Jain e Dubes, 1988).

Dada a variedade de métodos de agrupamento disponíveis, cada um com características específicas, os resultados obtidos para um mesmo conjunto de dados podem variar significativamente. Por essa razão, a validação desempenha um papel crucial ao permitir a comparação e a avaliação da qualidade das estruturas de agrupamento geradas (Hoef e Warrens, 2019). Essa etapa é indispensável para garantir que o método escolhido reflete adequadamente os padrões presentes nos dados.

A validação de agrupamentos é realizada por meio de índices estatísticos que mensuram aspectos como coesão interna, separação entre os clusters e relação com rótulos externos, quando disponíveis. Esses índices são aplicados com base em critérios de validação, que definem a estratégia utilizada para avaliar a qualidade dos agrupamentos. Segundo Faceli, Carvalho e Souto (2005), os critérios de validação classificam-se em três categorias principais: critérios internos, critérios externos e critérios relativos. Na sequência, apresentam-se esses critérios.

7.1 CRITÉRIOS INTERNOS

Os critérios internos avaliam a qualidade dos agrupamentos exclusivamente com base nos dados originais, sem necessidade de informações externas ou rótulos predefinidos. Esses critérios analisam, de forma independente, a estrutura interna de cada partição, focando em aspectos como a coesão entre os elementos dentro dos clusters e a separação entre os diferentes clusters. Por exemplo, um critério interno pode medir o grau em que uma partição gerada por um algoritmo de agrupamento é justificada pela matriz de similaridade ou pelas distâncias entre os elementos (Faceli, Carvalho e Souto, 2005).

Por serem aplicados individualmente a cada partição, os critérios internos permitem verificar se os grupos formados representam adequadamente a estrutura intrínseca dos dados. Segundo Distefano, Mameli e Poli (DISTEFANO; MAMELI; POLI), esses critérios são especialmente úteis em cenários exploratórios, onde não há referências presumidas sobre as partições, e baseiam-se na avaliação da homogeneidade dos elementos dentro de cada grupo e da heterogeneidade entre os diferentes clusters. Em outras palavras, esses critérios medem:

- coesão interna: Mede o quão próximos estão os elementos dentro de um mesmo

cluster.

- separação externa: Avalia o quão distintos são os clusters entre si.

Existe um grande número de índices de validação utilizados em critérios internos para avaliar um agrupamento. Alguns dos índices tradicionais mais comuns são: a Silhueta Média, o Índice de Dunn e o Índice de Calinski–Harabasz (CH).

7.1.1 Silhuetta Média

A medida de silhueta, proposta por Rousseeuw (1987), é um dos critérios internos mais empregados para avaliar a qualidade de uma partição. Esse índice quantifica, para cada observação, o quão bem ela está ajustada ao seu cluster em comparação com os clusters vizinhos, permitindo interpretar simultaneamente a coesão interna e a separação externa.

Conforme definido na Seção (6.4), o coeficiente de silhueta $s(i)$ é calculado a partir das quantidades $a(i)$ e $b(i)$, que representam, respectivamente, a distância média entre o ponto i e os demais elementos do seu próprio cluster, e a menor distância média entre i e os elementos de qualquer outro cluster. Maiores detalhes sobre essas quantidades, bem como suas formulações matemáticas, encontram-se apresentados naquela seção.

O coeficiente $s(i)$ assume valores no intervalo $[-1, 1]$, onde valores próximos de 1 indicam forte adequação ao cluster, valores próximos de 0 sugerem pontos na fronteira entre grupos, e valores negativos indicam possível alocação incorreta. Além disso, a silhueta média pode ser utilizada para determinar o número ideal de clusters K, como já mencionado no capítulo anterior.

7.1.2 Índice de Dunn

O índice de Dunn é uma medida amplamente utilizada para avaliar a qualidade de agrupamentos, especialmente no que diz respeito à identificação de clusters compactos e bem separados (Halkidi et al., 2002). Esse índice busca verificar se os indivíduos de um mesmo grupo são homogêneos entre si e, ao mesmo tempo, distintos dos indivíduos pertencentes a outros grupos. Em outras palavras, ele mede simultaneamente a coesão interna de cada agrupamento e a separação entre os diferentes clusters formados.

A fórmula do índice de Dunn é dada por:

$$D = \frac{\min_{1 \leq l,k \leq K, l \neq k} \left(\min_{i \in C_k, j \in C_l} d(i,j) \right)}{\max_{1 \leq k \leq K} \left(\max_{i,j \in C_k, i \neq j} d(i,j) \right)}, \tag{7.1}$$

Nesse contexto, K representa o número total de clusters produzidos pelo método de agrupamento, enquanto C_k denota o k-ésimo cluster. A função $d(i, j)$ corresponde à distância entre dois pontos i e j no espaço de dados considerado. O termo $\min_{i \in C_k, j \in C_l} d(i, j)$

expressa a menor distância observada entre dois elementos pertencentes a clusters distintos, constituindo, portanto, uma medida de separação entre os agrupamentos C_k e C_l. Já a expressão $\max_{i,j \in C_k, i \neq j} d(i,j)$ representa a maior distância entre dois elementos do mesmo cluster C_k, refletindo o grau de dispersão ou falta de compactação interna desse agrupamento.

Assim, o índice de Dunn estabelece uma razão entre a separação mínima existente entre os clusters e a maior dispersão interna observada em qualquer um deles. Valores elevados desse índice indicam agrupamentos mais bem definidos, caracterizados por clusters internamente compactos e externamente bem separados. Por essa razão, o índice de Dunn é amplamente empregado como critério interno de validação em análises de clusterização.

O índice de Dunn, portanto, estabelece uma relação entre a separação mínima entre os clusters e a maior distância observada dentro de qualquer cluster. Quanto maior o valor do índice de Dunn, melhor será a separação entre os clusters e maior será a compactação interna, indicando um agrupamento bem definido e distinto.

7.1.3 Índice de Calinski-Harabasz (CH)

O Índice de Calinski-Harabasz (CH), proposto por Caliński e Harabasz (1974), é amplamente reconhecido como um dos critérios internos mais eficazes para avaliar a qualidade de particionamentos em problemas de agrupamento. Esse índice recebeu destaque no estudo comparativo de Milligan e Cooper (1985), no qual, segundo Tibshirani, Walther e Hastie (2001), apresentou o melhor desempenho entre 30 métodos distintos de validação utilizados para estimar o número ideal de clusters. Sua eficiência decorre do equilíbrio que estabelece entre a separação dos grupos e a compactação interna de cada cluster.

O índice CH é definido pela razão entre a variabilidade entre os clusters e a variabilidade dentro deles, devidamente ponderadas pelo número de grupos. Matematicamente, ele é dado por:

$$CH(K) = \frac{\frac{B(K)}{K-1}}{\frac{W(K)}{n-K}}. \tag{7.2}$$

Nessa expressão, K representa o número de clusters formados e n denota o total de observações. As quantidades $B(K)$ e $W(K)$ correspondem, respectivamente, à soma dos quadrados das distâncias ao centróide global (variabilidade entre clusters) e à soma dos quadrados das distâncias aos centróides específicos de cada cluster (variabilidade interna). Suas definições formais são:

$$B(K) = \sum_{k=1}^{K} n_{C_k} \|\bar{x}_k - \bar{x}\|^2, \tag{7.3}$$

$$W(K) = \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \bar{x}_k\|^2, \tag{7.4}$$

em que n_{C_k} é o tamanho do cluster C_k, $\bar{x}$ é o vetor de médias global do conjunto de dados, $\bar{x}_k$ é o centróide do cluster C_k e x_i é o vetor de atributos da observação i. Assim, $B(K)$ mensura o quanto os centróides dos clusters se afastam da média global, enquanto $W(K)$ quantifica a dispersão interna de cada grupo em torno de seu centróide. Note que essas expressões constituem a decomposição clássica da soma total dos quadrados, de modo que $B(K)$ e $W(K)$ são diretamente análogos às somas dos quadrados entre e dentro dos grupos, respectivamente.

Quanto maior for o valor de $CH(K)$, melhor será a relação entre separação e compacidade dos clusters, indicando um particionamento mais adequado. Além disso, o índice pode ser utilizado para determinar o número ideal de grupos: o valor ótimo de K corresponde àquele que maximiza o índice Calinski-Harabasz.

7.2 CRITÉRIOS RELATIVOS

Os critérios relativos também são ferramentas fundamentais na validação de agrupamentos, utilizadas para comparar diferentes resultados obtidos a partir de um ou mais algoritmos de clustering. O objetivo é identificar qual solução apresenta o melhor ajuste aos dados ou qual algoritmo é mais adequado para representar a estrutura analisada, considerando diferentes configurações, como o número de clusters C_k (Faceli, Carvalho e Souto, 2005).

Como mencionado anteriormente, os critérios relativos comparam múltiplas soluções de agrupamento geradas sob condições variadas. Essas condições podem incluir:

- **Algoritmo utilizado:** Comparação entre métodos de agrupamentos como K-means, Método de Ward ou Misturas Gaussianas.
- **Número de clusters:** Avaliação para determinar a quantidade ideal de grupos (K).
- **Parâmetros específicos:** Ajustes que dependem das particularidades de cada algoritmo.

A análise consiste no cálculo de um índice de qualidade para cada agrupamento, gerando uma sequência de valores que é então analisada. O agrupamento mais adequado é determinado pela otimização do índice (máximo ou mínimo), ou pela identificação de um ponto de inflexão na curva resultante.

Diversos índices utilizados em critérios internos também podem ser aplicados como critérios relativos. Como observado por Jain e Dubes (1988), a diferença reside na forma de aplicação: a distinção entre critérios internos e relativos não está no índice em si, mas na maneira como ele é empregado. Quando o índice é calculado para múltiplas soluções de agrupamento, comparando-se seus valores em busca daquele que se destaca seja por

um pico, um vale ou uma mudança significativa na trajetória dos valores ao longo de diferentes configurações, ele passa a ser utilizado como um critério relativo. Por outro lado, quando avalia exclusivamente a qualidade de uma única solução de agrupamento, é considerado um critério interno. Essa abordagem relativa auxilia na escolha da melhor solução, alinhada à estrutura dos dados e aos objetivos do estudo. Por fim, é importante destacar que os índices discutidos anteriormente como critérios internos serão utilizados para comparar os métodos de agrupamento no contexto relativo e para auxiliar na escolha do número ideal de clusters.

7.3 CRITÉRIOS EXTERNOS

Os critérios externos de validação avaliam a qualidade de um agrupamento comparando os clusters gerados com uma estrutura de referência previamente conhecida ou estabelecida. Essa estrutura pode ser baseada em rótulos que representam a divisão ideal dos dados ou em partições construídas por especialistas da área, que utilizam conhecimento prévio sobre a natureza do conjunto de dados (Faceli et al., 2005).

O objetivo principal dos critérios externos é medir a correspondência entre os agrupamentos gerados pelo algoritmo e as partições esperadas, refletindo a intuição ou conhecimento do pesquisador sobre a estrutura subjacente nos dados. Esses critérios são particularmente úteis quando há uma expectativa clara sobre como os dados devem ser organizados. Por exemplo:

- Validação com rótulos conhecidos: Se os dados incluem rótulos categóricos conhecidos, como classes em problemas supervisionados, o critério externo verifica a precisão com que os clusters reproduzem essas categorias.
- Validação com conhecimento especializado: Para conjuntos de dados onde não há rótulos explícitos, mas existem agrupamentos sugeridos por especialistas, os critérios externos avaliam a proximidade entre a partição obtida e a estrutura sugerida.

Neste trabalho, a validação externa será realizada considerando rótulos conhecidos, uma vez que o conjunto de dados utilizado inclui a classificação verdadeira das observações. Nesse contexto, serão empregados dois critérios complementares: o Índice Rand Ajustado (ARI) e a acurácia obtida a partir da matriz de confusão, interpretada aqui no âmbito do agrupamento. Dessa forma, a acurácia atua como uma métrica adicional para avaliar o desempenho dos métodos de clustering, complementando a análise fornecida pelo ARI.

7.3.1 Índice Rand Ajustado (ARI)

O Índice de Rand Ajustado (ARI) é uma métrica amplamente utilizada para avaliar a similaridade entre duas partições de um conjunto de dados. Ele ajusta o Índice de Rand

(RI) para corrigir o impacto do acaso, proporcionando uma medida mais confiável na comparação de agrupamentos (Souza, 2010).

O RI mede a concordância entre duas partições analisando pares de elementos no conjunto de dados, considerando se esses pares estão no mesmo grupo ou em grupos diferentes em ambas as partições. No entanto, como o RI não corrige para concordâncias que poderiam ocorrer por pura sorte, seu valor pode ser inflado em situações de agrupamentos aleatórios. O ARI resolve essa limitação, ajustando o índice para remover o efeito do acaso.

Considere um conjunto de dados D particionado de duas maneiras diferentes:

- Partição $PC = \{C_1, C_2, \ldots, C_k\}$, com k grupos.
- Partição $PC' = \{C'_1, C'_2, \ldots, C'_{k'}\}$, com k' grupos.

Para cada par de elementos no conjunto D, as seguintes categorias são definidas:

- N_{11}: Número de pares que estão no mesmo grupo em ambas PC e PC'.
- N_{00}: Número de pares que estão em grupos diferentes em ambas PC e PC'.
- N_{10}: Número de pares que estão no mesmo grupo em PC, mas em grupos diferentes em PC'.
- N_{01}: Número de pares que estão no mesmo grupo em PC', mas em grupos diferentes em PC.

O número total de pares possíveis é dado por:

$$N = \frac{n(n-1)}{2},$$

onde n é o número total de elementos no conjunto de dados.

O Índice de Rand (RI) é então definido como:

$$RI = \frac{N_{11} + N_{00}}{N}.$$

O ARI ajusta esse índice considerando a expectativa do RI em agrupamentos aleatórios ($E[RI]$):

$$ARI = \frac{RI - E[RI]}{1 - E[RI]}. \tag{7.5}$$

Uma forma prática de calcular o ARI envolve o uso de uma tabela de contingência $[n_{ij}]$, em que $n_{ij} = |C_i \cap C'_j|$ representa o número de elementos pertencentes simultaneamente

ao cluster C_i da partição PC e ao cluster C'_j da partição PC'. A partir dessa tabela, o índice pode ser reescrito como:

$$ARI(PC, PC') = \frac{\sum_{ij} \binom{n_{ij}}{2} - \frac{\sum_i \binom{n_i}{2} \sum_j \binom{n'_j}{2}}{\binom{n}{2}}}{\frac{1}{2}\left(\sum_i \binom{n_i}{2} + \sum_j \binom{n'_j}{2}\right) - \frac{\sum_i \binom{n_i}{2} \sum_j \binom{n'_j}{2}}{\binom{n}{2}}}, \tag{7.6}$$

em que $n_i = \sum_j n_{ij}$ corresponde ao tamanho total do cluster C_i e $n'_j = \sum_i n_{ij}$ representa o tamanho total do cluster C'_j. Nessa formulação, cada termo da soma $\binom{n_{ij}}{2}$ contabiliza pares de elementos que pertencem simultaneamente ao mesmo cluster em ambas as partições, enquanto os demais termos ajustam o índice para descontar a concordância esperada pelo acaso.

O ARI assume valores no intervalo $[-1, 1]$. Valores próximos de 1 indicam alta concordância entre as partições, refletindo forte similaridade estrutural entre os agrupamentos. Valores próximos de 0 sugerem que a correspondência observada não difere da similaridade esperada ao acaso. Já valores negativos indicam que a concordância entre as partições é inferior ao esperado aleatoriamente, revelando discordância sistemática entre elas. Dessa forma, o ARI constitui uma métrica robusta para avaliar a concordância entre dois agrupamentos, especialmente adequada para cenários em que o número de clusters e os tamanhos dos grupos variam de forma significativa.

8 MÉTODO DE SEGMENTAÇÃO DE IMAGENS

A segmentação de imagens é uma técnica amplamente utilizada na análise de imagens digitais, cujo objetivo principal consiste em particionar uma imagem em regiões homogêneas com base em um ou mais atributos visuais, como cor, textura ou intensidade (Ballard e Brown, 1982). Os algoritmos anteriormente apresentados como métodos de agrupamento também possuem aplicações relevantes nesse contexto, podendo ser empregados tanto na segmentação quanto na compressão de imagens, devido à sua capacidade de agrupar pixels com características semelhantes. Para compreender melhor o funcionamento desses métodos no processo de segmentação de imagem, é importante antes revisar o conceito de imagem digital.

8.1 DEFINIÇÃO DE IMAGEM DIGITAL

De acordo com Gonzalez e Woods (2010), uma imagem digital pode ser definida como uma função bidimensional $f(x, y)$, na qual x e y representam coordenadas espaciais no plano, e o valor da função f em um dado par (x, y) corresponde à intensidade ou ao nível de cinza naquele ponto. Quando as variáveis x, y e os valores de intensidade assumem quantidades finitas e discretas, a função $f(x, y)$ caracteriza uma imagem digital propriamente dita.

Cada imagem digital é composta por um número finito de elementos denominados *pixels* (elementos de imagem). Em outras palavras, uma imagem pode ser entendida como uma matriz de pixels, na qual cada pixel armazena informações que determinam suas características visuais, como cor, brilho ou intensidade. A forma como esses pixels estão organizados e suas propriedades individuais são fundamentais para a aplicação de técnicas de processamento e análise, como a segmentação. Para ilustrar essa estrutura matricial, considere o exemplo hipotético de uma imagem em tons de cinza formada por uma matriz 3×3, cujos valores representam intensidades variando de 0 (preto) a 255 (branco):

$$I = \begin{bmatrix} 34 & 120 & 200 \\ 90 & 180 & 255 \\ 10 & 60 & 150 \end{bmatrix} \tag{8.1}$$

Nesse caso, cada elemento da matriz corresponde a um pixel da imagem. Por exemplo, o valor 34 indica um pixel mais escuro, enquanto o valor 255 representa um pixel completamente branco. Essa organização matricial permite interpretar imagens digitais como estruturas numéricas, possibilitando a aplicação de métodos computacionais para modificações, melhoramentos e, especialmente, segmentação.

Além disso, segundo Marques e Vieira (1999), uma imagem digital também pode ser modelada a partir da interação entre a iluminação incidente sobre um objeto e suas

propriedades de refletância. Nesse modelo, a função $f(x, y)$, que representa a intensidade da imagem no ponto (x, y), é expressa como o produto entre a iluminância $i(x, y)$ e a refletância $r(x, y)$, conforme a equação:

$$f(x, y) = i(x, y) \cdot r(x, y). \tag{8.2}$$

A função $i(x, y)$ descreve a quantidade de luz que incide sobre o objeto naquele ponto, sendo determinada pelas características da fonte de iluminação, com $0 < i(x, y) < \infty$. Já a função $r(x, y)$ representa a fração da luz incidente que é refletida pelo objeto no ponto (x, y), estando restrita ao intervalo $0 \leq r(x, y) \leq 1$. Um valor de $r(x, y) = 0$ indica absorção total da luz, enquanto $r(x, y) = 1$ corresponde à reflexão total. Esse modelo físico evidencia que a intensidade observada em cada pixel depende tanto das condições de iluminação quanto das propriedades dos objetos presentes na cena.

8.1.1 Imagens Coloridas e Quantização de Cores

As imagens digitais podem ser representadas de diferentes formas, dependendo da quantidade de informação armazenada em cada pixel. Entre os principais tipos estão as imagens binárias, em tons de cinza e as coloridas. As imagens binárias são as mais simples, possuindo apenas dois valores (preto e branco), sendo frequentemente utilizadas em tarefas básicas de segmentação entre objeto e fundo. Já as imagens em tons de cinza representam diferentes intensidades de luz entre o preto e o branco e são amplamente usadas quando a informação cromática não é essencial.

Neste trabalho, o foco recai sobre as imagens coloridas, que apresentam uma descrição mais rica e informativa. Nesse tipo de representação, cada pixel é definido por três componentes que expressam sua composição cromática. O modelo de cor mais comum é o RGB (Red, Green, Blue), no qual cada pixel é descrito por um vetor tridimensional de intensidades das cores primárias aditivas, usualmente codificadas em 8 bits cada (valores de 0 a 255) (Marques e Vieira, (1999)). Dessa forma, cada pixel pode ser representado por um triplo ordenado:

$$p = (R,\, G,\, B),$$

que indica suas intensidades de vermelho, verde e azul, respectivamente. A popularidade do modelo RGB decorre de sua compatibilidade direta com dispositivos de captura e exibição, como câmeras digitais e monitores.

Exemplo ilustrativo. Considere uma imagem composta por três pixels dispostos horizontalmente, cada um representando uma cor primária pura:

$$p_1 = (255,\, 0,\, 0) \quad \text{(vermelho)}, \qquad p_2 = (0,\, 255,\, 0) \quad \text{(verde)}, \qquad p_3 = (0,\, 0,\, 255) \quad \text{(azul)}.$$

Organizando esses pixels na forma de imagem, temos:

$$I = \begin{bmatrix} (255,0,0) & (0,255,0) & (0,0,255) \end{bmatrix}.$$

Essa representação evidencia que cada elemento da matriz da imagem não é um número único, mas sim um vetor tridimensional correspondente às três bandas (vermelho, verde e azul). Em termos computacionais, uma imagem RGB de dimensão $m \times n$ é armazenada como uma estrutura tridimensional:

$$I \in \mathbb{R}^{m \times n \times 3},$$

onde as duas primeiras dimensões representam a posição espacial do pixel, enquanto a terceira dimensão corresponde às componentes R, G e B. Assim, pode-se interpretar cada pixel como um ponto no espaço tridimensional de cores, no qual cores semelhantes estão próximas, e cores distintas encontram-se mais afastadas. Essa interpretação geométrica é fundamental para métodos de segmentação baseados em cor, pois algoritmos de agrupamento, como o k-means, tratam diretamente os pixels como vetores em $\mathbb{R}^3$.

A informação cromática desempenha papel relevante na segmentação, uma vez que regiões com valores próximos no espaço RGB tendem a ser agrupadas como pertencentes ao mesmo objeto ou superfície. Em muitos casos, essa diferenciação cromática é mais eficaz do que uma análise baseada apenas em intensidades de cinza, sobretudo quando objetos apresentam tonalidades semelhantes, mas cores distintas.

Apesar de a representação RGB padrão utilizar 24 bits (8 bits por canal), nem sempre é necessário preservar toda essa profundidade de cor. O processo de **quantização de cores** permite reduzir o número de cores possíveis na imagem, tipicamente para 256 (8 bits), por meio do uso de um *mapa de cores*. Nessa estrutura, cada pixel armazena apenas um índice que referencia uma tabela contendo as cores disponíveis. Conforme discutido por Gonzalez e Woods (2010), essa técnica reduz o tamanho do arquivo e o custo computacional, preservando a qualidade visual em muitas aplicações, especialmente em imagens com regiões homogêneas.

8.1.2 Processamento de Imagens

De acordo com Gonzalez e Woods (2010), o processamento digital de imagens não constitui uma tarefa isolada, mas um sistema composto por etapas interconectadas: aquisição, pré-processamento, segmentação, extração de características e, por fim, interpretação e classificação. A conversão de uma cena real em uma imagem digital ocorre inicialmente nas fases de aquisição e digitalização. A fase de aquisição consiste em transformar uma cena tridimensional em uma imagem analógica, geralmente por meio da captura da luz refletida pelos objetos (Queiroz e Gomes, 2001). Em seguida, essa imagem contínua é convertida para uma representação discreta, possibilitando seu armazenamento e processamento computacional (Conci, Azevedo e Leta, 2008).

Após a obtenção da imagem digital, realiza-se o pré-processamento, cujo objetivo principal é aprimorar a qualidade da imagem para favorecer as etapas subsequentes. Nesse

estágio, aplicam-se técnicas como realce de contraste, redução de ruído e isolamento de regiões de interesse. Ainda no contexto do pré-processamento, ocorre uma etapa essencial para a aplicação de algoritmos de segmentação baseados em agrupamento: a reorganização estrutural dos dados da imagem.

Imagens coloridas no modo RGB são armazenadas como matrizes tridimensionais de dimensão $m \times n \times 3$, em que as duas primeiras dimensões representam as coordenadas espaciais dos pixels e a terceira corresponde às três bandas de cor (vermelho, verde e azul). Entretanto, métodos como k-means, Misturas de Gaussianas e o método de Ward exigem que cada observação seja representada como um vetor em um espaço de características. Assim, cada pixel deve ser convertido em um vetor tridimensional (R, G, B), e a imagem deve ser reestruturada em uma matriz bidimensional de ordem $N \times 3$, sendo $N = m \times n$ o número total de pixels.

Essa conversão consiste em linearizar cada banda de cor e em seguida combiná-las em uma única matriz de características:

$$X = \begin{bmatrix} R_1 & G_1 & B_1 \\ R_2 & G_2 & B_2 \\ \vdots & \vdots & \vdots \\ R_N & G_N & B_N \end{bmatrix}.$$

Esse pré-processamento é fundamental, pois permite que cada pixel seja tratado como uma observação em $\mathbb{R}^3$. Dessa forma, cada linha da matriz X passa a representar um pixel, enquanto cada coluna corresponde a uma das três componentes de cor. Essa organização dos dados viabiliza a aplicação dos algoritmos de segmentação que serão discutidos nas seções posteriores deste trabalho.

A etapa seguinte, a segmentação, tem como finalidade subdividir a imagem em partes ou objetos constituintes, de modo que regiões homogêneas possam ser analisadas separadamente (Gonzalez e Woods, 2010). O objetivo central do processamento de imagens é extrair informações sobre os objetos representados. Para isso, a segmentação isola regiões de pixels com características semelhantes, enquanto os métodos de extração de atributos calculam parâmetros descritivos dessas regiões (Scuri, 2002). Posteriormente, a seleção de características identifica variáveis relevantes, fundamentais para diferenciar classes de objetos e auxiliar na classificação.

A fase final corresponde às etapas de reconhecimento e interpretação, nas quais os objetos são identificados e classificados com base nas características previamente extraídas. O reconhecimento consiste na atribuição de rótulos ou categorias aos objetos analisados, enquanto a interpretação busca compreender o significado dos padrões detectados no contexto da imagem. Dessa forma, objetos que compartilham propriedades semelhantes são agrupados e associados a uma mesma classe, encerrando o ciclo de processamento e

análise da informação visual.

8.2 SEGMENTAÇÃO DE IMAGENS

A segmentação de imagens é considerada uma das etapas mais desafiadoras do processamento digital de imagens. Trata-se do processo de dividir uma imagem em regiões ou objetos significativos, de modo a separar elementos de interesse do fundo ou de outras partes da cena. Em visão computacional, uma imagem segmentada corresponde ao agrupamento de pixels em unidades homogêneas de acordo com características comuns, como cor, textura ou bordas (Ballard e Brown, 1982).

O conceito de *image segmentation* surgiu na década de 1980 (Fu e Mui, 1980) e permanece até hoje como um campo essencial de pesquisa, pois está na base de todo o processamento de informações visuais. Segundo Mongelo (2012), seu principal objetivo é reduzir a complexidade da imagem, transformando-a em uma representação estruturada na qual os objetos ou regiões de interesse possam ser analisados de forma mais precisa. Na prática, a segmentação procura isolar áreas que compartilham características semelhantes, que podem ser definidas a partir de propriedades internas, como textura, cor e brilho, ou por propriedades externas, como fronteiras e contornos (Gonzalez e Woods, 2010). Uma Região de Interesse, por exemplo, pode ser obtida automaticamente a partir da própria imagem ou definida pelo usuário, sendo o local onde o processamento será concentrado. Diversas técnicas de segmentação foram propostas ao longo dos anos, cada uma mais adequada a determinados tipos de imagens. Não existe, entretanto, um método universal capaz de oferecer bons resultados para todos os tipos de imagens, uma vez que cada técnica apresenta vantagens e limitações, devendo ser escolhida conforme as características e o objetivo da aplicação.

A saída do processo de segmentação costuma ser um conjunto de regiões homogêneas, sem sobreposição, no qual cada pixel da imagem pertence a apenas uma região. Essa representação é especialmente útil em etapas posteriores, como reconhecimento, análise e interpretação, pois facilita a identificação de padrões e a classificação dos objetos (Conci, Azevedo e Leta, 2008). Assim, segmentar uma imagem significa transformá-la em partes coerentes que evidenciam informações relevantes para o domínio de aplicação, servindo como base para todo o processamento digital subsequente.

Durante a segmentação, o objetivo é reconhecer e isolar o objeto de interesse do restante da cena. Para isso, são selecionados atributos representativos da imagem, formando um vetor de características que a descreve de forma simplificada. Essa etapa visa reduzir a quantidade de informações necessárias para a classificação e, consequentemente, o tempo de processamento da tarefa (Castleman, 1996).

A Figura 4 ilustra um exemplo prático do processo de segmentação descrito anteriormente. A imagem foi segmentada utilizando o algoritmo K-means, com número de

grupos igual a 7, de modo que cada pixel foi representado por suas componentes de cor e agrupado conforme sua similaridade. O resultado evidencia como o método é capaz de particionar a imagem original em regiões homogêneas, destacando estruturas relevantes a partir da redução de variabilidade dentro de cada grupo. Esse tipo de abordagem demonstra a utilidade da segmentação baseada em agrupamento para simplificar cenas complexas e facilitar análises posteriores.

Figura 4 – Exemplo de segmentação de imagem utilizando o algoritmo K-means com $N = 7$.
Fonte: Elaboração própria (2025).

8.2.1 Técnicas de Segmentação

Neste trabalho, foi adotada a abordagem não supervisionada de segmentação de imagens, utilizando-se métodos de agrupamento já apresentados e descritos anteriormente: K-means, Misturas Finitas de Densidades e o método hierárquico de Ward.

Essas técnicas serão aplicadas com o objetivo de particionar a imagem em regiões homogêneas, de acordo com as características de cor e intensidade dos pixels. Nas subseções a seguir, são apresentados os procedimentos práticos de segmentação para cada um dos métodos, exemplificando suas aplicações.

8.2.2 Exemplificação do procedimento para o K-means

A segmentação de imagens por meio do método K-means foi realizada conforme os seguintes passos:

- **Representação dos pixels:** Cada pixel da imagem foi convertido em um vetor contendo os valores de intensidade nos três canais de cor (R, G e B).
- **Aplicação do algoritmo:** O algoritmo K-means foi executado sobre o conjunto de vetores, particionando os pixels em N grupos de acordo com a similaridade de cor.

- **Cálculo dos centróides:** Cada grupo é representado por um centróide que corresponde à média das cores dos pixels pertencentes àquele agrupamento.
- **Análise dos resultados:** Para valores menores de N, observaram-se grandes regiões homogêneas e bem definidas. À medida que N aumenta, a segmentação se torna mais detalhada, destacando variações sutis na coloração.

8.2.3 Exemplificação do procedimento para as Misturas Finitas de Densidades

A segmentação também foi realizada por meio das Misturas Finitas de Densidades, utilizando o modelo de Misturas Gaussianas implementado no pacote `mclust`. Esse método assume que os dados foram gerados por uma combinação de distribuições normais multivariadas, estimando os parâmetros de cada componente por meio do algoritmo EM (Expectation-Maximization).

O procedimento adotado seguiu as seguintes etapas:

- **Representação dos pixels:** Cada pixel da imagem foi convertido em um vetor contendo os valores de intensidade nos três canais de cor (R, G e B).
- **Aplicação do algoritmo:** Foram ajustados modelos de mistura para diferentes números de componentes N, permitindo avaliar a influência de N na qualidade da segmentação obtida.
- **Classificação dos pixels:** Cada pixel foi associado ao componente com maior probabilidade a posteriori, de acordo com os parâmetros estimados para as distribuições gaussianas.
- **Análise dos resultados:** as médias das distribuições associadas a cada grupo foram utilizadas para reconstruir a imagem segmentada.

8.2.4 Exemplificação do procedimento para o Método Hierárquico de Ward

A segmentação também foi realizada por meio do método hierárquico de Ward, o qual tem como princípio minimizar a variância interna dos grupos a cada etapa de fusão. Essa técnica é particularmente útil para identificar agrupamentos mais compactos e homogêneos, sendo uma alternativa não supervisionada eficiente para a segmentação de imagens coloridas.

- **Amostragem dos pixels:** Devido à alta dimensionalidade da imagem (com milhões de pixels), foi necessário selecionar uma amostra aleatória de aproximadamente 5.000 pixels, de modo a viabilizar o cálculo das distâncias euclidianas, cujo custo computacional é quadrático em relação ao número de observações.

- **Construção do dendrograma:** A partir dessa amostra, foi calculada a matriz de distâncias e aplicado o método de Ward para gerar o dendrograma hierárquico, representando o processo de fusão dos grupos com base na minimização da variância intragrupo.
- **Definição do número de grupos:** Foram testados diferentes valores de N, realizando o corte do dendrograma em cada caso para obter as partições correspondentes. Assim, foi possível analisar o comportamento da segmentação em diferentes níveis de granularidade.
- **Cálculo dos centróides e expansão:** Após definir os grupos na amostra, foram calculados os centróides médios (valores médios de R, G e B) de cada cluster. Em seguida, todos os pixels da imagem foram classificados com base na menor distância euclidiana ao centróide mais próximo, garantindo a reconstrução da segmentação completa.
- **Reconstrução da imagem:** Por fim, a imagem foi reconstruída substituindo as cores originais de cada pixel pelas cores médias de seus respectivos agrupamentos, permitindo visualizar o resultado da segmentação hierárquica.

9 APLICAÇÕES

O presente trabalho tem como objetivo comparar diferentes técnicas de agrupamento, incluindo o método hierárquico aglomerativo de Ward, o algoritmo K-means e as misturas de gaussianas. Neste capítulo, apresentam-se as aplicações práticas dessas técnicas, utilizando diferentes bases de dados para avaliar seu desempenho e suas particularidades. A seguir, descrevem-se os conjuntos de dados analisados e os resultados obtidos em cada estudo.

9.1 CONJUNTO DE DADOS *OLD FAITHFUL*

O conjunto de dados *Old Faithful* é um dos exemplos clássicos na literatura estatística para ilustração de técnicas de agrupamento e análise exploratória de dados. Ele contém observações referentes ao gêiser Old Faithful, localizado no Parque Nacional de Yellowstone, nos Estados Unidos, conhecido por suas erupções relativamente regulares. Esse conjunto é amplamente utilizado em estudos de agrupamento por apresentar uma estrutura naturalmente agrupada, com evidências claras da existência de dois padrões distintos de comportamento, o que o torna particularmente adequado para a aplicação e comparação de métodos de agrupamento não supervisionados visto no presente trabalho. Os dados estão disponíveis publicamente na literatura e em repositórios de software estatístico. Mais especificamente, o conjunto de dados pode ser acessado diretamente através do ambiente `R`, onde encontra-se incluído no pacote `datasets` sob o nome *faithful*.

9.1.1 Objetivo da Análise

O objetivo central desta análise é aplicar os métodos de agrupamento discutidos para identificar e caracterizar as subpopulações latentes nas erupções do gêiser. Busca-se segmentar as observações com base na relação bivariada entre a duração da erupção e o tempo de espera até o evento seguinte. Desta forma, pretende-se verificar a eficácia das técnicas na detecção da bimodalidade natural dos dados, separando as erupções em dois regimes distintos: o de curta duração com curtos intervalos de espera, e o de longa duração com longos intervalos.

9.1.2 Descrição das Variáveis

O conjunto de dados é composto por duas variáveis quantitativas, descritas a seguir:

- **Eruptions**: Duração da erupção do gêiser, em minutos;
- **Waiting**: Tempo de espera até a próxima erupção, em minutos.

9.1.3 Pré-processamento dos Dados e Análise Exploratória

Antes da aplicação dos algoritmos de agrupamento, realizou-se uma etapa de pré-processamento para garantir a qualidade e a comparabilidade dos dados. Inicialmente, verificou-se a integridade do conjunto, não sendo constatada a presença de valores ausentes ou inconsistências que exigissem tratamento ou remoção de observações. Em seguida, procedeu-se à análise exploratória para compreender o comportamento das variáveis. A Tabela 2 apresenta as estatísticas descritivas dos dados originais.

Tabela 2 – Estatísticas descritivas das variáveis numéricas do conjunto *Old Faithful*.

Variável	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Eruptions	1.60	2.16	4.00	3.49	4.45	5.10
Waiting	43.00	58.00	76.00	70.90	82.00	96.00

A análise visual é apresentada nas Figuras 5, 6 e 7, permitindo uma compreensão detalhada da distribuição e do relacionamento entre as variáveis.

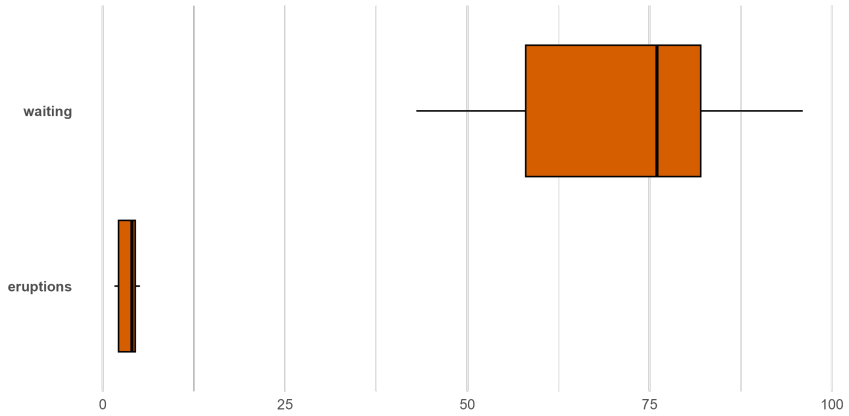

Figura 5 – Boxplots das variáveis numéricas do conjunto *Old Faithful*.
Fonte: Elaboração própria (2025).

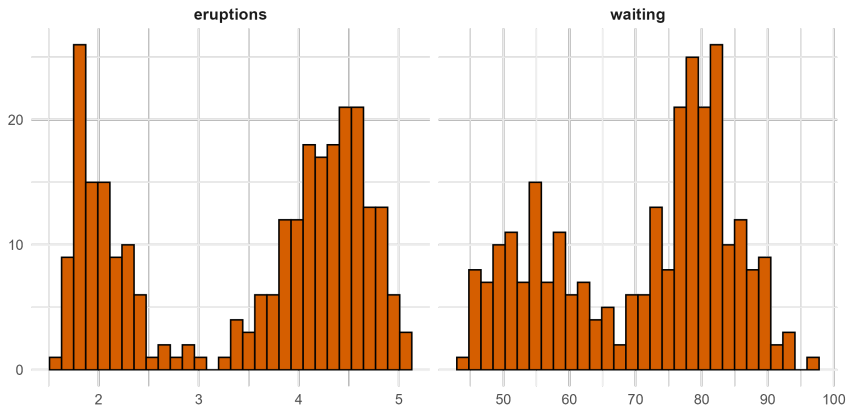

Figura 6 – Histogramas das variáveis do conjunto *Old Faithful*.
Fonte: Elaboração própria (2025).

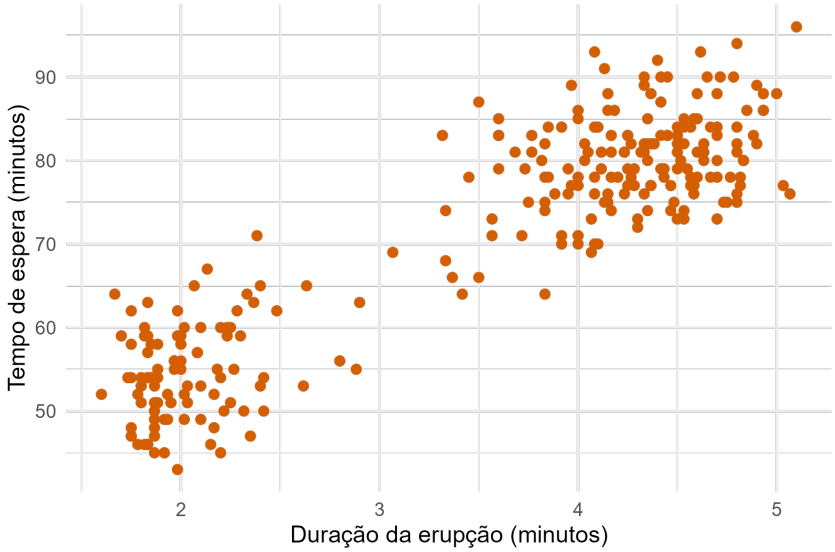

Figura 7 – Gráfico de dispersão entre *Eruptions* e *Waiting*.
Fonte: Elaboração própria (2025).

A análise dos histogramas (Figura 6) revela uma bimodalidade evidente em ambas as variáveis, indicando a existência de duas subpopulações latentes. O gráfico de dispersão (Figura 7) confirma visualmente essa estrutura, exibindo dois agrupamentos (*clusters*) densos e claramente separados. O gráfico de dispersão (Figura 7) confirma visualmente essa estrutura, exibindo dois agrupamentos claramente separados e densos, o que justifica a aplicação de métodos de agrupamento para segmentar essas observações.

Entretanto, ao observar a Figura 5 e a Tabela 2, nota-se uma discrepância significativa nas escalas: enquanto a variável *Eruptions* varia em intervalos curtos (aprox. 1,6 a 5,1), a variável *Waiting* apresenta valores em magnitude superior (aprox. 43 a 96). Tal diferença pode fazer com que a variável de maior magnitude domine o cálculo de distâncias (como a Euclidiana).

Para mitigar esse efeito, optou-se pela padronização dos dados pelo método do *Z-score*, resultando em variáveis com média zero e desvio padrão unitário. As estatísticas dos dados transformados são apresentadas na Tabela 9.

Tabela 3 – Estatísticas descritivas das variáveis padronizadas do conjunto *Old Faithful*.

Variável	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Eruptions	-1.65	-1.16	0.45	0.00	0.85	1.41
Waiting	-2.05	-0.95	0.38	0.00	0.82	1.85

A partir dessa padronização, torna-se viável e metodologicamente adequada a aplicação das técnicas de agrupamento discutidas nas seções subsequentes.

9.1.4 Escolha do Número de grupos

Conforme discutido no capítulo 6, a definição dessa quantidade constitui uma etapa fundamental nas técnicas de agrupamento, pois influencia diretamente a qualidade e a interpretabilidade dos resultados. Nesta aplicação, foram utilizados diferentes critérios para essa determinação: o método do cotovelo e o coeficiente de silhueta média para o *K-means*, o Critério de Informação Bayesiano (BIC) para as misturas de distribuições, e a análise do dendrograma, considerando os níveis de fusão e de similaridade, no caso do método hierárquico.

9.1.4.1 Método do Cotovelo

A Figura 8 evidencia uma redução significativa na variabilidade intra-grupos até $K = 2$. A partir desse ponto, as reduções tornam-se menos expressivas, sugerindo a estabilização da curva. Dessa forma, $K = 2$ é considerado o candidato adequado para o número de *clusters*.

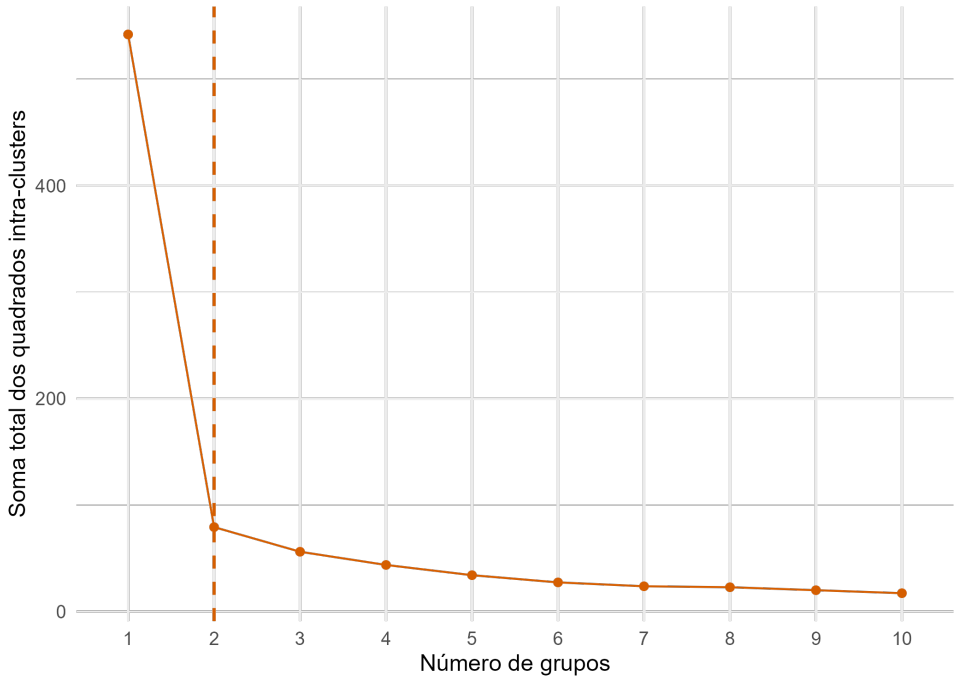

Figura 8 – Método do cotovelo para definição do número de clusters no K-means.
Fonte: Elaboração própria (2025).

9.1.4.2 Análise da Silhueta Média

O valor máximo da largura média da silhueta, observado na Figura 9, ocorre em $K = 2$. Este resultado indica que a qualidade da partição é maximizada nesta configuração, ou seja, a separação entre os grupos e a coesão interna são mais nítidas neste ponto.

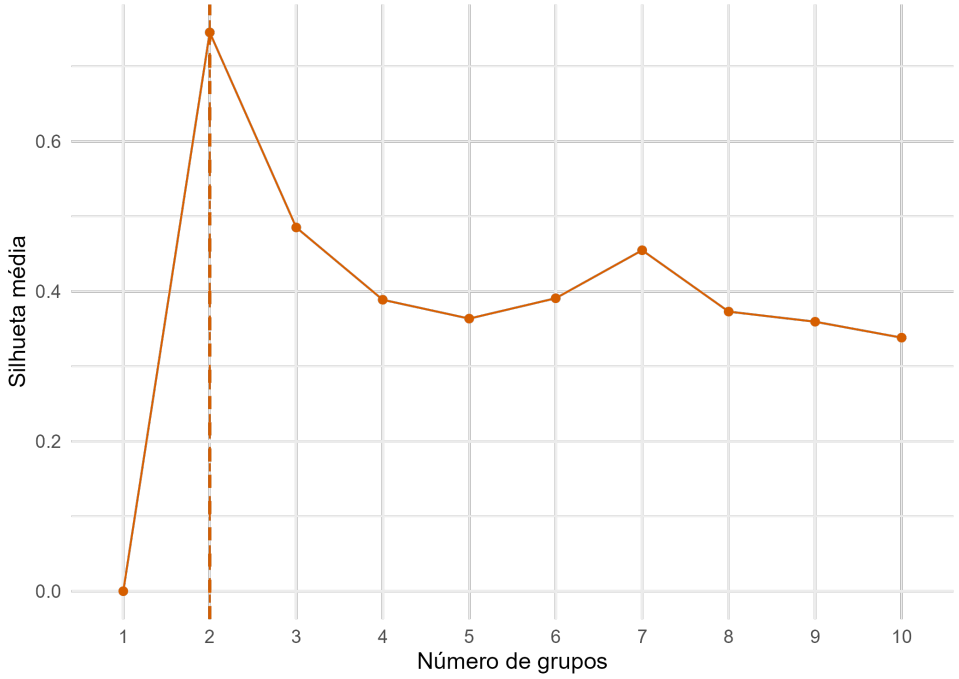

Figura 9 – Análise da silhueta média para diferentes números de clusters.
Fonte: Elaboração própria (2025).

9.1.4.3 Critério de Informação Bayesiano (BIC)

A avaliação dos modelos de Misturas Gaussianas foi realizada com base no BIC, a figura 10 apresenta os valores para todas as combinações de modelos e números de componentes ajustados. Para facilitar a identificação das melhores configurações, a tabela **??** destaca apenas os três maiores valores de BIC, correspondentes às soluções mais bem avaliadas segundo esse critério.

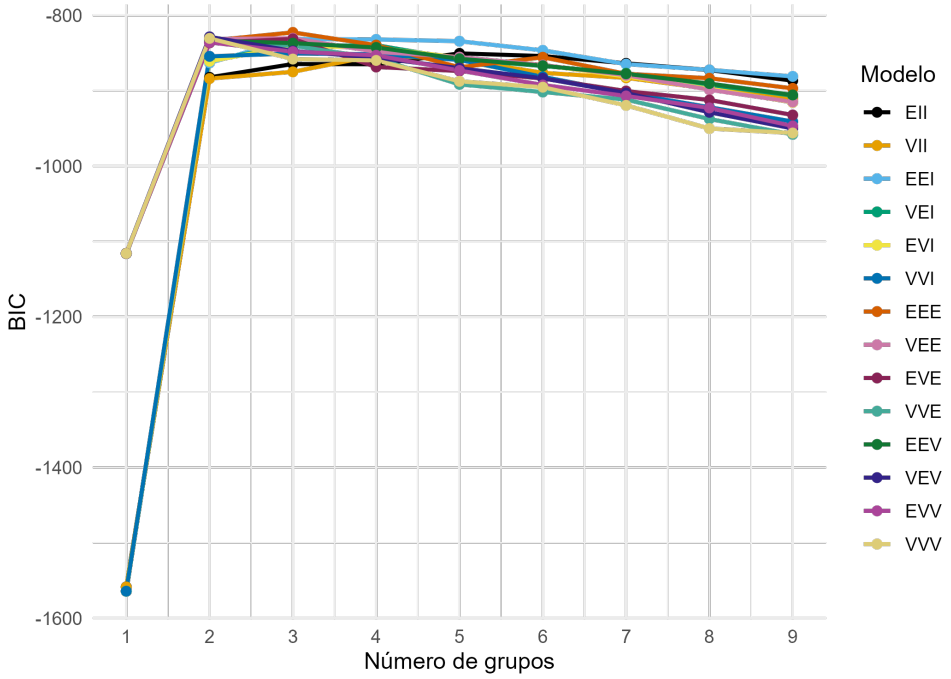

Figura 10 – Valores do Critério de Informação Bayesiano (BIC) para diferentes números de componentes e modelos de Misturas Gaussianas.
Fonte: Elaboração própria (2025).

Modelo	Número de Componentes	BIC
EEE	3	-822.692
VEV	2	-828.569
VEE	3	-830.381

Tabela 4 – Modelos com os melhores valores de BIC para Misturas Gaussianas.
Fonte: Elaboração própria (2025).

Observa-se que o modelo EEE com 3 componentes maximizou o critério BIC. No entanto, a diferença para o segundo colocado é muito pequena. Considerando o princípio da parcimônia e o objetivo de manter a consistência com os métodos de escolha de números de grupos anteriores (que indicaram $K = 2$), optou-se pela seleção do modelo VEV com 2 componentes. o modelo *VEV* admite volumes e formas variáveis entre os componentes, mantendo a mesma orientação, o que confere maior flexibilidade na representação da estrutura dos dados.

9.1.4.4 Análise do comportamento do nível de fusão (distância) e do nível de similaridade

No método hierárquico de Ward, a definição do número de grupos foi orientada pela análise conjunta dos níveis de fusão e de similaridade, apresentados no painel esquerdo (a) e no painel direito (b) da figura correspondente, respectivamente. O gráfico de saltos do nível de fusão permite identificar os pontos em que há aumentos mais expressivos nas distâncias entre grupos sucessivamente unidos, o que indica uma perda considerável

de homogeneidade interna. Paralelamente, o gráfico de similaridade mostra a queda acentuada na medida de semelhança entre os agrupamentos à medida que o número de grupos diminui.

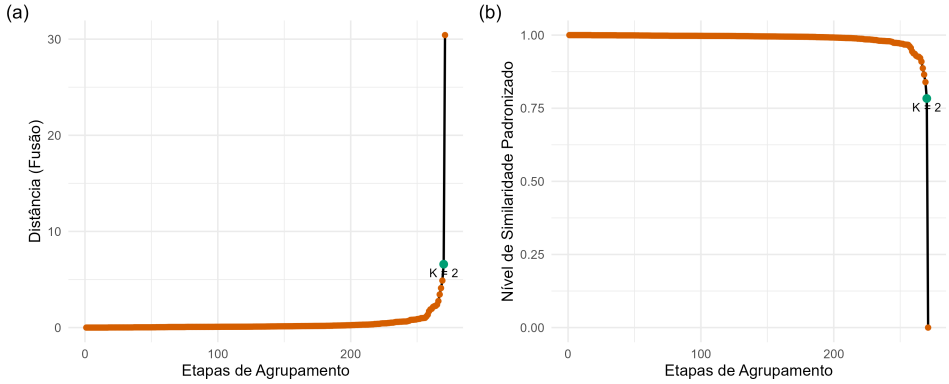

Figura 11 – Comportamento dos Níveis de Fusão e Similaridade.
Fonte: Elaboração própria (2025).

Observa-se que o maior salto no nível de fusão ocorre na transição de $K = 2$ para $K = 1$ grupos, indicando uma fusão abrupta e consequente perda significativa de similaridade entre os clusters. De forma consistente, o gráfico de similaridade também evidencia uma queda acentuada nesse mesmo ponto, reforçando que a estrutura natural dos dados é mais bem representada por dois grupos. Assim, tanto a análise do nível de fusão quanto a da similaridade sugerem que a escolha mais adequada para este conjunto de dados é $K = 2$.

9.1.5 Aplicação das Técnicas de Agrupamento

Nesta seção, são aplicados três métodos de agrupamento: o método hierárquico de Ward, o K-means e o modelo de Misturas Gaussianas. Primeiramente, apresentam-se os agrupamentos obtidos por cada técnica. Em seguida, os resultados são avaliados por meio de critérios internos, com o objetivo de identificar qual método apresenta maior consistência e coerência interna. A partir dessa análise, será selecionado o método com melhor desempenho para os dados em estudo, cujos clusters serão interpretados em detalhe.

9.1.5.1 Método Hierárquico de Ward

A Figura 12 apresenta o dendrograma obtido pelo método hierárquico de Ward. A inspeção visual do gráfico permite identificar a formação de agrupamentos bem definidos, refletindo a estrutura natural dos dados.

Com base nas análises do nível de fusão e da similaridade, discutidas na seção anterior, observa-se que a solução mais consistente corresponde à formação de $K = 2$ grupos. Essa configuração equilibra adequadamente a homogeneidade interna dos clusters e a

heterogeneidade entre eles, evitando tanto a fragmentação excessiva quanto a fusão de grupos distintos.

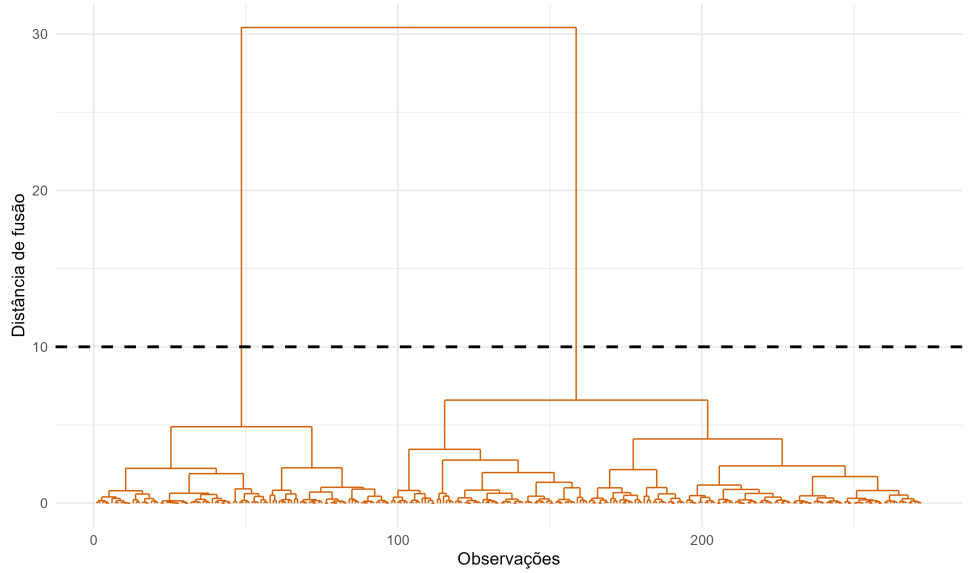

Figura 12 – Dendrograma obtido pelo método hierárquico de Ward.
Fonte: Elaboração própria (2025).

9.1.5.2 Método K-means

A Figura 13 ilustra a distribuição dos grupos obtidos, considerando $k = 2$. Para facilitar a visualização da variabilidade e dos limites de cada cluster, foram adicionadas elipses baseadas na distância Euclidiana. Essa escolha é consistente com a natureza do algoritmo K-means, que utiliza a distância Euclidiana para atribuir observações ao centróide mais próximo. Dessa forma, as elipses euclidianas fornecem uma representação geométrica coerente com a lógica de formação dos grupos, destacando a dispersão e a proximidade dos elementos em torno de seus centróides.

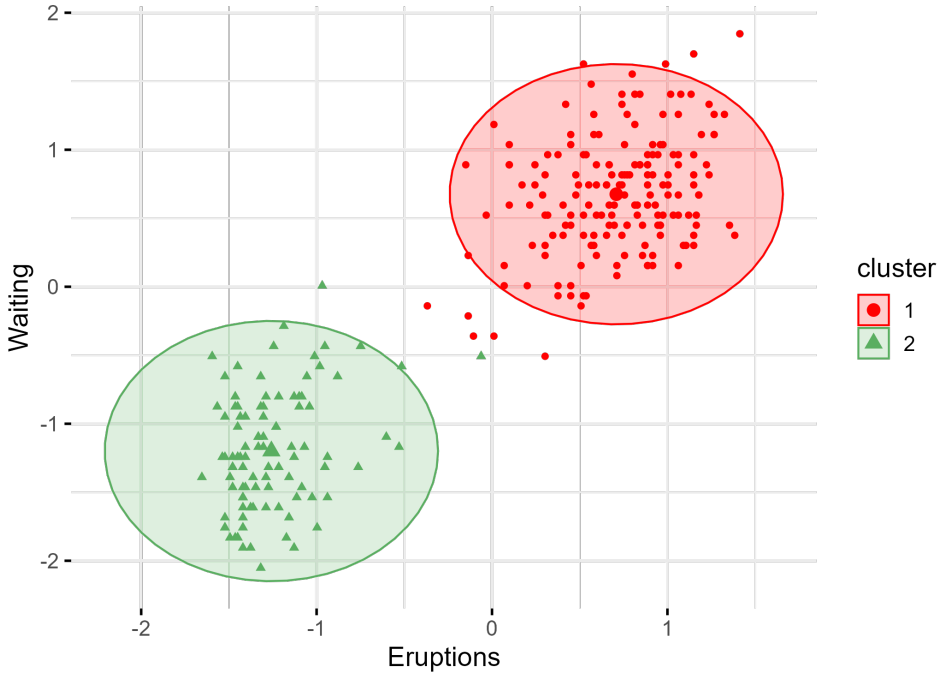

Figura 13 – Agrupamento obtido pelo k-means com dois grupos.
Fonte: Elaboração própria (2025).

9.1.5.3 Modelo de Misturas Gaussianas

A Figura 14 apresenta o resultado do ajuste do modelo de Mistura Gaussiana do tipo **VEV** (Elipsoidal, Volume Variável, Forma Igual e Orientação Variável) com 2 componentes. Diferente dos modelos esféricos, o modelo VEV permite que os grupos tenham tamanhos e rotações diferentes, adaptando-se melhor à correlação natural entre as variáveis.

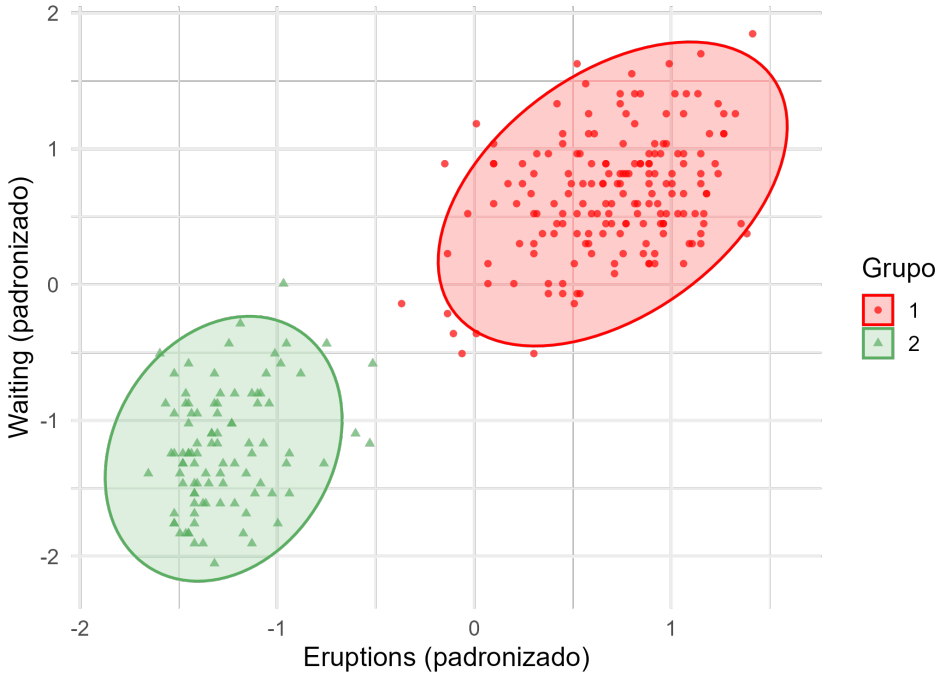

Figura 14 – Agrupamento obtido pelo modelo de Misturas Gaussianas (VEV, 2 componentes).
Fonte: Elaboração própria (2025).

As estimativas dos parâmetros são apresentadas na Tabela **??**. Além das médias vetoriais (μ_k) e dos desvios-padrão marginais (σ), o modelo estimou as proporções de mistura ($\hat{p}$) para cada componente como:

$$\hat{p} = (0,644;\ 0,356).$$

Isso indica que o Cluster 1 (associado a valores positivos na escala padronizada) detém aproximadamente 64,4% das observações, enquanto o Cluster 2 (valores negativos) contém 35,6%.

Tabela 5 – Estimativas de μ_j e σ_j do modelo de misturas finitas VEV ($\sigma_j = \sqrt{\text{diag}(\Sigma_j)}$).

Variáveis	**Cluster 1**		**Cluster 2**	
	μ_1	σ_1	μ_2	σ_2
Eruptions	0,702	0,362	-1,272	0,245
Waiting	0,667	0,457	-1,208	0,398

Fonte: Elaboração própria (2025).

Analisando a Tabela 13, nota-se que os clusters estão bem separados, o cluster 1 caracteriza-se por médias positivas em ambas as variáveis (maior tempo de erupção e espera) e apresenta maior variabilidade (desvios maiores). O Cluster 2 possui médias negativas (menor tempo de erupção e espera) e é mais compacto (menores desvios-padrão), o que é visualmente confirmado pelas elipses na Figura 14.

9.1.6 Resultados e Discussões

A Tabela 6 apresenta os resultados dos três critérios internos utilizados para avaliar a qualidade dos agrupamentos obtidos pelos métodos de Ward, K-means e Misturas Gaussianas. Como não se dispõe de rótulos verdadeiros das observações, a avaliação baseou-se exclusivamente em critérios relativos, isto é, medidas internas que permitem comparar o desempenho dos métodos a partir da estrutura dos próprios dados.

Tabela 6 – Comparação dos métodos de agrupamento segundo critérios internos

Método	**Silhueta**	**Calinski-Harabasz**	**Dunn**
Ward	0,746	1574,555	0,165
K-means	0,745	1575,784	0,057
Misturas	0,746	1574,555	0,165

Fonte: Elaboração própria (2025).

A Tabela 6 revela que os métodos de Ward e Misturas Gaussianas apresentaram desempenho idêntico e superior ao K-means na maioria das métricas avaliadas. Todos os algoritmos demonstraram alta coesão interna, com coeficientes de Silhueta em torno de 0,746.

O destaque da comparação reside no índice de Dunn. Enquanto o K-means obteve um valor baixo (0,057), sugerindo fronteiras pouco distintas ou clusters muito próximos, o método de Ward alcançou um índice significativamente maior (0,165). Isso indica que o agrupamento hierárquico foi mais eficiente em maximizar a separabilidade entre os grupos formados. Embora o K-means tenha apresentado um valor marginalmente superior no índice Calinski-Harabasz (reflexo de sua função objetivo de minimização de variância), essa vantagem não compensa a baixa separabilidade apontada pelo índice de Dunn.

Dessa forma, o método de Ward foi selecionado como o mais adequado, pois alia alta compactação (Silhueta elevada) à melhor definição de fronteiras entre os grupos (maior Dunn), além de ser um modelo menos dependente de suposições distributivas em comparação às Misturas Gaussianas. A Tabela 7 apresenta as estatísticas descritivas dos agrupamentos obtidos pelo método de Ward, as quais subsidiam a interpretação dos grupos formados.

Tabela 7 – Estatísticas descritivas dos agrupamentos (média e desvio padrão).

Variáveis	**Cluster 1**		**Cluster 2**	
	Média	DP	Média	DP
Eruptions	4,3	0,4	2,0	0,3
Waiting	80,0	6,0	54,5	5,8

A análise descritiva dos agrupamentos evidencia a formação de dois grupos distintos associados ao comportamento do gêiser. No Cluster 1, observam-se eventos de maior

magnitude, caracterizados por erupções de longa duração (média de 4,3 minutos) seguidas por um longo intervalo de recarga ou espera até o próximo evento (média de 80,0 minutos). O baixo desvio padrão em ambas as variáveis sugere que este é um padrão bastante consistente.

Em contrapartida, o Cluster 2 representa os eventos de menor intensidade, definidos por erupções curtas (média de 2,0 minutos) e um tempo de recuperação significativamente menor (média de 54,5 minutos). Essa distinção clara entre os grupos corrobora a relação física esperada: erupções mais duradouras liberam uma quantidade maior de energia e água, necessitando, consequentemente, de um período de espera mais extenso para que o sistema geotermal se reabasteça.

9.2 ANÁLISE DE AGRUPAMENTO NO CONJUNTO DE DADOS *WINES*

O conjunto de dados *Wines* é composto por informações químicas extraídas de amostras de vinho cultivadas na mesma região da Itália, conforme descrito por EVERITT *et al.*. A base de dados foi doada por Riccardo Leardi e é amplamente utilizada em estudos de classificação e agrupamento. Sua popularidade advém da estrutura clara e da presença de três grupos de vinhos distintos, previamente identificados, o que permite validar algoritmos de agrupamento não supervisionado por meio da comparação com os rótulos reais. O conjunto *Wines* está disponível publicamente no repositório de aprendizado de máquina da UCI (UCI Machine Learning Repository), podendo ser acessada em: **https://archive.ics.uci.edu/dataset/109/wine**

9.2.1 Objetivo da Análise

O objetivo desta análise é aplicar diferentes técnicas de agrupamento ao conjunto de dados *Wines*, desconsiderando inicialmente os rótulos reais das amostras. Embora esses grupos sejam conhecidos, eles são ignorados em um primeiro momento para permitir uma análise não supervisionada. A partir dos agrupamentos obtidos pelos métodos K-means, Método de Ward e Misturas Gaussianas, será realizada uma comparação com os grupos verdadeiros, a fim de avaliar a capacidade dos modelos em identificar estruturas naturais presentes nos dados. Essa abordagem permite investigar o desempenho dos métodos sob a perspectiva de critérios externos, fornecendo uma medida objetiva de sua eficiência em reproduzir padrões existentes.

9.2.2 Descrição das Variáveis

O conjunto de dados contém 13 variáveis quantitativas contínuas, que representam medidas químicas das amostras de vinho. As variáveis são as seguintes:

- **Alcohol**: teor alcoólico da amostra.

- **Malic acid**: concentração de ácido málico.
- **Ash**: teor de cinzas.
- **Alcalinity of ash**: alcalinidade das cinzas.
- **Magnesium**: concentração de magnésio.
- **Total phenols**: teor de fenóis totais.
- **Flavanoids**: teor de flavonoides.
- **Nonflavanoid phenols**: teor de fenóis não flavonoides.
- **Proanthocyanins**: concentração de proantocianidinas.
- **Color intensity**: intensidade da cor do vinho.
- **Hue**: tonalidade do vinho.
- **OD280/OD315 of diluted wines**: razão entre as absorvâncias a 280 e 315 nm.
- **Proline**: concentração de prolina.

9.2.3 Pré-processamento dos Dados e Análise Exploratória

Antes da aplicação das técnicas de agrupamento, foi realizado o pré-processamento dos dados com o objetivo de garantir a qualidade e a consistência das análises. Como o conjunto de dados *Wines* não apresenta valores ausentes, o foco concentrou-se na identificação de possíveis outliers e na padronização das variáveis.

Tabela 8 – Estatísticas descritivas das variáveis reais do conjunto de dados *Wines*.

Variável	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Alcohol	11.03	12.36	13.05	13.00	13.68	14.83
Malic	0.74	1.60	1.87	2.34	3.08	5.80
Ash	1.36	2.21	2.36	2.37	2.56	3.23
Alcalinity	10.60	17.20	19.50	19.49	21.50	30.00
Magnesium	70.00	88.00	98.00	99.74	107.00	162.00
Total Phenols	0.98	1.74	2.36	2.30	2.80	3.88
Flavanoids	0.34	1.20	2.13	2.03	2.88	5.08
Nonflavanoid	0.13	0.27	0.34	0.36	0.44	0.66
Proanthocyanins	0.41	1.25	1.56	1.59	1.95	3.58
Color	1.28	3.22	4.69	5.06	6.20	13.00
Hue	0.48	0.78	0.96	0.96	1.12	1.71
Dilution	1.27	1.94	2.78	2.61	3.17	4.00
Proline	278.00	500.50	673.50	746.89	985.00	1680.00

Fonte: Elaboração própria (2025).

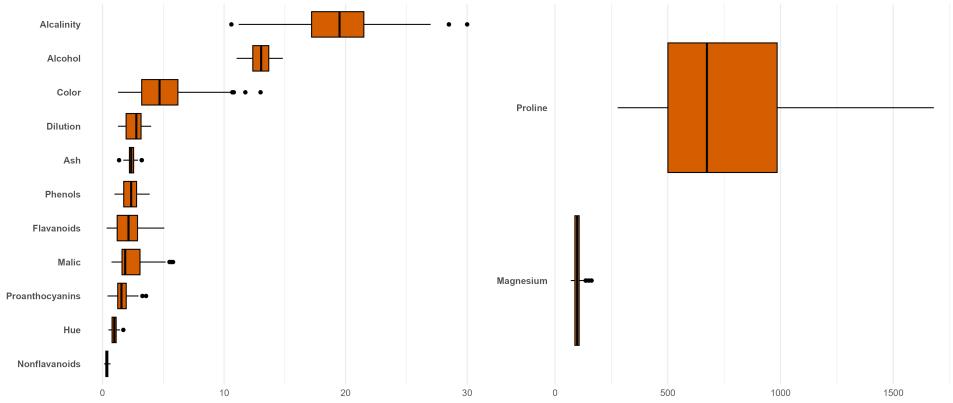

Figura 15 – Boxplots das variáveis numéricas do conjunto *Wines*.
Fonte: Elaboração própria (2025).

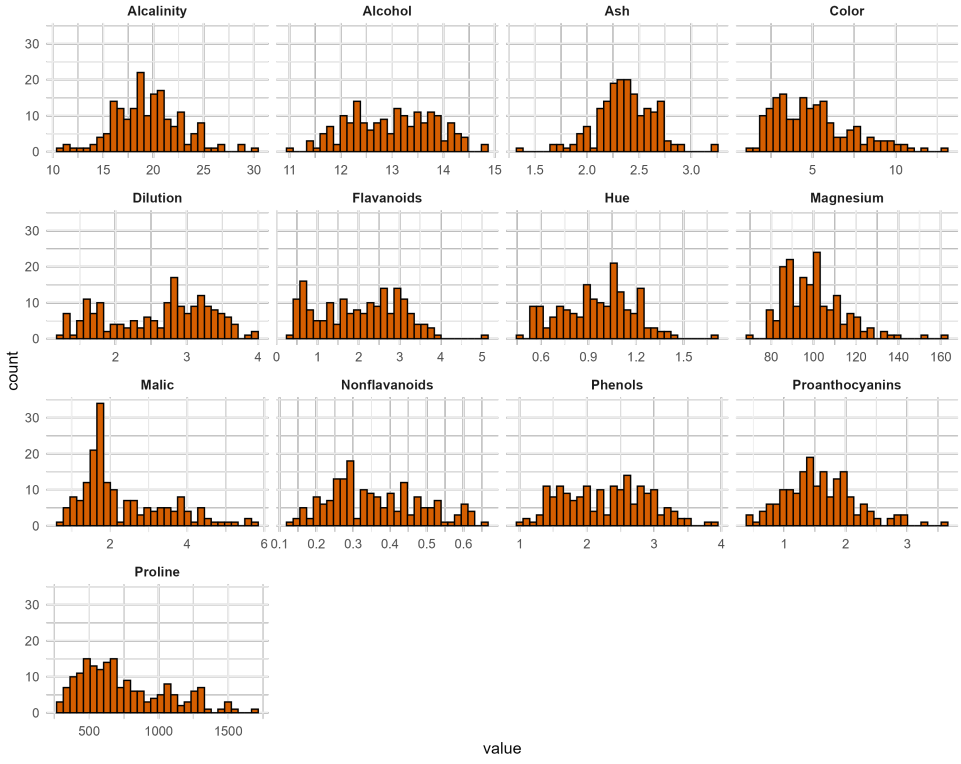

Figura 16 – Histogramas das variáveis numéricas do conjunto *Wines*.
Fonte: Elaboração própria (2025).

A inspeção visual por meio dos boxplots (Figura 15) e dos histogramas (Figura 16) permitiu identificar a presença de valores discrepantes em algumas variáveis. Observa-se, por exemplo, que a variável *Proline* apresenta uma escala consideravelmente superior às demais, o que reforça a necessidade da padronização. Ademais, enquanto variáveis como *Ash* e *Alcalinity* exibem distribuições próximas à normalidade, outras, como *Malic*, *Flavanoids*, *Color* e *Proline*, apresentam indícios de padrões multimodais.

Com o objetivo de mitigar os efeitos de escalas distintas e evitar que variáveis com maior variabilidade exercessem influência desproporcional nos métodos de agrupamento,

procedeu-se à padronização dos dados por meio do método de *z-score*. Após essa transformação, todas as variáveis passaram a apresentar média igual a zero e desvio-padrão unitário, assegurando comparabilidade direta entre elas. As estatísticas descritivas resultantes desse procedimento são apresentadas na Tabela 9.

Tabela 9 – Estatísticas descritivas das variáveis padronizadas (Z-score) do conjunto de dados *Wines*

Variável	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Alcohol	-2.43	-0.79	0.06	0.00	0.83	2.25
Malic	-1.43	-0.66	-0.42	0.00	0.67	3.10
Ash	-3.67	-0.57	-0.02	0.00	0.70	3.15
Alcalinity	-2.66	-0.69	0.00	0.00	0.60	3.15
Magnesium	-2.08	-0.82	-0.12	0.00	0.51	4.36
Phenols	-2.10	-0.88	0.10	0.00	0.81	2.53
Flavanoids	-1.69	-0.83	0.11	0.00	0.85	3.05
Nonflavanoids	-1.86	-0.74	-0.18	0.00	0.61	2.40
Proanthocyanins	-2.06	-0.60	-0.06	0.00	0.63	3.48
Color	-1.63	-0.79	-0.16	0.00	0.49	3.43
Hue	-2.09	-0.77	0.03	0.00	0.71	3.29
Dilution	-1.89	-0.95	0.24	0.00	0.79	1.96
Proline	-1.49	-0.78	-0.23	0.00	0.76	2.96

Fonte: Elaboração própria (2025).

Embora, em geral, valores de *z-score* superiores a 3 ou inferiores a -3 sejam considerados potenciais outliers, optou-se por manter todas as observações no conjunto de dados. Essa decisão visa preservar a variabilidade natural da amostra e avaliar o comportamento dos métodos de agrupamento diante da presença de discrepâncias moderadas, considerando que alguns algoritmos podem ser sensíveis a esses valores.

9.2.4 Escolha do Número de Grupos

A determinação do número de grupos seguiu os mesmos critérios utilizados na aplicação anterior, considerando os métodos do cotovelo e da silhueta média para o K-means, o Critério de Informação Bayesiano (BIC) para as Misturas Gaussianas e a inspeção do dendrograma no método de Ward utilizando o método do nível de fusão.

9.2.4.1 Método do Cotovelo

Na Figura 17, observa-se que a soma dos quadrados intra-grupos apresenta uma redução acentuada até o agrupamento com três clusters. A partir desse ponto, a taxa de decréscimo torna-se mais suave, caracterizando o formato típico de um "cotovelo". Dessa forma, definiu-se $K = 3$ como o número mais adequado de grupos para o método K-means.

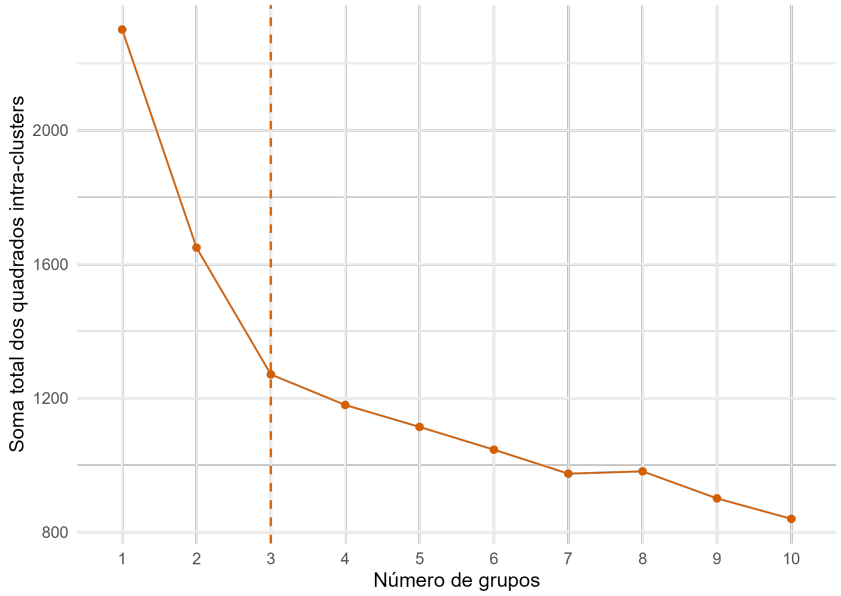

Figura 17 – Método do cotovelo para definição do número de clusters no K-means.
Fonte: Elaboração própria (2025).

9.2.4.2 Análise da Silhueta Média

A Figura 18 apresenta os valores do coeficiente de silhueta média para diferentes números de clusters. Verifica-se que o valor máximo ocorre em $K = 3$, indicando que essa configuração proporciona a melhor separação entre os grupos, com elevada coesão interna e boa distinção entre os clusters.

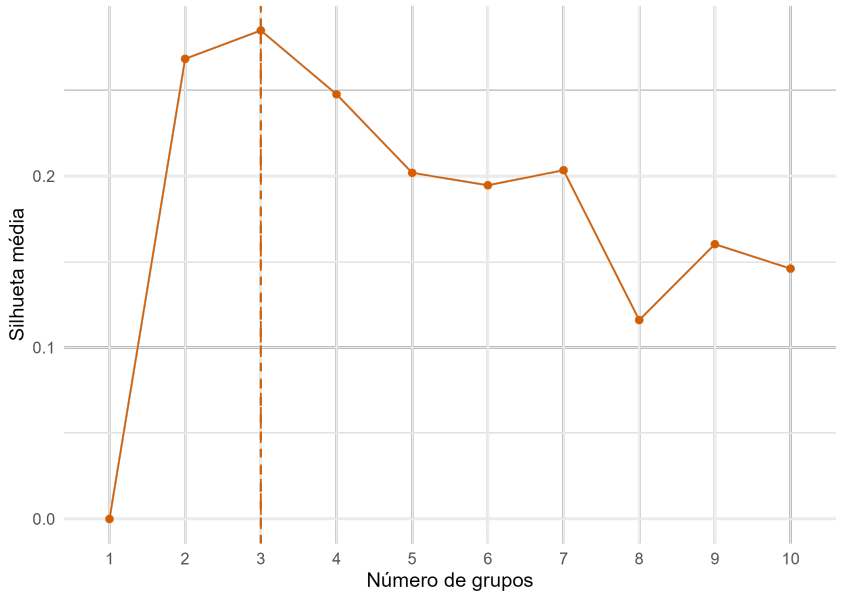

Figura 18 – Análise da silhueta média para diferentes números de clusters.
Fonte: Elaboração própria (2025).

9.2.4.3 Critério de Informação Bayesiano (BIC)

A avaliação dos modelos de Misturas Gaussianas foi realizada com base no BIC, a figura 19 apresenta os valores para todas as combinações de modelos e números de componentes ajustados. Para facilitar a identificação das melhores configurações, a Tabela 10 destaca apenas os três maiores valores de BIC. A seguir, apresenta-se essa seleção.

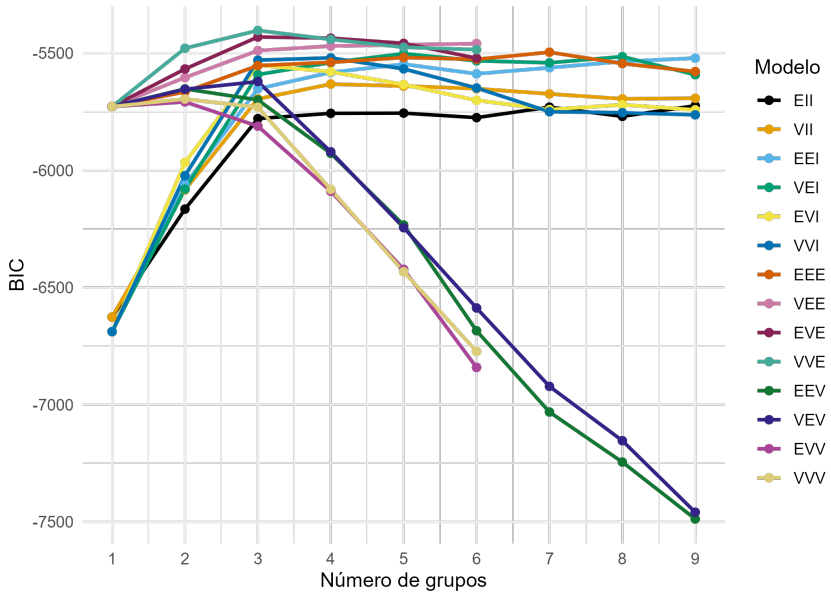

Figura 19 – Valores do Critério de Informação Bayesiano (BIC) para diferentes números de componentes e modelos de Misturas Gaussianas.
Fonte: Elaboração própria (2025).

Tabela 10 – Modelos com os melhores valores de BIC para Misturas Gaussianas.

Modelo	Número de Componentes	BIC
VVE	3	-5403,829
EVE	3	-5431,653
EVE	4	-5436,461

Fonte: Elaboração própria (2025).

Observa-se que o modelo *VVE* com três componentes apresentou o maior valor de BIC, sendo, portanto, considerado o mais adequado para representar a estrutura dos dados. Esse modelo assume que as distribuições que compõem a mistura possuem volumes e formas distintos, porém compartilham a mesma orientação da matriz de covariância. Em outras palavras, cada grupo pode apresentar diferentes níveis de dispersão e diferentes alongamentos elipsoidais, mas todos se alongam em uma direção comum no espaço multivariado.

No contexto do conjunto *Wines*, essa configuração faz sentido, pois as variáveis químicas dos vinhos tendem a apresentar padrões de correlação semelhantes entre si,

refletindo processos produtivos e características enológicas comuns, sao mesmo tempo em que os três tipos de vinho exibem variações próprias de magnitude e variabilidade interna. Assim, o modelo *VVE* consegue capturar de forma mais realista a heterogeneidade natural entre os grupos, ao permitir que cada um tenha volume e forma específicos, sem exigir orientações diferentes para cada componente.

9.2.4.4 Análise do comportamento do nível de fusão (distância) e do nível de similaridade

A Figura 20 mostra o comportamento dos níveis de fusão e de similaridade, apresentados no painel esquerdo (a) e no painel direito (b) da figura correspondente, respectivamente. Observa-se que, nas etapas iniciais, as fusões ocorrem entre grupos muito semelhantes, o que se reflete em baixos valores de distância e altos níveis de similaridade. À medida que o algoritmo progride, as distâncias de fusão aumentam de forma gradual até que, em determinado ponto, ocorre um salto acentuado na distância, indicando a união de grupos menos homogêneos.

Esse aumento repentino no nível de fusão coincide com uma queda expressiva na curva de similaridade, sugerindo a perda de homogeneidade entre os grupos e, portanto, um ponto adequado para interromper o processo de agrupamento. Nesse caso, o salto observado corresponde à formação de três grupos bem definidos, valor que também foi consistente com os resultados obtidos pelos métodos do cotovelo e da silhueta média.

Dessa forma, tanto a análise das distâncias quanto a da similaridade reforçam a escolha de $K = 3$ como número ideal de clusters para a base de dados *Wines*, evidenciando coerência entre as diferentes abordagens utilizadas para a definição do número de grupos.

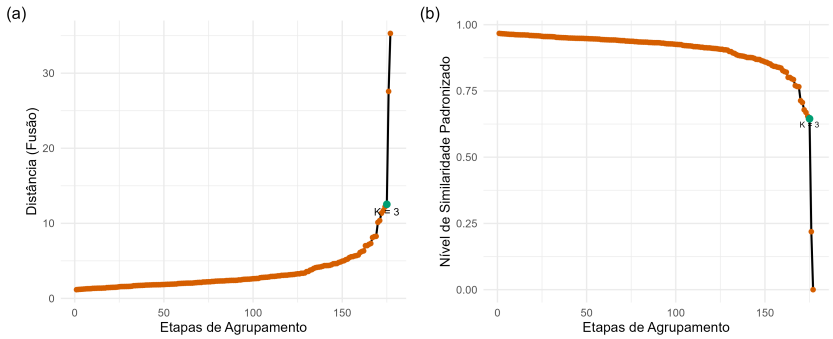

Figura 20 – Comportamento dos Níveis de Fusão e Similaridade.
Fonte: Elaboração própria (2025).

9.2.5 Aplicação das Técnicas de Agrupamento

Nesta etapa, aplicam-se os métodos de agrupamento de Ward, K-means e Misturas Gaussianas ao conjunto de dados *Wines*. Os resultados obtidos são comparados aos rótulos originais por meio da matriz de confusão, do Índice de Rand Ajustado e da acurácia

média. O método com melhor desempenho é então selecionado para análise e interpretação detalhada dos grupos formados.

9.2.5.1 Método Hierárquico de Ward

A Figura 21 apresenta o dendrograma obtido pelo método hierárquico de Ward. Observando o gráfico pode-se identificar a formação de três grupos bem definidos, o que é evidenciado pelos saltos expressivos nos níveis de fusão. Esses saltos indicam aumentos consideráveis nas distâncias entre os grupos a cada etapa de junção, sugerindo pontos de corte ideais para a determinação do número ótimo de clusters, conforme discutido anteriormente.

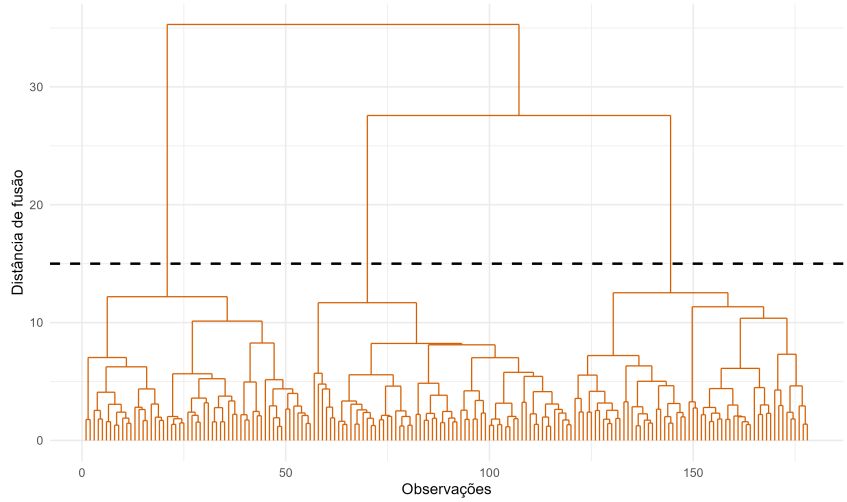

Figura 21 – Dendrograma obtido pelo método hierárquico de Ward.
Fonte: Elaboração própria (2025).

Para avaliar o desempenho do agrupamento em relação aos rótulos originais, foi construída a matriz de confusão apresentada a seguir. Essa matriz permite comparar as classificações obtidas pelo método de Ward com as classes reais do conjunto de dados, possibilitando uma análise quantitativa da coerência entre os agrupamentos identificados e as categorias verdadeiras.

Tabela 11 – Matriz de confusão entre os grupos obtidos pelo método Ward e os rótulos originais.

Rótulos Originais	Método Ward			
	Cluster 1	Cluster 2	Cluster 3	Total
Classe 1	59	0	0	59
Classe 2	5	58	8	71
Classe 3	0	0	48	48
Total	64	58	56	178

Fonte: Elaboração própria (2025).

Observa-se, a partir da matriz de confusão, que a correspondência entre os clusters formados e os rótulos reais é altamente consistente, podemos ver que das 59 observações pertencentes ao cluster 1 nos rótulos originais, o método de Ward classificou corretamente todas elas. Para o cluster 2, o desempenho foi inferior, isto é, das 71 observações, apenas 58 foram corretamente classificadas, evidenciando maior dificuldade na separação desse grupo. Já no cluster 3, o método obteve desempenho perfeito, classificando corretamente todas as 48 observações.

9.2.5.2 Método K-means

A partir da Figura 22, observa-se o resultado do agrupamento obtido pelo método K-means, considerando $k = 3$ grupos, valor determinado com base nos métodos do cotovelo e da silhueta média. A representação gráfica foi construída em duas dimensões por meio da Análise de Componentes Principais (ACP), uma vez que o conjunto de dados original possui 13 variáveis.

A ACP permitiu projetar os dados em um espaço bidimensional preservando a maior parte possível da variabilidade original. A primeira dimensão (PC1) explicou aproximadamente $36,2\%$ da variância total, enquanto a segunda dimensão (PC2) explicou $19,2\%$. Juntas, essas duas componentes capturam cerca de $55,4\%$ da estrutura dos dados, sendo suficientes para uma visualização clara da separação entre os grupos formados.

Além disso, foram incluídas elipses baseadas na distância euclidiana para representar a variabilidade interna de cada cluster, auxiliando na interpretação da compactação e da proximidade entre os grupos no espaço reduzido.

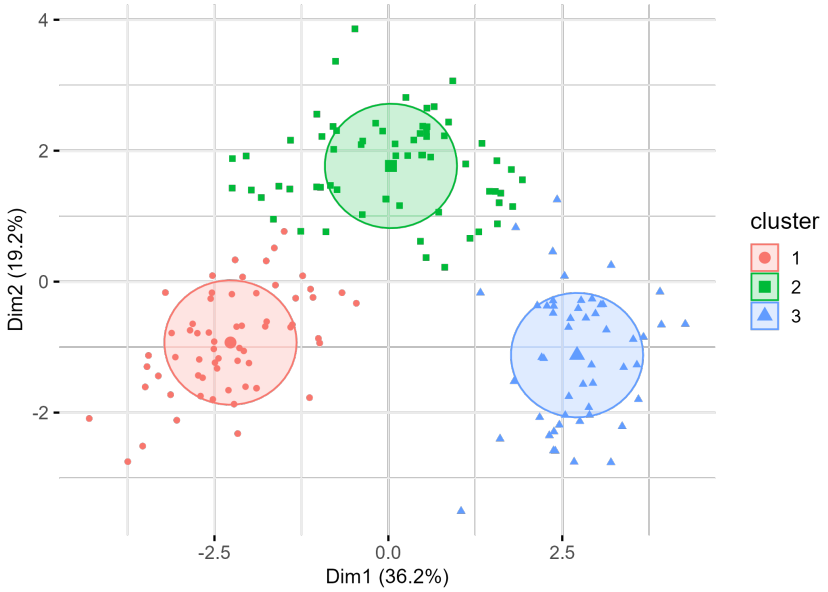

Figura 22 – Agrupamento obtido pelo k-means com três grupos.
Fonte: Elaboração própria (2025).

Com o objetivo de ampliar a análise visual, a Figura 25 apresenta uma representação tridimensional das três primeiras componentes principais. A inclusão da terceira componente (PC3), que explica $11,1\%$ da variabilidade total, permite observar a disposição espacial dos grupos em um contexto de maior dimensionalidade, preservando aproximadamente $66,5\%$.

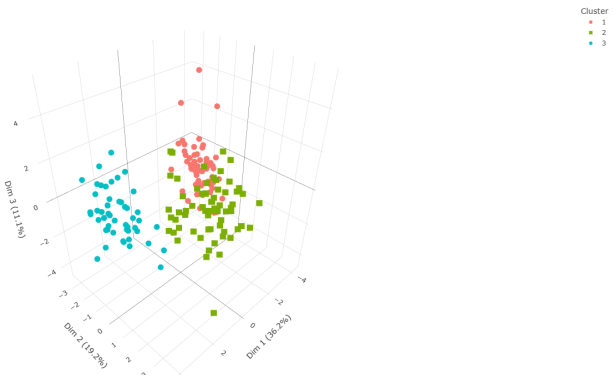

Figura 23 – Gráfico 3D do agrupamento obtido pelo k-means com três grupos.
Fonte: Elaboração própria (2025).

Para avaliar o desempenho do método em relação aos rótulos originais, apresenta-se, na Tabela 11, a matriz de confusão entre as classificações obtidas e as classes reais. Observa-se que o método apresentou resultados bastante consistentes, com elevada taxa de acerto na maioria dos grupos.

Tabela 12 – Matriz de confusão entre os grupos obtidos pelo método kmeans e os rótulos originais.

Rótulos Originais	Método k-means			
	Cluster 1	Cluster 2	Cluster 3	Total
Classe 1	59	0	0	59
Classe 2	3	65	3	71
Classe 3	0	0	48	48
Total	62	65	51	178

Fonte: Elaboração própria (2025).

Conforme se observa na matriz, das 59 observações pertencentes ao cluster 1 nos rótulos originais, o método K-means classificou todas corretamente. No cluster 2, composto por 71 observações, 65 foram corretamente alocadas, enquanto o cluster 3 apresentou classificação perfeita, com todas as 48 observações devidamente identificadas.

A maior dificuldade do K-means concentrou-se na separação entre o cluster 2 e os demais. Esse grupo ocupa uma região intermediária, situando-se entre os clusters 1 e 3. Essa proximidade entre os grupos pode ter favorecido pequenas confusões na classificação.

9.2.5.3 Modelo de Misturas Gaussianas

O modelo de Misturas Gaussianas com três componentes e estrutura de covariância do tipo *VVE* apresentou o melhor ajuste aos dados, segundo o critério de seleção adotado. Na Figura 24, é possível visualizar o resultado do agrupamento no espaço das duas primeiras componentes principais. As elipses representadas foram estimadas com base nos parâmetros do modelo, refletindo a forma, a orientação e a dispersão de cada grupo no espaço projetado.

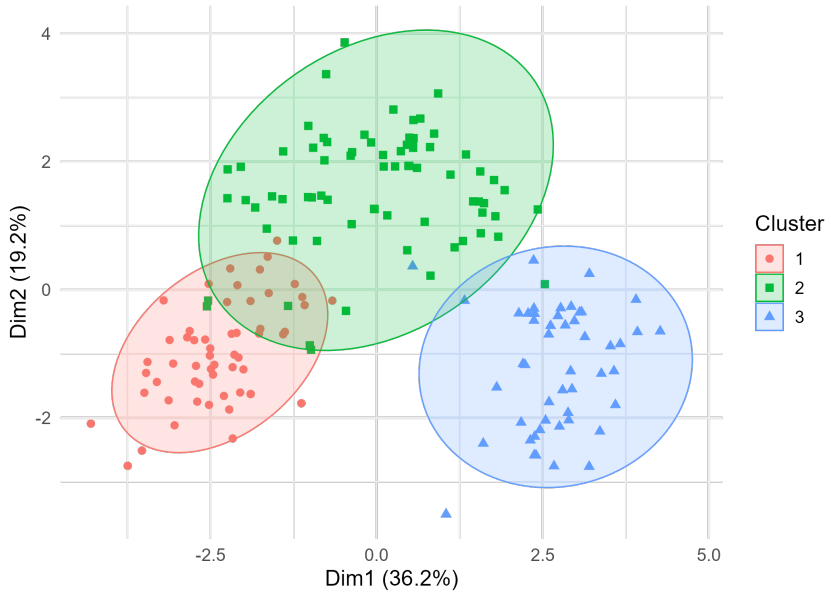

Figura 24 – Agrupamento obtido pelo modelo de Misturas Gaussianas (VVE, 3 componentes).
Fonte: Elaboração própria (2025).

O modelo de Misturas Gaussianas também foi representado no espaço tridimensional formado pelas três primeiras componentes principais. A Figura 25 apresenta essa visualização, permitindo observar a separação entre os grupos em um contexto de maior dimensionalidade. A terceira componente principal (PC3), responsável por $11,1\%$ da variabilidade total, acrescenta profundidade à análise gráfica ao revelar padrões que não são plenamente capturados pela representação bidimensional.

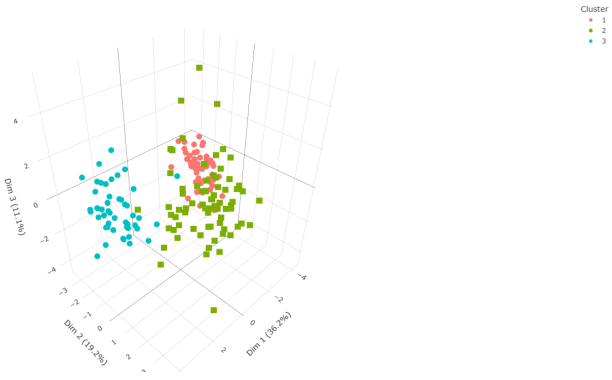

Figura 25 – Gráfico 3D do agrupamento obtido pelas misturas com três grupos.
Fonte: Elaboração própria (2025).

As estimativas dos parâmetros do modelo VVE com três componentes são apresentadas na Tabela 13. Além dos parâmetros de média e desvio-padrão marginal, o modelo também estima as proporções de cada componente, dadas por

$$\hat{p}_1 = 0.3131, \qquad \hat{p}_2 = 0.4114, \qquad \hat{p}_3 = 0.2755.$$

Essas proporções indicam a distribuição das observações entre os grupos, sugerindo que o segundo cluster é ligeiramente mais frequente, enquanto o terceiro é o menos representado. Para cada variável, apresentam-se as médias estimadas μ_k e os desvios-padrão marginais $\sigma_k = \sqrt{\text{diag}(\Sigma_k)}$, o que facilita a comparação direta entre os perfis internos dos clusters. Observa-se que os grupos exibem padrões distinctos, especialmente nas variáveis associadas à composição fenólica e às características físico-químicas do vinho, evidenciando que o modelo captura diferenças estruturais relevantes entre as amostras.

Tabela 13 – Estimativas de μ_j e σ_j do modelo de misturas finitas VVE ($\sigma_j = \sqrt{\text{diag}(\Sigma_j)}$).

Variáveis	**Cluster 1**		**Cluster 2**		**Cluster 3**	
	μ_1	σ_1	μ_2	σ_2	μ_3	σ_3
Ácido málico	-0.341	0.602	-0.323	1.011	0.871	1.075
Álcool	0.958	0.503	-0.833	0.767	0.155	0.752
Alcalinidade	-0.804	0.669	0.226	1.014	0.576	0.620
Cinza	0.260	0.726	-0.385	1.165	0.279	0.715
Fenóis totais	0.973	0.452	0.093	0.624	-1.244	0.697
Intensidade	0.238	0.470	-0.827	0.456	0.965	0.771
Magnésio	0.431	0.699	-0.341	1.170	0.019	0.722
Matiz (Hue)	0.476	0.474	0.426	0.862	-1.178	0.473
Não flavonoides	-0.620	0.670	0.044	0.914	0.639	0.862
OD280/OD315	0.776	0.481	0.274	0.640	-1.290	0.612
Phenols	0.886	0.548	-0.013	0.828	-0.987	0.596
Proantocianinas	0.548	0.700	0.084	0.998	-0.749	0.671
Prolina	1.237	0.648	-0.692	0.548	-0.372	0.410

Fonte: Elaboração própria (2025).

Por fim, apresenta-se a matriz de confusão, que relaciona os rótulos verdadeiros das observações com aqueles estimados pelo modelo de Misturas Gaussianas:

Tabela 14 – Matriz de confusão entre os grupos obtidos pelo método de misturas e os rótulos originais.

Rótulos Originais	Método Misturas			
	Cluster 1	Cluster 2	Cluster 3	Total
Classe 1	56	3	0	59
Classe 2	0	70	1	71
Classe 3	0	0	48	48
Total	56	73	49	178

Fonte: Elaboração própria (2025).

A matriz de confusão evidencia que, das 59 observações originalmente pertencentes ao cluster 1, o modelo classificou corretamente 56, alocando 3 incorretamente no cluster 2. Para o cluster 2, composto por 71 observações, 70 foram classificadas corretamente e apenas uma foi atribuída ao cluster 3. Já o cluster 3 apresentou desempenho perfeito, com todas as 48 observações corretamente identificadas.

Esses resultados indicam um desempenho bastante satisfatório do modelo de Misturas Gaussianas, sobretudo quando comparado aos métodos de Ward e K-means. Observa-se que o cluster 2, que havia apresentado maior dificuldade de separação nos métodos anteriores, foi classificado de forma quase totalmente correta, com apenas um erro de alocação. Esse comportamento reforça a flexibilidade do modelo de misturas, capaz de capturar estruturas de dados mais complexas e fronteiras não lineares entre os grupos.

9.2.6 Resultados e Discussões

Nesta seção, são apresentados os resultados obtidos a partir da aplicação dos critérios de validação externa às três técnicas de agrupamento estudadas. Tais critérios permitem avaliar o grau de concordância entre os agrupamentos gerados pelos métodos e os rótulos verdadeiros das observações, funcionando, portanto, como uma medida de desempenho dos modelos. Os índices utilizados foram o Índice de Rand Ajustado (ARI) e a acurácia média. A Tabela 13 apresenta os valores obtidos para cada técnica.

Tabela 15 – Critérios de validação externa aplicados aos métodos de agrupamento.

Método	ARI	Acurácia
Ward	0,77	0,92
K-means	0,90	0,97
Misturas	0,93	0,98

Fonte: Elaboração própria (2025).

O modelo de misturas finitas apresentou o melhor desempenho entre os três métodos avaliados, com um ARI de 0,93 e acurácia de 0,98, evidenciando uma excelente concordância com os rótulos reais. O método K-means também obteve resultados satisfatórios, com ARI de 0,90 e acurácia de 0,97, apresentando desempenho próximo ao das Misturas Finitas, embora ligeiramente inferior. Já o método hierárquico de Ward apresentou desempenho inferior aos demais, com ARI de 0,77 e acurácia de 0,92, indicando maior dificuldade em reproduzir corretamente os agrupamentos verdadeiros.

Com base nesses indicadores, conclui-se que o modelo de misturas finitas foi o mais eficaz na identificação das estruturas reais presentes nos dados, seguido de perto pelo k-means. O desempenho inferior do método de Ward pode estar relacionado à sua

natureza hierárquica, que, neste contexto, mostrou-se menos adequada para representar as fronteiras entre os grupos.

A partir desses resultados, prossegue-se com a interpretação dos grupos formados pelo modelo de misturas finitas, buscando identificar os padrões e características que distinguem cada agrupamento.

Tabela 16 – Análise descritiva dos agrupamentos encontrados pelo método de misturas finitas.

Variáveis	**Cluster 1**		**Cluster 2**		**Cluster 3**	
	Média	DP	Média	DP	Média	DP
Ácido málico	1,96	0,65	1,97	1,03	3,31	1,09
Álcool	13,78	0,45	12,32	0,56	13,13	0,56
Alcalinidade	16,82	2,31	20,26	3,36	21,42	2,23
Cinza	2,44	0,20	2,26	0,33	2,44	0,19
Fenóis totais	3,00	0,40	2,12	0,69	0,79	0,29
Intensidade	5,61	1,22	3,14	0,93	7,30	2,39
Magnésio	105,88	10,38	94,85	16,49	100,02	11,86
Matiz (Hue)	1,07	0,11	1,06	0,20	0,69	0,12
Não flavonoides	0,28	0,07	0,37	0,12	0,44	0,13
OD280/OD315	3,16	0,36	2,80	0,49	1,70	0,28
Proantocianinas	1,90	0,42	1,64	0,59	1,16	0,41
Prolina	1135,39	209,24	527,47	161,71	629,80	113,89

Fonte: Elaboração própria (2025).

A partir da Tabela 16, observa-se que os três agrupamentos obtidos pelo modelo de misturas finitas apresentam perfis químicos distintos, refletindo a capacidade do método em identificar estruturas bem definidas nos dados.

O Cluster 1 caracteriza-se por apresentar os maiores teores médios de álcool (13,78), fenóis totais (3,00), proantocianinas (1,90) e relação OD280/OD315 (3,16), além de elevados níveis de prolina (1135,39). Esses indicadores estão associados à maior qualidade e intensidade fenólica, sugerindo que este grupo corresponde a vinhos mais encorpados e de maior teor alcoólico. A coloração também se destaca, com valores intermediários de intensidade (5,61) e matiz (1,07), denotando equilíbrio entre pigmentos e compostos aromáticos.

O Cluster 2, por sua vez, apresenta valores intermediários de álcool (12,32) e fenóis totais (2,12), mas destaca-se por maior alcalinidade (20,26) e menor concentração de magnésio (94,85). Além disso, possui teores mais altos de compostos não flavonoides (0,37). Esse perfil sugere vinhos com características químicas mais suaves, possivelmente de menor corpo e estrutura tânica, refletindo uma categoria intermediária entre os grupos extremos.

Já o Cluster 3 apresenta um perfil bastante distinto, com teores reduzidos de fenóis totais (0,79), OD280/OD315 (1,70) e matiz (0,69), além de alta intensidade de cor (7,30) e elevado teor de ácido málico (3,31). Esses resultados indicam vinhos de acidez mais

pronunciada e menor maturação fenólica, o que pode estar relacionado a produtos de colheita precoce ou com menor grau de envelhecimento. Os níveis de prolina e magnésio também são inferiores aos do Cluster 1, reforçando a distinção em termos de complexidade química.

Essas diferenças demonstram a capacidade do modelo probabilístico de misturas gaussianas em representar a heterogeneidade da amostra, atribuindo probabilidades de pertencimento que refletem bem a estrutura subjacente dos dados.

9.3 SEGMENTAÇÃO DE IMAGENS: CONJUNTO *FRUITS*

O conjunto de dados *Fruits* consiste em uma imagem contendo diferentes frutas, tratada como uma matriz de pixels, em que cada observação é representada pelas intensidades das cores nas escalas de vermelho (R), verde (G) e azul (B). Dessa forma, cada pixel é considerado como uma unidade de análise com três variáveis associadas.

9.3.1 Objetivo da Análise

O objetivo desta aplicação é realizar a segmentação da imagem por meio de algoritmos de agrupamento, de modo a comparar o desempenho das técnicas de K-means, Misturas de Gaussianas e do método hierárquico de Ward. A análise busca avaliar como cada abordagem organiza os pixels em grupos de cores semelhantes e, consequentemente, como se dá a reconstrução da imagem a partir desses agrupamentos.

9.3.2 Pré-processamento dos Dados

O primeiro passo consistiu no pré-processamento da imagem digital. A figura foi carregada em formato matricial e decomposta em suas três componentes de cor no espaço RGB (vermelho, verde e azul). Cada pixel da imagem foi então representado como uma observação em um espaço tridimensional, definido pelos valores das intensidades de cor. Esse procedimento permitiu transformar a imagem em um conjunto de dados estruturado, adequado para aplicação dos algoritmos de agrupamento.

Para aplicar o método hierárquico de Ward, foi necessário realizar um ajuste metodológico em função do tamanho da imagem utilizada. A imagem original possui dimensão de 640**x**470 pixels, o que corresponde a 300.800 observações (pixels). Como o algoritmo hierárquico depende do cálculo de uma matriz de distâncias entre todos os pares de observações, o custo computacional cresce quadraticamente em relação ao número de pixels. Nesse caso, a matriz de distâncias exigiria centenas de gigabytes de memória, inviabilizando a análise. Assim, selecionou-se aleatoriamente uma amostra representativa de 5.000 pixels da imagem, sobre a qual foi calculada a matriz de distâncias e realizada a clusterização hierárquica. Posteriormente, os centróides obtidos em cada agrupamento

foram utilizados para rotular todos os demais pixels da imagem, por meio da atribuição ao centróide mais próximo. Esse procedimento permitiu aplicar o método de Ward de forma viável, mantendo a qualidade da segmentação e assegurando a representatividade da amostra. A escolha de uma amostra de 5.000 pixels baseou-se na necessidade de equilibrar representatividade e viabilidade computacional. Esse tamanho permite capturar adequadamente a variabilidade de cores presente na imagem, garantindo que todas as regiões sejam contempladas, ao mesmo tempo em que mantém o custo computacional do método de Ward dentro de limites operacionais. Com 5.000 observações, a matriz de distâncias apresenta 25 milhões de pares, valor administrável em memória e compatível com a execução eficiente do algoritmo, diferentemente do conjunto completo de 300.800 pixels, cuja matriz de distâncias exigiria centenas de gigabytes de RAM.

A análise foi conduzida considerando diferentes valores para o número de grupos K. Foram testados os valores de $K = 2, 3, \ldots, 8$, de forma a ilustrar como cada técnica se comporta em distintos níveis de segmentação da imagem. Essa variação é importante, pois possibilita observar tanto segmentações mais gerais (com poucos grupos, destacando grandes regiões homogêneas) quanto segmentações mais detalhadas (com maior número de grupos, capturando sutilezas de variação de cor).

9.3.3 Aplicação das Técnicas de Agrupamento

9.3.3.1 Método Hierárquico de Ward

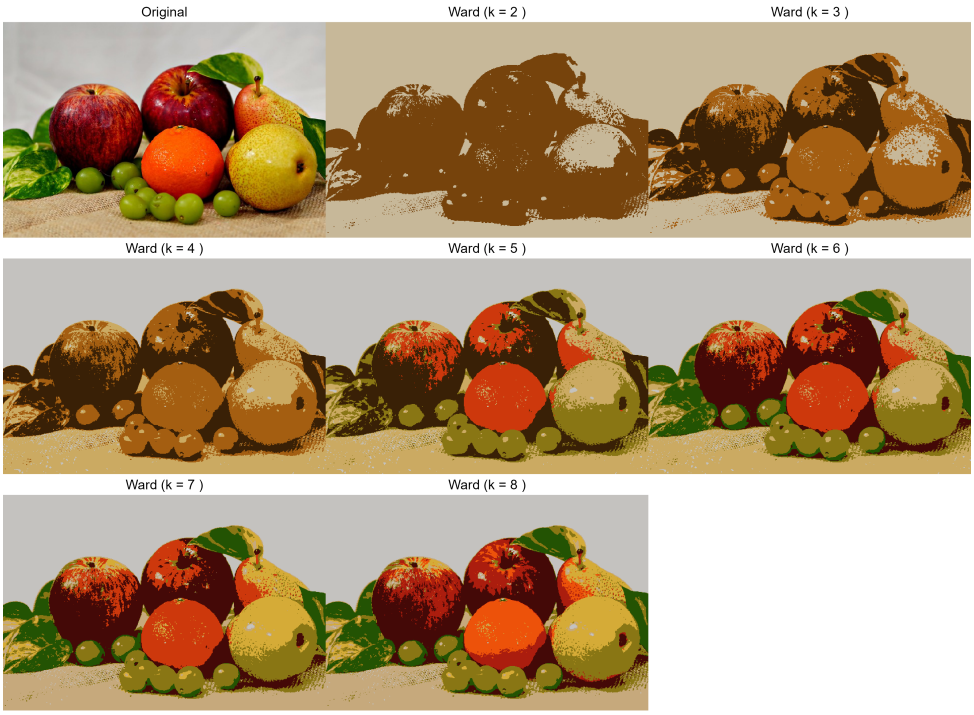

Figura 26 – Segmentação por Ward para vários k

9.3.3.2 Método K-means

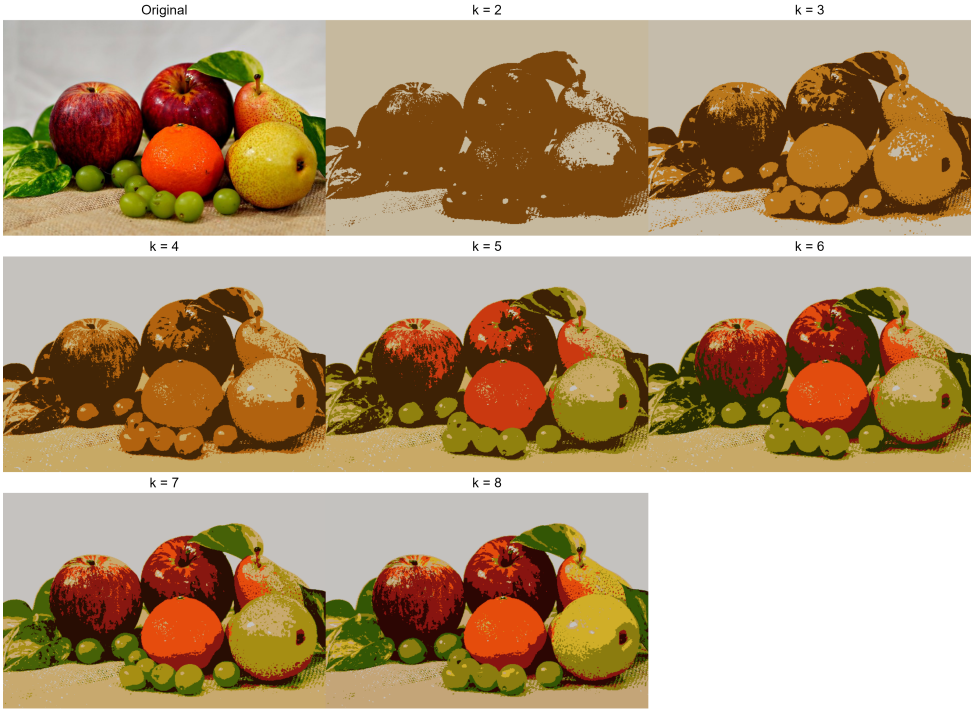

Figura 27 – Segmentação por K-means para vários k

9.3.3.3 Método Misturas Finitas de Gaussianas

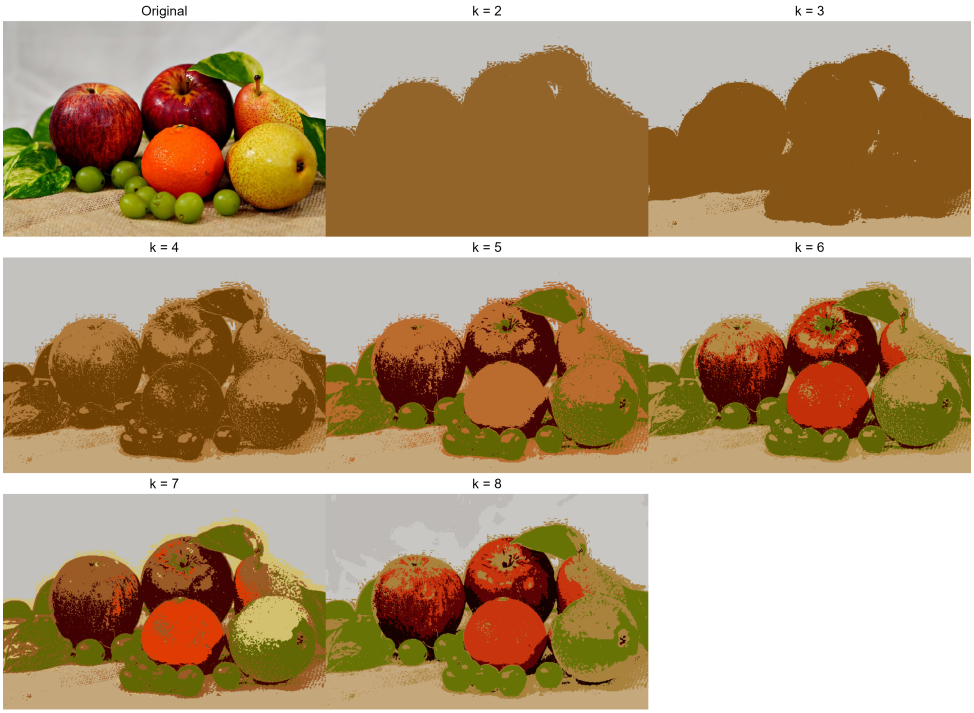

Figura 28 – Segmentação por Misturas de Gaussianas (GMM) para vários k

9.3.4 Resultados e Discussões

A segmentação da imagem de frutas possibilitou comparar o desempenho dos três métodos de agrupamento sob dois aspectos principais: *qualidade visual da segmentação* e *tempo de processamento*. Não foram utilizadas medidas internas de validação, como silhueta, Dunn ou Calinski–Harabasz, uma vez que a natureza dos dados, todos os pixels da imagem torna impraticável o cálculo de matrizes de distância completas, que ultrapassariam centenas de gigabytes de memória.

O K-means apresentou o desempenho mais equilibrado entre os métodos. Visualmente, produziu segmentações limpas, com boa separação das regiões de cor e baixo nível de ruído, especialmente para valores reduzidos de k. Além disso, foi o método mais rápido entre os três, levando aproximadamente 5,7 segundos para processar toda a imagem. No contexto analisado, destacou-se pela combinação de simplicidade, rapidez e resultados visualmente satisfatórios.

As Misturas de Gaussianas, embora ofereçam maior flexibilidade ao modelar distribuições elípticas de diferentes formas e orientações, apresentaram maior sensibilidade ao ruído e segmentações menos nítidas quando comparadas ao K-means.O modelo selecionado pelo critério BIC foi o VEV, que no Mclust representa uma estrutura de covariância que permite formas e orientações variáveis entre os componentes, porém com volumes iguais. Apesar de sua sofisticação, o tempo de processamento foi substancialmente maior, cerca de 233 segundos, sem apresentar ganhos visuais proporcionais.

O método de Ward, por sua vez, impôs limitações computacionais devido à necessidade de calcular a matriz de distâncias entre todos os pixels, o que é inviável para imagens de resolução típica. Para contornar essa restrição, foi utilizada uma amostra representativa dos pixels para construir a hierarquia, seguida da classificação dos demais pixels por proximidade aos centróides dos grupos formados. Essa estratégia tornou o método operacionalmente viável e permitiu produzir segmentações visualmente interpretáveis. O tempo total de execução foi de aproximadamente 2.78 segundos, tornando-o competitivo em termos de velocidade. Entretanto, sua dependência da etapa de amostragem reforça suas limitações quando aplicado diretamente a imagens completas.

Em síntese, a comparação visual e temporal indica que, para segmentação de imagens coloridas, o K-means oferece o melhor comprometimento entre custo computacional e clareza dos resultados, enquanto o GMM apresenta maior complexidade e instabilidade, e o método de Ward mostra-se limitado por questões de escalabilidade, embora apresente bom desempenho ao ser adaptado via amostragem.

10 CONCLUSÃO

O presente trabalho teve como objetivo comparar três técnicas de agrupamento amplamente utilizadas em estatística e aprendizado de máquina, sendo eles, o K-means, Método de Ward e Misturas de Gaussianas, avaliando seu desempenho em três contextos distintos de dados reais: classificação de vinhos, segmentação de imagens e análise dos padrões eruptivos do gêiser Old Faithful. Todas as análises foram integralmente implementadas em linguagem `R`, e o código completo foi disponibilizado em repositório público no GitHub, garantindo a reprodutibilidade e a transparência dos resultados.

Os experimentos demonstraram que não existe um método de agrupamento universalmente superior; a eficácia de cada técnica depende intrinsecamente das propriedades do conjunto de dados e da finalidade da análise. No conjunto de vinhos, as Misturas de Gaussianas apresentaram o melhor desempenho, capturando estruturas complexas de variabilidade. Na segmentação de imagens, o K-means sobressaiu-se devido à sua eficiência computacional para processar grandes volumes de dados (pixels), gerando regiões visualmente consistentes. Já na análise do Old Faithful, o Método de Ward mostrou-se o mais robusto, apresentando a melhor separabilidade entre os grupos e identificando com clareza a natureza bimodal das erupções.

De forma geral, os resultados reforçam a importância de uma análise criteriosa na escolha de técnicas de agrupamento. Cada método apresenta vantagens e limitações específicas, devendo a decisão ponderar tanto as características estatísticas dos dados quanto requisitos práticos. O estudo evidenciou o potencial do agrupamento como ferramenta versátil, aplicável desde a caracterização química de produtos e compressão de imagens até o entendimento de fenômenos naturais geofísicos.

Por fim, este trabalho aponta caminhos promissores para investigações futuras que visem superar as limitações aqui encontradas. Uma estratégia relevante para enriquecer a interpretação visual dos clusters, não abordada neste estudo, seria a aplicação do Escalonamento Multidimensional (MDS), que permitiria visualizar a separação dos grupos em dimensões reduzidas de forma mais intuitiva. Além disso, recomenda-se a utilização do Critério de Agrupamento Cúbico (CCC - Cubic Clustering Criterion) como uma métrica robusta para a determinação automática do número ideal de grupos. Outro ponto crucial para expansão da pesquisa é a comparação utilizando outras métricas de dissimilaridade, como a distância de Mahalanobis. Dado que este trabalho se limitou à distância Euclidiana, o uso de Mahalanobis seria especialmente benéfico para tratar a correlação entre variáveis e diferenças de escala sem depender exclusivamente da padronização prévia dos dados.

O trabalho também abre espaço para a exploração de outros paradigmas de clustering, tais como métodos baseados em densidade (DBSCAN, OPTICS) ou abordagens fundamentadas em redes neurais profundas. A inclusão de métricas de validação mais

sofisticadas, bem como a aplicação dos modelos em bases de dados maiores e mais heterogêneas, pode ampliar ainda mais a robustez das conclusões e contribuir significativamente para o avanço das técnicas de análise de agrupamentos.

REFERÊNCIAS

ANDERBERG, M. R. **Cluster analysis for applications**. [S.l.]: Academic Press, 1973.

ARABIE, P.; HUBERT, L. J. Cluster analysis in marketing research. In: BAGOZZI, R. P. (Ed.). **Advanced methods of marketing research**. Cambridge, MA: Blackwell, 1994. p. 160–189.

ARTES, R.; BARROSO, L. P. **Métodos multivariados de análise estatística**. São Paulo: Blucher, 2023.

BALLARD, D.; BROWN, C. **Computer Vision**. [S.l.]: Prentice-Hall Inc., 1982.

BANFIELD, J. D.; RAFTERY, A. E. Model-based gaussian and non-gaussian clustering. **Biometrics**, JSTOR, p. 803–821, 1993.

BASSO, R. M. **Misturas Finitas de Misturas de Escala Skew-Normal**. Dissertação (Mestrado) — Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica, Departamento de Estatística, Campinas, SP, 2009. Dissertação (Mestrado) — Orientador: Prof. Dr. Victor Hugo Lachos Dávila.

BENITES, L. *et al.* Finite mixture of regression models based on multivariate scale mixtures of skew-normal distributions. **Computational Statistics**, Springer, v. 40, p. 5163–5194, 2025.

BHOLOWALIA, P.; KUMAR, A. Ebk-means: A clustering technique based on elbow method and k-means in wsn. **International Journal of Computer Applications**, Citeseer, v. 105, n. 9, 2014.

BISHOP, C. M.; NASRABADI, N. M. **Pattern recognition and machine learning**. [S.l.]: Springer, 2006. v. 4.

BUSSAB, W. d. O.; MIAZAKI, S. E.; ANDRADE, D. F. Introdução à análise de agrupamento. In: **IX Simpósio Brasileiro de Probabilidade e Estatística**. São Paulo: IME-USP, 1990. p. 105.

CALIŃSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics - Theory and Methods**, Taylor & Francis, v. 3, n. 1, p. 1–27, 1974.

CASTLEMAN, K. R. **Digital Image Processing**. [S.l.]: Prentice Hall Inc., 1996.

CELEUX, G.; GOVAERT, G. Gaussian parsimonious clustering models. **Pattern Recognition**, Elsevier, v. 28, n. 5, p. 781–793, 1995.

CONCI, A.; AZEVEDO, E.; LETA, F. **Computação Gráfica: Teoria e Prática**. S.l.: Elsevier, 2008. v. 2.

CORMEN, T. H. *et al.* **Introduction to Algorithms**. 3. ed. [S.l.]: MIT Press, 2009.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm (with discussion). **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 39, p. 1–38, 1977.

DISTEFANO, V.; MAMELI, V.; POLI, I. Identifying spatial patterns with the bootstrap clustgeo technique. **Spatial Statistics**, Elsevier, v. 38, p. 100441, 2020.

DONI, M. V. Análise de cluster: métodos hierárquicos e de particionamento. **Universidade Presbiteriana Mackenzie**, 2004.

DRINEAS, P. *et al.* Clustering large graphs via the singular value decomposition. **Machine Learning**, v. 56, n. 1-3, p. 9–33, 1999.

DUBIEN, J. L.; WARDE, W. D. A mathematical comparison of the members of an infinite family of agglomerative clustering algorithms. **Canadian Journal of Statistics**, v. 7, p. 29–38, 1979.

DUDA, R. O.; HART, P. E. **Pattern Classification and Scene Analysis**. New York: Wiley, 1988.

DáVILA, V. H. L.; CABRAL, C. R. B.; ZELLER, C. B. **Finite Mixture of Skewed Distributions**. [S.l.]: Springer, 2018. (Springer Briefs in Statistics).

EVERITT, B. S. **Cluster Analysis, 3rd edn., Edward Arnold**. [S.l.]: University Press, Cambridge, UK, 1993.

EVERITT, B. S.; HAND, D. J. Mixtures of normal distributions. In: **Finite Mixture Distributions**. [S.l.]: Springer, 1981. p. 25–57.

EVERITT, B. S. *et al.* **Cluster Analysis**. 5. ed. Chichester, UK: Wiley, 2011.

FACELI, K.; CARVALHO, A. C. P. L. F. d.; SOUTO, M. C. P. d. **Validação de algoritmos de agrupamento**. São Carlos, SP, 2005.

FARIA, P. N. **Utilização de técnicas multivariadas na análise da divergência genética via modelo AMMI com reamostragem bootstrap**. Tese (Doutorado) — Universidade de São Paulo, 2012.

FÁVERO, L. P. L.; BELFIORE, P. **Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®**. Rio de Janeiro: Elsevier Brasil, 2017.

FISHER, L.; NESS, J. W. V. Admissible clustering procedures. **Biometrika**, Oxford University Press, v. 58, n. 1, p. 91–104, 1971.

FRALEY, C.; RAFTERY, A. **Mclust: software para agrupamento baseado em modelo, estimativa de densidade e análise discriminante**. [S.l.], 2002.

FRIGUI, H.; KRISHNAPURAM, R. Clustering by competitive agglomeration. **Pattern recognition**, Elsevier, v. 30, n. 7, p. 1109–1119, 1997.

FRÜHWIRTH-SCHNATTER, S. **Finite mixture and Markov switching models**. [S.l.]: Springer, 2006.

FU, K. S.; MUI, J. K. A survey on image segmentation. **Pattern Recognition**, [S.l.], v. 13, p. 3–16, 1980.

GAN, G.; MA, C.; WU, J. **Data clustering: theory, algorithms, and applications**. Philadelphia: Society for Industrial and Applied Mathematics, 2007. Disponível em: <https://epubs.siam.org/doi/abs/10.1137/1.9780898718348>. Acesso em: 24 jun. 2025. Disponível em: <https://epubs.siam.org/doi/abs/10.1137/1.9780898718348>.

GIOLO, S. R. **Introdução à análise de dados categóricos com aplicações**. 1. ed. [S.l.]: Blucher, 2017.

GONZALEZ, R. C.; WOODS, R. E. **Processamento Digital de Imagens**. 3. ed. S.l.: Pearson Prentice Hall, 2010.

HAIR, J. F. *et al.* **Análise multivariada de dados**. Porto Alegre: Bookman, 2005. 593 p.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. Cluster validity methods: Part ii. **SIGMOD Record**, v. 31, n. 3, p. 19–27, set. 2002.

HARTIGAN, J. Clustering algorithms. **John Wiley google schola**, v. 2, p. 25–47, 1975.

HARTIGAN, J. A.; WONG, M. A. Algorithm as 136: A k-means clustering algorithm. **Journal of the royal statistical society. series c (applied statistics)**, JSTOR, v. 28, n. 1, p. 100–108, 1979.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. **The elements of statistical learning: data mining, inference, and prediction**. 2. ed. New York: Springer, 2009. v. 2.

HOEF, H. Van der; WARRENS, M. J. Understanding information theoretic measures for comparing clusterings. **Behaviormetrika**, Springer, v. 46, n. 2, p. 353–370, 2019.

IWAYAMA, M.; TOKUNAGA, T. Cluster-based text categorization: A comparison of category search strategies. In: ACM. **Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '95)**. [S.l.], 1995. p. 273–280.

IZBICKI, R.; SANTOS, T. M. dos. **Aprendizado de máquina: uma abordagem estatística**. [S.l.]: Rafael Izbicki, 2020.

JAIN, A. K. Data clustering: 50 years beyond k-means. **Pattern recognition letters**, Elsevier, v. 31, n. 8, p. 651–666, 2010.

JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data**. Englewood Cliffs: Prentice Hall, 1988.

JAIN, A. K.; FLYNN, P. J. **Image segmentation using clustering**. [S.l.]: IEEE Press, Piscataway, NJ, 1996. 65–83 p.

JAMES, G. *et al.* **An Introduction to Statistical Learning: with Applications in R**. New York: Springer, 2013.

JARDINE, N.; SIBSON, R. The construction of hierarchic and non-hierarchic classifications. **The Computer Journal**, Oxford University Press, v. 11, n. 2, p. 177–184, 1968.

JOHNSON, R. A.; WICHERN, D. W. *et al.* **Applied multivariate statistical analysis**. 5. ed. ed. [S.l.]: Prentice hall Upper Saddle River, NJ, 2002.

JOHNSON, S. C. Hierarchical clustering schemes. **Psychometrika**, Springer, v. 32, n. 3, p. 241–254, 1967.

KALKSTEIN, L. S.; TAN, G.; SKINDLOV, J. A. An evaluation of three clustering procedures for use in synoptic climatological classification. **Journal of Applied Meteorology and Climatology**, v. 26, n. 6, p. 717–730, 1987.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding Groups in Data: An Introduction to Cluster Analysis**. 4. ed. New York: Wiley, 1990.

LANCE, G. N.; WILLIAMS, W. T. A general theory of classificatory sorting strategies: 1. hierarchical systems. **The Computer Journal**, Oxford University Press, v. 9, n. 4, p. 373–380, 1968.

LINDEN, R. Técnicas de agrupamento. **Revista de Sistemas de Informação da FSMA**, v. 4, n. 4, p. 18–36, 2009.

MARQUES, O.; VIEIRA, H. **Processamento Digital de Imagens**. [S.l.]: Brasport, 1999. ISBN 8574520098.

MARQUEZ, F. P. G. **Handbook of research on big data clustering and machine learning**. Hershey: IGI Global, 2019. Disponível em: <https://www.igi-global.com/book/handbook-research-big-data-clustering/223751>. Acesso em: 24 jun. 2025.

MCLACHLAN, G. J.; BASFORD, K. E. **Mixture models: Inference and applications to clustering**. [S.l.]: M. Dekker New York, 1988. v. 38.

MCLACHLAN, G. J.; KRISHNAN, T. **The EM Algorithm and Extensions**. 2. ed. Hoboken: John Wiley & Sons, 2007.

MCLACHLAN, G. J.; PEEL, D. **Finite mixture models**. New York: Wiley, 2000.

MEILA, M. The uniqueness of a good optimum for k-means. In: **Proceedings of the 23rd International Conference on Machine Learning**. [S.l.: s.n.], 2006. p. 625–632.

MILLIGAN, G.; COOPER, M. An examination of procedures for determining the number of clusters in a data set. **Psychometrika**, v. 50, p. 159–179, 1985.

MILLIGAN, G. W. Ultrametric hierarchical clustering algorithms. **Psychometrika**, Springer, v. 44, n. 3, p. 343–346, 1979.

MINGOTI, S. A. **Análise de dados através de métodos estatísticos multivariados: uma abordagem aplicada**. Belo Horizonte: Editora UFMG, 2005.

MIRFARAH, E.; NADERI, M.; CHEN, D.-G. Finite mixture of regression models for censored data based on scale mixtures of normal distributions. **Computational Statistics and Data Analysis**, v. 158, p. 107–182, 2021.

MONGELO, A. I. **Validação de método baseado em visão computacional para automação da contagem de viabilidade de leveduras em indústrias alcooleiras**. Dissertação (Dissertação (Mestrado em Biotecnologia)) — Universidade Católica Dom Bosco, Programa de Pós-Graduação em Biotecnologia, Campo Grande, Mato Grosso do Sul, 2012.

MORETTIN, P.; SINGER, J. da M. **Estatística e Ciência de Dados**. [S.l.]: LTC, 2022. ISBN 9788521638162.

MURPHY, K. P. **Machine learning: a probabilistic perspective**. Cambridge, MA: MIT Press, 2013. ISBN 9780262018029.

PEREIRA, J. R. G. **Um estudo sobre alguns métodos hierárquicos para análise de agrupamentos**. Dissertação (Tese (Mestrado em Estatística)) — Universidade Estadual de Campinas, Campinas, São Paulo, 1993. Orientadora: Profa. Dra. Gabriela Stangenhaus.

QUEIROZ, J.; GOMES, H. Introdução ao processamento digital de imagens. **VIII**, n. 1, 2001.

REN, Y. *et al.* Deep clustering: A comprehensive survey. **IEEE Transactions on Neural Networks and Learning Systems**, IEEE, 2024. Disponível em: <https://ieeexplore.ieee.org/document/10506283>. Acesso em: 24 jun. 2025.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, Elsevier, v. 20, p. 53–65, 1987.

SAHAMI, M. **Using machine learning to improve information access**. [S.l.]: stanford university, 1999.

SARLE, W. S. **Cubic Clustering Criterion**. Technical report a-108. Cary, NC: SAS Institute, 1983. v. 108. (SAS Technical Report, v. 108).

SCHWARZ, G. Estimating the dimension of a model. **Annals of Statistics**, v. 6, p. 461–464, 1978.

SCRUCCA, L. *et al.* mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. **The R Journal**, v. 8, n. 1, p. 289–317, ago. 2016.

SCRUCCA, L.; RAFTERY, A. E. Inicialização aprimorada de agrupamento baseado em modelo usando partições hierárquicas gaussianas. **Advances in Data Analysis and Classification**, Springer, v. 9, p. 447–460, 2015.

SCURI, A. E. **Fundamentos da Imagem Digital**. S.l.: s.n., 2002.

SEIDEL, E. J. *et al.* Comparação entre o método ward e o método k-médias no agrupamento de produtores de leite. **Ciência e Natura**, p. 07–15, 2008.

SHI, J.; MALIK, J. Normalized cuts and image segmentation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 22, n. 8, p. 888–905, 2000.

SOKAL, R. R.; SNEATH, P. H. A. **Principles of Numerical Taxonomy**. San Francisco: W. H. Freeman, 1963.

SOUZA, D. C. d. **Análise de agrupamentos para dados espaciais: estudo de simulação e aplicação a dados educacionais do estado do Paraná**. Dissertação (Mestrado) — Universidade Federal do Paraná, Curitiba, 2010. Dissertação (Mestrado em Ciências) — Curso de Pós-Graduação em Métodos Numéricos em Engenharia, Setor de Tecnologia e Ciências Exatas, Departamento de Construção Civil e Matemática. Orientador: Prof. Dr. Cesar Augusto Taconeli.

TIBSHIRANI, R.; WALTHER, G.; HASTIE, T. Estimating the number of clusters in a data set via the gap statistic. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 63, n. 2, p. 411–423, 2001.

TITTERINGTON, D. M. Common structure of smoothing techniques in statistics. **International Statistical Review/Revue Internationale de Statistique**, JSTOR, p. 141–170, 1985.

TITTERINGTON, D. M.; SMITH, A. F. M.; MAKOV, U. E. **Statistical Analysis of Finite Mixture Distributions**. [S.l.]: John Wiley & Sons, 1985.

WANG, Y. *et al.* Density peak clustering algorithms: A review on the decade 2014–2023. **Expert Systems with Applications**, Elsevier, v. 238, p. 121860, 2024. Disponível em: <https://doi.org/10.1016/j.eswa.2024.121860>. Acesso em: 24 jun. 2025.

WARD, J. H. J. Hierarchical grouping to optimize an objective function. **Journal of the American Statistical Association**, Taylor & Francis, v. 58, n. 301, p. 236–244, 1963.

WISHART, D. Note: An algorithm for hierarchical classifications. **Biometrics**, JSTOR, p. 165–170, 1969.

WONG, M. A. A hybrid clustering method for identifying high-density clusters. **Journal of the American Statistical Association**, Taylor & Francis, v. 77, n. 378, p. 841–847, 1982.

WONG, M. A.; LANE, T. A kth nearest neighbour clustering procedure. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley, v. 45, n. 3, p. 362–368, 1983.

ZELLER, C. B. *et al.* Finite mixture of regression models for censored data based on scale mixtures of normal distributions. **Advances in Data Analysis and Classification**, Springer, 2018.

ZHOU, S. *et al.* A comprehensive survey on deep clustering: Taxonomy, challenges, and future directions. **ACM Computing Surveys**, ACM, v. 57, n. 3, p. 1–38, 2024. Disponível em: <https://doi.org/10.1145/3652287>. Acesso em: 24 jun. 2025.