UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Corso di Laurea in Fisica

Tesi di Laurea

Produzione di entropia nel cervello a livello macroscopico

Relatore

Prof./Dr. Michele Allegra

Laureando

Gabriele Casagrande

Anno Accademico 2021/2022

Indice

Introduzione

È noto che una caratteristica dei sistemi viventi sia quella di operare, a livello microscopico, in uno stato stazionario di non-equilibrio termodinamico. Ciò significa che i processi molecolari e cellulari che supportano le funzioni biologiche negli organismi viventi avvengono lontano dalla condizione di equilibrio, violando la condizione di *bilancio dettagliato*. La rottura di tale condizione ha come conseguenza la produzione di entropia e l'emergere di una precisa *freccia temporale*. Tale fatto è stato evidenziato anche da Erwin Schrödinger in [1], dove scrive che un sistema vivente *"...'Si nutre di entropia negativa', attirando un flusso di entropia negativa su di sé, per compensare l'aumento di aumento di entropia che produce vivendo e quindi per mantenersi se stesso su un livello di entropia stazionario..."*.
Appurata quindi la rottura della condizione di equilibrio dettagliato su scala microscopica si potrebbe essere portati a pensare che, in tutti i sistemi viventi, ciò si riscontri anche osservando gradi di libertà macroscopici. Non è chiaro tuttavia come questo fatto si rifletta su scala macroscopica: non è evidente che osservabili macroscopiche, di fatto costituite da una media su molti gradi di libertà microscopici, possano mostrare una rottura dell'equilibrio su larga scala. Di conseguenza, come affermato anche da Lynn et al. in [2] è importante esaminare il ruolo della rottura della condizione di equilibrio dettagliato su larga scala.
Per fare ciò è necessario andare a valutare lo stato di equilibrio del sistema complessivo, indagando la dinamica microscopica del sistema. Questo può essere fatto attraverso diversi approcci diretti, come illustrano Gnesotto et al. in [3], che consistono nell'andare a perturbare in maniera controllata il sistema per ottenere una funzione di risposta che permetta di determinare in che misura la condizione di equilibrio venga violata. Queste tecniche tuttavia risultano essere intrinsecamente invasive e di conseguenza non adatte per indagare la dinamica microscopica di sistemi delicati. Per sistemi di questo tipo è stato sviluppato un metodo non invasivo che permette di rilevare il comportamento di non-equilibrio dimostrando una rottura della condizione di bilancio dettagliato e valutando l'ampiezza di tale rottura tramite una stima della produzione di entropia. Tale metodo si basa sull'utilizzo di tecniche che permettono di osservare la dinamica macroscopica del sistema in modo non invasivo - tipicamente, tecniche di imaging.

In questo elaborato ci si concentrerà nell'approfondire questa metodologia considerando come sistema di riferimento il cervello umano, seguendo l'esempio di lavori precedenti presentati in [2] e [4].
È difatti riconosciuto che le dinamiche neurali siano fenomeni di non equilibrio, tuttavia ciò risulta meno chiaro andando ad osservare la dinamica del cervello su scala macroscopica. Come accennato precedentemente, è possibile indagare questo fenomeno partendo da registrazioni di serie temporali ottenute tramite tecniche di neuro-imaging, quali ad esempio risonanza magnetica funzionale ($fMRI$) od elettrocorticografia ($ECoG$). Ciascuna registrazione può essere poi descritta come una successione di stati visitati dal sistema nella sua evoluzione temporale, a cui è associata una certa matrice P che contiene le probabilità di transizione tra i vari stati. Da tali matrici si può procedere poi ad una stima della produzione di entropia come verrà mostrato in seguito nei capitoli 1 e 2, considerando la dinamica macroscopica del cervello come un processo Markoviano.
Inizialmente verranno poste le basi teoriche che stanno alla base di tale metodo, andando a vedere nel dettaglio come sia possibile procedere ad una stima della produzione di entropia nel caso di processi Markoviani. Successivamente ci si concentrerà nel caso specifico del cervello umano, descrivendo come la stima può essere effettuata in questo caso e quali siano le tecniche sperimentali che permettano di realizzarla. Infine si andranno ad applicare questi strumenti, valutando la stima della produzione

di entropia in due casi differenti. In primo luogo si andrà ad effettuare un test su dati artificiali per verificare la bontà della stima proposta e le condizioni per cui questa diventa consistente. A seguire verrà applicato questo metodo a dati reali, analizzando la dinamica macroscopica del cervello proveniente da due diversi gruppi: il primo composto da individui in salute, il secondo da pazienti affetti da ictus.

Capitolo 1

Bilancio dettagliato e produzione di entropia per processi di Markov

In questo capitolo verranno affrontati i concetti teorici che permettono di arrivare ad una misura della rottura della condizione di equilibrio di un sistema stocastico. Dapprima verranno date delle definizioni generali delle grandezze in esame e successivamente si indagherà il legame tra la rottura della condizione di bilancio dettagliato e la produzione di entropia.

1.1 Definizione di entropia

Come detto che un metodo che permette di valutare la condizione di equilibrio termodinamico è quello di andare a valutare la produzione di entropia del sistema. Prima di arrivare a ciò è però necessario dare un'introduzione di quello che è il concetto di entropia.
Nell'amnito della Termodinamica, dal *secondo principio della termodinamica*, deriva che per ciascun sistema fisico si possa definire una variabile di stato detta **entropia** con delle precise caratteristiche. In particolare tale variabile (spesso indicata con la lettera S) è una funzione di stato ed una grandezza estensiva. Inoltre la variazione di entropia in processi infinitesimi è data dalla somma di contributi:

$$dS = \hat{d}_e S + \hat{d}_i S. \tag{1.1}$$

In particolare $\hat{d}_e S$ indica la variazione di entropia dovuta alle interazioni del sistema con il mondo esterno. Ad esempio, nel caso di un sistema chiuso, questa sarà uguale a

$$\hat{d}_e S = \frac{\hat{d}Q}{T}, \tag{1.2}$$

dove $\hat{d}Q$ è la quantità di calore infinitesima trasferita al sistema dal mondo esterno e T è la temperatura del sistema. L'altro termine, $\hat{d}_i S$, rappresenta invece la variazione di entropia dovuta ai processi che avvengono internamente al sistema. In questo caso non si ha un'espressione precisa, ma vale

$$\hat{d}_i S = \begin{cases} > 0 & \text{per processi naturali} \\ = 0 & \text{per processi quasi-stabili e reversibili} \\ < 0 & \text{per processi innaturali} \end{cases} \tag{1.3}$$

1.2 Stati stazionari di non-equilibrio

Un sistema fisico si trova in uno *stato stazionario* quando il valore di tutti i suoi parametri di stato è costante. Questa definizione è stata inizialmente introdotta da un punto di vista empirico come la configurazione termodinamica raggiunta in maniera spontanea da ciascun sistema isolato. Di conseguenza risulta che anche lo stato di equilibrio sia uno stato stazionario. Tuttavia molto più frequente,

soprattutto nei sistemi viventi, è una condizione di *non-equilibrio* costante nel tempo.
Una delle caratteristiche degli stati stazionari di non-equilibrio è che essi non possono esistere in sistemi isolati, ma possono essere mantenuti solamente se vi è un'interazione costante tra il sistema ed il mondo esterno. La condizione di stato stazionario richiede infatti che tutte le variabili di stato, tra cui anche l'entropia, abbiano valore costante. Di conseguenza si dovrà avere $dS = 0$, ma ricordando l'equazione 1.1 questo implica che

$$\hat{d}_e S = -\hat{d}_i S. \tag{1.4}$$

I sistemi isolati sono caratterizzati però dalla condizione $\hat{d}_e S = 0$ e questo implica necessariamente che $\hat{d}_i S = 0$, cioè il sistema si trova in uno stato di equilibrio.
Possiamo quindi notare come la differenza tra uno stato stazionario di equilibrio e uno di non-equilibrio stia proprio nel valore di entropia prodotta, che sarà nulla nel primo caso e positiva nel secondo.

1.3 Produzione di entropia per catene di Markov

Il primo ad introdurre il concetto di entropia associata ad un dato ensamble di processi stocastici è stato Shannon [5] nel 1948. Considerando una certa funzione di probabilità $p_i(t)$ al tempo t, su un set discreto di stati $i \in \Omega$, Shannon definisce l'entropia come

$$S(t) = -\sum_i p_i(t) \log p_i(t) \ , \tag{1.5}$$

ponendo per convenzione $x \log x = 0$ per $x = 0$.
Consideriamo ora di trattare processi di salto di Markov. Un processo Markoviano possiede un insieme discreto di possibili configurazioni appartenenti ad uno spazio delle fasi comune ed evolve dinamicamente tramite delle transizioni (cioè salti) spontanee e non correlate tra le varie configurazioni possibili secondo una certa probabilità di transizione. In particolare quindi ciascuno stato occupato al tempo t dipende solamente da quello occupato a $t-1$. Per un processo di salto di Markov continuo, le probabilità $p_i(t)$ dipendono da i ed evolvono nel tempo secondo

$$\dot{p}_i(t) = \sum_j p_j(t) W_{ji} \ , \tag{1.6}$$

con W_{ji} i tassi delle transizioni dallo stato j allo stato i.
Imponendo la condizione di Markov, per i tassi di transizione deve valere la condizione

$$\sum_i W_{ji} = 0 \ \forall j \tag{1.7}$$

e di conseguenza possiamo considerare $W_{ii} = -\sum_{j \neq i} W_{ij}$. Sfruttando 1.7 si può riscrivere 1.6

$$\dot{p}_i(t) = \sum_j p_j(t) W_{ji} = p_i(t) W_{ii} + \sum_{j \neq i} p_j(t) W_{ji} = -\sum_{j \neq i} p_i(t) W_{ij} + \sum_{j \neq i} p_j(t) W_{ji} \ . \tag{1.8}$$

In particolare si riconosce nella formula precedente la forma di una *Master equation.*
La variazione di entropia per un processo di salto continuo può essere derivata differenziando $S(t)$ di Shannon, presentata nella formula 1.5, rispetto al tempo [1]. In particolare ciò che si ottiene è

$$\dot{S}(t) = -\sum_i \dot{p}_i(t)(1 + \log p_i(t)) = -\sum_i \dot{p}_i(t) \log p_i(t) \ . \tag{1.9}$$

A questo punto possiamo sostituire il risultato per $\dot{p}_i$ trovato in 1.8 ottenendo

$$\begin{aligned} \dot{S}(t) = \sum_i \sum_{j \neq i} p_i(t) W_{ij} \log p_i(t) - \sum_i \sum_{j \neq i} p_j(t) W_{ji} \log p_i(t) = \\ \sum_i \sum_j [p_i(t) W_{ij} - p_j(t) W_{ji}] \log p_i(t) = \sum_{ij} p_i(t) W_{ij} \log p_i(t) - \sum_{ij} p_j(t) W_{ji} \log p_i(t) \ . \end{aligned} \tag{1.10}$$

[1] i concetti ed il formalismo utilizzati nella seguente parte riprendono il lavoro presentato in [6]

L'indice ij della sommatoria indica che si sta sommando su tutte le possibili coppie ij, con $i, j \in \Omega$. Di conseguenza possiamo invertire l'ordine degli indici nella seconda sommatoria e riscrivere 1.10 come

$$\dot{S}(t) = \sum_{ij} p_i(t) W_{ij} \log \frac{p_i(t)}{p_j(t)} \ . \tag{1.11}$$

A questo punto si procede moltiplicando e dividendo l'argomento del logaritmo per il medesimo termine, ottenendo

$$\dot{S}(t) = \sum_{ij} p_i(t) W_{ij} \log \frac{p_i(t) W_{ij} W_{ji}}{p_j(t) W_{ij} W_{ji}} = \dot{S}_i(t) + \dot{S}_e(t) \ . \tag{1.12}$$

In questo modo è stata ritrovata la stessa forma già vista in 1.1.
Si passa ora ad analizzare i termini di 1.12 separatamente partendo dal secondo. Questo è dato da

$$\dot{S}_e(t) = -\sum_{ij} p_i(t) W_{ij} \log \frac{W_{ij}}{W_{ji}} \ . \tag{1.13}$$

In particolare tale termine indica, come già accennato, il flusso di entropia *esterno*. Considerando l'altro termine si vede che questo è uguale a

$$\dot{S}_i(t) = \sum_{ij} p_i(t) W_{ij} \log \frac{p_i(t) W_{ij}}{p_j(t) W_{ji}} \ . \tag{1.14}$$

Ci riferiamo a questo come alla produzione di entropia *interna* al sistema.
Nel caso di sistemi in equilibrio vale la condizione di **bilancio dettagliato**:

$$p_i(t) W_{ij} = p_j W_{ji} \ , \tag{1.15}$$

dove $p_i \propto e^{-\beta E_i}$ è la distribuzione di Gibbs.
Ciò significa che le transizioni tra ciascuna coppia di stati sono bilanciate a coppie e di conseguenza tale condizione esprime la simmetria per inversione temporale che vale all'equilibrio. Diventa allora chiaro il perchè la produzione di entropia possa essere considerata come una misura della rottura della condizione di bilancio dettagliato. Se il sistema obbedisce tale condizione (quindi le probabilità di transizione sono simmetriche a coppie) la produzione di entropia $\dot{S}_i$ svanisce. Al contrario ciascuna violazione del bilancio dettagliato ha come conseguenza $\dot{S}_i > 0$.
Nel caso di sistemi stazionari di non equilibrio questa relazione viene meno ('si rompe il bilancio dettagliato'). In questo caso può essere definita una diversa relazione, detta **condizione locale di bilancio dettagliato** [7], che può essere scritta come una funzione delle energie degli stati E_i, E_j e del lavoro $\Delta \mathcal{W}_{ij}$, in particolare

$$\frac{W_{ij}}{W_{ji}} = \frac{e^{-\beta E_j}}{e^{-\beta E_i}} e^{\Delta \mathcal{W}_{ij}} \ . \tag{1.16}$$

Considerando 1.16, quindi per un sistema che soddisfa la condizione di bilancio dettagliato locale si vede come $\dot{S}_e$, che dipende da $\frac{W_{ij}}{W_{ji}}$, corrisponde alla variazione dell'entropia del reservoir associata alla transizione dallo stato i a j (nel caso di un sistema chiuso, considerando il *primo principio della Termodinamica*, questo termine è legato al calore scambiato).

Un discorso analogo, per arrivare ad una formulazione della produzione di entropia, può essere fatto nel caso in cui si consideri un processo Markoviano discreto. Anche in questo caso continua a valere la relazione di Shannon 1.5 per l'entropia, ma ciò che cambia è il modo con cui viene descritta la sua variazione. Nel caso di un processo Markoviano discreto infatti, invece che derivare rispetto al tempo, la variazione di entropia sarà

$$\Delta S = S(t+1) - S(t) = \sum_i p_i(t+1) \log p_i(t+1) - \sum_i p_i(t) \log p_i(t) \ . \tag{1.17}$$

Sviluppando i calcoli si arriva anche in questo caso ad un risultato analogo a quello ricavato precedentemente per il caso di processi di salto continui.

Capitolo 2

Produzione di entropia in dinamiche celebrali macroscopiche

2.1 Produzione di entropia nel cervello umano

Nel capitolo precedente si è visto come la rottura della condizione di bilancio dettagliato abbia come conseguenza la produzione di una quantità di entropia non nulla da parte del sistema stocastico considerato. Si considera ora come sistema stocastico il cervello, nell'ipotesi che le dinamiche neurali macroscopiche possano essere descritte come processi Markoviani.
In un sistema di questo tipo si può stimare la produzione di entropia andando ad osservare la successione degli stati occupati dalla dinamica macroscopica nel corso della sua evoluzione temporale. Per osservare tali successioni si possono andare ad osservare registrazioni provenienti da tecniche di neuro-imaging, le quali possono essere descritte tramite una sequenza di stati occupati nel tempo.
Si consideri in particolare la matrice P che contiene le probabilità di transizione congiunte tra gli stati del sistema. In particolare l'elemento P_{ij} di tale matrice indica la probabilità che il sistema passi dallo stato $x_{t-1} = i$ al tempo $t-1$ a quello $x_t = j$ al tempo t. Di conseguenza, se la dinamica è Markoviana, si può valutare la produzione di entropia tramite

$$\dot{S} = \sum_{ij} P_{ij} \ln \frac{P_{ij}}{P_{ji}}, \tag{2.1}$$

come visto anche in 1.14.
Il calcolo dell'entropia richiede quindi la stima delle probabilità di transizione tra i vari stati. Nei sistemi complessi tuttavia il numero di stati possibili è molto elevato e ciò rende la stima diretta di tali probabilità, e di conseguenza della produzione di entropia, di fatto impossibile. Per superare questa difficoltà sono quindi necessarie delle tecniche che permettano di ridurre la dimensionalità del set di dati a disposizione, andando quindi ad individuare un numero ridotto di stati possibili, ciascuno dei quali formato da gruppi di dati del set iniziale che condividono tra loro caratteristiche simili.
Tali tecniche vengono dette *tecniche di riduzione dimensionale*. Nel capitolo successivo verranno illustrate in particolare due di queste tecniche, utilizzate per individuare gli stati occupati dalla dinamica celebrale macroscopica, rifacendosi alle metodologie applicate da Lynn et al. in [2] e Sanz Perl et al. in [4].

2.2 Tecniche di riduzione dimensionale

L'utilizzo di dataset di grandi dimensioni è una pratica sempre più diffusa in varie discipline. Tuttavia, a causa della grande dimensionalità, questi risultano essere di difficile rappresentazione. È stato perciò necessario sviluppare tecniche che ne permettano di ridurre le dimensioni. Tali tecniche sono applicate in contesti in cui si pensa che le variabili misurate dipendano da un numero ristretto di variabili esplicative, ed il loro scopo è quello di identificare ed estrarre queste variabili. Nel seguito sono

presentate due tecniche che permettono di ridurre la dimensionalità comunemente usate : l' *analisi delle componenti principali* e gli *algoritmi di clustering*. Tali metodologie sono state successivamente usate per ridurre il numero di stati occupati possibili nella dinamica macroscopica del cervello al fine di descrivere ciascuna registrazione ottenuta tramite imaging come una successione di stati occupati.

2.2.1 Analisi delle Componenti Principali (PCA)

L'idea base dell' **analisi delle componenti principali (PCA)**, come spiegato anche in [8], è quella di trovare nuove variabili, che siano combinazioni lineari di quelle già esistenti, le quali permettano di massimizzare la varianza spiegata e non siano correlate tra loro.
Si consideri di avere una matrice delle osservazioni $\mathbf{X}$ di dimensione $N \times M$, dove n è il numero di osservazioni effettuate ed m quello di variabili misurate per ogni osservazione.
Generalmente, prima di procedere all'applicazione della PCA, la matrice $\mathbf{X}$ viene pre-processata. Si può ad esempio fare in modo che ogni colonna di $\mathbf{X}$ abbia media nulla. Se oltre a questo si divide ciascun elemento x_{nm} per la radice del numero delle osservazioni ($\sqrt{N}$ o $\sqrt{N-1}$), la matrice $\mathbf{X^T X}$ sarà una matrice di covarianza, per cui si considererà una *covariance* PCA.
Se le M variabili misurate in ciascuna osservazione sono espresse in unità di misura differenti si procede invece ad una normalizzazione, dividendo ciascuna variabile per la rispettiva norma. In questo caso la matrice $\mathbf{X^T X}$ è una matrice di correlazione, per cui ci si riferirà ad una *correlation* PCA.
La scomposizione in componenti principali è ottenuta a partire dalla *decomposizione a valori singolari* (**SVD** [9]) della matrice $\mathbf{X}$. Questa può essere scomposta come

$$\mathbf{X} = \mathbf{U\Sigma V^T}, \tag{2.2}$$

dove $\mathbf{U}$ è la matrice $N \times N$ degli *autovettori di sinistra*, $\mathbf{V}$ la matrice $M \times M$ degli *autovettori di destra* (entrambe ortogonali), e $\mathbf{\Sigma}$ la matrice $N \times M$ dei *valori singolari* con elementi lungo la diagonale $(\sigma_1 > \sigma_2 > \cdots > \sigma_{minN,M})$. In particolare, considerando $\mathbf{\Lambda}$ la matrice degli autovettori non nulli di $\mathbf{X^T X}$ e $\mathbf{X X^T}$, questa risulta essere uguale a $\mathbf{\Sigma^2}$.
I nuovi valori delle N osservazioni ottenute in seguito all'applicazione della PCA sono detti *factor scores*, la matrice $N \times M$ di questi, detta $\mathbf{F}$, si ottiene partendo dalla SVD nel seguente modo:

$$\mathbf{F} = \mathbf{U\Sigma}. \tag{2.3}$$

Questi *factor scores* possono essere visti come la proiezione dei valori delle osservazioni effettuate lungo le componenti principali. La matrice $\mathbf{V}$ restituisce i coefficienti delle combinazioni lineari usate per calcolare i factor scores, di conseguenza può essere interpretata come una *matrice di proiezione* che trasforma la matrice dei dati originali nei nuovi valori e viene per questo anche chiamata *matrice dei carichi*. In particolare si può vedere che

$$\mathbf{F} = \mathbf{U\Sigma} = \mathbf{U\Sigma V^T V} = \mathbf{XV}. \tag{2.4}$$

L'importanza di ciascuna componente è data dalla porzione dell'*inerzia totale* espressa dal fattore ad essa associata. Possiamo definire l'inerzia di una colonna come la somma degli elenti di questa elevati al quadrato, cioè

$$\gamma_m^2 = \sum_n^N x_{n,m}^2. \tag{2.5}$$

La somma delle inerzie relative a tutte le colonne è detta *inerzia totale*. Questa è anche uguale alla somma dei quadrati dei valori singolari della matrice $\mathbf{X}$. Per quanto riguarda le varie componenti principali si richiede che la prima sia quella che 'spiega' la maggior parte dell'inerzia totale (i.e. che abbia il maggior valore di inerzia). La seconda sarà quella avente il valore maggiore di inerzia tra quelle rimaste, con il vincolo aggiuntivo di essere ortogonale alla prima e così via per tutte le componenti di ordine successivo.
Al fine di diminuire le dimensioni dei dati osservati quello che si chiede è di estrarre solamente le informazioni principali, cioè scegliere un numero congruo di componenti principali da considerare.
Una procedura possibile è quella di considerare un grafico in cui venga riportata la percentuale di

inerzia spiegata da ciascuna componente. Questa può essere calcolata ad esempio per la componente i-esima con la formula

$$I_i = \frac{\gamma_i^2}{I_{tot}}. \tag{2.6}$$

La pendenza di questo grafico andrà man mano ad affievolirsi fino ad arrivare ad un valore circa costante e di conseguenza si prendono solaente le componenti i cui valori di inerzia sono precedenti a tale valore. Un'altro metodo è quello di considerare solamente le componenti il cui valore di inerzia spiegata sia superiore alla media. Quindi si considereranno solamente i valori per cui

$$I_i > \frac{1}{L}\sum_{i}^{L} I_i = \frac{1}{L} I_{tot}, \tag{2.7}$$

con $L = minN, M$.

2.2.2 K-means clustering gerarchico

Per ridurre il numero di possibili stati ossevati si può procedere utilizzando **algoritmi di clustering**. Questo metodo permette di raggruppare nel medesimo cluster oggetti aventi caratteristiche simili tra loro e differenti rispetto a quelli di altri cluster. Il clustering costituisce uno strumento molto utile per andare a esplorare grandi dataset, ed ha un grande numero di applicazioni in numerose discipline.

Tra le varie tecniche di clustering possibili una delle più basilari è quella del **k-means**.
Il k-means è un particolare tipo di *algoritmo di partizionamento* che permette di ricavare **k** cluster, con **k** fissato a priori. Ciascun dato è considerato come un'oggetto avente una certa posizione in uno spazio multidimensionale e si va a trovare una partizione per la quale gli oggetti assegnati al medesimo cluster sono il più vicini possibile. Tramite questa tecnica è possibile di conseguenza andare a ridurre il numero degli stati possibili del sistema.
Si consideri un set iniziale $D = \{x_1, x_2, \ldots, x_n\}$, con $x_i \in R^n$. Il k-means seleziona k semi iniziali e divide i dati tra questi andando a creare un set di k cluster $(C_1, C_2, \ldots, C_k)$, andando poi a calcolare il centroide di ciascun cluster. Il valore del centroide per la *i-esima* variabile del cluster C_j è dato da

$$\bar{x_i}^{(j)} = \frac{1}{n_j}\sum_{a \in C_j} x_{ai}, \tag{2.8}$$

quindi il vettore che rappresenta il centroide per il cluster C_j sarà $\bar{x}^{(j)} = (\bar{x_1}^{(j)}, \bar{x}_2^{(j)}, \ldots, \bar{x}_n^{(j)})$.
Dopo di che si riassegna ciascun punto al cluster 'più simile'. La scelta del concetto di simiglianza può essere fatta in modi differente. In questo caso, in conformità con [ ], è stata utilizzata la **somiglianza del coseno**.
Questa consiste, dati due vettore **x** e **y**, nel calcolare

$$sim(x, y) = \frac{x \cdot y}{\|x\|\|y\|} \tag{2.9}$$

dove $\|x\|$ è la *norma euclidea* del vettore $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n})$, definita come $\|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$.
Viene dunque valutato il coseno dell'angolo tra i due vettori **x** e **y**. Tanto più il valore è vicino ad 1, tanto più i due vettori sono vicini (i.e. simili) l'un l'altro.
Per ciascun punto del set D si valuta la somiglianza (nel modo appena definito) con i centroidi dei k cluster e si riassegna al cluster i tale per cui $\mathbf{sim(x_i, y) = sim(x_i, y)_{min}}$.
Dopo di che si calcolano nuovamente i centroidi dei k cluster e si itera questo procedimento finchè non vi è più alcuna variazione.
Tuttavia la partizione ottenuta tramite *k-means* risulta essere molto sensibile alla scelta dei k semi iniziali. In particolare è possibile che il *k-means* raggiunga una situazione in cui non si registrino pi cambiamenti tra i punti dei vari cluster, ma che esista una soluzione migliore per la partizione.
Per ovviare a questo problema viene utilizzata un'implementazione **gerarchica divisiva** [10]. Questa implementazione consiste nel considerare un numero di cluster crescenti facendo in modo di minimizzare la *dispersione* dei cluster.

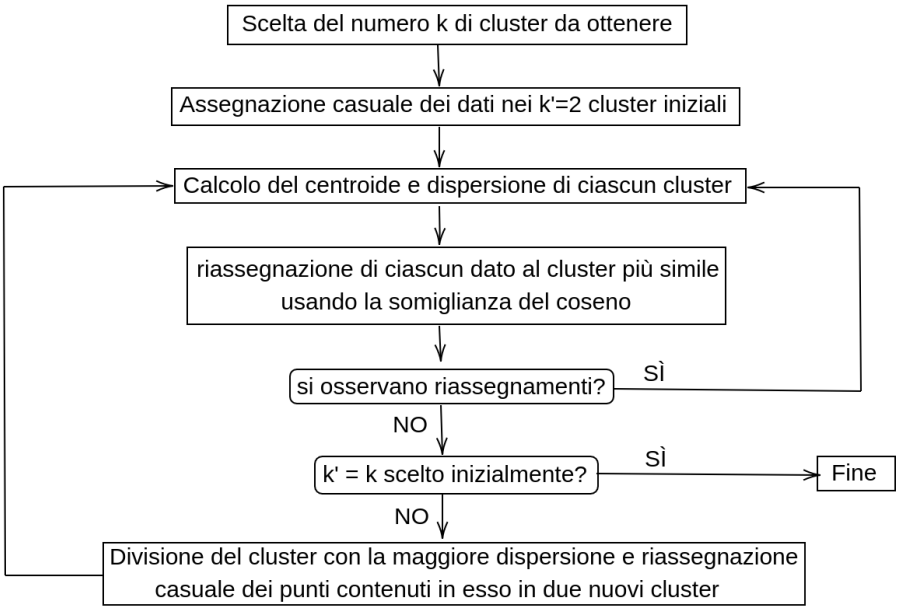


Figura 2.1: Schema del metodo k-means gerarchico

Anche in questo caso il numero di cluster k finali viene fissato a priori, tuttavia la partizione avviene in maniera differente. Si inizia considerando un numero di cluster $k' = 2$ e si applica il k-means per ottenere questi due. Successivamente si va a valutare quella che è la *dispersione* di ciascuno dei k' cluster così trovati. Questa viene valutata calcolando per ciascun C_j, con $j \in [1, k']$,la sommatoria delle similitudini tra il centroide del cluster *j-esimo* ed ognuno dei punti appartenenti ad esso.
Per un generico cluster C_i si avrà dunque

$$spread^{(i)} = \sum_{\mathbf{x} \in \mathbf{C_i}} sim(\bar{\mathbf{x}}^{(\mathbf{i})}, \mathbf{x}). \tag{2.10}$$

Calcolato questo valore per ciascuno dei k' cluster si sceglie quello avente la dispersione maggiore e si dividono i punti al suo interno in maniera casuale in 2 nuovi cluster. Successivamente si calcolano i centroidi dei nuovi cluster così trovati e si applica il *k-means*, considerando $k'' = 3$ cluster.
Questo processo viene poi iterato fino ad arrivare al numero di cluster k scelto.

Capitolo 3

Stima della produzione di entropia su dati artificiali

In questo capitolo si applicheranno le nozioni teoriche precedentemente proposte per effettuare una stima della produzione di entropia di un set di dati generati artificialmente. Questo procedimento è stato effettuato per valutare la bontà della stima della produzione di entropia, partendo da un campione creato *ad hoc* del quale si conosce il valore dell'aspettativa teorica $\dot{S}_{th}$.

3.1 Simulazioni

I dati artificiali sono stati generati partendo dalla simulazione di un processo Markoviano a tempo discreto che salta tra K stati differenti, con $K = 4$. Tale processo è associato ad una determinata matrice delle transizioni P che dipende da un parametro α, con $\alpha \in \{0.3, 0.5, 0.7, 0.9\}$. Nel primo caso il processo è stato simulato partendo da una matrice simmetrica, della forma

$$P = \frac{1}{4}\begin{pmatrix} \alpha & \frac{1-\alpha}{3} & \frac{1-\alpha}{3} & \frac{1-\alpha}{3} \\ \frac{1-\alpha}{3} & \alpha & \frac{1-\alpha}{3} & \frac{1-\alpha}{3} \\ \frac{1-\alpha}{3} & \frac{1-\alpha}{3} & \alpha & \frac{1-\alpha}{3} \\ \frac{1-\alpha}{3} & \frac{1-\alpha}{3} & \frac{1-\alpha}{3} & \alpha \end{pmatrix}. \tag{3.1}$$

Un sistema la cui dinamica macroscopica sia descritta da una matrice di questo tipo rispetta la condizione di bilancio dettagliato, infatti le probabilità di transizione tra i diversi stati rispettano la condizione 1.15 essendo simmetriche a coppie. Di conseguenza, ricordando la formula 2.1, la produzione di entropia sarà nulla. Altre simulazioni sono state realizzate considerando invece una matrice delle transizioni asimmetrica, in particolare

$$P = \frac{1}{4}\begin{pmatrix} \alpha & \frac{1-\alpha}{2} & \frac{1-\alpha}{4} & \frac{1-\alpha}{4} \\ \frac{1-\alpha}{4} & \alpha & \frac{1-\alpha}{2} & \frac{1-\alpha}{4} \\ \frac{1-\alpha}{4} & \frac{1-\alpha}{4} & \alpha & \frac{1-\alpha}{2} \\ \frac{1-\alpha}{2} & \frac{1-\alpha}{4} & \frac{1-\alpha}{4} & \alpha \end{pmatrix} \tag{3.2}$$

Un sistema dove le transizioni tra stati macroscopici siano descritte da una matrice di questo genere rompe la condizione di bilancio dettagliato (1.15), producendo una quantità di entropia non nulla, il cui valore varierà monotonicamente con α.

Ad ognuno dei K stati che compongono processo simulato è associata una distribuzione gaussiana $\mathcal{N}(\boldsymbol{\mu}_k, \sigma^2)$, $\quad k = 1, \ldots, K$, centrata in un punto differente $\boldsymbol{\mu}_k \in \mathbb{R}^2$. Pertanto, partendo dalla serie temporale $y_t,\ t = 1, \ldots, N$, dove ciascun $y_t \in \{1, \ldots, K\}$ è un indice discreto da 1 a k, viene generato in maniera casuale un punto partendo dalla rispettiva gaussiana, ottenendo una serie temporale di valori bi-dimensionali $X_t,\ t = 1, \ldots, N$, dove $X_t \in \mathbb{R}^2$. Per ogni valore di α sono state realizzate varie simulazioni variando il numero N dei punti simulati, con $N \in \{10^2, 10^3, 10^4, 10^5\}$.

I dati artificiali ottenuti tramite le simulazioni dei processi Markoviani sono stati poi analizzati utilizzando l'algoritmo di $k-means$ *gerarchico* presentato in 2.2.2, considerando di voler ottenere un numero finale di cluster $k = 4$, pari al numero di stati tra i quali salta il processo Markoviano. L'algoritmo di clustering assegna ciascun punto x_t a uno dei K clusters, per cui si ottiene una serie temporale $\hat{y}_t$ dove ciascun $\hat{y}_t \in \{1, \dots, K\}$ è un indice discreto da 1 a k. Idealmente (nel limite in cui l'algoritmo di clustering fosse perfettamente efficiente) la serie $\hat{y}_t$ dovrebbe coincidere con quella originaria y_t. A partire dalla serie $\hat{y}_t$ è stato possibile ricostruire una matrice delle transizioni sperimentale

$$\hat{P}_{ij} = \frac{1}{N} \sum_{t=1}^{N-1} \chi(\hat{y}_{t+1} = j, \hat{y}_t = i) \tag{3.3}$$

dove χ è la funzione indicatrice. Viene fatto notare che, data la stima (3.3), in alcuni casi si può ottenere $\hat{P}_{ij} = 0$ anche se $P_{ij} > 0$. Questa situazione corrisponde al caso in cui, sebbene $P_{ij} > 0$, nelle serie temporale ricostruita dai dati non si osserva alcuna transizione tra gli stati i e j. In tali circostanze non è possibile fissare $k = 4$, e occorre dimimuire il numero di cluster fino ad ottenere un k per cui risulti che tutte gli elementi Π_{ij} della matrice delle transizioni calcolata a partire dai dati artificiali siano non nulli. A partire dalla matrice delle transizioni stimata $\hat{P}$ è stato possibile calcolare una stima della produzione di entropia $\dot{S}_{est}$. Tale valore è stato poi confrontato con l'aspettativa teorica per valutare la bontà della stima ed eventuali contributi di errore che possono sorgere dall'analisi dei dati. In particolare ci si aspetta che al crescere di N il processo Markoviano raggiunga la convergenza con la relativa matrice delle transizioni e di conseguenza che la stima dell'entropia sia migliore per valori alti di N.

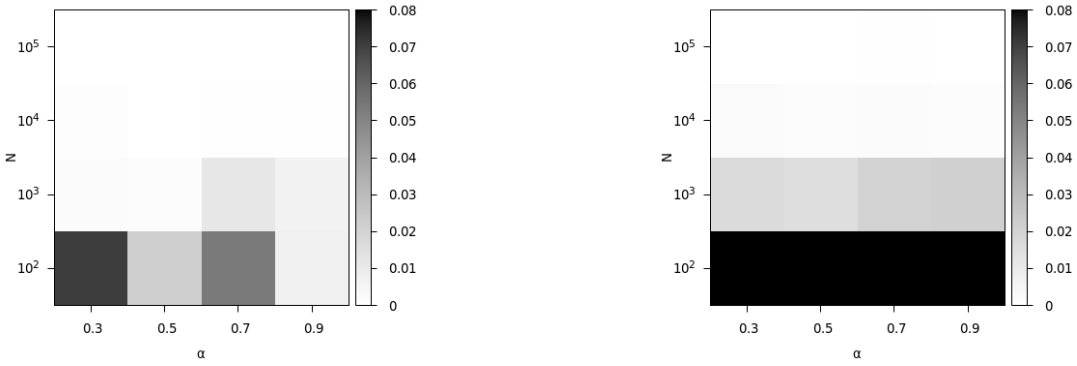
3.2 Considerazioni sulla lunghezza finita delle serie temporali

La lunghezza finita delle serie temporali limita l'accuratezza con la quale è possibile stimare la produzione di entropia. Per calcolare gli errori associati a questa stima si procede seguendo l'esempio di [2] ed applicando la tecnica delle *traiettorie bootstrap*. Il bootstrap è una tecnica statistica di ricampionamento che permette di approssimare la distribuzione campionaria di una statistica nel caso in cui non si conosce la distribuzione di tale statistica. Si consideri di osservare un set composto da n dati, ad esempio $x = (x_1, \dots, x_n)$. Da x procediamo a ricampionare un certo numero m di campioni formati dallo stesso numero di elementi n. Per farlo si estraggono in maniera casuale dati appartenenti al campione originale, andando eventualmente anche ad estrarre lo stesso dato più di una volta ottenendo così dei set di dati $x_i^* = (x_{i,1}^*, \dots, x_{i,n}^*)$. Data una funzione di interesse $f(x)$ calcolata a partire campione originale, è possibile ricalcolarla su ciascun campione ottenuto dal bootstrap, ottenendo un set di m valori $f(x_i^*)$. La dispersione delle $f(x_i^*)$ fornisce una stima dell'errore su $f(x)$.
In particolare, nel caso della produzione di entropia si consideri la lista delle transizioni osservate

$$I = \begin{pmatrix} i_1 & i_2 \\ i_2 & i_3 \\ \vdots & \vdots \\ i_{L-1} & i_L \end{pmatrix}, \tag{3.4}$$

dove L è la lunghezza della serie temporale ed i_l lo stato $l-esimo$. Costruendo delle traiettorie bootstrap di lunghezza L, andando a ricampionare le varie le righe della matrice I è possibile stimare l'errore associato alla misura finita delle serie temporali. In particolare è stato scelto di associare alla stima dell'entropia teorica un errore pari a 2 volte la deviazione standard calcolata a partire dalle traiettorie bootstrap.

La lunghezza finita delle serie temporali induce non soltanto un errore casuale a media nulla dovuto al campionamento - errore che si può stimare osservando la dispersione delle $f(x_i^*)$. Essa introduce anche un errore sistematico positivo ("fondo rumoroso"), dovuto alla presenza di lievi asimmetrie nella stima delle $\hat{P}_{ij}$ anche quanto le P_{ij} sono simmetriche. Tale fondo è presente anche se l'ordine temporale delle serie viene distrutto (come argomentato in [2]), e pertanto si può stimarne il valore facendo venire



(a) S_{est} (b) $\delta S_{bootstrap}$

Figura 3.1: Variazione dei valori di S_{est} e di $\delta S_{bootstrap}$ per $S_{th} = 0$, al variare di α ed N

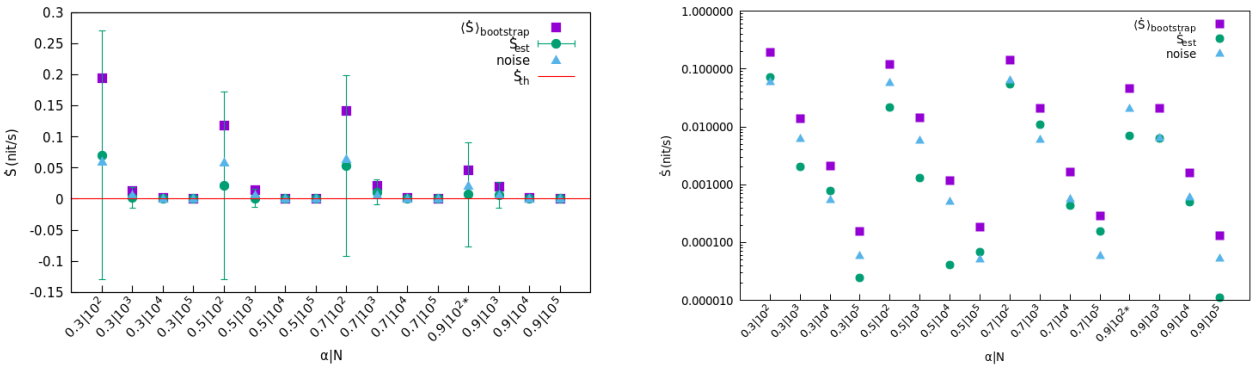
meno l'ordine temporale delle transizioni tra gli stati e calcolando la la produzione di entropia.
A tale scopo si costruiscono traiettorie bootstrap, andando questa volta a campionare singolarmente i vari stati i_l in modo che le transizioni non vengano preservate. Fatto ciò si stima il rumore di fondo andando a considerare una media della produzione di entropia sulle traiettorie bootstrap.
Si fa notare che sia per la stima dell'errore che per quella del fondo sono stati calcolati 100 set tramite ricampionamento con il metodo del bootstrap.

3.3 Analisi delle simulazioni

3.3.1 Caso della matrice delle transizioni simmetrica

Nel primo caso la matrice delle transizioni di riferimento risulta essere simmetrica. La dinamica macroscopica del sistema simulato soddisfa quindi la condizione di *bilancio dettagliato* e di conseguenza la produzione di entropia attesa è nulla, cioè $\dot{S}_{th} = 0\ nit/s$.
Per vedere quanto la stima calcolata si discosta dall'aspettativa teorica e come questa dipenda dai valori di α ed N è stata realizzata una mappa di calore, rappresentata in figura 3.1a. In ascissa ed in ordinata si trovano rispettivamente i valori di α ed N mentre le varie tonalità indicano il valore della produzione di entropia ottenuto.

Facendo riferimento a 3.1a può vedere che, come ci si aspetta, l'andamento della stima varia in maniera significativa all'aumentare del numero di dati artificiali simulati. In particolare al crescere di N il valore calcolato a partire dai dati si avvicina all'aspettativa teorica, cioè un valore nullo per la produzione di entropia. Inoltre si nota che osservando la mappa di calore 3.2b, nella quale la quantità prodotta è sostituita dall'errore stimato tramite traiettorie bootstrap (3.2), che anche questo diminuisce al crescere della taglia del campione originale. In particolare anche questo andamento è quello che ci si aspetterebbe poichè la deviazione standard è inversamente proporzionale a $\sqrt{N}$.
In 3.2 è stato spiegato come sono stati utilizzati dei ricampionamenti delle transizioni tra stati osservate, tramite il metodo bootstrap, per valutare l'errore associato all'utilizzo di un dataset di dimensione finita. Questa analisi ha permesso inoltre di calcolare la media del campione comprensivo delle 100 traiettorie bootstrap più il set di dati originali. Andando a confrontare il valore della media $\langle \dot{S}_{bootstrap} \rangle$ ottenuto in questo modo con la produzione di entropia stimata a partire dai dati originali, rappresentati entrambi per le possibili combinazioni di α ed N in 3.2a, si nota che i valori siano compatibili. Anche in questo caso si osserva come all'aumentare della dimensione dei campioni di dati in esame le stime tendano all'aspettativa teorica $\dot{S} = 0$ e di come in effetti l'errore sulla singola misura scali di conseguenza. In 3.2b si sceglie di considerare una raffigurazione in scala logaritmica degli stessi dati, scegliendo, ai fini della chiarezza, di non rappresentare l'errore associato alla stima della produzione di entropia. Ciò permette di indagare come i valori siano distribuiti nel caso di N elevato. In questo viene evidenziato che il valore della produzione di entropia mediata sulle traiettorie bootstrap risulti essere sempre maggiore rispetto alla stima calcolata a partire dal set di dati originale. La presenza di



(a) (b)

Figura 3.2: Confronto di $\dot{S}_{est}$ con $\langle \dot{S}_{bootstrap} \rangle$ e *noise*, per matrice delle transizioni simmetrica

questo *bias* sistematico può essere dovuta al fatto di utilizzare per la stima dell'errore casuale associato al calcolo della produzione di entropia il metodo del bootstrap. Le stime ottenute tramite tale metodo possono infatti essere soggette ad un errore sistematico, come spiegato anche in [11].

Nei grafici in figura 3.2 è rappresentato inoltre il valore del fondo calcolato sempre tramite traiettorie ottenute con il metodo del bootstrap. In particolare concentrandosi su 3.2a si può notare come i valori del fondo tendano a 0 all'aumentare della dimensione del campione in esame. In 3.2b si va invece ad analizzare in maniera più approfondita come siano disposti i valori del rumore rispetto la produzione di entropia calcolata, osservando come questi siano sempre confrontabili. In generale si nota che i valori si $\dot{S}_{est}$ risultino essere sempre compatibili con l'aspettativa teorica, cioè un valore nullo per la produzione di entropia, indipendentemente dal valore di N. Infatti per tutti i valori di $\dot{S}_{est}$ il valore atteso $\dot{S}_{th} = 0$ è compreso entro l'errore casuale, pari a due deviazioni standard, associato alla quantità calcolata. Tuttavia si nota anche che la stima migliore sia quella ottenuta a partire dai dataset di dimensione maggiore ($N > 10^4$).

3.3.2 Caso della matrice delle transizioni asimmetrica

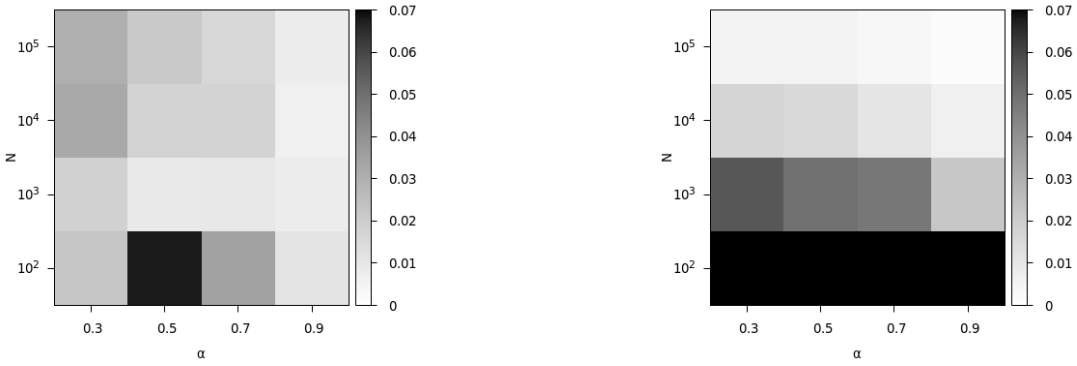
Considerando il set di dati in cui il processo Markoviano simulato aveva come riferimento la matrice delle transizioni asimmetrica, viene meno la condizione di *bilancio dettagliato* e di conseguenza si la produzione di entropia non sarà nulla. Questa inoltre varierà al variare del parametro α. Andando ad effettuare il calcolo si ottengono i valori riportati in tabella 3.1.

α	$\dot{S}_{th}$ (nit/s)
0.3	0.1213
0.5	0.0866
0.7	0.0520
0.9	0.0173

Tabella 3.1: Valori di $\dot{S}_{th}$

In analogia con l'analisi precedente anche in questo caso sono state realizzate delle mappe di calore, presentate in 3.3. In particolare in 3.3a è rappresentato il valore assoluto della differenza tra la stima della produzione di entropia effettuata e la sua aspettativa teorica. Osservando 3.3a si vede come le differenze assolute tra il valore della produzione di entropia calcolato e l'aspettativa teorica non seguano un preciso andamento. Invece gli errori stimati tramite i campioni bootstrap, raffigurati in 3.3b, diminuiscono all'aumentare della taglia del campione come ci si aspetta ($\propto \sqrt{N}$).

Nella figura 3.4 sono stati invece confrontati i valori di $\dot{S}_{est}$ con il valore medio e la misura del fondo calcolati tramite i campioni ottenuti dal bootstrap. Si può notare innanzitutto, osservando 3.4a, come $\dot{S}_{est}$ tenda al valore atteso per i vari α al crescere di N, come osservato anche nel caso della matrice



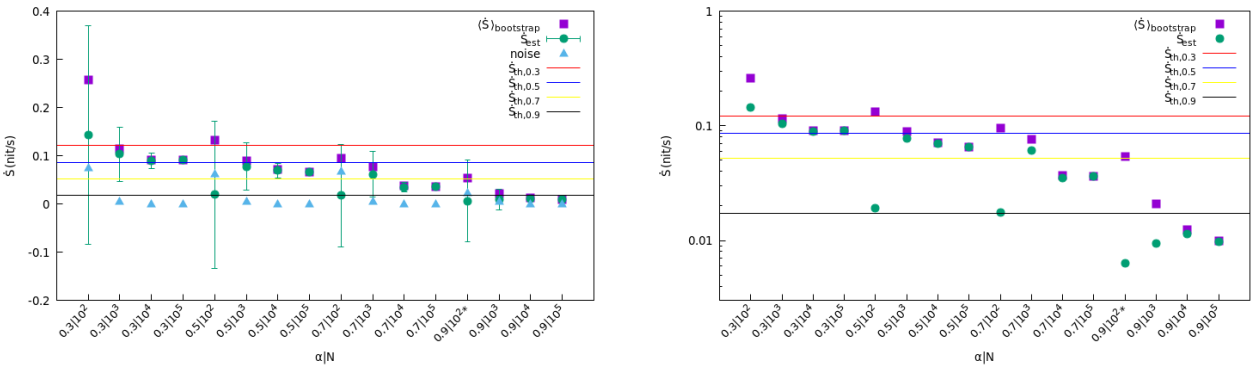
(a) $|\dot{S}_{est} - \dot{S}_{th}|$ (b) $\delta\dot{S}_{bootstrap}$

Figura 3.3: Variazione dei valori di $\dot{S}_{est}$ e di $\delta\dot{S}_{bootstrap}$ per $\dot{S}_{th} \neq 0$, al variare di α ed N

delle transizioni simmetrica.
Sempre dalla stessa figura si vede che la produzione di entropia calcolata risulta essere superiore al valore del fondo, ad eccezione del set con $\alpha = 0.9$. Per tale valore di α infatti l'aspettativa teorica della produzione di entropia $\dot{S}_{th}$ è minore rispetto agli altri casi, avvicinando di conseguenza al valore del rumore.
Si nota inoltre anche in questo caso come il valore della media calcolata sulle traiettorie bootstrap sia maggiore rispetto alla produzione di entropia stimata, soprattutto per i set composti da un numero minore di dati. Questo fatto, come detto anche nella sezione precedente, può essere dovuto all'utilizzo del metodo bootstrap per il ricampionamento.
Facendo sempre riferimento a 3.4a inoltre si può osservare che per i campioni di dati meno ampi ($N = 10^2$) non si possa affermare con sicurezza che il valore della produzione di entropia calcolato $\dot{S}_{est}$ sia effettivamente non nullo come ci si aspetterebbe. Per tali valori infatti il valore $\dot{S} = 0$ cade all'interno della barra di errore di due deviazioni standard. Un discorso analogo può essere fatto anche per i campioni di dimensione $N = 10^3$ per valori di α elevati.
In definitiva è possibile affermare che per ottenere valori della produzione di entropia calcolata $\dot{S}_{est}$ che siano diversi dal valore nullo con un sufficiente intervallo di confidenza (pari ad almeno tre deviazioni standard), per evitare quindi il caso di *falsi positivi*, è necessario tenere in considerazione campioni con un numero di componenti N sufficientemente alto. In particolare è sufficiente $N \geq 10^3$ per bassi valori di α, mentre al crescere di quest'ultima è necessario $N \geq 10^4$.
Per approfondire l'andamento dei dati con valori di N maggiori rispetto all'aspettativa teorica si è scelto di rappresentare i valori in esame in scala logaritmica (3.4b). In questo caso è stato scelto di non includere nella rappresentazione i valori del fondo in quanto molto minori rispetto agli altri (ad eccezione di $\alpha = 0.9$ come già detto). Dall'analisi di 3.4b si osserva che per valori di N elevato $(10^4, 10^5)$ la produzione di entropia calcolata S_{est} risulta essere inferiore rispetto ad $\dot{S}_{th}$. Questo fatto può fra sorgere il dubbio sulla presenza di un errore sistematico non considerato che sarà analizzato in 3.3.3.

3.3.3 Considerazioni su una possibile fonte di errore sistematico

Una possibile fonte di errore può essere riscontrata nel metodo stesso con cui i set di dati vengono creati, andando a simulare un processo Markoviano. In particolare in 3.1 si era descritto come ad ognuno dei K stati tra cui il sistema saltava fosse associata una distribuzione gaussiana $\mathcal{N}(\boldsymbol{\mu}_k, \sigma^2)$, $k = 1, \ldots, K$ centrata in un punto differente $\boldsymbol{\mu}_k \in \mathbb{R}^2$. Poichè i punti dello spazio bi-dimensionale x_t, $t = 1 \ldots N$, vengono generati casualmente partendo da tali gaussiane è possibile che questi vengano a trovarsi sulla coda della distribuzione normale.
Supponiamo che $x^k_{t,ext}$ sia uno di questi punti bi-dimensionali generato sulla coda della gaussiana, in particolare appartenente allo stato k, con $k \in \{1 \ldots K\}$. È possibile che tale punto risulti trovarsi più vicino al centroide $\mu_{k'}$, corrispondente alla distribuzione di un altro stato $k' \neq k$, piuttosto che a μ_k. In



(a) (b)

Figura 3.4: Confronto di $\dot{S}_{est}$ con $\langle \dot{S}_{bootstrap} \rangle$ e *noise*, per matrice delle transizioni asimmetrica

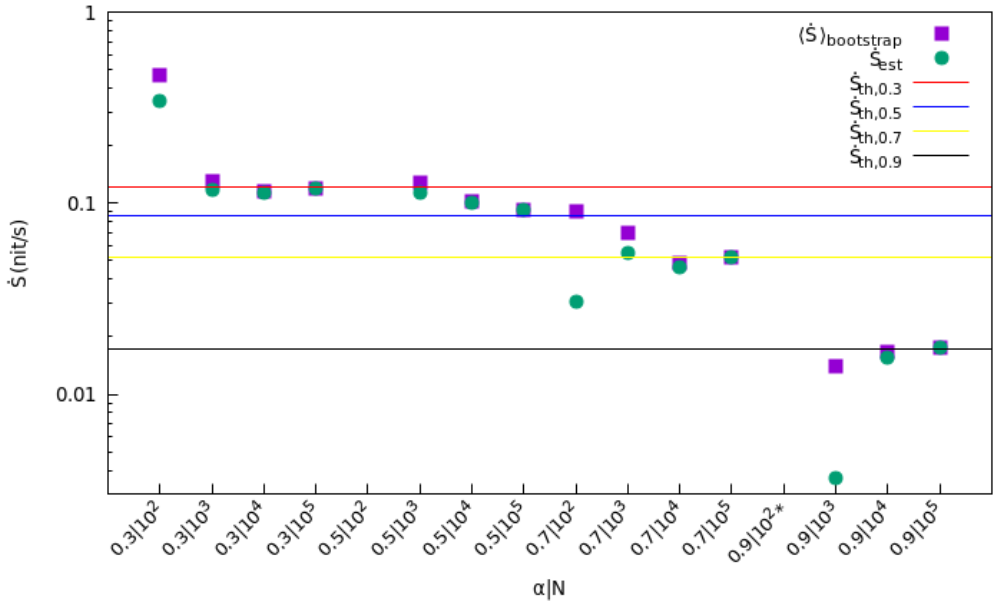


Figura 3.5: Confronto di $\dot{S}_{est}$ con $\langle \dot{S}_{bootstrap} \rangle$ utilizzando $X_t, clean$

questo caso andando ad applicare l'algoritmo del *k-means* tale punto $x^k_{t,ext}$ verrà associato al cluster relativo allo stato k', pur essendo stato generato dalla distribuzione gaussiana $\mathcal{N}(\boldsymbol{\mu}_k, \sigma^2)$, associata allo stato $k \neq k'$. Come conseguenza di questo fatto la serie temporale stimata $\hat{y}_t$, $t = 1, \ldots, N$, risulterà diversa da quella originaria y_t. Questo fatto può indurre un errore sistematico nella stima della produzione di entropia, poichè la formula 2.1 è valida nel caso specifico di processi di Markov.
Per indagare tale fenomeno si è proceduto alla simulazione di un ulteriore set di dati $X_t, clean$, con $t \in \{1, \ldots N\}$, dove $X_t, clean \in \mathbb{R}^2$. Questo è stato generato in modo identico ai precedenti, simulando una catena di Markov y_t associata alla matrice delle transizioni 3.2, facendo variare α ed N negli stessi intervalli già considerati. In questo caso tuttavia il valore della dispersione della gaussiana $\mathcal{N}(\boldsymbol{\mu}_k, \sigma^2)$ associata ad ognuno dei $K = 4$ stati è stato scelto quasi nullo, in modo da avere la certezza virtuale che la serie temporale $\hat{y}_t$, generata tramite l'applicazione dell'algoritmo di *k-means*, coincidesse con quella originala y_t.
Questo nuovo set è stato poi analizzato in maniera analoga a come fatto nelle sezioni precedenti. I valori calcolati sono riportati in scala logaritmica in figura 3.5.

In particolare si osserva come, utilizzando il set di dati $X_t, clean$, scompare il *bias* che si osservava in 3.4b. Infatti i valori della produzione di entropia calcolata $\dot{S}_{est}$, nel caso di N elevato, tendono al valore dell'aspettazione teorica S_{th} e non risultano inferiori a quest'ultimo, come visto invece nella sezione precedente.

Capitolo 4

Applicazione a dati reali

In questa sezione verranno applicate le nozioni esposte precedentemente per valutare la violazione della condizione di bilancio dettagliato, tramite il calcolo della produzione di entropia considerando due gruppi di dati reali. In particolare è stata analizzata la dinamica celebrale macroscopica a riposo di un gruppo di individui sani e di pazienti affetti da ictus celebrale, registrata tramite risonanza magnetica funzionale (fMRI) [12]. Ciascuna registrazione è composta da serie temporali di lunghezza variabile, provenienti da 90 diverse regioni celebrali. In ciascuna serie temporale, il valore corrisponde all'intensità del segnale BOLD (blood-oxygen-level dependent), campionato con frequenza 0.5 Hz.

4.1 Metodologia

Come detto le serie temporali presentano ad ogni tempo t il valore dell'intensità del segnale proveniente da ciascuna delle 90 regioni celebrali osservate. Di conseguenza questi set di dati individuano delle successioni di valori *90-dimensionali* difficili da trattare. In questo caso si possono utilizzare le tecniche presentate nel capitolo 2.
Si fa notare che le serie temporali all'interno di ciascun gruppo sono state concatenate per aumentarne l'evidenza statistica, formando così 2 dataset *'estesi'*. Come visto anche in 3 infatti i valori della produzione di entropia calcolata S_{est} per i campioni aventi un maggiore numero di dati erano quelli che meglio si avvicinavano al valore S_{th}. Tuttavia, sebbene le serie temporali siano concatenate, non sono state valutate le transizioni tra diversi individui al fine della stima delle probabilità di transizione, mantenendo perciò la divisione tra la dinamica neurale di ciascun soggetto.
Una volta stimata la matrice delle transizioni $\hat{P}$ si va ad osservare se tutte le transizioni tra stati possibili siano presenti almeno una volta, cioè nessun termine $\hat{P}_{ij}$ della matrice sia nullo (come visto anche in 3.1). Se tale condizione viene rispettata si procede al calcolo della produzione di entropia, tramite la formula 2.1.
L'analisi delle serie temporali è stata realizzata utilizzando separatamente le metodologie proposte in 2. Queste in particolare hanno permesso di identificare, partendo dalle serie temporali la successione degli stati occupati dalla dinamica celebrale nella sua evoluzione temporale.

4.2 Analisi tramite utilizzo dell'algoritmo di k-means gerarchico

Il primo metodo segue l'esempio del lavoro presentato in [2]. In questo caso viene applicato l'algoritmo del *k-means gerarchico* (2.2.2) ai campioni di dati *'estesi'*, considerando un numero k di cluster *90-dimensionali* crescente, in particolare con $k \in [2, 12]$. Ciascun valore all'interno del campione viene associato ad uno dei k cluster e viene osservata la successione di cluster occupati. Questa permette di andare a valutare la dinamica celebrale macroscopica e di conseguenza stimare la matrice delle transizioni dalla quale procedere al calcolo della produzione di entropia.
I risultati ottenuti tramite l'analisi dei campioni con l'algoritmo del *k-means gerarchico* sono rappresentati in figura 4.1. Dal grafico si può vedere come il valore della produzione di entropia calcolata

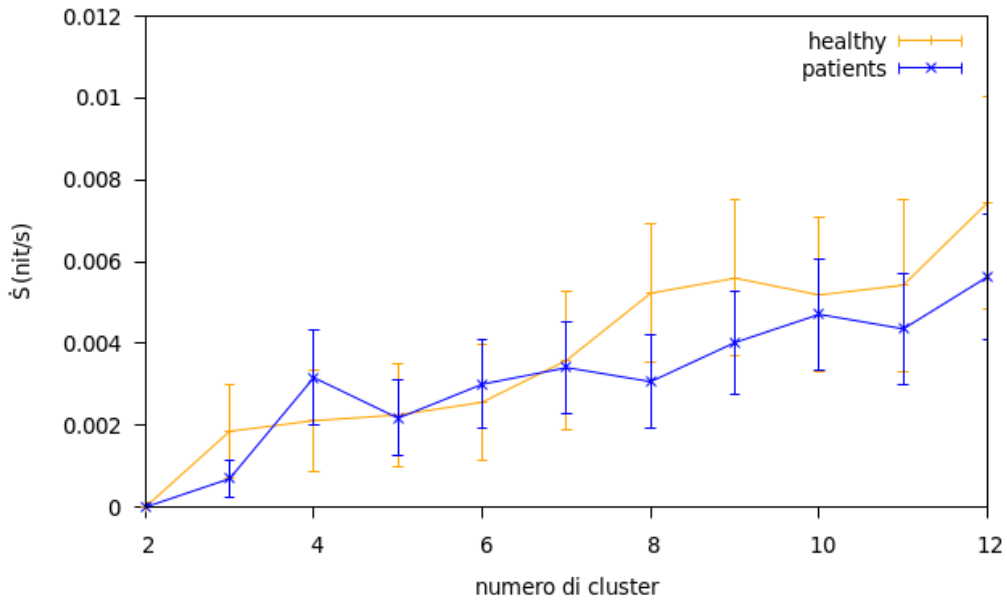


Figura 4.1: Confronto nel valore di $\dot{S}$ per pazienti ed individui sani al variare del numero di cluster

tenda ad aumentare generalmente con il numero k di cluster, fatto evidenziato anche in [2].
In generale potremmo dire che, almeno per $k \geq 4$, la produzione di entropia implica che la dinamica celebrale rompa la condizione di *bilancio dettagliato* su scala macroscopica. Poichè tale condizione viene realizzata per serie temporali registrate a partire da individui a riposo, si potrebbe pensare che ciò sia valido anche andando ad aumentare gli stimoli fisici e cognitivi, come di fatto mostrato in [2].
Inoltre si nota che, benchè ci si potessero aspettare delle differenze tra i valori si $\dot{S}$ appartenenti ai diversi campioni, questo non risulta evidente da 4.1. Per approfondire questo punto sarebbe di conseguenza necessaria un'analisi maggiormente approfondita, andando magari a dividere ulteriormente i pazienti a seconda della gravità della loro condizione. Queste ulteriori analisi non sono oggetto di questo elaborato, .

4.3 Analisi tramite PCA

Il secondo metodo è basato invece sul lavoro presentato in [4]. In questo caso le serie temporali *'estese'* sono state proiettate lungo le due dimensioni principali individuate tramite l'analisi delle componenti principali. Lo spazio bi-dimensionale così creato è stato poi diviso in $N \times N$ celle regolari, considerando un intervallo compreso tra -2 e 2 deviazioni standard dalla media, con N variabile. Considerando ciascuna cella come un singolo stato è stata valutata successivamente la dinamica celebrale osservando la successione degli stati occupati ed andando a stimare la matrice delle transizioni.
Innanzitutto si giustifica la scelta di aver proiettato i campioni lungo solamente due direzioni direzioni principali invece di sceglierne un numero maggiore. In particolare, osservando 4.2, si nota come per entrambi i campioni queste spieghino circa il 30% della varianza totale dell'attività neurale.
Procedendo con il calcolo della produzione di entropia risulta tuttavia che questo sia possibile solamente per una scelta di $N = 2$ celle. Per ciascun valore $N > 2$ risulta che la matrice delle transizioni stimata $\hat{P}$ abbia degli elementi $\hat{P}_{ij}$ di valore nullo. In particolare questo si osserva per stati i e j *'lontani'* tra loro, cioè non adiacenti. La probabilità di osservare transizioni di questo tipo è infatti minore rispetto a quella relativa a stati contigui ed inoltre tende a diminuire tanto più le celle (cioè gli stati) sono lontane tra loro. Come conseguenza di questo fatto con l'aumentare del numero delle celle N si osserva che la probabilità di ottenere transizioni tra stati non adiacenti scende velocemente a 0, rendendo di fatti impossibile calcolare la produzione di entropia.
Per $N = 2$ i valori ottenuti per i due campioni di individui sono rappresentati in 4.1.

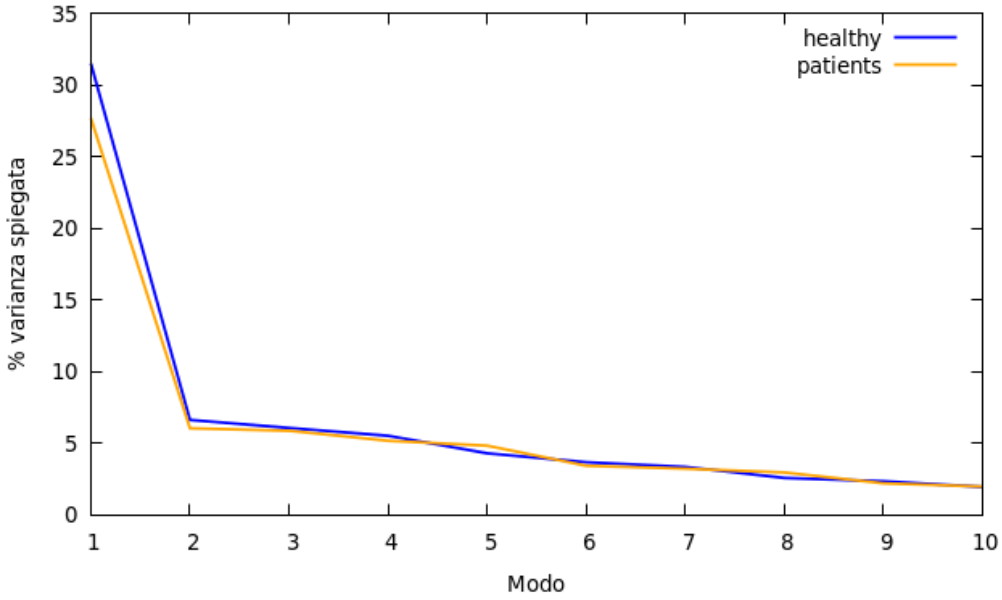


Figura 4.2: Varianza spiegata dai primi modi principali

	$\hat{S}$ (10^{-4} nit/s)
healthy	3 ± 5
patients	8 ± 6

Tabella 4.1: Valori di $\dot{S}$ per pazienti ed individui sani per $N = 2$ celle

Tali valori risultano essere confrontabili con quelli riportati in 4.1 per $k \leq 3$, tuttavia non si dispone di un numero sufficiente di dati per eseguire un confronto consistente tra le due metodologie utilizzate. Inoltre il numero esiguo di valori calcolati non è sufficiente per riuscire ad individuare, utilizzando tale metodo, la rottura o meno della condizione di *bilancio dettagliato* nel cervello su scala macroscopica.

Conclusioni

In questo elaborato è stato approfondito un metodo non invasivo che permette di valutare la rottura della condizione di *bilancio dettagliato* su scala macroscopica andando a quantificare la produzione di entropia. Tale metodo è basato su tecniche non invasive (tipicamente tecniche di imaging) tramite le quali è possibile osservare la dinamica macroscopica del sistema, andando ad ottenere dati organizzati in serie temporali . In particolare ci si è concentrati nel caso di processi Markoviani, mostrando come la rottura della condizione di *bilancio dettagliato* abbia come conseguenza una produzione di entropia non nulla.
Applicando tale metodo ai campioni di dati artificiali, realizzati simulando processi Markoviani, è stato poi possibile verificare come la stima della produzione di entropia sia confrontabile con l'aspettazione teorica ed in particolare tenda a questa per dataset di dimensione estesa.
Tale metodo apre la strada ad importanti sviluppi futuri. Ad esempio, come già accennato in 4.2, si potrebbero approfondire l'analisi proposta per verificare la presenza o meno di differenze nei valori della produzione di entropia tra pazienti ed individui sani.
Inoltre recenti studi hanno suggerito che il trasferimento di energia ed informazione tra le varie regioni del cervello umano possa essere facilitato dalla presenza di turbolenze [13]. Considerando quindi lo stretto legame tra rottura della condizione di equilibrio dettagliato e consumo di energia a livello cellulare e molecolare, si potrebbe andare ad indagare se la produzione di entropia sia associata ad un aumento del metabolismo neurale.
Infine, dato che la violazione del bilancio dettagliato su scala macroscopica può emergere come conseguenza di asimmetrie del sistema su scala miscroscopica [2], lavori futuri potrebbero essere rivolte all'indagine di una relazione tra la rottura della condizione di bilancio dettagliato e la connettività strutturale e funzionale tra varie aree del cervello.
Viene fatto notare inoltre come, benchè nel presente elaborato sia stato considerato il cervello come sistema di riferimento, tale metodologia può essere applicata a ciascun sistema per cui si possano ottenere dati registrati tramite serie temporali. In particolare tale metodo può essere usato per valutare la rottura della condizione di *bilancio dettagliato* in altri sistemi viventi complessi [14] [15], ma può essere considerato anche per indagare sistemi attivi non-biologici [16].

Bibliografia

[1] Erwin Schrödinger. *Che cos'è la vita? La cellula vivente dal punto di vista fisico*. Milano.

[2] Christopher W Lynn et al. «Broken detailed balance and entropy production in the human brain». In: (2021). DOI: 10.1073/pnas.2109889118/-/DCSupplemental.

[3] F. S. Gnesotto et al. «Broken detailed balance and non-equilibrium dynamics in living systems: a review». In: *Reports on progress in physics. Physical Society (Great Britain)* 81.6 (apr. 2018). DOI: 10.1088/1361-6633/AAB3ED. URL: https://pubmed.ncbi.nlm.nih.gov/29504517/.

[4] Yonatan Sanz Perl et al. «Nonequilibrium brain dynamics as a signature of consciousness». In: *Physical Review E* 104.1 (lug. 2021). DOI: 10.1103/PHYSREVE.104.014411. arXiv: 2012.10792.

[5] C. E. Shannon. «A mathematical theory of communication». In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[6] Luca Cocconi et al. «Entropy Production in Exactly Solvable Systems.» eng. In: *Entropy (Basel, Switzerland)* 22.11 (nov. 2020). ISSN: 1099-4300 (Electronic). DOI: 10.3390/e22111252.

[7] C Maes, K Netočn´ Netočn´y e B Wynants. «On and beyond Entropy Production: the Case of Markov Jump Processes». In: *Markov Processes Relat. Fields* 14 (2008), pp. 445–464. URL: http://itf.fys.kuleuven.be/%CB%9Cchrist,.

[8] Hervé Abdi e Lynne J. Williams. «Principal component analysis». In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.4 (lug. 2010), pp. 433–459. DOI: 10.1002/WICS.101. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/wics.101%20https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.101%20https://wires.onlinelibrary.wiley.com/doi/10.1002/wics.101.

[9] Hervé Abdi e Neil J Salkind. *Encyclopedia of measurement and statistics*. 2007.

[10] M Venkat Reddy, M Vivekananda e Hyderabad -Telangana. «Divisive Hierarchical Clustering with K-means and Agglomerative Hierarchical Clustering». In: *International Journal of Computer Science Trends and Technology (IJCST)* 5 (2013).

[11] G. A. Young e H. E. Daniels. «Bootstrap Bias». In: *Biometrika* 77.1 (1990), pp. 179–185. ISSN: 00063444. URL: http://www.jstor.org/stable/2336060.

[12] Richard B Buxton. «The physics of functional magnetic resonance imaging (fMRI)». In: *Reports on Progress in Physics* 76.9 (2013), p. 096601.

[13] Gustavo Deco e Morten L. Kringelbach. «Turbulent-like Dynamics in the Human Brain». In: *Cell Reports* 33.10 (2020), p. 108471. ISSN: 2211-1247. DOI: https://doi.org/10.1016/j.celrep.2020.108471. URL: https://www.sciencedirect.com/science/article/pii/S2211124720314601.

[14] Claudio Castellano, Santo Fortunato e Vittorio Loreto. «Statistical physics of social dynamics». In: *Rev. Mod. Phys.* 81 (2 mag. 2009), pp. 591–646. DOI: 10.1103/RevModPhys.81.591. URL: https://link.aps.org/doi/10.1103/RevModPhys.81.591.

[15] Gijsje H Koenderink et al. «An active biopolymer network controlled by molecular motors.» eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.36 (set. 2009), pp. 15192–15197. ISSN: 1091-6490 (Electronic). DOI: 10.1073/pnas.0903974106.

[16] Sriram Ramaswamy. «The Mechanics and Statistics of Active Matter». In: *Annual Review of Condensed Matter Physics* 1.1 (2010), pp. 323–345. DOI: 10.1146/annurev-conmatphys-070909-104101. eprint: https://doi.org/10.1146/annurev-conmatphys-070909-104101. URL: https://doi.org/10.1146/annurev-conmatphys-070909-104101.