Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea in Matematica

METODI TOPOLOGICI PER LA COMPARAZIONE LINGUISTICA

Relatore:
Prof.
MASSIMO FERRI

Presentata da:
ROBERTA
DE LAZZARI

Anno Accademico 2020 -2021

2

. .

Indice

Introduzione **3**

6

1 Linguistica **7**
1.1 Cos'è la linguistica? 7
1.2 Teorie linguistiche 8
1.3 Obiettivo dell'analisi linguistica 10

10

2 Strumenti Matematici **11**
2.1 Definizioni utili 11
2.2 Omologia persistente 14
2.3 Elementi computazionali 17
2.4 Elementi Probabilistici 19

21

3 Comparazione Linguistica **23**
3.1 Dati 23
3.2 PCA e costruzione dei complessi 24
3.3 Analisi topologica 24
3.3.1 H_0 24
3.3.2 H_1 25
3.3.3 Dimensione della famiglia 26
3.4 Esempio di analisi: lingue Indo-Europee 27
3.5 Confronti con altri metodi 30

Conclusione **33**

Introduzione

"Del triumvirato culturale dell'umanità composto da cultura, pensiero, e linguaggio, il linguaggio è il più accessibile ad un rigoroso studio intellettuale." [1]
Per questo motivo molti studiosi si sono posti l'obiettivo di indagare, mediante strumenti matematici, la struttura e l'evoluzione della lingua e del linguaggio. L'impiego della matematica può risultare insolito, data la natura umanistica dell'oggetto di studio, ma come si può vedere dalle innumerevoli situazioni simili, non solo la matematica si può applicare a questioni linguistiche, ma i risultati che si ottengono sono corretti.
Infatti, come si può leggere anche in [23], citando Giorgio Israel (storico della scienza 1996-2002): lo stesso fenomeno può essere rappresentato da diversi modelli, che sono spesso atti a offrire delle "prospettive diverse ma compatibili" di esso, e lo stesso modello "può essere impiegato per rappresentare fenomeni diversi, fra cui stabilisce una sorta di 'omologia' strutturale". [13]

Per un'introduzione alla teoria dei parametri sintattici su cui si basa la possibilità di applicare una qualsiasi teoria matematica alla linguistica, si rimanda a [1], dove il tema viene affrontato con un linguaggio accessibile anche ad un pubblico privo di conoscenze pregresse in merito alla linguistica.

Questo lavoro si pone l'obiettivo di illustrare come viene implicata l'omologia persistente nell'analisi dell'evoluzione delle lingue nel corso del tempo. Trattandosi di un contesto sviluppatosi recentemente, gran parte delle informazioni sono tratte da [27, 28] e da alcune altre pubblicazioni degli stessi autori. Va tenuto a mente che, trattandosi di un argomento interdisciplinare, sono necessarie diverse figure professionali per permettere una ricerca il più completa possibile e una soddisfacente interpretazione dei dati. Per questo motivo anche l'elaborato si compone di parti apparentemente disgiunte.

In particolare, nel primo capitolo si farà un'introduzione alla linguistica, riportando i concetti teorici necessari per la comprensione dei risultati ottenuti.
Nel secondo capitolo verranno illustrati gli strumenti matematici utilizzati nell'applicazione che se ne farà in seguito. Si approfondirà la parte di omologia persistente, dell'analisi delle componenti principali, e si daranno delle nozioni di base sui processi di Markov nascosti.
Nell'ultima parte si illustreranno le fasi da percorrere per un'efficace analisi delle componenti persistenti dei dati ottenuti dai database SSWL e LanGeLin, e se ne presenteranno gli esiti. Ci si soffermerà sui risultati dell'analisi effettuata nell'insieme delle lingue Indo-Europee. Infine si riassumeranno altri esempi di teorie matematiche implicate per lo stesso scopo che differiscono dall'omologia

persistente, ma che confermano la tesi dell'efficacia della matematica anche nelle applicazioni in ambiti distanti dalla teoria impiegata.

Capitolo 1

Linguistica

1.1 Cos'è la linguistica?

La linguistica è lo studio scientifico del linguaggio che si pone l'obiettivo di descrivere il maggior numero di lingue studiandole da diversi punti di vista. È di particolare interesse capire le relazioni che intercorrono tra le lingue nel loro cambiamento durante il corso del tempo, mediante uno studio matematico e computazionale della loro struttura.
Negli anni si sono sviluppate varie branche della linguistica: storica, del testo, tipologica, pragmatica e matematica. Quest'ultima utilizza dei metodi matematici per studiare non solo le lingue naturali ma anche quelle artificiali (ad esempio linguaggi formalizzati e di programmazione). Dagli anni Sessanta del secolo scorso avviene un notevole sviluppo della linguistica matematica, che vede come figura principale nei suoi studi N. Chomsky, un linguista americano del Novecento. Il suo lavoro più significativo era finalizzato a fornire delle definizioni matematiche di alcuni tipi di grammatiche e delle loro proprietà, codificate mediante valori numerici.

La struttura di una lingua può essere analizzata su vari livelli:

- fonologia: studio della struttura del suono, l'unità elementare è il fonema che viene considerato sotto l'aspetto fisiologico (come viene emesso il suono) e acustico
- morfologia: studio della costruzione delle parole, basata sul morfema, ovvero il minimo elemento di una frase o di una parola dotato di significato (come ad esempio una desinenza)
- sintassi: studio delle regole e dei processi che legano fonemi, morfemi e vocaboli nella formazione della frase (oggetto di studio privilegiato dalla linguistica matematica)
- semantica: studio di come viene trasmesso il significato mediante parole e frasi

Una delle applicazioni della linguistica matematica è la creazione di caldogrammi: alberi binari per la rappresentazione di informazioni inerenti l'evoluzione filogenetica delle famiglie di lingue.

Inoltre, dagli anni Sessanta, un obiettivo che si vuole raggiungere nell'ambito dell'intelligenza artificiale è quello di ottenere delle traduzioni fedeli e grammaticalmente corrette di testi completi. Per farlo si è studiato anche il processo di apprendimento del linguaggio, che avviene naturalmente nei bambini, confrontandolo con quello dell'adulto che si approccia ad una nuova lingua. Come spiegato in [1], è stata teorizzata una grammatica universale che permetterebbe ad una macchina di passare da una lingua all'altra mantenendo il significato del testo iniziale. In questo contesto, l'approccio matematico fornisce la strumentazione per formalizzare i risultati verso cui si tende e i processi intermedi.

1.2 Teorie linguistiche

Per codificare scientificamente le varie lingue esistite ed esistenti la maggior parte delle teorie utilizza il modello dei *Principi e Parametri* sviluppato da N. Chomsky e H. Lasnik negli anni Ottanta, per il quale si rimanda a [7]. Secondo questo modello i principi sono delle regole grammaticali generali, mentre i parametri sono delle variabili binarie utilizzate per la classificazione e distinzione delle lingue sulla base di strutture sintattiche, i quali a volte possono essere dipendenti tra loro. Con questo metodo si associa ad ogni lingua (o famiglia di lingue se pensata come un unicum) una stringa di variabili binarie, con valore $0/1$ oppure $\pm\ 1$, che contengono informazioni a proposito di caratteristiche presenti o meno nella lingua considerata.

Alcuni esempi sono: il *principio di preservazione della struttura* che permette di trasformare una frase nella sua forma passiva mantenendo la conformazione del livello profondo (si fa riferimento alla Transformational Grammar illustrata in seguito); il *principio di proiezione* che stabilisce quali proprietà lessicali si mantengono durante le trasformazioni; il lato in cui va messo il soggetto rispetto alla testa della frase o la possibilità di omettere il soggetto che sono invece dei parametri.

Tra i più conosciuti database che collezionano informazioni in questo modo si trovano il SSWL (*Syntactic Structures of the World Languages*) e il LanGeLin (*Language and Gene Lineages*) [18, 19]. Il primo mappa 252 lingue con 116 variabili sintattiche binarie (non esattamente quelle teorizzate da Chomsky nel modello dei *Principi e Parametri*). Da notare il fatto che alcune lingue sono parzialmente rappresentate perché riportano i dati solo per un numero inferiore di parametri. Il secondo database si basa sugli effettivi parametri sintattici e racchiude informazioni collezionate da Longobardi, linguista italiano e attualmente docente all'università di York, e dai suoi collaboratori, prevedendo la possibilità che i valori assunti siano 0 e ± 1. Partendo da questi dati, generalmente viene fatta un'analisi computazionale e successivamente vengono utilizzati degli strumenti matematici.

I risultati permettono di identificare delle strutture che mettono in relazione tutte le lingue, oppure quelle di una stessa famiglia linguistica, o ancora le diverse famiglie tra di loro. Da queste interazioni si possono dedurre informazioni sulle influenze che ci sono state nel corso degli anni e che hanno portato alle

lingue moderne come le si conosce al giorno d'oggi. Tuttavia molti dati non hanno ancora avuto un'interpretazione soddisfacente.

Si espongono ora, senza la pretesa di scendere nei particolari, alcuni dei principali modelli della Moderna Teoria Sintattica basandosi sul materiale del corso *Mathematical and Computational Linguistics* tenuto da M. Marcolli all'Università di Toronto [21].

Transformational Grammar

Una grammatica elaborata da N. Chomsky nel 1957 che postula l'esistenza nelle frasi di due livelli:

- profondo: più vicino al livello semantico e comune a più lingue
- superficiale: specifico della singola lingua

Il punto cardinale di questa teoria è il postulato secondo il quale il linguaggio è una manifestazione di una capacità innata della mente umana.

Government and Binding

Teoria sviluppata nel 1981 sempre da N. Chomsky, che definisce delle relazioni di dominanza tra elementi di un albero sintattico sfruttando la struttura binaria e lo sviluppo in verticale della proposizione. Il punto di partenza è il modello dei *Principi e Parametri*.

Minimalist Program

Non una teoria, ma un programma elaborato dallo stesso N. Chomsky nel 1993 all'interno dell'impostazione di *Principi e Parametri* [7]. Si postula l'esistenza di una semplice struttura computazionale responsabile della capacità linguistica nella mente umana, rifacendosi all'idea dell'esistenza di una grammatica universale.

Head-driven Phrase Structure Grammar

Grammatica prodotta nel 1987 da Carl Pollard, attualmente professore di Linguistica all'Università statale dell'Ohio, e Ivan Sag, che è stato un linguista, scienziato cognitivo e docente all'Università di Stanford. Questa grammatica è usata nel processare naturalmente le lingue e rappresenta le proprietà con matrici di attributi e valori. Permette di studiare le relazioni di dominanza immediata e di precedenza lineare tra gli elementi di una frase.

Lexical Functional Grammar

È un approccio che va a sostituire il modello dei *Principi e Parametri* , ed è stato teorizzato da Joan Bresnan e Ronald Kaplan, docenti all'Università di Stanford, nel 1982 per le lingue che non hanno una configurazione stabile. Non si ha una distinzione tra struttura profonda e superficiale, poiché è totalmente mancante, e c'è un ordine apparentemente libero delle parole.

Tree-adjoining Grammar

Sviluppatasi nel 1969 come una generalizzazione delle grammatiche libere dal contesto grazie a Aravind Joshi, professore di scienze cognitive nel dipartimento di Scienze Informatiche dell'Universtà della Pennsylvania. Questo tipo di grammatica sfrutta operazioni di sostituzione e unione che descrivono tutte le dipendenze sintattiche di una proposizione.

In tutte queste teorie del Novecento c'è una caratteristica comune: la resa grafica della struttura di una frase mediante alberi binari, che permette di analizzare il tipo di interazioni tra i vari elementi di base. Ciò che differenzia l'una dall'altra, invece, è il tipo di relazioni che si vanno a considerare. Inoltre alcune teorie sono modellate esplicitamente sulla configurazione della famiglia di lingue prese in considerazione (ad esempio quelle con una struttura libera).

1.3 Obiettivo dell'analisi linguistica

Nella linguistica matematica ci si pone l'obiettivo di trovare un metodo che permetta di classificare le lingue su base scientifica: si vuole rispettare l'evoluzione avvenuta finora, facendo corrispondere i risultati ottenuti con quelli storicamente accettati, ma è necessario anche permettere di fare delle previsioni attendibili sui cambiamenti futuri. Dunque bisogna conoscere i punti di contatto e le differenze che ci sono tra le varie lingue, per poter rappresentare graficamente il sunto dei dati di cui si dispone con dei caldogrammi. Questi forniscono informazioni riguardo i cambiamenti strutturali condivisi da un sottogruppo avvenuti nel corso del tempo.

Come già detto, la sintassi è quella che meglio si presta ad un approccio di questo tipo sia per l'impostazione schematica sia perché maggiormente connessa con questioni culturali ed è in grado di dare molte indicazioni su come si struttura il pensiero.

Si vedranno in seguito alcuni studi che analizzano le strutture sintattiche usando vari metodi matematici: topologico [27, 28], heat kernel [24], algebro-geometrico [31], spin glass model [32].

Capitolo 2

Strumenti Matematici

La scienza moderna ha continuamente a che fare con dati di vari tipi, prodotti nei modi più disparati, che si presentano prevalentemente come vettori d-dimensionali. Tuttavia si possono trarre informazioni interessanti anche considerando solo alcune delle componenti di questi vettori. Perciò è necessario l'utilizzo di algoritmi che siano in grado di fornire in output solo le coordinate necessarie.

Successivamente sono impiegate la geometria (studio quantitativo delle distanze) e la topologia (studio qualitativo sulla connessione, non basato su sistemi di riferimento) come strumenti per un'analisi efficace.
Per questo motivo ci si propone di fornire delle nozioni di topologia per poter interpretare i risultati ottenuti nei vari lavori di ricerca presi in considerazione.

Si assume che il lettore abbia familiarità con la teoria di base della Topologia Algebrica, per la quale si rimanda al testo [12] per eventuali approfondimenti.
I contenuti dei primi due paragrafi saranno basati su [6, 8, 9]. per il terzo paragrafo si fa riferimento al testo [14], l'organizzazione e le informazioni del quarto paragrafo sono tratte da [29, 10, 21, 5].

2.1 Definizioni utili

In questo paragrafo ci si propone di richiamare alla memoria alcune definizioni necessarie per la comprensione dell'omologia persistente, che verrà proposta successivamente.
La trattazione segue la struttura di [8].

Poiché i dati che saranno utilizzati successivamente saranno vettori con coefficienti ± 1, oppure 0 e 1, si dà la seguente definizione.

Definizione 1. Siano K un complesso simpliciale (finito), $p \in \mathbb{Z}$. Si chiama p-catena ogni combinazione lineare di p-simplessi con coefficienti in $\mathbb{Z}_2$.
Si indica con C_p lo spazio vettoriale libero su $\mathbb{Z}_2$ generato dalle p-catene.

Ogni p-catena identifica un insieme di p-simplessi di K, e la somma di due p-catene rappresenta la differenza simmetrica dei corrispondenti insiemi.

Si introduce ora, per ogni $p \in \mathbb{Z}$, un operatore lineare chiamato *operatore di bordo*

$$\partial_p : C_p \to C_{p-1},$$

che è sufficiente definire sui generatori di C_p, ovvero i p-simplessi, e poi estenderlo per linearità.

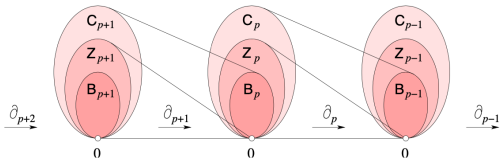


Figura 2.1: successione di complessi connessi da omomorfismi di bordo

Scrivendo $\sigma = [u_0, u_1, \ldots, u_p]$, si denota con $[u_0, \ldots, \hat{u}_j, \ldots, u_p]$ la faccia di σ generata da tutti i suoi vertici eccetto u_j, per $j = 0, \ldots, p$.
Si definisce quindi

$$\partial_p(\sigma) := \sum_{j=0}^{p} [u_0, \ldots, \hat{u}_j, \ldots, u_p]$$

È possibile provare che

$$\partial_p \partial_{p+1} = 0,$$

da cui segue che $B_p(K) \subset Z_p(K)$, con $B_p(K) := Im\partial_{p+1}$ e $Z_p(K) := Ker\partial_p$.

Osservazione 2. Si userà la notazione B_p e Z_p per riferirsi rispettivamente a $B_p(K)$ e $Z_p(K)$ nei casi in cui non ci sarà pericolo di confusione.

Osservazione 3. Gli elementi di B_p sono detti $p - bordi$, mentre gli elementi di Z_p sono detti $p - cicli$.

Definizione 4. Sia K un complesso simpliciale, il *p-esimo gruppo di omologia* è definito come il quoziente

$$\mathrm{H}_p(K) := Z_p(K)/B_p(K).$$

Intuitivamente si può dire che le classi non nulle di omologia sono rappresentate da cicli che non sono bordi.

Definizione 5. Siano X uno spazio topologico, spesso la realizzazione geometrica di un complesso simpliciale, e $f : X \to \mathbb{R}$ continua chiamata *funzione filtrante*. Per ogni $u \in \mathbb{R}$, si definisce *insieme di sottolivello* u l'insieme

$$X_u := \{x \in X \mid f(x) \leq u\}$$

Siano K un complesso simpliciale e $f : K \to \mathbb{R}$.
Diciamo che f è monotona se $f(\sigma) \leq f(\tau)$ per ogni σ faccia di τ in K.
Segue che l'insieme di sottolivello $K_a = f^{-1}(-\infty, a]$ è un sottocomplesso di K per ogni $a \in \mathbb{R}$. Se m è il numero di simplessi in K, si ottengono $n + 1 \leq m + 1$ diversi sottocomplessi, che si ordinano come una sequenza crescente

$$\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K$$

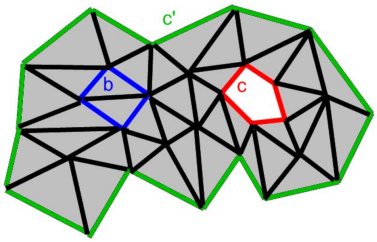


Figura 2.2: complesso simpliciale K, la catena blu è un 1-ciclo e un bordo; le catene rossa e verde sono 1-cicli ma non bordi, sono quindi 1-cicli non omologhi a zero. Sono inoltre fra loro omologhi, in quanto la loro differenza (mod 2) è il bordo di una catena 2-dimensionale.

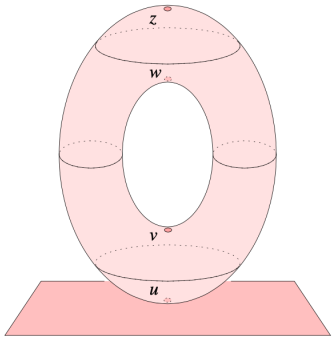


Figura 2.3: la funzione altezza sul toro con i punti critici u, v, w, z. Gli insiemi di livello sono tra i rispettivi valori di altezza

Definizione 6. La sequenza di complessi $\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K$ così ottenuta è detta *filtrazione* di f.

Per ogni $i \leq j$ si ha un'inclusione di mappe che induce l'omomorfismo

$$f_p^{i,j} : \mathrm{H}_p(K_i) \to \mathrm{H}_p(K_j),$$

per ogni dimensione p.
Così facendo la filtrazione corrisponde ad una sequenza di gruppi di omologia collegati da omomorfismi

$$0 = \mathrm{H}_p(K_0) \to \mathrm{H}_p(K_1) \to \cdots \to \mathrm{H}_p(K_n) = \mathrm{H}_p(K)$$

per ogni dimensione p.

Ciò che è di particolare interesse nella teoria dell'omologia persistente sono le classi che o si acquisiscono o si perdono (diventano banali o si uniscono ad altre) andando da K_i a K_j con $i < j$.

2.2 Omologia persistente

L'obiettivo di questo paragrafo è di illustrare i concetti basilari e gli oggetti della teoria dell'omologia persistente che ci saranno utili per la comprensione del capitolo successivo. A livello intuitivo si può dire che l'omologia persistente tiene traccia di quali caratteristiche topologiche permangono al variare di alcuni parametri dello spazio considerato.

Si consideri una situazione che sarà di aiuto per capire intuitivamente il meccanismo della persistenza:

Siano X uno spazio topologico connesso e $f : X \to \mathbb{R}$ una funzione filtrante. Per ogni $a, b \in \mathbb{R}$, con $a \leq b$, si ottiene una famiglia di sottolivelli annidati $X_a \subseteq X_b$. Si può pensare questa famiglia come un insieme di sottolivello X_a che cambia e cresce con l'aumentare del parametro a.
Per riuscire a rappresentare come cambia lo spazio durante questo procedimento si pensi ad ogni componente di X_a come un punto in un piano (Figura 2.4 a destra). Si ottiene così il grafico $G(f)$, disegnato a partire dal basso verso l'alto, pensando ad f come una funzione altezza.

Essendo X uno spazio connesso, le componenti di X_a, con l'aumentare di a, possono di volta in volta unirsi ma mai separarsi, allo stesso modo si comportano gli archi nel grafico (vedi Figura 2.4 a destra). In questo modo si ottiene alla fine una singola componente, ossia X, per un valore del parametro a abbastanza grande.

È importante notare che nel punto in cui due componenti si uniscono l'arco del grafico corrispondente alla più 'giovane' termina mentre sopravvive quello della più 'vecchia'.

Definizione 7. Con queste ipotesi si ha che $G(f)$ è un albero, chiamato *merge tree*.

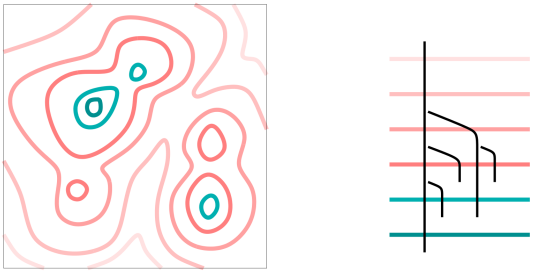


Figura 2.4: Sinistra: funzione sul quadrato unitario visualizzata disegnando sei insiemi di livello, colori più chiari indicano un valore minore del parametro. Destra: il merge tree della funzione.

Definizione 8. I *p-esimi gruppi di omologia persistente* sono le immagini degli omomorfismi indotti dalle inclusioni $\mathrm{H}_p^{i,j} = Im f_p^{i,j}$ per $0 \leq i \leq j \leq n$. Ovvero sono le classi di omologia di K_i che sono ancora vive in K_j:

$$\mathrm{H}_p^{i,j} := \mathrm{Z}_p(K_i)/(\mathrm{B}_p(K_j) \cap \mathrm{Z}_p(K_i))$$

per ogni dimensione p e per ogni coppia di indici $i \leq j$.

Definizione 9. Sia γ una classe in $\mathrm{H}_p(K_i)$

- si dice che γ *nasce* in K_i se $\gamma \notin \mathrm{H}_p^{i-1,i}$
- se γ è nata in K_i, si dice che *muore entrando in* K_j se si unisce con una classe più vecchia passando da K_{j-1} a K_j, ovvero

$$f_p^{i,j-1}(\gamma) \notin \mathrm{H}_p^{i-1,j-1} \text{ ma } f_p^{i,j}(\gamma) \in \mathrm{H}_p^{i-1,j}$$

Definizione 10. Se γ nasce in K_i e muore entrando in K_j, allora si chiama *persistenza* la differenza tra i valori delle funzioni, ovvero

$$\mathrm{pers}(\gamma) := a_j - a_i.$$

Più semplicemente, si può definire l'*indice di persistenza* della classe γ come la differenza tra gli indici: $j - i$.
Se γ nasce in K_i ma non muore mai, allora si fissa la sua persistenza e il suo indice di persistenza a infinito.

Definizione 11. I *p-esimi numeri di Betti persistenti* sono i ranghi dei gruppi di omologia

$$\beta_p^{i,j} := \mathrm{rank}\ \mathrm{H}_p^{i,j}.$$

Questi rappresentano il numero di classi di p-cicli di $\mathrm{H}_p(K_i)$ che "sopravvivono" in $\mathrm{H}_p(K_j)$.

Più precisamente, per ogni $i, j \in \mathbb{R}$, $i < j$, la *funzione del p-esimo numero di Betti persistente* assegna alla coppia (i, j) la dimensione della trasformazione lineare $f_p^{i,j}$, ovvero il suo numero di Betti persistente.

Definizione 12. È possibile visualizzare i numeri di Betti persistenti in un diagramma bidimensionale (in $\overline{\mathbb{R}}^2$) in cui le informazioni più importanti sono la posizione di alcuni punti (*cornerpoints*) e di alcune linee (*cornerlines*).

Le coordinate (i, j) di un *cornerpoint* rappresentano rispettivamente i momenti di "nascita" e di "morte" di un generatore. Analogamente l'ascissa di una *cornerline* è il momento in cui nasce un generatore (non muore mai).

La differenza $j - i$ tra le coordinate indica la *persistenza* del generatore.

L'insieme delle *cornerlines* (spesso sostituite dai loro *cornerpoints all'infinito*) con l'aggiunta di tutti i punti della diagonale $x = y$, è chiamato *diagramma di persistenza*.

È necessario precisare che per questa trattazione non si considerano le molteplicità di ogni *cornerpoint* e *cornerline*.

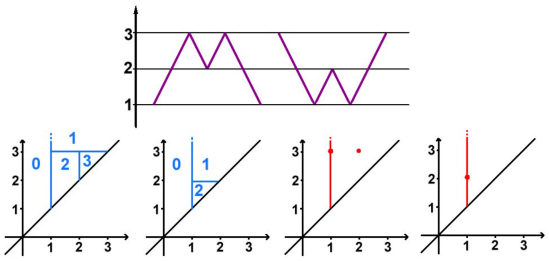


Figura 2.5: Sopra: gli oggetti di studio. In basso a sinistra: funzioni dei numeri di Betti 0-dimensionali di M e W. In basso a destra: i rispettivi diagrammi di persistenza.

Osservazione 13. L'importanza di tale diagramma deriva dal fatto che, in questo modo, è possibile distinguere anche due oggetti geometricamente equivalenti, ovvero che grazie ad una serie di movimenti rigidi o di trasformazioni continue risulterebbero uguali. Si riporta un esempio nella Figura 2.5

Definizione 14. Dati due diagrammi di persistenza X e Y, si definisce la *Bottleneck distance* tra i due come

$$W_\infty(X,Y) := \inf_{\eta: X \to Y} \sup_{x \in X} \|x - \eta(x)\|_\infty ,$$

dove $\eta : X \to Y$ è una biiezione, e se $x = (x_1, x_2)$ e $y = (y_1, y_2)$ allora $\|x - y\|_\infty = \max\{|x_1 - y_1|, |x_2 - y_2|\}$.

Definizione 15. Sia X un sottoinsieme finito di uno spazio metrico (Y, d) con distanza d. Il *complesso di Vietoris-Rips* di X relativo al parametro ε è definito come

$$VR(X,\varepsilon) := \{\sigma \subseteq X \mid \operatorname{diam}(\sigma) \leq 2\varepsilon\},$$

ovvero il complesso simpliciale astratto finito il cui insieme di vertici è X e dove $\{x_0, x_1, \ldots, x_k\}$ genera un k-simplesso se e solo se $d(x_i, x_j) \leq \varepsilon$ per ogni $0 \leq i \neq j \leq k$.

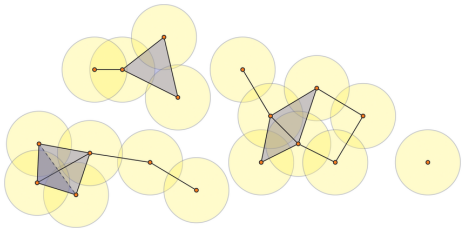


Figura 2.6: Esempio di un complesso di Vietoris-Rips

La necessità di definire il complesso di Vietoris-Rips è data dal fatto che si lavorerà con insiemi di punti (dati) e con la distanza euclidea, dunque X sarà

già dotato della topologia discreta, ma servirà trasformare i dati in un oggetto topologico interessante. Usando il complesso di V-R si riuscirà nell'intento, e di ogni complesso si potrà calcolare i rispettivi gruppi di omologia (per studiare il numero di "*buchi*" p-dimensionali nel dataset).

Definizione 16. Un modo per rappresentare graficamente i numeri di Betti persistenti di $VR(X, \varepsilon)$ è il cosiddetto *barcode*, che consiste in una serie di linee orizzontali in un piano il cui asse orizzontale corrisponde al parametro ε, mentre l'asse verticale indica l'ordine arbitrario dei generatori dell'omologia.

La peculiarità di questo grafico è di saper rappresentare qualitativamente caratteristiche significative (i *generatori* dei gruppi di omologia) dello spazio con cui si sta lavorando, filtrando il "rumore di fondo" che si genera con il complesso di Vietoris-Rips. [11]

2.3 Elementi computazionali

Nell'applicazione che sarà presentata successivamente, ma anche in tutte le altre, si lavora a partire da un insieme di dati che, per essere utilizzati, devono presentarsi sotto forma di complesso simpliciale. Per ottenere questo risultato si fa uso di una tecnica computazionale che permette di ridurre la dimensione dei dati: l'Analisi delle Componenti Principali (PCA).

La PCA è uno dei metodi utilizzati per la riduzione dimensionale derivante dall'ambito della statistica multivariata, il cui obiettivo è quello di costruire un sottospazio dove l'errore medio di ricostruzione del *training set* sia minimo, e di permettere una migliore interpretazione dei dati.

Algebricamente, le componenti principali sono delle particolari combinazioni lineari delle p variabili iniziali $X_1, X_2, \ldots, X_p$, ovvero dei dati. Geometricamente, queste nuove variabili rappresentano la selezione di un nuovo sistema di coordinate ottenuto dalla rotazione del sistema originale che aveva $X_1, X_2, \ldots, X_p$ come assi. I nuovi assi rappresentano le direzioni con la massima variabilità e forniscono una descrizione più semplice della struttura della covarianza.

Definizione 17. La *matrice di covarianza* $\Sigma = (\sigma_{i,j})$ del vettore di variabili $X^T = [X_1, X_2, \ldots, X_p] \in \mathbb{R}^{n \times p}$ è la matrice simmetrica con

$$\sigma_{i,j} := \mathrm{Cov}(X_i, X_j) = \frac{1}{n-1} \sum_{k=1}^{n} (x_{k,i} - \bar{x}_i)(x_{k,j} - \bar{x}_j),$$

dove $i, j = 1, 2, \ldots, p$, e $\bar{X} := [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p] \in \mathbb{R}^p$ è il vettore delle medie.

Σ è una misura di *dispersione* rispetto alla media dei dati.

Osservazione 18. Notiamo che gli elementi sulla diagonale di Σ sono

$$\sigma_{i,i} = \mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i).$$

Definizione 19. La *matrice di correlazione* $R = (r_{i,j})$ del vettore di variabili $X^T = [X_1, X_2, \ldots, X_p] \in \mathbb{R}^{n \times p}$ è data da

$$r_{i,j} := \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{Var}(X_i)\,\mathrm{Var}(X_j)}} = \frac{\sigma_{i,j}}{\sqrt{\sigma_{i,i}}\sqrt{\sigma_{j,j}}}$$

con $i, j = 1, 2, \ldots, p$.

R è la misura dell' *associazione lineare* tra le variabili.

Definizione 20. Sia Σ la matrice di covarianza associata al vettore delle variabili $X^T = [X_1, X_2, \ldots, X_p]$. Consideriamo le autocoppie di Σ: $(\lambda_1, v_1), (\lambda_2, v_2), \ldots, (\lambda_p, v_p)$ dove $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. Allora la *i-esima componente principale* è data da

$$Y_i = v_i^T X = v_{i,1}X_1 + v_{i,2}X_2 + \cdots + v_{i,p}X_p,$$

per $i = 1, 2, \ldots, p$.

Per fare in modo che le nuove variabili Y_i siano non correlate e rappresentino al meglio la varianza dei dati, è necessario che soddisfino le seguenti richieste:

$$\max_{v_i^T v_i = 1} \mathrm{Var}(Y_i) = \max_{v_i^T v_i = 1} v_i^T \Sigma v_i = \lambda_i, \quad i = 1, 2, \ldots, p \tag{2.1}$$

$$\mathrm{Cov}(Y_i, Y_j) = v_i^T \Sigma v_j = 0, \quad i \neq j \tag{2.2}$$

Se qualche λ_i è uguale, le scelte dei corrispondenti vettori v_i, e quindi di Y_i, non sono uniche.

Osservazione 21. Si noti che per la prima componente principale è necessario che $Y_1 = v_1^T X$ soddisfi solo la prima condizione, ovvero deve semplicemente massimizzare la varianza $\mathrm{V}ar(Y_1)$.

L'ultima uguaglianza in (2.1) è conseguenza dei seguenti risultati:

Teorema 22 (di Rayleigh-Ritz)**.** Sia A Hermitiana con $x^*Ax \in \mathbb{R}\ \forall\, x \in \mathbb{C}^n$. Siano $\lambda_1 \leq \lambda_1 \leq ... \leq \lambda_n$ gli autovalori di A. Vale

$$\lambda_1 \leq \frac{v^*Av}{v^*v} \leq \lambda_n \quad \forall\, 0 \neq v \in \mathbb{C}^n$$

dove $\frac{v^*Av}{v^*v}$ è detto *quoziente di Rayleigh.*

Da cui segue direttamente:

Teorema 23 (di Courant-Fischer)**.** Sia A una matrice reale simmetrica con autovalori $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$, sia $k \in \{1, 2, \ldots, n\}$. Allora

$$\min_{w_1,\ldots,w_{n-k} \in \mathbb{C}^n} \quad \max_{0 \neq x \in \mathbb{C}^n \ x \perp w_1,\ldots,w_{n-k}} \frac{x^*Ax}{x^*x} = \lambda_k$$

$$\max_{w_1,\ldots,w_{k-1} \in \mathbb{C}^n} \quad \min_{0 \neq x \in \mathbb{C}^n \ x \perp w_1,\ldots,w_{n-1}} \frac{x^*Ax}{x^*x} = \lambda_k$$

Il seguente è un risultato molto importante per lo scopo di riduzione dimensionale.

Teorema 24. Sia $X^T = [X_1, X_2, \dots, X_p]$ il vettore delle variabili con matrice di covarianza Σ, le cui autocoppie sono (λ_i, v_i) con $i = 1, \dots, p$, con gli autovalori in ordine decrescente e positivi. Siano $Y_i = v_i^T X_i$ le componenti principali con $i = 1, \dots, p$. Allora:

$$\sigma_{1,1} + \sigma_{2,2} + \cdots + \sigma_{p,p} = \sum_{i=1}^{P} Var(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} Var(Y_i)$$

Dal Teorema 24 si può direttamente dare la seguente definizione:

Definizione 25. La *frazione di varianza totale spiegata dalle prime k componenti principali* è:

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

Osservazione 26. Quest'ultima ci permette di troncare la PCA alle prime k variabili nel momento in cui queste spiegano l'80-90% della varianza totale. In questo modo $Y_1, \dots, Y_k$ saranno le nuove variabili (componenti principali).

2.4 Elementi Probabilistici

Il presente paragrafo è volto a dare una breve contestualizzazione dei modelli di Markov e in particolare dei modelli di Markov nascosti (Hidden Markov Models = HMM). Questi trovano un'interessante applicazione nell'analisi sintattica di cui si andranno ad esporre i risultati successivamente. In particolare saranno molto utili nello studio dei parametri sintattici e dei loro cambiamenti, ragionando su una struttura di albero binario con radice, in cui i vertici saranno i parametri e lungo ogni vertice si penserà una matrice di transizione.

Definizione 27. Un *processo stocastico a tempo discreto* su $(\Omega, \mathcal{F}, \mathbb{P})$ spazio di probabilità, è una famiglia $X = (X_n)_{n \in I}$ di variabili aleatorie, dove $I \subset \mathbb{N}_0$, ovvero:

$$X_n \in m\,\mathcal{F}\ \forall\, n \in I, \quad X_n : \Omega \to (\mathbb{R}^N, \mathcal{B}_N) \quad N > 0$$

Si considera $I = \mathbb{N}_0$ se non diversamente specificato.

Definizione 28. Un processo X a tempo discreto su $(\Omega, \mathcal{F}, (\mathcal{F}_n)_n, \mathbb{P})$, con $(\mathcal{F}_n)_n$ filtrazione, ha la *proprietà di Markov* se

- è adattato, ovvero $X_n \in m\,\mathcal{F}_n \quad \forall\, n \in \mathbb{N}_0$
- $\mathbb{E}[\varphi(X_{n+1})|\mathcal{F}_n] = \mathbb{E}[\varphi(X_{n+1})|X_n] \quad \forall \varphi \in m\,\mathcal{B}$ limitata, $\forall n \in \mathbb{N}_0$

Per fissare le idee si consideri un sistema di cui si può dire in ogni istante in quale degli N stati diversi $S_1, \dots, S_N$ si trova. Ad intervalli di tempo costanti e discreti, il sistema cambia stato (potendo rimanere nello stesso) secondo un insieme di probabilità associate allo stato. Chiamiamo $t = 1, 2, \dots$ gli istanti di cambiamento, e q_t l'effettivo stato al tempo t.

Definizione 29. Chiamiamo *catena di Markov discreta*, del primo ordine, il processo la cui descrizione probabilistica può essere troncata allo stato corrente e al predecessore:

$$\mathbb{P}(q_t = S_j | q_{t-1} = S_i,\ q_{t-2} = S_k, \dots) = \mathbb{P}(q_t = S_j | q_{t-1} = S_i) \tag{2.3}$$

Quindi si può dire che i modelli di Markov godono della proprietà di assenza di memoria.

Osservazione 30. Si considerano solo i processi in cui l'elemento di destra di (2.3) è indipendente dal tempo, in modo che l'insieme delle probabilità di transizione $a_{i,j}$ sia della forma:

$$a_{i,j} = \mathbb{P}(q_t = S_j | q_{t-1} = S_i), \quad 1 \leq i, j \leq N \tag{2.4}$$

Queste soddisfano i seguenti vincoli stocastici standard:

$$a_{i,j} \geq 0 \tag{2.5}$$

$$\sum_{j=1}^{N} a_{i,j} = 1 \tag{2.6}$$

Quello appena descritto può essere considerato un modello di Markov *osservabile* poiché l'output del processo è un insieme di stati ad ogni istante di tempo, e gli stati corrispondono ad eventi osservabili (Esempio: la registrazione dei dati relativi al meteo su base quotidiana in cui gli stati potrebbero essere {soleggiato, nuvoloso, piovoso}).

Ora si vuole estendere questo modello per includere i casi in cui le osservazioni sono funzioni di stato probabilistiche, e in cui non è noto q_t. Questi sono detti modelli di Markov *nascosti* perchè c'è un processo stocastico sottostante (la serie di stati) che non è rilevabile, ma influenza la sequenza di eventi osservati.

Definizione 31. Un *HMM* è caratterizzato da:

1. N ovvero il numero di stati nel modello, anche se nascosti, indicati con $S = \{S_1, S_2, \dots, S_N\}$
2. M ovvero il numero di simboli osservabili per ogni stato che corrispondono all'output fisico del sistema una volta modellizzato, indicati con $V = \{v_1, v_2, \dots, v_M\}$
3. $A = (a_{i,j})$ ovvero la matrice $N \times N$ di probabilità di transizione di stato dove
$$a_{i,j} = \mathbb{P}(q_{t+1} = S_j | q_t = S_i), \qquad 1 \leq i, j \leq N.$$
4. $B = \{b_j(k)\}$ insieme di distribuzioni di probabilità dei simboli osservati nello stato j dove
$$b_j(k) = \mathbb{P}(v_k \text{ al tempo } t | q_t = S_j), \quad 1 \leq j \leq N \ \ 1 \leq k \leq M$$
5. $\pi = (\pi_i)$ l'iniziale distribuzione di stato dove
$$\pi_i = \mathbb{P}(q_1 = S_i), \quad 1 \leq i \leq N$$

Assegnati dei valori appropriati di N, M, A, B, π, si ha che l'HMM può essere usato come un generatore per fornire una sequenza di osservazioni $O = O_1 O_2, \ldots, O_T$, dove T è il numero di osservazioni della sequenza, e ogni osservazione O_t è uno dei simboli di V. Questo può essere fatto nel seguente modo:

(a) si sceglie uno stato iniziale $q_1 = S_i$ in accordo alla distribuzione di stato iniziale π

(b) si fissa $t = 1$

(c) si sceglie $O_t = v_k$ basandosi sulla distribuzione di probabilità del simbolo nello stato S_i ovvero $b_i(k)$

(d) si transita in un nuovo stato $q_{t+1} = S_j$ secondo la distribuzione di probabilità di transizione di stato per lo stato S_i ovvero $a_{i,j}$

(e) si fissa $t = t + 1$ e si ritorna allo step (c) se $t < T$, altrimenti si termina la procedura

Osservazione 32. Per definire al completo un HMM sono necessari diversi parametri, come si è visto, quindi per comodità si usa la notazione compatta $\lambda = (A, B, \pi)$ per indicare la totalità dei parametri del modello

Prima di poter applicare un HMM al mondo reale è necessario risolvere i seguenti tre problemi:

Problema di valutazione: ovvero data una sequenza di osservazioni O e un modello λ, come si calcola efficientemente la probabilità $\mathbb{P}(O|\lambda)$? Questo ci permette di scegliere il modello che meglio combacia con le osservazioni

Problema di ottimizzazione: ossia data una sequenza O e un modello λ, come si sceglie la corrispondente sequenza di stati $Q = q_1 q_2 \ldots q_T$ che meglio 'spiega' le osservazioni? Questo ci permette di comprendere il significato fisico degli stati nascosti del modello

Problema di addestramento: data una sequenza O, come si adeguano i parametri del modello $\lambda = (A, B, \pi)$ per massimizzare $\mathbb{P}(O|\lambda)$? Questo ci permette di costruire un HMM che descrive al meglio come si ottiene una determinata sequenza di osservazioni

Capitolo 3

Comparazione Linguistica

Come già detto nel primo capitolo, l'analisi linguistica che risulta matematicamente più interessante è quella basata sul modello dei Principi e Parametri di Chomsky [7]. Questo approccio si presta particolarmente anche all'utilizzo delll'omologia persistente principalmente per due scopi: per identificare eventuali dipendenze tra gli stessi parametri [16, 22, 24, 25, 30, 32] e per scovare eventuali relazioni tra le lingue sulla base della distribuzione delle caratteristiche sintattiche. Quest'ultima questione è stata affrontata computazionalmente per ricostruire gli alberi filogenetici delle famiglie linguistiche [17, 18, 19, 20] e recentemente anche dal punto di vista algebrico-geometrico [31].

3.1 Dati

La base da cui partire è un database affidabile che sia in grado di mappare il maggior numero di lingue nel modo più completo possibile. Negli articoli [27, 28] a cui si fa maggior riferimento, sono utilizzati SSWL e LanGeLin, come già visto. Tuttavia emergono dei problemi con entrambi: in SSWL le lingue sono presenti con completezza variabile (alcune al 100%, altre con una percentuale minima di variabili mappate); il LanGeLin invece non ha questo limite ma si trovano i dati di un numero inferiore di lingue. Precedenti studi [24] hanno dimostrato che i due dataset si comportano diversamente in merito alle proprietà di clustering, e per quanto riguarda la ricostruzione filogenetica si ottengono risultati migliori con l'utilizzo di LenGeLin.

Una possibile motivazione potrebbe essere l'omogeneità della precisione dei dati di uno dei due, oppure la presenza in LanGeLin dei dialetti derivanti dal Greco e dal Latino, in quanto apportano delle microvariazioni intermedie tra lingue che sono più diffuse. Tuttavia, si potrebbe attribuire la responsabilità anche al procedimento di filtrazione dei dati attuato prima dell'applicazione della PCA, passaggio utile quando si sta lavorando con SSWL.

Si vedrà ora più nel dettaglio come viene applicata l'omologia persistente alla ricostruzione delle strutture gerarchiche sulla base delle caratteristiche sintattiche. Il risultato sarà un albero che permetterà di intuire l'evoluzione delle lingue in base alle mutazioni sintattiche.

3.2 PCA e costruzione dei complessi

Per poter procedere con l'analisi, i dati devono essere resi utilizzabili: in questa fase lo scopo è quello di ottenere una sottovarietà che li contenga. La PCA risulta particolarmente utile per ridurre la dimensione di questo spazio, la cui base sarà costituita dalle componenti principali. Si applica la PCA con diversi livelli di varianza (ad esempio 60% o 80%) anche se questo porterà a delle differenze nella suddivisione delle lingue all'interno delle sottofamiglie. Tuttavia i cluster principali non ne saranno influenzati.

Un problema che si pone è la correlazione dei parametri (ad esempio la posizione di soggetto e verbo in una proposizione). Per un modello dinamico dell'evoluzione linguistica non si possono considerare i parametri come variabili indipendenti e identicamente distribuite, ma saranno assegnati dei pesi: "peseranno" di più i parametri genuinamente indipendenti. Una scelta naturale è quella di utilizzare i coefficienti delle componenti principali che massimizzano la varianza.

Una volta ottenute delle variabili continue si misura la prossimità dei punti mediante la distanza Euclidea. È così possibile costruire dei complessi simpliciali per ogni raggio scelto: i complessi di Vietoris-Rips. Questo procedimento prevede il calcolo di un raggio critico r_c, ovvero il raggio minimo che permette di ottenere uno spazio connesso. Si fissa poi un parametro $\varepsilon = r_c/100$, e nell'algoritmo per la costruzione dei complessi si utilizzano i raggi $\varepsilon, 2\varepsilon, \ldots, N\varepsilon, \ldots$.

Successivamente sarà calcolata l'omologia persistente dell'insieme dei dati: si otterrà così un barcode e il relativo 1-scheletro. L'ultimo passaggio è quello della costruzione degli alberi delle componenti connesse persistenti, che tuttavia non vanno considerati come alberi filogenetici seppur contengano informazioni di questo tipo.

3.3 Analisi topologica

Partendo dai dati, gli autori di [28] si sono posti tre domande:

1 In che modo i generatori di H_0 possono essere utilizzati come metodo alternativo per ricostruire gli alberi filogenetici delle famiglie linguistiche?

2 Ci sono strutture di dimensione maggiore e che significato hanno in merito alla linguistica storica?

3 È possibile una stima dimensionale per le diverse famiglie linguistiche, dove la dimensione è una misura di quanto alcune delle proprie caratteristiche siano diffuse all'interno di una stessa famiglia?

3.3.1 H_0

Lo studio del gruppo omologico persistente H_0 dà informazioni in merito al numero di componenti connesse della relativa famiglia di complessi di Vietoris-Rips, al variare del raggio con cui vengono costruiti. Mediante la costruzione

degli alberi dati da queste componenti connesse persistenti, si possono studiare le suddivisioni in sottogruppi delle lingue iniziali e si possono osservare delle differenze con le informazioni storicamente accettate. Per questo motivo rappresentano solo un'organizzazione gerarchica di come si sono diffuse alcune caratteristiche sintattiche attraverso le lingue.
Infatti sono molto accurati nella suddivisione in sotto-famiglie, ma presentano anche sostanziali differenze. In particolare i nodi più interni degli alberi filogenetici vanno intesi come gli antenati delle lingue moderne, negli alberi "persistenti" indicano solo la suddivisione gerarchica dei clusters, ma non rappresentano relazioni temporali.

Inoltre, i numeri di Betti persistenti di H_0 danno informazioni sul numero di componenti connesse persistenti del complesso di Vietoris-Rips. Un valore basso sembra indicare che i parametri sintattici sono più concentrati e distribuiti omogeneamente tra le lingue della famiglia di cui si è calcolato il gruppo di omologia. Questi parametri sembrerebbero anche meno inclini ad essere condivisi con lingue di altri sottogruppi.

La costruzione dell'albero "persistente" riflette la struttura del barcode:

$$se\, C_r = \text{insieme di tutti i clusters di dimensione } r \quad C = \bigsqcup_r C_r$$

$$C_i \text{ figlio di } C_j \text{ se } C_i \subseteq C_j \text{ ma } \nexists\, C_k \,|\, C_i \subseteq C_k \subseteq C_j$$

Per un approfondimento a riguardo si faccia riferimento a [15].

3.3.2 H_1

Poiché non ci sono generatori persistenti non banali di H_k per $k \geq 2$, l'unica struttura significativa è quella del gruppo H_1 . Si possono trovare due tipi di generatori persistenti non banali di H_1: dovuti ad interazioni avvenute storicamente tra lingue non appartenenti alla stessa sottofamiglia; causati da scambi sintattici tra lingue, ma che non trovano spiegazione storica. In questo secondo caso o il generatore non è quello rilevante storicamente ma uno omologo, oppure è dovuto al fenomeno dell'omoplasia.

In linguistica, quando si parla di fenomeno omoplasia si fa riferimento a caratteristiche che una lingua ha acquisito da due rami diversi dell'albero filogenetico che non sono riconducibili ad un antenato comune. Quando si verifica questa situazione ci si aspetta di trovare un generatore del primo gruppo di omologia.

Dall'analisi condotta è emerso anche che generatori di H_1 persistente compaiono solo quando si considerano clusters molto grandi (almeno 30 lingue, nell'analisi che utilizza gli interi database). Per identificare il ciclo di lingue a cui si deve la nascita del generatore si procede *manualmente* eliminandone uno alla volta e ricalcolando l'omologia persistente dei rimanenti. Se ci sono più

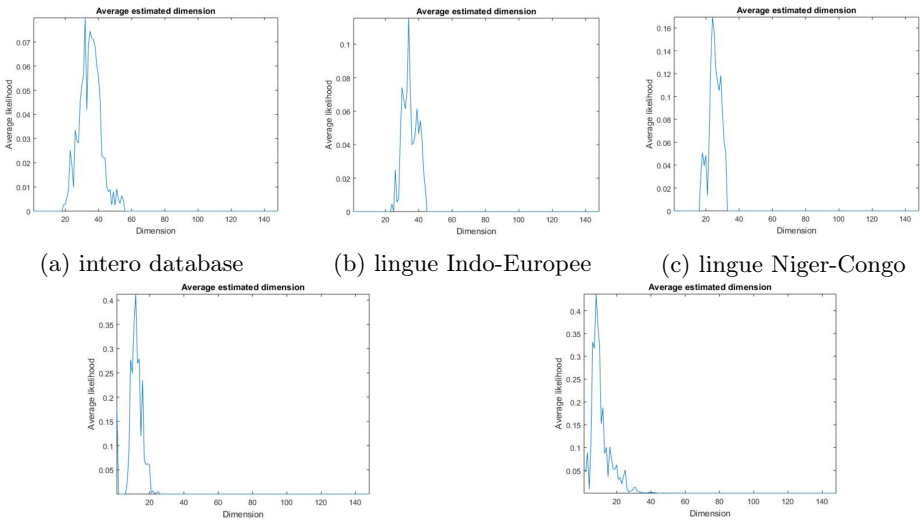
responsabili di un generatore allora questi sono omologhi e possono essere scelti entrambi. Se uno degli insiemi che si sta analizzando non è responsabile, allora questo sarà un bordo nel complesso di Vietoris-Rips.

3.3.3 Dimensione della famiglia

Quando ci si riferisce alla dimensione, è importante precisare che si sta parlando della dimensione dello spazio che contiene i dati. Questi ultimi vengono visti come vettori binari che mappano i parametri nelle varie lingue: nella matrice dei dati che si usa nella PCA, ogni dato rappresenta una variabile sintattica, mentre ogni entrata della matrice contiene le informazioni riguardanti una lingua.

La stima della dimensione dei dati è funzionale per avere una migliore comprensione delle relazioni e di eventuali dipendenze che esistono tra i parametri. Viene quindi costruito un algoritmo che permette di fare questa stima in modo da riuscire a trovare "la geometria della sintassi", che rimane una delle questioni aperte nel modello dei Principi e dei Parametri. Inoltre si può calcolare la densità dei dati. Questi calcoli ci permettono anche di stabilire quanto alcune lingue siano diffuse (in termini di densità e dimensione) all'interno di una specifica famiglia.

Si è visto che, lavorando con l'intero database SSWL, si ottiene una stima della dimensione circa attorno a 38. Lo stesso vale considerando solo le lingue Indo-Europee. La dimensione scende se ci si restringe alle famiglie linguistiche del Niger-Congo (23), Austro-Asiatiche (12) e Afro-Asiatiche (8) (Figura 3.1). Per ulteriori informazioni si fa riferimento alla pubblicazione [28]



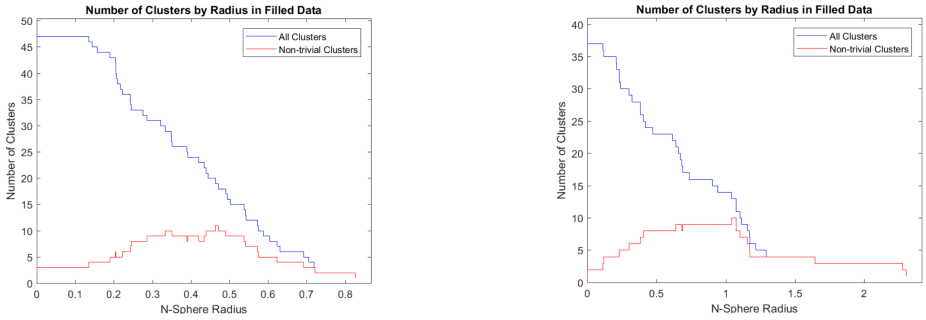
(a) intero database (b) lingue Indo-Europee (c) lingue Niger-Congo

(d) lingue Austro-Asiatiche (e) lingue Afro-Asiatiche

Figura 3.1: dimensione dello spazio delle lingue utilizzando il database SSWL

3.4 Esempio di analisi: lingue Indo-Europee

La famiglia Indo-Europea è la più consistente a livello numerico e di precisione, in entrambi i database. Inoltre è quella con il maggior numero di clusters non banali (Figura 3.2), da cui si può dedurre che le caratteristiche sintattiche sono distribuite meno omogeneamente.



(a) SSWL (b) LanGeLin

Figura 3.2: Numero di clusters in base al raggio per i dati sulle lingue Indo-Europee, con varianza percentuale per la PCA di 60%

In base ad ulteriori studi si è visto che il database di Longobardi fornisce risultati più simili agli alberi filogenetici già conosciuti dagli storici. Invece, SSWL ha troppe informazioni per poter essere usato interamente, quindi una volta fatta l'analisi con l'intero dataset, si procede ad analizzare le sottofamiglie per ottenere una maggiore accuratezza. Tuttavia permangono dei mal-posizionamenti dovuti alla presenza di lingue non sufficientemente mappate e a variabili incomplete. Per ovviare a questo problema si filtrano i dati mantenendo solo le lingue che sono complete almeno al 50%, e si restringe l'insieme alle variabili che sono completamente mappate in tutte le lingue rimaste.

Successivamente si calcola l'albero delle componenti persistenti con il metodo spiegato in precedenza, e si ottiene il risultato riportato nella Figura 3.3.

Si possono osservare alcuni particolari:

Cluster 52 Afrikaans, Olandese, Tedesco, e Fiammingo Occidentale: appartengono al ramo delle lingue Germaniche Occidentali. Riflette la struttura dell'albero filogenetico corrispondente tranne per l'Inglese che è stato posizionato nel Cluster 60 assieme all'Inglese di Singapore.

Cluster 80 Gotico, Islandese e Tardo Latino: è un indicatore di come questa procedura tenda a sbagliare il posizionamento delle lingue antiche

Cluster 70 Calabrese del Nord, Rumeno, Greco Cipriota, Greco, e Albanese: non corrisponde per niente alle informazioni storiche poiché contiene lingue sia Elleniche che Romanze

Tuttavia, come già visto, questi errori non sono dovuti ai dati del database ma ad una serie di altri fattori (PCA, incompletezza dei dati, omoplasia, significato diverso delle relazioni rispetto all'evoluzione naturale del linguaggio).

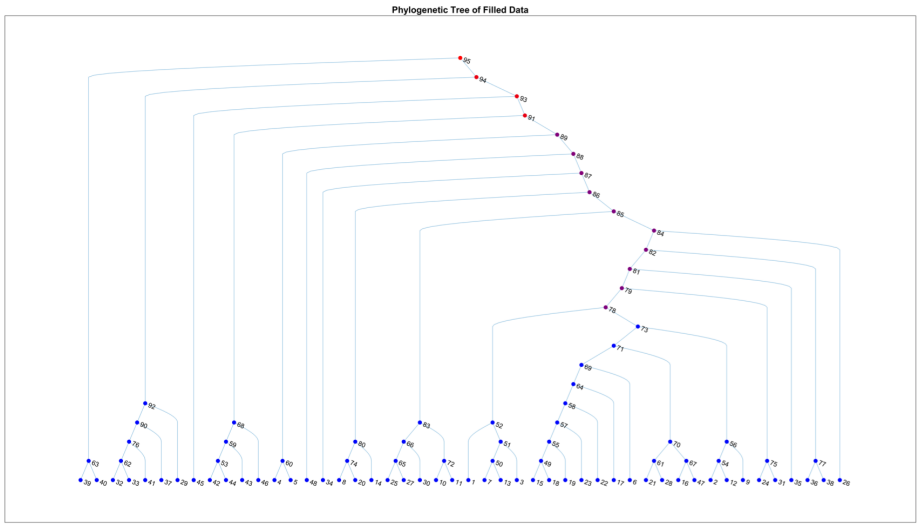


Figura 3.3: albero delle componenti persistenti della famiglia di lingue Indo-Europee sulla base dei dati di SSWL filtrati

Per quanto riguarda il dataset di Longobardi si ottengono dei risultati un po' diversi, come si può vedere nella Figura 3.4

In questo caso si può osservare che:

Cluster 60 si suddivide a sua volta in due grandi clusters (59 e 55) che contengono le lingue Romanze rispettivamente moderne (Italiano, Spagnolo, Francese, Portoghese, e Rumeno) e i dialetti del Sud Italia (Ragusa, Mussomeli, Aidone, Calabrese del Sud, Salentino, Calabrese del Nord, e Campano)

Cluster 72 lingue Elleniche, Germaniche e Slave: questo raggruppamento è attribuibile alla struttura di H_1

Cluster 67 Inglese Antico, Inglese, Olandese, Danese, Islandese, e Norvegese: anche in questo caso vediamo che l'Islandese non appartiene al gruppo delle lingue Nord-Germaniche ma a quelle Occidentali.

In questo caso però c'è una maggiore correlazione tra albero "persistente", suddivisione in clusters, e albero filogenetico. Infatti si può notare come le lingue antiche (ad esempio Gotico e Inglese Antico) siano correttamente posizionate nel sottogruppo delle lingue Germaniche occidentali.

La stessa procedura è stata fatta per le altre famiglie che non contengono lingue Indo-Europee, anche se nel database LanGeLin non sono quasi presenti, mentre nel SSWL sono meno complete. Quindi le famiglie Niger-Congo, Austro-Asiatiche e Afro-Asiatiche sono state indagate con i dati non filtrati di SSWL. Anche in questi casi si verificano degli errori nel posizionamento di alcune lingue. Un'analisi più approfondita si può trovare in [28].

Per quanto riguarda la struttura di H_1, considerando i dati filtrati di SSWL che si sono usati finora, si osserva un generatore persistente (con una piccola

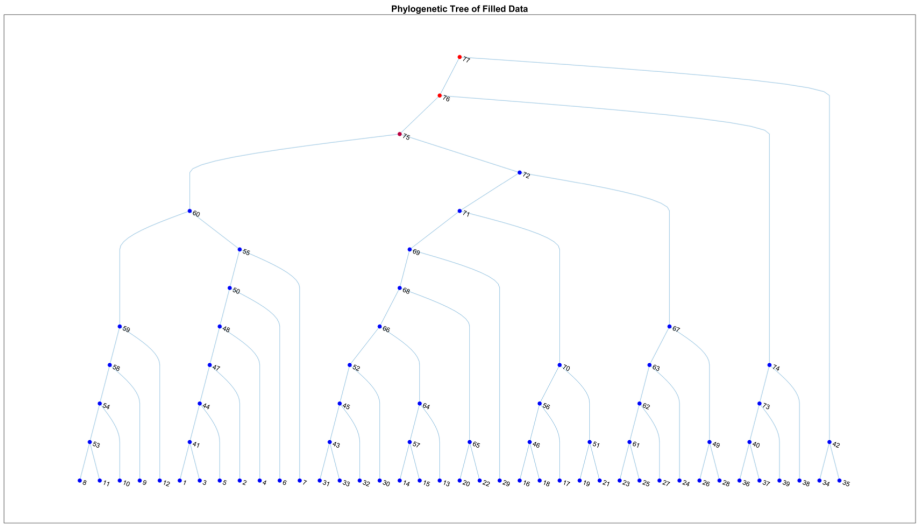


Figura 3.4: albero delle componenti persistenti della famiglia di lingue Indo-Europee sulla base dei dati di LanGeLin

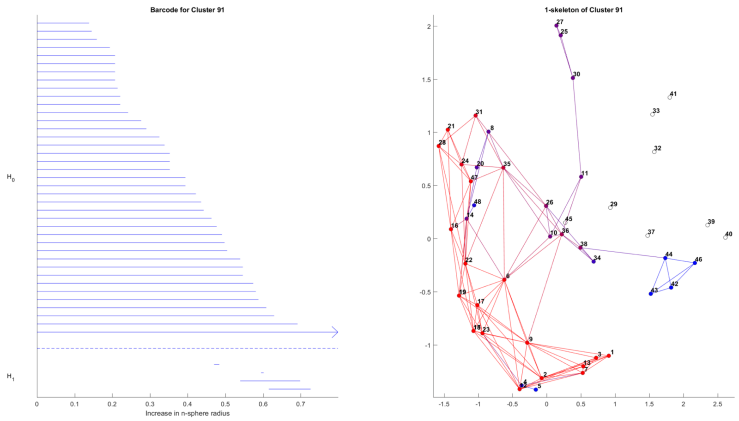
persistenza) a livello del Cluster 78, altri due più significativi emergono al numero 91 (Figura 3.5a), e un ultimo generatore si ritrova nel Cluster 95 (Figura 3.5b).

Per quanto riguarda i dati relativi al database LanGeLin, anche in questo caso sono presenti tre generatori persistenti di H_1 nel Cluster 77 (Figura 3.6b), con il primo che emerge al numero 75 (Figura 3.6a), e gli altri due al 76.

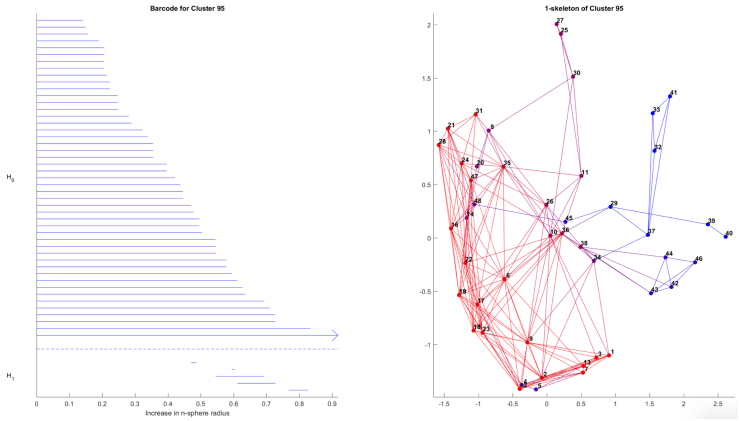
Infatti osservando il barcode delle lingue Indo-Europee che sfrutta i dati di SSWL (Figura 3.7) si possono vedere 2 generatori persistenti di H_0, che si possono ricondurre ai rami delle lingue Europee e Indo-Iraniane. Si nota anche un generatore persistente di H_1. La prima motivazione a cui si potrebbe pensare è il ponte Anglo-Normanno tra Francese e Inglese, ma questa è una correlazione lessicale e non sintattica. Con delle analisi più approfondite (come spiegato in precedenza) si giunge alla conclusione che il generatore è dovuto alle lingue greche antiche.

In questa fase è possibile trovare anche un esempio di omoplasia: l'Olandese Medio e il Tedesco Svizzero sono fortemente correlate, ma fanno parte di un ciclo non banale con il Lituano Balto-Slavo e il Ceco. Questa relazione non è storicamente giustificata da interazioni tra le lingue riportate, quindi l'unica spiegazione è che questo rappresenti appunto un caso di omoplasia.

Si rimanda a [26] per un testo che affronta l'argomento degli alberi filogenetici delle lingue Indo-Europee, in cui vengono riportati i risultati dovuti all'utilizzo di diversi database.



(a) Cluster 91



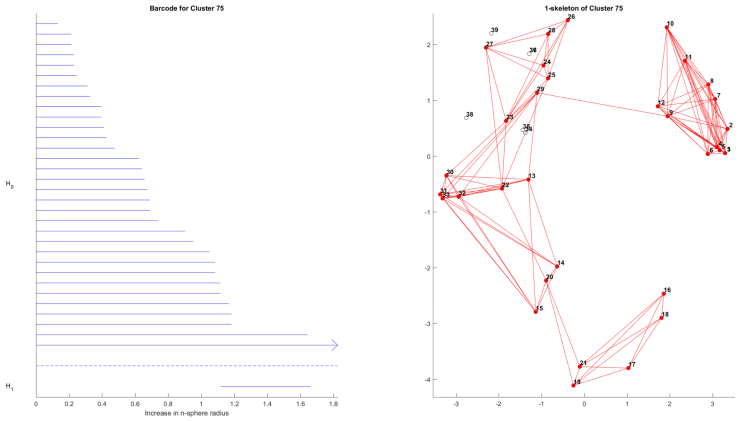
(b) Cluster 95

Figura 3.5: Struttura di H_1 persistente nei dati di SSWL filtrati delle lingue Indo-Europee e Ural-Altaiche, PCA60%, cluster 91 e 95

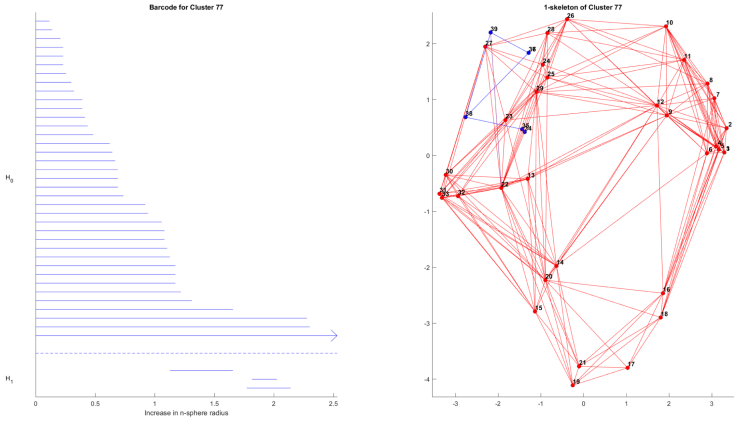
3.5 Confronti con altri metodi

L'analisi che si è riportata finora, come già detto, pone le sue basi su studi precedenti che affrontano lo stesso argomento con approcci diversi. I contributi più consistenti derivano dall'analisi di tipo algebrico-geometrico. Questa sfrutta gli stessi database che sono stati utilizzati con il metodo topologico, ma contemporaneamente. Alcuni dei vantaggi di questo approccio derivano dal fatto che all'albero venga assegnata una geometria, e quindi l'oggetto risultante conterrà più informazioni rispetto a quello ottenuto con altri metodi, ed inoltre è applicabile non solo agli alberi binari ma anche alle reti. Per ulteriori informazioni in merito ai dettagli del procedimento e ai risultati ottenuti si rimanda a [31].

Ciò che è interessante riportare è il fatto che, come si può vedere anche in [28], l'approccio algebrico-geometrico conduce a migliori risultati specialmente quando vengono prese in considerazione informazioni aggiuntive, non solo con variabili totalmente mappate. Infatti, anche nei casi delle lingue Germaniche o



(a) Cluster 75



(b) Cluster 77

Figura 3.6: Struttura di H_1 persistente nei dati di Longobardi delle lingue Indo-Europee e Ural-Altaiche, PCA60%, cluster 75 e 77

di quelle Romanze, i risultati ottenuti con questo procedimento riflettono l'evoluzione filogenetica storica, contrariamente all'albero persistente che colloca alcune lingue in posizioni errate.

Un'altra interessante modalità è riportata in [20]. In questo caso si fa un confronto tra analisi filogenetica basata su dati sintattici e su dati linguistici, e si può vedere come i risultati possano essere sovrapponibili anche se con minime differenze. Il metodo di comparazione parametrica (PCM) che viene utilizzato, sfrutta la distanza di Hamming (il numero di bit che hanno un valore diverso nelle due stringhe che si stanno comparando) come misura di quanto siano in relazione due lingue. Si è visto infine che i risultati ottenuti lavorando sulle lingue Indo-Europee sono affidabili.

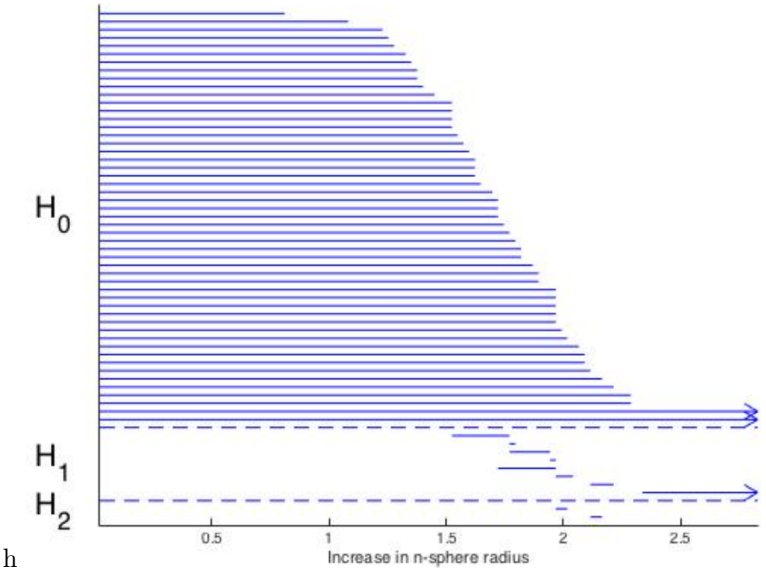


Figura 3.7: Barcode relativo alle lingue Indo-Europee sui dati SSWL

Sul modello dei Principi e Parametri sviluppato da Chomsky si basano ulteriori pubblicazioni come ad esempio [22, 25].

In [22] viene proposto un metodo per stimare quantitativamente l'entropia e la complessità delle famiglie linguistiche senza doverle calcolare per le singole lingue. Questo si basa sul metodo di comparazione parametrica e sulla teoria matematica dei codici di correzione degli errori [33] con lo scopo di unire le due teorie. Viene inoltre definita la funzione limite asintotico che permette di stabilire quanto siano diffusi o diversificati i parametri nella famiglia linguistica, il tutto stimando la posizione del punto bidimensionale relativo ad uno dei codici (stringa rappresentante una lingua), rispetto al limite asintotico.

In [25], invece, è sviluppato un metodo per studiare le relazioni di dipendenza tra parametri sintattici basandosi sulla memoria distribuita a bassa densità (*sparse distributed memory* o SDM). Si usa come punto di partenza la rete di Kanerva, ovvero un modello sviluppato da Penti Kanerva nel 1988 come modello matematico per la memoria umana a lungo termine. Qui i pensieri, le percezioni, le esperienze e anche i concetti sono rappresentati da vettori in uno spazio multidimensionale.
Questo risulta utile perché, nell'ottica di comprendere il meccanismo di acquisizione di una lingua e di come le strutture sintattiche vengono immagazzinate nel cervello umano, tale tipo di memorie sembra essere un ottimo candidato per la costruzione di modelli computazionali efficienti. Tuttavia servono ulteriori studi per confermare che la vicinanza sintattica individuata dalla rete di Kanerva corrisponda ad una prossimità storica delle lingue esaminate.

Altrettanto interessante è il modo in cui un modello fisico viene applicato ai dati dei parametri sintattici binari: lo *spin glass model* il cui utilizzo nella

linguistica viene sviluppato in [32, 30]. Lo spin è definito come l'orientazione dei poli magnetici (nord e sud) nello spazio tridimensionale. In questo contesto si considerano le lingue ai vertici di un grafo, in cui sono assegnate delle energie di interazione lungo i lati, e i parametri binari vengono dotati di spin. L'ipotesi necessaria è che un'alta energia di interazione tende a far allineare i parametri, come avviene con i materiali ferromagnetici, e questa energia è data dalla frequenza delle interazioni che avvengono tra le lingue (nell'MIT Media Lab si considera avvenuta un'interazione tra due lingue quando un utente è propenso ad editare un articolo di wikipedia in entrambe). In seguito si approfondisce l'evoluzione computazionale del modello, durante il quale l'energia di interazione rimane costante, e si fa costruendo un processo di Markov.
Il lavoro prosegue studiando la dinamica indotta sullo spazio dei parametri dal modello dei vetri di spin: la posizione del codice in questo spazio può essere vista come una misura di quanto le strutture sintattiche siano distribuite tra le lingue dell'insieme su cui viene calcolato. Inoltre, la posizione relativa ad alcune curve di riferimento fornisce informazioni sull'entropia e sulla complessità dell'insieme.

Per il metodo nel nucleo del calore riportato in [24], invece, si costruisce un grafo di prossimità e con l'algoritmo di Belkin Niyogi [2, 3] si assegnano dei pesi basandosi sulle soluzioni del nucleo del calore. In questo modo si dà più importanza alle variabili propense ad essere indipendenti tra loro. L'idea di base è quella di usare il laplaciano del grafo, tramite i suoi autovalori e autovettori, per ottenere una rappresentazione di dimensione minore rispetto a quella iniziale ma che mantenga le informazioni di prossimità. Molto spesso i risultati della suddivisione in clusters coincidono con quelli topologici, eventualmente con qualche posizionamento diverso.

Un ulteriore modo che viene impiegato per l'analisi linguistica è il machine learning [16], e anch'esso si basa sulla sintassi. Nel pratico vengono creati dei modelli delle dipendenze tra i parametri sintattici e vengono identificati eventuali gruppi di parametri i cui valori permettono di fare la suddivisione in famiglie linguistiche. Si cerca inoltre di costruire dei grafi per diminuire la dimensione dello spazio delle possibili grammatiche.

Infine si può vedere in [4] come lo stesso tipo di analisi dal punto di vista dell'omologia persistente si possa fare anche mediante l'uso di funzioni filtranti. Tuttavia in questo contesto non si hanno ancora i mezzi per dare un'interpretazione efficace dei risultati.

Conclusione

La presente tesi è servita per dimostrare come è possibile applicare strumenti matematici e computazionali, quali l'omologia persistente, i barcodes, gli alberi binari e l'analisi delle componenti principali, ad una questione lontana come quella dell'evoluzione linguistica. Sono stati presentati gli studi in merito al ruolo dei gruppi di omologia H_0 e H_1 nella costruzione degli alberi filogenetici, con un focus particolare sulle lingue Indo-Europee. Infine si sono riportati ulteriori metodi impiegati per lo stesso scopo conoscitivo.

Si possono quindi trarre delle conclusioni in merito all'apporto che ha la topologia nell'analisi linguistica:

- la topologia riesce a riconoscere dei fenomeni storici-linguistici già conosciuti dai linguisti, come ad esempio la suddivisione delle lingue in famiglie e sotto-famiglie;
- i diagrammi a barre del gruppo di omologia persistente H_0 riportano la suddivisione appena citata, e sono utili nel confronto con quanto storicamente ritenuto veritiero;
- la struttura di H_1 risulta più sensibile a fenomeni meno plateali che non vengono considerati nella struttura degli alberi filogenetici, come quanto visto per le influenze tra sotto-famiglie; si è visto che per alcune di queste interazioni si ritrova una possibile spiegazione nell'omoplasia.

Ad ogni modo rimangono dei punti da chiarire. Ad esempio, ci si chiede che interpretazioni si possano dare ad un generatore persistente di H_1 da un punto di vista storico-linguistico. Oppure quanto l'omologia persistente possa descrivere la distribuzione dei parametri sintattici nelle diverse famiglie linguistiche. Infine rimangono aperte le questioni delle lingue Ural-Altaiche e del loop Gotico-Slavo-Greco, le quali non hanno ancora trovato una spiegazione coerente.

Bibliografia

[1] Mark C Baker. *The atoms of language: The mind's hidden rules of grammar*. Basic books, 2001.

[2] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems*, 14, 2001.

[3] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[4] Mattia G Bergomi, Massimo Ferri, and Antonella Tavaglione. Steady and ranging sets in graph persistence. *arXiv preprint arXiv:2009.06897*, 2020.

[5] Hervé Bourlard and Samy Bengio. Hidden markov models and other finite state automata for sequence processing. Technical report, IDIAP, 2001.

[6] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[7] Noam Chomsky and Howard Lasnik. The theory of principles and parameters. In *Syntax*, pages 506–569. De Gruyter Mouton, 2008.

[8] Herbert Edelsbrunner and John Harer. *Computational topology: an introduction*. American Mathematical Soc., 2010.

[9] Massimo Ferri. Persistent topology for natural data analysis—a survey. In *Towards integrative machine learning and knowledge extraction*, pages 117–133. Springer, 2017.

[10] Sitanshu Gakkhar and Matilde Marcolli. Syntactic structures and the general markov models. *arXiv preprint arXiv:2104.08462*, 2021.

[11] Robert Ghrist. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

[12] Allen Hatcher. *Algebraic topology*. Tsinghua University Press Co., Ltd, 2005.

[13] Giorgio Israel. *La visione matematica della realtà: introduzione ai temi e alla storia della modellistica matematica*. GLF Editori Laterza, 2012.

[14] Richard Arnold Johnson, Dean W Wichern, et al. *Applied multivariate statistical analysis*, volume 6. Pearson London, UK:, 2014.

[15] Lida Kanari, Adélie Garin, and Kathryn Hess. From trees to barcodes and back again: theoretical and statistical perspectives. *Algorithms*, 13(12):335, 2020.

[16] Dimitar Lubomirov Kazakov, Guido Cordoni, Eyad Algahtani, Andrea Ceolin, Monica-Alexandrina Irimia, Shin-Sook Kim, Dimitris Michelioudakis, Nina Radkevich, Cristina Guardiano, and Giuseppe Longobardi. Learning implicational models of universal grammar parameters. 2018.

[17] Giuseppe Longobardi. Methods in parametric linguistics and cognitive history. *Linguistic Variation Yearbook*, 3(1):101–138, 2003.

[18] Giuseppe Longobardi. Principles, parameters, and schemata: a radically underspecified ug. *Linguistic Analysis*, pages 517–558, 2018.

[19] Giuseppe Longobardi and Cristina Guardiano. Evidence for syntax as a signal of historical relatedness. *Lingua*, 119(11):1679–1706, 2009.

[20] Giuseppe Longobardi, Cristina Guardiano, Giuseppina Silvestri, Alessio Boattini, and Andrea Ceolin. Toward a syntactic phylogeny of modern indo-european languages. *Journal of Historical Linguistics*, 3(1):122–152, 2013.

[21] Matilde Marcolli. Geometry of phylogenetic inference. 2015.

[22] Matilde Marcolli. Syntactic parameters and a coding theory perspective on entropy and complexity of language families. *Entropy*, 18(4):110, 2016.

[23] Panza Marco Molinini, Daniele. Sull'applicabilità della matematica. *La Matematica nella Società e nella Cultura. Rivista dell'Unione Matematica Italiana*, 7(3):367–395, 12 2014.

[24] Andrew Ortegaray, Robert C Berwick, and Matilde Marcolli. Heat kernel analysis of syntactic structures. *Mathematics in Computer Science*, pages 1–18, 2021.

[25] Jeong Joon Park, Ronnel Boettcher, Andrew Zhao, Alex Mun, Kevin Yuh, Vibhor Kumar, and Matilde Marcolli. Prevalence and recoverability of syntactic parameters in sparse distributed memories. In *International Conference on Geometric Science of Information*, pages 265–272. Springer, 2017.

[26] Asya Pereltsvaig and Martin W Lewis. *The Indo-European Controversy.* Cambridge University Press, 2015.

[27] Alexander Port, Iulia Gheorghita, Daniel Guth, John M Clark, Crystal Liang, Shival Dasu, and Matilde Marcolli. Persistent topology of syntax. *Mathematics in Computer Science*, 12(1):33–50, 2018.

[28] Alexander Port, Taelin Karidi, and Matilde Marcolli. Topological analysis of syntactic structures. *arXiv preprint arXiv:1903.05181*, 2019.

[29] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[30] Kevin Shu and Matilde Marcolli. Syntactic structures and code parameters. *Mathematics in Computer Science*, 11(1):79–90, 2017.

[31] Kevin Shu, Andrew Ortegaray, Robert C Berwick, and Matilde Marcolli. Phylogenetics of indo-european language families via an algebro-geometric analysis of their syntactic structures. *Mathematics in Computer Science*, pages 1–55, 2021.

[32] Karthik Siva, Jim Tao, and Matilde Marcolli. Spin glass models of syntax and language evolution. *arXiv preprint arXiv:1508.00504*, 2015.

[33] Scott A Vanstone and Paul C Van Oorschot. *An introduction to error correcting codes with applications*, volume 71. Springer Science & Business Media, 2013.