Universidade Federal de São Carlos– UFSCar
Centro de Ciências Exatas e de Tecnologia– CCET
Departamento de Computação– DC
Programa de Pós-Graduação em Ciência da Computação– PPGCC

Gustavo Henrique Chavari

Aprendizado não supervisionado de métricas utilizando geometria diferencial e o algoritmo ISOMAP no agrupamento de dados

São Carlos
2024

Gustavo Henrique Chavari

Aprendizado não supervisionado de métricas utilizando geometria diferencial e o algoritmo ISOMAP no agrupamento de dados

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências Exatas e de Tecnologia da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Mestrado em Ciência da Computação.

Área de concentração: Processamento de Imagens e Visão Computacional

Orientador: Alexandre Luís Magalhães Levada

São Carlos
2024

Resumo

O aprendizado não supervisionado de métricas consiste na construção de funções adaptativas de distância sem o conhecimento dos rótulos das classes e visa melhorar tanto o agrupamento quanto a classificação supervisionada de padrões. Normalmente, este processo pode ser realizado por múltiplos algoritmos de aprendizado de variedades, através da redução de dimensionalidade não linear. Recentemente, um novo algoritmo, conhecido como K-ISOMAP, foi proposto para esta finalidade. Ele utiliza medidas baseadas em geometria diferencial para substituir a distância euclidiana por medidas baseadas na curvatura local no método ISOMAP. Trata-se de um método que utiliza conceitos da geometria diferencial para construir uma função de distância intrínseca que mede as variações dos espaços tangentes locais ao longo dos caminhos mais curtos no grafo k-NN, motivado pelas equações de Frenet-Serret e a noção de curvatura. Este trabalho consiste em investigar a qualidade dos agrupamentos obtidos via GMM após o mapeamento dos dados para espaços de menor dimensão. Os resultados sobre diversos conjuntos de dados sugerem que o método K-ISOMAP é capaz de produzir agrupamentos melhores do que os produzidos pelo algoritmo ISOMAP padrão, sendo competitivo em relação ao estado-da-arte em aprendizado de métricas e variedades.

Palavras-chave: ISOMAP, curvatura, aprendizado de métricas não supervisionado, aprendizado de variedades, redução de dimensionalidade, agrupamento.

Abstract

Unsupervised metric learning consists of constructing adaptive distance functions without knowledge of class labels and aims to improve both clustering and supervised pattern classification. Typically, this process can be performed by multiple manifold learning algorithms, through nonlinear dimensionality reduction. Recently, a new algorithm, known as K-ISOMAP, has been proposed for this purpose. It uses differential geometry-based measures to replace the Euclidean distance with measures based on local curvature in the ISOMAP method. This method uses concepts from differential geometry to construct an intrinsic distance function that measures the variations of local tangent spaces along edges in the k-NN graph, motivated by the Frenet-Serret equations and the notion of curvature. This work investigates the quality of the clustering obtained via GMM after mapping the data to lower-dimensional spaces. The results on several datasets suggest that the K-ISOMAP method can produce better clustering than those produced by the standard ISOMAP algorithm, being competitive with the state-of-the-art in metric and manifold learning.

Keywords: ISOMAP, curvature, unsupervised metric learning, manifold learning, dimensionality reduction, clustering.

Lista de ilustrações

Lista de tabelas

Sumário

Capítulo 1

Introdução

A habilidade de associar similaridades entre diferentes objetos tem sido uma questão latente no aprendizado de máquina desde sua concepção. Para medir similaridades, é necessário introduzir uma métrica de distância que permite estabelecer quando um objeto é mais similar a outro. Métodos para aprendizado de métricas via redução de dimensionalidade buscam aprender uma métrica ou distância mais apropriada para dados em um espaço de alta dimensionalidade (BELLET; HABRARD; SEBBAN, 2014). Esses métodos são importantes, pois muitas vezes os dados não estão bem representados no espaço original, o que prejudica o desempenho de algoritmos de agrupamento e classificação. A escolha de uma métrica inapropriada pode levar a uma degradação no desempenho dos modelos, enquanto aprender uma métrica mais adequada pode melhorar a separação entre classes, a identificação de padrões e a generalização dos modelos (LI; TIAN, 2018).

Nos últimos anos, houve um considerável interesse na análise e processamento de dados em espaços de alta dimensão. A ideia mais aceita parte da intuição de que dados são gerados por sistemas estruturados com, potencialmente, muito menos graus de liberdade do que a dimensão ambiente sugere. Assim, diversos pesquisadores têm explorado o caso em que os dados residem em ou próximo a

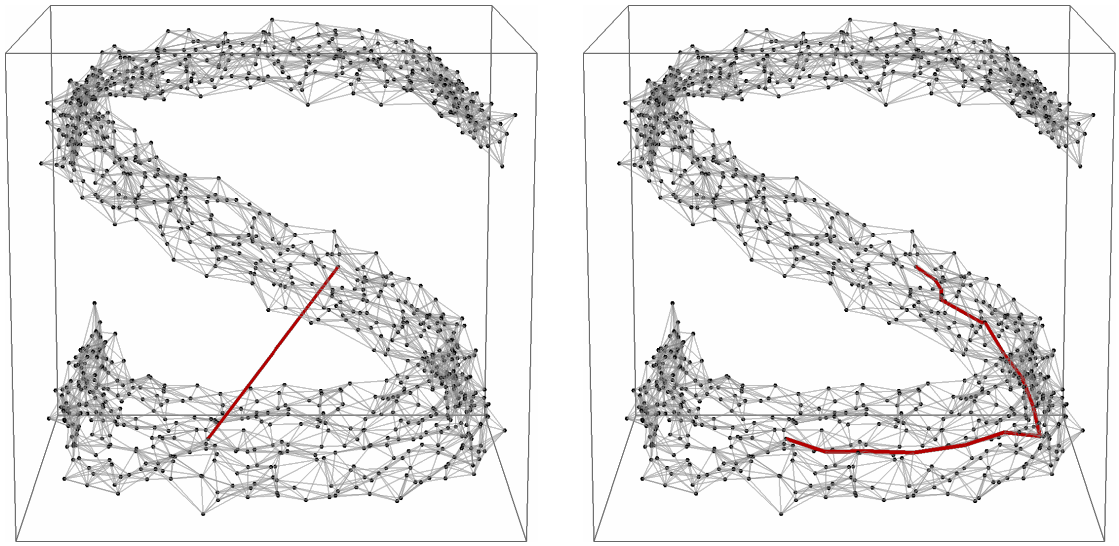
uma variedade do espaço ambiente (MEILă; ZHANG, 2023). O objetivo é estimar propriedades geométricas e topológicas a partir de dados situados sobre a variedade desconhecida. Esta consideração alavancou a criação de uma área de pesquisa, chamada de aprendizado de variedades, focada em descobrir tais estruturas. Entre os pioneiros da área, estão os métodos LLE e ISOMAP. Estes mostraram como era possível recuperar a variedade dentro do espaço euclidiano de dados, tanto localmente quanto globalmente. Essas técnicas encontraram aplicações diversas em problemas de classificação, agrupamento, recuperação de informações, bioquímica, física e mineração de dados (BELLMAN, 2003).

Sobre um ponto de vista vetorial, a alta dimensionalidade tem efeito direto sobre a noção de distância, implicando que padrões próximos estarão sempre distantes entre si. Como consequência, técnicas que funcionam eficientemente em baixa dimensão se tornam ineficazes em alta dimensão, devido à dificuldade em capturar relações significativas entre os dados. Neste contexto, o efeito negativo causado pelo aumento do número de dimensões é conhecido como *maldição da dimensionalidade* (NEGRI, 2021). Para contornar este problema, é sugerida uma redução no número de dimensões do espaço de atributos, de forma a preservar o máximo de informações originais possíveis. Este processo, chamado de *redução de dimensionalidade*, é utilizado como um passo no pré-processamento de dados e também auxilia na seleção de atributos relevantes para o problema. Métodos de redução de dimensionalidade fazem parte de classes de métodos mais gerais, conhecidos como aprendizado de métricas e aprendizado de variedades.

Embora sejam áreas distintas, o aprendizado de métricas e o aprendizado de variedades estão correlacionados e possuem relação direta com algoritmos de agrupamento. Em sua essência, o agrupamento envolve organizar pontos de dados semelhantes em grupos, utilizando funções de distância para descobrir padrões naturais ou similaridades que podem existir no conjunto de dados. O mais simples dos algoritmos de aprendizado de métricas é o algoritmo k-NN. Com ele, é possível induzir uma topologia no conjunto de dados de acordo com os vizinhos obtidos. Embora a distância euclidiana global possa não resumir propriedades importantes disponíveis nos dados, a distância euclidiana local no grafo k-NN serve de suporte para a maioria dos métodos de aprendizado de métricas e variedades. Em especial em dados de alta dimensionalidade, a busca por distâncias e espaços onde

as noções de similaridade ou localidade sejam preservadas e melhoradas é o foco principal das duas áreas (SUáREZ; GARCíA; HERRERA, 2021).

É conhecido que, no caso de amostras finitas, existe uma dimensionalidade ótima d^* de variáveis cuja acurácia média de classificação supervisionada é maior (HUGHES, 1968). Embora a questão da dimensionalidade seja de suma importância, ela necessita de métodos adequados de aprendizado de métricas e redução de dimensionalidade, além da boa descrição da tarefa e das hipóteses sobre os dados de entrada. O algoritmo ISOMAP (TENENBAUM; SILVA; LANGFORD, 2000) foi um dos primeiros a considerar um aprendizado de métricas local e global. Depois, vieram algoritmos como Laplacian Eigenmaps, t-SNE e UMAP que hoje, junto com o ISOMAP, compõem o estado da arte dos algoritmos de aprendizado de métricas.



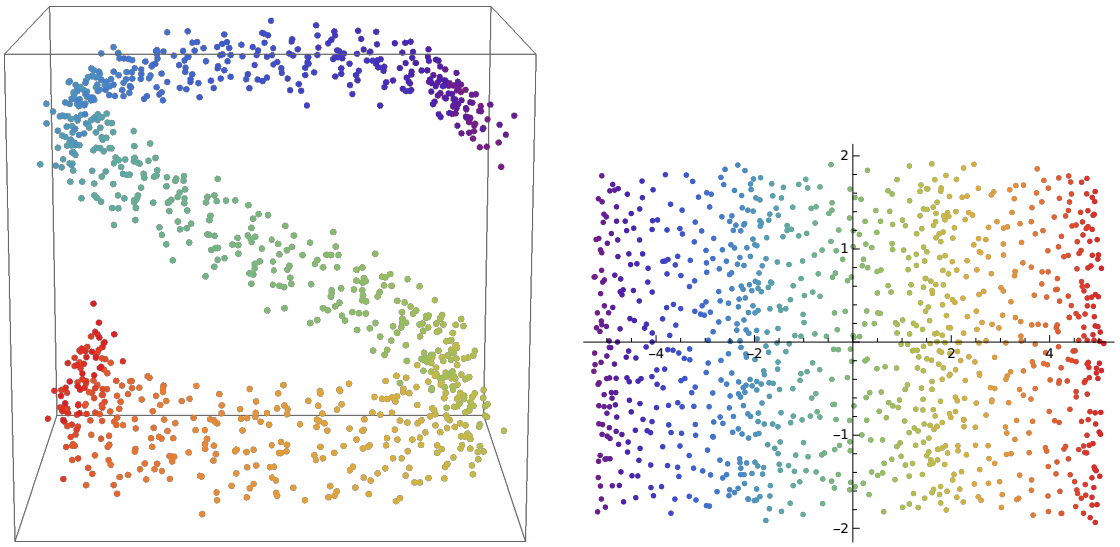
(a) (b)

Figura 1 – Diferença entre as distâncias (a) euclidiana e (b) geodésica em um grafo que representa a superfície S.

Uma grande limitação dos algoritmos de agrupamento tradicionais é a forte dependência da distância euclidiana. Muitos algoritmos, como K-Means, utilizam essa métrica para particionar dados em grupos com base no conceito de proximidade. Embora a distância euclidiana se alinhe bem com a suposição de que pontos dentro do mesmo cluster devem estar mais próximos uns dos outros do que

de pontos em clusters diferentes, essa dependência apresenta desafios em cenários onde os dados exibem relacionamentos não lineares e complexos. Nesses casos, a distância euclidiana pode não capturar com precisão as similaridades subjacentes, limitando a eficácia dos algoritmos de agrupamento.

Consequentemente, o desenvolvimento de métricas de distância adaptativas que suportam diversas estruturas se tornou uma área ativa de pesquisa. O aprendizado de métricas compreende um arcabouço de teorias e técnicas que capturam relações geométricas intrínsecas de conjuntos de dados. Sob a influência da maldição da dimensionalidade, é comum assumir que os pontos de dados estão imersos em um espaço de alta dimensão, geralmente com uma dimensão intrínseca significativamente menor. Essa abordagem, conhecida como *hipótese de variedade*, é frequentemente válida em aplicações do mundo real, e possui a capacidade de melhorar o desempenho de várias tarefas de aprendizado de máquina, incluindo classificação, agrupamento, recuperação de informações e mineração de dados (CARTER; RAICH; AU2, 2008).



(a) (b)

Figura 2 – Diferença entre as distâncias (a) euclidiana e (b) geodésica em um grafo que representa a superfície S.

O aprendizado de métricas também representa uma mudança de paradigma no aprendizado de máquina, visando descobrir automaticamente uma métrica de

distância apropriada diretamente dos dados, sem depender de informações rotuladas. Essa abordagem é fundamental em cenários onde os rótulos do problema estão ausentes, aproveitando a estrutura intrínseca dos dados para aprender similaridades significativas. Os Modelos de Mistura Gaussiana (GMM), por exemplo, utilizam a distância de Mahalanobis e a abordagem probabilística para generalizar o K-Means, permitindo que os grupos detectados assumam formas mais flexíveis além dos padrões circularmente simétricos (MAHMOOD; MEHMOOD; AL-ESSA, 2023).

Nos últimos anos, a incorporação de conceitos da geometria diferencial em métodos computacionais tem trazido benefícios significativos para diversas áreas da ciência. Em processamento de imagens e visão computacional, conceitos de fluxo, curvatura e tensores foram aplicados para segmentação, detecção de bordas, reconhecimento de objetos e reconstrução 3D. No aprendizado de máquina, a geometria diferencial fundamenta métodos de aprendizado de métricas, aprendizado de variedades e aprendizado profundo. No caso específico do aprendizado de métricas e aprendizado de variedades, o horizonte se mostra promissor, pois a geometria diferencial oferece diversas ferramentas para lidar com questões de métrica, dimensão e forma em espaços generalizados.

Recentemente, o K-ISOMAP foi proposto como um algoritmo de aprendizado de métricas não supervisionado baseado em conceitos da geometria diferencial. Sendo um método de redução de dimensionalidade não linear inspirado pelo ISOMAP, ele utiliza a variação do espaço tangente ao longo das arestas do grafo para construir uma função de distância intrínseca. Matemáticamente, tal variação diz respeito a curvatura. A principal vantagem do K-ISOMAP sobre o ISOMAP regular é a robustez da função de distância proposta contra outliers e seu poder discriminatório em espaços de alta dimensão em comparação com a distância euclidiana. Sabe-se que pontos de alta curvatura desempenham um papel dominante na percepção de forma por humanos (CHETVERIKOV, 2003). E com isso, vários métodos de aprendizado de máquina e reconhecimento de padrões não supervisionados tem começado a considerar esta informação geométrica intrínseca (RENGASWAMI; BOURNI; MAROULAS, 2024).

Neste trabalho, aplica-se o K-ISOMAP para aprender funções de distância adaptativas em agrupamento por GMM. Sua significância estatística em relação

ao ISOMAP padrão e outros métodos do estado-da-arte é testada e analisada. De acordo com os resultados, foi constatado que é possível melhorar a qualidade de grupos em comparação com o ISOMAP. Em relação ao estado-da-arte, o K-ISOMAP é competitivo, também se mostrando superior em muitos cenários. Sendo assim, faz-se necessário dissertar sobre a relação intrínseca entre aprendizado de métricas e agrupamento de dados. Estudando como a incorporação de funções de distância adaptativas baseadas na curvatura local pode ajudar a encontrar grupos, o presente trabalho também destaca direções que o K-ISOMAP pode seguir.

A estrutura está organizada da seguinte forma: o capítulo 2 apresenta os trabalhos relacionados que norteiam as direções futuras do trabalho. O capítulo 3 fundamenta os conceitos principais do aprendizado de variedades, aprendizado de métricas e agrupamento de dados utilizados no trabalho. O Capítulo 4 descreve e analisa o algoritmo K-ISOMAP. O Capítulo 5 apresenta a metodologia do experimento, da qual diversos conjuntos de dados são testados, junto da avaliação dos agrupamentos e da aplicação do testes de hipótese. Por fim, o Capítulo 5 traz as considerações finais, destacando as contribuições do trabalho e apontando as possíveis direções para trabalhos futuros.

Capítulo 2

Trabalhos relacionados

A Curvature Based Isometric Feature Mapping (LEVADA, 2022) O algoritmo K-ISOMAP proposto aplica conceitos de geometria diferencial para extrair informações intrínsecas de uma variedade de dados. Baseando-se nas equaçoes de Frenet-Serret e na noção de curvatura, o método consiste na criação de um grafo k-NN, chamado K-grafo, cujos pesos das arestas são calculados a partir de um *vetor de curvaturas*. Este vetor representa a extração de informação geométrica local a partir da variação dos espaços tangentes em cada vértice. Em sequência, é aplicado o processo do ISOMAP padrão: encontram-se os caminhos mínimos do K-grafo e a versão de baixa dimensão do conjunto de dados que preserva as mínimas distâncias. Experimentalmente, o K-ISOMAP mostrou-se vantajoso ao ser capaz de aumentar a performance de tarefas de classificação, se comparado a algoritmos de aprendizado de métricas e redução de dimensionalidade do estado-da-arte. A evidência empírica de que quanto maior a curvatura da aresta, maior o custo para se caminhar de um extremo a outro da aresta, sugere que regiões de alta curvatura podem ser vistas como regiões de decisão. Assim, o K-ISOMAP motiva diretamente aprofundamento no estudo das relações do aprendizado de métricas com geometria diferencial em problemas de classifiçao e agrupamento.

Curvature-based Clustering on Graphs (TIAN; LUBBERTS; WEBER, 2023) Em tarefas de detecção de comunidades em grafos, diversos métodos têm explorado a geometria do grafo para identificar subestruturas densamente conectadas que formam clusters ou comunidades (podem pertencer a mais de um grupo). Em particular, estudos recentes propõem algoritmos que utilizam curvaturas de Ricci discretas e fluxos geométricos associados, nos quais os pesos das arestas do grafo evoluem para revelar a estrutura comunitária. Esses algoritmos são analisados tanto para detecção de comunidades de filiação única, onde cada nó pertence a uma única comunidade, quanto para detecção de comunidades de filiação mista, onde comunidades podem se sobrepor. Para o caso de filiação mista, o uso do grafo linha (dual) é sugerido como uma abordagem vantajosa. Evidências teóricas e empíricas são apresentadas em favor da eficácia dos algoritmos baseados em curvatura. Além disso, os autores discutem a relação entre a curvatura de um grafo e a de seu grafo dual, possibilitando uma implementação eficiente da detecção de comunidades com filiação mista, contribuindo também para a análise de redes baseada em curvatura.

Spectral Curvature Clustering (SCC) (CHEN; LERMAN, 2009) Este trabalho situa-se sobre o *hybrid linear modeling*, que assume que um conjunto de dadods pode ser aproximado por uma mistura de variedades lineares com um grau de propabilidade entre cada uma. O método utiliza a curvatura polar para calcular a matriz de afinidade dos dados, aprimorando o desempenho de uma estrutura de agrupamento espectral voltada para a segmentação de subespaços afins. Especificamente, o artigo propõe um procedimento de amostragem iterativa para melhorar a estratégia de amostragem uniforme, um procedimento preciso de inicialização para o K-means e uma estratégia simples para isolar outliers. O algoritmo resultante, denominado Spectral Curvature Clustering (SCC) possui complexidade linear, e é apoiado por uma teoria que justifica seu desempenho bem-sucedido, orientando escolhas práticas do mundo real.

Kernel Spectral Curvature Clustering (KSCC) (CHEN; ATEV; LERMAN, 2009) Embora o problema geral de modelar dados como misturas de variedades seja extremamente desafiador, várias abordagens foram propostas para modelar dados como misturas de subespaços afins, incluindo o SCC. Na verdade, este trabalho é uma generalização do SCC utilizando o kernel trick. O algoritmo

resultante, denominado Kernel Spectral Curvature Clustering, utiliza kernels em dois níveis – tanto como um método de embedding implícito para linearizar variedades não planas quanto como um método fundamentado para converter um problema de afinidade variado em um problema de clustering espectral. A eficácia do método é demonstrada através de comparações com outros métodos do estado-da-arte, tanto em dados sintéticos quanto em um problema real.

Nonlinear Dimensionality Reduction for Clustering (TASOULIS; PAVLIDIS; ROOS, 2020) Este trabalho propõe uma abordagem não linear para o agrupamento hierárquico, sendo capaz de identificar clusters em variedades não lineares. O método utiliza o ISOMAP recursivamente para mergulhar subconjuntos de dados em uma dimensão menor, em seguida realiza uma partição binária projetada para evitar a divisão de clusters. Apresenta-se uma análise teórica das condições sob as quais clusters contínuos de alta densidade no espaço original são garantidos como separáveis no caso 1-dimensional. Experimentos extensivos em conjuntos de dados simulados e reais demonstram que algoritmos de clustering hierárquico divisivo derivados desta abordagem são eficazes.

Nonlinear subspace clustering using curvature constrained distances (BABAEIAN et al., 2015) Técnicas de modelagem de multi-variedades estão obtendo sucesso em inúmeros problemas em visão computacional e mineração de dados. Métodos tradicionais de subespaço, como PCA, assumem que os dados são distribuídos a partir de uma única variedade. No entanto, os dados podem estar distribuídos em várias variedades (ou superfícies) possivelmente. Isso motivou o desenvolvimento de novas técnicas não lineares para o agrupamento de subespaços em dados de alta dimensionalidade. O método proposto neste trabalho tem o objetivo de recuperar a estrutura de baixa dimensionalidade dos dados, focando em realizar agrupamentos efetivos. Para isso, propõe-se uma restrição de curvatura para encontrar o caminho mais curto entre pontos de dados e, em seguida, utilizá-la no ISOMAP para aprendizado de subespaços. O algoritmo seleciona aleatoriamente diversos nós de referência (landmarks) e verifica se existe um caminho restrito por curvatura entre cada nó de referência e todos os outros nós no grafo de vizinhança. A partir disso, ele constrói um vetor binário de características para cada ponto, onde cada entrada representa a conectividade daquele ponto a um determinado nó de referência. Esses vetores binários podem ser utilizados como

entrada em algoritmos de clustering convencionais, como o clustering hierárquico. Os experimentos realizados em conjuntos de dados sintéticos e reais confirmam o bom desempenho do algoritmo.

Ollivier's Ricci Curvature, Local Clustering and Curvature-Dimension Inequalities on Graphs (JOST; LIU, 2014) Neste artigo, explora-se a relação entre uma das propriedades mais elementares e importantes dos grafos, a presença e frequência relativa de triângulos, e uma noção combinatória de curvatura de Ricci. Em analogia com as noções de curvatura na geometria Riemanniana, a curvatura de Ricci é interpretada como um controle sobre a sobreposição entre as vizinhanças de dois vértices vizinhos. Assim, ela está naturalmente relacionada à presença de triângulos que contêm esses vértices, ou mais precisamente, ao coeficiente de agrupamento local, que é a proporção relativa de vizinhos conectados entre todos os vizinhos de um vértice. Com base nisso, deriva-se limites inferiores para a curvatura de Ricci em grafos em termos desses coeficientes de agrupamento local. Também é explorado desigualdades de curvatura-dimensão em grafos, baseando-nos em trabalhos anteriores de vários autores.

Multiple Manifold Clustering Using Curvature Constrained Path (BABAEIAN, 2018) O problema de agrupamento de múltiplas variedades é uma tarefa desafiadora, especialmente quando essas superfícies se intersectam. Métodos disponíveis, como o Isomap, não conseguem capturar corretamente a verdadeira forma da variedade nas proximidades da interseção, pois assumem que a variedade é conexa, resultando em agrupamentos incorretos para mais de uma variedade. O algoritmo ISOMAP utiliza o caminho mais curto entre os pontos, mas sua principal limitação, destacada neste artigo, está na ausência de uma restrição de curvatura, o que pode resultar na existência de caminhos entre pontos em superfícies diferentes. Este problema é abordado impondo uma restrição de curvatura ao algoritmo de caminho mínimos utilizado no ISOMAP. O algoritmo proposto seleciona aleatoriamente vários nós de referência e, em seguida, verifica se há um caminho restrito por curvatura entre cada nó de referência e todos os outros nós no grafo de vizinhança. A partir disso, é construído um vetor binário de características para cada ponto, onde cada entrada representa a conectividade desse ponto com um determinado nó de referência. Estes vetores binários podem então ser utilizados como entrada em algoritmos de agrupamento convencionais.

A Riemannian Approach to Graph Embedding (ROBLES-KELLY; HANCOCK, 2007) Este trabalho explora a relação entre o operador de Laplace-Beltrami e o Laplaciano de grafos com o objetivo de realizar a imersão de um grafo em uma variedade Riemanniana. Utilizando as propriedades dos campos de Jacobi, desenvolve-se uma matriz de pesos de arestas que reflete as curvaturas seccionais associadas as geodésicas na variedade. Em superfícies de curvatura seccional constante, as coordenadas de Kruskal são utilizadas para calcular pesos proporcionais às distâncias geodésicas entre os pontos, permitindo a imersão dos vértices do grafo na variedade. Uma técnica de centralização da matriz de Laplace possibilita a obtenção das coordenadas de imersão a partir dos autovetores. Com essas coordenadas, tarefas de manipulação em grafos, como o problema de correspondência, podem ser realizadas.

Capítulo 3

Fundamentação Teórica

3.1 Conceitos de Geometria Diferencial

Nesta seção, busca-se apresentar de forma resumida a linguagem da geometria diferencial moderna utilizada pelos métodos de aprendizado de métricas e variedades. Conhecimentos prévios de cálculo diferencial, álgebra linear e topologia serão assumidos pelo leitor. Para uma introdução aprofundada, veja (CARMO, 1992).

Definição 3.1.1. (Variedade) Um conjunto $\mathcal{M}^d$ é uma *variedade diferenciável* (ou simplesmente variedade) de dimensão d se for um espaço topológico Hausdorff e de base enumerável tal que:

1. Para cada ponto $p \in \mathcal{M}$, existe uma aplicação φ e uma vizinhança aberta $U \subseteq \mathcal{M}$ de p tal que $\varphi : U \to \varphi(U) \subseteq \mathbb{R}^d$ é bijetiva e tanto φ quanto φ^{-1} são suaves. O par (U, φ) é chamado de *carta*, e $\varphi^{-1} : \varphi(U) \subseteq \mathbb{R}^d \to \mathcal{M}$ é chamada de *coordenada local*. Assim, coordenadas locais associam suavemente pontos em $\mathbb{R}^d$ para pontos em $\mathcal{M}$.

2. Para quaisquer dois pontos $p, p' \in \mathcal{M}$ e duas cartas $(U, \varphi), (V, \phi)$ que os

contenham, se $U \cap V \neq \emptyset$, então a função de transição entre as cartas

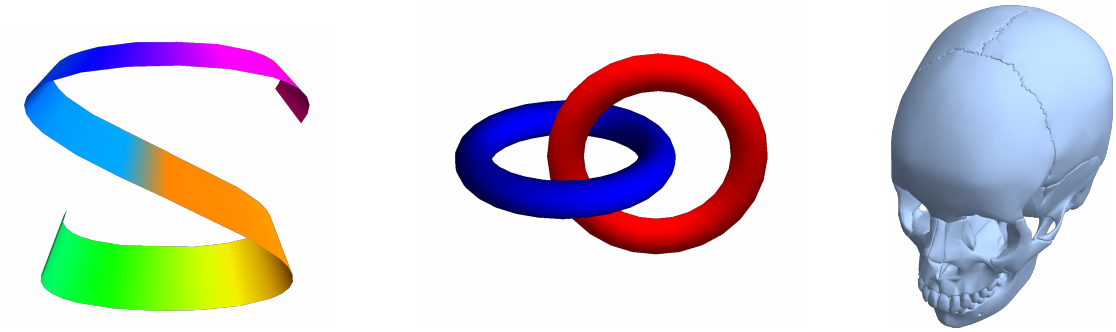
$$\varphi \circ \phi^{-1} : \varphi(U \cap V) \to \phi(U \cap V)$$

é suave e possui uma inversa suave. Isto significa que as mudanças de coordenadas locais não distorcem vizinhanças.

Em resumo, uma variedade suave é um conjunto de pontos que localmente se assemelha a um espaço euclidiano, embora globalmente isto não seja válido. Exemplos de variedades contemplam:

1. O próprio $\mathbb{R}^d$ é uma variedade de dimensão d.
2. Curvas planas ou espaciais são variedades de dimensão 1.
3. Superfícies em $\mathbb{R}^3$ como a esfera e o toro são variedades de dimensão 2.
4. O espaço das matrizes quadradas $M_n(\mathbb{R})$ é uma variedade de dimensão n^2.

Observe também que, mesmo com a existência de várias cartas para uma dada variedade $\mathcal{M}$, pela condição (2) acima, a dimensão d deve ser a mesma para todos os pontos em $\mathcal{M}$. Assim, d é chamado de *dimensão intrínseca* da variedade $\mathcal{M}$.



(a) (b) (c)

Figura 3 – Exemplos de variedades: (a) Superfície S: Uma variedade bidimensional sem buracos ou descontinuidades. (b) Hopf Link: União de dois toros bidimensionais entrelaçados. (c) Crânio humano: Uma variedade tridimensional com múltiplas componentes conexas e buracos, destacando a presença de características topológicas não triviais.

Observação (União disjunta de variedades)**.** Seja $\{\mathcal{M}_i\}_{i=1}^k$ uma coleção de variedades suaves de dimensão intrínseca d, onde cada $\mathcal{M}_i$ é uma variedade diferenciável e $\mathcal{M}_i \cap \mathcal{M}_j = \emptyset$ para $i \neq j$. A união disjunta $\mathcal{M} = \bigcup_{i=1}^k \mathcal{M}_i$ pode ser munida de uma estrutura diferenciável que a transforma em uma variedade d-dimensional.

3.1.1 Espaço Tangente e a Métrica Riemanniana

Outro objeto fundamental em geometria diferencial que fornece uma maneira de compreender o comportamento local de uma variedade em um ponto específico é o espaço tangente. Intuitivamente, o espaço tangente a uma variedade num ponto captura a noção de direções "infinitesimais" naquele ponto. Outra maneira de ver o espaço tangente é como uma aproximação linear da variedade em um ponto. Conceitos como distâncias e ângulos, ou seja, a geometria euclidiana, são transferidos de $\mathbb{R}^d$ para as variedades com o apoio dos espaços tangentes. Para começar a estudar este objeto, é necessário entender o conceito de vetor tangente.

Um vetor tangente em um ponto pode ser pensado como um par (p, v) que carrega duas informações: a *topológica*, que diz respeito a qual ponto o vetor é tangente, e a parte *vetorial*, sendo o vetor que é tangente de fato. É representado então por uma flecha, fixada em p, com o módulo, direção e sentido do vetor v. Matematicamente, fixando $p \in \mathcal{M}$, considere uma curva suave $\alpha : (-\varepsilon, \varepsilon) \to \mathcal{M}$ tal que $\alpha(0) = p$, para algum $\varepsilon > 0$. Um vetor tangente é definido como a avaliação de uma função sobre uma curva α ao redor de p. A ideia é que, ao avaliar diferentes funções sobre curvas em p, produzimos vetores de diferentes direções e magnitudes, ditos tangentes. E a junção destes objetos compreende o espaço dos vetores tangentes, um espaço vetorial com uma base natural cuja dimensão é a mesma da variedade. Para isso, considere $\mathcal{D}$ o conjunto das funções $f : \mathcal{M} \to \mathbb{R}$ diferenciáveis em p.

Definição 3.1.2. (Espaço Tangente) O *vetor tangente à curva* α em $t = 0$ é uma função $\alpha'(0) : \mathcal{D} \to \mathbb{R}$ dada por

$$\alpha'(0)(f) = \left.\frac{d(f \circ \alpha)}{dt}\right|_{t=0}.$$

Um *vetor tangente em p* é o vetor tangente em $t = 0$ de alguma curva α com $\alpha(0) = p$. O conjunto de todos os vetores tangentes a $\mathcal{M}$ em p é chamado de *espaço tangente em p* e será denotado por $T_p\mathcal{M}$.

Pode ser mostrado, utilizando uma carta (U, φ) ao redor de p com $\varphi = (x_1, \ldots, x_d)$, que o vetor $\alpha'(0)$ pode ser escrito em função a parametrização φ:

$$\alpha'(0) = \sum_{i=1}^{n} x_i'(0) \frac{\partial}{\partial x_i}\bigg|_{t=0}.$$

Note que, $\partial/\partial x_i|_{t=0}$ é o vetor tangente em p da *curva coordenada* x_i. Além disso, o conjunto $\{\frac{\partial}{\partial x_i}|_{t=0}\}_{i=1}^{d}$ forma uma base para o espaço vetorial $T_p\mathcal{M}$, consolidando-o como um espaço vetorial de dimensão d.

Definição 3.1.3. (Diferencial) Suponha que $f : \mathcal{M} \to \mathcal{N}$ seja uma aplicação de uma variedade $\mathcal{M} \subset \mathbb{R}^d$ em uma variedade $\mathcal{N} \subset \mathbb{R}^m$. A *diferencial* $df_p : T_p\mathcal{M} \to T_{f(p)}\mathcal{N}$ de f em um ponto $p \in \mathcal{M}$ é dada por

$$df_p(v) = (f \circ \alpha)'(0),$$

onde $\alpha : (-\epsilon, \epsilon) \to \mathcal{M}$ é qualquer curva tal que $\alpha(0) = p$ e $\alpha'(0) = v \in T_p\mathcal{M}$.

Muitas quantidades em variedades se resumem a cálculos sobre comprimentos e ângulos de vetores tangentes. Isto é possível pois, em $\mathbb{R}^m$, o produto escalar $\langle v, u\rangle = v^T u$ é suficiente para definir distâncias por $\|v - u\|^2 = \langle v - u, v - u\rangle$, e ângulos, por $\angle(v, u) = \cos^{-1}(\langle v, u\rangle/(\|v\|\|u\|))$. Por exemplo, ao calcular o comprimento de uma curva $\alpha : [a, b] \to \mathbb{R}^m$, note que na expressão

$$\mathcal{L}_a^b(\alpha) = \int_a^b \sqrt{\langle \alpha'(t), \alpha'(t)\rangle} dt. \tag{3.1}$$

o produto interno em $\mathbb{R}^m$ implicitamente varia ponto a ponto, pois o vetor tangente muda de origem para cada ponto da curva. Ou seja, está sendo calculado em $T_p\mathbb{R}^m \simeq \mathbb{R}^m$. Riemann generalizou esta ideia, observando que qualquer matriz $A \in \mathbb{R}^{m\times m}$ definida positiva pode induzir um *produto interno* por uma forma quadrática $\langle v, u\rangle_A = v^T A u$, onde $u, v \in T_p\mathcal{M}$. Este produto interno, quando definido para todo ponto em $\mathcal{M}$ é geralmente chamado de *tensor métrico*.

Definição 3.1.4. (Variedade Riemanniana) Uma *métrica Riemanniana* em uma variedade diferenciável M é uma correspondência que associa a cada ponto p de M um produto interno g_p (isto é, uma forma simétrica, bilinear e definida positiva) no espaço tangente T_pM, que varia diferenciavelmente da seguinte forma: Se $\varphi : U \subset \mathbb{R}^n \to M$ é uma carta ao redor de p, com $\varphi(x_1, x_2, \ldots, x_n) = q \in \varphi(U)$ e $\frac{\partial}{\partial x_i}(q) = dx_q(0, \ldots, 1, \ldots, 0)$, então $g_q(\frac{\partial}{\partial x_i}(q), \frac{\partial}{\partial x_j}(q)) = g_{ij}(x_1, \ldots, x_n)$ é uma função suave em U. O par (M, g) é chamado de variedade Riemanniana.

Vale observar que a métrica Riemanniana não é a mesma coisa que a métrica de distância usual $d(x, y)$. Na verdade, ela é uma ferramenta que permite cálculos de distâncias e outros elementos geométricos intrínsecos na variedade. Para ver isso, considere o caso de uma imersão $\varphi : U \subset \mathbb{R}^2 \to M \subset \mathbb{R}^3$ que representa uma superfície. Naturalmente, surge a questão de como medir o efeito de φ sobre vetores $u, v \in T_pU$ de maneira intrínseca. De acordo com a métrica

$$g(u, v) = \langle d\varphi_p(u), d\varphi_p(v) \rangle$$

é possível representar a métrica Riemanniana de M induzida pela imersão φ. Sendo uma forma bilinear positiva-definida conhecida como *primeira forma fundamental*, é representada por uma matriz 2×2, denotada por $\mathbf{I}$. E se descrito seu comportamento na base $\{\partial/\partial x_i\}$, por bilinearidade, será descrito seu comportamento sobre todos os vetores. Com isso, é possível expressar a primeira forma fundamental via Jacobiana de φ:

$$\mathbf{I}(u, v) = \langle d\varphi_p(u), d\varphi_p(v) \rangle = (J_\varphi u)^T (J_\varphi v) = u^T (J_\varphi^T J_\varphi) v \quad \text{para todo } u, v \in T_pU.$$

Portanto, em forma matricial $\mathbf{I} = (J_\varphi^T)(J_\varphi)$.

Em uma variedade Riemanniana, o comprimento de uma curva $\alpha : [a, b] \to \mathcal{M}$ é obtido utilizando apenas métrica

$$\mathcal{L}_a^b(\alpha) = \int_a^b \sqrt{g(\alpha'(t), \alpha'(t))} dt. \tag{3.2}$$

O ângulo que duas curvas se intersectam e volumes elementares gerados por dois vetores tangentes também envolvem apenas a métrica. Naturalmente, as *geodésicas* podem ser vistas como curvas de comprimento mínimo entre dois pontos p, q em M. Porém, sua interpretação mais geral é a de uma curva α cujo vetor aceleração

$\alpha''(t)$ é perpendicular ao plano tangente da variedade. Isto é, o comprimento do vetor tangente $\alpha'(t)$ é constante e igual para todo ponto em α. Nas aplicações em aprendizado de métricas e variedades, é suficiente considerar a noção de curvas que minimizam distâncias intrínsecas, o que leva a resolução de problemas de busca em grafos.

Outro conceito de suma importância na geometria Riemanniana, sendo muito utilizado como hipótese em métodos de aprendizado de variedades, é o seguinte:

Definição 3.1.5. (Isometria) Considere M e N duas variedades Riemannianas. Um difeomorfismo $f : M \to N$ é chamado de *isometria* se

$$\langle u, v \rangle_p = \langle df_p(u), df_p(v) \rangle_{f(p)} \quad \text{para todo } p \in M,\ u, v \in T_pM.$$

Além disso, M é dito ser *localmente isométrica* a N se, para todo ponto $p \in M$, existe uma vizinhança U de p em M e uma isometria $f : U \to f(U) \subset N$.

A partir da definição de isometria, observa-se que uma transformação que preserva a métrica entre duas variedades pode ser vista como uma forma natural de preservar a estrutura geométrica subjacente. Esta ideia é central em diversas abordagens de redução de dimensionalidade em aprendizado de variedades, onde se busca aprender um mergulho que mantenha as propriedades intrínsecas dos dados. Dois exemplos importantes que emergem desse conceito são o *Locally Linear Embedding* (LLE) e o *Isometric Feature Mapping* (ISOMAP), ambos baseados na preservação das distâncias geodésicas e das relações locais entre os pontos na variedade original.

3.1.2 Curvatura de Curvas

Neste trabalho aplica-se o conceito de curvatura de uma curva para calcular medidas de similaridade entre padrões. Mesmo assim, vale a pena mencionar a riqueza geométrica acerca do conceito de curvatura, fornecendo intuições sobre sua formulação sobre quaisquer variedades. Intuitivamente, a curvatura descreve o quanto um objeto se curva. Quando considerada extrínseca, se refere a taxa de mudança de algum referencial, no caso o espaço tangente. Quando intrínseca,

representa uma medida de o quanto uma região se desvia de ser um espaço euclidiano. Neste sentido, pode ser vista como uma descrição de como um objeto está situado no espaço (CARMO, 2016).

3.1.2.1 O triedo de Frenet-Serret

Considere uma curva plana α parametrizada por comprimento de arco. Em um ponto $\alpha(t)$, a curvatura mede a variação do ângulo da tangente em $\alpha(t)$, ou seja, mede o quanto a curva desvia de sua tangente no ponto a partir da expressão

$$k(t) = \langle N(t), \alpha''(t) \rangle,$$

sendo $N(t)$ o vetor normal de α em $\alpha(t)$. Em uma curva qualquer, nem sempre a o vetor aceleração é normal, mas em curvas parametrizadas por compreimento de arco, isto vale sempre. E portanto, $k(t) = \|\alpha''(t)\|$. Desta forma, conseguimos ter uma noção se a curva está virando para esquerda ou direita, em relação ao vetor tangente $\alpha'(t)$. Se o ponto varia, tem-se uma função $k(t)$ que mede o quanto (e a direção) que a curva desvia de sua reta tangente em cada ponto.

Se α está situada no espaço ambiente $\mathbb{R}^3$, existe uma outra direção na qual a curva pode torcer, pois além de direita e esquerda, há para cima e para baixo. Por isso, nasce uma outra quantidade, chamada de *torção*. Sendo assim, é possível construir uma base ortonormal $\{T, N, B\}$ que acompanha o movimento da curva α em cada ponto, chamado *triedo de Frenet-Serret*. O vetor T é o tangente a curva, N é o vetor normal a curva e B é o produto vetorial entre T e N. Na busca da taxa de variação destes vetores de acordo com α, obtem-se as fórmulas de Frenet-Serret

$$T' = kN \qquad N' = -kT - \tau B \qquad B' = \tau N.$$

onde k é a curvatura em $\alpha(t)$ a torção $\tau(t)$ tal que $B'(t) = \tau(t)N(t)$. Estas equações não dizem apenas como se relacionam a curvatura e a torção. Mais que isso, elas dizem que: se integrarmos essas equações para encontrar $\{T, N, B\}$, dados a curvatura k e a torção τ, é possível determinar o formato da curva no espaço. Na verdade, este é o *Teorema Fundamental da Teoria Local das Curvas*:

Teorema 3.1.1. (CARMO, 2016) Dada uma curvatura (positiva) $k(s)$ e uma torção $\tau(s)$, existe uma parametrização regular $\alpha : I \to \mathbb{R}^3$ tal que s é o comprimento

de arco, $k(s)$ é a curvatura e $\tau(s)$ é a torção da curva. Além disso, se outra curva satisfaz as mesmas condições, então ela é obtida através de α por um movimento rígido (translação, rotação) em $\mathbb{R}^3$.

3.1.2.2 O frame de Darboux

Similarmente, é possível obter um referencial para quando a curva está situada sobre uma superfície. Com o plano tangente e o vetor normal, é construído um referencial análogo ao de Frenet, chamado *frame de Darboux*. Dada uma curva α parametrizada por comprimento de arco sobre uma superfície S. Seja $\{T, N, B\}$ o referencial de Frenet ao longo de α, onde T, N e B denotam os campos vetoriais tangente, normal principal e binormal, respectivamente. No frame de Darboux, o vetor tangente $\mathbf{t}(t) = T(t)$ e o vetor normal unitário $\mathbf{n}(\alpha(t))$ formam os dois primeiros vetores do frame. O terceiro vetor, dado pelo produto vetorial

$$\mathbf{b}(t) = \mathbf{n}(\alpha(t)) \wedge \mathbf{t}(t), \tag{3.3}$$

é de fato tangente a superfície. E portanto, com o referencial $\{\mathbf{t}, \mathbf{b}, \mathbf{n}\}$ é possível mostrar que

$$\mathbf{t}' = \kappa_g \mathbf{b} + \kappa_n \mathbf{n} \qquad \mathbf{b}' = -\kappa_g \mathbf{t} + \tau_g \mathbf{n} \qquad \mathbf{n}' = -\kappa_n \mathbf{t} - \tau_g \mathbf{b}$$

onde κ_n, κ_g e τ_g são a *curvatura normal*, *curvatura geodésica* e a *torção geodésica* da curva α na superfície M na direção de $\mathbf{t}$, respectivamente, definidas por

$$\kappa_n = \langle \mathbf{t}', \mathbf{n} \rangle, \quad \kappa_g = \langle \mathbf{t}', \mathbf{b} \rangle, \quad \tau_g = \langle \mathbf{b}', \mathbf{n} \rangle.$$

Do ponto de vista extrínseco da superfície, a curvatura normal da curva é precisamente como a magnitude da projeção de $\alpha''(t)$ sobre $\mathbf{n}$. Ou seja, κ_n é o comprimento da projeção do vetor $\alpha''(t)$ sobre o vetor normal à superfície em p, com um sinal dado pela orientação $\mathbf{n}$. A curvatura geodésica é vista como o comprimento da projeção de $\alpha''(t)$ no plano tangente em p. Enquanto a torção geodésica mede a variação do plano tangente em relação ao vetor normal em p.

Nos caso em que a curva possui curvatura geodésica zero para todo t, chama-se a curva de *geodésica*. E no caso em que a curvatura normal é zero, a curva é chamada de *assintótica*. Por sua vez, se a torção geodésica é zero, a curva é chamada de

linha de curvatura, e representam as linhas cuja curvatura é a máxima ou mínima possível. Mais ainda, dado que o vetor tangente a curva é o mesmo do frame de Frenet-Serret, os vetores $\mathbf{n}$ e $\mathbf{b}$ são uma rotação dos vetores normal N e binormal B da curva α no plano ortogonal a $\mathbf{t}$.

3.1.3 Curvatura de Superfícies

Mais geralmente, pode-se considerar a aplicação normal

$$\mathbf{n} : \mathcal{M} \to S^2$$

chamada de *aplicação de Gauss*, que associa a cada ponto p da superfície a um vetor normal $\mathbf{n}(p)$ unitário. Sua diferencial $d\mathbf{n}_p$, conhecida como *operador de forma*, ou *aplicação de Weingarten*, descreve a variação do vetor normal ao longo de uma dada direção tangente.

Como a variação de um vetor normal unitário não pode ter componente na direção normal, $d\mathbf{n}_p(v)$ é sempre tangente à superfície. Por isso, se $\alpha : I \to \mathcal{M}$ é uma curva tal que $\alpha(0) = p$, por definição $\alpha'(t) \in T_pM$. Além disso,

$$\langle \alpha'(t), \mathbf{n}(\alpha(t)) \rangle = 0 \quad \text{para todo } t \in I.$$

Derivando esta expressão com respeito a t, por Leibniz, tem-se

$$\langle \alpha''(t), \mathbf{n}(\alpha(t)) \rangle + \langle \alpha'(t), d\mathbf{n}_{\alpha(t)}(\alpha'(t)) \rangle = 0. \tag{3.4}$$

O fato de que $d\mathbf{n}_p : T_pM \to T_pM$ é uma aplicação linear auto-adjunta nos permite associar a $d\mathbf{n}_p$ uma forma quadrática em $T_p(S)$

$$\mathbf{II}_p(\alpha'(t)) = -\langle \alpha'(t), d\mathbf{n}_{\alpha(t)}(\alpha'(t)) \rangle \tag{3.5}$$

conhecida como *segunda forma fundamental*. Em forma matricial, a segunda forma fundamental pode ser escrita via Jacobianas de φ e $\mathbf{n}$

$$\mathbf{II}_p = -(J_\varphi)^T (J_\mathbf{n}).$$

Para uma interpretação da segunda forma fundamental, Além disso, pelas equações acima, $\mathbf{II}_p(\alpha'(t)) = \kappa_n(p)$. Em outras palavras, o valor da segunda forma fundamental II_p para um vetor unitário $v \in T_p(S)$ é igual à curvatura normal de uma

curva regular que passa por p e é tangente a v. Em particular, obtem-se o seguinte resultado: Todas as curvas que estão em uma superfície M e têm, em um ponto dado $p \in M$, a mesma linha tangente, possuem neste ponto as mesmas curvaturas normais.

Considerando todos os vetores tangentes possíveis em p, os valores máximo e mínimo da curvatura normal em um ponto são chamados de *curvaturas principais*, κ_1 e κ_2, e as direções dos vetores tangentes correspondentes são chamadas de *direções principais*. Na verdade, pode ser visto que as direções principais são os *autovetores* de $d\mathbf{n}_p$, assim como as curvaturas principais são os *autovalores* de $d\mathbf{n}_p$. E a partir das equações de Weingarten, é possível obter uma versão matricial do operador de forma

$$d\mathbf{n}_p = -\mathbf{I}^{-1} \cdot \mathbf{II},$$

onde $\mathbf{I}$ e $\mathbf{I}$ representam as matrizes das primeira e segunda formas fundamentais.

Ao contrário das curvas, que não possuem curvatura intrínseca, mas têm curvatura extrínseca (elas só possuem curvatura quando mergulhadas em um espaço), superfícies podem ter curvatura intrínseca, independente de um mergulho. A *curvatura gaussiana* é igual ao produto das curvaturas principais, $\kappa_1\kappa_2$. A curvatura gaussiana determina se uma superfície é localmente convexa (quando é positiva), em forma de sela local (quando é negativa) ou em forma plana (quando é nula). Por outro lado, a curvatura média é uma medida extrínseca, definida pela média das curvaturas principais $(\kappa_1 + \kappa_2)/2$. De modo geral, a curvatura Gaussiana K e a curvatura média H podem ser definidas, respectivamente, por

$$K = \det(d\mathbf{n}_p) \quad \text{e} \quad H = -\frac{1}{2}\mathrm{tr}(d\mathbf{n}_p).$$

Um fato substancial em geometria diferencal é descrito por Gauss em seu *Teorema Egregium*: A curvatura Gaussiana K de uma superfície é invariante por isometrias locais. O mesmo não acontece com a curvatura média H, por ser uma medida extrínseca.

É natural perguntar se a primeira e a segunda formas fundamentais definem o formato de uma superfície no espaço, assim como acontece com uma curva no espaço. De fato, o *Teorema de Bonnet* afirma, de forma geral, que se nos forem dadas as formas fundamentais $\overline{\mathbf{I}}$ e $\overline{\mathbf{II}}$ que são propostas como a primeira e a segunda

forma fundamental da superfície, então, desde que satisfaçam as equações de Gauss e Codazzi (e Ricci em um cenário mais geral), localmente existe uma superfície no espaço Euclidiano tal que suas primeira e segunda formas fundamentais são $\mathbf{I} = \bar{\mathbf{I}}$ e $\mathbf{II} = \overline{\mathbf{II}}$, respectivamente. Tal superfície é única até um movimento rígido no espaço. Observe que as curvaturas principais κ_1, κ_2 são os autovalores do operador de forma $-\mathbf{I}^{-1}\mathbf{II}$. Conhecê-las é equivalente a conhecer a curvatura de Gauss e a curvatura média. Este conhecimento, no entanto, é insuficiente para garantir que duas superfícies são únicas, pois o problema seria indeterminado: dados os autovalores da matriz $-\mathbf{I}^{-1}\mathbf{II}$, há muita liberdade para escolher as matrizes $\mathbf{I}$ e $\mathbf{II}$.

Em variedades de dimensão maior que 2, explicitar o operador de forma não é tão simples quanto se parece. Riemann introduziu uma maneira abstrata e rigorosa de definir curvatura para essas variedades, conhecida como o *tensor de curvatura de Riemann*. Noções semelhantes encontraram aplicação em toda a matemática, e mais especialmente na física quando Einstein apresentou a Teoria da Relatividade. Hoje em dia, em problemas de aprendizado de máquina e reconhecimento de padrões, a curvatura de Ricci vem sendo utilizada em algoritmos de agrupamento, conforme (RENGASWAMI; BOURNI; MAROULAS, 2024).

Recentemente, um novo trabalho propôs o cáculo da curvatura gaussiana em um ponto da variedade de dados utilizando uma aproximação local do operador de forma (LEVADA; NIELSEN; HADDAD, 2024). Partindo da hipótese de que pontos com alta curvatura necessitam de mais vizinhos para melhor descrever sua vizinhança, o método busca uma escolha adaptativa do parâmetro k do número de vizinhos do classificador k-NN. Ele utiliza uma aproximação do tensor de curvatura a partir da consideração da inversa da matriz de covariância Σ_i e da matriz Hessiana em um ponto x_i. Neste sentido, há ainda mais espaço para o estudo e implementação das ferramentas da geometria diferencial, juntamente do seu impacto nas aplicações no aprendizado de máquina e reconhecimento de padrões.

3.2 Aprendizado de Variedades

Conhecido na literatura como um processo de redução de dimensionalidade não-linear, o *aprendizado de variedades* compreende um arcabouço de técnicas focadas em reconstruir um conjunto de dados na sua versão de baixa dimensionalidade

de acordo com informações geométricas intrínsecas. Métodos de aprendizado de variedades como KPCA, LLE e ISOMAP inspiraram e consolidaram uma rica área de pesquisa com aplicações na maioria das áreas da ciência. De forma geral, reduir a dimensão de um conjunto de dados com o PCA consiste em assumir que os dados originais estão próximos a um subepaço linear (ou afim) de dimensão d contido em $\mathbb{R}^m$, com $d << m$. O método busca encontrar o subespaço linear que mais se aproxima da disposição dos dados. As técnicas de aprendizado de variedades generalizam esta ideia, assumindo que os dados originais estão próximos a uma variedade diferenciável M de dimensão d contida em $\mathbb{R}^m$, com objetivo de criar funções para reconstruir as amostras em M buscando preservar aspectos intrínsecos da geometria dos dados.

Na geometria diferencial, um *embedding* (ou mergulho) é uma aplicação suave entre duas variedades $f : \mathcal{M} \to \mathcal{N}$, cuja inversa $f^{-1} : f(\mathcal{M}) \subset \mathcal{N} \to \mathcal{M}$ existe e também é suave. Assim, a variedade $\mathcal{M}$ fica identificada com sua imagem $f(\mathcal{M})$, chamada de *subvariedade* em $\mathcal{N}$. Nas aplicações em aprendizado de máquina, os dados de alta dimensionalidade são representados originalmente em $\mathcal{N} = \mathbb{R}^m$. Entretanto, considera-se que eles foram gerados aleatóriamente por uma variedade $\mathcal{M}$ desconhecida e enviados via mergulho f para a subvariedade $f(\mathcal{M})$. Neste cenário, os algoritmos de aprendizado de variedades podem ser vistos como a busca pelo mergulho inverso f^{-1} e a reconstrução dos pontos em $\mathcal{M}$, dadas as observações em $f(\mathcal{M}) \subset \mathcal{N}$.

Definição 3.2.1 (**Hipótese de Variedade**)**.** Considere um conjunto de dados $X = \{x_i\}_{i=1}^n$, onde cada $x_i \in \mathbb{R}^m$. A hipótese de variedade consiste em assumir que $x_i = f(y_i)$, onde cada $y_i \in \mathcal{M}^d \subset \mathbb{R}^d$ é gerado por uma função de densidade suave P, com d desconhecido, $d < m$, e $f : \mathcal{M} \to \mathbb{R}^m$ um mergulho suave.

Definição 3.2.2 (**Aprendizado de métricas**)**.**

Definição 3.2.3 (**Aprendizado de Variedades**)**.** Considere $\mathcal{M}^d \subset \mathbb{R}^d$ uma variedade e $f : \mathcal{M} \to \mathbb{R}^m$ um mergulho, com $d < m$. Um algoritmo de aprendizado de variedades consiste em recuperar $\mathcal{M}$ e f de acordo com a hipótese de variedade sobre as observações $\{x_i\} \in \mathbb{R}^m$.

Desta maneira, o problema ainda necessita de restrições sobre f para relacionar a geometria dos dados observados à estrutura das variáveis ocultas $\{y_i\}$ e ao próprio $\mathcal{M}$. Na maioria dos algoritmos de aprendizado de variedades, são assumidas diversas condições sobre f. Por exemplo, se f é uma isometria, então preserva distâncias e ângulos infinitesimais. Se f é um mergulho conforme, então preserva ângulos, mas não distâncias. Cada hipótese adicional sobre a geometria local ou global determina alguma classe de métodos de aprendizado de variedades. Em tarefas de redução de dimensionalidade, d é desconhecida a priori e definida como parâmetro de entrada pelo usuário, embora existam métodos para estimar seu valor.

A principal vantagem desta abordagem é que os mergulhos descrevem as variedades de forma global, sem precisar de múltiplas cartas. Além disso, o Teorema do Mergulho de Whitney (Lee, 2003) afirma que toda variedade $\mathcal{M}^d$ pode ser mergulhada em $\mathbb{R}^{2d}$. Portanto, se for encontrado um mergulho válido, uma redução significativa da dimensionalidade pode ser alcançada. E este é um dos principais objetivos dos algoritmos de aprendizado de variedades. Além disso, toda a teoria das variedades, construída sob a modelagem de espaços localmente Euclidianos, pode ser utilizada para o estudo de conjuntos de dados complexos.

3.2.1 Principal Component Analysis

No caso em que se assume que a variedade é um subespaço Euclidiano, e a redução de dimensionalidade desejada é linear, sendo conhecida como o método *Principal Component Analysis* (PCA). Neste cenário, a hipótese de isometria é simplificada, pois a métrica Euclidiana é preservada diretamente no subespaço projetado, facilitando a interpretação geométrica dos dados projetados. Na Análise de Componentes Principais, realiza-se uma transformação de atributos do espaço original de forma a capturar o nível de relevância dos novos eixos de referência, denominados *componentes principais*. Em termos geométricos, a PCA compreende

uma translação e rotação nos eixos do espaço de atributos original, transformando-o para o espaço definido por eixos ortogonais e centrado na média observada. Além disso, os novos eixos são ordenados em função da variabilidade observada ao repojetar os dados sobre tal eixo. A grosso modo, busca-se filtrar os atributos que mais descrevem a disposição dos dados dentro do espaço original.

Considere um conjunto de dados $X = \{x_i \in \mathbb{R}^n : i = 1, \ldots, m\}$ representado em forma matricial por $X_{n \times m}$, onde cada coluna representa um ponto x_i como uma matriz $n \times 1$. O objetivo do PCA é encontrar uma matriz W (operador de projeção), que mapeia os pontos em X para um subespaço $Y = WX$ em R^d com $d < m$. Tal operador W busca projetar ortogonalmente X nas d direções de maior variabilidade dos dados.

Se $Z = [T^T, S^T]$ é uma base ortonormal para $\mathbb{R}^m$, com $T = [w_1, w_2, \ldots, w_d]$ e $S = [w_{d+1}, \ldots, w_m]$ representando o subespaço PCA (das direções de maior espalhamento) e o subespaço ortogonal desconsiderado, respectivamente, a condição de ortogonalidade implica que

$$w_i^T \cdot w_j = \begin{cases} 1 & \text{se } i = j \\ 0 & \text{se } i \neq j \end{cases}.$$

Logo, cada $x \in X$ é representado nesta base como

$$x = \sum_{j=1}^{d} c_j w_j = \sum_{j=1}^{d} (w^T \cdot w_j) \cdot w_j.$$

Isto é, o operador W age tomando as d primeiras linhas de x nesta base

$$y^T = x^T W^T = \sum_{j=1}^{d} c_j w_j^T [w_1, \ldots, w_k] = (c_1, \ldots, c_k).$$

A priori, W é desconhecido. Mas, a tese é que este maximiza a variância dos dados, ou seja, que maximiza

$$J^{PCA}(W) = E(\|y\|^2) = E(y^T y) = \sum_{j=1}^{d} E(c_j^2)$$

Portanto, o problema de otimização a ser resolvido é

$$W^* = \arg\max_{w} \sum_{j=1}^{d} w_j^T \Sigma_x w_j$$

sujeito a $\|w_j\|^2 = 1$ para $j = 1, \ldots, d$. Este é um problema de otimização com restrição de igualdade. Sua soluçao é encontrada através de multiplicadores de Lagrange:

$$\mathcal{L}(W, \lambda_1, \ldots, \lambda_k) = \sum_{j=1}^{d} w_j^T \Sigma_x w_j - \sum_{j=1}^{d} \lambda_j (w_j^T w_j - 1).$$

Derivando a Lagrangeana em relação a w_j e igualando a zero:

$$\frac{\partial \mathcal{L}}{\partial w_j} = 2\Sigma_x w_j - 2\lambda_j w_j = 0 \Rightarrow \frac{\partial \mathcal{L}}{\partial w_j} = \Sigma_x w_j - \lambda_j w_j = 0.$$

Isto implica que $\Sigma_x w_j = \lambda_j w_j$. Dado que Σ_x é simétrica, é garantido que existem n autovalores distintos cujos autovetores são ortogonais entre si. E como Σ_x também é positiva semi-definida, os autovalores são todos não negativos. Substituindo $\Sigma_x w_j = \lambda_j w_j$ na função objetivo $J_{PCA}(W)$:

$$J_{PCA}(W) = \sum_{j=1}^{d} w_j^T \Sigma_x w_j = \sum_{j=1}^{d} \lambda_j w_j^T w_j = \sum_{j=1}^{d} \lambda_j, \quad \text{com } \|w_j\|^2 = 1.$$

Portanto, os vetores w_j da nova base que maximizam a variância dos dados transformados são os autovetores da matriz de covariância Σ_X. É importante notar que, após essa transformação, os dados se encontram decorrelacionados. Ou seja, a matriz de covariância Σ_Y, onde Y é o resultado da aplicação da transformação T em X, é diagonal pela decomposição em autovalores da matriz Σ_X, onde $\text{diag}(\lambda_1, \ldots, \lambda_n)$ é a matriz diagonal dos valores próprios de Σ_X. Em resumo, a PCA consiste em tomar os d autovetores da matriz de covariância associados aos d maiores autovalores.

Pode ser mostrado que o PCA também miniza o critério do MSE. Nessa abordagem, busca-se um conjunto de d vetores ortonormais de base que gerem um subespaço d-dimensional tal que o MSE entre o vetor original x e sua projeção nesse subespaço seja mínima. A seguir um pseudocódigo para a PCA.

3.2.2 Kernel PCA

Se o PCA busca por componentes principais no espaço de entrada, o KPCA (SCHÖLKOPF; SMOLA; MÜLLER, 1997) busca por componentes principais que estão relacionadas de forma não linear com as variáveis de entrada. Generalizando

Algoritmo 1 Principal Component Analysis (PCA)

1: **function** PCA(X,d)
2: Calcule a média e a matriz de covariância Σ_X.
3: Calcule os n autovalores λ_j e autovetores w_j de Σ_X.
4: Ordene os autovetores de tal forma que $\lambda_j \geq \lambda_{j+1}$.
5: Defina a matriz de transformação $W = [w_1, w_2, \ldots, w_d]$.
return $Y = WX$

a análise de componentes principais, a técnica incorpora transformações não lineares por meio do *kernel trick*. O kernel trick é a consideração de uma aplicação não linear que mapeia uma observação x em $\Phi(X)$, onde $\Phi : \mathbb{R}^m \to \mathbb{R}^{m'}$ onde $m' > m$, que é utilizada para calcular o produto interno das observações no espaço aumentado. Este método nos permite construir diferentes versões não lineares da PCA, mudando apenas função kernel utilizada. A primeira utilização desta ideia se deu com as *Support Vector Machines* (CORTES; VAPNIK, 1995) ao propor o aprendizado de regiões de decisão não lineares.

Considere um conjunto de dados X e um kernel positivo definido k em X, isto é, uma função de valor real $k : X \times X \to \mathbb{R}$ com a propriedade de que existe uma transformação $\Phi : X \to \mathcal{H}$ em um espaço de produto interno $\mathcal{H}$ tal que, para todos $x, y \in X$, tem-se $\langle \Phi(x), \Phi(y) \rangle = k(x, y)$. O núcleo k pode ser entendido como uma medida de similaridade não linear em um espaço com produto interno. O KPCA começa calculando as componentes principais dos pontos $\Phi(x_1), \ldots, \Phi(x_m)$.

Suponha, por enquanto, que a média dos dados no espaço de características seja zero. A matriz de covariância é dada por:

$$\Sigma_{\Phi(X)} = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i)^T \phi(x_i)$$

Mas como $\mathcal{H}$ pode ser de dimensão infinita, o problema de PCA precisa ser transformado em um problema que possa ser resolvido em termos do núcleo k. Com isso, para diagonalizar a matriz de covariâncias $\Sigma_{\Phi(X)}$, mesmo que $\mathcal{H}$ seja de dimensão infinita, observa-se primeiro que todas as soluções para

$$\Sigma_{\Phi(X)}(v) = \lambda v$$

com $\lambda \neq 0$ devem estar no conjunto gerado pelas imagens Φ de X. Assim, expande-

se a solução v como

$$v = \sum_{i=1}^{n} \alpha_i \Phi(x_i)$$

reduzindo, assim, o problema a encontrar os α_i. O último pode ser mostrado tomando a forma

$$n\lambda\alpha = K\alpha,$$

onde $\alpha = (\alpha_1, \ldots, \alpha_n)^T$ e K é uma matriz $n \times n$ tal que $K_{ij} = k(x_i, x_j)$. Além disso, para um ponto x, sua projeção na p-ésima componente principal é dada por

$$\phi(x)^T v_p = \sum_{i=1}^{n} \alpha_{pi} \phi(x)^T \phi(x_i) = \sum_{i=1}^{n} \alpha_{pi} k(x, x_i)$$

Em geral, não é valido assumir que a média das observações no espaço de característica é zero, e, portanto, é necessário subtrair a média $(1/n) \sum \Phi(x_i)$ de todos os pontos. Isso leva a um problema de autovalor ligeiramente diferente, onde diagonaliza-se

$$K' = (I - e^T e)K(I - e^T e),$$

onde $e = m^{-1/2}(1, \ldots, 1)$, ao invés de K.

Algoritmo 2 Kernel Principal Component Analysis (KPCA)

1: **function** KPCA(X,d,k)
2: Construa a matriz kernel normalizada K' dos dados.
3: Calcule os autovalores λ_j e autovetores v_j dessa matriz.
4: Ordene os autovetores de tal forma que $\lambda_j \geq \lambda_{j+1}$.
5: Para cada x, calcule a j-ésima componente principal de $\Phi(x)$ via

$$y_j = \sum_{i=1}^{n} \alpha_{ji} k(x, x_i), \quad \text{para } j = 1, \ldots, d.$$

return Y

A escolha do kernel apropriado depende da natureza dos dados e da relação a ser extraída. Entre várias possibilidades conhecidas de funções kernel, as mais comuns são: *linear, polinomial, gaussiano, cosseno* e *sigmoidal*. O kernel gaussiano, também conhecido como *kernel RBF* (Radial Basis Function), foi utilizado neste trabalho. Sendo definido por $K(x, x') = \exp(-\sigma\|x - x'\|^2)$, seus valores variam entre 0 e 1 e são decrescentes de acordo com a distância entre x e x', ou seja,

pontos próximos tendem a ser mais similares que pontos distantes. O hiperparâmetro σ controla o fator de escala na ponderação de distâncias locais. Quando não especificado pelo usuário, a biblioteca do `scikit-learn` considera $\sigma = 1/n$.

3.2.3 Locally Linear Embedding

O *Locally Linear Embedding* (LLE) (ROWEIS; SAUL, 2000) é um método de redução de dimensionalidade não linear local que busca encontrar as novas coordenadas de qualquer ponto x com base apenas na vizinhança desse ponto. A hipótese principal por trás do LLE é que, para uma densidade suficientemente grande de amostras, é esperado que o vetor x e seus vizinhos formem um patch linear, ou seja, pertençam todos a um subespaço euclidiano. Em outras palavras, o LLE busca encontrar isometrias locais na variedade de dados $\mathcal{M}^d$.

Partindo de um conjunto de observações $X \subset \mathbb{R}^m$ e uma dimensionalidade $d < m$, o funcionamento do LLE é sumarizado como segue:

(i) Para cada $x_i \in X$, encontre seus k vizinhos mais próximos.

(ii) Encontre a matriz de pesos W de tamanho $n \times n$ que minimiza o erro de reconstrução local, determinada pelo seguinte problema de otimização:

$$\min_W \sum_{i=1}^{n} \left\| x_i - \sum_{j=1}^{n} w_{ij} x_j \right\|^2 \tag{3.6}$$

$$\text{sujeito a: } \begin{cases} w_{ij} = 0 & \text{se } x_j \notin \mathcal{V}_k(x_i) \\ \sum_j w_{ij} = 1. \end{cases} \tag{3.7}$$

(iii) Encontre as coordenadas Y que minimizam o erro de reconstrução, resolvendo o seguinte problema de otimização:

$$E(w) = \min_Y \sum_{i=1}^{n} \left\| y_i - \sum_j w_{ij} y_j \right\|^2 \tag{3.8}$$

$$\text{sujeito a: } \begin{cases} \sum_i Y_{ij} = 0 & \text{para cada } j \\ Y^T Y = I. \end{cases} \tag{3.9}$$

3.2.3.1 Encontrar a matriz de reconstrução local

Minimizar $E(w)$ equivale a minimizar cada erro local. Dado x_i,

$$E(w) = \Big\| x_i - \sum_j w_j x_j \Big\|^2 \tag{3.10}$$

$$= \Big\| \Big(\sum_j w_j \Big) x_i - \sum_j w_j x_j \Big\|^2 \tag{3.11}$$

$$= \Big\| \sum_j w_j (x_i - x_j) \Big\|^2 \tag{3.12}$$

$$= \sum_k \sum_j w_j w_k (x_i - x_j)^T (x_i - x_k) = w^T C_i w \tag{3.13}$$

onde $C_i = (x_i - x_j)^T (x_i - x_k)$ representa a matriz de covariância local de x_i. Portanto, minimiza-se $E(w) = w^T C_i w$ para $i = 1, \ldots, n$, com a condição $\sum_j w_j = 1$, que garante a invariância por translação dos dados. Se considerarmos a restrição de que os pesos somam 1, ou seja, $\sum_j w_j = 1$, então geometricamente garante-se de que o erro de reconstrução é invariante a translação. De fato, seja $c \in \mathbb{R}^m$ e w o peso do erro $\|x_i + c\|$,

$$E(w) = \Big\| (x_i + c) - \sum_j w_j (x_j + c) \Big\|^2 \tag{3.14}$$

$$= \Big\| x_i + c - \sum_j w_j x_j - c \sum_j w_j \Big\|^2 \tag{3.15}$$

$$= \Big\| x_i + c - \sum_j w_j x_j - c \Big\|^2 = E(w). \tag{3.16}$$

O problema de estimar os pesos de reconstrução ótimos por mínimos quadrados pode ser expresso por n problemas de otimização independentes definidos por:

$$\arg\min_{w_i} w_i^T C_i w_i \quad \text{sujeito a} \quad \vec{1}^T w_i = 1 \quad \text{para } i = 1, \ldots, n$$

A solução tem uma fórmula fechada, dada por:

$$w_i = \frac{\lambda}{2} C_i^{-1} \vec{1}$$

Utilizando multiplicadores de Lagrange:

$$\mathcal{L}(w_i, \lambda) = w_i^T C_i w_i - \lambda(\vec{1}^T w_i - 1) \tag{3.17}$$

$$\Rightarrow \frac{\partial \mathcal{L}}{\partial w_i} = 0 \tag{3.18}$$

$$\Rightarrow 2C_i w_i - \lambda\vec{1} = 0 \tag{3.19}$$

$$\Rightarrow C_i w_i = \frac{\lambda}{2}\vec{1} \tag{3.20}$$

$$\Rightarrow w_i = \frac{\lambda}{2} C_i^{-1}\vec{1} \tag{3.21}$$

Neste caso, λ pode ser escolhido para garantir que $\sum_j w_{ij} = 1$. Mas na prática, considera-se n sistemas lineares $C_i w_i = \vec{1}$. E normaliza-se a solução depois, fazendo

$$w_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}.$$

fazendo com que o erro a ser minimizado seja dado por:

$$E(w) = \left\| x_i - \sum_j w_j x_j \right\|^2 + \alpha \sum_j w_j^2$$

onde α é um controle do nível de regularização. Nesse caso, para a estimação dos pesos de reconstrução ótimos, serão n problemas de otimização com restrições, dado por:

$$\arg\min_{w_i} w_i^T C_i w_i + \alpha w_i^T w_i \quad \text{sujeito a} \quad \vec{1}^T w_i = 1 \quad \text{para } i = 1, \ldots, n$$

Mostre que a solução tem fórmula fechada, dada por:

$$w_i = \frac{\lambda}{2}(C_i + \alpha I)^{-1}\vec{1}$$

A Lagrangeana fica:

$$\mathcal{L}(w_i, \lambda) = w_i^T C_i w_i + \alpha w_i^T w_i - \lambda(\vec{1}^T w_i - 1)$$

$$\Rightarrow \frac{\partial \mathcal{L}}{\partial w_i} = 0 \Rightarrow 2C_i w_i + 2\alpha w_i - \lambda\vec{1} = 0$$

$$\Rightarrow (C_i + \alpha I) w_i = \frac{\lambda}{2}\vec{1}$$

$$\Rightarrow w_i = \frac{\lambda}{2}(C_i + \alpha I)^{-1}\vec{1}$$

3.2.3.2 Encontrar coordenadas que minizam o erro de reconstrução

Para encontrar as coordenadas no $\mathbb{R}^d$, é minimizado o erro de reconstrução. Fixada a matriz de pesos W:

$$\Phi(Y) = \sum_{i=1}^{n} \left\| y_i - \sum_j w_{ij} y_j \right\|^2$$

Nas coordenadas de saída, é necessário que

$$\frac{1}{n} \sum_{i=1}^{n} y_i = 0$$

e que a matriz de covariância dos y_j seja a identidade. O que nos garante que $\langle y_i, y_j \rangle = \delta_{ij}$ (são ortogonais). Notando que não dá para minimizar $\Phi(Y)$ via minimização das n componentes da somatória, pois y_i está presente em cada componente, reescreve-se $\Phi(Y)$ na forma matricial:

$$\Phi(Y) = \sum_{i=1}^{n} \left(y_i - \sum_j w_{ij} y_j \right)^T \left(y_i - \sum_j w_{ij} y_j \right)$$

$$= \sum_{i=1}^{n} \left[y_i^T y_i - y_i^T \sum_j w_{ij} y_j - \left(\sum_j w_{ij} y_j \right)^T y_i + \sum_{j,k} w_{ij} w_{ik} y_j^T y_k \right]$$
$$= \sum_{i=1}^{n} y_i^T y_i - \sum_{i=1}^{n} \sum_{j=1}^{n} y_i^T w_{ij} y_j - \sum_{i=1}^{n} \sum_{j=1}^{n} y_j^T w_{ij} y_i + \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} y_j^T w_{ij} w_{ik} y_k.$$

É possível reescrever os somatórios em forma matricial considerando o operador traço de uma matriz. Utilizando propriedades deste mesmo operador, simplifica-se a escrita de $\Phi(Y)$:

$$\begin{aligned}
\Phi(Y) &= \mathrm{Tr}(Y^T Y) - \mathrm{Tr}(Y^T W Y) - \mathrm{Tr}(Y^T W^T Y) + \mathrm{Tr}(Y^T W^T W Y) \\
&= \mathrm{Tr}(Y^T Y) - \mathrm{Tr}(Y^T (W + W^T) Y) + \mathrm{Tr}(Y^T W^T W Y) \\
&= \mathrm{Tr}(Y^T [(I - W)^T (I - W)] Y) \\
&= \mathrm{Tr}(Y^T (I - W)^T (I - W) Y) \\
&= \mathrm{Tr}(Y^T M Y)
\end{aligned}$$

E o problema de otimização fica

$$\arg\min_{Y} \mathrm{Tr}(Y^T M Y) \quad \text{sujeito a} \quad \frac{1}{n} Y^T Y = I.$$

Após descobrir que a matriz $Y_{n\times k}$ deve conter os k autovetores de M associados aos menores autovalores, nota-se que o menor autovalor é sempre zero, com autovalor associado sendo o vetor $\vec{1} = [1, \ldots, 1] \in R^d$. De fato, dado que cada linha da matriz Wdeve somar 1, na forma matricial tem-se

$$W\vec{1} = \vec{1} \iff (I - W)\vec{1} = \vec{0} \implies (I - W)^T(I - W)\vec{1} = 0 \iff M\vec{1} = \vec{0}.$$

Mostrando assim que o vetor constante $\vec{1}$ é um autovetor de M com autovalor zero. Logo, a solução do problema estabelecido em (iii) pode ser obtida com a decomposição em autovetores da matriz $(I-W)^T(I-W)$, seguido pelo descarte do autovetor associado ao menor autovalor e, posteriormente, considerando apenas os autovetores associados aos d menores autovalores. Este processo define a matriz Y de dimensão $n \times d$, cujas linhas representam os padrões projetados. Por outro lado, a estimação dos pesos por mínimos quadrados é um problema mal-condicionado se o número de vizinhos k é maior que a dimensionalidade dos dados d. Nesse caso, regulariza-se o problema adicionando um termo de penalização.

Algoritmo 3 Locally Linear Embedding (LLE)

1: **function** LLE(X,d,k)
2: Encontre os k vizinhos mais próximos para cada x_i.
3: Encontre a matriz W de reconstrução de pesos locais.
4: Calcule os autovetores e autovalores de $(I - W)^T(I - W)$.
5: Ordene os autovetores de tal forma que $\lambda_j \leq \lambda_{j+1}$.
6: Descarte o autovetor com menor valor associado.
7: Defina $Y = [v_1, \ldots, v_d]$ com os d menores autovetores nas colunas.
return Y

3.2.4 Isometric Feature Mapping

O algoritmo *Isometric Feature Mapping* (ISOMAP) (TENENBAUM; SILVA; LANGFORD, 2000) foi proposto como um método de redução de dimensionalidade não linear que lida com o paradigma de isometria global da variedade de dados. Ele pode ser visto como uma extensão do método *Multidimensional Scaling* (MDS) (COX; COX, 2000), que preserva distâncias ao mesmo tempo que encontra uma representação em baixa dimensionalidade. A principal ideia do ISOMAP é utilizar distâncias sobre caminhos mínimos no grafo de vizinhanças para aproximar a distância geodésica na variedade. Deste modo, pode ser visto como um método que busca recuperar $\mathcal{M}^d$ a partir da hipótese de isometria global. Uma etapa crucial no algoritmo envolve estimar a desconhecida distância geodésica na variedade. Utilizando as distâncias euclidianas de algum grafo de vizinhanças construído com os pontos de dados, os caminhos mínimos do grafo são obtidos e utilizados para redução via MDS.

Na teoria dos Grafos, sempre é possível descobrir o caminho mínimo entre dois dados vértices de um grafo com pesos não-negativos. Isto é atingido com auxílio do algoritmo de Dijkstra (DIJKSTRA, 1959) ou o algoritmo de Floyd-Warshall (FLOYD, 1962). Em posse da árvore de menores caminhos de s até qualquer $v \in V$, define-se a matriz $D = \{d_{rs}^2\}$ para $r, s = 1, \ldots, n$ das distância do vértice r ao vértice s. Essa matriz possui distâncias a partir de um $\mathbb{R}^m$, e para estimar as coordenadas dos novos vértices em um $\mathbb{R}^d$ que preservem as distâncias em D, é utilizado o método MDS. Basicamente, tal método é divido em duas etapas:

(i) Encontrar a matriz de produtos internos $B = \{b_{rs}\}$ a partir da matriz D

(ii) Estimar as coordenadas em $\mathbb{R}^k$ que satisfazem B.

Em alto nível, diante de uma aproximação grande o suficiente da variedade de dados, o ISOMAP é capaz de recuperar a distância geodésica da variedade subjacente, como afirma o seguinte teorema (BERNSTEIN et al., 2000).

Teorema 3.2.1. (Convergência Assintótica) Sejam $\lambda_1, \lambda_2, \mu > 0$ quaisquer. Para um conjunto de dados suficientemente grande $x_1, \vec{x}_2, ..., x_n$, onde $x_i \in \mathbb{R}^m$,

temos que a desigualdade

$$(1-\lambda_1)d_M(x_i,x_j) \leq d_G(x_i,x_j) \leq (1+\lambda_2)d_M(x_i,x_j) \tag{3.22}$$

é válida, com probabilidade $(1-\mu)$, onde $d_G(x_i,x_j)$ é o comprimento dos menores caminhos no grafo e $d_M(x_i,x_j)$ é a verdadeira distância geodésica desconhecida na variedade.

3.2.4.1 Cálculo da matriz de produtos internos

Seja $D = \{d_{rs}^2\}$, para $r,s = 1,2,...,n$ a matriz de distâncias gedésicas ponto a ponto :

$$d_{rs}^2 = \|x_r - x_s\|^2 = (x_r - x_s)^T(x_r - x_s) \tag{3.23}$$

Seja B a matriz dos produtos internos, ou seja, $B = \{b_{rs}\}$, com $b_{rs} = x_r^T x_s$. É necessário assumir a hipótese de que os dados encontram-se centralizados, ou seja, possuem média nula:

$$\sum_{r=1}^{n} x_r = 0 \tag{3.24}$$

caso contrário haveriam infinitas soluções possíveis, uma vez que a aplicação de qualquer translação arbitrária no conjunto de pontos iria preservar as distâncias ponto-a-ponto. Da equação (3.23), aplicando a distributiva, vale:

$$d_{rs}^2 = x_r^T x_r + x_s^T x_s - 2x_r^T x_s \tag{3.25}$$

A partir da matriz D, calcula-se a média de uma coluna arbitrária s como:

$$\begin{aligned}
\frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 &= \frac{1}{n}\sum_{r=1}^{n} x_r^T x_r + \frac{1}{n}\sum_{r=1}^{n} x_s^T x_s - 2\frac{1}{n}\sum_{r=1}^{n} x_r^T x_s \\
&= \frac{1}{n}\sum_{r=1}^{n} x_r^T x_r + \frac{n}{n} x_s^T x_s - 2x_s\frac{1}{n}\sum_{r=1}^{n} x_r^T \\
&= \frac{1}{n}\sum_{r=1}^{n} x_r^T x_r + x_s^T x_s
\end{aligned} \tag{3.26}$$

Analogamente, calcula-se a média de uma linha arbitrária r como:

$$\begin{aligned}\frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 &= \frac{1}{n}\sum_{s=1}^{n} x_r^T x_r + \frac{1}{n}\sum_{s=1}^{n} x_s^T x_s - 2\frac{1}{n}\sum_{s=1}^{n} x_r^T x_s \\ &= \frac{n}{n} x_r^T x_r + \frac{1}{n}\sum_{s=1}^{n} x_s^T x_s - 2x_r^T \frac{1}{n}\sum_{s=1}^{n} x_s \\ &= x_r^T x_r + \frac{1}{n}\sum_{s=1}^{n} x_s^T x_s \end{aligned} \tag{3.27}$$

Por fim, é obtida a média dos elementos de D como:

$$\begin{aligned}\frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2 &= \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} x_r^T x_r + \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} x_s^T x_s - 2\frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} x_r^T x_s \\ &= \frac{1}{n}\sum_{r=1}^{n} x_r^T x_r + \frac{1}{n}\sum_{s=1}^{n} x_s^T x_s = \frac{2}{n}\sum_{r=1}^{n} x_r^T x_r \end{aligned} \tag{3.28}$$

Note que a partir da equação (3.25), é possível definir b_{rs} como:

$$b_{rs} = x_r^T x_s = -\frac{1}{2}(d_{rs}^2 - x_r^T x_r - x_s^T x_s) \tag{3.29}$$

Mas, da equação (3.27), é isolado o termo $-x_r^T x_r$ como:

$$-x_r^T x_r = -\frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 + \frac{1}{n}\sum_{s=1}^{n} x_s^T x_s \tag{3.30}$$

E da equação (3.26) isola-se o termo $-x_s^T x_s$ como:

$$-x_s^T x_s = -\frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 + \frac{1}{n}\sum_{r=1}^{n} x_r^T x_r \tag{3.31}$$

Calculando a diferença entre a equação (3.30) e a equação (3.31):

$$-x_r^T x_r - x_s^T x_s = -\frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 - \frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 + \frac{2}{n}\sum_{r=1}^{n} x_r^T x_r \tag{3.32}$$

Da equação (3.28) sabe-se que:

$$\frac{2}{n}\sum_{r=1}^{n} x_r^T x_r = \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2 \tag{3.33}$$

Finalmente, a matriz b_{rs} é expressada como uma função dos elementos da matriz de distâncias D:

$$b_{rs} = -\frac{1}{2}\left(d_{rs}^2 - \frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 - \frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 + \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2\right) \tag{3.34}$$

Fazendo $a_{rs} = -\frac{1}{2}d_{rs}$ é obtido:

$$a_{r.} = \frac{1}{n}\sum_{s=1}^{n} a_{rs} \qquad a_{.s} = \frac{1}{n}\sum_{r=1}^{n} a_{rs} \qquad a_{..} = \frac{1}{n}\sum_{r=1}^{n}\sum_{s=1}^{n} a_{rs} \tag{3.35}$$

de modo a ter $b_{rs} = a_{rs} - a_{r.} - a_{.s} + a_{..}$ Definindo a matriz $A = \{a_{rs}\}$, para $r, s = 1, 2, ..., n$ como $A = -\frac{1}{2}D$ e a matriz H como:

$$H = I - \frac{1}{n}\vec{1}\vec{1}^T, \tag{3.36}$$

é possível computar todos os elementos de B simultaneamente por $B = HAH$.

3.2.4.2 Estimação das coordenadas intrínsecas

A matriz de produtos internos B pode ser escrita como:

$$B_{n\times n} = X^T_{n\times m} X_{m\times n} \tag{3.37}$$

onde n e m são o número de amostras e o número de atributos e $X_{n\times m} = [x_1, x_2, ..., x_n]$ é a matriz de dados. Pode-se mostrar que a matriz B possui m autovalores não-negativos e $n - m$ autovalores nulos. Pela decomposição espectral de B, tem-se:

$$B = V\Lambda V^T \tag{3.38}$$

onde $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_m)$ é a matriz diagonal $m \times m$ dos dos autovalores não nulos de B e V é a matriz $n \times m$ cujas colunas são os m autovetores associados aos m autovalores não nulos. Assim, tem-se

$$B = X^T X = V\Lambda V^T = V\Lambda^{1/2}\Lambda^{1/2}V^T \tag{3.39}$$

o que leva finalmente a:

$$X = \Lambda^{1/2}V^T \tag{3.40}$$

onde $\Lambda^{1/2} = diag(\sqrt{\lambda_1}, \sqrt{\lambda_2}, ..., \sqrt{\lambda_m})$ denota a raiz quadrada dos autovalores (valores singulares). O algoritmo a seguir descreve os passos do ISOMAP.

Algoritmo 4 Isometric Feature Mapping

function ISOMAP(X)
    A partir da matriz de dados $X_{m\times n}$ construa um grafo k-NN.
    Compute a matriz de distâncias geodésicas do grafo $D_{n\times n}$.
    Compute $H = I - \frac{1}{n}U$, onde U é uma matriz $n \times n$ de 1's.
    Compute $B = -\frac{1}{2}HDH$.
    Encontre os autovalores e autovetores da matriz B.
    Selecione os $d < m$ maiores autovalores e seus autovetores
    Defina as matrizes Λ e V, conforme definidas anteriormente.
    return $\tilde{X} = \Lambda^{1/2}V^T$

Conforme a figura **??**, o método ISOMAP consegue capturar a estrutura métrica local e global da superfície S. Embora funcione neste conjunto de dados sintético, o método possui algumas limitações quando aplicado a conjunto de dados mais gerais. A quantidade k de vizinhos a ser considerada ou a presença de ruído nos dados pode afetar a conectividade do grafo e comprometer as distâncias geodésicas obtidas. Uma vez que a ponderação das arestas é calculada pela distância euclidiana, se k é muito pequeno, o grafo pode se tornar esparso demais para calcular boas aproximações das distâncias geodésicas. Se k é muito grande, pode acabar distorcendo a estrutura de variedade.

A complexidade computacional do ISOMAP é $O(n^3)$, com o maior custo computacional sendo o cálculo de todos os pares de distâncias do caminho mais curto. Como o ISOMAP trabalha com matrizes densas, essa complexidade de espaço não pode ser melhorada. Existem variantes do Isomap que melhoram o método de diferentes formas: o *Hessian Eigenmap* (DONOHO; GRIMES, 2003) permite lidar com dados não convexos, ao introduzir o uso do operador Hessiano.

3.2.5 Laplacian Eigenmaps

O Laplacians Eingenmaps (LE) originalmente foi proposto como um algoritmo de agrupamento, conhecido como *Agrupamento Espectral* (NG; JORDAN; WEISS, 2001). Sua fundamentação como método de aprendizado de variedades aconteceu logo depois por (BELKIN; NIYOGI, 2003). Em resumo, o Laplacian Eigenmaps busca preservar as localidades de um grafo de vizinhanças do espaço original $\mathbb{R}^m$ na sua transformação para no espaço de baixa dimensão $\mathbb{M}^d$. O método estuda os

autovetores da *Laplaciana do grafo* L, que graças a suas propriedades matemáticas, admite fórmula analítica fechada. A matriz L tem correspondência direta com o operador de Laplace-Beltrami na variedade, cujo espectro possui informações geométricas fundamentais.

O LE recebe um conjunto de dados $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^m$, a dimensão $d < m$ e o parâmetro do kernel t. Seus passos principais são:

(i) Construir um grafo $G = (V, E)$ baseando-se no método KNN ou ϵ-ball.

(ii) Definir a matriz de adjacência $W(t)$ do grafo G via Heat Kernel com parâmetro $t \in \mathbb{R}$

$$w_{ij} = \exp\left\{ -\frac{\|x_i - x_j\|^2}{t} \right\}$$

caso x_i e x_j estejam conectados por uma aresta ou $w_{ij} = 0$ caso contrário. Também poderiam ser definidos outros pesos, sem a necessidade do parâmetro t, como por exemplo $w_{ij} = 1$ se x_i e x_j estiverem conectados ou $w_{ij} = 0$ caso contrário.

(iii) Resolver o problema de otimização

$$\arg\min_y \; y^T L y \quad \text{sujeito a} \quad y^T D y = 1$$

para encontrar as coordenadas Y da imersão que são dadas pelos d autovetores associados aos d menores autovalores não nulos da matriz Laplaciana (não-normalizada) $L = D - W$. Aqui, D é a matriz diagonal que contém o grau de cada vertice do grafo, ou seja, $d_i = \sum_{j=1}^{n} w_{ij}$.

Note que, para todo $\vec{f} \in \mathbb{R}^n$, tem-se

$$\vec{f}^T L \vec{f} = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (f_i - f_j)^2.$$

De fato,

$$\begin{aligned}
\vec{f}^T L \vec{f} &= \vec{f}^T D \vec{f} - \vec{f}^T W \vec{f} \\
&= \sum_{i=1}^{n} d_i f_i^2 - \sum_{i=1}^{n}\sum_{j=1}^{n} f_i w_{ij} f_j \\
&= \frac{1}{2}\left(2\sum_{i=1}^{n} d_i f_i^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{n} f_i w_{ij} f_j\right) \\
&= \frac{1}{2}\left(\sum_{i=1}^{n} d_i f_i^2 + \sum_{j=1}^{n} d_j f_j^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{n} f_i w_{ij} f_j\right) \\
&= \frac{1}{2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} f_i^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{n} f_i w_{ij} f_j + \sum_{j=1}^{n}\sum_{i=1}^{n} w_{ij} f_j^2\right) \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(f_i^2 - 2 f_i f_j + f_j^2) \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(f_i - f_j)^2.
\end{aligned}$$

Logo, se x_i e x_j não estão conectados, $w_{ij} = 0$, e portanto $(f_i - f_j)^2$ não aparecerá na soma. Além disso, o termo $(f_i - f_j)^2$ representa a distância ao quadrado na reta real. Logo, ao mapear $x_i \to f_i \in \mathbb{R}$, de uma forma a minimizar a forma quadrática $\vec{f}^T L \vec{f}$, preserva-se a localidade, no sentido de que pontos do grafo que estão próximos permanecerão próximos na reta.

A solução para o problema de encontrar o vetor y que minimiza $J(y) = y^T L y$ sujeito à restrição $y^T D y = 1$ (para remover um fator de escala arbitrário) é que y deve ser o autovetor associado ao menor autovalor não-nulo de $D^{-1}L$. Note também que o menor autovalor da matriz laplaciana L é sempre zero e o autovetor correspondente é o vetor constante $\vec{1}$.

Resolvendo o problema de otimização com a restrição de igualdade

$$\arg\min y^T L y \quad \text{sujeito a} \quad y^T D y = 1,$$

obtem-se a Lagrangeana

$$\begin{aligned}
\mathcal{L}(y, \lambda) &= y^T L y - \lambda(y^T D y - 1) \\
\Rightarrow \frac{\partial \mathcal{L}}{\partial y} = 0 \Rightarrow 2Ly - 2\lambda D y = 0 \Rightarrow Ly &= \lambda D y
\end{aligned}$$

Portanto, a solução para o problema deve ser o autovetor y associado ao menor autovalor não-nulo de $D^{-1}L$ (Laplaciano Normalizado), sendo de descartado o autovetor $\vec{1}$ associado ao autovalor nulo.

No problema generalizado, em que deseja-se realizar uma imersão em $\mathbb{R}^d$, tem-se uma matriz $n \times d$

$$Y = [y_1, y_2, \ldots, y_n]^T$$

em que a i-ésima linha $y^{(i)}$ define as coordenadas de um ponto. A função objetivo a ser minimizada é generalizada para:

$$J(Y) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} \|y^{(i)} - y^{(j)}\|^2$$

Essa função objetivo pode ser expressa por

$$J(Y) = \mathrm{Tr}(Y^T L Y).$$

É possível mostrar que esta função objetivo é expressada como

$$J(Y) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} \|y_i - y_j\|^2 = \mathrm{Tr}(Y^T L Y).$$

De fato,

$$\begin{aligned} J(Y) &= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} \|y_i - y_j\|^2 \\ &= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} (y_i^T y_i - 2 y_i^T y_j + y_j^T y_j) \\ &= \mathrm{Tr}(Y^T L Y) \end{aligned}$$

$$\begin{aligned}
\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}\|y_i - y_j\|^2 &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(y_i^T y_i - 2y_i^T y_j + y_j^T y_j) \\
&= \frac{1}{2}\sum_{i=1}^{n} d_i y_i^T y_i - \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij} y_i^T y_j + \frac{1}{2}\sum_{j=1}^{n} d_j y_j^T y_j \\
&= \sum_{i=1}^{n} d_i y_i^T y_i - \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij} y_i^T y_j \\
&= \sum_{i=1}^{n} d_i y_i^T y_i - \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij} y_i^T y_j \\
&= \mathrm{Tr}(Y^T DY) - \mathrm{Tr}(Y^T WY) \\
&= \mathrm{Tr}(Y^T (D - W)Y) = \mathrm{Tr}(Y^T LY).
\end{aligned}$$

Isto implica que

$$\begin{aligned}
J(Y) &= \mathrm{Tr}(Y^T DY) - \mathrm{Tr}(Y^T WY) \\
&= \mathrm{Tr}(Y^T (DY - WY)) \\
&= \mathrm{Tr}(Y^T (D - W)Y) = \mathrm{Tr}(Y^T LY)
\end{aligned}$$

A Lagrangeana deste problema é dada por

$$\mathcal{L}(Y, \Lambda) = \mathrm{Tr}(Y^T LY) - \mathrm{Tr}(\Lambda(Y^T DY - I)),$$

onde $Y = [y_1, \ldots, y_n]^T_{n \times d}$. A derivada de $\mathrm{Tr}(Y^T LY)$ em relação a Y é dada por

$$LY + LY = 2LY,$$

pois L é simétrica. Logo,

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial Y} = 0 &\Rightarrow 2LY - 2\Lambda DY = 0 \\
&\Rightarrow LY = \Lambda DY \\
&\Rightarrow (D^{-1}L)Y = \Lambda Y
\end{aligned}$$

Portanto, as colunas de Y devem ser os d autovetores associados aos d menores autovalores de $D^{-1}L$.

Algoritmo 5 Laplacian Eigenmaps

function LE(X,d,k,t)
    Construa um grafo k-NN.
    Defina o kernel Gaussiano $W(t)$.
    Calcule a matriz Laplaciana normalizada $D^{-1}L$.
    Encontre os $d+1$ menores autovalores e autovetores da matriz $D^{-1}L$.
    Descarte o autovetor associado ao menor autovalor.
    Defina $Y = [v_1, \ldots, v_d]$ com os d menores autovetores nas colunas.
    return Y

3.2.6 t-Distributed Stochastic Neighbor Embedding

O t-SNE é uma versão adaptada de seu antecessor e método motivacional: Stochastic Neighbor Embedding (SNE) (HINTON; ROWEIS, 2002). O SNE começa convertendo distâncias Euclidianas entre pontos de dados $x_i \in \mathbb{R}^m$ para $i = 1, 2, \ldots, n$, em probabilidades condicionais como uma forma de representar similaridade. A similaridade entre dois pontos x_i e x_j é a probabilidade condicional $p_{j|i}$ de que x_i escolheria x_j como vizinho, se a seleção do vizinho for feita em proporção a uma densidade de probabilidade sob uma Gaussiana centrada em x_i.

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2/2\sigma_i^2\right)}$$

onde σ_i^2 é a variância da Gaussiana centrada em x_i. Note que, se x_i e x_j estão suficientemente próximos, então $p_{j|i}$ tem um valor significativamente alto, mas se estiverem amplamente separados, $p_{j|i}$ tende a zero. É possível calcular uma probabilidade condicional similar para os pontos transformados y_i e y_j em $\mathbb{R}^d$, denotada por $q_{j|i}$. Definindo a variância da Gaussiana para $1/\sqrt{2}$, a medida de similaridade é dada por

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)}.$$

A ideia é que as representações de baixa dimensionalidade y_i e y_j modelem corretamente as similaridades entre os vetores de alta dimensionalidade x_i e x_j. Logo, as probabilidades condicionais $p_{j|i}$ e $q_{j|i}$ devem ser iguais. O objetivo do SNE é encontrar uma representação de baixa dimensionalidade que minimize a distância entre essas duas probabilidades, tornando-as o mais próximo possível. Uma

medida estatística de quão próximas essas distribuições de probabilidade estão é a divergência de Kullback-Leibler, também conhecida como entropia relativa. O SNE minimiza uma soma de divergências de Kullback-Leibler em todos os pontos de dados usando um método de descida de gradiente, iniciando em uma semente qualquer. A função objetivo a ser minimizada é:

$$C = \sum_{i=1}^{n} KL(P_i \| Q_i) = \sum_{i=1}^{n} \sum_{j=1}^{n} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

onde P_i representa a distribuição de probabilidade condicional sobre todos os outros pontos de dados dado o ponto de dado x_i, e Q_i representa a distribuição de probabilidade condicional sobre todos os outros pontos no mapa dado o ponto y_i. Desta maneira, a função de custo do SNE foca em manter a localidade dos dados. Vale mencionar que o SNE utiliza a medida de perplexidade como parâmetro de entrada. Ela é utilizada para estimar a variância σ_i^2, ou seja, a densidade de vizinhos no espaço original.

3.2.6.1 O problema de *Crowding* e o t-SNE

Distâncias em espaços de alta dimensão tendem a se concentrar e tendem a ser quase idênticas, de modo que a preservação de distâncias nem sempre leva à preservação da estrutura de vizinhança. Em outras palavras, pontos que são vizinhos uns dos outros em altas dimensões não são sempre vizinhos em baixas dimensões. Para qualquer método de aprendizado de métricas e variedades, questiona-se a possibilidade de obter isometrias locais de $\mathbb{R}^m$ para a variedade $\mathcal{M}^d$. O *problema de crowding* consiste no seguinte fato: a área em $\mathbb{R}^d$ disponível para colocar os pontos distantes não será grande o suficiente em comparação com a área disponível para colocar pontos próximos. Em outras palavras, para modelar com precisão as pequenas distâncias, os pontos mais distantes de x_i terão que ser dispersos muito longe na variedade. Em termos de agrupamento de dados, este fenômeno pode gerar falsos grupos e relações nos dados. Portanto, os autores sugerem que uma maneira de aliviar esse problema é considerar, em baixa dimensionalidade, distribuições com caudas mais pesadas do que o modelo Gaussiano. É por isso que no t-SNE a distribuição t de Student é escolhida para substituir a distribuição Gaussiana tradicional usada no algoritmo SNE (MAATEN; HINTON, 2008).

A função de custo utilizada pelo t-SNE difere da usada pelo SNE de duas maneiras. A primeira, utiliza uma versão simetrizada da função de custo do SNE com uma simples computação da *descida do gradiente com momento*. Sendo assim, é minizada uma única divergência KL entre uma distribuição de probabilidade conjunta P no espaço de alta dimensionalidade e uma distribuição de probabilidade conjunta Q no espaço de baixa dimensionalidade. A segunda, utiliza uma distribuição t de Student, em vez de uma Gaussiana, para computar a similaridade entre dois pontos no espaço de baixa dimensionalidade. Ou seja, utiliza uma distribuição de cauda mais longa, aliviando o problema de crowding. No t-SNE, as similaridades no espaço de alta dimensionalidade são dadas por:

$$p_{ij} = \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma^2\right)}{\sum_{k \neq l} \exp\left(-\|x_k - x_l\|^2/2\sigma^2\right)}$$

onde a normalização na soma do denominador envolve todos os pares de pontos, não apenas os pontos vinculados a x_i, como no SNE original. Utilizando a distribuição t de Student com um grau de liberdade, as probabilidades conjuntas q_{ij} são definidas como:

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}}.$$

Note que ambos métodos podem sofrer de acordo com o ajuste do parâmetro de perplexidade que define o tamanho das vizinhanças. Mesmo assim, o sucesso do t-SNE inspirou variantes como o LargeVis (TANG et al., 2016) e o atual estado da arte, UMAP.

3.2.7 Uniform Manifold Approximation and Projection

O algoritmo *Uniform Manifold Approximation and Projection* (UMAP) está entre o estado-da-arte do aprendizado de métricas e variedades. É um método não supervisionado baseado em conceitos da geometria Riemanniana e topologia algébrica que assume que as amostras observadas são distribuídas uniformemente em uma variedade desconhecida $\mathcal{M}^d$. Sua fundamentação teórica é bastante complexa, sendo apenas resumida neste trabalho.

O UMAP (MCINNES et al., 2018) começa construindo um grafo com os k vizinhos mais próximos no espaço original dos dados $X \subset \mathbb{R}^m$. Representando o grafo

como um *conjunto fuzzy simplicial*, o UMAP busca encontrar uma representação de baixa dimensionalidade que aproxime o conjunto fuzzy simplicial de maneira a preservar as relações locais e globais presentes no espaço original. O processo de otimização envolve minimizar uma função de custo que penaliza discrepâncias entre as representações de alta e baixa dimensionalidade. Originalmente, a semente inicial utilizada em baixa dimensionalidade é dada via Laplacian Eigenmaps em X. Na etapa de otimização, o UMAP aplica o *Stochastic Gradient Descent* para otimizar a representação de baixa dimensionalidade, utilizando a entropia relativa entre as duas representações de X como função custo a ser minimizada. Nesta etapa, pontos similares no espaço original são atraídos uns aos outros no espaço de baixa dimensionalidade, enquanto pontos dissimilares são repelidos. Após o processo de otimização, obtem-se a representação final do conjunto de dados.

Embora seja efetivo ao preservar estruturas locais, o UMAP também sofre com a preservação da estrutura global da variedade. Seus parâmetros facilmente ajustáveis que controlam as forças de atração/repulsão, junto da taxa de aprendizado do gradiente, não são facilmente ajustados para capturar a estrutura global. Mesmo assim, podem ser utilizados para encontrar grupos naturais nos dados, ao mesmo tempo que suavida ruído e detecta outliers. Com complexidade $O(n^{1.14})$, torna-se um ótimo candidato para grandes conjuntos de dados. De fato, o método tem sido muito utilizado em visualização de dados e estruturas não lineares de alta dimensionalidade na biologia, química e neurociência (WANG et al., 2021).

3.2.8 Dimensionalidade Intrínseca via MLE

Um problema de interesse substancial na área de aprendizado de variedades é a determinação da dimensionalidade intrínseca dos dados. As abordagens existentes para estimativa deste valor podem ser divididas em dois grupos: métodos baseados em autovalores (ou projeções) e métodos geométricos. Os métodos de autovalores são baseados em PCA global ou local, com a dimensão intrínseca determinada pelo número de autovalores maiores que um determinado limiar. Estes podem ser uma boa ferramenta para análise exploratória de dados, onde se pode obter os autovalores e procurar por um limite bem definido, mas não para fornecer estimativas confiáveis da dimensão intrínseca. Os métodos geométricos exploram

a geometria intrínseca do conjunto de dados e são, muitas vezes, baseados em dimensões fractais, como a distância de correlação, ou distâncias entre vizinhos mais próximos, como .

Dos muitos métodos disponíveis para estimar a dimensão intrínseca, como a dimensão de correlação, foi o escolhido para compor este estudo a estimativa via MLE (LEVINA; BICKEL, 2004). De forma geral, este método analisa a distância dos vizinhos mais próximos como um processo homogêneo de Poisson que conta as observações de distância r de cada x. Para estimar a dimensão d a partir de observações, este método aplica uma estimativa via MLE sobre o processo observado. No que segue, é feita uma revisão dos principais conceitos do método.

Como de costume, as observações i.i.d. $X = \{x_1, \ldots, x_n\}$ em $\mathbb{R}^m$ representam uma imersão de uma amostra de dimensão inferior, isto é, $x_i = \varphi(y_i)$, onde y_i são amostras de uma densidade suave desconhecida em $\mathcal{M}^d$, com d desconhecido e $d \leq m$. A estimativa se baseia no seguinte fato. Considere a quantidade k de vizinhos mais próximos de um ponto fixado x, definindo $T_k(x)$ como a distância entre x e seu k-ésimo vizinho. Assumindo que $f(x) \approx \text{const}$ em uma pequena bola de raio $T_k(x)$ ao redor de x, tem-se o seguinte:

$$\frac{k}{n} \approx f(x) V_d(T_k(x))$$

onde $V_d(T_k(x)) = \frac{\pi^{d/2}}{\Gamma(1+\frac{d}{2})} \cdot [T_k(x)]^d$ é o volume da bola d-dimensional com raio $T_k(x)$ e $\Gamma(\cdot)$ é a função Gama. A fórmula mostra que, para amostras uniformes na variedade, uma bola de raio $T_k(x)$ centrada no ponto x conterá aproximadamente k pontos. Assim, o número de pontos de amostras em uma bola é aproximado pela densidade local $f(x)$ multiplicada pelo volume da bola.

Como d e $f(x)$ são desconhecidos na expressão acima, considera-se sobre as observações $x_i \in X$ um processo de Poisson em uma bola $B_x(r)$

$$N(t, x) = \sum_{i=1}^{N} \delta_{\{X_i \in B_x(t)\}}, \qquad 0 \leq t \leq r$$

que conta as observações de distância t de x. A partir de considerações sobre a função que governa o processo, a estimativa para $\hat{d}$ é obtida via MLE como

$$\hat{d}_r(x) = \left[\frac{1}{N(r,x)} \sum_{j=1}^{N(r,x)} \log \frac{r}{T_j(x)} \right]^{-1}.$$

Na prática, pode ser mais conveniente fixar o número de vizinhos k ao invés do raio da esfera r. Então, a estimativa da dimensionalidade em x torna-se

$$\hat{d}_k(x) = \left[\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{T_k(x)}{T_j(x)} \right]^{-1}.$$

Para obter a estimativa global, faz-se uma média entre as estimativas de cada x, retornando

$$\hat{d} = \frac{1}{n} \sum_{k=1}^{n} \hat{d}_k, \quad \text{onde} \quad \hat{d}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{d}_k(x_i).$$

Este método pode utilizar parâmetros ou pontos especiais, não necessariamente precisando do conjunto todo para análise. Em geral, para que esta estimativa funcione, é necessário que a esfera seja pequena e contenha um número suficiente de vizinhos, e por isso a escolha de k claramente afeta a estimativa. Pode ser o caso de que um conjunto de dados tenha diferentes dimensões intrínsecas em diferentes regiões. E com isso, a estimativa pode ser afetada por viés em diferentes componentes. Neste trabalho, assumimos que a a dimensão da variedade é igual em todo ponto. Além disso, o número de vizinhos utilizado na estimativa da dimensão intrínseca dos conjuntos de dados do experimento é $k = \sqrt{n}$.

3.3 Agrupamento de Dados

Em cenários onde não há um atributo objetivo, como em tarefas de classificação, o foco passa a ser deixar que os dados "falem por si". Modelos descritivos têm como objetivo identificar padrões subjacentes, como correlações, tendências, grupos, trajetórias e anomalias, que possam resumir relações dos dados. Neste contexto, as técnicas de agrupamento consistem em dividir objetos em grupos de acordo com similaridades presentes nos dados, sem a necessidade de rótulos pré-definidos. Essa metodologia é particularmente vantajosa em situações onde há pouco ou nenhum conhecimento prévio sobre a estrutura dos dados, permitindo a descoberta de padrões ou grupos naturais. Embora seu foco não seja de identificar classes e definir regras explícitas, os algoritmos de agrupamento são capazes de revelar estruturas intrínsecas entre os padrões no espaço de atributos.

De acordo com (NEGRI, 2021), os agrupamentos (ou grupos) são regiões do espaço de atributos que possuem alta densidade de padrões e que estão separadas

entre si por regiões de baixa densidade. Equivalentemente, é razoável admitir que padrões pertinentes a um determinado agrupamento demonstrem similaridade entre si e dissimilaridade em relação aos membros de outro agrupamento. Com isso, são elementos-chave, no processo de agrupamento, o uso de medidas que permitem a quantificação da dissimilaridade (ou similaridade) entre padrões. Neste sentido, a subjetividade é uma característica intrínseca à tarefa de agrupamento, uma vez que diferentes métodos ou critérios podem resultar em distintas formas de agrupar os mesmos dados. Essa variabilidade depende de como as similaridades entre os dados são definidas e das escolhas feitas durante o processo de análise, como o número de grupos ou a métrica de distância utilizada. Portanto, é essencial reconhecer que a interpretação dos resultados de um agrupamento pode variar conforme o algoritmo e os parâmetros adotados. Como ponto anterior à aplicação de técnicas de agrupamento, destaca-se a importância de se conhecer o problema em questão bem como as informações a serem extraídas. Uma vez que os diferentes métodos de agrupamento podem levar a diferentes interpretações, sua escolha deve ocorrer de forma consciente.

A literatura sobre agrupamento de dados é vasta e bem estabelecida, com inúmeras contribuições que abrangem desde algoritmos até métricas de similaridade. O algoritmo (HARTIGAN; WONG, 1979) foi o precursor da área, possuindo inúmeras generalizações e variações. Os modelos de mistura Gaussiana, inclusive, são uma destas. Há também uma diversidade de medidas de dissimilaridade bem fundamentadas, que são aplicáveis a diferentes tipos de dados, como dados numéricos, categóricos, nominais e binários (WANG; SUN, 2015).

Na visão matemática, os objetos podem ser representados como pontos em um espaço n-dimensional, onde cada dimensão (ou coordenada) corresponde a um atributo. O objetivo das técnicas de agrupamento é reduzir a dissimilaridade entre os objetos dentro de um grupo (intra-grupos) e aumentar a dissimilaridade entre os diferentes grupos (inter-grupos). Os métodos de agrupamento podem ser entendidos como métodos de classificação cujo paradigma de aprendizado é *não supervisionado*. Desse modo, considera-se que os métodos de agrupamento são funções $g : I \to G$, onde $I \subset \mathcal{X}$ é um conjunto de exemplos x observados no espaço de atributos $\mathcal{X}$, e $G = \{G_1, \ldots, G_c\}$ é uma partição de I em c subconjuntos. É de extrema importância destacar que não existe conhecimento algum sobre os rótulos

dos exemplos contidos em I. Decorrente do conceito de partição, é assegurado que $G_j \neq \emptyset$, para $j = 1, \ldots, c$, assim como $\bigcup_{j=1}^{c} G_j = I$. Esta definição estabelece claramente um processo de agrupamento *rígido*, para o qual os exemplos pertencem a um único agrupamento.

Outra abordagem de agrupamento, dita *nebulosa* (ou *fuzzy*), é estabelecida com base no conceito de conjuntos nebulosos (ZADEH, 1965). Nesta abordagem, os padrões desempenham diferentes graus de pertinência em relação a cada um dos agrupamentos. Formalmente, denota-se por $\lambda_{ij} \in [0, 1]$ como o grau de pertinência de um dado padrão x_i em relação ao agrupamento G_j, com $i = 1, \ldots, m$ e $j = 1, \ldots, k$. Além disso, deve estar assegurado que $\sum_{j=1}^{k} \lambda_{ij} = 1$ e $0 < \sum_{i=1}^{m} \lambda_{ij} < m$, para os possíveis i e j estabelecidos no problema. Enquanto a primeira restrição impõe um limite de pertinência com que cada padrão se associa aos diferentes grupos, a segunda restrição impede a existência de agrupamentos vazios. Segundo uma visão geral apresentada por (KOUTROUMBAS; THEODORIDIS, 2008), os métodos de agrupamento podem ser organizados em: hierárquicos; sequenciais, baseados em otimização de função custo; e outros. Na seção seguinte, será abordado o método de agrupamento via Mistura de Gaussianas que foi utilizado nos experimentos deste trabalho.

3.3.1 Gaussian Mixture Models

O Modelo de Mistura de Gaussianas (*Gaussian Mixture Model - GMM*) é um método probabilístico para modelagem de conjuntos de dados. O GMM assume que os padrões são gerados a partir de uma mistura de várias distribuições Gaussianas, cada uma representando um grupo nos dados (ZHAO; WEN; HAN, 2023). Os parâmetros do modelo incluem a média e a variância de cada componente Gaussiano, bem como os coeficientes de mistura que representam o peso relativo de cada componente na mistura. O algoritmo de Maximização da Expectativa (EM) é comumente empregado para estimar os parâmetros do GMM de forma iterativa. Por meio desse processo iterativo, o algoritmo maximiza a verossimilhança das observações, dados os parâmetros do GMM, identificando os grupos subjacentes nos dados.

Os GMMs são particularmente úteis para agrupamento quando os grupos pos-

suem formas complexas ou se sobrepõem, pois podem capturar a distribuição subjacente dos dados de maneira mais flexível em comparação com métodos como o agrupamento K-means (ZHANG et al., 2021). Além disso, os GMMs fornecem uma estrutura probabilística, possibilitando a quantificação da incerteza nas atribuições dos grupos, o que pode ser benéfico em várias aplicações, incluindo reconhecimento de padrões, segmentação de imagens e análise de dados (HU et al., 2024). Essa abordagem probabilística contrasta com métodos como o K-means, que dependem de atribuições rígidas de pontos a grupos únicos. Em vez disso, o GMMs estima a probabilidade de cada padrão pertencer a cada grupo, proporcionando insights valiosos sobre a possível sobreposição e incerteza (KASA; RAJAN, 2023). Essa técnica de aprendizado não supervisionado pode ser considerada uma generalização do algoritmo K-means, no qual, em vez de estimar apenas os centróides de cada grupo, também estimamos a forma (matriz de covariância) e a proporção de cada Gaussiana da mistura observada. Os GMMs atribuem probabilidades a cada ponto de dado pertencente a cada grupo com base nesses parâmetros, permitindo uma atribuição dos pontos de dados aos grupos,

Um *modelo de mistura* é um uma combinação linear de distribuibuições de probabilidade a fim de modelar um conjunto de dados $X \subset \mathbb{R}^m$. No GMM, assumimos que os padrões são variáveis i.i.d. de uma densidade $p(x)$ definida como uma mistura finita de c distribuições gaussianas multivariadadas

$$p(x) = \sum_{i=1}^{c} \pi_i N(x|\mu_i, \Sigma_i) \tag{3.41}$$

onde

$$N(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{m/2}|\Sigma_i|^{1/2}} \exp\left\{ -\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right\}, \tag{3.42}$$

μ_i é o vetor média e Σ_i é a matriz de covariância do i-ésimo componente. Os pesos $\pi_i \in [0,1]$ satisfazem $\sum_{i=1}^{c} \pi_i = 1$ e representam a probabilidade de que uma amostra aleatória x tenha sido gerada pelo i-ésimo componente. É importante notar que

$$\int_X p(x)dx = 1$$

dado que cada componente satisfaz $\int_X N(x|\mu_i, \Sigma_i)dx = 1$ e a ponderação feita por π_i não distorce o intervalo $[0,1]$. Neste sentido, busca-se encontrar os parâmetros $\vec{\theta} = (\pi_1, \ldots, \pi_c, \mu_1, \ldots, \mu_c, \Sigma_1, \ldots, \Sigma_c)$ que melhor representa X.

3.3.2 O algoritmo EM

O algoritmo EM é uma procedimento genérico para a modelagem probabilística de um conjunto de dados. Basicamente, o algoritmo otimiza os parâmetros de uma função de distribuição de probabilidades de forma a representar os dados da forma mais verossímil possível em casos onde as equações não podem ser resolvidas analiticamente (KURODA; GENG, 2023). Tipicamente, isso ocorre porque tais modelos envolvem variáveis latentes além dos parâmetros desconhecidos e dos dados observados. Em outras palavras, existem dados ausentes nas observações ou o modelo pode ser reformulado de maneira mais simples assumindo a existência de variáveis não observadas (latentes), que geralmente especificam o componente da mistura ao qual cada amostra pertence. Para encontrar uma solução de máxima verossimilhança, é necessário calcular as derivadas da função de log-verossimilhança em relação a todos os parâmetros desconhecidos e variáveis latentes e resolver simultaneamente todas as equações obtidas. Em modelos estatísticos com variáveis latentes, como o modelo GMM, isso é geralmente impossível (YOU; LI; DU, 2023).

Para compreender o tratamento sobre a distribuição $p(x)$, considere uma variável latente $Z = [z_1, \ldots, z_c]^T$ tal que apenas uma das componentes é igual a 1. Esta variável é considerada para identificar as desconhecidas componentes da mistura responsáveis por gerar x. Definindo $\pi_i = p(z_i)$ como a probabilidade a priori do padrão x ser gerado pela i-ésima Gaussiana, tem-se $\pi_i \in [0, 1]$ e $\sum_{i=1}^{c} \pi_i = 1$. Note que a i-ésima Gaussiana corresponde a distribuição condicional de x dado um valor particular, isto é,

$$p(x|z_i = 1) = N(x|\mu_i, \Sigma_i) \tag{3.43}$$

Das distribuições $p(x)$ e $p(x|y_j)$ tem-se a distribuição conjunta $p(x, z_i) = p(z_i)p(x|z_i)$. Com isso, a distribuição $p(x)$ é obtida por

$$p(x) = \sum_{i=1}^{c} p(x|z_i) = \sum_{i=1}^{c} \pi_i N(x|\mu_i, \Sigma_i).$$

Com estes ingredientes, calcula-se a probabilidade de pertinência de cada x_i a classe z_j como

$$p(z_j|x_i, \vec{\theta}) = \frac{\pi_j p(x_i|\mu_j, \Sigma_j)}{\sum_{k=1}^{c} \pi_k p(x_i|\mu_k, \Sigma_k)}.$$

Dentre diferentes técnicas existentes para obtenção dos parâmetros, a Estimação por Máxima Verossimilhança (*Maximum Likelihood Estimation - MLE*) é a

utilizada pelo GMM. Ela consiste em determinar o valores que maximizam a função de verossimilhança $L(\vec{\theta}; X, Z) = p(X, Z; \vec{\theta})$, que é a probabilidade conjunta em função do parâmetro $\vec{\theta}$. Nesta configuração, (FISHER, 1992) determinou que os parâmetros ótimos que modelam $p(X, \vec{\theta})$, em relação a um conjunto de observações X, correspondem aos parâmetros que maximizam a função de verossimilhança marginal

$$L(\vec{\theta}; X) = \int_Z p(X, Z, \vec{\theta}) dZ.$$

No entanto, a não linearidade dessa expressão torna a obtenção da solução praticamente intratável via $dL(\vec{\theta}) = 0$. Como alternativa, o algoritmo EM procura encontrar o MLE da log-verossimilhança marginal iterativamente aplicando os seguintes dois passos:

Expectativa: Calcule o valor esperado da log-verossimilhança para a distribuição condicional de Z dado X, usando a estimativa atual dos parâmetros, denotada por $\vec{\theta}^{(t)}$.

$$Q(\vec{\theta}|\vec{\theta}^{(t)}) = E_{Z|X,\vec{\theta}^{(t)}} \left[\log L(\vec{\theta}; X, Z) \right] \tag{3.44}$$

Maximização: Encontrar os parâmetros que maximizam a quantidade calculada no primeiro passo:

$$\vec{\theta}^{(t+1)} = \arg\max_{\vec{\theta}} Q(\vec{\theta}|\vec{\theta}^{t}) \tag{3.45}$$

Pode-se mostrar que, no caso dos Modelos de Mistura Gaussiana, após a inicialização dos parâmetros, os passos de expectativa e maximização são dados por:

Passo E: Calcule as probabilidades de pertencimento de classe $T_{j,i}^{(t)}$ por:

$$T_{j,i}^{(t)} = \frac{\pi_j p(x_i|\mu_j, \Sigma_j)}{\sum_{k=1}^{K} \pi_k p(x_i|\mu_k, \Sigma_k)} \tag{3.46}$$

Passo M: Atualize os parâmetros por:

$$n_j = \sum_{i=1}^{n} T_{j,i}^{(t)} \tag{3.47}$$

$$\mu_j^{(t+1)} = \frac{1}{n_j} \sum_{i=1}^{n} T_{j,i}^{(t)} x_i \tag{3.48}$$

$$\Sigma_j^{(t+1)} = \frac{1}{n_j} \sum_{i=1}^{n} T_{j,i}^{(t)} (x_i - \mu_j^{(t+1)})(x_i - \mu_j^{(t+1)})^T \tag{3.49}$$

$$\pi_j^{(t+1)} = \frac{n_j}{n} \tag{3.50}$$

A convergência do algoritmo EM depende essencialmente do comportamento da função de log-verossimilhança. Quando a log-verossimilhança se torna constante, o algoritmo converge para os estimadores de máxima verossimilhança do modelo GMM.

3.3.3 Medidas de Qualidade

Em analogia à avaliação de resultados de classificação, o particionamento efetuado por um método de agrupamento pode ser avaliado segundo sua capacidade de descrever a organização dos dados contemplados pelo problema em si (NEGRI, 2021). A avaliação do particionamento efetuado por um método de agrupamento pode ser conduzida por medidas baseadas em observações de referência ou de forma autônoma. Dentre diferentes propostas existentes na literatura, o Coeficiente Silhueta (*Silhouette Coefficient – SC*), o Índice de Calinski-Harabasz (*Calinski-Harabasz Index – CHI*), e o Índice de Davies-Bouldin são medidas que permitem avaliar resultados de agrupamentos com base em indicadores estatísticos. O *SC* mede o quão semelhante cada ponto é ao seu próprio grupo comparado com outros grupos, enquanto o *CHI* considera a razão entre a dispersão intra-grupo e inter-grupo para avaliar a qualidade do agrupamento. Já o *DBI* avalia a relação entre a coesão dentro dos grupos e a separação entre os grupos, sendo uma medida útil para determinar a compactação e a separação dos clusters.

Por outro lado, medidas que comparam o agrupamento gerado com uma referência conhecida também são amplamente usadas. A medida *V* contempla uma forma de avaliação baseada em exemplos de referência, medindo a precisão e a

completude do agrupamento em relação a rótulos conhecidos. O *Índice de Rand* (RI) é outra medida popular, que computa a similaridade entre dois agrupamentos considerando acertos aleatórios, sendo ajustada para reduzir o impacto do acaso. Da mesma forma, a medida *Fowlkes-Mallows* avalia a precisão e a completude dos clusters, focando na proporção de pares de pontos que são agrupados corretamente, tanto dentro do mesmo cluster quanto em clusters distintos.

Essas métricas proporcionam uma visão abrangente sobre a qualidade dos agrupamentos, permitindo tanto uma avaliação baseada em critérios internos, relacionados às características dos próprios dados, quanto em comparação com rótulos de referência conhecidos. No que segue, encontra-se um resumo das principais definições das métricas.

3.3.3.1 Rand Index

O Índice de Rand (*Rand Index* - RI) (RAND, 1971) é baseado na comparação de pares de elementos. Resumidamente, ele verifica se os pares são classificados de forma consistente, ou seja, se estão no mesmo agrupamento, em ambos os agrupamentos ou se estão em agrupamentos diferentes. A medida é calculada com base na contagem de pares:

a : de mesma classe e mesmo grupo

b : de mesma classe e grupos distintos

c : de classes distintas e mesmo grupo

d : de classes distintas e grupos distintos.

Sua equação é definida por

$$RI = \frac{a + d}{a + c + b + d}.$$

Quando os pares de agrupamentos não têm similaridades, ou seja, a e b próximos de zero, o RI se aproxima de zero. Quando os agrupamentos são idênticos, ou seja, c e b próximos de zero, o RI se aproxima de um.

3.3.3.2 Calisnki-Harabasz

Baseando-se nas variabilidades externas e internas dos grupos resultantes, o Índice de Calinski-Harabasz (CH) (CALINSKI; HARABASZ, 1974) é construído

a partir da "soma dos quadrados".

$$CHI = \frac{\mathrm{Tr}(V_E)\#I - k}{\mathrm{Tr}(V_I)\frac{k-1}{k}}$$

sendo que V_I e V_E, conforme já definidos pelas equações acima, quantificam a variabilidade interna dos agrupamentos e a separação entre agrupamentos.

A definição matemática do índice CH, para um conjunto de dados $X = \{x_1, x_2, ..., x_N\}$ é dada por:

$$CH = \left[\frac{\sum_{k=1}^{K} n_k \|c_k - c\|^2}{K-1}\right]\left[\frac{N-K}{\sum_{k=1}^{K}\sum_{i=1}^{n_k}\|x_i - c_k\|^2}\right]$$

onde n_k e c_k denotam o número de amostras e o centróide do k-ésimo agrupamento, c é o centróide global (média dos centróides) e N é o número total de amostras. Os valores expressos por esta medida tendem a aumentar ao passo que os agrupamentos se tornam densos e com maior separabilidade entre si.

Os valores expressos por esta medida tendem a aumentar ao passo que os agrupamentos se tornam densos e com maior separabilidade entre si. No entanto, vale destacar que, devido à presença de V_I em sua composição, a medida CH tende a diminuir quando os agrupamentos identificados não são compactos.

3.3.3.3 Davies-Bouldin Index

O Índice Davies–Bouldin (DAVIES; BOULDIN, 1979) é uma métrica para avaliar algoritmos de agrupamento. Esta é uma técnica de avaliação interna, onde a validação de quão bem o agrupamento foi feito é realizada usando quantidades e características inerentes ao conjunto de dados. Isso tem a desvantagem de que um bom valor reportado por este método não implica a melhor recuperação de informação.

Dado n pontos dimensionais, seja C_i um cluster de pontos de dados. Seja X_j uma amostra de dimensão m atribuído ao cluster C_i:

$$S_i = \left(\frac{1}{T_i}\sum_{j=1}^{T_i}\|X_j - A_i\|_p^q\right)^{1/q}$$

Aqui, A_i é o centróide de C_i e T_i é o tamanho do cluster i. S_i é a raiz q-ésima do momento q-ésimo dos pontos no cluster i em relação à média. Se $q = 1$, então S_i

é a distância média entre os vetores de características no cluster i e o centróide do cluster. Normalmente, o valor de p é 2, o que torna a distância uma *função de distância Euclidiana*. Muitas outras métricas de distância podem ser usadas, no caso de *variedades* e dados de alta dimensionalidade, onde a distância euclidiana pode não ser a melhor medida para determinar os clusters. É importante notar que esta métrica de distância precisa estar alinhada com a métrica usada no próprio esquema de clustering para resultados significativos.

$$M_{i,j} = \|A_i - A_j\|_p = \left(\sum_{k=1}^{m} |a_{k,i} - a_{k,j}|^p \right)^{1/p}$$

$M_{i,j}$ é uma medida de separação entre o cluster C_i e o cluster C_j. $a_{k,i}$ é o k-ésimo elemento de A_i, e há m tais elementos em A, pois é um centróide de dimensão m. Aqui, k indexa as características dos dados, e isso é essencialmente a *distância Euclidiana* entre os centros dos clusters i e j quando $p = 2$.

3.3.3.4 Fowlkes–Mallows

O Índice de Fowlkes–Mallows (FOWLKES; MALLOWS, 1983) é um método de avaliação externa que é usado para determinar a similaridade entre dois agrupamentos, e também uma métrica para medir *matrizes de confusão*. Esta medida de similaridade pode ser entre dois agrupamentos hierárquicos ou entre um agrupamento e uma classificação de referência. Um valor mais alto para FM indica uma maior similaridade entre os agrupamentos e as classificações de referência.

Quando os resultados de dois algoritmos de agrupamento são usados para avaliar os resultados, a medida FM é definido como:

$$FM = \sqrt{PPV \cdot TPR} = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}}$$

onde TP é o número de *verdadeiros positivos*, FP é o número de *falsos positivos*, e FN é o número de *falsos negativos*. TPR é a *taxa de verdadeiros positivos*, também chamada de *sensibilidade* ou *recall*, e PPV é a *taxa de predição positiva*, também conhecida como *precisão*.

O valor mínimo possível do *Índice de Fowlkes–Mallows* é 0, o que corresponde à pior classificação binária possível, onde todos os elementos foram classificados

incorretamente. E o valor máximo possível do *Índice de Fowlkes–Mallows* é 1, o que corresponde à melhor classificação binária possível, onde todos os elementos foram classificados perfeitamente.

3.3.3.5 Silhuette-Score

O método SC (ROUSSEEUW, 1987) é capaz de encontrar partições dos dados de modo que a variabilidade interna dos agrupamentos seja minimizada enquanto a separação entre os agrupamentos seja maximizada. Nestes casos, a quantificação das variabilidades interna e externa dos agrupamentos é tomada como referencial. Considere o conjunto de padrões X é particionado entre os agrupamentos $G_1, \ldots, G_k$. A medida SC é expressa por:

$$SC = \frac{S_E - S_I}{\max\{S_I, S_E\}}$$

onde

$$S_I = \frac{2}{k} \sum_{j=1}^{k} \frac{k m_j (m_j - 1)}{1} \sum_{x_i, x_\ell \in G_j} \|x_i - x_\ell\|$$

e

$$S_E = \frac{1}{k} \sum_{j=1}^{k} \frac{1}{m_j m_{p_j}} \sum_{x_i \in G_j} \sum_{x_\ell \in G_{p_j}} \|x_i - x_\ell\|$$

e $p_j = \arg\max_{p_j=1,\ldots,k, p_j \neq j} \|\mu_j - \mu_{p_j}\|$, sendo μ_j e μ_{p_j} os vetores médios (i.e., os centróides) que representam, respectivamente, os agrupamentos G_j e G_{p_j}, para um $j \neq p_j \in \{1, \ldots, k\}$, de cardinalidades m_j e m_{p_j}, respectivamente. Os valores de SC estão limitados ao intervalo $[-1, 1]$, cujos limites inferior e superior indicam pior e melhor avaliação possível. A quantidade S_I, conforme representada, fornece a distância euclidiana média entre cada par de observação pertencente a um mesmo agrupamento; já S_E expressa a distância euclidiana média entre cada padrão e todos os padrões que ocupam o agrupamento mais próximo.

3.3.3.6 V-Score

A medida V (ROSENBERG; HIRSCHBERG, 2007) realiza uma avaliação do particionamento feito sobre $\mathcal{I}$ com base em duas características: a *homogeneidade* (H_ω) e a *completude* (C_ω). A primeira propriedade, de homogeneidade, vislumbra

que os agrupamentos identificados sejam compostos por exemplos de uma única classe. Por outro lado, a completude estabelece que exemplos de uma mesma classe devam estar associados a um mesmo agrupamento.

A verificação dessas propriedades é baseada em medidas de entropia. Para um particionamento composto por k agrupamentos, a entropia da classe ω_j condicionada à existência de k classes, são calculadas:

$$H(\Omega) = -\sum_{j=1}^{\ell} \frac{m_j}{m} \log\left(\frac{m_j}{m}\right)$$

$$H(\Omega|G) = -\sum_{j=1}^{\ell}\sum_{\ell=1}^{k} \frac{m_{j\ell}}{m_j} \log\left(\frac{m_{j\ell}}{m_j}\right)$$

em que $H(\Omega)$ se refere à entropia das classes e $H(\Omega|G)$ é a entropia das classes condicionada aos agrupamentos. Posteriormente, são determinadas expressões para a homogeneidade, a completude e, em último, para a medida V:

$$H_\omega = 1 - \frac{H(\Omega|G)}{H(\Omega)}$$

$$C_\omega = 1 - \frac{H(G|\Omega)}{H(G)}$$

$$V = \frac{2H_\omega C_\omega}{H_\omega + C_\omega}.$$

Os valores retornados pela medida V estão contidos no intervalo $[0, 1]$, no qual 1 aponta para um ajuste perfeito entre os agrupamentos e as classes expressas definidas através das informações de referência. Por outro lado, diante de agrupamentos aleatórios, essas medidas podem levar a valores distintos entre si, ambos diferentes de zero. Além da avaliação baseada em amostras de referência, a medida V pode ser empregada na comparação de aderência entre particionamentos gerados por dois métodos distintos ou em diferentes arranjos de parâmetros desses intervenientes. Neste caso, basta assumir Ω como um conjunto de referência cujos indicadores de classe associados aos padrões foram gerados de maneiras distintas ou diferentes.

3.4 Teste de Hipóteses

Os testes de hipóteses são uma ferramenta fundamental na estatística inferencial, amplamente utilizados em diversas áreas de pesquisa para tomada de decisão com base em dados amostrais. O processo de teste de hipóteses permite avaliar a veracidade de uma afirmação ou hipótese sobre uma população, utilizando-se de evidências extraídas de amostras. Neste contexto, os elementos centrais de um teste de hipóteses incluem a *hipótese nula* (H_0), a *hipótese alternativa* (H_1), o *p-valor* e os níveis de significância.

De acordo com (HOLLANDER; WOLFE; CHICKEN, 2015), o primeiro passo no desenvolvimento de um teste de hipóteses é a formulação de duas hipóteses mutuamente exclusivas: a *hipótese nula*, H_0, que representa o estado inicial, e a *hipótese alternativa*, H_1, que é a proposição que se deseja testar. Em termos formais, a hipótese nula geralmente sugere que não há efeito ou diferença significativa, enquanto a hipótese alternativa sugere o oposto, ou seja, a existência de um efeito, diferença ou relação estatisticamente significante:

H_0: Não há diferença significativa ou efeito na população.

H_1: Há uma diferença significativa ou efeito na população.

A decisão a respeito de qual das hipóteses aceitar é tomada com base nas evidências fornecidas pelos dados da amostra.

O *nível de significância* α é um valor previamente definido no teste, geralmente fixado em 5% ($\alpha = 0.05$). Este valor representa a probabilidade máxima de se cometer um *erro tipo I*, ou seja, rejeitar a hipótese nula quando esta é verdadeira. Em termos práticos, ao definir α, o teste aceita uma chance de α de rejeitar H_0 incorretamente.

Além disso, outro erro possível em testes de hipóteses é o *erro tipo II*, que ocorre quando se falha em rejeitar H_0 quando H_1 é verdadeira. A probabilidade de cometer um erro tipo II é geralmente representada por β, e o complemento $(1 - \beta)$ é denominado *poder do teste*, que reflete a capacidade do teste detectar uma diferença quando ela realmente existe.

O *p-valor* é a medida probabilística que indica a força da evidência contra a hipótese nula. Mais precisamente, ele é a probabilidade de obtermos um resultado

tão extremo quanto, ou mais extremo do que, o observado nos dados, assumindo que a hipótese nula seja verdadeira. Em um teste de hipóteses, compar-se o p-valor com o nível de significância α:

(i) Se o *p-valor* $\leq \alpha$, rejeita-se H_0 em favor de H_1.

(ii) Se o *p-valor* $> \alpha$, não há evidências suficientes para rejeitar H_0.

É importante notar que um p-valor pequeno não implica, necessariamente, que a hipótese alternativa é verdadeira, mas sim que os dados observados são inconsistentes com a hipótese nula. No entanto, o p-valor sozinho não mede o tamanho do efeito ou a importância prática dos resultados. Dependendo da natureza da hipótese e do tipo de dados envolvido, diferentes tipos de testes de hipóteses podem ser utilizados.

No contexto deste trabalho, onde se busca comparar as medidas de qualidade de agrupamentos obtidas por diferentes métodos de aprendizado de variedades utilizando o modelo de Mistura de Modelos Gaussianos (GMM), é necessário empregar um método estatístico que permita a comparação de múltiplos métodos de forma robusta. Como os dados a serem comparados envolvem medidas repetidas (ou amostras dependentes), a análise estatística deve levar em consideração a dependência existente entre as amostras.

O *teste de Friedman* é uma alternativa não-paramétrica ao teste ANOVA de medidas repetidas, adequado para situações onde não se pode assumir que os dados seguem uma distribuição normal. Este teste avalia se há diferenças estatisticamente significativas entre três ou mais grupos relacionados (neste caso, os diferentes métodos de aprendizado de variedades), comparando as medianas das distribuições. Como estamos lidando com algoritmos de agrupamento aplicados a diferentes métodos de aprendizado de variedades, o teste de Friedman é apropriado, pois ele leva em consideração a dependência entre as execuções dos experimentos em cada método.

Após a detecção de diferenças significativas pelo teste de Friedman, surge a necessidade de identificar quais métodos de aprendizado de variedades apresentam diferenças entre si. Para isso, utiliza-se um teste pós-hoc, que permite a comparação par a par entre os métodos. O *teste post-hoc de Nemenyi* é uma escolha

comum neste contexto, pois é indicado para comparar todos os pares de grupos após um teste de Friedman. Ele verifica quais métodos diferem significativamente uns dos outros, ajustando para o fato de que múltiplas comparações estão sendo realizadas, e sem a necessidade de assumir a normalidade dos dados.

Assim, a combinação do teste de Friedman e do teste post-hoc de Nemenyi proporciona uma abordagem estatística robusta para avaliar se há diferenças significativas entre as medidas de qualidade de agrupamento dos diferentes métodos de aprendizado de variedades, e, em caso afirmativo, identifica especificamente quais métodos se diferenciam.

Capítulo 4

O algoritmo K-ISOMAP

Com a noção de preservação de distâncias em espaços de alta dimensionalidade, o método ISOMAP busca encontrar uma imersão isométrica de um conjunto de dados para um novo espaço euclidiano $\mathbb{R}^d$ tal que $d < m$. No início deste processo, é extraído um grafo de vizinhanças de acordo com a regra ε-*ball* ou k-NN. Esta estrutura permite a interpretação de uma topologia da variedade de dados, onde o ISOMAP busca pelas distâncias mais próximas. Mudanças nos parâmetros ε ou k afetam diretamente a noção de distância percebida pelo grafo. Além disso, a presença de ruído e/ou outliers pode degradar ainda mais a estrutura subjacente.

Na tentativa de aliviar a sensibilidade a ruído e outliers que o ISOMAP possui, o método K-ISOMAP foi proposto. Ele utiliza o conceito de curvatura para criar uma nova função de custo para as arestas, explorando a hipótese de que arestas com alta curvatura indicam transição entre grupos/classes. Além disso, o K-ISOMAP abre espaço para um maior aprofundamento sobre a relação entre algoritmos de agrupamento e aprendizado de métricas/variedades. Este capítulo busca explicar a fundamentação teórica do K-ISOMAP e exemplificar seu funcionamento.

4.1 O K-grafo

O método K-ISOMAP (LEVADA, 2022) utiliza a base ortonormal que define o espaço tangente em um ponto da variedade como fonte de informação geométrica. Como a variação do campo tangente é expressa apenas em termos da curvatura, a proposta consiste em quantificar como o espaço tangente varia ao redor da vizinhança de cada ponto, para construir uma função de distância intrínseca a ser utilizada na ponderação das arestas do grafo KNN (LEVADA, 2022). Este grafo k-NN definido em termos da curvatura é chamado de *K-grafo*.

Considere $X = \{x_1, x_2, \ldots, x_n\}$, com $x_i \in \mathbb{R}^m$, a matriz de dados. O primeiro passo no algoritmo K-ISOMAP consiste na obtenção de um grafo k-NN a partir dos dados. Neste estágio inicial, adota-se a distância euclidiana para computar os vizinhos mais próximos de cada amostra x_i. A partir disso, o K-ISOMAP propõe a atualização do grafo k-NN para o K-grafo antes de calcular as distâncias geodésicas. Chamando de η_i o sistema de vizinhança de x_i, um patch P_i é definido como o conjunto $\{x_i \cup \eta_i\}$. Note que o número de elementos de P_i é $k+1$, para $i = 1, 2, ..., n$. Em notação matricial, um patch P_i é dado por:

$$P_i = [x_i, x_{i1}, x_{i2}, ..., x_{ik}] = \begin{bmatrix} x_i(1) & x_{i1}(1) & \ldots & x_{ik}(1) \\ x_i(2) & x_{i1}(2) & \ldots & x_{ik}(2) \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ldots & \vdots \\ x_i(m) & x_{i1}(m) & \ldots & x_{ik}(m) \end{bmatrix}_{m \times (k+1)} \tag{4.1}$$

Para cada patch P_i, calcula-se o subespaço PCA associado que representa uma aproximação do plano tangente à variedade de dados em x_i. Após calcular a matriz de covariâncias Σ_i, decompondo-a em seus autovetores e autovalores, obtém-se

$$\Sigma_i = U_i Q_i U_i^T$$

onde $U_i = [u_{i1}, u_{i2}, \ldots, u_{im}]$ é uma matriz $m \times m$ cujas colunas representam os autovetores de Σ_i e Q_i é uma matriz diagonal com os autovalores de Σ_i.

Utilizando a base ortonormal do espaço PCA, que é constituída pelos autovetores do patch P_i, o K-ISOMAP atualiza os pesos das arestas do grafo k-NN pela variação local dos espaços tangentes em cada aresta. Considere $U_i = [u_{i1}, \ldots, u_{im}]$

e $U_j = [u_{j1}, \ldots, u_{jm}]$ os espaços PCA que aproximam os espaços tangentes em cada vértice da aresta (x_i, x_j). Uma maneira de calcular a variação entre os espaços tangentes é utilizando a norma da diferença entre os respectivos vetores das bases. Desta maneira, fica definido um vetor $\vec{K}_{ij}$, chamado de vetor de curvaturas principais, cujas componentes são

$$\vec{K}_{ij}(l) = \|u_{il} - u_{jl}\| \quad \text{para } l = 1, 2, \ldots, m. \tag{4.2}$$

A norma $\|\vec{K}_{ij}\|$ é considerada como o novo peso da aresta (x_i, x_j). O método K-ISOMAP finaliza encontrando os caminhos mínimos do K-grafo e as coordenadas intrínsecas na dimensão desejada.

Originalmente, o método utiliza todas as m componentes principais ordenadas de forma decrescente no cálculo do vetor de curvaturas de uma aresta. Entretanto, neste trabalho, utiliza-se apenas as d primeiras componentes de $\vec{K}_{ij}$ para ponderar o K-grafo, sendo d a dimensão intrínseca da variedade de dados. Esta consideração é motivada pelo simples fato de que o espaço tangente tem a mesma dimensão que a variedade. Portanto, ao considerar a taxa de variação do espaço tangente, ou seja, a curvatura, basta comparar as d componentes principais. A Figura 4 ilustra os espaços tangentes em uma aresta (x_i, x_j).

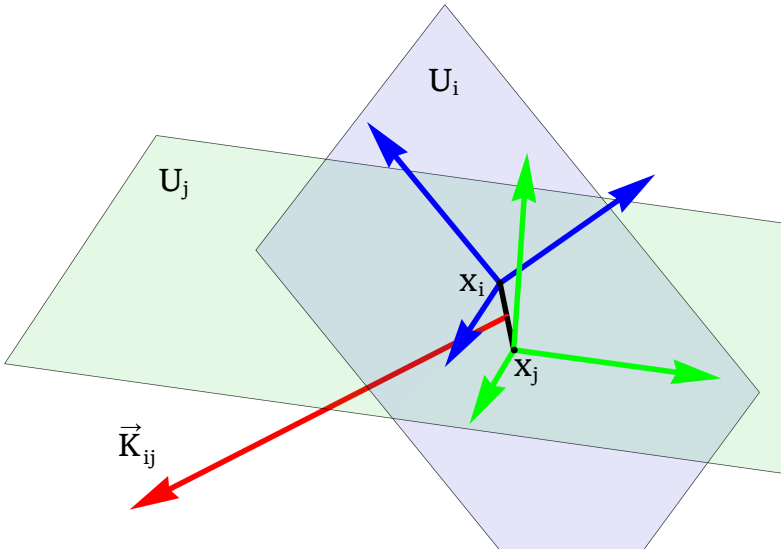


Figura 4 – Vetor de curvaturas principais da aresta (x_i, x_j).

Além da norma euclidiana, é possível utilizar um parâmetro f que determina a medida de similaridade a ser extraída de $\vec{K}_{ij}$. Por padrão, a métrica f_0 é definida como a norma do vetor de curvaturas $\|\vec{K}_{ij}\|$. A métrica f_1 fornece a curvatura associada à primeira componente de $\vec{K}_{ij}$, enquanto f_2 fornece a curvatura relacionada à última componente. A métrica f_3 calcula a média entre as curvaturas do primeiro e do último componentes de $\vec{K}_{ij}$. A métrica f_4 representa a curvatura média de cada componente. A métrica f_5 fornece o valor máximo, enquanto a métrica f_6 fornece o valor mínimo de $\vec{K}_{ij}$. A métrica f_7 calcula o produto entre os valores mínimo e máximo de $\vec{K}_{ij}$, enquanto a métrica f_8 calcula a diferença entre eles. A métrica f_9 aplica um kernel exponencial negativo na curvatura média. A métrica f_{10} combina o ISOMAP e K-ISOMAP na fórmula

$$\hat{A}[i,j] = (1-\alpha)\frac{A[i,j]}{\sum_{l=1}^{m} A[i,l]} + \alpha\|\vec{K}_{ij}\| \tag{4.3}$$

onde $A[i,j]$ é o peso da aresta (i,j) no gráfico k-NN e $\hat{A}[i,j]$ é o peso atualizado. Quando α é igual a 0, o algoritmo produz o ISOMAP padrão até um fator de escala local, enquanto definir α como 1 resulta na métrica f_0 sendo aplicada. O algoritmo a seguir descreve os passos do K-ISOMAP.

Algoritmo 6 Curvature Based Isometric Feature Mapping

function K-ISOMAP(X,d,k,f,α)
    A partir da matriz de dados X construa um grafo k-NN.
    Compute as matrizes de covariância Σ_i para cada patch P_i.
    Compute o K-grafo A ponderado por f e $\vec{K}_{ij}$.
    Compute a matriz de distâncias geodésicas do K-grafo $D_{n\times n}$.
    Compute $H = I - \frac{1}{n}U$, onde U é uma matriz $n \times n$ de 1's.
    Compute $B = -\frac{1}{2}HDH$.
    Encontre os autovalores e autovetores da matriz B.
    Selecione os $d < m$ maiores autovalores e seus autovetores.
    Defina as matrizes Λ e V, conforme definidas no ISOMAP.
    return $\tilde{X} = \Lambda^{1/2}V^T$

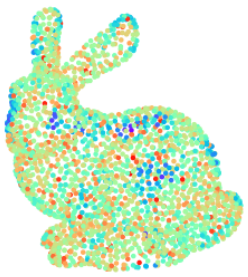


Figura 5 – Curvatura média ao redor de cada vértice do K-grafo no Stanford Bunny (k=7).

4.2 Propriedades

A principal limitação do K-ISOMAP diz respeito ao seu alto custo computacional em conjuntos de dados com grandes amostras e/ou características. O principal problema é que, para cada um dos n patches do gráfo k-NN, tem-se que calcular o PCA para obter a aproximação do espaço tangente naquele ponto. Como a complexidade computacional do PCA é $O(nm^2 + m^3)$, mesmo em casos em que o número de amostras nos patches é pequeno, o gargalo é no número de características. O custo total do PCA no K-ISOMAP se torna $O(n^2m^2 + nm^3)$, o que é inviável para grandes conjuntos de dados.

Por ser um algoritmo baseado em grafo de vizinhanças, o K-ISOMAP também sofre com a definição da quantidade de vizinhos e a presença de ruído nos dados. Na Figura 6, por exemplo, é observado o efeito do ruído e aumento do número de vizinhos no mergulho pelo K-ISOMAP. Além disso, para aproximar os espaços tangentes da variedade é necessária uma quantidade considerável de vizinhos, o que aumenta significativamente a complexidade do algoritmo.

Ainda, ao considerar a curvatura como medida empírica de similaridade no grafo, não é garantida a convergência teórica para a distância geodésica da variedade de dados no K-ISOMAP. Por este motivo, o K-ISOMAP pode ser considerado um método focado no aprendizado de métricas, embora utilize o aprendizado de variedades na consideração do K-grafo. Mesmo assim, há muito espaço para estudo do comportamento do espaço tangente e a curvatura no método, tanto computacionalmente quanto teoricamente.

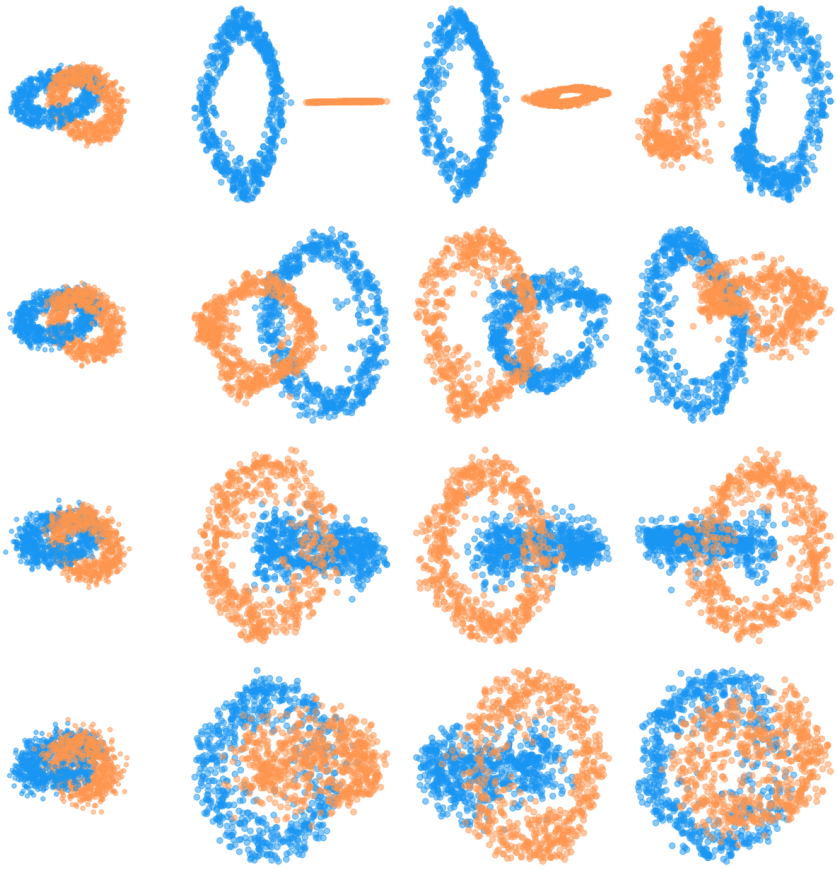


Figura 6 – Double Torus e sua redução por K-ISOMAP de acordo com uma quantidade de ruído (eixo vertical) e quantidade de vizinhos (eixo horizontal).

As limitações do K-ISOMAP confrontam com uma excelente performance em tarefas de classificação (LEVADA, 2022). O cálculo do espaço tangente em cada ponto age como um filtro de ruído das diferentes distribuições em cada região da variedade de dados. Mesmo assim, a presença de ruído pode aumentar a diferença $\vec{K}_{ij} = \|u_{il} - u_{jl}\|$ entre as bases dos espaços tangentes em x_i e x_j. Por outro lado, o K-ISOMAP lida com outliers melhor do que o ISOMAP. De fato, no K-ISOMAP a distância só é relevante na construção do grafo k-NN. Os menores caminhos no K-grafo são, na verdade, caminhos de mínima curvatura.

Assim, a suposição de que a curvatura afeta o desempenho dos algoritmos de classificação e agrupamento é baseada na observação de que pontos de alta curvatura tendem a se concentrar em bordas, que são os principais componentes

na definição de classes/grupos. A curvatura no K-grafo também indica uma medida de quanto a aresta se desvia de ser plana. Regiões de alta curvatura se associam a limites entre grupos, e quando o peso da aresta (borda) é alto, ele sinaliza um limite entre grupos.

No que segue, a metodologia e os resultados do experimento realizado com o K-ISOMAP e os demais métodos de aprendizado de métricas/variedades discutidos é apresentada. Em resumo, deseja-se testar a qualidade de agrupamentos obtidos com o GMM após aplicação do ISOMAP, K-ISOMAP, Kernel PCA, LLE, Laplacian Eigenmaps, t-SNE e UMAP.

Capítulo 5

Experimentos e Resultados

Tendo como objetivo avaliar a performance do K-ISOMAP em problemas de agrupamento, conjuntos de dados reais e sintéticos foram selecionados para compor as duas baterias do experimento realizado. São utilizadas medidas de qualidade intrínsecas e extrínsecas para capturar quantitativamente a estrutura de agrupamentos obtidos após redução de dimensionalidade. O experimento consiste na comparação das métricas de qualidade de cada agrupamento obtido com o GMM segundo o K-ISOMAP, métodos clássicos e o estado-da-arte do aprendizado de métricas e redução de dimensionalidade. Por fim, são aplicados testes de hipóteses para comparar os resultados RAW e K-ISOMAP com os demais métodos de RD. Uma análise gráfica e descritiva é conduzida para evidenciar as características dos resultados obtidos.

5.1 Metodologia e procedimento experimental

Deseja-se testar se é possível melhorar a qualidade de agrupamentos via GMM a partir do K-ISOMAP. Para fazer isso, utiliza-se outros métodos de redução de dimensionalidade como referência: ISOMAP, KPCA, LLE, Laplacian Eigenmaps,

t-SNE e UMAP. Em (LEVADA, 2022), os testes realizados supõem a dimensionalidade intrínseca como sendo $d = 2$. Como a maioria dos conjuntos de dados possui uma estrutura de variedade de dimensão alta, neste trabalho, incorpora-se no experimento uma estimativa da dimensionalidade intrínseca d^* de cada conjunto de dados, conforme a seção **??**. Este parâmetro será utilizado por cada método de RD na etapa de embedding, e também na seleção das d^* curvaturas principais do K-grafo. No K-ISOMAP, a alta dimensionalidade do espaço original pode gerar ruídos no vetor de curvaturas principais. Por isso, é feita a ponderação no K-grafo de acordo com as d^* direções de maior curvatura.

De modo a testar os métodos de RD sob uma mesma estrutura, define-se o número de vizinhos[1] do grafo k-NN como $k = \lfloor \sqrt{n} \rfloor$. Na teoria dos grafos, é conhecido que a performance teórica de um classificador k-NN é maximizada quando, sob condições especiais (CHAUDHURI; DASGUPTA, 2014),

$$k \to \infty \quad \text{e} \quad k/n \to 0 \quad \text{para} \quad n \to \infty. \tag{5.1}$$

Mesmo se k for uma variável independente de n, mas que satisfaz (5.1), é possível maximizar a performance do classificador. Nos métodos de RD, busca-se preservar estruturas locais e/ou globais do dados a partir de um grafo. Logo, $k = \lfloor \sqrt{n} \rfloor$ é considerada uma heurística na inicialização dos métodos, pois garante uma quantidade de vizinhos adequada para computar similaridades.

Após calcular a qualidade de agrupamentos, para cada bateria do experimento, aplica-se o teste de hipóteses não-paramétrico de Friedman, seguido do teste *post-hoc* de Nemenyi para validar se os resultados obtidos pelo K-ISOMAP são significativos em relação aos resultados dos demais métodos. Quando o teste de Friedman indica um resultado significativo, o teste de Nemenyi post-hoc identifica as diferenças específicas entre os pares de grupos. Neste experimento, comparam-se os resultados do K-ISOMAP com os resultados RAW e os demais métodos, formando seis testes para cada métrica. Foram testados 40 conjuntos de dados disponíveis publicamente no portal <www.openML.org>, separados em duas baterias que refletem características semelhantes dos dados. Ademais, foram adicionados dois conjuntos de dados sintéticos para compor a primeira bateria do experimento. A seguir, descreve-se o processo experimental a ser realizado:

[1] Não se aplica ao Kernel PCA e ao t-SNE.

Algoritmo 7 Procedimento Experimental

Input: Datasets $\mathcal{X} = \{X_1, X_2, ..., X_{42}\}$
Métodos $\mathcal{M} = \{\text{RAW}, \text{KISOMAP}, \text{ISOMAP}, \text{UMAP}, \text{LLE}, \text{LE}, \text{t-SNE}, \text{KPCA}\}$
Métricas $\mathcal{Q} = \{\text{RI,FM,VS,CH,SS,DB}\}$

procedure EXPERIMENTO($\mathcal{X}, \mathcal{M}, \mathcal{Q}$)
  for $X_i \in \mathcal{X}$ **do**
    $n_i \leftarrow \text{NSamples}(X_i)$ ▷ Número de amostras
    $k_i \leftarrow \lfloor \sqrt{n_i} \rfloor$ ▷ Número de vizinhos
    $c_i \leftarrow \text{NLabels}(X_i)$ ▷ Número de classes de X_i
    $Y_i \leftarrow \text{Labels}(X_i)$ ▷ Classes de X_i
    $X_i \leftarrow \text{Normalize}(X_i)$ ▷ Normalização de X_i
    $d_i \leftarrow \text{IntrinsicDimension}(X_i)$ ▷ Estimativa via MLE (3.2.8)
    for $M \in \mathcal{M}$ **do**
      if $M = \text{KISOMAP}$ **then** ▷ Calcula o máximo dentre as opções f
        for $j \leftarrow 0$ **to** 10 **do**
          $\tilde{X}_{M,i,j} \leftarrow M(X_i, k_i, d_i, f_j)$ ▷ Execução do K-ISOMAP
          $\tilde{Y}_{M,i,j} \leftarrow \text{GMM}(\tilde{X}_{M,i}, c_i)$ ▷ Classes do GMM
        $\text{RI}_{M,i} \leftarrow \max\{\text{RandIndex}(Y_i, \tilde{Y}_{M,i,j})\}_j$
        $\text{FM}_{M,i} \leftarrow \max\{\text{FowlkesMallows}(Y_i, \tilde{Y}_{M,i,j})\}_j$
        $\text{VS}_{M,i} \leftarrow \max\{\text{VScore}(Y_i, \tilde{Y}_{M,i,j})\}_j$
        $\text{CH}_{M,i} \leftarrow \max\{\text{CalinskiHarabasz}(\tilde{X}_{M,i,j}, \tilde{Y}_{M,i,j})\}_j$
        $\text{SS}_{M,i} \leftarrow \max\{\text{Silhouette}(\tilde{X}_{M,i,j}, \tilde{Y}_{M,i,j})\}_j$
        $\text{DB}_{M,i} \leftarrow \max\{\text{DaviesBouldin}(\tilde{X}_{M,i,j}, \tilde{Y}_{M,i,j})\}_j$
      else
        if $M = \text{RAW}$ **then**
          $\tilde{X}_{M,i} \leftarrow X_i$ ▷ Não há redução de dimensionalidade
        else
          $\tilde{X}_{M,i} \leftarrow M(X_i, k_i, d_i)$ ▷ t-SNE e KPCA não utilizam k_i
        $\tilde{Y}_{M,i} \leftarrow \text{GMM}(\tilde{X}_{M,i}, c_i)$ ▷ Classes do GMM
        $\text{RI}_{M,i} \leftarrow \text{RandIndex}(Y_i, \tilde{Y}_{M,i})$
        $\text{FM}_{M,i} \leftarrow \text{FowlkesMallows}(Y_i, \tilde{Y}_{M,i})$
        $\text{VS}_{M,i} \leftarrow \text{VScore}(Y_i, \tilde{Y}_{M,i})$
        $\text{CH}_{M,i} \leftarrow \text{CalinskiHarabasz}(\tilde{X}_{M,i}, \tilde{Y}_{M,i})$
        $\text{SS}_{M,i} \leftarrow \text{SilhouetteScore}(\tilde{X}_{M,i}, \tilde{Y}_{M,i})$
        $\text{DB}_{M,i} \leftarrow \text{DaviesBouldin}(\tilde{X}_{M,i}, \tilde{Y}_{M,i})$
  for $M \in \mathcal{M} \setminus \{\text{RAW,KISOMAP}\}$ **do** ▷ Testes de hipóteses
    for $Q \in \mathcal{Q}$ **do**
      $\mathcal{F}_{M,Q} \leftarrow \text{Friedman}(\{Q_{\text{RAW}}\}, \{Q_{\text{KISOMAP}}\}, \{Q_M\})$
      $\mathcal{N}_{M,Q} \leftarrow \text{Nemenyi}(\{Q_{\text{RAW}}\}, \{Q_{\text{KISOMAP}}\}, \{Q_M\})$
  return p-valores $\{\mathcal{F}_{M,Q}\}$ e $\{\mathcal{N}_{M,Q}\}$ para cada método M e métrica Q

5.1.1 Primeira bateria de experimentos

Cada uma das baterias foi pensada para conter dados com descrições parecidas, como por exemplo área do problema, número de dimensões, quantidade de amostras e quantidade de classes. Na primeira bateria de experimentos, 27 conjuntos de dados considerados de baixo porte foram selecionados. O número de amostras é reduzido em alguns casos para não sofrer com o custo computacional elevado do K-ISOMAP. Uma breve descrição de cada conjunto de dados seguida dos resultados gerais é apresentada abaixo.

Dataset	n	m	c	d	k
Engine1	383	5	3	3	19
Hopf-Link	1352	3	2	2	36
Pen-Digits (25%)	1873	16	10	4	43
PhishingWebsites (10%)	1105	30	2	2	33
Repliclust-Archetype	1100	3	6	2	33
breast-tissue	106	9	4	2	10
car-evaluation	1728	21	4	4	41
cmc	1473	9	3	2	38
conference_attendance	246	6	2	2	15
diggle_table_a2	310	8	2	2	17
fri_c4_250_100	250	100	2	33	15
hayes-roth	132	4	2	2	11
heart-h	294	13	5	2	17
heart-statlog	270	13	2	6	16
led24 (20%)	640	24	10	13	25
newton_hema	140	3	2	2	11
prnn_synth	250	2	2	2	15
qsar-biodeg	1055	41	2	2	32
rabe_131	50	5	2	3	7
satimage (25%)	1607	36	6	7	40
servo	167	4	2	3	12
spambase (25%)	1150	57	2	2	33
steel-plates-fault	1941	27	7	4	44
tic-tac-toe	958	9	2	8	30
visualizing_environmental	111	3	2	2	10
wisconsin	194	32	2	7	13
xd6	973	9	2	2	31

Tabela 1 – Descrição dos conjuntos de dados da primeira bateria do experimento.

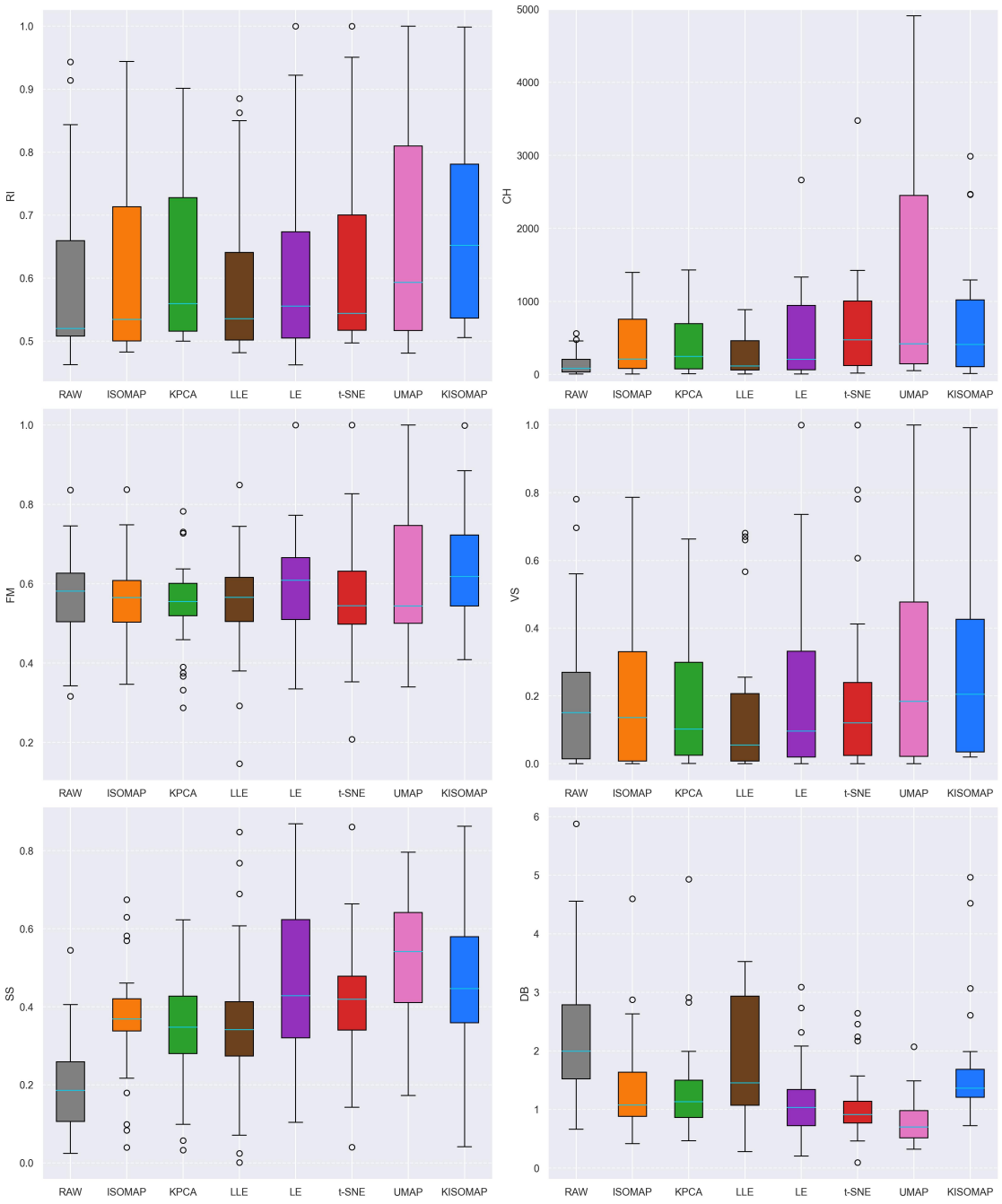


Figura 7 – Boxplots de cada métrica de qualidade de agrupamento dos resultados da primeira bateria de experimentos.

Método	Métrica	p-valor F	p-valor N	Teste F	Teste N
ISOMAP	CH	0.000	0.002	1	1
ISOMAP	DB	0.000	0.000	1	1
ISOMAP	FM	0.000	0.960	1	0
ISOMAP	RI	0.000	0.990	1	0
ISOMAP	SS	0.000	0.002	1	1
ISOMAP	VS	0.000	0.912	1	0
Kernel PCA	CH	0.000	0.001	1	1
Kernel PCA	DB	0.000	0.000	1	1
Kernel PCA	FM	0.000	1.000	1	0
Kernel PCA	RI	0.000	0.438	1	0
Kernel PCA	SS	0.000	0.001	1	1
Kernel PCA	VS	0.000	0.438	1	0
LLE	CH	0.000	0.521	1	0
LLE	DB	0.011	0.292	1	0
LLE	FM	0.000	0.362	1	0
LLE	RI	0.000	0.180	1	0
LLE	SS	0.000	0.012	1	1
LLE	VS	0.000	0.054	1	0
Laplacian Eigenmaps	CH	0.000	0.012	1	1
Laplacian Eigenmaps	DB	0.000	0.000	1	1
Laplacian Eigenmaps	FM	0.001	0.232	1	0
Laplacian Eigenmaps	RI	0.000	0.997	1	0
Laplacian Eigenmaps	SS	0.000	0.000	1	1
Laplacian Eigenmaps	VS	0.000	0.521	1	0
UMAP	CH	0.000	0.000	1	1
UMAP	DB	0.000	0.000	1	1
UMAP	FM	0.002	0.990	1	0
UMAP	RI	0.000	0.399	1	0
UMAP	SS	0.000	0.000	1	1
UMAP	VS	0.003	0.292	1	0
t-SNE	CH	0.000	0.000	1	1
t-SNE	DB	0.000	0.000	1	1
t-SNE	FM	0.000	0.912	1	0
t-SNE	RI	0.000	0.479	1	0
t-SNE	SS	0.000	0.000	1	1
t-SNE	VS	0.001	0.232	1	0
Total				36/36	16/36

Tabela 2 – Resultados dos testes na primeira bateria de experimentos.

Ao revisar o resultado da Tabela 2, observa-se que apenas as métricas de qualidade internas obtiveram significância estatística segundo o teste de Nemenyi. Isto acontece em todos os métodos, exceto para o LLE. O teste de Friedman confirma que todas as distribuições são distintas em relação ao K-ISOMAP, mas o teste de Nemenyi aponta que apenas as métricas CH, DB e SS de fato apresentam resultados significativamente distintos. Entretanto, de acordo com a Figura 7, os resultados do K-ISOMAP são ligeiramente maiores do que todos os outros métodos de RD nas métricas externas. Isto indica que há alinhamento entre os clusters gerados e os rótulos verdadeiros.

Na métrica DB, o K-ISOMAP perde apenas para o LLE e RAW. Neste caso onde o K-ISOMAP obteve um desempenho pior após a aplicação do método de RD, o embedding pode ter eliminado informações relevantes ou não preservado adequadamente a estrutura dos dados necessária para a formação de clusters coesos segundo a métrica DB. Uma possível explicação é que os dados brutos já possuíam uma dimensionalidade intrínseca que favorecia a formação de clusters sem a necessidade de redução, ou que o método de redução utilizado não foi o mais adequado para capturar a estrutura subjacente dos dados. Mesmo assim, em grande maioria, o K-ISOMAP produziu clusters mais coesos e bem separados que os demais métodos, sugerindo uma estrutura interna consistente nos dados.

Em relação ao ISOMAP, o K-ISOMAP apresentou média e mediana consideravelmente maiores. Mesmo que o teste de Nemenyi tenha apontado apenas as métricas externas com resultados significativos, fica evidente que houve uma melhoria nos resultados do K-ISOMAP. Este fato corrobora a hipótese de que a incorporação de uma função de distância intrínseca baseada na curvatura local no algoritmo ISOMAP é capaz de melhorar a qualidade de agrupamentos obtidos via GMM. Ainda, o K-ISOMAP compete com os métodos clássicos e o estado-da-arte.

Vale destacar que a escolha dos parâmetros para este experimento está associada a uma necessidade de comparar os métodos sob a mesma estrutura. Pode ocorrer que, em diferentes regiões da variedade de dados, o grafo considerado ou a dimensionalidade intrínseca não sejam adequados e introduzam degradações na topologia da variedade, como acontece, por exemplo, no *efeito ferradura* (**??**). Neste caso, o ajuste de parâmetros individuais deve ser considerado para melhores comparações entre métodos e conjuntos de dados específicos.

5.1.2 Segunda bateria de experimentos

No segunda bateria de experimentos, foram selecionados 15 conjuntos de dados reais com um grande número de características, alguns deles com mais características do que amostras. Antes da redução de dimensionalidade, em alguns casos, utiliza-se o PCA para reduzir o número de características, mantendo ainda uma alta explicabilidade dos dados. Em conjuntos com uma grande quantidade de amostras, apenas uma pequena parcela foi escolhida, a fim de não extrapolar o custo computacional do K-ISOMAP. A maioria dos conjuntos desta bateria é utilizada em tarefas de classificação e agrupamento de imagens. Uma breve descrição de cada conjunto é seguida dos resultados e é apresentada abaixo.

Dataset	n	m	**PCA features**	c	d	k
AP-Omentum-Kidney	337	10937	90	2	9	18
AP_Breast_Kidney	604	10935	500	2	16	24
AP_Endometrium_Breast	405	10935	400	2	19	20
AP_Ovary_Lung	324	10935	100	2	12	18
COIL-20	1440	16384	80	20	3	37
F-MNIST (5%)	3000	784	100	10	8	54
MNIST (5%)	3000	784	100	10	8	54
OVA_Uterus	1545	10935	100	2	9	39
Olivetti-Faces	400	4096	100	40	6	20
cnae-9	1080	856	50	9	2	32
eating	945	6373	100	7	12	30
har (10%)	1029	561	100	6	13	32
leukemia	72	7129	40	2	7	8
micro-mass	360	1300	100	10	4	18
oh5.wc	918	3012	40	10	10	30

Tabela 3 – Descrição dos conjuntos de dados da segunda bateria do experimento.

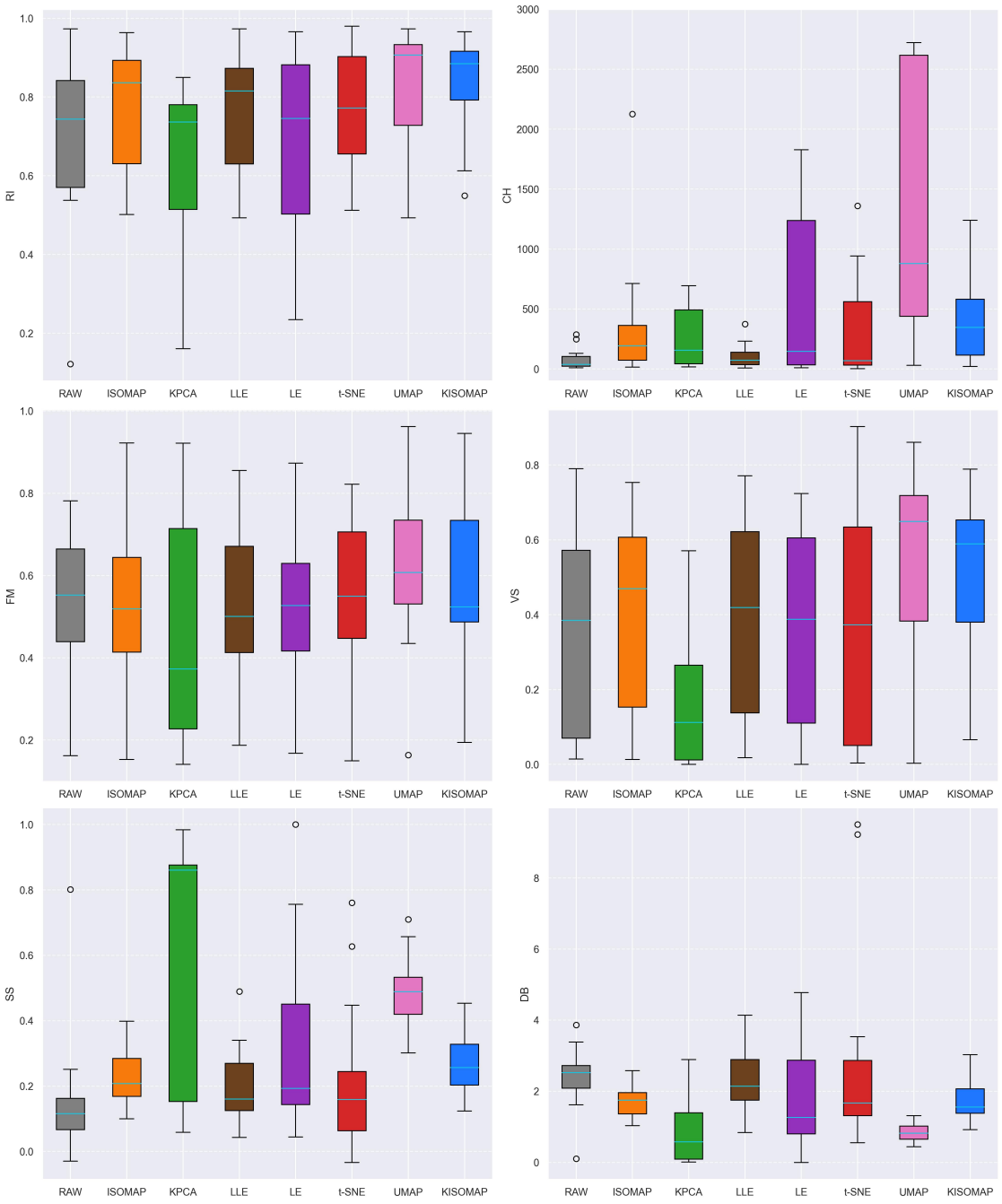


Figura 8 – Boxplots de cada métrica de qualidade de agrupamento dos resultados da segunda bateria de experimentos.

Método	Métrica	p-valor F	p-valor N	Teste F	Teste N
ISOMAP	CH	0.000	0.010	1	1
ISOMAP	DB	0.001	0.001	1	1
ISOMAP	FM	0.002	0.982	1	0
ISOMAP	RI	0.001	0.745	1	0
ISOMAP	SS	0.000	0.029	1	1
ISOMAP	VS	0.004	0.517	1	0
Kernel PCA	CH	0.000	0.017	1	1
Kernel PCA	DB	0.001	0.001	1	1
Kernel PCA	FM	0.015	0.848	1	0
Kernel PCA	RI	0.001	0.228	1	0
Kernel PCA	SS	0.001	0.002	1	1
Kernel PCA	VS	0.000	0.017	1	1
LLE	CH	0.000	0.408	1	0
LLE	DB	0.005	0.228	1	0
LLE	FM	0.041	1.000	1	0
LLE	RI	0.002	0.848	1	0
LLE	SS	0.002	0.017	1	1
LLE	VS	0.038	0.929	1	0
Laplacian Eigenmaps	CH	0.001	0.046	1	1
Laplacian Eigenmaps	DB	0.085	0.110	0	0
Laplacian Eigenmaps	FM	0.022	0.517	1	0
Laplacian Eigenmaps	RI	0.002	0.982	1	0
Laplacian Eigenmaps	SS	0.006	0.029	1	1
Laplacian Eigenmaps	VS	0.005	0.228	1	0
UMAP	CH	0.000	0.000	1	1
UMAP	DB	0.000	0.000	1	1
UMAP	FM	0.017	0.046	1	1
UMAP	RI	0.038	0.110	1	0
UMAP	SS	0.000	0.000	1	1
UMAP	VS	0.015	0.017	1	1
t-SNE	CH	0.000	0.029	1	1
t-SNE	DB	0.014	0.310	1	0
t-SNE	FM	0.022	0.228	1	0
t-SNE	RI	0.005	0.310	1	0
t-SNE	SS	0.002	0.982	1	0
t-SNE	VS	0.017	0.982	1	0
Total				35/36	16/36

Tabela 4 – Resultados dos testes na segunda bateria de experimentos.

Com base nos resultados da Tabela 4, percebe-se que o K-ISOMAP mais uma vez obtém significância estatística sobre o ISOMAP padrão nas métricas internas CH, DB e SS. Entretanto, desta vez o padrão não se repete nos demais métodos, pois o teste de Nemenyi não aponta a mesma significância. O teste de Friedman confirma que todas as distribuições são distintas, exceto a do Laplacian Eigenmaps. As métricas CH e SS foram as que apresentaram um desempenho consistente em comparação ao ISOMAP. Para a métrica DB, entretanto, pode ser considerado um empate por conta das médias muito próximas. Nesta bateria, os resultados internos em comparação com os externos não são favoráveis para o K-ISOMAP. Isto demonstra que, em grandes conjuntos de dados, o método pode não ser uma boa escolha, tanto computacionalmente quanto quantitativamente. Porém, as métricas externas parecem promissoras, ao passo que o K-ISOMAP em geral é melhor que todos os outros métodos de RD, exceto o UMAP.

O método K-ISOMAP continua a demonstrar competitividade, evidenciada pela melhoria na qualidade dos agrupamentos em diversos conjuntos de dados. Essa melhoria sugere que o K-ISOMAP é capaz de capturar eficientemente as relações locais dos dados por meio da incorporação de uma métrica de curvatura, superando limitações do ISOMAP tradicional em cenários onde a preservação de similaridades globais é menos eficaz. Em particular, quando as métricas externas do K-ISOMAP apresentam valores elevados, a introdução da medida de curvatura demonstra potencial para identificar regiões de decisão entre classes ou grupos, destacando sua capacidade de aprimorar a discriminação de estruturas complexas. Esses resultados reforçam a hipótese de que a integração de medidas de curvatura ao ISOMAP pode otimizar a qualidade de agrupamentos via GMM, tornando o K-ISOMAP uma escolha robusta para tarefas que demandam a preservação de similaridades intrínsecas dos dados, mesmo em contextos onde métodos tradicionais falham ou apresentam desempenho inferior.

De acordo com os boxplots apresentados na Figura 8, observa-se que o K-ISOMAP supera os métodos KPCA, LLE e Laplacian Eigenmaps na maioria das métricas avaliadas. No entanto, ao comparar especificamente com o KPCA, o K-ISOMAP demonstra superioridade em todas as métricas, exceto no Silhouette Score, onde o KPCA se destaca em relação aos demais métodos. A exceção observada no Silhouette Score sugere que, embora o KPCA possa gerar clusters

internamente mais coesos em determinados cenários, o K-ISOMAP apresenta um desempenho global superior, destacando-se como uma ferramenta promissora para tarefas que exigem a preservação de estruturas complexas e relações não lineares nos dados. Esses resultados reforçam que o K-ISOMAP é uma alternativa robusta e eficaz para a análise de dados de alta dimensionalidade, especialmente quando comparado a métodos clássicos e ao estado da arte.

5.2 Discussão

A partir dos resultados obtidos, nota-se que, mesmo em dados com uma alta quantidade de características (m), o K-ISOMAP consegue superar o ISOMAP na qualidade dos clusters resultantes com GMM na maioria dos conjuntos de dados analisados. Além disso, sua performance é competitiva em relação ao estado-da-arte e os algoritmos clássicos. Por outro lado, a principal limitação do K-ISOMAP é o seu alto custo computacional. De fato, o K-ISOMAP possui complexidade $O(n^3 + n^2m^2 + nm^3)$. Os métodos ISOMAP, Kernel PCA, LLE e Laplacian Eigenmaps possuem complexidade computacional de $O(n^3)$ devido ao cáculo de autovalores e autovetores de matrizes de similaridades. O algoritmo t-SNE utilizado possui complexidade de $O(n^2)$. Por sua vez, o algoritmo UMAP possui complexidade $O(n^{1.14})$, sendo o mais adequado para conjuntos de dados de grande escala.

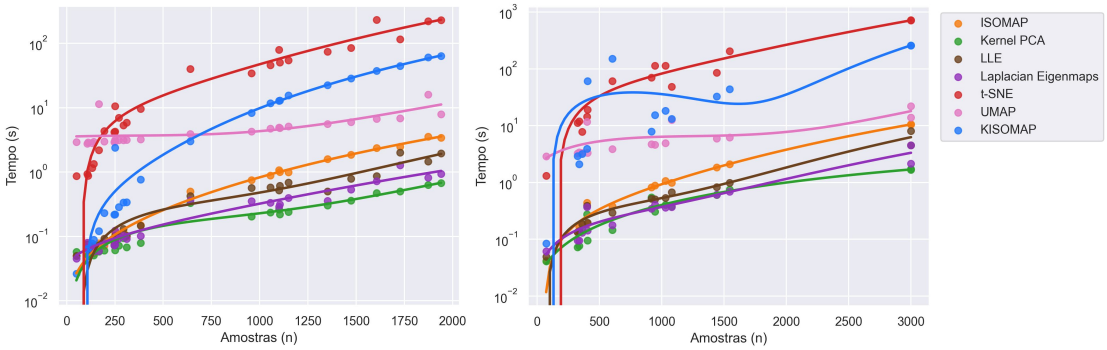


Figura 9 – Tempo de execução de cada método de RD nos conjuntos de dados da primeira e segunda bateria do experimento, respectivamente.

A Figura 9 compara os tempos de execução dos métodos. Note que, na segunda bateria, o gargalo do K-ISOMAP é evidente por conta da alta quantidade de

características. Observa-se também que o tempo de execução do t-SNE foi o mais alto. Isto se deve ao fato que não existe solução analítica para o critério que está sendo otimizado. Em vez disso, a solução é aproximada utilizando gradiente descendente. Pelo mesmo motivo, o UMAP acaba sendo o terceiro mais lento. Porém, é notável sua escalabilidade em conjuntos de dados com alta quantidade de amostras e características (MCINNES et al., 2018).

Por fim, os resultados evidenciam que o desempenho dos métodos de redução de dimensionalidade é fortemente influenciado pela métrica de similaridade adotada e pelas métricas de qualidade utilizadas para avaliar o agrupamento. Essas variações destacam a importância de selecionar algoritmos adequados para diferentes tarefas e tipos de conjuntos de dados, considerando suas características intrínsecas. Nesse contexto, o K-ISOMAP emerge como um método promissor, com potencial para otimizações tanto teóricas, como no cálculo do K-grafo, quanto computacionais, visando um equilíbrio entre eficiência e desempenho na redução de dimensionalidade. Alternativas supervisionadas e semi-supervisionadas podem ser estudadas, unificando similaridades extrínsicas e intrínsecas. Tais aprimoramentos podem consolidar o K-ISOMAP como uma ferramenta robusta e versátil para a análise de dados complexos e de alta dimensionalidade.

Todas as figuras e gráficos apresentados nesta dissertação foram desenvolvidos pelo autor. O código utilizado para gerar os resultados pode ser encontrado em <https://github.com/gustavochavari/curvature-based-isomap>.

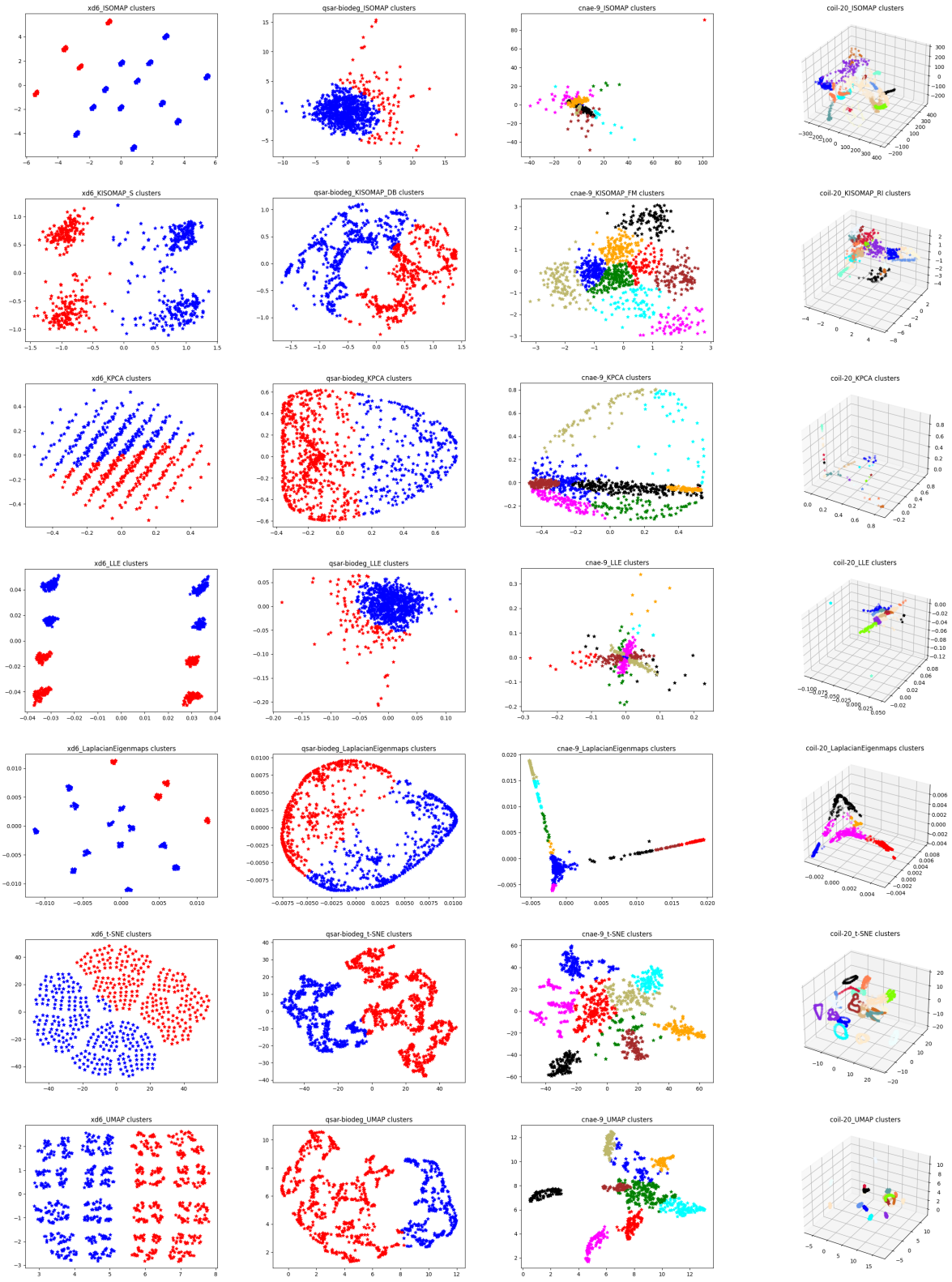


Figura 10 – Redução de dimensionalidade e agrupamentos obtidos com o GMM dos conjunto xd6, qsar-biodeg, cnae-9 e coil-20.

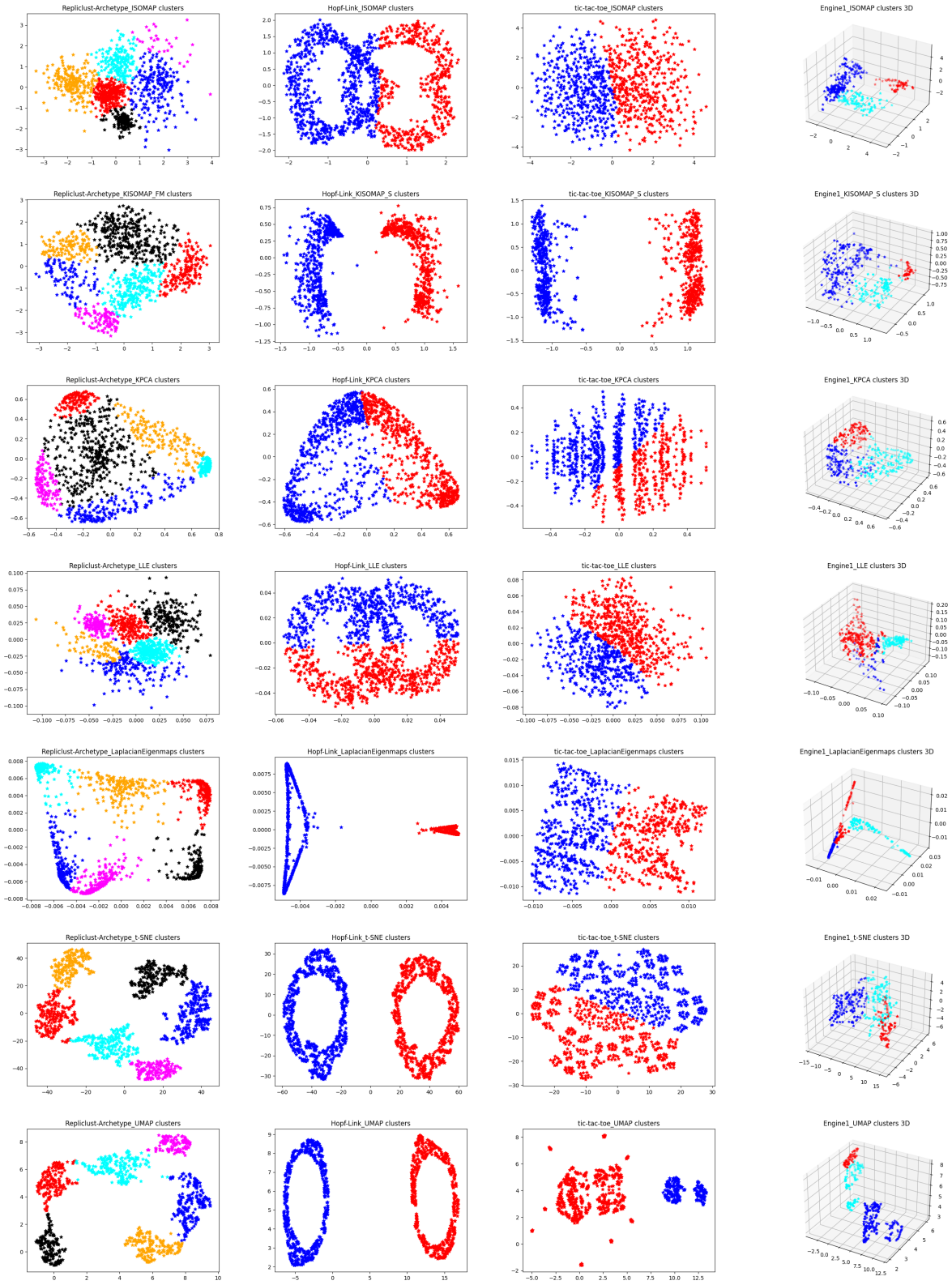


Figura 11 – Redução de dimensionalidade e agrupamentos obtidos com o GMM dos conjunto Repliclust-Archetype, Hopf-Link, tic-tac-toe e Engine-1.

Capítulo 6

Conclusão

Neste trabalho, foi discutida a atuação do método K-ISOMAP em redução de dimensionalidade em tarefas de agrupamento. Utilizando uma nova definição de similaridade por curvatura, o método explora a geometria das curvas da variedade de dados ao aprender uma métrica de distância adaptada para capturar a curvatura local. Constatou-se que o K-ISOMAP trouxe melhorias significativas no desempenho de agrupamentos via GMM, em comparação com o método ISOMAP padrão. Além disso, o K-ISOMAP se mostrou competitivo ao ser comparado com algoritmos de aprendizado de variedades do estado-da-arte. Mesmo sendo uma medida empírica, ela se mostrou consistente em medir similaridades. Isto motiva o aprofundamento no conceito de curvatura em caminhos mínimos em grafos, buscando outras medidas de similaridade e aplicações na área.

A investigação sobre dados de alta dimensão revelou os desafios impostos pela maldição da dimensionalidade, principalmente sobre o custo computacional do K-ISOMAP. Mesmo assim, a principal vantagem do algoritmo K-ISOMAP proposto, em comparação com outros métodos de aprendizado de métricas e variedades, é a utilização da curvatura como uma forma de medir similaridade. Algoritmos tradicionais muitas vezes sofrem com métricas convencionais de distância que pecam

ao preservar propriedades de conjuntos de dados complexos na imersão em um espaço menor. Ao realizar que a curvatura é um invariante por isometrias locais, o K-ISOMAP utiliza esta medida intrínseca para encontrar similaridades nos dados no espaço original. E, portanto, os dados transformados pelo MDS preservam as similaridades no espaço reduzido.

Por meio de extensivas avaliações empíricas em conjuntos de dados de referência, demonstra-se o desempenho superior do K-ISOMAP em comparação ao seu método motivador ISOMAP. Melhorias significativas na qualidade dos agrupamentos via GMM foram observadas, especialmente em cenários de dados complexos e de alta dimensão. Foi discutida a sensibilidade do método ao lidar com outliers e ruído, embora tal sensibilidade seja muito menor do que a do ISOMAP. Em regiões de alta curvatura, uma quantidade de vizinhos maior é necessária para descrever os planos tangentes precisamente. Embora possa ser adaptado para melhorar o grafo de vizinhanças neste sentido, a flexibilidade desta estrutura é um dos seus principais pontos fortes, tornando-a aplicável a conjuntos de dados de diferentes domínios e com diferentes medidas de curvatura. Sua natureza geométrica permite uma integração com algoritmos de agrupamento existentes, o que facilita a escolha do aprendizado de métricas não supervisionado.

Olhando para o futuro, várias direções se apresentam para investigações em aprendizado de métricas e variedades. A melhoria da escalabilidade e eficiência do nosso método para lidar com conjuntos de dados maiores permanece um objetivo primordial. Novas estimativas de curvatura podem ser exploradas a partir da roupagem da geometria diferencial, que inclusive já é utilizada em métodos do estado-da-arte. Além disso, investigar os fundamentos teóricos do aprendizado não supervisionado com métricas de curvatura pode fornecer informações importantes sobre o funcionamento do K-ISOMAP. Compreender como a curvatura se comporta em diferentes geometrias e distribuições de dados é crucial para o ajuste do método. Outro aspecto a ser explorado é a ampliação da estrutura para incorporar paradigmas de aprendizado semissupervisionados, aumentando sua aplicabilidade em cenários onde há escassez de dados rotulados ou onde a experiência humana pode ser necessária para orientar o processo de agrupamento. Outra possibilidade é a generalização do K-ISOMAP para problemas *out-of-sample*, a partir de sua consideração como um método de *kernel*.

Em suma, este trabalho apresenta os fundamentos da área de aprendizado de métricas e variedades, oferecendo uma abordagem fundamentada para reconhecer padrões complexos em conjuntos de dados. Ao pairar sobre as teorias da geometria diferencial e aprendizado de variedades, o K-ISOMAP fornece uma maneira poderosa para analisar e explorar dados de alta dimensão. Abrindo caminho para avanços futuros no aprendizado de máquina e no reconhecimento de padrões, o método alavanca a consideração da curvatura no agrupamento de dados em outras tarefas de aprendizado de máquina.

Apêndice A

Tabelas do experimento

No Apêndice desta dissertação, encontram-se as tabelas referentes às duas baterias de experimentos realizados. Em cada tabela, os valores em negrito destacam os maiores resultados obtidos para cada conjunto de dados, indicando o melhor desempenho entre os métodos avaliados. Adicionalmente, os valores sublinhados indicam se o método K-ISOMAP ou ISOMAP foi o melhor para aquele conjunto de dados. Estas tabelas fornecem uma visão abrangente e comparativa dos resultados, reforçando as análises apresentadas ao longo do trabalho.

Dataset	RAW	ISOMAP	KPCA	LLE	LE	t-SNE	UMAP	KISOMAP
Engine1	0.518	0.535	0.599	0.559	0.635	0.556	0.532	**0.653**
Hopf-Link	0.504	0.689	0.568	0.500	**1.000**	**1.000**	**1.000**	0.999
Pen-Digits	0.914	0.897	0.901	0.885	0.856	0.951	**0.954**	0.928
PhishingWebsites	0.500	0.500	0.500	0.500	0.500	0.500	0.500	**0.506**
Repliclust-Archetype	0.943	0.944	0.855	0.863	0.922	0.940	0.936	0.901
breast-tissue	0.621	0.632	0.574	0.609	0.622	0.645	0.639	**0.652**
car-evaluation	0.519	0.483	**0.544**	0.482	0.487	0.532	0.481	0.525
cmc	0.463	0.530	**0.559**	0.507	0.462	0.520	0.520	0.543
conference_attendance	0.513	0.617	0.502	0.617	0.522	0.522	**0.705**	0.661
diggle_table_a2	0.742	0.737	0.770	0.559	0.703	0.751	0.770	**0.799**
fri_c4_250_100	0.515	0.500	**0.516**	0.508	0.504	0.504	0.514	**0.516**
hayes-roth	0.500	0.500	0.513	0.498	0.506	**0.519**	0.499	0.516
heart-h	**0.639**	0.602	0.578	0.536	0.591	0.580	0.593	0.604
heart-statlog	0.666	0.747	0.629	0.511	0.499	0.511	0.629	0.736
led24	0.844	**0.873**	0.850	0.813	0.866	0.839	0.870	**0.873**
newton_hema	0.514	0.498	**0.556**	0.538	**0.556**	0.542	**0.556**	**0.556**
prnn_synth	0.592	0.499	**0.730**	0.665	0.725	0.498	0.498	0.588
qsar-biodeg	0.500	0.534	0.504	0.503	0.500	**0.568**	0.516	0.585
rabe_131	0.628	0.493	0.726	0.850	0.628	0.497	0.850	**0.885**
satimage	0.819	0.849	0.857	0.840	0.861	0.857	**0.865**	0.862
servo	0.499	0.497	0.530	0.497	0.548	0.498	**0.866**	0.660
spambase	0.654	0.559	0.501	0.543	0.520	0.656	0.521	**0.763**
steel-plates-fault	0.732	0.744	0.753	0.712	0.644	0.745	0.732	**0.763**
tic-tac-toe	0.512	0.514	0.515	0.500	0.500	0.544	**0.672**	0.517
visualizing_environmental	0.571	0.558	0.552	0.521	0.521	0.558	0.525	**0.586**
wisconsin	0.501	0.506	**0.521**	0.499	0.511	0.517	0.504	0.517
xd6	0.520	0.514	0.502	0.518	**0.601**	0.518	0.518	0.530
Média	0.609	0.613	0.619	0.597	0.622	0.625	0.658	**0.675**
Mediana	0.520	0.535	0.559	0.536	0.556	0.544	0.593	**0.652**
Mínimo	0.463	0.483	0.500	0.482	0.462	0.497	0.481	**0.506**
Máximo	0.943	0.944	0.901	0.885	**1.000**	**1.000**	**1.000**	0.999

Tabela 5 – Avaliaçao dos agrupamentos do GMM de acordo com o Índice de Rand na primeira bateria de experimentos.

Dataset	RAW	ISOMAP	KPCA	LLE	LE	t-SNE	UMAP	KISOMAP
Engine1	131.2	292.0	246.2	115.9	177.3	474.9	**958.3**	325.2
Hopf-Link	476.2	827.5	760.7	885.2	1333.3	3478.0	**12921.3**	2989.4
Pen-Digits	345.9	1198.4	1145.1	525.4	5045.5	1088.5	**12644.7**	1293.5
PhishingWebsites	149.8	774.1	917.1	657.9	919.4	1015.8	**2875.2**	1216.4
Repliclust-Archetype	458.8	538.6	735.0	260.9	**2663.5**	1217.1	3251.7	1011.9
breast-tissue	12.2	87.1	69.0	81.6	**143.3**	96.3	129.8	114.9
car-evaluation	118.2	737.3	478.8	534.1	597.5	664.3	419.7	**786.6**
cmc	266.7	1397.3	1430.0	6163.1	**9405.2**	1424.7	4599.4	2462.8
conference_attendance	33.3	210.6	100.4	**838.7**	205.9	271.3	77.8	267.1
diggle_table_a2	560.4	1155.3	540.2	165.4	255.0	1141.4	382.9	**2468.3**
fri_c4_250_100	6.6	9.5	9.3	5.4	6.1	**96.7**	56.7	12.0
hayes-roth	29.3	79.0	79.0	64.5	67.7	118.5	**163.5**	107.7
heart-h	43.0	455.7	384.8	254.0	**1104.8**	511.4	2025.4	485.6
heart-statlog	44.3	102.4	72.6	24.8	43.0	53.6	**231.4**	107.8
led24	16.0	37.7	26.5	15.5	41.2	33.6	**224.3**	49.7
newton_hema	69.2	63.4	**152.3**	90.3	84.0	125.2	84.4	91.2
prnn_synth	167.7	185.3	148.3	136.6	169.4	**447.6**	397.2	279.5
qsar-biodeg	81.3	251.6	350.0	89.7	698.9	**761.3**	728.6	747.9
rabe_131	12.8	5.9	23.1	18.3	19.3	17.0	**51.1**	31.2
satimage	344.8	1133.4	1036.2	254.0	490.8	1089.2	**4910.5**	915.2
servo	32.3	91.7	63.2	58.3	60.1	185.2	**228.9**	89.7
spambase	34.1	74.1	**1167.8**	63.9	970.2	995.7	664.6	1026.2
steel-plates-fault	243.9	838.1	656.6	395.7	8267.0	928.5	**8522.5**	1147.7
tic-tac-toe	102.6	105.5	104.4	70.4	95.6	153.1	**1716.6**	409.8
visualizing_environmental	81.3	103.5	83.1	57.6	52.8	**140.9**	122.0	105.7
wisconsin	42.7	55.1	47.0	16.3	23.3	35.5	**116.1**	61.2
xd6	121.4	454.0	471.6	649.8	457.3	**654.1**	592.3	467.1
Média	149.1	417.2	418.5	462.7	1236.9	637.8	**2188.8**	706.4
Mediana	81.3	210.6	246.2	115.9	205.9	**474.9**	419.7	409.8
Mínimo	6.6	5.9	9.3	5.4	6.1	17.0	**51.1**	12.0
Máximo	560.4	1397.3	1430.0	6163.1	9405.2	3478.0	**12921.3**	2989.4

Tabela 6 – Avaliaçao dos agrupamentos do GMM de acordo com o índice de Calisnki-Harabasz na primeira bateria de experimentos.

Dataset	RAW	ISOMAP	KPCA	LLE	LE	t-SNE	UMAP	KISOMAP
Engine1	0.491	0.527	0.570	0.557	0.609	0.526	0.526	**0.636**
Hopf-Link	0.504	0.697	0.569	0.499	**1.000**	**1.000**	**1.000**	0.999
Pen-Digits	0.607	0.544	0.536	0.566	0.553	0.760	**0.777**	0.649
PhishingWebsites	0.619	0.572	0.558	0.571	**0.635**	0.563	0.561	0.578
Repliclust-Archetype	0.836	**0.837**	0.582	0.675	0.772	0.827	0.813	0.714
breast-tissue	0.316	0.362	0.390	0.380	0.335	0.377	0.354	**0.409**
car-evaluation	0.423	0.388	0.459	0.391	0.400	0.444	0.375	**0.646**
cmc	0.478	0.392	0.367	0.429	**0.479**	0.395	0.395	0.427
conference_attendance	0.642	0.745	0.626	0.745	0.653	0.653	**0.814**	0.790
diggle_table_a2	0.746	0.741	0.782	0.619	0.708	0.755	0.782	**0.808**
fri_c4_250_100	0.535	0.543	0.538	0.657	0.696	**0.707**	0.517	0.549
hayes-roth	0.510	0.510	0.528	0.510	0.515	0.534	0.534	**0.571**
heart-h	0.581	0.442	0.375	0.472	0.426	0.376	0.443	**0.461**
heart-statlog	0.669	**0.748**	0.637	0.571	0.509	0.516	0.631	0.746
led24	0.343	0.376	0.287	0.147	0.365	0.208	0.372	**0.408**
newton_hema	0.582	0.565	0.555	0.534	0.567	0.538	0.567	**0.583**
prnn_synth	0.635	0.497	**0.730**	0.663	0.725	0.496	0.496	0.680
qsar-biodeg	0.589	0.583	**0.630**	0.613	0.609	0.590	0.541	0.618
rabe_131	0.644	0.621	0.727	0.849	0.668	0.500	0.849	**0.885**
satimage	0.604	0.586	0.607	0.587	0.628	0.610	**0.634**	0.620
servo	0.569	0.595	0.594	0.593	0.610	0.592	**0.892**	0.732
spambase	0.665	0.677	0.517	0.652	0.701	0.668	0.703	**0.776**
steel-plates-fault	0.364	0.347	0.332	0.292	0.385	0.353	0.340	**0.450**
tic-tac-toe	0.543	0.549	0.548	0.525	0.523	0.565	**0.716**	0.539
visualizing_environmental	0.594	0.567	0.563	0.531	0.534	0.564	0.525	**0.614**
wisconsin	0.505	0.511	0.521	0.568	0.511	0.515	0.504	**0.520**
xd6	0.546	0.588	0.529	0.544	**0.664**	0.545	0.544	0.570
Média	0.561	0.560	0.543	0.546	0.584	0.562	0.600	**0.629**
Mediana	0.581	0.565	0.555	0.566	0.609	0.545	0.544	**0.618**
Mínimo	0.316	0.347	0.287	0.147	0.335	0.208	0.340	**0.408**
Máximo	0.836	0.837	0.782	0.849	**1.000**	**1.000**	**1.000**	0.999

Tabela 7 – Avaliaçao dos agrupamentos do GMM de acordo com o índice de Folkes-Mallows na primeira bateria de experimentos.

Dataset	RAW	ISOMAP	KPCA	LLE	LE	t-SNE	UMAP	KISOMAP
Engine1	0.164	0.228	0.228	0.215	**0.339**	0.215	0.232	0.325
Hopf-Link	0.007	0.328	0.102	0.000	**1.000**	**1.000**	**1.000**	0.992
Pen-Digits	0.697	0.675	0.663	0.660	0.683	0.809	**0.825**	0.697
PhishingWebsites	0.002	0.000	0.006	0.001	0.003	0.003	0.004	**0.019**
Repliclust-Archetype	0.781	**0.786**	0.609	0.681	0.736	0.781	0.768	0.690
breast-tissue	0.137	0.181	0.198	0.136	0.169	0.208	0.216	**0.256**
car-evaluation	0.175	0.023	0.181	0.142	0.023	0.169	0.003	**0.050**
cmc	**0.020**	0.012	0.029	0.008	0.020	0.005	0.005	**0.020**
conference_attendance	0.009	0.000	0.001	0.000	0.010	0.010	0.023	**0.037**
diggle_table_a2	0.393	0.383	0.541	0.198	0.325	0.413	0.541	**0.584**
fri_c4_250_100	0.019	0.000	0.020	0.007	0.000	0.007	**0.021**	0.020
hayes-roth	0.008	0.007	0.017	0.001	0.013	**0.024**	0.000	0.020
heart-h	0.156	0.205	0.176	0.054	0.184	0.178	0.184	**0.209**
heart-statlog	0.252	**0.392**	0.243	0.012	0.000	0.023	0.209	0.386
led24	0.410	0.439	0.297	0.124	0.420	0.233	0.414	**0.447**
newton_hema	0.042	0.004	0.088	0.060	0.096	0.066	0.096	**0.108**
prnn_synth	0.212	0.001	**0.373**	0.255	0.364	0.000	0.000	0.156
qsar-biodeg	0.048	**0.183**	0.049	0.033	0.100	0.120	0.024	0.146
rabe_131	0.375	0.136	0.357	0.671	0.218	0.017	0.671	**0.673**
satimage	0.560	0.601	0.622	0.567	0.631	0.607	**0.641**	0.623
servo	0.000	0.155	0.093	0.158	0.091	0.162	**0.648**	0.205
spambase	0.258	0.050	0.002	0.023	0.037	0.246	0.034	**0.405**
steel-plates-fault	0.281	0.334	0.302	0.221	0.303	**0.369**	0.339	0.336
tic-tac-toe	0.007	0.008	0.010	0.000	0.000	0.087	**0.223**	0.025
visualizing_environmental	0.151	0.107	0.098	0.036	0.046	0.103	0.042	**0.193**
wisconsin	0.006	0.016	0.033	0.000	0.019	0.028	0.010	**0.032**
xd6	0.033	0.000	0.004	0.031	**0.096**	0.035	0.032	0.023
Média	0.193	0.195	0.198	0.159	0.219	0.219	0.267	**0.284**
Mediana	0.151	0.136	0.102	0.054	0.096	0.120	0.184	**0.205**
Mínimo	0.000	0.000	0.001	0.000	0.000	0.000	0.000	**0.019**
Máximo	0.781	0.786	0.663	0.681	**1.000**	**1.000**	**1.000**	0.992

Tabela 8 – Avaliaçao dos agrupamentos do GMM de acordo com o V-Score na primeira bateria de experimentos.

Dataset	RAW	ISOMAP	KPCA	LLE	LE	t-SNE	UMAP	KISOMAP
Engine1	0.241	0.417	0.376	0.331	0.437	0.377	**0.579**	0.471
Hopf-Link	0.264	0.359	0.347	0.391	0.636	0.618	**0.795**	0.626
Pen-Digits	0.242	0.372	0.348	0.300	0.699	0.446	**0.701**	0.402
PhishingWebsites	0.254	0.582	0.455	0.690	**0.755**	0.472	0.664	0.593
Repliclust-Archetype	0.335	0.368	0.380	0.278	0.611	0.470	**0.623**	0.447
breast-tissue	0.097	0.381	0.368	0.414	0.424	0.381	**0.459**	0.443
car-evaluation	0.099	0.372	0.279	0.287	0.384	0.287	0.231	**0.759**
cmc	0.234	0.446	0.433	0.768	**0.869**	0.521	0.750	0.566
conference_attendance	0.137	0.630	0.304	**0.848**	0.590	0.500	0.437	0.684
diggle_table_a2	0.545	0.675	0.623	0.395	0.595	0.664	0.541	**0.863**
fri_c4_250_100	0.025	0.040	0.033	0.024	0.155	**0.861**	0.173	0.042
hayes-roth	0.186	0.371	0.359	0.320	0.332	0.419	**0.659**	0.433
heart-h	0.254	0.569	0.450	0.424	0.683	0.469	**0.709**	0.508
heart-statlog	0.143	0.256	0.210	0.351	0.165	0.155	**0.409**	0.283
led24	0.033	0.083	0.057	0.001	0.120	0.040	**0.232**	0.173
newton_hema	0.400	0.368	**0.495**	0.406	0.412	0.444	0.437	0.393
prnn_synth	0.406	0.424	0.398	0.343	0.433	0.571	0.556	**0.650**
qsar-biodeg	0.272	0.366	0.288	0.482	0.578	0.373	0.413	**0.461**
rabe_131	0.214	0.179	0.316	0.269	0.309	0.222	**0.439**	0.360
satimage	0.101	0.359	0.449	0.268	0.428	0.357	0.580	0.358
servo	0.136	0.361	0.282	0.284	0.287	0.472	**0.558**	0.448
spambase	0.034	0.461	0.427	0.607	**0.811**	0.393	0.797	0.544
steel-plates-fault	0.168	0.324	0.250	0.183	0.804	0.324	0.408	**0.852**
tic-tac-toe	0.095	0.099	0.099	0.071	0.104	0.142	**0.615**	0.279
visualizing_environmental	0.374	0.416	0.428	0.341	0.427	**0.484**	0.471	0.436
wisconsin	0.180	0.217	0.192	0.242	0.127	0.152	**0.340**	0.224
xd6	0.111	0.352	0.308	0.412	0.359	0.376	0.372	0.323
Média	0.207	0.365	0.332	0.360	0.464	0.407	**0.517**	0.468
Mediana	0.186	0.368	0.348	0.341	0.428	0.419	**0.541**	0.447
Mínimo	0.025	0.040	0.033	0.001	0.104	0.040	**0.173**	0.042
Máximo	0.545	0.675	0.623	0.848	**0.869**	0.861	0.797	0.863

Tabela 9 – Avaliaçao dos agrupamentos do GMM de acordo com o Sihuette Score na primeira bateria de experimentos.

Dataset	RAW	ISOMAP	KPCA	LLE	LE	t-SNE	UMAP	KISOMAP
Engine1	**1.627**	0.886	1.132	1.484	1.303	0.913	0.500	1.323
Hopf-Link	**1.611**	1.190	1.244	1.086	0.708	0.614	0.321	0.974
Pen-Digits	**1.729**	0.971	1.103	1.878	0.586	0.852	0.437	1.422
PhishingWebsites	**1.907**	0.881	0.619	1.030	0.520	0.824	0.519	1.283
Repliclust-Archetype	0.934	0.878	1.059	**3.047**	0.526	0.739	0.491	1.079
breast-tissue	**2.124**	0.856	0.680	0.742	0.806	0.815	0.710	1.488
car-evaluation	**2.749**	1.077	1.483	1.391	1.209	1.396	1.489	1.967
cmc	**1.431**	0.917	0.776	0.280	0.203	0.739	0.468	0.908
conference_attendance	**2.462**	0.621	1.421	0.385	1.034	0.849	0.521	1.162
diggle_table_a2	0.661	0.418	0.467	**1.052**	0.840	0.460	0.664	0.723
fri_c4_250_100	**5.878**	4.598	4.931	3.524	1.504	0.096	2.070	4.965
hayes-roth	**2.004**	1.143	1.134	1.202	1.247	0.937	0.553	1.202
heart-h	1.319	0.856	0.785	0.782	0.737	0.843	0.404	**4.521**
heart-statlog	2.384	1.557	1.865	**3.090**	2.318	2.169	1.049	1.989
led24	**3.564**	2.630	2.828	3.515	2.082	2.644	1.477	2.608
newton_hema	1.174	1.210	0.847	1.055	1.000	0.957	0.890	**1.305**
prnn_synth	0.904	1.016	1.180	**1.230**	1.093	0.677	0.701	1.121
qsar-biodeg	**3.132**	1.723	0.707	2.838	0.744	1.071	1.050	1.234
rabe_131	1.682	**1.786**	1.331	1.514	1.136	1.571	0.893	1.374
satimage	**2.876**	1.007	0.876	1.968	1.051	0.939	0.557	1.288
servo	**2.118**	1.175	1.561	1.453	1.575	0.772	0.616	1.431
spambase	**4.556**	2.380	0.898	3.157	0.228	0.947	0.798	1.611
steel-plates-fault	1.752	1.131	1.515	**2.594**	0.587	1.154	0.915	1.363
tic-tac-toe	2.942	2.875	2.912	**3.492**	3.090	2.456	0.515	3.069
visualizing_environmental	0.946	0.864	0.997	1.132	1.377	0.768	0.819	**1.378**
wisconsin	1.997	1.716	1.994	**3.028**	2.734	2.244	1.209	1.757
xd6	**2.827**	1.065	1.279	1.221	1.031	1.127	1.189	1.216
Média	2.196	1.386	1.394	**1.821**	1.158	1.095	0.808	1.695
Mediana	1.997	1.077	1.134	**1.453**	1.034	0.913	0.701	1.363
Mínimo	0.661	0.418	0.467	0.280	0.203	0.096	0.321	**0.723**
Máximo	5.878	4.598	4.931	3.524	3.090	2.644	2.070	**4.965**

Tabela 10 – Avaliaçao dos agrupamentos do GMM de acordo com o índice de Davies-Bouldin na primeira bateria de experimentos.

Dataset	RAW	ISOMAP	KPCA	LLE	LE	t-SNE	UMAP	KISOMAP
AP-Omentum-Kidney	0.738	0.511	0.570	0.703	0.684	0.772	0.722	**0.931**
AP_Breast_Kidney	0.544	0.921	0.509	0.848	0.868	0.767	**0.942**	0.936
AP_Endometrium_Breast	0.538	0.601	0.737	0.761	0.515	0.565	**0.942**	0.613
AP_Ovary_Lung	0.734	0.751	0.521	0.499	0.500	0.539	0.726	**0.821**
COIL-20	0.959	0.936	0.262	0.933	0.234	**0.980**	0.973	0.937
F-MNIST	0.878	0.888	0.844	0.895	0.890	**0.904**	0.907	0.898
MNIST	0.836	0.893	0.777	0.849	0.878	0.902	**0.909**	0.902
OVA_Uterus	0.597	0.502	**0.850**	0.557	0.506	0.512	0.576	0.624
Olivetti-Faces	0.973	0.964	0.165	0.973	0.966	0.971	**0.974**	0.966
cnae-9	0.122	0.660	0.767	0.542	0.454	0.908	**0.924**	0.880
eating	0.744	0.742	0.161	0.737	0.745	0.753	0.731	**0.763**
har	0.833	0.836	0.788	0.863	0.841	0.746	0.845	**0.866**
leukemia	0.540	**0.559**	0.525	0.493	0.496	0.540	0.493	0.549
micro-mass	0.849	0.893	0.742	0.883	0.905	0.900	**0.909**	0.901
oh5.wc	0.797	0.847	0.785	0.815	**0.886**	0.856	0.867	0.885
Média	0.712	0.767	0.600	0.757	0.691	0.774	0.829	**0.831**
Mediana	0.744	0.836	0.737	0.815	0.745	0.772	**0.907**	0.885
Mínimo	0.122	0.502	0.161	0.493	0.234	0.512	0.493	**0.549**
Máximo	0.973	0.964	0.850	0.973	0.966	**0.980**	0.974	0.966

Tabela 11 – Avaliaçao dos agrupamentos do GMM de acordo com o índice de Rand na segunda bateria de experimentos.

Dataset	RAW	ISOMAP	KPCA	LLE	LE	t-SNE	UMAP	KISOMAP
AP-Omentum-Kidney	42.7	55.5	26.8	35.6	34.6	38.8	**208.0**	143.3
AP_Breast_Kidney	26.1	366.6	46.7	36.8	32.7	94.2	**2300.0**	427.5
AP_Endometrium_Breast	25.7	54.7	38.8	17.5	16.0	2.6	**348.7**	56.7
AP_Ovary_Lung	24.8	90.0	27.4	18.6	21.9	25.1	**456.5**	100.2
COIL-20	129.7	2126.0	4130.5	5542.5	5423.0	4669.4	**16721.1**	3520.0
F-MNIST	249.0	713.3	407.0	373.9	1827.2	942.0	**7840.1**	946.1
MNIST	105.8	307.0	694.6	231.2	1080.9	638.6	**2721.0**	464.4
OVA_Uterus	103.4	360.6	272.0	150.1	161.5	483.9	**588.8**	396.4
Olivetti-Faces	18.2	75.1	**1.9 mi**	117.3	166.0	31.1	422.0	72.9
cnae-9	37.0	314.3	577.7	128.1	**4819.2**	1359.5	2511.6	1239.1
eating	58.7	127.7	156.2	66.7	96.6	66.6	**878.5**	131.3
har	287.4	642.6	179.9	82.1	92.8	31.3	**34170.4**	697.0
leukemia	9.8	15.8	17.5	8.5	9.9	32.7	**30.2**	21.0
micro-mass	17.3	192.0	108.5	71.9	1395.4	78.4	**1043.3**	346.5
oh5.wc	21.7	68.2	112.0	58.1	146.5	68.4	**463.4**	150.4
Média	77.2	367.3	**128 k**	462.6	1021.5	570.8	**4713.6**	580.9
Mediana	37.0	192.0	156.2	71.9	146.5	68.4	**878.5**	346.5
Mínimo	9.8	15.8	17.5	8.5	9.9	2.6	**30.2**	21.0
Máximo	287.4	2126.0	**1,9 mi**	5542.5	5423.0	4669.4	**34170.4**	3520.0

Tabela 12 – Avaliaçao dos agrupamentos do GMM de acordo com o índice de Calinski-Harabasz na segunda bateria de experimentos.

Dataset	RAW	ISOMAP	KPCA	LLE	LE	t-SNE	UMAP	KISOMAP
AP-Omentum-Kidney	0.782	0.632	0.718	0.749	0.733	0.822	0.767	**0.946**
AP_Breast_Kidney	0.687	0.922	0.711	0.855	0.873	0.771	**0.943**	0.937
AP_Endometrium_Breast	0.642	0.695	0.857	0.828	0.622	0.694	**0.962**	0.705
AP_Ovary_Lung	0.741	0.760	0.717	0.611	0.527	0.550	0.732	**0.838**
COIL-20	0.624	0.443	0.225	0.446	0.246	**0.806**	0.737	0.424
F-MNIST	0.466	0.479	0.356	0.501	0.505	0.530	**0.543**	0.504
MNIST	0.413	0.519	0.232	0.379	0.475	0.531	**0.565**	0.524
OVA_Uterus	0.741	0.656	**0.922**	0.707	0.657	0.667	0.725	0.763
Olivetti-Faces	0.475	0.283	0.141	0.463	0.350	0.390	**0.478**	0.295
cnae-9	0.330	0.324	0.229	0.258	0.358	0.616	**0.667**	0.493
eating	0.162	0.153	**0.373**	0.187	0.168	0.150	0.163	0.194
har	0.552	0.546	0.533	**0.635**	0.553	0.270	0.607	0.610
leukemia	0.556	0.576	**0.707**	0.604	0.636	0.718	0.518	0.566
micro-mass	0.485	0.514	0.222	0.471	0.564	0.504	**0.571**	0.514
oh5.wc	0.331	0.384	0.217	0.293	0.489	0.340	0.435	**0.481**
Média	0.532	0.526	0.477	0.532	0.517	0.557	**0.628**	0.586
Mediana	0.552	0.519	0.373	0.501	0.527	0.550	**0.607**	0.524
Mínimo	0.162	0.153	0.141	0.187	0.168	0.150	0.163	**0.194**
Máximo	0.782	0.922	0.922	0.855	0.873	0.822	**0.962**	0.946

Tabela 13 – Avaliaçao dos agrupamentos do GMM de acordo com o índice de Folkes-Mallows na segunda bateria de experimentos.

Dataset	RAW	ISOMAP	KPCA	LLE	LE	t-SNE	UMAP	KISOMAP
AP-Omentum-Kidney	0.425	0.088	0.055	0.422	0.362	0.336	0.391	**0.758**
AP_Breast_Kidney	0.149	0.753	0.000	0.629	0.664	0.428	**0.808**	0.789
AP_Endometrium_Breast	0.061	0.216	0.004	0.345	0.119	0.010	**0.716**	0.207
AP_Ovary_Lung	0.375	0.392	0.009	0.041	0.000	0.071	0.375	**0.560**
COIL-20	0.790	0.639	0.189	0.643	0.233	**0.903**	0.860	0.635
F-MNIST	0.562	0.564	0.428	0.580	0.586	0.612	**0.619**	0.589
MNIST	0.460	0.575	0.274	0.419	0.576	0.607	**0.649**	0.612
OVA_Uterus	0.022	0.060	0.001	0.018	0.053	0.031	0.045	**0.066**
Olivetti-Faces	0.785	0.673	0.112	0.771	0.700	0.731	**0.788**	0.674
cnae-9	0.014	0.292	0.256	0.129	0.388	0.672	**0.718**	0.561
eating	0.030	0.013	0.016	0.073	0.038	0.021	0.017	**0.097**
har	0.583	0.550	0.571	0.652	0.624	0.148	**0.655**	0.633
leukemia	0.079	0.091	0.031	**0.145**	0.103	0.004	0.003	0.087
micro-mass	0.660	0.670	0.305	0.614	**0.724**	0.656	0.719	0.671
oh5.wc	0.385	0.470	0.230	0.368	0.555	0.373	**0.558**	0.553
Média	0.359	0.403	0.165	0.390	0.382	0.373	**0.528**	0.499
Mediana	0.385	0.470	0.112	0.419	0.388	0.373	**0.649**	0.589
Mínimo	0.014	0.013	0.000	0.018	0.000	0.004	0.003	**0.066**
Máximo	0.790	0.753	0.571	0.771	0.724	**0.903**	0.860	0.789

Tabela 14 – Avaliaçao dos agrupamentos do GMM de acordo com o V-Score na segunda bateria de experimentos.

Dataset	RAW	ISOMAP	KPCA	LLE	LE	t-SNE	UMAP	KISOMAP
AP-Omentum-Kidney	0.200	0.228	**0.861**	0.155	0.139	0.105	0.521	0.296
AP_Breast_Kidney	0.151	0.352	**0.876**	0.077	0.075	0.118	0.710	0.360
AP_Endometrium_Breast	0.074	0.134	**0.863**	0.154	0.044	-0.007	0.544	0.124
AP_Ovary_Lung	0.072	0.208	**0.876**	0.263	0.108	0.063	0.489	0.224
COIL-20	0.251	0.398	0.865	0.490	**1.000**	0.627	0.657	0.454
F-MNIST	0.115	0.257	0.124	0.160	**0.520**	0.249	0.461	0.270
MNIST	0.032	0.144	0.151	0.043	0.377	0.184	**0.408**	0.203
OVA_Uterus	0.049	0.206	**0.984**	0.276	0.147	0.240	0.431	0.251
Olivetti-Faces	0.174	0.252	**0.885**	0.326	0.382	0.159	0.492	0.257
cnae-9	**0.802**	0.340	0.119	0.340	0.756	0.447	0.553	0.379
eating	0.062	0.162	**0.974**	0.123	0.213	0.008	0.400	0.170
har	0.122	0.176	0.059	0.182	0.178	-0.034	**0.480**	0.175
leukemia	0.116	0.189	0.687	0.233	0.179	**0.761**	0.302	0.204
micro-mass	0.124	0.312	0.290	0.127	**0.643**	0.220	0.519	0.417
oh5.wc	-0.029	0.100	0.154	0.065	0.193	0.063	0.393	**0.259**
Média	0.154	0.230	0.584	0.201	0.330	0.214	**0.491**	0.269
Mediana	0.116	0.208	**0.861**	0.160	0.193	0.159	0.489	0.257
Mínimo	-0.029	0.100	0.059	0.043	0.044	-0.034	**0.302**	0.124
Máximo	0.802	0.398	0.984	0.490	**1.000**	0.761	0.710	0.454

Tabela 15 – Avaliaçao dos agrupamentos do GMM de acordo com o Silhouette Score na segunda bateria de experimentos.

Dataset	RAW	ISOMAP	KPCA	LLE	LE	t-SNE	UMAP	KISOMAP
AP-Omentum-Kidney	2.665	1.963	2.890	2.691	**2.896**	2.417	1.139	1.382
AP_Breast_Kidney	2.614	1.178	0.085	3.403	**3.941**	2.479	0.440	1.418
AP_Endometrium_Breast	3.865	2.578	0.093	4.139	4.775	**9.502**	0.507	3.033
AP_Ovary_Lung	3.387	1.753	0.095	3.141	**3.609**	3.535	0.799	1.773
COIL-20	**1.618**	1.132	0.578	0.838	0.000	0.552	0.462	0.922
F-MNIST	**2.155**	1.645	1.366	1.632	0.740	1.319	0.819	1.422
MNIST	**2.447**	1.730	1.430	2.037	0.862	1.636	0.986	1.558
OVA_Uterus	2.788	1.964	0.012	2.771	**2.857**	1.671	1.229	2.225
Olivetti-Faces	**1.623**	1.168	0.093	0.950	0.921	1.398	0.683	1.380
cnae-9	0.103	1.554	1.129	**3.009**	0.452	0.734	0.626	0.935
eating	2.525	1.765	0.022	2.271	1.711	**3.215**	1.052	1.915
har	2.191	1.871	1.709	1.944	2.017	**9.222**	0.923	2.424
leukemia	**2.530**	1.970	0.224	2.141	1.022	0.813	1.312	2.236
micro-mass	**2.024**	1.034	1.288	1.319	0.515	1.313	0.759	1.133
oh5.wc	**2.866**	2.015	1.871	1.872	1.267	2.522	0.964	1.776
Média	2.360	1.688	0.859	2.277	1.839	**2.822**	0.847	1.702
Mediana	**2.525**	1.753	0.578	2.141	1.267	1.671	0.819	1.558
Mínimo	0.103	**1.034**	0.012	0.838	0.000	0.552	0.440	0.922
Máximo	3.865	2.578	2.890	4.139	4.775	**9.502**	1.312	3.033

Tabela 16 – Avaliaçao dos agrupamentos do GMM de acordo com o índice de Davies-Bouldin na segunda bateria de experimentos.

Referências

BABAEIAN, A. **Multiple Manifold Clustering Using Curvature Constrained Path**. 2018. Disponível em: <https://arxiv.org/abs/1812.02327>.

BABAEIAN, A. et al. Nonlinear subspace clustering using curvature constrained distances. **Pattern Recognition Letters**, v. 68, p. 118–125, 2015. ISSN 0167-8655. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167865515002974>.

BELKIN, M.; NIYOGI, P. Laplacian eigenmaps for dimensionality reduction and data representation. **Neural Computation**, v. 15, n. 6, p. 1373–1396, 2003.

BELLET, A.; HABRARD, A.; SEBBAN, M. **A Survey on Metric Learning for Feature Vectors and Structured Data**. 2014. Disponível em: <https://arxiv.org/abs/1306.6709>.

BELLMAN, R. **Dynamic Programming**. New York: Dover Publications, 2003. 2319–2323 p.

BERNSTEIN, M. et al. **Graph Approximations to Geodesics on Embedded Manifolds**. [S.l.], 2000.

CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics**, Taylor & Francis, v. 3, n. 1, p. 1–27, 1974. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/03610927408827101>.

CARMO, M. do. **Riemannian Geometry**. Birkhäuser, 1992. (Mathematics (Boston, Mass.)). ISBN 9783764334901. Disponível em: <https://books.google.com.br/books?id=uXJQQgAACAAJ>.

CARMO, M. P. do. **Differential Geometry of Curves and Surfaces: Revised and Updated Second Edition**. Estados Unidos: Dover Publications, 2016.

CARTER, K. M.; RAICH, R.; AU2, A. O. H. I. **An Information Geometric Framework for Dimensionality Reduction**. 2008. Disponível em: <https://arxiv.org/abs/0809.4866>.

CHAUDHURI, K.; DASGUPTA, S. Rates of convergence for nearest neighbor classification. In: GHAHRAMANI, Z. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2014. v. 27. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2014/file/db957c626a8cd7a27231adfbf51e20eb-Paper.pdf>.

CHEN, G.; ATEV, S.; LERMAN, G. Kernel spectral curvature clustering (kscc). In: **2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops**. [S.l.: s.n.], 2009. p. 765–772.

CHEN, G.; LERMAN, G. Spectral curvature clustering (scc). **International Journal of Computer Vision**, Springer, v. 81, n. 3, p. 317–330, 2009. Disponível em: <https://doi.org/10.1007/s11263-008-0178-9>.

CHETVERIKOV, D. A simple and efficient algorithm for detection of high curvature points in planar curves. In: PETKOV, N.; WESTENBERG, M. A. (Ed.). **Computer Analysis of Images and Patterns**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003. p. 746–753.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, Sep 1995. ISSN 1573-0565. Disponível em: <https://doi.org/10.1007/BF00994018>.

COX, T.; COX, M. **Multidimensional Scaling**. 2. ed. [S.l.]: Chapman and Hall/CRC, 2000.

DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, PAMI-1, p. 224–227, 1979. Disponível em: <https://api.semanticscholar.org/CorpusID:13254783>.

DIJKSTRA, E. W. A note on two problems in connexion with graphs. **Numer. Math.**, Springer-Verlag, Berlin, Heidelberg, v. 1, n. 1, p. 269–271, dez. 1959. ISSN 0029-599X. Disponível em: <https://doi.org/10.1007/BF01386390>.

DONOHO, D. L.; GRIMES, C. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. **Proceedings of the National Academy of Sciences of the United States of America**, v. 100, n. 10, p. 5591–5596, 2003.

FLOYD, R. W. Algorithm 97: Shortest path. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 5, n. 6, p. 345, jun. 1962. ISSN 0001-0782. Disponível em: <https://doi.org/10.1145/367766.368168>.

FOWLKES, E. B.; MALLOWS, C. L. A method for comparing two hierarchical clusterings. **Journal of the American Statistical Association**, v. 78, p. 553–569, 1983. Disponível em: <https://api.semanticscholar.org/CorpusID:122161673>.

HARTIGAN, J. A.; WONG, M. A. Algorithm as 136: A k-means clustering algorithm. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, [Royal Statistical Society, Oxford University Press], v. 28, n. 1, p. 100–108, 1979. ISSN 00359254, 14679876. Disponível em: <http://www.jstor.org/stable/2346830>.

HINTON, G. E.; ROWEIS, S. Stochastic neighbor embedding. In: BECKER, S.; THRUN, S.; OBERMAYER, K. (Ed.). **Advances in Neural Information Processing Systems**. MIT Press, 2002. v. 15. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf>.

HOLLANDER, M.; WOLFE, D. A.; CHICKEN, E. **Nonparametric Statistical Methods**. Wiley, 2015. ISSN 1940-6347. ISBN 9781119196037. Disponível em: <http://dx.doi.org/10.1002/9781119196037>.

HU, C. et al. Joint unsupervised contrastive learning and robust gmm for text clustering. **Information Processing & Management**, v. 61, n. 1, p. 103529, 2024. ISSN 0306-4573.

HUGHES, G. On the mean accuracy of statistical pattern recognizers. **IEEE Transactions on Information Theory**, v. 14, n. 1, p. 55–63, January 1968.

JOST, J.; LIU, S. Ollivier's ricci curvature, local clustering and curvature-dimension inequalities on graphs. **Discrete & Computational Geometry**, Springer, v. 51, n. 2, p. 300–322, 2014. Disponível em: <https://doi.org/10.1007/s00454-013-9558-1>.

KASA, S.; RAJAN, V. Avoiding inferior clusterings with misspecified gaussian mixture models. **Scientific Reports**, Nature, v. 13, 2023.

KOUTROUMBAS, K.; THEODORIDIS, S. **Pattern recognition**. [S.l.]: Academic Press, 2008.

KURODA, M.; GENG, Z. Acceleration of the em algorithm. **WIREs Computational Statistics**, v. 15, n. 6, p. e1618, 2023.

LEVADA, A. L. M. A curvature based isometric feature mapping. In: **2022 26th International Conference on Pattern Recognition (ICPR)**. Montreal, QC, Canada: [s.n.], 2022. p. 557–563.

LEVADA, A. L. M.; NIELSEN, F.; HADDAD, M. F. C. **Adaptive k-nearest neighbor classifier based on the local estimation of the shape operator**. 2024. Disponível em: <https://arxiv.org/abs/2409.05084>.

LEVINA, E.; BICKEL, P. Maximum likelihood estimation of intrinsic dimension. In: SAUL, L.; WEISS, Y.; BOTTOU, L. (Ed.). **Advances in Neural Information Processing Systems**. MIT Press, 2004. v. 17. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2004/file/74934548253bcab8490ebd74afed7031-Paper.pdf>.

LI, D.; TIAN, Y. Survey and experimental study on metric learning methods. **Neural Networks**, v. 105, p. 447–462, 2018.

MAATEN, L. van der; HINTON, G. Visualizing data using t-sne. **Journal of Machine Learning Research**, v. 9, n. 86, p. 2579–2605, 2008. Disponível em: <http://jmlr.org/papers/v9/vandermaaten08a.html>.

MAHMOOD, H.; MEHMOOD, T.; AL-ESSA, L. A. Optimizing clustering algorithms for anti-microbial evaluation data: A majority score-based evaluation of k-means, gaussian mixture model, and multivariate t-distribution mixtures. **IEEE Access**, v. 11, p. 79793–79800, 2023.

MCINNES, L. et al. Umap: Uniform manifold approximation and projection. **Journal of Open Source Software**, The Open Journal, v. 3, n. 29, p. 861, 2018. Disponível em: <https://doi.org/10.21105/joss.00861>.

MEILă, M.; ZHANG, H. **Manifold learning: what, how, and why**. 2023. Disponível em: <https://arxiv.org/abs/2311.03757>.

NEGRI, R. G. **Reconhecimento de Padrões: um estudo dirigido**. São Paulo: Blucher, 2021.

NG, A.; JORDAN, M.; WEISS, Y. On spectral clustering: Analysis and an algorithm. In: DIETTERICH, T.; BECKER, S.; GHAHRAMANI, Z. (Ed.). **Advances in Neural Information Processing Systems**. MIT Press, 2001.

v. 14. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2001/file/801272ee79cfde7fa5960571fee36b9b-Paper.pdf>.

RAND, W. Objective criteria for the evaluation of clustering methods. **Journal of the American Statistical Association**, Taylor and Francis, v. 66, n. 336, p. 846–850, 1971.

RENGASWAMI, S.; BOURNI, T.; MAROULAS, V. **On the Relation between Graph Ricci Curvature and Community Structure**. 2024. Disponível em: <https://arxiv.org/abs/2407.06197>.

ROBLES-KELLY, A.; HANCOCK, E. R. A riemannian approach to graph embedding. **Pattern Recognition**, v. 40, n. 3, p. 1042–1056, 2007. ISSN 0031-3203. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0031320306002421>.

ROSENBERG, A.; HIRSCHBERG, J. V-measure: A conditional entropy-based external cluster evaluation measure. In: EISNER, J. (Ed.). **Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)**. Prague, Czech Republic: Association for Computational Linguistics, 2007. p. 410–420. Disponível em: <https://aclanthology.org/D07-1043>.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, p. 53–65, 1987. ISSN 0377-0427. Disponível em: <https://www.sciencedirect.com/science/article/pii/0377042787901257>.

ROWEIS, S. T.; SAUL, L. K. Nonlinear dimensionality reduction by locally linear embedding. **Science**, v. 290, n. 5500, p. 2323–2326, 2000. Disponível em: <https://www.science.org/doi/abs/10.1126/science.290.5500.2323>.

SCHÖLKOPF, B.; SMOLA, A.; MÜLLER, K.-R. Kernel principal component analysis. In: GERSTNER, W. et al. (Ed.). **Artificial Neural Networks — ICANN'97**. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997. p. 583–588. ISBN 978-3-540-69620-9.

SUáREZ, J. L.; GARCíA, S.; HERRERA, F. A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges. **Neurocomputing**, v. 425, p. 300–322, 2021. ISSN 0925-2312. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231220312777>.

TANG, J. et al. Visualizing large-scale and high-dimensional data. In: **Proceedings of the 25th International Conference on World Wide Web**. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2016. (WWW '16), p. 287–297. ISBN 9781450341431. Disponível em: <https://doi.org/10.1145/2872427.2883041>.

TASOULIS, S.; PAVLIDIS, N. G.; ROOS, T. Nonlinear dimensionality reduction for clustering. **Pattern Recognition**, v. 107, p. 107508, 2020. ISSN 0031-3203. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0031320320303113>.

TENENBAUM, J. B.; SILVA, V. d.; LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. **Science**, v. 290, p. 2319–2323, 2000.

TIAN, Y.; LUBBERTS, Z.; WEBER, M. **Curvature-based Clustering on Graphs**. 2023. Disponível em: <https://arxiv.org/abs/2307.10155>.

WANG, F.; SUN, J. Survey on distance metric learning and dimensionality reduction in data mining. **Data Min. Knowl. Discov.**, Kluwer Academic Publishers, USA, v. 29, n. 2, p. 534–564, mar. 2015. ISSN 1384-5810. Disponível em: <https://doi.org/10.1007/s10618-014-0356-z>.

WANG, Y. et al. **Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMAP, and PaCMAP for Data Visualization**. 2021. Disponível em: <https://arxiv.org/abs/2012.04456>.

YOU, J.; LI, Z.; DU, J. A new iterative initialization of em algorithm for gaussian mixture models. **PLoS ONE**, v. 18, p. e0284114, 2023.

ZADEH, L. Fuzzy sets. **Information and Control**, v. 8, n. 3, p. 338–353, 1965. ISSN 0019-9958. Disponível em: <https://www.sciencedirect.com/science/article/pii/S001999586590241X>.

ZHANG, Y. et al. Gaussian mixture model clustering with incomplete data. **ACM Trans. Multimedia Comput. Commun. Appl.**, Association for Computing Machinery, New York, NY, USA, v. 17, n. 1s, mar 2021. ISSN 1551-6857.

ZHAO, B.; WEN, X.; HAN, K. Learning semi-supervised gaussian mixture models for generalized category discovery. In: **2023 IEEE/CVF International Conference on Computer Vision (ICCV)**. Los Alamitos, CA, USA: IEEE Computer Society, 2023. p. 16577–16587.