UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI FISICA E ASTRONOMIA "GALILEO GALILEI"
DIPARTIMENTO DI MATEMATICA
CORSO DI LAUREA IN FISICA

ANALISI DELLA NOZIONE DI TRANSFER ENTROPY E APPLICAZIONI

LAUREANDO:
DAVIDE FRISON

RELATORE:
PROF. MARCO FAVRETTI

ANNO ACCADEMICO 2014-2015

Indice

Introduzione

Una grandezza fisica X espressa come variabile aleatoria è una funzione $X\colon \Omega \to S$, con Ω spazio campionario e S spazio degli eventi. In questa tesi considereremo variabili aleatorie discrete, cioè con S insieme discreto finito. L' entropia è una funzione che esprime l'incertezza di una variabile aleatoria. Essa fu introdotta in via del tutto generale da C. Shannon, nell'ambito della teoria dell'informazione, come

$$H(X) = -p(x)\sum p(x)\ln p(x),$$

e dipende dalla sola distribuzione di probabilità di X.

Consideriamo due variabili X e Y. A partire dalle distribuzioni scelte sono possibili ulteriori diverse definizioni di entropia, quali l'entropia congiunta $H(X,Y)$, che rappresenta l'incertezza sul sistema formato da entrambe le variabili, e l'entropia condizionata $H(X \mid Y)$, che rappresenta l'incertezza su X nota la variabile Y. La differenza

$$I(X;Y) = H(X) - H(X \mid Y)$$

viene detta informazione mutua di X e Y, e rappresenta la diminuzione di incertezza su X dovuta all' osservazione di Y.

É possibile estendere questa definizione ad un caso generale, in cui vengono utilizzate più di due variabili. Data una sequenza di valori di una variabile X, e una sequenza di valori della variabile Y, misurati in sequenza temporale discreta da $t = 0$ a t, cioè due processi stocastici discreti, è possibile calcolare la riduzione di incertezza sulla variabile X a un certo istante t, noti gli m valori precedenti della variabile stessa (chiamati $\bar{X}$), ed l valori precedenti della variabile Y (chiamati $\bar{Y}$).

In questo caso è di interesse la differenza tra l'incertezza su $X = X_t$, condizionata $\bar{X}$ e $\bar{Y}$, e l' incertezza su $X = X_t$, condizionata $\bar{X}$, cioè

$$T_{Y\to X}(m,l) = H(X \mid \bar{X},\bar{Y}) - H(X,\bar{X})$$

, che viene detta transfer entropy, introdotta da T. Schreiber nell'articolo *Measuring Information Transfer*, del 2000. Il suo scopo è di analizzare il rapporto causa effetto tra Y e X. Più specificatamente, un valore di $T_{Y\to X}(m,l)$ diverso da zero dice che il flusso di informazione sulla variabile X al tempo t dovuto a $\bar{Y}$, è maggiore rispetto al contributo dato dalla sola $\bar{X}$, cioè che Y è la causa più importante dei valori ottenuti di X. Gli utilizzi della transfer entropy sono svariati, ad esempio nella finanza, come studio dell'influenza del Dow Jones sul Dax 30, ma anche nelle neuroscienze, meteorologia, ecc..

In questa tesi analizzeremo, attraverso alcuni test, qual è il rapporto causa effetto tra due processi stocastici X e Y, in base ai valori ottenuti di transfer entropy. Inoltre mostreremo un'applicazione pratica in ambito biologico, dove studieremo la possibilità di inferire presenza di una specie di piante in una zona di foresta, sulla base di dati di presenza di altre specie.

Capitolo 1

Entropia, covarianza e informazione mutua

1.1 Entropia di Shannon

Sia (Ω, P) spazio di probabilità, con partizione $\Omega = \cup A_i$, $A_i \cap A_j \neq 0$ per $i \neq j$. Sia $p_i = P(A_i)$ la probabilità dell'evento A_i. Vogliamo esprimere l'incertezza associata al verificarsi di uno dei possibili eventi A_i, $i = 1, \ldots, n$.

Nel seguito consideriamo il caso di una partizione di Ω associata a una variabile aleatoria discreta $X \colon \Omega \to S = \{x_1, \ldots, x_n\} \subset \mathbb{R}$. Allora $p(X = x_i) = P(X^{-1}(x_i)) = p_i$, $i = 1, \ldots, n$ e $A_i = X^{-1}(x_i)$.

Possiamo associare a p la funzione $H_n(p)$, che descrive l'incertezza associata a $p_1, \ldots, p_n$, supponendo che $H_n(p)$ abbia la stessa forma funzionale per ogni n. Il teorema di Shannon dice che $H_n(p)$, una volta definiti alcuni caratteri assiomatici che deve ragionevolmente soddisfare, è univocamente definita.

1.1Teorema(C. Shannon) *Sia $H_n(p)$ una funzione definita, per ogni n, sul dominio $p_i \geq 0$, $\sum_{i=1}^n p_i = 1$. Supponiamo*

1) $H_n(p_1, \ldots, p_n)$ continua nelle p_i, per ogni n;

2)$H_n(p_1, \ldots, p_n)$ funzione simmetrica dei suoi argomenti;

3)$H_{n+1}(p_1, \ldots, p_n, 0) = H_n(p_1, \ldots, p_n)$;

4)$H_n(p_1, \ldots, p_n) \leq H_n(\frac{1}{n}, \ldots, \frac{1}{n})$;

5)data $p_i = p_1^{(i)} + \cdots + p_m^{(i)} con \sum_{i=1}^n \sum_{j=1}^m p_j^{(i)} = 1$,

$H_{nm}(p_1^{(1)}, \ldots, p_m^{(n)}) = H_n(p_1, \ldots, p_n) + \sum_{i=1}^n p_i H_m(\frac{p_1^{(i)}}{p_k}, \ldots, \frac{p_m^{(i)}}{p_i})$.

allora

$$H_n(p_1, \ldots, p_n) = -\lambda \sum_{i=1}^{N} p_i \ln p_i. \tag{1.1}$$

I requisiti 1) e 2) sono naturali, poiché gli eventi $X_1, \ldots, X_n$ sono mutualmente esclusivi. La 3) dice che se consideriamo un evento in aggiunta, la cui probabilità di realizzarsi è nulla, allora l' incertezza non cambia. La 4) dice che se la probabilità di ciascun evento è $p_i = \frac{1}{n}$ per ogni i, allora non è possibile distinguere un evento dall'altro e l'incertezza è massima. La 5) afferma che che se ogni evento A_k di probabilità p_k, $k = 1, \ldots, n$, è decomposto in $p_j^{(k)}$ sotto alternative, $j = 1, ..., m$, allora l'incertezza sul sistema è l'incertezza dovuta agli n eventi , più l'incertezza dovuta all'aver considerato m sotto alternative per ogni evento.

Spesso utilizzeremo come notazione $H(X)$ invece di $H(p)$, confondendo $X \sim p(x)$, ricordando comunque che l' entropia è funzione della probabilità di ogni evento.

Proprietà di H

Nel dominio di definizione $D = \{p \in \mathbb{R} \colon p_i \geq 0\}$, $H_n(p)$ è positiva e infinitamente derivabile, grazie alla continuità in zero, $\lim_{p \to 0} p \ln p = 0$, con $HessH(p) = -Diag[\frac{1}{p_i}]$.

Ciò assicura che $H(p)$ sia una funzione concava in D e abbia massimo assoluto in $p_i = \frac{1}{n}, H(\frac{1}{n}) = \lambda \ln n$.

La funzione $H(p)$ è detta entropia di Shannon e rappresenta una descrizione generalizzata del concetto di entropia fisica. Infatti secondo il principio della massima entropia (PME), l'entropia fisica $S(c)$ coincide

con il valore di massimo (unico) di $H(p)$; in questo caso p rende massima l' incertezza $S(p)$, data descrizione del sistema sotto forma di vincoli $f(p) = c$.

L'entropia può essere espressa con basi diverse del logaritmo, e in tal caso subisce solo una trasformazione di scala. Infatti se a e b sono due basi diverse per il logaritmo, allora

$$H_b(p) = \log_b a \log_a p. \tag{1.2}$$

Se la base del logaritmo utilizzata è e, l' entropia viene espressa in nats. Solitamente, quando le variabili hanno come possibili esiti 0 e 1, si utilizza la base 2 del logaritmo, e l' entropia viene espressa in bits.

Entropia come valore di aspettazione

Se una variabile aleatoria discreta X è descritta da una distribuzione $p(x)$, allora il valore di aspettazione di una variabile $g(X)$ è

$$E_p g(X) = \sum_{x \in X} g(x)p(x).$$

In questo caso l'entropia di Shannon rappresenta il valore di aspettazione del reciproco del logaritmo,

$$g(x) = \frac{1}{\ln p(x)}$$

e

$$H(X) = E_p \frac{1}{\ln p(x)}.$$

Nel seguito scriveremo $E(X)$ per rappresentare il valore di aspettazione di una variabile aleatoria X, sottintendendo la distribuzione associata ad X.

Entropia relativa

Consideriamo due distribuzioni di probabilità p e q. L'entropia relativa $D(p \parallel q)$ misura l'inefficienza data dall'aver utilizzato q come distribuzione, sapendo che quella che descrive il fenomeno è p.

Esplicitamente si definisce

$$D(p \parallel q) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}, \tag{1.3}$$

L'entropia relativa è una quantità non negativa. Per dimostrarlo faremo uso della *Disuguaglianza di Jensen*.

1.2 Teorema (Disuguaglianza di Jensen) *Se f è una funzione convessa e X una variabile casuale, allora vale*

$$E(f(X)) \geq f(E(X)) \tag{1.4}$$

In più, se f è strettamente convessa, cioè ha derivata seconda strettamente positiva, l'uguaglianza vale se e solo se

$$X = E(X) \tag{1.5}$$

Notiamo che se f è funzione convessa, allora $-f$ è concava. Perciò il teorema si può estendere anche a funzioni concave, invertendo la disuguaglianza. Se f concava vale

$$E(f(X)) \leq f(E(X)) \tag{1.6}$$

E se $-f$ è strettamente convessa, vale l'uguaglianza sempre per (1.5).

Ora possiamo dimostrare

1.3 Teorema (Disuguaglianza dell'informazione) *Siano $p(x)$ e $q(x)$, $x \in S$ $H_n(p)$, due distribuzioni di probabilità per X. Allora vale*

$$D(p \parallel q) \geq 0 \tag{1.7}$$

con uguaglianza se e solo se

$$p(x) = q(x) \quad \forall x \tag{1.8}$$

Dim.: Applicando la (1.6) alla (1.3), per la funzione $f(t) = \ln t$*, che è strettamente concava, si ottiene*

$$\begin{aligned} -D(p \parallel q) &= -\sum p(x) \ln \frac{p(x)}{q(x)} && (1.9)\\ &= \sum p(x) \ln \frac{q(x)}{p(x)} && (1.10)\\ &\leq \ln \sum p(x) \frac{q(x)}{p(x)} && (1.11)\\ &= \ln \sum q(x) && (1.12)\\ &= \ln 1 && (1.13)\\ &= 0. && (1.14) \end{aligned}$$

1.2 Covarianza e correlazione

Anche se si tratta di nozioni elementari della probabilità, riportiamo le definizioni di covarianza e correlazione, per confrontarle con quelle di informazione mutua e transfer entropy (vedi sezioni successive).

Covarianza

La covarianza tra due variabili aleatorie X e Y, che si indica con $Cov(X,Y)$ o K_{XY} è

$$\begin{aligned} Cov(X,Y) &= E\{[X - E(X)][Y - E[Y]\} \\ &= E(XY) - E(X)E(Y). \end{aligned} \tag{1.15}$$

Se le due variabili sono indipendenti allora $p(x,y) = p(x)p(y)$ e

$$E(XY) = E(X) \cdot E(Y) \tag{1.16}$$

$$Cov(X,Y) = 0 \tag{1.17}$$

e la covarianza risulta nulla. Viceversa se la covarianza è nulla, non è detto che le due variabili siano indipendenti.

Ad esempio consideriamo la variabile X e $Y = g(X) = X^2$ e supponiamo che X abbia una distribuzione simmetrica rispetto allo 0, cioè

$$\begin{aligned} X\colon \Omega \to S = \{0, \pm 1, \pm 2, \ldots, \pm k\} \qquad g\colon S \to K = \{0, 1, 4, \ldots, k^2\} \\ Y = (g \circ X)\colon \Omega \to K = \{0, 1, 4, \ldots, k^2\} \end{aligned} \tag{1.18}$$

per cui

$$p_X(x) = p_X(-x) \qquad p_Y(y = x^2) = p_X(x) + p_X(-x)$$

Allora

$$\begin{aligned} Cov(X,Y) &= E(X \cdot X^2) - E(X) \cdot E(X^2) \\ &= \sum p(x)x^3 - (\sum p(x)x)(\sum p(x)x^2) \\ &= 0 - 0 \cdot (\sum p(x)x^2) = 0 \end{aligned} \tag{1.19}$$

Dunque l'annullarsi della covarianza è una condizione necessaria ma non sufficiente affinché due variabili siano statisticamente indipendenti.

Vediamo come si esprime la varianza di una variabile in funzione della covarianza.

Consideriamo le variabili

$$X_1, \ldots, X_n \qquad Z = F(X_1, \ldots, X_n)$$

Supponiamo che $F(X_1, \ldots, X_n)$ si possa espandere in serie di Taylor attorno a un punto $(\bar{X}_1, \ldots, \bar{X}_n)$. Allora

$$F(X_1, \ldots, X_n) = (\bar{X}_1, \ldots, \bar{X}_n) + \sum_i \frac{\partial F}{\partial X_i}(X_i - \bar{X}_i).$$

Se definiamo

$$Var(Z) = Var(F) = E(Z^2) - E(Z)^2,$$

allora

$$\begin{aligned}
Var(Z) &\approx E\{[(\bar{X_1},\ldots,\bar{X_n}) + \sum_i \frac{\partial F}{\partial X_i}(X_i - \bar{X_i})]^2\} \\
&- [E(\bar{X_1},\ldots,\bar{X_n}) + \sum_i \frac{\partial F}{\partial X_i}(X_i - \bar{X_i})]^2 \\
&= E[(\bar{X_1},\ldots,\bar{X_n})^2] + E\{[\sum_i \frac{\partial F}{\partial X_i}(X_i - \bar{X_i})]^2\} \\
&+ 2E[(\bar{X_1},\ldots,\bar{X_n})\sum_i \frac{\partial F}{\partial X_i}(X_i - \bar{X_i})] \\
&- [E(\bar{X_1},\ldots,\bar{X_n})]^2 - 2E\sum_i \frac{\partial F}{\partial X_i}(X_i - \bar{X_i})E(\bar{X_1},\ldots,\bar{X_n}) \\
&- [E\sum_i \frac{\partial F}{\partial X_i}(X_i - \bar{X_i})]^2
\end{aligned}$$

Usando il fatto che $(\bar{X_1},\ldots,\bar{X_n})$ è composti da elementi costanti, per cui

$$E[(\bar{X_1},\ldots,\bar{X_n})^2] - [E(\bar{X_1},\ldots,\bar{X_n})]^2 = 0,$$

$$+2E[(\bar{X_1},\ldots,\bar{X_n})\sum_i \frac{\partial F}{\partial X_i}(X_i - \bar{X_i})] - 2E\sum_{substacki} \frac{\partial F}{\partial X_i}(X_i - \bar{X_i})E(\bar{X_1},\ldots,\bar{X_n}) = 0.$$

Dunque

$$\begin{aligned}
Var(Z) &\approx E[\sum_i \frac{\partial F}{\partial X_i}(X_i - \bar{X_i})]^2 - [E\sum_i \frac{\partial F}{\partial X_i}(X_i - \bar{X_i})]^2 \\
&= \sum_i E[\frac{\partial F}{\partial X_i}(X_i - \bar{X_i})]^2 + 2E\sum_{i<j} \frac{\partial F}{\partial X_i}\frac{\partial F}{\partial X_j}(X_i - \bar{X_i})(X_j - \bar{X_j}) \\
&- \sum_i \{E[\frac{\partial F}{\partial X_i}(X_i - \bar{X_i})]\}^2 - 2\sum_{i<j} E[\frac{\partial F}{\partial X_i}\frac{\partial F}{\partial X_j}(X_i - \bar{X_i})(X_j - \bar{X_j})] \\
&= E\sum_i \{[\frac{\partial F}{\partial X_i}(X_i - \bar{X_i})]^2 - E[\frac{\partial F}{\partial X_i}(X_i - \bar{X_i})]^2\} \\
&+ 2E\sum_{i<j} \{[\frac{\partial F}{\partial X_i}\frac{\partial F}{\partial X_j}(X_i - \bar{X_i})(X_j - \bar{X_j})] - E[\frac{\partial F}{\partial X_i}\frac{\partial F}{\partial X_j}(X_i - \bar{X_i})(X_j - \bar{X_j})]\}
\end{aligned}$$

Ricordiamo che i $\frac{\partial F}{\partial X_i}$ sono valutati a $X_i = \bar{X_i}$, perciò costanti e possono essere portati fuori dall'operatore di aspettazione

$$\begin{aligned}
Var(Z) &\approx \sum_i (\frac{\partial F}{\partial X_i})^2 E[X_i^2 + \bar{X_i}^2 - 2X_i\bar{X_i} - E(X_i^2 + \bar{X_i}^2 - 2X_i\bar{X_i})] \\
&+ 2\sum_{i<j} \frac{\partial F}{\partial X_i}\frac{\partial F}{\partial X_j} E[X_iX_j + \bar{X_i}\bar{X_j} \\
&- 2X_i\bar{X_j} - 2X_j\bar{X_i} - E([X_iX_j + \bar{X_i}\bar{X_j} - 2X_i\bar{X_j} - 2X_j\bar{X_i})]. \\
Var(Z) &= \sum_i (\frac{\partial F}{\partial X_i})^2 Var(X) + 2\sum_{i<j} \frac{\partial F}{\partial X_i}\frac{\partial F}{\partial X_j} Cov(X_i, X_j) \qquad (1.20)
\end{aligned}$$

Per esprimere in modo compatto la 1.20 introduciamo la matrice delle covarianze degli X_i,

$$V_{i,j} = E(X_iX_j) - E(X_i)E(X_j)$$

di ordine $N \times N$ i cui elementi (i, j) generici sono le covarianze, e in particolare gli elementi in diagonale sono le varianze degli X_i. Se $\frac{\partial F}{\partial X_i} = F_i$, e F_i^T è il trasposto, si ha

$$Var(Z) \approx \sum_{i,j} F_i^T V_{i,j} F_i \tag{1.21}$$

Ad esempio consideriamo il caso di relazione lineare, $Z = aX + bY$. Allora

$$\frac{\partial F}{\partial X} = a, \frac{\partial F}{\partial Y} = b$$

$$\frac{\partial F}{\partial X}\frac{\partial F}{\partial Y} = ab$$

Perciò

$$Var(Z) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y),$$

poiché nel caso lineare l'espansione di Taylor è esatta.

Con $Z = \sum_{i=1}^{n} a_i X_i$ si possono utilizzare il vettore $A = (a_1, \ldots, a_n)$ e il suo trasposto A^T, per esprimere

$$Var(z) = \sum_{i,j} A_i^T V_{i,j} A_j \tag{1.22}$$

Covarianza campionaria

$$Cov_{campionaria}(X, Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N - 1}, \tag{1.23}$$

dove $\bar{x}$ e $\bar{y}$ sono i valori medi di X e Y estratti da un campione di N dati,

$$\bar{x} = \sum_{i=1}^{N} \frac{x_i}{N} \qquad \bar{y} = \sum_{i=1}^{N} \frac{y_i}{N}.$$

A differenza della covarianza definita precedentemente, che viene detta teorica in contrapposizione, la covarianza campionaria non assume nota la distribuzione di probabilità delle variabili X e Y, e per il calcolo utilizza le frequenze relative.

Correlazione

La correlazione tra due variabili aleatorie fornisce informazioni riguardo l'andamento di una variabile rispetto all' altra.

Il coefficiente di correlazione lineare, detto anche coefficiente di correlazione di Pearson, è definito come

$$r_{x,y} = Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(X, Y)}{\sigma_x \sigma_y} \tag{1.24}$$

ed è adimensionale. Risulta nullo se la covarianza è nulla.

Inoltre dato che

$$Var(X) \geq 0, Var(Y) \geq 0,$$

se prendiamo una variabile casuale Z definita come

$$Z = \sigma_x Y - \sigma_y X,$$

si ha

$$Var(Z) = \sigma_x^2 Var(Y) + \sigma_y^2 Var(X) - 2\sigma_x \sigma_y Cov(X, Y) \geq 0,$$

dunque

$$\begin{aligned} \sigma_x \sigma_y Cov(X, Y) &\leq Var(X)Var(Y), \\ r_{x,y}\sigma_x^2 \sigma_y^2 &\leq Var(X)Var(Y), \\ r_{x,y} &\leq 1 \end{aligned}$$

Ripetendo per $Z = \sigma_x Y + \sigma_y X$ si trova

$$r_{x,y} \geq -1$$

Perciò $r_{x,y}$ ha come valori estremi -1 e 1. Per questo viene detto anche coefficiente di correlazione lineare: misura la dipendenza di due variabili con dipendenza generica e verifica se questa è vicina alla linearità. Quando $r_{x,y}$ raggiunge uno dei valori estremi ± 1, significa che la dipendenza è lineare. Ad esempio se $Y = a + bX$, allora

$$\begin{aligned} E(Y) &= a + bE(X), \\ Var(y) &= b^2 Var(x), \\ Cov(X,Y) &= E(XY) - E(X) \cdot E(Y) \\ &= E(aX + bX^2) - aE(X) - bE^2(X) \\ &= b[E(X^2) - E^2(X)] \end{aligned}$$

e il coefficiente di correlazione è

$$r_{x,y} = \frac{bVar(X)}{|b|\, Var(X)} = \pm 1 \tag{1.25}$$

a seconda che b sia positivo o negativo.

L'esempio mostra che il coefficiente d' angolo della retta è proprio il coefficiente di correlazione.

Uno strumento d' analisi per verificare la dipendenza lineare è quello dell'interpolazione lineare: si grafica una variabile in funzione dell'altra e se i dati si distribuiscono in modo quasi-lineare, si fittano con una retta secondo l'algoritmo di interpolazione lineare, che si basa sul metodo dei minimi quadrati. Il coefficiente d'angolo trovato è il coefficiente di correlazione.

Per utilizzare l'interpolazione devono essere soddisfatte le seguenti ipotesi:

1. esiste una relazione $Y = a + bX$;
2. X è affetta da errori trascurabili, da cui l' uguaglianza sopra;
3. gli errori quadratici medi delle possibili alternative di Y, ignoti, sono tutti uguali, o comunque molto simili e sono casuali;
4. i possibili eventi di una variabile sono indipendenti tra loro;

Il coefficiente $r_{x,y}$ che si ottiene non rappresenta il coefficiente di correlazione tra gli errori di X e Y, bensì il coefficiente di correlazione per l' insieme di punti (X_i, Y_i), cioè rappresenta la correlazione tra una coppia e l' altra. Grazie ad esso è possibile ricavare l' errore a posteriori sulla grandezza Y. Interpolare $Y(X)$ o $X(Y)$ fornisce lo stesso risultato finale, se trascuriamo gli errori della variabile nelle ascisse.

Se l' incertezza degli eventi X_i non è trascurabile, si utilizza l' interpolazione lineare pesata, e in questo caso il fit cambia a seconda che si utilizzi X o Y nelle ascisse. In questo caso nel calcolo dell'errore a posteriori intervengono gli errori sugli Y_i calcolati per propagazione

$$Var(Y) = \frac{\partial F}{\partial X} Var(X) = b^2 Var(X).$$

Se X e Y non hanno una relazione lineare, si può provare a linearizzarla con un cambio di parametro: ad esempio se $Y = a \log X$, si può prendere l'esponenziale dei valori assunti da Y e utilizzare $Z = e^Y$ e fare un fit con coppie di valori (X, Z). Una volta calcolato l' errore a posteriori su Z, si potrà risalire all'errore su Y per propagazione. Nell'esempio

$$Var(Y) = \frac{\partial \log(Z)}{\partial Z} Var(Z) = \frac{1}{Z} Var(Z)$$

.

Esistono altre procedure di fit, ad esempio quello polinomiale, che applicano il metodo dei minimi quadrati a funzioni generiche candidate a rappresentare la dipendenza tra le variabili aleatorie.

Per controllare qual è il fit che meglio descrive la dipendenza dei dati si fa riferimento alle tabelle del χ^2. A seconda del valore del χ^2 ottenuto e del numero di gradi di libertà, si ottiene un range di probabilità per cui il fit risulta più o meno corretto.

1.3 Informazione mutua

Introduciamo l'*entropia congiunta*,

$$H(X \cap Y) = H(X, Y) = -\sum p(x, y) \ln p(x, y). \tag{1.26}$$

Essa rappresenta l'incertezza associata a due variabili aleatorie X e Y.

L'incertezza su X, sapendo che Y ha avuto esito y, è:

$$H(X \mid Y = y) = -\sum p(x \mid y) \ln p(x \mid y), \tag{1.27}$$

dove

$$p(x \mid y) = \frac{p(x, y)}{p(y)}$$

è la probabilità condizionata.

Si può definire l'*entropia condizionata* come l'incertezza media ottenuta dall'osservazione di X, noto $Y = y$, pesato su tutti i possibili micro-stati della variabile Y:

$$\begin{aligned} H(X \mid Y) &= -\sum p(y) H(X \mid Y = y) \\ &= -\sum p(x, y) \ln p(x \mid y). \end{aligned} \tag{1.28}$$

In particolare da queste definizioni discende che

$$\begin{aligned} H(X, Y) &= -\sum p(x, y) \ln \frac{p(x, y)p(y)}{p(y)} \\ &= -\sum p(x, y) \ln \frac{p(x, y)p(x)}{p(x)} \\ &= -\sum p(x, y) \ln \frac{p(x, y)}{p(y)} - \sum p(x, y) \ln p(y) \\ &= -\sum p(x, y) \ln \frac{p(x, y)}{p(x)} - \sum p(x, y) \ln p(x) \\ &= H(X \mid Y) + H(Y) = H(Y \mid X) + H(X). \end{aligned} \tag{1.29}$$

Se due variabili sono s.i., $p(x, y) = p(x)p(y)$, allora l' entropia congiunta diventa

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

In generale nota $p(x, y)$, è possibile calcolare le marginali $p(x)$ e $p(y)$. Il viceversa è falso. Dunque $p(x, y)$ contiene più informazione su X, Y rispetto a $p(x)$ e $p(y)$.

La differenza tra queste due è detta *informazione mutua*, ed è definita come

$$I(X; Y) = H(X) + H(Y) - H(X, Y) \tag{1.30}$$

L'informazione mutua è una funzione che misura la riduzione di incertezza su X dovuta alla conoscenza di Y. Si nota subito, scambiando X con Y, che l'informazione mutua è simmetrica

$$I(X; Y) = I(Y; X) \tag{1.31}$$

Un' altra forma in cui si può esprimere l'informazione mutua è

$$I(X; Y) = -\sum p(x) \ln p(x) - \sum p(y) \ln p(y) + \sum p(x, y) \ln p(x, y) \tag{1.32}$$

$$= \sum p(x, y) \ln \frac{p(x, y)}{p(x)p(y)} \tag{1.33}$$

$$= D(p(x, y) \parallel p(x)p(y)) \tag{1.34}$$

e rappresenta il guadagno di informazione dovuta dall'aver assunto $p(x)p(y)$ come distribuzione per X e Y, sapendo che quella vera è $p(x,y)$. Equivalentemente essa rappresenta il guadagno di informazione dato dall'aver osservato contemporaneamente X e Y. Infatti dato che $H()$ è una quantità non negativa, il termine $H(X,Y)$ riduce l'incertezza dovuta ai termini $H(X)$ e $H(Y)$.

Si nota facilmente, usando (1.33), che altre due forme equivalenti sono

$$I(X;Y) = H(X) - H(X \mid Y) \tag{1.35}$$
$$= H(Y) - H(Y \mid X) \tag{1.36}$$

Segue che se $Y = X$

$$I(X;X) = H(X) - H(X \mid X) = H(X). \tag{1.37}$$

L'informazione mutua è *non negativa*, e deriva da (1.7) applicata a (1.34), perciò

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \geq 0 \tag{1.38}$$

Inoltre da (1.34) *è nulla se e solo se* $p(x,y) = p(x)p(y)$*, cioè* X *e* Y *sono s.i.*.

Grazie a (1.38) possiamo notare che

$$H(Y) \geq H(Y \mid X) \tag{1.39}$$
$$H(X) \geq H(X \mid Y) \tag{1.40}$$

cioè l'entropia condizionata è minore di quella non condizionata.

In generale data una variabile X e un' altra variabile $Y = g(X)$, appena è noto il valore di X lo è anche quello di $g(X)$, ma non viceversa, cioè

$$H(g(X) \mid X) = 0 \qquad H(X \mid g(X)) \neq 0 \tag{1.41}$$

Perciò applicando (1.29), si ha

$$\begin{aligned} H(X, g(X)) &= H(X) + H(g(X) \mid X) = H(g(X)) + H(X \mid g(X)) \\ &= H(X) = H(g(X)) + H(X \mid g(X)) \end{aligned}$$

Ora utilizzando (1.40) con $Y = g(X)$ si ottiene facilmente

$$H(g(X)) \leq H(X) \tag{1.42}$$

In questo capitolo abbiamo trovato che due variabili sono s.i. se e solo se l'informazione mutua è nulla. Inoltre sappiamo anche che se le variabili sono s.i., anche la covarianza è nulla. Dunque

$$I(X;Y) = 0 \Rightarrow Cov(X,Y) = 0 \tag{1.43}$$
$$\tag{1.44}$$

Invece non vale l'implicazione opposta, ed è dovuto al fatto che $Cov(X,Y) = 0$ è una condizione necessaria ma non sufficiente affinché due variabili siano s.i..

Queste proprietà si possono riassumere in

$$X, Y \text{ s.i.} \Rightarrow Cov(X,Y) = 0 \qquad Cov(X,Y) = 0 \not\Rightarrow X, Y \text{ s.i.} \tag{1.45}$$
$$I(X;Y) = 0 \iff X, Y \text{ s.i.} \tag{1.46}$$
$$\tag{1.47}$$

Informazione mutua condizionata

Date tre variabili X, Y, Z, l'informazione mutua condizionata esprime la differenza tra l'incertezza presente su X, noto Z, e l'incertezza su X noti Y e Z.

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z) \tag{1.48}$$
$$= H(X \mid Z) - H(X) + H(X) - H(X \mid Y, Z) \tag{1.49}$$
$$= I(X; Y, Z) - I(X; Z) \tag{1.50}$$

Nel caso di n variabili $X_1, \dots, X_n$, l'entropia congiunta si può esprimere con la *regola della catena per l' entropia:*

$$\begin{aligned}
H(X_1, \dots X_n) &= -\sum p(x_1, \dots, x_n) \ln p(x_1, \dots, x_n) \\
&= H(X_1) + H(X_2, \dots, X_n \mid X_1) \\
&= H(X_1) - \sum p(x_1, \dots, x_n) \ln \frac{p(x_1, \dots, x_n)}{p(x_1)} \frac{p(x_1, x_2)}{p(x_1, x_2)} \\
&= H(X_1) - \sum p(x_1, \dots, x_n) \ln \frac{p(x_1, \dots, x_n)}{p(x_1, x_2)} - \sum p(x_1, \dots, x_n) \ln \frac{p(x_1, x_2)}{p(x_1)} \\
&= H(X_1) + H(X_2 \mid X_1) + H(X_3, \dots, X_n \mid X_1 \cap X_2) \\
&= H(X_1) + H(X_2 \mid X_1) - \sum p(x_1, \dots, x_n) \ln \frac{p(x_1, \dots, x_n)}{p(x_1, x_2)} \frac{p(x_1, x_2, x_3)}{p(x_1, x_2, x_3)} \\
&= H(X_1) + H(X_2 \mid X_1) + H(X_3 \mid X_1, X_2) + H(X_4, \dots, X_n \mid X_1, X_2, X_3).
\end{aligned}$$

Iterando nell'ultimo termine si arriva alla forma

$$\begin{aligned}
H(X_1, \dots X_n) &= H(X_1) + H(X_2 \mid X_1) + \cdots + H(X_n \mid X_1, \dots X_{n-1}) \\
&= \sum_{i=1}^{N} H(X_i \mid X_1, X_2, \dots, X_{i-1}).
\end{aligned} \tag{1.51}$$

L'espressione (1.50) diventa, nel caso siano noti m valori di Y e l valori di Z:

$$\begin{aligned}
I(X; Y_1, \dots, Y_m \mid Z_1, \dots, Z_n) &= H(X \mid Z_1, \dots, Z_l) - H(X \mid Y_1, \dots, Y_m, Z_1, \dots, Z_l) && (1.52) \\
&= H(X \mid Z_1, \dots, Z_l) - H(X) \\
&\quad + H(X) - H(X \mid Y_1, \dots, Y_m, Z_1, \dots, Z_l) \\
&= I(X; Y_1, \dots, Y_m, Z_1, \dots, Z_l) - I(X; Z_1, \dots, Z_l) && (1.53)
\end{aligned}$$

e servirà per la definizione di transfer entropy. In particolare anche l' informazione mutua condizionata è non negativa. Per notarlo scriviamo in forma compatta

$$\bar{Y} = (Y_1, \dots, Y_l) \qquad \bar{Z} = (Z_1, \dots, Z_l),$$

e scriviamo in forma estesa la (1.52) utilizzando la (1.28) nel caso di molte variabili. Si ottiene

$$\begin{aligned}
I(X; \bar{Y} \mid \bar{Z}) &= -\sum p(x, \bar{z}) \ln p(x \mid \bar{z}) + \sum p(x, \bar{y}, \bar{z}) \ln p(x \mid \bar{y}, \bar{z}) && (1.54) \\
&= \sum p(x, \bar{y}, \bar{z}) \ln \frac{p(x \mid \bar{y}, \bar{z})}{p(x \mid \bar{z})} && (1.55) \\
&= \sum p(x, \bar{y}, \bar{z}) \ln \frac{p(x, \bar{y}, \bar{z})}{p(\bar{y}, \bar{z}) p(x \mid \bar{z})} && (1.56) \\
&= D(p(x, \bar{y}, \bar{z}) \parallel p(\bar{y}, \bar{z}) p(x \mid \bar{z})) && (1.57)
\end{aligned}$$

Di nuovo, grazie a (1.7), si vede subito che anche l'informazione mutua condizionata è non negativa. In particolare è nulla se e solo se

$$p(x \mid \bar{y}, \bar{z}) = p(x \mid \bar{z}) \tag{1.58}$$

cioè X *è indipendente da* $\bar{Y}$ *condizionatamente* $\bar{Z}$.

Capitolo 2

Transfer Entropy

L'informazione mutua è una quantità simmetrica, dunque non ha direzionalità. Essa permette di dire se c'è dipendenza tra due variabili X ed Y e quantifica l'informazione condivisa tra di esse. Non permette però di sapere se Y contiene più informazione su X o viceversa, cioè se è più importante il contributo di Y a X o di X ad Y.

Supponiamo di avere due variabili aleatorie di cui effettuiamo misure in sequenza temporale discreta a step di unità temporali τ. In questo modo la X e la Y rappresentano due processi stocastici discreti, in cui ognuna delle due, considerata a un certo tempo t fissato, rappresenta una variabile aleatoria. Se iniziamo a misurare a t_0, avremo $t_m = t_0 + m\tau$ al generico istante m-esimo. Per semplicità poniamo $t_0 = 0$ e $\tau = 1$. In questo modo si avanza a step di tempi unitari dall'istante zero. Poniamo

$$X = X_t \qquad \bar{X} = (X_{t-m}, \dots, X_{t-1}) \qquad \bar{Y} = (Y_{t-l}, \dots, Y_{t-1})$$

.

La transfer entropy è definita come

$$T_{Y\to X}(m,l) = H(X \mid \bar{X}) - H(X \mid \bar{X}, \bar{Y}) \tag{2.1}$$
$$= I(X; \bar{X}, \bar{Y}) - I(X; \bar{X}) \tag{2.2}$$

facendo riferimento a (1.53). Esplicitamente

$$T_{Y\to X}(m,l) = \sum p(x, \bar{x}, \bar{y}) \ln \frac{p(x \mid \bar{x}, \bar{y})}{p(x \mid \bar{x})} \tag{2.3}$$
$$= \sum p(x, \bar{x}, \bar{y}) \ln \frac{p(x, \bar{x}, \bar{y}) p(\bar{x})}{p(\bar{x}, \bar{y}) p(x, \bar{x})} \tag{2.4}$$
$$= \sum p(x, \bar{x}, \bar{y}) \ln \frac{p(x, \bar{x}, \bar{y})}{p(\bar{x}, \bar{y}) p(x \mid \bar{x})} \tag{2.5}$$
$$= D(p(x, \bar{x}, \bar{y}) \parallel p(\bar{x}, \bar{y}) p(x \mid \bar{x})) \tag{2.6}$$

Dunque è definita *non negativa* ed è *nulla se e solo se*

$$p(x \mid \bar{x}, \bar{y}) = p(x \mid \bar{x}), \tag{2.7}$$

Questo *significa che X è indipendente da $\bar{Y}$ condizionatamente a $\bar{X}$.*

Scriviamo questa proprietà come

$$T_{Y\to X} = 0 \iff X \text{ indipendente da } \bar{Y} \text{ condizionatamente a } \bar{X} \tag{2.8}$$
$$T_{X\to Y} = 0 \iff Y \text{ indipendente da } \bar{X} \text{ condizionatamente a } \bar{Y} \tag{2.9}$$

Vediamo alcuni casi particolari in cui è possibile calcolare la transfer entropy dalla relazione tra X e Y.

1. Sia X indipendente da $\bar{Y}$, e $\bar{X}$ qualsiasi. Allora vale la (2.7) e la transfer entropy $T_{Y\to X}(m,l)$ è nulla. Dunque

$$X, \bar{Y} \ \ s.i. \Rightarrow T_{Y\to X} = 0 \tag{2.10}$$
$$Y, \bar{X} \ \ s.i. \Rightarrow T_{X\to Y} = 0 \tag{2.11}$$

2. Se invece $X = g(\bar{X})$ e $\bar{Y}$ qualsiasi, allora

$$H(X \mid \bar{X}) = H(g(\bar{X}) \mid \bar{X}) = 0 \qquad H(X \mid \bar{X}, \bar{Y}) = 0$$

per (1.41). Inserendo in (2.1) si ha

$$T_{Y\to X}(m,l) = 0 \tag{2.12}$$

Perciò $T_{Y\to X}(m,l)$ *nulla è una condizione necessaria ma non sufficiente affinché* $\bar{Y}$ *e* X *siano indipendenti.* Dunque

$$T_{Y\to X} = 0 \Rightarrow X, \bar{Y} \ \textit{ s.i. condizionatamente } \ \bar{X} \tag{2.13}$$
$$T_{X\to Y} = 0 \Rightarrow X, \bar{Y} \ \textit{ s.i. condizionatamente } \ \bar{X} \tag{2.14}$$

3. Sia X indipendente da $\bar{X}$, e $\bar{Y}$ generica. Allora

$$p(x \mid \bar{x}, \bar{y}) = p(x \mid \bar{y}) \quad p(x \mid \bar{x}) = p(x) \quad p(x, \bar{x}, \bar{y}) = p(x, \bar{y})p(\bar{x})$$

Per cui vale

$$T_{Y\to X}(m,l) = \sum p(x, \bar{y})p(\bar{x}) \ln \frac{p(x \mid \bar{y})}{p(x)} \tag{2.15}$$
$$= \sum p(x)p(x, \bar{y}) \ln \frac{p(x \mid \bar{y})}{p(x)} \tag{2.16}$$
$$= \sum p(x, \bar{y}) \ln p(x \mid \bar{y} - \sum p(x, \bar{y}) \ln p(x) \tag{2.17}$$
$$= H(X) - H(X \mid \bar{Y}) \tag{2.18}$$
$$= I(X; \bar{Y}) \tag{2.19}$$

Notiamo che anche se vale

$$I(X; \bar{Y}) = I(\bar{Y}; X),$$

invece

$$T_{Y\to X}(m,l) \neq T_{X\to Y}(m,l),$$

e ciò è dovuto alla definizione utilizzata per la transfer entropy. Infatti sotto le stesse condizioni si ha che

$$T_{X\to Y}(m,l) = \sum p(y, \bar{y}, \bar{x}) \ln \frac{p(y \mid \bar{y}, \bar{x})}{p(\bar{y} \mid \bar{x})} \tag{2.20}$$

Dato che X è indipendente da $\bar{X}$, ma $\bar{Y}$ è generica, non si può ridurre la (2.20) alla (2.19).

4. Consideriamo due variabili X e Y in cui ogni esito di X è indipendente da ogni altro suo esito, e lo stesso per Y. Invece imponiamo che $X_t = g(Y_{t-k}) =$ per $k \geq 1$. Allora valgono $p(x \mid \bar{x}, \bar{y}) = p(x \mid \bar{y})$, $p(x \mid \bar{y}) = p(x)$. Notiamo che anche $\bar{X}$ dipende da $\bar{Y}$ in generale. Allora se $\bar{y} = (y_{t-1}, \ldots, y_{t-l})$, se $l \geq k$ il valore x è contenuto nel vettore $\bar{y}$ e sono dipendenti. Invece se $l < k$ $\bar{y}$ non contiene x e x e $\bar{y}$ sono s.i.. Perciò

   (a) $l < k$
   $p(x \mid \bar{x}, \bar{y}) = p(x \mid \bar{y}) = p(x)$. Allora la transfer entropy diventa

$$\begin{aligned} T_{Y\to X}(m,l) &= \sum p(x, \bar{x}, \bar{y}) \ln \frac{p(x \mid \bar{x}, \bar{y})}{p(x \mid \bar{x})} \\ &= \sum p(x, \bar{x}, \bar{y}) \ln \frac{p(x)}{p(x))} \\ &= \sum p(x, \bar{x}, \bar{y}) \ln 1 \\ &= 0 \end{aligned}$$

(b) $l \geq k$

$p(x \mid \bar{x}, \bar{y}) = p(x \mid \bar{y}) = p(g(y_{t-k} \mid \bar{y}) = 1$. In questo caso la transfer entropy è

$$\begin{aligned} T_{Y\to X}(m,l) &= \sum p(x,\bar{x},\bar{y}) \ln \frac{p(x \mid \bar{x},\bar{y})}{p(x \mid \bar{x})} \\ &= \sum p(x,\bar{x},\bar{y}) \ln \frac{1}{p(x)} \\ &= -\sum p(x) \ln p(x) \\ &= H(X) \\ &= H(g(Y_{t-k})) \end{aligned}$$

Calcoliamo ora $T_{X\to Y}(m,l)$. Dato che le variabili sono dipendenti tramite $X_t = Y_{t-k}$, risulta anche $Y_t = X_{t+k}$. Perciò X_{t+k} non è presente in $\bar{X}$, e Y e $\bar{X}$ sono s.i.. Inoltre anche Y e $\bar{Y}$ sono s.i. per definizione. Perciò $p(y \mid \bar{y}, \bar{x}) = p(y)$ e $p(y \mid \bar{y}) = p(y)$. Dunque si ha

$$\begin{aligned} T_{X\to Y}(m,l) &= \sum p(y,\bar{y},\bar{x}) \ln \frac{p(y \mid \bar{y},\bar{x})}{p(y \mid \bar{y})} \\ &= \sum p(y,\bar{y},\bar{x}) \ln \frac{p(y)}{p(y)} \\ &= \sum p(y,\bar{y},\bar{x}) \ln 1 \\ &= 0 \quad \forall (m,l) \end{aligned}$$

2.1 Test con variabili indipendenti: Lancio di due monete

In questo test vogliamo calcolare informazione mutua e transfer entropy tra due variabili s.i.. In particolare vogliamo verificare che entrambe sono nulle, perché X non dipende da Y e da $\bar{Y}$, e viceversa per le Y.

Consideriamo una coppia di variabili aleatorie indipendenti X e Y che rappresentano l' esito del lancio di due monete. L'esito testa sarà rappresentato dal valore 1, l' esito croce dal valore 0. La probabilità congiunta degli eventi di X e Y deve soddisfare $p(x,y) = p(x)p(y)$ per l'indipendenza.

$p_X(0) = p_X(1) = \frac{1}{2}$
$p_Y(0) = p_Y(1) = \frac{1}{2}$
$p_{X,Y}(0,0) = p_{X,Y}(0,1) = p_{X,Y}(1,0) = p_{X,Y}(1,1) = \frac{1}{4}$

Dalle definizioni di informazione mutua e covarianza, e transfer entropy deriva subito che

$$\begin{aligned} Cov(X,Y) &= 0, \\ I(X;Y) &= 0 \\ T_{Y\to X}(m,l) &= 0 \\ T_{Y\to X}(m,l) &= 0 \end{aligned}$$

In particolare per variabili s.i., le transfer entropy sono nulle per qualsiasi coppia di valori m, l.
Simuliamo tredici lanci di due monete con un programma C++:

X	1	0	0	1	0	1	0	1	1	1	0	0	1
Y	0	1	0	1	0	0	0	0	1	1	1	0	0

Grazie al *Teorema Centrale del Limite*, sappiamo che all'aumentare del numero N di lanci, le frequenze relative si avvicineranno alle probabilità date dalla distribuzione.

	p_X	p_Y
0	$\frac{6}{13}$	$\frac{8}{13}$
1	$\frac{7}{13}$	$\frac{5}{13}$

Le frequenze $p_{X,Y}(x,y)$ sono le frequenze relative di eventi contemporanei, cioè eventi di X e Y che accadono allo stesso tempo. Ad esempio la coppia formata dalla casella 3 di X e casella 3 di Y è $(0,0)$ e si ripete 4 volte in $n = 13$ eventi.

$p_{X,Y}$			
$(0,0)$	$\frac{4}{13}$	$(1,0)$	$\frac{4}{13}$
$(0,1)$	$\frac{2}{13}$	$(1,1)$	$\frac{3}{13}$

Ora calcoliamo $I(X;Y)$ utilizzando (1.33).

$$\begin{aligned}
I(X;Y) &= \sum p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} \\
&= p_{X,Y}(0,0) \ln \frac{p_{X,Y}(0,0)}{p_X(0)p_Y(0)} + p_{X,Y}(0,1) \ln \frac{p_{X,Y}(0,1)}{p_X(0)p_Y(1)} \\
&+ p_{X,Y}(1,0) \ln \frac{p_{X,Y}(1,0)}{p_X(1)p_Y(0)} + p_{X,Y}(1,1) \ln \frac{p_{X,Y}(1,1)}{p_X(1)p_Y(1)} \\
&= \frac{4}{13} \ln \frac{\frac{4}{13}}{\frac{6}{13} \cdot \frac{8}{13}} + \frac{2}{13} \ln \frac{\frac{2}{13}}{\frac{6}{13} \cdot \frac{5}{13}} \\
&+ \frac{4}{13} \ln \frac{\frac{4}{13}}{\frac{7}{13} \cdot \frac{8}{13}} + \frac{3}{13} \ln \frac{\frac{3}{13}}{\frac{7}{13} \cdot \frac{5}{13}} \\
&= 0.006 \; nats.
\end{aligned}$$

In base due diventa

$$I(X;Y) = 0.009 \; bits.$$

Il valore ottenuto suggerisce scambio d'informazione tra le variabili X e Y.

Dunque ipotizziamo che ci sia un errore legato alla finitezza dei dati in esame che altera il risultato di $I(X;Y)$.

Vediamo quali sono i risultati di transfer entropy.
Le calcoleremo in due modi diversi:

1. considerando solo lo stato di X e Y appena precedente ad ogni valore di X, usando $m = l = 1$;
2. considerando $m = 6$ valori precedenti di X e $l = 5$ valori precedenti, allo stato X_t, di Y.

Ci aspettiamo di trovare in entrambi i casi transfer entropy nulle, per quanto detto a proposito di transfer entropy e variabili s.i., indipendentemente da m e l.

1. **Primo modo** $(m = l = 1)$

   In questo caso per il calcolo di $p_{X,\bar{X},\bar{Y}}(x,\bar{x},\bar{y})$ si considera l'evento $(X,\bar{X},\bar{Y})$ di lunghezza tre: i valori all'istante $t+1$ e t di X, e il valore all'istante t di Y. Se prendiamo la prima cella di X, questa non ha stati precedenti, perciò dovremmo iniziare a contare gli eventi $(X,\bar{X},\bar{Y})$ a partire dalla seconda cella, per il calcolo delle frequenze, perciò il numero di eventi possibili è $n-1 = 13-1 = 12$.

$p_{X,\bar{X},\bar{Y}}$			
$(0,0,0)$	0	$(1,0,0)$	$\frac{4}{12}$
$(0,0,1)$	$\frac{2}{12}$	$(1,0,1)$	0
$(0,1,0)$	$\frac{2}{12}$	$(1,1,0)$	$\frac{1}{12}$
$(0,1,1)$	$\frac{2}{12}$	$(1,1,1)$	$\frac{1}{12}$

   Le sequenze $(X,\bar{X})$ hanno lunghezza $m+1$ ciascuna. In questo caso avremo due valori: l' istante a $t+1$ e a t. Per lo stesso motivo di prima, il numero di eventi possibili di questo tipo è $n-1 = 12$. Le sequenze $(\bar{X},\bar{Y})$ hanno invece lunghezza pari al massimo tra m e l. In questo caso, $m = l = 1$,$(\bar{X},\bar{Y})$ è formato dal valore di X e Y allo stesso istante. Inoltre avremo $n = 13$ possibili coppie e non $n-1 = 12$.

	$p_{X,\bar{X}}$	$p_{\bar{X},\bar{Y}}$
$(0,0)$	$\frac{2}{12}$	$\frac{4}{13}$
$(0,1)$	$\frac{4}{12}$	$\frac{2}{13}$
$(1,0)$	$\frac{4}{12}$	$\frac{4}{13}$
$(1,1)$	$\frac{2}{12}$	$\frac{3}{13}$

Infine le $p_{\bar{X}}(\bar{x})$, che in questo caso coincidono con le frequenze relative di X poiché consideriamo solo un valore di storia.

$$p_X(0) = \tfrac{8}{13} \quad p_X(1) = \tfrac{5}{13}$$

Utilizzando (2.3) si calcola

$$T_{Y\to X}(1,1) = 0.467\ bits.$$

Ora facciamo il calcolo per $T_{X\to Y}(1,1)$

$p_{Y,\bar{Y},\bar{X}}$			
$(0,0,0)$	$\frac{3}{12}$	$(1,0,0)$	$\frac{1}{12}$
$(0,0,1)$	$\frac{1}{12}$	$(1,0,1)$	$\frac{2}{12}$
$(0,1,0)$	$\frac{2}{12}$	$(1,1,0)$	0
$(0,1,1)$	$\frac{1}{12}$	$(1,1,1)$	$\frac{2}{12}$

	$p_{Y,\bar{Y}}$	$p_{\bar{Y},\bar{X}}$
$(0,0)$	$\frac{4}{12}$	$\frac{4}{13}$
$(0,1)$	$\frac{3}{12}$	$\frac{2}{13}$
$(1,0)$	$\frac{3}{12}$	$\frac{4}{13}$
$(1,1)$	$\frac{3}{12}$	$\frac{3}{13}$

$$p_Y(0) = \tfrac{6}{13} \quad p_Y(1) = \tfrac{7}{13}$$

Segue

$$T_{X\to Y}(1,1) = 0.244\ bits.$$

I risultati ottenuti portano a pensare che lo scambio di informazione tra X e $\bar{Y}$ noto $\bar{X}$, sia maggiore di quello tra Y e $\bar{X}$ noto $\bar{Y}$.

Dovremmo supporre che ci sia un errore sistematico che altera i risultati, o che i dati in esame sono pochi.

2. **Secondo modo** ($m = 6$, $l = 5$)

   Il procedimento è lo stesso di prima, ma cambia la lunghezza degli esiti considerati. Ad esempio $(X, \bar{X}, \bar{Y})$ ha lunghezza dodici, $m+l+1 = 12$, perciò le frequenze di questi eventi vengono contate dalla cella 7 di X, perché le celle precedenti non hanno almeno 6 valori di storia.

   Quindi il numero di volte in cui l'evento $(X, \bar{X}, \bar{Y})$ si può ripetere è $n - m$. Notiamo che se fosse stato $n > l$, il numero di eventi possibili sarebbe stato $n - l$. Per il calcolo delle altre frequenze il ragionamento è lo stesso. Negli eventi in cui non compare lo stato a tempo t avremo $n - m + 1$ (o o $n - l + 1$ ripetizioni possibili se $l > m$).

Programma C++ per la transfer entropy

Aumentando i valori d m e l e il numero di dati, i calcoli diventano molto più laboriosi. Per questo ho creato un programma in C++ che calcola informazione mutua e transfer entropy tra due variabili, in cui si possono scegliere il numero di dati da inserire (i lanci per ogni moneta), e i parametri m e l. Il programma restituisce medie aritmetiche delle due variabili, covarianza teorica e campionaria, informazione mutua, transfer entropy e effective transfer entropy (che sarà introdotta più avanti), nei sensi da Y a X e viceversa.

Il programma sfrutta delle matrici le cui colonne sono gli $(x, \bar{x}, \bar{y})$. Attraverso un "match" tra le colonne (le parole in codice), restituisce le probabilità di ciascuna. Queste vengono utilizzate per i calcoli delle funzioni sopra.

I risultati sono

$$T_{Y\to X}(6,5) = 0.100\ bits$$
$$T_{X\to Y}(6,5) = 0.400\ bits$$

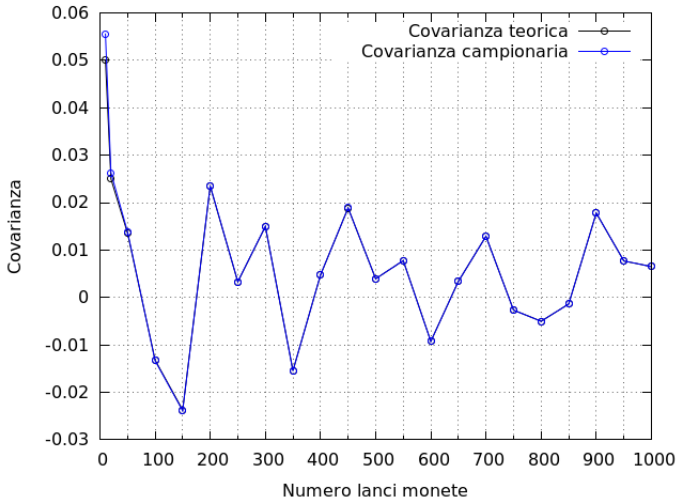
e si ripresenta il problema del caso $m = l = 1$.

Dato che i risultati ottenuti finora sono insoddisfacenti, aumentiamo il numero di dati in esame, considerando molti più lanci di monete.

2.1.1 Andamento della transfer entropy al variare del numero di lanci

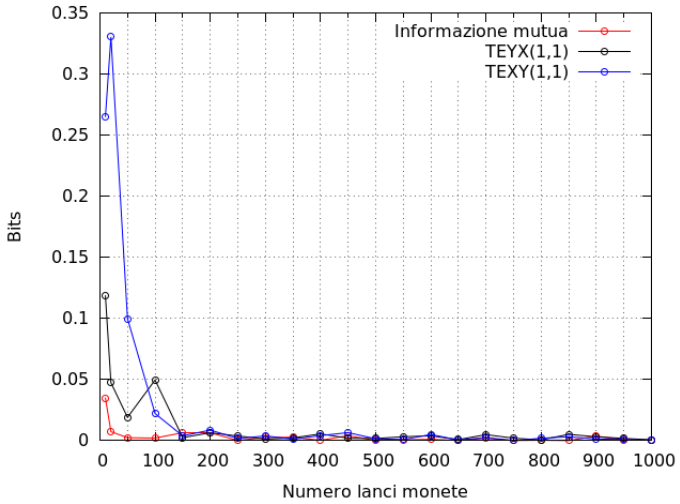
Consideriamo nuovamente il caso $m = l = 1$, e ricalcoliamo covarianza, informazione mutua e transfer entropy con un numero lanci che varia da zero a mille.

Figura 2.1: *Grafico covarianza teorica e campionaria*



Le covarianze sono vicine allo zero, e sono quasi sovrapposte nel grafico. Inoltre all' aumentare dei dati si nota che fluttuano sempre di meno attorno allo zero.

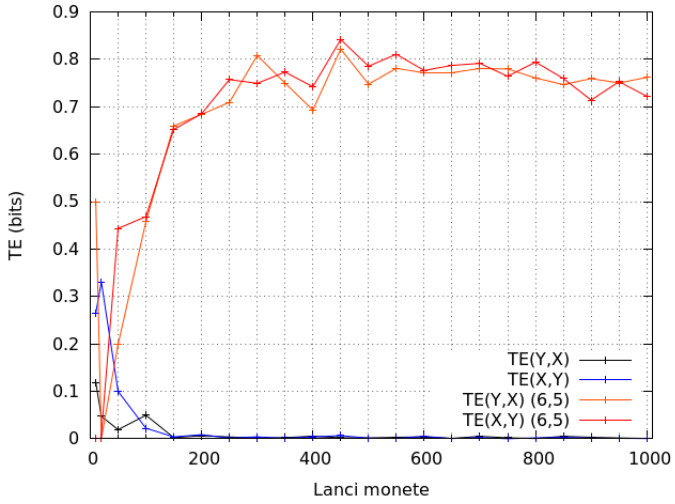
Figura 2.2: *Grafico transfer entropy* $(1,1)$ *e informazione mutua*



Transfer entropy e informazione mutua vanno a zero assieme per $N \geq 200$.

Ora consideriamo un altro valore dei parametri, ad esempio $m = 6$, $l = 5$, e per i calcoli utilizziamo lo stesso set di dati. Ci aspettiamo di ottenere un risultato simile a $m = l = 1$, perché le variabili sono s.i., e il risultato non dovrebbe dipendere dalla storia considerata.

Figura 2.3: *Grafico transfer entropy* $(1,1)$ *e transfer entropy* $(6,5)$



Invece di andare a zero, le transfer entropy tendono a un valore di circa 0.7, in disaccordo con la teoria, il che porta a pensare:

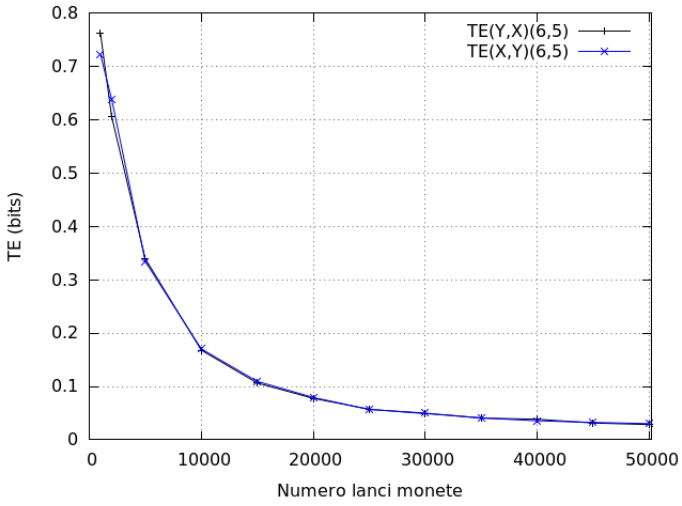
1. che i dati per $m = 6$, $l = 5$ siano ancora insufficienti per valutare se la transfer entropy è nulla;
2. che il calcolo con $m = l = 1$ non considera una storia sufficiente a discriminare l' indipendenza delle variabili;
3. che ci sia un errore sistematico nel calcolo con $m = 6$, $l = 5$, che devia i risultati di $T_{Y\to X}$ e $T_{X\to Y}$ dal loro valore vero.

Tra le proposte possiamo già pensare di scartare la 2., perché sappiamo che i dati sono stati generati da una distribuzione uniforme. La proposta 3. sembra essere probabile, perché $T_{Y\to X}$ e $T_{X\to Y}$ sono molto vicine e sembrano essere affette dallo stesso errore sistematico.

Per quanto riguarda la proposta 1., date due variabili aleatorie X e Y qualsiasi, più i parametri m e l sono grandi, più combinazioni sono possibili per gli elementi con $\bar{x}$ e $\bar{y}$, perciò servono molti più dati affinché le frequenze relative si avvicinino alle distribuzioni di probabilità. Quindi per ci aspettiamo che per ottenere valori della transfer entropy a $m = 6$ e $l = 5$ simili a quella con $m = l = 1$, servano molti più dati.

Verifichiamo graficamente quanto detto utilizzando set di dati che vanno da 10 a 50000.

Figura 2.4: *Grafico* $T_{Y\to X}(6,5)$ *e* $T_{X\to Y}(6,5)$



Dal grafico si nota che le transfer entropy vanno a zero e allo stesso modo, perciò se c' è un errore sistematico , dev'essere diminuito all'aumentare del numero di lanci.

Facciamo un'ulteriore verifica considerando il valore fisso $m = 1$ e aumentando l da uno fino a sei (e viceversa). Il numero di lanci andrà da dieci a mille.

Figura 2.5: *Grafico* $T_{Y \to X}(m, l)$

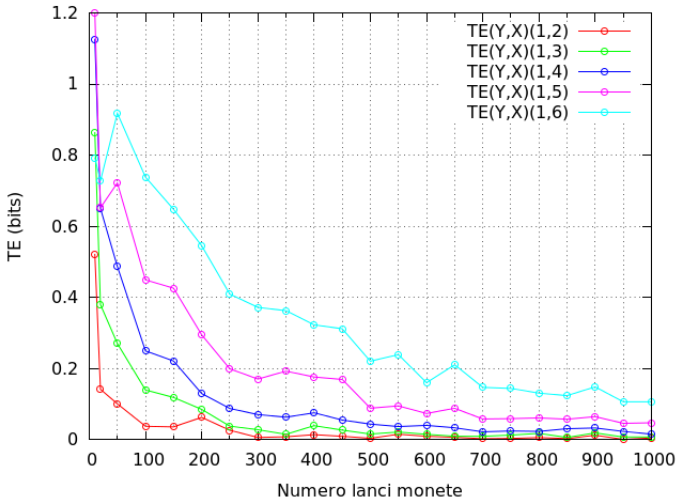
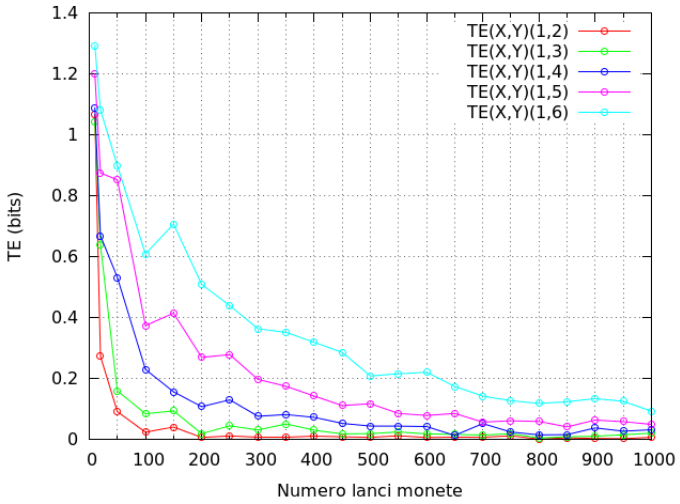


Figura 2.6: *Grafico* $T_{X \to Y}(m, l)$



Si nota che, a parità di N numero di dati, le transfer entropy con valore di l più alto sono più distanti dallo zero, e quindi presentano un errore maggiore.

2.1.2 Effective Transfer Entropy

Esiste una modifica dell'originale transfer entropy, detta effective transfer entropy, che permette di eliminarle gli errori sistematici, almeno in parte. In generale, per calcolare le vere transfer entropy tra due variabili, dovremmo considerare l'intera storia di un processo, e quindi mandare i parametri m e l all' infinito.

L'effective transfer entropy è stata introdotta da R. Marschinski e H. Kantz (*Eur. Phys. J.,2002*) ed è definita come

$$T_{Y\to X}(m,l)_{eff} = T_{Y\to X}(m,l) - T_{Y_{shuffled}\to X}(m,l) \tag{2.21}$$

$$= \sum p(x,\bar{x},\bar{y}) \ln \frac{p(x \mid \bar{x},\bar{y})}{p(x \mid \bar{x})} \tag{2.22}$$

$$= -\sum p(x,\bar{x},y_{\bar{shuff}}) \ln \frac{p(x \mid \bar{x},y_{\bar{shuff}})}{p(x \mid \bar{x})} \tag{2.23}$$

Essa è pari alla differenza di due transfer entropy:

1. la prima è quella classica, calcolata con i dati di X e Y ordinati temporalmente;

2. la seconda è calcolata rimescolando in modo casuale solo i dati della variabile di cui si vuole calcolare l'influenza, in questo caso la Y.

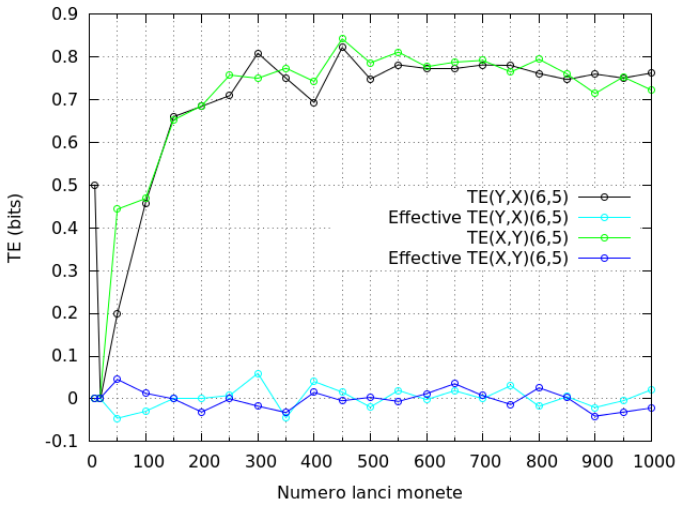
Si nota che rimescolando la Y, viene eliminata la relazione di causa-effetto tra X e Y.

Possiamo pensare quindi che $T_{Y\to X}(m,l)$ contenga un errore residuo, presente anche in quella rimescolata. Sottraendo le due si elimina questo tipo di errore. In particolare se indichiamo con $T_{Y\to X}(m,l)_{res}$ l'errore residuo sulla transfer entropy, deve valere

$$T_{Y\to X}(m,l)_{res} = T_{Y_{shuffled}\to X}(m,l) \tag{2.24}$$

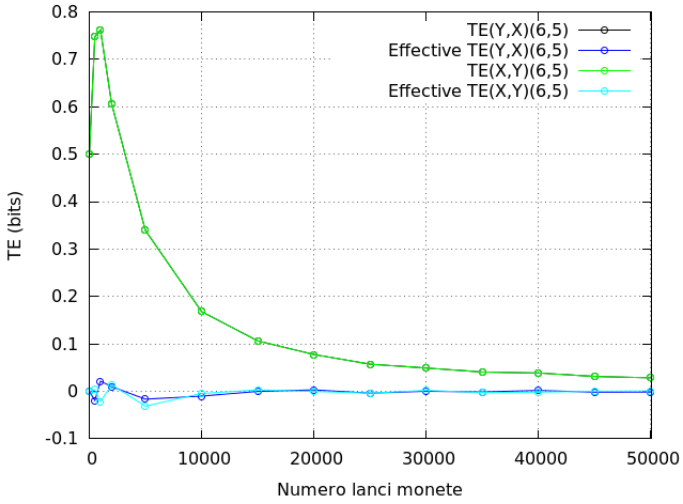
Dal grafico (2.5) si nota che la transfer entropy, calcolata con $m = 6$ e $l = 5$, non tende a zero, ma ad un valore maggiore. Dunque ipotizziamo che l'errore sistematico sia diminuito all'aumentare di N, ma non sia stato eliminato. Per questo caso facciamo un confronto tra transfer entropy e effective transfer entropy, e vediamo se in quest'ultima l'errore sistematico sparisce.

Figura 2.7: *Grafico* $T_{Y\to X}(6,5)$, $T_{X\to Y}(6,5)$ *classiche ed efficaci*



Vediamo qual è l'effetto su larga scala

Figura 2.8: *Grafico* $T_{Y\to X}(6,5)$, $T_{X\to Y}(6,5)$ *classiche ed efficaci*



L'effective transfer entropy è nulla, e, come si vede dal grafico, tra le due c'è una differenza, che supponiamo sia la transfer entropy residua. Sembra che nell'effective transfer entropy siamo riusciti a togliere l'errore sistematico.

Possiamo concludere dicendo che usando l' effective transfer entropy e N sufficientemente elevato, i dati sperimentali sono coerenti con il modello X,Y s.i.. In particolare, a parità di N, l'effective transfer entropy è più vicina allo zero della transfer entropy, e stima meglio il valore vero.

2.2 Test con variabili dipendenti

In questo test vogliamo mostrare un esempio in cui $T_{Y\to X}(m,l) \neq T_{X\to Y}(m,l)$.

Definiamo Y_t con esiti equiprobabili

$$\boxed{p_Y(0) = p_Y(1) = \tfrac{1}{2}}$$

e X tale che $X_t = Y_{t-6}$. Allora per $t \neq t'$ valgono

$$Y_t \neq Y_{t'} \quad X_t \neq X_{t'}$$

Dunque per $X_t \neq Y_{t-6}$ vale

$$\boxed{\begin{aligned} &p_X(0) = p_X(1) = \tfrac{1}{2} \\ &p_Y(0) = p_Y(1) = \tfrac{1}{2} \\ &p_{X,Y}(0,0) = p_{X,Y}(0,1) = p_{X,Y}(1,0) = p_{X,Y}(1,1) = \tfrac{1}{4} \end{aligned}}$$

Segue da (1.34) e (1.17) che

$$\begin{aligned} I(X;Y) &= 0 \\ Cov(X,Y) &= 0 \end{aligned}$$

Per le transfer entropy, possiamo fare riferimento al caso 4. a pag. 20. Infatti questo test è lo stesso esempio, con $X_t = g(Y_{t-6}) = Y_{t-6}$, cioè $k = 6$. Allora si hanno i due casi

1. $l < 6$,

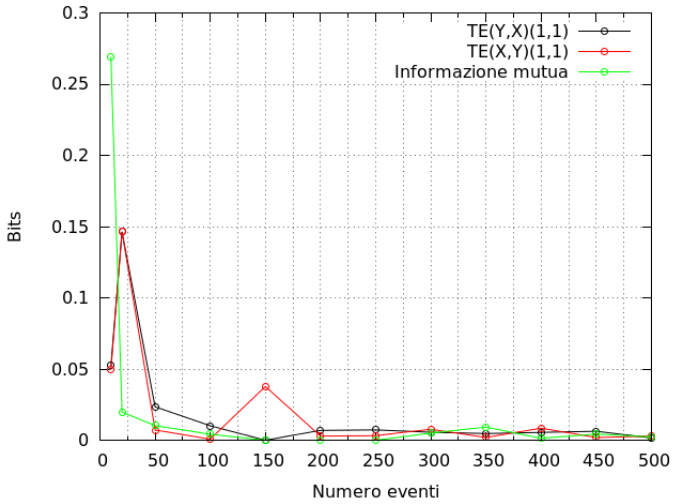
$$\begin{aligned} T_{Y\to X}(m,l) &= 0 \\ T_{Y\to X}(m,l) &= 0 \end{aligned}$$

2. $l \geq 6$,

$$\begin{aligned} T_{Y\to X}(m,l) &= H(Y) \\ &= \log_2 2 = 1\ bit \end{aligned}$$

Analizziamo i dati generati dal programma C++:

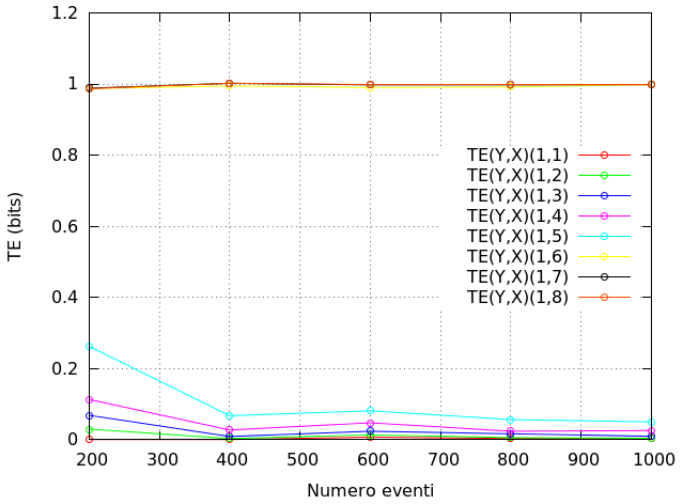
Figura 2.9: *Grafico informazione mutua e transfer entropy*



Ora facciamo variare il parametro l da uno a otto e vediamo cosa accade.

Vediamo prima l' andamento delle $T_{Y\to X}(m,l)$:

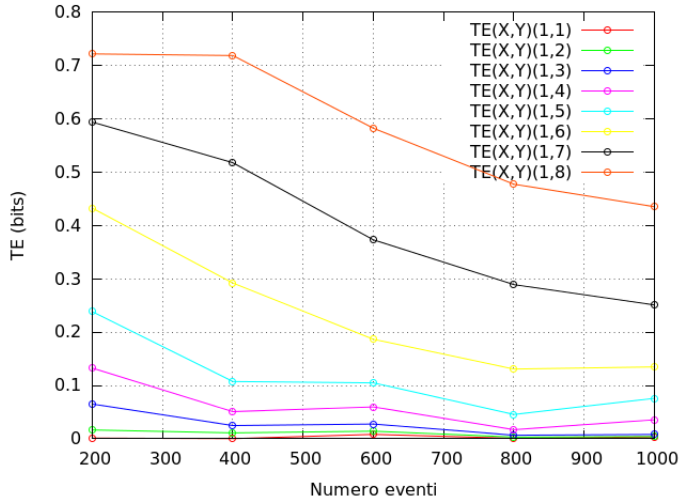
Figura 2.10: *Grafico transfer entropy* $Y \to X$



Per $l \geq 6$ le transfer entropy si assestano a un valore di circa un bit, dunque sembrano confermare che $\bar{Y}$ non è indipendente da X, condizionatamente a $\bar{X}$, a distanza di più di sei unità temporali.

Ora vediamo le $T_{X\to Y}(m,l)$:

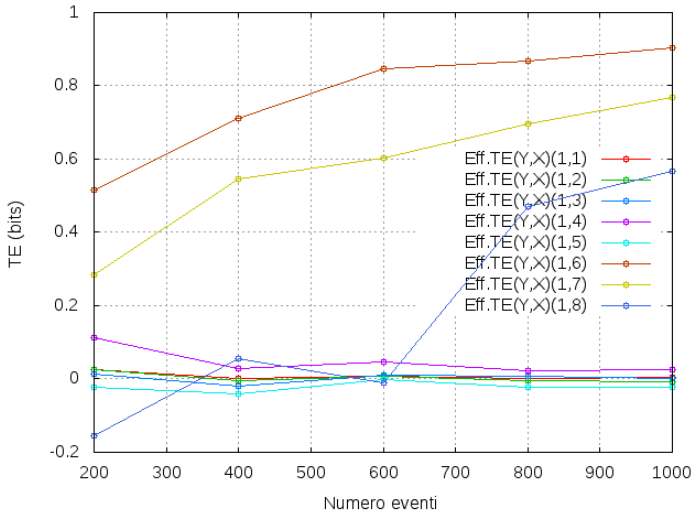
Figura 2.11: *Grafico transfer entropy* $X \to Y$



In questo caso le transfer entropy tandono allo zero, ma più aumenta l, più il loro valore è distante da esso. Notiamo lo stesso andamento nelle figure (2.6) e (2.7) del test per variabili s.i.. Dato che in quel caso la tranfser entropy aveva valore vero nullo, probabilmente significa che anche qui si stanno avvicinando allo zero.

Vediamo le effective transfer entropy.

Figura 2.12: *Grafico* $T_{Y\to X}(m,l)_{eff}$



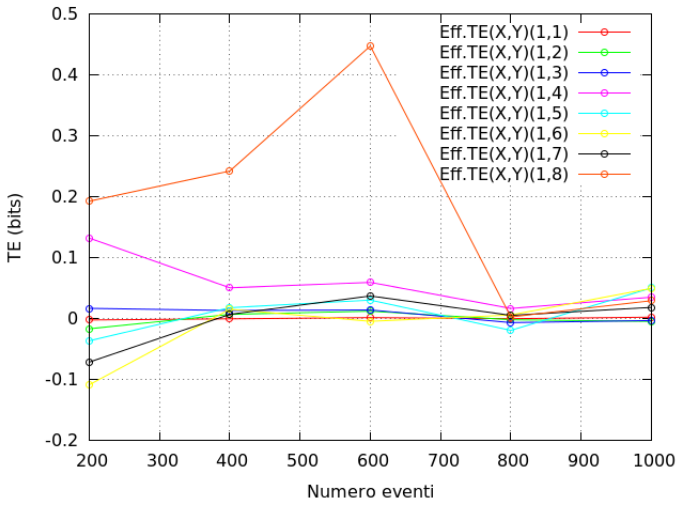
Confrontando con la figura (2.10) si nota che fissati N, m e l, si ha $T_{Y\to X}(m,l)_{eff} < T_{Y\to X}(m,l)$. Inoltre:

1. per $l < 6$ le effective transfer entropy tendono allo zero all'aumentare di N, ma più è alto il valore di l, più il loro valore fluttua e l'andamento a zero è ritardato;

2. per $l \geq 6$ le effective transfer entropy aumentano all' aumentare di N, ma più è alto il valore di l, più l'andamento crescente viene ritardato. Ad esempio notiamo che $T_{Y\to X}(1,7) < T_{Y\to X}(1,6)$ a parità di N. Abbiamo comunque verificato che anche le effective transfer entropy raggiungono il valore di un bit, ma per un numero di dati molto maggiore, che aumenta all' aumentare di l.

   Possiamo concludere che per $l \geq 6$ X non è indipendente da $\bar{Y}$ condizionatamente a $\bar{X}$.

Figura 2.13: *Grafico transfer entropy $X \to Y$ efficaci*



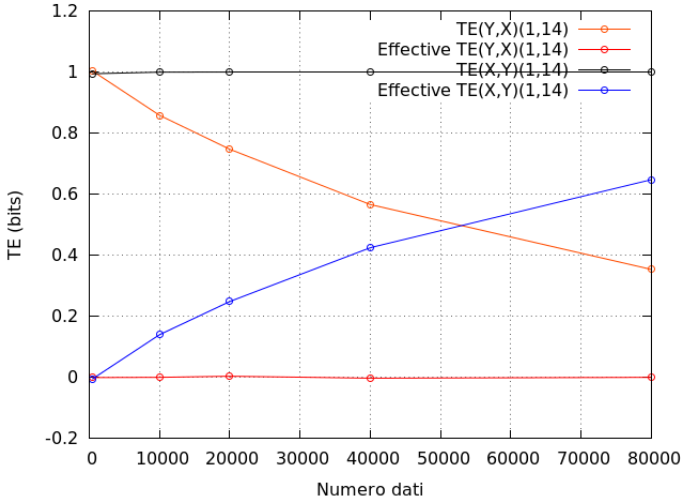
Si verifica che le effective transfer entropy $T_{X\to Y}(m,l)_{eff}$ sono molto più vicine allo zero rispetto al grafico (2.11), e ancora una volta, fissato N e aumentando l, il loro valore fluttua maggiormente. Possiamo concludere che X è indipendente da $\bar{Y}$ condizionatamente $\bar{X}$.

Se notiamo il grafico (2.12), per $l \geq 6$, sembra accadere il seguente fenomeno: quando X è indipendente da $\bar{Y}$ condizionatamente a $\bar{X}$, l'effective transfer entropy tende asintoticamente alla transfer entropy, in questo caso al valore di un bit..

Ripetiamo lo stesso test, ma con X e Y dipendenti secondo la relazione $X_t = Y_{t-14}$, per verificare se accade nuovamente lo stesso comportamento. Sappiamo già che per $l \geq 14$ la transfer entropy $T_{Y\to X}(m,l)$ ha valore vero di un bit.

Grafichiamo le transfer entropy e le effective transfer entropy con $m = 1$ e $l = 14$

Figura 2.14: *Grafico transfer entropy per grandi N*



Il grafico evidenzia il comportamento asintotico dell'effective transfer entropy, al valore di un bit.

Notiamo che fissato N numero di dati, quando $T_{X\to Y}(1,14)$ decresce, $T_{Y\to X}(1,14)_{eff}$ aumenta di circa la stessa quantità. Spieghiamo il perché.

Dalla definizione di effective transfer entropy (che riportiamo per comodità)

$$T_{Y\to X}(m,l)_{eff} = T_{Y\to X}(m,l) - T_{Y_{shuffled}\to X}(m,l)$$

si nota che se inseriamo il valore ottenuto nel grafico (2.14), $T_{Y\to X}(m,l) \approx 1$ *bit*, per ogni N, allora l' andamento asintotico è dovuto al solo termine di rimescolamento.

Ora notiamo che anche $T_{X\to Y}(m,l)$ ha comportamento asintotico allo zero (il suo valore vero), poiché Y non dipende da $\bar{X}$. In più possiamo pensare che la transfer entropy da X in Y si comporti come le $T_{X\to Y}(m,l)$ del test nella sezione 2.1, asintotiche allo zero. Infatti dato che in questo test, e nel test in 2.1, le variabili hanno gli stessi esiti (0 e 1), allora $T_{X\to Y}(m,l)$ deve avere andamento simile alle transfer entropy delle monete, tendente allo zero, poichè non c' è rapporto causa-effetto tra Y e $\bar{X}$. Dato che per le monete se scambiamo X con Y otteniamo un risultato simile di transfer entropy, cioè vale

$$T_{Y\to X}(m,l) \approx T_{X\to Y}(m,l)$$

possiamo fare ancora questa approssimazione, ma sostituendo il primo termine con $T_{Y_{shuffled}\to X}(m,l)$, in cui viene perso anche il rapporto causa-effetto tra X e $\bar{Y}$. Perciò otteniamo

$$T_{Y_{shuffled}\to X}(m,l) \approx T_{X\to Y}(m,l)$$

e sostituendo nell'effective transfer entropy si ha

$$T_{Y\to X}(1,14)_{eff} \approx T_{Y\to X}(1,14) - T_{X\to Y}(1,14) \tag{2.25}$$
$$\approx 1 - T_{X\to Y}(1,14) \tag{2.26}$$

In questo modo abbiamo spiegato il comportamento asintotico di $T_{Y\to X}(1,14)_{eff}$, dovuto al fatto che $T_{X\to Y}(1,14)$ è asintotica.

Dall'analisi di questo caso particolare possiamo dire che in generale $T_{Y\to X}(m,l)_{eff}$ non stima meglio il valore vero di transfer entropy rispetto a $T_{Y\to X}(m,l)$. Invece $T_{Y\to X}(m,l)_{eff}$ è più efficace per identificare variabili il cui valore vero di transfer entropy è nullo. Infatti in questo caso l'effective transfer entropy elimina l'errore di fondo dovuto alla finitezza dei dati, che allontana la transfer entropy dal valore zero.

Capitolo 3

Applicazione della transfer entropy: analisi della presenza/assenza di specie di piante

In questo capitolo vogliamo mostrare un'applicazione pratica della transfer entropy.

Abbiamo a disposizione un set di dati proveniente da una foresta di $1000x500m^2$. La foresta è stata suddivisa in 5000 celle, che vengono numerate da sinistra verso destra partendo dal basso. Abbiamo a disposizione i dati di 320 specie di piante, e ognuno rappresenta la presenza (indicata con 1) o l'assenza (indicata con 0) della specie nella determinata cella.

Vogliamo verificare se la presenza/assenza di una specie, in una cella di foresta, è dovuta più alla presenza/assenza di un' altra specie (rispetto alla presenza di se stessa), nella cella precedente.

Come nella sezione 2.1 calcoleremo le p_X, p_Y, $p_{X,Y}$, $p_{\bar{X},\bar{Y}}$,ecc. per ogni coppia di specie. La presenza/assenza in una cella è analoga al risultato del lancio di una monete (0 per assenza, 1 per presenza).

Dunque, ad esempio:

1. $p_{X,Y}(0,0)$ = probabilità che le specie X e Y non sono presenti nella stessa cella;
2. $p_{X,Y}(1,0)$ = probabilità che la specie X è presente e Y non è presente nella stessa cella;
3. $p_{X,Y}(0,1)$ = probabilità che la specie X non è presente e Y è presente nella stessa cella;
4. $p_{X,Y}(1,1)$ = probabilità che le specie X e Y sono presenti nella stessa cella;

Per fare un confronto con i risultati che otterremo, abbiamo calcolato l'errore standard della popolazione per informazione mutua e le transfer entropy da dieci set di dati per X e Y s.i., cioè come due monete. Dato che per le transfer entropy abbiamo i due risultati $T_{Y\to X}(m,l)$ e $T_{X\to Y}(m,l)$, li abbiamo trattato come se fossero i risultati di un' unica transfer entropy calcolata da 20 set di dati diversi, poiché X e Y sono definite allo stesso modo e i risultati sono simili scambiandole. Per l'effective transfer entropy abbiamo fatto lo stesso. Considereremo gli errori centrati attorno al valore vero, zero, per tutte e tre le funzioni. Abbiamo ottenuto:

1. $\sigma = 0.000136$ bits per $I(X;Y)$;
2. $\sigma = 0.000125$ bits per le transfer entropy, classiche ed efficaci;
3. $\sigma = 0.000246$ bits per l' effective transfer entropy.

Dato che l'effective transfer entropy dovrebbe essere minore della transfer entropy, nel caso di variabili s.i., e il σ della transfer entropy efficace è maggiore di quello della transfer entropy, utilizziamo per entrambe il maggiore dei due. Definiamo i seguenti range, attorno allo zero:

1. $[0, 0.000136]$ bits per $I(X;Y)$;
2. $[0, 0.000246]$ bits per le transfer entropy, classiche ed efficaci;

Se i valori che otterremo ricadranno all'interno del range, potremmo affermare con più certezza se le specie di piante X, Y, si comportano come variabili s.i., o se, ad esempio X non dipende da $\bar{Y}$ condizionatamente a $\bar{X}$

Questi range sono indicativi solo per i risultati con $m = l = 1$ e $N = 5000$.

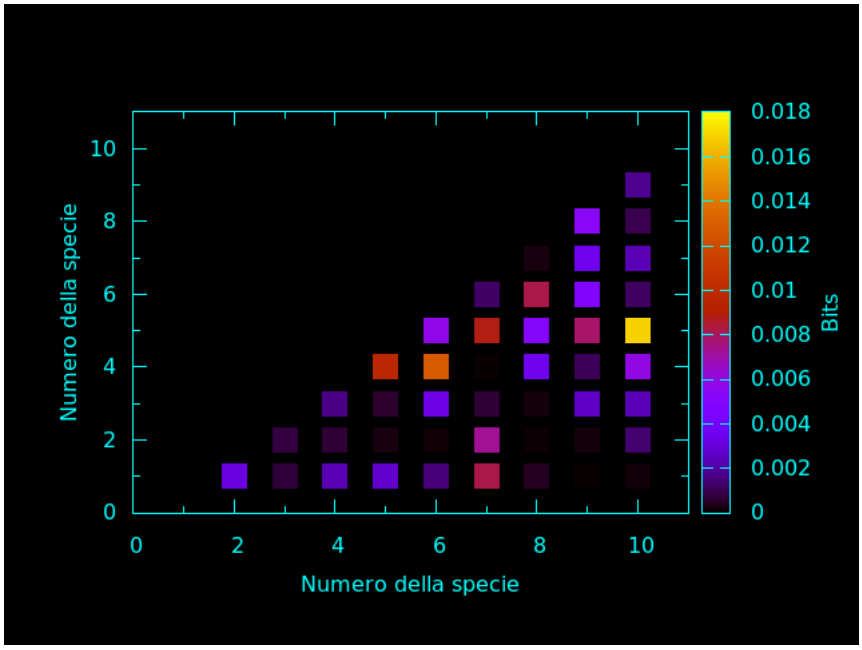
Selezioniamo le dieci specie con numero totale di presenze vicino a 2500.

Specie	Presenze totali		Specie	Presenze totali
40	2461		1	2461
95	1987		2	1987
109	2927		3	2927
129	3041		4	3041
227	2211	$\longrightarrow$	5	2211
229	2101		6	2101
240	2833		7	2833
269	2611		8	2611
277	2211		9	2211
282	2351		10	2351

A sinistra la tabella con la numerazione originale delle specie, a destra la numerazione adottata per facilitare la visione dei grafici. Questa volta presenteremo i grafici con gli assi x e y rappresentano il numero della specie, e il colore del generico punto (x, y) rappresenta il valore in bits, cioè come *heat map*.

Riportiamo il grafico dell'informazione mutua.

Figura 3.1: *Grafico informazione mutua*



L'informazione mutua, per ogni coppia X, Y, non rientra nel range $[0, 0.000136]$ del caso di variabili s.i., e i più alti valori ottenuti si avvicinano a $\approx 100\sigma$. Se però confrontiamo i valori con il valore di un bit della $T_{X\to Y}(1, 6)$ ottenuto nel caso di variabili dipendenti, allora i risultati sono ancora molto bassi, cioè almeno 100 volte più piccoli. Possiamo concludere che le specie sono dipendenti a due a due, ma molto debolmente. Il valore più alto di informazione mutua, in giallo, è tra la specie 5 e 10.

Per quanto riguarda le transfer entropy, le graficheremo in modo che la coppia (x, y) nel grafico rappresenti il valore di $T_{Y\to X}(1, 1)$. Dunque a sinistra della diagonale $y = x$ si troveranno le $T_{Y\to X}(1, 1)$; a destra le $T_{X\to Y}(1, 1)$. Ad esempio l' elemento in $(6, 8)$ rappresenterà $T_{8\to 6}(1, 1)$ e l' elemento in $(8, 6)$ rappresenterà $T_{6\to 8}(1, 1)$. Utilizzeremo lo stesso range, in bits, per transfer entropy e effective transfer entropy, in modo da renderci subito conto della differenza tra le due dalla variazione di colore della cella (x, y).

Vediamo le transfer entropy e effective transfer entropy con $m = l = 1$.

Figura 3.2: *Grafico* $T_{Y\to X}(1,1)$ *e* $T_{X\to Y}(1,1)$

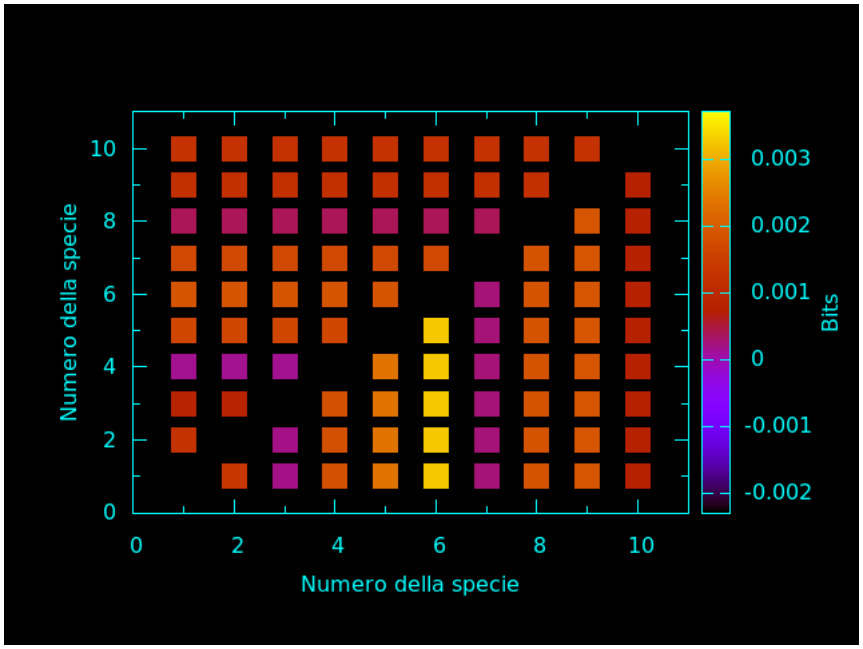
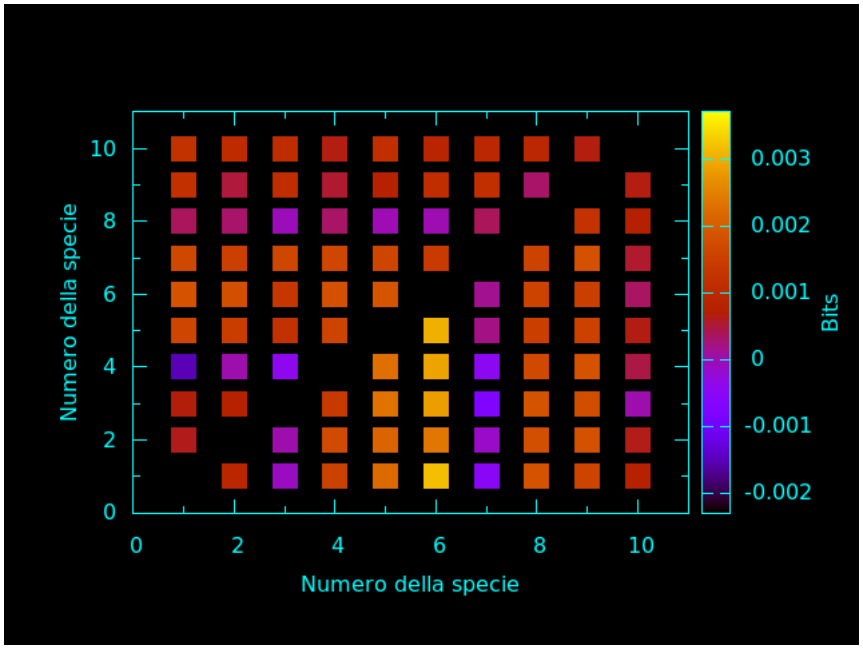


Figura 3.3: *Grafico* $T_{Y\to X}(1,1)_{eff}$ *e* $T_{X\to Y}(1,1)_{eff}$



Transfer entropy ed effective transfer entropy presentano valori molto simili tra loro e vicini allo zero. Notiamo che anche le transfer entropy non rientrano nel range definito, e sono maggiori di un ordine di grandezza rispetto al 0.000246 bits (10 volte maggiori). Dunque possiamo dire che probabilmente la presenza/assenza di una specie (X) in una cella è determinata, molto debolmente, dalla presenza di un'altra specie ($\bar{Y}$) nella cella precedente , più che da sé stessa ($\bar{X}$). In particolare la specie 6 sembra dipendere più dalle specie $1, 2, 3, 4, 5$ rispetto a sé stessa. Viceversa, la specie 6 influenza più debolmente le altre piante, rispetto a sé stesse.

L'analisi presentata in questo capitolo è preliminare, e (sia per i risultati ottenuti, che per il metodo di selezione scelto) indica che probabilmente i valori più alti di transfer entropy sono dovuti alla maggiore dipendenza di una specie dall'altra (condizionatamente), che potrà essere confermata o meno con analisi più dettagliate. Ad esempio l'influenza tra le specie potrebbe essere determinata in modo più forte da

altri caratteri, che non riguardano la sola presenza/assenza delle specie, ma ad esempio la presenza di fiori, frutti,ecc., quindi attraverso una selezione più specifica delle variabili. Dunque, in questo caso, i risultati della transfer entropy permettono di selezionare i processi stocastici che presentano un maggiore rapporto causa-effetto, e fornisce una base dalla quale è possibile proseguire con un'analisi più dettagliata.

Appendice A

Entropia in teoria dell'informazione

In teoria dell'informazione si estende il concetto di variabile aleatoria a una funzione $X\colon \Omega \to S = \{x_1, \ldots, x_n\}$ dove S è un insieme generico e non un sottoinsieme di $\mathbb{R}$, e viene detto range di X.

Si definisce *source code* o *codice sorgente* di una variabile aleatoria X la mappa $C : S \to D^*$.

D^* è un insieme di stringhe di lunghezza finita, formato da elementi di un insieme finito D di simboli, detto D-alfabeto. Di seguito supporremo sempre $D = \{0, 1, \ldots, D-1\}$ composto da un numero D di simboli. $C(x)$ rappresenta la *parola in codice* di un elemento $x \in S$ e $l(x)$ rappresenta la lunghezza di $C(x)$.

Una *sequenza* è una stringa di parole in codice, separate l'un l' altra da una virgola. Assumeremo che ad ogni valore di X venga associato un unica parola in codice. Ciò equivale a dire che la mappa C è iniettiva, cioè se $x \neq x'$, allora $C(x) \neq C(x')$.

Un codice è *istantaneo* se nessuna parola in codice è il prefisso di un' altra. In pratica è possibile decifrare la parola in codice istantaneamente, appena arriva, perché non dipende da altre parole in codice.

Dato che nel linguaggio informatico sono note le parole in codice e non i valori associati tramite la mappa C, è possibile conoscere solo la frequenza con cui si ripete ciascuna parola in codice.

Grazie alla iniettività di C, è possibile definire il concetto di *probabilità di una parola in codice* come la probabilità dell'evento associato. In particolare tale probabilità coinciderà con la frequenza della parola in codice.

Ad esempio $C(\text{rosso}) = 00$ e $C(\text{blu}) = 11$ sono le parole in codice per $\S = \{\text{rosso, blu}\}$, con alfabeto $D = \{0, 1\}$ e le due stringhe, 00 e 11,hanno entrambe di lunghezza due.

Si definisce *lunghezza media* di un codice $C(x)$,

$$L(C) = \sum_{x \in X} p(x) l(x). \tag{A.1}$$

per una variabile aleatoria X, con distribuzione di probabilità $p(X)$.

Consideriamo una variabile aleatoria X con seguente distribuzione di probabilità, codice e lunghezza

$p_X(1) = \frac{1}{2}$	$C(1) = 0$	$l(1) = 1$
$p_X(2) = \frac{1}{4}$	$C(2) = 10$	$l(2) = 2$
$p_X(3) = \frac{1}{8}$	$C(3) = 110$	$l(3) = 3$
$p_X(4) = \frac{1}{8}$	$C(4) = 111$	$l(4) = 3$

La lunghezza media del codice è

$$\begin{aligned} L(C) &= \sum_{x \in X} p(x) l(x) \\ &= \frac{1}{2} \cdot 1bit + \frac{1}{4} \cdot 2bits + 2 \cdot \frac{1}{8} \cdot 3bits \\ &= 1.75bits \end{aligned}$$

L' entropia di X è

$$\begin{aligned} H(X) &= -\sum_{x\in X} p(x)\log_2 p(x) \\ &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - 2\frac{1}{8}\log_2\frac{1}{8} \\ &= 1.75 bits \end{aligned}$$

ed ha lo stesso valore, ma ciò non è sempre vero.

Ad esempio per la seguente distribuzione e codice

$p_X(1) = \frac{1}{3}$	$C(1) = 0$	$l(1) = 1$
$p_X(2) = \frac{1}{3}$	$C(2) = 10$	$l(1) = 2$
$p_X(3) = \frac{1}{3}$	$C(3) = 110$	$l(3) = 3$

si hanno

$$\begin{aligned} L(C) &= 1.66 bits \\ H(X) &= 1.58 bits \leq L(C) \end{aligned}$$

Dimostriamo che l'entropia è la lunghezza minima di una parola in codice che permette di esprimere una variabile. Per farlo useremo il

1.5 Teorema (Kraft inequality) *Sia D alfabeto finito di cardinalità D, e C codice istantaneo per la variabile X. Allora*

$$\sum_{i=1}^{\infty} D^{-l_i} \leq 1,$$

dove $l_i = l(x_i)$ è la lunghezza della parola in codice i-esima

Dim: Sia $y_1 \dots y_{l_i}$ l' i-esimo codice. Allora se D è la cardinalità del D-alfabeto, e il numero di elementi è dieci,

$$0.y_1 \dots y_{l_i} = \sum_{j=1}^{l_i} D^{-j} y_j$$

è il valore numerico decimale del codice.

Questa codifica corrisponde all'intervallo

$$(0.y_1 \dots y_{l_i}, 0.y_1 \dots y_{l_i} + \frac{1}{D^{l_i}})$$

Dato che questo è un sotto-intervallo di $[0, 1]$ e l'intervallo è minore uguale a uno, ciò prova che la somma

$$\sum_{i=1}^{N} D^{-l_i} l e 1.$$

1.6 Teorema: *La lunghezza L di un codice istantaneo di alfabeto D per una variabile aleatoria X , è maggiore uguale all' entropia di X,*

$$L \geq H_D(X), \tag{A.2}$$

con uguaglianza verificata per $D^{-l_i} = p_i$.

Dim.:

$$\begin{aligned} L - H_D(X) &= \sum p_i l_i + \sum p_i \log_D p_i \\ &= -\sum p_i \log_D D^{-l_i} + \sum p_i \log_D p_i \\ &= +\sum p_i \log_D \frac{p_i \sum D^{-l_j}}{D^{-l_i} \sum D^{-l_j}}. \end{aligned} \tag{A.3}$$

Scrivendo

$$r = \frac{D^{-l_i}}{\sum D^{-l_j}}, c = \sum D^{-l_j}$$

si ha

$$\begin{aligned} L - H_D(X) &= + \sum p_i \log_D \frac{p}{r} - + \sum p_i \log_D \sum D^{-l_j} \\ &= D(p \parallel r) + \log_D \frac{1}{c} \geq 0, \end{aligned} \tag{A.4}$$

dato che l'entropia relativa è non negativa e $c \leq 1$ (Kraft inequality).

Bibliografia

[1] Cardin F., Favretti M., *Modelli fisico matematici*, Padova, Cleup, 2013

[2] Thomas M. Cover, Joy A. Thomas,*Elements of Information Theory*, New York, John Wiley & Sons, 1991

[3] P. Jizba, H. Kleinert, M. Shefaat, *Rényi' s information transfer between financial time series*, Physica A 391, 2012

[4] R. Marschinski, H. Kantz, *Analysing the information flow between financial time series*, Eur. Phys. J. B 30, 275-281, 2002

[5] C. E. Shannon, W. Weaver, *The Mathematical Theory of Communication*, New York, 1949

[6] T. Schreiber, *Measuring Information Transfer*, Phys. Rev. Lett. 85, 461, 2000

[7] M. Loreti, *Teoria degli errori e fondamenti di statistica*, 2006

[8] R. J. Barlow, *Statistics: a guide to the use of statistical methods in the physical sciences*, J. Wiley & Sons, 1997

M-x reftex-find-duplicate-labels