Bollettino
Unione Matematica Italiana

Sezione A – La Matematica nella Società e nella Cultura

Igor Prünster

Misure di probabilità aleatorie derivate da processi additivi crescenti e loro applicazione alla statistica bayesiana

Bollettino dell'Unione Matematica Italiana, Serie 8, Vol. **7-A**—*La Matematica nella Società e nella Cultura* (2004), n.3, p. 563–566.

Unione Mastematica Italiana

<http://www.bdim.eu/item?id=BUMI_2004_8_7A_3_563_0>

Bollettino dell'Unione Matematica Italiana, Unione Mastematica Italiana, 2004.

Bollettino U. M. I.
La Matematica nella Società e nella Cultura
Serie VIII, Vol. VII-A, Dicembre 2004, 563-566

Misure di probabilità aleatorie derivate da processi additivi crescenti e loro applicazione alla statistica bayesiana.

Igor Prünster

1. – Premessa.

Alla base dell'impostazione bayesiana all'inferenza vi è la nozione di misura di probabilità aleatoria (m.p.a.) la cui legge riveste il ruolo di distribuzione iniziale. Da un punto di vista concettuale, al paradigma bayesiano è stato dato un assetto rigoroso fin dagli anni '30 mediante l'introduzione del concetto di scambiabilità e del celebre teorema di rappresentazione dovuti a Bruno de Finetti. Per quanto concerne l'approccio comunemente detto non parametrico, in cui la m.p.a. non dipende da un parametro finito-dimensionale, le difficoltà di natura tecnica nel proporre esempi concreti ne hanno a lungo impedito lo sviluppo. Soltanto a partire dall'introduzione della la m.p.a. di Dirichlet in [3], si è potuto assistere ad un proliferare di ricerche volte sia allo studio delle proprietà del processo di Dirichlet sia all'individuazione di m.p.a. alternative ad essa.

2. – Il processo di Dirichlet.

Sembra opportuno richiamare brevemente la definizione e le principali proprietà del processo di Dirichlet, dato anche il ruolo predominante che svolge tuttora in ambito bayesiano. Ai fini della trattazione seguente, tra le diverse possibili costruzioni della m.p.a di Dirichlet, la procedura che porta a definire il processo di Dirichlet come normalizzazione di un processo gamma riveste particolare importanza. Sia dunque $\Gamma := \{\Gamma_t : t \geq 0\}$ un processo gamma, i.e. un processo stocastico definito su uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$ t.c. i suoi incrementi sono indipendenti, stazionari e distribuiti secondo una legge gamma. Sia, inoltre, α una misura su $\mathcal{B}(\mathbb{R}R)$ t.c. $0 < \alpha(\mathbb{R}R) := a < +\infty$ e con corrispondente funzione di distribuzione A. Allora, mediante la riparametrizzazione $t = A(x)$ si ottiene un processo $\Gamma_A = \{\Gamma_{A(x)} : x \in \mathbb{R}R\}$ ad incrementi indipendenti (ma non stazionari) t.c. $\Gamma_{A(x)} - \Gamma_{A(y)}$ ha distribuzione gamma di parametri $A(x) - A(y)$ e 1, per ogni $y < x$. Inoltre, $\Gamma_a := \lim_{x \to +\infty} \Gamma_a(x)$ è strettamente positivo e, grazie alla riparametrizzazione, finito q.c.-$\mathbb{P}$. Di conseguenza, l'operazione di normalizzazione è ben definita e $\widetilde{F}(\cdot) = \Gamma_{A(\cdot)}/\Gamma_a$ rappresenta una funzione di ripartizione aleatoria. La corrispondente m.p.a., $\widetilde{\mathcal{D}}_\alpha$, è detta processo di Dirichlet.

Passiamo ora in rassegna alcune delle principali proprietà del processo di Dirichlet. Per quanto concerne i suoi momenti, di solito usati per incorporare le in-

formazioni di cui si dispone a priori, si ha, ad esempio, per ogni $B \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{E}[\widetilde{\mathcal{Q}}_\alpha(B)] = \frac{\alpha(B)}{a} \qquad \mathrm{Var}[\widetilde{\mathcal{Q}}_\alpha(B)] = \frac{\alpha(B)(a - \alpha(B))}{a^2(a+1)}.$$

Per quanto concerne lo studio di quantità a posteriori, si assuma che $(\Omega, \mathcal{F}, \mathbb{P})$ supporti anche una successione $X = (X_n)_{n \geq 1}$ di variabili aleatorie scambiabili, i.e. di variabili aleatorie indipendenti e identicamente distribuite dato $\widetilde{\mathcal{Q}}_\alpha$. La distribuzione predittiva di $\widetilde{\mathcal{Q}}_\alpha$ assume la forma

$$\mathbb{P}(X_{n+1} \in \cdot \,|\, X_1, \ldots, X_n) = \frac{a}{a+n} \frac{\alpha(\cdot)}{a} + \frac{n}{a+n} \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(\cdot).$$

Inoltre, la distribuzione di Dirichlet risulta godere dell'attraente proprietà di coniugio: infatti, la distribuzione finale di $\widetilde{\mathcal{Q}}_\alpha$, dati $X_1, \ldots, X_n$, è ancora Dirichlet con parametro $\alpha^* = \alpha + \sum_{i=1}^{n} \delta_{X_i}$.

Un'altra interessante problematica, affrontata per la prima volta in , è costituita dallo studio della distribuzione di medie del processo di Dirichlet, i.e. di funzionali lineari $\widetilde{\mathcal{Q}}_\alpha(f) := \int f \mathrm{d}\widetilde{\mathcal{Q}}_\alpha$ ove f è una qualsiasi funzione misurabile a valori reali. In [1] si dimostra che $\widetilde{\mathcal{Q}}_\alpha(f)$ è finito q.c.-$\mathbb{P}$ se $\int_{\mathbb{R}} \log(1 + \lambda |f(x)|)\alpha(\mathrm{d}x) < +\infty$ per ogni $\lambda > 0$ e che la funzione di ripartizione di $\widetilde{\mathcal{Q}}_\alpha(f)$ è data da

$$\mathbb{F}(\sigma) = \frac{1}{2} - \frac{1}{\pi} \int_{(0, +\infty)} \frac{1}{t} \mathrm{Im} \left\{ \exp\left[- \int_{\mathbb{R}} \log[1 + it(\sigma - f(x))]\, \alpha(\mathrm{d}x) \right] \right\} \mathrm{d}t$$

per ogni $\sigma \in \mathbb{R}R$ ove $\mathrm{Im}\, z$ denota la parte immaginaria di $z \in \mathbb{R}C$.

3. – Misure aleatorie ad incrementi indipendenti normalizzate.

Nel lavoro è stata introdotta ed analizzata una nuova classe di m.p.a., le misure aleatorie ad incrementi indipendenti normalizzate (RMI normalizzate). Nell'ambito di questa classe di m.p.a. che contiene il processo di Dirichlet come caso particolare, si è cercato di capire, da un punto di vista descrittivo, le ragioni profonde della semplicità delle espressioni analitiche cui dà luogo il processo di Dirichlet e, da un punto di vista costruttivo, di proporre delle alternative concrete.

L'idea che porta alla definizione di una RMI normalizzata è semplice: nella precedente costruzione di $\widetilde{\mathcal{Q}}_\alpha$ si tratta di sostituire il processo gamma con un generico processo additivo crescente (PAC). Un PAC $\xi := \{\xi_t : t \geq 0\}$, definito su uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$, è essenzialmente un processo stocastico ad incrementi indipendenti (non necessariamente stazionari) con traiettorie crescenti e continue da destra e t.c. $\xi_0 = 0$ q.c.-$\mathbb{P}$. Dato un PAC ξ, dalla riparametrizzazione mediante una misura finita α si ottiene un PAC ξ_α finito q.c.-$\mathbb{P}$. Si dimostra che la trasformata di Laplace di $\xi_\alpha(B)$, per ogni $B \in \mathcal{B}(\mathbb{R}R)$, è data da

$$\mathbb{E}\left[\mathrm{e}^{-\lambda \xi_\alpha(B)}\right] = \exp\left[- \int_{B \times (0, +\infty)} (1 - \mathrm{e}^{-\lambda v})\, \nu_\alpha(\mathrm{d}x, \mathrm{d}v) \right]$$

ove ν_α denota la misura di intensità di ξ_α. Tale misura di intensità riveste un ruolo cruciale nella derivazione dei risultati seguenti, in quanto individua univoca-

mente ξ_α. Per poter normalizzare ξ_α, bisogna garantire che $\xi_a > 0$ q.c.-$\mathbb{P}$. Si mostra che $\nu_\alpha(\mathbb{R}R, (0, +\infty)) = +\infty$ rappresenta una condizione necessaria e sufficiente affinché ciò avvenga. Garantita la liceità dell'operazione di normalizzazione, la m.p.a.

$$\tilde{P}(\cdot) = \frac{\xi_\alpha(\cdot)}{\xi_a}$$

è detta RMI normalizzata.

Al contrario del caso del processo di Dirichlet, il calcolo dei momenti sia iniziali sia finali di una RMI normalizzata è tutt'altro che agevole, poiché, in generale, le sue distribuzioni finito-dimensionali non sono note. A tal fine è stata sviluppata una tecnica, basata sulla solo conoscenza della trasformata di Laplace di ξ_α, che permette di derivare i momenti di una RMI normalizzata. Ad esempio, si ha che

$$\mathbb{E}\big[\tilde{P}(B)\big] = \int_{(0,+\infty)} \left\{ \int_{(0,+\infty)} v\, \mathrm{e}^{-uv}\, \mathrm{e}^{-\int_{\mathbb{R}\times(0,+\infty)}(1-\mathrm{e}^{-\lambda u})\,\nu_\alpha(\mathrm{d}x,\,\mathrm{d}v)}\,\mathrm{d}u \right\} \nu_\alpha(B,\,\mathrm{d}v).$$

Nel caso di una RMI normalizzata caratterizzata da una misura di intensità del tipo $\nu_\alpha(\mathrm{d}x,\,\mathrm{d}v) = \alpha(\mathrm{d}x)\,\nu(\mathrm{d}v)$, le espressioni dei momenti si riducono, e.g., a

$$\mathbb{E}\big[\tilde{P}(B)\big] = \frac{\alpha(B)}{a} \qquad \mathrm{Var}\,[\tilde{P}(B)] = \frac{\alpha(B)(a-\alpha(B))}{a^2}\,\Im,$$

ove

$$\Im := a \int_{(0,+\infty)} u \exp\left\{ -a \int_{(0+\infty)} (1-\mathrm{e}^{-\lambda v})\,\nu(\mathrm{d}v) \right\} \int_{(0,+\infty)} v^2\, \mathrm{e}^{-uv}\,\nu(\mathrm{d}v)\,\mathrm{d}u\,.$$

Per quanto concerne le distribuzioni predittive, si assuma la scambiabilità delle osservazioni, si indichi con $X_1^*, \ldots, X_k^*$ le k osservazioni distinte presenti nel campione con n_j osservazioni uguali a X_j^*. Allora, se α è una misura diffusa, si ha

$$\mathbb{P}(X_{n+1} \in \mathrm{d}x \,|\, X_1, \ldots, X_n) = w^{(n)}\,\alpha(\mathrm{d}x) + \frac{1}{n}\sum_{j=1}^{k} w_j^{(n)}\,\delta_{X_j^*}(\mathrm{d}x)$$

ove i pesi, ricavati in forma esplicita, dipendono dalla misura di intensità e dalle osservazioni. La regola predittiva è dunque una combinazione lineare tra la misura α e la distribuzione empirica pesata in base alla frequenza delle osservazioni generalizzando, quindi, la distribuzione predittiva del processo di Dirichlet in maniera intuitiva.

Con riferimento alla distribuzione finale, si giunge ad un risultato negativo, in quanto si dimostra che il processo di Dirichlet è l'unica RMI normalizzata che gode della proprietà di coniugio. Ciononostante, risulta possibile caratterizzare la distribuzione finale in termini di mistura.

Interessanti risultati si ottengono per la distribuzione di medie di RMI normalizzate $\tilde{P}(f)$. Infatti, si mostra che $\tilde{P}(f)$ è finita q.c.-$\mathbb{P}$ se e solo se vale

$$\int_{\mathbb{R}\times(0,+\infty)} [1 - \exp(-\lambda v |f(x)|)]\,\nu_\alpha(\mathrm{d}x,\,\mathrm{d}v) < +\infty \quad \text{per ogni } \lambda > 0.$$

Quindi, ricavando la funzione caratteristica di un funzionale lineare di un PAC e sfruttando opportunamente la formula d'inversione di Gurland, si ottiene anche

la distribuzione di $\tilde{P}(f)$, la cui funzione di ripartizione è data da

$$\mathbb{F}(\sigma) = \frac{1}{2} - \frac{1}{\pi} \lim_{T \uparrow +\infty} \int_0^T \frac{1}{t} \operatorname{Im} \left\{ \exp \left[- \int_{\mathbb{R} \times (0, +\infty)} [1 - e^{itv(f(x) - \sigma)}] \, \nu_\alpha(dx, dv) \right] \right\} dt$$

per ogni $\sigma \in \mathbb{R}R$. Inoltre, si ottengono espressioni per la distribuzione finale di medie di RMI normalizzate in termini dell'integrale frazionario di Liouville-Weyl.

L'analisi delle RMI normalizzate è stata poi completata considerando vari casi particolari di rilevanza statistica. In quei casi, le espressioni generali si semplificano notevolmente permettendone l'impiego, diretto oppure mediante opportune approssimazioni numeriche, a problemi inferenziali concreti.

4. – Ulteriori risultati.

Le RMI normalizzate sono state quindi ulteriormente generalizzate a misure aleatorie normalizzate dirette da PAC, le quali contengono il processo mistura di Dirichlet, introdotto in [4], come caso speciale. Sono fornite condizioni per la loro esistenza. In particolare, si ottengono risultati per la distribuzione di loro medie sia iniziali sia finali e, mediante l'introduzione di adeguate variabili latenti, si propone un algoritmo di simulazione.

Infine, sono state considerate anche le cosiddette m.p.a. neutrali a destra introdotte in [2]. Le loro medie possono essere rappresentate come funzionali esponenziali di PAC. Questo fatto viene sfruttato per fornire condizioni sufficienti per la finitezza di una media e per l'assoluta continuità della sua distribuzione. Inoltre, si ricavano espressioni per i suoi momenti di ogni ordine. Ricorrendo all'algoritmo di massima entropia, si ottiene un'approssimazione della densità di una media di un m.p.a. neutrale a destra.

BIBLIOGRAFIA

[1] CIFARELLI D.M. e REGAZZINI E., *Distribution functions of means of a Dirichlet process*, Ann. Statist., **18** (1990), 429-442.

[2] DOKSUM K., *Tailfree and neutral random probabilities and their posterior distributions*, Ann. Probab., **2** (1974), 183-201.

[3] FERGUSON T.S., *A Bayesian analysis of some nonparametric problems*, Ann. Statist., **1** (1973), 209-230.

[4] LO A.Y., *On a class of Bayesian nonparametric estimates: I. Density estimates*, Ann. Statist., **12** (1984), 351-357.

Dipartimento di Economia Politica e Metodi Quantitativi
Università degli Studi di Pavia; e-mail: igor.pruenster@unipv.it
Dottorato in Statistica Matematica
(sede amministrativa: Università di Pavia) - Ciclo XIV
Direttore di ricerca: Prof. Eugenio Regazzini, Università degli Studi di Pavia