

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Geradores de homologia persistente e aplicações

Carlos Henrique Venturi Ronchi

Dissertação de Mestrado do Programa de Pós-Graduação em
Matemática (PPG-Mat)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Carlos Henrique Venturi Ronchi

Geradores de homologia persistente e aplicações

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Matemática. *EXEMPLAR DE DEFESA*

Área de Concentração: Matemática

Orientador: Prof. Dr. Marcio Fuzeto Gameiro

USP – São Carlos
Junho de 2018

Carlos Henrique Venturi Ronchi

Persistent homology generators and applications

Dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP – in accordance with the requirements of the Mathematics Graduate Program, for the degree of Master in Science.
EXAMINATION BOARD PRESENTATION COPY

Concentration Area: Mathematics

Advisor: Prof. Dr. Marcio Fuzeto Gameiro

USP – São Carlos
June 2018

RESUMO

RONCHI, C. H. V. **Geradores de homologia persistente e aplicações**. 2018. 66 p. Dissertação (Mestrado em Ciências – Matemática) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

a.

Palavras-chave: Modelo, Monografia de qualificação, Dissertação, Tese, Latex.

ABSTRACT

RONCHI, C. H. V. **Persistent homology generators and applications**. 2018. [66](#) p.
Dissertação (Mestrado em Ciências – Matemática) – Instituto de Ciências Matemáticas e
de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

a.

Keywords: Template, Qualification monograph, Dissertation, Thesis, Latex.

LISTA DE ILUSTRAÇÕES

| | |
|---|----|
| Figura 1 – Representação do pipeline para a utilização da homologia persistente com um conjunto de dados. | 22 |
| Figura 2 – Exemplos de k -simplexos para $k \in \{0, 1, 2, 3\}$ | 23 |
| Figura 3 – Exemplo em que a interseção de dois simplexos não é um simplexo. . . | 24 |
| Figura 4 – Exemplo de filtração para um complexo simplicial K | 24 |
| Figura 5 – Exemplo de um complexo simplicial abstrato e sua realização geométrica | 26 |
| Figura 6 – Exemplo de um complexo de Čech para um raio r fixado. | 26 |
| Figura 7 – Exemplo do complexo de Vietoris-Rips com os mesmos pontos utiliza- dos para a construção na Figura 6. | 27 |
| Figura 8 – Diagrama de Voronoi de três pontos no plano. | 28 |
| Figura 9 – Complexo Alpha para um conjunto de pontos no plano. | 29 |
| Figura 10 – Exemplo da filtração de um complexo simplicial e o barcode e diagramas de persistência associados. | 31 |
| Figura 11 – 0- e 1- Diagramas de Persistência da Figura 4. | 33 |
| Figura 12 – Filtração de um complexo simplicial. | 34 |
| Figura 13 – 0- e 1- diagramas de persistência da filtração mostrada na Figura 12. | 35 |
| Figura 14 – Exemplo de um diagram de persistência de um módulo de persistência q -tame com um quadrante em destaque. | 39 |
| Figura 15 – Representação por intervalo (esquerda), pela função rank (meio) e pelo ponto decorado (direita) do módulo intervalar $\mathbf{k}[1, 3) = \mathbf{k}(1^-, 3^-)$ | 42 |
| Figura 16 – Pontos decorados que são detectados pela medida aplicada no retângulo R. | 48 |
| Figura 17 – Representação gráfica da Proposição 3.19 | 49 |
| Figura 18 – Demonstração da proposição através da figura. | 51 |
| Figura 19 – Pontos extraídos de um círculo com ruídos. | 57 |
| Figura 20 – Diagrams de persistência do círculo X . Em laranja o diagrama de persistência de dimensão 1, em azul o de dimensão 0. A filtração de Vietoris-Rips foi usada para calcular o complexo simplicial. | 57 |
| Figura 21 – Seis imagens de persistência do diagrama de dimensão 1 da Figura 20. | 58 |
| Figura 22 – Esquema de uma rede neural artificial. O número de vértices na camada escondida é determinado pelo tamanho da matriz A_i | 60 |

LISTA DE ALGORITMOS

| | |
|--|----|
| Algoritmo 1 – Redução da matriz bordo ∂ | 32 |
|--|----|

LISTA DE CÓDIGOS-FONTE

| | |
|---|----|
| algoritmos/std_alg.jl | 65 |
|---|----|

LISTA DE TABELAS

SUMÁRIO

| | | |
|-------|---|----|
| 1 | INTRODUÇÃO | 19 |
| 2 | HOMOLOGIA PERSISTENTE 101 | 21 |
| 2.1 | Filtrações | 22 |
| 2.1.1 | <i>Complexo de Čech</i> | 25 |
| 2.1.2 | <i>Complexo de Vietoris-Rips</i> | 27 |
| 2.1.3 | <i>Complexo Alpha</i> | 28 |
| 2.2 | A matriz de bordo ∂ | 30 |
| 2.3 | Redução da matriz ∂ | 32 |
| 2.4 | Calculando a homologia persistente | 33 |
| 3 | MÓDULOS DE PERSISTÊNCIA | 37 |
| 3.1 | Módulos de persistência e decomposições | 37 |
| 3.1.1 | <i>Índices e posets</i> | 39 |
| 3.1.2 | <i>Categoria de módulos</i> | 40 |
| 3.1.3 | <i>Módulos Intervalares</i> | 40 |
| 3.1.4 | <i>Decomposição em módulos intervalares</i> | 42 |
| 3.1.5 | <i>Cálculos com quivers</i> | 44 |
| 3.2 | Medidas retangulares | 47 |
| 3.2.1 | <i>A medida de persistência</i> | 47 |
| 3.2.2 | <i>r-medidas abstratas</i> | 50 |
| 3.2.3 | <i>Equivalência de medidas e diagramas</i> | 51 |
| 3.3 | Comportamento de módulos e exemplos | 52 |
| 3.4 | <i>Interleaving</i> | 52 |
| 3.5 | O teorema de isometria | 52 |
| 4 | GERADORES ÓTIMOS E OUTROS CONCEITOS | 53 |
| 4.1 | Geradores ótimos | 53 |
| 4.2 | Vetorização do diagrama de persistência | 53 |
| 4.2.1 | <i>Estabilidade da Imagem de Persistência</i> | 53 |
| 4.2.2 | <i>Exemplos de Imagens de Persistência</i> | 56 |
| 4.3 | Mapper | 58 |
| 5 | APLICAÇÕES | 59 |

| | | |
|-----------------------|--|----|
| 5.1 | Geradores ótimos em classificadores de imagens | 59 |
| 5.1.1 | <i>Redes Neurais Convolucionais (CNN)</i> | 59 |
| 5.2 | Imagens de persistência aplicadas a proteínas | 60 |
| 6 | CONCLUSÃO | 61 |
| REFERÊNCIAS | | 63 |
| APÊNDICE A | ALGORITMO <i>STANDARD</i> E FUNÇÕES AUXILIARES | 65 |

INTRODUÇÃO

HOMOLOGIA PERSISTENTE 101

A topologia sempre foi vista como uma área de abstração da matemática, sem espaço para aplicações. Ela é usada para o estudo de diversos espaços em sua forma abstrata, auxiliando matemáticos em diversas demonstrações de teoremas e dando uma base fundamental para grande parte da teoria matemática usada no dia a dia ([POINCARÉ, 1895](#)).

Certas propriedades dos espaços topológicos são estudadas através da topologia algébrica, dando algumas informações, como o número de componentes conexas por caminhos de um espaço e buracos. A princípio esta é uma área altamente abstrata da matemática, nos últimos anos esta visão foi mudando, com o desenvolvimento da Homologia Persistente e Análise Topológica de Dados.

Um conjunto de dados, geralmente um subconjunto finito de algum espaço métrico, pode ser estudado através da homologia persistente e assim obtemos informações topológicas do objeto em estudo.

O pipeline da análise topológica de dados pode ser dividido nos seguintes passos:

1. A entrada do algoritmo pode ser um conjunto de pontos ou alguma matriz de distância/similaridade do conjunto de dados.
2. A construção de um objeto combinatorial em cima do conjunto de dados ou da matriz de distância. Geralmente uma filtração ou um complexo simplicial.
3. A partir da filtração ou do complexo simplicial é possível extrair informações topológicas e geométricas do conjunto de dados, por exemplo o número de componentes conexas, como um algoritmo de Clustering.
4. Por fim a interpretação dos dados obtidos e possível pós processamento para a utilização em outros algoritmos, como os de classificação ou regressão.

Neste capítulo descrevemos de forma ingênua a homologia persistente, começando com filtrações, passando pelos espaços vetoriais associados aos complexos simpliciais e chegando ao algoritmo de homologia persistente. Mostraremos também como interpretar os resultados obtidos. A [Figura 1](#) mostra os passos para utilizar esta ferramenta em um conjunto de dados.

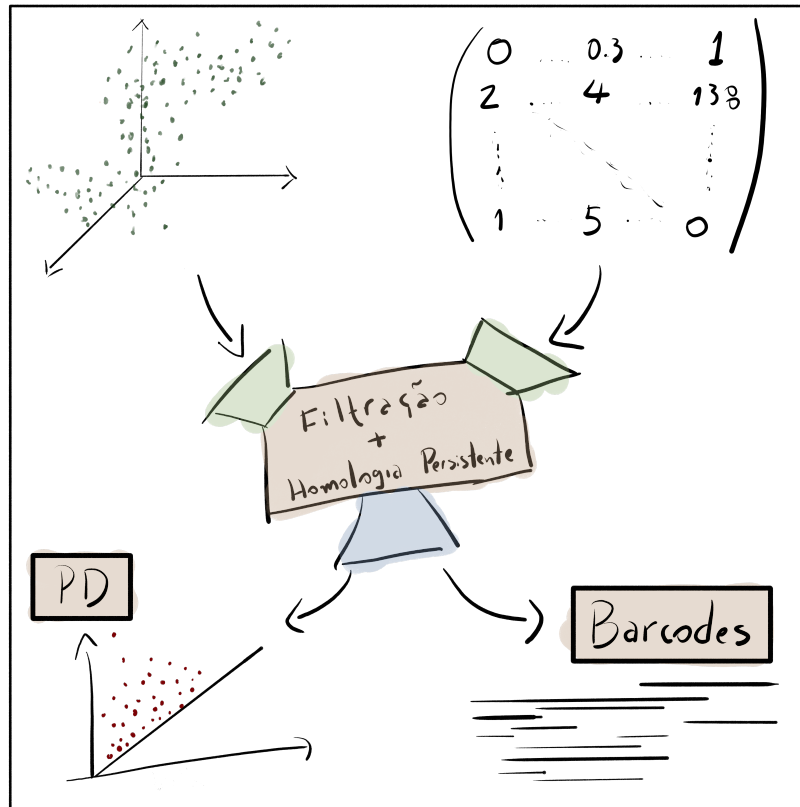


Figura 1 – Representação do pipeline para a utilização da homologia persistente com um conjunto de dados.

Fonte: Elaborada pelo autor.

2.1 Filtrações

A filtração de um conjunto de dados é o primeiro passo na nossa sequência apresentada na [Figura 1](#). Dado um conjunto de dados precisamos construir um objeto combinatorial de forma que possa ser analisado do ponto de vista da topologia assim como computacionalmente. A filtração é este objeto que captura as mudanças do conjunto dada uma escala.

Algumas definições se fazem necessárias para entendermos o que é a filtração e qual o seu papel na análise topológica de dados. Começamos definindo um simplexo, primeiro objeto combinatorial que é a base da filtração.

Definição 2.1. Sejam $v_0, v_1, \dots, v_k \in \mathbb{R}^n$ linearmente afins, ou seja $\{v_1 - v_0, \dots, v_k - v_0\}$ é um conjunto linearmente independente. O k -simplexo definido pelos pontos acima, chamados de vértices, é a envoltória convexa, definida na abaixo.

$$\left\{ \sum_{i=0}^k \lambda_i v_i \mid \sum_{i=0}^k \lambda_i = 1 \text{ e } \lambda_i \geq 0, \forall i \right\}. \quad (2.1)$$

Denotamos o k -simplexo por $\langle v_0, \dots, v_k \rangle$.

Note que para $k = 0$, temos um único vértice. Para $k = 1$, temos uma reta, já para $k = 2$ temos um triângulo preenchido. E no caso $k = 3$, um tetraedro. Os simplexos podem ser vistos na [Figura 2](#). Além disso, dizemos que a dimensão do k -simplexo é k . A envoltória convexa de qualquer subconjunto dos vértices de um simplexo σ é chamado de face de σ .

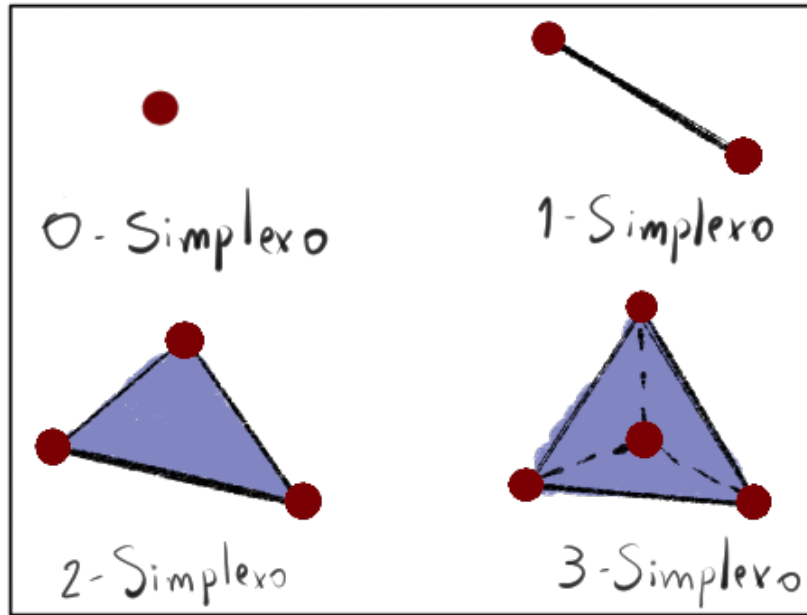


Figura 2 – Exemplos de k -simplexos para $k \in \{0, 1, 2, 3\}$.

Fonte: Elaborada pelo autor.

Tendo definido os k -simplexos, podemos definir o complexo simplicial.

Definição 2.2. Um complexo simplicial K é uma coleção de simplexos satisfazendo as seguintes relações:

- Dado $\sigma \in K$, temos que para toda face $\tau \subset \sigma$ vale $\tau \in K$.
- A interseção de dois simplexos é face de ambos os simplexos, em outras palavras, $\sigma, \tau \in K$ implica que $\sigma \cap \tau \subset \sigma$ e $\sigma \cap \tau \subset \tau$.

Nessa definição utilizamos o símbolo \subset para indicar que uma face. Usaremos esse símbolo com essa denotação quando falarmos sobre simplexos e faces. A segunda condição

é necessária para evitar casos patológicos como mostrado na Figura 3. Dizemos que a dimensão do complexo simplicial K é a maior dimensão dentre os simplexos em K . Podemos definir agora a filtração de um complexo simplicial.

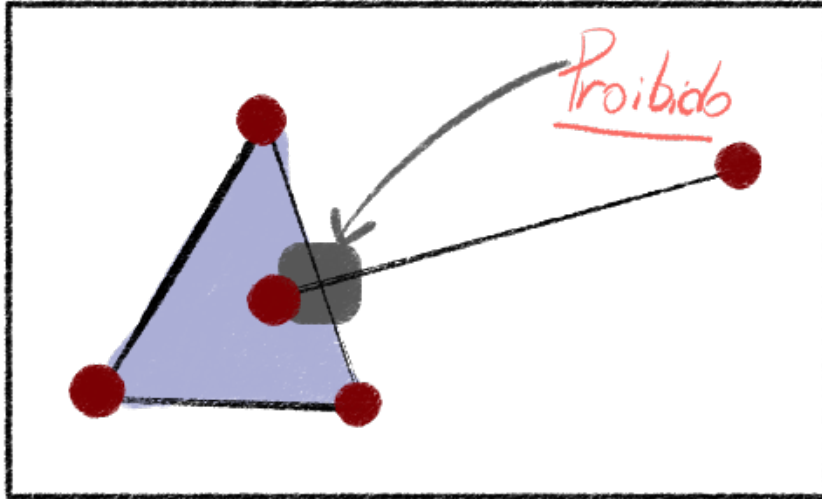


Figura 3 – Exemplo em que a interseção de dois simplexos não é um simplexo.

Fonte: Elaborada pelo autor.

Definição 2.3. Seja K um complexo simplicial. Definimos uma filtração de K sendo uma sequência de subconjuntos $K_i \subset K$, com $i \in \{1, \dots, n\}$, de tal forma que K_i é um complexo simplicial para todo i e vale que

$$K_1 \subset \dots \subset K_{n-1} \subset K_n = K.$$

Na Figura 4 temos um exemplo de filtração para um complexo simplicial.

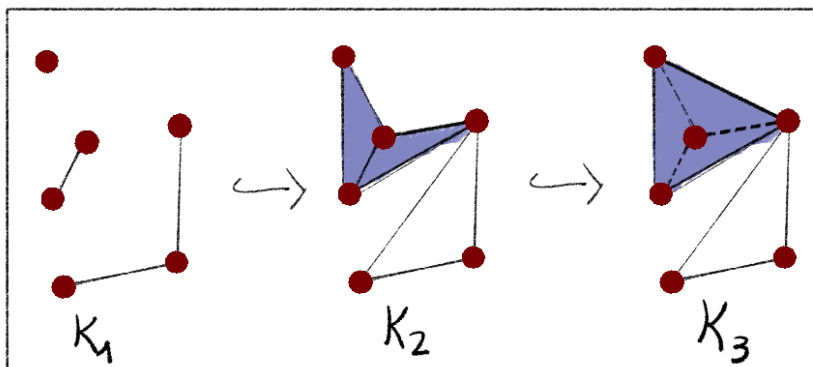


Figura 4 – Exemplo de filtração para um complexo simplicial K .

Fonte: Elaborada pelo autor.

2.1.1 Complexo de Čech

Para construir complexos simpliciais a partir dos dados, precisamos abstrair a noção de um simplexo simplicial. Na definição dada anteriormente, temos uma representação geométrica do que é um simplexo, mas podemos abstrair tal noção dando origem aos *complexos simpliciais abstratos*. As definições para os complexos definidos nesta seção e nas próximas foram retiradas de (EDELSBRUNNER, 2010).

Definição 2.4. Seja X um conjunto finito com pontos quaisquer. Seja F um conjunto de subconjuntos não-vazios de X . Dizemos que F é um complexo simplicial abstrato de X se a seguinte condição é satisfeita.

- Se para todo $\sigma \in F$, temos que para todo subconjunto $\sigma' \subset \sigma$ está em F também.

Cada elemento $\sigma \in F$ é chamado de simplexo. Denotamos um k -simplexo σ por $\langle x_{i_0}, \dots, x_{i_k} \rangle$, onde x_{i_j} são elementos de X .

Exemplo 2.5. Seja $X = \{a, b, c\}$ e considere $F = \{\{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}\}$. Precisamos mostrar que F é um complexo simplicial abstrato. Seja $\sigma = \{a, c\}$. Note que seus subconjuntos são $\{a\}$ e $\{c\}$, além disso ambos pertencem a F . De forma análoga, mostramos que para qualquer outro simplexo, suas faces (subconjuntos) estão em F .

Podemos realizar os complexos simpliciais abstratos geometricamente, ou seja, apesar de trabalharmos com conjuntos de elementos quaisquer, podemos incluir esses complexos em algum \mathbb{R}^n e assim visualiza-los. Para obtermos o complexo simplicial *geométrico*, associamos a cada simplexo abstrato σ um simplexo geométrico. Por exemplo, se adotarmos o complexo simplicial abstrato F acima mostrado, teríamos que sua realização geométrica seria um triângulo sem preenchimento, como é mostrado na Figura 5.

Observe que se o nosso conjunto X for um subconjunto finito de \mathbb{R}^d , podemos ter simplexos de dimensão maiores do que d , ou seja, não podem ser realizados (ou visualizados) em \mathbb{R}^d necessariamente. Um exemplo dessa situação pode ser visto no complexo simplicial final da Figura 4, considerando que os pontos vermelhos são a realização geométrica dos pontos de X , onde X é um subconjunto do \mathbb{R}^2 .

Essa é uma grande diferença entre os complexos simpliciais geométricos e abstratos. Uma vez tendo definido os complexos simpliciais abstratos, podemos definir o *complexo de Čech*.

Definição 2.6. Seja X um conjunto de pontos $\{x_1, \dots, x_n\}$ em \mathbb{R}^d . O complexo de Čech de X para um valor real $r > 0$ é o conjunto $C^r(X)$, onde $\sigma = \langle x_{i_1}, \dots, x_{i_k} \rangle \in C^r(X)$ se, e somente se vale a seguinte condição

$$\bigcap_{j=1}^k B(x_{i_j}, r) \neq \emptyset.$$

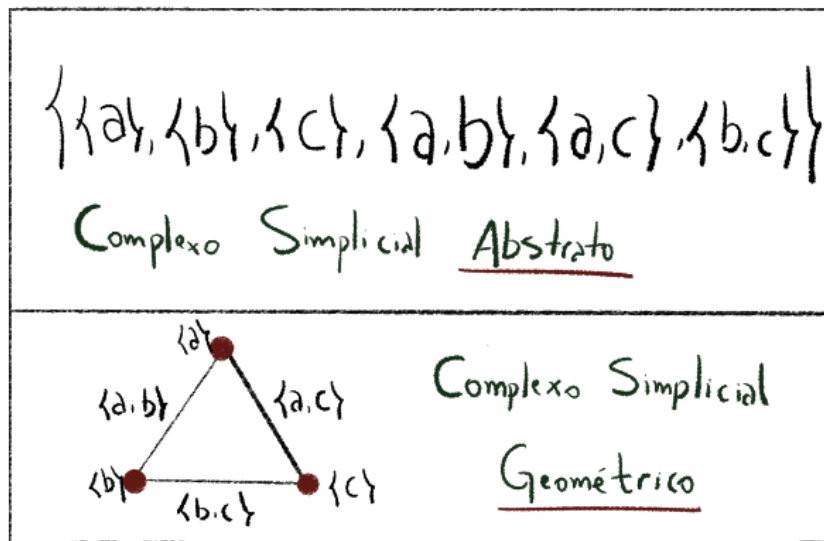


Figura 5 – Exemplo de um complexo simplicial abstrato e sua realização geométrica

Fonte: Elaborada pelo autor.

A definição acima nos diz que quando temos k pontos cujas bolas de raio r centradas neles se intersectam, adicionamos um k simplexo no complexo simplicial abstrato, o que seria apenas o conjunto desses pontos. Geometricamente falando, se duas bolas se intersectam, adicionamos uma aresta. Se três bolas se intersectam, adicionamos um triângulo preenchido, e assim por diante. Na [Figura 6](#) temos um exemplo do complexo simplicial de Čech.

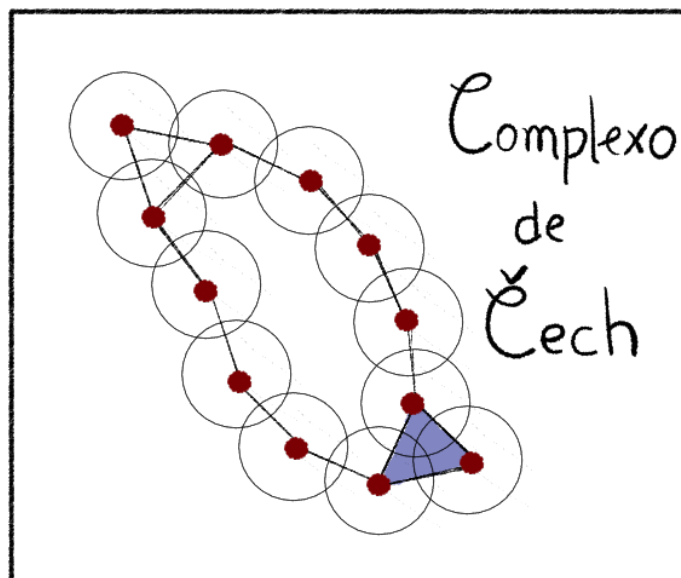


Figura 6 – Exemplo de um complexo de Čech para um raio r fixado.

Fonte: Elaborada pelo autor.

Da mesma forma que definimos a filtração para um complexo simplicial geométrico, o mesmo vale para o caso abstrato.

2.1.2 Complexo de Vietoris-Rips

O complexo de Vietoris-Rips possui uma construção similar ao complexo de Čech, porém computacionalmente é um método mais barato, já que analisa apenas distância entre pontos dois a dois.

Definição 2.7. Seja X um conjunto de pontos $\{x_1, \dots, x_n\}$ em \mathbb{R}^d . O complexo de Vietoris-Rips de X para um valor real $r > 0$ é o conjunto $V^r(X)$, onde o simplexo $\sigma = \langle x_{i_1}, \dots, x_{i_k} \rangle \in V^r(X)$ se, e somente se vale a seguinte condição

$$d(x_{i_k}, x_{i_j}) < r \quad \forall j, l \in 1, \dots, k.$$

A [Figura 7](#) é um exemplo do complexo de Vietoris-Rips. Uma das diferenças que a construção dos dois complexos já definidos nos dá é que no caso do complexo de Čech temos triângulos preenchidos, e isso não ocorre para Vietoris-Rips.

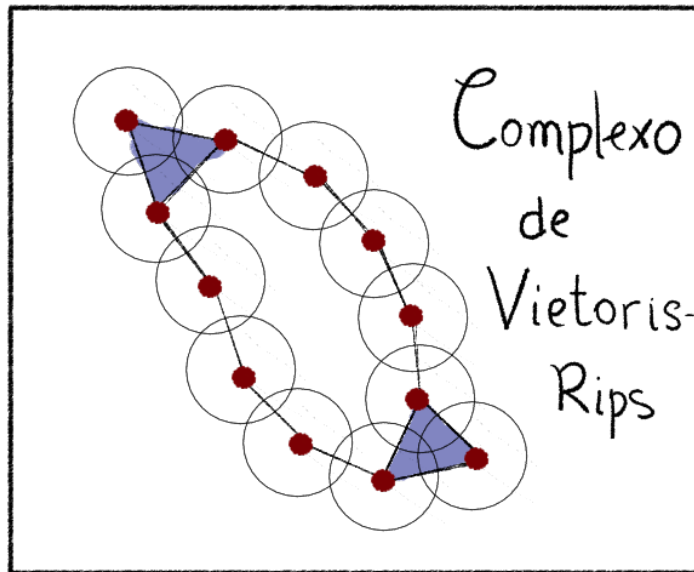


Figura 7 – Exemplo do complexo de Vietoris-Rips com os mesmos pontos utilizados para a construção na [Figura 6](#).

Fonte: Elaborada pelo autor.

Mesmo com as regras diferentes para a construção de complexos, temos a seguinte relação entre os dois complexos.

$$C^r(X) \subset V^r(X) \subset C^{2r}(X) \quad (2.2)$$

A primeira inclusão segue do fato que se k bolas se intersectam então elas se intersectam dois a dois com a mesma distância. A segunda inclusão segue da desigualdade triangular da métrica sendo usada e o fato que as bolas se intersectam duas a duas.

2.1.3 Complexo Alpha

E como uma terceira opção para a construção de um complexo simplicial através de pontos no \mathbb{R}^n , temos o complexo Alpha. A construção é similar ao complexo de Čech, porém os conjuntos centrados nos pontos são uma interseção de bolas no \mathbb{R}^n com conjuntos convexos especiais, as células de Voronoi. Nesta subseção utilizaremos o \mathbb{R}^n para as definições, porém elas podem ser generalizadas para qualquer espaço métrico.

O diagrama de Voronoi é um tipo especial de decomposição de um espaço métrico, um conjunto que possui uma distância associada a ele. Dado um subconjunto $X \subset \mathbb{R}^n$ finito, onde $X = \{x_1, \dots, x_k\}$, definimos a célula de Voronoi associada ao ponto x_i sendo o seguinte conjunto

$$V_i = \{x \in \mathbb{R}^n \mid d(x_i, x) \leq d(x_j, x), \forall j \in 1, \dots, k\},$$

em que d é a distância euclidiana usual. A Figura 8 mostra um exemplo de diagram de Voronoi para três pontos no \mathbb{R}^2 . Podemos agora definir o complexo simplicial Alpha.

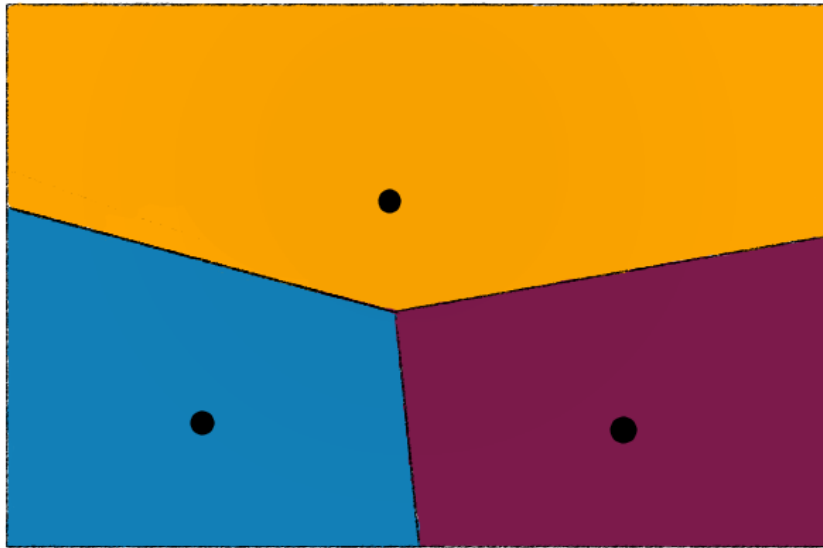


Figura 8 – Diagrama de Voronoi de três pontos no plano.

Fonte: Elaborada pelo autor.

Definição 2.8. Seja X um conjunto de pontos $\{x_1, \dots, x_n\}$ em \mathbb{R}^d . O complexo Alpha de X para um valor real $r > 0$ é o conjunto $A^r(X)$, onde o simplexo $\sigma = \langle x_{i_1}, \dots, x_{i_k} \rangle \in A^r(X)$ se, e somente se vale a seguinte condição

$$\bigcap_{j=1}^k R(x_{i_j}, r) \neq \emptyset,$$

onde $R(x_{ij}, r) = B(x_{ij}, r) \cap V_{ij}$, para todo $j \in \{1, \dots, k\}$.

Na Figura 9 temos o exemplo de um complexo Alpha. É interessante notar que o Alpha é um subcomplexo do complexo de Čech, ou seja, para $r > 0$, $A^r(X) \subset C^r(X)$. Além disso esse complexo herda uma propriedade importante dos diagramas de Voronoi, a realização geométrica no espaço em que os pontos se encontram, isto é, se os pontos em \mathbb{R}^d satisfazem a condição de posição geral, então o complexo simplicial abstrato Alpha pode ser realizado geometricamente no \mathbb{R}^d , ou seja o complexo simplicial geométrico pode ser construído no \mathbb{R}^d ! Isso é fundamental computacionalmente, já que diminui a complexidade dos cálculos e aumenta a velocidade para obtenção do complexo.

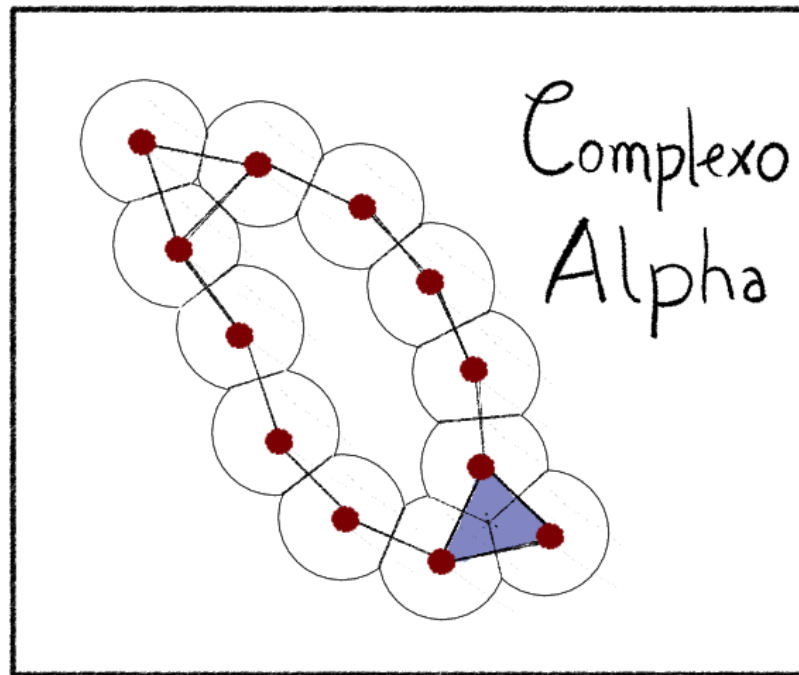


Figura 9 – Complexo Alpha para um conjunto de pontos no plano.

Fonte: Elaborada pelo autor.

Uma variação muito importante do complexo Alpha é a versão com peso. Ao invés de considerar um raio fixo para cada bola ao redor de um ponto, podemos dar um *peso* para cada ponto. Seja $X = \{x_1, \dots, x_n\}$ o nosso conjunto de pontos finitos e $\{w_1, \dots, w_n\}$ conjunto de valores maiores ou iguais a zero, que serão os pesos associados a cada ponto. Para cada x_i , ao invés de associar a bola usual do complexo Alpha, associamos a seguinte bola.

$$R_{w_i}(x_i, r) = B(x_i, r + w_i^2) \cap V_i$$

Esse é um complexo muito usado em aplicações biomoleculares, em que o conjunto de pontos são átomos de uma molécula e o peso para cada átomo é o seu respectivo raio de Van der Waals.

2.2 A matriz de bordo ∂

Agora vamos para o terceiro passo descrito na lista anteriormente. Uma vez com os dados, podemos construir uma filtração de um complexo simplicial criado a partir deles que irá capturar diversas informações, como os buracos que um conjunto de dados tem e o quanto eles persistem na nossa filtração.

A ferramenta matemática utilizada para extrair essas informações da filtração são os grupos de homologia. Para uma filtração $K_1 \subset \dots \subset K_m = K$ e um p fixo, a p -ésima homologia persistente de K é o par

$$(\{H_p(K_i)\}_{1 \leq i \leq m}, \{f_{i,j}\}_{1 \leq i \leq j \leq m}), \quad (2.3)$$

em que para todo $i, j \in \{1, \dots, m\}$, $f_{i,j}$ são aplicações lineares entre os espaços vetoriais $H_p(K_i)$ e $H_p(K_j)$. Mais especificamente, os espaços vetoriais $H_p(K_i)$ são grupos de homologia com coeficientes em um espaço vetorial. No nosso caso usamos o espaço vetorial \mathbb{Z}_2 . Consulte (EDELSBRUNNER, 2010) para uma introdução à teoria de homologia nesse contexto.

A homologia persistente dá informações topológicas sobre a filtração do complexo simplicial. Os elementos das bases de cada $H_p(K_i)$ correspondem a ciclos p -dimensionais, podendo ser buracos. Ciclos são os nomes dados aos representantes dos elementos da base do espaço vetorial em questão. No caso $p = 0$, temos que cada elemento da base corresponde à uma componente conexa, $p = 1$ cada elemento corresponde a um buraco. Portanto, considere os elementos da base de $H_p(K_i)$. Para cada um deles, desenhe um ponto. Se $f_{i,i+1}(u) = 0$, então desenhe um intervalo que termina em $i + 1$. Se $f_{i,i+1} = v$, onde v é um elemento da base de $H_p(K_{i+1})$, então desenhe uma reta que liga u ao ponto que representa v no próximo passo da filtração. Dessa forma vamos anotando os ciclos, que são os elementos da base, ao longo da filtração. Na Figura 10 temos um exemplo para uma filtração.

Podemos falar também que $u \in H_p(K_i)$ nasceu no tempo i da filtração se u não é imagem de nenhum elemento de $H_p(K_{i-1})$ sobre $f_{i-1,i}$. Dizemos também que $u \in H_p(K_j)$ morreu em j se j é o menor índice tal que $f_{i,j}(u) = 0$, onde $j > i$. A persistência do ponto u pode ser representada pelo intervalo $[i, j)$. Além disso, se u nasce no tempo i e nunca morre, denotamos o intervalo associado à essa informação como $[i, +\infty)$.

Existem duas formas de visualizar esses intervalos, através dos *barcodes* ou dos *diagramas de persistência (PD)*. No barcode desenhamos uma barra do comprimento do intervalo $[i, j)$. Já no diagrama de persistência representamos com um ponto (i, j) no plano. A Figura 10 possui o barcode e o diagrama de persistência para o conjunto de dados associado.

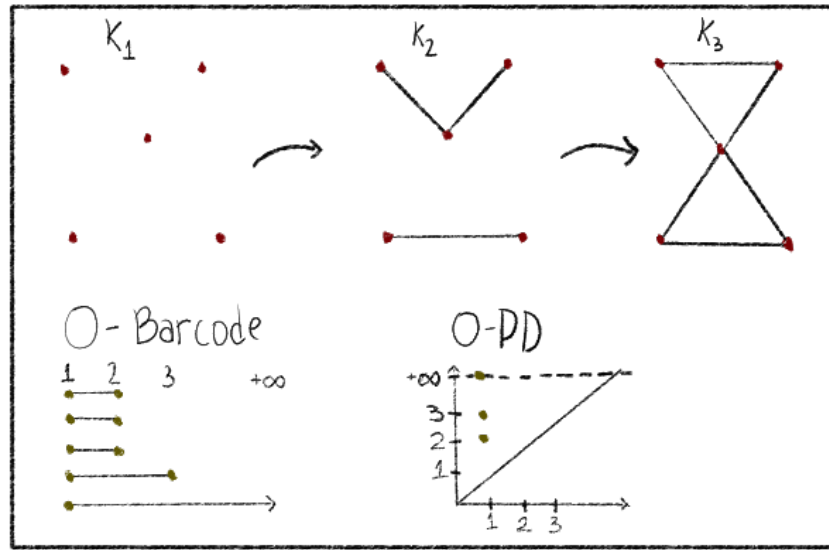


Figura 10 – Exemplo da filtração de um complexo simplicial e o barcode e diagramas de persistência associados.

Fonte: Elaborada pelo autor.

Tendo essa ferramenta, como podemos traduzi-la para o contexto dos dados, e como calcular os pares dos diagramas de persistência? Abaixo segue uma lista dos primeiros passos que devem ser feitos para a obtenção do diagrama de persistência.

1. Dado um conjunto de dados, determinar alguma filtração;
2. Listar todos os simplexes na filtração;
3. Ordenar os simplexes satisfazendo duas regras:
 - a) A face um simplexo o precede na ordenação;
 - b) Um simplexo no complexo K_i precede os simplexes em K_j , $j > i$, que não pertençam a K_i ;
4. Construir a matriz de bordo.

A matriz de bordo é quem vai armazenar as informações topológicas importantes das quais iremos extrair mais tarde.

Definição 2.9. Seja K um complexo simplicial, $K_1 \subset \dots \subset K_m$ uma filtração e $\sigma_1, \dots, \sigma_n$ uma ordenação dos simplexes de K satisfazendo as regras acima mencionadas. A *matriz de bordo* de K , denotada por ∂ , é uma matriz de tamanho $n \times n$, em que cada entrada tem o seguinte valor

$$\delta(i, j) = \begin{cases} 1, & \text{se o simplexo } \sigma_i \text{ é face de } \sigma_j \text{ e } \dim(\sigma_j) = \dim(\sigma_i) + 1 \\ 0, & \text{caso contrário.} \end{cases}$$

Com a matriz construída, podemos utilizar um método de eliminação de Gauss para a redução da matriz.

2.3 Redução da matriz ∂

O algoritmo que será descrito aqui é conhecido como algoritmo *standard* para a redução da matriz ∂ (EDELSBRUNNER; LETSCHER; ZOMORODIAN, 2000). Estamos trabalhando sobre \mathbb{Z}_2 , ou seja, $1 + 1 = 0$. Durante o processo de redução da matriz será apenas necessário somar colunas.

Dado $j \in \{1, \dots, n\}$, denotamos por $low(j)$ o maior inteiro $i \in \{1, \dots, n\}$ tal que $\delta(i, j) = 1$. Note que $i < j$, pois segundo as regras de construção da matriz de bordo, temos que $\delta(i, j) = 1$ só quando σ_i é face de codimensão 1 de σ_j . Assim temos o Algoritmo 1 para reduzir a matriz de bordo.

Algoritmo 1 – Redução da matriz bordo ∂ .

- 1: Dados os simplexes $\sigma_1, \dots, \sigma_n$ e a matriz de bordo ∂ correspondente.
 - 2: **para** $j \in \{1, \dots, n\}$ **faça**
 - 3: **enquanto** existe i tal que $low(i) = low(j)$ **faça**
 - 4: Some a coluna i a coluna j .
 - 5: **fim enquanto**
 - 6: **fim para**
-

Dizemos que a matriz está reduzida quando $low(j) \neq low(j_0)$ para quaisquer colunas j, j_0 não nulas. Observe que uma coluna j pode ser zerada, dizemos então que $low(j)$ é indefinido. Além disso, a matriz reduzida, denotada por R , é escrita como a multiplicação de duas matrizes.

$$R = \partial \cdot V \tag{2.4}$$

A matriz V é uma matriz triangular superior que acumula a informação dos ciclos. Uma vez com a matriz ∂ reduzida a R , podemos interpretar as colunas de R da seguinte forma.

1. A coluna j é nula. Dizemos que o simplexo σ_j é *positivo*, pois dá vida a um ciclo.
2. A coluna j é não-nula. Seja i tal que $low(j) = i$. Dizemos então que σ_j é um simplexo *negativo*, pois quando ele é adicionado temos a morte de um ciclo. Ainda mais, esse ciclo nasceu com a adição do simplexo σ_i .

A nomenclatura de simplexes positivos e negativos vêm da teoria clássica de homologia que estuda propriedades homológicas de um complexo simplicial ao invés de toda a filtração. Para mais detalhes, consulte (EDELSBRUNNER, 2010).

Agora podemos construir o diagrama de persistência para a filtração dada utilizando a matriz de redução. Denotamos por Dgm_p o p -ésimo diagrama de persistência, com $p \in \{0, \dots, k-1\}$ onde k é a maior dimensão dentre os simplexes σ_i . Cada p representa a dimensão dos grupos de homologia descritos anteriormente. Se $p = 0$, então o diagrama de persistência nos dirá quais componentes conexas apareceram na filtração e o quão persistente elas são. Para $p = 1$, teremos buracos 1-dimensionais, por exemplo, um círculo vazado tem um buraco. Já um toro, tem dois buracos, um visível e outro por dentro da superfície.

Seja p fixado e σ_j um simplexo de dimensão $p+1$ tal que $low(j) = i$. Dessa forma, adicionamos o ponto (a_i, a_j) ao multiconjunto Dgm_p , em que a_i e a_j são os menores índices tais que $\sigma_i \in K_{a_i}$ e $\sigma_j \in K_{a_j}$, por exemplo, se σ_i é adicionado na filtração em K_l e σ_j é adicionado em K_q , então $a_i = l$ e $a_j = q$. Se tivermos um simplexo σ_i de dimensão p tal que $low(i)$ é indefinido, então adicionamos o ponto $(a_i, +\infty)$ à Dgm_p . Observe na Figura 11 os diagramas de persistência de dimensão 0 e 1 da respectiva filtração.

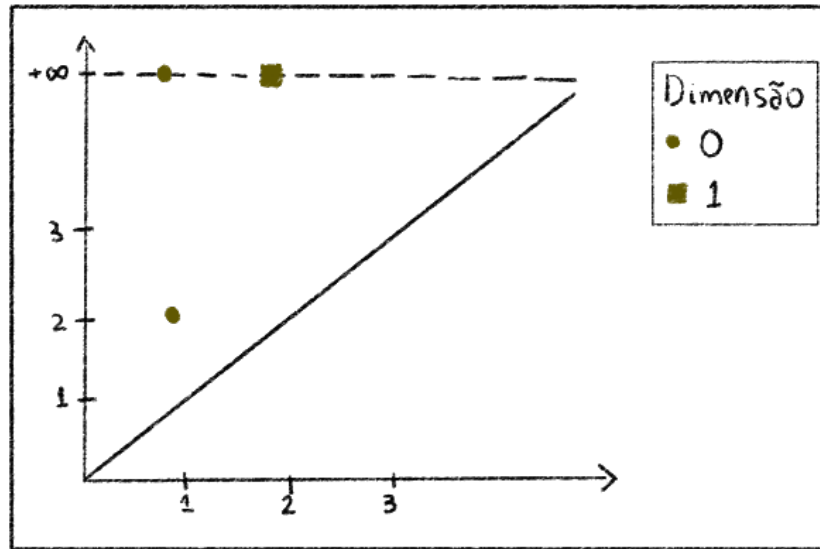


Figura 11 – 0- e 1- Diagramas de Persistência da Figura 4.

Fonte: Elaborada pelo autor.

2.4 Calculando a homologia persistente

Nesta seção iremos calcular a homologia persistente de uma filtração já dada, além disso apresentaremos uma implementação para redução da matriz ∂ em Julia.

Considere a filtração da Figura 12. Observe que temos 4 vértices ($\sigma_1, \dots, \sigma_4$), 5 arestas ($\sigma_5, \dots, \sigma_9$) e 1 triângulo (σ_{10}) ao total, temos ao total 10 simplexes. Diretamente da figura já podemos extrair os diagrams de persistência de dimensão 0 e 1. Note que no

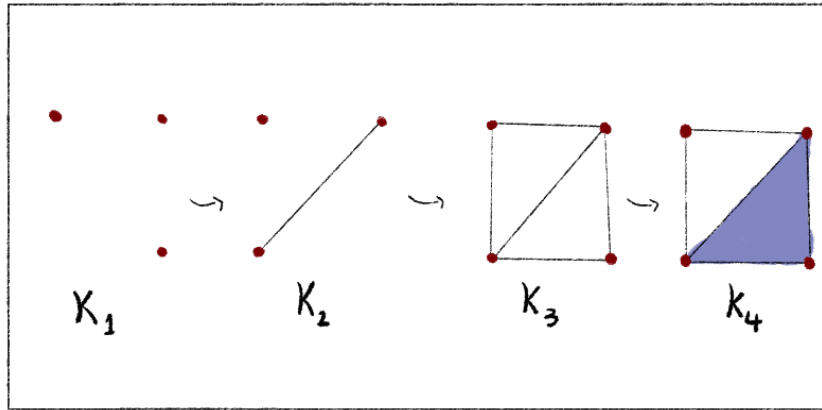


Figura 12 – Filtração de um complexo simplicial.

Fonte: Elaborada pelo autor.

primeiro passo da filtração temos 2 componentes conexas, sendo que uma delas morre no terceiro passo e a outra sobrevive até o final. Temos portanto dois intervalos e logo dois pontos no 0-diagrama de persistência: $[1, +\infty)$ e $[1, 3)$.

Quando temos um intervalo infinito, geralmente se representa o acima dos índices da filtração no momento em que ele nasceu, como pode ser visto na [Figura 13](#). Já para $p = 1$, constatamos dois intervalos, que representam os dois buracos unidimensionais que surgiram. O primeiro buraco é o que aparece no passo 2 da filtração com a introdução dos simplexos σ_3 e σ_6 e não morre, ou seja, se mantém até o final da filtração, enquanto o segundo buraco surge no passo 3 da filtração com a introdução dos simplexos σ_8 e σ_9 e morre no passo 4 com o nascimento do triângulo σ_{10} . Logo, nossos intervalos são $[2, +\infty)$ e $[3, 4)$.

Sendo assim, podemos construir os dois diagramas de persistência, que podem ser vistos na [Figura 13](#).

Agora que calculamos intuitivamente os diagramas de persistência, vamos construir a matriz de bordo ∂ da filtração mostrada na [Figura 12](#) e utilizar implementações no *Julia* para verificar os resultados. A matriz de bordo pode ser visualizada abaixo, note

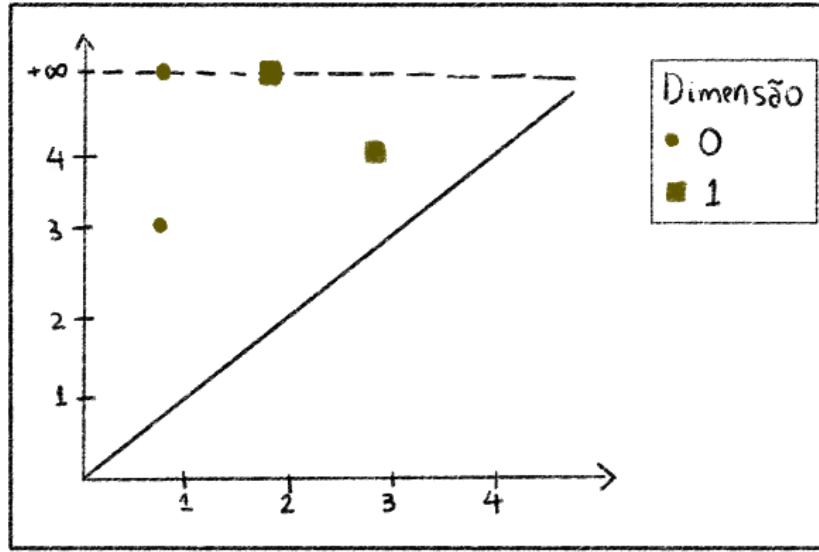


Figura 13 – 0- e 1- diagramas de persistência da filtração mostrada na Figura 12.

Fonte: Elaborada pelo autor.

que as regras para construção da matriz são satisfeitas.

$$\partial = \begin{matrix} & \begin{matrix} \sigma_1 & \sigma_2 & \sigma_3 & \sigma_4 & \sigma_5 & \sigma_6 & \sigma_7 & \sigma_8 & \sigma_9 & \sigma_{10} \end{matrix} \\ \begin{matrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \sigma_4 \\ \sigma_5 \\ \sigma_6 \\ \sigma_7 \\ \sigma_8 \\ \sigma_9 \\ \sigma_{10} \end{matrix} & \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{matrix} \quad (2.5)$$

Para reduzir vamos realizar as seguintes operações:

1. somar a coluna 7 com a coluna 6,
2. somar a coluna 7 com a coluna 5, assim zerando a coluna 7,
3. somar a coluna 9 com a coluna 6,
4. somar a coluna 9 com a coluna 8,
5. somar a coluna 9 com a coluna 5, assim zerando a coluna 9.

Note então que após esses passos a matriz ∂ está reduzida e com a seguinte forma.

$$R = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (2.6)$$

Vamos interpretar a matriz agora, para isso temos que parear os simplexos. Utilizando as regras de pareamento descritas anteriormente, temos os pares:

- σ_5 com σ_2 ,
- σ_6 com σ_4 ,
- σ_8 com σ_3 ,
- σ_{10} com σ_9 .

Além disso, existem simplexos que não foram pareados, como os simplexos σ_1 e σ_7 . Note que eles representam o nascimento de p -ciclos que não morrem ao longo da filtração, ou seja, σ_1 corresponde à componente conexa que nasce no primeiro passo da filtração e não morre e σ_7 corresponde ao buraco que nasce no segundo passo e não morre até o final da filtração. Portanto, temos os seguintes intervalos:

- $[a_2, a_5) = [1, 1)$, que não seria adicionado ao diagrama,
- $[a_4, a_6) = [2, 2)$, que não seria adicionado ao diagrama,
- $[a_3, a_8) = [1, 3)$,
- $[a_9, a_{10}) = [3, 4)$,
- $[a_1, +\infty) = [1, +\infty)$,
- $[a_7, +\infty) = [2, +\infty)$,

Logo, $Dgm_0 = \{(1, 3), (1, +\infty)\} \cup \Delta$ e $Dgm_1 = \{(3, 4), (2, +\infty)\} \cup \Delta$, onde $\Delta = \{(x, x) \mid x \in \mathbb{R}^+\}$. Note que obtemos o mesmo resultado, como esperado! O algoritmo *standard* está implementado e pode ser visto no [Apêndice A](#).

MÓDULOS DE PERSISTÊNCIA

A homologia persistence teve seu início em uma intersecção entre as ciências da computação e a matemática. Os primeiros artigos mostravam algoritmos sobre espaços topológicos simples, como esferas (EDELSBRUNNER; LETSCHER; ZOMORODIAN, 2000). No entanto, a teoria foi se desenvolvendo ao longo dos anos ao ponto em que as linguagens utilizadas para tratar da homologia persistente é a teoria de categorias conjuntamente com a teoria de representações (CHAZAL *et al.*, 2016).

Neste capítulo tratamos do desenvolvimento da homologia persistente sob a luz dessas linguagens. Na primeira seção definimos o que são os módulos de persistência e suas relações com os diagramas de persistência. Na segunda seção descrevemos a medida retangular, usada para abstrair o conceito de diagrama de persistência e poder estudar o quão *tame* ele o é. Apresentamos na terceira seção alguns exemplos do comportamento dos módulos de persistência e exemplos. A quarta seção é fundamental, pois mostramos como comparar dois módulos de persistência, através do *interleaving*. E finalmente, apresentamos o teoria de isometria e mostramos uma das implicações com a teoria desenvolvido neste capítulo.

3.1 Módulos de persistência e decomposições

Nesta seção iremos definir os módulos de persistência, apresentar teoremas de decomposição dos módulos e introduzir a notação de quiver, que será utilizada para as próximas seções e demonstrações de outros resultados.

Fixaremos aqui o corpo \mathbf{k} para todos os espaços vetoriais apresentados neste texto.

Definição 3.1. Um módulo de persistência \mathfrak{M} sobre os números reais \mathbb{R} é uma família

indexada sobre \mathbb{R} de espaços vetoriais

$$(V_t \mid t \in \mathbb{R}),$$

e uma família de aplicações lineares duplamente indexadas

$$(v_t^s: V_s \rightarrow V_t \mid s \leq t)$$

que satisfazem a seguinte relação de composição

$$v_t^s \circ v_s^r = v_t^r,$$

em que a função v_r^r é considerada a função identidade.

O módulo de persistência pode ser visto como um funtor entre a categoria dos números reais com o morfismo $s \rightarrow t$, em que $s \leq t$ e a categoria de espaços vetoriais.

Vamos dar um exemplo de módulo de persistência que se encontra no contexto de análise topológica de dados. Seja X um espaço vetorial e $f: X \rightarrow \mathbb{R}$ uma função, não necessariamente contínua e considere os conjuntos de nível

$$X^t = (X, f)^t = \{x \in X \mid f(x) \leq t\}.$$

Temos uma sequência de conjuntos encaixados, X^t com $t \in \mathbb{R}$, ou seja, existe uma função inclusão $\iota_t^s: X^s \hookrightarrow X^t$ que satisfaz trivialmente a lei de composição e existe uma função identidade. Chamamos esta sequência de conjuntos e funções de filtração de subníveis de (X, f) , denotada por \mathfrak{X}_{sub} ou \mathfrak{X}_{sub}^f .

Dada a sequência acima, podemos transforma-la em um módulo de persistência utilizando qualquer funtor da categoria de espaços topológicos para a categoria de espaços vetoriais. Neste caso utilizamos o funtor de homologia $H = H_k(-, \mathbf{k})$ de dimensão k com coeficientes em \mathbf{k} . Assim, podemos definir o seguinte módulo de persistência \mathfrak{V}

$$V_t = H(X^t) \quad v_t^s = H(\iota_t^s): V_s \rightarrow V_t.$$

Podemos também escrever $\mathfrak{V} = H(\mathfrak{X}_{sub})$.

Um exemplo na análise topológica de dados é quando X é um complexo simplicial finito e X^t é um subcomplexo. Devido as propriedades dos complexos, existem finitos valores críticos onde há mudanças em X . Suponha que os valores sejam $a_1 < \dots < a_n$. Então toda a informação do módulo de persistência é dada pela seguinte sequência de espaços vetoriais de dimensão finita

$$H(X^{a_1}) \rightarrow \dots \rightarrow H(X^{a_n}).$$

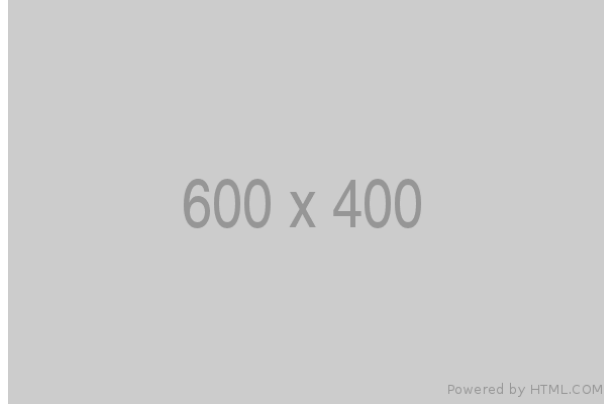


Figura 14 – Exemplo de um diagram de persistência de um módulo de persistência q -tame com um quadrante em destaque.

Neste caso, $H(\mathfrak{X}_{sub})$ admite uma descrição compacta, existe um algoritmo eficiente para o seu calculo e por último, a descrição é contínua com relação a f , ou seja, é estável sob uma métrica.

A descrição mencionada acima é o diagrama de persistência ou barcode. A estrutura é dada por uma lista de intervalos da forma $[b, d) = [a_i, a_j)$ ou $[a_i, +\infty)$. Cada intervalo representa um ciclo, uma propriedade, que nasce em b e morre em d .

Iremos mostrar aqui que é possível associar um diagram de persistência para módulos de persistência \mathfrak{V} q -tame. Um módulo de persistência é q -tame se

$$r_t^s = \text{rank}(v_t^s) < \infty \text{ para } s < t.$$

Intuitivamente falando, um módulo é q -tame se para todo quadrante que pegamos com a origem na diagonal, existem finitos pontos do diagram de persistência neste quadrante como pode ser visto na [Figura 14](#).

3.1.1 Índices e posets

No início desta seção definimos o módulo de persistência com o conjunto de índices sendo os reais. No entanto, é possível definir utilizando quaisquer conjuntos parcialmente ordenados da mesma forma que com os reais. Seja \mathbf{T} um poset, a coleção de espaços vetoriais e aplicações lineares que satisfazem as leis de composição e identidade é chamada de \mathbf{T} -módulo de persistência, ou módulo de persistência sobre \mathbf{T} .

Além disso, podemos restringir o poset \mathbf{T} para um subconjunto $\mathbf{S} \subset \mathbf{T}$ de forma a obter o \mathbf{S} -módulo de persistência, que são os espaços vetoriais e aplicações lineares cujos índices são elementos de \mathbf{S} . Esta é a restrição de \mathfrak{V} em \mathbf{S} e pode ser denotada por $\mathfrak{V}_{\mathbf{S}}$ ou $\mathfrak{V}|_{\mathbf{S}}$.

3.1.2 Categoria de módulos

Com a definição de módulos de persistência sobre um poset \mathbf{T} qualquer, podemos definir homomorfismos entre módulos. Sejam $\mathfrak{U}, \mathfrak{V}$ \mathbf{T} -módulos de persistência. Um homomorfismo Φ entre \mathfrak{U} e \mathfrak{V} é uma família de aplicações lineares $(\phi_t: U_t \rightarrow V_t \mid t \in \mathbf{T})$ tal que o seguinte diagrama comuta para todo $s \leq t$.

$$\begin{array}{ccc} U_s & \xrightarrow{u_t^s} & U_t \\ \phi_s \downarrow & & \downarrow \phi_t \\ V_s & \xrightarrow{v_t^s} & V_t \end{array}$$

A composição de dois homomorfismos Φ, Ψ é dada por cada índice $t \in \mathbf{T}$, ou seja, $\Phi \circ \Psi$ é a coleção de aplicações lineares $(\phi_t \circ \psi_t: U_t \rightarrow W_t \mid t \in \mathbf{T})$, onde Φ é homomorfismo entre \mathfrak{U} e \mathfrak{V} e Ψ entre \mathfrak{V} e \mathfrak{W} . A identidade é definida de forma trivial. Portanto, temos a categoria dos módulos. Definamos os seguintes conjuntos

$$\begin{aligned} \text{Hom}(\mathfrak{U}, \mathfrak{V}) &= \{ \text{homomorfismos } \mathfrak{U} \rightarrow \mathfrak{V} \}, \\ \text{End}(\mathfrak{V}) &= \{ \text{homomorfismos } \mathfrak{V} \rightarrow \mathfrak{V} \}. \end{aligned}$$

3.1.3 Módulos Intervalares

A relação entre os diagramas de persistência e módulos de persistência são fundamentadas pelos módulos intervalares. Eles são a base da teoria de homologia persistente.

Um intervalo em um conjunto totalmente ordenado \mathbf{T} é um subconjunto $J \subset \mathbf{T}$ tal que se $r \in J$ e $t \in J$ tal que $r < s < t$, então $s \in J$. Portanto, para qualquer intervalo $J \subset \mathbf{T}$, o módulo intervalar $\mathfrak{I} = \mathbf{k}^J$ é definido como o \mathbf{T} -módulo de persistência com a seguinte família de espaços vetoriais

$$I_t = \begin{cases} \mathbf{k} & \text{se } t \in J \\ 0 & \text{caso contrário,} \end{cases}$$

e as aplicações lineares

$$\dot{i}_t^s = \begin{cases} id & \text{se } s, t \in J \\ 0 & \text{caso contrário.} \end{cases}$$

Como mencionado anteriormente, os intervalos seriam as propriedades representadas no diagrama de persistência, ou seja, o módulo intervalar \mathbf{k}^J representa uma propriedade que persiste por todo intervalo J .

Devida a sua importância, módulos intervalares com índices em subconjuntos de \mathbb{R} possuem uma notação especial. Para distinguir os vários casos de intervalos, usamos uma supernotação: $+$ e $-$, a decoração dos pontos. Para intervalos finitos adota-se o seguinte dicionário

$$(p^-, q^-) = [p, q)$$

$$(p^-, q^+) = [p, q]$$

$$(p^+, q^-) = (p, q)$$

$$(p^+, q^+) = (p, q]$$

O dicionário acima vale para $p < q$. No caso em que $p = q$, representamos o intervalo por $(p^-, p^+) = [p, p]$. Para intervalos infinitos, usamos o símbolo $-\infty^+$ e $+\infty^-$ com definição similar à acima e com a adição do seguinte intervalo

$$(-\infty^+, +\infty^-) = (-\infty, +\infty).$$

Quando queremos referenciar um ponto decorados mas não sabemos sua decoração, denotamos por p^* , podendo ser p^- ou p^+ .

Podemos estender os reais para os reais decorados, um conjunto totalmente ordenado com as seguintes relações

$$p^- < p < p^+ < q^- < q < q^+,$$

para todo $p < q$. Definimos o semiplano diagonal superior em \mathbb{R}^2 como

$$\mathcal{H} = \{ (p, q) \mid p \leq q \}$$

O semiplano diagonal superior $\tilde{\mathcal{H}}$ é a união de \mathcal{H} com os pontos no infinito.

Portanto, um módulo intervalar pode ser representado de diversas formas, visualizados também na [Figura 15](#)

- Como um intervalo na reta real;
- como uma função $\mathcal{H} \rightarrow \{0, 1\}$ definida por $(s, t) \mapsto \text{rank}(i_t^s)$;
- como um ponto $(p, q) \in \mathcal{H}$ e um traço representando a respectiva decoração.

Os traços representando a decoração são dados por

$$(p^-, q^-): \swarrow$$

$$(p^-, q^+): \searrow$$

$$(p^+, q^-): \swarrow$$

$$(p^+, q^+): \nearrow$$

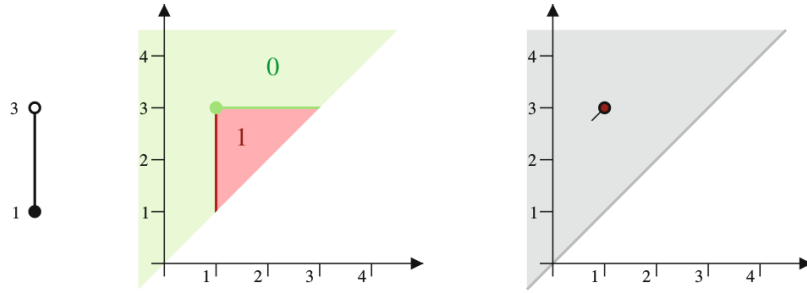


Figura 15 – Representação por intervalo (esquerda), pela função rank (meio) e pelo ponto decorado (direita) do módulo intervalar $\mathbf{k}[1, 3) = \mathbf{k}(1^-, 3^-)$.

3.1.4 Decomposição em módulos intervalares

Definição 3.2. A soma direta $\mathfrak{W} = \mathfrak{U} \oplus \mathfrak{V}$ de dois módulos de persistência \mathfrak{U} e \mathfrak{V} é definida por

$$W_t = U_t \oplus V_t, \quad w_t^s = u_t^s \oplus v_t^s$$

Esta definição generaliza-se para somar arbitrárias, tanto finitas como infinitas. Vamos agora definir a indecomponibilidade de um módulo de persistência.

Definição 3.3. Um módulo de persistência \mathfrak{W} é indecomponível se dada uma decomposição $\mathfrak{U} \oplus \mathfrak{V}$, então $\mathfrak{U} = 0$ ou $\mathfrak{W} = \mathfrak{V} = 0$ ou \mathfrak{W} .

Podemos estudar os módulos de persistência através de sua decomposição por módulos intervalares. Dado uma sequência de intervalos $(J_l \mid l \in L)$,

$$\mathfrak{W} \cong \bigoplus_{l \in L} \mathbf{k}^{J_l}.$$

Neste caso, podemos pensar que cada intervalo J_l representa uma propriedade. Esta decomposição acaba sendo muito importante por este motivo. Mas a questão que fica é: quais módulos são decomponíveis em intervalos? E porque decompõe-se em módulos intervalares?

A resposta para a primeira pergunta é o Teorema 3.7. Já para a segunda questão, os módulos intervalares são indecomponíveis, como mostramos na Proposição 3.5.

Proposição 3.4. Seja $\mathfrak{J} = \mathbf{k}_T^J$ um módulo intervalar sobre $\mathbf{T} \subset \mathbb{R}$. Então $\text{End}(\mathfrak{J}) = \mathbf{k}$.

Demonstração. Vamos definir uma função Φ entre $\text{End}(\mathfrak{J})$ e \mathbf{k} que será um isomorfismo de anéis. Seja $\Phi: \mathbf{k} \rightarrow \text{End}(\mathfrak{J})$ definida por

$$\alpha \mapsto \varphi^\alpha$$

onde φ^α é um endomorfismo de \mathfrak{J} tal que $\varphi_t^\alpha: I_t \rightarrow I_t$ e $\varphi_t^\alpha(x) = \alpha x$. É fácil ver que a aplicação é um homomorfismo de anéis. Vamos definir a inversa de Φ . Para isso, note primeiro que

qualquer endomorfismo de \mathfrak{J} age como multiplicação por escalar em qualquer I_t não nulo. Precisamos mostrar que dados s, t , temos que o escalar definido é o mesmo para ambos os casos:

$$\begin{aligned}\Psi^{-1}:\mathfrak{J} &\rightarrow \mathbf{k} \\ \varphi &\mapsto \alpha,\end{aligned}$$

Vamos mostrar que a aplicação está bem definida.

Primeiro, pela observação acima, dados s, t tais que $I_s, I_t \neq 0$, temos que vale o seguinte para $\varphi \in \mathfrak{J}$.

$$\begin{aligned}\varphi_s:\mathbf{k} &\rightarrow \mathbf{k} \\ x &\mapsto \alpha x\end{aligned}$$

e

$$\begin{aligned}\varphi_t:\mathbf{k} &\rightarrow \mathbf{k} \\ x &\mapsto \beta x\end{aligned}$$

Precisamos mostrar que $\alpha = \beta$, demonstrando a proposição. Mas isso segue pelo diagrama comutativo dos homomorfismos entre módulos de persistência, como podemos ver na (3.1), assumindo que $s \leq t$.

$$\begin{array}{ccc} I_s & \xrightarrow{id} & I_t \\ \varphi_s \downarrow & & \downarrow \varphi_t \\ I_s & \xrightarrow{id} & I_t \end{array} \quad (3.1)$$

No caso acima temos a identidade entre I_s e I_t , já que ambos são \mathbf{k} . Logo, segue que $\alpha = \beta$, provando a Proposição. \square

Proposição 3.5. Módulos intervalares são indecomponíveis.

Demonstração. Suponha que exista uma decomposição $\mathfrak{J} = \mathfrak{U} \oplus \mathfrak{V}$. Considere agora as projeções sob \mathfrak{U} e \mathfrak{V} . Ambas são homomorfismos idempotentes. Mas como $\text{End}(\mathfrak{J})$ é isomorfo a \mathbf{k} e os únicos idempotentes de \mathbf{k} são 0 e 1, segue que \mathfrak{J} é indecomponível. \square

Teorema 3.6. (Krull-Remak-Schmidt-Azumaya) Suponha que um módulo de persistência $\mathbf{T} \subset \mathbb{R}$ pode ser escrito como soma direta de módulos intervalares de duas formas diferentes

$$\mathfrak{V} \cong \bigoplus_{l \in L} \mathbf{k}^{J_l} \cong \bigoplus_{m \in M} \mathbf{k}^{K_m},$$

então existe uma bijeção $\sigma:L \rightarrow M$ tal que $J_l = K_{\sigma(l)}$ para todo $l \in L$.

Demonstração. A demonstração segue do Teorema 1 (AZUMAYA, 1950) com a observação de que se $\mathbf{k}^J \cong \mathbf{k}^L$, então $J = K$. Só é necessário verificar uma condição de localidade para aplicarmos o teorema: se $\psi, \phi \in \text{End}(\mathfrak{J})$ são não isomorfismos, então $\psi + \phi$ não é isomorfismo. Mas pela proposição anterior, isso segue do fato que a única aplicação que não é isomorfismo em $\text{End}(\mathfrak{J})$ é a aplicação nula. \square

Teorema 3.7. (Gabriel, Auslander, Ringel-Tachikawa, Webb, Crawley-Boevey) Seja \mathfrak{V} um módulo de persistência sobre $\mathbf{T} \subset \mathbb{R}$. Então \mathfrak{V} pode ser decomposto como uma soma direta de módulos intervalares sob as seguintes condições:

- \mathbf{T} é um conjunto finito;
- cada V_t é um espaço vetorial de dimensão finita.

Por outro lado, existe um módulo de persistência sob \mathbb{Z} que não admite uma decomposição intervalar.

Demonstração. Detalhes podem ser vistos em (CHAZAL et al., 2016), página 22, **Teorema 2.8.** \square

Se um módulo de persistência indexado sobre \mathbb{R} pode ser decomposto

$$\mathfrak{V} \cong \bigoplus_{l \in L} (p_l^*, q_l^*),$$

então o diagrama de persistência decorado é definido pelo multiconjunto

$$\text{Dgm}(\mathfrak{V}) = \text{Int}(\mathfrak{V}) = \{ (p_l^*, q_l^*) \mid l \in L \}$$

e o diagrama de persistência não decorado é o multiconjunto

$$\text{dgm}(\mathfrak{V}) = \text{int}(\mathfrak{V}) = \{ (p_l, q_l) \mid l \in L \} - \Delta,$$

onde Δ é a diagonal no plano.

Note que ambos os diagramas definidos não dependem da escolha da decomposição, devido ao Teorema 3.6. Além disso, o diagrama dgm é o diagrama de pontos não decorados e sem a diagonal, sendo encontrado com frequência em exemplos práticos de análise de dados. Para a definição da distância bottleneck acaba sendo mais importante.

3.1.5 Cálculos com quivers

Vamos agora definir uma notação para trabalhar com módulos de persistência sobre conjuntos de índices finitos. Um módulo de persistência \mathfrak{V} sobre um conjunto finito de índices

$$\mathbf{T}: \quad a_1 < \dots < a_n$$

da reta real pode ser visto como um diagrama de n espaços vetoriais e $n - 1$ aplicações lineares, como mostrado abaixo

$$\mathfrak{V}: V_{a_1} \longrightarrow \dots \longrightarrow V_{a_n}.$$

O diagrama acima é a representação do seguinte **quiver**:

$$\bullet \longrightarrow \bullet \longrightarrow \dots \longrightarrow \bullet$$

Vimos que podemos decompor alguns módulos de persistência em módulos intervalares. Para estes podemos representa-los com quivers da seguinte forma.

Exemplo 3.8. Seja $a < b < c$. Existem 6 módulos intervalares diferentes sobre este intervalo.

$$\begin{aligned} \mathbf{k}[a, a] &= \bullet_a \text{---} \circ_b \text{---} \circ_c & \mathbf{k}[a, b] &= \bullet_a \text{---} \bullet_b \text{---} \circ_c & \mathbf{k}[a, c] &= \bullet_a \text{---} \circ_b \text{---} \bullet_c \\ \mathbf{k}[b, b] &= \circ_a \text{---} \bullet_b \text{---} \circ_c & \mathbf{k}[b, c] &= \circ_a \text{---} \bullet_b \text{---} \bullet_c \\ \mathbf{k}[c, c] &= \circ_a \text{---} \circ_b \text{---} \bullet_c \end{aligned}$$

Os círculos \bullet representam uma cópia do espaço vetorial \mathbf{k} unidimensional. O círculo \circ representa o espaço vetorial nulo. A aplicação linear entre dois \bullet é a identidade e qualquer aplicação contendo \circ é a nula.

Esta notação pode ser usada para representar a multiplicidade dos módulos intervalares da decomposição de um módulo de persistência sobre um conjunto de índices finito, essa quando existe. Seja \mathfrak{V} um módulo de persistência sobre o conjunto $\mathbf{T} = \{a_1, \dots, a_n\}$. Definimos a multiplicidade de $[a_i, a_j] \subseteq \mathbf{T}$ em $\mathfrak{V}_{\mathbf{T}}$ como o número de cópias do módulo $\mathbf{k}[a_i, a_j]$ na decomposição de $\mathfrak{V}_{\mathbf{T}}$.

Exemplo 3.9. Seja \mathfrak{V} módulo de persistência sobre $\mathbf{T} = \{a, b, c\}$. Escrevemos

$$\langle [b, c] | \mathfrak{V}_{\mathbf{T}} \rangle \text{ ou } \langle \circ_a \text{---} \bullet_b \text{---} \bullet_c | \mathfrak{V} \rangle$$

para representar a multiplicidade de $\circ_a \text{---} \bullet_b \text{---} \bullet_c$ no seguinte módulo de 3 termos

$$\mathfrak{V}: V_{a_1} \longrightarrow \dots \longrightarrow V_{a_n}.$$

Exemplo 3.10. Considere o módulo com dois espaços vetoriais e uma única aplicação linear $\mathfrak{V}: V_a \xrightarrow{\mu} V_b$. Então os invariantes de μ são

$$\begin{aligned} \text{rank}(\mu) &= \langle \bullet_a \text{---} \bullet_b | \mathfrak{V} \rangle, \\ \text{nulidade}(\mu) &= \langle \bullet_a \text{---} \circ_b | \mathfrak{V} \rangle, \\ \text{conulidade}(\mu) &= \langle \circ_a \text{---} \bullet_b | \mathfrak{V} \rangle. \end{aligned}$$

Basta notar que para V_a, V_b espaços de dimensão finita, existe uma decomposição das suas bases

$$e_1, \dots, e_r, f_1, \dots, f_n \text{ e } g_1, \dots, g_r, h_1, \dots, h_m$$

tais que $\mu(e_i) = g_i$, $\mu(f_j) = 0$ para todo i, j . Assim, os espaços vetoriais unidimensionais gerados pelos elementos das bases geram uma decomposição do módulo \mathfrak{V} nos seguintes intervalos

$$\begin{aligned} (\text{span}(e_i) &\rightarrow \text{span}(g_i)) \\ (\text{span}(f_j) &\rightarrow 0) \\ (0 &\rightarrow \text{span}(h_k)) \end{aligned}$$

que são isomorfos respectivamente à $\bullet_a \text{---} \bullet_b$, $\bullet_a \text{---} \circ_b$ e $\circ_a \text{---} \bullet_b$.

Proposição 3.11. Suponha que podemos decompor um módulo de persistência \mathfrak{V} como uma soma direta

$$\mathfrak{V} = \bigoplus_{l \in L} \mathfrak{V}^l,$$

então

$$\langle [a_i, a_j] \mid \mathfrak{V}_{\mathbf{T}} \rangle = \sum_{l \in L} \langle [a_i, a_j] \mid \mathfrak{V}_{\mathbf{T}}^l \rangle$$

para qualquer conjunto de índices $\mathbf{T} = \{a_1, \dots, a_n\}$ e intervalos $[a_i, a_j] \subseteq \mathbf{T}$.

Demonstração. Segue do fato que a decomposição intervalar de $\mathfrak{V}_{\mathbf{T}}$ é a soma direta das decomposições intervalares de cada $\mathfrak{V}_{\mathbf{T}}^l$ para todo $l \in L$. \square

Proposição 3.12. (Princípio da restrição) Sejam \mathbf{S}, \mathbf{T} conjuntos de índices com $\mathbf{S} \subset \mathbf{T}$. Então

$$\langle I \mid \mathfrak{V}_{\mathbf{S}} \rangle = \sum_J \langle J \mid \mathfrak{V}_{\mathbf{T}} \rangle,$$

onde a soma é sobre todos os intervalos $J \subseteq \mathbf{T}$ que se restringe a I sobre \mathbf{S} .

Demonstração. Tome uma decomposição intervalar arbitrária de $\mathfrak{V}_{\mathbf{T}}$. Então uma decomposição intervalar é induzida em $\mathfrak{V}_{\mathbf{S}}$. Note agora que para $I \subseteq \mathbf{S}$, temos diversos intervalos $J \subseteq \mathbf{T}$ tais que $J \cap \mathbf{S} = I$. Devido a linearidade da soma direta, temos que os intervalos de $\mathfrak{V}_{\mathbf{S}}$ do tipo I são os intervalos de $\mathfrak{V}_{\mathbf{T}}$ do tipo J acima. \square

Exemplo 3.13. Seja $a < p < b < q < c$. Então temos os seguintes exemplos para os conjuntos de índices:

- $\mathbf{T}_1 = \{a, b, q, c\}$, $S_1 = \{a, b, c\}$, $I_1 = [b, c]$.

$$\langle \circ_a \text{---} \bullet_b \text{---} \bullet_c \rangle = \langle \circ_a \text{---} \bullet_b \text{---} \bullet_q \text{---} \bullet_c \rangle$$

- $\mathbf{T}_2 = \{a, p, b, c\}$, $S_2 = \{a, b, c\}$, $I_2 = [b, c]$.

$$\langle \circ_a \text{---} \bullet_b \text{---} \bullet_c \rangle = \langle \circ_a \text{---} \circ_p \text{---} \bullet_b \text{---} \bullet_c \rangle + \langle \circ_a \text{---} \bullet_p \text{---} \bullet_b \text{---} \bullet_c \rangle$$

- $\mathbf{T}_3 = \{a, b, q, c\}$, $S_2 = \{a, b, c\}$, $I_2 = [c, c]$.

$$\langle \circ_a \text{---} \circ_b \text{---} \bullet_c \rangle = \langle \circ_a \text{---} \circ_b \text{---} \circ_q \text{---} \bullet_c \rangle + \langle \circ_a \text{---} \circ_b \text{---} \bullet_q \text{---} \bullet_c \rangle$$

3.2 Medidas retangulares

Na seção anterior discutimos módulos de persistência decomponíveis e seus diagramas de persistência, Dgm e dgm. No entanto, nem sempre os módulos são decomponíveis, não sendo possível definir os diagramas de persistência. Para definir-los, podemos nos guiar pela seguinte ideia: se soubermos contar o número de pontos do Dgm pertence em cada retângulo do semiplano, então conhecemos Dgm.

Iremos nos inspirar na teoria da medida para construir uma função que nos dá um valor inteiro ou infinito e que podemos associar um diagrama de persistência com ela. A ideia é que para módulos bem comportados podemos avaliar esta função em retângulos e extrair um conjunto discreto de pontos, que juntamente com sua multiplicidade gerará o diagrama de persistência. No caso que o módulo de persistência for decomponível, as definições são iguais. Caso contrário seguimos com a teoria normalmente.

No resto deste capítulo iremos tratar apenas de medidas finitas no semiplano diagonal sem considerar os pontos no infinito. Os argumentos usados podem ser facilmente estendidos para o caso de medidas infinitas através de um processo de limite e para o semiplano diagonal com os pontos no infinito podemos usar o truque de colocar tudo dentro de um retângulo com a função \arctan .

3.2.1 A medida de persistência

Definição 3.14. Seja \mathfrak{V} um módulo de persistência. Então a medida de persistência de \mathfrak{V} é a função

$$\mu_{\mathfrak{V}}(R) = \langle \circ_a \text{---} \bullet_b \text{---} \bullet_c \text{---} \circ_d \mid \mathfrak{V} \rangle$$

definida no retângulo $R = [a, b] \times [c, d]$ no plano com $a < b \leq c < d$.

Veremos como a medida tem uma relação com módulos decomponíveis. Abaixo um resultado para módulos intervalares.

Proposição 3.15. Seja $\mathfrak{V} = \mathbf{k}^J$, em que $J = (p^*, q^*)$ é um intervalo real. Seja $R = [a, b] \times [c, d]$ tal que $a < b \leq c < d$. Então

$$\mu_{\mathfrak{V}}(R) = \begin{cases} 1 & \text{se } [b, c] \subseteq J \subseteq (a, d) \\ 0 & \text{caso contrário.} \end{cases}$$

Demonstração. Como \mathbf{k}^J restrito a $\{a, b, c, d\}$ é só um intervalo ou o módulo nulo, temos que $\mu_{\mathfrak{V}}(R) \leq 1$, pois apenas teríamos a função identidade, cujo rank é um, ou a função nula, cujo rank é zero.

Como a medida tem valores em $\{0, 1, \dots\} \cup \infty$, vamos averiguar quando acontece $\mu_{\mathfrak{M}}(R) = 1$. Note que $\mu_{\mathfrak{M}}(R) = 1$ quando

$$\mathbf{k}_{\{a,b,c,d\}}^J = \circ_a \text{---} \bullet_b \text{---} \bullet_c \text{---} \circ_d.$$

E esta restrição vale se e somente se $b, c \in J$ e $a, d \notin J$. \square

A Proposição 3.15 pode ser representada visualmente. Considere o módulo intervalo como um ponto decorado (p^*, q^*) no semiplano diagonal superior. Então se o ponto estiver no interior do retângulo R , ele será detectado independente da decoração. Mas se estiver na borda, apenas aqueles cuja decoração apontem para dentro do retângulo serão detectados, como pode ser visto na Figura 16



Figura 16 – Pontos decorados que são detectados pela medida aplicada no retângulo R .

Definição 3.16. Seja $R = [a, b] \times [c, d]$, em que $a < b \leq c < d$ e considere o ponto decorado (p^*, q^*) com $p^* < q^*$. Definimos $(p^*, q^*) \in R$ se alguma das condições equivalentes é verdade:

- Se $p^* \in [a, b]$ e $q^* \in [c, d]$;
- Se $a < p^* < b$ e $c < q^* < d$ na ordem total definida anteriormente;
- Se $a^+ \leq p^* \leq b^-$ e $c^+ \leq q^* \leq d^-$;
- Se o intervalo satisfaz: $[b, c] \subseteq (p^*, q^*) \subseteq (a, d)$;
- O ponto com o traço (p^*, q^*) está dentro do retângulo R .

Definição 3.17. Definimos com o **r-interior** do retângulo R o conjunto

$$R^\times = \{ (p^*, q^*) \mid (p^*, q^*) \in R \}.$$

Também podemos definir o **interior** de R como o conjunto

$$R^\circ = (a, b) \times (c, d).$$

A expressão $|_R$ indica a restrição de um multiconjunto de pontos decorados no retângulo R .

Corolário 3.18. Suponha que \mathfrak{V} seja um módulo de persistência decomponível sobre R .

$$\mathfrak{V} = \bigoplus_{l \in L} k(p_l^*, q_l^*).$$

Então

$$\mu_{\mathfrak{V}}(R) = \text{card}(\text{Dgm}(\mathfrak{V})|_R)$$

para todo retângulo $R = [a, b] \times [c, d]$ com $a < b \leq c < d$.

Demonstração. A demonstração segue direto das Proposições 3.15 e 3.11. \square

A função μ é chamada de medida pois é aditiva em relação a divisão dos retângulos. Vamos provar este fato agora.

Proposição 3.19. $\mu_{\mathfrak{V}}$ é aditiva sobre divisão vertical e horizontal dos retângulos:

$$\begin{aligned} \mu_{\mathfrak{V}}([a, b] \times [c, d]) &= \mu_{\mathfrak{V}}([a, p] \times [c, d]) + \mu_{\mathfrak{V}}([p, b] \times [c, d]) \\ \mu_{\mathfrak{V}}([a, b] \times [c, d]) &= \mu_{\mathfrak{V}}([a, b] \times [c, q]) + \mu_{\mathfrak{V}}([a, b] \times [q, d]) \end{aligned}$$

para todo $a < p < b \leq c < q < d$. Esta propriedade pode ser visualizada na Figura 17.

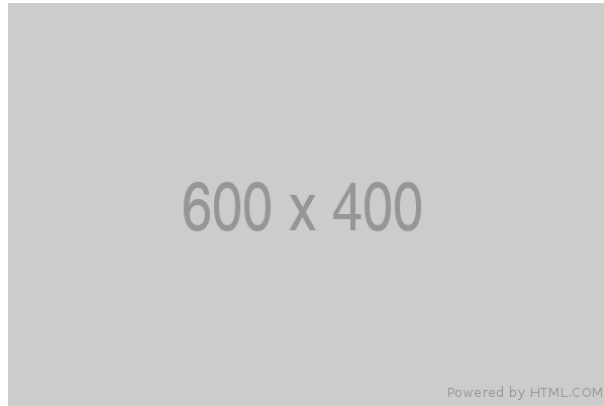


Figura 17 – Representação gráfica da Proposição 3.19

Demonstração. A demonstração segue direto da Proposição 3.12: para a aditividade na divisão horizontal temos que

$$\begin{aligned} \mu_{\mathfrak{V}}([a, b] \times [c, d]) &= \langle \circ_a \text{---} \bullet_b \text{---} \bullet_c \text{---} \circ_d \rangle \\ &= \langle \circ_a \text{---} \bullet_p \text{---} \bullet_b \text{---} \bullet_c \text{---} \circ_d \rangle + \langle \circ_a \text{---} \circ_p \text{---} \bullet_b \text{---} \bullet_c \text{---} \circ_d \rangle \\ &= \langle \circ_a \text{---} \bullet_p \text{---} \bullet_c \text{---} \circ_d \rangle + \langle \text{---} \circ_p \text{---} \bullet_b \text{---} \bullet_c \text{---} \circ_d \rangle \\ &= \mu_{\mathfrak{V}}([a, p] \times [c, d]) + \mu_{\mathfrak{V}}([p, b] \times [c, d]). \end{aligned}$$

De forma análoga vale o resultado para a divisão vertical. \square

3.2.2 *r-medidas abstratas*

Até agora trabalhamos com módulos de persistência e uma medida associada. Porém, podemos trabalhar de forma mais abstrata, sem mencionar os módulos.

Definição 3.20. Seja $\mathcal{D} \subseteq \mathbb{R}^2$. Defina

$$\text{Rect}(\mathcal{D}) = \{ [a, b] \times [c, d] \subset \mathcal{D} \mid a < b \text{ e } c < d \}.$$

A **r-medida** ou **medida retangular** em \mathcal{D} é uma função

$$\mu: \text{Rect}(\mathcal{D}) \rightarrow \{0, 1, \dots\} \cup \{\infty\},$$

que também é aditiva na divisão horizontal e vertical dos retângulos.

Proposição 3.21. Seja μ uma **r-medida** em $\subseteq \mathbb{R}^2$. Então

- Se $R \in \text{Rect}(\mathcal{D})$ pode ser escrito como a união de retângulos com interior disjuntos, $R = R_1 \cup \dots \cup R_n$, então $\mu(R) = \sum \mu(R_i)$;
- Se $R \subseteq S$, então $\mu(R) \leq \mu(S)$.

Demonstração. (*Finitamente aditiva*) Seja $R = [a, b] \times [c, d]$. Por indução e pela propriedade de divisão vertical, temos que a aditividade finita vale para decomposições da forma

$$R = \bigcup_i R_i,$$

em que cada $R_i = [a_i, a_{i+1}] \times [c, d]$, com $a = a_1 < \dots < a_m = b$. De forma análoga, vale para divisões horizontais e portanto a aditividade vale para decomposições da forma

$$R = [a, b] \times [c, d] = \bigcup_{i,j} R_{ij},$$

onde $R_{ij} = [a_i, a_{i+1}] \times [c_j, c_{j+1}]$ com $a = a_1 < \dots < a_m = b$ e $c = c_1 < \dots < c_n = d$. Para uma decomposição arbitrária $R = R_1 \cup \dots \cup R_k$ o resultado segue considerando uma decomposição em que cada R_i é decomposto em intervalos da forma acima.

(*Monotonicidade*) Decomponha S em uma coleção de retângulos R e R_1, \dots, R_k que possuem interiores disjuntos. Portanto, da propriedade de aditividade e que $\mu \geq 0$

$$\begin{aligned} \mu(S) &= \mu(R) + \mu(R_1) + \dots + \mu(R_k) \\ &\geq \mu(R). \end{aligned}$$

□

Proposição 3.22. (Subaditividade) Seja μ uma **r-medida** em $\mathcal{D} \subseteq \mathbb{R}^2$. Se um retângulo $R \in \text{Rect}(\mathcal{D})$ está contido numa união finita de retângulos $R_i \in \text{Rect}(\mathcal{D})$

$$R \subseteq R_1 \cup \cdots \cup R_k,$$

então

$$\mu(R) \leq \mu(R_1) + \cdots + \mu(R_k).$$

Demonstração. Seja $a_1 < \cdots < a_m$ sequência de todos os valores do eixo x de cada vértice dos retângulos. Considere também $c_1 < \cdots < c_n$ sequência de todos os valores do eixo y de cada vértice dos retângulos. Então cada retângulo R_k pode ser decomposto como união dos seguintes subretângulos

$$[a_i, a_{i+1}] \times [c_j, c_{j+1}],$$

que possuem interiores disjuntos por construção e sua medida é a soma das medidas de cada subretângulo. Como cada subretângulo de R pertence a um ou mais retângulos R_i , segue da aditividade o resultado. A Figura 18 mostra uma demonstração visual desta proposição. \square



Figura 18 – Demonstração da proposição através da figura.

3.2.3 Equivalência de medidas e diagramas

Teorema 3.23.

3.3 Comportamento de módulos e exemplos

3.4 *Interleaving*

3.5 O teorema de isometria

GERADORES ÓTIMOS E OUTROS CONCEITOS

4.1 Geradores ótimos

4.2 Vetorização do diagrama de persistência

Dado uma sequência de conjuntos de dados X_i e os respectivos diagramas de persistência tem-se uma variação no seus tamanhos, devido a natureza do algoritmo de homologia persistente. Além de variação entre os tamanhos, cada diagrama é um multi-conjunto, sendo mais difícil de analisar-los. Ao utilizar algoritmos de machine learning, assume-se entradas com tamanhos fixos no conjunto inteiro de dados. Portanto diagramas de persistências descrevendo uma sequência de proteínas, por exemplo, precisam ser vetorizados de alguma forma antes de podermos utilizar em conjunto com outros algoritmos, como redes neurais ou regressão linear.

Existem várias formas de vetorização de um diagrama de persistência, como *Persistence landscapes* (BUBENIK, 2015) e Imagem de persistência (*Persistence Image*) (ADAMS *et al.*, 2017). Neste trabalho apresentamos a imagem de persistência e alguns exemplos.

4.2.1 Estabilidade da Imagem de Persistência

Imagem de persistência é uma vetorização de forma que o respectivo diagrama é representado como uma imagem de tamanho fixo (n, m) . De forma intuitiva esse método é uma forma de suavização do diagrama de persistência, em que uma gaussiana é centrada em cada ponto e depois são somadas com peso. Abaixo descrevemos formalmente

o processo para obter uma imagem de persistência.

Seja $D = \{(b_i, d_i)\}_i$ um diagrama de persistência em alguma dimensão e considere $T(x, y) = (x, y - x)$ uma transformação linear em \mathbb{R}^2 . Seja $T(B)$ o multiconjunto decorrente da transformação linear T aplicada em B onde cada ponto $(x, y) \in B$ corresponde ao ponto $(x, y - x) \in T(B)$. Considere agora uma função de probabilidade diferenciável $\phi_u: \mathbb{R}^2 \rightarrow \mathbb{R}$ com média $u = (u_x, u_y)$.

Fixe agora uma função peso $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ de tal forma que ela é zero no eixo horizontal, contínua e diferenciável por partes. É importante que essas condições sejam satisfeitas, pois elas garante a estabilidade da imagem de persistência sob a distância 1-Wasserstein. Dessa forma temos a seguinte definição.

Definição 4.1. Para um diagrama de persistência B , a correspondente superfície de persistência $\rho_B: \mathbb{R}^2 \rightarrow \mathbb{R}$ é a função dada por

$$\rho_B(z) = \sum_{u \in T(B)} f(u) \phi_u(z).$$

Entretanto, um computador não consegue utilizar uma função para fazer cálculos e estimativas, ela precisa ser vetorizada (ou discretizada) de alguma forma. Desta forma, vamos discretizar ρ_B em um domínio específico, que depende de $T(B)$. Em específico, fixamos um grid e o valor de cada pixel é dado pela integral nessa região.

Definição 4.2. Seja B um diagrama de persistência. A imagem de persistência de B é a coleção de pixels

$$I(\rho_B)_p = \iint_p \rho_B dy dx.$$

Na vetorização do diagrama alguns parâmetros precisam ser estabelecidos. Em (ADAMS *et al.*, 2017) mostra-se que as imagens são robustas sob a escolha da resolução (tamanho do grid). A outra escolha é a distribuição e dependendo a variância. Em (ADAMS *et al.*, 2017) a distribuição gaussiana é utilizada com variância dependendo do problema e sendo assim o usuário a escolhe. Por último, a escolha da função peso, que pode variar de problema pra problema. A função abaixo é um exemplo utilizado por (ADAMS *et al.*, 2017). Observe que para pontos com valores altos de persistência tem um valor maior também. Mas com problemas que pontos de baixa ou média persistência são importantes, a utilização de outros pontos se faz necessária.

$$w_b(t) = \begin{cases} 0 & \text{if } t \leq 0, \\ \frac{t}{b} & \text{if } 0 < t < b, \\ 1 & \text{if } t \geq b, \end{cases} \quad (4.1)$$

onde b é considerado o valor de maior persistência em $T(B)$.

Em vários conjuntos é normal que apresentem ruídos e algumas variações, assim dando diagramas de persistência diferentes. Entretanto, há uma medida para avaliar a distância entre eles.

Definição 4.3. A distância p -Wasserstein definida entre dois diagramas de persistência B e B' é dada por

$$W_p(B, B') = \inf_{\gamma: B \rightarrow B'} \left(\sum_{u \in B} \|u - \gamma(u)\|_\infty^p \right)^{\frac{1}{p}},$$

onde $1 \leq p < \infty$ e γ é bijeção entre B e B' .

Seja $h: \mathbb{R}^2 \rightarrow \mathbb{R}$ uma função diferenciável. Denote $|\nabla h| = \sup_{z \in \mathbb{R}^2} \|\nabla h(z)\|_2$. Pelo teorema do valor médio, temos que

$$|h(u) - h(v)| \leq |\nabla h| \|u - v\|_2. \quad (4.2)$$

Seja $u, v \in \mathbb{R}^2$ e considere as duas distribuições diferenciáveis ϕ_u, ϕ_v . Como o supremo e a derivada de direção maximal de uma distribuição de probabilidade diferenciável são invariantes por translação, podemos denotar $|\nabla \phi_u|$ por $|\nabla \phi|$ e $\|\phi_u\|_\infty$ por $\|\phi\|_\infty$. E observe ainda devido a invariância pela translação, temos que

$$\|\phi_u - \phi_v\|_\infty \leq |\nabla \phi| \|u - v\|_2. \quad (4.3)$$

Vamos enunciar um lema agora que será utilizado nas provas de estabilidade das imagens de superfície e persistência.

Lema 4.4. Sejam $u, v \in \mathbb{R}^2$, temos que $\|f(u)\phi_u - f(v)\phi_v\| \leq (\|f\|_\infty |\nabla \phi| + \|\phi\|_\infty |\nabla f|) \|u - v\|_2$.

Demonstração. Seja $z \in \mathbb{R}^2$ qualquer, então

$$\begin{aligned} |f(u)\phi_u(z) - f(v)\phi_v(z)| &= |f(u)(\phi_u(z) - \phi_v(z)) + (f(u) - f(v))\phi_v(z)| \\ &\leq \|f\|_\infty |\phi_u(z) - \phi_v(z)| + \|\phi\|_\infty |f(u) - f(v)| \\ &\leq \|f\|_\infty |\nabla \phi| \|u - v\|_2 + \|\phi\|_\infty |\nabla f| \|u - v\|_2 \quad \text{por 4.3 e 4.2} \\ &= (\|f\|_\infty |\nabla \phi| + \|\phi\|_\infty |\nabla f|) \|u - v\|_2. \end{aligned}$$

□

Teorema 4.5. A superfície de persistência é estável em relação a distância 1-Wasserstein. Dados B, B' diagramas de persistência finitos, temos que

$$\|\rho_B - \rho_{B'}\|_\infty \leq \sqrt{10} (\|f\|_\infty |\nabla \phi| + \|\phi\|_\infty |\nabla f|) W_1(B, B')$$

Demonstração. Por hipótese, B e B' são finitos, logo existe uma bijeção entre B e B' que atinge o ínfimo da distância de Wasserstein. Portanto

$$\begin{aligned}
\|\rho_B - \rho_{B'}\|_\infty &= \left\| \sum_{u \in T(B)} f(u) \phi_u - \sum_{u \in T(B)} f(\gamma(u)) \phi_{\gamma(u)} \right\|_\infty \\
&\leq \sum_{u \in T(B)} \|f(u) \phi_u - f(\gamma(u)) \phi_{\gamma(u)}\| \\
&\leq (\|f\|_\infty |\nabla \phi| + \|\phi\|_\infty + |\nabla f|) \sum_{u \in T(B)} \|u - \gamma(u)\|_2 \quad \text{por 4.4} \\
&\leq \sqrt{2} (\|f\|_\infty |\nabla \phi| + \|\phi\|_\infty + |\nabla f|) \sum_{u \in T(B)} \|u - \gamma(u)\|_\infty \quad \text{já que } \|\cdot\|_2 \leq \sqrt{2} \|\cdot\|_\infty \text{ em } \mathbb{R}^2 \\
&\leq \sqrt{10} (\|f\|_\infty |\nabla \phi| + \|\phi\|_\infty + |\nabla f|) \sum_{u \in T(B)} \|u - \gamma(u)\|_\infty \quad \text{já que } \|T(\cdot)\|_2 \leq \sqrt{5} \|\cdot\|_\infty \\
&= \sqrt{10} (\|f\|_\infty |\nabla \phi| + \|\phi\|_\infty + |\nabla f|) W_1(B, B').
\end{aligned}$$

A última desigualdade é necessária, pois a distância de Wasserstein é definida sobre os pontos do diagrama de persistência que são da forma nascimento e morte, não nascimento e persistência. \square

E por fim, temos que as imagens de persistência são estáveis.

Teorema 4.6. A imagem de persistência é estável em relação a distância 1-Wasserstein. Se A é o valor máximo dentre todos os pixels da imagem, então

$$\|I(\rho_B) - I(\rho_{B'})\|_\infty \leq \sqrt{10} (\|f\|_\infty |\nabla \phi| + \|\phi\|_\infty + |\nabla f|) W_1(B, B'). \quad (4.4)$$

Demonstração. A demonstração segue do Teorema 4.5 e do fato que para um pixel p qualquer

$$|I(\rho_B)_p - I(\rho_{B'})_p| \leq A(p) \|\rho_B - \rho_{B'}\|_\infty,$$

em que $A(p)$ representa a área do píxel p . \square

4.2.2 Exemplos de Imagens de Persistência

Considere X um conjunto de pontos extraídos de um círculo com ruído, como pode ser visto na Figura 19. Na Figura 20 tem-se os diagramas de persistência do círculo de dimensão 0 e 1, assim mostrando as componentes conexas e buracos. Note que existem dois pontos longe da diagonal, um representando a componente conexa e o outro o buraco do círculo. Vamos agora analisar as imagens de persistência para cada uma das imagens. Escolhemos a distribuição gaussiana dada por

$$g_u(x, y) = \frac{1}{2\pi\sigma^2} e^{-((x-u_x)^2 + (y-u_y)^2)/2\sigma^2}.$$

A Equação 4.1 é utilizada como função peso. Para calcular as imagens de persistência definimos três variâncias: 0.001, 0.1, 1.0, e dois tamanhos de imagem: 10×10 e 50×50 . O resultado pode ser visto na Figura 21.

Observe a diferença entre os tamanhos escolhidos para as imagens. Com um tamanho maior, a informação fica mais fina, porém a imagem fica esparsa. Além disso, com

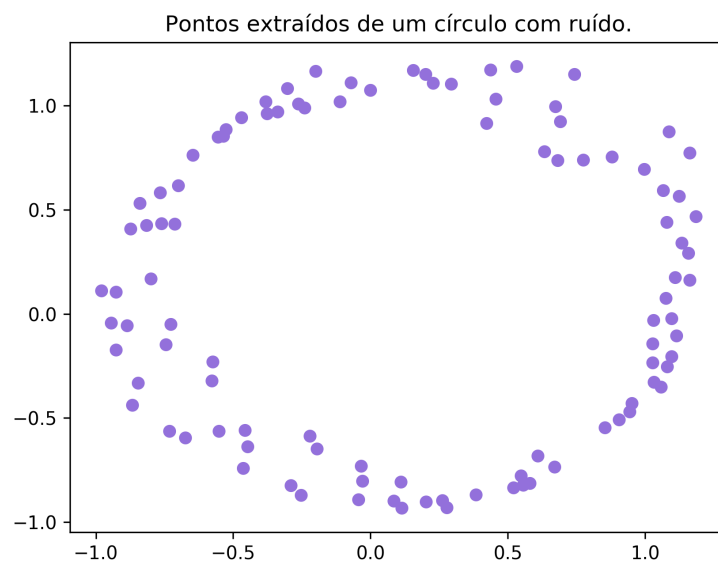


Figura 19 – Pontos extraídos de um círculo com ruídos.

Fonte: Elaborada pelo autor.

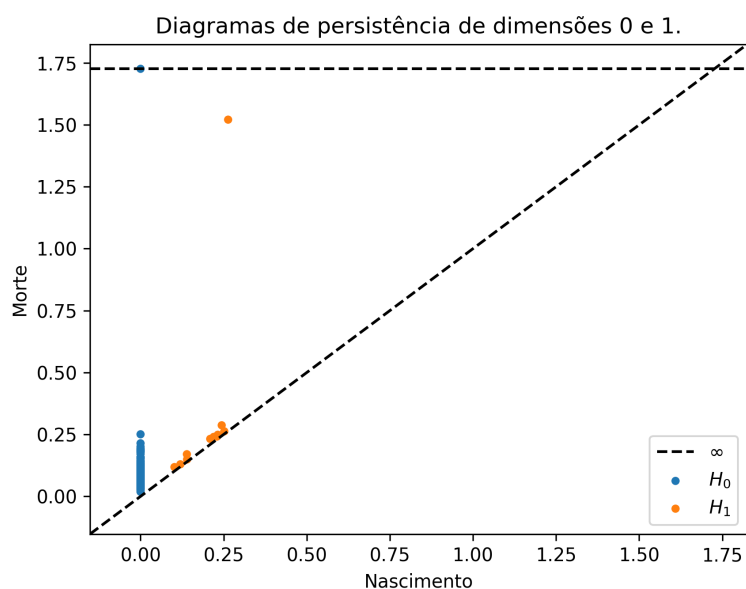


Figura 20 – Diagramas de persistência do círculo X . Em laranja o diagrama de persistência de dimensão 1, em azul o de dimensão 0. A filtração de Vietoris-Rips foi usada para calcular o complexo simplicial.

Fonte: Elaborada pelo autor.

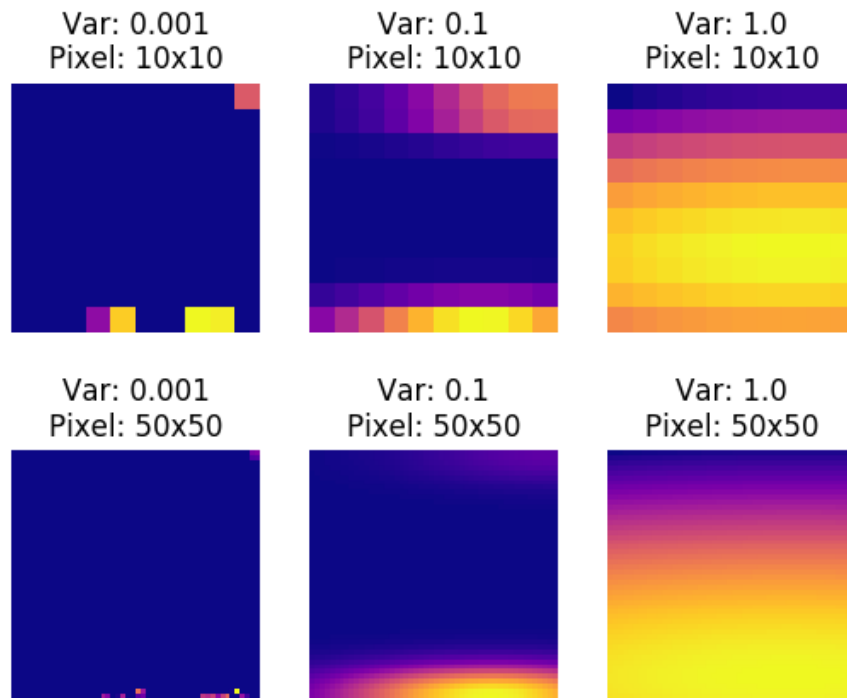


Figura 21 – Seis imagens de persistência do diagrama de dimensão 1 da Figura 20.

Fonte: Elaborada pelo autor.

uma variância mais baixa os pontos ficam mais concentrados, enquanto para valores mais altos há uma troca contínua entre as regiões dos pontos com maior frequência.

Todo o código para gerar o círculo com ruído, calcular os diagramas e imagens se encontram no *Jupyter Notebook* no repositório da dissertação: <<https://github.com/chronchi/dissertacao>> na pasta *jupyter_notebook*.

4.3 Mapper

APLICAÇÕES

Neste capítulo serão descritas algumas aplicações utilizando geradores ótimos e imagens de persistência.

5.1 Geradores ótimos em classificadores de imagens

Utilizando imagens e rótulos associados a elas é possível criar classificadores, algoritmos que decidem os rótulos dada uma imagem. Alguns deles são Redes Neurais (MCCULLOCH; PITTS, 1943), SVM (CORTES; VAPNIK, 1995), Redes Neurais Convolucionais (abreviado por CNN, sigla em inglês) (LECUN *et al.*, 1989) e *Generative Adversarial Networks (GAN)* (GOODFELLOW *et al.*, 2014).

Nesta seção será descrito as redes neurais convolucionais e como obteve-se um classificador de imagens utilizando-as. Além disso, será descrito como outros classificadores foram gerados utilizando informações disponibilizadas pelos geradores ótimos para obter-se um classificador com melhor acurácia do que a rede neural convolucional original.

5.1.1 Redes Neurais Convolucionais (CNN)

O algoritmo de redes neurais artificiais é o precursor da CNN. Um rede neural artificial é uma composição de funções f_n que tem como contra domínio algum \mathbb{R}^m . O seu domínio é dado pela dimensão dos dados disponíveis, por exemplo, se temos uma imagem de tamanho 10x10, a dimensão do domínio é 100. Logo, a rede neural pode ser descrita como uma função $Ann: \mathbb{R}^p \rightarrow \mathbb{R}^m$

$$Ann(x) = f_n(\dots f_2(A_2 * f_1(A_1 * x + b_1) + b_2), \quad (5.1)$$

onde A_i é uma matrix de tamanho arbitrário e $b_i \in \mathbb{R}$. Na Figura 22, temos uma imagem clássica para redes neurais.

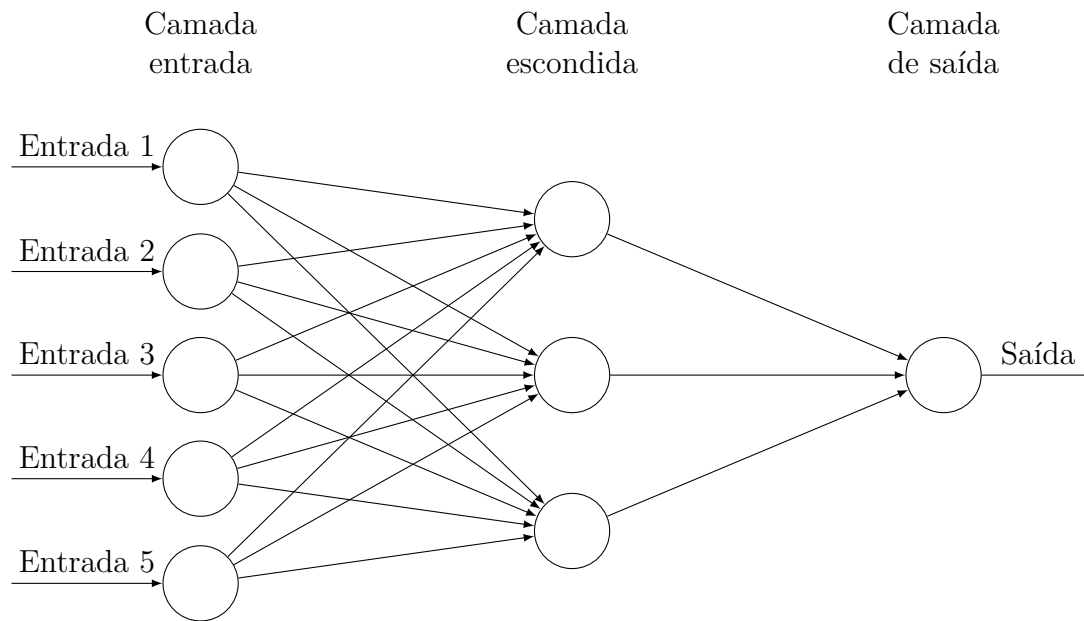


Figura 22 – Esquema de uma rede neural artificial. O número de vértices na camada escondida é determinado pelo tamanho da matriz A_i

5.2 Imagens de persistência aplicadas a proteínas

CONCLUSÃO

REFERÊNCIAS

ADAMS, H.; EMERSON, T.; KIRBY, M.; NEVILLE, R.; PETERSON, C.; SHIPMAN, P.; CHEPUSHTANOVA, S.; HANSON, E.; MOTTA, F.; ZIEGELMEIER, L. Persistence images: A stable vector representation of persistent homology. **Journal of Machine Learning Research**, v. 18, n. 8, p. 1–35, 2017. Disponível em: <<http://jmlr.org/papers/v18/16-337.html>>. Citado nas páginas 53 e 54.

AZUMAYA, G. Corrections and supplementaries to my paper concerning krull-remak-schmidt's theorem. **Nagoya Mathematical Journal**, Cambridge University Press (CUP), v. 1, p. 117–124, jun. 1950. Disponível em: <<https://doi.org/10.1017/s002776300002290x>>. Citado na página 44.

BUBENIK, P. Statistical topological data analysis using persistence landscapes. **Journal of Machine Learning Research**, v. 16, n. 3, p. 77–102, 2015. Disponível em: <<http://jmlr.org/papers/v16/bubenik15a.html>>. Citado na página 53.

CHAZAL, F.; SILVA, V. de; GLISSE, M.; OUDOT, S. **The Structure and Stability of Persistence Modules**. Springer International Publishing, 2016. Disponível em: <<https://doi.org/10.1007/978-3-319-42545-0>>. Citado nas páginas 37 e 44.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, Springer Nature, v. 20, n. 3, p. 273–297, set. 1995. Disponível em: <<https://doi.org/10.1007/bf00994018>>. Citado na página 59.

EDELSBRUNNER, H. **Computational topology : an introduction**. Providence, R.I: American Mathematical Society, 2010. ISBN 0821849255. Citado nas páginas 25, 30 e 32.

EDELSBRUNNER, H.; LETSCHER, D.; ZOMORODIAN, A. Topological persistence and simplification. In: **Proceedings 41st Annual Symposium on Foundations of Computer Science**. IEEE Comput. Soc, 2000. Disponível em: <<https://doi.org/10.1109/sfcs.2000.892133>>. Citado nas páginas 32 e 37.

GOODFELLOW, I. J.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A.; BENGIO, Y. Generative adversarial nets. In: **Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2**. Cambridge, MA, USA: MIT Press, 2014. (NIPS'14), p. 2672–2680. Disponível em: <<http://dl.acm.org/citation.cfm?id=2969033.2969125>>. Citado na página 59.

LECUN, Y.; BOSER, B.; DENKER, J. S.; HENDERSON, D.; HOWARD, R. E.; HUBBARD, W.; JACKEL, L. D. Backpropagation applied to handwritten zip code recognition. **Neural Computation**, MIT Press - Journals, v. 1, n. 4, p. 541–551, dez. 1989. Disponível em: <<https://doi.org/10.1162/neco.1989.1.4.541>>. Citado na página 59.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The Bulletin of Mathematical Biophysics**, Springer Nature, v. 5, n. 4, p.

115–133, dez. 1943. Disponível em: <<https://doi.org/10.1007/bf02478259>>. Citado na página 59.

POINCARÉ, H. Analysis situs. **Journal de l'École Polytechnique**, p. 1–123, 1895. Citado na página 21.

ALGORITMO *STANDARD* E FUNÇÕES AUXILIARES

```

1 # Retorna o valor low(column) descrito no texto.
2 # Exemplo:
3 # a = [1,1,0,0,1,0]
4 # get_low(a) retorna 5, maior índice tal que o valor de a nesse índice é 1.
5 function get_low(column)
6     a = findlast(column .== 1)
7     if a == nothing
8         return 0
9     else
10        return a
11    end
12 end
13
14 # Algoritmo Standard. Input: Matriz bordo P. Output: Matriz reduzida R.
15 function reduce_matrix(boundary)
16     # número de simplexes no complexo simplicial
17     nb_simplex = size(boundary, 2)
18     # matriz R
19     reduced = boundary
20     # matrix V
21     cycles = one(boundary)
22     # Acessa cada coluna para reduzi-la
23     lowest_ones = []
24     for col in 1:nb_simplex
25         # obtém os low(i) para i < col
26         lowest_ones = [get_low(reduced[:,k]) for k in 1:(col-1)]
27         # enquanto tiver algum j_0 tal que low(j_0) = low(col) e low(col) != 0,
28         # reduza a matriz
29         while sum(get_low(reduced[:,col]) .== lowest_ones) != 0 && get_low(
30             reduced[:,col]) != 0
31             # opera a soma nas colunas que possuem o mesmo valor low.
32             for k in 1:length(lowest_ones)
33                 if get_low(reduced[:,k]) == get_low(reduced[:,col])
34                     reduced[:,col] = rem.(reduced[:,col] + reduced[:,k], 2)

```

```
33         cycles[:,col] = rem.(cycles[:,col] + cycles[:,k], 2)
34     end
35 end
36 end
37 end
38 # retorna a matriz reduzida R e a matriz V
39 return reduced, cycles
40 end
41
42 # E por fim a matriz do nosso exemplo
43 P = [0 0 0 0 1 1 0 0 0 0;
44      0 0 0 0 1 0 1 1 0 0;
45      0 0 0 0 0 0 0 1 1 0;
46      0 0 0 0 0 1 1 0 1 0;
47      0 0 0 0 0 0 0 0 0 0;
48      0 0 0 0 0 0 0 0 0 0;
49      0 0 0 0 0 0 0 0 0 1;
50      0 0 0 0 0 0 0 0 0 1;
51      0 0 0 0 0 0 0 0 0 1;
52      0 0 0 0 0 0 0 0 0 0]
53
54 # aplicando o algoritmo de redução
55 R, V = reduce_matrix(P)
```

