Sistemi complessi

Lorenzo Monacelli

17 febbraio 2016

Indice

Prefazione

Gli appunti raccolti sono frutto delle lezioni del corso del professor Loreto, nell'anno accademico 2015-2016 università di Roma "La sapienza".

Qualunque errore o svista è da imputare all'autore degli appunti, che non ha ancora sottoposto il testo al professore per una sua approvazione dei contenuti. Per chiarimenti, segnalazioni o alre comunicazione è possibile contattarmi al seguente indirizzo e-mail.

mesonepigreco@gmail.com

Lorenzo Monacelli.

La password del corso è SC.2000 per accedere ai contenuti protetti da copyright. La mail di riferimento del corso è: sistemicomplessi.sapienza@gmail.com.

Capitolo 1

Distribuzioni a coda larga

Le distribuzioni a coda larga sono leggi di probabilità che tendono all'infinito con un andamento polinomiale. La caratteristica di queste distribuzioni è l'inesistenza dei momenti della distribuzione da un certo ordine in poi, per cui non vale il generale teorema del limite centrale. Fanno la loro comparsa in quasi tutti i sistemi complessi, e il loro studio riveste un ruolo fondamentale.

1.1 Introduzione di calcolo delle probabilità

Se abbiamo una distribuzione di probabilità $\rho(x)$ devono valere le seguenti proposizioni:

$$\rho(x) \geq 0 \qquad \int_D \rho(x)dx = 1$$

Il differenziale della funzione di distribuzione è la probabilità che la variabile sia compresa nell'intervallo $[x, x + dx]$:

$$\rho(x)dx = P\{x < X < x + dx\} \tag{1.1}$$

La variabile indicata con il carattere maiuscolo X rappresenta l'insieme di tutti i possibili valori della variabile casuale, mentre x è il singolo valore assunto da X.

$$P(x > x_0) = P_>(x_0) = \int_{x_0}^{b} \rho(x)dx \qquad P(x < x_0) = P_<(x_0) = \int_{a}^{x_0} \rho(x)dx$$

Dalla definizione di distribuzione di probabilità si può passare alla definizione della media di un osservabile $O(x)$. Il modo più semplice è campionare questo osservabile tante volte, e poi calcolare la media algebrica:

$$\bar{O} = \frac{1}{N}\sum_{i=1}^{N} O_i$$

Possiamo raggruppare le O_i in base al loro valore, in modo da chiamare f_z il numero di volte che è uscito il valore O_i (se O_i è una variabile continua si può immaginare raggruppare i dati in modo da generare un istogramma):

$$\bar{O} = \frac{1}{N}\sum_{z=1}^{Z} f_z O_z$$

Dove z è l'indice con cui abbiamo suddiviso il campionamento, Z il numero totale di bins introdotti. Se prendiamo il limite per N che va a infinito:

$$\bar{O} = \lim_{N\to\infty} \sum_z \frac{f_z}{N} O_z$$

Per il principio dei grandi numeri il valore f_z/N è proprio la probabilità:

$$\bar{O} = \sum_z^Z p_z O_z$$

Nel limite per $Z \to \infty$ otteniamo[1]:

$$\bar{O} = \int dp(z) O(z) = \int \rho(z) dz O(z)$$

I valori attesi di alcuni osservabili possono essere interessanti da studiare, come l'osservabile potenza di x:

$$O_n(x) = x^n$$

Il valore medio di O_n è il momento n-esimo della distribuzione.

Ci sono distribuzioni per le quali la media non è definita, vuol dire che se facciamo un numero infinito di campionamenti la media diverge. Una distribuzione di questo tipo è la Lorentziana:

$$\rho(x) = \frac{c}{b^2 + x^2}$$

Questa distribuzione converge (può essere normalizzata), ma la sua media è problematica:

$$\langle x \rangle = \int_{-\infty}^{\infty} \frac{cx}{b^2 + x^2} dx$$

Questa distribuzione non ha la media[2]. Quando calcoliamo la media otteniamo un valore che si distribuisce intono allo zero, ma con una varianza che non può mai essere annullata, anche con un numero infinito di misure.

1.1.1 Cambiamenti di variabile

Un problema molto utile da discutere è quello di chiedersi come cambia la distribuzione di probabilità di una variabile casuale quando facciamo un cambiamento di variabile:

$$X \to \rho(x) \qquad y = f(x)$$

Se X ha una distribuzione $\rho(x)$, qual è la distribuzione della variabile casuale Y ottenuta attraverso la trasformazione $y = f(x)$?

La probabilità di trovare la x in un intervallo $[x_1, x_2]$ deve essere uguale a quella di trovare la Y in $[f(x_1), f(x_2)]$ se la funzione è monotona (Figura 1.1).

$$\rho(x)dx = \tilde{\rho}(y)dy$$

[1] Per $Z \to \infty$ il valore di p_z diventa la probabilità di avere un evento con z compreso tra z e $z + dz$, che può essere riscritto come $\rho(z)dz$ (equazione 1.1)

[2] In realtà se prendiamo la parte principale di Cauchy dell'integrale la media viene pari a zero (proprio come ci aspettiamo). Tuttavia se calcoliamo il momento secondo (la varianza) questo integrale diverge anche con la parte principale di Cauchy.

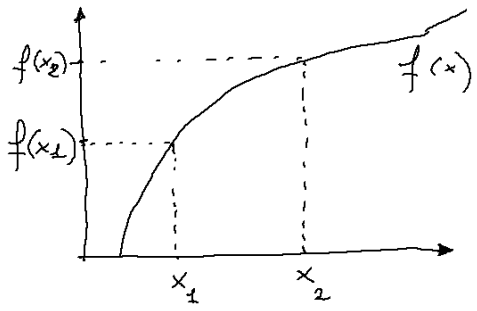


Figura 1.1: Schema di come cambia la probabilità di trovare la variabile x o y in un dato intervallo.

$$\tilde{\rho}(y) = \rho(x)\frac{dx}{dy} = \rho(x)\frac{1}{f'(x)}$$

Poiché vogliamo la la distribuzione espressa in y dobbiamo invertire la relazione che lega x a y:

$$\tilde{\rho}(y) = \frac{\rho\left[f^{-1}(y)\right]}{|f'\left[f^{-1}(y)\right]|} \tag{1.2}$$

Se la funzione non è monotona dobbiamo spezzare le probabilità in intervalli in cui la funzione è monotona e ripetere questo conto per ciascun intervallo. Il modulo sulla derivata è stato aggiunto per per evitare che la distribuzione di probabilità diventi negativa in caso di funzioni decrescenti.

Questo permette idealmente di generare delle variabili casuali distribuite secondo qualsiasi distribuzione. Immaginiamo che y sia come probabilità la $P_<(x)$:

$$y = f(x) = P_<(x) = \int_a^x \rho(x')dx'$$

Con questa scelta la $\tilde{\rho}(y)$ è banale:

$$f'(x) = \rho(x)$$

$$\tilde{\rho}(y) = \frac{\rho(x)}{f'(x)} = 1$$

Se costruiamo la y con la comulativa della $\rho(x)$, questa è distribuita in modo uniforme nell'intervallo. È possibile invertire questa relazione per passare da variabili distribuite in modo uniforme a variabili distribuite secondo una certa distribuzione $\rho(x)$:

$$x = f^{-1}(y)$$

È chiaro che per estrarre numeri distribuiti con una certa legge $\rho(x)$ occorre conoscere sia la sua primitiva $f(x)$, che la sua inversa.

1.1.2 Funzione caratteristica

Nelle funzioni di distribuzione a coda larga non vale il teorema del limite centrale, perché la varianza non è finita, dobbiamo pertanto costruire un suo ana-

logo. Per farlo è comodo lavorare con la funzione caratteristica, definita come la *trasformata di Fourier* della funzione di distribuzione della probabilità:

$$\hat{\rho}(z) = \int_D e^{izx}\rho(x)dx \leftrightarrow \rho(x) = \frac{1}{2\pi}\int e^{-izx}\hat{\rho}(z)dz$$

Questa funzione può essere definita sui numeri complessi, quindi l'integrale può essere fatto su cammini particolari del piano complesso. Se la $\rho(x)$ è simmetrica rispetto alla x la $\rho(z)$ è reale. Questo può essere visto banalmente:

$$\hat{\rho}(z) = \int_{-\infty}^{\infty} e^{izx}\rho(x)dx = \int_{-\infty}^{\infty} \cos(zx)\rho(x)dx + i\int_{-\infty}^{\infty} \sin(zx)\rho(x)dx$$

Se $\rho(x)$ è simmetrica il secondo termine è l'integrale di una funzione globalmente antisimmetrica, su un intervallo pari, quindi si annulla. Si nota facilmente che dalla normalizzazione della $\rho(x)$ segue immediatamente che:

$$\hat{\rho}(z=0) = 1$$

Una delle caratteristiche interessanti della $\hat{\rho}(z)$ è il significato delle sue derivate. Le derivate rispetto a z di questa funzione, otteniamo oggetti proporzionali ai momenti della distribuzione.

$$\partial_z^n \hat{\rho}(z) = \int_D (ix)^n e^{izx}\rho(x)$$

$$\partial_z^n \hat{\rho}(z=0) = i^n \langle x^n \rangle$$

Possiamo anche definire un altro oggetto: il *cumulante*.

$$c_n = [-i\partial_z^n]^n \ln \hat{\rho}(z=0)$$

Possiamo calcolarne i primi esplicitamente per capirne il significato:

$$c_1 = \frac{\partial_z \hat{\rho}(z)}{i\hat{\rho}(z)}_{z=0} = \langle x \rangle$$

$$c_2 = \partial_z \left[\frac{\partial_z \hat{\rho}(z)}{\hat{\rho}(z)}\right] = \frac{\partial_z^2 \hat{\rho}(z) - [\partial_z \hat{\rho}(z)]^2}{\hat{\rho}^2(z)}_{z=0} = \left[\langle x^2 \rangle - \langle x \rangle^2\right] = \sigma^2$$

Se si va avanti si ottiene:

$$c_3 = \langle (x - \langle x \rangle)^3 \rangle$$

$$c_4 = \langle (x - \langle x \rangle)^4 \rangle - 3\sigma^4$$

c_1 rappresenta il valore atteso della distribuzione, c_2 la varianza, il termine c_3 si chiama *skiwness* che ci dice quanto non è simmetrica rispetto al massimo questa funzione. In genere la *skewenss* è definita adimensionalmente come λ_3:

$$\lambda_3 = \frac{c_3}{\sigma^3}$$

$$\lambda_4 = \frac{c_4}{\sigma^4}$$

La λ_4 si chiama *curtosi* e ci dice quanto ci allontaniamo dalla gaussiana, più è appuntita la funzione maggiore è la curtosi, più è arrotondata, più la curtosi è negativa.

La gaussiana è una distribuzione che ha la caratteristica di possedere solo i primi due momenti.

$$\rho_G = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

La funzione caratteristica della gaussiana è:

$$\hat{\rho}_G(z) = e^{i\mu z - \frac{1}{2}\sigma^2 z^2}$$

Calcoliamone il logaritmo

$$\ln \hat{\rho}_G(z) = i\mu z - \frac{1}{2}\sigma^2 z^2$$

$$c_1 = \mu \qquad c_2 = \sigma^2 \qquad c_3 = 0 \qquad c_4 = 0 \qquad \cdots$$

La funzione gaussiana ha soltanto i primi due momenti non nulli. La gaussiana può essere introdotta come distribuzione che massimizza l'entropia di una variabile casuale di cui si fissano media e varianza. Per farlo dobbiamo definire l'entropia.

L'entropia ci dice il contenuto di informazione di una sorgente di numeri casuali, con distribuzione di probabilità p_k:

$$S\{p_k\} = -\sum_k p_k \log_2 p_k$$

Si usa il logaritmo in base 2 perché Shannon introdusse questa definizione per segnali in base due. Il $-$ davanti al logaritmo fa in modo che tutto sia positivo. Il significato intuitivo di entropia è il numero di bit necessari a codificare l'informazione data dalla sorgente. Calcoliamo ad esempio l'entropia di una monetina, che ha 50% e 50% di probabilità di generare testa o croce:

$$S = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

Per n lanci ho bisogno di n bit per scrivere il messaggio contenuto nella sorgente. Se avessi una sorgente che ha delle correlazioni la sua entropia sarebbe minore.

Data una variabile casuale, con media e varianza fissata, qual è la distribuzione che massimizza l'entropia?

Rispondiamo passo passo, iniziamo con il calcolare la densità di probabilità che massimizza l'entropia di un sistema in cui non fissiamo ne media ne varianza.

$$S = -\int_a^b \rho(x)\ln\rho(x)dx \qquad \int_a^b \rho(x) = 1$$

Usando il metodo dei moltiplicatori di Lagrange, definiamo la funzione da minimizzare con vincoli:

$$\Phi = S - \mu\left[\int_a^b \rho(x)dx - 1\right]$$

$$\frac{\delta\Phi}{\delta\rho(x')} = \rho(x')\ln\rho(x') - \mu\rho(x') = 0$$

Da cui si ottiene:

$$\ln\rho(x') = \mu \qquad \rho(x') = e^{\mu} = cost$$

Imponendo la condizione di normalizzazione (o imponendo che $\frac{\delta\Phi}{\delta\mu} = 0$) si ottiene:

$$\int_a^b e^{\mu}dx = 1$$

$$e^{\mu} = \frac{1}{b-a}$$

$$\rho(x) = \frac{1}{b-a}$$

Quindi la distribuzione uniforme massimizza l'entropia. Se fissiamo la media della distribuzione otteniamo un altro vincolo

$$\int_a^b x\rho(x)dx = m$$

In questo caso la funzione da ottimizare è:

$$\Phi = S - \mu\left[\int_a^b \rho(x)dx - 1\right] - \lambda\left[\int_a^b x\rho(x)dx - m\right]$$

$$\frac{\delta\Phi}{\delta\rho(x')} = -\rho(x')\ln\rho(x') - \mu\rho(x') - x\rho(x')\lambda = 0$$

$$\ln\rho(x') = \lambda x + \mu$$

$$\rho(x') = e^{\mu}e^{\lambda x}$$

Imponendo la normalizzazione e il fatto che la media vale m (oppure calcolando le altre due derivate funzionali rispetto a μ e λ e imponendo che is annullino) otteniamo:

$$\rho(x) = me^{-\frac{x}{m}}$$

Questo rimane vero anche nel caso in cui l'integrale è definito tra due variabili a e b a patto che il valor atteso non sia il punto medio dell'intervallo $[a, b]$, in quel caso ritroviamo (ovviamente) la distribuzione uniforme.

Se fissiamo anche il momento secondo:

$$\int_a^b x^2\rho(x)dx = \sigma^2$$

Allora otteniamo la gaussiana. Ripetiamo anche in questo caso il conto:

$$\Phi[\rho] = -\int \rho(x)\ln\rho(x)dx$$

Con vincoli su ρ

$$\int \rho(x)dx = 1 \qquad \int x\rho(x)dx = m \qquad \int x^2\rho(x)dx = \sigma^2$$

Costruiamo la funzione[3]:

$$\Phi[\rho(x), \mu, \lambda, \gamma] = -\int \rho(x') \ln \rho(x') dx' + \mu \int \rho(x') dx' + \lambda \int x' \rho(x') dx' + \gamma \int x'^2 \rho(x') dx'$$

$$\frac{\delta \Phi}{\delta \rho(x)} = -\ln \rho(x) - 1 + \mu + \lambda x + \gamma x^2 = 0$$

Questa derivata funzionale corrisponde a fare la derivata della $\rho(x)$-esima variabile di cui è funzione la Φ, in cui la Φ è funzione di un infinito contiuo di variabili. Quindi tra tutti i termi dell'integrale $\rho(x')$ l'unico termine della derivata che non si annulla è $\rho(x)$.

Da cui otteniamo:

$$\ln \rho(x) = -1 + \mu + \lambda x + \gamma x^2$$

$$\rho(x) \propto e^{\lambda x + \gamma x^2} \propto e^{-\frac{(x-m)^2}{2\sigma^2}}$$

I moltiplicatori possono essere determinati imponendo le condizioni al contorno.

Per capire se effettivamente la distribuzione gaussiana è un massimo o un minimo dell'entropia dobbiamo fare la derivata seconda funzionale:

$$\frac{\delta^2 \Phi}{\delta \rho(x) \delta \rho(x')} = -\frac{1}{\rho(x')} \delta(x - x')$$

Quindi la derivata seconda è sempre negativa, e quindi in corrispondenza della gaussiana abbiamo un massimo dell'entropia. La gaussiana è importante perché esce fuori dal teorema del limite centrale: se abbiamo una variabile casuale che ha una certa distribuzione con varianza finita, la media si distribuisce secondo una distribuzione gaussiana.

1.1.3 Teorema del limite centrale

Il teorema del limite centrale afferma che una qualunque variabile casuale, ottenuta dalla media di variabili gaussiane, nel limite $N \to \infty$ si distribuisce in modo gaussiano.

Chiediamoci ad esempio come si ricava la distribuzione di probabilità della somma di variabili casuali. Immaginiamo di avere due dadi, qual'è la distribuzione che esca 8?

$$p_8 = p_2 p_6 + p_3 p_5 + p_4 p_4 + p_5 p_3 + p_6 p_2$$

$$p_8 = \sum_{k=1}^{6} p_k p_{8-k} \qquad p_7 = 0$$

Questo risultato può essere esteso a variabili continue. Supponaiamo di avere la variabile x_1 distribuita con la $\rho_1(x_1)$ e la variabile x_2 secondo la probabilità $\rho(x_2)$. La variabile casuale x continua è:

$$x = x_1 + x_2 \qquad \rho(x) = \int_{D_1} \rho_1(x_1) \rho_2(x - x_1) dx_1$$

[3]Possiamo droppare le costanti nei vincoli, che sono determinanti solo per valutare i valori dei moltiplicatori di Lagrange. Tuttavia possiamo poi imporre direttamente i vincoli sul risultato ottenuto per riottenere gli stessi risultati.

Questo è un prodotto di convoluzione. Se ne facciamo la trasformata di fourier diventa un prodotto semplice. La funzione caratteristica è data semplicemente dal prodotto delle funzioni caratteristiche.

$$\hat{\rho}(z) = \hat{\rho}_1(z)\hat{\rho}_2(z)$$

Questa proprietà è molto buona, perché facendo il logaritmo ad entrambi i membri otteniamo una somma:

$$\ln \hat{\rho}(z) = \ln \hat{\rho}_1(z) + \ln \hat{\rho}_2(z)$$

A questo punto possiamo ottenere come si comportano i comulanti della distribuzione somma:

$$c_n = \left[\frac{d}{idz}\right]^n \hat{\rho}(z=0)$$

I cumulanti della distribuzione somma sono pari alla somma dei cumulanti delle rispettive funzioni, ed in particolare per media e varianza.

$$m = m_1 + m_2 \qquad \sigma^2 = \sigma_1^2 + \sigma_2^2$$

Questo risultato è immediato ottenere l'estensione ad n variabili distribuite tutte con la stessa distribuzione:

$$x = \sum x_i \qquad \rho_i(x) = \rho_j(x)$$

$$\ln \hat{\rho}(z) = \sum_{i=1}^{N} \ln \hat{\rho}_i(z) = N \ln \hat{\rho}_i(z)$$

$$\hat{\rho}(z) = [\hat{\rho}_i(z)]^N$$

Allo stesso modo per i cumulanti otteniamo:

$$m = Nm_0 \qquad \sigma^2 = N\sigma_0^2$$

Il cumulante di ordine n della somma di N elementi va come:

$$c_{n,N} = Nc_{n,1}$$

Dove indichiamo nel primo pedice l'ordine del comulante, nel secondo la la distribuzione che genera quel comulante. Adesso possiamo farne la media, basta introdurre una nuova variabile:

$$y = \frac{x}{N}$$

Dobbiamo capire come si distribuisce la y (equazione 1.2).

$$\tilde{\rho}(y)dy = \rho(x)dx \qquad \tilde{\rho}(y)$$

$$\hat{\tilde{\rho}}(z) = \int e^{izy}\tilde{\rho}(y)dy = \int e^{iz\frac{x}{N}}\rho(x)dx = \hat{\rho}\left(\frac{z}{N}\right) \tag{1.3}$$

Il cumulante n-esimo della media sarà dunque dato da:

$$c_{n,\langle\rangle} = \frac{1}{N^n}Nc_{n,1} = N^{1-n}c_{n,1} \tag{1.4}$$

Il valore della media rimane quindi uguale:

$$c_{1,\langle\rangle} = c_{1,1}$$

Il valore aspettato della media deve necessariamente essere uguale al valore aspettato della singola variabile.

$$c_{2,\langle\rangle} = \frac{c_{2,1}}{N} \qquad \sigma^2 = \frac{\sigma_0^2}{N}$$

Tutti gli altri cumulanti vanno a zero con esponenti via via più alti. Se N va all'infinito muoiono tutti i cumulanti, e la distribuzione tende ad una delta. Quello che si fa è passare ad un altra variabile.

$$v = (y - m_0)\,\frac{\sqrt{N}}{\sigma_0} \tag{1.5}$$

Abbiamo shiftato la curva in modo che $c_{1,v} = 0$. Da cui possiamo ricavare allo stesso modo i cumulanti:

$$c_{n,v} = N^{1-n} c_{n,1} \left(\frac{\sqrt{N}}{\sigma_0}\right)^n$$

$$c_{n,v} = N^{1-\frac{n}{2}} \frac{c_{n,1}}{\sigma_0^n}$$

In particolare

$$c_{2,v} = 1$$

Tutti i cumulanti da 3 in poi dipendono da una potenza negativa di N. Nel limite per N che tende ad infinito il resto dei cumulanti vanno a zero. Quindi l'unica funzione di distribuzione con soli i primi due cumulanti è proprio la gaussiana. Questo è valido solo fin che la funzione è derivabile infinite volte. Se la ρ ha una legge a potenza ad un certo punto ci sarà un momento in cui il cumulante n-esimo diverge.

Nella definizione della funzione caratteristica abbiamo

$$\hat{\rho}(z) = \int e^{izx}\rho(x)dx = \int \left(1 + izx - \frac{1}{2}z^2x^2 + O(z^3)\right)\rho(x)dx$$

Quiando sviluppiamo la somma otteniamo:

$$\hat{\rho}(z) = 1 + iz\,\langle x\rangle - \frac{1}{2}z^2\,\langle x^2\rangle + o(z^2)$$

Shiftiamo la funzione in modo che $\langle x\rangle = 0$

$$\hat{\rho}(z) = 1 - \frac{1}{2}z^2\,\langle x^2\rangle + o(z^2)$$

Se facciamo la somma di più variabili otteniamo:

$$\hat{\rho}_\Sigma(z) = \left(1 - \frac{1}{2}z^2\sigma^2 + o(z^2)\right)^N \qquad x = \sum_{i=1}^{N} x_i$$

$$\hat{\rho}_{\langle\rangle}(z) = \left(1 - \frac{1}{2}\frac{z^2}{N^2}\sigma^2 + o(z^2)\right)^N$$

Dobbiamo passare alla variabile v

$$v = \frac{y\sqrt{N}}{\sigma}$$

Bisogna sostituire a z con $\frac{z\sqrt{N}}{\sigma}$:

$$\hat{\rho}_v(z) = \left(1 - \frac{1}{2}\frac{z}{N} + o(z^2)\right)^N$$

Se facciamo il limite per $N \to \infty$

$$\lim_{N\to\infty} \hat{\rho}_v(z) = e^{-\frac{1}{2}z^2}$$

Non è così ovvio che v sia effettivamente distribuita con una gaussiana. Infatti non è detto che la trasformata di Fourier della funzione genetratrice converga uniformemente:

$$\rho(v) = \frac{1}{2\pi}\int e^{-izv}\hat{\rho}(z)dz$$

Abbiamo dimostrato la convergenza puntuale alla gaussiana della $\hat{\rho}$, perché l'integrale sia correttamente definito occorre tuttavia che la convergenza sia uniforme. Questa convergenza è assicurata per fortuna da un teorema matematico.

1.2 Leggi a potenza

Molte distribuzioni presentano l'andamento a legge di potenza per alti valori delle variabili che si studiano. Definiamo ora una variabile che rispetta la legge di potenza nel seguente modo.

$$\rho(x) = \begin{cases} cx^{-\alpha} & x \geq m \\ 0 & x < m \end{cases} \qquad \alpha > 1$$

La costante c può essere determinanta dalla normalizzazione:

$$\int_m^\infty cx^{-\alpha}dx = 1 \qquad c\left[\frac{x^{1-\alpha}}{1-\alpha}\right]_m^\infty = 1 \qquad -\frac{1}{1-\alpha}m^{1-\alpha}c = 1$$

$$c = (\alpha - 1)m^{\alpha-1}$$

La funzione di distribuzione diventa:

$$\rho(x) = (\alpha - 1)m^{\alpha-1}x^{-\alpha}$$

La distribuzione comulativa è semplice da trovare analiticamente:

$$P_>(x) = \int_x^\infty \rho(x)dx = \left(\frac{x}{m}\right)^{1-\alpha}$$

$$P_<(x) = \int_m^x \rho(x)dx = 1 - \left(\frac{x}{m}\right)^{1-\alpha}$$

Possiamo sfruttare quanto visto nella sezione precedente per ricavare come si estraggono numeri casuali distribuiti a legge di potenza. Calcoliamo la primitiva:

$$r = P_<(x) = \int_m^x (\alpha - 1)m^{\alpha-1}x^{-\alpha}dx$$

$$r = 1 - \left(\frac{x}{m}\right)^{1-\alpha}$$

r è la variabile distribuita tra 0 e 1 uniformemente. Invertiamo la relazione per ottenere x con distribuzione a legge di potenza.

$$\left(\frac{x}{m}\right)^{1-\alpha} = 1 - r \qquad \frac{x}{m} = (1-r)^{\frac{1}{1-\alpha}}$$

$$x = m(1-r)^{\frac{1}{1-\alpha}}$$

Siccome r è estratto uniformemente tra 0 e 1, possiamo sostituire r con $1 - r$, e ottenere dati estratti a legge di potenza da un computer con il seguente sistema:

$$x = mr^{\frac{1}{1-\alpha}}$$

1.2.1 Proprietà estremali

Supponiamo di estrarre N numeri distribuiti a legge di potenza, qual è il numero x_n tale ché in media ci sia una sola estrazione con $x \geq x_n$? Rispondere a questa domanda corrisponde a chiedersi come è distribuito il valore più alto ottenuto in una serie di estrazioni. Se troviamo un certo x_n tale che

$$P_>(x_n) \approx \frac{1}{N}$$

Vuol dire che in N estrazioni capita al più una singola volta un numero maggiore di x_n. Quindi x_n è detto estremale della distribuzione:

$$N\left(\frac{x_n}{m}\right)^{1-\alpha} \approx 1 \qquad x_n \approx N^{\frac{1}{\alpha-1}} \tag{1.6}$$

La cosa interessante è che anche x_n ha un andamento a legge di potenza con il numero di estrazioni N. Questa è una caratteristica tipica delle funzioni a coda larga. Questo può essere ricavato in modo più rigoroso, calcolando la distribuzione di probabilità di x_n.

Prendiamo un numero $M \gg m$, sulla coda della distribuzione.

$$P_<(M) = \int_m^M \rho(x)dx$$

Questa è la probabilità che la variabile casuale x sia minore di M. Dopo N estrazioni, qual è la probabilità che tutti i numeri siano contenuti in questa regione?

$$P_<(M)^N = (1 - P_>(M))^N \approx e^{-NP_>(M)}$$

Questa è la probabilità che dopo N estrazioni tutte le variabili siano minori di M (e in particolare la più grande tra loro x_n lo sia). Questa è la distribuzione comulativa di x_n.

$$\rho(x_n) \approx \frac{\partial}{\partial n} e^{-NP_>(M)} = N\rho(x_n)e^{-NP_>(x_n)}$$

$$\rho(x_n) = x_n^{-\alpha} e^{-N\left(\frac{x_n}{x_{min}}\right)^{1-\alpha}}$$

Per valori di $x_n \gg m$ si ottiene

$$\rho(x_n) \approx x_n^{-\alpha}$$

Abbiamo dimostrato che il valore maggiore di una serie di estrazioni ha esattamente la stessa distribuzione della singola estrazioni (per alti valori di x_n). La distribuzione completa del valor estremo è detta distriuzione di Fréchet.

1.2.2 Legge di Zipf

La distribuzione a coda larga è anche detta distribuzione di Zipf. In genere questo si ottiene quando si misura la distribuzione sperimentale delle parole nei testi. Una funzione studiata per caratterizzare la linguisticaSe contiamo il numero di parole differenti che leggiamo (funzione dizionario) otteniamo una legge di potenza:

$$D(t) \sim t^{\gamma} \qquad 0 < \gamma < 1$$

Questa è la legge di Heaps. Se abbiamo una legge del tipo $y = kx^{-\alpha}$ la legge di potenza può essere riconosciuta in un grafico doppio-logaritmico (Figura 1.2).

$$\log y = \log k - \alpha \log x$$

Figura 1.2: Dati distribuiti in una legge di potenza su un grafico doppio-logaritmico.

Spesso però questi grafici sono molto sporcati. Per ovviare al problema si può plottare la cumulativa che abbatte il rumore.

Esiste anche un altro metodo moltousato, il *frequency rank*. Immaginiamo di avere il numero di abitanti per ogni città.

Ordiniamo i dati in classifica, mettendo al primo posto la città con più abitanti, e così via. Andiamo a plottare il logaritmo della grandezza studiata in

funzione del logaritmo del rank (posizione in classifica). I punti si mettono su una legge a potenza. Questo è un metodo molto efficacie.

Per dimostrare questa proprietà dobbiamo capire chi è il rank. Rank pari ad 1 vuol dire che il numero di città che hanno un numero di abitanti maggiore o uguale al suo valore è 1. Il ranking è il numero di elementi maggiori o uguali di un certo valori fissato.

$$NP_{>}(x) = r$$
$$r \propto x^{1-\alpha}$$
$$x \propto r^{-\frac{1}{\alpha-1}}$$

Se plottiamo il logaritmo del valore in funzione del logaritmo del ranking otteniamo una retta che ha come coefficiente angolare $1/(\alpha-1)$. Se facciamo questo per le parole di un testo otteniamo:

$$f \approx r^{-1}$$

Che implica che $\alpha \approx 2$. Questa legge è detta *legge di Zipf*. Questa legge semplice permette di valutare il coefficiente γ della legge di Heaps.

Quando si legge una parola nuova, nell'intero testo, fin dove siamo arrivati a leggere, è apparsa con frequenza di $1/N$. Possiamo ricollegare la frequenza alla legge di Zipf:

$$f = \frac{1}{N} = r^{-\beta} \qquad r \approx N^{\frac{1}{\beta}} \qquad \gamma = \frac{1}{\beta} = \alpha - 1$$

La pendenza della frequency rank per alti rank (sulla coda) è legato alla legge di Heaps. Per testi molto lunghi l'andamento della legge del rank vs valore cambia bruscamente pendenza, per questo è necessario prendere il valore sulla coda (che può essere diverso dal valore iniziale).

1.2.3 Stima della pendenza della retta nella legge di potenza

I minimi quadrati sono spesso pessimi per studiare questi oggetti. Infatti l'informazione più importante si trova sulle code. Per ovviare a questo problema si usa la divergenza di *Kullback-Leibler*, che rappresenta quanti bit extra ci servono per codificare la variabile $\{q_k\}$.

$$D_{KL} = \sum_k p_k \log_2 \frac{p_k}{q_k} = \sum_k p_k \log_2 p_k - \sum_k p_k \log_2 q_k = -S(p) + S(p,q) \quad (1.7)$$

$$D_{KL}(p,q) \neq D_{KL}(q,p)$$

Questo oggetto non è una distanza, però ci da una idea di quanto siano diverse le due funzioni, e viene utilizzata per ricavare il principio di massima verosimiglianza.

Usiamo questa funzione per ricavare i parametri della distribuzione che vogliamo studiare. Immaginiamo di assumere che un fenomeno, distribuito secondo una funzione $\rho(x|\theta_0)$, sia distribuito a legge di potenza. Possiamo chiederci quali sono i parametri θ che massimizzano la "sovrapposizione" tra le due funzioni. Per confrontarle usiamo la divergenza di Kullback-Leibler.

La funzione di distribuzione che misuriamo sia $\rho(x|\theta)$. Possiamo scrivere la D_{KL} come una sommatoria:

$$D_{KL} = \sum \rho(x|\theta_0) \log \frac{\rho(x|\theta_0)}{\rho(x|\theta)}$$

Se la funzione di distribuzione del fenomeno è esattamente una legge di potenza allora esiste un set di parametri θ che annullano la D_{KL}. Tuttavia questo non è mai vero[4], e la cosa migliore che possiamo fare è trovare per quali valori di θ la D_{KL} è minima.

Si può interpretare la definizione di D_{kL} (1.7) come il calcolo di una media:

$$D_{KL} = \left\langle \log_2 \frac{\rho(x,\theta_0)}{\rho(x|\theta)} \right\rangle = \frac{1}{N} \sum_{i=1}^{N} \log_2 \frac{\rho(x_i,\theta_0)}{\rho(x_i|\theta)}$$

Poiché vogliamo minimizzare rispetto a θ l'unico termine non costante è (spezzando il logaritmo)

$$-\sum_i \log(\rho(x_i|\theta)$$

Trovare il minimo di D_{kl} equivale a massimizzare la funzione:

$$\sum_i \log(\rho(x_i|\theta) = \log \nu(\theta) \tag{1.8}$$

Dove ν è detta *verosimiglianza* del problema.

Sostituiamo adesso con ρ la legge di potenza.

$$y = \begin{cases} \left(\frac{\alpha-1}{m}\right)\left(\frac{x}{m}\right)^{\alpha} & x > m \\ \\ 0 & x \leq m \end{cases}$$

La funzione di verosimiglianza diventa:

$$\nu(\alpha) = \prod_{i=1}^{N} \left(\frac{\alpha-1}{m}\right)\left(\frac{x_i}{m}\right)^{-\alpha}$$

Calcoliamo il logaritmo (1.8):

$$\ln \nu(\alpha) = N \ln\left(\frac{\alpha-1}{m}\right) - \alpha \sum_{i=1}^{N} \ln\left(\frac{x_i}{m}\right)$$

Massimizziamola rispetto al parametro α.

$$\frac{\partial}{\partial\alpha} \ln \nu(\alpha) = \frac{N}{\alpha-1} - \sum_{i=1}^{N} \ln\left(\frac{x_i}{m}\right) = 0$$

$$\frac{\alpha-1}{N} = \frac{1}{\sum_{i=1}^{N} \ln\left(\frac{x_i}{m}\right)}$$

[4]Per piccoli x possono essere presenti altri termini che decadono esponenzialmente, effetti di cut-off a grandi x, o semplicemente l'andamento misurato non è una legge di potenza.

$$\alpha = 1 + \frac{N}{\sum_{i=1}^{N} \ln\left(\frac{x_i}{m}\right)} = 1 + \frac{1}{\langle \ln\left(\frac{x}{m}\right)\rangle}$$

Introducendo b per semplificare la notazione, l'esponente della legge di potenza diventa:

$$\alpha = 1 + \frac{N}{b} \qquad b = \sum_{i=1}^{N} \ln\left(\frac{x_i}{m}\right)$$

Possiamo stimare l'errore commesso nel calcolo di α con un metodo empirico. Tanto migliore è il modello, tanto più la funzione di verosimiglianza sarà piccata (Figura 1.3)

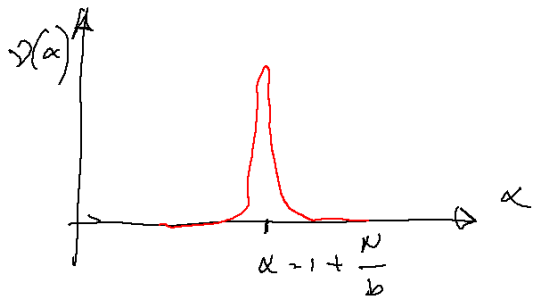


Figura 1.3: Schema della funzione di verosimiglianza ν. Il massimo della ν coincide con il miglior valore di α per approssimare i dati. Una stima dell'errore di α può essere ottenuta cercando di valutare la larghezza di questa funzione.

Possiamo stimare la larghezza di ν facendo finta che sia a sua volta una distribuzione calcolandone la deviazione standard.

$$\langle \alpha \rangle = \frac{\int \alpha \nu(\alpha) d\alpha}{\int \nu(\alpha) d\alpha}$$

$$\langle \alpha \rangle = \frac{\int \alpha \left(\frac{\alpha-1}{m}\right)^N e^{-\alpha b} d\alpha}{\int \left(\frac{\alpha-1}{m}\right)^N e^{-\alpha b} d\alpha}$$

Svolgendo i calcoli diventa:

$$\langle \alpha \rangle = 1 + \frac{N}{b} \tag{1.9}$$

La deviazione standard può essere calcolata:

$$\sigma = \sqrt{\langle \alpha^2 \rangle - \langle \alpha \rangle^2}$$

Svolgendo i calcoli si ottiene:

$$\sigma = \sqrt{\langle \alpha^2 \rangle - \langle \alpha \rangle^2} = \frac{\sqrt{N+1}}{b}$$

Il coefficiente b può essere ricavato dall'espressione del valor medio (1.9):

$$b = \frac{N}{\langle \alpha \rangle - 1} \qquad \sigma = \frac{\langle \alpha \rangle - 1}{\sqrt{N}}$$

Nell'ipotesi $N \gg 1$. Questo ci piace perché all'aumentare di N la deviazione standard di α diminuisce, permettendo di stimare, con un opportuno numero di dati, il coefficiente α con la precisione voluta.

1.3 Teorema del limite centrale generalizzato

Cerchiamo di estendere il teorema del limite centrale, dimostrato per distribuzioni i cui momenti sono sempre finiti in Sez. 1.1.3, alle leggi di potenza. Cambiamo leggermente la definizione di legge di potenza per renderla simmetrica rispetto a $x = 0$ (e avere una funzione caratteristica reale).

$$y = \begin{cases} \frac{1}{2}\left(\frac{\alpha-1}{m}\right)\left(\frac{|x|}{m}\right)^{-\alpha} & |x| > m \\ \\ 0 & |x| < m \end{cases}$$

Calcoliamo la funzione caratteristica:

$$1 - \hat{\rho}(t) = 1 - \int_{-\infty}^{\infty} e^{itx}\rho(x)dx = 1 - \int_{-\infty}^{-m} e^{itx}\rho(x)dx - \int_{m}^{\infty} e^{itx}\rho(x)dx$$

Usiamo la condizione di normalizzazione della distribuzione per portare quell'1 dentro l'integrale:

$$1 = \int_{-\infty}^{-m} \rho(x)dx + \int_{m}^{\infty} \rho(x)dx$$

$$1 - \hat{\rho}(t) = \int_{-\infty}^{-m} \left(1 - e^{itx}\right)\rho(x)dx + \int_{m}^{\infty} (1 - e^{itx})\rho(x)dx$$

Facciamo un cambio di variaili nel primo integrale:

$$1 - \hat{\rho}(t) = \int_{m}^{\infty} (1 - e^{-itx})\rho(x)dx + \int_{m}^{\infty} (1 - e^{itx})\rho(x)dx$$

Dove abbiamo sfruttato la parità[5] della $\rho(x)$. Possiamo mettere tutto sotto lo stesso segno di integrale:

$$1-\hat{\rho}(t) = \int_{m}^{\infty} \left[2 - \left(e^{-itx} + e^{itx}\right)\right]\rho(x)dx = \int_{m}^{\infty} 2\left[1 - \cos(tx)\right]\frac{1}{2}\left(\frac{\alpha-1}{m}\right)\left(\frac{|x|}{m}\right)^{-\alpha}dx$$

$$1 - \hat{\rho}(t) = (\alpha-1)m^{\alpha-1}\int_{m}^{\infty} \frac{1-\cos(tx)}{x^{\alpha}}dx$$

In realtà $\hat{\rho}(t)$ dipende solo da $|t|$ (è una funzione pari). Possiamo esplicitare questa proprietà con un cambiamento di variabile nell'integrale:

$$u = |t|x \qquad x = \frac{u}{|t|}$$

$$1 - \hat{\rho}(t) = (\alpha-1)m^{\alpha-1}|t|^{\alpha-1}\int_{|t|m}^{\infty} \frac{1-\cos u}{u^{\alpha}}du$$

[5] $\rho(x) = \rho(-x)$

$$1 - \hat{\rho}(t) = (\alpha - 1)(m|t|)^{\alpha - 1} \int_{m|t|}^{\infty} \frac{1 - \cos u}{u^{\alpha}} du$$

Troviamo subito una condizione affinché la $\hat{\rho}(t)$ sia bene definita: $\alpha > 1$, altrimenti l'integrale non converge.

Per piccoli valori di t il numeratore dell'integranda si può approssimare:

$$1 - \cos u \approx \frac{1}{2} u^2$$

$$\int_{m|t|}^{\infty} \frac{1 - \cos u}{u^{\alpha}} du \overset{t \to 0}{\longrightarrow} \int_{0}^{\infty} \frac{1 - \cos u}{u^{\alpha}} du - \int_{0}^{m|t|} \frac{1}{2} u^{2-\alpha} du = C + \tilde{C}|t|^{3-\alpha}$$

Dove C e $\tilde{C}$ sono due costanti.

Se $\alpha < 3$ per piccoli t domina il termine costante (C)

$$\int_{m|t|}^{\infty} \frac{1 - \cos u}{u^{\alpha}} du \approx C \qquad \alpha < 3$$

Se $\alpha < 3$ per piccoli t il termine che moltiplica $\tilde{C}$ diverge, e possiamo trascurare la costante.

$$\int_{m|t|}^{\infty} \frac{1 - \cos u}{u^{\alpha}} du \approx \tilde{C}|t|^{3-\alpha} \qquad \alpha > 3$$

Il caso $\alpha = 3$ è molto particolare e raro, non sarà discusso in questi appunti[6].

Possiamo sostituire quindi nell'espressione della funzione generatrice:

$$1 - \hat{\rho}(t) \approx (\alpha - 1) m^{\alpha - 1} C |t|^{\alpha - 1} \qquad \alpha < 3$$

$$\hat{\rho}(t) \approx 1 - k|t|^{\alpha - 1} \qquad \alpha < 3$$

Si nota che la derivata seconda della ρ diverge in zero per $\alpha < 3$. La varianza della distribuzione non è definita.

Nel caso $\alpha > 3$

$$1 - \hat{\rho}(t) = (\alpha - 1) m^{\alpha - 1} |t|^{\alpha - 1} \tilde{C} |t|^{3-\alpha} \approx |t|^2 \tilde{k} \qquad \alpha > 3$$

Per $\alpha > 3$ l'esponente della t non dipende più da α. Esistono Quindi sparisce la dipendenza da α. In questo caso esiste la varianza.

Calcoliamo ora nel caso $\alpha < 1$ la funzione generatrice della somma di più variabili distribuite a potenza:

$$\hat{\rho}_{\Sigma}(t) = [\hat{\rho}(t)]^N = \left(1 - k|t|^{\alpha - 1}\right)^N$$

Il limite per $N \to \infty$ non porta direttamente a nulla. Dobbiamo trovare una variabile ausiliaria che permetta di poter fare semplicemente questo limite.

Passiamo ad una variabile s (ricordiamo come si trasformano le funzioni generatrici, eq. 1.3):

$$s = x N^{-\frac{1}{\alpha - 1}} \qquad \hat{\rho}_s(t) = \hat{\rho}\left(\frac{t}{N^{-\frac{1}{\alpha - 1}}}\right)$$

[6] In quel caso la distribuzione diventa logaritmica. Tuttavia basta che α si discosti per un ε piccolo a piacere da 3 che la nostra trattazione è consistente.

$$\hat{\rho}_s(t) = \left(1 - k\left|\frac{t}{N^{\frac{1}{\alpha-1}}}\right|^{\alpha-1}\right)^N = \left(1 - \frac{k}{N}|t|^{\alpha-1}\right)^N$$

Riconosciamo che il fattore $N^{\frac{1}{\alpha-1}}$ è la stima del massimo valore estratto dopo N tentativi (1.6).

$$s = \frac{\sum_i x_i}{x_N} \qquad x_N = \max\{x_i\} = N^{\frac{1}{\alpha-1}}$$

$$\hat{\rho}_s(t) \overset{N\to\infty}{\longrightarrow} e^{-k|t|^{\alpha-1}} \tag{1.10}$$

L'equazione (1.10) rappresenta il teorema del limite centrale per le leggi a potenza nel caso $\alpha < 3$. In generale l'antitrasformata di questa espressione non è analitica, e prende il nome di funzione di Lévy.

Se α è 2, la $\hat{\rho}_s(t)$ ha un andamento esponenziale, e la sua antitrasformata è una lorentziana. Ma anche la $\rho(x)$ con $\alpha = 2$ è una lorentziana. Inoltre la variabile s in questo caso rappresenta proprio la media aritmetica delle misure.

$$s(\alpha = 2) = \frac{\sum x_i}{N}$$

La distribuzione della media di variabili lorentziane è una lorentziana, la cui larghezza non dipende dal numero di misure fatte (come era invece nel limite centrale).

Vediamo ora il caso $\alpha > 3$.

$$\alpha = 3 + \varepsilon \qquad \varepsilon > 0$$

La distribuzione ha un andamento del tipo:

$$\rho(x) \sim x^{-(3+\varepsilon)}$$

In questo caso la varianza esiste. Vediamo il momento terzo della distribuzione:

$$\langle x^3 \rangle \sim \int_m^\infty x^3 x^{-3-\varepsilon} dx = ?$$

A seconda del valore di ε l'integrale può divergere o meno. Tuttavia siamo interessati a stimare il momento terzo a partire da una media fatta su N misure:

$$\langle x^3 \rangle_N = \frac{1}{N}\sum_i x_i^3 \overset{N\to\infty}{\longrightarrow} \langle x^3 \rangle$$

Questo tenderà al momento terzo per $N \to \infty$. Su un numero finito di misure tuttavia le x_i sono limitate dalla x_N.

$$x_N = \max\{x_i\} \sim N^{\frac{1}{\alpha-1}}$$

Tutte le $x > x_N$ vanno quindi tolte dal dominio di integrazione nella stima di $\langle x^3 \rangle_N$:

$$\langle x^3 \rangle_N \sim \int_m^{x_N} x^3 x^{-3+\varepsilon}\, dx \sim x_N^{1-\varepsilon} \sim \left(N^{\frac{1}{2+\varepsilon}}\right)^{1-\varepsilon}$$

$$\langle x^3 \rangle_N \sim N^{\frac{1-\varepsilon}{2+\varepsilon}}$$

Il comulante di terzo ordine ha lo stesso andamento del momento terzo:

$$c_{3,1} \sim N^{\frac{1-\varepsilon}{2+\varepsilon}}$$

Calcoliamo il cumulante della somma di N variabili

$$c_{3,N} \sim N \cdot N^{\frac{1-\varepsilon}{2+\varepsilon}}$$

Passiamo alla media (I passaggi tra cumulanti sono stati ricavati nell'equazione 1.4):

$$c_{3,\langle\rangle} \sim \left(\frac{1}{N}\right)^3 N N^{\frac{1-\varepsilon}{2+\varepsilon}}$$

Dobbiamo passare alla variabile v (la stessa variabile definita per il limite centrale standard, nell'equazione 1.5)

$$c_{3,v} \sim (\sqrt{N})^3 \left(\frac{1}{N}\right)^3 N^{\frac{1-\varepsilon}{2+\varepsilon}} \sim N^{-\frac{3}{2}\frac{\varepsilon}{2+\varepsilon}} \qquad v = \langle\rangle \frac{\sqrt{N}}{\sigma}$$

Quindi il terzo cumulante della variabile v tende a zero per N che va all'infinito, e con lui tutti quelli di ordine superiore. Quindi per $\alpha > 3$ esiste sempre la varianza, e tutti i cumulanti di ordine superiore al secondo sono nulli. Abbiamo ritrovato il teorema del limite centrale[7]. Se ε è molto piccolo bisogna avere N molto grande prima di avere convergenza alla gaussiana.

Conclusione

Il teorema del limite centrale esteso per funzioni a coda larga prevede che, per $N \to \infty$:

- $\alpha < 3$ la media generalizzata s segue la distribuzione di Lévy (con varianza costante).
- $\alpha > 3$ la media algebrica è distribuita con una gaussiana (con varianza che diminuisce al crescere di N).

1.4 Invarianza di scala

Molti fenomeni in fisica, o in altri ambiti della vita, sono privi di scala.

Una funzione *scale-free* può essere definita matematicamente dalla seguente relazione:

$$g(ax) = h(a)g(x) \qquad \forall x, a \in \mathbb{R} \tag{1.11}$$

L'equazione esprime il fatto che, applicando un opportuno allargamento su entrambi gli assi, la funzione $g(x)$ rimane invariata.

Poiché l'equazione deve valere per ogni x e a possiamo scegliere un caso particolare

$$x = 1$$

[7]Infatti proprio come per il teorema del limite centrale tradizionale l'unica funzione di distribuzione che ha solo i primi due cumulanti non nulli è la gaussiana.

con cui ricavare l'espressione di $h(a)$:

$$g(a) = h(a)g(1) \qquad h(a) = \frac{g(a)}{g(1)} \tag{1.12}$$

$$g(ax) = \frac{g(a)}{g(1)}g(x)$$

Deriviamo tutto rispetto al parametro a:

$$xg'(xa) = g'(a)\frac{g(x)}{g(1)}$$

Scegliamo ora $a = 1$:

$$xg'(x) = \frac{g'(1)}{g(1)}g(x)$$

$$xg'(x) = kg(x)$$

Questa equazione si risolve per separazione di variabili.

$$g(x) = cx^k \qquad k = \frac{g'(1)}{g(1)} \tag{1.13}$$

Abbiamo dimostrato che le uniche funzioni continue invarianti per scala sono le leggi di potenza.

1.4.1 Legge di Benford

Nell'ottocento si era osservato che la prima cifra del numero rappresentante la lunghezza di un fiume ha una distribuzione di probabilità precisa, data da:

$$P(n) \propto \log \frac{n+1}{n} \tag{1.14}$$

Dove n è il valore della cifra.

La cosa interessante è che questa distribuzione non dipende dall'unità di misura utilizzata per esprimere la lunghezza del fiume.

Questo fenomeno prende il nome di *legge di Benford*. La stessa regola vale in molti altri campi, come ad esempio la dimensione dei crateri sulla Luna, o il prezzo delle azioni in borsa (indipendentemente dalla valuta).

Da questa considerazione possiamo ricavare informazioni importanti sulla densità di probabilità $\rho(M)$ dove M esprime l'intero numero che stiamo analizzando (la lunghezza dei fiumi o il prezzo dell'azione).

Possiamo scrivere la funzione di probabilità $P(n)$ integrando la $\rho(M)$ su tutti i numeri in cui l'ultima cifra è fissata.

$$n \cdot 10^k < x < (n+1) \cdot 10^k$$

Dove k esprime il numero totale di cifre che compongono M (se è decimale, k può essere negativo). La probabilità che la prima cifra del numero sia n vale:

$$P(n) = \sum_k \int_{n10^k}^{(n+1)10^k} \rho(M)dM$$

Ora poiché il fenomeno non dipende dall'unità di misura usata, se cambiamo $M \to aM$, la probabilità deve rimanere la stessa:

$$\sum_k \int_{n10^k}^{(n+1)10^k} \rho(M)dM = \sum_k \int_{n10^k}^{(n+1)10^k} \rho(aM)d(aM)$$

Una condizione sufficiente perché i due integrali siano uguali è che siano uguali gli integrandi

$$a\rho(aM) = \rho(M)$$
$$\rho(aM) = a^{-1}\rho(M)$$

Confrontando questa relazione con la definizione di fenomeno *scale-free* (equazioni 1.11) otteniamo una stima della $\rho(M)$. Dall'equazione (1.11) otteniamo $h(a)$:

$$h(a) = a^{-1}$$

Con l'equazione (1.12) ricaviamo l'espressione della $g(a)$:

$$g(a) = g(1)a^{-1} \qquad g'(a) = -g(1)a^{-2}$$

Da cui ricaviamo l'esponente k della legge di potenza (eq. 1.13):

$$k = \frac{g'(1)}{g(1)} = -1$$

Da cui ricaviamo la distribuzione della $\rho(M)$.

$$\rho(M) \sim M^{-1}$$

Questa distribuzione sembra *non-normalizzabile* (l'integrale diverge all'infinito). Tuttavia questi fenomeni hanno un cut-off[8] che tronca la funzione ρ oltre un certo integrale.

Possiamo a questo punto calcolare la probabilità iniziale $P(n)$ e dimostrare che rispetta la legge di Benford.

$$\frac{P(n)}{P(1)} = \frac{\sum_k \int_{n10^k}^{(n+1)10^k} M^{-1}dM}{\sum_k \int_{10^k}^{2\cdot 10^k} M^{-1}dM} = \frac{\sum_k \left(\ln \frac{n+1}{n}\right)}{\sum_k \log 2} = \frac{\log \frac{n+1}{n}}{\log 2}$$

Nell'ultimo passaggio abbiamo eliminato le due sommatorie, infatti gli argomenti delle sommatorie non dipendono da k (e possono essere portati fuori). Questo risultato coincide proprio con quanto osservato nell'ottocento (1.14).

Distribuzioni del tipo M^{-1} possono scaturire da processi moltiplicativi.

$$\sum_i x_i \sim Gaussiana \qquad \prod_i x_i = \sum_i \log x_i$$

$$\log\left(\prod_i x_i\right) \to Gaussiana \qquad x_i \sim \text{log normale}$$

[8]La lunghezza di un fiume è limitata dalla superficie terrestre, il diametro dei crateri dalla circonferenza lunare e il prezzo delle azioni dal capitale complessivo.

La funzione log normale è una simil gaussiana:

$$P(x) = \frac{1}{x} e^{-\frac{(\log x - \mu)^2}{2\sigma_l}} \tag{1.15}$$

Intorno al valore $x = e^{\mu}$ l'esponenziale è costante e la funzione che domina è l'andamento x^{-1}.

Nel caso del prezzo delle azioni questo comportamento è ben giustificato, perché le azioni fluttuano sui valori relativi[9].

Se prendiamo un segmento e iniziamo a tagliuzzarlo in tanti pezzettini a caso, la lunghezza dei vari segmentini è distribuita con la log normale. I crateri sulla luna sono distribuiti log normale perché gli asteroidi si sono scontrati più volte frammentandosi come i segmentini.

1.4.2 Distribuzione log normale

La distribuzione log normale riveste particolare importanza nello studio dei sistemi complessi, perché è il limite a cui tende la distribuzione di probabilità del prodotto di tante variabili casuali:

$$\log x = \sum_{i=1}^{N} \log y_i$$

Per tante variabili y_i vale il teorema del limite centrale:

$$\rho(\log x) \to e^{-\frac{(\log x - N\mu_l)^2}{2\sigma_l^2 N}}$$

Dove

$$\langle \log y \rangle = \mu_l \qquad \langle \log^2 y \rangle - \mu_l^2 = \sigma_l^2$$

Calcoliamo la distribuzione delle x

$$\tilde{\rho}(x)dx = \rho(\log x) d\log x$$

$$\tilde{\rho}(x) = \rho(\log x)\frac{1}{x}$$

Da cui si ottiene la distribuzione log normale:

$$\tilde{\rho}(x) = \frac{1}{x} e^{-\frac{(\log x - \mu_l N)^2}{2\sigma_l^2 N}}$$

$$\log \tilde{\rho} = -\log x - (\log x - \mu_l N)^2/(2\sigma_l N) + \log C$$

A priori questa distribuzine sembra molto diversa da una polinomiale, infatti il grafico presenta una deviazione quadratica (Figura 1.4).

Tuttaia il termine quadratico si riesce ad apprezzare solo se varia in questo range:

$$0 < \frac{(\log x - N\mu_l)^2}{2\sigma_l^2 N} < 1$$

$$N\mu_l < \log x < N\mu_l + \sqrt{2\sigma^2 N}$$

$$1 < x e^{N\mu_l} < e^{\sqrt{2\sigma_l^2 N}}$$

Se $N \sim 100$ $\sigma_l^2 \sim 1$, questo oggetto deve variare su sei ordine di grandezza prima di poter apprezzare la curvatura sul grafico doppiolog (e quindi distinguerlo dalla legge di potenza).

[9] Il valore delle azioni si aggiorna moltiplicandovi una variabile casuale.

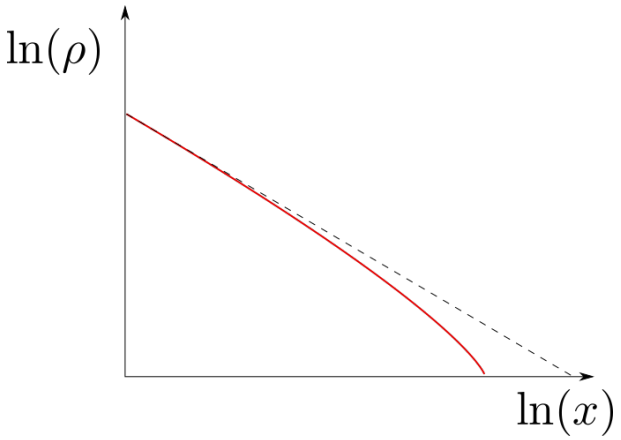


Figura 1.4: Grafico doppio-logaritmico della distribuzione log normale. In principio è molto differente rispetto a quella di una distribuzione polinomiale, tuttavia il termine di curvatura è apprezzabile solo dopo molti ordini di grandezza.

1.5 Variabili nascoste nelle leggi di potenza

Spesso le leggi di potenza possono essere ricondotte a leggi più fondamentali, caratterizzate da variabili che si distribuiscono in modo gaussiano o esponenziale, attraverso il riconoscimento di opportune variabili nascoste nel problema.

Immaginiamo di avere una variabile x con distribuzione $\rho(x)$, in modo che esista $\rho(0)$. Come è distribuita la sua inversa?

$$y = \frac{1}{x} \qquad \tilde{\rho}(y)dy = \rho(x)dx$$

$$\tilde{\rho}(y) = \frac{\rho(x)}{\left|\frac{dy}{dx}\right|} = \frac{\rho\left(\frac{1}{y}\right)}{y^2}$$

$$\tilde{\rho}(y) = y^{-2}\rho\left(\frac{1}{y}\right)$$

La cosa interessante è che per quando y va all'infinito la $\rho(1/y)$ tende ad una costante e quindi la y va come una legge di potenza di esponente -2.

Un esempio tipico è il sistema di Ising paramagnetico. In questa fase la magnetizzazione spontanea è circa nulla, se misuriamo le fluttuazioni relative di magnetizzazione:

$$\frac{\Delta m}{m} = \frac{m_t - m_{t-T}}{m_t + m_{t+T}} 2$$

Otteniamo un processo distribuito asintoticamente con una legge di potenza di esponente $\alpha = -2$.

C'è un altro processo matematico interessante. Se abbiamo una variabile x che è distribuita esponenzialmente, e abbiamo una y legata alla x con un altra costante:

$$\rho(x) = c_1 e^{-ax} \qquad y = c_2 e^{bx}$$

Possiamo chiederci come è distribuita la y.

$$\tilde{\rho}(y) = \frac{\rho(x)}{c_2 b e^{bx}} = \frac{c_1 e^{-ax}}{by} = k y^{-1} e^{-ax}$$

$$\tilde{\rho}(y) = k y^{-1} e^{-bx\frac{a}{b}} = k y^{-1-\frac{a}{b}}$$

Quindi anche in questo caso abbiamo una legge di potenza. Da un andamento a legge di potenza è possibile estrarre quindi una variabile casuale x distribuita esponenzialmente. Queste sono i modelli a variabile nascosta che consentono di passare da distribuzioni a legge di potenza a distribuzioni esponenziale.

1.6 Modello di Yule-Simon

Il modello di Yule-Simon cerca di spiegare l'insorgere la legge di Zipf nella distribuzione delle parole nei testi. Immaginiamo di scrivere un testo aggiungendo una parola alla volta. Il numero di parole totali è indicato dal "tempo" t. Con una certa probabilità p inventiamo una parola completamente nuova. Con probabilità $1-p$ peschiamo nel mucchio di parole che abbiamo già scritto una parola a caso. Questo fenomeno produce una situazione detta *rich gets richer* ed è alla base dell'insorgere della legge a potenza. Se abbiamo una parola che si ripete più volte nel testo questa ha più probabilità di essere ripescata. La crescita del dizionario ovviamente è data dall'andamento:

$$D \sim pt$$

Questa non è una previsione consistente con la teoria perché nella realtà la crescita avviene un modo sublineare.

$$D \sim t^{\gamma} \qquad \gamma < 1$$

Quindi il modello non spiega la legge di Heaps.

Scriviamo la *master equation* del modello per vedere se almeno è spiegata la legge di Zipf. Facciamo l'ipotesi di passare dal discreto al continuo. Sia N_k il numero di parole occorse k volte, kN_k rappresenta quante parole occorse N_k volte sono presenti nel testo. Questa è una proprietà che dipende dal tempo. Il numero di parole occorse k volte nel testo aumenta se al passaggio successivo estraiamo una parola dal testo occorsa $k-1$ volte, e diminuisce se estraiamo una parola occorsa k volte (entrambi questi casi si verificano se aggiungiamo una parola già presente nel mucchio, con probabilità $1-p$). Se $k=1$, N_1 aumenta con probabilità p (estraiamo una nuova parola).

L'equazione che sintetizza questo processo è:

$$N_k(t+1) = N_k(t) + (1-p)\cdot \underbrace{\frac{(k-1)N_{k-1}(t)}{t}}_{\substack{\text{Prob. di estrarre una parola} \\ \text{occorsa } k-1 \text{ volte}}} - (1-p)\frac{kN_k(t)}{t} + p\delta_{k,1}$$

Passiamo al continuo.

$$N_k(t+1) - N_k(t) \approx \frac{\partial N}{\partial t}(t)$$

$$kN_k(t) - (k-1)N_{k-1}(t) \approx \frac{\partial(kN_k)}{\partial k}$$

La *master equation* nel caso continuo diventa:

$$\frac{\partial N_k}{\partial t} = -\frac{1-p}{t}\frac{\partial(kN_k)}{\partial k}$$

La δ sparisce se studiamo la soluzione per $k \neq 0$. Ipotizziamo il numero di parole che occorrono k volte nel testo diviso il numero di parole totali tendi ad un valore asintotico e costante per $t \to \infty$.

$$N_k(t) \overset{t\to\infty}{\longrightarrow} tp_k$$

Sostituendo questa espressione asintotica si ottiene:

$$\partial_t(tp_k) = -\frac{1-p}{t}\partial_k(ktp_k) = -(1-p)\partial_k(kp_k)$$

Come si vede l'equazione non ha più dipendenza esplicita dal tempo, questo è consistente con l'ipotesi di andamento asintontico indipendente da t.

$$p_k = -(1-p)\left[p_k + kp_k'\right]$$

$$p_k(2-p) = (p-1)kp_k'$$

$$kp_k' = -p_k \underbrace{\frac{2-p}{1-p}}_{\alpha}$$

$$k\frac{dp_k}{dk} = -\alpha p_k$$

$$\frac{dp_k}{p_k} = -\alpha\frac{dk}{k}$$

$$p_k \sim k^{-\alpha} \qquad \alpha = \frac{2-p}{1-p} = 1 + \frac{1}{1-p}$$

Abbiamo ritrovato una legge di tipo polinomiale (simile alla legge di Zipf). Questo esponente tuttavia non descrive realmente quello che avviene nelle lingue (infatti è facile dimostrare che, nel modello di Yale-Simon, $\alpha > 2$, mentre nell'inglese $\alpha < 2$). Quindi per il modello $\alpha > 2$. Se si prende l'inglese otteniamo che $\alpha < 2$. Mandelbroot criticò molto questo modello, a causa di questi pessimi risultati numerici, mentre Simon sostenne che il suo modello coglie il motivo per il quale nasce la legge di Zipf, pur mancando del dettaglio necessario a prevedere i giusti esponenti.

Se chiamiamo β dal frequency range

$$\beta = \frac{1}{\alpha - 1} = 1 - p$$

E il γ di Hips è pari a

$$\gamma = \frac{1}{1-p} > 1$$

Mentre la legge di Heaps è sublineare. Il modello di Simon può essere ulteriormente migliorato se peschiamo aggiungiamo nel mucchio di parole pescabili tutta una serie di nuove parole ogni volta che peschiamo una parola a casaccio.

1.7 Tempo di primo ritorno e funzioni generatrici

Analizziamo un altro modellino, il camminatore aleatorio (*random-walk*). Il camminatore può spostarsi in due posizioni per dimensione[10], e sceglie direzione e verso in cui spostarsi aleatoriamente, con probabilità uniforme. In una o due dimensioni è facile dimostrare che il camminatore, in un tempo finito, ritornerà all'origine. Siamo interessati al numero medio di passi che fa il camminatore prima di tornare all'origine per la prima volta. Questa grandezza è definita *tempo medio di primo ritorno*, e si distingue dal *tempo medio di ritorno totale* che rappresenta la probabilità di tornare all'origine dopo n passi (metà degli spostamenti da una parte e metà dall'altra, n deve essere pari).

Per calcolare u_{2n}, probabilità che dopo $2n$ passi il camminatore si trovi nell'origine usiamo la distribuzione binomiale.

$$u_{2n} = \binom{2n}{n} \left(\frac{1}{2}\right)^n \left(\frac{1}{2}\right)^n$$

Se $n \gg 1$ possiamo usare l'approssimazione di Stirling.

$$n! \approx n^n e^{-n} \sqrt{2\pi n} \tag{1.16}$$

$$u_{2n} = \frac{(2n)!}{n!n!} 2^{-2n} \approx \frac{(2n)^{2n} e^{-2n} \sqrt{4\pi n}}{(n^{2n} e^{-2n} 2\pi n)} 2^{-2n} \propto n^{-\frac{1}{2}} \tag{1.17}$$

Calcoliamo la probabiltià che ci sia un generico ritorno, indipendentemente da n:

$$p_{ritorno} \propto \sum_{n=1}^{\infty} n^{-\frac{1}{2}} = \infty$$

Aspettando un tempo sufficientemente lungo prima o poi il camminatore ritornerà all'origine. Anche nel caso bidimensionale questo è vero, infatti possiamo a ciascuna direzione due random-walk unidimensionali. La probabilità di far ritorno dopo $2n$ passi è pari alla probabilità che, simultaneamente, i due random-walk ritornino nell'origine:

$$u_{2n}^{d=2} \sim \left(n^{-\frac{1}{2}}\right)^2 = n^{-1}$$

Che diverge come la serie armonica. È facile dedurre che il random-walk in qualunque dimensione maggiore di 2 non abbia questa divergenza, per cui esiste una probabilità non nulla che il camminatore non torni più nell'origine.

Stimare il tempo di primo ritorno è più interessante. Un esempio di applicazione è legato ai modelli di estinzione delle categorie *tassonomiche*[11]. Possiamo

[10] Se il sistema è a una sola dimensione il camminatore può spostarsi solo verso destra o sinistra (2), se è in due dimensioni verso Nord, Sud, Est e Ovest (4), e così via: il numero di posizioni accessibili è pari al doppio delle dimensioni del sistema.

[11] Le categorie tassonomiche sono insieme di specie che si sono evolute da un progenitore comune. Nel tempo l'intera categoria acquista un numero di specie che possono aumentare o diminuire, tanto che il numero di specie presenti nella categoria è assimilabile alla posizione del nostro camminatore. Poiché il numero di specie non può mai essere negativa, appena $n = 0$ avviene l'estinzione totale, e il processo si arresta. Il tempo di primo ritorno corrisponde proprio al tempo di estinzione della categoria.

definire una funzione generatrice $U(z)$

$$U(z) = \sum_{n=0}^{\infty} u_{2n} z^n$$

La u_{2n} è data dalla (1.17). La serie può essere sommata:

$$U(z) = (1-z)^{-\frac{1}{2}}$$

Vogliamo legare la probabilità di un ritorno generico u_{2n} con quella di fare un primo ritorno f_{2n}. La probabilità di trovarci nell'origine dopo $2n$ passi può essere scritta come la probabilità di trovarci nell'origine dopo $2n-2m$ passi per la probabilità di fare un primo ritorno dopo $2m$ passi, sommando su tutti i possibili valori di m (Figura 1.5).

$$u_{2n} = \sum_{m \leq n} f_{2m} u_{2n-2m} \tag{1.18}$$

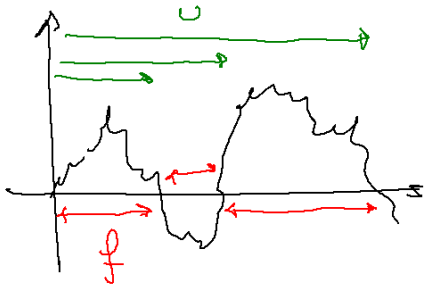


Figura 1.5: La u è la probabiltià di ritporno (in verde), mentre la f è la probabilità di primo ritorno.

Anche la probabilità di primo ritorno avrà una generatrice:

$$F(z) = \sum_{n=0}^{\infty} f_{2n} z^n$$

Usiamo la fattorizzazione (1.18) per trovare un legame tra le due funzioni generatrici:

$$U(z) = \sum_{n=0}^{\infty} u_{2n} z^n = 1 + \sum_{n \geq 1} \sum_{m \leq n} f_{2m} u_{2n-2m} z^n = 1 + \overbrace{\sum_{m \geq 1} f_{2m} z^m}^{F(z)} \underbrace{\sum_{n \geq m} u_{2n-2m} z^{n-m}}_{U(z)}$$

$$U(z) = 1 + F(z)U(z)$$

$$F(z) = 1 - \frac{1}{U(z)} = 1 - \sqrt{1-z}$$

Possiamo notare che, apparte costanti moltiplicative, vale la seguente relazione:

$$F(z) = 1 - \sqrt{1-z} \propto \int (1-z)^{-\frac{1}{2}} dz = \int U(z) dz$$

Da cui è facile ricavare un espressione per la f_{2n} (a meno di costanti moltiplicative)

$$F(z) = \int U(z) dz = \int \sum u_{2n} z^n dz = \sum \frac{u_{2n}}{n+1} z^{n+1}$$

Possiamo dire che

$$f_{2n} \approx \frac{u_{2n}}{n} \overset{n\to\infty}{\longrightarrow} n^{-\frac{3}{2}}$$

$$f_{2n} \propto n^{-\alpha} \qquad \alpha = 1.5$$

Questo esponente può essere ottenuto sperimentalmente per i processi di estinsione in categorie tassonomiche:

$$\alpha = 1.7 \pm 0.3$$

Che è compatibile con il modello del random walk.

1.8 Percolazione

La percolazione è un fenomeno di transizione di fase che studia la formazione di cluster in un sistema. Immaginiamo di avere un reticolo, in cui ciascun sito può essere acceso o spento. Un cluster rappresenta un insieme di siti, tra loro primi vicini, tutti accesi.

Possiamo stimare la dimensione media del cluster in funzione in funzione della probabilità che un sito sia acceso o spento. Esiste un valore critico P_c tale che la dimensione del cluster diverge (Figura 1.6)

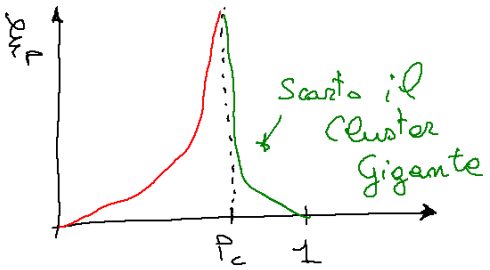


Figura 1.6: Grafico della dimensione dei cluster in funzione della probabilità di avere un sito attivo. Esiste un punto critico P_c in cui la dimensione del cluster più grande diverge.

La lunghezza tipica del cluster ξ_p si avvicina al punto critico con una legge di potenza di questo tipo:

$$\xi_p \sim |p - p_c|^{-\nu}$$

La cosa interessante è che l'esponente ν non dipende dal tipo di reticolo ma solo dalla dimensione.

Possiamo studiare il problema della percolazione studiando l'invarianza di scala al punto critico. Qualunque trasformazione di scala (raggruppamento di cluster) deve mantenere invariata la probabilità P_c. L'operazione di riscalamento avviene nel seguente modo:

$$p' = R_l(p) \tag{1.19}$$

Nel caso unidimensionale un esempio di raggruppamento di scala è mostrato in Figura 1.8.

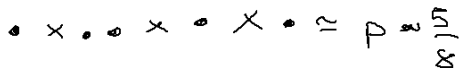


Figura 1.7: Configurazione di siti unidimensionale.

Nel caso unidimensionale il punto critico è con $p_c = 1$, tutti i siti accesi[12], l'invarianza di scala è garantita solo se tutti i siti sono nello stesso stato (accesi o spenti). La percolazione avviene nel sistema che massimizza la dimensione del cluster.

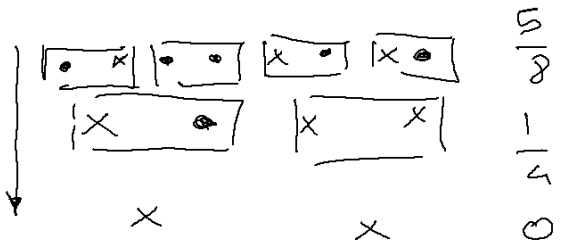


Figura 1.8: Trasformazioni di scala in sistemi unidimensionali.

A seconda di come cambiuamo scala la p cambia. Il punto critico è quello che garantisce invarianza di scala.

$$p_c = R_l(p_c)$$

[12]Basta un solo sito spento perché il sistema perda l'invarianza di scala.

Il punto critico è una soluzione istabile dell'equazione[13].

Possiamo definire la R_l indipendentemente dalla scala l. Un blocco che percola deve avere tutti i suoi siti accesi. La probabilità che un blocchetto percoli alla nuova scala è quella che tutti i siti da cui è composto siano accesi:

$$R_l(p) = p^l$$

Imponiamo l'invarianza di scala per il punto critico.

$$p_c^l = p_c \qquad \forall l > 0$$

Che ha per soluzioni:

$$p_c = 0 \qquad p_c = 1$$

Se grafichiamo questa equazione mettendo p^l sull'asse delle y e p su quella delle x otteniamo che il punto istabile si ha per $p = 1$ (Figura 1.9).

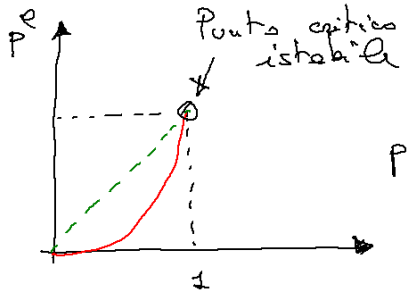


Figura 1.9: Schema delle soluzioni per il caso unidimensionale.

Vediamo come funziona l'avvicinamento al punto critico, sviluppiamo l'equazione (1.19) attorno a p_c

$$p' - p_c \approx R_l'(p_c)(p - p_c)$$

Prendiamo il valore assoluto:

$$|p' - p_c| \approx l|p - p_c|$$

Eleviamo alla ν e passiamo alla dimensione del cluster.

$$\xi' \approx l^{-\nu}\xi$$

Ma ξ è una grandezza del sistema, in una dimensione si riduce di un fattore:

$$\xi' = l^{-1}\xi$$

Per confronto troviamo:

$$\nu = 1$$

Possiamo estendere questo ragionamento al caso bidimensionale con qualche approssimazione. Immaginiamo di avere un reticolo triagnolare (Figura 1.10).

Figura 1.10: Reticolo triangolare.

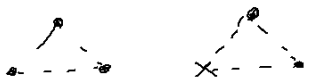


Figura 1.11: Schema della percolazione per un reticolo triangolare.

In questo caso per fare un cambio scala il sistema percola se abbiamo trangoli con due siti attivi (Figura 1.11)

La probabilità che questo si verifichi per un cambiamento di scala è pari a:

$$R_l(p) = p^3 + 3p^2(1-p) = p$$

Graficando questa equazione si ottiene la Figura 1.12.

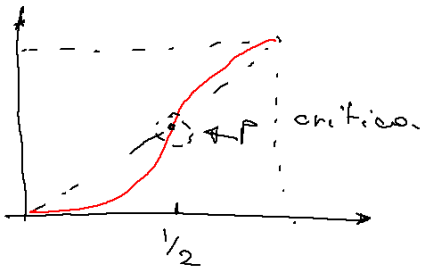


Figura 1.12: Schema delle soluzioni per un reticolo triangolare.

Ripetendo il calcolo della ξ si arriva ad un eseponente che assomiglia al valore sperimentale circa pari a 4/3. Non si ottiene a 4/3 perché non abbiamo considerato tutti i possibili riscalamenti, solo uno particolare.

[13] L'invarianza di scala è una proprietà instabile del sistema nel punto critico.

Capitolo 2

Teoria dell'informazione

In questo capitolo vogliamo sviluppare gli strumenti di teoria dell'informazione per arrivare a definire rigorosamente il significato dell'attributo "complessità", introducendo una misura matematica della complessità di un sistema.

2.1 Definizione termodinamica dell'entropia

L'entropia è una funzione di stato termodinamica. La differenza di entropia tra due stati A e B è definita come l'integrale di Clausius lungo una qualunque trasformazione *reversibile*. Se l'integrale è fatto lungo una trasformazione irreversibile si ottiene la disuguaglianza di Clausius:

$$S(B) - S(A) \geq \int_A^B \frac{\delta Q}{T}$$

È facile mostrare che l'entropia in un sistema isolato è una funzione crescente nel tempo, infatti $\delta Q = 0$ e:

$$S(B) - S(A) \geq 0$$

Per avere un'interpretazione fisica dell'entropia possiamo studiare l'espansione di un gas. Aumentiamo il volume di un gas perfetto da V_1 a V_2. L'espansione reversibile possiamo ottenerla attaccando un pistone ad una molla dura, che fa avvenire il processo molto lentamente (senza dissipazioni). Per semplicità immaginiamo che il gas sia costantemente in equilibrio termico con una sorgente, con la quale scambia calore (trasformazione isoterma):

$$dU = \delta Q - dW \qquad dU = 0$$

$$\delta Q = \delta W = nRT \ln \frac{V_2}{V_1}$$

Da cui si ottiene la variazione di entropia del gas:

$$\Delta S_{gas} = \frac{\Delta Q}{T} = nR \ln \frac{V_2}{V_1}$$

Cosa succede alla sorgente? La sorgente ha perso una quantità di calore pari a ΔQ, da cui la variazione di entropia della sorgente

$$\Delta S_{sorgente} = -\Delta S_{gas}$$

$$\Delta S_{tot} = 0$$

Il livello di disordine del sistema non sta aumentando, in questo caso abbiamo immagazzinato una densità di energia pari al lavoro ottenuto.

Completamente diverso è il caso di un espansione libera di un gas. In questo caso abbiamo un gas che si trova inizialmente in un volume V_1, aprendo una valvola facciamo defluire il gas all'interno di un volume V_2, mantenendo il sistema con pareti adiabatiche. Il sistema non fa lavoro e non scambia calore con l'esterno, e la temperatura finale rimane invariata. L'entropia è una trasformazione di stato, poiché lo stato finale del gas è lo stesso se avessimo fatto l'espansione isoterma reversibile, la sua variazione di entropia è

$$\Delta S_{gas} = \frac{\Delta Q}{T} = nR \ln \frac{V_2}{V_1}$$

Tuttavia qui non c'è scambio di calore con alcuna sorgente, e la variazione totale di entropia è positiva:

$$\Delta S_{tot} \geq 0$$

L'entropia del gas monoatomico perfetto può essere ricavata:

$$S(n, U, V) = nR \ln \left[\gamma_0 \left(\frac{U}{n} \right)^{\frac{3}{2}} \frac{V}{n} \right]$$

Questo può anche essere riscritto come

$$S(n, U, V) = Nk_b \ln \left[\gamma_1 \left(\frac{U}{N} \right)^{\frac{3}{2}} \frac{V}{N} \right]$$

Il temrine V/N è il volume disponibile per ogni particella, ossia la fluttuazione spaziale media di una particella, U/N è l'energia media per particella

$$\frac{V}{N} = \Delta x^3 \qquad U = \frac{\langle p^2 \rangle}{2m} = \frac{\langle p_x^2 \rangle + \langle p_y^2 \rangle + \langle p_z^2 \rangle}{2m} = \frac{3}{2m} \langle p_x^2 \rangle$$

$$\Delta p^2 = \langle p^2 \rangle - \langle p \rangle^2$$

Per sistemi isotropi $\langle p \rangle = 0$, il temrine $\langle p_z^2 \rangle$ rappresenta la fluttuazione tipica della particella. Se li risostituiamo si ottiene:

$$S = k_B N \ln \left[\gamma_2 (\Delta x)^3 (\Delta p)^3 \right] = k_b \ln \left[\frac{(\Delta x)(\Delta p)}{\gamma} \right]^{3N}$$

Il termine $\Delta x^3 \Delta p^3$ è lo spazio delle fasi disponibile per la singola particella. Se Γ è il volume nello spazio delle fasi disponibile a tutte le N particelle, l'entropia può essere riscritta in questo modo (il coefficiente γ è una costante additiva, poiché l'entropia è una funzione di stato può essere tranquillamente eliminato).

$$S = k_b \ln \Gamma(N, U, V)$$

Dove Γ è il volume dello spazio delle fasi, ossia il numero di configurazioni microscopiche corrispondono allo stato macroscopico del sistema.

$$\Gamma(N,U,V) = \int_{U-\frac{\delta U}{2} \leq H \leq U+\frac{\delta U}{2}} \frac{dx^{3N}dp^{3N}}{N!h^{3N}}$$

Possiamo definire qual è la probabilità che il sistema sia all'interno di un volume a $d\Gamma_N$

$$dP_{\Gamma_N} = \frac{d\Gamma_N}{\Gamma(N,U,V)}$$

Se consideriamo tutte le configurazioni equiprobabili (microcanonico) possiamo definire una densità di probabilità ρ:

$$\rho(\vec{x}_1, \cdots, \vec{x}_n, \vec{p}_1, \cdots, \vec{p}_n) = \frac{1}{\Gamma(U,V,N)}$$

Con questa definizione di ρ si può riscrivere l'entropia:

$$S = k_b \int dP_{\Gamma_N}(-\ln \rho) = -k_b \langle \ln \rho \rangle$$

Questa è la definizione di Shannon dell'entropia. È una definizione che rispetta il significato termodinamico, ma che può essere estrapolata a qualunque funzione di probabilità ρ

L'entropia può essere quindi definita usando la formula di Shannon, prendendo una generica distribuzione di probabilità ρ.

$$S = -k_b \langle \ln \rho \rangle$$

Consideriamo un sistema che si può trovare in un certo numero di stati, etichettabili con degli indici i. A ciascuno di questi stati può essere associata una probabilità di occorrenza:

$$p_i = \frac{e^{-E_i/k_bT}}{Z} \qquad Z = \sum_i e^{-\frac{E_i}{k_bT}}$$

Possiamo definire l'entropia di questa distribuzione:

$$\langle \ln p_i \rangle = \sum_i p_i \ln p_i = \langle \ln \left[\frac{e^{-\frac{E_i}{k_bT}}}{Z} \right] \rangle = -\frac{\langle E \rangle}{k_bT} - \ln Z$$

$$k_bT \langle \ln p_i \rangle = -\langle E \rangle \underbrace{-k_bT \ln Z}_{F}$$

La F è detta energia libera di Helmholtz. La definizione di energia libera è che

$$F = \langle E \rangle - TS$$

Questo è il legame tra entropia macroscopica ed entropia microscopica del sistema.

$$S = k_b \langle \ln p_i \rangle$$

Anche in questo ensamble. Si può partire attraverso principi di massimizzazione ritrovare le proprietà dell'Ensamble microcanonico o canonico.

L'ensamble microcanonico è un sistema in cui l'energia è fissata. Vogliamo massimizzare l'entropia, con un vincolo.

$$\sum_i p_i = 1$$

Usiamo la tecnica dei moltiplicatori di Lagrange, definiamo un funzionale

$$\Lambda = -k_b T \sum_i p_i \ln p_i + \lambda k_b \sum_i p_i$$

I due gradienti devono essere paralleli e opposti tra loro (la funzione non deve cambiare se ci muoviamo sul vincolo, per trovare il massimo vincolato i gradienti devono essere paralleli.

$$\frac{\partial \Lambda}{\partial p_i} = 0$$

Queste sono n equazioni

$$\frac{\partial \Lambda}{\partial p_i} = -k_b \ln p_i - k_b + \lambda k_b = 0$$

$$\ln p_i = \lambda - 1$$
$$p_i = e^{\lambda - 1}$$

Questo dimostra che la probabilità non dipende dallo stato del sistema. Dobiamo trovare il vincolo:

$$\sum_i p_i = 1 \qquad N e^{\lambda - 1} = 1$$

$$e^{\lambda - 1} = \frac{1}{N}$$
$$p_i = \frac{1}{N}$$

Abbiamo ritrovato la probabilità dell'ensamble microcanonico. Per trovare l'ensamble caninico dobbiamo imporre non solo che le probabilità siano normalizzate, ma che l'energia media sia fissato.

$$\sum_i p_i E_i = \langle E \rangle$$

$$\Lambda = -k_b T \sum_i p_i \ln p_i + \lambda k_b \sum_i p_i + \beta k_b \sum_i p_i E_i$$

$$\frac{\partial \Lambda}{\partial p_i} = -k_b \ln p_i - k_b + \lambda k_b + \beta k_b E_i = 0$$

$$p_i = e^{\lambda - 1} e^{\beta E_i}$$

Imponendo la normalizzazione

$$p_i = e^{\lambda - 1} \sum_{i=1}^{N} e^{\beta E_i} = 1$$

E la conservazione del valore medio dell'energia:

$$\langle E \rangle = e^{\lambda - 1} \sum_{i=1}^{N} E_i e^{\beta E_i}$$

$$e^{\lambda - 1} = \frac{1}{\sum_{i=1}^{N} e^{\beta E_i}}$$

Da cui si ricava che

$$p_i = \frac{e^{\beta E_i}}{\sum_i e^{\beta E_i}}$$

Dove β è una costante che definisce opportunamente l'energia media.

2.2 Definizione di Shannon dell'entropia

Immaginiamo di avere un sistema che si può trovare in due stati, con due rispettive probabilità p_1 e p_2

Vogliamo cercare una funzione (incertezza) che abbia come prorpietà

- Se $p_i = \delta_{ii^*}$, allora la funzione è nulla $S = 0$.
- Se $p_i = \frac{1}{N}$ allora abbiamo il massimo di incertezza.
- Funzione additiva

Immaginiamo di avere un sistema diviso in due sottosistemi, che possono trovarsi in N e M possibili stati. L'incertezza sul sistema globale deve essere la somma delle incertezze

$$S(N \cdot M) = S(N) + S(M)$$

L'unica funzione continua che ha questa proprietà è il logaritmo

$$S = k \log N$$

Proviamo ad applicare queste richieste in un sistema che può trovarsi in due stati p_1 e p_2. Possiamo ricavare una formula per l'entropia immaginando di avere N replice del sistema. Per la regola di additività l'incertezza del sistema grosso formato N repliche può essere scritta:

$$S_{TOT} = N S(p_1, p_2)$$

Per grandi N avremo Np_1 sistemi sullo stato 1, e Np_2 sistemi sullo stato 2. Il numero di possibili stati corrisponde al numero di combinazioni con Np_1 sistemi in 1 e Np_2 sistemi in 2.

$$\# = \frac{N!}{(Np_1)!(Np_2)!}$$

Da cui possiamo calcolare l'entropia dello stato totale:

$$S_{TOT} = k \log \left[\frac{N!}{(Np_1)!(Np_2)!} \right]$$

Usiamo lo sviluppo di Sterling

$$\log N! \approx N \log N - N$$

$$S = k\left[N \log N - N - Np_1 \log(Np_1) + Np_1 - Np_2 \log(Np_2) + Np_2\right]$$

$$S = k\left[N \log N - Np_1 \log N - Np_1 \log p_1 - Np_2 \log N - Np_2 \log p_2\right]$$

$$S = -kN\left[p_1 \log p_1 + p_2 \log p_2\right]$$

Da cui si ricava che l'entropia del singolo sistema può essere scritto come:

$$S(p_1, p_2) = -k\left[p_1 \log p_1 + p_2 \log p_2\right]$$

Questa relazione è possibile estenderla in più stati.

$$S(p_i) = -k \sum_i p_i \log p_i = -k \left\langle \log p_i \right\rangle$$

Questa corrisponde proprio alla definizione che avevamo dato in termodinamica, ora però l'entropia rappresenta l'incertezza dell'esito della produzione di una sorgente di informazione (possono essere caratteri, messagi o generici stati). Shanon era partito da un problema completamente differente, voleva descrivere matematicamente la comunicazione. L'essenza della comunicazione è divisa in fasi:

- Produzione del messaggio: **sorgente**
- **codifica** del messaggio (essenzialmente in codifica binaria).
- Invio del messaggio attraverso un **canale** di comunicazione (che può essere rumoroso o avere una certa capacità)
- **decodifica** del messaggio tramite un **destinatario**.

La comunicazione non è mai diretta. Si pensa ad un messaggio, si codifica il messaggio in un linguaggio, e poi attraverso un canale di comunicazione si propaga fino al destinatario, e deve avvenire la decondifica e comprensione del messaggio.

In questo capitolo studieremo soprattutto la sorgente, nell'ottica di quantificare la complessità del segnale emesso, e descriveremo brevemente la codifica (e la compressione) del segnale.

2.3 Entropia della variabile aleatoria

La variabile aleatoria X è identificata da una funzione di densità di probabilità $p(x)$ così definita:

$$P(X) = \{p_X(x), x \in \chi = (x_1, \cdots, x_M)\}$$

La X rappresenta la variabile aleatoria, la x sono i possibili valori che questa variabile può assumere, secondo la distribuzione $p(x)$. L'entropia di una variabile aleatoria è definita da

$$H(X) = -\sum_{x \in \chi} p(x) \log_2 p(x)$$

Se si usa il logaritmo in base due l'entropia della variabile casuale rappresenta il numero di bit necessari per descrivere la variabile.

$$H(X) = \langle -\log_2 p(x) \rangle$$

Questa entropia è un funzionale, non dipende direttamente dai valori che la variabile può assumere, ma da una funzione di questi valori (la distribuzione di probabilità). Per un sistema a due stati l'entropia è

$$H = -p \log_2 p - (1-p) \log_2 (1-p)$$

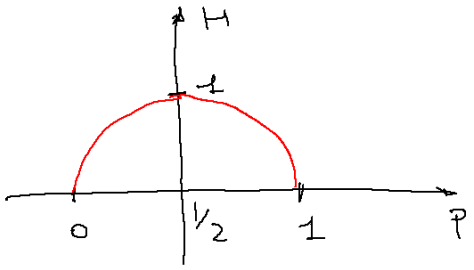


Figura 2.1: Schema dell'entropia per una variabile aleatoria con due possibili valori (0 o 1). L'entropia è nulla agli estremi (totale certezza) e massima quando la probabilità è esattamente un mezzo.

Immaginiamo di avere la sorgente, e diciamo che la sorgente può trovarsi in M possibili stati (un semaforo 3, una tastiera può trovarsi in un centinaio di stati differenti). Immaginiamo di volerlo codificare in un codice binario, con due stati possibili nel sistema. La sorgente emette una sequenza di stati.

$$S^N = \{S_1, S_2, S_3, \cdots S_N\}$$

Il numero di bit necessari per codificare la specifica sequenza S^N è

$$l(S^N) = \log_2 \frac{1}{P(S^N)}$$

È possibile costruire l'entropia a blocchi del sistema, ossia l'entropia della sorgente misurata su squenze di lunghezza N. Data una qualunque sequenza di N caratteri, calcoliamo il numero medio di bit necessari per codificarla:

$$\langle L \rangle = \sum_{\{S^N\}} l(S^n) P(S^n) = -\sum_{\{S^N\}} P(S^N) \log_2 P(S^N)$$

$$H_N\left(\{S^N\}\right) = -\sum_{\{S^N\}} P(S^N) \log_2 P(S^N)$$

L'entropia a blocchi rappresenta il numero di bit medio per codificare una generica sorgente di lunghezza N.

In teoria dell'informazione si suppone sempre che la sorgente abbia distribuzione di probabilità nota (o che emetta un infinità di messaggi). L'entropia è sempre misurata sulla sorgente che emette l'informazione, mai sulla singola sequenza.

In casi reali questo non è possibile, perché non conosciamo le probabilità della sorgente e abbiamo un numero limitato di messaggi. Possiamo al più cercare di inferire la sua distribuzione di probabilità. Qualche anno dopo le scoperte di Shannon è nata una nuova branca: la teoria della complessità algoritmica.

La complessità algoritmica si occupa di misurare la complessità della singola sequenza, ignorando il comportamento della sorgente. Complessità algoritmica e teoria dell'informazione si ricongiungono nel limite di sequenze infinitamente lunghe.

Possiamo definire anche un entropia differenziale:

$$h_N = H_{N+1} - H_N$$

Questa rappresenta è il numero di caratteri che vi serve per codificare l'extra-carattere. Sappiamo che

$$h_{N+1} \leq h_N$$

Aggiungere un bit non può far perdere informazione. Ora possiamo arrivare all'espressione completa dell'entropia di Shannon:

$$h = \lim_{N\to\infty} \frac{H_N}{N} = \lim_{N\to\infty} -\frac{1}{N} \sum_{\{S^N\}} P(S^n) \log_2 P(S^n) \tag{2.1}$$

Come si stima l'entropia di Shannon? Si può calcolare l'entropia a blocchi e estrapolare per $N \to \infty$, ma questo metodo è molto poco efficacie in quanto per $N > 7$ occorrono sequenze lunghissime per avere una stima credibile dell'entropia (si immagini di dover contare in un testo quante volte compare una data sequenza di più di sette lettere, sequenze del genere sono molto rare, e appariranno anche nei testi più lunghi al più una volta, da cui è molto difficile fare una stima accurata della probabilità effettiva di quella data sequenza).

I prototipi di sistemi a memoria finita sono le catene di Markov. Una catena di Markof emette tanti caratteri la cui probabilità dipende dagli m stati passati. In generale possiamo avere memoria variabile (ciascun carattere dipende dagli m caratteri precedenti).

La catena di Markof è definita da una matrice di transizione ($m = 2$):

$$A = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1M} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NM} \end{pmatrix}$$

La probabilità che capiti una data sequenza è:

$$P(S^N) = \mu_{s1} p_{S_1,S_2} p_{S_2,S_3} \cdots p_{S_{N-1,N}}$$

Dove μ è la probabilità di stare nello stato iniziale S_1. La probabilità di stare in uno stato iniziale σ è la probabilità di essere in τ allo stato precedente per la probabilità di transire da τ a σ, sommato su tutti i possibili valori dello stato τ di partenza:

$$\mu_\sigma = \sum_\tau \mu_\tau P_{\tau\sigma}$$

$$\vec{\mu} = A^T \vec{\mu}$$

Questa è un'equazione agli autovalori per la matrice di Markof (con autovalore 1). L'autovettore della matrice di Markof rappresenta le probabilità iniziali per ciascuno stato. Possiamo calcolare l'entropia di una catena di Markof. Partiamo dalla definizione a blocchi di entropia:

$$h_N = \sum_{S_1 \cdots S_{N-1}} P_{S_1} \cdots P_{S_{N-1}} \left[-\sum_{S_N} P(S_N|S_1 \cdots S_{N-1}) \log_2 P(S_N|S_1 \cdots S_{N-1}) \right]$$

Le ultime probabilità sono gli elementi della catena di Markof di ordine $N-1$. Immaginiamo che la memoria sia solo del primo ordine:

$$h_N = \sum_{S_{N-1}} P_{S_{N-1}} \left[-\sum_{S_N} P(S_N|S_{N-1}) \log_2 P(S_N|S_{N-1}) \right] = h_2$$

Questo risultato è del tutto generale, infatti una qualunque memoria finita di lunghezza m ha:

$$h_n = h_m \qquad \forall n > m$$

Quando abbiamo un sistema di una certa memoria si può dividere il sistema in N/m blocchi. In ciascuno di questi blocchi possiamo calcolare l'entropia H_m

$$H_N = H_m \frac{N}{m}$$

Quando si fanno i limiti

$$h_N = H_{N+1} - H_N = \frac{H_m}{m}(N+1) - \frac{H_m}{m} N = \frac{H_m}{m}$$

Per sistemi di memoria finita il limite per $N \to \infty$ può essere fatto prendendo $N > m$.

Teorema 2.3.1 (Shannon-McMillan-Breiman (1948-53)) *Siano $x_1, \cdots, x_n$ variabili indipendenti e indipendentemente distribuite.*

$$-\frac{1}{N} \log_2 p(x_1, \cdots, x_n) \overset{N\to\infty}{\longrightarrow} H(x) \text{ in probabilità}$$

Con

$$x_n \to x \text{ se } \forall \varepsilon > 0 \qquad \lim_{n\to\infty} P(|x_n - x| < \varepsilon) = 1$$

$$\lim_{N\to\infty} P\left(\left| -\frac{1}{N} \log_2 P(x_1, \cdots, x_n) - H(x) \right| < \varepsilon \right) = 1$$

Detto anche teorema AEP (Asymtotic Equipartition Property)

Il corollario di questo teorema è che esiste un insieme di "sequenze tipiche" A_ε^N . Queste sono tutte sequenze la cui probabilità è compreso tra due estremi

$$2^{-N(H(x)-\varepsilon)} < P(x \in A_\varepsilon^N) < 2^{-N(H(x)+\varepsilon)}$$

La probabilità di tutto l'insieme è dato da

$$1 - \varepsilon < p(\{A_\varepsilon^N\}) < 1$$

La cardinalità di queste squenze

$$(1-\varepsilon)2^{N(H(x)-\varepsilon)} \leq |A_\varepsilon^N| \leq 2^{N(H(x)+\varepsilon)}$$

Tutte queste squenze tipiche coprono l'insieme delle probabilità. L'insieme delle squenze che contano per il calcolo dell'entropia è un numero estremamente piccolo, infatti è dell'ordine di $2^{NH(x)}$ che è normalmente esponenzialmente più piccolo di 2^N (numero totale di sequenze).

Il teorema di Shannon in pratica afferma che il numero di bit necessari per codificare il singolo carattere di una sequenza tende all'entropia della sorgente per $N \to \infty$. Nel caso specifico di variabili scorrelate è estremamente semplice da dimostrare, anche se è stato successivamente dimostrato in casi più generali (in cui le variabili sono correlate).

Definizione 2.3.1 (Sequenza tipica) *Una sequenza si dice "tipica" se la sua probabilità di occorrenza è limitata da:*

$$2^{-N(H(x)+\varepsilon)} \leq P(x_1, \cdots, x_n) \leq e^{-N(H(x)-\varepsilon)}$$

Ossia, indicando con A_ε^N l'insieme delle sequenze tipiche a N caratteri con tolleranza ε, vale la seguente proprietà:

$$(x_1, \cdots, x_n) \in A_\varepsilon^N \qquad \Longrightarrow \qquad H(x) - \varepsilon \leq -\frac{1}{N}\log_2 P(x_1, \cdots, x_n) \leq H(x) + \varepsilon$$

Si può dimostrare (e lo faremo a breve) che la probabilità di estrarre una sequenza di N caratteri contenuta nell'insieme A_ε^N è inferiormente limitata:

$$P(A_\varepsilon^N) > 1 - \varepsilon \tag{2.2}$$

Ossia l'insieme A_ε^N ricopre, al diminuire di ε, l'intero spazio delle probabilità. Se prova a fare un conteggio delle sequenze si scopre che il numero di sequenze tipiche è limitato dalle disuguaglianze.

$$(1-\varepsilon)2^{N(H-\varepsilon)} \leq |A_\varepsilon^N| \leq 2^{N(H(x)+\varepsilon)} \tag{2.3}$$

In un alfabeto con K caratteri si ha un totale di K^N possibili sequenze, il valore massimo dell'entropia H si ottiene quando ciascun carattere dell'alfabeto è equiprobabile.

$$H_{max} = -\sum_{i=1}^{K} \frac{1}{K}\log_2 \frac{1}{K} = \log_2 K$$

Per una sorgente completamente caotica il numero di sequenze tipiche è pari all'intero insieme di sequenze possibili.

$$N_{eff} \sim 2^{NH(x)} = 2^{N\log_2 K} = K^N$$

Per una sorgente caotica tutte le sequenze sono tipiche. Tuttavia se la sorgente ha un certo ordine la situazione è molto diversa:

$$H(x) < \log_2 K$$

$$N_{eff} \sim 2^{NH(x)} \ll K^N$$

Il numero di sequenze tipiche è una frazione esponenzialmente più piccola rispetto al numero di tutte le possibili sequenze[1], tanto minore è l'entropia della sorgente, tanto più piccolo è l'insieme di sequenze tipiche che coprono l'intero output probabile della sorgente.

Dimostriamo la relazione (2.2). Il teorema AEP ci assicura che, scelto N sufficientemente grande, il logaritmo dell'inverso della probabilità della sequenza tende all'entropia. Scelta una soglia δ possiamo trovare una lunghezza specifica N_0 per cui tutte le sequenze di lunghezza maggiore di N_0 hanno una probabilità di essere tipiche maggiore di $1-\delta$:

$$\forall \delta > 0 \; \exists N_0 \; : \; N > N_0 \qquad \Rightarrow \qquad \underbrace{P\left(\left|-\frac{1}{N}\log_2 P(x_1, \cdots, x_n) - H(x)\right| < \varepsilon\right)}_{\text{Probabilità che la sequenza emessa dalla sorgente sia tipica}} > 1-\delta$$

Dimostriamo ora la cardinalità dell'insieme delle sequenze tipiche (2.3) Possiamo maggiorare il numero di sequenze tipiche:

$$1 = \sum_{\{s^n\}} P(s^n) \geq \sum_{\{s^N\} \in A_\varepsilon^N} P(s^N) \tag{2.4}$$

Dalla definizione di sequenza tipica si ottiene la seguente minorazione

$$P(s^n) \overset{s^n \in A_\varepsilon^n}{\geq} 2^{-N(H(x)+\varepsilon)}$$

Sostituendo nella (2.4) si ottiene:

$$1 \geq \sum_{\{s^N\} \in A_\varepsilon^N} 2^{-N(H(x)+\varepsilon)} = 2^{-N(H(x)+\varepsilon)} \overbrace{\sum_{\{s^N\} \in A_\varepsilon^N} 1}^{|A_\varepsilon^n|}$$

Che può essere riscritta dividendo per il prefattore esponenziale:

$$|A_\varepsilon^n| \leq 2^{N(H(x)+\varepsilon)}$$

L'altro pezzo della disequazione (2.3) può essere dimostrata analogamente. Abbiamo visto che per l'equazione (2.2) la probabilità di pescare una sequenza in A_ε^N è minorata da $1-\varepsilon$. Sfruttando l'altra minorazione della definizione di sequenza tipica possiamo riscrivere:

$$1 - \varepsilon < P(A_\varepsilon^N) \leq \sum_{\{s^N\} \in A_\varepsilon^N} 2^{-N(H(x)-\varepsilon)} = 2^{-N(H-\varepsilon)} |A_\varepsilon^N|$$

Da cui si ottiene l'ultima relazione.

$$(1-\varepsilon) 2^{N(H-\varepsilon)} < |A_\varepsilon^N|$$

Possiamo prendere l'insieme di tutte le sequenze, e all'interno di questo c'è un sottoinsieme delle sequenze tipiche.

[1] L'entropia si trova all'esponente, ridurre anche di poco un esponente causa una drastica riduzione del valore della funzione esponenziale.

Cerchiamo ora di fare una stima di quanti bit occorrono per codificare una sequenza tipica. Un metodo semplice e del tutto generale è quello di ordinare le sequenze (secondo un certo ordine, ad esempio lessilografico) fino ad arrivare alla cardinalità, e poi associare ciascuna sequenza tipica alla sua posizione. Il numero di Bit necessari è il logaritmo in base 2 della cardinalità più 1.

$$[N(H(x)+\varepsilon)]+1$$

Se la sequenza non è tipica A_ε in generale occorrono molti più bit:

$$N\log_2 K+1$$

Per codificare un messaggio, che può contenere o meno sequenze tipiche occorre aggiungere un bit davanti alla sequenza che individua se la sequenza è tipica o meno. Pertanto il numero di bit per codificare le sequenze sono:

$$[N(H(x)+\varepsilon)]+2 \quad \text{Sequenza tipica}$$

$$N\log_2 K+2 \qquad \text{Sequenza non tipica}$$

Quindi il valore aspettato della lunghezza in bit del messaggio è:

$$E[l(s^n)]=\sum_{\{s^n\}}P(s^n)l(s^n)=\sum_{\{s^n\}\in A_\varepsilon^N}P(s^n)\left[N(H+\varepsilon)+2\right]+\sum_{\{s^n\}\notin A_\varepsilon^n}P(s^n)\left[N\log_2 K+2\right]$$

Possiamo minorare questa quantità

$$E[l(s^n)]\leq N\left(H+\varepsilon\right)+2+\varepsilon\left[N\log_2 K+2\right]\leq NH(x)+N\varepsilon+2+2\varepsilon+\varepsilon N\log_2 K$$

Dove abbiamo usato le due probabilità che la sequenza sia o meno caratteristica.

$$E[l(s^n)]\leq N\underbrace{\left[\varepsilon+\frac{2+\varepsilon}{N}+\varepsilon\log_2 k\right]}_{\varepsilon'}+NH=N(H+\varepsilon')$$

2.4 Complessità algoritmica

La teoria della complessità algoritmica nasce per l'esigenza di descrivere l'informazione contenuta all'interno di sequenze date, senza conoscere la sorgente che le emette. Questo caso infatti è il più realistico, se si analizza la Divina Commedia di Dante Alighieri abbiamo una sequenza finita di caratteri, ma non conosciamo la sorgente dentro la mente del poeta fiorentino che ha potuto generare questo capolavoro.

Questa teoria fu sviluppata per la prima volta da i matematici russi Kolmogorov e Sinai a partire dagli anni 50 (poco dopo la formulazione di Shannon della teoria dell'informazione).

L'obbiettivo è misurare la complessità della singola sequenza.

La complessità algoritmica è definita (a chiacchere) come *la lunghezza espressa in bit del programma più corto che produce la sequenza in uscita e si ferma subito dopo.* Ad esempio la complessità della Divina Commedia è la lunghezza del più corto programma capace di produrre l'intero testo Dantesco in output.

Un modo banale può essere codificare l'intera Divina Commedia nel programma e fargliela stampare. Tuttavia questo metodo non è molto intelligente, e si può pensare di comprimere il messaggio utilizzando codifiche furbe per le parole più usate. La complessità algoritmica corrisponde alla grandezza del file più compresso che è possibile ottenere dalla sequenza originale (più il comando di stampa e di arresto).

È importante specificare che il programma si debba fermare: immaginiamo di scrivere tutti i possibili programmi ed eseguirli. Man mano che i programmi si fermano selezioniamo solo quelli che danno in output la sequenza voluta. Se tenessimo in considerazione programmi che non si fermano, non sapremmo mai se esiste un programma più corto (ma di esecuzione più lunga) di quelli che si sono già fermati capace di riprodurre la sequenza. In questo modo possiamo fissare un limite di esecuzione T, e definire la complessità algoritmica come la lunghezza del più corto programma che da in output la sequenza voluta e si arresta in un tempo $t \leq T$ nel limite $T \to \infty$ (l'aver ignorato programmi che non si arrestano mai assicura il comportamento sensato di questo limite).

Una definizione più formale della complessità algoritmica K è data dalla seguente espressione:

$$K_U(x) = min_{p:U(p)=x} l(p)$$

Dove U è il computer universale, capace di eseguire qualunque programma. È importante avere un computer universale per poter confrontare correttamente la lunghezza dei programmi. Gli algoritmi di compressione cercano di togliere la ridondanza delle sequenze, senza perdere informazione. Immaginiamo come comando stampare i primi N bit della radice di e. La sequenza di N bit così ottenuta può essere ottenuta quindi con un programma cortissimo.

Il computer universale ci si riferisce alle macchine di Turing. Il computer universale riesce a simulare qualunque altro computer. Un computer analogiche sfrutta un fenomeno fisico. Un computer analogico non è un computer universale, i computer digitali sono buone approssimazione. Quello che ha dimostrato Kolmogorof possiamo calolare è che la complessità algoritmica ottenuta per misurare un computer universale.

$$K_U(x) \leq K_A(x) + C_A$$

La complessità di un computer universale è sempre minore o uguale alla complessità algoritmica di un computer analogica più un pezzo C_A che dipende unicamente dal computer (ad esempio sistema operativo più linguaggio di programmazione). Se dividiamo per N

$$\frac{K_U}{N} \leq \frac{K_A}{N} + \frac{C_A}{N}$$

Se N è molto grande il pezzo aggiuntivo non c'è.

Inoltre notiamo che se la complessità algoritmica cresce con N con potenza minore di 1, per N molto grande la complessità per carattere va a zero. Le uniche sequenze davvero complesse sono quelle per cui

$$K \propto N$$

Qual è la complessità media di tutte le possibili sequenze con N caratteri (contenute in S^N)?

$$\lim_{N\to\infty} \frac{\langle S^N \rangle}{N} = \lim_{N\to\infty} \sum_{\{S^n\}} k_s^N P(S^n) = h$$

Il termine k_s^N rappresenta la complessità della sequenza s, che per $N \to \infty$ corrisponde al numero minimo di bit necessari per codificare la sequenza. Ma questa è proprio la definizione di Shannon di entropia.

2.5 Paradosso di Gödel

Il paradosso di Gödel asserisce che in ogni sistema formale esistono assersioni indecidibili.

Se proviamo a valutare la correttezza dell'affermazione "questa affermazione è falsa" entreremmo in un loop infinito: se l'affermazione fosse vera risulterebbe falsa, se fosse falsa risulterebbe vera.

Il problema contro cui lavoro Gödel era quello di riuscire a costruire un sistema formale da cui, a partire da assiomi, linguaggio, preposizioni e regole logiche, si potessero ottenere tutti i possibili teoremi con dimostrazioni.

Gödel tentò di smontare questa aspirazione positivistica di tardo 800 mostrando che all'interno di ogni sistema formale esistono assersioni indecidibili (la cui veridicidità non è dimostrabile all'interno del sistema stesso).

Immaginiamo di trovare una affermazione del tipo "questa proposizione non è dimostrabile"; se il sistema formale la reputa vera vuol dire che esistono proposizioni non dimostrabili o decidibili, se la reputa falsa vuol dire che è possibile dimostrare questa affermazione, e che quindi nel sistema esistono proposizioni non dimostrabili.

La consequenza diretta del teorema di Gödel è che la complessità algoritmica non è computabile.

Teorema 2.5.1 (Primo teorema di Gödel) *Ogni programma minimale è necessariamente casuale*

Il programma minimale p non può essere ulteriormente compresso. Supponiamo per assurdo che il programma minimale p possa essere compresso, e che esista un programma p' tale che $l(p')$ sia più piccola di p, che dia come output p. Questo è un assurdo perché posso produrre una sequenza x (output di p) attraverso il programma p':

- Eseguo p' (genero il programma p)
- Eseguo p (ho come output x)
- Arresto

In questo modo ho trovato un programma che può produrre x in uscita più corto di p (gli ultimi due passaggi non dipendonon dalla complessità di x, e la loro lunghezza può essere trascurata per grandi N), ma questo è impossibile perché p era il programma più corto esistente per produrre x.

Teorema 2.5.2 (Secondo teorema di Gödel) *In un sistema formale di complessità N è impossibile dimostrare che una particolare successione di cifre binarie ha complessità maggiore di $N + c$, dove c è una costante che dipende dal sistema formale.*

Non possiamo dimostrare che la complessità di un numero è più grossa di un sistema formale.

Consequenza diretta di questo è che la complessità algoritmica è un problema indecidibile, infatti non abbiamo un maggiorante "a priori" della generica sequenza, e quindi il nostro sistema formale non può dimostrare la complessità (è possibile che la complessità della sequenza sia maggiore della complessità del sistema formale).

Chiaramente il problema può essere decidibile per specifiche sequenze (posso maggiorarne la complessità trovando delle particolari compressioni della sequenza) ma non è possibile deciderlo *a prescindere* dalla sequenza in esame.

In altre parole, quando si cerca di computare la complessità di una sequenza con un sistema formale, non sappiamo mai se è possibile farlo prima di averci provato (e aver trovato un maggiorante più piccolo della dimensione del sistema).

Questo teorema può essere compreso attraverso affermazioni di tipo logico: *Trova il primo numero positivo che può essere provato che non può essere specificato da un programma con massimo di 22 parole.*

Questa affermazione rientra nel sistema formale con un massimo di 22 parole (sono 22 parole esatte), e sta chiedendo di calcolare un numero che non rientra in questo sistema formale (poiché ha complessità maggiore di 22 parole). Una tale affermazione è fattibile all'interno del sistema formale (può essere però compiuta se si toglie il vincolo delle 22 parole, ossia allargando il sistema formale). Questo esempio mostra come il vincolo sul sistema formale può rendere indecidibile alcune preposizioni. E il calcolo della complessità algoritmica è sempre reso indecidibile qualora il sistema formale è ristretto a un numero di bit inferiore della complessità del numero da calcolare (per lo stesso motivo per cui è indecidibile l'affermazione di cui sopra).

Allo stesso modo un esempio di affermazione indecidibile è se un programma è in grado di fermarsi o meno (certamente sul singolo programma possiamo analizzarlo e vedere se si ferma o ha dei loop infiniti, tuttavia non esiste un programma che, avendo un codice in input, è in grado di stabilire se quel codice si fermerà o meno).

2.6 Entropia relativa

Supponiamo di avere una sorgente A di bit 0 e 1 scorrelati tra loro. L'entropia della sorgente è

$$h_A = -p_a \log_2 p_a - (1 - p_a) \log_2 (1 - p_a)$$

Se conosciamo la sorgente A il numero minimo di bit per codificarla è questa. Esisterà la codifica migliore anche per B

$$h_b = -p_b \log_2 p_b - (1 - p_b) \log_2 (1 - p_b)$$

Supponiamo di usare per A la migliore codifica di B.

$$-p_a \log_2 p_b - (1 - p_a) \log_2 (1 - p_b)$$

Stiamo codificando con B il codice emesso da A. Si dimostra facilmente che questo è maggiore di h_A.

Questo oggetto è definita come la *cross-entropy*.

La cross entropy di B su A o A su B è differente. Quanti bit stiamo sprecando per usare la codifica sbagliata?

$$d(A|B) = -p_a \log_2 p_b - (1-p_a)\log_2(1-p_b) - [-p_a \log_2 p_a - (1-p_a)\log_2(1-p_a)]$$

$$d(A|B) = p_a \log_2 \frac{p_a}{p_b} + (1-p_a)\log_2 \frac{1-p_a}{1-p_b}$$

La $d_(A|B)$ è definita entropia relativa.

Stiamo misurando una sorta di similarità tra le due sorgenti. Non è una vera distanza perché non è simmetrica e non soddisfa la disuguaglianza triangolare.

La sorgente A e B sono due lingue. Questo mi sta dicendo quando sono simili italiano e inglese: se conosciamo l'inglese quanto è facile apprendere l'italiano, quando conosciamo l'italiano quanto è facile apprendere l'inglese (che non sono uguali).

Queste distanze sono informatiche. Questo si può generalizzare. Possiamo definire la *cross-entropy* a blocchi:

$$\hat{H}_N(A|B) = -\sum_{\{s^N\}} p_a(s^N) \log_2 p_b(s^n)$$

Da cui possiamo definire la vera cross-entropy:

$$\tilde{h}(A|B) = \lim_{N\to\infty} \frac{1}{N} \tilde{H}_N(A|B)$$

L'entropia relativa è anche definita come la divergenza di Kullback-Leibler.

$$D_N(A|B) = -\sum_{\{s^N\}} p_a(s^n) \log_2 \frac{p_b(s^N)}{p_a(s^n)}$$

$$d(A|B) = \lim_{N\to\infty} D_N(A|B)$$

Bisogna fare attensione a che le probabilità delle sequenze non siano nulle.

Possiamo definire la mutua informazionetra due variabili stocastiche X e Y:

$$I(X:Y) = \sum_{x\in X\ y\in Y} P(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

La mutua informazione ci definisce quanto una variabile ci dice rispetto all'altra. Questo è l'entropia relativa fra la probabilità congiunta delle due variabili e le probabilità marginale. Se i due oggetti sono indipendenti allora la mutua informazione è zero.

$$I(X:Y) = D\left(P(x,y)|P(x)P(y)\right)$$

Questa può essere riscritta come

$$p(x,y) = p(x|y)p(y)$$

$$I(X:Y) = \sum_{x,y} P(x,y) \log_2 \frac{P(x|y)}{p(x)} = -H(x|y) + H(x) = -H(y|x) + H(y)$$

L'entropia di x è essenzialmente tutto quello che so di x condizionato a y più la mutua informazione.

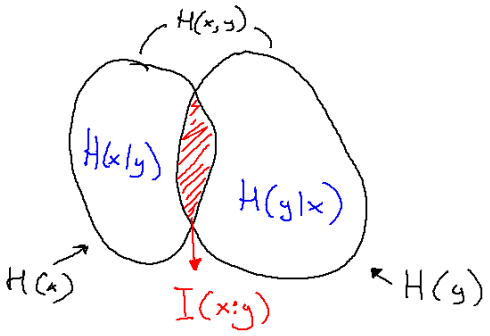


Figura 2.2: Schema dell'entropia tra due variabili.

2.7 Misure di entropia

Il problema che ci poniamno adesso è quello di misurare l'entropia o la complessità di una data seguenza. Una prima idea per calcolare l'entropia è quella di definire un entropia a blocchi.

$$H_N = -\sum_{\{s^n\}} P(s^n) \log_2 p(s^n)$$

Possiamo prendere sottosequenze della sequenza M e vediamo di contare tutte le volte in cui appare la sequenza 001. Possiamo stimare tutte le volte che vediamo la generica sequenza:

$$\hat{p}_N(s^n) = \frac{n(s_1 \cdots s_n)}{M - N}$$

Questo mi da una stima della probabilità di occorrenza. Per la legge di grandi numeri questa per $M \to \infty$. Questi numeri non sono grandi (M al massimo di tutti i testi digitalizzati è pari a 10^8 caratteri), il problema fondamentale è che supponiamo di avere una sequenza lunga, se la probabilità è così alta che io trovo la sequenza una volta sola non so quanto arrivo. Qual è la probabilità che occorra una sequenza di lunghezza N se abbiamo un alfabeto casuale di K caratteri? il limite lo abbiamo quando:

$$K^N \approx M$$

Facciamo l'esempio di un testo. $K = 26$ in inglese, $M \sim 10^8$, qual è il massimo valore di N?

$$N = \frac{\log M}{\log K} = \frac{8}{1.3} \approx 7$$

Riusciamo a capire soltanto le correlazione delle parole e sfuggono completamente le correlazioni all'interno delle frasi, paragrafi e sezioni.

Il fatto di fare scommesse e vincere le scommesse possiamo avere una misura della complessità del fenomeno. Immaginiamo di voler indovinare al tempo $t+1$.

$$S_{t+1} = u_t S_t$$

Dove u_t è una frazione.

$$S_{t+1} = u_t u_{t-1} S_{t-1}$$

La strategia delle persone è avere un certo capitale e investirlo in ogni istante in una certa azione. Si può decidere che frazione investire del capitale, quello che non si conosce sono queste u_t. Ad ogni istante non investo una parte del capitale (l_t). I miei soldi al tempo $t+1$ è

$$S_{t+1} = (1 - l_t)S_t + l_t S_t u_t = [1 + (u_t - 1)l_t]S_t$$

Il problma dello scommettiutore è decidere ogni istante il valore di l_t.
Il prezzo al tempo t di un azione è dato da:

$$S_t = S_0 e^{\lambda t}$$

Dove

$$\lambda = \lim_{t\to\infty} \frac{1}{t} \ln \frac{S_t}{S_0} = \langle \log [1 + (u-1)l] \rangle$$

Possiamo adesso reinterpretare la media sulle probabilità che occorra ciascuno singolo u.

$$\lambda = \sum_{i=1}^{m} \log [1 + (u_i - 1)l]\, p_i$$

Supponiamo di avere una tattica di investimento costante (l uguale ad ogni tempo). Se facciamo l'ottimizzazione del processo (ottimizziamo l).

$$\lambda(l^*) = \sum_{i=1}^{m} p_i \ln [1 + (u_i - 1)l^*] = \sum_{i=1}^{N} p_i \log \frac{p_i}{q_i}$$

$$q_i = \frac{p_i}{1 + (u_i - 1)l^*}$$

Facciamo un caso semplice il caso in cui $m = 2$.

$$u_1 - 1 = u_1 - u_2 = \alpha$$

$$\lambda(l) = p_1 \ln(1 + \alpha l) + p_2 \log(1 - \alpha l)$$
$$\lambda(l) = p \ln(1 + \alpha l) + (1 - p) \log(1 - \alpha l)$$

De vogliamo trovare la l ottimale occorre massimizzare λ

$$\frac{d\lambda}{dl} = \frac{p}{1 + \alpha l}\alpha - \frac{1 - p}{1 - \alpha l}\alpha = 0$$

$$p - p\alpha l^* = 1 + \alpha l^* - p - p\alpha l^*$$
$$2p - 1 = \alpha l^*$$
$$l^* = \frac{2p - 1}{\alpha}$$

Adesso

$$\lambda(l^*) = p \ln(2p) + (1 - p) \ln [2(1 - p)]$$
$$\lambda(l^*) = \ln 2 - h$$

Per questo processo il meglio che posso fare è avere un rate di crescita.

Se il processo è completamente casuale non possiamo guadagnare. Se non abbiamo informazioni sul sistema, d'altra parte se esistono delle correlazioni all'interno del sistema. Si può stimare l'entropia di un sistema stimando il guadagno massimo ottenuto da uno scommettitore (sarà quello che si è avvicinato maggiormente all'entropia del sistema).

Immaginiamo di avere un testo e voler scommettere sul carattere successivo. Bisogna investire tutte le lettere.

Possiamo definire delle grandezze che si chiamano.

$$\sum_{x_{k+1}} q(x_{k+1}|x_k \cdots x_1) = 1$$

Ad ogni istante il massimo che possiamo guadagnare è l'alfabeto $\tilde{k}$ volte il capitale puntato sulla lettera vincente

$$S_{k+1} = \tilde{k} q(x_{k+1}|x_1 \cdots x_k) S_k$$

Questo processo è noto come *gambling proporzionale*.Si può dimostrare ceh

$$\left(\tilde{k} - \langle \log_d S_k \rangle\right) \log_2 d \geq K_A(x_1 \cdots x_k)$$

Dove K_A è la complessità algoritmica. C'è un parallelo molto diretto tra la complessità algoritmica di una sequenza e la capacità di fare previsione. Se si fanno previsioni su quello che accadrà un'attimo dopo, tanto più siamo bravi e abbiamo capito come funziona il fenomeno.

Con schemi di questo tipo è dell'ordine dei 2.3 (molto minore degli 8 di codifica ascii).

2.7.1 Misure attraverso gli algoritmi di compressioni

Il problema della compressione dei dati c'è una letteratura sconfinata dal punto di vista delle applicazioni. La compressioni ha due grandi branche, la compressione senza perdita ed esistono meccanismi con perdita (che sono dietro mp3) in cui si tagliano una quantità di informazioni che non sono percepibile dall'essere umano. Se vogliamo utilizzare i compressore come misuratore di complessità. La complessità algoritmica fissa il numero di bit minimo, qualuqnue compressione ci da una stima maggiorando la complessità algoritmica.

I compressori possono essere statistici. Questi compressori presumono di conoscere la statistica della sorgente. Un esempio è il codice Morse. Prende il carattere dell'alfabeto inglese e li trasforma in una sequenza. Il carattere più diffuso è codificato con meno lettere. Il compressore è statistico perché si usa una conoscenza a priori della sorgente. Immaginiamo di avere sequenze particolari. Esistono dei compressori che si adattano alla sequenza. Cerchiamo di imparare dalla sequenza stessa la statistica.

Un algoritmo di questo tipo è l'algoritmo di Lemper-Ziv (LZ77). Questo algoritmo sta alla base di *gzip*.

Immaginiamo di avere una sequenza

$$qwhhABCDhhABCDcABCDhhz$$

L'algoritmo inizia a leggere la sequenza all'inizio e poi si sposta verso destra. L'algoritmo guarda a vanti e cerca la sequenza più lunga che è già occorsa nel passato. L'algoritmo sostituisce al posto di questa sequenza due numeri (quanto è lunga la sequenza e quanto indietro bisogna andare per trovare la stessa occorrenza. Il compressore sostituisce un puntatore al posto della sequenza. Si sta cercando di imparare dal sistema ricorrenze per le stesse sequenze. Potremmo cercare se la sequenza è palindroma o meno, oppure spesso in biologia. Esiste un teorema che dimostra che se questa sequenza sequenza fosse infinita, la stima che l'algoritmo fa della complessità converge esattamente alla complessità.

$$\lim_{n\to\infty} \frac{l(\text{testo compresso})}{N} = h$$

Questo è dovuto al fatto che esistono le sequenze tipiche nel ssitema, e codificado le sequenze tipiche il valore tende ad h. Il numero di bit per codificare la sequenza è fatto da due pezzi: la lunghezza della sequenza è dell'ordine di n, di quanto devo andare indietro (al massimo N) Di quanti bit abbiamo bisogno per codificarli?

$$n = \frac{\log_n N}{h}$$

$$l(\text{testo compresso}) = \frac{\#_{bits} - sequenza}{n} = \frac{\log_2 n + \log_2 N + O(\log\log n)}{\log_2 \frac{N}{h}} = h + \frac{\log\log N}{\log N} + \cdots$$

Questo algoritmo converge all'entropia di shannon della sorgente che ha emesso quella sorgente.

Capitolo 3

Reti complesse

Il campo delle reti complesse è molto legato al tempo in cui si è sviluppato. Il fatto che si sia sviluppato in questi anni non è casuali, l'avento di internet ha stravolto questo modo di pensare.

Una rete è un insieme i nodi collegati da *link*. I collegamenti possono essere direzionali: un link mi connette il nodo A al nodo B senza che il nodo B sia connesso ad A. Un esempio è il link diretto.

Il problema delle reti è stato studiato inizialmente nell'ambito della sociologia. Questo studio nasce negli anni 50 quando i sociologi iniziano a voler studiare le comunità di persone. Si inizia a stuiare come si uniscano le persone. I sociologi facevano ricerche andando da delle persone a chiedere direttamente le loro amicizie. Le reti sociali che si avevano erano molto piccole. Anche con reti sociali povere erano venute fuori cose interessanti. Un effetto particolare era l'effetto "small world". Per trovare un percorso che prendono due nodi anche molto lontani servono pochi passi. È importante che ci sia un effetto di questo tipo, perché le informazioni fluiscono molto facilmente. Reti grandi hanno un diametro molto piccolo. La cosa fu sperimentata nel 1967, quando fu fatto un esperimento molto famose. Dette un certo numero di lettere di carta dando a tutti un indirizzo a cui consegnalre. Quando le raccolse a Boston vide che in media queste persone ci mettevano 5 passi. La rete degli statuiniti aveva un diametro di circa 5.

Il diametro di una rete è la distanza minima che lega due nodi scelta a caso.

Negli anni 50 due matematici Erdos e Reny sviluppò il primo modello di rete. Questo modello non doveva essere regolare (deve contenere del disordine), e deve avere la proprietà dello small word. Il modello fatto è una rete casuale. Immaginiamo di prendere N nodi e ciascuna coppia di nodi con una probabilià p tiriamo un link. Si ottiene una rete con i nodi connessi a caso, e ogni nodo mediamente ha $p(N-1)$ vicini.

Bastano questi pochi ingredienti per ottenere una formula analitica. In una rete costruita con questo modello si ottiene che il diametro è proporzionale al logaritmo del numero di nodi.

$$d \sim \ln N$$

Oltre a questa carattersitica dello Small word le reti hanno un'altra caratteristica interessanti: il *clustering*. È facile che il numero di triplette chiuse di una rete diviso le triplette connesse è un coefficiente finito.

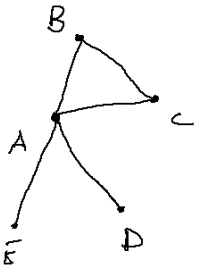


Figura 3.1: Schema di tripletta chiusa (ABC) e tripletta connessa (ADE). Una tripletta chiusa è sempre connessa.

Il coefficiente di clustering

$$C = \frac{\#_{\text{triplette chiuse}}}{\#_{\text{tiny triplette connesse}}}$$

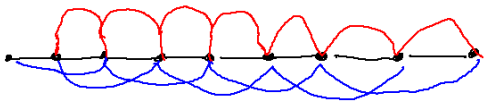
Il modello di E-R non basta più perché si ha

$$C \ll 1 \qquad \text{per} \qquad N \to \infty$$

Mentre per le reti sociali il coefficiente di Clustering ha un valore finto anche per $N \to \infty$ (Nella rete dei manager italiani il coefficiente di clustering è 0.59).

Questi elementi sembrano essere molto contraddittorie. Il clustering sembra dire che le reti sono molto dense, però poi questo effetto viene superato dalla caratteristica dello small world. Per tenere insieme questi effetti il modello successivo è di Watts e Strogatz, pubblicato su *Nature* nel 1998.

Immaginiamo di partire da un reticolo unidimensionale, ma immaginiamo di avere un link ai primi e secondi vicini.



Con una probabilità p prendiamo ciascuno di questi *link* e lo riattacchiamo a caso, e questo lo facciamo su tutti i nodi con probabilità p.

Se $p = 1$ otteniamo il modello di ER, per $p = 0$ otteniamo un modello regolare. L'interesse di questo modello sta nel fatto che per pochi valori dei link siano staccati e riattaccati. Questo modello ci genera reti che hanno un

forte raggruppamento, perché si parte da un reticolo regolare (con clustering altissimo), spostando pochi link rimane alto il clustering ma otteniamo un effeto di small world.

In questi modelli contano solo due parametri: il numero di *link* e il numero di *nodi*. I sociologi già dagli anni 50 si sono messi a cercare di trovare questi dati.

Per misurare il numero di *link* e nodi avevano scoperto che su grandi numeri è molto difficile dare un'idea specifica perché i numeri veri non si possono misurare.

Definizione 3.0.1 (Grado) *Il grado di una rete si definisce come il numero medio di contatti di un individuo nella rete.*

Il grado sembrava in origine essere non costante. Si può studiare la distribuzione statistica dei grado dei nodi in una rete. In caso di una rete direzionale bisogna distinguere tra grado uscente (k^{OUT}) e grado entrante (k^{IN}).

Venne fuori una cosa molto particoalre, la distribuzione del nodo è una legge di potenza:

$$P(k) \sim k^{-\alpha}$$

Con

$$2 < \alpha < 3$$

Questa distribuzione è molto particolare perché la varianza è infinita:

$$\int_0^\infty P(k)k^2 dk \sim \int_0^\infty k^{2-\alpha} dk = \infty$$

L'integrale non converge. Queste leggi di potenza sono oggetti molto interessanti per i fisici perche spiegano il comportamento nelle transizioni di fase.

Se queste reti vengono fuori così, deve esistere un principio profondo e universale che non riguarda il dettaglio della struttura della rete, ma la fisica alla base delle reti. Queste distribuzioni non sono riprodote nei modelli tradizionali di ER. Nel modello ER la distribuzione che esce fuori è una distribuzione binomiale, che ha valor medio e varianza ben definito (addirittura per $N \to \infty$ diventa una gaussiana). Questo ovviamente vale anche nel modello di WS, che è un caso intermedio tra reticolo regolare (con varianza nulla) al modello di ER (con varianza finita), quindi non è in grado di replicare la varianza infinta delle reti reali.

Nel 1999 si iniziano a studiare le reti *scale-free*. Barabasi e Albert fanno un modellino in cui le reti sono *dinamiche* e *auto-organizzate*.

Immaginiamo che la rete evolve nel tempo. Ad ogni istante di tempo entra un nuovo nodo nella rete, il nodo viene attaccato con il *preferential attachment*. Il nodo si attacca prevalemtemente al nodo con più amici (con probabilità proporzionale al grado del nodo dato). Questo principio è presente in tante reti reali, e può essere in grado di spiegare la distribuzione dei nodi.

Il modello di BA riproduce una legge di potenza a:

$$p(k) \sim k^{-3}$$

Immaginiamo che un nodo entri nella reti, e questo manda un solo *link*. $k_i(t)$ è il grado del nodo i al tempo t e $\Pi_i(t)$ è la proabilità di ottenere un nuovo link al tempo successivo. Il modello ci dice

$$\Pi_i \sim k_i$$

$$\Pi_i = \frac{k_i}{\sum_j k_j}$$

$$k_i(t+1) = \begin{cases} k_i(t) + 1 & \text{con probabilità } \Pi_i(t) \\ k_i(t) & \text{con probabilità } 1 - \Pi_i(t) \end{cases}$$

Possiamo calcolare l'aumento medio del grado

$$\langle k_i(t+1) - k_i(t) \rangle = \langle \Pi_i \rangle$$

Passando al continuo

$$\dot{k}_i(t) = \frac{k_i}{\sum k_j}$$

Il fattore di normalizzazione è la somma di tutti i gradi della rete (il doppio dei link introdotti dal tempo zero al tempo t):

$$\sum_j k_j = 2mt = 2t$$

$$\dot{k}(t) = \frac{k}{2t}$$

$$\frac{\dot{k}}{k} = \frac{1}{2t}$$

$$\ln k = \frac{1}{2} \ln t + c$$

$$k = At^{\frac{1}{2}}$$

Cercando il grado mendio dell'$i$esimo nodo possiamo stimare A.

$$k_i(t) = \left(\frac{t}{i}\right)^{\frac{1}{2}}$$

Dato che tutti nodi crescono allo stesso modo la differenza è dettata soltanto dall'istante di ingresso. Questo ci permette di passare dall'istante in cui abbiuamo introdotto un grado al suo grado.

$$\dot{k}_i = \frac{k_i}{2t}$$

$$k_i(t) = At^{\frac{1}{2}}$$

Sostituiamo la condizione inziiale:

$$k_i(i) = 1$$

Il nodo i viene introdotto dopo un tempo i.

$$k_i(t) = \left(\frac{t}{i}\right)^{\frac{1}{2}}$$

Questa formula ci dice che il grado dei nodi più vecchi è maggiore del grado dei nodi più giovani (e crescono anche più velocemente). Questo vuol dire che

ad ogni età del nodo possiamo far corrispondere il grado e vice versa. Questo ci dice

$$i(k) = tk^{-2}$$

La probabilità di estrarre a caso un nodo che abbia $k_i > k$ equivale ad estrarre un nodo poiù vecchio di una certa età. Questo ci permette di calcolare facilmente questa probabilità

$$P(k_i > k) = P\left(i < i(k)\right) = \frac{i(k)}{t} = k^{-2}$$

La probabilità che un nodo a caso abbia una data di nascita minore di un valore $i(k)$, la probabilità è il numero di nodi che vengono pirma di i (che sono proprio i) diviso il numero di nodi totali (t).

Per ottenere la densità di probabilità dobbiamo derivare la comulativa:

$$p(k) = -\frac{d}{dk}P(k) \sim k^{-3}$$

Il modello di Barabasi può essere modificarto per rendere la rete migliore. Se $m > 1$ si ottengono i triangoli, se poi modifichiamo questa probabilità

$$\Pi_i \sim k_i + c$$

Otteniamo esponenti che vanno da 2 a 3. Se invece si usa una probabilità

$$\Pi_i \sim k_i f_i$$

Con la fitness (un numero generato casuale) si rompe la relazione tra età e grado.

Un altro modello interessante è il *configuration model*. Questo modello è usato dal punto di vista computazionale per generare una rete computazionalmente con una certa sequenza. Vogliamo riprodurre una rete che abbia i nodi esattamente con il grado dato.

Immaginiamo di voler riprodure una re fatta da soli quattro nodi in cui due abbiano grado pari a due e due grado pari a 4.

Creiamo i nodi con i link da attaccare. Estraiamo a sorte due nodi e li uniamo (dobbiamo avere l'accortezza di partire da gradi maggiori) Questo modello è interessante per creare tante reti che abbiano tutte la stessa sequenza k. Quello che cambia tra una rete e l'altra non è la distribuzione di probabilità ma come questi nodi sono collegati tra loro.

Una rete è *complessa* quando ha un grado distribuito a legge di potenza, alto clustering e piccolo diametro.

3.1 Robustezza delle reti

Perché quando abbiamo reti autoorganizzate emerge questa forma di reti complesse? Queste reti complesse non sono scelte a caso, ma perché hanno delle caratteristiche di robustezza molto speciali (che le ha permesse di sopravvivere nel tempo).

Cosa intendiamo per robustezza? Immaginiamo di togliere a caso link, esiste una frazione minima di link rimuovibili prima che la rete si inizia a frammentare.

Se si studia il modello ER si può vedere qual è la percentuale di link da togliere alla rete prima di vederla frammentarsi. Avviene qualcosa di simile

al modello percolativo. Esiste una frazione critica che se togliamo dalla rete frammentano la rete in tanti pezzi molto più piccoli. In una rete complessa, se prendiamo prima i nodi più connessi, abbiamo una frammentazione rapidissima. Se i nodi vengono rimossi a caso la rete non si frammenta mai.

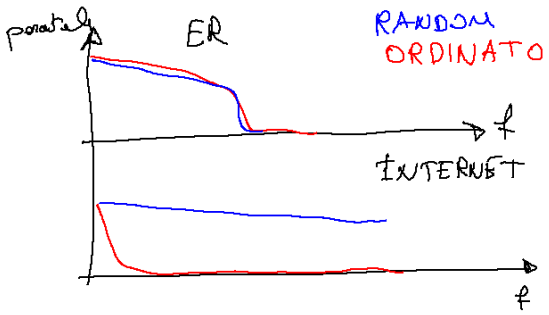


Figura 3.2: Sull'asse delle x troviamo la frazione di nodi rimossi, su quella delle y il numero di nodi nel cluster più grande della rete (rispetto al totale dei nodi).

Il modello di Barabasi-Albert descrive correttamente questo fenomeno, possiamo togliere una frazione molto grande di nodi casualmente.

Un altra cosa notata in queste reti riguardava il comportamento delle epidemie all'interno delle reti. Quello che osservarono Pastor-Satornas e Vespignani è che le infezioni virali informatiche inzialmente si diffondono a partire da un numero di server molto piccolo, in realtà i virus continuano a infettare i computer per tantisimo tempo.

I virus sparisono si esponenzialmente ma con costanti di tempo dell'ordine di anni (nonostante le patch vengono rilasciate pochi giorni dopo l'uscita del virus).

Loro simularono l'epidemia con il SIS serve a simulare le epidemie su larga scala. Su un reticolo regolare esiste una soglia di infettività critica al di sotto del quale il virus muore. Su reti complesse la soglia di infettività va a zero, quindi bastano pochissimi infetti che il numero di computer infettati sia una frazione finita del totale.

La struttura dei link nella rete ha delle caratteristiche speciali.

Lo studio principale per capire le reti è quello di studiare la percolazione all'interno dell reti. Nel 2000 viene studiato per la prima volta la percolazione sulla rete complessa. Su un reticolo esiste una soglia critica di percolazione.

Su una rete complessa si può studiare la percolazione analiticamente. Dobbiamo calcolare la soglia percolativa, quando si forma un cluster gigante?

Sia q la probabilità che un link scelto a caso non porti ad un nodo che fa parte di un cluster gigante.

Se $q = 1$ prendendo un qualunque link, non arrivo al al GC, se $q < 1$ esiste un GC. Possiamo calcolare la q in modo ricorsivo. La probabilità che la probabilità che un *link* non porti al GC, è la probabilità che un link mi porti ad un nodo non collegato al GC. Quindi tutti i link di questo nodo non devono portare al GC.

$$q = \sum_k P_{\rightarrow}(k) q^{k-1}$$

Ossia q è la probabiltià di scegliere un nodo di grado k per la probabiltà che tutti i link di questo nodo non portano al GC.

Il termine $P_{\rightarrow}(k)$ è la probabilità che un link a caso mi porti su un nodo di grado k.

$$P_{\rightarrow}(k) = \frac{P(k)k}{\sum'_k P(k')k'} = \frac{P(k)k}{\langle k \rangle}$$

Questa probabilità è il numero di nodi di grado k per il numero di link collegati a quel nodo.

Ritornando all'equazione originale si ottiene:

$$q = \sum_k \frac{kP(k)}{\langle k \rangle} q^{k-1} = F(q)$$

Possiamo risolvere l'equazione per via grafica.

$$F(0) = \frac{P(1)}{\langle k \rangle} > 0 \qquad F(1) = 1$$

$$F'(q) > 0 \qquad F''(q) > 0$$

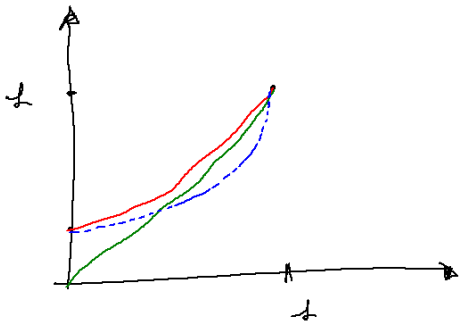


Figura 3.3: Esiste una soluzione nel sistema solo se la $F(q)$ ha una derivata in 1 maggiore di 1.

La condizione per l'esistenza di una soluzione non banale è

$$F'(q = 1) \geq 1$$

Possiamo calcolare tutto e questo equivale a dire che

$$\frac{\langle k^2 \rangle}{\langle k \rangle} \geq 2$$

Quindi se la rete ha questa probabilità la rete è connessa altrimenti la rete è frammentata. Per essere una rete connessa deve valere questa caratteristica statistica.

Quello che possiamo vedere è quanti nodi bisogna togliere ad una rete complessa per far diventare questo valore più piccolo di 2.

Questo è detto criterio di Mollloy-Reed.

La percolazione su queste reti hanno un comportamento del tutto diverso da quello che avviene su reticolo regolare. Tutti i fenomeni si diffondono in modo velocissimo. Il criterio di Molloy Reed ha un impotesi principale, il grado di un nodo è del tutto scorrelato dal grado dei suoi vicini. Ogni nodo è indipendente dai suoi vicini (questa è detta approssimazione di campo medio). Questa approssimazione è esatta con un β-lattice, in cui non ci sono cammini chiusi. Ad esempio questo non vale nel caso della classica percolazione in un reticolo quadrato, dove il numero di link non è indipendente dalla posizione del sito. Immaginiamo che la rete abbia una certa distribuzione del grado e immaginare a togliere dei nodi. La distribuzione cambia, sia $P_f(k)$ la distribuzione dei nodi dopo averne tolti una frazione f.

$$P_f(k) \qquad \frac{\langle k^2 \rangle}{k}$$

Dopo che abbiamo tolto un nodo alcuni link spariscono. La probabilità che un nodo di grado k_0 passi a un grado k è:

$$P(k_0 \to k) = \binom{k_0}{k} (1-f)^k f^{k_0-k}$$

La distribuzione del grado P' dopo la rimozione di f nodi va pesata sulla probabilità che parti da un nodo di grado k_0, che diventi di grado k dopo la rimozione.

$$P'(k) = \sum_{k_0} P(k_0) \binom{k_0}{k} (1-f)^k f^{k_0-k}$$

Possiamo calcolare il valore medio del grado e la varianza:

$$\langle k \rangle_f = \sum_k k P'(k) = (1-f) \langle k \rangle_0$$

$$\langle k^2 \rangle_f = \sum_k k^2 P'(k) = (1-f)^2 \langle k \rangle_0^2 + bho$$

$$\frac{\langle k^2 \rangle_f}{\langle k \rangle_f} = (1-f) \frac{\langle k^2 \rangle_0}{\langle k \rangle_0} + f$$

Il criterio di Molloy reed può essere riscritto per trovare la soglia critica f_c.

$$(1-f_c) \frac{\langle k^2 \rangle_0}{\langle k \rangle_0} + f_c = 2$$

Applichiamo questo criterio alla rete complessa:

$$p(k) = ck^{-\alpha}$$

Da cui possiamo calcolare il valore atteso:

$$\langle k \rangle_0 = \int_m^{k_{max}} P(k) k dk =$$

Il grado massimo del nodo può essere stimato facilmente dalla relazione

$$\int_{k_{max}}^{\infty} P(k)dk = \frac{1}{N}$$

Da cui possiamo integriamo

$$\langle k \rangle_0 = \frac{1}{2-\alpha}\left(k_{max}^{2-\alpha} - m^{2-\alpha}\right)$$

$$\langle k^2 \rangle_0 = \int_m^{k_{max}} P(k)k^2 dk = \frac{c}{3-\alpha}\left[k_{max}^{3-\alpha} - m^{3-\alpha}\right]$$

Da cui possiamo calcolare il fattore

$$z_0 = \frac{\langle k^2 \rangle_0}{\langle k \rangle_0}$$

Se la calcoliamo per

$$k_{max} \gg m \qquad \alpha > 3$$

$$z_0 \to m$$

Nel caso invece in cui

$$2 < \alpha < 3$$

$$z_0 \to bho$$

Se

$$1 < \alpha < 2$$

$$z_0 \to k_{max}$$

Se andiamo a insierire z_0 nella soglia ottieniamo

$$f_c = 1 - \frac{1}{m-1} \qquad \alpha > 3$$

$$f_c = 1 \qquad \alpha < 3$$

Questo vuol dire che la soglia tende ad 1 per le reti che ci interessano maggiormente. Se la distribuzione del grado ha un esmponente più piccolo di 3, la rete non può mai essere frammentata.

Nel caso delle reti complesse i fenomeni percolativi si comportano in maniera qualitativamente differente rispetto alle reti regolari.

3.2 Reti sociali

Le reti sociali sono leggermente differenti dalle altri. In sociologia esiste *l'assortative mixing* ossia i gradi dei nodi vicini sono fortemente correlati.

La correlazione in queste reti si misura sia con il coefficiente r di Pearson, ossia

$$r = \frac{\sum_{jk}(e_{ik} - q_j q_k)}{\sum_k k^2 q_k - \left(\sum_k q_k\right)^2}$$

Oppure si prende un nodo e si guarda qual è il grado medio dei suoi vicini e si fa la media su tutti i nodi, e si vede se il grado dei vincini è crescente c'è un

mixing assortativo se è decrescente c'è un mixing disassortativo (in funzione di k).

Quello che si vete è che per tutte le reti standard (internet, neuroni, interazione tra proteine, Web, ecosistemi) il coefficiente r è negativo, mentre per tutte le reti sociali r è sempre positivo.

Il motivo per cui le reti sono state molto importanti è il fatto che le reti sociali i nodi sono costruiti sulla base di comuntà gerarchiche. Le reti sociali sono diverse perché sociali o perché collaborative? Newman e Park hanno suggerito che sono le comunità a stabilire. Prendiamo N nodi con un cert numero G di comunità e creare connessioni a caso all'interno delle comunità.

QUesto modello può essere studiato analiticamente, e la correlazione in questo tipo di reti viene sempre positiva.

Siccome le relaizoni scientifiche è gisutificato correttamente, mentre quella per i manager il valore di r preisto da una rete collaborativo è differente. Le reti sociali possono avere infatti anche rapporto di mutua antipatia hanno caratteristica opposta, e si possono studiare entrambi.

3.3 Algoritmi spettrali

Questi sono un'insieme di metodi di studio delle reti attraverso lo studio delle matrici che definiscono le reti.

L'algoritmo spettrale più famoso è *PageRank*. Prima di Google i motori di ricerca facevano analisi sul contenuto della pagina. Google per la prima volta capì che l'importanza della pagina non era tanto nel contenuto, quanto nei collegamenti. Google trova che se una pagina è più importante su un ipotesi. Immaginiamo che il web sia una rete diretta, e i visitatori siano utenti che camminano sulla rete seguendo i link, l'ipotsi di google è che la rilevanza di una pagina web è la probabilità di visitarla camminando a caso sulla rete. Questa probabilità (chiamata *page rank*) immaginarono l'algoritmo.

Questa è la distribuzione stazionaria di un random walk sulla rete a tempi discreti. Studiamo il cammino aleatorio su questa rete, quando questo cammino sarà durato abbastanza tempo le probabilità di trovare un cammino aleatorio.

Immaginiamo di avere quattro nodi. A ciscuno di questi link associamo la probabilità di trasferimento (dato dal numero di link su ciascun nodo). Possiamo costruire la matrice A_{ij} con la probabilità di andare dal nodo j al nodo i. Sia $\bar{\nu}$ la distribuzione iniziale di tanti random walk che partono a tutte le pagine, qual è la probabilità di trovare un random walk sul nodo i al tempo t sarà $\nu_i^{(t)}$. Vogliamo calcolarla per $t \to \infty$.

L'evoluzione temprale di questo random walk è

$$\nu_i^{(t+1)} = \sum_j A_{ij} \nu_j^{(t)}$$

Possiamo calcolare lo stato stazionario:

$$\nu_i^{(\infty)} = \sum_j A_{ij} \nu_j^{(\infty)}$$

Questa è un equazione agli autovalori

$$\nu^{(\infty)} = A\nu^{(\infty)}$$

Il *page rank* è l'autostato con autovalore 1 della matrice dei link A del web. Google per misurare la rilevanza delle varie pagine guardava la rete, costruiva la matrice A calcolava l'autovettore principale e le componenti di questo vettore corrispondeva al page rank che era associata a questa pagina. Ossia immaginava un miliardo di random walk sulla pagina.

Questo algoritmo funziona fin al momento in cui non abbiamo una pagina senza link (o loop chiusi). In questo modo il *page rank* sarebbe 1 per la pagina senza link e zero per tutto il resto della rete. Allora l'algoritmo ha una probabilità p il random walk fa un salto casuale in un altro punto. Per aggiungere questa proabilità si modifica leggermente la matrice:

$$M = (1-p)A + pB$$

Dove la matrice B è definita:

$$B = \frac{1}{n}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

In questo modo curiamo il problema delle patologie. In questo modo la matrice M è ergodico e vale il teprema di Perron-Frobenius. Il teorema ci dice che c'è sempre un autovalore 1 di molteplicità 1, ed è il più grande. Questo autovettore associato all'autovalore ha semrpe elementi positivi.

Se guardiamo la matrice A^T al posto di A. Anche questa matrice ha un'autovalore pari ad 1 e questo ha tutte componenti costanti.

Se la matrice A è data da gruppi sconnessi otteniamo una matrice a blocchi. Quello che succede è che la matrice trasposta di questa rete ha un autovettore (di autovalore 1) di molteplicità pari al numero di sotto comunità di node sconnesse tra loro.

Se non abbiamo tre autovalori pari a 1 ma ne abbiamo tre di valore prossimo a 1 (0.98) e di valore leggermente diverso vuol dire che abbiamo comunità ben definite che hanno pochi collegamenti tra loro. Studiando gli autovalori della matrice trasposta ci da informazioni sulla struttura comunitaria della rete. Questo ha una grande applicazione tecnologica. Riuscire ad individuare le comunità di pagine web permetteva ai motori di ricerca di fare una ricerca per comunità.

3.4 Analisi dati

Il sito *SNAP* è un grosso database con tantissimi dati su reti grandi.

Capitolo 4

Sistemi economici

4.1 Fatti stilizzati di economia

L'economia è molto differente dalle scenze naturali, perché le leggi sono auto imposto. Le scienze sociali mutato sia qualitativamente che quantitativamente in modo irreversibile nel tempo storico. La storia nella scienza è stata introdotta da Darwin. Siccome c'è un elemento storico nella dinamica che crea condizioni economiche in cui gli agenti operano non possiamo pensare di affrontare il sistema con delle equazioni di equilibrio o con elementi di minimizzazione. Descrivere gli agenti come un sistema di gas perfetto all'equilibrio è molto diverso dal considerare il sistema fortemente fuori equilibrio. Dobbiamo guardare le cose da un punto di vista storico, considerandole nel loro evolvere.

Questa cosa è cruciale perchè in economia ci si è illusi di poter descrivere gli ensamble economici come ensamble di equilibrio. Questo ha portato ad una serie di descrizione errate della realtà. Non bisogna pensare che gli economisti non siano riusciti a catturare la complessità del mondo. Gli economisti sono come dei medici, hanno bisogno di ricette. L'atteggiamento dell'economista cerca di risolvere i problemi, ha costruito un costrutto teorico molto potente che viene dalla filosofia che gli permette di risolvere i problemi con delle ricette.

Netwon negli ultimi anni della sua vita affrontò un problema molto serio. In inghilterra si erano sviluppate due monete, una per le colonie e una per il commercio interno. Non era definito un tasso di cambio tra le due. Nel momento in cui la finanza si inizia a sviluppare, e non c'è un tasso di cmbio ben definito. Se oggi si iniziano a emettere i future (ipotesi o vincoli su futuro) questo crea una istabilità sul vincolo economico. Il tasso di cambio imposto da Newton crea un problema, la percezione dell'operatore è diversa da quello di Newton, la percezione è diversa da quello che la realtà si presenta si fa un incetta di moneta d'argento perché si crede siano inferiori al cambio, questo crea una scarsità dell'argento. Solo nel 1821 l'inghilterra introdusse il gold standard, che permetteva l'aquisto di una quantità di oro standard con una ben deinita quantità di monete.

Un sistema complesso è un sistema con una rete di interazioni e una rete di agenti irriducibile. In un sistema semplice se abbiamo due variabili e soserviamo che c'è una correlazione tra due variabili possiamo introdurre un principio di causalità forte. In un sistema complesso abbiamo molta difficolta nel ricavare

le cause e gli effetti. Quando non riuscimo a ridurre il problema e non è chiaro il sistema, perché non riusciamo a fare delle operazioni di taglio.

La storia del pensiero economico si fa risalire a d Adam Smith, autore di *the welth of nation* e inizia a porsi il problema di perché esistono nazioni più avanzate rispetto ad un altra. Simth rinuncia ad una trattazione matematica del problema, questo perché la matematica sviluppata dalla fisica non era idonea a descrivere questo tipo di sistemi. I sistemi non sono tutti in equilibrio. Una volta fatta l'ipotesi riduzionista si possono fare esperimenti, perché possiamo isolare il sistema da studiare dall'ambiente. Primo di Galileo il mondo era autoritario, non abbiamo piùà l'elemento autoritativo che mi dice chi è la legge, se la legge non viene dettata da qualcuno, come facciamo a sapere se una legge è giusta? Con il metodo riduzionista.

Il pensiero economico è fortemente legato alle rivoluzioni scientifiche. Un agente economico avrà una tendenza ad ottimizzare le sue azioni che possiamo prevedere, e quindi il sistema è stabile. L'ipotesi ragionevole è che possiamo immaginare di descrivere il mondo attraverso una dinamica relativamente stabile. Fino a subito prima della crisi del 2008 ha prevalso il modo di descrivere la dinamica attraverso il principio della minimizzazione. La domanda della regina, che alla normal school of economics.

È avvenuta un istabilità del sistema negata strutturalmente dalla teoria economica. Il frutto di un'istabilità intera, La domanda della regina è questa: perché nessuno l'aveva notato? Perché nessuno l'aveva immaginato? La risposta arrivò nove mesi dopo, in cui diranno che, lavorando in un ipotsi riduzionista, il mondp era stato diviso in tanti piccoli pezzi, in qualche modo si era persa la pittura globale, non c'era nessuno che aveva guardato il sistema nella sua complessità, mancando l'elemento della complessità.

Ci sono stati quattro capitalismi: il capitalismo coloniale, che accompagna l'inizio della rivluzione industriale. È un capitalismo che lavora sullo spostare le merci, e ha bisogno di una finanza molto aggressiva. Il carico viene prevenduto dalle navi. Viene spostata nel futuro un'azione dell'oggi. Si crea la borsa, viene creata da una famiglia veneziana che si chiama Borsa. C'è un capitalizmo industriale che invece lavora per trasformare le merci. Questo capitalismo porterà alla guerre mondiali, che sono guerre di comquiste. Il capitalismo finanziario nasce nel dopoguerra perché il sistema ha bisogno di crescere, e si fa attraverso la distribuzione del debito, si distribuisce il rischio tra chi assume il pezzo del debito. La finanzia significa spostare nel futuro un pagamento. Il capitalismo finanziario lavora sul denaro, la moltiplicazione del denaro. QUesta moltiplicazione funziona distribuendo il rischio, io mi prendo un debito e ti chiedo in prestito del denaro. Ora c'è il capitalismo digitale, è un capitalismo che lavora solo sull'informazione non più sul denaro, in questo capitalismo gli schemi sono cambiati radicalmente. In un economia in cui siamo contemporaneamente fornitori e clienti il ruolo del denaro viene meno. L'economia è lo studio di risorse limitate in uno stato con divisione del lavoro. La divisione del lavoro è cruciale in un sistema economico.

L'economia è lo studio della distribuzione delle risorse limitate. Nel pensiero econimco è importante e diverso l'approccio tra i fisici e gli econmisti. L'utilizzo della matematica in economia, molto formale e molto pesante, viene usata per verificare la coerenza interna di un ragionamento che viene fatto. Non viene usata per descrivere la realtà. L'impostazione è il tentativo di un estrema rigore,

Nasce il concetto di mano invisiibile, Pareto scrive delle equazione (ottimo Paretiano) l'ottimo raggiunto da un sistema di agenti economici, un punto in cui non possiamo far muovere nulla senza che alcuno agente perda qualcosa, ci si immagina che ci sia una mano invisibile. Da li in poi nasce la sintesi neoclassica: viviamo in un mondo di agenti ottimizzanti, queste azioni ci portano inevitabilmente verso un miglioramento. Questo è il momento di svolta dell'economia. Ci si immagina di descrivere l'economia con le leggi della fisica classica. Kanse immagina che non ci sia questo equilibrio, però sostiene che tanto sul lungo termine saremo tutti morti. Dobbiamo studiare quindi necessariamente sistemi fuori equilibrio. Quando affrontiamo un rischio, queto può essere giudicato. Black e School inventano un equazione con il rumore. Questa equazione è così importante perché permette di fare i derivati. Un derivato è una scommessa su quello che accadrà domani. Per impaccettare tutte queste scommesse devo avere uno strumento matematico che mi dice quanto vale. Black e school scrivono un equazione che permette di descreivere quale sarà il prezzo all'indomani. Questo fa un impotesi di stabilità. Dobbiamo fare delle ipotesi, che il mondo che descrivo sia sufficientemente non correlato.

Con due ipotensi della razionalità dell'agente (agisce ottimizzando un'ignota funzione di "utilità") e della efficienza dei mercati (tutta l'informazione di un bene è concentrata del tempo) possiamo costurire un fondamento teorico che tratta il sistema matematicamente con una formulazione stotastica. Questa stocastica non è dovuta all'istabilità delle azioni interne.

Gli elementi della teoria neoclassica è l'ipotesi di un equilibrio statico di sottofondo. Non si interessa di come avviene il flusso dinamico dell'informazione del sistema.

Nell'economia avvengono transizioni strutturali nell'economia.

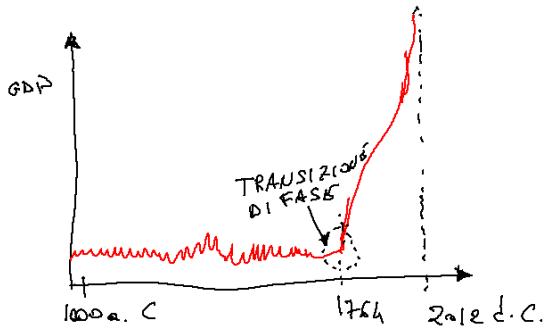


Figura 4.1: Transizione di fase della rivoluzione industriale.

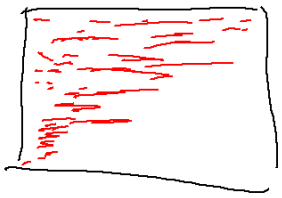
Le regole dell'economia possono essere bypassate, perché scritti dall'essere umano. Questo fa si che quando si affronta l'economia bisogna sempre ricordarci che è possibile un processo di modellazione, e di analisi empirica, tuttavia questi modelli hanno un limite intrinseco perché possono cambiare le leggi di interazione degli agenti.

4.2 La metrica: Fitness

Il punto di partenza è l'eterogenità. Supponiamo di voler definire una misura di complessità a partire da quanto diverse sono le matite e quanto sono complesse: Prendiamo i 17 paesi più poveri nel 1960. Con tutti un GDP sempre uguali. Tra questi ci sono anche la China e il Burundi. Questi sistemi partono con condizioni iniziali molto simili succedono evoluzioni estremamente differenti.

Ciascun essere è univocamente determinato dall'informazione genetica. Questa si manifesta nei fenotipi. Questo determinano la fitnes del nostro ambiente. Il passaggio dall'informazione genetica al fenotipi si perde informazione. Tuttavia dall'osservazione dei fenotipi è possibile inferire molta dell'informazione gentica.

Vogliamo identificare cosa è il DNA di un paese, cosa è il suo fenotipo e qual è la fitness di un paese. L'informazione genetica di una nazione è un ste di capabilities (il sistema bancario o finanziario, sistema scolastico, sanitario, inquinamento, risorse naturali, ...) Scrivere questa lista è impossibile e ancora più complicato è scrivere l'interazione tra queste *capabilities*. Come interagisce come stiamo migliorando il sistema educativo, e questo come agisce sull'industria? Non sappiamo scrivere l'informazione genetica di un paese, ma siappiamo il suo fenotipo. Il fenotipo è quello che il paese riesce a produrre. Questi dati non sono coerente e omogenei, ma esistono dati molti coerenti e affidabili degli *export*. I dati di *export* sono raccolte dalle dogane, tutte le merci che passano dalle dogane devono pagere dei dazi e quindi sono standardizzate. Si può dimostrare che sia una sonda affidabile di ciò che il paese produce. SI ha una lista dei paesi e una lista dei prodotti. Si mette un *link* tra un paese e un prodoto se il paese è un produttore (o esportatore) rilevante di quel prodotto. In questo schema non ci sono link tra paesi e link tra prodotti. Questa è una rete bipartita, poiché possiamo dividerla in due sottosistemi in cui non ci sono link tra nodi dello stesso ordine. Difatto le *capabilities* rappresentano un layer intermedio che collega i paesi ai prodotti. Quella rete bipartita si può rappresentare con una matrice di adiacensa (ogni riga un paese e colonne prodotti) Se si riordinano righe e colonne delle matrice otteniamo un organizzazione triangolare. Questo da molte informazioni econimiche. Le Le righe in alto sono i paesi più sviluppa-



ti, le colonne a destra sono i prodotti complessi (elettronica) mentre a sinistra abbiamo i prodotti sono prodotti pochi complessi (come materie prime).

Le nazioni che hanno maggior successo in questo momento sono quelle più diversificate. Quello che emerge è che i paesi competono in un ambiente dina-

mico, per cui la specializzazione funziona su un tempo limitato, e la robustezza è data dalla diversifiazione.. Vogliamo provare a scrivere delle equazioni. Se diciamo che un prodotto è fatto da un paese diversificato non stiamo dicendo molto. Se un prodotto è fatto da un paese poco diversificato, stiamo dando molte informazioni su questo prodotto, perché questo prodotto avrà bassa complessità. Allo stesso modo un paese che produce un prodotto molto ubiquo non stiamo dicendo molto sul paese, mentre se il paese fa un prodotto fatto da molti piochi paesi, il paese è molto sviluppato. Questo si traduce in un algoritmo

$$\tilde{F}_c^{(n)} = \sum_p M_{cp} Q_p^{(n-1)} \qquad \tilde{Q}_p^{(n)} = \frac{1}{\sum_c M_{cp} \frac{1}{F_c^{(n-1)}}}$$

Dove F_c è la fitness del paese, M_{cp} è la matrice paese-prodotto, e Q_p è la complessità del prodotto:

$$F_c^{(n)} = \frac{\tilde{F}_c^{(n)}}{\langle \tilde{F}_c^{(n)} \rangle_c} \qquad Q_p^{(n)} = \frac{\tilde{Q}_p^n}{\langle \tilde{Q}_p^{(n)} \rangle_p}$$

Questo è un algoritmo iterativo, alla fine abbiamo un vettore di fitness e un vettore di complessità e la complessità dei prodotti. La formula è non linearia, la complessità del prodotto è dello stesso ordine di grandezza della fitness del peggior stato.

La fitness e complexity è possibile applicarli anche agli ecosistemi complessi: di quanti alimenti si ciba una data specie e ogni alimento di quante specie è cibato, possiamo usare una fitness e complexity per determinare la criticità di quell'alimento o specie, in modo che se si togliesse quell'alimento o specie come crollerebbe il sistema.

Supponiamo di avere una machina di 800 cavalli con 180 km/h e una che va a 80 cavalli alla stessa velocità. Adesso usando GDP e la fitness possiamo dire qualcosa sui potenziali di crescita a lungo termine dei due paesi. Costruiamo anno per anno i ranking delle nazioni per quattro paesi (per i paesi BRIC, Brasile, Russia India e Cina). Se guardiamo le funzioni in termine di GDP questo era tutto abbastanza coerenti.

Possiamo mettere uno scatter plot, possiamo mettere Fitness e GDP, c'è un po' di correlazione. Le deviazioni dalla linea sono molto significative. Perché oggetti con grande GDP ma basso Fittness hanno enorme probabilità di fare un crack. Si può studiare la dinamica dei paesi su questi piano. Ci sono due zone una di flussi laminari, in cui loi spostamento è prevedibile, e altre in cui succede un gran casino. La mappa che ci interessa di più è una mappa che ci da la predicilbilità dell'atmosferica, misurando quando sono prevedivbili gli eventi metereologiche.

Quello che emerge è che la predicibilità molto diversa, e queste zone sono clusterizzate. Questo esatamente come nel grafico scuttering. La dinamica economica ideale vive in uno spazio ad alta dimensionalità. Quindi abiamo bassa dimensionalità nell spazio della dinamica, quindi quando proiettiamo su questo piano cioè che è vicino è vicino anche nello spazio vero. Invece nella regione caotiche due oggetti che sembrano vicini, possono essere molto lontani, Ossia fitness e GDP sono ottimi autovettori nella principal component analisys, mentre nella zona caotica esistono altri autovettori con autovalori confrontabili, quindi la risudizione in fitness e GDP non basta. Possiamo costruire anche

le distribuzione di probabilità di dove finiranno i paesi. Potremmo pensare di usare queste funzioni per fare delle distribuzioni a lungo termine, e si trovano cose molto interessanti. In un economia di sussistenza se riversiamo una certa quantità di investimenti, esiste la poverty trap in cui anche se insiriamo dei soldi non succede nulla. Esiste la Middle-income trap. Sono paesi emergenti, in cui rimangono in un limbo intermedio, ma non riescono a colmare l'ultimo gap tenologico. Corea del Sud sono riusciti a colmare questa gap, ma paesi che sono rimasti invischiati sono ad esempio il sud africa. Dopo una fase di sviluppo non è riuscito a fare. Queste trappole vengono definiti in termini di GDP pro capite. Bisogna dare dei criteri di trappole che sfruttano la fitness. Se proviamo ad aumnetare la dimensionalità dello spazio mettendoci un asse di fitness. Quello che viene fuori è che non tutti i paesi stanno in mittle-trap.

In realtà c'è un continuo di trappole economiche. I paesi petrolifici sono sottosoglia. Si può tentare di fare un fit lineari.

4.3 Evidenze empiriche e modelli finanziari

L'econofisica è un termine abusato, un sistema complesso per definizione è il comportamento umano. Tanti trader umani presentano una dinamica estremamente complicata. Per l'analisi dei mercati finanziaria ci sono tre livelli di comprensione del sistema:

- Analisi fenomenologica: Possiamo dare descrizione quantitativa ma descrittiva del sistema (fatti stilizzati)
- Possiamo azzardare l'equivalente di un modello fisico del mercato finanziario, questo si fa con i modelli ad agente. Sono delle metafore con cui si ispira,
- Comprensione teorica, comprensione degli esponenti critici del sistema con il gruppo di rinormalizzazione.

La prime introduzioni di modelli usati nella fisica per la finanza. Il moto browniano è stato applicato per la prima volta alla borsa da Bachelier nel 1900. Nei tempi più recenti c'è stata una crescita rapida. Il fatto che ci sia poco feedback con gli esperimenti (che non sono riproducibili visto il fatto che l'intero sistema evolve fuori dall'equilibrio pertanto le situazioni non ricapitano mai uguali nella borsa). Si tende pertanto a preferire modelli semplici e solubili analiticamente piuttosto che modelli complicato.

Non è sicuro che esistano leggi universali in ambito economico. La visione standard dell'economia riflette la presenza di agenti razionali (gli agenti hanno la capacità di prendere la decisione giusta, prendere le informazioni disponibili sul mercato e elaborarle sempre nel modo migliore). Il prezzo rappresenta complettamente l'informazione del mercato. L'agente dovrebbe avere in mente il valore realte del prezzo e quindi l'agente compra o vende se il prezzo è minore o maggiore di quello percepito. Non è possibile fare profitto senza rischi se tutti gli agenti hanno la stessa informazione.

In questi modelli non è possibile fare una previsione accurata sul prezzo perché gli aggiustamenti sono quasi istantanei. Il prezzo è fortemente collegato alle notizie disponibili. Ci sono diversi problemi: le stesse proprietà statistiche sono viste a scale molto differenti. Questa oscillazione dei prezzi si vede a scale

temporali molto diverse. Andando a scale temporali del secondo o di anni si vedono statistiche molto differenti.

Esistono una serie di analisi empiriche che fanno vedere come notizie scambiate e volumi dei prezzi giocano un ruolo minore.

Un modello di ordine zero che si può pensare per i prezzi è il random walk. Il problema è che spesso si notano in borsa fluttuazioni molto più grandi, ossia a legge di potenza.

$$f(x) \to x^{-\alpha}$$

Questa cosa fu notata per la prima volta da Mandelbrot nel prezzo del cotone. La legge a potenza non può essere descritta da un random walk gaussiano. Possiamo provare a introdurre dei derivati del processo gaussiano per dare fluttuazioni più grandi.

Esiste un breakdown dell'invarianza di scala che è quello in cui andiamo ad una scala temporale troppo piccola (ad un livello tick by tick che è un effetto di discretizzazione). Possiamo provare a rilassare l'ipotesi di stazionarietà gaussiana ed estrarre con valore atteso e varianza che variano. L'ipotesi che si può fare è che esista una funzione di correlazione che lega la realizzazione a due tempi diversi, funzione soltanto della differenza di questi due tempi.

Possiamo definire una memoria del sistema. Se la funzione decade esponenzialmente del tempo abbiamo una dipendenza esponenziale:

$$R(\tau) = e^{-\frac{\tau}{\tau_c}}$$

Dove τ_c è il tempo di correlazione. Nei sistemi complessi:

$$R(\tau) \sim \tau^{\alpha-1}$$

Se questo è il caso questa funzione può non essere integrabile e il sistema presenta una memoria infinita.

Si può vedere ad occhio che esistono momenti di agitazione del mercato e momenti tranquilli, il coefficiente di diffusione del random-walk non è stazionario.

La quantità principale che viene studiata son i return:

$$r_t = p_t - p_{t-\Delta t}$$

Ossia l'incremento del prezzo dopo un certo tempo Δt. Spesso questa quantità è definita con i logaritmi che mi danno la variazione relativa del prezzo.

$$r_t = \log p_t - \log p_{t-\Delta t}$$

Il primo fatto empirico che può essere visto nel return è che i return non sono gaussiani. Abbiamo degli eventi che vanno fino a 160 sigma, sotto un ipotesi gaussiana erano del tutto impossibili. Possiamo concludere che le fluttuazioni dei mercati finanziari hanno una caratteristica non gaussiana.

La funzione è asimmetrica, infatti i grossi eventi negativi sono più frequenti dei grossi eventi positivi, percui la distribuzione ha una scewness non nulla. Siamo molto fuori dal regime di Levy, quindi non si esclude che aggregando il return si ritorni ad una gaussiana. Guardando la kurtosys (che da la presenza di code grasse), aggregando sempre di più il valore della k scende (ma per quanto ne sappiamo non va a zero). La correlazione delle r va a zero molto rapidamente (a scale temporali molto piccole abbiamo una leggera anticorrelazione).

Abbiamo visto che c'è una clusterizzazione tra momenti di grande euforia e momenti di grande clustering, possiamo studiare l'autocorrelazione per i valori assoluti. Quanto è probabile che se abbiamo avuto oggi una fluttuazione in modulo di un certo valore domani abbiamo la stessa fluttuazione. Questo decade a zero con un esponente addiritutra minore di 1 (quindi non è integrale). Questo effetto è detto *volatility clustering*: le fluttuazioni della stessa misura tendono a clusterizzarsi. I return non sono correlati ma non sono indipendenti.

Una cosa interessante da dire è la presenza dei gap durante la notte. Una cosa molto dibattuta è se questi esponenti hanno o meno una classe di universalità.

4.4 Theoretical Models

I modelli possibili sono quelli di scrivere un processo stocastico per i prezzi, come ad esempio il random walk. Non ci interessiamo del perché il prezzo vada da un lato o l'altro, ma studiamo attraverso un processo stocastico. Questi metodi sono detti *econometrici*.

Possiamo cercare di capire perché succedono quello che vedono, ossia costruire un mercato virtuale da simulare, ossia mettiamo una causa al modello.

L'ulitimo approccio è l'order book, ossia il funzionamento vero del movimento dei prezzi, come fanno gli agenti per mettersi daccordo su che prezzo vendere o comprare (approccio microscopico). Questo permette di avere una grande mole di dati.

Una possiile correzione al random walk è quello del random walk geometrico. Ossia in cui la variabile stocastica è la variazione relativa di prezzo:

$$P(r) = \frac{1}{\sigma r\sqrt{2\pi}} e^{-(\ln r - \mu)^2/2\sigma^2}$$

Anche questo non è sufficiente per spiegare i fatti normali. Questa non presenta i volatility clustering, perché i return sono indipendenti per quanto riguarda il segno.

Un modello ARCH ha vinto il nobel nel 2003. In questo modello la varianza varia con la fluttuazione:

$$p_i = p_{i-1} + \sigma_i \xi_i$$

$$\sigma_i^2 = \alpha + \sum_{j=1}^{k} \alpha_j \eta_j^2$$

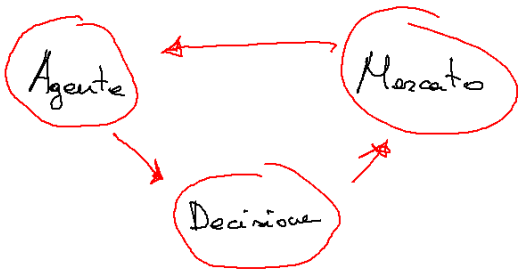
Questo da un volatility cluster ma non a legge di potenza.

Pi si mette dinamica nel modello più parametri possiamo aggiustare per fare il fit più possiamo aggiustare. Potremo scrivere un processo stocastico complicatissimo che fitta perfettamente ma non da un insight della motivazione di questa fluttuazione.

Si possono fare dei modelli stocastici per cercare di capire il "sentimento" del mercato. Esiste il modello cook, in cui il return dipende dalla diztnza della media mobile (la media degli m passi precedenti). In base la segno possiamo definire due possibili strategie, quella de trend follower, oppure il trend aderse, che ha una tendenza più stabilizzante. Questa non ha la pretesa di stimare i fatti stilizati ma permette di fare dei fit per capire se nel mercato ci sono questi trend.

4.4.1 Modelli ad agente

I modelli si basano sullo studio del singolo agente, perché non fare dei modelli che non agiscono seguendo le ipotesi della gente razionale ma hanno strategie proprie, possono farsi prendere dal panico. Abbiamo una struttura del tipo



L'agente attua una strategia che porta ad una decisione che a sua volta modifica il prezzo del mercato, da cui l'agente impara per migliorare la propria strategia.

Un dei modelli più efficacie è quello di Lux-Marchesi che prevede l'esistenza di agenti razionali (fondamentalisti), che tendono a riportare il prezzo verso il prezzo fondamentale (stabilizzano). I Chjaritisti che si dividono in ottimisti e pessimisti, dopo aver visti la serie storica decidono se vendere o comprare (hanno un effetto destabilizzante).

Un altro modello importante è la presenza dell'heriding, possiamo sfruttare la dinamica di opinioni, se vediamo tante persone che hanno una tale strategia cercheremo di rivincerla. Il modello è molto difficile da risolvere analiticamente.

Una simulazione di questo modello da esattamente i fatti stilizzati che si evincono. Riusciamo a spiegare i fatti stilizzati con il modello di Lux Marchesi.

La scelta dei parametri è molto particolare. Se prendiamo un numero di agenti $N = 5000$ non da luogo ai fatti stilizzati, $N = 500$ si. Quindi non si capisce bene se questo riproduzioni sono dovute ad una accidente scelta buona dei parametri.

Possiamo studiare per quanto riguarda gli intermediari (aggregazioni di broker) la correlazione tra il return è la inventory variation (il valore scambiato quando si compra meno il valore scambiato quando si vende, return del prortafoglio dell'agente).

Possiamo assumere un assunzione lineare in cui la inventory variation è lineare nella $r(t)$, il segnodi γ.

Si può studiare questo γ al variare della grandezza della firm (associazione di broker). Le firm che tendono a seguire il mercato sono poche e sono molto grandi (con γ positivo). Quelli reverse sono molti di più e molto eterogenei (γ negativo). Le firm che fanno un revercing tendomo molto a copiarsi fra loro.

La dinamica di opinioni ha effetti tangibili sul mercato.

Capitolo 5

Strumenti di analisi di sistemi complessi

5.1 Machine learning

La *machine learning* è un oggetto che impara in modo raffinato dal sistema, in grado di gestire situazioni in cui un essere umano non saprebbe cosa prendere.

Il machine learning ci insegna qualcosa sulle variabili rilevanti del sistema. Una regressione lineare è un esempio di machine learning. Un altro problema è quello della classificazione. Esistono due grandi classi di approcci, esistono casi supervisionati, stadi del sistema per cui si conosce la risposta giusta. Un altra cosa che si può fare è l'andamento non supervisionato. Si possono raggruppare i punti.

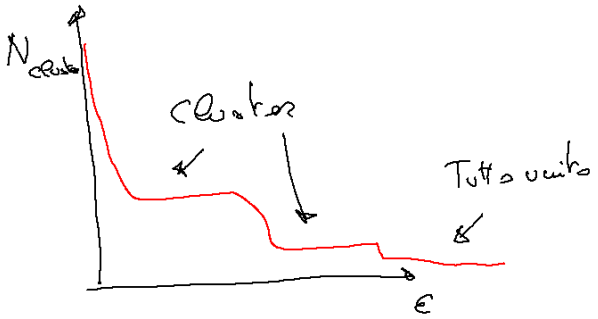
Si può tentare di minimizzare la somma delle distanze tra i punti all'interno dei cluster. Questo algoritmo è detto k means. Il primo problema è definire il numero di clusters, questo algoritmo cresce quadraticamente quindi è necessaria fare quealche approssimazione.

Si scelgono a caso due punti. si clusterizzando tutti i punti più vicini ad un dato cluster. Dopo di che ricalibra il centro del cluster come baricentro.

Questo metodo sostanzialmente trova il piano che ottimizza l'algoritmo.

Un altro algoritmo si basa sulla densità. Questo algoritmo si chiama *DBSCAN*. Bisogna scegliere due parametri, un raggio epsilon. Disegnamo circonferenze dei nostri punti e tutti quelli che cadono nella circonferenza risultano collegati. I punti si uniscono usando dei network. Nel algoritmo DBSCAN c'è un altro parametro: il numero n di punti che devono essere all'interno di un raggio ε.

In questo modo si escludono punti isolati. Questo algoritmo costruisce un network ed è in grado di dividere a prescindere dalla forma dei cluster. per capire i parametri ε e n si possono fare dei plot del numero di cluster in funzione di ε. I cluster corretti sono i platoux



5.2 Riconoscimento testo

Uno dei metodi perfetti è il teorema di Bayes (*Bayes naive classifier*). Permette di collegare un set di cause in un osservazione è tornare indietro. QUesti modelli sono molto generali. Si cerca di predire una classe C dato un vettore di Features. Si suppone che tutte le Features sono indipendenti. Dobbiamo farci un idea di come sono fatte le probabilità $P(F_i|C)$. Il dataset $MNIST$ permette di

$$P(C|F_1 \cdots F_i) \propto P(C) \prod_{i=1}^{n} P(F_i|C)$$

Un altro algoritmo è l'albero decisionale, il modello è un sistema ad albero che pone delle domande ai dati. Se chiediamo la classe in cui viaggiava il passeggero quanto possiamo discriminare i passeggeri e dividiamo in gruppi. Verifichiamo quanto questo gruppo divide bene la nostra variabile. Un metodo corretto è quello di calcolare l'entropia della divisione operata dal singolo campo.

All'interno di tutte le domande scegliamo la decisione migliore. La soglia giusta dell'età in cui dividere i casi. Siamo riusciti a realizzare un modello compatto che tiene conto delle correlazioni tra variabili ed è un modello parlante, da informazioni sulle variabili importanti.

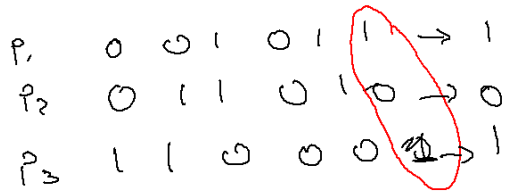
Un paese è un osservabile di uno stato di variabili che produce o meno un determinato prodotto. Vogliamo trovare la probabilità che un dato prodotto passi ad essere zero o uno come funzione degli altri prodotti. Il numero di possibili stato. Vorremo cercare di capire quali prodotti infulenzano veramente il prodotto che ci interessa. Ad esempio non sappiamo se le automobili sono collegate alla produzione di grano.

Il metodo del random tree è

Tuttavia se il numero di colonne è enorme esiste grossa probabilità di vedere correlazini casuali. Per ovviare a questo problema si prendono a caso poche righe e colonne e si costruisce un albero. A questo punto si possono costruire tanti alberi casuali.

Si può fare la media delle risposte dei singoli alberi. Si costruiscono dei predittori deboli ma si usano tanti tutti assieme. Questo permette di far sparire una correlazione casuale che avrebbe altrimenti avuto una conseguenza catastrofica.

Esiste un pacchetto python scikit-learn.



Capitolo 6

Dinamica sociale

QUando si parla di sistemi complessi ci si chiede se c'è ordine o un disordine, oppure se esistono dei pattern del sistema che si ripetono.

Si può cercare di capire se all'interno di una dinamica di opinioni o consenso se si converge ad un consenso generale.

Si raggiunge sempre l'ordine? Esistono transizioni di fase tra stati ordinati e no? Esistono forme di patterns nel modello? Qual è la dinamica pirma della convergenza verso lo stato asintotico. Quanto tempo ci si mette ad arrivare nello stato asintotico. Lo stato asintotico è uno stato statico (assorbente) oppure uno stato stazionario (proprietà statistiche costatni ma con una certa dinamica) o c'è una dinamica lenta (come i processi di aging)?

La dimensione come influenza il sistema? La topologia ha ruolo?

6.1 Modelli

Vedremo due modelli, come il modello di Voter che si basa sulla imitazione o il modello spin flip cinetico di ising (in cui un individuo).

Il mo9dello di Voter si prende un nodo della rete e quel nodo copia l'opinione di uno dei vicini a caso. Il modello di Ising invece campiona tutti i vicini del nodo e associa il valore della maggioranza dei nodi vicini (campo medio).

Il Courseling è un ottimo metodo per studiare la formazione di domini:

$$C(r,t) = \langle s_i(t) s_{i+r}(t) \rangle - \langle s_i(t) \rangle^2$$

La dimensione media dei domini è dell'ordine

$$L \sim t^{\frac{1}{z}}$$

Per il modello di Ising si introducono le grandezze tempo necessario e densità iniziale x di spin + per raggiungere il consenso in un sistema grande N. $E(x)$ è la *exit probability*. La probabilità che data una configurazione iniziale con x spin up si raggiunge uno stato finale tutto fatto da spin up.

In dimensione 1

$$E(x) = x \qquad d = 1$$

$$E(x) = \theta\left(x - \frac{1}{2}\right) \qquad d > 1$$

$$T \sim N^{\frac{2}{d}}$$

Questo a $T = 0$.

Il modello di Voter. Questo modello ha una dimensione molto diversa, ed è un modello paradigmatico. Il parametro d'oridine (se siamo sopra o sotto la transizoine di fase). Questo parametro è dato dal numero di link attivi, tutti i link tra due nodi con due stati diversi.

$$\rho = \frac{1}{\sum_i k_i} \left(\sum_{\langle ij \rangle} \frac{1 - s_i s_j}{2} \right)$$

Ci sono delle varianti. Scegliamo prima l'individuo che copia o l'individuo che è copiato? In un reticolo bidimensionale non succede nulla, in una rete complessa cambia tutto, infatti cambia l'ordine. Oppure si può scegliere una coppia e scegliere con probabilità 1/2.

Rispetto al modello di Ising il modello di Voter genera bordi molto frastagliati. Il modello di Ising introduce una tensione superficiale che rende smussati i bordi, mentre nel modello di Voter. Il modo in cui si arriva a convergenza nel Voter e nel modello di Ising è completamente diverso. La exit probability per il modello di Ising è deterministica. la situazione $E(x) = x$ è presente a prescindere dalla dimensione.

Il modello Voter in una sola dimensione può essere trattato analiticamente.

In una dimensione può avvenire qualcosa solo alle interfaccie. possiamo dimenticarci del valore degli spin, e rappresentare il modellod i voter con delle palline che si muovono. Il modello di voter è dato dal moto di queste interfaccie, che possono spostarsi. Quindi il modello di Voter 1D corrisponde ad un camminatore aleatorio. Ogni volta che due camminatori si incontrano si annichilano.

Per capire quanto tempo impieghiamo basta capire quanto tempo impiegano i camminatori a annichilire.Quanto tempo ci mettono due camminatori a collidere?

$$\Delta t = l^2/D$$

La densità di camminatori cambia del numero di camminatori:

$$\Delta\rho \sim -\rho$$

$$\dot{\rho} = \frac{\Delta\rho}{\Delta t} \sim -D\rho^3$$

$$\rho(t) \sim \frac{1}{\sqrt{Dt}}$$

Questo è la stessa dinamica microscopica del modello di ising a temperatura zero.

A dimensione maggiore di 1 cambia drasticamente. Il sistema non si ordina ma rimane in uno stato stazionario nel limite termdinamico. Sul sistema finito in realtà esistono fluttuazioni sufficientemente grosse capaci di riordinare il modello.

6.1.1 Soluzione analitica al modello di Voter

Vogliamo calcolare un equazione funzione del tempo che il sistema si trovi in una configurazione $\rho(t)$

In ogni aggiornamento possiamo andare da ρ. La densità può o crescere o diminuire:

$$\rho \to \rho \pm \delta\rho$$

Dove $\delta\rho = \frac{1}{N}$. Sia $R(\rho)$ la probabiltià che la densità aumenti

$$R(\rho) = P\left[\rho \to \rho + \delta\rho\right] = \rho(1-\rho)$$

$$L(\rho) = P\left[\rho \to \rho - \delta\rho\right] = (1-\rho)\rho$$

Proviamo a scrivere l'equazione per la probabilità di trovare il sistema su uno stato ρ al tempo $t + \delta t$ in modo ricorsivo.

$$P(\rho, t+\delta) = P(\rho, t)\left[1 - R(\rho) - L(\rho)\right] + P(\rho - \delta\rho, t)R(\rho - \delta\rho) + P(\rho + \delta\rho, t)L(\rho + \delta\rho)$$

Sviluppiamo tutti i termini

$$P(\rho \pm \delta\rho, t) = P(\rho, t) \pm \frac{\partial P}{\partial \rho}\delta\rho + \frac{1}{2}\frac{\partial^2 P}{\partial \rho^2}\delta\rho^2 + O(\delta\rho^3)$$

$$P(\rho, t+\delta t) = P(\rho, t) + \frac{\partial P}{\partial t}\delta t$$

$$R(\rho - \delta\rho) = R(\rho) - \frac{\partial R}{\partial \rho}\delta\rho + \frac{1}{2}\frac{\partial^2 R}{\partial \rho^2}\delta\rho^2 + O(\delta\rho^3)$$

$$L(\rho + \delta\rho) = L(\rho) + \frac{\partial L}{\partial \rho}\delta\rho + \frac{1}{2}\frac{\partial^2 L}{\partial \rho^2}\delta\rho^2 + O(\delta\rho^3)$$

A questo punto possiamo sostituire all'interno:

$$P(\rho, t) + \frac{\partial P}{\partial t}\delta t = P(\rho, t) - P(\rho, t)L(\rho) - P(\rho, t)R(\rho) + A + B$$

$$A = \left[P(\rho, t) - \partial_\rho P\delta\rho + \frac{1}{2}\partial_\rho^2 P\delta\rho^2\right]\left[R(\rho) - \partial_\rho R\delta\rho + \frac{1}{2}\partial_\rho^2 R\delta\rho^2\right]$$

E analogamente il termine B con L, facciamo tutte le moltiplicazione tenendo tutto fino all'oridne $\delta\rho^2$ e δt.

$$\begin{aligned}\partial_t P\delta t = &-P(\rho, t)\partial_\rho R\delta\rho + \frac{1}{2}P(\rho, t)\partial_\rho^2 R\delta\rho^2 - \\ &- R(\rho)\partial_\rho P\delta\rho + \partial_\rho P\partial_\rho R\delta\rho^2 + \frac{R}{2}\partial_\rho^2 P\delta\rho^2 + P(\rho, t)\partial_\rho L\delta\rho + \\ &+ \frac{1}{2}P(\rho, t)\partial_\rho^2 L\delta\rho^2 + L(\rho)\partial_\rho P\delta\rho + \partial_\rho\partial_\rho L\delta\rho^2 + \frac{1}{2}L\partial_\rho^2 P\delta\rho^2\end{aligned}$$

Riunendo i termini in $\delta\rho$ e $\delta\rho^2$. Definiamo prima

$$V(\rho) = \left[R(\rho) - L(\rho)\right]\frac{\delta\rho}{\delta t}$$

$$D(\rho) = \frac{1}{2}\frac{\delta\rho^2}{\delta t}\left[R(\rho) + L(\rho)\right]$$

Riunendo tutti i conti si ottiene il risultato.

$$\partial_t P(\rho,t) = -\partial_\rho \left[V(\rho)P(\rho,t)\right] + \partial_\rho^2 \left[D(\rho)P(\rho,t\right]$$

Questa equazione è proprio l'equazione di Fokker -Plank. Il primo termine è detto termine di Drift, mentre il secondo è data dalla diffusione. Il primo termine spinge il consenso in una direzione. Nel modello di Voter la parte di drift è nulla, ed il modello è prettamente diffusivo. Il modello di Ising invece nel limite termodinamico la parte diffusiva tende a zero, mentre è dominante il termine di drift. Nel caso del Voter model il termine di Drift non c'è perché $R(\rho) = L(\rho)$. L'equazione di campo medio diventa:

$$\partial_t P(\rho,t) = \frac{1}{N}\partial_\rho^2 \left[\rho(1-\rho)P(\rho,t)\right]$$

Questo è il modello voter su una topologia infinitodimensionale, ogni punto del sistema può interagire con infiniti altri punti del sistema.

Per $\rho = 0$ e $\rho = 1$ abbiamo gli stati assorbenti del ssitema. Questa equazione si può riscrivere con y

$$y = 2\rho - 1$$

$$P(y,t) = \sum_{n=0}^{\infty} A_n C_n^{3/2} \exp\left[-\left(\frac{n(n+3)+2}{2N}\right)t\right]$$

Quello che ci interessa è la *exit probability*. Possiamo scrivere un'equazione ricorsiva: la probabilità di arrivare a 1 partendo da ρ è la stessa probabilità che se in un passo non succede nulla.

$$E(\rho) = E(\rho)\left[1 - R(\rho) - L(\rho)\right] + E(\rho+\delta\rho)R(\rho) + E(\rho-\delta\rho)L(\rho)$$

Sostituendo tutti i pezzi si ottiene:

$$V(\rho)\partial_\rho E(\rho) + D(\rho)\partial_\rho^2 E = 0$$

Nel modello di Vouther il primo termine nullo e otteniamo un equazione differenziale. Sappiamo come condizione iniziale

$$E(\rho=0) = 0 \qquad E(\rho=1) = 1$$

Siccome la derivata seconda è nulla quindi deve essere lineare:

$$E(\rho) = \rho$$

Il tempo per arrivare a convergenza è dato da:

$$T(\rho) = T(\rho)\left[1 - L(\rho) - R(\rho)\right] + R(\rho)T(\rho+\delta\rho) + L(\rho)T(\rho-\delta\rho) + \delta t$$

La soluzione diventa

$$V(\rho)\partial_\rho T + D(\rho)\partial_\rho^2 T = -1$$

Nel Vougher model il termine di drift è nullo. In questo caso le condizioni al contorno sono diverse

$$T(\rho=0) = 0 \qquad T(\rho=1) = 0$$

(Il tempo per convergere in uno stato assorbente è nullo (abbiamo già converso)

$$\partial_\rho T = -N \int d\rho \frac{1}{\rho(1-\rho)} \propto \ln\left(\frac{\rho}{1-\rho}\right)$$

Integrando un altra volta viene

$$T(\rho) = N\left[-\rho \ln \rho - (1-\rho)\ln(1-\rho)\right] = NS(\rho)$$

Ossia il tempo di convergenza proporzionale all'entropia della configurazione iniziale.

6.1.2 Soluzione al modello di Ising

Nel caso di Ising le R e L sonodiverse:

$$R(\rho) = \theta\left(\rho - \frac{1}{2}\right) = 1 - L(\rho)$$

$$V(\rho) = 1 - 2\theta\left(\frac{1}{2} - \rho\right) \qquad D(\rho) = \frac{1}{2N} \overset{N\to\infty}{\longrightarrow} 0$$

Il termine diffusivo nel limite termodinamico va a zero. Se vogliamo l'equazione della *exit probability*

$$V(\rho)\partial_\rho E(\rho) = 0$$

La soluzione di questa equazione è una costante

$$E(\rho) = 0 \qquad \rho < \frac{1}{2}$$

$$E(\rho) = 1 \qquad \rho > \frac{1}{2}$$

Per il Voughter model è possibile avere anche una soluzione analitica esatta.

Mettiamoci in d-dimensione. Uno stato del sistema determinato dalla variabile

$$S = \{s_i\}$$

Chiamiamo la configurazione S^k la stessa configurazione di S con il k-esimo spin invertito. Sia la probabilità che il k-esimo spin giri:

$$w_k(S) = w_k\left(S_k \to -S_k\right) = \frac{1}{2}\left(1 - \frac{1}{2d} s_k \sum_{nn} s_j\right)$$

Il numero di primiv nvicini è 2d

$$\partial P(s,k) = \sum_k \left[w_k(s^k)P(s^k,t) - w_k(s)p(s,t)\right]$$

Ossia la probabilità di trovarmi in un posto è quello di prendere una configurazione con il k-esimo spin girato e mi chiedo la probabilità di rigirare quello spin meno quella di girare un altro qualunque spin a partire dalla configurazione giusta.

Possiamo concentrarci in uno spin e ci chiediamo in funzione del tempo quanto vale.

$$\langle s_i \rangle = \sum_s s_i P(s,t)$$

$$\partial_t \langle s_i \rangle = \partial_t \sum_s P(s,t) s_i = \sum_s \sum_k \left[w_k(s^k) P(s^k,t) - w_k(s) P(s,t) \right] s_i$$

$$w_k(s) = w_k^0 + s_k w_k^1 \qquad w_k^0 = \frac{1}{2} \qquad w_k^1 = -\frac{1}{4d} \sum_n n s_j$$

$$\partial_t \langle s_i \rangle = \sum_s \sum_k \left[(w_k^0 - s_k w_k^1) P(s^k,t) - (w_k^0 + s_k w_k^1) P(s,t) \right] s_i$$

$$= \sum_{s\ k} \left[s_k w_k^0 \left(P(s^k,t) - P(s,t) \right) - w_k^1 s_i s_k \left(P(s^k,t) + P(s,t) \right) \right]$$

Se prendiamo questa somma:

$$\sum_s s_i w_k^0 \left(P(s^k,t) - P(s,t) \right)$$

Proviamo a vedere $k \neq i$, in questo caso è nullo perché i due spin sono diverse solo nello spin k , quindi tutte le configurazione con $k \neq i$ sono nulle. Quindi in realtà la sommatoria su k la possiamo buttare. Questo vale anche per la seconda sommatoria.

$$= \sum_s \left[s_i w_i^0 \left(P(s^i,t) - P(s,t) \right) - w_i^1 s_i s_i \left(P(s^i,t) + P(s,t) \right) \right]$$

Il primo termine mi da

$$\partial_t \langle s_i \rangle = -2 \langle s_i \rangle w_i^0 - 2 \langle w_i^1 \rangle = -2 \langle s_i w_i \rangle$$

$$\partial_t \langle s_i \rangle = \frac{\left[-2s \langle s_i \rangle + \sum_{nn} \langle s_j \rangle \right]}{2d} = \frac{1}{2d} \nabla_i^2 \langle s_i \rangle$$

Dove ∇_i^2 è il laplaciano discreto. Questa è una equazione diffusiva discreta. Ora vogliamo sapere il valore medio di $\langle s \rangle$

$$\langle s \rangle = \frac{1}{N} \sum_i \langle s_i \rangle$$

$$\partial_t \langle s \rangle = \frac{1}{N} \sum_i \partial_t \langle s_i \rangle = \frac{1}{N} \sum_i \left[-2d \langle s_i \rangle + \sum_{j \in nn} \langle s_j \rangle \right] = 0$$

Questa è la media del laplaciano discreto.

$$\langle s \rangle = cost$$

La magnetizzazione media del sistema rimane costante e dipende dalla configurazione iniziale. Cerchiamo la exit probability. Possiamo chiederci la densità del sistema.

Facciamo tante simulazioni in parallelo

$$\rho(0) = 1 \cdot E(\rho(0)) + 0 \cdot [1 - E(\rho(0))]$$

Il numero di domini è dato da:

$$n = \begin{cases} t^{\frac{1}{2}} & d = 1 \\ \frac{1}{lnt} & d = 2 \\ a - bt^{-\frac{d}{2}} & d > 2 \end{cases}$$

6.2 Modelli di dinamiche di opinione

Abbiamo una popolazione di individui, ogni idnividuo può trovarsi in un certo numero di stati. Esiste una certa interazione, e l'interazione può far cambiare idea alle persone.

I modelli possono essere classificati in base a come la scelta è discreta di opinioni, situazione di stati continui, o situazioni in cui si hanno un vettore di opinioni.

La connettività degli individui (detti agenti). Il tipo di rete è un reticolo, un grafo completo o uno scale free). Interazione tra individui. In tutti i modelli si parte dal principio dell'accordo. Ci sono delle situazioni di bounded confidence, ossia interagiscono solo due agenti sufficientemente simili, o ci avviciniamo solo se sono verificate condizone sul gruppo.

Iniziano a studiare anche situazioni in cui possono essere disaccordi o persone che si allontanano con effetto dell'interazione. Non esiste solo l'individui ma esiste anche un campo esterno (informazione) che crea una certa pressione. Questo modellizza l'effetto dei mass media.

Il majority ryle model è si prende un gruppo a caso nel modello e tutto il gruppo viene messo uguale alla maggioranza. Esiste una soglia critica in cui se mi trovo sopra arrivo a accordo positivo, sotto negativo.

Si possono calcolare i tempi di consenso.

Esiste il modello di Sznajd , in cui se due individui che sono in accordo tra loro convincono tutti i loro primi vicini. Il deffuant model è il primo modello in cui le variabili sono continue. Si scelgono due individui, si mette una soglia poer interagire e quando le persone interagiscono, e quando interagiscono le due opinioni si avvicniano di un fattore costante.

Quello che nascie da questo modello è che partire da questo modello si creano delle opinioni differenti.

Il modello di Axelrod è quello di provare a modellizzare in che modo interagiscono le culture. Se abbiamo tante culture in interazione che succede? Andremo verso situazioni di globalizzazione oppure no? L'agente è caratterizzato da un vettore di features, Ogni elemento del vettore può assumere q possibili valori. La probabilità con due individui possono interagire in modo proporzionale all'overlap di Features. Se interagiscono predono una qualunque features che avevano in comune la si pone uguale.

La distribuzione di cluster è una legge di potenza. Una transizione del secondo ordine le curve si sovrappongono a diverse dimensioni mentre per una al primo ordine diventa più tagliente per crescere di N.

6.2.1 Effetti dell'informazione esterna

Campagne di informazioni molto aggressiva tendono a portare a segregazione, e quindi non è benerifiante. Il primo modello è un modello di ising in campo random, un articolo molto bello di Bouchand. Ogni individuo ha una posizione binaria, in interazione con un certo numero di primi vicini. Il primo ingrediante è la propensione individuale. Sono numeri congelati che fanno parte della persiona (definito su tutto $\mathbb{R}$). Abbiamo un informazione esterna che può dipendere dal tempo $F(t)$ (anche questa può dipendere dal tempo.

Per ogni spin facciamo il segno di una somma fatta da tre comtributi:

$$S_i(t) = \text{sign}\left[\phi_i + F(t) + \sum_{j\in\nu_i} J_{ij}S_j(t-1)\right]$$

Come dipende l'opinione media da tutti questi parametri?
Cominciamo con un caso facile, individui indipendenti.

$$M = \frac{1}{N}\sum_i S_i = \frac{1}{N}\sum_i \text{sign}\left[\phi_i + F\right]$$

$$M = -\int_{-\infty}^{-F} d\phi p(\phi) + \int_{-F}^{\infty} d\phi p(\phi)$$

Dove $p(\phi)$ è la distribuzione di probabilità delle propensioni individuali

$$-R(-F) + [1 - R(-F)] = 1 - 2R(-F)$$

$$R(x) = \int_{-\infty}^{x} d\phi p(\phi)$$

ossiamo fare il caso un po più complicato del campo medio.

$$S_i(t) = \text{sign}\left[\phi_i + F(t) + \frac{J}{N}\sum_{j\in\nu_i} S(t-1)\right] = \text{sign}\left[\phi_i + F(t) + JM\right]$$

Possiamo ripetere i conti e ottenere una quantità:

$$M = 1 - 2R\left(-F - MJ\right)$$

Questa è un'equazione autoconsistente. Mettiamoci nel caso in cui gli individui sono poco accoppiati tra loro.

$$J \ll 1$$

$$1 - 2R(-F - MJ) = 1 - 2R(-F) + 2\frac{dR}{d\phi}(-F)JM = 1 - 2R(-F) + 2p(-F)JM$$

Sostituendo nell'equazione si ottiene.

$$M = \frac{M_0}{1 - 2p(-F)J}$$

Quello che si ottiene è che tanto più grande è il picco di p tanto più abbiamo un amplificazione della magnetizzazione globale del sistema.

Proviamo invece a fare il caso di J grande.

$$M = 1 - 2R(-F - JM)$$

Assumiamo di avere una gaussana come $p(x)$. Concentriamoci intorno ad $F = 0$ ($M \approx 0$)

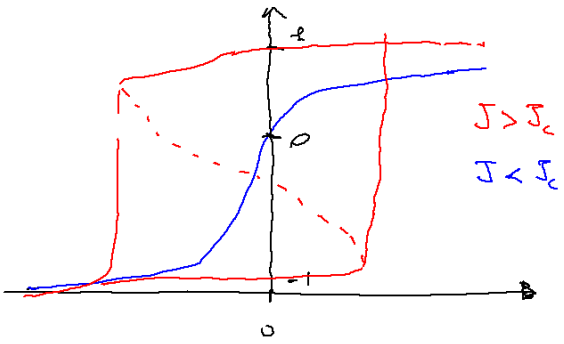
$$M = 1 - 2R(-JM) = 1 - 2R(0) + 2p(0)JM = 2p(0)JM$$

Quindi per avere intersezioni diverse da zero la pendenza della curva a destra deve essere maggiore. Abbiamo un valore critico per il coefficiente di accoppiamento che mi distingue due casi. Anche per F pari a zero esistono soluzioni per M:

$$2p(0)J_c = 1$$

$$J_c = \frac{1}{2p(0)}$$

Questo è il valore critico di accoppaimento. Sopra questo valore il sistema si magnetizza anche in assenza di campo esterno di informazioni.



Il processo assomiglia ad una curva di Isteresi. Per $J < J_c$ è interessante studiare come è fatta la derivata di $M(f)$, l'altezza di questa curcva scala con una larghezza di tipo:

$$h \sim w^{-\frac{2}{3}}$$

Vicino a J_c. Andiamo a vedere i dati reali. Questo esponente è dell'ordine di .73, un po' più alti dell'esponente di Ising.

Un altro modello più efficiacie è quello che contempla la possibilità di essere in disaccordo. Esiste un'informazione esterna, le variabili sono continue ed esistono diverse scelte per una particolare opinione. Esiste un informazione esterna, e gli individui possono o essere in accordo o in disaccordo e allontanarsi. Ogni in dividuo ha un vettore di probabilità per ogni scelta.

$$x = \begin{pmatrix} p_1 & p_2 & \cdots & p_n \end{pmatrix}$$

$$sum_{i=1}^{N} p_k = 1$$

Un modo per capire quato sono simili è detta il cosine similarity:

$$o^{ij} = \frac{x^i \cdot x^j}{|x^i||x^j|}$$

Dati due individui i e j vogliamo definire la probabilità che siano in accordo. La probabilità di accordo è proporzionale all'overlap (con una piccola correzione per evitare che a zero overlap gli individui non interagiscono. Se due individui sono in accordo si avvicninano altrimenti si allontanano, di un parametro α.

L'informazione eserna la identifichiamo come un individuo esterno. Ha un vettore con N scelte. l'unica differenza è che queste probabilità sono fissate. Con probabilità p_i ad ogni istante può interagire con questo individuo esterema.

In questo modo possiamo modellizzare sia la presenza di messaggi estremi o messaggi più modulati su una cosa più delicata. È importante cercare di partire con overlap differenti. Vogliamo sapere quanto è omogenea la popolazione.

Sembra che se il numero di scelte è alto è più facile far convergere il grafico. Il numero di cluster, partiamo da diverse condizione di overlap. Se l'overlap iniziale è alto andiamo verso un unico cluster.

Il punto di transizione è più spostato avanti se il numero di parametri sono grandi (sembra assurdo ma è così)

Si può cambiare anche quasi con continuità il ruolo dell'informazione. Possiamo vedere il numero di cluster che ci sono nel sistema. Quello che avviene è che con poco overlap l'informazione esterna non influenza minimamente il sistema. Se l'informazione esterna è aggressiva la gente tende ad allontanarsi dall'informazione esterna, se l'informazione è molto moderata è facile far convergere le persone verso la loro opinione.

Con informazione estrema ma non troppo presa a piccole dosi, ha un effetto di assunzione immediata (ma di rigetto a lungo termine).

6.3 Dinamica del linguaggio

Nel caso di dinamica del linguaggio la situazione è differente, rispetto ai modelli di opinion dynamics. Il sistema di comunicazione deve essere efficacie e efficiente. Tipicamente si vuole eliminare il dissenso. In questi modelli si vuole vedere come ha fatto la lingua ad emergere nella nostra specie? Nomi, categorie e strutture sintattiche.

Il linguaggio è un sistema che evolve costantemente e che quindi si autoorganizza. Le scale di tempo sono scale ti tempo culturali. Quello che avviene è che il tempo di vita medio è per i vari individui, esiste una scala di tempo biologiche, le informazioni sono trasmesse in modo verticale e orizzontali. La scala di tempo culturale è la trasmissione orizzontale. Il sistema di comunicazione del linguaggio deve essere stabile attraverso le generazioni. Il linguaggio dei segni Nicaraguanse è molto interessante, non esistevano scuole per bambini sordomuti, e ognuno aveva sviluppato un linguaggio dei segni locali, poi è stata aperta per la prima volta un centro per bambini sordomuti. Hanno preso i bambini da angoli del paesi e gli hanno messi tutti nella stessa scuola. Nell'arco di pochissimo tempo hanno sviluppato un nuovo linguaggio dei segni completamente nuovo che adesso è considerata una lingua, studiata come se fosse una lingua, questo nonostante il fatto che ci fosse una lingua esterna che si cercasse di imporre. Stanno cercando trasmetendo, è andato verso una semplificazione, abbiamo meno segni fondamentali che possono essere combinati.

Negli ultimi anni è nata la linguistica "in silico", usando i computer per fare analisi. Sia per quello che riguardano aspetti teorici di modellizzazione che aspetti sperimentali.

Vengono messi dei robot in una stanza, si poteva simulare una popolazione grande cambiando la struttura generica. ad ogni passo c'era uno speaker e un hearer che doveva descrivere una lavagna.

Si parte da un interazione di coppia. La telecamera deve guardare una lavagna segmentare tutto quello che vede. Bisogna concettualizzare. Lo speaker inveta una parola per indicare un oggeto. C'era un feedback elettronico che insegna all hearer l'oggetto "wabaku". Poi si cambia coppia di robot e così via

fino a convergenza. Si è andati avanti verso situazioni motlo più complicato si cerca di nominare oggetti che si trovano in determinate posizioni con i robot che hanno prospettive differenti della stessa scena.

Oppure abbiamo la Field simulation. Si portano in laboratorio individui e si sottopongonon a task linguistici. Ha fatto dei videogiochi.

Le domande che ci possiamo porre in questo caso sono se esistono requisiti minimi all'interno della popolazione perché emerga una struttura con delle carateristiche linguistiche condivise? Esiste uno stato stazionario o abbiamo una dinamica molto lenta? La taglia del sistema che ruolo ha? La topologia?

Vogliamo capire come funziona un fenomeno collettivo a partire dalla conoscenza del fenomeno elementare.

6.3.1 The naming game

Questo modello è il problema di avere N individui che devono associare un nome a degli oggetti. Abbiamo una popolazione di N individui. Ogni individuo è caratterizzato da una lista di associazioni tra nomi ed oggetto. Per ora ci limitiamo al caso in cui l'oggetto è solo uno (mettiamo da parte il problema dell'omonimia). Gli individui vogliono condividere un lessico condiviso. Non ci sono i cheaters (quelli che imbrogliano, dicono A per farsi capire B). Esiste l'interazione tra coppie, speaker e hearer. Le interazioni sono locali e diadi. Esiste l'oggeto, lo speaker identifica l'oggetto e associa un nome. Può o inventare un nome nuovo se non ha nessun nome in quell'oggetto, oppure avere una scelta casuale tra diverse alternativa. Una volta detto il nome lo pronunciamo e la persona cerca di fare l'associazione tra il nome e l'oggetto. L'interazione può essere un fallimento, ma dopo l'interazione viene registrata dall'healer, esiste una persona che associa la parola vecchia nel nuovo repertorio.

L'interazione può avere successo. Se lo speaker pronuncia una parola che l'hearer ha già nel suo repertorio, entrambi gli agenti cancellano tutte le parole. I repertori inizialmente sono vuoti.

Quanto può crescere l'inventory?

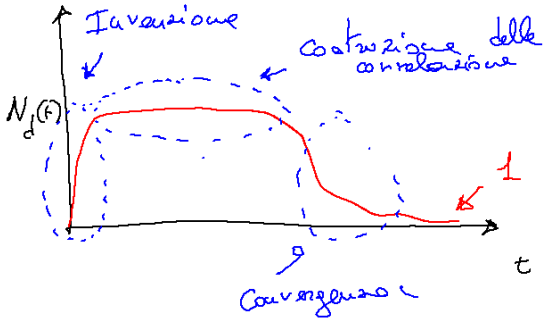


Figura 6.1: Numero di parole memorizzate per ogni agente

Possiamo chiederci quanto tempo ci vuole per arrivare a convergenza.

$$\frac{dN_w(t)}{dt} \approx \left(1 - \frac{2cN_A^\alpha}{N_A}\right) - \underbrace{\frac{2cN_A^\alpha}{N_A}}_{\text{Probabilità di successo}} 2cN_A^\alpha$$

Possiamo trovare il massimo

$$\frac{dN_w(t)}{dt} = 0 \qquad \alpha = \frac{1}{2}$$

Anche il tempo

$$t_{max} \sim N_A^\beta$$

Si può riscrivere

$$\frac{dN_w(t)}{dt} \approx \left(1 - \frac{kt}{N_A^2}\right) - \underbrace{\frac{kt}{N_A}^2}_{\text{Probabilità di successo}} 2cN_A^{\frac{1}{2}}$$

$$\beta = \frac{3}{2}$$

Dove abbiamo modellizzato la probabilità di successo come kt/N_A^2. Il fatto che cambia la dimensione del dizionario. Possiamo vedere che succede per lo spazio delle parole. Possiamo calcolare la frequenza con cui vengono dette determinate parole, facciamo il grafico di Zipf. Tutto va come una condensazione di Bose-Einstein.

Si può fare un network e collegare individui che condividono una parola.

Che succede in funzione della topologie, in cui si ha una rete di interazione differenti. Ogni nodo è un agente ogni link un interazione. Fino ad ora abbiamo ragionato in una rete semplicemente connessa. Il sistema converge sempre. Quello che può succedere è che i tempi di scaling possono dipendere dalla topologia di reti.

Su reticoi regoari il tempo va come

$$t_{conv} \sim N^\beta \qquad \beta = 1 + \frac{2}{d}$$

Il nomero di parole totale del sistema nel repertorio di ogni individuo avremo un numero finito del sistema.

Con un grafo small-world da un convergenza veloce e una memoria piccola. Questa sembra una situazione ideale.

Possiamo introdurre un parametro β che ci dice che quando due agenti hanno successo solo con probabilità β si cancella la memoria di entrambi.

Possiamo chiederci come cmbiano il numero di persone che conosce A, B e entrambi i termini:

$$\frac{dn_A}{dt} = -n_A n_b + \beta n_{AB}^2 + \frac{3\beta - 1}{2} n_A n_{AB}$$

$$\frac{dn_A}{dt} = -n_A n_b + \beta n_{AB}^2 + \frac{3\beta - 1}{2} n_B n_{AB}$$

Possiamo chiederci quali sono i punti fissi. Scopriamo che ci sono 3 punti fissi, oppure esiste una situazione di coesistenza:

$$n_A = n_B = b(\beta) \qquad n_{AB} = 1 - 2b(\beta)$$

$$\frac{d(n_A - n_B)}{dt} = \frac{3\beta - 1}{4}(n_A - n_B)n_{AB}$$

Quindi a seconda del segno di questa derivata dipende in che valore misto arriviamo. E si vede che il valore critico è

$$\beta_c = \frac{1}{3}$$

Possiamo fare la matrice di stabilità del consenso critico. Per la situaizone di consenso si ottiene un valore:

$$\lambda_3 = -\frac{3}{2}(\beta - \beta_c)$$

Quindi il punto fisso diventa istabile nel caso in cui gli autovalori sono positivi.

$$\frac{d}{dt}\begin{pmatrix}\delta n_a \\ \delta n_b\end{pmatrix} = \begin{pmatrix}\partial_{n_a}\dot{n}_a & \partial_{n_b}\dot{n}_a \\ \partial_{n_a}\dot{n}_b & \partial_{n_b}\dot{n}_b\end{pmatrix})\begin{pmatrix}\delta n_a \\ \delta n_b\end{pmatrix}$$

Se studiamo il tempo di convergenza andiamo in una situazion in cui coesistono sempre più condizioni nel sistema. Sotto il valore $\beta = \frac{1}{3}$ abbiamo altre transizione sulla coesistenza di più parole. Questo sembra che esistano diverse convenzione.

6.3.2 Lingue creole

Le lingyue creole sono nate dal contatto di tante lingue preesistenti. Le lingue creole è dato dalla lingua genitrici, la grammatica e la sintassi ha una struttura originale. Esistono tutti una serie di statiche in alcuni casi. Abbiamo bisogni di una struttura molto particolare della popolazione. Si mettono in comune una serie di schiavisti e colonarizzatore, abbiamo una serie di persone intermedie, mulatti, che fanno da intermdiazione tr aschiavi e colonialisti. Esistono tre abbiamo pochissime interazione tra padroni e schiavi unidirezionale, tra mulatti e schiavi bidirezzionale.

6.3.3 Category game

La categorizzazione è un nome aperto per avere una categoria. Nominare i colori è interessanti perché abbiamo uno spettro continui, bisogna creare una partizione all'interno dello spettro a cui assegnare i nomi. Per ogni individuoi abbiamo uno spazio reale da 0 a 1. Ogni istante giochiamo una specie di naming game, vengono proposti due topic (due numeri casuali). Se lo speaker prima crea una partizione tra a e b e poi pronuncia il nome per quella partizione. e l'hearer deve capire dal nome che colore ha scelto lo speaker.

La categorizzazione dipende molto dall'ambiente esposto.