CHAPTER 5

Maximum Likelihood e Minimi Quadrati

5.1 La Verosomiglianza - Likelihood

Date N misure $\{x_i\}_i^N$ queste si definiscono IID quando sono:

- **indipendenti:** l'esito di un campionamento non è influenzato da nessuno degli altri
- **identicamente distribuiti:** Tutte le misure seguono la stessa funzione di distribuzione di pobabilità

$$pdf_x(x, \underline{\theta}) : R \to R^+$$

La funzione di probabilità congiunta (joint-pdf) di N campionamenti IID è il prodotto delle singole probabilità (poichè indipendenti tra loro):

$$pdf_{set}(x_1....., x_N, \underline{\theta}) = \prod_i^N pdf_x(x_i, \underline{\theta_t})$$

essa rappresenta la densità di probabilità da associare all'evento casuale consistente nell'estrarre un particolare set di dati, ed è una funzione definita su uno spazio N-dimensionale.

Se si sostituisce al valore vero $\underline{\theta_t}$ il generico valore $\underline{\hat{\theta}}$ stimato dalle N misure e se esse vengono considerate non più variabili casuali, ma costanti che sono state determinate dalle operazioni di misura, la precedente funzione prende il nome di *funzione di verosomiglianza* o **likelihood**.

$$L(\underline{x}, \underline{\hat{\theta}}) = \prod_i^N pdf_x(x_i, \underline{\hat{\theta}})$$

Rappresenta la densità di probabilità da associare all-evento casuale consistente nell'essere un certo $\underline{\hat{\theta}}$ il valore vero del nostro parametro, nell'ipotesi di avere già ottenuto un set di N misure.

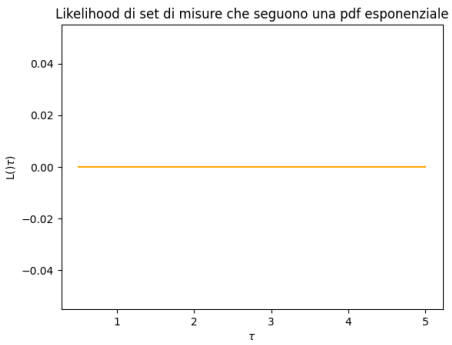
È possibile definire anche la **loglikelihood** ovvero:

$$l(\underline{\theta}) = \log(L(\underline{x}, \underline{\hat{\theta}})) = \log(\prod_i^N pdf_x(x_i, \underline{\hat{\theta}})) = \sum_i^N \log(pdf_x(x_i, \underline{\theta})$$

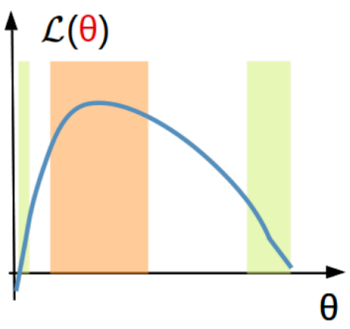
5.2 Comportamento della likelihood

Raccolte delle misure $\{x_i\}_i^N$ IID, la funzione di likelihood $L(\theta)$ restituisce la probabilità che si aveva di raccogliere tale campione assunta la conoscenza del parametro θ (che per comodità in questo capitolo assumiamo in una sola dimensione).

Se la $L(\theta)$ è all'incirca piatta vuol dire che si ha la stessa probabilità di ottenere un campionamento $\{x_1, ..., x_N\}$ per ogni valore di θ, ciò significa che i campionamenti non forniscono molte informazioni sul parametro θ.



Se la $L(\theta)$ è una campana vuol dire si avrà una probabilità:



- Alta se il θ_t del campione è nell'area arancione in figura
- bassa se θ_t è nella regione verde
- massima se θ_t coincide con il massimo della funzione di likelihood

di aver raccolto il set di dati in esame. Dunque i campionamenti forniscono informazioni su θ. Più è stretta la campana e più piccolo è il range dei valori che massimizzano la probabilità di ottenere il campione raccolto. La larghezza della campana è legata all'incertezza con cui è possibile determinare il valore vero di θ.

Quando definiamo una $pdf(x, \theta)$ senza conoscere θ si descrive un fascio di distribuzioni di probabilità e dunque un'infinità di modelli diversi. Cambiando la varianza di una distribuzione di cambia anche la funzione di likelihood.

5.3 Minimum Variance Bound

La funzione di likelihood $L(\theta)$ può essere utilizzata per misurare l'informazione che i campionamenti contengono relativamente al parametro θ che descrive il modello. L'informazione così ottenuta consente di valulatare la minima varianza raggiungibile di uno stimatore di $\hat{\theta}$, ovvero dato un insieme di misure e un modello esiste un limite inferiore alla varianza raggiungibile.

5.3.1 Informazione di Fischer

Vogliamo costruire una metrica che $I(\theta)$ che fornisca l'ammontare d'informazione contenuta in una variabile casuale osservabile x, relativamente a un parametro non osservabile θ, da cui dipende la pdf(x). Tale funzione prende il nome d'**informazione di Fischer**.

$$I(\theta) = E[(\frac{\partial}{\partial\theta} log(L(\theta)))^2] = -E[\frac{\partial^2}{\partial\theta^2} log(L(\theta)]$$

essa gode delle seguenti proprietà:

- Associa un valore nullo se i dati sono irrilevanti per la stima di θ.
- Aumenta al crescere della dimensione del campione (purchè i dati siano rilevanti per la stima di θ).
- È legata alla precisione della stima: se l'informazione aumenta la varianza minima raggiungibile da uno stimatore $\hat{\theta}$ diminuisce.

Più i valori assunti dalla likelihood risultano essere sparpagliati peggiore è l'informazione che si ha a disposizione e dunque la precisione di stima di uno stimatore ne risentirà in egual modo. Viceversa più i valori risultano essere raggruppati in un certa regione di spazio e migliore sarà la precisione con cui si ottiene lo stimatore.

5.3.2 Teorema di Rao - Cramér

Per uno stimatore $\hat{\theta}(x)$ consistente, bias $b_n(\hat{\theta})$, $V[\hat{\theta}] < \infty$ e non dipendente da x, definito su un dominio Ω_θ si ha che il **MVB (Minimum Variance Bound)** definisce una relazione tra la varianza di un qualsiasi stimato $\hat{\theta}$ e la likelihood $L(\underline{X}, \theta)$ nel seguente modo:

$$V[\hat{\theta}] \geq \frac{(1 - \frac{\partial}{\partial\theta} b_n(\theta))^2}{E[(\frac{\partial}{\partial\theta} log(L(\theta)))^2]} = V[\hat{\theta}]_{min} \tag{5.1}$$

Al numeratore si ha una quantità che dipende dallo specifico stimatore $\hat{\theta}$, ma solo se questo è biased, mentre a denominatore si ha l'informazione di Fischer, dunque una quantità che non dipende dallo specifico stimatore ma solo da modello e dati.

Diciamo che uno stimatore è **efficiente** se la grandezza:

$$\xi(\hat{\theta}) = \frac{V[\hat{\theta}]_{min}}{V[\hat{\theta}]} = 1 \tag{5.2}$$

5.4 Maximum Likelihood

Considerato un campione di N misure IID, si definiscono stimatori di **Maximum Likelihood (ML)** dei parametri quei valori che massimizzano la funzione di likelihood. Data una likelihood differenziabile rispetto $\underline{\theta}$ e il cui punto di massimo non è ai margini del range dei parametri, gli stimatori $\underline{\hat{\theta}}$ sono dati dalla soluzione delle equazioni differenziali:

$$\frac{\partial L}{\partial \theta_i} = 0 \quad i = 1, ..., N \tag{5.3}$$

Se esiste più di un massimo, viene considerato il maggiore.

Con questa definizione di ML si considera il valore dello stimatore in corrispondenza del quale la probabilità associata al campionamento è la massima possibile.
Le proprietà degli stimatori ottenute con il metodo di ML valgono asintoticamente, ovvero con un numero sufficientemente elevato di misure. Alcune **proprietà notevoli** sono:

1. Consistente
2. Asintoticamente efficiente
3. Asintoticamente non distorto
4. Proprietà d'invarianza: $\hat{\theta}_{ML}$ stimatore di $\theta_t \Rightarrow g(\hat{\theta}_{ML})$ è stimatore di $g(\theta_t)$.

Esempio proprietà d'invarianza

Consideriamo la distribuzione esponenziale $f(x,\tau) = \frac{1}{\tau}e^{\frac{-t}{\tau}}$ e si abbiamo N misure IID la likelihood rispetto a τ è $L(\tau) = \prod_i^N \frac{1}{\tau}e^{\frac{-t_i}{\tau}}$.

Applichiamo il metodo di ML alla loglikelihood di τ, dunque si ha che:

$$\sum_i^N(-\frac{1}{\tau} + \frac{t_i}{\tau^2}) = 0 \iff N = \sum_i^N \frac{t_i}{\tau} \iff \tau_{ML} = \frac{1}{N}\sum_i^N t_i$$

Posto $\lambda(\tau) = \frac{1}{\tau}$ possiamo riscrivere la pdf come $f(x,\lambda) = \lambda e^{-\lambda t}$ data la sostituzione vogliamo verificare che $\lambda(\tau_{ML}) = \frac{1}{\tau_{ML}}$.

Dim.

$$\frac{\partial l}{\partial \theta} = 0 \iff \sum_i^N(\frac{1}{\lambda} - t_i) = 0 \iff \frac{N}{\lambda} = \sum_i^N t_i \iff \lambda = \frac{1}{\tau}$$

Se si dimostra che $\hat{\tau}$ è uno stimatore non distorto, è anche vero che $\hat{\lambda}$ è anch'esso non distorto ?

Poichè la relazione $\hat{\lambda}(\hat{\tau})$ può essere non lineare non è detto che anche $\hat{\lambda}$ sia non distorto.
Si può dimostrare che $\hat{\lambda}$ è uno stimatore distorto di λ e dunque solo asintoticamente non lo è.

5.5 Varianza dello stimatore di ML - Metodo Grafico

Espandiamo fino al secondo ordine con un polinomio di Taylor la loglikelihood in un'intorno del parametro $\hat{\theta}$ ottenuto con il metodo di ML.

$$log(L(\theta)) \approx log(L(\hat{\theta})) + \frac{1}{2}\frac{d^2}{d\theta^2}(log(L(\theta))|_{\theta=\hat{\theta}}(\theta - \hat{\theta})^2 \tag{5.4}$$

Se assumiamo che lo stimatore sia unbiased (il che per le proprietà precedente asintoticamente può esserlo) ed efficiente si ha dal teorema di Rao-Cramer che il secondo addendo è equivalente a:

$$\frac{d^2}{d\theta^2}(log(L(\theta))|_{\theta=\hat{\theta}} = -\frac{1}{\sigma_{\hat{\theta}_{ML}}}$$

E dunque possiamo riscrivere l'equazione (8.6) come:

$$log(L(\theta)) \approx log(L(\hat{\theta})) - \frac{1}{2\sigma_{\hat{\theta}_{ML}}}(\theta - \hat{\theta})^2 \tag{5.5}$$

Consideriamo un cambio di variabile da $\hat{\theta}$ a $\hat{\theta} \pm \sigma_{\hat{\theta}}$ dunque possiamo riscrive la (8.7) come:

$$log(L(\hat{\theta} \pm \sigma_{\hat{\theta}}) = log(L(\hat{\theta})) - \frac{1}{2} \tag{5.6}$$

dunque da tale cambia di variabile si ha che la loglikelihood diminuisce della metà del suo valore di massimo. Si può dimostrare che la funzione di loglikelihood diventa una parabola (e la funzione di likelihood diventa una Gaussiana) per valori grandi del campione. Anche se la log(L) non è parabolica si può utilizzare l'equazione (8.8) come la definizione di errore

statistico. Determinando i punti d'intersezione con quanto definito in (8.8) si ottiene una stima della varianza dello stimatore $\hat{\theta}$ usando il metodo di ML.

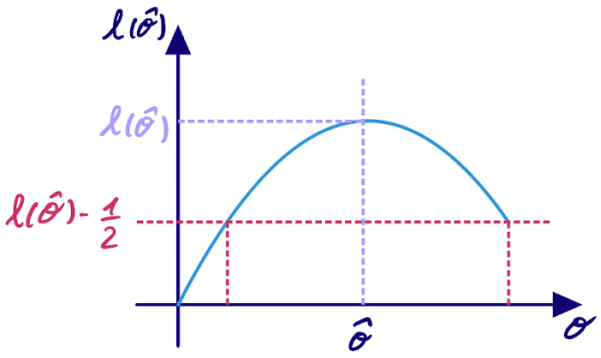


Figure 5.1: funzione di loglikelihood, stima della varianza

Con un numero finito di misure l(θ) non è simmetrico e al crescere del numero di misure la forma funzionale assume un aspetto parabolico. Per simmetria la likelihood L(θ) diventa una Gaussiana.

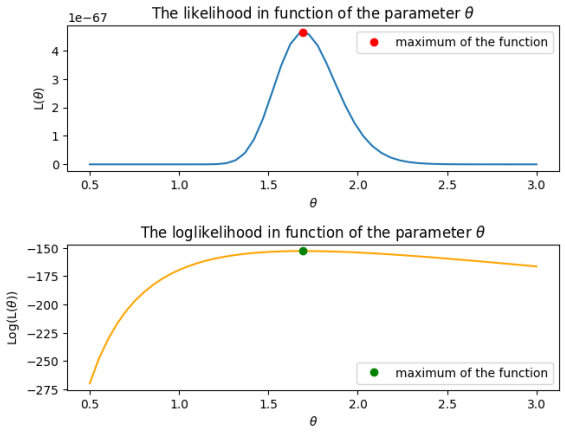


Figure 5.2: likelihood per un campione di 800 misure

5.6 Extended likelihood

Consideriamo una variabile aleatoria x distribuita secondo una pfd f(x,$\underline{\theta}$), dove i parametri $\underline{\theta}$ non li conosciamo e ipotizziamo di avere un campione di N misure $\{x_1, ..., x_N\}$. Spesso si ha che il numero di osservazioni n del campione è una variabile aleatoria che segue una pdf Poissoniana con frequenza media λ. Il risultato di un'esperimento può essere definito dalla variabile n (che è la dimensione del campione) e dalle n misure raccolte $\{x_1, ..., x_n\}$.
Dunque possiamo definire la likelihood dei parametri $\underline{\theta}$, dato un numero di venti medio λ rispetto ad un campione di n misure.

$$L(\lambda, \theta) = \frac{\lambda^{-n}}{n!} e^{-\lambda} \prod_{i=1}^{n} f(x_i, \underline{\theta}) = \frac{e^{-\lambda}}{n!} \prod_{i=1}^{n} \lambda_i f(x_i, \underline{\theta}) \tag{5.7}$$

Tale funzione prende il nome di **extended likelihood.**

5.7 Varianza di uno stimatore per più parametri

La varianza di $V[\hat{\theta}]$ di un vettore di parametri è definita dalla matrice di covarianza dove le sue componenti sono date da:

$$V_{ij} = \Big[- \frac{\partial^2}{\partial\theta_i \partial\theta_j} l(\hat{\theta}) \Big]^{-1} \quad \text{per} \quad i, j = 1, ..., n \tag{5.8}$$

dove i termini per $i \neq j$ sono dati da $Cov[\theta_i, \theta_j]$.

5.8 Intervallo di confidenza

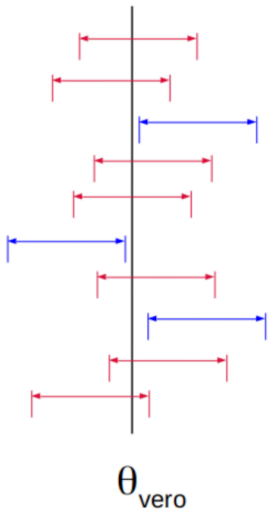
Si è costruito uno stimatore $\hat{\theta}$ per un parametro θ che è una variabile aleatoria di conseguenza esiste una distribuzione pdf($\hat{\theta}$) che è caratterizzata da una media E[$\hat{\theta}$] e una varianza V[$\hat{\theta}$]. La **stima** è il valore che lo stimatore assume in corrispondenza di un campione di dati e lo definiamo $\hat{\theta}^*$.

L'**intervallo di confidenza** è solitamente indicato come:

$$\hat{\theta}^* \pm \sigma \quad \text{oppure} \quad \hat{\theta}^{*\,+\sigma_1}_{\,-\sigma_1} \tag{5.9}$$

la seconda notazione viene usata quando l'intervallo è asimmetrico.

In generale il risultato della stima è un intervallo [a,b], poichè è una variabile aleatoria si ha che gli estremanti sono variabili aleatorie e dunque per diversi campioni si hanno diversi intervalli di confidenza. A ciascun intervallo viene associata una probabilità (**livello di confidenza**) che misura quanto è buona la stima del valore vero.



$$P(\theta_t \in [a,b])$$

dove essa rappresenta la frazione delle volte in cui ripetendo l'esperimento, la stima restituisce un intervallo che contiene il valore vero.

In generale si vuole trovare un metodo che individui un intervallo [a,b] tale per cui la probabilità che $\theta_t \in [a,b]$ sia pari a un certo valore β detto **livello di confidenza**. Un intervallo di confidenza così definito prende il nome di **intervallo di condifenza di Neyman**. Un intervallo particolare è quello dato da $[\hat{\theta}^* - \sigma, \hat{\theta}^* + \sigma]$ e si definisce **intervallo centrale**.

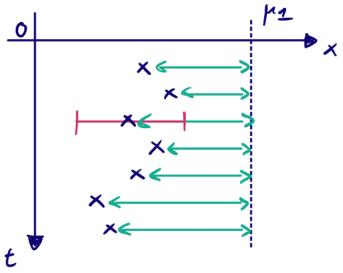
5.9 Metodo dei minimi quadrati

Si parte da un modello $y = f(x, \underline{\theta})$ che definisce una relazione tra due grandezze fisiche x ed y e che dipende da N parametri $\underline{\theta}$, ipotizziamo di avere un campione di misure formato da punti $\{(x_i, y_i)\}_i^N$ ciascuno di essi segue una pdf(x) e una pdf(y) di conseguenza sono variabili aleatorie che

hanno una loro incertezza σ_x e σ_y. Vogliamo trovare un metodo che ci permetta di stimare i parametri ignoti partendo dai dati sperimentali. Tale metodo prende il nome di **interpolazione** o **fit** dei dati. Per semplicità consideriamo il caso il cui l'incertezza sulle misure x sia trascurabile.

5.9.1 Funzionale Q^2

Partiamo considerando solo la variabile aleatoria x e ci domandiamo per quali valori del campione, si minimizza la distanza dalla media della popolazione $E[x] = \mu$; per rispondere a tale domanda definiamo la misura della distanza definendo un funzionale:



$$Q^2\Big(\{x_i\}_1^N, \mu\Big) = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{\sigma_i^2} \tag{5.10}$$

Dunque ciascuna distanza è normalizzata alla larghezza della $pdf(x_i)$ e quindi tiene conto delle fluttuazioni statistiche di quel determinato campionamento.
Per determinare il parametro cerchiamo quel valore $\hat{\mu}_{MQ}$ per cui Q^2 ammette un minimo assoluto nello spazio dei parametri (che nel nostro caso ha dimensione 1):

$$\frac{d}{d\mu}Q^2\Big(\{x_i\}_1^N, \mu\Big) = 0 \tag{5.11}$$

5.9.2 Valore di aspettazione di $\hat{\mu}_{MQ}$

Derivando la funzione di Q^2 si ha:

$$\frac{d}{d\mu}Q^2 = -2\sum_{i=1}^{N}\frac{(x-\hat{\mu}_{MQ})}{\sigma_i^2} = 0 \iff \hat{\mu}_{MQ} = \frac{\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{N}\frac{1}{\sigma_i^2}} = \overline{x}$$

dunque per un campione di N misure la stima del valore atteso della popolazione coincide con la media aritmetica pesata del campione.

5.9.3 Varianza dello stimatore $\hat{\mu}_{MQ}$

Consideriamo $\hat{\mu}_{MQ} = \phi(x_1, ..., x_N)$ e che il campione sia costituito da misure statisticamente indipendenti rispetto la variabile aleatoria x, la varianza dello stimatore sará data da:

$$V[\hat{\mu}_{MQ}] = \sum_{i=1}^{N} \Big(\frac{\partial \phi}{\partial x_i}\Big)^2 \sigma_i^2 = \Big[\sum_{i=1}^{N} \frac{1}{\sigma_i^2}\Big]^{-2} \sum_{i=1}^{N} \Big(\frac{1}{\sigma^2}\Big)^2 \sigma_i^2 = \frac{1}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}} < \sigma_i^2 \quad \forall i$$

L'incertezza è dominata dalle misure con le σ_i più piccole.

5.10 Varianza di un stimatore usando i MQ - Metodo grafico

Con il metodo di maximum likelihood si sono cercati i valori dei parametri che rendevano massima la probabilità, dato un modello, di osservare i dati campionati. Con il metodo dei minimi quadrati invece si ceca il valore dei parametri che rende minima la distanza tra i dati campionati e il modello. Quando le $pdf(x_i)$ sono Gaussiane i due stimatori coincidono.

$$L(\hat{\theta}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\Big[-\frac{(x_i - \hat{\theta})^2}{2\sigma_i^2}\Big] = cost \cdot e^{-\frac{Q^2}{2}}$$

se passiamo alla loglikelihood

$$\log(L(\hat{\theta})) = \log(\text{cost}) - \frac{Q^2(\hat{\theta})}{2}$$

Stimando $\hat{\theta}$ con il metodo di Maximum Likelihood si ha

$$\frac{\partial \log(L(\hat{\theta}))}{\partial \hat{\theta}} = -\frac{1}{2}\frac{\partial Q^2(\hat{\theta})}{\partial \hat{\theta}} \tag{5.12}$$

Quandole misure seguono una pdf Gaussiana e lo stimatore $\hat{\theta}$ è efficiente e non distorto sappiamo che il MVB coincide con la varianza dello stimatore

ottenuto con il metodo di ML.

$$V[\hat{\theta}_{ML}] = -\frac{1}{\frac{\partial^2 \log(L(\theta))}{\partial\theta^2}\Big|_{\theta=\hat{\theta}_{ML}}} \tag{5.13}$$

Dalla relazione (5.12) possiamo riscrivere l'uguaglianza (5.13) come:

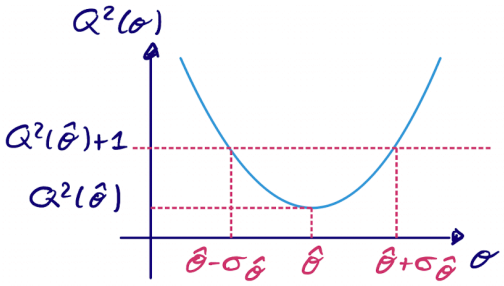
$$V[\hat{\theta}_{MQ}] = \frac{2}{\frac{\partial^2 Q^2(\theta)}{\partial\theta^2}\Big|_{\theta=\hat{\theta}_{MQ}}} \tag{5.14}$$

Sviluppando con Taylor al secondo ordine la funzione Q^2 in un intorno di $\hat{\theta}_{MQ}$ si ha:

$$Q^2(\theta) \approx Q^2(\hat{\theta}_{MQ}) + \frac{1}{2}\frac{d^2Q^2(\theta)}{d\theta^2}\Big|_{\theta=\hat{\theta}_{MQ}}(\theta - \hat{\theta}_{MQ})^2 = Q^2(\hat{\theta}_{MQ}) + \frac{(\theta - \hat{\theta}_{MQ})^2}{V[\theta]}$$

Effettuando un cambio di coordinate $\theta = \hat{\theta}_{MQ} \pm \sigma_{\hat{\theta}}$ si ha:

$$Q^2(\hat{\theta}_{MQ} \pm \sigma_{\hat{\theta}}) = Q^2(\hat{\theta}_{MQ}) + 1$$



Dunque considerando graficamente l'intersezione con $Q^2(\theta)$ si determinano $\hat{\theta}_{MQ} \pm \sigma_{\hat{\theta}}$. Poichè per ipotesi le misure $\{x_i\}_i^N$ seguono una distribuzione di probabilità Gaussiana e sono IID si ha che $Q^2(\theta)$ segue la distribuzione di probabilità del chi quadro $\chi^1(N-k)$ per E$[Q^2]$ = N-k gradi di libertà. Se questo non avviene è perchè:

- si è avuta una fluttuazione statistica sfavorevole
- il modello non descrive i dati
- il modello è corretto, ma i dati sono stati raccolti in modo errato

- il modello è corretto, ma le incertezze attribuite ai dati sono sbagliate (sono sopra o sottovalutate).

5.11 Modelli lineari nei parametri

Un modello si definisce **lineare** se due grandezza fisiche x ed y sono legate da una relazione lineare rispetto a parametr $\underline{\theta}$.

$$y_0^i = \sum_{j=1}^{N} \theta_j h_j(x_i) \tag{5.15}$$

Assumiamo che l'incerteza sulla variabile x sia ininfluente, mentre y è una variabile aleatoria le cui misure sono IID con varianza $V[y_i] = \sigma_i^2$. Definiamo una variabile aleatoria asuiliaria ϵ che possiede la stessa distribuzione di probabilità di y. Riscriviamo (5.15) come:

$$y_i = \sum_{j=1}^{N} \theta_j h_j(x_i) + \epsilon_i \tag{5.16}$$

Assumiamo che per le ϵ valgano le seguenti proprietá:

$$E[\epsilon_i] = 0 \qquad V[\epsilon_i] = V[y_i] = \sigma_i^2$$

In questo modo la pdf(y) coincide con la pdf(ϵ). La variabile ϵ rappresenta l'errore statistico sulla misura, mentre le y_0 sono il valore vero della misura restituito dal modello.
Applichiamo il metodo dei minimi quadrati, andando a stimare quei valori dei parametri $\underline{\theta}$ che minimizzano la distanza tra il valore vero restituito dal modello lineare e i dati campionati sperimentalmente, pesando gli scarti quadratici rispetto la varianza delle distribuzioni di probabilità delle singole y_i.

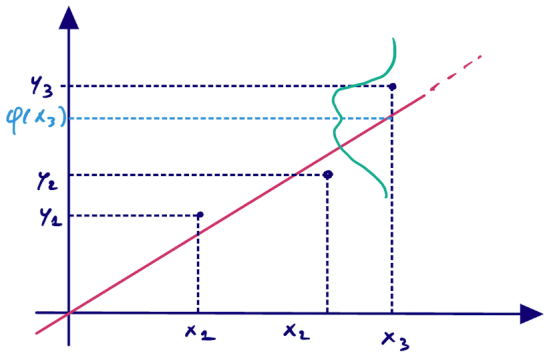


Figure 5.3: $\psi(x_3)$ rappresenta il valore vero della misura rispetto al modello, mentre y_3 il dato campionato su cui è presente un errore statistico.

Per farlo utilizziamo il funzionale Q^2 riscrivendolo come:

$$Q^2 = \sum_{i=1}^{N} \frac{(y_i - y_0(x_i, \theta))^2}{\sigma_i^2} = \sum_{i=1}^{N} \frac{(y_i - \sum_{j=1}^{N} \theta_j h_j(x_i))^2}{\sigma_i^2} = \sum_{i=1}^{N} \frac{\epsilon_i}{\sigma_i^2} \tag{5.17}$$

5.11.1 Stime dei parametri $\underline{\theta}$

Per un vettore di parametri $\underline{\theta}$ di dimensione N, avremo un sistema di K equazioni:

$$\nabla Q^2(\underline{x}, \underline{\theta}) = \begin{cases} \dfrac{\partial Q^2}{\partial \theta_1} = 0 \\ \vdots \\ \dfrac{\partial Q^2}{\partial \theta_n} = 0 \end{cases} \tag{5.18}$$

Per una singola riga si ha:

$$\frac{\partial Q^2}{\partial \theta_j} = \sum_{i=1}^{N} \frac{-2}{\sigma_j^2} \Big(\frac{y_i - \sum_{j=1}^{N} \theta_j h_j(x_i)}{\sigma_i} \Big)^2 h_j(x_i) = 0 \quad \forall i = 1, ..., N \tag{5.19}$$

Per comodità di esposizione riscriviamo l'espressione (5.16) in forma vetto-

riale:

$$\underline{y} = H\underline{\theta} + \underline{\epsilon} \tag{5.20}$$

ed essendo una modello a più parametri si ha la matrice di covarianza $V[\epsilon]$, possiamo considerare la matrice in funzione di ϵ poichè V[y] e $V[\epsilon]$ sono legate da una traslazione rigida e dunque risultano essere identiche. In generale tale matrice non è diagonale, ma poichè si è assunto che le misure siano IID la matrice è diagonale.

$$V[\epsilon] = \begin{bmatrix} \sigma_i^2 & \cdots\cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

Riscrivendo $Q^2 = \underline{\epsilon}^T V^{-1} \underline{\epsilon}$ in forma matriciale e risolvendo il sistema (5.19) si stimano i parametri come:

$$\underline{\hat{\theta}}_{MQ} = (H^T V^{-1} H)^{-1} \cdot H^T V^{-1} \underline{y} \tag{5.21}$$

dove $V(\underline{\hat{\theta}}) = (H^T V^{-1} H)^{-1}$ dove nella matrice H è contenuta l'informazione delle misure.
Resta da verificare che lo stimatore sia non distorto ovvero $E(\underline{\hat{\theta}}) = \underline{\theta}_t$:

$$E[\underline{\hat{\theta}}_{MQ}] = (H^T V^{-1} H)^{-1} \cdot H^T V^{-1} E[\underline{y}] = (H^T V^{-1} H)^{-1} \cdot (H^T V^{-1} H)\underline{\theta_t}$$

dove:

$$E[\underline{y}] = E[H\underline{\hat{\theta}} + \underline{\epsilon}] = E[H\underline{\hat{\theta}}] + E[\underline{\epsilon}] = HE[\underline{\hat{\theta}}] = H\underline{\theta}_t$$

duque lo stimatore ottenuto è non distorto.
A differenza del metodo di ML che gode di buone proprietà asintoticamente si ha che quello dei MQ le ha per un numero finito di misure.

5.12 Sovrastima degli errori

Ipotizziamo che le incertezze siano sovrastimate di un fatto α ovvero $\sigma_i^* = \alpha \cdot \sigma_i$, di conseguenza possiamo riscrivere la matrice di covarianza come:

$$W[y^*] = \alpha^2 \cdot \begin{bmatrix} \sigma_i^2 & \cdots\cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix} = \alpha^2 \cdot V[y]$$

Se sostituiamo tale ipotesi all'interno dell'equazione (5.21) si ha:

$$\underline{\hat{\theta}} = (H^T W^{-1} H)^{-1} \cdot H^T W^{-1} \underline{y} = \alpha^2 \cdot \frac{1}{\alpha^2} \cdot (H^T V^{-1} H)^{-1} \cdot H^T V^{-1} \underline{y} = \quad (5.22)$$

$$= (H^T V^{-1} H)^{-1} \cdot H^T V^{-1} \underline{y}$$

dunque la stima del parametro $\underline{\hat{\theta}}_{MQ}$ rimane invariata anche se si sono sovrastimate le incertezze. Quella che cambia è la varianza del parametro stimato infatti:

$$V[\underline{\hat{\theta}}_{MQ}] = (H^T W^{-1} H)^{-1} = \alpha^2 (H^T V^{-1} H)^{-1} \quad (5.23)$$

e dunque viene sovrastimata di un fattore α^2.

5.12.1 Relazione tra il numero di parametri e il campione

Se si hanno K parametri e campione di dimensione N dove K=N si ha che:

- il sistema ammette soluzione, e si ha una curva che passa per tutti i punti;
- se il sistema non ammette soluzione allora i dati falsificano il modello.

5.12.2 Incertezze sulla variabile indipendente

Nel caso in cui il modello sia lineare rispetto ai parametri, per esempio una retta $y(x, a, b) = a + bx$ e siano presenti delle incertezze sulla variabile indipendente x, si propaga l'errore sulle y_i l'errore di σ_{x_i} ottenendo un errore

complessivo $\sigma_i^2 = \sigma_{y_i}^2 + b\sigma_{x_i}^2$ e dunque il funzionale Q^2 si riscrive come:

$$Q^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{(y_i - y)^2}{\sigma_{y_i}^2 + b\sigma_{x_i}^2} \tag{5.24}$$

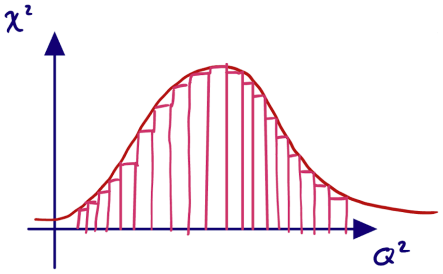
5.12.3 Stima del fattore di sovrastima α

Ipotizziamo che gli $\underline{\epsilon}$ seguano una distribuzione di probabilità Gaussiana e di avere N misure indipendenti allora il funzione Q^2 definito come l'ultima uguaglianza della (5.17) è segue la distribuzione di $\chi^2(N-K)$ per N-K gradi di libertà dove K è il numero di parametri da stimare. In forma matriciale possiamo riscriverlo rispetto alla matrice di sovrastima W come:

$$\hat{Q^2} = \epsilon^T W^{-1}\epsilon = \frac{1}{\alpha^2}\cdot\epsilon^T W^{-1}\epsilon = \alpha^{-2}\cdot Q^2 \tag{5.25}$$

Vogliamo determinare α^2 affinchè $\hat{Q}^2_{min} = E[\hat{Q^2}] = N - K$ e dunque:

$$\alpha^2 = \frac{\hat{Q}^2_{min}}{N-K} \tag{5.26}$$



Il fattore di scala così determinato non costituisce la sovrastima "reale" di cui si sono sbagliate le misure, poichè $\hat{Q^2}_{min}(\underline{x}, \underline{\hat{\theta}}_{MQ})$ per $\underline{\hat{\theta}}_{MQ}$ fissato è anch'esso una variabile aleatoria che segue la distribuzione di $\chi^2(N-K)$ di conseguenza la (5.26) stima un solo valore rispetto al campione a disposizione, ma α^2 è una variabile aleatoria che segue la medesima distribuzione di $\hat{Q^2}_{min}(\underline{x}, \underline{\hat{\theta}}_{MQ})$ e dunque per determinarlo bisogna ricostruire la distribuzione del χ^2.

5.13 Teorema di Gauss-Markov

Si consideri un insieme di variabili aleatorie statisticamente indipendenti $\{(x_i, y_i)\}_i^N$ tali che sono legate tra loro da una relazione lineare rispetto ai parametri $y_i = \psi(x_i, \underline{\theta}) + \epsilon_i$ dove $E[\epsilon_i] = 0$ e $V[\epsilon_i]$ finita $\forall i$ (proprietà di

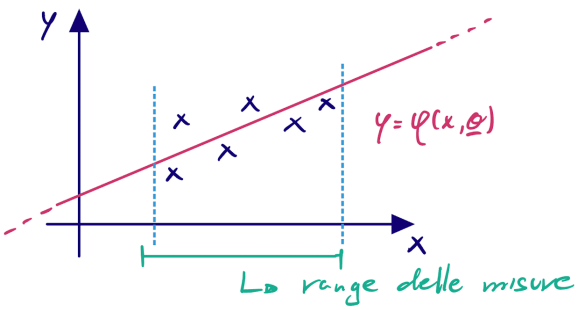
omoschedasticità) ed inoltre y_i indipendenti dai parametri $\underline{\theta} \Rightarrow$ si ha che lo stimatore $\underline{\hat{\theta}}$ ottenuto con il metodo dei minimi quadrati è non distorto e ha varianza minima fra tutti gli stimatori lineari.

Osservazioni

- ϵ_i non è necessario che siano distribuite secondo una Gaussiana
- il teorema non ci dice che si è trovato lo stimatore più **efficiente** fra tutti gli stimatori possibili, ma sicuramente quello pi'u efficiente tra quelli lineari.

È possibile trovare uno stimatore più efficiente per i parametri $\underline{\theta}$ però dovremmo determinarlo con una funzione non lineare e ammesso che questa esista non è garantito che lo stimatore ottenuto sia non distorto.
Gli stimatori descritti dal teorema di Gauss-Markov vengono definiti **B.L.U.E (Best Linear Unbiased Estimator)**.

5.14 Interpolazione ed Estrapolazione



Assumendo che i parametri $\underline{\theta}$ siano già stati determinati. **L'interpolazione** è la determinazione del valore di y mediante la funzione $\psi(x, \underline{\theta})$ per una misura x contenuta all'interno del range delle misure, ovvero l'intervallo contente i dati del campione. Si definisce **estrapolazione** il valore di y per qualsiasi altro valore di x non compreso all'interno del range delle misure definito sperimentalmente. Per una valore ottenuto da una retta di regressione lineare, l'errore è legato all'incertezza nella stima dei parametri di $\underline{\theta}$. Di conseguenza:

$$V[y]_i = \nabla\psi(x, \underline{\theta})^T \cdot Cov[\underline{\hat{\theta}}_{MQ}] \cdot \nabla\psi(x, \underline{\theta}) \tag{5.27}$$

Per esempio in un caso a due parametri si avrà:

$$V[y]_i = \begin{bmatrix} \frac{\partial \psi}{\partial \theta_1} & \frac{\partial \psi}{\partial \theta_2} \end{bmatrix} \cdot \begin{bmatrix} \sigma^2_{\theta_1} & Cov(\theta_1, \theta_2) \\ Cov(\theta_2, \theta_1) & \sigma^2_{\theta_2} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial \psi}{\partial \theta_1} \\ \frac{\partial \psi}{\partial \theta_2} \end{bmatrix}$$

che coincide con la formula di propagazione degli errori presente al capitolo 3. In generale l'errore aumenta allontanandosi dalla regione di campionamento (quindi quando si passa dall'interpolazione all'estrapolazione).

Esempio di fit lineare

Consideriamo di avere un set di misure $\{(x_i, y_i)\}_i^N$ IID con incertezza σ_{y_i} sulla variabile aleatoria y, e che il modello è quello di una retta y = a+bx, stimiamo i parametri a e b usando il metodo dei minimi quadrati. Avremo che:

$$\vec{\theta} = \begin{bmatrix} a \\ b \end{bmatrix} \qquad \vec{h} = \begin{bmatrix} 1 \\ x \end{bmatrix}$$

e poichè le misure sono indipendenti tra loro, la matrice di covarianza è diagonale. Applicando il metodo dei MQ si stimano:

$$\hat{a} = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}} \sum_{i=1}^{N} \frac{y_i - \hat{b} x_i}{\sigma_i^2} = \overline{y} - \hat{b}\overline{x}$$

$$\hat{b} = \frac{1}{\sum_{i=1}^{N} \left(\frac{x_i}{\sigma_i}\right)^2} \sum_{i=1}^{N} \frac{y_i x_i - \hat{a} x_i}{\sigma_i^2} = \frac{\overline{x}\overline{y} - \overline{xy}}{\overline{x^2} - \overline{x}^2}$$

e le incertezze associate alle stime sono date dalla matrice di covarianza associate ai parametri:

$$V[\underline{\hat{\theta}}] = \begin{bmatrix} \sigma_a^2 & Cov[a,b] \\ Cov[b,a] & \sigma_b^2 \end{bmatrix} = \frac{\sigma^2}{N(\overline{x^2} - \overline{x}^2)} \cdot \begin{bmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{bmatrix}$$

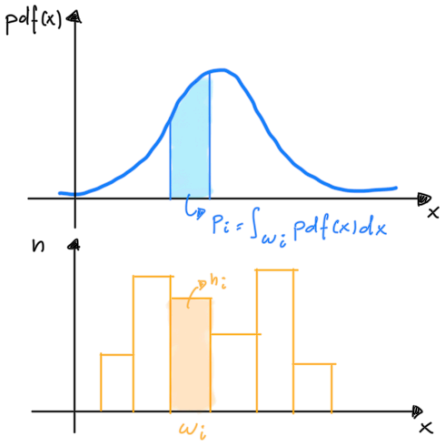
I termini in [redacted] viene definito **dispersione delle misure**. Si osserva che non necessariamente i parametri sono scorrelati tra loro. Più le msiure sono sparpagliate meigliore è la precisione con cui si stimano i parametri, poichè la banda che definisce le possibili rette che fittano i dati diventa più sottile, inoltre anche il numero di misure fornisce un contributo alla precisione con cui si determinano gli stimatori.

Se calcoliamo il valore assunto dalla funzione in punto x_0 rispetto ai parametri $\hat{a}$ e $\hat{b}$ si ricava che l'errore su y_0 è dato da:

$$V[y_0] = \frac{\sigma^2}{N}\left[1 + \frac{(x_0 - \overline{x})^2}{(\overline{x^2} - \overline{x}^2)^2}\right]$$

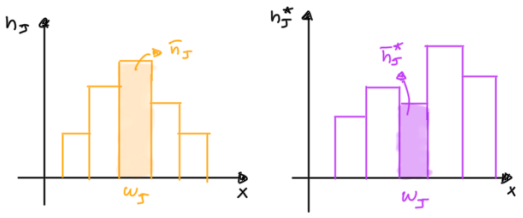
Più x_0 si allontana da $\overline{x}$ più è grande la varianza su y_0.

5.15 Fit d'istogrammi



Consideriamo un insieme di N misure $\{x_i\}_i^N$ IID che seguono la pdf $f(x_i, \vec{\theta}) : \Omega \subseteq \mathbb{R} \to \mathbb{R}$, raccogliamo i dati in modo da dividere l'asse reale in k intervalli $\omega_j = [x_i, x_{i+1})$. tali che $\omega_j \cap \omega_i = \varnothing \quad \forall i, j$ e $\bigcup_{j=1}^{k} \omega_j = \Omega$. Tali intervalli vengono defini 'bin', per ciascuno di essi contiamo il numero di misure n_j che cadono all'interno dell'intervallo. Complessivamente avremo che $\sum_{j=1}^{N} n_j = N$ numero totale di misure. Poichè le misure sono IID a ciascuna bin associamo una probabilità:

$$p_j = \int_{\omega_j} pdf(x, \vec{\theta})dx \quad \forall j = 1, ...k$$



Partendo dalla probabilità associata a ciascun bin ω_j costruiamo un istogramma i cui conteggi attesi sono dati da i $E[n_j] = N \cdot p_j(\vec{\theta})$. Le misure di conteggio dell'istogramma definito dai dati del campione seguono una joint-pdf $pdf(n_1, ..., n_k, p_1, ...p_k)$ che definisce una distribuzione multinomiale. Per $p \to 0$ le misure di conteggio sono statisticamente indipendenti tra loro di conseguenza la multinomiale può essere riscritta come prodotto di distribuzioni di probabilità binomiali.

$$pdf(n_1, ..., n_k, p_1, ..., p_k) = \prod_{i=1}^{k} pdf(n_i, p_i)$$

Nel caso in cui il valore di aspettano del conteggio dei bin sia costante possiamo approssimare le singole pdf binomiali come delle Poissoniane dove $\lambda_i(\vec{\theta}) = N * p_i(\vec{\theta})$ rappresenta la frequenza media di eventi della distribuzione di Poisson di conseguenza la joint-pdf diventa :

$$\prod_{i=1}^{k} pdf(n_i, p_i) = \prod_{i=1}^{k} \frac{\lambda_i(\vec{\theta})^{n_i}}{n_i!} e^{-\lambda_i} = L(n_1, ..., n_k, \theta)$$

Che coincide con la extended likelihood.
Non solo si conosce la pdf associata ai dati ma anche il loro valore di aspettazione e varianza di conseguenza è possibile applicare il metodo di ML e dei MQ.

Osservazioni

Utilizzare le misure binnate ci permette di ridurre le dimensioni dei dati con cui si lavora, anche se si ha una perdita d'informazione questa viene bilanciata da una maggiore semplicità e dalla possibilità di usare differenti tecniche per la stima dei parametri, come per esempio quella di ML che per campioni molto grandi risulta essere da un punto di vista computazionale dispendiosa.

5.15.1 Binned Data - Minimi Quadrati

Applichiamo il metodo dei minimi quadrati confrontando i conteggi dell'istogramma dei dati con l'istogramma dei valori attesi. Dato un campione $\{x_i\}_i^N$ IID dopo averli rappresentati in un istogramma, associamo al valore atteso di conteggi per ciascun bin l'espressione:

$$\mu_i = E[n_j] = N \cdot \int_{\omega_j} pdf(x, \vec{\theta}) dx = N \cdot p_j(\vec{\theta})$$

Scrivendo il funzionale Q^2 come:

$$Q^2 = \sum_{i=1}^{N} \frac{(n_i - E[n_i])^2)}{V[n_i]} = \sum_{i=1}^{N} \frac{(n_i - Np_i(\vec{\theta}))^2)}{Np_i(\vec{\theta})}$$

Assumendo che i conteggi n_i seguano una pdf Poiss(n_i, μ_i) (la discussione che segue risulterebbe verificata ugualmente anche se le pdf fossero binomiali). Per $n_i \to \infty$ possiamo approssimare la poissoniana a una gaussiana con valore atteso e varianza pari a μ_i di conseguenza Q^2 dipendendo da variabili aleatorie distribuite secondo una gaussiana, ed essendo lui stesso una variabile casuale si ha che segue la distribuzioni di $\chi^2(k-s)$ dove s è il numero di parametri. Definiamo dunque il χ^2 nella forma di **Neyman**:

$$\chi^2_{Neyman} = \sum_{i=1}^{N} \frac{(n_i - \mu_i(\vec{\theta})^2)}{n_i} = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{O_i} \tag{5.28}$$

Dove per n_i grande si è approssima $V[n_i] \approx n_i$. La necessità di approssimare i dati a una Gaussiana e la varianza al numero di conteggi di un bin, è determinata dal fatto che così facendo il teorema di Gauss-Markov risulta soddisfatto poichè tra le ipotesi è necessario che i momenti non dipendano dai parametri. In conclusione di procede a minimizzare il χ^2_{Neyman} per determinare i parametri $\vec{\theta}$.

Procedura di fit di un istogramma

La procedura di fit di un istogramma con il metodo del χ^2 di Neyman prevede quindi di:

- verificare che sia rispetta l'ipotesi pdf(n_i) gaussiana;
- associare a n_i un errore $\sqrt{n_i}$
- calcolare per ogni valore del parametro θ_i i valori attesi μ_i dei conteggi in ciascun bin;
- minimizzare il χ^2_{Neyman} trovando il valore di θ per cui l'accordo valori misurati - valori attesi è il migliore;
- all'occorrenza effettuare un test del χ^2

5.15.2 Binned Data - Maximum Likelihood

Fare il fit di un'istogramma quando non vale l'approssimazione Gaussiana prevede necessariamente l'utilizzo della binned Maximum Likelihood. Analogamente a quanto discusso a inizio sezione si procede a passare discretizzare il modello. Dove la funzione di likelihood associata al campione è data da

$$L(n_1, ..., n_k, \vec{\theta}) = \prod_i^k Bin(n_i, p_i(\vec{\theta}, N)$$

e si procede a massimizzare la funzione di likelihood:

$$\frac{\partial L(\vec{n}, \vec{\theta})}{\partial \theta_j} = 0 \qquad \forall j = 1, ..., k$$

Ad un bin-size deve essere associata una probabilità $p_i(\vec{\theta})$ piccola. Non è necessario che i bin abbiano tutti la stessa dimensione, possiamo sceglierle in modo che $N \cdot p_i$ sia grande e quindi valga l'approssimazione Gaussiana.

L'operazione di binning, fa perdere informazione, questo si può riflettere sull'incertezza del parametro stimato , ma anche sulla capacità di verificare che i dati siano ben descritti dal modello. In questi casi è utili studiare la dipendenza del risultato dal binning.