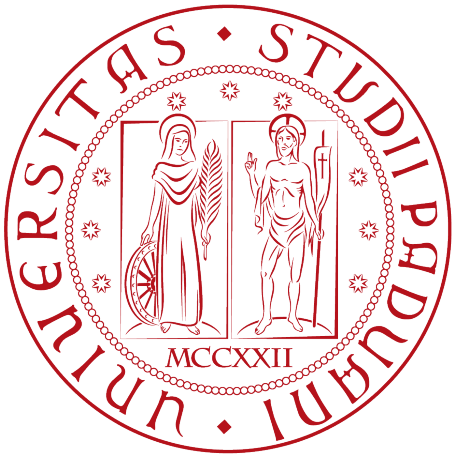
TOPOLOGICAL DATA ANALYSIS

MARCO TARANTINO



Prof. Silvana Bazzoni
Classe di Scienze Naturali
Scuola Galileiana di Studi Superiori

Novembre 2016

Marco Tarantino: *Topological Data Analysis,* Novembre 2016

A Olga.

SOMMARIO

Lo studio dei Big Data è più che mai attuale vista la grande quantità di dati che viene prodotta dai più disparati settori delle scienze applicate e dell'industria, e vi è sempre più necessità di tecniche di analisi di tipo qualitativo, oltre che quantitativo.

Da un lato vi è stato lo sviluppo di strumenti di analisi estremamente generali, come quelle tecniche che insieme vengono chiamate Machine Learning, per i quali esistono pacchetti software estremamente moderni che possono essere usati in ambiente production, oltre che per la ricerca. Dall'altro, l'efficienza dei moderni processori ha consentito l'applicazione di queste tecniche ad ampio spettro su enormi quantità di dati non strutturati.

Un nuovo tipo di problema, tuttavia, è emerso: ora che abbiamo a disposizione tutti questi dati non strutturati, che spesso sono collezioni di punti in spazi vettoriali di dimensione elevata, vorremmo capirne la struttura globale.

Le tecniche attualmente utilizzate tendono a non essere sufficientemente sensibili alla struttura dei dati o a non individuare con accuratezza le loro proprietà geometriche. Per risolvere questo problema sono state sviluppate una serie di tecniche, che insieme vengono chiamate Topological Data Analysis (TDA), alla cui base c'è l'omologia persistente, che noi ci occuperemo di presentare.

L'omologia persistente consente di studiare gli invarianti topologici dei dati *su tutte le scale* allo stesso tempo, e può essere usata da sola per scoprire strutture all'interno dei dati, ad esempio componenti connesse o presenza di buchi (n-dimensionali), o in congiunzione con tecniche di Machine Learning come le (*Support Vector Machines*), mediante la definizione di opportuni kernel nello spazio dei diagrammi persistenti.

RINGRAZIAMENTI

Ringrazio la Prof.ssa Silvana Bazzoni che mi ha guidato e continua ancora a guidarmi con pazienza e dedizione nella scoperta del meraviglioso mondo della Matematica.

INDICE

ELENCO DELLE FIGURE

1 INTRODUZIONE

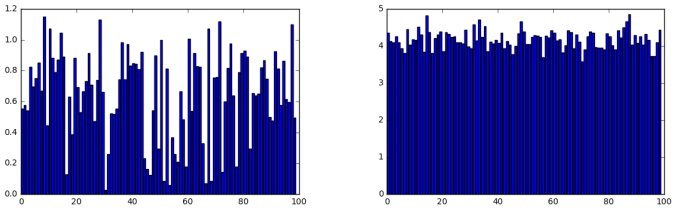
Negli ultimi anni molte branche della scienza e dell'industria si trovano a produrre e analizzare insiemi di dati sempre più grandi e complessi. Questi Big Data sono difficilmente analizzabili usando tecniche standard a causa di problemi sia tecnici che teorici.

Ad esempio, dover processare grandi quantità di dati complessi può avere facilmente un costo troppo elevato in termini di tempo e memoria rispetto alle possibilità della tecnologia odierna. Vi sono poi problemi che si annidano nell'analisi di dati in elevate dimensioni, in cui bisogna tenere conto di alcune particolarità, come il fatto che la distanza fra due punti smette di essere significativa.

Si può visualizzare questa cosa confrontando l'estrazione casuale di punti dal quadrato in dimensione 2 o dall'ipercubo in dimensione 100: estratti due punti x e y dalla distribuzione uniforme sull'ipercubo unitario in $\mathbb{R}^d$, la distanza fra loro è

$$d(x,y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2},$$

allora per la disuguaglianza di Hoeffding la distribuzione di $d(x,y)$ è concentrata attorno al valore medio per dimensioni d elevate. In fig. 1 si può vedere la differenza: i due grafici mostrano l'andamento della distanza di un punto casuale dell'ipercubo da altri 100 punti casuali per $d = 2$ e $d = 100$. Si può notare come all'aumentare della dimensione i valori si schiacciano attorno al valor medio.



(a) $d = 2$ (b) $d = 100$

Figura 1. Distanza fra punti casuali nell'ipercubo

Ci interessa quindi sviluppare tecniche in grado di rivelare informazioni in spazi metrici di dimensione possibilmente elevata, in particolare vorremmo poter riassumere in un numero finito (piccolo) di variabili l'informazione che ci serve.

Utilizzeremo tecniche di topologia algebrica, e le informazioni che queste ci consentiranno di catturare riguardano la *forma* dei dati. In dimensioni elevate è difficile visualizzare un insieme di punti e quindi comprendere se questo abbia una forma interessante, tuttavia utilizzando la topologia è possibile produrre una rappresentazione sintetica di uno spazio complesso mediante una quantità finita di informazioni, e tale che calcolando opportuni invarianti di questa rappresentazione si ricavino informazioni sulla *forma* dello spazio. La topologia consente anche di esprimere mediante quantità precise cosa intendiamo per *forma*.

Prendendo ad esempio la corona circolare in fig. 2, la proprietà che vogliamo essere in grado di catturare è il fatto che essa contiene essenzialmente un solo loop non banale a meno di omotopia, o in maniera equivalente potremmo dire che ha un unico foro.

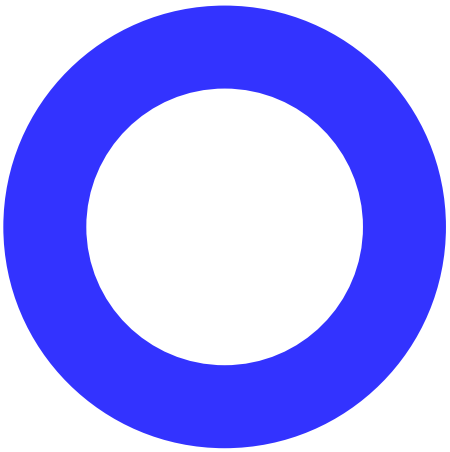


Figura 2. Corona circolare

Vorremmo anche essere in grado di rappresentare sinteticamente uno spazio mediante simplessi, conservando quanta più informazione possibile sulla sua forma. Ad esempio passando dalla circonferenza al quadrato in fig. 3 perdiamo informazioni sulla curvatura, tuttavia le proprietà topologiche sono invariate.

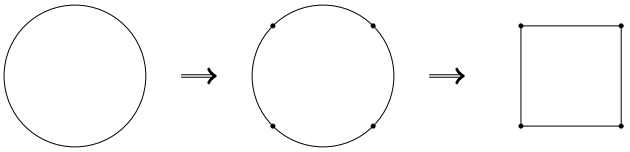


Figura 3. Rappresentazione di una circonferenza

Nel tipo di applicazione che ci interessa, tuttavia, l'oggetto di studio non è una varietà come nei casi precedenti, ma una nube di punti, cioé un sottinsieme finito di $\mathbb{R}^d$ o più in generale uno spazio metrico

finito $X = \{x_1, \ldots, x_n\}$. Anche in questo caso lo spazio può esibire una forma interessante, ad esempio in fig. 4 i dati si dispongono chiaramente in tre gruppi, mentre in fig. 5 hanno chiaramente una forma circolare.

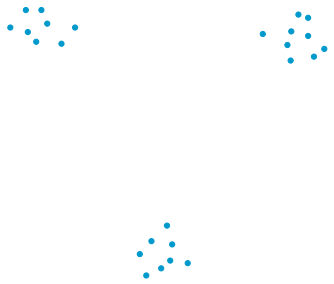


Figura 4. Dati divisi in più cluster

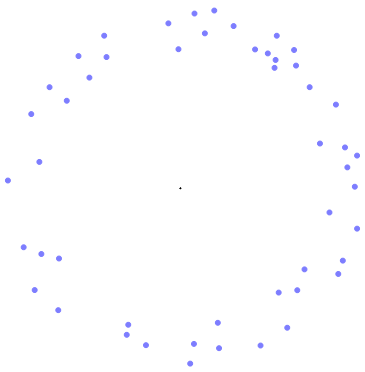


Figura 5. Campionamento da una corona circolare

Come si possono esprimere queste osservazioni mediante tecniche algebriche? Se i nostri dati sono collezionati nell'insieme finito $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$, possiamo considerare

$$\widehat{X}_\varepsilon = \bigcup_{x \in X} B_\varepsilon(x)$$

dove $B_\varepsilon(x)$ è la palla di raggio ε e centro x. Allora per qualche opportuno $\varepsilon > 0$ l'insieme $\widehat{X}_\varepsilon$ associato alla fig. 4 diventa del tipo raffigurato in fig. 6.

Analogamente al variare di ε vediamo gli spazi mostrati in fig. 7.

A questo punto abbiamo a che fare con spazi più o meno noti e, per opportune scelte di ε, si possono calcolare i gruppi di omologia

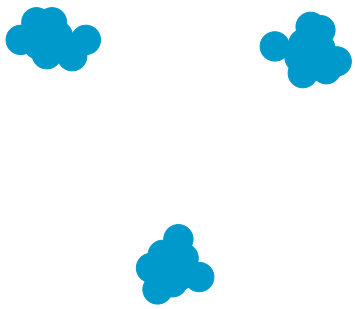


Figura 6. $\widehat{X}_\varepsilon$

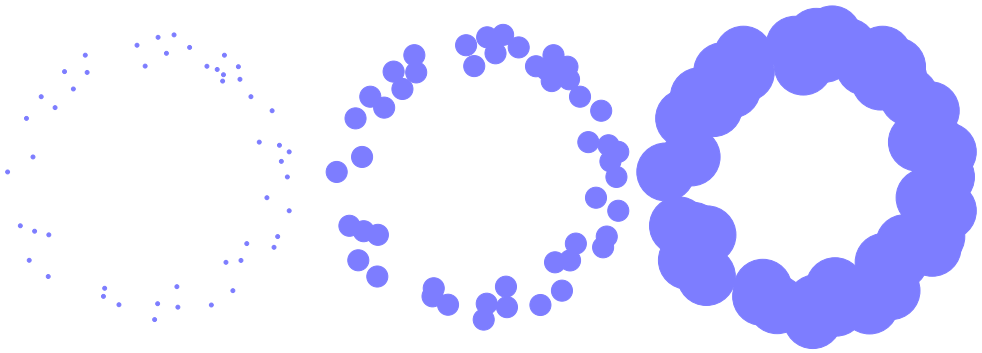


Figura 7. $\widehat{Y}_\varepsilon$ al variare di ε

$H_0(\widehat{X}_\varepsilon)$ e $H_1(\widehat{Y}_\varepsilon)$ che racchiudono rispettivamente l'informazione sul numero di cluster in fig. 4 e di loop in fig. 5.

Tuttavia in entrambi questi casi c'è bisogno di selezionare un parametro di scala ε per cui $\widehat{X}_\varepsilon$ contenga esattamente le informazioni che vogliamo trovare, e questa scelta non sempre è evidente. Ad esempio potrebbe succedere che i dati sono in dimensione talmente alta che non è possibile visualizzarli, oppure proprietà che sono evidenti ad una certa scala scompaiono al variare di ε, mentre ne compaiono di nuove. Quindi siamo alla ricerca di metodi che consentano non solo di dedurre informazioni sulla forma dei dati senza doversi preoccupare della selezione di una scala, ma che riescano anche a mettere in relazione le variazioni di queste proprietà al variare della scala di riferimento. L'insieme delle tecniche che consentono di fare questo e molto altro vanno sotto il nome di Topological Data Analysis (TDA).

Lo sviluppo della TDA è stato iniziato da Carlsson et al. nel 2009 in [3], con l'obiettivo di riuscire a ricavare a partire da un insieme di dati informazioni *qualitative*, *indipendenti dalla scelta di coordinate* e prediligendo una visione d'insieme. Vi sono stati molti progressi sia nella teoria sia nell'applicazione, ad esempio nello sviluppo della teoria della persistenza multidimensionale [6, 5, 1, 4], della fondazione categorica della persistenza [8], dell'utilizzo della persistenza nell'inferenza statistica [2, 12], e la ricerca è tuttora in corso. Alcuni risultati conseguiti utilizzando la TDA sono la determinazione di nuove variabili che influenzano l'attività neurale [16], la classificazione di traiettorie in robotica [14], l'identificazione di nuovi tipi di cancro al seno [13].

Lo strumento principale della TDA è la persistenza omologica, cioé l'assegnazione ad uno spazio metrico finito di una filtrazione di spazi vettoriali $\{V_r\}_{r\in\mathbb{R}}$ in cui gli spazi vettoriali V_r e le mappe fra essi $V_r \to V_s$ per $r < s$ contengano informazioni sulla *forma* dello spazio in esame.

In particolare, gli spazi vettoriali V_r sono gruppi di omologia (e quindi saranno in dicati con $H_q(X_r)$) e il parametro r in genere è preso in $\mathbb{R}^+$ perché identifica una scala di riferimento. Questi gruppi di omologia persistenti vengono visualizzati mediante codici a barre persistenti (fig. 8), dove ogni linea traccia l'evoluzione di una singola *feature* topologica e la lunghezza ne è la durata.

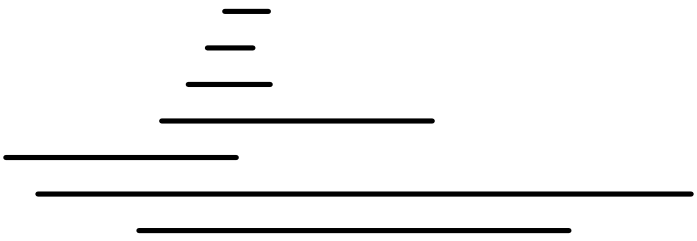


Figura 8. Esempio di codice a barre persistente

La proprietà fondamentale dei gruppi di omologia persistenti è che

essi tengono traccia del legame che vi è fra le proprietà topologiche a diverse scale.

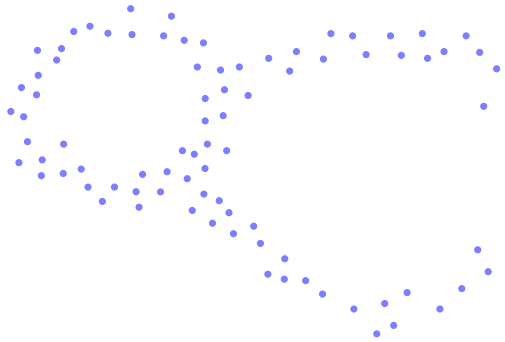
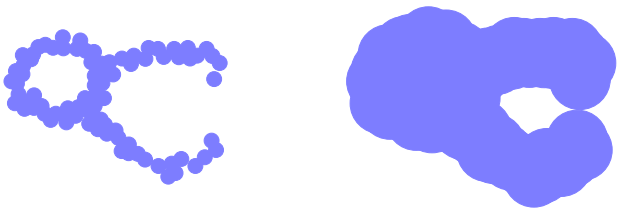


Figura 9. Un doppio anello

Si consideri ad esempio la nube di punti in fig. 9. Al variare di r compaiono due loop come mostrato in fig. 10.



(a) $\widehat{X}_{r_1}$ (b) $\widehat{X}_{r_2}$

Figura 10. Variazione delle proprietà di $\widehat{X}_r$

La presenza dei due loop è rappresentata dall'omologia persistente dal fatto che $H_1(\widehat{X}_{r_1})$ e $H_1(\widehat{X}_{r_2})$ hanno entrambi dimensione 1. Tuttavia, l'omologia persistente consente anche di distinguere i due loop. Infatti, poiché $r_1 < r_2$ l'omologia persistente viene con una mappa $\psi : H_1(\widehat{X}_{r_1}) \to H_1(\widehat{X}_{r_2})$ e il fatto che i due loop sono distinti lo si deduce dal fatto che $\psi = 0$.

Questa proprietà dell'omologia persistente, detta *funtorialità*, è l'elemento essenziale che rende queste tecniche estremamente utili nelle applicazioni.

L'altra proprietà fondamentale dell'omologia persistente è enunciata nel Teorema 1 che garantisce che ogni gruppo di omologia persistente associato ad uno spazio metrico finito può essere rappresentato con un numero finito (idealmente piccolo) di parametri. Questo fatto

consente di visualizzare con pochi parametri un insieme di dati potenzialmente complesso, inoltre lo spazio di questi parametri può essere reso uno spazio metrico completo (come in [12]), aprendo spazio all'utilizzo di tecniche di statistica e machine learning.

Nel capitolo 2 ci occuperemo di sviluppare l'omologia persistente, richiamando velocemente le nozioni necessarie di topologia algebrica e in particolare l'omologia simpliciale, mentre nel capitolo 3 mostreremo alcuni esempi su dati sintetici di codici a barre persistenti.

2

TEORIA

2.1 OMOLOGIA SIMPLICIALE

Fissato un campo k, ad ogni spazio topologico possiamo associare una successione di k-spazi vettoriali $H_i(X;k)$ detti *gruppi di omologia*. Per la precisione si tratta di una successione di funtori $H_*(-;k) : \mathbf{Top} \to \mathbf{Vect}_k$. Nel resto della sezione ci occuperemo di definire in modo operativo questi gruppi.

Esistono diversi modi di calcolare i gruppi di omologia. Il più generale e potente è l'omologia singolare, tuttavia questa risulta scomoda da usare in pratica perché richiede di lavorare con quozienti di spazi vettoriali di dimensione più che numerabile. Per aggirare il problema definiremo soltanto l'omologia simpliciale.

In virtù degli assiomi di Eilenberg-Steenrod [9, 10] le due definizioni sono equivalenti (almeno per gli spazi che andremo a considerare). Per una trattazione più dettagliata si vedano [11] o [15].

Si procederà associando all'insieme di dati X uno spazio topologico (detto complesso simpliciale) per ogni parametro $\varepsilon \in \mathbb{R}^+$ e si definirà l'omologia solo per questi particolari spazi topologici.

Gli spazi $\widehat{X}_\varepsilon$ usati nell'introduzione, sebbene comodi per introdurre un'idea intuitiva di persistenza, non sono l'ambiente naturale in cui lavorare, quindi non verranno più usati nella trattazione formale. Si osservi, però, che l'omologia di $\widehat{X}_\varepsilon$ è equivalente all'omologia del complesso simpliciale di Čech di parametro ε associato a X. Per questioni di comodità, tuttavia, noi lavoreremo principalmente con il complesso di Vietoris-Rips.

2.1.1 *Complessi simpliciali*

I complessi simpliciali sono particolari spazi topologici che hanno una descrizione combinatorica. Dato un insieme di punti $S = \{s_0, \dots, s_n\}$ in $\mathbb{R}^k$ diremo che sono in posizione generale se non sono contenuti in nessun sottospazio affine di dimensione minore di n. Se S è in posizione generale, il suo inviluppo convesso $\sigma(S)$ è detto *(n-)simplesso generato da* S. I punti s_i di S si chiamano *vertici* di $\sigma(S)$. Se $\emptyset \neq T$ è un sottinsieme di S, $\sigma(T)$ è detto *faccia* di $\sigma(S)$.

Con questi ingredienti possiamo dare la seguente

Definizione 1. Un *complesso simpliciale* (finito) K è una famiglia finita di simplessi in uno spazio euclideo tali che:

1. Se $\sigma \in K$ e τ è una faccia di σ, allora $\tau \in K$.
2. Se $\sigma, \tau \in K$, allora il simplesso $\sigma \cap \tau$ è una faccia sia di σ sia di τ.

È chiaro che un complesso simpliciale è determinato essenzialmente da proprietà combinatoriche dell'insieme dei suoi vertici, che motiva la seguente costruzione astratta.

Definizione 2. Un *complesso simpliciale astratto* X è il dato della coppia $(V(X), \Sigma(X))$, dove $V(X)$ è un insieme finito, i cui elementi sono i *vertici* di X, e $\Sigma(X)$ è una famiglia di sottinsiemi non vuoti di $V(X)$, i cui elementi sono detti *simplessi* di X, tale che:

1. se $v \in V(X)$ allora $\{v\} \in \Sigma(X)$, e
2. se $\sigma \in \Sigma(X)$ e $\emptyset \neq \tau \subseteq \sigma$, allora $\tau \in \Sigma(X)$.

Un simplesso $\sigma \in \Sigma(X)$ è detto q-simplesso se $|\#\sigma| = q + 1$. Indichiamo con X_q l'insieme dei q-simplessi di X.

Osservazione 1. Ogni complesso simpliciale K determina un complesso simpliciale astratto $\widehat{K}$ tale che $V(\widehat{K})$ è l'insieme dei vertici dei simplessi di K e un sottinsieme di $V(\widehat{K})$ è in $\Sigma(\widehat{K})$ se e solo se è l'insieme dei vertici di un simplesso di K.

Possiamo anche definire i morfismi $f : X \to Y$ fra complessi simpliciali astratti come le mappe $f_V : V(X) \to V(Y)$ fra i sottostanti insiemi di vertici e tali che $f_V(\sigma) \in \Sigma(Y)$ per ogni $\sigma \in \Sigma(X)$.

(NMDC: inserire qualche disegno, eventualmente un riferimento a qualche testo sugli oggetti simpliciali)

Ad ogni complesso simpliciale astratto X si può associare un complesso simpliciale $|X|$ detto la *realizzazione geometrica* di X (NMDC: aggiungere un riferimento) e tale che i morfismi $f : X \to Y$ siano mandati in mappe continue $|f| : |X| \to |Y|$ in maniera funtoriale, cioé $|g \circ f| = |g| \circ |f|$. Inoltre, ogni complesso simpliciale K è omeomorfo alla realizzazione geometrica del complesso simpliciale astratto $\widehat{K}$ ad esso associato.

2.1.2 *Omologia simpliciale*

Ora definiremo i gruppi di omologia associati a un complesso simpliciale astratto.

Definizione 3. Sia k un gruppo o un campo o un anello e sia $X = (V(X), \Sigma(X))$ un complesso simpliciale astratto, insieme con un ordine totale sull'insieme $V(X)$. Il gruppo dei q-*cicli* di X su k è il k-modulo $C_q(X, k)$ generato dagli elementi

- $[v_0, \ldots, v_q]$ con $v_0, \ldots, v_q \in V(X)$ e tali che $\{v_0, \ldots, v_q\} \in \Sigma(X)$

modulo le seguenti relazioni:

1. $[v_0, \dots, v_q] = 0$ se $v_i = v_j$ per qualche $i \neq j$,
2. $[v_{\sigma(0)}, \dots, v_{\sigma(q)}] = \mathrm{sign}(\sigma)[v_0, \dots, v_q]$ per ogni permutazione σ di $\{0, \dots, q\}$.

Scriveremo $C_*(X, k)$ per indicare tutti i gruppi dei cicli in tutti i gradi, eventualmente sottintendendo il campo k.

Osservazione 2. Nella definizione precedente si è tenuta traccia dell'ordine dei vertici dei complessi simpliciali perché esso è importante ai fini dell'omologia, dunque d'ora in poi consideriamo sempre i complessi simpliciali ordinati.

Lemma 1. Dati due complessi simpliciali astratti X e Y e un morfismo f fra essi, allora f induce un omomorfismo $f_q : C_q(X) \to C_q(Y)$ per ogni $q \in \mathbb{N}$ definito da:

$$\begin{aligned} f_q : C_q(X) &\longrightarrow C_q(Y) \\ [v_0, \dots, v_q] &\mapsto [f(v_0), \dots, f(v_q)]. \end{aligned}$$

Scriveremo $f_* : C_*(X) \to C_*(Y)$ per indicare tutti questi morfismi.

Osservazione 3. Il motivo per cui il gruppo dei q-cicli è stato introdotto mediante generatori e relazioni è che in questo modo è più facile sollevare mappe di complessi simpliciali astratti a omomorfismi fra i gruppi dei cicli. Se avessimo semplicemente definito il gruppo dei q-cicli come il k-modulo generato da X_q sarebbe stato necessario modificare la definizione di f_* in modo che i cicli mappassero correttamente, perdendo notevolmente in eleganza.

Definizione 4. Sia X un complesso simpliciale astratto, allora esistono morfismi $\partial_q^X : C_q(X) \to C_{q-1}(X)$ per ogni $q \in \mathbb{N}_0$ definiti come:

$$\partial_q([v_0, \dots, v_q]) = \sum_{i=0}^{q} (-1)^i [v_0, \dots, \widehat{v}_i, \dots, v_q]$$

dove la notazione $\widehat{v}_i$ significa che l'i-mo elemento non è presente. Scriveremo ∂ invece che ∂^X quando non è necessario specificare il complesso simpliciale sottostante.

Lemma 2. Nelle notazioni precedenti, $\partial_{q-1} \circ \partial_q = 0$.

Osservazione 4. Il lemma precedente ci dice che per ogni complesso simpliciale astratto X i suoi cicli $C_*(X)$ formano un complesso di k-moduli. Allora data una mappa di complessi simpliciali astratti $f : X \to Y$, il suo sollevamento $f_* : C_*(X) \to C_*(Y)$ è un morfismo di

complessi di k-moduli, cioé $f_{q-1} \circ \partial_q^X = \partial_q^Y \circ f_q$. Quest'ultima proprietà può essere espressa dicendo che il seguente diagramma è commutativo (cioé che qualsiasi percorso si segua componendo i morfismi non cambia il risultato della composizione).

$$\begin{array}{ccc} C_q(X) & \xrightarrow{\partial_q^X} & C_{q-1}(X) \\ \downarrow{\scriptstyle f_q} & & \downarrow{\scriptstyle f_{q-1}} \\ C_q(Y) & \xrightarrow{\partial_q^Y} & C_{q-1}(Y) \end{array}$$

Possiamo visualizzare $C_*(X)$ come la seguente successione:

$$\ldots \longrightarrow C_2 \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0$$

a cui per completezza aggiungeremo il morfismo $\partial_0 : C_0 \to 0$.

Definizione 5. Dato un complesso impliciale astratto X, per ogni $q \in \mathbb{N}_0$ vale $\operatorname{Im} \partial_q \subseteq \operatorname{Ker} \partial_{q-1}$. Definiamo il q-mo gruppo di omologia di X come il quoziente

$$H_q(X) = \frac{\operatorname{Ker} \partial_q}{\operatorname{Im} \partial_{q+1}}.$$

Il seguente esempio serve a mostrare intuitivamente come l'omologia è collegata alle proprietà geometriche di un complesso simpliciale. Consideriamo il complesso simpliciale astratto $X = (V(X), \Sigma(X))$ con:

$$V(X) = (A, B, C, D, E)$$
$$\Sigma(X) = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{A, B\}, \{C, D\}, \{D, E\}, \{C, E\}\}$$

rappresentato in figura fig. 11. La scrittura $(A, \ldots, E)$ indica che l'ordine è $A < \cdots < E$.

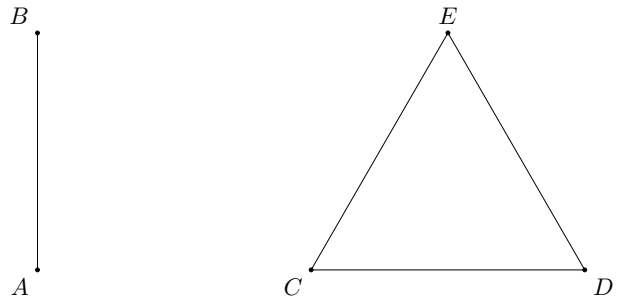


Figura 11. Un complesso simpliciale

Allora C_0 è lo spazio vettoriale generato da $[A], [B], [C], [D], [E]$ e C_1 da $[A, B], [C, D], [D, E], [C, E]$. Applicando il differenziale ∂_1 ai generatori di C_1 otteniamo:

$$\begin{aligned}\partial_1([A, B]) &= [B] - [A]\\ \partial_1([C, D]) &= [D] - [C]\\ \partial_1([D, E]) &= [E] - [D]\\ \partial_1([C, E]) &= [E] - [C]\end{aligned}$$

Per semplicità, si può anche scrivere ∂_1 in forma matriciale come

$$\begin{array}{c} \\ [A]\\ [B]\\ [C]\\ [D]\\ [E]\end{array}\begin{array}{c}\begin{array}{cccc}[A, B] & [C, D] & [D, E] & [C, E]\end{array}\\ \begin{pmatrix} -1 & 0 & 0 & 0\\ 1 & 0 & 0 & 0\\ 0 & -1 & 0 & -1\\ 0 & 1 & -1 & 0\\ 0 & 0 & 1 & 1\end{pmatrix}\end{array},$$

da cui risulta evidente che $\operatorname{Im}\partial_1$ ha dimensione 3, quindi $H_0(X) = \operatorname{Ker}\partial_0/\operatorname{Im}\partial_1$ è un k-spazio di dimensione 2. Si osservi che X ha esattamente 2 componenti connesse.

Analogamente $\operatorname{Ker}\partial_1$ ha dimensione 1 e C_2 è lo spazio banale, quindi $H_1(X)$ ha dimensione 1, che rappresenta il fatto che c'è un loop nella spezzata $[C, D] + [D, E] + [E, C]$ (infatti $\operatorname{Ker}\partial_1$ è generato proprio da questo vettore).

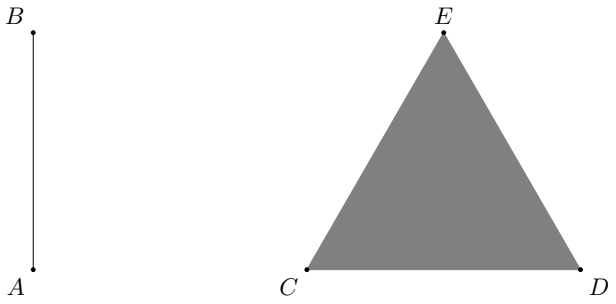


Figura 12. Il complesso X con l'aggiunta di un 2-simplesso

Se aggiungiamo a X il simplesso $\{C, D, E\}$ come in fig. 12 abbiamo che C_2 è generato da $[C, D, E]$ e $\partial_2([C, D, E]) = [D, E] - [C, E] + [C, D] = [C, D] + [D, E] + [E, C]$. Quindi chiudendo il loop con un simplesso abbiamo che $\operatorname{Im}\partial_2 = \operatorname{Ker}\partial_1$ e $H_1(X)$ è lo spazio banale.

Osservazione 5. L'assegnazione di $H_*(X)$ a un complesso simpliciale astratto X è in realtà un funtore, cioé per ogni morfismo di complessi simpliciali astratti $f : X \to Y$ esiste un omomorfismo di k-spazi vettoriali $f_* : H_*(X) \to H_*(Y)$ (e in realtà f_* è il morfismo indotto da $f_* : C_*(X) \to C_*(Y)$) e $(g \circ f)_* = g_* \circ f_*$.

2.2 OMOLOGIA PERSISTENTE

Nella sezione precedente è stata definita l'omologia simpliciale e si è visto con alcuni esempi come questa sia collegata a proprietà topologiche del complesso simpliciale in analisi. Quello che resta per poter usare l'omologia per studiare una nube di punti è un modo di costruire un simplesso simpliciale a partire dai dati, il che essenzialmente si riduce a studiare i dati ad una certa scala di grandezza. Tuttavia, non è necessariamente ovvio come fare e, come si è visto nell'introduzione, talvolta non esiste una scelta di scala che consenta di catturare tutte le proprietà (omologiche) dei nostri dati.

Per questo motivo si introduce il concetto di persistenza, che è un modo di controllare diverse scale contemporaneamente e le relazioni tra queste.

Definizione 6. Un *insieme persistente* è una famiglia di insiemi $\{X_r\}_{\mathbb{R}}$ indicizzata da $\mathbb{R}$ e tale che per ogni coppia di numeri reali $r < s$ esiste una funzione $f_{s,r} : X_r \to X_s$ e queste funzioni sono tali che se $r < s < t$ allora $f_{t,r} = f_{t,s} \circ f_{s,r}$. Più in generale un oggetto persistente in una categoria $\mathcal{C}$ è un funtore $(R, \leqslant) \to \mathcal{C}$, quindi si può parlare di *spazi vettoriali persistenti*, *spazi topologici persistenti*, *complessi simpliciali persistenti*, ecc.

A questo punto è possibile associare un complesso simpliciale astratto ai nostri dati nel seguente modo.

Definizione 7. Sia X uno spazio metrico finito. Fissato un numero reale positivo r il *complesso di Vietoris-Rips* di X è un complesso simpliciale astratto $VR(X, r)$ così definito:

- l'insieme dei vertici di $VR(X, r)$ è X,
- la collezione $\{x_0, \dots, x_n\} \subseteq X$ è un simplesso di $VR(X, r)$ se e solo se

$$d(x_i, x_j) \leqslant r \text{ per tutti gli } i, j \in \{0, \dots, n\}.$$

Osservazione 6. Se $r < s$ esiste un'ovvia iniezione $VR(X, r) \to VR(X, s)$, siccome gli insiemi dei vertici coincidono e ogni simplesso di $VR(X, r)$ è anche un simplesso di $VR(X, s)$, quindi la famiglia $\{VR(X, r)\}_{r \in \mathbb{R}}$ forma un complesso simpliciale persistente.

Ad esempio si consideri un esagono regolare come in fig. 13. Se R è la misura del lato dell'esagono, le tre figure rappresentano $VR(X, r)$ rispettivamente con $0 \leqslant r < R$, $R \leqslant r < \sqrt{3}R$, e $r \geqslant \sqrt{3}R$. I complessi di Vietoris-Rips sono elementari nelle prime due figure. Si osservi, invece, che nella terza vi sono 8 2-simplessi, sebbene intuitivamente sia sufficiente usarne 4 per vedere che lo spazio diventa banale a quella scala.

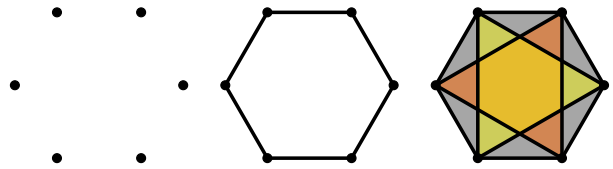


Figura 13. Esempio di complesso di Vietoris-Rips

A questo punto applicando $H_*(-,k)$ a $\{VR(X,r)\}_{\mathbb{R}}$ otteniamo una famiglia di spazi vettoriali $\{H_*(VR(X,r),k)\}$ che insieme alle mappe $H_*(f_{s,r},k) : H_*(VR(X,r)) \to H_*(VR(X,s))$ formano uno spazio vettoriale persistente. Osserviamo inoltre che essendo partiti da uno spazio metrico finito X ci saranno solo un numero finito di complessi di Vietoris-Rips al variare di r.

Gli spazi $\{H_*(VR(X,r),k)\}_{r\in\mathbb{R}}$ racchiudono l'informazione topologica dello spazio metrico X, tuttavia questa definizione astratta potrebbero sembrare di difficile utilizzo: assumendo di poter calcolare tutti gli $H_q(VR(X,r),k)$ per tutti gli r e per piccoli valori di q, la domanda successiva diventa cosa si può dire delle interazioni fra i vari complessi su scale diverse?

In linea di principio dobbiamo considerare per ogni $f_{s,r} : VR(X,r) \to VR(X,s)$ non banale l'associato morfismo di spazi vettoriali $H_q(f_{s,r}) : H_q(VR(X,r)) \to H_q(VR(X,s))$ e tenere traccia delle dimensioni degli spazi $H_q(VR(X,r))$, $H_q(VR(X,s))$, e $\operatorname{Im} H_q(f_{s,r})$.

Al fine di poter meglio comprendere come sono fatti i gruppi di omologia persistenti, studiamo meglio gli spazi vettoriali persistenti. D'ora in poi per semplicità diremo che V è uno spazio vettoriale persistente invece che indicare esplicitamente tutta la famiglia $\{V_r\}$. Qualora fosse necessario esplicitare i morfismi $f_{s,r} : V_r \to V_s$ li indicheremo con la notazione (V,f).

Definizione 8. Sia $I \subseteq \mathbb{R}$. Diremo che I è un intervallo di $\mathbb{R}$ se dati comunque $r,s \in I$ tali che $r < s$ e t tale che $r < t < s$, allora anche $t \in I$. Definiamo *moduli intervallo* gli spazi vettoriali persistenti k_I, dove I è un intervallo di $\mathbb{R}$, tali che $k_I(r) = k$ se $r \in I$ e 0 altrimenti.

Definizione 9. Si definisce *somma diretta* di due spazi vettoriali persistenti (V,f^V) e (W,f^W) lo spazio vettoriale persistente U definito come $U_r = V_r \oplus W_r$ per ciascun r e $f^U_{s,r} = f^V_{s,r} \oplus f^W_{s,r}$, e lo si indica come $(V \oplus W, f^V \oplus f^W)$.

Poiché i moduli intervallo sono associati a intervalli di $\mathbb{R}$ li si può rappresentare come barre lunghe quanto l'intervallo ad essi associati (fig. 14a), e similmente una somma diretta finita di moduli intervallo $k_{I_1} \oplus k_{I_2} \oplus \cdots \oplus k_{I_n}$ può essere rappresentata con quelli che sono detti

(a) k_I (b) $\bigoplus_{I \in \mathcal{D}} k_I$

Figura 14. Esempi di codici a barre

codici a barre (fig. 14b). Il seguente teorema ci dice che in realtà tutti gli spazi vettoriali persistenti si posso rappresentare in questo modo.

Teorema 1 (Decomposizione degli spazi vettoriali persistenti [7])**.** Sia (V, f) uno spazio vettoriale persistente tale che V_r ha dimensione finita per ogni $r \in \mathbb{R}$, allora esiste un multi-insieme $\mathcal{D}$ di intervalli, cioé una famiglia di intervalli con ripetizioni, tale che

$$V \cong \bigoplus_{I \in \mathcal{D}} k_I.$$

2.2.1 *Codice a barre persistente di una nube di punti*

Con la teoria sviluppata finora è possibile, a partire da una collezione di dati sotto forma di nube di punti, costruire un codice a barre associato alla sua omologia persistente. Sia X una nube di punti, cioé un sottinsieme finito di $\mathbb{R}^n$, possiamo costruire una filtrazione di complessi simpliciali astratti

$$VR(X, r_0) \hookrightarrow VR(X, r_1) \hookrightarrow VR(X, r_2) \hookrightarrow \dots$$

e applicarvi $H_q(-, k)$ ottenendo

$$H_q(VR(X, r_0)) \to H_q(VR(X, r_1)) \to H_q(VR(X, r_2)) \to \dots$$

Il codice a barre persistente di X allora è il codice a barre associato a $H_q(VR(X, -))$ mediante il Teorema 1. Le barre lunghe di questo codice a barre sono considerate delle proprietà topologiche robuste di X perché restano su più scale, mentre barre corte sono considerate rumore. Ad esempio nella **??** mostrata nell'introduzione il codice a barre persistente di H_0 presenta una lunga barra, il che significa che l'insieme da un certo punto in poi è connesso, mentre il codice a barre di H_1 contiene una barra che compare intorno alla distanza tipica fra due punti vicini e scompare intorno a $\sqrt{3}$ volte il raggio, a rappresentare il fatto che l'insieme ha essenzialmente un buco. Il fatto che le barre corrispondenti ai loop in H_1 scompaiano a circa $\sqrt{3}$

volte il raggio lo si intuisce dalla fig. 13 disegnando il codice a barre di H_1 dell'esagono come si vede in fig. 15.

Tuttavia, è possibile che un codice a barre persistente non mostri questa dualità fra barre lunghe e corte, ma presenti anche barre di lunghezza diversa.

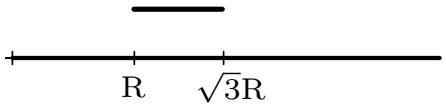


Figura 15. Codice a barre persistente di H_1 per un esagono regolare di lato R

Per chiarezza esplicitiamo il procedimento seguendo [8].

Definizione 10 (Persistenza di una nube di punti)**.** Il processo di persistenza di una nube di punti è l'insieme delle seguenti operazioni.

1. Sia X una nube di punti, cioé un sottinsieme finito di $\mathbb{R}^n$.
2. A partire da X si costruisce una famiglia crescente di complessi simpliciali (astratti) $\{X_r\}$ (ad esempio con il complesso di Vietoris-Rips) la cui omologia H_q produce per ogni $q \geqslant 0$ uno spazio vettoriale persistente:
$$H_q(X_{r_0}) \to H_q(X_{r_1}) \to H_q(X_{r_2}) \to \cdots$$
3. Il Teorema 1 fornisce un multi insieme di intervalli, che può essere visualizzato come codice a barre persistente.

Utilizzando i codici a barre persistenti si riescono a individuare buchi n-dimensionali e a tracciarne l'evoluzione su più scale, o più in generale si possono distinguere due nubi di punti sulla base di proprietà omologiche dei complessi persistenti associati. Nel capitolo 3 mostreremo come è possibile utilizzare l'omologia persistente per fare questo tipo di analisi.

2.3 PERSISTENZA FUNZIONALE

Nella sezione 2.2 si è visto come calcolare l'omologia persistente di una nube di punti, tuttavia non tutte le forme di enteresse vengono evidenziate dai codici a barre persistenti in questo modo. Si consideri ad esempio la fig. 16 in cui sono presenti tre rami che si distaccano da un nucleo centrale.

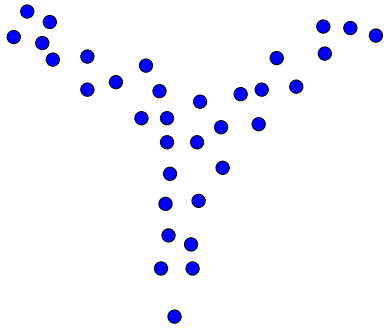


Figura 16. Esempio di dati disposti in rami

È possibile costruire un complesso simpliciale persistente la cui omologia evidenzi questa ramificazione? La risposta è fortunatamente positiva. Di seguito mostreremo la procedura più generale detta persistenza funtoriale, e vedremo come il caso in esame segue come suo caso particolare.

Sia $f : X \to \mathbb{R}$ una funzione reale definita su uno spazio metrico finito X. Allora per ogni $r \in \mathbb{R}$ possiamo considerare le preimmagini $X_r = f^{\leftarrow}([-\infty, r])$. Fissato un parametro $0 < \rho \in \mathbb{R}$ una filtrazione di complessi di Vietoris-Rips

$$VR(X_{r_0}, \rho) \hookrightarrow VR(X_{r_1}, \rho) \hookrightarrow VR(X_{r_2}, \rho) \hookrightarrow \cdots$$

e, applicando $H_q(-)$ a questi complessi, si ottiene lo spazio vettoriale persistente

$$H_q(VR(X_{r_0}, \rho)) \to H_q(VR(X_{r_1}, \rho)) \to H_q(VR(X_{r_2}, \rho)) \to \cdots$$

Infine, anche in questo caso il teorema 1 ci consente di costruire un codice a barre persistente.

Per rendere la persistenza funzionale sensibile a forme come quella in fig. 16 si può usare una funzione che misuri l'eccentricità di un punto $x \in X$ come

$$e_p(x) = \Big(\sum_{y \in X} d(x, y)^p \Big)^{\frac{1}{p}}$$

o, in caso $p = \infty$, $e_\infty(x) = \max_{y \in X} d(x, y)$. Poiché queste funzioni assumono valori alti per punti distanti dal "centro", prendiamo la rinormalizzazione

$$\hat{e}_p(x) = \frac{e_p^{max} - e_p(x)}{e_p^{max} - e_p^{min}}$$

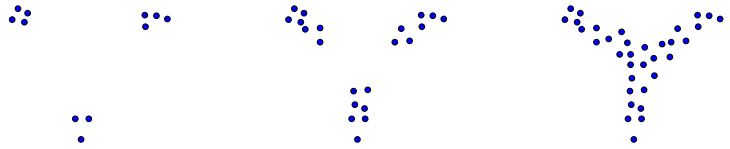


Figura 17. X_r al variare di $r \in [0, 1]$

che assume valori in $[0, 1]$ e raggiunge entrambi gli estremi. Inoltre i punti distanti dal centro adesso hanno valori bassi, come si può vedere in fig. 17.

Il codice a barre persistente corrispondente è mostrato in fig. 18.

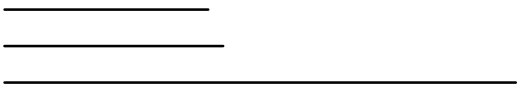


Figura 18. Persistenza funzionale

2.4 PERSISTENZA MULTIDIMENSIONALE

Nell'ultimo esempio nella sezione 2.3 si può notare che abbiamo costruito una filtrazione di complessi simpliciali $(VR(X_r, \rho), \phi)$ al variare di $r \in [0, 1]$, mantenendo però ρ costante.

Tuttavia, sarebbe altrettanto valido fissare un parametro r e costruire la filtrazione al variare di ρ e calcolare l'omologia di questa. Più in generale, data qualunque funzione $f : X \to \mathbb{R}$ è possibile considerare le filtrazioni $VR(f^{\leftarrow}(-\infty, r], \rho)$ per valori fissati di r o ρ e calcolare il codice a barre persistente, tuttavia questo processo va contro il motivo iniziale per cui era stata introdotta l'omologia persistente: evitare di scegliere la scala di grandezza.

Si osservi che se $r < s$ e $\rho < \sigma$, allora $VR(f^{\leftarrow}(-\infty, r], \rho) \hookrightarrow VR(f^{\leftarrow}(-\infty, s], \sigma)$ induce una mappa $H_q(VR(f^{\leftarrow}(-\infty, r], \rho)) \to H_q(VR(f^{\leftarrow}(-\infty, s], \sigma))$, quindi abbiamo a che fare essenzialmente con una filtrazione a più parametri, detta *multifiltrazione* (vd. definizione 11). Lo studio dell'omologia di una multifiltrazione è detto *persistenza multidimensionale* di cui mostreremo brevemente i principali risultati seguendo [5]. Anticipiamo già che non è possibile trovare risultati come il Teorema 1 che garantiscono l'esistenza di invarianti completi.

L'oggetto di studio della persistenza multidimensionale sono i moduli persistenti indicizzati da un insieme con un ordine quasi-parziale.[1]

Dati $u, v \in \mathbb{N}^n$, diciamo che $u \lesssim v$ se $u_i \leqslant v_i$ per ogni $1 \leqslant i \leqslant n$.

Un *monomio* in $x_1, \ldots, x_n$ è un prodotto della forma

$$x_1^{v_1} \cdot x_2^{v_2} \cdots x_n^{v_n}$$

con $v_i \in \mathbb{N}$, che denoteremo x^v con $v = (v_1, \ldots, v_n) \in \mathbb{N}^n$). Un *polinomio* in $x_1, \ldots, x_n$ a coefficienti in un campo k è una combinazione lineare finita $\sum_v \alpha_v x^v$ con $\alpha_v \in k$. L'insieme di questi polinomi è indicato con $k[x_1, \ldots, x_n]$ ed ha un'ovvia struttura di anello.

Un *anello* n*-graduato* è un anello R con una decomposizione in gruppi abeliani $R \cong \oplus_v R_v$, $v \in \mathbb{N}^n$ tale che $R_u \cdot R_v \subseteq R_{u+v}$. L'anello dei polinomi $A_n = k[x_1, \ldots, x_n]$ è un anello n-graduato da $k[x_1, \ldots, x_n] \cong \oplus_v A_v$ dove $A_v = kx^v$, $v \in \mathbb{N}^n$.

Un *modulo* n*-graduato* su un anello n-graduato R è un R-modulo M insieme a una decomposizione $M \cong \oplus_v M_v$, $v \in \mathbb{N}^n$ tale che $R_u \cdot M_v \subseteq M_{u+v}$.

Definizione 11. Dato uno spazio X, una *multifiltrazione* $\{X_v\}_{v \in \mathbb{N}^n}$ è una collezione di sottinsiemi $X_v \subseteq X$ tali che $X_u \subseteq X_v$ per ogni $u \lesssim v$ e tali che il diagramma

$$\begin{array}{ccc} X_u & \longrightarrow & X_{v_1} \\ \downarrow & & \downarrow \\ X_{v_2} & \longrightarrow & X_w \end{array}$$

commuta per ogni $u \lesssim v_1, v_2 \lesssim w$.

I complessi di Vietoris-Rips $VR(f^{\leftarrow}(-\infty, r_i], \rho_j)$ dalla sezione 2.3 sono un esempio di multifiltrazione.

Definizione 12. Un *modulo persistente* M è una famiglia di k-moduli $\{M_v\}_{v \in \mathbb{N}^n}$ insieme con morfismi di moduli $f_{v,u} : M_u \to M_v$ per tutti gli $u \lesssim v$ e tali che $f_{w,v} \circ f_{v,u} = f_{w,u}$ ogniqualvolta $u \lesssim v \lesssim w$.

L'omologia di una multifiltrazione di complessi simpliciali è un modulo persistente. Ad ogni modulo persistente può essere associata una struttura di modulo n-graduato nel seguente modo:

Definizione 13. Dato un modulo persistente M, definiamo il seguente modulo n-graduato su $A_n = k[x_1, \ldots, x_n]$

$$\alpha(M) = \bigoplus_v M_v,$$

dove la struttura di k-modulo è data dalla somma diretta e $x^{v-u} : M_u \to M_v$ è $f_{u,v}$ ogniqualvolta $u \lesssim v$.

1 Si confronti con definizione 6, dove l'insieme degli indici è totalmente ordinato, $\mathbb{R}$

Il [5, Teorema 1], che riportiamo di seguito, ci dice che questa assegnazione è un'equivalenza di categorie.

Teorema 2. La corrispondenza α nella definizione 13 è un'equivalenza di categorie fra la categoria dei moduli persistenti finiti su k e la categoria dei moduli n-graduati finitamente generati su $A_n = k[x_1, \dots, x_n]$.

Quindi il problema della classificazione dei moduli persistenti si riconduce alla classificazione dei moduli n-graduati finitamente generati. Tuttavia i moduli n-graduati con $n > 1$ sono sostanzialmente differenti dal caso $n = 1$ della persistenza omologica, in particolare Carlsson e Zomorodian dimostrano sempre in [5] che non esiste un invariante completo, discreto e indipendente da k che classifichi i moduli n-graduati.

Quello che si può trovare è un invariante parziale che sia discreto e indipendente da k, detto invariante di rango che definiamo di seguito.

Definizione 14. Sia $\dot{\mathbb{N}} = \mathbb{N} \cup \{\infty\}$. Sia $\mathbb{D}^n \subset \mathbb{N}^n \prod \dot{\mathbb{N}}^n$ la sovradiagonale, cioé $\mathbb{D}^n = \{(u, v) | u \in \mathbb{N}^n, v \in \dot{\mathbb{N}}^n, u \lesssim v\}$. Per $(u, v), (u', v') \in \mathbb{D}^n$ definiamo $(u, v) \preceq (u', v')$ se $u \lesssim u'$ e $v \lesssim v'$.

Definizione 15 (Invariante di rango ρ_M)**.** Se M è un A_n-modulo n-graduato finitamente generato, definiamo $\rho_M : \mathbb{D}^n \to \mathbb{N}$ la mappa $\rho_M(u, v) = \mathrm{rank}(x^{v-u} : M_u \to M_v)$.

Osservazione 7. ρ_M rispetta l'ordine dato da $\preceq$, cioé se $(u, v) \preceq (u', v')$, allora $\rho_M(u, v) \leqslant \rho_M(u', v')$.

Nel caso di una multifiltrazione di complessi simpliciali possiamo dare una definizione dell'invariante di rango direttamente in termini dei complessi.

Definizione 16. Se $X = \{X_v\}_{v \in \mathbb{N}^n}$ è una multifiltrazione di complessi simpliciali, l'invariante di rango di $H_i(X_v, k)$ si può scrivere come

$$\rho_{X,i}(u, v) = \mathrm{rank}(H_i(X_u, k) \to H_i(X_v, k))$$

L'invariante di rango si restringe a un invariante completo nel caso di moduli 1-graduati, tuttavia anche nel caso di moduli n-graduati può essere utilizzato per trovare proprietà topologiche persistenti cercando punti $(u, v) \in \mathbb{D}^n$ che sia lontani dalla diagonale. La lontananza dalla diagonale significa che la proprietà è persistente, ed è l'equivalente di cercare barre lunghe nell'omologia persistente.

3

ESEMPI E APPLICAZIONI

NMDC: aggiugnere esempi come statistical machine learning usando persistence diagrams

3.1 SEZIONE 1 CAPITOLO 3

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

BIBLIOGRAFIA

[1] Adcock, A., and Carlsson, G. The Ring of Algebraic Functions on Persistence Bar Codes. 1–21.

[2] Bubenik, P. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research 16* (2015), 25.

[3] Carlsson, G. *Topology and data*, vol. 46. 2009.

[4] Carlsson, G., Singh, G., and Zomorodian, A. Computing multidimensional persistence. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2009).

[5] Carlsson, G., and Zomorodian, A. The theory of multidimensional persistence. *Discrete and Computational Geometry 42*, 1 (2009), 71–93.

[6] Cerri, A., Di Fabio, B., Ferri, M., Frosini, P., and Landi, C. Betti numbers in multidimensional persistent homology are stable functions. *Mathematical Methods in the Applied Sciences 36*, 12 (2013), 1543–1557.

[7] Crawley-Boevey, W. Decomposition of pointwise finite-dimensional persistence modules. 1–7.

[8] Curry, J. M., and Mar, A. T. Topological Data Analysis and Cosheaves. 1–30.

[9] Eilenberg, S., and Steenrod, N. Axiomatic approach to homology theory. *... of the National Academy of Sciences of ...* (1945).

[10] Eilenberg, S., and Steenrod, N. E. *Foundations of algebraic topology*.

[11] Hatcher, A. Algebraic Topology.

[12] Kwitt, R., Huber, S., Niethammer, M., Lin, W., and Bauer, U. Statistical Topological Data Analysis - A Kernel Perspective. *Advances in Neural Information Processing Systems* (2015), 3070–3078.

[13] Lum, P. Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J., and Carlsson, G. Extracting insights from the shape of complex data using topology. *Scientific reports 3* (2013), 1236.

[14] Pokorny, F. T., Hawasly, M., and Ramamoorthy, S. Multiscale Topological Trajectory Classification with Persistent Homology. *Robotics: Science and Systems* (2014).

[15] Rotman, J. J. *An Introduction to Algebraic Topology*, vol. 119 of *Graduate Texts in Mathematics*. Springer New York, New York, NY, 1988.

[16] Spreemann, G., Dunn, B., Botnan, M. B., and Baas, N. A. Using persistent homology to reveal hidden information in neural data.