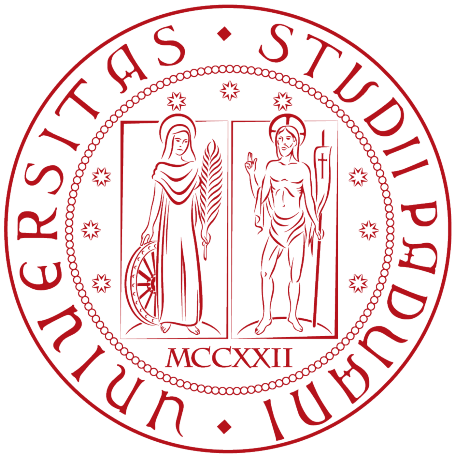
TOPOLOGICAL DATA ANALYSIS

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 di tecniche per analizzare grandi dataset 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 sono quotidianamente progressi sia teorici che pratici.

Da un lato vi è stato lo sviluppo di strumenti di analisi estremamente generali, come quelle tecniche che insieme vengono chiamate Machine Learning (ML), 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.

Vi sono già alcune tecniche sviluppate a questo scopo, come PCA (*Principal Component Analysis*), che tuttavia possono non essere sufficientemente sensibili alla struttura dei dati, oppure i dati possono essere in un formato che difficilmente può essere analizzato usando queste tecniche, ad esempio anziché punti di spazi vettoriali potremmo avere semplicemente una funzione distanza fra i punti del campione che stiamo analizzando, e da questa metrica vorremmo ricostruire la struttura dei dati. (NMDC: add picture)

Per risolvere questo problema sono state sviluppate una serie di tecniche, che insieme vengono chiamate Topological Data Analysis (TDA). Di queste noi ci occuperemo di presentare l'omologia persistente e l'algoritmo Mapper. La prima si occupa 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 ML come SVM (*Support Vector Machines*), mediante la definizione di opportuni kernel nello spazio dei persistence diagrams. Il secondo produce un riassunto topologico dei dati sotto forma di grafo e può essere usato da solo o per preprocessare i dati prima di procedere con una pipeline PCA.

RINGRAZIAMENTI

Ringrazio la Professoressa 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 TEORIA

1.1 INTRODUZIONE

La TDA si prefigge come obiettivo quello di trovare strutture complesse in un insieme di dati. Finora sono stati conseguiti risultati come la determinazione di nuove variabili che influenzano l'attività neurale [7], la classificazione di traiettorie in robotica [5], l'identificazione di nuovi tipi di cancro al seno [4], e molti altri.

La novità della TDA sta nel provare a catturare la *forma* dei dati e, in questa, cercare proprietà topologiche interessanti che costituiscano un segnale anziché un rumore.

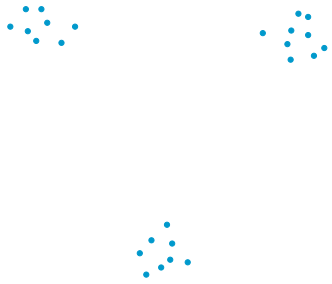


Figura 1. Dati divisi in più cluster

Ad esempio, consideriamo un insieme di dati come in fig. 1. È chiaro a chi osserva che vi sono tre gruppi di punti, tuttavia la formalizzazione matematica di questa osservazione non è immediata.

Il modo usato dalla TDA è l'*omologia persistente*, di cui diamo un'introduzione informale per poi riprenderlo in (NMDC: inserire riferimento al capitolo). Sia

$$X = \{x_1, \ldots, x_n\}$$

il nostro insieme di dati. Consideriamo per $\varepsilon > 0$ l'insieme

$$\widehat{X}_\varepsilon = \bigcup_{0 \leqslant i \leqslant n} B(x_i, \varepsilon)$$

dove $B(x_i, \varepsilon)$ è la palla di centro x_i e raggio ε. Osserviamo che esiste un intervallo di valori $a \leqslant \varepsilon \leqslant b$ per cui $\widehat{X}_\varepsilon$ appare come in fig. 2.

Allora possiamo possiamo calcolare il numero di componenti connesse di $\widehat{X}_\varepsilon$, o equivalentemente la dimensione del gruppo di omolo-

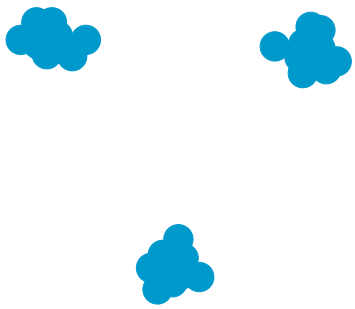


Figura 2. $\widehat{X}_\varepsilon$

gia $H_0(\widehat{X}_\varepsilon; k)$, dove k è un campo. L'osservazione che X è composto essenzialmente da tre componenti è espressa dal fatto che

$$\dim_k(H_0(\widehat{X}_\varepsilon; k)) = 3$$

per un intervallo notevole di valori di ε, e da un certo $\overline{\varepsilon}$ in poi diventa 1.

Ovviamente, non è necessario parlare di dimensione del gruppo di omologia H_0 per discutere del numero di componenti connesse. Tuttavia, se consideriamo l'insieme di dati X come in fig. 3, possiamo chiederci come formalizzare l'intuizione che essi sono disposti in forma circolare.

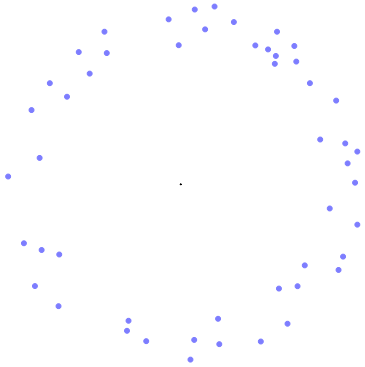


Figura 3. Campionamento da una corona circolare

Ancora una volta possiamo considerare l'insieme $\widehat{X}_\varepsilon$ per diversi valori di ε come in fig. 4 e considerare stavolta il gruppo di omologia $H_1(\widehat{X}_\varepsilon; k)$.

La dimensione di $H_1(\widehat{X}_\varepsilon; k)$ varia fino a stabilizzarsi su 1. Questo ci dice che c'è essenzialmente un buco 1-dimensionale nei dati.

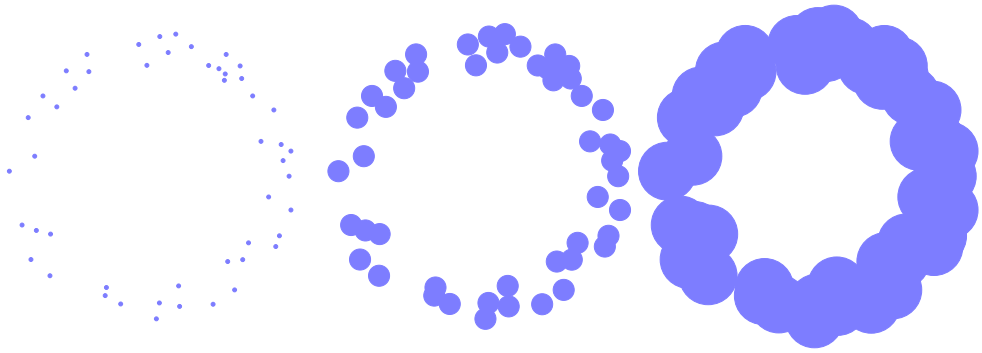


Figura 4. $\widehat{X}_\varepsilon$ al variare di ε

Possiamo anche osservare variazioni di struttura al variare della scala di riferimento. Ad esempio, in fig. 5 possiamo vedere che i tre cluster sulla sinistra collassano in un unico cluster se condiseriamo $\widehat{X}_\varepsilon$ per $\varepsilon \gtrsim e_2/2$.

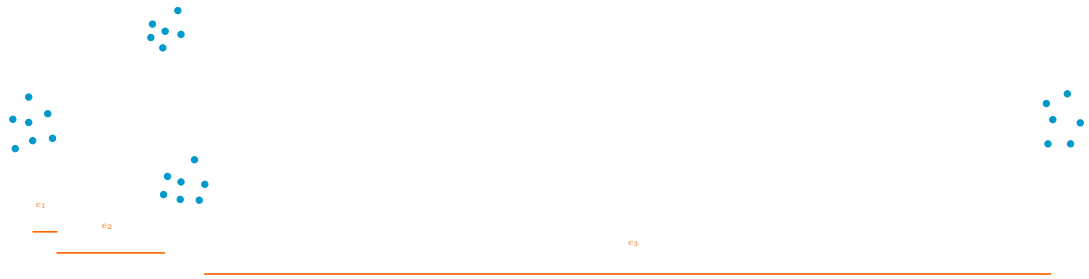


Figura 5. Cluster su scale diverse

Per visualizzare l'andamento di $\dim_k(\widehat{X}_\varepsilon)$ al variare di ε usiamo un tipo di grafico chiamato *persistence barcode*: per ogni *feature* presente nei dati, disegnamo un segmento orizzontale lungo quanto l'intervallo di lunghezze di ε in cui la feature persiste, come in fig. 6.

Il grafico fa interpretato nel seguente modo: all'inizio vi sono 26 punti distinti, al crescere di ε questi punti vengono uniti ad altri e quindi il numero si riduce, finché per $\varepsilon \gtrsim e_1/2$ restano essenzialmente 4 cluster, da $e_2/2$ ne restano solo due e da $e_3/2$ il gruppo di omologia $H_0(\widehat{X}_\varepsilon; k)$ diventa banale. (NMDC: aggiusta il grafico del barcode)

Figura 6. Un esempio di persistence barcode

Possiamo sempre usare questa rappresentazione grazie alla decomposizione dei persistence barcodes garantita dal teorema (NMDC: aggiungere riferimento).

Un altro aspetto che l'omologia persistente cattura è la relazione fra i gruppi di omologia nelle diverse scale, in particolare $H_*(\widehat{X}_\varepsilon; k)$ è funtoriale rispetto all'ordine di $(\mathbb{R}, \leqslant)$, cioé se $\varepsilon_1 \leqslant \varepsilon_2$ allora c'è una mappa $H_*(\widehat{X}_{\varepsilon_1}; k) \to H_*(\widehat{X}_{\varepsilon_2}; k)$ e se $\varepsilon_1 \leqslant \varepsilon_2 \leqslant \varepsilon_3$, allora la mappa associata a $\varepsilon_1 \leqslant \varepsilon_3$ è uguale alla composizione delle due mappe associate a $\varepsilon_1 \leqslant \varepsilon_2$ e $\varepsilon_2 \leqslant \varepsilon_3$.

Questo ci consente di catturare proprietà come quelle che si osservano in fig. 7.

In fig. 8 osserviamo che al variare di ε la dimensione di $H_1(\widehat{X}_\varepsilon; k)$ resta 1, tuttavia è chiaro che i due anelli sono due proprietà distinte dei dati. Questa distinzione non è racchiusa nel gruppo di omologia H_1, mentre la si vede dal fatto che la mappa

$$H_1(\widehat{X}_{\varepsilon_1}; k) \xrightarrow{0} H_1(\widehat{X}_{\varepsilon_2}; k)$$

associata a $\varepsilon_1 \leqslant \varepsilon_2$ è il morfismo nullo. Questo ci dice che non ci sono relazioni fra i due gruppi di omologia.

(NMDC:dire qualcosa sull'algoritmo Mapper?)

Nel resto del capitolo ci occuperemo di costruire l'omologia persistente, con attenzione all'aspetto computazionale.

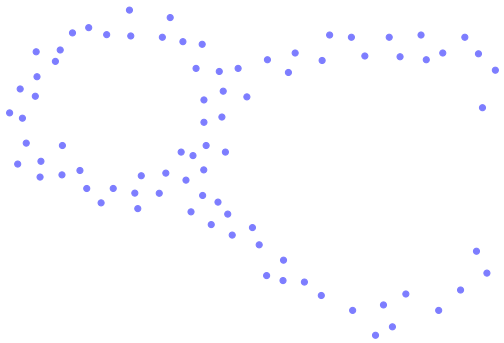
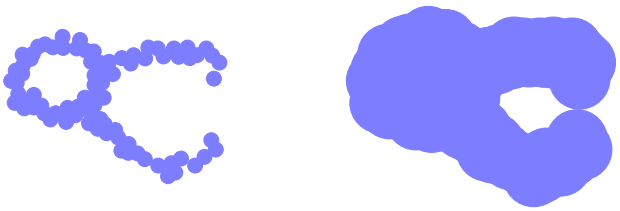


Figura 7. Un doppio anello



(a) $\widehat{X}_{\varepsilon_1}$ (b) $\widehat{X}_{\varepsilon_2}$

Figura 8. Variazione delle proprietà di $\widehat{X}_\varepsilon$

1.2 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 [1, 2] le due definizioni sono equivalenti (almeno per gli spazi che andremo a considerare). Per una trattazione più dettagliata si vedano [3] o [6].

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.

1.2.1 *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 $\tau \in \Sigma(X)$.

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.

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

Scriveremo C_* per indicare tutti i gruppi dei cicli in tutti i gradi.

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:

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

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:

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

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).

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

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

2

CAPITOLO 2

2.1 SEZIONE 1 DEL CAPITOLO 1

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.

3

CAPITOLO 3

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

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

[3] Hatcher, A. Algebraic Topology.

[4] 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.

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

[6] Rotman, J. J. *An introduction to algebraic topology*. Springer-Verlag, 1988.

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