1 Preliminari

- Sia $E[x] = \mu$ e $Var(x) = E[(x-\mu)(x-\mu)^T]$. Queste sono la **media vettoriale** e la **varianza matriciale**. Poichè la **traccia** ha la proprietà $tr(AB) = tr(BA)$, si ottiene facilmente la varianza scalare come $\sigma^2 = E[||x-\mu||^2] = E[(x-\mu)^T(x-\mu)] = tr(E[(x-\mu)^T(x-\mu)]) = tr(E[(x-\mu)(x-\mu)^T] = tr(Var(x))$. In maniera analoga si ha che $tr(E[xx^T]) = E[||x||^2]$. La varianza matriciale dà informazioni anche sulla **varianza delle singole componenti** del vettore aleatorio, tramite i suoi elementi diagonali, mentre gli elementi non-diagonali rappresentano la correlazione tre le varie componenti

- Si ha (a costante) $E[(x-a)(x-a)^T] = E[(x-\mu+\mu-a)(x-\mu+\mu-a)^T] = E[(x-\mu)(x-\mu)^T] + (\mu-a)(\mu-a)^T$ essendo nulla l'aspettazione dei due termini incrociati (essendo nulla $E[x-\mu]$). Quindi $E[(x-a)(x-a)^T] = Var(x) + (\mu-a)(\mu-a)^T$ che si traduce anche nella versione scalare $E[||x-a||^2] = Var_{scalare}(x) + ||\mu-a||^2$. Nel caso particolare di una v.a. x e di $a=0$ essa implica la nota formula $E[x^2] = \sigma^2 + \mu^2$. Tale formula giustifica la **decomposizione in BIAS e Varianza del MSE**, scegliendo $x = \hat{\theta}(y)$, e definendo il BIAS come $E[\hat{\theta}(y)] - \theta$. Si ha quindi

$$MSE(\hat{\theta}(y)) := E[(\hat{\theta}(y)-\theta)(\hat{\theta}(y)-\theta)^T] = Var(\hat{\theta}(y)) + BIAS \cdot BIAS^T \Rightarrow MSE_{scalare} = VAR_{scalare} + ||BIAS||^2$$

- Le aspettazioni vanno fatte rispetto a θ, cioè $E_\theta[\cdot]$, in quanto le densità di probabilità dipendono da θ, ed il valore VERO θ_0 (dell'approccio Fisheriano) non è noto. θ verrà spesso omesso come pedice dell'aspettazione per motivi di semplicità, tendendolo quindi sottointeso

- si verificano facilmente (ricordando che $v^T M w = w^T M^T v$), se $M = M^T$ (simmetrica), le seguenti formule per Gradiente ed Hessiano

$$\nabla_\theta v^T M \theta = \nabla_\theta \theta^T M v = Mv,\ \nabla_\theta \theta^T M \theta = 2M\theta,\ \nabla^2_\theta \theta^T M \theta = 2M$$

- MSE (Mean-Square-Error, in senso matriciale e di conseguenza anche scalare) è il parametro-base per valutare la bontà o meno di uno stimatore. Uno stimatore è migliore di un altro nel senso MSE se **per ogni** θ ha MSE non-superiore[1] (quindi il migliore vale in senso **uniforme** rispetto a θ). Uno stimatore di una certa classe (ad esempio gli **Unbiased o Non-Polarizzati o Corretti**, quelli per cui vale $E_\theta[\hat{\theta}(y)] = \theta$ per ogni θ) è detto **MVE (a minima varianza d'errore)** se non esiste nella stessa classe uno stimatore migliore. Restringersi ad una certa classe è obbligatorio, altrimenti gli stimatori costanti (che **non** sfruttano l'informazione delle misure effettuate) paradossalmente renderebbero impossibile l'esistenza di uno stimatore MVE. Infatti per $\hat{\theta}(y) = a$ vale $MSE = (a-\theta)^2$, mentre qualunque stimatore non-costante ha $MSE(\theta) > 0$ per ogni θ, in quanto solo la sua varianza è sufficiente a renderlo tale, in quanto la varianza di $f(y)$ può essere nulla solo se $f(y)$ è costante. Gli stimatori costanti non sono nè migliori nè peggiori degli altri, **nel senso MSE uniforme** in θ, in quanto per θ tendente all'infinito MSE di uno stimatore costante diverge, ma è invece nullo per θ pari alla costante stessa, mentre gli stimatori non-banali hanno spesso un $MSE(\theta)$ limitato (o comunque inferiore a quello di stimatori costanti per θ tendente all'infinito) ma mai nullo (questo almeno nel caso in cui sia $\Theta = \mathbb{R}^k$, o quantomeno sia un insieme illimitato). Si ricorda che uno stimatore **non** può dipendere da θ (che non è noto), ma solo da y (altrimenti θ stesso sarebbe uno stimatore costante MVE, essendo identicamente nullo MSE)

- **uno stimatore corretto non esiste sempre** (e quindi a maggior ragione non esiste sempre uno stimatore corretto MVE), come provato da $p(y,\theta) = \theta e^{-\theta y}$, $\theta > 0, y \geq 0$. Infatti se esistesse un tale $g(y)$ sarebbe

$$E_\theta[g(y)] = \int_0^{+\infty} g(y)\theta e^{-\theta y} dy = \theta \Rightarrow \int_0^{+\infty} g(y) e^{-\theta y} dy = 1 \Rightarrow \left(\frac{\partial}{\partial \theta}\right) \Rightarrow \int_0^{+\infty} y g(y) e^{-\theta y} dy = 0$$

  dove tutte le funzioni in gioco sono non-negative, per cui dev'essere $yg(y) = 0$ che implica $g(y) = 0$ che contraddice la prima equazione (che diventa $\theta = 0$)

[1] In senso matriciale, $A \geq B$ (**utilizzato solo per matrici simmetriche**) significa $A - B \geq 0$, cioè **semidefinita positiva**. Ciò ovviamente implica $tr(A) \geq tr(B)$, e più precisamente si ha $tr(A) = tr(B)$ **se e solo se** $A = B$

2 Minimi Quadrati (Least Squares) Non Probabilistici

Modello (vettoriale) con rango(ϕ) = k, con ϕ matrice $n \times k$ (e quindi per forza $n \geq k$), dove e è il vettore degli errori di misura (senza nessuna ipotesi probabilistica)

$$y = \Phi\theta + e$$

Il problema è minimizzare (W è la matrice dei pesi diagonale, in particolare può essere $W = I$)

$$\hat{\theta}^{LS} = argmin_\theta \ J(\theta), \ J(\theta) := (y - \phi\theta)^T W (y - \phi\theta)$$

Facendo il gradiente e ponendo uguale a zero (per cercare il minimo), dove LS sta per Least Squares

$$\nabla_\theta \ J(\theta) = -2\phi^T W y + 2\phi^T W \phi\theta = 0 \ \Rightarrow \ \hat{\theta}^{LS} = (\phi^T W \phi)^{-1}(\phi^T W) y$$

e l'Hessiano risulta

$$\nabla^2_\theta \ J(\theta) = 2\phi^T W \phi > 0$$

nelle ipotesi sul rango di ϕ e con $W > 0$ (pesi tutti positivi), per cui trattasi di un minimo relativo, che è anche assoluto in quanto per θ tendente all'infinito $J(\theta)$ diverge (con andamento parabolico).

3 Minimi Quadrati (Least Squares) Probabilistici

Modello (vettoriale) con rango(ϕ) = k, con ϕ matrice $n \times k$ (e quindi per forza $n \geq k$), dove e è il vettore degli errori di misura, assunto questa volta un vettore aleatorio a media nulla e varianza (matriciale) $\Sigma > 0$

$$y = \Phi\theta + e$$

Il problema è minimizzare MSE (Mean Square Error), nelle due ipotesi seguenti

- stimatore **lineare** della forma $\hat{\theta}^{LS}(y) = Ay$
- stimatore **corretto (unbiased)**, cioè $E_\theta[\hat{\theta}^{LS}(y)] = \theta$, per ogni $\theta \in \Theta \subseteq \mathbb{R}^k$

Si vede facilmente che è Unbiased se e solo se $A\phi = I$, equazione che ammette almeno una soluzione (dalle ipotesi sul rango di ϕ), ma in generale ne ammette infinite. Se A_0 è una soluzione particolare di $A\phi = I$, la soluzione generale è $A = A_0 + B$, per ogni B tale che $B\phi = 0$. Calcolando MSE (matriciale) che coincide con la varianza (matriciale) essendo Corretto, si ha

$$Var(\hat{\theta}^{LS}(y)) = E_\theta[Aee^TA^T] = AE_\theta[ee^T]A^T = A\Sigma A^T = (A_0 + B)\Sigma(A_0 + B)^T = A_0\Sigma A_0^T + 2B\Sigma A_0^T + B\Sigma B^T$$

Se risulta $B\Sigma A_0^T = 0$, per ogni B tale che $B\phi = 0$, il che equivale a dire che esiste M tale che $\Sigma A_0^T = \phi M$, in tal caso si semplifica in

$$Var(\hat{\theta}^{LS}(y)) = A_0\Sigma A_0^T + B\Sigma B^T \geq A_0\Sigma A_0^T$$

per cui la varianza è **minima**[2] e vale $A_0\Sigma A_0^T$. Ma $\Sigma A_0^T = \phi M$ implica $A_0 = M^T\phi^T\Sigma^{-1}$, ed imponendo $A_0\phi = I$ si ottiene $M^T(\phi^T\Sigma^{-1}\phi) = I$, da cui necessariamente $M^T = (\phi^T\Sigma^{-1}\phi)^{-1}$ e quindi $A_0 = (\phi^T\Sigma^{-1}\phi)^{-1}\phi^T\Sigma^{-1}$, da cui anche $A_0\Sigma A_0 = (\phi^T\Sigma^{-1}\phi)^{-1}$. In conclusione, la soluzione cercata è

$$\hat{\theta}^{LS}(y) = (\phi^T\Sigma^{-1}\phi)^{-1}\phi^T\Sigma^{-1}y, \ MSE(\hat{\theta}^{LS}(y)) = Var(\hat{\theta}^{LS}(y)) = (\phi^T\Sigma^{-1}\phi)^{-1}$$

Ovviamente MSE scalare coincide con VAR scalare e basta fare la **traccia** ($tr((\phi^T\Sigma^{-1}\phi)^{-1})$). Si noti che tale stimatore (lineare, unbiased, ed a MSE minimo) ha esattamente la stessa forma del caso non-probabilistico, con la sostituzione di W con Σ^{-1}. I calcoli necessari alla deduzione sono invece stati completamente diversi. Si notino i due casi particolari seguenti

- ϕ quadrata (ed invertibile per le ipotesi sul suo rango): in tal caso $\hat{\theta}^{LS}(y) = \phi^{-1}y$
- $\Sigma = \sigma^2 I$ (v.a. IID, cioè **indipendenti ed identicamente distribuite**): in tal caso $\hat{\theta}^{LS}(y) = (\phi^T\phi)^{-1}\phi^T y$
- quello che abbiamo trovato è lo Stimatore Lineare Unbiased MVE (a minima varianza d'errore, cioè con minimo MSE)

[2] $B\Sigma B^T$ è **almeno semidefinita positiva**, ed è nulla se e solo se $B = 0$, quindi ogni $B \neq 0$ conduce ad una varianza scalare **superiore**

4 Fisher e Cramer-Rao per Stimatori Corretti

Si definisca (MENO-LOG-VEROSIMIGLIANZA o meno-log-LIKELIHOOD), dato y con densità di probabilità $p(y,\theta)$ (NOTA: tutto vale invariato per vettori aleatori DISCRETI, ma qui ci limitiamo ai CONTINUI descritti da una densità di probabilità)

$$\ell(\theta, y) = -\log p(y,\theta)$$

Sia $\hat{\theta}(y)$ un QUALSIASI stimatore corretto di θ, e consideriamo il vettore aleatorio

$$v = \begin{bmatrix} \hat{\theta}(y) - \theta \\ \nabla_\theta \ell(\theta, y) \end{bmatrix}$$

Tale vettore ha media nulla (il primo termine per definizione, il secondo come conseguenza dei seguenti passaggi, fatti per semplicità nel caso di un solo parametro)

$$E_\theta[\frac{\partial}{\partial\theta}\ell(\theta,y)] = \int_{-\infty}^{+\infty} \frac{\partial}{\partial\theta}\ell(\theta,y)p(y,\theta)dy = -\int_{-\infty}^{+\infty} \frac{\partial}{\partial\theta}p(y,\theta)dy = -\frac{\partial}{\partial\theta}\int_{-\infty}^{+\infty} p(y,\theta)dy = -\frac{\partial}{\partial\theta}1 = 0$$

Con calcoli assolutamente analoghi, si vede (si ricordi che $\hat{\theta}(y)$ NON dipende da θ), sfruttando il fatto che $\hat{\theta}(y)$ è corretto

$$E_\theta[\hat{\theta}(y)\frac{\partial}{\partial\theta}\ell(\theta,y)] = \ldots = -\frac{\partial}{\partial\theta}\int_{-\infty}^{+\infty} \hat{\theta}(y)p(y,\theta)dy = -\frac{\partial}{\partial\theta}E_\theta[\hat{\theta}(y)] = -\frac{\partial}{\partial\theta}\theta = -1$$

e quindi facilmente (θ è costante rispetto alla variabile di integrazione y) anche

$$E_\theta[(\hat{\theta}(y) - \theta)\frac{\partial}{\partial\theta}\ell(\theta,y)] = -1$$

da cui la matrice varianza di v

$$Var(v) = \begin{bmatrix} Var(\hat{\theta}(y)) & -1 \\ -1 & I(\theta) \end{bmatrix} \geq 0,\ I(\theta) := Var(\frac{\partial}{\partial\theta}\ell(\theta,y))$$

ed applicando il Test di Sylvester (determinante non-negativo)

$$Var(\hat{\theta}(y)) \geq I^{-1}(\theta)$$

che si generalizza al caso VETTORIALE (**senza DIM**) pur di definire

$$I(\theta) = Var(\nabla_\theta \ell(\theta, y))$$

La precedente matrice è detta matrice di INFORMAZIONE di Fisher, e la precedente disuguaglianza è detta LIMITE INFERIORE di Cramer-Rao, in quanto fornisce un minimo sotto il quale la varianza di nessuno stimatore corretto può andare. La situazione MIGLIORE possibile è quella in cui $Var(\hat{\theta}(y)) = I^{-1}(\theta)$, nel qual caso lo stimatore corretto si dice EFFICIENTE. Qualche nota finale

- i ragionamenti effettuati sono validi se si può invertire l'ordine tra integrazione in y e derivata rispetto a θ. Ciò accade quasi sempre, ma NON qualora gli estremi di integrazione dipendano da θ, come accade ad esempio per le v.a. UNIFORMI ($\mathcal{U}(0,\theta)$). In tal caso $\ell(\theta, y) = \log\theta$ il cui gradiente (scalare) elevato al quadrato è θ^{-2}, costante rispetto ad y, per cui la sua aspettazione fornisce $I(\theta) = \theta^{-2}$ ed il limite di Cramer-Rao sembrerebbe essere θ^2, mentre lo stimatore CORRETTO $2y$ ha varianza inferiore, pari a $\frac{\theta^2}{3}$. Non c'è contraddizione, semplicemente la derivazione di Cramer-Rao non vale in questo caso
- sempre se vale questa inversione integrale-derivata, si dimostra con conti analoghi che vale anche (**senza DIM**)

  $$I(\theta) = E_\theta[\nabla^2_\theta \ell(\theta, y)]$$

  Tale formula in genere richiede calcoli più semplici rispetto alla formula precedente
- nel caso $y_1, y_2, \ldots, y_n$ IID, la densità fattorizza nel prodotto di densità, il logaritmo diventa una somma, le aspettazioni sono di n termini identici (ciascuno relativo ad una diversa y_i), e quindi si ottiene $I_n(\theta) = nI(\theta)$, dove $I(\theta)$ è relativo ad UNA v.a. e $I_n(\theta)$ relativo a tutte le n v.a.. Ne consegue che il limite di Cramer-Rao diventa n volte più piccolo ($\frac{I^{-1}(\theta)}{n}$), con n v.a. IID

5 Il Caso Lineare Gaussiano

Riprendiamo il problema

$$y = \phi\theta + e$$

dove ora assumiamo che e sia Gaussiano, precisamente che sia $e \sim \mathcal{N}(0, \Sigma)$. Notiamo che questo è ASSOLUTAMENTE EQUIVALENTE ad assumere per y la descrizione probabilistica $y \sim \mathcal{N}(\phi\theta, \Sigma)$, quindi modello lineare e descrizione probabilistica sono equivalenti nel caso GAUSSIANO. Allora ($\det 2\pi\Sigma = (2\pi)^n \det\Sigma$)

$$p(y, \theta) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} e^{-\frac{(y-\phi\theta)^T\Sigma^{-1}(y-\phi\theta)}{2}}$$

Semplici conti conducono a (dove K è una costante rispetto a θ, quindi irrilevante)

$$\ell(\theta, y) = K + \frac{(y-\phi\theta)^T\Sigma^{-1}(y-\phi\theta)}{2} \;\Rightarrow\; \nabla_\theta\ell(\theta, y) = -\phi^T\Sigma^{-1}y + \phi^T\Sigma^{-1}\phi\theta = \phi^T\Sigma^{-1}(\phi\theta - y)$$

Il vettore $\phi\theta - y$ ha media nulla e varianza Σ. Da un lato è confermata la media nulla (si riveda la prova di Cramer-Rao) di $\nabla_\theta\ell(\theta, y)$, dall'altro la sua varianza vale $\phi^T\Sigma^{-1}E_\theta[(\phi\theta-y)^T(\phi\theta-y)]\Sigma^{-1}\phi = \phi^T\Sigma^{-1}\Sigma\Sigma^{-1}\phi = \phi^T\Sigma^{-1}\phi$, da cui

$$I(\theta) = \phi^T\Sigma^{-1}\phi \;\Rightarrow\; MSE(\hat\theta(y)) = Var(\hat\theta(y)) \geq (\phi^T\Sigma^{-1}\phi)^{-1}$$

per OGNI stimatore corretto. Ma in una precedente Section abbiamo provato che (indipendentemente dalla Gaussianità o meno), $I^{-1}(\theta)$ è proprio MSE dello Stimatore Lineare Unbiased a Minima Varianza d'Errore, quindi questo significa che nessuno Stimatore Unbiased può comportarsi meglio di lui. Lo Stimatore Lineare Unbiased nel caso Gaussiano è lo Stimatore Unbiased MVE in assoluto (nessuno stimatore non-lineare potrebbe fare meglio). Si noti che la matrice di Fisher poteva essere calcolata anche ricorrendo all'Hessiano (che è costante rispetto a θ, per cui la sua aspettazione è l'Hessiano stesso)

$$\nabla^2_\theta\ell(\theta, y) = \phi^T\Sigma^{-1}\phi \;\Rightarrow\; E_\theta[\nabla^2_\theta\ell(\theta, y)] = \phi^T\Sigma^{-1}\phi = I(\theta)$$

Teorema. (senza DIM) Se esiste uno Stimatore Unbiased MVE, allora esso è necessariamente unico.

Questo teorema conferma che non solo nessun altro stimatore corretto può essere migliore di quello Lineare (nel senso MVE), ma non può neppure eguagliarne le prestazioni. Inoltre, l'uguaglianza in Cramer-Rao prova che tale stimatore lineare è non solo MVE, ma è addirittura EFFICIENTE. Interessante è il caso particolare in cui

$$\phi = [\,1 \quad 1 \quad \dots \quad 1\,]^T,\; \Sigma = \sigma^2 I$$

che corrisponde al caso di variabili IID, in cui θ è la media da stimare, mentre σ^2 è supposta nota. Particolarizzando le equazioni precedenti a tale caso particolare, si trova facilmente

$$\hat\theta^{LS}(y) = \frac{[\,1 \quad 1 \quad \dots \quad 1\,]\,y}{n} = \frac{1}{n}\Sigma_1^n\, y_i,\; I(\theta) = \frac{n}{\sigma^2},\; MSE = VAR = I^{-1}(\theta) = \frac{\sigma^2}{n}$$

che prova che la MEDIA CAMPIONARIA è lo stimatore Unbiased ottimo (in senso MVE) della media, essendo addirittura EFFICIENTE.

6 Stima della Varianza nel Caso Gaussiano

Vogliamo estendere l'analisi del caso particolare appena analizzato (variabili IID) considerando la situazione in cui anche σ^2 sia da stimare (consideriamo cioè $\theta_1 = \theta, \theta_2 = \sigma^2$). Ricalcoliamo in tal caso

$$\ell(\theta, \sigma^2, y) = K + \frac{n}{2}\log\sigma^2 + \frac{\Sigma_1^n\,(y_i-\theta)^2}{2\sigma^2} \;\Rightarrow\; \nabla_{\theta,\sigma^2}\ell(\theta, \sigma^2, y) = \begin{bmatrix} \frac{1}{\sigma^2}\Sigma_1^n\,(\theta - y_i) \\ \frac{n\sigma^2 - \Sigma_1^n\,(y_i-\theta)^2}{2\sigma^4} \end{bmatrix} \;\Rightarrow$$

$$\nabla^2_{\theta,\sigma^2}\ell(\theta, \sigma^2, y) = \begin{bmatrix} \frac{n}{\sigma^2} & \frac{1}{\sigma^4}\Sigma_1^n\,(y_i-\theta) \\ \frac{1}{\sigma^4}\Sigma_1^n\,(y_i-\theta) & \frac{2\Sigma_1^n\,(y_i-\theta)^2 - n\sigma^2}{2\sigma^6} \end{bmatrix} \;\Rightarrow\; I(\theta, \sigma^2) = E[\nabla^2_{\theta,\sigma^2}\ell(\theta, \sigma^2, y)] = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

Quindi il limite di Cramer-Rao è

$$I^{-1}(\theta, \sigma^2) = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

Per stimare la varianza, distinguiamo ora due casi

- θ è noto (cioè la media è nota e solo la varianza va stimata). In tal caso pare NATURALE ricorrere alla VARIANZA CAMPIONARIA (del tutto analoga alla media campionaria)

$$\hat{\sigma}^2(y) = \frac{1}{n}\Sigma_1^n\ (y_i - \theta)^2\ \Rightarrow\ E[\hat{\sigma}^2(y)] = \sigma^2,\ MSE(\hat{\sigma}^2) = Var(\hat{\sigma}^2(y) = \frac{2\sigma^4}{n}$$

dove il calcolo della varianza è un pò laborioso (richiedendo l'utilizzo del momento centrale $M_4 = 3\sigma^4$ della Gaussiana). Tuttavia il risultato è potente: trattasi ancora una volta di Stimatore Unbiased MVE essendo EFFICIENTE (eguaglia il limite di Cramer-Rao)

- θ va a sua volta stimato. In tal caso non possiamo inserire θ nell'espressione dello Stimatore, ma possiamo molto ragionevolmente sostituire θ con la sua stima, quindi ricorrere a

$$\hat{\sigma}^2(y) = \frac{1}{n}\Sigma_1^n\ (y_i - \hat{\theta}(y))^2\ \Rightarrow\ E[\hat{\sigma}^2(y)] = \frac{n-1}{n}\sigma^2,\ MSE(\hat{\sigma}^2(y)) = Var(\hat{\sigma}^2(y)) + Bias(\hat{\sigma}^2(y)) = \left(2 - \frac{1}{n}\right)\frac{\sigma^4}{n}$$

Questa volta i conti sono complessi sia per media che per varianza, ma il risultato è che lo stimatore risulta BIASED, ed ha un MSE addirittura leggermente inferiore al limite di Cramer-Rao (leggermente nel senso che per n grande $\frac{1}{n}$ diventa trascurabile). Inoltre a rigore non avrebbe senso questa comparazione con Cramer-Rao, non essendo lo stimatore corretto, ma la comparazione è stata fatta per evidenziare che MSE non solo si scosta poco dal minimo per gli stimatori Unbiased, ma che addirittura in questo caso abbiamo individuato uno stimatore BIASED che è MIGLIORE di qualunque stimatore Unbiased (nel senso MSE), in quanto va sotto il limite inferiore di Cramer-Rao: talvolta rinunciare alla CORRETTEZZA può risultare addirittura un beneficio. Modificare lo stimatore in modo da renderlo Biased è comunque semplicissimo, bastando moltiplicarlo per $\frac{n}{n-1}$. Così facendo otteniamo un secondo stimatore

$$\hat{\sigma}^2(y) = \frac{1}{n-1}\Sigma_1^n\ (y_i - \hat{\theta}(y))^2\ \Rightarrow\ E[\hat{\sigma}^2(y)] = \sigma^2,\ MSE(\hat{\sigma}^2(y)) = Var(\hat{\sigma}^2(y)) = \frac{2\sigma^4}{n-1}$$

che ha MSE leggermente PEGGIORE rispetto a Cramer-Rao, ed ovviamente a maggior ragione peggiore del precedente stimatore BIASED

- **Nota.** Si noti l'enorme differenza tra questa situazione e quella già vista a proposito degli stimatori COSTANTI, che possono comportarsi meglio di stimatori corretti: in quel caso l'essere MIGLIORE non era in senso UNIFORME rispetto a θ, ma solo in un intervallo di valori, il che è quanto dire che uno stimatore è migliore nel senso MSE di un altro per alcuni valori di θ, e peggiore per altri valori. Qui invece il MIGLIORE vale in senso uniforme, cioè per ogni θ la varianza campionaria biased si comporta meglio della sua versione correttificata

7 Stima a Massima Verosimiglianza (Maximum Likelihood)

Indicata con ML (Maximum Likelihood), fornisce un efficiente strumento per costruire Stimatori molto potenti. Essa si basa sulla seguente considerazione intuitiva: dato y (cioè misurato un certo valore di y), qual è il valore di θ che è maggiormente IN ACCORDO con y? Intuitivamente, quello che rende MASSIMA $p(y, \theta)$. In sostanza, indicheremo con $\hat{\theta}^{ML}(y)$ quel valore di θ che massimizza la probabilità che si sia ottenuto proprio quel valore di y. Siccome log è monotona crescente, questo equivale a minimizzare $\ell(\theta, y)$, quindi

$$\hat{\theta}^{ML}(y) = argmin_\theta\ \ell(\theta, y)$$

Chiaramente questo richiede un'analisi dei punti critici (quelli in cui $\nabla_\theta \ell(\theta, y) = 0$) ed il comportamento sulla frontiera di Θ. Normalmente, se il punto critico è uno solo e l'Hessiano è definito positivo (minimo locale), esso individua anche il minimo assoluto, ma in generale un'analisi del comportamento anche sulla frontiera può essere necessario. Ovviamente per avere senso occorre che tale minimo assoluto sia univocamente determinato per ogni y, in modo che possa essere espresso come una funzione di y.

1. Il caso Lineare Gaussiano. Nel caso di $y = \phi\theta + e$, con $e \sim \mathcal{N}(0, \Sigma)$ (o equivalentemente nel caso di $y \sim \mathcal{N}(\phi\theta, \Sigma)$), abbiamo già visto che si ha

$$\nabla_\theta \ell(\theta, y) = -\phi^T\Sigma^{-1}y + \phi^T\Sigma^{-1}\phi\theta = \phi^T\Sigma^{-1}(\phi\theta - y),\ \nabla^2_\theta \ell(\theta, y) = \phi^T\Sigma^{-1}\phi > 0$$

e ponendo $\nabla_\theta \ell(\theta, y) = 0$ si ottiene facilmente lo stesso risultato della stima Least Squares, cioè

$$\hat{\theta}^{ML}(y) = \hat{\theta}^{LS}(y) = (\phi^T \Sigma^{-1} \phi)^{-1} \phi^T \Sigma^{-1} y, \; MSE = VAR = (\phi^T \Sigma^{-1} \phi)^{-1} = I^{-1}(\theta)$$

dove il fatto che sia un minimo assoluto è garantito dall'essere l'Hessiano definito positivo (minimo locale) e dal comportamento divergente di $\ell(\theta, y)$ per $||\theta|| \to +\infty$. Di conseguenza valgono tutti i risultati già visti a proposito: la stima MVE, l'EFFICIENZA, la CORRETTEZZA, che si estendono al caso particolare della media campionaria. Questo è in essenza il **Teorema di Gauss-Markov (lo Stimatore ML nel caso Lineare Gaussiano è lineare, corretto, MVE, ed efficiente)**.

2. Stima della Varianza nel caso Gaussiano IID. Nel caso di $y \sim \mathcal{N}(\mu \, [1 \quad 1 \quad \dots \quad 1]^T, \sigma^2 I)$, abbiamo già visto che

$$\nabla_{\mu,\sigma^2} \ell(\mu, \sigma^2, y) = \begin{bmatrix} \frac{1}{\sigma^2} \Sigma_1^n \, (\mu - y_i) \\ \frac{n\sigma^2 - \Sigma_1^n \, (y_i - \mu)^2}{2\sigma^4} \end{bmatrix}, \; \nabla^2_{\mu,\sigma^2} \ell(\mu, \sigma^2, y) = \begin{bmatrix} \frac{n}{\sigma^2} & \frac{1}{\sigma^4} \Sigma_1^n \, (y_i - \mu) \\ \frac{1}{\sigma^4} \Sigma_1^n \, (y_i - \mu) & \frac{2\Sigma_1^n \, (y_i - \mu)^2 - n\sigma^2}{2\sigma^6} \end{bmatrix}$$

Uguagliando a zero $\nabla_{\mu,\sigma^2} \ell(\mu, \sigma^2, y)$ si ottengono la media e la varianza campionaria (nel caso di media μ NOTA, la derivata da fare è una sola e porge la varianza campionaria CORRETTA ed EFFICIENTE in tale situazione). Si nota che quindi la ML non fornisce sempre stimatori corretti (la varianza campionaria non lo è), ma comunque con ottime prestazioni (la stima della varianza ha MSE addirittura inferiore a Cramer-Rao). Valutando l'Hessiano nel punto $(\hat{\mu}, \hat{\sigma}^2)$ si ottiene l'Hessiano pari ad $I(\mu, \sigma^2)$, valutato in $\sigma = \hat{\sigma}^2$ (quindi simile alla matrice di Fisher, ma con una sostanziale differenza: esso è funzione di y e non di μ, σ^2, in quanto nessuna aspettazione va fatta e si valuta l'Hessiano in un punto funzione degli stimatori e quindi di y). Sulla frontiera $\ell(\mu, \sigma^2)$ tende ad infinito (per μ è evidente, per σ^2 tendente a zero prevale σ^{-2} rispetto a $\log \sigma^2$, per σ^2 tendente ad infinito prevale il termine $\log \sigma^2$).

3. Stima di una Variabile Uniforme. Sia $y \sim \mathcal{U}(0, \theta)$, con ovviamente $\theta > 0$ (si ricordi media $\frac{\theta}{2}$ e varianza $\frac{\theta^2}{12}$). Osservato y, certamente θ non può essere inferiore ad y (altrimenti sarebbe nulla la densità valutata in (y, θ)), quindi $\theta \geq y$. Tuttavia $p(y, \theta) = \theta^{-1}$ implica che il massimo valore si ottenga quando θ è il più piccolo possibile. Quindi $\hat{\theta}^{ML}(y) = y$. Lo stimatore NON è corretto, ed ha MSE (calcolando Bias e Varianza) pari a $\frac{\theta^2}{3}$. La sua versione *correttificata* $\hat{\theta}(y) = 2y$ ha MSE=VAR pari allo stesso valore precedente. Tuttavia, studiando $\hat{\theta}(y) = ay$ al variare di a si vede che il minimo MSE si ottiene per $a = \frac{3}{2}$, e vale $\frac{\theta^2}{4}$. In sostanza, in tale caso nè una stima lineare corretta nè la stima ML sono ottimali. Esiste una stima lineare BIASED e diversa da quella ML, che rende MSE inferiore, ed è quindi migliore di entrambe. Questo prova che la stima ML non è sempre ottimale, tuttavia il suo MSE non si discosta troppo da quello dello stimatore lineare ottimo (si ricordi il precedente discorso sulla non-validità di Cramer-Rao in questo caso).

4. Stima di una Variabile Esponenziale Unilatera. Sia $p(y, \theta) = \theta^{-1} e^{-y\theta^{-1}}, \; \theta > 0$ (si ricordi media θ e varianza θ^2). Si ha

$$\ell(\theta, y) = \log \theta + \frac{y}{\theta} \; \Rightarrow \; \nabla_\theta \ell(\theta, y) = \theta^{-1} - y\theta^{-2} \; \Rightarrow \; \hat{\theta}^{ML}(y) = y$$

Calcolando $E[\nabla^2_\theta \ell(\theta, y)] = \theta^{-2} = I(\theta)$, si trova il limite di Cramer-Rao $I^{-1}(\theta) = \theta^2$, che coincide con MSE della stima CORRETTA ML. Quindi in tal caso ML fornisce correttezza, MVE ed efficienza. Si noti che se avessimo fatto i conti con $p(y, \theta) = \theta e^{-y\theta}$ avremmo trovato lo STESSO limite di Cramer-Rao, ma sappiamo che in tal caso non è applicabile in quanto tale variabile EXP non ammette NESSUNO stimatore corretto. Con la ML si troverebbe $\hat{\theta}^{ML}(y) = y^{-1}$, che ha facilmente media e varianza infinite, quindi anche MSE infinito, chiaramente del tutto insoddisfacente. Eppure lo scambio $\theta \leftrightarrow \theta^{-1}$, apparentemente innocuo, ha creato un pandemonio: si passa da uno stimatore BIASED ed a MSE infinito, ad uno corretto, MVE ed efficiente. Ciò significa che stimare θ oppure θ^{-1} non è, come potrebbe apparire a prima vista, la *stessa cosa*, e che la SCELTA del parametro da stimare può essere determinante. In altri termini, usare una parametrizzazione oppure un'altra non è ininfluente, come dimostra questo caso limite, e scegliere una *buona* riparametrizzazione può sistemare le cose. Si noti che comunque i due stimatori ML sono legati dallo stesso cambio di variabile (inversione). Ciò non è un caso, in quanto vale il

Teorema di Invarianza della Stima ML (senza DIM). Date $p(\theta, y)$ e $p(\tau(\mu), y)$, con $\theta = \tau(\mu)$ (la stessa densità parametrizzata da un cambio di variabile $\mu \to \theta = \tau(\mu)$), le stime ML sono legate dallo STESSO cambio di variabile

$$\hat{\theta}^{ML}(y) = \hat{\tau(\mu)}^{ML}(y) = \tau(\hat{\mu}^{ML}(y))$$

5. Stima di Variabili Bernoulliane. Date n v.a. y_i IID con probabilità discreta $p(y,\theta) = \theta^y(1-\theta)^{(1-y)}$, con $0 \leq \theta \leq 1$ e $y = 0, 1$ (in altri termini $\theta = P[y=1], 1-\theta = P[y=0]$), si vuole stimare θ. Si ha (ponendo per semplicità $S = \Sigma_1^n\ y_i,\ 0 \leq S \leq n$)

$$\ell(\theta, y) = -S\log\theta - (n-S)\log(1-\theta) \ \Rightarrow\ \frac{\partial}{\partial\theta}\ell(\theta,y) = -\frac{S}{\theta} + \frac{n-S}{1-\theta}$$

Occorre distinguere dei casi: se $S = 0$ oppure $S = n$ la derivata non si annulla mai, quindi il minimo va ricercato negli estremi. Se $S = 0$ il minimo si ha per $\theta = 0$, mentre per $\theta = 1$ se $S = n$. Se invece $0 < S < n$ la derivata si annulla per $\theta = \frac{S}{n}$, e trattasi di un minimo in quanto $\ell(\theta, y)$ diverge per θ tendente a 0 oppure ad 1. Tale formula per θ è in accordo con i due casi particolari evidenziati, per cui $\theta = \frac{S}{n}$ è in ogni caso il punto di minimo, da cui

$$\hat{\theta}^{ML}(y) = \frac{1}{n}\Sigma_1^n\ y_i$$

Trattasi della solita media campionaria, e da $E[y] = \theta$ segue che è stimatore corretto. Da

$$I(\theta) = E[\frac{\partial^2}{\partial\theta^2}\ell(\theta,y)] = \frac{E[S]}{\theta^2} + \frac{n-E[S]}{(1-\theta)^2} = \frac{n}{\theta} + \frac{n}{1-\theta} = \frac{n}{\theta(1-\theta)} \ \Rightarrow\ I^{-1}(\theta) = \frac{\theta(1-\theta)}{n}$$

Ricordando che $Var(y) = \theta(1-\theta)$, ed essendo y_i IID, segue facilmente che la varianza dello stimatore ML coincide con il limite di Cramer-Rao, quindi lo Stimatore ML è ancora una volta corretto, MVE ed efficiente.

8 Note Finali

1. MSE matriciali e scalari. Finora abbiamo concentrato l'attenzione soprattutto su MSE matriciale, in quanto per determinare se uno stimatore fosse efficiente o meno, fosse migliore di un altro o meno, dovevamo comparare la matrice Varianza (nel caso Unbiased) con la matrice di Fisher, oppure verificare che una matrice MSE di uno stimatore fosse maggiore od uguale a quella di un altro (si ricorda che per definizione $A \geq B$ significa $A - B \geq 0$, cioè che la matrice simmetrica $A - B$ è SEMIDEFINITA POSITIVA: in questo senso vanno intesi i $\geq$ in senso matriciale), ma ricordiamo che l'effettivo MSE è sempre in senso SCALARE, in quanto $MSE(\hat{\theta}(y)) = E_\theta[||\hat{\theta}(y) - \theta||^2]$, cioè rappresenta una sorta di DISTANZA EUCLIDEA mediata (nel senso dell'aspettazione). Ma passare dal MSE matriciale a quello scalare è semplice, bastando fare la traccia. Tuttavia MSE matriciale contiene più informazione, in quanto da esso si può risalire anche al MSE delle SINGOLE componenti del vettore $\hat{\theta}(y)$. Vediamo alcuni esempi

- la relazione $MSE = VAR + BIAS^2$ si scrive, nei casi matriciale e scalare

  $E[(\hat{\theta}(y)-\theta)(\hat{\theta}(y)-\theta)^T] = Var[\hat{\theta}(y)] + (E[\hat{\theta}(y)]-\theta)(E[\hat{\theta}(y)]-\theta)^T \Rightarrow E[||\hat{\theta}(y)-\theta||^2] = tr[Var(\hat{\theta}(y)] + ||E[\hat{\theta}(y)]-\theta||^2$

  In effetti, ciascun termine si ottiene valutando la traccia, in quanto si dimostra che $tr(AB) = tr(BA)$, ed invertendo l'ordine MSE e BIAS matriciale diventano degli scalari (equivale a rimpiazzare matrici della forma vv^T con scalari $v^Tv = ||v||^2$, dove v è un vettore)

- dalla varianza matriciale si risale alle varianze delle singole componenti. Ad esempio

  $$MSE = \begin{bmatrix} 100 & 3 \\ 3 & 1 \end{bmatrix} \Rightarrow MSE_{scalare} = 101, MSE(\hat{\theta}_1(y)) = 100, MSE(\hat{\theta}_2(y)) = 1$$

  In tal modo si vede che globalmente $MSE = 101$ (traccia), ma che θ_1 e θ_2 vengono stimate con precisione BEN DIVERSA: la varianza della stima di θ_1 è 100 volte più grande di quella di θ_2, che quindi è stimato in modo molto più preciso rispetto ad θ_1 (le singole varianze sono ovviamente gli elementi sulla diagonale. Il 3 fuori della diagonale non ha invece alcun ruolo particolare, in quanto indica solo la presenza di una CORRELAZIONE non nulla tra i due stimatori di θ_1 e θ_2). La stessa cosa vale per varianza e bias: guardando le componenti diagonali delle varianza e del $BIAS^2$ matriciali, si risale alle varianza ed al $BIAS^2$ delle singole componenti

2. Regioni di Confidenza. Spesso la stima deve essere accompagnata da una INFO aggiuntiva, nel senso che la stima in se non ha senso se non si sa quanto precisa essa sia. Tale info è in effetti contenuta nella varianza dello stimatore, ma spesso si preferisce conoscere una REGIONE cui il valore VERO θ_0

appartiene con una certa probabilità. Ad esempio, un conto è dire che $\hat{\theta} = 10$, ben altro dire che θ_0 appartiene ad un certo intervallo centrato in 10 con una certa probabilità, ad esempio che θ_0 sta nell'intervallo [9 11] con probabilità 0.95 (del 95%). Si parla quindi di REGIONI DI CONFIDENZA, nel senso che si confida che θ_0 abbia, con elevata probabilità, un valore che sta all'interno di una certa regione. Il caso più semplice è quello Gaussiano, in cui è noto ad esempio che $P[x \in [\ \mu - 2\sigma\ \mu + 2\sigma\ ]] \simeq 0.95$, mentre $P[x \in [\ \mu - 2.6\sigma\ \mu + 2.6\sigma\ ]] \simeq 0.99$. Quindi se ad esempio $\hat{\theta}(y)$ è uno stimatore corretto di θ con varianza σ^2, possiamo dire che $P[\hat{\theta}(y) \in [\ \theta_0 - 2\sigma\ \theta_0 + 2\sigma\ ]] = 0.95$. Ma siccome θ_0 è IGNOTO, è molto più utile riscrivere la precedente inclusione nella forma $\theta_0 \in [\ \hat{\theta}(y) - 2\sigma\ \hat{\theta}(y) + 2\sigma\ ]$, in quanto la stima si calcola e la varianza è qui supposta nota. Tuttavia non è corretto dire θ_0 **appartiene ad un certo intervallo con una certa probabilità**, in quanto θ_0 NON è una v.a. Quello che possiamo dire è che **un certo intervallo contiene θ_0 con una certa probabilità**, nel senso che è l'INTERVALLO ad essere un insieme ALEATORIO, infatti l'intervallo ha estremi che sono funzioni di $\hat{\theta}(y)$, che è la VERA v.a. in questo contesto. Scriveremo quindi

$$P[\theta_0 \in C(y)] = \alpha\ (\alpha = 0.95, 0.99,\ \text{ad esempio})$$

dove $C(y)$ è una regione (aleatoria) di livello di confidenza α. In generale, θ_0 è un vettore e quindi la regione $C(y)$ un sottoinsieme (aleatorio) di $\mathbb{R}^k$, tuttavia nel caso interessante scalare essa si riduce ad un intervallo. Nel caso particolare Gaussiano scalare, ed in presenza di stimatori corretti, possiamo quindi dire che $C(y) = [\ \hat{\theta}(y) - 2\sigma\ \hat{\theta}(y) + 2\sigma\ ]$ è un intervallo di confidenza di livello 95%, mentre $C(y) = [\ \hat{\theta}(y) - 2.6\sigma\ \hat{\theta}(y) + 2.6\sigma\ ]$ un intervallo di confidenza di livello 99%. Tali intervalli (aleatori, funzioni di y, che diventano intervalli deterministici solo dopo aver OSSERVATO la misura - realizzazione - di y), indicano un range di valori in cui cade il valore vero θ_0 con una certa probabilità (tipicamente 95% o 99%), e vengono anche indicati con il termine intervallo di confidenza a 2 SD (oppure a 2.6 SD), in quanto SD sta per standard deviation, cioè σ (da non confondere con la varianza, che è σ^2). Nell'esempio precedente, supponendo lo stimatore corretto, Gaussiano, e di aver effettuato una certa misura

$$VAR = \begin{bmatrix} 100 & 3 \\ 3 & 1 \end{bmatrix},\ \hat{\theta}(y) = \begin{bmatrix} 22 \\ 10 \end{bmatrix}\ \Rightarrow \theta_1 \in [\ 2\ 42\ ],\ \theta_2 \in [\ 8\ 12\ ]$$

determinano gli intervalli di confidenza a 2 SD per θ_1, θ_2. Non sorprendente che θ_2 sia collocato in modo molto più preciso rispetto a θ_1, osservando le ben diverse varianze.

3. Proprietà asintotiche. Lo stimatore ML gode di ottime proprietà, come abbiamo visto, e si rivela spesso un buon stimatore. Tuttavia la massima espressione della potenza della stima ML sta nelle sua proprietà asintotiche, nel senso che sotto ipotesi molto generali ed ipotizzando un numero n di v.a. IID, la stima non solo migliora all'aumentare di n (si pensi alla media campionaria ed alla sua varianza $\frac{\sigma^2}{n}$, che diventa infinitesima per $n \to +\infty$, con ovvie conseguenze sull'intervallo di confidenza che diventa sempre più piccolo e localizza θ_0 con sempre maggior precisione), ma soddisfa certe condizioni espresse dal teorema che segue. Dobbiamo prima dare qualche DEF, relativa ad una SEQUENZA di stimatori (stimatori che utilizzano un numero n crescente di misure, quindi di v.a.). Per semplicità indicheremo con $\hat{\theta}_n(y)$ $(= \hat{\theta}_n(y_1, y_2, \ldots, y_n))$ tale sequenza di stimatori

- **Stimatore asintoticamente corretto (unbiased).** Se vale $lim_{n\to+\infty}\ E[\hat{\theta}_n(y)] = \theta$
- **Stimatore asintoticamente a varianza nulla.** Se vale $lim_{n\to+\infty}\ Var(\hat{\theta}_n(y)) = 0$
- **Stimatore consistente.** Se vale $lim_{n\to+\infty}\ P[||\hat{\theta}_n(y) - \theta|| < \epsilon] = 1$ (per ogni $\epsilon > 0$ ed uniformemente in θ). Si noti che ciò significa in termini pratici che lo stimatore converge in probabilità al valore VERO θ_0.

  È immediato verificare (con la disuguaglianza di Chebishev) che asintotica correttezza e varianza nulla implicano la consistenza
- **Stimatore asintoticamente efficiente (e Normale, cioè Gaussiano).** Se vale che $\sqrt{n}[\hat{\theta}_n(y) - \theta]$ tende in distribuzione alla Gaussiana a media nulla e varianza $I^{-1}(\theta)$ (Fisher riferita ad UNA SOLA v.a. y_i, supposte IID), cioè $lim_{n\to+\infty}\ \hat{\theta}_n(y) \sim \mathcal{N}(\theta_0, \frac{I^{-1}(\theta_0)}{n})$. È stato indicato θ_0 e non θ, in quanto l'asintotica efficienza implica sia la correttezza che la varianza nulla asintotica, quindi la consistenza
- **Teorema. (senza DIM)** Lo stimatore ML (sotto ragionevoli ipotesi), nel caso di v.a. IID, è sia consistente che asintoticamente efficiente

- In conclusione, non importa se ML conduce ad una stima non corretta, se è o meno efficiente nel caso corretto, se le v.a. considerate sono Gaussiane o meno. Al limite, spunta l'asintotica correttezza, Gaussianità, consistenza, efficienza. In particolare, si ricordi che $nI(\theta)$ è Fisher relativo ad n v.a. IID, quindi tende a diventare Gaussiano con media θ_0 e varianza pari a quella data dal limite di Cramer-Rao

9 Qualche Esempio

Il **primo esempio** dimostra come il precedente teorema NON valga nel caso cada l'ipotesi IID. Sia

$$y_1 = \theta_1\theta_2 + e_1,\ y_i = \theta_1 + e_i,\ (i = 2, 3, \ldots, n),\ e_i \sim \mathcal{N}(0, 1) \text{ ed indipendenti } (i = 1, 2, \ldots, n)$$

Le y_i sono indipendenti, Gaussiane e con la stessa varianza, ma non con la stessa media (la prima è $\theta_1\theta_2$, le altre valgono θ_1), quindi non sono IID. Calcoliamo

$$\nabla_\theta \ell = \begin{bmatrix} \theta_2(\theta_1\theta_2 - y_1) + \Sigma_2^n\ (\theta_1 - y_i) \\ \theta_1(\theta_1\theta_2 - y_1) \end{bmatrix},\ \nabla^2_\theta \ell = \begin{bmatrix} n - 1 + \theta_2^2 & 2\theta_1\theta_2 - y_1 \\ 2\theta_1\theta_2 - y_1 & \theta_1^2 \end{bmatrix},\ I_n(\theta) = E[\nabla^2_\theta \ell] = \begin{bmatrix} n - 1 + \theta_2^2 & \theta_1\theta_2 \\ \theta_1\theta_2 & \theta_1^2 \end{bmatrix}$$

avendo sfruttato $E[y_1] = \theta_1\theta_2$. Come si vede $I_n(\theta)$ NON ha la forma $nI(\theta)$, tipica del caso IID. Calcolando l'inversa

$$I_n^{-1}(\theta) = \begin{bmatrix} \frac{1}{n-1} & -\frac{1}{n-1}\frac{\theta_2}{\theta_1} \\ -\frac{1}{n-1}\frac{\theta_2}{\theta_1} & \frac{1}{\theta_1^2} + \frac{1}{n-1}\frac{\theta_2^2}{\theta_1^2} \end{bmatrix} \to \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{\theta_1^2} \end{bmatrix} \text{ per } n \to +\infty$$

che NON tende a zero come nel caso IID. Abbiamo

$$MSE_n(\theta) = tr(I_n^{-1}(\theta)) = \frac{1}{\theta_1^2} + \frac{1}{n-1}(1 + \frac{\theta_2^2}{\theta_1^2}),\ MSE_n(\theta_1) = \frac{1}{n-1},\ MSE_n(\theta_2) = \frac{1}{\theta_1^2} + \frac{1}{n-1}\frac{\theta_2^2}{\theta_1^2}$$

come limiti inferiori alla varianza di uno stimatore corretto, rispettivamente in senso globale (θ_1 e θ_2) e per i singoli parametri θ_1, θ_2. Mentre per θ_1 va tutto bene (MSE tende a zero), per θ_2 no (MSE tende a $\frac{1}{\theta_1^2}$). Ponendo $\nabla_\theta \ell = 0$ per trovare lo stimatore ML, si trovano due soluzioni

$$\theta_1 = 0, \theta_2 = -\frac{\Sigma_2^n\ y_i}{y_1} \Rightarrow \ell = \frac{1}{2}\Sigma_1^n\ y_i^2,\ \text{ma anche } \theta_1 = \mu, \theta_2 = \frac{y_1}{\mu} \Rightarrow \ell = \frac{1}{2}\Sigma_2^n\ (y_i - \mu)^2,\ \text{dove } \mu = \frac{1}{n-1}\Sigma_2^n\ y_i$$

Per $\theta \to \infty$ la funzione ℓ diverge, quindi il minimo assoluto va ricercato tra i due punti trovati. Ma sviluppando i conti si trova

$$\frac{1}{2}\Sigma_2^n\ (y_i - \mu)^2 = \frac{1}{2}(\Sigma_2^n\ y_i^2 - (n-1)\mu^2) \leq \frac{1}{2}\Sigma_1^n y_i^2$$

disuguaglianza che vale in senso stretto, eccetto che nel caso $y_i = 0$ per ogni i (evento a probabilità nulla, che fa perdere senso allo stimatore: a livello puramente matematico, ci sarebbero in tal caso infiniti punti di minimo in $\theta_1 = 0, \theta_2 =$ qualsiasi). La stima ML deve far riferimento al punto di minimo assoluto, e quindi

$$\hat{\theta}_1^{ML}(y) = \mu,\ \hat{\theta}_2^{ML}(y) = \frac{y_1}{\mu},\ \text{con } \mu = \frac{1}{n-1}\Sigma_2^n\ y_i$$

La stima per θ_1 è la solita media campionaria (eccetto per il primo campione y_1) di v.a. IID, quindi la sua media e varianza sono evidenti: media θ_1 e varianza $\frac{1}{n-1}$ (oltre alla Gaussianità), da cui la correttezza e l'efficienza (si veda l'inversa di Fisher) per ogni n quindi a maggior ragione al limite per $n \to +\infty$. Si ha quindi consistenza per tale stimatore, che tende al valore VERO $\theta_{1,0}$. Per θ_2 non possiamo dire nulla: valutare $E[\hat{\theta}_2]$ è un'impresa, come pure valutare la sua varianza. Tuttavia la prima equazione, che è l'unica che coinvolge θ_2, può essere rivista per n grande, nel senso che possiamo sostituire a θ_1 il suo valore vero $\theta_{1,0}$ (che è stato stimato con un'approssimazione tanto migliore quanto più n è grande), per cui essa diventa ($\simeq$ in quanto stiamo approssimando $\theta_{1,0}$ con la sua stima)

$$\frac{y_1}{\theta_{1,0}} \simeq \theta_2 + \frac{e_1}{\theta_{1,0}}$$

che è un modello lineare $\frac{y_1}{\theta_{1,0}} \sim \mathcal{N}(\theta_2, \frac{1}{\theta_{1,0}^2})$, da cui

$$\hat{\theta}_2^{(n)}(y) = \frac{y_1}{\hat{\theta}_1^{(n)}(y)} \simeq \frac{y_1}{\theta_{1,0}} \sim \mathcal{N}(\theta_2, \frac{1}{\theta_{1,0}^2})$$

da cui l'asintotica Gaussianità, correttezza ed efficienza di tale stimatore (si veda Fisher per n grande), ma NON la consistenza. Come spiegare tale fenomeno? È quasi evidente: la prima misura non dà nessuna informazione su θ_1, essendo coinvolto anche θ_2, ma tutte le altre si, e permettono di stimare θ_1 in modo consistente. Ma a questo punto la prima equazione, essendo stato ricavato $\theta_{1,0}$ con ottima precisione, diventa un'unica misura che coinvolge θ_2, e da questa si stima θ_2 come usuale in un modello lineare Gaussiano. Ma la misura è UNA SOLA, le altre non danno alcuna informazione su θ_2, quindi la varianza di tale stimatore è ben precisa e non può tendere a zero.

Il **secondo esempio** riguarda invece situazioni IID in cui il precedente teorema sulla stima ML può essere applicato. Supponiamo che si sia calcolata

$$\hat{\theta}^{ML}(y) = f(y)$$

dove $f(y) = f(y_1, y_2, \ldots, y_n)$ abbia un'espressione talmente complicata da non permetterci di valutarne media e varianza (e quindi la sua correttezza o meno ed il suo MSE). Valutando la matrice di Fisher $I(\theta)$ relativa ad UNA SOLA misura, sappiamo che per n grande si avrà che $\hat{\theta}^{ML}(y) \sim \mathcal{N}(\theta, \frac{I^{-1}(\theta)}{n})$, quindi avremo asintotica Normalità, correttezza ed MSE pari alla traccia di $I^{-1}(\theta)$ divisa per n. Quindi abbiamo informazioni sul comportamento asintotico dello stimatore ML, senza dover effettuare alcuna aspettazione sull'espressione $f(y)$ o $f^2(y)$ (si ricorda che, nel caso scalare, $E[f^2(y)] = Var(f(y)) + (E[f(y)])^2$, quindi media e varianza dello stimatore si riducono al calcolo di $E[f(y)], E[f^2(y)]$).

Il **terzo esempio** si riferisce a situazioni in cui è descritto un modello di misura della forma $y_i = f(x_i, e_i, \theta)$, dove e_i sono v.a., mentre x_i sono NUMERI (misurati, ma NUMERI, cioè privi di una descrizione probabilistica), e supponiamo che sia richiesta la stima ML di θ corrispondente a certe misure (x_i, y_i) assegnate. La prima cosa da fare è **sostituire le misure** x_i (e NON le y_i), in modo da ottenere ora un modello del tipo $y_i = g(e_i, \theta)$, per il quale si può calcolare $\hat{\theta}^{ML}(y_1, y_2, \ldots, y_n)$, ed INFINE sostituire i valori di y_i per trasformare lo stimatore in una STIMA. Questo per chiarire che espressioni molto complesse possono solo confondere le acque, ma non intaccano la sostanza della stima, che è sempre e solo basata su y, non sulla presenza di eventuali variabili aggiuntive. Come esempio, si abbia

$$y_i = \frac{\log x_i}{x_i}\theta + e_i, \ e_i \text{ Gaussiane a media nulla e varianza } \sigma^2, \text{ IID}$$

e si abbiano le misure $(y_1, x_1) = (5, 2), (y_2, x_2) = (4, 3)$. Sostituendo dapprima le x_i, otteniamo il modello LINEARE GAUSSIANO IID

$$y = \begin{bmatrix} \frac{\log 2}{2} \\ \frac{\log 3}{3} \end{bmatrix} \theta + e \ \Rightarrow \ \hat{\theta}(y) = (\phi^T \phi)^1 \phi^T y = \frac{\frac{\log 2}{2} y_1 + \frac{\log 3}{3} y_2}{\frac{(\log 2)^2}{4} + \frac{(\log 3)^2}{9}}$$

da cui possiamo calcolare $I(\theta)$ e tutto quello che ci serve, ed alla fine **sostituire** i valori di y_i per trasformare lo STIMATORE in una STIMA, cioè

$$\hat{\theta}^{ML}\left(\begin{bmatrix} 5 \\ 4 \end{bmatrix}\right) = \frac{\frac{5\log 2}{2} + \frac{4\log 3}{3}}{\frac{(\log 2)^2}{4} + \frac{(\log 3)^2}{9}} := \hat{\theta}$$

Tale numero rappresenta la STIMA ML corrispondente ai dati misurati, e non ha nulla a che vedere con lo STIMATORE, che rimane una v.a. funzione di y_1, y_2. Ciò per dimostrare come problemi che all'apparenza si presentano complessi, sono riducibili a banali modelli lineari Gaussiani, pur di utilizzare le misure nel modo corretto: prima si sostituiscono i valori di x_i, poi si costruisce lo stimatore ML (funzione di y_i), infine si valuta lo stimatore in y_i, per ottenere la STIMA di θ corrispondente a quelle misure. Stesso discorso se si volesse valutare una regione di confidenza: calcolata la varianza dello stimatore, sia ν^2, si avrebbe un intervallo centrato in $\hat{\theta}$ di ampiezza 4ν oppure 5.2ν (a seconda che sia richiesta confidenza a 2 SD oppure a 2.6 SD)