Lezione n.4: Riassunti eusastivi e distribuzioni iniziali

Roma, 13 marzo 2003

Brunero Liseo
Dipartimento di studi geoeconomici, linguistici, statistici e storici per l'analisi regionale
Università di Roma "La Sapienza"
Rome, Italy
brunero.liseo@uniroma1.it
tel. 06-49766110

Riassunto esaustivo

Analizzando il modo in cui è determinata la distribuzione finale, si nota come essa dipenda dal risultato osservato, cioè dai dati, solo attraverso la funzione di verosimiglianza $L(\theta; z)$, o meglio il *nucleo* della stessa. Se $L(\theta; z)$ dipende solo da alcune funzioni dei dati $t_1(z), \cdots, t_k(z)$, queste rappresentano le uniche funzioni di cui abbiamo bisogno per aggiornare l'informazione. L'osservazione dell'intero campione, o delle funzioni $t_j(\cdot)$, è equivalente. Possiamo formalizzare quanto detto in una

Definizione 1 *La funzione* $T = t(X_1, \cdots, X_n)$ *si dice un* **riassunto esaustivo** *per il modello*

$$\mathcal{E} = \{\mathcal{X}, \mathcal{F}, \mathcal{P}\}, \mathcal{P} = \{p(\cdot \mid \theta), \theta \in \Theta\}$$

se, per qualunque distribuzione iniziale $h(\theta)$ *si ha*

$$h(\theta \mid x_1, \cdots, x_n) = h(\theta \mid t),$$

per ogni realizzazione $z = (x_1, \cdots, x_n)$ *tale che* $t(x_1, \cdots, x_n) = t$.

Un modo diverso per dire la stessa cosa è

$$\Theta \perp (X_1, \cdots, X_n) \mid T.$$

Esempio [Poisson]: Siano $X_1, \cdots, X_n \mid \theta$ i.i.d. Poisson(θ), cioè

$$p(X_i = k \mid \theta) = \frac{e^{-\theta}\theta^k}{k!}$$

e sia $h(\theta)$ una generica distribuzione iniziale. La verosimiglianza è

$$L(\theta; z) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta_i^x}{x_i!} \propto e^{-n\theta}\theta^{\sum x_i} = e^{-n\theta}\theta^t$$

dove $t = \sum x_i$.
La distribuzione finale è

$$h(\theta \mid z) = \frac{h(\theta)\theta^t \exp\{-n\theta\}}{\int_\Theta h(\theta)\theta^t \exp\{-n\theta\}},$$

che dipende dai dati solo attraverso t

Questa definizione non è operativa: esiste però una versione modificata del Criterio di fattorizzazione di Neyman

Teorema 1 *T è un riassunto esaustivo per $\mathcal{E}$ se e solo se la densità congiunta $p(z \mid \theta)$ può essere fattorizzata nel modo seguente:*

$$p(x_1, \cdots, x_n \mid \theta) = g(t, \theta) K(x_1, x_2, \cdots, x_n), \qquad \forall \theta \in \Theta \text{ e } \forall x_1, \cdots, x_n \in \mathrm{R}_n$$

dove $t = t(x_1, \cdots, x_n)$, e le funzioni $g(t, \theta) \geq 0$ e $K(x_1, x_2, \cdots, x_n) \geq 0$ non sono univocamente determinate, g è funzione dei dati solo attraverso θ e K non dipende da θ.

No dimostrazione

Esempio [Distr. uniforme]: Siano $X_1, \cdots, X_n \mid \theta$ i.i.d. U(θ), cioè

$$p(x_i \mid \theta) = \frac{1}{\theta}, \qquad 0 \geq x_i \geq \theta.$$

La verosimiglianza è

$$L(\theta) \propto \frac{1}{\theta^n} I_{[T,\infty]}(\theta),$$

dove $T = \max(X_1, \cdots, X_n)$: perciò T è un riassunto esaustivo per θ.

Scelta della distribuzione iniziale

La scelta della iniziale è prettamente soggettiva. Questo non esclude che possano essere suggeriti dei criteri generali per arrivare all'elicitazione di una singola distribuzione che rappresenti in modo formale le nostre informazioni sul parametro. Cominciamo ad elencare alcuni metodi **soggettivi**

1. Metodo dell'istogramma
2. Metodo del confronto relativo
3. Specifica forma funzionale

- Istogramma

Si suddivide lo spazio Θ in sottoinsiemi e ci chiediamo quale probabilità assegnare ai singoli intervalli. Questo equivale all'elicitazione di alcuni quantili della distribuzione $h(\theta)$. In seguito occorre scegliere una forma funzionale che sia il più possibile coerente con la precedente elicitazione

- Confronto relativo

Sia ad esempio $\Theta = [0, 1]$. Ci chiediamo quali siano il valore più probabile e il meno probabile di Θ; supponiamo che siano $\theta_+ = 5/6$ e $\theta_- = 0$, e che inoltre θ_+ sia tre volte più probabile di θ_-. Continuiamo poi a confrontare altri punti fino a ottenere un grafico che possa poi essere approssimato da una data forma funzionale.

- Forma funzionale

Si sceglie una forma funzionale specifica e si determinano i k parametri di tale distribuzione attraverso l'imposizione di vincoli su

- k parametri
- k momenti

Distribuzioni coniugate

Siano $X_1, X_2, \cdots, X_n$, n v. aleatorie i.i.d. dato $\theta \in \Theta$. Assumiamo che le v.a. siano dotate di densità o distribuzione di probabilità indicata con $p(x \mid \theta)$. La verosiglianza per θ è

$$L(\theta) \propto \prod_{j=1}^{n} p(x_j \mid \theta).$$

Una distribuzione di probabilità iniziale $h(\theta)$ si dice coniugata al modello utilizzato o, equivalentemente, alla verosimiglianza $L(\theta)$, se la forma funzionale della distribuzione iniziale e della distribuzione finale sono uguali.

Esempio [Beta-binomiale]**:** Abbiamo già visto nel capitolo precedente che, dato un modello bernoulliano, l'uso di una distribuzione iniziale di tipo Beta(α, β) implica che la distribuzione finale sia ancora di tipo Beta con parametri modificati dalle osservazioni campionarie, cosicché $\alpha^* = \alpha + k$ e $\beta^* = \beta + n - k$, dove k è il numero di successi nelle n prove effettuate.

Esempio [Esponenziale-Gamma]: Siano $X_1, \cdots, X_n$ n i.i.d. Exp(λ), ovvero, per $j = 1, \cdots, n$

$$f(x_j \mid \lambda) = \lambda \exp\{-\lambda x_j\}.$$

La verosimiglianza è

$$L(\lambda) \propto \lambda^n \exp\left\{-\lambda \sum_{j=1}^{n} x_j\right\},$$

e una espressione per la distribuzione iniziale di λ che sia coniugata con $L8\theta)$ è data dalla distribuzione di tipo Gamma(α, ν) con densità

$$h(\lambda) = \frac{\alpha^\nu}{\Gamma(\nu)} \exp\{-\alpha\lambda\} \lambda^{\nu-1}.$$

Si vede facilmente che la distribuzione a posteriori risultante

è proporzionale a

$$h(\lambda \mid \mathbf{x}) \propto \lambda^{n+\nu-1} \exp\{-\lambda(\alpha + n\bar{x})\}.$$

Dunque $h(\lambda \mid \mathbf{x})$ è ancora di tipo Gamma(α^*, ν^*), con parametri aggiornati dall'informazione sperimentale in

$$\alpha^* = \alpha + n\bar{x} \quad \text{e} \quad \nu^* = \nu + n.$$

Alcuni commenti sulle distrib. coniugate

- Elicitazione limitata ai parametri
- In genere è possibile ottenere distribuzioni non informative per particolari scelte dei parametri.

Distribuzioni non informative

E' molto frequente nelle applicazioni l'uso di distribuzioni *improprie*, ovvero tali che

$$\int_{\Theta} h(\theta) d\theta = \infty;$$

In principio, tali leggi non dovrebbero essere consentite dalle regole del calcolo delle probabilità . Tuttavia ci sono diverse ragioni che conducono alla determinazione di distribuzioni di tipo *convenzionale*

- Adottare uno spirito frequentista all'interno di uno schema bayesiano (compromesso fondazionale)
- Formalizzazione matematica di uno **stato di ignoranza** su θ
- Approssimazione di vere distribuzioni di probabilità

Il metodo più generale per la costruzione di distribuzioni non informative è quello proposto da Jeffreys nel suo testo del 1961. Il suo risultato generale è il seguente

$$h(\theta) \propto \sqrt{\det I(\theta)}.$$

Le motivazioni complete per arrivare a tale risultato sono però complesse.

Teoria delle distribuzioni noninformative

Consideriamo

$$\mathcal{E}_k = (\mathcal{X}_k, \mathbf{\Theta}, \mathcal{P}) \tag{1}$$

il consueto esperimento statistico. dove Θ è lo spazio dei parametri. Con il simbolo $\pi^N(\theta)$ indicheremo una generica distribuzione non informativa su $\mathbf{\Theta}$.

In questo corso considereremo soltanto il caso in cui il problema inferenziale sia un problema di stima. La determinazione di distribuzioni convenzionali per problemi di verifica di ipotesi e scelta del modello statistico è un problema più complesso.

La distribuzione uniforme

Evoluzione del concetto Consideriamo inizialmente la situazione sperimentale più semplice in cui lo spazio parametrico sia composto da un numero finito di possibili valori, cioè

$$\boldsymbol{\Theta} = \{\theta_1, \theta_2, \cdots, \theta_H\}$$

In tal caso è intuitivo ritenere che una distribuzione iniziale non informativa debba pesare allo stesso modo le H possibili alternative cosicché

$$h^N(\theta_j) = H^{-1}, \quad j = 1, \cdots, H.$$

L'intuizione è stata guidata dal *Principio di ragione Insufficiente* (PRI). Nonostante la semplicità del contesto, anche questo problema nasconde delle insidie: se ω_1 viene suddiviso in due diversi valori, diciamo θ_{11} e θ_{12}, la

cardinalità di $\boldsymbol{\Theta}$ diviene $H+1$ e, di conseguenza, il PRI indurrebbe ad utilizzare una distribuzione a priori uniforme sugli $H+1$ valori, modificando totalmente la legge iniziale.

La situazione si complica ancora di più quando $\boldsymbol{\Omega}$ ha una cardinalità numerabile o addirittura superiore.

Esempio []: k replicazioni $(X_1, \cdots, X_k)$ di una v.a.

$$X \sim Be(\omega), \quad \omega \in [0,1],$$

In questo esempio lo spazio parametrico, pur avendo potenza del continuo, è compatto e una naturale estensione del PRI porterebbe a supporre che la legge iniziale di default per ω sia uniforme su $\mathbf{\Theta}$, ovvero

$$\pi^N(\theta) = 1 \; I_{[0,1]}(\theta).$$

Sebbene tale scelta sia ragionevole, in diverse impostazioni si preferiranno distribuzioni differenti.

Il problema essenziale legato all'uso della legge uniforme (costante) su un sottoinsieme di R è la mancanza di invarianza rispetto a riparametrizzazioni bigettive (uno a uno) di θ. **Esempio ??** (continua). In alcune applicazioni, il parametro di interesse potrebbe non essere θ bensì una trasformazione, ad esempio, $\lambda = -\log\theta$: la legge uniforme su θ induce, per via dello Jacobiano, la seguente legge sul nuovo parametro λ:

$$\pi^N(\lambda) = \exp\{-\lambda\}\ I_{[0,\infty)}(\lambda),$$

distribuzione esponenziale di parametro 1. Di contro una legge uniforme su λ non ricondurrebbe alla legge uniforme su θ.

Quali rimedi adottare allora nel caso in cui $\boldsymbol{\Theta}$ è illimitato? Ne riparleremo: intanto

- I difensori e gli utilizzatori del PRI sostengono che, prima di poter utilizzare tale principio, occorre scegliere con cura la cardinalità e la parametrizzazione rispetto alle quali sarà poi ragionevole adottare una distribuzione uniforme.

- Il fatto che la legge di probabilità sia impropria non è, in genere, un vereo problema. Leggi improprie sono teoricamente giustificabili in termini di additività finita. L'importante è verificare che la corrispondente distribuzione a posteriori ottenuta mediante applicazione del teorema di Bayes risulti propria, qualunque sia il risultato campionario osservato!!

Il metodo di Jeffreys

La mancanza di invarianza della legge uniforme rispetto a trasformazioni dei parametri indusse Jeffreys a formulare un nuovo criterio di costruzione di distribuzioni a priori non informative. La definizione è piuttosto semplice: dato il modello statistico di base, la distribuzione non informativa di Jeffreys è

$$h^J(\theta) \propto \sqrt{\det(I(\theta))}, \tag{2}$$

dove $I(\theta)$ rappresenta la matrice d'informazione attesa di Fisher relativa ad una singola osservazione

$$H(\theta) = \left\{ H_{a,b} = -\,\mathrm{E}_\omega \left(\frac{\partial^2}{\partial\theta_a \partial\theta_b} \log p(x|\theta \right) \right\},$$

$a, b = 1, \cdots, d$ L'uso di h^J è ovviamente subordinato all'esistenza di tale matrice e al suo essere definita positiva.

Esempio (continua). La densità relativa ad una singola osservazione è

$$p(x \mid \theta) = \theta^x (1-\theta)^{1-x} I_{(0,1)}(x);$$

ne segue che

$$\begin{aligned}
&\log p(x \mid \theta) = x \log \theta + (1-x) \log(1-\theta); \\
&\frac{\partial}{\partial \theta} \log p(x \mid \theta) = \frac{x}{\theta} - \frac{1-x}{1-\theta}; \\
&-\frac{\partial^2}{\partial \theta^2} \log p(x \mid \theta) = \frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}; \\
&-\mathrm{E}_\omega \left(\frac{\partial^2}{\partial \theta^2} \log p(x|\theta \right) = \frac{1}{\theta} - \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)};
\end{aligned}$$

la legge di Jeffreys è dunque

$$h^J(\theta) = \frac{1}{\pi}\theta^{-1/2}(1-\theta)^{-1/2}. \qquad (3)$$

Si tratta di una distribuzione di tipo Beta$(\frac{1}{2}, \frac{1}{2})$, simmetrica e dalla caratteristica forma ad U. La distribuzione (3) è invariante per riparametrizzazioni biunivoche di θ: sia $\lambda = \lambda(\theta)$ tale che $I(\lambda)$ esiste ed è definita positiva, allora,

$$h^J_\theta(\theta) = h^J_\lambda(\lambda(\theta)) \mid \det(\frac{\partial\lambda}{\partial\theta}) \mid$$

L'idea di Jeffreys fu quella di generalizzare il concetto di invarianza già popolare nel caso di modelli di posizione (o di scala, o di posizione e scala) come il modello gaussiano con μ e/o σ parametri incogniti.

Esempio [Parametro di posizione]: Siano $X_1, \cdots, X_n \sim f(x-\mu)$ e indipendenti.
La famiglia $\{f_\mu, \mu \in \mathrm{R}\}$ è invariante per traslazione, nel senso che $Y = X + a$ ha una distribuzione ancora del tipo f, qualunque sia a.
μ è allora un parametro di posizione e come tale si richiede una "naturale" condizione d'invarianza

$$\pi(\mu) = \pi(\mu - a), \quad \forall a \in \mathrm{R},$$

che ha soluzione $\pi(\theta) =$ cost.

Derivazione della legge di Jeffreys

Dato $\mathcal{E} = \{\mathcal{X}, \mathcal{P}, \mathbf{\Theta}\}$, consideriamo la **distanza simmetrica di Kullback-Leibler** tra due elementi di $\mathcal{E}$, indicizzati da θ e θ':

$$J(\theta, \theta') = \int_{\mathcal{X}} \log \frac{p(x|\theta')}{p(x|\theta)} (p(x|\theta') - p(x|\theta)) dx.$$

È facile verificare che, per $\theta' \to \theta$, la distanza simmetrica di Kullback e Leibler può essere approssimata dalla metrica di Fisher, cosicché

$$\frac{J(\theta, \theta')}{(\theta' - \theta)^2} \to H(\theta).$$

Inoltre, se λ rappresenta un'altra parametrizzazione biunivoca di θ, risulta

$$H(\theta) = H(\lambda(\theta)) \left(\frac{d\lambda}{d\theta} \right)^2,$$

ovvero scritto in una forma alternativa,

$$\sqrt{H(\theta)}d\theta = \sqrt{H(\lambda)}d\lambda.$$

La distribuzione di Jeffreys

$$h_J(\theta) \propto \sqrt{H(\theta)},$$

dunque,**pesa** sottoinsiemi di $\mathcal{E}$ in modo proporzionale al valore della matrice H, indipendentemente dalla parametrizzazione adottata.

In altri termini, la metrica naturale per il modello $\mathcal{E}$ non è quella euclidea bensì quella indotta dalla matrice di informazione H, e la distribuzione iniziale di Jeffreys può essere interpretata come la distribuzione uniforme su $\mathcal{E}$, secondo tale metrica. Il metodo di Jeffreys è ancora oggi quello più comunemente utilizzato quando $d = 1$.
Tuttavia lo stesso Jeffreys suggerì alcune modifiche alla sua regola generale nel caso di parametro multidimensionale.