11 Il metodo della massima verosimiglianza

Supponiamo di avere di fronte a noi due urne, una con 90 palline bianche e 10 nere, e l'altra con 10 nere e 90 bianche; identifichiamo le urne con la proporzione di palline bianche che contengono (denotata con p), in modo tale che la prima urna corrisponda a $p = .9$ e la seconda a $p = .1$. Viene estratta una singola pallina e non sappiamo da quale urne venga; osserviamo però che la pallina è bianca. Dovendo cercare di risalire all'urna di provenienza, sembra naturale scegliere quella che rende più probabile osservare quello che abbiamo effettivamente osservato, cioè l'urna che contiene la maggiore proporzione di palline bianche; in altre parole, dall'osservazione "è stata estratta una pallina bianca" sembra naturale far discendere lo stimatore $\widehat{p} = 0.9$.

Supponiamo ora che invece di una sola pallina ne siano estratte 5, e che risultino essere 3 bianche e 2 nere; immaginiamo altresì che l'alternativa non sia solo tra due urne, ma tra un continuo di urne con tutte le possibili proporzioni di palline bianche tra 0 ed 1, identificate sempre con p. In questo caso, la probabilità di osservare 3 bianche e 2 nere è data da una legge binomiale:

$$\Pr\{\text{3 bianche e 2 nere}\} = \binom{5}{3} p^3 (1-p)^2,$$

ed è facilmente verificabile che questa probabilità è massimizzata prendendo $\widehat{p} = \frac{3}{5}$.

Questo esempio molto semplice dovrebbe aiutare a capire l'idea che è alla base degli stimatori di massima verosimiglianza - si tratta di costruire la funzione di probabilità (o di densità, nel caso continuo) relativa ad un certo campione aletaorio, e vederla quindi non più come funzione del campione, ma come funzione (aleatoria) dei parametri, prendendo come date le osservazioni. In altre parole, si ha la seguente definizione:

Definition 92 *(Funzione di verosimiglianza) Sia $X_1, ..., X_n$ un campione aleatorio con legge (=funzione di probabilità o densità, a seconda del caso discreto o continuo) f_θ, $\theta \in \mathbb{R}^p$. La funzione di verosimiglianza $L : \mathbb{R}^p \to \mathbb{R}$ è definita da*

$$L(\theta; X_1, ..., X_n) := \prod_{i=1}^{n} f(X_i; \theta) = f(X_1, ..., X_n; \theta) \ .$$

Remark 93 *Con un leggero abuso di notazione, a sinistra della prima uguaglianza abbiamo scritto f sia per la legge univariata che per qualla congiunta; inoltre usiamo $f(.)$ sia per il caso discreto che per quello continuo. Il punto più importante è però un altro: la verosimiglianza è funzione del parametro prendendo i valori di $X_1, ...X_n$ come dati, mentre la densità congiunta prende il parametro θ come dato ed è funzione dei possibili valori delle osservazioni $x_1, ..., x_n$.*

Remark 94 *Ovviamente si ha (nel caso continuo)*

$$\int_{\mathbb{R}^n} f(x_1, ..., x_n; \theta) dx_1 ... dx_n = 1 \ ;$$

non c'è però nessun motivo di aspettarsi l'uguaglianza

$$\int_{\mathbb{R}^p} L(\theta; X_1, ..., X_n) d\theta \stackrel{???}{=} 1 \ ,$$

anzi in generale questo integrale può divergere. La funzione di verosimiglianza pertanto NON identifica una densità di probabilità.

Example 95 *(Valor medio nel caso Gaussiano) Sia $X_1, ..., X_n$ un campione aleatorio estratto da una legge Gaussiana $N(\mu, 1)$; è immediato verificare che la funzione di verosmiglianza prende la forma*

$$L(\mu; X_1, ..., X_n) = \frac{1}{(2\pi)^{n/2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)^2 \right\} \ .$$

In questo caso (e in quelli a seguire) è conveniente lavorare con il logaritmo della funzione di verosimiglianza; trattandosi di una trasformazione monotona crescente, il valore che massimizza la funzione (cioè l'argmax) non cambia, e pertanto nemmeno lo stimatore. Si ha in particolare

$$\begin{aligned} \ell(\mu; X_1, ..., X_n) &: = \log L(\mu; X_1, ..., X_n) \\ &= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)^2, \end{aligned}$$

funzione che è evidentemente massimizzata in

$$\widehat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} X_i \ .$$

Example 96 *(Variabili Bernoulliane) Sia $X_1, ..., X_n$ un campione aleatorio di variabili $Ber(p)$; la probabilità di k successi per $k = 0, ..., n$ è evidentemente data da $p^k (1-p)^{n-k}$, e la verosimiglianza è pertanto*

$$L(\mu; X_1, ..., X_n) = p^{\sum_{i=1}^n X_i} (1-p)^{n - \sum_{i=1}^n X_i}.$$

La massimizzazione della log-verosimiglianza ci fa ottenere facilmente

$$\begin{aligned} \frac{\partial \log L(\mu; X_1, ..., X_n)}{\partial p} &= (\sum_{i=1}^{n} X_i) \frac{\partial}{\partial p} \log p + (n - \sum_{i=1}^{n} X_i) \frac{\partial}{\partial p} \log(1-p) \\ &= \frac{\sum_{i=1}^n X_i}{p} - \frac{(n - \sum_{i=1} X_i)}{1-p} \ ; \end{aligned}$$

imponendo che la derivata sia pari a zero, si vede facilmente che

$$\widehat{p}_{MLE} = \frac{1}{n} \sum_{i=1} X_i \ .$$

Remark 97 *Si sarebbe potuto scrivere la verosimiglianza come*

$$L(\mu; X_1, ..., X_n) = \binom{n}{\sum_{i=1}^n X_i} p^{\sum_{i=1}^n X_i} (1-p)^{n-\sum_{i=1}^n X_i}.$$

Quale è la versione "giusta" della verosimiglianza? Come sarebbero cambiate le stime utilizzando questa seconda versione? Si tratta di un risultato generale o c'è un principio generale sottostante?

11.0.1 Consistency and Asymptotic Gaussianity

To study consistency, we need first to introduce the Kullback-Leibler distance between two probility densities/probability mass function. This is defined by

Definition 98 *Siano f, g funzioni di densità o funzioni di probabilità. La distanza di Kullback-Leibler tra f è definita da*

$$D(f,g) = \int \log(\frac{f(x,\theta)}{g(x,\theta)}) f(x;\theta) dx$$

se f è assolutamente continua rispetto a g; altrimenti la distanza viene posta pari a $+\infty$. Analogamente, nel caso discreto

$$D(p,q) = \sum_{x_i} \log(\frac{p(x_i,\theta)}{q(x_i,\theta)}) p(x_i;\theta) \ .$$

Remark 99 *E' facilmente verificabile che la distanza di Kullback-Leibler è sempre non-negativa; infatti, utilizzando la disuguaglianza di Jensen*

$$\begin{aligned} D(f,g) &= E_f[\log \frac{f(X)}{g(X)}] = -E_f[\log \frac{g(X)}{f(X)}] \\ &\geq \log\left(\int \frac{g(x,\theta)}{f(x,\theta)}) f(x;\theta) dx\right) = \log\left(\int g(x,\theta) dx\right) = 0 \ . \end{aligned}$$

D'altra parte la distanza di Kullback-Leibler non identifica una metrica, ad esempio perché non è simmetrica negli argomenti (e non vale la disuguaglianza triangolare).

Quale è la relazione tra funzione di verosimiglianza e distanza di Kullback-Leibler? Focalizziamoci come al solito sulla log-verosimiglianza; considerando che la stimatore è invariante se la funzione è trasformata linearmente con coefficienti che non dipendono dal parametro, possiamo passare da $\log L = \sum_{i=1}^n \log f(X_i;\theta)$ a

$$M_n(\theta) := \frac{1}{n}\sum_{i=1}^n \log f(X_i;\theta) - \frac{1}{n}\sum_{i=1}^n \log f(X_i;\theta_0) = \frac{1}{n}\sum_{i=1}^n \log \frac{f(X_i;\theta)}{f(X_i;\theta_0)} \ ,$$

dove θ_0 denota il "vero" valore del parametro. Sotto alcune delle condizioni di regolarità che abbiamo visto precedentemente affinchè valga la legge dei grandi numeri, abbiamo che, *ad un valore di θ fissato:*

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\log\frac{f(X_i;\theta)}{f(X_i;\theta_0)} &\rightarrow {}_pE[\log\frac{f(X_1;\theta)}{f(X_1;\theta_0)}] \\ &= \int\log\frac{f(x;\theta)}{f(x;\theta_0)}f(x;\theta_0)dx \\ &= -D(f_{\theta_0},f_\theta)\ .\end{aligned}$$

L'idea euristica dietro alla dimostrazione della consistenza è dunque la seguente: $M_n(\theta)$ converge a $-D(f_{\theta_0},f_\theta)$, che è sempre negativa tranne quando $\theta=\theta_0$, dove vale zero. Poiché lo stimatore di massima verosimiglianza è definito come il valore di θ che massimizza $M_n(\theta)$, a sua volta esso convergerà al valore che massimizza $-D(f_{\theta_0},f_\theta)$, cioè $\theta=\theta_0$. Nel seguito, scriviamo per brevità $M(\theta)=-D(f_{\theta_0},f_\theta)$; abbiamo la seguente proposizione:

Proposition 100 *(Consistenza) Assumiamo che valgano le seguenti condizioni di regolarità:*

a) (Legge dei grandi numeri uniforme) Si ha che

$$\sup|M_n(\theta)-M(\theta)|=o_p(1)\ \textit{per}\ n\rightarrow\infty\ ;$$

b) (Convessità) Si ha che per ogni $\varepsilon>0$ esiste $\eta_\varepsilon>0$ tale che

$$|\theta-\theta_0|>\varepsilon\Longrightarrow M(\theta_0)-M(\theta)>\eta_\varepsilon\ .$$

Allora $\widehat{\theta}_{MLE;n}\rightarrow_p\theta_0$.

Remark 101 *L'ipotesi a) è più forte delle leggi dei grandi numeri che abbiamo dimostrato in precedenza, perché richiede che la convergenza sia uniforme in θ, il che non è garantito anche se la convergenza avviene puntualmente per ogni valore fisso di θ. Le leggi dei grandi numeri uniformi sono ampiamente trattate nei corsi della magistrale (Teoremi di Glivenko-Cantelli) ma non fanno parte del programma di questo corso. L'ipotesi b) è essenzialmente una condizione di identificabilità: se esistesse un insieme di valori diversi di θ a distanza di Kullback-Leibler 0 da θ_0 il problema della consistenza sarebbe evidentemente irrisolvibile (e lo stimatore potrebbe non essere nemmeno ben definito).*

Proof. L'idea è di mostrare che per ogni $\varepsilon>0$ fissato si ha $\Pr\left\{M(\theta_0)-M(\widehat{\theta}_n)>\eta_\varepsilon\right\}\rightarrow 0$; grazie alla condizione di convessità abbiamo

$$\Pr\left\{|\theta_0-\widehat{\theta}_n|>\varepsilon\right\}\leq\Pr\left\{M(\theta_0)-M(\widehat{\theta}_n)>\eta_\varepsilon\right\}$$

e conseguentemente la proposizione sarà dimostrata. A questo proposito notiamo che, per definizione di stimatore di massima verosimiglianza

$$M_n(\widehat{\theta}_n)-M_n(\theta_0)>0$$

e quindi

$$\begin{aligned} M(\widehat{\theta}_n) - M(\theta_0) &= M_n(\theta_0) - M(\widehat{\theta}_n) + M(\theta_0) - M_n(\theta_0) \\ &\leq M_n(\widehat{\theta}_n) - M(\widehat{\theta}_n) + M(\theta_0) - M_n(\theta_0) \ . \end{aligned}$$

Segue quindi che

$$\begin{aligned} \Pr\left\{ M(\theta_0) - M(\widehat{\theta}_n) > \eta_\varepsilon \right\} &\leq \Pr\left\{ |M(\theta_0) - M_n(\theta_0)| > \frac{\eta_\varepsilon}{2} \right\} + \Pr\left\{ |M_n(\widehat{\theta}_n) - M(\widehat{\theta}_n)| > \frac{\eta_\varepsilon}{2} \right\} \\ &\to 0 \ , \text{ per } n \to \infty \ , \end{aligned}$$

usando per il primo termine la legge dei grandi numeri standard e per il secondo quella uniforme (che implica ovviamente quella standard). ■

11.0.2 La funzione punteggio

Definition 102 *(Funzione Punteggio) La funzione punteggio (score function in inglese) è definita da*

$$s_n(\theta; X_1, ..., X_n) = \frac{\partial}{\partial \theta} \log L(\theta; X_1, ..., X_n) \ ,$$

assumendo ovviamente che la derivata esista. In genere, si tratta di una funzione da $\mathbb{R}^p$ in $\mathbb{R}^p$.

Nella discussione che segue inizieremo dal caso $p = 1$. Notiamo innanzitutto che la funzione punteggio è essa stessa una variabile aleatoria, per ogni valore di θ fissato. Ha quindi senso domandarsi quale sia il suo valor medio e la sua varianza, ammesso che esistano. Iniziamo quindi a imporre alcune condizioni di regolarità.

Definition 103 *(Condizioni di regolarità) Valgono le seguenti ipotesi:*

- *Le condizioni di consistenza*
- *Le osservazioni $X_1, ...X_n$ sono indipendenti ed identicamente distribuite*
- *La log-verosimiglianza ammette due derivate continue, $\frac{\partial^2}{\partial \theta^2} \log L_n \in C^2$*
- *La derivata seconda ha momento secondo finito, $E[|\frac{\partial^2}{\partial \theta^2} \log L_n|] < \infty$*
- *Il vero valore del parametro θ_0 appartiene all'interno dell'insieme dei valori ammissibili, $\theta_0 \in Int(\Theta)$*
- *E' possibile scambiare due volte la derivata con l'integrale nel valor medio della log-verosimiglianza*

Remark 104 *Queste condizioni sono più forti del necessario per qullo che riguarda i risultati sulla funzione punteggio, ma preferiamo enunciarle una volta per tutte in funzione del risultato sulla asinotitca Gaussianità degli stimatori ML.*

Lemma 105 *Sotto le condizioni di regolarità, abbiamo che*

$$E_{\theta_0}[s_n(\theta_0; X_1, ..., X_n)] = 0 \ .$$

Proof. Per definizione abbiamo che

$$\begin{aligned} E_{\theta_0}[s_n(\theta_0; X_1, ..., X_n)] &= \int \left\{\log' L(\theta_0; x_1, ..., x_n)\right\} f(x_1, ..., x_n; \theta_0) dx_1 ... dx_n \\ &= \int \frac{\frac{\partial f(x_1,...,x_n;\theta_0)}{\partial \theta}\Big|_{\theta=\theta_0}}{f(x_1, ..., x_n; \theta_0)} f(x_1, ..., x_n; \theta_0) dx_1 ... dx_n \\ &= \int \left.\frac{\partial f(x_1, ..., x_n; \theta_0)}{\partial \theta}\right|_{\theta=\theta_0} dx_1 ... dx_n \ . \end{aligned}$$

Poiché per ipotesi possiamo scambiare derivate ed integrali, abbiamo che

$$\begin{aligned} \int \left.\frac{\partial f(x_1, ..., x_n; \theta_0)}{\partial \theta}\right|_{\theta=\theta_0} dx_1 ... dx_n &= \left.\frac{\partial}{\partial \theta} \int f(x_1, ..., x_n; \theta) dx_1 ... dx_n\right|_{\theta=\theta_0} \\ &= \left.\frac{\partial}{\partial \theta} 1\right|_{\theta=\theta_0} = 0 \ . \end{aligned}$$

■

Example 106 *Se consideriamo il caso della verosimiglianza Gaussiana, abbiamo facilmente (assumendo varianza nota e pari a 1, X_i iid $N(\mu_0, 1)$)*

$$s_n(\mu; X_1, ..., X_n) = \sum_{i=1}^{n} (X_i - \mu) \ ,$$

ed è evidentemente che il valor medio si annulla prendendo $\mu = \mu_0$:

$$E[s_n(\mu; X_1, ..., X_n)] = n(\mu - \mu_0) \ .$$

Lemma 107 *Sotto le condizioni di regolarità, abbiamo che*

$$\begin{aligned} & E\left[\left.\frac{\partial^2 \log L(\theta; x_1, ..., x_n)}{\partial \theta^2}\right|_{\theta=\theta_0}\right] \\ = & \int \left\{\log'' L(\theta_0; x_1, ..., x_n)\right\} f(x_1, ..., x_n; \theta_0) dx_1 .... dx_n \\ = & -\int \left\{\log' L(\theta_0; x_1, ..., x_n)\right\}^2 f(x_1, ..., x_n; \theta_0) dx_1 .... dx_n \\ = & -E\left[\left\{\left.\frac{\partial \log L(\theta; x_1, ..., x_n)}{\partial \theta}\right|_{\theta=\theta_0}\right\}^2\right] \ . \end{aligned}$$

Proof. Si noti innanzitutto che, per qualsiasi valore di θ

$$H(\theta) := \int \left\{\log' L(\theta; x_1, ..., x_n)\right\} f(x_1, ..., x_n; \theta) dx_1 ....dx_n \equiv 0 \ .$$

Di nuovo per le condizioni di regolarità possiamo allora considerare

$$\begin{aligned}
0 &= \frac{\partial H(\theta)}{\partial \theta} = H'(\theta) \\
&= \frac{\partial}{\partial \theta} \int \left\{\log' L(\theta; x_1, ..., x_n)\right\} f(x_1, ..., x_n; \theta) dx_1 ....dx_n \\
&= \int \left\{\log'' L(\theta; x_1, ..., x_n)\right\} f(x_1, ..., x_n; \theta) dx_1 ....dx_n \\
&\quad + \int \left\{\log' L(\theta; x_1, ..., x_n)\right\} f'(x_1, ..., x_n; \theta) dx_1 ....dx_n \\
&= \int \left\{\log'' L(\theta; x_1, ..., x_n)\right\} f(x_1, ..., x_n; \theta) dx_1 ....dx_n \\
&\quad + \int \left\{\log' L(\theta; x_1, ..., x_n)\right\} \frac{f'(x_1, ..., x_n; \theta)}{f(x_1, ..., x_n; \theta)} f(x_1, ..., x_n; \theta) dx_1 ....dx_n \\
&= \int \left\{\log'' L(\theta; x_1, ..., x_n)\right\} f(x_1, ..., x_n; \theta) dx_1 ....dx_n \\
&\quad + \int \left\{\log' L(\theta; x_1, ..., x_n)\right\}^2 f(x_1, ..., x_n; \theta) dx_1 ....dx_n \ ,
\end{aligned}$$

ed il risultato è dimostrato. ■

Remark 108 *E' importante notare che in $\theta = \theta_0$*

$$\begin{aligned}
&\int \left\{\log' L(\theta_0; x_1, ..., x_n)\right\}^2 f(x_1, ..., x_n; \theta_0) dx_1 ....dx_n \\
= \ & E[(s_n(\theta_0))^2] = Var(s_n(\theta_0)) \ .
\end{aligned}$$

La varianza della funzione punteggio calcolata nel valore vero del parametro è pertanto uguale al reciproco della derivata seconda della log-verosimiglianza nello stesso punto.

Definition 109 *(Informazione di Fisher) La varianza della funzione punteggio è nota come informazione di Fisher e si indica con $I_n(\theta_0)$. In generale, si tratta di una matrice di dimensione $p \times p$.*

11.0.3 L'asintotica Gaussianità degli stimatori MLE

Theorem 110 *Sotto le condizioni di regolarità di cui sopra, si ha che*

$$\sqrt{n}\left\{\widehat{\theta}_n - \theta_0\right\} \rightarrow_d N(0, I_1^{-1}(\theta_0)) \ .$$

Equivalentemente, abbiamo che

$$I_n^{1/2}(\theta_0)\left\{\widehat{\theta}_n - \theta_0\right\} \rightarrow_d N(0, Id_p) \ ,$$

dove Id_p indica la matrice di identità di ordine p.

Proof. Consideriamo per semplicità il caso $p = 1$. Per il teorema del valor medio di Lagrange, esiste $\overline{\theta}_n$ intermedio tra $\widehat{\theta}_n$ e θ_0 tale per cui vale l'uguaglianza

$$0 = \log' L(\widehat{\theta}_n) = \log L'(\theta_0) + \log L''(\overline{\theta}_n)(\widehat{\theta}_n - \theta_0) \ ,$$

da cui

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -\frac{\log L'(\theta_0)/\sqrt{n}}{\log L''(\overline{\theta}_n)/n} \ .$$

Per il numeratore abbiamo una somma di variabili aleatorie IID con valor medio nullo e varianza finita; siamo quindi nel dominio di applicabilità del teorema del limite centrale ed otteniamo

$$\log L'(\theta_0)/\sqrt{n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left. \frac{\partial \log f(X_i;\theta)}{\partial \theta} \right|_{\theta=\theta_0} \rightarrow_d N(0, I_1(\theta_0)) \ .$$

Per il denominatore abbiamo una somma di variabili IID con valor medio finito, quindi per la legge dei grandi numeri su variabili uniformemente integrabili ed il teorema di Slutzky otteniamo

$$\log L''(\overline{\theta}_n)/n = \frac{1}{n} \sum_{i=1}^{n} \left. \frac{\partial^2 \log f(X_i;\theta)}{\partial \theta^2} \right|_{\theta=\overline{\theta}_n} \rightarrow_p I_1(\theta_0) \ .$$

Combinando i due risultati ed usando di nuovo Slutzky si arriva all'enunciato del Teorema. ■

Remark 111 *Questa dimostrazione rende evidente come la consistenza degli stimatori sia un prerequisito necessario perché abbia senso la domanda sulla loro asintotica Gaussianità.*

12 Il limite inferiore di Cramér-Rao

Un risultato notevole riguarda la determinazione della varianza minima degli stimatori non-distorti, sotto condizioni di regolarità; in particolare, emerge che tale varianza minima coincide con quella degli stimatori di massima verosimiglianza, sotto condizioni di regolarità. Per tale motivo, gli stimatori di massima verosimiglianza sotto opportune ipotesi (che coprono gran parte delle distribuzioni di uso comune, almeno nei casi più semplici) risultano essere non solo consistenti ed asinotitcamente Gaussiani, ma anche efficienti in senso assoluto.

Remark 112 *E' evidente che porsi la questione sulla varianza minima ha senso solo per stimatori che siano non-distorti; altrimenti qualsiasi stimatore con valore identicamente costante non potrebbe essere migliorato, avendo varianza identicamente pari a zero.*