UNIVERSITÀ DI TRIESTE

Dipartimento di Fisica

Laurea Triennale in Fisica

TESI DI LAUREA

Un nuovo approccio per lo studio di correlazioni non lineari in dataset multidimensionali con applicazione a segnali elettroencefalografici.

Relatore:
Prof. Fabio Anselmi
Correlatore:
Prof. Fabio Benatti

Laureando:
Giacomo Amerio

A. A. 2022/2023

Abstract

Lo studio di sistemi fisici descritti da dati multidimensionali rumorosi è una delle sfide delle scienze moderne. In quest'epoca di grande progresso tecnologico, sono richiesti strumenti che facilitino la comprensione di grandi moli di dati rendendone accessibile la struttura e le interconnessioni. L'analisi delle correlazioni, siano esse lineari o non lineari, è un esempio di tali strumenti.
In questa tesi è presentato un metodo innovativo basato sull'algoritmo *Two nearest-neighbors* che permette l' analisi delle correlazioni non lineari nei dati tramite la stima della loro dimensionalità intrinseca.
Tale approccio, nel presente lavoro, è stato impiegato per la caratterizzazione della connettività strutturale in dati elettroencefalografici (EEG) registrati durante l'esecuzione di compiti motori e cognitivi.
Applicando la tecnica proposta, siamo riusciti ad identificare le correlazioni più pronunciate tra i segnali degli elettrodi ed a ricondurle a specifiche aree cerebrali che governano la pianificazione e l'esecuzione di movimenti degli arti. Così facendo, si è definito una nuova metodologia di analisi che consente di caratterizzare statisticamente la natura non lineare delle correlazioni di serie temporali associate a misure di voltaggio.
I risultati ottenuti evidenziano l'utilità di questa tecnica nel rilevare sottili interdipendenze fra segnali neurali, non accessibili ai tradizionali metodi di correlazione lineare come per esempio il Coefficiente di Correlazione di Pearson. L'analisi tramite la stima della dimensione intrinseca del data-set si è dimostrata così uno strumento promettente per decodificare l'informazione comune ai segnali provenienti da diverse popolazioni di neuroni.

Indice

Elenco delle figure

Capitolo 1

Introduzione

Le correlazioni descrivono l'esistenza di relazioni statistiche tra variabili e sono studiate per descrivere sistemi caratterizzati da un alto grado di complessità. Sebbene storicamente si sia fatto ampio ricorso a misure di correlazione lineare, come il coefficiente di Pearson, tali metodologie presentano limitazioni nel catturare interdipendenze non lineari in data-set multidimensionali.
I metodi proiettivi per la semplificazione di dati multivariati, come l'Analisi delle Componenti Principali [1], tentano di ovviare a questi problemi proiettando i dati in spazi a dimensionalità ridotta. Tuttavia, ciò comporta inevitabilmente una perdita di informazioni sulle strutture correlazionali non lineari caratterizzanti molti data-sets.
In questo lavoro presentiamo una tecnica per analizzare le correlazioni non lineari tra variabili, basata sul concetto di dimensione intrinseca (ID). L'idea fondamentale è che la complessità geometrica dei dati, quantificata dalla ID, possa rilevare interdipendenze non lineari tra variabili.
Stimando la dimensionalità intrinseca di un insieme di dati e confrontandola con quella di versioni casualmente permutate dei dati stessi, è possibile identificare correlazioni non lineari altrimenti difficili da rilevare. La presenza di correlazioni emerge da differenze significative tra la dimensionalità intrinseca originale e la media di quelle delle versioni permutate, dove eventuali correlazioni sono state distrutte. Per identificare tale differenza, si conduce uno Z-test che valuta se l'ID originale è significativamente inferiore a quella delle versioni non correlate. Una ID inferiore indica la presenza di correlazioni tra le variabili del dataset originale [17].
Il nostro studio ha usato i dati appartenenti al EEG Motor Movement/Imagery Dataset[21][19] consistenti in segnali EEG misurati da 64 elettrodi in 109 soggetti durante l'esecuzione di quattro task cognitivi diversi. I segnali sono stati raccolti con un sampling rate di 160 Hz. L'applicazione dell'algoritmo ha dato risultati interessanti in due studi diversi: il primo, riguardante l'analisi su un singolo task svolto da tre pazienti diversi, ha evidenziato pattern di correlazione non lineare comuni a tutti i soggetti. Il secondo, riguardante invece un singolo individuo impegnato nei quattro task di cui sopra, ci ha permesso di trovare le variazioni delle correlazioni dettate dalla diversa natura delle attività eseguite.
Al fine di agevolare la visualizzazione di questi risultati, sono state create diverse mappe

di correlazione in cui sono presenti pattern distinti tra i pazienti. Successivamente sono state riportate le correlazioni più importanti tra i segnali EEG su mappe di elettrodi sovrapposte allo scalpo. Ciò ha permesso di avvalorare l'ipotesi per cui le correlazioni più forti tra segnali sono da ricercare in zone ben localizzate del cervello [6]. Risultato interessante, col nuovo metodo, sono state identificate correlazioni invisibili al metodo di Pearson.

La tesi è organizzata come segue: Il Capitolo 2 delinea il background matematico della dimensionalità intrinseca [15] [20] e dell'algoritmo *Two-nearest neighbors* [20] e motiva l'utilizzo della stima dell'ID nell'analisi di correlazioni non lineari [17]. Il Capitolo 3 descrive il codice progettato per l'analisi, le proprietà dei data-set su cui è possibile applicare l'algoritmo in questione ed i dettagli dei test statistici svolti sui risultati. Il Capitolo 4 presenta l'applicazione sperimentale della tecnica proposta sull'*EEG Motor Movement/Imagery Data-set* [19][21], illustrando i dettagli di acquisizione dei dati ed i risultati sperimentali sulla connettività strutturale e funzionale di varie reti neurali. Inoltre viene integrato un confronto tra l'analisi dati condotta tramite la nuova tecnica e quella tradizionale che adotta il coefficiente di Pearson [18]. Il capitolo 5 discute l'impatto e le prospettive future di questo approccio topologico allo studio della connettività neurale. Infine, il capitolo 6 riporta il materiale supplementare ed in particolare, i grafici di correlazione relativi a tutti i soggetti dello studio.

Capitolo 2

Apparato teorico

2.1 Analisi Dati Topologica

L'analisi dati topologica è una branca della scienza che si concentra sulla comprensione delle proprietà topologiche e geometriche dei dati, piuttosto che basarsi solo su misure statistiche o algebriche.
Questo approccio si è dimostrato estremamente utile in una vasta gamma di campi, dalla cosmologia alla biologia computazionale e alla fisica dei sistemi complessi.
In questo paragrafo, verranno esplorate le principali nozioni dell'analisi dati topologica, le sue proprietà fondamentali, con un confronto diretto rispetto a metodi più tradizionali, come l'Analisi delle Componenti Principali (PCA)[1].

2.1.1 Concetti Fondamentali

L'analisi dati topologica (TDA) si basa sul concetto di topologia, che studia le proprietà geometriche e spaziali degli oggetti indipendentemente da misure quantitative. In questo contesto i dati vengono rappresentati come punti in uno spazio multidimensionale, dove le relazioni di vicinanza tra i punti riflettono la struttura topologica dei dati stessi.

Le proprietà chiave di questo tipo di analisi includono:

- *Invarianza topologica*: La TDA è invariante rispetto a trasformazioni continue dello spazio dei dati, il che significa che la struttura topologica viene preservata anche quando i dati vengono traslati, ruotati o scalati.

- *Rilevamento di caratteristiche multi-scala*: L'analisi può identificare caratteristiche rilevanti a diverse scale, permettendo di catturare sia dettagli locali che strutture globali nei dati.

- *Resistenza al rumore*: Questo metodo è robusto al rumore nelle misure, in quanto le proprietà topologiche stabili non sono sensibili a fluttuazioni casuali.

2.1.2 Confronto con l'Analisi delle Componenti Principali PCA

La PCA è un metodo di proiezione che cerca una trasformazione lineare degli input che massimizzi la varianza dei dati nello specifico, cerca una rotazione ortogonale degli input che renda massima la varianza.
Questo algoritmo ha due problemi principali: cambiando le unità di misura delle variabili, vengono modificate anche le componenti principali poiché questo cambiamento produce varianze diverse per le variabili trasformate. Di conseguenza cambiano gli autovalori della matrice di covarianza. In secondo luogo, per determinate distribuzioni di dati, la proiezione del data-set sulle componenti principali non apporta alcun miglioramento, ad esempio nel caso di dati bidimensionali distribuiti uniformemente lungo un cerchio centrato nell'origine, non esistono direzioni che massimizzino la varianza, sono dunque presenti autovalori degeneri che rendono l'applicazione di questa tecnica futile.
Come già riportato, la TDA è invariante per riscalamento delle variabili, inoltre, come verrà dimostrato nel seguito di questo studio, una distribuzione non lineare dei dati non ostacola la raccolta di importanti informazioni sulla natura del data-set.

2.2 Dimensione Intrinseca

La nozione di dimensione intrinseca ha le sue radici nella teoria dei sistemi dinamici, dove una dissertazione integrale sullo spazio delle configurazioni è impossibile. Per agevolare lo studio di questi sistemi, è necessario utilizzare proiezioni su un numero limitato di variabili, metodo che semplifica significativamente l'analisi e porta a risultati approssimati, ma utili.

La dimensione intrinseca è un concetto polisemico, molti autori ne hanno dato la propria definizione. Alcuni di loro la definiscono come il numero minimo di parametri necessari per generare una rappresentazione dei dati mantenendo la struttura intrinseca dell'insieme [2]. Bishop [5] afferma che l'ID è la dimensione dello spazio in cui i dati risiedono interamente, senza perdita di informazioni, mentre secondo Fukanaga [3], l'ID è il numero minimo di parametri necessari per descrivere accuratamente le caratteristiche salienti di un sistema.

L'effettuazione di una misura così cruciale è resa difficile dalla sua dipendenza dalle variazioni di densità del data-set. Tuttavia, esistono due principali metodologie ed entrambe sono capaci di produrre risultati adeguati:

- *metodi proiettivi* il cui obiettivo è trovare una proiezione lineare dei dati su uno sottospazio della rappresentazione mediante trasformazioni lineari o non lineari dell'input. L'Analisi delle Componenti Principali [1] e la Multidimensional Scaling [22] sono alcune delle tecniche utilizzate per raggiungere tale scopo;
- *metodi geometrici* che si concentrano invece sulla topologia del data-set, in particolare utilizzano le distanze tra i punti all'interno dello stesso per effettuare una

stima dell'ID. Esempi di tecniche che ricadono in questa metodologia sono basati sulle dimensioni frattali o su algoritmi di tipo *nearest neighbors*

Quest'ultima tecnica permette di definire le distribuzioni degli intorni dei punti come funzioni della dimensionalità d dell'intero insieme, assumendo però, che i punti vicini siano estratti da ipersfere d-dimensionali sufficientemente piccole.

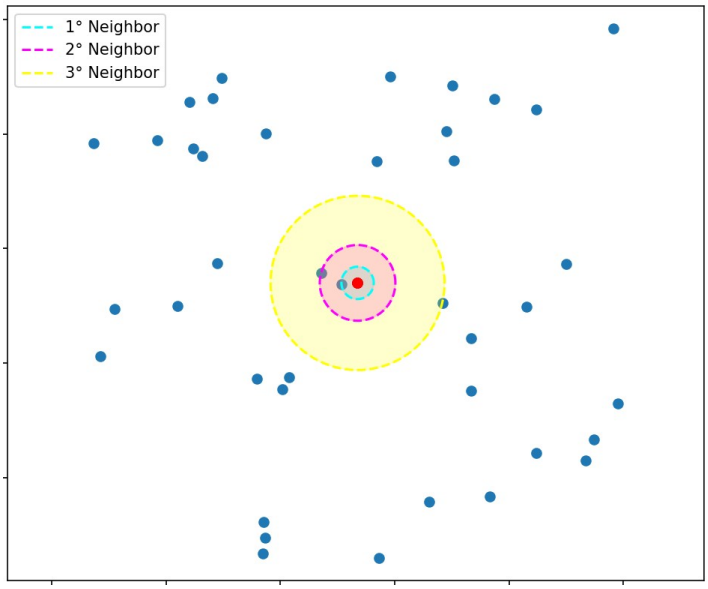


Figura 2.1: Proiezione sul piano delle tre ipersfere date dalle prime tre distanze più brevi da un determinato punto.

Nel prossimo paragrafo, sarà introdotto uno stimatore di ID che utilizza un algoritmo di *nearest neighbors*: il modello **TWO-Nearest Neighbors**

2.3 Algoritmo TWO-Nearest Neighbors

Lo stimatore di dimensione intrinseca **TWO-Nearest Neighbors** è un modello basato sul metodo geometrico dei *nearest neighbors*. Esso sfrutta le due distanze più brevi da ciascun punto ai suoi vicini, minimizzando la dipendenza dell'ID dalla densità. Assumere che la densità sia uniforme fino al secondo vicino consente al modello di calcolare sia la distribuzione che la funzione cumulativa del rapporto tra la seconda e la prima distanza. Come proveremo in seguito, queste distribuzioni dipendono dalla dimensione intrinseca d ma sono indipendenti dalla densità.
Approssimando la distribuzione cumulativa con la funzione cumulativa empirica calcolata sul data-set, è possibile stimarne la dimensione intrinseca. La legittimità di questo metodo, nel caso si abbia a che fare con insiemi caratterizzati da densità non uniforme e curvatura ad esempio il set di dati MNIST [11], è stata dimostrata in FACCO et al.[20], e un'analisi del data-set menzionato è fornita nella sezione 2.3.2.

2.3.1 I risultati matematici dell'algoritmo TWO-NN

Supponendo che i sia un punto generico all'interno del data-set, possiamo esaminare l'elenco ordinato dei primi k vicini più prossimi ad i, e denotarne le distanze come $r_1, r_2, ..., r_k$, dove r_1 indica la distanza dal punto più vicino, r_2 dal secondo e così via. In questa definizione r_0 è posto uguale a zero.

La formula per il calcolo del volume del guscio ipersferico situato tra due vicini consecutivi, indicati come $l-1$ ed l, può essere definita come segue:

$$\Delta\nu_l = \omega_d(r_l^d - r_{l-1}^d) \tag{2.1}$$

dove d rappresenta la dimensione dello spazio in cui esistono i punti ed ω_d è il volume dell'ipersfera di raggio unitario.

Se la densità $\rho(x) = \rho$ per ogni x o, in maniera analoga, se il processo di Poisson che definisce la distribuzione dei punti è omogeneo [15], allora tutte le grandezze $\Delta\nu_l$ sono indipendenti ed identicamente distribuite come variabili aleatorie esponenziali con il parametro di intensità pari alla densità ρ:

$$P(\Delta\nu_l \in [\nu, \nu + d\nu]) = \rho e^{-\rho\nu} d\nu. \tag{2.2}$$

Dalle precedenti ipotesi, e definendo $R = \frac{\Delta\nu_i}{\Delta\nu_j}$, è possibile calcolare la funzione di distribuzione di probabilità (PDF) di R come:

$$\begin{aligned} P(R \in [R', R' + dR']) &= \int_0^\infty d\nu_i \int_0^\infty \rho^2 e^{-\rho(\nu_i+\nu_j)} 1_{\{\frac{\nu_j}{\nu_i} \in [R', R'+dR']\}}, d\nu_j \\ &= dR' \frac{1}{(1+R')^2}, \end{aligned}$$

in cui $1_{\{\frac{\nu_j}{\nu_i} \in [R', R'+dR']\}}$ è la funzione caratteristica dell'intorno $[R', R' + dR']$. Dividendo rispetto a dR', la PDF diviene:

$$g(R) = \frac{1}{(1+R)^2}. \tag{2.3}$$

Successivamente, viene definito il rapporto μ come $\mu \doteq \frac{r_2}{r_1} \in [1, +\infty]$. Questo passaggio consentirà di scrivere la funzione di distribuzione cumulativa (CDF) in modo che dipenda esplicitamente da d. A questo punto, notando che:

$$R = \mu^d - 1, \tag{2.4}$$

è immediato determinare che la funzione di distribuzione di μ sia:

$$f(\mu) = d\mu^{-d-1} 1_{[1,+\infty]}(\mu), \tag{2.5}$$

mentre, per definizione, la distribuzione cumulativa è ottenibile integrando la PDF:

$$F(\mu) = (1 - \mu^{-d}) 1_{[1,+\infty]}(\mu). \tag{2.6}$$

Questo risultato evidentemente non dipende dalla densità locale, ma dipende esplicitamente dalla dimensione intrinseca d. Pertanto, per stimare la dimensione intrinseca, si può derivare la seguente formula:

$$d = -\frac{\log(1 - F(\mu))}{\log(\mu)}. \tag{2.7}$$

Questa equazione implica che l'insieme di punti $S \doteq \{(log(\mu), -log(1 - F(\mu)))\} \subset \mathbf{R^2}$ giace sulla linea $l \doteq \{(x, y)|y = d * x\}$ passante per l'origine.

Come conseguenza dell'equazione (2.7), il data-set dovrà avere densità uniforme soltanto *localmente*, ovvero dovrà essere uniforme nell'intervallo del secondo vicino. La validità del risultato persiste anche nel caso in cui il data-set sia composto da un numero limitato di punti, infatti le simulazioni numeriche condotte da Facco et al [20] dimostrano che, anche per un numero finito di punti, la stima dell'ID tramite TWO-NN in data-set non uniformi produce misure accurate.

Richiedere l'uniformità locale nell'intervallo del secondo vicino consente un'analisi più flessibile ed adattiva. Al contrario, i metodi che richiedono l'uniformità su scale più ampie, impongono condizioni più rigide sui dati. Questa rigidità può limitare l'applicabilità dei modelli d'analisi, specialmente quando i dati presentano variazioni su scale maggiori.
La dimostrazione qui riportata è liberamente ispirata al lavoro di [20].

2.3.2 Stima dell'ID applicata al MNIST database

In questa sezione viene proposta un'analisi sui dati MNIST [11] per illustrare il comportamento dell'algoritmo TWO-NN in dataset più complessi.

Il MNIST è un ampio database di immagini di cifre scritte a mano, ciascuna composta da 28x28 pixel, accompagnate dalle rispettive etichette. È ampiamente utilizzato per la verifica delle prestazioni su Reti Neurali Artificiali. Il database è costituito da 60.000 esempi di addestramento e 10.000 esempi di test.

Per semplicità, il metodo TWO-NN è stato applicato ad un sottoinsieme di questi dati, precisamente a tutte le immagini corrispondenti all'etichetta "7". Dai risultati mostrati nella sezione 2.3.1, è possibile definire un algoritmo per trovare l'ID nel seguente modo:

- Calcolare la distanza componente per componente su ciascun punto nel data-set $i = 1, ..., N$, dove la distanza è definita come:

$$dist(x_i, x_j) = ||x_i - x_j + \epsilon||_p \tag{2.8}$$

dove $||\cdot||_p$ è la p-norma: $||x||_p = (\sum_{i=1}^{N} |x_i^p|)^{\frac{1}{p}}$

- Trovare per ogni punto le due distanze più brevi r_1 ed r_2.
- Calcolare $\forall\ i$, $\mu_i = \frac{r_2}{r_1}$
- Disporre i punti in ordine crescente attraverso una permutazione σ.
- Definire e misurare la funzione cumulativa empirica :

$$F^{emp}(\mu_{\sigma_i}) \doteq \frac{i}{N} \tag{2.9}$$

- Graficare i valori sul piano dato da $\{(\log(\mu_i), -\log(1 - F^{emp}(\mu_i)|i = 1, ..., N\}$ ed eseguire un fit con una retta passante per l'origine.

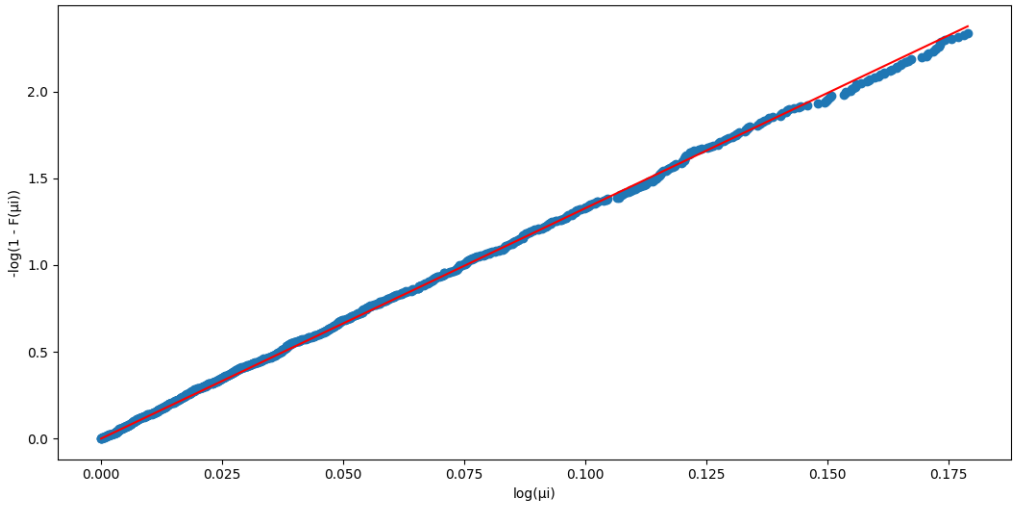


Figura 2.2: Sono stati utilizzati 1032 esempi di cifre scritte "7", ognuno dei quali ha 784 componenti, per mostrare i risultati nella sezione 2.3.1. La pendenza di questa retta rappresenta la dimensione intrinseca (ID) del sottoinsieme, ed è pari a $d = 13,27$. Questo valore è in linea con lavori precedenti [7] e [8], i quali stabiliscono che l'ID si trova tra 12 e 14.

È importante notare che il 10% dei punti con i valori più elevati di μ è stato scartato al fine di migliorare il processo di fit. Implementando l'algoritmo sull'intero set di dati, la dimensione stimata è risultata essere 14,9. La causa di questa leggera sovrastima va ricercata nella presenza di rumore.

La tecnica TWO-NN ha dimostrato una ragionevole robustezza alle fluttuazioni di densità e, nel complesso, si è rivelata un accurato stimatore dell'ID.

2.4 Analisi di correlazione tramite dimensionalità intrinseca

Per indagare il grado di somiglianza tra diversi segnali, è necessario quantificare la corrispondenza tra i due canali di input che compongono ciascuna matrice. Al fine di raggiungere questo obiettivo, è essenziale calcolare le correlazioni tra gli elementi che caratterizzano ciascun segnale. A causa della natura non-lineare di tali caratteristiche e del fatto che questi insiemi sono costituiti da variabili a più componenti, la tecnica basata sul coefficiente di correlazione di Pearson R^2 non può essere implementata correttamente nel trovare pattern di connettività. Questa limitazione richiede un metodo in grado di aggirare tali vincoli e misurare correttamente la correlazione tra gli input.
La dimensione intrinseca di un data-set è fondamentalmente legata al rapporto tra le diverse proprietà che caratterizzano i dati. Queste correlazioni stabiliscono la rappresentazione associata al sottospazio entro cui risiedono i punti, e la dimensione di questa varietà rappresenta l'ID dell'insieme di dati.

2.5 Stima dell'ID per un data-set a due variabili

In questo paragrafo viene illustrato il procedimento che stabilisce l'esistenza di correlazioni in un data-set a due variabili tramite osservazioni svolte sulla variazione di dimensione intrinseca. Questa analisi è tratta da [17].
Per quanto detto finora, nel caso in cui due grandezze non presentino alcuna correlazione lineare, il coefficiente di Pearson tenderà a zero ed i punti del data-set saranno distribuiti su di un piano, viceversa, se una delle variabili è funzione lineare dell'altra, allora R^2 assumerà approssimativamente il valore uno, ovvero la dimensione intrinseca del data-set corrisponderà ad uno.
Qualora le grandezze in gioco non abbiano alcuna relazione di natura lineare, i vantaggi del metodo basato sulla dimensione intrinseca diventano più espliciti. Infatti, si ha che $R^2 \approx 0$, mentre la stima della dimensione intrinseca è invariante.
L'approccio che permette di inferire la presenza di correlazioni in un data-set osservando il comportamento dell'ID può essere descritto in quattro punti principali:

- Dato un data-set a due variabili, viene stimata la sua dimensione intrinseca.

- Si permuta l'ordine degli elementi di una delle due variabili, di fatto eliminando qualunque correlazione ci fosse tra le componenti dei due sotto-insiemi.

- Si misura la dimensione intrinseca dei vari data-set con variabili non più correlate e se ne calcolano la media e la rispettiva deviazione standard.

- Si conduce uno Z-test ad una coda per determinare la probabilità che la dimensione intrinseca del data-set originale sia significativamente inferiore alla media dei data-set generati con permutazioni. Infatti, rendendo i punti del data-set

meno correlati tra loro la ID corrispondente aumenta poiché essi potranno essere posizionati al di fuori della varietà topologica di partenza.

Permutando uno dei due set di variabili, le distribuzioni di probabilità delle incognite $p(x_1)$ e $p(x_2)$ rimangono inalterate, mentre la distribuzione di probabilità congiunta $p(x_1, x_2)$ è tale che $p(x_1, x_2) = p(x_1)p(x_2)$ se non sono presenti correlazioni. Tuttavia questa definizione non vale nel caso in cui le variabili presentino correlazioni non lineari. Pertanto, esaminando la distribuzione di probabilità congiunta prima e dopo le permutazioni, si può studiare l'esistenza di correlazioni tra le coordinate.

Di seguito è riportato un esempio tratto da [17] che espone come questo metodo permetta di rilevare correlazioni in un data-set non lineare.

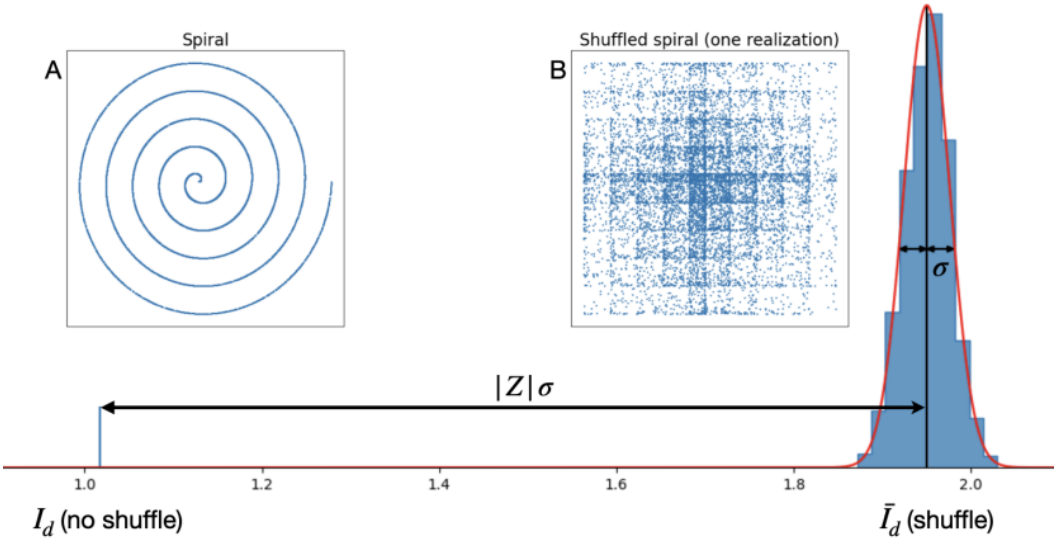


Figura 2.3: I dati sintetici in A non hanno correlazione lineare, permutando i punti sull'asse y, la distribuzione congiunta cambia e la dimensione intrinseca del nuovo data-set aumenta. Questo è rappresentato dalla distribuzione dei valori delle ID rilevate sul secondo data-set. Per definizione di Z-score, l'intervallo tra ID ed $ID_{shuffle}$ è la distanza tra le dimensioni dei due insiemi di dati.

Capitolo 3

Metodi

3.1 L'algoritmo

Lo scopo di questo progetto di ricerca è quello di ottenere un algoritmo che analizzi grandi quantità di dati, e sia in grado di determinare l'esistenza di correlazioni non-lineari attraverso lo studio della dimensione intrinseca della varietà geometrica su cui giacciono i punti del data-set.
Il programma riportato in questo paragrafo è stato progettato al fine di sintetizzare i modelli matematici descritti nei paragrafi precedenti in un codice implementabile su computer.

3.1.1 Il codice

Il codice scritto in linguaggio Python [10] implementa un'analisi per valutare la correlazione non lineare tra coppie di variabili in un data-set multivariato, attraverso il calcolo ed il confronto statistico della dimensionalità intrinseca.
Esso può essere descritto in pochi passaggi:

- Viene letto il data-set multivariato da file CSV (*Comma Separated Values*). I nomi dei file vengono estratti e filtrati secondo un pattern.

- Vengono definite le variabili d'interesse per l'analisi, ovvero le colonne del data-set su cui calcolare l'ID. Inoltre, viene inizializzata una struttura dati (Dataframe Pandas) per contenere i risultati dell'analisi su tutte le possibili coppie di variabili.

- Vengono definite due funzioni: una per calcolare l' ID sui dati originali, una per calcolare l'ID su versioni randomizzate dei dati.

  In particolare, il comando **.compute.id.2NN** fa parte della libreria *DADApy* [16] ed utilizza l'algoritmo TWO-NN del paragrafo 2.3.

- Il file è caricato nel dataframe, vengono poi generate N versioni randomizzate del data-set. Successivamente sono calcolate le dimensioni intrinseche delle versioni randomizzate e dei dati originali.
- Vengono derivate statistiche descrittive dalle ID randomizzate, nello specifico, sono calcolate media e deviazione standard di queste ID.
- Si calcola uno Z-score per quantificare la deviazione dell'ID originale dalla distribuzione delle ID rilevate sui data-set casuali.
- Lo Z-score ed il rispettivo *p-value* vengono inseriti nel dataframe per la coppia di variabili analizzata.
- La struttura dati contenente i risultati viene visualizzata come heatmap ed esportata in una tabella su file CSV.

Da notare che è stato usato un comando della libreria **joblib** di Python per parallelizzare le operazioni di calcolo su tutte le CPU virtuali del computer. In questo modo il tempo d'esecuzione del programma è stato ridotto esponenzialmente.

3.1.2 Heatmap

Una *heatmap* è una rappresentazione visiva di dati in forma di mappa di calore, comunemente usata per visualizzare i valori di una matrice.
Le proprietà di una heatmap sono:

- Rappresenta i dati in forma di griglia rettangolare, in cui ogni cella corrisponde ad un valore.
- Usa una scala di colori per codificare i valori delle celle
- Ha due assi cartesiani che definiscono le etichette delle righe e delle colonne della griglia.
- Permette di visualizzare pattern nei dati, come cluster, outlier e gradienti.
- Può rappresentare dati sia simmetrici che antisimmetrici rispetto alla diagonale.
- Può includere una maschera per nascondere elementi non significativi.

In definitiva, i vantaggi delle heatmap sono la capacità di visualizzare in modo intuitivo grandi quantità di dati multidimensionali, identificare pattern complessi e comunicare efficacemente relazioni tra variabili.

3.1.3 Proprietà dei data-set per l'analisi di correlazioni non lineari

L'algoritmo presentato per lo studio della correlazione tra coppie di variabili attraverso il calcolo della dimensionalità intrinseca si presta ad essere applicato a data-set con specifiche caratteristiche.
I dati devono essere numerici e preferibilmente in un numero elevato di campioni.

L'algoritmo può tecnicamente essere applicato anche ad insiemi di dati ridotti, ma l'importanza di avere una grande mole di punti risiede nel comportamento del modello TWO-NN che, se si hanno a disposizione più dati (con distanze medie tra il punto in considerazione ed il suo secondo vicino minori), calcola una dimensione intrinseca più accurata. Infatti in caso di un numero ridotto di punti, la ID tenderebbe inevitabilmente ad uno, un risultato in linea con il noto assioma geometrico per il quale due punti sono sempre contenuti in una linea.
I dati dovrebbero idealmente rappresentare una serie temporale, dove sia presente una dipendenza sequenziale tra i campioni. La presenza di rumore, sia esso derivante dal processo generativo o da quello di misura, è utile in quanto enfatizza le differenze tra segnale originale e randomizzato nel calcolo dell'ID.
Infine per ottenere risultati statisticamente significativi è necessario che i data-set contengano un numero sufficiente di osservazioni.

3.1.4 Lo Z-test per l'analisi statistica

Lo Z-test è uno strumento statistico utilizzato per verificare ipotesi e fare inferenza su una popolazione sulla base di un campione. In particolare, consente di determinare se la media di un campione differisce in modo statisticamente significativo dalla media di una popolazione di riferimento.
Lo Z-test si basa sulla normalizzazione della differenza tra la media campionaria e la media di popolazione, attraverso la divisione per la deviazione standard σ. Questa normalizzazione permette di esprimere la distanza tra i valori in termini di σ. Questa divisione ne consente un uso efficace anche in analisi di grandi moli di dati multidimensionali.
Il risultato è uno Z-score, un punteggio adimensionale che quantifica, in termini normalizzati, di quante deviazioni standard la media del campione si discosta da quella della popolazione. Maggiore è lo Z-score in valore assoluto, maggiore è la significatività della differenza tra le medie.
Ipotizzando che la variabile analizzata segua una distribuzione normale, è possibile derivare il *p-value*, che quantifica la probabilità di ottenere per caso uno Z-score almeno maggiore di quello osservato. Convenzionalmente si assume significativo un *p-value* minore di 0.05.
La validazione statistica delle misure di correlazione è spesso definita in letteratura adottando una soglia arbitraria predefinita [18]. Il vantaggio dell'algoritmo descritto in questa tesi è dato dal fatto che lo z-score ed il corrispettivo p-value forniscono una soglia di significatività della correlazione definita dalla distribuzione della stima dell'ID, in questa maniera il criterio di validazione possiede basi empiriche piuttosto che essere scelto arbitrariamente.

Nel contesto dell'algoritmo proposto, lo Z-test consente di quantificare in modo standardizzato la deviazione σ dell'ID di un segnale originale rispetto a 35 distribuzioni surrogate generate mediante randomizzazione. I segnali così prodotti sono coerenti con l'ipotesi nulla di assenza di correlazione. Ciò permette di mappare visivamente e identificare interconnessioni più significative tra coppie di variabili.

Capitolo 4

Applicazione sperimentale

L'algoritmo presentato è stato applicato nell'analisi di correlazioni tra i segnali di coppie di elettrodi in registrazioni EEG (elettroencefalografia) effettuate su soggetti umani durante l'esecuzione di specifici compiti.

4.1 Il data-set EEG per l'analisi di correlazione

L'EEG rileva l'attività elettrica cerebrale tramite elettrodi posizionati sullo scalpo, fornendo una misura indiretta della dinamica dei segnali neurali con risoluzione temporale dell'ordine dei millisecondi. Un segnale EEG è una sequenza temporale di valori di potenziali elettrici. Il potenziale è misurato rispetto ad un elettrodo di riferimento. La connettività strutturale tra popolazioni di neuroni, stimata tramite correlazioni tra segnali, riflette il grado di interazione reciproca tra le regioni cerebrali sottostanti. Tali interazioni neurali su ampia scala sottendono l'integrazione di informazioni necessarie per funzioni cognitive complesse.
Per questo lavoro è stato utilizzato il *EEG Motor Movement/Imagery Data-set* [19][21]. Questo data-set consiste in oltre 1500 registrazioni EEG di uno o due minuti ciascuna, ottenute da 109 volontari.

4.1.1 Protocollo Sperimentale

I soggetti hanno eseguito diverse attività motorie/ immaginarie durante la registrazione dell' elettroencefalogramma a 64 canali utilizzando il sistema BCI2000 [21]. Ciascun paziente ha eseguito 14 sessioni sperimentali: due sessioni di un minuto come baseline e tre sessioni di due minuti per ognuna delle quattro seguenti attività:

- Compare un target a sinistra o a destra dello schermo. Il soggetto apre e chiude il pugno corrispondente fino a quando il bersaglio scompare. Successivamente il soggetto si rilassa.
- Compare un target a sinistra o a destra dello schermo. Il soggetto immagina di aprire e chiudere il pugno corrispondente fin quando il bersaglio scompare.

- Compare un bersaglio in alto o in basso allo schermo. Il soggetto apre e chiude entrambi i pugni o entrambi i piedi a seconda che il target sia in alto o in basso.
- Compare un bersaglio in alto o in basso allo schermo. Il soggetto immagina di aprire e chiudere entrambi i pugni o entrambi i piedi a seconda che il target sia in alto o in basso.

Sono presenti 109 file corrispondenti ad ogni paziente, ognuno dei file contiene le 14 sessioni ed ognuna delle sessioni comprende i 64 segnali degli elettrodi ognuno con una frequenza di campionamento di 160 misure al secondo.
Gli elettrodi sono posizionati sul cranio nella maniera esposta nell'immagine che segue:

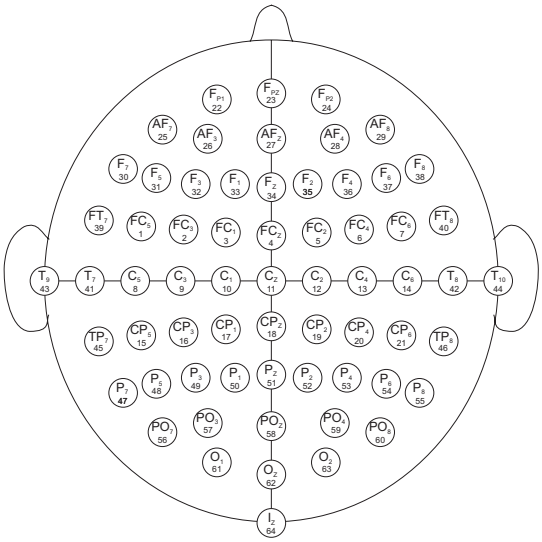


Figura 4.1: Gli elettrodi sono disposti secondo il sistema internazionale 10-10 [13]

4.2 Risultati Sperimentali

In questa sezione sono illustrati i risultati dello studio svolto sull'*EEG Motor Movement/Imagery Data-set* [19][21]. In particolare, sono presentati i risultati derivanti da due metodologie diverse: l'analisi di correlazione tramite l'indice di Pearson e l'approccio basato sulla stima della dimensionalità intrinseca.
Successivamente, discuteremo i risultati sperimentali ottenuti e il confronto tra i due modelli d'analisi.

4.2.1 Analisi di correlazione tramite stima dell'ID

L'algoritmo di analisi presentato è stato applicato nello studio della connettività strutturale e funzionale di dati EEG registrati in due diversi contesti sperimentali. La connettività strutturale può essere intesa come la rappresentazione delle traiettorie degli impulsi elettrici che attraversano le reti neurali, mentre il concetto di connettività funzionale è definito in termini di connessioni statistiche [4] tra diverse attività cerebrali in posizioni anatomiche differenti. Quest'ultima è studiata in analisi di correlazione o covarianza dei dati EEG [18]. Nel primo studio, l'obiettivo è di caratterizzare le dinamiche neurali emergenti in più soggetti durante un compito motorio specifico: la terza attività motoria riportata nel paragrafo 4.1.1. Sono state utilizzate le registrazioni dei segnali EEG rilevate durante l'esecuzione ripetuta della medesima attività. Lo studio ha reso possibile la rilevazione di importanti correlazioni nei segnali di elettrodi situati sulla corteccia motoria primaria. In particolare, l'analisi tramite la stima della ID ha evidenziato un aumento significativo della complessità delle interazioni funzionali tra le reti neurali responsabili dell'esecuzione dei movimenti rispetto a quanto misurato in condizioni di riposo.
Nel secondo studio, è stato analizzato un singolo soggetto sottoposto a diverse condizioni sperimentali le quali includono sia compiti cognitivi che motori. La comparazione delle coppie di elettrodi che mostrano un più alto grado di correlazione ha consentito di discriminare reti neurali distinte attivate da diversi compiti.
La scelta di condurre due studi diversi, uno focalizzato sulla variazione inter-soggetto (Studio A) e l'altro sulla variazione inter-attività (Studio B), è motivata dall'intento di dissociare le correlazioni rivelate da un'analisi relativa alle caratteristiche neurofisiologiche individuali da quelle derivanti dalle diverse dinamiche dei compiti cognitivi.
Nel primo studio, l'applicazione del medesimo paradigma motorio in più soggetti ha permesso di identificare pattern di connettività comuni, emergenti durante l'esecuzione del compito e così attribuibili principalmente alla rete neurale responsabile della funzione motoria condotta.
Al contrario, nello studio con singolo soggetto la variazione sistematica dell'attività svolta dal paziente ha evidenziato distinti schemi di connessione, indicativi di configurazioni di segnali cerebrali diversi.
Nonostante le differenze di pattern di connettività tra studi, in entrambi i casi l'approccio presentato in questo lavoro ha consentito di identificare correttamente le zone in cui i segnali riportano un'elevata correlazione nelle aree della corteccia cerebrale dedite alla pianificazione ed all'esecuzione di attività motorie e cognitive: la corteccia

motoria primaria e l'area motoria supplementare localizzata nella faccia mediale del lobo frontale [6].

4.2.2 Analisi di correlazione tramite il coefficiente di Pearson

Un'ulteriore analisi sul data-set è stata condotta usando il coefficiente di correlazione lineare di Pearson. Questo metodo ha fornito un benchmark classico per confrontare i risultati ottenuti con nuovo metodo .
Il coefficiente di Pearson è definito come :

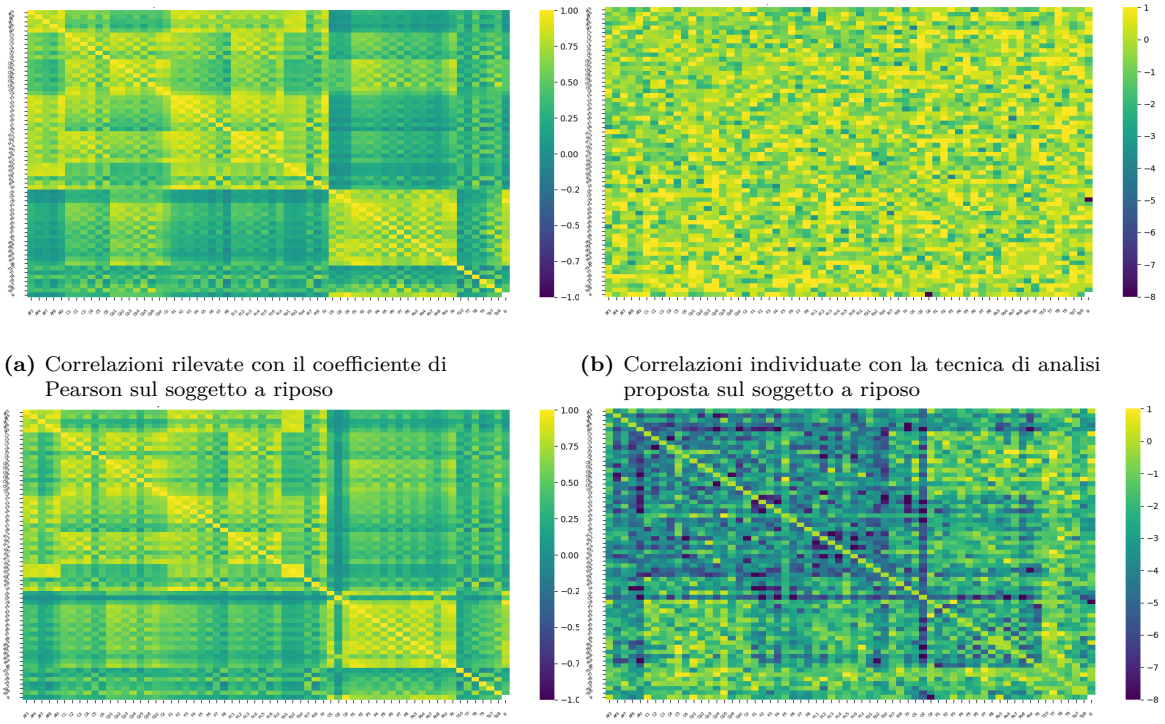
$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{4.1}$$

Dove σ_{XY} è la covarianza tra X ed Y variabili statistiche e σ_X, σ_Y sono le rispettive deviazioni standard. Per definizione il coefficiente può assumere valori tra -1 ed 1. Un valore di 1 o -1 corrisponde a punti che si trovano interamente su di una retta, quindi corrispondono ad una correlazione lineare totale.
L'approccio qui riportato si basa sull'assunzione che la relazione tra le variabili sia approssimativamente lineare. Questo significa che qualsiasi deviazione dalla linearità può portare ad una stima errata della correlazione.
Il calcolo è stato eseguito tra tutte le possibili coppie di elettrodi relativi agli stessi soggetti dell'analisi di correlazione tramite stima della dimensione intrinseca.

4.2.3 Risultati dello Studio Inter-soggetto

I segnali EEG presentano inevitabilmente un alto grado di rumore, non stazionarietà e non-linearità, che riflettono la complessità dell'attività celebrale. L'analisi tramite la stima della dimensionalità intrinseca fornisce una misura robusta di tali interrelazioni. Confrontando infatti le correlazioni catturate durante gli stati di riposo con quelle misurate durante l'esecuzione del compito motorio, si nota come la tecnica di analisi dati tramite coefficiente di Pearson non sia in grado di discriminare con sufficiente accuratezza la variazione di attività neurale determinata dall'effettuazione della task, quest'osservazione è in linea con quanto riportato nel lavoro di Sakkalis V. et al. [23] . Pertanto l'implementazione di questo algoritmo sui dati EEG ha permesso di caratterizzare le dinamiche di connessione indotte da uno specifico compito motorio, rivelando schemi di correlazione tra i segnali neurali multicanale ed isolando le casistiche con successo. Da notare che le heatmap dei coefficienti di Pearson 4.2(a) e 4.2(c) sono tuttavia più robuste al rumore rispetto a quelle che si basano sulla analisi proposta, per questo motivo infatti, la heatmap di correlazione riportata in 4.2(a) riesce a trovare pattern anche nelle misure effettuate in condizioni di riposo.
Nell'approccio presentato, la quantificazione delle correlazioni tra le coppie di segnali viene effettuata tramite uno Z-test che confronta la dimensionalità intrinseca originale con la distribuzione delle ID di versioni degli stessi dati, ma modificate mediante

permutazione casuale. Valori negativi dello z-score indicano che l'ID originale è significativamente inferiore alla media delle ID surrogate. Ciò implica che la randomizzazione ha distrutto le correlazioni presenti nei dati originali. Quindi a z-score negativi corrispondono correlazioni più intense tra le serie temporali in esame, ciò rivela un grado maggiore di interazione funzionale tra le reti neurali associate a determinati segnali.



(a) Correlazioni rilevate con il coefficiente di Pearson sul soggetto a riposo

(b) Correlazioni individuate con la tecnica di analisi proposta sul soggetto a riposo

(c) Correlazioni rilevate con il coefficiente di Pearson sul soggetto attivo

(d) Correlazioni individuate con la tecnica di analisi proposta sul soggetto attivo

Figura 4.2: Le heatmap sono caratterizzate da una scala monocromatica riportata in legenda. Nelle suddette mappe, le celle visualizzate con tonalità più scure indicano valori di z-score più negativi, riflettendo quindi una correlazione maggiore tra i segnali delle coppie che rappresentano. In contrasto, nelle heatmap relative all'analisi tramite coefficiente di Pearson, una correlazione lineare più elevata è rappresentata da celle con colorazioni più chiare. Sugli assi cartesiani sono riportati in ordine alfabetico le etichette dei 64 elettrodi.
(a) Heatmap dei coefficienti di Pearson ottenuti dalle misure svolte sul soggetto 2 in condizioni di riposo. (b) Heatmap delle correlazioni rilevate tramite stima della dimensione intrinseca. Le misure sono relative allo stato di riposo osservato nel soggetto 2. (c) Heatmap dei coefficienti di Pearson ottenuti dalle misure svolte sul soggetto 2 durante il compito motorio. (d) Heatmap delle correlazioni tra coppie di elettrodi stimate con il metodo della stima dell'ID.
Le misure sono state rilevate durante l'attività motoria del soggetto 2. I risultati dell'analisi proposta relativi ai soggetti rimanenti sono visualizzabili nel capitolo 6.

Il blocco di z-score ad elevata intensità nel primo quadrante delle heatmap 4.2(c) e 4.2(d) evidenzia un buon grado di correlazione tra le popolazioni di neuroni nella corteccia motoria primaria, come ci si aspettava da segnali tratti durante un compito motorio [6]. A riprova di questa affermazione, la figura 4.3 illustra le connessioni con intensità di correlazione maggiore, misurate con entrambe le tecniche. I risultati presentano similitudini nella fascia centrale di elettrodi, tuttavia in figura 4.3(b) si notano distintamente correlazioni tra gli elettrodi delle fasce laterali. Questi elettrodi sono stati collegati all'attivazione di reti neurali che governano il movimento delle mani in Saby et al. [12].

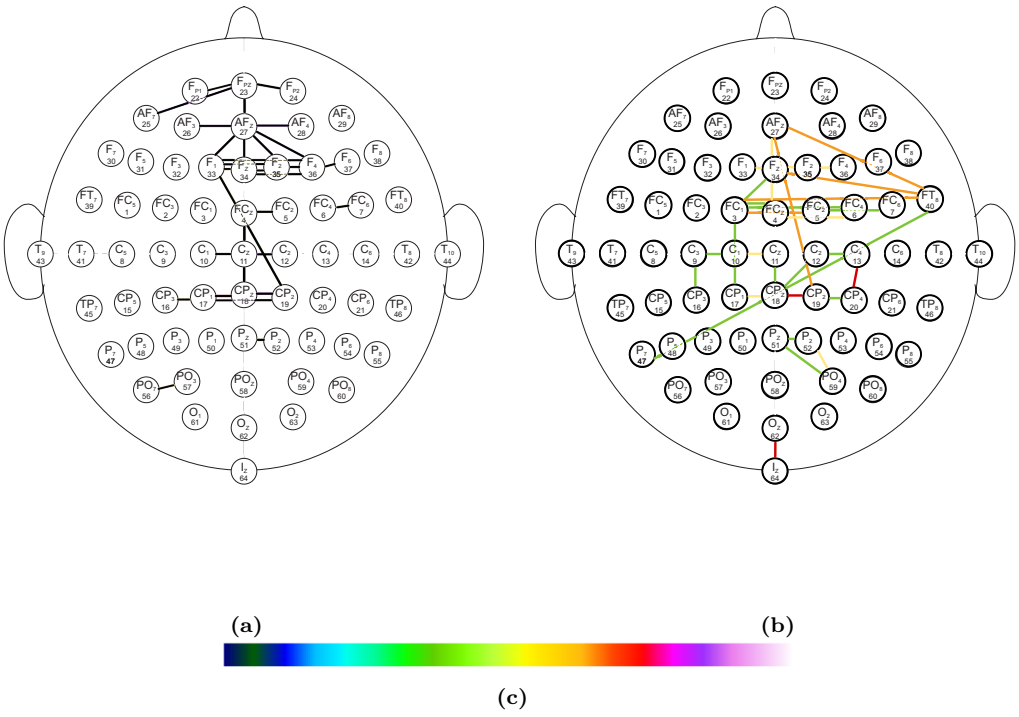
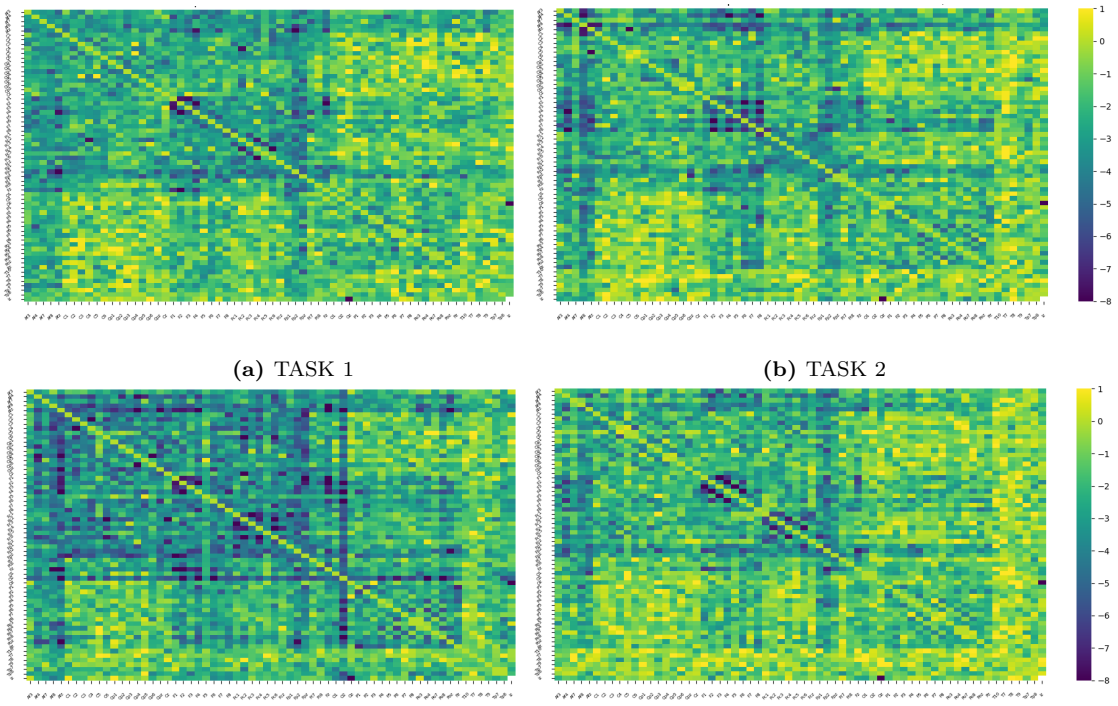


Figura 4.3: (a) Mappa delle connessioni più intense rivelate dall'analisi tramite indice di Pearson. (b) Mappa delle connessioni più significative rivelate dall'analisi tramite stima dell'ID, la scala di colori indica l'intensità della correlazione secondo la colormap riportata in (c).

4.2.4 Risultati dello Studio Inter-attività

La scelta di integrare l'analisi inter-soggetto sul singolo compito con uno studio inter-attività sul singolo soggetto permette di sottolineare le variazioni delle misure di correlazione tra tipi di dataset diversi.

Le heatmap 4.4(b) e 4.4(d) relative ai compiti immaginari mostrano una minore strutturazione delle connessioni nelle regioni corticali centrali rispetto ai compiti motori. In particolare, si nota l'assenza di correlazioni forti tra elettrodi della fascia "F" che sono i sensori la cui attivazione è generalmente collegata al movimento di mani e piedi. Ciò suggerisce un ruolo critico di queste aree associate all'integrazione sensomotoria effettiva come d'altronde suggerito in Dowman R. et al.[6].



(a) TASK 1 **(b)** TASK 2

(c) TASK 3 **(d)** TASK 4

Figura 4.4: (a) Heatmap raffigurante l'intensità degli z-score rilevati dal soggetto 2 durante la prima attività motoria. (b) Heatmap raffigurante l'intensità degli z-score rilevato sul soggetto 2 durante la prima attività cognitiva. (c) Heatmap raffigurante l'intensità degli z-score rilevato sul soggetto 2 durante la seconda attività motoria. (d) Heatmap raffigurante l'intensità degli z-score rilevato sul soggetto 2 durante la seconda attività cognitiva.

Per attività cognitive differenti, ci si aspetta una distribuzione spaziale diversa dei segnali con maggior correlazione. Questo risultato è riportato in figura 4.5 in cui sono illustrate le mappe di connessione maggiore tra elettrodi nel caso di compiti motori e in quello di compiti in cui era richiesto al soggetto di pianificare l'azione motoria. Tra i due grafici è possibile distinguere reti di correlazioni diverse in numero di connessioni, intensità e distribuzione spaziale. Nonostante le misure condotte tramite elettroencefalografia siano spesso di difficile interpretazione a causa dell'abbondanza di rumore nei segnali e del *Volume Conduction Problem* [9], ovvero la diffusione globale dei segnali elettrici attraverso tessuti conduttivi nel cervello, i risultati presentati identificano la corteccia motoria primaria e l'area motoria supplementare come le zone in cui i segnali dimostrano un maggiore grado di connessione tra le reti neurali sottostanti.

Le mappe di elettrodi riportate di seguito considerano rispettivamente la somma delle correlazioni tra i segnali delle due attività motorie e quella delle due attività immaginarie.

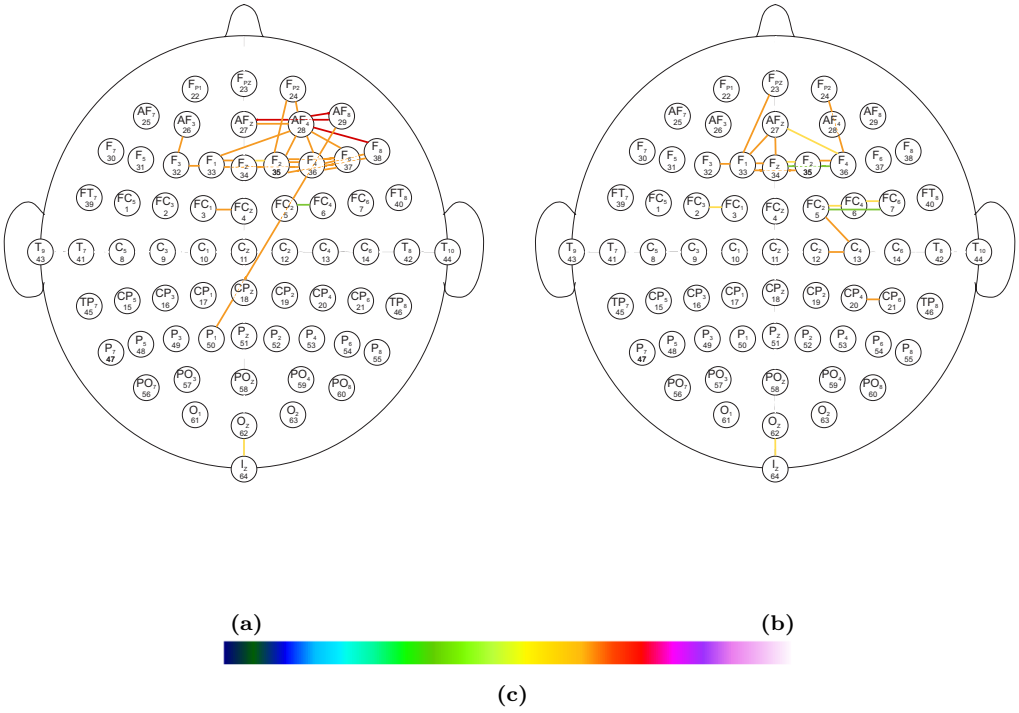


Figura 4.5: (a) Mappa delle connessioni più significative rivelate dall'analisi tramite stima dell'ID durante le attività motorie, la scala di colori indica l'intensità della correlazione secondo la colormap riportata in (c). (b) Mappa delle connessioni più significative rivelate dall'analisi tramite stima dell'ID durante le attività immaginarie.

Capitolo 5

Conclusioni

In questa tesi è stata presentata e convalidata una metodologia innovativa per l'analisi delle correlazioni non lineari applicando un nuovo algoritmo su misure elettroencefalografiche. I risultati ottenuti dimostrano l'efficacia di questo approccio basato sulla stima della dimensione intrinseca del dataset nel rilevare la connettività funzionale associata a specifiche attività cognitive.
In particolare, l'implementazione dell'algoritmo su dati associati a più soggetti ha rivelato pattern comuni di correlazione tra segnali derivanti da misure in zone localizzate del cervello. Ciò potrebbe indicare un coinvolgimento coordinato di regioni neurali associate a compiti motori.
Nel secondo studio, invece, è stata sottolineata la variazione sistematica dei pattern correlazionali in base alla natura del compito, quantificando la dissociazione tra compiti mentali e motori.

Sebbene vi siano opinioni contrastanti sull'accuratezza dell'EEG nel localizzare l'attività cerebrale, la nostra analisi ha consentito l'individuazione di correlazioni importanti tra aree coinvolte nell'esecuzione e pianificazione di compiti cognitivi semplici, fornendo utili informazioni sulla connettività tra regioni neurali distinte.
Tuttavia, vi sono ancora margini di miglioramento ed espansione di questa tecnica.
In primis, per motivi computazionali, lo studio è stato condotto su di un ridotto campione di soggetti. Per aumentare la significatività statistica è necessario implementare l'algoritmo sull'intera popolazione di pazienti disponibile.
Una sfida da affrontare è la riduzione del rumore nei dati EEG dovuto al *Volume Conduction Problem*. Tecniche di filtraggio che impongono vincoli anatomici alle misure effettuate potrebbero attenuare questo effetto.
Inoltre, l'analisi potrebbe essere estesa al dominio delle frequenze, applicando la trasformata di Fourier alle serie temporali e studiando come la connettività si modula in specifiche bande di oscillazione. La comparazione con segnali elettrofisiologici più diretti, come quelli registrati con elettrodi intracranici [14], consentirebbe una validazione più robusta dei pattern di connettività inferiti.
Infine, considerando intere regioni di elettrodi anziché coppie, si potrebbero carat-

terizzare reti neurali complesse e dinamiche associate a funzioni cognitive superiori. Integrando questi sviluppi con modelli che prendano in considerazione soltanto determinate aree cerebrali, sarà possibile decifrare più compiutamente i meccanismi neurali alla base dei compiti cognitivi presi in considerazione in questo studio.

Bibliografia

[1] H. Hotelling. «Analysis of a complex of statistical variables into principal components». In: *Journal of Educational Psychology* 24(6) (1933), pp. 417–441. DOI: https://doi.org/10.1037/h0071325.

[2] Robert S. Bennett. «The intrinsic dimensionality of signal collections». In: *IEEE Trans. Inf. Theory* 15 (1969), pp. 517–525. URL: https://api.semanticscholar.org/CorpusID:21242497.

[3] Keinosuke Fukanaga. *Introduction to statistical pattern recognition*. New York, Academic Press, 1972.

[4] Karl J. Friston. «Functional and effective connectivity in neuroimaging: A synthesis». In: *Human Brain Mapping* 2 (1994). URL: https://api.semanticscholar.org/CorpusID:11977447.

[5] Christopher M. Bishop. «Neural networks for pattern recognition». In: 1995. URL: https://api.semanticscholar.org/CorpusID:60563397.

[6] Robert Dowman e Stephanie Schell. «Innocuous-related sural nerve-evoked and finger-evoked potentials generated in the primary somatosensory and supplementary motor cortices». In: *Clinical Neurophysiology* 110 (1999), pp. 2104–2116. URL: https://api.semanticscholar.org/CorpusID:13311132.

[7] Matthias Hein e Jean-Yves Audibert. «Intrinsic Dimensionality Estimation of Submanifolds in Rd». In: *Proceedings of the 22nd International Conference on Machine Learning*. ICML '05. Bonn, Germany: Association for Computing Machinery, 2005, pp. 289–296. ISBN: 1595931805. DOI: 10.1145/1102351.1102388. URL: https://doi.org/10.1145/1102351.1102388.

[8] J. A. Costa e Alfred O. Hero. «Determining Intrinsic Dimension and Entropy of High-Dimensional Shape Spaces». In: *Statistics and Analysis of Shapes*. 2006. URL: https://api.semanticscholar.org/CorpusID:115174647.

[9] Schoffelen J.M. e Gross J. «Source connectivity analysis with MEG and EEG.» In: *Hum Brain Mapp.* (2009). DOI: 10.1002/hbm.20745.

[10] Guido Van Rossum e Fred L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.

[11] Li Deng. «The mnist database of handwritten digit images for machine learning research». In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.

[12] Saby J., Meltzoff A.N. e Marshall P.J. «Neural body maps in human infants: Somatotopic responses to tactile stimulation in 7-month-olds». In: *NeuroImage* (2015). DOI: https://doi.org/10.1016/j.neuroimage.2015.05.097.

[13] Nuwer M.R. «10-10 Electrode Placement System». In: *Clinical Neurophysiology* (2018). DOI: 10.1016/j.clinph.2018.01.065.

[14] Papadopoulou M., Friston K. e Marinazzo D. «Estimating Directed Connectivity from Cortical Recordings and Reconstructed Sources». In: *Brain Topogr* 32 (2019). DOI: https://doi.org/10.1007/s10548-015-0450-6.

[15] Francesco Denti et al. «Distributional Results for Model-Based Intrinsic Dimension Estimators». In: 2021. URL: https://api.semanticscholar.org/CorpusID:233423601.

[16] Aldo Glielmo et al. «DADApy: Distance-based analysis of data-manifolds in Python». In: *Patterns* (2022), p. 100589. ISSN: 2666-3899. DOI: https://doi.org/10.1016/j.patter.2022.100589. URL: https://www.sciencedirect.com/science/article/pii/S2666389922002070.

[17] Lorenzo Basile et al. «Relating Implicit Bias and Adversarial Attacks through Intrinsic Dimension». In: *ArXiv* abs/2305.15203 (2023). URL: https://api.semanticscholar.org/CorpusID:258865920.

[18] Giovanni Chiarion et al. «Connectivity Analysis in EEG Data: A Tutorial Review of the State of the Art and Emerging Trends». In: *Bioengineering* 10 (2023). URL: https://api.semanticscholar.org/CorpusID:257644289.

[19] Goldberger A. et al. *PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals*. URL: https://physionet.org/content/eegmmidb/1.0.0/.

[20] Facco E. et al. «Estimating the intrinsic dimension of datasets by a minimal neighborhood information». In: *Sci Rep.* (September 2017). DOI: 10.1038/s41598-017-11873-y..

[21] Schalk G. et al. «BCI2000: a general-purpose brain-computer interface (BCI) system». In: *IEEE Trans Biomed Eng.* (2004 Jun), pp. 1034–43. DOI: 10.1109/TBME.2004.827072..

[22] Norman Pigden. «Multidimensional Scaling: History, Theory and Applications». In: *ournal of the Royal Statistical Society Series D: The Statistician* 37 (March 1988), pp. 90–92. DOI: https://doi.org/10.2307/2348396.

[23] Sakkalis V., Doru Giurc Neanu C. e Xanthopoulos P. «Assessment of linear and nonlinear synchronization measures for analyzing EEG in a mild epileptic paradigm.» In: *IEEE Transactions on Information Technology in Biomedicine : a Publication of the IEEE Engineering in Medicine and Biology Society.* (2009 Jul). DOI: 10.1109/titb.2008.923141..

Capitolo 6

Materiale supplementare

Questo capitolo è organizzato nelle seguenti sezioni: il paragrafo 6.0.1 illustra i grafici delle correlazioni stimate tramite l'analisi della dimensione intrinseca relativi al terzo task motorio. I grafici si riferiscono ai soggetti 2, 4, 5 dell' *EEG Motor Movement/Imagery Data-set* [19][21]. Il paragrafo 6.0.2 riporta i grafici delle correlazioni misurate sui segnali catturati durante le condizioni di riposo degli stessi soggetti. Il paragrafo 6.0.3 integra i grafici delle prime due sezioni sottraendo allo z-score relativo al task quello relativo alle condizioni di riposo. Il paragrafo 6.0.4 contiene i grafici ricavati dalle analisi svolte con il coefficiente di Pearson sui tre soggetti in esame durante la task motoria 3. Infine, il paragrafo 6.0.5 include alcuni grafici che hanno aiutano la comprensione dell'analisi della connettività strutturale tra le reti neurali dei soggetti presi in considerazione.

6.0.1 Grafici relativi alla terza attività motoria

I grafici esposti in questa sezione riportano i risultati dell'analisi condotta sui segnali di voltaggio misurati esclusivamente nelle sessioni sperimentali relative al terzo task motorio.
Di grande rilevanza è la presenza di pattern riconoscibili e comuni a tutti e tre i soggetti presi in considerazione.

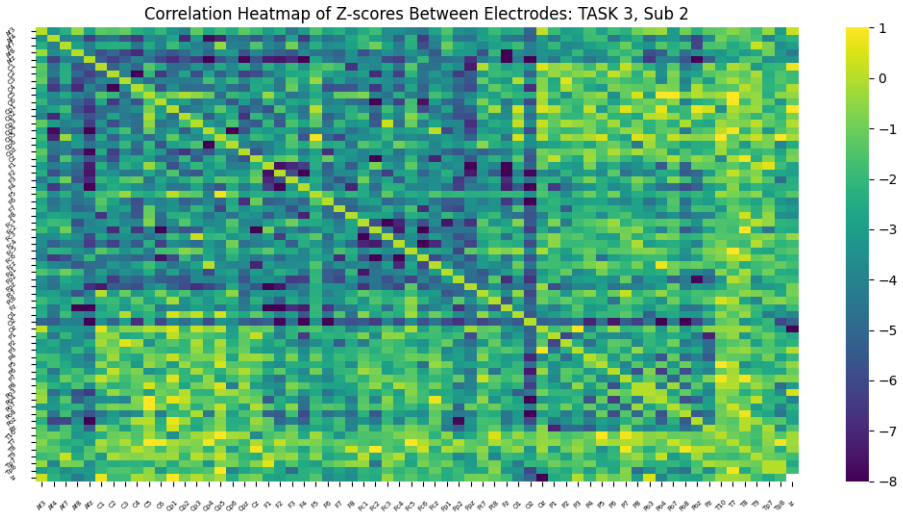


Figura 6.1: Heatmap relativa al paziente 2

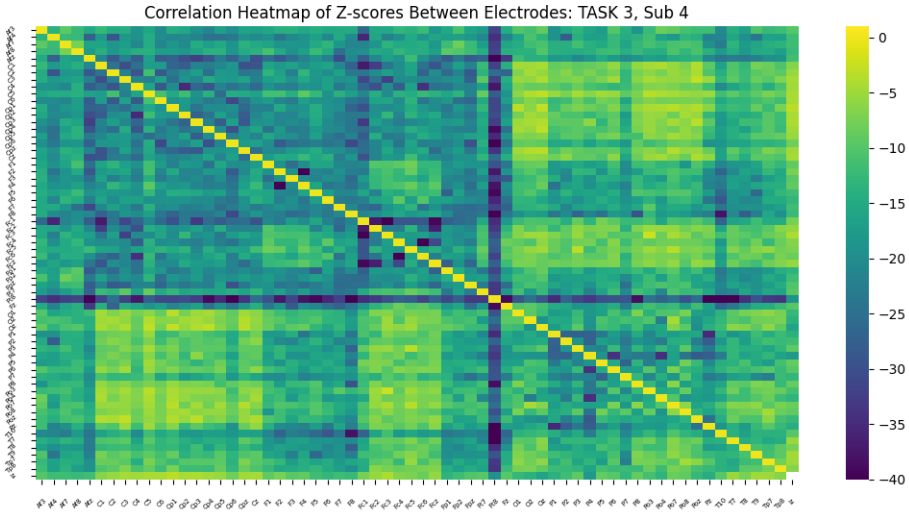


Figura 6.2: Heatmap relativa al paziente 4

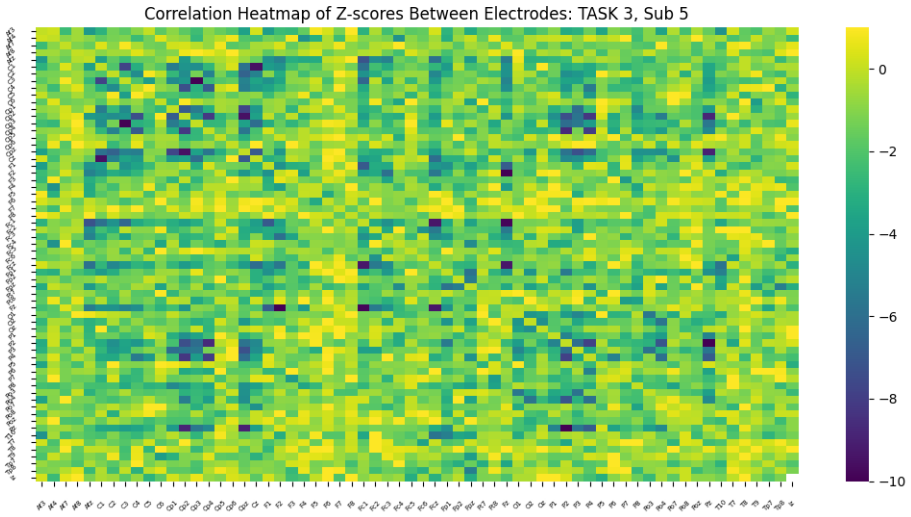


Figura 6.3: Heatmap relativa al paziente 5

6.0.2 Grafici relativi allo stato di riposo

Sono stati studiati anche i segnali relativi agli stati di riposo dei soggetti.

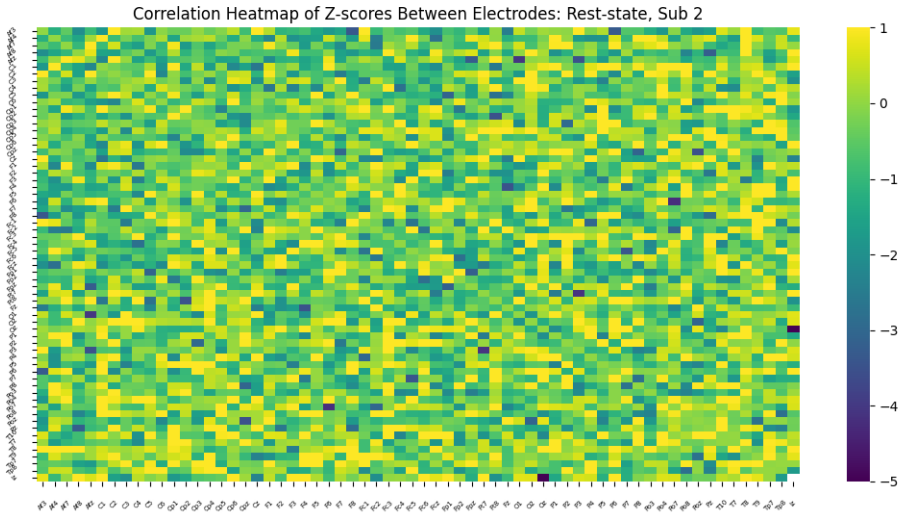


Figura 6.4: Heatmap relativa allo stato di riposo del soggetto 2

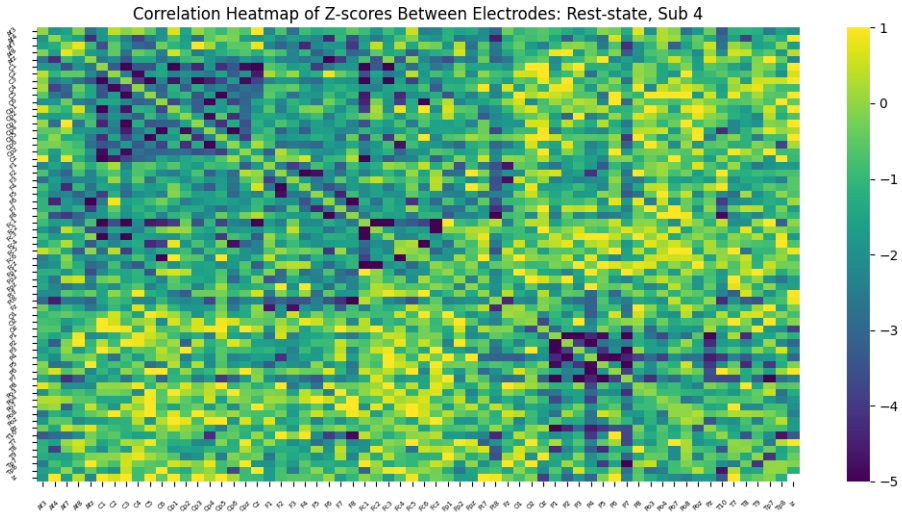


Figura 6.5: Heatmap relativa allo stato di riposo del soggetto 4

Figura 6.6: Heatmap relativa allo stato di riposo del soggetto 5

Osservando la scala di colore dei grafici sopra esposti è palese l'assenza di pattern riconoscibili ed anzi, la quasi totale non-correlazione tra le coppie di elettrodi, come d'altronde ci si aspettava dagli stati di riposo. Il soggetto 4 si conferma essere quello che più degli altri mostra un'elevata correlazione spontanea tra le zone del cervello.Tuttavia, in proporzione, i risultati di questo paziente sono in linea con quelli degli altri.

6.0.3 Grafici integrativi

Al fine di avere una visualizzazione più corretta dei risultati, in questa sezione sono esposte le heatmap in cui è stato sottratto il valore di z-score misurato in condizioni di riposo al valore di z-score del soggetto attivo. In questi grafici, si continuano ad osservare pattern simili tra i soggetti dello studio. Essi sono resi ancor più significativi da questa sottrazione.

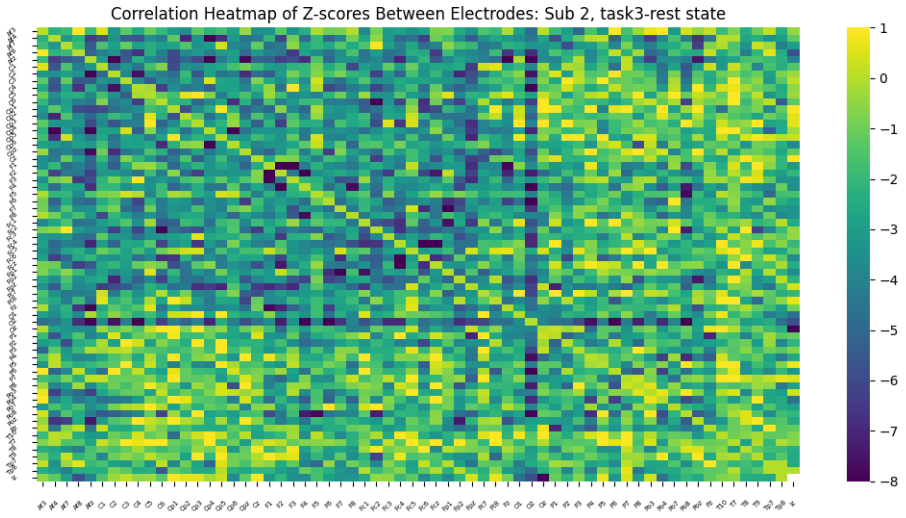


Figura 6.7: Heatmap che integra gli z-score del soggetto 2 in attività con quelli dello stesso soggetto a riposo.

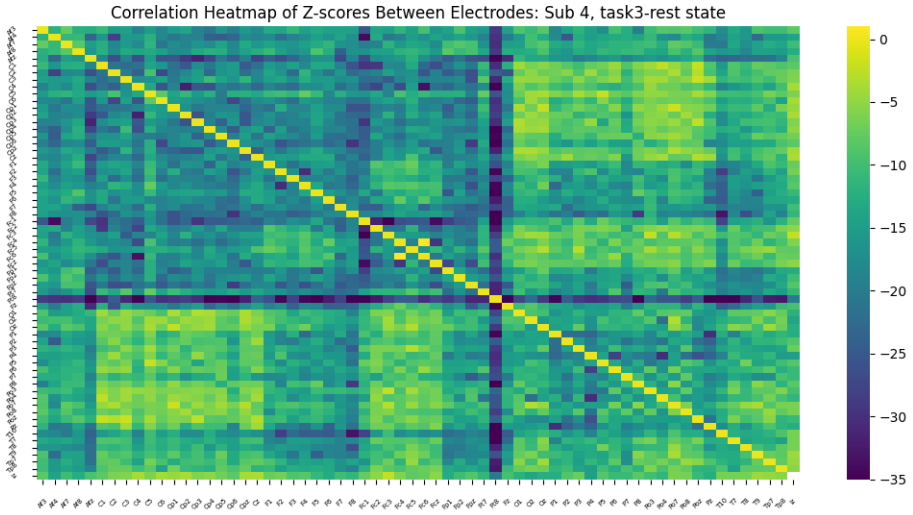


Figura 6.8: Heatmap che integra gli z-score del soggetto 4 in attività con quelli dello stesso soggetto a riposo.

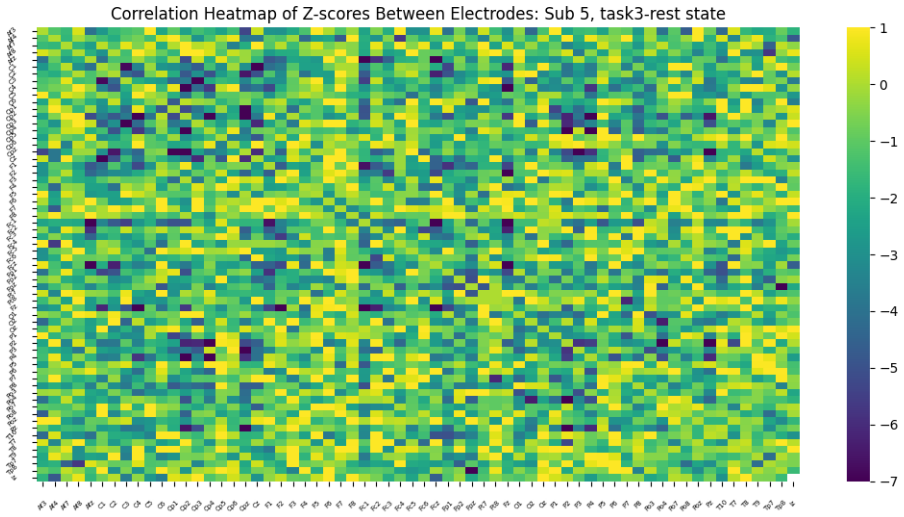
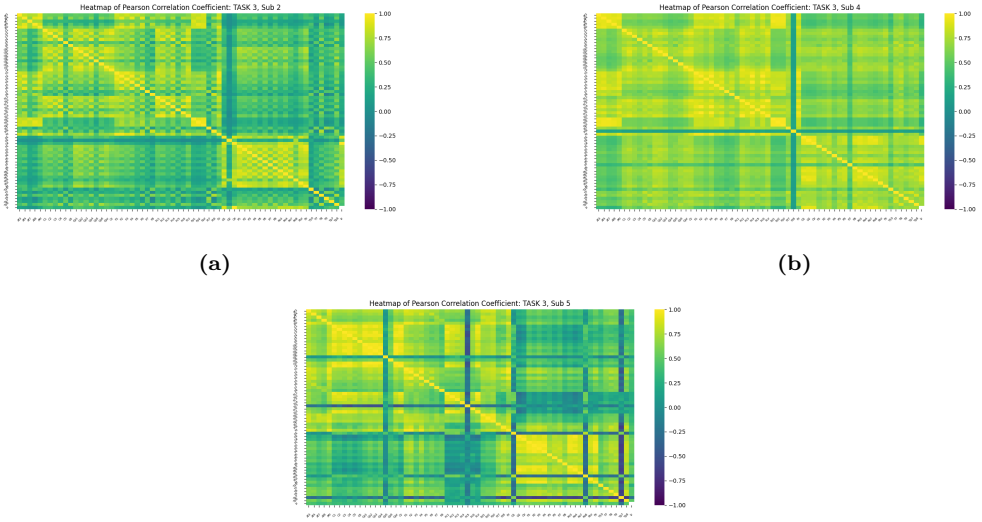


Figura 6.9: Heatmap che integra gli z-score del soggetto 5 in attività con quelli dello stesso soggetto a riposo.

6.0.4 Grafici del coefficiente di correlazione di Pearson

Nei seguenti grafici sono riportate le correlazioni misurate sui segnali relativi al task 3 attraverso il coefficiente di Pearson.



(a) **(b)**

(c)

Figura 6.10: (a) Heatmap dei coefficienti di correlazione di Pearson associati al soggetto 2. (b) Heatmap dei coefficienti di correlazione di Pearson associati al soggetto 4. (c) Heatmap dei coefficienti di correlazione di Pearson associati al soggetto 5.

6.0.5 Grafici di correlazione spaziale

In questo paragrafo sono fornite due ulteriori rappresentazioni dei risultati che coadiuvano la visualizzazione delle zone del cervello in cui è presente maggior correlazione. Per uno studio globale, i valori di z-score sono stati riscalati usando la funzione di Python *MinMaxScaler* così da poter estendere le considerazioni oltre il singolo soggetto.
Gli z-score scalati, relativi ad ogni ognuno dei tre pazienti, sono stati sommati componente per componente ed i risultati sono rappresentati sotto forma di *network map*. I nodi di questa rete corrispondono ai 64 elettrodi posti sul cranio, mentre le linee che li collegano rappresentano l'intensità di correlazione tra la coppia. Per facilitare l'osservazione dei risultati sono stati riportati soltanto gli z-score che eccedevano un determinato *threshold* di 15.5. Considerando che il massimo valore d'intensità è 17, soltanto le correlazioni più importanti sono esposte in figura.

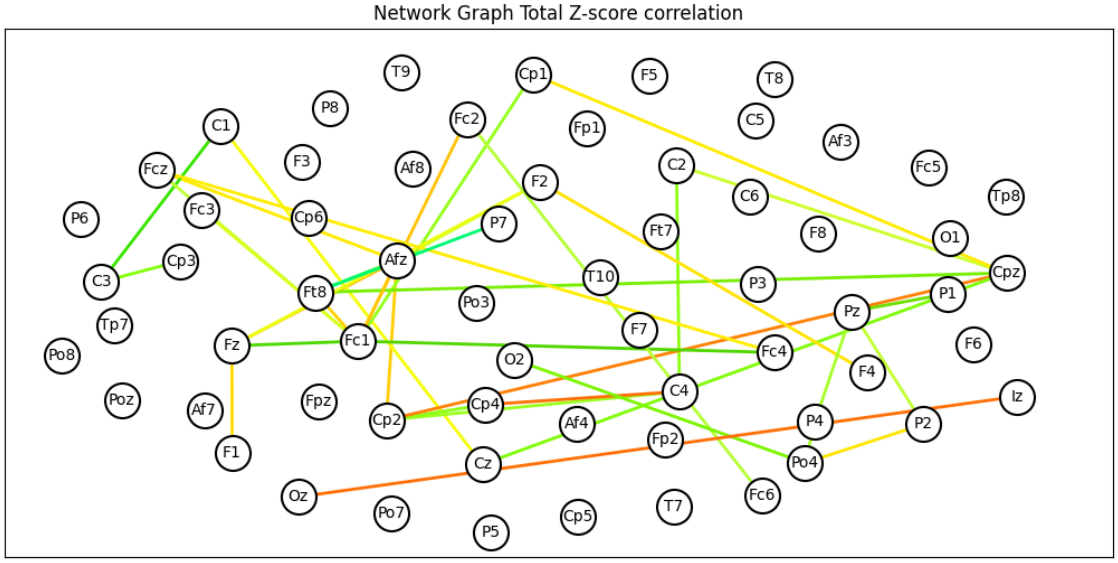


Figura 6.11: Network Map. I collegamenti in verde simboleggiano degli z-score di circa 15.5, mentre quelli in giallo ed arancione rappresentano z-score sempre maggiori. Come si può osservare, non tutte le zone del cervello correlano sufficientemente bene.

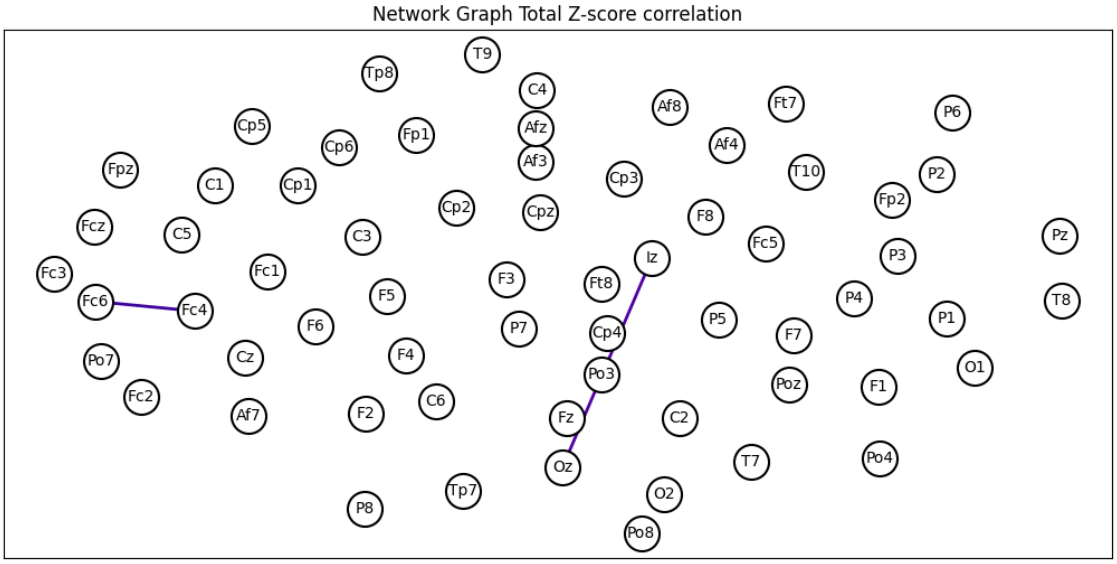


Figura 6.12: Network Map delle correlazioni rilevate duranti le sessioni di riposo. Da notare l' assenza della maggior parte dei collegamenti rispetto alla figura precedente. Questo risultato è in linea con quanto ci si potesse aspettare.

Questi risultati possono essere studiati anche attraverso un'altra rappresentazione. Di seguito sono esposte delle mappe di gradiente che riportano gli elettrodi che presentano più collegamenti in figura 4.3(b).
Le celle di colore più scuro illustrano gli elettrodi che hanno un numero di collegamenti elevato, mentre quelle di colore più chiaro (qui in giallo) simboleggiano gli elettrodi da cui partono meno linee.

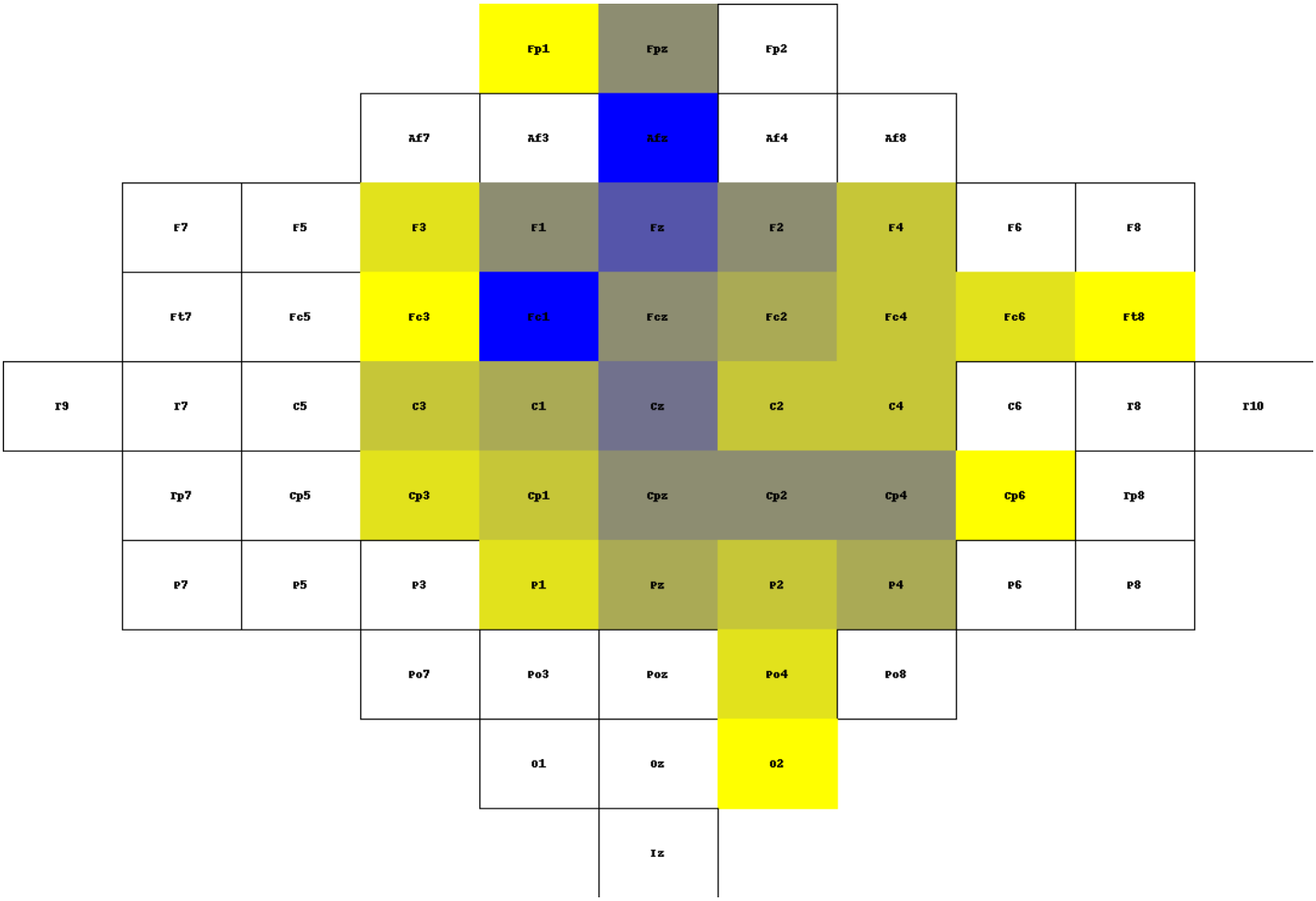


Figura 6.13: La disposizione degli elettrodi richiama l'ordine in cui sono rappresentati in figura 4.1. Gli elettrodi sono identificabili dall'etichetta al centro della cella.