Analisi di Sequenze di Dati: complementi

E. Caglioti

30 maggio 2010

1 Catene di Markov: Teorema H e convergenza all'equilibrio

Una proprietà generale delle catene di Markov è il fatto che l'entropia relativa tra due misure di probabilità che si evolvono con la stessa catena di Markov diminuisce (non cresce). Da questa proprietà discende anche, sotto opportune ipotesi sulla matrice di transizione, l'esistenza di una unica misura limite e la convergenza ad essa.

Teorema 1.1 *Teorema H.*
Consideriamo una catena di Markov a stati discreti con matrice di transizione A ad elementi tutti positivi. Siano p e q due misure di probabilità e $q' = A^T q$, $p' = A^T p$ le loro evolute. Allora

$$D(p'|q') \leq D(p|q),$$

dove il segno di equaglianza si ha solo se $p = q$.

Dimostrazione. Ricordiamo che le probabilità evolvono con la trasposta della matrice di transizione A : cioè

$$p'_i = \sum_j A_{j,i} p_i.$$

Definiamo inoltre $\phi(x) = x \log x$.
Allora

$$
\begin{aligned}
D(p'|q') &= \sum_i p'_i \log \frac{p'_i}{q'_i} \\
&= \sum_i q'_i \frac{p'_i}{q'_i} \log \frac{p'_i}{q'_i} \\
&= \sum_i q'_i \phi\left(\frac{p'_i}{q'_i}\right) \\
&= \sum_i q'_i \phi\left(\sum_j \frac{1}{q'_i} A_{j,i} p_j\right) \\
&= \sum_i q'_i \phi\left(\sum_j \frac{p_j}{q_j} \frac{q_j A_{j,i}}{q'_i}\right) \\
&\leq \sum_i q'_i \sum_j \phi\left(\frac{p_j}{q_j}\right) \frac{q_j A_{j,i}}{q'_i} \\
&= \sum_j q_j \phi\left(\frac{p_j}{q_j}\right) \equiv D(p'|q)
\end{aligned}
$$

Nella sesta riga abbiamo applicato Jensen utilizzando il fatto che, per ogni i,

$$\sum_j \frac{q_j A_{j,i}}{q'_i} = 1;$$

mentre nella settima abbiamo sommato su i.

Inoltre, essendo ϕ strettamente convessa e q_i positiva for any i, l'eguaglianza vale solo se $p_i = cq_i$ per ogni i ed ovviamente, essendo p e q misure di probabilità la $c = 1$.

1.1 Convergenza all'equilibrio

Possiamo notare che come immediato corollario del Teorema precedente otteniamo che se q è stazionaria: cioè se $q = A^T q$, allora l'entropia relativa tra $(A^T)^t p$ e q decresce con t (a meno che per qualche t non sia $p = q$).

Sembra quindi ragionevole che una misura di probabilità che evolve secondo una catena di Markov tenda per $n \to \infty$ all'equilibrio e che questo equilibrio sia unico. Qui dimostreremo questo teorema che sarà essenzialmente un corollario del Teorema precedente.

Teorema 1.2 *Convergenza all'equilibrio.*

Consideriamo una catena di Markov a stati finiti con matrice di transizione A ad elementi tutti positivi. Sia p la misura di probabilità iniziale e sia

$$p_t = (A^T)^t p$$

la misura al tempo t.

Allora,

$$\lim_{t\to\infty} p_t = \tilde{p},$$

dove $\tilde{p}$ è l'unica soluzione di

$$\tilde{p} = A^T \tilde{p}.$$

Dimostrazione.
Per il Teorema H sappiamo che

$$D(p_t|p_{t+1}$$

è una sequenza non crescente. Quindi, dato che è anche limitata dal basso, esiste

$$\bar{D} = \lim_{t\to\infty} D(p_t|p_{t+1}.$$

Sia quindi (P, P') un punto di accumulazione per la coppia p_t, p_{t+1}.

Per la continuità dell'entropia relativa deve essere

$$\bar{D} = D(\bar{p}, \bar{p'}).$$

Vogliamo far vedere che $\bar{D} = 0$ e che quindi $P = P'$ (notiamo infatti che $\bar{D} = 0$ solo se $P = P'$).

Supponiamo per assurdo che non lo sia.

Allora, chiamando $Q = A^T P$ e $Q' = A^T P'$ troviamo

$$D(Q|Q') < D(P|P').$$

Ma sia Q che Q' sono punti di accumulazione di (p_t, p_{t+1}), quindi troviamo

$$\bar{D} = D(Q, Q') < D(P, P') = \bar{D}$$

il che è assurdo.

2 Filogenesi a partire da matrici di distanze

Diamo qui due definizioni equivalenti di matrice di distanze additive.

2.1 Distanze additive

Definizione 2.1 *Additività.*
Una matrice di distanze $d_{i,j} : i, j \in \{1...n\}$ si dice additiva se esiste un albero T sul quale, per ogni coppia di foglie i, j, la distanza su T tra i e j coincide con $d_{i,j}$.

Definizione 2.2 *Condizione a Quattro Punti.*
Una matrice di distanze $d_{i,j} : i, j \in \{1...n\}$ si dice soddisfare la condizione a quattro punti se, per ogni quartetto di foglie i, j, k, l, chiamando $D_1 = d_{i,j} + d_{k,l}$, $D_2 = d_{i,k} + d_{j,l}$, $D_3 = d_{i,l} + d_{j,k}$, vale una delle tre seguenti:

- $D_1 < D_2 = D_3$
- $D_2 < D_1 = D_3$
- $D_3 < D_1 = D_2$

Teorema 2.1 *Una matrice di distanze è additiva se e solo se soddisfa alla condizione a quattro punti.*

2.2 Neighbour Joining

Dimostriamo che l'algoritmo Negihbor-Joining ricostruisce esattamente la filogenesi quando la matrice è additiva. L'algoritmo iterativamente sceglie una coppia, la sostituisce con un altra foglia, di cui ricalcola le distanze, fino a quando non si rimangono solo 3 foglie.

Per dimostrare qundi che l'algortitmo è corretto basta dimostrare che la coppia di foglie che minimizza $D_{i,j}$ è un *cherry*, cioè una coppia di foglie che sono attaccate allo stesso nodo.

Infatti, una volta che ciò sia assicurato, possiamo notare che l'operazione di sostituzione delle due foglie con una foglia trasforma la matrice additiva in una matrice additiva.

Teorema 2.2 *La coppia i, j che minimizza $D_{i,j}$ é un cherry.*

Dimostrazione. Prima di tutto definiamo, per ogni nodo, k, la quantità

$$Q_k = \sum_i d_{k,i} \tag{2.1}$$

cioè la somma delle distanze di questo nodo con tutte le foglie.

Data una foglia i definiamo g_i il nodo genitore di i.

Valgono i due seguenti Lemmata.

Lemma 1. Siano i e j due foglie. Allora

- se $g_i == g$,
$$-(N-2)D_{i,j} = Q_g$$
- altrimenti,
$$-(N-2)D_{i,j} < Q_{g_1} + Q_{g_2}$$

Lemma 2. Il massimo di Q_k al variare di k sui nodi interni è preso su di un nodo che e' collegato a due foglie.

Dimostriamo prima di tutto che come conseguenza del Lemma 1 e del Lemma 2 otteniamo il Teorema. Sia infatti k il nodo interno che massimizza Q_k che per il Lemma 2 sappiamo essere un nodo collegato a due foglie. Chiamiamo queste due foglie $\bar{i}$ e $\bar{j}$. Per il Lemma 1 sappiamo che $-(N-2)g_{\bar{i},\bar{j}} = Q_k$. Dobbiamo quindi far vedere che per ogni altra coppia i, j tali che $g_i \neq g_j$ si ha $D_{i,j} > D_{\bar{i}.\bar{j}}$. Ciò si ottiene facilmente notando che, sempre per il Lemma 1, se $g_i \neq g_j$, allora

$$-(N-2)D_{i,j} < \frac{Q_{g_i} + Q_{g_j}}{2} \leq \max_k Q_k = Q_{\bar{k}} = -(N-2)D_{\bar{i},\bar{j}}$$

Dimostrazione Lemma 1. Prima di tutto notiamo che data una foglia i la sua distanza da una foglia f, $f \neq i$ soddisfa

$$d_{i,f} = d_{g_i,f} + d_{i,g_i},$$

mentre se $i = f$ ovviamente

$$0 = d_{i,i} = d_{g,i} - d_{g,i}.$$

Sommando quindi su tutte le foglie otteniamo

$$Q_i = Q_f + (N-2)d_{i,g_i}.$$

Se consideriamo quindi due foglie i, j troviamo

$$Q_i + Q_j = Q_{g_i} + Q_{g_j} + (N-2)d_{i,g_i} + (N-2)d_{j,g_j}. \tag{2.2}$$

D'altronde,

$$-(N-2)D_{i,j} = \sum_f d_{i,f} + \sum_f d_{j,f} - (N-2)d_{i,j} = Q_i + Q_j - d_{i,j} \tag{2.3}$$

Quindi, utilizzando le (2.2), (2.3) otteniamo

$$-(N_2)D_{i,j} = Q_i + Q_j - (N-2)d_{i,j} = Q_{g_i} + Q_{g_j} - (N-2)d_{i,j} + (N-2)d_{i,g_i} + (N-2)d_{j,g_j} \tag{2.4}$$

Il Lemma si ottiene quindi notando che se $g_i = g_j$ allora $d_{i,j} = d_{i,g_i} + d_{j,g_j}$, mentre se $d_{i,j} < d_{i,g_i} + d_{j,g_j}$.

Dimostrazione Lemma 2. Dato un nodo k siano k_1, k_2, k_3 i tre nodi cui k è collegato, e siano T_1, T_2, T_3 i 3 alberi che hanno come nodo genitore k_1, k_2, k_3 rispettivamente. Ovviamente l'albero T_i potrebbe consistere del solo nodo k_i. Qui vogliamo appunto dimostrare che per il nodo k che massimizza Q_k, due tra k_1, k_2, k_3 sono foglie. Chiamiamo N il numero di foglie dell'albero T ed N_i il numero di foglie dell'albero T_i. Si verifica facilmente che per ogni $i = 1, 2, 3$,

$$Q_k = Q_{k_i} - (N - 2N_i)|T_i|d_{k,k_i} \quad , \tag{2.5}$$

infatti le foglie dell'albero T_i sono più vicine a al nodo k_i di una distanza d_{k,k_i} mentre per le foglie non contenute in T_i vale il contrario. Dato che $N_1 + N_2 + N_3 = N$ possiamo notare che almeno uno di questi è minore o uguale di $N/3$ e quindi minore di $N/2$. Quindi, chiamando i quel nodo troviamo $Q_k < Q_{k_i}$ il che ci porta ai seguenti tre casi:

- se k è collegato solo a nodi interni allora non puó massimizzare Q_k;
- se k è collegato ad una sola foglia (e quindi anche a due nodi interni) allora, chiamando senza perdita di generalità k_1 la foglia troviamo $N_1 = 1$. Ciò vuol dire che $N_2 + N_3 = N - 1$ e che quindi almeno uno tra N_1 ed N_2 è minore di $N/2$. Quindi, di nuovo Q_k non può essere massimo.
- Rimane quindi solo il caso in cui k sia collegato a due foglie che era quanto volevamo dimostrare.

2.3 Formula di Pauplin

Teorema 2.3 *Formula di Pauplin*
Sia T con n foglie e sia L la lunghezza totale dell'albero, cioè la somma delle lunghezze dei suoi legami. Allora vale la seguente formula

$$L = \sum_{a,b} 2^{-t_{a,b}} d_{a,b} \tag{2.6}$$

dove $t_{a,b}$ é il numero di nodi che bisogna attraversare per andare dalla foglia a alla foglia b.

Dimostrazione. Dato un legame k chiamiamo l_k la sua lunghezza. Chiamiamo $W_{a,b}$ l'insieme dei legami sul cammino tra la foglia a e la foglia b. Chiamando

P la somma di Pauplin abbiamo

$$\begin{aligned} P = & \quad \sum_{a,b} 2^{-t_{a,b}} d_{a,b} && (2.7)\\ = & \quad \sum_{a,b} \sum_{k \in W_{a,b}} 2^{-t_{a,b}} l_k && (2.8)\\ = & \quad \sum_k l_k \sum_{a,b} 1_{k \in W_{a,b}} 2^{-t_{a,b}} && (2.9)\\ & && (2.10) \end{aligned}$$

Possiamo ora notare che se il legame k è in $W_{a,b}$ allora a si trova da una parte di k mentre b è dall'altra. Quindi. a meno di rinominare le foglie, possiamo ottenere che le a sono sempre da una parte (la parte sinistra per esempio), mentre le b sono sempre dall'altra (quella destra). Quindi

$$\sum_{a,b} 1_{k \in W_{a,b}} 2^{-t_{a,b}} = \sum_{a \in sinistra} \sum_{b \in destra} 2^{-t_{a,b}} \tag{2.11}$$

dove abbiamo chiamato S l'albero che ha come radice il nodo sinistro del legame k (che chiamiamo s) e D l'albero che ha come radice il nodo destro del legame k che chiamiamo d.

Possiamo allora notare che

$$t_{a,b} = S_a + D_b$$

dove S_a è la profondità del nodo a nell'albero S e dove D_b è la profondità del legame b nell'albero D. Allora

$$\sum_{a \in S} \sum_{b \in D} 2^{-t_{a,b}} = \sum_{a \in S} \sum_{b \in D} 2^{-S_a - D_b} = \sum_{a \in S} 2^{-S_a} \sum_{b \in D} 2^{-D_B} = 1 \cdot 1 = 1 \tag{2.12}$$

grazie all'eguaglianza di Kraft !

Un altra importante proprietà della lunghezza di Pauplin è che essa, vista come funzione della topologia t, è minima sull'albero vero.

Qui ci limiteremo a mostrare che la lunghezza di Pauplin è minimo locale sull'albero giusto.

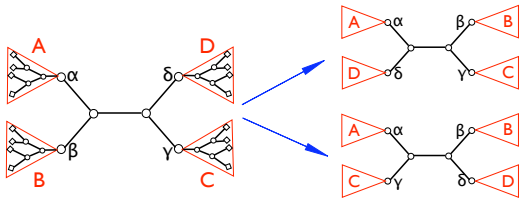
Per fare ciò dobbiamo prima introdurre il concetto di locale.

Definizione 2.3 *Mosse ammissibili.*

Dato un albero T consideriamo un suo legame k che separa in due parti l'albero in modo che da ogni parte del legame vi siano almeno due foglie. Chiamiamo s e d i nodi connessi dal legame k. Chiamiamo α, β, i figli di s; γ, δ, i figli di d.

Chiamiamo A e B i due alberi figli di s, e C e D i due alberi figli di d.

Allora una mossa di tipo Q è una mossa per la quale la struttura interna di A, B, C, D rimane inalterata, ma in cui A è collegato con C e B è collegato con D. Ovviamente si può anche collegare A con D e B con C.



Vale allora il seguente teorema.

Teorema 2.4 *Sia T l'albero esatto e sia T' un qualunque albero che si può ottenere a partire dall'albero T con una sola mossa di tipo Q. Allora*

$$P(T') > P(T)$$

Dimostrazione. Sia k il legame che tagliamo e siano A, B, C, D, i sotto alberi definiti sopra. Dato che T è l'albero esatto, per ogni $a \in A, b \in B, c \in C, d \in D$ vale

$$d_{a,b} + d_{c,d} < d_{a,c} + d_{b,d} = d_{a,d} + d_{b,c}.$$

Consideriamo la mossa per la quale A viene collegato con C e B viene collegato con D ottenendo così l'albero T'. Definiamo per $X \neq Y \in \{A, B, C, D\}$

$$P_{X,Y} = \sum_{x \in X, y \in Y} d_{x,y} 2^{-l_x - l_y}$$

dove l_x ed L_y sono rispettivamente le profondità di x ed y nel sottoalbero X e nel sottoalbero Y. Chiamiamo inoltre $P(X, X)$ per $X \in \{A, B, C, D\}$ la lunghezza di Pauplin per il sottoalbero X :

$$P_{X,X} = \sum_{x_1 \in X, x_2 \in X} d_{x_1,x_2} 2^{-t_{x_1,x_2}}.$$

Possiamo allora notare che

$$P(T) = P(A)+P(B)+P(C)+P(D)+\frac{1}{2}(P_{A,B} + P_{C,D})+\frac{1}{4}(P_{A,C} + P_{B,D} + P_{A.D} + P_{B,C}) \tag{2.13}$$

mentre

$$P(T') = P(A)+P(B)+P(C)+P(D)+\frac{1}{2}(P_{A,C} + P_{B,D})+\frac{1}{4}(P_{A,B} + P_{C,D} + P_{A.D} + P_{B,C}) \tag{2.14}$$

Sottraendo $P(T)$ da $P(T')$ troviamo

$$P(T') - P(T) = \frac{1}{4}(P_{A,C} + P_{B,D} - P_{A.B} + P_{C,D}) \tag{2.15}$$

Concludiamo facendo vedere che $P_{A,C}+P_{B,D}-P_{A.B}-P_{C,D}$ è la combinazione convessa di $d_{a,c}+d_{b,d}-d_{a,b}-d_{c,d}$ con opportuni pesi e quindi è positiva.

In formule

$$P_{A,C}+P_{B,D}-P_{A.B}-P_{C,D} = \sum_{a\in A, b\in B, c\in C, d\in D} 2^{-l_a-l_b-l_c-l_d}\left(d_{a,c}+d_{b,d}-d_{a,b}-d_{c,d}\right) < 0, \tag{2.16}$$

dove abbiamo usato il fatto che $\sum_{x\in X} 2^{-t_x} = 1$, e quindi, per esempio, che

$$P_{A,C} = \sum_{a\in A, c\in C} 2^{-t_a+t_c} d_{a,c} = \sum_{a\in A, b\in B, c\in C, d\in D} 2^{-l_a-l_b-l_c-l_d} d_{a,c}$$

dato che le somme su b e d danno 1.