Antonella Bodini

Un nuovo approccio ai problemi dell'inferenza statistica coerente col punto di vista bayesiano

Bollettino dell'Unione Matematica Italiana, Serie 8, Vol. **6-A**—*La Matematica nella Società e nella Cultura* (2003), n.2, p. 227–230.

Unione Matematica Italiana

<http://www.bdim.eu/item?id=BUMI_2003_8_6A_2_227_0>

Bollettino dell'Unione Matematica Italiana, Unione Matematica Italiana, 2003.

Bollettino U. M. I.
La Matematica nella Società e nella Cultura
Serie VIII, Vol. VI-A, Agosto 2003, 227-230

Un nuovo approccio ai problemi dell'inferenza statistica coerente col punto di vista bayesiano.

ANTONELLA BODINI

1. – Brevi considerazioni generali.

È concezione largamente diffusa che uno degli scopi principali dell'introduzione di un modello statistico nell'analisi di un fenomeno osservabile sia permettere di fare previsioni sullo svolgersi del fenomeno stesso. Tale capacità è esplicabile, in parte, nell'ambito di un'analisi statistica frequentista, ma è solo con l'impostazione bayesiana che si fornisce una regola generale d'apprendimento dall'esperienza, il teorema di Bayes appunto, grazie alla quale la risposta al problema della previsione si esprime, in termini di probabilità condizionali, coerentemente con i principi della teoria della probabilità.

All'interno stesso dell'ambiente bayesiano, diverso è il rilievo dato alle possibilità di impiego della legge delle realizzazioni future del fenomeno: in modo più o meno preminente, questa interviene nell'analisi usualmente in fase di convalida o di scelta del modello, dunque *dopo* la fase di stima dei parametri incogniti da cui quelle distribuzioni dipendono. È stato, tuttavia, argomentato in [1] che, se l'obiettivo primario dell'analisi è la previsione, allora l'attenzione dello statistico deve principalmente rivolgersi alla individuazione di metodi che permettano di confrontare, proprio da questo punto di vista, eventuali diversi modelli proposti, piuttosto che rivolgersi alla stima dei parametri o alla verifica delle ipotesi. Infatti, sottolineare l'importanza di queste fasi dell'inferenza porta a rivestire i parametri di un «*inappropiate existential meaning*».

Parrebbe dunque che la determinazione di buoni stimatori e la determinazione di buoni modelli previsivi vengano considerati come problemi diversipiuttosto che come fasi successive di un'unica procedura.

La precedente citazione conduce, inevitabilmente, ad un altro tipo di riflessione. Usualmente il teorema di Bayes viene applicato introducendo distribuzioni di probabilità per i parametri incogniti, distribuzioni che dovrebbero riflettere le conoscenze disponibili su tali entità le quali, tuttavia, sono generalmente non osservabili, dipendendo, nella loro esistenza e specificazione, dal particolare processo di modellizzazione utilizzato. Che l'assegnazione di valutazioni di probabilità per enti privi di un significato empirico sia una questione cruciale soprattutto dal punto di vista dell'interpretazione dei risultati ottenuti, è un fatto generalmente riconosciuto non solo dagli statistici classici, che ne fanno una delle principali critiche all'impostazione bayesiana, ma anche all'interno della stessa comunità bayesiana.

Osserviamo, infine, che il ricorso allo schema bayesiano convenzionale viene spesso giustificato dalla difficoltà computazionale cui si va incontro volendo determinare la legge delle osservazioni future a partire direttamente dalla legge congiunta delle osservazioni realizzate e di quelle che si vogliono prevedere, cioè con approcci di tipo diretto, ma la crescente efficienza degli strumenti di calcolo e delle tecniche di simulazione sta, forse, togliendo progressivamente terreno a questa argomentazione.

Accogliendo tutti questi spunti, in questo lavoro di tesi abbiamo studiato un approccio all'inferenza statistica secondo una concezione a nostro parere completamente bayesiana, che, se non del tutto originale, è stata finora, per quanto a nostra conoscenza, poco esplorata: essa si propone di finalizzare l'analisi alla previsione fin dalla fase di stima e di coniugare la capacità di previsione dei metodi bayesiani con l'approccio frequentista al trattamento dei parametri.

2. – Il metodo proposto.

La possibilità di realizzare, sul fenomeno osservato, accurate previsioni deriva, in generale, dalla combinazione di due fattori: un ragionamento logico-deduttivo che porta all'introduzione di un modello che, in quanto semplificazione della dinamica del fenomeno stesso, dipende da parametri incogniti; un procedimento di apprendimento dall'esperienza che definisce il modo in cui la previsione soggiace a quanto l'osservazione segnala come rilevante.

Abbiamo già avuto modo di dire che il nostro atteggiamento riguardo la stima dei parametri sarà, per così dire, di tipo frequentista: ogni informazione disponibile verrà convogliata, infatti, nella formulazione di leggi per le sole quantità osservabili. Assunzione concettualmente necessaria, secondo noi, per riuscire in questo è che esistano, sul fenomeno, studi autorevoli, intendendo con ciò studi condotti da esperti del settore, o fondati su consolidate teorie, grazie ai quali sia possibile fissare, a meno di un parametro incognito θ, un modello teorico affidabile ai fini della previsione, modello che potrà essere formulato sia direttamente in termini di una legge $\varphi_{1,\theta}$ per la generica osservazione, sia in termini di particolari caratteristiche (momenti, medie condizionali, ecc.) della legge stessa. Supporremo d'ora in poi che l'insieme $\mathbf{E}$ di tutte le determinazioni possibili di una qualunque osservazione sia uno spazio metrico, separabile e completo, cioè uno spazio *polacco*, e che $\mathcal{E} = \mathcal{B}(\mathbf{E})$ sia la sua σ-algebra di Borel. Posto $\mathbf{E}^{N-1} = \otimes_{1 \leq j < N} \mathbf{E}_j$, per $N \leq +\infty$, ove $\mathbf{E}_j = \mathbf{E}$ per ogni indice j; indicando con $\mathcal{E}^{N-1}$ la corrispondente σ-algebra prodotto e con $x^{(N-1)}$ un qualunque elemento $(x_i)_{1 \leq j < N}$ di $\mathbf{E}^{N-1}$, rappresentiamo l'osservazione i-ma con la proiezione coordinata, X_i, che associa ad ogni elemento $x^{(N-1)}$ la sua i-ma componente x_i. La famiglia di variabili aleatorie così costruita prende il nome di *processo canonico*. Poichè il modello teorico esprime la legge di una qualunque osservazione, richiediamo che *la famiglia sia equidistribuita, con legge comune* $\varphi_{1,\theta}$. Fissata, per ogni valore del parametro θ, una successione, $\Phi_\theta = \{\varphi_{j,\theta}\}_{j \geq 2}$, di *nuclei di probabilità* da $\mathbf{E}^{j-1}$ a $\mathcal{E}$, che garantiscano la condizione di equidistribuzione, per il teorema di Ionescu Tulcea [2] resta univocamente individuata una misura di probabilità P_θ su $(\mathbf{E}^N, \mathcal{E}^N)$ rispetto alla quale X_1 ha legge $\varphi_{1,\theta}$ e $\varphi_{n,\theta}(x^{(n-1)}, \bullet)$ è la distribuzione condizionale di X_n dato $X^{(n-1)}$, cioè una *distribuzione predittiva*.

La scelta di Φ_θ può essere, quindi, correttamente interpretata come specificazione, necessariamente di carattere soggettivo, delle modalità con cui si ritiene che le informazioni relative alle osservazioni realizzate debbano influire sulla previsione delle realizzazioni future.

ESEMPIO 1. – *Sia* **E** *uno spazio polacco qualunque. Fissata una legge* $\varphi_{1,\theta}$ *sia* α *un numero reale strettamente positivo. Si consideri la successione di nuclei di probabilità cosìdefinita:*

$$\varphi_{n+1,\theta}(x^{(n)}, A) = \frac{\alpha}{\alpha + n}\varphi_{1,\theta}(A) + \frac{n}{\alpha + n}\, e_n(A) \tag{1}$$

ove con e_n *indichiamo la misura empirica* $e_n(A) = e_{x^{(n)}}(A) = \sum_{i=1}^{n} \delta_{x_i}(A)/n$.

Con questa scelta la successione $(X_1, X_2, \ldots)$ *è scambiabile con misura di de Finetti data dalla legge di un processo di Ferguson-Dirichlet con parametro* $\alpha(\bullet) = \alpha\varphi_{1,\theta}(\bullet)$.

Naturalmente ogni problema inferenziale è completamente risolto quando venga stimato θ, restando così fissata la legge, P_θ, dell'intera famiglia delle osservazioni. Ricordando quanto detto nel paragrafo precedente, cominciamo con l'osservare che, a questo punto, abbiamo a disposizione due strumenti di previsione, il modello teorico ed il processo induttivo che da esso prende origine. Ovviamente il secondo può presentare una notevole complessità rispetto al primo, *in primis* dal punto di vista computazionale. Sembra evidente, dunque, la necessità di determinare criteri che legittimino l'uso del modello teorico al posto di quello statistico. Questo problema non è certamente nuovo in statistica e spesso viene trattato come un problema di selezione del modello, ma quello che noi proponiamo è che la sua soluzione diventi la chiave anche per la stima. In effetti, la sostituzione di cui sopra risulta appropriata se la discrepanza tra le previsioni fatte col modello teorico $\varphi_{1,\theta}$ e quelle realizzate grazie alla legge P_θ è la più piccola possibile. Definiamo, quindi, la perdita L cui si è soggetti usando il modello $\varphi_{1,\theta}$ al posto della legge predittiva $\varphi_{n+1,\theta}$, quando il risultato sperimentale sia dato dal vettore $x^{(n)}$, come una funzione $L : \mathbb{P} \times \mathbb{P} \to [0, +\infty)$, essendo $\mathbb{P}$ un sottinsieme di misure di probabilità su $(\mathbf{E}, \mathcal{E})$, per la quale risulti $L(p, p) = 0$ per ogni p in $\mathbb{P}$. Secondo quanto detto sopra *uno stimatore di* θ *sarà un qualunque valore* $\widehat{\theta}$ *per cui*:

$$L\big(\varphi_{1,\widehat{\theta}}, \varphi_{n+1,\widehat{\theta}}(x(n))\big) = \min_{\theta \in \Theta} L\big(\varphi_{1,\theta}, \varphi_{n+1,\theta}(x(n))\big)$$

Si è tacitamente assunto che L sia finita per ogni θ e $x^{(n)}$ in $\mathbf{E}^n$ e che L sia nulla quando le leggi predittive non siano dipendenti dall'osservazione.

Il metodo proposto è piuttosto generale potendo incorporare ogni modello assegnato, anche con spazio dei parametri che non sia un sottinsieme dell'ordinario spazio euclideo, e potendo applicarsi anche ad osservazioni che non si presentino in forma puntuale, ma, p.es. in forma raggruppata. Per concludere questa nota e chiarire meglio quanto esposto, esamineremoo brevemente questo caso in maggior dettaglio. Nel contesto presentato nell'Esempio 1, sia $\mathbf{E} = \mathbb{R}$ e sia $C = \{C_1, \ldots, C_k\}$ una prefissata partizione misurabile di $\mathbb{R}$. Supponiamo, come detto, che l'informazione ricavata dall'osservazione di un campione di taglia n, $(X_1, \ldots, X_n)$, consista nel numero $n_i = \sum_{j=1}^{n} I_{C_i}(X_j)$ di osservazioni cadute in ciascun elemento C_i e di voler prevedere il valore dell'osservazione $(n+i)$-ma. Poichè in letteratura si è soliti considerare i dati raggruppati come una rappresentazione semplificata di un gran numero di osservazioni puntuali, riteniamo di maggior interesse poter dare valutazioni di probabilità per gli eventi della forma $\{X_{n+i} \in A\}$, per qualunque insieme misurabile A, piuttosto che limitarci a quelli generati dagli elementi di C. Per questo assumiamo come modello teorico $\varphi_{1,\theta}$ una legge definita su $\mathbb{R}$ e non su una sua discretizzazione indotta da C. Per l'ipotesi di scambiabilità possiamo fissare $i=1$. La legge di X_{n+1} condizionatamente a $G=(n_1, \ldots, n_k)$ è data da

$$\varphi^*_{n+1,\theta}(A) = \frac{n}{\alpha+n} \sum_{j=1}^{k} \frac{n_j}{n} \frac{\varphi_{1,\theta}(A \cap C_j)}{\varphi_{1,\theta}(C_j)} + \frac{\alpha}{\alpha+n} \varphi_{1,\theta}(A) \qquad (A \in \mathcal{E}), \tag{2}$$

se $\varphi_{1,\theta}(C_j) > 0$ per ogni j. Se $\varphi_{1,\theta}(C_j) = 0$ per qualche j, allora la legge predittiva non è univocamente determinata, ma se ne possono comunque determinare delle versioni coerenti con (2). Significative funzioni di perdita sono date dalla distanza in variazione totale tra $\varphi_{1,\theta}$ e $\varphi^*_{n+1,\theta}$ e la divergenza di Kullback-Leibler di $\varphi_{1,\theta}$ su $\varphi^*_{n+1,\theta}$ che si esprimono, rispettivamente, come:

$$d_{TV}(\varphi_{1,\theta}, \varphi^*_{n+1,\theta}) = \frac{n}{2(\alpha+n)} \sum_{j=1}^{k} \left| \varphi_{1,\theta}(C_j) - \frac{n_j}{n} \right|, \tag{3}$$

$$d_{KL}(\varphi_{1,\theta}, \varphi^*_{n+1,\theta}) = \sum_{j=1}^{k} \varphi_{1,\theta}(C_j) \operatorname{Log} \frac{\varphi_{1,\theta}(C_j)(\alpha+n)}{n_j + \alpha\varphi_{1,\theta}} . \tag{4}$$

Nella tesi, abbiamo applicato i risultati ottenuti in questo contesto all'analisi del reddito delle famiglie italiane.

Osserviamo che, se diversi sono i modelli plausibili per il fenomeno in questione (anche con distinti spazi del parametro), questi risultano automaticamente confrontabili, quando il processo di apprendimento dall'esperienza sia considerato fissato, coerentemente con la procedura di stima introdotta: dopo n osservazioni il modello *preferibile* risulterà essere quello che rende minima la perdita in corrispondenza al valore stimato del parametro.

Segnaliamo, per concludere, che abbiamo studiato il metodo proposto anche in relazione a problemi di regressione.

BIBLIOGRAFIA

[1] GEISSER S., *Predictive Inference: an Introduction*, *Chapmanand Hall*, *Londra* (1993).
[2] KALLENBERG O., *Foundations of Modern Probability*, *Springer* (1997).

Dipartimento di Matematica, Università di Pavia
e-mail: anto@mi.imati.cnr.it
Dottorato in Statistica Matematica (sede amministrativa: Pavia) - Ciclo XIII
Direttore di ricerca: Prof. Eugenio Regazzini, Università di Pavia