Bollettino
Unione Matematica Italiana

Sezione A – La Matematica nella Società e nella Cultura

Laura M. Sangalli

Alcune misure di probabilità aleatorie e loro applicazioni in statistica bayesiana

Bollettino dell'Unione Matematica Italiana, Serie 8, Vol. **10-A**—*La Matematica nella Società e nella Cultura* (2007), n.2, p. 339–342.

Unione Matematica Italiana

<http://www.bdim.eu/item?id=BUMI_2007_8_10A_2_339_0>

L'utilizzo e la stampa di questo documento digitale è consentito liberamente per motivi di ricerca e studio. Non è consentito l'utilizzo dello stesso per motivi commerciali. Tutte le copie di questo documento devono riportare questo avvertimento.

Articolo digitalizzato nel quadro del programma
bdim (Biblioteca Digitale Italiana di Matematica)
SIMAI & UMI
http://www.bdim.eu/

Bollettino dell'Unione Matematica Italiana, Unione Matematica Italiana, 2007.

Bollettino U. M. I.
La Matematica nella Società e nella Cultura
Serie VIII, Vol. X-A, Agosto 2007, 339-342

Alcune misure di probabilità aleatorie e loro applicazioni in statistica bayesiana

Laura M. Sangalli

Nella tesi vengono considerate due classi di misure di probabilità aleatorie e loro applicazioni in statistica bayesiana.

Nella prima parte della tesi sono considerate probabilità aleatorie che costituiscono una generalizzazione delle misure aleatorie con incrementi indipendenti normalizzate, studiate ad esempio in [3]. Le successioni di variabili aleatorie indipendenti e identicamente distribuite (i.i.d.), aventi per comune distribuzione una misura aleatoria nella classe definita, costituiscono modelli per il campionamento delle specie. Vengono derivate espressioni esplicite per le distribuzioni finito-dimensionali e per le distribuzioni predittive delle successioni di variabili aleatorie.

Nella seconda parte della tesi sono considerate probabilità aleatorie definite tramite opportune funzioni di processi di diffusione. Le successioni di variabili aleatorie i.i.d., aventi per comune distribuzione una misura aleatoria in tale classe, costituiscono modelli per l'analisi della sopravvivenza in cui il tasso di mortalità è determinato dal processo di diffusione. Viene mostrato come questi modelli possano essere efficientemente trattati tramite tecniche Markov chain Monte Carlo (McMC).

1. – Misure aleatorie con incrementi indipendenti normalizzate condizionate e modelli per il campionamento delle specie.

Sia μ una misura aleatoria con incrementi indipendenti (RMI) su R, ovvero tale che, per ogni n ed ogni famiglia misurabile $\{A_1, \ldots, A_n\}$ di sottoinsiemi di R a due disgiunti, le variabili aleatorie $\mu(A_1), \ldots, \mu(A_n)$ sono stocasticamente indipendenti. Se la massa totale $T := \mu(\mathrm{R})$ della RMI è quasi certamente finita e strettamente positiva, normalizzando la RMI si ottiene una misura di probabilità aleatoria, $\varphi = \mu/\mu(\mathrm{R})$, detta misura aleatoria con incrementi indipendenti normalizzata (NRMI). Tali probabilità aleatorie, di cui il processo di Dirichlet è un esempio notevole, sono state studiate ad esempio in [3]. Sia ora h una funzione misurabile non-negativa su R^+ tale che $\int_{\mathrm{R}^+} h(\sigma)\, Q\{\mu(\mathrm{R}) \in \mathrm{d}\sigma\} = 1$, dove Q è la legge di μ. Si consideri quindi la misura di probabilità aleatoria τ tale che, per ogni n ed ogni famiglia misurabile $\{B_1, \ldots, B_n\}$ di sottoinsiemi di R, il vettore aleatorio $(\tau(B_1), \ldots, \tau(B_n))$ ha distribuzione data da

$$(1) \qquad q_{B_1,\ldots,B_n}(d\theta_1, \ldots, d\theta_n) := \int_{\mathrm{R}^+} Q\{\varphi(B_1) \in d\theta_1, \ldots, \varphi(B_n) \in d\theta_n \,|\, \mu(\mathrm{R}) = \sigma\}\, \mathrm{h}(\sigma)\, \mathrm{Q}\{\mu(\mathrm{R}) \in \mathrm{d}\sigma\}.$$

La probabilità aleatoria τ viene chiamata NRMI condizionata. Si noti come sia ottenuta mediante una opportuna deformazione della legge della RMI μ, attraverso la funzione h. Se $h(t) = 1$ per ogni t, l'equazione (\refleggetau) determina la legge della NRMI φ, cosicché la classe delle NRMI è inclusa in quella delle NRMI condizionate.

Si prenda ora una successione $(X_n)_n$ di variabili aleatorie condizionatamente i.i.d., data la NRMI condizionata τ, aventi comune legge τ. Viene cercata un'espressione esplicita delle leggi finito-dimensionali di $(X_n)_n$, in termini della funzione di deformazione h e della misura $\nu(dx\,ds)$, su $\mathrm{R} \times \mathrm{R}^+$, che caratterizza la legge della RMI μ secondo la rappresentazione di Lévy-Khintchine. Si assume che $\nu(dx\,ds) = \eta(dx, s)\,\rho(ds)$, dove $\rho(ds)$ è una misura σ-finita su R^+ e $\eta(dx, s)$ è una misura aleatoria σ-finita su R. Per semplificare la notazione, sia $f_T(\,\cdot\,) := Q\{T \in \cdot\}$ e $l_i(B)(ds) := s^i\, 1_{s>0}\, \eta(B, s)\,\rho(ds)$; inoltre si denoti con $*$ la convoluzione di misure. Tramite applicazione di una versione multidimensionale della formula di Faà di Bruno, viene dimostrato che, per ogni n e ogni misurabile $B_1, \ldots, B_n$,

$$
\begin{aligned}
(2) \qquad & P\{X_1 \in B_1, \ldots, X_n \in B_n\} \\
& = \sum_{\pi(n)} \int_{\mathrm{R}^+} h(t)\, \frac{1}{t^n}\, \big(f_T * l_{n_1}(\cap_{i\in\pi_1} B_i) * \cdots * l_{n_{q(\pi)}}(\cap_{i\in\pi_{q(\pi)}} B_i)\big)(dt)
\end{aligned}
$$

dove la somma è estesa su tutte le partizioni $\pi = \{\pi_1, \ldots, \pi_{q(\pi)}\}$ di $\{1, \ldots, n\}$ e n_j denota il numero di elementi in π_j. Viene inoltre data l'espressione semplificata per il caso in cui alcuni dei sottoinsiemi $B_1, \ldots, B_n$ fossero coincidenti. Nel caso speciale in cui $\nu(dx\,ds) = a(dx)\,\rho(ds)$, con $a(dx)$ misura finita su R, partendo dalle leggi finito-dimensionali (\refleggesucc), vengono derivate le distribuzioni predittive della successione: se $x_1, \ldots, x_q$ sono numeri reali distinti, $n_1 + \ldots + n_q = n$, e $(\tilde{x}_1, \ldots, \tilde{x}_n)$ è un qualsiasi riordinamento di $(\underbrace{x_1, \ldots, x_1}_{n_1}, \ldots, \underbrace{x_q, \ldots, x_q}_{n_q})$, allora, per ogni misurabile B,

$$
P\{X_{n+1} \in B | X_1 = \tilde{x}_1, \ldots, X_n = \tilde{x}_n\} = \sum_{j=1,\ldots,q} p_j 1(x_j \in B) + p_{q+1}\, \frac{a(B)}{a(\mathrm{R})}
$$

con $p_j \geq 0$, per $j = 1, \ldots, q+1$, e $\sum_{j=1}^{q+1} p_j = 1$; l'espressione dei pesi p_j non è qui riportata per brevità. Le successioni di variabili aleatorie, con distribuzioni predittive che hanno questa attraente forma di combinazione lineare convessa di una misura empirica pesata e di una misura parametro, costituiscono modelli naturali per il problema del campionamento delle specie, nel quale $(X_1, \ldots, X_n)$ è un campione casuale da una grande popolazione di individui di varie specie e X_i rappresenta la specie dell'i-esimo individuo campionato. Questi modelli sono stati molto studiati in letteratura nel caso in cui la misura a è diffusa. Si veda ad esempio [2] e la bibliografia lì citata. Nella tesi viene cosiderato il caso più generale in cui a ammette degli atomi (ovvero $\eta(\cdot, s)$ ammette degli atomi, per i risultati in cui non si richiede $\nu(dx\,ds) = a(dx)\,\rho(ds)$). Mentre nel caso di a diffusa i pesi p_j dipendono solamente dalle numerosità campionarie $(n_1, \ldots, n_q)$ delle specie osservate, quando a è discreta i pesi p_j dipendono anche da quali specie siano state effettivamente osservate, e cioè da $\{x : a(\{x\}) > 0, X_i = x \text{ per qualche } i \leq n\}$. Vengono inoltre studiate altre quantità

d'interesse nel problema del campionamento delle specie, quali la legge della partizione del campione in gruppi di individui aventi la stessa specie e la legge del numero di specie differenti. Viene infine discussa una immediata estensione della classe di misure aleatorie considerata, che comprende forme molto studiate di modelli gerarchici, quali le misture di processi di Dirichlet. Quando possibile, sono ottenute espressioni in termini di polinomi esponenziali di partizioni di Bell, facilmente calcolabili mediante l'uso di ordinatori. Tutti i risultati sono illustrati sviluppando un modello che include come casi particolare alcuni modelli classici quali il Poisson-Dirichlet con due parametri, il Ferguson-Dirichlet ed il γ-stabile.

2. – Misure aleatorie basate su processi di diffusione e modelli per l'analisi della sopravvivenza.

Si consideri il processo di diffusione X, soluzione dell'equazione differenziale stocastica

$$dX_t = \beta(X_t, \theta) + \sigma dB_t\,, \quad 0 \le t \le T \le \infty\,, \quad X_0 = x_0\,, \tag{3}$$

dove $B = \{B_t : 0 \le t \le T\}$ è un moto browniano standard ed il drift $\beta(x, \theta)$, che è funzione del parametro aleatorio d-dimensionale Θ, soddisfa opportune condizioni di regolarità che assicurano l'esistenza ed unicità in senso debole della soluzione di (\refSDE). Si denoti con W_σ la legge di σB, e con P_θ la legge di X. Per il teorema di Girsanov's, la derivata di Radon-Nikodym di P_θ, rispetto a W_σ, è data da

$$g(x|\theta) := \frac{d\mathrm{P}_\theta}{d\mathrm{W}_\sigma}(x) = \exp\left\{ \int_0^T \frac{\beta(x_t, \theta)}{\sigma^2}\, dx_t - \frac{1}{2} \int_0^T \frac{\beta(x_t, \theta)^2}{\sigma^2}\, dt \right\}.$$

Sulla base del processo di diffusione X, viene definita una misura di probabilità aleatoria $\psi_{X,h}$, su $[0, T]$, ponendo

$$\psi_{X,h}\big([0,t]\big) := \frac{1 - \exp\Big\{ - \int_0^t h(X_s) ds \Big\}}{1 - \exp\Big\{ - \int_0^T h(X_s) ds \Big\}} \qquad 0 \le t \le T \tag{4}$$

dove h è un'opportuna funzione misurabile non-negativa e continua.

Si prenda ora una successione $(Y_n)_n$ di variabili aleatorie condizionatamente i.i.d., data la probabilità aleatoria $\psi_{X,h}$, aventi comune legge $\psi_{X,h}$. Da (\refDDRM), si ha che la distribuzione di $Y_1, \ldots, Y_n$, data $X = x$, ha densità, rispetto alla misura di Lebesgue n-dimensionale $\mathcal{L}^n$, data da

$$l(y_1, \ldots, y_n|x) := \left[\prod_{j=1}^n h(x_{y_j}) \right] \frac{\exp\Big\{ - \sum_{j=1}^n \int_0^{y_j} h(x_s) ds \Big\}}{\Big[1 - \exp\Big\{ - \int_0^T h(x_s) ds \Big\} \Big]^n}\,.$$

Le variabili aleatorie $Y_1, \ldots, Y_n$ possono essere viste come tempi di sopravvivenza, modellati attraverso un processo di diffusione latente che ne determina il tasso di mortalità. In particolare, il tasso istantaneo di mortalità al tempo t, ovvero il limite per $\Delta t \downarrow 0$ di $P(t \leq Y_i < t + \Delta t | Y_i \geq t)/\Delta t$, è dato da $h(X_t)$ nel caso di orizzonte temporale T infinito e da $h(X_t)/\left(1 - \exp\left\{-\int_0^T h(X_s)ds\right\}\right)$ per T finito. L'idea di utilizzare processi di diffusione nella costruzione di modelli per l'analisi della sopravvivenza risale alla fine degli anni '70 e risponde al pensiero che "quando si modellizzano dati di sopravvivenza può essere d'interesse immaginare un processo sottostante che conduca all'evento in questione" ([1]). In letteratura era già stato studiato un modello per l'analisi della sopravvivenza in cui il tasso di mortalità istantaneo è dato dal quadrato di un processo di Ornstein-Uhlenbeck. Si veda ad esempio [1] e la bibliografia lì citata. Nella tesi viene mostrato come questi modelli diffusivi latenti possano essere efficientemente trattati mediante tecniche McMC, per scelte generali del processo di diffusione e della funzione h. In particolare, vengono cercate stime McMC delle distribuzioni predittive di $(Y_n)_n$, o, equivalentemente, delle distribuzioni di sopravvivenza. Per far ciò viene sviluppato uno schema Hastings-within-Gibbs per la simulazione dalla distribuzione di (Θ, X) condizionata a $(Y_1, \ldots, Y_n)$, che ha densità, rispetto alla misura prodotto $\mathcal{L}^d \otimes \mathrm{W}_\sigma$, proporzionale a $p_\Theta(\theta)g(x|\theta)l(y_1, \ldots, y_n|x)$, dove p_Θ è la distribuzione a priori di Θ. Una speciale attenzione è dedicata alla mossa di aggiornamento della diffusione X: per aumentare la probabilità di accettazione di tale mossa, viene proposto un aggiornamento della diffusione solo su di un piccolo sottointervallo di $[0, T]$ alla volta, tenendo fisso il resto della traiettoria, mediante un sistema di ponti browniani su intervalli sovrapposti. Viene inoltre mostrato come lo schema McMC possa essere reso particolarmente robusto rispetto alla scelta dei parametri del modello, mediante una riparametrizzazione del modello stesso in termini di $(\Theta; \widetilde{X}; Y_1, \ldots, Y_n)$, dove $\widetilde{X}_t = 1(t \leq Y_{[n]})X_t + 1(t > Y_{[n]})\left[B_t - B_{Y_{[n]}}\right]$, con $Y_{[n]} = \max\{Y_1, \ldots, Y_n\}$. Con questa parametrizzazione si esprime l'idea che i dati non portano informazione circa il sottostante processo X per tempi successivi al più grande fra i tempi di sopravvivenza osservati. Viene infine anche considerato il caso in cui si hanno molteplici gruppi di osservazioni, come avviene tipicamente quando si effettuano studi clinici.

RIFERIMENTI BIBLIOGRAFICI

[1] Aalen O.O. e Gjessing H.K., *Survival models based on the Ornstein-Uhlenbeck process*, Lifetime Data Anal., **10** (2004), 407-423.

[2] Pitman J., *Poisson-Kingman partitions*, IMS Lecture Notes Monogr. Ser., **40** (2003), 1-34.

[3] Regazzini E., Lijoi A. e Prünster I., *Distributional results for means of normalized random measures with independent increments*, Ann. Statist., **31** (2003), 560-585.

Dipartimento di Matematica, Politecnico di Milano
e-mail: laura.sangalli@polimi.it
Dottorato in Matematica e Statistica
(sede amministrativa: Università degli Studi di Pavia) - Ciclo XVIII
Direttori di ricerca: Prof. Eugenio Regazzini, Università degli Studi di Pavia,
e Prof. Gareth O. Roberts, Lancaster University (UK)