UNIVERSITÀ DEGLI STUDI DI FIRENZE

FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI
Corso di Laurea Specialistica in Matematica.

TESI DI LAUREA SPECIALISTICA

Metodi di Geometria Algebrica per la ricostruzione statistica di alberi filogenetici

Candidato:
Luigi Biondi

Relatori:
Prof. Giovanni Marchetti

Prof. Giorgio Ottaviani

Anno Accademico 2010-2011

Indice

Elenco delle figure

Introduzione

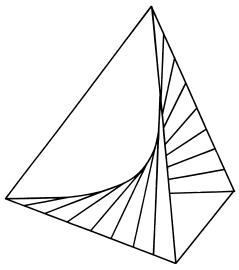
L'idea alla base di questo lavoro di tesi può essere riassunta efficacemente in una frase che S. Sullivant in [22] definisce *il mantra della statistica algebrica*:

i modelli statistici sono varietà algebriche.

Per capire il significato di questa espressione un po' criptica prendiamo come esempio la relazione di indipendenza fra due variabili aleatorie discrete X e Y a valori rispettivamente in $\{1,\dots,n_1\}$ e $\{1,\dots,n_2\}$:

$$\mathbf{P}[X=i,Y=j]=\mathbf{P}[X=i]\mathbf{P}[Y=j]=$$
$$=\left(\sum_{h=1}^{n_2}\mathbf{P}[X=i,Y=h]\right)\left(\sum_{k=1}^{n_1}\mathbf{P}[X=k,Y=j]\right)\qquad i=1,\dots,n_2\, j=1\dots n_2.$$

Ora, una distribuzione di probabilità discreta è caratterizzata dai valori $p_{ij}=\mathbf{P}[X=i,Y=j]$ delle probabilità di tutti possibili stati: possiamo perciò vedere l'insieme di tutte le distribuzioni per due variabili (con un numero di stati fissato a priori) come il sottoinsieme di uno spazio affine reale di dimensione $n_1\times n_2$ dei punti a coordinate non negative e a somma 1 (*simplesso di probabilità*).
L'insieme delle distribuzioni in cui X e Y sono indipendenti corrisponde quindi al luogo degli zeri dei polinomi $p_{ij}-\left(\sum_{h=1}^{n_2}p_{ih}\right)\left(\sum_{k=1}^{n=1}p_{kj}\right)$, al variare di i e j.



Superficie delle distribuzioni di coppie di variabili indipendenti nel caso $n_1=n_2=2$.

Da questo esempio si capiscono quindi quali siano le motivazioni dietro lo sviluppo negli ultimi anni della cosiddetta *statistica algebri-*

ca, disciplina che coniuga la statistica, il calcolo delle probabilità e la geometria algebrica.
La statistica algebrica studia le varietà algebriche definite dai modelli statistici, associando a vincoli come l'indipendenza e l'indipendenza condizionata degli ideali. Grazie a questo approccio possiamo, per usare ancora una volta le parole di Sullivant, "ricondurre problemi statistici a problemi di geometria algebrica": questa operazione non fornisce necessariamente soluzioni, ma già avere un nuovo punto di vista può dare nuove idee per trovarle.

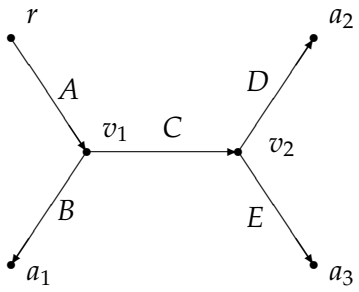
Nella prima parte di questo lavoro (capitoli 1 e 2) daremo le basi per la traduzione nel linguaggio della geometria algebrica delle proprietà delle distribuzioni di probabilità: troveremo nei prodotti tensoriali uno strumento per rappresentare le distribuzioni congiunte, grazie alla possibilità di trasformare l'indipendenza in una condizione sul rango di alcune particolari matrici dette *flattening*.
Una volta trovato un modo di associare ai concetti basilari del calcolo delle probabilità degli ideali di un anello di polinomi, saremo in grado di descrivere in termini geometrici lo spazio delle distribuzioni che soddisfano le condizioni poste da modelli statistici, con particolare riferimento ai modelli grafici.
Otterremo perciò una corrispondenza fra modelli statistici e varietà algebriche, ritrovando in molti casi varietà già note, come le *varietà di Segre* e le *varietà secanti*. Poter utilizzare i risultati conosciuti da tempo in geometria fornisce uno stimolo ulteriore a proseguire in questa direzione.

La seconda parte (capitoli 3 e 4) sarà invece interamente dedicata allo studio delle *varietà filogenetiche*, la prima applicazione concreta della teoria fin qui sviluppata. Buona parte di quanto presentiamo è tratta dai lavori di E. Allman e J. Rhodes ([2] del 2003 e [3] del 2008), che a loro volta riprendono idee introdotte per la prima volta da J. Cavender e J. Felsenstein in [5] nel 1987 e, indipendentemente, da J.A. Lake in [16], sempre nel 1987.
Il problema in esame è la ricostruzione della storia evolutiva di n specie viventi a partire da un allineamento di stringhe del loro DNA. Ciò

che noi vogliamo ottenere è una struttura grafica (*albero filogenetico*) in cui le specie siano tanto più vicine fra di loro quanto lo sono evolutivamente.
Per prima cosa sarà quindi necessario studiare i possibili modelli "candidati": ogni albero binario definisce un insieme di relazioni di indipendenza e indipendenza condizionata fra le variabili aleatorie associate ai suoi vertici, quindi anche una varietà algebrica che chiameremo **Varietà Filogenetica**.
Daremo all'inizio una descrizione parametrica di un insieme che genera questa varietà: ad ogni arco dell'albero assoceremo una matrice (detta *di transizione*) i cui elementi sono le probabilità di passaggio fra gli stati delle variabili associate ai suoi estremi. Ciò che otterremo saranno dei polinomi omogenei negli elementi di queste matrici:

$$p_{i_1 \dots i_n} = \sum_H \pi_{b_r} M_{e_1}(b_{v_1}, b_{v_2}) \dots M_{e_m}(b_{v_{k-1}}, b_{v_k}),$$

dove H è l'insieme delle stringhe $(b_{v_1}, b_{v_2}, \dots, b_{v_k})$, tutti gli stati delle variabili (comprese quelle interne) compatibili con gli stati $i_1, \dots, i_n$ delle n variabili relative alle specie in esame. Dunque, al variare delle matrici di transizione e della probabilità sul vertice iniziale, otteniamo tutte le possibili distribuzioni definite dall'albero, secondo un modello chiamato *Modello generale di Markov*.



Albero filogenetico binario con 4 foglie. Le lettere maiuscole indicano le matrici di transizione.

Il passaggio successivo sarà quello di trovare delle equazioni che caratterizzino la varietà filogenetica, dette *invarianti filogenetici*. Vedremo che, accorpando vertici, possiamo considerare alberi complicati

come strutture più semplici corrispondenti ai casi di base: a partire dalle equazioni per le varietà secanti alle varietà di Segre saremo in grado di trovare gli invarianti per ogni varietà filogenetica.
Presenteremo poi due algoritmi per scegliere l'albero filogenetico che più si adatta alle sequenze di DNA a nostra disposizione: il primo fa un uso diretto degli invarianti, valutandoli per ogni possibile albero binario e scegliendo quello le cui equazioni sono "meglio soddisfatte" dai dati. Il secondo, descritto da N. Eriksson in [9], adotta un approccio semplificato, cercando di ricostruire gli *split* dell'albero, ovvero le partizioni dell'insieme delle variabili ottenuti dalla cancellazione di un arco: per fare ciò, dovremo stabilire quando certi flattening hanno rango piccolo, obiettivo che raggiungeremo sfruttando le proprietà della fattorizzazione SVD.
Concluderemo il nostro lavoro (capitolo 5) presentando i risultati dell'implementazione in MATLAB di questi due metodi.
Prima ne verificheremo la correttezza, applicandoli a dati ottenuti simulando una distribuzione costruita a partire da un albero noto: avremo così modo di stabilire se l'algoritmo ha ricostruito correttamente o no l'albero giusto.
Infine, useremo il secondo algoritmo per analizzare dei dati reali che ci sono stati cortesemente forniti dal Dipartimento di biologia evoluzionistica dell'Università di Firenze: confronteremo quindi i risultati ottenuti dal nostro metodo con quelli che fornisce il software MEGA, normalmente utilizzato dai biologi anche per la ricostruzione filogenetica con i metodi di massima verosimiglianza, neighbour joining e minima evoluzione.

Notazione

- $[n] = \{1, \ldots, n\}$.
- $\Delta_n = \left\{(x_1, \ldots, x_n) \in \mathbb{R}^n : \sum_{i=1}^n x_i = 1,\ x_i \geq 0\ \forall i\right\}$.
- $V(F) = \{x \in \mathbb{K}^n \colon f(x) = 0,\ \forall f \in F\}$, se $F \subseteq \mathbb{K}[x_1, \ldots, x_n]$.
- $I(V) = \{f \in \mathbb{K}[x_1, \ldots, x_n] \colon f(x) = 0,\ \forall x \in V\}$, se $V \subseteq \mathbb{K}^n$.
- $\langle F \rangle$: ideale generato da $F \subseteq \mathbb{K}[x_1, \ldots, x_n]$.

1 Richiami di Geometria Algebrica

1.1 Flattening di Tensori

Definizione. Siano $V_1, \dots, V_n$ spazi vettoriali di dimensione finita su un campo $\mathbb{K}$ di caratteristica 0, ognuno dei quali munito di base $\mathcal{B}_j = \{e_s^j\}_{s=1,\dots,k_j}$, e sia $\{\epsilon_{i_1,\dots,i_n}\} = \{e_{i_1}^1 \otimes \cdots \otimes e_{i_n}^n : (i_1, \dots, i_n) \in [k_1] \times \cdots \times [k_n]\}$ la base indotta su $V_1 \otimes \cdots \otimes V_n$.
Sappiamo che $V \otimes W \simeq \mathrm{Hom}(V^*, W)$, e dunque che, più in generale, se A, B è una partizione di $[n]$, $V_1 \otimes \cdots \otimes V_n \simeq \mathrm{Hom}((V_A)^*, V_B)$ con $V_A = \bigotimes_{a\in A} V_a$ e $V_B = \bigotimes_{b\in B} V_b$. Possiamo scrivere esplicitamente questo isomorfismo, detto *contrazione rispetto alle coordinate in A*:

$$\begin{aligned} \psi_{A,B} : V_1 \otimes \cdots \otimes V_n &\longrightarrow \mathrm{Hom}((V_A)^*, V_B) \\ \varphi = \sum_{i_1,\dots,i_n} \varphi_{i_1\dots i_n} \epsilon_{i_1\dots i_n} &\longmapsto \beta_\varphi, \end{aligned}$$

dove

$$\beta_\varphi : \bigotimes_{a\in A} V_a^* \longrightarrow \bigotimes_{b\in B} V_b$$

è l'omomorfismo che sull'elemento $\epsilon_{i_\alpha}^*$ della base duale della base di $\bigotimes_{a_\in A} V_a$ vale

$$\beta_\varphi(\epsilon_{i_\alpha}^*) = \epsilon_{i_\alpha}^*\Big(\sum_{i_1,\dots,i_n} \varphi_{i_1\dots i_n}\epsilon_{i_1\dots i_n}\Big) = \sum_{i_1,\dots,i_n} \varphi_{i_1\dots i_n}\epsilon_{i_\beta},$$

dove i_β è il multi-indice ottenuto da $(i_1, \dots, i_n)$ eliminando gli indici in α. $\psi_{A,B}(\varphi)$ è detto *flattening* di φ rispetto alle coordinate di A e lo indichiamo come $\mathrm{Flat}_A(\varphi)$ o $\mathrm{Flat}_{A,B}(\varphi)$.

Anche se la definizione di flattening è abbastanza complicata, c'è un modo semplice per vedere questa operazione sui tensori: ordiniamo lessicograficamente i multi-indici $(i_\alpha)_{\alpha\in A}$ e $(i_\beta)_{\beta\in B}$. Il flattening di φ rispetto alle coordinate (A, B) è la matrice che ha come elemento (I, J)-esimo $\varphi_{\tau_A(I),\tau_A(J)}$, dove τ_i è la biezione fra i multi-indici i_α e $[\prod_{\alpha\in A} k_\alpha]$ e τ_i è la biezione fra i multi-indici i_β e $[\prod_{\beta\in B} k_\beta]$. In sostanza, il flattening è una matrice che ha gli stessi elementi del tensore ordinati in modo da raggruppare gli indici di A e gli indici di B (da qui il nome, "appiattimento").

Poniamo per definizione $\text{Flat}_{A,\emptyset}(\varphi) = \text{Flat}_{\emptyset,A}(\varphi)$ il vettore di $\mathbb{K}^N$ che contiene tutti gli elementi di φ ordinati lessicograficamente.

Esempio 1.1. Se $\varphi = \sum_{ijk} \varphi_{ijk}\epsilon_{ijk}$ è un tensore $2 \times 2 \times 2$ con le seguenti sezioni

$$\varphi_{1\cdot\cdot} = \begin{pmatrix} \varphi_{111} & \varphi_{112} \\ \varphi_{121} & \varphi_{122} \end{pmatrix} \qquad \varphi_{2\cdot\cdot} = \begin{pmatrix} \varphi_{211} & \varphi_{212} \\ \varphi_{221} & \varphi_{222} \end{pmatrix},$$

abbiamo

$$\text{Flat}_1(\varphi) = \begin{pmatrix} \varphi_{111} & \varphi_{112} & \varphi_{121} & \varphi_{122} \\ \varphi_{211} & \varphi_{212} & \varphi_{221} & \varphi_{222} \end{pmatrix}$$
$$\text{Flat}_2(\varphi) = \begin{pmatrix} \varphi_{111} & \varphi_{112} & \varphi_{211} & \varphi_{212} \\ \varphi_{121} & \varphi_{122} & \varphi_{221} & \varphi_{222} \end{pmatrix}$$
$$\text{Flat}_3(\varphi) = \begin{pmatrix} \varphi_{111} & \varphi_{121} & \varphi_{211} & \varphi_{221} \\ \varphi_{112} & \varphi_{122} & \varphi_{212} & \varphi_{222} \end{pmatrix}.$$

Lemma 1.1. *Sia $\varphi \in V_1 \otimes \cdots \otimes V_n$ e supponiamo che il flattening di φ rispetto alla prima coordinata abbia rango 0 o 1. Allora $\exists\, v_1 \in V_1$ e $\psi \in V_2 \otimes \cdots \otimes V_n$ tali che $\varphi = v_1 \otimes \psi$.*

Dimostrazione. Se il rango è 0 allora $\varphi = 0$ e non c'è niente da dimostrare: supponiamo dunque che il flattening abbia rango 1. Sia $v_1 \in V_1$ un generatore dell'immagine di $\text{Flat}_1(\varphi)$ e completiamolo ad una base $v_1, \ldots, v_{k_1}$ di V_1. Possiamo quindi scrivere $\varphi = v_1 \otimes \psi_1 + \ldots + v_{k_1} \otimes \psi_{k_1}$ per opportuni $\psi_t \in V_2 \otimes \cdots \otimes V_n$. Poiché $\text{Flat}_1(\varphi)(\beta) \in \text{span}(v_1)\ \forall \beta \in V_2^* \otimes \cdots \otimes V_n^*$, abbiamo $\psi_t = 0\ \forall t \geq 2$ e quindi la tesi. □

Definizione. Siano $V_1, \ldots, V_n$ spazi vettoriali su un campo $\mathbb{K}$. Un tensore $\varphi \in V_1 \otimes \cdots \otimes V_n$ si dice *decomponibile* se $\exists v_i \in V_i,\ i = 1 \ldots n$ tali che $\varphi = v_1 \otimes \cdots \otimes v_n$.

Dal precedente lemma segue la

Proposizione 1.2. *$\varphi \in V_1 \otimes \cdots \otimes V_n$ è decomponibile $\iff$ il rango di ogni flattening di φ è 0 o 1.*

Dimostrazione. Sia $\varphi = v_1 \otimes \cdots \otimes v_n \in V_1 \otimes \cdots \otimes V_n$ un tensore decomponibile: abbiamo $\text{Flat}_i(\varphi)(\beta) = \beta(v_1 \otimes \cdots \otimes v_{i-1} \otimes v_{i+1} \otimes \cdots \otimes$

$v_n)v_i$, quindi $\text{Im}(\text{Flat}_i(\varphi)) = span(v_i)$ e il flattening ha perciò rango 1.
Viceversa supponiamo che ogni flattening di φ abbia rango 1. Procediamo per induzione su n. Per $n = 2$, dal lemma 1.1 otteniamo immediatamente che $\varphi = v_1 \otimes v_2$. Supponiamo quindi vero l'enunciato fino a $n-1$ e sia $\varphi \in V_1 \otimes \cdots \otimes V_n$. Ancora per il lemma 1.1 sappiamo che $\exists\, v_1 \in V_1$ e $\psi \in V_2 \otimes \cdots \otimes V_n$ tali che $\varphi = v_1 \otimes \psi$. È facile verificare che anche i flattening di ψ hanno rango 0 o 1: per ipotesi induttiva esistono quindi $v_2 \in V_2, \ldots, v_n \in V_n$ e $\psi = v_2 \otimes \cdots \otimes v_n$. Otteniamo perciò $\varphi = v_1 \otimes \psi = v_1 \otimes (v_2 \otimes \cdots \otimes v_n) = v_1 \otimes v_2 \otimes \cdots \otimes v_n$ e dunque φ è decomponibile. □

Osservazione. È chiaro dalla dimostrazione che in realtà è sufficiente che $n-1$ flattening, comunque scelti fra gli n possibili, abbiano rango 0 o 1 per poter dire che φ è decomponibile.

Definizione. Il rango di un tensore $\varphi \in V_1 \otimes \cdots \otimes V_n$ è il minimo $k \in \mathbb{N}$ per il quale esistono $\varphi_1, \ldots, \varphi_k \in V_1 \otimes \cdots \otimes V_n$ decomponibili tali che $\varphi = \sum_{i=1}^{k} \varphi_i$.

Notiamo che nel caso bidimensionale il rango tensoriale non è altro che il rango matriciale.

1.2 Morfismo e Varietà di Segre

Definizione. Siano $V_1, \ldots, V_n$ spazi vettoriali di dimensione finita su $\mathbb{C}$. L'immersione

$$\begin{aligned} \phi : V_1 \times \cdots \times V_n &\hookrightarrow V_1 \otimes \cdots \otimes V_n \\ (v_1, \ldots, v_n) &\longmapsto v_1 \otimes \cdots \otimes v_n \end{aligned} \tag{1.1}$$

definisce un morfismo proiettivo

$$\begin{aligned} s : \mathbb{P}(V_1) \times \cdots \times \mathbb{P}(V_n) &\hookrightarrow \mathbb{P}(V_1 \otimes \cdots \otimes V_n) \\ ([v_1], \ldots, [v_n]) &\longmapsto [v_1 \otimes \cdots \otimes v_n] \end{aligned}$$

detto *morfismo di Segre*. La sua immagine è detta *Varietà di Segre*.

La proposizione 1.2 ci dice che un elemento di $\mathbb{P}(V_1 \otimes \cdots \otimes V_n)$ appartiene alla varietà di Segre se e soltanto se i suoi flattening hanno rango minore o uguale a 1.
Le equazioni che definiscono la varietà di Segre sono pertanto quadratiche e sono date dall'annullamento dei minori 2×2 dei flattening.

Esempio 1.2. Siano V_1, V_2 e V_3 spazi vettoriali complessi 2-dimensionali. Allora $\varphi \in V_1 \otimes V_2 \otimes V_3$ appartiene alla varietà di Segre se e soltanto se i minori 2×2 dell'esempio 1.1 si annullano, ovvero

$$\begin{aligned}
&\varphi_{111}\varphi_{212} - \varphi_{112}\varphi_{211} = 0, \qquad &&\varphi_{111}\varphi_{221} - \varphi_{121}\varphi_{211} = 0,\\
&\varphi_{111}\varphi_{222} - \varphi_{122}\varphi_{211} = 0, \qquad &&\varphi_{112}\varphi_{221} - \varphi_{121}\varphi_{212} = 0,\\
&\varphi_{112}\varphi_{222} - \varphi_{122}\varphi_{212} = 0, \qquad &&\varphi_{121}\varphi_{222} - \varphi_{122}\varphi_{221} = 0.
\end{aligned}$$

$$\begin{aligned}
&\varphi_{111}\varphi_{122} - \varphi_{112}\varphi_{121} = 0, \qquad &&\varphi_{111}\varphi_{221} - \varphi_{211}\varphi_{121} = 0,\\
&\varphi_{111}\varphi_{222} - \varphi_{212}\varphi_{121} = 0, \qquad &&\varphi_{112}\varphi_{221} - \varphi_{211}\varphi_{122} = 0,\\
&\varphi_{112}\varphi_{222} - \varphi_{212}\varphi_{122} = 0, \qquad &&\varphi_{211}\varphi_{222} - \varphi_{212}\varphi_{221} = 0.
\end{aligned}$$

$$\begin{aligned}
&\varphi_{111}\varphi_{122} - \varphi_{121}\varphi_{112} = 0, \qquad &&\varphi_{111}\varphi_{212} - \varphi_{211}\varphi_{112} = 0,\\
&\varphi_{111}\varphi_{222} - \varphi_{221}\varphi_{112} = 0, \qquad &&\varphi_{121}\varphi_{212} - \varphi_{211}\varphi_{122} = 0,\\
&\varphi_{121}\varphi_{222} - \varphi_{221}\varphi_{122} = 0, \qquad &&\varphi_{211}\varphi_{222} - \varphi_{221}\varphi_{212} = 0.
\end{aligned}$$

Notiamo che i 18 polinomi dell'esempio 1.2 non sono tutti indipendenti, coerentemente con quanto osservato prima, ovvero che è sufficiente che 2 dei 3 flattening di φ abbiano rango 1.
Un'ulteriore conferma di questo fatto segue considerando la decomposizione di $S^2(V_1 \otimes V_2 \otimes V_3)$:

$$\begin{aligned}
S^2(V_1 \otimes V_2 \otimes V_3) = {} & S^2(V_1) \otimes S^2(V_2) \otimes S^2(V_3) \bigoplus\\
& S^2(V_1) \otimes \Lambda^2(V_2) \otimes \Lambda^2(V_3) \bigoplus\\
& \Lambda^2(V_1) \otimes S^2(V_2) \otimes \Lambda^2(V_3) \bigoplus\\
& \Lambda^2(V_1) \otimes \Lambda^2(V_2) \otimes S^2(V_3).
\end{aligned}$$

Ora, $S^2(V_1 \otimes V_2 \otimes V_3)$ ha dimensione 36, mentre $S^2(V_1) \otimes S^2(V_2) \otimes$

$S^2(V_3)$ ha dimensione 27 e non contiene elementi dell'ideale della varietà di Segre $\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1$. Segue quindi che questo ideale è generato dalle 9 quadriche che generano $S^2(V_1) \otimes \Lambda^2(V_2) \otimes \Lambda^2(V_3) \bigoplus \Lambda^2(V_1) \otimes S^2(V_2) \otimes \Lambda^2(V_3) \bigoplus \Lambda^2(V_1) \otimes \Lambda^2(V_2) \otimes S^2(V_3)$, ovvero è sufficiente considerare 2 flattening su 3.

1.3 Join di Varietà

Siano V e W varietà affini (complesse). Il *join* di V e W è la varietà

$$\mathcal{J}(V,W) = \overline{\{\lambda v + (1-\lambda)w \colon v \in V, w \in W, \lambda \in \mathbb{C}\}}$$

dove la chiusura è secondo la topologia di Zariski.
In questo lavoro di tesi ci occuperemo spesso di varietà secanti, ovvero il join di una varietà con se stessa: $\sigma_2(V) = \mathcal{J}(V,V)$.
Questa varietà è costituita dalle rette passanti per coppie di punti di V. Possiamo costruire induttivamente la varietà *k-secante* a V, o *k*-esima varietà secante a V, come l'unione dei sottospazi k-dimensionali di $\mathbb{C}^N$ generati da $k+1$ punti indipendenti di V:

$$\sigma_1(V) = V \qquad \sigma_k(V) = \mathcal{J}(\sigma_{k-1}(V), V) \text{ per } k \geq 1.$$

Se V e W sono varietà proiettive, anche il loro join è la chiusura proiettiva del join dei loro coni, definizione che si estende alle varietà secanti.

Proposizione 1.3. *Sia $\varphi \in \sigma_k(V_1 \times \cdots \times V_n) \subseteq V_1 \otimes \cdots \otimes V_n$ la k-esima varietà secante all'immagine dell'immersione (1.1). Allora il rango di ogni flattening di φ è $\leq k$.*

Dimostrazione. Siano $\varphi^1, \cdots, \varphi^k \in V_1 \times \cdots \times V_n$ tali che $\varphi = \sum_{i=1}^k \varphi^i$. Sappiamo dalla proposizione 1.2 che il rango dei flattening rispetto alle singole coordinate di ogni φ^i è 0 o 1. Dunque otteniamo

$$\mathrm{rk}(\mathrm{Flat}_j(\varphi)) = \mathrm{rk}\left(\mathrm{Flat}_j\left(\sum_{i=1}^k \varphi^i\right)\right) = \mathrm{rk}\left(\sum_{i=1}^k \mathrm{Flat}_j(\varphi^i)\right) \leq \sum_{i=1}^k \mathrm{rk}(\mathrm{Flat}_j(\varphi^i)) \leq k.$$

□

La proposizione è valida anche se consideriamo la proiettivizzazione di $\sigma_k(V_1 \times \cdots \times V_n)$: abbiamo quindi delle condizioni sugli elementi della varietà di Segre $\mathbb{P}(V_1) \times \cdots \times \mathbb{P}(V_n)$.

Per studiare i modelli statistici che incontreremo più avanti sarà necessario introdurre una struttura algebrica più generale:

Definizione. Sia $\mathbb{R}[t_1, \ldots, t_k]$ l'anello dei polinomi nelle indeterminate $t_1, \ldots, t_k$ a coefficienti reali. Un insieme *semi-algebrico* è un'unione finita della forma

$$\Theta = \bigcup_{i=1}^{m} \left\{ \vartheta \in \mathbb{R}^k : f(\vartheta) = 0 \ \forall f \in F_i, h(\vartheta) > 0 \ \forall h \in H_i \right\}$$

dove $F_i, H_i \subseteq \mathbb{R}[t_1, ..., t_k]$ e H_i è finito per ogni i.

Siano V, W insiemi semi-algebrici. La loro *mistura* è l'insieme

$$\text{Mixt}(V, W) = \left\{ \lambda v + (1 - \lambda) w : v \in V, w \in W, \lambda \in [0, 1] \right\}$$

cioè l'unione delle combinazioni convesse di punti in V e W. In maniera analoga alle varietà secanti definiamo la mistura di un insieme come

$$\text{Mixt}^1(V) = V \qquad \text{Mixt}^k(V) = \text{Mixt}(\text{Mixt}^{k-1}(V), V)$$

Notiamo che per k sufficientemente grande questo insieme diventa l'inviluppo convesso di V.
Il legame fra misture e secanti è dato dalla seguente proposizione, che possiamo ritrovare in [8].

Proposizione 1.4. *Sia V un insieme semi-algebrico. Allora $\sigma_k(\overline{V}) = \overline{\text{Mixt}^s(V)}$, dove la chiusura è secondo la topologia di Zariski.*

Siano $V_1, \ldots, V_m$ $\mathbb{C}$-spazi vettoriali di dimensioni rispettivamente $n_1, \cdots, n_m$ e consideriamo la varietà di Segre X definita dall'immersione di questi in $\mathbb{P}^N$ con $N = \left(\prod_{i=1}^{m} n_i\right) - 1$. La varietà $\sigma_k(X)$ è costituita dalle combinazioni lineari di k elementi di X, che come abbiamo detto sono tensori di rango 1; ciò ci permette di caratterizzare questa varietà k-secante come l'insieme dei tensori di rango $\leq k$.

In [17] troviamo un viceversa della proposizione 1.3 che vale nel caso delle varietà 2-secanti.

Teorema 1.5. *$I(\sigma_2(V_1 \times \cdots \times V_n)$ è generato dai minori 3×3 dei flattening. In particolare, se $\varphi \in V_1 \otimes \cdots \otimes V_n$, allora $\varphi \in \sigma_2(V_1 \times \cdots \times V_n)$ se e soltanto se $\mathrm{rk}(\mathrm{Flat}_i(\varphi)) \leq 2$ per ogni coordinata i.*

1.4 Dimensione delle varietà secanti

La domanda che ci poniamo ora è la seguente:

> *Qual è la dimensione della varietà k-secante alla varietà di Segre?*

Con il semplice conteggio dei parametri è possibile dare una prima stima grezza:

Definizione. La *dimensione aspettata* di una varietà $k-$secante ad una varietà $X \subseteq \mathbb{P}^N$ è

$$d_e = \min\{N, k\dim(X) + k - 1\} \tag{1.2}$$

Diremo che la varietà è *difettiva* se la sua dimensione è strettamente minore della dimensione aspettata.

Proposizione 1.6. *Per ogni varietà k-secante ad una varietà proiettiva X si ha $d = \dim(\sigma_k(X)) \leq d_e$.*

Dimostrazione. $d \leq N$ segue banalmente dal fatto che $\sigma_k(X)$ è immersa in $\mathbb{P}^N$.
Sia $\sigma^k(X) = \overline{\{(x_1, \ldots, x_k, y) : y \in span(x_1, \ldots, x_k)\}} \subseteq \underbrace{X \times \cdots \times X}_{\text{k volte}} \times \mathbb{P}^N$
Abbiamo che $\sigma_k(X) = \pi(\sigma^k(X))$, dove π è la proiezione su $\mathbb{P}^N$, e dunque $d \leq \dim(\sigma^k(X)) = k\dim(X) + (k-1)$. □

Esempio 1.3. Siano V, W spazi vettoriali di dimensione 3. La varietà di Segre X definita dall'immersione $\phi : \mathbb{P}(V) \times \mathbb{P}(W) \longrightarrow \mathbb{P}^8$ ha dimensione 4 e quindi la dimensione aspettata della varietà 2-secante è $\min\{8, 2 \times 4 + 2 - 1\} = 8$. Sappiamo però che questa è costituita dall'insieme delle matrici (proiettivizzate) 3×3 di rango ≤ 2, cioè la varietà algebrica definita dall'equazione $\det(A) = 0$ e avente perciò dimensione 7: $\sigma_2(X)$ è dunque difettiva.

La dimensione delle varietà secanti alle varietà Segre non è nota in generale, ma in alcuni casi possiamo dire molto.
Un primo risultato riguarda il caso $\mathbb{P}^n \times \cdots \times \mathbb{P}^n$. (trattato ad esempio in [1].)

Teorema 1.7. *Sia* $X = Seg(\underbrace{\mathbb{P}^n \times \cdots \times \mathbb{P}^n}_{k \text{ volte}}) \subseteq \mathbb{P}^{(n+1)^k-1}$ *con* $k \geq 3$. *Siano inoltre*

$$s_k = \left\lfloor \frac{(n+1)^k}{nk+1} \right\rfloor, \qquad \delta_k \equiv s_k \mod (n+1), \qquad \delta_k \in \{0,1,\ldots,n\}.$$

- *Se* $s \leq s_k - \delta_k$ *allora* $\dim(\sigma_s(X)) = d_e$.
- *Se* $s \geq s_k - \delta_k + n + 1$ *allora* $\sigma_s(X) = \mathbb{P}^{(n+1)^k-1}$.

Per il caso generale abbiamo invece una congettura che compare in [1].

Definizione. Sia $\mathbf{n} = (n_1, \ldots, n_k) \in \mathbb{N}^k$ con $n_1 \leq \cdots \leq n_k$.

- $\mathbf{n}$ è detta *bilanciata* se $n_k \leq \prod_{i=1}^{k-1}(n_i+1) - \sum_{i=1}^{k-1} n_i$.
- $\mathbf{n}$ è detta *sbilanciata* se $n_k - 1 \geq \prod_{i=1}^{k-1}(n_i+1) - \sum_{i=1}^{k-1} n_i$.

Congettura 1.8. $\sigma_s(Seg(\mathbb{P}^{n_1} \times \cdots \times \mathbb{P}^{n_k}))$ *è difettiva se e soltanto se*

- $(n_1, \ldots, n_k)$ *è sbilanciata;*
- $k = 3$ *e* $(n_1, n_2, n_3) = (2, n, n)$ *con n pari;*
- $k = 3$ *e* $(n_1, n_2, n_3) = (2, 3, 3)$*;*
- $k = 4$ *e* $(n_1, n_2, n_3, n_4) = (1, 1, n, n)$.

2 Statistica Algebrica

2.1 I modelli statistici sono varietà algebriche

Il frequente ricorso a modelli polinomiali in statistica ha negli ultimi anni fornito lo spunto per l'utilizzo di metodi geometrici nello studio di problemi legati a questo ambito. In questa sezione ci occuperemo di ridefinire alcuni concetti fondamentali del Calcolo delle Probabilità in modo da consentire la formulazione in termini di varietà algebriche delle proprietà dei modelli statistici.

Una variabile aleatoria discreta X a valori in $[n]$ è caratterizzata dal vettore $p = (p_1, \dots, p_n)$, detto *distribuzione di probabilità* di X, per il quale abbiamo $p_i = \mathbf{P}[X = i]$ per ogni i. I vincoli $p_i \geq 0 \, \forall i$ e $\sum_{i=1}^{n} p_i = 1$ definiscono un sottoinsieme di $\mathbb{R}^n$ detto *simplesso di probabilità* ed indicato con Δ_{n-1}. Le variabili aleatorie a valori in $[n]$ sono quindi in corrispondenza $1-1$ con gli elementi di $\Delta_n \subset \mathbb{R}^n$.

Definizione. Un *modello statistico (discreto)* $\mathcal{M}$ è un sottoinsieme del simplesso di probabilità:

$$\mathcal{M} \subseteq \Delta_n.$$

Un modello statistico è semplicemente una generica collezione di variabili aleatorie, eventualmente vettoriali.
Consideriamo infatti una famiglia di variabili aleatorie finite $X_1, \dots, X_n$ a valori rispettivamente negli insiemi $[k_i]$ per $i = 1, \dots, n$; ponendo $p_{j_1,\dots,j_n} = \mathbf{P}[X_1 = j_1, \dots, X_n = j_n]$ definiamo un tensore $p \in \mathbb{R}^{k_1} \otimes \cdots \otimes \mathbb{R}^{k_n}$ (detto *tensore di probabilità*), che rappresenta le probabilità congiunte delle variabili.
Dal momento che $\sum_{i_1,\dots,i_n} p_{i_1 \dots i_n} = 1$ e $p_{i_1 \dots i_n} \geq 0$ possiamo vedere anche p come un punto del simplesso di probabilità $\Delta_{k_1 \times \cdots \times k_n}$.

Esempio 2.1. Se X e Y sono variabili aleatorie che rappresentano lanci di monete (non necessariamente eque) a valori in $\{T, C\}$, il tensore di probabilità è la matrice $\begin{pmatrix} p_{TT} & p_{TC} \\ p_{CT} & p_{CC} \end{pmatrix} \in \Delta_{2\times 2} = \Delta_4$. Se vogliamo studiare le distribuzioni per le quali $\mathbf{P}[X = T, Y = T] = \mathbf{P}[X = C, Y = C]$, dobbiamo prendere in considerazione il modello $\mathcal{M} = \{(p_{TT}, p_{TC}, p_{CT}, p_{CC}) \in \Delta_4 \colon p_{TT} = p_{CC}\}$.

Definizione. Chiameremo *modello statistico algebrico* un modello statistico nella forma

$$\mathcal{M} = V_\Delta(F) = V(F) \cap \Delta_n$$

dove $F \subset \mathbb{R}[\mathbf{p}]$ è una famiglia di polinomi nelle indeterminate $\mathbf{p} = \{p_{i_1 \dots i_n}\}$ a coefficienti reali.

L'ideale $I_\mathcal{M} = \langle F \rangle$ generato da F in $\mathbb{R}[\mathbf{p}]$ è detto *ideale del modello*, e vale $V_\Delta(F) = V_\Delta(I_\mathcal{M})$.

Spesso un modello statistico è definito anche a partire da dei parametri che variano in un opportuno sottoinsieme di $\mathbb{R}^k$:

Definizione. $\mathcal{M}$ è un *modello statistico algebrico parametrico*[1] se esistono un insieme semi-algebrico $\Theta \subseteq \mathbb{R}^k$ e una mappa razionale $\phi : \mathbb{R}^k \to \mathbb{R}^n$ per i quali

$$\mathcal{M} = \phi(\Theta) \subseteq \Delta_n.$$

Incontreremo in seguito modelli definiti in entrambi i modi, ognuno dei quali presenta vantaggi e svantaggi: ad esempio la forma parametrica è più semplice da trovare poiché in genere è ottenuta prendendo in considerazione distribuzioni semplici e immergendole in spazi di dimensione più grande, ma per questo motivo è più difficile da confrontare con i dati. Buona parte del nostro lavoro successivo consisterà nel trovare le equazioni per modelli parametrici.

Probabilità marginali

Un tensore di probabilità contiene tutte le informazioni relative al comportamento di una famiglia di variabili aleatorie, dal momento che tiene conto di tutti i possibili stati di esse. È naturale allora cercare un modo di dedurre da esso informazioni sui sottoinsiemi delle variabili.

Siano $X_1, \dots, X_n$ variabili aleatorie discrete con tensore di probabilità p e sia $X_{i_1} \dots X_{i_m}$ una sottofamiglia di esse: da p possiamo

[1] Vedi Sullivant [22].

costruire un tensore $\tilde{p}$ per le variabili della sottofamiglia (*probabilità marginali*) ponendo

$$\tilde{p}_{j_1\dots j_k} = \sum_{t_1,\dots,t_n} \mathbf{P}[X_1 = t_1, \dots, X_{i_1} = j_1, \dots, X_{i_k} = j_k, \dots, X_n = t_n].$$

Abbiamo cioè ottenuto un tensore di tipo $k_{i_1} \times \cdots \times k_{i_m}$ sommando sulle probabilità delle variabili al di fuori della sottofamiglia in esame. Per enfatizzare questo fatto e semplificare la notazione scriveremo $p_{+\cdots i_1 \cdots i_m \cdots +}$ per indicare il tensore di probabilità relativo alle variabili $X_{i_1}, \dots, X_{i_m}$, ponendo quindi un "+" in luogo del pedice relativo alle altre variabili.
Ad esempio, con questa notazione avremo $\mathbf{P}[X_i = j] = p_{+\cdots \underset{i}{j} \cdots +} = p^i_j$.

2.2 Indipendenza di variabili aleatorie

La prima relazione fra variabili aleatorie che consideriamo è quella di *indipendenza*: due variabili sono indipendenti se non hanno nessuna influenza l'una sull'altra. Più formalmente:

Definizione. Due variabili aleatorie X_1 e X_2 si dicono *indipendenti* se $p_{j_1 j_2} = \mathbf{P}[X_1 = j_1, X_2 = j_2] = p_{j_1 +} p_{+ j_2}$ per ogni $(j_1, j_2) \in [k_1] \times [k_2]$, ovvero

$$p = p^1 (p^2)^t = \begin{pmatrix} \vdots & & \vdots & & \vdots \\ p^2_1 p^1 & \dots & p^2_j p^1 & \dots & p^2_{k_2} p^1 \\ \vdots & & \vdots & & \vdots \end{pmatrix}.$$

Le colonne di p sono in questo caso tutte multiple di p^1, quindi possiamo affermare che se X_1 e X_2 sono indipendenti la matrice delle probabilità ha rango 1.

Proposizione 2.1. *Sia $P \in \Delta_{m \times n}$ una matrice a coefficienti positivi e a somma 1. Se $\mathrm{rk}(P) = 1$ allora esistono $p_1 \in \Delta_m$ e $p_2 \in \Delta_n$ tali che $P = p_1 p_2^t$.*

Dimostrazione. Poiché $\mathrm{rk}(P) = 1$ esistono $\mu = (\mu_1, \dots, \mu_m) \in \mathbb{R}^m$ e $\lambda = (\lambda_1, \dots, \lambda_n) \in \mathbb{R}^n$ per i quali $P_{ij} = \lambda_i \mu_j$ per ogni i e j. Po-

niamo $p_1 = \left(\sum_{i=1}^{n} \lambda_i\right)\mu$ e $p_2 = \left(\sum_{j=1}^{m} \mu_j\right)\lambda$ ottenendo $\sum_j (p_1)_j = \sum \lambda_i \sum \mu_j = \sum \lambda_i \mu_j = 1$, e quindi, dal momento che chiaramente $(p_1)_j \geq 0 \, \forall j$, $p_1 \in \Delta_m$ (analogamente $p_2 \in \Delta_n$). Infine $(p_1 p_2^t)_{uv} = \lambda_u \mu_v \sum_{i,j} \lambda_i \mu_j = \lambda_u \mu_v = P_{uv} \, \forall u, v$. □

Questo fornisce le equazioni per un modello statistico:

$$\mathcal{M}_{\perp\!\!\!\perp} = \left\{ p \in \Delta_{k_1 \times k_2} \subseteq \mathbb{R}^{k_1 \times k_2} \colon \operatorname{rk} p = 1 \right\} =$$
$$= \left\{ p \in \Delta_{k_1 \times k_2} \colon \det \bar{p} = 0, \, \forall \bar{p} \text{ sottomatrice } 2 \times 2 \text{ di } v \right\},$$

detto *modello di indipendenza*.
Se chiamiamo F l'insieme dei minori 2×2 di $\mathbf{p}$, possiamo scrivere $\mathcal{M}_{\perp\!\!\!\perp} = \Delta_{k_1 k_2} \cap V(F)$: il modello di indipendenza per 2 variabili aleatorie è un modello statistico algebrico. L'ideale $I_{\perp\!\!\!\perp}$ generato da F è detto *ideale di indipendenza*.

Esempio 2.2. Rappresentiamo il lancio di due monete con due variabili aleatorie X e Y a valori in $\{0, 1\}$. Il tensore di probabilità è la matrice $p = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} \in \Delta_{2\times 2}$. Se le variabili sono indipendenti il rango di p è 1, e dunque $I_{\perp\!\!\!\perp} = \langle p_{00}p_{11} - p_{01}p_{10} \rangle \subseteq \mathbb{R}[p_{00}, p_{01}, p_{10}, p_{11}]$.

L'indipendenza di due variabili è ovviamente un caso particolare di un modello generale:

Definizione. $X_1, \ldots, X_n$ variabili aleatorie si dicono *mutuamente indipendenti* se $p_{j_1,\ldots,j_n} = \prod_{l=1}^{n} p_{j_l}^{l}$ per ogni $(j_1, \ldots, j_n) \in [k_1] \times \cdots \times [k_n]$.

Come prima, questa condizione è equivalente a richiedere $p = p^1 \otimes \cdots \otimes p^n$, cioè che p sia decomponibile; la proposizione 1.2 ci permette di affermare che $X_1, \ldots, X_n$ sono mutuamente indipendenti se e soltanto se tutti flattening del tensore di probabilità hanno rango 1. Ancora come sopra, se chiamiamo F l'insieme dei minori 2×2 dei flattening di p, abbiamo $\mathcal{M}_{\perp\!\!\!\perp} = \Delta_K \cap V(F)$, con $K = \prod_{i=1}^{n} k_i$.

Spesso saremo interessati all'indipendenza congiunta di famiglie di variabili aleatorie:

Definizione. Siano $X_1, \dots, X_n$ variabili aleatorie e sia (A, B, C) una partizione di $[n]$. Chiamiamo $X_A = \{X_i : i \in A\}$ e $X_B = \{X_i : i \in B\}$ le due variabili vettoriali definite da A e B. Diremo che le due sottofamiglie A e B sono indipendenti se lo sono le due variabili X_A e X_B, cioè se $\mathrm{Flat}_{A,B}(p_{\cdot\cdot+})$ ha rango 1, dove $p_{\cdot\cdot+}$ è il tensore ottenuto sommando sulle variabili in C.

Osservazione. Non è detto che se $X_i \perp\!\!\!\perp X_j \; \forall i \in A, j \in B$ allora $X_A \perp\!\!\!\perp X_B$. Infatti, consideriamo 3 variabili aleatorie binarie X_1, X_2 e X_3 a valori in $\{0, 1\}$ con la seguente distribuzione congiunta p_{ijk}

$$p_{\cdot\cdot 0} = \begin{pmatrix} 0.013 & 0.167 \\ 0.047 & 0.073 \end{pmatrix} \qquad p_{\cdot\cdot 1} = \begin{pmatrix} 0.087 & 0.333 \\ 0.053 & 0.227 \end{pmatrix}$$

e siano $A = \{1, 2\}$ e $B = \{3\}$. La matrice delle probabilità marginali per X_1 e X_3

$$\begin{pmatrix} 0.180 & 0.120 \\ 0.420 & 0.280 \end{pmatrix}$$

ha rango 1, quindi $X_1 \perp\!\!\!\perp X_3$. La matrice delle probabilità marginali per X_2 e X_3

$$\begin{pmatrix} 0.060 & 0.240 \\ 0.140 & 0.560 \end{pmatrix}$$

ha rango 1 e quindi anche $X_2 \perp\!\!\!\perp X_3$. D'altra parte però $0.013 = p_{000} \neq p_{00+} p_{++0} = 0.03$, quindi X_A non è indipendente da X_B.

Proposizione 2.2. *Due famiglie di variabili aleatorie X_A e X_B sono indipendenti se e soltanto se esistono $\phi \in \mathbb{R}^{k_1} \otimes \cdots \otimes \mathbb{R}^{k_j}$ e $\psi \in \mathbb{R}^{k_{j+1}} \otimes \cdots \otimes \mathbb{R}^{k_r}$ tali che $p_{\cdot\cdot+} = \phi \otimes \psi$.*

Dimostrazione. Se X_A e X_B sono indipendenti $\mathrm{rk}(\mathrm{Flat}_{A,B}(p_{\cdot\cdot+})) = 1$, esistono dunque $v \in \Delta_{K_1}$ e $w \in \Delta_{K_2}$ tali che $\mathrm{Flat}_{A,B} = vw^t$, dove $K_1 = \prod_{i=1}^{j} k_i$ e $K_2 = \prod_{i=j+1}^{r} k_i$. Raggruppando opportunamente gli elementi di v in un tensore $\phi \in \mathbb{R}^{k_1} \otimes \cdots \otimes \mathbb{R}^{k_j}$ e gli elementi di w in un tensore $\psi \in \mathbb{R}^{k_{j+1}} \otimes \cdots \otimes \mathbb{R}^{k_r}$ abbiamo la tesi. Viceversa, se $p_{\cdot\cdot+} = \phi \otimes \psi$ abbiamo $\mathrm{Flat}_{A,B}(p_{\cdot\cdot+}) = (\mathrm{Flat}_{A,\varnothing}(\phi))(\mathrm{Flat}_{\varnothing,B}(\psi))^t$, cioè $\mathrm{rk}(\mathrm{Flat}_{A,B}(p_{\cdot\cdot+})) = 1$. □

2.3 Indipendenza condizionata

Abbiamo detto che l'indipendenza è una situazione nella quale le variabili non si influenzano l'una con l'altra; un concetto più generale è l'indipendenza condizionata che, come suggerisce il nome, descrive distribuzioni di probabilità su variabili che, pur non essendo necessariamente indipendenti, lo sono se conosciamo le realizzazioni di altre variabili.

Definizione. Siano X, Y, Z variabili aleatorie finite. Diremo che X è indipendente da Y dato Z se per ogni $i \in [k_x]$, $j \in [k_y]$ e $z \in [k_z]$ si ha

$$p_{ijz} = p_{i+z}p_{+jz}. \tag{2.1}$$

Equivalentemente, $X \perp\!\!\!\perp Y|Z$ se, $\forall z \in [k_z]$, $p_{\cdot\cdot z}$ ha rango 1.

Se chiamiamo C una relazione di indipendenza condizionata $X \perp\!\!\!\perp Y|Z$, l'ideale I_C generato dal polinomio nell'equazione (2.1) ci permette di definire un modello $M_C = \Delta_n \cap V(I_C)$ detto *modello di indipendenza condizionata*.
Analogamente, se $C_1, \dots, C_m$ sono relazioni di indipendenza condizionata, l'ideale $I = I_{C_1} + \cdots + I_{C_m}$ definisce un modello di indipendenza condizionata.

Possiamo estendere la definizione a gruppi di variabili:

Definizione. Sia $X_1, \dots, X_n$ una famiglia di variabili aleatorie, e sia (A, B, C) una partizione di un sottoinsieme di$\{1, \dots, n\}$. Diremo che X_A è indipendente da X_B dato X_C (e scriveremo $X_A \perp\!\!\!\perp X_B|X_C$) se, per ogni x_A, x_B, z_C stati delle variabili rispettivamente di X_A, X_B e X_C, si ha

$$\mathbf{P}[X_A = x_A, X_B = x_B|X_C = z_C] = \mathbf{P}[X_A = x_A|X_C = z_C]\mathbf{P}[X_B = x_B|X_C = z_C].$$

Proposizione 2.3. *Siano $X_1, \dots, X_n$ variabili aleatorie e sia (A, B, C) una partizione di $\{1, \dots, n\}$. Se $X_A \perp\!\!\!\perp X_B|X_C$ allora*

$$p_{x_\alpha x_\beta z_\gamma} p_{y_\alpha y_\beta z_\gamma} = p_{x_\alpha y_\beta z_\gamma} p_{y_\alpha x_\beta z_\gamma} \tag{2.2}$$

per ogni x_α, y_α stati di X_A, x_β, y_β stati di X_B e z_γ stato di X_C.

Dimostrazione. Raggruppiamo le variabili contenute in A in una singola variabile aleatoria Y_A che assume valori in $\prod_{a\in A}[k_a]$ con probabilità pari alle probabilità congiunte delle variabili X_A, e lo stesso facciamo per B e C, e sia q il tensore tridimensionale di probabilità per Y_A, Y_B e Y_C. Se $X_A \perp\!\!\!\perp X_B|X_C$ avremo che, fissato $z_\gamma \in \prod_{c\in C}[k_c]$, la matrice q_{z_γ} definita da $q_{z_\gamma}(i,j) = q_{ijz_\gamma}$ ha rango 1. Dunque ogni sua sottomatrice quadrata ha determinante nullo, e perciò , per ogni $x_\alpha, x_\beta, y_\alpha, y_\beta$ come in ipotesi, abbiamo

$$\begin{aligned} 0 &= q_{z_\gamma}(x_\alpha, y_\alpha)q_{z_\gamma}(x_\beta, y_\beta) - q_{z_\gamma}(x_\beta, y_\alpha)q_{z_\gamma}(x_\alpha, y_\beta) = \\ &= q_{x_\alpha y_\alpha z_\gamma} q_{x_\beta y_\beta z_\gamma} - q_{x_\beta y_\alpha z_\gamma} q_{x_\alpha y_\beta z_\gamma}. \end{aligned}$$

Poiché $q_{xyz} = \mathbf{P}[Y_A = x, Y_B = y, Y_C = z]$, "separando" le variabili si ha la tesi. □

2.4 Mistura di Modelli

Grazie alla definizione di indipendenza condizionata è possiamo descrivere adesso situazioni in cui alcune variabili sono mancanti, cioè se non conosciamo la distribuzione di tutte le variabili, ma soltanto il modello indotto dalla marginalizzazione rispetto ad un sottoinsieme delle variabili.
Iniziamo con un caso semplice: siano X e Y due variabili aleatorie a valori rispettivamente in $[n]$ e $[s]$, dove Y è "nascosta", ovvero è una variabile per la quale non è possibile raccogliere dati. In questo caso supponiamo però di avere a disposizione La distribuzione congiunta per X e Y è data da

$$\mathbf{P}[X = i, Y = j] = p_{ij} = \pi_j p_i^j,$$

dove con p^j indichiamo il vettore contenente le probabilità per la variabile X condizionate al valore j di Y e con π indichiamo la distribuzione di Y. Dal momento che i dati di Y non sono ottenibili l'unica distribuzione osservabile è quella marginale di X, data da

$$\mathbf{P}[X = i] = \sum_{j\in[s]} \pi_j p_i^j.$$

Notiamo che la distribuzione marginale di X è combinazione convessa delle s probabilità condizionate $p^1, \ldots, p^s$, pesate dalla distribuzione π.
Questo ci dà l'idea per la seguente

Definizione. Sia $\mathcal{P} \subseteq \Delta_r$. L'*s-esima mistura di modelli* per $\mathcal{P}$, o *s-esimo modello mistura* per $\mathcal{P}$, è

$$\mathrm{Mixt}^s(\mathcal{P}) = \left\{ \sum_{j \in [s]} \pi_j p^j : \pi \in \Delta_s \; e \; p^j \in \mathcal{P} \; \forall j \right\}$$

Più semplicemente, una mistura di modelli è un modello indotto dalla marginalizzazione rispetto ad una variabile latente ad ogni stato della quale è associata una distribuzione dal modello $\mathcal{P}$. Notiamo che, coerentemente con la notazione utilizzata, l's-esimo modello mistura per un modello $\mathcal{P}$ è l's-esima mistura di insiemi per $\mathcal{P}$.

Esempio 2.3. Sia $\mathcal{M}_{\perp\!\!\!\perp}^{(2)}$ il modello di indipendenza di due variabili, a valori rispettivamente in $[r]$ e $[c]$, dato dalle matrici di rango 1 in $\Delta_{r \times c}$.
$\mathrm{Mixt}^s(\mathcal{M}_{X \perp\!\!\!\perp Y}^{(2)})$ è dato dalle matrici A in $\Delta_{r \times c}$ di *rango non-negativo* $\leq s$, dove con rango non negativo si intende il minimo numero s di matrici a coefficienti non negativi di rango 1 che sommate diano A.

Una mistura di modelli molto importante è il *modello a classi latenti*, una generalizzazione dell'esempio precedente: sia $\mathcal{P} = \mathcal{M}_{\perp\!\!\!\perp}$ il modello di *completa indipendenza* per delle variabili $X_1, \ldots, X_n$, ovvero $X_1, \ldots, X_n$ mutuamente indipendenti. Il modello a classi latenti è la mistura di modelli su $\mathcal{P}$, e corrisponde alla situazione nella quale un insieme di variabili tra di loro correlate sono indipendenti se condizionate ad una singola variabile latente.

Definizione. $\varphi \in V_1 \otimes \cdots \otimes V_n$ ha *rango non negativo* r se esistono $\varphi^1, \ldots, \varphi^r \in V_1 \otimes \cdots \otimes V_n$ decomponibili a coefficienti non negativi tali che $\varphi = \sum_{j=1}^r \varphi^j$. Scriveremo per indicare ciò $\mathrm{rk}_+(\varphi) = r$.

Teorema 2.4. *Se $\mathcal{M}_{\perp\!\!\!\perp}$ è il modello di completa indipendenza per n variabili a valori in $[k_1], \ldots, [k_n]$, si ha*

$$\mathrm{Mixt}^s(\mathcal{M}_{\perp\!\!\!\perp}) = \{ p \in \Delta_K \colon \mathrm{rk}_+(p) \leq s \}, \tag{2.3}$$

con $K = \prod k_i$.

Purtroppo non esistono algoritmi generali per il calcolo del rango non negativo di un tensore, anzi, determinarlo risulta essere addirittura un problema *NP-Hard* e pertanto sarà necessario percorrere altre vie per trovare le equazioni del modello a classi latenti.

2.5 Modelli Grafici

Per rappresentare modelli più complessi useremo dei grafi, grazie all'ampia teoria esistente sul loro utilizzo per lo studio delle relazioni n-arie, come, ad esempio, l'indipendenza e l'indipendenza condizionata.

Elementi di teoria dei grafi

Un *grafo* è una coppia ordinata G di insiemi finiti (V, E): gli elementi di V sono punti detti *vertici* mentre gli elementi di E sono coppie di vertici e sono detti *archi*. Un grafo si dice *orientato* se le coppie in E sono ordinate, *non orientato* altrimenti. Due vertici $v, w \in V$ si dicono *consecutivi* se $(v, w) \in E$ o $(w, v) \in E$.
Se $e = (v, w) \in E$ diremo che v e w sono gli *estremi* di e. Se G è orientato diremo inoltre che v è un *genitore* di w, w è *figlio* di v e e è *incidente* in w. A volte v è chiamata anche *coda* di e e w *testa* di e. Due archi $(v_1, w_1), (v_2, w_2)$ si dicono *consecutivi* se hanno un estremo comune.
Una collezione di archi consecutivi $\{(v_i, w_i)\}_{i=1\dots n}$ è detta *cammino* in G da v_1 a w_n; inoltre, se $w_i = v_{i+1}$ per ogni $i = 1, \dots n-1$ il cammino si dice *orientato*. Se esiste un cammino orientato da v a w in G, v si dice *antenato* di w mentre w si dice *discendente* di v. Un cammino orientato da un vertice in se stesso si dice *ciclo*.
Se e, f sono archi consecutivi in un cammino τ incidenti in uno stesso vertice $v \in V$ diremo che v è una *collisione* per τ.
Ad ogni grafo orientato G possiamo associare un grafo non orientato $\tilde{G}$ "dimenticando" il verso degli archi (ovvero l'ordine delle coppie): in questo caso $\tilde{G}$ si dice *soggiacente* a G (*skeleton graph*).
Un grafo non orientato G si dice *connesso* se per ogni $v, w \in V$ esiste un cammino da v a w in G. Un grafo orientato G si dice *connesso* se

per ogni $v, w \in V$ esiste un cammino orientato da v a w in G, si dice invece *debolmente connesso* se è connesso il grafo non orientato ad esso soggiacente.

Rappresentazione di Indipendenze condizionate

Sia $G = (V, E)$ un grafo orientato aciclico, ovvero privo di cicli (comunemente indicato con DAG, *directed acyclic graph*), e associamo ad ogni vertice $v \in V$ una variabile aleatoria discreta X_v.

Definizione. Una coppia (G, p), dove G è un DAG e p un tensore di probabilità per le variabili $\{X_v\}_{v \in V}$, è detta *rete bayesiana* su G se

$$p_{i_1 \dots i_n} = \prod_{v \in V} \mathbf{P}[X_v = i_v | X_{\mathrm{pa}(v)} = i_{\mathrm{pa}(v)}],$$

dove pa(v) è l'insieme dei genitori di v.

Definizione. Diremo che la famiglia di variabili $\{X_v\}_{v \in V}$ di una rete bayesiana soddisfa la *proprietà locale di Markov per* G se valgono le seguenti indipendenze:

$$\mathrm{local}(G) = \{X_v \perp\!\!\!\perp X_{\mathrm{nd}(v)} | X_{\mathrm{pa}(v)} : v \in V\},$$

dove nd(v) è l'insieme dei vertici di G che non sono discendenti di v. Riprendendo le notazioni di [13], chiamiamo $I_{local(G)}$ l'ideale generato da queste relazioni.

Sia $C \subseteq V$ e consideriamo un cammino γ in G. Diciamo che γ è d-separato da C se vale una delle seguenti condizioni

- γ contiene un vertice $v \in C$ che non è una collisione per γ
- γ contiene un vertice $v \notin C$ collisione per γ che non ha discendenti in C.

Se A, B, C insiemi disgiunti di vertici diremo che C d-separa A e B se C d-separa ogni cammino da un vertice di A a un vertice di B.

Definizione. Dato un DAG $G = (V, E)$, diremo che la rete bayesiana su G $\{X_v\}_{v \in V}$ soddisfa la *proprietà globale di Markov* per G se $X_A \perp\!\!\!\perp X_B | X_C$ se C d-separa A e B in G.
Chiamiamo $I_{global(G)}$ l'ideale generato da queste relazioni.

Vale il seguente teorema di fattorizzazione:

Teorema 2.5. $V_\Delta(I_{local(G)}) = V_\Delta(I_{global(G)})$,

riportato nel testo classico di S. Lauritzen sui modelli grafici [18].

2.6 Varietà Algebriche dei Modelli Statistici

Abbiamo visto che per il modello a classi latenti le difficoltà che incontriamo nel trovare le equazioni ci spingono a cercare modi alternativi di studiare geometricamente i modelli statistici.
Il primo passo consiste nel passaggio alla chiusura proiettiva, così da poterci ricondurre alle varietà studiate comunemente in geometria algebrica.

Modello di completa indipendenza

Il modello di completa indipendenza è l'immagine dell'applicazione

$$\begin{aligned} \Delta_{k_1} \times \cdots \times \Delta_{k_n} &\longrightarrow \Delta_K \\ (p_1, \ldots, p_n) &\mapsto p_1 \otimes \cdots \otimes p_n \end{aligned} \tag{2.4}$$

ed è quindi un modello statistico parametrico.
È anche un modello grafico, corrispondente al grafo privo di archi.
L'ideale $I_{\perp\!\!\!\perp}$ definisce la varietà dei tensori decomponibili, che, come abbiamo visto nel paragrafo 1.2 di questo lavoro, è la varietà di Segre, immagine del morfismo proiettivo

$$\begin{aligned} \phi : \mathbb{P}^{k_1} \times \cdots \times \mathbb{P}^{k_n} &\hookrightarrow \mathbb{P}^K \\ ([p_1], \ldots, [p_n]) &\mapsto [p_1 \otimes \cdots \otimes p_n]. \end{aligned} \tag{2.5}$$

.

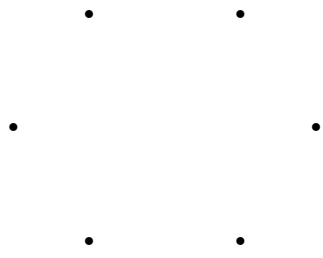


Figura 2.1: Grafo per il modello di completa indipendenza.

Misture e modello a classi latenti

Se $\mathcal{P}$ è un modello parametrico $\phi(\Theta)$, anche $\text{Mixt}^s(\mathcal{P})$ è parametrico descritto dall'applicazione

$$\Theta^s \times \Delta_s \longrightarrow \Delta_K$$
$$((\theta_1, \ldots, \theta_s), \pi) \longmapsto \sum_{i=1}^{s} \pi_i \phi(\theta_i).$$

Se $\mathcal{P}$ è il modello di completa indipendenza, il grafo corrispondente al modello comprendente anche la variabile latente è il cosiddetto grafo a stella o ad artiglio:

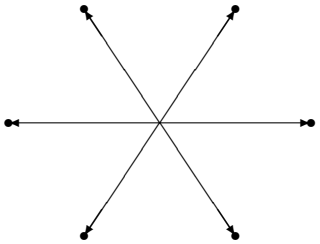


Figura 2.2: Grafo a stella.

Grazie alla proposizione 1.4 possiamo trovare la varietà corrispondente a questo modello:

$$\overline{\text{Mixt}^s(\mathcal{M}_{\perp\!\!\!\perp})} = \sigma_s(\overline{\mathcal{M}_{\perp\!\!\!\perp}}) = \sigma_s(\mathbb{P}^{k_1-1} \times \cdots \times \mathbb{P}^{k_n-1}).$$

Lo studio delle varietà secanti alle varietà di Segre è un campo di

ricerca ancora attivo nella geometria algebrica ma che comunque fornisce numerosi risultati che potremo utilizzare. Vedremo poi che per i nostri modelli avremo bisogno delle equazioni di $\sigma_4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)$, non ancora interamente note ma oggetto di ricerca negli ultimi anni. Notiamo che il grafo in figura 2.2 non corrisponde al modello a classi latenti ma al modello con 7 variabili contenente anche quella latente rispetto alla quale si deve poi marginalizzare.

3 Alberi Filogenetici

Lo scopo ultimo di questo lavoro è fornire dei metodi algebrici per la ricostruzione degli alberi filogenetici, diagrammi usati in biologia per lo studio della storia evoluzionistica delle specie viventi. Nella nostra interpretazione questi sono modelli grafici definiti da grafi chiamati alberi: in questo paragrafo studieremo le varietà algebriche relative a questi modelli.

3.1 Modelli definiti da alberi

Definizione. Un *albero* è un grafo non orientato T aciclico e connesso. Se un vertice v di T ha valenza 1 è detto *foglia*, altrimenti è un *vertice interno*.
Un *sottoalbero* di un albero T è un albero ottenuto considerando un sottoinsieme dei vertici e degli archi di T.

Dato un albero T, otteniamo un grafo orientato scegliendo un vertice r (detto *radice*) e orientando gli archi in direzione uscente da esso. Notiamo che con questa scelta ogni vertice ha al più un genitore: parleremo in questo caso di *albero orientato*.
Considereremo d'ora in avanti alberi i cui vertici interni abbiano valenza costante $\tau \geq 3$, con particolare attenzione al caso $\tau = 3$, corrispondente agli alberi *binari*, chiamati così in quanto, con l'orientazione sopra descritta, ogni vertice interno diverso dalla radice ha 2 figli.

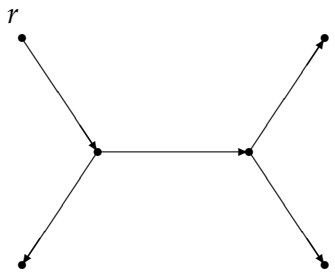


Figura 3.1: Esempio di albero binario con 4 foglie.

Sia $T = (V, E)$ un albero binario con n foglie e m vertici interni: ad esso possiamo associare un modello grafico con $n + m$ variabili

aleatorie definito dalla proprietà di Markov locale. Dal momento che ogni vertice ha al più 1 genitore, possiamo definire il modello nella maniera seguente:

- Sia π_r la distribuzione di probabilità per X_r.
- Per ogni $v \in V$, sia $w \in V$ il genitore di v e sia e l'arco di estremi w e v (notiamo che è l'unico entrante in v). Ad e possiamo associare una matrice M_e $l_1 \times l_2$, dove l_1 è il numero di stati di X_w e l_2 il numero di stati di X_v, detta *di transizione*, tali che

$$M_e(i,j) = \mathbf{P}[X_v = i | X_w = j].$$

Proposizione 3.1. *π_r e $\{M_e\}_{e\in E}$ individuano il modello associato all'albero T.*

Dimostrazione. Vogliamo vedere che le matrici di transizione e la distribuzione su r sono sufficienti a costruire il tensore di probabilità P. Siano dunque $v_1, \dots, v_k$ i vertici diversi dalla radice in T e sia $p_{\rho i_1 \dots i_k} = \mathbf{P}[X_r = \rho, X_{v_1} = i_1, \dots, X_{v_k} = i_k]$. Si ha allora

$$p_{\rho i_1 \dots i_k} = p_{i_1 \dots i_k | \rho} p_\rho = p_{i_3 \dots i_k | i_1 i_2 \rho} p_{i_1 i_2 | \rho} p_\rho = p_{i_3 \dots i_k | i_1 i_2 \rho} p_{i_1 | \rho} p_{i_2 | \rho} p_\rho = \\ = p_{i_3 \dots i_k | i_1 i_2 \rho} M_{r v_1}(\rho, i_1) M_{r v_2}(\rho, i_2) \pi_\rho$$

e l'ultima uguaglianza segue da $X_{v_1} \perp\!\!\!\perp X_{v_2} | X_r$. Iterando questo procedimento riusciamo a scrivere il tensore di probabilità a partire dagli elementi di π e delle matrici $\{M_e\}$. □

Questa proposizione ci permette di vedere i modelli grafici associati agli alberi come modelli parametrici, descritti al variare degli elementi delle matrici di transizione e dalla distribuzione sulla radice.
Spesso, come nel caso dei modelli filogenetici, gli alberi rappresentano situazioni nelle quali i vertici interni sono soltanto delle tappe intermedie in processi evolutivi che hanno come risultato le foglie. È naturale dunque considerare il modello come relativo soltanto alle foglie e le variabili nei vertici interni come nascoste tramite un'operazione di marginalizzazione, esattamente come avevamo visto nel modello a classi latenti.

D'ora in avanti, perciò, quando parleremo di modello definito da un albero, ci riferiremo sempre al modello relativo alle foglie.

Alberi Filogenetici

Anche se ancora non abbiamo parlato direttamente di filogenesi, possiamo definire di già cosa sia un albero filogenetico, premettendo che per il momento lo vediamo soltanto come una particolare struttura grafica, senza nessun significato biologico.
Sia T un albero con n foglie, m vertici interni e radice r. Come al solito associamo ad ogni vertice una variabile aleatoria X_v che supponiamo essere a valori in $[q]$ per ogni vertice. Questo valore q è il numero di stati che le variabili nel modello possono assumere: ci occuperemo principalmente dei casi $q = 2$ e $q = 4$, il primo perché è il più semplice e trattabile mentre il secondo perché corrisponde ai siti di DNA, nei quali le variabili sono a valori nell'insieme delle basi $\{A, C, G, T\}$. Questo modello grafico è detto *albero filogenetico* per n specie.

Definizione. Due alberi filogenetici con n foglie T_1, T_2 si dicono *topologicamente equivalenti* se, esiste un omeomorfismo che porta T_1 in T_2 e i vertici di T_1 nei corrispondenti vertici di T_2 Le classi di equivalenza rispetto a questa relazione sono dette *topologie* di alberi con n foglie.

Ad esempio, esiste una sola topologia di alberi binari con 3 foglie, mentre ne esistono tre di alberi con 4 foglie (figura 3.2).

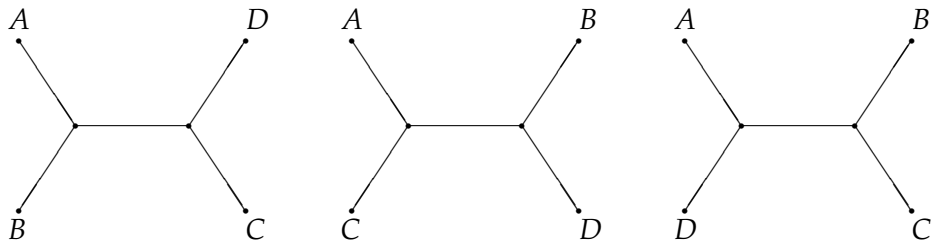


Figura 3.2: Topologie di alberi binari con 4 foglie.

A seconda della struttura delle matrici di transizione e della distribuzione sulla radice si identificano le seguenti classi di modelli, ognuna delle quali ha un particolare significato biologico di cui però non ci occuperemo.

Classi di modelli filogenetici

Modello generale di Markov: Non ci sono condizioni ulteriori sugli elementi delle matrici di transizione, dunque il numero totale dei parametri è $q(q-1)|E|+q-1$. È il modello più generale e la varietà corrispondente include tutte le altre.

Kimura a 3 parametri: $q = 4$ e le matrici sono nella forma

$$M = \begin{pmatrix} 1-(\alpha+\beta+\gamma) & \alpha & \beta & \gamma \\ \alpha & 1-(\alpha+\beta+\gamma) & \gamma & \beta \\ \beta & \gamma & 1-(\alpha+\beta+\gamma) & \alpha \\ \gamma & \beta & \alpha & 1-(\alpha+\beta+\gamma) \end{pmatrix}.$$

Kimura a 2 parametri: Come il modello a 3 parametri con $\beta = \gamma$.

Jukes-Cantor: Come il modello a 3 parametri con $\alpha = \beta = \gamma$.

Modello "Strand Symmetric": $q = 4$, $\pi_1 = \pi_4, \pi_2 = \pi_3$ e le matrici sono nella forma:

$$M = \begin{pmatrix} \theta_{AA} & \theta_{AC} & \theta_{AG} & \theta_{AT} \\ \theta_{CA} & \theta_{CC} & \theta_{CG} & \theta_{CT} \\ \theta_{CT} & \theta_{CG} & \theta_{CC} & \theta_{CA} \\ \theta_{AT} & \theta_{AG} & \theta_{AC} & \theta_{AA} \end{pmatrix}.$$

3.2 Il modello generale di Markov: parametrizzazione

Il modello che studieremo in questo lavoro è il modello generale di Markov, così come viene trattato nel lavoro di E. Allman e J. Rhodes [3].

Lo spazio dei parametri

Sappiamo che per descrivere interamente il modello sono sufficienti la distribuzione π_r sulla radice e le matrici di transizione, perciò lo spazio dei parametri sarà

$$S = \Delta_q \times \mathcal{M}^{|E|},$$

dove $\mathcal{M} = \left\{ M \in M(q, q, \mathbb{R}) \colon \sum_j M_{ij} = 1, M_{ij} \geq 0 \right\}$ è l'insieme delle *matrici stocastiche*.

L'applicazione

Possiamo descrivere adesso in modo esplicito la parametrizzazione del modello filogenetico:

$$\begin{aligned} \phi_r : S &\longrightarrow \Delta_{q^n} \\ s &\mapsto P, \end{aligned}$$

con

$$\begin{aligned} P(i_1, \ldots, i_n) &= \mathbf{P}[X_1 = i_1, \ldots, X_n = i_n] = \\ &= \sum_{b_v \in H} \left(\pi_r(b_r) \prod_e M_e(b_{f(e)}, b_{s(e)}) \right), \end{aligned} \tag{3.1}$$

dove il prodotto è calcolato al variare degli archi e di T orientati uscenti da r, ognuno dei quali parte dal vertice $s(e)$ e termina nel vertice $f(e)$. L'insieme H sul quale sommiamo è invece così definito:

$$H = \left\{ (b_v)_{v \in V} \colon b_v \in [q] \text{ se } v \neq a_j,\ b_v = i_j \text{ se } v = a_j \right\},$$

ovvero le possibili configurazioni delle variabili compatibili con gli stati delle foglie.
L'immagine $\phi_r(S)$ di questa applicazione è un modello statistico parametrico detto *modello filogenetico* per l'albero T e indicato con $\mathcal{M}_T$. La chiusura di $\phi_r(S)$ in $\mathbb{C}^{q^n}$ è detta *varietà filogenetica affine* ed è indicata con $V(T)$. La chiusura proiettiva di questa varietà è la *varietà* filogenetica proiettiva e la indicheremo ancora con $V(T)$, specificando quale delle due consideriamo qualora vi sia possibilità di confusione.

È possibile definire il tensore P in maniera alternativa e più semplice con la seguente procedura induttiva. Sia $\{a_1, \ldots, a_n\}$ un ordinamento delle foglie di T e chiamiamo *ciliegia* di T una coppia di foglie distinte a_{i_1}, a_{i_2} unite ad un vertice interno comune in T. Per $n \geq 3$, sia T_n^r l'albero con n foglie e radice nel vertice r; indichiamo con T_{n-1}^r l'albero ottenuto da T_n^r cancellando una ciliegia, che supponiamo abbia

come foglie a_{n-1} e a_n, e sostituendo il vertice interno appena eliminato con una nuova foglia b. In questo modo otteniamo una sequenza $T_n^r, T_{n-1}^r, \ldots, T_2^r$ di sottoalberi di T^r.
L'albero con 2 foglie ha un solo arco e e quindi una sola matrice di transizione M: lo spazio dei parametri in questo caso è $S_2 = \Delta_q \times \mathcal{M}$ e la parametrizzazione è

$$\phi_r^{(2)}(s) = P = \text{diag}(\pi_r)M,$$

dove $\text{diag}(\pi_r)$ è la matrice diagonale che ha come diagonale la distribuzione sulla radice.
Prendiamo adesso un albero con n foglie, cancelliamo una ciliegia, per esempio quella formata da a_{n-1} e a_{n-2}, e sostituiamo al vertice interno eliminato una foglia b come sopra ottenendo un albero con $n-1$ foglie T_{n-1}; sia $\tilde{P} = \phi_r^{(n-1)}(\tilde{s})$ l'immagine tramite la parametrizzazione di T_{n-1} di $\tilde{s} \in S_{n-1}$ ottenuto escludendo da $s \in S_n$ le matrici relativi agli archi della ciliegia cancellata e_1 ed e_2. Allora definiamo $\phi_r(s) = \phi_r^{(n)}(s) = P$, tensore avente sezioni

$$P(i_1, i_2, \ldots, i_{n-2}, \cdot, \cdot) = M_{e_2}^t \,\text{diag}(v) M_{e_1},$$

dove $v = \tilde{P}(i_1, \ldots, i_{n-2}, \cdot)$. Si può verificare che questa definizione è indipendente dalla scelta delle ciliegie e coincide con la precedente, e inoltre è indipendente dalla posizione della radice.

Un esempio con 3 foglie

Vediamo una parametrizzazione semplice, con variabili binarie. Le

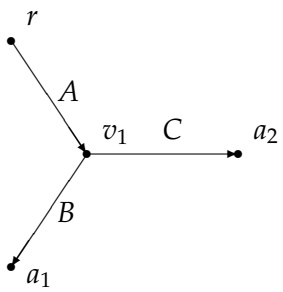


Figura 3.3: Albero filogenetico binario con 3 foglie.

matrici di transizione di questo albero filogenetico sono:

$$A = \begin{pmatrix} A_{00} & A_{01} \\ A_{10} & A_{11} \end{pmatrix} \qquad \text{per l'arco } rv_1,$$
$$B = \begin{pmatrix} B_{00} & B_{01} \\ B_{10} & B_{11} \end{pmatrix} \qquad \text{per l'arco } v_1a_1,$$
$$C = \begin{pmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{pmatrix} \qquad \text{per l'arco } v_1a_2.$$

La variabile aleatoria X_r è distribuita secondo il vettore $\pi = (\pi_0, \pi_1)$. Con queste notazioni abbiamo

$$\tilde{P} = \operatorname{diag}(\pi)A = \begin{pmatrix} \pi_0 & 0 \\ 0 & \pi_1 \end{pmatrix} \begin{pmatrix} A_{00} & A_{01} \\ A_{10} & A_{11} \end{pmatrix} = \begin{pmatrix} \pi_0 A_{00} & \pi_0 A_{01} \\ \pi_1 A_{10} & \pi_1 A_{11} \end{pmatrix}.$$

Notiamo che $\tilde{P}$ soddisfa

$$\tilde{P}(i,j) = \mathbf{P}[X_r = i, X_{v_1} = j].$$

Posti $v_0 = (\pi_0 A_{00}, \pi_0 A_{01})$ e $v_1 = (\pi_1 A_{10}, \pi_1 A_{11})$, il tensore $P = \phi_r(\pi, A, B, C)$ ha le due seguenti sezioni:

$$P(0,\cdot,\cdot) = B^t \operatorname{diag}(v_0)C = \begin{pmatrix} B_{00} & B_{10} \\ B_{01} & B_{11} \end{pmatrix} \begin{pmatrix} \pi_0 A_{00} & 0 \\ 0 & \pi_0 A_{01} \end{pmatrix} \begin{pmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{pmatrix} =$$
$$= \begin{pmatrix} \pi_0(A_{00}B_{00}C_{00} + A_{01}B_{10}C_{10}) & \pi_0(A_{00}B_{00}C_{01} + A_{01}B_{10}C_{11}) \\ \pi_0(A_{00}B_{01}C_{00} + A_{01}B_{11}C_{10}) & \pi_0(A_{00}B_{01}C_{01} + A_{01}B_{11}C_{11}) \end{pmatrix}$$

e

$$P(1,\cdot,\cdot) = B^t \operatorname{diag}(v_1)C = \begin{pmatrix} B_{00} & B_{10} \\ B_{01} & B_{11} \end{pmatrix} \begin{pmatrix} \pi_1 A_{10} & 0 \\ 0 & \pi_1 A_{11} \end{pmatrix} \begin{pmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{pmatrix} =$$
$$= \begin{pmatrix} \pi_1(A_{10}B_{00}C_{00} + A_{11}B_{10}C_{10}) & \pi_1(A_{10}B_{00}C_{01} + A_{11}B_{10}C_{11}) \\ \pi_1(A_{10}B_{01}C_{00} + A_{11}B_{11}C_{10}) & \pi_1(A_{10}B_{01}C_{01} + A_{11}B_{11}C_{11}) \end{pmatrix}.$$

L'immagine dell'applicazione $\phi_r : \Delta_2 \times \mathcal{M}^3 \longrightarrow \Delta_8$ è il modello filogenetico $\mathcal{M}_T$, e la varietà filogenetica associata è $V(T) = \sigma_2(\mathbb{P}^1 \times$

$\mathbb{P}^1 \times \mathbb{P}^1$) che, per il teorema 1.7, riempie tutto $\mathbb{P}^7$.

Un esempio con 4 foglie

Vediamo adesso una parametrizzazione più complicata, corrispondente ad un modello non studiato finora. Le matrici di transizione di

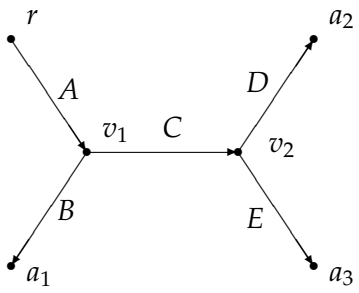


Figura 3.4: Albero filogenetico binario con 4 foglie.

questo albero filogenetico sono:

$$A = \begin{pmatrix} A_{00} & A_{01} \\ A_{10} & A_{11} \end{pmatrix} \quad \text{per l'arco } rv_1,$$

$$B = \begin{pmatrix} B_{00} & B_{01} \\ B_{10} & B_{11} \end{pmatrix} \quad \text{per l'arco } v_1a_1,$$

$$C = \begin{pmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{pmatrix} \quad \text{per l'arco } v_1v_2,$$

$$D = \begin{pmatrix} D_{00} & D_{01} \\ D_{10} & D_{11} \end{pmatrix} \quad \text{per l'arco } v_2a_2,$$

$$E = \begin{pmatrix} E_{00} & E_{01} \\ E_{10} & E_{11} \end{pmatrix} \quad \text{per l'arco } v_2a_3.$$

La variabile aleatoria X_r è distribuita secondo il vettore $\pi = (\pi_0, \pi_1)$. Chiamiamo $\tilde{P}$ il tensore $2 \times 2 \times 2$ dell'esempio precedente: posto $s = (\pi, A, B, C, D, E)$, cerchiamo il tensore $P = \phi_r(s)$.
Poniamo $v_{00} = \tilde{P}(0,0,\cdot)$, $v_{01} = \tilde{P}(0,1,\cdot)$, $v_{10} = \tilde{P}(1,0,\cdot)$, $v_{11} =$

$\tilde{P}(1,1,\cdot)$. Allora P è il tensore $2\times 2\times 2\times 2$ le cui sezioni sono

$$P(i,j,\cdot,\cdot) = D^t \operatorname{diag}(v_{ij}) E.$$

Ad esempio

$$P(0,0,\cdot,\cdot) = \\ = \begin{pmatrix} D_{00} & D_{10} \\ D_{01} & D_{11} \end{pmatrix} \begin{pmatrix} \pi_0(A_{00}B_{00}C_{00} + A_{01}B_{10}C_{10}) & 0 \\ 0 & \pi_0(A_{00}B_{00}C_{01} + A_{01}B_{10}C_{11}) \end{pmatrix} \begin{pmatrix} E_{00} & E_{01} \\ E_{10} & E_{11} \end{pmatrix}.$$

Da cui

$$\begin{aligned} P(0,0,0,0) =& \pi_0((A_{00}B_{00}C_{00} + A_{01}B_{10}C_{10})D_{00}E_{00} + (A_{00}B_{00}C_{01} + A_{01}B_{10}C_{11})D_{10}E_{10}), \\ P(0,0,0,1) =& \pi_0((A_{00}B_{00}C_{00} + A_{01}B_{10}C_{10})D_{00}E_{01} + (A_{00}B_{00}C_{01} + A_{01}B_{10}C_{11})D_{10}E_{11}), \\ P(0,0,1,0) =& \pi_0((A_{00}B_{00}C_{00} + A_{01}B_{10}C_{10})D_{01}E_{00} + (A_{00}B_{00}C_{01} + A_{01}B_{10}C_{11})D_{11}E_{10}), \\ P(0,0,1,1) =& \pi_0((A_{00}B_{00}C_{00} + A_{01}B_{10}C_{10})D_{01}E_{01} + (A_{00}B_{00}C_{01} + A_{01}B_{10}C_{11})D_{11}E_{11}). \end{aligned}$$

La parametrizzazione $\phi_r : \Delta_2 \times \mathcal{M}^5 \longrightarrow \Delta_{16}$ definisce il modello filogenetico M_T, la cui chiusura proiettiva è la varietà filogenetica $V(T) \subseteq \mathbb{P}^{15}$.

Parametrizzazione alternativa

La parametrizzazione fin qui introdotta è molto intuitiva ma con una piccola modifica è possibile introdurne un'altra senza fare uso di condizioni non omogenee come quella sulle matrici stocastiche.
Sia $U = \mathbb{C}^{|E|q^2}$ l'insieme delle famiglie matrici $q \times q$ a coefficienti complessi associate agli archi di T. Definiamo una mappa polinomiale omogenea $\psi : U \to \mathbb{C}^{q^n}$ in maniera del tutto analoga a quanto fatto con ϕ_r:

- Per $n = 2$ la mappa ψ è l'identità $\mathbb{C}^{q^2} \to \mathbb{C}^{q^2}$.
- Per n generico ricalchiamo la costruzione fatta per ϕ_r.

Proposizione 3.2. $\overline{\psi(U)} = CV(T)$, *il cono sulla varietà filogenetica* $V(T)$.

Dimostrazione. • $CV(T) \subseteq \overline{\psi(U)}$

Vediamo che $\phi_r(S) \subseteq \psi(U)$. Sia $s = (\pi, \{M_e\}) \in S$, sia e_0 l'arco di T_2 e poniamo $M'_{e_0} = \mathrm{diag}(\pi)M_{e_0}$. Si ha $\phi_r(s) = \psi(u)$ con $u = \left(M_{e_0}, \{M_e\}_{e \neq e_0}\right)$ e quindi $V(T) \subseteq \overline{\psi(U)}$.
Poiché $\psi(U)$ è un cono si ha $CV(T) \subseteq \overline{\psi(U)}$.

- $\overline{\psi(U)} \subseteq CV(T)$

  Supponiamo per semplicità che T sia binario.
  Mostriamo che esiste un sottoinsieme $A \subseteq U$ aperto non vuoto (e quindi denso) tale che $\psi(A) \subseteq CV(T)$.
  Se $n = 2$ allora $\phi_r(S)$ contiene le matrici di probabilità bidimensionali le cui righe hanno sempre somma non nulla. Infatti, data una matrice (v_{ij}) così fatta, basta prendere $(\pi, M) \in S$ così fatte:

$$\pi = \left(\sum_i v_{1i}, \ldots, \sum_i v_{qi}\right) \qquad M = \begin{pmatrix} \frac{v_{11}}{\sum_i v_{1i}} & \cdots & \frac{v_{1q}}{\sum_i v_{1i}} \\ \vdots & \ldots & \vdots \\ \frac{v_{q1}}{\sum_i v_{qi}} & \cdots & \frac{v_{qq}}{\sum_i v_{qi}} \end{pmatrix}.$$

  Ora, l'insieme delle matrici in $U = \mathbb{C}^2$ la cui somma degli elementi è diversa da zero e aventi somma delle righe diversa da zero è un aperto in U la cui immagine è contenuta nel cono su $\phi_r(S)$, ed è quindi l'aperto che stavamo cercando.
  Procedendo per induzione, siano e_1 e e_2 gli archi di T_n non appartenenti a T_{n-1}, e sia e_3 l'arco che li unisce al resto dell'albero. Possiamo supporre che r non sia il vertice comune a e_1, e_2 e e_3.
  Esiste un aperto $A_1 \subseteq U$ tale che per ogni $u \in A_1$ la somma delle righe delle matrici $M_{e_1}, M_{e_2} \in u$ è sempre diversa da zero. Sia D_i la matrice diagonale i cui elementi diagonali sono la somma delle righe di M_{e_i}, si ha come prima $M_{e_i} = D_i M'_{e_i}$ e la somma delle righe di ogni M'_{e_i} è 1. Sia $M'_{e_3} = M_{e_3} D_1 D_2$. Allora, per ogni $u \in A_1$, sia $u' = \left(\{M_e\}_{e \neq e_1, e_2, e_3}, M'_{e_1}, M'_{e_2}, M'_{e_3}\right) \in A_1$. Osserviamo che $\psi(u) = \psi(u')$ e che $\omega : u \mapsto u'$ è una mappa razionale.
  Siano $\psi_{n-1} : U_{n-1} \to \mathbb{C}^{q^{n-1}}$ e $\phi^r_{n-1} : S_{n-1} \to \mathbb{C}^{q^{n-1}}$ le parametrizzazioni associate a T_{n-1}. Per ipotesi induttiva esiste un insieme $\tilde{A} \subseteq U_{n-1}$ per il quale $\psi_{n-1}(\tilde{A})$ è contenuto nel cono su $\phi^r_{n-1}(S_{n-1})$. Allora $A = \omega^{-1}(\tilde{A} \times \mathbb{C}^{2q^2})$ è l'aperto che

cercavamo.

□

Grazie a questa nuova costruzione considereremo d'ora in avanti lo spazio dei parametri U e la parametrizzazione ψ per definire la varietà filogenetica $V(T)$.

3.3 Il modello generale di Markov: invarianti

La descrizione di un modello statistico tramite parametrizzazione è spesso la più naturale e la più semplice, ma ha il grande svantaggio di non essere direttamente applicabile ai dati. Per valutare se i dati a nostra disposizione provengono da un certo modello, infatti, dovremmo rintracciare a partire da essi i parametri di partenza, cosa molto complicata quando non impossibile.
Inoltre, se la parametrizzazione non è iniettiva, avremo anche il fenomeno della *non identificabilità*, ovvero la possibilità che a più parametri differenti corrisponda la stessa distribuzione, con conseguente impossibilità di sapere quali siano quelli originari.
Infine, ricordiamo che il nostro obiettivo è la scelta di un modello appropriato ai dati, non la ricostruzione dei parametri: dovendo scegliere fra numerosi modelli, l'identificazione diventa quindi, oltre che ancora più complicata, superflua.
Perciò è molto importante riuscire a trovare equazioni che caratterizzino gli elementi del modello indipendentemente dai parametri.

Definizione. Un polinomio $f \in \mathbb{C}[\mathbf{p}]$ si dice *invariante filogenetico* per l'albero T se $f(p) = 0$ per ogni $p \in V(T)$. L'ideale degli invarianti filogenetici per un albero T è detto *ideale filogenetico* I_T.

L'idea alla base del nostro approccio alla ricerca degli invarianti è vedere l'albero T come un modello grafico non solo per le variabili alle foglie anche per i propri sottoalberi. Sostanzialmente, cercheremo di ricavare gli invarianti per T considerando di volta in volta l'albero come un modello grafico più semplice relativo a variabili più complicate. Unendo poi le informazioni ottenute da ognuna di queste "semplificazioni" otterremo una descrizione completa di T.

Flattening rispetto agli archi

Per prima cosa concentriamoci sugli archi.

Definizione. Sia T un albero con n foglie e sia p il tensore di probabilità $q \times q \times \cdots \times q$ associato: fissiamo $e \in E$ e supponiamo di cancellarlo da T. La conseguente sconnessione induce una partizione $\{A, B\}$delle foglie di T, detta *split* rispetto ad e.

Siano $\mathcal{A} = \{X_v \colon v \in A\}$ e $\mathcal{B} = \{X_v \colon v \in B\}$ le due famiglie di variabili corrispondenti a questa partizione, e siano X_A eX_B le variabili a valori rispettivamente in $q^{|A|}$ e $q^{|B|}$ aventi come distribuzione le distribuzioni congiunte delle variabili in $\mathcal{A}$ e $\mathcal{B}$.

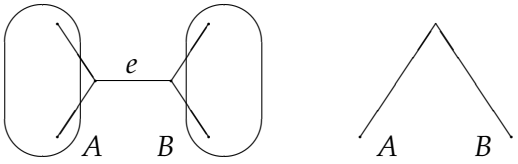


Figura 3.5: Flattening rispetto ad un arco.

Come si nota dalla figura 3.5 il modello per le variabili X_A e X_B è il modello a classi latenti con variabile latente che assume valori in $[q]$. Se $\tilde{p}$ è il tensore di probabilità per questo modello, abbiamo allora

$$\mathrm{rk}(\tilde{p}) \leq q$$

e, poiché $\tilde{p} = \mathrm{Flat}_{A,B}(p)$, otteniamo

$$\mathrm{rk}(\mathrm{Flat}_{A,B}(p)) \leq q \tag{3.2}$$

per ogni split $\{A, B\}$ dell'albero T. Per comodità di notazione indicheremo questo particolare flattening con $\mathrm{Flat}_e(p)$ e lo chiameremo *flattening rispetto all'arco e*.
La disequazione 3.2 ci fornisce un primo insieme di invarianti per T, poiché i polinomi (omogenei) dell'insieme

$$\mathcal{F}_{edge}(T) = \{\text{minori } (q+1) \times (q+1) \text{ di } \mathrm{Flat}_e(p) \colon e \in E\}$$

si annullano, per quanto visto finora, per ogni $p \in V(T)$.
L'ideale $I_{edge}(T)$ generato da $\mathcal{F}_{edge}(T)$ è detto *ideale degli archi di T* e è una componente di I_T.
In generale l'inclusione $I_{edge}(T) \subseteq I_T$ è stretta, ma esiste un caso particolare in cui vale l'uguaglianza, quello delle variabili binarie, la cui dimostrazione compare in [3]:

Teorema 3.3. *Sia T un albero binario con n foglie. Se $q = 2$ l'ideale I_T è generato da $\mathcal{F}_{edge}(T) = \{$minori 3×3 di $\text{Flat}_e(p) : e \in E\}$.*

Questo non è vero in generale per $q > 2$: ad esempio, se T è un albero binario con 3 foglie e $q = 4$, l'ideale generato dai flattening rispetto agli archi è $\{0\}$, ma $V(T) = \sigma_4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3) \neq \mathbb{P}^{63} = V(\{0\})$ in quanto $\dim(V(T)) \leq d_e(V(T)) = 39$.

Flattening rispetto ai vertici

Dal momento che considerare soltanto i flattening rispetto gli archi non è sufficiente, estendiamo le considerazioni esposte nello scorso paragrafo alle partizioni prodotte dalla cancellazione dei vertici.
Sia quindi T un albero (che supporremo per semplicità binario) e $\{A, B, C\}$ la partizione delle foglie indotta dalla sconnessione di T dovuta alla cancellazione di un vertice interno.
Definiamo le 3 variabili aleatorie X_A, X_B, X_C in maniera analoga ai flattening rispetto agli archi. Sia $\tilde{p}$ il tensore $q^{|A|} \times q^{|B|} \times q^{|C|}$ associato al modello per queste tre variabili.

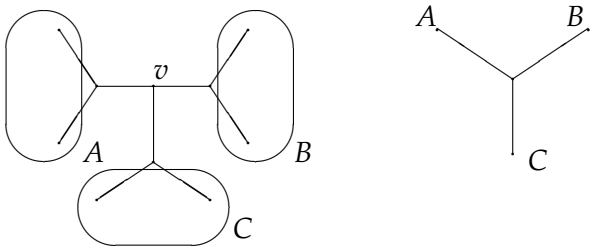


Figura 3.6: Flattening rispetto ad un vertice.

Come si nota dalla figura 3.6, $\tilde{p}$ è un elemento del modello a classi latenti con 3 variabili e variabile latente a valori in $[q]$. La condizione

nota

$$\mathrm{rk}(\tilde{p}) \leq q$$

per questo modello però , a differenza del caso precedente, non dà direttamente delle equazioni per il calcolo degli invarianti, dal momento che $\tilde{p}$ non è una matrice ma un tensore 3-dimensionale. Denoteremo $\tilde{p}$ con $\mathrm{Flat}_v(p)$ e lo chiameremo *flattening rispetto al vertice v*.
Nel caso che T non sia binario e che v abbia valenza maggiore di 3, avremo tensori di dimensione maggiore, ma la condizione sul rango è la stessa.
Continueremo la nostra ricerca degli invarianti concentrandoci sugli alberi detti *a stella*, ovvero quelli associati al modello a classi latenti.

3.4 Il modello generale di Markov: alberi a stella

Definizione. Un albero è detto *a stella* se esiste un unico vertice interno.

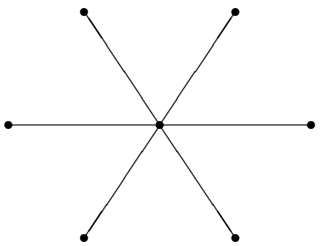


Figura 3.7: Un albero a stella.

Scegliendo come radice il vertice interno r abbiamo il grafo per il modello a classi latenti.
Supponiamo questa volta che soltanto X_r sia a valori in $[q]$, mentre le variabili alle n foglie possono assumere valori $l_1, \ldots, l_n$. Denotiamo con $V(q; l_1, \ldots, l_n)$ la varietà filogenetica associata a quest'albero. Mostreremo che con semplici operazioni possiamo ricondurre lo studio di tutte queste varietà al caso $V(q; q, \ldots, q)$.
Per prima cosa definiamo un'operazione fra tensori:

Definizione. Sia Q un tensore $l_1 \times \cdots \times l_n \times k$ e R un tensore $k \times r_1 \times \cdots \times r_m$. Definiamo $Q * R$ il tensore $m+n$-dimensionale per cui

$$(Q * R)(i_1, \ldots, i_n, j_1, \ldots, j_m) = \sum_{\alpha=1}^{k} Q(i_1, \ldots, i_n, \alpha) R(\alpha, j_1, \ldots, j_m).$$

Abbiamo scelto l'ultima coordinata di Q e la prima di R per semplicità di notazione, ma chiaramente questa operazione si può definire per ogni coppia di indici che variano nello stesso insieme. Notiamo che se Q e R sono matrici l'operazione $*$ coincide con l'usuale prodotto di matrici.

Definiamo inoltre un'operazione fra alberi:

Definizione. Siano T' un albero con n foglie $\{a_1, \ldots a_n\}$ e T'' un albero con m foglie $\{b_1, \ldots, b_m\}$. $T' * T''$ è l'albero con $n+m-2$ foglie ottenuto identificando le foglie a_n e b_1, cancellandole insieme agli archi ad esse incidenti e congiungendo i vertici adiacenti agli archi appena cancellati con un nuovo arco.

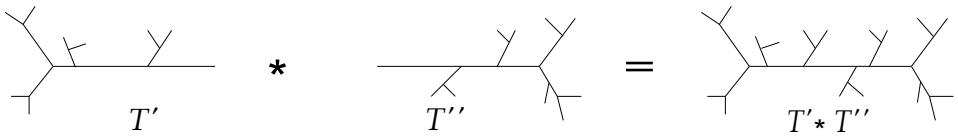


Figura 3.8: Operazione sugli alberi.

Se poniamo la radice di T' in a_1 e quella di T'' in b_1, $T' * T''$ ha un'orientazione indotta coerente con quella che solitamente utilizziamo.
Siano U', U'' e U gli spazi dei parametri rispettivamente di T', T'' e $T' * T''$, e ψ', ψ'' e ψ le relative parametrizzazioni. Dati $u' \in U'$ e $u'' \in U''$ otteniamo un elemento $u' * u'' \in U$ mantenendo le matrici originali per ogni arco diverso da quello di congiunzione e associando a quest'ultimo il prodotto delle matrici di transizione per gli archi cancellati.
Con questa scelta vale la seguente

Proposizione 3.4. - $\psi(u' * u'') = \psi'(u') * \psi''(u'')$

- $CV(T' * T'') = \overline{CV(T') * CV(T'')}$.

Un caso particolare molto importante di questa operazione si verifica quando T'' è un albero con 2 foglie. In questo casto $T' * T''$ ha la stessa topologia di T' e possiamo vedere $\psi'(u') * \psi''(u'')$ come l'estensione dell'arco incidente nell'n-esima foglia.
Se restringiamo U'' a $GL(q, \mathbb{C})$ possiamo parlare di *azione* del gruppo $GL(q, \mathbb{C})$ su U' e $\psi'(U')$. Ovviamente questa azione non è limitata all'n-esima foglia ma può essere rispetto ad un qualsiasi indice di $\psi'(U')$.

Questa costruzione è necessaria per il seguente teorema, che circoscrive la ricerca degli invarianti filogenetici, la cui lunga dimostrazione è riportata in [3].

Teorema 3.5. *Sia T un albero con n foglie. Supponiamo $l_1, \dots, l_n \geq q$ e sia $\mathcal{F}$ un insieme di polinomi tali che $V(\mathcal{F}) = V(q; q, \dots, q)$.*
Per $k = 1, \dots, n$ siano $Z_k = (z_{ij}^k)$ matrici $l_k \times q$ di indeterminate. Sia P un tensore $l_1 \times \cdots \times l_n$ di indeterminate e sia $\tilde{P}$ il tensore $q \times \cdots \times q$ ottenuto facendo agire ogni Z_k sul k-esimo indice di P.
Indichiamo con $\tilde{\mathcal{F}}$ l'insieme dei polinomi ottenuti valutando i polinomi in $\mathcal{F}$ negli elementi di $\tilde{P}$ ed estraendo i coefficienti degli z_{ij}^k.
Allora $V(\tilde{\mathcal{F}} \cup \mathcal{F}_{edge}(T)) = V(q; l_1, \dots, l_n)$.
Inoltre, se $\mathcal{F} = I(V(q; q, \dots, q))$, $\tilde{\mathcal{F}} \cup \mathcal{F}_{edge}(T)$ genera $I(V(q; l_1, \dots, l_n))$.

Il teorema ci dice che per ricavare gli invarianti per qualsiasi albero a stella è sufficiente conoscere gli invarianti per il caso $V(q; q, \dots, q) = \sigma_q(\mathbb{P}^{q-1} \times \cdots \times \mathbb{P}^{q-1})$, che sono quindi le varietà fondamentali anche per la costruzione di tutti gli invarianti filogenetici.

Varietà per gli alberi binari a stella

Consideriamo i casi fondamentali quando $n = 3$, studiati ad esempio in [17].

Caso $q = 2$. La varietà $V(2; 2, 2, 2) = \sigma_2(\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1)$ riempie tutto $\mathbb{P}^7$, pertanto il suo ideale è quello banale $\{0\}$.

Caso $q = 3$ L'ideale della varietà $V(3; 3, 3, 3) = \sigma_3(\mathbb{P}^2 \times \mathbb{P}^2 \times \mathbb{P}^2)$ è generato da 27 polinomi di grado 4.

Caso $q = 4$ Questo è il caso più importante perché sarà quello che sfrutteremo per la ricostruzione filogenetica. È stata dimostrato recentemente ([11]) che la varietà $V(4;4,4,4)$ è definita da un insieme di polinomi di grado 5, 6 e 9. Non è ancora noto invece se questo insieme di polinomi generi l'ideale della varietà: la prof.ssa E. Allman ha promesso a chiunque riesca a stabilire se ciò sia vero un salmone da lei personalmente pescato, affumicato e spedito (da qui il nome di *congettura del salmone*).

Invarianti per $V(q;q,q,q)$

Dal momento che abbiamo ricondotto il nostro studio alla determinazione di $V(q;q,\dots,q)$, è necessario trovare i polinomi che definiscano queste varietà . Abbiamo già detto che il caso $q = 2$ è banale, quindi supponiamo $q > 3$. Inoltre, visto che ci concentreremo sugli alberi binari, considereremo solo il caso $n = 3$.
Sia $P \in V(q;q,q,q)$. Denotiamo con $P_{\cdot\cdot i}$ la matrice ottenuta fissando a i il terzo indice di P e definiamo $P_{\cdot\cdot +} = \sum_{i=1}^{q} P_{\cdot\cdot i}$.

Proposizione 3.6. *Per ogni* $i, j \in [q]$ *si ha*

$$P_{\cdot\cdot i}(P_{\cdot\cdot +})^{-1}P_{\cdot\cdot j} = P_{\cdot\cdot j}(P_{\cdot\cdot +})^{-1}P_{\cdot\cdot i} \tag{3.3}$$

Questo risultato ci permette di trovare un primo insieme di invarianti detti di *commutazione*, introdotti per la prima volta da Strassen in [21] e ricavati indipendentemente anche in [2].
Per ogni coppia (i,j) abbiamo infatti q^2 polinomi di grado $q+1$ contenuti nell'ideale filogenetico.
Per $q = 3$ inoltre questo insieme di invarianti definisce interamente l'ideale, mentre per $q = 4$ questo fornisce soltanto la componente di grado 5.
Usando argomenti simili in [11] e [12] si mostra come ricavare anche i polinomi di grado 6 e 9.

3.5 Invarianti per gli alberi binari

Lo studio degli alberi a stella conclude molto probabilmente la ricerca degli invarianti.

Congettura 3.7 ([3], Congettura 5)**.** *Per ogni q e per ogni numero di foglie n di un albero filogenetico binario T, l'ideale filogenetico I_T è la somma degli ideali associati ai flattening di P nei vertici di T.*

Se $q = 2$ ritroviamo il teorema 3.3, in quanto l'ideale associato al flattening ad un vertice v è la somma degli ideali associati agli archi incidenti in v, come dimostrato in [17] da Landsberg e Manivel.
Dimostriamo adesso che i polinomi prodotti dai flattening rispetto a lati e vertici definiscono la varietà filogenetica.

Sia T un albero con n foglie e per ogni $v \in V$ sia $\mathcal{F}_v(T)$ l'insieme di polinomi $\tilde{\mathcal{F}} \cup \mathcal{F}_{edge}$ definiti nel teorema 3.5 associati al flattening di P in v.
Poniamo $\mathcal{F}(T) = \bigcup_{v \in V} \mathcal{F}_v(T)$ e sia $V_f(T) = V(\mathcal{F}(T))$.

Lemma 3.8. *Siano T', T'' alberi filogenetici rispettivamente con n e m foglie, e sia $T = T' * T''$. Se $Q \in CV_f(T')$ e $R \in CV_f(T'')$, allora $Q * R \in CV_f(T)$.*

Lemma 3.9. *Siano T', T'' alberi filogenetici rispettivamente con n e m foglie, e sia $T = T' * T''$. Se $P \in CV_f(T)$ allora esistono $Q \in CV_f(T')$ e $R \in CV_f(T'')$ tali che $P = Q * R$.*

Siamo ora pronti a dimostrare il teorema:

Teorema 3.10. *Sia T un albero con n foglie. Allora $V_f(T) = V(T)$.*

Dimostrazione. Supponiamo per semplicità che T sia binario.
Abbiamo già visto che $\mathcal{F} \subseteq I_T$ e quindi $V(T) \subseteq V_f(T)$. Per dimostrare l'inclusione opposta procediamo per induzione su n numero di foglie dell'albero.

I casi $n = 2$ e $n = 3$ sono banali, in quanto i flattening rispetto ai lati e rispetto ai vertici interni restituiscono il tensore di partenza.

Sia T un albero binario con $n \geq 4$ foglie. Siano T_{n-1} e T_3 sottoalberi di T rispettivamente con $n-1$ e 3 foglie e tali che $T = T_{n-1} * T_3$ (li troviamo ad esempio scegliendo una ciliegia di T).
Per ogni $P \in CV_f(T)$ abbiamo, per il lemma 3.9, $Q \in CV_f(T_{n-1})$ e $R \in CV_f(T_3)$ tali che $P = Q * R$. Dunque, per la proposizione 3.4,

l'applicazione

$$\begin{aligned} \mu : CV_f(T_{n-1}) \times CV_f(T_3) &\longrightarrow CV_f(T) \\ (Q, R) &\longmapsto Q * R \end{aligned}$$

è suriettiva.
Indichiamo con ψ_{n-1} e ψ_3 le parametrizzazioni di $CV_f(T_{n-1})$ e di $CV_f(T_3)$, e siano U_{n-1} e U_3 i rispettivi spazi dei parametri. Allora, per il lemma il seguente diagramma commuta:

$$\begin{array}{ccc} U_{n-1} \times U_3 & \xrightarrow{\psi_{n-1} \times \psi_3} & CV_f(T_{n-1}) \times CV_f(T_3) \\ \downarrow \alpha & & \downarrow \mu \\ U_n & \xrightarrow{\psi_n} & CV_f(T) \end{array} ,$$

dove $\alpha(u_{n-1}, u_3) = u_{n-1} * u_3$.
Poiché α e μ sono suriettive, e, per ipotesi induttiva, $\overline{\psi_{n-1} \times \psi_3(U_{n-1} \times U_3)} = CV_f(T_{n-1}) \times CV_f(T_3)$, abbiamo $\overline{\psi_(U_n)} = CV_f(T)$ e dunque $V_f(T) = V(T)$.

□

In conclusione, la procedura per costruire un insieme di invarianti filogenetici per un albero filogenetico T le cui variabili abbiano valori in $[q]$ è la seguente.

Sia p un tensore $\underbrace{q \times q \times \cdots \times q}_{n}$ di indeterminate.

- Per ogni arco e di T, sia $\mathcal{F}_e(T)$ l'insieme dei minori $(q+1) \times (q+1)$ di $\text{Flat}_e(p)$.
- Per ogni vertice v di T, sia $\mathcal{F}_v(T)$ l'insieme dei polinomi ottenuti applicando le equazioni di $V(q; l_1, l_2, l_3)$ agli elementi di $\text{Flat}_v(p)$.
- $\mathcal{F}(T) = (\bigcup_{e \in E} \mathcal{F}_e(T)) \cup (\bigcup_{v \in V} \mathcal{F}_v(T))$ è l'insieme di invarianti filogenetici cercato.

Esempio di invarianti filogenetici

Vediamo un esempio di invarianti filogenetici nel caso dell'albero in figura 3.9.

Supponiamo che le variabili alle foglie $a_1, \ldots, a_5$ siano binarie.

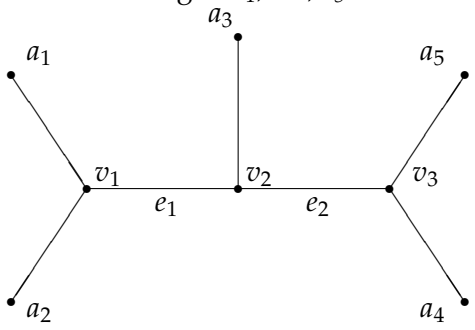


Figura 3.9: Albero filogenetico con 5 foglie.

Per il teorema 3.3 l'ideale del modello è generato dai flattening agli archi e_1 e e_2. Sia $p_{i_1 i_2 i_3 i_4 i_5}$ un tensore $2 \times 2 \times 2 \times 2 \times 2$: le equazioni cercate sono date dall'annullarsi dei minori 3×3 delle matrici

$$\mathrm{Flat}_{e_1}(p) = \begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} & p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & p_{01011} & p_{01100} & p_{01101} & p_{01110} & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & p_{10011} & p_{10100} & p_{10101} & p_{10110} & p_{10111} \\ p_{11000} & p_{11001} & p_{11010} & p_{11011} & p_{11100} & p_{11101} & p_{11110} & p_{11111} \end{pmatrix}$$

e

$$\mathrm{Flat}_{e_2}(p) = \begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} \\ p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & p_{01011} \\ p_{01100} & p_{01101} & p_{01110} & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & p_{10011} \\ p_{10100} & p_{10101} & p_{10110} & p_{10111} \\ p_{11000} & p_{11001} & p_{11010} & p_{11011} \\ p_{11100} & p_{11101} & p_{11110} & p_{11111} \end{pmatrix}.$$

Ad esempio, alcuni invarianti sono

$$f_1 = p_{00000}(p_{01001}p_{10000} - p_{01010}p_{10001}) - p_{01000}(p_{00001}p_{10010} - p_{00010}p_{10001}) + \\ + p_{10000}(p_{00001}p_{01010} - p_{00010}p_{01001})$$

$$f_2 = p_{01100}(p_{10001}p_{10110} - p_{10010}p_{10101}) - p_{10000}(p_{01101}p_{10110} - p_{01110}p_{10101}) + \\ + p_{10100}(p_{01101}p_{10010} - p_{00111}p_{10001})$$

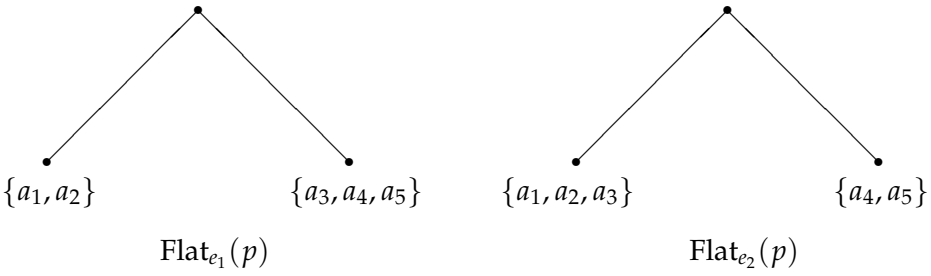


Figura 3.10: Flattening rispetto agli archi e_1 e e_2.

Poiché le variabili sono binarie, non c'è bisogno di considerare i flattening ai vertici, che infatti, essendo $\sigma_2(\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1) = \mathbb{P}^7$, non danno alcun contributo all'ideale.

4 Metodi Statistici per la ricostruzione di alberi filogenetici

Ci occuperemo d'ora in avanti dell'applicazione dei modelli fin qui studiati all'*analisi filogenetica*, ovvero quella branca della biologia che si occupa della ricostruzione della storia evolutiva delle specie viventi.
Supponiamo di avere a nostra disposizione i dati relativi alle sequenze di DNA di n diverse specie esistenti: la domanda alla quale vorremmo poter rispondere è

> *È possibile ricostruire l'ordine delle mutazioni avvenute nel passato e che hanno portato alle specie attuali?*

Per risolvere questo problema è necessario in primo luogo formulare delle ipotesi riguardo la storia evolutiva delle specie in esame e confrontarle con i dati in modo da escludere quelle che non sono compatibili con l'evidenza sperimentale. Ad ognuna delle ipotesi da verificare corrisponde un determinato albero filogenetico, in cui i vertici interni rappresentano le specie estinte, per le quali non abbiamo quindi dati disponibili, mentre le foglie rappresentano le specie esistenti che vogliamo studiare. Le mutazioni avvengono lungo gli archi dell'albero, che collegano le specie più vicine fra di loro evolutivamente.

Supponiamo di avere a disposizione per ognuna delle n specie in esame una sequenza di DNA, ovvero una sequenza di lettere dell'alfabeto $\{A, C, G, T\}$ di lunghezza fissata l. Tramite un'operazione chiamata *allineamento* è possibile disporre le sequenze in modo da far corrispondere verticalmente posizioni nella sequenza relative ad uno stesso sito di DNA.
Per quanto riguarda il nostro studio, dunque, un allineamento è una matrice $n \times l$ di caratteri $\{A, C, G, T\}$. Ogni colonna è una realizzazione indipendente dalle altre di una n-upla $X_1, \ldots, X_n$ di variabili aleatorie, a valori in $\{A, C, G, T\}$, corrispondenti alle specie in esame: possiamo quindi vedere l'allineamento come un insieme di l esperimenti indipendenti.
Esistono tecniche per allineare opportunamente sequenze di DNA,

che si basano sulla minimizzazione del numero di mutazioni per sito, ma di questo non ci occuperemo. È sufficiente sapere che a partire da campioni di DNA è possibile ricavare un allineamento.
Supponiamo infine di poter interpretare la relazione di vicinanza evolutiva fra le n specie come relazioni di indipendenza rappresentabili da un albero filogenetico con n foglie.
Grazie allo studio appena compiuto sugli invarianti saremo in grado di trovare algoritmi in grado stabilire quale modello meglio si adatti al un dato allineamento di DNA sotto queste ipotesi.

4.1 Ricostruzione di alberi filogenetici tramite invarianti

Per prima cosa dobbiamo trovare una stima $\hat{p}$ per p, il tensore di probabilità $q \times \cdots \times q$ relativo alle variabili $X_1, \dots X_n$, a partire dall'allineamento.
Per fare questo utilizzeremo il tensore delle frequenze:

$$\hat{p}_{i_1,\dots,i_n} = \frac{\text{Numero di occorrenze della colonna } i_1 \dots i_n}{l}.$$

Ad esempio, se le specie sono 4 e la colonna $ACAG$ compare 10 volte in un allineamento di lunghezza 1000, avremo $\hat{p}_{ACAG} = 0,01$.
L'idea alla base del metodo degli invarianti è confrontare $\hat{p}$ con gli invarianti di *ogni possibile albero filogenetico* poiché sappiamo che se $p \in V(T)$ allora $f(p) = 0$ per ogni $f \in \mathcal{F}(T)$. Purtroppo noi abbiamo a disposizione soltanto una stima di p, perciò è praticamente impossibile che gli invarianti si annullino effettivamente in p.
Ciò che faremo allora sarà valutare gli invarianti negli elementi di $\hat{p}$ e scegliere l'albero per il quale questi risultano più "piccoli".

Algoritmo 1. *Metodo degli invarianti*

Input: Un allineamento di dati genomici da n specie da un alfabeto Σ con q stati.

Output: Un albero binario con n foglie.

Passo 1: Calcolare le probabilità empiriche $\hat{p}_{i_1 \dots i_n}$ e scriverle in un tensore $\hat{p}$.

Passo 2: Per ogni T_i topologia di alberi binari con n foglie

- determinare $\mathcal{F}(T_i)$ insieme di invarianti filogenetici.
- Calcolare
$$t_i = \sum_{f \in \mathcal{F}(T_i)} |f(\hat{p})|.$$

Passo 3: Scegliere T_i corrispondente al minimo t_i.

Poiché, per $l \to \infty$, $\hat{p} \to p$ e le funzioni t_i sono continue, l'algoritmo ricostruisce correttamente l'albero filogenetico tanto meglio quanto più sono lunghe le sequenze.

Purtroppo questo algoritmo ha un costo computazionale elevato: infatti, al crescere di n numero di specie aumentano rapidamente sia il numero di alberi binari da confrontare sia la cardinalità dell'insieme degli invarianti per albero.
In particolare, si può dimostrare che il numero di topologie di alberi binari con n foglie è dato dal numero di Schröder $(2n-5)!! = 1 \times 3 \times \cdots \times (2n-7) \times (2n-5)$ (Tabella 4.1).

Dover determinare l'insieme degli invarianti per ogni topologia diventa quindi computazionalmente impraticabile quando n è grande ed è quindi necessario trovare metodi che permettano di selezionare opportunamente solo alcuni alberi e invarianti.
Sappiamo però (si veda per esempio [2]) che non è sempre necessario considerare l'intero insieme degli invarianti per ritrovare la corretta topologia dell'albero: quali sono quindi quelli che si comportano meglio? Quali sono sufficienti per la ricostruzione filogenetica? È possibile trovare una norma migliore della norma $L_1 = \sum |\cdot|$ (usata nell'algoritmo) per valutare gli invarianti?

n	$\sharp$ alberi
3	1
4	3
5	15
6	105
7	945
8	10395
9	135135
10	2027025
$\vdots$	$\vdots$

Tabella 4.1: Numero di alberi binari con n foglie.

Rispondere a queste domande permetterebbe di migliorare notevolmente le prestazioni e soprattutto l'effettiva applicabilità dell'algoritmo.

In [4], riprendendo la costruzione degli alberi filogenetici introdotta in [7], è presentato un risultato molto importante grazie al quale possiamo concentrarci solo sugli invarianti degli archi:

Teorema 4.1. *Sia $\mathcal{T}$ l'insieme delle topologie di alberi binari con n foglie. Per ogni $T \in \mathcal{T}$ esiste un aperto U_T tale che, se $p \in \bigcup_{T\in\mathcal{T}} U_T$, allora*

$$p \in V(T_0) \iff f(p) = 0 \; \forall f \in \mathcal{F}_{edge}(T_0).$$

4.2 Ricostruzione di alberi filogenetici con il metodo SVD

In [9] è presentato un algoritmo alternativo per la ricostruzione filogenetica, che ha il vantaggio di non richiedere il confronto di tutte le possibili topologie di alberi binari, così da essere applicabile ad allineamenti con un numero di specie assai superiore rispetto a quanto è possibile fare con il metodo degli invarianti.
L'idea di base è cercare di ritrovare gli split dell'albero filogenetico: il seguente teorema di Buneman (equivalenza degli split) ci permette infatti di affermare che ciò è sufficiente a risalire all'albero.

Definizione. Due bipartizioni $\{A_1, B_1\}, \{A_2, B_2\}$ delle foglie di un albero T si dicono *compatibili* se almeno uno degli insiemi $A_1 \cap A_2$, $A_1 \cap B_2$, $B_1 \cap A_2$, $B_1 \cap B_2$ è vuoto.

Osservazione. Due split di un albero sono sempre compatibili fra di loro. Notiamo inoltre che ogni albero binario con n foglie ha $2n-3$ archi e quindi esistono altrettanti split per le sue foglie.

Teorema 4.2 (Equivalenza degli split, [19], teorema 2.35)**.** *Sia V un insieme di n vertici e sia S una collezione di $2n-3$ bipartizioni di V. Allora gli elementi di S sono due a due compatibili se e soltanto se S è l'insieme degli split di un albero binario T.*
Inoltre, tale albero è unico.

Dobbiamo adesso trovare un modo per stabilire quando una partizione delle foglie è uno split.

Definizione. Sia $\{A, B\}$ una partizione di $[n]$. La *distanza* fra $\{A, B\}$ e lo split più vicino è $e(A, B) = \sharp$ archi in $T_A \cap T_B$, dove T_A e T_B sono rispettivamente i sottoalberi generati da A e da B.

Osserviamo che $e(A, B) = 0$ se e soltanto se $\{A, B\}$ è uno split.
Consideriamo $T_A \cap T_B$ come sottoalbero di T_A; coloriamo di rosso i vertici di $T_A \cap T_B$ e di blu gli altri vertici di T_A. Diremo che un vertice è *monocromatico* se ha lo stesso colore di tutti i vertici ad esso adiacenti. Sia $\text{mono}(A)$ il numero di vertici monocromatici rossi.

Teorema 4.3. *Sia $\{A, B\}$ una partizione di $[n]$, sia T un albero filogenetico binario per il modello generale di Markov con n foglie su un alfabeto di q lettere. Allora*

$$\text{rk}(\text{Flat}_{A,B}(p)) = \min\{q^{e(A,B)+1-mono(A)}, q^{e(A,B)+1-mono(B)}, q^{|A|}, q^{|B|}\}. \tag{4.1}$$

Se $\{A, B\}$ è uno split, $Flat_{A,B}(p)$ è il flattening di p relativo ad un arco, ed infatti dal teorema ritroviamo un risultato già noto

Corollario 4.4. *Se $\{A, B\}$ è uno split di T allora* $\text{rk}\, Flat_{A,B}(P) \leq q$.

Il seguente risultato ci sarà utile per creare un algoritmo di ricostruzione degli alberi:

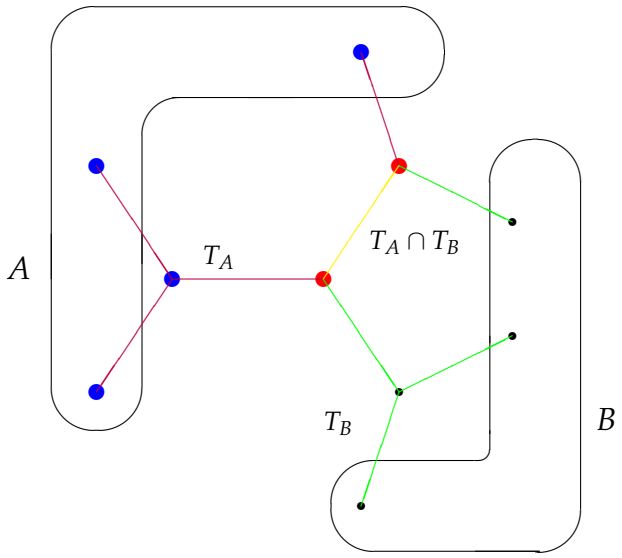


Figura 4.1: Vertici rossi e blu.

Corollario 4.5. *Se* $\{A, B\}$ *non è uno split nell'albero e* $|A|, |B| \geq 2$ *allora* $\mathrm{rk}\, Flat_{A,B}(p) \geq q^2$.

Poiché le matrici che abbiamo a disposizione sono costruite a partire da dati reali in generale non potremo fare affidamento dal calcolo del rango per creare un algoritmo affidabile per la costruzione degli alberi filogenetici a partire dai risultati finora enunciati. È necessario quindi trovare un modo per determinare quando una matrice è "più vicina ad essere di rango k" di un'altra. Per fare ciò useremo la decomposizione ai valori singolari, o *SVD* (singular value decomposition).

Definizione. Una *decomposizione ai valori singolari* di una matrice $m \times n$ A ($m \geq n$) è una fattorizzazione $A = U\Sigma V^t$ dove U è una matrice $m \times n$ che soddisfa $U^tU = Id_n$, V è una matrice $n \times n$ che soddisfa $V^tV = Id_n$ e $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ con $\sigma_1 \geq \sigma_2 \cdots \geq \sigma_n \geq 0$, detti *valori singolari* di A.

Definiamo prima di tutto due norme matriciali.

Definizione. Sia $A = (a_{ij})$ una matrice $m \times n$. La *norma di Frobenius*

di A è

$$\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2},$$

mentre la *norma* L_2 di A è

$$\|A\|_2 = \max_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \left\{ \frac{\|Ax\|}{\|x\|} \right\}, \tag{4.2}$$

dove $\|x\| = \sqrt{\sum_{i=1}^n x_i^2}$.

Grazie al seguente teorema (lo si ritrova ad esempio in [23]) possiamo trovare un modo per stabilire quanto una matrice è lontana dall'avere un rango fissato:

Teorema 4.6. *La distanza di una matrice A dalla più vicina matrice di rango k è*

$$\min_{\mathrm{rk}(B) \leq k} \|A - B\|_F = \sqrt{\sum_{i=k+1}^{n} \sigma_i^2} \tag{4.3}$$

nella norma di Frobenius e

$$\min_{\mathrm{rk}(B) \leq k} \|A - B\|_2 = \sigma_{k+1} \tag{4.4}$$

nella norma L_2.

Questo ci permette di costruire l'algoritmo:

Algoritmo 2. *Costruzione degli alberi tramite SVD*

Input: Un allineamento di dati genomici da n specie da un alfabeto Σ con q stati.

Output: Un albero binario con n foglie.

Passo 1: Calcolare le probabilità empiriche $p_{i_1 \dots i_n}$ e scriverle in un tensore P.

Passo 2: Per $k = n \searrow 4$ effettuare le seguenti operazioni:

- Per ognuna delle $\binom{k}{2}$ coppie di specie (i, j) scrivere $Flat_{\{i,j\},[n]\setminus\{i,j\}}(P)$ e trovarne la decomposizione SVD.
- Scegliere la coppia rispetto alla quale il $\sqrt{\sum_{i \geq k+1} \sigma_i^2}$ sia minimo e unirla ad un unico vertice dell'albero. Considerare nei successivi passi queste due variabili come un'unica a valori in $[m_i] \times [m_j]$ con m_i e m_j stati rispettivamente di X_i e X_j.

Grazie a questo algoritmo possiamo ricavare un albero filogenetico senza dover necessariamente confrontare tutti i possibili alberi con n foglie.

Esempio di ricostruzione

Usando MATLAB simuliamo un allineamento di lunghezza $l = 1000$ proveniente da un modello distribuito secondo l'albero con 6 foglie in figura 4.2. Alle foglie $\{a_1, \dots, a_6\}$ sono associate le variabili $X_1, \dots X_6$ a valori in $\{A, C, G, T\}$.

Dopo aver calcolato $\hat{p}$, per ogni coppia di variabili X_u, X_v riportiamo nella tabella 4.2 il valore di $t_{ij} = \sqrt{\sum_{i=5}^{16} \sigma_i^2}$, dove i σ_i sono i valori singolari di $Flat_{\{u,v\}}(\hat{p})$.
Il minimo di questa tabella corrisponde alla partizione $\{5, 6\}, \{1, 2, 3, 4\}$: uniamo quindi le foglie a_5 e a_6 in un'unico vertice w_1 nell'albero.
Ripetiamo questo procedimento considerando le variabili X_1, X_2, X_3, X_4, X_7, quest'ultima associata al vertice w_1. Notiamo che, a differenza delle altre, X_7 può assumere 16 stati. La tabella 4.3 riporta i risultati ottenuti. Poiché il minimo è in corrispondenza della partizione

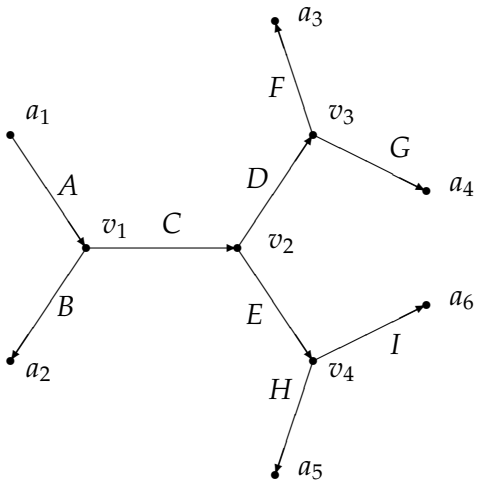


Figura 4.2: Albero filogenetico binario con 6 foglie usato nell'esempio.

	2	3	4	5	6
1	17.5968	17.8593	18.5077	19.0437	18.8461
	2	18.5405	18.6460	18.1039	17.9513
		3	17.7858	19.2187	19.2316
			4	19.1195	19.4773
				5	16.9615

Tabella 4.2: Distanze fra $\text{Flat}_{ij}(\hat{p})$ e la più vicina matrice di rango 4.

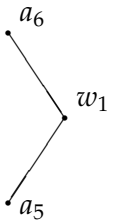


Figura 4.3: Situazione dopo la prima iterazione del ciclo.

$\{1,2\},\{3,4,7\}$, uniamo le foglie a_1 e a_2 in un unico vertice w_2 nell'albero.

Sono rimaste soltanto 4 variabili libere X_3, X_4, X_7, X_8, quest'ultima corrispondente al vertice w_2, quindi per completare l'albero è suf-

	2	3	4	7
1	17.5968	17.8593	18.5077	27.3605
2		18.5405	18.6460	24.8974
3			17.7858	27.7195
4				28.1875

Tabella 4.3: Distanze fra $\text{Flat}_{ij}(\hat{p})$ e la più vicina matrice di rango 4 dopo la prima iterazione.

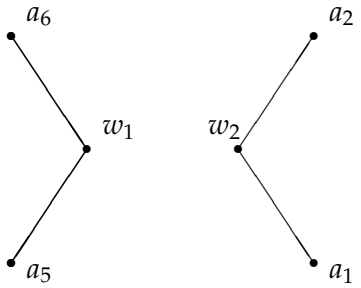


Figura 4.4: Situazione dopo la seconda iterazione del ciclo.

ficiente ripetere l'operazione una sola volta ancora. In tabella 4.4 sono riportati i risultati.

	4	7	8
3	17.7858	27.7195	28.1875
4		28.1875	27.7195
7			22.3822

Tabella 4.4: Distanze fra $\text{Flat}_{ij}(\hat{p})$ e la più vicina matrice di rango 4 dopo la terza iterazione.

Il minimo è ottenuto in corrispondenza della coppia $\{a_3, a_4\}$: uniamo queste foglie in un vertice interno w_3. Ora che rimangono solo 3 vertici liberi concludiamo la nostra ricostruzione unendoli in un unico vertice interno trivalente.

L'algoritmo ha quindi ricostruito la corretta topologia dell'albero, ovvero quella rispetto alla quale avevamo generato i dati: infatti, i due alberi hanno come split $\{1,2\}, \{3,4\}, \{5,6\}$ e perciò per il teorema di Buneman hanno la stessa topologia.

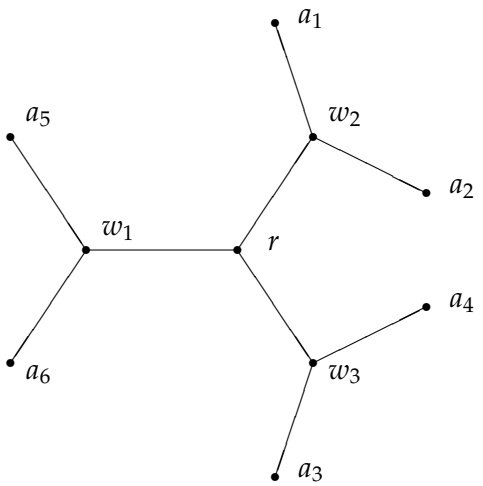


Figura 4.5: Albero filogenetico binario ricostruito con il metodo SVD.

5 Simulazioni

Abbiamo studiato infine il comportamento degli algoritmi 1 e 2 sui dati per verificarne l'affidabilità . I listati dei programmi MATLAB usati per le simulazioni si trovano in appendice A.

5.1 Confronto invarianti e SVD su alberi con variabili binarie

Per prima cosa abbiamo controllato la correttezza dei due metodi: abbiamo generato casualmente 10000 set di parametri per l'albero in figura 5.1 con variabili binarie, costruendo poi i relativi tensore di probabilità p. Abbiamo applicato gli algoritmi di ricostruzione a questi dati *esatti*: entrambi hanno ricostruito l'albero corretto nel 100% dei casi.

Abbiamo poi simulato con MATLAB dati da 1000 alberi con la topologia illustrata in figura 5.1, con 5 variabili binarie, ripetendo il processo 3 volte, con 3 diversi set di parametri. Per ogni simulazione abbiamo poi eseguito entrambi gli algoritmi di ricostruzione sui dati ottenuti. Questa procedura è stata poi ripetuta diverse volte varian-

do la lunghezza delle sequenze generate.

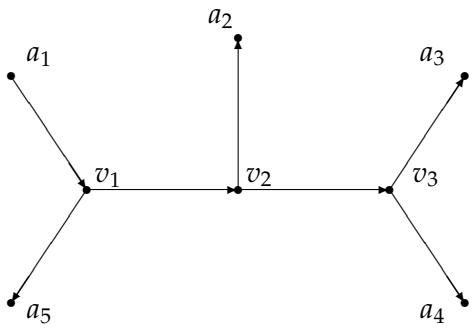


Figura 5.1: Albero filogenetico usato nella simulazione.

Parametri stocastici

Le matrici di transizione usate in questa simulazione sono state costruite generando matrici stocastiche M con probabilità uniforme. I risultati sono riassunti nella tabella 5.1 ed in figura 5.2:

	lunghezza sequenze					
	100	250	500	750	1000	1500
invarianti	5.9%	8.9%	14.8%	17%	22.8%	25.6%
SVD	5.2%	6%	8.1%	9.6%	11.8%	14%
	2000	3000	4000	5000	6000	7000
invarianti	33.4%	45.5%	50.3%	54.1%	58.1%	66.1%
SVD	16.9%	20.5%	21.6%	22.5%	26.3%	28.5%

Tabella 5.1: Percentuali di corretta ricostruzione degli alberi filogenetici nella prima simulazione.

Parametri stocastici a predominanza diagonale

Le matrici di transizione usate in questa simulazione sono state costruite generando matrici stocastiche M con probabilità uniforme ed utilizzando nel programma

$$\tilde{M} = \frac{1}{2}M + \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

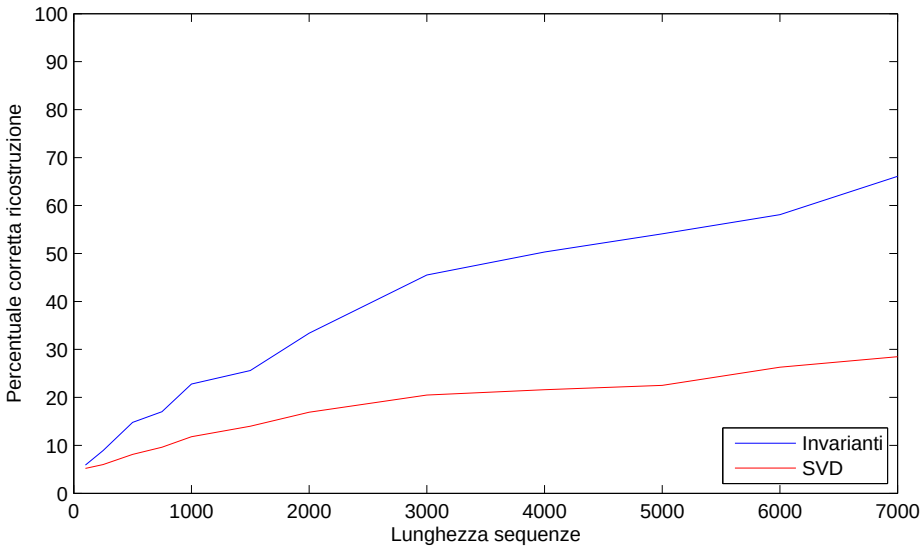


Figura 5.2: Confronto fra i due metodi con il primo set di parametri.

così da avere delle matrici in cui il massimo di ogni riga si trovi sulla diagonale.
I risultati sono riassunti nella tabella 5.2 ed in figura 5.3:

	lunghezza sequenze					
	100	250	500	750	1000	1500
invarianti	14.9%	17.9%	18.4%	18.7%	19.6%	23.2%
SVD	23.4%	31%	33.5%	34%	39.2%	41.7%

	2000	3000	4000	5000	6000	7000
invarianti	25%	31.7%	38.8%	44.3%	47.4%	55.1%
SVD	47.1%	51.9%	59.1%	67.8%	71.6%	80.1%

Tabella 5.2: Confronto fra i due metodi con il secondo set di parametri.

Parametri stocastici a forte predominanza diagonale

Le matrici di transizione usate in questa simulazione sono state costruite generando matrici stocastiche M con probabilità uniforme ed

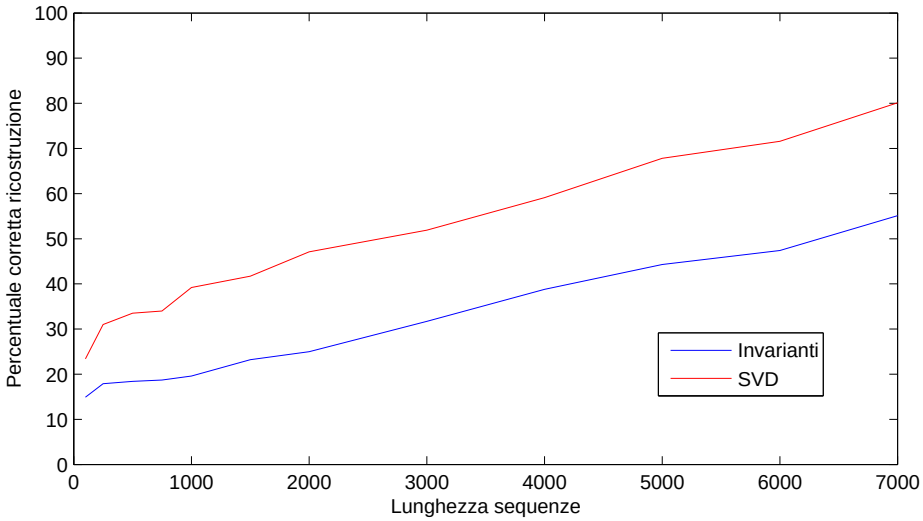


Figura 5.3: Confronto fra i due metodi con il secondo set di parametri.

utilizzando nel programma

$$\tilde{M} = \frac{1}{10}M + \frac{9}{10}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

in modo da avere matrici in cui il massimo di ogni riga si trovi sulla diagonale e sia almeno 9 volte maggiore della somma degli altri elementi.I risultati sono riassunti nella tabella 5.3 ed in figura 5.4:

	lunghezza sequenze					
	100	250	500	750	1000	1500
invarianti	74.8%	88.1%	97.5%	99.5%	100%	100%
SVD	71%	90%	98.4%	99.6%	99.8%	100%

	2000	3000	4000	5000	6000	7000
invarianti	100%	100%	100%	100%	100%	100%
SVD	100%	100%	100%	100%	100%	100%

Tabella 5.3: Confronto fra i due metodi con il terzo set di parametri.

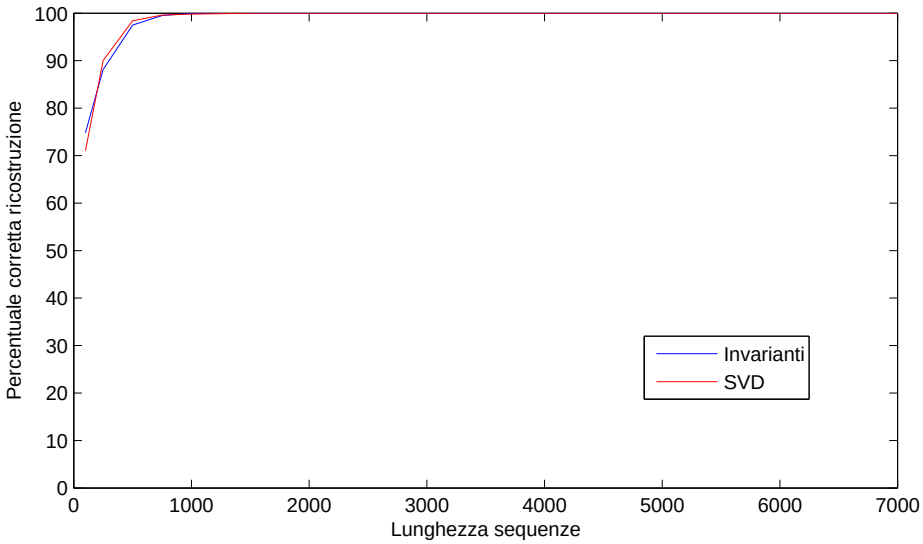


Figura 5.4: Confronto fra i due metodi con il terzo set di parametri.

Analisi delle prestazioni

Notiamo innanzi tutto che l'accuratezza dei due metodi cresce con la lunghezza delle sequenze, come era logico attendersi, visto che $\lim_{l\to\infty} \hat{p} = p$.
Meno ovvia è la pesante dipendenza dalla struttura dei parametri di entrambi i metodi. Il metodo degli invarianti si rivela più stabile, ma la performance del metodo SVD cambia drasticamente a seconda del peso degli elementi diagonali delle matrici di transizione.
Non è chiaro quale possa essere la causa di questo comportamento, ma vale la pena notare che quest'ultimo algoritmo dipende da un numero maggiore di scelte arbitrarie, nelle quali, come si nota ad esempio dalla tabella 4.2, la differenza fra un accoppiamento giusto e uno sbagliato può essere di uno o più ordini di grandezza inferiore ai dati analizzati.
Anche il rendimento del metodo degli invarianti, nonostante la maggiore stabilità, è migliorabile: in [10] è presente un primo studio sul comportamento dei singoli invarianti (nel caso del modello di Jukes-Cantor) in cui si evidenzia come esista una grande eterogeneità tra di essi. Ci sono infatti invarianti "buoni", che permettono di identificare la giusta topologia spesso (i.e. sono piccoli per l'albero giusto), e altri

meno buoni che addirittura sono più piccoli se l'albero non è quello corretto.
Evidentemente aver considerato in blocco l'intero insieme degli invarianti senza una norma che pesasse di più quelli migliori ha peggiorato la performance dell'algoritmo. Come possiamo vedere però nell'articolo di Eriksson, ci sono notevoli margini di miglioramento.
Fortunatamente, inoltre, nella maggior parte dei casi si può supporre che i dati provenienti da sequenze estratte da specie esistenti siano generati da matrici simili a quelle utilizzate nella terza simulazione, per le quali entrambi i metodo hanno avuto un ottimo rendimento. Generalmente, infatti, la probabilità di una mutazione, specialmente se consideriamo specie simili, è piuttosto bassa. Questo fa ben sperare riguardo all'utilizzo di entrambi i metodi su dati biologicamente significativi.

5.2 Utilizzo del metodo SVD su dati reali.

Il dipartimento di biologia evoluzionistica dell'università di Firenze ci ha gentilmente fornito un allineamento di 15 sequenze di DNA da analizzare con i nostri metodi. Si tratta di sequenze di lunghezza 1350 di 16s rRNA, estratte da alcuni ceppi di Escherichia coli .
Vista la dimensione dei dati, abbiamo potuto utilizzare soltanto il metodo SVD, che ha dato come risultato l'albero in figura 5.5.

Utilizzando il metodo di massima verosimiglianza messo a disposizione da MEGA abbiamo ottenuto dagli stessi dati l'albero in figura 5.6. I due alberi sono diversi, ma presentano comunque delle somiglianze. Notiamo ad esempio che i 3 sottoalberi principali del primo albero corrispondono alla tripartizione $\{3,4,5,6,11,12,13,14,15\}, \{2,9,10\}, \{1,7,8\}$, mentre i 3 sottoalberi principali del secondo corrispondono alla tripartizione $\{4,5,11,13,14,15\}, \{2,3,9,10,12\}, \{1,6,7,8\}$, dunque solamente 4 foglie su 15 risultano essere in sottoalberi diversi. Possiamo dunque essere moderatamente soddisfatti del funzionamento di questo algoritmo.

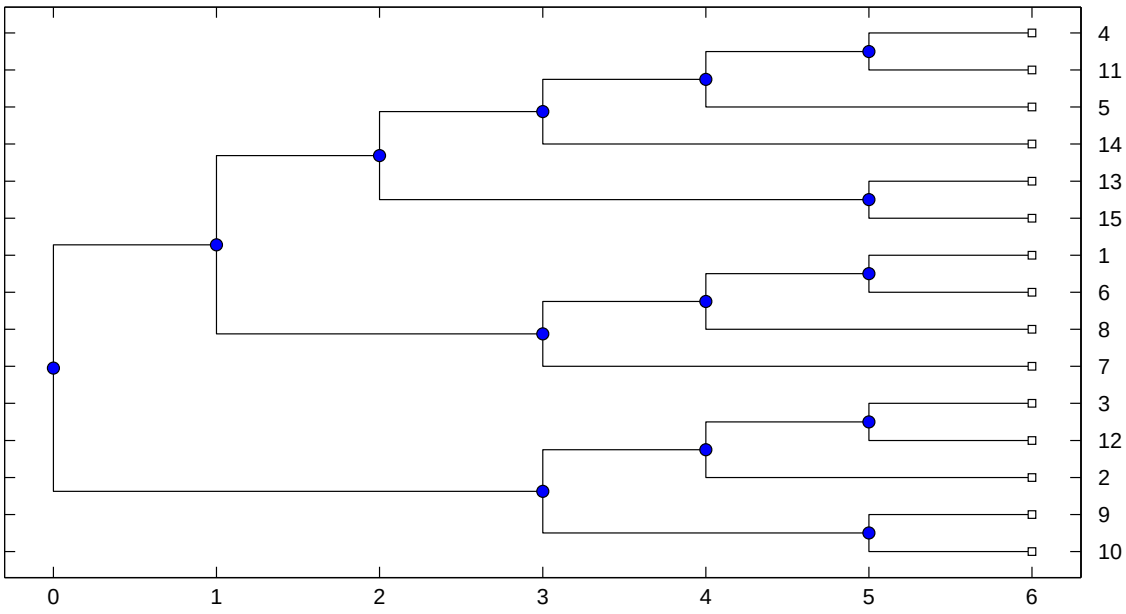


Figura 5.5: Albero con 15 foglie ricostruito con il metodo SVD.

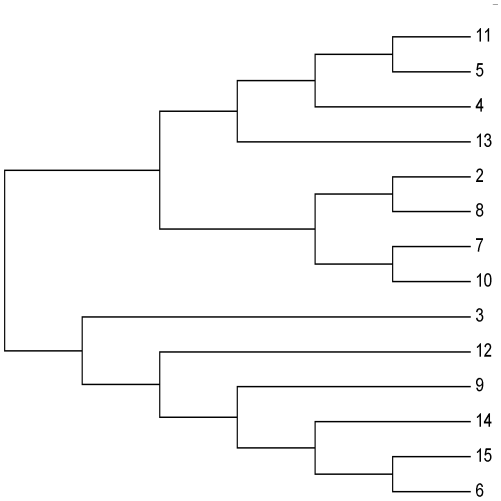


Figura 5.6: Albero con 15 foglie ricostruito da MEGA
con il metodo di massima verosimiglianza.

A Codici MATLAB

A.1 Simulazione Albero 5 foglie

```
function x=discprob(v)
%Simula il comportamento di una variabile aleatoria
      discreta con
%distribuzione v a valori in {0,1, ... , n-1}
if  prod(double(abs(v)==v))==0 || norm(v)==0
    disp('v deve essere non nullo a valori positivi')
return
end
x=0;
y=sum(v);
v=v/y;
w=cumsum(v);
u=rand;
for i=1:size(w,1)
    if u < w(i)
        x=i-1;
    break
    end
end

function A=randstoc(n,m)
%genera una matrice stocastica
A=abs(randn(n,m));
c=sum(A);
for i=1:m
    A(:,i)=A(:,i)/c(i);
end

function [q,X]=sim4bin1(A,B,C,D,E,p,n)
%Simula dati binari provenienti da una distribuzione
      definita dall'albero
%con 4 foglie ((1,2),(3,4))
X=zeros(4,n);
t=zeros(2,2,2,2);
for j=1:n
    r=discprob(p);
    X(1,j)=r;
    switch r
        case 0
            f=discprob(A(1,:)');
```

```
            switch f
                case 0
                    X(2,j)=discprob(B(1,:)');
                    g=discprob(C(1,:)');
                    X(3,j)=discprob(D(g+1,:)');
                    X(4,j)=discprob(E(g+1,:)');
                case 1
                    X(2,j)=discprob(B(2,:)');
                    g=discprob(C(2,:)');
                    X(3,j)=discprob(D(g+1,:)');
                    X(4,j)=discprob(E(g+1,:)');
            end
        case 1
            f=discprob(A(2,:)');
            switch f
                case 0
                    X(2,j)=discprob(B(1,:)');
                    g=discprob(C(1,:)');
                    X(3,j)=discprob(D(g+1,:)');
                    X(4,j)=discprob(E(g+1,:)');
                case 1
                    X(2,j)=discprob(B(2,:)');
                    g=discprob(C(2,:)');
                    X(3,j)=discprob(D(g+1,:)');
                    X(4,j)=discprob(E(g+1,:)');
            end
    end
    t(X(1,j)'+1,X(2,j)'+1,X(3,j)'+1,X(4,j)'+1)=t(X(1,j)
        '+1,X(2,j)'+1,X(3,j)'+1,X(4,j)'+1)+1;
end
q=t/n;

function [q,X]=sim5bin11(A,B,C,D,E,F,G,p,n)
%Simula dati binari provenienti da una distribuzione
    definita dall'albero
%con 5 foglie (((1,5),2),(3,4))
X=zeros(5,n);
t=zeros(2,2,2,2,2);
[~,Y]=sim4bin1(A,B,C,D,E,p,n);
X([2 3 4],:)=Y([2 3 4],:);
for j=1:n
    X(1,j)=discprob(F(Y(1,j)+1,:)');
    X(5,j)=discprob(G(Y(1,j)+1,:)');
    t(X(1,j)+1,X(2,j)+1,X(3,j)+1,X(4,j)+1,X(5,j)+1)=t(X(1,
        j)+1,X(2,j)+1,X(3,j)+1,X(4,j)+1,X(5,j)+1)+1;
```

```
end
q=t/n;
```

A.2 Programmi algoritmo SVD variabili binarie

```
function [i,j]=Phylflatbin(P)
%Trova la coppia di indici di un tensore P corrispondente
      al flattening che
%pi si avvicina ad avere rango 2
n=size(size(P),2);
%n=numero di specie, ovvero la dimensione di P
d=zeros(n);
m=1:n;
for i=1:n-1
    for j=i+1:n;
        if size(P,i)~=1 && size(P,j)~=1
        T=permute(P,[[i,j],removerows(m',[i,j])']);
        F=reshape(T,(size(P,i)*size(P,j)), (numel(P))/((
              size(P,i)*size(P,j))));
        [~,S]=svd(F);
        s=diag(S);
        d(i,j)=norm(s(3:end));
        else
            continue
        end
    end
end
d(~d) = nan;
[~, ind]=min(d(:));
[i,j]=ind2sub(size(d),ind);

function M=binsvd(P)
%Algoritmo di costruzione di alberi filogenetici tramite
      SVD
%Input: P array n-dimensionale delle frequenze delle basi
%Output: Matrice M per disgnare l'albero con il comando
      phytree
%VARIABILI BINARIE
n=size(size(P),2);
%n: numero di specie in esame
M=[];
m=1:n;
d=2*ones(1,n);
```

```
%d: vettore contenente le dimensioni di P.
c=0;
%contatore iterazioni ciclo
for k=n:-1:4
    c=c+1;
    [i,j]=Phylflatbin(P);
    % Determiniamo la coppia rispetto alla quale il
        flattening ha rango pi
    % vicino a 4
    mi=size(P,i);
    mj=size(P,j);
    M=[M;[i,j]];
    %Inseriamo la coppia nella matrice M, cosicch nell'
        albero finale le
    %due foglie saranno unite da un arco
    %Varys uccide Kevan e Pycelle
    d(i)=1;
    d(j)=1;
    d=[d,mi*mj];
    %nuove dimensioni dell'array
    P=permute(P,[removerows(m',[i,j])',[i,j]]);
    P=reshape(P,d);
    %nuova forma dell'array
    m=[m,n+c];
    %nuovi indici dell'array
end
q=(removerows((1:n+c-1)',M(:)))';
M=[M;q;n+c n+c+1];
%Prendiamo le foglie non ancora selezionate e le uniamo al
    resto
%dell'albero
tree=phytree(M);
view(tree)
```

A.3 Programmi algoritmo SVD variabili ACGT

```
function x=code(a)
%trasforma un carattere A,C,G,T in 1,2,3,4
switch a
    case 'A'
        x=1;
    case 'C'
```

```
        x=2;
    case 'G'
        x=3;
    case 'T'
        x=4;
end

function x=conv(v)
%converte un vettore di cifre al numero intero da esse
      costituito
n=size(v,2);
m=n-1:-1:0;
y=(10*ones(1,n)).^m;
x=v*y';

function [Z,T]=trad(X)
%dato un allineamento X restituisce un vettore Z con i
      codici dei pattern
%di X e un vettore T con le frequenze relative ai codici
      in Z
if iscell(X)==0
X=num2cell(X);
end
Z=[];
T=[];
[n,m]=size(X);
Y=zeros(n,m);
for j=1:m
    for i=1:n
        Y(i,j)=code(X{i,j});
    end
    w=conv(Y(:,j)');
    [p k]=quick(Z,w);
    if k==1
        T(p)=T(p)+1;
    else
        Z=shift(Z,p);
        Z(p)=w;
        T=shift(T,p);
        T(p)=1;
    end
end

function w=insert(v,M)
%Inserisce un elemento M in un vettore ORDINATO v (se non
      gi presente)
```

```
%mantenendo l'ordine
[p i]=quick(v,M);
if i==1
    w=v;
else
    w=shift(v,p);
    w(p)=M;
end

function [i,j]=smflat2(Z,T,G,L)
%Trova la coppia di indici corrispondente al flattening
      che
%pi si avvicina ad avere rango 4, considerando la matrice
      dei dati invece
%dell'intero tensore.
%L  la lista delle variabili libere, G  la matrice in cui
      sono salvati
%gli split precedenti.
%n=numero di specie, ovvero la dimensione di P
n=ceil(log10(Z(1,1)+1));
d=zeros(n);
l=size(L,2);
for i=1:l-1
    for j=i+1:l
        F=smfrequenze2(Z,T,L(i),L(j),G);
        S=svd(F);
        d(L(i),L(j))=norm(S(5:end));
    end
end
d(~d) = nan;
[~, ind]=min(d(:));
[i,j]=ind2sub(size(d),ind);

function M=smsvd2(Y)
%Ricostruisce l'albero filogenetico usando il metodo SVD a
      partire da un
%allineamento Y.
M=[];
G=[];
n=size(Y,1);
c=0;
[Z,T]=trad(Y);
L=1:n;
for k=n:-1:4
    c=c+1;
```

```
    [i,j]=smflat2(Z,T,G,L);
    M=[M;[i,j]];
    L=removerows(L',[quick(L,i),quick(L,j)])';
    L=[L,n+c];
    G=[M,(n+1:n+size(M,1))'];
end
q=(removerows((1:n+c-1)',M(:)))';
M=[M;q;n+c n+c+1];
tree=phytree(M);
view(tree)
```

A.4 Programmi algoritmo invarianti 5 variabili binarie

```
function [m,R]=inv5bin(P)
%dato un tensore p 2x2x2x2x2 calcola gli invarianti per
    ogni possibile
%albero con 5 foglie. Nel vettore R sono salvati i
    punteggi di ogni albero,
%m  l'indice corrispondente all'albero scelto con il
    metodo degli
%invarianti.
R=zeros(1,15);
p12=P;
P12=reshape(p12,4,8);
d12=0;
p13=permute(P,[1 3 2 4 5]);
P13=reshape(p13,4,8);
d13=0;
p14=permute(P,[1 4 2 3 5]);
P14=reshape(p14,4,8);
d14=0;
p15=permute(P,[1 5 3 4 2]);
P15=reshape(p15,4,8);
d15=0;
p23=permute(P,[2 3 1 4 5]);
P23=reshape(p23,4,8);
d23=0;
p24=permute(P,[2 4 1 3 5]);
P24=reshape(p24,4,8);
d24=0;
p25=permute(P,[2 5 1 3 4]);
P25=reshape(p25,4,8);
```

```
d25=0;
p34=permute(P,[3 4 1 2 5]);
P34=reshape(p34,4,8);
d34=0;
p35=permute(P,[3 5 1 2 4]);
P35=reshape(p35,4,8);
d35=0;
p45=permute(P,[4 5 1 2 3]);
P45=reshape(p45,4,8);
d45=0;

for i=1:2
    for j=i+1:3
        for k=j+1:4
            for l=1:6
               for m=l+1:7
                    for n=m+1:8
                        d12=d12+abs(det(P12([i j k],[l m n
                            ])));
                        d13=d13+abs(det(P13([i j k],[l m n
                            ])));
                        d14=d14+abs(det(P14([i j k],[l m n
                            ])));
                        d15=d15+abs(det(P15([i j k],[l m n
                            ])));
                        d23=d23+abs(det(P23([i j k],[l m n
                            ])));
                        d24=d24+abs(det(P24([i j k],[l m n
                            ])));
                        d25=d25+abs(det(P25([i j k],[l m n
                            ])));
                        d34=d34+abs(det(P34([i j k],[l m n
                            ])));
                        d35=d35+abs(det(P35([i j k],[l m n
                            ])));
                        d45=d45+abs(det(P45([i j k],[l m n
                            ])));
                    end
                end
            end
        end
    end
end
R(1)=d34+d15;
```

```
R(2)=d34+d25;
R(3)=d12+d35;
R(4)=d12+d45;
R(5)=d15+d24;
R(6)=d13+d25;
R(7)=d24+d35;
R(8)=d13+d45;
R(9)=d15+d23;
R(10)=d14+d25;
R(11)=d14+d35;
R(12)=d23+d45;
R(13)=d12+d34;
R(14)=d13+d24;
R(15)=d14+d23;
[~,m]=min(R);
```

A.5 Codice riassuntivo simulazioni

```
%% Alberi filogenetici

%% Il  caso con 4 foglie e variabili ACGT

n=1000;
%  n  la lunghezza delle sequenze
[A,B,C,D,E,p]=randpar;
% parametri casuali ACGT
[q,Y]=simDNA4_1(A,B,C,D,E,p,n);
smsvd2(Y);
% ricostruzione tramite SVD
%% Il caso con 5 foglie e variabili binarie
n=1000;
% n  la lunghezza delle frequenze
[A,B,C,D,E,F,G,p]=randpar5bin;
[q,Y]=sim5bin11(A,B,C,D,E,F,G,p,n);
% Simulazione per l'albero ((1,5),2,(3,4))
binsvd(q);
% Albero ricostruito con SVD
inv5bin(q)
% Codice albero ricostruito con metodo degli invarianti: 1
        il codice
% dell'albero secondo cui sono stati simulati i dati
```

B Dati

B.1 Parametri delle simulazioni: esempio albero con 5 foglie

Parametri stocastici

$$A = \begin{pmatrix} 0.9039 & 0.0961 \\ 0.8465 & 0.1535 \end{pmatrix} \quad B = \begin{pmatrix} 0.2463 & 0.7537 \\ 0.4678 & 0.5322 \end{pmatrix} \quad C = \begin{pmatrix} 0.7271 & 0.2729 \\ 0.5381 & 0.4519 \end{pmatrix}$$

$$D = \begin{pmatrix} 0.5836 & 0.4164 \\ 0.4823 & 0.6177 \end{pmatrix} \quad E = \begin{pmatrix} 0.3165 & 0.6835 \\ 0.7367 & 0.2633 \end{pmatrix} \quad F = \begin{pmatrix} 0.8975 & 0.1025 \\ 0.3694 & 0.6306 \end{pmatrix}$$

$$G = \begin{pmatrix} 0.7160 & 0.2849 \\ 0.6549 & 0.3451 \end{pmatrix} \quad \pi_r = \begin{pmatrix} 0.2899 \\ 0.7101 \end{pmatrix}$$

Parametri stocastici a predominanza diagonale

$$A = \begin{pmatrix} 0.7182 & 0.2818 \\ 0.2845 & 0.7155 \end{pmatrix} \quad B = \begin{pmatrix} 0.8359 & 0.1641 \\ 0.1505 & 0.8495 \end{pmatrix} \quad C = \begin{pmatrix} 0.7223 & 0.2777 \\ 0.1487 & 0.8513 \end{pmatrix}$$

$$D = \begin{pmatrix} 0.7525 & 0.2475 \\ 0.4832 & 0.5168 \end{pmatrix} \quad E = \begin{pmatrix} 0.6040 & 0.3960 \\ 0.2482 & 0.7518 \end{pmatrix} \quad F = \begin{pmatrix} 0.9589 & 0.0411 \\ 0.2608 & 0.7392 \end{pmatrix}$$

$$G = \begin{pmatrix} 0.5022 & 0.4978 \\ 0.3373 & 0.6627 \end{pmatrix} \quad \pi_r = \begin{pmatrix} 0.1680 \\ 0.8320 \end{pmatrix}$$

Parametri stocastici a forte predominanza diagonale

$$A = \begin{pmatrix} 0.9849 & 0.0151 \\ 0.0110 & 0.9890 \end{pmatrix} \quad B = \begin{pmatrix} 0.9951 & 0.0049 \\ 0.0244 & 0.9756 \end{pmatrix} \quad C = \begin{pmatrix} 0.9200 & 0.0800 \\ 0.0179 & 0.9821 \end{pmatrix}$$

$$D = \begin{pmatrix} 0.9122 & 0.0878 \\ 0.0639 & 0.9361 \end{pmatrix} \quad E = \begin{pmatrix} 0.9400 & 0.0600 \\ 0.0069 & 0.9931 \end{pmatrix} \quad F = \begin{pmatrix} 0.9425 & 0.0575 \\ 0.0254 & 0.9746 \end{pmatrix}$$

$$G = \begin{pmatrix} 0.9137 & 0.0863 \\ 0.0401 & 0.9599 \end{pmatrix} \quad \pi_r = \begin{pmatrix} 0.3794 \\ 0.6206 \end{pmatrix}$$

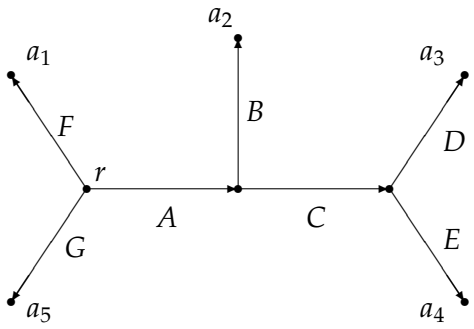


Figura B.1: Albero filogenetico usato nella simulazione.

B.2 Simulazioni esempio 6 foglie

$$A = \begin{pmatrix} 0.5489 & 0.1669 & 0.2056 & 0.0785 \\ 0.0663 & 0.7721 & 0.0902 & 0.0713 \\ 0.1667 & 0.1290 & 0.5629 & 0.1414 \\ 0.2123 & 0.0185 & 0.2092 & 0.5600 \end{pmatrix} \qquad B = \begin{pmatrix} 0.5140 & 0.1678 & 0.1587 & 0.1596 \\ 0.0794 & 0.6428 & 0.0849 & 0.1929 \\ 0.0957 & 0.2026 & 0.6423 & 0.6840 \\ 0.0471 & 0.1263 & 0.1425 & 0.6840 \end{pmatrix}$$

$$C = \begin{pmatrix} 0.5854 & 0.0646 & 0.2351 & 0.1149 \\ 0.0391 & 0.5907 & 0.1646 & 0.2056 \\ 0.0524 & 0.1237 & 0.6636 & 0.1603 \\ 0.2563 & 0.0089 & 0.0489 & 0.6860 \end{pmatrix} \qquad D = \begin{pmatrix} 0.6739 & 0.1764 & 0.1374 & 0.0123 \\ 0.1570 & 0.6440 & 0.0009 & 0.1982 \\ 0.1549 & 0.0748 & 0.5454 & 0.2249 \\ 0.1962 & 0.0059 & 0.0996 & 0.6983 \end{pmatrix}$$

$$E = \begin{pmatrix} 0.6998 & 0.0111 & 0.1930 & 0.0961 \\ 0.1111 & 0.7461 & 0.0645 & 0.0783 \\ 0.0296 & 0.1368 & 0.6177 & 0.2159 \\ 0.3114 & 0.1069 & 0.0388 & 0.5429 \end{pmatrix} \qquad F = \begin{pmatrix} 0.7030 & 0.0417 & 0.0283 & 0.2270 \\ 0.1559 & 0.6350 & 0.1618 & 0.0472 \\ 0.0409 & 0.2213 & 0.7179 & 0.0199 \\ 0.0868 & 0.3106 & 0.0801 & 0.5225 \end{pmatrix}$$

$$G = \begin{pmatrix} 0.5097 & 0.2275 & 0.0517 & 0.2112 \\ 0.1771 & 0.6871 & 0.0211 & 0.1147 \\ 0.0711 & 0.1407 & 0.6148 & 0.1733 \\ 0.1753 & 0.0199 & 0.2189 & 0.5860 \end{pmatrix} \qquad H = \begin{pmatrix} 0.6929 & 0.1527 & 0.1368 & 0.0177 \\ 0.1496 & 0.7041 & 0.0206 & 0.1257 \\ 0.2356 & 0.0669 & 0.6117 & 0.0857 \\ 0.1107 & 0.0643 & 0.2594 & 0.5656 \end{pmatrix}$$

$$I = \begin{pmatrix} 0.5892 & 0.2610 & 0.0450 & 0.1048 \\ 0.3285 & 0.6058 & 0.0587 & 0.0070 \\ 0.2144 & 0.1812 & 0.5563 & 0.0481 \\ 0.0049 & 0.0563 & 0.3771 & 0.5615 \end{pmatrix} \qquad \pi_r = \begin{pmatrix} 0.0789 \\ 0.1112 \\ 0.2882 \\ 0.5218 \end{pmatrix}$$

Riferimenti bibliografici

[1] H. Abo, G. Ottaviani, C. Peterson, *Induction for secant varieties of Segre varieties*, Trans. Amer. Math. Soc. 361 (2009), no. 2, 767-792

[2] E. S. Allman, J. A. Rhodes, *Phylogenetic ideals and varieties for the general Markov model of sequence mutation*, Math. Biosci. **186** (2003), 113-144.

[3] E. S. Allman, J. A. Rhodes, *Phylogenetic ideals and varieties for the general Markov model*, Adv. in Appl. Math. **40** (2008), no. 2, 127-148.

[4] M. Casanellas, J. Fernandez-Sanchez, *Relevant phylogenetic invariants of evolutionary models*, ArXiv e-prints 2009.

[5] J. Cavender, J. Felsenstein, *Invariants of phylogenies in a simple case with discrete states*, Journal of Classification, 4:57-71, 1987.

[6] L. Chiantini, *Introduzione alla Statistica Algebrica*, `http://www.smfn.unisi.it/smfn_lauree/view_matdid.php?id=1419`

[7] J. Draisma, J. Kuttler, *On the ideals of equivariant tree models*, Mathematische Annalen, **344**, 619-644, 2009.

[8] M. Drton, B. Sturmfels, S. Sullivant, *Lectures on Algebraic Statistics*, Springer 2008

[9] N. Eriksson, *Tree construction using singular value decompsition*, in L. Pachter and B. Sturmfels (eds.), *Algebraic Statistics for Computational Biology*, capitolo 19, pagg. 347-358. Cambridge University Press, Cambridge, UK, 2005.

[10] N. Eriksson, *Using invariants for phylogenetic tree construction*, in M. Putinar and S. Sullivant (eds.), *Emerging Applications of Algebraic Geometry*, I.M.A. Volumes in Mathematics and its Applications, Springer 2009.

[11] S. Friedland, *On tensors of border rank l in $\mathbb{C}^{n+m+l}$*, arXiv:1003.1968.

[12] S. Friedland, E. Gross, *A proof of the set-theoretic version of the salmon conjecture*, arXiv:1104.1776

[13] L. D. Garcia, M. Stillman, B. Sturmfels, *Algebraic geometry of Bayesian networks*, J. Symbolic Comput. 39, no. 3-4, 331-355, 2005.

[14] J. Harris, *Algebraic Geometry, a first course*, Springer 1992

[15] R. Hartshorne, *Algebraic Geometry*, Springer 1977

[16] J.A. Lake, *A rate independent technique for analysis of nucleic acid sequences: Evolutionary Parsimony.*, Mol. Bio. Evol. 4(2): 167-191, 1987.

[17] J.M. Landsberg, L. Manivel, *On the Ideals of secant varieties of Segre varieties*, Found. comput. Math., 4(4):397-422, 2004.

[18] S. Lauritzen, *Graphical Models*, Oxford University Press, New York, 1996.

[19] L. Pachter and B. Sturmfels (eds.), *Algebraic Statistics for Computational Biology*, Cambridge University Press, Cambridge, UK, 2005.

[20] I. Shafarevich, *Basic Algebraic Geometry*, Springer-Verlag, 1977.

[21] V. Strassen, *Rank and optimal computation of generic tensors*, Linear Algebra Appl. 52/53, 645-685, 1983.

[22] S. Sullivant, *Statistical models are algebraic varieties*, `http://www.math.harvard.edu/~seths/lecture1.pdf`

[23] L. M. Trefethen, D. Bau, *Numerical Linear Algebra*, SIAM, 1997.