UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

MARIA SOCORRO LIRA LEITE

Algoritmos de Agrupamento Espectral para Formas Planas

Recife

2025

MARIA SOCORRO LIRA LEITE

Algoritmos de Agrupamento Espectral para Formas Planas

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Estatística.

Área de Concentração: Estatística Aplicada

Orientador (a): Getúlio José Amorim do Amaral

Coorientador (a): Marcelo Rodrigo Portela Ferreira

Recife

2025

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Leite, Maria Socorro Lira.
Algoritmos de agrupamento espectral para formas planas / Maria Socorro Lira Leite. - Recife, 2025.
81f.: il.

Dissertação (Mestrado) - Universidade Federal de Pernambuco, Ciências Exatas e da Natureza, Pós-Graduação em Estatística, 2025.
Orientação: Getúlio José Amorim do Amaral.
Coorientação: Marcelo Rodrigo Portela Ferreira.
Inclui referências.

1. Análise de agrupamento; 2. Estatística de forma; 3. Agrupamento Espectral. I. Amaral, Getúlio José Amorim do. II. Ferreira, Marcelo Rodrigo Portela. III. Título.

UFPE-Biblioteca Central

MARIA SOCORRO LIRA LEITE

ALGORITMOS DE AGRUPAMENTO ESPECTRAL PARA FORMAS PLANAS

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestra em Estatística.

Aprovado em: 20 de fevereiro de 2025.

BANCA EXAMINADORA

gov.br
Documento assinado digitalmente
GETULIO JOSE AMORIM DO AMARAL
Data: 25/02/2025 11:38:51-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Getúlio José Amorim do Amaral
Presidente (Orientador), UFPE

gov.br
Documento assinado digitalmente
JODAVID DE ARAUJO FERREIRA
Data: 25/02/2025 11:24:57-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Jodavid de Araújo Ferreira
Examinador Interno, UFPE

gov.br
Documento assinado digitalmente
ANDERSON LUIZ ARA SOUZA
Data: 24/02/2025 18:11:17-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Anderson Luiz Ara Souza
Examinador Externo, UFPR

Dedico aos meus pais e a tudo que escapa ao tempo.

AGRADECIMENTOS

Gostaria de registrar minha profunda gratidão aos meus pais, Antonio e Maria, que, sob o sol, me permitiram chegar até aqui na sombra. Agradeço também a todos que fizeram parte da minha trajetória acadêmica e de vida: professores, amigos e demais familiares.

RESUMO

Algoritmos de agrupamento são ferramentas essenciais para explorar estruturas de dados e encontram aplicação em diversas áreas do conhecimento. Entre eles, o agrupamento espectral, baseado na teoria dos grafos, destaca-se por seu desempenho em dados não convexos e tem sido alvo de intensas pesquisas. Com os avanços tecnológicos e a crescente disponibilidade de dados geométricos, a análise estatística de formas surge como uma abordagem promissora para lidar com esse tipo de informação. Este trabalho propõe extensões dos algoritmos espectrais para a análise de formas 2D conforme definido por Kendall. Versões incorporando kernel gaussiano, métodos hierárquicos, k-vizinhos mais próximos e medidas de similaridade e dissimilaridade entre grupos mostraram desempenho encorajador. Os algoritmos foram testados em diversos cenários de simulação no espaço de pré-formas e com dados reais disponíveis na literatura. A avaliação do desempenho dos algoritmos foi realizada através do índice de Rand corrigido, demonstrando potencial para aplicações futuras.

Palavras-chaves: Análise de agrupamento; Estatística de forma; Agrupamento Espectral.

ABSTRACT

Clustering algorithms are essential tools for exploring data structures and have applications in many fields of knowledge. Among them, spectral clustering, based on graph theory, stands out for its performance on non-convex data and has been the focus of extensive research. With technological advancements and the growing availability of geometric data, statistical shape analysis emerges as a promising approach to handle this type of information. This work proposes extensions of spectral algorithms for the analysis of 2D shapes as defined by Kendall. Versions of incorporating Gaussian kernels, hierarchical methods, k-nearest neighbors, and similarity and dissimilarity measures between groups showed encouraging performance. The algorithms were tested in various simulation scenarios within the pre-shape space and on real data available in the literature. Algorithm performance was evaluated using the adjusted Rand index and demonstrates potential for future applications.

Keywords: Cluster analysis; Shape statistics; Spectral clustering.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO E MOTIVAÇÃO

Atualmente, a geração de grandes volumes de dados é um fenômeno em constante expansão, impulsionado pela digitalização de processos, pela conectividade global e pelos avanços tecnológicos. Esse fluxo contínuo de informações representa não apenas um desafio, mas também uma oportunidade: compreender e processar esses dados de forma eficiente possibilita insights valiosos para áreas como negócios, ciência, saúde e governança. A mineração de dados, que integra técnicas de aprendizado de máquina, estatística e análise computacional, tem como objetivo explorar grandes conjuntos de dados em busca de padrões, tendências e informações relevantes. Um dos métodos fundamentais nesse contexto é o agrupamento (JAIN, 2010), que consiste em organizar os dados em grupos com características semelhantes. Essa abordagem é especialmente útil para explorar conjuntos complexos, identificar comportamentos e revelar estruturas subjacentes, contribuindo para análises mais profundas e tomadas de decisão mais eficazes. O interesse por técnicas de agrupamento tem crescido em diversas áreas, sendo aplicado em contextos como psicologia, pesquisa de mercado, reconhecimento de padrões, processamento de imagens, planejamento urbano e biologia (GUPTA; GUPTA; MISHRA, 2011).

Vários tipos de algoritmos de agrupamento foram desenvolvidos na literatura (AGGARWAL; REDDY, 2014). Métodos clássicos de agrupamento como k-means (MACQUEEN, 1967) e suas variantes (SU; CHOU, 2001), (WANG et al., 2014), (JABI, 2019), (XIA, 2020) foram extensivamente explorados e têm um bom desempenho no particionamento, apesar de apresentar limitações quando diante de dados não lineares. A fim de obter um efeito de agrupamento de alta qualidade em amostras de dados de formas arbitrárias, o algoritmo de agrupamento espectral (SHI; MALIK, 2000), (NG; JORDAN; WEISS, 2001), (LUXBURG, 2007) foi proposto.

O agrupamento espectral foi desenvolvido sob a teoria dos grafos, transforma o problema de agrupamento em um problema de particionamento de grafo, tratando cada objeto de dados como um nó de grafo e sua similaridade como pesos de aresta (NG; JORDAN; WEISS, 2001). Em comparação com o k-means, que apresenta bom desempenho na análise de dados com distribuição aproximadamente gaussiana, o agrupamento espectral é particularmente eficaz na detecção de agrupamentos em dados com estruturas não convexas (MONNEY et al., 2020a) e tem sido utilizado em diversas aplicações tais como segmentação de imagens (WANG; DONG, 2012), (ALSHAMMARI; TAKATSUKA, 2019a), detecção de estruturas de comunidade em redes

sociais (DENG et al., 2024), (LAASSEM et al., 2022), (LIERDE; CHOW; CHEN, 2019), (WANG; LIN; WANG, 2017), em bioinformática (YU et al., 2020), (BABICHEV; YASINSKA-DAMRI; LIAKH, 2023), (XIA; GU; ZHANG, 2020), (QI et al., 2021) e em análise de imagens de sensoriamento remoto (TAŞDEMIR, 2013), (TAŞDEMIR; YALÇIN; YILDIRIM, 2015).

Diversos estudos têm se dedicado a aprimorar o desempenho do agrupamento espectral, seja por meio do aperfeiçoamento da matriz de similaridade (MONNEY et al., 2020b), da matriz laplaciana (LI; WEI; ZHAO, 2022) ou do próprio processo de agrupamento (ALSHAMMARI; TAKATSUKA, 2019b), utilizando dados mais representativos. Mais adiante, serão discutidos em maior detalhe alguns trabalhos que abordam essas estratégias de aprimoramento. De modo geral, um algoritmo de agrupamento espectral segue as seguintes etapas: inicialmente, os dados são representados por uma matriz de similaridade, a partir da qual se constrói a matriz laplaciana. Em seguida, realiza-se a decomposição espectral da matriz laplaciana para mapear cada ponto de dado a uma representação de baixa dimensão. Por fim, o conjunto de dados é particionado em duas ou mais classes com base nesses vetores representativos.

Imagens de objetos estão disponíveis em muitas áreas do conhecimento. Avanços na tecnologia tem levado à coletas de rotinas de informações geométricas e o estudo de formas de objeto é cada vez mais importante. A análise de formas é uma área bastante útil e sólida para lidar com estudos de formas de objetos e informação geométrica. Seus métodos têm se mostrado relevantes em vários campos e aplicações tais como biologia, medicina, análise de imagens, arqueologia, geografia, geologia, agricultura, genética, imagens biomédicas, reconhecimento de padrões militares (DRYDEN; MARDIA, 2016).

A forma de um objeto, definida na área de análise estatística de formas, é a informação que permanece quando os efeitos de locação, escala e rotação são removidos através de operações matemáticas, como descrito em (DRYDEN; MARDIA, 2016). Quando são removidos os efeitos de locação e escala, os dados são chamados de dados de pré-formas e são descritos em uma hiperesfera complexa, ou seja, no espaço não Euclidiano. Quando a informação de rotação é removida do espaço de pré-formas obtem-se as formas dos objetos no espaço de formas que também é não Euclidiano. Os primeiros trabalhos na área da análise estatística de formas foram relatados por (KENDALL, 1977); (BOOKSTEIN, 1986); (KENDALL, 1984) que traziam os conceitos fundamentais da área.

Em diversas ocasiões em análise de formas, é necessário agrupar um conjunto de dados em grupos, por exemplo, quando é necessário detectar o número de diferentes espécies em um conjunto, análise de agrupamento pode ser utilizado para detectar os grupos de cada espécie

(AMARAL; DRYDEN; WOOD, 2007). Uma limitação fundamental dos algoritmos de agrupamento é que eles só são projetados para espaços Euclidianos (EVERITT et al., 2011). Assim, o desenvolvimento de algoritmos para o espaço não Euclidiano é necessário quando forem utilizados dados no espaço de pré-formas. Observando a relevância de agrupamentos para analisar formas de objetos, alguns trabalhos têm sido propostos (GEORGESCU, 2009); (AMARAL et al., 2010); (NABIL; GOLALIZADEH, 2016); (VINUÉ; SIMÓ; ALEMANY, 2016); (BRUSE et al., 2017); (HUANG; STYNER; ZHU, 2015); (KLASSEN et al., 2004).

(GEORGESCU, 2009) propõe uma generalização do algoritmo k-means, adaptado para integrar métricas de Procrustes e estimativas completas da forma média, com o objetivo de agrupar objetos com contornos múltiplos ou difusos. (VINUÉ; SIMÓ; ALEMANY, 2016) adaptam o algoritmo clássico k-means para o contexto da Análise de Forma, com foco no caso tridimensional. (NABIL; GOLALIZADEH, 2016) consideraram quatro medidas de distância não Euclidianas combinadas com diferentes métodos de agrupamentos aglomerativos no espaço de forma 2D. Ao aplicar um determinado conjunto de métricas de distância e função de ligação, produzindo agrupamento significativos, (BRUSE et al., 2017) desenvolveram um trabalho que combina segmentação automática, modelagem estatística de formas e agrupamento hierárquico aglomerativo para classificar automaticamente 60 modelos anatômicos em grupos de controles saudáveis e indivíduos com cardiopatias congênitas, incluindo subgrupos definidos por diagnóstico clínico. (KLASSEN et al., 2004) apresenta o uso da geometria diferencial computacional para análise de formas, abordando curvas e espaços de curvas. Especificamente, os autores derivam representações geométricas diferenciais para formas representadas por curvas planas e desenvolvem algoritmos para calcular caminhos geodésicos entre formas arbitrárias nos espaços de forma resultantes.

Considerando que estudos sobre algoritmos baseados em teoria de grafos para agrupar dados de forma é um tema em aberto e buscando acrescer no âmbito da análise de agrupamento, em particular, estudos sobre algoritmos de agrupamento espectral e análise estatística de formas, visto sua vasta aplicabilidade, esse trabalho propõe desenvolver versões de algoritmos de agrupamento para análise de dados de formas baseados em agrupamento espectral. Para validar os métodos, foram realizados experimentos com conjuntos de dados simulados de formas planas e conjuntos de dados reais disponíveis na literatura. A avaliação do desempenho dos algoritmos de agrupamentos para os conjuntos de dados foi medido através do índice de Rand corrigido (HUBERT; ARABIE, 1985).

1.1 OBJETIVOS

As principais contribuições deste estudo estão resumidas a seguir.

1. Explorar a teoria dos grafos e métodos de agrupamento espectral: Apresentar os conceitos fundamentais de teoria dos grafos e destacar pesquisas relevantes que abordam métodos de agrupamento espectral, enfatizando suas aplicações e limitações em diferentes contextos;

2. Desenvolver algoritmos de agrupamento espectral adaptados à análise de dados de formas planas, com ênfase na construção de métodos apropriados para espaços não Euclidianos, dado que os dados no espaço de formas não pertencem ao espaço Euclidiano tradicional;

3. Fornecer avaliações experimentais conduzidas em conjuntos de dados sintéticos e reais com objetivo de comparar o desempenho dos algoritmos propostos em termos de robustez, destacando suas vantagens e limitações.

1.2 ORGANIZAÇÃO DA DISSERTAÇÃO

Além do capítulo introdutório, que apresenta a motivação da pesquisa ao contextualizar a importância e a aplicabilidade do agrupamento espectral e da análise estatística de formas, esta dissertação é composta por mais cinco capítulos. No capítulo 2, abordam-se os principais métodos de agrupamento, incluindo os métodos de partição e hierárquicos, bem como as medidas de distância mais utilizadas na literatura. O capítulo também apresenta uma introdução ao agrupamento espectral e aos seus principais fundamentos matemáticos. Por fim, é realizada uma breve revisão da literatura, destacando estudos relevantes que evidenciam a eficácia do agrupamento espectral e os desafios que ainda motivam pesquisas futuras na área.

No capítulo 3 apresentamos uma revisão geral sobre análise de formas planas fazendo uso das coordenadas de Kendall, aqui definimos o que é uma configuração matemática, espaço de pré-formas, forma média, por fim, definiremos a distância entre formas, importante conceito no cenário de análise de agrupamento. Além disso, introduzimos a definição da distribuição Bingham Complexa, suas propriedades e método de simulação. Essa distribuição será utilizada

para realizar o estudo numérico do trabalho. Finalizamos o capítulos apresentando o algoritmo de agrupamento k-means para formas planas.

No capítulo 4 apresentaremos os algoritmos de agrupamentos propostos: algoritmo espectral clássico (NG; JORDAN; WEISS, 2001), algoritmo espectral hierárquico (LIU L; CHEN, 2013), algoritmo espectral com critério de similaridade e dissimilaridade (WANG et al., 2017), e por fim, o algoritmo espectral baseado em k-NN (LUXBURG, 2007) em suas versões adaptadas para dados de formas planas.

No capítulo 5, apresentamos a avaliação numérica e os resultados da pesquisa, com a descrição dos conjuntos de dados simulados e reais, considerando o espaço de pré-formas utilizado na validação dos métodos propostos. Em seguida, realizamos uma análise detalhada dos resultados obtidos.

Por fim, no capítulo 6, são apresentadas as considerações finais da pesquisa, acompanhadas de sugestões para trabalhos futuros.

2 MÉTODOS DE AGRUPAMENTO

A crescente digitalização e conectividade global têm gerado volumes massivos de dados, o que representa tanto um desafio quanto uma oportunidade. Processar esses dados de forma eficiente pode revelar insights valiosos para diversas áreas, como negócios, ciência e saúde. A mineração de dados, que combina aprendizado de máquina, estatística e análise computacional, busca identificar padrões e estruturas úteis em grandes conjuntos de dados. Um de seus métodos centrais é o agrupamento, que organiza dados com características semelhantes, sendo fundamental para exploração, descoberta de comportamentos e apoio à tomada de decisões.

Agrupar dados semelhantes é um processo de aprendizagem não supervisionado, nesse caso, não há rótulos associados aos dados. Assim, o algoritmo busca identificar padrões ou estruturas inerentes ao conjunto de dados. As técnicas de agrupamento diferem de várias maneiras e podem ser categorizadas em dois grandes grupos: métodos de agrupamento não hierárquico e métodos hierárquico. Os métodos não hierárquicos tem como objetivo encontrar diretamente uma partição de n elementos em k_g grupos. Esse método funciona organizando os dados de modo que cada elemento pertença exatamente a um grupo, com base em critérios de similaridade. O agrupamento hierárquico busca construir uma hierarquia de agrupamentos por meio de dois processos: aglomerativo e divisivo. No método aglomerativo cada observação começa em seu próprio agrupamento, e pares de agrupamentos são fundidos à medida que se sobe na hierarquia. No método divisivo, todas as observações começam em um único grupo, e divisões são realizadas de forma recursiva à medida que se desce na hierarquia.

Os métodos não hierárquicos diferem dos hierárquicos em diversos aspectos. Primeiramente, é necessário pré-definir o número de grupos k_g, ao contrário das técnicas hierárquicas, que não exigem essa informação. Além disso, em cada etapa do processo, os agrupamentos podem ser reorganizados, permitindo a divisão de grupos anteriormente combinados, o que inviabiliza a construção de um dendrograma. Por fim, os métodos não hierárquicos tendem a ser mais eficientes na análise de conjuntos de dados de maior porte.

Uma questão central na análise de agrupamento é o critério utilizado para determinar até que ponto dois elementos de um conjunto de dados podem ser considerados semelhantes. Para isso, é fundamental adotar medidas que quantifiquem essa similaridade, sendo as medidas de distância as mais comuns. Com elas, é possível avaliar o grau de proximidade entre os elementos e, assim, agrupar aqueles que apresentam menor distância entre si. Dessa forma, a escolha

da medida de distância torna-se um fator crucial, pois influencia diretamente a forma como a similaridade é avaliada, impactando a formação dos grupos e o desempenho do algoritmo de agrupamento. A seguir, veremos algumas distâncias que são comumente utilizadas em análise de agrupamento. Sejam os vetores de dados $\mathbf{a} = (\mathbf{a_1}, \ldots, \mathbf{a_n}), \quad \mathbf{b} = (\mathbf{b_1}, \ldots, \mathbf{b_n})$, algumas das medidas de dissimilaridade mais comuns entre os pontos são:

1. **Distância Euclidiana:** seja um espaço euclidiano d-dimensional, a distância euclidiana entre os pontos é definida como

$$d_1(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{j=1}^{n} (\mathbf{a_j} - \mathbf{b_j})^2}.$$

2. **Distância de Manhattan:** baseia-se na soma das diferenças absolutas entre os pontos, a distância de Manhanttan é definida como

$$d_2(\mathbf{a}, \mathbf{b}) = \sum_{j=1}^{n} |\mathbf{a_j} - \mathbf{b_j}|.$$

3. **Distância de Minkowski:** generaliza as distâncias Euclidiana e de Manhattan com um parâmetro p, se $p = 2$ temos a distância Euclidiana e se $p = 1$ temos a distância de Manhattan. A distância de Minkowshi é definida como

$$d_3(\mathbf{a}, \mathbf{b}) = \left(\sum_{j=1}^{n} |\mathbf{a_j} - \mathbf{b_j}|^p \right)^{1/p}.$$

4. **Distância de Mahalanobis:** considera a correlação entre as variáveis, sendo útil para dados com escalas diferentes. A distância de Mahalanobis é definida como segue

$$d_4(\mathbf{a}, \mathbf{b}) = \sqrt{(\mathbf{a} - \mathbf{b})^T \mathbf{S}^{-1} (\mathbf{a} - \mathbf{b})},$$

em que $\mathbf{S}$ é a matriz de covariância entre $\mathbf{a}$ e $\mathbf{b}$.

2.1 MÉTODOS DE PARTICIONAMENTO

Os métodos de partição constituem uma abordagem amplamente utilizada na análise de agrupamento. Dado um conjunto de dados com n objetos, o objetivo desses métodos é dividir os dados em k_g partições disjuntas, onde k_g é um número pré-definido de grupos, satisfazendo

$k_g \leq n$. Inicialmente, o algoritmo gera uma partição dos dados e, em seguida, aplica uma estratégia de reatribuição iterativa, realocando elementos entre os clusters com o intuito de melhorar o valor de um critério de agrupamento específico. Como a busca pela otimização global nesse contexto é um problema conhecido por ser NP-difícil (BARIONI et al., 2014), a maioria dos algoritmos adota abordagens heurísticas. Dentre as mais comuns, destacam-se os métodos baseados em centróide, como o k-means, nos quais cada cluster é representado pela média das posições dos elementos em cada dimensão, e os métodos baseados em elementos representativos, k-medoids, em que cada cluster é representado por seu medoide, ou seja, o elemento mais próximo do centro do grupo.

2.1.1 Método rígido: k-means

O k-means é um dos algoritmos mais populares e amplamente utilizados em análise de agrupamento para particionar dados em k_g grupos. É o mais clássico algoritmo de particionamento do tipo rígido e foi introduzido por (MACQUEEN, 1967). O grupo é representado por seu centróide, que geralmente é a média de pontos dentro de um grupo. A função objetivo usada para k-means é a soma das distâncias quadradas entre um ponto e seu centróide expresso através de uma distância apropriada. A minimização do critério de agrupamento se dará pela formação das partições que apresentem maior similaridade dentro dos grupos e menor similaridade entre os grupos obtidos pelos valores das distâncias dos objetos aos centróides.

Embora o algoritmo de agrupamento k-means seja amplamente utilizado, ele ainda apresenta algumas limitações: se o valor de k_g for inadequado, o algoritmo tende a convergir para um ótimo local ((YUAN; YANG, 2019); (STEMMER, 2021); (QURESHI; AHAMAD, 2018); (YIN et al., 2023)), nesse caso, o ótimo local pode corresponder a uma partição que não reflete bem a estrutura real dos dados, resultando em agrupamentos pouco representativos; o algoritmo é sensível a ruídos e outliers, tornando o resultado do agrupamento ineficientes (GAN; NG, 2017); apresenta resultados ineficientes quando diante de dados não linearmente separáveis, dados sobrepostos ou dados de forma arbitrária não convexo (WANG et al., 2017); (DU et al., 2015); (LUCIŃSKA; WIERZCHOŃ, 2012); (MONNEY et al., 2020a); (YIN et al., 2023).

Devido à sua ampla aplicação em diversas áreas, o algoritmo k-means tem sido objeto de numerosos estudos que visam aperfeiçoar sua formulação original, proposta por (MACQUEEN, 1967). Nesse contexto, (HUANG et al., 2005) introduziram o algoritmo weighted k-means, que incorpora uma etapa adicional ao procedimento clássico para atualizar iterativamente os

pesos das variáveis com base na partição atual, conferindo maior flexibilidade à modelagem da importância dos atributos. Por sua vez, (AMORIM, 2012) propuseram o algoritmo Minkowski k-means, no qual são atribuídos pesos específicos a cada atributo dentro de cada cluster, interpretados como fatores de redimensionamento, de modo a adaptar a métrica de distância à estrutura local dos dados. Ademais, (JOTHI; MOHANTY; OJHA, 2019) desenvolveram uma variante com inicialização determinística, cujo objetivo é eliminar a aleatoriedade na escolha dos centróides iniciais, mitigando, assim, a suscetibilidade do algoritmo a convergência em mínimos locais.

Com o objetivo de mitigar a influência de dados ruidosos no desempenho do algoritmo k-means, (WANG; SU, 2011) propuseram uma versão aprimorada do método, na qual é aplicada uma etapa de pré-processamento para filtrar os dados de entrada e remover instâncias consideradas ruidosas. Essa filtragem impede que tais observações interfiram na determinação dos protótipos iniciais dos clusters. A exclusão prévia dos dados ruidosos contribui para uma melhora significativa na qualidade da partição gerada, reduzindo de forma eficaz o impacto negativo do ruído e resultando em agrupamentos mais precisos e estáveis.

Ao final do capítulo 3, apresentaremos o algoritmo k-means introduzido por (MACQUEEN, 1967) para dados de formas planas utilizando como base o trabalho desenvolvido por (AMARAL et al., 2010).

2.2 MÉTODO HIERÁRQUICO

A concepção de hierarquia consiste em uma ordem de prioridade entre os elementos de um conjunto. Sendo assim, algoritmos de agrupamentos baseados no método hierárquico organizam um conjunto de dados em uma estrutura hierárquica de acordo com a proximidade entre os indivíduos (JR, 1963). Na maioria das vezes são utilizadas em análise exploratória de dados ou com o intuito de identificar possíveis agrupamentos e o valor provável do número de grupos, k_g, no agrupamento não hierárquico. Os algoritmos hierárquicos são divididos em aglomerativos ou divisivos. Algoritmos aglomerativos começam com cada indivíduo no conjunto de dados sendo representado por um único grupo. Os algoritmos divisivos operam na direção oposta ao método aglomerativo. No começo, o conjunto de dados inteiro está em um só grupo e um procedimento que divide os grupos vai ocorrendo sucessivamente até que cada indivíduo seja um único grupo.

Nos algoritmos hierárquicos, em cada estágio do algoritmo, cada novo conglomerado for-

mado é um agrupamento de conglomerados formados nos estágios anteriores, isto é, se dois elementos aparecem juntos num mesmo grupo em algum estágio anterior, eles aparecerão juntos em todos os estágios subsequentes. Devido a isso, os resultados de um algoritmo hierárquico são normalmente mostrados como uma árvore binária ou um dendograma. A raiz do dendograma representa o conjunto de dados inteiro e os nós representam os indivíduos. Os nós intermediários representam a magnitude da proximidade entre os indivíduos. A altura do dendograma expressa a distância entre um par de indivíduos ou entre um par de grupos ou ainda entre um indivíduo e um grupo. O resultado do agrupamento pode ser obtido cortando-se o dendograma em diferentes níveis. Esta forma de representação fornece descrições informativas e visualização para as estruturas de grupos em potencial, especialmente quando há realmente relações hierárquicas nos dados (MURTAGH, 1983).

Para decidir quais agrupamentos devem ser aglomerados ou divididos, é necessária uma medida de dissimilaridade entre conjuntos de observações, obtida por meio de uma medida de distância apropriada entre observações individuais do conjunto de dados, e um critério de ligação, que especifica a dissimilaridade entre conjuntos como uma função das distâncias entre pares de observações nesses conjuntos. A seguir, descreveremos os métodos de agrupamento hierárquico aglomerativos mais comuns.

1. **Método da ligação simples:** Nesse método, a similaridade entre dois conglomerados é definida pelos dois elementos mais parecidos entre si (SIBSON, 1973). Este método também é conhecido como agrupamento dos vizinhos mais próximos, ou seja, é utilizada a menor distância entre os elementos dos grupos. Sejam dois grupos A e B, em que $\mathbf{x}_i \in A$ e $\mathbf{x}_j \in B$. A distância entre dois grupos será definida como:

$$d_{LS}(A, B) = \min\{d(\mathbf{x_i}, \mathbf{x_j}), \quad \mathbf{x_i} \in A, \quad \mathbf{x_j} \in B\}.$$

2. **Método da Ligação Completa:** Nesse método, a similaridade entre dois grupos é definida pelos elementos que são menos semelhantes entre si (DEFAYS, 1977). Sejam dois grupos A e B, em que $\mathbf{x}_i \in A$ e $\mathbf{x}_j \in B$. A distância entre dois grupos será definida como:

$$d_{LC}(A, B) = \max\{d(\mathbf{x_i}, \mathbf{x_j}), \quad \mathbf{x_i} \in A, \quad \mathbf{x_j} \in B\}.$$

3. **Método Ligação Média:** Este método trata a distância entre dois grupos como a média das distâncias entre todos os pares de elementos que podem ser formados com os elementos dois dois grupos que estão sendo comparados. Se o grupo A tem n_1 elementos e o grupo B tem n_2 elementos, a distância entre eles será definida como:

$$d_{LM}(A,B) = \frac{1}{n_1 n_2} \sum_{i \in A} \sum_{j \in B} d(\mathbf{x_i}, \mathbf{x_j}).$$

4. **Método do centróide:** Neste método, a distância entre dois grupos é definida como sendo a distância entre os vetores de médias (centróides) dos grupos que estão sendo comparados. Assim, a distância entre dois grupos A e B é definida como:

$$d_C(A,B) = d(\bar{\mathbf{x}}_A, \bar{\mathbf{x}}_B),$$

em que $\bar{\mathbf{x}}_A$ é a média do grupo A e $\bar{\mathbf{x}}_B$ é a média do grupo B.

5. **Método Ligação Ward:** Este método é conhecido como variância mínima e se fundamenta nos seguintes princípios (JR, 1963): inicialmente, cada elemento é considerado como um único grupo; em cada passo do algoritmo de agrupamento calcula-se a soma de quadrados dentro de cada grupo. Esta soma é o quadrado da distância Euclidiana de cada elemento amostral pertencente ao grupo em relação ao correspondente vetor de médias do grupo, isto é

$$SS_i = \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i.})^T (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i.}),$$

em que n_i é o número de elementos do grupo A_i quando está no passo k_s do processo de agrupamento, $\mathbf{x}_{ij}$ é o vetor de observações do j-ésimo elemento que pertence ao i-ésimo grupo, $\bar{\mathbf{x}}_i$ é o centroide do grupo A_i e SS_i representa a soma de quadrados correspondente ao grupo A_i. No passo k_s, a soma de quadrados total dentro dos grupos é

$$SSR = \sum_{i=1}^{g_{k_s}} SS_i,$$

em que g_{k_s} é o número de grupos existentes quando se está no passo k_s. A distância entre os grupos A_i e A_j é

$$d_W(A_i, A_j) = \frac{n_i n_j}{n_i + n_j}(\bar{\mathbf{x}}_{j\cdot} - \bar{\mathbf{x}}_{i\cdot})^T(\bar{\mathbf{x}}_{j\cdot} - \bar{\mathbf{x}}_{i\cdot}).$$

Em cada passo do algoritmo de agrupamento, os dois grupos que minimizam essa distância são combinados. É possível mostrar que essa distância é a diferença entre o valor de SSR depois e antes de se combinar os grupos A_i e A_j em um único grupo (MINGOTI, 2007). Assim, o método Ward combina os dois grupos que resultam no menor valor de SSR.

2.3 AGRUPAMENTO ESPECTRAL

O agrupamento espectral foi proposto com base na teoria dos grafos (NG; JORDAN; WEISS, 2001). A ideia principal por trás do agrupamento espectral é transformar o problema de particionamento de dados em um problema de corte de grafo. A primeira pesquisa sobre agrupamento espectral remonta a 1973, quando (DONATH; HOFFMAN, 1973) propuseram o uso de autovetores da matriz de Similaridade para particionamento de grafo. No mesmo ano, (FIEDLER, 1973) provou a relação entre conectividade do grafo e o segundo menor autovalor da matriz Laplaciana. (HAGEN; KAHNG, 1992) apresentaram o método Ratio cut (*RadioCut*) em 1992, conectando agrupamento, particionamento de grafo e autovetores de matrizes de Similaridade. (SHI; MALIK, 2000) propuseram o método de corte normalizado (*NCut*) em 2000, considerando conexões entre grupos e intra-grupos para particionamento balanceado de grupos. O corte Min/Max de (DING et al., 2001) e algoritmo NJW proposto por (NG; JORDAN; WEISS, 2001), foram desenvolvidos com base na teoria de grafos.

Nessa seção faremos uma introdução sobre os objetos matemáticos utilizados na estrutura do agrupamento espectral: grafos, matriz de similaridade, matriz laplaciana e métodos de particionamento do grafo. Para maiores detalhes sobre as definições abaixo veja (DING et al., 2024) e (LUXBURG, 2007).

2.3.1 Grafos

Seja $G = (V, E)$ um grafo com conjunto de vértices finito e não vazio $V = \{v_1, \ldots, v_n\}$ e $E \subset V \times V$ é um conjunto de arestas entre os vértices.

Definição: Se $(v_i, v_j) \in E$, essa aresta pode ser representada por e_{ij}. Além disso, se

E é simétrica, isto é, $e_{ij} \in E$ se somente se $e_{ji} \in E$ vale para todas as arestas, então G é chamado de grafo não direcionado.

Nesse momento, não definiremos grafo direcionado já que, para fins de análise de agrupamento, o trabalho se concentrará em grafos não direcionados.

Definição: Seja $G' = (V, E')$ um grafo. Agora substitua E' por $E = E' \times \mathbb{R}_0^+$ de modo que cada aresta $e_{ij} \in E'$ carrega um valor não negativo $w_{ij} \geq 0$. Então, $G = (V, E)$ é chamado de grafo ponderado e w_{ij} é chamado peso da aresta que conecta v_i e v_j.

A partir daqui, procuraremos descrever grafo como matrizes e estudar suas propriedades.

Definição: Seja G um grafo não direcionado e ponderado, o grau do vértice $v_i \in V$ é definido como

$$\mathbf{d}_i = \sum_{j=1}^{n} w_{ij}.$$

Observe que, na verdade, esta soma só percorre todos os vértices adjacentes a v_i, pois para todos os outros vértices v_j o peso w_{ij} é 0. A matriz de graus $\mathbf{D}$ é definida como a matriz diagonal com os graus $\mathbf{d_1}, \ldots, \mathbf{d_n}$ na diagonal.

Além disso, o vértice $\{i | v_i \in A\}$ no subconjunto $A \subset V$ é abreviado como $i \in A$. Sejam $A, B \subset V$ dois conjuntos arbitrários, e o peso da conexão entre A e B é representado por $W(A, B)$

$$W(A, B) = \sum_{i \in A, j \in B} w_{ij}.$$

Consideramos duas maneiras diferentes de medir o tamanho de um subconjunto $A \subset V$:

1. $|A|$ = número de vértices em A;
2. $vol(A) = \sum_{i \in A} \mathbf{d}_i$.

Intuitivamente, $|A|$ mede o tamanho de A pelo seu número de vértices, enquanto $vol(A)$ mede o tamanho de A somando os pesos de todas as arestas anexadas aos vértices em A.

Definição: Um subconjunto $A \subset V$ de um grafo é conectado se quaisquer dois vértices em A puderem ser unidos por um caminho tal que todos os pontos intermediários também estejam em A.

Definição: Seja G um grafo e o subconjunto $A \subset V$. Se existe um caminho entre quaisquer dois pontos em A e o caminho está completamente em A, isto é, todos os nós no

caminho pertencem a A, então A é dito ser conectado. Se A estiver conectado e não houver conexão entre A e seu complemento $\bar{A}$, então A é dito um componente conectado.

Os conjuntos não vazios $A_1, \ldots, A_k$ formam uma partição do grafo se $A_i \cap A_j = \emptyset$ e $A_i \cup \ldots, \cup A_k = V$.

Definição: Seja $A \subset V$ um subconjunto de um grafo. Então o vetor indicador de A é definido como $\mathbb{1}_{\mathbf{A}} = (\mathbf{f_1}, \ldots, \mathbf{f_n}) \in \mathbb{R}^n$ em que

$$\mathbf{f}_i = \begin{cases} 1, & \text{se } v_i \in A \\ 0, & \text{caso contrário.} \end{cases}$$

2.3.2 Matriz de Similaridade do Grafo

Seja um conjunto de dados $\mathbf{x_1}, \ldots, \mathbf{x_n}$ e alguma similaridade $s_{ij} \geq 0$ entre todos os pares de dados $\mathbf{x_i}$ e $\mathbf{x_j}$. Em seguida, construa um grafo $G = (V, E)$ em que os dados correspondem aos vértices do grafo e os valores da similaridade são os pesos de cada aresta, ou seja, $w_{ij} = s_{ij}$. Então, G é chamado de grafo de similaridade dos dados.

Definição: Seja um grafo $G = (V, E)$ não direcionado e ponderado cujo valor do peso é w_{ij}. Se $(v_i, v_j) \notin E$, defina $w_{ij} = 0$. Então, $\mathbf{W} = (w_{ij})$ é chamada matriz de similaridade ou adjacência do grafo. Como G é um grafo não direcionado, existe $w_{ij} = w_{ji}$ para todos os pesos, o que implica que $\mathbf{W}$ é uma matriz simétrica.

Ao construir a matriz de similaridade do grafo o objetivo é modelar o relacionamento da vizinhança local entre os dados. Abaixo é apresentado as principais formas de construção da matriz de similaridade.

1. **Matriz de k-vizinhos mais próximos**: o objetivo é conectar o vértice v_i com o vértice v_j, se v_j estiver entre os k-vizinhos mais próximos de v_i. No entanto, esta definição leva a um grafo direcionado, como a relação de vizinhança não é simétrica. Há duas maneiras de tornar este grafo não direcionado. A primeira maneira é ignorar as direções das arestas, ou seja, conecte v_i e v_j com uma aresta não direcionada se v_i estiver entre os k-vizinhos mais próximos de v_j ou se v_j estiver entre os k-vizinhos mais próximos de v_j. A matriz resultante é o que geralmente é chamado de matriz de k-vizinho mais próximo. A segunda opção é conectar os vértices v_i e v_j se v_i estiver entre os k-vizinhos

mais próximos de v_j; e v_j estiver entre os k-vizinhos mais próximos de v_i. A matriz resultante é chamado de matriz de k-vizinhos mais próximos mútuo.

2. **Matriz de conexão completa:** Aqui conectamos todos pontos com similaridade positiva entre si, e pondera todas as arestas por s_{ij}. Como a matriz deve representar as relações locais de vizinhança, esta construção só é útil se a própria função de similaridade modelar a vizinhança. Um exemplo de tal função de similaridade é a função kernel gaussiana

$$s(\mathbf{x_i}, \mathbf{x_j}) = \exp\left(-\frac{d^2(\mathbf{x_i}, \mathbf{x_j})}{2\sigma^2}\right), \tag{2.1}$$

em que $d^2(\mathbf{x_i}, \mathbf{x_j}) = ||\mathbf{x_i}, \mathbf{x_j}||^2$ mede a dissimilaridade entre duas observações, geralmente, usa-se a distância Euclidiana e o parâmetro σ controla a largura da vizinhança.

2.3.3 Matriz Laplaciana do Grafo

Uma das principais ferramentas para agrupamento espectral são as matrizes laplacianas do grafo. O laplaciano é uma matriz que representa um grafo, é o objeto principal da teoria dos grafos espectrais, existe todo um campo dedicado ao estudo dessas matrizes, para maiores detalhes veja (BUTLER; CHUNG et al., 2006). Nesse momento vamos definir diferentes matrizes laplacianas e apontar suas propriedades mais importantes.

Assumindo que G é um grafo ponderado e não direcionado, com matriz de similaridade $\mathbf{W}$, em que $w_{ij} = w_{ji} \geq 0$, e matriz de graus $\mathbf{D}$.

Definição: A matriz laplaciana não normalizada do grafo é definida como

$$\mathbf{L} = \mathbf{D} - \mathbf{W}.$$

Embora a estrutura do grafo laplaciano seja simples, ele possui muitas propriedades na teoria dos grafos.

Proposição 1: Seja $\mathbf{L}$ uma matriz laplaciana do grafo, então vale as seguintes propriedades:

1. Para cada vetor $\mathbf{f} \in \mathbb{R}^n$, temos

$$\mathbf{f}'\mathbf{L}\mathbf{f} = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(\mathbf{f}_i - \mathbf{f}_j)^2.$$

2. $\mathbf{L}$ é simétrica e positiva semidefinida.

3. O menor autovalor de $\mathbf{L}$ é 0, e o correspondente autovetor é o vetor constante 1.

4. $\mathbf{L}$ tem n autovalores não negativos com valor real $0 = \lambda_1 \leq \ldots \leq \lambda_n$.

Observe que a matriz laplaciana não normalizada não depende dos elementos diagonais da matriz de similaridade $\mathbf{W}$. Cada matriz de similaridade que coincide com $\mathbf{W}$ em todas as posições fora da diagonal leva a mesma matriz laplaciana não normalizada $\mathbf{L}$. Em particular, as arestas próprias em um grafo não altera o grafo laplaciano correspondente. A matriz laplaciana não normalizada e seus autovalores e autovetores podem ser usados para descrever muitas propriedades de grafos, um exemplo que será importante para agrupamento espectral é o seguinte:

Proposição 2: Seja G um grafo não direcionado e ponderado. Então a diversidade geométrica k do autovalor 0 de $\mathbf{L}$ é igual ao número de componentes conectados $A_1, \ldots, A_k$ do grafo. O espaço de características do autovalor 0 é gerado pelos vetores indicadores $\mathbb{1}_{\mathbf{A_1}}, \ldots, \mathbb{1}_{\mathbf{A_k}}$ dos componentes conectados.

A matriz laplaciana que conhecemos até agora é chamado de não normalizado, o que sugere que pode ser normalizada em algum sentido. A seguir apresentaremos duas matrizes que são chamadas de matrizes laplacianas normalizadas na literatura. Ambas as matrizes estão intimamente relacionadas entre si e são definidos como

$$\mathbf{L}_{sym} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2},$$

$$\mathbf{L}_{rw} = \mathbf{D}^{-1}\mathbf{L}.$$

A primeira matriz é chamada por $\mathbf{L_{sym}}$, pois é uma matriz simétrica, e a segunda por $\mathbf{L_{rw}}$, pois está intimamente relacionada a passeio aleatório. A seguir resumimos diversas propriedades de $\mathbf{L_{sym}}$ e $\mathbf{L_{rw}}$, que podem ser encontradas no trabalho de (BUTLER; CHUNG et al., 2006) sobre grafos laplacianos normalizados.

Observe que, para tornar correta a definição dessas matrizes, é necessário atender que $\mathbf{D} = \prod_i \mathbf{d}_i \neq 0$, caso contrário não haverá matriz inversa. Isso é equivalente a dizer que $\mathbf{d}_i \neq 0$ para todo i. Tal como na Proposição 1, algumas propriedades importantes desses dois tipos de matrizes laplaciana serão mostradas abaixo.

Proposição 3: Sejam $\mathbf{L_{sym}}$ e $\mathbf{L_{rw}}$ duas matrizes laplacianas normalizadas definidas acima, respectivamente, temos as seguintes propriedades:

1. Para cada vetor $\mathbf{f} \in \mathbb{R}^n$, temos

$$\mathbf{f}'\mathbf{L_{sym}}\mathbf{f} = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}\left(\frac{\mathbf{f}_i}{\sqrt{\mathbf{d}_i}} - \frac{\mathbf{f}_j}{\sqrt{\mathbf{d}_j}}\right)^2.$$

2. Se λ é um autovalor de $\mathbf{L_{rw}}$ e o autovetor correspondente é $\mathbf{u}$, então λ também é um autovalor de $\mathbf{L_{sym}}$ e o autovetor correspondente é $\mathbf{w} = \mathbf{D}^{1/2}\mathbf{u}$.

3. Se λ é um autovalor de $\mathbf{L_{rw}}$ e o autovetor correspondente é $\mathbf{u}$, então λ e $\mathbf{u}$ podem resolver o seguinte problema de autovalor generalizado: $\mathbf{Lu} = \lambda\mathbf{Du}$.

4. Zero é um autovalor de $\mathbf{L_{rw}}$ e o autovetor correspondente é um vetor $\mathbb{1}$ (contém apenas números 1). 0 é também o autovalor de $\mathbf{L_{sym}}$ e o correspondente autovetor é $\mathbf{D}^{1/2}\mathbb{1}$.

5. $\mathbf{L_{sym}}$ e $\mathbf{L_{rw}}$ são semidefinidas positivas e têm n autovalores reais não negativos $0 = \lambda_1 \leq \ldots \leq \lambda_n$.

Assim como para a matriz laplaciana não normalizada, para a matriz Laplaciana normalizado podemos obter um resultado semelhante ao da Proposição 2.

Proposição 4: Seja G um grafo ponderado não direcionado, então a diversidade geométrica k dos autovalores 0 de $\mathbf{L_{sym}}$ e $\mathbf{L_{rw}}$ é igual ao número dos componentes conectados $A_1, \ldots, A_k$ no grafo. Para $\mathbf{L_{rw}}$, o espaço de recurso de 0 é gerado pelo vetor indicador $\mathbb{1}_{A_i}$ desses componentes conectados; Para $\mathbf{L_{sym}}$, o espaço de recursos de 0 é gerado pelo vetor $\mathbf{D}^{1/2}\mathbb{1}_{A_i}$.

As provas para as proposições de 1 a 4 podem ser encontradas em (LUXBURG, 2007).

2.3.4 Métodos de partição do grafo

A intuição do agrupamento é separar o conjunto de dados em grupos diferentes de acordo com suas semelhanças. Para dados fornecidos em forma de um grafo, este problema pode ser reformulado como um problema de corte do grafo: queremos encontrar uma partição do grafo tal que as arestas entre os diferentes grupos têm um peso muito baixo, o que significa que os pontos em diferentes grupos são diferentes entre si, e as arestas dentro de um grupo têm alta

peso, o que significa que os pontos dentro do mesmo grupo são semelhantes entre si. Veremos, a partir daqui, como o agrupamento espectral pode ser derivado como uma aproximação para tal problema de particionamento de grafos.

O particionamento de grafos é um problema clássico na teoria dos grafos, há muitos métodos tradicionais tais como: Método do Corte Mínimo (HOFMEYR, 2016), Método de Ajuste de Corte Proporcional (NICA, 2020), Método de Ajuste de Corte Normal (YANG et al., 2022), Método de Corte min-max (NIE et al., 2010). Para uma discussão sobre vantagens e desvantagens de cada um desses métodos ver (DING et al., 2024).

Dado um grafo com a matriz de Similaridade $\mathbf{W}$, a maneira mais simples e direta de construir uma partição do grafo é resolver o problema $mincut$, isto é, minimizar uma função objetivo. Para um determinado número k_g de subconjuntos, a abordagem mincut consiste simplesmente em escolher uma partição $A_1, \ldots, A_{k_g}$ que minimiza

$$cut(A_1, \ldots, A_{k_g}) = \frac{1}{2} \sum_{i=1}^{k_g} cut(A_i, \bar{A}_i).$$

Aqui introduzimos o fator $1/2$ para consistência notacional, caso contrário, contaríamos cada aresta duas vezes no corte. Em particular para $k_g = 2$, $mincut$ é um problema relativamente fácil e pode ser resolvido de forma eficiente, ver (STOER; WAGNER, 1997) e a discussão nele contida. Contudo, na prática muitas vezes não leva a partições satisfatórias. O problema é que em muitos casos, a solução de $mincut$ simplesmente separa um vértice individual do resto do grafo. Isso não é o que queremos alcançar em agrupamento, já que os grupos devem ser razoavelmente grandes. Uma maneira de contornar este problema é solicitar explicitamente que os conjuntos $A_1, \ldots, A_{k_g}$ sejam razoavelmente grandes.

As duas funções objetivo mais comuns são $RatioCut$ (HAGEN; KAHNG, 1992) e o corte normalizado $NCut$ (SHI; MALIK, 2000). No $RatioCut$, o tamanho de um subconjunto A de um grafo é medido pelo seu número de vértices $|A|$, enquanto em $NCut$ o tamanho é medido pelos pesos de suas arestas $vol(A)$. As definições são:

$$RadioCut(A_1, \ldots, A_{k_g}) = \sum_{i=1}^{k_g} \frac{cut(A_i, \bar{A}_i)}{|A_i|}, \tag{2.2}$$

$$NCut(A_1, \ldots, A_{k_g}) = \sum_{i=1}^{k_g} \frac{cut(A_i, \bar{A}_i)}{vol(A_i)}. \tag{2.3}$$

Observe que ambas as funções objetivo assumem um valor pequeno se os grupos A_i não são muito pequenos. Em particular, o mínimo da função $\sum_{i=1}^{k_g}(1/|A_i|)$ é alcançado se todos

$|A_i|$ coincidem, e o mínimo de $\sum_{i=1}^{k_g}(1/vol(A_i))$ é alcançado se todos $vol(A_i)$ coincidirem. Portanto, o que ambas as funções objetivo tentam alcançar é que os grupos sejam equilibrados, conforme medido pelo número de vértices ou pesos das arestas, respectivamente. Infelizmente, a introdução de condições de equilíbrio faz com que o problema do corte mínimo, anteriormente simples de resolver, se torne NP-difícil, ver (WAGNER; WAGNER, 1993) para uma discussão. O agrupamento espectral é uma maneira para resolver versões relaxadas desses problemas, relaxar o $NCut$ leva ao agrupamento espectral normalizado, enquanto relaxar o $RatioCut$ leva ao agrupamento espectral não normalizado.

Embora a solução de muitas funções objetivo de agrupamento de grafos seja NP-difícil, por meio de derivação matemática, se forem escritas na forma de entropia de Rayleigh, a solução de relaxação da função objetivo de particionamento do grafo pode ser obtida pelas propriedades de entropia de Rayleigh (MOHAR, 1997), isso porque o valor mínimo, o segundo valor mínimo, ..., e o valor máximo da entropia de Rayleigh corresponde ao autovalor mínimo, o segundo autovalor mínimo, ..., e o autovalor máximo da matriz Laplaciana, respectivamente, e o valor extremo é obtido no autovetor correspondente. Pode ser visto que o autovetor da matriz laplaciana do grafo contém as informações de categoria dos vértices. Portanto, o problema de otimização da divisão de grafos pode ser convertido em um problema de otimização numérica calculando os autovalores e autovetores da matriz laplaciana, de modo que possa ser resolvido em um curto espaço de tempo (EMIROV et al., 2022). O agrupamento espectral é uma maneira para resolver versões relaxadas desses problemas, a derivação do agrupamento espectral normalizado como relaxamento de minimizar $Ncut$ pode ser encontrada em (LUXBURG, 2007) e (SHI; MALIK, 2000).

Geralmente, qualquer agrupamento espectral consiste em três partes: pré-processamento, representação espectral e agrupamento. Primeiro, os dados são representados como a matriz de Similaridade do grafo, e a correspondente matriz Laplaciana é construída. Então, a auto-composição da matriz Laplaciana é usada para mapear cada ponto de dados para um valor representativo de baixa dimensão. Finalmente, o conjunto de dados é dividido em duas ou mais classes baseadas nos novos pontos representativos. Para maiores detalhes sobre agrupamento espectral ver (LUXBURG, 2007) e suas referências, que foi base para construção teórica apresentada até aqui.

2.3.5 Revisão da Literatura

Na seção anterior, foi apresentada uma breve formalização matemática introdutória sobre a estrutura do agrupamento espectral, envolvendo grafos, matriz de similaridade, matriz laplaciana e métodos de particionamento de grafos. O termo agrupamento espectral refere-se, na verdade, a uma abordagem que engloba uma família de algoritmos de agrupamento baseados em propriedades espectrais. Como um dos objetivos deste trabalho é propor algoritmos de agrupamento espectral adaptados a dados de formas planas, tornou-se necessário revisar propostas já existentes, com o intuito de selecionar aquelas que representem as principais linhas de pesquisa da área. Muitos estudos concentram-se em aprimorar distintos componentes do método — seja a matriz de similaridade, a matriz laplaciana ou o processo de agrupamento em si. Nesta seção, destacamos algumas aplicações relevantes e os principais desafios que limitam a adoção do agrupamento espectral em áreas multidisciplinares de grande impacto, como mineração de texto, segmentação de imagens, detecção de comunidades em redes sociais e bioinformática.

Em mineração de texto, o agrupamento espectral pode ser usado para agrupar textos e descobrir tópicos ou estruturas semânticas latentes em coleções de textos. Isso tem aplicações práticas em tarefas como recuperação de informações, classificação de documentos e modelagem de tópicos. Alguns estudos recentes foram desenvolvidos nesse âmbito (JANANI; VIJAYARANI, 2019), (MENON; ASHOK; ARYA, 2022), (ZHANG; LI; WANG, 2017), (ROY; BASU, 2022). A pesquisa sobre agrupamento espectral para mineração de texto não é abrangente, e há ainda muito espaço para melhorias, por exemplo, quanto a ruídos e valores discrepantes em dados de texto.

O agrupamento espectral é amplamente utilizado na segmentação de imagens, especialmente quando há texturas e estruturas complexas na imagem (ZHANG et al., 2021), (CHEN; GUO, 2017), (ZHANG et al., 2021), (CHALLA et al., 2020). Ainda assim, nesse campo, esse tipo de agrupamento apresenta algumas deficiências como: alta complexidade computacional, especialmente quando se trata de dados de imagem em grande escala, além de ser sensível a ruídos e às mudanças locais na imagem. (ZHANG et al., 2021) propuseram um método de agrupamento de subamostras baseado em agrupamento espectral linear acelerado, que usa recursos de imagem e distância de Manhattan para melhorar a distância métrica, com o objetivo de melhorar a precisão da segmentação de superpixel de imagens não convexas. (CHALLA et al., 2020) propõe um algoritmo que melhora a eficiência computando uma solução aproximada

para agrupamento espectral.

Em redes sociais, o agrupamento espectral pode ser usado para detectar a estrutura da comunidade, encontrando subgrupos estreitamente conectados na rede, isso é importante para entender as relações sociais, descobrir grupos sociais potenciais e prevendo a disseminação de informações. (DENG et al., 2024) propuseram um algoritmo de agrupamento espectral baseado em amostragem para redes sociais de grande escala. (WANG; LIN; WANG, 2017) desenvolveram um algoritmo de agrupamento espectral ponderado para detectar a estrutura da comunidade da rede sem saber o número de grupos com antecedência. (LAASSEM et al., 2022) usa o espectro do laplaciano normalizado com base na matriz de Coulomb; primeiro incorpora os vértices do grafo em um vetor no espaço de baixa dimensão e, em seguida, executa o agrupamento k-means nos vértices projetados. Este método é usado para construir uma boa matriz de similaridade em detecção de comunidades. (LIERDE; CHOW; CHEN, 2019) propôs o agrupamento espectral estendido para endereçar comunidades sobrepostas e grandes redes na detecção de comunidades. Além disso, o agrupamento espectral na detecção de comunidades também enfrenta problemas de alta dimensionalidade e sensibilidade ao ruído, complexidade de construção de grafos e falta de adaptabilidade a redes dinâmicas.

No vasto campo da bioinformática, o agrupamento espectral tem sido utilizado para identificar padrões em expressão gênica e conjuntos de genes relacionados. Estudos desenvolvidos por (YU et al., 2020), (BABICHEV; YASINSKA-DAMRI; LIAKH, 2023), (XIA; GU; ZHANG, 2020), (QI et al., 2021) demonstram como o agrupamento espectral é uma ferramenta promissora nesse campo de pesquisa. A fim de integrar vários conjuntos de dados genômicos para identificar estruturas e reduzir o ruído, (JOHN et al., 2020) propuseram um método de agrupamento espectral para dados ôhmicos complexos. Um algoritmo de agrupamento espectral baseado em redes neurais profundas para lidar com a alta dimensionalidade na Eletroencefalografia foi desenvolvido por (CHAKLADAR; SAMANTA; ROY, 2022). No entanto, (DING et al., 2024) pontuam que alguns problemas relacionados à bioinformática precisam ser considerados: os dados de bioinformática são frequentemente heterogêneos, incluindo múltiplos tipos de entidades biológicas e diferentes condições experimentais. O agrupamento espectral lida bem com estruturas latentes, mas pode exigir modelagem adicional em casos de heterogeneidade. Na bioinformática, a interpretação biológica dos grupos formados é fundamental e depende de conhecimento especializado na área.

Diversos estudos têm se dedicado a aprimorar o desempenho do agrupamento espectral, seja por meio da melhoria da matriz de similaridade, da matriz laplaciana ou do próprio pro-

cesso de agrupamento, utilizando dados mais representativos. De modo geral, um algoritmo de agrupamento espectral segue três etapas principais: primeiramente, os dados são representados por uma matriz de similaridade, a partir da qual se constrói a matriz laplaciana correspondente; em seguida, realiza-se a decomposição espectral dessa matriz, projetando cada ponto de dado em um espaço de menor dimensão; por fim, os dados são agrupados com base nessas novas representações. Na prática, o desempenho do agrupamento espectral é impactado principalmente por dois fatores: a escalabilidade do algoritmo para grandes volumes de dados e a precisão na formação dos agrupamentos. O método mais básico requer tempo da ordem de $O(n^2)$ para a construção do grafo e das matrizes envolvidas, e $O(n^3)$ para o cálculo dos autovalores e autovetores da matriz laplaciana, o que limita sua aplicação direta em conjuntos de dados de grande escala (DING et al., 2024).

Uma etapa fundamental do agrupamento espectral é a escolha de uma medida de distância adequada para capturar a estrutura intrínseca dos dados. Para um bom desempenho, pontos pertencentes à mesma classe devem apresentar alta similaridade e manter coerência espacial. Nesse contexto, a construção da matriz de similaridade torna-se um elemento central, influenciando diretamente a eficácia do agrupamento espectral (DING et al., 2024). A função kernel gaussiano, frequentemente utilizada para medir similaridade entre pontos, apresenta uma limitação importante: sua falta de adaptabilidade, devido ao parâmetro de escala fixo (σ). Nessa abordagem, a similaridade entre dois pontos é determinada unicamente pela distância euclidiana, sem considerar a densidade ou o contexto local dos dados. Essa limitação pode comprometer a precisão do agrupamento, resultando em particionamentos subótimos (NIE; WANG; HUANG, 2014).

(ZELNIK-MANOR; PERONA, 2004) propuseram o uso de um parâmetro que permite o autoajuste das distâncias entre pontos com base nas estatísticas locais das vizinhanças, por meio dos k-vizinhos mais próximos. Inspirados por esse trabalho, diversos autores propuseram variantes do agrupamento espectral que buscam determinar parâmetros de escala local na função kernel gaussiano, como em (LI et al., 2007), (YE; SAKURAI, 2015) e (CAO et al., 2013). Nesse contexto, Zhang et al. (ZHANG; LI; YU, 2011) introduziram um método adaptativo baseado na densidade local para medir similaridades, aprimorando a função kernel gaussiano ao considerar a densidade entre os pontos de dados, o que resulta em maior similaridade intra-classe. Além disso, os mesmos autores (ZHANG; YOU, 2011) propuseram uma abordagem baseada em passeio aleatório para processar a matriz de similaridade derivada do kernel gaussiano.

É possível construir a matriz de similaridade por meio do método k-NN. É um método con-

siderado simples, mas que apresenta algumas desvantagens. Estudos utilizando k-NN mútuo e suas versões modificadas foram explorados por (LUCIŃSKA; WIERZCHOŃ, 2012); (VEENSTRA; COOPER; PHELPS, 2016); (TAN; ZHANG; WU, 2020); (KIM et al., 2021); (YIN et al., 2023); (ALSHAMMARI; STAVRAKAKIS; TAKATSUKA, 2021).

Trabalhos voltados para melhorar a eficiência na construção da matriz de similaridade no espaço do kernel ou utilizando vários kernels foram desenvolvidos por (YE; SAKURAI, 2018), (WANG et al., 2019), (JI et al., 2019), (MONNEY et al., 2020b), (KANG et al., 2019). No estudo realizado por (NATALIANI; YANG, 2019), os autores ajustaram a similaridade ao estimarem a largura de banda do kernel gaussiano com base no valor máximo do vizinho mais próximo; apesar de melhorar a estabilidade dos resultados, o método baseia-se em matrizes totalmente conectadas, exigindo recursos computacionais substanciais. (WANG et al., 2011) propõem um modelo de agrupamento espectral que constrói uma matriz de similaridade adequada, incorporando informações geométricas locais para garantir baixa similaridade entre pontos de diferentes classes. A matriz de similaridade é facilmente corrompida por pontos de ruído, levando a um agrupamento incorreto. Algumas pesquisas propuseram versões do algoritmo espectral robusto a ruídos (WANG; DING; JIA, 2019), (TAO et al., 2019), (HESS et al., 2019), (ZHU et al., 2018b), (ZHU et al., 2018a), (KIM et al., 2021), (TAN; ZHANG; WU, 2020).

O agrupamento espectral sob informação supervisionada é geralmente uma extensão do agrupamento espectral não supervisionado tradicional, melhorando o efeito do agrupamento ao incorporar informações supervisionadas adicionais. Essas informações supervisionadas são adicionadas à matriz de similaridade para aumentar a conectividade intra-cluster e a dissimilaridade entre clusters. Trabalhos sob essa perspectiva foram realizados por (YE et al., 2017); (PENG; ZHANG; YI, 2013); (BAI; QI; LIANG, 2023). No estudo desenvolvido por (ZHANG et al., 2014) é proposto um agrupamento espectral baseado em regressão de mínimos quadrados.

Após a construção da matriz de similaridade, o próximo passo é estabelecer a matriz laplaciana correspondente de acordo com diferentes métodos de divisão de grafos. A seleção do método de partição de grafos e o estabelecimento da matriz laplaciana terá um impacto importante nos resultados do agrupamento (SAADE; KRZAKALA; ZDEBOROVÁ, 2014). O trabalho de (LUO et al., 2010) melhora o agrupamento espectral com base no operador ρ-laplaciano. Eles usaram um método eficaz de otimização de gradiente descendente para obter o espaço de incorporação global do operador ρ-laplaciano, este espaço contém as informações de partição multiclasses do conjunto de dados. Com o objetivo de combinar as informações dos atributos e relacionamentos em formação nos dados, (RENGASAMY; MURUGESAN, 2022) melhorou o

algoritmo de (WANG; DING; LI, 2009) alterando a matriz diagonal para medição de uma matriz laplaciana mais eficaz. Para explorar as relações internas dos subclusters, (LI; WEI; ZHAO, 2022) impôs restrições à classificação da matriz laplaciana para garantir que o número de componentes conectados obtidos da matriz de similaridade seja igual ao número de subclusters fornecidos.

Logo após a autocomposição da matriz laplaciana, a seleção de autovetores é crucial. Nem todo autovetor contém informações úteis para agrupamento (TRILLOS et al., 2014). O algoritmo de (NG; JORDAN; WEISS, 2001) usa os maiores k_g autovetores para particionar os dados. Um autovalor maior significa que o autovetor correspondente contém mais informações, o que significa que tem uma capacidade mais forte de distinguir dados. No agrupamento espectral, os primeiros k_g maiores autovetores são usados para representar os dados originais porque possuem maior discriminação em comparação com outros autovetores, que podem alcançar melhor resultados no agrupamento. Enquanto isso, esses autovetores também são os mais distintos recursos extraídos da estrutura do grafo, logo, a seleção dos autovetores é um fator importante para garantir a eficácia do agrupamento (DING et al., 2024). Mas para alguns problemas de agrupamento, selecionar os maiores k_g autovetores não garante detectar a estrutura de dados real. (REBAGLIATI; VERRI, 2011) descobriram que para obter a divisão ótima de k_g grupos usando a matriz laplaciana, a diferença entre o k_g-ésimo e (k_g+1)-ésimos autovalores devem ser suficientemente grande. Eles propuseram que a partição ótima de k_g grupos pode ser obtida a partir do subespaço k_g-dimensional dos autovetores $m_g(m_g > k_g)$ anteriores, com m_g sendo um parâmetro definido pelo usuário. (ALSHAMMARI; TAKATSUKA, 2019b) propuseram um agrupamento espectral aproximado para resolver o problema de seleção de autovetores da matriz laplaciana. O algoritmo usa o índice Davies-Bouldin para estimar a capacidade do autovetor de particionar o grafo, com os menores autovalores tendo maior probabilidade de seleção.

(VORA; RAMAN, 2018) abordou a localização de objetos na aprendizagem não supervisionada usando agrupamento espectral iterativo. O algoritmo calcula o operador laplaciano regularizado com base na matriz de similaridade e seleciona o segundo menor autovetor para classificação binária. Múltiplas iterações de agrupamento espectral são realizados para obter resultados altamente localizados. (KANAAN-IZQUIERDO; ZIYATDINOV; PERERA-LLUNA, 2018) usaram a análise de Componentes Principais Comuns (CPC) para calcular o autovetor comum da entrada matriz Laplaciana de múltiplas visualizações. (KADAVANKANDY; COUILLET, 2019) analisaram os atributos de distribuição dos autovetores para agrupamento espectral de grafos

aleatórios e agrupamento espectral de kernel de vetores aleatórios gaussianos de alta dimensão. Em grafos aleatórios, o comportamento do autovetor principal isolado torna-se semelhante ao dos vetores aleatórios gaussianos.

No agrupamento espectral, o número de agrupamentos geralmente precisa ser definido previamente pelo usuário, sendo assim, é um parâmetro de entrada muito sensível. Como estimar com precisão o número de grupos é um dos principais desafios no agrupamento espectral (ALSHAMMARI; TAKATSUKA, 2019b). Depois de obter uma nova representação dos dados com os autovetores, k-means é usado para obter resultados de agrupamento diretamente. A maioria dos métodos existentes determina o número ideal de grupos com base no princípio de minimizar a similaridade total entre grupos ou maximizar a similaridade total dentro dos grupos. (TEPPER et al., 2011) propuseram um método de agrupamento orientado pela percepção que considera apenas a distância entre os pontos de dados. Determina automaticamente o número de grupos definindo um limite de detecção, independente da dimensão dos dados originais e forma de classe, evitando o desastre dimensional. (FANG; WANG, 2012) desenvolveram um esquema baseado em bootstrap para avaliar a instabilidade de agrupamento, selecionando o número de classes para minimizar instabilidade. (ZHANG; YE; SAKURAI, 2022) determinam de forma adaptativa o número ideal de subclusters combinando grafos compartilhados do vizinho mais próximo com aprendizagem em conjunto e agrupamento espectral. (LAHMAR et al., 2020) empregou um algoritmo de aprendizagem em conjunto, que inclui seleção de recursos, agrupamento espectral aprimorado e um novo índice de validade de agrupamento difuso para avaliar o número preciso de subclusters selecionados.

Apesar de funcionar muito bem em tarefas complexas, o agrupamento espectral apresenta algumas limitações. Seu desempenho é sensível à matriz de similaridade de dados e sua alta complexidade computacional limita suas aplicações em problemas de larga escala. Para reduzir a complexidade computacional do agrupamento espectral, foram propostos algumas abordagens de aceleração, por exemplo, com base na equivalência entre agrupamento espectral e kernel k-means (DHILLON; GUAN; KULIS, 2007); (LIU, 2017). Algoritmos de agrupamento espectral restrito também foram desenvolvidos em (WANG; QIAN; DAVIDSON, 2014); (LIU; TAO; FU, 2017); (CAI; CHEN, 2014); (WANG et al., 2023). Atualmente, uma série de métodos aprimorados de agrupamento espectral foram desenvolvidos para aumentar a eficácia e eficiência do agrupamento espectral, alguns estudos recentes são (ALSHAMMARI; STAVRAKAKIS; TAKATSUKA, 2021); (WANG et al., 2023); (ZHONG; PUN, 2023) (YU et al., 2024); (YIN et al., 2023).

3 ANÁLISE DE FORMAS

Dados geométricos de objetos são coletados rotineiramente em diversas situações do cotidiano, desempenhando um papel essencial em várias áreas do conhecimento, como diagnóstico de doenças, reconhecimento facial e identificação de proteínas. A análise estatística de formas, um campo relativamente recente, dedica-se ao estudo das características geométricas de objetos, com o objetivo de compará-las e descrevê-las no espaço. O desenvolvimento de métodos para analisar dados relacionados à forma tem se tornado cada vez mais relevante, com aplicações amplas em genética, medicina, análise de imagens, arqueologia, bioinformática, geologia e geografia, conforme discutido por (DRYDEN; MARDIA, 2016).

O primeiro passo para se obter a forma de um objeto, é adicionar pontos no contorno da imagem, chamados marcos anatômicos (DRYDEN; MARDIA, 2016). Quando, através de transformações matemáticas adequadas, são removidos os efeitos de escala e locação a partir das coordenadas de um objeto no espaço dos marcos anatômicos, um novo conjunto de coordenadas de um objeto pode ser obtido. Este conjunto é chamado de coordenadas de pré-formas e o novo sistema de coordenadas recebe o nome de espaço das pré-formas. A forma é, finalmente, obtida removendo a informação de rotação das coordenadas pré-formas do objeto. A informação de rotação é eliminada rotacionando um objeto para que ele fique tão próximo quanto possível de um molde. O novo conjunto de coordenadas do objeto está dentro de um novo espaço, que é chamado espaço de formas. Assim, podemos definir o que é a forma de um objeto. Segundo (KENDALL, 1977): Forma é toda a informação geométrica que permanece quando efeitos de escala, localização e rotação são removidos de um objeto.

Na sequência, serão detalhadas abordagens específicas para representar formas de objetos. Apresentaremos o que é uma configuração matemática, o que é uma pré-forma, forma média e definiremos um importante conceito no âmbito na análise de agrupamento: a distância entre formas. Essas definições nos permitiram lidar com dados geométricos de maneira eficaz. As definições abaixo foram extraídas do livro de (DRYDEN; MARDIA, 2016).

3.1 REPRESENTAÇÃO MATEMÁTICA DAS FORMAS

A fim de descrever uma forma, inicialmente, localizaremos um número finito de pontos em um objeto que são chamados de marcos ou pontos de referência. Na literatura existem vários

sinônimos para pontos de referência, incluindo vértices, pontos de controle, locais, pontos-chave, nós, pontos de modelo, marcadores e assim por diante. O marco é a principal fonte de dados para a descrição dos formas. A posição dos pontos estão associadas às coordenadas cartesianas do objeto.

Definição: Marcos ou Pontos de referência são pontos de correspondência identificados em cada objeto, utilizados para estabelecer relações tanto entre os objetos quanto dentro das populações.

Existem três tipos básicos de marcos em nossas aplicações: matemáticos, pseudo-marcos e científicos ou anatômicos.

Definição: Marcos matemáticos são pontos definidos em um objeto com base em suas propriedades matemáticas ou geométricas específicas.

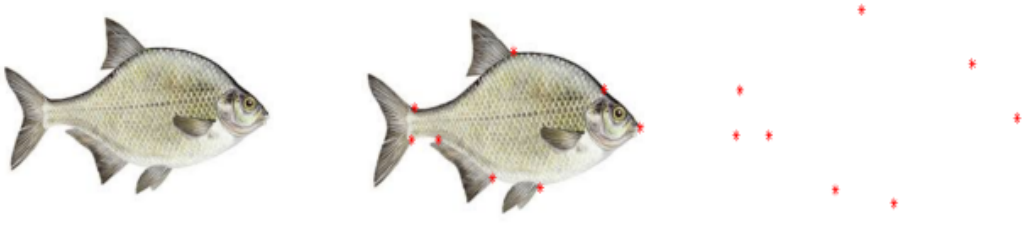
Definição: Pseudo-marcos de referência são pontos construídos em um objeto, posicionados ao longo do contorno ou situados entre marcos científicos ou matemáticos.

A utilização de pseudo-marcos é mais para conceituar a forma do objeto, são úteis no campo biológico, na correspondência de superfícies, quando os pontos podem ser localizados em uma grade sobre cada superfície, por exemplo, o cortical, superfície do cérebro ou a superfície do hipocampo.

Definição: Um marco científico ou anatômico é um ponto definido por um especialista que apresenta relevância científica ou biológica significativa em um contexto específico.

Podemos citar, por exemplo, o canto do olho ou o encontro de duas suturas em um crânio como marco anatômico. (OLIVEIRA, 2016) elabora uma imagem interessante para melhor entendimento sobre forma e marcos anatômicos expressa na Figura 1. Pode-se notar a representação do contorno de peixes, os marcos anatômicos e a configuração da forma.

Figura 1 – Representação de formas e marcos anatômicos.



(a) Contorno e forma (b) Marcos anatômicos (c) 8 marcos da configuração da forma

Fonte: (OLIVEIRA, 2016).

A fim de descrever a forma de um objeto é necessário especificar um sistema de coordenadas. Existem vários sistemas de coordenadas, nesse trabalho, vamos considerar o sistema de coordenadas de (KENDALL, 1977). Em (KENDALL, 1984) foi formalizado e definido os conceitos básicos para a análise de formas. Um dos pontos mais importantes do trabalho de Kendall foi a proposta dos sistemas de coordenadas, que possuem finalidade de obter a forma de um objeto.

Quando digitalizado os marcos, os objetos geralmente possuem tamanhos, posições e rotações diferentes, dentro do equipamento de medição. Por isso, é necessário retirar os efeitos de locação, escala e rotação do conjunto de dados em ordem para trazê-los para um tamanho, orientação e posição padronizados antes de uma análise mais aprofundada. Para isso, é preciso definir uma configuração dos pontos de referência e um espaço de configuração.

Definição: A configuração é o conjunto de pontos de referência de um determinado objeto. A matriz de configuração $\mathbf{X}$ é a matriz $k \times m$ de coordenadas cartesianas de k pontos de referência em m dimensão. O espaço de configuração é o espaço de todos as coordenadas dos marcos.

A matriz de configuração $\mathbf{X}$, também chamada de configuração matemática é representada da seguinte forma

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & \dots & x_{1,m} \\ x_{2,1} & \dots & x_{2,m} \\ \vdots & \ddots & \vdots \\ x_{k,1} & \dots & x_{k,m} \end{pmatrix}$$

.

Nesta dissertação serão considerados os casos onde $k \geq 3$ e $m = 2$, o que corresponde os espaço dos marcos anatômicos no $\mathbb{R}^2$. Assim, a matriz de configuração $\mathbf{X}$ resume-se ao caso com duas colunas.

3.2 ESPAÇO DE PRÉ-FORMAS

Queremos obter a forma de um objeto, então, algumas transformações matemáticas devem ser feitas na matriz $\mathbf{X}$ para remover os efeitos de locação, escala e rotação. Para $m = 2$, a configuração matemática deve ser reescrita como um vetor complexo. Defina um vetor

complexo $(k \times 1)$ tal que

$$\mathbf{z^0} = (\mathbf{z_1^0}, \ldots, \mathbf{z_k^0})^T = (x_{1,1} + ix_{1,2}, \ldots, x_{k,1} + ix_{k,2})^T, \tag{3.1}$$

o qual corresponde as coordenadas complexas para os marcos.

Para remover a locação do vetor complexo devemos definir a sub-matriz de Helmert $(\mathbf{H})$. A matriz de Helmert $(\mathbf{H}^F)$ é uma matriz quadrática ortogonal $k \times k$ com a primeira linha de elementos igual a $1/\sqrt{k}$, então a j-ésima linha possui $(j-1)$ elementos iguais a $-1/\sqrt{j(j-1)}$ seguido por um elemento igual a $(j-1)\times 1/\sqrt{j(j-1)}$ e $(k-j)$ zeros. A sub-matriz de Helmert é a matriz de Helmert sem a primeira linha. Por exemplo, para $k = 4$ a matriz de Helmert é

$$\mathbf{H}^F = \begin{pmatrix} 1/2 & 1/2 & 1/2 & 1/2 \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 & 0 \\ -1/\sqrt{6} & -1/\sqrt{6} & 2/\sqrt{6} & 0 \\ -1/\sqrt{12} & -1/\sqrt{12} & -1/\sqrt{12} & 3/\sqrt{12} \end{pmatrix}$$

e a sub-matriz de Helmert é

$$\mathbf{H} = \begin{pmatrix} -1/\sqrt{2} & 1/\sqrt{2} & 0 & 0 \\ -1/\sqrt{6} & -1/\sqrt{6} & 2/\sqrt{6} & 0 \\ -1/\sqrt{12} & -1/\sqrt{12} & -1/\sqrt{12} & 3/\sqrt{12} \end{pmatrix}$$

Para remover a locação do vetor complexo $\mathbf{z^0}$, basta multiplicar o vetor pela sub-matriz de Helmert $(\mathbf{H})$ de dimensão $(k-1) \times k$. A configuração Hermertizada é dada por

$$\mathbf{w}_{(k-1\times 1)} = \mathbf{H}_{(k-1\times k)}\mathbf{z^0}_{k\times 1}, \tag{3.2}$$

onde $\mathbf{w}$ representa a configuração $\mathbf{z^0}$ sem o efeito de locação.

Dessa forma, pré-multiplicando o vetor $\mathbf{w}$ por $\mathbf{H}^T$ obtém-se a configuração centrada

$$\mathbf{H}^T\mathbf{w} = \mathbf{H}^T\mathbf{H}\mathbf{z^0} = (I_k - \frac{1}{k}1_k 1_k^T)\mathbf{z^0} = \mathbf{z^0} - \frac{1}{k}\sum_{j=1}^{k}\mathbf{z^0_{(j)}}1_k.$$

Para exemplificar, (DRYDEN; MARDIA, 2016) apresenta a configuração matemática de um indivíduo obtido dos dados de gorilas machos, em que $\mathbf{z}^0 = (53 + 220i, 46 - 35i, 0 + 0i, 0 + 37i, 12 + 122i, 58 + 204i, 93 + 117i, 103 + 28i)^T$. Vale notar que a configuração Helmertizada perde a dimensão original dos dados. Este problema é corrigido com a multiplicação por $\mathbf{H^T}$.

Para remover o efeito de escala deve-se dividir a configuração Helmertizada, obtida na expressão (3.2) pela sua norma:

$$\mathbf{z}_{(k-1\times 1)} = \frac{\mathbf{w}}{|\mathbf{w}|} = \frac{\mathbf{w}}{\sqrt{\mathbf{w}^*\mathbf{w}}} = \frac{\mathbf{H}_{(k-1\times k)}\mathbf{z^0}_{(k\times 1)}}{\sqrt{(\mathbf{H}_{(k-1\times k)}\mathbf{z^0}_{(k\times 1)})^*\mathbf{H}_{(k-1\times k)}\mathbf{z^0}_{(k\times 1)}}}. \tag{3.3}$$

onde $\mathbf{w}^*$ é o transposto do conjugado de $\mathbf{w}$ e $|.|$ denota a norma complexa de $\mathbf{w}$. O vetor $\mathbf{z}$, de acordo com (KENDALL, 1984), é chamado de pré-forma da configuração complexa $\mathbf{z^0}$. É importante notar que a pré-forma é uma forma com o efeito de rotação retido.

Devido à importância da pré-forma no estudo das coordenadas de Kendall, alguns conceitos importantes devem ser considerados.

Definição: A pré-forma de uma matriz de configuração $\mathbf{X}$ é dada por

$$\mathbf{z}^F_{(k-1\times m)} = \frac{\mathbf{H}_{(k-1)\times k}\mathbf{X}_{(k\times m)}}{|\mathbf{HX}|}. \tag{3.4}$$

o qual é invariante sob locação e escala da configuração original. A partir dessa Equação pode-se obter as pré-formas centralizadas de forma que

$$\mathbf{z_{Ce}}_{(k\times m)} = \frac{\mathbf{Ce}_{(k\times k)}\mathbf{X}_{(k\times m)}}{|\mathbf{CeX}|},$$

desde que $\mathbf{Ce} = \mathbf{H}^T\mathbf{H}$. Note que $\mathbf{z}^F$ é uma matriz $(k-1)\times m$ enquanto que $\mathbf{z_{Ce}}$ é uma matriz $k\times m$. A vantagem em usar $\mathbf{z^F}$ é por ser de posto completo e sua dimensão é menor do que a de $\mathbf{z_{Ce}}$. Por outro lado, a vantagem de trabalhar com a pré-forma centralizada $\mathbf{z_{Ce}}$ é que a representação das coordenadas Cartesianas é coerente com a configuração original, como pontua (DRYDEN; MARDIA, 2016).

O espaço das pré-formas é o espaço de todas possíveis pré-forma $\mathbf{z}^F$. Ou seja, o espaço de todos os possíveis vetores de dimensão $(k-1)$ que não possuem a informação da locação e escala. Para pré-formas planas, este espaço é uma hiperesfera complexa de dimensão $(k-1)$, isto é

$$\mathbb{C}S^{k-2} = \{\mathbf{z^F} : \mathbf{z^{F*}z^F} = 1, \mathbf{z^F} \in \mathbb{C}^{k-1}\}, \tag{3.5}$$

em que $\mathbb{C}^{k-1}$ é o espaço complexo de dimensão $(k-1)$ e $\mathbf{z^{F*}}$ é o transposto do conjugado de $\mathbf{z^F}$.

Definição: A forma de uma matriz de configuração $\mathbf{X}$ é toda a informação geométrica sobre $\mathbf{X}$ que é invariante sobre locação, rotação e escala. A forma pode ser representada como

$$[\mathbf{z}] = \{e^{i\theta}\mathbf{z^F} : \theta \in [0, 2\pi]\},$$

em que θ é o grupo especial ortogonal de rotações e $\mathbf{z^F}$ é a pré-forma de $\mathbf{X}$, podendo ser qualquer uma de suas versões rotacionadas. Para $m = 2$ o espaço da forma é o espaço projetivo complexo $\mathbb{C}P^{k-2}$, o espaço de linhas complexas que passam pela origem conforme (KENDALL, 1984).

3.3 FORMA MÉDIA

A obtenção da estimativa da forma média das formas dos objetos é um importante conceito relacionado com a análise dos dados em análise estatística de formas. Para estimar a forma média de uma amostra aleatória de configurações, considere $\mathbf{z_1}^0, \ldots, \mathbf{z_n}^0$ como uma amostra aleatória de configurações complexas de uma população de n objetos ou indivíduos o qual $\mathbf{z_i}^0$ foi definido pela Equação (3.1).

Resultado: A forma média Procrustes completa $[\hat{\mu}]$ pode ser encontrada como o autovetor correspondente ao maior autovalor da soma quadrática complexa e matriz produto

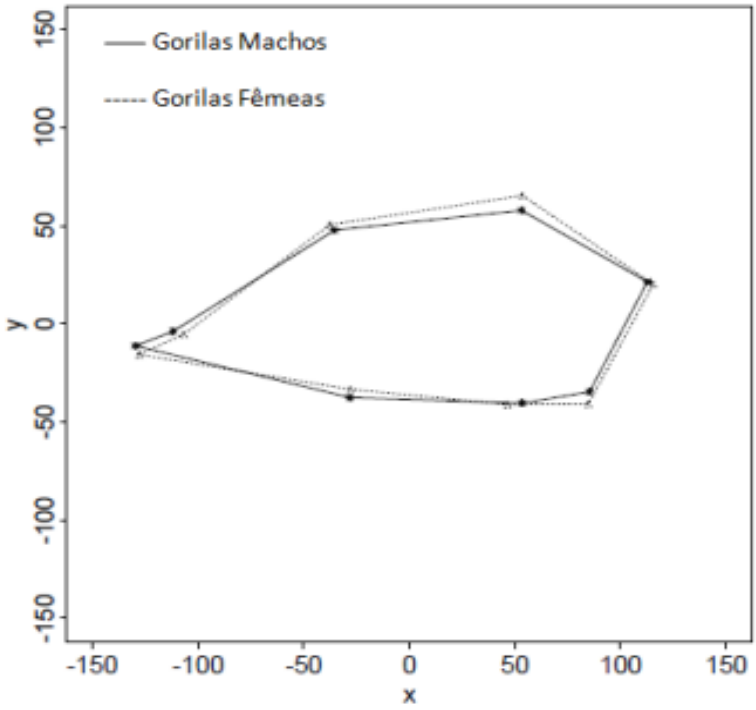
$$S = \sum_{i=1}^{n} \frac{\mathbf{w_i w_i^*}}{\mathbf{w_i^* w_i}} = \sum_{i=1}^{n} \mathbf{z_i z_i^*},$$

onde $\mathbf{z_i} = \mathbf{w_i}/||\mathbf{w_i}||$, $i = 1, \ldots, n$ são as pré-formas.

Assim, $[\hat{\mu}]$ é dado pelo autovetor complexo correspondente ao maior autovalor, ou autovetor dominante de S. O autovetor é único (até uma rotação - todas as rotações de $[\hat{\mu}]$ são também soluções, mas todos estes correspondem a mesma forma), desde que exista um único autovalor maior de S.

Para representar de maneira didática a forma média, a Figura 2, retirada do livro de (DRYDEN; MARDIA, 2016), ilustra a forma média de gorilas machos e fêmeas. Para obter essa ilustração foi considerado um conjunto de dados de crânios de gorilas que tem 29 gorilas machos e 30 gorilas fêmeas.

Figura 2 – Forma média Procrustes completa de crânios de gorilas machos e fêmeas.



Fonte: (DRYDEN; MARDIA, 2016).

As novas coordenadas resultantes do ajustamento das pré-formas dos objetos à sua forma média, para uma dada amostra de pré-formas $\mathbf{z_i}$, são chamadas ajustes Procrustes ou coordenadas Procrustes. Seja uma amostra aleatória de pré-formas e sejam as correspondentes configurações helmertizadas $\mathbf{w_1}, \ldots, \mathbf{w_n}$. De acordo com (DRYDEN; MARDIA, 2016) as configurações têm uma rotação arbitrária, assim antes de continuar com a análise estatística de formas, é necessário rotacionar todas as configurações de tal maneira que estejam o mais próximo possível da forma média amostral e isto é feito pela seguinte equação:

$$\mathbf{w_i}^P = \frac{\mathbf{w_i^*}\hat{\mu}\mathbf{w_i}}{\mathbf{w_i^*}\mathbf{w_i}} = \mathbf{z_i^*}\hat{\mu}\mathbf{z_i}, \quad i = 1, \ldots, n$$

onde cada $\mathbf{w_i^P}$ é o ajuste Procrustes completo de $\mathbf{w_i}$ em $\hat{\mu}$. A forma média Procrustes completa pode ser obtida por tomar a média aritmética das coordenadas Procrustes completa, ou seja, $\frac{1}{n}\sum_{i=1}^{n}\mathbf{w_i^P}$ tem a mesma forma como a forma média Procrustes $[\hat{\mu}]$.

3.4 DISTÂNCIA ENTRE FORMAS

Um conceito de distância entre duas formas é necessário para definir completamente o espaço métrico de forma não Euclidiana. Considere duas matrizes de configuração de k pontos

e dimensão $m = 2$, $\mathbf{X}$ e $\mathbf{Y}$, e suas configurações centradas e de tamanho unitário (pré-forma centrada) $\mathbf{z_x} = (\mathbf{z_{x1}}, \ldots, \mathbf{z_{xk}})^T$ e $\mathbf{z_y} = (\mathbf{z_{y1}}, \ldots, \mathbf{z_{yk}})^T$, de duas configurações $\mathbf{X}$ e $\mathbf{Y}$ onde $||\mathbf{z_x}|| = 1 = ||\mathbf{z_y}||$ e $\mathbf{z_x^*} 1_k = 0 = \mathbf{z_y^*} 1_k$ Dessa forma, a distância Procrustes completa entre duas formas $\mathbf{z_x}$ e $\mathbf{z_y}$

$$d_F = \sqrt{1 - |\mathbf{z_x^* z_y}|^2}, \quad 0 \leq d_F \leq 1 \tag{3.6}$$

Esta distância é invariante aos efeitos de locação, escala e rotação. Consequentemente, podemos considerar $\rho = (1 - d_F)^{1/2}$. Para dados no plano o espaço pré-forma é uma esfera complexa $\mathbb{C}S^{k-2}$ de raio unitário em dimensão complexa $k - 1$ como expresso em (3.5). O ângulo entre as pré-formas complexas $\mathbf{z_x}$ e $\mathbf{z_y}$ é

$$\rho = arccos(|\mathbf{z_x^* z_y}|), \quad 0 \leq \rho \leq \pi/2 \tag{3.7}$$

Essa quantidade também denominada como geodésica, é definida como o caminho mais curto entre $\mathbf{z_x}$ e $\mathbf{z_y}$ na hiperesfera da pré-forma e não é afetada pela rotação. Consequentemente, pode-se ver explicitamente que a distância Procrustes ρ é o ângulo entre as pré-formas $\mathbf{z_x}$ e $\mathbf{z_y}$, é também chamada de distância Riemanniana. Desde que, o raio da esfera da pré-forma seja 1 pode-se considerar ρ como sendo a distância ótima no círculo na esfera da pré-forma.

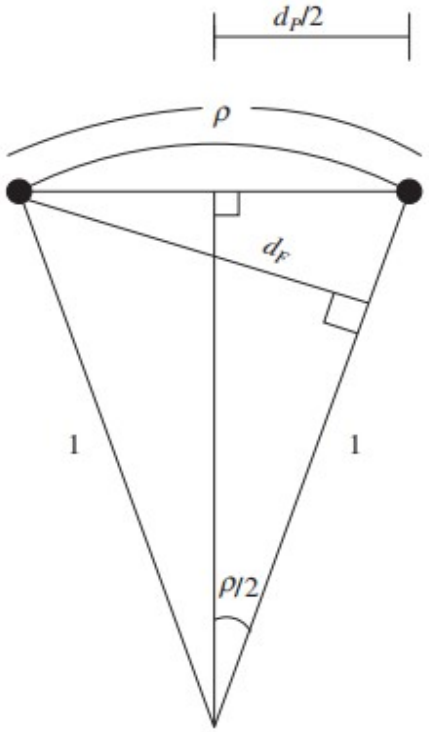
A distância Procrustes parcial também é invariante quanto a rotação entre $\mathbf{z_x}$ e $\mathbf{z_y}$, é definida por

$$d_P = \sqrt{2(1 - |\mathbf{z_x^* z_y}|)} = \sqrt{2(1 - \cos\rho)}, \quad 0 \leq d_p \leq \sqrt{2} \tag{3.8}$$

A representação das distâncias definidas em (3.6), (3.7) e (3.8) na hiperespera da pré-forma pode ser observada por meio da Figura 3.

Um resumo com as distâncias (3.6), (3.7) e (3.8) é apresentada na Tabela 1. Nos algoritmos de agrupamento desenvolvidos ao longo da pesquisa, usamos uma das distâncias resumidas na Tabela 1 como critério de dissimilaridade entre as formas.

Figura 3 – Ilustração da relação entre as distâncias d_F, d_P, e ρ na esfera da pré-forma.



Fonte: (DRYDEN; MARDIA, 2016).

Tabela 1 – Distâncias no espaço de formas.

Distâncias	**Notação**	**Fórmula**	**Intervalo**
Distância Procrustes Completa	d_F	$d_F = \sqrt{1 - \lvert\mathbf{z}_\mathbf{x}^* \mathbf{z}_\mathbf{y}\rvert^2}$	$0 \leq d_F \leq 1$
Distância Procrustes Parcial	d_P	$d_P = \sqrt{2(1 - \lvert\mathbf{z}_\mathbf{x}^* \mathbf{z}_\mathbf{y}\rvert)}$	$0 \leq d_p \leq \sqrt{2}$
Distância Riemanniana	ρ	$\rho = arccos(\lvert\mathbf{z}_\mathbf{x}^* \mathbf{z}_\mathbf{y}\rvert)$	$0 \leq \rho \leq \pi/2$

3.5 DISTRIBUIÇÃO BINGHAM COMPLEXA

A distribuição Bingham Complexa introduzida por (KENT, 1994) é uma das principais distribuições para análise de formas dos marcos em duas dimensões. Foi obtida a partir da distribuição Bingham real que é utilizada para dados esféricos. A distribuição Bingham Complexa é uma distribuição no espaço de vetores unitários complexo ou, na esfera unitária complexa. Vamos considerar o caso de uma distribuição de probabilidade na espera de pré-formas $\mathbb{C}S^{k-1}$, em que $\mathbb{C}S^{k-1}$ é uma esfera unitária complexa em $k-1$ dimensão. No caso da análise de formas, com k marcos anatômicos e $m = 2$ dimensão, considerando $\mathbf{z}^\mathbf{F}$ como defindido em (3.4), a distribuiçao Bingham Complexa em $\mathbb{C}S^{k-1}$ denotada por $\mathbb{C}B_{k-1}(A)$ tem função densidade de probabilidade dada por

$$f(\mathbf{z}^{\mathbf{F}}) = c(A)^{-1} \exp(\mathbf{z}^{F*} A \mathbf{z}^{\mathbf{F}}), \quad \mathbf{z}^{\mathbf{F}} \in \mathbb{C}S^{k-1} \tag{3.9}$$

em que, $\mathbf{z}^{F*}$ é o conjulgado transposto complexo de $\mathbf{z}^{\mathbf{F}}$, A é uma matriz Hermetiana $k \times k$ e $c(A)$ é uma constante de normalização dada por

$$c(A) = 2\pi^{k-1} \sum_{i=1}^{k-1} a_i \exp(\lambda_i), \quad a_i^{-1} = \prod_{i \neq j} (\lambda_j - \lambda_i)$$

em que $\lambda_1 < \lambda_2 < \ldots \lambda_{k-1} = 0$ denota os autovalores de A. Note que $c(A) = c(\wedge)$ depende somente dos autovalores de A, com $\wedge = diag(\lambda_1, \ldots, \lambda_{k-1})$, a prova para esse resultado pode ser encontrada em (DRYDEN; MARDIA, 2016). Além disso, se $A = I$, então $f(\mathbf{z}^{\mathbf{F}})$ torna-se uma distribuição uniforme sobre $\mathbb{C}S^{k-1}$ devido a restrição $\mathbf{z}^{F*}\mathbf{z}^{\mathbf{F}} = 1$. Uma vez que $\mathbf{z}^{F*}\mathbf{z}^{\mathbf{F}} = 1$ para $\mathbf{z}^{\mathbf{F}} \in \mathbb{C}S^{k-1}$ pode-se notar que as matrizes A e $A + \alpha I$ definem a mesma distribuição Bingham Complexa com $c(A + \alpha I) = c(A)e^{\alpha}$, com α sendo um número complexo.

A distribuição Bingham Complexa tem a importante propriedade de simetria complexa, em que $\mathbf{z}^{\mathbf{F}}$ e $e^{i\theta}\mathbf{z}^{\mathbf{F}}$ tem a mesma distribuição, isto é

$$f(e^{i\theta}\mathbf{z}^{\mathbf{F}}) = f(\mathbf{z}^{\mathbf{F}}). \tag{3.10}$$

Devido a essa propriedade, a distribuição Bingham Complexa é invariante em relação as rotações das pré-forma $\mathbf{z}^{\mathbf{F}}$, fazendo com que seja adequada para o cenário da análise de dados de formas em duas dimensões (KENT, 1994). Para mais detalhes sobre as propriedade da distribuição Bingham Complexa ver (KENT, 1994).

Para simular a distribuição Bingham complexa, vários métodos foram propostos por (KENT; CONSTABLE; ER, 2004). Nesta dissertação, usaremos o método Truncação pelo simplex. Inicialmente, gera-se $(k-1)$ exponenciais truncadas $T\exp(\lambda_j)$, pelo método de aceitação e rejeição e, então, essas variáveis aleatórias são expressas em coordenadas polares para obter uma distribuição Bingham Complexa. O Algoritmo 1 representa os passos para simular a distribuição exponencial truncada multivariada através do método de aceitação e rejeição.

Algoritmo 1: Simulação da distribuição exponencial truncada

Output: Retorna Distribuição exponencial truncada $S_j \sim T\exp(\lambda_j), \quad j = 1, \ldots, k-1$.

1. Simule uma variável aleatória $U_j \sim U[0,1], \quad j = 1, \ldots, k-1$.

2. Seja $S_j' = -(1/\lambda_j)\log(1 - U_j(1 - e^{-\lambda_j}))$, $S' = (S_1', \ldots, S_{k-1}')^T$ de modo que S_j' são amostras aleatórias independentes $T\exp(\lambda_j)$.

3. Se $\sum_{j=1}^{k-1} S_j' < 1$, estabeleça $S = S'$. Por outro lado, rejeita-se S' e retorne ao passo 1.

O método para simular uma distribuição Bingham Complexa usa $(k-1)$ exponenciais truncadas para gerar um vetor k de uma distribuição Bingham Complexa. Seja $\lambda_1 \geq \ldots \geq \lambda_k = 0$ os autovalores de $-A$, com $\lambda = (\lambda_1, \ldots, \lambda_{k-1})$ sendo o vetor dos primeiros $k-1$ autovalores, o Algoritmo 2 retorna um vetor k, com $\mathbf{z^F} = (\mathbf{z_1^F}, \ldots, \mathbf{z_k^F})$ tem distribuição Bingham Complexa.

Algoritmo 2: Simulação da Distribuição Bingham complexa

Output: Retorna Distribuição Bingham complexa.

1. Gere $S = (S_1, \ldots, S_{k-1})$ onde $S_j \sim Texp(\lambda_j)$ são variáveis aleatórias independentes usando Algoritmo 1.

2. Se $\sum_{j=1}^{k-1} S_j' < 1$, escreva $S_k = 1 - \sum_{j=1}^{k-1} S_j'$. Por outro lado volte ao passo 1.

3. Gere ângulos independentes $\theta_j \sim U[0, 2\pi], \quad j = 1, \ldots, k$;

4. Calcule $z_j^F = S_j^{1/2}\exp(i\theta_j), \quad j = 1, \ldots, k$.

Com base na fundamentação teórica apresentada até aqui, abaixo descreveremos o algoritmo k-means para formas planas, utilizado neste trabalho como referência comparativa para os métodos de agrupamento espectral propostos.

3.6 ALGORITMO k-MEANS PARA FORMAS PLANAS

O algoritmo k-means possui como objetivo particionar n observações dentre k_g grupos de modo que cada observação pertença ao grupo cuja distância entre essa observação e o protótipo do grupo é mínima. Adaptamos o algoritmo k-means introduzido por (MACQUEEN, 1967) para dados de formas planas utilizando como base o trabalho desenvolvido por (AMARAL et al., 2010).

Seja um conjunto de n objetos ou indivíduos a ser agrupados em um conjunto de k_g grupos, $C = (C_r, r = 1, \ldots, k_g)$. O algoritmo k-means encontra uma partição minimizando um critério que mede distância entre pré-formas de grupos e a formas média:

$$J(C_r) = \sum_{i \in C_r} dist^2(\mathbf{z}_i, \mu_\mathbf{r}), \tag{3.11}$$

em que $dist^2$ é uma medida de distância geral como as definidas pelas Equações (3.6), (3.7) e (3.8). O objetivo do k-means é minimizar a soma do erro quadrático sobre o grupo k_g

$$J = \sum_{r=1}^{k_g} \sum_{i \in C_r} dist^2(\mathbf{z}_i, \mu_\mathbf{r}).$$

Uma descrição do algoritmo k-means (k_{FP}) adaptada para trabalhar com dados de análise estatística de formas é apresentada no Algoritmo 3. A usual distância Euclidiana foi substituída por uma das distâncias Procrustes para os dados no espaço de pré-formas.

Algoritmo 3: Agrupamento k-means para formas planas

Input: Pré-forma definida em (3.3) ou (3.4), número de grupos k_g e alocação inicial

Output: Grupos C_r $(1 \leq r \leq k_g)$

1. Obtenha a forma média para cada grupo;

2. Atribua cada objeto à forma média do grupo mais próximo, através das Equações (3.6), (3.7) ou (3.8);

3. Calcule a forma média de cada grupo;

4. Repita os passos 2 e 3 até que a forma média não mude ou um valor ótimo da Equação (3.11) seja encontrado.

Este algoritmo move os objetos entre os agrupamentos até que a função objetivo não se altere ou se altere muito pouco, ou até que o número de iterações máxima pré determinado seja alcançado. O resultado é um conjunto de grupos com indivíduos com características homogêneas dentro dos grupos e com características heterogêneas entre os grupos. Apesar do algoritmo convergir rapidamente para uma solução, essa solução encontrada depende da alocação inicial, logo o método pode convergir para um ótimo local. A partir daqui, chamaremos esse algoritmo de k_{FP}, algoritmo k-means para formas planas.

4 CONTRIBUIÇÕES ALGORÍTMICAS

O agrupamento é uma das técnicas mais consolidadas na análise multivariada. O desenvolvimento de novos métodos nessa área é essencial para aprofundar a compreensão e o tratamento de dados complexos em diversas disciplinas. Com o crescimento exponencial no volume e na diversidade dos dados, os métodos tradicionais muitas vezes se mostram insuficientes para capturar padrões intrínsecos, especialmente em contextos com alta dimensionalidade, presença de ruído ou estruturas não lineares. Abordagens inovadoras possibilitam a identificação de agrupamentos mais representativos, aumentando a precisão das análises e proporcionando resultados mais confiáveis, o que contribui para uma tomada de decisão mais eficaz em múltiplas áreas do conhecimento.

Os avanços tecnológicos têm levado à coleta rotineira de informações geométricas, tornando o estudo das formas dos objetos cada vez mais relevante. A análise de formas é uma área promissora e eficaz para lidar com dados geométricos e estruturas morfológicas. A formalização desse campo, proposta por (KENDALL, 1984), define a forma como uma representação em um sistema de coordenadas baseado em marcos anatômicos, situadas em um espaço não Euclidiano. Em muitos contextos da análise de formas, há interesse em agrupar conjuntos de dados; no entanto, uma limitação significativa dos algoritmos de agrupamento convencionais é que eles foram projetados para operar em espaços Euclidianos. Isso se deve ao fato de que tais algoritmos, em geral, utilizam a distância Euclidiana como medida de dissimilaridade, o que pressupõe que os dados estejam inseridos em um espaço Euclidiano. Diante disso, torna-se necessário o desenvolvimento de algoritmos específicos capazes de operar em espaços não Euclidianos, especialmente quando se trabalha com dados no espaço de formas.

O desenvolvimento de novos algoritmos de agrupamento espectral tem revolucionado a análise de dados complexos, permitindo a identificação de estruturas intrínsecas em conjuntos de dados de alta dimensionalidade. As novas abordagens incorporam métodos dinâmicos para lidar com estruturas de dados diverso, expandindo a aplicabilidade do agrupamento espectral em várias áreas. Nesta pesquisa foram propostos algoritmos de agrupamento espectral adaptados para particularidades dos dados da área de análise estatística de formas de objetos. Os algoritmos de agrupamento espectral de (NG; JORDAN; WEISS, 2001), (LIU L; CHEN, 2013), (WANG et al., 2017), (LUXBURG, 2007) ganharam novas versões utilizando uma nova medida de distância para dados de formas planas. A seguir, apresentaremos os algoritmos que foram

desenvolvidos ao longo desse trabalho.

4.1 ALGORITMO DE AGRUPAMENTO ESPECTRAL CLÁSSICO PARA FORMAS PLANAS

O primeiro algoritmo de agrupamento espectral adaptado para dados da área de análise estatística de formas de objetos, que chamaremos de algoritmo espectral clássico para formas planas (AEC_{FP}), foi baseado no algoritmo proposto por (NG; JORDAN; WEISS, 2001).

Nesse trabalho, o particionamento espectral do grafo é obtido utilizando o segundo autovetor da matriz Laplaciana. Aqui, o autovetor é visto como uma solução para uma relaxação do problema de particionamento do grafo, em que é possível mostrar que cortes baseados no segundo autovetor fornecem uma aproximação garantida do corte ideal (CHUNG, 1997). Os autores se basearam no trabalho de (WEISS, 1999) e (MEILA; SHI, 2000).

O algoritmo de (NG; JORDAN; WEISS, 2001) constrói a matriz de Similaridade utilizando a função kernel gaussiana definida em (2.1). Em posse da matriz de Similaridade, constrói a matriz Laplaciana ($\mathbf{L}$) e extrai os k_g autovetores associados aos k_g maiores autovalores de $\mathbf{L}$, normalizando-os em seguida; O agrupamento é obtido usando o algoritmo k_g-means clássico, já que na etapa de obtenção dos autovetores de $\mathbf{L}$, mapea-se os dados para dimensão $\mathbb{R}^{k_g}$, formando grupos mais compactos, podendo obter um bom resultado no agrupamento utilizando o k-means. Esse algoritmo difere do método proposto por (MEILA; SHI, 2000), neste os autores normalizam a matriz de Similaridade e usam seus autovetores em vez de $\mathbf{L}$ para obter o agrupamento, além disso, não normalizam os autovetores obtidos.

No trabalho de (JAYASUMANA et al., 2013) é apresentado o núcleo gaussiano de Procrustes, um núcleo definido positivo comprovadamente na variedade de formas. Para utilizar o espaço de Hilbert é preciso, de acordo com o teorema de Mercer, que a função kernel seja definida positiva para definir um espaço válido. Sendo definida positiva, esta função kernel permite incorporar a variedade de formas em um espaço de Hilbert de alta dimensão. A vantagem de tal incorporação é permitir o uso de métodos de reconhecimento bem estabelecidos que requerem geometria linear, além de incorporar um espaço de dimensão inferior em um espaço de dimensão superior, ajudando a identificar padrões complexos em determinado distribuição de dados.

O kernel proposto por (JAYASUMANA et al., 2013) é inspirado na função de base radial gaussiano (RBF), eficaz em espaços Euclidianos. Mais especificamente, substituem a distância

Euclidiana no RBF pela distância Procrustes completa, definida em (3.6), na variedade de formas. Isso produz o kernel $k_P(\mathbf{z}_i, \mathbf{z}_j) = \exp(-d_F^2(\mathbf{z}_i, \mathbf{z}_j)/2\sigma^2)$, denominado Kernel gaussiano de Procrustes, que é positivo definido.

Teorema: O kernel gaussiano de Procrustes, $k_P : (SP_m \times SP_m) \to \mathbb{R}$:

$$k_p(\mathbf{z}_i, \mathbf{z}_j) = \exp(-d_F^2(\mathbf{z}_i, \mathbf{z}_j)/2\sigma^2), \tag{4.1}$$

é um kernel definido positivo para todo $\sigma \in \mathbb{R}$, em que d_F é a distância Procrustes completa entre duas formas.

Para detalhes quanto a prova desse teorema ver (JAYASUMANA et al., 2013). A proposta do trabalho consiste na adaptação do cálculo da matriz de similaridade no algoritmo de (NG; JORDAN; WEISS, 2001) com base na função kernel, usando a distância de Procrustes completa, uma vez que esta é a única distância que satisfaz as condições de Mercer, como provado em (JAYASUMANA et al., 2013).

O Algoritmo 4 resume o passo a passo para obter o agrupamento de formas utilizando o algoritmo espectral clássico para formas planas.

Algoritmo 4: Agrupamento espectral clássico para formas planas

Input: Pré-forma definida em (3.3) ou (3.4) e número de grupos k_g

Output: Grupos C_r $(1 \leq r \leq k_g)$

1. Calcule a matriz de Similaridade $W = (w_{ij}) \in \mathbb{R}^{n \times n}$, utilizando (4.1);

2. Calcule a matriz Laplaciana normalizada;

3. Encontre os k_g autovetores de $\mathbf{L}$ associados aos seus k_g maiores autovalores;

4. Construa uma matriz $\mathbf{Z}$ concatenando os k_g autovetores associados aos k_g maiores autovalores de $\mathbf{L}$;

5. Obtenha a matriz $\mathbf{Y}$ normalizada através de $\mathbf{Z}$, aplicando $y_{ij} = z_{ij}/\sum_{r=1}^{k_g} z_{ir}^2$, fazendo com que todas as linhas de $\mathbf{Z}$ tenham norma unitária;

6. Usando $\mathbf{Y}$, obter os grupos utilizando o algoritmo k-means clássico.

4.2 ALGORITMO DE AGRUPAMENTO ESPECTRAL COMBINADO COM AGRUPAMENTO HIERÁRQUICO PARA FORMAS

O segundo algoritmo de agrupamento espectral adaptado para dados da área de análise estatística de formas de objetos, que chamaremos de algoritmo espectral hierárquico para formas planas (AEH_{FP}), foi baseado no algoritmo desenvolvido por (LIU L; CHEN, 2013).

No trabalho de (LIU L; CHEN, 2013), o agrupamento é obtido utilizando método hierárquico aglomerativo para agrupar o conjunto de dados. Os autores consideram que ao utilizarem o método hierárquico, o agrupamento espectral considera a relação entre a vizinhança e lida de forma mais eficaz quando diante de dados de vizinhos ruidosos. Sendo assim, o algoritmo espectral hierárquico adaptado para dados de formas planas desenvolvido nesse pesquisa é semelhante ao algoritmo espectral clássico para formas planas apresentado anteriormente diferenciando-se, porém, no método de agrupamento utilizado na etapa final do algoritmo, sendo utilizado um dos método de agrupamento hierárquico aglomerativo discutidos anteriormente.

O Algoritmo 5 resume o passo a passo para obter o agrupamento de formas utilizando o algoritmo espectral hierárquico para formas planas.

Algoritmo 5: Agrupamento espectral hierárquico para formas planas

Input: Pré-forma definida em (3.3) ou (3.4) e número de grupos k_g

Output: Grupos C_r $(1 \leq r \leq k_g)$

1. Calcule a matriz de Similaridade $W = (w_{ij}) \in \mathbb{R}^{n \times n}$, utilizando (4.1);

2. Calcule a matriz Laplaciana normalizada;

3. Encontre os k_g autovetores de $\mathbf{L}$ associados aos seus k_g maiores autovalores;

4. Construa uma matriz $\mathbf{Z}$ concatenando os k_g autovetores associados aos k_g maiores autovalores de $\mathbf{L}$;

5. Obtenha a matriz $\mathbf{Y}$ normalizada através de $\mathbf{Z}$, aplicando $y_{ij} = z_{ij} / \sum_{r=1}^{k} z_{ir}^2$, fazendo com que todas as linhas de $\mathbf{Z}$ tenham norma unitária;

6. Usando $\mathbf{Y}$, obter os grupos utilizando o algoritmo hierárquico aglomerativo.

A ideia principal do algoritmo é, portanto, utilizar o agrupamento hierárquico em vez do k-means no agrupamento espectral clássico com o intuito de eliminar as informações enganosas de vizinhos ruidosos nos dados. Esse algoritmo apresenta limitações quando diante de grandes

bases de dados.

4.3 ALGORITMO DE AGRUPAMENTO ESPECTRAL BASEADO EM CRITÉRIO DE SIMILARIDADE E DISSIMILARIDADE PARA FORMAS PLANAS

O terceiro algoritmo de agrupamento espectral adaptado para dados da área de análise estatística de formas de objetos, que chamaremos de algoritmo espectral similaridade e dissimilaridade para formas planas ($AESD_{FP}$), foi baseado no algoritmo desenvolvido por (WANG et al., 2017).

O trabalho de (WANG et al., 2017) propõe um algoritmo de agrupamento espectral baseado em um critério de similaridade e dissimilaridade, incorporando um critério de dissimilaridade no critério de corte normalizado do grafo. O critério de partição desempenha um papel importante no desempenho de agrupamento espectral. Esses critérios têm suas vantagens e desvantagens (DING et al., 2024), no entanto, são projetados de acordo com o critério de similaridade.

No critério de corte normalizado (SHI; MALIK, 2000) a similaridade entre grupos tem um impacto negativo em maximizar a similaridade dentro do grupo. Para reduzir esse efeito negativo, (WANG et al., 2017) introduz o conceito de dissimilaridade no critério de corte normalizado para tornar a semelhança e dissimilaridade das amostras mais claras e melhorar o desempenho do agrupamento.

Suponha que um grafo G seja particionado em dois subgrafos X_1 e X_2 onde $X_1 \cup X_2 = V$ e $X_1 \cap X_2 = \varnothing$. A Equação (2.3) pode, também, ser escrita como:

$$Ncut(X_1, X_2) = \frac{cut(X_1, X_2)}{assoc(X_1, V)} + \frac{cut(X_1, X_2)}{assoc(X_2, V)}, \tag{4.2}$$

onde $cut(X_1, X_2) = \sum_{x_u \in X_1} \sum_{x_v \in X_2} W_{uv}$ e $assoc(X_i, V) = \sum_{x_u \in X_i} \sum_{x_p \in V} W_{uv}$.

Na verdade, $cut(X_1, X_2)$ pode ser considerado como a similaridade entre grupos, e $assoc(X_i, V)$ é a soma da similaridade entre grupos e similaridade dentro do grupo. O critério de corte normalizado pode não apenas maximizar o similaridade dentro dos grupos, mas também maximizar a dissimilaridade entre grupos. O que é mais importante é que o corte normalizado também pode evitar com eficiência a preferência de partição de pequena região que ocorre frequentemente no critério $minCut$. Outra forma de representar 4.2 é:

$$min_y = \frac{\mathbf{y^T(D - W)y}}{\mathbf{y^T D y}}. \tag{4.3}$$

onde $\mathbf{D}$ é uma matriz diagonal cujos elementos $\mathbf{d}_{ii} = \sum_{j=1}^{n} W_{ij}$. A solução de 4.3 é equivalente a solução de $(\mathbf{D} - \mathbf{W})\mathbf{y} = \lambda \mathbf{D}\mathbf{y}$.

O objetivo do corte normalizado é maximizar a similaridade dentro do grupo e minimizar a semelhança entre grupos, mas o denominador do objetivo em (4.3), pode ser visto aproximadamente como a soma da similaridade dentro do grupos e da similaridade entre grupos. É um contradição que maximizamos $\mathbf{y^T D y}$ enquanto faz o a similaridade entre grupos seja minimizada. Note que, se a similaridade dentro do grupo for grande o suficiente, o efeito da similaridade entre grupos sobre o denominador pode ser ignorado. Mas a maximização similaridade entre grupos no denominador ainda tem alguns efeitos negativos no desempenho do agrupamento.

A suposição de agrupamento é maximizar a similaridade dentro do grupo e, simultaneamente, minimizar a similaridade entre grupos, que também pode ser descrita como minimizar a dissimilaridade dentro do grupo e maximizar a dissimilaridade entre os grupo, chamado de critério da dissimilaridade. Seja $\mathbf{Q}$ a matriz de dissimilaridade, onde $\mathbf{Q}_{ij}$ está em proporção direta com a distância, quanto menor é a semelhança, maior a dissimilaridade. Se (2.1) for selecionado para medir a similaridade, então a medida da dissimilaridade é dado por

$$\mathbf{Q}_{ij} = 1 - \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2}\right).$$

Importante destacar que (WANG et al., 2017) não considera um valor fixo para σ^2, como nos algoritmos anteriores, σ^2 é calculado de acordo com (ZELNIK-MANOR; PERONA, 2004). Portanto, $\mathbf{Q}_{ij}$ é reescrito como

$$\mathbf{Q}_{ij} = 1 - \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\sigma_\mathbf{i}\sigma_\mathbf{j}}\right). \tag{4.4}$$

Queremos reduzir o efeito da similaridade entre grupos no denominador do objetivo em (4.3) no corte normalizado. Assim, introduzimos o critério de dissimilaridade no denominador do objetivo em (4.3), obtemos:

$$min_y = \frac{\mathbf{y^T}(\mathbf{D} - \mathbf{W})\mathbf{y}}{(\mathbf{1} - \mathbf{m})\mathbf{y^T D y} - \mathbf{m}\mathbf{y^T Q y}}, \tag{4.5}$$

onde $0 < m \leq 1$ é um fator de compensação que determina o influência do critério de dissimilaridade. Quanto maior o fator m, maior será a influência. Se $m = 0$, então o método proposto aqui é totalmente igual ao agrupamento espectral baseado em corte normalizado.

Teorema : Dado $0 < m \leq 1$ em (4.5), a similaridade entre grupos no denominador de (4.5) tem menor peso que a similaridade entre os grupos no denominador de (4.3).

A prova do Teorema se encontra em (WANG et al., 2017). Esse Teorema afirma que o efeito da similaridade entre grupos na maximização da similaridade dentro dos grupos pode ser reduzido desde que o critério de dissimilaridade seja introduzido no critério de corte normalizado.

Note que, quando $\mathbf{Q}_{ij}$ é muito pequeno, contribui menos para o segundo termo do denominador e as duas amostras $\mathbf{x_i}$ e $\mathbf{x_j}$ tendem a ser particionados no mesmo grupo. Ao mesmo tempo, a semelhança $\mathbf{W}_{ij}$ no numerador é grande. Quando $\mathbf{Q}_{ij}$ é muito grande, as duas amostras $\mathbf{x_i}$ e $\mathbf{x_j}$ tendem ser particionadas em grupos diferentes. Devido ao efeito de $\mathbf{Q}_{ij}$, o valor de $m\mathbf{Q}_{ij}$ no denominador também é muito grande. Neste caso, embora tenhamos aumento no valor de $(y_i - y_j)^2$ para fazer os pontos de dados estando longe um do outro, não teria um enorme efeito na otimização do objetivo. Fazendo isso, podemos aumentar a dissimilaridade entre os grupos.

Podemos, agora, apresentar o agrupamento espectral similaridade e dissimilaridade para formas planas expresso no Algoritmo 6.

Algoritmo 6: Agrupamento espectral similaridade e dissimilaridade para formas planas

Input: Pré-forma definida em (3.3) ou (3.4) e número de grupos k_g

Output: Grupos C_r $(1 \leq r \leq k_g)$

1. Calcule a matriz de Similaridade $W = (w_{ij}) \in \mathbb{R}^{n \times n}$, utilizando (4.1) e (ZELNIK-MANOR; PERONA, 2004) e gere Q de acordo com (4.4);

2. Encontre os k_g autovetores associados aos seus k_g maiores autovalores $(\mathbf{D} - \mathbf{W})$ e $((1-m)\mathbf{D} - m\mathbf{Q})$;

3. Construa uma matriz $\mathbf{Z}$ concatenando os k_g autovetores associados aos k_g maiores autovalores de $\mathbf{L}$;

4. Obtenha a matriz $\mathbf{Y}$ normalizada através de $\mathbf{Z}$, aplicando $y_{ij} = z_{ij} / \sum_{r=1}^{k} z_{ir}^2$, fazendo com que todas as linhas de $\mathbf{Z}$ tenham norma unitária;

5. Usando $\mathbf{Y}$, obter os grupos utilizando o algoritmo k-means clássico.

4.4 ALGORITMO DE AGRUPAMENTO ESPECTRAL BASEADO EM K-NN PARA FORMAS PLANAS

O quarto algoritmo de agrupamento espectral adaptado para dados da área de análise estatística de formas de objetos, que chamaremos de algoritmo espectral baseado em k-NN para formas planas ($AEkNN_{FP}$), constrói a matriz de similaridade dos dados por meio de k-vizinhos mais próximos (LUXBURG, 2007).

O algoritmo k-vizinhos mais próximos (k-NN) é um método, frequentemente, usado em classificação desenvolvido por (FIX, 1985). Um objeto é classificado considerando seus k_v vizinhos mais próximos, sendo assim, o uso do algoritmo k-NN envolve a definição de uma métrica de distância a ser calculada a fim de determinar os vizinhos mais próximos; e um valor para o parâmetro k_v, que determina o número de vizinhos considerados próximos ao objeto.

Seja um conjunto de dados $\mathbf{x_1}, \ldots, \mathbf{x_n}$ e alguma similaridade $s_{ij} \geq 0$ entre todos os pares de dados $\mathbf{x}_i$ e $\mathbf{x}_j$. Em seguida, construa um grafo $G = (V, E)$ em que os dados são os vértices do grafo e os valores da similaridade são os pesos de cada aresta. O objetivo é conectar o vértice v_i com o vértice v_j, se v_j estiver entre os k-vizinhos mais próximos de v_i, isto é: conecte v_i e v_j com uma aresta não direcionada se v_i estiver entre os k-vizinhos mais próximos de v_j ou se v_j estiver entre os k-vizinhos mais próximos de v_j. A matriz resultante é o que geralmente é chamado de matriz de k-vizinho mais próximo.

Seja $\{x_i^p \mid p \in \{1, \ldots, n-1\}\}$ a sequência dos $n-1$ vizinhos mais próximos do ponto x_i, ordenados da menor para a maior distância, segundo uma métrica apropriada. Aqui, x_i^p denota o índice do p-ésimo vizinho mais próximo de x_i no conjunto de dados. A matriz de similaridade $\mathbf{W}^{(k)}$, associada ao grafo dos k-vizinhos mais próximos, é definida por:

$$\mathbf{W}_{ij}^{(k)} = \begin{cases} 1, & \text{se } j \in \mathcal{N}_k(i) \text{ ou } i \in \mathcal{N}_k(j) \\ 0, & \text{caso contrário,} \end{cases} \tag{4.6}$$

Note que, j representa o índice de um ponto do conjunto de dados — assim como i; $\mathbf{W}_{ij}^{(k)}$ tem dimensão $n \times n$ onde n é o tamanho da amostra e $\mathcal{N}_k(i)$ denota o conjunto dos k-vizinhos mais próximos do ponto x_i e $\mathcal{N}_k(j)$ denota o conjunto dos k vizinhos mais próximos do ponto x_j. No contexto de análise de formas, considere x_i^p como definido em (3.3) ou (3.4).

O Algoritmo 7 resume o passo a passo para obter o agrupamento de formas utilizando o algoritmo espectral baseado em k-NN para formas planas.

Algoritmo 7: Agrupamento espectral baseado em k-NN para formas planas

Input: Pré-forma definida em (3.3) ou (3.4) e número de grupos k_g

Output: Grupos C_r $(1 \leq r \leq k_g)$

1. Calcule a matriz de Similaridade $W^k \in \mathbb{R}^{n \times n}$, utilizando (4.6);

2. Calcule a matriz Laplaciana normalizada;

3. Encontre os k_g autovetores de $\mathbf{L}$ associados aos seus k_g maiores autovalores;

4. Construa uma matriz $\mathbf{Z}$ concatenando os k_g autovetores associados aos k_g maiores autovalores de $\mathbf{L}$;

5. Obtenha a matriz $\mathbf{Y}$ normalizada através de $\mathbf{Z}$, aplicando $y_{ij} = z_{ij} / \sum_{r=1}^{k} z_{ir}^2$, fazendo com que todas as linhas de $\mathbf{Z}$ tenham norma unitária;

6. Usando $\mathbf{Y}$, obter os grupos utilizando o algoritmo k-means clássico.

Ao utilizar o método k-NN para a construção da matriz de similaridade dos dados no agrupamento espectral, a escolha de k_v pode ter uma influência significativa no precisão dos grupos detectados (ZHANG et al., 2017). Em seu trabalho, (VEENSTRA; COOPER; PHELPS, 2016) apresenta os benefícios de construir uma matriz de similaridade unindo o grafo k-NN e a árvore geradora mínima da matriz de Similaridade negada, desenvolvendo uma versão do algoritmo espectral menos dependente da escolha de k_v. Nesse trabalho, testamos diversos valores para $k_v = \{5, 6, 7, 8, 9, 10\}$, consideramos $k_v = 6$, uma vez que para os demais valores, não houve mudança no resultado do agrupamento.

5 AVALIAÇÃO NUMÉRICA E RESULTADOS

Para validar os métodos de agrupamento espectral para análise estatística de formas planas, sete aplicações com os conjuntos de dados reais de Crânios Gorilas, Crânios de Orangotangos, Crânios de Chimpanzés, Crânios de Macacos, Cérebros de adultos saudáveis, Vértebras de ratos e Cérebros de pacientes esquizofrênicos e não esquizofrênicos. Além dos dados reais, foram utilizados conjuntos de dados simulados com diferentes concentrações, utilizando experimento Monte Carlo no espaço de pré-formas que corresponde a uma esfera no espaço complexo através da distribuição Bingham complexa, em que a distância Procustes completa substituiu as medidas de distância comumente usadas em algoritmos de agrupamento no espaço Euclidiano.

Os algoritmos utilizados nos experimentos foram implementados na linguagem de programação (R, 2024) (ver http://cran.r-project.org) utilizando o ambiente de programação RStudio na sua versão 4.3.2 disponível em https://www.rstudio.com/. Da linguagem de programação (R, 2024), foi utilizado o pacote shapes desenvolvido por (DRYDEN; MARDIA, 2016) que possui diversas funções que nos auxiliaram no desenvolvimento dos algoritmos para análise estatística de formas. A avaliação do desempenho dos algoritmos nos conjuntos de dados foi realizada através do índice de Rand corrigido (IRC) (HUBERT; ARABIE, 1985).

5.1 ÍNDICE DE RAND CORRIGIDO

Em (HUBERT; ARABIE, 1985), os autores definiram índices através da comparação de duas partições. Quando duas partições são independentes, no sentido em que uma depende dos valores no conjunto de dados e a outra não, a distribuição de alguns desses índices pode ser estabelecida a partir de uma perspectiva teórica.

O índice de Rand corrigido é um índice externo de avaliação que mede a similaridade entre uma partição rígida a priori e uma partição obtida fornecida por um algoritmo de agrupamento. O índice de Rand corrigido tem seus valores contidos no intervalo [-1,1], tal que quanto mais próximos de 1, mais semelhante a partição encontrada pelo método e à partição definida a priori; enquanto valores perto de 0 (ou mesmo negativos) correspondem a partições obtidas por acaso. Faremos uma definição mais formal do índice de Rand corrigido a seguir.

Seja $U = \{u_1, u_2, \ldots, u_i, \ldots, u_R\}$ e $V = \{v_1, v_2, \ldots, v_j, \ldots, v_C\}$ duas partições do mesmo conjunto de dados tendo respectivamente R e C grupos. O índice de Rand corrigido é dado

pela Equação (5.1):

$$IRC = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}}{\frac{1}{2}[\sum_{i=1}^{R}\binom{n_{i.}}{2} + \sum_{j=1}^{C}\binom{n_{.j}}{2}] - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}} \tag{5.1}$$

onde $\binom{n}{2} = \frac{n(n-1)}{2}$ e n_{ij} representam o número de objetos em comum que estão nos grupos u_i e v_i; n_i indica o número de objetos no grupo u_i; $n_{.j}$ indica o número de objetos no grupo v_j; e n é o número total de objetos no conjunto de dados.

5.2 AVALIAÇÃO NUMÉRICA - SIMULAÇÕES

O conjunto de dados simulados no espaço de pré-formas foi gerado com distribuição Bingham complexa, cuja densidade é definida em (3.9). Foram utilizados três marcos anatômicos $(k = 3)$ para amostras com número total de objetos igual a $n = 30$ e $n = 50$. Diferentes valores para a concentração $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ foram utilizados, gerando três cenários: $\boldsymbol{\lambda_1} = (900, 100, 1)$, $\boldsymbol{\lambda_2} = (600, 50, 1)$ e $\boldsymbol{\lambda_3} = (100, 40, 1)$, em que a concentração dos dados vai diminuindo, isto é: $\boldsymbol{\lambda_1} > \boldsymbol{\lambda_2} > \boldsymbol{\lambda_3}$. Foi considerado 100 réplicas de Monte Carlo. Em cenários de alta concentração dos dados (baixa variabilidade), as observações tendem a se agrupar em torno de valores centrais ou regiões específicas do espaço de dados, com pouca dispersão entre elas. Essa característica indica que os dados apresentam um grau elevado de homogeneidade, o que pode simplificar algumas análises ao reduzir a necessidade de lidar com outliers ou padrões altamente complexos.

Um dos grupos das amostras foi rotacionado para obter dois grupos com médias diferentes. E para tanto, o grau de rotação para as análises foi de 0.05π. Para garantir que as amostras geradas possuem diferenças estatísticas significativas, foram considerados os testes apresentados por (AMARAL; DRYDEN; WOOD, 2007) para diferenças de médias. Ou seja, iremos testar se as médias dos dois grupos são diferentes, se o p-valor do teste for menor que o nível de significância ($\alpha = 5\%$, adotado) rejeita-se a hipótese nula e conclui-se que os grupos possuem médias diferentes.

5.2.1 Resultado das simulações no espaço de pré-formas

As Tabelas 2, 3 e 4 apresentam os resultados obtidos pelos algoritmos propostos no espaço de pré-formas, considerando os diversos cenários de concentração dos dados e para amostras

de tamanho $n = 30$ e $n = 50$.

Tabela 2 – Médias dos índices de Rand corrigido para o cenário $\lambda = (900, 100, 1)$.

Método	$n = 30$	$n = 50$
k_{FP}	0,6356 (0,4764)	0,6286 (0,4814)
AEC_{FP}	0,5447 (0,1281)	0,5672 (0,096)
AEH_{FP}	0,5140 (0,1208)	0,5271 (0,1107)
$AESD_{FP}$	0,5053 (0,1123)	0,5241 (0,0901)
$AEkNN_{FP}$	0,4811 (0,1455)	0,5063 (0,1194)

Fonte: Elaborado pela autora.

A Tabela 2 apresenta as médias dos índices de Rand corrigido (com o respectivo desvio padrão entre parênteses) para os algoritmos de agrupamento espectral e o algoritmo k-means para dados de formas planas considerando o cenário $\lambda = (900, 100, 1)$. Os resultados para o método k_{FP} mostram desempenho semelhante entre os tamanhos amostrais, mas com alta variabilidade quando comparado aos métodos espectrais. Os algoritmos espectrais adaptados para formas planas AEC_{FP}, AEH_{FP}, $AESD_{FP}$ e $AEkNN_{FP}$ apresentaram desempenho consistente entre os dois tamanhos amostrais, com variações nas médias e desvios padrão que sugerem maior estabilidade e desempenho aprimorado para $n = 50$. O método k_{FP} destacou-se com os maiores índices médios, ainda que os métodos espectrais tenham mostrado desempenhos competitivos. O método $AEkNN_{FP}$ apresentou os menores valores médios dentre os algoritmos espectrais propostos.

Tabela 3 – Médias dos índices de Rand corrigido dos algoritmos para o cenário $\lambda = (600, 50, 1)$.

Método	$n = 30$	$n = 50$
k_{FP}	0,0387 (0,113)	0,023 (0,169)
AEC_{FP}	0,2748 (0,121)	0,2972 (0,085)
AEH_{FP}	0,2723 (0,119)	0,2847 (0,1086)
$AESD_{FP}$	0,2824 (0,112)	0,2986 (0,1051)
$AEkNN_{FP}$	0,2918 (0,1339)	0,3177 (0,1138)

Fonte: Elaborado pela autora.

Os resultados dos índices de Rand corrigido para diferentes algoritmos aplicados ao cenário $\lambda = (600, 50, 1)$, em que a concentração dos dados é menor que o cenário anterior são mostrados na Tabela 3. O método k_{FP} obteve os menores valores médios, indicando baixa consistência e alta variabilidade. Já os métodos espectrais adaptados para o estudo de formas planas apresentaram valores médios melhores em comparação com o método k_{FP}. Entre os

métodos espectrais, o método $AEkNN_{FP}$ se destaca com o melhor desempenho geral entre os métodos espectrais.

Ao compararmos os resultados das Tabelas 2 e 3 observa-se que os índices de Rand corrigido diminuíram para todos os métodos no cenário $\lambda = (600, 50, 1)$, indicando pior desempenho dos algoritmos em um cenário de menor concentração de dados. No entanto, métodos espectrais mantiveram melhor desempenho relativo em comparação ao método k-means adaptado para formas planas. Isso já era esperado, já que o método k-means apresenta bons desempenhos em cenários de alta concentração de dados (baixa variabilidade), pois esse contexto favorece a convergência para o ótimo global.

Tabela 4 – Médias dos índices de Rand corrigido dos algoritmos para o cenário $\lambda = (100, 40, 1)$.

Método	$n = 30$	$n = 50$
k_{FP}	0,0329 (0,101)	0,0526 (0,121)
AEC_{FP}	0,1012 (0,102)	0,1590 (0,075)
AEH_{FP}	0,089 (0,085)	0,1196 (0,079)
$AESD_{FP}$	0,1289 (0,098)	0,1535 (0,088)
$AEkNN_{FP}$	0,1473 (0,103)	0,1597 (0,084)

Fonte: Elaborado pela autora.

A Tabela 4 apresenta os resultados dos índices de Rand corrigido para diferentes algoritmos no cenário $\lambda = (100, 40, 1)$. O método k_{FP} continua com os menores valores médios entre os algoritmos. Esses resultados indicam fraco desempenho e alta variabilidade para o método k_{FP} em cenários em que a concentração dos dados vai diminuindo e sua variabilidade aumentando. Comparado aos cenários anteriores, o desempenho geral no cenário $\lambda = (100, 40, 1)$ é o mais baixo, indicando que a menor concentração e maior dispersão dos dados impactaram negativamente todos os métodos. Contudo, os métodos espectrais continuam se destacando como os melhores algoritmos em termos de robustez, mesmo em condições desafiadoras. De modo geral, os métodos espectrais $AEkNN_{FP}$ e $AESD_{FP}$ mostram maior consistência e adaptabilidade, enquanto que o k_{FP} apresenta desempenho fraco em cenários de menor concentração.

5.3 AVALIAÇÃO NUMÉRICA - DADOS REAIS

Além dos conjuntos de dados simulados, sete conjuntos de dados reais de formas de objetos foram utilizados nesta pesquisa: Crânios de Gorilas, Crânios de Orangotangos, Crânios de

Chimpanzés, Crânios de macacos, Cérebro de adultos saudáveis, Vértebra de Ratos e Cérebros de esquifrênicos e não esquizofrênicos. Esses conjuntos de dados são clássicos da literatura da análise estatística de formas. Além do interesse em realizar análise de agrupamento com o intuito de agrupar os dados em grupos, em que os objetos dentro de um mesmo grupo sejam mais similares entre si, por vezes queremos investigar como a forma muda durante o crescimento ou evolução, como a forma está relacionada ao tamanho, como a forma é afetado por doenças, como a forma está relacionada a outras covariáveis, como gênero, idade ou condições ambientais e como descrever a variabilidade da forma. A seguir, faremos uma breve descrição dos conjuntos de dados utilizados para validar os métodos propostos. Os conjuntos de dados estão disponíveis para acesso via pacotes shapes, e as informações descritas abaixo podem ser acessadas em (DRYDEN, 2023)

1. **Crânios de Gorilas:** O conjunto de dados Crânios de Gorilas pode ser acessado via pacote shapes. Contém 59 objetos, sendo 29 de gorilas machos e 30 de gorilas fêmeas, que foram obtidos em uma investigação para avaliar as diferenças entre os sexos cranianos de gorilas adultos (DRYDEN; MARDIA, 2016). Oito marcos anatômicos localizados por um biólogo especialista foram escolhidos para cada crânio. Nesse contexto, um pesquisador poderia estar interessadoem saber se existem diferenças de forma entre os sexos nas regiões cerebrais, ou ainda, quanto a descrições geométricas da diferença de forma, e como a forma se relaciona com o tamanho e outras covariáveis.

2. **Crânios de Orangotangos:** Nesta análise, são considerados 8 marcos em duas dimensões de 60 orangotangos, sendo 30 fêmeas e 30 machos. No contexto de análise de agrupamento queremos separar em grupos distintos os orangotangos machos dos orangotandos fêmeas. Essa separação poderia ajudar em estudos relacionados gênero das espécies em Biologia.

3. **Crânios de Chimpanzés:** O conjunto de dados dessa aplicação refere-se a dados de crânios de grandes primatas, como chimpanzés. São investigados crânios de 26 chimpanzés fêmeas e 28 chimpanzés machos. Foram considerados 8 marcos em duas dimensões para análise do crânio de chimpanzés.

4. **Crânios de Macacos:** Busca-se investigar sobre as diferenças de gênero a partir de registros cranianos de uma espécie de macaco `Macaca fasciularis`. Foram obtidas amostras aleatórias de 9 crânios de machos e 9 crânios de fêmeas. Um subconjunto de

sete marcos anatômicos foi localizado em cada crânio dos macacos e as coordenadas tridimensionais de cada ponto foram registradas. A análise utilizando os métodos para dados de formas planas, considerou os dados em duas dimensões.

5. **Cérebros de adultos saudáveis:** Este conjunto de dados considera 24 marcos localizados em 58 cérebros adultos saudáveis, em que os grupos são definidos pelo gênero. A análise utilizando os métodos para dados de formas planas, considerou os dados em duas dimensões. Nesse cenário, o agrupamento poderia ser útil em segmentar os grupos para avaliar como o gênero é afetado por presença ou não de alguma doença.

6. **Vértebras de Ratos:** Num experimento para avaliar os efeitos da seleção para o peso corporal na forma de vértebras de camundongos, foram obtidos três grupos de camundongos: Controle, Grande e Pequeno. O grupo Controle contém camundongos não selecionados, o grupo Grande contém camundongos selecionados a cada geração de acordo com o maior peso corporal e o grupo Pequeno foi selecionado para peso corporal menor. Seis marcos anatômicos localizados em 30 ossos controles, 23 ossos grandes e 23 ossos pequenos. O objetivo poderia ser avaliar se existe uma diferença de tamanho e forma entre os três grupos e fornecer descrições de quaisquer diferenças.

7. **Cérebros de pacientes esquizofrênicos e não esquizofrênicos:** Esse conjunto de dados corresponde a uma amostra de 28 configurações contendo 13 marcos anatômicos extraídos de imagens de ressonância magnética de cérebros de 14 pacientes saudáveis (controle) e 14 pacientes com esquizofrenia. Nesse contexto, além do interesse em realizar uma análise de agrupamento afim de separar pacientes esquizofrênicos e não esquizofrênicos, seria interessante estudar quaisquer diferenças de forma no cérebro entre os dois grupos, seja na forma média ou na variabilidade de forma. Havendo diferenças de forma entre os dois grupos, então isso deverá permitir aos investigadores obter uma maior compreensão sobre a condição.

Na Tabela 5 é apresentado uma análise descritiva dos dados reais considerados no estudo. A maioria dos dados foram divididos em dois grupos, exceto para o caso do estudo com Vértebras de Ratos. A maior partes das divisões dos grupos foi feita considerando o gênero dos objetos envolvidos. A mediana do número de marcos anatômicos utilizados foi igual a 8 e, o maior e menor tamanho de amostra consideradas foi quanto ao conjunto de dados sobre Vértebras de Ratos e Crânios de Macacos, respectivamente.

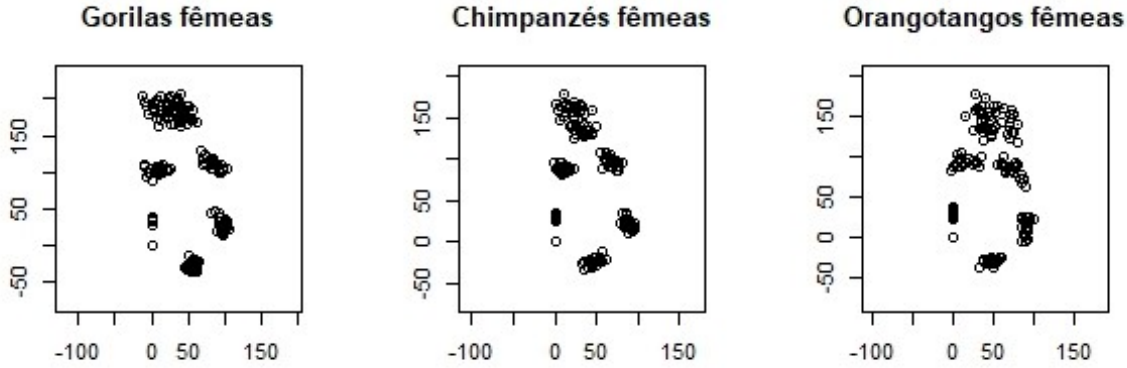
Tabela 5 – Análise descritiva dos dados reais.

Dados Reais	**Tamanho da amostra (n)**	**Grupos**	**Marcos anatômicos**
Crânios de Gorilas	59	2	8
Crânios de Orangotangos	60	2	8
Crânios de Chimpanzés	54	2	8
Crânios de Macacos	18	2	7
Cérebros de adultos saudáveis	58	2	24
Vértebra de Ratos	76	3	6
Cérebros de pacientes esquizofrênicos e não esquizofrênicos	28	2	13

Fonte: Elaborado pela autora.

Nas Figuras 4 e 5 são apresentados os seis marcos anatômicos dos Crânios de Gorilas, Chimpanzés e Orangotangos fêmeas e machos, respectivamente. É possível notar diferença de tamanho e formato entre os gêneros.

Figura 4 – Marcos anatômicos de Gorilas, Chimpanzés e Orangotangos fêmeas.



Diferenças quanto ao formato das vértebras dos três tipos de ratos: ratos Controle (30), ratos Grandes (23) e ratos Pequenos (23), a partir da visualização dos seis marcos anatômicos é apresentado na Figura 6.

Figura 5 – Marcos anatômicos de Gorilas, Chimpanzés e Orangotangos machos.

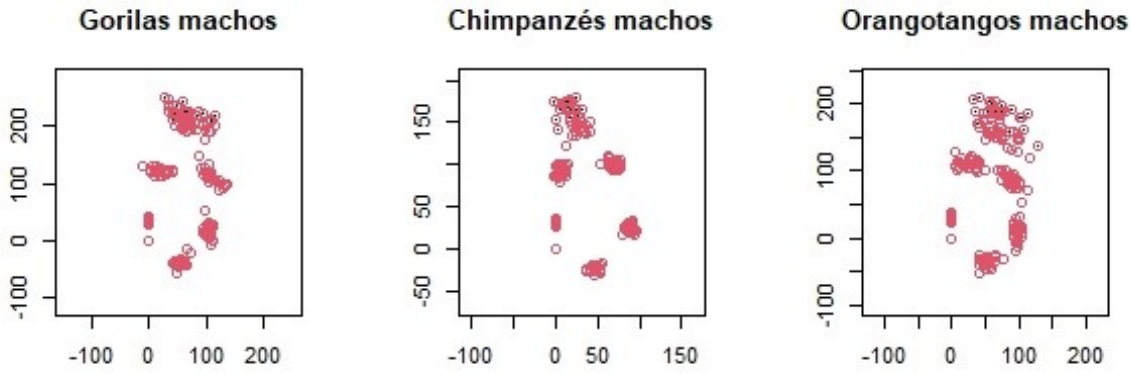
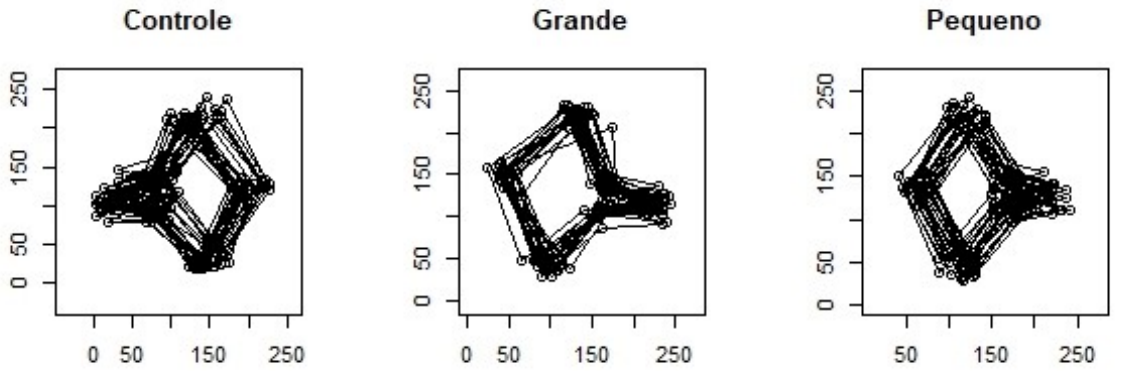


Figura 6 – Marcos anatômicos de vértebras de ratos para o grupo Controle, Grande e Pequeno.



5.3.1 Resultado dos dados reais

A Tabela 6 apresenta os índices de Rand corrigido obtidos pelos algoritmos propostos aplicados a conjuntos de dados reais.

A partir dos resultados apresentados na Tabela 6 observa-se que o método espectral AEC_{FP} destacou-se como o mais consistente, apresentando os maiores índices em cinco dos sete conjuntos de dados. Isso sugere que ele é mais robusto para diferentes tipos de dados reais. O método $AESD_{FP}$ também teve bom desempenho em alguns casos, como nos Crânios de Gorilas e Orangotangos. Todos os métodos espectrais apresentaram bom desempenho diante dos dados de Crânios de Gorilas, com destaque para o $AESD_{FP}$ que alcançou o maior índice de Rand corrigido (0,803); os demais métodos espectrais AEC_{FP}, AEH_{FP} e $AEkNN_{FP}$,

também tiveram valores elevados (0,743) e superiores ao k_{FP}.

Para os dados de Crânios de Orangotangos, o método k_{FP} se destacou com um índice de 0,657, seguido por $AESD_{FP}$ (0,597). Métodos como AEC_{FP} e AEH_{FP} apresentaram índices moderados (0,486), enquanto o $AEkNN_{FP}$ teve o menor desempenho (0,339). Comparando os resultados para métodos espectrais AEC_{FP} e AEH_{FP}, observamos que o algoritmo AEH_{FP} apresentou valores iguais ou inferiores aos de AEC_{FP}. Quando diante de amostras grandes esses métodos se equivalem, em amostras menores (Cérebros de adultos saudáveis e Crânios de Macacos) o método AEC_{FP} apresenta resultados vantajosos.

Tabela 6 – Médias dos índices de Rand corrigido obtido a partir dos algoritmos considerados para dados reais.

Dados/ Método	k_{FP}	AEC_{FP}	AEH_{FP}	$AESD_{FP}$	$AEkNN_{FP}$
Crânios de Gorilas	0,684	0,743	0,743	**0,803**	0,743
Crânios de Macacos	-0,003	**0,412**	-0,021	-0,055	0,055
Crânios de Chimpanzés	0,071	**0,182**	**0,182**	**0,149**	0,070
Cérebros de adultos saudáveis	**0,103**	**0,103**	**0,103**	0,080	0,041
Crânios de Orangotango	**0,657**	0,486	0,486	**0,5974**	0,339
Cérebros de pacientes Esquizofrênicos e não escrizofrênicos	-0,016	**0,857**	0,009	0,009	-0,014
Vértebra de Ratos	0,268	**0,373**	0,151	0,128	0,173

Fonte: Elaborado pela autora.

A Tabela 7 apresenta os índices de Rand corrigido obtidos a partir de métodos de agrupamento hierárquico aglomerativo utilizados quando aplicado o método AEH_{FP}, considerando diferentes estratégias de ligação (completa, simples, média e Ward) aplicadas a dados reais. Os índices de Rand corrigido variam dependendo do conjunto de dados e do método utilizado, refletindo as particularidades dos dados e a adequação das estratégias de agrupamento.

Tabela 7 – Índices de rand para dados reais considerando os métodos de agrupamento hierárquico aglomerativo.

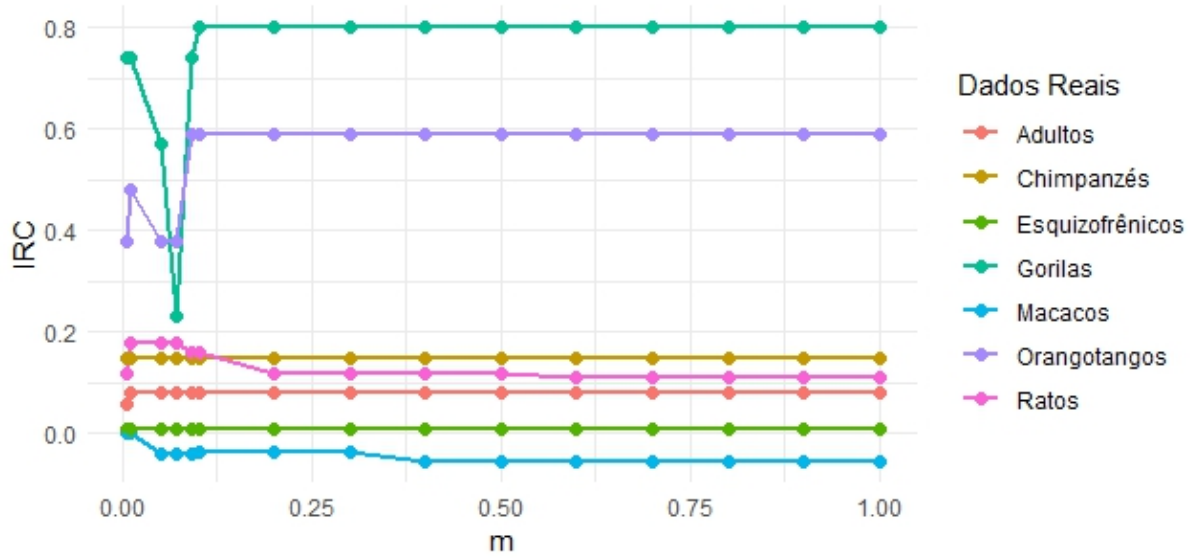
Dados/ Método	Ligação Completa	Ligação Simples	Ligação Média	Ligação Ward
Crânios de Gorilas	0,743	0,743	0,743	0,743
Crânios de Macacos	-0,021	0,027	-0,021	-0,021
Crânios de Chimpanzés	0,123	0,072	0,183	0,051
Cérebros de adultos saudáveis	0,014	0,021	0,103	0,1032
Crânios de Orangotango	0,434	0,486	0,486	0,486
Cérebros de pacientes Esquizofrênicos e não escrizofrênicos	0,0094	0,0094	-0,016	0,0094
Vértebra de Ratos	0,1508	0,0126	0,146	0,146

Fonte: Elaborado pela autora.

Para dados de Crânios de Gorilas todos os métodos de ligação apresentaram o mesmo índice de Rand corrigido (0,743); a ligação média obteve o maior índice (0,183) para dados de Crânios de Chimpanzés; para dados de Cérebros de Adultos Saudáveis, os métodos de ligação média e Ward apresentaram os melhores índices (0,103); a ligação completa apresentou menor índice (0,434) para dados de Crânios de Orangotango. Para os dados de Vértebra de Ratos, a ligação simples não é indicada.

Na Figura 7 apresenta os índices de Rand corrigido (IRC) em função do parâmetro m, para diferentes conjuntos de dados reais. Cada linha representa a evolução do IRC para um conjunto de dados específico à medida que o valor de m varia.

Figura 7 – Variação do Índice de Rand corrigido para os dados reais em função de m, utilizando o algoritmo $AESD_{FP}$.



O gráfico demonstra que o impacto do parâmetro m no desempenho dos agrupamentos varia entre os conjuntos de dados, principalmente quando m é muito próximo de zero. Dados como Crânios de Gorilas e Crânios de Orangotangos mostram uma melhora inicial no IRC antes de estabilizar em valores altos, indicando que o parâmetro m pode ser ajustado para melhorar a qualidade dos agrupamentos. Em contrapartida, dados como Cérebros de Adultos saudáveis, Crânios de Chimpanzés e Cérebros de pacientes Esquizofrênicos e não esquizofrênicos têm IRC constantes, sugerindo pouca sensibilidade ao parâmetro m. De modo geral, há variações nos resultados do índice de Rand corrigido em valores de m no intervalo entre 0 e 0.1. Nesse intervalo os resultados do algoritmo espectral $AESD_{FP}$ são equivalentes ao AEC_{FP} para os dados de Crânios de Gorilas e Crânios de Orangotangos.

6 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

6.1 CONSIDERAÇÕES FINAIS

Nesta dissertação foram apresentados métodos de agrupamentos baseados em teoria dos grafos adaptados para trabalhar com dados da área de análise estatística de formas. Consideramos, inicialmente, que estudos sobre algoritmos baseados em teoria de grafos para agrupar dados de forma plana é um tema em aberto, além disso, buscamos acrescer estudos sobre algoritmos de agrupamento espectral, visto sua vasta aplicabilidade. Dessa forma, propomos novas opções de algoritmos de agrupamento para analisar formas planas de objetos e, assim, contribuir com alternativas de agrupamento de dados para a literatura de área análise estatística de formas e agrupamento espectral.

Com os avanços tecnológicos, a análise de formas tornou-se cada vez mais relevante para lidar com dados geométricos. As formas planas são representadas por coordenadas baseadas em marcos anatômicos, localizadas em um espaço não Euclidiano, o que exige métodos específicos de análise. Como algoritmos tradicionais de agrupamento operam em espaços Euclidianos, esta dissertação propõe adaptações que permitem lidar com a natureza não Euclidiana dos dados de forma.

Foram apresentados quatro algoritmos de agrupamento espectral voltados para formas planas. O primeiro, AEC_{FP}, constrói a matriz de similaridade utilizando a função kernel gaussiana e realiza o agrupamento a partir dos k autovetores da matriz Laplaciana, aplicando o algoritmo k-means. O segundo, AEH_{FP} segue a mesma estrutura, mas utiliza o método hierárquico para o agrupamento. O terceiro, $AESD_{FP}$ introduz um critério de dissimilaridade no corte normalizado do grafo. Por fim, o quarto, $AEkNN_{FP}$, constrói a matriz de similaridade com base nos k-vizinhos mais próximos e realiza o agrupamento também por meio do k-means. O algoritmo k-means para formas planas (AMARAL et al., 2010) foi utilizado como referência comparativa para os métodos de agrupamento espectral propostos.

Os algoritmos propostos foram testados em diversos cenários de simulação no espaço de pré-formas e com dados reais disponíveis na literatura, em que a avaliação do desempenho dos algoritmos foi realizada através do índice de Rand corrigido. Simulações da distribuição Bingham que possui suporte na esfera unitária e é bastante utilizada em análise de formas planas também foi estudada. Foram considerados cenários de alta concentração $\lambda = (900, 100, 1)$, concentração intermediária $\lambda = (600, 50, 1)$ e baixa concentração $\lambda = (100, 40, 1)$, para

$n = 30$ e $n = 50$. No cenário de alta concentração o algoritmo k-means para formas planas apresentou os maiores índices de Rand corrigido, embora os algoritmos espectrais adaptados mostraram resultados competitivos. Nos cenários de concentração intermediária e baixa, os métodos espectrais mantiveram melhor desempenho relativo em comparação com o método k-means adaptado para formas planas.

Além dos conjuntos de dados simulados, sete conjuntos de dados reais de formas de objetos foram utilizados nesta pesquisa: Crânios de Gorilas, Crânios de Orangotangos, Crânios de Chimpanzés, Crânios de macacos, Cérebro de adultos saudáveis, Vértebra de Ratos e Cérebros de esquifrênicos e não esquizofrênicos. Observa-se que o método espectral AEC_{FP} destacou-se como o mais consistente, apresentando os maiores índices de Rand corrigido em cinco dos sete conjuntos de dados. O método espectral $AESD_{FP}$ também teve bom desempenho em alguns casos, como nos Crânios de Gorilas e Orangotangos. De modo geral, os algoritmos espectrais propostos demonstraram resultados encorajadores.

Por fim, este trabalho contribuiu para a literatura teórica dos métodos de agrupamento para análise estatística de formas, desenvolvendo algoritmos espectrais adaptados para dados de forma plana. Colaborou também com estudos sobre métodos de agrupamento, em particular, exploramos novas aplicações para agrupamento espectral. Esperamos que essa pesquisa possa motivar trabalhos futuros sobre o tema e que os algoritmos aqui desenvolvidos possam ser utilizados em diversas aplicações reais.

6.2 TRABALHOS FUTUROS

Estudos sobre algoritmos baseados em teoria de grafos para agrupar dados de forma é um tema em aberto. Esta pesquisa buscou, inicialmente, investigar potencialidades quanto ao uso de agrupamento espectral para análise de formas planas. Devido a sua capacidade na detecção de estruturas complexas de dados, o agrupamento espectral tem motivado diversas pesquisas recentes, sendo novas versões mais robustas do algoritmo propostas anualmente. Assim, há muitas extensões de algoritmos espectrais que podem ser adaptadas para análise de formas planas, sendo algumas dessas pesquisas exploradas no final do capítulo 2 quando apresentamos uma breve revisão da literatura sobre o tema. Dessa forma, algumas pesquisas futuras podem explorar os seguintes pontos:

1. Desenvolver novos algoritmos de agrupamento espectral e suas versões para dados de

forma com foco na construção de matriz de Similaridade ou matriz Laplaciana mais eficientes;

2. Adaptar versões mais recentes e robustas do algoritmo espectral para dados de formas planas e comparar o desempenho dos algoritmos propostos em termos de eficiência com os algoritmos adaptados nessa pesquisa;

3. Comparar o desempenho dos algoritmos espectrais para formas planas com outros algoritmos de agrupamento adaptados para o mesmo tipo de dados;

4. Nesse trabalho, utilizamos a função do kernel gaussiano para construir a matriz de similaridade em três dos quatro algoritmos propostos. No entanto, a função kernel gaussiano é sensível a ruídos e valores discrepantes. Sob essa ótica, trabalhos podem ser desenvolvidos com o intuito de construir melhores matrizes de similaridade, por exemplo, utilizando o conceito de multi-kernel como desenvolvido por (MONNEY et al., 2020b), adaptando a dados de formas planas e comparando com os resultados dessa pesquisa;

5. Um dos algoritmos propostos utilizou o método k-NN para construir a matriz de similaridade dos dados. No entanto, o método k-NN mútuo é mais eficiente segundo (TAN; ZHANG; WU, 2020). Nesse sentido, estudo sobre algoritmos k-NN e suas variantes adaptadas para dados de formas planas poderia gerar resultados motivadores;

6. Agrupamento espectral associado com agrupamento de conjuntos tem sido pauta em diversas pesquisas recentes. Sob essa perspectiva, desenvolver novos algoritmos e adaptá-los para dados de formas planas. Esses algoritmos podem apresentar resultados mais robustos do que os obtido nessa pesquisa;

7. Considerar simulações para coordenadas no espaço tangente e comparar com os resultados para coordenadas de Kendall.

REFERÊNCIAS

AGGARWAL, C. C.; REDDY, C. K. Data clustering. *Algorithms and applications. Chapman&Hall/CRC Data mining and Knowledge Discovery series, Londra*, 2014.

ALSHAMMARI, M.; STAVRAKAKIS, J.; TAKATSUKA, M. Refining a k-nearest neighbor graph for a computationally efficient spectral clustering. *Pattern Recognition*, Elsevier, v. 114, p. 107869, 2021.

ALSHAMMARI, M.; TAKATSUKA, M. Approximate spectral clustering with eigenvector selection and self-tuned k. *Pattern Recognition Letters*, Elsevier, v. 122, p. 31–37, 2019.

ALSHAMMARI, M.; TAKATSUKA, M. Approximate spectral clustering with eigenvector selection and self-tuned k. *Pattern Recognition Letters*, Elsevier, v. 122, p. 31–37, 2019.

AMARAL, G. A.; DRYDEN, I.; WOOD, A. T. A. Pivotal bootstrap methods for k-sample problems in directional statistics and shape analysis. *Journal of the American Statistical Association*, Taylor & Francis, v. 102, n. 478, p. 695–707, 2007.

AMARAL, G. J.; DORE, L. H.; LESSA, R. P.; STOSIC, B. K-means algorithm in statistical shape analysis. *Communications in Statistics—Simulation and Computation®*, Taylor & Francis, v. 39, n. 5, p. 1016–1026, 2010.

AMORIM, R. C. D. Constrained clustering with minkowski weighted k-means. In: IEEE. *2012 IEEE 13th International Symposium on Computational Intelligence and Informatics (CINTI)*. [S.l.], 2012. p. 13–17.

BABICHEV, S.; YASINSKA-DAMRI, L.; LIAKH, I. A hybrid model of cancer diseases diagnosis based on gene expression data with joint use of data mining methods and machine learning techniques. *Applied Sciences*, MDPI, v. 13, n. 10, p. 6022, 2023.

BAI, L.; QI, M.; LIANG, J. Spectral clustering with robust self-learning constraints. *Artificial Intelligence*, Elsevier, v. 320, p. 103924, 2023.

BARIONI, M. C. N.; RAZENTE, H.; MARCELINO, A. M.; TRAINA, A. J.; JR, C. T. Open issues for partitioning clustering methods: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 4, n. 3, p. 161–177, 2014.

BOOKSTEIN, F. L. Size and shape spaces for landmark data in two dimensions. *Statistical science*, Institute of Mathematical Statistics, v. 1, n. 2, p. 181–222, 1986.

BRUSE, J. L.; ZULUAGA, M. A.; KHUSHNOOD, A.; MCLEOD, K.; NTSINJANA, H. N.; HSIA, T.-Y.; SERMESANT, M.; PENNEC, X.; TAYLOR, A. M.; SCHIEVANO, S. Detecting clinically meaningful shape clusters in medical image data: metrics analysis for hierarchical clustering applied to healthy and pathological aortic arches. *IEEE Transactions on Biomedical Engineering*, IEEE, v. 64, n. 10, p. 2373–2383, 2017.

BUTLER, S.; CHUNG, F. et al. Spectral graph theory. *Handbook of linear algebra*, Florida: CRC Press, p. 47, 2006.

CAI, D.; CHEN, X. Large scale spectral clustering via landmark-based sparse representation. *IEEE transactions on cybernetics*, IEEE, v. 45, n. 8, p. 1669–1680, 2014.

CAO, J.; CHEN, P.; ZHENG, Y.; DAI, Q. A max-flow-based similarity measure for spectral clustering. *Etri Journal*, Wiley Online Library, v. 35, n. 2, p. 311–320, 2013.

CHAKLADAR, D. D.; SAMANTA, D.; ROY, P. P. Multimodal deep sparse subspace clustering for multiple stimuli-based cognitive task. In: IEEE. *2022 26th International Conference on Pattern Recognition (ICPR)*. [S.l.], 2022. p. 1098–1104.

CHALLA, A.; DANDA, S.; SAGAR, B. D.; NAJMAN, L. Power spectral clustering. *Journal of Mathematical Imaging and Vision*, Springer, v. 62, n. 9, p. 1195–1213, 2020.

CHEN, C.; GUO, J. A general approach for handwritten digits segmentation using spectral clustering. In: IEEE. *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*. [S.l.], 2017. v. 1, p. 547–552.

CHUNG, F. R. *Spectral graph theory*. [S.l.]: American Mathematical Soc., 1997. v. 92.

DEFAYS, D. An efficient algorithm for a complete link method. *The computer journal*, Oxford University Press, v. 20, n. 4, p. 364–366, 1977.

DENG, J.; HUANG, D.; DING, Y.; ZHU, Y.; JING, B.; ZHANG, B. Subsampling spectral clustering for stochastic block models in large-scale networks. *Computational Statistics & Data Analysis*, Elsevier, v. 189, p. 107835, 2024.

DHILLON, I. S.; GUAN, Y.; KULIS, B. Weighted graph cuts without eigenvectors a multilevel approach. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 29, n. 11, p. 1944–1957, 2007.

DING, C. H.; HE, X.; ZHA, H.; GU, M.; SIMON, H. D. A min-max cut algorithm for graph partitioning and data clustering. In: IEEE. *Proceedings 2001 IEEE international conference on data mining*. [S.l.], 2001. p. 107–114.

DING, L.; LI, C.; JIN, D.; DING, S. Survey of spectral clustering based on graph theory. *Pattern Recognition*, Elsevier, p. 110366, 2024.

DONATH, W. E.; HOFFMAN, A. J. Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, IBM, v. 17, n. 5, p. 420–425, 1973.

DRYDEN, I.; MARDIA, K. *Statistical shape analysis with applications in R*. [S.l.: s.n.], 2016. Second edition.

DRYDEN, I. L. *shapes: Statistical Shape Analysis*. [S.l.], 2023. R package version 1.2.6. Disponível em: <https://cran.r-project.org/package=shapes>.

DU, L.; ZHOU, P.; SHI, L.; WANG, H.; FAN, M.; WANG, W.; SHEN, Y.-D. Robust multiple kernel k-means using l21-norm. In: *Twenty-fourth international joint conference on artificial intelligence*. [S.l.: s.n.], 2015.

EMIROV, N.; CHENG, C.; SUN, Q.; QU, Z. Distributed algorithms to determine eigenvectors of matrices on spatially distributed networks. *Signal Processing*, Elsevier, v. 196, p. 108530, 2022.

EVERITT, B.; LANDAU, S.; LEESE, M.; STAHL, D. Cluster analysis. Wiley, 2011.

FANG, Y.; WANG, J. Selection of the number of clusters via the bootstrap method. *Computational Statistics & Data Analysis*, Elsevier, v. 56, n. 3, p. 468–477, 2012.

FIEDLER, M. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, Institute of Mathematics, Academy of Sciences of the Czech Republic, v. 23, n. 2, p. 298–305, 1973.

FIX, E. *Discriminatory analysis: nonparametric discrimination, consistency properties*. [S.l.]: USAF school of Aviation Medicine, 1985. v. 1.

GAN, G.; NG, M. K.-P. K-means clustering with outlier removal. *Pattern Recognition Letters*, Elsevier, v. 90, p. 8–14, 2017.

GEORGESCU, V. Clustering of fuzzy shapes by integrating procrustean metrics and full mean shape estimation into k-means algorithm. In: CITESEER. *IFSA/EUSFLAT Conf.* [S.l.], 2009. p. 1679–1684.

GUPTA, A.; GUPTA, A.; MISHRA, A. Research paper on cluster techniques of data variations. *International Journal of Advance Technology & Engineering Research*, v. 1, n. 1, p. 39–47, 2011.

HAGEN, L.; KAHNG, A. B. New spectral methods for ratio cut partitioning and clustering. *IEEE transactions on computer-aided design of integrated circuits and systems*, IEEE, v. 11, n. 9, p. 1074–1085, 1992.

HESS, S.; DUIVESTEIJN, W.; HONYSZ, P.; MORIK, K. The spectacl of nonconvex clustering: A spectral approach to density-based clustering. In: *Proceedings of the AAAI conference on artificial intelligence*. [S.l.: s.n.], 2019. v. 33, n. 01, p. 3788–3795.

HOFMEYR, D. P. Clustering by minimum cut hyperplanes. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 39, n. 8, p. 1547–1560, 2016.

HUANG, C.; STYNER, M.; ZHU, H. Clustering high-dimensional landmark-based two-dimensional shape data. *Journal of the American Statistical Association*, Taylor & Francis, v. 110, n. 511, p. 946–961, 2015.

HUANG, J. Z.; NG, M. K.; RONG, H.; LI, Z. Automated variable weighting in k-means type clustering. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 27, n. 5, p. 657–668, 2005.

HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of classification*, Springer, v. 2, p. 193–218, 1985.

JABI, M. e. a. Deep clustering: On the link between discriminative models and k-means. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 43, n. 6, p. 1887–1896, 2019.

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

JANANI, R.; VIJAYARANI, S. Text document clustering using spectral clustering algorithm with particle swarm optimization. *Expert Systems with Applications*, Elsevier, v. 134, p. 192–200, 2019.

JAYASUMANA, S.; SALZMANN, M.; LI, H.; HARANDI, M. A framework for shape analysis via hilbert space embedding. In: *Proceedings of the IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2013. p. 1249–1256.

JI, M.; RAO, H.; LI, Z.; ZHU, J.; WANG, N. Partial multi-view clustering based on sparse embedding framework. *IEEE Access*, IEEE, v. 7, p. 29332–29343, 2019.

JOHN, C. R.; WATSON, D.; BARNES, M. R.; PITZALIS, C.; LEWIS, M. J. Spectrum: fast density-aware spectral clustering for single and multi-omic data. *Bioinformatics*, Oxford University Press, v. 36, n. 4, p. 1159–1166, 2020.

JOTHI, R.; MOHANTY, S. K.; OJHA, A. Dk-means: a deterministic k-means clustering algorithm for gene expression analysis. *Pattern Analysis and Applications*, Springer, v. 22, p. 649–667, 2019.

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

KADAVANKANDY, A.; COUILLET, R. Asymptotic gaussian fluctuations of spectral clustering eigenvectors. In: IEEE. *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. [S.l.], 2019. p. 694–698.

KANAAN-IZQUIERDO, S.; ZIYATDINOV, A.; PERERA-LLUNA, A. Multiview and multifeature spectral clustering using common eigenvectors. *Pattern Recognition Letters*, Elsevier, v. 102, p. 30–36, 2018.

KANG, Z.; WEN, L.; CHEN, W.; XU, Z. Low-rank kernel learning for graph-based clustering. *Knowledge-Based Systems*, Elsevier, v. 163, p. 510–517, 2019.

KENDALL, D. G. The diffusion of shape. *Advances in applied probability*, Cambridge University Press, v. 9, n. 3, p. 428–430, 1977.

KENDALL, D. G. Shape manifolds, procrustean metrics, and complex projective spaces. *Bulletin of the London mathematical society*, Wiley Online Library, v. 16, n. 2, p. 81–121, 1984.

KENT, J. T. The complex bingham distribution and shape analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, Oxford University Press, v. 56, n. 2, p. 285–299, 1994.

KENT, J. T.; CONSTABLE, P. D.; ER, F. Simulation for the complex bingham distribution. *Statistics and Computing*, Springer, v. 14, p. 53–57, 2004.

KIM, J.-H.; CHOI, J.-H.; PARK, Y.-H.; LEUNG, C. K.-S.; NASRIDINOV, A. Knn-sc: novel spectral clustering algorithm using k-nearest neighbors. *IEEE Access*, IEEE, v. 9, p. 152616–152627, 2021.

KLASSEN, E.; SRIVASTAVA, A.; MIO, M.; JOSHI, S. H. Analysis of planar shapes using geodesic paths on shape spaces. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 26, n. 3, p. 372–383, 2004.

LAASSEM, B.; IDARROU, A.; BOUJLALEB, L.; IGGANE, M. A spectral method to detect community structure based on coulomb's matrix. *Social Network Analysis and Mining*, Springer, v. 13, n. 1, p. 3, 2022.

LAHMAR, I.; ZAIER, A.; YAHIA, M.; BOUALLEGUE, R. A new self adaptive fuzzy unsupervised clustering ensemble based on spectral clustering. In: IEEE. *2020 17th International Multi-Conference on Systems, Signals & Devices (SSD)*. [S.l.], 2020. p. 1–5.

LI, X.; WEI, T.; ZHAO, Y. Deep spectral clustering with constrained laplacian rank. *IEEE Transactions on Neural Networks and Learning Systems*, IEEE, v. 35, n. 5, p. 7102–7113, 2022.

LI, Z.; LIU, J.; CHEN, S.; TANG, X. Noise robust spectral clustering. In: IEEE. *2007 IEEE 11th International Conference on Computer Vision*. [S.l.], 2007. p. 1–8.

LIERDE, H. V.; CHOW, T. W.; CHEN, G. Scalable spectral clustering for overlapping community detection in large-scale networks. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 32, n. 4, p. 754–767, 2019.

LIU, H.; TAO, Z.; FU, Y. Partition level constrained clustering. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 40, n. 10, p. 2469–2483, 2017.

LIU, H. e. a. Spectral ensemble clustering via weighted k-means: Theoretical and practical evidence. *IEEE transactions on knowledge and data engineering*, IEEE, v. 29, n. 5, p. 1129–1143, 2017.

LIU L; CHEN, X. e. a. Hsc: A spectral clustering algorithm combined with hierarchical method. *Neural Network World*, n. 23, p. 499–521, 2013.

LUCIŃSKA, M.; WIERZCHOŃ, S. T. Spectral clustering based on k-nearest neighbor graph. In: SPRINGER. *Computer Information Systems and Industrial Management: 11th IFIP TC 8 International Conference, CISIM 2012, Venice, Italy, September 26-28, 2012. Proceedings 11*. [S.l.], 2012. p. 254–265.

LUO, D.; HUANG, H.; DING, C.; NIE, F. On the eigenvectors of p-laplacian. *Machine Learning*, Springer, v. 81, p. 37–51, 2010.

LUXBURG, U. V. A tutorial on spectral clustering. *Statistics and computing*, Springer, v. 17, p. 395–416, 2007.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*. [S.l.: s.n.], 1967.

MEILA, M.; SHI, J. Learning segmentation by random walks. *Advances in neural information processing systems*, v. 13, 2000.

MENON, R. R.; ASHOK, A.; ARYA, S. Document cluster analysis based on parameter tuning of spectral graphs. In: *Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2021*. [S.l.]: Springer, 2022. p. 401–413.

MINGOTI, S. A. Análise de dados através de métodos estatística multivariada: uma abordagem aplicada. In: *Análise de dados através de métodos estatística multivariada: uma abordagem aplicada*. [S.l.: s.n.], 2007. p. 295–295.

MOHAR, B. Some applications of laplace eigenvalues of graphs. In: *Graph symmetry: Algebraic methods and applications*. [S.l.]: Springer, 1997. p. 225–275.

MONNEY, A.; ZHAN, Y.; JIA, H.; BENUWA, B.-B. Co-regularized discriminative spectral clustering with adaptive similarity measure in dual-kernel space. *IEEE Access*, IEEE, v. 8, p. 46427–46439, 2020.

MONNEY, A.; ZHAN, Y.; JIANG, Z.; BENUWA, B.-B. A multi-kernel method of measuring adaptive similarity for spectral clustering. *Expert Systems with Applications*, Elsevier, v. 159, p. 113570, 2020.

MURTAGH, F. A survey of recent advances in hierarchical clustering algorithms. *The computer journal*, The British Computer Society, v. 26, n. 4, p. 354–359, 1983.

NABIL, M.; GOLALIZADEH, M. On clustering shape data. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 86, n. 15, p. 2995–3008, 2016.

NATALIANI, Y.; YANG, M.-S. Powered gaussian kernel spectral clustering. *Neural Computing and Applications*, Springer, v. 31, p. 557–572, 2019.

NG, A.; JORDAN, M.; WEISS, Y. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, v. 14, 2001.

NICA, B. Cut ratios and laplacian eigenvalues. *Linear Algebra and its Applications*, Elsevier, v. 593, p. 18–28, 2020.

NIE, F.; DING, C.; LUO, D.; HUANG, H. Improved minmax cut graph clustering with nonnegative relaxation. In: SPRINGER. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part II 21*. [S.l.], 2010. p. 451–466.

NIE, F.; WANG, X.; HUANG, H. Clustering and projected clustering with adaptive neighbors. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2014. p. 977–986.

OLIVEIRA, R. A. de. Algoritmos para determinação do número de grupos em estudos de formas planas. *Dissertação de Mestrado*, Universidade Federal de Pernambuco, 2016.

PENG, X.; ZHANG, L.; YI, Z. Scalable sparse subspace clustering. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2013. p. 430–437.

QI, R.; WU, J.; GUO, F.; XU, L.; ZOU, Q. A spectral clustering with self-weighted multiple kernel learning method for single-cell rna-seq data. *Briefings in Bioinformatics*, Oxford University Press, v. 22, n. 4, p. bbaa216, 2021.

QURESHI, M. N.; AHAMAD, M. V. An improved method for image segmentation using k-means clustering with neutrosophic logic. *Procedia computer science*, Elsevier, v. 132, p. 534–540, 2018.

R, C. T. R: A language and environment for statistical computing. In: . [S.l.: s.n.], 2024.

REBAGLIATI, N.; VERRI, A. Spectral clustering with more than k eigenvectors. *Neurocomputing*, Elsevier, v. 74, n. 9, p. 1391–1401, 2011.

RENGASAMY, S.; MURUGESAN, P. K-means–laplacian clustering revisited. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 107, p. 104535, 2022.

ROY, A. K.; BASU, T. Postimpact similarity: a similarity measure for effective grouping of unlabelled text using spectral clustering. *Knowledge and Information Systems*, Springer, v. 64, n. 3, p. 723–742, 2022.

SAADE, A.; KRZAKALA, F.; ZDEBOROVÁ, L. Spectral clustering of graphs with the bethe hessian. *Advances in neural information processing systems*, v. 27, 2014.

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.

SIBSON, R. Slink: an optimally efficient algorithm for the single-link cluster method. *The computer journal*, Oxford University Press, v. 16, n. 1, p. 30–34, 1973.

STEMMER, U. Locally private k-means clustering. *Journal of Machine Learning Research*, v. 22, n. 176, p. 1–30, 2021.

STOER, M.; WAGNER, F. A simple min-cut algorithm. *Journal of the ACM (JACM)*, ACM New York, NY, USA, v. 44, n. 4, p. 585–591, 1997.

SU, M.-C.; CHOU, C.-H. A modified version of the k-means algorithm with a distance based on cluster symmetry. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 23, n. 6, p. 674–680, 2001.

TAN, M.; ZHANG, S.; WU, L. Mutual knn based spectral clustering. *Neural Computing and Applications*, Springer, v. 32, p. 6435–6442, 2020.

TAO, X.; WANG, R.; CHANG, R.; LI, C.; LIU, R.; ZOU, J. Spectral clustering algorithm using density-sensitive distance measure with global and local consistencies. *Knowledge-Based Systems*, Elsevier, v. 170, p. 26–42, 2019.

TAŞDEMIR, K. A hybrid similarity measure for approximate spectral clustering of remote sensing images. In: IEEE. *2013 IEEE International Geoscience and Remote Sensing Symposium-IGARSS*. [S.l.], 2013. p. 3136–3139.

TAŞDEMIR, K.; YALÇIN, B.; YILDIRIM, I. Approximate spectral clustering with utilized similarity information using geodesic based hybrid distance measures. *Pattern Recognition*, Elsevier, v. 48, n. 4, p. 1465–1477, 2015.

TEPPER, M.; MUSÉ, P.; ALMANSA, A.; MEJAIL, M. Automatically finding clusters in normalized cuts. *Pattern Recognition*, Elsevier, v. 44, n. 7, p. 1372–1386, 2011.

TRILLOS, N. G.; SLEPCEV, D.; BRECHT, J. von; LAURENT, T.; BRESSON, X. Consistency of cheeger and ratio graph cuts. *arXiv preprint arXiv:1411.6590*, 2014.

VEENSTRA, P.; COOPER, C.; PHELPS, S. Spectral clustering using the knn-mst similarity graph. In: IEEE. *2016 8th Computer Science and Electronic Engineering (CEEC)*. [S.l.], 2016. p. 222–227.

VINUÉ, G.; SIMÓ, A.; ALEMANY, S. The k-means algorithm for 3d shapes with an application to apparel design. *Advances in Data Analysis and Classification*, Springer, v. 10, n. 1, p. 103–132, 2016.

VORA, A.; RAMAN, S. Iterative spectral clustering for unsupervised object localization. *Pattern Recognition Letters*, Elsevier, v. 106, p. 27–32, 2018.

WAGNER, D.; WAGNER, F. Between min cut and graph bisection. In: SPRINGER. *Mathematical Foundations of Computer Science 1993: 18th International Symposium, MFCS'93 Gdańsk, Poland, August 30–September 3, 1993 Proceedings 18*. [S.l.], 1993. p. 744–750.

WANG, B.; ZHANG, L.; WU, C.; LI, F.-z.; ZHANG, Z. Spectral clustering based on similarity and dissimilarity criterion. *Pattern Analysis and Applications*, Springer, v. 20, p. 495–506, 2017.

WANG, F.; DING, C.; LI, T. Integrated kl (k-means–laplacian) clustering: a new clustering approach by combining attribute data and pairwise relations. In: SIAM. *Proceedings of the 2009 SIAM International Conference on Data Mining*. [S.l.], 2009. p. 38–48.

WANG, J.; SU, X. An improved k-means clustering algorithm. In: IEEE. *2011 IEEE 3rd international conference on communication software and networks*. [S.l.], 2011. p. 44–46.

WANG, J.; WANG, J.; SONG, J.; XU, X.-S.; SHEN, H. T.; LI, S. Optimized cartesian k-means. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 27, n. 1, p. 180–192, 2014.

WANG, L.; DING, S.; JIA, H. An improvement of spectral clustering via message passing and density sensitive similarity. *IEEE access*, IEEE, v. 7, p. 101054–101062, 2019.

WANG, L.; DONG, M. Multi-level low-rank approximation-based spectral clustering for image segmentation. *Pattern Recognition Letters*, Elsevier, v. 33, n. 16, p. 2206–2215, 2012.

WANG, R.; CHEN, H.; LU, Y.; ZHANG, Q.; NIE, F.; LI, X. Discrete and balanced spectral clustering with scalability. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, 2023.

WANG, T.-S.; LIN, H.-T.; WANG, P. Weighted-spectral clustering algorithm for detecting community structures in complex networks. *Artificial Intelligence Review*, Springer, v. 47, p. 463–483, 2017.

WANG, X.; QIAN, B.; DAVIDSON, I. On constrained spectral clustering and its applications. *Data Mining and Knowledge Discovery*, Springer, v. 28, p. 1–30, 2014.

WANG, X.-D.; CHEN, R.-C.; YAN, F.; ZENG, Z.-Q.; HONG, C.-Q. Fast adaptive k-means subspace clustering for high-dimensional data. *IEEE Access*, IEEE, v. 7, p. 42639–42651, 2019.

WANG, Y.; JIANG, Y.; WU, Y.; ZHOU, Z.-H. Spectral clustering on multiple manifolds. *IEEE Transactions on Neural Networks*, IEEE, v. 22, n. 7, p. 1149–1161, 2011.

WEISS, Y. Segmentation using eigenvectors: a unifying view. In: IEEE. *Proceedings of the seventh IEEE international conference on computer vision*. [S.l.], 1999. v. 2, p. 975–982.

XIA, K.; GU, X.; ZHANG, Y. Oriented grouping-constrained spectral clustering for medical imaging segmentation. *Multimedia Systems*, Springer, v. 26, n. 1, p. 27–36, 2020.

XIA, S. e. a. Ball k k-means: Fast adaptive clustering with no bounds. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 44, n. 1, p. 87–99, 2020.

YANG, J.; YANG, X.; ZHOU, Z.-B.; LIU, Z.-Y. Graph matching based on fast normalized cut and multiplicative update mapping. *Pattern Recognition*, Elsevier, v. 122, p. 108228, 2022.

YE, W.; ZHOU, L.; SUN, X.; PLANT, C.; BÖHM, C. Attributed graph clustering with unimodal normalized cut. In: SPRINGER. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10*. [S.l.], 2017. p. 601–616.

YE, X.; SAKURAI, T. Spectral clustering using robust similarity measure based on closeness of shared nearest neighbors. In: IEEE. *2015 International joint conference on neural networks (IJCNN)*. [S.l.], 2015. p. 1–8.

YE, X.; SAKURAI, T. Spectral clustering with adaptive similarity measure in kernel space. *Intelligent Data Analysis*, IOS Press, v. 22, n. 4, p. 751–765, 2018.

YIN, L.; LV, L.; WANG, D.; QU, Y.; CHEN, H.; DENG, W. Spectral clustering approach with k-nearest neighbor and weighted mahalanobis distance for data mining. *Electronics*, MDPI, v. 12, n. 15, p. 3284, 2023.

YU, F.; ZHAO, R.; SHI, Z.; LU, Y.; FAN, J.; ZENG, Y.; MAO, J.; LI, W. Boosting spectral clustering on incomplete data via kernel correction and affinity learning. *Advances in Neural Information Processing Systems*, v. 36, 2024.

YU, T.; ZHAO, Y.; HUANG, R.; LIU, S.; ZHANG, X. Fast approximate spectral clustering via adaptive filtering of random graph signals. In: IEEE. *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. [S.l.], 2020. p. 511–514.

YUAN, C.; YANG, H. Research on k-value selection method of k-means clustering algorithm. *J*, MDPI, v. 2, n. 2, p. 226–235, 2019.

ZELNIK-MANOR, L.; PERONA, P. Self-tuning spectral clustering. *Advances in neural information processing systems*, v. 17, 2004.

ZHANG, C.; ZHU, G.; LIAN, B.; CHEN, M.; CHEN, H.; WU, C. Image segmentation based on multiscale fast spectral clustering. *Multimedia Tools and Applications*, Springer, v. 80, p. 24969–24994, 2021.

ZHANG, L.; LI, J.; WANG, C. Automatic synonym extraction using word2vec and spectral clustering. In: IEEE. *2017 36th Chinese Control Conference (CCC)*. [S.l.], 2017. p. 5629–5632.

ZHANG, S.; LI, X.; ZONG, M.; ZHU, X.; CHENG, D. Learning k for knn classification. *ACM Transactions on Intelligent Systems and Technology (TIST)*, ACM New York, NY, USA, v. 8, n. 3, p. 1–19, 2017.

ZHANG, W.; YE, X.; SAKURAI, T. Ensemble learning for cluster number detection based on shared nearest neighbor graph and spectral clustering. In: IEEE. *2022 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2022. p. 1–8.

ZHANG, X.; LI, J.; YU, H. Local density adaptive similarity measurement for spectral clustering. *Pattern Recognition Letters*, Elsevier, v. 32, n. 2, p. 352–358, 2011.

ZHANG, X.; YOU, Q. An improved spectral clustering algorithm based on random walk. *Frontiers of Computer Science in China*, Springer, v. 5, p. 268–278, 2011.

ZHANG, X.-Y.; WANG, L.; XIANG, S.; LIU, C.-L. Retargeted least squares regression algorithm. *IEEE transactions on neural networks and learning systems*, IEEE, v. 26, n. 9, p. 2206–2213, 2014.

ZHONG, G.; PUN, C.-M. Self-taught multi-view spectral clustering. *Pattern Recognition*, Elsevier, v. 138, p. 109349, 2023.

ZHU, X.; ZHANG, S.; HE, W.; HU, R.; LEI, C.; ZHU, P. One-step multi-view spectral clustering. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 31, n. 10, p. 2022–2034, 2018.

ZHU, X.; ZHANG, S.; LI, Y.; ZHANG, J.; YANG, L.; FANG, Y. Low-rank sparse subspace for spectral clustering. *IEEE Transactions on knowledge and data engineering*, IEEE, v. 31, n. 8, p. 1532–1543, 2018.