Bollettino
Unione Matematica Italiana

Sezione A – La Matematica nella Società e nella Cultura

Alessandra Guglielmi

Risultati sulle distribuzioni di medie di un processo di Dirichlet

Bollettino dell'Unione Matematica Italiana, Serie 8, Vol. **1-A**—*La Matematica nella Società e nella Cultura* (1998), n.1S (Supplemento Tesi di Dottorato), p. 125–128.

Unione Matematica Italiana

<http://www.bdim.eu/item?id=BUMI_1998_8_1A_1S_125_0>

Bollettino dell'Unione Matematica Italiana, Unione Matematica Italiana, 1998.

La matematica nella Società e nella Cultura
Bollettino U. M. I.
(8) 1-A Suppl. (1998), pag. 125-128

Risultati sulle distribuzioni di medie di un processo di Dirichlet.

ALESSANDRA GUGLIELMI

Il problema affrontato nella tesi è il calcolo di distribuzioni di variabili aleatorie del tipo

$$\Gamma_\alpha^g = \int_{\Re} g(x)\, P_\alpha(dx) \tag{1}$$

dove P_α è una misura di probabilità aleatoria distribuita secondo la legge di Ferguson-Dirichlet con parametro α, essendo α una misura finita sulla retta reale, sotto ipotesi per le quali l'espressione (1) è finita con probabilità 1.

Tale studio prende spunto da problemi di inferenza bayesiana non parametrica. Considerata una successione indefinitamente proseguibile di osservazioni della stessa quantità, ottenute in condizioni ambientali analoghe, si può ritenere che la sua legge di probabilità sia invariante per l'ordine dei risultati, ovvero che le osservazioni, rappresentate da una successione di variabili aleatorie $\{X_n\}_{n\geq 1}$, siano scambiabili. Questa proprietà permette di rappresentare la legge della successione $\{X_n\}_{n\geq 1}$ in modo piuttosto semplice: condizionatamente alla misura di probabilità aleatoria $\tilde{p}$, limite della successione della legge empirica (al divergere del numero di osservazioni), le X_n sono indipendenti ed identicamente distribuite secondo $\tilde{p}$. I problemi statistici che nascono in questo contesto sono di due tipi: la previsione del risultato di osservazioni future condizionata a quello di altre osservazioni e l'inferenza su $\tilde{p}$.

Descritto sommariamente il tipo di problemi affrontati nella tesi, riporto per completezza l'enunciato del teorema di rappresentazione (dovuto a Bruno de Finetti) sopra richiamato. Ci si limita a considerare il caso di variabili aleatorie X_n reali, definite su uno spazio di probabilità $(\Omega, \mathcal{F}, P)$. Con $\mathcal{P}$ si denota lo spazio delle misure di probabilità sui boreliani di $\Re$, munito della topologia della convergenza debole.

Una successione di variabili aleatorie reali $\{X_n\}_{n\geq 1}$ è scambiabile se, e solo se, esiste una misura di probabilità q sui boreliani di $\mathcal{P}$ tale che

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = \int_{\mathcal{P}} \prod_{i=1}^{n} p(A_i)\, q(dp), \quad A_i \in \mathcal{B}(\Re), \quad i = 1, \ldots, n, \tag{2}$$

per ogni intero positivo n. La misura q è la legge limite della distribuzione empirica $(1/n)\sum_1^n \delta_{X_i(\omega)}$, *per* $n \to +\infty$. *Inoltre la rappresentazione* (2) *è unica.*

La legge q viene detta *misura di de Finetti* della successione scambiabile $\{X_n\}_n$ o *misura iniziale* di $\tilde{p}$.

Negli ultimi venti anni si è andata affermando un'*impostazione bayesiana non parametrica*, caratterizzata dalla scelta di misure di de Finetti non ristrette a sottoclassi parametriche di $\mathcal{P}$. Tra quelle più usate si può citare la misura di Ferguson-Dirichlet, legge di una misura di probabilità aleatoria P_α, detta *processo di Ferguson-Dirichlet* (o, più semplicemente, di Dirichlet), caratterizzata dalla seguente condizione: se $A_1, \ldots, A_n$ è una qualsiasi partizione finita e misurabile di $\mathfrak{R}$, con $\alpha(A_j) > 0$ per ogni $j = 1, \ldots, n$, il vettore $(P_\alpha(A_1), \ldots, P_\alpha(A_n))$ ha distribuzione di Dirichlet con parametri $(\alpha(A_1), \ldots, \alpha(A_n))$, dove α è misura finita su $\mathfrak{R}$, detta parametro del processo (per una definizione completa, si veda [3]).

Si può ricordare che esiste una distribuzione condizionale regolare di P_α, dato $(X_1, \ldots, X_n)$ - la cosiddetta distribuzione finale - del tipo Ferguson-Dirichlet, con parametro $\alpha + \sum_1^n \delta_{X_i}$, dove δ_x indica la misura che assegna massa 1 a $\{x\}$ e 0 altrove. Questa proprietà risulta comoda in vista del calcolo di inferenze su P_α.

Di rilevante interesse statistico è, in generale, il calcolo della distribuzione di «valori caretteristici» della misura di probabilità aletoria $\tilde{p}$ (il cui significato è posto in luce dal teorema di rappresentazione di de Finetti sopra richiamato) come, ad esempio, i momenti, ed in particolare la media aritmetica (momento primo). Nella tesi mi sono occupata innanzitutto della distribuzione di Γ_α^g con $g(x) = x$, $x \in \mathfrak{R}$, che d'ora in poi chiamerò *media di Dirichlet*, indicandola con Γ_α. In effetti, è praticamente immediato ricavare la legge di Γ_α^g per una g arbitraria da quella di Γ_α. Il problema considerato è stato studiato da diversi autori; in particolare in [1] è stata ricavata l'espressione della funzione di ripartizione M_α di Γ_α, attraverso un procedimento piuttosto complesso dal punto di vista analitico, ma autonomo rispetto ad altri, basato sulla trasformata di Stieltjes di tale distribuzione. L'espressione di M_α così determinata non si presta ad essere valutata numericamente senza difficoltà, a meno che la misura α sia concentrata su un insieme finito. Nella tesi viene proposto anzitutto un procedimento più semplice per ricavare l'espressione della trasformata di Stieltjes di Γ_α, utilizzando un risultato in [2] che caratterizza la misura di Ferguson-Dirichlet come legge invariante di una particolare catena markoviana, le cui realizzazioni sono misure di probabilità su $\mathfrak{R}$. Difatti, ciò permette di scrivere immediatamente, quando α è misura con supporto finito, un'equazione integrale per la funzione generatrice dei momenti m_α di Γ_α,

$$tm_\alpha(t) = \int_0^t (x/t)^{\alpha-1} m_\alpha(x)\, \widehat{\alpha}(t-x)\, dx\,, \qquad t \in \mathfrak{R}\,, \tag{3}$$

dove $\widehat{\alpha}(u) = \int_{\Re} e^{-ux}\, dA(x)$. Si dimostra che (3) è equivalente a

$$\int_{[\tau,\, +\infty)} (x+s)^{-a}\, dM_a(x) = \exp\left(- \int_{[\tau,\, T]} \log(s+x)\, dA(x) \right), \qquad s > -\tau\,,$$

dove: $[\tau, T]$ è un intervallo arbitrario che include il supporto di α, A è la funzione di distribuzione associata alla misura α con massa complessiva uguale ad $a > 0$ e il membro di sinistra è la trasformata di Stieltjes $\mathcal{S}_\alpha$ di ordine a della funzione di ripartizione M_a. La facilità con cui si ricava la trasformata di Stieltjes attraverso questo procedimento garantisce, in un certo senso, che questa trasformata è lo strumento «naturale» per lo studio della distribuzione di Γ_a. Dall'espressione di $\mathcal{S}_\alpha$ è facile ricavare la corrispondente funzione di ripartizione, attraverso una formula «classica» di inversione. L'espressione trovata può essere estesa, poi, a casi più generali se, con probabilità uguale a 1, le traiettorie di P_α sono misure di probabilità con media finita. In questo caso, infatti, la variabile aleatoria Γ_a è il limite in distribuzione della media di un processo di Dirichlet il cui parametro è una misura a supporto finito, ottenuta da α attraverso un procedimento di «troncamento» e «discretizzazione», quando detto parametro converge opportunamente ad α. Più precisamente, se il supporto di α è il compatto $[\tau, T]$ e $\mathcal{D}_n$ è una sua partizione, $\mathcal{D}_n = \{[x_{k-1}^{(n)}, x_k^{(n)}),\ k = 1, \ldots, n\}$, con $\tau = x_0^{(n)} < x_1^{(n)} < \ldots < x_n^{(n)} = T$, e $\|\mathcal{D}_n\| = \max_{1 \leq k \leq n} |x_k^{(n)} - x_{k-1}^{(n)}|$, $n = 1, 2, \ldots$, allora, definite due misure $\underline{\alpha}_n$ e $\overline{\alpha}_n$, $\underline{\alpha}_n = \alpha(\{x_0^{(n)}\})\,\delta_{x_0^{(n)}} + \sum_1^{n-1} \alpha((x_{j-1}^{(n)}, x_j^{(n)}])\,\delta_{x_j^{(n)}}$, $\overline{\alpha}_n = \alpha([x_0^{(n)}, x_1^{(n)}])\,\delta_{x_0^{(n)}} + \sum_1^{n} \alpha((x_j^{(n)}, x_{j+1}^{(n)}])\,\delta_{x_j^{(n)}}$ per ogni n, si ha

$$M_{\underline{\alpha}_n}(x) \uparrow M_\alpha(x), \qquad M_{\overline{\alpha}_n}(x) \downarrow M_\alpha(x), \qquad \text{per } \|\mathcal{D}_n\| \downarrow 0, \qquad \text{uniformemente in } x.$$

Se, invece, α ha supporto finito, e τ e T sono tali che $-\infty < \tau < T < +\infty$, definita $\alpha_{[\tau,\, T]}(A) = \alpha(A \cap [\tau, T)) + \alpha((T, +\infty))\,\delta_T(A)$ per ogni A boreliano di $\Re$, si ottiene:

$$\text{se } \int_{\Re} |t|\, P_\alpha(dt;\, \omega) < +\infty \text{ con probabilità } 1\,,$$

$$M_{\alpha_{[\tau,\, T]}}(x) \to M_\alpha(x),\ \tau \to -\infty\,,\ T \to +\infty\,, \qquad \text{uniformemente in } x\,.$$

Pertanto, nel caso in cui il supporto di α sia inferiormente limitato da una costante τ, si ottiene, per $x > \tau$,

$$M(x) = \frac{2^{a-1}}{\pi} (x-\tau)^a \int_0^{\pi} \left(\cos\frac{y}{2} \right)^{a-1} \cos\left[\frac{a+1}{2}\, y - \int_{\Re} \arg\{(x-\tau)\, e^{iy} + u\}\, dA(\tau+u) \right] \cdot$$

$$\cdot \exp\left(- \int_{\Re} \log|(x-\tau) e^{iy} + u|\, dA(\tau+u) \right) dy\,.$$

Questo procedimento di troncamento e discretizzazione suggerisce l'approssimazione di $M_\alpha(x)$ con $M_\beta(x)$, essendo M_β la funzione di ripartizione di una media di Dirichlet con parametro uguale ad una misura con supporto finito. Nella tesi largo spazio è dedicato alla scelta di β tra le misure ottenute da α per troncamento e discretizzazione, in funzione dell'errore di approssimazione. Gli esempi numerici riportati sono stati ottenuti utilizzando, per il calcolo di $M_\beta(x)$, ruotine di librerie Fortran. La parte del lavoro che riguarda la valutazione dell'errore di approssimazione è del tutto nuova in letteratura.

La procedura che ha portato a ricavare l'espressione di M_α è stata applicata anche al caso vettoriale, cioè a quelle situazioni in cui l'indagine statistica considera diversi valori di sintesi della distribuzione «incognita» $\tilde{p}$ contemporaneanemente, del tipo

$$\left(\int_{\Re} g_1(x)\, P_\alpha(dx), \ \ldots, \int_{\Re} g_m(x)\, P_\alpha(dx) \right), \tag{4}$$

oppure indici diversi dalla media, come, per esempio la varianza

$$\int_{\Re} (x - \Gamma_\alpha)^2 P_\alpha(dx). \tag{5}$$

In questi casi sono stati ottenuti solo risultati parziali come, ad esempio: una equazione integrale per la funzione caratteristica del vettore (4) ed una sua soluzione quando α ha supporto finito; una espressione per la funzione caratteristica di (5) quando α ha supporto limitato.

BIBLIOGRAFIA

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

[2] FEIGIN P.D.and TWEEDIE R.L., *Linear functionals and Markov chains associated with Dirichlet processes*, Mathematical Proceeding of the Cambridge Philosophical Society, **105** (1989), 579-585.

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

CNR-IAMI, via Ampère 56 - 20131 Milano
Dottorato in Matematica (sede amministrativa: Milano) - Ciclo VIII
Direttore di ricerca: Prof. E. Regazzini, IMQ, Università «L. Bocconi», Milano