FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI

CORSO DI LAUREA MAGISTRALE IN FISICA

Tesi di Laurea in Fisica Teorica

Analisi del Flusso di Informazione in Sistemi Complessi

Relatori:
Prof. Sebastiano STRAMAGLIA
Prof. Roberto BELLOTTI

Laureando:
Ruggiero SANTORO

ANNO ACCADEMICO 2012-2013

Indice

Introduzione

Cos'è un *sistema complesso*? Probabilmente questa è la prima domanda che sorgerebbe, leggendo il titolo di questa tesi. Col termine *sistema complesso* si identifica un sistema le cui parti, considerate singolarmente, interagiscono tra loro provocando cambiamenti della struttura complessiva. In tal maniera, dunque, si può osservare che le proprietà del singolo elemento appartenente al sistema vengono meno in luce di una maggiore informazione fornita dalle relazioni che legano gli elementi stessi. Sistemi complessi tipici possono essere il sistema nervoso, i sistemi sociali ed economici e gli ecosistemi. Questi sono solo alcuni esempi di sistemi complessi, tali da mettere in luce la grande vastità di applicazioni possedute dalla teoria relativa ad essi.

Lo scopo di questo lavoro di tesi risulta essere l'illustrazione di una classe di approcci utili per effettuare lo studio dei sistemi complessi in termini di relazioni causali. Per raggiungere tale obiettivo si è cercato, tramite tali classi, di determinare le relazioni di causa-effetto tra le varie componenti di un sistema generico, applicando,infine, le procedure di analisi generate da tali approcci di indagine a dei casi concreti per testarne l'utilizzo.

Nel primo capitolo di questa tesi viene fornita una descrizione di tali sistemi, visti in generale. In detto capitolo viene trattato lo studio dei sistemi complessi attraverso la teoria dei grafi, introducendo concetti quali la distribuzione dei gradi di connessioni nel sistema complesso (anche chiamato *network*) o le proprietà di Small World utili a comprendere le basi teoriche dei sistemi complessi. Fatta questa doverosa introduzione, nel capitolo due e tre ci si è focalizzati sull'informazione presente nei sistemi complessi e come tale informazione passi da un elemento ad un altro. Nel capitolo due viene

trattata la teoria dell'informazione la cui nascita è fatta coincidere, secondo opinione comune, con l'articolo di Claude Shannon del 1948 [Sh48]. In tale trattazione è stata introdotta inizialmente la versione di Shannon dell'entropia, ossia l'entropia relativa ad una variabile aleatoria generica, generalizzandola successivamente a variabili aleatorie multidimensionali. All'interno del capitolo viene descritta, inoltre, la relazione tra entropia d'informazione e Secondo Principio della Termodinamica. È stato poi trattato il concetto di mutua informazione e di entropia relativa che rispettivamente rappresentano l'informazione relativa ad un ente X contenuta in un ente Y e una misura dell'inefficienza nella strategia di utilizzo di una distribuzione q al posto di una distribuzione p per descrivere la variabile aleatoria X. Nel capitolo tre è stato introdotto il concetto di causalità, prima sotto l'aspetto della Fisica in generale e successivamente introducendo la metodologia statistica di calcolo della causalità stessa: la causalità di Granger. In seguito è stata trattata una misura direzionale più accurata della causalità tra due enti attraverso il concetto di Tranfer Entropy. Relativamente alla causalità di Granger successivamente è stato applicato il Partial Conditioning, che risulta indispensabile per avere dati informativi nel caso di studio di sistemi di cui si possegga un numero di informazioni esiguo.

Nel capitolo quattro è stato applicato il Partial Conditioning per il calcolo della causalità di Granger a due set di dati: la coltura cellulare ""HeLa"" e la Risonanza Magnetica Funzionale fatta ad un uomo mentre dorme. I risultati hanno messo in luce come l'applicazione del Partial Conditioning era indispensabile, poiché il calcolo della causalità di Granger multivariata, dando valori tutti pari a zero, risultava per nulla informativo a monte di risultati informativi ottenuti tramite il Partial Conditioning.

Capitolo 1

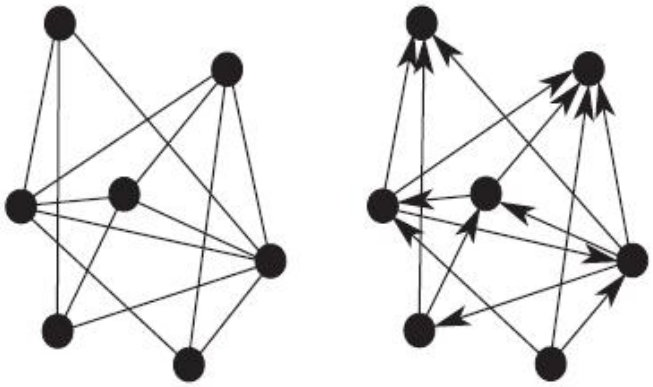
Reti Complesse

La natura, le strutture economiche e sociali ed Internet sono sistemi organizzati in modo tale da avere un numero enorme di elementi che interagiscono tra loro. In questi sistemi è possibile osservare che le proprietà dei singoli elementi, che ne fanno parte, debbono essere poste in secondo piano in luce di una maggiore informazione fornita dalle relazioni che legano gli elementi stessi. Strutture di questa natura vengono chiamate sistemi complessi. È possibile descrivere in termini visivi tali sistemi attraverso le reti complesse, che risultano essere delle strutture geometriche composte da linee e punti, che mettono in risalto le relazioni tra gli elementi del sistema complesso cui fanno riferimento.

1.1 Networks

La teoria che studia il sistema dei network è la teoria dei grafi. Infatti è osservabile che un qualsivoglia sistema complesso può essere visto come un grafo. Si definisce un grafo attraverso l'insieme $G = (\mathscr{N}, \mathscr{L})$ che risulta essere la coppia degli insiemi $\mathscr{N}$ ed $\mathscr{L}$. L'insieme $\mathscr{N} = \{n_1, n_2, \ldots, n_N\}$, supposto diverso dell'insieme vuoto, rappresenta l'insieme degli elementi, anche chiamati *nodi*, che interagiscono all'interno del grafo, e l'insieme $\mathscr{L} = \{l_1, l_2, \ldots, l_K\}$ contiene le coppie di elementi di $\mathscr{N}$ che interagiscono tra di loro, meglio conosciuti col termine di *link*. Due nodi collegati tramite un link verrano detti *vicini*. A seconda che i network risultino *ordinati*

(o diretti) o *non-ordinati (o non-diretti)* i link saranno coppie ordinate o non ordinate. Nella prima eventualità non prenderemo in considerazione il legame in sè, ma focalizzeremo l'attenzione sul nodo da cui parte la relazione, pertanto avremo un legame direzionale tra i nodi che compongono il link. Nel secondo caso, invece, sarà importante osservare il legame stesso e non la direzione di quest'ultimo, per cui non ci interesserà l'ordine della coppia, ma semplicemente gli elementi al suo interno. N e K rappresenteranno rispettivamente il numero di nodi e link.



(a) *Grafo non-diretto.* **(b)** *Grafo diretto.*

Figura 1.1: Tipologie di grafi osservabili con $N = 7$ e $K = 14$.

Ci si riferisce a un nodo, etichettandolo col suo indice: n_i. Alla stessa maniera si indicherà il singolo link attraverso l_{ij} o attraverso la coppia (i, j). Come è facile vedere, nel caso di networks diretti si ha che $l_{ij} \neq l_{ji}$, mentre nel caso di networks non-diretti si osserva che $l_{ij} = l_{ji}$. Come mostrato in figura 1.1, i link sono rappresentati graficamente da linee che collegano dei punti che rappresentano i nodi. Il numero di link in un network con un numero di nodi pari ad N varia da un minimo di 0 ad un massimo di $N(N-1)/2$, eventualità nella quale il grafo, rappresentante il network, viene detto *completo*. Nell'eventualità che un grafo sia completo lo stesso verrà denotato con l'espressione K_N. Inoltre il grafo verrà detto sparso se $K << N^2$ e denso se $K \approx o(N^2)$.

Un sottografo $G' = (\mathscr{N}', \mathscr{L}')$ di $G = (\mathscr{N}, \mathscr{L})$ è un grafo in cui $\mathscr{N}' \subseteq \mathscr{N}$ e $\mathscr{L}' \subseteq \mathscr{L}$. Qualora un sottografo contenga tutti i link che legano i nodi di $\mathscr{N}'$ presenti in $\mathscr{L}$ allora tale sottografo verrà detto *indotto* e verrà denotato

come $G = G[\mathcal{N}']$. Un caso particolare che presenta questa caratteristica è il sottografo dei vicini del nodo i, denotato tramite G_i.

Una proprietà importante della teoria dei grafi è quella relativa alla raggiungibilità di un nodo del grafo, cioè come un nodo sia raggiungibile partendo da un altro nodo preso a caso. Infatti, due nodi che non risultano essere vicini possono essere ugualmente collegati tra di loro. Per questo scopo vanno definiti alcuni concetti. Un *cammino* dal nodo i al nodo j risulterà essere un insieme di nodi e link che partono da i e terminano in j. La lunghezza di detto cammino sarà pari al numero di link appartenenti al cammino stesso. Un cammino generico potrà toccare più volte alcuni nodi o attraversare più volte alcuni link in quanto alcuna restrizione è stata posta a riguardo. Per questo motivo è opportuno definire il *trial* e il *path* dove il primo rappresenta un percorso nel quale nessun link viene ripetuto, mentre il secondo rappresenta un percorso nel quale nessun nodo viene raggiunto più di una volta. Definiremo *percorso minimo* tra due nodi la lunghezza minima che un percorso deve avere per connetterli. Sfruttando la definizione di percorso, osserveremo che i grafi potranno essere *connessi* o *non-connessi*. Risulteranno essere connessi qualora per ogni coppia di nodi i e j vi sarà un percorso che li connette, saranno, invece, non-connessi qualora questa proprietà non vi sia.

Come ultima osservazione è possibile definire la rappresentazione matriciale di un grafo attraverso l'utilizzo della matrice di *connettività (o adiacenza)* $\mathscr{A}$. Detta matrice risulterà essere una matrice $N \times N$ le cui componenti a_{ij} $(i, j = 1, \ldots, N)$ saranno pari ad 1 qualora l_{ij} esista e pari a 0 in caso contrario. Inoltre, le componenti appartenenti alla diagonale, a_{ii}, saranno tutte pari a 0. La matrice di connettività sarà banalmente simmetrica in caso di grafi non-diretti. Oltre alla matrice di connettività la rappresentazione matriciale può essere ottenuta tramite la matrice d'*incidenza* $\mathscr{B}$, la quale risulta essere una matrice $N \times K$ i cui termini b_{ij} saranno pari ad 1, qualora il nodo i fosse toccato dal link j e pari a 0 altrimenti.

1.1.1 Grado del Nodo e Distribuzione dei Gradi

Preso un nodo appartenente ad un grafo definiremo *grado (o grado di connettività)* k_i di detto nodo il numero di link che lo toccano. Utilizzando la matrice di adiacenza si otterrà che:

$$k_i = \sum_{j \in \mathcal{N}} a_{ij}. \tag{1.1.1}$$

Nel caso in cui il grafo considerato risulti essere diretto, allora il grado di un nodo che ne fa parte verrà suddiviso in due componenti $k_i^{out} = \sum_j a_{ij}$ e $k_i^{in} = \sum_j a_{ji}$, dove il primo è riferito ai link che partono dal nodo i mentre il secondo è riferito ai link che arrivano al nodo i. Naturalmente il grado totale k soddisferà alla relazione $k_i = k_i^{out} + k_i^{in}$. Utilizzando i gradi dei nodi di tutto il network, sarà possibile definire la *distribuzione dei gradi* $P(k)$ come la probabilità che un nodo scelto a caso tra tutti i nodi del grafo abbia grado k o come la frazione dei nodi presenti nel grafo che hanno grado k. Nel caso di grafi diretti verranno considerate due distribuzioni relative al grado k^{out} e k^{in} pari rispettivamente a: $P(k^{out})$ e $P(k^{in})$. Come per ogni variabile aleatoria di cui è conosciuta la distribuzione di probabilità, possiamo calcolare per il grado di un nodo i momenti relativi. Il momento di ordine n di $P(k)$ risulterà pari a:

$$\langle k^n \rangle = \sum_k k^n \, P(k). \tag{1.1.2}$$

Il momento di ordine uno sarà il grado di connettività medio nel grafo, mentre il momento di ordine due misurerà le fluttuazioni del grado di connettività. Diremo che un network è *correlato* se la probabilità relativa alla possibilità che un nodo di grado k risulti collegato ad un nodo di grado k' dipende da k. Nel caso in cui detta probabilità risulti indipendente da k, allora detto network verrà definito *non-correlato*. Qualora un network risulti non-correlato, allora la distribuzione dei gradi determinerà completamente le proprietà statistiche dello stesso. In caso contrario, ovvero qualora il network risulti essere correlato, sarà necessario introdurre la probabilità condizionata $P(k'|k)$ definita come la probabilità che un link, che tocca un

nodo di grado k, metta detto nodo in relazione con un nodo di grado k'. Ovviamente, dovendo detta probabilità condizionata soddisfare la regola di normalizzazione, avremo che $\sum_{k'} p(k'|k) = 1$. Ulteriormente la probabilità condizionata deve soddisfare la condizione di bilancio dettagliato:

$$k\,P(k'|k)\,P(k) \;=\; k'\,P(k|k')\,P(k') \tag{1.1.3}$$

la quale stabilisce che il numero di link, che collegano un nodo di grado k' ad un nodo di grado k, deve essere uguale al numero di link, che collegano un nodo di grado k ad un nodo di grado k'. Questa proprietà è molto importante in quanto stabilisce forti restrizioni relativamente alla forma che $P(k'|k)$ può assumere qualora sia già assegnata una $P(k)$.

Sebbene il grado di correlazione possa essere sintetizzato attraverso la probabilità condizionata $P(k'|k)$, sorge un problema relativamente alla valutazione della stessa. Nel caso dei network reali, a causa delle dimensioni finite del grafo, la stima relativa a $P(k'|k)$ risulta essere poco precisa. Per questo motivo si introduce il *grado medio dei primi vicini* di un nodo i pari a:

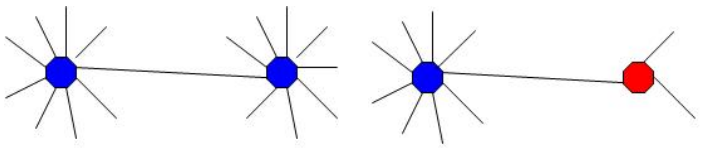
$$k_{nn,i} = \frac{1}{k_i}\sum_{j\in\mathscr{N}_i} k_j = \frac{1}{k_i}\sum_{j=1}^{N} a_{ij}k_j \tag{1.1.4}$$

dove la somma varia sui nodi appartenenti a $\mathscr{N}_i$, insieme degli elementi appartenenti al sottografo dei vicini del nodo i, G_i. Detto valore può essere riscritto esplicitando la dipendenza rispetto al grado k invece che la dipendenza dal nodo i, trasformando la relazione in:

$$k_{nn}(k) = \sum_{k'} k'P(k'|k). \tag{1.1.5}$$

È facile vedere che, se un grafo risultasse non-correlato, la $P(k'|k)$ risulterebbe pari a $k'P(k')/\langle k\rangle$, e che la $k_{nn}(k)$ risulterebbe pari a $\langle k^2\rangle / \langle k\rangle$ (nel primo caso basta considerare (1.1.3) e porre $P(k'|k)$ indipendente da k, nel secondo caso basta sostituire la prima relazione (1.1.3) in (1.1.5), tenendo sempre conto dell'indipendenza di $P(k'|k)$ da k). Nel contempo, per quanto riguarda i grafi correlati, definiremo quelli in cui $k_{nn}(k)$ cresce al crescere di k come *assortativi* e quelli in cui $k_{nn}(k)$ decresce al crescere di k come

disassortativi. I grafi assortativi saranno dunque grafi in cui i nodi con un grado k tenderanno in media a collegarsi con nodi di grado simile, mentre i grafi disassortativi saranno grafi in cui i nodi con un grado k molto basso tenderanno in media a collegarsi con nodi di grado più alto (figura 1.2).



(a) *Grafo assortativo.* **(b)** *Grafo disassortativo.*

Figura 1.2: Assortatività nei grafi rappresentanti sistemi complessi.

1.1.2 Percorso Minimo, Diametro e Betweenness

Il percorso minimo è un concetto indispensabile, quando si parla di trasporto o comunicazione all'interno di un network. Un esempio pratico può essere il voler scambiare un'informazione da un nodo di internet ad un altro. Il percorso minimo risulta essere un'ottima scelta per effettuare lo scambio, volta a minimizzare i tempi e rendere più veloce la connessione. Una maniera per rappresentare il percorso minimo è definire la matrice $N \times N$ $\mathscr{D}$, le cui componenti d_{ij} risultano essere pari alla lunghezza del percorso minino che va dal nodo i a quello j. Il valore massimo che le d_{ij} possono assumere verrà detto *diametro* del grafo. Una misura della separazione media tra due nodi presente all'interno di un grafo è fornita dalla *lunghezza media del percorso minimo*, o *lunghezza di percorso caratteristica*, definita come valor medio delle lunghezze di percorso minimo su tutti i nodi del grafo:

$$L = \frac{1}{N(N-1)} \sum_{i,j \in \mathscr{N}, i \neq j} d_{ij}. \tag{1.1.6}$$

C'è una grossa problematica relativa a questa definizione. Qualora vi fossero nodi disconnessi L divergerebbe. Per evitare questa evenienza, è possibile restringere la sommatoria solo relativamente a nodi che risultano connessi. Un'altra soluzione è considerare l'*efficienza* di G:

$$E = \frac{1}{N(N-1)} \sum_{i,j \in \mathscr{N}, i \neq j} \frac{1}{d_{ij}} \tag{1.1.7}$$

la quale differisce da (1.1.6) solo per il fatto che quest'ultima rappresenta la media dei percorsi minimi, mentre (1.1.7) rappresenta la media armonica dei percorsi minimi. È banalmente osservabile che il problema relativo alla divergenza in (1.1.6) viene risolto in (1.1.7), poichè le componenti disconnesse nell'efficienza daranno contributo nullo.

È lecito supporre che la comunicazione tra nodi che non risultano essere vicini (supponiamo siano gli elementi i e j) dipenda dai nodi appartenenti ai percorsi che connettono i due elementi i e j stessi. Sfruttando questa idea sarebbe ideale supporre di calcolare il numero di percorsi, che connettono i a j e che passano per un nodo prefissato. In tal maniera si potrà definire la *betweenness* di un nodo. Questa proprietà, insieme al grado del nodo e alla *prossimità* di un nodo (definita come l'inverso della distanza media dagli altri nodi), rappresenta uno strumento per il calcolo della *centralità di un nodo*. La *betweenness* b_i di un nodo i è definita come:

$$b_i = \sum_{j,k \in \mathcal{N}, j \neq k} \frac{n_{jk}(i)}{n_{jk}} \tag{1.1.8}$$

dove n_{jk} rappresenta il numero di percorsi minimi che connettono j a k, mentre $n_{jk}(i)$ rappresenta il numero di percorsi minimi che connettono j a k passanti per il nodo i. Il concetto di betweenness può essere ulteriormente esteso relativamente ai link, sostituendo, all'interno della stessa nozione, il nodo, per cui deve passare il percorso minimo, con il link, per cui deve passare il percorso stesso.

1.2 Clustering

Il *clustering*, conosciuto anche col termine di *transitività*, è una proprietà tipica delle reti di conoscenze. Infatti, questa proprietà segue l'idea che, se due persone hanno un amico in comune, è probabile che si conoscano. Dal punto di vista della teoria dei grafi, detta proprietà si traduce con la presenza nel network di un alto numero di sottografi completi con 3 nodi, anche chiamati triangoli, denotati con l'espressione K_3.

La quantificazione del clustering avviene attraverso l'utilizzo della *transitività* T del grafo che risulta essere la frazione di triadi (triple di nodi

connessi) che formano un triangolo (K_3):

$$T = \frac{3 \times \#\, di\, triangoli\, in\, G}{\#\, di\, triadi\, in\, G}. \tag{1.2.1}$$

Il 3 a numeratore serve a compensare il fatto che ogni triangolo conta per tre triadi connesse, ognuna delle quali parte da uno dei tre nodi che forma il triangolo stesso. Si può osservare che il valore di T è compreso tra 0 ed 1 e che il valore $T = 1$ si avrà solo nel caso in cui il grafo sia completo.

C'è un'alternativa all'osservazione del clustering, ottenibile grazie al *coefficiente di clustering*. S'introduce preliminarmente la quantità c_i (il *coefficiente di clustering locale relativo al nodo* i). Per farlo calcoliamo la probabilità che il termine a_{jm} della matrice di connettività $\mathscr{A}$ possa essere pari ad 1, ovvero che j ed m siano vicini tra loro, posto che j ed m risultino essere due vicini per il nodo i. Il valore di detta probabilità si ottiene mettendo a rapporto il numero e_i dei link presenti in G_i (che ricordiamo essere il sottografo dei vicini del nodo i) col numero dei link presenti in G_i stesso, qualora quest'ultimo fosse completo, ossia il numero massimo di link ammissibile per G_i. Ciò detto, dunque, il coefficiente di clustering locale relativo al nodo i, c_i, sarà pari a:

$$c_i = \frac{2\, e_i}{k_i(k_i - 1)} = \frac{\sum_{j,m} a_{jm}}{k_i(k_i - 1)} \tag{1.2.2}$$

ove la somma è fatta relativamente ai nodi j ed m presenti in G_i. Il coefficiente di clustering C sarà dunque facilmente calcolabile a partire dai valori di c_i attraverso la relazione:

$$C = \langle c \rangle = \frac{1}{N} \sum_{i \in \mathscr{N}} c_i. \tag{1.2.3}$$

Dalla definizione traspare che tanto C quanto c_i prenderanno valore all'interno dell'intervallo $[0, 1]$. Può essere, inoltre, utile definire il coefficiente di clustering della classe di connettività pari a k, denotato tramite $c(k)$. Esso sarà calcolato prendendo il valor medio delle c_i relativamente a tutti i nodi di grado k. Come ultima misura del clustering del grafo G possiamo vedere

l'*efficienza locale* definita come:

$$E_{loc} = \frac{1}{N} \sum_{i \in \mathcal{N}} E(G_i) \tag{1.2.4}$$

dove $E(G_i)$ è l'efficienza di G_i definita nella relazione (1.1.7).

1.3 La proprietà di Small-World

Sovente i network reali risultano organizzati in diverse aree ad alta connettività che risultano connesse tra loro da link che fungono da ponte tra le varie aree del grafo, utili a velocizzare la comunicazione tra un nodo ed un altro, minimizzando in tal modo il valore del minimo percorso.

A partire da questa proprietà è interessante osservare come i network reali si comportano. Si consideri un reticolo D dimensionale a forma ipercubica. Si può osservare relativamente a detto network che il numero medio di nodi che vanno toccati, per connettere due nodi qualsiasi del grafo, cresce relativamente alla crescita del numero N di nodi totali presenti nel grafo come $N^{1/2}$. Questa proprietà geometrica ottenibile per i grafi random viene meno nei network reali per i quali è possibile vedere che la lunghezza media del percorso minimo, eq. (1.1.6), dipende da N al massimo logaritmicamente. Questa proprietà viene definita *proprietà di small-world*.

Una prima analisi di detto comportamento la si osserva nello studio effettuato da Milgram [Mi67] il quale studiò il numero medio di step da effettuare per collegare due nodi in un network di conoscenze. Il suo esperimento consisteva nello scegliere persone a caso nel Nebraska e chiedere loro di inviare una lettera ad un destinatario lontano che si trovava a Boston. Lo scambio della lettera poteva avvenire solo tra persone interne alla propria sfera di conoscenze, scegliendo la persona che più sembrava candidata a conoscere, anche in maniera indiretta, il destinatario finale. Volendo fare una supposizione sul possibile risultato dell'esperimento, verrebbe naturale pensare che il numero medio degli scambi della lettera debba essere un numero abbastanza alto, data la natura del setup sperimentale. Risultò abbastanza sorprendente osservare che il numero medio degli scambi della lettera ottenuto tramite l'esperimento fosse pari a sei. Questo tipo di esperimento è

più conosciuto nella modalità nella quale, non è lo scambio di una lettera a mettere in relazione i nodi del network, ma le strette di mano. Inoltre in tale modalità i nodi da collegare risultano essere un uomo a caso e il presidente degli Stati Uniti d'America.

La proprietà di small-world è stata osservata in svariati networks reali, inclusi quelli biologici. A differenza dei networks random, i networks reali sono spesso associati alla presenza di clustering, denotato da alti valori del coefficiente di clustering definito nell'equazione (1.2.3). È sfruttando questa evenienza che Watts e Strogatz [Wa98] hanno definito i networks small-world come quei networks che hanno un alto coefficiente di clustering C. In termini di efficienza questa definizione equivale a dire che detti networks devono avere un alto valore di efficienza globale E_{glob}, eq. (1.1.7), ed un alto valore di efficienza locale E_{loc}, eq. (1.2.4), cioè network estremamente efficienti nello scambio di informazione sia su scala locale che su scala globale.

1.3.1 Modello di Watts-Strogatz

Come già specificato, il *modello di Watts-Strogatz* è un metodo utilizzato per costruire grafi, che denoteremo come $G^{WS}_{N,K}$, aventi proprietà di small-world. Il modello si basa sulla creazione di networks a partire da grafi random in cui ognuno dei possibili link si conserva con probabilità $1-p$ e cambia estremità con probabilità p. Si parte da un anello di N nodi. In detto anello ogni nodo

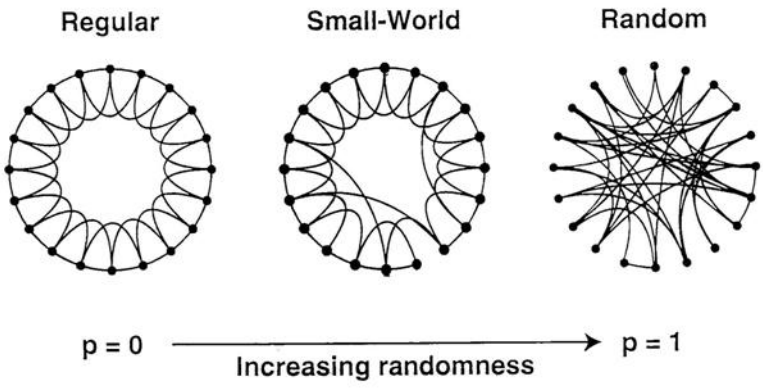


Figura 1.3: Struttura dell'anello di N nodi al variare delle probabilità p.

risulta connesso ai suoi $2m$ vicini (il 2 sta per la simmetria destra sinistra). In tal modo vi saranno un numero totale di link pari a $K = mN$. Per questo anello così definito, si fissa in maniera ciclica un nodo (supponiamo di procedere in senso orario) e si considera sempre in questa maniera ogni link. Quest'ultimo sarà reindirizzato verso un altro nodo con probabilità p, mentre resterà uguale con probabilità $1 - p$. A seconda del valore che p assume, avremo un reticolo regolare per $p = 0$ (in questo caso alcun link verra modificato) e un reticolo completamente random per $p = 1$, con la restrizione che ogni nodo di tale network dovrà avere un minimo grado di connessione: $k_{min} = m$ (Fig. 1.3). Per valori di p intermedi la procedura genera grafi con le proprietà di small-world.

Possiamo cercare di vedere cosa capita, in tali tipi di networks, relativamente alla lunghezza di percorso minimo medio e al valore di clustering al variare dei valori di p. Relativamente alla lunghezza del percorso minimo si

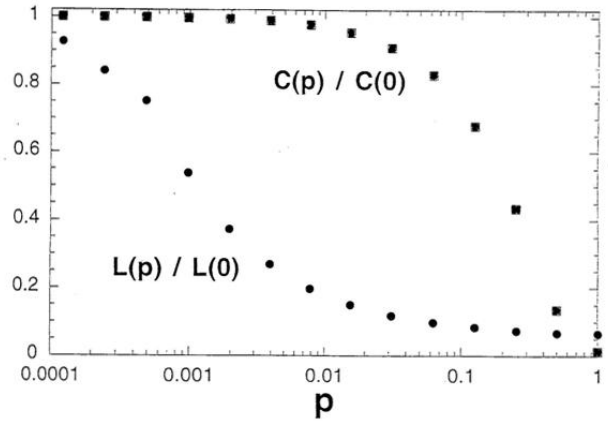


Figura 1.4: Valori del rapporto $C(p)/C(0)$ e del rapporto $L(p)/L(0)$ al variare dei valori di p, probabilità che un link venga reindirizzato.

osserva che esso varia da un massimo valore pari a $N/2K$, quando $p = 0$, ad un minimo, ottenuto quando $p = 1$, pari a $\ln N / \ln K$. Per valori di probabilità intermedi si osserva che il valore di minino si ottiene rapidamente all'aumentare di p (figura 1.4). Relativamente alla dipendenza del parametro di *clustering*, $C(p)$, dalla probabilità p si parte dall'osservazione che per $p = 0$ $C(0) = 3(m-1)/2(2m-1)$. A causa del fatto che con probabilità $(1 - p)$ i link non vengono reindirizzati, due vicini che risultavano collegati

a $p = 0$ restano connessi con probabilità $(1 - p)$ e quindi un triangolo che risultava esistere a $p = 0$ continuerà ad esistere con probabilità $(1 - p)^3$, per cui risulterà una dipendenza di C da p pari a:

$$C(p) \approx \tilde{C}(p) = \frac{3\,(m-1)}{2\,(2m-1)}(1-p)^3 \tag{1.3.1}$$

dove $\tilde{C}(p)$ è ridefinito come il rapporto tra il numero medio di link tra i vicini di un vertice e il numero totale di link possibili tra i vicini e un vertice. Relativamente alla distribuzione del grado di connettività, questa per $p = 0$ risulta essere una delta di Dirac posizionata su $2m$, mentre per $p = 1$ risulta pari a una distribuzione di Poisson. Per valori medi di p la distribuzione risulta essere pari a:

$$P(k) = \sum_{i=0}^{min(k.m,m)} \binom{m}{i} (1-p)^i\, p^{m-i}\, \frac{(pm)^{k-m-i}}{(k-m-i)!}\, e^{pm} \tag{1.3.2}$$

per $k \geq m$ ed è uguale a zero per valori di k più piccoli di m.

1.4 Distribuzioni indipendenti dalla scala

L'approccio iniziale nello studio dei sistemi complessi fu quello di pensare che gli oggetti d'interesse fossero tutti networks omogenei. L'omogeneità implica che le proprietà topologiche di ogni nodo del grafo sono equivalenti, caratteristica tipica dei reticoli regolari o dei grafi random. In questi ultimi ognuno dei possibili $N(N-1)/2$ link è presente con eguale probabilità e perciò, nel limite di grandi dimensioni del grafo, la distribuzione del grado dei nodi è una distribuzione binomiale o poissoniana. Quando si passò a studiare i networks reali ci si aspettò che il comportamento di questi ultimi dovesse portare a distribuzioni centrate intorno ad un valor medio con una ben definita deviazione standard. Tale aspettativa fu disattesa in quanto si osservò che quasi tutti i networks reali avevano una distribuzione dei gradi del nodo a legge di potenza, cioè della forma $P(k) \approx A\,k^{-\gamma}$. Si osservò inoltre che l'esponente γ rispettava la restrizione $2 < \gamma < 3$. Il grado di connessione medio $\langle k \rangle$ in un grafo, che segue una distribuzione a legge di potenza, sotto le restizioni della γ precedentemente poste, risulta ben definito. Stessa cosa

non può dirsi invece per la varianza $\sigma^2 = \langle k^2 \rangle - \langle k \rangle^2$, in quanto il momento di ordine due diverge con il divergere del limite superiore di integrazione:

$$\langle k^2 \rangle = \int_{k_{min}}^{k_{max}} k^2 \, P(k) \approx k_{max}^{3-\gamma}.$$

Un network che soddisfi la proprietà di avere una distribuzione dei gradi dei nodi a legge di potenza viene detto *network indipendente dalla scala* o *scale-free*. Questo perché è possibile osservare che una distribuzione a legge di potenza risulta invariante rispetto a qualunque cambiamento di scala[1]. La conseguenza principale relativamente alla distribuzione a legge di potenza dei gradi dei nodi per i networks reali risulta essere attinente alla struttura di questi ultimi. Infatti risulta che tali grafi sono definiti in maniera tale che pochi nodi, detti *hubs*, risultano avere grande connettività mentre la maggior parte di essi risulta connessa con pochi elementi.

1.4.1 Modello di Barabasi-Albert

Il *modello di Barabasi-Albert* [Ba99] è un modello di crescita dei network che, basandosi sul meccanismo di *preferential attachment*, crea networks con distribuzioni indipendenti dalla scala. Questi tipi di networks, già osservati nel paragrafo precedente, risultano essere utili nello studio del World Wide Web, in quello dei networks di citazioni e in quello dei networks sociali. Il *preferential attachment* si basa sull'idea che un qualcosa da ripartire verrà distribuito in maniera tale che chi ha di più continuerà ad avere di più. È per questo motivo che in passato ci si riferiva a detta caratteristica col termine di "*effetto San Matteo*" ("*Perché a chiunque ha sarà dato e sarà nell'abbondanza; ma a chi non ha sarà tolto anche quello che ha.*" Matteo, 25:29). Questa idea è tipica del mondo di internet in cui siti con un alto grado di connettività tenderanno ad acquisire nuovi link in scala maggiore rispetto a siti con un basso grado di connettività. Stessa cosa vale per i networks sociali che studiano la distribuzione di capitale; in essi si osserva come una maggiore quantità di liquidità vada verso zone con maggior quan-

[1] Supponiamo che $f(x)$ sia una distribuzione a legge di potenza. Presumendo di scalare i valori di x rispetto ad un fattore α, risulta che la distribuzione a legge di potenza è l'unica che soddisfa la relazione $f(\alpha x) = \beta f(x)$. Ciò implica che sotto cambio di scala descrittiva il sistema può essere analizzato sempre nella stessa maniera.

tità di denaro a discapito di zone che ne posseggono meno. La procedura per ottenere networks con distribuzione a legge di potenza, o indipendenti dalla scala, è la seguente: si parte inizialmente da un numero pari ad m_0 di nodi in un network completo e ad ogni istante $t = 1, 2, 3, \ldots, N - m_0$ un nuovo nodo j con grado di connettività pari ad $m \leq m_0$ è aggiunto a dato network, in maniera tale che la probabilità per la quale un nuovo link, che parte da j, tocchi un nodo pre-esistente i risulti linearmente proporzionale al grado di connessione del nodo i nell'istante immediatamente precedente l'introduzione del nuovo link, cioè:

$$\prod_{j \to i} = \frac{k_i}{\sum_l k_l}. \tag{1.4.1}$$

Chiedendosi quali sono le caratteristiche del modello, l'attenzione si riverserebbe sulla distribuzione dei gradi di connettività, sulla lunghezza media del percorso minimo e sul parametro di clustering di quest'ultimo [Al02]. Relativamente alla distribuzione del gradi di connettività è possibile osservare che al tendere di $t \to \infty$ detta distribuzione avrà la classica forma tipica dei modelli scale-free ossia una legge di potenza del tipo:

$$P(k) \approx k^{-\gamma} \tag{1.4.2}$$

ove il parametro γ sarà pari a 3. Qualora nella procedura per generare il network non si fosse utilizzata una probabilità di connessione tra il nodo i ed il nodo j pari a (1.4.1), ossia non avessimo utilizzato il preferential attachment, ma fosse stata utilizzata una probabilità costante, tipo $\prod_{j \to i} = 1/(m_0 + t - 1)$, la distribuzione ottenuta sarebbe stata del tipo: $P(k) = e/m\, exp(-k/m)$, probabilità che non gode della proprietà di invarianza di scala. Tutto ciò implica la necessita di utilizzo del preferential attachment all'interno della procedura di generazione di un network scale-free. La lunghezza media del percorso minimo nel modello di Barabasi-Albert cresce logaritmicamente al crescere della grandezza del network, con una dipendenza analitica pari a:

$$L \approx \frac{log\, N}{log(log\, N)} \tag{1.4.3}$$

Il coefficiente di clustering invece varia al variare della grandezza del network come:

$$C \approx N^{-0.75} \tag{1.4.4}$$

caratterizzando dunque un decadimento più lento rispetto ai network casuali, ove $C \approx N^{-1}$, ma molto differente dai modelli small-world dove il coefficiente di clustering risulta costante e dunque indipendente da N.

1.5 Modularità

La modularità è una funzione usata nell'analisi dei grafi o delle reti. Alcuni esempi di applicazione possono essere le reti di computer o i social-networks. Il suo valore quantifica la qualità della divisione della rete in varie comunità. Otterremo per una buona suddivisione alti valori di modularità ed un basso valore della stessa in caso contrario. All'interno dei moduli la densità di link sarà alta ma, fra un modulo e l'altro ci saranno pochi collegamenti e quindi una densità inferiore. L'uso più comune della modularità è relativo all'ottimizzazione delle tecniche per determinare le comunità nelle reti.

L'idea principale alla base di questo concetto è che i link all'interno di un modulo (o comunità) siano di più di quello che ci si aspetterebbe qualora il network fosse di tipo random, sotto la restrizione che abbia lo stesso numero di link totali di partenza e che mantenga il grado di connettività per ogni singolo nodo. Per tal scopo si utilizza una funzione, denotata come Q, che calcola la differenza tra il numero reale di link in una comunità e il valore aspettato. Per far ciò si usa la matrice di connettività $\mathscr{A}$, che si ricorda essere la matrice i cui elementi a_{ij} sono pari al numero di link che connettono i nodi i e j (nel caso in cui i link non possano ripetersi, pari ad 1 se i nodi i e j sono connessi e 0 altrimenti). Oltre a detta matrice definiamo la matrice $\mathscr{P}$ i cui elementi, p_{ij}, denotano il numero aspettato di link, che collegano il nodo i a quello j, nel caso in cui il network fosse di tipo random. Si definisce ulteriormente la funzione g_i che associa ad ogni nodo i la comunità a cui il nodo stesso appartiene. Detto ciò è possibile definire la funzione Q come:

$$Q = \frac{1}{2m} \sum_i \sum_j (a_{ij} - p_{ij}) \delta(g_i, g_j) \tag{1.5.1}$$

dove la delta è la classica delta di Dirac, ed m rappresenta il numero totale di link nel grafo. Il fattore $1/2m$ è posto in maniera del tutto convenzionale e non è una necessità nella definizione della modularità in quanto rappresenta una costante del grafo.

Bisogna dunque cercare di specificare la forma che la matrice $\mathscr{P}$ deve assumere in quanto non è stata, al momento, ancora esplicitata. Essa deve sottostare ad alcuni requisiti dettati dalla procedura di definizione della modularità stessa. Come detto, il numero totale di link del network random deve essere pari al numero di link totali del network reale di partenza per cui si avrà la restrizione:

$$\sum_{ij} a_{ij} \;=\; \sum_{ij} p_{ij} \;=\; 2m. \tag{1.5.2}$$

Inoltre, si richiede che il grado di ogni nodo debba essere uguale nel network reale di partenza e nel network random, per cui:

$$\sum_{j} p_{ij} \;=\; k_i. \tag{1.5.3}$$

Ulteriore osservazione può essere posta relativamente alla probabilità che un link si connetta al nodo i. Essa dovrà dipendere esclusivamente dal grado di connettività k_i del nodo i stesso. Dunque la probabilità che due nodi si connettano è data dalla probabilità che un link tocchi sia il nodo i sia quello j, per cui si avrà:

$$p_{ij} \;=\; f(k_i)f(k_j). \tag{1.5.4}$$

Combinando i risultati delle espressioni (1.5.3) e (1.5.4) si ottiene:

$$\sum_{j} p_{ij} \;=\; f(k_i)\sum_{j} f(k_j) \;=\; k_i. \tag{1.5.5}$$

Non sapendo la forma esplicita della funzione $f(k_i)$ assumiamo, basandoci sull'osservazione posta in (1.5.5) e ricordandoci che $\sum_j f(k_j)$ è una costante, che $f(k_i) = Ck_i$, per cui la relazione (1.5.4) diventa:

$$p_{ij} \;=\; C^2 k_i k_j. \tag{1.5.6}$$

Combinando questo risultato con la relazione (1.5.2) si ottiene:

$$\begin{aligned} \sum_{ij} p_{ij} \;&=\; \sum_{ij} C^2 k_i k_j = 2m \\ C^2 \sum_i k_i \sum_j k_j &= 2m \\ C^2 4m^2 &= 2m \\ C^2 &= \frac{1}{2m} \end{aligned} \tag{1.5.7}$$

ove è stato sfruttato il fatto che $\sum_i k_i = 2m$. Perciò, sostituendo il risultato ottenuto in (1.5.7) con la restrizione posta in (1.5.4) sulla forma di $f(k_i)$, si ottiene:

$$p_{ij} \;=\; \frac{k_i k_j}{2m}. \tag{1.5.8}$$

Abbiamo così ottenuto la forma esplicita dei termini della matrice $\mathscr{P}$ che possiamo dunque sostituire in (1.5.1), ottenendo l'usuale forma del coefficiente di modularità presente in letteratura:

$$Q = \frac{1}{2m} \sum_i \sum_j \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(g_i, g_j). \tag{1.5.9}$$

Da questa espressione si definisce la matrice di modularità $\mathscr{B}$ i cui termini b_{ij} risultano pari a:

$$b_{ij} \;=\; a_{ij} - \frac{k_i k_j}{2m}. \tag{1.5.10}$$

Capitolo 2

La Teoria dell'Informazione

Nello studio della relazione tra Teoria Statistica dei Gas e Termodinamica, Boltzmann, partendo dall'estensività della variabile termodinamica entropia e dall'indipendenza stocastica dei microstati di due sistemi posti a contatto termico, pervenne alla relazione:

$$S = k_B \ln \Omega$$

dove k_B rappresenta la costante di Boltzmann e Ω rappresenta il numero di microstati compatibili con le variabili macroscopiche misurate per il sistema, di cui stiamo calcolando l'entropia. Si può osservare che detta formula è del tutto equivalente alla relazione (di Gibbs):

$$S = -k_B \sum_i f_i \ln f_i$$

dove le f_i rappresentano le probabilità di osservare un microstato compatibile con l'i-esimo stato macroscopico (stiamo dividendo l'insieme Ω in i sottoinsiemi disgiunti).

Tramite queste equazioni si può osservare che l'entropia e l'ordine statistico vanno di pari passo. Premettiamo che all'aumentare dei microstati compatibili l'entropia aumenta e, viceversa, al diminuire dei microstati compatibili, l'entropia diminuisce. Ciò posto si osserva che quanto più piccolo è il numero di microstati compatibili tanto maggiore sarà l'informazione relativa al sistema oggetto di studio. Infatti, selezionare microstati con fissate

proprietà escluderà tutti quelli privi delle stesse riducendo il numero di microstati scelti; per esempio, dire che i microstati compatibili non devono avere una proprietà A fa diminuire il numero di essi e fornisce informazione (non soddisfano la proprietà A). Un esempio concettuale può essere dato da un vaso che si rompe. I cocci di un vaso rotto possono disporsi in svariati modi compatibili, ma, partendo da essi, non si può capire il disegno che vi era su di esso. Ciò implica, quindi, minore informazione sul vaso. Il vaso intero avrà un solo modo compatibile, cioè se stesso, e in tal caso sarà facile vedere qual'è il disegno su di esso. Ciò implicherà maggiore informazione. Quest'idea relativa alla stretta correlazione tra informazione ed entropia ha portato ad estendere l'utilizzo di quest'ultima anche fuori dal campo termodinamico.

Lo sviluppo delle telecomunicazioni e della trasmissione dei dati ha posto sin dagli albori domande circa la massima semplificazione dello scambio di quest'ultimi e, dunque, questioni relative alla minimizzazione delle dimensioni del pacchetto di informazione da scambiare. È opinione comune far coincidere la nascita della teoria dell'informazione con la pubblicazione nel 1948 dell'articolo "A matematical Theory of Communication" [Sh48] ad opera di Claude E. Shannon. In questo articolo viene definita l'*entropia di informazione*, così chiamata da Shannon sotto consiglio di John von Neumann[1]. Questa quantità rappresenta il limite inferiore di compressione del pacchetto dei dati da trasmettere, senza che vi sia alcuna perdita di informazione, e risulta strettamente connessa all'entropia definita in ambito termodinamico.

2.1 L'Entropia d'Informazione

Introduciamo il concetto di entropia come misura dell'incertezza di una variabile aleatoria. Definiamo una variabile aleatoria X con dominio $\mathscr{X}$ e

[1]La mia più grande preoccupazione era come chiamarla. Quando discussi della cosa con John Von Neumann, lui ebbe un'idea migliore. Mi disse che avrei dovuto chiamarla entropia, per due motivi: Innanzitutto, la tua funzione d'incertezza è già nota nella meccanica statistica con quel nome. [...]. In secondo luogo, e più significativamente, nessuno sa cosa sia con certezza l'entropia, così in una discussione sarai sempre in vantaggio.

distribuzione di probabilità data dalla funzione $p(x) = Pr\{X = x\}$ con $x \in \mathscr{X}$.

Definizione 2.1 (Entropia di informazione)**.** L'entropia $H(X)$ della variabile aleatoria X è definita come:

$$H(X) = -\sum_{x \in \mathscr{H}} p(x) \log_2 p(x) \tag{2.1.1}$$

La definizione di entropia di informazione suggerisce che la stessa possa essere vista come funzionale della probabilità. Il logaritmo è calcolato in base 2 in maniera tale da poter esprimere l'entropia in *bit*. Ovviamente non è escluso il possibile utilizzo di basi diverse ottenendo una relazione analoga a meno di un fattore moltiplicativo. A seconda della base utilizzata misureremo l'entropia in *nats* (base e), *trits* (base 3) o *Hartleys* (base 10). Risulterà conveniente utilizzare il logaritmo in base 2 in modo tale da misurare l'entropia in *bit*, cosa che faremo ora in avanti. Ulteriore osservazione può essere posta basandoci sulla continuità di $p(x) \log p(x)$: porremo che $0 \log 0 = 0$. Data la definizione di entropia, cerchiamo di capirne il funzionamento tramite un esempio concettuale, mostrando che la definizione stessa di entropia di informazione deriva dalla volontà di calcolare in media quale dev'essere la lunghezza più piccola possibile della descrizione di una variabile aleatoria per conoscerne il comportamento.

Esempio 2.1. Supponiamo di considerare una variabile aleatoria che assuma valori:

$$X = \begin{cases} a & \text{con probabilità } \frac{1}{2} \\ b & \text{con probabilità } \frac{1}{4} \\ c & \text{con probabilità } \frac{1}{8} \\ d & \text{con probabilità } \frac{1}{8}. \end{cases}$$

L'entropia di detta variabile risulterà essere pari a:

$$H(X) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{8}\log_2\frac{1}{8} = \frac{7}{4}\,bits.$$

Supponiamo ora di voler stabilire il valore di X con il minor numero possibile di domande binarie. Un esempio può esser dato dall'effettuare come prima

domanda: ''X è uguale ad a?''. Ciò divide i casi in due. Se la risposta non fosse affermativa passeremmo alla domanda: ''X è uguale ad b?''. In caso di ulteriore risposta negativa passeremmo alla domanda: ''X è uguale ad c?''. Con questo si chiuderebbe il cerchio delle domande utili per sapere il valore di X in qualsiasi caso. Il valore dell'entropia ci dice che il numero minimo di domande che si possono effettuare coincide con 1.75. Ovviamente non si può avere un numero non intero di domande binarie. Infatti, è possibile dimostrare che detto numero deve essere interno all'intervallo $[H(X), H(X)+1]$. Come ultima osservazione, notiamo che se X assumesse uno dei quattro valori con probabilità 1, risulterà che $H(X)$ varrà 0. Questo equivale a dire che non avremo bisogno di informazione (domande binarie da fare in tal caso) per sapere che valore la variabile aleatoria avrà assunto, dato che sappiamo a priori il risultato dell'esperimento.

Ricordiamo ciò che abbiamo precedentemente accennato, cioè che l'entropia d'informazione risulta essere un funzionale della probabilità. Questo implica che l'entropia e, dunque, l'informazione relativa ad una variabile aleatoria, dipende esclusivamente dalla distribuzione di probabilità utilizzata per descrivere la variabile aleatoria e cambia, cambiando la distribuzione utilizzata. Ciò risulta abbastanza comprensibile, in quanto una previsione su un evento dipende esclusivamente dalla probabilità che esso si verifichi.

Ciò detto, risulterà conveniente estendere il concetto di entropia anche a casi più generali, ossia casi in cui non sarà solo una la variabile aleatoria da tenere in conto, ma più di una. Il caso più banale risulta essere quello con due variabili aleatorie.

Definizione 2.2 (Entropia congiunta)**.** Consideriamo una coppia di variabili aleatorie (X,Y), dove $x \in \mathscr{X}$ e $y \in \mathscr{Y}$, e sia data la loro probabilità congiunta pari a $p(x,y)$. L'entropia congiunta $H(X,Y)$ sarà data dalla relazione:

$$H(X,Y) = -\sum_{x\in\mathscr{H}}\sum_{y\in\mathscr{Y}} p(x,y)\,\log_2 p(x,y) \tag{2.1.2}$$

Detta definizione può essere estesa anche a casi più generali applicando la stessa procedura, ma considerando un numero n a piacere di variabili aleatorie. Com'è facile osservare, risulta che la definizione appena data non

aggiunge nulla di più alla definizione 2.1, in quanto si può notare che è del tutto equivalente, per ottenere l'equazione (2.1.2), usare l'equazione (2.1.1) in luogo di uno scambio della variabile aleatoria X con il vettore aleatorio (X, Y). A partire da detta definizione sarà facile definire anche l'entropia condizionata sfruttando le proprietà della teoria della probabilità.

Definizione 2.3 (Entropia condizionata)**.** Consideriamo la coppia di variabili aleatorie $X \in \mathscr{X}$ e $Y \in \mathscr{Y}$ e la loro probabilità congiunta $p(x, y)$, introdotta nella definizione 2.2. Allora l'entropia condizionata $H(Y|X)$ sarà data da:

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathscr{H}} p(x)\, H(Y|X = x) \\ &= -\sum_{x \in \mathscr{H}} p(x) \sum_{y \in \mathscr{Y}} p(y|x)\, \log_2 p(y|x) \\ &= -\sum_{x \in \mathscr{H}} \sum_{y \in \mathscr{Y}} p(x, y)\, \log_2 p(y|x) \\ &= -E_{p(x,y)} \left[\log_2 p(Y|X)\right]. \end{aligned} \tag{2.1.3}$$

Il significato delle definizioni appena date è presto detto e deriva strettamente dalla teoria della probabilità. L'entropia congiunta risulta essere la minima lunghezza (in bit in questo caso) dell' informazione da trasmettere per sapere con certezza il valore assunto dalla coppia di variabili X e Y. L'entropia condizionata, invece, rappresenta la minima lunghezza dell'informazione da trasmettere per venire a conoscenza del valore assunto dalla variabile Y, premesso che il valore che X assume sia già conosciuto. Possiamo osservare che questi due tipi di entropia appena definiti possono esser messi in relazione attraverso la *regola della catena*.

Teorema 2.1 (Regola della catena)**.** *Date due variabili aleatorie X e Y con $x \in \mathscr{X}$ e $y \in \mathscr{Y}$ caratterizzate da probabilità $p(x)$ e $p(y)$ rispettivamente e da probabilità congiunta $p(x, y)$ avremo:*

$$H(X, Y) = H(X) + H(Y|X) \tag{2.1.4}$$

Dimostrazione.

$$\begin{aligned} H(Y,X) &= -\sum_{x\in\mathscr{H}}\sum_{y\in\mathscr{Y}} p(x,y)\,\log_2 p(x,y) \\ &= -\sum_{x\in\mathscr{H}}\sum_{y\in\mathscr{Y}} p(x,y)\,\log_2 p(x)p(y|x) \\ &= -\sum_{x\in\mathscr{H}}\sum_{y\in\mathscr{Y}} p(x,y)\,\log_2 p(x) - \sum_{x\in\mathscr{H}}\sum_{y\in\mathscr{Y}} p(x,y)\,\log_2 p(y|x) \\ &= -\sum_{x\in\mathscr{H}} p(x)\,\log_2 p(x) - \sum_{x\in\mathscr{H}}\sum_{y\in\mathscr{Y}} p(x,y)\,\log_2 p(y|x) \\ &= H(X) + H(Y|X). \end{aligned}$$

□

La relazione (2.1.4) mette in luce che l'entropia condizionata $H(Y|X)$ altro non è che l'informazione che X non fornisce su Y e che deve essere pertanto aggiunta ad $H(X)$ per ottenere una completa informazione previsionale congiunta di X ed Y, pari all'entropia d'informazione congiunta $H(Y,X)$.

Le definizioni finora date sono state ottenute esplicitando una base 2 per il logaritmo in modo tale che l'entropia possa essere misurata in *bits*. Naturalmente, come già accennato, si possono cambiare base ed unità di misura per l'entropia, aggiungendo un fattore moltiplicativo all'entropia stessa.

2.1.1 Approccio assiomatico

Oltre al metodo già osservato, per ottenere una definizione dell'entropia di informazione, esiste anche un approccio assiomatico molto più simile al metodo utilizzato da Shannon nel suo articolo [Sh48].

Vogliamo sviluppare una misura dell'informazione (la chiameremo $I(p)$) necessaria per sapere se un evento, che avviene con probabilità p, sia accaduto o meno. Detta quantità è anche conosciuta col termine di *autoinformazione*.

Non ci interessa la specificità dell'evento in sé, ma solo se l'evento stesso viene osservato o meno.
Ciò che richiediamo relativamente a questa misura di informazione è che:

1. Detta misura debba essere non negativa: $I(p) \geq 0$ (non ha senso parlare di numero di bit negativo);

2. Se un evento possiede probabilità 1, non abbiamo bisogno di informazioni relativamente alla possibilità che l'evento accada: $I(p) = 0$ (so di certo che l'evento è avvenuto o avverrà, non ho bisogno di informazioni per saperlo);

3. Se si studiano due eventi indipendenti, allora l'informazione di cui abbiamo bisogno relativamente alla possibilità che gli eventi accadano è: $I(p_1 * p_2) = I(p_1) + I(p_2)$ (se due eventi sono indipendenti, è ovvio che ho bisogno di tutta l'informazione relativa ad entrambi per venire a conoscenza con certezza dei risultati dei due esperimenti indipendenti);

4. Vogliamo che la nostra misura dell'informazione risulti essere una funzione continua della probabilità (piccole variazioni nella probabilità portano a piccole variazioni nella misura di informazione).

Cerchiamo la formulazione analitica di $I(p)$. La dimostrazione di ciò è abbastanza semplice:

- $I(p^2)$ è uguale per la proprietà (3) a $I(p) + I(p) \Rightarrow I(p^2) = 2\,I(p)$;
- per induzione ottengo che $I(p^n) = n\,I(p)$ ove $n \in \mathbb{N}$;
- $I(p) = I((p^{\frac{1}{m}})^m) = m\,I(p^{\frac{1}{m}})$ dunque $I(p^{\frac{1}{m}}) = \frac{1}{m}\,I(p)$ e quindi in generale $I(p^{\frac{n}{m}}) = \frac{n}{m}\,I(p)$;
- per la proprietà di continuità, per $0 < p \leq 1$ e per $a > 0$ con $a \in \mathbb{R}$ otteniamo che $I(p^a) = a\,I(p)$.

Da questa breve discussione possiamo concludere che l'autoinformazione $I(p)$ soddisfa le proprietà tipiche della funzione logaritmo, perciò:

$$I(p) = -\log_b(p) = \log_b\left(\frac{1}{p}\right).$$

In questa definizione abbiamo usato una generica base b. Come è facile osservare, il risultato è del tutto analogo a quello ottenuto nella definizione 2.1 (eq. 2.1.1).

2.2 Entropia Relativa e Mutua Informazione

Come osservato, l'entropia di una variabile aleatoria è la misura della quantità di informazione minima da possedere per descrivere, in maniera previsionale, la variabile aleatoria stessa. Detto ciò, si definiscono due ulteriori quantità.

La prima è l'entropia relativa $D(p||q)$, la quale risulta essere la misura della distanza tra due distribuzioni p e q utilizzate per descrivere la stessa variabile aleatoria. Essa misura l'inefficienza nella strategia di utilizzo di una distribuzione q al posto di una distribuzione p per descrivere la variabile aleatoria X. Volendo esprimere il concetto attraverso un esempio, avremmo che, se utilizzassimo la distribuzione p per descrivere la variabile aleatoria oggetto della nostra trasmissione di dati, potremmo ottenere, calcolando l'entropia $H(p)$ attraverso l'equazione (2.1.1), il valore della quantità minima di dati da inviare per avere informazione previsionale sulla variabile aleatoria stessa. Se, invece della distribuzione p, avessimo utilizzato per descrivere la variabile aleatoria X un'altra distribuzione, per esempio q, il numero di bit necessari per descrivere previsionalmente la variabile stessa sarebbe stato in media pari a $H(p) + D(p||q)$.

Definizione 2.4 (Entropia relativa)**.** L'*entropia relativa* o *distanza di Kullback-Leibler* tra due distribuzioni di probabilità p e q è definita come:

$$\begin{aligned} D(p||q) &= \sum_{x \in \mathscr{H}} p(x) \log_2 \left(\frac{p(x)}{q(x)} \right) \\ &= -E_p \log_2 \left(\frac{p(x)}{q(x)} \right). \end{aligned} \tag{2.2.1}$$

Sfruttando la continuità di $p \log_2\left(\frac{p}{q}\right)$, assumeremo che $0 * \log_2\left(\frac{0}{q}\right) = 0$ e che $p * \log_2\left(\frac{p}{0}\right) = \infty$. Si può osservare che $D(p||q)$ è sempre diversa da 0 tranne nel caso $p = q$. Inoltre l'entropia relativa non è una vera e propria distanza in quanto non soddisfa le proprietà che una distanza deve verificare; ciononostante continueremo a pensarla come ad una distanza.

Introduciamo adesso la mutua informazione che risulta essere una misura della quantità di informazione, relativa ad una variabile aleatoria X, che un'altra variabile aleatoria Y contiene.

Definizione 2.5 (Mutua informazione)**.** Consideriamo due variabili aleatorie $X \in \mathscr{X}$ e $Y \in \mathscr{Y}$ con probabilità rispettivamente $p(x)$ e $p(y)$ e con distribuzione di probabilità congiunta $p(x,y)$. La *mutua informazione* $I(X;Y)$ è l'entropia relativa (ovvero la distanza di Kullback Leibler) tra la distribuzione congiunta e la distribuzione prodotto, cioè:

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathscr{H}} \sum_{y\in\mathscr{Y}} p(x,y)\, \log_2\left(\frac{p(x,y)}{p(x)p(y)}\right) \\ &= D(p(x,y)||p(x)p(y)). \end{aligned} \tag{2.2.2}$$

Dunque, la mutua informazione misura l'informazione aggiuntiva da considerare per descrivere previsionalmente il comportamento congiunto di X ed Y, nel caso supponessimo che le due variabili aleatorie siano indipendenti. È inoltre banalmente osservabile dalla definizione 2.5 che $I(X;Y)$ risulta simmetrica rispetto allo scambio di variabili. Il teorema seguente mette in luce una stretta relazione tra la mutua informazione e l'entropia.

Teorema 2.2 (Mutua informazione ed entropia)**.**

$$I(X;Y) = H(X) - H(X|Y) \tag{2.2.3}$$

$$I(X;Y) = H(Y) - H(Y|X) \tag{2.2.4}$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{2.2.5}$$

$$I(X;Y) = I(Y;X) \tag{2.2.6}$$

$$I(X;X) = H(X) \tag{2.2.7}$$

Questi risultati vengono tradotti visivamente attraverso il diagramma di Venn rappresentato in figura 2.1.

Cerchiamo di dare un significato alle formule nel teorema 2.2. L'equazione (2.2.3) stabilisce che $I(X;Y)$ rappresenta la quantità media di informazione, che Y fornisce relativamente ad X. Infatti, (2.2.3) stabilisce che $H(X)$, cioè l'informazione da avere per descrivere previsionalmente X, è pari alla somma di $H(X|Y)$, quantità d'informazione che Y non fornisce su X, più $I(X;Y)$. Pertanto quest'ultima quantità rappresenterà l'informazione che Y fornisce su X. La relazione (2.2.3) si ottiene facilmente, sostituendo la definizione d'entropia d'informazione (2.1.1) e la definizione di entropia

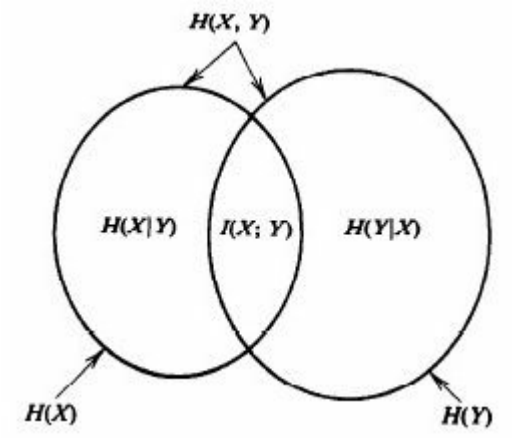


Figura 2.1: Relazione tra mutua informazione ed entropia.

d'informazione condizionata (2.1.3) all'interno della definizione di mutua informazione (2.2.2). L'equazione (2.2.4) deriva dalla simmetria della (2.2.3). Queste due relazioni stabiliscono che la quantità media di informazione che Y fornisce su X è uguale alla quantità media di informazione che X fornisce su Y, ossia più semplicemente X dice di Y quanto Y dice di X. L'equazione (2.2.5) deriva dalla definizione di entropia congiunta (def. 2.2, pag. 22) e viene sintetizzata visivamente dalla figura 2.1. Traspare da essa che l'informazione totale da possedere per descrivere previsionalmente sia X che Y è pari alla somma dell'informazione da possedere per descrivere X e dell'informazione da possedere per descrivere Y meno l'informazione che X fornisce di Y o viceversa (questo perché altrimenti detta quantità verrebbe considerata due volte). L'equazione (2.2.6) descrive la simmetria della mutua informazione che traspare dalla relazione che la definisce (2.2.3) e ristabilisce quanto osservato nella descrizione delle relazioni (2.2.3) e (2.2.4). L'equazione (2.2.7) afferma, invece, che la mutua informazione tra una variabile aleatoria e se stessa è l'entropia della variabile aleatoria, cioè la quantità media di informazione che la variabile aleatoria X fornisce su se stessa è uguale ad $H(X)$ (banalmente stiamo dicendo che X dà informazione totale relativamente a se stessa, cioè ci fornisce tutta l'informazione necessaria per prevederla). Questa è la ragione per la quale l'entropia è anche chiamata *self-information*.

Cerchiamo ora di estendere i risultati finora ottenuti tentando di ren-

derli più generali possibile. Una prima osservazione è che l'entropia di una collezione di variabili aleatorie è pari alla somma delle entropie condizionate.

Teorema 2.3 (Regola della catena generalizzata)**.** *Sia $X_1, X_2, \ldots, X_n$ una collezione di variabili aleatorie la cui probabilità congiunta sia data dalla funzione $p(x_1, x_2, \ldots, x_n)$. L'entropia congiunta di dette variabili sarà pari a:*

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1). \tag{2.2.8}$$

Estendo ora il concetto di mutua informazione al caso del calcolo della quantità media di informazione in meno richiesta per descrivere X, qualora non solo un'unica variabile Y sia conosciuta, ma le variabili conosciute siano X e Y.

Definizione 2.6 (Mutua informazione condizionata)**.** La mutua informazione condizionata di due variabili aleatorie X e Y, eliminata la variabile aleatoria Z, è definita da:

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\ &= E_{p(x,y,z)} \log_2 \left(\frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)} \right). \end{aligned} \tag{2.2.9}$$

Detta definizione, nel contesto in cui nel sistema siano presenti tre variabili aleatorie, risulta essere pari all'informazione che solo Y fornisce su X (fig. 2.2).

Si nota che anche per la mutua informazione esiste una regola della catena.

Teorema 2.4 (Regola della catena generalizzata per l'informazione)**.** *Sia $X_1, X_2, \ldots, X_n$ una collezione di variabili aleatorie la cui probabilità congiunta sia data dalla funzione $p(x_1, x_2, \ldots, x_n)$ e sia Y una ulteriore variabile aleatoria. La mutua informazione di dette variabili sarà pari a:*

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, \ldots, X_1). \tag{2.2.10}$$

Dopo aver generalizzato i concetti di entropia e mutua informazione, non ci resta che generalizzare il concetto di entropia relativa.

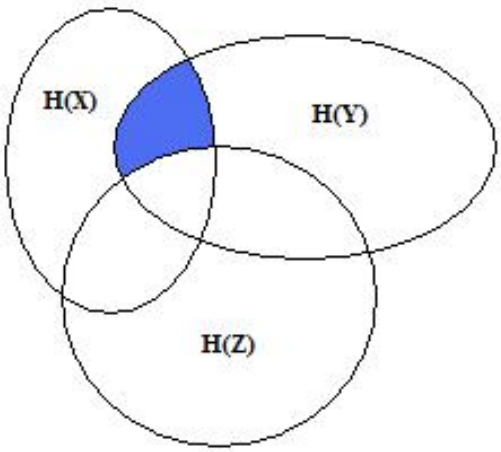


Figura 2.2: La zona blu rappresenta visivamente la mutua informazione condizionata $I(X;Y|Z)$, ossia l'informazione che solo Y fornisce su X.

Definizione 2.7 (Entropia relativa condizionata)**.** L'entropia relativa condizionata $D(p(y|x)||q(y|x))$ è pari alla media delle entropie relative (distanze di Kullback-Leibler) tra le distribuzioni di probabilità condizionate $p(y|x)$ e $q(y|x)$ calcolata sulla distribuzione $p(x)$ della variabile aleatoria condizionante:

$$\begin{aligned} D(p(y|x)||q(y|x)) &= \sum_x p(x) \sum_y p(y|x) \log_2\left(\frac{p(y|x)}{q(y|x)}\right) \\ &= E_{p(x,y)} \log_2\left(\frac{p(Y|X)}{q(Y|X)}\right). \end{aligned} \tag{2.2.11}$$

Oltre che per l'entropia e per la mutua informazione, posso vedere che anche per l'entropia relativa esiste una regola della catena.

Teorema 2.5 (Regola della catena per l'entropia relativa)**.** *L'entropia relativa tra due distribuzioni di probabilità congiunte di una coppia di variabili aleatorie X ed Y, $p(x,y)$ e $p(x,y)$, può essere calcolata come la somma dell'entropia relativa tra le marginali di dette distribuzioni e dell'entropia relativa condizionata tra le distribuzioni condizionate calcolate a partire da $p(x,y)$ e $p(x,y)$, cioè:*

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x)). \tag{2.2.12}$$

Sfruttando la disuguaglianza di Jensen si ottengono alcune disuguaglianze che ampliano ulteriormente lo spettro delle proprietà viste finora.

Teorema 2.6 (Disuguaglianza dell'informazione)**.** *Siano $p(x)$ e $q(x)$ ($x \in \mathscr{X}$) due distribuzioni di probabilità. Allora l'entropia relativa tra le due distribuzioni soddisfarrà la relazione:*

$$D(p||q) \geq 0 \tag{2.2.13}$$

dove l'uguaglianza vale se e solo se:

$$p(x) = q(x) \qquad \forall x \in \mathscr{H}$$

Da questo teorema derivano alcuni corollari che elenchiamo di seguito.

Corollario (Non-negatività della mutua informazione)**.** *Per qualsiasi coppia di variabili aleatorie X ed Y:*

$$I(X;Y) \geq 0 \tag{2.2.14}$$

con uguaglianza se e solo se X ed Y sono indipendenti.

Corollario (Non-negatività dell'entropia relativa condizionata)**.** *Siano $p(y|x)$ e $q(y|x)$ due distribuzioni di probabilità condizionata, allora è valida la relazione:*

$$D(p(y|x)||q(y|x)) \geq 0 \tag{2.2.15}$$

con uguaglianza se e solo se $p(y|x) = q(y|x)$ per ogni x ed y con $p(x) > 0$.

Corollario (Non-negatività della mutua informazione condizionata)**.** *Per qualsiasi coppia di variabili aleatorie X ed Y e per qualsiasi variabile aleatoria Z:*

$$I(X;Y|Z) \geq 0 \tag{2.2.16}$$

con uguaglianza se e solo se X ed Y sono, esclusi i valori in cui risultano entrambi condizionati con Z, indipendenti.

Altra proprietà da osservare è che la distribuzione di probabilità che massimizza l'entropia $H(X)$ di una variabile aleatoria X, definita sull'insieme $\mathscr{X}$, è la distribuzione uniforme su tale insieme. Detta proprietà può

essere ottenuta in tal maniera: si definisce la distribuzione uniforme su $\mathscr{X}$ della variabile aleatoria X pari a $u(x) = \frac{1}{|\mathscr{X}|}$, ove $|\mathscr{X}|$ rappresenta il numero di elementi interni all'insieme $\mathscr{X}$. Calcoliamo l'entropia relativa tra la distribuzione uniforme appena introdotta e una distribuzione $p(x)$ generica:

$$D(p||u) = \sum p(x) \log_2\left(\frac{p(x)}{u(x)}\right) = \log_2 |\mathscr{H}| - H(X). \tag{2.2.17}$$

Sfruttando l'equazione appena ottenuta e l'equazione (2.2.13), otterremo che l'entropia di informazione $H(X)$ soddisferà la condizione:

$$H(X) \leq \log_2 |\mathscr{H}| \tag{2.2.18}$$

con uguaglianza valida nel solo caso in cui X è distribuito uniformemente su $\mathscr{X}$. Pertanto la distribuzione uniforme massimizzerà l'entropia $H(X)$.

Com'è intuitivamente lecito supporre, è possibile verificare che la conoscenza di un'altra variabile aleatoria può solo ridurre la quantità di informazione da trasmettere per descrivere la variabile aleatoria di partenza. Ciò viene reso più formale dal seguente teorema.

Teorema 2.7 (Il condizionamento riduce l'entropia)**.** *Siano X e Y due variabili aleatorie la cui distribuzione di probabilità congiunta sia $p(x, y)$. Varrà la seguente relazione:*

$$H(X|Y) \leq H(X) \tag{2.2.19}$$

dove l'uguaglianza vale se e solo se X e Y sono indipendenti.

Come ultimo risultato vediamo il limite massimo per l'entropia congiunta di una collezione di variabili aleatorie.

Teorema 2.8 (Limite di indipendenza per l'entropia)**.** *Sia $X_1, X_2, \ldots, X_n$ una collezione di variabili aleatorie la cui distribuzione di probabilità congiunta sia $p(x_1, x_2, \ldots, x_n)$. Varrà allora la seguente relazione:*

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i) \tag{2.2.20}$$

dove l'uguaglianza vale se e solo se le X_i sono indipendenti.

Tutti questi risultati saranno utili per mettere in luce la stretta connessione sussistente tra entropia di informazione e seconda legge della termodinamica.

2.3 Relazione con la seconda Legge della Termodinamica

Il secondo principio della termodinamica, storicamente introdotto tramite i due principi di *Clausius* e *Kelvin-Planck*, ha un'ulteriore formulazione statistica: l'entropia di un sistema isolato risulta essere non decrescente. Cerchiamo di vedere la relazione tra questo principio e l'entropia di informazione. Come già visto, la definizione che Boltzmann ha dato di entropia lega la stessa al numero di microstati compatibili con le variabili mascroscopiche misurate, che il sistema oggetto di studio può assumere. Questa definizione corrisponderebbe alla definizione di entropia d'informazione, se tutti gli stati fossero equiprobabili.

Cerchiamo di osservare la relazione tra entropia d'informazione e secondo principio della termodinamica, facendo vedere come oltre all'entropia fisica anche l'entropia d'informazione aumenta nel tempo. Per fare questo consideriamo un sistema descrivibile come un processo markoviano, le cui transizioni siano governate dalle leggi fisiche cui il sistema obbedisce. Otterremo varie modalità per interpretare la relazione tra l'entropia d'informazione e il secondo principio della termodinamica (ossia l'aumento della entropia d'informazione stessa).

1. *L'entropia relativa $D(\mu_n||\mu'_n)$ decresce con n.* Siano μ_n e μ'_n due distribuzioni di probabilità sullo spazio dei campioni della catena markoviana al tempo n. Siano $\mu(x_n, x_{n+1})$ e $\mu'(x_n, x_{n+1})$ rispettivamente le distribuzioni di probabilità congiunte al tempo n ed $n+1$ per le distribuzioni μ_n e μ'_n. Alla stessa maniera definiamo le distribuzioni di probabilità condizionate $\mu(x_{n+1}|x_n)$ e $\mu'(x_{n+1}|x_n)$, che risulteranno essere le probabilità di transizione sulla catena markoviana. Sfruttando

il teorema 2.5 ottengo la relazione:

$$\begin{aligned} &D(\mu(x_n, x_{n+1})||\mu'(x_n, x_{n+1})) = \\ &\quad = D(\mu(x_n)||\mu'(x_n)) + D(\mu(x_n|x_{n+1})||\mu'(x_n|x_{n+1})) \\ &\quad = D(\mu(x_{n+1})||\mu'(x_{n+1})) + D(\mu(x_{n+1}|x_n)||\mu'(x_{n+1}|x_n)). \end{aligned}$$

Focalizziamo l'attenzione sulle probabilità condizionate $\mu(x_{n+1}|x_n)$ e $\mu'(x_{n+1}|x_n)$. Le distribuzioni μ_n e μ'_n sono relative allo spazio dei campioni in cui la catena markoviana prende valori. Essa non da assolutamente prescrizioni relativamente alla probabilità di transizione nel processo markoviano (che è una caratteristica indipendente dalla distribuzione degli stati che il processo può assumere) per cui risulterà che entrambe le probabilità condizionate saranno uguali alla distribuzione di probabilità di transizione $r(x_n|x_{n+1})$, scorrelata ovviamente dalla distribuzione sullo spazio dei campioni. Ciò detto, otterremo che la probabilità di transizione resterà sempre la stessa indipendentemente dalla scelta della distribuzione μ_n o μ'_n per gli stati che il processo può assumere, per cui:

$$D(\mu(x_n|x_{n+1})||\mu'(x_n|x_{n+1})) = D(\mu(x_{n+1}|x_n)||\mu'(x_{n+1}|x_n)) = 0.$$

Sfruttando la non-negatività dell'entropia relativa condizionata (corollario al teorema 2.6), risulta che:

$$D(\mu_n||\mu'_n) \geq D(\mu_{n+1}||\mu'_{n+1}). \tag{2.3.1}$$

Dunque, sintetizzando, risulta che la distanza tra due distribuzioni di probabilità sullo spazio dei campioni di uno stesso sistema, descritto come un processo markoviano, diminuisce col tempo. Per semplificare, prendiamo in considerazione un sistema di tassazione. Supponiamo che in due paesi A e B vi sia lo stesso sistema di tassazione, ma all'istante iniziale la distribuzione degli stipendi nei due paesi sia diversa (il sistema di tassazione costituisce la catena di Markov). L'osservazione precedente dimostra che col passare del tempo le distribuzioni si

avvicineranno sempre più.

2. *L'entropia relativa $D(\mu_n||\mu)$ tra una distribuzione sullo spazio dei campioni della catena makoviana μ_n calcolata al tempo n e una distribuzione stazionaria μ sulla stessa catena markoviana decresce con n.* Osserviamo la relazione precedente (eq. 2.3.1). In essa la distribuzione μ'_n può essere una qualsiasi distribuzione al tempo n. Sia al tempo n la distribuzione μ'_n pari ad una distribuzione stazionaria μ. Ovviamente, essendo stazionaria, la stessa distribuzione varrà in ogni istante $m \geq n$. Dunque:

$$D(\mu_n||\mu) \geq D(\mu_{n+1}||\mu). \tag{2.3.2}$$

Questa relazione stabilisce che col passare del tempo la distribuzione tende a divenire stazionaria.

3. *L'entropia aumenta se la distribuzione stazionaria è uniforme.* Non è detto che la tendenza a diminuire dell'entropia relativa tra una distribuzione qualsiasi e la distribuzione stazionaria del processo markoviano implichi che l'entropia della variabile aleatoria debba aumentare. Un esempio può essere dato da un processo markoviano che risulta avere una distribuzione di probabilità stazionaria non uniforme. Abbiamo visto che l'entropia d'informazione di un sistema risulta massima quando le probabilità sullo spazio dei campioni risultano uguali per ogni elemento interno a detto insieme, cioè quando la distribuzione di probabilità risulta essere uniforme. Dunque, supponendo che la distribuzione di partenza del sistema in oggetto risulti essere uniforme, utilizzando l'osservazione precedente, avremo che il sistema tenderebbe a distribuirsi, col passare del tempo, in maniera tale da essere descritto tramite la propria distribuzione stazionaria. Essendo però la distribuzione stazionaria non uniforme, ciò implicherebbe una minore entropia d'informazione, che porterebbe ad una diminuzione della stessa col passare del tempo. In tale caso, dunque, l'entropia aumenterebbe invece che diminuire.

Se invece la distribuzione stazionaria del sistema fosse uniforme, allora

l'entropia relativa risulterebbe pari a:

$$D(\mu_n||\mu) = \log_2 |\mathscr{H}| - H(X_n)$$

dove $H(X_n)$ risulta essere l'entropia d'informazione, qualora considerassimo come distribuzione di probabilità per il processo markoviano la distribuzione μ_n, e $\mathscr{H}$ è una costante, pari al numero di possibili valori che il processo markoviano stesso può assumere (la distribuzione uniforme è pari a $u(x) = 1/|\mathscr{H}|$). Dunque per ottenere la diminuzione di $D(\mu_n||\mu)$ all'aumentare di n bisogna avere allo stesso tempo un aumento di $H(X_n)$ all'aumentare di n.

4. *L'entropia condizionata $H(X_n|X_1)$ aumenta con n per un processo markoviano stazionario.* Qualora il processo di Markov fosse stazionario, l'entropia d'informazione $H(X_n)$ risulterebbe costante. Allo stesso tempo l'entropia condizionata $H(X_n|X_1)$ aumenterebbe con n. Infatti risulta:

$$\begin{aligned} H(X_n|X_1) &\geq H(X_n|X_1, X_2) && \text{(Il condizionamento riduce l'entropia)} \\ &= H(X_n|X_2) && \text{(Per la markovianità)} \\ &= H(X_{n-1}|X_1) && \text{(Per la stazionarietà).} \end{aligned}$$

Ciò implicherà che col passare del tempo trascorso, i valori passati del processo markoviano daranno sempre meno informazioni sui valori attuali del processo stesso.

2.4 Il Principio di Massima Entropia

Sappiamo dalla Termodinamica che un sistema nel giungere all'equilibrio tenderà a massimizzare la propria entropia. Cerchiamo di estendere il risultato fisico ad una classe più amplia di casi, ponendoci all'interno della teoria dell'informazione. Abbiamo visto nel paragrafo precedente che l'entropia d'informazione segue il secondo principio della termodinamica, per cui detta quantità, calcolata per una variabile aleatoria, aumenterà col passare del tempo. Detto ciò, quindi, l'entropia d'informazione soddisferà il

principio di massimizzazione per cui l'entropia tenderà a massimizzare il proprio valore. Nell'articolo di Jaynes del 1957 [Ja57], l'autore afferma che lo scopo della teoria dell'informazione è quello di costituire un'utile procedura per creare delle distribuzioni di probabilità relative a oggetti di cui si ha solo una conoscenza non completa, procedura che coincide con il principio di massima entropia.
Consideriamo inizialmente il seguente problema. Si voglia massimizzare l'entropia $h(f)$ considerando tutte le densità di probabilità f per le quali:

1. $f(x) \geq 0$ con uguaglianza al di fuori del supporto S;
2. $\int_S f(x)\,dx = 1$;
3. $\int_S f(x)\,r_i(x)\,dx = \alpha_i$ con $0 \leq i \leq m$.

Dunque f è una generica densità di probabilità costruita su un supporto S tale che abbia come vincolo i momenti $\alpha_1, \alpha_2, \ldots, \alpha_m$. Il nostro problema si traduce in una richiesta di massimizzazione dell'entropia di informazione ($H(f) = -\int f(x) \ln f(x) dx$) sotto i vincoli di normalizzazione e i vincoli relativi ai valori assegnati ai momenti della densità di probabilità. Utilizzeremo il metodo dei moltiplicatori di Lagrange per risolvere la questione. Introduciamo il lagrangiano:

$$L = -\int f(x)\,\ln f(x) dx + \lambda_0 \left(\int f(x) dx - 1 \right) + \sum_{i=1}^{m} \lambda_i \left(\int f(x)\, r_i(x) dx - \alpha_i \right)$$

deriviamo il lagrangiano rispetto ad $f(x)$ per un valore fissato di x:

$$\frac{\partial L}{\partial f(x)} = -\ln f(x) - 1 + \lambda_0 + \sum_{i=1}^{m} \lambda_i r_i(x).$$

Ponendo uguale a zero quest'ultima relazione, otterremo la forma della distribuzione che massimizza l'entropia:

$$f(x) = e^{\lambda_0 - 1 + \sum_{i=1}^{m} \lambda_i r_i(x)} \qquad (x \in S)$$

dove $\lambda_0, \lambda_1, \ldots, \lambda_m$ sono scelti in maniera tale che $f(x)$ soddisfi le restrizioni.

Vediamo tramite un esempio le implicazioni di tale descrizione:

Esempio 2.2. L'esempio è tratto da "Statistical mechanics of money" di Dragulescu [Dr00]. Consideriamo che in un sistema vi sia una quantità finita di moneta pari ad M dollari con un numero fisso di agenti pari ad N. Supponiamo che il sistema economico abbia questo meccanismo: in istanti fissati, un agente seleziona in maniera casuale un altro agente e gli trasferisce un dollaro. Cercheremo la distribuzione di capitale a lungo termine. Ovviamente questo esempio non risulta essere del tutto realistico, in quanto il tasso di crescita sarà pari a zero e ci sarà solo una ridistribuzione del capitale iniziale. La distribuzione a lungo termine può essere scritta a partire da un set di probabilità dato dall'insieme $\{p_i\}$. Sia $\{n_i\}$ il set che descrive gli agenti che hanno capitale pari ad i dollari. Varranno le restrizioni (definendo $p_i = \frac{n_i}{N}$):

$$\sum_i p_i \; i = \frac{M}{N}$$

e

$$\sum_i p_i = 1.$$

Applicando il metodo dei moltiplicatori di Lagrange, abbiamo:

$$L = -\sum_i p_i \ln p_i - \lambda\left(\sum_i p_i \; i - \frac{M}{N}\right) - \mu\left(\sum_i p_i - 1\right)$$

dalla quale otteniamo, derivando rispetto alla probabilità p_i:

$$\frac{\partial L}{\partial p_i} = -\ln p_i - 1 - \lambda_i - \mu = 0.$$

Risolvendo e calcolando i moltiplicatori, otterremo la distribuzione:

$$p_i = \frac{1}{T} e^{-\frac{i}{T}}$$

ove $T = \frac{M}{N}$, ovvero la quantità media di capitale per agente. Il risultato risulta essere una distribuzione di Boltzmann-Gibbs nella quale possiamo pensare alla quantità media di capitale come alla temperatura. Per questo motivo un'economia di questa natura viene anche chiamata economia di Boltzmann.

Esempio 2.3. Consideriamo un esempio descritto nell'articolo "A Statisti-

cal Equilibrium Model of Wealth Distribution" di Milakovic [Mi01]. Supponiamo che un sistema economico sia composto da un numero a di agenti e che ognuno di essi abbia al tempo t un capitale pari a $w(a,t)$. Il capitale totale cui il sistema economico può usufruire al tempo t risulterà pari a: $W(t) = \sum_a w(a,t)$. Per semplificare l'esempio, supponiamo che i possibili valori di capitale che un agente può avere costituiscano un set discreto $\{w_i\}$ con probabilità associate ad ogni valore pari a $\{p_i\}$. Cerchiamo di calcolare le $\{p_i\}$. Cominciamo col considerare delle variabili globali che caratterizzino le caratteristiche degli elementi del sistema. Un esempio può essere considerare i tassi di crescita economici: $R_i = w_i(t_1)/w_i(t_0)$ valido per l'i-esimo valore di capitale. Il tasso di crescita medio viene calcolato con la media geometrica ponderata la quale, utilizzando la relazione tra media geometrica e media aritmetica dei logaritmi, diventa:

$$\sum_i p_i \ \ln(R_i) = \ln R$$

relazione che insieme alla relazione:

$$\sum_i p_i = 1$$

rappresenta i vincoli del nostro problema di massimizzazione. Definiamo dunque la lagrangiana del problema:

$$L = -\sum_i p_i \ \ln p_i + \lambda\left(\sum_i p_i \ \ln(R_i) - \ln R\right) + \mu\left(\sum_i p_i - 1\right)$$

Derivando rispetto alle probabilità, otterremo:

$$\frac{\partial L}{\partial p_i} = -\ln p_i - 1 + \lambda \ \ln(R_i) + \mu = 0$$

Risolvendo e calcolando i moltiplicatori, avremo la distribuzione:

$$p_i = \frac{R_i^\lambda}{\sum_i R_i^\lambda}$$

ottenendo così una distribuzione a legge di potenza.

Capitolo 3

Causalità

Con termine causalità ci si riferisce al particolare rapporto sussistente tra due eventi o enti dei quali uno, chiamato *effetto*, dipende dall'altro, chiamato *causa*. Detta dipendenza può essere estesa ad eventualità, in cui da più cause discendano uno o più effetti. Lo studio del collegamento tra causa ed effetto è da ritenersi fondamentale nello sviluppo della ricerca in quasi tutti i campi del sapere scientifico ed ha sempre impegnato le menti degli uomini più brillanti da Aristotele fino agli scienziati moderni.
Il concetto di causa è stato sviscerato sin dagli albori della storia del pensiero e si fonda sull'esperienza umana della produzione di effetti o fatti per azione volontaria. Un primo esempio di caratterizzazione del concetto di causa venne dato da Aristotele, il quale lo divise in 4 categorie: *materiale, formale, efficiente e finale*[1]. Risulta facile osservare che Aristotele, nell'operazione di caratterizzazione del concetto di causa, si sia ricondotto ad una dimensione tipicamente umana, ossia l'idea che ogni ente sia costituito da qualcosa, abbia una data forma, abbia un creatore e serva a qualcosa. Con l'evoluzione del pensiero gli studiosi spostarono preminentemente la loro idea di filosofia sullo studio della causa efficiente e finale. Con la nascita della scienza la

[1] Le cause secondo Aristotele possono essere divise in 4 tipi:

- *Causa materiale*: indica la materia di cui è fatto un ente;
- *Causa formale*: indica la forma o l'essenza di un ente;
- *Causa efficiente*: indica ciò che ha prodotto;
- *Causa finale*: indica il fine che un ente deve realizzare con la sua esistenza.

causa efficiente passò in primo piano e finì col diventare l'unico significato comunemente inteso quando si parla di causa. È con l'evoluzione dell'idea di causa, la quale portò, come detto, ad uno sbilanciamento in favore della causa efficiente, che si giunge all'impostazione kantiana del principiò di causalità, la quale tratta quest'ultimo come legge di connessione tra causa ed effetto, fonte di tutti i mutamenti osservabili in natura.

3.1 La causalità in Fisica

È indispensabile, quando si effettua una trattazione fisica, stabilire i rapporti di causa-effetto definendo cosa in un sistema risulti causa e cosa risulti effetto. Ad esempio nel caso della meccanica classica, più nello specifico nel caso della seconda legge della dinamica, risulta possibile stabilire che l'accelerazione cui è soggetto un corpo è l'effetto della forza agente sullo stesso che risulta esserne la causa. Il rapporto causa-effetto stabilito con la seconda legge farebbe supporre che esso valga sempre e più in generale farebbe supporre che qualsiasi sia il rapporto causa-effetto trovato esso valga sempre. Così non è. Infatti i rapporti di causa-effetto risultano essere relativi alla teoria fisica utilizzata. Considerando la relatività generale, l'effetto della causa "forza" non risulta essere più l'accelerazione ma lo scostamento dalla geodetica.

Oltre all'individuazione della causa e dell'effetto e all'osservazione di come causa ed effetto cambino a seconda della teoria osservata, risulta considerevolmente importante stabilire la posizione relativa che causa ed effetto devono avere all'interno della scala temporale. La restrizione temporale relativa alla causa di un effetto particolare, all'interno della meccanica classica, risulta essere la richiesta che essa preceda temporalmente l'effetto. Nel contesto della teoria relativistica di Einstein detto vincolo risulta più particolare: l'evento causa (le sue coordinate spazio-temporali) deve trovarsi all'interno del cono-luce del passato dell'evento effetto; ossia non è possibile che un qualsiasi evento X in una posizione e ad un tempo a scelta possa influenzare un evento Y, dunque esserene causa. Ciò sarà possibile solo se X si troverà all'interno del cono-luce passato di Y.

Forse, però, il problema fisico più importante collegato al principio di causalità è la maniera nella quale la causa determina l'effetto. La posizione iniziale, figlia di tempi in cui gli esperimenti di dinamica mettevano in risalto un contatto tra causa (oggetto causante la forza) ed effetto (oggetto che subiva la variazione di moto, "accelerazione", dovuta alla forza), fu un disinteressamento totale relativamente a quali fossero gli enti mediatori tra causa ed effetto, o meglio cosa permettesse, partendo dalla causa, di ottenere l'effetto. Con l'avvento della teoria newtoniana sorse il problema di comprendere come un corpo attraesse un altro nello spazio vuoto. Quando Newton delineò le basi della teoria della gravitazione universale non riuscì a comprendere come il Sole potesse, senza mezzi mediatori, attrarre a se la Terra. Quest'idea, molto semplice da comprendere ed immaginare al giorno d'oggi, era del tutto rivoluzionaria nel periodo di massimo fiorimento della teoria dinamica nella quale si osservava che l'azione di una forza era sempre mediata da un corpo che la esercitava. Come spesso si fa in Fisica, Newton sorvolò sulla questione e introdusse come giustificazione del problema il concetto di *azione a distanza*. Sebbene i dati sperimentali fossero totalmente a favore della teoria di Newton il problema epistemologico sorto relativamente all'interpretazione del concetto di *azione a distanza* non fu superato. L'impasse fu risolta con l'avvento della teoria quantistica dei campi tramite l'introduzione dei bosoni vettori che risultano essere i mediatori delle forze elementari. L'azione a distanza non diviene altro che lo scambio di un bosone vettore tra l'oggetto causante la forza elementare e l'oggetto che la subisce.

3.2 Causalità di Granger

La trattazione vista finora pone delle basi filosofiche ed epistemologiche relativamente al concetto di causalità. I collegamenti causa-effetto tra enti o eventi sono stati trattati a partire dalle leggi di natura che li collegano, piuttosto che a partire da una pura analisi sui dati a disposizione. Per meglio dire, non abbiamo evinto dalla trattazione un test statistico che permettesse di comprendere se una serie storica potesse essere causa di un altra (test indispensabile più che nello studio della Fisica nello studio statistico-

economico). È appunto in tale ambito che nasce una trattazione matematica della causalità.

Potrebbe essere opportuno chiedersi se sia possibile utilizzare la correlazione tra due serie storiche per ricercare l'esistenza di causalità tra di esse. Ciò deriva dall'idea che se le variazioni delle due serie storiche risultano presentarsi allo stesso tempo e con un andamento identico, allora sarebbe possibile un influenza di un sistema su un altro. Per osservare come detta procedura non è corretta, usiamo un esempio presente in [Su12]. Consideriamo due equazioni ricorsive accoppiate:

$$\begin{aligned} X(t+1) &= X(t)[r_x - r_x X(t) - \beta_{x,y} Y(t)] \\ Y(t+1) &= Y(t)[r_y - r_y Y(t) - \beta_{y,x} X(t)] \end{aligned} \tag{3.2.1}$$

e supponiamo di assumere i seguenti valori: $r_x = 3.8$, $r_y = 3.5$, $\beta_{x,y} = 0.02$ e $\beta_{y,x} = 0.1$. Come si osserva dalla figura 3.1, a seconda delle condizioni iniziali poste, i sistemi risultano correlati in maniera diversa o con correlazione che va e viene nel tempo. Tutto ciò può farci presagire che l'utilizzo della correlazione come metodologia di ricerca della causalità risulta piuttosto fallace e pone dunque la questione relativa alla formulazione di un altro strumento.

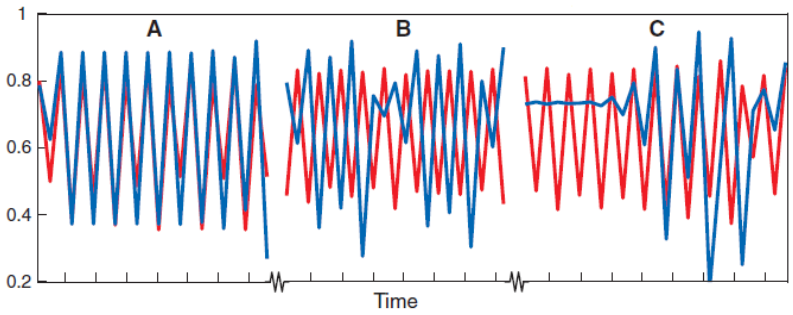


Figura 3.1: Correlazione tra il sistema X (blu) ed il sistema Y (rosso): (**A**) sistemi correlati, (**B**) sistemi anticorrelati e (**C**) in vari periodi i sistemi risultano correlati, anticorrelati o non correlati.

Ad ovviare al problema è stato Clive Granger il quale, nel suo articolo *"Investigating causal relations by econometric models and cross-spectral methods"*, pubblicato nel 1969 [Gr69], ha introdotto il concetto statistico di *causalità* basandosi sul concetto di predizione. La causalità di Granger

asserisce che, considerate due serie storiche $\{X_{1,t}\}$ e $\{X_{2,t}\}$, qualora i valori passati di $\{X_{1,t}\}$ diano informazioni predittive su $\{X_{2,t}\}$ oltre a quelle già fornite dai valori passati di $\{X_{2,t}\}$ stessa, migliorandone ovviamente la precisione previsionale, allora $\{X_{1,t}\}$ Granger-causa $\{X_{2,t}\}$. Da un punto di vista prettamente matematico potremmo affermare che $\{X_{1,t}\}$ Granger-causa $\{X_{2,t}\}$ se:

$$\begin{aligned} E[X_{2,t} - E(X_{2,t}|\cdot)|X_{2,t-1}, X_{2,t-2}, \dots; X_{1,t-1}, X_{1,t-2}, \dots]^2 \leq \\ \leq E[X_{2,t} - E(X_{2,t}|\cdot)|X_{2,t-1}, X_{2,t-2}, \dots]^2, \end{aligned} \tag{3.2.2}$$

ossia se la varianza di $\{X_{2,t}\}$ calcolata al tempo t condizionata dai valori precedenti di $\{X_{1,t}\}$ risulta minore rispetto alla varianza di $\{X_{2,t}\}$, calcolata al tempo t e non condizionata dai valori precedenti di $\{X_{1,t}\}$. Dunque, a tutti gli effetti la causalità di Granger non coincide con la causalità in genere. Dire che $\{X_{1,t}\}$ Granger-causa $\{X_{2,t}\}$ significa che $\{X_{1,t}\}$ contiene informazioni predittive su $\{X_{2,t}\}$, non che ne sia la causa. Quindi se $\{X_{1,t}\}$ rappresenta la serie storica del tasso di disoccupazione di una nazione e $\{X_{2,t}\}$ rappresenta la serie storica del PIL della stessa, l'affermazione "$\{X_{1,t}\}$ Granger-causa $\{X_{2,t}\}$" implica che il tasso di disoccupazione è utile per prevedere l'andamento del PIL e non che il tasso di disoccupazione causa una variazione del PIL stesso.

Volendo fare un esempio generico possiamo pensare al caso nel quale consideriamo tre serie storiche $\{X_t\}$, $\{Y_t\}$ e $\{W_t\}$. Inizialmente effettuiamo uno studio previsionale sul valore di X_{t+1} utilizzando solo i valori X_t e W_t calcolati al tempo t. Successivamente effettuiamo ancora una volta il calcolo previsionale sul valore di X_{t+1}, questa volta utilizzando oltre a X_t e W_t anche Y_t. Qualora la seconda previsione risulti più precisa della prima, allora risulterebbe che Y_t contiene informazioni utili per prevedere il valore di X_{t+1}, che non risultano contenute né in X_t né in W_t[2]. Nulla vieta l'eventualità che W_t abbia contenute in sé informazioni utili a fare previsioni sui possibili valori di X_{t+1}; l'osservazione posta dall'evenienza che $\{Y_t\}$ Granger-causi $\{X_t\}$ è che $\{Y_t\}$ contenga informazioni non presenti in $\{W_t\}$ e tanto meno

[2]Nell'effettuare la seconda previsione dobbiamo, ovviamente, tenere ancora una volta in conto la serie storica W_t.

in $\{X_t\}$. Dunque i requisiti affinché $\{Y_t\}$ Granger-causi $\{X_t\}$ si sintetizzano nella richiesta che Y_t preceda X_{t+1} e che $\{Y_t\}$ contenga informazioni utili per effettuare previsioni sulla serie storica $\{X_t\}$, non presenti in altre serie storiche.

Dopo aver definito il concetto di causalità di Granger risulta interessante vederne la trattazione matematica. Il contesto in cui la Granger-causalità viene più utilizzata è quello delle regressioni lineari, sebbene nuovi lavori abbiano ampliato il campo di utilizzo di questo strumento statistico. Supponiamo dunque di considerare un modello autoregressivo lineare bivariato di due variabili X_1 e X_2:

$$\begin{aligned} X_{1,t} &= \sum_{j=1}^{p} A_{11,j}\, X_{1,t-j} \;+\; \sum_{j=1}^{p} A_{12,j}\, X_{2,t-j} \;+\; E_{1,t} \\ X_{2,t} &= \sum_{j=1}^{p} A_{21,j}\, X_{1,t-j} \;+\; \sum_{j=1}^{p} A_{22,j}\, X_{2,t-j} \;+\; E_{2,t} \end{aligned} \tag{3.2.3}$$

dove p rappresenta il numero di step temporali, intercorsi dagli ultimi valori presi in considerazione nel modello al valore di cui vogliamo fare la previsione; la matrice A_j contiene i coefficienti che rappresentano il peso che o X_1 o X_2, considerati al tempo $t-j$, hanno nella previsione di X_1 o X_2 al tempo t; infine, E_1 ed E_2 rappresentano gli errori relativi alla predizione effettuata col modello regressivo rispettivamente su X_1 e su X_2. Qualora la previsione su $X_{1,t}$ sia migliorata attraverso l'introduzione nel modello regressivo dei valori precedenti a t di $\{X_{2,t}\}$, allora l'errore nel modello regressivo ($E_{1,t}$) dovrebbe diminuire. Questa circostanza non si verificherebbe in alcun modo se $A_{12,1} = A_{12,2} = \cdots = A_{12,p} = 0$; pertanto sarà conveniente effettuate un test di ipotesi con ipotesi nulla $H_0 : A_{12,1} = A_{12,2} = \cdots = A_{12,p} = 0$ con H_1 ipotesi complementare. La procedura che si segue è l'effettuazione di un test-F sull'ipotesi H_0. Si calcolano inizialmente le somme dei residui al quadrato per l'ipotesi H_0 ed H_1, pari rispettivamente a:

$$SRQ_1 = \sum_{t=1}^{T} E_{1,t}$$

$$SRQ_0 = \sum_{t=1}^{T} E^*_{1,t}$$

dove $E^*_{1,t}$ deriva da:

$$X_{1,t} = \sum_{j=1}^{p} A_{11,j}\, X_{1,t-j} \; + \; E^*_{1,t}$$

cioè dal modello regressivo in cui risulta verificata l'ipotesi H_0. Una volta calcolata:

$$F \; = \; \frac{(SRQ_0 - SRQ_1)/p}{SRQ_1/(T-2p-1)}$$

dove T è pari al numero di misure sperimentali effettuate sulle due serie storiche, si verifica se detto valore risulta maggiore rispetto al valore critico del 5% per la distribuzione di Fisher-Snedecor $F(p, T-2p-1)$. Se questa eventualità risulta verificata, rifiuteremo l'ipotesi nulla, cioè che $\{X_{2,t}\}$ non Granger-causa $\{X_{1,t}\}$; ciò implica che per F sufficientemente grandi affermeremo che $\{X_{2,t}\}$ Granger-causa $\{X_{1,t}\}$[3]. Un ulteriore approccio [Ma08] è quello di calcolare la causalità di Granger utilizzando la seguente espressione:

$$\mathscr{F}_{X_2 \to X_1} \; = \; ln \frac{\epsilon(X_1)}{\epsilon(X_1|X_2)} \tag{3.2.4}$$

ove $\epsilon(X_1|X_2)$ rappresenta la varianza di X_1 condizionata dai valori di X_2, mentre $\epsilon(X_1)$ è la varianza di X_1. Come si vede in tal caso, il parametro dipenderà dal tempo, a meno che non si abbiano distribuzioni stazionarie. La differenza sostanziale tra questo approccio e quello precedente è che quest'ultimo non si basa su un test parametrico che verifica la presenza di causalità, ma cerca di dare una misura dell'aumento previsionale che la serie storica $\{X_{2,t}\}$ dà relativamente a $\{X_{1,t}\}$. È possibile osservare che è definibile una procedura del tutto equivalente a quella appena definita presente in [Ma08]:

$$\delta_{X_2 \to X_1} \; = \; \frac{\epsilon(X_1) - \epsilon(X_1|X_2)}{\epsilon(X_1|X_2)} \tag{3.2.5}$$

che risulta collegabile a (3.2.4) tramite la trasformazione:

$$\mathscr{F}_{X_2 \to X_1} \; = \; ln[1 - \delta_{X_2 \to X_1}]. \tag{3.2.6}$$

[3]Per uno studio approfondito si rimanda al testo di Hamilton [Ha95].

3.2.1 Estensione del concetto di Causalità di Granger

Da ciò che si è osservato nel paragrafo precedente, si nota come si possa definire una procedura attraverso la quale verificare se una serie storica sia utile a prevedere l'andamento di un'altra. È stato necessario introdurre il concetto di *Causalità di Granger* per ottenere questo risultato. Relativamente però al suo utilizzo sorgono alcune problematiche. La chiave della Granger-causalità è la *separabilità* delle serie storiche causa ed effetto, cioè, l'idea per la quale l'informazione riguardante il fattore causa sia indipendente dal fattore causato, per meglio dire, l'informazione sulla causa non può essere estrapolata dai valori della variabile causata (per esempio l'informazione relativa al PIL non è contenuta all'interno della variabile numero disoccupati). Ciò implica la possibilità di eliminare facilmente dal modello la variabile causa in modo tale da applicare l'algoritmo di Granger. Qualora la separabilità non sia una proprietà del sistema oggetto di studio, insorgono problemi relativi all'applicazione della metodologia della causalità di Granger. Un tipico esempio può essere dato da sistemi accoppiati debolmente, come ad esempio i sistemi biologici. In un sistema biologico in cui sono presenti due specie (predatore-preda), le informazioni relative alla preda sono presenti e non eliminabili all'interno delle informazioni relative al predatore e viceversa. Per questo motivo è utile introdurre un ulteriore approccio che copra i settori in cui la causalità di Granger non può essere utilizzata. Questo metodo si chiama *Convergent Cross Mapping (CCM)*. Per introdurlo seguiamo la procedura definita in [Su12]. Consideriamo le equazioni accoppiate di Lorenz che ricordiamo essere definite dalle equazioni differenziali accoppiate:

$$\begin{cases} \dot{x} = \sigma(y - x) \\ \dot{y} = \rho x - xz - y \\ \dot{z} = xy - \beta z. \end{cases} \tag{3.2.7}$$

Risulta comodo introdurre questo caso per la possibilità di utilizzare in maniera più semplice la teoria della dinamica dei sistemi. Detto ciò, si considera l'attrattore del sistema, M (Figura 3.2), che corrisponde al luogo dei punti $\underline{m}(t) = [x(t), y(t), z(t)]$ ottenuti al variare di t. A partire da esso creiamo gli attrattori replicanti M_X e M_Y tramite questa procedura:

consideriamo il valore di x al tempo t e definiamo un intervallo temporale τ. Preso questo valore creiamo il vettore $\underline{x}(t) = [x(t), x(t-\tau), x(t-2\tau)]$: il luogo dei punti $\underline{x}(t)$ ottenuti al variare di t caratterizzerà M_X. Seguiamo la stessa procedura per definire M_Y. È possibile osservare che ad ogni $\underline{m}(t)$ corrisponde uno ed un solo $\underline{x}(t)$. Una corrispondenza uno ad uno si osserva ovviamente anche tra $\underline{m}(t)$ e $\underline{y}(t)$. Il fatto che ad un solo $\underline{m}(t)$ corrispondano due soli $\underline{x}(t)$ e $\underline{y}(t)$ implicherà facilmente che ad ogni $\underline{x}(t)$ corrisponderà un solo $\underline{y}(t)$ e viceversa.

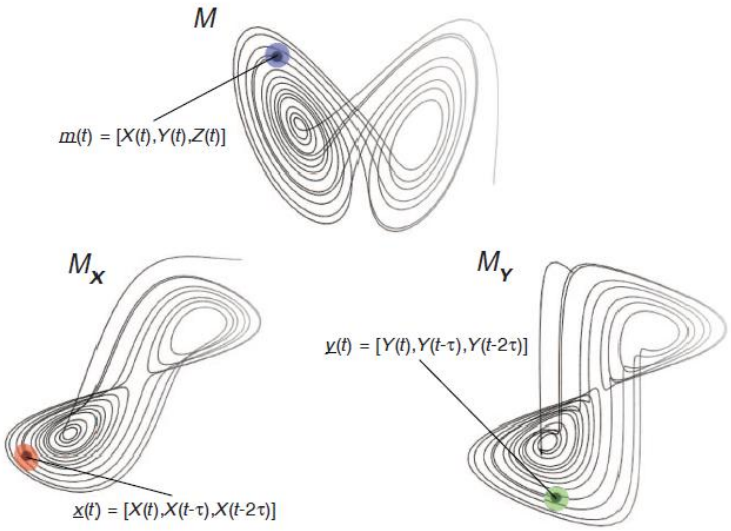


Figura 3.2: Attrattore di Lorenz M e attrattori fantasma M_X e M_Y.

L'idea di base del CCM è che qualora X e Y fossero accoppiati, i punti nelle vicinanze di $\underline{x}(t)$ (zona rossa intorno allo stesso, vedi fig. 3.2) corrisponderanno ai punti nelle vicinanze di $\underline{y}(t)$ (zona verde intorno allo stesso, vedi fig. 3.2). Abbiamo utilizzato l'idea presente nella teoria della dinamica dei sistemi, per la quale due serie storiche X e Y sono casualmente connesse se condividono un attrattore comune: ciò implica che ognuna delle due variabili può identificare lo stato dell'altra. Risulta semplice osservare che con l'aumentare delle osservazioni i primi vicini di $\underline{x}(t)$ e $\underline{y}(t)$ aumenteranno migliorando così la precisione. Una volta posta la base concettuale cerchiamo di introdurre un algoritmo che possa definire la procedura da seguire per effettuare i calcoli. Consideriamo inizialmente due serie storiche

$\{X\} = \{X(1) \dots X(L)\}$ e $\{Y\} = \{Y(1) \dots Y(L)\}$ di lunghezza L. Successivamente si definisce l'attrattore replicante M_X, considerando che esso è dato dall'insieme dei punti $\underline{x}(t) = [x(t), x(t-\tau), x(t-2\tau), \dots, x(t-(E-1)\tau)]$. Si noti che la dimensione di M_X risulta pari ad E. Terminata questa procedura iniziale, si calcola la stima data dal cross-mapping di $Y(t)$, che si denoterà come $\hat{Y}(t)|M_X$, ottenuta dalla seguente procedura: si prendono i primi $E+1$ vicini di $\underline{x}(t)$ e li si etichetta tramite i parametri $t_1, \dots, t_{E+1}$, dove l'indice varia dal più vicino al più lontano. Utilizzando la relazione uno ad uno tra $\underline{x}(t)$ e $\underline{y}(t)$, si identificano le $y(t_i)$ corrispondenti alle $x(t_i)$. Si usa, quindi, per la stima la relazione:

$$\hat{Y}(t)|M_X = \sum_{i=1}^{E+1} w_i \, Y(t_i) \tag{3.2.8}$$

in cui risulta che:

$$w_i = \frac{u_i}{\sum_i u_i}$$

e

$$u_i = e^{\frac{d[\underline{x}(t), \underline{x}(t_i)]}{d[\underline{x}(t), \underline{x}(t_1)]}}$$

con $d[\underline{x}(t), \underline{x}(t_i)]$ distanza euclidea tra i due vettori. Nella stessa maniera si definisce il cross-mapping in senso opposto. Se X ed Y sono dinamicamente accoppiati allora gli $Y(t_i)$ saranno ordinati secondo il pedice i dal più vicino al più lontano. Ciò implicherà un'ottima stima $\hat{Y}(t)|M_X$, in quanto i termini $Y(t_i)$ nella sommatoria sono opportunamente pesati. Inoltre, con l'aumentare del valore di L, aumentando i punti dell'attrattore, le distanze dai primi vicini si accorceranno, portando, qualora sussista l'accoppiamento, ad una convergenza tra $\hat{Y}(t)|M_X$ e $Y(t)$ e viceversa. Per testare il valore di detta convergenza, si effettua un test regressivo (indice di correlazione di Pearson, ρ), vedendo quanto $\hat{Y}(t)|M_X$ e $Y(t)$ risultano correlati. Un esempio pratico può essere fornito dal sistema descritto dalle equazioni (3.2.1) con i valori dei parametri posti in precedenza. Utilizzando la procedura appena definita per questo sistema, si ottiene un indice di correlazione tra $\hat{Y}(t)|M_X$ e $Y(t)$ e tra $\hat{X}(t)|M_Y$ e $X(t)$ descritto in figura 3.3.

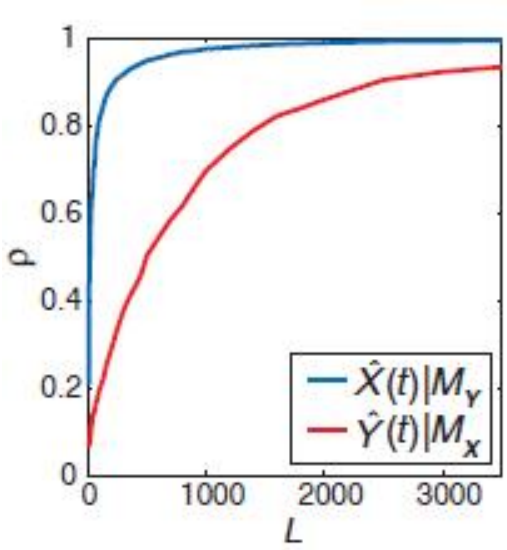


Figura 3.3: Indice di correlazione tra $\hat{Y}(t)|M_X$ e $Y(t)$ (linea rossa) e tra $\hat{X}(t)|M_Y$ e $X(t)$ (linea blu). La più rapida convergenza della correlazione tra $\hat{X}(t)|M_Y$ e $X(t)$ è dovuta al fatto che gli effetti di X su Y sono più forti che il contrario. Infatti: $\beta_{y,x} > \beta_{x,y}$. Dunque utilizzare Y per prevedere i valori di X risulta molto conveniente in quanto i valori di Y sono fortemente influenzati da quelli di X.

Traspare, come già precedentemente affermato, che con l'aumentare degli eventi considerati la correlazione aumenti sempre più fino a raggiungere un plateau. Per il caso in oggetto risulta che la correlazione tra $\hat{X}(t)|M_Y$ e $X(t)$ è più forte rispetto a quella tra $\hat{Y}(t)|M_X$ e $Y(t)$. Si viene a creare dunque una distinzione tra le situazioni che si possono incontrare:

1. *Causalità bidirezionata*: è il caso in cui si hanno effetti della variabile X su Y e viceversa. Un esempio è dato dalle equazioni (3.2.1) (fig. 3.3). In tal caso, sebbene non simmetricamente, risulta che le due variabili influiscono vicendevolmente l'una sull'altra. Infatti come abbiamo visto (fig. 3.3), qualora $\beta_{y,x} > \beta_{x,y}$ si avrebbe una diversa influenza di X su Y rispetto al contrario. Ciò permette di ottenere una più veloce convergenza nella predizione di X rispetto a quella relativa ad Y e, dunque, permette di vedere come X "causi" Y (il fatto che Y migliori la previsione di X implica che i valori di Y sono fortemente influenzati dai valori di X, $\beta_{y,x} > \beta_{x,y}$) in maniera maggiore rispetto a come X "causa" Y. Può essere interessante vedere come la direzione dell'informazione vari al variare dei parametri $\beta_{y,x}$ e $\beta_{x,y}$. Per far

questo fissiamo la lunghezza L a 400 e facciamo i calcoli al variare dei parametri. I risultati riportati in figura 3.4 ci fanno osservare come nei casi in cui si abbia una disuguaglianza tra $\beta_{y,x}$ e $\beta_{x,y}$, la correlazione tra stima di una variabile e valore effettivo della stessa risulta migliore o peggiore rispetto all'altra variabile.

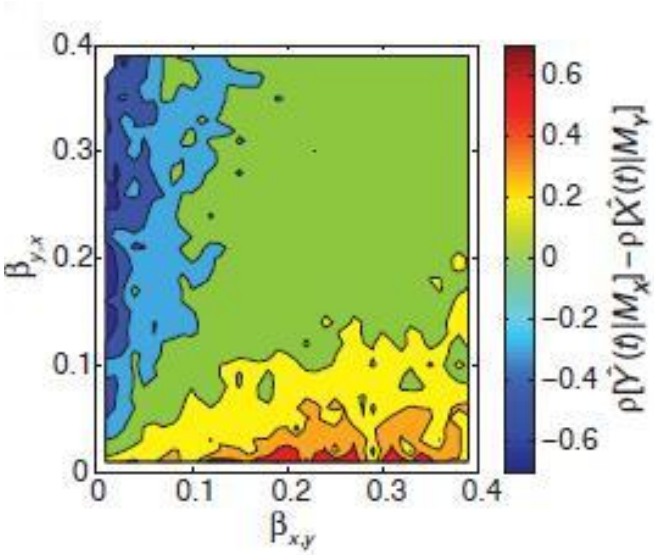
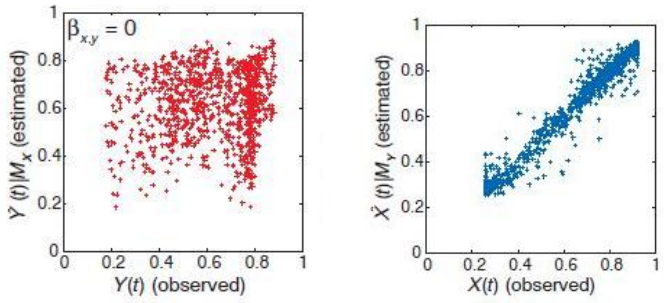


Figura 3.4: Differenza del valore di correlazione tra $\hat{Y}(t)|M_X$ e $Y(t)$ e $\hat{X}(t)|M_Y$ e $X(t)$ al variare dei parametri $\beta_{y,x}$ e $\beta_{x,y}$ ($L = 400$).

2. *Causalità unidirezionata*: è il caso in cui la variabile X influenza la dinamica di Y, ma non è valido il contrario. È come se non ci fosse più l'effetto "feedback". Un esempio è fornito in figura 3.5 dove, considerando il solito modello dato dalle equazioni (3.2.1), abbiamo posto $\beta_{x,y} = 0$, in modo tale che Y non abbia alcuni effetti predittivi su X e, quindi, considerato il caso in cui Y non sia causa di X.

   C'è una problematica, però, relativa alla causalità unidirezionata. Com'è facile supporre, anche qualora $\beta_{x,y} = 0$, e quindi Y non abbia alcuni effetti predittivi su X, nell'eventualità in cui gli effetti predittivi di X su Y (descritti dal parametro $\beta_{y,x}$) risultino molto forti, risulterebbe verificato il caso della "sincronia", che porterebbe ancora una volta ad una causalità bivariata. Un esempio di calcolo è dato dalla figura 3.6 basata sul solito modello delle equazioni ricorsive accoppiate (3.2.1),



(a) **(b)**

Figura 3.5: Confronto tra valori stimati e valori osservati nel caso in cui: (a) Y non ha alcuni effetti predittivi su X ($\beta_{x,y} = 0$) e (b) Y ha effetti predittivi su X ($\beta_{x,y} \neq 0$).

in cui si è posto $\beta_{x,y} = 0$ e si son fatti variare i parametri $\beta_{y,x}$ ed L.

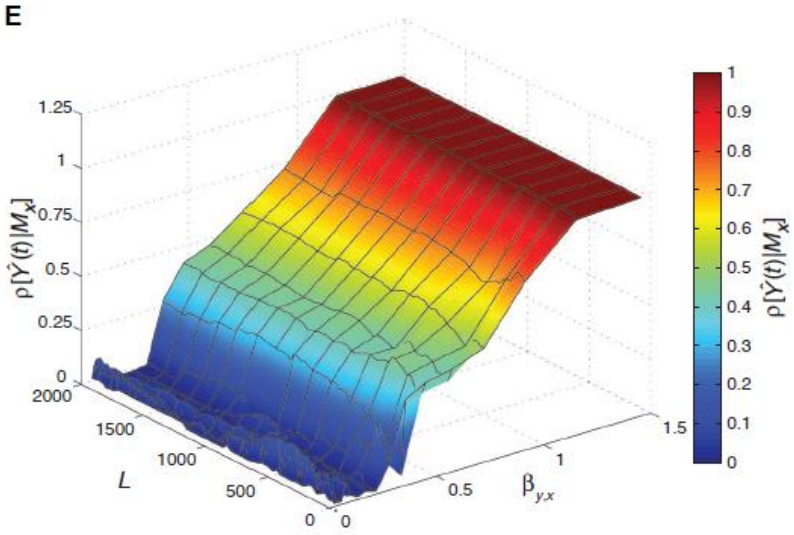


Figura 3.6: Valore della correlazione tra $\hat{Y}(t)|M_X$ e $Y(t)$ al variare dei parametri L e $\beta_{y,x}$ posto $\beta_{x,y} = 0$.

Come precedentemente detto, ciò che si osserva è la presenza di una causalità bidirezionale, qualora ci trovassimo innanzi ad una forte influenza di X su Y.

Effettuate queste premesse, possiamo osservare come, considerando un sistema complesso, il CCM sia un ottimo sistema di indagine. Prendiamo in analisi un sistema che funzioni come quello descritto in figura 3.7, nel quale

le specie 1, 2 e 3 interagiscono tra loro e la loro interazione è causa delle specie 4 e 5.

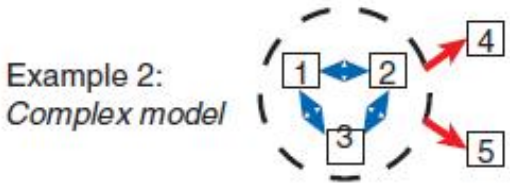


Figura 3.7: Struttura di un sistema complesso.

Applicando a questo sistema il metodo del cross mapping, si ottengono i valori di correlazione (figura 3.8), che misurano, come già detto, la causalità nel sistema.

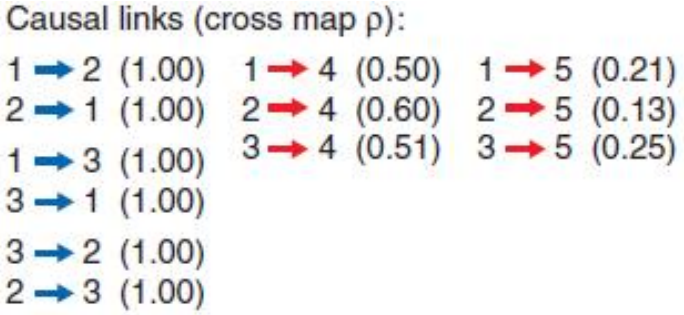


Figura 3.8: Risultati del cross mapping relativi al sistema complesso (sono stati considerati solo i link significativi cioè quelli che hanno opportuna probabilità di interagire).

Come si può notare, la metodologia CCM è molto utile anche nello studio di sistemi complessi.

Ciò premesso, risulta opportuno chiarire tutto ciò che è stato appena definito tramite un esempio pratico.

Esempio 3.1 (Relazioni di causa tra il sistema sardine e il sistema alici)**.** Usiamo un esempio presente in [Su12]. Si vuole studiare la relazione esistente tra la pesca di sardine (nel Pacifico) e la pesca delle alici (nel Mare del Nord). Relativamente a questo studio ci sono state molte ipotesi, ma nessun metodo statistico di rivelazione di interconnessioni casuali era mai stato utilizzato. Come punto di partenza si considera la serie temporale che denota il numero delle sardine e delle alici pescate rappresentata in figura 3.9.

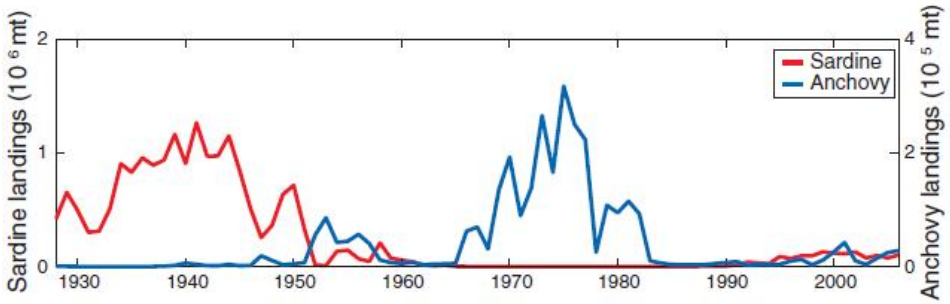


Figura 3.9: Numero delle sardine (rosso) e delle alici (blu) pescate nei vari anni.

Utilizzando questi parametri, possiamo applicare il CCM per vedere se c'è una relazione di causalità tra la quantità di sardine e di alici pescate. Come si evince dalla figura 3.10, non vi è alcun segnale tra alici e sardine che indica una totale mancanza di interazione tra le due specie.

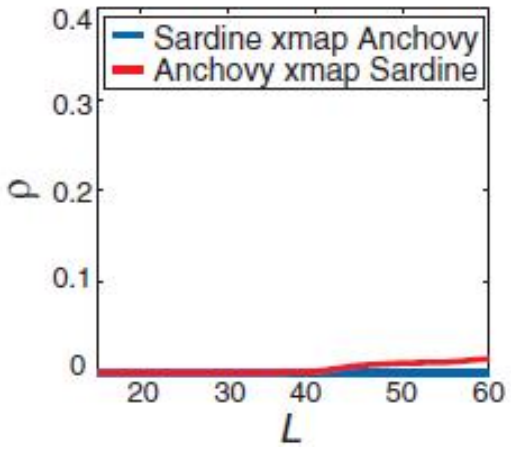
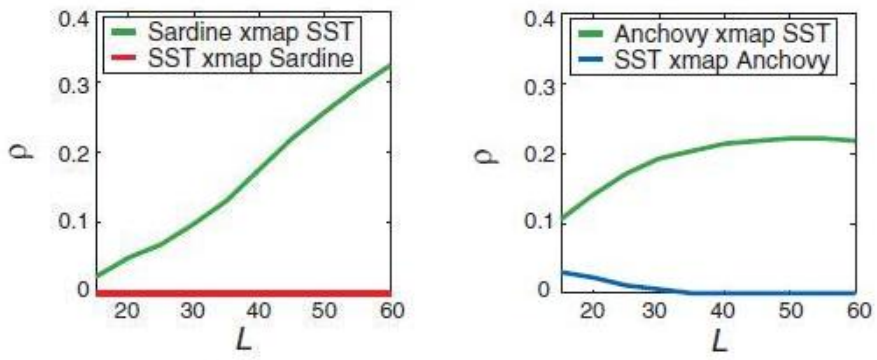


Figura 3.10: Correlazione tra valore stimato, calcolato col cross mapping con i valori relativi alle alici, e valore effettivo delle sardine (blu). Correlazione tra valore stimato, calcolato col cross mapping con i valori relativi alle sardine, e valore effettivo delle alici (rosso).

Ulteriore osservazione può essere quella relativa alla ricerca di causalità tra il numero delle sardine o alici pescate e la temperatura superficiale del mare. Per far ciò si applica il CCM tra la serie storica del numero di sardine (o alici) pescate e la serie storica della temperatura superficiale del mare. I risultati riportati in figura 3.11 fanno trasparire una casualità unidirezionata, evidenziando la presenza di informazione relativa alla temperatura

superficiale del mare all'interno dei valori delle sardine pescate e delle alici pescate. Ciò porterà all'osservazione della presenza di un debole accoppiamento unidirezionale tra temperatura e volume del pescato, che permetterà di osservare una debole relazione di causalità della temperatura superficiale del mare sul volume di pesca.



(a) *Cross mapping numero sardine pescate con temperatura superficiale del mare (verde) e cross mapping temperatura superficiale del mare con numero sardine pescate (rosso).*

(b) *Cross mapping numero alici pescate con temperatura superficiale del mare (verde) e cross mapping temperatura superficiale del mare con numero alici pescate (blu).*

Figura 3.11: Ricerca della connessione di causalità tra quantità di sardine pescate e temperatura superficiale dell'acqua (a) e tra quantità di alici pescate e temperatura superficiale dell'acqua (b).

Dunque, sebbene il sistema sardine non interagisce col sistema alici, detti sistemi risultano debolmente forzati dalla temperatura superficiale del mare.

3.3 Transfer Entropy

Sinora abbiamo visto soltanto se una serie storica risulta essere causa di un'altra. Nessun accenno è stato fatto relativamente a modalità di studio tramite le quali stabilire la quantità d'informazione trasferita da un processo ad un'altro. Utilizzeremo a tal scopo la definizione di *transfer entropy* data da Schreiber [Sc00].

La definizione di detta quantità è particolare e merita attenzione. Per la sua introduzione conviene partire dalle problematiche che la mutua informazione possiede relativamente alla richiesta che possa essere misura dell'informazione scambiata. In primis ricordiamo che la mutua informazione non

rappresenta altro che l'informazione contenuta in una variabile aleatoria X relativamente ad una variabile aleatoria Y. Detta quantità (2.2.2) risulta deficitaria nella richiesta di essere una misura dell'informazione scambiata, a causa dell'assenza di termini dinamici e di direzione relativamente all'informazione (la mutua informazione è simmetrica). Potremmo pensare di ridefinire la mutua informazione, introducendo un ritardo temporale tra le due variabili studiate, in maniera tale che la mutua informazione divenga pari a:

$$M_{IJ} = \sum p(i_n, j_{n-\tau})\, log_2 \frac{p(i_n, j_{n-\tau})}{p(i_n)\, p(j_{n-\tau})}.$$

In tal caso quello che effettueremmo è uno studio relativo all'eventualità in cui i al tempo n risulti indipendente da j al tempo $n-\tau$. Questa misura mi fornirebbe la quantità d'informazione che $j_{n-\tau}$ possiede relativamente a i_n, dando solo una descrizione relativa all'indipendenza dei due segnali a tempi ritardati, senza esplicitare se la misurazione effettuata sia da riferire all'informazione scambiata piuttosto che ad una risposta dovuta da un segnale comune. Bisogna perciò cercare di ottenere uno strumento di calcolo più opportuno.

La *transfer entropy* risulta essere ottima a tal scopo. Come punto di partenza per l'introduzione di questa entropia consideriamo probabilità che tengano conto della dinamica del sistema. Per semplificare il problema, supponiamo che il sistema in oggetto possa essere approssimato tramite un processo di Markov di ordine k, cioè che la probabilità di trovare i_{n+1} al tempo $n+1$ sia pari a $p(i_{n+1}|i_n, \ldots, i_{n-k+1}) = p(i_{n+1}|i_n, \ldots, i_{n-k})$, cioè che il valore ottenuto a $n+1$ sia condizionato solo dagli ultimi k eventi precedenti. Introduciamo la notazione sintetica $i_n^{(k)} = (i_n, \ldots, i_{n-k+1})$. Utilizzando l'entropia condizionata, possiamo definire l'*entropy rate*, ossia la quantità d'informazione che tutti gli stati del sistema, precedenti il tempo $n+1$, non forniscono relativamente allo stato che il sistema assume al tempo $n+1$, come:

$$h_I = \sum p(i_{n+1}, i_n^{(k)})\, log_2\, p(i_{n+1}|i_n^{(k)}).$$

Questa misura considererebbe solo un unico processo e non metterebbe in luce il possibile trasferimento d'informazione tra un processo ed un altro.

Detto ciò, è necessaria una misura che mi dica quanta informazione di un processo passi ad un altro. Supponiamo di avere due processi I e J e che il processo I risulti essere markoviano di ordine k. Qualora i valori passati del processo J, rispetto al tempo $n+1$, non fornissero alcuna informazione relativa al valore che I assume al tempo $n+1$, allora risulterebbe che i valori di J non hanno alcun influsso nella probabilità condizionata relativa a i_{n+1}. Per meglio dire risulterebbe che (considerando influenti solo gli ultimi l valori passati di J):

$$p(i_{n+1}|i_n^{(k)}) = p(i_{n+1}|i_n^{(k)}, j_n^{(l)}).$$

L'eventualità appena accennata coincide con il caso in cui non vi sia alcun flusso di informazione tra J ed I. L'errore nel supporre che il flusso di informazione non vi sia e dunque nel non considerare i valori passati di J porta ad un aumento dell'informazione necessaria per codificare il segnale I al tempo $n+1$. La quantità aggiuntiva di informazione necessaria coincide proprio col flusso di informazione che da J passa ad I. Per calcolare questa quantità viene in aiuto l'entropia relativa (distanza di Kullback), con la quale calcoleremo l'informazione aggiuntiva da apportare, nell'eventualità in cui considerassimo i valori passati, rispetto al tempo $n+1$, di J ininfluenti per la previsione del valore i_{n+1}. Questa quantità risulterà pari a:

$$\mathscr{T}_{J\to I} = \sum p(i_{n+1}, i_n^{(k)}, j_n^{(l)})\, log_2 \frac{p(i_{n+1}|i_n^{(k)}, j_n^{(l)})}{p(i_{n+1}|i_n^{(k)})}. \tag{3.3.1}$$

Detta relazione può essere facilmente riscritta per ottenere una relazione analoga (ponendo $k=1$ e $l=1$):

$$\mathscr{T}_{J\to I} = H(i_{n+1}|i_n) - H(i_{n+1}|i_n, j_n). \tag{3.3.2}$$

C'è una problematica relativa all'applicazione della transfer entropy per i sistemi continui. La definizione data pone l'accento solo circa sistemi discretizzati. Per estendere il concetto a casi continui abbiamo due possibili strade: o la discretizzazione del campione continuo, suddividendo l'intervallo in cui possiamo trovare i valori del sistema in uno o più sottointervalli, rendendo dunque discreto il problema ed applicando successivamente la definizione

posta in (3.3.1); oppure sostituendo la somma con un integrale

$$\mathscr{T}_{J\to I} = \int p(i_{n+1}, i_n^{(k)}, j_n^{(l)})\, log_2 \frac{p(i_{n+1}|i_n^{(k)}, j_n^{(l)})}{p(i_{n+1}|i_n^{(k)})}\, di\, d^k i_n^{(k)}\, d^l j_n^{(l)},$$

cercando di capire come calcolare dal campione discreto misurato la distribuzione di probabilità congiunta, dalla quale successivamente calcolare le probabilità condizionate e le densità marginali. L'idea principale è di calcolare la distribuzione congiunta attraverso la *kernel density estimation* di cui un esempio può essere dato (ponendo $k = 1$ e $l = 1$) da

$$\hat{p}_r(i_{n+1}, i_n, j_n) = \frac{1}{N} \sum_{n'} \vartheta \left(r - \left| \begin{matrix} i_{n+1} - i_{n'+1} \\ i_n - i_{n'} \\ j_n - j_{n'} \end{matrix} \right| \right)$$

dove N è il numero di campioni estratti dalla serie storica continua, ϑ è la nostra *kernel function*, che in tal caso coincide con la θ di Heaviside, ed r è un parametro di grandezza che indica la scala di approssimazione rispetto alla quale calcoliamo la distribuzione, anche chiamato *parametro di kernel*. L'idea sarebbe porre il parametro r a 0, ma ciò non può essere fatto in quanto i risultati risulterebbero distorti. Altri kernel possono essere utilizzati relativamente al calcolo di $\hat{p}_r(i_{n+1}, i_n, j_n)$. Le procedure di discretizzazione del campione e di calcolo della distribuzione di probabilità tramite *kernel density estimation* possono inoltre essere utilizzate contemporaneamente. Un esempio di applicazione di detto concetto può essere il calcolo della transfer entropy tra le serie storiche, che misurano la frequenza istantanea del battito del cuore e la frequenza istantanea del respiro in un uomo che dorme. Il campione è espresso in figura 3.12. I risultati espressi in figura 3.13 mostrano come la mutua interazione tra le due serie storiche sia simmetrica (cosa che ci aspettavamo) e come per un significativo range di valori di r (parametro del kernel) ci sia un maggiore flusso di informazione dal cuore al respiro piuttosto che il contrario. Ciò indica una causalità bidirezionata asimmetrica tra i due sistemi con maggiore influsso del cuore sul respiro che viceversa. Inoltre è possibile osservare che al variare di r a zero la transfer entropy si annulla a causa della dimensione finita dei campioni.

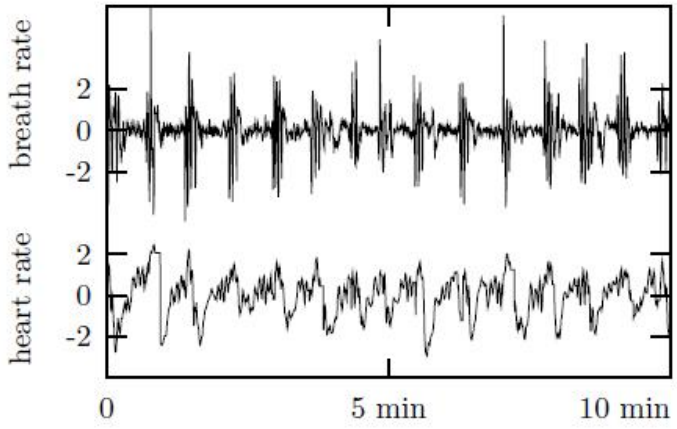


Figura 3.12: Serie storiche della frequenza del respiro e della frequenza del battito in un uomo che dorme.

In ultima analisi possiamo chiederci adesso se esista una relazione tra la transfer entropy e la causalità di Granger. Come abbiamo visto nella trattazione della Granger-causalità, essa poggia la sua definizione sull'idea di previsione, cioè sull'idea che una variabile X Granger-causerà Y se i valori di X sono utili per migliorare la previsione sul futuro di Y. La tranfer entropy si poggia su un idea diversa: essa è definita in termini di riduzione dell'incertezza sulla variabile Y. Un modo alternativo per definirla potrebbe essere l'idea per la quale $\mathscr{T}_{X\to Y}$ è il grado di riduzione dell'ambiguità sul futuro di Y. È lecito domandarsi se l'approccio ''predittivo" possa essere messo in relazione con l'approcio di ''riduzione dell'ambiguità''. A risolvere la questione ci ha pensato L. Barnett che nel suo lavoro, *Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables* [Ba09], ha posto in stretto legame transfer entropy e Granger-causa. Ponendosi nel caso, come accennato dal titolo dell'articolo, dello studio di variabili gaussiane, è pervenuto al risultato:

$$\mathscr{F}_{X\to Y} = 2\,\mathscr{T}_{J\to I} \tag{3.3.3}$$

dove $\mathscr{F}_{X\to Y}$ sfrutta la definizione di causalità di Granger presente nell'equazione (3.2.4).

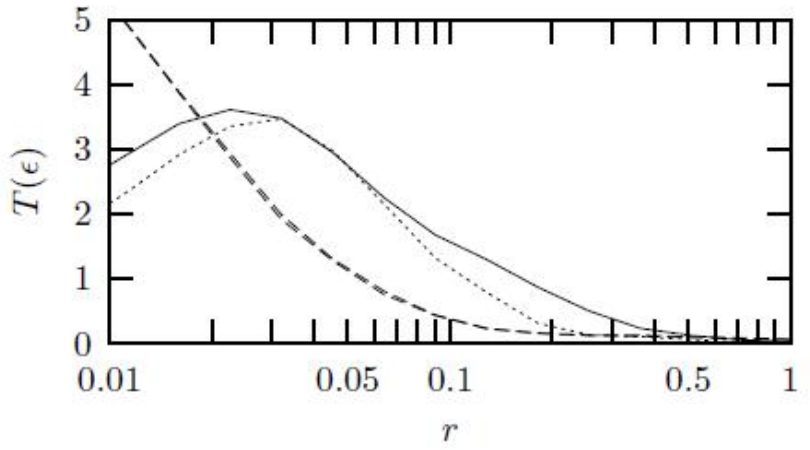


Figura 3.13: Transfer entropies $\mathscr{T}(cuore \rightarrow respiro)$ (linea continua), $\mathscr{T}(respiro \rightarrow cuore)$ (linea punteggiata) e mutua informazione ritardata M_τ con $\tau = 0.5s$ (linea tratteggiata).

3.4 Partial Conditioning

La descrizione posta nelle relazioni (3.2.4) e (3.2.5) presenta all'atto pratico alcune problematiche. L'introduzione di dette relazioni è stata necessaria soprattutto per lo studio di contesti multivariati nei quali non sono unicamente due le variabili (X_1 ed X_2 o X ed Y che dir si voglia) ad esser presenti, ma un numero ben superiore. Infatti, si può osservare che l'applicazione della causalità di Granger può essere non ideale nel caso multivariato. Il principio di questo problema sta nel fatto che l'utilizzo della Granger-causalità porta a misure non corrette in quanto non permette di eliminare gli effetti di causalità indiretta presenti nel network. Per ovviare a detta problematica è stata definita una procedura di *partial conditioning* [Ma12], che si basa sull'idea per la quale condizionare su poche variabili, scelte in maniera tale che risultino le più informative relativamente alla variabile guida, invece che su tante, permette di eliminare le interazioni indirette sparse nel network. Inoltre il condizionamento su un alto numero di variabii porta alla necessità di avere un gran numero di campioni per ottenere risultati significativi. Riducendo il numero di variabili condizionanti, si possono ottenere risultati ragionevoli anche con piccoli campioni.

In generale la Granger-causalità non definisce alcuna procedura per sce-

gliere quali siano le variabili da scartare, mentre nella teoria dell'informazione è possibile individuare le variabili più informative scegliendole una ad una attraverso la mutua informazione. Prima di definire la procedura di lavoro introduciamo il contesto di studio: consideriamo una serie storica $\{x_\alpha(t)\}_{\alpha=1,\dots,n}$. A partire da data serie temporale si determina il vettore di stato $X_\alpha(t)$, pari a:

$$X_\alpha(t) = (x_\alpha(t-m), \dots, x_\alpha(t-1))$$

dove m rappresenta la finestra temporale del cross-mapping. Denominiamo con $\epsilon(x_\alpha|\mathbf{X})$ l'errore quadratico medio relativo alla previsione del valore di x_α, calcolato al tempo t, condizionata dal valore assunto negli istanti precedenti il tempo t da tutte le variabili facenti parte del network $\mathbf{X}$. I valori passati verranno considerati all'interno della finestra temporale di lunghezza pari ad m precedente l'istante t. Supponiamo di voler calcolare l'influenza che X_β ha su X_α. Si denota data influenza come $X_\beta \to X_\alpha$, ove X_β rappresenta una o più variabili interne ad $\mathbf{X}$. In tali circostanze le relazioni (3.2.4) e (3.2.5) verranno ridefinite come:

$$\mathscr{F}_{X_\beta \to X_\alpha} = ln \frac{\epsilon(x_\alpha|\mathbf{X}/X_\beta)}{\epsilon(x_\alpha|\mathbf{X})} \tag{3.4.1}$$

per (3.2.4), e:

$$\delta_{X_\beta \to X_\alpha} = \frac{\epsilon(x_\alpha|\mathbf{X}/X_\beta) - \epsilon(x_\alpha|\mathbf{X})}{\epsilon(x_\alpha|\mathbf{X}/X_\beta)} \tag{3.4.2}$$

per (3.2.5), ove $\mathbf{X}/X_\beta$ significa che considero tutte le variabili in $\mathbf{X}$ eccetto quelle interne a X_β. Descriviamo la procedura da seguire supponendo di voler calcolare la causalità $X_\beta \to X_\alpha$. Fissiamo inizialmente il numero di variabili condizionanti da utilizzare. Chiamiamo tale valore n_d. Successivamente denotiamo con $\mathbf{Z} = (X_{i,1}, X_{i,2}, \dots, X_{i,n_d})$ il set di variabili più informative per X_β. In altre parole $\mathbf{Z}$ massimizza la mutua informazione (eq. (2.2.2)) $I(X_\beta; \mathbf{Z})$ calcolata variando su tutti i possibili set di n_d variabili, scelte nell'insieme $\mathbf{X}$. Successivamente si calcola la causalità che risulterà pari a:

$$\mathscr{F}_{X_\beta \to X_\alpha} = ln \frac{\epsilon(x_\alpha|\mathbf{Z})}{\epsilon(x_\alpha|\mathbf{Z} \cup X_\beta)}. \tag{3.4.3}$$

Esiste una procedura ottimizzata per costruire il set $\mathbf{Z}$ delle variabili che massimizzano la mutua informazione. Detta procedura si basa su una scelta iterativa delle migliori variabili da utilizzare: inizialmente si calcola la mutua informazione tra la variabile X_β e tutte le altre variabili del sistema e si sceglie tra tutte queste quella che massimizza la mutua informazione. La seconda variabile si sceglie, tra le variabili rimanenti, in maniera tale che risulti essere quella per la quale, congiuntamente alla variabile precedentemente selezionata, si massimizzi la mutua informazione con la variabile X_β. Successivamente si procede iterativamente alla stessa maniera. Cioè, prese le $k-1$ variabili $\mathbf{Z}_{k-1}$, definite con la procedura iterativa precedentemente descritta, il set $\mathbf{Z}_k$ è ottenuto aggiungendo alle variabili contenute in $\mathbf{Z}_{k-1}$ la variabile, scelta tra le variabili rimanenti del sistema, che massimizzi la mutua informazione congiuntamente alle variabili contenute in $\mathbf{Z}_{k-1}$, ossia la variabile che dia maggior apporto d'informazione.

Capitolo 4

Applicazioni

Il partial conditioning, introdotto nel paragrafo 3.4, rappresenta la maniera ideale per ricercare i rapporti di causalità in sistemi in cui la Granger-causalità non risulta essere un ottimo approccio descrittivo. Considereremo da un punto di vista pratico questa affermazione attraverso l'utilizzo sia della causalità di Granger sia del partial conditioning relativamente a due set di dati: i geni ‘‘HeLa” e le zone celebrali in un paziente che dorme.

4.1 La coltura cellulare ‘‘HeLa’’

La coltura cellulare ‘‘HeLa” è una famosa linea cellulare isolata da un cancro della cervice uterina. Il suo principale utilizzo si ha nella ricerca biomedica relativa alla risposta che le colture hanno relativamente ad infezioni da virus. La caratteristica principale di detta coltura risulta essere l'immortalità. Le ‘‘HeLa” sono state le prime cellule umane a presentare questa proprietà. Questa caratteristica, propria non solo delle ‘‘HeLa”, ma di numerose altre colture cellulari, implica che le ‘‘HeLa” possono dividersi un numero illimitato di volte, sempre che le condizioni di sopravvivenza vengano mantenute. Questa anomalia è dovuta ad una mutazione della *telomerasi*, che previene l'accorciamento del *telomero* durante la replicazione. È l'accorciamento del telomero a provocare nelle cellule normali la morte dopo un numero fissato di riproduzioni e, dunque, il mancato accorciamento dovuto all'anomalia nella telomerasi provoca l'immortalità osservata nelle ‘‘HeLa”.

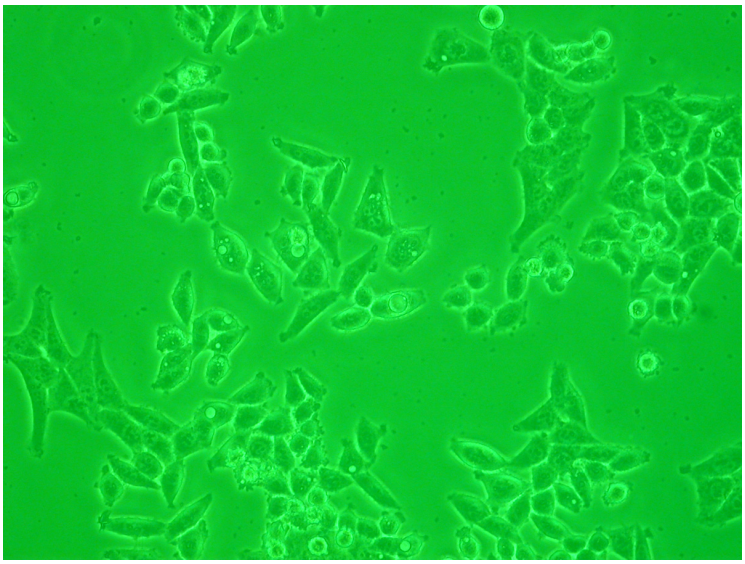


Figura 4.1: Cellule HeLa al microscopio a contrasto di fase.

Questa linea cellulare fu isolata e propagata da George Otto Gey nel 1951. Gey ricevette una porzione di tessuto contenente le ‘‘HeLa” da un medico dell'ospedale Johns Hopkins di Baltimora, che aveva ottenuto il lembo di tessuto da una biopsia della massa tumorale di una paziente, Henrietta Lacks, attuata a fini diagnostici l'8 Febbraio del 1951. La paziente morì nell'arco di pochi mesi (precisamente il 4 Ottobre 1951) a causa delle complicazioni della neoplasia. Gey moltiplicò le cellule tumorali ricevute *in vitro* senza il consenso della Lacks e dei suoi parenti (all'epoca non era necessario il consenso informato per l'utilizzo dei prelievi bioptici in quanto essi erano considerati di proprietà dell'ospedale[1]); presto si accorse di avere per le mani un tipo di cellule di particolare interesse e iniziò a donarle, insieme a tutti i brevetti per gli strumenti e i processi ideati dal suo laboratorio, a qualunque scienziato li richiedeva, tutto questo per il bene della scienza. Queste donazioni si sono succedute nel tempo e, nonostante le cellule che si riproducono *in vitro* nelle varie parti del mondo continuino a mutare, tutte loro discendono dall'unico tessuto primordiale estratto dalla Lacks. Si conta che al giorno d'oggi dette cellule risultano molto più numerose rispetto a tutte

[1]Relativamente a questa storia ci fu una controversia risolta dalla Corte Suprema della California, caso *Moore vs. Regents of the University of California*, nella quale si stabilì che un tessuto prelevato da un paziente non è di suo possesso e può essere commercializzato (per la legge della California).

le cellule presenti nel corpo di Henrietta Lacks al momento del prelievo.

Il nome ‘‘HeLa’’ deriva dalle prime lettere del nome e del cognome della paziente, che in un primo momento, per mantenere l’anonimato, vennero cambiati in Helen Lane o Helen Larson. Successivamente fu reso noto il suo vero nome.

Sebbene, come accennato la paziente da cui fu effettuato il prelievo morì per il cancro, le cellule tumorali a lei asportate risultano ancora vive e vengono utilizzate nei laboratori ancora oggi per scopi che vanno dalla ricerca per i vaccini alla ricerca sul cancro. Le cellule ‘‘HeLa” possono dividersi molte più volte rispetto alle altre cellule; è come se avessero il meccanismo di morte cellulare programmata (*apoptosi*) spento. Esse sono oggi coltivate in laboratori di tutto il mondo per fini scientifici e, sebbene siano trattate come cellule tumorali, posseggono delle caratteristiche uniche, che le differenziano dalle altre cellule di questo tipo. Le ‘‘HeLa” sono molto più resistenti delle altre cellule tumorali e sono in grado di sopravvivere in condizioni che altre cellule non possono tollerare; sono in grado di vivere per un periodo relativamente lungo anche in assenza di terreno di coltura.

Queste cellule presentano una forte differenziazione dalle cellule umane. In tali tipi di cellule sono presenti un numero di cromosomi che varia da 76 a 80 (contro un valore pari a 46 per un uomo sano). A causa di ciò si è supposto di rendere dette cellule una specie a se stante, col nome di Helacyton gartleri (in onore di Stanley Gartler). Detta proposta non è stata accettata dalla comunità scientifica.

Applicheremo ad alcuni geni di tali cellule la procedura del partial conditioning. Il set di dati [Fu07] [Wh02] corrisponde a 94 geni (tabella 4.1) di cui si è misurata la concentrazione in 48 tempi, scanditi da un intervallo temporale pari ad un’ora. I 94 geni sono stati selezionati sulla base della loro funzionalità nella regolazione del ciclo cellulare e dello sviluppo tumorale. I dati ottenuti da tali misurazioni sono stati successivamente filtrati e si è applicato a tale set inizialmente la procedura per il calcolo delle Granger-causalità e successivamente la procedura del partial conditioning.

Tabella 4.1: Nomi dei geni. Essi risultano etichettati in modo crescente (da 1 a 94) partendo da NFkB e procedendo lungo le righe.

$NFkB$	$IL-1b$	$IFN-g$	$IL-1RA$
$IL-6$	$IL-8$	$MCP-1$	$TNF-a$
$ICAM-1$	$VCAM-1$	$COX-2$	$Bcl-XL$
$Faz(CD95)$	IAP	$A20$	$c-myc$
$IRF-2$	$IkappaBa$	$JunB$	$P53$
$P21$	$GADD45$	$B99$	$TSP1$
$BAI-1$	$MASPIN$	PAI	$PIDD$
Fas	$Killer/DR5$	$Noxa$	$PUMA$
$E2F-1$	$PERP$	$IGF-BP3$	$Ribonucleotide\,reductase$
$Wip1$	$STAT3$	$Bcl-2$	$Mcl-1$
$CyclinD1$	$CyclinE1$	$c-jun$	$c-fos$
$IRF-1$	PKR	$STAT-1$	$RECK$
$COL6A1$	$BRCA1$	$BRCA2$	$PDX1$
$OCT4$	$PDGFRA$	$VEGF$	$FGFR1$
$FGF7$	$FGF1$	$FGF5$	$FGF18$
$FGFR4$	$FGFR3$	$FGF2$	$FGFR2$
$FGF12B$	$FGFRL1$	$FGFR1$	$FGF11$
$FGF12B$	$FGF7$	$FRAG1\,FGF$	$FGF9$
$FOP\,FGFR1$	$FGF20$	$THBS1$	$SERPINA1$
$STK12$	$CDK7$	$DIP1$	$TRF4-2$
$TGFBR3$	$EKI1$	$TPD52L$	$RGS5$
$GPR51$	$PKIG$	$SERPINA4$	$ADAM12$
$CAMK2B$	$RNAHP$	$TMSB10$	$NRBP$
MET	$GDF1$		

4.1.1 Risultati computazionali

Per prima cosa, si è applicato al set di dati precedentemente descritto la procedura per il calcolo della Granger-causalità multivariata. I risultati ottenuti hanno messo in luce come quest'ultima procedura non risulta essere informativa. Infatti tutti i valori ottenuti risultano pari a a zero. Questa osservazione, già accennata nella motivazione relativa all'introduzione del partial conditioning (par. 3.4), è stata dunque osservata oltre che dal lato teorico anche da un punto di vista pratico. Si è visto, come precedentemente detto, che i valori ottenuti per la causalità di Granger multivariata, nel caso di relazioni di causalità tra i 94 geni delle cellule ‘‘HeLa’’ da noi selezionati, risultavano essere pari tutti a zero (fig. 4.2). Questo risultato è dovuto ai

pochi dati a disposizione e alla mancata possibilità di eliminare le relazioni di causalità indiretta tra i geni.

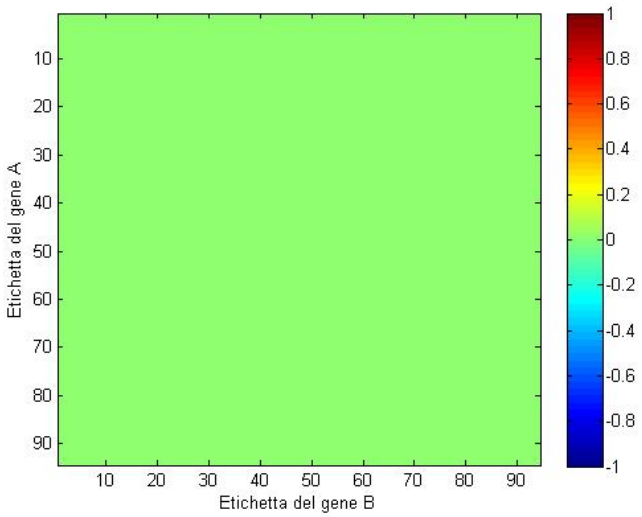


Figura 4.2: Valori della Granger-causalità multivariata (o condizionata) calcolati per il set di valori relativi ai 94 geni più informativi delle callule ‘‘HeLa’’. Le etichette usate per ogni gene sono state definite nella tabella 4.1.

Per ovviare a tale risultato negativo si è applicata la procedura del partial conditioning allo stesso set di dati. Detta procedura è stata impiegata variando il parametro n_d da un valore minimo pari a 0, caso in cui studieremo la causalità di Granger bivariata, ad un massimo valore pari a 5 (si ricordi che il parametro n_d è pari al numero di variabili considerate che risultano essere le più informative per la variabile condizionante e che il calcolo della causalità di Granger tramite l'uso del partial conditioning sfrutta l'equazione 3.4.3). I risultati ottenuti tramite tale processo risultano informativi rispetto ai precedenti (in quanto non sono tutti pari a zero) e mettono in luce le relazioni di causalità che sussistono tra i geni selezionati nelle cellule ‘‘HeLa”.

Inizialmente si sono calcolati i valori della causalità di Granger, con l'utilizzo del partial conditioning, tra un gene ed un altro del sistema, variando il numero di variabili più informative da selezionare (da 0 a 5). In figura 4.7 sono mostrate le rappresentazioni grafiche della matrice $\mathscr{P}$, i cui elementi p_{ij}, sono pari alla causalità (sotto applicazione del partial conditioning, cal-

colati tramite la relazione (3.4.3)) che il nodo j del network di geni esercita sul nodo i dello stesso. I valori risultano molto prossimi a zero eccetto che per pochi geni (per i quali detti valori non superano mai lo 0.3).

Detto ciò, è risultato interessante osservare la quantità d'informazione in ingresso ed in uscita per ogni gene. Tali quantità potranno essere ottenute sommando sulle righe della matrice $\mathscr{P}$, ottenendo in tal modo l'informazione che arriva al nodo i, e sommando sulle colonne della stessa, ottenendo l'informazione che parte dal nodo j. Tramite tale procedura si ottengono due funzioni che ad ogni gene associano rispettivamente l'informazione in uscita e quella in entrata (fig. 4.8). In tal maniera sarà possibile inoltre indicare quale sarà il gene più informativo e quale il gene che riceverà più informazione in assoluto. Il gene più informativo in assoluto risulta essere il $c-myc$ mentre quello che riceve più informazione risulta essere il $BRCA1$. Il $c-myc$ risulta essere un gene molto importante all'interno di una cellula. Esso è un oncogene dovuto ad una mutazione del gene Myc, che codifica proteine che si legano a sequenze di DNA appartenenti al dominio di altri geni. A causa dell'importanza di tale gene, la variazione di quest'ultimo porta ad un'espressione disordinata di molti geni alcuni dei quali dedicati alla proliferazione cellulare, da cui l'insorgenza di neoplasie. D'altro canto invece il $BRCA1$ risulta essere un gene oncosoppressore, che codifica una proteina, che interviene nel ciclo cellulare. Dunque all'interno del network di geni selezionati nelle "HeLa" il gene più informativo risulta essere il gene, nella sua variante oncogena, che codifica proteine adatte a codificare un numero molto alto di geni. Mentre il gene che riceve più informazione risulta essere un gene oncosoppressore che interviene nel ciclo cellulare.

Per capire se il comportamento di un singolo gene risulta più propendente al ricevere piuttosto che al fornire informazione è opportuno calcolare per ognuno di essi il rapporto tra informazione in uscita e informazione in ingresso. I risultati al variare di n_d sono mostrati in figura 4.9. Da quanto emerge dai risultati ottenuti da data procedura il $c-myc$ risulta essere il gene il cui rapporto tra informazione in uscita ed in ingresso è il più alto.

In ultima analisi si è osservato come si comporta mediamente il sistema al variare di n_d, ossia se in media i geni, che rappresentano i nodi del network da noi considerato, risultano essere più propensi a cedere o acquisire

informazione. Ciò che si è osservato è che i valori ottenuti (fig. 4.3) sono prossimi a 1, risultato che implica mediamente un comportamento analogo in acquisizione e rilascio d'informazione.

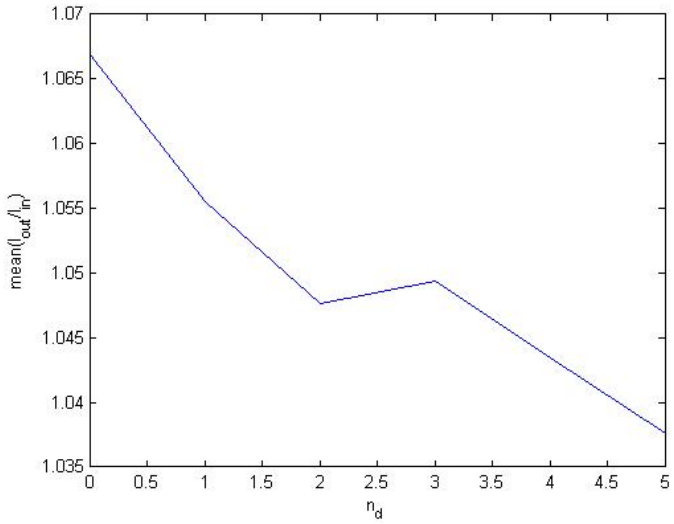


Figura 4.3: Variazione della media di I_{out}/I_{in} al variare di n_d per i dati genetici.

4.2 Attività celebrale

La Risonanza Magnetica Funzionale, abbreviata RMF o fMRI (Functional Magnetic Resonance Imaging), è una tecnica di imaging biomedico che consiste nell'uso dell'imaging a risonanza magnetica per valutare la funzionalità di un organo o un apparato, in maniera complementare all'imaging morfologico. Sebbene col termine risonanza magnetica funzionale si possa indicare una qualsiasi tecnica di imaging a risonanza magnetica che dia informazioni aggiuntive rispetto alla semplice analisi morfologica, essa è spesso usata per indicare la risonanza magnetica funzionale neuronale, ossia una delle tecniche di neuroimaging funzionale di sviluppo più recente. Questa tecnica è in grado di visualizzare la risposta emodinamica correlata all'attività neuronale del cervello o del midollo spinale nell'uomo o in altri animali.

Con lo sviluppo della scienza medica si è osservato che le variazioni del flusso sanguigno e dell'ossigenazione sanguigna nel cervello (emodinamica) sono strettamente correlate all'attività neurale. Quando le cellule nervose

sono attive, consumano l'ossigeno trasportato dall'emoglobina dei globuli rossi che attraversano i capillari sanguigni locali. Effetto di questo consumo di ossigeno è un aumento del flusso sanguigno nelle regioni dove si verifica maggiore attività neurale, che avviene con un ritardo da 1 a 5 secondi circa. Tale risposta emodinamica raggiunge un picco in 4-5 secondi, prima di tornare a diminuire fino al livello iniziale (in genere scende anche sotto di esso): si hanno così, oltre che variazioni del flusso sanguigno cerebrale, anche modificazioni localizzate del volume sanguigno cerebrale e della concentrazione relativa di emoglobina ossigenata ed emoglobina non ossigenata.

L'emoglobina risulta essere diamagnetica quando ossigenata e paramagnetica quando non ossigenata dunque, il segnale dato dal sangue nella risonanza magnetica nucleare (RMN) varia in funzione del livello di ossigenazione. Questi differenti segnali possono essere rilevati usando un'appropriata sequenza di impulsi RMN, ad esempio il contrasto Blood Oxygenation Level Dependent (BOLD). Maggiori intensità del segnale BOLD derivano da diminuzioni nella concentrazione di emoglobina non ossigenata. Mediante analisi con scanner per imaging a risonanza magnetica, usando parametri sensibili alla variazione della suscettività magnetica, è possibile stimare le variazioni del contrasto BOLD, che possono risultare di segno positivo o negativo in funzione delle variazioni relative del flusso sanguigno cerebrale e del consumo d'ossigeno. Incrementi del flusso sanguigno cerebrale, in proporzione superiori all'aumento del consumo d'ossigeno, porteranno ad un maggiore segnale BOLD; viceversa, diminuzioni nel flusso, di maggiore entità rispetto alle variazioni del consumo d'ossigeno, causeranno minore intensità del segnale BOLD.

Dopo questa breve introduzione, volta a spiegare su quali principi si basa la risonanza magnetica funzionale, delineiamo il caso oggetto del nostro studio. Considereremo l'esame di un paziente effettuato mentre dorme, suddividendo il cervello in 90 regioni corticali e subcorticali (tabella 4.2). Per ogni regione campioneremo l'attività celebrale, con frequenza pari ad 1 secondo, ottenendo un set di dati pari a 500 tempi per ognuna delle 90 regioni [Ma10]. Successivamente filtreremo le misure e applicheremo a tale set di dati la procedura per il calcolo della Granger-causalità e la procedura del partial conditioning.

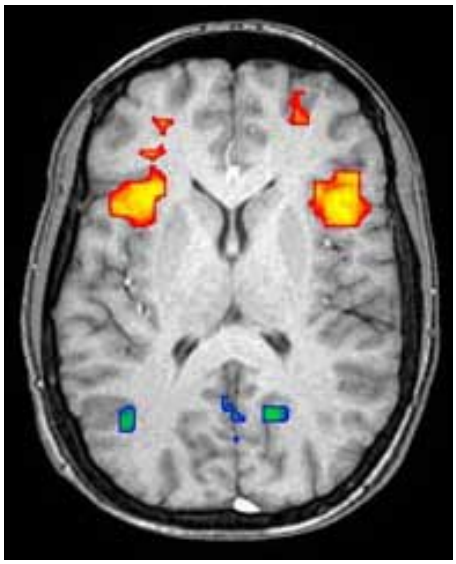


Figura 4.4: Esempio di imaging ottenuto tramite la risonanza magnetica funzionale con la tecnica di contrasto BOLD (Blood Oxygenation Level Dependent). Le zone rosse denotano zone con un'aumentata attività celebrale mentre le zone blu denotano zone con attività celebrale in diminuzione.

4.2.1 Risultati computazionali

In via preliminare si è provveduto ad eseguire una procedura di decimazione sulla variabile temporale del set di dati. Il motivo di tale procedimento sul dataset è stato l'interesse nel verificare come il sistema si comporta al variare della scelta della scala temporale di misurazione dell'attività celebrale. Per meglio dire l'interesse nel verificare come i risultati sarebbero cambiati a seconda che si fosse scelto di misurare l'attività celebrale a cadenza di un secondo o di dieci secondi o di una scala temporale diversa. Per tale motivo si è fatto variare il parametro di decimazione δ, ponendolo pari ad 1, 5, 10 e 20. Questa procedura consiste nel sostituire per un set di dati, relativi ad un gene, pari a δ, considerati sulla variabile temporale come i primi δ, i secondi δ, eccetera, il valor medio degli stessi. In tal maniera si otterrà un nuovo set di dati, dove ad ogni gene non corrisponderanno più 500 tempi ma $500/\delta$ tempi. Nel caso in cui ponessimo $\delta = 1$, considereremmo il campione intero di dati. Per ogni set di valori ottenuti tramite tale procedimento si è applicata la medesima metodologia utilizzata per il set di dati dei geni ''HeLa''.

In prima analisi si è calcolata la Granger-causalità multivariata tra le varie aree celebrali, considerando tutti i set ottenuti tramite decimazione del campione di dati iniziali. Anche in tal caso tutti i valori della causalità di Granger multivariata risultavano pari a zero (fig. 4.5). I motivi di tali risultati sono del tutto analoghi al caso precedente, cioè i pochi dati a disposizione e l'impossibilità di eliminare le relazioni di causalità indiretta.

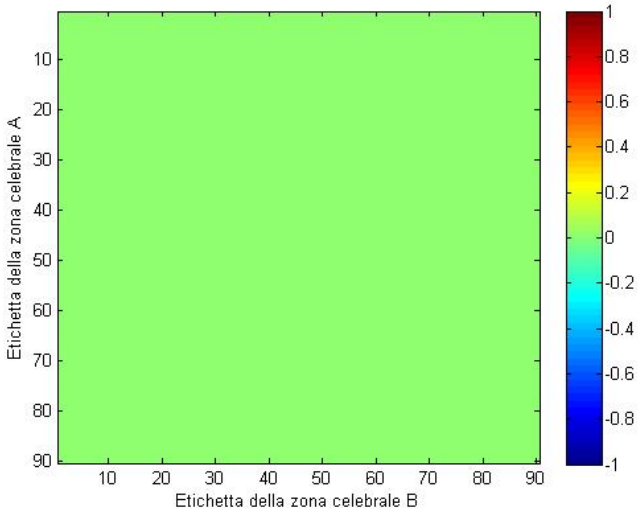


Figura 4.5: Valori della Granger-causalità multivariata (o condizionata) calcolati per il set di valori relativi alle 90 regioni in cui abbiamo suddiviso il cervello di un paziente dormiente. Le etichette usate per ogni gene sono state definite nella tabella 4.2. I risultati risultano essere indipendenti dalla decimazione effettuata sul campione.

Successivamente si sono calcolati per gli stessi set di dati i valori di Granger-causalità, sotto applicazione della metodologia di partial conditioning, osservando come detta causalità variasse al variare dei valori di n_d. I calcoli sono stati effettuati su tutti i set di dati ottenuti tramite decimazione. Una volta terminati i calcoli è stato possibile osservare poche interazioni interessanti tra le varie regioni celebrali a causa del fatto che la maggior parte delle interazioni davano valori della causalità di Granger pari o prossimi a zero (fig. 4.10).

Ancora una volta, dunque, per ottenere valori più significativi si è proceduto a calcolare l'informazione in uscita ed in ingresso per ogni singola area celebrale. I risultati di tale procedura (fig. 4.11) hanno messo in luce le zone celebrali che risultavano dare o ricevere più informazione. Per decimazioni

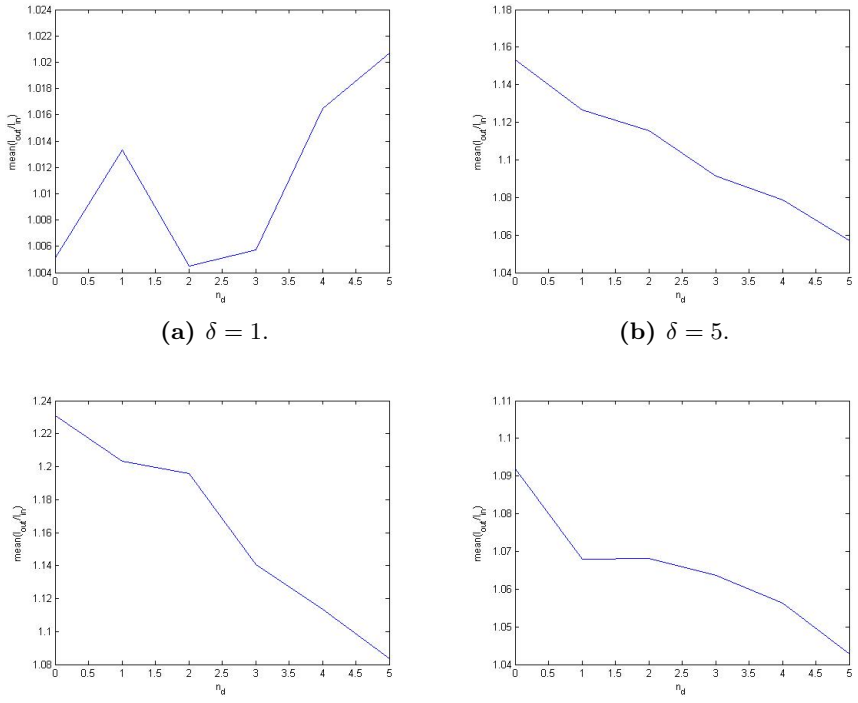
con parametri non molto elevati ($\delta = 1$ e $\delta = 5$) si è osservato che la regione più informativa risultava essere la $SupraMarginal_R$, mentre quella che riceveva più informazione in assoluto risultava la $Parietalsup_R$. La prima zona risulta, secondo le conoscenze attuali, collegata alla percezione e allo sviluppo della parola. Infatti, si è osservato che data zona risulta essere attivata quando si richiede al paziente, che si sottopone alla risonanza magnetica funzionale, di cercare le similitudini tra due parole o di spiegare il significato di una parola o di una frase. Data zona si attiva sia nel caso in cui si faccia leggere la parola o la frase sia nel caso in cui la si faccia ascoltare, il che fa presagire appunto una stretta connessione con la capacità di elaborazione della parola. La seconda, invece, risulta collegata all'orientamento spaziale e riceve, in un paziente sveglio, un gran numero di input visivi. Tutto ciò può far presagire la bontà dei risultati ottenuti.

Risulta interessante, però, osservare come per parametri di decimazione più alti ($\delta = 10$ o $\delta = 20$) i risultati vengano distorti portando, oltre che a zone informative diverse rispetto ai dati ottenuti per decimazioni più basse, anche a valori diversi al variare del parametro di partial conditioning n_d. Ciò potrebbe far presagire la possibilità di una dimensione caratteristica, come per la decimazione nel modello di Ising, sorpassata la quale l'informazione viene distorta.

Successivamente si è proceduto ad osservare come variasse il valore del rapporto tra informazione in uscita ed in ingresso per ogni area celebrale cercando quale fosse la zona che avesse valori più alti per dato rapporto. I risultati sono riportati in figura 4.12. In tal caso si è osservato che essendo i valori molto prossimi tra loro la scelta del parametro di partial conditioning n_d condizionava i risultati. Nonostante tale condizionamento, si è potuto osservare che la zona, che risultava fornire informazione in luogo di una minore ricezione della stessa, risultava essere la $Fusiform_L$. Essa risulta essere la zona celebrale che elabora le informazioni sui colori, processa i riconoscimenti facciali e i riconoscimenti delle parole. Tale risultato può essere confortante a causa della situazione di sonno del paziente da cui è stato prelevato il campione. A tutti gli effetti, durante la fase REM il cervello elabora le informazioni riguardanti il mondo in cui si svolgerà il sogno, per cui è plausibile che la zona $Fusiform_L$ possa essere vista come un centro

di elaborazione nel quale ad una piccola quantità d'informazione in ingresso corrisponde una quantità d'informazione in uscita molto più elevata.

In ultima analisi si è proceduto all'osservazione di come la media del rapporto tra informazione in uscita e informazione in ingresso su tutte le zone celebrali variasse al variare del parametro n_d. I risultati (fig. 4.6) sempre molto prossimi ad uno hanno mostrato, in media, un comportamento analogo di ogni area celebrale in ricezione e in uscita di informazione. Ciò implica il fatto che il cervello di un uomo che dorme è costituito in media da zone celebrali che forniscono tanta informazione quanta ne ricevono.



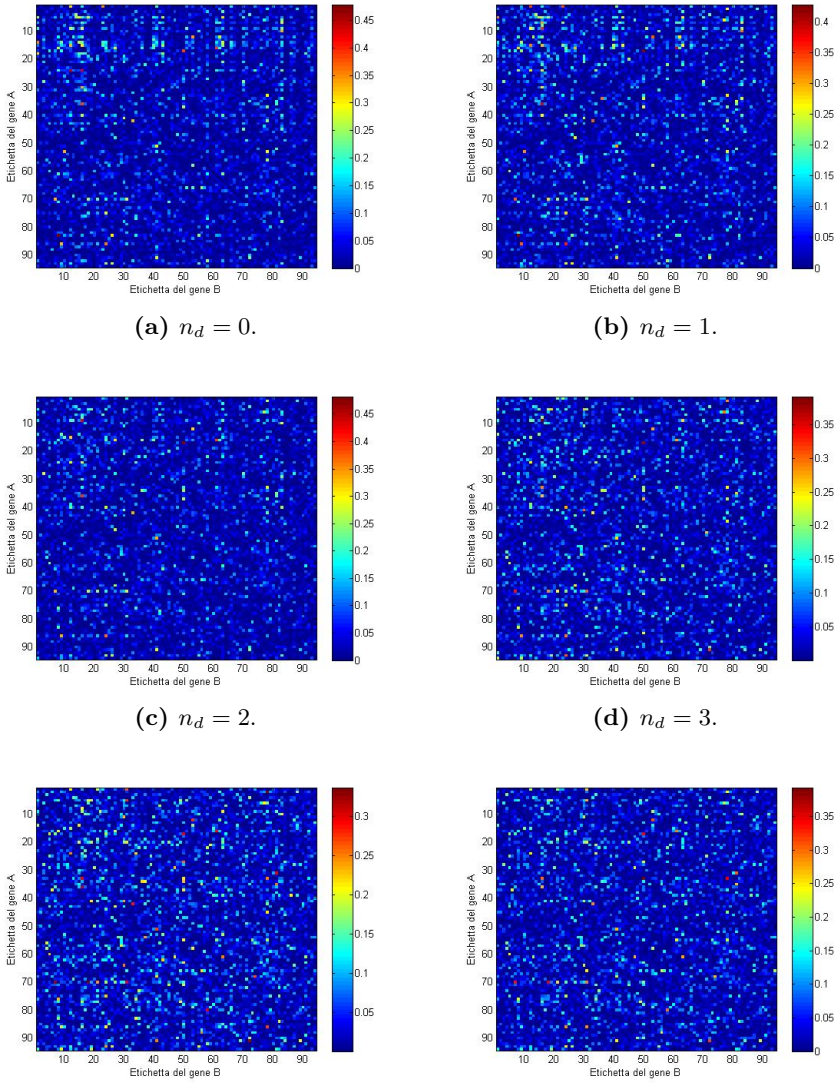
(a) $\delta = 1$. **(b)** $\delta = 5$.

(c) $\delta = 10$. **(d)** $\delta = 20$.

Figura 4.6: Variazione della media del rapporto tra informazione in uscita ed informazione in ingresso al variare del parametro n_d, a seconda del parametro di decomazione δ utilizzato.

Tabella 4.2: Nomi delle aree celebrali. Esse risultano etichettate in modo crescente (da 1 a 90) partendo da $Amygdala_L$ e procedendo lungo le righe.

$Amygdala_L$	$Amygdala_R$	$Angular_L$
$Angular_R$	$Calcarine_L$	$Calcarine_R$
$Caudate_L$	$Caudate_R$	$Cingulum_{Ant,L}$
$Cingulum_{Ant,R}$	$Cingulum_{Mid,L}$	$Cingulum_{Mid,R}$
$Cingulum_{Post,L}$	$Cingulum_{Post,R}$	$Cuneus_L$
$Cuneus_R$	$Frontal_{Inf,Oper,L}$	$Frontal_{Inf,Oper,R}$
$Frontal_{Inf,Orb,L}$	$Frontal_{Inf,Orb,R}$	$Frontal_{Inf,Tri,L}$
$Frontal_{Inf,Tri,R}$	$Frontal_{Med,Orb,L}$	$Frontal_{Med,Orb,R}$
$Frontal_{Mid,Orb,L}$	$Frontal_{Mid,Orb,R}$	$Frontal_{Mid,L}$
$Frontal_{Mid,R}$	$Frontal_{Sup,Medial,L}$	$Frontal_{Sup,Medial,R}$
$Frontal_{Sup,Orb,L}$	$Frontal_{Sup,Orb,R}$	$Frontal_{Sup,L}$
$Frontal_{Sup,R}$	$Fusiform_L$	$Fusiform_R$
$Heschl_L$	$Heschl_R$	$Hippocampus_L$
$Hippocampus_R$	$Insula_L$	$Insula_R$
$Lingual_L$	$Lingual_R$	$Occipital_{Inf,L}$
$Occipital_{Inf,R}$	$Occipital_{Mid,L}$	$Occipital_{Mid,R}$
$Occipital_{Sup,L}$	$Occipital_{Sup,R}$	$Olfactory_L$
$Olfactory_R$	$Pallidum_L$	$Pallidum_R$
$ParaHippocampal_L$	$ParaHippocampal_R$	$Paracentral_{Lobule,L}$
$Paracentral_{Lobule,R}$	$Parietal_{Inf,L}$	$Parietal_{Inf,R}$
$Parietal_{Sup,L}$	$Parietal_{Sup,R}$	$Postcentral_L$
$Postcentral_R$	$Precentral_L$	$Precentral_R$
$Precuneus_L$	$Precuneus_R$	$Putamen_L$
$Putamen_R$	$Rectus_L$	$Rectus_R$
$Rolandic_{Oper,L}$	$Rolandic_{Oper,R}$	$SuppMotorArea_L$
$SuppMotorArea_R$	$SupraMarginal_L$	$SupraMarginal_R$
$Temporal_{Inf,L}$	$Temporal_{Inf,R}$	$Temporal_{Mid,L}$
$Temporal_{Mid,R}$	$Temporal_{Pole_Mid_L}$	$Temporal_{Pole_Mid_R}$
$Temporal_{Pole_Sup_L}$	$Temporal_{Pole_Sup_R}$	$Temporal_{Sup_L}$
$Temporal_{Sup_R}$	$Thalamus_L$	$Thalamus_R$



(a) $n_d = 0$.

(b) $n_d = 1$.

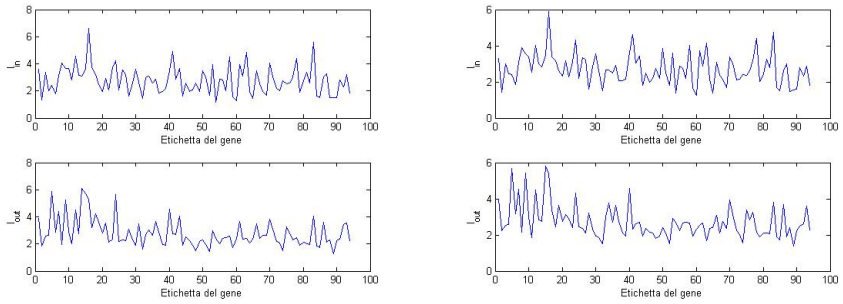
(c) $n_d = 2$.

(d) $n_d = 3$.

(e) $n_d = 4$.

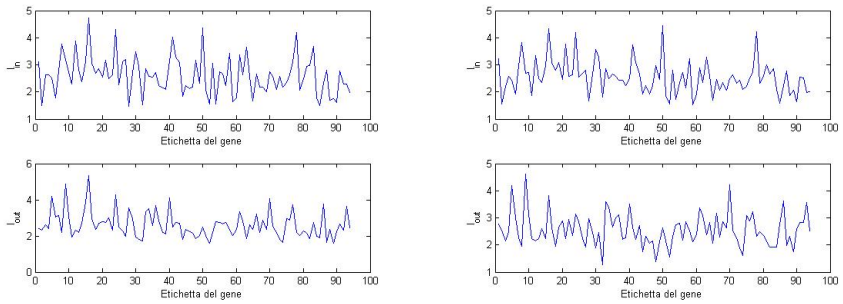
(f) $n_d = 5$.

Figura 4.7: Valori della Granger-causality con l'utilizzo del partial conditioning calcolata per diversi valori di n_d, numero di variabili più informative selezionate. Le etichette usate per ogni gene sono state definite nella tabella 4.1.



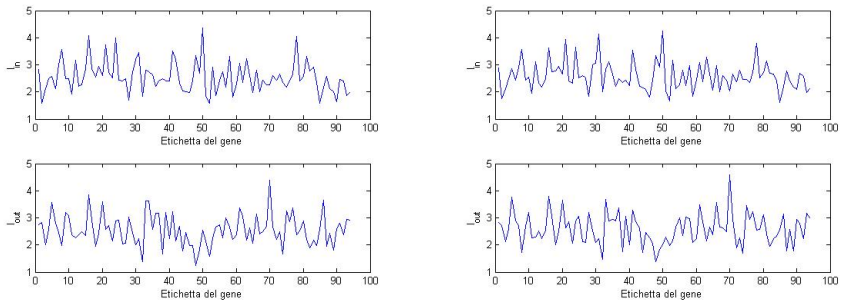
(a) $n_d = 0$.

(b) $n_d = 1$.



(c) $n_d = 2$.

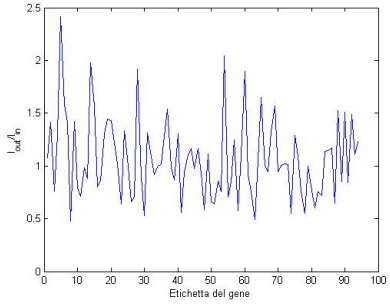
(d) $n_d = 3$.



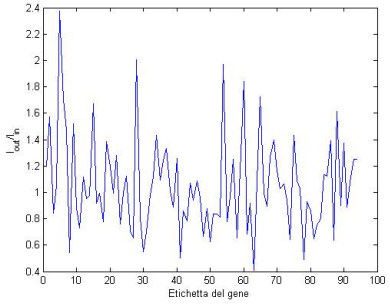
(e) $n_d = 4$.

(f) $n_d = 5$.

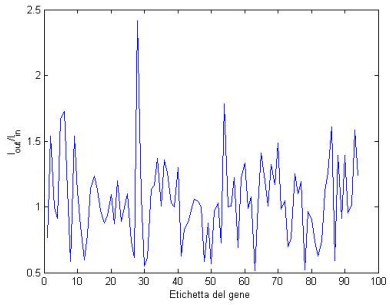
Figura 4.8: Valori dell'informazione in ingresso ed uscita per ogni gene calcolati per diversi valori di n_d, numero di variabili più informative selezionate. Le etichette usate per ogni gene sono state definite nella tabella 4.1.



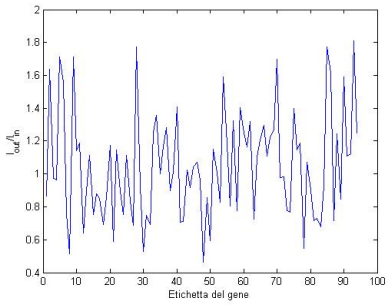
(a) $n_d = 0$.



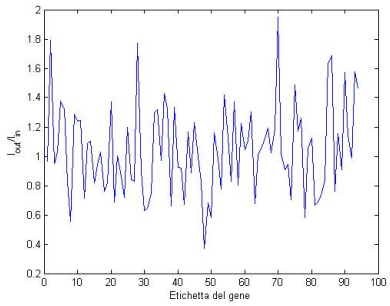
(b) $n_d = 1$.



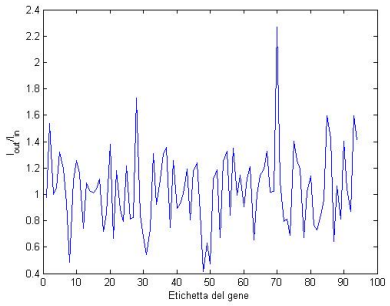
(c) $n_d = 2$.



(d) $n_d = 3$.

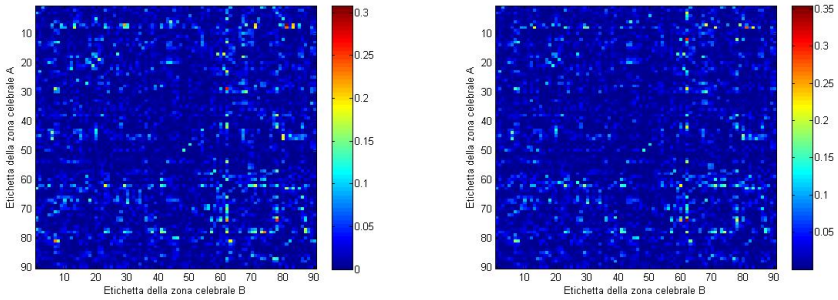


(e) $n_d = 4$.



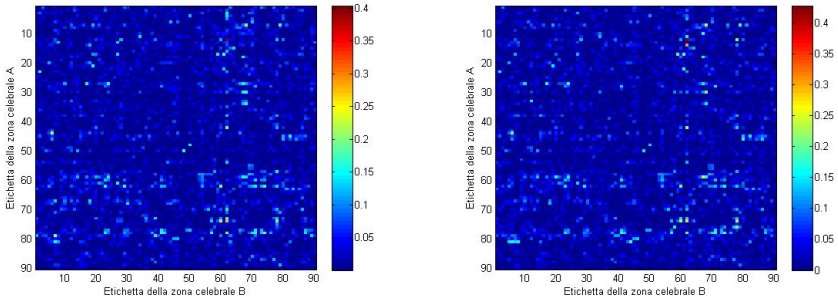
(f) $n_d = 5$.

Figura 4.9: Valori del rapporto tra informazione in uscita ed informazione in ingresso calcolati per ogni gene usando la procedura di partial conditioning per diversi valori di n_d, numero di variabili più informative selezionate. Le etichette usate per ogni gene sono state definite nella tabella 4.1.



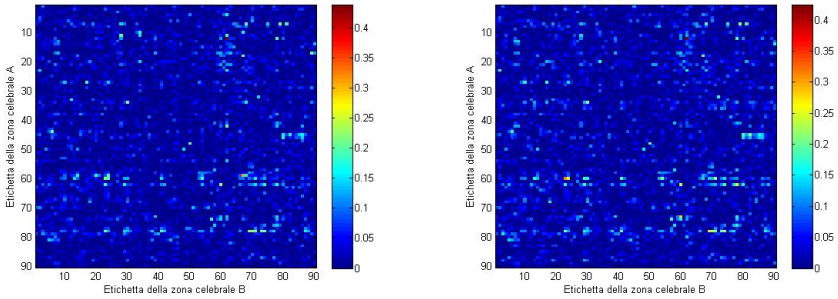
(a) $n_d = 0$.

(b) $n_d = 1$.



(c) $n_d = 2$.

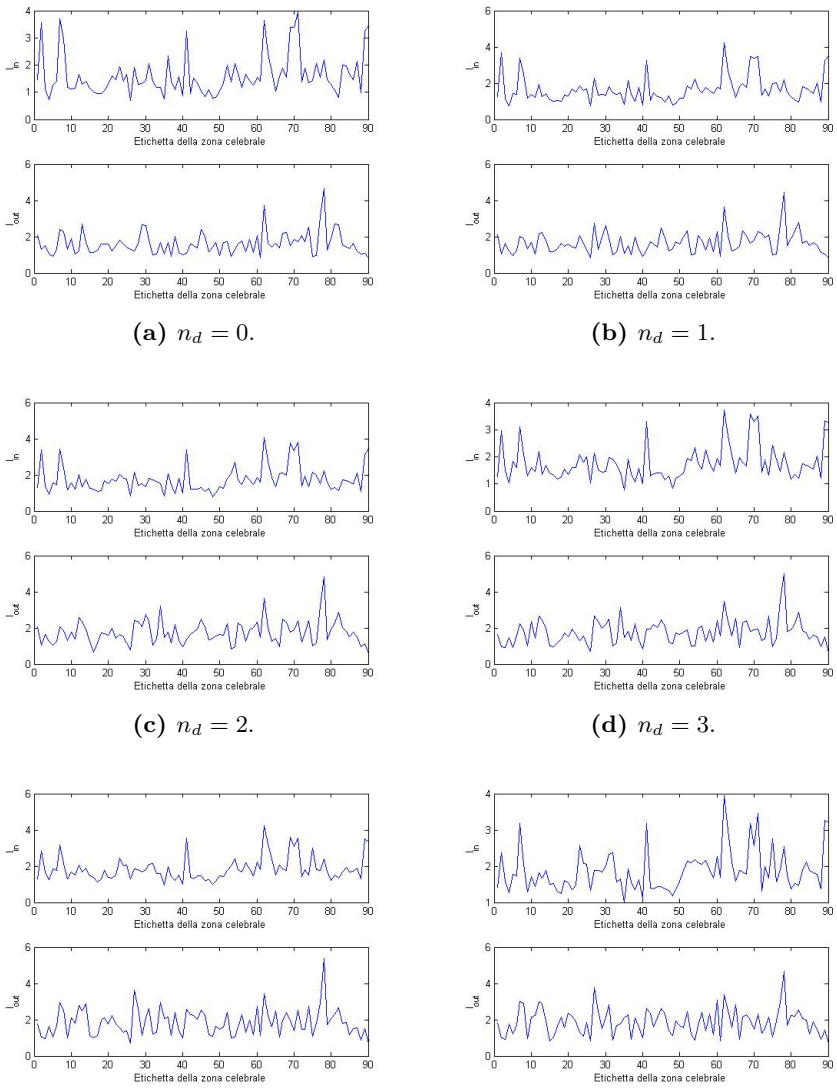
(d) $n_d = 3$.



(e) $n_d = 4$.

(f) $n_d = 5$.

Figura 4.10: Valori della Granger-causality, con l'utilizzo del partial conditioning, calcolati per diversi valori di n_d relativamente al campione di misurazioni della risonanza magnetica funzionale con parametro di decimazione δ pari ad uno, cioè considerando il campione completo. Le etichette usate per ogni area celebrale sono state definite nella tabella 4.2.



(a) $n_d = 0$.

(b) $n_d = 1$.

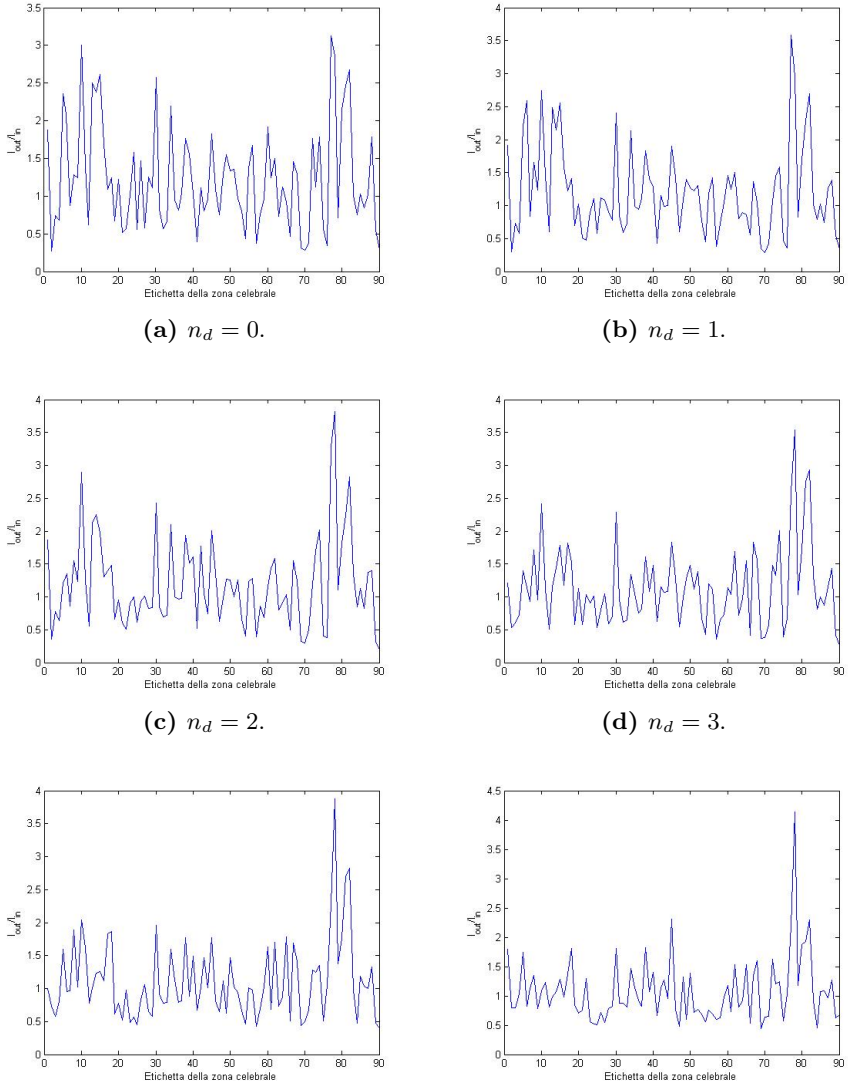
(c) $n_d = 2$.

(d) $n_d = 3$.

(e) $n_d = 4$.

(f) $n_d = 5$.

Figura 4.11: Valori dell'informazione in ingresso ed uscita per ogni zona celebrale calcolati per diversi valori di n_d relativamente al campione ottenuto tramite procedura di decimazione effettuate con parametro $\delta = 5$. Le etichette usate per ogni area celebrale sono state definite nella tabella 4.2.



(a) $n_d = 0$. **(b)** $n_d = 1$.

(c) $n_d = 2$. **(d)** $n_d = 3$.

(e) $n_d = 4$. **(f)** $n_d = 5$.

Figura 4.12: Rapporto tra informazione in uscita ed informazione in ingresso, per ogni zona celebrale, calcolati per diversi valori di n_d. Per procedere nei calcoli si è applicata al campione la procedura di decimazione con parametro $\delta = 20$. Le etichette usate per ogni area celebrale sono state definite nella tabella 4.2.

Capitolo 5

Conclusioni

Durante l'evoluzione di un sistema complesso l'informazione interna allo stesso viene trasferita tra i vari elementi che appartengono al network, creando relazioni di causalità tra i vari enti a seconda che l'informazione arrivi o esca. In questo lavoro di tesi si è cercato di dare un senso a questa affermazione, osservando gli strumenti opportuni per misurare lo scambio d'informazione.

Nella parte iniziale di questa tesi è stato introdotto il concetto stesso di sistema complesso, focalizzando l'attenzione sulla procedura descrittiva utilizzata per descriverlo. A tal fine è stata osservata la teoria dei grafi che mette in luce, attraverso strutture geometriche composte da linee e punti, le relazioni tra gli elementi del sistema complesso cui la struttura geometrica fa riferimento. Tramite tale teoria sono stati introdotti, dunque, tutti gli strumenti atti a descrivere un sistema geometrico qualsiasi.

Successivamente, attraverso la teoria dell'informazione e degli strumenti statistici di calcolo della causalità in un sistema (causalità di Granger e transfer entropy), sono state definire le procedure di calcolo per osservare i flussi di informazione tra le varie parti di un sistema. L'introduzione dell'entropia d'informazione e della mutua informazione ha permesso di vedere rispettivamente la procedura da seguire per calcolare l'informazione necessaria per prevedere il comportamento di una variabile e l'informazione che una variabile dà di un'altra. La mutua informazione, però, essendo simmetrica, ha portato alla necessità di un'estensione della teoria in quanto non permet-

teva di osservare il flusso direzionale di informazione tra due elementi interni al sistema complesso. A tal fine è stata avviata una trattazione relativa alla teoria della causalità. Preliminarmente veniva preso in prestito dalla teoria dell'econometria il concetto di causalità di Granger, a dispetto dell'opinione dello stesso Granger, il quale sosteneva che l'utilizzo dalla sua versione di causalità dovesse restare all'interno dell'ambito econometrico. Successivamente veniva introdotta la transfer entropy, che rappresenta un migliore strumento per effettuare tale calcolo in quanto la causalità di Granger risulta valida solo per variabili che seguono la distribuzione normale. Relativamente alla causalità di Granger è stata inoltre definita la procedura di partial conditioning che risulta essere utile nel caso in cui il campione in possesso fornisca pochi valori.

Infine si è proceduto all'applicazione di questi concetti in ambito materiale, applicandoli a dei dati relativi alle concentrazioni dei geni nella coltura cellulare ""HeLa"" e all'attività di alcune aree celebrali misurata tramite la Risonanza Magnetica Funzionale. Entrambe le applicazioni hanno messo in luce la necessità dell'utilizzo del partial conditioning per ottenere risultati informativi relativamente allo scambio di informazione nei sistemi complessi. Tutto ciò ha portato, dunque, all'osservazione della genuinità dell'introduzione della procedura di partial conditioning nel calcolo della causalità di Granger.

Bibliografia

[Al02] Albert, R. - Barabasi, A. *Statistical mechanics of complex networks.* Reviews of Modern Physics, vol. 74, pp. 47-97, 2002.

[Ba09] Barnett, L. *Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables.* Physical Review Letters, vol. 103, 238701, 2009.

[Ba99] Barabasi, A. - Albert, R. *Emergence of scaling in random networks.* Science, vol. 286, pp. 509-512, 1999.

[Bo06] Boccaletti, S. - Latora, V. *Complex networks: Structure and dynamics.* Physical Reports, vol. 424, pp. 175-308, 2006.

[Co91] Cover, T. *Elements of Information Theory.* John Wiley & Sons, Inc., 1991.

[Dr00] Dragulescu, E. T. *Statistical mechanics of money.* The European Physical Journal, Springer-Verlag, 2000.

[Fu07] Fujita, A. et al. BCM Syst. Biol., vol. 1, 39, 2007.

[Gr69] Granger, C. W. J. *Investigating causal relations by econometric models and cross-spectral methods.* Econometrica 37, pp. 424-438, 1969.

[Ha95] Hamilton, J. D. *Econometria delle serie storiche.* Monduzzi Editore, 1995.

[Ja57] Jaynes, E. T. *Information Theory and Statistical Mechanics.* The Physical Review, vol. 106, pp. 620-630, 1957.

[Ma08] Marinazzo, D. - Pellicoro, M. - Stramaglia, S. *Kernel Method for Nonlinear Granger Causality.* Phys. Rev. Let. 100, art. 14, 2008.

[Ma10] Marinazzo, D. - Liao, W.- Pellicoro, M. - Stramaglia, S. *Grouping time series by pairwise measures of redundancy.* Phys. Rev. Let. A, 374, 39, 4040-4044, 2010.

[Ma12] Marinazzo, D. - Pellicoro, M. - Stramaglia, S. *Causal Information Approach to Partial Conditioning in Multivariate Data Sets.* Computational and Mathematical Methods in Medicine, vol. 2012, Hindawi, 2012.

[Mi01] Milakovic, M. *A Statistical Equilibrium Model of Wealth Distribution.* Computing in Economics and Finance 2001, Society for Computational Economics, 2001.

[Mi67] Milgram, S. *The Small-World Problem.* Psychol. Today, vol. 1, 1967.

[Sc00] Schreiber, T. *Measuring Information Transfer.* Phys. Rev. Let. 85, pp. 461-464, 2000.

[Sh48] Shannon, C. *A Mathematical Theory of Communication.* Bell System Technical Journal, vol. 27, pp. 379-423, 623-656, July, October, 1948.

[Su12] Sugihara, G. - May, R. et al. *Detecting Causality in Complex Ecosystems.* Science, vol. 338, pp. 496-500, October, 2012.

[Wa98] Watts, D. J. - Strogatz, S. H. *Collective Dynamics of Small-World Networks.* Nature, vol. 393, 1998.

[Wh02] Whitfield, M. L. et al. Mol. Biol. Cell., vol. 13, 2002.

[Ze70] Zemansky, M. *Calore e Termodinamica.* Zanichelli, Bologna, 1970.