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.

Spiegazione Semplice

Cos'è il Limite di Cramér-Rao (CRLB)?
Questo capitolo introduce un concetto fondamentale: l'efficienza.

- Abbiamo visto nel Capitolo 9 che uno stimatore "buono" deve essere *non-distorto* (onesto, la sua media è il valore vero) e avere una *varianza piccola* (preciso, i suoi valori non "ballano" molto).
- La domanda è: "Quant'è piccola la varianza più piccola che possiamo mai sperare di ottenere?"

Il limite di Cramér-Rao (CRLB) è la risposta. È un **limite di velocità teorico** per la precisione. Dice che, per una data distribuzione, *nessun* stimatore non-distorto può essere più preciso (avere una varianza più bassa) di un certo numero. Questo numero è $I_n(\theta_0)^{-1}$, l'inverso dell'Informazione di Fisher (vista nel Cap. 11).
Il testo anticipa che lo stimatore di Massima Verosimiglianza (MLE) è "efficiente in senso assoluto" perché, per n grande, **raggiunge** questo limite di velocità teorico.

Remark 113. *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.*

Spiegazione Semplice

Perché solo "Non-Distorti"?
Questa nota è importante. Se volessimo solo "minimizzare la varianza", avremmo uno stimatore perfetto: $T_n = 5$.
Se il vero valore θ fosse 7, il nostro stimatore $T_n = 5$ sarebbe sempre 5. La sua varianza è 0 (non "balla" affatto). È il più preciso possibile! Ma è anche terribilmente *sbagliato* (è distorto, o "biased").
Il limite di Cramér-Rao si applica solo alla "gara" degli stimatori *onesti* (non distorti): tra tutti quelli che "in media ci azzeccano" (cioè $\mathbb{E}[W_n] = \theta$), qual è il più preciso (con varianza minima)?

Consideriamo ora uno stimatore generico di un parametro $\theta \in \mathbb{R}$,

$$W_n = W_n(X_1, ..., X_n)$$

e supponiamo che la funzione $\psi(\theta) := E_{f_\theta}[W(X_1, ..., X_n)]$ sia di classe C^1 per ogni n; assumiamo inoltre che la densità congiunta f_θ sia tale per cui, per ogni $h \in C^1$ $h : \mathbb{R}^n \to \mathbb{R}$ con $E_\theta[|h(X_1, ..., X_n)|] < \infty$ valga la scambiabilità

$$\frac{d}{d\theta} E_\theta[h(X_1, ..., X_n)] = \int_{\mathbb{R}^n} h(x_1, ..., x_n) \frac{\partial}{\partial \theta} f(x_1, ..., x_n; \theta) dx_1 ... dx_n$$

Chiamiamo ora $b(\theta) = E[W_n] - \theta$ il bias della nostra statistica, e $b'(.)$ la sua derivata rispetto a θ. Abbiamo allora il seguente

Theorem 114 (Cramér-Rao)**.** *Per $\theta \in \mathbb{R}$*

$$Var[W_n] \geq \frac{\{1 + b'(\theta_0)\}^2}{I_n(\theta_0)}$$

(Nota: il testo originale riporta $\{b'(\theta_0)\}^2$, ma la forma standard include $1 + b'(\theta)$ per la stima di θ, derivando $E[W_n] = \theta + b(\theta)$ rispetto a θ).

Remark 115. *Per semplicità, abbiamo enunciato e dimostreremo il teorema solo nel caso $p = 1$. La generalizzazione a p generico è comunque abbastanza semplice; consideriamo lo stimatore W_n come un vettore colonna $p \times 1$, con valor medio $\Psi(\theta)$ che ha matrice Jacobiana $p \times p$ denotata con $J\Psi()$. Allora la matrice di varianza e covarianza di W_n soddisfa la disuguaglianza*

$$Var[W_n] \geq J\Psi(\theta_0) I_n^{-1}(\theta_0) J\Psi(\theta_0)^T.$$

La disuguaglianza va intesa none solito senso di disuguaglianza tra matrici simmetriche e non-negative definite: la loro differenza deve essere non-negativa definita.

Dimostrazione. (caso $p = 1$). L'idea di fondo è utilizzare la disuguaglianza di Cauchy-Schwartz, scrivendola come

$$Var[Y] \geq \frac{Cov^2(X, Y)}{Var[X]}$$

Prendiamo come X la funzione punteggio $\frac{d}{d\theta} \log L(\theta; X_1, ..., X_n)|_{\theta=\theta_0}$; sappiamo che $E[X] = 0$ da cui $Cov(X, Y) = E[XY]$. Prendiamo quindi $Y = W_n$, e osserviamo che

$$\begin{aligned}
E[XY] &= \int_{\mathbb{R}^n} W(x_1, ..., x_n) \frac{d}{d\theta} \log L(\theta; x_1, ..., x_n)|_{\theta=\theta_0} f(x_1, ..., x_n; \theta_0) dx_1 ... dx_n \\
&= \int_{\mathbb{R}^n} W(x_1, ..., x_n) \frac{\frac{d}{d\theta} f(\theta; x_1, ..., x_n)|_{\theta=\theta_0}}{f(x_1, ..., x_n; \theta_0)} f(x_1, ..., x_n; \theta_0) dx_1 ... dx_n \\
&= \int_{\mathbb{R}^n} W(x_1, ..., x_n) \frac{d}{d\theta} f(\theta; x_1, ..., x_n)|_{\theta=\theta_0} dx_1 ... dx_n \\
&= \frac{d}{d\theta} \int_{\mathbb{R}^n} W(x_1, ..., x_n) f(\theta; x_1, ..., x_n) dx_1 ... dx_n|_{\theta=\theta_0} \\
&= \frac{d}{d\theta} E_\theta[W_n]|_{\theta=\theta_0}
\end{aligned}$$

Questo conclude la dimostrazione nel caso con densità; il caso discreto è identico. □

Spiegazione Semplice

Come funziona la dimostrazione (in breve)?
La dimostrazione è un trucco matematico molto elegante che usa la disuguaglianza di Cauchy-Schwarz (vista nel Capitolo 1).
Questa disuguaglianza lega la covarianza tra due variabili Y e X alla loro varianza: $Var[Y] \geq Cov(X, Y)^2 / Var[X]$.
L'idea della prova è scegliere Y e X in modo furbo:

- $Y = W_n$ (lo stimatore che stiamo valutando).
- $X = s_n(\theta_0)$ (la Funzione Punteggio, cioè la derivata della log-verosimiglianza, vista nel Cap. 11).

La prova consiste nel calcolare i due pezzi del lato destro:

1. $Var[X]$: Per definizione (Def. 109), la varianza dello Score è l'Informazione di Fisher, $I_n(\theta_0)$. (Questo è il denominatore).
2. $Cov(X, Y)$: Con un po' di algebra (scambiando derivata e integrale), si dimostra che questa covarianza è uguale alla derivata della media dello stimatore, $\frac{d}{d\theta} \mathbb{E}_\theta[W_n]$. (Questo è il numeratore).

Mettendo insieme i pezzi, si ottiene $Var[W_n] \geq (\frac{d}{d\theta} \mathbb{E}_\theta[W_n])^2 / I_n(\theta_0)$, che è la tesi (nel teorema, $\frac{d}{d\theta} \mathbb{E}_\theta[W_n] = \frac{d}{d\theta}(\theta + b(\theta)) = 1 + b'(\theta)$).

Remark 116. *E' interessante studiare cosa succede nel caso in cui le condizioni di regolarità, ed in particolare la possibilità di scambiare la derivata con l'integrale nel valor medio, non siano soddisfatte. Prendiamo ad esempio un campione di variabili i.i.d., uniformi in* $[0, \theta]$*; si verifica facilmente che la funzione di verosimiglianza prende la forma*

$$L(\theta; X_1, ..., X_n) = \prod_{i=1}^{n} \frac{1}{\theta} I_{[0,\theta]}(X_i) = \frac{1}{\theta^n} I_{[0,\theta]}(X_{(n)})$$

dove $X_{(n)}$ *indica la più grande delle n osservazioni; abbiamo inoltre*

$$\hat{\theta}_{ML;n} = X_{(n)}.$$

Ora, si vede anche facilmente che, per ogni $\epsilon > 0$

$$\begin{aligned} Pr\{\theta_0 - X_{(n)} > \epsilon\} &= \prod_{i=1}^{n} Pr\{\theta_0 - X_i > \epsilon\} \\ &= \{Pr\{\theta_0 - \epsilon > X_1\}\}^n \\ &= \{\frac{1}{\theta_0}(\theta_0 - \epsilon)\}^n = (1 - \frac{\epsilon}{\theta_0})^n. \end{aligned}$$

Poichè queste probabilità sono sommabili, abbiamo che lo stimatore è completamente convergente (e pertanto converge quasi certamente). Inoltre abbiamo che

$$\begin{aligned} Pr\{n(\theta_0 - X_{(n)}) > \epsilon\} &= \{Pr\{\theta_0 - \frac{\epsilon}{n} > X_1\}\}^n \\ &= (1 - \frac{\epsilon}{n\theta_0})^n \to \exp(-\frac{\epsilon}{\theta_0}) \end{aligned}$$

per $n \to \infty$*; abbiamo quindi una forma di superconsistenza, perché la convergenza avviene a velocità* n^{-1} *invece che* $n^{-1/2}$ *come nel teorema precedente. D'altra parte la distribuzione limite è esponenziale invece che Gaussiana. Questo esempio mostra come, quando le condizioni di regolarità vengono a mancare, non necessariamente le proprietà degli stimatori di massima verosimiglianza debbano peggiorare: possono addirittura essere superiori, sia come modalità che come velocità di convergenza.*

Spiegazione Semplice

Cosa succede se le "regole" non valgono? (Esempio di Superconsistenza)
Il limite di Cramér-Rao e l'Asintotica Gaussianità dell'MLE (Cap. 11) funzionano solo sotto "condizioni di regolarità" (essenzialmente, la distribuzione $f(x;\theta)$ deve essere "liscia" e non avere "salti" o "spigoli" che dipendono da θ).
Questo esempio usa la distribuzione Uniforme $U[0, \theta]$. Qui, il punto finale θ *dipende* dal parametro. Le condizioni di regolarità **non** valgono (non si può scambiare derivata e integrale).
Cosa succede?

- **Lo stimatore MLE:** Lo stimatore di massima verosimiglianza per θ è $\hat{\theta}_{ML} = X_{(n)}$, cioè il valore *massimo* osservato nel campione. (Intuitivo: se il massimo che ho visto è 8.5, la mia stima migliore per θ è 8.5).
- **Velocità di convergenza:** Il CLT (Teorema 110) dice che gli stimatori "regolari" convergono con velocità $\sqrt{n}$ (o $n^{-1/2}$). Questo stimatore, invece, converge con velocità n (o n^{-1}). È *molto più veloce*. Questa è chiamata **superconsistenza**.
- **Distribuzione limite:** L'errore "zoomato" $n(\theta_0 - X_{(n)})$ non converge a una Gaussiana, ma a una distribuzione Esponenziale.

Morale: quando le condizioni di regolarità falliscono, il framework "standard" (Gaussiana, limite di Cramér-Rao) crolla, e possono succedere cose diverse, a volte (come in questo caso) persino migliori.