Alma Mater Studiorum - Università di Bologna

Scuola di Dottorato in Scienze Economiche e Statistiche

Dottorato di Ricerca in
Metodologia Statistica per la Ricerca Scientifica
XXIII ciclo

INDICATORI DI CORRELAZIONE E DI DISORDINE BASATI SUL CONCETTO DI ENTROPIA

Fabio Bertozzi

Dipartimento di Scienze Statistiche "P. Fortunati"
Marzo 2011

Alma Mater Studiorum - Università di Bologna

Scuola di Dottorato in Scienze Economiche e Statistiche

Dottorato di Ricerca in
Metodologia Statistica per la Ricerca Scientifica
XXIII ciclo

INDICATORI DI CORRELAZIONE E DI DISORDINE BASATI SUL CONCETTO DI ENTROPIA

Fabio Bertozzi

Coordinatore:
Prof.ssa Daniela Cocchi:

Tutor:
Prof. Rodolfo Rosa

Settore Disciplinare: SECS-S/01

Dipartimento di Scienze Statistiche "P. Fortunati"
Marzo 2011

Indice

Scopo e struttura della tesi

Negli ultimi anni sono stati sviluppati diversi indicatori e diversi metodi che utilizzano i concetti di entropia mutuati dalla teoria dell'informazione. Alcuni di essi sono finalizzati alla misura del livello di dipendenza, di autocorrelazione e di irregolarità (imprevedibilità e comportamento casuale) delle fluttuazioni di una serie o di più serie fra di loro (dipendenza, correlazione e sincronismo).

Lo sviluppo e l'utilizzo di questo tipo di indicatori è motivato dal fatto che le misure e i test statistici di dipendenza più comunemente usati fanno in genere riferimento ad opportune funzioni di correlazione basate su relazioni lineari che coinvolgono variabili continue e/o processi Gaussiani. Tali misure tendono però a risultare inefficaci nel caso di variabili discrete, relazioni non lineari o processi non Gaussiani.

Lo scopo della presente tesi è quindi quello di verificare il comportamento di alcuni di questi indicatori basati sul concetto di entropia, applicandoli a serie binarie opportunamente generate al fine di simulare situazioni le più diverse fra loro. In particolare, verranno presi in considerazione alcuni indicatori che hanno una lunga storia ed una letteratura ben consolidata alle spalle, come la mutua informazione, l'entropia congiunta e l'entropia relativa o distanza di Kullback Leibler, ed altri di più recente introduzione, quale l'entropia di Granger, Maasoumi e Racine, indicata con $S\rho$ o di utilizzo più settoriale, come l'entropia approssimata di Pincus (Approximate Entropy, indicata come $ApEn$).

Si tratterà quindi di comparare in maniera sistematica il comportamento di questi indicatori mettendone in luce limiti e potenzialità ed individuando, volta per volta, quelli più affidabili e performanti. Il nostro obiettivo, comunque, non è tanto quello di valutarne la bontà in termini di capacità di

discriminare fra dipendenza/indipendenza, ancorché non lineare. Questo, infatti, in ispecie per la mutua informazione e l'entropia di Granger, Maasoumi e Racine, è già stato ampiamente dibattuto e dimostrato in letteratura, come ad esempio in [10][20][29][5]. Qui si vorrebbe invece più propriamente verificare se e in che modo questi indicatori sono in grado di fornirci informazioni utili sulla struttura delle serie cui vengono applicati, sulla loro complessità e sul loro grado di disordine, con particolare attenzione alla *ApEn*, l'entropia approssimata di Pincus, che dovrebbe essere in grado di fornire indicazioni proprio riguardo il livello di casualità, di regolarità e di complessità di una serie. Per ultimo, infine, si vorrebbe approfondire l'analisi ed il significato della *ApEn* applicata non più ad una stessa serie, bensì a due serie diverse fra loro, misura che in letteratura prende il nome di *cross-ApEn* [27][41][43][44][56], nonchè proporne eventuali modifiche ed affinamenti. Il significato e l'interpretazione dei risultati forniti dalla *cross-ApEn*, infatti, non è sempre né facile né univoco e può facilmente portare a conclusioni fuorvianti, specie se utilizzati con leggerezza ed in maniera acritica.

Si è scelto di considerare unicamente dati binari in quanto questo tipo di dati è spesso utilizzato in letteratura per una prima verifica delle prestazioni e delle proprietà degli indicatori che si vanno caratterizzando. Ciò trova giustificazione nel fatto che le serie utilizzate, essendo costituite da due soli valori distinti, possono essere più agevolmente modulate sulle diverse esigenze sperimentali ed i risultati più facilmente previsti e verificati. Ciò si traduce, in sostanza, in un maggiore e più preciso controllo dei fattori sperimentali.

L'analisi delle serie binarie riveste inoltre una notevole importanza di per sè stessa. Non sono pochi infatti i casi in cui un determinato fenomeno può o deve essere dicotomizzato in moda da renderne più agevole o possibile l'analisi. Ad esempio serie storiche dicotomizzate vengono spesso utilizzate nell'analisi finanziaria per lo studio e la previsione del segno delle variazioni (guadagni o perdite) dei titoli azionari e dei mercati in genere [2], nonché della volatilità degli stessi [2][6]. Nel primo caso, il valore della variazione dell'indice considerato, misurato fra due generici istanti t e $t+1$, viene sostituito con un uno oppure con uno zero a seconda che si tratti di un guadagno (variazione positiva) o di una perdita (variazione negativa), nel secondo caso la serie viene dicotomizzata classificando le variazioni osservate come ad "alta volatilità" o a "bassa volatilità" in funzione del loro valore assoluto. In questo modo si può ridurre enormemente la complessità del sistema senza per altro pregiudicare il potenziale informativo dei dati stessi.

Un altro esempio di utilizzo di serie binarie si ha nel Data Mining, in particolare quando la mole dei dati, vuoi per il numero di serie storiche coinvolte, vuoi per la lunghezza delle stesse, è talmente elevata da rendere praticamente impossibile un'analisi adeguata a causa dei limiti fisici legati alla disponibilità di memoria ed alle prestazioni degli elaboratori elettronici normalmente disponibili. In questo caso la dicotomizzazione può avvenire, ad esempio, trasformando i valori delle serie in sequenze di zero e di uno a seconda che

ciascun dato si posizioni al di sopra o al di sotto della relativa media [3][54]. A differenza delle serie originarie, le serie storiche così trasformate possono essere compresse e manipolate in maniera molto efficiente.[1] Anche in questo caso la perdita di potere informativo dei dati è generalmente trascurabile: Bagnall, Ratanamahatana, Keogh e Janacek hanno dimostrato che, sotto alcune condizioni, il potere discriminante delle serie così dicotomizzate è asintoticamente equivalente a quello ottenibile con le serie originarie [3].

A seconda dei risultati ottenuti e delle considerazioni che ne deriveranno, potrà poi essere presa in considerazione la possibilità di estendere gli esperimenti qui realizzati anche a serie caratterizzate da dati di tipo diverso (interi, non interi, ecc.).

Per quanto concerne la sua struttura, la tesi si compone di quattro parti principali.

Nella prima parte (Parte I) verranno presentati i concetti fondamentali che sono alla base della teoria dell'informazione approfondendo, in particolar modo, alcuni indicatori più direttamente legati alla definizione di entropia e, quindi, alla distribuzione di probabilità degli elementi delle serie da analizzare, quali:

- la mutua informazione;
- l'entropia congiunta;
- l'entropia relativa, anche detta distanza di Kullback Leibler;
- la misura di Granger, Maasoumi e Racine, indicata come S_ρ;

nonchè

- l'entropia approssimata (Approximate Entropy), indicata come $ApEn$, e la Cross Approximated Entropy, indicata come $cross\text{-}ApEn$;

che si basano invece sulla presenza e sulla permanenza di determinati pattern (insiemi di elementi consecutivi) all'interno delle serie considerate.

Nella seconda parte (Parte II) gli indicatori di cui sopra verranno utilizzati per lo studio (autocorrelazione e livello di irregolarità) di serie binarie appositamente create, quali:

- serie perfettamente periodiche a cui sono state applicate diversi livelli di perturbazione casuale;
- serie perfettamente periodiche semplici con diversa lunghezza del periodo;

[1] Il fattore minimo di compressione è di 1 : 32 dato che, per il solo effetto dell'operazione di dicotomizzazione, la quantità di memoria altrimenti necessaria per rappresentare un generico elemento di una serie come numero reale, normalmente codificato con 4 byte (32 bit), si riduce ad un solo bit.

- serie complesse periodiche e non periodiche;
- serie generate da un processo stocastico;
- serie generate da un processo caotico.

Nella terza parte (Parte III), gli stessi indicatori verranno utilizzati per lo studio (correlazione e sincronismo) di coppie di serie binarie generalmente diverse fra di loro. In particolare, verranno incrociate fra loro tutte le serie ottenute a partire da una stessa serie perfettamente periodica cui sono stati applicati diversi livelli di perturbazione casuale.

Nella quarta parte (Parte IV), infine, si cercherà di modificare l'algoritmo che sta alla base del calcolo della *cross-ApEn* in modo che i suoi risultati possano essere utilizzati anche come misura di correlazione e non solo come generici indicatori di "sincronia".

Per una maggiore chiarezza e correttezza formale si fa presente che, nel corso della presente tesi, i termini correlazione ed autocorrelazione verranno utilizzati con un significato del tutto generale e svincolato da quello usuale di correlazione lineare. In particolare, con il termine "correlazione" si intenderà una relazione tra due variabili casuali tale che a ciascun valore della prima variabile corrisponda, con una certa regolarità, un valore della seconda e viceversa. Non si tratta necessariamente di un rapporto di causa ed effetto, ma semplicemente della tendenza di una variabile a variare in associazione con un'altra. La correlazione potrà essere diretta o positiva quando variando una variabile in un senso anche l'altra varierà nello stesso senso, inversa o negativa quando variando una variabile in un senso l'altra varierà in senso inverso. Una correlazione si dice spuria se lega due fenomeni che non hanno alcun nesso causale, indiretta quando due variabili X e Y sono correlate perché in realtà correlate entrambe ad una terza variabile Z.

Parte I

Entropia e fondamenti di teoria dell'informazione

CAPITOLO 1

Concetti generali

1.1 Premesse

L'uso della probabilità per quantificare l'avverabilità dei possibili esiti[1] di un determinato fenomeno aleatorio implica sempre un certo grado di incertezza. Se si esclude il caso banale in cui tutta la probabilità si addensa su di un unico risultato, rendendo così tale risultato certo e tutti gli altri impossibili, a priori non è infatti mai possibile pronosticare con esattezza il risultato di un fenomeno aleatorio.

Nota che sia la distribuzione di probabilità del fenomeno, saremo però in grado di esprimere, con minore o maggiore sicurezza, delle previsioni in merito ai possibili risultati del fenomeno stesso.

Intuitivamente, quando la maggior parte della probabilità è addensata su un unico o su pochi casi (possibili risultati), si avrà più "fiducia" sul loro verificarsi rispetto ad altri casi caratterizzati da bassa o bassissima probabilità. Se si dovesse scommettere su un possibile risultato, si sarebbe naturalmente portati a scommettere una cifra più alta sui risultati caratterizzati dalle probabilità più elevate piuttosto che su quelli dotati di probabilità più bassa.

Per contro, in un fenomeno i cui possibili esiti abbiano tutti la stessa identica probabilità (distribuzione uniforme) o probabilità molto simili fra loro, il dubbio sull'esito finale sarebbe sicuramente maggiore, per cui diffi-

[1] Per semplicità di esposizione, senza comunque perdere di generalità e se non altrimenti specificato, tali esiti si supporranno sempre essere di tipo discreto e in numero finito.

cilmente si sarebbe razionalmente portati a scommettere cifre consistenti su di un risultato piuttosto che su di un altro. Si può quindi affermare che alcune distribuzioni di probabilità esprimono maggiore incertezza, maggiore aleatorietà, rispetto ad altre.

Un problema che diversi studiosi si sono posti è stato quello di esprimere quantitativamente questo concetto di incertezza, definendo una qualche entità matematica che, data un certa distribuzione di probabilità, vi associasse un valore numerico che potesse essere interpretato come misura dell'incertezza espressa dalla distribuzione stessa.

Nelle pagine seguenti verranno brevemente illustrate alcune delle principali entità matematiche che sono state sviluppate per cercare di dare una risposta a questo problema.

1.2 L'entropia di Shannon

Il concetto di entropia (entropia termodinamica) fu introdotto da Rudolf Clausius. Egli per primo, nel suo "Trattato sulla teoria meccanica del calore" del 1864, utilizzò la parola entropia (dal greco $\varepsilon\nu$, "dentro" e $\tau\rho o\pi\eta$, "cambiamento") ad indicare che se un sistema, alla temperatura assoluta T (in gradi Kelvin), riceve una quantità di calore dQ, la sua "entropia" aumenta di una quantità $dS = dQ/dT$. L'entropia rappresenta quindi una misura del disordine di un sistema.

La meccanica statistica correla ancora più intimamente l'entropia al concetto di ordine tramite la relazione $S = k \log W$, dove k è la costante di Boltzmann e W (thermodynamische Wahrscheinlichkeit, probabilità termodinamica) rappresenta il numero di microstati distinti corrispondenti al medesimo stato macroscopico. La definizione meccanico-statistica è considerata la definizione fondamentale di entropia, dato che tutte le altre possono essere da essa matematicamente derivate, ma non viceversa.

In teoria dell'informazione, l'entropia è invece definita come la quantità di incertezza o di informazione associata ad una variabile aleatoria.

All'inizio degli anni 40 del '900 si pensava che aumentando la velocità di trasmissione di informazioni su di un canale di comunicazione dovesse aumentare di pari passo anche la probabilità di errore. Con i suoi studi riguardanti il modo in cui diversi segnali potevano essere compressi e trasmessi in maniera efficiente, Claude Shannon per primo dimostrò che ciò non era vero purché la velocità di trasmissione non fosse superiore alla capacità del canale. Egli intuì inoltre che processi casuali come la musica o il discorso avevano un determinato livello di complessità al di sotto del quale il segnale con cui tali grandezze venivano trasmesse non poteva essere ulteriormente compresso.

Shannon, su consiglio di John Von Neumann, chiamò tale complessità entropia, in omaggio al significato che questa parola aveva assunto in termodinamica.

Shannon racconto più tardi che una delle sue più grandi preoccupazioni fu come chiamare questo suo risultato. In un primo momento pensò di chiamarlo informazione, ma la parola era già fin troppo usata, così decise di chiamarlo incertezza. Quando discusse della cosa con Von Neumann, quest'ultimo ebbe un'idea migliore e gli suggerì di chiamarlo entropia, e questo 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" [45].

Nel suo lavoro sulla teoria matematica delle telecomunicazioni pubblicato nel 1948 sulla rivista "Bell System Technical Journal"[47], Shannon diede una definizione generale della misura H di tale entropia, ovvero dell'informazione o dell'incertezza associata ad un insieme di n possibili eventi, ciascuno caratterizzato da una determinata probabilità p_i.

Una tale misura $H(p_1, p_2, \ldots, p_n)$ avrebbe dovuto possedere almeno le tre seguenti fondamentali proprietà:

1. avrebbe dovuto essere una funzione continua delle probabilità p_i. In questo caso una qualsiasi piccola variazione nella distribuzione di probabilità $p_1, p_2, \ldots, p_n$ avrebbe dovuto condurre ad una corrispondente piccola variazione di H.

2. Se tutte le probabilità p_i fossero state uguali (distribuzione uniforme), cioè $p_i = 1/n$, con $i = 1, 2, \ldots, n$, allora H avrebbe dovuto essere una funzione monotonica crescente di n. Nel caso di eventi ugualmente probabili, infatti, tanto maggiore è il numero di eventi possibili, tanto maggiore è l'incertezza sul possibile risultato.

3. Se l'esito (evento elementare) associato ad un determinato fenomeno avesse potuto essere scomposto nella realizzazione di due scelte successive, l'entropia complessiva H del fenomeno avrebbe dovuto corrispondere alla somma pesata dei valori delle entropie relative ai vari eventi in cui il fenomeno era stato scomposto. In altre parole, H avrebbe dovuto essere funzione unicamente della distribuzione di probabilità riferita agli eventi elementari e non di come gli stessi avrebbero potuto venire eventualmente raggruppati all'interno della distribuzione.

Da un punto di vista matematico il punto 3 può essere meglio descritto e formalizzato mediante il seguente esempio:

Si consideri l'estrazione di un generico oggetto da un insieme di n possibili oggetti, ciascuno caratterizzato da una propria probabilità p_i. Si supponga poi che tale estrazione possa essere "scomposta" nella realizzazione di due

successive estrazioni, con una prima estrazione individuante un'urna j fra k possibili, con ciascuna urna caratterizzata da una probabilità di estrazione pari a w_j e contenente n_j degli n oggetti. In questo caso si avrebbero n possibili eventi elementari (l'estrazione di uno qualsiasi fra gli n oggetti), caratterizzati ciascuno da una probabilità p_i, con $i = 1, 2, \ldots, n$, ma raggruppati in k gruppi (le k urne) di probabilità w_j, con $j = 1, 2, \ldots, k$, con:

$$w_1 = \sum_{i=1}^{n_1} p_i; \quad w_2 = \sum_{i=n_1+1}^{n_2} p_i; \quad \ldots \quad w_k = \sum_{i=n_{k-1}+1}^{n} p_i \quad \text{e} \quad \sum_{j=1}^{k} n_j = n$$

allora:

$$H(p) = H(w) + \sum_{j=1}^{k} w_j H[\{p_i|w_j\}_j]$$

dove $H(p)$ è l'entropia della distribuzione di probabilità degli eventi elementari non aggregati, $H(w)$ è l'entropia riferita ai gruppi e $\{p_i|w_j\}_j$ esprime la distribuzione di probabilità condizionata relativa agli eventi elementari appartenenti all'i-esimo gruppo.

Shannon dimostrò che la sola misura, H, che soddisfa alle tre sopra enunciate proprietà ha forma [47]:

$$H = -k \sum_{i=1}^{n} p_i \log p_i$$

con k costante positiva che dipende, semplicemente, dall'unità di misura scelta. Ponendo $k = 1$ ed utilizzando il logaritmo in base due, Shannon definì la misura $H = -\sum_{i=1}^{n} p_i \log_2 p_i$ come *entropia* dell'insieme di probabilità $\{p_1, p_2, \ldots, p_n\}$.

Nel caso in cui X sia una variabile casuale, $H(X)$ ne indicherà pertanto l'entropia. In questo caso X non rappresenta l'argomento di una funzione, ma un simbolo necessario a differenziare, ad esempio, l'entropia della variabile casuale X, $H(X)$, da quella di Y, $H(Y)$.

Se si utilizza il logaritmo in base due, l'unità di misura dell'entropia sarà espressa in *bit* e l'entropia potrà essere pensata, anche, come il numero medio di cifre binarie necessarie a codificare la variabile casuale considerata.

L'entropia H è inoltre caratterizzata dalle seguenti importanti proprietà:

1. l'entropia di una distribuzione di probabilità $\{p_i\}$, con $i = 1, 2, \ldots, n$, è uguale a zero se e solo se tutte le probabilità p_i sono uguali a zero eccetto una sola che vale invece uno. Ciò significa che H assume valore zero unicamente quando vi è certezza del risultato mentre, in caso contrario, H è positiva;

2. l'entropia di una distribuzione di probabilità $\{p_i\}$, con $i = 1, 2, \ldots, n$, è massima ed uguale a $\log_2 n$ quando tutte le probabilità p_i sono uguali fra loro, ovvero quando $p_i = 1/n$ (distribuzione uniforme), situazione questa che esprime la massima incertezza;

3. ogni modifica del valore delle probabilità $\{p_i\}$ orientata verso il loro livellamento (ovvero verso il caso di equidistribuzione o distribuzione uniforme), determina un incremento del valore di H.

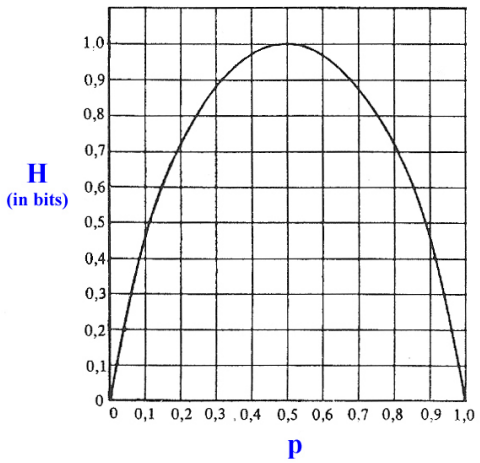

Figura 1.1: Entropia nel caso di due eventi di probabilità p e $(1-p)$ [48, p. 55].

1.2.1 L'entropia di Shannon come entropia termodinamica

Si può agevolmente dimostrare che l'entropia di Shannon è equivalente alla classica definizione di entropia utilizzata in termodinamica:

$$S(E) = \log N(E)$$

dove $N(E)$ indica il numero di microstati accessibili in funzione dell'energia E. Poiché si assume che i possibili microstati di uguale energia siano tutti ugualmente probabili, la probabilità dell'i-esimo microstato è pari a $p(i) = 1/N(E)$. Sostituendo nella formula di Shannon $1/N(E)$ a $p(i)$ si ha:

$$H = -\sum_{i=1}^{N} \frac{1}{N} \log \frac{1}{N}$$

che equivale, per le proprietà del logaritmo, a:

$$H = \sum_{i=1}^{N} \frac{1}{N} \log N \quad \text{e quindi} \quad H = \log N \sum_{i=1}^{N} \frac{1}{N}$$

che, dopo aver risolto la sommatoria, si riduce a:

$$H = \log N$$

che è proprio l'entropia termodinamica.

1.2.2 L'entropia di Shannon come sorpresa media

Se indichiamo con X una variabile casuale atta a rappresentare un generico fenomeno aleatorio, la quantità $-\log_2 p(x_i)$ è spesso indicata come "sorpresa" relativa al possibile evento x_i di X. Infatti, se $p(x_i)$ è piccola, noi saremmo certamente alquanto "sorpresi" dall'assistere al verificarsi dell'evento x_i ed invero la quantità $-\log_2 p(x_i)$ sarà grande. Se invece la probabilità $p(x_i)$ è alta, la "sorpresa" non potrà che essere bassa e ciò è correttamente indicato da un più basso valore della quantità $-\log_2 p(x_i)$.

E' quindi ragionevole assumere la quantità $-\log_2 p(x_i)$ come misura della "sorpresa" relativa al verificarsi del generico evento x_i, in quanto $-\log_2 p(x_i)$ è tanto più elevato quanto più è bassa la probabilità $p(x_i)$ associata ad x_i e viceversa.

L'entropia di Shannon può quindi essere vista anche come valore atteso della sorpresa, o sorpresa media, riferita ai possibili esiti di un determinato fenomeno aleatorio. Essa può essere infatti riscritta come:

$$H(X) = \sum_i \left[-\log_2 p(x_i)\right] p(x_i) = E\left[-\log_2 p(x_i)\right]$$

L'entropia indica quindi, in media, la sorpresa che avremmo nell'apprendere l'effettivo risultato di un determinato fenomeno casuale. Ciò rafforza il concetto di H come misura dell'incertezza associata ad una determinata distribuzione di probabilità: maggiore è l'incertezza associata ad un determinato fenomeno, maggiore sarà la sorpresa, in media, nell'apprendere il reale esito del fenomeno stesso.

1.3 Entropia di una variabile aleatoria multidimensionale: l'entropia congiunta

Date due variabili casuali X e Y che prevedano m possibili eventi la prima ed n possibili eventi la seconda, la probabilità congiunta che si verifichi contemporaneamente l'evento x_i per la variabile casuale X e l'evento y_j per la variabile Y è indicata con $p(x_i, y_j)$.

Estendendo alle variabili a due dimensioni i concetti sviluppati per le variabili aleatorie unidimensionali, l'entropia della variabile congiunta X, Y è definita nel modo seguente [8, p. 15]:

$$H(X,Y) = -\sum_i \sum_j p(x_i, y_j) \log_2 p(x_i, y_j)$$

Le entropie delle singole variabili X e Y possono poi calcolarsi come:

$$\begin{aligned} H(X) &= -\sum_i \sum_j p(x_i, y_j) \log_2 \sum_j p(x_i, y_j) \\ H(Y) &= -\sum_i \sum_j p(x_i, y_j) \log_2 \sum_i p(x_i, y_j) \end{aligned}$$

essendo $p(x_i) = \sum_j p(x_i, y_j)$ e $p(y_j) = \sum_i p(x_i, y_j)$ le probabilità marginali rispettivamente di X e di Y. Un ragionamento analogo vale per variabili n-dimensionali.

1.4 Entropia condizionale

Nel caso di due variabili casuali X ed Y aventi distribuizione di probabilità congiunta $p(x_i, y_j)$, la probabilità condizionata $p(y_j|x_i)$ che, fissato un determinato evento x_i di X, si verifichi un qualsiasi evento y_j di Y è data da:

$$p(y_j|x_i) = \frac{p(x_i, y_j)}{\sum_j p(x_i, y_j)}$$

ovvero

$$p(y_j|x_i) = \frac{p(x_i, y_j)}{p(x_i)}$$

dato che $p(x_i) = \sum_j p(x_i, y_j)$.

In questo contesto viene comunemente definita entropia condizionale di Y dato X [14, p. 16], $H(Y|X)$, la media pesata dell'entropia di Y dato ogni possibile evento x_i di X calcolata con pesi proporzionali alla probabilità di ottenere quel particolare evento x_i, ovvero:

$$H(Y|X) = -\sum_i \sum_j p(x_i, y_j) \log_2 p(y_j|x_i) \tag{1.1}$$

L'entropia condizionale $H(Y|X)$ misura l'incertezza media di Y quando è noto il valore di X. Sostituendo nella (1.1) a $p(y_j|x_i)$ la notazione equivalente $\frac{p(x_i, y_j)}{\sum_j p(x_i, y_j)}$ si ha:

$$\begin{aligned} H(Y|X) &= -\sum_i \sum_j p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{\sum_j p(x_i, y_j)} = \\ &= -\sum_i \sum_j p(x_i, y_j) \log_2 p(x_i, y_j) - \left[-\sum_i \sum_j p(x_i, y_j) \log_2 \sum_j p(x_i, y_j) \right] = \\ &= H(X, Y) - H(X) \end{aligned}$$

o, anche:

$$H(X, Y) = H(X) + H(Y|X)$$

e

$$H(X,Y) = H(Y) + H(X|Y)$$

L'incertezza (entropia) della variabile aleatoria congiunta X, Y è pertanto pari alla somma dell'incertezza della variabile X e di quella di Y quando X è noto. In generale si ha inoltre che $H(X|Y) \neq H(Y|X)$.

Nel caso di n variabili casuali $X_1, X_2, \ldots, X_n$ aventi densità di probabilità congiunta $p(x_1, x_2, \ldots, x_n)$, per la regola della catena per l'entropia [8, p. 21], si ha poi che:

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

1.5 Mutua informazione

Un'indicazione importante del legame eventualmente esistente fra due variabili aleatorie X ed Y è dato dalla mutua informazione [8, p. 18]:

$$I(X;Y) = \sum_i \sum_j p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \tag{1.2}$$

Applicando le regole dei logaritmi, la (1.2) può infatti essere riscritta nella forma:

$$\begin{aligned}
I(X;Y) &= \sum_i \sum_j p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \\
&= \sum_i \sum_j p(x_i, y_j) \log_2 p(x_i, y_j) + \sum_i \sum_j p(x_i, y_j) \log_2 \frac{1}{p(x_i)} + \\
&\quad + \sum_i \sum_j p(x_i, y_j) \log_2 \frac{1}{p(y_j)} \\
&= \sum_i \sum_j p(x_i, y_j) \log_2 p(x_i, y_j) - \sum_i p(x_i) \log_2 p(x_i) - \sum_j p(y_j) \log_2 p(y_j) \\
&= -H(X,Y) + H(X) + H(Y)
\end{aligned}$$

ovvero

$$\begin{aligned}
I(X;Y) &= H(X) + H(Y) - H(X,Y) \\
&= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X)
\end{aligned} \tag{1.3}$$

E' evidente, quindi, che la mutua informazione tra due variabili casuali non è nient'altro che la riduzione di incertezza di una variabile dovuta alla conoscenza dell'altra. Se la conoscenza di Y riduce la nostra incertezza su X, allora si dice che Y porta informazioni su X.

Se X e Y sono indipendentemente distribuite, ovvero se $p(x,y) = p(x)p(y)$, la mutua informazione fra le due variabili è zero. Inoltre, tale misura è simmetrica, cioè $I(X;Y) = I(Y;X)$, ed è sempre non negativa.

Si ha poi che

$$I(X;X) = H(X)$$

per cui, a volte, la mutua informazione è detta anche *autoinformazione*.

Si può calcolare, infine, anche la *mutua informazione condizionale* [8, p. 22] fra due variabili aleatorie X e Y data una terza variabile casuale Z:

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = \sum_i \sum_j \sum_k p(x_i, y_j, z_k) \log_2 \frac{p(x_i, y_j|z_k)}{p(x_i|z_k)p(y_j|z_k)}$$

Per la mutua informazione condizionale vale la regola della catena per l'informazione, ovvero [8, p. 22]:

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, X_{i-2}, \ldots, X_1)$$

1.5.1 Il condizionamento riduce l'entropia

Dalla precedente (1.3) si ha che $I(X;Y) = H(X) - H(X|Y)$. Poiché la mutua informazione è sempre non negativa possiamo scrivere:

$$H(X) - H(X|Y) \geq 0$$

per cui

$$H(X) \geq H(X|Y)$$

Ne discende, quindi, che la conoscenza di una variabile non può mai aumentare l'incertezza relativa all'altra variabile anzi, tale incertezza non potrà che diminuire a meno che X ed Y non siano fra loro indipendenti, nel qual caso essa non subisce variazioni. Da quanto sopra deriva immediatamente che:

$$H(X,Y) \leq H(X) + H(Y)$$

dove il segno di uguaglianza vale solamente nel caso di eventi fra loro indipendenti, ovvero quando $p(x_i, y_j) = p(x_i)p(y_j)$. Infatti

$$H(X,Y) = H(X) + H(Y|X) \leq H(X) + H(Y)$$

poiché

$$H(Y|X) \leq H(Y)$$

Più in generale [8, p. 28], date n variabili casuali $X_1, X_2, \ldots, X_n$ aventi funzione di probabilità congiunta $p(x_1, x_2, \ldots, x_n)$ si ha:

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

1.5.2 Disuguaglianza dell'elaborazione dell'informazione

Se X, Y, Z formano una catena di Markov, per definizione la distribuzione condizionale di Z dipende solo da Y essendo condizionalmente indipendente da X, ovvero $X \to Y \to Z$.

In questo caso si ha che [8, p. 32]:

$$I(X;Y) \geq I(X;Z)$$

Se $Z = g(Y)$, poiche è $X \to Y \to Z$, si ha che:

$$I(X;Y) \geq I(X;g(Y))$$

Ciò significa che nessuna elaborazione su Y, sia di tipo casuale che deterministico, può far aumentare l'informazione che Y contiene su X.

1.5.3 Disuguaglianza di Fano

Date due variabili casuali X ed Y, la disuguaglianza di Fano mette in relazione la probabilità di errore che si commette nel cercare di individuare il valore di X, noto che sia quello di Y, all'entropia condizionale $H(X|Y)$.

Poiché l'entropia condizionale di una variabile casuale X, fissata un'altra variabile casuale Y, è zero se e solo se X è funzione di Y, sarà possibile stimare X a partire da Y con una probabilità di errore pari a zero se e solo se $H(X|Y) = 0$. Analogamente, ci si potrà aspettare di poter stimare X con una bassa probabilità di errore se e solo se $H(X|Y)$ è bassa.

Si supponga perciò di voler stimare X a partire dalle osservazioni di una variabile aleatoria Y, correlata alla X tramite la distribuzione di probabilità condizionale $p(x|y)$ e sia $\widehat{X} = g(Y)$ tale stima. Può essere allora desiderabile calcolare gli estremi della probabilità per cui $\widehat{X} \neq X$.

Sia $P_e = P(\widehat{X} \neq X)$, ovvero la probabilità di errore nella stima, allora [8, p. 39]:

$$H(P_e) + P_e \log_2(|\mathcal{X}| - 1) \geq H(X|Y)$$

o anche, ma con un'approssimazione maggiore:

$$1 + P_e \log_2 |\mathcal{X}| \geq H(X|Y) \quad \text{ovvero} \quad P_e \geq \frac{H(X|Y) - 1}{\log_2 |\mathcal{X}|}$$

con $|\mathcal{X}|$ numero di valori distinti che la variabile casuale X può assumere.

1.6 Entropia relativa o distanza di Kullback Leibler

L'entropia relativa $D(p||q)$[2] è una misura della distanza fra due distribuzioni o, anche, una misura dell'inefficienza che si riscontra se si assume che la

[2] In questo contesto si utilizza il simbolo || ad indicare il confronto tra due distribuzioni di probabilità che, se si usasse il simbolo |, questo potrebbe essere facilmente confuso con quello di condizionamento di una variabile rispetto all'altra.

distribuzione di probabilità della variabile aleatoria X sia q quando invece la vera distribuzione è p.[3]

L'entropia relativa, o distanza di Kullback Leibler, fra due distribuzioni di probabilità p e q è definita[4] come [8, p. 18]:

$$D(p||q) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)}$$

L'entropia relativa è sempre non negativa, ovvero $D(p||q) \geq 0$, ed è zero se e solo se $p(x) = q(x)$ per ogni $x \in \mathcal{X}$.

L'entropia relativa non è tuttavia una vera misura di distanza in quanto non è simmetrica e non soddisfa la disuguaglianza triangolare $||\overrightarrow{a} + \overrightarrow{b}|| \leq ||\overrightarrow{a}|| + ||\overrightarrow{b}||$. Ciò non di meno è spesso utile pensare all'entropia relativa come alla "distanza" fra due distribuzioni.

La mutua informazione fra due variabili aleatorie X ed Y qualsiasi

$$I(X;Y) = \sum_i \sum_j p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$

altro non è che l'entropia relativa fra la distribuzione di probabilità congiunta delle variabili X e Y ed il prodotto delle relative distribuzioni di probabilità marginali, ovvero:

$$I(X;Y) = D\left[p(x,y)||p(x)p(y)\right]$$

L'entropia relativa condizionale è definita invece come [8, p. 22]:

$$D\left[p(y|x)||q(y|x)\right] = \sum_i p(x_i) \sum_j p(y_j|x_i) \log_2 \frac{p(y_j|x_i)}{q(y_j|x_i)}$$

Data l'entropia relativa e l'entropia relativa condizionale, vale la regola della catena per l'entropia relativa [8, p. 23], ovvero:

$$D\left[p(x,y)||q(x,y)\right] = D\left[p(x)||q(x)\right] + D\left[p(y|x)||q(y|x)\right]$$

1.7 Densità di entropia

Fino ad ora si è focalizzata l'attenzione sul concetto di entropia associato a variabili casuali ed a distribuizioni di probabilità. E' però possibile applicare tale concetto anche a sequenze di simboli (stringhe e serie storiche) [14].

[3] Se, per esempio, si conoscesse la vera distribuzione di probabilità p della variabile aleatoria X, per descrivere tale variabile si potrebbe costruire un codice di lunghezza media $H(p)$. Se si usasse invece un codice basato sulla distribuzione q, avremmo bisogno di $H(p) + D(p||q)$ bits in media per descrivere adeguatamente X.

[4] Si assume che $0 \log \frac{0}{q} = 0$ e $p \log \frac{p}{0} = +\infty$, in analogia al caso continuo dove $\lim_{p\to 0} p \log \frac{p}{q} = 0$ e $\lim_{q\to 0} p \log \frac{p}{q} = +\infty$.

Si consideri una stringa di lunghezza infinita $\overleftrightarrow{S} = \ldots X_{-1}, X_0, X_1, X_2, \ldots$, dove le X_i identificano altrettante variabili casuali aventi come possibili valori i simboli di un determinato alfabeto finito $\mathcal{X}$, cioè $X_i = x_i \in \mathcal{X}$.

Si indichi con $S_n = X_1, \ldots, X_n$ un blocco consecutivo di n variabili e sia $p(x_1, x_2, \ldots, x_n) = p(s_n)$ la probabilità congiunta dei blocchi di n simboli consecutivi.

Tale probabilità si assume invariante alle traslazioni, ovvero:

$$p(x_i, x_{i+1}, \ldots, x_{i+n-1}) = p(x_1, x_2, \ldots, x_n) \quad \text{per ogni } i \text{ ed } n$$

il che equivale a considerare tali simboli come generati da un processo stocastico stazionario.

La stringa si può inoltre suddividere in due parti, in una parte sinistra, $\overleftarrow{S}$, corrispondente al "passato", e in una parte destra, $\overrightarrow{S}$, corrispondete al "futuro", nel modo seguente:

$$\overleftarrow{S} \equiv \ldots, X_{-3}, X_{-2}, X_{-1} \quad \text{ed} \quad \overrightarrow{S} \equiv X_0, X_1, X_2, \ldots$$

Per calcolare l'entropia dell'intera sequenza $\overleftrightarrow{S}$ si potrebbe cominciare calcolando l'entropia di Shannon, $H(S_n)$, per blocchi di n simboli consecutivi:

$$H(S_n) = -\sum_{x_1 \in \mathcal{X}} \ldots \sum_{x_n \in \mathcal{X}} p(x_1, \ldots, x_n) \log_2 p(x_1, \ldots, x_n)$$

per poi far tendere n all'infinito, ma si vedrebbe come anche $H(S_n)$ tenderebbe all'infinito positivo, come d'altronde è abbastanza intuitivo supporre, avendo a che fare con un numero infinito di variabili.

Come si può quindi confrontare l'entropia di serie composte da un numero infinito di variabili se tale entropia è infinita? Una soluzione [14, p. 14] potrebbe essere quella di calcolare la cosiddetta *densità di entropia*[5], h_μ, ovvero:

$$h_\mu := \lim_{n \to +\infty} \frac{H(Sn)}{n}$$

detta anche, a seconda dei contesti, "tasso di entropia", "entropia metrica", ecc. La densità di entropia può essere riscritta anche come entropia condizionale:

$$h_\mu = \lim_{n \to +\infty} H(X_n | X_0, X_1, \ldots, X_{n-1})$$

Per processi stazionari, la densità di entropia può essere pensata anche come l'entropia o l'incertezza associata, in media, ad un dato simbolo se tutti i simboli precedenti della serie sono conosciuti.

Se esiste una forte correlazione fra i simboli considerati, la conoscenza di tutti i simboli precedenti ridurrà grandemente l'incertezza riguardo al valore

[5] Si può dimostrare che il limite h_μ esiste, quanto meno, per tutti i processi stocastici stazionari, vedi [8, pp. 63-66].

del successivo. La densità di entropia può essere espressa anche nell'ulteriore forma [14, p. 15]:

$$h_\mu = \lim_{n \to +\infty} [H(S_n) - H(S_{n-1})]$$

da cui risulta chiaro che h_μ esprime il tasso di crescita dell'entropia man mano che si considerano blocchi di simboli sempre più lunghi. La densità di entropia, in definitiva, misura quanto un sistema sia prevedibile, ma non fornisce alcuna indicazione riguardo alla possibile difficoltà nella previsione stessa. Si considerino infatti le due stringhe $\overleftrightarrow{S}_A$ ed $\overleftrightarrow{S}_B$ seguenti:

$$\overleftrightarrow{S}_A = \ldots 10101010101010101010101 \ldots$$

$$\overleftrightarrow{S}_B = \ldots 10101100101011001010110 0 \ldots$$

Entrambe le stringhe sono periodiche, ovvero vi è un blocco di simboli che si ripete indefinitivamente: la prima ha periodo due (il blocco che si ripete, "10", è composto di soli due simboli), mentre la seconda ha periodo 8 (il blocco che si ripete, "10101100", è composto invece da 8 simboli). Per queste loro caratteristiche, l'evoluzione di entrambe le serie può essere prevista con certezza e la densità di entropia di entrambe, di conseguenza, è pari a zero.

Chiaramente, però, la complessità delle due serie non è la stessa. Il periodo di $\overleftrightarrow{S}_B$ è più lungo di quello di $\overleftrightarrow{S}_A$, così sarebbe ragionevole attendersi che il comportamento di $\overleftrightarrow{S}_B$ sia in un certo qual modo più difficile da prevedere di quello di $\overleftrightarrow{S}_A$. Questa però è una distinzione che h_μ non è in grado di evidenziare.

CAPITOLO 2

Complessità di Kolmogorov

Le misure di entropia fino ad ora analizzate cercano tutte di quantificare il livello di disordine o la vicinanza ad una condizione di massima casualità, ma non sono in grado di descrivere la complessità intrinseca di un fenomeno, potendosi infatti osservare uguali livelli di entropia in fenomeni anche molto diversi fra loro dal punto di vista della complessità descrittiva.

In questo capitolo tratteremo quindi del concetto di complessità elaborato da Kolmogorov, concetto strettamente legato a quello di entropia e che, come tale, più volte compare in teoria dell'informazione, ma che non deve essere confuso, come invece alcune volte avviene, con quello proprio di entropia.

Nel 1965 Kolmogorov formulò la sua definizione della complessità intrinseca descrittiva di una sequenza di simboli come la lunghezza del più corto programma binario per calcolatore in grado di descrivere quella determinata sequenza[1]. In questo modo, la complessità di Kolmogorov riferita ad una determinata sequenza non è direttamente vincolata alla distribuzione di probabilità che la caratterizza.

Kolmogorov dimostrò come la sua definizione di complessità fosse sostanzialmente indipendente dal calcolatore utilizzato e come la lunghezza attesa, in bit, del più breve programma binario atto a descrivere una stringa generata da una variabile casuale fosse approssimativamente uguale all'entropia della variabile casuale stessa.

[1] Il fatto che si parli di programma più corto non implica necessariamente che esso sia anche il più efficiente dal punto di vista computazionale e che, quindi, ne minimizzi anche il tempo di esecuzione, condizione questa che, invece, in genere non si realizza

Nella pratica non è però possibile utilizzare concretamente il concetto di più corto programma binario per computer perchè ci potrebbe volere un tempo indefinitivamente lungo per riuscire ad individuare tale programma minimale. Il concetto di complessità di Kolmogorov può essere tuttavia considerato come un modo di pensare e di affrontare i problemi in accordo al principio che la spiegazione più semplice è sempre la migliore.

Alcuni esempi [8, p. 145] potranno meglio chiarire il concetto sopra esposto. Si considerino le tre stringhe seguenti:

1. 01010101010101010101010101010101010101010101010101010101010101010101
2. 01101010000010011110011001100111111100111011110011001001000010000
3. 11011110011101011111011011111011101011011110000101110010100111011

Quale potrebbe essere il più corto programma binario atto a descrivere (stampare) ciascuna di queste sequenze? La prima sequenza è evidentemente molto semplice in quanto è la ripetizione, per trentadue volte, della stessa coppia di simboli "01". Un semplice programma[2] atto a descrivere tale sequenza potrebbe essere, ad esempio, del tipo:

$K = 32$
stampa K volte la stringa "01"

Si può notare come, in questo caso, la complessità di Kolmogorov non vari al variare della lunghezza della stringa, in quanto il programma atto a rappresentarla rimane sempre lo stesso, essendo sufficiente sostituire a 32 il numero effettivo di ripetizioni della coppia "01" che caratterizzano la stringa.

La seconda sequenza sembra casuale, ed infatti supera molti test di casualità, ma in realtà è l'espansione binaria di $\sqrt{2} - 1$, e quindi è ancora una sequenza "semplice". Un programma atto a rappresentarla potrebbe essere:

$K = 64$
calcola[3] $\sqrt{2} - 1$
converti il risultato in un numero binario
stampane le prime K cifre

Anche in questo caso la lunghezza del programma è costante non essendo dipendente dal numero dei simboli che caratterizzano la stringa (le prime K cifre della trasformazione binaria del numero reale $\sqrt{2} - 1$).

Anche la terza sequenza sembra random, a parte il fatto che la proporzione di 1 non è prossima a quella attesa del 50% (ci sono infatti quarantatre 1 e solo ventuno 0). In questo caso, per mezzo di opportuni algoritmi di compressione, la sequenza potrebbe essere espressa con all'incirca

[2] Per esigenze di comprensione esso è scritto in linguaggio discorsivo, ma il concetto che ne è alla base non cambia.

[3] Tutti gli elaboratori dispongono di una funzione per il calcolo della radice quadrata

$\log_2 n + nH(k/n)$ bit[4], dove n identifica la lunghezza della sequenza e k rappresenta il numero degli 1 presenti. Anche questa volta, quindi, è possibile esprimere la sequenza con un numero di bit inferiore ad n ed il vantaggio di questa rappresentazione sarà tanto più elevato quanto più le proporzione dei simboli presenti nella sequenza (in questo caso 0 ed 1) si discostano da quelle di perfetta casualità. Si può concludere che, anche in questo caso, la sequenza, nonostante appaia come casuale, è relativamente "semplice", anche se non così semplice come le due precedenti che potevano essere rappresentate per mezzo di programmi di lunghezza costante. In questo caso, invece, la lunghezza del programma, e quindi la complessità di Kolmogorov associata alla sequenza, è proporzionale ad n.

Consideriamo infine una sequenza completamente casuale 010100101110 ... 011010010 quale potrebbe essere, ad esempio, quella generata dal lancio di una moneta non truccata. In questo caso possiamo ottenere 2^n possibili sequenze di lunghezza n tutte ugualmente probabili. E' inoltre molto probabile che una sequenza di questo tipo non sia comprimibile (ovvero che non sia possibile descriverla con un numero di bit inferiore al numero di cifre che la compongono) e che, quindi, per rappresentare tale sequenza non vi sia programma migliore di uno del tipo:

stampa la stringa 010100101110 ... 011010010

e che, quindi, la complessità descrittiva di una sequenza binaria completamente casuale sia pari alla lunghezza della sequenza stessa.

2.1 Indipendenza della nozione di complessità dal tipo di calcolatore considerato

La nozione di intrinseca complessità è indipendente dal calcolatore che viene utilizzato per definirla, ovvero la lunghezza del più corto programma binario realizzabile per descrivere una determinata sequenza di simboli non dipende dal calcolatore considerato, in quanto tale lunghezza rimane sempre la stessa a meno di una costante addittiva che caratterizza i diversi tipi di calcolatore. Inoltre, nel caso di lunghe sequenze ad elevata complessità, questa costante addittiva, che rappresenta, in sostanza, la lunghezza del preprogramma che consente ad un determinato calcolatore di simularne un altro, diventa trascurabile.

Il concetto di cui sopra si traduce nel fatto che ogni calcolatore, anche il più semplice, può essere considerato come universale, nel senso che, a partire da un qualsiasi calcolatore, mediante l'utilizzo di programmi più o meno complessi, è sempre possibile simulare il comportamento di tutti gli altri.

[4] $H(k/n)$ è l'entropia di una distribuzione di probabilità binaria del tipo: $P(X = 0) = (n - k)/n$ e $P(X = 1) = k/n$

Come calcolatore di riferimento per il calcolo della complessità di Kolmogorov si considera, in genere, la Macchina Universale di Turing o UTM, definita teoricamente nel 1937 da Alan Turing come un insieme di operazioni elementari atte a simulare il comportamento computazionale umano. Essa rappresenta, concettualmente, il più semplice calcolatore universale e, come tale, ogni nuovo sistema di calcolo può esservi sempre ricondotto. Di conseguenza, ogni calcolatore può essere sempre simulato da e simulare una Macchina Universale di Turing .

2.2 Definizione rigorosa di complessità di Kolmogorov

Data una stringa binaria x di lunghezza finita e un calcolatore universale $\mathcal{U}$, denotata con $l(x)$ la lunghezza della stringa x e con $\mathcal{U}(p)$ l'output del calcolatore $\mathcal{U}$ quando gli viene sottoposto il programma p, si definisce *complessità di Kolmogorov* (o complessità algoritmica) $K_{\mathcal{U}(x)}$ di una stringa x rispetto al calcolatore universale $\mathcal{U}$ la descrizione di x avente lunghezza minima, ovvero [8, p. 147]:

$$K_{\mathcal{U}}(x) = \min_{p:\mathcal{U}(p)=x} l(p),$$

e cioè la minima lunghezza fra tutti i possibili programmi interpretati dal calcolatore $\mathcal{U}$ che stampano la stringa x.

La complessità di Kolmogorov può essere pensata anche come segue: "se una persona può descrivere una sequenza di simboli ad un'altra persona in maniera tale da condurre, senza ambiguità e in un tempo finito, al calcolo o alla rappresentazione di tale sequenza, il numero di bit impiegati nella descrizione rappresenta un limite superiore per la complessità di Kolmogorov". Ad esempio, con 70 caratteri si potrebbe scrivere "Stampa le prime 1.239.875.981.825.931 cifre della radice quadrata di e". Considerando 8 bit per carattere (ASCII), la complessità di Kolmogorov di quel numero enorme non è più grande di $8 * 70 = 560$ bit. La maggior parte di sequenze di questa lunghezza avrebbe invece una complessità di Kolmogorov pari a 1.239.875.981.825.931 bit, ma nel caso specifico esiste un semplice algoritmo che ci consente di calcolare direttamente la radice quadrata di e consentendoci, quindi, un notevole risparmio in termini di complessità descrittiva.

Nella definizione precedente non si fa menzione della lunghezza della stringa x. Se si assume che il calcolatore conosca già la lunghezza $l(x)$ della stringa x, si può definire la *complessità condizionale di Kolmogorov*, nota $l(x)$, come [8, p. 148]:

$$K_{\mathcal{U}}[x|l(x)] = \min_{p:\mathcal{U}[p,l(x)]=x} l(p),$$

inoltre, se $\mathcal{U}$ è un calcolatore universale, allora, per ogni altro calcolatore $\mathcal{A}$, si ha che:

$$K_{\mathcal{U}}(x) \leq K_{\mathcal{A}}(x) + c_{\mathcal{A}}$$

per tutte le stringhe $x \in \{0,1\}$ e dove la costante $c_{\mathcal{A}}$ non dipende da x (universalità della complessità di Kolmogorov)[8, p. 148].

La costante $c_{\mathcal{A}}$ può essere anche molto grande, in quanto legata alle dimensioni del programma che consente ad un calcolatore di simularne un altro, e quindi di riprodurne tutte le routine e le funzioni. Il punto fondamentale, però, è che la lunghezza del programma di simulazione, e quindi la dimensione della costante $c_{\mathcal{A}}$ è indipendente dalla lunghezza di x, la stringa che deve essere rappresentata. Per sequenze sufficientemente lunghe, la lunghezza del programma di simulazione diventa trascurabile, motivo per cui spesso si può trattare la complessità di Kolmogorov trascurando tale costante.

Se $\mathcal{U}$ e $\mathcal{A}$ sono entrambi calcolatori universali, si ha sempre che:

$$|K_{\mathcal{U}}(x) - K_{\mathcal{A}}(x)| \leq c$$

Nel seguito sono riportati gli enunciati di alcuni teoremi utili a descrivere diverse proprietà della complessità di Kolmogorov.

- La complessità condizionale non è maggiore della lunghezza della sequenza [8, p. 149], ovvero:

$$K[x|l(x)] \leq l(x) + c$$

- Limite superiore della complessità di Kolmogorov [8, p. 149]:

$$K(x) \leq K[x|l(x)] + 2\log_2 l(x) + c$$

- Limite inferiore della complessità di Kolmogorov [8, p. 150]. Il numero di stringhe x aventi complessità $K(x) < k$ soddisfa la seguente disuguaglianza:

$$|\{x \in \{0,1\} : K(x) < k| < 2^k$$

- La complessità di Kolmogorov di una stringa binaria x è limitata da [8, p. 153]:

$$K(x_1 x_2 \ldots x_n|n) \leq nH_0\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + 2\log_2 n + c$$

con $H_0(p) = -p\log_2 p - (1-p)\log_2(1-p)$

Si avrà successo nel comprimere la stringa, ovvero nel rappresentarla mediante un numero di bit inferiore al numero delle cifre che la compongono, solo se $l(p) < l(x)$, dove $l(p)$ è la lunghezza del programma di compressione, ovvero se $K(x) < l(x)$.

In generale, quando la lunghezza $l(x)$ della sequenza è piccola, la costante che compare nell'espressione della complessità di Kolmogorov sovrasterà di gran lunga il contributo di $l(x)$, quindi i teoremi di cui sopra sono utili soprattutto quando $l(x)$ è molto grande. In questo caso la costante, che non dipende da $l(x)$, potrà essere tranquillamente trascurata.

2.3 Complessità di Kolmogorov ed entropia

In generale, come già precedentemente accennato, il valore atteso della complessità di Kolmogorov di una sequenza casuale è prossimo all'entropia di Shannon. Una definizione più rigorosa della relazione fra complessità di Kolmogorov ed entropia può essere espressa come segue [8, p. 154]: dato un processo stocastico $\{X_i\}$ costituito da variabili casuali i.i.d. con densità di probabilità $f(x), x \in \mathcal{X}$, dove $\mathcal{X}$ è un alfabeto finito e sia

$$f(x^n) = \prod_{i=1}^{n} f(x_i)$$

allora, per ogni n esiste una costante c tale che

$$H(X) \leq \frac{1}{n} \sum_{x^n} f(x^n) K(x^n|n) \leq H(X) + \frac{|\mathcal{X}| \log n}{n} + \frac{c}{n}$$

di conseguenza $\lim_{n \to +\infty} \frac{1}{n} E\left[K(X^n|n)\right] = H(X)$

2.4 Casualità algoritmica e sequenze incomprimibili

Come si è potuto osservare nei paragrafi precedenti, esistono un certo numero di lunghe sequenze che possono essere considerate relativamente "semplici", tuttavia la maggior parte delle sequenze non è di tipo semplice. Questo significa che, se si osserva una sequenza generata in maniera casuale, è molto probabile che tale sequenza sia di tipo complesso. Si può infatti dimostrare che la probabilità che una sequenza generata in maniera casuale possa essere compressa per più di k bit non è superiore a 2^{-k} e che, quindi, la maggior parte delle sequenze abbia una complessità prossima alla propria lunghezza.

In simboli, date n variabili casuali di Bernoulli di parametro $1/2$, $X_1, X_2, \ldots, X_n$, si ha che [8, p. 157]:

$$P[K(X_1 X_2 \ldots X_n|n) < n - k] < 2^{-k}$$

Molto importanti sono poi i concetti di *sequenza algoritmicamente casuale* e di *sequenza incomprimibile*:

- una sequenza $x_1, x_2, \ldots, x_n$ è detta *algoritmicamente casuale* se [8, p. 157] $K(x_1 x_2 \ldots x_n | n) \geq n$
- una sequenza x di lunghezza infinita è detta *incomprimibile* se [8, p. 157] $\lim_{n\to\infty} \frac{K(x_1 x_2 \ldots x_n | n)}{n} = 1$

Se una stringa binaria $x_1 x_2 \ldots x_n$ è incomprimibile, allora soddisfa la legge dei grandi numeri nel senso che [8, p. 157] :

$$\frac{1}{n} \sum_{i=1}^{n} x_i \to \frac{1}{2} \text{ in probabilità}$$

Ciò significa che le proporzioni di 0 e di 1 in una qualsiasi sequenza incomprimibile sono all'incirca uguali. Inoltre, la complessità di Kolmogorov di una sequenza di variabili casuali binarie i.i.d. aventi distribuzione di Bernoulli di parametro θ è prossima all'entropia $H(\theta)$. Si può infatti dimostrare che [8, p. 159], date n variabili casuali i.i.d. $X_1, X_2, \ldots, X_n$ distribuite secondo una $\mathcal{B}er(\theta)$, allora:

$$\frac{1}{n} K(X_1, X_2, \ldots, X_n | n) \to H(\theta) \text{ in probabilità}$$

CAPITOLO 3

Alcuni indicatori di dipendenza basati sul concetto di entropia

Le misure e i test statistici di dipendenza più comunemente usati sono opportune funzioni di correlazione giustificate da relazioni lineari che coinvolgono variabili continue e/o processi Gaussiani. Questo tipo di misure tendono ad essere inefficaci nel caso di variabili discrete o quando trattano relazioni non lineari o processi non Gaussiani.

Recentemente sono stati sviluppati diversi metodi che utilizzano i concetti di entropia mutuati dalla teoria dell'informazione. L'entropia è infatti definita sull'insieme delle *distribuzioni* che sono alla base dei concetti di dipendenza/indipendenza sia nel caso discreto che continuo, è adimensionale e si può applicare sia al caso univariato che a quello multivariato.

In questo capitolo verranno analizzati nel dettaglio alcuni di questi indicatori ed, in particolare, la mutua informazione, l'entropia relativa o distanza di Kullback Leibler, l'entropia di Granger, Maasoumi e Racine o S_ρ, l'entropia Approssimata (Approximate Entropy) o *ApEn* e una sua variante, da utilizzarsi nel confronto di serie diverse, la *cross-ApEn*. I primi tre indicatori sono funzioni esplicite delle distribuzioni di probabilità marginali e della distribuzione di probabilità congiunta delle variabili aleatorie che caratterizzano i fenomeni oggetto di studio, distribuzioni che dovranno essere stimate a partire dai dati campionari, ovvero dalle serie storiche considerate. La *ApEn* e la *cross-ApEn* si basano invece sul modo con cui i singoli elementi si susseguono nelle serie considerate per cui, in questo caso, non è necessario stimare alcuna distribuzione di probabilità. La *ApEn* misura infatti

il logaritmo della frequenza con cui blocchi di elementi (pattern) di una determinata lunghezza m che sono "vicini"[1] fra loro lo rimangano anche se si aumenta di un elemento la dimensione di tali blocchi.

3.1 Mutua informazione

Come già accennato nella precedente sezione 1.5, la mutua informazione $I(X;Y) = \sum_i \sum_j p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$, fornisce un'indicazione del legame eventualmente esistente fra due variabili aleatorie X ed Y. In particolare essa:

- assume valore zero se solo se $p(x, y) = p(x)p(y)$, ovvero se le variabili X ed Y sono fra loro indipendenti;
- è sempre non negativa, quindi $I(X;Y) \geq 0$;
- nel caso continuo si ha che $I(X;Y) = +\infty$ se $Y = g(X)$, ovvero l'indicatore tende all'infinito positivo nel caso in cui vi sia una perfetta relazione, ancorché non lineare, fra X e Y.

La mutua informazione è stato proposta, ad esempio, come criterio su cui basare test di indipendenza [11][5], per lo studio del livello di dipendenza (non lineare) nelle serie storiche [9][5], per l'analisi delle componenti principali [25], ecc.

La principale difficoltà nel calcolo della mutua informazione, così come della distanza di Kullback Leibler e dell'entropia di Granger, Maasoumi e Racine (S_ρ), che tratteremo più avanti, risiede nel fatto che le distribuzioni di probabilità marginali e la distribuzione di probabilità congiunta sono generalmente sconosciute. Inoltre, a parte pochi casi in cui tali distribuzioni sono particolari funzioni di forma nota, non è quasi mai possibile determinare la distribuzione della mutua informazione in forma analitica.

Un metodo standard utilizzato per stimare le distribuzioni di probabilità coinvolte è quello degli istogrammi (a celle equidistanti o a celle equiprobabili). In generale, però, tale metodo conduce a notevoli sovrastime o sottostime in relazione al particolare tipo di distribuzione considerata. In generale, si possono ottenere risultati migliori utilizzando un sistema basato su istogrammi adattativi, ovvero istogrammi che sono in grado di adattarsi nel migliore dei modi alle diverse distribuzione di probabilità [9]. Ciò si può realizzare utilizzando una definizione della mutua informazione basata sulle partizioni. Una partizione finita di $\mathcal{R}^d$ è un qualsiasi insieme finito di sotto insiemi disgiunti di $\mathcal{R}^d$ la cui unione costituisce l'intero $\mathcal{R}^d$. Tali sottoinsiemi

[1] Ovvero blocchi per i quali il massimo del valore assoluto della differenza fra elementi corrispondenti è inferiore ad un determinato valore di soglia r.

sono spesso chiamati celle della partizione e, in pratica, hanno forma rettangolare (ipercubi di $\mathcal{R}^d$). Ogni cella può essere indicata come $C_k = A_k B_k$, dove A_k è la proiezione ortogonale di C_k sul sottospazio in cui è definita la variabile X e B_k è la proiezione ortogonale di C_k sul sottospazio in cui è definita la variabile Y. Si può quindi dimostrare[2] che la mutua informazione è l'estremo superiore di tutte le possibili partizioni finite di $\mathcal{R}^d$[11]:

$$I(X;Y) = \sup_{\{C_k\}} \sum_k P_{X,Y}(C_k) \log_2 \frac{P_{X,Y}(C_k)}{P_X(A_k) P_Y(B_k)}$$

dove $\{C_k\}$ è l'insieme di tutte le possibili partizioni costituite dalle celle C_k, $P_{X,Y}(C_k)$ è la probabilità che la coppia (X, Y) assuma il suo valore nella cella C_k, $P_X(A_k)$ è la probabilità che X assuma il suo valore in A_k e $P_Y(B_k)$ assuma il suo valore in B_k. Si può inoltre dimostrare che, costruendo una sequenza di partizioni costituite da celle sempre più piccole, i corrispondenti valori della mutua informazione crescono in maniera monotonica. Il processo di crescita tende ad esaurirsi quando all'interno di ogni cella della partizione si è ottenuta una distribuzione dei punti pressoche uniforme. Per mezzo di un test di uniformità come ad esempio il χ^2 è quindi possibile decidere quando terminare la procedure ricorsiva di partizionamento [9][31].

Un altro metodo consolidato e molto utilizzato nella stima delle distribuzioni necessarie per il calcolo della mutua informazione, così come della distanza di Kullback Leibler e della S_ρ, è quello degli stimatori di densità del kernel o KDE [31].

La letteratura statistica riporta che gli stimatori kernel sono in genere superiori al metodo degli istogrammi in quanto presentano errore quadratico medio inferiore, sono insensibili alla scelta di un'origine per la determinazione delle celle ed è possibile specificare celle di forma più sofisticata da utilizzarsi per il conteggio delle frequenze incognite. Non ci sono infatti ragioni formali per utilizzare ipercubi come celle sulle quali stimare la frequenza relativa degli eventi. Il metodo della densità del kernel utilizza pertanto un peso generalizzato, o funzione di kernel, per caratterizzare ogni cella. In particolare, se un determinato evento $y = [y_1, y_2, \ldots, y_d]^T$ è individuato da un vettore casuale di dimensioni d di cui si vuole stimare la probabilità e $y_i = [y_{1i}, y_{2_i}, \ldots, y_{di}]^T$, con $i = 1, n$, sono gli n vettori campionari, si ha che:

$$\widehat{p}(y) = \frac{1}{n} \sum_{i=1}^{n} K(u_i)$$

con

$$u_i = \frac{(y - y_i)^T S^{-1} (y - y_i)}{h^2}$$

dove:

[2] R. L. Dobrushin. General formulation of Shannon's main theorem in information theory. Uspekhi Mat. Nauk (in russo), 14:3-104, 1959. Tradotto in Am. Math. Soc. Trans., 33:323-428, 1959. Citato da [5]

- $\widehat{p}(y)$ è una stima della distribuzione di probabilità della variabile vettoriale y;
- $K(u)$ è una funzione kernel multivariata;
- h è l'ampiezza di banda del kernel che corrisponde, sostanzialmente, alla dimensione dei lati dell'ipercubo del metodo degli istogrammi;
- S è la matrice di varianze e covarianze del vettore campionario y_i.

La funzione di kernel $K(u)$ deve essere una funzione di probabilità valida. A tal fine viene spesso utilizzata la funzione normale multivariata

$$K(u) = \frac{1}{(2\pi)^{d/2} h^d \det(S)^{1/2}} e^{-u/2}$$

In questo contesto, $K(u)$ rappresenta il peso dato ad una osservazione y_i, peso che è basato sulla distanza tra y ed y_i (in genere una distanza euclidea modificata per tenere conto della covarianza fra le coordinate dei vari punti). Lo stimatore di densità del kernel è, in definitiva, una media pesata delle frequenze relative delle osservazioni nelle vicinanze del punto da stimare. La funzione di kernel $K(u)$ fornisce il peso relativo mentre h determina l'insieme dei valori su cui è calcolata la media.

Fra gli altri possibili metodi si possono poi citare quello dei k più prossimi vicini (o KNN) [25] e l'espansione di Edgeworth, o EDGE, riportata in [50]. L'approccio basato sul metodo dei k più prossimi vicini sembra avere prestazioni migliori di quello basato sulla densità del kernel purchè k sia scelto in maniera appropriata. Un valore di k troppo piccolo (o troppo grande) conduce ad uno stimatore con una piccola (grande) distorsione ed una grande (piccola) varianza. In ogni caso, la scelta di un k appropriato in modo da ottimizzare sia la distorsione che la varianza dello stimatore è un processo alquanto difficile da realizzarsi nella pratica. Il metodo di Edgeworth, invece, funziona bene quando le distribuzioni di probabilità da stimare sono prossime alla distribuzione normale, altrimenti lo stimatore risulta distorto [31].

Poiché $0 \leq I(X;Y) \leq +\infty$, il confronto fra campioni diversi risulta difficile. Per superare questo limite diversi autori, tra cui Joe [23], hanno suggerito una versione standardizzata della mutua informazione, il coefficiente di correlazione globale, definito come $\lambda(X;Y) = \sqrt{1 - e^{2I(X;Y)}}$, che varia tra 0 e 1 ed è così confrontabile con il coefficiente di correlazione lineare r. Il coefficiente λ esprime la dipendenza globale, quindi sia quella lineare che non lineare, fra le variabili X ed Y.

Nello studio dell'autocorrelazione, la mutua informazione ci consente di individuare lag caratterizzati da una eventuale presenza di dipendenza non lineare [31], anche se non ci fornisce nessuna indicazione riguardo al tipo di relazione eventualmente riscontrata [11].

La mutua informazione, inoltre, può essere utilizzata nella scelta di un appropriato parametro di ritardo per la ricostruzione (embedding) dello spazio delle fasi a partire da serie temporali sperimentali. Lo spazio delle fasi è costruito sviluppando un sistema rappresentativo d-dimensionale a partire dai valori della serie storica scalare $x(t)$. Nello spazio delle fasi d-dimensionale un punto al tempo t è rappresentato dal vettore $\beta_t = \{x_t, x_{t+\tau}, x_{t+2\tau}, \ldots, x_{t+(d-1)\tau}\}$, dove τ rappresenta il tempo di ritardo. Nella pratica, la scelta di un valore di τ appropriato è un passo importante in quanto la qualità della ricostruzione dipende anche da tale valore. Se τ è troppo piccolo, infatti, non ci sarà praticamente differenza fra i diversi elementi del vettore $\beta_t = \{x_t, x_{t+\tau}, x_{t+2\tau}, \ldots, x_{t+(d-1)\tau}\}$, in quanto tutti i punti si addenseranno attorno alla bisettrice dello spazio ricostruito. Nel caso in cui i dati siano affetti da rumore, inoltre, i vettori così costituiti non avranno praticamente significato se le variazioni del segnale nell'intervallo temporale $d\tau$ dovessero essere inferiori rispetto al livello di rumore osservato. D'altro canto, se si scegliesse un τ troppo grande, le diverse coordinate potrebbero risultare praticamente incorrelate. In questa situazione, l'attrattore ricostruito potrebbero diventare molto complicato anche nel caso in cui il sottostante "vero" attrattore fosse semplice. Ciò si manifesta, in particolare, nei sistemi caotici, dove la funzione di autocorrelazione decade molto velocemente [24, p. 130]. Quale criterio di scelta per τ, Fraser e Swinney [16], proposero di utilizzare il primo minimo incontrato nella mutua informazione $I(X_t; X_{t-\tau})$. Questo criterio può dare risultati migliori rispetto alla funzione di autocorrelazione, poiché quest'ultima misura solo la dipendenza lineare, mentre $I(X_t; X_{t-\tau})$ riesce a cogliere anche la relazione non lineare fra le due variabili.

In ogni caso, nessun criterio fornisce sempre i migliori risultati in tutte le possibili situazioni. Ciò non di meno, $I(X;Y)$ è utile per indagare la dipendenza tra le coordinate e per identificare altre variabili che possono essere utili per la ricostruzione degli attrattori.

3.2 Entropia relativa o distanza di Kullback Leibler

L'Entropia Relativa o distanza di Kullback Leibler, $D(p||q) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)}$, è comunemente utilizzata in statistica come misura di similarità tra due distribuzioni di probabilità. La distanza di Kullback Leibler soddisfa, tra le altre, le tre seguenti proprietà:

- autosimilarità: $D(p||p) = 0$;
- autoidentificazione: $D(p||q) = 0$ se e solo se $p = q$;
- positività: $D(p||q) \geq 0$ per ogni p, q.

La distanza di Kullback Leibler è utilizzata in diversi contesti come misura appropriata della distanza fra due modelli statistici e riveste un ruolo fondamentale per ottenere i cosiddetti "criteri di informazione". Tali criteri sono alla base della costruzione di procedure per la scelta fra modelli statistici diversi quali, ad esempio, lo AIC (Akaike Information Criteria) [12]. E' inoltre utilizzata in molti aspetti del riconoscimento del discorso e delle immagini [21].

La distanza di Kullback Leibler non è una misura simmetrica. Dato che, però, questa caratteristica può essere desiderabile in diverse situazioni, è stata sviluppata anche una sua forma simmetrizzata, che prende il nome di misura di Jefferys Kullback Leibler ed è così definita[3] [28][53]:

$$J(p||q) = D(p||q) + D(q||p) = \sum_{i=1}^{n}(q_i - p_i)(\log_2 q_i - \log_2 p_i)$$

Anche in questo caso, come già per la mutua informazione, le distribuzioni di probabilità coinvolte sono normalmente molto complesse e quindi, salvo che in pochi casi elementari, non è generalmente possibile calcolare per via analitica la distanza di Kullback Leibler, motivo per cui, generalmente, si ricorre ad una soluzione numerica per mezzo dei metodi Monte Carlo.

L'idea è di estrarre un campione x_i a partire dalla distribuzione di probabilità p così che $E_p\left[\log_2 \frac{p(x_i)}{q(x_i)}\right] = D(p||q)$. Utilizzando n campioni indipendenti ed identicamente distribuiti $\{x_i\}$, $i = 1, n$, si ha che [12][21][53]:

$$D(p||q)_{\text{MonteCarlo}} = \frac{1}{n}\sum_{i=1}^{n}\left[p(x_i)\log_2 \frac{p(x_i)}{q(x_i)}\right]$$

con $D(p||q)_{\text{MonteCarlo}} \to D(p||q)$ per $n \to +\infty$.

Nel caso in cui p e q siano due distribuzioni di probabilità normali in R^d con medie μ_p e μ_q e matrici di varianze e covarianze $\sum_p$ e $\sum_q$, la misura di Kullback Leibler esiste in forma chiusa ed ha forma:

$$D(p||q) = \frac{1}{2}\left[\log_2 \frac{|\Sigma_q|}{|\Sigma_p|} + \ \text{Tr}\left(\Sigma_q^{-1}\Sigma_p\right) - d + (\mu_p - \mu_q)^T\Sigma_q^{-1}(\mu_p - \mu_q)\right]$$

dove $|\Sigma_p|$ è il determinante della matrice Σ_p, Tr è la funzione traccia e d è la dimensione dello spazio in cui sono definite p e q.

3.3 Entropia di Granger, Maasoumi e Racine

Fra i vari indici basati sul concetto di entropia citati in letteratura, vale la pena menzionare quello proposto da Granger, Maasoumi e Racine, più spesso

[3] H. Jefferys. Proc. Roy. Soc. Lon. Serie A 186, 453-461, 1946 e S. Kullback and R. A. Leibler. Ann. Math. Statist. 22, 79-86, 1951. Citati da [53]

indicato come S_ρ, e definito come:

$$S_\rho = \frac{1}{2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left(f^{\frac{1}{2}} - f_1^{\frac{1}{2}} f_2^{\frac{1}{2}}\right)^2 dxdy$$

dove f è la distribuzione di densità congiunta delle variabili casuali X e Y, mentre f_1 ed f_2 sono le loro rispettive distribuzioni di densità marginali. Se X ed Y sono indipendenti, $S_\rho = 0$ altrimenti la S_ρ è positiva e maggiore di zero.

Granger, Maasoumi e Racine svilupparono la S_ρ a partire dalla misura di Hellinger-Battacharya-Matusita [20][29]. In particolare, il coefficiente di Battacharya su due generiche distribuzioni f_a e f_b è definito come [4][52]:

$$\rho^*(f_a, f_b) = \int_{-\infty}^{+\infty}(f_a f_b)^{\frac{1}{2}} dx$$

e costituisce una misura di divergenza fra distribuzioni. Tale coefficiente non ha una struttura metrica in quanto non soddisfa tutti e tre gli assiomi necessari. E' invece una distanza metrica la sua versione modificata, la distanza di Battacharya $B(f_a, f_b) = \sqrt{1 - \rho^*(f_a, f_b)}$.

Il discriminante di Hellinger, anche conosciuto con il nome di misura di Matusita, ha invece la forma seguente [52][30]:

$$M(f_a, f_b) = \int_{-\infty}^{+\infty}(f_a^{\frac{1}{2}} - f_b^{\frac{1}{2}})^2 dx$$

Fra il discriminante di Hellinger e la distanza di Battacharya esiste la seguente relazione:

$$M(f_a, f_b) = 2B(f_a, f_b)^2$$

Sia $M(f_a, f_b)$ che $B(f_a, f_b)$, caso raro fra le misure di divergenza, hanno la peculiare caratteristica di soddisfare, tra l'altro, la disuguaglianza triangolare e quindi di potersi considerare misure metriche, così come la S_ρ dato che

$$S_\rho = 1 - \rho^*(f_a, f_b) = B(f_a, f_b)^2$$

quando a f_a si sostituisca la distribuzione di densità congiunta delle variabili casuali X ed Y, $f = f(X, Y)$ ed a f_b il prodotto $f_1 f_2$ delle rispettive densità marginali. La S_ρ ha inoltre una precisa relazione con la famiglia di entropie di ordine k di Havrda e Charvat:

$$\begin{aligned} H_k(p) &= \frac{1}{k-1}\left[1 - E(p^{k-1})\right], \text{ per } k \neq 1, \\ &= \text{ entropia di Shannon, } -E(\log p), \text{ per } k \to 1 \end{aligned}$$

Per due qualsiasi funzioni di densità f_a ed f_b la famiglia di entropie di ordine k (asimmetrica rispetto a f_b) diventa:

$$D_k(f_a||f_b) = \frac{1}{k-1}\left[\int_{-\infty}^{+\infty}(f_b^k f_a^k)\mathrm{d}F_a - 1\right], \text{ per } k \neq 1$$

con $D_k(f_a||f_b) \to D(f_a||f_b)$, distanza di Kullback Leibner, per $k \to 1$. Se si mediano le due misure asimmetriche $D_k(f_a||f_b)$ e $D_k(f_b||f_a)$ si ottiene una misura simmetrica che, per $k \to 1$ coincide con la misura di Jefferys Kullback Leibler $J(f_a||f_b)$.

Si consideri infine la misura simmetrica di classe k per $k = 1/2$. In questo caso si ha:

$$D_{1/2} = D_{1/2}(f_a||f_b) + D_{1/2}(f_b||f_a) = 2M(f_a, f_b) = 4B(f_a, f_b)^2 \quad (3.1)$$

Sostituendo nella (3.1) la distribuzione congiunta $f = f(X, Y)$ al posto di f_a e il prodotto delle distribuzioni marginali $f_1(X)f_2(Y)$ al posto di f_b, si ottiene nuovamente la S_ρ, ovvero l'entropia di Granger, Maasoumi e Racine.

Con la S_ρ gli autori cercavano un indicatore che non si limitasse a verificare la dipendenza di per sè (sia lineare che non lineare), ma che potesse misurare anche il grado di scostamento dalla condizione di indipendenza e che fosse robusta nei confronti di possibili (ma sconosciuti) processi non lineari e non gaussiani. La mutua informazione e la distanza di Kullback Leibler, così come quasi tutte le altre entropie, non sono però misure di tipo "metrico", in quanto esse violano o il principio di simmetria, o la regola della disuguaglianza triangolare, o entrambe. Ciò significa che esse costituiscono misure di divergenza, non di distanza, mentre la S_ρ, essendo una misura di tipo "metrico", ha invece il vantaggio di consentire comparazioni fra fenomeni e modelli statistici diversi.

L'entropia di Granger, Maasoumi e Racine normalizzata al suo massimo soddisfa formalmente le sei seguenti proprietà [29]:

- è definita sia per variabili continue che per variabili discrete (basta sostituire al simbolo di integrale quello di sommatoria);
- assume valore zero nel caso di indipendenza fra X e Y e varia fra 0 e $+1$;
- assume valore uguale all'unità nel caso di un'esatta e misurabile (non necessariamente lineare) relazione $Y = m(X)$ fra le variabili,
- è uguale oppure ha una semplice relazione con il coefficiente di correlazione lineare nel caso di distribuzione normale bivariata,
- è una misura metrica, ovvero una vera misura di distanza e non solo di divergenza,
- è invariante rispetto ad una trasformazione continua e strettamente crescente $\psi(.)$. Ciò è utile in quanto X e Y sono indipendenti se e solo se $\psi(X)$ e $\psi(Y)$ sono indipendenti. La proprietà di invarianza è quindi importante perchè, altrimenti, trasformazioni volute o non volute su X e Y potrebbero evidenziare diversi livelli di dipendenza.

Come è stato fatto notare dagli stessi Granger, Maasoumi e Racine [20], l'approssimazione asintotica normale derivata per la S_ρ non consente di ottenere risultati affidabili per eseguire inferenze. E' pertanto necessario calcolare gli intervalli di confidenza sotto l'ipotesi nulla di indipendenza ricorrendo ai metodi Monte Carlo, così da poter verificare eventuali e significativi scostamenti dallo zero. In particolare, applicando un rimescolamento casuale ai dati disponibili, è possibile simulare delle replicazioni che siano fra loro serialmente indipendenti ma che conservino le distribuzioni marginali dei dati originali. Infatti, il riordinamento casuale dei dati lascia intatte le distribuzioni marginali generando, nel contempo, una distribuzione bivariata indipendente. Il rimescolamento può quindi essere utilizzato per calcolare la statistica S_ρ con dati generati in accordo all'ipotesi nulla di indipendenza. Questo procedimento può essere ripetuto un numero molto elevato di volte in modo da generare una distribuzione empirica della S_ρ sotto l'ipotesi di indipendenza. A questo punto, data la distribuzione empirica, è possibile ottenere i valori critici da applicare a dati provenienti da campioni finiti.

Per derivare le densità coinvolte e pervenire così ad una stima consistente della S_ρ, Granger, Maasoumi e Racine [20] suggeriscono l'uso degli stimatori kernel originariamente proposti da Parzen [33]. Gli stessi autori affermano che la S_ρ si è dimostrata particolarmente indicata per individuare forme generiche di associazione e di dipendenza seriale presenti nelle serie storiche analizzate.

3.4 Approximate Entropy

Si possono osservare due comportamenti caratteristici tramite i quali le serie appaiono deviare da una situazione di cosiddetta *regolarità*:

- quando manifestano un'elevata deviazione standard;
- quando appaiono caratterizzate da un elevato grado di disordine o di casualità.

Questi due aspetti sono sostanzialmente differenti e necessitano, per essere trattati, di strumenti investigativi diversi. Nel primo caso, ovvero quando si voglia quantificare una deviazione da una tendenza centrale, lo strumento più appropriato rimane la misura della deviazione standard. Per studiare il grado di disordine, invece, sarebbe auspicabile poter disporre di un indicatore in grado di misurare il livello di irregolarità, nel senso di imprevedibilità, delle fluttuazioni della serie.

Agli inizi degli anni 90 del XX secolo Steve Pincus [34] [40] [41] ha sviluppato un indicatore da lui chiamato Approximate Entropy (entropia approssimata), indicato con *ApEn*, e finalizzato alla quantificazione dell'*irregolarità* in sequenze di dati. La *ApEn*, inizialmente indirizzata allo studio di data set relativamente piccoli ed affetti da "rumore", è un indicatore, variabile

in maniera continua, in grado di caratterizzare una serie storica in base al suo *livello di irregolarità* a partire dalla condizione di ordine totale (perfetta prevedibilità, ovvero $ApEn = 0$) fino a quella di completa irregolarità (serie completamente random[4]).

Il calcolo della *ApEn* non si basa su assunzioni particolari o modelli specifici, ma è determinato unicamente dall'ordine e dalla frequenza con cui gli elementi si succedono all'interno della serie. E' quindi applicabile anche a singole sequenze ed a prescindere da modelli teorici di riferimento.

A detta dell'autore la *ApEn* può essere utilizzata per:

- verificare se una determinata sequenza di dati possa o meno essere considerata di tipo random;
- cogliere sottili differenze fra serie caratterizzate da un alto grado di irregolarità;
- misurare una sorta di "distanza" dalla condizione di massima irregolarità, consentendo così di "ordinare" serie diverse in base al loro grado di disordine.

Sebbene la *ApEn* sia stata originariamente sviluppata in ambienti matematici, il suo impiego è tutt'ora sostanzialmente limitato al campo delle scienze naturali, in particolare in medicina e fisiologia, mentre ha avuto un'attenzione assolutamente marginale in ambito statistico. Fra gli impieghi più significativi della *ApEn* si possono citare quelli effettuati, ad esempio, per cercare di discriminare fra neonati sani e neonati malati sulla base della quantificazione del livello di regolarità osservato negli elettrocardiogrammi (differenze fra battiti). L'ipotesi, che sembrerebbe effettivamente confermata dai risultati sperimentali, è che il battito cardiaco dei bambini più sani abbia un grado di casualità maggiore, e che quindi appaia meno regolare, rispetto a quello dei bambini malati. Negli studi effettuati, i bambini malati hanno manifestato infatti valori della *ApEn* mediamente più bassi rispetto a quelli riscontrati nei gruppi di controllo [37], quando addirittura non manifestavano valori della *ApEn* inferiori a tutti quelli relativi ai neonati sani [38].

3.4.1 Definizione

Tramite la *ApEn* si assegna a ciascuna sequenza, o serie storica analizzata, un numero non negativo. Ad un valore della *ApEn* più elevato corrisponde una

[4]Per la teoria della complessità algoritmica, una serie di simboli è considerata imprevedibile, ovvero random, se le informazioni contenute nelle serie non possono essere ulteriormente compresse. Per quanto riguarda le serie binarie, secondo l'approccio assiomatico di Kolmogorov-Uspenskii, Martin-Löf, Chaitin ed altri, una sequenza di cifre binarie è detta random, nel senso che manifesta la massima irregolarità e complessità, se la lunghezza del più corto programma binario necessario a descrivere questa sequenza è almeno lungo quanto la lunghezza della serie stessa. Infatti, tutte le sequenze non random di pari lunghezza possono essere compresse e quindi rappresentate da programmi più corti.

maggiore casualità apparente del processo, una maggiore irregolarità della serie, mentre valori più bassi sono associati alla presenza di caratteristiche identificabili e quindi ad una maggiore regolarità di comportamento.

Il valore della *ApEn* associato ad una specifica sequenza dipende, oltre che dalle caratteristiche della serie analizzata, da due parametri fondamentali che devono essere sempre specificati:

- la dimensione del pattern di confronto m,
- il margine di tolleranza r.

La *ApEn* è definita come segue [36]. Dati:

- un numero intero positivo N che identifica la lunghezza della sequenza,
- una dimensione del pattern di confronto m, con $m \leq N$,
- un margine di tolleranza positivo r,
- una sequenza di numeri reali $u := (u_1, u_2, \ldots, u_N)$,

definiti:

- la distanza fra due blocchi x_i e x_j come $d(x_i, x_j) = \max_{k=1,2,\ldots,m} |u_{i+k-1} - u_{j+k-1}|$, dove $x_i = (u_i, u_{i+1}, \ldots, u_{i+m-1})$ e $x_j = (u_j, u_{j+1}, \ldots, u_{j+m-1})$, per cui due blocchi x_i ed x_j qualsiasi si considerano "vicini" se $d(x_i, x_j) \leq r$,
- $C_i^m(r) = \frac{\text{numero di } j \leq N-m+1 \text{ tali che } d(x_i,x_j) \leq r}{N-m+1}$,

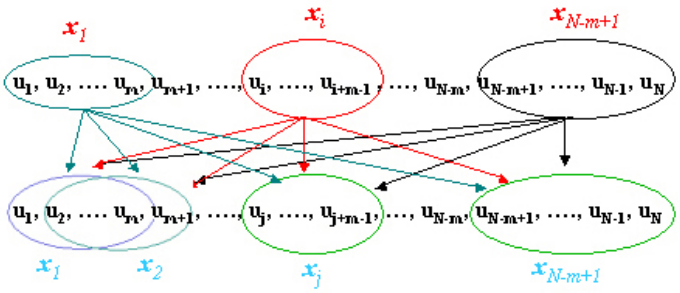

Figura 3.1: Modalità di calcolo della *ApEn*: confronto fra blocchi x_i ed x_j.

- $\Phi^m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \log C_i^m(r)$.

allora la *ApEn*[5] di parametri m ed r per una sequenza di N elementi è definita nel modo seguente:

[5] Negli esperimenti condotti nella seconda parte della tesi, per coerenza con l'entropia di Shannon e le altre entropie analizzate, verrà utilizzato il logaritmo in base due.

$$ApEn(m,r,N)(u) = \Phi^m(r) - \Phi^{m+1}(r), \quad m \geq 1$$

$$ApEn(0,r,N)(u) = -\Phi^1(r)$$

La $ApEn(m,r,N)(u)$ misura quindi il logaritmo della frequenza con cui blocchi (pattern) di lunghezza m che sono "vicini" fra loro rimangano "vicini" anche se si aumenta di un elemento la dimensione dei blocchi stessi.

La quantità $\Phi^{m+1}(r) - \Phi^m(r) = -ApEn(m,r,N)(u)$ rappresenta inoltre la media sulle i del logaritmo della probabilità condizionale che $|u_{j+m} - u_{i+m}| \leq r$ qualora $|u_{j+k} - u_{i+k}| \leq r$ per ogni $k = 1, 2, \ldots, m-1$ [13][34].

Dalla definizione emerge chiaramente che la $ApEn$ è una misura che dipende dai parametri m ed r, motivo per cui i confronti fra sequenze (serie) diverse sono giustificati solo a parità di tali parametri. La $ApEn$, poi, è praticamente insensibili al "rumore" di intensità inferiore al valore del parametro r, che qui agisce come un vero e proprio filtro. L'autore ha inoltre dimostrato come la $ApEn$ risulti robusta ed insensibile agli "outliers". Valori estremamente alti o estremamente bassi, se non si presentano in maniera molto frequente, hanno infatti un impatto assai ridotto sul calcolo della $ApEn$ [36].

3.4.2 Coerenza e scelta dei parametri m ed r

Un metodo si può considerare coerente se fornisce risultati non discordanti a seconda del modo in cui viene applicato. Poiché non ci sono rigorose indicazioni riguardo all'uso dei parametri r ed m, la coerenza è tuttora il maggior problema nell'utilizzo della $ApEn$. In altre parole, un ricercatore che scegliesse un determinata coppia (m,r) per analizzare i suoi dati, potrebbe, potenzialmente, ottenere il risultato opposto rispetto ad un altro ricercatore che avesse utilizzato una diversa coppia (n,s). Ad esempio, il processo A potrebbe apparire più random del processo B per molte coppie (m,r), ma non necessariamente per tutte. Fortunatamente, per molti processi A e B tale evenienza non si verifica. In questo caso si parla di coppie di processi completamente coerenti. Due coppie di processi si dicono completamente coerenti se ogni volta che $ApEn(m,r)(A) < ApEn(m,r)(B)$, per una qualsiasi scelta di m ed r, allora $ApEn(n,s)(A) < ApEn(n,s)(B)$ per tutte le possibili scelte di n e di s. In questo caso si può affermare che il processo B è più irregolare del processo A senza bisogno di indicare m ed r.

Si dice poi che vi è coerenza relativa su un range statisticamente valido di coppie (m,r), analogamente a quanto accade per le coppie completamente coerenti, se ogni qualvolta che $ApEn(m,r)(A) < ApEn(m,r)(B)$ per una coppia di (m,r), allora $ApEn(n,s)(A) < ApEn(n,s)(B)$ per tutte le coppie (n,s) del range [41][35]. In particolare, per sistemi privi di rumore, se l'entropia di Kolmogorov-Sinai del processo $A \leq$ dell'entropia di Kolmogorov-Sinai del processo B, allora $ApEn(A) \leq ApEn(B)$. Ciò si realizza, inoltre,

per la mappa di Enon e per la mappa Logistica, due dei sistemi caotici maggiormente studiati oltre che, in generale, per i processi stocastici, i quali spesso manifestano entropia di Kolmogorov-Sinai infinita [38].

Per evitare contributi significativi da parte di eventuali "rumori", il parametro r dovrebbe essere scelto in modo da risultare più elevato della maggior parte dei disturbi. In genere, si possono ottenere risultati significativi per valori di r pari o superiori a tre volte la media del rumore. Sulla base di calcoli teorici confortati dall'evidenza empirica, Pincus ha tratto la conclusione preliminare che per $m = 2$ ed $N = 1000$, la $ApEn$ fornisce risultati statisticamente ragionevoli per valori di r compresi fra il 10 % ed il 25% della deviazione standard della serie storica considerata. Per valori di r più piccoli, di solito si ottengono stime scadenti delle probabilità condizionali, mentre per valori di r più elevati si perdono troppe informazioni [34][38].

Poiché, in genere, una diminuzione del valore della $ApEn$ è spesso associata ad una diminuzione della deviazione standard della serie, l'esprimere r in termini percentuali rispetto alla deviazione standard consente altresì di ottenere un indicatore decorrelato da tale misura. Tale operazione consente inoltre l'invarianza di scala per la $ApEn$. Nel caso in cui il livello del rumore dovesse essere comunque molto elevato, con un rapporto segnale/rumore < 3, la validità della $ApEn$, così come quella di molte altre statistiche risulterebbe fortemente compromessa [39].

Per quanto riguarda la scelta di m, Pincus suggerisce $m = 2$ o $m = 3$ così da assicurare valori ragionevoli della probabilità condizionale stimata a partire dagli N punti della serie. Calcoli teorici mostrano che valori ragionevoli di tali probabilità condizionali si ottengono per un numero di punti compreso fra 10^m e 30^m [38].

Visti gli esiti ottenuti sui dati sperimentali, l'autore afferma che la $ApEn$ sembra fornire buoni risultati già a partire da $N \geq 60$.

3.4.3 Forme analitiche e distribuzione asintotica

Sebbene sia estremamente difficile rappresentare la $ApEn$ in forma analitica, esistono alcuni casi in cui ciò è possibile come, ad esempio, per molti processi stocastici [34]:

- dato un processo stocastico stazionario u_i avente spazio degli stati continuo, data la misura $\mu(x, y)$ della probabilità congiunta stazionaria in $\Re^2$ del processo e la probabilità di x all'equilibrio $\pi(x)$ si ha:

$$ApEn(1, r) = -\int \mu(x, y) \log \left[\int_{z=y-r}^{y+r} \int_{w=x-r}^{x+r} \mu(w, z) dw dz / \int_{w=x-r}^{x+r} \pi(w) dw \right] dx dy$$

- dato un qualsiasi processo composto da variabili aleatorie i.i.d. con densità di probabilità $\pi(x)$, per ogni $m \geq 1$ si ha:

$$ApEn(m, r) = -\int \pi(x) \log \left[\int_{z=x-r}^{x+r} \pi(z) dz \right] dx$$

- data una catena di Markov del primo ordine stazionaria con spazio degli stati discreto X, fissato $r < \min|x-y|, x \neq y$, con x ed y valori qualsiasi appartenenti allo spazio degli stati X, per ogni m, si ha:

$$ApEn(m,r) = -\sum_{x\in X}\sum_{y\in X}\pi(x)p_{xy}\log p_{xy}$$

  con $\pi(x)$ vettore di probabilità all'equilibrio e p_{xy} probabilità di transizione da un generico punto x ad un generico punto y di X.

Alhakim e Molchanov [1] e Rukhin [46] hanno inoltre dimostrato in maniera indipendente l'uno dall'altro che, sotto l'ipotesi di completa casualità della serie, la *ApEn* converge asintoticamente ad una distribuzione χ^2. In particolare, Alhakim e Molchanov hanno dimostrato che, data una sequenza di lunghezza N i cui elementi rappresentino altrettante realizzazioni di variabili binarie i.i.d., data una dimensione m del pattern di confronto e fissata la costante $c_0 = -\frac{3}{\ln 2}$, la statistica

$$N\frac{ApEn(m,N)-1}{c_0 2^{m-2}}$$

converge, per $N \to \infty$, ad una distribuzione χ^2 con 2^{m-1} gradi di libertà.

3.4.4 Distorsione e Sample-Approximate Entropy

La $ApEn(m,r,N)$ è purtroppo una statistica distorta ed il suo valore atteso aumenta fino ad un limite finito all'aumentare di N. Per questa ragione, ai fini di un corretto confronto fra due gruppi di dati, le serie considerate dovrebbero avere lo stesso numero N di componenti [37].

La distorsione della statistica $ApEn(m,r,N)$, che in molti casi risulta però essere asintoticamente corretta, è dovuta principalmente a due distinti fattori, di cui il secondo è di gran lunga il più importante [39]:

- la concavità e la non linearità della funzione logaritmo che compare nella sua definizione;
- il fatto che, nella definizione di $C_i^m(r)$, viene considerato anche il confronto del pattern x_i con se stesso (self match). Ciò trova la sua giustificazione nel fatto che, in questo modo, i calcoli, che coinvolgono i logaritmi, rimangono finiti, ma ha come conseguenza una sottostima delle probabilità condizionate che sono alla base del calcolo della *ApEn*. Nel caso in cui ci siano pochi vettori x_j "vicini" al pattern di confronto x_i per molti pattern x_i, ciò può portare a distorsioni anche dell'ordine del 20-30%.

Un tentativo per ridurre la distorsione imputabile al confronto dei pattern con se stessi è stato fatto da Richman e Moorman nel 2000 [43] con l'introduzione della *Sample-ApEn*. Essa, a differenza della *ApEn*, non considera i self matches e calcola in maniera differente le probabilità condizionali.

In particolare, per il calcolo della *ApEn* si considera la media dei logaritmi del numero di riscontri dei pattern di controllo di lunghezza m all'interno della serie e poi se ne esegue la differenza con la media calcolata sul numero di pattern di lunghezza $m+1$. Per la *Sample-ApEn*, invece, Richman e Moorman considerano la media dei riscontri di pattern di lunghezza, rispettivamente, m ed $m+1$ e poi calcolano il logaritmo del loro rapporto.

In particolare, ai fini del calcolo della *Sample-ApEn* si definiscono le seguenti quantità:

- $B_i^m(r) = \dfrac{\text{numero di } j \leq N-m \text{ ed } i \neq j \text{ tali che } d(x_i, x_j) \leq r}{N-m-1}$, dove x_i ed x_j sono vettori di m elementi consecutivi;
- $B^m(r) = \dfrac{\sum_{i=1}^{N-m} B_i^r(r)}{N-m}$

Nello stesso modo di definiscono $A_i^m(r)$ ed $A^m(r)$ dove però x_i ed x_j sono vettori composti da $m+1$ elementi consecutivi. $B^m(r)$ indica quindi la probabilità che due serie abbiano in comune m punti consecutivi mentre $A^m(r)$ indica la probabilità che ne abbiano $m+1$. Si definisce quindi la *Sample-ApEn* come:

$$\textit{Sample-ApEn} = \lim_{N \to +\infty} -\log \frac{A^m(r)}{B^m(r)}$$

che viene stimata attraverso la statistica *Sample-ApEn*(m, r, N). La *Sample-ApEn* non è definita per $B^m(r) = 0$, ovvero nel caso in cui non sia avuto alcun riscontro di ordine m, oppure quando $A^m(r) = 0$, valore che corrisponde ad una probabilità condizionata pari a zero e quindi ad un valore della *Sample-ApEn* tendente all'infinito. La *Sample-ApEn* non è quindi definita a meno che un pattern di confronto non ricorra almeno una volta vuoi per una lunghezza pari ad m che ad $m+1$.

Rispetto alla *ApEn*, gli autori dichiarano per la *Sample-ApEn* un campo di variazione più ampio, una coerenza relativa maggiore ed una distorsione minore. D'altro canto, l'applicazione pratica di entrambe le misure è rivolta principalmente a confronti fra serie diverse a parità di m, r ed N, per cui la distorsione non rappresenta un handicap molto grave in quanto non si vuole tanto ottenere una stima corretta del parametro incognito del processo, quanto piuttosto una graduatoria delle serie rispetto al loro grado di irregolarità.

3.4.5 Approximate Entropy ed entropia di Kolmogorov-Sinai

La *ApEn* deriva direttamente dalla formula per il calcolo dell'entropia di Kolmogorov-Sinai[6] sviluppata da Eckmann e Ruelle [13]:

$$\lim_{r\to 0}\lim_{m\to+\infty}\lim_{N\to+\infty}\left[\Phi^m(r)-\Phi^{m+1}(r)\right]$$

Uno dei limiti maggiori della misura di Eckmann-Ruelle e delle misure ad essa collegate risiede nel fatto che esse sono intese quali strumenti teorici e non come mezzi per discriminare sistemi dinamici a partire da dati sperimentali, ovvero insiemi di dati limitati ed affetti da rumore. Per convergere, infatti, queste formule hanno bisogno di una quantità enorme e praticamente non ottenibile di dati. Inoltre, il grado di "complessità" o di disordine di molti processi stocastici, che possono rappresentare validi modelli per diversi fenomeni fisici, varia al variare dei valori dei parametri di controllo. Tali variazioni non sono però colte da queste misure che rimangono invece costanti, spesso attestate su valori pari a zero o ad infinito. Ad esempio, l'entropia di Eckmann-Ruelle di un processo MIX[7] risulta infinita per qualsiasi valore del parametro p quando $p > 0$.

La formula di Eckmann-Ruelle e le sue varianti sono invero utili per la classificazione dei sistemi caotici a bassa dimensione: un valore diverso da zero assicura, infatti, la caoticità di un sistema conosciuto come deterministico.

Dato che il valore della formula di Eckmann-Ruelle calcolato a partire da processi affetti da rumore di qualsiasi intensità tende all'infinito, perchè tale formula possa essere applicata a dati sperimentali essa dovrà essere parametrizzata mediante opportuni valori di r. In questo caso, però, i risultati ottenuti saranno di più difficile lettura.

Pur se molto simile alla formula di Eckmann-Ruelle, da cui per altro deriva, la *ApEn* è stata costruita con uno scopo diverso e ben preciso: fornire una formula statisticamente valida e largamente applicabile per distinguere data set in relazione al loro grado di irregolarità. In particolare, fissati m ed r, la

[6] L'entropia di Kolmogorov-Sinai consente, tra l'altro, di classificare i sistemi dinamici in caotici e non caotici sulla base del loro tasso di informazione.

[7] MIX(p) è un modello misto stocastico deterministico così definito:
MIX$(p)_j = (1 - Z_j)X_j + Z_jY_j$, con:
p parametro tale che $0 \leq p \leq 1$;
$X_j = \sqrt{2}\sin(2\pi j/12)$ per ogni j;
Y_j variabili casuali uniformi i.i.d. sull'intervallo $[-\sqrt{3}, \sqrt{3}]$;
Z_j variabili casuali i.i.d. tali che $Z_j = 1$ con probabilità p e $Z_j = 0$ con probabilità $1 - p$.
All'aumentare del valore del parametro p, il processo diventa apparentemente più irregolare, imprevedibile e complesso. Il MIX(p) ha media pari a zero e standard error pari ad uno per qualsiasi valore di p, per cui questi momenti non sono in grado di discriminare differenti MIX l'uno dall'altro, così come l'entropia di Kolmogorov-Sinai, che tende all'infinito per $p > 0$ ed assume valore pari a zero per MIX(0). E questo a prescindere dalla quantità di dati disponibili.

$ApEn(m,r)$ rappresenta un parametro del processo generatore dei dati, mentre la quantità $ApEn(m,r,N)$, calcolata su un insieme di dati di numerosità N, costituisce una statistica per la $ApEn(m,r)$. La $ApEn(m,r)$ è definita come $\lim_{N\to+\infty} ApEn(m,r,N)$ in quanto, per tutti i processi ragionevoli, esiste virtualmente un unico limite con probabilità uno [39].

La *ApEn* non deve quindi essere pensata come un'approssimazione o una forma semplificata dell'entropia di Kolmogorov-Sinai, anche se, per m ed N che tendono all'infinito ed r che tende a zero la *ApEn*, per costruzione, converge alla formula di Eckmann-Ruelle. Per serie finite, infatti, questa approssimazione si realizza solo per N molto elevati, attrattori di piccole dimensioni ed m sufficientemente grandi.

La *ApEn* assume sempre valori finiti e si può applicare a processi stocastici, deterministici affetti o meno da rumore e misti. Inoltre, il fatto che la *ApEn*, a differenza della formula di Eckmann-Ruelle, assuma valori finiti per processi MIX(p), rinforza il concetto che la *ApEn* non debba essere intesa come una stima dell'entropia di Kolmogorov-Sinai, ma piuttosto come una quantità autonoma. L'utilizzo primario della *ApEn* non è quindi quello di certificare il caos, cosa che non è in grado di fare, quanto quello di riconoscere sottili alterazioni o anormalità di lungo periodo che non sarebbero altrimenti visibili. Non è invece indicata per individuare acuti cambiamenti associati a valori piuttosto infrequenti. Inoltre, se la serie non è stazionaria, ovvero se presenta uno o più trend pronunciati, crescenti o decrescenti, poco potrà essere inferito sia dalla *ApEn*, sia dalle statistiche basate sui momenti, sia dall'analisi spettrale, in quanto il trend tenderà a sovrastare le altre caratteristiche. In questo caso sarà quindi necessario depurare le serie dal trend prima di poter fare delle considerazioni sui risultati ottenuti con i vari indicatori.

La *ApEn* è inoltre intimamente legata all'entropia condizionale. Se X è una variabile casuale discreta a stati finiti avente distribuzione di probabilità $P(X = x_i) = p_i$ allora, per $r < \min_{i\neq j} |x_i - x_j|$, si ha che $ApEn(m,r) = H(X_{m+1}|X_1,\ldots,X_m)$.

3.4.6 Approximate Entropy per sequenze binarie finite

In questo paragrafo, che riporta alcuni degli elementi teorici sviluppati sull'argomento da Pincus [36], si focalizzerà l'attenzione su un particolare insieme di sequenze, ovvero sulle sequenze binarie, cioè su quelle sequenze i cui elementi sono costituiti unicamente dalle cifre (simboli) 0 ed 1, con l'avvertenza che i concetti trattati possono essere agevolmente estesi anche al caso di serie costituite da elementi appartenenti ad alfabeti k-dimensionali.

Nel seguito, il margine di tolleranza r sarà fissato ad un qualsiasi valore $r < 1$. Nel caso di sequenze binarie, infatti, fissare un valore di $r < 1$ equivale a considerare "vicini" due elementi se e solo se questi sono uguali fra loro in quanto il valore assoluto della loro differenza può assumere solo il valore 0,

nel caso di elementi uguali, ed 1, nel caso di elementi diversi. Un valore di $r \geq 1$ non avrebbe pertanto senso in quanto la disuguaglianza $d(x_i, x_j) \leq r$ risulterebbe sempre soddisfatta. Con le restrizioni di cui sopra (sequenze binarie ed $r \leq 1$), il valore della *ApEn* risulta pertanto indipendente da r e verrà quindi indicato semplicemente come $ApEn(m, N)$.

Le sequenze binarie possono essere classificate in base al loro comportamento rispetto alla *ApEn* e, di conseguenza, rispetto al livello di irregolarità che esprimono. In particolare:

- *sequenza $\{m, N\}$-irregolare*
  una sequenza di lunghezza N, u_*, è detta $\{m, N\}$-irregolare se $ApEn(m, N)(u_*) = \max_u ApEn(m, N)(u)$, dove il massimo è valutato su tutte le 2^N possibili sequenze u di lunghezza N;

- *sequenza massimamente irregolare*
  una sequenza u_* è detta massimanente irregolare, o N-irregolare, se è $\{m, N\}$-irregolare per $m = 0, 1, 2, \ldots, m_{crit}(N)$, con $m_{crit}(N)$ definito come il massimo valore di m per cui $2^{2^m} \leq N$.

L'introduzione di $m_{crit}(N)$ è motivata da un'applicazione del metodo di Ornstein e Weiss [32] per cui se

$$u = (u_1, u_2, \ldots, u_N), \quad N \geq 1$$

è una tipica realizzazione di un processo bernoulliano, allora

$$\lim_{N \to \infty} ApEn[m_{crit}(N), N](u) = h$$

dove h è l'entropia del processo.

Riguardo alle relazioni fra sequenze diverse si ha che:

- una sequenza binaria u_* di lunghezza N è detta *maggiormente $\{m, N\}$-irregolare* rispetto alla sequenza v_* se $ApEn(m, N)(u_*) > ApEn(m, N)(v_*)$;

- una sequenza binaria u_* di lunghezza N è detta *maggiormente N-irregolare* rispetto alla sequenza v_* se $ApEn(m, N)(u_*) \geq ApEn(m.N)(v_*)$, per $m = 0, 1, , 2, \ldots, m_{crit}(N)$, con la disuguaglianza in senso stretto valida per almeno un valore di $m \leq m_{crit}(N)$.

Importante è poi il concetto di *deficit dalla massima regolarità*, che indica quanto una sequenza u sia prossima, in termini di casualità, alla sequenza $\{m, N\}$-irregolare nonchè a quella massimamente irregolare. Data una sequenza binaria u di lunghezza N, Pincus [36] definisce *deficit dalla sequenza $\{m, N\}$-irregolare* la quantità:

$$\mathbf{def}_m[u] := \max_{|v|=N} ApEn(m, N)(v) - ApEn(m, N)(u)$$

dove le v sono tutte le possibili sequenze binarie di lunghezza N, mentre il *deficit dalla sequenza massimamente irregolare* è definito come:

$$\mathbf{De}[u] := \max_{m \leq m_{crit}(N)} \mathbf{def}_m[u]$$

3.4.7 Approximate Entropy per sequenze binarie non finite

Pincus [36] estende poi i concetti sviluppati per le sequenze binarie finite a sequenze binarie non finite. Data infatti una sequenza binaria non finita $u = (u_1, u_2, \ldots)$ e un valore di $r < 1$, egli estrae la stringa di lunghezza finita N, $u^{(N)} = (u_1, u_2, \ldots, u_N)$, per cui $ApEn(m, N)(u) := ApEn(m, N)[u^{(N)}]$, e definisce quale $ApEn$ della sequenza u il limite, assumendo che tale limite esista, a cui tende la $ApEn$ della stringa $u^{(N)}$ per N che diverge all'infinito, ovvero:

$$ApEn(m)(u) := \lim_{N \to \infty} ApEn(m, N)[u^{(N)}]$$

Egli definisce inoltre come *computazionalmente random* una sequenza binaria non finita u, indicata come *C-random*, se e solo se $ApEn(m)(u) = \log 2$ per tutti gli $m \geq 0$. Il caso binario può essere esteso senza difficoltà ad un alfabeto di k elementi. In questo caso una sequenza u è chiamata *C-random* se e solo se $ApEn(m)(u) = \log k$ per tutti gli $m \geq 0$.

Si può osservare come, data una sequenza infinita di variabili casuali binarie $\{X_i\}$ aventi distribuzione di probabilità $P(X = 1) = 0.5$ e $P(X = 0) = 0.5$, l'assunzione di indipendenza congiunta così come definita dalla teoria classica delle probabilità fa si che tale sequenza sia da considerarsi C-random con probabilità uno, ovvero $ApEn(m)(u) = \log 2$ per ogni m.

3.4.8 Indipendenza e numeri normali

Anche nel caso di sequenze composte da realizzazioni di variabili i.i.d. discrete ed uniformi di probabilità $p = 1/k$, l'assunzione di indipendenza congiunta fa si che esse siano da considerarsi C-random con probabilità uno, ovvero $ApEn(m)(u) = \log k$ per ogni m. Inoltre, data una sequenza infinita u, la funzione $\mathbf{def}_m[u^{(N)}]$ (deficit dalla massima irregolarità) misura la "distanza" della sequenza $u^{(N)}$ di lunghezza N dalla condizione di massima irregolarità.

In questo modo la $ApEn$ rappresenta, a differenza della complessità algoritmica, una alternativa computazionalmente applicabile per identificare sequenze completamente random.

Tramite la $ApEn$ e il concetto di deficit dalla massima irregolarità, Pincus ha sviluppato anche una definizione alternativa di numero normale. Convenzionalmente, infatti, un numero è detto normale in una data base b se nel suo sviluppo in tale base le cifre e le successioni finite di cifre appaiono tutte con la stessa frequenza. Dato un numero reale x, indicata con s una successione

finita di cifre in una determinata base b, $(b > 1)$ e con $\#(s, N)$ il numero di apparizioni di s nelle prime N cifre di x, allora x è normale nella base b se $\lim_{N\to+\infty} \#(s, N)/N = 1/b^k$ per ogni successione s di lunghezza k. Per Pincus la normalità di un numero qualsiasi si può ricondurre, semplicemente, alla condizione $ApEn(m)(u) = \log k$ per ogni m, dove k è il numero di cifre del sistema di numerazione adottato, ovvero $\mathbf{def}_m[u^{(N)}] \to 0$ per $N \to +\infty$ [40]. La condizione di normalità si realizza quindi in $\lim_{N\to+\infty} \mathbf{def}_m[u^{(N)}] = 0$ per ogni $m \geq 0$. La normalità è però una condizione limite. Per "porzioni" più o meno lunghe di numeri normali è pertanto possibile valutare solamente il loro grado di irregolarità rispetto all'irregolarità massima ammissibile per sequenze di pari lunghezza.

Pincus ha anche utilizzato la $ApEn$ e la $\mathbf{def}_m$ per studiare il comportamento di e, π, $\sqrt{2}$ e $\sqrt{3}$, numeri che si suppongono normali. Egli ha applicato la $ApEn$ alle prime N cifre dell'espansione sia binaria che decimale di e, π, $\sqrt{2}$ e $\sqrt{3}$ considerando un numero di cifre molto elevato ($N \leq 300.000$ per l'espansione binaria e $N \leq 1.000.000$ per l'espansione decimale).

I risultati ottenuti (vedi figure 3.2 e 3.3), comunque tutti molto prossimi alla condizione di massima irregolarità, mostrano, in base due, differenze considerevoli fra e, π, $\sqrt{2}$ e $\sqrt{3}$, specialmente per $m = 2$. Per N elevato, inoltre, $\sqrt{3}$ è risultato essere molto meno irregolare di π. In base dieci queste differenze permangono, ma appaiono molto meno pronunciate.

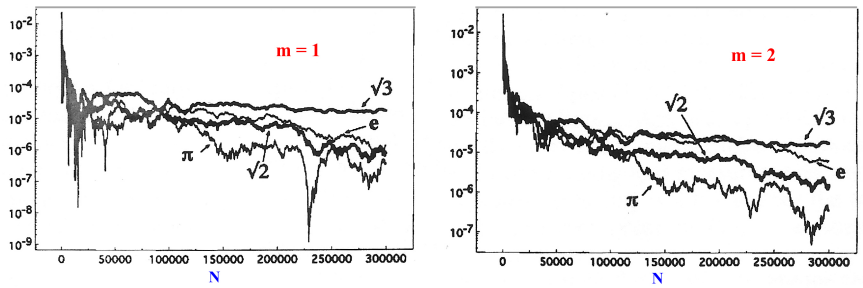

Figura 3.2: Deficit dalla massima irregolarità per l'espansione binaria di e, π, $\sqrt{2}$ e $\sqrt{3}$. $m = 1$ ed $m = 2$.

Rappresentazioni differenti di uno stesso numero possono comunque produrre sequenze con caratteristiche anche completamente diverse, motivo per cui le irregolarità riscontrate in una determinata rappresentazione non sono necessariamente esplicative riguardo alle irregolarità dello stesso numero espresso per mezzo di un'altra rappresentazione.

Per poter impostare un test statistico relativo alla condizione di completa casualità[8] basato sulla $ApEn$ è necessario derivarne la distribuzione limite

[8] Il concetto di numero casuale è fondamentale in molti settori fra cui, ad esempio, la criptografia. Gli algoritmi di criptaggio, infatti, si basano su generatori di numeri

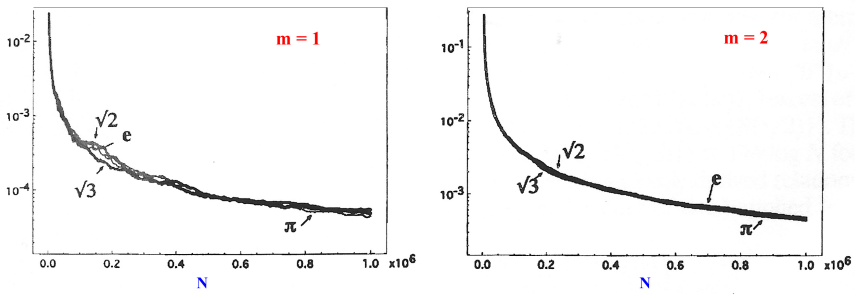

Figura 3.3: Deficit dalla massima irregolarità per l'espansione decimale di e, π, $\sqrt{2}$ e $\sqrt{3}$. $m = 1$ ed $m = 2$.

sotto tale condizione. Rukhin [46] ha dimostrato che, fissato m e posto $r = 0$, per N tendente all'infinito si ha:

$$2N[\log k - ApEn(m)] \to \chi^2(k^{m+1} - k^m) = \chi^2(k-1)k^m$$

con k numero di simboli dell'alfabeto. Dato un valore osservato della $ApEn(m)$, definito $\chi^2(\text{obs}) = 2N[\log k - ApEn(m)]$, il P-value pertanto è:

$$P_N(m) = 1 - \Gamma[2^{m-1}, \chi^2(\text{obs})/2]$$

dove il simbolo Γ indica la funzione gamma incompleta.

3.5 Cross Approximate Entropy

Il termine correlazione è diventato nel tempo sinonimo di "associazione" e "corrispondenza" anche se, matematicamente parlando, si tratta di una misura di dipendenza fra le tante possibili.

La correlazione, ed in particolar modo la correlazione lineare, è la sintesi naturale della dipendenza nelle distribuzioni normali multivariate e nei sistemi lineari. Il vastissimo utilizzo come modello della distribuzione normale multivariata e lo sviluppo di teorie matematiche da essa derivate ha elevato, di fatto, la correlazione ad indicatore "universale" di associazione.

Nel caso in cui, però, la correlazione venga utilizzata con modelli e dati di tipo più generale, per i quali il legame fra correlazione ed associazione non è così rigoroso, si possono avere problemi significativi ed interpretazioni fuorvianti.

Per questo motivo, nel 1996, Pincus e Singer [41] proposero una misura alternativa, definita *cross-ApEn* che, a detta degli autori, sarebbe in grado

(pseudo) casuali. La verifica della casualità di tali generatori diventa fondamentale per l'industria delle telecomunicazioni dove la firma digitale e la gestione delle chiavi sono di vitale importanze per l'elaborazione e la sicurezza delle informazioni.

di cogliere il concetto di associazione fra due variabili, sia da un punto di vista teorico che empirico, in maniera più generale rispetto alla correlazione. In particolare, la *cross-ApEn* rappresenta una misura della "asincronia" o "irregolarità condizionale" fra due sequenze u e v.

Come l'*ApEn*, da cui deriva, la *cross-ApEn* costituisce una famiglia di statistiche caratterizzate da due parametri:

- la dimensione del pattern di confronto m;
- il margine di tolleranza o "errore" r ammissibile nel confronto fra elementi corrispondenti appartenenti alle due sequenze in esame.

Da un punto di vista computazionale, l'algoritmo di calcolo è sostanzialmente lo stesso di quello della *ApEn*. L'unica differenza consiste nel fatto che le corrispondenze del pattern di confronto e del suo elemento successivo non vengono più verificate all'interno di una stessa serie u, ma all'interno di una serie distinta v. In particolare, la *cross-ApEn* è definita come segue [36]. Dati:

- le sequenze finite $u = (u_1, u_2, \ldots, u_N)$ e $v = (v_1, v_2, \ldots, v_N)$ composte da N elementi ciascuna;
- i parametri m ed r;
- i blocchi (pattern) di lunghezza m $x_i = (u_i, u_{i+1}, \ldots, u_{i+m-1})$ ed $y_j = (v_j, v_{j+1}, \ldots, v_{j+m-1})$ tratti, rispettivamente, da u e da v;

definiti:

- per ogni $i \leq N - m + 1$, la quantità

$$C_i^m(r)(v||u) = \frac{\text{numero di } j \leq N - m + 1 \text{ tali che } d(x_i, x_j) \leq r}{N - m + 1}$$

  come la misura, con una tolleranza massima pari ad r, della regolarità, o frequenza, con cui un determinato pattern x_i di u risulta simile ai vari possibili pattern y_j di lunghezza m di v e dove, analogamente al caso della *ApEn*, $d(x_i, y_j) = \max_{k=1,2,\ldots,m} |u_{i+k-1} - v_{j+k-1}|$ è la massima differenza fra le rispettive componenti scalari dei due blocchi x_i ed y_j;

- $\Phi^m(r)(v||u) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \log C_i^m(r)(v||u)$;

si definisce allora *cross-ApEn* di parametri m, r la quantità:

$$\textit{cross-ApEn}(m, r, N)(v||u) = \Phi^m(r)(v||u) - \Phi^{m+1}(r)(v||u), \quad m \geq 1;$$

$$\textit{cross-ApEn}(0, r, N)(v||u) = -\Phi^1(r)(v||u)$$

La *cross-ApEn* quantifica quindi la sincronia relativa di due segnali interconnessi. In questo contesto la sincronia si riferisce alla similarità di pattern, ovvero a quante volte un pattern di una serie che si presenta nell'altra, si ritrova ancora se si aumenta la sua dimensione di un elemento. Non si tratta pertanto di una sincronia di tipo temporale. Un minore sincronismo fra le serie implica una minore corrispondenza fra patterns e quindi valori di *cross-ApEn* più elevati. Tale metodologia consentirebbe quindi un approccio diverso alla valutazione di segnali bivariati e dei loro cambiamenti senza richiedere la conoscenza di un modello multivariato completo.

Pincus consiglia di utilizzare, per serie storiche standardizzate, valori di $m = 1$ oppure $m = 2$ ed $r = 0.2$, in quanto tali valori assicurerebbero buone proprietà di replicabilità.

Per sistemi caratterizzati da una dipendenza debole, è stato osservato come valori di $m \geq 3$ ed r piccolo, portino a risultati di scarsa coerenza, mentre un numero consistente di "eventi rari" si tradurrebbe, inevitabilmente, in un aumento della distorsione della statistica. Per questo tipo più generale di serie storiche è pertanto fondamentale procedere con prudenza scegliendo valori di m piccoli (ad esempio $m = 2$) e valori di r moderati.

Poiché il valore della *cross-ApEn*, come d'altronde quello della *ApEn*, in genere aumenta all'aumentare del rumore, è necessario comparare serie che presentino livelli di rumore dello stesso ordine. Per ovviare a problemi legati all'ordine di grandezza delle variabili considerate, è inoltre opportuno procedere alla standardizzazione degli elementi delle serie che si vogliono comparare: se $u_1, u_2, \ldots, u_n$ è la serie originaria e $u_1^*, u_2^*, \ldots, u_n^*$ è la serie standardizzata, allora $u_i^* = \dfrac{u_1 - \overline{u}}{\sigma_u}$.

La *cross-ApEn* è stata applicata con successo, ad esempio, in biologia e fisiologia per misurare la variazione congiunta dei livelli di secrezione ormonale in alcune patologie [27][44], per valutare la relazione fra l'attività simpatica del nervo renale e la pressione arteriosa nei topi [56], ecc.

3.5.1 Limiti intrinseci della cross Approximate Entropy: cross Sample Approximate Entropy

Uno dei limiti più evidenti della *cross-ApEn* è insito nella sua stessa definizione. E' infatti sufficiente che anche una sola delle quantità $C_i^m(r)(v||u)$ assuma valore zero, ovvero che un qualsiasi pattern x_i della serie u non trovi riscontro nella serie v, perchè la *cross-ApEn* non sia calcolabile dato che, altrimenti, si avrebbe $\log C_i^m(r)(v||u) \to -\infty$. Nel caso della *ApEn* questo problema non si poneva in quanto, confrontando una serie con se stessa, vi era sempre almeno un riscontro per ogni pattern, quello del pattern con se stesso.

Per ovviare a questo problema, Richman e Moorman [43] hanno elaborato la *cross Sample-ApEn*, un indicatore, derivato direttamente dalla *Sample-*

ApEn discussa nella precedente sezione 3.4.4, che si basa sul logaritmo del rapporto della media dei riscontri di pattern di lunghezza, rispettivamente, m ed $m+1$.

In particolare, ai fini del calcolo della *cross Sample-ApEn* Richman e Moorman definiscono le seguenti quantità:

- $B_i^m(r)(v||u) = \dfrac{\text{numero di } j \leq N-m \text{ tali che } d(x_i, x_j) \leq r}{N-m}$, dove x_i ed x_j sono vettori di m elementi consecutivi tratti, rispettivamente, dalle serie u e v;
- $B^m(r)(v||u) = \dfrac{\sum_{i=1}^{N-m} B_i^r(r)(v||u)}{N-m}$

Allo stesso modo essi definiscono $A_i^m(r)(v||u)$ ed $A^m(r)(v||u)$, dove però x_i ed x_j sono vettori composti da $m+1$ elementi consecutivi. $B^m(r)(v||u)$ indica quindi la probabilità che due serie u e v abbiano in comune m punti consecutivi mentre $A^m(r)(v||u)$ indica la probabilità che ne abbiano $m+1$. Gli autori definiscono quindi la *cross Sample-ApEn* come:

$$\textit{cross Sample-ApEn} = \lim_{N \to +\infty} -\log \frac{A^m(r)(v||u)}{B^m(r)(v||u)}$$

che viene stimata attraverso la statistica *cross Sample-ApEn*(m, r, N). La *Sample-ApEn* risulta quindi definità se $A^m(r)(v||u) \neq 0$, ovvero quando vi sia almeno una corrispondenza di pattern di lunghezza $m+1$ fra le due serie u e v. Condizione, questa, molto meno restrittiva rispetto a quella richiesta per la *cross-ApEn*, per la quale era necessario che ciascuno degli $N-m$ possibili pattern di lunghezza $m+1$ della serie u tovassero almeno una corrispondenza nella serie v.

La *cross Sample-ApEn* risulta inoltre simmetrica, ovvero *cross Sample-ApEn*$(v||u)$ = *cross Sample-ApEn*$(u||v)$, cosa che, in genere, non si verifica per la *cross-ApEn*.

Parte II

Analisi dell'autocorrelazione e del livello di disordine in serie binarie

CAPITOLO 4

Introduzione

Scopo di questa seconda parte della tesi è quello di effettuare una serie di esperimenti con gli indicatori precedentemente introdotti, ovvero con:

- la misura di Granger, Maasoumi e Racine, indicata come S_ρ;
- l'entropia congiunta;
- la mutua informazione;
- l'entropia relativa, anche detta distanza di Kullback Leibler;
- l'entropia approssimata o Approximated Entropy, indicata come $ApEn$;

utilizzandoli nello studio (autocorrelazione e livello di irregolarità) di serie binarie appositamente create ed, in particolare su:

- serie perfettamente periodiche a cui sono stati applicati diversi livelli di perturbazione casuale (rispettivamente il 25%, 50%, 75% e 100% degli elementi della serie originaria viene sostituito mediante estrazione casuale di nuovi elementi scelti fra i due possibili: 0 oppure 1);
- serie semplici perfettamente periodiche con diverse lunghezze del periodo;
- serie deterministiche complesse perfettamente periodiche;
- serie deterministiche complesse non periodiche;

- serie generate da un processo stocastico;
- serie generate da una mappa logistica in regime caotico.

Dato che le forme normalizzate dell'entropia congiunta e della mutua informazione coincidono, come si avrà modo di osservare nel corso del successivo capitolo 5, a partire dal capitolo 6 l'analisi sperimentale dell'entropia congiunta verrà abbandonata per concentrarci sulla mutua informazione che, a parità di altre condizioni, mostra varianza inferiore.

Per l'esecuzione degli esperimenti previsti sono state utilizzate serie composte da 1000 elementi. Mediante l'esecuzione di test preliminari si è visto infatti che tale lunghezza risultava più che sufficiente per far emergere la tendenza di fondo dei fenomeni considerati. Una lunghezza di 1000 elementi rappresenta quindi, in questo caso, un buon compromesso fra l'attendibilità dei risultati, i tempi di elaborazione necessari al calcolo degli indicatori (tempo che per la *ApEn* cresce esponenzialmente all'aumentare della lunghezza della serie) e le dimensioni delle serie che normalmente si possono incontrare nella pratica, dove non è quasi mai possibile disporre di serie indefinitamente lunghe.

Per quanto riguarda il numero di replicazioni, si è osservato che già con qualche centinaio di casi si otteneva una buona stabilità dei risultati. In via cautelativa e per ottenere stime massimamente attendibili, si è comunque convenuto di considerare un numero di simulazioni pari a 5000, numero ridotto a 1000 nel caso della *ApEn*, che è caratterizzata da tempi di elaborazione sensibilmente più elevati rispetto a quelli relativi agli altri indicatori ma che appare, nel contempo, anche più stabile.

4.1 Generazione di numeri pseudocasuali

Nel corso della trattazione si è fatto largo uso di funzioni per la generazione di numeri pseudocasuali a partire da una probabilità data. In particolare è stata utilizzata la funzione "sample" di R. Si è quindi ritenuto opportuno "testare" gli indicatori utilizzati al fine di verificare che i risultati ottenuti a partire da serie pseudocasuali generate dal computer concordassero con quelli ottenuti su serie sicuramente casuali. A tale scopo sono stati applicati gli indicatori precedentemente citati a:

- una serie di controllo "veramente" casuale, scaricata dal sito web *www.random.org* e composta da dati binari (1000 elementi) di origine meteorologica;
- su 5000 serie (1000 nel caso della *ApEn*) di 1000 elementi ciascuna generate in maniera pseudocasuale mediante la funzione "sample(c(0,1), N, c(0.5,0.5), replace = TRUE)" di "R", dove gli elementi 0 ed 1 avevano pari probabilità di esservi inclusi.

Tabella 4.1: Valori dell'entropia congiunta, della mutua informazione, della S_ρ e della *ApEn* calcolati su di una serie binaria "veramente" casuale (dati di origine meteorologica).

Lunghezza della serie = 1000

Frequenza di zeri = 0.509

Lag	Sp	Entropia Congiunta	Mutua Informazione	ApEn
1	0,000695	1,998359	0,001174	0,998730
2	0,001142	1,997604	0,001928	0,996158
3	0,000058	1,999434	0,000098	0,995492
4	0,000065	1,999423	0,000110	0,986006
5	0,001325	1,997294	0,002239	0,954174
6	0,000239	1,999128	0,000405	0,924141
7	0,000165	1,999253	0,000279	0,872314
8	0,000177	1,999234	0,000298	0,748415
9	0,000327	1,998980	0,000553	0,582550
10	0,000105	1,999355	0,000177	0,389506

Tabella 4.2: Valori dell'entropia congiunta, della mutua informazione, della S_ρ e della *ApEn* e relativi intervalli di confidenza al 95% calcolati a partire da 5000 serie pseudocasuali generate dal computer.

Lunghezza delle serie = 1000, numero di casi = 5000 (1000 nel caso dell'ApEn)

Lag	S_ρ		Entropia Congiunta		Mutua Informazione		ApEn	
1	0,000413	0,000001 0,003170	1,997846	1,991511 1,999948	0,000695	0,000001 0,003353	0,999590	0,995910 1,000975
2	0,000434	0,000002 0,003391	1,997811	1,991468 1,999948	0,000729	0,000001 0,003538	0,998069	0,992927 1,000639
3	0,000429	0,000003 0,003306	1,997820	1,991631 1,999948	0,000720	0,000001 0,003574	0,995120	0,987931 0,999501
4	0,000428	0,000003 0,003481	1,997822	1,991635 1,999937	0,000717	0,000001 0,003632	0,989299	0,980197 0,995920
5	0,004380	0,000003 0,003521	1,997807	1,991302 1,999948	0,000732	0,000001 0,003740	0,977354	0,963595 0,987586
6	0,000432	0,000003 0,003447	1,997820	1,991455 1,999948	0,000720	0,000001 0,003615	0,952366	0,934143 0,967592
7	0,000423	0,000003 0,003064	1,997836	1,991680 1,999942	0,000704	0,000001 0,003585	0,897223	0,869938 0,922625
8	0,000427	0,000003 0,003314	1,997830	1,991573 1,999942	0,000709	0,000001 0,003526	0,773809	0,741075 0,806146
9	0,000428	0,000003 0,003634	1,997831	1,991440 1,999948	0,000709	0,000001 0,003531	0,573604	0,539532 0,603616
10	0,000442	0,000003 0,003572	1,997808	1,991634 1,999948	0,000732	0,000001 0,003605	0,365523	0,332746 0,397083

A partire dai risultati ottenuti su ciascuna singola serie pseudocasuale è stata poi calcolata, per ogni indicatore, la distribuzione empirica con il valore medio ed il relativo l'intervallo di confidenza al 95%. I risultati sono riportati nelle tabelle 4.1 e 4.2.

Come era logico aspettarsi, visto che la maggior parte delle routine per la generazione di numeri pseudocasuali sono estremamente collaudate e performanti, vi è sostanziale identità di risultati fra i due set di prove, a conferma dell'affidabilità della routine di R utilizzata e degli indicatori studiati.

CAPITOLO 5

Analisi di serie perfettamente periodiche affette da rumore

Sono state effettuate una serie di simulazioni a partire da cinque diverse condizioni iniziali, ovvero da cinque tipi di serie perfettamente periodiche di 1000 elementi ciascuna e così strutturate:

- *serie 1*: serie di periodo due, pattern "01", caratterizzata dal 50% di 0 e dal 50% di 1;
- *serie 2*: serie di periodo quattro, pattern "0011", caratterizzata dal 50% di 0 e dal 50% di 1;
- *serie 3*: serie di periodo dieci, pattern "0000011111", caratterizzata anch'essa dal 50% di 0 e dal 50% di 1;
- *serie 4*: serie di periodo dieci, pattern "0000000111", caratterizzata dal 70% di 0 e dal 30% di 1;
- *serie 5*: serie di periodo dieci, pattern "0000000001", caratterizzata dal 90% di 0 e dal 10% di 1.

Tali serie possono essere considerate fra loro pressoché equivalenti dal punto di vista della complessità algoritmica (vedi precendete capitolo 2), in quanto possono essere tutte descritte per mezzo del seguente programma minimale:

$K = \ldots$
$J = \ldots$
$T = \ldots$
scrivi K zero e poi J uno T volte.

Per ciascuna condizione iniziale sono stati calcolati i seguenti indicatori:

- l'entropia di Shannon;
- l'entropia di Granger, Maasoumi e Racine (S_ρ);
- l'entropia congiunta;
- la mutua informazione;
- le distanze di Kullback Leibner;
- la Approximate Entropy ($ApEn$).

Il valore della S_ρ, dell'entropia congiunta, della mutua informazione e delle distanze di Kullback Leibner sono state calcolate considerando covariate con lag da 0 a 10 in modo da avere almeno un lag di ampiezza uguale ad un multiplo del periodo della serie originaria (condizione di perfetta autocorrelazione). Per lo stesso motivo, per il calcolo della $ApEn$ è stato considerato un valore di m variabile da 0 a 10.

5.1 Modalità di esecuzione dell'esperimento

Le serie di periodo due, quattro e dieci, caratterizzate da una percentuale iniziale di 0 e di 1 pari al 50%, sono state "perturbate" mediante la sostituzione, rispettivamente, del 25, 50, 75 e 100% delle loro componenti, scelte a caso, con una sequenza di 0 e di 1 generati in maniera casuale con probabilità pari a $p = q = 0.5$, dove $p = P(X = 1)$ e $q = P(X = 0)$.

Le due serie di periodo 10 caratterizzate da una percentuale iniziale di 0 pari, rispettivamente, al 70% e al 90%, sono state anch'esse perturbate mediante la sostituzione del 25, 50, 75 e 100% delle loro componenti con una sequenza di 0 e di 1 generati casualmente con probabilità pari, per la prima serie, a $q = 0.7$, $p = 0.3$ e, per la seconda serie, a $q = 0.9$, $p = 0.1$. In questo modo, in media, si è mantenuta inalterata la percentuale iniziale di 0 e di 1 caratterizzante ciascuna serie. Per quanto riguarda le serie perturbate al 100%, esse non sono altro che serie pseudocasuali interamente generate dal computer a partire da una distribuzione di probabilità caratterizzata, a seconda dei casi, da $p = 0.5$, $p = 0.3$ o $p = 0.1$.

Per ciascuna modalità (serie iniziale e livello di perturbazione) sono stati simulati al calcolatore, in maniera indipendente l'uno dall'altro, 5000 casi. Nel caso della $ApEn$, per ragioni di ottimizzazione del tempo macchina, si

sono comunque considerate "solo" le prime 1000 simulazioni. Per ogni caso sono stati calcolati gli indicatori di cui sopra e poi ne è stata fatta la media per ottenere dei valori di "modalità" che potessero esprimere l'andamento atteso del fenomeno.

Oltre al valor medio, per ciascuna modalità sono stati calcolati, a partire dalle distribuzioni empiriche degli indicatori così ottenute, i valori corrispondenti ai percentili 0.25 e 99.75, in modo da ottenere intervalli di confidenza al 95% per la media, e la varianza.

5.2 Risultati dell'esperimento

5.2.1 Entropia di Shannon

Com'era prevedibile, essendo direttamente legata alla distribuzione di probabilità del fenomeno (vedi precedente capitolo 1.2), l'entropia di Shannon, $H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$, si mantiene stabile e, dato p, molto prossima al suo valore teorico, poiché rimane stabile, in media, la proporzione di 0 e di 1 delle serie al variare del livello di randomizzazione considerato. L'entropia di Shannon, che indica il livello di "informazione" insito in una distribuzione di probabilità associata ad una variabile casuale, non è pertanto in grado, per definizione, né di discriminare il grado di dipendenza degli elementi di una serie, né il grado di complessità (ad esempio quella algoritmica) della serie stessa.

Tabella 5.1: Valori dell'entropia di Shannon e relativi intervalli di confidenza al 95% in funzione del pattern e del livello di randomizzazione considerato.

Lunghezza della serie = 1000, numero di casi = 5000

		Livello di randomizzazione									
Serie	Valore teorico	0%	25%		50%		75%		100%	
p=0.5 periodo 2	1	1	0,999689	0,998473 1,000000	0,999449	0,997402 1,000000	0,999328	0,996662 0,999997	0,999304	0,996662 0,999997
p=0.5 periodo 4	1	1	0,999683	0,998473 1,000000	0,999452	0,997225 1,000000	0,999290	0,996257 1,000000	0,999295	0,996463 0,999997
p=0.5 periodo 10	1	1	0,999686	0,998337 1,000000	0,999463	0,997225 1,000000	0,999307	0,996463 1,000000	0,999295	0,996463 0,999997
p=0.3	0,881291	0,881291	0,881209	0,858162 0,902193	0,880562	0,849943 0,909736	0,880565	0,845737 0,911832	0,880732	0,842896 0,912870
p=0.1	0,468996	0,468996	0,468831	0,429759 0,508899	0,468234	0,416119 0,517753	0,468471	0,409187 0,523584	0,468245	0,405693 0,526480

5.2.2 Entropia di Granger, Maasoumi e Racine

La quantità $S_\rho = \frac{1}{2}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\left(\sqrt{f(x,y)} - \sqrt{g(x)}\sqrt{h(y)}\right)^2$ è una misura metrica della dipendenza fra due serie o fra gli elementi di una stessa serie (vedi precedente capitolo 3.3), dove:

- $f(X,Y)$ indica la distribuzione congiunta delle variabili casuali X ed Y;
- $g(X)$ indica la distribuzione di X;
- $h(y)$ indica la distribuzione di Y.

Nel nostro caso, trattandosi di serie binarie, una stima corretta delle distribuzioni marginali può essere calcolata direttamente mediante il rapporto fra il numero di 1 e di 0 caratterizzanti le singole serie e il numero complessivo, n, di elementi delle stesse. Per quanto riguarda, invece, la distribuzione congiunta, tramite "R" è stato effettuato il conteggio del numero di realizzazioni di ciascuna delle quattro possibili modalità "00", "01", "10", "11", realizzazioni che sono poi state rapportate, anche in questo caso, al numero totale degli elementi della serie.

La quantità S_ρ è particolarmente indicata per valutare l'indipendenza in distribuzione di una serie e, per costruzione, indica indipendenza se e solo se $S_\rho = 0$. Il valore relativo alla condizione di massima autocorrelazione varia invece da serie a serie in funzione della sua distribuzione di probabilità e si può ottenere confrontando la serie con se stessa (lag = 0). In tale situazione avremo, ovviamente, distribuzioni marginali uguali:

$$g(x) = h(y) \text{ con } g(x) = \begin{cases} q & \text{se } x = 0 \\ p & \text{se } x = 1 \end{cases}$$

Inoltre, poiché ogni elemento è associato sempre e solo con se stesso, la distribuzione di probabilità congiunta $f(x,y)$ sarà del tipo:

$$f(x,y) = \begin{cases} 0 & \text{se } x \neq y \\ q & \text{se } x = y = 0 \\ p & \text{se } x = y = 1 \end{cases}$$

Sostituendo i valori delle probabilità di cui sopra nell'espressione

$$S_\rho = \frac{1}{2}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\left(\sqrt{f(x,y)} - \sqrt{g(x)}\sqrt{h(y)}\right)^2$$

e svolgendo i calcoli, si ottiene la seguente formula per il massimo della S_ρ:

$$S_{\rho(\max)} = \frac{1 + p(1-2\sqrt{p}) + q(1-2\sqrt{q})}{2} = 1 - p\sqrt{p} - q\sqrt{q}$$

Per le serie di periodo due, quattro e dieci, caratterizzate da $p = 0.5$, tale valore corrisponde a circa 0.2929, mentre per le restanti serie di periodo dieci, a $p = 0.3$ corrisponde un valore pari a circa 0.25 e a $p = 0.1$ circa 0,1146. Come prevedibile, nel caso di serie perfettamente periodiche, il valore di massima autocorrelazione si osserva anche in corrispondenza di lag multipli del periodo della serie (vedi tabelle da 5.2 a 5.6). Dai dati tabulati si può poi osservare come i valori di $S\rho$ relativi a serie diverse, oppure a lag diversi all'interno di una stessa serie, non siano praticamente più distinguibile fra loro già ad un livello di randomizzazione del 75%. A partire da tale livello, inoltre, essi risultano tutti compatibili con l'ipotesi di indipendenza distributiva ($\alpha = 0.05$).

Bisogna dire, comunque, che il test utilizzato (in sostanza l'intervallo di confidenza ottenuto dalla distribuzione empirica) è di tipo bidirezionale quando, probabilmente, nel caso del livello di randomizzazione massima (100%), sarebbe stato forse più opportuno considerare un test monodirezionale. Il limite inferiore che l'indicatore può assumere in tale circostanza è infatti noto a priori e vale zero. Per una questione di omogeneità, si è preferito però utilizzare il medesimo criterio di costruzione degli intervalli di confidenza per tutti i livelli di randomizzazione. Nel caso di randomizzazione pari al 100% ciò si traduce, inevitabilmente, in una leggera sovrastima del limite superiore dell'intervallo di confidenza, fatto che comunque non inficia i risultati ed i ragionamenti di tipo più generale che si vanno costruendo.

Tabella 5.2: Valori della S_ρ e relativi intervalli di confidenza al 95% per diversi valori di lag e di randomizzazione. Serie di periodo 2 e $p = 0.5$.

Pattern '01', P=0.5, lunghezza della serie = 1000, numero di casi = 5000

	Livello di randomizzazione									
Lag	0%	25%		50%		75%		100%		
0	0,292893	0,292779	0,292332 0,292893	0,292691	0,291938 0,292893	0,292646	0,291667 0,292892	0,292638	0,291667 0,292892
1	0,292893	0,044737	0,035387 0,054827	0,008126	0,004326 0,012803	0,000638	0,000007 0,001998	0,000127	0,000000 0,000637
2	0,292893	0,044354	0,034978 0,054962	0,008057	0,004182 0,012624	0,000606	0,000004 0,002013	0,000126	0,000001 0,000614
3	0,292893	0,044562	0,035116 0,054878	0,008141	0,004383 0,012717	0,000628	0,000006 0,002045	0,000125	0,000001 0,000605
4	0,292893	0,044422	0,035264 0,054952	0,008042	0,004275 0,012868	0,000598	0,000005 0,001980	0,000129	0,000001 0,000642
5	0,292893	0,044652	0,035391 0,055014	0,008134	0,004380 0,012736	0,000630	0,000008 0,002010	0,000125	0,000001 0,000611
6	0,292893	0,044409	0,034979 0,054947	0,008043	0,004277 0,012620	0,000616	0,000005 0,001982	0,000125	0,000001 0,000640
7	0,292893	0,044633	0,035364 0,054935	0,008128	0,004281 0,012668	0,000639	0,000009 0,001985	0,000128	0,000001 0,000618
8	0,292893	0,044416	0,034978 0,054858	0,008043	0,004279 0,012692	0,000608	0,000007 0,001975	0,000129	0,000001 0,000615
9	0,292893	0,044661	0,035628 0,055038	0,008156	0,004322 0,012753	0,000645	0,000007 0,002061	0,000128	0,000001 0,000624
10	0,292893	0,044396	0,034977 0,054550	0,008021	0,004329 0,012402	0,000606	0,000006 0,001987	0,000128	0,000001 0,000621

Tabella 5.3: Valori della S_ρ e relativi intervalli di confidenza al 95% per diversi valori di lag e di randomizzazione. Serie di periodo 4 e $p = 0.5$.

Pattern '0011', P=0.5, lunghezza della serie = 1000, numero di casi = 5000

	Livello di randomizzazione									
Lag	0%	25%		50%		75%		100%		
0	0,292893	0,292777	0,292332 0,292893	0,292696	0,291874 0,292893	0,292639	0,291594 0,292893	0,292624	0,291518 0,292893
1	0	0,000023	0,000000 0,000118	0,000073	0,000000 0,000353	0,000106	0,000000 0,000518	0,000124	0,000001 0,000619
2	0,292893	0,044747	0,035364 0,055128	0,008169	0,004290 0,012925	0,000631	0,000005 0,001995	0,000124	0,000001 0,000639
3	0	0,000025	0,000000 0,000126	0,000072	0,000001 0,000362	0,000114	0,000001 0,000561	0,000125	0,000001 0,000634
4	0,292893	0,044459	0,034979 0,054935	0,008099	0,004286 0,012723	0,000606	0,000005 0,001990	0,000123	0,000001 0,000640
5	0	0,000025	0,000000 0,000124	0,000072	0,000001 0,000367	0,000108	0,000001 0,000532	0,000127	0,000001 0,000648
6	0,292893	0,044706	0,035607 0,055084	0,008177	0,004436 0,012782	0,000635	0,000008 0,001995	0,000130	0,000001 0,000649
7	0	0,000025	0,000000 0,000125	0,000072	0,000001 0,000366	0,000110	0,000001 0,000550	0,000129	0,000001 0,000626
8	0,292893	0,044478	0,035228 0,054957	0,008087	0,004277 0,012793	0,000619	0,000007 0,002057	0,000128	0,000001 0,000642
9	0	0,000024	0,000000 0,000113	0,000072	0,000001 0,000351	0,000112	0,000001 0,000543	0,000125	0,000001 0,000622
10	0,292893	0,044706	0,035411 0,055011	0,008152	0,004339 0,012896	0,000638	0,000009 0,002003	0,000131	0,000001 0,000650

Tabella 5.4: Valori della S_ρ e relativi intervalli di confidenza al 95% per diversi valori di lag e di randomizzazione. Serie di periodo 10 e $p = 0.5$.

Pattern '0000011111', P=0.5, lunghezza della serie = 1000, numero di casi = 5000

Lag	Livello di randomizzazione									
	0%	25%		50%		75%		100%		
0	0,292893	0,292780	0,292332 0,292893	0,292692	0,291874 0,292893	0,292639	0,291594 0,292893	0,292633	0,291594 0,292893
1	0,051317	0,014854	0,010638 0,019354	0,002901	0,000962 0,005370	0,000299	0,000001 0,001228	0,000125	0,000001 0,000618
2	0,005064	0,001634	0,000649 0,002757	0,000398	0,000003 0,001288	0,000131	0,000001 0,000651	0,000128	0,000001 0,000649
3	0,005064	0,001655	0,000690 0,002837	0,000416	0,000005 0,001358	0,000138	0,000001 0,000654	0,000124	0,000001 0,000638
4	0,051317	0,014943	0,010734 0,019588	0,002972	0,001016 0,005648	0,000315	0,000001 0,001261	0,000127	0,000001 0,000625
5	0,292893	0,044708	0,035498 0,055183	0,008144	0,004402 0,012726	0,000629	0,000007 0,002008	0,000129	0,000001 0,000618
6	0,051317	0,014945	0,010692 0,019571	0,002967	0,001062 0,005524	0,000316	0,000001 0,001326	0,000129	0,000001 0,000651
7	0,005064	0,001655	0,000693 0,002837	0,000407	0,000005 0,001313	0,000144	0,000001 0,000708	0,000123	0,000001 0,000605
8	0,005064	0,001634	0,000677 0,002816	0,000387	0,000004 0,001243	0,000134	0,000001 0,000657	0,000123	0,000001 0,000589
9	0,051317	0,014870	0,010495 0,019541	0,002926	0,001044 0,005459	0,000293	0,000001 0,001208	0,000129	0,000001 0,000618
10	0,292893	0,044495	0,034990 0,055366	0,008006	0,004367 0,012419	0,000613	0,000007 0,002014	0,000129	0,000001 0,000621

Tabella 5.5: Valori della S_ρ e relativi intervalli di confidenza al 95% per diversi valori di lag e di randomizzazione. Serie di periodo 10 e $p = 0.3$.

Pattern '0000000111', P=0.3, lunghezza della serie = 1000, numero di casi = 5000

Lag	Livello di randomizzazione									
	0%	25%		50%		75%		100%		
0	0,250021	0,249934	0,241376 0,257849	0,249881	0,238474 0,259774	0,249731	0,236991 0,260902	0,249692	0,236991 0,261283
1	0,033791	0,010370	0,006997 0,014022	0,002174	0,000604 0,004329	0,000247	0,000001 0,001045	0,000129	0,000000 0,000649
2	0,000279	0,000142	0,000000 0,000573	0,000110	0,000000 0,000556	0,000119	0,000001 0,000589	0,000129	0,000001 0,000656
3	0,055285	0,008764	0,005804 0,012343	0,001662	0,000382 0,003520	0,000225	0,000001 0,001013	0,000129	0,000001 0,000665
4	0,055285	0,008798	0,005323 0,012977	0,001648	0,000299 0,003635	0,000215	0,000001 0,000961	0,000127	0,000001 0,000626
5	0,055285	0,008867	0,004493 0,014311	0,001702	0,000235 0,004062	0,000242	0,000001 0,001127	0,000125	0,000001 0,000631
6	0,055285	0,008799	0,005164 0,013092	0,001662	0,000314 0,003685	0,000223	0,000001 0,001033	0,000122	0,000001 0,000618
7	0,055285	0,008716	0,005725 0,012273	0,001651	0,000379 0,003536	0,000217	0,000001 0,000978	0,000129	0,000001 0,000630
8	0,000279	0,000145	0,000001 0,000589	0,000112	0,000001 0,000565	0,000118	0,000001 0,000567	0,000128	0,000001 0,000661
9	0,033791	0,010361	0,007032 0,014043	0,002142	0,000575 0,004259	0,000245	0,000002 0,001060	0,000128	0,000001 0,000626
10	0,250021	0,039669	0,030449 0,049936	0,007543	0,003964 0,012216	0,000584	0,000007 0,001948	0,000128	0,000001 0,000649

Tabella 5.6: Valori della S_ρ e relativi intervalli di confidenza al 95% per diversi valori di lag e di randomizzazione. Serie di periodo 10 e $p = 0.1$.

Pattern '0000000001', P=0.1, lunghezza della serie = 1000, numero di casi = 5000

Lag	Livello di randomizzazione									
	0%	25%		50%		75%		100%		
0	0,114562	0,114485	0,102947 0,125722	0,114441	0,098970 0,129344	0,114491	0,096962 0,131137	0,114404	0,096962 0,132029
1	0,005279	0,000855	0,000040 0,002489	0,000264	0,000001 0,001214	0,000143	0,000000 0,000716	0,000128	0,000000 0,000652
2	0,005279	0,000888	0,000041 0,002642	0,000259	0,000000 0,001214	0,000139	0,000000 0,000720	0,000134	0,000000 0,000686
3	0,005279	0,000884	0,000044 0,002565	0,000262	0,000000 0,001192	0,000145	0,000000 0,000768	0,000132	0,000000 0,000651
4	0,005279	0,000895	0,000047 0,002642	0,000262	0,000001 0,001161	0,000142	0,000000 0,000736	0,000133	0,000000 0,000671
5	0,005279	0,001098	0,000010 0,005064	0,000319	0,000001 0,001495	0,000164	0,000001 0,000871	0,000131	0,000000 0,000655
6	0,005279	0,000896	0,000050 0,002642	0,000258	0,000001 0,001185	0,000142	0,000001 0,000742	0,000130	0,000001 0,000687
7	0,005279	0,000871	0,000051 0,002540	0,000266	0,000001 0,001239	0,000143	0,000001 0,000733	0,000138	0,000001 0,000695
8	0,005279	0,000898	0,000041 0,002642	0,000260	0,000001 0,001194	0,000142	0,000001 0,000746	0,000135	0,000001 0,000695
9	0,005279	0,000859	0,000041 0,002565	0,000262	0,000001 0,001161	0,000138	0,000001 0,000695	0,000129	0,000001 0,000609
10	0,114562	0,021499	0,014231 0,030281	0,005009	0,001906 0,009191	0,000511	0,000004 0,001820	0,000137	0,000001 0,000713

Al fine di studiare il comportamento della S_ρ in relazione ai diversi livelli di randomizzazione applicati a serie massimamente correlate, focalizzeremo adesso la nostra attenzione sui valori della S_ρ calcolati in corrispondenza del lag 10. Il lag 10 è infatti multiplo del periodo delle serie 1, 3, 4 e 5 per cui, in assenza di randomizzazione (perfetta periodicità), ciascuna serie esprime un valore della S_ρ pari a quello di massima autocorrelazione (ad ogni 1 corrisponde sempre un 1 e, viceversa, ad ogni 0 corrisponde sempre e solo uno 0). Per quanto riguarda la serie 2 (periodo 4 e pattern "0011"), il lag 10, così come i lag 2, 6, ecc., sono multipli del semiperiodo della stessa e, visto il particolare tipo di pattern che la caratterizza, in corrispondenza di questi lag ogni elemento viene ad essere associato al suo complementare (ad ogni 1 è associato uno 0 e viceversa), determinando così una situazione di perfetta ancorchè inversa autocorrelazione (vedi precedente tabella 5.3). La stessa cosa si ha per la serie 3 (periodo 10 e pattern "0000011111") in corrispondenza dei lag 5, 15, 25, ecc. (vedi precedente tabella 5.4) ed, in generale, in corripondenza di lag coincidenti con il semiperiodo della serie nel caso di sequenze perfettamente periodiche che alternano sempre uno stesso numero di 0 e di 1.

Dai dati riportati nella tabella 5.7 appare evidente che, in media, in corrispondenza di lag multipli del periodo della serie il valore della S_ρ aumenta,

anche se in maniera via via decrescente, al tendere di p a q. Dagli stessi dati si rileva, inoltre, come la S_ρ non appaia in grado di discriminare le serie in relazione alla loro complessità algoritmica. Infatti, se in corrispondenza del lag 10 osserviamo il comportamento dell'indicatore, si potrà notare come, in assenza di randomizzazione, esso assuma valori significativamente diversi a seconda del livello di probabilità p che contraddistingue le diverse serie considerate, e ciò in contrasto con l'assunto che esse abbiano tutte la stessa complessità algoritmica.

Tabella 5.7: Valori della S_ρ e relativi intervalli di confidenza al 95% calcolati in corrispondenza del lag 10

	Livello di randomizzazione									
Serie	0%	25%		50%		75%		100%		
p=0.5 periodo 2	0,292893	0,044396	0,034977 0,054550	0,008021	0,004329 0,012402	0,000606	0,000006 0,001987	0,000128	0,000001 0,000621
p=0.5 periodo 4	0,292893	0,044706	0,035411 0,055011	0,008152	0,004339 0,012896	0,000638	0,000009 0,002003	0,000131	0,000001 0,000650
p=0.5 periodo 10	0,292893	0,044495	0,034990 0,055366	0,008006	0,004367 0,012419	0,000613	0,000007 0,002014	0,000129	0,000001 0,000621
p=0.3	0,250021	0,039669	0,030449 0,049936	0,007543	0,003964 0,012216	0,000584	0,000007 0,001948	0,000128	0,000001 0,000649
p=0.1	0,114562	0,021499	0,014231 0,030281	0,005009	0,001906 0,009191	0,000511	0,000004 0,001820	0,000137	0,000001 0,000713

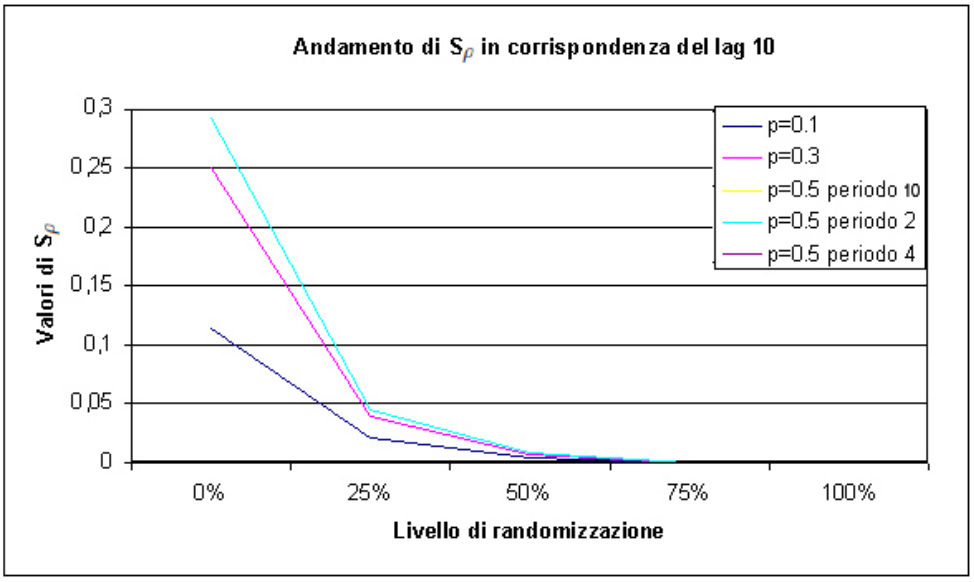

Figura 5.1: Andamento dei valori della S_ρ in corrispondenza del lag 10

Per consentire confronti rispetto ai livelli di dipendenza osservati in serie diverse, i valori della S_ρ possono essere "normalizzati" dividendoli per i rispettivi valori massimi. Dai dati normalizzati (vedi tabella 5.8) si può osservare come, per lag multipli del periodo della serie, il valore della S_ρ diminuisca in maniera molto evidente e regolare, all'incirca di un ordine di grandezza, all'aumentare del livello di randomizzazione. In particolare, in assenza di randomizzazione, vi è perfetta corrispondenza fra tutti gli elementi della serie e della sua covariata, ovvero massima autocorrelazione, fatto questo che è correttamente indicato da un valore della S_ρ normalizzata pari a 1. All'aumentare del livello di randomizzazione, la struttura della serie, in origine perfettamente periodica, si fa via via sempre più confusa, fino a perdersi del tutto in corrispondenza dei livelli di randomizzazione più elevati, dove la S_ρ rapportata al suo valore massimo tende ad assumere valori molto prossimi a 0, coerentemente con l'ipotesi di indipendenza distributiva. Si può poi

Tabella 5.8: Valori della S_ρ normalizzata e relativi intervalli di confidenza al 95% calcolati in corrispondenza del lag 10

	Livello di randomizzazione									
Serie	0%	25%		50%		75%		100%	
p=0.5 periodo 2	1,000000	0,151635	0,119486 0,186319	0,027406	0,014790 0,042372	0,002070	0,000022 0,006789	0,000439	0,000003 0,002121
p=0.5 periodo 4	1,000000	0,152698	0,120949 0,187894	0,027853	0,014826 0,044060	0,002179	0,000031 0,006844	0,000448	0,000003 0,002220
p=0.5 periodo 10	1,000000	0,151976	0,119508 0,189105	0,027353	0,014921 0,042429	0,002095	0,000024 0,006881	0,000439	0,000003 0,002121
p=0.3	1,000000	0,158718	0,121828 0,199798	0,030186	0,015864 0,048886	0,002338	0,000027 0,007799	0,000513	0,000003 0,002599
p=0.1	1,000000	0,187790	0,124303 0,264498	0,043769	0,016655 0,080313	0,004460	0,000036 0,015896	0,001201	0,000011 0,006232

osservare come la S_ρ normalizzata applicata alle diverse serie esprima, in media:

- valori che oscillano fra il 15 e il 20% del valore di dipendenza massima (massima autocorrelazione) ad un livello di randomizzazione del 25%;
- valori oscillanti fra il 2 e il 5% ad un livello di randomizzazione del 50%;
- valori molto prossimi a zero, indicanti una condizione di sostanziale indipendenza, in corrispondenza di una randomizzazione $\geq$ al 75%.

Sembrerebbe poi che, a parità di livello di randomizzazione, il valore della S_ρ normalizzata diminuisca all'approssimarsi di p a q, indicando così un minor livello di autocorrelazione in accordo al fatto che, al tendere di p a q, l'entropia aumenta. Tuttavia, i confronti fra i valori normalizzati della S_ρ relativi a serie diverse ma riferiti ad uno stesso livello di randomizzazione non appaiono significativi, con la sola eccezione del livello di randomizzazione

del 25%, dove il valore della S_ρ corrispondente alla serie con distribuzione $p = 0.1$ risulta essere significativamente diverso dagli altri ($\alpha = 0.05$).

Si può inoltre osservare come, per serie perturbate al 100%, a cui dovrebbe corrispondere una condizione di perfetta indipendenza distributiva, il valore della S_ρ, per quanto basso, non è mai nullo. Ciò è dovuto al fatto che, nei nostri esperimenti, l'indicatore, che può assumere solo valori positivi, è in realtà una media dei singoli valori della S_ρ calcolati a partire da tutti i casi simulati. Tali casi generalmente comprendono anche serie caratterizzate da probabilità di essere generate piuttosto basse ma esprimenti livelli di dipendenza relativamente elevati e che, quindi, fanno si che il valore medio della S_ρ sia, per quanto piccolo, comunque sempre maggiore di zero.

In ogni caso, il potere discriminante della S_ρ in relazione al grado di perturbazione indotta appare difficile da graduare e piuttosto limitato, in quanto circoscritto a livelli di randomizzazione non troppo elevati ($\leq 50\%, 60\%$). Infatti, gli esperimenti condotti a partire da serie in origine perfettamente periodiche e, quindi, massimamente autocorrelate, mostrano che, in genere, tali serie vengono classificate dalla S_ρ come completamente indipendenti, per $\alpha = 0.05$, già ad un livello di randomizzazione del 75%, (anche se con valori molti prossimi al valore di soglia), condizione, questa, per altro verificata, il più delle volte, anche dal test del χ^2, il quale presenta una modalità di calcolo comunque molto simile alla S_ρ.

5.2.3 Entropia congiunta

L'entropia congiunta, $H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(x,y)$, (vedi precedente capitolo 1.3) mostra un andamento opposto a quello della S_ρ. In particolare, quando la S_ρ è prossima a zero, l'entropia congiunta assume valori positivi e, com'è ovvio aspettarsi, molto vicini a quelli riferiti alla condizione di massima casualità (indipendenza distributiva) e quindi, compatibilmente con le distribuzioni marginali, di massima entropia (vedi tabelle da 5.9 a 5.13).

Tabella 5.9: Valori dell'entropia congiunta e dei relativi intervalli di confidenza al 95% per diversi valori di lag e di randomizzazione. Serie di periodo 2 e $p = 0.5$.

Pattern '01', P=0.5, lunghezza della serie = 1000, numero di casi = 5000

	Livello di randomizzazione									
Lag	0%	25%		50%		75%		100%		
0	1	0,999690	0,998473 1,000000	0,999451	0,997225 1,000000	0,999301	0,996463 1,000000	0,999294	0,996463 0,999997
1	1	1,755551	1,704416 1,804691	1,952420	1,926123 1,974348	1,994965	1,985841 1,999648	1,997877	1,991823 1,999937
2	1	1,756391	1,705604 1,804781	1,952703	1,927155 1,974457	1,995103	1,985651 1,999728	1,997863	1,991927 1,999948
3	1	1,755281	1,703814 1,803182	1,952238	1,925597 1,973962	1,995038	1,985555 1,999739	1,997875	1,991893 1,999942
4	1	1,756571	1,705186 1,804861	1,953002	1,926715 1,974362	1,995106	1,985748 1,999742	1,997888	1,992082 1,999948
5	1	1,755677	1,703524 1,803182	1,952460	1,926643 1,973798	1,994912	1,985291 1,999717	1,997875	1,991995 1,999937
6	1	1,756606	1,704919 1,803255	1,952502	1,927140 1,973681	1,995074	1,985850 1,999727	1,997860	1,991875 1,999937
7	1	1,755503	1,704674 1,802902	1,952391	1,925698 1,973779	1,994947	1,985493 1,999682	1,997855	1,992009 1,999948
8	1	1,756466	1,705579 1,804690	1,952857	1,925664 1,974457	1,995112	1,986024 1,999712	1,997871	1,992008 1,999942
9	1	1,755432	1,703370 1,803186	1,952382	1,925273 1,973897	1,994921	1,985394 1,999670	1,997881	1,991757 1,999948
10	1	1,756456	1,705487 1,804899	1,952737	1,925918 1,974340	1,995130	1,985877 1,999729	1,997838	1,991766 1,999925

Tabella 5.10: Valori dell'entropia congiunta e dei relativi intervalli di confidenza al 95% per diversi valori di lag e di randomizzazione. Serie di periodo 4 e $p = 0.5$.

Pattern '0011', P=0.5, lunghezza della serie = 1000, numero di casi = 5000

	Livello di randomizzazione									
Lag	0%	25%		50%		75%		100%		
0	1	0,999683	0,998473 1,000000	0,999452	0,997225 1,000000	0,999290	0,996257 1,000000	0,999294	0,996463 0,999997
1	2	1,999225	1,996702 1,999983	1,998505	1,993979 1,999965	1,997915	1,991724 1,999960	1,997877	1,991823 1,999937
2	1	1,756186	1,705090 1,804865	1,952146	1,926705 1,973783	1,995001	1,985613 1,999717	1,997863	1,991927 1,999948
3	2	1,999224	1,996673 1,999983	1,998495	1,994026 1,999965	1,997939	1,991531 1,999948	1,997875	1,991893 1,999942
4	1	1,757000	1,705462 1,806324	1,952823	1,926670 1,974749	1,995026	1,985391 1,999724	1,997888	1,992082 1,999948
5	2	1,999223	1,996766 1,999983	1,998467	1,993919 1,999960	1,997951	1,991460 1,999954	1,997875	1,991995 1,999937
6	1	1,755943	1,703373 1,804767	1,952220	1,926334 1,973622	1,994971	1,985458 1,999688	1,997860	1,991875 1,999937
7	2	1,999231	1,996713 1,999983	1,998518	1,993916 1,999965	1,997950	1,991680 1,999960	1,997855	1,992009 1,999948
8	1	1,756839	1,703482 1,804623	1,952861	1,926646 1,974275	1,995122	1,985755 1,999716	1,997871	1,992008 1,999942
9	2	1,999223	1,996709 1,999983	1,998509	1,994068 1,999960	1,997951	1,991511 1,999954	1,997881	1,991757 1,999948
10	1	1,755653	1,701228 1,803258	1,952189	1,925698 1,973702	1,994900	1,985298 1,999682	1,997838	1,991766 1,999925

Tabella 5.11: Valori dell'entropia congiunta e dei relativi intervalli di confidenza al 95% per diversi valori di lag e di randomizzazione. Serie di periodo 10 e $p = 0.5$.

Pattern '0000011111', P=0.5, lunghezza della serie = 1000, numero di casi = 5000

Lag	Livello di randomizzazione									
	0%	25%		50%		75%		100%		
0	1	0,999686	0,998337 1,000000	0,999463	0,997225 1,000000	0,999307	0,996463 1,000000	0,999294	0,996463 0,999997
1	1,721928	1,915356	1,889345 1,939308	1,982024	1,966769 1,993258	1,996878	1,989484 1,999902	1,997877	1,991823 1,999937
2	1,970951	1,990071	1,983143 1,995742	1,996605	1,990248 1,999811	1,997852	1,991770 1,999948	1,997863	1,991927 1,999948
3	1,970951	1,989908	1,982851 1,995675	1,996569	1,990206 1,999792	1,997816	1,991706 1,999960	1,997875	1,991893 1,999942
4	1,721928	1,914879	1,888738 1,938787	1,981649	1,966198 1,993474	1,996796	1,989343 1,999885	1,997888	1,992082 1,999948
5	1	1,755650	1,701978 1,804028	1,952210	1,925292 1,974213	1,994934	1,985388 1,999680	1,997875	1,991995 1,999937
6	1,721928	1,914892	1,888349 1,939280	1,981700	1,966440 1,992982	1,996782	1,989309 1,999901	1,997860	1,991875 1,999937
7	1,970951	1,989933	1,982843 1,995566	1,996502	1,989999 1,999763	1,997806	1,991722 1,999937	1,997855	1,992009 1,999948
8	1,970951	1,990014	1,982963 1,995593	1,996642	1,990048 1,999821	1,997873	1,991758 1,999948	1,997871	1,992008 1,999942
9	1,721928	1,915542	1,888643 1,940028	1,981969	1,966802 1,993447	1,996880	1,989646 1,999913	1,997881	1,991757 1,999948
10	1	1,757049	1,703404 1,805789	1,952844	1,927077 1,974034	1,995078	1,985784 1,999706	1,997838	1,991766 1,999925

Tabella 5.12: Valori dell'entropia congiunta e dei relativi intervalli di confidenza al 95% per diversi valori di lag e di randomizzazione. Serie di periodo 10 e $p = 0.3$.

Pattern '0000000111', P=0.3, lunghezza della serie = 1000, numero di casi = 5000

Lag	Livello di randomizzazione									
	0%	25%		50%		75%		100%		
0	0,881291	0,881209	0,858162 0,902193	0,880562	0,849943 0,909736	0,880565	0,845737 0,911832	0,880732	0,842896 0,912870
1	1,570951	1,702095	1,652983 1,749079	1,748563	1,685421 1,806314	1,759698	1,690714 1,822925	1,760719	1,685745 1,825306
2	1,760964	1,761582	1,714660 1,804333	1,760522	1,698412 1,818390	1,760427	1,690177 1,823664	1,760718	1,685781 1,825704
3	1,570951	1,715667	1,666847 1,762221	1,751950	1,690023 1,810142	1,759872	1,691094 1,823130	1,760737	1,685671 1,825424
4	1,570951	1,715209	1,663369 1,763266	1,751754	1,687961 1,809929	1,759838	1,690902 1,823123	1,760735	1,685783 1,825556
5	1,570951	1,714801	1,661249 1,764521	1,751558	1,688184 1,811235	1,759783	1,689960 1,822938	1,760763	1,685575 1,825549
6	1,570951	1,715125	1,665128 1,761938	1,751761	1,688783 1,810060	1,759912	1,690929 1,823431	1,760753	1,685930 1,825447
7	1,570951	1,715774	1,666960 1,763211	1,751861	1,688774 1,809342	1,759919	1,691353 1,822726	1,760751	1,685559 1,824997
8	1,760964	1,761576	1,714947 1,804334	1,760447	1,698406 1,817931	1,760480	1,691102 1,823573	1,760766	1,685947 1,824932
9	1,570951	1,702227	1,653253 1,749312	1,748563	1,685271 1,806867	1,759737	1,690708 1,822631	1,760764	1,685786 1,825091
10	0,881291	1,540124	1,477405 1,601279	1,717079	1,650498 1,776236	1,757715	1,688664 1,821240	1,760755	1,685586 1,824881

Tabella 5.13: Valori dell'entropia congiunta e dei relativi intervalli di confidenza al 95% per diversi valori di lag e di randomizzazione. Serie di periodo 10 e $p = 0.1$.

Pattern '0000000001', P=0.1, lunghezza della serie = 1000, numero di casi = 5000

Lag	Livello di randomizzazione									
	0%	25%		50%		75%		100%		
0	0,468996	0,468831	0,429759 0,508899	0,468234	0,416119 0,517753	0,468471	0,409187 0,523584	0,468245	0,405693 0,526480
1	0,921928	0,933426	0,851912 1,010699	0,935138	0,832238 1,035491	0,936127	0,818177 1,047111	0,935747	0,811361 1,052997
2	0,921928	0,933473	0,852714 1,011494	0,935172	0,832238 1,035485	0,936107	0,818172 1,047111	0,935687	0,811361 1,053817
3	0,921928	0,933399	0,853682 1,011339	0,935150	0,832133 1,035491	0,936089	0,818306 1,047113	0,935731	0,811361 1,055101
4	0,921928	0,933416	0,851906 1,012156	0,935143	0,832238 1,035356	0,936088	0,817965 1,047169	0,935749	0,811768 1,052979
5	0,921928	0,932893	0,851912 1,011347	0,934806	0,833127 1,035356	0,935979	0,818306 1,047321	0,935710	0,814224 1,052957
6	0,921928	0,933398	0,851913 1,011343	0,935084	0,833127 1,035477	0,936072	0,818927 1,047113	0,935715	0,811768 1,052957
7	0,921928	0,933442	0,853069 1,011018	0,935110	0,832251 1,035070	0,936096	0,820642 1,047112	0,935686	0,813816 1,053008
8	0,921928	0,933378	0,853054 1,011490	0,935207	0,833104 1,035077	0,936093	0,818306 1,047155	0,935706	0,812802 1,053900
9	0,921928	0,933404	0,853682 1,011375	0,935160	0,832238 1,035477	0,936081	0,818346 1,047093	0,935722	0,813856 1,054541
10	0,468996	0,797141	0,717871 0,877788	0,903515	0,803034 1,005218	0,933530	0,815614 1,045399	0,935705	0,814209 1,055101

Nel caso di serie perfettamente periodiche (serie non perturbate), il valore di massima autocorrelazione si raggiunge, anche in questo contesto, in corrispondenza di un lag multiplo del periodo della serie e coincide con il valore minimo ammissibile per l'entropia congiunta. Tale valore, che si può ottenere confrontando la serie con se stessa (lag 0), è pari al valore dell'entropia di Shannon calcolata sulla distribuzione marginale della serie. Infatti, data una serie binaria con distribuzione $P(X = 1) = p$ e $P(X = 0) = q$, nel caso di massima autocorrelazione (ad un valore 1 nella serie corrisponde sempre e solo un valore 1 nella sua covariata e viceversa) si avrà:

$$\begin{aligned} p(x_i, x_{i+lag}) &= 0 \text{ se } x_i \neq x_{i+lag} \\ p(x_i, x_{i+lag}) &= p \text{ se } x_i = x_{i+lag} = 1 \\ p(x_i, x_{i+lag}) &= q \text{ se } x_i = x_{i+lag} = 0 \end{aligned}$$

così che il calcolo dell'entropia congiunta si riduce, appunto, al calcolo dell'entropia di Shannon sulla distribuzione della serie originale.

Per definizione, invece, il valore massimo dell'entropia congiunta si raggiunge in corrispondenza della condizione di completa casualità (livello di randomizzazione pari al 100%) e, poiché la completa casualità implica l'indipendenza in distribuzione, tale massimo coincide con il valore dell'entropia calcolata sulla distribuzione di probabilità congiunta data la condizione di indipendenza $p(x, y) = p(x)p(y)$.

Sostituendo l'espressione $p(x)p(y)$ a $p(x,y)$ nella formula dell'entropia congiunta si ottiene:

$$H(X,Y)_{max} = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x)p(y)\log_2 p(x)p(y)$$

ovvero, poiché $p(x)$ e $p(y)$ rappresentano la stessa distribuzione di probabilità:

$$\begin{aligned}
H(X,Y)_{max} &= -\sum_{x_i\in\mathcal{X}}\sum_{x_j\in\mathcal{X}} p(x_i)p(x_j)\log_2 p(x_i)p(x_j) \\
&= -\sum_{x_i\in\mathcal{X}}\sum_{x_j\in\mathcal{X}} p(x_i)p(x_j)\log_2 p(x_i) - \sum_{x_i\in\mathcal{X}}\sum_{x_j\in\mathcal{X}} p(x_i)p(x_j)\log_2 p(x_j) \\
&= -\sum_{x_j\in\mathcal{X}} p(x_j)\sum_{x_i\in\mathcal{X}} p(x_i)\log_2 p(x_i) - \sum_{x_i\in\mathcal{X}} p(x_i)\sum_{x_j\in\mathcal{X}} p(x_j)\log_2 p(x_j) \\
&= -\sum_{x_i\in\mathcal{X}} p(x_i)\log_2 p(x_i) - \sum_{x_j\in\mathcal{X}} p(x_j)\log_2 p(x_j) \\
&= -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x) - \sum_{x\in\mathcal{X}} p(x)\log_2 p(x) = 2H(X)
\end{aligned}$$

Per cui il valore massimo dell'entropia congiunta calcolata su serie aventi la medesima distribuzione di probabilità, massimo che si raggiunge nel caso di indipendenza distributiva, è pari a due volte l'entropia di Shannon calcolata sulla distribuzione marginale.

L'entropia congiunta non pare poi in grado di discriminare le serie in relazione alla loro complessità algoritmica. Infatti, nonostante che, in assenza di randomizzazione, esse siano caratterizzate dallo stesso livello di complessità, l'entropia congiunta assume comunque valori significativamente differenti a seconda del livello di probabilità p che le contraddistingue.

Si può inoltre osservare che, per ogni serie considerata, l'entropia congiunta si stabilizza su valori fra loro praticamente uguali e molto prossimi a $2H(X)$ già a partire da un livello di randomizzazione del 50%, attestando con ciò una ipotetica condizione di indipendenza la quale, almeno per quanto concerne il livello di randomizzazione del 50%, è stata decisamente smentita dai risultati ottenuti con il test del χ^2. Da questo punto di vista, l'entropia congiunta appare quindi meno "sensibile" e, di conseguenza, meno affidabile rispetto alla S_ρ.

Tabella 5.14: Andamento dei valori dell'entropia congiunta e dei relativi intervalli di confidenza al 95% in corrispondenza del lag 10

Serie	0%	25%		50%		75%		100%	
p=0.5 periodo 2	1	1,756456	1,705487 1,804899	1,952737	1,925918 1,974340	1,995130	1,985877 1,999729	1,997838	1,991766 1,999925
p=0.5 periodo 4	1	1,755653	1,701228 1,803258	1,952189	1,925698 1,973702	1,994900	1,985298 1,999682	1,997838	1,991766 1,999925
p=0.5 periodo 10	1	1,757049	1,703404 1,805789	1,952844	1,927077 1,974034	1,995078	1,985784 1,999706	1,997838	1,991766 1,999925
p=0.3	0,881291	1,540124	1,477405 1,601279	1,717079	1,650498 1,776236	1,757715	1,688664 1,821240	1,760755	1,685586 1,824881
p=0.1	0,468996	0,797141	0,717871 0,877788	0,903515	0,803034 1,005218	0,933530	0,815614 1,045399	0,935705	0,814209 1,055101

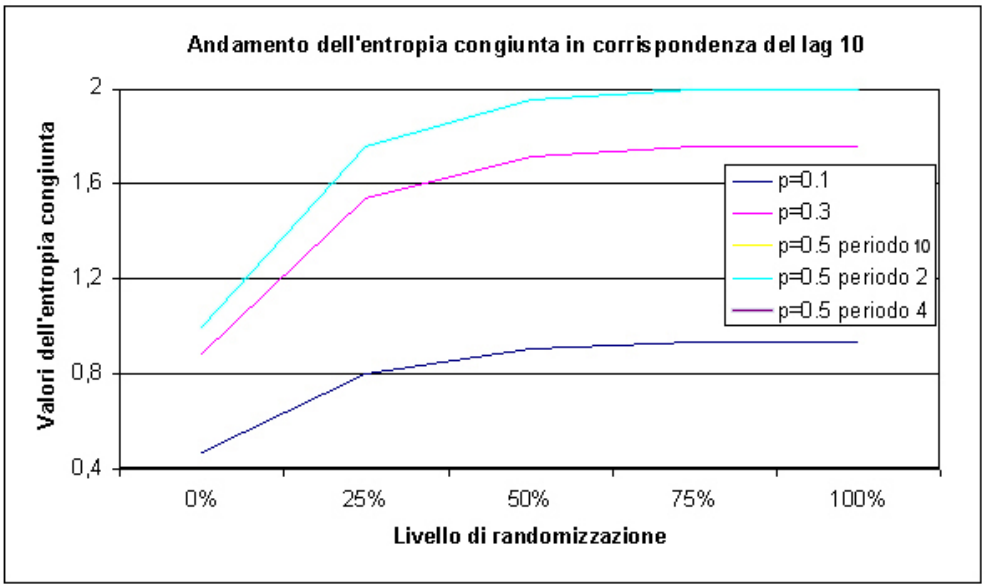

Figura 5.2: Andamento dei valori dell'entropia congiunta in corrispondenza del lag 10

Anche in questo caso, per poter confrontare l'andamento dell'entropia congiunta in serie caratterizzate da diverse distribuzioni di probabilità o, per una stessa serie, rispetto ad altri indicatori quali la S_ρ, è opportuno normalizzare l'indicatore in modo che esso vari nell'intervallo $[0, 1]$, ovvero:

$$\begin{aligned} H(X,Y)_{norm} &= 1 - \frac{H(X,Y) - H(X,Y)_{min}}{H(X,Y)_{max} - H(X,Y)_{min}} \\ &= \frac{H(X,Y)_{max} - H(X,Y)}{H(X,Y)_{max} - H(X,Y)_{min}} = \frac{2H(X) - H(X,Y)}{2H(X) - H(X)} \\ &= \frac{2H(X) - H(X,Y)}{H(X)} = 2 - H(X,Y)/H(X) \end{aligned}$$

Nella tabella seguente (tabella 5.15) sono riportati i valori normalizzati dell'entropia congiunta in corrispondenza del lag 10. Si può quindi osservare

Serie	0%	25%		50%		75%		100%	
p=0.5 periodo 2	1	0,243544	0,195101 0,294513	0,047263	0,025660 0,074082	0,004870	0,000271 0,014123	0,002162	0,000075 0,008234
p=0.5 periodo 4	1	0,244347	0,196742 0,298772	0,047811	0,026298 0,074302	0,005100	0,000318 0,014702	0,002162	0,000075 0,008234
p=0.5 periodo 10	1	0,242951	0,194211 0,296596	0,047156	0,025966 0,072923	0,004922	0,000294 0,014216	0,002162	0,000075 0,008234
p=0.3	1	0,252423	0,183031 0,323590	0,051632	0,000000 0,127182	0,005522	0,000000 0,083875	0,002073	0,000000 0,087368
p=0.1	1	0,300322	0,128366 0,469343	0,073510	0,000000 0,287758	0,009512	0,000000 0,260935	0,004875	0,000000 0,263931

Tabella 5.15: Andamento dei valori dell'entropia congiunta normalizzata e dei relativi intervalli di confidenza al 95% in corrispondenza del lag 10

come, in media, l'entropia congiunta normalizzata applicata alle diverse serie esprima:

- valori che oscillano fra il 25 e il 30% del valore di dipendenza massima (massima autocorrelazione) ad un livello di randomizzazione del 25%;
- valori oscillanti fra il 5 e il 7% ad un livello di randomizzazione del 50%;
- valori abbastanza prossimi a zero, indicanti una condizione che tende ad una sostanziale indipendenza, in corrispondenza di una randomizzazione $\geq$ al 75%.

L'entropia congiunta normalizzata presenta tuttavia valori di varianza talmente elevati da sconsigliarne l'uso nella pratica. L'indicatore, sia nella sua forma normalizzata che non, non è infatti in grado di discriminare livelli di perturbazione superiori al 50% nonostante che, fra i valori relativi ai diversi livelli di randomizzazione, vi possano essere differenze anche di un ordine di grandezza.

5.2.4 Mutua informazione

La mutua informazione fra due variabili aleatorie X e Y, definita (vedi precedente capitolo 1.5) come $I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$, misura la riduzione di "incertezza" di una variabile dovuta alla conoscenza dell'altra. Se le due variabili sono indipendentemente distribuite, la mutua informazione vale zero, così come vale zero se una delle due variabili ha entropia nulla.

La mutua informazione si annulla infatti, al pari della S_ρ, in coincidenza della condizione di indipendenza ed è massima, pari all'entropia di Shannon calcolata sulla distribuzione marginale, nel caso di serie uguali (massima correlazione). Il suo comportamento appare inoltre, a meno di un fattore di scala, quasi completamente sovrapponibile a quello della S_ρ (vedi tabelle da 5.16 a 5.20).

Tabella 5.16: Valori della mutua informazione e dei relativi intervalli di confidenza al 95% per diversi valori di lag e di randomizzazione. Serie di periodo 2 e $p = 0.5$.

Pattern '01', P=0.5, lunghezza della serie = 1000, numero di casi = 5000

Lag	Livello di randomizzazione									
	0%	25%		50%		75%		100%		
0	1	0,999690	0,998473 1,000000	0,999451	0,997225 1,000000	0,999301	0,996463 1,000000	0,999294	0,996463 0,999997
1	1	0,243830	0,195125 0,294821	0,046481	0,024782 0,072691	0,003634	0,000030 0,011499	0,000711	0,000001 0,003496
2	1	0,242990	0,194560 0,294144	0,046199	0,024558 0,071414	0,003496	0,000023 0,011593	0,000725	0,000001 0,003704
3	1	0,244101	0,195721 0,294825	0,046665	0,024893 0,072634	0,003563	0,000029 0,011739	0,000715	0,000001 0,003517
4	1	0,242811	0,194613 0,294355	0,045901	0,024553 0,072037	0,003495	0,000020 0,011476	0,000701	0,000001 0,003595
5	1	0,243704	0,196744 0,295468	0,046442	0,025110 0,072357	0,003690	0,000024 0,011558	0,000715	0,000001 0,003718
6	1	0,242775	0,195370 0,294318	0,046401	0,025414 0,071395	0,003530	0,000026 0,011187	0,000732	0,000001 0,003658
7	1	0,243877	0,196296 0,294698	0,046511	0,025501 0,072531	0,003656	0,000034 0,011659	0,000736	0,000001 0,003629
8	1	0,242915	0,194972 0,294195	0,046044	0,024490 0,073171	0,003490	0,000024 0,011456	0,000719	0,000001 0,003577
9	1	0,243948	0,196739 0,295839	0,046519	0,025247 0,073152	0,003682	0,000030 0,011598	0,000709	0,000001 0,003465
10	1	0,242924	0,193512 0,294338	0,046163	0,024984 0,072329	0,003472	0,000024 0,011263	0,000751	0,000001 0,003788

La maggiore differenza rispetto alla S_ρ che è emersa dagli esperimenti effettuati riguarda il comportamento rispetto alla serie di periodo 10, pattern "0000000111" e $p = 0.3$. In particolare, come si può osservare dalla tabella 5.21, in assenza di randomizzazione (ma il discorso è valido anche per livelli di randomizzazione diversi da zero) è possibile individuare quattro diversi tipi di distribuzione congiunta (in tabella 5.21 sono indicati con colori differenti). La S_ρ riesce correttamente ad individuare e graduare tali diversi

Tabella 5.17: Valori della mutua informazione e dei relativi intervalli di confidenza al 95% per diversi valori di lag e di randomizzazione. Serie di periodo 4 e $p = 0.5$.

Pattern '0011', P=0.5, lunghezza della serie = 1000, numero di casi = 5000

	Livello di randomizzazione									
m	0%	25%		50%		75%		100%		
0	1	0,999683	0,998473 1,000000	0,999452	0,997225 1,000000	0,999290	0,996257 1,000000	0,999294	0,996463 0,999997
1	0	0,000142	0,000000 0,000734	0,000397	0,000000 0,001959	0,000665	0,000001 0,003233	0,000711	0,000001 0,003496
2	1	0,243180	0,194045 0,294697	0,046757	0,025118 0,072346	0,003579	0,000025 0,011557	0,000725	0,000001 0,003704
3	0	0,000143	0,000000 0,000721	0,000408	0,000000 0,002108	0,000640	0,000001 0,003208	0,000715	0,000001 0,003517
4	1	0,242366	0,193257 0,294257	0,046080	0,024203 0,072161	0,003553	0,000025 0,011365	0,000701	0,000001 0,003595
5	0	0,000143	0,000000 0,000724	0,000435	0,000000 0,002260	0,000628	0,000001 0,003113	0,000715	0,000001 0,003718
6	1	0,243423	0,193843 0,296460	0,046683	0,025419 0,072394	0,003606	0,000020 0,011729	0,000732	0,000001 0,003658
7	0	0,000135	0,000000 0,000688	0,000386	0,000000 0,001949	0,000628	0,000000 0,003125	0,000736	0,000001 0,003629
8	1	0,242527	0,194624 0,296157	0,046043	0,024824 0,072337	0,003456	0,000025 0,010996	0,000719	0,000001 0,003577
9	0	0,000143	0,000000 0,000738	0,000396	0,000000 0,001950	0,000627	0,000001 0,003130	0,000709	0,000001 0,003465
10	1	0,243713	0,195412 0,298147	0,046717	0,025642 0,072755	0,003677	0,000026 0,011905	0,000751	0,000001 0,003788

Tabella 5.18: Valori della mutua informazione e dei relativi intervalli di confidenza al 95% per diversi valori di lag e di randomizzazione. Serie di periodo 10 e $p = 0.5$.

Pattern '0000011111', P=0.5, lunghezza della serie = 1000, numero di casi = 5000

	Livello di randomizzazione									
Lag	0%	25%		50%		75%		100%		
0	1	0,999686	0,998337 1,000000	0,999463	0,997225 1,000000	0,999307	0,996463 1,000000	0,999294	0,996463 0,999997
1	0,278072	0,084015	0,059820 0,109142	0,016903	0,006044 0,031436	0,001737	0,000003 0,007200	0,000711	0,000001 0,003496
2	0,029049	0,009301	0,003918 0,015856	0,002322	0,000012 0,007507	0,000763	0,000001 0,003742	0,000725	0,000001 0,003704
3	0,029049	0,009464	0,003831 0,016294	0,002359	0,000017 0,007656	0,000800	0,000001 0,003996	0,000715	0,000001 0,003517
4	0,278072	0,084494	0,060725 0,110366	0,017278	0,005867 0,032293	0,001818	0,000003 0,007526	0,000701	0,000001 0,003595
5	1	0,243723	0,195331 0,296969	0,046718	0,025110 0,073282	0,003680	0,000029 0,011893	0,000715	0,000001 0,003718
6	0,278072	0,084481	0,060087 0,111502	0,017228	0,006129 0,032188	0,001832	0,000003 0,007460	0,000732	0,000001 0,003658
7	0,029049	0,009439	0,003987 0,016323	0,002427	0,000026 0,007677	0,000808	0,000001 0,003995	0,000736	0,000001 0,003629
8	0,029049	0,009358	0,003949 0,016106	0,002287	0,000012 0,007624	0,000741	0,000001 0,003723	0,000719	0,000001 0,003577
9	0,278072	0,083830	0,059802 0,110145	0,016958	0,005841 0,031703	0,001735	0,000003 0,007151	0,000709	0,000001 0,003465
10	1	0,242323	0,193508 0,296377	0,046082	0,025020 0,071585	0,003538	0,000026 0,011828	0,000751	0,000001 0,003788

Tabella 5.19: Valori della mutua informazione e dei relativi intervalli di confidenza al 95% per diversi valori di lag e di randomizzazione. Serie di periodo 10 e $p = 0.3$.

Pattern '0000000111', P=0.3, lunghezza della serie = 1000, numero di casi = 5000

	Livello di randomizzazione									
Lag	**0%**	**25%**		**50%**		**75%**		**100%**		
0	0,881291	0,881209	0,858162 0,902193	0,880562	0,849943 0,909736	0,880565	0,845737 0,911832	0,880732	0,842896 0,912870
1	0,191631	0,060321	0,041108 0,081788	0,012564	0,003381 0,025316	0,001430	0,000002 0,006348	0,000733	0,000001 0,003467
2	0,001618	0,000831	0,000001 0,003400	0,000616	0,000001 0,003033	0,000699	0,000001 0,003459	0,000719	0,000001 0,003639
3	0,191631	0,046751	0,031375 0,063682	0,009174	0,002164 0,019236	0,001261	0,000001 0,005731	0,000721	0,000001 0,003574
4	0,191631	0,047208	0,028889 0,069180	0,009356	0,001766 0,020801	0,001287	0,000002 0,005973	0,000729	0,000001 0,003553
5	0,191631	0,047603	0,025197 0,075179	0,009545	0,001222 0,022872	0,001350	0,000001 0,006075	0,000724	0,000001 0,003525
6	0,191631	0,047280	0,029601 0,068315	0,009338	0,001862 0,020980	0,001219	0,000002 0,005400	0,000726	0,000001 0,003761
7	0,191631	0,046623	0,031513 0,063866	0,009224	0,002198 0,019466	0,001225	0,000001 0,005400	0,000726	0,000001 0,003686
8	0,001618	0,000813	0,000001 0,003415	0,000640	0,000001 0,003071	0,000671	0,000001 0,003453	0,000720	0,000001 0,003635
9	0,191631	0,060171	0,040559 0,081739	0,012530	0,003275 0,025230	0,001417	0,000002 0,006054	0,000723	0,000001 0,003701
10	0,881291	0,222290	0,173780 0,275483	0,044008	0,022871 0,069480	0,003463	0,000023 0,011185	0,000728	0,000001 0,003511

Tabella 5.20: Valori della mutua informazione e dei relativi intervalli di confidenza al 95% per diversi valori di lag e di randomizzazione. Serie di periodo 10 e $p = 0.1$.

Pattern '0000000001', P=0.1, lunghezza della serie = 1000, numero di casi = 5000

	Livello di randomizzazione									
Lag	**0%**	**25%**		**50%**		**75%**		**100%**		
0	0,468996	0,468831	0,429759 0,508899	0,468234	0,416119 0,517753	0,468471	0,409187 0,523584	0,468245	0,405693 0,526480
1	0,016063	0,004222	0,000233 0,011301	0,001334	0,000003 0,005941	0,000790	0,000001 0,003925	0,000716	0,000001 0,003667
2	0,016063	0,004181	0,000287 0,010995	0,001333	0,000001 0,005848	0,000773	0,000001 0,003771	0,000763	0,000001 0,003839
3	0,016063	0,004241	0,000299 0,011615	0,001358	0,000001 0,005941	0,000768	0,000001 0,003771	0,000733	0,000001 0,003772
4	0,016063	0,004204	0,000228 0,010992	0,001362	0,000003 0,005903	0,000764	0,000001 0,003924	0,000727	0,000001 0,003628
5	0,016063	0,004722	0,000060 0,015391	0,001666	0,000001 0,007504	0,000862	0,000001 0,004391	0,000745	0,000001 0,003879
6	0,016063	0,004210	0,000287 0,010992	0,001382	0,000002 0,005791	0,000783	0,000001 0,003949	0,000725	0,000001 0,003603
7	0,016063	0,004174	0,000287 0,010992	0,001379	0,000001 0,006043	0,000753	0,000001 0,003952	0,000739	0,000001 0,003876
8	0,016063	0,004231	0,000287 0,011148	0,001306	0,000003 0,005668	0,000751	0,000001 0,003772	0,000734	0,000001 0,003603
9	0,016063	0,004203	0,000261 0,011409	0,001343	0,000002 0,006023	0,000773	0,000001 0,003760	0,000717	0,000001 0,003531
10	0,468996	0,140507	0,095202 0,191782	0,032994	0,012193 0,059273	0,003280	0,000009 0,011446	0,000750	0,000001 0,003840

livelli mediante quattro distinti valori, mentre la mutua informazione ne riesce a distinguere solo tre. La mutua informazione (come d'altronde anche l'entropia congiunta) non riesce infatti a discriminare la situazione di cui ai lag 1 e 9 da quella dei lag da 3 a 7 mostrando quindi, in questo contesto, una minore sensibilità rispetto alla $S\rho$. Dalla tabella 5.22, che sintetizza il com-

Tabella 5.21: Distribuzione congiunta della serie "0000000111" e $p = 0.3$ in funzione del lag ed in assenza di randomizzazione.

Pattern '0000000111', P=0.3, lunghezza della serie = 1000, numero di casi = 5000

Lag	0	1	2	3	4	5	6	7	8	9	10
Distribuzione congiunta	00	01	01	01	00	00	00	00	00	00	00
	00	00	01	01	01	00	00	00	00	00	00
	00	00	00	01	01	01	00	00	00	00	00
	00	00	00	00	01	01	01	00	00	00	00
	00	00	00	00	00	01	01	01	00	00	00
	00	00	00	00	00	00	01	01	01	00	00
	00	00	00	00	00	00	00	01	01	01	00
	11	10	10	10	10	10	10	10	11	11	11
	11	11	10	10	10	10	10	10	10	11	11
	11	11	11	10	10	10	10	10	10	10	11
Sρ	0,250021	0,033791	0,000279	0,055285	0,055285	0,055285	0,055285	0,055285	0,000279	0,033791	0,250021
Mutua informazione	0,881291	0,191631	0,001618	0,191631	0,191631	0,191631	0,191631	0,191631	0,001618	0,191631	0,881291

portamento della mutua informazione relativamente alle cinque diverse serie analizzate (lag 10), si può facilmente osservare come anche questo indicatore non appaia, in genere, in grado di discriminare le serie in relazione al loro grado di complessita algoritmica. Infatti, pur in assenza di randomizzazione, anche la mutua informazione assume valori differenti a seconda del livello di probabilità p che caratterizza le diverse serie.

Infine, ancora come nel caso della S_ρ, gli esperimenti condotti a partire da serie perfettamente periodiche mostrano che, in genere, la mutua informazione non è in grado di distingure fra serie perturbate al 75% e al 100%.

Tabella 5.22: Valori della mutua informazione e dei relativi intervalli di confidenza al 95% in corrispondenza del lag 10

Serie	0%	25%		50%		75%		100%	
p=0.5 periodo 2	1	0,242924	0,193512 0,294338	0,046163	0,024984 0,072329	0,003472	0,000024 0,011263	0,000751	0,000001 0,003788
p=0.5 periodo 4	1	0,243713	0,195412 0,298147	0,046717	0,025642 0,072755	0,003677	0,000026 0,011905	0,000751	0,000001 0,003788
p=0.5 periodo 10	1	0,242323	0,193508 0,296377	0,046082	0,025020 0,071585	0,003538	0,000026 0,011828	0,000751	0,000001 0,003788
p=0.3	0,881291	0,222290	0,173780 0,275483	0,044008	0,022871 0,069480	0,003463	0,000023 0,011185	0,000728	0,000001 0,003511
P=0.1	0,468996	0,140507	0,095202 0,191782	0,032994	0,012193 0,059273	0,003280	0,000009 0,011446	0,000750	0,000001 0,003840

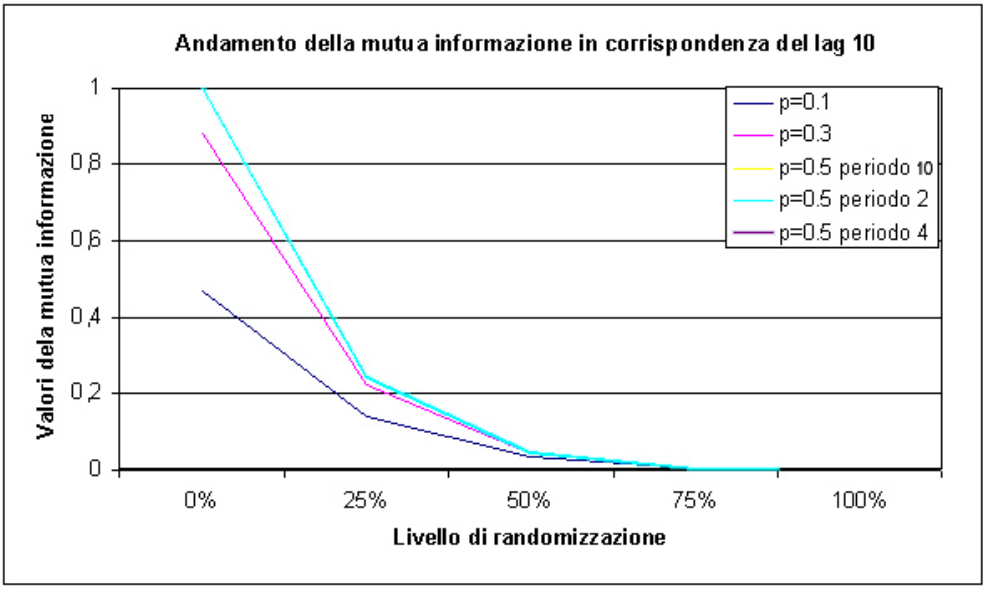

Figura 5.3: Andamento dei valori della mutua informazione in corrispondenza del lag 10

Può essere poi opportuno normalizzare la mutua informazione in modo da consentirne il confronto con altri indicatori e, mediante la comparazione di serie caratterizzate da diversi valori della probabilità p, studiarne le variazioni al variare del livello di randomizzazione.

In particolare, come già accennato all'inizio del paragrafo, nel caso di serie aventi uguale distribuzione di probabilità la mutua informazione assume il suo valore massimo quando le serie sono uguali (auto informazione) e tale valore massimo è pari all'entropia di Shannon calcolata sulla distribuzione marginale. In questo caso la mutua informazione normalizzata assumerà la

forma seguente:

$$\begin{aligned} I(X;Y)_{norm} &= I(X;Y)/I(X;Y)_{max} \\ &= I(X;Y)/I(X;X) \\ &= I(X;Y)/H(X) \end{aligned}$$

Dalla 1.3 del precedente capitolo 1.5 si ha che, in generale:

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

che si può scrivere, visto che X ed Y hanno medesima distribuzione, come:

$$I(X;Y) = 2H(X) - H(X,Y)$$

per cui, dividendo entrambi i membri per $H(X)$, la mutua informazione normalizzata diventa:

$$I(X;Y)_{norm} = 2 - H(X,Y)/H(X)$$

che altro non è che la forma normalizzata dell'entropia congiunta.

Nella tabella 5.23, che riporta i valori della mutua informazione normalizzata in corrispondenza del lag 10, si può osservare come, in effetti, i valori della mutua informazione qui riportati siano sostanzialmente uguali ai corrispondenti valori dell'entropia congiunta normalizzata (vedi precedente tabella 5.15). Tuttavia, la mutua informazione presenta intervalli di confidenza decisamente più ridotti e quindi minore varianza, motivo questo che la rende sicuramente preferibile, quale indicatore di dipendenza, all'entropia congiunta. Per questa ragione, nel proseguo della presente tesi, l'entropia congiunta non verrà più calcolata.

Tabella 5.23: Valori della mutua informazione normalizzata e dei relativi intervalli di confidenza al 95% in corrispondenza del lag 10

Serie	0%	25%		50%		75%		100%	
p=0.5 periodo 2	1	0,242924	0,193512 0,294338	0,046163	0,024984 0,072329	0,003472	0,000024 0,011263	0,000751	0,000001 0,003788
p=0.5 periodo 4	1	0,243713	0,195412 0,298147	0,046717	0,025642 0,072755	0,003677	0,000026 0,011905	0,000751	0,000001 0,003788
p=0.5 periodo 10	1	0,242323	0,193508 0,296377	0,046082	0,025020 0,071585	0,003538	0,000026 0,011828	0,000751	0,000001 0,003788
p=0.3	1	0,252256	0,197206 0,312619	0,049977	0,025973 0,078904	0,003933	0,000026 0,012702	0,000826	0,000001 0,003986
p=0.1	1	0,299697	0,203061 0,409065	0,070465	0,026040 0,126589	0,007002	0,000020 0,024433	0,001602	0,000002 0,008200

5.2.5 Entropia relativa o Distanza di Kullback Leibner

L'entropia relativa $D(h||g) = \sum_{x\in\mathcal{X}} h(x) \log \frac{h(x)}{g(x)}$ misura la distanza fra due distribuzioni di probabilità h e g (vedi precedente capitolo 3.2). L'entropia

relativa è sempre non negativa ed assume valore zero se e solo se $h(x) = g(x)$ per ogni $x \in \mathcal{X}$.

Nei casi studiati, le distanze di Kullback Leibner assumono sempre valori pari o molto prossimi a zero. Questo perchè ogni serie viene confrontata con se stessa a meno di uno slittamento di un numero di elementi pari al lag scelto. Ovviamente, tale slittamento, considerando un lag variabile da 1 a 10, può comportare l'introduzione di un numero massimo di "nuovi" elementi pari al lag stesso (quindi al massimo 10) e questo, su un totale di 1000 elementi, non comporta variazioni significative nelle distribuzioni marginali. Da qui la sostanziale uguaglianza, in distribuzione, delle serie confrontate. Tale misura non appare quindi in grado di cogliere il livello di dipendenza fra gli elementi di una stessa serie né il loro grado di disordine, né, tanto meno, il loro grado di complessità.

Inoltre, poiché in questa seconda parte della tesi le distribuzioni marginali delle serie poste confronto saranno sostanzialmente sempre uguali e, di conseguenza, le distanze di Kullback Leibler assumeranno sempre valori pari o molto prossimi a zero, nel proseguo dell'analisi dell'autocorrelazione anche le distanze di Kullback Leibler non verranno più calcolate.

5.2.6 Approximate Entropy

A differenza degli indicatori precedenti, la *ApEn* non è calcolata a partire dalla distribuzione di probabilità della serie, ma si considera l'ordine con cui gli elementi si susseguono nella serie stessa (vedi precedente capitolo 3.4). Essa assume valore zero nel caso di entropia nulla, ovvero di perfetta regolarità, e raggiunge il suo massimo in corrispondenza della massima entropia (massima irregolarità o massimo disordine).

In realtà, però, la condizione di perfetta regolarità viene individuata solo in presenza di valori del parametro m (la dimensione dei pattern di confronto) pari o comunque prossimi alla lunghezza del periodo della serie. Purtroppo però, all'aumentare di m, diminuisce in genere il potere informativo dell'indicatore il quale, in linea di massima, tende ad assumere valori via via sempre più bassi a prescindere dall'effettivo grado di regolarità incontrato.

Dall'analisi dei valori della *ApEn* calcolati sulle serie completamente randomizzate (condizione di massima entropia), si può osservare come questi valori si approssimino al massimo teorico previsto per la *ApEn*, ovvero al valore dell'entropia di Shannon calcolata a partire dalla distribuzione di probabilità della serie, solo per m abbastanza piccoli (1,2,3 e 4) e come se ne discostino in maniera sostanziale e del tutto fuorviante già per valori di m uguali a 7 o 8. Questo comportamento conferma quanto affermato da Pincus [38], l'ideatore della *ApEn*, che suggerisce di utilizzare, in presenza di serie di lunghezza limitata, valori di m pari, al massimo, a 2 o a 3.

Il problema che si ha nell'utilizzare valori di m relativamente elevati, e quindi pattern di confronto di una certa dimensione, è infatti implicito

nell'algoritmo di calcolo della *ApEn*. Questo indicatore misura infatti il logaritmo della verosimiglianza che insiemi di pattern fra loro "vicini" lo rimangano anche qualora se ne aumenti la dimensione di una unità includendovi il successivo elemento della serie. All'aumentare delle dimensioni del pattern, la probabilità di incontrare due pattern "vicini" (uguali nel caso di serie binarie), su cui poi andare a valutare la probabilità condizionata che anche il successivo elemento della serie sia "vicino", diminuisce in maniera esponenziale, potendosi anche ridurre, in serie di lunghezza limitata, alla sola coincidenza del pattern con se stesso. Studi teorici indicherebbero che, dato m, per ottenere risultati attendibili si dovrebbero considerare serie con un numero di elementi compreso fra 10^m e 30^m [38].

Dai dati riportati nelle tabelle da 5.24 a 5.28 si può notare come, per $m = 0$ e qualsiasi livello di randomizzazione, la *ApEn* approssimi il valore dell'entropia di Shannon della serie.

Nel caso di serie binarie, ma la stessa cosa vale per serie generate da distribuzioni categoriche caratterizzate da un numero finito di eventi, si può infatti dimostrare quanto segue: dalla formula generale per la *ApEn*, posto $r = 0$ e data la generica serie binaria u di N elementi di cui k uno e $N - k$ zeri, si ha che:

$$
\begin{aligned}
&ApEn(m, N)(u) = \Phi^m - \Phi^{m+1}, \quad m \geq 1 \\
&ApEn(0, N)(u) = -\Phi^1 \\
&\Phi^m = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \log C_i^m \\
&C_i^m = \frac{\text{numero di } j \leq N-m+1 \text{ tali che } d(x_i, x_j) = 0}{N-m+1}
\end{aligned}
$$

che, per $m = 0$ diventa:

$$
\begin{aligned}
&ApEn(0, N)(u) = -\Phi^1 \\
&\Phi^1 = \frac{1}{N} \sum_{i=1}^{N} \log C_i^1 \\
&C_i^1 = \frac{\text{numero di } j \leq N \text{ tali che } d(x_i, x_j) = 0}{N}
\end{aligned}
$$

Poiché, per ipotesi, la serie è caratterizzata da k uno, nel caso in cui $x_i = 1$ si avranno k match esatti, ovvero la condizione $d(x_i, x_j) = 0$ sarà soddisfatta k volte e, quindi, $C_i^1 = k/N$. Nel caso in cui, invece, $x_i = 0$, si avranno $N - k$ match esatti per cui $C_i^1 = (N - k)/N$.

Analogamente si avranno k valori dell'indice i per cui $C_i^1 = k/N$ e $N - k$ valori per cui $C_i^1 = (N - k)/N$, di conseguenza:

$$
\Phi^1 = \frac{1}{N} \left[k \log \frac{k}{N} + (N-k) \log \frac{(N-k)}{N} \right]
$$

e, quindi,

$$\Phi^1 = \frac{k}{N}\log\frac{k}{N} + \frac{N-k}{N}\log\frac{(N-k)}{N}$$

Posto $k/N = p$ e $(N-k) = q$ si ha, infine

$$ApEn(0,N)(u) = -(p\log p + q\log q)$$

che è proprio l'entropia di Shannon della serie.

Tabella 5.24: Valori della $ApEn$ e relativi intervalli di confidenza al 95% per una serie di periodo 2 e $p = 0.5$ in funzione di diversi valori di m e di randomizzazione

Pattern='01', P = 0.5, lunghezza della serie = 1000, numero di casi = 5000

	Livello di randomizzazione									
m	0%	25%		50%		75%		100%		
0	1	1,000689	0,999336 1,001003	1,000445	0,998558 1,001000	1,000327	0,997457 1,001001	1,000294	0,997646 1,000999
1	0	0,756202	0,702036 0,807631	0,954472	0,927985 0,976008	0,996643	0,988799 1,000771	0,999590	0,995910 1,000975
2	0	0,673130	0,622072 0,728253	0,925114	0,892887 0,952009	0,992881	0,982815 0,999296	0,998069	0,992927 1,000639
3	0	0,611534	0,555108 0,669550	0,899286	0,862949 0,930954	0,987283	0,973720 0,996571	0,995120	0,987931 0,999501
4	0	0,579764	0,525011 0,637778	0,877324	0,836321 0,911240	0,979530	0,963616 0,991435	0,989299	0,980197 0,995920
5	0	0,544359	0,488930 0,601943	0,850151	0,806604 0,887731	0,965136	0,946810 0,979863	0,977354	0,963595 0,987586
6	0	0,514919	0,464119 0,569322	0,813895	0,766153 0,854435	0,939106	0,915817 0,958733	0,952366	0,934143 0,967592
7	0	0,484803	0,444197 0,528272	0,746791	0,699027 0,791298	0,881636	0,851628 0,909056	0,897223	0,869938 0,922625
8	0	0,460049	0,422857 0,498576	0,649529	0,608235 0,686140	0,759347	0,724074 0,793416	0,773809	0,741075 0,806146
9	0	0,435330	0,401152 0,471806	0,541489	0,508953 0,571012	0,571096	0,538364 0,603111	0,573604	0,539532 0,603616
10	0	0,409931	0,377848 0,442731	0,439073	0,408177 0,471127	0,377770	0,345482 0,412043	0,365523	0,332746 0,397083

Tabella 5.25: Valori della $ApEn$ e relativi intervalli di confidenza al 95% per una serie di periodo 4 e $p = 0.5$ in funzione di diversi valori di m e di randomizzazione

Pattern='0011', P = 0.5, lunghezza della serie = 1000, numero di casi = 5000

	Livello di randomizzazione									
m	0%	25%		50%		75%		100%	
0	0,999995	1,000672	0,999597 1,001002	1,000414	0,998037 1,001003	1,000309	0,997458 1,001001	1,000298	0,997240 1,000999
1	0,999998	1,000557	0,999231 1,001006	1,000066	0,997202 1,000993	0,999690	0,996577 1,000972	0,999590	0,995910 1,000975
2	0	0,756488	0,704155 0,803395	0,952829	0,926286 0,974274	0,995201	0,987236 1,000048	0,998069	0,992927 1,000639
3	0	0,752119	0,698549 0,799641	0,949626	0,922899 0,971551	0,992240	0,983239 0,998640	0,995120	0,987931 0,999501
4	0	0,666442	0,610574 0,720250	0,916090	0,881184 0,945088	0,984025	0,970721 0,994123	0,989299	0,980197 0,995920
5	0	0,653771	0,596135 0,709502	0,904285	0,870143 0,935084	0,972176	0,956146 0,984674	0,977354	0,963595 0,987586
6	0	0,568889	0,509767 0,623258	0,854424	0,812210 0,892323	0,944877	0,922547 0,962087	0,952366	0,934143 0,967592
7	0	0,535142	0,480958 0,586240	0,796556	0,752671 0,840750	0,889788	0,861372 0,915322	0,897223	0,869938 0,922625
8	0	0,470594	0,425128 0,514984	0,679861	0,640018 0,720029	0,766679	0,732090 0,798941	0,773809	0,741075 0,806146
9	0	0,432672	0,388667 0,474729	0,552178	0,518690 0,582852	0,573439	0,540648 0,603315	0,573604	0,539532 0,603616
10	0	0,380573	0,342626 0,415498	0,412933	0,378100 0,446060	0,370943	0,341460 0,405009	0,365523	0,332746 0,397083

Tabella 5.26: Valori della $ApEn$ e relativi intervalli di confidenza al 95% per una serie di periodo 10 e $p = 0.5$ in funzione di diversi valori di m e di randomizzazione

Pattern='0000011111', P = 0.5, lunghezza della serie = 1000, numero di casi = 5000

	Livello di randomizzazione									
m	0%	25%		50%		75%		100%	
0	1,000662	1,000487	0,999186 1,001002	1,000324	0,998399 1,001002	1,000298	0,997652 1,000999	0,999995	0,997240 1,000999
1	0,721909	0,916802	0,891691 0,941402	0,983861	0,968995 0,995232	0,998661	0,992989 1,000946	0,999531	0,995806 1,000968
2	0,649036	0,915907	0,889675 0,939817	0,982186	0,967317 0,993814	0,997220	0,990767 1,000538	0,998072	0,993241 1,000698
3	0,550995	0,890906	0,853795 0,924271	0,976435	0,959096 0,990744	0,994209	0,985980 0,999251	0,995014	0,988183 0,999283
4	0,400010	0,825853	0,777373 0,868796	0,956885	0,930899 0,978301	0,987078	0,975172 0,994833	0,989215	0,980273 0,995872
5	0	0,674248	0,617889 0,730059	0,910696	0,875901 0,939968	0,972652	0,955562 0,984785	0,977156	0,964289 0,987464
6	0	0,631472	0,575671 0,684948	0,877223	0,840784 0,912123	0,946687	0,926367 0,963443	0,952750	0,933918 0,967748
7	0	0,589009	0,538195 0,637014	0,819699	0,777875 0,861097	0,891577	0,860996 0,917997	0,897619	0,871107 0,921637
8	0	0,536319	0,491561 0,581605	0,711048	0,672120 0,749354	0,769892	0,736222 0,802069	0,773695	0,737745 0,804520
9	0	0,472585	0,432936 0,511240	0,561330	0,527390 0,590092	0,573238	0,542228 0,603224	0,572755	0,539900 0,604420
10	0	0,379719	0,341879 0,415423	0,400332	0,365312 0,434975	0,368625	0,337685 0,401272	0,365135	0,333118 0,396719

Tabella 5.27: Valori della $ApEn$ e relativi intervalli di confidenza al 95% per una serie di periodo 10 e $p = 0.3$ in funzione di diversi valori di m e di randomizzazione

Pattern='0000000111', P = 0.3, lunghezza della serie = 1000, numero di casi = 5000

	Livello di randomizzazione									
m	0%	25%		50%		75%		100%		
0	0,87956	0,881199	0,858458 0,901565	0,880912	0,848811 0,908168	0,880040	0,844515 0,914438	0,881396	0,845979 0,912351
1	0,689638	0,820986	0,790916 0,850740	0,868667	0,835093 0,898621	0,879734	0,844477 0,911971	0,880337	0,846096 0,912312
2	0,590023	0,816327	0,782087 0,848687	0,867491	0,833910 0,898109	0,878399	0,842969 0,911259	0,878853	0,844352 0,910794
3	0,360972	0,751660	0,709356 0,789746	0,853669	0,818334 0,884127	0,874736	0,839323 0,907403	0,875875	0,841431 0,908789
4	0,324522	0,726057	0,679552 0,766868	0,841662	0,803413 0,874282	0,868091	0,832462 0,900521	0,869892	0,832008 0,903377
5	0,275496	0,691151	0,643845 0,734610	0,822804	0,784520 0,857577	0,854967	0,817361 0,887425	0,857169	0,819170 0,891799
6	0,200005	0,638991	0,587346 0,687635	0,790203	0,750062 0,826764	0,829001	0,792121 0,863015	0,831246	0,793900 0,867167
7	0	0,556068	0,501250 0,605336	0,741078	0,702752 0,777623	0,783576	0,747389 0,816488	0,785198	0,750505 0,820305
8	0	0,515345	0,460813 0,564127	0,683071	0,649954 0,716014	0,715165	0,682323 0,745597	0,713673	0,682060 0,745294
9	0	0,461297	0,413302 0,505800	0,600799	0,568091 0,634272	0,622240	0,589827 0,653092	0,617353	0,585914 0,646452
10	0	0,378162	0,335739 0,416529	0,489177	0,453687 0,525690	0,508995	0,468219 0,548364	0,507049	0,468622 0,541415

Tabella 5.28: Valori della $ApEn$ e relativi intervalli di confidenza al 95% per una serie di periodo 10 e $p = 0.1$ in funzione di diversi valori di m e di randomizzazione

Pattern='0000000001', P = 0.1, lunghezza della serie = 1000, numero di casi = 5000

	Livello di randomizzazione									
m	0%	25%		50%		75%		100%		
0	0,466454	0,467998	0,427434 0,507051	0,468135	0,417136 0,518931	0,468796	0,410213 0,527600	0,468517	0,406740 0,524777
1	0,452918	0,464810	0,423762 0,505172	0,466784	0,415163 0,515386	0,466002	0,408692 0,521660	0,468731	0,406959 0,524364
2	0,434856	0,459674	0,419508 0,500875	0,464606	0,412164 0,514005	0,464448	0,407791 0,520756	0,467313	0,406710 0,523609
3	0,414159	0,453439	0,412232 0,493504	0,461655	0,409735 0,509571	0,461904	0,404718 0,519902	0,464603	0,404078 0,520709
4	0,390024	0,446200	0,407050 0,485602	0,457667	0,407059 0,505233	0,457976	0,400414 0,514510	0,460312	0,400486 0,516307
5	0,360972	0,437274	0,398493 0,476670	0,452404	0,400846 0,500099	0,452710	0,396145 0,506474	0,454582	0,396097 0,510059
6	0,324522	0,427069	0,387558 0,466308	0,446198	0,395647 0,491792	0,446056	0,392587 0,496994	0,447444	0,389931 0,500179
7	0,275496	0,414646	0,376374 0,455627	0,438841	0,391980 0,484764	0,437669	0,384851 0,487623	0,438326	0,381730 0,490969
8	0,200005	0,399663	0,359134 0,440348	0,430288	0,383982 0,473938	0,428034	0,374257 0,476144	0,427517	0,374330 0,477363
9	0	0,380276	0,333027 0,423112	0,420628	0,377275 0,461047	0,416832	0,367085 0,462721	0,415050	0,364593 0,461504
10	0	0,277910	0,226736 0,330718	0,378730	0,332877 0,419949	0,399995	0,353539 0,443178	0,401311	0,354507 0,444148

Si può poi osservare come la massima variazione nel valore della $ApEn$ si abbia passando da serie perfettamente periodiche a serie perturbate al 25% (cosa comune, per altro, anche agli altri indicatori), mentre nel passaggio da tale livello ai livelli successivi la variazione risulti assai più contenuta e, a volte, difficile da discriminare. In particolare, a parte alcune eccezioni relative alla serie di periodo 2 e $p = 0.5$, i valori calcolati in corrispondenza dei livelli di randomizzazione del 75% e del 100% sono praticamente indistinguibili fra loro.

Ancora, la ApEn è apparsa in grado di individuare con buona precisione la periodicità delle serie, per arrivare a ciò, però, è stato necessario aumentare il valore del parametro m oltre i valori consigliati in letteratura relativamente alla lunghezza delle serie considerate. Inoltre, per $m = 1$ ed in assenza di randomizzazione, la serie di pattern "01" viene "classificata" dalla $ApEn$ come perfettamente regolare ($ApEn = 0$), mentre quella di pattern "0011" è indicata come massimamente irregolare ($ApEn = 1$), seguita, in termini di disordine, da quella di pattern "0000011111", "0000000111" e "0000000001". Per $m = 2$, invece, la serie "0011" risulta essere perfettamente regolare ($ApEn = 0$), mentre la "0000011111" continua a risultare più irregolare della "0000000111", a sua volta più irregolare della "0000000001". Come si è visto, quindi, la scelta di un determinato valore di m, può potenzialmente condurre a risultati anche completamente diversi (grafici 5.4 e 5.4). A parte questo, comunque, il comportamento della $ApEn$ nei confronti del livello di disordine (livello di randomizzazione) parrebbe non discostarsi troppo da quello osservato per gli altri indicatori. Ovviamente anche la $ApEn$ può essere normalizzata rapportandola, di volta in volta, al suo massimo assoluto che, per una distribuzione binaria, corrisponde a $\log_2 2 = 1$, oppure al suo massimo relativo, che coincide con l'entropia di Shannon della serie considerata. In questo modo l'indicatore assumerà sempre il valore uno in corrispondenza della condizione di massimo disordine, condizione che risulterà così immediatamente riconoscibile (tabelle 5.29 e 5.30).

Ovviamente, nel primo caso l'indicatore esprimerà la distanza dal massimo disordine possibile, che si realizza sotto la condizione di massima entropia, ovvero per distribuzioni marginali uniformi ($p = q = 0.5$), mentre nel secondo indicherà più propriamente la distanza dal massimo disordine ammissibile data la particolare distribuzione marginale della serie considerata.

Poiché il calcolo della ApEn non è vincolato alla stima delle distribuzioni marginali e congiunte, essa parrebbe più adatta sia della mutua informazione che della S_ρ ad operare con serie di lunghezza limitata. Nel caso di campioni di ampiezza (lunghezza delle serie) ridotta, infatti, la stima delle distribuzioni marginali e congiunte potrebbero differire in maniera anche considerevole dall'effettiva distribuzione del fenomeno. D'altronde, la S_ρ, la mutua informazione e l'entropia congiunta sono di più agevole ed univoca interpretazione e permettono di determinare il livello di autocorrelazione anche per lag relativamente grandi senza con ciò perdere in affidabilità o incorrere nei problemi

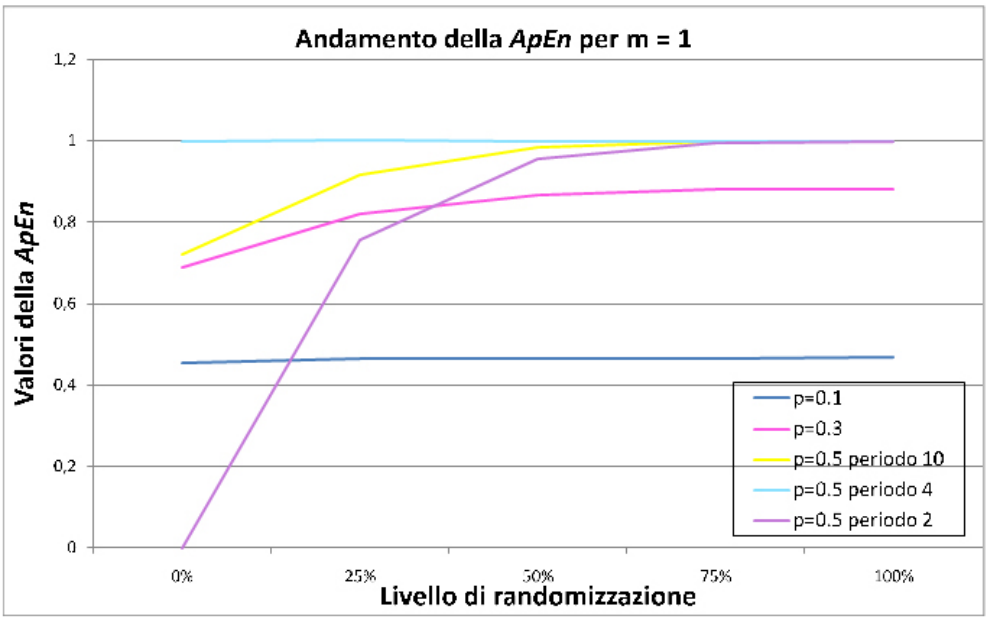

Figura 5.4: Andamento dei valori della $ApEn$ per $m = 1$

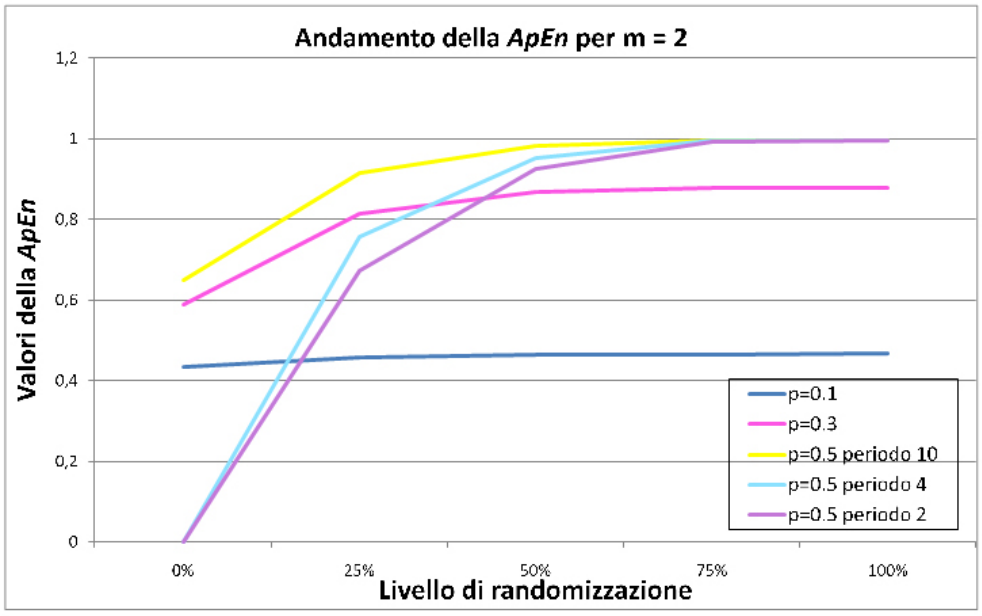

Figura 5.5: Andamento dei valori della $ApEn$ per $m = 2$

che si avrebbero utilizzando la $ApEn$ con valori di m elevati.

Tabella 5.29: Valori della $ApEn$ e relativi intervalli di confidenza al 95% calcolati per $m = 1$

	ApEn, m = 1	Livello di randomizzazione								
	0%	25%		50%		75%		100%		
p=0.5 periodo 2	0	0,756202	0,702036 0,807631	0,954472	0,927985 0,976008	0,996643	0,988799 1,000771	0,999590	0,995910 1,000975	
p=0.5 periodo 4	0,999998	1,000557	0,999231 1,001006	1,000066	0,997202 1,000993	0,999690	0,996577 1,000972	0,999590	0,995910 1,000975	
p=0.5 periodo 10	0,721909	0,916802	0,891691 0,941402	0,983861	0,968995 0,995232	0,998661	0,992989 1,000946	0,999531	0,995806 1,000968	
p=0.3	0,689638	0,820986	0,790916 0,850740	0,868667	0,835093 0,898621	0,879734	0,844477 0,911971	0,880337	0,846096 0,912312	
p=0.1	0,452918	0,464810	0,423762 0,505172	0,466784	0,415163 0,515386	0,466002	0,408692 0,521660	0,468731	0,406959 0,524364	

Tabella 5.30: Valori della $ApEn$ normalizzata e relativi intervalli di confidenza al 95% calcolati per $m = 1$

	ApEn normalizzata, m = 1	Livello di randomizzazione								
	0%	25%		50%		75%		100%		
p=0.5 periodo 2	0	0,756202	0,702036 0,807631	0,954472	0,927985 0,976008	0,996643	0,988799 1,000771	0,999590	0,995910 1,000975	
p=0.5 periodo 4	0,999998	1,000557	0,999231 1,001006	1,000066	0,997202 1,000993	0,999690	0,996577 1,000972	0,999590	0,995910 1,000975	
p=0.5 periodo 10	0,721909	0,916802	0,891691 0,941402	0,983861	0,968995 0,995232	0,998661	0,992989 1,000946	0,999531	0,995806 1,000968	
p=0.3	0,782532	0,931572	0,897452 0,965334	0,985676	0,947579 1,019664	0,998233	0,958227 1,034813	0,998918	0,960064 1,035200	
p=0.1	0,965719	0,991075	0,903552 1,077136	0,995284	0,885217 1,098914	0,993617	0,871420 1,112292	0,999436	0,867725 1,118057	

CAPITOLO 6

Analisi di serie semplici perfettamente periodiche in funzione della lunghezza del periodo

Si considerino ora le serie seguenti, caratterizzate da una alternanza di zeri e di uno costante, e così strutturate:

- *serie 1*: 500 zero e poi 500 uno;
- *serie 2*: 250 zero e poi 250 uno;
- *serie 3*: 100 zero e poi 100 uno;
- *serie 4*: 50 zero e poi 50 uno;
- *serie 5*: 25 zero e poi 25 uno;
- *serie 6*: 10 zero e poi 10 uno;
- *serie 7*: 5 zero e poi 5 uno.

Tali serie sono perfettamente periodiche ed esprimono tutte, sia al limite che per la lunghezza considerata (1000 elementi), una proporzione di zero, q, e di uno, p, costante e pari a 0.5. Da un punto di vista della complessità algoritmica esse sono molto semplici e, soprattutto, equivalenti, potendosi tutte descrivere mediante un semplice programma del tipo:

$K = \ldots$
$T = \ldots$
scrivi K zero e poi K uno T volte

6.1 Risultati dell'esperimento

6.1.1 Entropia di Granger, Maasoumi e Racine

Come si può vedere dai dati normalizzati riportati nella tabella 6.1, il livello di autocorrelazione espresso dalla S_ρ, a parità di lag, diminuisce al diminuire della lunghezza del periodo e ciò si verifica in maniera decisamente più pronunciata in corrispondenza delle serie caratterizzate dai periodi più brevi.

Si può poi osservare come, in corrispondenza del lag 1, la S_ρ relativa alla serie di periodo 500 indichi un'autocorrelazione pari a circa l'89% del massimo ammissibile, mentre per la serie di periodo 250, tale valore si riduca a circa l'85%, valori che appaiono relativamente bassi se si considera che, nel primo caso, si ha perfetta corrispondenza per 998 valori su 1000 e nel secondo per 996 su 1000.

Tabella 6.1: Serie perfettamente periodiche: valori della S_ρ normalizzati in funzione della lunghezza del periodo.

P = 0.5, lunghezza delle serie = 1000

	Alternanza di zeri e di uno						
Lag	**500**	**250**	**100**	**50**	**25**	**10**	**5**
0	1	1	1	1	1	1	1
1	0,894449	0,852145	0,770680	0,682843	0,565934	0,360448	0,175206
2	0,852145	0,793742	0,682843	0,565934	0,415739	0,175206	0,017289
3	0,820249	0,750065	0,618335	0,482188	0,313172	0,072018	0,017289
4	0,793742	0,714015	0,565934	0,415739	0,235865	0,017289	0,175206
5	0,770680	0,682843	0,521295	0,360448	0,175206	0	1
6	0,750065	0,655138	0,482188	0,313172	0,126833	0,017289	0,175206
7	0,731305	0,630064	0,447289	0,272048	0,088204	0,072018	0,017289
8	0,714015	0,607074	0,415739	0,235865	0,057717	0,175206	0,017289
9	0,697926	0,585789	0,386937	0,203786	0,034315	0,360448	0,175206
10	0,682843	0,565934	0,360448	0,175206	0,017289	1	1

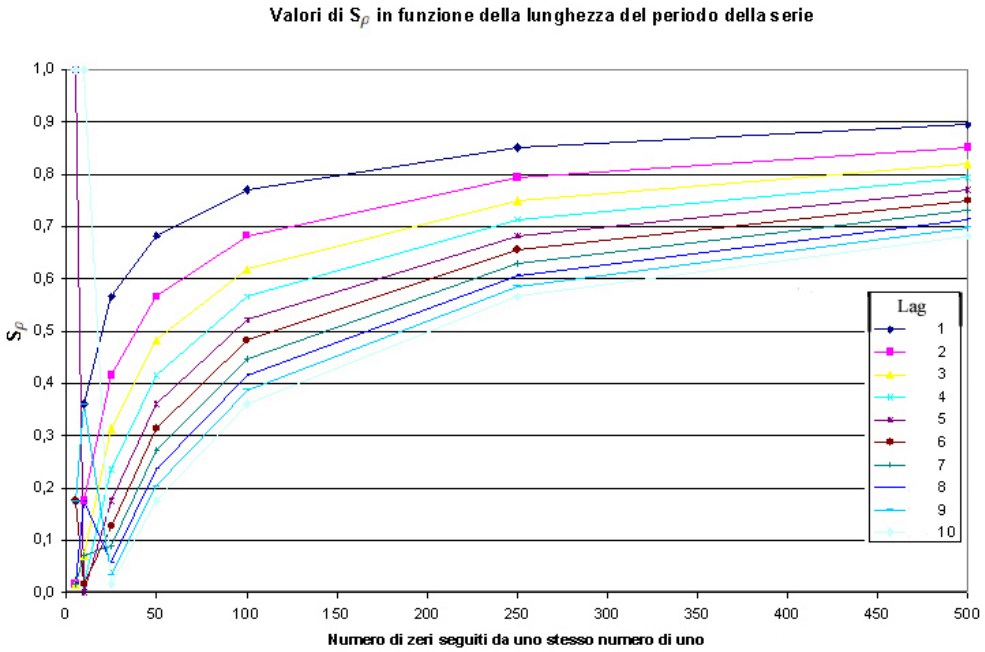

Figura 6.1: Valori della S_ρ normalizzati in funzione della lunghezza del periodo della serie.

La tabella 6.2 descrive invece il comportamento della distribuzione congiunta al variare della lunghezza della serie con riferimento ai lag 0, 1 e 2. Vi si può osservare come il numero iniziale di coppie '00' e '11' cali molto lentamente, e di conseguenza il numero di coppie del tipo '10' e '01' aumenti altrettanto lentamente, al diminuire della lunghezza del periodo delle serie.

Tabella 6.2: Variazione della distribuzione congiunta in funzione del lag e della lunghezza del periodo della serie.

P = 0.5, lunghezza della serie = 1000

	Tipo di serie	Alternanza di zero e di uno								
		500			100			10		
	Lag	0	1	2	0	1	2	0	1	2
Simboli distribuzione congiunta	'00'	500	499	498	500	495	490	500	450	400
	'10'	0	1	2	0	5	10		50	100
	'01'	0	1	2	0	5	10		50	100
	'11'	500	499	498	500	495	490	500	450	400

A parità di lunghezza del periodo, poi, il livello di autocorrelazione evidenziato dalla S_ρ diminuisce all'aumentare del lag, in quanto ci si sta allontanando dalla situazione di perfetta corrispondenza fra gli elementi della

serie. Ciò si verifica fino a che non si raggiunge un lag pari al semiperiodo della serie (vedi, ad esempio, le serie di periodo 5 e 10), superato il quale la S_ρ ricomincia a crescere fino a toccare nuovamente il suo massimo in corrispondenza di un lag pari al periodo della serie stessa.

6.1.2 Mutua informazione

Per la mutua informazione, i cui valori normalizzati sono riportati in tabella 6.3, vale sostanzialmente lo stesso discorso fatto per la S_ρ. Tuttavia, a parità di lag, nel passare da una serie all'altra i valori della mutua informazione relativi alle serie di periodo maggiore (serie da 1 a 4) mostrano diminuzioni meno marcate rispetto alla S_ρ, mentre si osservano diminuzioni più marcate per le serie di periodo più corto (serie 6 e 7), comportamento questo che parrebbe più consono all'effettivo andamento del fenomeno.

Una tendenza analoga si può osservare, inoltre, all'interno di una stessa serie per le variazioni che si hanno passando da un lag all'altro. Per tutte le serie la mutua informazione mostra, nel passare dal lag 0 al lag 10, un intervallo di variazione più contenuto rispetto alla S_ρ. Tale differenza è particolarmente evidente nel passaggio dal lag 0 al lag 1 e, soprattutto, nelle le serie di periodo più lungo. Ciò in accordo al fatto che, in presenza di serie caratterizzate da sequenze di simboli uguali molto lunghe, lo slittamento di un solo elemento di confronto (o di pochi elementi) non incide in maniera sostanziale sul valore dell'autocorrelazione.

Tabella 6.3: Serie perfettamente periodiche: valori della mutua informazione normalizzata in funzione della lunghezza del periodo.

P = 0.5, lunghezza delle serie = 1000

	Alternanza di zeri e di uno						
Lag	**500**	**250**	**100**	**50**	**25**	**10**	**5**
0	1	1	1	1	1	1	1
1	0,979186	0,962378	0,919207	0,858559	0,757708	0,531004	0,278072
2	0,962378	0,932778	0,858559	0,757708	0,597821	0,278072	0,029049
3	0,947085	0,906222	0,805608	0,672555	0,470639	0,118709	0,029049
4	0,932778	0,881650	0,757708	0,597821	0,365690	0,029049	0,278072
5	0,919207	0,858559	0,713603	0,531004	0,278072	0	1
6	0,906222	0,836654	0,672555	0,470639	0,204960	0,029049	0,278072
7	0,893726	0,815739	0,634076	0,415761	0,144549	0,118709	0,029049
8	0,881650	0,795675	0,597821	0,365690	0,095619	0,278072	0,029049
9	0,869941	0,776358	0,563530	0,319923	0,057317	0,531004	0,278072
10	0,858559	0,757708	0,531004	0,278072	0,029049	1	1

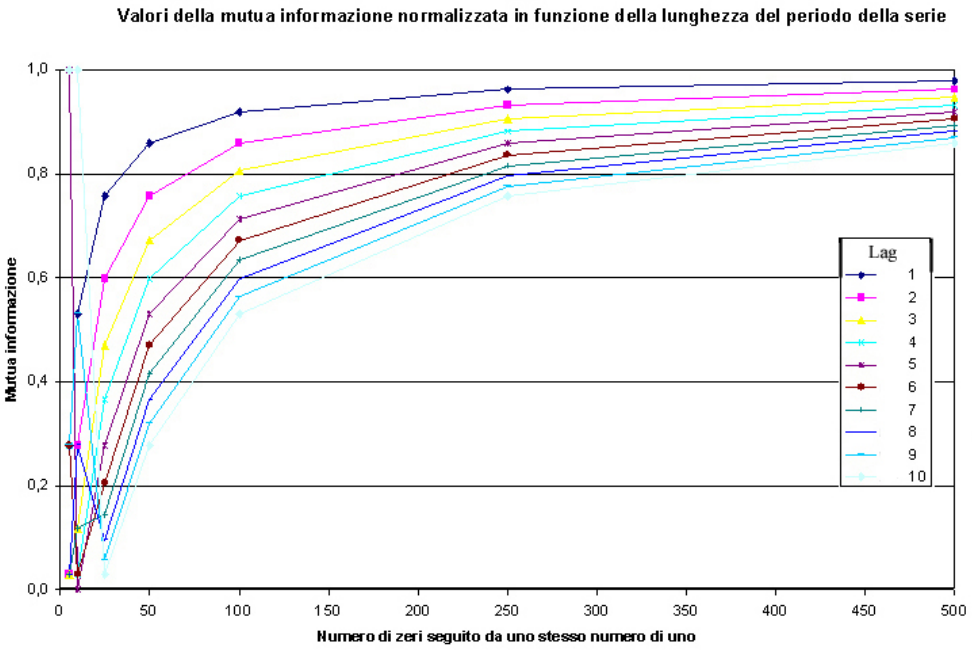

Figura 6.2: Valori della mutua informazione normalizzata in funzione della lunghezza del periodo della serie.

6.1.3 Approximate Entropy

Per le serie caratterizzate da una lunghezza del periodo pari a 500, 250 e 100 elementi, la *ApEn* manifesta un valore molto prossimo a zero, attestando così la sostanziale "regolarità" delle serie stesse. Si può inoltre osservare, almeno fino alla serie di periodo 50, una sostanziale costanza dei valori della *ApEn* all'aumentare del valore del parametro m. Questo perchè tali serie sono costituite da un numero molto elevato di elementi consecutivi uguali, ragion per cui la probabilità condizionata che, dati due pattern uguali, questi lo rimangano anche includendovi il successivo elemento della serie, rimane sempre molto elevata.

Una condizione di questo tipo (sostanziale costanza della *ApEn* al crescere del valore del parametro m), se incontrata nella pratica, potrà quindi far pensare ad una qualche costanza di lungo periodo degli elementi di una serie o, nel caso di serie non binarie, ad una serie che varia abbastanza lentamente.

Vediamo poi come per serie di periodo più limitato (periodo 5 o 10), la *ApEn* non ci fornisca alcuna informazione sulla regolarità delle stesse almeno fino a quando il valore di m non arriva a coincidere con la lunghezza del periodo (tabella 6.4), risultato che però è stato ottenuto violando il dettato teorico-pratico che postula l'utilizzo di valori di m non superiori a 3 nel caso di serie, come le nostre, composte da non più di un migliaio di elementi.

Tabella 6.4: Serie perfettamente periodiche: valori della *ApEn* in funzione della lunghezza del periodo.

P = 0.5, lunghezza delle serie = 1000

	Alternanza di zeri e di uno						
m	**500**	**250**	**100**	**50**	**25**	**10**	**5**
0	0,999929	0,999922	0,999923	0,999922	0,999922	0,999922	1
1	0,010514	0,029486	0,074644	0,136855	0,239504	0,468988	0,721909
2	0,010504	0,029468	0,074516	0,136314	0,237170	0,452924	0,649036
3	0,010507	0,029460	0,074381	0,135755	0,234740	0,434868	0,550995
4	0,010508	0,029445	0,074276	0,135202	0,232201	0,414165	0,400010
5	0,010498	0,029413	0,074132	0,134589	0,229511	0,390026	0
6	0,010500	0,029423	0,074034	0,134045	0,226765	0,360957	0
7	0,010490	0,029389	0,073882	0,133415	0,223814	0,324514	0
8	0,010492	0,029385	0,073760	0,132814	0,220705	0,275490	0
9	0,010493	0,029351	0,073645	0,132178	0,217450	0,200002	0
10	0,010484	0,029359	0,073471	0,131516	0,213951	0	0

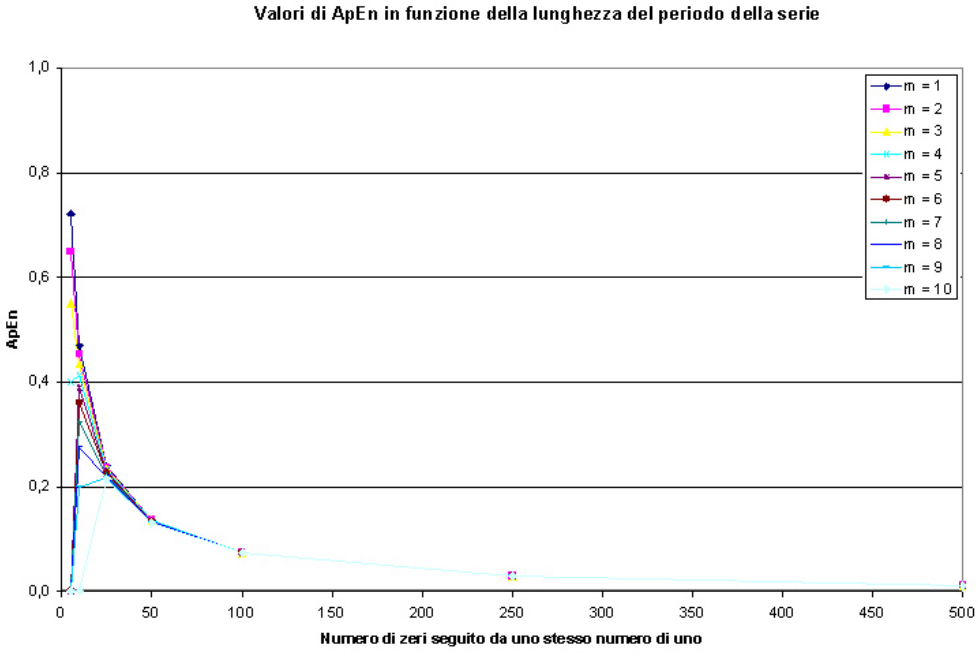

Figura 6.3: Valori della *ApEn* in funzione della lunghezza del periodo della serie.

6.2 Serie caratterizzate da lunghe sequenze di elementi uguali

Per cercare di meglio comprendere il comportamente dei nostri indicatori in presenza di serie caratterizzate da lunghe sequenze di elementi uguali, si considerino ora le serie seguenti:

- *serie 1*: serie composta da 500 zero seguiti da 500 uno;
- *serie 2*: serie composta da 750 zero seguiti da 250 uno;
- *serie 3*: serie composta da 900 zero seguiti da 100 uno;
- *serie 4*: serie composta da 950 zero seguiti da 50 uno;
- *serie 5*: serie composta da 975 zero seguiti da 25 uno;
- *serie 6*: serie composta da 990 zero seguiti da 10 uno;
- *serie 7*: serie composta da 995 zero seguiti da 5 uno;
- *serie 8*: serie composta da 999 zero seguiti da 1 uno.

Queste serie, a differenza delle precedenti, sono caratterizzate da una percentuale di zeri e di uno variabile da serie a serie e, di conseguenza, da differenti valori di entropia (vedi tabella 6.5). Da un punto di vista della

Tabella 6.5: Valori teorici dell'entropia di Shannon calcolati su distribuzioni caratterizzate da diverse probabilità P(X=0).

Valore teorico dell'entropia di Shannon H(X)

P(X=0)	0.5	0.75	0.9	0.95	0.975	0.990	0.995	0.999
H(X)	1,0	0,811278	0,468996	0,286397	0,168661	0,080793	0,045415	0,011408

complessità algoritmica sono però ancora equivalenti fra loro e, sostanzialmente, anche alle serie precedenti. Esse infatti possono essere descritte nel modo seguente:

$K = \ldots$
$T = \ldots$
scrivi K zero e poi T - K uno

Inoltre, per tutte le serie, al lag 1 vi è perfetta corrispondenza per 998 elementi su 1000 e, pertanto, il numero di elementi perfettamente coincidenti è lo stesso per ciascuna serie. Ciò nonostante, dai dati delle tabelle 6.6

e 6.7, si può vedere come i valori normalizzati della S_ρ e della mutua informazione decrescano più che proporzionalmente all'aumentare del numero di zeri consecutivi.

Al lag 1, in particolare, la serie 1 (500 zeri consecutivi) è caratterizzata da un valore della S_ρ pari a circa 0.89, che si riduce a 0.79 per la serie 4 (950 zeri consecutivi), a 0.59 per la serie 6 (990 zeri consecutivi) per scendere praticamente a zero (0.000341) nel caso della serie 8 (999 zeri consecutivi). Per le stesse serie, la mutua informazione assume invece i valori: 0.98, 0.94, 0.80 e 0 (0.000127). In questo contesto, la mutua informazione cala quindi molto più lentamente della S_ρ fornendo in questo modo informazioni più aderenti alla realtà del fenomeno. Tuttavia anch'essa, per la serie 7 (995 zeri) e 8 (999 zeri) fornisce dei valori, 0.67 e 0.000127, che, se presi in maniera acritica, potrebbero risultare del tutto fuorvianti.

Un comportamento analogo lo si osserva per una stessa serie al variare del lag. Ad esempio, la serie 7, caratterizzata da 995 zeri consecutivi, mostra un valore della S_ρ pari a 0.45 in corrispondenza del lag 1 (pari al 45% dell'autocorrelazione massima ammissibile) ed addirittura un valore molto prossimo a zero, condizione di completa indipendenza, a partire dal lag 4. E questo nonostante si abbia perfetta corrispondenza per 998 elementi su 1000 al lag 1, per 992 elementi su 1000 al lag 4 (99.2% di corrispondenze) e per 980 elementi su 1000 al lag 10 (98% di corrispondenze). La mutua informazione, pur manifestando valori leggermente più elevati, non si comporta in maniera molto diversa.

Tabella 6.6: Serie perfettamente periodiche: valori della S_ρ normalizzati in funzione del numero di zeri consecutivi in serie di 1000 elementi.

	Numero di zeri consecutivi su 1000 elementi							
Lag	500	750	900	950	975	990	995	999
0	1	1	1	1	1	1	1	1
1	0,894449	0,881574	0,839906	0,790308	0,720804	0,588616	0,450090	0,000341
2	0,852145	0,834300	0,776839	0,709062	0,615304	0,441475	0,267061	0,000341
3	0,820249	0,798734	0,729744	0,648967	0,538433	0,338039	0,145411	0,000341
4	0,793742	0,769232	0,690920	0,599833	0,476415	0,257414	0,057114	0,000341
5	0,770680	0,743605	0,657386	0,557713	0,423917	0,191564	0,001756	0,000341
6	0,750065	0,720732	0,627612	0,520583	0,378205	0,136452	0,001756	0,000341
7	0,731305	0,699947	0,600690	0,487240	0,337659	0,089814	0,001756	0,000341
8	0,714015	0,680816	0,576028	0,456902	0,301226	0,050447	0,001756	0,000341
9	0,697926	0,663036	0,553216	0,429026	0,268176	0,018286	0,001756	0,000341
10	0,682843	0,646390	0,531953	0,403215	0,237982	0,003599	0,001756	0,000341

Tabella 6.7: Serie perfettamente periodiche: valori della mutua informazione normalizzata in funzione del numero di zeri consecutivi in serie di 1000 elementi.

	Numero di zeri consecutivi su 1000 elementi							
Lag	500	750	900	950	975	990	995	999
0	1	1	1	1	1	1	1	1
1	0,979186	0,974857	0,958774	0,935733	0,896665	0,800933	0,669487	0,000127
2	0,962378	0,954654	0,926111	0,885542	0,817412	0,653382	0,435109	0,000127
3	0,947085	0,936321	0,896702	0,840732	0,747491	0,526774	0,244802	0,000127
4	0,932778	0,919207	0,869418	0,799458	0,683782	0,414879	0,092738	0,000127
5	0,919207	0,903000	0,843725	0,760844	0,624811	0,315016	0,000798	0,000127
6	0,906222	0,887520	0,819306	0,724376	0,569721	0,225957	0,000798	0,000127
7	0,893726	0,872645	0,795954	0,689716	0,517955	0,147409	0,000798	0,000127
8	0,881650	0,858289	0,773520	0,656621	0,469126	0,080066	0,000798	0,000127
9	0,869941	0,844389	0,751894	0,624915	0,422953	0,026560	0,000798	0,000127
10	0,858559	0,830895	0,730990	0,594459	0,379229	0,001804	0,000798	0,000127

Tabella 6.8: Serie perfettamente periodiche: valori della $ApEn$ in funzione del numero di zeri consecutivi in serie di 1000 elementi.

	Numero di zeri consecutivi su 1000 elementi							
m	500	750	900	950	975	990	995	999
0	0,999929	0,799048	0,439498	0,244142	0,113274	0	0	0
1	0,010514	0,011103	0,011368	0,011448	0,011486	0,011508	0	0
2	0,010504	0,011098	0,011370	0,011447	0,011485	0,011507	0	0
3	0,010507	0,011101	0,011364	0,011445	0,011483	0,011505	0	0
4	0,010508	0,011098	0,011365	0,011444	0,011482	0,011504	0	0
5	0,010498	0,011094	0,011363	0,011442	0,01148	0,011502	0	0
6	0,010500	0,011096	0,011363	0,011440	0,011478	0,011501	0,011508	0
7	0,010490	0,011090	0,011358	0,011439	0,011477	0,011500	0,011507	0
8	0,010492	0,011093	0,011358	0,011437	0,011476	0,011498	0,011505	0
9	0,010493	0,011086	0,011356	0,011437	0,011474	0,011496	0,011504	0
10	0,010484	0,011085	0,011356	0,011434	0,011472	0,011495	0,011502	0,011508

Tabella 6.9: Serie perfettamente periodiche: valori della $ApEn$ normalizzata in funzione del numero di zeri consecutivi in serie di 1000 elementi.

	Numero di zeri consecutivi su 1000 elementi							
m	500	750	900	950	975	990	995	999
0	0,999929	0,984925	0,937104	0,852460	0,671608	0	0	0
1	0,010514	0,013686	0,014012	0,014111	0,014158	0,014185	0	0
2	0,010504	0,013680	0,014015	0,014110	0,014157	0,014184	0	0
3	0,010507	0,013683	0,014008	0,014107	0,014154	0,014181	0	0
4	0,010508	0,013680	0,014009	0,014106	0,014153	0,014180	0	0
5	0,010498	0,013675	0,014006	0,014104	0,014151	0,014178	0	0
6	0,010500	0,013677	0,014006	0,014101	0,014148	0,014176	0,014185	0
7	0,010490	0,013670	0,014000	0,014100	0,014147	0,014175	0,014184	0
8	0,010492	0,013673	0,014000	0,014098	0,014146	0,014173	0,014181	0
9	0,010493	0,013665	0,013998	0,014098	0,014143	0,014170	0,014180	0
10	0,010484	0,013664	0,013998	0,014094	0,014141	0,014169	0,014178	0,014185

I risultati di cui sopra consiglierebbero quindi grande prudenza nell'utilizzo di questi indicatori in presenza di casi limite quali quelli studiati. Fra i due, comunque, la mutua informazione parrebbe più stabile e in grado di rispondere in maniera più adeguata rispetto alla S_ρ.

Per quanto riguarda la $ApEn$, sia nella sua forma semplice che normalizzata (tabelle 6.8 e 6.9), si osservano invece valori simili e molto prossimi a zero per tutte le serie e per tutti e 10 i valori del parametro m considerati. In questo caso l'indicatore sembrerebbe quindi in grado di rilevare con successo la sostanziale regolarita di tutte le serie analizzate.

Una cosa che risulta evidente, infine, è che nessuno degli indicatori considerati è in grado di fornirci informazioni rispetto al grado di complessità delle serie analizzate in quanto, come si può vedere chiaramente nelle tabelle 6.8 e 6.10, serie che possono a ragione essere considerate equivalenti da un punto di visto della complessità algoritmica, sono caratterizzate da valori anche molto differenti della S_ρ, della mutua informazione e della $ApEn$.

Tabella 6.10: Serie perfettamente periodiche: valori non normalizzati della S_ρ e dell'entropia congiunta in corrispondenza del lag 0.

Lag 0, valori non normalizzati

	Numero di zeri consecutivi su 1000 elementi							
	500	750	900	950	975	990	995	999
Sp	0,292893	0,225481	0,114562	0,062874	0,033312	0,013962	0,007137	0,001468
Mutua informazione	1	0,811278	0,468996	0,286397	0,168661	0,080793	0,045415	0,011408

CAPITOLO 7

Analisi di serie deterministiche complesse periodiche e non periodiche

Si considerino ora le due serie periodiche:

- *serie 1*: serie di periodo 10 e pattern "1100111000";
- *serie 2*: serie di periodo 20 e pattern "11001110000001110011";

e le due serie non periodiche:

- *serie 3*: serie di pattern "100111000011111..."
- *serie 4*: serie di pattern "101100111000...";

Le serie 1 e 2 differiscono da quelle del precedente capitolo 6 in virtù di un disegno del periodo più complesso. Nelle serie di cui al capitolo precedente, infatti, il periodo era sempre costituito da due soli blocchi di simboli consecutivi, il primo composto da zeri, il secondo da uno. La lunghezza dei blocchi variava da serie a serie a formare periodi di varia lunghezza e disegno. Nelle serie che si stanno considerando adesso, invece, sequenze di zero e di uno di diversa lunghezza si alternano all'interno del periodo.

Le serie 3 e 4 sono invece generate da due diversi algoritmi. La serie 3 è composta da un uno a cui fanno seguito due zeri, poi tre uno, poi quattro zeri, e così via. L'elemento n-esimo della serie sarà pertanto costituito da

una sequenza di n simboli uguali, l'$n+1$-esimo da $n+1$ simboli uguali ma di tipo opposto a quelli dell'elemento precedente e via dicendo. La serie 4, invece, è composta da un uno seguito da uno zero, da due uno seguiti da due zeri, da tre uno seguiti da tre zeri e così via. L'n-esimo elemento sarà quindi composto da una sequenza di $n/2$ simboli, se n n è pari, da $(n+1)/2$ se n è dispari. Anche in questo caso, i simboli che compongono un determinato elemento della serie saranno sempre di tipo opposto rispetto a quelli che caratterizzano l'elemento precedente e quello successivo. Per n tendente all'infinito, tutte e quattro le serie esprimono una proporzione di zero, q, e di uno, p, costante e pari a 0.5. Per la lunghezza considerata (1000 elementi), invece, le prime due serie sono caratterizzate da una proporzione di zeri e di uno perfettamente uguale ($p = q = 0.5$), mentre la serie 3 è caratterizzata da $p = 0.494$ e la serie 4 da $p = 0.504$.

Da un punto di vista della complessità algoritmica, la serie 1 può essere considerata come la più semplice, seguita dalla 2, il cui periodo, di lunghezza doppia, è composto dalla sequenza che caratterizza il periodo della serie 1 seguita dalla medesima sequenza riscritta però in ordine inverso. Si può poi senz'altro affermare che la serie 3 è più complessa della serie 2 e che la serie 4 ha una complessità algoritmica pari o superiore a quella della serie 3.

7.1 Risultati dell'esperimento

7.1.1 Entropia di Granger, Maasoumi e Racine

Come si può vedere dai dati normalizzati riportati nella tabella 7.1, le serie 1 e 2 sono entrambe caratterizzate da 2 distinti valori della S_ρ che si ripresentano ad intervalli regolari al variare del lag (0.017289 e 0.175206 per la serie 1, 0.072018 e 0.017289 per la serie 2). Un tale comportamento, se incontrato nella pratica, potrebbe quindi far pensare ad una qualche regolarità nel disegno della serie.

Le serie 3 e 4 mostrano invece valori della S_ρ che, prossimi a 0.5 in corrispondenza del lag 1, vanno via via diminuendo all'aumentare del valore del lag, come ad indicare una mancanza di regolarità o una regolarità, non evidenziata, di periodo più lungo.

I valori della S_ρ calcolati sulle serie 1 e 2 si manterrebbero costanti anche nel caso in cui si aumentasse la lunghezza delle serie. Infatti, purchè tale lunghezza fosse un multiplo del periodo (ma per serie sufficientemente lunghe questa condizione non è necessaria), la struttura della serie, e quindi la sua distribuzione di probabilità, non cambierebbe, così come non cambierebbe la distribuzione congiunta e, di conseguenza, il valore della S_ρ.

Nel caso delle serie 3 e 4, invece, i valori della S_ρ tenderebbero ad aumentare perchè, per come sono strutturati gli algoritmi generatori, al crescere della lunghezza delle serie aumenterebbe anche la dimensione dei gruppi di simboli consecutivi uguali. Di conseguenza, con valori di lag > 0, se per serie

Tabella 7.1: Serie deterministiche complesse periodiche e non periodiche: valori della S_ρ normalizzata.

Lunghezza della serie = 1000

	Pattern della serie			
Lag	1100111000	1100111000000111 0011	1001110000111 11....	1011001110 00....
0	1	1	1	1
1	0,017289	0,072018	0,547277	0,474892
2	0,175206	0,017289	0,395307	0,310971
3	0,175206	0,072018	0,296072	0,208348
4	0,017289	0,017289	0,222578	0,136780
5	0,175206	0,017289	0,167237	0,086486
6	0,017289	0,017289	0,122555	0,051862
7	0,175206	0	0,089073	0,027213
8	0,175206	0,017289	0,061844	0,011867
9	0,017289	0,017289	0,041239	0,003172
10	1	0,017289	0,025966	0,000088

sufficientemente lunghe le distribuzioni marginali rimarrebbero sostanzialmente costanti, nella distribuzione congiunta aumenterebbero invece le percentuali di corrispondenze esatte (coppie "00" e "11") rispetto a quelle non esatte (coppie "10" e "01"), il cui numero diventerebbe via via sempre più trascurabile rispetto al totale avvicinandosi, attraverso valori della S_ρ crescenti, al valore di perfetta autocorrelazione in corrispondenza della quale la S_ρ esprime il suo massimo.

7.1.2 Mutua informazione

La mutua informazione (vedi tabella 7.2) presenta un comportamento del tutto analogo a quello della S_ρ. L'unica differenza osservabile è che la mutua informazione esprime valori mediamente più elevati sia per le serie periodiche che per quelle non periodiche.

Tabella 7.2: Serie deterministiche complesse periodiche e non periodiche: valori della mutua informazione normalizzata.

Lunghezza della serie = 1000

	Pattern della serie			
Lag	1100111000	1100111000000111 0011	1001110000111 11....	1011001110 000....
0	1	1	1	1
1	0,029049	0,118709	0,739637	0,664692
2	0,278072	0,029049	0,573611	0,467774
3	0,278072	0,118709	0,448074	0,326521
4	0,029049	0,029049	0,346881	0,220233
5	0,278072	0,029049	0,266215	0,141814
6	0,029049	0,029049	0,198349	0,086070
7	0,278072	0	0,145905	0,045540
8	0,278072	0,029049	0,102262	0,019950
9	0,029049	0,029049	0,068646	0,005330
10	1	0,029049	0,043394	0,000102

7.1.3 Approximate Entropy

La *ApEn* (vedi tabella 7.3) mostra invece, per tutte e 4 le serie, un andamento decrescente all'aumentare del valore di m. Tuttavia, per le serie 1 e 2 esso parte da valori molto alti e prossimi a uno (rispettivamente 0.971 e 0.881), mantenendosi a livelli relativamente elevati (≥ 0.4) per i primi 4 valori di m, per poi scendere a zero o a valori prossimi a zero per $m = 5$, nel caso della serie 1 e per $m = 8$ nel caso della serie 2. La *ApEn* non riesce quindi ad individuare chiaramente la regolarità insita nelle serie 1 e 2 almeno per quei valori di m ($m = 2$ o $m = 3$) che lo stesso Pincus suggerisce di utilizzare.

Tabella 7.3: Serie periodiche complesse e serie deterministiche: valori della *ApEn*.

Lunghezza della serie = 1000

	Pattern della serie			
m	1100111000	1100111000000111 0011	1001110000111 11....	1011001110 000....
0	1	1	0,999633	0,999998
1	0,970931	0,881269	0,262308	0,333819
2	0,550997	0,689677	0,254894	0,326610
3	0,400013	0,461442	0,253462	0,323285
4	0,400010	0,414624	0,246131	0,316061
5	0	0,375892	0,244673	0,308713
6	0	0,199980	0,237370	0,301488
7	0	0,199978	0,235907	0,294320
8	0	0,098991	0,228666	0,287172
9	0	0	0,227181	0,279993
10	0	0	0,219988	0,272933

Le serie 3 e 4 presentano invece, al variare di m, valori della *ApEn* più omogenei e, soprattutto, più bassi. Per $m = 1$ la *ApEn* è pari a 0.262 per la serie 3 ed a 0.339 per la serie 4, ad indicare un elevato grado di

regolarità, maggiore per la serie 3 che non per la serie 4. In questo caso, la maggiore regolarità rilevata dalla *ApEn* dipende dal fatto che la lunghezza delle sequenze di simboli uguali cresce più velocemente nella serie 3 che non nella 4 e, di conseguenza, all'aumentare della lunghezza complessiva della serie, aumenta anche la probabilità condizionata che, trovati uguali due gruppi di simboli consecutivi $\{x_i, \ldots, x_{i+n}\}$ e $\{x_j, \ldots, x_{j+n}\}$, lo siano anche i successivi elementi x_{i+n+1} ed x_{j+n+1}.

Anche in questo caso, poi, la *ApEn* non è in grado di graduare le serie analizzate in base al loro livello di complessità algoritmica. Considerando i valori della *ApEn* relativi a $m = 1$, rispettivamente 0.970931, 0.881269, 0.262308 e 0.333819, la serie 3 verrebbe indicata come la meno complessa, seguita dalla 4, dalla 2 e dalla 1, classifica che non rispecchia assolutamente la complessità reale delle serie considerate.

7.2 Serie con un numero limitato di elementi

Per verificare il comportamento della S_ρ, della mutua informazione e della *ApEn* in presenza di serie di lunghezza limitata e disegno deterministico non periodico, gli indicatori sono stati applicati alle serie 3 e 4 considerandone però solo i primi 25, 50 e 100 elementi.

In corrispondenza del lag 1 tutte le serie così ottenute presentano valori della S_ρ e della mutua informazione (vedi tabelle 7.4 e 7.5) che partono da un livello molto basso e crescono all'aumentare del numero degli elementi della serie, in accordo con le considerazioni già espresse nella precedente sezione 7.1.1. Per i lag successivi non si riesce, invece, ad individuare un comportamento di tipo generale e questo, probabilmente, proprio a causa della limitata lunghezza delle serie considerate e della conseguente difficoltà di ottenere stime attendibili delle distribuzioni di frequenza sia marginali che congiunte.

Tabella 7.4: Serie deterministiche non periodiche: valori della S_ρ normalizzata in funzione della lunghezza della serie.

	N. elementi serie 100111000011111...			N. elementi serie 101100111000...		
Lag	25	50	100	25	50	100
0	1	1	1	1	1	1
1	0,126476	0,205536	0,292411	0,033725	0,105484	0,187705
2	0,008588	0,046016	0,116444	0,026883	0,002881	0,038381
3	0,018712	0,003428	0,039610	0,111840	0,021431	0,001197
4	0,064266	0,008986	0,006773	0,158624	0,080255	0,016451
5	0,089594	0,031945	0,001743	0,085168	0,115806	0,055253
6	0,129433	0,080574	0,015945	0,011991	0,080255	0,085126
7	0,060259	0,092275	0,035582	0,028632	0,052627	0,112636
8	0,019854	0,107518	0,065393	0,071112	0,009062	0,103297
9	0,003370	0,088204	0,088548	0,071112	0,009066	0,078827
10	0,009126	0,046016	0,097293	0,033725	0,019665	0,040665

Tabella 7.5: Serie deterministiche non periodiche: valori dell'entropia congiunta normalizzata in funzione della lunghezza della serie.

	N. elementi serie 1001110000011111...			N. elementi serie 1011001110000...		
Lag	25	50	100	25	50	100
0	1	1	1	1	1	1
1	0,204406	0,320727	0,442548	0,051733	0,170888	0,296432
2	0,009756	0,075313	0,188488	0,040163	0,003689	0,063770
3	0,012405	0,004629	0,065713	0,163857	0,031354	0,000851
4	0,100447	0,010480	0,010247	0,214316	0,121996	0,025065
5	0,146546	0,042603	0,000339	0,098204	0,178289	0,089099
6	0,204406	0,120726	0,024107	0,000794	0,121996	0,137114
7	0,095118	0,145668	0,058110	0,028730	0,077191	0,178669
8	0,028699	0,173600	0,107617	0,098204	0,004739	0,161668
9	0,001068	0,144549	0,145045	0,098204	0,004739	0,119632
10	0,010743	0,075313	0,158659	0,051733	0,028386	0,057494

La $ApEn$, a sua volta, parte da valori molto elevati e prossimi alla condizione di massima casualità (serie di 25 elementi) per diminuire all'aumentare della lunghezza delle serie (vedi tabella 7.6). La diminuzione è più sensibile per la serie 3 che per la serie 4 in quanto, come già osservato nella precedente sezione 7.1.3, l'algoritmo generatore della serie 3 fa si che tale serie venga caratterizzata da lunghe sequenze consecutive di simboli uguali più velocemente rispetto alla serie 4 e che, quindi, l'algoritmo di calcolo della $ApEn$ la identifichi come maggiormente regolare rispetto all'altra. In questo contesto, valori della $ApEn$ calcolati per $m > 3$ non sono

Tabella 7.6: Serie deterministiche: valori della $ApEn$ in funzione della lunghezza della serie.

	N. elementi serie 1001110000011111...			N. elementi serie 1011001110000...		
m	25	50	100	25	50	100
0	0,970950	1	0,996792	0,970950	0,998196	1
1	0,918296	0,721929	0,566015	0,950978	0,848088	0,721930
2	0,633985	0,602996	0,536123	0,767319	0,743956	0,661687
3	0,583659	0,577799	0,520016	0,500652	0,665959	0,603438
4	0,266667	0,468228	0,460544	0,316992	0,552497	0,545444
5	0,266667	0,450000	0,444407	0,266667	0,437744	0,489788
6	0,133333	0,375489	0,388098	0,133333	0,318872	0,434581
7	0,133333	0,337744	0,371896	0	0,250000	0,382293
8	0	0,218873	0,316272	0	0,187744	0,330827
9	0	0,168872	0,299930	0	0,168872	0,279105
10	0	0,100000	0,248495	0	0,100000	0,257992

assolutamente informativi a causa dell'estrema brevità delle serie e grossi interrogativi dovrebbero porsi, specie per le serie di 25 e 50 elementi, anche per $m = 2$ ed $m = 3$. In ogni caso, per serie di lunghezza limitata quali quelle considerate, il potere informativo degli indicatori utilizzati sembra essere piuttosto limitato.

CAPITOLO 8

Analisi di serie generate da un processo stocastico

Si consideri il semplice processo generatore di una serie stocastica rappresentato dallo schema di figura 8.1.

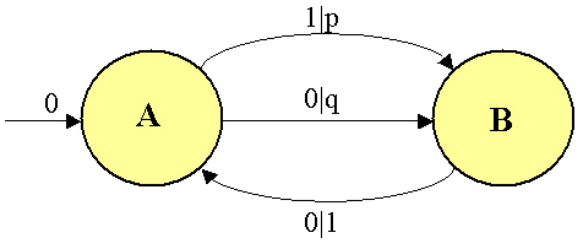

Figura 8.1: Esempio di processo generatore per una serie stocastiche

Al tempo t_0 ci si trova in una modalità di partenza A e in uno stato $x_{t_0} = 0$,. Al successivo istante t_1 il sistema passa nella modalità B mantenendo lo stato 0, $x_{t_1} = 0$, con probabilità q, oppure assumendo lo stato 1, $x_{t_1} = 1$, con probabilità p. Una volta in B il sistema è obbligato, all'istante successivo, a ritornare nella modalità A, caratterizzata dallo stato $x_t = 0$. All'istante ancora successivo, il sistema ritorna necessariamente alla modalità B mantenendo lo stato 0 con probabilità q oppure assumendo lo stato 1

con probabilità p. In sostanza, il sistema passa, ad ogni istante successivo, dalla modalità A alla B e viceversa (modalità ricorrenti) potendo assumere:

- nella transizione da A a B, alternativamente lo stato $x_t = 0$ con probabilità q oppure lo stato $x_t = 1$ con probabilità pari a p;
- nella transizione da B a A, unicamente lo stato $x_t = 0$.

8.1 Modalità di esecuzione dell'esperimento

Per mezzo del processo stocastico precedentemente descritto sono stati generati tre insiemi di 5000 simulazioni ciascuno. Ogni insieme di simulazioni era caratterizzato da una diversa coppia di probabilità di trasizione, ovvero:

- *insieme 1*: $p = 0.25$ e $q = 0.75$;
- *insieme 2*: $p = 0.5$ e $q = 0.5$;
- *insieme 3*: $p = 0.75$ e $q = 0.25$;

Ciascuna simulazione consisteva, data una determinata coppia di probabilità di trasizione, nella generazione di una serie di 1000 elementi. Per ogni serie è stato poi calcolato il valore della S_ρ, della mutua informazione e della $ApEn$ (in questo caso solo sulle prime 1000 simulazioni di ogni insieme). A partire dai valori della S_ρ, della mutua informazione e della $ApEn$ calcolati sulle singole simulazioni, è stato possibile ottenere le relative distribuzioni empiriche e, di consequenza, calcolarne i valori medi, la varianza ed i quantili.

Nella tabella 8.1 sono riportate le proporzioni di uno attese ed empiriche per ciascuno dei tre livelli di probabilità considerati, mentre nella tabella 8.2 sono evidenziati i valori teorici ed empirici dell'entropia di Shannon. Dati:

- $P(B_{t+1}, A_t) = 0.5$ la probabilità di passare, al tempo t, dalla modalità A alla B;
- $P(A_{t+1}, B_t) = 0.5$ la probabilità di passare, al tempo t, dalla modalità B alla A;
- $P(x_{t+1} = 1 | B_{t+1}, A_t) = p$, la probabilità che, nel passare dalla modalità A alla B lo stato assuma valore 0;
- $P(x_{t+1} = 1 | A_{t+1}, B_t) = 0$, la probabilità che, nel passare dalla modalità B alla A lo stato assuma valore 1;

allora la proporzione attesa di uno può essere calcolata come segue:

$$P(B_{t+1}, A_t)P(x_{t+1} = 1 | B_{t+1}, A_t) + P(A_{t+1}, B_t)P(x_{t+1} = 1 | A_{t+1}, B_t)$$

ovvero $0.5p$.

Tabella 8.1: Percentuale teorica ed empirica (con intervalli di confidenza al 95%) di uno sul totale della serie in funzione della probabilità di transizione p.

Lunghezza della serie = 1000

	Probabilità p					
	0.25		0.50		0.75	
Teorica	0,125000		0,250000		0,375000	
Empirica	0,125011	0,106000 0,144000	0,250084	0,228000 0,272000	0,375142	0,356000 0,394000

Tabella 8.2: Valori teorici ed empirici (con intervalli di confidenza al 95%) dell'entropia di Shannon in funzione della probabilità di transizione p.

Lunghezza della serie = 1000

	Probabilità p					
	0.25		0.50		0.75	
Teorica	0,543564		0,811278		0,954434	
Empirica	0,542985	0,487732 0,594619	0,810936	0,774509 0,844320	0,954245	0,939313 0,967332

8.2 Risultati dell'esperimento

8.2.1 Entropia di Granger, Maasoumi e Racine

Come si può vedere dai dati riportati nella tabella 8.3, per ciascuna serie, lo stimatore S_ρ assume sempre e solo due valori i quali risultano essere significativamente diversi tra loro. Il valore più alto si presenta inoltre, contrariamente a quanto ci si poteva aspettare, in corrispondenza dei lag dispari. Infatti, per costruzione, le serie sono caratterizzate dall'avere uno zero come primo simbolo, zero che si ripresenta poi sempre, in maniera costante, ogni due simboli. Ogni serie avrà quindi una struttura del tipo: "0x0x0x0x0x0x0x0x0x0x.....", dove il simbolo "x" può assumere, indifferentemente, il valore 0 oppure 1 con una probabilità che dipende solo dal parametro p. In presenza di lag pari, quindi, vi è sicura corrispondenza fra almeno la metà degli elementi della serie, in quanto gli zeri fissi si incontrano sempre fra di loro.

Inoltre, come si può osservare, per $p = 0.25$ i valori della S_ρ sono tutti molto bassi, di conseguenza la S_ρ parrebbe indicare una condizione abbastanza prossima a quella di indipendenza nonostante che, in media per i lag pari, si abbia perfetta corrispondenza per almeno il 50% degli elementi della serie. Il valore della S_ρ tende poi ad aumentare, anche se molto lentamente, al crescere di p e, quindi, al conseguente aumento della percentuale di uno. La S_ρ raggiunge infine il suo massimo in corrispondenza di $p = 1$, condizione

in cui si ha il 50% di zero e il 50% di uno che si ripetono alternativamente (ovvero "01010101....") e, quindi, perfetta regolarità (periodicità di periodo 2) e massima autocorrelazione.

Dai valori riportati nella tabella 8.3 risulta inoltre evidente come la S_ρ riesca a discriminare chiaramente le 3 diverse condizioni di partenza ($p = 0.25, p = 0.50$ oppure $p = 0.75$).

Tabella 8.3: Valori della S_ρ normalizzati e relativi intervalli di confidenza al 95% in funzione della probabilità di transizione p.

Lunghezza della serie = 1000

Lag	Valori di p					
	0.25		0.50		0.75	
0	1	0,896356 1,100013	1	0,954884 1,042602	1	0,983744 1,014718
1	0,061473	0,045962 0,078455	0,163260	0,138000 0,189874	0,343431	0,306880 0,381603
2	0,015036	0,004305 0,028865	0,055900	0,037204 0,077264	0,178437	0,142394 0,215782
3	0,061479	0,045962 0,078145	0,163258	0,138000 0,190363	0,343421	0,306880 0,381603
4	0,015006	0,004457 0,028654	0,056036	0,036983 0,077080	0,178742	0,143859 0,216031
5	0,061483	0,045962 0,078145	0,163261	0,138000 0,190370	0,343423	0,306880 0,381603
6	0,015045	0,004322 0,029140	0,056122	0,036844 0,077724	0,179141	0,144298 0,215643
7	0,061485	0,045962 0,078145	0,163258	0,138000 0,190501	0,343431	0,306880 0,381603
8	0,014987	0,004302 0,028497	0,055928	0,036886 0,077809	0,178923	0,144346 0,216211
9	0,061489	0,045962 0,077963	0,163260	0,138000 0,190363	0,343425	0,306880 0,381603
10	0,015241	0,004574 0,029494	0,056007	0,037057 0,077767	0,178709	0,144186 0,215621

8.2.2 Mutua informazione

Per quanto riguarda la mutua informazione (vedi tabella 8.4) possiamo osservare, come già negli esperimenti precedenti, un comportamento del tutto analogo a quello della S_ρ. L'unica differenza risiede nell'intensità dei valori. Se in corrispondenza dei lag dispari i valori osservati sono sostanzialmente uguali, nel caso di lag pari i valori della S_ρ risultano significativamente inferiori rispetto a quelli assunti dalla mutua informazione confermando così, anche in questo caso, la tendenza della mutua informazione a mantenere

valori più elevati rispetto alla S_ρ al diminuire del livello di autocorrelazione nelle serie osservate.

Tabella 8.4: Valori della mutua informazione normalizzata e relativi intervalli di confidenza al 95% in funzione della probabilità di transizione p.

Lunghezza della serie = 1000

Lag	Valori di p					
	0.25		0.50		0.75	
0	1	0,915112 1,080787	1	0,963679 1,034089	1	0,987011 1,011734
1	0,047992	0,035518 0,061773	0,151839	0,126651 0,178686	0,365576	0,322285 0,411200
2	0,023683	0,006595 0,045665	0,091572	0,061426 0,125554	0,282513	0,229044 0,336820
3	0,047997	0,035518 0,061297	0,151833	0,126651 0,179498	0,365560	0,322285 0,411200
4	0,023623	0,006849 0,045518	0,091792	0,061008 0,125392	0,282968	0,231507 0,337271
5	0,047999	0,035518 0,061297	0,151837	0,126651 0,179498	0,365561	0,322285 0,411200
6	0,023686	0,006634 0,046124	0,091926	0,060672 0,126443	0,283562	0,231900 0,336663
7	0,047997	0,035518 0,061297	0,151828	0,126651 0,179498	0,365574	0,322285 0,411197
8	0,023586	0,006626 0,045133	0,091605	0,060672 0,126448	0,283238	0,231961 0,337385
9	0,047999	0,035518 0,061297	0,151829	0,126649 0,179498	0,365562	0,322282 0,411200
10	0,023992	0,007061 0,046827	0,091734	0,061015 0,126478	0,282920	0,231694 0,336354

8.2.3 Approximate Entropy

La *ApEn* raggiunge il suo valore più elevato, circa 0.69 per $m = 1$, in corrispondenza di $p = 0.5$ (proporzione di uno nella serie pari al 25%), mentre per $p = 0,25$ e per $p = 0,75$ i valori riscontrati sono entrambi più bassi e pari rispettivamente, sempre per $m = 1$, a circa 0.52 e 0.61 (vedi tabella 8.5).

Anche in questo caso, pur in presenza di un andamento diverso rispetto agli altri due indicatori, le 3 condizioni iniziali sono chiaramente discriminate. La *ApEn* parrebbe inoltre fornire indicazioni più attendibili circa la struttura della serie, con un grado di disordine riscontrato variabile dal 50% (per $p = 0.25$) a circa il 70% (per $p = 0.50$) del massimo teorico ($\log_2 2 = 1$). La relativa costanza del valore della *ApEn* all'aumentare del valore del parametro m, in ispecie nel caso di $p = 0.25$, rafforza ulteriormente l'ipotesi della presenza di un qualche tipo di struttura regolare all'interno delle serie.

Si può inoltre notare come il valore più elevato della $ApEn$ non si osservi in corrispondenza della serie che esprime il valore dell'entropia di Shannon più elevato (serie caratterizzata da $p = 0.75$), ma in corrispondenza di quella che presenta il più alto valore dell'entropia riferita alla sola componente aleatoria (serie caratterizzata da $p = 0.5$).

Tabella 8.5: Valori della $ApEn$ e relativi intervalli di confidenza al 95% in funzione della probabilità di transizione p.

Lunghezza della serie = 1000

	Valori di p					
m	**0.25**		**0.50**		**0.75**	
0	0,544122	0,488449 0,596283	0,810560	0,773996 0,844738	0,955313	0,941850 0,965442
1	0,517933	0,470228 0,560921	0,688129	0,676028 0,694692	0,605789	0,307538 0,392385
2	0,508981	0,460665 0,549813	0,655179	0,636648 0,671940	0,542297	0,218564 0,321409
3	0,489207	0,447804 0,522691	0,593331	0,585449 0,597686	0,453964	0,307538 0,392385
4	0,481637	0,440599 0,515521	0,575739	0,559043 0,588507	0,437375	0,220914 0,321839
5	0,466703	0,430101 0,496445	0,544502	0,533777 0,550836	0,414770	0,307538 0,392385
6	0,458850	0,421741 0,489148	0,533727	0,519737 0,544596	0,408659	0,221289 0,321259
7	0,447644	0,413553 0,474551	0,517868	0,506536 0,524831	0,402182	0,307538 0,392383
8	0,437944	0,403992 0,466583	0,508317	0,495417 0,518734	0,395947	0,221348 0,321947
9	0,429463	0,396353 0,455910	0,500066	0,488094 0,508297	0,393907	0,307536 0,392385
10	0,414935	0,383372 0,441818	0,486090	0,470165 0,497908	0,382182	0,221093 0,320964

La situazione cambia se si considera, come per gli altri indicatori, la $ApEn$ nella sua forma normalizzata (tabella 8.6). In questo caso il livello di disordine più elevato, come già la minor correlazione per la S_ρ e per la mutua informazione, si osserva in corrispondenza di $p = 0.25$ ($ApEn$ normalizzata $\simeq 0.953$), seguito da $p = 0.5$ ($ApEn$ normalizzata $\simeq 0.848$) e da $p = 0.75$ ($ApEn$ normalizzata $\simeq 0.635$).

Tabella 8.6: Valori della $ApEn$ normalizzata e relativi intervalli di confidenza al 95% in funzione della probabilità di transizione p.

Lunghezza della serie = 1000

	Valori di p					
m	0.25		0.50		0.75	
0	1,001027	0,898604 1,096988	0,999115	0,954045 1,041244	1,000921	0,986815 1,011533
1	0,952846	0,865083 1,031932	0,848204	0,833288 0,856293	0,634710	0,322221 0,411118
2	0,936377	0,847490 1,011496	0,807589	0,784747 0,828248	0,568187	0,228999 0,336753
3	0,899999	0,823830 0,961600	0,731353	0,721638 0,736721	0,475637	0,322221 0,411118
4	0,886072	0,810574 0,948408	0,709669	0,689089 0,725408	0,458256	0,231461 0,337204
5	0,858598	0,791261 0,913316	0,671166	0,657945 0,678973	0,434572	0,322221 0,411118
6	0,844151	0,775881 0,899890	0,657884	0,640640 0,671282	0,428169	0,231854 0,336596
7	0,823535	0,760817 0,873036	0,638336	0,624368 0,646919	0,421383	0,322221 0,411116
8	0,805690	0,743228 0,858377	0,626563	0,610662 0,639404	0,414850	0,231915 0,337318
9	0,790087	0,729174 0,838743	0,616393	0,601636 0,626538	0,412713	0,322218 0,411118
10	0,763360	0,705292 0,812817	0,599166	0,579536 0,613733	0,400428	0,231648 0,336288

CAPITOLO 9

Analisi di serie generate da un processo caotico

La mappa logistica $X_{t+1} = rX_t(1-X_t)$, $0 < X_0 < 1$ è una semplice funzione non lineare (in questo caso un polinomio di secondo grado) spesso utilizzata in demografia per esprimere, in un sistema chiuso ed in funzione del tempo t, l'effetto combinato della:

- crescita esponenziale della popolazione (effetto più visibile quando la popolazione è piccola);
- competizione intraspecifica, quando la popolazione è numerosa, che si traduce in una mortalità aggiuntiva dovuta alla competizione degli individui tra loro per assicurarsi il cibo e i necessari mezzi di sostentamento. Tale comportamento è tradotto matematicamente dal termine quadratico di segno negativo.

La mappa logistica è caratterizzata dal parametro r, che rappresenta un tasso combinato tra spinta riproduttiva e necessità di risorse (competizione intraspecifica). Essa è nota anche come esempio di quanto possa essere complesso e "caotico" il comportamento risultante da una semplice equazione dinamica non lineare. In particolare, un sistema si dice caotico se presenta le seguenti caratteristiche:

- forte dipendenza dalle condizioni iniziali, ovvero a partire da input molto simili si possono ottenere, in breve tempo, risultati anche molto

diversi fra loro (a variazioni infinitesime degli ingressi corrispondono variazioni finite in uscita);

- l'andamento reale del sistema non può essere previsto in anticipo a partire da valori iniziali misurati strumentalmente;
- i valori che assume il sistema rimangono tutti confinati entro certi limiti, ovvero il sistema non evolve verso l'infinito per nessun valore iniziale.

In definitiva, il termine caos deterministico indica il comportamento di un sistema la cui evoluzione, pur essendo regolata da leggi precise (deterministiche), risente così tanto da eventuali piccolissime imprecisioni nella determinazione delle condizioni iniziali da risultare, su lunghi periodi, praticamente imprevedibile. Da un punto di vista matematico, un sistema caotico è prevedibile quanto qualunque altro ma, nel mondo reale, tutte le grandezze sono conosciute a meno di un errore. Così, anche le condizioni iniziali non sono mai conosciute perfettamente. In un sistema caotico due condizioni iniziali che, alla luce delle nostre misure, ci sembrano identiche, possono portare a due evoluzioni completamente diverse.

La mappa logistica, in particolare, è caratterizzata da una grande sensibilità alle condizioni iniziali, ovvero, per certi valori di r, due valori iniziali x_0 fra loro anche molto vicini possono divergere nel tempo in maniera esponenziale pressochè imprevedibile restando comunque sempre confinati all'interno dell'intervallo $[0, 1]$. Al variare del parametro r, la mappa logistica è caratterizzata dai seguenti comportamenti:

- per $0 < r \leq 1$, tende a zero indipendentemente dai valori di ingresso;
- per $1 < r < 3$, tende ad un unico valore asintotico $\frac{r-1}{r}$ indipendentemente dai valori di ingresso;
- per $r = 3$, si realizza una biforcazione e la mappa oscilla sempre fra due valori dipendenti dal valore di partenza;
- al crescere del valore di r compaiono sempre nuove biforcazioni e la mappa oscilla tra un numero sempre maggiore di valori asintotici;
- per r approssimativamente pari a 3.57 si ha l'insorgenza del caos, minime variazioni del valore iniziale portano a risultati fra loro anche molto differenti;
- la maggior parte dei valori di r compresi fra $3.57 < r \leq 4$ esibiscono un comportamento caotico, ma ci sono comunque ancora dei valori isolati di r che mostrano comportamenti non caotici detti *isole di stabilità*;
- per $r > 4$ i valori di uscita sono esterni all'intervallo $[0, 1]$ e divergono per quasi tutti i valori iniziali.

9.1 Modalità di esecuzione dell'esperimento

Sono stati realizzati otto gruppi di 5000 simulazioni ciascuno. Ogni gruppo è caratterizzato da uno stesso valore di r, che varia invece da gruppo a gruppo a partire dal valore 3.65 per giungere a 4.00 con incrementi di 0.05. Si è inoltre eseguito un ulteriore gruppo di simulazioni in corrispondenza del valore $r = 3.82847$, all'incirca $1 + \sqrt{8}$, che corrisponde all'inizio di un'isola di stabilità.

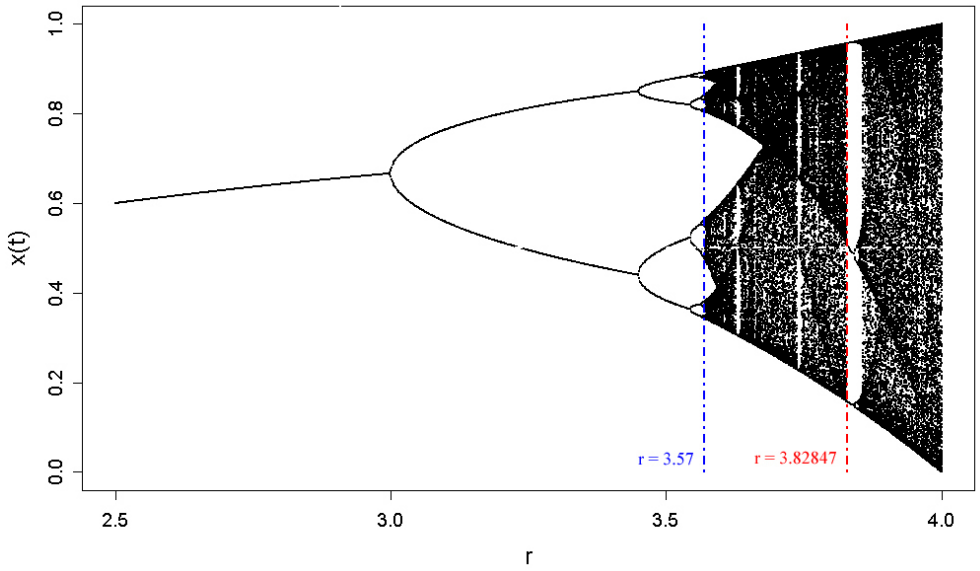

Figura 9.1: Comportamento della mappa logistica non dicotomizzata in funzione del parametro r.

Ogni simulazione ha prodotto una sequenza di 1000 elementi Y_t ottenuti dicotomizzando come segue i risultati della mappa logistica:

$$Y_t = \begin{cases} 0, & 0 \leq X_t \leq 0.5 \\ 1, & 0.5 < X_t \leq 1 \end{cases}$$

$$\text{con } X_{t+1} = rX_t(1 - X_t), \qquad 0 < X_0 < 1$$

Ciascuna sequenza è stata ottenuta a partire da un valore iniziale, x_0, scelto in maniera casuale all'interno dell'intervallo $]0, 1[$, e scartando i primi 100 valori, corrispondenti ad un comportamento transitorio iniziale. Per ogni gruppo è stato poi calcolato il valore della S_ρ normalizzata, della mutua informazione normalizzata e della $ApEn$ (in questo caso solo sulle prime 1000 simulazioni). A partire dai valori della S_ρ, della mutua informazione e della $ApEn$ calcolati sulle singole simulazioni, è stato possibile ottenerne la

distribuzione empirica e, di consequenza, i rispettivi valori medi, la varianza ed i quantili.

9.2 Risultati

9.2.1 Entropia di Granger, Maasoumi e Racine

Dai valori della S_ρ riportati in tabella 9.1 si può osservare come il comportamento dell'indicatore sembri suggerire una qualche forma di struttura regolare per valori del parametro r pari a 3.65, 3.82847, 3.85 e, in parte, anche per $r = 3.70$. In particolare, per $r = 3.65$ si può osservare come il valore della S_ρ relativo al lag 1, 0.2129, si ripeta con cadenza regolare ogni 2 lag. D'altronde, come si può vedere dalla figura 9.2, non si è ancora entrati nella zona caotica. Per $r = 3.70$, in corrispondenza del lag 1 si ha invece un valore della S_ρ pari a 0.1364, quindi più basso del precedente, ma comunque ancora indicativo di un certo livello di autocorrelazione, livello tanto più significativo se si considera la non linearità della S_ρ e il fatto che l'indicatore tende a zero molto rapidamente già a partire da livelli di perturbazione non particolarmente elevati. Tale valore si ripete anche in corrispondenza dei lag 3 e 5, dopo di che tale accenno di regolarità non è più osservabile.

Per $r = 3.82847$, che si situa proprio all'inizio di un'isola di stabilità, si rileva molto chiaramente un andamento periodico di periodo 3. Infatti, in corrispondenza dei lag 3, 6 e 9 si ha un valore della S_ρ costante e pari a circa 0.92, un valore molto elevato e prossimo alla condizione di massima autocorrelazione. Anche i restanti valori, tutti pari a circa 0.25, esprimono un livello di autocorrelazione non trascurabile.

Per $r = 3.85$, valore che si situa all'interno della medesima isola di stabilità, ma in una posizione dall'andamento un pò più articolato e complesso, l'indicatore fornisce ancora una chiara indicazione di struttura, anche se meno forte della precedente. Si possono infatti osservare due valori consecutivi praticamente uguali e relativamente bassi, 0.055, che si ripetono regolarmente ogni 3 lag (lag 1 e 2, 4 e 5, 7 ed 8, 10 ecc.), altri due valori di intensità più che doppia, circa 0.1396, in corrispondenza dei lag 3 e 9 ed infine un valore pari a 0.34 in corrispondenza del lag 6, equidistante rispetto ai due valori $S_\rho = 0.1396$ precedenti.

In corrispondenza dei valori di r pari a 3.90, 3.95 e 4.00 non si osserva, invece, alcuna regolarità apparente. Inoltre, il valore della S_ρ si riduce in maniera evidente e regolare nel passare da $r = 3.90$ ad $r = 4.00$, dove un valore molto prossimo a zero, (circa 0.0004), attesta il comportamento caotico del sistema, nonchè l'indipendenza fra gli elementi della serie osservata. Tale risultato concorda inoltre con l'andamento dell'entropia di Shannon che aumenta fino a raggiungere un valore molto prossimo ad 1 in corrispondenza di $r = 4.00$, nonchè con la percentuale media di uno (o di zero) rilevata nelle serie, percentuale che si approssima sempre più a 0.5, valore imprescindibile

Tabella 9.1: Valori della S_ρ normalizzata e relativi intervalli di confidenza al 95% in funzione del parametro r.

Lunghezza della serie = 1000, numero di casi = 5000

Lag	Valori di r							
	3.65		3.70		3.75		3.80	
0	1	0,975188 1,024179	1	0,955057 1,042883	1	0,971178 1,026866	1	0,971557 1,026162
1	0,212914	0,191195 0,236057	0,136445	0,116998 0,156831	0,196685	0,174682 0,219081	0,193778	0,173031 0,215028
2	0,040166	0,025244 0,057481	0,035480	0,020492 0,052944	0,019752	0,011505 0,029561	0,004028	0,000423 0,009315
3	0,212921	0,191195 0,236057	0,136447	0,116998 0,156831	0,020731	0,011486 0,032012	0,000587	0,000004 0,002240
4	0,066201	0,049016 0,084852	0,023698	0,011997 0,037673	0,005790	0,000322 0,014584	0,001499	0,000011 0,005112
5	0,212927	0,191195 0,236057	0,136447	0,116998 0,156831	0,027348	0,011036 0,047355	0,005650	0,000727 0,012992
6	0,102640	0,082767 0,123631	0,012332	0,003539 0,023979	0,016818	0,004829 0,032756	0,012189	0,003816 0,023054
7	0,212919	0,191195 0,236057	0,003347	0,000039 0,010783	0,010936	0,003193 0,021554	0,001796	0,000016 0,005693
8	0,070504	0,049102 0,093939	0,006167	0,000552 0,014557	0,006500	0,001177 0,013889	0,001060	0,000008 0,003911
9	0,212920	0,191195 0,235514	0,004188	0,000082 0,012006	0,001070	0,000010 0,004166	0,000854	0,000007 0,003242
10	0,118406	0,092921 0,146039	0,002509	0,000025 0,008058	0,008537	0,001322 0,019678	0,000864	0,000007 0,003258

Lag	Valori di r									
	3.82847		3.85		3.90		3.95		4.00	
0	1	0,996098 1,001598	1	0,998591 1,002053	1	0,987698 1,010133	1	0,985867 1,013256	1	0,997636 1,000895
1	0,268121	0,262048 0,269690	0,055397	0,054163 0,056310	0,056018	0,043962 0,069096	0,018437	0,011687 0,026252	0,000419	0,000003 0,001612
2	0,249982	0,150729 0,269690	0,055381	0,054148 0,056556	0,005251	0,001906 0,009323	0,008284	0,002894 0,015106	0,000430	0,000004 0,001578
3	0,928693	0,569537 1,001598	0,139643	0,128248 0,147594	0,008283	0,002895 0,015173	0,006761	0,001768 0,013499	0,000426	0,000004 0,001638
4	0,253376	0,168668 0,269690	0,055367	0,053993 0,056620	0,006502	0,002077 0,011921	0,001282	0,000019 0,003859	0,000429	0,000004 0,001646
5	0,249828	0,152213 0,269690	0,055421	0,054148 0,056556	0,002251	0,000091 0,006041	0,003116	0,000101 0,008240	0,000429	0,000005 0,001567
6	0,924444	0,547026 1,001598	0,340871	0,323337 0,369408	0,000603	0,000006 0,002303	0,002200	0,000060 0,006084	0,000440	0,000006 0,001684
7	0,251681	0,158723 0,269690	0,055363	0,053960 0,056620	0,001606	0,000021 0,004884	0,000502	0,000005 0,001939	0,000429	0,000005 0,001628
8	0,249343	0,147592 0,269690	0,055430	0,053997 0,056658	0,003938	0,000190 0,010243	0,000807	0,000008 0,002976	0,000442	0,000006 0,001654
9	0,921966	0,533230 1,001598	0,139698	0,128248 0,146726	0,001630	0,000020 0,005369	0,000599	0,000006 0,002214	0,000422	0,000007 0,001568
10	0,250891	0,152325 0,269690	0,055342	0,053826 0,056716	0,000558	0,000007 0,002123	0,000591	0,000007 0,002187	0,000443	0,000007 0,001693

per un comportamento che deve approssimare quello di completa casualità (vedi tabella 9.2).

Tabella 9.2: Percentuale di uno ed entropia di Shannon con intervalli di confidenza al 95% in funzione del parametro r.

Lunghezza della serie = 1000, numero di casi = 5000

	Valori di r							
	3.65		3.70		3.75		3.80	
Entropia di Shannon	0,872817	0,852711 0,893173	0,765388	0,731816 0,796699	0,855457	0,831240 0,877593	0,852181	0,829749 0,873832
Proporzione di uno	0,706620	0,690000 0,722000	0,776931	0,759000 0,795000	0,719803	0,703000 0,737000	0,722211	0,706000 0,738000

	Valori di r									
	3.82847		3.85		3.90		3.95		4.00	
Entropia di Shannon	0,917789	0,912870 0,918961	0,991719	0,990272 0,993886	0,985109	0,973242 0,994149	0,966807	0,953694 0,979051	0,999284	0,996463 0,999997
Proporzione di uno	0,667163	0,666000 0,672000	0,446556	0,442000 0,454000	0,570506	0,545000 0,596000	0,606311	0,585000 0,626000	0,499962	0,470000 0,530000

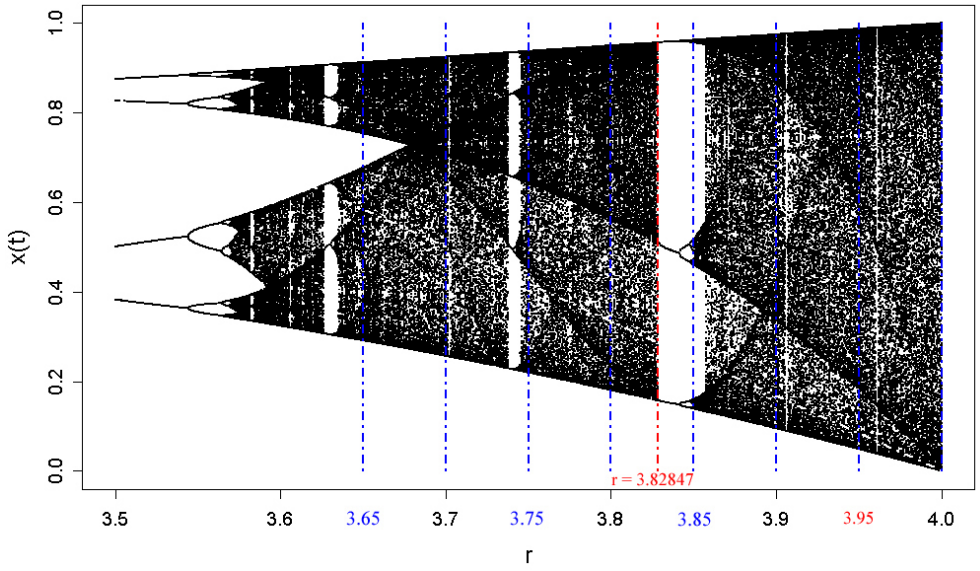

Figura 9.2: Comportamento della mappa logistica non dicotomizzata per valori del parametro r compresi fra 3.5 e 4.0.

Tabella 9.3: Valori della mutua informazione normalizzata e relativi intervalli di confidenza al 95% in funzione del paramtro r.

Lunghezza della serie = 1000, numero di casi = 5000

	Valori di r							
Lag	3.65		3.70		3.75		3.80	
0	1	0,980104 1,019321	1	0,963749 1,034375	1	0,976865 1,021480	1	0,977166 1,020921
1	0,207936	0,184820 0,232799	0,122786	0,104184 0,142489	0,189306	0,166331 0,212899	0,185992	0,164695 0,208299
2	0,066245	0,041896 0,094303	0,058007	0,033604 0,086263	0,032593	0,018980 0,048676	0,006613	0,000689 0,015300
3	0,207947	0,184820 0,232796	0,122788	0,104184 0,142489	0,031820	0,018051 0,048142	0,000943	0,000004 0,003593
4	0,108616	0,081078 0,138560	0,038693	0,019559 0,061449	0,009133	0,000521 0,022717	0,002391	0,000014 0,008101
5	0,207956	0,184820 0,232796	0,122786	0,104184 0,142489	0,045107	0,018251 0,077806	0,009292	0,001185 0,021382
6	0,166733	0,135449 0,199464	0,020052	0,005704 0,039110	0,025907	0,007686 0,049429	0,018949	0,006085 0,035287
7	0,207941	0,184820 0,232796	0,005087	0,000050 0,016232	0,018025	0,005245 0,035552	0,002939	0,000021 0,009363
8	0,115502	0,081143 0,152692	0,009979	0,000860 0,023642	0,010253	0,001891 0,021607	0,001727	0,000007 0,006419
9	0,207942	0,184820 0,232796	0,006341	0,000116 0,017905	0,001712	0,000009 0,006643	0,001367	0,000005 0,005202
10	0,191353	0,151771 0,233523	0,004032	0,000030 0,013037	0,014061	0,002152 0,032562	0,001403	0,000005 0,005298

	Valori di r									
Lag	3.82847		3.85		3.90		3.95		4.00	
0	1	0,996881 1,001277	1	0,998877 1,001636	1	0,990183 1,008073	1	0,988714 1,010566	1	0,998116 1,000713
1	0,273137	0,265918 0,274532	0,091176	0,089259 0,092560	0,091635	0,072060 0,112609	0,030411	0,019339 0,043151	0,000706	0,000003 0,002722
2	0,263382	0,201397 0,274532	0,091149	0,089259 0,092932	0,008833	0,003211 0,015674	0,013748	0,004829 0,024980	0,000723	0,000003 0,002662
3	0,964759	0,761523 1,001277	0,224554	0,207024 0,236763	0,013920	0,004875 0,025458	0,011331	0,002963 0,022623	0,000714	0,000003 0,002766
4	0,265906	0,218280 0,274532	0,091125	0,088948 0,093116	0,010891	0,003491 0,019957	0,002138	0,000026 0,006436	0,000719	0,000003 0,002777
5	0,263402	0,202859 0,274532	0,091212	0,089259 0,092932	0,003784	0,000150 0,010156	0,005188	0,000165 0,013691	0,000717	0,000003 0,002646
6	0,961220	0,740077 1,001277	0,506269	0,483963 0,542116	0,001008	0,000004 0,003867	0,003686	0,000098 0,010202	0,000734	0,000003 0,002837
7	0,264657	0,209296 0,274532	0,091118	0,088948 0,093116	0,002691	0,000026 0,008160	0,000835	0,000004 0,003241	0,000714	0,000003 0,002747
8	0,262967	0,198436 0,274532	0,091224	0,088990 0,093116	0,006617	0,000318 0,017210	0,001343	0,000006 0,004966	0,000734	0,000003 0,002778
9	0,959051	0,726156 1,001277	0,224636	0,207024 0,235294	0,002728	0,000025 0,009000	0,000998	0,000004 0,003708	0,000701	0,000003 0,002637
10	0,263966	0,203047 0,274532	0,091082	0,088666 0,093311	0,000929	0,000003 0,003562	0,000983	0,000005 0,003666	0,000733	0,000003 0,002847

9.2.2 Mutua informazione

Il comportamento della mutua informazione (vedi precedente tabella 9.3) è perfettamente sovrapponibile a quello della S_ρ.

Si notano tuttavia valori normalizzati generalmente più elevati, in particolare per $r = 3.85$, dove la mutua informazione registra, in corrispondenza del lag 6, un valore massimo pari a 0.506 contro lo 0.34 della S_ρ e, in corrispondenza dei lag 3 e 9, un valore di 0.22 contro i circa 0.14 della S_ρ.

9.2.3 Approximate Entropy

Dai valori di tabella 9.4 si può vedere come la *ApEn* fornisca indicazioni generalmente in accordo con quelle fornite sia dalla S_ρ che dalla mutua informazione.

Si può osservare come, per $m = 1$, i valori della *ApEn* per r pari a 3.65, 3.75 e 3.80 siano sostanzialmente uguali e pari a circa 0.69, mentre in corrispondenza di $r = 3.70$ si osserva un valore della *ApEn* significativamente più basso e pari a circa 0.67. Se per $r = 3.70$ tale comportamento può essere imputato all'accenno di struttura e, quindi, di regolarità messa in evidenza dalla S_ρ e dalla mutua informazione, meno chiaro appare invece il comportamento della *ApEn* in corrispondenza di $r = 3.65$, dove la presunta condizione di maggiore regolarità rilevata dagli altri due indicatori viene qui evidenziata solo a partire da $m = 3$.

Per il valore di soglia $r = 3.82847$ $(1 + \sqrt{8})$ si registra, per $m = 1$, un valore della *ApEn* prossimo a 0.67, valore che scende a circa 0.02 per $m = 2$ ed assume valori via via ancora più bassi al crescere di m, mettendo così in luce una struttura fortemente regolare e periodica di periodo ≥ 2, come per altro già evidenziato dalla S_ρ e dalla mutua informazione.

In corrispondenza degli altri valori di r, 3.85, 3.90, 3.95 e 4.00, si osservano invece, per $m = 1$, valori della *ApEn* generalmente ≥ 0.90 (0.90, 0.94, 0.9999). Ciò starebbe ad indicare, per $r = 3.90, 3.95$ e 4.00 una condizione di sostanziale (perfetta in corrispondenza di $r = 4.00$) caoticità, condizione confermata dal fatto che i valori della *ApEn* si mantengono sempre molto elevati anche per i livelli di m successivi al primo. Un discorso a parte va fatto invece per $r = 3.85$ la cui *ApEn*, pur partendo da un valore pari a 0.90 per $m = 1$, scende prima a 0.62 per $m = 2$ e poi a 0.22 per m compreso tra 3 ed 8 ad indicare una certa regolarità di periodo 3 o superiore, in accordo a quanto già evidenziato dagli altri indicatori.

Tabella 9.4: Valori della $ApEn$ e relativi intervalli di confidenza al 95% in funzione del paramtro r.

Lunghezza della serie = 1000, numero di casi = 1000

m	Valori di r							
	3.65		3.70		3.75		3.80	
0	0,873657	0,853278 0,893027	0,765323	0,734558 0,794804	0,855942	0,833055 0,879655	0,853303	0,831536 0,873330
1	0,692012	0,686047 0,694936	0,671688	0,656081 0,684139	0,694255	0,691105 0,695614	0,694413	0,692213 0,695620
2	0,681766	0,667368 0,691523	0,651669	0,638488 0,663980	0,692259	0,687356 0,695160	0,690669	0,683469 0,694650
3	0,556112	0,541249 0,569167	0,594658	0,589018 0,598035	0,681720	0,671158 0,689844	0,690036	0,682227 0,694261
4	0,553250	0,535296 0,567933	0,592383	0,586344 0,596596	0,639562	0,613188 0,658459	0,682917	0,669874 0,691895
5	0,434980	0,425954 0,439744	0,541818	0,532114 0,549162	0,585134	0,552739 0,609672	0,675076	0,659512 0,687171
6	0,420348	0,405874 0,431152	0,526344	0,505826 0,541546	0,576556	0,542554 0,602726	0,629013	0,602799 0,650554
7	0,420336	0,405847 0,431149	0,521735	0,499776 0,538317	0,549054	0,514754 0,577836	0,621094	0,594238 0,643393
8	0,396911	0,381023 0,410222	0,514886	0,492282 0,532887	0,538005	0,502407 0,567643	0,611745	0,583651 0,635575
9	0,396909	0,381146 0,410150	0,512650	0,488790 0,531021	0,514425	0,475653 0,546688	0,595275	0,566064 0,621776
10	0,374058	0,356043 0,389432	0,499177	0,474483 0,519832	0,481044	0,438983 0,515084	0,568877	0,534297 0,597382

m	Valori di r									
	3.82847		3.85		3.90		3.95		4.00	
0	0,918792	0,914098 0,919225	0,992747	0,991423 0,995031	0,985991	0,973685 0,994762	0,967537	0,953689 0,980101	1,000300	0,997390 1,001009
1	0,667749	0,667339 0,672190	0,902280	0,899490 0,907068	0,895780	0,875045 0,912731	0,938047	0,924118 0,950256	0,999631	0,996093 1,000970
2	0,021936	0,000000 0,188687	0,616904	0,608705 0,631073	0,842178	0,821224 0,858248	0,874658	0,866314 0,879457	0,998241	0,993667 1,000675
3	0,017074	0,000000 0,147813	0,219862	0,209217 0,224346	0,756670	0,739195 0,771647	0,870771	0,860032 0,877728	0,995300	0,988820 0,999261
4	0,015322	0,000000 0,132410	0,219871	0,209217 0,224341	0,753155	0,734762 0,768556	0,862906	0,848445 0,873180	0,989420	0,980760 0,995727
5	0,012203	0,000000 0,114718	0,219868	0,209217 0,224336	0,741463	0,717103 0,761255	0,834765	0,814692 0,850501	0,977366	0,964094 0,987426
6	0,010400	0,000000 0,099293	0,219554	0,206818 0,224223	0,730446	0,703716 0,752869	0,819337	0,796752 0,838056	0,952748	0,935607 0,967990
7	0,009076	0,000000 0,090576	0,219572	0,207131 0,224225	0,709620	0,677326 0,736811	0,795537	0,767117 0,816949	0,897607	0,869963 0,920992
8	0,007527	0,000000 0,080252	0,219582	0,207130 0,224225	0,677247	0,640084 0,709330	0,746364	0,714075 0,776467	0,773708	0,740236 0,804984
9	0,006270	0,000000 0,064062	0,112678	0,108964 0,115729	0,616375	0,575084 0,652099	0,653403	0,619453 0,686382	0,572991	0,540365 0,603847
10	0,005249	0,000000 0,054412	0,112661	0,108965 0,115799	0,539603	0,498292 0,574412	0,520122	0,488295 0,549021	0,364909	0,335101 0,395077

9.3 Analisi di isole di stabilità

Al fine di meglio valutare il comportamento della S_ρ, della mutua informazione e della $ApEn$ in corrispondenza di intervalli di periodicità, è stato realizzato un'altro set di simulazioni con riferimento all'isola di stabilità che inizia in prossimità di $r = 3.82847$ (vedi figura 9.3) ed in particolare:

- una serie di simulazioni volte a studiare il comportamento nella zona di "salto" compresa nell'intervallo $3.82847 < r < 3.85$. In tale intervallo, infatti, in corrispondenza del lag 1 la S_ρ passa da circa 0.27 a 0.06, la mutua informazione da 0.27 a 0.09 e la $ApEn$ da 0.67 a 0.90. La zona di salto è stato alla fine circoscritta fra $3.844568 < r < 3.844569$;
- in corrispondenza del valore r = 3.828, che si situa subito prima dell'inizio dell'isola di stabilità considerata;
- in corrispondenza del valore r = 3.8565, che si situa poco prima della fine dell'isola di stabilità;
- in corrispondenza del valore r = 3.8575, che si situa subito al di la dell'isola di stabilità.

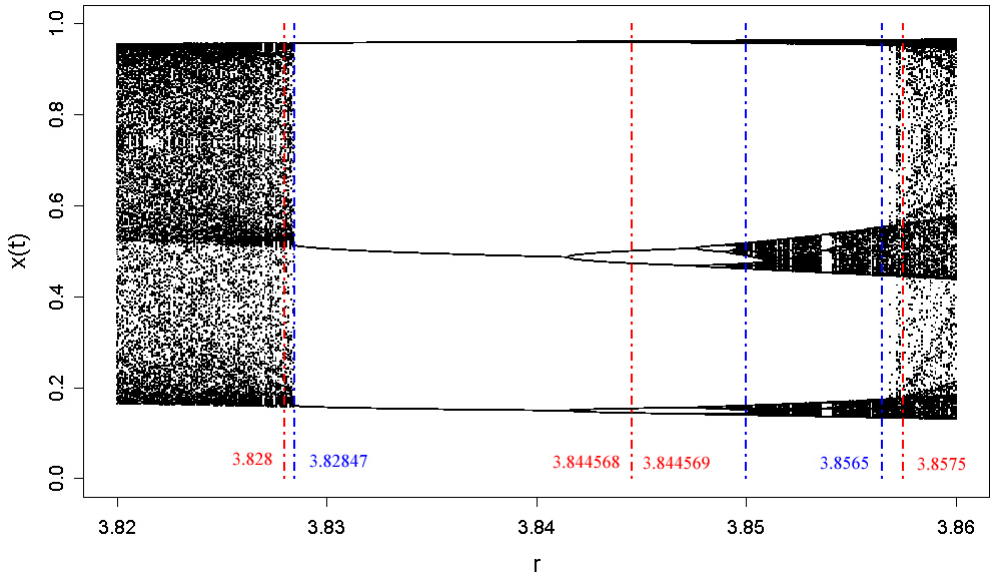

Figura 9.3: Isola di stabilità $r = 3.82847$.

Dai risultati delle simulazioni effettuate e riportati in tabella 9.5, si può vedere come gli indicatori considerati, specie se valutati assieme, riescano a caratterizzare in maniera abbastanza soddisfacente il comportamento della

Tabella 9.5: Valori della S_ρ,della mutua informazione e della $ApEn$ relativi all'isola di stabilità $r = 3.82847$.

% di uno

Valori di r						
3.828	3.82847	3.844568	3.844569	3.85	3.8565	3.8575
0.684169	0,667163	0.333380	0.499933	0,446556	0.483821	0.553939

Sp normalizzato

	Valori di r						
Lag	3.828	3.82847	3.844568	3.844569	3.85	3.8565	3.8575
0	1	1	1	1	1	1	1
1	0,243027	0,268121	0,268638	0,049184	0,055397	0,049800	0,043675
2	0,025866	0,249982	0,268626	0,049141	0,055381	0,049800	0,020940
3	0,141233	0,928693	0,999450	0,049147	0,139643	0,200446	0,150971
4	0,048404	0,253376	0,268619	0,049187	0,055367	0,049831	0,034156
5	0,024602	0,249828	0,268644	0,049147	0,055421	0,049802	0,015163
6	0,100286	0,924444	0,999446	0,998131	0,340871	0,213750	0,118706
7	0,032143	0,251681	0,268629	0,049193	0,055363	0,049834	0,030693
8	0,018930	0,249343	0,268656	0,049130	0,055430	0,049804	0,010846
9	0,073688	0,921966	0,999477	0,049260	0,139698	0,210154	0,091250
10	0,022817	0,250891	0,268641	0,049199	0,055342	0,049828	0,027776

Mutua Informazione normalizzata

	Valori di r						
Lag	3.828	3.82847	3.844568	3.844569	3.85	3.8565	3.8575
0	1	1	1	1	1	1	1
1	0,243177	0,273137	0,273915	0,081726	0,091176	0,082712	0,072195
2	0,040404	0,263382	0,273886	0,081655	0,091149	0,082669	0,034958
3	0,225906	0,964759	0,999803	0,081874	0,224554	0,314987	0,241551
4	0,073373	0,265906	0,273894	0,081730	0,091125	0,082718	0,056692
5	0,038546	0,263402	0,273902	0,081665	0,091212	0,082670	0,025373
6	0,162652	0,961220	0,999801	0,999513	0,506269	0,334160	0,192109
7	0,049845	0,264657	0,273902	0,081740	0,091118	0,082722	0,051016
8	0,029897	0,262967	0,273910	0,081637	0,091224	0,082673	0,018177
9	0,120530	0,959051	0,999805	0,081849	0,224636	0,328973	0,149071
10	0,035845	0,263966	0,273912	0,081750	0,091082	0,082711	0,046217

ApEn

	Valori di r						
m	**3.828**	3.82847	3.844568	3.844569	3.85	3.8565	3.8575
0	0,900794	0,918792	0,919234	1,001020	0,992747	1,000075	0,991760
1	0,681404	0,667749	0,667349	0,919238	0,902280	0,917342	0,920393
2	0,523576	0,021936	0	0,667456	0,616904	0,661763	0,806210
3	0,488900	0,017074	0	0,001158	0,219862	0,327696	0,564737
4	0,477935	0,015322	0	0,001043	0,219871	0,327686	0,558519
5	0,458547	0,012203	0	0,000948	0,219868	0,327680	0,520083
6	0,446564	0,010400	0	0,000139	0,219554	0,323742	0,493147
7	0,438134	0,009076	0	0,000104	0,219572	0,323739	0,476622
8	0,423202	0,007527	0	0,000096	0,219582	0,323736	0,442385
9	0,402351	0,006270	0	0,000078	0,112678	0,309455	0,414898
10	0,374298	0,005249	0	0,000073	0,112661	0,309449	0,385294

mappa logistica in presenza di un'isola di stabilità come quella considerata. Infatti, per $r = 3.828$ ed $r = 3.8575$, che si situano subito prima l'inizio e subito dopo la fine dell'isola di stabilità, la S_ρ e la mutua informazione non evidenziano chiari segni di struttura, anche se, dall'analisi comparata con i dati relativi alle altre condizioni simulate (r che va da 3.82847 a 3.8565) si potrebbe forse già intravvedere un accenno del comportamento che invece caratterizzerà gli altri valori di r. D'altronde, osservando la figura 9.3, si può osservare come, in corrispondenza di tali valori di r, i punti sul grafico non si distribuiscano in maniera uniforme, ma si rilevino delle zone più dense (le zone più scure), dove i punti si addensano con maggiore frequenza e, quindi, con maggiore probabilità, caratterizzando così in qualche modo il comportamento della mappa logistica.

I livelli di autocorrelazione osservati non risultano, in ogni caso, particolarmente elevati (sempre ≤ 0.15), se si esclude lo 0.26 per $r = 3.828$ rilevato in corrispondenza del lag 1, e comunque sempre notevolmente più bassi rispetto a quelli riscontrati in corrispondenza degli altri valori di r.

La $ApEn$ attribuisce a $r = 3.828$ ed $r = 3.8575$, ed indipendentemente dai valori di m considerati, il grado di disordine più elevato fra i gruppi di simulazioni considerati, classificando inoltre la condizione $r = 3.828$ come maggiormente "random" rispetto a $r = 3.8575$, risultato non in contrasto con i dati forniti dalla S_ρ e dalla mutua informazione. Continuando nell'analisi, la S_ρ e la mutua informazione mostrano lo stesso comportamento per $r = 3.82847$ e $r = 3.844568$, identificando per entrambi i livelli una struttura quasi perfettamente periodica di periodo 3. Infatti, in corrispondenza dei lag 3, 6 e 9 osserviamo valori della mutua informazione e della S_ρ pari a circa 0.96 e 0.92 per $r = 3.82847$ e di 0.9998 e 0.9994 per $r = 3.844568$, valori assai prossimi (per $r = 3.844568$ praticamente coincidenti) con quelli di massima autocorrelazione. Anche i valori riferiti ai rimanenti lag mostrano livelli di dipendenza relativamente elevati, (intorno a 0.26) e, soprattutto, costanti.

I risultati di cui sopra trovano conferma anche nella $ApEn$, il cui valore si azzera completamente già a partire da $m = 2$ per $r = 3.844568$, mentre assume un valore pari a 0.022 per $r = 3.82847$.

Nel passaggio da $r = 3.844568$ a $r = 3.844569$, la S_ρ e la mutua informazione mettono invece in risalto un sostanziale mutamento nel comportamento della mappa logistica dicotomizzata. Si passa infatti da una perfetta regolarità di periodo 3 ad un'altrettanto quasi perfetta regolarità di periodo 6 (lag 6, mutua informazione = 0.9995 ed $S_\rho = 0.998$) accompagnata da un livello di autocorrelazione molto basso e, soprattutto, costante rilevato in corrispondenza di tutti gli altri lag (mutua informazione = 0.082 ed $S_\rho = 0.049$).

Con $r = 3.82847, 3.844568$ e 3.8844569 abbiamo esplorato la parte iniziale e più semplice dell'isola di periodicità, caratterizzata, come si può ben vedere in figura 9.3, prima da 3 e poi da 6 biforcazioni. Con $r = 3.85$ ed $r = 3.8565$, che si situano, rispettivamente, in una zona dove le biforcazioni iniziano ad

aumentare sensibilmente ed in prossimità della fine dell'isola di stabilità, si vuole invece analizzare la seconda e più complessa parte dell'isola stessa.

Per $r = 3.85$ si osserva chiaramente una forma di struttura ben definita, con un valore massimo (mutua informazione ed S_ρ pari rispettivamente a 0.506 e 0.341) abbastanza elevato in corrispondenza del lag 6, due valori ancora significativamente elevati, uguali e simmetrici rispetto al valore massimo (lag 9 e 3) e valori molto bassi e costanti per i rimanenti lag (mutua informazione = 0.091 ed $S_\rho = 0.049$). Tali risultati non sono in contrasto con quelli forniti dalla $ApEn$ che mostra, per $m = 1$, un livello di disordine molto elevato e pari a 0.9022, valore che si riduce a 0.6169 per $m = 2$ per scendere a circa 0.22 per m da 3 a 8 evidenziando così una regolarità di livello superiore abbastanza elevata e che trova conferma nella perfetta e prolungata coincidenza del valore 0.22.

Per $r = 3.8565$, infine, si osserva ancora un comportamento che non può che definirsi regolare: in corrispondenza dei lag 3, 6 e 9 si osservano infatti valori abbastanza simili, ancorchè non uguali (mutua informazione pari rispettivamente a 0.315, 0.334 e 0.329) ed indicanti un livello significativo di autocorrelazione, mentre per i restanti lag si hanno valori molto più bassi e costanti (mutua informazione pari a circa 0.0827). Anche la $ApEn$, a fronte di un valore iniziale pari a circa 0.92, scende a 0.66 in corrispondenza di $m = 2$ per stabilizzarsi su 0.328 per livelli di m che vanno da 3 a 8, indicando con ciò una sicura regolarità di comportamento, anche se leggermente inferiore rispetto al livello di r precedente.

A titolo di conferma dei risultati ottenuti, si è analizzato il comportamento degli stessi indicatori applicandoli ad un'altra isola di stabilità (vedi figura 9.4), i cui comportamento è sintetizzato nella tabella 9.6.

Dai risultati ottenuti si può vedere come, per valori di r esterni all'isola di stabilità ($r = 3.73$ ed $r = 3.75$), sia la mutua informazione che la S_ρ non mostrino un comportamento tale da far pensare ad una qualche forma di regolarità. Essi inoltre assumono valori generalmente abbastanza bassi ed indicanti una scarsa o scarsissima autocorrelazione. Analogamente, la $ApEn$ presenta valori abbastanza alti (rispettivamente 0.68 e 0.69 per $m = 1$ ed ancora 0.58 e 0.59 per $m = 5$) ed abbastanza prossimi fra loro per tutti i livelli di m, indicando in questo un grado di disordine relativamente elevato.

Per $r = 3.741$, valore situato nella porzione iniziale dell'isola di stabilità, osserviamo un comportamento chiaramente regolare e periodico di periodo 5. In corrispondenza dei lag 5 e 10, infatti, la mutua informazione assume un valore pari a circa 0.99 (0.977 per la S_ρ) e quindi praticamente coincidente a quello di autocorrelazione massima. Si osservano inoltre valori praticamente costanti ed abbastanza elevati che si ripetono ad intervalli regolari (lag 1, 4, 6 e 9, mutua informazione pari a circa 0.43) e valori altrettanto costanti ma molto più bassi che si ripetono anch'essi in modo regolare (rimanenti lag 2, 3, 7 ed 8, mutua informazione pari a circa 0.02). La $ApEn$, invece, parte da

Tabella 9.6: Valori della S_ρ, della mutua informazione e della *ApEn* relativi all'isola di stabilità $r = 3.738$ - $r = 3.745$.

% di uno

Valori di r				
3.73	3.741	3.742	3.743	3.75
0.766280	0.600664	0.699753	0.663750	0,719803

Sp normalizzato

	Valori di r				
Lag	3.73	3.741	3.742	3.743	3.75
0	1	1	1	1	1
1	0,146516	0,395296	0,221452	0,273474	0,196685
2	0,019771	0,012317	0,001173	0,005043	0,019752
3	0,146513	0,012131	0,001119	0,004999	0,020731
4	0,006219	0,389387	0,218148	0,271268	0,005790
5	0,000995	0,976710	0,135747	0,271685	0,027348
6	0,001314	0,390468	0,218728	0,271634	0,016818
7	0,002006	0,012215	0,001153	0,005018	0,010936
8	0,001458	0,012246	0,001131	0,005028	0,006500
9	0,000919	0,390197	0,218499	0,271409	0,001070
10	0,002832	0,976751	0,977372	0,414762	0,008537

Mutua Informazione normalizzata

	Valori di r				
Lag	3.73	3.741	3.742	3.743	3.75
0	1	1	1	1	1
1	0,133576	0,430673	0,217827	0,279586	0,189306
2	0,032325	0,020649	0,001930	0,008392	0,032593
3	0,133570	0,020337	0,001842	0,008319	0,031820
4	0,010102	0,428068	0,216289	0,278658	0,009133
5	0,001565	0,989997	0,218341	0,415534	0,045107
6	0,002108	0,428670	0,216567	0,278851	0,025907
7	0,003239	0,020478	0,001896	0,008350	0,018025
8	0,002256	0,020531	0,001861	0,008366	0,010253
9	0,001443	0,428755	0,216551	0,278808	0,001712
10	0,004349	0,990047	0,991363	0,597201	0,014061

ApEn

	Valori di r				
m	3.73	3.741	3.742	3.743	3.75
0	0,784743	0,971503	0,882471	0,922079	0,855942
1	0,680010	0,553088	0,690223	0,664333	0,694255
2	0,673262	0,406308	0,676168	0,643996	0,692259
3	0,592979	0,404317	0,676175	0,636222	0,681720
4	0,581688	0,007899	0,402522	0,362877	0,639562
5	0,579520	0,004058	0,007646	0,138242	0,585134
6	0,573418	0,003606	0,006967	0,137787	0,576556
7	0,567855	0,002386	0,005854	0,137146	0,549054
8	0,561195	0,001362	0,005166	0,136817	0,538005
9	0,551617	0,001088	0,004636	0,136516	0,514425
10	0,528902	0,000860	0,001451	0,135523	0,481044

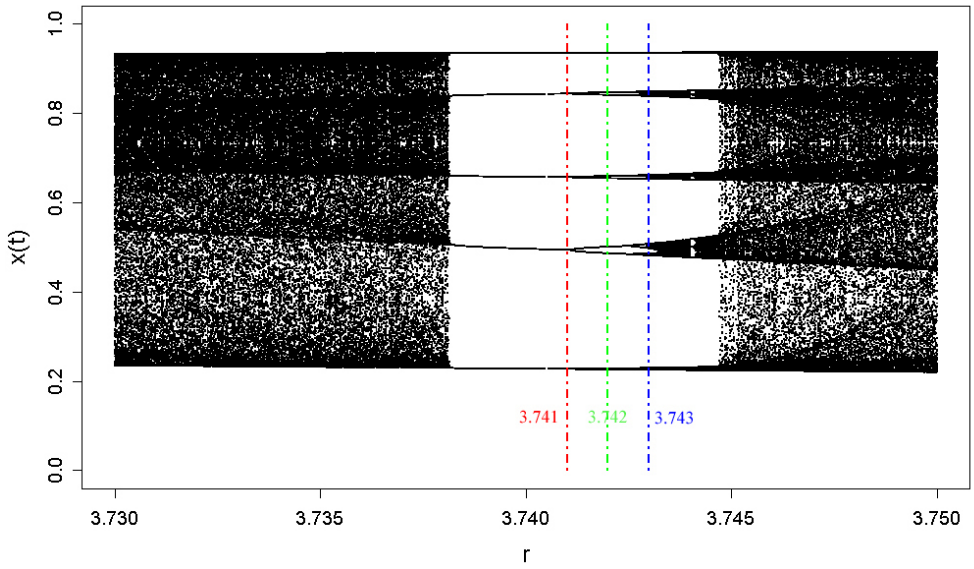

Figura 9.4: Isola di stabilità $r = 3.738$ - $r = 3.745$.

0.553, scende a poco più di 0.4 in corrispondenza di $m = 2$ ed $m = 3$ e si attesta su un valore molto prossimo a zero per tutti i successivi livelli di m.

Per $r = 3.742$, valore che si trova in corrispondenza di un numero di biforcazioni relativamente più elevato, si nota un comportamento simile alla condizione precedente. Anche in questo caso, infatti, i valori della mutua informazione e della S_ρ lasciano chiaramente intravvedere una sottostante struttura periodica anche se, questa volta, di periodo 10 (lag 10, mutua informazione $= 0.991$, $S_\rho = 0.977$). Si ha inoltre la ripetizione, ad intervalli regolari (gli stessi del caso precedente), di valori costanti e relativamente elevati (mutua informazione pari a circa 0.217 in corrispondenza dei lag 1, 4, 6, 9 e, in questo caso, anche del lag 5) alternati a valori altrettanto costanti ma decisamente più bassi e molti prossimi alla condizione di indipendenza (lag 2, 3, 7, ed 8, mutua informazione pari a circa 0.0019). La $ApEn$ parte, in questo caso, da 0.69, si mantiene su valori intorno a 0.676 per $m = 2$ ed $m = 3$, cala a 0.403 per $m = 3$ e poi praticamente si azzera a partire da $m = 5$.

La condizione $r = 3.742$ presenta quindi anch'essa una struttura ben identificabile ma di periodo più lungo e leggermente meno regolare rispetto a quella osservabile per $r = 3.741$.

In corrispondenza di $r = 3.743$, infine, pur senza individuare un comportamento perfettamente periodico, si può ancora osservare una struttura ben definita caratterizzata, anche questa volta, da valori costanti e relativamente elevati che si ripetono ad intervalli regolari con lo stesso disegno di massima

dei due casi precendeti (lag 1, 4, 6 e 9, valori della mutua informazione pari a circa 0.279) alternati a valori altrettanto costanti ma decisamente più bassi e tendenti a zero (lag 2, 3, 7 ed 8, mutua informazione pari a circa 0.0083). Si osservano poi due valori decisamente più elevati (mutua informazione pari a 0.416 e 0.597) in corrispondenza dei lag 5 e 10, lag che caratterizzavano il periodo delle due condizioni precedenti. In questo caso la $ApEn$ parte da 0.66, scende a circa 0.64 per $m = 2$ ed $m = 3$, a 0.362 per $m = 4$ per poi stabilizzarsi su un valore prossimo a 0.13 per i successivi livelli di m, lasciando intravvedere anch'essa un comportamento regolare anche se di intensità inferiore rispetto ai due casi precedenti.

Ricollegandoci ai dati più generali riportati nelle tabelle 9.1, 9.3 e 9.4, si può quindi affermare che la S_ρ, la mutua informazione e, in misura minore, la $ApEn$, appaiono in grado di descrivere il comportamento della mappa logistica dicotomizzata in relazione al livello di dipendenza e di regolarità che essa esprime. Tali indicatori evidenziano infatti un aumento del livello di irregolarità all'aumentare di r riuscendo inoltre a caratterizzare con chiarezza le isole di stabilità, mettendo in luce una qualche regolarità di struttura anche nei tratti in cui le biforcazioni sono più numerose e il sistema, di conseguenza, più confuso.

CAPITOLO 10

Considerazioni finali

Come si è potuto vedere nel corso di questa esposizione, la S_ρ, la mutua informazione, l'entropia congiunta e la *ApEn* sono indicatori utilizzabili per lo studio del livello di autocorrelazione e di regolarità di una serie, mentre la distanza di Kullback Leibner, o entropia relativa, non è risultata di nessuna utilità in quanto, per costruzione, essa tende ad assumere sempre valori pari a zero quando le serie poste a confronto sono caratterizzate, come in questo caso, dalla medesima distribuzione di probabilità.

Si è visto inoltre come la mutua informazione e l'entropia congiunta, nella loro forma normalizzata, corrispondano ad una stessa misura. Tuttavia la mutua informazione presenta varianza minore e quindi, a parità di altre condizioni, è sicuramente da preferirsi.

Si è potuto osservare, poi, come la mutua informazione e la S_ρ abbiano un comportamento molto simile, differendo sostanzialmente solo per un fattore di scala. La mutua informazione tende però ad assumere valori più alti, in particolar modo in corrispondenza dei livelli di irregolarità più elevati e tende a zero in maniera più graduale rispetto alla S_ρ. Ciò ci consente di graduare maggiormente il comportamento delle serie più irregolari. La S_ρ presenta invece, in genere, livelli di varianza leggermente inferiori. Comunque, ai fini pratici, i due indicatori appaiono sostanzialmente equivalenti per cui, se del caso, il ricercatore potrà utilizzare quello a lui più familiare.

Si è visto poi che, in genere, i risultati ottenuti con la S_ρ e con la mutua informazione concordano con quelli forniti dalla *ApEn*, e non potrebbe essere altrimenti in quanto orientati, in linea di massima, alla misura di un medesimo fenomeno.

Dai risultati ottenuti appare chiaro, tuttavia, come né la S_ρ, né la mutua informazione né, tanto meno, la $ApEn$ possano essere considerati indicatori universali in quanto essi non sono in grado di fornire indicazioni utili ed univoche in tutte le possibili situazioni. Essi dovrebbero essere pertanto utilizzati con molta prudenza e valutandone sempre limiti e caratteristiche.

La S_ρ e la mutua informazione sono sicuramente più indicati quali indicatori di indipendenza e di correlazione, ma possono fornire utili informazioni anche nei riguardi dell'eventuale presenza di una qualche forma di struttura. La $ApEn$, invece, è sostanzialmente un indicatore del livello di disordine e misura l'autocorrelazione solo indirettamente, tramite il livello di regolarità della serie e le modalità con cui un determinato disegno, o pattern, si ripete nella sequenza dei dati.

Tutti e tre gli indicatori sono in grado di identificare con precisione strutture periodiche o quasi periodiche purchè il numero di lag per la S_ρ e la mutua informazione, o la dimensione del parametro m per la $ApEn$, siano sufficientemente elevati. Questa è una caratteristica peculiare di tali indicatori, oltre che uno dei loro limiti principali. Un'eventuale periodicità viene infatti rilevata solo in corrispondenza di valori del lag o di m multipli del periodo della serie. Per la $ApEn$ questo risultato si può avere anche in prossimità di tale multiplo ed è caratterizzato dalla persistenza a zero dell'indicatore per tutti i livelli di m successivi. Tuttavia, mentre per la S_ρ e per la mutua informazione la relazione fra il numero di lag considerati e la lunghezza della serie necessaria al fine di ottenere stime attendibili è lineare, per la $ApEn$ tale relazione è di tipo esponenziale. Cio comporta che, in presenza di serie di lunghezza limitata, ad esempio composte da qualche migliaio di elementi, i valori di m che possono essere considerati senza tema di incorrere in stime completamente fuorvianti, si riducono ai primi due, al massimo ai primi tre. Tale limite ovviamente non esiste per la S_ρ e la mutua informazione.

Ciò non di meno, potrebbe essere comunque utile considerare un numero maggiore di m (nel presente studio si è considerato $1 \leq m \leq 10$). Questo perchè un'eventuale regolarità di comportamento dei valori della $ApEn$ al variare di m (ad esempio la persistenza di uno stesso valore o di blocchi di valori uguali), può darci informazioni riguardo all'eventuale struttura della serie, ancorchè non perfettamente periodica oppure di periodo elevato.

Ovviamente solo un valore della S_ρ e della mutua informazione normalizzata prossimo ad uno, oppure della $ApEn$ prossimo a zero sono sinonimi di perfetta regolarità e massima autocorrelazione. Tuttavia la mancanza di un tale riscontro non può e non deve escludere, se del caso, la ricerca di forme di struttura diverse, più complesse o di periodo più lungo.

Si è visto poi come gli indicatori in questione, soprattutto la S_ρ e la mutua informazione, siano particolarmente efficaci nell'analisi della serie logistica dicotomizzata (vedi capitolo 9). Non appare invece particolarmente soddisfacente la caratterizzazione di serie periodiche in funzione del grado di rumore indotto (vedi capitolo 5).

Gli indicatori analizzati, infine, non sono assolutamente adatti a fornire indicazioni riguardo la complessità delle serie studiate. Uno stesso livello della S_ρ, della mutua informazione o della $ApEn$ può riferirsi, infatti, a serie anche molto diverse fra loro dal punto di vista della complessità algoritmica.

Più che per misure assolute, gli indicatori studiati, ed in particolare la $ApEn$, parrebbero maggiormente indicati, pur con tutte le cautele del caso, nella caratterizzazioni di serie diverse ma afferenti ad uno stesso fenomeno, variabile nel tempo o nello spazio. Sarebbe inoltre opportuno, per quanto possibile, procedere sempre al calcolo congiunto della mutua informazione (o della S_ρ) e della $ApEn$, unitamente ad altri indicatori più specificatamente legati alla materia oggetto di studio, in modo da avere una conferma dei risultati ottenuti ed un aiuto nell'interpretazione degli stessi, cosa il più delle volte tutt'altro che banale.

Un'attenzione particolare andrebbe poi prestata in tutte quelle situazioni in cui si debbano confrontare fra loro numeri molto piccoli che, se pur presentando differenze a volte anche di un ordine di grandezza, risultano comunque molto prossimi allo zero e di intensità assolutamente marginale rispetto all'intervallo di variazione, $[0, 1]$, degli indicatori considerati. In tali situazioni, quand'anche il confronto dovesse risultare statisticamente significativo, ad una meccanica applicazione di tecniche dovrà preferirsi un approccio critico che tenga conto, volta per volta, delle caratteristiche del fenomeno trattato e delle relative condizioni al contorno.

Parte III

Analisi della correlazione e del livello di disordine fra serie binarie

CAPITOLO 11

Introduzione

Scopo di questa terza parte della tesi è quello di effettuare alcuni esperimenti con gli indicatori precedentemente considerati, ovvero con:

- la misura di Granger, Maasoumi e Racine, indicata come S_ρ;
- la mutua informazione;
- l'entropia relativa, anche detta distanza di Kullback Leibler;
- l'entropia approssimata fra serie diverse, anche detta cross Approximated Entropy ed indicata con *cross-ApEn*;

al fine di studiare la correlazione ed il livello di irregolarità riscontrabile fra serie binarie diverse fra loro ed appositamente create allo scopo.

In particolare, in analogia a quanto già sperimentato nel precedente capitolo 5, sono stati generati, in maniera del tutto indipendente l'uno dall'altro, due gruppi di serie perfettamente periodiche a cui sono stati successivamente applicati diversi livelli di perturbazione casuale. In particolare ed in maniera indipendente per i due gruppi, sono stati di volta in volta sostituiti il 25%, 50%, 75% e 100% degli elementi delle serie originarie con uno qualsiasi dei due simboli "0" oppure "1" estratti in maniera casuale in base ad una probabilità data.

L'esperimento è quindi consistito nel confrontare le serie del primo gruppo, dette serie di confronto e caratterizzate, di volta in volta, da un determinato livello di perturbazione, con le serie del secondo gruppo, dette serie

obiettivo, a cui saranno successivamente applicati tutti i livelli di perturbazione a partire dalla condizione di perfetta regolarità (livello di perturbazione uguale a zero) fino a quella di massima casualità (livello di perturbazione pari al 100%). Operando in questo modo si sono ottenuti venticinque diversi gruppi di confronti fra loro indipendenti (cinque livelli di perturbazione per le serie di confronto associati ai cinque livelli di perturbazione previsti per le serie obiettivo). Unicamente per una questione di semplicità, si è convenuto di confrontare la prima serie del primo gruppo con la prima serie del secondo gruppo e così via fino al completo svuotamento dei due gruppi (figura 11.1).

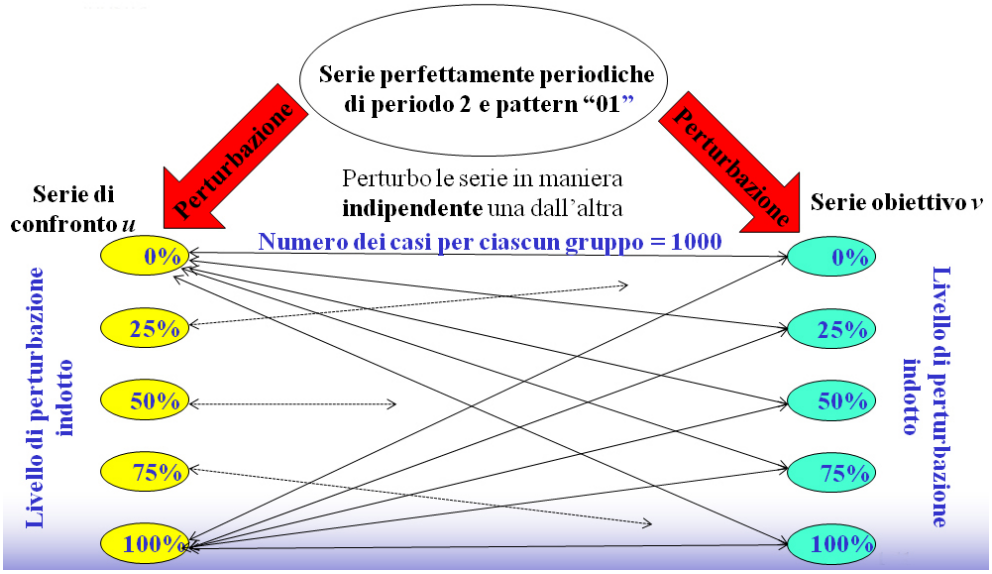

Figura 11.1: Confronto fra serie perfettamente periodiche a cui sono stati applicati vari livelli di perturbazione.

Ogni gruppo era costituito da 1000 serie (casi) generate dal calcolatore in maniera indipendente l'una dall'altra. A differenza di quanto stabilito per gli esperimenti realizzati in precedenza, si è qui convenuto di ridurre il numero dei casi da 5000 a 1000, in modo da velocizzare le operazioni di calcolo e di simulazione senza per questo causare sostanziali ripercussioni sulla stabilità e sulla conseguente affidabilità dei risultati.

Ogni serie era infine composta, come per lo studio dell'autocorrelazione, da 1000 elementi.

Per ogni singolo confronto fra una serie del primo gruppo con una serie del secondo gruppo sono stati calcolati gli indicatori precedentemente citati, ottenendone così la distribuzione empirica, e poi ne è stata fatta la media per ottenere, per ciascuno dei venticinque gruppi di confronto, anche dei valori che potessero esprimere l'andamento atteso del fenomeno. Oltre al

valor medio sono stati calcolati, a partire dalle distribuzioni empiriche così ottenute, anche la varianza ed i valori corrispondenti ai percentili 0.25 e 99.75 in modo da ottenere, per la media, intervalli di confidenza al 95%.

Diversamente da quanto effettuato per lo studio dell'autocorrelazione, dove si confrontavano unicamente serie fra loro uguali, la *cross-ApEn*, che è stata in questo caso utilizzata in luogo della $ApEn$, è stata calcolata solo per i livelli di m da 1 a 3, come per altro suggerito in letteratura, in modo da limitare i problemi di distorsione e di interpretazione dei risultati.

CAPITOLO 12

Analisi di serie perfettamente periodiche affette da rumore.

Per cominciare, si sono considerati due gruppi di serie caratterizzate da una medesima condizione iniziale, ovvero serie perfettamente periodiche di periodo due e pattern "01", e da un processo perturbativo caratterizzato da una probabilità di inserire, indifferentemente, uno zero oppure un uno in una qualsiasi posizione delle serie pari a 0.5. Sia nella condizione iniziale di perfetta periodicità che in corrispondenza dei vari livelli di perturbazione, i due gruppi di serie erano quindi sempre caratterizzati da una frequenza media di 0 e di 1 pari all'incirca a 0.5. In ragione di ciò ci si poteva ragionevolmente aspettare valori della $S\rho$, della mutua informazione, della distanza di Kullback Leibler e della *cross-ApEn* simmetrici rispetto ai livelli di perturbazione indotti. Ciò significa, ad esempio, che se si indica con u_{25} una serie del primo gruppo (serie di confrontop) perturbata al 25% e con v_{75} una serie del secondo gruppo (serie obiettivo) perturbata al 75%, in media ci si poteva aspettare che $S\rho(u_{25}; v_{75}) \simeq S\rho(u_{75}; v_{25})$ e cosi via. Inoltre, la distanza di Kullback Leibler si poteva supporre molto prossima a zero. Infatti, avendosi distribuzioni di probilità marginali medie uguali nei due gruppi posti a confronto, la "distanza" fra le stesse, in media, sarebbe stata nulla e la misura. pertanto, avrebbe dovuto tendere a zero.

12.1 Risultati dell'esperimento

12.1.1 Entropia di Granger, Maasoumi e Racine

Nella tabella 12.1 si può osservare come, concordemente alle aspettative, i valori della $S\rho$ risultino simmetrici rispetto alla diagonale principale e come essi diminuiscano all'aumentare del livello di randomizzazione delle due serie, ovvero man mano che ci si sposta verso il basso e verso destra nella tabella. Come ci si poteva aspettare, l'ultima riga e l'ultima colonna della tabella presentano valori della S_ρ praticamente coincidenti fra loro e comunque molto prossimi a zero, ovvero alla condizione di massima indipendenza. Ciò in accordo al fatto che almeno una delle due serie, pur conservando la distribuzione marginale di partenza, è di tipo completamente casuale. Si può inoltre rilevare come anche il valore della S_ρ relativo a serie entrambe perturbate al 75% non si discosti significativamente dalla condizione di indipendenza.

Tabella 12.1: Valori della S_ρ e relativi intervalli di confidenza al 95% riferiti a due serie di periodo 2 in funzione di diversi valori di randomizzazione caratterizzati da $p = q = 0.5$.

Serie sottostanti di periodo due e pattern 01 Lunghezza delle serie: 1000 Numero di casi: 1000

		Livello di perturbazione serie di confronto									
		0%		25%		50%		75%		100%	
Liv. di perturbazione serie obiettivo	0%	1	1 1	0,304037	0,276006 0,331751	0,117045	0,098241 0,137258	0,028178	0,018746 0,038837	0,000880	0,000034 0,002687
	25%	0,303297	0,276018 0,334025	0,152228	0,130234 0,175048	0,064227	0,048649 0,080808	0,016293	0,008961 0,025000	0,001004	0,000048 0,003100
	50%	0,117124	0,099357 0,137990	0,064137	0,049536 0,079047	0,028166	0,018444 0,039532	0,007967	0,002837 0,014757	0,001087	0,000045 0,003445
	75%	0,027966	0,018178 0,038166	0,016170	0,009157 0,024990	0,007897	0,002902 0,013959	0,002932	0,000304 0,007095	0,001223	0,000063 0,003745
	100%	0,000810	0,000034 0,002497	0,001032	0,000054 0,002938	0,001139	0,000056 0,003491	0,001260	0,000056 0,003952	0,001292	0,000058 0,004157

12.1.2 Mutua informazione

La mutua informazione (tabella 12.2) presenta un comportamento sostanzialmente analogo a quello della S_ρ. In questo caso, però, il valore relativo a serie entrambe perturbate al 75%, pur se estremamente basso, risulta significativamente diverso dai valori dell'ultima riga e dell'ultima colonna a riprova, ancora una volta, della maggiore "sensibilità" della mutua informazione rispetto alla S_ρ.

Tabella 12.2: Valori della mutua informazione e relativi intervalli di confidenza al 95% riferiti a due serie di periodo 2 in funzione di diversi valori di randomizzazione caratterizzati da $p = q = 0.5$.

Serie sottostanti di periodo due e pattern 01 Lunghezza delle serie: 1000 Numero di casi: 1000

		Livello di perturbazione serie di confronto									
		0%		25%		50%		75%		100%	
Liv. di perturbazione serie obiettivo	0%	1	1 1	0,457926	0,421072 0,494361	0,189194	0,160069 0,220280	0,046473	0,030837 0,064163	0,000749	0,000003 0,002778
	25%	0,456816	0,421072 0,497299	0,242834	0,209734 0,276477	0,105210	0,080302 0,131616	0,026390	0,014148 0,040732	0,000735	0,000003 0,002740
	50%	0,189197	0,161469 0,220780	0,105085	0,081149 0,128906	0,046033	0,030024 0,064929	0,012206	0,003943 0,023537	0,000668	0,000003 0,002427
	75%	0,046150	0,030239 0,063285	0,026103	0,014436 0,039557	0,012128	0,004167 0,021917	0,003539	0,000142 0,009067	0,000749	0,000003 0,002583
	100%	0,000648	0,000003 0,002775	0,000664	0,000002 0,002560	0,000698	0,000003 0,002714	0,000678	0,000003 0,002641	0,000674	0,000003 0,002413

12.1.3 Entropia relativa o distanza di Kullback Leibler

Si è considerata una sola distanza di Kullback Leibler (serie di confronto verso serie obiettivo) in quanto, trattando nel caso specifico serie aventi distribuzione marginale simmetrica del tipo $p = q = 0.5$, entrambe le distanze D_{xy} e D_{yx} risultano uguali per definizione.

Nella tabella 12.3 si può notare come i valori della distanza di Kullback Leibler siano tutti molto bassi e sostanzialmente uguali. Le differenze osservate non risultano essere in questo caso statisticamente significative, ma imputabili alle fluttuazioni casuali che si possono riscontrare nelle frequenze di zeri e di uno caratterizzanti ciascuna serie perturbata. E' quindi ragionevole ipotizzare che tali differenze, così come i valori dell'indicatore, possano ridursi ulteriormente aumentando la lunghezza delle serie, visto che ciò dovrebbe portare ad una diminuzione dell'intensità media delle fluttuazioni della frequenza di zeri e di uno attorno al loro valore atteso.

In ogni caso, appurato che i valori osservati attestano tutti una "distanza" media fra le serie considerate sostanzialmente nulla, si può osservare come la distanza di Kullback Leibler vada aumentando al crescere del livello di randomizzazione. Ciò è imputabile al fatto che, al crescere del livello di perturbazione indotta, aumenta di pari passo anche il "peso" degli elementi perturbati sul totale e, di conseguenza, la probabilità di avere distribuzioni marginali maggiormente variabili. Poiché il limite inferiore della distanza di Kullback Leibler è fisso e pari a zero, tale maggiore variabilità non può che tradursi in una "distanza" media maggiore, anche se, comunque, sempre molto bassa.

Tabella 12.3: Valori della distanza di Kullback Leibler e relativi intervalli di confidenza al 95% riferiti a due serie di periodo 2 in funzione di diversi valori di randomizzazione caratterizzati da $p = q = 0.5$.

Serie sottostanti di periodo due e pattern 01 Lunghezza delle serie: 1000 Numero di casi: 1000

		Livello di perturbazione serie N. uno									
		0%		25%		50%		75%		100%	
Liv. di perturbazione serie N. due	0%	0	0 0	0,000321	0,000003 0,001273	0,000534	0,000003 0,001951	0,000683	0,000003 0,002428	0,000738	0,000003 0,002957
	25%	0,000316	0,000003 0,001155	0,000626	0,000003 0,002429	0,000873	0,000003 0,003336	0,000981	0,000003 0,003952	0,000962	0,000003 0,003741
	50%	0,000546	0,000003 0,002107	0,000834	0,000003 0,003158	0,001075	0,000003 0,004168	0,001215	0,000003 0,004878	0,001169	0,000003 0,004390
	75%	0,000617	0,000003 0,002266	0,001052	0,000003 0,003969	0,001173	0,000003 0,004619	0,001410	0,000012 0,005589	0,001317	0,000003 0,004864
	100%	0,000722	0,000003 0,002610	0,001080	0,000003 0,004171	0,001226	0,000012 0,004621	0,001452	0,000003 0,005609	0,001510	0,000003 0,005850

12.1.4 Cross Approximate Entropy e cross Sample Approximate Entropy

Nel precedente capitolo 3.5 abbiamo visto che la *cross-ApEn*, così come la *ApEn*, misura il logaritmo della frequenza con cui blocchi (pattern) di lunghezza m che sono "vicini" fra loro rimangano "vicini" anche se si aumenta di un elemento la dimensione di tali blocchi. L'unica differenza risiede nel fatto che con la *ApEn* si confronta sempre una medesima serie con se stessa, mentre con la *cross-ApEn* si confrontano due serie non necessariamente uguali fra loro.

Da ciò ne discende che, nel caso della *ApEn*, il logaritmo della frequenza con cui blocchi (pattern) di lunghezza m (o $m+1$) risultano fra loro "vicini", ovvero:

$$\log C_i^m(r) = \log \frac{\text{numero di } j \leq N - m + 1 \text{ tali che } d(x_i, x_j) \leq r}{N - m + 1},$$

è sempre definito, in quanto il numeratore di un generico $C_i^m(r)$ assume sempre almeno il valore uno (confronto del pattern con se stesso). Questa condizione, invece, non si realizza necessariamente nella *cross-ApEn*. Se un determinato pattern di indice i della serie di confronto non trova alcun riscontro nella serie obiettivo, la quantità $C_i^m(r)$ assumerà valore zero. Conseguentemente il suo logaritmo risulterà non definito e la *cross-ApEn* non calcolabile. Ciò è proprio quello che si è verificato nel corso dell'esperimento. Infatti, in molti casi caratterizzati da $m > 1$ e livelli di randomizzazione diversi da zero, la *cross-ApEn* è risultata non calcolabile. Per ovviare a tale problema si è quindi convenuto di sostituire, nel corso di tutto l'esperimento, la *cross-ApEn* con la *cross Sample-ApEn* introdotta nello stesso capitolo 3.5.

Nella tabella 12.4, colonne bianco-azzurre, sono riportati i valori della *cross Sample-ApEn* per m da 1 a 3 e per i diversi valori di randomizzazione considerati. La colonna in giallo riporta invece i valori della *cross-ApEn* calcolati in corrispondenza di una serie di confronto perfettamente periodica e serie obiettivo variamente randomizzate (l'unico set di confronti che si è riuscito a calcolare interamente per mezzo della *cross-ApEn*). Come si può ben vedere, i valori forniti dai due indicatori sono sostanzialmente coincidenti. Questo si verifica, in particolare, per serie caratterizzate, come in questo caso, da distribuzioni marginali del tipo $p = q = 0.5$, mentre in altri contesti, per serie di lunghezza limitata, i valori dei due indicatori possono divergere in maniera anche abbastanza significativa. In ogni caso, a meno di ragioni particolari, è comunque preferibile utilizzare la *cross Sample-ApEn* che, oltre ad essere affetta solo in maniera marginale dal problema delle mancate coincidenze, presenta anche una distorsione minore rispetto alla *cross-ApEn* [43].

Dai dati riportati nella tabella 12.4 si può osservare come la *cross Sample-ApEn* aumenti man mano che ci si sposta verso il basso e verso destra nella tabella. La *cross Sample-ApEn*, e di conseguenza la *cross-ApEn*, aumentano quindi correttamente al crescere del livello di randomizzazione indotto, di volta in volta, nella serie di confronto, nella serie obiettivo o in entrambe, mostrando con ciò un comportamento tendenzialmente analogo a quello riscontrato con la S_ρ e con la mutua informazione.

E' inoltre interessante osservare come, per i livelli di randomizzazione più bassi, la *cross Sample-ApEn* diminuisca in maniera abbastanza evidente al crescere del valore del parametro m. Diminuzione che va via via riducendosi fino ad annullarsi del tutto in corrispondenza dei livelli di randomizzazione più elevati, dove si osservano, per ogni valore di m, valori della *cross Sample-ApEn* prossimi a uno, condizione che, per la *ApEn* corrisponde a quella di massima entropia e, quindi, di massimo disordine. La diminuzione del valore della *cross Sample-ApEn* che si registra, al crescere di m, in corrispondenza dei livelli di randomizzazione più bassi non è tuttavia necessariamente imputabile ad una maggiore regolarità di periodo più elevato. Infatti, viste le caratteristiche intrinseche delle serie considerate, è più verosimile che ciò sia da ricondursi ad una stima carente delle probabilità condizionali che pattern "vicini" per una data lunghezza m lo rimangano anche qualora si consideri l'elemento successivo della serie, ovvero qualora si porti la lunghezza del pattern di confronto a $m + 1$. Questo fatto consiglia ancora una volta di procedere con estrema cautela nell'interpretazione dei risultati forniti dalla *cross Sample-ApEn* (e dalla *ApEn* in generale), specie in mancanza di altre informazioni a contorno.

Nelle tabelle 12.5, 12.6 e 12.7 è evidenziato il comportamento della *cross Sample-ApEn* in corrispondenza di specifici valori di m. Dai risultati tabulati si può chiaramente osservare come, per ogni valore di m, il comportamento di fondo dell'indicatore sia ovviamente sempre lo stesso. Quello che cambia

Tabella 12.4: Valori di *cross Sample-ApEn* e relativi intervalli di confidenza al 95% riferiti a due serie di periodo 2 in funzione di diversi valori di randomizzazione caratterizzati da $p = q = 0.5$.

Serie sottostanti di periodo due e pattern 01 — Lunghezza delle serie: 1000 — Numero di casi: 1000

Livello di perturbazione serie obiettivo			Livello di perturbazione serie di confronto											
			Cross-ApEn	Sample Cross-ApEn										
		m	0%	0%		25%		50%		75%		100%		
	0%	1	0	0	0 0	0,356286	0,305127 0,404922	0,678867	0,599601 0,752075	0,914386	0,827173 1,004131	1,000217	0,911638 1,090925
		2	0	0	0 0	0,217776	0,182175 0,254186	0,515280	0,441714 0,593663	0,842666	0,735478 0,960941	1,003013	0,885226 1,133748
		3	0	0	0 0	0,197281	0,163555 0,233797	0,452138	0,379401 0,538190	0,789591	0,657669 0,945393	1,008004	0,860433 1,198048
	25%	1	0,357046	0,358076	0,310482 0,408750	0,604538	0,559385 0,650566	0,810598	0,758608 0,856071	0,949873	0,898483 1,000763	1,000133	0,950893 1,053444
		2	0,219051	0,219281	0,185353 0,255381	0,435023	0,387474 0,480995	0,682394	0,622525 0,739917	0,904896	0,839310 0,976375	1,001972	0,934443 1,082916
		3	0,197785	0,198350	0,166061 0,233720	0,380535	0,336873 0,421485	0,610914	0,546867 0,676085	0,867801	0,785764 0,957541	1,003642	0,913960 1,101843
	50%	1	0,675767	0,678909	0,606201 0,749433	0,810471	0,762915 0,860130	0,913005	0,881710 0,942668	0,978128	0,952563 1,002384	0,999805	0,978075 1,023520
		2	0,515930	0,516364	0,444330 0,589630	0,683006	0,623810 0,747432	0,840316	0,792692 0,884143	0,956926	0,919165 0,993448	0,999694	0,968878 1,033465
		3	0,451183	0,451871	0,376143 0,537059	0,610316	0,546040 0,680704	0,785383	0,729891 0,839845	0,937157	0,888874 0,983642	1,000316	0,960474 1,039447
	75%	1	0,912439	0,912729	0,834685 1,004171	0,950636	0,898526 1,001647	0,977402	0,953002 0,999783	0,994442	0,983757 1,001334	0,999858	0,992497 1,006779
		2	0,843064	0,840989	0,725906 0,966278	0,906641	0,836492 0,975615	0,956284	0,922696 0,988935	0,988981	0,973396 1,000076	0,999760	0,989160 1,009682
		3	0,790146	0,790807	0,649650 0,950172	0,869986	0,784910 0,952772	0,937296	0,893074 0,978502	0,983678	0,963969 0,998411	0,999752	0,986653 1,012451
	100%	1	1,001585	1,000987	0,911656 1,090818	1,001478	0,950034 1,055335	0,999511	0,977877 1,021739	0,999969	0,992611 1,007786	1,000126	0,995838 1,004656
		2	1,004683	1,001850	0,879004 1,136669	1,001962	0,933366 1,074890	0,999187	0,968626 1,029911	0,999940	0,989774 1,010315	1,000105	0,994440 1,006443
		3	1,006299	1,006812	0,836961 1,201427	1,003358	0,916424 1,096099	0,999658	0,960583 1,039519	0,999966	0,987704 1,012830	1,000322	0,991897 1,009215

sono le intensità dei valori osservati, i quali risultano mediamente più elevati in corrispondenza dei livelli di m più bassi.

Per $m = 1$ (tabella 12.5), i valori associati ai livelli di randomizzazione del 75% e del 100% risultano fra loro statisticamente indistinguibili, cosa che non si verifica, in genere, per $m = 2$ ed $m = 3$ ponendosi, anche in questo caso, problemi nell'interpretazione dei risultati e nella scelta del valore del parametro m da utilizzare.

Tabella 12.5: Valori di *cross Sample-ApEn* e relativi intervalli di confidenza al 95% riferiti a due serie di periodo 2 per $m = 1$ in funzione di diversi valori di randomizzazione caratterizzati da $p = q = 0.5$.

Serie sottostanti di periodo due e pattern 01 Lunghezza delle serie: 1000 Numero di casi: 1000

		Livello di perturbazione serie di confronto					
		Cross-ApEn	Sample Cross-ApEn				
		0%	0%	25%	50%	75%	100%
Liv. di perturbazione serie obiettivo	0%	0	0 0 0	0,356286 0,305127 0,404922	0,678867 0,599601 0,752075	0,914386 0,827173 1,004131	1,000217 0,911638 1,090925
	25%	0,357046	0,358076 0,310482 0,408750	0,604538 0,559385 0,650566	0,810598 0,758608 0,856071	0,949873 0,898483 1,000763	1,000133 0,950893 1,053444
	50%	0,675767	0,678909 0,606201 0,749433	0,810471 0,762915 0,860130	0,913005 0,881710 0,942668	0,978128 0,952563 1,002384	0,999805 0,978075 1,023520
	75%	0,912439	0,912729 0,834685 1,004171	0,950636 0,898526 1,001647	0,977402 0,953002 0,999785	0,994442 0,983757 1,001334	0,999858 0,992497 1,006779
	100%	1,001585	1,000987 0,911656 1,090818	1,001478 0,950034 1,055335	0,999511 0,977877 1,021739	0,999969 0,992611 1,007786	1,000126 0,995838 1,004656

Tabella 12.6: Valori di *cross Sample-ApEn* e relativi intervalli di confidenza al 95% riferiti a due serie di periodo 2 per $m = 2$ in funzione di diversi valori di randomizzazione caratterizzati da $p = q = 0.5$.

Serie sottostanti di periodo due e pattern 01 Lunghezza delle serie: 1000 Numero di casi: 1000

		Livello di perturbazione serie di confronto					
		Cross-ApEn	Sample Cross-ApEn				
		0%	0%	25%	50%	75%	100%
Liv. di perturbazione serie obiettivo	0%	0	0 0 0	0,217776 0,182175 0,254186	0,515280 0,441714 0,593663	0,842666 0,735478 0,960941	1,003013 0,885226 1,133748
	25%	0,219051	0,219281 0,185353 0,255381	0,435023 0,387474 0,480995	0,682394 0,622525 0,739917	0,904896 0,839310 0,976375	1,001972 0,934443 1,082916
	50%	0,515930	0,516364 0,444330 0,589630	0,683006 0,623810 0,747432	0,840316 0,792692 0,884143	0,956926 0,919165 0,993448	0,999694 0,968878 1,033465
	75%	0,843064	0,840989 0,725906 0,966278	0,906641 0,836492 0,975615	0,956284 0,922696 0,988935	0,988981 0,973396 1,000076	0,999760 0,989160 1,009682
	100%	1,004683	1,001850 0,879004 1,136669	1,001962 0,933366 1,074890	0,999187 0,968626 1,029911	0,999940 0,989774 1,010315	1,000105 0,994440 1,006443

Tabella 12.7: Valori di *cross Sample-ApEn* e relativi intervalli di confidenza al 95% riferiti a due serie di periodo 2 per $m = 3$ in funzione di diversi valori di randomizzazione caratterizzati da $p = q = 0.5$.

Serie sottostanti di periodo due e pattern 01 Lunghezza delle serie: 1000 Numero di casi: 1000

		Livello di perturbazione serie di confronto					
		Cross-ApEn	Sample Cross-ApEn				
		0%	0%	25%	50%	75%	100%
Liv. di perturbazione serie obiettivo	0%	0	0 0 0	0,197281 0,163555 0,233797	0,452138 0,379401 0,538190	0,789591 0,657669 0,945393	1,008004 0,860433 1,198048
	25%	0,197785	0,198350 0,166061 0,233720	0,380535 0,336873 0,421485	0,610914 0,546867 0,676085	0,867801 0,785764 0,957541	1,003642 0,913960 1,101843
	50%	0,451183	0,451871 0,376143 0,537059	0,610316 0,546040 0,680704	0,785383 0,729891 0,839845	0,937157 0,888874 0,983642	1,000316 0,960474 1,039447
	75%	0,790146	0,790807 0,649650 0,950172	0,869986 0,784910 0,952772	0,937296 0,893074 0,978502	0,983678 0,963969 0,998411	0,999752 0,986653 1,012451
	100%	1,006299	1,006812 0,836961 1,201427	1,003358 0,916424 1,096099	0,999658 0,960583 1,039519	0,999966 0,987704 1,012830	1,000322 0,991897 1,009215

12.2 Applicazione della cross Sample Approximate Entropy a serie fra loro uguali

Fino ad ora abbiamo confrontato fra loro serie omogenee sia per quanto riguarda la distribuzione marginale che il tipo di fenomeno sottostante, rappresentato, in questo caso, dal disegno della serie. Tali serie però non erano uguali. Per avere un quadro completo del comportamento della *cross Sample-ApEn* e, di conseguenza, della *cross-ApEn*, è opportuno applicare la *cross Sample-ApEn* anche a coppie di serie fra loro uguali ma caratterizzate, di volta in volta, da diversi livelli di randomizzazione. Dalla definizione di cui al capitolo 3.5 discende immediatamente che il calcolo della *cross-ApEn* (*cross Sample-ApEn*) su due serie uguali equivale al calcolo della *ApEn* (*Sample-ApEn*) su una delle due serie.

Il comportamento della *cross Sample-ApEn*, della S_ρ e della mutua informazione applicate a serie fra loro uguali è riassunto nella tabella 12.8. Dai risultati ottenuti si può osservare come sia la S_ρ che la mutua informazione assumano sempre, in questo caso, uno stesso valore, sostanzialmente pari all'unità, indipendentemente dal livello di randomizzazione considerato.

Tale valore, come ci si poteva aspettare, corrisponde a quello teorico di perfetta correlazione, dato che vi è sempre un'esatta corrispondenza fra ogni elemento delle serie poste a confonto. Il valore riscontrato risulta inoltre diverso dai valori presenti sulle diagonali principali delle precedenti tabelle 12.1 (S_ρ) e 12.2 (mutua informazione), dove venivano confrontate fra loro serie generalmente diverse, anche se caratterizzate da un medesimo livello di randomizzazione, e quindi non necessariamente correlate.

Diverso è invece il discorso per la *cross Sample-ApEn*, i cui valori non solo

Tabella 12.8: Valori di S_ρ, mutua informazione e *cross Sample-ApEn* e relativi intervalli di confidenza al 95% riferiti a serie uguali di periodo 2 in funzione di diversi livelli di randomizzazione caratterizzati da $p = q = 0.5$.

Serie sottostanti di periodo due e pattern 01 **Lunghezza delle serie: 1000** **Numero di casi: 1000**

	Livello di perturbazione serie di confronto e serie obiettivo									
	0%	25%		50%		75%		100%		
Sp	1	0,999611	0,998403 0,999996	0,999292	0,997160 0,999996	0,999181	0,996954 0,999996	0,999128	0,996740 0,999996
Mutua informazione	1	0,999690	0,998727 0,999997	0,999436	0,997737 0,999997	0,999347	0,997572 0,999997	0,999305	0,997402 0,999997
cross-Sample ApEn m = 1	0	0,602116	0,532310 0,665795	0,910772	0,863877 0,950905	0,991739	0,975323 0,999430	0,997158	0,989715 0,999927
m = 2	0	0,433127	0,368670 0,494411	0,838091	0,768735 0,899452	0,983356	0,960544 0,996795	0,994309	0,984027 0,999258
m = 3	0	0,377470	0,321652 0,434430	0,779912	0,702638 0,849114	0,972298	0,944554 0,991560	0,988775	0,975406 0,996786

variano al variare del valore del parametro m e del livello di randomizzazione, ma risultano anche sostanzialmente uguali a quelli presenti sulla diagonale principale della precedente tabella 12.4. Ciò significa che la *cross Sample-ApEn*, e la *cross-ApEn* da cui deriva, non sono in grado di evidenziare una condizione di perfetta correlazione, ovvero di perfetta uguaglianza fra tutti gli elementi delle due serie poste a confronto né, tanto meno, di distinguere, a parità di livello di randomizzazione e di distribuzione marginale, fra due serie uguali e fra due serie anche completamente diverse fra loro. La *cross Sample-ApEn* e la *cross-ApEn*, infatti, si basano sul principio della similarità fra pattern, ovvero su quante volte un pattern di una serie che si ritrova anche nell'altra, indipendentemente dalla sua posizione, vi si ritrova ancora se si aumenta la sua dimensione di un elemento. Non si ha quindi a che fare con una sincronia di tipo temporale o posizionale, e di ciò bisognerebbe essere ben consapevoli sia al momento di scegliere tali strumenti sia al momento di interpretarne i risultati pena, altrimenti, il rischio di finire completamente fuori strada.

CAPITOLO 13

Confronto fra serie perfettamente periodiche di disegno diverso

In questo capitolo verranno poste a confronto le serie perfettamente periodiche già studiate singolarmente nel precedente capitolo 5. Verrà considerato un unico caso per ogni confronto, in quanto trattasi di serie deterministiche perfettamente individuate, noto il pattern e la lunghezza del periodo. Ci si limiterà poi a considerare il solo caso $lag = 0$, in quanto il comportamento dei nostri indicatori al variare del lag è già stato ampiamente analizzato nei capitoli precedenti, mentre la *cross Sample-ApEn* verrà calcolata, come già nel precedente capitolo 12, per valori del parametro m variabili da 1 a 3.

Gli indicatori utilizzati, inoltre, non saranno normalizzati in quanto il nostro interesse è qui focalizzato sulle differenze fra i diversi fenomeni (il diverso disegno delle serie poste a confronto) in una determinata situazione (perfetta periodicità ed assenza di rumore) e non tanto sul comportamento dei diversi fenomeni in relazione a particolari stimoli esterni (ad esempio il livello di perturbazione), come invece è avvenuto nel capitolo 5.

13.1 Risultati dell'esperimento

13.1.1 Entropia di Granger, Maasoumi e Racine

Nella tabella 13.1 si può osservare come i valori della S_ρ relativi alla condizione di perfetta correlazione (confronto fra serie uguali) posti sulla diagonale principale della tabella varino al variare della distribuzione marginale

delle serie considerate e quindi, di per sé, non forniscano un risultato univoco ed immediatamente esplicativo di tale situazione.

Tabella 13.1: Valori della S_ρ non normalizzata: confronto fra serie perfettamente periodiche aventi disegno diverso.

Lunghezza delle serie: 1000

		Serie di confronto				
		'01'	'0011'	'0000011111'	'0000000111'	'0000000001'
Serie obiettivo	'01'	0,292893	0	0,005064	0,006118	0,030684
	'0011'	0	0,292893	0	0	0
	'0000011111'	0,005064	0	0,292893	0,104963	0,030684
	'0000000111'	0,006118	0	0,104963	0,250021	0,048770
	'0000000001'	0,030684	0	0,030684	0,048770	0,114562

Normalizzando l'indicatore, ovvero dividendo ogni volta i valori osservati per il massimo ammissibile date le distribuzioni marginali delle serie confrontate, la S_ρ assumerebbe valore uno ogni qualvolta la distribuzione congiunta esprimesse, fissate le distribuzioni marginali, una condizione di correlazione pari al massimo ammissibile. Tale valore, però, non evidenzierebbe necessariamente una condizione di perfetta correlazione (match di ogni elemento con se stesso), ma solamente il raggiungimento della massima correlazione ammissibile, al limite anche molto bassa, date le condizioni a contorno. Così facendo si individuerebbero immediatamente le serie perfettamente correlate ma non le si potrebbero distinguere da quelle che presentano massima correlazione ammissibile, ma che uguali non sono. Ancora una volta, quindi, emerge in maniera chiara il fatto che un indicatore, normalizzato o non normalizzato, non può essere applicato e considerato meccanicamente, ma deve essere valutato criticamente ed in associazione a tutte le informazioni a contorno che debbano e possano essere assunte. In questo caso, ad esempio, le distribuzioni marginali che, se uguali, indicherebbero una condizione di perfetta correlazione, altrimenti solo la massima correlazione ammissibile.

Appaiono invece correttemente ed immediatamente evidenziati i casi di indipendeza ($S_\rho = 0$). Si può poi inoltre notare come i restanti valori tendano a graduare le situazioni intermedie assegnando, correttamente, valori più bassi a quei confronti le cui distribuzioni congiunte tendano maggiormente ad una distribuzione uniforme e valori più alti a quelle che presentano, invece,

una predominanza di coppie di simboli uguali “00” e 11) od opposti (“10” e “01”).

13.1.2 Mutua informazione

La mutua informazione, i cui valori sono riportati nella tabella 13.2, presenta un comportamento del tutto analogo alla S_ρ. L'unica differenza risiede nell'intensità dei valori, per la mutua informazione tutti molto più alti a causa delle diverse modalità di calcolo.

Tabella 13.2: Valori della mutua informazione non normalizzata: confronto fra serie perfettamente periodiche aventi disegno diverso.

Lunghezza delle serie: 1000

		Serie di confronto				
		'01'	'0011'	'0000011111'	'0000000111'	'0000000001'
Serie obiettivo	'01'	1	0	0,029049	0,034852	0,108032
	'0011'	0	1	0	0	0
	'0000011111'	0,029049	0	1	0,395816	0,108032
	'0000000111'	0,034852	0	0,395816	0,881291	0,193507
	'0000000001'	0,108032	0	0,108032	0,193507	0,468996

13.1.3 Entropia relativa o distanza di Kullback Leibler

I valori delle distanze di Kullback Leibler sono riportati nelle tabelle 13.3 e 13.4. Come si può chiaramente osservare, in generale l'entropia relativa non è simmetrica. Lo è solamente nel caso di distribuzioni marginali uguali oppure opposte, cioè $p_u = q_v$ e $p_v = q_u$. Poiché la distanza di Kullback Leibler misura la “distanza” fra due distribuzioni, essa assume valore zero solo nel caso di marginali uguali. Essa, inoltre, una volta fissate le distribuzioni marginali, assume sempre lo stesso valore a prescindere dalla forma della distribuzione congiunta. Per tale motivo la distanza di Kullback Leibler non è grado di caratterizzare la relazione fra due serie al di la di una generica indicazione di “distanza” fra le distribuzioni marginali delle serie stesse, potendosi uno stesso valore riferire a situazioni anche molto diverse fra loro.

Tabella 13.3: Valori della distanza di Kullback Leibler fra serie di confronto e serie obiettivo (D_{xy}): confronto fra serie perfettamente periodiche aventi disegno diverso.

Lunghezza delle serie: 1000

		Serie di confronto				
		'01'	'0011'	'0000011111'	'0000000111'	'0000000001'
Serie obiettivo	'01'	0	0	0	0,118709	0,531004
	'0011'	0	0	0	0,118709	0,531004
	'0000011111'	0	0	0	0,118709	0,531004
	'0000000111'	0,125769	0,125769	0,125769	0	0,167817
	'0000000001'	0,736966	0,736966	0,736966	0,221690	0

Tabella 13.4: Valori della distanza di Kullback Leibler fra serie obiettivo e serie di confronto (D_{yx}): confronto fra serie perfettamente periodiche aventi disegno diverso.

Lunghezza delle serie: 1000

		Serie obiettivo				
		'01'	'0011'	'0000011111'	'0000000111'	'0000000001'
Serie di confronto	'01'	0	0	0	0,125769	0,736966
	'0011'	0	0	0	0,125769	0,736966
	'0000011111'	0	0	0	0,125769	0,736966
	'0000000111'	0,118709	0,118709	0,118709	0	0,221690
	'0000000001'	0,531004	0,531004	0,531004	0,167817	0

13.1.4 Cross Sample Approximate Entropy

Nelle tabelle 13.5, 13.6 e 13.7 sono riportati i valori della *cross Sample-ApEn* per livelli del parametro m rispettivamente pari a 1, 2 e 3. Come si può immediatamente notare, la *cross Sample-ApEn* risulta simmetrica, a meno di approssimazioni calcolatorie, per ogni valore del parametro m.

Come già evidenziato nel capitolo 5.2.6, nel caso di serie perfettamente

Tabella 13.5: Valori della *cross Sample-ApEn* per $m = 1$: confronto fra serie perfettamente periodiche aventi disegno diverso.

Lunghezza delle serie: 1000

m = 1		Serie di confronto				
		'01'	'0011'	'0000011111'	'0000000111'	'0000000001'
Serie obiettivo	'01'	0	1,001433	2,324822	2,325396	2,333239
	'0011'	1,001433	0	0,999123	0,998542	0,997962
	'0000011111'	2,324823	0,999127	0,555365	0,555397	0,554153
	'0000000111'	2,325403	0,998550	0,555403	0,464618	0,398756
	'0000000001'	2,333264	0,997961	0,554168	0,398765	0,310840

Tabella 13.6: Valori della *cross Sample-ApEn* per $m = 2$: confronto fra serie perfettamente periodiche aventi disegno diverso.

Lunghezza delle serie: 1000

m = 2		Serie di confronto				
		'01'	'0011'	'0000011111'	'0000000111'	'0000000001'
Serie obiettivo	'01'	0	---	---	---	0,998555
	'0011'	---	0	1,325700	1,326855	2,333832
	'0000011111'	---	1,325696	0,626959	0,627343	0,562813
	'0000000111'	---	1,326853	0,627334	0,484880	0,432979
	'0000000001'	0,998547	2,333850	0,562811	0,432960	0,341218

periodiche la *ApEn* si annulla quando il valore del parametro m è pari o prossimo al periodo della serie. Ciò è quanto si osserva anche con la *cross Sample-ApEN*, per tutti e tre i valori di m considerati, confrontando fra loro due serie di disegno "01" (periodo due) e due serie di disegno "0011" (periodo quattro). Nel caso delle serie "0000011111", "0000000111" e "0000000001" di periodo 10, invece, la *cross Sample-ApEn* di parametro $m = 1$ assume valori

Tabella 13.7: Valori della *cross Sample-ApEn* per $m = 3$: confronto fra serie perfettamente periodiche aventi disegno diverso.

Lunghezza delle serie: 1000

m = 3		Serie di confronto				
		'01'	'0011'	'0000011111'	'0000000111'	'0000000001'
Serie obiettivo	'01'	0	---	---	---	---
	'0011'	---	0	0,998547	0,998547	---
	'0000011111'	---	0,998549	0,651686	0,653007	0,716386
	'0000000111'	---	0,998549	0,653012	0,448513	0,508954
	'0000000001'	---	---	0,716379	0,508956	0,379013

positivi e via via decrescenti pari, rispettivamente a 0.5554, 0.4646 e 0.3108 (vedi diagonale principale della tabella 13.5).

Da quanto sopra risulta evidente che la *cross Sample-ApEn* non è in grado di evidenziare una condizione di perfetta correlazione (e quindi di perfetta sincronia temporale) quale quella che si realizza, necessariamente, nel confronto di una serie con se stessa poiché, come si può ben vedere, i valori posti sulle diagonali principali delle tabelle 13.5, 13.6 e 13.7 risultano diversi fra loro. E ciò non può essere sicuramente imputato alla mancanza di normalizzazione, che i primi tre confronti riguardavano tutti serie caratterizzate dalla medesima distribuzione marginale $p = q = 0.5$.

Si deve inoltre rilevare come, per $m = 2$ ed ancora di più per $m = 3$, diversi confronti non abbiano soluzione in quanto la quantità $A^m(r)(v||u)$, posta a numeratore nella formula per il calcolo della *cross Sample-ApEn*, assume valore zero (nessun riscontro fra pattern di lunghezza $m + 1$ appartenenti alle due serie u e v). Questo problema, già osservato nel caso della *cross-ApEn* (capitolo 12.1.4), si ripresenta adesso anche per la *cross Sample-ApEn*. Bisogna comunque dire che, rispetto alla *cross-ApEn*, la *cross Sample-ApEn*, ne è affetta in misura molto minore. In ogni caso, il sapere che $A^m(r)(v||u) = 0$, il più delle volte può essere informativo tanto quanto un valore numerico. Limitandoci infine a considerare il caso $m = 1$ (tabella 13.5), si può vedere come, a differenza della *ApEn*, per la *cross-ApEn* e per la *cross Sample-ApEn* siano possibili anche valori maggiori di uno. Nella *ApEn*, infatti, il massimo valore ammissibile, che si raggiunge nel caso di completa randomicità, è pari a $\log_2 k$, con k che indica il numero di simboli dell'alfabeto considerato (nel caso di serie binarie $k = 2$). Dalla tabella 13.5, invece, si può osservare come i valori della *cross Sample-ApEn* riferiti ai confronti fra la serie di pattern "01" (periodo due) e le serie di pattern "0000011111", "0000000111" e "0000000001" (periodo dieci) siano al-

l'incirca pari a 2.32, un valore assai più alto di quello, pari all'incirca ad uno, che la stessa *cross Sample-ApEn* assume, in media, nel caso del confronto fra due serie completamente casuali caratterizzate dalle medesime distribuzioni marginali, $p = q = 0.5$, delle serie di pattern "01" e "0000011111" (tabella 12.4).

13.2 Confronto fra serie indipendenti

Si vuole ora verificare se la *cross Sample-ApEn* sia o meno in grado di evidenziare, in maniera univoca, un'eventuale condizione di indipendenza fra due serie. A questo scopo andremo a confrontare quattro diverse coppie di serie indipendenti caratterizzate però sempre dalle stesse distribuizioni marginali. In particolare:

- *prima coppia*: entrambe le serie sono periodiche di periodo 100 e costruite in modo da risultare fra loro indipendenti. Infatti, il disegno della serie di confronto è costituito da 70 zeri consecutivi seguiti da 30 uno ($p_u = 0.3$ e $q_u = 0.7$), mentre quello della serie obiettivo è formato da 28 zeri seguiti da 42 uno, poi da 12 zeri seguiti da 18 uno ($p_v = 0.6$ e $q_v = 0.4$);

- *seconda coppia*: la serie di confronto, periodica di periodo 100, è la stessa della prima coppia, mentre la serie obiettivo è una serie completamente casuale avente distribuzione marginale $p_v = 0.6$ e $q_v = 0.4$;

- *terza coppia*: la serie di confronto è una serie completamente casuale avente distribuzione marginale $p_u = 0.3$ e $q_v = 0.7$, mentre la serie obiettivo, periodica di periodo 100, è la stessa della prima coppia;

- *quarta coppia*: entrambe le serie sono completamente casuali ed aventi distribuzione marginale, rispettivamente, $p_u = 0.3$, $q_u = 0.7$ e $p_v = 0.6$ e $q_v = 0.4$.

Dai risultati riportati nella tabella 13.8, si rileva subito che la $S\rho$, la mutua informazione e le distanze di Kullback Leibler non variano al variare del tipo di serie in ingresso poiché non variano le distribuzioni marginali che caratterizzano le serie stesse. Inoltre, essendo tutte e quattro coppie di serie indipendenti fra loro, tale condizione viene correttamente individuata sia dalla S_ρ che dalla mutua informazione, le quali assumono sempre un valore molto prossimo a zero.

La *cross Sample-ApEn* assume invece valori differenti per ciascuna coppia ad eccezione della seconda e della quarta, i cui valori non differiscono fra loro in maniera significativa ($\alpha = 0.05$) pur risultando significativamente diversi dagli altri. In particolare, nel caso del confronto fra serie entrambe

Tabella 13.8: Confronti fra serie fra loro indipendenti in funzione del tipo di serie in ingresso.

		Serie di confronto $p_u = 0.3$, $q_u = 0.7$				Serie obiettivo $p_v = 0.6$, $q_v = 0.4$					
		Sp		Mutua informazione		Dxy		Dyx		*Cross Sample-ApEn*	
Serie confrontate	Periodica vs Periodica	0	0 0	0	0 0	0,265148	0,265148 0,265148	0,277058	0,277058 0,277058	0,092081	0,092081 0,092081
	Periodica vs Casuale	0,000124	0 0,000621	0,000713	0,000001 0,003592	0,265292	0,212727 0,320726	0,277025	0,223896 0,332021	1,069765	0,970920 1,168326
	Casuale vs Periodica	0,000123	0 0,000614	0,000710	0,000001 0,003528	0,265553	0,218813 0,318549	0,277834	0,225161 0,339298	0,896609	0,814109 0,983641
	Casuale vs Casuale	0,000130	0 0,000598	0,000749	0,000001 0,003458	0,265211	0,193566 0,346874	0,277212	0,201014 0,366260	1,120107	1,076882 1,170320

periodiche, ma comunque indipendenti fra loro, la *cross Sample-ApEn* assume un valore molto basso, pari a circa 0.092, e prossimo allo zero, indice di massima regolarità. Negli altri tre confronti, invece, l'indicatore assume un valore decisamente più elevato e pari, a seconda dei casi, a circa 1.07, 0.90 e 1.12.

Da qui, ancora, i dubbi sul significato e sull'interpretazione dei risultati della *cross Sample-ApEn* e della *cross-ApEn*.

13.3 Confronto fra serie perfettamente complementari

Per caratterizzare ulteriormente il comportamento dei nostri indicatori, si è eseguito un ulteriore esperimento confrontando fra loro quattro gruppi di serie perfettamente complementari (ad uno zero nella serie di confronto corrisponde sempre un uno nella serie obiettivo e viceversa) aventi disegno e distribuzioni marginali diverse da gruppo a gruppo. Dai risultati ottenuti, riassunti nella tabella 13.9, si può notare come sia la S_ρ che la mutua informazione non normalizzate, come già nel caso della perfetta correlazione, assumano valori diversi in relazione alle diverse distribuzioni marginali delle serie confrontate. Tuttavia i valori osservati coincidono con quelli relativi alla condizione di perfetta correlazione riportati sulla diagonale principale delle precedenti tabelle 13.1 e 13.2 e riferiti alle serie il cui pattern è evidenziato in nero nella tabella 13.9. E' evidente, quindi, che tali indicatori, pur individuando una condizione di regolarità fra le serie poste a confronto, non sono in grado di distinguere fra correlazione diretta e correlazione inversa, dato che assumono, in entrambi i casi, i medesimi valori. Per quanto riguarda la *cross Sample-ApEn*, può avere significato un valore di 3.18, e quindi molto elevato, relativo al confronto fra due serie che a prima vista esprimono

Tabella 13.9: Confronto fra serie binarie perfettamente complementari.

Lunghezza delle serie: 1000

	Sp	Mutua informazione	Dxy	Dyx	cross Sample APEN
'0101010101' vs '1010101010'	0,292893	1	0	0	0
'1011001100' vs '0100110011'	0,292893	1	0	0	0,943869
'0000011111' vs '1111100000'	0,292893	1	0	0	0,555361
'0000000111' vs '1111111000'	0,250021	0,881291	0,488957	0,488957	0,691878
'0000000001' vs '1111111110'	0,114562	0,468996	2,535940	2,535940	3,181222

poca o nulla sincronia, così come il valore zero associato al primo confronto, dove seppur sfalsati di un lag, gli elementi della serie di confronto si ripetono con perfetta regolarità nella serie obiettivo. Di più difficile interpretazione appaiono invece i valori relativi ai tre confronti centrali. In ogni caso, tali risultati appaiono completamente slegati dal concetto di correlazione inversa intesa, nel caso binario, come perfetta coincidenza di ciascun elemento con il suo complementare.

capitolo 14

Considerazioni finali

Nel corso degli esperimenti effettuati nella terza parte della tesi, si è avuto modo di verificare come la S_ρ e la mutua informazione si prestino in maniera abbastanza soddisfacente allo studio delle relazioni eventualmente esistenti fra serie diverse. Esse presentano un comportamento ed una dinamica molto simile e riescono ad individuare efficacemente una eventuale situazione di indipendenza fra le serie considerate: in presenza di una tale condizione, entrambi gli indicatori assumono, in media, un valore molto prossimo a zero.

Essi riescono inoltre a rilevare una situazione di perfetta correlazione, anche se tale condizione può non essere immediatamente riconoscibile. Infatti, i valori relativi alla condizione di perfetta correlazione (o di massima correlazione ammissibile) variano in funzione delle distribuzioni marginali e corrispondono ai massimi teorici assumibili dagli indicatori nelle diverse condizioni. Nel caso in cui il valore rilevato corrisponda al massimo teorico e le distribuzioni marginali siano uguali, si è in presenza di una condizione di perfetta correlazione (serie uguali), se invece si è raggiunto il massimo teorico ma le marginali non sono uguali, si è invece in presenza della condizione di massima correlazione compatibilmente con le marginali date.

Nel caso si normalizzino gli indicatori rapportandoli al loro massimo, rendendoli così variabili fra zero ed uno, la loro lettura ne risulterebbe certamente assai semplificata, ciò non di meno però, per distinguere fra serie perfettamente correlate e serie massimamente correlate bisognerebbe comunque analizzare le distribuzioni marginali e verificare se queste siano uguali o meno.

La S_ρ e la mutua informazione, inoltre, non sono in grado di distinguere fra correlazione diretta e correlazione inversa, tanto che, nel caso di perfetta

correlazione, sia diretta che inversa, esse assumono il medesimo valore.

Si è visto poi che la distanza di Kullback Leibler non è assolutamente in grado di qualificare le serie in base al loro grado di relazione o di struttura al di la di una generica indicazione di "distanza" fra le loro distribuzioni marginali. Infatti, a parità di tali distribuzioni, essa assume sempre il medesimo valore anche in presenza di serie di disegno diverso.

La *cross Sample-ApEn* e la *cross-ApEn*, infine, misurano una generica sincronia fra gli elementi delle due serie poste a confronto, ovvero sintetizzano la frequenza con cui un generico pattern di lunghezza m appartenente alla serie di confronto e che si ritrova nella serie obiettivo, vi si ritrova ancora se si aumenta la sua dimensione di una unità. Al di la della loro definizione, il significato di questi indicatori non è sempre chiaro né facile da interpretare. Intanto, a differenza della *ApEn*, il cui massimo assoluto, pari a uno per le serie binarie, è esplicativo della condizione di massima casualità e, in definitiva, di indipendenza, la *cross Sample-ApEn* e la *cross-ApEn* possono assumere anche valori maggiori di uno. Dagli esperimenti condotti si evince poi chiaramente che, confrontando diversi tipi di serie fra loro indipendenti, la *cross Sample-ApEn* può assumere valori anche molto diversi e questo a prescindere dalle distribuzioni marginali considerate.

La *cross Sample-ApEn* non è inoltre in grado di rilevare una condizione di perfetta o massima correlazione. Si sono viste, infatti, coppie di serie perfettamente correlate esprimere valori diversi di *cross Sample-ApEn*. Allo stesso modo, a parità di distribuzioni marginali e di livello di perturbazione indotto, serie non correlate e serie perfettamente correlate potevano presentare, in media, un medesimo valore. A maggior ragione, la *cross Sample-ApEn* non è in grado di rilevare un'eventuale condizione di correlazione inversa.

Per quanto sopra, appare lecito esprimere seri dubbi sul significato reale della *cross Sample-ApEn* e della *cross-ApEn*, nonchè sulle possibili interpretazioni dei loro risultati al di la di una generica indicazione di maggiore o minore sincronia non posizionale fra gli elementi delle serie considerate. Appare quindi più che mai necessario utilizzare questi indicatori con estrema cautela ed avendo bene chiaro in mente che tipo di informazioni si vogliono e si possono ottenere.

Parte IV

Modifiche alla ***cross Sample-***$ApEn$, l'indice di corrispondenza

CAPITOLO 15

Introduzione

Come abbiamo visto nei capitoli precedenti, sia la *cross Sample-ApEn* che la *cross-ApEn* non sono in grado di:

- evidenziare una eventuale condizione di perfetta correlazione, sia essa diretta o indiretta. A parità di distribuzioni marginali essi possono infatti assumere i medesimi valori sia nel caso di serie fra loro incorrelate, perchè indipendenti, che nel caso di serie perfettamente correlate perchè uguali;
- rilevare in maniera non ambigua un'eventuale situazione di indipendenza. Tali indicatori, infatti, possono assumere valori diversi, anche a parità di distribuzioni marginali, nel caso di coppie di serie indipendenti ma di disegno diverso.

In questa quarta parte della tesi si cercherà pertanto di caratterizzare un indicatore, che chiameremo "indice di corrispondenza" o I_{crs}, in modo da colmare almeno in parte le lacune di cui sopra rendendo la *cross Sample-ApEn* e la *cross-ApEn* maggiormente informative senza per questo appesantirne o complicarne oltre misura l'algoritmo di calcolo.

L'indice di corrispondenza, che consiste sostanzialmente in un contatore che tenga conto del numero di corrispondenze fra pattern di lunghezza m aventi la stessa posizione all'interno delle serie confrontate, verrà quindi utilizzato per gli stessi esperimenti già illustrati nella Parte II e III della tesi. I risultati ottenuti saranno poi confrontati con quelli della S_ρ, della mutua informazione, della *ApEn* e della *cross-ApEn* per valutarne le prestazioni e la reale utilità.

CAPITOLO 16

L'indice di corrispondenza

Nel precedente capitolo 3.5.1 abbiamo visto che le quantità fondamentali che entrano in gioco nel calcolo della *cross Sample-ApEn*[1] fra due serie qualunque u e v sono le seguenti:

- $B_i^m(r)(v||u) = \dfrac{\text{numero di } j \leq N-m+1 \text{ tali che } d(x_i, x_j) \leq r}{N-m+1}$, dove x_i ed x_j sono vettori di m elementi consecutivi tratti, rispettivamente, dalle serie u e v;
- $B^m(r)(v||u) = \dfrac{\sum_{i=1}^{N-m+1} B_i^r(r)(v||u)}{N-m+1}$

Allo stesso modo gli autori, Richman e Moorman [43], definiscono le quantità $A_i^m(r)(v||u)$ ed $A^m(r)(v||u)$, dove però x_i ed x_j sono vettori composti da $m+1$ elementi consecutivi. $B^m(r)(v||u)$ indica quindi la probabilità che due serie u e v abbiano in comune m elementi consecutivi, comunque posizionati, mentre $A^m(r)(v||u)$ indica la probabilità che ne abbiano $m+1$. A partire da tali quantità essi definiscono la *cross Sample-ApEn* come:

$$\textit{cross Sample-ApEn} = \lim_{N \to +\infty} -\log \frac{A^m(r)(v||u)}{B^m(r)(v||u)}$$

La *cross Sample-ApEn* può essere quindi stimata per mezzo dalla statistica $\textit{cross Sample-ApEn}^m(r)(v||u) = -\log \dfrac{A^m(r)(v||u)}{B^m(r)(v||u)}$.

Ad integrazione di quanto sopra, si propone ora quanto segue:

[1] Un discorso del tutto analogo può essere fatto anche per la *cross-ApEn* e la *ApEn*

- $M_i^m(r)(v||u) = 1$ se $d(x_i, x_j) \leq r$ quando $i = j$, dove x_i ed x_j sono vettori di m elementi consecutivi tratti, rispettivamente, dalle serie u e v. $M_i^m(r)(v||u)$ è quindi una variabile indicatrice che può assumere:
  - valore 1 se e solo se i pattern di indice i tratti dalle due serie u e v sono fra loro "vicini" (uguali a meno della quantità r),
  - valore 0 altrimenti;
- $M^m(r)(v||u) = \dfrac{\sum_{i=1}^{N-m+1} M_i^m(r)(v||u)}{N-m+1}$. Il contatore $M^m(r)(v||u)$ indica pertanto il numero complessivo di pattern corrispondenti che risultano uguali a meno della quantità r rapportati al numero totale, $N-m+1$, dei confronti fra pattern corrispondenti.

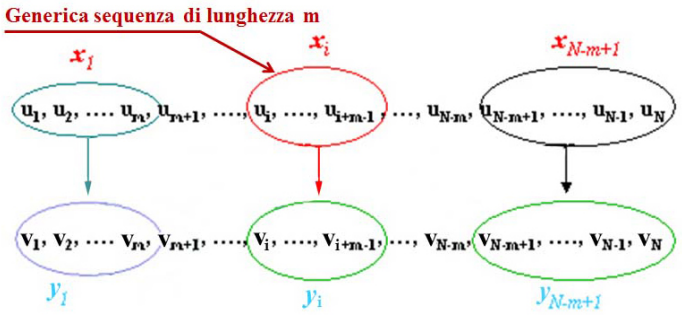

Figura 16.1: Modalità di calcolo dell'indice di corrispondenza: confronto fra blocchi x_i ed y_i.

La quantità $M^m(r)(v||u)$, che chiameremo "indice di corrispondenza di ordine m" o, anche, I_{crs}^m, potrà quindi assumere un valore compreso fra zero ed uno. Assumerà valore uno se e solo se tutti i pattern confrontati risulteranno "vicini" e quindi se le due serie sono uguali fra loro a meno del fattore di tolleranza r (condizione di perfetta correlazione). Assumerà invece valore zero se non vi è alcuna corrispondenza fra pattern aventi la stessa posizione. Nel caso di serie binarie un valore zero indicherà una condizione di perfetta ancorchè negativa correlazione in cui ogni zero è associato ad un uno e viceversa. In serie più generali, ovviamente, ciò indicherà solamente che ciascun elemento non risulterà mai associato a sè stesso nella distribuzione congiunta espressa dalla due serie considerate.

Se $I_{crs}^m = 1$ per un qualche valore di m, allora I_{crs}^m non potrà che assumere valore uno per tutti i possibili valori del parametro m. Infatti ciò implica una corrispondenza uno ad uno fra tutti gli elementi delle serie considerate, pertanto, qualunque sia la lunghezza del pattern di confronto, gruppi di m elementi consecutivi aventi la stessa posizione all'interno delle due serie non potranno che risultare uguali in quanto lo sono, singolarmente, gli elementi

che li compongono. Se invece $I^m_{crs} = 0$ per un qualche valore $m = k$, allora l'indice assumerà valore zero anche per tutti i successivi valori di $m > k$ in quanto, se non vi è nessuna corrispondenza fra pattern di lunghezza k, ciò significa che essi differiscono già per almeno un elemento, e quindi non vi potrà comunque essere corrispondenza se a tali pattern si aggiungerà un ulteriore elemento, qualunque esso sia.

16.1 Massimi e minimi dell'indice di corrispondenza

I valori zero ed uno sono dei massimi assoluti raggiungibili solo in particolari condizioni. Nello specifico, il valore uno può essere raggiunto solo nel caso di distribuzioni marginali fra loro uguali, mentre il valoro zero è compatibile unicamente con distribuzioni marginali fra loro opposte. In tutti gli altri casi, si avranno dei minimi e dei massimi relativi variabili in funzione delle distribuzioni marginali che caratterizzano le serie confrontate.

In particolare, per $m = 1$ e date due serie aventi distribuzione marginale rispettivamente $P_u(X = 1) = p_u, P_u(X = 0) = q_u$ e $P_v(X = 1) = p_v, P_v(X = 0) = q_v$:

- il massimo relativo dell'indice di corrispondenza assumerà il valore:

  $$\min(p_u, p_v) + \min(q_u, q_v)$$

  Da ciò deriva immediatamente che, nel caso di distribuzioni marginali uguali $p_u = p_v = p$ e $q_u = q_v = q$, il massimo relativo avrà forma $p+q = 1$ e, quindi, coinciderà con il massimo assoluto e con la condizione di perfetta correlazione;

- analogamente, il minimo relativo si avrà in corrispondenza del valore

  $$|p_u - q_v| \text{ (oppure } |p_v - q_u|, \text{ ma il risultato non cambia).}$$

  Da ciò deriva immediatamente che, nel caso di distribuzioni marginali opposte $p_u = q_v$ e $q_u = p_v$, il minimo relativo assumerà la forma $|p_u - q_v| = 0$, coincidendo con il minimo assoluto che, nel caso di serie binarie, esprime la condizione di perfetta correlazione inversa.

16.2 L'indice di corrispondenza e l'indipendenza in distribuzione

L'indice di corrispondenza è in grado di rilevare anche un'eventuale situazione di indipendenza fra due serie. In particolare, condizione necessaria e sufficiente perchè fra due serie binarie qualsiasi vi sia indipendenza, e quindi

incorrelazione, è che si abbia:

$$I^1_{crs} = p_u p_v + q_u q_v$$

La necessarietà deriva dal fatto che, nel caso di indipendenza, la probabilità che due qualsiasi elementi corrispondenti siano uguali coincide con la somma delle probabilità che tali elementi risultino essere, rispettivamente, due uno oppure due zero (non vi è iterazione), ovvero:

$$\begin{aligned} P_{uv}[(1,1)\bigcup(0,0)] &= P_{uv}(1,1) + P_{uv}(0,0) \\ &= P_u(X=1)P_v(X=1) + P_u(X=0)P_v(X=0) \\ &= p_u p_v + q_u q_v \end{aligned}$$

Nel caso in cui le due serie abbiano medesima distribuzione marginale del tipo $p = q = 0.5$, la condizione di cui sopra si riduce a $I^1_{crs} = p^2 + q^2$ per cui $I^1_{crs} = 0.5^2 + 0.5^2 = 0.25 + 0.25 = 0.5$.

Per dimostare la sufficienza si può procedere come segue. Data la tabella 16.2, riassuntiva di una generica distribuzione congiunta e delle due relative distribuzioni marginali

Tabella 16.1:

	p_v	q_v
p_u	$P_{uv}(1,1)$	$P_{uv}(1,0)$
q_u	$P_{uv}(0,1)$	$P_{uv}(0,0)$

Poiché, per ipotesi è $P_{uv}(1,1) + P_{uv}(0,0) = p_u p_v + q_u q_v$, la tabella 16.2 può essere riscritta come:

Tabella 16.2:

	p_v	q_v
p_u	$p_u p_v + q_u q_v - P_{uv}(0,0)$	$P_{uv}(1,0)$
q_u	$P_{uv}(0,1)$	$P_{uv}(0,0)$

da cui deriva che

$$\begin{cases} p_u p_v + q_u q_v - P_{uv}(0,0) + P_{uv}(0,1) = p_v \\ P_{uv}(0,1) + P_{uv}(0,0) = q_u \end{cases}$$

e quindi

$$\begin{cases} p_u p_v + q_u q_v - P_{uv}(0,0) + P_{uv}(0,1) + P_{uv}(0,1) + P_{uv}(0,0) = p_v + q_u \\ P_{uv}(0,1) + P_{uv}(0,0) = q_u \end{cases}$$

$$\begin{cases} 2P_{uv}(0,1) = p_v + q_u - p_u p_v - q_u q_v \\ P_{uv}(0,1) + P_{uv}(0,0) = q_u \end{cases}$$

$$\begin{cases} 2P_{uv}(0,1) = p_v(1-p_u) + q_u(1-q_v) \\ P_{uv}(0,1) + P_{uv}(0,0) = q_u \end{cases}$$

$$\begin{cases} 2P_{uv}(0,1) = p_v q_u + p_v q_u \\ P_{uv}(0,1) + P_{uv}(0,0) = q_u \end{cases}$$

$$\begin{cases} P_{uv}(0,1) = p_v q_u \\ P_{uv}(0,0) = q_u - p_v q_u \end{cases}$$

$$\begin{cases} P_{uv}(0,1) = p_v q_u \\ P_{uv}(0,0) = q_u(1-p_v) = q_u q_v \end{cases}$$

da cui, infine:

Tabella 16.3:

	p_v	q_v
p_u	$p_u p_v$	$p_u q_v$
q_u	$q_u p_v$	$q_u q_v$

che è proprio la tabella di indipendenza ottenuta a partire dalle distribuzioni marginali date. Analoga dimostrazione può essere fatta anche per $m = 2$ ed $m = 3$ e, comunque, per $m > 1$.

Nel resto della trattazione ci limiteremo comunque a considerare l'indice di corrispondenza di ordine uno ovvero I^1_{crs}, nel seguito indicato semplicemente come I_{crs}.

Sotto l'ipotesi nulla di indipendenza e per un numero n di elementi costituenti le serie sufficientemente elevato, l'indice di corrispondenza di ordine uno I_{crs} si distribuisce come una normale $\mathcal{N}\left[p_{ind}, \sqrt{p_{ind}(1-p_{ind})/n}\right]$, con $p_{ind} = p_u p_v + q_u q_v$. Infatti, sotto tale ipotesi, ogni elemento della distribuzione congiunta ottenuta a partire dalle serie considerate può essere pensato come una realizzazione indipendente di una variabile aleatoria di tipo bernulliano dove all'evento successo corrispondono, convenzionalmente, le due coppie di simboli uguali (1,1) e (0,0) con probabilità p_{ind}, mentre all'evento insuccesso sono associate le due coppie di simboli diversi (1,0) e (0,1) aventi probabilità $1 - p_{ind}$. Poiché il numero complessivo di successi conseguiti in n ripetizioni indipendenti di una variabile di bernoulli di probabilità p_{ind} si distribuisce secondo una variabile aleatoria binoc miale $\mathcal{B}in(n, p_{ind})$, per il terorema del limite centrale si ha che, per $n \to \infty$, $I_{crs} \sim \mathcal{N}\left[p_{ind}, \sqrt{p_{ind}(1-p_{ind})/n}\right]$.

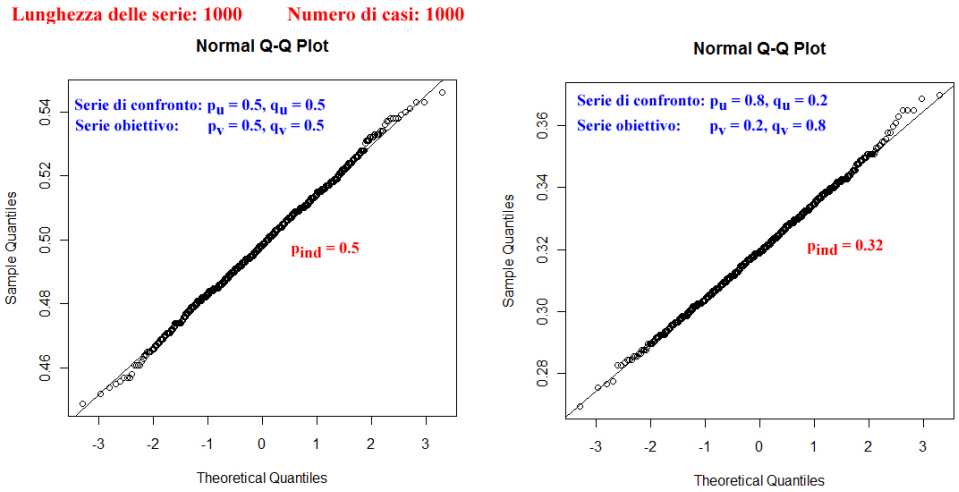

Figura 16.2: Distribuzione empirica dell'indice di corrispondenza per valori attesi dell'indice pari, rispettivamente, a 0.5 e 0.32. Serie composte da 1000 elementi.

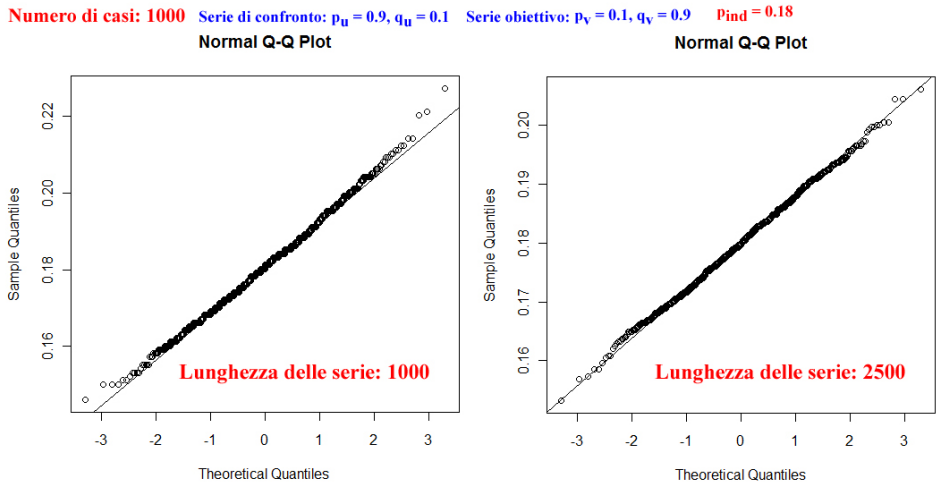

Figura 16.3: Distribuzione empirica dell'indice di corrispondenza, valore atteso pari a 0.18, in funzione del numero di elementi delle serie (1000 e 2500 elementi).

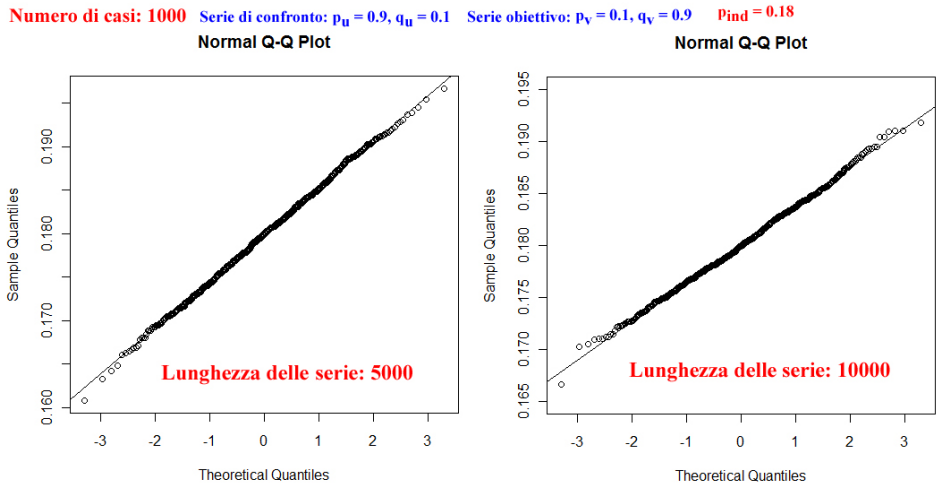

Figura 16.4: Distribuzione empirica dell'indice di corrispondenza, valore atteso pari a 0.18, in funzione del numero di elementi delle serie (5000 e 10000 elementi).

Nella figura 16.2 sono riportati i grafici del "qqplot"[2] dei valori empirici dell'indice di corrispondenza calcolato su due serie aventi distribuzioni marginali uguali fra loro e simmetriche ($p = q = 0.5$) e su due serie aventi distribuzioni marginali opposte e mediamente asimmetriche ($p_u = q_v = 0.8, q_u = p_v = 0.2$). Come si può vedere dai grafici, già con serie di 1000 elementi ciascuna, con le marginali considerate la distribuzione empirica dell'indice di corrispondenza approssima in maniera abbastanza soddisfacente la distribuzione normale.

Per serie aventi distribuzioni marginali maggiormente asimmetriche come ad esempio $p_u = q_v = 0.9, q_u = p_v = 0.1$, si può vedere (tabelle 16.4 e 16.4) che la convergenza è più lenta, ma che, comunque, già con 5000 elementi si ha una buona approssimazione della distribuzione normale.

16.3 L'indice di corrispondenza normalizzato

Per far si che l'indice di corrispondenza vari fra uno e meno uno ed assuma valore zero in corrispondenza della condizione di indipendenza, è possibile normalizzarlo in modo da ottene un indice standardizzato in senso assoluto o in senso relativo. Nella normalizzazione assoluta, l'indice sarà rapportato al suo massimo e minimo assoluti e quindi assumerà valore uno solo nel caso di perfetta correlazione diretta e meno uno nel caso di perfetta correlazione inversa. Esso esprimerà quindi, a partire dalla condizione di indipendenza

[2] Funzioni qqnorm(x) e qqline(x) di R

assunta come origine, la percentuale di corrispondenze (dirette o inverse) rispetto alla condizione di perfetta correlazione.

In simboli, indicando con I_{Acrs} l'indice di corrispondenza di ordine uno normalizzato in forma assoluta, si avrà:

$$\begin{cases} \text{se } I_{crs} - p_{ind} > 0, \text{ allora } I_{Acrs} = \frac{I_{crs} - p_{ind}}{1 - p_{ind}} \\ \text{se } I_{crs} - p_{ind} = 0, \text{ allora } I_{Acrs} = 0 \\ \text{se } I_{crs} - p_{ind} < 0, \text{ allora } I_{Acrs} = \frac{I_{crs} - p_{ind}}{p_{ind}} \end{cases}$$

Nella normalizzazione relativa invece, l'indice sarà rapportato al massimo ed al minimo relativi associati a quella particolare coppia di distribuzioni marginali caratterizzante le serie considerate. L'indice normalizzato in forma relativa assumerà quindi valore uno nel caso di massima correlazione diretta ammissibile e meno uno nel caso di massima correlazione inversa ammissibile. Esso esprimerà quindi, a partire dalla condizione di indipendenza assunta come origine, la percentuale di corrispondenze (dirette o inverse) rispetto alla condizione di massima correlazione ammissibile compatibilmente con le marginali date.

In simboli, indicando con I_{Rcrs} l'indice di corrispondenza di ordine uno normalizzato in forma relativa, si avrà:

$$\begin{cases} \text{se } I_{crs} - p_{ind} > 0, \text{ allora } I_{Rcrs} = \frac{I_{crs} - p_{ind}}{\min(p_u, p_v) + \min(q_u, q_v) - p_{ind}} \\ \text{se } I_{crs} - p_{ind} = 0, \text{ allora } I_{Rcrs} = 0 \\ \text{se } I_{crs} - p_{ind} < 0, \text{ allora } I_{Rcrs} = \frac{I_{crs} - p_{ind}}{p_{ind} - |p_u - q_v|} \end{cases}$$

Ovviamente, per calcolare l'indice normalizzato è necessario conoscere le distribuzioni marginali delle serie considerate o, comunque, una loro stima.

Nel proseguo della trattazione, quando si farà riferimento in maniera generica all'indice di corrispondenza normalizzato si intenderà sempre quello normalizzato in forma assoluta I_{Acrs}.

16.4 L'indice di corrispondenza associato a lag diversi da zero

Il procedimento precedentemente descritto, che ha portato alla definizione di $M_i^m(r)(v||u)$ e dell'indice di corrispondenza $I_{crs}^m = M^m(r)(v||u)$, può essere esteso in modo da considerare, nello stesso algoritmo, anche il caso di lag diversi da zero. Se si confrontano infatti i pattern di indice i della serie u con quelli di indice $i + lag$, con $lag = 0, \ldots, max\_lag$ della serie v, si possono ottenere delle quantità $M_{i,lag}^m(r)(v||u)$ e $M_{lag}^m(r)(v||u)$ che indicano, rispettivamente:

- se vi è (valore uno) o non vi è (valore zero) corrispondenza fra il generico pattern di posto i della serie di confronto u e il corrispondente pattern di posto $i + lag$ della serie obiettivo v;

- il numero totale di corrispondenze fra gli elementi che occupano il generico posto i nella serie di confronto e quelli che occupano il corrispondente posto $i + lag$ nella serie obiettivo.

Ciò può essere di particolare interesse nello studio dell'autocorrelazione, in associazione alla *ApEn*, o della correlazione qualora si possa supporre che serie storiche distinte possano presentare analogie di comportamento a meno di un determinato sfasamento temporale $lag = \Delta t$. Ovviamente $M^m_{lag}(r)(v||u) = M^m(r)(v||u)$ quando $lag = 0$.

Un approccio di questo tipo, ovvero il calcolo dell'indice di corrispondenza per lag da 1 a max_lag all'interno dello stesso algoritmo di calcolo della *cross-ApEn* consente un notevole risparmio di tempo rispetto al ricalcolo, ogni volta, della *cross-ApEn* (o della *ApEn*) e dell'indice di corrispondenza fra due serie dove la seconda sia stata scalata, di volta in volta, di un elemento (il secondo elemento diventa il primo, il terzo diventa il secondo, l'n-esimo diventa l'*n-1*esimo e così via). Tra l'altro, il calcolo della *cross-ApEn* o della *ApEn* in funzione del lag non rivestirebbe grande interesse in quanto, in media, tali indicatori non variano se gli elementi della serie obiettivo v vengono fatti scalare di una posizione verso sinistra.

Infatti, considerando un processo di scorrimento a sinistra quale quello riportato in figura 16.5, si può dimostrare quanto segue:

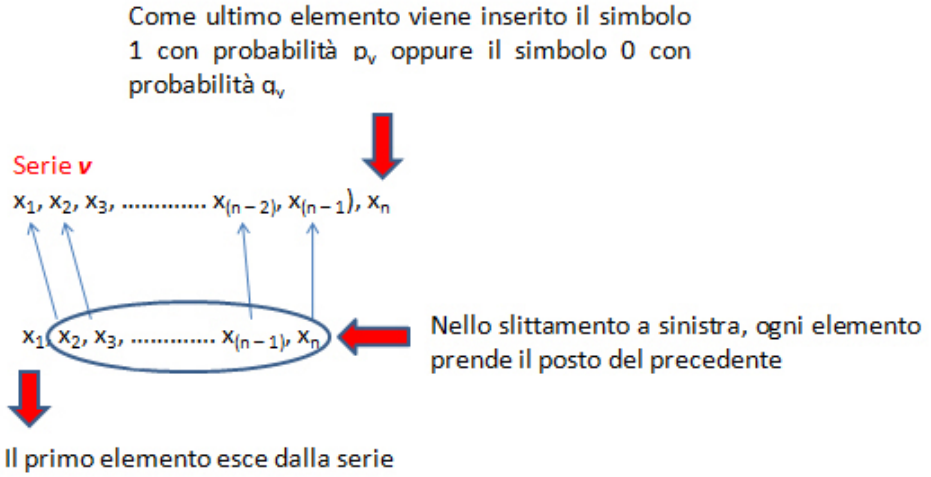

Figura 16.5: Schema di slittamento a sinistra.

Si consideri un generico elemento $B^m_i(r)(v||u) = k$ con $m = 1$ e si supponga poi di scalare la serie obiettivo v di lunghezza n di un elemento. L'elemento $B^1_i(r)(v||u)$ calcolato sugli $n-1$ elementi della serie scalata potrà assumere i seguenti valori:

- $k-1$ con probabilità $p_u p_v + q_u q_v$, nel caso in cui il primo elemento della serie obiettivo v originaria, che si è "perso" nel processo di slittamento, fosse uguale al primo elemento della serie di confronto u;

- k con probabilità $p_u q_v + p_v q_u$, nel caso in cui il primo elemento della serie obiettivo v originaria fosse diverso dal primo elemento della serie di confronto u;

con p_u e $q_u = 1 - p_u$ che indicano la probabilità di avere un 1 oppure uno 0 nella generica posizione i della serie di confronto u, mentre p_v e $q_v = 1 - p_v$ rappresentano le analoghe probabilità per la serie obiettivo v.

Si consideri ora l'n-esimo elemento aggiunto in coda alla serie scalata v, tale elemento potrà assumere valore 1 con probabilità p_v e 0 con probabilità q_v, di conseguenza esso potrà essere uguale all'n-esimo elemento della serie di confronto u con probabilità $p_u p_v + q_u q_v$ (coppie 1,1 oppure 0,0), oppure diverso con probabilità $p_u q_v + p_v q_u$ (coppie 1,0 oppure 0,1).

L'elemento $B_i^1(r)(v||u)$ calcolato su tutti gli n elementi della serie scalata v potrà quindi assumere i seguenti valori:

- $k-1$ con probabilità $(p_u p_v + q_u q_v)(p_u q_v + p_v q_u)$ se il primo elemento della serie v originaria, che si è "perso", era uguale al primo elemento della serie di confronto u, mentre il nuovo elemento inserito in coda alla serie v è diverso dal corrispondente n-esimo valore della serie u;

- k con probabilità $(p_u p_v + q_u q_v)^2$ se il primo elemento della serie v originaria, che si è "perso", era uguale al primo elemento della serie di confronto u, così come il nuovo elemento inserito in coda alla serie v è uguale al corrispondente n-esimo valore della serie u;

- k con probabilità $(p_u q_v + p_v q_u)^2$ se il primo elemento della serie v originaria, che si è "perso", era diverso dal primo elemento della serie di confronto u, così come il nuovo elemento inserito in coda alla serie v è diverso dal corrispondente n-esimo valore della serie u;

- $k+1$ con probabilità $(p_u q_v + p_v q_u)(p_u p_v + q_u q_v)$ se il primo elemento della serie v originaria, che si è "perso", era diverso dal primo elemento della serie di confronto u, mentre il nuovo elemento inserito in coda è uguale al corrispondente n-esimo valore della serie u.

Il valor medio di $B_i^1(r)(v||u)$ calcolato sulla serie scalata ($lag = 1$) sarà pertanto:

$$\begin{aligned} E[B_i^1(r)(v||u)] &= (k-1)(p_u p_v + q_u q_v)(p_u q_v + p_v q_u) + k(p_u p_v + q_u q_v)^2 + \\ &\quad + k(p_u q_v + p_v q_u)^2 + (k+1)(p_u q_v + p_v q_u)(p_u p_v + q_u q_v) \\ &= (k - 1 + k + 1)(p_u p_v + q_u q_v)(p_u q_v + p_v q_u) + \\ &\quad + k(p_u p_v + q_u q_v)^2 + k(p_u q_v + p_v q_u)^2 \end{aligned}$$

$$
\begin{aligned}
&= 2k(p_u p_v + q_u q_v)(p_u q_v + p_v q_u) + k(p_u p_v + q_u q_v)^2 \\
&\quad + k(p_u q_v + p_v q_u)^2 \\
&= k[(p_u p_v + q_u q_v) + (p_u q_v + p_v q_u)]^2 \\
&= k[(p_u p_v + p_u q_v) + (q_u q_v + p_v q_u)]^2 \\
&= k[p_u(p_v + q_v) + q_u(q_v + p_v)]^2 \\
&= k(p_u + q_u)^2 \\
&= k
\end{aligned}
$$

Quindi, in media, il valore di $B_i^1(r)(v||u)$ non cambia al variare del lag e questo per ciascun valore di i, di conseguenza rimane costante, in media, anche la quantità $B^1(r)(v||u)$. In maniera analoga si può dimostrare che, per $m = 2$, rimane costante anche la quantità $A^m(r)(v||u)$ e, di conseguenza, la *cross Sample-ApEn* e la *cross-ApEn*.

16.5 L'indice di corrispondenza come indicatore "stand alone"

Fino ad ora si è considerato l'indice di corrispondenza all'interno dell'algoritmo di calcolo della *cross-ApEn* e come "upgrade" di quest'ultima. Esso, tuttavia, ha un significato specifico quale indicatore di correlazione, gode di notevoli proprietà ed è caratterizzato da una propria distribuzione asintotica. L'indice di corrispondenza si presta pertanto ad essere considerato e calcolato anche in maniera indipendente e distinta dalla *cross-ApEn*. Inoltre, la richiesta di tempo macchina e di risorse necessarie al calcolo della *cross-ApEn* (e della *ApEn* in generale) cresce in maniera esponenziale al crescere della lunghezza delle serie considerate e del valore del parametro m, mentre per l'indice di corrispondenza, così come per gli altri indicatori considerati nella presente tesi, tali necessità crescono in maniera lineare.

In particolare, nel caso binario e per $m = 1$, il calcolo dell'indice di corrispondenza risulta estremamente semplice ed altamente performante. Se N indica la lunghezza delle serie considerate e U e V sono gli array di lunghezza N che contengono, rispettivamente, gli elementi binari della serie di confronto u e della serie obiettivo v allora, in R, l'algoritmo di calcolo dell'indice di corrispondenza di ordine 1 si riduce a:

$$I_{crs} \leftarrow sum(!xor(U + V))/N$$

CAPITOLO 17

Analisi di serie perfettamente periodiche affette da rumore

Per valutare il comportamento dell'indice di corrispondenza relativamente a condizioni e fenomeni già studiati, si è applicato lo I_{crs} all'esperimento di cui al precedente capitolo 12.

In particolare, si sono considerati due gruppi di serie caratterizzate da una medesima condizione iniziale, ovvero serie perfettamente periodiche di periodo due e pattern "01". Ciascun gruppo era formato da 1000 serie (casi) generate dal calcolatore in maniera indipendente l'una dall'altra, mentre ciascuna serie era composta da 1000 elementi.

L'esperimento consisteva nel confrontare le serie del primo gruppo, dette serie di confronto e caratterizzate, di volta in volta, da un determinato livello di perturbazione, con le serie del secondo gruppo, dette serie obiettivo, a cui sono stati successivamente applicati tutti i livelli di perturbazione a partire dalla condizione di perfetta regolarità (livello di perturbazione uguale a zero) fino a quella di massima casualità (livello di perturbazione pari al 100%).

L'operazione di perturbazione comportava la sostituzione di una determinata percentuale di elementi delle serie, scelti a caso, con altrettanti valori 0 o 1 generati in maniera casuale con probabilità pari a 0.5.

Operando in questo modo si sono ottenuti venticinque diversi gruppi di confronti fra loro indipendenti (cinque livelli di perturbazione per le serie di confronto associati ai cinque livelli di perturbazione previsti per le serie obiettivo) le cui serie erano sempre caratterizzate da una frequenza media di 0 e di 1 pari all'incirca a 0.5.

Nella tabella 17.1 sono riportati i valori dell'indice di corrispondenza, mentre nella tabella 17.2 sono riportati i valori dell'indice di corrispondenza normalizzato in forma assoluta.

Tabella 17.1: Valori dell'indice di corrispondenza e relativi intervalli di confidenza al 95% riferiti a due serie di periodo 2 in funzione di diversi valori di randomizzazione caratterizzati da $p = q = 0.5$.

Serie sottostanti di periodo due e pattern 01 Lunghezza delle serie: 1000 Numero di casi: 1000

Liv. di perturbazione serie obiettivo	Livello di perturbazione serie di confronto									
	0%		25%		50%		75%		100%	
0%	1	1 1	0,874586	0,857858 0,889890	0,749957	0,727728 0,772773	0,625490	0,598599 0,653654	0,500109	0,469469 0,531532
25%	0,874576	0,857858 0,889890	0,781337	0,760736 0,802803	0,687052	0,661662 0,711737	0,594138	0,565541 0,621647	0,500116	0,467467 0,532533
50%	0,749877	0,726727 0,771772	0,687771	0,664640 0,710736	0,625193	0,597573 0,651652	0,562489	0,532508 0,592593	0,499821	0,468468 0,531532
75%	0,625916	0,599575 0,652678	0,593444	0,561537 0,619620	0,563075	0,535536 0,592593	0,531110	0,502478 0,561587	0,500366	0,470445 0,531532
100%	0,500471	0,470445 0,530556	0,500930	0,469469 0,532533	0,499892	0,465465 0,531557	0,499969	0,468468 0,528529	0,500397	0,469444 0,530556

In una situazione quale quella considerata, essendo le distribuzioni marginali simmetriche e uguali fra loro ($p = q = 0.5$), i valori dell'indice di corrispondenza normalizzato in forma assoluta e in forma relativa coincidono, in quanto il minimo ed il massimo relativo risultano pari, rispettivamente, a zero e ad uno, mentre il valore caratteristico associato alla condizione di indipendenza, p_{ind}, vale 0.5.

Nella tabella 17.2 si può osservare come i valori dell'indice di corrispondenza normalizzato risultino simmetrici rispetto alla diagonale principale e come il valore dell'indice diminuisca all'aumentare del livello di randomizzazione, ovvero man mano che ci si sposta verso il basso e verso destra nella tabella. In particolare, si può notare come il valore relativo al confronto fra le due serie non perturbate, e quindi uguali (intersezione fra la prima riga e la prima colonna), assuma valore uno ad indicare perfetta correlazione. L'ultima riga e l'ultima colonna presentano invece valori praticamente coincidenti fra loro e comunque molto prossimi a zero, ovvero alla condizione di indipendenza. Ciò concordemente al fatto che almeno una delle due serie, pur conservando la distribuzione marginale d'origine, è di tipo completamente casuale. Si può inoltre rilevare come anche il valore relativo a serie entrambe perturbate al 75% non si discosti significativamente dalla condizione di massima indipendenza.

In definitiva, nella particolare situazione analizzata, l'indice di corrispondenza, pur assumendo valori diversi e diversamente scalati, presenta un com-

Tabella 17.2: Valori dell'indice di corrispondenza normalizzato e relativi intervalli di confidenza al 95% riferiti a due serie di periodo 2 in funzione di diversi valori di randomizzazione caratterizzati da $p = q = 0.5$.

Serie sottostanti di periodo due e pattern 01 Lunghezza delle serie: 1000 Numero di casi: 1000

Liv. di perturbazione serie obiettivo		Livello di perturbazione serie di confronto									
		0%		25%		50%		75%		100%	
0%	1	1 1	0,749172	0,715716 0,779780	0,499914	0,455456 0,545546	0,250980	0,197198 0,307308	0,000218	-0,061062 0,063064
25%	0,749152	0,715716 0,779780	0,562674	0,521472 0,605606	0,374104	0,323324 0,423474	0,188276	0,131082 0,243294	0,000232	-0,065066 0,065066
50%	0,499754	0,453454 0,543544	0,375542	0,329280 0,421472	0,250386	0,195146 0,303304	0,124978	0,065016 0,185186	-0,000358	-0,063064 0,063064
75%	0,251832	0,199150 0,305356	0,186888	0,123074 0,239240	0,126150	0,071072 0,185186	0,062220	0,004956 0,123174	0,000732	-0,059110 0,063064
100%	0,000942	-0,059110 0,061112	0,001860	-0,061062 0,065066	-0,000216	-0,069070 0,063114	-0,000062	-0,063064 0,057058	0,000794	-0,061112 0,061112

portamento tendenziale molto simile a quanto già osservato per la S_ρ e la mutua informazione

E' interessante infine osservare come i valori dell'indice di corrispondenza relativi alla prima riga ed alla prima colonna, ottenuti dal confronto fra una serie perfettamente periodica e la stessa serie a cui sono stati via via applicati tutti i livelli di randomizzazione, corrispondano al complemento ad uno della percentuale di randomizzazione applicato. In questo caso, quindi, l'indice rileva esattamente, in termini percentuali, l'incidenza del livello di rumore a partire da una condizione di perfetta regolarità.

17.1 Valori dell'indice di corrispondenza in funzione del *lag*

Nella tabella 17.3 sono riportati i valori dell'indice di corrispondenza normalizzato, per *lag* da 0 a 10, relativi al confronto fra una serie perfettamente periodica di periodo due e pattern "01" (serie di confronto) e la stessa serie a cui sono stati via via applicati tutti i livelli di randomizzazione.

Come si può vedere, in assenza di randomizzazione, l'indice di corrispondenza normalizzato assume:

- valore 1 in corrispondenza dei lag pari, situazione per cui vi è perfetta corrispondenza fra tutti gli elementi delle due serie, e quindi perfetta correlazione;

- valore -1, ovvero perfetta correlazione inversa, in corrispondenza dei lag dispari dove, dato il particolare disegno delle serie, ogni elemento è sempre associato al suo complementare.

Tabella 17.3: Valori dell'indice di corrispondenza normalizzato in funzione del lag: confronto fra una serie perfettamente periodica e la stessa serie a cui sono stati applicati successivamente tutti i livelli di randomizzazione.

Serie sottostanti di periodo due e pattern 01 — Lunghezza delle serie: 1000 — Numero di casi: 1000

Serie obiettivo: livello di randomizzazione	Serie di confronto (non perturbata)										
	Lag										
	0	1	2	3	4	5	6	7	8	9	10
	Valori dell'Indice di corrispondenza normalizzato in forma assoluta										
0%	1	-1	1	-1	1	-1	1	-1	1	-1	1
25%	0,750402	-0,750364	0,750354	-0,750308	0,750298	-0,750340	0,750318	-0,750330	0,750326	-0,750360	0,750374
50%	0,500032	-0,500038	0,500024	-0,500002	0,499962	-0,499940	0,499940	-0,500016	0,500064	-0,500042	0,500054
75%	0,249642	-0,249646	0,249644	-0,249592	0,249616	-0,249652	0,249660	-0,249622	0,249662	-0,249704	0,249620
100%	-0,002328	0,002336	-0,002270	0,002262	-0,002228	0,002210	-0,002192	0,002262	-0,002236	0,002272	-0,002342

L'indice di corrispondenza evidenzia inoltre chiaramente la diminuzione del livello di sincronia fra le due serie, in origine uguali (lag pari) o perfettamente complementari (lag dispari), all'aumentare del livello di disordine indotto nella serie obiettivo. Dai valori 1 o -1, infatti, si arriva progressivamente a zero, valore che esprime la condizione di indipendenza, in corrispondenza del livello di randomizzazione del 100%.

Osservando le righe della tabella in corrispondenza dei livelli di randomizzazione compresi fra il 25 ed il 75%, si può invece osservare come uno stesso valore si ripeta ogni due lag, con valori positivi per i lag pari (correlazione diretta) e negativi per quelli dispari (correlazione inversa).

In questo caso, quindi, l'indice di corrispondenza evidenzia un comportamento di fondo tendenzialmente periodico di periodo due la cui evidenza va via via scemando al crescere del livello di rumore. Lo stesso comportamento, anche se qui non evidenziato, si riscontra anche per l'indice di corrispondenza del secondo e del terzo ordine ($m = 2$ ed $m = 3$).

17.2 Confronti fra serie dello stesso tipo uguali e diverse

Nella tabella 17.4 sono riportati i valori dell'indice di corrispondenza ($lag = 0$) normalizzato in forma assoluta relativi a confronti fra serie fra loro uguali

(prima colonna) e serie fra loro diverse (seconda colonna) a parità di livello di randomizzazione e di serie di origine (serie perfettamente periodiche di periodo due e pattern "01").

Tabella 17.4: Valori dell'indice di corrispondenza normalizzato nel confronto fra coppie di serie uguali e serie diverse a parità di livello di randomizzazione e di serie di origine.

Serie sottostanti di periodo due e pattern 01
Lunghezza delle serie: 1000 Numero di casi: 1000

Livello di randomizzazione	Serie uguali	Serie diverse
0%	1	1
25%	1	0,562170
50%	1	0,249976
75%	1	0,062592
100%	1	0,002134

Come si può chiaramente osservare, nel caso di confronti fra serie uguali l'indice di corrispondenza assume sempre valore uno a prescindere dal livello di randomizzazione applicato, mentre assume valori distinti e proporzionali al numero di elementi corrispondenti uguali nel caso di serie fra loro diverse. Com'era prevedibile, esso è quindi in grado, a differenza della *cross Sample-ApEn*, di distinguere, a parità di distribuzioni marginali, fra serie uguali e serie diverse.

capitolo 18

Confronto fra serie perfettamente periodiche di disegno diverso

Si è poi ripetuto con l'indice di corrispondenza l'esperimento di cui al precedente capitolo 13 ponendo a confronto fra loro le serie perfettamente periodiche di disegno diverso già considerate nel capitolo 5.

Nella tabella 18.1 sono riportati i valori dell'indice di corrispondenza non normalizzato relativo ai confronti fra le serie considerate. Come si può ben vedere, i valori situati sulla diagonale principale, e quindi riferiti ad una condizione di perfetta correlazione (confronto di una serie con se stessa), sono tutti pari ad uno, ad indicare, correttamente, completa corrispondenza fra tutti gli elementi delle serie. Si può inoltre notare come i valori posizionati sulla seconda colonna e sulla seconda riga, ad eccezione di quelli posti sulla diagonale principale, siano tutti approssimatamente pari a 0.5, ovvero al valore corrispondente alla condizione di indipendenza, $p_{ind} = p_u p_v + q_u q_v$, date le particolari distribuzioni marginali, così come evidenziato anche dalla S_ρ e dalla mutua informazione. Infine, a meno di approssimazioni i valori relativi alle celle in grigio corrispondono al massimo ammissibile dell'indice di corrispondenza. Tale massimo è sempre minore o uguale ad uno ed è pari a $\min(p_u, p_v) + \min(q_u, q_v)$, mentre il minimo, sempre maggiore o uguale a zero, è pari a $|p_u - q_v|$. Il massimo assoluto, pari ad uno, è raggiungibile unicamente nel caso di serie caratterizzate entrambe da una distribuzione marginale simmetrica $p = q = 0.5$, mentre il minimo assoluto, pari a zero, si può raggiungere solo nel caso di serie aventi distribuzioni marginali opposte $p_u = q_v$ e $q_u = p_v$.

Tabella 18.1: Valori dell'indice di corrispondenza non normalizzato: confronto fra serie perfettamente periodiche aventi disegno diverso.

Lunghezza delle serie: 1000

		Serie di confronto				
		'01'	'0011'	'0000011111'	'0000000111'	'0000000001'
Serie obiettivo	'01'	1	0,500000	0,600402	0,599398	0,599398
	'0011'	0,500000	1	0,500000	0,498996	0,498996
	'0000011111'	0,600402	0,500000	1	0,800201	0,601406
	'0000000111'	0,599398	0,498996	0,800201	1	0,801205
	'0000000001'	0,599398	0,498996	0,601406	0,801205	1

Se si normalizzano in forma assoluta i valori dell'indice di corrispondenza di cui alla precedente tabella 18.1, essa assume l'aspetto seguente:

Tabella 18.2: Valori dell'indice di corrispondenza normalizzato: confronto fra serie perfettamente periodiche aventi disegno diverso.

Lunghezza delle serie: 1000

		Serie di confronto				
		'01'	'0011'	'0000011111'	'0000000111'	'0000000001'
Serie obiettivo	'01'	1	0,000000	0,200804	0,198796	0,198796
	'0011'	0,000000	1	0,000000	-0,002008	-0,002008
	'0000011111'	0,200804	0,000000	1	0,600402	0,202812
	'0000000111'	0,198796	-0,002008	0,600402	1	0,415309
	'0000000001'	0,198796	-0,002008	0,202812	0,415309	1

18.1 Confronto fra serie indipendenti

L'indice di corrispondenza è stato poi applicato a quattro diverse coppie di serie indipendenti caratterizzate sempre dalle stesse distribuizioni marginali (vedi precedente capitolo 13.2). In particolare:

- *prima coppia*: entrambe le serie sono periodiche di periodo 100 e costruite in modo da risultare fra loro indipendenti. Infatti, il disegno della serie di confronto è costituito da 70 zeri consecutivi seguiti da 30 uno ($p_u = 0.3$ e $q_u = 0.7$), mentre quello della serie obiettivo è formato da 28 zeri seguiti da 42 uno, poi da 12 zeri seguiti da 18 uno ($p_v = 0.6$ e $q_v = 0.4$);
- *seconda coppia*: la serie di confronto, periodica di periodo 100, è la stessa della prima coppia, mentre la serie obiettivo è una serie completamente casuale avente distribuzione marginale $p_v = 0.6$ e $q_v = 0.4$;
- *terza coppia*: la serie di confronto è una serie completamente casuale avente distribuzione marginale $p_u = 0.3$ e $q_v = 0.7$, mentre la serie obiettivo, periodica di periodo 100, è la stessa della prima coppia;
- *quarta coppia*: entrambe le serie sono completamente casuali ed aventi distribuzione marginale, rispettivamente, $p_u = 0.3$, $q_u = 0.7$ e $p_v = 0.6$ e $q_v = 0.4$.

Tabella 18.3: Confronti fra serie fra loro indipendenti in funzione del tipo di serie in ingresso.

		Serie di confronto $p_u = 0.3$, $q_u = 0.7$		Serie obiettivo $p_v = 0.6$, $q_v = 0.4$	
		Sρ	Mutua informazione	*Cross Sample-ApEn*	Indice di corrispondenza
Serie confrontate	Periodica vs Periodica	0 (0, 0)	0 (0, 0)	0,09208 (0,092081, 0,092081)	0 (0, 0)
	Periodica vs Casuale	0,000124 (0, 0,000621)	0,000713 (0,000001, 0,003592)	1,069765 (0,970920, 1,168326)	-0,000046 (-0,029130, 0,024833)
	Casuale vs Periodica	0,000123 (0, 0,000614)	0,000710 (0,000001, 0,003528)	0,896609 (0,814109, 0,983641)	-0,000557 (-0,028261, 0,021852)
	Casuale vs Casuale	0,000130 (0, 0,000598)	0,000749 (0,000001, 0,003458)	1,120107 (1,076882, 1,170320)	0,000322 (-0,029565, 0,026296)

Come si può vedere dalla tabella 18.3, l'indice di corrispondenza normalizzato in forma assoluta, analogamente alla $S\rho$ ed alla mutua informazione, assume in tutti e quattro i casi un valore molto prossimo a zero ad indicare,

correttamente, indipendenza fra le serie considerate, cosa che, come abbiamo già avuto modo di vedere, non è invece in grado di fare la *cross Sample-ApEn*.

18.2 Confronto fra serie perfettamente complementari

Analogamente a quanto effettuato con gli altri indicatori, si è eseguito un ulteriore esperimento confrontando fra loro quattro gruppi di serie perfettamente complementari (ad uno zero nella serie di confronto corrisponde sempre un uno nella serie obiettivo e viceversa) aventi disegno e distribuzioni marginali diverse da gruppo a gruppo. Dai risultati ottenuti, riassunti nella tabella 18.4, si nota immediatamente che l'indice di corrispondenza normalizzato in forma assoluta assume valore -1, corrispondente alla condizione di perfetta correlazione inversa, per tutti i confronti effettuati.

Tabella 18.4: Confronto fra serie binarie perfettamente complementari.

Lunghezza delle serie: 1000

	Sp	Mutua informazione	cross Sample APEN	Indice di corrispondenza normalizzato
'0101010101' vs '1010101010'	0,292893	1	0	-1
'1011001100' vs '0100110011'	0,292893	1	0,943869	-1
'0000011111' vs '1111100000'	0,292893	1	0,555361	-1
'0000000111' vs '1111111000'	0,250021	0,881291	0,691878	-1
'0000000001' vs '1111111110'	0,114562	0,468996	3,181222	-1

In definitiva, quindi, nel caso di serie binarie l'indice di corrispondenza è l'unico indicatore fra quelli considerati in grado di evidenziare in maniera chiara ed univoca sia la condizione di perfetta correlazione diretta (serie uguali) che inversa (serie complementari). La S_ρ e la mutua informazione, infatti, non distinguono fra correlazione diretta e correlazione inversa: i valori di cui alla tabella 18.4 coincidono infatti con i valori assunti da tali indicatori nel caso del confronto di due serie uguali aventi, indifferentemente, disegno pari a quello della serie di confronto o a quello della serie obiettivo, mentre la *cross Sample-ApEn* assume i valori più diversi risultando, in questo caso, assolutamente non informativa.

Tramite l'indice di corrispondenza è inoltre possibile verificare la condizione di indipendenza fra due serie (o fra gli elementi di una stessa serie). In questo caso, però, la conoscenza del solo valore dell'indicatore non è più sufficiente, dovendosi confrontare il valore empirico ottenuto per mezzo dei dati sperimentali con il valore teorico di riferimeno p_{ind} o con una sua stima. A tal fine è quindi necessario disporre delle distribuzioni marginali delle serie considerate, o di una loro stima, in modo da poter calcolare in via analitica la distribuzione dell'indice di corrispondenza sotto l'ipotesi nulla di indipendenza. In alternativa è comunque possibile ricavare la distribuzione empirica dell'indice di corrispondenza ricorrendo ai metodi Monte Carlo.

Ciò non di meno, l'indice di corrispondenza ha una validità sia di per se, quale mezzo per investigare il livello di correlazione fra le serie o all'interno di una stessa serie, sia in associazione alla *cross-ApEn* o alla *cross Sample-ApEn*, dove può essere integrato nell'algoritmo di calcolo in maniera efficace ed efficiente.

CAPITOLO 19

Analisi di serie semplici perfettamente periodiche in funzione della lunghezza del periodo

Si considerino ora le serie seguenti caratterizzate da una alternanza di zeri e di uno costanti e così strutturate:

- *serie 1*: 500 zero e poi 500 uno;
- *serie 2*: 250 zero e poi 250 uno;
- *serie 3*: 100 zero e poi 100 uno;
- *serie 4*: 50 zero e poi 50 uno;
- *serie 5*: 25 zero e poi 25 uno;
- *serie 6*: 10 zero e poi 10 uno;
- *serie 7*: 5 zero e poi 5 uno.

Tali serie sono perfettamente periodiche ed esprimono tutte, sia al limite che per la lunghezza considerata (1000 elementi), una proporzione di zero, q, e di uno, p, costante e pari a 0.5. Esse sono inoltre caratterizzate dallo stesso livello di complessità algoritmica, come già dimostrato nel capitolo 6.

Tabella 19.1: Serie perfettamente periodiche: valori dell'indice di corrispondenza normalizzato in funzione della lunghezza del periodo.

p = q = 0.5, Lunghezza della serie: 1000

	Alternanza di zeri e di uno						
Lag	500	250	100	50	25	10	5
0	1	1	1	1	1	1	1
1	0,996	0,992	0,980	0,960	0,920	0,800	0,600
2	0,992	0,984	0,960	0,920	0,840	0,600	0,200
3	0,988	0,976	0,940	0,880	0,760	0,400	-0,200
4	0,984	0,968	0,920	0,840	0,680	0,200	-0,600
5	0,980	0,960	0,900	0,800	0,600	0	-1
6	0,976	0,952	0,880	0,760	0,520	-0,200	-0,600
7	0,972	0,944	0,860	0,720	0,440	-0,400	-0,200
8	0,968	0,936	0,840	0,680	0,360	-0,600	0,200
9	0,964	0,928	0,820	0,640	0,280	-0,800	0,600
10	0,960	0,920	0,800	0,600	0,200	-1	1

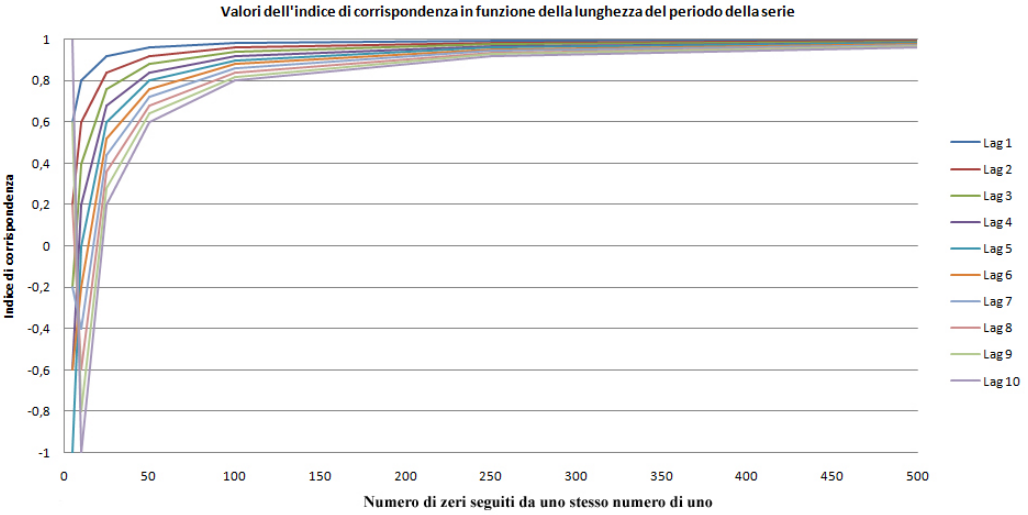

Figura 19.1: Valori dell'indice di corrispondenza normalizzato in funzione della lunghezza del periodo.

Come si può vedere dai valori normalizzati riportati nella tabella 19.1, in corrispondenza di un determinato livello di periodicità le differenze fra i valori dell'indice calcolate fra due lag consecutivi risultano costanti. In particolare si può osservare una differenza pari a 4 nel caso della serie 1, 8 per la serie 2, 20 per la serie 3 e così via. Un comportamento di questo tipo non può che far pensare ad una struttura delle serie regolare e periodica tanto che, in questo caso particolare ($p = q = 0.5$), dal valore di tali differenze è possibile risalire al periodo delle serie stesse. Infatti, nel caso specifico si ha:

$$periodo \ = 2/\Delta I_{Acrs} \ fra \ due \ lag \ consecutivi,$$

dove la costante 2 corrisponde al fattore di scala utilizzato nel normalizzare in forma assoluta l'indice di corrispondenza in modo che esso possa variare nell'intervallo $[1, -1]$ invece che fra $[1, 0]$.

A parità di lag, invece, si assiste ad una diminuzione proporzionale dell'indice di corrispondenza, e quindi della correlazione, al diminuire della lunghezza del periodo delle serie in quanto, man mano che si scorre la serie obiettivo verso sinistra, aumenta necessariamente il numero di mancate corrispondenze.

Nel caso delle serie 6 e 7, il cui periodo è pari o inferiore al numero di lag considerati, si può inoltre notare quanto segue:

- per la serie 6, di periodo 10, l'indice di corrispondenza si annulla (condizione di indipendenza) in corrispondenza del lag 5 per poi diventare negativo e raggiungere il valore -1 in corrispondenza del lag 10, dove ogni elemento è confrontato con il suo complementare (condizione di perfetta correlazione inversa);

- per la serie 7, di periodo 5, l'indice di corrispondenza dapprima diminuisce fino a raggiungere il valore -1 in corrispondenza del lag 5, senza però passare per il valore 0, e poi ricomincia a crescere per tornare al valore 1 in corrispondenza del lag 10, quando ogni elemento è nuovamente confrontato con se stesso.

In questo caso l'indice di corrispondenza è quindi in grado di fornirci informazioni abbastanza precise riguardo la struttura della serie e la lunghezza del periodo. Non è invece risultato di alcuna utilità, al pari della S_ρ, della mutua informazione e della $ApEn$, al fine di discriminare le serie in base al loro grado di complessità algoritmica. Tali serie sono infatti caratterizzate da uno stesso livello di complessità, mentre l'indice di corrispondenza assume valori distinti per ciascuna serie a prescindere dal lag considerato.

19.1 Serie caratterizzate da lunghe sequenze di elementi uguali

Per cercare di meglio comprendere il comportamente dell'indice di corrispondenza in presenza di serie caratterizzate da lunghe sequenze di elementi uguali e distribuzioni marginali asimmetriche, si considerino ora le serie seguenti:

- *serie 1*: 500 zero e poi 500 uno;
- *serie 2*: 750 zero e poi 250 uno;
- *serie 3*: 900 zero e poi 100 uno;
- *serie 4*: 950 zero e poi 50 uno;
- *serie 5*: 975 zero e poi 25 uno;
- *serie 6*: 990 zero e poi 10 uno;
- *serie 7*: 995 zero e poi 5 uno;
- *serie 8*: 999 zero e poi 1 uno.

Queste serie, a differenza delle precedenti, sono caratterizzate da una percentuale di zeri e di uno variabile da serie a serie e, di conseguenza, da differenti distribuzioni marginali. Da un punto di vista della complessità algoritmica sono però ancora equivalenti fra loro e, sostanzialmente, anche alle serie precedenti.

Nella tabella 19.2 sono riportati i valori dell'indice di corrispondenza non normalizzato. Come si può vedere, a parità di lag l'indice assume lo stesso valore per ciascuna delle serie da 1 a 6. La serie 7, invece, condivide con le serie precedenti i valori dell'indice calcolati in corrispondenza dei lag da 1 a 5, dopo di che tale valore diventa costante e pari 0.990. La serie 8, infine, condivide con le altre il solo valore relativo al primo lag 1, dato che, poi, il valore dell'indice rimane costante e pari a 0.998 per tutti i valori del lag da 1 a 10.

Questo comportamento è perfettamente corretto. L'indice di corrispondenza non normalizzato esprime infatti la percentuale di elementi coincidenti nelle due serie confrontate, percentuale che, nel caso di serie uguali ($lag = 0$), è pari al 100%. Nelle serie in esame, all'aumentare del lag, ovvero man mano che si fa scorrere la serie obiettivo verso sinistra inserendovi uno zero in coda, aumenta di pari passo anche il numero di uno della serie obiettivo che vengono ad essere associati agli zeri della serie di confronto determinando così una proporzionale diminuzione del valore dell'indice di corrispondenza.

Questo almeno nel caso di serie in cui il numero di uno consecutivi è pari o superiore al numero massimo di lag considerati. Nel caso delle serie

7 ed 8, infatti, l'indice di corrispondenza diminuisce fin tanto che il numero di uno consecutivi risulta inferiore o uguale al valore del lag, dopo di che esso rimane costante in quanto ormai tutti gli uno si sono accoppiati con degli zero, motivo per cui un ulteriore slittamento verso sinistra, e quindi l'inserimento di altro zero in coda (uno zero entra, uno zero esce), determina solamente una variazione nella posizione delle coppie di elementi non uguali ma non del loro numero. Cosa che si verifica a partire dal lag 5 per la serie 7, il cui disegno vede 995 zeri consecutivi seguiti da 5 uno, ed al lag 1 per la serie 8, composta da 999 zeri consecutivi seguiti da un solo uno. A partire da tali lag, i componenti delle serie 7 ed 8 risultano essere fra loro indipendenti ed incorrelati, e ciò è attestato da fatto che, per ciascuna serie, il valore assunto dall'indice di corrispondenza coincide con quello tipico della condizione di indipendenza, che dipende dalle distribuzioni marginali, ed è riportato nell'ultima riga (in azzurro) della stessa tabella.

Tabella 19.2: Serie perfettamente periodiche: valori dell'indice di corrispondenza non normalizzato in funzione del numero di zeri consecutivi in serie di 1000 elementi.

Indice di corrispondenza non normalizzato

	Numero di zeri consecutivi su 1000 elementi							
Lag	500	750	900	950	975	990	995	999
0	1	1	1	1	1	1	1	1
1	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998
2	0,996	0,996	0,996	0,996	0,996	0,996	0,996	0,998
3	0,994	0,994	0,994	0,994	0,994	0,994	0,994	0,998
4	0,992	0,992	0,992	0,992	0,992	0,992	0,992	0,998
5	0,990	0,990	0,990	0,990	0,990	0,990	0,990	0,998
6	0,988	0,988	0,988	0,988	0,988	0,988	0,990	0,998
7	0,986	0,986	0,986	0,986	0,986	0,986	0,990	0,998
8	0,984	0,984	0,984	0,984	0,984	0,984	0,990	0,998
9	0,982	0,982	0,982	0,982	0,982	0,982	0,990	0,998
10	0,980	0,980	0,980	0,980	0,980	0,980	0,990	0,998
p_{ind}	0,5	0,625	0,82	0,905	0,95125	0,9802	0,99005	0,998002
min I_{crs}	0	0,5	0,8	0,9	0,95	0,98	0,99	0,998

Nella tabella 19.3 sono invece riportati i valori dell'indice di corrispondenza normalizzato in forma assoluta. Come si può vedere, esso diminuisce man mano che ci si sposta verso la destra della tabella, ovvero man mano che il numero di uno consecutivi sul totale degli elementi della serie diminuisce, mentre aumenta il valore del fattore di normalizzazione p_{ind}. Tale diminuzione risulta più evidente all'aumentare del lag. L'indice di corrispondenza assume inoltre un valore molto prossimo a zero (condizione di indipendenza) in corrispondenza del lag 10 per la serie 6, a partire dal lag 5 per la serie

7 e dal lag 1 per la serie 8, in accordo a quanto già osservato nella tabella precedente nel caso dell'indice non normalizzato.

Se si considera invece l'andamento dell'indice in funzione del lag, si può vedere come le differenze fra i valori calcolati su una stessa serie e riferiti a due lag consecutivi rimangano costanti (vedi i valori riportati nell'ultima riga, in azzurro, della tabella 19.3) almeno fino a che non si raggiunge, eventualmente, la condizione di indipendenza, dopo di che tali differenze, ovviamente, si annullano.

Come già nel caso delle serie semplici perfettamente periodiche studiate precedentemente, un tale comportamento lascia intravvedere una regolarità di fondo molto forte nella struttura della serie anche se, in questo caso, risulta assai meno evidente il nesso esistente fra valore dell'indice e disegno del periodo.

Tabella 19.3: Serie perfettamente periodiche: valori dell'indice di corrispondenza normalizzato in forma assoluta in funzione del numero di zeri consecutivi in serie di 1000 elementi.

Indice di corrispondenza normalizzato in forma assoluta

	Numero di zeri consecutivi su 1000 elementi							
Lag	500	750	900	950	975	990	995	999
0	1	1	1	1	1	1	1	1
1	0,996	0,994667	0,988889	0,978947	0,958974	0,898990	0,798995	-0,000002
2	0,992	0,989333	0,977778	0,957895	0,917949	0,797980	0,597990	-0,000002
3	0,988	0,984000	0,966667	0,936842	0,876923	0,696970	0,396985	-0,000002
4	0,984	0,978667	0,955556	0,915789	0,835897	0,595960	0,195980	-0,000002
5	0,980	0,973333	0,944444	0,894737	0,794872	0,494949	-0,000051	-0,000002
6	0,976	0,968000	0,933333	0,873684	0,753846	0,393939	-0,000051	-0,000002
7	0,972	0,962667	0,922222	0,852632	0,712821	0,292929	-0,000051	-0,000002
8	0,968	0,957333	0,911111	0,831579	0,671795	0,191919	-0,000051	-0,000002
9	0,964	0,952000	0,900000	0,810526	0,630769	0,090909	-0,000051	-0,000002
10	0,960	0,946667	0,888889	0,789474	0,589744	-0,000204	-0,000051	-0,000002
ΔI_{Acrs}	0,0040	0,005333	0,011111	0,021053	0,041026	0,101010	0,201005	0

Anche in questo contesto l'indice di corrispondenza normalizzato presenta un andamento abbastanza simile a quello della S_ρ e della mutua informazione, pur presentando variazioni più contenute e maggiormente proporzionali al livello di autocorrelazione effettivamente presente. Anch'esso, inoltre, non appare in grado di caratterizzare le serie considerate in ragione del loro livello di complessità algoritmica.

CAPITOLO 20

Analisi di serie deterministiche complesse periodiche e non periodiche

In analogia agli esperimenti di cui al precedente capitolo 7, si considerino adesso le due serie periodiche:

- *serie 1*: di periodo 10 e pattern "1100111000";
- *serie 2*: di periodo 20 e pattern "11001110000001110011";

e le due serie non periodiche:

- *serie 3*: di pattern "100111000011111..."
- *serie 4*: di pattern "101100111000...";

Per n che tende all'infinito tutte e quattro le serie esprimono una proporzione di zero, q, e di uno, p, costante e pari a 0.5. Per la lunghezza considerata (1000 elementi), invece, le prime due serie sono caratterizzate da una proporzione di zeri e di uno perfettamente uguale ($p = q = 0.5$), mentre la serie 3 è caratterizzata da $p = 0.494$ e la serie 4 da $p = 0.504$.

Da un punto di vista della complessità algoritmica, la serie 1 può essere considerata come la più semplice, seguita dalla 2, il cui periodo, di lunghezza doppia, è composto dalla sequenza che caratterizza il periodo della serie 1 seguita dalla medesima sequenza riscritta, però, in ordine inverso. Si può

poi senz'altro affermare che la serie 3 è più complessa della serie 2 e che la serie 4 ha una complessità algoritmica pari o superiore a quella della serie 3.

Come si può vedere dai risultati riportati nella tabella 20.1, per la serie 1 i valori dell'indice di corrispondenza normalizzato risultano essere simmetrici rispetto al lag 5. Esso assume inoltre i soli valori 0.2 e 0.6 che si ripetono con regolarità, ora con segno positivo, ora con segno negativo evidenziando, senza dubbio, una qualche forma di regolarità nel disegno della serie. L'indice di corrispondenza assume inoltre valore 1 (perfetta correlazione) in corrispondena del lag 10 ad indicare che la serie è perfettamente periodica di periodo 10.

Nella serie 2 l'indice di corrispondenza assume solo 3 valori distinti, 0.4, 0.2, con segni ora positivi ora negativi, e 0. In questo caso la regolarità sottostante risalta con minore evidenza, pur tuttavia, un andamento di questo tipo potrebbe suggerire ulteriori indagini ed il considerare un numero di lag più ampio. Così facendo si potrebbe senz'altro appurare che i valori dell'indice di corrispondenza sono simmetrici rispetto al lag 10 e che, in corrispondenza del lag 20, l'indice assume valore 1 ad indicare perfetta periodicità di periodo 20.

Tabella 20.1: Serie deterministiche complesse periodiche e non periodiche: valori dell'indice di corrispondenza normalizzato in forma assoluta.

Lunghezza della serie = 1000

	Pattern della serie			
Lag	1100111000	11001110000001110011	100111000011111....	101100111000....
0	1	1	1	1
1	0,2	0,4	0,912	0,876
2	-0,6	-0,2	0,826	0,758
3	-0,6	-0,4	0,744	0,646
4	0,2	-0,2	0,664	0,538
5	0,6	0,2	0,588	0,436
6	0,2	0,2	0,512	0,342
7	-0,6	0	0,442	0,250
8	-0,6	-0,2	0,372	0,166
9	0,2	-0,2	0,306	0,086
10	1	-0,2	0,244	0,012

Le serie 3 e 4 mostrano invece valori dell'indice di corrispondenza che, prossimi a 0.9 in corrispondenza del lag 1, vanno via via diminuendo all'aumentare del valore del lag, come ad indicare una mancanza di regolarità o una regolarità, non evidenziata, di periodo più lungo.

I valori dell'indice di corrispondenza calcolati sulle serie 1 e 2, come già quelli della S_ρ e della mutua informazione si mantengono costanti anche nel caso in cui si aumenti la lunghezza delle serie. Infatti, purchè tale lunghezza sia un multiplo del periodo (ma per serie sufficientemente lunghe questa

condizione non è necessaria), la struttura della serie non cambia, così come non cambia la distribuzione congiunta e, di conseguenza, il valore dell'indice.

Nel caso delle serie 3 e 4, invece, i valori dell'indice di corrispondenza tendono ad aumentare perchè, per come sono strutturati gli algoritmi generatori, al crescere della lunghezza delle serie aumenta anche la dimensione dei gruppi di simboli consecutivi uguali. Di conseguenza, al variare del lag, se per serie sufficientemente lunghe le distribuzioni marginali rimangono sostanzialmente costanti, nella distribuzione congiunta aumenta invece la percentuale di corrispondenze esatte (coppie "00" e "11") rispetto a quelle non esatte (coppie "10" e "01"), il cui numero diventa via via sempre più trascurabile rispetto al totale. Per questo motivo, all'aumentare della lunghezza delle serie 3 e 4, l'indice di corrispondenza tenderà ad avvicinarsi sempre più al valore uno, a prescindere dal valore del lag, ad indicare una condizione sempre più prossima a quella di perfetta correlazione.

Anche in questo caso l'indice di corrispondenza presenta un andamento molto simile a quello della S_ρ e della mutua informazione. Si osserva però una differenza sostanziale: mentre la S_ρ e la mutua informazione presentavano solo valori positivi non distinguendo fra correlazione (o autocorrelazione) diretta o inversa, l'indice di corrispondenza riesce a discriminare queste due situazioni assumendo appunto, in corrispondenza di una situazione di correlazione inversa, valori negativi.

Infine, anche l'indice di corrispondenza, come già gli altri indicatori, non pare in grado di fornire indicazioni in merito al grado di complessità algoritmica delle serie studiate.

CAPITOLO 21

Analisi di serie generate da un processo stocastico

Si applicherà adesso l'indice di corrispondenza ad alcuni esempi di serie generate dal semplice processo stocastico di figura 21.1, il quale è stato dettagliatamente studiato nel precedente capitolo 8.

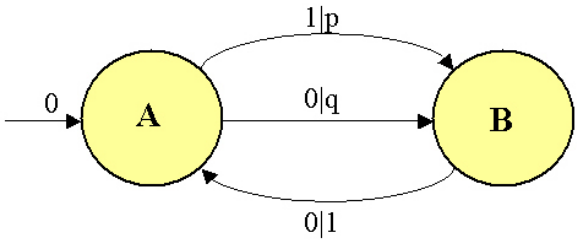

Figura 21.1: Esempio di processo generatore per una serie stocastiche

In particolare, al tempo t_0 ci si trova in una modalità di partenza A e in uno stato $x_{t_0} = 0$. Al successivo istante t_1 il sistema passa nella modalità B mantenendo lo stato 0, $x_{t_1} = 0$, con probabilità q, oppure assumendo lo stato 1, $x_{t_1} = 1$, con probabilità p. Una volta in B il sistema è obbligato, all'istante successivo, a ritornare nella modalità A, caratterizzata dallo stato $x_t = 0$. All'istante ancora successivo, il sistema ritorna necessariamente alla modalità B mantenendo lo stato 0 con probabilità q oppure assumendo

lo stato 1 con probabilità p. In sostanza, il sistema passa, ad ogni istante successivo, dalla modalità A alla B e viceversa (modalità ricorrenti) potendo assumere:

- nella transizione da A a B, alternativamente lo stato $x_t = 0$ con probabilità q oppure lo stato $x_t = 1$ con probabilità pari a p;
- nella transizione da B a A, unicamente lo stato $x_t = 0$.

Per mezzo di tale processo stocastico sono stati generati tre insiemi di 1000 simulazioni ciascuno. Ogni insieme di simulazioni era caratterizzato da una diversa coppia di probabilità di trasizione, ovvero:

- *insieme 1*: $p = 0.25$ e $q = 0.75$;
- *insieme 2*: $p = 0.5$ e $q = 0.5$;
- *insieme 3*: $p = 0.75$ e $q = 0.25$;

Ciascuna simulazione consisteva, data una determinata coppia di probabilità di transizione, nella generazione di una serie di 5000 elementi. Si è scelto di considerare serie di 5000 elementi, a differenza di quanto realizzato con gli altri indicatori, in quanto si è osservato che, con serie di soli 1000 elementi, gli intervalli di confidenza risultavano eccessivamente ampi e, quindi, di scarso significato. Gli intervalli di confidenza si riducevano notevolmente anche portando i casi a 5000 e considerando serie di lunghezza standard (1000 elementi), ma ciò avveniva in misura inferiore rispetto all'aumento della lunghezza della serie. Il valore medio dell'indice di corrispondenza osservato, invece, non differiva sostanzialmente aumentando la lunghezza della serie o il numero dei casi.

Nelle tabelle 21.1 e 21.2 sono riportati, rispettivamente, i valori dell'indice di corrispondenza e dell'indice di corrispondenza normalizzato in forma assoluta relativi ai tre diversi valori della probabilità di transizione p.

Come si può vedere dai dati di entrambe le tabelle, per ciascuna serie l'indice di corrispondenza assume sempre e solo due valori i quali risultano essere significativamente diversi tra loro. Il valore più alto si presenta inoltre, conformemente a quanto ci si poteva aspettare, in corrispondenza dei lag pari, diversamente da quanto rilevato per la S_ρ e la mutua informazione. Infatti, per costruzione, le serie sono caratterizzate dall'avere uno zero come primo simbolo, zero che si ripresenta poi sempre, in maniera costante, ogni due simboli. Ogni serie avrà quindi una struttura del tipo: "0x0x0x0x0x0x0x0x0x0x.....", dove il simbolo "x" può assumere, indifferentemente, il valore 0 oppure 1, con una probabilità che dipende solo dal parametro p. In presenza di lag pari, quindi, vi è sicura corrispondenza fra almeno la metà degli elementi della serie, in quanto gli zeri fissi si incontrano sempre fra di loro. Ciò significa che la probabilità di trovare elementi corrispondenti uguali è più elevata nei lag pari che nei lag dispari, motivo per

Tabella 21.1: Valori dell'indice di corrispondenza non normalizzato e relativi intervalli di confidenza al 95% in funzione della probabilità di transizione p.

Indice di corrispondenza non normalizzato
Lunghezza della serie = 5000, N. di casi = 1000

	Valori di p					
Lag	0.25		0.50		0.75	
0	1	1 1	1	1 1	1	1 1
1	0,749761	0,732400 0,766400	0,499842	0,480400 0,518000	0,249953	0,232790 0,267610
2	0,812400	0,801395 0,824405	0,750088	0,740200 0,759400	0,812543	0,801600 0,824000
3	0,749760	0,732400 0,766600	0,499838	0,480395 0,518000	0,249963	0,232595 0,267805
4	0,812330	0,801400 0,823405	0,749986	0,740395 0,759800	0,812344	0,800800 0,823810
5	0,749764	0,732400 0,766600	0,499846	0,480395 0,517805	0,249961	0,232790 0,267805
6	0,812310	0,801000 0,823605	0,749749	0,739595 0,759605	0,812423	0,800400 0,824205
7	0,749759	0,732400 0,766605	0,499847	0,480400 0,517810	0,249966	0,232790 0,267805
8	0,812111	0,800800 0,823400	0,749841	0,740000 0,759805	0,812576	0,800800 0,823600
9	0,749757	0,732595 0,766605	0,499840	0,480400 0,518000	0,249958	0,232595 0,267800
10	0,812524	0,801800 0,823200	0,750094	0,740195 0,759605	0,812630	0,801195 0,823805

Tabella 21.2: Valori dell'indice di corrispondenza normalizzato in forma assoluta e relativi intervalli di confidenza al 95% in funzione della probabilità di transizione p.

Indice di corrispondenza normalizzato in forma assoluta
Lunghezza della serie = 5000, N. di casi = 1000

	Valori di p					
Lag	0.25		0.50		0.75	
0	1	1 1	1	1 1	1	1 1
1	-0,040306	-0,062528 -0,019008	-0,200253	-0,231360 -0,171200	-0,529500	-0,561807 -0,496264
2	0,142400	0,092091 0,197280	0,333568	0,307200 0,358400	0,600092	0,576747 0,624533
3	-0,040307	-0,062528 -0,018752	-0,200259	-0,231368 -0,171200	-0,529481	-0,562174 -0,495896
4	0,142080	0,092114 0,192709	0,333296	0,307720 0,359467	0,599667	0,575040 0,624128
5	-0,040302	-0,062528 -0,018752	-0,200246	-0,231368 -0,171512	-0,529485	-0,561807 -0,495896
6	0,141989	0,090286 0,193623	0,332664	0,305587 0,358947	0,599836	0,574187 0,624971
7	-0,040308	-0,062528 -0,018746	-0,200245	-0,231360 -0,171504	-0,529476	-0,561807 -0,495896
8	0,141079	0,089371 0,192686	0,332909	0,306667 0,359480	0,600162	0,575040 0,623680
9	-0,040311	-0,062278 -0,018746	-0,200256	-0,231360 -0,171200	-0,529491	-0,562174 -0,495906
10	0,142967	0,093943 0,191771	0,333584	0,307187 0,358947	0,600277	0,575883 0,624117

cui, in corrispondenza dei lag pari, il valore dell'indice di corrispondenza presenta valori più elevati ad indicare un maggiore livello di autocorrelazione. In particolare, in corrispondenza dei lag pari, in media, ci si potrà aspettare una percentuale di $0,5 + (p^2 + q^2)/2$ elementi uguali, pari a 0.8125, 0.75 e 0,8125 relativi, rispettivamente, a $p = 0.25$, $p = 0.5$ e $p = 0.75$, che sono proprio, a meno delle inevitabili fluttuazioni casuali, i valori dell'indice non normalizzato riportati in tabella.

Dai valori dell'indice di corrispondenza normalizzato in forma assoluta (tabella 21.2), si può notare come i valori relativi ai lag dispari siano considerati come inversamente correlati. Il livello di correlazione diretta o inversa esplicitato dall'indice aumenta inoltre all'aumentare del valore della probabilità di transizione p per arrivare, in corrispondenza di $p = 1$ (serie del tipo "01010101..."), ai valori uno e meno uno indicanti, rispettivamente, perfetta correlazione diretta (lag pari) e perfetta correlazione inversa (lag dispari).

Dai risultati tabulati si può infine rilevare come l'indice di corrispondenza, normalizzato o meno che sia, riesca chiaramente a discriminare le 3 diverse condizioni di partenza ($p = 0.25, p = 0.50$ oppure $p = 0.75$) e come abbia un comportamento, in linea generale, non dissimile da quello della S_ρ e della mutua informazione.

CAPITOLO 22

Analisi di serie generate da un processo caotico

In questo capitolo verrà studiato l'indice di corrispondenza mediante la sua applicazione alla mappa logistica analizzata nel precedente capitolo 9.

Nello specifico, la nostra attenzione di concentrerà sull'isola di stabilità che inizia in prossimità di un valore del parametro r pari a 3.82847, $\simeq 1+\sqrt{8}$ (vedi figura 22.1) ed in particolare:

- in corrispondenza del valore r = 3.828, che si situa subito prima dell'inizio dell'isola di stabilità considerata;
- in corrispondenza del valore r = 3.82847, che si situa subito dopo l'inizio dell'isola di stabilità considerata;
- in corrispondenza dei valori 3.844568 e 3.844569, che identificavano una zona di "salto" sia per la S_ρ che per la mutua informazione e la $ApEn$, ovvero punti nei quali gli indicatori menzionati modificavano in maniera sostanziale il loro comportamento;
- in corrispondenza del valore r = 3.85, che si situa in una zona centrale dove le biforcazioni tendono a farsi più numerose e l'andamento più confuso;
- in corrispondenza del valore r = 3.8565, che si situa poco prima della fine dell'isola di stabilità;

- in corrispondenza del valore r = 3.8575, che si situa subito al di la dell'isola di stabilità considerata.

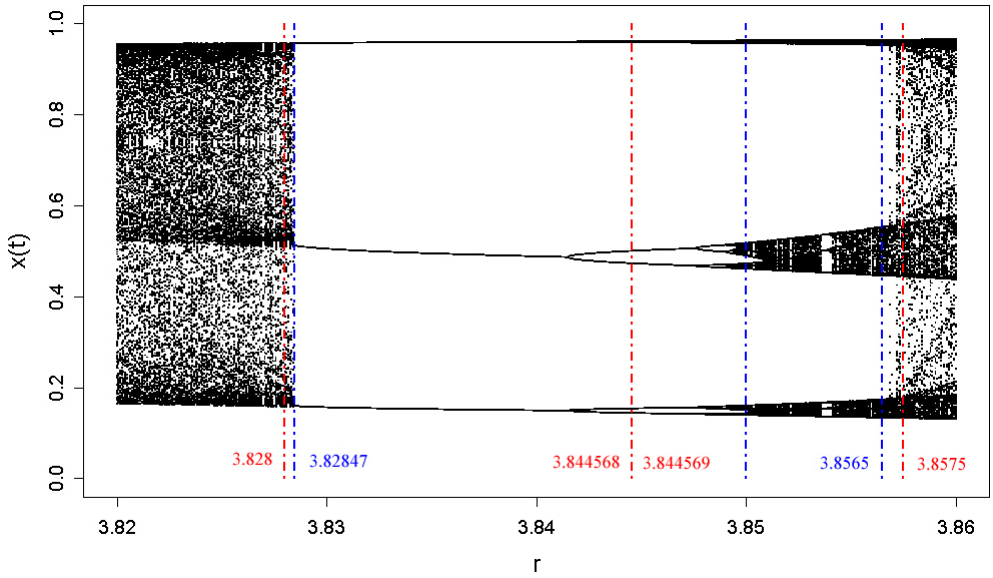

Figura 22.1: Isola di stabilità $r = 3.82847$.

Al fine dell'esperimento, per ogni valore del parametro r considerato sono state generate 1000 serie (casi) composte da 5000 elementi. Ciascun elemento, indicato con Y_t, è stato ottenuto dicotomizzando come segue i risultati della mappa logistica:

$$Y_t = \begin{cases} 0, & 0 \leq X_t \leq 0.5 \\ 1, & 0.5 < X_t \leq 1 \end{cases}$$

$$\text{con } X_{t+1} = rX_t(1 - X_t), \qquad 0 < X_0 < 1$$

Come si può osservare dai dati riportati nelle tabella 22.1, l'indice di corrispondenza sembra riuscire a caratterizzare in maniera abbastanza soddisfacente l'andamento della mappa logistica in presenza di un'isola di stabilità. Confrontando i valori della tabella 22.1 (indice di corrispondenza non normalizzato) con quelli della tabella 9.5 (S_ρ, mutua informazione e $ApEn$) di cui al precedente capitolo 9.3, risulta evidente l'analogia di comportamento dell'indice non normalizzaro con la S_ρ e la mutua informazione, mentre l'indice normalizzato in forma assoluta (tabella 22.2), pur mantenendo lo stesso andamento tendenziale, è in grado di evidenziare anche il segno della correlazione.

Infatti, per $r = 3.828$ ed $r = 3.8575$, valori che si situano subito prima l'inizio e subito dopo la fine dell'isola di stabilità, l'indice di corrispondenza

Tabella 22.1: Mappa logistica dicotomizzata: valori dell'indice di corrispondenza non normalizzato relativi all'isola di stabilità $r = 3.82847$.

Indice di corrispondenza non normalizzato
Lunghezza della serie = 5000, numero di casi = 1000

	Valori di r						
Lag	3.828	3.82847	3.844568	3.844569	3.85	3.8565	3.8575
0	1	1	1	1	1	1	1
1	0,368488	0,333542	0,333332	0,333328	0,333335	0,333334	0,352929
2	0,477240	0,334067	0,333339	0,333339	0,333333	0,333328	0,399572
3	0,796284	0,998985	0,999991	0,666725	0,775187	0,817388	0,782468
4	0,446309	0,333905	0,333335	0,333330	0,333335	0,333342	0,370725
5	0,478382	0,334102	0,333339	0,333335	0,333329	0,333318	0,415608
6	0,762561	0,998811	0,999990	0,999993	0,891481	0,826241	0,753933
7	0,466316	0,334004	0,333335	0,333321	0,333340	0,333342	0,377458
8	0,489367	0,334117	0,333338	0,333346	0,333326	0,333323	0,430120
9	0,735050	0,998734	0,999990	0,666719	0,775192	0,824163	0,725096
10	0,481817	0,334048	0,333334	0,333324	0,333341	0,333339	0,383643

Tabella 22.2: Mappa logistica dicotomizzata: valori dell'indice di corrispondenza normalizzato in forma assoluta relativi all'isola di stabilità $r = 3.82847$.

Indice di corrispondenza normalizzato
Lunghezza della serie = 5000, numero di casi = 1000

	Valori di r						
Lag	3.828	3.82847	3.844568	3.844569	3.85	3.8565	3.8575
0	1	1	1	1	1	1	1
1	-0,351093	-0,399720	-0,399998	-0,333344	-0,341101	-0,334010	-0,302310
2	-0,159581	-0,398776	-0,399985	-0,333322	-0,341105	-0,334022	-0,210104
3	0,528589	0,997716	0,999980	0,333450	0,545008	0,634404	0,559782
4	-0,214050	-0,399067	-0,399992	-0,333340	-0,341101	-0,333994	-0,267130
5	-0,157569	-0,398713	-0,399985	-0,333330	-0,341113	-0,334042	-0,178403
6	0,450552	0,997324	0,999978	0,999986	0,780372	0,652128	0,502036
7	-0,178818	-0,398889	-0,399992	-0,333358	-0,341091	-0,333994	-0,253820
8	-0,138225	-0,398686	-0,399987	-0,333308	-0,341119	-0,334032	-0,149714
9	0,386890	0,997151	0,999978	0,333438	0,545018	0,647968	0,443679
10	-0,151520	-0,398810	-0,399994	-0,333352	-0,341089	-0,334000	-0,241593

lascia forse intravvedere solo qualche debole accenno di struttura, che si concretizza principalmente nella costanza dell'alternanza dei segni. D'altronde, osservando la figura 22.1, si può osservare come, in corrispondenza di tali valori di r, i punti sul grafico non si distribuiscano in maniera uniforme, ma si rilevino delle zone più dense (le zone più scure), dove i punti si addensano con maggiore frequenza e, quindi, con maggiore probabilità, caratterizzando così in qualche modo il comportamento della mappa logistica dicotomizzata anche al di fuori dell'isola di stabilità. I livelli di autocorrelazione osservati risultano comunque sempre decisamente più bassi rispetto a quelli registrati in aree dove la mappa logistica è caratterizzata da un comportamento decisamente più regolare ($r = 3.82847$, $r = 3.844568$, $r = 3.844569$ e $r = 3.85$).

Per $r = 3.82847$ e $r = 3.844568$ l'indice di corrispondenza identifica, per entrambi i livelli di r, una struttura perfettamente periodica di periodo 3. Infatti, in corrispondenza dei lag 3, 6 e 9 si possono osservare valori dell'indice sostanzialmente pari ad 1, il che significa che vi è, in pratica, perfetta coincidenza fra tutti gli elementi della serie originaria e della serie scalata verso sinistra di 3, 6 o 9 elementi, mentre per tutti gli altri lag si osserva sempre uno stesso valore, che per l'indice normalizzato è pari a circa -0.399, ad indicare un certo livello, costante, di correlazione inversa.

Nel passaggio da $r = 3.844568$ a $r = 3.844569$, come già la S_ρ e la mutua informazione, anche l'indice di corrispondenza mette in risalto un sostanziale mutamento nel comportamento della mappa logistica dicotomizzata. Si passa infatti da una perfetta regolarità di periodo 3 ad un'altrettanto perfetta regolarità di periodo 6, ($r = 3.844569$, lag 6, $I_{Acrs} = 0.99998$), mentre il valore relativo ai lag 3 e 9 si riduce ad un terzo ($I_{Acrs} = 0.3334$) e quello dei lag rimanenti si attesta su di un valore costante e pari a -0.3333, non molto dissimile da quelli osservati nel caso di $r = 3.82847$ e $r = 3.844568$.

Con $r = 3.82847, 3.844568$ e 3.8844569 abbiamo esplorato la parte iniziale e più semplice dell'isola di periodicità caratterizzata, come si può ben vedere in figura 22.1, prima da 3 e poi da 6 biforcazioni. Con $r = 3.85$ ed $r = 3.8565$, che si situano, rispettivamente, in una zona dove le biforcazioni iniziano ad aumentare sensibilmente ed in prossimità della fine dell'isola di stabilità, si analizzerà invece la seconda e più complessa parte dell'isola stessa.

Per $r = 3.85$ e limitatamente ai lag considerati, si osserva ancora chiaramente una forma di struttura ben definita, con un valore massimo dell'indice di corrispondenza normalizzato molto elevato (0.78) in corrispondenza del lag 6, due valori ancora relativamente elevati (0.545), uguali e simmetrici rispetto al valore massimo (lag 9 e 3) e valori negativi e costanti per i rimanenti lag (-0.3411). Un comportamento analogo si può osservare per $r = 3.8565$, dove l'indice di corrispondenza normalizzaro assume un valore praticamente costante e pari a circa 0.64 in corrispondenza dei lag 3,6 e 9 ed un valore negativo e costante, pari a -0.334, in corrispondenza di tutti gli altri.

Si è poi realizzata una serie di ulteriori prove in corrispondenza di $r = 4$,

valore che è associato ad un comportamento della mappa logistica completamente caotico. In tale situazione ci si aspetterebbe un valore dell'indice di corrispondenza che attesti la sostanziale incorrelazione delle serie generate. In effetti, come si può osservare nella tabella 22.3, dove sono riportati i valori dell'indice di corrispondenza normalizzato e non normalizzato, quest'ultimo assume valore 0.5 in corrispondenza di tutti i lag considerati. Tale valore, per una serie caratterizzata da una distribuzione marginale di indifferenza $p = q = 0.5$ quale quella che dovrebbe identificare una serie completamente caotica, è proprio il valore corrispondente alla condizione di indipendenza e, quindi, di incorrelazione. In ragione di ciò l'indice standardizzato assume correttamente il valore 0.

Tabella 22.3: Mappa logistica dicotomizzata: valori dell'indice di corrispondenza normalizzato in forma assoluta relativi all'isola di stabilità $r = 3.82847$.

Indice di corrispondenza calcolato in corrispondenza di r = 4
Lunghezza della serie = 5000, N. di casi = 1000

	r = 4, Indice di corrispondenza			
Lag	non normalizzato		normalizzato	
0	1	1 1	1	1 1
1	0,500243	0,487590 0,514800	0,000486	-0,024820 0,029600
2	0,500514	0,486595 0,514600	0,001028	-0,026810 0,029200
3	0,499925	0,486000 0,513605	-0,000150	-0,028000 0,027210
4	0,499991	0,486395 0,513410	-0,000018	-0,027210 0,026820
5	0,500134	0,486400 0,513800	0,000268	-0,027200 0,027600
6	0,500056	0,486195 0,513605	0,000112	-0,027610 0,027210
7	0,500613	0,487795 0,514400	0,001226	-0,024410 0,028800
8	0,500200	0,486595 0,513405	0,000400	-0,026810 0,026810
9	0,499586	0,485385 0,513005	-0,000828	-0,029230 0,026010
10	0,499981	0,485800 0,514600	-0,000038	-0,028400 0,029200

22.1 Sensibilità ai valori iniziali

A conclusione degli esperimenti condotti con la mappa logistica, si è realizzata un'ultima serie di simulazioni volta a verificare la sensibilità alle condizioni iniziali degli indicatori considerati. A tal fine si sono applicate la S_ρ, la mutua informazione, la *cross Sample-ApEn* e l'indice di corrispondenza a due serie generate nella zona caotica della mappa logistica ($r = 4$) e caratterizzate da un valore iniziale x_0 uguale fino alla quinta cifra decimale ($x_{u0} = 0.287564$ e $x_{v0} = 0.287565$), in modo da stabilire se tali indicatori fossero in grado di rilevare e quantificare il processo di divergenza fra le due serie.

Nella tabella 22.4 sono riportati i valori degli indicatori non normalizzati applicati alle due serie di cui sopra in funzione del numero di elementi delle serie stesse. Come si può vedere, l'indice di corrispondenza evidenza chiaramente il momento in cui le due serie cominciano a divergere passando dal valore uno, indicante perfetta coincidenza fra tutti gli elementi delle due serie considerate, ad un valore inferiore ad uno. Anche la S_ρ e la mutua informazione riescono ad evidenziare tale situazione, anche se in maniera meno evidente ed intuitiva. Infatti, fintanto che le due serie di mantengono uguali, sia la S_ρ che la mutua informazione si mantengono pressoche costanti per poi calare in maniera più o meno pronunciata in corrispondenza dell'inizio del processo di divergenza.

Sia l'indice di corrispondenza, che la S_ρ e la mutua informazione, con la diminuzione dei rispettivi valori evidenziano il progredire del processo di divergenza fino ad attestare una sostanziale indipendenza fra le due serie già a partire da 500 elementi.

La *cross Sample-ApEn* appare invece molto meno informativa. Essa non appare in grado di evidenziare l'inizio del processo di divergenza né, tantomeno, la sua evoluzione, mantenendosi sempre molto prossima ad uno.

Per valutare ulteriormente la sensibilità alle condizioni iniziali, si è ripetuto l'esperimento precedente considerando due valori iniziali ancora più vicini fra loro (uguali fino alla dodicesima cifra decimale), ovvero $x_{u0} = 0.28756371$ 64524 e $x_{v0} = 0.2875637164525$. I risultati sono riportati nella tabella 22.5. Come si può osservare, il comportamento di fondo degli indicatori non cambia. Essi inoltre, com'era logico aspettarsi, evidenziano un maggior ritardo nell'inizio del processo di divergenza proprio in ragione della maggiore vicinanza dei valori iniziali considerati.

Analogamente a quanto già osservato per la S_ρ e la mutua informazione, anche l'indice di corrispondenza appare quindi in grado di descrivere il comportamento della mappa logistica dicotomizzata in relazione al livello di dipendenza e di regolarità che essa esprime. Esso riesce a caratterizzare adeguatamente le isole di stabilità mettendo in luce una qualche regolarità di struttura anche nei tratti in cui le biforcazioni sono più numerose e il sistema, di conseguenza, più confuso. E' inoltre in grado di evidenziare con estrema chiarezza il processo di divergenza fra serie caotiche caratterizzate

Tabella 22.4: Mappa logistica dicotomizzata: sensibilità ai valori iniziali (lag 0).

Serie caotiche, sensibilità ai valori iniziali in funzione del numero di elementi delle serie (x_{u0} = 0.287564 e x_{v0} = 0.287565)

	Numero di elementi delle serie						
	10	25	50	100	250	500	1000
Sp	0,282259	0,072988	0,025656	0,006168	0,001800	0,000006	0,000019
Mutua Informazione	0,970950	0,378090	0,143543	0,035333	0,010367	0,000036	0,000109
Cross Sample-ApEn	0,965235	0,990219	0,998801	1,006495	1,005451	1,002805	1,002001
I_{crs}	1,000000	0,840000	0,720000	0,610000	0,560000	0,504000	0,506000

	Numero di elementi delle serie				
	15	16	18	19	20
Sp	0,282259	0,288746	0,279758	0,182531	0,189454
Mutua Informazione	0,970950	0,988699	0,964079	0,720528	0,744484
Cross Sample-ApEn	0,973033	0,987722	0,952382	0,966447	0,984733
I_{crs}	1,000000	1,000000	1,000000	0,947368	0,950000

Tabella 22.5: Mappa logistica dicotomizzata: sensibilità ai valori iniziali (lag 0).

Serie caotiche, sensibilità ai valori iniziali in funzione del numero di elementi delle serie
(x_{u0} = 0.2875637164524 e x_{v0} = 0.2875637164525)

	Numero di elementi delle serie						
	10	25	37	42	43	45	50
Sp	0,282260	0,282260	0,291149	0,292292	0,222835	0,136056	0,105821
Mutua Informazione	0,970950	0,970950	0,995253	0,998364	0,861600	0,650050	0,531021
Cross Sample-ApEn	0,965235	0,961526	0,977973	0,991489	0,993487	0,994051	1,002391
I_{crs}	1,000000	1,000000	1,000000	1,000000	0,976744	0,933333	0,900000

da condizioni iniziali molto vicine. Infine, esso può fornisce informazioni oltre che sull'intensità, anche sul segno della correlazione, cosa che invece né la S_ρ, né la mutua informazione né, tanto meno, la *ApEn* o la *cross-ApEn* sono in grado di fare.

CAPITOLO 23

Considerazioni finali

Come si è avuto modo di appurare nella Parte III della tesi, pur fornendo una certa misura del grado di "vicinanza" o di "sincronismo" fra due serie, la *cross-ApEn* e la *cross Sample-ApEn* non sono in grado di fornire informazioni univoche sull'eventuale indipendenza fra le serie stesse né, tantomeno, sul loro livello di correlazione. Per ovviare a tali lacune si è modificato l'algoritmo di calcolo originario in modo da associarvi un ulteriore valore numerico che abbiamo chiamato *indice di corrispondenza* o I_{crs}.

A fronte di un incremento del tempo di elaborazione sostanzialmente trascurabile, l'indice di corrispondenza, che calcola la frequenza di elementi corrispondenti uguali fra le due serie, è in grado di dirci se le serie considerate possano essere ritenute uguali fra loro a meno di un margine di tolleranza r, cosa che la *cross-ApEn* non è in grado di fare fornendoci, nel contempo, anche un valore numerico proporzionale all'intensità del livello di correlazione eventualmente esistente fra di esse. In particolare, l'indice di corrispondenza, che in questa forma per essere calcolato non necessità della conoscenza di alcuna distribuzione, varia fra zero, che nel caso di serie binarie corrisponde alla condizione di perfetta correlazione inversa, ed uno, che attesta la perfetta coincidenza fra tutti gli elementi delle serie considerate.

Un importante passo avanti nell'interpretazione dei risultati forniti può essere ottenuto normalizzando l'indice di corrispondenza. Per fare ciò è comunque necessario ricavare il suo valore caratteristico p_{ind}, che rappresenta il valore che l'indice assume in corrispondenza della condizione di indipendenza e dipende dalla forma delle distribuzioni marginali delle serie considerate. In particolare, se $\{p_u, q_u\}$ rappresenta la distribuzione marginale della serie

di confronto u e $\{p_v,\, q_v\}$ quella della serie obiettivo v, $p_{ind} = p_u p_v + q_u q_v$. Per calcolare tale valore non è comunque strettamente necessario disporre delle distribuzioni marginali o di una loro stima, in quanto la distribuzione empirica dell'indice data la condizione di indipendenza, così come una stima di p_{ind} possono essere ottenute tramite simulazione per mezzo dei metodi Monte Carlo. Noto che sia il valore di p_{ind} è quindi possibile calcolare un indice normalizzato in forma assoluta, ovvero rispetto al suo minimo ed al suo massimo assoluti, zero ed uno, ed alla condizione di indipendenza, variabile da serie a serie a seconda delle distribuzioni marginali. In particolare si avrà:

$$\begin{cases} \text{se } I_{crs} - p_{ind} > 0, \text{ allora } I_{Acrs} = \frac{I_{crs} - p_{ind}}{1 - p_{ind}} \\ \text{se } I_{crs} - p_{ind} = 0, \text{ allora } I_{Acrs} = 0 \\ \text{se } I_{crs} - p_{ind} < 0, \text{ allora } I_{Acrs} = \frac{I_{crs} - p_{ind}}{p_{ind}} \end{cases}$$

In questa forma, l'indice di corrispondenza, I_{Acrs}, varierà nell'intervallo $[1, -1]$ assumendo i valori:

- 1, nel caso di perfetta correlazione positiva;
- 0, nel caso di indipendenza;
- -1, nel caso di perfetta correlazione inversa

Si è inoltre dimostrato che (vedi capitolo 16.2), sotto la condizione di indipendenza, l'indice di corrispondenza (non normalizzato) si distribuisce, se il numero n degli elementi delle serie è sufficientemente elevato, come una normale di media $\mu = p_{ind}$ e deviazione standard $\delta = \sqrt{p_{ind}(1 - p_{ind})/n}$, ovvero:

$$I_{crs} \sim \mathcal{N}\left[p_{ind}, \sqrt{p_{ind}(1 - p_{ind})/n}\right].$$

L'indice così costruito è corretto, in quanto somma di due frequenze (la frequenza delle coppie di elementi corrispondenti uguali $(0, 0)$ e $(1, 1)$) ed indipendente dal tipo di codifica adottata. Ovviamente, nel caso di serie non binarie, il valore zero (meno uno nel caso dell'indice normalizzato) non corrisponderà più alla condizione di perfetta correlazione inversa, ma a quella particolare condizione dove un qualsiasi elemento della serie di confronto non è mai associato a se stesso nella serie obiettivo. L'indice di corrispondenza, al pari della *ApEn*, è poi insensibile a disturbi di intensità inferiore al valore del livello di soglia r e risente in maniera minimale di eventuali (pochi) valori anomali (outliers) in quanto non è la loro intensità, bensì il loro numero che si riflette sul valore finale dell'indicatore. Esso può essere inoltre utilizzato sia nel caso di relazioni lineari che non lineari e, al pari della *ApEn* da cui prende spunto, sia con dati di tipo binario, che di tipo discreto o continuo.

Applicando l'indice di corrispondenza agli esperimenti già effettuati con la S_ρ, la mutua informazione, le distanze di Kullback Leibler e la *ApEn* o

la *cross-ApEn* si è potuto osservare, in generale, un comportamento tendenzialmente molto simile a quello della S_ρ e della mutua informazione, con l'eccezione del caso delle serie semplici perfettamente periodiche (capitolo 19), dove l'indice di corrispondenza ha mostrato una maggiore fedeltà all'evidenza teorica e al modello studiato.

In definitiva, l'indice di corrispondenza presenta alcune interessanti proprietà:

- per il suo calcolo non è necessaria la stima di alcuna distribuzione, che il valore caratteristico p_{ind}, così come la sua distribuzione sotto la condizione di indipendenza può essere ricavata mediante simulazione;
- la sua forma asintotica, data la condizione di indipendenza, è una funzione semplice e, note le distribuzioni marginali, perfettamente definita avendosi $I_{crs} \sim \mathcal{N}\left[p_{ind}, \sqrt{p_{ind}(1-p_{ind})/n}\right]$;
- risulta direttamente proporzionale all'intensità della correlazione o dell'autocorrelazione esistente fra le serie considerate;
- presenta valori mediamente più elevati e maggiormente spaziati rispetto sia alla S_ρ che alla mutua informazione;
- nella sua forma normalizzata evidenza e quantifica in maniera immediata la correlazione inversa per tramite dei valori negativi;
- è applicabile anche a serie caotiche;
- si integra perfettamente nell'algoritmo di calcolo della *ApEn* e della *cross-ApEn*;
- ha comunque un significato proprio ed indipendente da quello della *cross-ApEn* e, in quanto tale, può essere calcolato e valutato anche in maniera disgiunta da essa. In questo caso il suo algoritmo risulta estremamente semplice e performante.

Esso presenta ovviamente anche delle limitazioni, la principale delle quali consiste nell'avere una varianza decisamente più elevata rispetto a quella della S_ρ e della muta informazione, ragion per cui, per avere stime affidabili, si dovrebbe poter disporre di serie più lunghe. In ogni caso 5000 elementi appaiono una lunghezza già più che sufficiente per ottenere risultati ampiamente accettabili nella maggior parte delle situazioni che si possono incontrare nella pratica.

Nella tabella 23.1 sono riportati i valori della deviazione standard media della S_ρ, della mutua informazione, dell'entropia congiunta, della *cross Sample-ApEn* e dell'indice di corrispondenza calcolati in funzione della lunghezza delle serie sotto la condizione di indipendenza, distribuzioni di probabilità marginali simmetriche ($p = q = 0.5$) ed insiemi di 1000 simulazioni

(casi). Come si può vedere, la deviazione standard della S_ρ risulta inferiore di oltre due ordini di grandezza rispetto a quella dell'indice di corrispondenza. Bisogna dire, però, che sia la S_ρ che la mutua informazione, specie in corrispondenza di livelli di relezione non particolarmente elevati, tendono ad assumere valori molto bassi e la deviazione standard, di conseguenza, non può che riflettere questa situazione.

Tabella 23.1: Deviazioni standard in funzione della lunghezza delle serie.

Indicatori non normalizzati: deviazione standard calcolata su 1000 casi

	Numero di elementi delle serie				
	1000	2500	5000	10000	15000
Sp	0,000185	0,000067	0,000033	0,000016	0,000014
Mutua Informazione	0,000930	0,000380	0,000220	0,000110	0,000069
Entropia Congiunta	0,001640	0,000700	0,000380	0,000169	0,000117
Cross Sample-ApEn	0,001400	0,000600	0,000280	0,000140	0,000091
I_{crs}	0,015600	0,010000	0,007500	0,005000	0,004000

In ogni caso, considerato anche l'esiguo sforzo necessario, è consigliabile calcolare l'indice di corrispondenza ogni qual volta si calcolino la *cross-ApEn* e la *cross Sample-ApEn*, sia per confermare eventuali tendenze messe in luce da questi indicatori, sia per meglio specificare il comportamento delle serie considerate, in particolar modo per quanto riguarda il livello di correlazione eventualmente esistente fra le stesse.

L'indice di corrispondenza non è comunque in grado di stabilire, al di la di una eventuale indicazione di direzione e di intensità, il tipo di relazione che intercorre fra due fenomeni né, tanto meno, al pari degli altri indicatori analizzati, di fornire informazioni riguardo alla complessità algoritmica delle serie studiate.

Conclusioni e prospettive future

Con questo lavoro si è cercato di:

- realizzare uno studio di tipo applicativo, comparato ed esaustivo, fra alcuni degli indicatori di indipendenza e di correlazione basati sul concetto di entropia noti in letteratura (una delle loro caratteristiche peculiari consiste nel poter essere applicati anche a relazioni non lineari);
- valutare il comportamento dell'*Approximate Entropy*, più comunemente nota come *ApEn*, un indicatore di irregolarità e di dipendenza poco utilizzato in statistica ma che ha visto una certa diffusione nel campo delle scienze mediche e biologiche, in modo da comprenderne appieno limiti e potenzialità;
- proporre un'integrazione all'algoritmo di calcolo della *cross-ApEn* (la versione della *ApEn* applicabile al confronto fra serie distinte) in modo da colmare alcune lacune dell'indicatore e renderlo più esaustivo e completo.

Si sono pertanto confrontati fra loro i seguenti indicatori:

- l'entropia di Granger, Maasoumi e Racine, più spesso indicata con S_ρ;
- la mutua informazione;
- l'entropia congiunta;
- la misura di Kullback Leibler o entropia relativa;
- la *Approximate Entropy*, più comunemente nota come *ApEn* e la *cross Approximate Entropy*, anche detta *cross-ApEn*;

applicandoli a diversi insiemi di serie binarie appositamente create allo scopo. Si è preferito ricorrere ad un alfabeto binario perchè, se da un lato ciò può limitare la possibilità di generalizzare tout court i risultati ottenuti, dall'altro permette di ridurre al minimo la complessità del sistema. In questo modo è possibile definire precise relazioni fra le variabili e creare serie rappresentative di una pluralità di situazioni, anche complesse, riuscendo nel contempo a mantenere sotto controllo ed a valutare correttamente i diversi fattori sperimentali considerati.

Dopo aver riassunto alcuni concetti basilari relativi all'entropia ed alla teoria dell'informazione ed aver riassunto lo stato dell'arte degli indicatori sopra menzionati (Parte I della tesi), si sono impiegati gli stessi allo scopo di cercare di individuare e descrivere la struttura interna di una serie, nonchè il suo grado di irregolarità o di dipendenza ed, eventualmente, per ordinare i diversi tipi di serie in base alla loro complessità algoritmica (Parte II della tesi). Si è successivamente cercato di evidenziare le relazioni (regolarità, correlazione ed indipendenza) eventualmente esistenti fra due serie distinte ma di forma nota a priori (Parte III della tesi) per poi proporre ed applicare ai casi precedentemente studiati (Parte IV della tesi) un ulteriore indicatore, da associare alla *cross-ApEn*, in modo da colmarne le lacune ed integrare le informazioni che essa è in grado di fornire. Questo ulteriore indicatore, definito nel corso dell'esposizione come *indice di corrispondenza* o I_{crs}, basandosi sul concetto di distanza esemplificato dal parametro r proprio della *ApEn*, si integra perfettamente nell'algoritmo di calcolo della *cross-ApEn*. Esso rappresenta sicuramente un valore aggiunto per la *cross-ApEn* e per la sua forma computazionalmente più stabile, la *cross Sample-ApEn* potendo fornire precise informazioni riguardo ad un'ipotetica condizione di indipendenza fra le serie, oltre che sul livello di correlazione, diretta o inversa, eventualmente esistente fra le stesse.

23.1 risultati ottenuti

Per mezzo degli esperimenti eseguiti si è potuto immediatamente verificare come la mutua informazione e l'entropia congiunta, nella loro forma normalizzata, corrispondano ad una stessa misura. Cosa che è stata poi dimostrata analiticamente nel capitolo 5.2.3. Tuttavia la mutua informazione presenta varianza minore e quindi, a parità di altre condizioni, è sicuramente da preferirsi.

Si è visto inoltre come la mutua informazione e la S_ρ manifestino un comportamento molto simile, differendo sostanzialmente solo per un fattore di scala. La mutua informazione tende però ad assumere valori più alti, in particolar modo in corrispondenza dei livelli di irregolarità più elevati, e tende a zero (condizione di indipendenza) in maniera più graduale rispetto alla S_ρ. Quest'ultima presenta però, in genere, livelli di varianza leggermente infe-

riori. Ai fini pratici e per lo studio di serie binarie, comunque, i due indicatori appaiono sostanzialmente equivalenti per cui, se del caso, il ricercatore potrà utilizzare quello a lui più congeniale.

La distanza di Kullback Leibler, che misura una sorta di "distanza" fra le distribuzioni marginali delle serie considerate, non si è rivelata invece di alcuna utilità nello studio dell'autocorrelazione e della struttura di una singola serie in quanto, avendo a che fare con distribuzioni marginali praticamente coincidenti, assumeva sempre un valore molto prossimo a zero. Da questo punto di vista risulta scarsamente informativa anche nello studio delle relazioni esistenti fra serie distinte in quanto, dipendendo unicamente dalla forma delle distribuzioni marginali, essa può fornire uno stesso valore anche a fronte di condizioni e di strutture delle serie le più diverse fra loro.

Per quanto riguarda la $ApEn$, si è notato che, in genere, i risultati osservati concordano con quelli ottenuti con la S_ρ e con la mutua informazione. Questi ultimi sono sicuramente più indicati quali indicatori di indipendenza e correlazione, ma possono fornire informazioni utili anche per quanto riguarda la presenza di un'eventuale forma di struttura. La $ApEn$, invece, rappresenta sostanzialmente un indicatore del livello di disordine e di irregolarità e misura l'autocorrelazione solo indirettamente, tramite il livello di regolarità presente nella serie e le modalità con cui un determinato pattern si ripete nella sequenza dei dati.

Tutti e tre gli indicatori sono in grado di identificare con precisione strutture periodiche o quasi periodiche purchè il numero di lag per la S_ρ e la mutua informazione o la dimensione del parametro m per la $ApEn$ siano sufficientemente elevati. Un'eventuale periodicità viene infatti rilevata solo in corrispondenza di valori di lag o di m multipli del periodo della serie. Nel caso di lag multipli del periodo, infatti, non si fa altro che confrontare la serie con una sua copia speculare e, di conseguenza, la S_ρ e la mutua informazione assumono il loro valore massimo che, nota la distribuzione marginale della serie è calcolabile e, pertanto, riconoscibile. Per la $ApEn$, un'eventuale periodicità può essere rilevata anche in prossimità di un valore di m pari al periodo della serie ed è caratterizzata dalla persistenza a zero dell'indicatore per tutti gli m successivi. Tuttavia, il numero di elementi della serie necessari per ottenere stime attendibili cresce in maniera esponenziale all'aumentare di m per cui nella pratica, con serie di qualche migliaio di elementi, la letteratura sconsiglia di utilizzare valori di m superiori a 2 o a 3. Ciò non di meno, nello studio della struttura interna di una singola serie, può essere a volte utile considerare valori di m maggiori. Questo perchè un'eventuale regolarità di comportamento della $ApEn$ al variare di m, come la persistenza di un medesimo valore o di blocchi di valori uguali, potrebbe fornirci una qualche indicazione riguardo all'eventuale struttura della serie, ancorchè non perfettamente periodica o di periodo più elevato.

Si è visto, poi, come gli indicatori in questione, e soprattutto la S_ρ e la mutua informazione, siano particolarmente efficaci nell'analisi della serie

logistica dicotomizzata (capitolo 9). Non appare invece particolarmente soddisfacente la caratterizzazione di serie periodiche in funzione del grado di rumore indotto (capitolo 5).

Tutti gli indicatori considerati non sono infine risultati assolutamente adatti a fornire indicazioni riguardo alla complessità delle serie studiate. Valori diversi della S_ρ, della mutua informazione o della *ApEn* possono riferirsi, infatti, a serie caratterizzate dalla medesima complessità algoritmica e viceversa.

Per quanto riguarda l'analisi ed il confronto di serie diverse (Parte III della tesi), si è potuto verificare come sia la S_ρ che la mutua informazione si prestino, in qualche modo, anche allo studio del livello di correlazione. Esse, inoltre, come già appurato nel caso dell'analisi di una singola serie, presentano un comportamento ed una dinamica molto simile riuscendo ad individuare efficacemente una eventuale condizione di indipendenza, in corrispondenza della quale entrambi gli indicatori assumono valore zero. Esse sono anche in grado di rilevare un'eventuale condizione di perfetta o massima correlazione, corrispondente al massimo raggiungibile dagli indicatori date le particolari distribuzioni marginali, condizione che diventerebbe comunque evidente solo nel caso di normalizzazione degli indicatori rispetto a detto massimo. La S_ρ e la mutua informazione, comunque, non sono in grado di distinguere fra correlazione diretta e correlazione inversa, assumendo in entrambe le situazioni il medesimo valore.

Il discorso si fa più complesso nel caso della *ApEn* o, meglio, della *cross-ApEn*, che altro non è che la *ApEn* applicata a serie distinte. Intanto, se la *ApEn* applicata ad una singola serie è sempre definita (vedi capitolo 12.1.4), la *cross-ApEn*, invece non lo è. E' infatti possibile, e in molti casi assai probabile, che un determinato pattern x_i di lunghezza $m+1$ appartenente alla serie di confronto non abbia riscontro nella serie obiettivo, mentre nella *ApEn* ogni pattern veniva invece sempre confrontato anche con se stesso. In questo caso il logaritmo del numero di riscontri non è definito e, di conseguenza, la *cross-ApEn* non è calcolabile. Per ovviare a questo problema, Richman e Moorman (capitolo 3.5.1) hanno sviluppato una versione modificata della *cross-ApEn*, denominata *cross Sample-ApEn* la quale, invece di considerare la somma dei logaritmi del numero di riscontri di tutti i possibili pattern x_i, calcola il logaritmo della somma finale. Così facendo, la *cross Sample-ApEn* risulta definita ogni qual volta vi sia almeno un riscontro per un qualsiasi generico pattern x_i di lunghezza $m+1$. La modifica apportata da Richman e Moorman rappresenta quindi una condizione assai meno restrittiva della precedente e consente una maggiore applicabilità dell'indicatore così modificato. Per tale motivo, nel corso della Parte III della tesi si è preferito utilizzare la *cross Sample-ApEn* in luogo della *cross-ApEn*.

La *cross Sample-ApEn* e la *cross-ApEn*, comunque, misurano entrambe una generica sincronia fra gli elementi delle serie poste a confronto, ovvero sintetizzano la frequenza con cui un generico pattern di lunghezza m ap-

partenente alla serie di confronto che si ritrova in una qualsiasi posizione nella serie obiettivo, vi si ritrovi ancora se si aumenta la sua dimensione di una unità. Al di la della loro definizione, il significato di questi indicatori non è sempre né chiaro né facile da interpretare.

Intanto, a differenza della *ApEn*, il cui massimo assoluto, pari a uno nel caso di serie binarie[1], è esplicativo della condizione di massima casualità e, in definitiva, di indipendenza degli elementi di una serie, la *cross Sample-ApEn* e la *cross-ApEn* possono assumere anche valori maggiori di uno. Inoltre, dagli esperimenti condotti si evince chiaramente che, confrontando coppie diverse di serie fra loro indipendenti, la *cross Sample-ApEn* può assumere una pluralità di valori diversi anche in presenza delle medesime distribuzioni marginali.

La *cross Sample-ApEn* e la *cross-ApEn* non sono poi in grado di rilevare una condizione di perfetta correlazione. Si sono viste, infatti, coppie di serie perfettamente correlate esprimere valori diversi di *cross Sample-ApEn*. Allo stesso modo, a parità di distribuzioni marginali e di livello di perturbazione indotto, serie assolutamente non correlate e serie perfettamente correlate potevano presentare, in media, uno stesso valore di *cross Sample-ApEn*. A maggior ragione, la *cross Sample-ApEn* non è in grado di rilevare una condizione di perfetta correlazione inversa o, comunque, di correlazione inversa in generale.

Per quanto sopra, appare lecito esprimere seri dubbi sul significato reale della *cross Sample-ApEn* e della *cross-ApEn*, nonchè sulle possibili interpretazioni dei loro risultati al di la di una generica indicazione di maggiore o minore sincronia non posizionale fra gli elementi delle serie considerate.

Per colmare le lacune evidenziate ed integrare le informazioni fornite dalla *cross Sample-ApEn* e dalla *cross-ApEn*, si è ritenuto utile modificarne gli algoritmi di calcolo in modo che essi potessero fornire un ulteriore valore numerico definito *indice di corrispondenza* o I_{crs}.

L'indice di corrispondenza, che altro non è che la frequenza di elementi uguali che occupano la medesima posizione all'interno delle serie confrontate, è in grado di dirci se due serie sono fra loro uguali a meno di un margine di tolleranza r, cosa che la *cross-ApEn* non può fare, fornendoci nel contempo anche un valore numerico proporzionale all'intensità del livello di correlazione eventualmente esistente fra le serie stesse. In particolare, in questa forma più semplice, l'indice di corrispondenza varia fra zero, che nel caso di serie binarie corrisponde alla condizione di perfetta correlazione inversa, ed uno, che attesta la perfetta coincidenza fra tutti gli elementi delle serie considerate.

Normalizzando l'indice rispetto al suo minimo e massimo assoluti (i valori zero ed uno) e al suo valore caratteristico p_{ind}, esso diventa maggiormente informativo e di più facile lettura, centrando sul valore zero un'eventuale condizione di indipendenza ed assumendo valori positivi, variabili nell'intervallo

[1] $\max ApEn = \log_2 k$, con k numero di simboli dell'alfabeto considerato

$[1, 0[$, nel caso di correlazione diretta e valori negativi, variabili nell'intervallo $]0, -1]$, nel caso di correlazione inversa.

Per normalizzare l'indice è però necessario conoscerne il valore caratteristico p_{ind}, che corrisponde al valore che l'indice assume in corrispondenza della condizione di indipendenza e che dipende dalla forma delle distribuzioni marginali delle serie considerate. In particolare, se $\{p_u, q_u\}$ è la distribuzione marginale della serie di confronto u e $\{p_v, q_v\}$ quella della serie obiettivo v, $p_{ind} = p_u p_v + q_u q_v$. Per calcolare tale valore non è comunque strettamente necessario disporre delle distribuzioni marginali o di una loro stima. La distribuzione empirica dell'indice data la condizione di indipendenza, così come una stima di p_{ind} possono essere infatti ottenute anche tramite simulazione per mezzo dei metodi Monte Carlo e, nello specifico, tramite la realizzazione di un congruo numero di permutazioni delle serie originarie in modo da simulare altrettante condizioni di indipendenza distributiva su cui poi calcolare il valore dell'indice di corrispondenza.

Si è inoltre dimostrato (capitolo 16.2) che, sotto la condizione di indipendenza, l'indice di corrispondenza non normalizzato relativo a dati binari si distribuisce, se il numero n degli elementi delle è serie sufficientemente elevato, come una normale di media $\mu = p_{ind}$ e deviazione standard $\delta = \sqrt{p_{ind}(1 - p_{ind})/n}$

L'indice così costruito è corretto, in quanto somma di due frequenze (la frequenza delle coppie di elementi corrispondenti uguali $(0, 0)$ e $(1, 1)$) ed è indipendente dal tipo di codifica adottata con l'avvertenza che, nel caso di serie non binarie, il valore zero (meno uno nel caso dell'indice normalizzato) non corrisponderà più alla condizione di perfetta correlazione inversa, ma bensì a quella particolare condizione dove un qualsiasi elemento della serie di confronto non è mai associato allo stesso simbolo nella serie obiettivo.

L'indice di corrispondenza gode inoltre di numerose altre proprietà. Al pari della *ApEn* esso è insensibile al rumore di intensità inferiore al valore di soglia r e risente in maniera minimale di eventuali (pochi) valori anomali (outliers) in quanto non è la loro intensità, bensì il loro numero che si riflette sul valore finale. Esso può essere utilizzato sia nel caso di relazioni lineari che non lineari e sia con dati di tipo binario, che di tipo discreto o continuo. Per il suo calcolo non è necessaria la stima di alcuna distribuzione. Nel caso di normalizzazione, il valore caratteristico p_{ind}, così come la distribuzione sotto la condizione di indipendenza, possono essere ricavati sia per via analitica che mediante simulazione al computer. Esso è inoltre direttamente proporzionale all'intensità dell'eventuale correlazione esistente fra le serie considerate e, nella sua forma normalizzata, evidenza e quantifica in maniera immediata la correlazione inversa per tramite dei valori negativi, proprietà queste che né la S_ρ né la mutua informazione posseggono. L'indice di corrispondenza, infine, può essere utilizzato anche in presenza di serie caotico-deterministiche e si integra perfettamente nell'algoritmo di calcolo della *cross-ApEn* e della *cross Sample-ApEn*.

L'indice di corrispondenza, tuttavia, anche perchè non soggetto a trasformazioni numeriche volte a ridurne la varianza, presenta una variabilità decisamente più elevata rispetto a quella della S_ρ e della muta informazione (vedi tabella 23.1 di cui al precedente capitolo 23), ragion per cui, per avere stime affidabili, si dovrebbe poter disporre di serie mediamente più lunghe. In ogni caso, 5000 elementi appaiono una lunghezza già più che sufficiente per ottenere risultati ampiamente accettabili nelle più diverse situazioni.

Ripetendo con l'indice di corrispondenza gli esperimenti descritti nelle Parti II e III della tesi, si è infatti potuto osservare, in generale, un comportamento tendenziale molto simile a quello della S_ρ e della mutua informazione, con l'eccezione del caso delle serie semplici perfettamente periodiche (capitolo 19), dove l'indice di corrispondenza ha mostrato una maggiore fedeltà all'evidenza teorica e al modello studiato.

L'indice di corrispondenza non è comunque in grado di stabilire, al di la di una eventuale indicazione di direzione e di intensità ed al pari degli altri indicatori studiati, il tipo di relazione che intercorre fra due fenomeni né, tanto meno, di fornire informazioni riguardo alla complessità algoritmica delle serie studiate.

In ogni caso, considerato anche l'esiguo sforzo necessario, è consigliabile calcolare l'indice di corrispondenza ogni qual volta si calcoli la *cross-ApEn* e la *cross Sample-ApEn*, sia per confermare eventuali tendenze messe in luce da tali indicatori, sia per meglio specificare il comportamento delle serie considerate, in particolar modo per quanto riguarda il livello di correlazione eventualmente esistente fra le stesse. Il fatto che l'algoritmo di calcolo ben si integri con quello della *cross-ApEn* nulla toglie al fatto che l'indice di corrispondenza possa essere calcolato anche "stand alone", nel qual caso l'algoritmo di calcolo risulta estremamente semplice e performante e, comunque, per un numero di lag a piacere.

Per concludere, dai risultati ottenuti appare chiaro come tutti gli indicatori studiati, per quanto potenzialmente utilizzabili, non possano essere considerati indicatori universali, tanto meno la *ApEn* e, soprattutto, la *cross-ApEn*. Essi non sono infatti in grado di fornire indicazioni utili ed univoche in tutte le possibili situazioni ma, se considerati tutti assieme, possono fornire ciascuno un utile contributo alla conoscenza generale del fenomeno.

Sarebbe quindi opportuno, per quanto possibile, procedere sempre al calcolo congiunto della mutua informazione o della S_ρ, della *ApEn* e dell'indice di corrispondenza, unitamente ad altri indicatori più specificatamente legati alla materia oggetto di studio, in modo da avere una conferma dei risultati ottenuti ed un aiuto nell'interpretazione degli stessi, cosa il più delle volte tutt'altro che banale. Appare ovvio, comunque, come l'applicazione degli indicatori studiati non possa avvenire in maniera meccanica e svincolata dai singoli contesti analizzati, pena altrimenti l'ottenere risultati anche completamente fuorvianti. Risulta quindi più che mai necessario utilizzare tali

indicatori con estrema cautela ed avendone ben chiari limiti e caratteristiche, oltre al tipo di informazioni che si vogliono e si possono ottenere.

Un'attenzione particolare andrebbe poi prestata in tutte quelle situazioni in cui si debbano confrontare fra loro numeri molto piccoli che, se pur presentando differenze a volte anche delle dimensioni di un ordine di grandezza, risultano comunque molto prossimi allo zero e di intensità assolutamente marginale rispetto all'ampiezza dell'intervallo di variazione degli indicatori considerati. In tali situazioni, quand'anche il confronto dovesse risultare statisticamente significativo, ad una meccanica applicazione di tecniche dovrà preferirsi un approccio critico che tenga conto, volta per volta, delle caratteristiche del fenomeno trattato e delle relative condizioni al contorno.

23.2 Prospettive future

I problemi tutt'ora aperti e che si vorrebbero ancora indagare sono legati soprattutto all'indice di corrispondenza.

In primo luogo, vi è la necessità di studiarne le proprietà e le caratteristiche in un ambito più generale, che ricomprenda quindi anche serie caratterizzate da dati di tipo discreto e continuo. In particolare, anche se la sua distribuzione empirica e una stima del suo valore caratteristico p_{ind} sono sempre ricavabili mediante simulazioni con i metodi Monte Carlo, sarebbe sicuramente importante, anche nel caso di serie non binarie, poter calcolare per via analitica la distribuzione dell'indice sotto la condizione di indipendenza nonché il suo valore caratteristico.

In particolare, qualora si considerino valori del parametro r maggiori di zero, il calcolo per via analitica dei parametri caratterizzanti la distribuzione dell'indice e, quindi, anche di p_{ind}, comporterebbe delicati aspetti di coerenza fra il metodo utilizzato per il calcolo della frequenza di elementi uguali fra le serie analizzate, in quanto vengono considerati uguali due elementi se la loro differenza, in valore assoluto, è inferiore o uguale al valore del parametro r, e quello utilizzato per la determinazione delle distribuzioni marginali, dove r non compare affatto.

Potrebbe poi essere particolarmente interessante valutare eventuali possibili applicazioni dell'indice di corrispondenza alle serie storiche finanziarie dicotomizzate, in particolar modo, ad esempio, all'analisi del segno delle variazioni dell'aggregato economico considerato (perdite o guadagni) e della sua volatilità ("alta" o "bassa").

Bibliografia

[1] A. Alhakim and S. Molchanov. Some markov chains on abelian groups and applications to approximate entropy and to the testing of random number generator. In *Proceeding of the Workshop on Random Walks and Geometry*, Vienna, 2001. Erwin Shrodinger Institute.

[2] Stanislav Anatolyev and Nikolay Gospodinov. Financial return dynamics via decomposition. *Journal of Business and Economic Statistics*, (28):232–245, 2010.

[3] Anthony Bagnall, Chotirat Ratanamahatana, Eamonn Keogh, and Janacek Gareth. A bit level representation for time series data mining with shape based similarity. *Data Mining and Knowledge Discovery*, 13(1):11–40, 2006.

[4] Binay K. Bhattacharya and Godfried T. Toussaint. An upper bound on the probability of misclassification in terms of matusita's measure of affinity. *Ann. Inst. Statist. Math. A*, (34):161–165, 1982.

[5] David R. Brillinger. Some data analyses using mutual information. *Brazilian Journal of Probability and Statistics*, (18):163–182, 2004.

[6] Peter F. Christoffersen and Francis X. Diebold. Financial asset return, direction of change forecasting, and volatility dynamics. *Management Science*, 52(8):1273–1287, 2006.

[7] Giuseppe Cicchitelli. *ProbabilitÃ e statistica*. Maggioli Editore, seconda edizione edition, 1991.

[8] T.M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications, 1991.

[9] Georges A. Darbellay and Diethelm Wuertz. The entropy as a tool for analysing statistical dependences in financial time series. *Physica A*, (287):429–439, 2000.

[10] Andreia Dionisio, Rui Menezes, and Diana A. Mendes. Mutual information: a dependence measure for nonlinear time series. *Physica A: Statistical Mechanics and its Application*, 344(1-2):326–329, 2004.

[11] Andreia Dionisio, Rui Menezes, and Diana A. Mendes. Entropy-based indipendence test. *Nonlinear Dynamics*, (44):351–357, 2006.

[12] Minh N. Do. Fast approximation of kullback-leibler distance for dependence trees and hidden markov model. *IEEE Signal Processing Letters*, 10(4):115–118, 2003.

[13] J. P. Eckmann and D. Ruelle. Ergodic theory of caos and strange attractors. *Reviews of Modern Physics*, 3(3):617–656, 1985.

[14] David P. Feldman. A brief introduction to: Information theory, excess entropy and computational mechanics. Technical report, College of Atlantic, 105 Eden street, Bar Harbour, ME 04609 http:hornacek.coa.edu/dave/, April 1998. revised October 2002.

[15] David P. Feldman. *Computational Mechanics of Classical Spin Systems*. PhD thesis, University of California at Davis, 1998.

[16] A. M. Fraser and H. L. Swinney. Indipendent coordinates for strange attractors from mutual information. *Physical Review A*, (33):1134, 1986.

[17] Diego Luis Gonzales, Simone Giannerini, and Rodolfo Rosa. Detecting structure in parity binary sequences. *IEEE Engineering in Medicine and Biology Magazine*, 25(1):69–81, 2006.

[18] Diego Luis Gonzales, Simone Giannerini, and Rodolfo Rosa. Strong short-range correlations and dichotomic codon classes in coding dna sequences. *Physical Review*, 78(5):051918, 2008.

[19] C.W. Granger, E. Maasoumi, and J. Racine. A dependence metric for nonlinear time series. In *Econometric Society World Congress 2000 Contributed Papers*. Econometric Society, 2000.

[20] C.W. Granger, E. Maasoumi, and J. Racine. A dependence metric for possibly non linear processes. *Journal of Time Series Analysis*, 25(5):649–669, 2004.

[21] Jonh R. Hershey and Peder A. Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP 2007)*, Honolulu, Hawai'i, 2007. Signal Processing Society.

[22] Marcus Hutter. Distribution of mutual information. Technical Report IDSIA-13-01, IDSIA - Swiss Institute for Artificial Intelligence, Galleria 2, CH-6928 Manno-Lugano, Switzerland, December 2001.

[23] H. Joe. Relative entropy measures of multivariate dependence. *Journal of American Statistical Association*, (84):157–164, 1989.

[24] Holger Kantz and Thomas Schreiber. *Nonlinear time series analysis.* Cambridge University Press, 1997.

[25] Alexander Kraskov, Harald Stogbauer, and Peter Grassberger. - estimating mutual information. *Physical Review E*, 69(6):6138, 2004.

[26] A.J. Lawrence and R.C. Wolff. Binary time series generated by chaotic logistic maps. *Stochastics and Dynamics*, 3(4):529–544, 2003.

[27] Peter Y. Liu, Steve M. Pincus, Daniel M. Keenan, Ferdinand Roelfsema, and Johannes D. Veldhuis. Analysis of bidirectional pattern synchrony of concentration-secretion pairs: implementation in the human testicular and adrenal axes. *The American Journal of Physiology*, (288):440–446, 2005.

[28] Esfandiar Maasoumi. A compendium to information theory in economics and econometrics. *Econometrics Reviews*, 12(2):137–181, 1993.

[29] Esfandiar Maasoumi and Jeff Racine. Entropy and predictability of stock market returns. *Journal of Econometrics*, 107(1-2):291–312, 2002.

[30] Kameo Matusita. Decision rules, based on the distance, for problems of fit, two samples, and estimation. *Ann. of Math. Statist.*, 26(4):631–640, 1955.

[31] Young-Il Moon, Balaji Rajagopalan, and Upmanu Lall. Estimation of mutual information using kernel density estimators. *Physical Review E*, 52(3):2318–2321, 1995.

[32] D. Ornstein and B. Weiss. How sampling reveals a process. *Ann. Prob.*, (18):905–930, 1990.

[33] E. Parzen. On estimation of a probability density function and mode. *Ann. of Math. Statist.*, (33):1065–1076–640, 1962.

[34] Steve M. Pincus. Approximate entropy as a measure of system complexity. *Proc. Natl. Acad. Sci. USA*, 88:2297–2301, March 1991.

[35] Steve M. Pincus. Approximate entropy (*apen*) as a complexity measure. *Chaos*, 5(1):110–117, 1995.

[36] Steve M. Pincus. Approximate entropy as an irregularity measure for financial data. *Econometric Reviews*, 27(4-6):329–362, 2008.

[37] Steve M. Pincus, Theodore R. Cummins, and Gabriel G. Haddad. Heart rate control in normal and aborted-sids infants. *The American Journal of Physiology*, (264):638–646, 1993.

[38] Steve M. Pincus, Igor M. Gladstone, and Richard A. Ehrenkranz. A regularity statistic for medical data analysis. *Journal of Clinical Monitoring and Computing*, 7:335–345, 1991.

[39] Steve M. Pincus and Ary Goldberger. Physiological time-series analysis: what does regularity quantify? *The American Journal of Physiology*, (266):1643–1656, 1994.

[40] Steve M. Pincus and Rudolf E. Kalman. Not all (possibly) "random" sequences are created equal. *Proc. Natl. Acad. Sci. USA*, 94:3513–3518, April 1997.

[41] Steve M. Pincus and Burton H. Singer. Randomness and degrees of irregularity. *Proc. Natl. Acad. Sci. USA*, 93:2083–2088, March 1996.

[42] Steve M. Pincus and Burton H. Singer. A recipe for randomness. *Mathematics*, 95:10367–10372, 1998.

[43] Joshua S. Richman and Randall J. Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *The American Journal of Physiology*, (278):2039–2049, 2000.

[44] Ferdinand Roelfsema, Steven M. Pincus, and Johannes D. Veldhuis. Patients with cushing's disease secrete adrenocorticotropin and cortisol jointly more asynchronously than healthy subjects. *Journal of Clinical Endocrinology and Metabolism*, 88(2):688–692, 1998.

[45] Rodolfo Rosa. Massima entropia: E. t. jaynes e dintorni. *Statistica*, Anno XLV(2):181–208, 1985.

[46] Andrew L. Rukhin. Approximate entropy for testing randomness. *Journal of App. Prob.*, 37(1):88–100, 2000.

[47] Claude E. Shannon. A mathematical theory of communication. *Bell System Technology Journal*, (27):379–423, 1948. Come ristampato e

tradotto in "La Teoria Matematica delle Telecomunicazioni", Claude E. Shannon and Warren Weaver, ETAS LIBRI (1983).

[48] Claude E. Shannon and Warren Weaver. *La Teoria Matematica delle Telecomunicazioni*. Il mondo dell'informatica. ETAS LIBRI, seconda edizione italiana edition, Aprile 1983.

[49] Ritei Shibata. Bootstrap estimate of kullback leibler information for model selection. Technical report, Department of Statistics. University of California, Berkeley, California, January 1995.

[50] Taiji Suzuki, Masashi Sugiyama, Jun Sese, and Takafumi Kanamori. Approximating mutual information by maximum likelihood density ration estimation. In *Proceedings of the Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery 2008 (FSDM2008)*, volume 4, pages 5–20. Journal of Machine Learning Research, 2008.

[51] Inder Jet Taneja. *Generalized Information Measures and Their Applications*. On-line book:www.mtm.ufsc.br/∼taneja/book/book.html. Departamento de Matematica. Universidade Federal de Santa Caterina 88040-900 Florianopolis, SC, Brazil, 2001.

[52] N. A. Thacker, F. J. Aherne, and P. I. Rocket. The bhattacharyya metric as an absolute similarity measure for frequency coded data. *Kybernetika*, 34(4):363–368, 1997.

[53] Michele Tumminello, Fabrizio Lillo, and Rosario N. Mantegna. Kullback leibler distance as a measure of the information filtered from multivariated data. *Physical Review E*, 76(3):1123, 2007.

[54] Li Weimin, Liu Liangxu, and Le Jianjin. Clustering streaming time series using CBC. In *Y. Shi et al. (Eds): ICCS 2007, Part III, LNCS 4489*, pages 629–636. Springer-Verlag Berlin Heidelberg, 2007.

[55] Edward N. Zalta. Turing machine. Technical report, Stanford Encyclopedia of Philosophy. Summer 2003 edition, URL=http://plato.stanford.edu/archives/sum2003/entries/turing-machine/.

[56] Tao Zhang, Zhuo Yang, and John H. Coote. Cross-sample entropy as a measure of complexity and regularity of renal sympathetic nerve activity in the rat. *Experimental Physiology*, 92(4):659–669, 2007.