UNIVERSITÀ DEGLI STUDI DI GENOVA

SCUOLA DI SCIENZE MATEMATICHE, FISICHE E NATURALI

CORSO DI LAUREA MAGISTRALE IN MATEMATICA

TESI DI LAUREA

Approcci basati sulla teoria dei grafi per confronti multiclasse in connettività cerebrale strutturale

Relatrici:
Dott.ssa Sara Garbarino
Dott.ssa Sara Sommariva

Correlatrice:
Prof.ssa Cristina Campi

Candidato:
Andrea Pedemonte

Anno accademico 2023/2024

Indice

Introduzione

I temi di ricerca che vengono affrontati in questa Tesi trovano collocazione nell'ambito della matematica applicata alla medicina e, in particolare, nello studio della connettività cerebrale in persone affette da malattie neurodegerative. Si tratta di malattie caratterizzate da un peggioramento cognitivo costante nel tempo, come, ad esempio, la malattia di Alzheimer, la sclerosi multipla e il morbo di Parkinson.

L'interesse verso questo ramo deriva da una sempre maggiore collaborazione tra la matematica e le scienze mediche. Infatti l'introduzione di tecniche statistiche in aree di competenza medica sta diventando di fondamentale aiuto al campo delle suddette discipline, nella costante ricerca di diagnosi precoci e personalizzate dei pazienti.

Il punto di partenza dello studio é l'analisi del cervello umano, costituito da unità fondamentali, i nueuroni, i quali, connettendosi tra loro, contribuiscono a costruire una rete particolarmente complessa.

A partire dalla seconda metà del secolo scorso la necessità di studiare e analizzare queste tipologie di strutture ha portato a sviluppare ed approfondire le conoscenze sulle reti per poter comprendere nel dettaglio caratteristiche e funzioni degli elementi costituenti del cervello.

Vari studi han portato alla consapevolezza che i neuroni interagiscono tra loro, cooperando e presentando alcune similitudini. In particolar modo si é capito che sistemi complessi differenti condividono alcuni principi organizzativi chiave, i quali possono essere quantificati dagli stessi parametri. Diventa quindi necessario andare a studiare queste similarità per poter comprendere meglio struttura e funzionamento del cervello.

Il lavoro di questa Tesi si concentra sull'esplorazione delle reti di connettività cerebrale dal punto di vista strutturale, sfruttando la teoria dei grafi, all'interno della quale una rete viene descritta come un insieme di punti, detti nodi o vertici, connessi tra loro tramite archi o spigoli [6].

Per questo motivo, nel primo capitolo, vengono fornite le nozioni base sui grafi e vengono definite diverse metriche di connettività. Quest' ultime sono misure applicabili alle matrici di adiacenza associate ai grafi al fine di analizzare quantitativamente proprietà inerenti l'intero grafo e le sue sottounità, ossia i nodi e gli archi.
Il secondo capitolo si sposta invece sull'ambito statistico, con una descrizione di tutti i test statistici più importanti che si possono applicare sui risultati ottenuti dalle suddette metriche, al fine di confronatare la connettività in diversi gruppi di soggetti. In questa sezione diventa fondamentale la distinzione tra le metriche che studiano la rete nel suo complesso, ovvero quelle globali, per cui é possibile applicare test univariati, e quelle locali, che agiscono, invece, sui singoli nodi o archi del grafo, costringendo all'utilizzo di test multivariati più sofisticati.
Nell'ultimo capitolo, infine, i concetti teorici presentati nei primi due vengono applicati a dati reali, con l'obiettivo di validarne l'efficacia e la rilevanza pratica. In particolare, sono stati analizzati i dati di 164 pazienti, suddivisi in quattro gruppi a seconda del disturbo cognitivo diagnosticato. Attraverso l'utilizzo delle metriche di connettività e l'applicazione dei test statistici appropriati, si è cercato di individuare differenze significative tra i gruppi di pazienti, al fine di trovare nuove chiavi interpretative per la comprensione delle malattie neurodegenerative.

Capitolo 1

Teoria dei grafi per la connettività

In questo primo capitolo si vuole fornire una descrizione teorica e dettagliata delle metriche di connettività. Questo termine si riferisce a delle misure che descrivono quanto e come i nodi di un grafo sono interconnessi tra loro. Per questo motivo il capitolo presenta una prima sezione in cui vengono fornite nozioni base relative al concetto di grafo. Nella seconda sezione vengono descritte le proprietà delle categorie in cui si possono suddividere le metriche di connettività, con un approfondimento per le più rilevanti di ciascun gruppo, con esempi annessi. Infine il capitolo termina con l'approfondimento di una rete più complessa, priva di scala e avente distribuzione dei gradi in legge di potenza: il modello di Barabàsi-Albert.

1.1 Nozioni base sui grafi

Definizione 1.1.1. *Un grafo é una coppia $G = (V, E)$, nella quale V é un insieme finito non vuoto di elementi, detti vertici o nodi, ed E é un insieme finito di coppie non ordinate di V, detti archi o spigoli [19].*

Generalmente gli archi vengono indicati come coppie (u, v) e, nel caso in cui $u = v$, l'arco é detto cappio. Nel caso in cui un arco sia presente più di una volta in E si parla di arco multiplo; viceversa si parla di arco semplice.
Vengono ora fornite le definizioni di alcune caratteristiche che possono avere i grafi, le quali saranno utili in seguito.

Definizione 1.1.2. *Un grafo $G = (V, E)$ si dice orientato se gli archi (u, v) sono coppie ordinate, cioé hanno una direzione.*

Definizione 1.1.3. *Un grafo $G = (V, E)$ si dice pesato se ad ogni arco (u, v) é attribuito un valore numerico che descrive la rilevanza della connessione tra due nodi.*

Definizione 1.1.4. *Un grafo $G = (V, E)$ si dice completo se esiste un arco tra ogni coppia di nodi.*

Nella figura 1.1 viene fornita una rappresentazione grafica di un grafo indiretto non pesato.

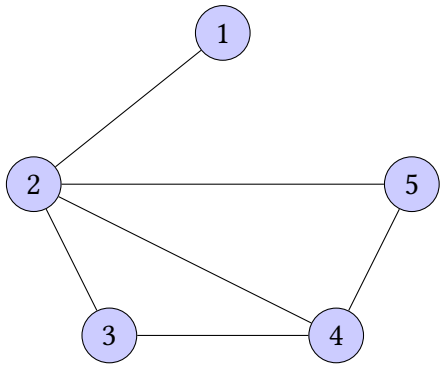


Figura 1.1: Esempio di grafo indiretto non pesato e non completo.

Tale grafo é costituito da 5 vertici e da 6 archi, con l'insieme dei nodi $V = \{1, 2, 3, 4, 5\}$ e con l'insieme degli spigoli $E = \{(1, 2), (2, 3), (3, 4), (4, 5), (5, 3), (5, 2)\}$.
Un concetto importante da introdurre, che verrà ampiamente utilizzato nell'applicazione della teoria dei grafi alle reti neurali, é quello di grado.
Per farlo é necessario fornire la definizione di arco incidente ad un nodo.

Definizione 1.1.5. *Un arco (u, v) si dice incidente ad un nodo se tale nodo è uno dei due estremi dell'arco*

Segue la definizione di grado di un vertice:

Definizione 1.1.6. *Il numero di archi incidenti in un nodo v é detto grado di v.*

Nella figura 1.1, ad esempio, il nodo 5 ha grado 3, in quanto sono 3 gli archi ad esso incidenti.
Se un vertice ha grado 0 si dice isolato, mentre se ha grado 1 é detto terminale.
Nel caso di grafo orientato é possibile fare una distinzione dei gradi in due tipologie:

- grado In: numero di archi diretti che arrivano in quel nodo;

- grado Out: numero di archi diretti che partono da quel nodo.

Nel caso in cui ogni vertice del grafo abbia lo stesso grado, il grafo si dice regolare. Ad esempio, se ogni nodo ha grado r, il grafo è regolare di grado r o r-regolare. Definito il concetto di grado di un nodo é possibile descrivere una rappresentazione del grafo sotto forma di una matrice, detta matrice di adiacenza:

Definizione 1.1.7. *La matrice di adiacenza A è una matrice quadrata in cui ogni entrata descrive la presenza o l'assenza di un arco tra due nodi del grafo.*

Tale matrice, che verrà approfondita nel terzo capitolo, può essere binaria o pesata a seconda che si consideri solo la presenza o assenza di un link o la forza di tale connessione. Descritte le proprietà base di un grafo é possibile introdurre il concetto di sottografo mediante la seguente definizione:

Definizione 1.1.8. *Sia $G = (V, E)$ un grafo con un insieme di vertici V e un insieme di archi E. Un grafo $H = (V', E')$ si dice sottografo di un grafo G se e solo se:*

1. *V' è un sottoinsieme di V;*
2. *E' è un sottoinsieme di E;*
3. *Per ogni arco e in E' i suoi vertici sono in V'.*

La terza condizione indica quindi che se uno arco fa parte di H, allora anche i vertici che collega devono far parte di H.
Generalmente un sottografo si ottiene andando a rimuovere determinati vertici e spigoli dal grafo originario, come riportato nella figura 1.2.

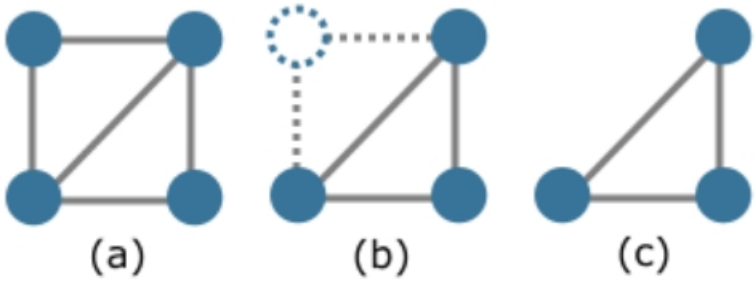


Figura 1.2: Rimozione di alcuni vertici o archi (b) dal grafo (a) per ottenere il sottografo (c). Immagine presa da *symbio6.nl*.

Risulta infine essere utile introdurre il concetto di cammino e cammino semplice:

Definizione 1.1.9. *Un cammino di un grafo é una sequenza di vertici tale che ogni coppia consecutiva sia collegata da un arco* (u, v) *del grafo.*

Definizione 1.1.10. *Un cammino di un grafo si dice semplice se tutti i vertici che lo compongono sono distinti, eccetto, eventualmente, il primo e l'ultimo se coincidono.*

Date tali definizione é possibile descrivere la lunghezza di un cammino come segue:

Definizione 1.1.11. *Il numero di archi* (u, v) *che compongono un cammino é detto lunghezza del cammino.*

Da tale definizione segue che se un cammino é costituito da n archi, la sua lunghezza sarà $n - 1$.
Infine é possibile dare una definizione di connettività legata alla teoria dei grafi:

Definizione 1.1.12. *La capacità di raggiungere un nodo partendo da un altro vertice del grafo, tramite uno solo o più archi, é detta connettività.*

1.2 Metriche di connettività

Dopo aver descritto le proprietà base di un grafo é possibile introdurre le metriche di connettività: strumenti che descrivono come e quanto i diversi nodi di un grafo sono interconnessi tra loro.
L'obiettivo di questa sezione é quello di descrivere le principali categorie in cui si possono raggruppare queste metriche, fornendo poi un approfondimento di alcune di esse, le quali saranno oggetto di studio nei capitoli successivi. La sezione termina infine con un'applicazione pratica di tali metriche su un grafo, per facilitare la compresensione del loro utilizzo.
Le metriche di connettività si possono, innanzitutto, suddividere, a seconda delle caratteristiche, in quattro diverse categorie: *centralità*, *segregazione*,*integrazione* e *robustezza*.

1.2.1 Centralità

La prima sezione che viene descritta é quella relativa alle metriche di centralità [4]. Queste misure valutano l'importanza o l'influenza di un nodo o di un arco all'interno della rete. Tra le più rilevanti vi sono il *grado*, la *betweenness centrality*, detta *edge betweenness*, se applicata agli archi, e la *centralità degli autovalori*.

Grado

Il grado è una misura di centralità che viene calcolata a partire dai legami tra i diretti vicini di un vertice, valutando il numero di connessioni che ogni nodo presenta con gli altri vertici. Nella teoria dei grafi si calola come segue.
Si consideri il generico nodo $i \in V$ con $G = (V, E)$ un grafo, il grado di tale vertice é:

$$k_i = \frac{1}{N-1} \sum_{j=1}^{|V|} a_{ij}, \tag{1.1}$$

con N il numero totale dei vertici e a_{ij} l'elemento (i, j) della matrice di adiacenza A.

Betweenness centrality

La betweenness centrality (BC) è una misura che indica quanto un nodo sia in grado di controllare il flusso di informazioni tra gli altri nodi della rete. Un valore elevato indica che il vertice ha una grande capacità di facilitare l'interazione tra gli altri nodi, il che significa che la rete senza tale nodo si dividerebbe in sottoreti.
Si calcola mediante la seguente formula:

$$BC(v_i) = \frac{\sum_{i \neq j \neq k} \sigma_{i,j}(v_k)}{\sum_{j=1}^{N} \sigma_{i,j}},$$

con N il numero di vertici, $\sigma_{i,j}$ il numero totale di percorsi geodetici, ovvero i percorsi più brevi che collegano due vertici u e v in un grafo G.
La distanza tra essi, indicata con $d(u, v)$, é detta distanza geodetica.

Centralità degli autovalori

La centralità degli autovalori (EC) descrive l'idea per cui un nodo è tanto più centrale quanto più è connesso a nodi importanti, ossia aventi molti legami. La centralità del nodo è quindi proporzionale alla centralità dei nodi a cui è connesso, ossia è proporzionale alla sua posizione vicino ai nodi o alle comunità più significative di un grafo. Sia $A = \{a_{i,j}\}$ matrice di adiacenza del grafo, la centralità di autovalore x_i relativa al nodo i si calcola:

$$x_i = \frac{1}{\lambda} \sum_{k=1}^{N} a_{ki} x_k, \tag{1.2}$$

con $\lambda \neq 0$ costante. In forma matriciale si può scrivere:

$$\lambda x = xA$$

Quindi il vettore EC x è l'autovettore di sinistra di A associato all'autovalore λ. La scelta migliore per λ è il più grande autovalore della matrice dei valori assoluti di A. Questa metrica presenta tuttavia alcune problematiche. Infatti é possibile avere valori nulli per un nodo in caso di assenza di legami in entrata in (1.2), il che porta a un contributo nullo della metrica di centralità degli altri nodi.

1.2.2 Segregazione

Le metriche di segregazione sono misure che indicano quanto la rete è divisa in gruppi o comunità. Le più note sono il *coefficiente di clustering*, la *transitività dei percorsi* e la *modularità* [3].

Coefficiente di clustering

I nodi più vicini di una rete si possono raggruppare tra loro andando a formare un *cluster*. Per quantificare il numero di connessioni esistenti tra questi nodi vicini in proporzione al numero massimo di connessioni possibili viene calcolato il coefficiente di clustering. Generalmente le reti *random*, ossia reti in cui le connessioni tra nodi sono casuali, hanno una clusterizzazione bassa, mentre quelle complesse hanno una clusterizzazione elevata, dovuta alla densità di collegamenti relativamente alta. Esistono due tipologie di coefficienti di clustering: locale e globale.
Per descrivere la prima tipologia sia $E = \{e_{ij}\}_{i,j=1}^{n}$ l'inieme dei collegamenti dal veritice v_i al vertice v_j, $N_i = \{v_j : (e_{ji}, e_{ij}) \in E^2\}$ l'insieme dei nodi direttamente connessi a un vertice v_i e sia $k_i = |N_i|$.
Il coefficiente di clustering locale per grafi diretti é:

$$C_i = \frac{2 \cdot |\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}, \tag{1.3}$$

con $\frac{k_i(k_i-1)}{2}$ il numero di collegamenti possibili tra i membri della rete.

Transitività del percorso

La transitività del percorso valuta la probabilità che due nodi connessi a un terzo nodo siano connessi tra di loro. La transitività perfetta implica che, se x è connesso (attraverso un arco) a y e y è connesso a z, allora x è connesso anche a z.

Questa proprietà può essere quantificata come segue: un loop di lunghezza tre è una sequenza di nodi x, y, z, x tale che (x, y), (y, z) e (z, x) sono archi del grafo e formano il più piccolo ciclo possibile in un grafo indiretto che inizia e finisce nello stesso nodo. Un percorso di lunghezza due è una sequenza di nodi x, y, z tale che (x, y) e (y, z) sono spigoli (lo spigolo (z, x) può essere presente o meno). Si osservi che il loop x, y, z, x è diverso da y, z, x, y. Il coefficiente di transitività T di una rete, è quindi il rapporto tra il numero di loop di lunghezza tre e il numero di percorsi di lunghezza due. Rappresenta pertanto la frequenza dei loop di lunghezza tre nella rete.
Ovvero in formule si ha che:

$$T = \frac{|\text{numero loop di lunghezza 3}|}{|\text{numero di cammini di lunghezza 2}|}$$

I due casi limite sono:

- $T = 1$: perfetta transitività;
- $T = 0$: assenza di percorsi di lunghezza 3.

Modularità

La modularità é invece utilizzata per valutare la forza della divisione della rete in moduli o comunità, ovvero in gruppi. Si ricava mediante la seguente formula:

$$Q = \frac{1}{2|E|} \sum_{i,j} \left(a_{ij} - \gamma \frac{k_i k_j}{2|E|} \right) \delta(v_i, v_j) \tag{1.4}$$

con k_i il grado pesato del nodo i, γ il parametro di risoluzione, ovvero un valore che serve per controllare quanto grandi o piccole sono le comunità che si trovano in una rete e, infine, $\delta(v_i, v_j)$ la delta di Kronecker con valore 1 se i vertici sono nella stessa comunità, altrimenti 0.

1.2.3 Integrazione

Le metriche di integrazione misurano quanto è facile spostarsi da un nodo all'altro. La metrica più usata in questo ambito é l'efficienza che si suddivide tra *efficienza globale* e *efficienza di percorso o locale* [17]. Con il termine efficienza si descrive un parametro che stima le distanze topologiche degli elementi della rete ed é inversamente proporzionale alla lunghezza media dei percorsi che collegano i nodi.

Efficienza globale

L'efficienza globale è una metrica che descrive l'efficienza del trasferimento di informazioni a distanza in una rete. Per ricavarla si calcola il numero più breve di passi necessari per andare dal nodo i ad ogni altro nodo della rete per poi ripetere tale operazione per tutti gli altri nodi. Successivamente viene calcolato, separatamente per ogni nodo, il numero medio di passi più brevi per spostarsi verso tutti gli altri vertici della rete. L'inverso del numero medio di passi più brevi per ogni nodo viene poi sommato per tutti i vertici e questa somma viene, infine, normalizzata tenendo conto del numero totale di connessioni che potrebbero esistere nella rete.
In formule:

$$E_{global} = \frac{1}{N(N-1)} \sum_{j,k} \frac{1}{L_{j,k}} \tag{1.5}$$

con N l'insieme di tutti i nodi della rete e $L_{j,k}$ la distanza minima, in termini di archi percorsi, tra i nodi j e k della rete. L'efficienza globale è una misura scalare che va da 0 a 1, dove 1 indica la massima efficienza globale della rete.

Efficienza locale

L'efficienza locale misura, invece, l'efficienza media del trasferimento di informazioni all'interno di sottografi ed è definita come l'inverso della lunghezza media del percorso più breve di tutti i vicini di un dato nodo tra di loro. É pertanto una metrica molto simile a quella descritta in precedenza, con la differenza che si concentra non su tutto il grafo ma solamente su parti di esso.
L'efficienza locale E_{loc} é la media delle efficienze di ciascun sottografo G_i costituito dai vertici vicini di ciascun nodo i della rete.
In formule si ha:

$$E_{local} = \frac{1}{N} \sum_{i \in N} E(G_i),$$

con N il numero totale di nodi della rete. Anche l'efficienza locale varia tra 0 e 1 ed é massima quando é pari a 1.
Si può osservare che sotto determinate condizioni i valori forniti da efficienza gobale e locale risultano essere molto simili o, in certi casi, identici. Ad esempio se un grafo risulta essere quasi completo, ovvero molti nodi sono connessi direttamente, le distanze tra qualsiasi coppia di nodi sono piccole. Questo significa che i nodi possono comunicare rapi-

damente sia a livello globale che locale. Di conseguenza, sia l'efficienza globale che quella locale saranno alte e simili. Se invece il grafo é completo non c'è differenza tra efficienza globale e locale, dato che ogni nodo è direttamente raggiungibile. Inoltre se il grafo é regolare ogni nodo ha un numero fisso di connessioni, e pertanto la distanza tra i nodi è simile in tutto il grafo. Questa simmetria nelle connessioni locali si riflette nell'efficienza locale che tende ad essere molto vicina a quella globale.
Per questi motivi, dovuti a proprietà spesso presenti nelle reti, nelle applicazioni pratiche che verranno esposte nei successivi capitoli non verrà più effettuata una distinzione tra efficienza globale e locale e verrà considerata semplicemente la metrica relativa all'efficienza di una rete.

1.2.4 Robustezza

Le metriche di robustezza, infine, valutano la capacità della rete di mantenere la connettività sotto vari tipi di fallimenti o attacchi. All'interno di questa questa categoria le più rilevanti sono la *forza* e l'*assortatività* [16].

Forza

Nella teoria dei grafi, la forza di una rete é un parametro che calcola le partizioni in cui si può suddividere l'insieme dei vertici e individua le zone in cui vi é un'alta concentrazione di archi. Formalmente corrisponde al rapporto minimo tra archi rimossi e componenti di vertici in cui é stato decomposto il grafo di interesse. La sua finalità é quella di valutare la robustezza del grafo, in particolare in seguito alla rimozione di alcuni vertici.
Di seguito viene fornita una sua formulazione matematica:

$$\sigma(G) = \min_{\pi \in \Pi} \frac{|\partial \pi|}{|\pi| - 1}, \tag{1.6}$$

con $G = (V, E)$ grafo indiretto, Π insieme delle partizioni dei vertici di V e $\partial\pi$ l'insieme degli archi che incrociano la partizione di vertici $\pi \in \Pi$.

Assortatività

Nella teoria delle reti, l'assortatività descrive la tendenza dei vertici delle reti a connettersi ad altri nodi che possono avere proprietà simili o dissimili. Generalmente viene considerata in base al grado dei vertici. Per valutare tale metrica viene calcolato il coefficiente di

assortatività, il quale è definito come il coefficiente di correlazione dei nodi in base al loro grado ed é ricavato con la seguente formula:

$$r = \frac{S_1 N_3 - S_2^2}{S_1 S_3 - S_2^2} \tag{1.7}$$

dove

$$S_k = \sum_{i=1}^{n} d_i^k, k = 1, 2, 3, \qquad N_3 = d^T A d = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} d_i d_j, \tag{1.8}$$

con A matrice di adiacenza della rete e d_i gradi dei nodi.
Tale parametro varia tra -1 e 1 e presenta questi tre casi limite:

- $r = 1$: rete perfettamente assortativa ;
- $r = 0$: rete non assortativa;
- $r = -1$: rete completamente disassortativa.

Analizzando i due casi limite, con $r = 1$ la rete è generalmente costituita da due o più componenti regolari connesse, mentre con $r = -1$ la rete ha una struttura biregolare, ossia tutti i vertici sullo stesso lato della bipartizione hanno lo stesso grado. L'esempio più semplice di tale struttura è un grafo a stella. Viene riportata una rappresentazione grafica dei due casi nella figura 1.3.

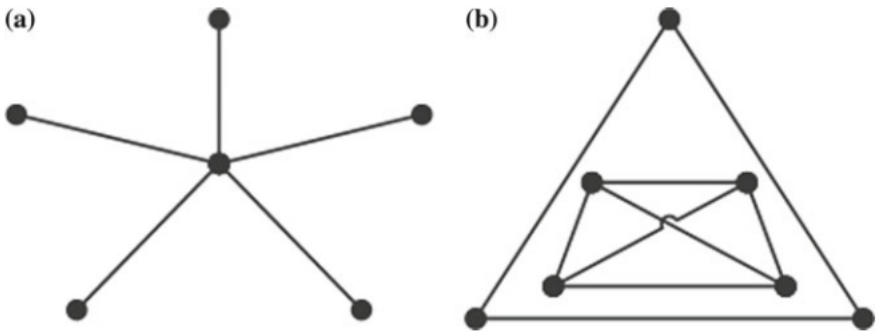


Figura 1.3: Esempi di rete completamente disassortativa (a) e perfettamente assortativa (b). Immagine tratta da [12]

1.2.5 Esempio

A conclusione della sezione sulle metriche di connettività viene fornito un esempio pratico. Si consideri il grafo riportato in figura 1.4.

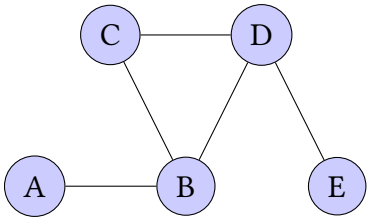


Figura 1.4: Esempio di grafo con cinque nodi e cinque archi.

Partendo dalle metriche di centralità si può osservare che il nodo A ed il nodo E hanno grado 1, il nodo C ha grado 2, mentre i restanti nodi B ed E han grado 3.
La betweenness centrality misura quante volte un nodo appare sui percorsi più brevi tra tutte le coppie di nodi nel grafo. Pertanto é necessario considerare tutte le coppie di nodi, cercare il percorso più breve tra essi (ad esmpio per la coppia C-E il percorso più breve é C-D-E) e infine contare quante volte un nodo compare in quel percorso. Ripetendo per tutte le coppie di vertici si ottiene che il nodo B ha il valore più elevato (3), ed é quindi il nodo più importante come ponte tra le altre parti del grafo, i nodi C e D hanno valore 2 e, infine, i nodi A ed E non hanno alcun ruolo intermedio nei percorsi tra altri nodi, quindi la loro betweenness centrality è 0.
Per calcolare la centralità degli autovalori é necessario, invece, determinare la matrice di adiacenza A. Essendo il grafo non orientato la matrice é simmetrica ed é di taglia 5×5, in quanto sono 5 i nodi.
La sua forma é la seguente:

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Successivamente é necessario andare a calcolare gli autovalori della matrice risolvendo la seguente equazione:

$$\det(A - \lambda I) = 0,$$

tramite la quale si ottengono i seguenti autovalori: $\lambda_1 \approx -1.6180, \lambda_2 \approx -1.3028, \lambda_3 \approx 0.0, \lambda_4 \approx 0.6180, \lambda_5 \approx 2.3028$.

La corrispondente matrice degli autovettori é la seguente:

$$V = \begin{pmatrix} -0.3717 & 0.3262 & 0.5773 & 0.6015 & -0.2453 \\ 0.6015 & -0.4250 & 8.4814e-17 & 0.3717 & -0.5650 \\ 1.5067e-16 & 0.6525 & -0.5773 & -9.9904e-18 & -0.4907 \\ -0.6015 & -0.4250 & 1.9083e-17 & -0.3717 & -0.5650 \\ 0.3717 & 0.3262 & 0.5773 & -0.6015 & -0.2453 \end{pmatrix},$$

in cui ogni colonna corrisponde a un autovettore associato al corrispondente autovalore. Si osserva che l'autovalore massimo é λ_5 a cui corrisponde l'autovettore $(-0.2453, -0.5650, -0.4907, -0.5650, -0.2453)$, ovvero la quinta colonna di V. Ciascuno di tali valori viene normalizzato consentendo di ottenere i seguenti valori 0.1161, 0.2676, 0.2324, 0.2676, 0.1161, i quali indicano la centralità di autovalore dei nodi a partire da A fino ad E. Si può notare che, secondo questa metrica, i nodi più centrali risultano essere B e D, seguiti da C, mentre quelli più marginali sono i nodi A ed E.

Analizzando le metriche di segregazione, il coefficiente di clustering per un nodo misura la tendenza dei suoi vicini a essere collegati tra loro; é quindi necessario calcolare il numero di triangoli che includono il nodo i e i suoi collegamenti ai nodi vicini per poter applicare la formula 1.3. Calcolando per ciascun nodo, i vertici A ed E hanno un solo nodo vicino, e, pertanto, non potendo formare triangoli hanno coefficiente 0. I nodi B e D hanno tre vicini e fan parte di un triangolo, quindi il loro coefficiente vale $\frac{1}{3}$. Infine facendo parte dello stesso triangolo ma essendo connesso a 2 nodi, l'ultimo vertice C ha coefficiente $\frac{1}{2}$. Il coefficiente di transitività T di una rete, è, invece, il rapporto tra il numero di loop di lunghezza tre e il numero di percorsi di lunghezza due. Nell'esempio ci sono 6 loop di lunghezza 3 (2 per C, 2 per B e 2 per D) e 14 percorsi di lunghezza 2 (2 per A, 3 per B, 4 per C, 3 per D e 2 per E), quindi $T = \frac{6}{14} = \frac{3}{7}$.

Infine per calcolare la modularità, si deve prima definire una divisione in comunità. Si considerano, ad esempio, due comunità: $\{A, B, C\}$ e $\{D, E\}$. Si applica la formula 1.4 e si sfrutta il fatto che che la rete è piccola e non ci sono molte interconnessioni tra comunità, pertanto la modularità sarà positiva e relativamente alta (~ 0.3), indicando una buona divisione della rete in moduli.

Per quanto riguarda le metriche di integrazione, concentrandosi sull'efficienza, é necessario calcolare la distanza minima, in numero di archi percorsi, tra ogni nodo del grafo. In particolare il nodo A ha distanza 1 da B, 2 da C ed D e 3 da E; il nodo B ha distanza 1 dai

nodi A,C e D e 2 da E; il nodo C dista 2 da A ed E e 1 da B e C; il nodo D dista 2 da A e 1 da B,C ed E e, infine, il nodo E ha distanza 3 da A, 2 da B e C e 1 da D.
Sapendo che il numero totale di nodi della rete é 5 e applicando la formula 1.5 si ottiene:

$$E = \frac{1}{20} \cdot (\frac{7}{3} + \frac{7}{2} + 3 + \frac{7}{2} + \frac{7}{3}) = \frac{1}{20} \cdot \frac{44}{3} = \frac{11}{15}.$$

Infine é possibile andare a calcolare le metriche di robustezza. Per calcolare la forza é necessario scegliere una partizione dei vertici, ad esempio si sceglie di formare un primo gruppo costituito dai vertici A, B e C e un secondo composto da D ed E. In questo caso si può osservare che due spigoli attraversano le partizioni collegandole, ovvero BD e CD. Applicando la formula 1.6 si ottiene $\sigma(G) = \frac{2}{2-1} = 2$.
Infine per calcolare l'assortatività é necessario applicare le formule 1.8 per calcolare il coefficiente r, mediante la formula 1.7. Sfruttando i gradi e la matrice di adiacenza calcolati in precedenza, si ottengono i seguenti risultati: $S_1 = 10, S_2 = 24, S_3 = 64, N_3 = 54$. Applicando 1.7 si ottiene il coefficiente di assortatività $r = -0.41$.
Tale valore indica che il grafo ha una tendenza disassortativa, ovvero i nodi con un alto grado tendono a connettersi a nodi aventi un basso grado.

1.3 Modello di Barabàsi-Albert

Per completare il capitolo viene proposta un'ultima sezione in cui viene descritta una particolare tipologia di rete, detta di Barabàsi-Albert.
Si tratta di un modello che descrive reti complesse caratterizzate da nodi connessi in modo eterogeneo, ovvero in cui alcuni di essi acquisiscono molte più connessioni rispetto ad altri, come ad esempio internet.
A causa della sua rilevanza e del suo ampio utilizzo in ambito pratico, questo modello viene descritto a completamento del capitolo; si osservi comunque che si tratta di un approfondimento e di un esempio di rete complessa che non verrà ripreso nei capitoli successivi.
Per poter introdurre la rete di Barabasi-Albert é necessario riprendere la definzione di grado, fornita in formule in 1.1.
Definito il concetto di grado di un nodo é possibile valutare la probabilità che un nodo arbitrario abbia esattamente grado k, detta distribuzione del grado, denotata con $\mathbb{P}(k)$, e

ricavata con la seguente formula:

$$\mathbb{P}(k) = \frac{n_k}{|V|},$$

con n_k numero di nodi aventi grado k.

La distribuzione del grado di una rete può assumere forme differenti a seconda della tipologia di rete. Alcune reti presentano connessioni tra nodi tutte ugualmente probabili e pertanto la distribuzione dei gradi é rappresentata da una curva uniforme.

Altre, invece, presentano distribuzioni differenti, che possono essere, ad esempio, gaussiane simmetriche o, viceversa, non simmetriche, con code caratterizzate da gradi elevati.

Più precisamente alcune reti presentano una distribuzione in legge di potenza, ovvero:

$$\mathbb{P}(k) \sim k^{-\gamma}, \tag{1.9}$$

con il parametro $\gamma \in [2, 3]$. In tal caso la rete é detta *free-scale*.

Uno dei modelli più famosi di costruzione di una rete priva di scala, é il modello di Barabàsi-Albert (1999) [5] che, data una rete avente m_0 nodi, ciascuno avente almeno un link con un altro vertice, viene costruita mediante i due seguenti passaggi:

- *crescita*: ad ogni passo t viene aggiunto un nodo avente m archi ($m \leq m_0$) che lo connettono a m nodi già esistenti nella rete;
- *attaccamento preferenziale*: la probabilità $\Pi(k)$ che un link del nuovo nodo si colleghi al nodo i dipende dal grado k_i secondo la formula:

$$\Pi(k) = \frac{k_i}{\sum_j k_j} \tag{1.10}$$

.

L'equazione 1.10 implica che se un nuovo nodo ha la possibilità di scegliere, ad esempio, tra un nodo di grado due e uno di grado quattro, è due volte più probabile che si connetta al nodo di grado quattro. Ovvero ciascun nodo aggiunto ha maggiore probabilità di connettersi ad un nodo avente grado elevato. Dopo t iterazioni il modello di Barabàsi-Albert genera una rete con $N = t + m_0$ nodi e $M = m_0 + mt$ connessioni.

Per la rete descritta dal modello di Barabàsi-Albert vale il seguente risultato:

Proposizione 1.3.1. *Nel modello di Barabàsi-Albert la distribuzione dei gradi segue una legge di potenza con esponente 3, ovvero* $\mathbb{P}(k) \sim k^{-3}$.

Per arrivare a questa proprietà é necessario concentrarsi sull'evoluzione temporale del modello di Barabàsi-Albert, analizzando inizialmente il grado di un singolo nodo in funzione del tempo. Nel modello proposto un nodo esistente può aumentare il suo grado ad ogni iterazione t ogni volta che un nuovo nodo entra nella rete; quest'ultimo si collega a m dei nodi $N(t)$ già presenti nel sistema. L'equazione 1.10 descrive la probabilità che uno di questi collegamenti si colleghi al nodo i. Si approssima il grado k_i con una variabile reale continua, che rappresenta il suo valore di aspettativa su molte realizzazioni del processo di crescita. La velocità con cui un nodo esistente i acquisisce collegamenti come risultato di nuovi nodi che si collegano ad esso è:

$$\frac{dk_i}{dt} = m\Pi(k_i) = m\frac{k_i}{\sum_{j=1}^{N-1} k_j} \tag{1.11}$$

Il coefficiente m indica che ogni nuovo nodo presenta m collegamenti. Pertanto, il nodo i ha m possibilità di essere scelto. La sommatoria nel denominatore della 1.11 si riferisce a tutti i nodi della rete tranne il nuovo nodo aggiunto, quindi si può riscrivere:

$$\sum_{j=1}^{N-1} k_j = 2mt - m \tag{1.12}$$

Sostituendo 1.12 in 1.11 si ottiene

$$\frac{dk_i}{dt} = \frac{k_i}{2t-1}.$$

Per t grandi il termine (-1) può essere trascurato al denominatore, ottenendo così:

$$\frac{dk_i}{dt} = \frac{k_i}{2t}$$

Tale equazione può essere riscritta separando le variabili e mettendola a sistema con la condizione iniziale $k_i(t_i) = m$, ossia il nodo i si unisce alla rete al tempo t_i con m connessioni. In questo modo si ottiene il seguente problema di Cauchy:

$$\begin{cases} \frac{dk_i}{k_i} = \frac{dt}{2t} \\ k_i(t_i) = m \end{cases}$$

Risolvendolo si ricava la seguente soluzione:

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\beta} \tag{1.13}$$

con $\beta = \frac{1}{2}$ detto *esponente dinamico*.

L'equazione 1.13 porta ad alcuni interessanti risultati:

- il grado di ogni nodo aumenta seguendo una legge di potenza con lo stesso esponente dinamico $\beta = \frac{1}{2}$, ossia tutti i nodi seguono la stessa legge dinamica;
- la crescita dei gradi è sublineare (cioè $\beta \leq 1$). Infatti ogni nuovo nodo ha più vertici a cui collegarsi rispetto al nodo precedente e, pertanto, al passare del tempo i nodi esistenti competono per i collegamenti con un pool crescente di altri nodi;
- tanto prima viene aggiunto il nodo i, tanto più alto è il suo grado $k_i(t)$;
- la velocità con cui il nodo i acquisisce nuovi collegamenti è data dalla derivata dell'equazione 1.13 e vale:
$$\frac{dk_i(t)}{dt} = \frac{m}{2}\frac{1}{\sqrt{t_i t}}$$
.

Dall'ultima osservazione si ottiene che in ogni passo temporale i nodi più vecchi acquisiscono più collegamenti e, inoltre, la velocità con cui un nodo acquisisce i collegamenti diminuisce con il passare del tempo come $t^{-1/2}$. Pertanto, un numero sempre minore di collegamenti procede verso un nodo.
Per tornare, infine, al risultato descritto nella proposizione, sfruttando diversi strumenti analitici, é possibile prevedere la distribuzione dei gradi della rete di Barabàsi-Albert:

$$p(k) \sim 2m^{\frac{1}{\beta}} k^{-\gamma}, \tag{1.14}$$

con

$$\gamma = \frac{1}{\beta} + 1 = 3.$$

Capitolo 2

Test statistici per le metriche di connettività

Nel secondo capitolo la discussione si sposta nell'ambito statistico.
Questo capitolo è, infatti, dedicato alla descrizione dei test di verifica d'ipotesi utilizzati nel prosieguo della tesi per verificare l'uguaglianza del valore delle metriche di connettività calcolate in diversi gruppi di pazienti.
Più precisamente questo capitolo inizia con una sezione relativa ai test di normalità, che verificano se determinate variabili, nel caso di studio le metriche di connettività, seguono o meno una distribuzione Gaussiana. Nella sezione centrale si introduce l'Analisi della Varianza (ANOVA) e la sua variante multivariata (MANOVA), due tecniche che verificano se esistono differenze significative tra le medie di più gruppi, rispettivamente, per una singola variabile o considerando contemporaneamente più variabili dipendenti. Successivamente viene introdotta la categoria dei test non parametrici e, in particolare, il test di Kruskal-Wallis, variante non parametrica di ANOVA utilizzabile quando la variabile risposta non segue un andamento gaussiano.
Infine, nell'ultima sezione, viene approfondito un problema che emerge nell'applicazione di più test contemporaneamente, ovvero il problema dei confronti multipli.

2.1 Test di normalità

La distribuzione normale, o Gaussiana, è una delle distribuzioni di probabilità più importante e utilizzata in statistica, a cui è possibile ricondursi, grazie al teorema del limite centrale,

ogni qualvolta si ha disposizione una grande quantità di dati.
La densità che descrive questa tipologia di distribuzione é data dalla seguente formula:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R},$$

con μ media e σ deviazione standard.
Diversi test statistici, tra cui ANOVA, richiedono l'assunzione di normalità per essere applicati. Pertanto, nella successiva sezione, si descrivono due test, comunemente utilizzati, per verificare se un dato campione è stato estratto da una popolazione avente distribuzione Gaussiana, ossia il test di Kolmogorov-Smirnov e il test di Shapiro-Wilk.

2.1.1 Test di Kolmogorov-Smirnov

Il test di Kolmogorov-Smirnov é un test non parametrico utilizzato per stabilire se un campione proviene da una popolazione avente una distribuzione specifica. Tale test, pertanto, verifica l'ipotesi nulla H_0 per cui i dati seguono una distribuzione specifica contro l'ipotesi alternativa H_1 per cui, viceversa, i dati non seguono la distribuzione considerata.
La statistica test è costruita come segue.
Siano $x_1 < x_2 < .. < x_n$ le osservazioni ordinate indipendenti e identicamente distribuite.
La funzione di distribuzione empirica è definita come:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(x_i),$$

con

$$I_{(-\infty,x]}(x_i) = \begin{cases} 1 & \text{se } x_i \leq x \\ 0 & \text{se } x_i > x \end{cases}$$

funzione indicatrice.
Indicata con $F(x)$ la funzione di distribuzione cumulata della variabile di interesse, la statistica test é definita come segue:

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \tag{2.1}$$

Si tratta quindi della più grande differenza, in valore assoluto, tra le due funzioni di distribuzione valutate su tutti i valori assunti da x.
Nella figura 2.1 è fornito un esempio di grafico di funzione di distribuzione cumulata empirica, la curva a scalini in blu, confrontata con una funzione di distribuzione cumulata

normale, entrambe ottenuto valutando 100 osservazioni campionarie nell'intervallo compreso tra -2.5 e 2. La distanza massima tra le due curve fornisce la statistica test di Kolmogorov-Smirnov.

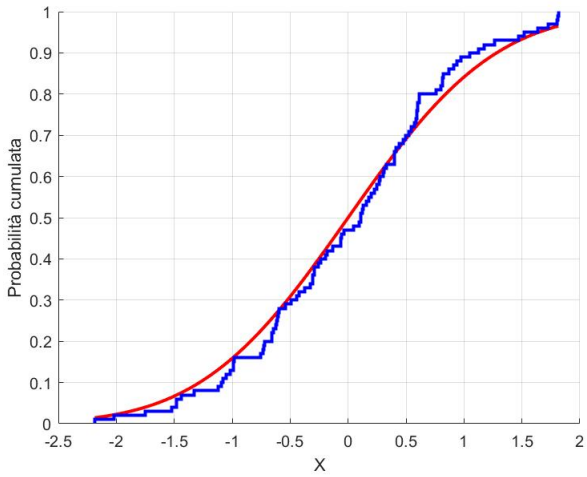


Figura 2.1: Rappresentazione grafica di una funzione di distribuzione cumulata empirica (in blu) e di una funzione di distribuzione cumulata normale (in rosso).

L'ipotesi relativa alla forma viene rifiutata se la statistica del test D_n è maggiore del valore critico ottenuto da una tabella apposita dei valori critici di Kolmogorov-Smirnov. In letteratura esistono diverse varianti di questa tabella, che fornisce i valori soglia oltre i quali si può rifiutare l'ipotesi nulla in base alla statistica D calcolata.
Si osservi che il test di Kolmogorov Smirnov può essere utilizzato non solo per verificare se un campione estratto da una popolazione segue una certa distribuzione, ma anche per confrontare le distribuzioni di più campioni, come approfondito in [18].
A conclusione viene fornito un esempio riassuntivo.

Esempio 1. *Si generano 1000 numeri casuali per quattro differenti distribuzioni: normale, esponenziale, t-student con 3 gradi di libertà e lognormale. Si effettuano cioè quattro test (uno per campione) in cui l'ipotesi nulla H_0 è che il campione sia realizzazione di variabil normali, mentre H_1 afferma che non lo è. Vengono poi fissati il livello di significatività $\alpha = 0.05$, il valore critico 0.04301 e la regione critica: si rifiuta H_0 se $D > 0.04301$.*
Dai quattro campioni estratti si ottengono i seguenti valori:

- *D_1= 0.0241492 per il campone estratto da una distribuzione normale;*
- *D_2 = 0.0514086 per il campone estratto da una distribuzione esponenziale;*

- *D_3 = 0.0611935 per il campone estratto da una distribuzione t-student;*
- *D_4 = 0.5354889 per il campone estratto da una distribuzione logaritmica.*

Di conseguenza, come previsto, l'ipotesi nulla non viene rifiutata per i dati distribuiti normalmente, ma viene rifiutata per le altre tre serie di dati che non sono distribuite secondo una gaussiana.

2.1.2 Test di Shapiro-Wilk

Nel caso in cui vengano considerati piccoli campioni, un test più potente del test di Kolmogorov-Smirnov é il test di Shapiro-Wilk, un test di normalità, descritto in un articolo pubblicato nel 1965 da Samuel Shapiro e Martin Wilk [15], che verifica se un campione proviene o meno da una popolazione normalmente distribuita.
Per descriverlo é necessario introdurre una statistica test, generalmente indicata W, e ottenuta mediante i passaggi successivamente illustrati.
Sia $m^T = (m_1, ..., m_n)$ il vettore dei valori attesi delle statistiche normali standard, ovvero aventi media 0 e deviazione standard 1, e sia $V = (v_{ij})_{i,j=1}^n$ la matrice di covarianza di ordine $n \times n$ corrispondente. Pertanto, indicato con $x_1 \leq x_2 \leq ..x_n$ un campione ordinato di taglia n estratto dalle variabili aleatorie $X_1, .., X_n$, ciascuna delle quali segue una distribuzione Gaussiana, si ha:

$$\mathbb{E}(X)_i = m_i \quad (i = 1, 2, ..., n),$$

$$\text{cov}(X_i, X_j) = v_{ij} \quad (i, j = 1, 2, ..., n).$$

Sia ora $y^T = (y_1, ..., y_n)$ un vettore di osservazioni casuali ordinate. L'obiettivo é sviluppare un test per verificare l'ipotesi per cui si tratta di un campione estratto da una distribuzione normale di media μ e varianza σ^2 sconosciute.
Nel caso in cui $\{y_i\}_{i=1}^n$ sia un campione normale allora si può esprimere come:

$$y_i = \mu + \sigma x_i \quad (i = 1, 2, ..., n).$$

Sfruttando il teorema ai minimi quadrati si ricava che la migliore stima lineare non distorta per i parametri μ e σ é data dalle quantità che minimizzano la seguente forma quadratica:

$$(y - \mu 1 - \sigma m)^T V^{-1} (y - \mu 1 - \sigma m), \tag{2.2}$$

con $1^T = (1, ..., 1)$ vettore unitario. Le stime ottenute per minimizzare 2.2 sono rispettivamente:

$$\hat{\mu} = \frac{m^T V^{-1}(m1^T - 1m^T)V^{-1}y}{1^T V^{-1}1m^T V^{-1}m - (1^T V^{-1}m)^2}$$

$$\hat{\sigma} = \frac{1^T V^{-1}(1m^T - m1^T)V^{-1}y}{1^T V^{-1}1m^T V^{-1}m - (1^T V^{-1}m)^2}.$$

Nel caso di distribuzione simmetrica si ha $1^T V^{-1}1 = 0$ e, pertanto:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}, \quad \hat{\sigma} = \frac{m^T V^{-1} y}{m^T V^{-1} m}.$$

Sia, inoltre, $S^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$ il termine che denota la stima simmetrica non distorta del fattore $(n-1)\sigma^2$.

La statistica test W é data dalla formula:

$$W = \frac{R^4\hat{\sigma}^2}{C^2 S^2} = \frac{b^2}{S^2} = \frac{(a^t y)^2}{S^2} = \frac{\left(\sum_{i=1}^{n} a_i y_i\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

con:

- $R^2 = m^T V^{-1} m$;
- $C^2 = m^T V^{-1} V^{-1} m$;
- $a^T = (a_1, ..., a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{\frac{1}{2}}}$
- $b = \frac{R^2\hat{\sigma}}{C}$

Si noti che la costante C è definita in modo tale da normalizzare i coefficienti lineari.

Definita la statistica test é possibile calcolare il p-value e dare un'interpretazione del test. L'ipotesi nulla H_0 è che la popolazione sia distribuita normalmente. Pertanto, se il p-value è inferiore al livello di significatività α scelto, l'ipotesi nulla viene rifiutata e si ha la prova che i dati analizzati non sono distribuiti normalmente. D'altro canto, se il valore del p-value è superiore al livello α scelto, l'ipotesi nulla non può essere respinta.

Come la maggior parte dei test di significatività statistica, se la dimensione del campione è sufficientemente grande, questo test può rilevare anche scostamenti insignificanti dall'ipotesi nulla costringendo a svolgere ulteriori indagini.

2.2 Test di confronto tra gruppi

Dopo aver descritto due test che è possibile utilizzare per per verificare se un campione é estratto da una popolazione che segue una distribuzione Gaussiana, é possibile introdurre alcuni test che permettono di confrontare due o più gruppi. Nel seguito verranno presentati sia test univariati che test multivariati.
La necessità di effettuare una distinzione relativa a queste due categorie di indagini statistiche é dovuta alle metriche di connettività introdotte nel primo capitolo e riassunte nella tabella 2.1.

Metrica	Tipologia	Categoria
Grado	Locale	Centralità
Assortatività	Globale	Robustezza
Betweenness	Locale	Centralità
Coefficiente di clustering	Locale	Segregazione
Betweenness tra archi	Locale	Centralità
Centralità autovalori	Locale	Centralità
Efficienza	Globale	Integrazione
Modularità	Globale	Segregazione
Transitività percorsi	Locale	Segregazione
Forza	Locale	Robustezza

Tabella 2.1: Tipologie e categorie delle metriche di connettività utilizzate.

La seconda colonna di questa tabella indica se la metrica considerata è locale o, a seconda che tale misura quantifichi proprietà relative a singoli nodi (o gruppi ristretti di questi con relativi archi), oppure proprietà della rete nel suo complesso.
Questa distinizione é particolarmente rilevante in quanto l'applicazione di due tipologie differenti di metriche di connettività ad una generica matrice dei dati fornisce diverse forme di risultati. Più precisamente, indicato con n il numero di componenti (ad esempio il numero di pazienti su cui viene testata la misura), le metriche globali, agendo sull'intera matrice, restituiscono un risultato $\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$, ossia un vettore univariato.

Viceversa il risultato ottenuto applicando una metrica di connettività locale, la quale lavora sui singoli nodi della matrice, coinvolge più variabili ed é del tipo $\mathbf{y} = \begin{pmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_n^T \end{pmatrix}$, $\mathbf{y} \in \mathbb{R}^{n \times r}, \mathbf{y}_i \in \mathbb{R}^r$, con r che rappresenta il numero di nodi o di archi a seconda della misura considerata.

Ovvero ad ogni soggetto considerato viene assegnato un vettore o una matrice.

Per questo motivo diventa cruciale la distinzione tra test statistici univariati e multivariati.

2.2.1 ANOVA

Per avere un'overview completa sui test statistici univariati si consiglia la lettura di [18].

Nello studio proposto l'analisi si é concentrata su uno dei test statistici univariati più famoso e usato: ANOVA (*analysis of variance*) [7]

Si tratta di una tecnica introdotta da R.A. Fisher come approccio statistico di verifica delle ipotesi per il confronto di due o più medie di distinte popolazioni per una singola variabile dipendente, nato a partire da una generalizzazione del t-test, un altro test statistico che confronta le medie di due gruppi per determinare se c'è una differenza significativa tra di esse.

In particolare, con ANOVA é possibile testare l'ipotesi nulla:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k,$$

dove μ_i é la media dell'i-esimo gruppo, con $i = 1, ..., k$.

L'ipotesi alternativa H_1 é, viceversa, che ci sono almeno due medie di gruppi distinti che differiscono in modo statisticamente significativo l'una dall'altra.

Per quantificare la variabilità tra le medie dei gruppi, ANOVA calcola la cosiddetta *sum of squares between groups* (SSB), mediante la seguente formula:

$$SSB = \sum_{i=1}^{k} (\bar{y}_i - \bar{y})^2,$$

con $\bar{y}_i$ la media campionaria del gruppo i-esimo, $\bar{y}$ la media campionaria di tutti i dati e k il numero totale di gruppi.

Un presupposto fondamentale di ANOVA è che se non ci sono differenze significative tra un insieme di medie della popolazione, o del gruppo, allora la corrispondente somma dei

quadrati tra i gruppi si comporta come la somma dei quadrati dell'errore casuale. Tale termine in ANOVA viene calcolato come la somma dei quadrati delle differenze tra i valori osservati e le loro medie previste, e serve per stimare la componente di errore residua nel modello.

Nota come *sum of squares within groups* (SSE) si ottiene in formule:

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{\kappa_i} (y_{ij} - \bar{y}_i)^2,$$

con y_{ij} l'osservazione j-esima nel gruppo i-esimo e κ_i il numero di osservazioni per ciascun gruppo.

Le somme dei quadrati derivano da una decomposizione lineare dell'osservazione j-esima nel gruppo i-esimo, y_{ij}, per la quale si può scrivere:

$$y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i). \tag{2.3}$$

Considerando, ad esempio, un vettore di osservazioni $(y_{11}, y_{12}, y_{13}, y_{21}, y_{22}, y_{23})$, in cui ci sono $k = 2$ e $\kappa_i = 3$ osservazioni per ciascun gruppo, é possibile riscrivere 2.3 come segue:

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} = \begin{pmatrix} \bar{y} \\ \bar{y} \\ \bar{y} \\ \bar{y} \\ \bar{y} \\ \bar{y} \end{pmatrix} + \begin{pmatrix} \bar{y}_1 - \bar{y} \\ \bar{y}_1 - \bar{y} \\ \bar{y}_1 - \bar{y} \\ \bar{y}_2 - \bar{y} \\ \bar{y}_2 - \bar{y} \\ \bar{y}_2 - \bar{y} \end{pmatrix} + \begin{pmatrix} y_{11} - \bar{y}_1 \\ y_{12} - \bar{y}_1 \\ y_{13} - \bar{y}_1 \\ y_{21} - \bar{y}_2 \\ y_{22} - \bar{y}_2 \\ y_{23} - \bar{y}_2 \end{pmatrix} \tag{2.4}$$

Ritornando all'equazione 2.3 e sottraendo a sinistra e destra per $\bar{y}$ si ottiene:

$$y_{ij} - \bar{y} = (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i) \tag{2.5}$$

Da tale equazione si evince che la differenza tra un'osservazione e la media di tutti i dati $(y_{ij} - \bar{y})$ può essere separata in due componenti: la differenza tra la media del gruppo e la media totale $(\bar{y}_i - \bar{y})$ e la differenza tra l'osservazione e la media del gruppo $(y_{ij} - \bar{y}_i)$. Elevando al quadrato entrambi i lati dell'equazione 2.5 e sommando tutti i gruppi e le osservazioni all'interno dei gruppi, si ottiene la somma dei quadrati aggiustata, ovvero la *total adjusted sum of squares* (SST):

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{\kappa_i} (y_{ij} - \bar{y})^2.$$

Nella tabella 2.2 sono riportati i risultati appena descritti relativi all'analisi della varianza, con l'aggiunta dei gradi di libertà (df), fattori che indicano il numero di valori indipendenti che possono variare in un'analisi statistica e la media quadratica, ovvero la somma dei quadrati divisa per il numero appropriato di gradi di libertà.
Ad esempio, il quadrato medio tra i gruppi è $MSB = \frac{SSB}{k-1}$.

Source of variation	Sum of Squares	df	Mean Squares
Between groups (SSB)	$\sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2$	$k-1$	$MSB = \frac{SSB}{k-1}$
Within groups (SSE)	$\sum_{i=1}^{k}\sum_{j=1}^{\kappa_i}(y_{ij} - \bar{y}_i)^2$	$k(\kappa_i - 1)$	$MSE = \frac{SSE}{k(\kappa_i - 1)}$
Total (SST)	$\sum_{i=1}^{k}\sum_{j=1}^{\kappa_i}(y_{ij} - \bar{y})^2$	$k\kappa_i - 1$	

Tabella 2.2: Tabella dell'analisi della varianza.

Per l'applicazione di ANOVA, é necessario che le tre seguenti assunzioni siano verificate.

- la variabile dipendente, cioé quella osservata, è distribuita normalmente in ogni gruppo che viene confrontato;
- c'è omogeneità delle varianze, ovvero le varianze della popolazione in ciascun gruppo sono uguali;
- indipendenza delle osservazioni.

Si osserva che la prima condizione é strettamente legata al tema introdotto nella prima sezione di questo capitolo, ovvero la necessità di eseguire test di normalità per verificare la distribuzione di una variabile.
Per avere conclusioni relative al test di ipotesi, é necessario valutare il p-value ottenuto, confrontandolo con il livello di significatività α, generalmente fissato a 0.05. Nel caso in cui sia superiore a 0.05 non ci sono differenze statisticamente significative tra le medie dei gruppi, mentre, in caso contrario si può affermare che vi é una differenza significativa tra di esse. In questo caso é spesso utile specificare quali gruppi presentano tali differenze, tramite l'applicazione di test post-hoc, come il test di Tuckey.

Bartlett test

Per applicare ANOVA é necessario anche verificare l'ipotesi relativa all'omogeneità delle varianze dei gruppi confrontati. Per fare ciò, viene applicato il test di Bartlett, un'analisi statistica basata sulla statistica χ^2 (una distribuzione statistica che misura quanto una serie di dati osservati si discosta da ciò che ci si aspetta teoricamente), con $k-1$ gradi di libertà, dove k indica il numero di gruppi considerati. In altre parole il test di Bartlett verifica se k popolazioni hanno uguale varianza.
L'ipotesi nulla é quindi:

$$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2,$$

contro l'ipotesi alternativa per cui le varianze di almeno due gruppi sono differenti.
Per indagare la significatività delle differenze tra le varianze di k popolazioni normalmente distribuite, si estraggono campioni indipendenti da ognuna delle popolazioni. Sia poi S_j^2 la varianza di un campione di n_j elementi estratti dal j-esimo gruppo ($j = 1, ..., k$).
La statistica test é espressa come:

$$B = \frac{(N-k)ln\frac{\sum_{i=1}^{k}(n_i-1)s_i^2}{N-k} - \sum_{i=1}^{k}(n_i-1)ln(s_i^2)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i-1} - \frac{1}{N-k}\right)},$$

dove N indica la somma delle taglie di tutti i campioni.
Ottenuta la statistica test, é, infine, possibile calcolare il p-value e confrontarlo con il livello critico di significatività α; nel caso in cui esso sia maggiore non ci sono differenze significative tra le varianze dei gruppi e pertanto l'assunzione di omogeneità delle varianze dei gruppi confrontati, richiesta da ANOVA, risulta essere soddisfatta.

2.2.2 MANOVA

Dopo aver descritto una delle principali tecniche di analisi statistica univariata è possibile passare all'analisi multivariata. Per avere un'overview completa sui test statistici multivariati si consiglia la lettura di [18].
Nello studio proposto l'analisi si é concentrata sul corrispondente test multivariato di ANOVA, ovvero MANOVA(*multivariate analysis of variance*) [2].
Si tratta di una tecnica utilizzata per confrontare le medie di più variabili dipendenti simultaneamente tra gruppi, particolarmente utile quando tali varabili sono ben correlate tra loro.

Tale metodo si applica quindi ad un modello multivariato del tipo:

$$Y = X\beta + \epsilon,$$

con

- Y: matrice delle variabili risposta di dimensione $n \times r$, con n osservazioni e r variabili dipendenti;
- X: matrice dei predittori di dimensione $n \times p$ con p predittori o variabili indipendenti;
- β: matrice dei coefficienti da stimare di dimensione $p \times r$ le cui colonne $\beta_1, ..., \beta_r$ rappresentano i coefficienti;
- ϵ: matrice errore di dimensione $n \times r$ le cui colonne $\epsilon_1, ..., \epsilon_r$ rappresentano le variabili aleatorie errore.

All'interno di questo modello, in cui si assume che le variabili risposta siano indipendenti, si richiede che $\epsilon_i \sim \mathcal{N}(0_n, \Sigma)$ e $Y_i \sim \mathcal{N}(x_i^T\beta, \Sigma)$, con Σ matrice di covarianza delle $n \times r$ variabili aleatorie campionarie.
Per questo motivo per applicare MANOVA le variabili dipendenti devono seguire due ipotesi:

- seguono distribuzioni normali multivariate;
- hanno varianze e covarianze uguali nei diversi gruppi (omoscheasticità o omogeneità delle covarianze).

La prima condizione richiede che dato un vettore y la sua densità sia della forma:

$$f(y) = \frac{1}{(\sqrt{2\pi})^k \cdot |\Sigma|^{\frac{1}{2}}} e^{\frac{-(y-\mu)^T \Sigma^{-1}(y-\mu)}{2}},$$

con k numero di variabili risposta.
Tuttavia, generalmente, il vettore delle risposte osservate ed il termine di errore del modello, non presentano la distribuzione prevista. Per aggirare questo problema si trasformano i dati fino a raggiungere la normalità, rischiando, però, di complicare l'interpretazione dei risultati finali.
La seconda condizione, legata a matrici Σ che misurano il grado di variazione simultanea

di due variabili, é dovuta al fatto che ci si aspetta che le variabili di risposta siano non correlate o moderatamente correlate, in quanto se così non fosse, si rischierebbe di avere una variabile il cui effetto è già stato spiegato da un'altra nel modello.
Come anticipato precedentemente, il test MANOVA esamina se i vettori delle medie per due o più gruppi provengono dalla stessa distribuzione campionaria. Viene quindi valutata l'ipotesi nulla

$$H_0 : \mu_1 = ... = \mu_k, \tag{2.6}$$

contro l'ipotesi alternativa

$$H_1 : \mu_i \neq \mu_j,$$

con k gruppi, $i < j$ e $j = 1, 2, ..., k$.
Per fare ciò sono state proposte, negli anni, diverse statistiche test, come ad esempio quelle fornite da Wilks (1932), Pillai (1955) e Roy (1957), le quali seppur aventi caratteristiche differenti conducono agli stessi risultati.
Di seguito vengono forniti dettagli relativi a tali statistiche, modificate nel caso di set di dati con risposte intrinsecamente non gaussiane, sviluppate troncando le espansioni in serie di ciascuna delle tre statistiche test originali.
Lambda di Wilks: Samuel Wilks nel 1932 sviluppò una statistica test per misurare le differenze tra i centroidi, ovvero i centri di un insieme di punti in uno spazio multidimensionale, delle medie delle variabili indipendenti.
Successivamente viene riportata la forma di tale statistica test:

$$\Lambda = \prod_{i=1}^{s} \frac{1}{1 + \lambda_i}, \tag{2.7}$$

con $s = \min(k - 1, p)$, dove k indica i trattamenti effettuati e p il numero di risposte ai trattamenti e λ_i l'i-esimo autovalore della matrice $E^{-1}H$, dove E, detta matrice di errore, descrive la varianza all'interno dei gruppi, ovvero misura quanto le osservazioni di ciascun gruppo variano rispetto al loro centro. Si dice quindi che E misura la *unexplained variance*, cioè la parte della varianza che non si può spiegare con le differenze tra i gruppi. Viceversa, H, detta matrice delle ipotesi, misura la *explained variance*, infatti valuta quanto i gruppi stessi si differenziano l'uno dall'altro sulle spazio delle variabili di risposta.
La statistica test 2.7 può essere espressa come:

$$\Lambda = \prod_{i=1}^{s} (1 + \lambda_i)^{-1} \tag{2.8}$$

e, usando l'espansione in serie di Taylor, 2.8 diventa:

$$\Lambda = \prod_{i=1}^{s}(1 - \lambda_i + \lambda_i^2 - \lambda_i^3 + \lambda_i^4 - \lambda_i^5 + ...) \tag{2.9}$$

La forma della statistica in 2.9 produce un'approssimazione numerica della statistica test in 2.8. Pertanto, ne consegue che la sommatoria nella parentesi della 2.9 può essere scomposta consentendo di ottenere le tre seguenti statistiche di Wilks troncate:

$$W_1 = \prod_{i=1}^{s}(1 - \lambda_i)$$

$$W_2 = \prod_{i=1}^{s}(1 - \lambda_i + \lambda_i^2)$$

$$W_3 = \prod_{i=1}^{s}(1 - \lambda_i + \lambda_i^2 - \lambda_i^3).$$

Traccia di Pillai: nel 1955 Sreedharan Pillai sviluppò una statistica test avente la seguente forma:

$$\nu = \prod_{i=1}^{s}\frac{\lambda_i}{1 + \lambda_i}, \tag{2.10}$$

la quale, come fatto precedentemente, può essere scritta come segue:

$$\nu = \prod_{i=1}^{s}\lambda_i(1 + \lambda_i)^{-1} \tag{2.11}$$

e, usando l'espansione in serie di Taylor, l'equazione 2.11 diventa:

$$\nu = \prod_{i=1}^{s}(\lambda_i - \lambda_i^2 + \lambda_i^3 - \lambda_i^4 + \lambda_i^5 - \lambda_i^6 + ...) \tag{2.12}$$

Sfruttando 2.12 le somme dei primi due, tre e quattro termini, rispettivamente, consentono di ottenere le tre forme troncate della statistica di Pillai, ovvero:

$$P_1 = \prod_{i=1}^{s}(\lambda_i - \lambda_i^2)$$

$$P_2 = \prod_{i=1}^{s}(\lambda_i - \lambda_i^2 + \lambda_i^3)$$

$$P_3 = \prod_{i=1}^{s}(\lambda_i - \lambda_i^2 + \lambda_i^3 - \lambda_i^4).$$

Radice caratteristica di Roy: viene infine riportata la statistica test sviluppata da Samarendra Roy nel 1957, con la sua versione modificata.
La radice caratteristica di Roy é fornita dalla seguente espressione:

$$\theta_{max} = \frac{\lambda_{max}}{1 + \lambda_{max}} \tag{2.13}$$

Come fatto in precedenza l'equazione 2.13 può essere scritta come segue:

$$\theta_{max} = \lambda_{max}(1 + \lambda_{max})^{-1} \tag{2.14}$$

e, sfruttando nuovamente l'espansione in serie di Taylor, la formula 2.14 diventa:

$$\theta_{max} = \lambda_{max} - \lambda_{max}^2 + \lambda_{max}^3 - \lambda_{max}^4 + \lambda_{max}^5 - \lambda_{max}^6 + ... \tag{2.15}$$

I primi due, tre, quattro termini, rispettivamente, della serie 2.15 consentono di trovare le tre statistiche troncate di Roy:

$$R_1 = \lambda_{max} - \lambda_{max}^2$$

$$R_2 = \lambda_{max} - \lambda_{max}^2 + \lambda_{max}^3$$

$$R_3 = \lambda_{max} - \lambda_{max}^2 + \lambda_{max}^3 - \lambda_{max}^4.$$

L' efficacia di tutte le nove statistiche di test troncate proposte per la modellizzazione di dati MANOVA con risposte non normali viene successivamente determinata e verificata attraverso un'estesa analisi effettuata tramite i metodi Monte-Carlo, che non vengono qua approfonditi.
Una volta calcolata una delle statistiche test descritte, il passo successivo del test consiste nel determinare il p-value associato a tale statistica, ovvero, la probabilità di ottenere un valore della statistica test uguale o più estremo di quello osservato, assumendo che l'ipotesi nulla 2.6 sia vera .
Se il p-value è inferiore a un livello di significatività predefinito, solitamente $\alpha = 0.05$ si rifiuta l'ipotesi nulla. Viceversa non si rifiuta l'ipotesi nulla e quindi, in questo caso, non ci sono prove sufficienti per affermare che esistono differenze significative tra i gruppi.
Si osserva che, sebbene le statistiche di Roy, Wilks e Pillai, possono avere sensibilità diverse a determinati aspetti dei dati, generalmente forniscono risultati coerenti, consentendo di ottenere, tramite MANOVA, risultati particolarmente solidi.

Il coefficiente di correlazione di Kendall

Per completare la sezione relativa al test MANOVA viene riportata la descrizione teorica di un coefficiente, che risulterà essere molto utile nell'applicazione pratica descritta nel terzo capitolo, fornendo una possibilità di ridurre il numero di metriche locali da analizzare, ovvero il coefficiente di correlazione di Kendall [1].
Si tratta di un coefficiente che valuta il grado di somiglianza tra due serie di ranghi assegnati a uno stesso insieme di oggetti.
Si considera il seguente insieme:

$$\mathbb{P} = \{a, b, ..., x, y\}.$$

Tale insieme, costituito da N oggetti ordinati può essere decomposto in $\frac{1}{2}N(N-1)$ coppie ordinate.
Per confrontare due insiemi ordinati, $\mathbb{P}_1$ e $\mathbb{P}_2$, aventi lo stesso insieme di oggetti, l'approccio di Kendall consiste, innanzittutto, nel contare il numero di coppie diverse tra di essi. Tale numero fornisce una distanza tra gli insiemi, detta distanza di differenza simmetrica, indicata con $d_\Delta(\mathbb{P}_1, \mathbb{P}_2)$.
Il coefficiente di correlazione si ottiene normalizzando la differenza simmetrica in modo tale che assuma valori compresi tra -1 e +1, con +1 corrispondente alla minima distanza possibile, ovvero gli insiemi possono essere riconosciuti come uguali, mentre -1, viceversa, corrisponde alla massima distanza possibile.
Sapendo che il numero massimo di coppie che possono differire tra due insiemi aventi $\frac{1}{2}N(N-1)$ coppie ordinate, equivale a $N(N-1)$, si ottiene il seguente coefficiente di correlazione di Kendall:

$$\tau = \frac{\frac{1}{2}N(N-1) - d_\Delta(\mathbb{P}_1, \mathbb{P}_2)}{\frac{1}{2}N(N-1)} = 1 - \frac{2 \cdot d_\Delta(\mathbb{P}_1, \mathbb{P}_2)}{N(N-1)}.$$

Al suo aumentare, aumenta il livello di somiglianza dei due gruppi che vengono confrontati. Tale coefficiente verrà calcolato in relazione alle metriche locali descritte in precedenza per poter ricercare possibili somiglianze tra di esse. Infatti in caso buona correlazione tra due sarà possibile applicare il test MANOVA ad una sola di esse, riducendo così il numero di test da applicare e gli errori che ne possono conseguire, problematica che verrà approfondita nella sezione conclusiva di questo capitolo.

2.2.3 Test di Kruskal-Wallis

La sezione relativa ai test di confronti tra gruppi termina con una descrizione di una categoria di test: i test non parametrici. In particolare, vengono descritte le caratteristiche principali di questi metodi con i loro vantaggi e svantaggi e, a conclusione, viene fornito un particolare esempio di test non parametrico, ovvero il test di Kruskal-Wallis.
I metodi statistici tradizionali, come il test della t-student e l'analisi della varianza (ANOVA), descritto precedentemente, richiedono alcune ipotesi sulla distribuzione della popolazione o del campione. In particolare nelle tecniche parametriche, come visto, deve essere soddisfatta l'ipotesi di normalità, ovvero che le medie del gruppo di campioni siano normalmente distribuite, e l'ipotesi di uguale varianza, ossia che le varianze dei campioni e della popolazione corrispondente siano uguali. Tuttavia, se questi presupposti non sono soddisfatti, i test parametrici non possono essere effettuati e, pertanto, si deve far ricorso a tecniche statistiche non parametriche.
L'analisi statistica non parametrica si differenzia notevolmente da quella parametrica in quanto utilizza solo il rango delle dimensioni dei dati o i segni $\pm$, invece che i loro valori originali. Più precisamente l'analisi non parametrica si concentra sull'ordine delle dimensioni dei dati piuttosto che sui valori da essi assunti.
Per chiarire con un esempio, si suppone di avere cinque dati per una variabile X, riportati nella tabella 2.3.

X_1	X_2	X_3	X_4	X_5
32	47	32	18	99

Tabella 2.3: Valori di esempio per le variabili da X_1 a X_5.

Dopo aver elencato i dati nell'ordine delle loro dimensioni, ogni istanza di dati viene classificata da uno a cinque; il dato con il valore più basso (18) viene classificato 1, mentre quello con il valore maggiore (99) viene classificato 5. Ci sono due istanze di dati con valori pari a 32, a cui viene quindi assegnato un rango di 2.5 (essendo due valori a pari merito al secondo posto in ordine crescente). Inoltre, i segni assegnati a ogni istanza di dati sono $+$ per i valori maggiori di un valore di riferimento fissato e $-$ per quelli inferiori ad esso.
Mentre l'analisi parametrica si concentra sulla differenza delle medie dei gruppi da confrontare, l'analisi non parametrica si concentra sul rango, ponendo quindi maggiore enfasi

sulle differenze dei valori mediani rispetto alla media.

Questa tipologia di indagine statistica ha una maggiore potenza rispetto all'analisi parametrica quando i dati non sono distribuiti normalmente. Infatti, come mostrato nell'esempio precedente, una caratteristica distintiva di questi test è che sono meno influenzati dai valori estremi rispetto ai test parametrici. Questo accade perché, nel processo di classificazione, solo il rango o il segno di un valore viene considerato, e non la sua grandezza effettiva. Nell'esempio, anche se il valore massimo è 99, il suo impatto sui risultati è limitato poiché viene semplicemente classificato come il rango più alto (5) o riceve un segno positivo. La specifica grandezza di 99 non altera direttamente il risultato del test, rendendo l'analisi non parametrica più robusta in presenza di dati con valori estremi

Nella tabella 2.4 vengono riportati i test parametrici generalmente applicati nel caso di uno, due o più campioni, con il corrispondente test statistico non parametrico.

Campioni	Test parametrico	Test non parametrico
1	t-test per un campione	Segno - Wilcoxon
2	t-test per campioni accoppiati	Segno - Wilcoxon
2	t-test per campioni non accoppiati	Mann-Whitney - Kolmogorov-Smirnov
$k > 2$	ANOVA	Kruskal-Wallis - Jonckheere

Tabella 2.4: Confronto tra test parametrici e non parametrici.

La sezione termina quindi con un approfondimento del corrispettivo test non parametrico di ANOVA, ovvero il test di Kruskal-Wallis.

Dalla tabella si osserva che una possibile alternativa a tale test é il test di Jonckheer; tuttavia in questa sezione l'analisi si concentra solamente sul test di Kruskal-Wallis, in quanto é una tecnica più versatile e che richiede meno assunzioni a priori.

Il test di Kruskal-Wallis è una tecnica non parametrica per analizzare la varianza. Esso analizza se esiste una differenza nei valori mediani di tre o più campioni indipendenti, andando a classificare i valori dei dati originali. Per far ciò raccoglie tutte le istanze di dati dai campioni e le classifica in ordine crescente. Se due punteggi sono uguali, viene utilizzata la media dei due ranghi. Le somme dei ranghi vengono quindi calcolate e la statistica del test di Kruskal-Wallis, indicata con H viene calcolata secondo la seguente equazione:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \left(\frac{R_j^2}{n_j} \right) - 3(N+1),$$

con k il numero di gruppi, R_j la somma del rango di ciascun campione, n_j il numero di campioni per ciascun gruppo e $N = \sum_{j=1}^{k} n_j$.
In questo test l'ipotesi nulla H_0 afferma che le mediane dei gruppi sono uguali, ovvero non ci sono differenze significative tra i gruppi in termini delle loro posizioni mediane nella distribuzione complessiva. Viceversa l'ipotesi alternativa descrive che almeno una delle mediane dei gruppi differisce dalle altre, implicando la presenza di una differenza significativa tra le mediane di almeno due dei gruppi considerati.
Per completare il test e verificare se rifiutare o meno l'ipotesi nulla si confronta la statistica H con un valore critico ottenuto dalla distribuzione chi-quadro. Pertanto fissato un livello di significatività (generalmente $\alpha = 0.05$) si consulta la tavola della distribuzione chi-quadro per trovare il valore critico corrispondente ai gradi di libertà $k - 1$, con k numero di gruppi ed al livello di significatività scelto. Se H è maggiore o uguale al valore critico allora si rifiuta H_0 e, pertanto, c'è evidenza sufficiente per affermare che almeno una delle mediane dei gruppi è significativamente diversa dalle altre. Viceversa se H è minore allora non si può rifiutare l'ipotesi nulla e di conseguenza non si può concludere che le mediane dei gruppi differiscano significativamente.

2.3 Il problema dei confronti multipli

Per completare il capitolo viene riportato un problema che spesso caratterizza i test statistici, ovvero il problema dei confronti multipli (*multiple comparison problem*).
Questo problema si verifica quando si considera simultaneamente un insieme di inferenze statistiche: maggiore è il numero di inferenze fatte, maggiore è la probabilità di effettuare inferenze errate. Per chiarire con un esempio, se un test viene eseguito al livello di significatività del 5% e l'ipotesi nulla corrispondente è vera, il rischio di rifiutare erroneamente l'ipotesi nulla è solamente del 5% (errore di tipo I). Tuttavia, se 100 test sono condotti contemporaneamente ciascuno al livello del 5% e tutte le ipotesi nulle corrispondenti sono vere, il numero atteso di falsi positivi, cioé gli errori che si commettono se si rifiuta l'ipotesi nulla pur essendo vera, è 5. Pertanto, se i test sono statisticamente indipendenti l'uno dall'altro, cioè sono eseguiti su campioni indipendenti, la probabilità di avere almeno un rifiuto errato è di circa il 99.4%.
Di conseguenza quando si eseguono molti test, anche con un basso livello di significatività,

la probabilità di ottenere casualmente risultati significativi (rifiuto H_0) diventa molto alta, il che può portare a conclusioni errate se non si applicano correzioni adeguate.
Di seguito vengono quindi descritte due possibili soluzioni per ovviare a tale problema.

Correzione di Bonferroni

La correzione di Bonferroni [14] é la tecnica più semplice ed immediata per risolvere i rischi dovuti ai confronti multipli nei test statistici.
Tale correzione consiste nel modificare il livello critico di significatività dividendolo per il numero di test statistici eseguiti.
In formule si avrà quindi:

$$\alpha_{Bonf} = \frac{\alpha}{n} \tag{2.16}$$

con n il numero totale di test statistici eseguiti.

La correzione di Bonferroni presenta tuttavia alcuni difetti, in quanto, é una modifica che tende ad essere troppo conservativa, in quanto riduce drasticamente la probabilità di ottenere falsi positivi, portando contemporaneamente ad un aumento notevole della probabilità di avere falsi negativi o errori del secondo tipo (si accetta erroneamente H_0), rendendo più difficile rilevare effetti reali quando si eseguono molti test. Infatti se viene eseguito un numero elevato di test l'aumento del termine al denominatore porta ad ottenere un nuovo livello critico di significatività particolarmente piccolo, cosicchè pochi o nessuno dei test risulterà significativo dopo l'applicazione della correzione.

False discovery rate

Un'altra soluzione per affrontare il problema dei confronti multipli, senza essere troppo conservativi come nel caso della correzione di Bonferroni, é la tecnica statistica del *false discovery rate* (FDR) [13].
Questa tecnica calcola un tasso, detto FDR, e lo utilizza per controllare il livello critico di significatività α. Per far ciò questo metodo identifica un insieme di potenziali risultati positivi del test, più probabili di altri, da indagare ulteriormente. Questi controlli sono meno conservativi degli accorgimenti descritti nella correzione di Bonferroni, al costo, tuttavia, di avere un tasso di falsi positivi più alto.
FDR dipende in parte dal fatto che le statistiche da testare siano o meno correlate tra loro. Si considera dapprima il caso in cui siano indipendenti o positivamente correlate e si de-

scrive il procedimento ricavato dagli studi di Benjamini e Hochberg.
Fissato n il numero di risultati del test si pongono i loro rispettivi valori di p-value in ordine crescente, cioé $p(1), ..., p(k), ..., p(n)$. Successivamente viene cercato il valore di k, indicato con $k^{'}$, tale che:

$$p(k^{'}) \leq \frac{\alpha}{n}k. \tag{2.17}$$

I primi k test sono quelli con risultati significativi, ovvero quelli per cui l'ipotesi nulla viene rifiutata.
Per chiarire con un esempio:

Esempio 2. *Si considera un caso in cui vengono effettuati 6 test,* $n = 6$*, i loro corrispondenti valori di* p*-value, in ordine crescente, sono* $0.001, 0.005, 0.009, 0.020, 0.080, 0.100$ *e* $\alpha = 0.05$*. Applicando la parte destra della formula 2.17 si ottengono, rispettivamente per* $k = 1, 2, 3, 4, 5, 6$*, i seguenti valori: 0.008, 0.017, 0.025, 0.033, 0.042, 0.050. I primi 4 valori di* p *sono inferiori ai corrispondenti valori ottenuti con* $\frac{\alpha}{k}n$*, ad esempio* $0.001 < 0.008, ..., 0.020 < 0.033$*, mentre, viceversa, sono maggiori per gli ultimi due casi. Pertanto, il* k *critico assume valore 4, di conseguenza le prime 4 ipotesi nulle si considerano rifiutate.*

Nel caso in cui le statistiche siano correlate negativamente é necessario considerare anche il termine $c(n) = \sum_{i=1}^{n} \frac{1}{i}$ per cui la formula 2.17 diventa:

$$p(k^{'}) \leq \frac{\alpha}{n \cdot c(n)}k.$$

Dopodiché si procede analogamente a quanto descritto sopra.

Capitolo 3

Confronti di metriche applicate su pazienti affetti da malattie neurodegenerative

Nell'ultimo capitolo l'intento é quello di mettere in pratica nel concreto gli aspetti teorici introdotti nei primi due capitoli, descrivendo un analisi statistica condotta per confrontare la connettività strutturale cerebrale stimata in pazienti affetti da demenza con diversi livelli di decadimento cognitivo..

Per fare ciò nella prima sezione vengono descritti gli aspetti più rilevanti della connettività strutturale cerebrale, con un focus su alcune tecniche innovative, come la *diffusion Magnetic Resonance Imaging* (dMRI) e la *Diffusion Tensor Imaging* (DTI). Dopo averle descritte vengono forniti esempi di problemi recenti in cui vengono utilizzate.

Tutto il resto del capitolo é invece incentrato sulla descrizione dell'applicazione delle tecniche introdotte nel secondo capitolo su un dataset pubblico che raccoglie i dati relativi a una coorte di 164 pazienti, suddivisi in quattro gruppi a seconda delle diversa condizione cognitiva. Dopo aver descritto il setting dei dati iniziali, vengono presentate due sezioni, una relativa alle metriche globali e una relativa a quelle locali, delle quali vengono descritti i diversi test statistici applicati sui risultati forniti da ciascuna metrica, con l'obiettivo finale di ricercare differenze significative tra i gruppi di pazienti nel caso univariato e multivariato.

3.1 Connettività strutturale cerebrale da risonanza a diffusione

La connettività strutturale cerebrale rappresenta l'insieme delle connessioni anatomiche che collegano diverse aree del cervello, principalmente attraverso le fibre di sostanza bianca. Per poterla analizzare in modo non invasivo esistono varie tecniche di recente sviluppo, tra le quali una delle più efficaci risulta essere la diffusion MRI.
Nella seguente sezione viene descritta tale tecnica e una sua variante, la DTI, con relative caratteristiche ed applicazioni.

3.1.1 Dati di DTI

La Diffusion MRI è una tecnica di risonanza magnetica che viene utilizzata per misurare la diffusione delle particelle d'acqua nei tessuti biologici, principalmente nel cervello. Questa tecnica è particolarmente utile per studiare la microstruttura dei tessuti e le connessioni cerebrali, permettendo di ottenere informazioni dettagliate che non sono visibili con altre tecniche di imaging. Come detto, é soprattutto utilizzata per tessuti strutturati come il cervello, perchè, in quella zona, la diffusione è anisotropica, ovvero le molecole d'acqua diffondono più facilmente lungo la direzione delle fibre nervose; viceversa, in caso di uguale diffusione in ogni direzione si parla di diffusione isotropica.
La diffusione delle molecole d'acqua può essere modellata matematicamente attraverso un tensore di diffusione [10], ossia una matrice simmetrica definita positiva che descrive come la diffusione varia nelle diverse direzioni.
Tale tecnica avanzata per analizzare l'imaging per diffusione é detta DTI, ovvero una tecnica di imaging a tensori di diffusione.
Nello spazio tridimensionale tale tensore di diffusione, indicato con D, é generalmente rappresentato da una matrice quadrata di taglia 3 nel seguente modo:

$$D = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{yx} & D_{yy} & D_{yz} \\ D_{zx} & D_{zy} & D_{zz} \end{bmatrix} = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{xy} & D_{yy} & D_{yz} \\ D_{xz} & D_{yz} & D_{zz} \end{bmatrix},$$

essendo D simmetrica.

Le componenti del tensore di diffusione D_{ij} rappresentano il grado di diffusione dell'acqua lungo la direzione i influenzata dal gradiente di concentrazione lungo la direzione j. Più precisamente le componenti diagonali rappresentano la diffusività dell'acqua lungo gli assi x, y e z. Ad esempio, D_{xx} indica quanto le molecole d'acqua si diffondono lungo la direzione x. Viceversa le componenti fuori diagonale rappresentano l'interazione della diffusione tra le direzioni. Ad esempio, D_{xy} rappresenta quanto la diffusione lungo la direzione x è influenzata dal gradiente lungo la direzione y.
La misurazione delle componenti del tensore di diffusione viene effettuata tramite una serie di scansioni dMRI con gradienti applicati in diverse direzioni. Da queste misurazioni, è possibile calcolare il tensore di diffusione utilizzando metodi di fitting lineare o non lineare. In generale, si risolve un sistema di equazioni basato sul modello di Stejskal-Tanner:

$$S(g) = S_0 e^{-bg^t Dg}$$

con:

- $S(g)$ il segnale misurato con un gradiente applicato nella direzione g;
- S_0 il segnale senza gradiente di diffusione;
- b un valore che dipende dai parametri del gradiente, come intensità e direzione di diffusione;
- g un vettore unitario che rappresenta la direzione del gradiente di diffusione;
- D il tensore di diffusione.

3.1.2 Costruzione di un grafo da dati di DTI

Ogni voxel, ovvero ogni unità volumetrica, nell'immagine ottenuta dalla tecnica DTI contiene un tensore di diffusione, che descrive il movimento delle molecole d'acqua nei tre assi principali. Grazie a questo tensore, è possibile inferire la direzione predominante lungo la quale l'acqua si diffonde, che coincide con l'orientamento delle fibre nervose. Utilizzando questi dati, diventa quindi possibile ricostruire le traiettorie delle fibre nervose nel cervello attraverso una tecnica chiamata trattografia.
La trattografia è uno strumento fondamentale per mappare le connessioni tra le diverse aree del cervello. Utilizzando algoritmi di tracciamento delle fibre, essa segue le direzioni

principali della diffusione dell'acqua nei tessuti cerebrali, ricostruendo le traiettorie delle fibre di sostanza bianca. Il risultato è una rappresentazione dettagliata della connettività strutturale del cervello sotto forma di mappe o modelli tridimensionali che illustrano le vie di comunicazione tra le diverse regioni cerebrali.

Per effettuare un'analisi accurata della connettività cerebrale, è necessario suddividere il cervello in regioni anatomiche definite, un processo noto come segmentazione o parcellizzazione del cervello. Questa suddivisione può essere basata su diversi atlanti cerebrali, che suddividono la corteccia e le strutture sottocorticali in un insieme di aree funzionalmente o strutturalmente distinte. Ogni regione segmentata viene trattata come un nodo all'interno del grafo, e le connessioni tra queste regioni, determinate tramite la trattografia, costituiscono gli archi del grafo.
Nella figura 3.1 viene mostrato un esempio di segmentazione del cervello.

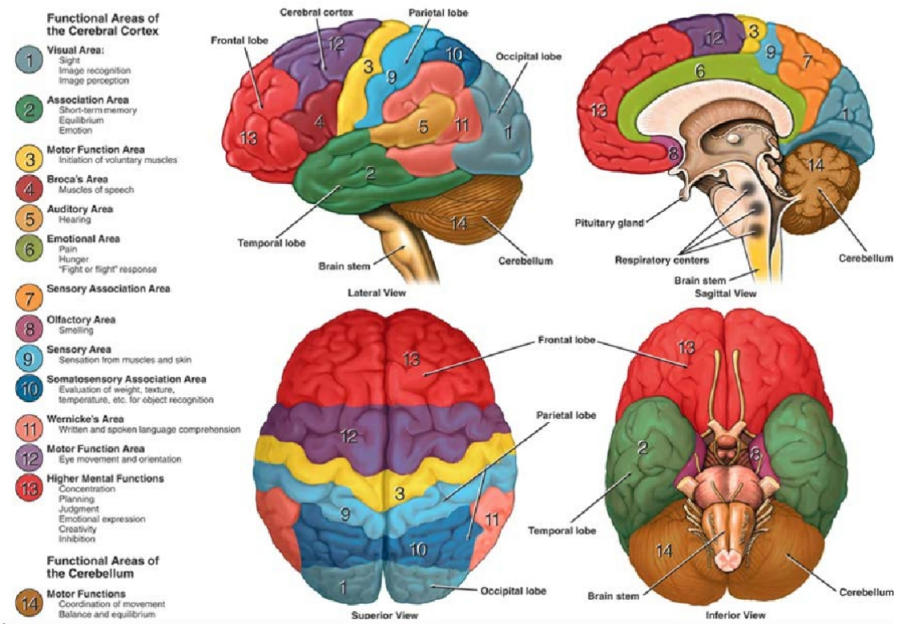


Figura 3.1: Esempio di segmentazione del cervello. Immagine presa da www.semanticscholar.org.

Le mappe ottenute dalla trattografia, una volta segmentato il cervello in regioni di interesse (ROIs), possono essere utilizzate per costruire matrici di di connettività strutturale cerebrale.
In una matrice di connettività strutturale ciascuna entrata rappresenta la forza o la probabilità delle connessioni tra la corrispondente coppia di nodi, dove quest'ultimi corrispondono alle regioni cerebrali identificate durante la segmentazione. In questo contesto, la forza di

connessione tra due nodi può essere determinata sulla base di parametri come il numero e la lunghezza delle fibre che collegano le regioni o la densità di connessione tra le aree.

Successivamente, per applicare i concetti della teoria dei grafi, è comune trasformare questa matrice ponderata in una matrice di adiacenza binaria, introdotta nella prima parte del primo capitolo. Per far ciò si sceglie una soglia sui valori della matrice di connessione: se il valore del link tra due nodi la supera, si considera esistente un collegamento tra quei due nodi, e il corrispondente valore nella matrice assume valore 1; se invece é inferiore, il collegamento viene considerato assente, e il valore corrispondente è nullo. Il risultato finale è quindi una matrice di adiacenza binaria, che definisce la presenza o l'assenza di collegamenti tra i nodi, e rappresenta la struttura di base di un grafo non ponderato.
Infine è possibile applicare i principi della teoria dei grafi per analizzare la connettività del cervello. In questo contesto, i nodi del grafo rappresentano le diverse regioni cerebrali, mentre gli archi le connessioni tra queste regioni. Il grafo così ottenuto può essere analizzato per identificare varie proprietà strutturali della connettività cerebrale.
Per chiarire i concetti appena descritti, nella figura 3.2 viene riportata una rappresentazione grafica del connettoma cerebrale (la mappa completa delle connessioni neurali all'interno di un cervello o di una parte di esso) di un bambino sano di 4 anni.
Si possono notare i seguenti passaggi:

1. definizione dei nodi della rete, con la parcellizzazione anatomica dell'immagine ad alta risoluzione pesata;

2. stima di una misura continua di associazione tra i nodi attraverso la connettività strutturale, ottenuta con la trattografia;

3. generazione della matrice di associazione, ottenuta analizzando tutte le associazioni a coppie tra i nodi.

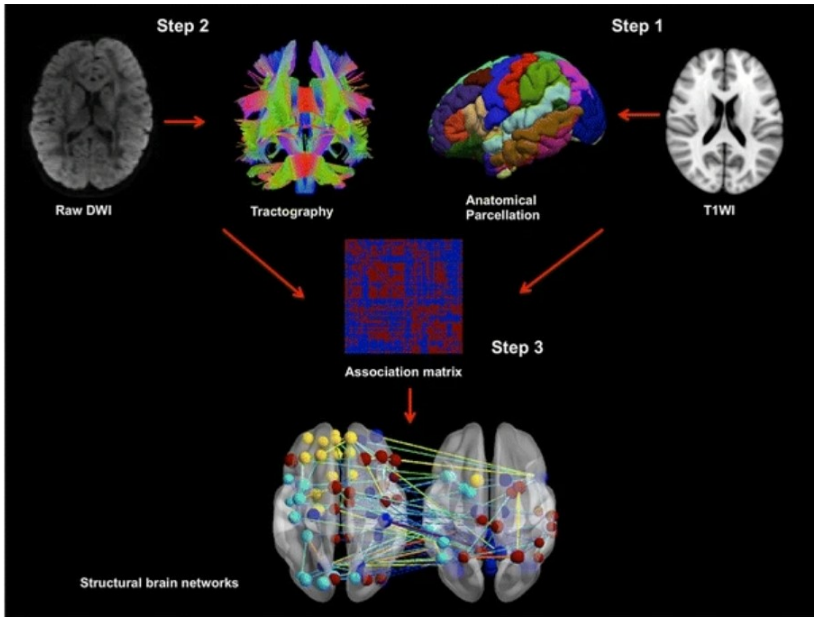


Figura 3.2: Esempio di connettoma cerebrale strutturale di un bambino sano di 4 anni, esplorato utilizzando la teoria dei grafi. Immagine presa da link.springer.com.

3.1.3 Applicazioni

La DTI e la successiva costruzione di grafi per l'analisi della connettività cerebrale stanno assumendo un ruolo cruciale nello studio e nella classificazione di pazienti affetti da malattie neurodegenerative [11]. Queste patologie, come l'Alzheimer, il Parkinson e la sclerosi multipla, sono caratterizzate da alterazioni progressive della struttura cerebrale, in particolare della sostanza bianca e delle connessioni tra le regioni cerebrali. La DTI permette di analizzare tali cambiamenti, fornendo un quadro dettagliato delle alterazioni della connettività strutturale, che può essere utilizzato per distinguere i pazienti affetti da queste malattie rispetto ai soggetti sani.

Uno degli obiettivi principali della DTI, per la classificazione dei pazienti, è l'identificazione di biomarcatori strutturali che possano predire la progressione della malattia o distinguere tra diverse patologie. La costruzione di grafi e l'analisi della connettività consentono di quantificare con precisione i cambiamenti nelle reti cerebrali, offrendo potenziali biomarker per la diagnosi precoce.

Un'area in forte espansione é l'integrazione dei dati di DTI con algoritmi di machine learning per la classificazione automatizzata dei pazienti con malattie neurodegenerative. Infatti, grazie alla grande quantità di dati ricavati dalla connettività cerebrale e dalle metriche derivate dalla teoria dei grafi, il machine learning permette di identificare pattern comples-

si e non lineari che possono sfuggire all'analisi tradizionale, consentendo di raggiungere un'accuratezza diagnostica superiore rispetto alle tecniche tradizionali di imaging.

Gli algoritmi di apprendimento supervisionato, ovvero quelli in cui vengono forniti alla rete i dati di input e i corrispettivi output, vengono spesso utilizzati per addestrare modelli capaci di distinguere tra pazienti con diverse patologie neurodegenerative o per differenziare pazienti da soggetti sani. Oltre all'apprendimento supervisionato, gli algoritmi di apprendimento non supervisionato, come il clustering o le tecniche di riduzione dimensionale, come la *principal component analysis* (PCA), possono essere utilizzati per scoprire sottogruppi di pazienti o nuovi fenotipi di malattia. Questi metodi permettono di identificare pattern latenti nei dati di connettività che possono riflettere differenze sottili nella progressione della malattia o nell'efficacia delle terapie.

3.2 Dataset analizzato

Dopo aver descritto gli aspetti tecnici relativi alla connettività cerebrale é possibile descrivere un'applicazione pratica, con la quale si vuole mostrare nel concreto gli aspetti teorici che sono stati descritti nei primi due capitoli.
Tutta l'applicazione che viene descritta in questo capitolo é stata eseguita mediante il programma Matlab, mentre i dati iniziali sono stati forniti in tabelle Excel.
In questo primo paragrafo viene descritto il setting dei dati iniziali.
Sono stati considerati 164 pazienti, suddivisi in 4 gruppi, aventi le seguenti sigle:

- **AD** (*Alzheimer's Disease*): soggetti affetti da morbo di Alzheimer, una forma comune di demenza caratterizzata da un declino progressivo delle funzioni cognitive e della memoria dei pazienti;

- **CN** (*Cognitively Normal*): gruppo di controllo che include individui cognitivamente normali, ovvero persone che non mostrano segni evidenti di deterioramento cognitivo o demenza;

- **EMCI** (*Early Mild Cognitive Impairment*): soggetti affetti da una compromissione cognitiva lieve precoce, ovvero uno stadio intermedio tra il normale invecchiamento e la demenza;

- **LMCI** (*Late Mild Cognitive Impairment*): soggetti affetti da una compromissione cognitiva lieve tardiva, simile all'EMCI, ma con sintomi di deterioramento cognitivo più evidenti.

Per ciascuno dei pazienti é stata fornita una matrice di connettività strutturale, avente dimensione 84×84, in quanto il cervello é stato segmentato in 84 distinte aree . In termini di grafi ciascuna di essa rappresenta un grafo pesato, ovvero avente connessioni ponderate dai pesi, e non diretto, ossia gli archi tra i nodi sono privi di direzione.
Su ciascuna matrice di ogni gruppo sono state applicate le metriche di connettività illustrate nel primo capitolo che vengono qua riprese: grado, assortatività, betweenness, coefficiente di clustering, edge betweenness, centralità degli autovalori, efficienza, modularità, transitività del percorso e forza della rete. Come anticipato nel secondo capitolo tali metriche si possono suddividere in due gruppi: metriche locali e metriche globali, a seconda che lavorino, rispettivamente, sui nodi della rete o sulla rete nel suo complesso. Da questo punto in avanti questa distinzione risulta essere fondamentale, in quanto ogni analisi statistica con conseguenti risultati é stata effettuata in relazione alla globalità o località delle metriche considerate.
Infine, a seconda del tipo di misura effettuata, sono stati applicati i test statistici descritti nel secondo capitolo con l'obiettivo di trovare differenze significative o meno tra i diversi gruppi di pazienti.

3.3 Risultati sulle metriche globali

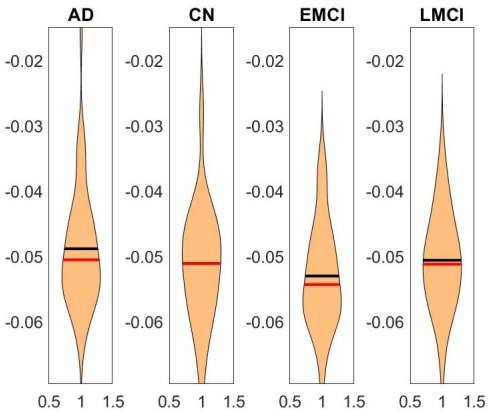
Le metriche globali sono misure che descrivono le proprietà strutturali complessive di un grafo o di una rete.
Le tre metriche globali considerate sono: assortatività, efficienza e modularità.
Dopo aver applicato tali misure alle matrici di connessione, descritte nella precedente sezione, é stata studiata la distribuzione dei risultati ottenuti per poter scegliere il test statistico da applicare, a seconda che la richiesta di normalità fosse soddisfatta o meno.
Per fare ciò sono state seguite due strade, una grafica con la realizzazione dei violin plot e una analitica, con l'esecuzione del test di Kolmogorov-Smirnov, applicato sia con la correzione di Bonferroni che con la tecnica del False Discovery Rate (FDR).
Nel caso in cui il test di normalità venga superato é stato poi applicato il test statisti-

co ANOVA, viceversa Kruskal-Wallis, per ricercare differenze significative tra i gruppi di pazienti.

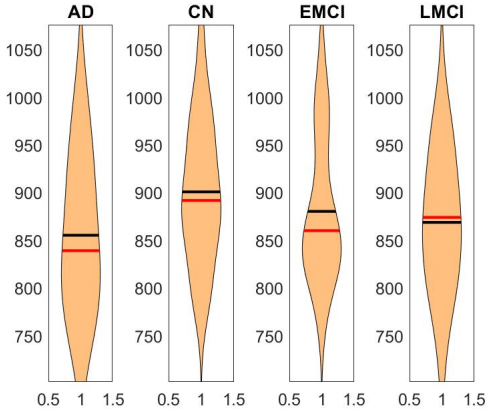
3.3.1 Analisi delle distribuzioni

I violin plot [9] sono uno strumento di visualizzazione dei dati, utilizzato per rappresentare la distribuzione di probabilità di una variabile continua. Risultano essere particolarmente utili quando vengono confrontati più gruppi contemporaneamente e sono così chiamati perché assumono una caratteristica forma a violino, in cui la sagoma mostra la densità stimata della distribuzione dei dati, con la parte a sinistra e a destra del violino speculari a riflettere la stessa densità dei dati.
Tale analisi grafica é stata utilizzata per condurre una prima analisi esplorativa visiva sul comportamento delle metriche globali così da meglio comprendere i successivi risultati dei test statistici.
Nei grafici 3.3a, 3.3b e 3.3c vengono riportati i violin plot delle metriche assortatività, efficienza e modularità.

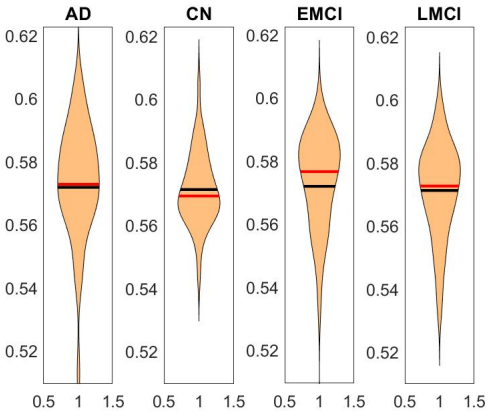
Dall'analisi dei violin plot si può osservare che, per l'assortatività, le figure sono relativamente simmetriche in tutti i gruppi, con una forma simile, a campana, per ciascuna categoria. In tutti i gruppi la media si avvicina alla mediana, pertanto le distribuzioni sembrano avvicinarsi a quelle di una gaussiana.
Nel caso dell'efficienza le distribuzioni mostrano una certa variabilità tra i gruppi, specialmente nel gruppo EMCI, avente una coda più pronunciata, a differenza delle altre che risultano essere più regolari. Di conseguenza il gruppo EMCI dell'efficienza potrebbe non seguire una distribuzione gaussiana.
Infine, dall'analisi dei violin plot relativi alla modularità, si osserva una certa simmetria in gruppi come CN e LMCI, mentre nei gruppi AD ed EMCI si notano distribuzioni leggermente più allungate e asimmetriche, il che potrebbe suggerire una deviazione dalla distribuzione normale.



(a) Assortatività



(b) Efficienza



(c) Modularità

Figura 3.3: Violin plot delle metriche assortatività, efficienza e modularità applicate ai quattro gruppi di pazienti. In nero la linea della media, in rosso quella della mediana.

3.3.2 Test di normalità

Dopo aver analizzato la normalità delle variabili mediante un'interpretazione visiva con i violin plot, si conduce un analisi confirmatoria presentando i risultati del test di normalità di Kolmogorov-Smirnov. Per ovviare al problema dei confronti multipli sono state applicate, separatamente, sia la correzione di Bonferroni, che quella con il false discovery rate. A causa del dataset sufficientemente grande é stato applicato solo il test di Kolmogorov-Smirnov e non quello di Shapiro-Wilk, più potente nel caso di piccoli campioni.
Nella tabella 3.1 vengono mostrati i risultati ottenuti con la correzione di Bonferroni.

Metrica	Gruppo	p-value	α	α_{Bonf}	Esito
Assortatività	AD	0.2129	0.0500	0.0125	non rifiuto H_0
	CN	0.4753	0.0500	0.0125	non rifiuto H_0
	EMCI	0.2300	0.0500	0.0125	non rifiuto H_0
	LMCI	0.5000	0.0500	0.0125	non rifiuto H_0
Efficienza	AD	0.1112	0.0500	0.0125	non rifiuto H_0
	CN	0.5000	0.0500	0.0125	non rifiuto H_0
	EMCI	0.0091	0.0500	0.0125	rifiuto H_0
	LMCI	0.5000	0.0500	0.0125	non rifiuto H_0
Modularità	AD	0.0027	0.0500	0.0125	rifiuto H_0
	CN	0.3200	0.0500	0.0125	non rifiuto H_0
	EMCI	0.0137	0.0500	0.0125	non rifiuto H_0
	LMCI	0.1844	0.0500	0.0125	non rifiuto H_0

Tabella 3.1: Risultati del test di normalità con correzione di Bonferroni.

La correzione di Bonferroni riduce il livello critico di significatività, fissato a priori a 0.05, a 0.0125. Tale valore é stato ottenuto applicando la formula 2.16, riportata nel secondo capitolo, considerando 4 test, ovvero il numero di gruppi di pazienti ai quali sono state applicate le metriche globali. Questa scelta, relativa al numero di test considerati, ovvero 4 test sui 4 gruppi di pazienti, fissata la metrica, é stata effettuata per mantenere basso l'errore di II specie, che avviene quando non si rifiuta H_0, cioé l'ipotesi di Gaussianità, pur essendo essa falsa.
L'ipotesi nulla H_0 afferma che i dati seguono una distribuzione normale e, per decidere se

rifiutarla o meno, é necessario confrontare il p-value con il nuovo α corretto.
Si osservi che il calcolatore ha fissato a 0.5 i p-value che superano tale valore.
Per la variabile assortatività si può osservare che in nessun caso il p-value è inferiore alla soglia di α_{bonf} e, pertanto, non si rifiuta l'ipotesi nulla per nessuno dei quattro gruppi, ovvero non ci sono prove statistiche per rifiutare la normalità della distribuzione di tale metrica per tutti i gruppi.
Per questa metrica é quindi possibile applicare il test statistico ANOVA.
Nel caso dell'efficienza, invece, si osserva che nel gruppo EMCI il p-value é inferiore al valore di α_{bonf}, e, di conseguenza, in quel caso, si rifiuta l'ipotesi nulla, ovvero ci sono prove statistiche sufficienti per concludere che la distribuzione della metrica efficienza nel gruppo EMCI non segue una distribuzione normale. Nonostante uno solo dei quattro gruppi non supera il test, questo é sufficiente a rendere necessaria l'applicazione del test di Kruskal-Wallis, che non richiede l'ipotesi della normalità, per tale metrica.
Tale test risulta necessario anche relativamente alla metrica modularità, infatti il gruppo di pazienti AD non supera il test di Kolmogorov-Smirnov, dato che il p-value é inferiore al livello critico di significatività corretto con Bonferroni.
Nella tabella 3.2 vengono invece riportati i risultati ottenuti applicando il test di normalità con la correzione mediante il false discovery rate.

Per applicare la correzione FDR é stato necessario calcolare la soglia FDR. Per fare ciò il livello critico di significatività α, fissato a priori a 0.05, é stato inizialmente diviso per il numero di test per ciascuna metrica sui gruppi, ovvero 4, e poi moltiplicato per un numero tra 1 e 4. Per scegliere tale valore i p-value dei gruppi di pazienti per ogni metrica sono stati ordinati in ordine crescente, ovvero da 1 (più piccolo) a 4 (più alto), e, il valore precedentemente ottenuto, é stato moltiplicato per la posizione di ciascun gruppo nella classifica dei p-value.
Successivamente il p-value é stato confrontato con la soglia FDR per stabilire se rifiutare o meno l' ipotesi nulla H_0: i dati seguono una distribuzione normale.
Si osserva che i risultati ottenuti sono i medesimi di quelli del test di normalità applicato con la correzione di Bonferroni. Infatti per la metrica assortatività, applicata a tutti i gruppi, il p-value rimane superiore alla soglia FDR e, pertanto, non si rifiuta H_0, ovvero ci sono prove statistiche sufficienti per concludere che la distribuzione di tale metrica segue una distribuzione normale.

Metrica	Gruppo	p-value	α	FDR threshold	Esito
Assortatività	AD	0.2129	0.0500	0.0125	non rifiuto H_0
	CN	0.4753	0.0500	0.0375	non rifiuto H_0
	EMCI	0.2300	0.0500	0.0250	non rifiuto H_0
	LMCI	0.5000	0.0500	0.0500	non rifiuto H_0
Efficienza	AD	0.1112	0.0500	0.0250	non rifiuto H_0
	CN	0.5000	0.0500	0.0375	non rifiuto H_0
	EMCI	0.0091	0.0500	0.0125	rifiuto H_0
	LMCI	0.5000	0.0500	0.0500	non rifiuto H_0
Modularità	AD	0.0027	0.0500	0.0125	rifiuto H_0
	CN	0.3200	0.0500	0.0500	non rifiuto H_0
	EMCI	0.0137	0.0500	0.0250	rifiuto H_0
	LMCI	0.1844	0.0500	0.0375	non rifiuto H_0

Tabella 3.2: Risultati del test di normalità con correzione FDR.

Viceversa, nel caso dell'efficienza, per il gruppo EMCI, e della modularità, applicata ai gruppi AD ed EMCI, il p-value é inferiore alla soglia FDR, portando al rifiuto dell'ipotesi nulla. Come concluso in precedenza questo fatto é sufficiente a rendere necessario l'utilizzo del test di Kruskal-Wallis per entrambe le metriche, mentre per l'assortatività, avendo superato il test, é possibile applicare ANOVA.
Si può osservare che i risultati ottenuti mediante le due correzioni sono, in questo caso, identici e vengono, inoltre, confermate le interpretazioni visive effettuate interpretando i violin plot.
Risulta comunque necessario sottolineare l'importanza del livello critico di significatività scelto, dato che un valore diverso avrebbe potuto comportare risultati differenti.

3.3.3 Test ANOVA e test di Kruskal-Wallis

Dopo aver descritto i risultati ottenuti mediante il test di normalità con le due correzioni, é possibile applicare i test statistici per verificare la presenza o meno di differenze significative tra i gruppi di pazienti.
Come osservato nella sezione precedente la variabile assortatività supera il test di nor-

malità e, pertanto, soddisfa una delle assunzioni necessarie per poter applicare ANOVA. Tuttavia, come osservato nel capitolo 2, é necessario verificare anche l'ipotesi relativa all'omogeneità delle varianze tra i gruppi. Per fare ciò é stato applicato il test di Bartlett, il quale ha restituito un p-value di 0.50847, maggiore, quindi, del livello critico di significatività, fissato a 0.05. Pertanto l'ipotesi nulla, che afferma che le varianze tra i gruppi sono uguali, non viene rifiutata e, quindi, é possibile applicare ANOVA.
Viceversa, dato che le metriche efficienza e modularità non hanno superato il test di normalità, é necessario applicare, per tali variabili, il test di Kruskal-Wallis, dato che esso non richiede l'assunzione di normalità.
Nella tabella 3.3 vengono riportati i risultati ottenuti applicando i test statistici sopracitati.

Metrica	Test	Source	SS	df	MS	F o χ^2	$p > F$ o χ^2	α_{Bonf}	α_{FDR}
Assortatività	ANOVA	Groups	0.00036	3	0.00012	1.72	0.1648	0.0167	0.0333
		Error	0.01126	160	0.00007				
		Total	0.01163	163					
Efficienza	K-W	Groups	16338.1	3	5446.02	7.25	0.0645	0.0167	0.0167
		Error	351226.9	160	2195.17				
		Total	367565	163					
Modularità	K-W	Groups	4506.3	3	1502.09	2	0.5728	0.0167	0.0500
		Error	927375.7	160	5796.1				
		Total	973075.9	163					

Tabella 3.3: Risultati test ANOVA e test di Kruskal-Wallis.

Come si può notare nelle ultime due colonne, anche nel caso dei test ANOVA e Kruskal-Wallis sono state applicate le correzioni di Bonferroni e FDR per superare il problema dei confronti multipli. Tali accorgimenti sono stai effettuati considerando tutti e tre i test applicati. Il livello di significatività corretto é stato poi confrontato con il p-value per valutare se rifiutare o meno H_0: non vi sono differenze significative tra i gruppi. Infatti l'obiettivo di questi due test statistici é quello di verificare se, relativamente alle metriche globali considerate, é possibile rilevare o meno distinzioni tra i diversi gruppi di pazienti.
Si può tuttavia osservare che in nessuna delle tre metriche vi sono differenze statisticamente significative tra i gruppi analizzati, poiché tutti i valori di p-value sono superiori ai livelli di significatività corretta, sia con Bonferroni che con FDR. Pertanto in nessuno dei tre casi si rifiuta l'ipotesi nulla.

Si può quindi concludere che l'analisi statistica delle metriche globali non ha consentito di ricavare conclusioni significative, dato che per assortatività, efficienza e modularità non é possibile rilevare distinzioni importanti tra i diversi gruppi di pazienti.
Solamente nel caso dell'efficienza, tuttavia, la presenza di un p-value particolarmente piccolo suggerisce la presenza di un trend verso una differenza tra i gruppi, che potrebbe diventare significativa, ad esmpio, con campioni più grandi. É comunque necessario sottolineare che nonostante i dati indichino una tendenza in quella direzione non si può concludere con certezza, dato che il p-value rimane comunque più elevato dell' α corretto.

3.4 Risultati sulle metriche locali

Le metriche locali in un grafo servono a descrivere proprietà specifiche dei singoli nodi o dei collegamenti nel loro intorno locale, cioè come ogni nodo o arco si comporta rispetto ai suoi vicini. Le metriche locali considerate sono: grado, betweenness centrality, coefficiente di clustering, edge betweeness, centralità degli autovalori, transitività del percorso e forza della rete.
L'obiettivo é lo stesso di quello descritto nella precedente sezione relativa alle metriche globali. Tuttavia, essendo misure che agiscono sui singoli nodi, restituendo come risultati vettori di vettori o vettori di matrici, é necessario applicare il test statistico multivariato MANOVA, descritto nel secondo capitolo.
Prima di applicarlo lo studio si é tuttavia concentrato su un altro test, ovvero il test della tau di Kendall, con l'obiettivo di trovare possibili similitudini tra i risultati forniti dalle metriche, col fine di ridurre il numero di test da applicare.

3.4.1 Test della tau di Kendal

L'obiettivo del test della tau di Kendall, come descritto nel secondo capitolo, è valutare la correlazione tra due variabili , misurando la forza e il segno della loro associazione. Il test verifica se esiste una relazione tra di esse al fine di valutare se i risultati di una delle due, pur dando diverse informazioni, risultano essere ridondanti.
Nella figura 3.4 sono riportati i coefficienti di correlazione di Kendall ottenuti dai vari confronti di coppie di variabili globali, poste su asse x e asse y. I valori sono stati ottenuti considerando tutti i pazienti, senza distinzione di gruppo, per ogni metrica considerata.

Si può notare che sono stati calcolati i coefficienti di correlazione per tutti gli incroci di metriche possibili, tranne che per le variabili edge betweenness e transitività del percorso, dato che, a differenza delle altre metriche che restituiscono vettori di vettori, queste due misure restituiscono vettori di matrici, risultando pertanto confrontabili solamente tra loro. Inoltre in tale figura le 7 variabili locali sono state riportate sugli assi e alcune di loro sono state abbreviate con l'uso di etichette che vengono qua spiegate per facilitare la comprensione: Clust=coefficiente di clustering, C.Aut=centralità degli autovalori, BetC=betweenness centrality, EdgeB=edge betweenness e Tr.Perc=transitività del percorso.
La figura 3.4, appena introdotta, é una *heatmap*, ovvero una matrice in cui a ciascun valore é assegnato un colore che, come indicato dalla *colorbar* a destra, varia a seconda che la correlazione tra le due metriche si avvicini a +1 o a -1. Imponendo una soglia ottimale di 0.6, si può osservare che alcune metriche presentano una correlazione significativa tra loro. In particolare, emerge una buona correlazione tra la forza e il coefficiente di clustering, tra la forza e la centralità degli autovalori, e tra la centralità degli autovalori e il coefficiente di clustering. Queste correlazioni indicano che, nonostante ciascuna metrica fornisca informazioni distinte sulla rete, vi è una sovrapposizione tra i loro risultati. Pertanto, per l'applicazione del test statistico MANOVA, é sufficiente includere una sola di queste tre metriche, riducendo così il numero di test da applicare.

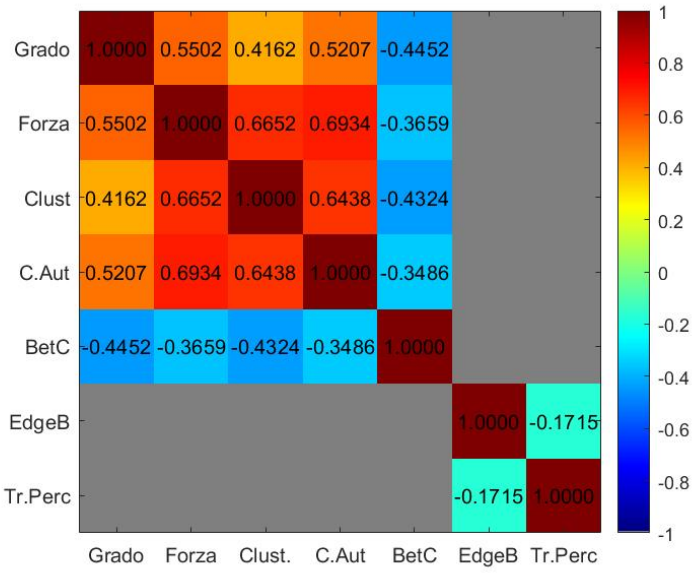


Figura 3.4: Heatmap Kendall tau. In grigio i riquadri relativi alle variabili per cui non é stato possibile effettuare un confronto.

3.4.2 MANOVA

Dopo aver applicato il test della tau di Kendall, i risultati forniti dalle metriche locali, sono stati confrontati mediante il test statistico MANOVA.

Come osservato nella sezione precedente la correlazione a coppie tra tre variabili: forza della rete, coefficiente di clustering e centralità degli autovalori risulta buona (superiore ad una soglia ottimale fissata a 0.6). Pertanto, al fine di evitare test ridondanti, é stata considerata una sola di queste tre metriche di connettività, ovvero la forza della rete. Si noti che tale scelta é arbitraria, infatti il test provato anche con una delle altre due opzioni ha fornito medesimi risultati.

I risultati ottenuti vengono riportati in due tabelle differenti, a seconda della forma dei risultati forniti dalle metriche di connettività, in quanto sono state effettuate operazioni preliminari differenti prima di applicare MANOVA.

Nella tabella 3.4 vengono riportati i risultati ottenuti applicando MANOVA alle tre metriche che, applicate alle matrici di adiacenza, restituiscono un vettore per ciascun paziente. Ciascun vettore, di dimensione 1×84, é stato ridotto ad un vettore 1×42 sfruttando la sua simmetria. Infatti, in ciascuno, la prima metà é relativa all'emisfero sinistro del cervello, mentre la seconda a quello destro. Pertanto ogni vertice della parte sinistra ha un suo corrispondente destro (ad esempio il nodo 1 e il nodo 43, il nodo 2 e il 44 e così via). Pertanto ciascun coppia é stata sostituita dalla media dei due valori, dimezzando così il vettore originario. Sapendo che il numero dei pazienti é 164 (n, le osservazioni) e il nuovo numero di nodi é 42 (r, le variabili dipendenti), é stato applicato MANOVA ad una matrice Y di dimensione 164×42.

Inoltre, per prevenire ad errori legati al fatto che le variabili potrebbero non seguire una distribuzione normale multivariata, i dati sono stati a priori log-normalizzati.

Metrica	Pillai	Roy	α_{Bonf}	Esito
Grado	0.0817	0.3377	0.0100	non rifiuto H_0
Forza	0.1763	0.6468	0.0100	non rifiuto H_0
Betweenness centrality	0.5302	0.6125	0.0100	non rifiuto H_0

Tabella 3.4: Risultati p-value MANOVA per le metriche locali che restituiscono vettori per ciascun paziente.

Nella tabella 3.4 vengono riportati i p-value ottenuti applicando MANOVA con le statistiche test di Pillai e di Roy, descritte nel secondo capitolo. Non é stata, invece, considerata la statistica della lambda di Wilks, essendo più debole all'assenza di normalità multivariata, che non é stata verificata nel caso in esame.
Dal confronto tra i p-value e i corrispondenti livelli critici di significatività α, con la correzione di Bonferroni, applicata considerando i 5 test effettuati su tutte le metriche locali, si osserva che per tutte e tre le metriche, con entrambe le statistiche test, il p-value rimane superiore a α_{Bonf}. Pertanto in tutti i casi non si rifiuta H_0, ovvero ci sono le evidenze statistiche sufficienti per concludere che non ci sono differenze significative tra i gruppi considrati, in relazione a queste tre variabili locali.
Come già affermato in precedenza, é importante ricordare che gli stessi risultati sono stati ottenuti con le metriche coefficiente di custering e centralità degli autovalori, che risultano quindi non essere rilevanti per avere una distinzione tra i pazienti.
Nella tabella 3.5 vengono, invece, riportati i risultati ottenuti applicando MANOVA alle due metriche che, applicate alle matrici di adiacenza, restituiscono una matrice per ciascun paziente, ovvero transitività del percorso ed edge betweenness. Per poter applicare MANOVA, senza considerare un numero eccessivamente grande di nodi, si é deciso di considerare solo alcuni di essi, ovvero i nodi delle regioni temporali del cervello. Per precisione si tratta dei nodi 8, 14, 29, 32, 50, 56, 71 e 74 (simmetrici tra emisfero sinistro ed emisfero destro). Avendo scelto queste 8 regioni, le nuove matrici associate ad ogni paziente sono diventate di dimensione 8×8. Successivamente é stata sfruttata la loro simmetria, per cui le informazioni fornite dalle entrate sopra e sotto la diagonale risultano essere ridondanti. Sono state quindi considerate solamente le 28 componenti sopra la diagonale nulla, le quali sono diventate le variabili dipendenti della matrice Y del modello multivariato (avente quindi dimensione 164×28).

Metrica	Pillai	Roy	α_{Bonf}	Esito
Transitività del percorso	0.7443	0.7864	0.0100	non rifiuto H_0
Edge betweenness	0.8291	0.9213	0.0100	non rifiuto H_0

Tabella 3.5: Risultati p-value MANOVA per le metriche locali che restituiscono matrici per ciascun paziente.

Dall'osservazione della tabella 3.5 e dal confronto tra i p-value e il livello critico di signi-

ficatività, corretto mediante Bonferroni, si giunge alle medesime conclusioni effettuate in precedenza. Infatti in ogni caso, con entrambe le statistiche test, le due metriche transitività del percorso ed edge betweenness non evidenziano differenze significative tra i gruppi di pazienti. Infatti, essendo il p-value sempre superiore al corrispondente α_{Bonf}, non si rifiuta mai l'ipotesi nulla H_0.

Si può quindi concludere che nessuna metrica locale ha consentito di trovare differenze statistiche significative tra i gruppi di pazienti analizzati.

Risulta comunque importante notare che l'analisi condotta é soggetta a tanti fattori. Per citarne uno, ad esempio, essa può variare a seconda del numero di nodi considerati. Infatti si é osservato che applicando MANOVA a situazioni differenti, i risultati presentavano importanti variazioni, il che suggerisce la possibilità di seguire nuove strade per condurre ulteriori indagini tramite l'applicazione di differenti test statistici d'ipotesi.

Conclusioni

Lo studio proposto in questa tesi é iniziato dall'interpretazione del cervello tramite la teoria dei grafi, che consente di descriverlo come una struttura formata da nodi connessi da archi. Per poterne analizzare le proprietà sono state introdotte le metriche di connettività e, successivamente, sono stati presentati i test statistici d'ipotesi necessari per verificare le uguaglianze dei risultati da esse ottenuti.
Nell'ultimo capitolo, dopo aver descritto le tecniche più importanti per ricavare i dati necessari per analizzare la connettività strutturale cerebrale, é stata descritta un'applicazione pratica avente l'obiettivo di ricercare differenze significative tra gruppi di pazienti soggetti a diversi disturbi cognitivi.
Come osservato, sia le metriche globali che quelle locali, non hanno consentito di trovare differenze significative. Infatti solamente la metrica efficienza della rete presenta un p-value particolarmente piccolo, il quale suggerisce un *trend* verso una differenza significativa tra i gruppi, relativamente a tale misura.
E'importante sottolineare come i risultati ottenuti sono soggeti a diversi fattori. Ad esempio, per le metriche edge betweenness e transitività del percorso, l'analisi MANOVA é stata applicata considerando le sole regioni temporali e trascurando le altre aree cerebrali. Di conseguenza una differente scelta di tali nodi di interesse potebbe portare a risultati differenti. Inoltre, relativamente a MANOVA, é possibile applicare una sua alternativa più resistente a deviazioni dalla normaltà multivariata, ossia PERMANOVA, una tecnica che non é stata approfondita nello studio proposto.
Tutta la ricerca effettuata risulta, quindi, essere solamente preliminare e si presta a numerose possibili prosecuzioni ed approfondimenti per poter espandere tale ricerca.
Innanzittutto l'applicazione di metriche di connettività consiste in un procedimento efficace e computazionalmente poco costoso, che però spesso non consente di avere un'analisi dettagliata della rete e dei risultati ottenuti. A tal proposito un approccio possibile consiste

nel confronto diretto tra grafi, tramite tecniche come la *Graph Edit Distance* (GED), che misura la distanza tra due grafi in termini di modifiche necessarie per trasformare un grafo nell'altro, tramite aggiunta o rimozione di nodi e archi. Questa tecnica rientra nell'ambito del problema dell'isomorfismo di grafi, ovvero se è possibile "mappare" un grafo sull'altro preservando tutte le connessioni.

Un naturale sviluppo del lavoro proposto riguarda l'implementazione di metodi di *Network-Based Statistics* (NBS) [8], ovvero tecniche che consentono di individuare sottoreti (*sub-networks*), in cui le connessioni differiscono significativamente tra gruppi, riducendo al minimo il rischio di falsi positivi. La loro applicazione può migliorare la potenza statistica dell'analisi, soprattutto nel caso di dataset con un numero elevato di connessioni e pazienti, favorendo la scoperta di pattern globali e locali di alterazione nella connettività cerebrale.

Infine, un ultimo sviluppo possibile riguarda l'integrazione dei grafi, costruiti dalle matrici di adiacenza, con dati provenienti da altre modalità di imaging, come la risonanza magnetica funzionale (fMRI), una tecnica di imaging cerebrale che misura l'attività neuronale indirettamente, rilevando i cambiamenti nel flusso sanguigno. Il suo utilizzo può consentire di avere un quadro più completo delle alterazioni cerebrali, permettendo di esplorare le relazioni tra struttura e funzione del cervello, al fine di migliorare la comprensione dei disturbi neurodegenerativi.

Bibliografia

[1] H. Abdi. The kendall rank correlation coefficient, 2007.

[2] A. L. Adeleke, W. B. Yahya, and A. Usman. A comparison of some test statistics for multivariate analysis of variance model with non-normal responses. Natural Sciences Research, 5, 2015.

[3] M. Arston, C. Onder, C. Herly, and P. Raeger. Remarks on path-transitivity in finite graphs. European Journal of Combinatorics, 17:371–378, 1996.

[4] S. M. H. Bamakana, I. Nurgalieva, and Q. Qu. Opinion leader detection: a methodological review. Expert Systems with Applications, 115:200–222, 2019.

[5] A.-L. Barabási. The Barabasi-Albert model. Cambridge University Press, 2016.

[6] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. Nature Reviews Neuroscience, 10:186–198, 2009.

[7] A. J. Dobson. An Introduction to Generalized Linear Models. Chapman Hall/CRC, 2001.

[8] O. Fornito, A. Zalesky, and E. Bullmore. Network-based statistic: Identifying differences in brain networks. NeuroImage, 52:1059–1069, 2010.

[9] D. W. Gerbing. The integrated violin-box-scatter (vbs) plot to visualize the distribution of a continuous variable. 2024.

[10] P. B. Kingsley. Introduction to diffusion tensor imaging mathematics: Part i. tensors, rotations, and eigenvectors. Concepts in Magnetic Resonance Part A, pages 101–122, 2006.

[11] C. Lenglet, J. Campbell, and M. Descoteaux. Mathematical methods for diffusion mri processing, 2010.

[12] T. Radivilova, D. Ageyev, and N. Kryvinska. Data-Centric Business and Applications. Springer, 2021.

[13] A. J. Rogers and S. Weiss. Epidemiologic and population genetic studies. Proceedings of the American Thoracic Society, 20:289–299, 2010.

[14] P. M. Sedgwick. Multiple significance tests: the bonferroni correction. British Medical Journal, 344, 2012.

[15] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality. Biometrika, pages 591–611, 1965.

[16] V. Shergin, S. Udovenko, and L. Chala. Assortativity properties of barabási-albert networks. CEUR Workshop Proceedings, pages 55–67, 2021.

[17] M. L. Stanley, S. L. Simpson, D. Dagenbach, R. G. Lyday, J. H. Burdette, and P. J. Laurienti. Changes in brain network efficiency and working memory performance in aging. Frontiers in Human Neuroscience, 9, 2015.

[18] L. Wasserman. All of Statistics: A Concise Course in Statistical Inference. Springer, 2004.

[19] R. J. Wilson. Introduction to Graph Theory. Longman, London, UK, 1972.