Bollettino
Unione Matematica Italiana

Sezione A – La Matematica nella Società e nella Cultura

Antonio Pievatolo

Simulazione mediante processi stocastici per la probabilità applicata e per la statistica

Bollettino dell'Unione Matematica Italiana, Serie 8, Vol. **5-A**—*La Matematica nella Società e nella Cultura* (2002), n.1, p. 143–162.

Unione Matematica Italiana

<http://www.bdim.eu/item?id=BUMI_2002_8_5A_1_143_0>

Bollettino dell'Unione Matematica Italiana, Unione Matematica Italiana, 2002.

Bollettino U. M. I.
La Matematica nella Società e nella Cultura
Serie VIII, Vol. V-A, Aprile 2002, 143-162

Simulazione mediante processi stocastici per la probabilità applicata e per la statistica.

ANTONIO PIEVATOLO

Nella prefazione a [6] si legge: *Nel formulare un modello stocastico per descrivere un fenomeno reale, nel passato si soleva fare un compromesso tra un modello che fosse una copia della realtà e uno che fosse analiticamente trattabile. In altre parole, non pareva esserci alcun vantaggio nello scegliere un modello che si conformasse fedelmente al fenomeno in esame se non era possibile analizzarlo con strumenti matematici. [...] Tuttavia, l'avvento relativamente recente di risorse di calcolo veloci e poco costose ha aperto la strada ad un altro approccio: cercare di formulare un modello il più possibile fedele del fenomeno e fare affidamento su uno studio di simulazione per analizzarlo.*

I più semplici modelli stocastici sono relativi a fenomeni che si manifestano sotto forma di un numero reale e la cui legge non varia al passare del tempo. Appartengono a questa categoria gli esperimenti ripetibili in identiche condizioni e in modo indipendente uno dall'altro, come l'estrazione del primo numero su una ruota del lotto, il lancio di un dado, le prove di laboratorio nelle quali si possa assumere ragionevolmente che tutte le condizioni sperimentali siano sotto controllo. Il numero reale che costituisce la manifestazione del fenomeno, essendo ignoto a priori, viene chiamato *variabile aleatoria*; la funzione che assegna a ciascun intervallo immaginabile la probabilità che la variabile aleatoria vi appartenga è detta *distribuzione di probabilità*.

Se una variabile aleatoria x può assumere tutti i valori di un intervallo $[a, b]$, la probabilità che appartenga ad un intervallo infinitesimo dx è approssimata da $f(x)\,dx$, dove la *densità di probabilità* f è una funzione non negativa e definita sull'intero asse reale (e nulla al di fuori di $[a, b]$) a integrale 1. Ipotizzata una forma per f, la manife-

stazione del fenomeno aleatorio è frequentemente sintetizzata attraverso il suo *valore atteso*:

$$\mathrm{E}\, x = \int_{-\infty}^{+\infty} x f(x)\, dx\,,$$

cioè una «somma» dei valori assumibili da x «pesati» con la probabilità associata a ciascuno di essi. (La lettera E è l'iniziale della parola *expectation*). L'incertezza dovuta all'aleatorietà del risultato è quantificata esaurientemente da f, ma è spesso riassunta nella *varianza*:

$$\mathrm{var}\, x = \int_{-\infty}^{+\infty} (x - \mathrm{E}\, x)^2 f(x)\, dx\,,$$

cioè il valore atteso del quadrato degli scarti (anch'essi aleatori) dal valore atteso.

Vi sono situazioni nelle quali il valore atteso, la varianza, o altre sintesi della distribuzione di probabilità non sono valutabili analiticamente. Questo avviene specialmente quando, come abbiamo rilevato sopra, non si vuole rinunciare a complicare il modello per ottenere una rappresentazione il più fedele possibile della realtà. È qui che i metodi di simulazione, o di Montecarlo, trovano applicazione. La simulazione delle variabili aleatorie di più comune utilizzo (come quelle con densità gaussiana) è stata trattata esaurientemente nel passato [4]. Al giorno d'oggi, sono studiati prevalentemente problemi multidimensionali (x è un vettore aleatorio) o nei quali si analizzino contemporaneamente fenomeni tra loro dipendenti.

Per questa ragione, sulle fondamenta ormai ben consolidate della simulazione di variabili aleatorie più comuni, in questi ultimi dieci anni si sono venuti affermando in statistica i metodi *Markov chain Monte Carlo* (in breve, MCMC), grazie ai quali è diventato possibile trattare anche modelli stocastici estremamente complessi. Il problema maggiormente affrontato (noto come integrazione Montecarlo) è quello di valutare, tramite simulazione, la media di una funzione di un vettore aleatorio x, che indicheremo con $\mathrm{E}\, h(x)$. Inoltre, i metodi di simulazione vengono applicati con successo a problemi di ottimizzazione.

Il «metodo di Montecarlo» per antonomasia, valuta $\mathrm{E}\, h(x)$ con un

esperimento simulato costituito da prove indipendenti; nei metodi MCMC, le prove sono invece dipendenti. Facciamo l'esempio di un sistema a due stati, indicati con 0 e 1, osservato ad intervalli regolari. Si indichi con x_n il dato osservato all'istante n e si suppongano note e costanti nel tempo le probabilità di transizione da uno stato all'altro p_{ij}. Poniamo quindi $p_{01} = \alpha$ e $p_{10} = \beta$; allora $p_{00} = 1 - \alpha$ e $p_{11} = 1 - \beta$.

Iniziamo col determinare lo stato stazionario del sistema, cioè il suo comportamento per $n \to \infty$, esaminando per prima cosa $p_{11}^{(n)}$, la probabilità di trovarsi nello stato 1 all'istante n essendo da questo partiti. Allo stato 1 nell'istante n si può arrivare in due modi, partendo da 0 oppure da 1 all'istante $n-1$, quindi:

$$p_{11}^{(n)} = p_{10}^{(n-1)} p_{01} + p_{11}^{(n-1)} p_{11} =$$

$$p_{10}^{(n-1)} \alpha + p_{11}^{(n-1)} (1 - \beta) = (1 - \alpha - \beta)\, p_{11}^{(n-1)} + \alpha\,.$$

Con la condizione iniziale $p_{11}^{(0)} = 1$ (lo stato iniziale è 1), la soluzione dell'equazione ricorsiva è:

$$p_{11}^{(n)} = \frac{\alpha}{\alpha + \beta} + \frac{\beta}{\alpha + \beta} (1 - \alpha - \beta)^n\,,$$

il cui limite per $n \to \infty$ è $\alpha/(\alpha + \beta)$. Lo stesso valore limite si ottiene per $p_{01}^{(n)}$, quindi poniamo $p_1^{(\infty)} = \alpha/(\alpha + \beta)$ e $p_0^{(\infty)} = \beta/(\alpha + \beta)$.

Supponiamo ora di voler determinare la media di $h(x)$ dove $\Pr(x = 0) = p_0^{(\infty)}$ e $\Pr(x = 1) = p_1^{(\infty)}$, cioè:

$$\mathrm{E}\, h(x) = h(0)\, p_0^{(\infty)} + h(1) p_1^{(\infty)} = h(0)\, \frac{\beta}{\alpha + \beta} + h(1)\, \frac{\alpha}{\alpha + \beta}$$

e di non essere in grado di fare il calcolo direttamente (ovviamente non è questo il caso). Generata col calcolatore una successione di n variabili aleatorie uniformi e indipendenti [1] $\{u_i\}_{1 \leq i \leq n}$ nell'intervallo $(0, 1)$, possiamo servircene in due modi.

[1] *Uniforme in un intervallo*: detto di variabile aleatoria la cui probabilità di assumere valori in un qualsiasi sottointervallo infinitesimo dx è proporzionale a dx. *Indipendenti*: detto di un insieme di variabili aleatorie per le quali la conoscenza del valore assunto da un qualsiasi numero di esse non dà informazioni sul valore assunto dalle restanti.

– Metodo di Montecarlo: se $u_i \leq \beta/(\alpha+\beta)$, poniamo $x_i = 0$ e $x_i = 1$ altrimenti, per $i = 1, \ldots, n$, e calcoliamo la media campionaria

$$\overline{h}_n = \frac{1}{n} \sum_{i=1}^{n} h(x_i)$$

come approssimazione di $\mathrm{E}\, h(x)$; se n è elevato, per la legge dei grandi numeri, ci si può attendere che la frazione di uni nel campione sia molto vicina a $\alpha/(\alpha+\beta)$.

– Metodo MCMC: fissato un valore iniziale x_0 (per esempio $x_0 = 0$), generiamo una *catena di Markov*, cioè una successione di valori *dipendenti* $\{x_i\}_{1 \leq i \leq n}$ servendoci delle probabilità di transizione precedentemente introdotte:

$$x_{i+1} = \mathbb{I}_{(0,\,\alpha)}(u_i) \quad \text{se } \mathrm{x_i} = 0$$

$$x_{i+1} = \mathbb{I}_{(0,\,1-\beta)}(u_i) \quad \text{se } x_i = 1\,,$$

dove $\mathbb{I}_{(a,\,b)}$ è la funzione indicatrice dell'intervallo (a, b). Calcoliamo quindi $\overline{h}_n$ come sopra. Se lo stato corrente al passo i è 0, la probabilità α governa la transizione allo stato successivo, mentre se lo stato corrente è 1 la transizione è governata da β.

Poiché, come abbiamo visto, le probabilità di trovarsi nello stato 1 al passo i tendono a $\alpha/(\alpha+\beta)$, ci aspettiamo che la frequenza osservata di uni nella simulazione sia molto vicina a questo valore e che $\overline{h}_n$ sia un'approssimazione accurata della media. In figura 1 abbiamo tracciato l'evoluzione di $\overline{h}_n$ con $h(x) = x$ in una simulazione di 1000 iterazioni, ponendo $\alpha = 3/4$ e $\beta = 1/4$ ($\alpha/(\alpha+\beta) = 3/4$).

Confrontando i due algoritmi, si può osservare che la dipendenza tra i valori generati del secondo è il prezzo pagato per avere un metodo più semplice per la generazione di ciascuna osservazione, infatti nel primo caso occorre calcolare $\alpha/(\alpha+\beta)$, mentre nel secondo si usano α o β separatamente. Avremmo potuto stimare $\mathrm{E}\, h(x)$ con un'altra catena di Markov con la medesima *distribuzione di equilibrio* $(p_0^{(\infty)}, p_1^{(\infty)})$; nel seguito vedremo come scegliere questa catena che non è l'oggetto di interesse, ma è *solamente uno strumento* per ottenere la distribuzione di equilibrio voluta e calcolare medie come $\mathrm{E}\, h(x)$ in casi complicati, come quando x è multidimensionale.

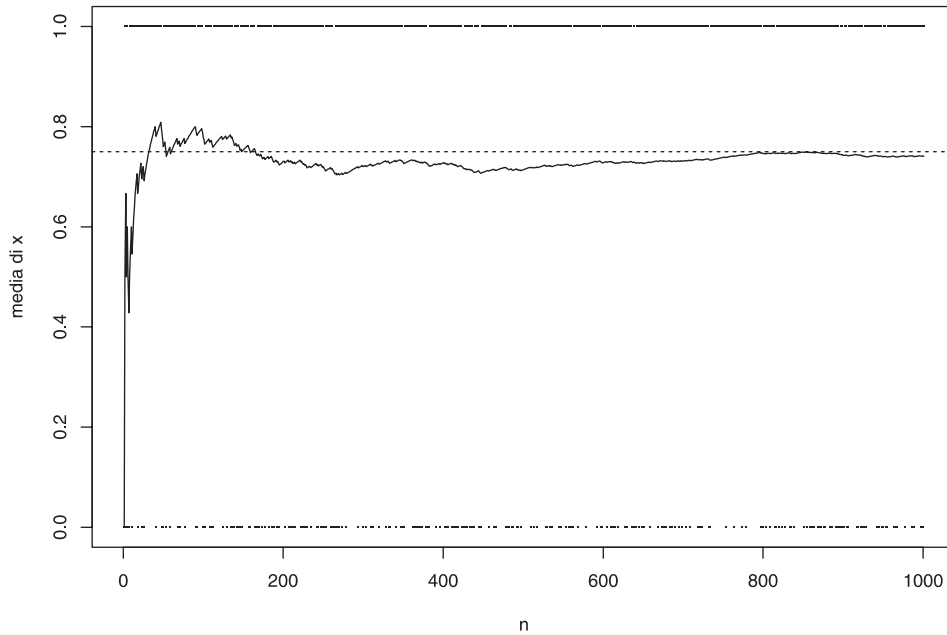

Figura 1. – Tracciato di $\overline{h}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ dalla catena di Markov a due stati (linea continua) e stati successivamente occupati dalla catena (puntini) quando $\alpha/(\alpha+\beta)=3/4$.

Nella sezione 2 ci serviamo del famoso modello di Ising per introdurre i metodi MCMC sotto le sembianze dell'algoritmo di Metropolis-Hastings. Il motivo della scelta è duplice: l'algoritmo di Metropolis è apparso per la prima volta per risolvere problemi relativi a questo modello, che costituisce anche un buon esempio di come difficoltà legate alla simulazione di variabili di dimensionalità elevata possano essere superate grazie alla dipendenza insita nei metodi MCMC.

Nella sezione 3 trattiamo di un caso particolare dell'algoritmo, detto *Gibbs sampler*, ormai diffusissimo tra gli statistici, grazie anche alla possibilità di realizzarlo quasi automaticamente. Ne approfittiamo per accennare al problema dell'inferenza bayesiana, dove questo algoritmo trova frequentemente applicazione.

Nella sezione 4 diamo qualche indicazione sull'entità dell'errore commesso nel valutare $\mathrm{E}\, h(x)$ con lo strumento MCMC. A causa della simulazione, il problema di integrazione viene infatti trasfor-

mato in un problema di stima, e ogni buona analisi statistica, oltre alla stima, deve indicare anche l'errore ad essa associato. Mentre è banale verificare che per campioni di variabili aleatorie indipendenti questo errore è $O(1/\sqrt{n})$, vedremo che l'abbandono dell'ipotesi di indipendenza ne rende più difficile la valutazione.

Concludiamo illustrando una tra le molte possibili applicazioni dell'integrazione Montecarlo ad un problema non banale di affidabilità e con un esempio di ottimizzazione stocastica.

2. – Le origini.

Un modello paradigmatico della meccanica statistica è anche un ottimo strumento per mostrare l'efficacia dei metodi MCMC: si tratta del modello di Ising, che prende il nome dal fisico tedesco E. Ising, che nel 1925 lo utilizzò per spiegare alcuni comportamenti osservati nella magnetizzazione dei materiali.

Una versione del modello di Ising in due dimensioni prevede la presenza di m^2 dipoli collocati sugli incroci tra le maglie di un reticolo quadrato (siti) di dimensione $m \times m$. Ciascun dipolo può disporsi in due versi, indicati rispettivamente con -1 e 1; l'insieme $x = \{x_s\}$, dove $s \in \{(i, j),\ i, j = 1, \ldots, m\}$, contiene gli orientamenti dei dipoli collocati sugli m^2 siti.

In assenza di un campo magnetico esterno, l'energia del sistema è rappresentata dalla funzione

$$H(x) = -\beta \sum_{\substack{s,\,t \\ s \sim t}} x_s x_t, \qquad \beta > 0\,, \tag{1}$$

dove la somma è estesa alle coppie di siti adiacenti ($s \sim t$) in direzione orizzontale o verticale. Ad ogni configurazione dei dipoli, viene associata una probabilità tramite la seguente distribuzione discreta:

$$P(x) = \frac{\exp\{-H(x)\}}{Z(\beta)}\,, \qquad Z(\beta) = \sum_x \exp\{-H(x)\}\,. \tag{2}$$

La probabilità $P(x)$ favorisce configurazioni a bassa energia, che, in questa formulazione del modello, sono quelle in cui dipoli vicini hanno lo stesso orientamento. Si osservi inoltre che le caratteristiche macrosco-

piche del sistema sono determinate dall'insieme delle relazioni microscopiche tra le coppie di dipoli vicini, come appare chiaro dalla (1).

Le configurazioni tipiche del sistema possono essere visualizzate attraverso una serie di simulazioni dalla distribuzione $P(x)$. Poiché quelle ammissibili sono 2^{m^2}, non è possibile, se non per m molto ridotto, preparare una tabella contenente le probabilità di ciascuna configurazione e procedere alla simulazione Montecarlo diretta. Il metodo di Metropolis, nella sua versione «una componente alla volta» ([5], sez. 5.2.3 e 7.1.4), considera invece un sito per volta assieme ai soli siti ad esso vicini, generando una successione $\{x^{(n)}\}$ di vettori aleatori dipendenti. Per semplificare la notazione, associamo ad ogni sito s un intero $i = i(s)$, $i = 1, \ldots, m^2$, in modo che possiamo scrivere $x = (x_1, \ldots, x_{m^2})$. Se $x^{(n-1)}$ è il vettore simulato al passo $n-1$, per il sito con indice $i = 1$ si propone un nuovo orientamento y_1 per il dipolo ad esso associato, scegliendolo a caso tra -1 e 1; si calcola quindi la differenza di energia ΔH_1 tra la configurazione $x' = (y_1, x_2^{(n-1)}, \ldots, x_{m^2}^{(n-1)})$ e quella corrente $x^{(n-1)}$; estratto infine un numero u da una distribuzione uniforme sull'intervallo $(0, 1)$, se $u < \exp\{-\Delta H_1\} = P(x')/P(x^{(n-1)})$, si pone $x_1^{(n)} = y_1$, altrimenti $x_1^{(n)} = x_1^{(n-1)}$. Queste operazioni vanno ripetute per ogni sito tenendo conto delle modifiche precedenti, cosicché, se y_i è il nuovo orientamento proposto per il sito i,

$$
\begin{aligned}
\Delta H_i = \quad & H((x_1^{(n)}, \ldots, x_{i-1}^{(n)}, y_i, x_{i+1}^{(n-1)}, \ldots, x_{m^2}^{(n-1)})) \\
& - H((x_1^{(n)}, \ldots, x_{i-1}^{(n)}, x_i^{(n-1)}, x_{i+1}^{(n-1)}, \ldots, x_{m^2}^{(n-1)})) \\
= & -\beta \sum_{\substack{j<i \\ j\sim i}} x_j^{(n)}(y_i - x_i^{(n-1)}) - \beta \sum_{\substack{j>i \\ j\sim i}} x_j^{(n-1)}(y_i - x_i^{(n-1)}),
\end{aligned} \tag{3}
$$

dove le somme sono fatte rispetto all'indice j e per i soli siti adiacenti al sito i interessato dalla modifica. Una volta visitati tutti i siti, il passo n dell'algoritmo è terminato e la nuova configurazione è $x^{(n)}$ e si procede oltre finché un qualche indicatore di interesse, come la frazione di dipoli con orientamento -1, non si sia stabilizzato.

La procedura basata sulla (3) implica che una nuova configurazione viene sempre accettata se porta ad una diminuzione dell'energia, cioè ad un aumento della probabilità ad essa associata, mentre viene

rifiutata tanto più facilmente quanto più l'incremento dell'energia è consistente. La visita a ciascun sito richiede di considerare solo i valori delle variabili associate ai siti adiacenti, una caratteristica decisiva, che beneficia della particolare struttura dell'energia (1).

La successione $\{x^{(n)}\}$ è una catena di Markov omogenea ([2]) con distribuzione di equilibrio data da P e, quando n è sufficientemente elevato, $x^{(n)}$ può essere considerato come un campione da P. L'algoritmo di Metropolis è stato generalizzato da Hastings ([5], cap. 6), che ha mostrato che la convergenza alla distribuzione di equilibrio si mantiene anche quando, nella transizione da $x^{(n-1)}$ a $x^{(n)}$, y_i viene estratto da una qualunque distribuzione $q(y_i)$, a patto che il nuovo valore per il sito i sia accettato se

$$(4) \quad u \leqslant \frac{q(x_i^{(n-1)})\, P((x_1^{(n)}, \ldots, x_{i-1}^{(n)}, y_i, x_{i+1}^{(n-1)}, \ldots, x_{m^2}^{(n-1)}))}{q(y_i)\, P((x_1^{(n)}, \ldots, x_{i-1}^{(n)}, x_i^{(n-1)}, x_{i+1}^{(n-1)}, \ldots, x_{m^2}^{(n-1)}))}.$$

È facile vedere che, se $q(-1) = q(1) = 1/2$, si riottiene l'algoritmo di Metropolis.

In figura 2 mostriamo l'evoluzione della catena per $m = 256$ a partire dallo stato iniziale di eguale orientamento -1 di tutti i dipoli, fino allo stato $x^{(n)}$ che si ottiene dopo $n = 100$ passi, per $\beta = 1/2{,}269$ ([3]). Visivamente, i riquadri nella seconda riga differiscono meno tra di loro di quelli nella prima, indicando una progressiva stabilizzazione della successione $\{x^{(n)}\}$.

Questa successione, come nell'esempio dell'introduzione, è utilizzabile per approssimare la media di funzioni di x:

$$(5) \qquad \mathrm{E}\, h(x) = \sum_x h(x)\, P(x) \simeq \frac{1}{n} \sum_{i=0}^{n-1} h(x^{(i)}) = \overline{h}_n,$$

dove h può ad esempio essere la frazione di dipoli con orientamento -1. L'espressione esatta di $\mathrm{E}\, h(x)$ richiede di sommare 2^{m^2} termini,

([2]) La legge di transizione da $x^{(n-1)}$ a $x^{(n)}$ è sempre la medesima indipendentemente da n.

([3]) Questo è il valore cui corrisponde la cosiddetta *transizione di fase*: quando β supera questo valore, dipoli adiacenti tendono ad assumere lo stesso orientamento e si osservano grandi aree omogenee; quando β decresce, i dipoli tendono invece a comportarsi indipendentemente, facendo assumere al sistema un aspetto caotico.

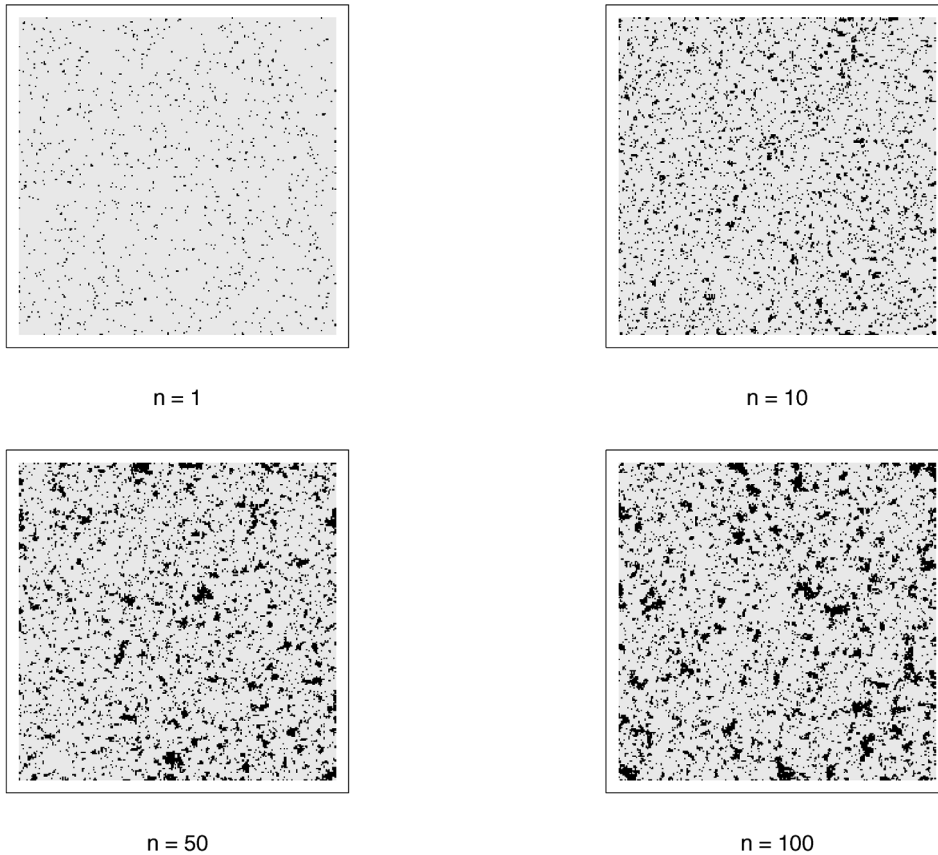

Figura 2. – Simulazione del modello di Ising su un reticolo di dimensioni 256×256 a partire dallo stato di uguale orientamento -1 (colore bianco) per tutti i dipoli, dopo 1, 10, 50 e 100 passi.

mentre se ne sommano n per ottenere l'approssimazione Montecarlo, dove il valore di n richiesto per rendere trascurabile l'errore di approssimazione dovuto alla simulazione può essere notevolmente inferiore, come vedremo nella sezione 4.

3. – MCMC nella statistica bayesiana.

Nella sezione precedente, per mezzo del modello di Ising, abbiamo visto che la caratteristica principale dei metodi MCMC è di generare

un campione di variabili dipendenti da una distribuzione su uno spazio di dimensione anche molto elevata utilizzando, per un singolo passo, una minima parte dell'informazione contenuta nella distribuzione stessa. Inoltre, per ottenere stime affidabili di funzioni di variabili aleatorie, è sufficiente che la catena di Markov simulata visiti solo una porzione dello spazio degli stati, allo stesso modo in cui, in statistica, si ottengono stime precise dei parametri incogniti di modelli probabilistici da un campione di dimensione relativamente ridotta. Queste caratteristiche hanno favorito, a partire dalla fine degli anni '80 (quando risorse di calcolo consistenti si sono rese disponibili a basso costo), una notevole diffusione di questi metodi in statistica, poiché è diventato facile fare uso di modelli complessi su spazi di dimensione elevata e risolvere i conseguenti problemi di integrazione anche quando la strada dei metodi di Montecarlo tradizionali (come il campionamento per importanza, [5] sez. 3.3) non sia percorribile.

Questo è stato particolarmente vero per l'inferenza bayesiana. Il tipico modello bayesiano parametrico interpreta i dati y, provenienti da un esperimento o dall'osservazione di un fenomeno, alla luce di un modello appartenente ad una famiglia parametrica di densità di probabilità $\{f(y;\,\theta);\,\theta \in \Theta\}$, dove Θ è uno spazio euclideo di dimensione k. Per fare inferenza sul parametro non osservato θ, si sintetizzano le informazioni iniziali su di esso tramite una densità di probabilità *a priori* $p(\theta)$, e si studia come queste si modificano, in seguito all'osservazione di y, attraverso la densità *a posteriori*

$$p(\theta \mid y) = \frac{f(y;\,\theta)\,p(\theta)}{Z(y)}\,, \qquad Z(y) = \int_{\Theta} f(y;\,\theta)\,p(\theta)\,d\theta\,. \tag{6}$$

La (6) è conosciuta come *formula di Bayes* e consente di valutare la distribuzione di probabilità di una quantità non osservata (cioè θ) a partire dall'osservazione di una seconda quantità da essa dipendente (cioè y).

Il parametro θ, per esempio, può rappresentare la probabilità di ottenere testa lanciando una determinata moneta. Chiunque si aspetterebbe che questa probabilità sia attorno a $1/2$, quindi $p(\theta)$ sarà una densità di probabilità sull'intervallo $(0,\,1)$ con media $1/2$, come la densità uniforme $p(\theta) = \mathbb{I}_{(0,\,1)}(\theta)$. Data una successione di m

lanci $y = (y_1, \ldots, y_m)$ ($y_i = 1$ se il lancio è testa, 0 altrimenti), se θ fosse noto, la probabilità di osservare proprio y sarebbe

$$f(y;\, \theta) = \prod_{i=1}^{m} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{\sum_{i=1}^{m} y_i}(1-\theta)^{m-\sum_{i=1}^{m} y_i}.$$

Infatti, se θ è noto, l'esito di un lancio non dà informazioni sui successivi, quindi i lanci possono essere considerati indipendenti per l'osservatore. Per la formula di Bayes, la densità a posteriori di θ, è

$$p(\theta\,|\,y) = (m+1)\binom{m}{\sum y_i}\theta^{\sum y_i}(1-\theta)^{m-\sum y_i}.$$

La nostra valutazione a priori di θ era espressa da $p(\theta)$; coerentemente, dopo aver osservato il risultato dei lanci, sarà espressa da $p(\theta\,|\,y)$, sintetizzabile con la propria media a posteriori:

$$\mathrm{E}\,[\theta\,|\,y] = \int_0^1 \theta p(\theta\,|\,y)\, d\theta = \frac{\sum y_i + 1}{m+2}\,.$$

Dall'esempio è chiaro che in problemi più complessi (specie con θ multidimensionale), si è spesso interessati a valutare la media a posteriori di una funzione $h(\theta)$:

$$\mathrm{E}(h(\theta)\,|\,y) = \int_{\Theta} h(\theta)\, p(\theta\,|\,y)\, d\theta\,,$$

che, specialmente a causa della completa libertà nello specificare $p(\theta)$, risulta quasi sempre non determinabile analiticamente. A questo punto è facile fare un'analogia tra θ nella (6) e x nella (2) e pensare di utilizzare il metodo di Metropolis-Hastings e la (5) per valutare $E(h(\theta)\,|\,y)$, costruendo una catena di Markov $\{\theta^{(n)}\}$ che abbia $p(\theta\,|\,y)$ come distribuzione di equilibrio.

La versione dell'algoritmo di Metropolis-Hastings che ne ha decretato il successo nella comunità statistica è il cosiddetto *Gibbs sampler*. Se $\theta^{(n-1)} = (\theta_1^{(n-1)}, \ldots, \theta_k^{(n-1)})$ è lo stato corrente, il nuovo stato $\theta^{(n)}$ viene raggiunto aggiornando una componente di $\theta^{(n-1)}$ per volta come per $x^{(n)}$ nel modello di Ising. Fissato un indice i, $i = 1, \ldots, k$, si propone un nuovo valore $\theta_i^{(n)}$ ponendo $q(\theta_i^{(n)})$ pari al-

la cosiddetta distribuzione completamente condizionata di θ_i:

$$p(\theta_i^{(n)} \mid \theta_1^{(n)}, \ldots, \theta_{i-1}^{(n)}, \theta_{i+1}^{(n-1)}, \ldots, \theta_k^{(n-1)}, y) \tag{7}$$

e si procede quindi all'accettazione/rifiuto secondo la regola data dalla (4), trovando che questa avviene sempre con probabilità 1. Il *Gibbs sampler* si riduce quindi al campionamento in successione dalle distribuzioni condizionate complete (7), e, come Metropolis, visita lo spazio degli stati muovendosi a turno nella direzione di ciascuna coordinata del sistema cartesiano di riferimento, come mostrato in figura 3, dove la distribuzione di equilibrio è gaussiana bidimensionale con vettore medio nullo, varianze unitarie e correlazione $\varrho = 0,7$:

$$p(\theta_1, \theta_2) = \frac{1}{2\pi\sqrt{1-\varrho^2}} \exp\left\{ -\frac{1}{2(1-\varrho^2)} (\theta_1^2 - 2\varrho\theta_1\theta_2 + \theta_2^2) \right\}. \tag{8}$$

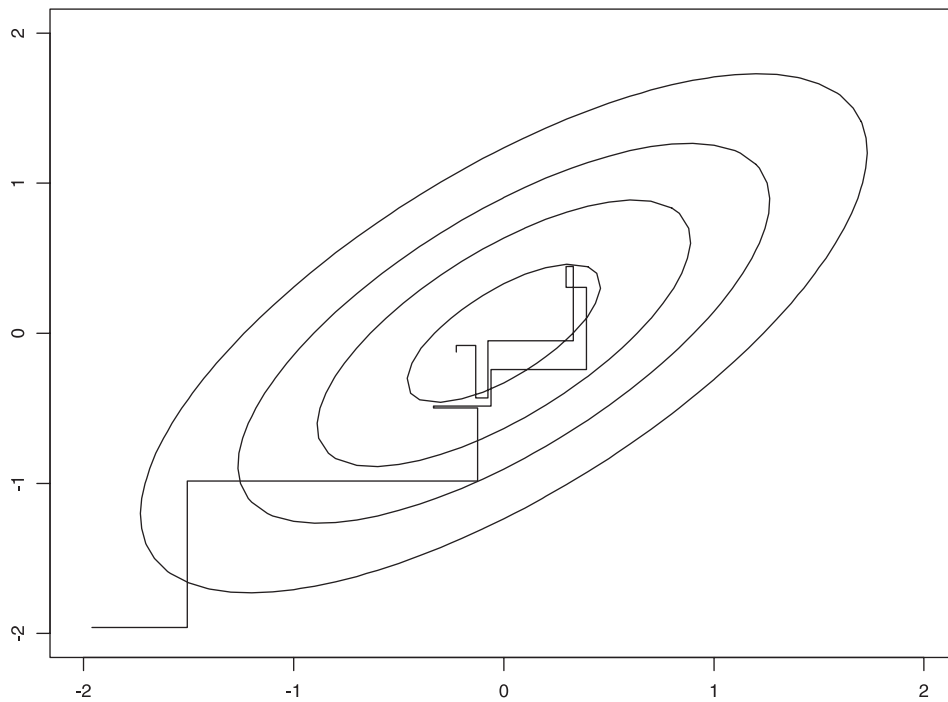

Figura 3. – Curve di livello della gaussiana bivariata (8) con $\varrho = 0,7$ e percorso di 10 passi del *Gibbs sampler* con valore iniziale $(\theta_1^{(0)}, \theta_2^{(0)}) = (-2, -2)$.

La correlazione ϱ misura la forza della dipendenza tra θ_1 e θ_2 e corrisponde al coefficiente del termine di rotazione $\theta_1 \theta_2$ nell'equazione delle ellissi di figura 3.

4. – Convergenza all'equilibrio e precisione delle stime.

Perché l'approssimazione (5) converga al vero valore atteso e la distribuzione di equilibrio sia quella voluta, è necessario che l'algoritmo di Metropolis-Hastings sia costruito con attenzione. Pensando ad uno spazio degli stati numerabile e adottando la notazione usata per il modello di Ising come notazione generale, occorre che:

1. la catena di Markov sia irriducibile, vale a dire ciascuno stato x tale che $P(x) > 0$ deve essere raggiungibile da ogni altro stato in un numero finito di passi con probabilità positiva;

2. la catena sia ricorrente, cioè ciascuno stato x tale che $P(x) > 0$ è visitato infinitamente spesso nel corso di una simulazione di durata infinita; questo consente che il rapporto delle frequenze delle visite a due stati qualunque sia limitato superiormente e inferiormente;

3. la catena sia aperiodica, cioè deve esistere almeno uno stato tale che il massimo comun divisore del numero di transizioni necessarie per ritornare in quello stato sia 1; in tal modo si escludono situazioni nelle quali la catena visita ciclicamente gli stati, impedendo di valutarne la probabilità.

Se valgono queste condizioni, diremo che catena di Markov è *ergodica.*

La struttura stessa del meccanismo di accettazione/rifiuto dell'algoritmo di Metropolis-Hastings garantisce l'ergodicità a patto che la distribuzione $q(y_i)$ per la proposta di un nuovo stato sia scelta in modo da non confinare la catena di Markov in un sottoinsieme dello spazio degli stati. Sotto queste condizioni, la media campionaria della (5) converge al vero valore atteso di una qualunque funzione h.

Una questione molto più delicata riguarda l'errore di approssimazione dovuto alla simulazione che si commette con la (5), conside-

rato che l'algoritmo non può continuare indefinitamente. Trascuriamo il fatto che esso non viene mai inizializzato «in equilibrio» (cioè $x^{(0)}$ non può essere un campione da P per ovvi motivi), assumendo di aver compiuto un numero di passi sufficiente a rendere ininfluente la nostra scelta di $x^{(0)}$; a questo punto riazzeriamo il contatore e iniziamo a raccogliere i dati simulati per calcolare $\overline{h}_n$.

Valutiamo l'errore associato ad $\overline{h}_n$ con la radice quadrata della varianza:

$$\operatorname{var} \overline{h}_n = \mathrm{E}\,[\overline{h}_n - \mathrm{E}\; h(x)]^2\,,$$

dove la media interna è calcolata rispetto alla distribuzione di equilibrio P, mentre quella esterna, a causa dei prodotti incrociati che risultano dallo sviluppo del quadrato, deve tener conto anche del meccanismo di transizione. Considerato che quest'ultimo si ripete in modo identico ad ogni passo, si ottiene

$$\begin{aligned}\operatorname{var} \overline{h}_n &= \frac{1}{n^2}\left[\sum_{i=0}^{n-1} \operatorname{var} h(x^{(i)}) + 2 \sum_{i=0}^{n-2} \sum_{j=i+1}^{n-1} \operatorname{cov}\,(h(x^{(i)}),\, h(x^{(j)}))\right] \\ &= \frac{1}{n}\gamma_0 + \frac{2}{n}\sum_{i=1}^{n-1}\left(1 - \frac{i}{n}\right)\gamma_i, \qquad \gamma_i = \operatorname{cov}\,(h(x^{(k)}),\, h(x^{(k+i)}))\,,\end{aligned} \tag{9}$$

dove

$$\operatorname{cov}\,(h(x^{(i)}),\, h(x^{(j)})) = \mathrm{E}\,(h(x^{(i)}) - \mathrm{E}\; h(x))(h(x^{(j)}) - \mathrm{E}\; h(x))\,.$$

Senza svolgere tutti i passaggi, notiamo solo che le quantità γ_i, dette *covarianze*, hanno lo stesso significato della correlazione ϱ precedentemente introdotta per la distribuzione gaussiana bivariata.

La varianza Montecarlo (9) di $\overline{h}_n$ è dunque infinitesima per $n \to \infty$ se le covarianze γ_i sono a loro volta infinitesime per $i \to \infty$, per il teorema di Cesàro sulla media delle successioni. Tuttavia questo non dà informazioni sulla velocità di convergenza. Infatti, mentre il primo addendo della (9) è $O(1/n)$, come per i metodi di Montecarlo tradizionali basati sulla simulazione di variabili aleatorie indipendenti, nulla si può dire sul secondo, dovuto alla dipendenza tra gli stati della catena di Markov. Se aggiungiamo l'ipotesi che $\sum_j \gamma_j < \infty$, si mostra facilmente che $\operatorname{var} \overline{h}_n = O(1/n)$. Con ciò, l'errore associato

a $\overline{h}_n$ non dipende dalla dimensione di x (come per i metodi di quadratura), ma solamente dalla lunghezza della simulazione. Va comunque detto che un algoritmo troppo «lento» nel visitare lo spazio degli stati (ad esempio a causa di una dipendenza troppo marcata tra $x^{(n)}$ ed $x^{(n+1)}$), potrebbe richiedere un valore di n inaccettabile per raggiungere la precisione desiderata.

La condizione $\sum_j \gamma_j < \infty$ vale sempre per una catena di Metropolis-Hastings ergodica su uno spazio degli stati finito, mentre non è semplice verificarla in generale, ed è necessario imporre vincoli più o meno restrittivi sulla distribuzione $q(y_i)$. I vincoli servono ad assicurare che la transizione tra due stati x e y, quando x è lo stato corrente, avvenga con probabilità uniformemente limitata inferiormente rispetto ad x (anche se essa richiede più di un passo). Si tratta cioè una condizione che chiameremo di «uniforme raggiungibilità» dell'intero spazio degli stati a partire da qualsiasi punto. Esistono anche condizioni meno stringenti, ma troppo complesse per essere enunciate senza ricorrere a formalismi.

L'uniforme raggiungibilità è ovvia per spazi degli stati finiti; nell'esempio dell'introduzione, la probabilità di transizione è limitata inferiormente dal minimo tra le probabilità di transizione tra le quattro coppie di stati. Nel caso numerabile o continuo, se lo spazio degli stati è il prodotto cartesiano di quelli delle singole coordinate, la verifica può essere ancora possibile; altrimenti, può capitare di doversi accontentare dell'ergodicità della catena, senza avere garanzie sul comportamento dell'errore.

5. – Integrazione e ottimizzazione.

Dalla discussione fatta finora, si evince che i metodi MCMC trovano applicazione in problemi nei quali compaiono spazi di dimensione elevata oppure quando il calcolo diretto di quantità di interesse, anche se possibile in teoria, è proibitivo in termini di tempo.

Al di là della semplice simulazione di un sistema complesso, l'integrazione Montecarlo è senza dubbio il campo dove i metodi MCMC vengono applicati più di frequente. Un esempio spesso usato

nella letteratura statistica riguarda i guasti di 10 pompe in un impianto nucleare, tenute sotto osservazione per periodi di durata diversa, come riportato nella seguente tabella:

Pompa	1	2	3	4	5	6	7	8	9	10
Num. guasti	5	1	5	14	3	19	1	1	4	22
Periodo	94,32	15,72	62,88	125,76	5,24	31,44	1,05	1,05	2,10	10,48

Nella notazione della sezione 3, si ipotizza che il numero di guasti y_i per l'i-esima pompa segua una distribuzione di Poisson con media $\theta_i t_i$, $i = 1, \ldots, 10$, dove t_i è il periodo di osservazione della pompa e θ_i è il numero atteso di guasti per unità di tempo. Allora, se le pompe funzionano indipendentemente,

$$f(y; \theta) = \prod_{i=1}^{10} \frac{1}{y_i!} e^{-\theta_i t_i} (\theta_i t_i)^{y_i}.$$

A ciascun θ_i viene attribuita una distribuzione a priori $\mathcal{G}(\alpha, \beta)$ (gamma di parametri α e β)

$$p(\theta_i; \alpha, \beta) = \frac{1}{\Gamma(\alpha)} \beta^{\alpha} \theta_i^{\alpha - 1} e^{-\beta\theta_i}, \qquad \alpha, \beta > 0$$

e i θ_i vengono considerati indipendenti, perciò

$$p(\theta; \alpha, \beta) = \prod_{i=1}^{10} p(\theta_i; \alpha, \beta).$$

Quest'impostazione è frequentemente usata per creare un modello bayesiano per i guasti di unità simili, ma con differenze individuali non osservabili: il numero atteso di guasti per unità di tempo di ciascuna pompa proviene infatti dalla medesima distribuzione a priori $\mathcal{G}(\alpha, \beta)$, ma differisce da pompa a pompa. Sul parametro incognito β viene infine imposta un'ulteriore distribuzione a priori $\mathcal{G}(\gamma, \delta)$. Posti $\alpha = 1{,}8$, $\gamma = 0{,}01$ e $\delta = 1$ ([5], sez. 7.1.6) ([4]), la distribuzione a posteriori dei para-

([4]) L'assegnazione dei valori indicati a β, γ e δ determina la distribuzione a priori dei parametri incogniti. Mentre da un punto di vista concettuale non c'è assolutamente nulla da dire al proposito, vi sono tuttora controversie su come questo dovrebbe essere fatto in pratica.

metri incogniti $p(\theta, \beta | y)$ è proporzionale, a meno della costante $Z(y)$ (si veda la (6)), a $f(y; \theta)\, p(\theta; \alpha, \beta)\, p(\beta)$. Conservando, in quest'ultima quantità, solo i termini che dipendono da θ_i, abbiamo che la distribuzione completamente condizionata di θ_i (si veda la (7)) è $\mathcal{G}(y_i + \alpha, t_i + \beta)$, $i = 1, \ldots, 10$; facendo lo stesso per β, si ottiene che la sua distribuzione completamente condizionata è $\mathcal{G}(\gamma + 10\alpha, \delta + \sum_i \theta_i)$. Siccome esistono metodi standard per simulare direttamente dalla distribuzione gamma ([5], cap. 2), possiamo realizzare un *Gibbs sampler* che generi una catena $(\theta^{(n)}, \beta^{(n)})$ con distribuzione di equilibrio $p(\theta, \beta | y)$. Dalla catena potremo poi valutare, per esempio, il numero atteso a posteriori di guasti delle 10 pompe nell'unità di tempo:

$$\mathrm{E}\left(\sum_{i=1}^{10} \theta_i \,|\, y\right) = \mathrm{E}\left(h(\theta, \beta) \,|\, y\right) \simeq \frac{1}{n} \sum_{i=0}^{n-1} h(\theta^{(i)}, \beta^{(i)}).$$

In [5], sez. 7.1.6, viene mostrato che per questo *Gibbs sampler* vale l'uniforme raggiungibilità dello spazio degli stati, della quale si è parlato nella sezione precedente, e dunque l'errore Montecarlo nella stima della media è $O(1/\sqrt{n})$.

Utilizzando il programma BUGS [7] ([5]), abbiamo trovato che la media a posteriori del numero di guasti per le dieci pompe nell'unità di tempo è 6,465. (La corrispondente quantità osservata è $\sum y_i / t_i = 7{,}393$). L'errore Montecarlo (stimato) dopo 1000 passi del *Gibbs sampler* è pari a 0,05422, mentre con 2000 passi scende a 0,03859. Il rapporto tra il secondo e il primo errore è 0,71173, un valore molto prossimo a $\sqrt{1000/2000} = 0{,}70711$.

I metodi MCMC possono anche risolvere problemi di ottimizzazione. Per vedere come ciò sia possibile, consideriamo ancora uno spazio discreto finito e il problema di minimizzare una funzione a valori reali $H(x)$. In modo equivalente, possiamo invece massimizzare $\exp\{-H(x)\}$, cioè trovare i punti di massimo (le mode) della distri-

([5]) Liberamente reperibile nella rete telematica su http://www.mrc-bsu.cam.ac.uk/bugs

buzione di probabilità discreta

$$P(x) = \frac{\exp\{-H(x)\}}{\sum_x \exp\{-H(x)\}}.$$

La trasformazione $T \mapsto P^{\frac{1}{T}}(x) = P_T(x)$, dove T è chiamata «temperatura», produce una distribuzione discreta uniforme sui punti di massimo globale di $P(x)$, quando $T \to 0$. Infatti, supponendo che i punti di massimo globale siano M, se x^* è uno di questi, possiamo scrivere:

$$P_T(x) = \frac{\exp\left\{-\frac{1}{T}(H(x) - H(x^*))\right\}}{\sum_x \exp\left\{-\frac{1}{T}(H(x) - H(x^*))\right\}}; \tag{10}$$

dunque, quando $T \to 0$, $P_T(x) \to 0$ se x non è un punto di massimo globale, mentre $P_T(x) \to 1/M$ altrimenti.

L'idea dell'algoritmo detto *simulated annealing* è di generare una catena di Markov la cui distribuzione di equilibrio sia appunto la distribuzione uniforme sull'insieme dei punti di massimo. Questo si ottiene simulando una catena di Markov non omogenea: stabilita una successione decrescente $\{T_n\}$ per T, si esegue l'algoritmo di Metropolis-Hastings impiegando nella (4) P_{T_n} anziché P per l'intero passo n. In tal modo, se la temperatura viene ridotta abbastanza lentamente, al crescere di n, la catena di Markov si ritroverà a visitare soltanto i punti di massimo, visto che le probabilità degli altri stati saranno divenute trascurabili. La riduzione di T deve avvenire gradualmente per evitare di rimanere bloccati in un punto di massimo locale, il che è garantito se $T_n = D/\log n$. La costante D dipende dall'entità dei dislivelli della superficie di P (si veda [5], sez. 5.2.3) e di solito non è calcolabile; inoltre, la riduzione di T su scala logaritmica è troppo graduale e richiede una simulazione troppo lunga. Per questi motivi, in pratica si procede empiricamente usando un tasso geometrico, con $T_n = r^n T_0$, $0 < r < 1$, dove r e T_0 vengono scelte in modo che la superficie di P_{T_1}, da utilizzare nel primo passo dell'algoritmo

di Metropolis, sia sufficientemente «piatta» da permettere di raggiungere qualunque stato con probabilità non trascurabile.

Servirsi di un metodo approssimato per l'ottimizzazione in uno spazio finito può sembrare artificioso, tuttavia la dimensione enorme di taluni spazi può rendere vantaggioso quest'approccio. Un caso è il modello di Ising su reticoli di grandi dimensioni con un campo magnetico esterno, dove l'energia è data da,

$$H(x) = -\alpha \sum_{s} x_s - \beta \sum_{\substack{s,t \\ s \sim t}} x_s x_t, \qquad \beta > 0\ ,\ \alpha \neq 0$$

per il quale può essere interessante trovare la configurazione più probabile (cioè quella con energia minima).

6. – Commento conclusivo.

I metodi MCMC sono uno strumento molto potente per la simulazione di modelli stocastici complessi ed il calcolo di quantità ad essi legate, ma, a parte i casi più semplici, richiedono di essere ben calibrati per produrre risultati affidabili su spazi degli stati numerabili o continui. In particolare, non è facile valutare l'errore di approssimazione dovuto alla simulazione (come si intuisce dalla sezione 4) in funzione della sua lunghezza e quindi non è sempre chiaro quanti passi sono necessari per far scendere l'errore sotto la soglia desiderata. Sono stati quindi sviluppati numerosi strumenti che, con l'ausilio di rappresentazioni grafiche e di verifiche basate su proprietà formali delle catene di Markov generate, aiutano a compiere delle scelte, ma sempre con un margine d'errore. Come per l'inferenza statistica, si tratta infatti di stabilire quanto distanti si è dall'obiettivo attraverso un campione comunque non «esaustivo».

Per l'approfondimento di questi aspetti, tuttora oggetto di ricerca, rimandiamo il lettore a [5], cap. 8. Per maggiori informazioni sui metodi MCMC in generale, suggeriamo infine di visitare il sito internet www.statslab.cam.ac.uk/~mcmc/.

Il lettore interessato all'impostazione bayesiana all'inferenza statistica, può invece consultare, in ordine crescente di difficoltà, [3], [2] e [1].

REFERENCES

[1] J. M. BERNARDO - A. F. M. SMITH, *Bayesian Theory*, Wiley, New York (1994).
[2] G. E. P. BOX - G. C. TIAO, *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, Massachusetts (1973).
[3] F. CAROTI GHELLI, *Statistica Bayesiana*, F. Angeli, Milano (1978).
[4] L. DEVROYE, *Non-uniform random variate generation*, Springer-Verlag, New York (1985).
[5] C. P. ROBERT - G. CASELLA, *Monte Carlo statistical methods*, Springer-Verlag, New York (1999).
[6] S. M. ROSS, *Simulation*. Academic Press, San Diego (1997).
[7] D. J. SPIEGELHALTER, *et alias*, *BUGS: Bayesian Inference Using Gibbs Sampling, Version 1.3 for MS Windows*, MRC Biostatistics Unit, Cambridge (2000).

Antonio Pievatolo, Consiglio Nazionale delle Ricerche, Istituto per le Applicazioni della Matematica e dell'Informatica, via Ampère 56, 20131 Milano