Introduzione alla statistica bayesiana

Marco Regis*[1]
Università degli studi di Torino e Istituto Nazionale di Fisica Nucleare
via P. Giuria 1, I–10125 Torino, Italy

January 17, 2020

[1*]Email: regis@to.infn.it

Contents

[1]

[1] NB: alcuni degli esempi mostrati di volta in volta a lezione non sono contenuti in queste note.
Come fonti esterne si può far riferimento a G. Cowan "Statistical Data Analysis" (Oxford Science Publications), R. M. Neal "Probabilistic Inference Using Markov Chain Monte Carlo Methods", e Baldi and Brunak "Bioinformatics" (MIT).

Chapter 1

Introduzione alla probabilità

Il concetto di probabilità è intimamente legato a quello di incertezza. La probabilità può infatti essere definita come la quantificazione del livello di casualità di un evento, dove viene detto casuale ciò che non è noto o non può essere predetto con livello assoluto di certezza. Le diverse interpretazioni della probabilità nascono fondamendalmente dal suo duplice significato: epistemico, ossia come incertezza relativa alla (limitata) conoscenza umana, ed empirico, ossia come incertezza intrinseca dei fenomeni (come per esempio nell'interpretazione della scuola di Copenaghen della meccanica quantistica).

Faremo principalmente riferimento ad una concezione epistemica della probabilità (propria dell'interpretazione bayesiana), ossia ad una probabilità figlia della nostra ignoranza e non della natura intrinseca dei sistemi (considerati deterministici).

In questo senso la probabilità entra nell'analisi di diversi sistemi come ad esempio: fenomeni con un numero elevato di gradi di libertà (descritti facendo uso di valori medi relativi al comportamento collettivo di poche variabili), fenomeni complessi in cui le molteplici interazioni in gioco non sono computabili (ad esempio, un moto browniano descritto attraverso processi stocastici), situazioni di caos deterministico in cui le condizioni iniziali non sono sotto controllo, o fenomeni non ripetibili.

1.1 Concetto di probabilità e possibili interpretazioni (frequentista e bayesiano)

Definizioni di probabilità:

- **Classica** (de Moivre, Laplace): la probabilità di un evento è il rapporto tra il numero di casi favorevoli e quelli possibili, supposto che tutti gli eventi siano equiprobabili, ossia $\mathcal{P}_A = N_A/N$ (dove N è il numero di casi possibili e N_A è il numero di casi favorevoli per l'evento A).

  La definizione classica di probabilità, basata su eventi discreti e di numero finito, è difficilmente estendibile al caso di variabili continue. Altro elemento debole è la condizione ideale di perfetta uniformità, dove tutti i possibili esiti (lo spazio degli eventi) sono noti in precedenza e tutti sono ugualmente probabili. Inoltre quest'ultima condizione viene imposta prima di aver definito la nozione di probabilità (dando luogo a una circolarità nella definizione). La definizione classica conduce a paradossi quando un fenomeno ammette un numero infinito di possibilità (paradosso di Bertrand).

- **Frequentista** (von Mises): la probabilità di un evento è il limite cui tende la frequenza relativa dell'evento, al tendere all'infinito del numero delle prove effettuate, ossia $\mathcal{P}_A = \lim_{N\to\infty} N_A(N)/N$. Tale definizione può essere applicata anche senza conoscere a priori lo spazio degli eventi e senza assumere la condizione di equiprobabilità degli eventi elementari. Si assume però che l'esperimento sia ripetibile più volte, idealmente infinite, sotto le stesse condizioni.

  Più concretamente: $\lim_{N\to\infty} \mathcal{P}(|N_A(N)/N - \mathcal{P}_A| > \epsilon) = 0$. Questa definizione però contiene nuovamente una circolarità legata al fatto che bisogna decidere quanto piccolo dev'essere ϵ (quanto buona dev'essere l'approssimazione), ossia, in termini pratici, quanto grande deve essere N. Inoltre non tutti gli esperimenti sono ripetibili o possono essere ripetuti nelle stesse condizioni, e la probabilità si applica esclusivamente a fenomeni che si presentano su larga scala, mentre non si può parlare di probabilità di eventi singoli (o mai verificatisi).

  Un importante avanzamento rispetto alla concezione classica in cui la probabilità è stabilita a priori, prima di guardare i dati, risiede nel fatto che nella concezione frequentista la probabilità è invece

ricavata a posteriori, dall'esame dei dati. Sono però entrambe probabilità oggettive e il punto di vista frequentista non è di una incertezza epistemica ma al contrario legato ad una visione empirica della probabilità. Quella frequentista è una probabilità diretta, ossia una probabilità che, a partire dalla conoscenza della legge di probabilità di una popolazione, si assegna a un campione tratto da essa (come verrà discusso meglio in seguito).

- **Bayesiana**: Nell'approccio bayesiano, la proababilità è una misura del grado di plausibilità di una proposizione. Questa definizione è applicabile a qualsiasi evento. Si può vedere come un mapping del livello di plausibilità in un numero reale nell'intervallo [0,1]. Una proposizione complessa può essere costruita a partire da proposizioni elementari attraverso gli operatori logici di congiunzione (AND), disgiunzione (OR) e negazione (NOT).

  La probabilità bayesiana è una probabilità inversa, ossia consiste nel risalire dalle frequenze osservate alla probabilità. Nell'approccio bayesiano si utilizzano considerazioni "personali" per assegnare la probabilità ad un dato evento prima di fare l'esperimento. La probabilità a priori è quindi legata al grado di credibilità dell'evento, stabilito in maniera soggettiva (nota: si usano credibilità e plausibilità come sinonimi). Il teorema di Bayes consente in seguito, alla luce delle frequenze osservate, di "aggiustare" la probabilità a priori, per arrivare alla probabilità a posteriori. Quindi, tramite tale approccio, si usa una stima del grado di credibilità di una data ipotesi prima dell'osservazione dei dati, al fine di associare un valore numerico al grado di credibilità di quella stessa ipotesi successivamente all'osservazione dei dati. Essendo basata su un informazione a priori, non è una probabilità assoluta, ma sempre condizionata (alla conoscenza pregressa).

  Nell'approccio frequentista si determina quanto volte l'osservazione cade in un certo intervallo, mentre in quello bayesiano si attribuisce direttamente una probabilità di verità all'intervallo.

  Bayesianamente, non c'è distinzione relativamente all'origine dell'incertezza, cioè tra incertezza "statistica" (dovuta alla precisione finita dello strumento di misura) e incertezza "sistematica" (legata a effetti deterministici solo parzialmente noti, per esempio di calibrazione). In effetti entrambi sono legati alla mancanza di informazione e il carattere "casuale" dell'incertezza statistica è semplicemente dovuto alla non-conoscenza delle esatte condizioni in cui si trova il sistema. D'altra parte, come abbiamo visto, da un punto di vista frequentista si considerano gli esperimenti come casuali, ma stabilire la casualità dell'esperimento (per sistemi deterministici) porta a una certa circolarità nella definizione di probabilità. Un altro modo per vedere la questione è che la statistica bayesiana considera solo i dati veramente osservati, mentre nell'approccio frequentista bisogna fare assunzioni sulla distribuzione di possibili dati non osservati.

  Si può mostrare che per casi in cui un risultato frequentista esiste e nel limite di un campionamento molto grande, i risultati bayesiano e frequentista coincidono. Mentre ovviamente ci sono casi in cui l'approccio frequentista non è possibile.

Altre possibili interpretazioni, varianti delle due principali scuole, frequentista e bayesiana, quali ad esempio l'interpretazione propensista (Popper), logicista (Keynes) o assiomatica (Kolmogorov), non verranno discusse nel corso in quanto raramente utilizzate nell'ambito della ricerca scientifica e un eventuale approfondimento è lasciato all'interesse dello studente.

1.2 Proprietà generali della probabilità

La teoria della probabilità è una disciplina matematica relativamente recente, formalizzata solo nel XX secolo. Si basa su tre assiomi (formulazione alla Kolmogorov). Chiamando S lo spazio degli eventi (ossia l'insieme di tutti i possibili risultati dell'esperimento) e considerando un evento A (ossia un sottoinsieme di S, $A \subset S$), la probabilità $\mathcal{P}_A$ associata ad A è un numero reale tale che:

a) $\mathcal{P}_S = 1$,

b) $\mathcal{P}_A \geq 0 \ \forall A$,

c) $\mathcal{P}(A \cup B) = \mathcal{P}_A + \mathcal{P}_B$, $\forall B$ tale che $A \cap B = \emptyset$.

Da questi assiomi si possono dimostrare una serie di proprietà tra cui: $\mathcal{P}(\bar{A}) = 1 - \mathcal{P}(A)$ (dove $\bar{A}$ è il complemento di A), $\mathcal{P}(B) \leq \mathcal{P}(A)$ se $B \subset A$, $\mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(B) - \mathcal{P}(A \cap B)$, ecc.. Questi assiomi costituiscono le fondamenta della probabilità ma non dicono nulla su come questa debba essere interpretata.

Si definisce **probabilità condizionata** $\mathcal{P}(A|B)$ (ossia la probabilità di A noto B):

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(A \cap B)}{\mathcal{P}(B)} \tag{1.1}$$

Due sottoinsiemi A e B sono detti **indipendenti** se $\mathcal{P}(A \cap B) = \mathcal{P}(A)\mathcal{P}(B)$. Notare che dire che due sottoinsiemi sono indipendenti non significa dire che sono disgiunti (ossia $A \cap B = \emptyset$).

$\mathcal{P}(A, B) = \mathcal{P}(A \cap B)$ viene chiamata **probabilità congiunta** mentre $\mathcal{P}(A \cup B)$ è detta probabilità disgiunta. Si può anche dimostrare che $\mathcal{P}(A) = \sum_B \mathcal{P}(A, B)$ e qui $\mathcal{P}(A)$ viene detta **probabilità marginale** di A.

Dalla definizione di probabilità congiunta e usando Eq. 1.1, si può dimostrare il **teorema di Bayes**:

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(B|A)\,\mathcal{P}(A)}{\mathcal{P}(B)} \ . \tag{1.2}$$

Una variabile X che assume uno specifico valore per ogni elemento di S è detta variabile casuale e costituisce un mapping dell'insieme degli eventi S nei numeri reali. Ad essa può essere associata una funzione di **densità di probabilità** f_X e una funzione di distribuzione cumulativa $F_X(x) \equiv \mathcal{P}_X(X \leq x)$. Il suo inverso, che viene detto quantile di ordine α, $F_X^{-1}(\alpha) \equiv inf(x : F(x) \geq \alpha)$ (dove $0 \leq \alpha \leq 1$) è il valore di x al di sotto del quale risiede una frazione α (o spesso espresso come percentile$= 100\alpha$) della probabilità totale. La mediana di una distribuzione è il valore che corrisponde al 50esimo percentile e non necessariamente corrisponde alla media (media di X pesata sulla densità di probabilità) o alla moda (valore per cui la densità di probabilità è massima). Per una variabile casuale continua si può scrivere la probabilità come un'integrale della densità di probabilità: $\mathcal{P}((a, b)) = \int_a^b f_X(x)\ dx$ (per la probabilità che X risieda nell'intervallo delimitato da a e b). La distribuzione cumulativa è quindi $F_X(x) = \int_{-\infty}^{x} f_X(t)\ dt$.

Tali concetti possono essere facilmente generalizzati al caso di un vettore casuale N-dimensionale $\mathbf{X} = (X_1, X_2, ..., X_N)$. Tipicamente si è interessati solamente ad alcune delle componenti di tale vettore. Consideriamo per esempio una componente X_i; integrando la distribuzione congiunta $f_{\mathbf{X}}(\mathbf{x})$ si definisce una funzione di **distribuzione marginale** di X_i:

$$f_{X_i}(x) = \int_{-\infty}^{+\infty} dx_1 ... dx_{i-1} dx_{i+1} ... dx_N \ f_{\mathbf{X}}(x_1, ..., x_{i-1}, x, x_{i+1}, ..., x_N) \ . \tag{1.3}$$

Consideriamo X e Y essere due variabili casuali continue con distribuzioni congiunta $f_{XY}(x, y)$ e marginalizzate $f_X(x)$ e $f_Y(y)$. Si definiscono:

- Valore di aspettazione di X: $E[X] = \mu_X = \int\ x f_X(x)\ dx$.
- Varianza di X: $Var(X) = \sigma_X^2 = E[(X - E[X])^2] = \int\ (x - \mu_X)^2 f_X(x)\ dx$.
- Covarianza di X e Y: $Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = \int\ (x - \mu_X)\,(y - \mu_Y)\, f_{XY}(x, y)\ dx\ dy$.
- Correlazione di X e Y: $\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X\, \sigma_Y}$.

Generalizzando il concetto precedentemente introdotto, X e Y sono variabili casuali indipendenti se la probabilità congiunta soddisfa: $f_{XY}(x, y) = f_X(x)\, f_Y(y)\ \forall (x, y)$ (il che implica $Cov(X, Y) = 0$ e $\rho_{XY} = 0$).

Sia l'approccio frequentista sia l'approccio bayesiano hanno in comune gli assiomi della probabilità e gli aspetti matematici della statistica. Anche il teorema di Bayes ha validità per entrambi gli approcci così come il fatto che in entrambi i casi si parla solitamente di statistica parametrica. Ciò che cambia è il significato da dare al concetto di probabilità, all'atteggiamento nel confronto dell'idea di una probabilità soggettiva e di conseguenza l'utilizzo e l'importanza che si dà al teorema di Bayes.

1.3 Distribuzioni di probabilità: il caso Gaussiano

In questa sezione verrà brevemente illustrata la distribuzione di probabilità gaussiana. Per altre distribuzioni (binomiale, Poissoniana, χ^2, ecc.) si faccia riferimento ai corsi degli anni precedenti.

La funzione di distribuzione di probabilità gaussiana di una variabile casuale x definita tra $-\infty$ e $+\infty$ è data da:

$$G(\mu, \sigma) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\,\sigma^2}} \ . \tag{1.4}$$

Usando le definizioni di cui sopra, non è difficile mostrare che il valore di aspettazione è μ e la varianza è σ^2. Nel caso gaussiano inoltre moda, media e mediana coincidono e sono date da μ.

Ci focalizziamo sulla gaussiana per via del **teorema del limite centrale**:
Qualsiasi variabile casuale che sia somma (o meglio combinazione lineare) di variabili indipendenti ($X =$

$\sum_i^n x_i$) segue una distribuzione Gaussiana con media $\sum_i \mu_i$ e varianza $\sigma_X^2 = \sum_i \sigma_i^2$ nel limite $n \to \infty$ indipendentemente dalle distribuzioni di probabilità delle x_i.

In pratica il teorema del limite centrale può essere applicato a diverse situazioni sperimentali in cui l'errore è somma di molte ed indipendenti cause: rumore strumentale, errori di lettura, contaminazioni, ecc.. In questi casi tipicamente si assume quindi che la misura sia distribuita gaussianamente.

Inoltre, con un cambio di variabile $z = (x - \mu)/\sigma$ la Gaussiana si traforma in distribuzione normale $N(z) = \exp(-z^2/2)/\sqrt{2\,\pi}$. Il caso gaussiano è anche il limite della distribuzione poissoniana quando il numero medio degli eventi diventa grande.

1.4 Inferenza frequentista: hypothesis testing e intervalli di confidenza

L'inferenza statistica è il procedimento attraverso cui si ottengono le caratteristiche di una popolazione dall'osservazione di una parte di essa, detta campione. Lo scopo dell'inferenza statistica è sostanzialmente interpretare i risultati di un espertimento in termini di un modello determinando una stima dei parametri del modello e relativi errori. E' un problema di inversione e, in un certo senso, c'è un rovesciamento di punto di vista rispetto al calcolo delle probabilità. Infatti quest'ultimo ha lo scopo di valutare la probabilità di un evento, mentre nell'inferenza si vuole ricostruire la distribuzione di probabilità in base all'osservazione degli eventi.

Quindi nell'inferenza statistica si vuole quantificare il livello di plausibilità di un'ipotesi (modello) rispetto ai dati sperimentali. Quando si considera una distribuzione come funzione dei parametri del modello a fissato risultato sperimentale, tale distribuzione viene detta funzione di verosimiglianza dei parametri (perchè indica quanto verosimilmente i valori dei parametri si accordano al risultato osservato), o **likelihood**. [1] La likelihood è l'elemento centrale della statistica frequentista, ed è importante distinguerla dalla distribuzione di probabilità come funzione del risultato sperimentale, a fissati parametri del modello (che invece è propria della statistica bayesiana, come vedremo).

Supponiamo di fare un esperimento composto di una serie di misure che ci forniscono dei dati $\mathbf{x}$ e che la probabilità congiunta dei dati sia una funzione di alcuni parametri $f(\mathbf{x}|\boldsymbol{\theta})$, con $\boldsymbol{\theta} = (\theta_1, ..., \theta_M)$ Valutando questa funzione con i dati ottenuti e considerandola come sola funzione dei parametri, si ha la likelihood $\mathcal{L}(\boldsymbol{\theta}) = f(\boldsymbol{x}|\boldsymbol{\theta})$.

Più precisamente, consideriamo una variabile casuale osservativa x con una certa distribuzione di probabilità $f(x|\boldsymbol{\theta})$ che dipende da alcuni parametri teorici incogniti $\boldsymbol{\theta}$. Questa probabilità è detta **probabilità condizionata** di avere la variabile x dati i parametri $\boldsymbol{\theta}$. Se la misura viene ripetuta N volte ottenendo $x_1, x_2, ..., x_N$ e le misure sono tra loro indipendenti, la probabilità congiunta (si vedano leggi enunciate nei paragrafi precedenti) di avere x_1 nell'intervallo dx_1, x_2 nell'intervallo dx_2, ecc.. è data da $\prod_{i=1}^N f(x_i|\boldsymbol{\theta})dx_i$. La definizione di likelihood è semplicemente data dalla distribuzione di probabilità congiunta:

$$\mathcal{L}(\boldsymbol{x};\boldsymbol{\theta}) = f(x_1|\boldsymbol{\theta})\, f(x_2|\boldsymbol{\theta}) \ldots f(x_N|\boldsymbol{\theta})\ . \tag{1.5}$$

Si opera però un cambio di prospettiva. Si considerano gli x_i fissati (cioè l'esperimento terminato) e la probabilità congiunta degli x_i come funzione solo dei parametri $\boldsymbol{\theta}$. La likelihood viene in effetti spesso semplicemente indicata come $\mathcal{L}(\boldsymbol{\theta})$, in quanto funzione solo di $\boldsymbol{\theta}$.

Data questa definizione è logico assumere come valore più plausibile per θ quello che massimizza $\mathcal{L}$. Il metodo della **maximum likelihood** (ML) per inferenza di parametri consiste quindi nel determinare il valore dei parametri che massimizza la likelihood. Assumendo che $\mathcal{L}$ sia una funzione differenziabile e che il massimo non si trovi agli estremi del range possibile per i parametri, il metodo consiste nel risolvere il sistema dato da:

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_j} = 0\ ,\ j = 1, ..., M\ . \tag{1.6}$$

Chiamiamo le soluzioni di questo sistema $\hat{\theta}_j$ (che sono i **maximum likelihood estimator** (MLE) dei θ_j). Essendo funzioni dei dati x_i sono variabili casuali. L'approccio frequentista è appunto quello di determinare la distribuzione dei $\hat{\theta}_j$ nota quella degli x_i. E' importante non confondere gli stimatori coi parametri: nell'approccio frequentista non è definita una distribuzione di probabilità per i parametri $\boldsymbol{\theta}$.

[1] In statistica la differenza tra probabilità e verosimiglianza è che la prima è usata prima che i dati siano disponibili per descrivere i possibili risultati sperimentali a fissato valore dei parametri, mentre la seconda è usata dopo che i dati sono disponibili come funzione dei parametri a fissato risultato sperimentale.

Non è garantito che il metodo della maximum likelihood sia quello ottimale, ma in pratica funziona molto bene in diverse situazioni (per una discussione più formale si vedano le referenze suggerite). Un aspetto importante è che lo stimatore è asintoticamente unbiased, ossia si può dimostrare che asintoticamente il valore di aspettazione dello stimatore è uguale al valore vero del parametro per qualsiasi valore vero. Inoltre, per grandi set di dati, è tipicamente efficiente (cioè la varianza ad esso associata è vicina al minimo).

Per ricondurci a qualcosa di familiare notiamo che è equivalente massimizzare la likelihood o il logaritmo della likelihood, in quanto il logaritmo è una funzione monotonicamente crescente. Consideriamo N variabili casuali x_i distribuite gaussianamente. La likelihood corrispondente ha la forma:

$$\mathcal{L}(x_1, ..., x_N; \boldsymbol{\theta}) = \prod_{i=1}^{N} \exp\left(-\frac{(x_i - \mu_i(\boldsymbol{\theta}))^2}{2\,\sigma_i^2}\right) . \tag{1.7}$$

Prendendo il logaritmo si ha:

$$-2\ln\mathcal{L}(x_1, ..., x_N; \boldsymbol{\theta}) = -2\sum_{i=1}^{N}\left(-\frac{(x_i - \mu_i(\boldsymbol{\theta}))^2}{2\,\sigma_i^2}\right) = \chi^2(\boldsymbol{\theta}) . \tag{1.8}$$

Quindi nel caso (molto comune) in cui si possono assumere distribuzioni gaussiane, la massimizzazione della likelihood corrisponde alla minimizzazione del χ^2 e il metodo maximum likelihood corrisponde al metodo dei minimi quadrati.

L'**intervallo di confidenza** $[\theta_1, \theta_2]$ di un certo parametro θ al livello α è definito come $F(\theta_1 < \hat{\theta} < \theta_2) = \alpha$, dove F è la cumulativa della likelihood. Si può mostrare che nel limite di un campionamento grande, la likelihood segue una distribuzione gaussiana centrata nel maximum likelihood estimator $\mathcal{L}(\theta) = \mathcal{L}_{max}\exp[-(\theta - \hat{\theta})^2/(2\sigma^2_{\hat{\theta}})]$ dove la varianza è in generale data da $Cov[\hat{\theta}_i, \hat{\theta}_j]^{-1} = E[-\frac{\partial^2 \ln\mathcal{L}}{\partial\theta_i\,\partial\theta_i}]$. Quindi il livello di confidenza a N-deviazioni standard è dato da: $\ln\mathcal{L}(\hat{\theta} \pm N\sigma_{\hat{\theta}}) = \ln\mathcal{L}_{max} - N^2/2$. In questo caso, la procedura è del tutto analoga al metodo standard in cui si utilizza la differenza di χ^2 e, per esempio, considerando un solo parametro, il livello di confidenza a 68% è dato da $\chi^2(\theta) - \chi^2_{min} = 1$.

Spesso uno è interessato anche al **goodness-of-fit**, ossia al livello di compatibilità tra l'ipotesi fatta (il modello) e i dati osservati. Questo può essere determinato costruendo un test statistico. Il cosiddetto **P-value** è la probabilità P di ottenere un risultato con un livello di compatibilità con l'ipotesi H_0 uguale o inferiore rispetto a quello osservato (assumendo H_0 sia vera). Da notare che non è la probabilità dell'ipotesi (non definita da un punto di vista frequentista) e che il P-value è una variabile casuale (essendo funzione dei dati) distribuita tra 0 e 1. Nel caso semplice di una variabile casuale x distribuita gaussianamente questo è dato dall'integrale della distribuzione χ^2:

$$P = \int_{\chi^2_{obs}}^{+\infty} f_{\chi^2}(x; n_{dof})\; dx , \tag{1.9}$$

dove n_{dof} è il numero di gradi di libertà (dati sperimentali - parametri stimati).

Nel caso del metodo della ML, si può considerare il valore della likelihood al suo massimo $\mathcal{L}_{max}$ come test del goodness-of-fit. D'altra parte, però, questo può non essere semplice da calcolare perchè, anche conosocendo la distribuzione della likelihood, non è nota a priori la distribuzione di $\mathcal{L}_{max}$. In questo caso, essa viene determinata attraverso un Monte-Carlo in cui si generano dei "dati" partendo dai valori determinati attraverso la ML applicata ai dati veri. Dalla distribuzione ottenuta si può quindi ottenere il P-value sostituendo tale distribuzione alla f_{χ^2} di Eq. 1.9, ossia calcolando il valore della distribuzione cumulativa associata per $\mathcal{L}_{max}$ pari al risultato osservativo.

Un test statistico utile quando si vogliono comparare due modelli è il **rapporto tra le likelihood**. Da esso si calcola il P-value e si decide quando un modello può essere rigettato in favore di un altro. Nel limite di grandi campionamenti, il logaritmo del likelihood ratio si riduce a una differenza di χ^2 e il teorema di Wilks afferma che (nel caso di modelli nested) tale test statistico è asintoticamente distribuito come una variabile χ^2 con numero di gradi di libertà dato dalla differenza tra il numero dei parametri dei due modelli.

Nell'inferenza frequentista si usa spesso la **profile likelihood**, cioè anzichè marginalizzare sui parametri a cui non si è interessati (**nuisance**) per determinare un intervallo di confidenza, si proietta la likelihood nel sottospazio dei parametri che si vuole investigare massimizzando il suo valore nelle direzioni (parametri) nuisance. Non ne discutiamo in dettaglio in quanto non esiste un analogo bayesiano, nel cui contesto questa procedura di massimizzazione non è definita perchè la risultante profile posterior non sarebbe interpretabile come una distribuzione di probabilità.

Chapter 2

Statistica bayesiana

Come abbiamo visto l'approccio frequentista può essere riassunto in: determinare un maximum likelihood estimator (che è funzione dei dati e quindi una variabile casuale) e la sua distribuzione come funzione delle distribuzioni assunte per i dati. Nella maggior parte dei casi questo step viene effettuato numericamente generando dei mock datasets e determinando per ognuno di essi il ML estimator. La likelihood e il metodo Monte-Carlo sono quindi alla base dell'approccio frequentista.

In questa procedura però non c'è spazio per una conoscenza pregressa, ossia per un prior. Essa in effetti non migliorerebbe molto le cose quando un esperimento è estremamente migliore del precedente o quando un esperimento misura un solo parametro (ossia il risultato non dipende da altri parametri poco noti). Queste caratterstiche si applicano a molti esperimenti, ma non a tutti.

La scelta dei prior è invece un ingrediente fondamentale della statistica Bayesiana. Ciò può essere considerato come una limitazione perchè introduce un elemento soggettivo, ma il principio guida della teoria bayesiana è che non esiste inferenza senza assunzioni e la scelta del prior mira a riflettere il più accuratamente (oggettivamente) possibile le assunzioni e la conoscenza del problema prima dell'esperimento. Inoltre, in numerose applicazioni scientifiche, la perdita di informazioni pregresse (ad esempio di esperimenti precedenti) può limitare la capacità di un esperimento di fornire informazioni.

Nell'inferenza bayesiana si assume che un modello teorico (descritto da certi parametri $\boldsymbol{\theta}$) sia vero e si cerca di ottenere la distribuzione di $\boldsymbol{\theta}$ noti i dati $\boldsymbol{x}$. Questa distribuzione è chiamata probabilità a posteriori, ossia la probabilità che i parametri assumano certi valori, una volta effettuato l'esperimento e avendo fatto una serie di assunzioni. Da tale posterior si calcolano poi i valori di aspettazione e errori dei parametri.

Come già detto, il teorema di Bayes è una semplice conseguenza degli assiomi della probabilità e dal punto di vista matematico non dà luogo a nessuna controversia. Il dibattito è invece relativo alla questione se esso debba essere o meno utilizzato come elemento base dell'inferenza.

2.1 Dai prior ai posterior

Dal teorema di Bayes in Eq. 1.2, la probabilità a posteriori (brevemente, **posterior**) è: [1]

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{x})} \ . \tag{2.1}$$

Abbiamo già visto che $p(\boldsymbol{x}|\boldsymbol{\theta})$ viene detta **likelihood**. Per esempio, nel caso gaussiano unidimensionale essa è data da:

$$\mathcal{L}(\boldsymbol{\theta}) \equiv p(x|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\,\sigma^2}} \exp\left(-\frac{(x-\mu(\boldsymbol{\theta}))^2}{2\,\sigma^2}\right). \tag{2.2}$$

La funzione $p(\boldsymbol{\theta})$ è detta probabilità a priori (brevemente, **prior**), e esprime ciò che è già noto relativamente ai parametri prima che l'esperimento venga condotto. Può essere relativa ad un esperimento precedentemente condotto o un'assunzione teorica (per esempio, variabili tipo vita media di un oggetto devono avere valori positivi). La letteratura relativa alla scelta dei prior è molto vasta.

Consideriamo qui il caso di assenza di informazioni. L'idea è di associare una probabilità uguale a uguali stati di conoscenza (e, in questo caso, siamo totalmente ignoranti). Una scelta standard è quella di

[1] Per comodità, in questo e nei prossimi capitoli, le distribuzioni di probabilità verranno tipicamente indicate con p anzichè f come nelle sezioni precedenti.

prior uniformi (flat prior). Siccome i prior entrano nella determinazione di una distribuzione di probabilità devono essere opportunamente normalizzati (per avere la probabilità normalizzata a 1). Nel caso di flat prior unidimensionale si ha $p(\theta) = (\theta_{max} - \theta_{min})^{-1}$. D'altra parte, però, se il parametro può variare di diversi ordini di grandezza e vogliamo dare lo stesso peso a tutti gli ordini, la scelta del prior è $p(\theta) \propto 1/\theta$, ossia flat in scala logaritmica. E' da notare che anche la scelta di flat prior può avere non banali conseguenze se il risultato dipende da una funzione non-lineare del parametro $f(\theta)$. Infatti $p(f) = p(\theta)d\theta/df$, e quindi il prior non è flat su f (e nel tentativo di introdurre una "non-informazione" su θ si può introdurre una forte informazione su f). Detto in altri termini, una differente parametrizzazione del problema può condurre a posterior non equivalenti.

La funzione $p(\boldsymbol{x})$ in Eq. 2.1 viene detta **evidenza** (o likelihood marginale):

$$p(\boldsymbol{x}) = p(\boldsymbol{x}) \int p(\boldsymbol{\theta}|\boldsymbol{x}) \mathrm{d}\boldsymbol{\theta} = \int p(\boldsymbol{x}) p(\boldsymbol{\theta}|\boldsymbol{x}) \mathrm{d}\boldsymbol{\theta} = \int \mathcal{L}(\boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \; . \tag{2.3}$$

Per la stima dei parametri (che viene fatta a fissato modello), essa è semplicemente una costante che normalizza la probabilità (quindi la probabilità relativa dei parametri non dipende da essa) e può essere ignorata. Diventa invece cruciale quando si vogliono confrontare diversi modelli, come vedremo.

Usiamo ora una notazione che evidenzia la possibilità che due osservatori, Tizio (T) e Caio (C), possano avere prior differenti:

$$p(\boldsymbol{\theta}|\boldsymbol{x}, I_i) = \frac{\mathcal{L}(\boldsymbol{\theta}) p(\boldsymbol{\theta}|I_i)}{p(\boldsymbol{x})} \quad (i = T, C) \; , \tag{2.4}$$

dove I_i è un'informazione ritenuta essere vera da i. Questa situazione può essere comune in scienza. La posterior però deve convergere ad un'unica soluzione per quanto riguarda l'inferenza. Vediamolo con un esempio. Supponiamo che Tizio e Caio abbiano un differente prior gaussiano su un certo parametro $p(\theta|I_i) = \exp[-(\theta - m_i)^2/(2\, s_i^2)]$ con $i = T, C$ e conducano insieme l'esperimento in cui misurano il parametro assumendo una risposta gaussiana dell'apparato sperimentale, quindi la loro likelihood sarà $\mathcal{L}(\theta) = \mathcal{L}_0 \exp[-(x - \theta)^2/(2\, \sigma^2)]$. Non è difficile calcolare la posterior (esercizio) che è nuovamente una gaussiana con media e varianza:

$$\bar{\mu}_i = \frac{x + m_i \, (\sigma/s_i)^2}{1 + (\sigma/s_i)^2} \quad , \quad \bar{\sigma}_i^2 = \frac{\sigma^2}{1 + (\sigma/s_i)^2} \; . \tag{2.5}$$

Ripetendo più volte l'esperimento, possiamo sostituire a x la sua media μ e alla varianza σ^2 la varianza sulla media $\sigma_\mu = \sigma^2/N$, dove N è il numero di misure. Siccome la varianza decresce al crescere delle misure come $1/N$, il rapporto σ_μ/s_i diventa sempre più piccolo (e μ si avvicina sempre più a $\bar{\mu}$). Quindi anche partendo da prior molto differenti Tizio e Caio convergeranno (quando la quantità di dati sarà sufficiente) alla stessa posterior.

Bisogna invece essere cauti nei casi in cui una scelta differente di prior teorici conduce a posterior differenti. Strettamente parlando significa che i dati non sono sufficientemente buoni (ossia l'informazione derivante non è sufficientemente grande da superare quella precedentemente nota).

Un'altra questione che è lecito porsi analizzando i risultati di un esperimento A è se sia equivalente aggiungere l'informazione di un esperimento passato B come prior rispetto a considerare i due esperimenti come un unico esperimento. Per dimostrarlo chiamiamo rispettivamente x, x' e xx' i dati dell'esperimento B, A e A+B, e I e xI l'informazione a priori dell'esperimento B e A (l'ultima tiene conto dell'esperimento B, da cui la notazione xI). La posterior dell'esperimento A, considerando il prior da B è:

$$p(\boldsymbol{\theta}|\boldsymbol{x}', \boldsymbol{x}I) = \frac{p(\boldsymbol{x}'|\boldsymbol{\theta}, \boldsymbol{x}I) p(\boldsymbol{\theta}|\boldsymbol{x}I)}{p(\boldsymbol{x}'|\boldsymbol{x}I)} = \frac{p(\boldsymbol{x}'\boldsymbol{x}|\boldsymbol{\theta}, I) p(\boldsymbol{\theta}|\boldsymbol{x}I)}{p(\boldsymbol{x}'|\boldsymbol{x}I) p(\boldsymbol{x}|\boldsymbol{\theta}, I)} = \frac{p(\boldsymbol{x}'\boldsymbol{x}|\boldsymbol{\theta}, I) p(\boldsymbol{\theta}|I)}{p(\boldsymbol{x}'|\boldsymbol{x}I) p(\boldsymbol{x}|I)} = \frac{p(\boldsymbol{x}'\boldsymbol{x}|\boldsymbol{\theta}, I) p(\boldsymbol{\theta}|I)}{p(\boldsymbol{x}'\boldsymbol{x}|I)} = p(\boldsymbol{\theta}|\boldsymbol{x}'\boldsymbol{x}, I) \tag{2.6}$$

dove l'ultima espressione è il teorema di Bayes considerando A e B come un unico esperimento (e si sono utilizzate le definizioni di probabilità congiunta nel secondo e quarto passaggio e il teorema di Bayes nel terzo).

2.2 Intervalli di credibilità

I più comuni stimatori $\hat{\theta}$ del valore di un parametro θ sono il **picco** (massimo, cioè la moda) della **posterior** (analogo al MLE) e la media. Quest'ultima è definita come $\hat{\theta} = \int \mathrm{d}\theta \, \theta \, p(\theta|\boldsymbol{x})$.

Per quanto riguarda l'errore associato si definisce una regione R_α di credibilità α come $\int_{R_\alpha} p(\boldsymbol{\theta}|\boldsymbol{x}) \mathrm{d}\boldsymbol{\theta} = \alpha$. Nel caso in cui $\boldsymbol{\theta}$ sia un vettore, si può far riferimento a due valori, l'**errore condizionale** e l'**errore**

marginale. L'errore condizionale è la minima barra di errore che si può associare a θ_i se tutti gli altri parametri θ_j sono noti ed è legato alla distribuzione della posterior a fissati valori di θ_j. Raramente questo valore viene quotato. Come abbiamo già visto marginalizzare consiste nell'integrare su un certo numero di parametri e la posterior marginale di un parametro θ_i è data da $p(\theta_i|\boldsymbol{x}) = \int \mathrm{d}\theta_1 ... \mathrm{d}\theta_{i-1}\mathrm{d}\theta_{i+1}\mathrm{d}\theta_N p(\boldsymbol{\theta}|\boldsymbol{x})$. Spesso si fa anche riferimento a livelli di credibilità 2D (ellissi) dove 2 parametri sono lasciati fuori dall'integrazione.

Da un punto di vista bayesiano si può definire una distribuzione predittiva per un evento non osservato. Chiamando x_{N+1} il caso non osservato e $x_1, ..., x_N$ quelli osservati, essa è data da:

$$p(x_{N+1}|x_1, ..., x_N) = \int p(x_{N+1}|\theta)p(\theta|x_1, ..., x_N)\mathrm{d}\theta \ , \tag{2.7}$$

ed è una media delle predizioni pesata su tutti i possibili valori del parametro θ con la rispettiva probabilità.

2.3 Selezione di modelli

L'evidenza, definita in Eq. 2.3, è lo strumento principale per confrontare modelli diversi e selezionare il migliore. Usando una notazione in cui la dipendenza dal modello M è esplicita:

$$p(\boldsymbol{x}|M) = \int p(\boldsymbol{x}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)\mathrm{d}\boldsymbol{\theta} \ . \tag{2.8}$$

Dall'evidenza si passa alla posterior del modello fissati i dati attraverso il teorema di Bayes: $p(M|\boldsymbol{x}) \propto p(M)p(\boldsymbol{x}|M)$ (la costante di normalizzazione legata ai dati è irrilevante perchè identica per qualsiasi modello). Solitamente non si parte con una preferenza per un particolare modello e i prior $p(M)$ sono dati da $1/N$, dove N è il numero di modelli considerati.

Comparando due modelli M_A e M_B il rapporto tra le posterior è dato da:

$$\frac{p(M_A|\boldsymbol{x})}{p(M_B|\boldsymbol{x})} = \frac{p(M_A)p(\boldsymbol{x}|M_A)}{p(M_B)p(\boldsymbol{x}|M_B)} \ . \tag{2.9}$$

Il **fattore di Bayes** è definito come il rapporto tra le evidenze $B_{AB} = p(\boldsymbol{x}|M_A)/p(\boldsymbol{x}|M_B)$ e (soprattutto per standard prior $p(M_A) = p(M_B)$) ci permette di decidere se un modello è migliore di un altro. Un valore del fattore di Bayes > 1 (< 1) supporta (sfavorisce) il modello A rispetto al B. La scala di riferimento per la selezione dei modelli è detta "Jeffreys' scale". Riassumendo si ha che il test è inconclusivo per $|\ln B| < 1$, di debole evidenza per $1 < |\ln B| < 2.5$ e di forte evidenza per $|\ln B| > 5$.

Notare che la selezione di modelli e l'inferenza parametrica sono trattate in modo decisamente separato. L'inferenza è calcolata all'interno di ciascun modello separatamente. La selezione dei modelli invece estende la valutazione delle ipotesi alla luce dei dati ottenuti allo spazio dei modelli teorici possibili.

Il metodo bayesiano in generale non fornisce un goodness-of-fit, solo probabilità relative (quando si fa inferenza il modello è assunto essere vero). Solitamente si procede comunque al calcolo del χ^2 per verificare che il fit sia sensato.

2.4 Limite frequentista

Nel caso di prior uniformi si ha $p(\boldsymbol{\theta}|\boldsymbol{x}) \propto \mathcal{L}(\boldsymbol{\theta})$ e l'inferenza bayesiana coincide con quella frequentista. Lo stimatore dato dal massimo nella posterior coincide col maximum likelihood estimator. Inoltre la selezione di modelli attraverso il fattore di Bayes si avvicina al metodo del likelihood ratio, utilizzando però la likelihood marginale (cioè integrata nello spazio dei parametri). In altre parole, la differenza si riduce a confrontare il valor medio delle due likelihood anzichè i picchi.

2.5 Connessione tra inferenza bayesiana e meccanica statistica

Storicamente le tecniche Markov-chain Monte-Carlo che vedremo nel prossimo capitolo sono state inizialmente sviluppate per risolvere problemi di meccanica statistica. Attualmente sono lo strumento principale di calcolo nell'inferenza bayesiana. Ciò è dovuto al fatto che esiste una stretta analogia tra meccanica statistica e inferenza parametrica bayesiana.

Finora abbiamo discusso di modelli dipendenti da una serie di parametri $\theta_1, ..., \theta_N$. Consideriamo per esempio un ensemble canonico, ossia un sistema termodinamico con temperatura T, volume e numero di

costituenti fissati e costanti, e immaginiamo che ciascun set di possibili valori dei parametri corrisponda ad un microstato $s(\boldsymbol{\theta})$. Assegnamo all'energia E_s del sistema in tale microstato l'espressione: $E_s(\boldsymbol{\theta}) = -T \ln[p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})]$ (ossia, a meno di una costante, il logaritmo della likelihood moltiplicata per il prior).

La funzione di partizione canonica è quindi:

$$Z_C = \sum_s \exp(-E_s/T) \to \int \mathrm{d}\boldsymbol{\theta} \exp(-E_s(\boldsymbol{\theta})/T) = \int \mathrm{d}\boldsymbol{\theta} p(\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta}) = p(\boldsymbol{x}) \ . \tag{2.10}$$

Quindi corrisponde all'evidenza bayesiana, ossia alla quantità fondamentale per selezionare modelli.

La distribuzione canonica di probabilità è infine data:

$$\frac{1}{Z_C} \exp(-E_s(\boldsymbol{\theta})/T) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x})} = p(\boldsymbol{\theta}|\boldsymbol{x}) \ , \tag{2.11}$$

che corrisponde alla posterior bayesiana. Quindi l'aspettazione rispetto alla posterior della statistica bayesiana equivale all'aspettazione rispetto alla distribuzione canonica nella meccanica statistica.

Chapter 3

Tecniche Markov-chain Monte-Carlo per inferenza di parametri

La soluzione generale per un qualsiasi problema di inferenza statistica bayesiana è stata discussa nei paragrafi precedenti. Ci sono pochi casi dove questa soluzione è semplice, come per prior gaussiani e likelihood gaussiana. Nella maggior parte dei casi, invece, può essere molto complicato o addirittura impossibile risolvere analiticamente l'equazione per la probabilità a posteriori. Anche per questa ragione in passato l'inferenza bayesiana era una corrente minoritaria in statistica. La possibilità di ricorrere a simulazioni numeriche ha invece portato, negli ultimi anni, a una rapidissima crescita della sua popolarità ed in molti ambiti scientifici ed economici è ormai considerata come lo strumento da utilizzare, essendo più flessibile e meno restrittiva dell'inferenza frequentista.

In particolare, le tecniche Markov Chain Monte Carlo (MCMC) sono attualmente il metodo computazionale maggiomente utilizzato per risolvere un problema di inferenza Bayesiana. Le discutiamo in questo capitolo. Esse hanno permesso di mappare numericamente la posterior, anche in casi molto complicati in cui, per esempio, la likelihood è ottenuta attraverso simulazioni numeriche, lo spazio dei parametri ha un numero elevato di dimensioni e la posterior ha una struttura complessa con molti picchi. Se il numero dei parametri è limitato (diciamo 3 o 4) allora può essere fattibile valutare la posterior con una griglia sufficientemente fine da individuare il picco e stimare gli errori, ma al crescere della dimensionalità N dello spazio dei parametri i punti della griglia necessari per avere una buona stima crescerebbero esponenzialmente con N, rendendo questo metodo inattuabile. Attraverso una tecnica MCMC si cerca invece di campionare più efficientemente la posterior concentrandosi nelle regioni dove essa è più grande e tralasciando lo spazio dei parametri dove la probabilità è molto piccola, come vedremo. In questo modo il numero di punti da calcolare cresce circa linearmente anzichè esponenzialmente con N.

L'idea di fondo è una procedura iterativa in cui si crea una catena di valori nello spazio dei parametri dove il valore θ_i è aggiornato al valore θ_{i+1} secondo un algoritmo tale che la distribuzione finale della catena segua la distribuzione di probabilità a cui si è interessati.

3.1 Catene di Markov

L'inferenza probabilistica attraverso catene di Markov consiste nel costruire sequenze di punti (le "catene") nello spazio dei parametri, la cui densità è proporzionale alla distribuzione di probabilità a posteriori a cui siamo interessati (Eq. 2.1). Esistono infatti catene di Markov che convergono ad un'unica e stazionaria distribuzione di probabilità e quindi possono essere utilizzate per stimare i valori di aspettazione di variabili rispetto a tale distribuzione. Il termine Monte-Carlo si riferisce al fatto che per la computazione si ricorre ad un ripetuto campionamento casuale (attraverso la generazioni di sequenze di numeri casuali).

Una catena di Markov è una sequenza di variabili casuali $Y_1, Y_2, Y_3, ..$ tale che la dipendenza della distribuzione di Y_{i+1} dai valori di $Y_1, .., Y_i$ è interamente codificata dal valore di Y_i, ossia
a) $\mathcal{P}(y_{i+1}|y_i, \{y_j, j = 0, ..., i-1\}) = \mathcal{P}(y_{i+1}|y_i)$,
cioè il passaggio ad uno stato del sistema dipende unicamente dallo stato immediatamente precedente e non dal come si è giunti a tale stato (dalla storia). La probabilità di transizione da uno stato y allo step i ad uno stato y' allo step $i+1$ è descritta attraverso una matrice $T_i(y, y')$, detta **matrice di transizione**. Essa può essere intesa come la distribuzione condizionale di Y' noto Y.

Si definisce inoltre una **distribuzione iniziale** λ tale che

b) $\mathcal{P}(y_0) = \lambda(y_0)$,
ossia una distribuzione marginale di Y_0 che fornisce la probabilità iniziale degli stati.

Una sequenza di variabili che rispetti a) e b) è una **catena di Markov**. Si può facilmente dimostrare che $\mathcal{P}(y_0, y_1, .., y_i) = \lambda(y_0)\, T_0(y_0, y_1)\, T_1(y_1, y_2) \ldots T_{i-1}(y_{i-1}, y_i)$ e che $T_{i+j}(y, y') = \sum_{\tilde{y}} T_i(y, \tilde{y}) T_j(\tilde{y}, y')$ (equazione di Chapman-Kolmogorov). Dalle definizioni di cui sopra si deduce che la probabilità di uno stato y al tempo i è legato alla corrispondente probabilità al tempo $i-1$ da $p_i(y) = \sum_{y'} p_{i-1}(y')\, T_{i-1}(y', y)$.

Una matrice di transizione si dice irriducibile se ogni stato comunica con tutti gli altri stati (ossia partendo da uno stato si può raggiungere un qualsiasi altro stato). Se uno stato y è tale che $T_i(y, y) = 1$, lo stato si dice assorbitore.

Se la probabilità di transizione dipende solamente dallo stato del sistema in cui si trova la catena, ma non dal numero di step effettuati, ossia la matrice di transizione è costante $T \equiv T(y, y')$ (cioè, rispetto alla notazione precedente, il pedice non è più rilevante), la catena di Markov è detta **omogenea** (o **stazionaria** visto che gli step sono spesso identificati come "tempi"). In questo caso la catena di Markov è interamente specificata una volta definite le probabilità iniziali λ dei vari stati (la probabilità marginale dei possibili Y_0) e la probabilità di transizione T da uno stato all'altro della catena (la probabilità condizionata di Y_{i+1} dati i possibili valori di Y_i).

Una distribuzione di probabilità $\pi(y)$ è **invariante** rispetto a una data catena di Markov con probabilità di transizione $T_i(y, y')$ se per ogni i vale:
$\pi(y) = \sum_{y'} \pi(y')\, T_i(y', y)$ (se la catena è omogenea semplicemente $\pi(y) = \sum_{y'} \pi(y')\, T(y', y)$),
ossia una volta che la catena raggiunge tale distribuzione questa non muta più al variare della catena.

Una catena di Markov omogenea è **reversibile** se (e solo se) soddisfa la condizione di **bilancio dettagliato**: $\pi(y)\, T(y, y') = \pi(y')\, T(y', y)$. Questa condizione implica che la distribuzione è invariante. Infatti:

$$\sum_{y'} \pi(y')\, T(y', y) = \sum_{y'} \pi(y)\, T(y, y') = \pi(y) \sum_{y'} T(y, y') = \pi(y)\ . \tag{3.1}$$

Ogni catena di Markov ha almeno una distribuzione invariante ma può averne anche più di una. Saremo interessati a catene di Markov che ammettono un'unica distribuzione invariante, detta di **equilibrio**. L'unicità è soddisfatta se la catena è **ergodica**, ed è verificata se le probabilità $p_i(y)$ convergono alla distribuzione invariante $\pi(y)$ quando $i \to \infty$, cioe' $\lim_{i\to\infty} p_i(y) = \pi(y)\ \forall y$ e indipendentemente dalla scelta delle probabilità iniziali $p_0(y)$. Assumeremo che tutte le catene considerate siano ergodiche (per dimostrazioni più rigorose si vedano le referenze consigliate).

In generale siamo interessati a catene di Markov che convergono ad una soluzione unica e stazionaria e in cui gli elementi della catena sono campioni dalla distribuzione di interesse che, nel caso di inferenza bayesiana, è la posterior $p(\theta|d)$. La generazioni di elementi di una catena ha una natura probabilistica e esistono diversi algoritmi per costruire catene di Markov. Nelle prossime sezioni esamineremo gli algoritmi Metropolis–Hastings, Gibbs e hybrid Monte-Carlo, e discuteremo brevemente un algoritmo in cui non si fa ricorso a catene di Markov (nested sampling). La scelta ottimale dell'algoritmo dipende dallo specifico problema e dalla posterior che si vuole esplorare.

Due aspetti da tenere in considerazione sotto questo punto di vista sono il periodo di burn-in e le correlazioni tra punti. Al crescere degli step della catena, la distribuzione di target viene sempre meglio approssimata. All'inizio del campionamento però la distribuzione può essere significativamente lontana da quella stazionaria. Ci vuole appunto un certo tempo prima di raggiungere la distribuzione stazionaria di equilibrio e tale periodo è detto di **burn-in**. I campioni provenienti da tale parte iniziale della catena vanno tipicamente scartati perchè possono non rappresentare accuratamente la distribuzione desiderata.

Normalmente un algoritmo MCMC genera catene di Markov di campioni, ognuno dei quali è autocorrelato a quelli generati immediatamente prima e dopo di lui. Conseguentemente campioni successivi non sono indipendenti ma formano una catena di Markov con un certo grado di **correlazione** (se si vogliono campioni indipendenti si può ad esempio considerare solo quelli generati ogni un certo numero, ad esempio ogni cento, e scartare il resto).

L'arte dei diversi algoritmi MCMC risiede nel rendere il meccanismo efficiente, il che implica anche la riduzione al minimo del tempo di burn-in e della correlazione tra diversi campioni. Ciò può dipendere ovviamente dalle specifiche caratteristiche del problema e dalla distribuzione che si vuole esplorare.

Una volta che la catena di Markov è costruita diventa abbastanza semplice ottenere stime di valori di aspettazione per qualsiasi funzione dei parametri in gioco. La media a posteriori è data da (dove $\langle\rangle$ denota il valore di aspettazione rispetto alla posterior)

$$\langle\theta\rangle = \int p(\theta|\boldsymbol{x})\theta \mathrm{d}\theta = \frac{1}{N}\sum_{i=1}^{N} \theta^{(i)}\ . \tag{3.2}$$

In generale si può ottenere il valore di aspettazione rispetto alla posterior di una funzione qualsiasi dei parametri $f(\boldsymbol{\theta})$:

$$\langle f(\theta)\rangle = \frac{1}{N}\sum_{i=1}^{N} f(\theta^{(i)}) \; . \tag{3.3}$$

Uno degli scopi principali dell'inferenza parametrica è determinare gli intervalli di credibilità dei parametri. Questo può essere fatto attraverso la probabilità marginale unidimensionale del parametro di interesse. Chiamando quest'ultimo θ_j (dove $\boldsymbol{\theta}$ è il vettore dei parametri di dimensione M), la sua probabilità marginale a posteriori $p(\theta_j|\boldsymbol{x})$ è data da:

$$p(\theta_j|\boldsymbol{x}) = \int p(\boldsymbol{\theta}|\boldsymbol{x})\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_{j-1}\mathrm{d}\theta_{j+1}\ldots\mathrm{d}\theta_M \; . \tag{3.4}$$

Questa può essere ottenuto dalla catena, dividendo il range di θ_j in una serie di bin e contando il numero di campioni all'interno di ciascun bin (ignorando i valori delle altre variabili θ_i), normalizzati al numero totale di campioni. Infatti gli elementi della catena sono campioni della probabilità a posteriori totale $p(\boldsymbol{\theta}|\boldsymbol{x})$. La probabilità ottenuta integrando tra due estremi (ossia sommando gli elementi di un certo numero di bin) ci dà l'associato intervallo di credibilità (per esempio, il 68%). Similmente si possono ottenere probabilità marginali a posteriori bidimensionali (le "ellissi").

3.2 Algoritmo Metropolis-Hastings

Come altri metodi MCMC, lo scopo dell'algortimo Metropolis–Hastings è di generare una collezione di stati che seguano una determinata distribuzione (nel caso bayesiano, la posterior). Consideriamo $\boldsymbol{\theta}$ un vettore di parametri con una likelihood $\mathcal{L}(\boldsymbol{\theta})$ e prior $p_{prior}(\boldsymbol{\theta})$. Come abbiamo visto, in una catena di Markov ci si muove da una posizione nello spazio dei parametri $\boldsymbol{\theta}_1$ a una nuova posizione $\boldsymbol{\theta}_2$ con probabilità di transizione $T(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, la quale è scelta in modo tale che la catena di Markov abbia $\pi(\boldsymbol{\theta})$ come distribuzione asintotica stazionaria. Per comodità definiamo $p_{ij} \equiv T(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$.

Nell'algoritmo Metropolis-Hastings questo è realizzato prendendo una arbitraria distribuzione (proposal density o jumping distribution) $p_{step}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i+1})$ per proporre un nuovo punto $\boldsymbol{\theta}_{i+1}$ a partire dal punto $\boldsymbol{\theta}_i$ in cui si trova la catena (per brevità $p^0_{ij} \equiv p_{step}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$) e il punto proposto è accettato con probabilità:

$$\alpha_{i,i+1} \equiv \alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i+1}) = min\left[1, \frac{p^0_{i+1,i}\,\pi(\boldsymbol{\theta}_{i+1})}{p^0_{i,i+1}\,\pi(\boldsymbol{\theta}_i)}\right] \; . \tag{3.5}$$

In altre parole si definisce la matrice di transizione come $p_{ij} = \alpha_{ij}\, p^0_{ij}$, con α da Eq. 3.5. L'accettanza α corregge per il fatto che la proposal density non è la target distribution.

Focalizziamoci ora sul caso particolare dell'algoritmo di Metropolis-Hastings in cui p_{step} è simmetrica (ossia $p^0_{ij} = p^0_{ji}$, da cui $\alpha_{i,i+1} = min\,[1, \pi(\boldsymbol{\theta}_{i+1})/\pi(\boldsymbol{\theta}_i)]$. Con questa costruzione è semplice mostrare che vale il principio del bilancio dettagliato $\pi(\boldsymbol{\theta}_{i+1})\,T(\boldsymbol{\theta}_{i+1}, \boldsymbol{\theta}_i) = \pi(\boldsymbol{\theta}_i)\,T(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i+1})$ e quindi la distribuzione di equilibrio è $\pi(\boldsymbol{\theta})$. Infatti:

$$\begin{aligned}
\pi(\boldsymbol{\theta}_{i+1})\,p_{i+1,i} &= \pi(\boldsymbol{\theta}_{i+1})\,\alpha_{i+1,i}\,p^0_{i+1,i} = min\,[\pi(\boldsymbol{\theta}_{i+1}), \pi(\boldsymbol{\theta}_i)]\,p^0_{i+1,i} = min\,[\pi(\boldsymbol{\theta}_i), \pi(\boldsymbol{\theta}_{i+1})]\,p^0_{i,i+1} \\
&= \pi(\boldsymbol{\theta}_i)\,\alpha_{i,i+1}\,p^0_{i,i+1} = \pi(\boldsymbol{\theta}_i)\,p_{i,i+1}
\end{aligned}$$

Nella pratica, l'algoritmo consiste in:

Inizializzazione: Scegliere un punto arbitrario $\boldsymbol{\theta}_0$ come primo campione e scegliere una densità di probabilità arbitraria $p_{step}(\boldsymbol{\theta}_{i+1}, \boldsymbol{\theta}_i)$ che suggerisce un candidato per il nuovo campionamento di $\boldsymbol{\theta}_{i+1}$, dato il precendete valore $\boldsymbol{\theta}_i$.

Ad ogni iterazione i: Partendo dalla distribuzione $p_{step}(\tilde{\boldsymbol{\theta}}_{i+1}, \boldsymbol{\theta}_i)$ generare un candidato $\tilde{\boldsymbol{\theta}}_{i+1}$ per il prossimo campionamento. Calcolare l'accettanza $\alpha = \mathcal{L}(\tilde{\boldsymbol{\theta}}_{i+1})p_{prior}(\tilde{\boldsymbol{\theta}}_{i+1})/(\mathcal{L}(\boldsymbol{\theta}_i)p_{prior}(\boldsymbol{\theta}_i))$ per decidere se il candidato è accettato o rifiutato. Se $\alpha \geq 1$ il candidato è più probabile di $\boldsymbol{\theta}_i$ ed è automaticamente accettato, $\boldsymbol{\theta}_{i+1} = \tilde{\boldsymbol{\theta}}_{i+1}$. Altrimenti il candidato è accettato con una probabilità α. Tipicamente questa condizione è implementata generando ogni volta un numero random α_m con distribuzione uniforme tra 0 e 1, e il candidato è accettato se $\alpha \geq \alpha_m$ o rigettato in caso contrario (per cui c'è una certa probabilità di rimanere nello stesso stato, ossia $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i$).

La scelta della proposal density è arbitraria. In linea di massima qualsiasi distribuzione che fa muovere nello spazio dei parametri funziona, ma le performance computazionali (accuratezza e velocità) possono

essere molto differenti. La scelta più semplice è una distribuzione gaussiana centrata nel valore precedente $\boldsymbol{\theta}_i$. Se approssimativamente si conosce la forma della distribuzione di probabilità da campionare, si possono però definire proposal density più efficienti.

3.3 Algoritmo di Gibbs

Nel caso di distribuzioni **multivariate** può non essere semplice determinare quale sia la proposal density da utilizzare (in particolare se il numero di dimensioni è elevato) perchè le diverse dimensioni potrebbero comportarsi in maniera differente. L'algoritmo di Gibbs cerca di ovviare a questa difficoltà scegliendo un nuovo campionamento per ogni dimensione separatamente dalle altre anzichè fare uno step in tutte le dimensioni contemporaneamente. Il punto fondamentale è che, per una distribuzione multivariata, può essere più semplice campionare partendo dalle **distribuzioni condizionate** piuttosto che dalla distribuzione congiunta o dalle distribuzioni marginali (che devono essere ottenute tramite un'integrazione della distribuzione congiunta).

Consideriamo una variabile N-dimensionale $\boldsymbol{\theta} = \theta_1, ..., \theta_N$. L'algoritmo consiste in:

Inizializzazione: Scegliere un valore arbitrario per ciascuna variabile del vettore $\boldsymbol{\theta}_0$ come primo campione e una densità di probabilità condizionata $p_{step}(\theta_j|\theta_1, ..., \theta_{j-1}, \theta_{j+1}, ..., \theta_N)$ che suggerisce un candidato per il nuovo campionamento di θ_j, dato il precendete valore.

Ad ogni iterazione i: Partendo dalla distribuzione $p_{step}(\tilde{\theta}_j^{(i+1)}|\theta_1^{(i+1)}, ..., \theta_{j-1}^{(i+1)}, \theta_{j+1}^{(i)}, ..., \theta_N^{(i)})$ generare un candidato $\tilde{\theta}_j^{(i+1)}$ per il prossimo campionamento di ogni variabile partendo dalla distribuzione condizionata di quella variabile sulle altre variabili. Dopodichè il valore della variabile è aggiornato non appena campionata.

La sequenza del campionamento costituisce una catena di Markov la cui distribuzione stazionaria corrisponde alla distribuzione di probabilità congiunta di interesse (nel caso di inferenza bayesiana, alla posterior congiunta). Questo è assicurato dal fatto che le transizioni avvengono seguendo la distribuzione condizionata. Se la distribuzione condizionata $\pi(\theta_j|\theta_1, ..., \theta_{j-1}, \theta_{j+1}, ..., \theta_N)$ non è nota, per la definizione di p_{step} si può introdurre una proposal density arbitraria e utilizzare, per esempio, l'algoritmo Metropolis–Hastings per effettuare il campionamento.

L'algoritmo di Gibbs è particolarmente utile quando è difficile campionare direttamente la distribuzione congiunta, ma la distribuzione condizionata di ogni variabile è nota e semplice da campionare. Per esempio, reti bayesiane[1] sono spesso specificate in termini di distribuzioni condizionate, e quindi in questo caso il campionamento di Gibbs è ideale per ottenere la posterior.

Esistono variazioni all'algoritmo base appena descritto (si vedano referenze suggerite), che si adattano meglio in specifiche situazioni. Più genericamente la definizione dell'algoritmo di Gibbs può essere intesa come un campionamento di distribuzioni multivariate (con numero di dimensioni elevate) in cui si campiona una variabile alla volta (e, come già detto, l'algoritmo Metropolis–Hastings può essere implementato per tali passaggi).

Gli algoritmi appena descritti (Metropolis–Hastings e Gibbs) sono potenti e abbastanza generali. Esistono però situazioni in cui essi devono essere maneggiati con cura. Specialmente per uno spazio dei parametri a molte dimensioni con una posterior multi-modale, tali tecniche possono presentare problemi. In generale la questione è sempre trovare un buon compromesso tra un campionamento fine (per avere una buona ricostruzione della posterior) e un campionamento efficiente (per avere un tempo computazionale accessibile). Se per esempio, però, esistono nello spazio dei parametri alcune "isole" molto piccole ma con una probabilità grande, gli algoritmi di cui sopra potrebbero essere inaccurati (non mappano tali isole) o inefficienti (per riuscire a mapparle si fanno step molto piccoli). Gli algoritmi seguenti offrono possibili vie di uscita in tali casi.

3.4 Hamiltonian (o Hybrid) Monte-Carlo

Questa tecnica è basata sull'analogia con un sistema fisico definito da un campo di potenziale coerente, che è il principale responsabile della dinamica del sistema, a cui si sommano effetti random cinetici. L'idea è di rendere il "walk meno random" in modo tale da riuscire ad effettuare step grandi (cosicchè la catena abbia un buon mixing tra grandi e piccoli step), ma in modo che la probabilità che lo step sia accettato sia comunque alta (cioè non si finisca spesso in regioni a bassa probabilità).

[1] Una rete bayesiana è un grafo orientato aciclico in cui i nodi sono le variabili casuali (l'equivalente dei parametri del modello nell'inferenza bayesiana). Ad ogni nodo è associata una funzione di probabilità condizionata che, noti i valori delle variabili dei nodi "genitori", assegna una probabilità alla variabile relativa al nodo "figlio".

Nell'HMC il campionamento avviene su uno spazio dei parametri più grande rispetto a quello che si vuole esplorare. Infatti per un modello a N parametri, vengono introdotte N variabili ausiliare (una per ogni parametro) e il metodo campiona una distribuzione con $2N$ dimensioni. In un certo senso l'HMC tratta i parametri come coordinate, la target distribution come un potenziale effettivo in questo sistema di coordinate e genera per ogni coordinata un generico momento. Lo spazio è esplorato trattando il problema come un sistema dinamico ed evolvendo lo spazio delle fasi risolvendo le equazioni dinamiche. Transizioni deterministiche (dettate dalla dinamica) vengono alternate con le usuali transizioni stocastiche. Alla fine, i momenti vengono ignorati (cioè marginalizzati) e si ottiene la distribuzione di interesse.

Più concretamente, lo schema è il seguente. Si definisce un "potenziale" attraverso la distribuzione di interesse: $U(\boldsymbol{\theta}) = -\ln \pi(\boldsymbol{\theta})$ (quindi nell'inferenza bayesiana $U(\boldsymbol{\theta}) = -\ln[\mathcal{L}(\boldsymbol{\theta})\, p_{prior}(\boldsymbol{\theta}|I)]$). Per ogni "coordinata" θ_α viene generato un "momento" u_α tipicamente da una distribuzione normale con media 0 e varianza 1. La distribuzione N-dimensionale del momento è quindi una semplice gaussiana multivariata $\mathcal{N}(\mathbf{u})$. L'"energia cinetica" è $K(\mathbf{u}) = \mathbf{u}^T\mathbf{u}/2$ e l'hamiltoniana è $H(\boldsymbol{\theta}, \mathbf{u}) = U(\boldsymbol{\theta}) + K(\mathbf{u})$.

Si può quindi definire una distribuzione target estesa $\pi_e(\boldsymbol{\theta}, \mathbf{u})$ data da:

$$\pi_e(\boldsymbol{\theta}, \mathbf{u}) = \exp(-H(\boldsymbol{\theta}, \mathbf{u})) = \exp(-U(\boldsymbol{\theta}))\exp(-K(\mathbf{u})) \propto \pi(\boldsymbol{\theta})\,\mathcal{N}(\mathbf{u})\ . \tag{3.6}$$

Marginalizzando la $\pi_e(\boldsymbol{\theta}, \mathbf{u})$ su $\mathbf{u}$ (cioè semplicemente ignorando la coordinata $\mathbf{u}$ associatata ad ogni punto della catena) si ottiene la $\pi(\boldsymbol{\theta})$ di interesse.

Il punto è che richiedendo che le equazioni di Hamilton ($\dot{\theta}_\alpha = \partial H/\partial u_\alpha$ e $\dot{u}_\alpha = -\partial H/\partial\theta_\alpha$) siano soddisfatte si avrebbe una H invariante. In questo modo anche π_e non varierebbe e l'accettanza sarebbe uguale a 1. Questo però richiederebbe di conoscere il potenziale (cioè ciò che vogliamo ottenere). Tipicamente invece si utilizza un'approssimazione analitica del potenziale (ottenuta per esempio runnando una breve catena MCMC fittata con una gaussiana multivariata) i cui gradienti possono quindi essere determinati analiticamente. Questo però fa sì ovviamente che la simulazione non sia esatta. Per curare tale aspetto il punto raggiunto dalla dinamica come un candidato per una transizione, anzichè come una transizione certa. La decisione relativa può essere presa utilizzando nuovamente l'algoritmo di Metropolis-Hastings, con accettanza definita da $\alpha = \Big[1, \exp(-H(\tilde{\boldsymbol{\theta}}_{i+1}, \mathbf{u}_{i+1}))/\exp(-H(\boldsymbol{\theta}_i, \mathbf{u}_i))\Big]$. Quindi il candidato può anche essere rigettato. Notare tra l'altro che in questo modo continua a valere il bilancio dettagliato.

Per un certo numero di step si evolve il sistema attraverso la dinamica hamiltoniana, per poi aggiornare il momento stocasticamente. Questo punto viene registrato come nuovo elemento della catena (e quindi gli elementi della catena non corrispondono al numero degli step effettuati). Uno dei dettagli rilevanti risiede proprio nella scelta del numero di step da effettuare prima di generare un nuovo elemento della catena.

I vantaggi di questo approccio sono l'alta probabilità di accettanza dei nuovi punti e la minore correlazione tra stati successivi (grazie all'introduzione di $K(\mathbf{u})$), il che può velocizzare la convergenza dell'algoritmo alla $\pi(\boldsymbol{\theta})$ cercata.

Un altro algoritmo molto utilizzato in caso di distribuzioni multimodali e di necessità di far compiere alla catena step lunghi è il Metropolis Coupled MCMC (MC^3). Brevemente, esso consiste nel costruire diverse catene simultaneamente e "riscaldare" ad ogni step una di esse, ossia sottrarre alla probabilità un certo offset in modo tale che i picchi della distribuzione siano più bassi. In questo modo per la catena diventa meno improbabile uscire dalla regione di un picco una volta entrati e questa catena viene utilizzata come "scout" di regioni ad alta probabilità.

3.5 Nested sampling

Come già accennato nel caso di distribuzioni multimodali l'ultilizzo di catene di Markov deve essere effettuato con cautela. Per completezza menzioniamo brevemente un algoritmo molto efficace in questo caso e nel quale si risolve il problema di statistica bayesiana senza far ricorso a una catena di Markov.

Differentemente dagli algoritmi visti finora, i quali miravano ad ottenere la posterior, nel nested sampling viene calcolata l'evidenza $Z \equiv p(\mathbf{x})$. Ciò viene effettuato attraverso l'integrale $Z = \int_0^1 \mathcal{L}(X)\, dX$ dove $dX = p_{prior}(\theta)\, d\theta$. Senza entrare nei dettagli, l'idea di fondo di questa procedura è di "ordinare" iterativamente $\mathcal{L}$ rispetto a X in modo via via più accurato, in modo tale da calcolare l'integrale come semplice somma $\sum_i \mathcal{L}_i\, \Delta X_i$. Come by-product di questa integrazione si ottiene anche la posterior (da $p_i = \mathcal{L}_i\, \Delta X_i/Z$). Si può dire che il nested sampling progressivamente trasforma la distribuzione a priori in quella a posteriori richiedendo che i campionamenti abbiano likelihood crescente. Questo viene fatto cominciando con un campionamento casuale di un certo numero di punti nello spazio dei parametri effettuato attraverso distribuzione a priori. Ad ogni iterazione il punto con la likelihood più bassa viene scartato e viene sostituito con un punto campionato sempre con la prior ma soggetto alla richiesta di avere

una likelihood maggiore. In questo modo si costruisce una lista di campionamenti (quelli "scartati") con i quali è possibile calcolare l'integrale sopra menzionato, riducento un integrale multidimensionale (com'è genericamente il calcolo dell'evidenza) in un integrale unidimensionale in termini di frazioni di volumi di prior.

Chapter 4

Esempi

4.1 Catena di Markov per sistema a 2 stati

Consideriamo un sistema a due stati 1 e 2 schematizzato in Fig. 4.1. Quando il sistema si trova nello stato 1 ha una probabilità α di transire allo stato 2 (e quindi $1-\alpha$ di rimanere nello stato 1), mentre quando il sistema si trova nello stato 2 ha una probabilità β di transire allo stato 1 (e quindi $1-\beta$ di rimanere nello stato 2). Intuitivamente si può notare che la probabilità di trovarsi nello stato 1 sarà $\beta/(\alpha+\beta)$, mentre di trovarsi nello stato 2 sarà $\alpha/(\alpha+\beta)$. Verifichiamo che una catena di Markov conduce a tale conclusione.

La matrice di transizione per tale sistema è data da:

$$T = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} \tag{4.1}$$

dove gli autovalori di tale matrice sono $y_1 = 1$ e $y_2 = 1-\alpha-\beta$ (ottenuti tramite l'usuale $det(T - y\mathbb{I}) = 0$) e chiamiamo T_d la matrice diagonale formata da tali autovalori.

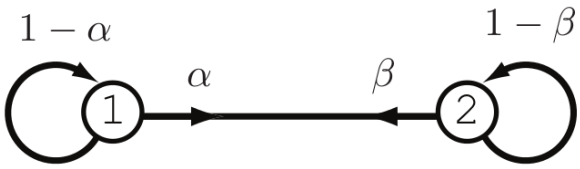

Figure 4.1: Sistema a due stati.

La soluzione generale dopo n-step è dunque $T^{(n)} = U\, T_d^n\, U^{-1}$ (dove U è la matrice con gli autovettori di T per colonne). Consideriamo la probabilità per cui partendo da 1 si torna a 1 dopo n-step (cioè l'elemento (1,1) della matrice), $T_{11}^{(n)} = A + B(1-\alpha-\beta)^n$ (con A e B che derivano dal prodotto con le matrici U e U^{-1}). Si può determinare A e B notando che $T_{11}^{(0)} = 1$ e $T_{11}^{(1)} = 1-\alpha$ e eguagliandoli all'espressione generale, da cui ottiene $A = \beta/(\alpha+\beta)$ e $B = \alpha/(\alpha+\beta)$. Quindi la probabilità di essere nello stato 1 partendo da 1 è data da:

$$\mathcal{P}_1(Y_n = 1) = T_{11}^{(n)} = \frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}(1-\alpha-\beta)^n \ . \tag{4.2}$$

Facendo analoghe considerazioni per i casi in cui si parte dallo stato 2 o si vuole determinare la probabilità di trovarsi in 2 (cioè per $T_{12}^{(n)}$, $T_{21}^{(n)}$ e $T_{22}^{(n)}$), si ottiene:

$$T_{ij}^{(n)} = \begin{bmatrix} \frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}(1-\alpha-\beta)^n & \frac{\alpha}{\alpha+\beta} - \frac{\alpha}{\alpha+\beta}(1-\alpha-\beta)^n \\ \frac{\beta}{\alpha+\beta} - \frac{\beta}{\alpha+\beta}(1-\alpha-\beta)^n & \frac{\alpha}{\alpha+\beta} + \frac{\beta}{\alpha+\beta}(1-\alpha-\beta)^n \end{bmatrix} \to \begin{bmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \\ \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{bmatrix} , \tag{4.3}$$

dove nel secondo passaggio si è preso il limite per $n \to \infty$. Abbiamo quindi verificato che la catena di Markov (nel limite di grande campionamento) ci dà una probabilità di trovarsi nello stato 1 uguale a $\beta/(\alpha+\beta)$ e di trovarsi nello stato 2 uguale a $\alpha/(\alpha+\beta)$, indipendentemente dallo stato di partenza. Il rate di convergenza è esponenziale.

4.2 Alberi filogenetici

Un albero filogenetico è un diagramma rappresentante le relazioni di discendenza tra diversi gruppi di organismi. La ricostruzione di alberi filogenetici è alla base della prova dell'evoluzionismo. Quest'ultimo prevede infatti che lo sviluppo delle forme di vita sia avvenuto a partire da un progenitore comune (la "radice"), il quale ha dato origine per speciazione a diverse linee di discendenza, fino ad arrivare alle specie attualmente esistenti (le "foglie").

Diverse caratteristiche degli organismi (morfologiche e biochimiche) sono state utilizzate nel corso degli anni (secoli) per la derivazione degli alberi filogenetici. Attualmente l'analisi delle sequenze molecolari gioca un ruolo fondamentale in tale ricostruzione.

Nella prima sottosezione discutiamo come costruire un modello probabilistico di evoluzione delle sequenze. Nella seconda viene descritto come assemblare un albero filogenetico che descriva tale evoluzione, ossia come fare inferenza. La probabilità a posteriori di un albero filogenetico T_d, nota una matrice di sequenze molecolari $\boldsymbol{L}$ è data da:

$$\mathcal{P}(T_d|\boldsymbol{L}) = \frac{\mathcal{P}(\boldsymbol{L}|T_d)\,\mathcal{P}(T_d)}{\sum_{d'=1}^{D_s}\mathcal{P}(\boldsymbol{L}|T_{d'})\,\mathcal{P}(T_{d'})}\ , \tag{4.4}$$

dove $\mathcal{P}(\boldsymbol{L}|T_d)$ è la likelihood dell'albero T_d, mentre $\mathcal{P}(T_d)$ è la probabilità a priori dell'albero T_d. Il numero totale di alberi D_s è determinato (usando calcolo combinatorio) dal numero di specie e dal tipo di albero (con/senza radice). Un prior non informativo che viene spesso utilizzato è $\mathcal{P}(T_d) = 1/D_s$.

4.2.1 Modello filogenetico

In un albero, qualsiasi coppia di punti è unita da un unico percorso (non esistono loop). Considereremo alberi biforcati (dove ogni vertice ha uno o tre vicini) e radicati (l'albero si sviluppa a partire da un unico nodo, detto radice) come nell'esempio in Fig. 4.2.

Assumiamo le sostituzioni di lettere nelle sequenze molecolari avere le seguenti proprietà:

1. Non esistono altri processi oltre a sostituzioni (come inserzioni o delezioni), quindi tutte le sequenze hanno uguale lunghezza.

2. Le sostituzioni in una determinata posizione sono indipendenti dalle sostituzioni nelle altre posizioni.

3. La probabilità di sostituzione dipende esclusivamente dallo stato in cui ci si trova e non dalla storia passata.

4. La probabilità di sostituzione è indipendente dalla posizione.

La proprietà 3) significa considerare l'evoluzione come un processo Markoviano. La probabilità di sostituzione di una lettera $\chi^i(t)$ nella posizione i della sequenza in un tempo di evoluzione $t > 0$ è data da

$$\mathcal{P}^i_{XY}(t) = \mathcal{P}(\chi^i(t+s) = Y|\chi^i(s) = X)\ . \tag{4.5}$$

Considereremo inoltre $\mathcal{P}_{XY}$ indipendente da $s \geq 0$, ossia una catena omogenea. Dalla proprietà 4) possiamo anche eliminare la dipendenza dalla posizione i, $\mathcal{P}^i_{XY}(t) = \mathcal{P}_{XY}(t)$.

Nella realtà nessuna delle 4 proprietà enunciate è realizzata in natura (da un DNA o altre sequenze). Possono però essere considerate come un'utile approssimazione di ordine zero.

Notare che la probabilità di sostituzione gioca il ruolo della matrice di transizione descritta nel capitolo precedente, ossia $T(X, Y)$ (la dipendenza dal tempo di T era stata in un certo senso omessa considerando intervalli di "tempo" sempre uguali tra i punti i e $i+1$ e catene omogenee). Valgono quindi per $\mathcal{P}_{XY}$ tutte le proprietà enunciate introducendo la matrice di transizione.

Come già menzionato, una catena omogenea è totalmente specificata dalla probabilità iniziale e dalla matrice di transizione "iniziale". Vediamolo in questo caso. La probabilità di sostituzione al tempo iniziale $t = 0$ è banalmente data dalla matrice $\mathcal{P}_{XY}(0) = \mathbb{I}$, ossia dalla matrice identità $M \times M$, dove M è il numero di lettere dell'alfabeto A. Per esempio, nel caso del DNA, l'alfabeto è composto da $A = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ e la matrice della probabilità di sostituzione è una 4×4. Il rate di sostituzione può essere calcolato come la derivata a $t = 0$:

$$Q = \mathcal{P}'_{XY}(0) = \lim_{t\to 0^+}\ \frac{\mathcal{P}_{XY}(t) - \mathbb{I}}{t}\ . \tag{4.6}$$

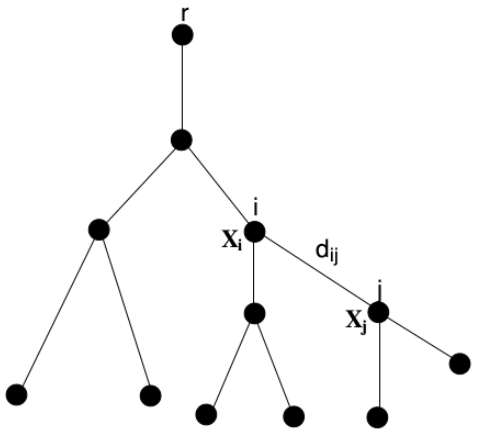

Figure 10.1: A Simple Binary Phylogenetic Tree. r is the root; d_{ji} is the time distance between i and j; X_i is the letter assigned to the hidden vertex i. The observed letters are at the leaves at the bottom. The probability of the substitution from vertex i to vertex j is $p_{X_j X_i}(d_{ji})$.

Figure 4.2: Esempio di albero filogenetico (da "Bioinformatics", Baldi and Brunak).

Da questa relazione e dalle proprietà della matrice di transizione si ottiene $\mathcal{P}_{XY}(t+dt) = \mathcal{P}_{XY}(t)\mathcal{P}_{XY}(dt) = \mathcal{P}_{XY}(t)(\mathcal{P}_{XY}(0) + Q\,dt) = \mathcal{P}_{XY}(t)(\mathbb{I} + Q\,dt)$. Combinando il primo e l'ultimo termine si ha:

$$\mathcal{P}'_{XY}(t) = \frac{\mathcal{P}_{XY}(t+dt) - \mathcal{P}_{XY}(t)}{dt} = \mathcal{P}_{XY}(t)\,Q \quad \Rightarrow \quad \mathcal{P}_{XY}(t) \propto e^{Q\,t}\ . \tag{4.7}$$

L'evoluzione è quindi definita dalla matrice dei rate di sostituzione Q. In generale, si possono avere $M \times M$ diverse espressioni del rate di sostituzione per ognuna delle possibili transizione da una lettera all'altra. Per semplicità, assumiamo il rate di sostituzione all'unità di tempo per ogni lettera essere identico e dato da una costante λ (in generale, il modello può essere molto più complicato e λ potrebbe anche variare da ramo a ramo dell'albero). Denotiamo invece come $\mathcal{P}_X$ la probabilità che la lettera X sia scelta quando la sostituzione avviene. λ e $\mathcal{P}_X$ possono essere ottenuti direttamente dai dati.

Non è difficile intuire che la probabilità di NON sostituzione è data da $e^{-\lambda\,t}$, mentre la probabilità di sostituzione è:

$$\mathcal{P}_{XY}(t) = e^{-\lambda\,t}\,\delta_{XY} + \left(1 - e^{-\lambda\,t}\right)\mathcal{P}_Y\ . \tag{4.8}$$

Il primo termine è relativo alla non-sostituzione, mentre il secondo descrive la transizione da X a $Y \neq X$. Notare che abbiamo definito un processo reversibile, per cui vale il principio del bilancio dettagliato: $\mathcal{P}_{YX}(t)\mathcal{P}_Y = \mathcal{P}_{XY}(t)\mathcal{P}_X$.

4.2.2 Costruzione della likelihood e inferenza bayesiana

Un modello filogenetico è dato da un modello probabilistico dell'evoluzione, cioè da una descrizione stotastica della sostituzione di lettere come appena descritto, combinato con un modello di albero filogenetico.

Vogliamo ora determinare quale siano la topologia dell'albero filogenetico e la lunghezza dei rami più probabili dal confronto con le sequenze molecolari osservate. Si tratta quindi di un problema di inferenza dove il modello è dato dalla posizione dei nodi (ossia dalla topologia dell'albero e dalla lunghezza dei rami) e i dati sono appunto le sequenze. Nel caso in cui i rate di sostituzione non siano noti possono anch'essi entrare a far parte dei parametri del modello.

$$\boldsymbol{L} = \{L_{kn}\} = \begin{array}{c} \text{Specie 1} \\ \text{Specie 2} \\ .. \\ \text{Specie K} \end{array} \left\{ \begin{array}{cccccc} \text{A} & \text{A} & \text{C} & .. & \text{C} & \text{T} \\ \text{A} & \text{A} & \text{C} & .. & \text{G} & \text{T} \\ .. & .. & .. & .. & .. & .. \\ \text{A} & \text{C} & \text{C} & .. & \text{C} & \text{T} \end{array} \right\} \tag{4.9}$$

La costruzione della likelihood può essere effettuata partendo dalla descrizione probabilistica della sostituzione di lettere appena discussa. Consideriamo K sequenze di lunghezza N costituite dalle lettere di un alfabeto A (per esempio nel caso del DNA si ha una sequenza di miliardi di nucleotidi A, C, G, T), ciascuna relativa ad una specie. In generale le sequenze potrebbero avere lunghezza diversa, ma, come

già menzionato, considereremo l'approssimazione in cui questa non varia. Possiamo quindi costruire una matrice L di dimensioni $K \times N$ combinando le sequenze a disposizione. Chiamiamo L_{kn} la lettera alla posizione n-esima nella sequenza k-esima e denotiamo con $\boldsymbol{L}_k$ la sequenza k-esima. Nel caso del DNA, $L_{kn} =$ può essere A, C, G o T. Consideriamo inoltre un albero filogenetico T con una certa radice r, un certo numero di nodi I e chiamiamo d_{ij} la distanza tra i nodi i e j. La likelihood è data da $\mathcal{P}(\boldsymbol{L}_1, ..., \boldsymbol{L}_K|T)$. La proprietà 2) delle sostituzioni introdotta precedentemente implica l'indipendenza di una colonna di $\boldsymbol{L}$ dalle altre, ossia:

$$\mathcal{P}(\boldsymbol{L}_1, ..., \boldsymbol{L}_K|T) = \prod_{n=1}^{N} \mathcal{P}(L_{1n}, ..., L_{Kn}|T) \ . \tag{4.10}$$

Possiamo quindi studiare separatamente le varie colonne e le probabilità $\mathcal{P}(L_{1n}, ..., L_{Kn}|T)$. I "parametri" del modello, ossia le variabili casuali in senso bayesiano, in questo problema sono dati dai vertici, o meglio, ogni vertice i ha una lettera associata χ_i e tali lettere sono le variabili. Notare che non sono quantità osservative. Un albero filogenetico può essere visto come una rete bayesiana. Infatti, una rete bayesiana è un grafo aciclico orientato in cui i nodi rappresentano le variabili casuali (in senso bayesiano) e ai rami è associata una probabilità condizionata tra i nodi connessi. La rete fornisce la distribuzione di probabilità congiunta delle variabili.

Nel nostro caso la probabilità condizionata del nodo j noto il nodo "genitore" i è data da:

$$\mathcal{P}(\chi_j = Y|\chi_i = X) = \mathcal{P}_{XY}(d_{ij}) \ , \tag{4.11}$$

dove $\mathcal{P}_{XY}$ è la probabilità di sostituzione in Eq. 4.8 e la distanza d_{ij} misura il tempo.

Attraverso questa procedura[1] è possibile quindi stabilire la probabilità (la verosimiglianza) di un determinato albero rispetto alle sequenze misurate.

Abbiamo dunque la base per poter definire un processo di inferenza statistica. Una tecnica statistica molto utilizzata anche in filogenetica computazionale è quella della maximum likelihood descritta nel primo capitolo. La likelihood globale viene costruita partendo dagli argomenti sopracitati. Un altro metodo molto comune è quello della massima parsimonia, dove gli alberi filogenetici favoriti sono quelli che implicano il minor numero di eventi di evoluzione (sostituzioni) per spiegare i dati.

Computazionalmente il problema dell'inferenza in ambito filogenetico è molto complesso. Si pensi, ad esempio, che il numero dei possibili alberi cresce in modo superesponenziale al crescere del numero di specie considerate. In questo senso, l'applicazione di tecniche MCMC e conseguente inferenza bayesiana può in molti casi portare ad un significativo avanzamento. L'inferenza bayesiana permette inoltre di introdurre una informazione a priori fornita per esempio da indipendenti evidenze biologiche (per esempio legate alle proprietà dei nucleotidi) o morfologiche, o prior di tipo tassonomico.

Per algoritmi ispirati al metodo Metropolis-Hastings visto nel precedente capitolo, l'idea consiste nel generare un albero candidato ad ogni step attraverso una determinata procedura, valutare la probabilità di transizione rispetto all'albero in cui si trova la catena (determinando la probabilità attraverso le probabilità di sostituzione come accennato precedentemente e tenendo conto di un eventuale prior) [2] e stabilire un criterio per accettare/rigettare il candidato. L'algoritmo MCMC deve ovviamente avere la distribuzione a posteriori dei nodi dell'albero come distribuzione invariante. L'arte risiede nel definire metodi efficienti per generare modelli di albero ad ogni step della catena e nel criterio per accettare/rigettare tali modelli.

In generale i metodi MCMC utilizzati in questo ambito possono essere molto complessi e costituiscono attualmente un tema importante nell'ambito della ricerca in filogenetica.

4.3 Codice numerico

Qui di seguito riportiamo l'esempio di un codice $C++$ che determina la posterior di un set di parametri utilizzando l'algoritmo Metropolis-Hastings. Per semplicità consideriamo un caso unidimensionale con una likelihood gaussiana e flat (o gaussian) prior. Tale scheletro può essere esteso all'analisi di problemi più complessi.

[1] Non abbiamo visto nel dettaglio un algoritmo di rete bayesiana ma la letteratura a rigurdo è molto vasta e chi fosse interessato può fare riferimento alle referenze suggerite.

[2] Come già accennato, in generale, la probabilità di sostituzione può essere anch'essa considerata parte del modello, aggiungendo quindi parametri da esplorare attraverso le catene.

File: /home/marco/Didattica/Statistica/esempio_MCMC_MH.cc Page 1 of 2

```
#include <cmath>
#include <cstdlib>
#include <cstdio>
#include <fstream>
#include <iostream>
#include <string>
#include <vector>
using namespace std;
// Esempio di MCMC attraverso l'algoritmo di Metropolis
// NB: per un caso più complesso rivedere la generazione di numeri casuali

const string title("Metropolis-Hastings Algorithm Example");

double Likel (double theta) { // likelihood
  double mu = 0.0, sigma=1.0;
  return exp(-pow((theta-mu)/sigma,2)/2.0);
}

double Prior (double theta) { // prior
//  if(theta<=0) return exp(-pow(theta/0.01,2)/2.0); // esempio di prior
  return 1.0;
}

double P (double theta) {return Likel(theta)*Prior(theta);} // posterior


double theta0 = 0.0, thetaWalker = theta0; // posizione iniziale (introdotta in input)
double delta = 0.1; // max step size

bool MarkovStep ( ) { // Metropolis-Hastings method
  double thetaTrial = thetaWalker + delta*(2.0*rand()/(1.0*RAND_MAX)-1.0); //
candidato //NB: delta*(2.0*rand()/(1.0*RAND_MAX))-1.0 genera un valore random tra -
delta e delta.
  double ratio = P(thetaTrial)/P(thetaWalker);
//    cout<< "thetaWalker = "<< thetaWalker<< "thetaTrial = "<< thetaTrial<<" ratio =
"<<ratio<<" P = "<<P(thetaWalker)<< endl;
  if (rand()/(1.0*RAND_MAX) < ratio) { // condizione per accettanza // NB: rand()/
(1.0*RAND_MAX) genera numero random tra 0 e 1.
    thetaWalker = thetaTrial;
    return true;}
  else return false;
}


int stepsToDiscard = 10; // steps da scartare per evitare correlazione
unsigned long steps = 0; // steps proposti (nella catena finale)
unsigned long accepts = 0; // steps accettati


void MCMCStep ( ) { // Step della catena (cioè per evitare correlazione si può
scartare un certo numero di step effettuati, quindi step nella catena < step
effettuati)
  for (int i = 0; i < stepsToDiscard; i++) MarkovStep();
  if (MarkovStep()) ++accepts;
  ++steps;
}


double thetaMax = 10.0; // estremo dell'intervallo per l'istogramma della probabilita'
const int bins = 100; // numero di bin nell'istogramma
vector<double> histogram(bins); // istogramma
int stepsMC = 20000; // step in una singola catena  (introdotti in input)
```

```
int stepsBurnin=100; // step di burn-in da scartare in una singola catena
int nmc=1; // numero di catene


void MCMC ( ) {  // catena
  thetaWalker = theta0;
  for (int i = 0; i <= stepsMC; i++) {
    MCMCStep();
    int bin = int((thetaWalker+thetaMax)/(2.0*thetaMax)*bins); // bin da incrementare
    if (bin >= 0 && bin < bins && i > stepsBurnin) ++histogram[bin]; }
  ofstream data_file("histogram.dat"); // file con istogramma
  for (int bin = 0; bin < bins; bin++) {
    double theta = -thetaMax + (bin + 0.5) * (2.0*thetaMax)/bins; // valore centrale
del bin
    data_file << theta << " " << histogram[bin]/(stepsMC*nmc*(2.0*thetaMax)/bins*1.0)
<< '\n';
  }
  data_file.close();
  double percent = int(accepts / double(steps) * 100.0);
  cout <<" Metropolis-Hastings steps = " << steps << ", accepted = " << percent <<
"%"<<endl;
}


int main () {
cout << title << endl;
//cout<<"How big is RAND_MAX ? "<<RAND_MAX<<endl; // per controllare di non aver
impostato troppi presunti random step
cout << " Posizione da cui partire: theta0 = "; cin >> theta0; // posizione da cui
iniziare la catena
cout << " Numero di step per catena = "; cin >> stepsMC; // step in una singola catena

// set random number generator seed
srand (time(NULL));
for (int i = 0; i < nmc; i++) MCMC();
}
```

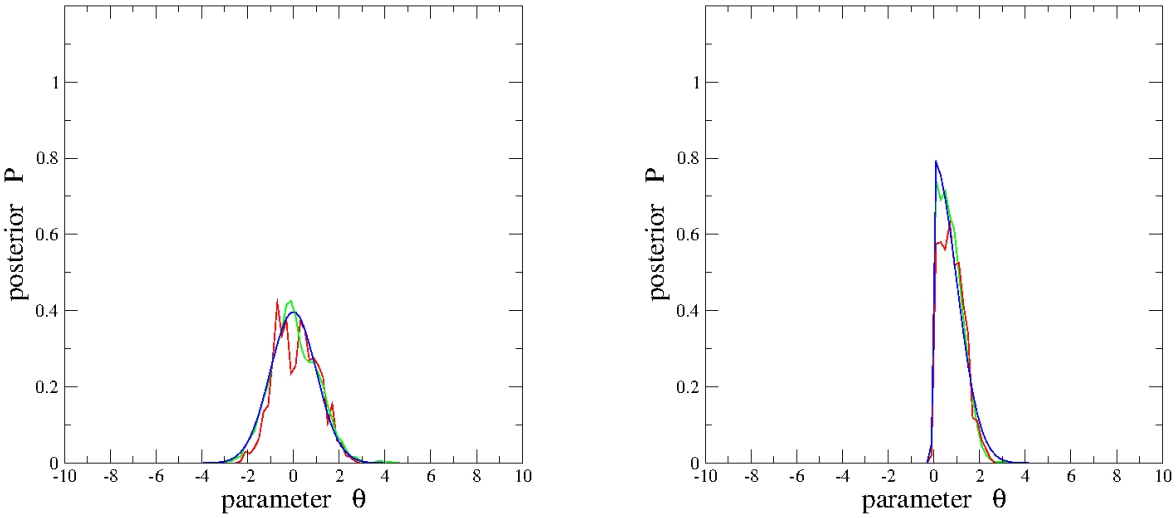

Figure 4.3: Output per la posterior probability dell'esempio unidimensionale gaussiano con flat (sinistra) e gaussian (per valori negativi, destra) prior, come descritto nel codice. La distribuzione viene sempre meglio approssimata al crescere del numero dei campionamenti (il numero dei punti della catena cresce andando da rosso a verde a blu).