Appunti di teoria della stima
Approccio classico

A. Garzelli, L. Capobianco

Teoria della stima per gli attuali sistemi di *signal processing*

Elaborazione e l’estrazione di informazione da un insieme di dati.

Esempi:

- Radar
- Sonar
- Riconoscimento vocale
- Analisi di immagini

Caratterizzazione del problema

Stimare il valore di uno o più parametri significativi da un *set* di dati.

Caratterizzazione statistica.

Matematicamente, noto un vettore di dati di N elementi dipendenti da un parametro sconosciuto θ, vogliamo determinare θ utilizzando i dati a disposizione, che in altri termini significa definire uno stimatore

$$\hat{\theta} = f(x[0], x[1], \ldots, x[N-1])$$

dove f indica una funzione e **x** è il vettore dei dati.

Caso di un sistema Radar

Determinare la posizione dell'oggetto in esame, a partire dal ritardo τ_0 al quale si riceve l'eco dell'oggetto, mediante lo studio dell'equazione

$$\tau_0 = \frac{2R}{c} \ .$$

Fenomeni aleatori: perdite di propagazione, disturbi, sorgenti di rumore, ritardi aggiuntivi dovuti alle distorsioni di canale.

Tramite la teoria della stima si valuta l'approssimazione τ_0 mediante un valore $\hat{\tau_0}$.

Caso di un sistema sonar

il parametro di interesse è ancora la posizione di un oggetto in esame, si stima il valore dell'angolo di vista β (definito da $\beta = \arccos(\frac{v\tau_0}{d})$) mediante il valore $\hat{\beta}$.

Per determinare uno stimatore che produca risultati affidabili, è necessario come primo passo effettuare una buona modellizzazione matematica dei dati in esame. Tale modellizzazione deve essere effettuata mediante una descrizione probabilistica, e più precisamente attraverso una densità di probabilità (PDF), ovvero:

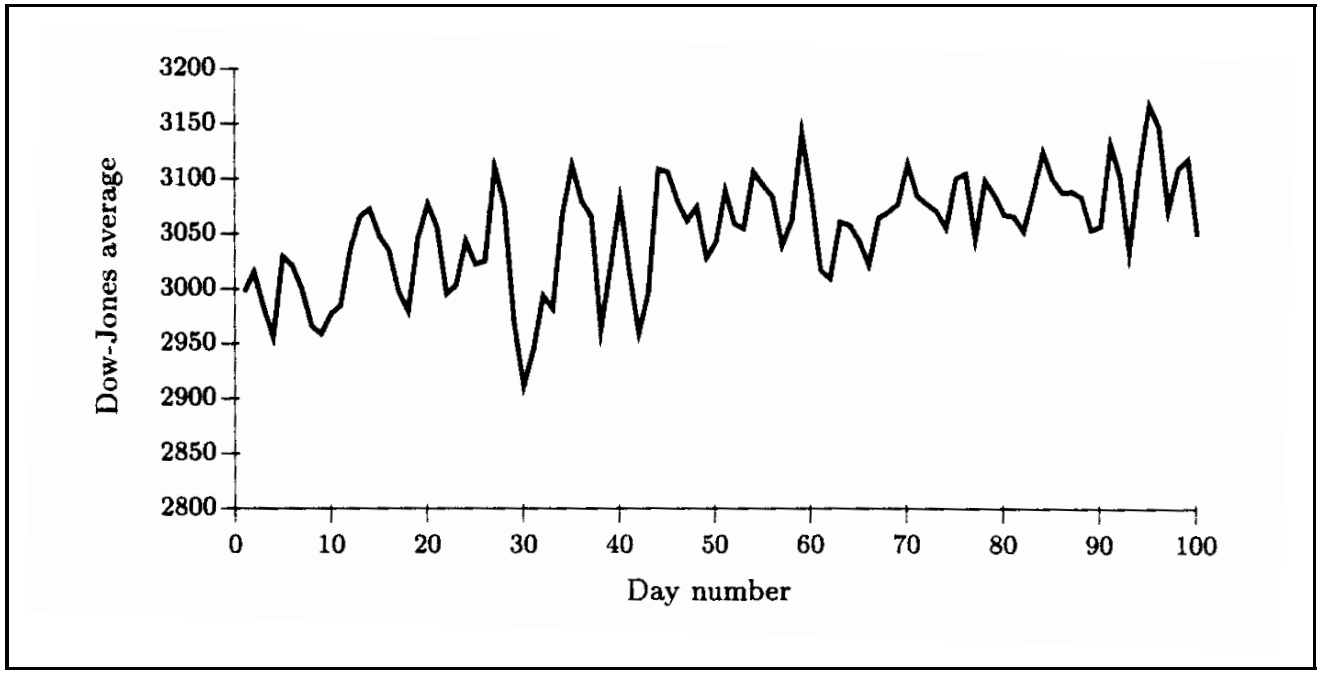
$$p(\mathbf{x}; \theta) = p(x[0], x[1], \ldots, x[N-1]; \theta) \ .$$

Approccio classico

La PDF viene parametrizzata tramite la variabile sconosciuta θ; ciò che si ottiene è una famiglia di PDF, ognuna per un dato valore di θ.

$$p(\mathbf{x}; \theta)$$

Esempio: andamento media indice *Dow-Jones*



Fluttuante, ma con aumento medio piuttosto costante, apparentemente lineare. Assumiamo

$$\mathbf{x}[n] = A + Bn + w[n] \quad n = 0, 1, \ldots, N-1 \, .$$

e rumore WGN: ogni campione di $w[n]$ segue una PDF Gaussiana, con media nulla e varianza σ^2, campioni $w[n]$ scorrelati

$\theta = [AB]^T$, $\mathbf{x} = [x[0], x[1], \ldots, x[N-1]]$, allora la PDF può essere espressa come

$$p(\mathbf{x}; \theta) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} exp[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A - Bn)^2] \, .$$

L’assunzione di rumore WGN

Modello matematico per espressione dello stimatore in forma chiusa

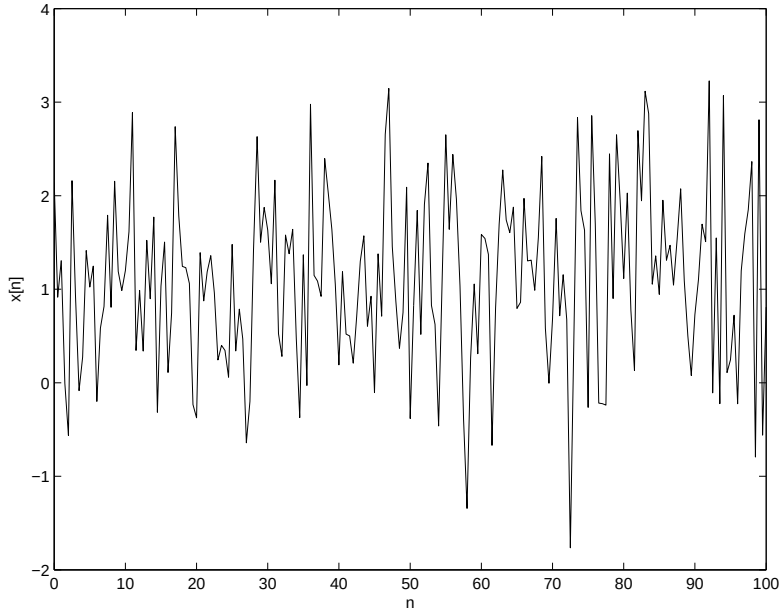
Affidabilità dello stimatore $\Longleftrightarrow$ scelta della PDF

Una volta specificata la PDF, il problema si sposta sulla possibilità di specificare uno stimatore ottimo per la curva dei dati: uno stimatore può infatti dipendere da altri parametri, a condizione però che questi siano noti.

stimatore: regola che assegna un valore a θ per ogni realizzazione di **x**

stima di θ: valore di θ ottenuto per una data realizzazione di **x**

Prestazioni di uno stimatore



$x[n]$ composto da un segnale in continua a cui si sovrappone un certo livello di rumore

$$x[n] = A + w[n]$$

dove $w[n]$ è un processo di rumore a media nulla. Basandoci su questa ipotesi, il nostro interesse è rivolto ad una stima di A. Intuitivamente, dato che A rappresenta il livello medio di $x[n]$, una scelta ragionevole per stimare A può essere:

$$\hat{A} = \frac{1}{N} \sum_{n=0}^{N-1} x[n] \quad .$$

- Con che precisione $\hat{A}$ approssima A?
- Esistono stimatori migliori di quello proposto?

Per i dati dell'esempio, si ha $\hat{A} = 1.1$, che è molto vicino al valore reale utilizzato $A = 1$.

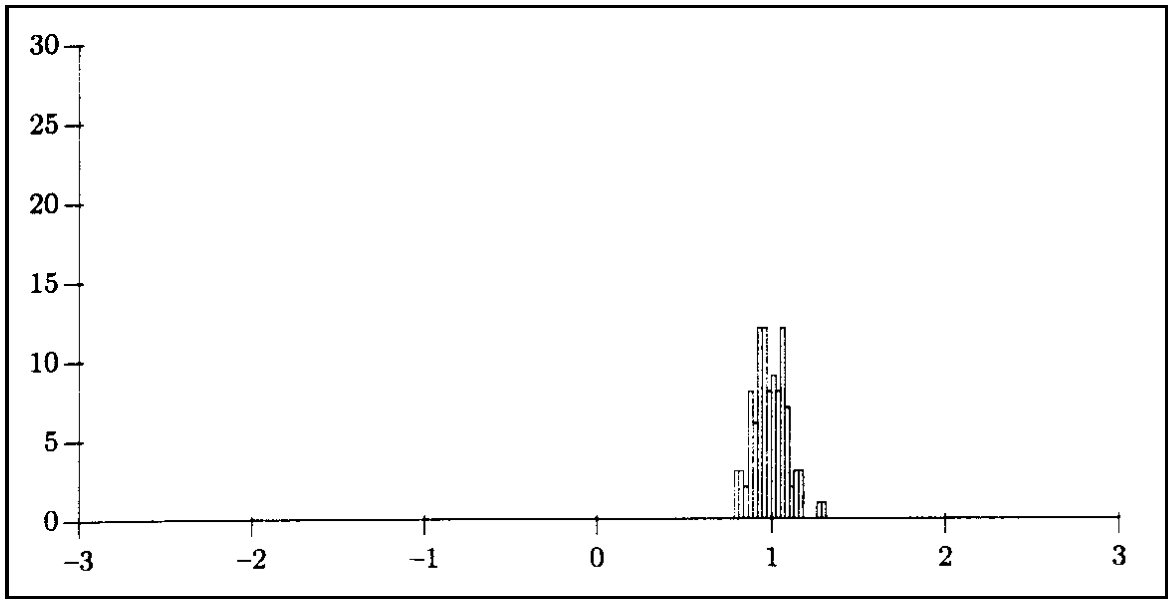
Altro stimatore per lo stesso problema $\bar{A} = x[0]$

- non affidabile
- non fa uso di tutti i dati e della informazione in essi contenuta
- non utilizza una operazione di media che riduca l'effetto del rumore

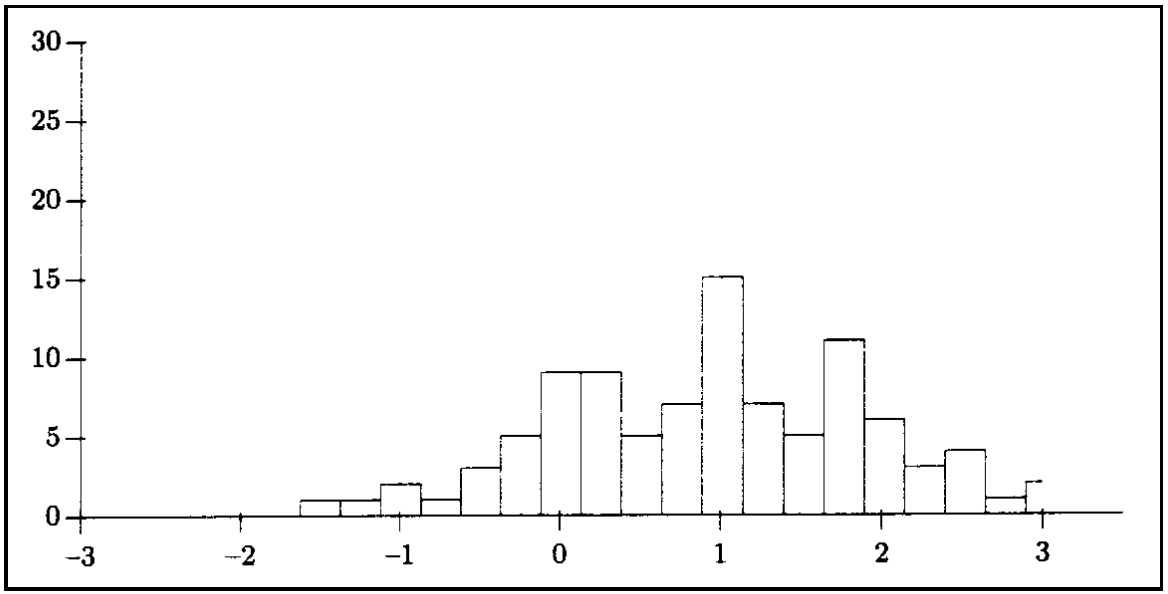
In questo caso, $\bar{A} = 2.1$, valore molto lontano dalla media reale. Anche se $\bar{A}$ avesse fornito un valore vicino al valore reale, non sarebbe stato comunque uno stimatore affidabile.

Immaginiamo di ripetere l'esperimento da cui sono stati tratti i dati, e di applicare i due stimatori proposti ai vari set di dati ottenuti. Supponiamo inoltre di fissare $A = 1$ e di sommare differenti realizzazioni di rumore $w[n]$; riportando su un istogramma i valori dei due stimatori proposti per ogni insieme di dati, possiamo ricostruire una approssimazione della PDF che descrive il numero di volte che uno stimatore ha prodotto un certo valore.

Istogramma di $\hat{A}$



Istogramma di $\bar{A}$



I grafici delle due figure riportano gli istogrammi relativi a 100 realizzazioni. E' evidente che $\hat{A}$ è uno stimatore migliore, perché i valori ottenuti sono più concentrati attorno al valore $A = 1$.

Valutazione statistica degli stimatori

Assumendo che il rumore abbia media nulla e varianza σ^2, si ha

$$E(\hat{A}) = E(\frac{1}{N}\sum_{n=0}^{N-1} x[n]) = \frac{1}{N}\sum_{n=0}^{N-1} E(x[n]) = A$$

$$E(\bar{A}) = E(x[0]) = A \quad ,$$

dunque in media gli stimatori producono il valore A.

La varianza degli stimatori è, essendo $w[n]$ scorrelato:

$$\text{var}(\hat{A}) = \text{var}(\frac{1}{N}\sum_{n=0}^{N-1} x[n]) = \frac{1}{N^2}\sum_{n=0}^{N-1} \text{var}(x[n]) = \frac{1}{N^2} N\sigma^2 = \frac{\sigma^2}{N}$$

$$\text{var}(\bar{A}) = \text{var}(x[0]) = \sigma^2 > \text{var}(\hat{A}) \ .$$

L'esempio sottolinea due aspetti importanti della teoria della stima che devono essere tenuti ben chiari in mente:

- Uno stimatore è una variabile casuale e pertanto può essere completamente descritto solo statisticamente o dalla sua PDF
- L'uso di una simulazione al computer può non essere completa: la stima delle prestazioni può al massimo essere valutata con un

certo grado di accuratezza o, nel peggiore dei casi, si possono ottenere risultati sbagliati, dovuti ad un insufficiente numero di esperimenti o ad una cattiva caratterizzazione del problema.

Abbiamo anticipato che dalla scelta della PDF dipende l'accuratezza e l'affidabilità dello stimatore: vogliamo sottolineare inoltre che tale scelta del modello deve ovviamente essere guidata dall'analisi di alcune caratteristiche fondamentali che la PDF deve presentare:

- compatibile con i vincoli del problema
- compatibile con la conoscenza a priori dei dati
- matematicamente maneggevole.

Approccio classico / Approccio bayesiano

Approccio *classico* $p(\mathbf{x};\theta)$

Parametro di interesse assunto sconosciuto, ma deterministico

Approccio *Bayesiano*

Parametro da stimare inteso come una realizzazione di una variabile casuale, le cui informazioni a priori sono contemplate in $p(\theta)$:

$$p(\mathbf{x},\theta) = p(\mathbf{x}|\theta)p(\theta)$$

$p(\theta)$ è la PDF che riassume le informazioni riguardo θ, note prima dell'osservazione dei dati, mentre la $p(\mathbf{x}|\theta)$ è la PDF che riassume la conoscenza a posteriori, offerta dall'osservazione dei dati condizionati da θ.

Approccio Bayesiano nell'esempio relativo all'indice Dow Jones.

Osservazione dei dati $\Rightarrow$ media intorno al valore 3000.

Inutile cercare un valore di A inferiore a 2000 o superiore a 4000.
Possiamo allora restringere la ricerca all'intervallo $[2800, 3200]$,
introducendo questa informazione nota nella PDF di A: ovvero
possiamo assumere che A non sia deterministica, ma una variabile
con una sua PDF, possibilmente uniforme nell'intervallo $[2800, 3200]$.

Analisi degli stimatori che, in media, producono il valore vero del parametro in esame

$\Rightarrow$ *stimatori non polarizzati*

Fra questi, cercheremo quello a minore variabilità. $\Rightarrow$ *MVU, minimum variance unbiased* estimator

Non sempre esiste lo stimatore non polarizzato a minima varianza

Se esiste, per trovarlo

- metodo *lower bound Cramer-Rao*,

Se non esiste, o in caso di fallimento del metodo, si possono utilizzare dei vincoli lineari per ottenere una semplice implementazione subottima dello stimatore.

Stimatore non polarizzato

Uno stimatore è non polarizzato se:

$$E(\hat{\theta}) = \theta \quad \forall\, \theta \in [a, b].$$

Esempio - Stimatore non polarizzato per livello di corrente continua in rumore WGN

Si consideri l'osservazione:

$$x[n] = A + w[n] \;\; n = 0, 1, ..., N-1$$

- A parametro da stimare
- $w[n]$ vettore WGN

Il valore di A può essere compreso nell'intervallo $-\infty < A < \infty$.

Dunque, una scelta ragionevole per la valutazione del valore medio è

$$\hat{A} = \frac{1}{N} \sum_{n=0}^{N-1} x[n] \; .$$

Per le proprietà di linearità dell'operatore $E[\cdot]$, abbiamo

$$E(\hat{A}) = E\left[\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right] = \frac{1}{N}\sum_{n=0}^{N-1} E(x[n]) = A$$

per qualsiasi A

Consideriamo invece, per la stessa osservazione, lo stimatore

$$\bar{A} = \frac{1}{2N} \sum_{n=0}^{N-1} x[n] \, .$$

Applicando l'operatore di media, si ha

$$\begin{aligned} E(\bar{A}) &= E[\frac{1}{2}A] \\ &= A \;\; se \;\; A = 0 \\ &\neq A \;\; se \;\; A \neq 0 \end{aligned}$$

Lo stimatore modificato funziona solo se $A = 0$, pertanto esso è polarizzato.

Combinazione di stimatori

$\{\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_n\}$

n stimatori per uno stesso parametro θ

Sia

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_i$$

$\hat{\theta}_i$ non polarizzati, con stessa varianza, e mutuamente scorrelati $\Rightarrow$

$$E(\hat{\theta}) = \theta$$

$$\text{var}(\hat{\theta}) = \frac{1}{n^2} \sum_{i=1}^{n} \text{var}(\hat{\theta}_i) = \frac{\text{var}(\hat{\theta}_i)}{n}$$

Se n cresce, la varianza dello stimatore così ottenuto decresce e passando al limite, se $n \to \infty$, $\hat{\theta} \to \theta$, con varianza nulla

Tuttavia, se gli stimatori sono polarizzati, ovvero $E(\hat{\theta}_i) = \theta + b(\theta)$, dove $b(\theta)$ è il valore di polarizzazione (*bias*) dello stimatore allora

$$\begin{aligned} E(\hat{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} E(\hat{\theta}_i) \\ &= \theta + b(\theta) \end{aligned}$$

$\hat{\theta}$ non converge al valore vero, indipendentemente dal numero di stimatori utilizzati nella media, cioè

$$\lim_{n\to\infty} \hat{\theta} \neq \theta$$

Prestazioni: criterio a minima varianza

Nella ricerca di uno stimatore dobbiamo ovviamente stabilire un criterio di ottimalità per valutarne le prestazioni. Possiamo utilizzare l'*errore quadratico medio* (MSE), definito come

$$mse(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] \ .$$

Questo indice misura la deviazione quadratica media dello stimatore dal valore vero. In generale l'adozione di questo criterio porta a stimatori che non possono essere scritti esclusivamente in funzione dei dati.

Riscriviamo l'espressione dell'mse come

$$\begin{aligned} mse(\hat{\theta}) &= E\{[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2\} \\ &= \text{var}(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2 \\ &= \text{var}(\hat{\theta}) + b^2(\theta) \end{aligned}$$

da cui si evidenzia come l'MSE contenga un errore dovuto alla varianza degli stimatori così come al valore di polarizzazione (*bias*).

Esempio. Consideriamo, nel problema *DC level in WGN*, lo stimatore modificato

$$\bar{A} = a\frac{1}{N}\sum_{n=0}^{N-1} x[n]$$

dove cercheremo di determinare la costante a in modo che risulti minimo l'mse.

Siccome $E(\bar{A}) = aA$ e var$(\bar{A}) = a^2\sigma^2/N$, abbiamo,

$$mse(\bar{A}) = \frac{a^2\sigma^2}{N} + (a-1)^2A^2$$

Derivando rispetto ad a otteniamo,

$$\frac{mse(\bar{A})}{da} = 2\frac{a\sigma^2}{N} + 2(a-1)A^2$$

che uguagliata a zero fornisce un valore di minimo per

$$a_{opt} = \frac{A^2}{A^2 + \sigma^2/N} \quad .$$

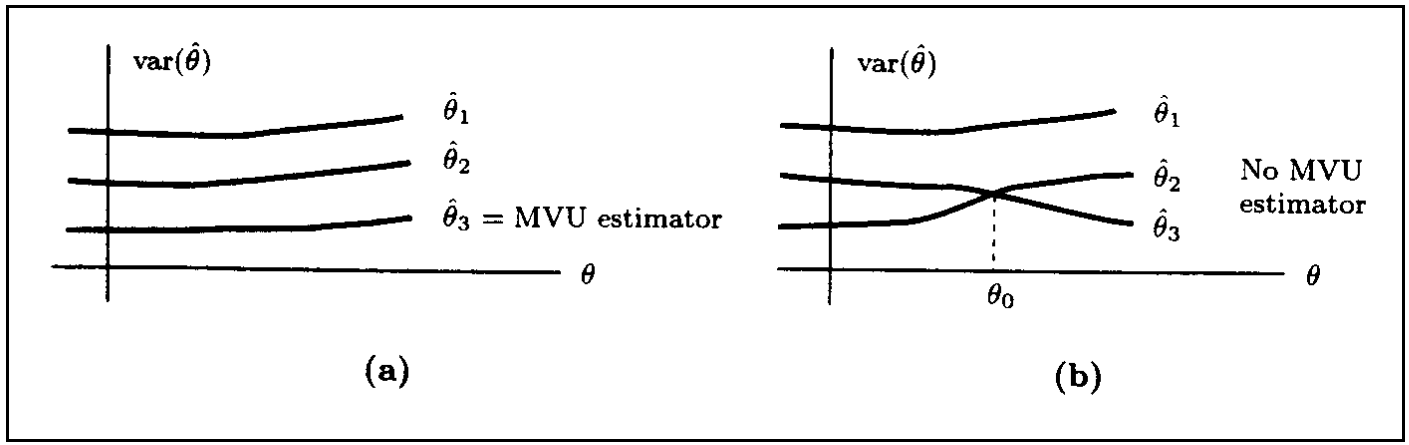
Per valori finiti di N, il valore ottimo di a dipende dal valore sconosciuto di A, pertanto lo stimatore MVU non è realizzabile.

Esistenza dello stimatore MVU

Criterio per valutare se esiste uno stimatore MVU per ogni θ appartenente all'intervallo di esame.

In generale l'MVU non esiste

Esempio



Nel caso a, stimatore MVU individuato dalla curva θ_3

Nel caso b, non si può individuare uno stimatore ottimo, in quanto per valori di θ inferiori a θ_0, θ_2 è migliore, mentre per valori $\theta > \theta_0$, lo stimatore θ_3 risulta migliore secondo il criterio MVU.

Per situazioni simili al caso a, θ_3 viene chiamato *stimatore non polarizzato a varianza uniforme minima*.

Esempio. Inesistenza dell'MVU

Se la funzione densità di probabilità (PDF) cambia al variare di θ, possiamo aspettarci che con θ possa cambiare anche lo stimatore da utilizzare.

Date due successive osservazioni, $x[0]$ e $x[1]$, con PDF

$$
\begin{aligned}
x[0] &\sim \mathcal{N}(\theta, 1) \\
x[1] &\sim \begin{cases} \mathcal{N}(\theta, 1) & se \quad \theta \geq 0 \\ \mathcal{N}(\theta, 2) & se \quad \theta < 0 \end{cases}
\end{aligned}
$$

dove $\mathcal{N}$ indica una PDF normale.

Si può mostrare che gli stimatori

$$\begin{aligned}\hat{\theta}_1 &= \frac{1}{2}(x[0] + x[1]) \\ \hat{\theta}_2 &= \frac{2}{3}x[0] + \frac{1}{3}x[1]\end{aligned}$$

non sono polarizzati. La varianza risulta essere rispettivamente

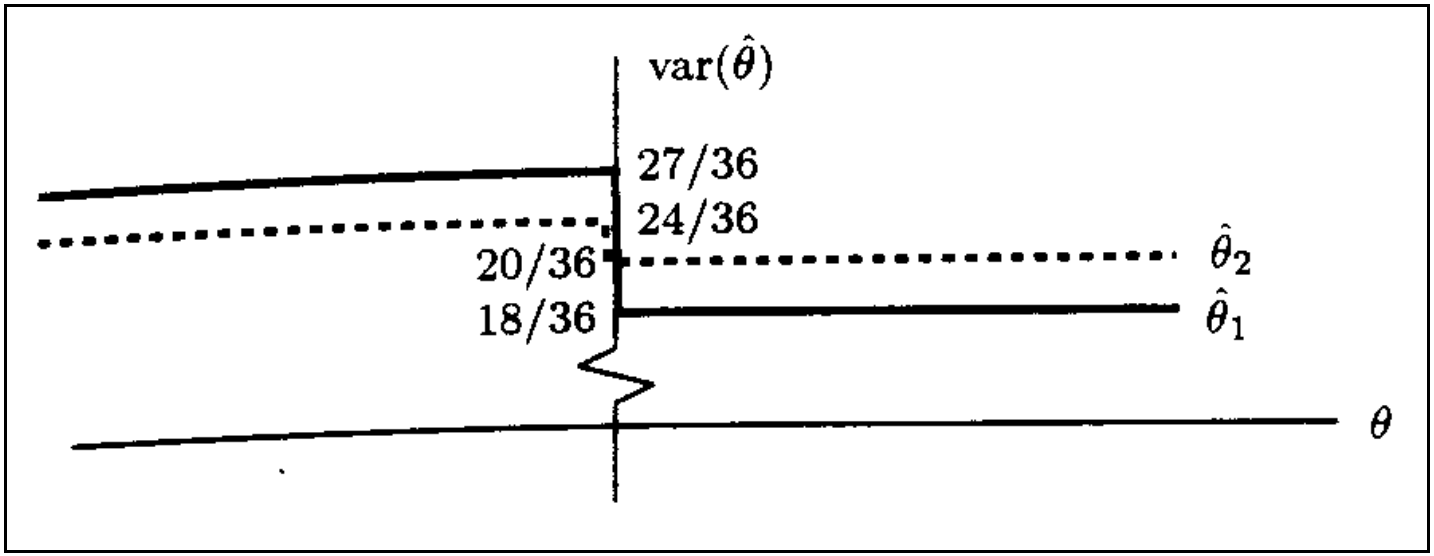
$$\begin{aligned}\mathsf{var}(\hat{\theta}_1) &= \frac{1}{4}(\mathsf{var}(x[0]) + \mathsf{var}(x[1])) \\ \mathsf{var}(\hat{\theta}_2) &= \frac{4}{9}\mathsf{var}(x[0]) + \frac{1}{9}\mathsf{var}(x[1])\end{aligned}$$

per cui,

$$\text{var}(\hat{\theta}_1) = \begin{cases} \frac{18}{36} = \frac{1}{2} & se \quad \theta \geq 0 \\ \frac{27}{36} = \frac{3}{4} & se \quad \theta < 0 \end{cases}$$

$$\text{var}(\hat{\theta}_2) = \begin{cases} \frac{20}{36} = \frac{5}{9} & se \quad \theta \geq 0 \\ \frac{24}{36} = \frac{2}{3} & se \quad \theta < 0 \end{cases}$$

La figura seguente mostra il risultato.



Si può dimostrare che in generale per $\theta \geq 0$ il valore minimo possibile per la varianza di un stimatore non polarizzato è $18/36$, mentre per $\theta < 0$ è $24/36$.

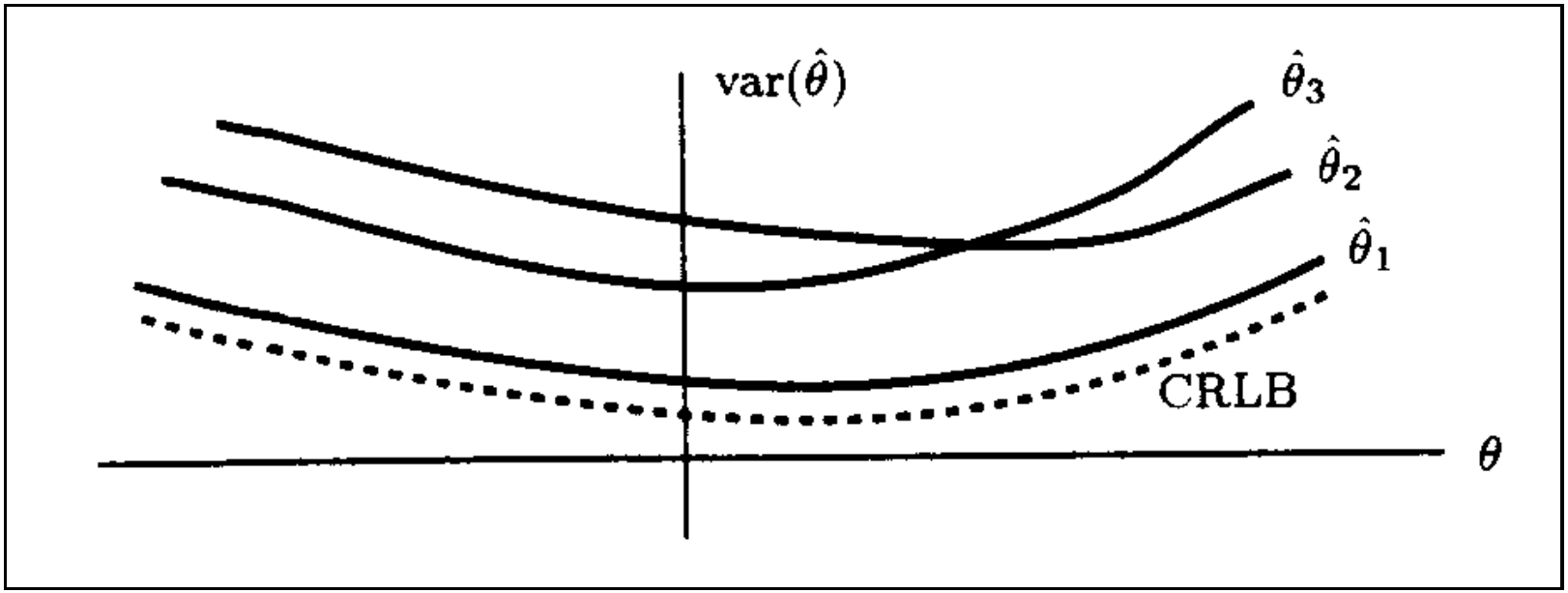
Ricerca dell’MVU

Non sempre esiste lo stimatore MVU.

Anche qualora esista, non esiste un metodo che dia la certezza di poterlo trovare. Abbiamo tre possibilità:

- Determinare il *Cramer-Rao lower bound*) (CRLB), e verificare l’esistenza di uno stimatore che lo soddisfi;
- Applicare il teorema Rao-Blackwell-Lehmann-Scheffe, (RBLS);
- Ricavare una classe di stimatori lineari e non-polarizzati dallo stimatore iniziale, e successivamente trovare in questo insieme l’MVU.

Ovviamente, mentre il primo ed il secondo metodo possono produrre lo stimatore MVU, il terzo lo produrrà solo se questo è lineare nei dati.



Il metodo CRLB fornisce un limite inferiore.

Dato un insieme di dati, non è possibile trovare uno stimatore non polarizzato con varianza strettamente minore di un valore che rappresenta il limite inferiore per quel dato problema di stima.

Se troviamo uno stimatore con varianza uguale al valore dato dal metodo CRLB $\Longrightarrow$ è l'MVU che stiamo cercando.

Cramer-Rao Lower Bound

La possibilità di determinare un limite inferiore (*lower bound*) per la varianza di un qualsiasi stimatore non polarizzato risulta essere estremamente utile nella pratica.

Analisi delle performance dello stimatore in esame da un confronto diretto fra i valori teorici e i valori ottenuti

Nel migliore dei casi, questo metodo ci permette di identificare lo stimatore MVU.

Considerazioni sull’accuratezza di uno stimatore

Stima basata su dati osservati e PDF che li caratterizza

L’accuratezza della stima dipende direttamente dalla PDF

Se la PDF dipende solo debolmente dal parametro che stiamo cercando di stimare o addirittura non dipende affatto da esso, non possiamo certamente aspettarci una corretta stima di tale parametro

In generale, quanto più la PDF è influenzata dal parametro sconosciuto, maggiore è la precisione con cui possiamo effettuare la stima

PDF vista come una funzione del parametro sconosciuto, con **x** fissato

$\Rightarrow$ viene detta *funzione di verosimiglianza* (likelihood function)

Intuitivamente, la precisione con cui possiamo stimare il parametro dipende dalla presenza di variazioni brusche della funzione di verosimiglianza: maggiore è l'irregolarità della funzione di verosimiglianza (intuitivamente: *grandi valori di curvatura*), maggiormente sarà accurata la stima.

L’irregolarità della funzione può essere misurata dal valore massimo, cambiato di segno, della derivata seconda del logaritmo della funzione di verosimiglianza, ovvero dalla *curvatura* ***media*** della *log-likelihood function*. La misura della curvatura può essere fornita dalla funzione

$$-E\left[\frac{\partial^2 \ln p(\mathbf{x}; A)}{\partial A^2}\right]$$

che è appunto una misura della curvatura media della *log-likelihood function*

Maggiore risulta la quantità espressa nell’espressione precedente, minore risulterà la varianza dello stimatore.

Cramer Rao Lower Bound - parametro scalare

Hp) La PDF $p(\mathbf{x};\theta)$ soddisfa la condizione di regolarità,

$$E\left[\frac{\partial \ln p(\mathbf{x};\theta)}{\partial\theta}\right] = 0 \quad \forall\theta.$$

$\Longrightarrow$ La varianza di un qualsiasi stimatore $\hat{\theta}$ non polarizzato deve soddisfare la disequazione

$$\text{var}(\hat{\theta}) \geq \frac{1}{-E\left[\frac{\partial^2 \ln p(\mathbf{X};\theta)}{\partial\theta^2}\right]} \quad \forall\theta.$$

dove la derivata è valutata sul valore vero di θ, e il valore atteso su $p(\mathbf{x};\theta)$.

Inoltre è possibile trovare uno stimatore non polarizzato per cui valga l'uguaglianza per ogni θ, **se** e **solo se** la derivata rispetto a θ della log-likelihood function (ovvero la *score function*) soddisfa l'uguaglianza

$$\left[\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta}\right] = I(\theta)(g(\mathbf{x}) - \theta)$$

per qualche coppia di funzioni g e I. Tale stimatore, che risulta essere proprio l'MVU, è $\hat{\theta} = g(\mathbf{x})$ e la funzione che esprime il valore di varianza minima è $1/I(\theta)$.

L’espressione esplicita del valore atteso è data da:

$$E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta^2}\right] = \int \frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta^2} p(\mathbf{x};\theta) d\mathbf{x}$$

in quanto la derivata seconda è una variabile casuale dipendente da **x**. Esso rappresenta, come già anticipato, il valor medio della curvatura della *log-likelihood function*.

Esempio - Stimatore non polarizzato per livello di corrente continua in rumore WGN

Riprendiamo l'esempio

$$x[n] = A + w[n] \;\; n = 0, 1, ..., N-1$$

dove $w[n]$ rappresenta un rumore WGN con varianza σ^2. Per determinare il CRLB per la costante A, abbiamo, essendo i campioni scorrelati,

$$\begin{aligned} p(\mathbf{x}; A) &= \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x[n] - A)^2\right] \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - A)^2\right] \end{aligned}$$

Effettuando la derivata prima

$$\begin{aligned}
\frac{\partial \ln p(\mathbf{x}; A)}{\partial A} &= \frac{\partial}{\partial A}\left[-\ln\left[(2\pi\sigma^2)^{\frac{N}{2}}\right] - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right] \\
&= \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A) \\
&= \frac{N}{\sigma^2}(\bar{x}-A)
\end{aligned}$$

dove $\bar{x}$ rappresenta il valor medio.

Effettuando nuovamente la derivata

$$\frac{\partial^2 \ln p(\mathbf{x}; A)}{\partial A^2} = -\frac{N}{\sigma^2}$$

e notando che la derivata seconda è costante, si ottiene il CRLB

$$\text{var}(\hat{A}) \geq \frac{\sigma^2}{N}$$

Dimostriamo adesso che quando viene assunto il valore CRLB, si ha

$$\text{var}(\hat{\theta}) = \frac{1}{I(\theta)}$$

dove

$$I(\theta) = -E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right].$$

Dal teorema CRLB si ottiene

$$\text{var}(\hat{\theta}) = \frac{1}{-E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right]}$$

$$\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta} = I(\theta)(\hat{\theta} - \theta)$$

$$\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2} = \frac{\partial I(\theta)}{\partial \theta}(\hat{\theta} - \theta) - I(\theta)$$

per cui il valore atteso cambiato di segno diventa

$$\begin{aligned} -E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right] &= -\frac{\partial I(\theta)}{\partial \theta}\left[E(\hat{\theta}) - \theta\right] + I(\theta) \\ &= I(\theta) \end{aligned}$$

ed infine

$$\mathrm{var}(\hat{\theta}) = \frac{1}{I(\theta)}.$$

Il CRLB non è sempre soddisfatto

Uno stimatore che contemporaneamente sia non polarizzato e rispetti il criterio CRLB, viene denominato *efficiente*, in quanto utilizza efficientemente i dati, sfruttando gran parte dell'informazione in essi contenuta.

Intuitivamente, maggiore è l'informazione, minore sarà il valore del *bound* in quanto minore è l'incertezza sulla misura. *Si noti che non è detto che uno stimatore MVU sia anche efficiente, in quanto pur essendo lo stimatore a varianza minima, può non rispettare il criterio CRLB.*

La quantità $I(\theta)$ è detta *Informazione di Fisher*, ed è una quantità non-negativa in quanto può essere dimostrata l'equivalenza

$$-E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta^2}\right] = E\left[\left(\frac{\partial \ln p(\mathbf{x};\theta)}{\partial\theta}\right)^2\right]$$

Un'altra importante caratteristica della $I(\theta)$ è la proprietà di additività per osservazioni indipendenti. Questa proprietà porta direttamente alla conclusione che il CRLB per N osservazioni IID (Indipendenti e Identicamente distribuite) è pari a $1/N$ volte il CRLB per una singola osservazione. Ovvero si ha:

$$I(\theta) = Ni(\theta)$$

dove

$$i(\theta) = -E\left[\frac{\partial^2 \ln p(x[n];\theta)}{\partial\theta^2}\right]$$

è la *Informazione di Fisher* per un campione. Per osservazioni non indipendenti, ci aspettiamo che l'informazione sia minore della quantità $Ni(\theta)$; per osservazioni completamente dipendenti, in cui per esempio $x[0] = x[1] = x[2] = ... = x[N-1]$, avremo $I(\theta) = i(\theta)$.

CRLB per un segnale immerso in rumore bianco Gaussiano

Osservazione di un segnale deterministico, immerso in un segnale WGN e dipendente da un parametro θ

$$x[n] = s[n, \theta] + w[n] \quad n = 0, 1, ..., N-1$$

La funzione di verosimiglianza è

$$p(\mathbf{x}; \theta) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} \left(x[n] - s[n; \theta]\right)^2\right].$$

Effettuando la derivata prima della *log-likelihood function*, si ha:

$$\frac{\partial \ln p(\mathbf{x}; \theta)}{\partial \theta} = \frac{1}{\sigma^2} \sum_{n=0}^{N-1} \left(x[n] - s[n; \theta]\right) \frac{\partial s[n; \theta]}{\partial \theta},$$

mentre effettuando la derivata seconda

$$\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left[(x[n]-s[n;\theta])\frac{\partial^2 s[n;\theta]}{\partial \theta^2} - \left(\frac{\partial s[n;\theta]}{\partial \theta}\right)^2\right].$$

L'operatore di media produce

$$E\left(\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right) = -\frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left(\frac{\partial s[n;\theta]}{\partial \theta}\right)^2$$

per cui in definitiva abbiamo:

$$\text{var}(\hat{\theta}) \geq \frac{\sigma^2}{\sum_{n=0}^{N-1}\left(\frac{\partial s[n;\theta]}{\partial \theta}\right)^2}.$$

L'espressione del *bound* dimostra la forte dipendenza del segnale da θ. Tale metodo produce stimatori molto accurati nel caso dell'analisi di segnali che variano rapidamente al variare di θ (in quanto c'è una forte dipendenza dalla derivata seconda). I casi in cui si abbia

$$\begin{aligned} s[n;A] &= A \\ s[n;\varphi] &= A\cos(2\pi f_0 n + \varphi) \end{aligned}$$

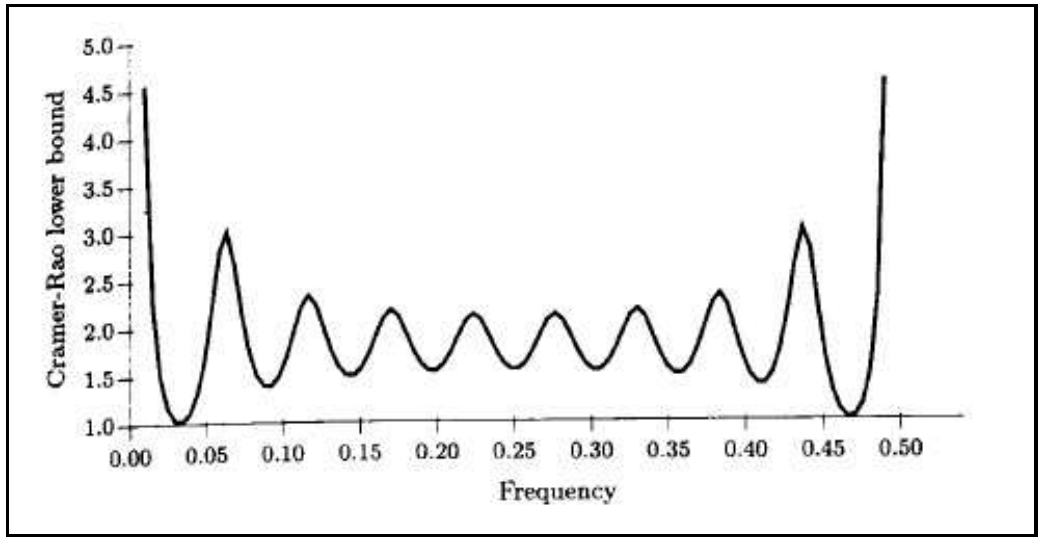
risultano essere dei casi particolari del caso appena studiato.

Esempio - Stima della frequenza di un segnale sinusoidale

$$s[n; f_0] = A\cos(2\pi f_0 n + \phi) \quad 0 < f_0 < \frac{1}{2}$$

con A e ϕ note. Si ottiene

$$\text{var}(\hat{f}_0) \geq \frac{\sigma^2}{A^2 \sum_{n=0}^{N-1} \left[2\pi n \sin(2\pi f_0 n + \phi)\right]^2}.$$



Andamento del CRLB in funzione della frequenza, con un SNR $A^2/\sigma^2 = 1$, lunghezza del record dati $N = 10$ e fase nulla. Si noti che se $f_0 \to 0$, il CRLB tende ad infinito, in quanto per valori di f_0 prossimi a zero, piccole variazioni di frequenza non alterano significativamente il segnale.

Trasformazione di Parametri

Nella pratica accade frequentemente che il parametro che interessa stimare sia una funzione di altri parametri fondamentali. Ad esempio, piuttosto che essere interessati alla stima dell'ampiezza A di una sinusoide o del suo segno, potremmo essere interessati alla potenza del segnale, legata al quadrato di A, A^2. Se conosciamo il CRLB per A, possiamo facilmente risalire al CRLB per A^2 o più in generale, per qualsiasi funzione di A.

Stima di $\alpha = g(\theta)$ invece che di θ.

Si può dimostrare che

$$\text{var}(\hat{\alpha}) \geq \frac{\left(\frac{\partial g}{\partial \theta}\right)^2}{-E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right]} \qquad (a)$$

Per l'esempio appena riportato, si avrebbe $\alpha = g(A) = A^2$ e

$$\text{var}(A^2) \geq \frac{(2A)^2}{N/\sigma^2} = \frac{4A^2\sigma^2}{N}.$$

Se $\alpha = g(\theta)$ è una funzione non lineare e θ è uno stimatore *efficiente*, allora $g(\theta)$ non è uno stimatore efficiente, mentre conserva questa caratteristica nel caso in sui $\alpha = g(\theta)$ sia una funzione lineare (o affine), come si può dimostrare dalla (a).

Tuttavia l'efficienza è mantenuta *approssimativamente* per trasformazioni non lineari con N molto grande. Infatti, quando N cresce, la PDF di $g(\theta)$ diventa più concentrata intorno alla media $\theta = E(\hat{\theta})$. Dunque possiamo linearizzare g intorno alla media con una buona approssimazione, tanto migliore quanto più stretta è la PDF, ovvero quanto più lungo è il record dati:

$$g(\hat{\theta}) \approx g(\theta) + \frac{dg(\theta)}{d\theta}(\hat{\theta} - \theta)$$

Questo porta alle equazioni approssimate

$$\begin{aligned} E\left[g(\hat{\theta})\right] &\approx g(\theta) \\ \text{var}\left[g(\hat{\theta})\right] &\approx \left[\frac{dg(\theta)}{d\theta}\right]^2 \text{var}(\hat{\theta}) \end{aligned}$$

Stima di più parametri - caso vettoriale

Vettore di parametri $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_p]^T$ nell'ipotesi di $\boldsymbol{\theta}$ non polarizzato

Si può dimostrare che il CRLB, che ci permette di assegnare un *bound* per la varianza dell'elemento i-esimo di $\boldsymbol{\theta}$, può essere trovato come l'elemento $[i, i]$ della matrice inversa

$$\text{var}(\boldsymbol{\theta}_i) \geq [I^{-1}(\underline{\theta})]_{ii} \ ,$$

dove $I(\boldsymbol{\theta})$ è la $p \times p$ *matrice dell'informazione di Fisher* $p \times p$

La matrice dell'informazione di Fisher è definita come

$$[\mathbf{I}(\boldsymbol{\theta})]_{ij} = -E\left[\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}\right]$$

per $i = 1, 2, ..., p$, $j = 1, 2, ..., p$. Si noti che per $p = 1$, si ritorna al caso scalare. Per il calcolo dell'espressione precedente viene utilizzato il valore vero di $\boldsymbol{\theta}$.

Esempio - DC level in WGN

Generalizziamo

$$x[n] = A + w[n] \;\; n = 0, 1, ..., N-1$$

supponendo che oltre ad A sia sconosciuta anche la varianza σ^2. Il vettore, con $p = 2$, diventa dunque $\boldsymbol{\theta} = [A \; \sigma^2]$. La matrice 2×2, di *Informazione di Fisher*, è

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} -E\left[\dfrac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A^2}\right] & -E\left[\dfrac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A \partial \sigma^2}\right] \\ -E\left[\dfrac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \sigma^2 \partial A}\right] & -E\left[\dfrac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial {\sigma^2}^2}\right] \end{bmatrix}$$

Simmetrica e semidefinita positiva.

La funzione di verosimiglianza è

$$\ln p(\mathbf{x};\boldsymbol{\theta}) = -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2$$

Effettuando le derivate e l'operazione di media, la matrice diventa

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix}$$

Nonostante non sia vero in generale, in questo caso la matrice è diagonale e può essere facilmente invertita, trovando

$$\text{var}(\hat{A}) \geq \frac{\sigma^2}{N}$$

$$\text{var}(\hat{\sigma^2}) \geq \frac{2\sigma^4}{N}$$

Notiamo che il CRLB per $\hat{A}$ è lo stesso rispetto al caso in cui A sia l'unico parametro da stimare, ma questo non può essere generalizzato, come vediamo nell'esempio successivo.

Esempio - Line fitting

Consideriamo il problema di *line fitting* (o regressione lineare) per una data osservazione

$$x[n] = A + Bn + w[n] \quad n = 0, 1, ..., N-1$$

in cui $w[n]$ è un rumore WGN; vogliamo determinare il CRLB per la coppia $\boldsymbol{\theta} = [A\ B]^T$. La matrice di Fisher è

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} -E\left[\dfrac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A^2}\right] & -E\left[\dfrac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A \partial B}\right] \\ -E\left[\dfrac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial B \partial A}\right] & -E\left[\dfrac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial B^2}\right] \end{bmatrix}.$$

La funzione di verosimiglianza è

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A-Bn)^2\right\}.$$

Effettuando le operazione di derivate e di media, si ottiene la matrice

$$\begin{aligned}\mathbf{I}(\boldsymbol{\theta}) &= \frac{1}{\sigma^2}\begin{bmatrix} N & \sum_{n=0}^{N-1} n \\ \sum_{n=0}^{N-1} n & \sum_{n=0}^{N-1} n^2 \end{bmatrix} \\ &= \frac{1}{\sigma^2}\begin{bmatrix} N & \frac{N(N-1)}{2} \\ \frac{N(N-1)}{2} & \frac{N(N-1)(2N-1)}{2} \end{bmatrix}. \end{aligned} \quad (1)$$

Invertendo la matrice,

$$\mathbf{I^{-1}}(\boldsymbol{\theta}) = \sigma^2 \begin{bmatrix} \frac{2(2N-1)}{N(N+1)} & -\frac{6}{(N(N+1))} \\ -\frac{6}{(N(N+1))} & \frac{12}{(N(N^2-1))} \end{bmatrix}$$

IL CRLB è allora

$$\text{var}(\hat{A}) \geq \frac{2(2N-1)\sigma^2}{N(N+1)}$$

$$\text{var}(\hat{B}) \geq \frac{12\sigma^2}{N(N^2-1)}.$$

Cramer Rao Lower Bound - Caso vettoriale

Si assuma che la PDF di $p(\mathbf{x};\boldsymbol{\theta})$ soddisfi la condizione di regolarità

$$E\left[\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \mathbf{0} \quad \text{per ogni } \boldsymbol{\theta}$$

dove la media è realizzata su $p(\mathbf{x};\boldsymbol{\theta})$ rispetto ad **x**. La matrice di covarianza di un qualsiasi stimatore $\hat{\boldsymbol{\theta}}$ soddisfa l'equazione

$$\boldsymbol{C}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{I}^{-1}(\boldsymbol{\theta}) \geq \mathbf{0}$$

dove la disequazione indica che la matrice è semidefinita positiva. La *Fisher information matrix* $\mathbf{I}(\boldsymbol{\theta})$ è data da

$$[\mathbf{I}(\boldsymbol{\theta})]_{ij} = -E\left[\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}\right]$$

dove le derivate sono effettuate rispetto al valore vero di $\boldsymbol{\theta}$ e l'operazione di media è realizzata rispetto a $p(\mathbf{x};\boldsymbol{\theta})$. Inoltre lo stimatore raggiunge il *bound* $\boldsymbol{C_{\bar{\theta}} = I^{-1}(\theta)}$ **se** e **solo se**

$$\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{I}(\boldsymbol{\theta})(\mathbf{g}(\mathbf{x}) - \boldsymbol{\theta})$$

per qualche funzione *p-dimensionale* $\mathbf{g}$, e qualche matrice $\mathbf{I}$ di dimensioni $p \times p$. Lo stimatore MVU così ottenuto è $\hat{\boldsymbol{\theta}} = \mathbf{g}(\mathbf{x})$, e la sua matrice di covarianza $\mathbf{I}^{-1}(\boldsymbol{\theta})$.

MVU per modelli lineari

Valutazione dello stimatore MVU agevolata da modelli

modelli lineari $\Longrightarrow$ immediato trovare lo stimatore una volta identificato il modello

costruzione del modello lineare $\Longrightarrow$ proprietà $\Longrightarrow$ soluzione

Teorema per identificare e caratterizzare la procedura di stima per un problema lineare

MVU per un Modello Lineare

Dati osservati modellati come

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

con $\mathbf{x}$ *vettore di osservazione* $N \times 1$

$\mathbf{H}$ *matrice di osservazione* $N \times p$ $(N > p)$ e rango p

$\boldsymbol{\theta}$ vettore di parametri $p \times 1$ da stimare

$\mathbf{w}$ vettore di rumore $N \times 1$ distribuito $\mathcal{N}(0, \sigma^2 I)$

$\Longrightarrow$ stimatore MVU $\quad \hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$

e matrice di covarianza di $\hat{\boldsymbol{\theta}}$ $\quad \mathcal{C}_{\hat{\boldsymbol{\theta}}} = \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}$

Inoltre per i modelli lineari lo stimatore MVU è *efficiente* in quanto la sua varianza soddisfa il CRLB.

Si noti che l'inversa di $(\mathbf{H}^T\mathbf{H})$ esiste in quanto il rango di $\mathbf{H}$ è p, le colonne sono linearmente indipendenti.

Inoltre non solo la media e la varianza, ma tutta la statistica di $\hat{\boldsymbol{\theta}}$ è completamente specificata, in quanto $\hat{\boldsymbol{\theta}}$ è una trasformazione lineare di un vettore Gaussiano $\mathbf{x}$ e dunque

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}(0, \sigma^2(\mathbf{H}^T\mathbf{H})^{-1})$$

Verifica

$$\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{I}(\boldsymbol{\theta})(\mathbf{g}(\mathbf{x}) - \boldsymbol{\theta})$$

$\hat{\boldsymbol{\theta}} = \mathbf{g}(\mathbf{x})$ è MVU con matrice di covarianza $\mathcal{C}_{\hat{\boldsymbol{\theta}}} = \mathbf{I}^{-1}(\boldsymbol{\theta})$

Nel caso in esame

$$\begin{aligned}\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{\partial}{\partial \boldsymbol{\theta}}\left[-\ln(2\pi\sigma^2)^{\frac{N}{2}} - \frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})^T(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})\right] = \\ &= -\frac{1}{2\sigma^2}\frac{\partial}{\partial \boldsymbol{\theta}}\left[\mathbf{x^T x} - 2\mathbf{x^T H}\boldsymbol{\theta} + \boldsymbol{\theta^T H^T H}\boldsymbol{\theta}\right] =\end{aligned}$$

$$
\begin{aligned}
\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{1}{\sigma^2}\left[\mathbf{H^T x} - \mathbf{H^T H}\boldsymbol{\theta}\right] \\
&= \frac{(\mathbf{H^T H})(\mathbf{H^T H})^{-1}}{\sigma^2}\left[\mathbf{H^T x} - \mathbf{H^T H}\boldsymbol{\theta}\right] = \\
&= \frac{(\mathbf{H^T H})}{\sigma^2}\left[(\mathbf{H^T H})^{-1}\mathbf{H^T x} - \boldsymbol{\theta}\right] = \\
&= \mathbf{I}(\boldsymbol{\theta})(\mathbf{g}(\mathbf{x}) - \boldsymbol{\theta})
\end{aligned}
$$

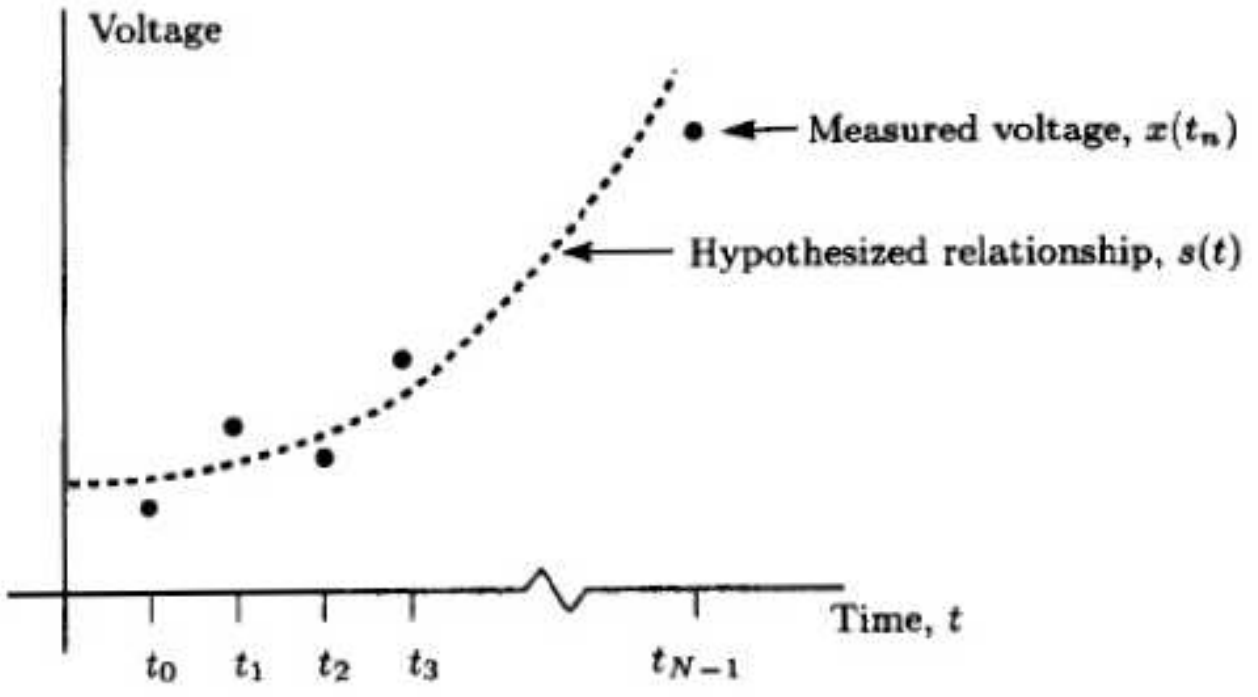
$\Rightarrow$

$$
\begin{aligned}
\hat{\boldsymbol{\theta}} &= \mathbf{g}(\mathbf{x}) = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} \\
\mathcal{C}_{\hat{\boldsymbol{\theta}}} &= \mathbf{I}^{-1}(\boldsymbol{\theta}) = \sigma^2(\mathbf{H^T H})^{-1}
\end{aligned}
$$

Esempio - Curve fitting

In molte situazioni sperimentali si cerca di determinare delle relazioni empiriche fra due o più variabili.

In figura sono visibili i risultati di un esperimento di misura di tensione agli istanti $t = t_0, t_1, t_2, ..., t_{N-1}$.



Dalla misura si nota un andamento quadratico in funzione del tempo.

Pertanto un'ipotesi ragionevole per il modello dei dati può essere

$$x(t_n) = \theta_1 + \theta_2 t_n + \theta_3 t_n^2 + w(t_n) \quad n = 0, 1, 2, ..., N-1.$$

Assumiamo che $w(t_n)$ siano variabili casuali gaussiane indipendenti e identicamente distribuite (iid), con media nulla e varianza σ^2 o che siano campioni di un rumore WGN. Il modello utilizzabile è

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

dove

$$\mathbf{x} = [x(t_0)x(t_1)...x(t_{N-1})]^T,$$

lo stimatore è

$$\hat{\boldsymbol{\theta}} = [\theta_1\theta_2\theta_3]^T$$

posto

$$\mathbf{H} = \begin{bmatrix} 1 & t_0 & t_0^2 \\ 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ \vdots & \vdots & \vdots \\ 1 & t_{N-1} & t_{N-1}^2 \end{bmatrix}.$$

Più in generale, se vogliamo adattare un polinomio di ordine $(p-1)$ ai dati sperimentali,

$$x(t_n) = \theta_1 + \theta_2 t_n + \theta_3 t_n^2 + \ldots + \theta_p t_n^{p-1} + w(t_n) \quad n = 0, 1, 2, ..., N-1,$$

la soluzione è ancora

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x}$$

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

dove

$$\mathbf{x} = [x(t_0)x(t_1)...x(t_{N-1})]^T,$$

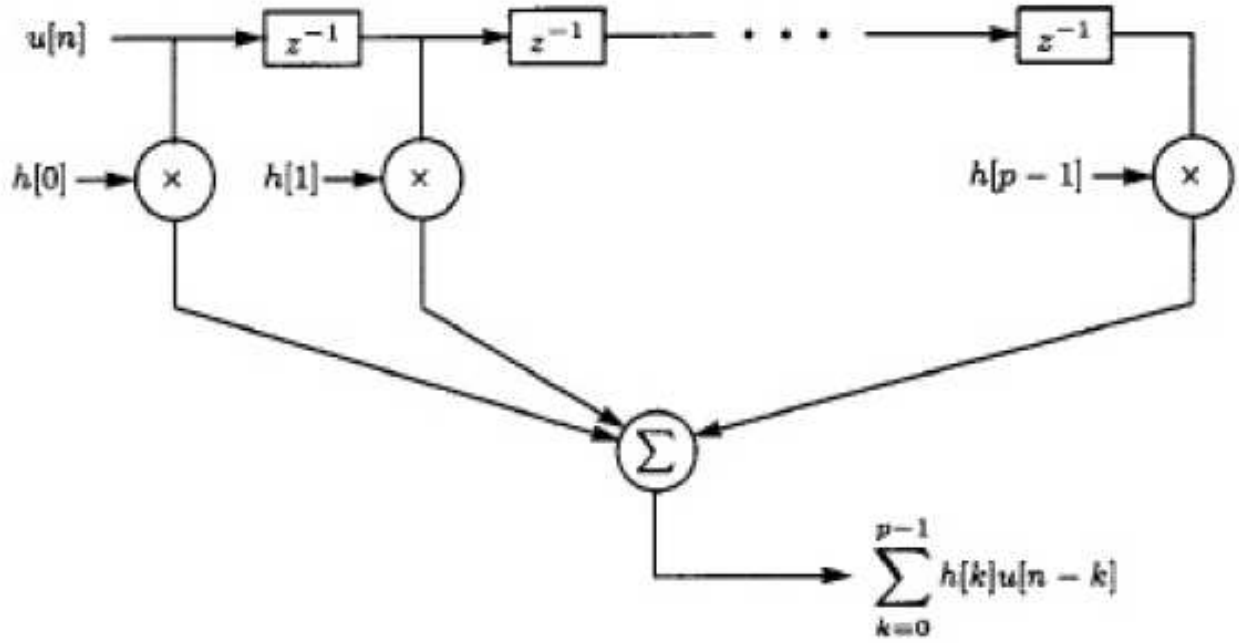
$$\mathbf{H} = \begin{bmatrix} 1 & t_0 & \cdots & t_0^{p-1} \\ 1 & t_1 & \cdots & t_1^{p-1} \\ 1 & t_2 & \cdots & t_2^{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_{N-1} & \cdots & t_{N-1}^{p-1} \end{bmatrix}.$$

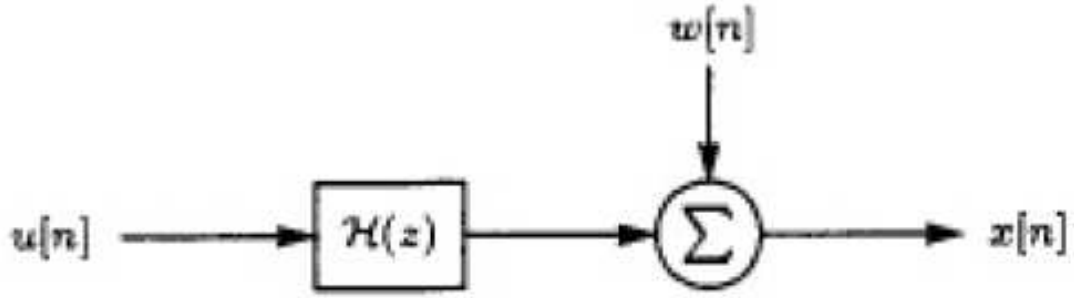
La matrice di osservazione in questo caso ha la forma di una *matrice di Vandermonde*, e la risultante curva è

$$\hat{s}(t) = \sum_{i=1}^{p} \hat{\theta}_i t^{i-1}.$$

Esempio - Sistema di identificazione

Identificare un modello di un sistema attraverso l’analisi dei dati in ingresso e uscita.





$$\mathcal{H}(z) = \sum_{k=0}^{p-1} h[k] z^{-k}$$

Osserviamo l'esempio in figura che implementa un filtro FIR: il sistema è pilotato da un ingresso $u[n]$, che serve per testare il sistema.

Idealmente, all'uscita la sequenza $\sum_{k=0}^{p-1} h[k]u[n-k]$ ci permette di stimare la risposta all'impulso del filtro. Nella pratica, tuttavia, l'uscita è corrotta dal rumore, pertanto assumere un modello che presenta un rumore AWGN risulta più corretto. Avremo

$$x[n] = \sum_{k=0}^{p-1} h[k]u[n-k] + w[n] \quad n = 0, 1, ..., N-1$$

dove si assume che $u[n] = 0$ per $n < 0$.

In forma matriciale avremo

$$\mathbf{x} = \underbrace{\begin{bmatrix} u[0] & 0 & \cdots & 0 \\ u[1] & u[0] & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u[N-1] & u[N-2] & \cdots & u[N-p] \end{bmatrix}}_{\mathbf{H}} \underbrace{\begin{bmatrix} h[0] \\ h[1] \\ \vdots \\ h[p-1] \end{bmatrix}}_{\boldsymbol{\theta}} + \mathbf{w}$$

Forma caratteristica di un modello lineare, per cui lo stimatore MVU per la risposta all'ingresso $u[n]$ è

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$$

e la matrice di covarianza di $\hat{\boldsymbol{\theta}}$ è

$$\mathcal{C}_{\hat{\boldsymbol{\theta}}} = \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}$$

Quale segnale pilota $u[n]$ devo scegliere per minimizzare la varianza dello stimatore?

La varianza di $\hat{\theta}_i$ è

$$\text{var}(\hat{\theta}_i) = e_i^T \mathcal{C}_{\hat{\theta}} e_i$$

dove $e_i = [00...010...00]^T$ ('1' occupa la i-esima posizione)

Partiamo dalla disuguaglianza di Cauchy-Schwartz

$$(\xi_1^T \xi_2)^2 \leq \xi_1^T \xi_1 \xi_2^T \xi_2$$

Poiché $\mathcal{C}_{\hat{\theta}}^{-1}$ può essere fattorizzato come $\mathcal{D}^T\mathcal{D}$ con $\mathcal{D}$ matrice invertibile $p \times p$ e imponendo $\xi_1 = \mathcal{D}e_i$ e $\xi_2 = \mathcal{D}^{T^{-1}}e_i$

$$(e_i^T\mathcal{D}^T\mathcal{D}^{T^{-1}}e_i)^2 = 1,$$

si ha

$$\begin{aligned} 1 &\leq (e_i^T\mathcal{D}^T\mathcal{D}e_i)(e_i^T\mathcal{D}^{-1}\mathcal{D}^{T^{-1}}e_i) \\ &= (e_i^T\mathcal{C}_{\hat{\theta}}^{-1}e_i)(e_i^T\mathcal{C}_{\hat{\theta}}e_i) \end{aligned}$$

e in definitiva

$$\text{var}(\hat{\theta}_i) \geq \frac{1}{(e_i^T\mathcal{C}_{\hat{\theta}}^{-1}e_i)} = \frac{\sigma^2}{[\mathbf{H}^T\mathbf{H}]_{ii}}.$$

Vale l'uguaglianza (e dunque l'MVU viene raggiunto) *se* e *solo se* $\xi_1 = c\xi_2$ per una data costante c, ovvero se

$$\mathcal{D}e_i = c_i \mathcal{D}^{T^{-1}} e_i$$

o, equivalentemente, se

$$\mathcal{D}^T \mathcal{D} e_i = c_i e_i \quad i = 1, 2, ..., p.$$

Siccome si ha

$$\mathcal{D}^T \mathcal{D} = \mathcal{C}_{\hat{\theta}}^{-1} = \frac{\mathbf{H}^T \mathbf{H}}{\sigma^2}$$

abbiamo

$$\frac{\mathbf{H}^T \mathbf{H}}{\sigma^2} e_i = c_i e_i.$$

Si ottiene, dalla combinazioni delle precedenti espressioni in forma matriciale, che la condizione per il raggiungimento dell'MVU è esprimibile in funzione della matrice $\mathbf{H}$ come

$$\mathbf{H}^T\mathbf{H} = \sigma^2 \begin{bmatrix} c_1 & 0 & \cdots & o \\ 0 & c_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c_p \end{bmatrix}.$$

$\Longrightarrow$ condizione sull'ingresso pilota $u[n]$ per minimizzare la varianza dello stimatore MVU. Siccome $[\mathbf{H}]_{ij} = u[i-j]$,

$$[\mathbf{H}^T\mathbf{H}] = \sum_{n=1}^{N} u[n-i]u[n-j] \;\; i = 1, 2, ..., p \; j = 1, 2, ..., p$$

e per N molto grande si ha

$$[\mathbf{H}^T\mathbf{H}] \approx \sum_{n=0}^{N-1-|i-j|} u[n]u[n+\mid i-j \mid]$$

in cui può essere riconosciuta una sequenza di autocorrelazione di una sequenza deterministica $u[n]$. Inoltre con questa approssimazione $\mathbf{H}^T\mathbf{H}$ diventa una matrice simmetrica di autocorrelazione di *Toeplitz*

$$\mathbf{H}^T\mathbf{H} = N \begin{bmatrix} r_{uu}[0] & r_{uu}[1] & \cdots & r_{uu}[p-1] \\ r_{uu}[1] & r_{uu}[0] & \cdots & r_{uu}[p-2] \\ \vdots & \vdots & \ddots & \vdots \\ r_{uu}[p-1] & r_{uu}[p-2] & \cdots & r_{uu}[0] \end{bmatrix}$$

con

$$r_{uu}[k] = \frac{1}{N} \sum_{n=0}^{N-1-k} u[n]u[n+k]$$

che può essere vista come l'autocorrelazione di $u[n]$. Perché $\mathbf{H}^T\mathbf{H}$ sia diagonale è necessario che

$$r_{uu}[k] = 0 \;\; k \neq 0,$$

che è una condizione approssimativamente vera se usiamo una sequenza PRN in ingresso.

$\Longrightarrow \mathbf{H}^T\mathbf{H} = N r_{uu}[0] I$, e dunque

$$\text{var}(\hat{h}[i]) = \frac{1}{N r_{uu}[0]/\sigma^2} \quad i = 0, 1, ..., p-1.$$

Scegliendo una sequenza PRN allora abbiamo ottenuto lo stimatore MVU come

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$$

in cui $\mathbf{H}^T\mathbf{H} = Nr_{uu}[0]I$. Si ottiene

$$\hat{h}[i] = \frac{1}{Nr_{uu}[0]} \sum_{n=0}^{N-1} u[n-i]u[n] \tag{2}$$

$$= \frac{\frac{1}{N} \sum_{n=0}^{N-1-i} u[n]x[n+i]}{r_{uu}[0]} \tag{3}$$

in quanto $u[n] = 0$ se $n < 0$. Il numeratore è la crosscorrelazione $r_{ux}[i]$ fra le sequenze di ingresso e uscita, per cui, se usiamo una

sequenza PRN per identificare il sistema, lo stimatore MVU (per alti valori di N) è

$$\hat{h}[i] = \frac{r_{ux}[i]}{r_{uu}[0]} \quad i = 0, 1, ..., p-1$$

dove

$$r_{ux}[i] = \frac{1}{N} \sum_{n=0}^{N-1-i} u[n]x[n+i]$$

e

$$r_{uu}[0] = \sum_{n=0}^{N-1} u^2[n]$$

Statistiche sufficienti

I modelli lineari rendono semplice la valutazione del CRLB e dunque dello stimatore MVU.

Se non esiste uno stimatore efficiente o non è possibile ricondursi a un modello lineare, come verificare l’esistenza di uno stimatore MVU?

Teorema Rao-Blackwell-Lehmann-Scheffe (concetto di statistica sufficiente)

Si può determinare lo stimatore MVU da un’ispezione della PDF

Per il problema della stima di un livello A di corrente DC immerso in rumore WGN abbiamo trovato lo stimatore MVU

$$\hat{A} = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$$

con varianza minima σ^2/N

Quali campioni portano la maggiore informazione al problema della stima?

Esiste un insieme o sottoinsieme di campioni *sufficiente* ai fini della stima?

Caso in esame, set di dati sufficienti:

$$\begin{aligned} S_1 &= \{x[0], x[1], ..., x[N-1]\} \\ S_2 &= \{x[0] + x[1], ..., x[N-1]\} \\ S_3 &= \left\{ \sum_{n=0}^{N-1} x[n] \right\}. \end{aligned}$$

Il set di dati che contiene il minor numero di campioni pur essendo sufficiente, viene chiamato *set minimo* o *minima statistica sufficiente*

S_3 è la *minima statistica sufficiente*

Per la stima di A infatti, una volta nota S_3, non abbiamo più bisogno di conoscere i valori dei singoli campioni, in quanto tutta l'informazione necessaria è contenuta nella loro somma.

Per estendere questo concetto, si consideri la PDF dei dati

$$\begin{aligned} p(\mathbf{x}; A) &= \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x[n]-A)^2\right] \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right] \end{aligned}$$

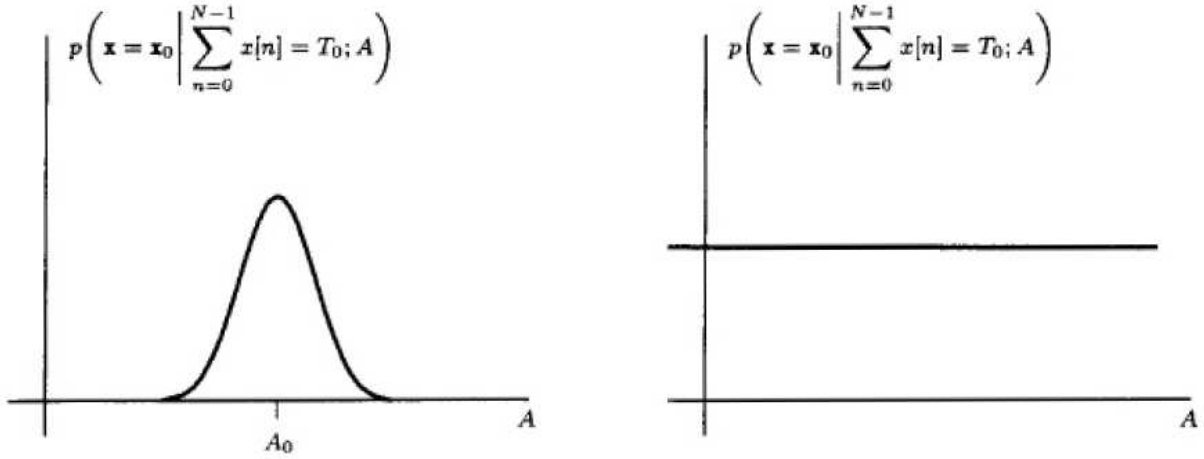
in cui si assume che $T(\mathbf{x}) = \sum_{n=0}^{N-1} x[n] = T_0$ sia stata osservata

PDF $\Longrightarrow$ PDF condizionata

$$p(\mathbf{x}|\sum_{n=0}^{N-1} x[n] = T_0; A)$$

Statistica sufficiente per la stima di A $\Longrightarrow$ PDF condizionata non dipendente dal valore di A.

Se fosse dipendente da A, potremmo dedurre dai dati una maggiore informazione su A, oltre a quella già fornita dalla statistica che riteniamo sufficiente, e ciò vorrebbe dire che la statistica non è affatto sufficiente



Determinare la Statistica Sufficiente

Teorema della fattorizzazione di Neymann-Fisher

$T(\mathbf{x})$è una statistica sufficiente per la variabile θ **se** e **solo se** è possibile fattorizzare la PDF $p(\mathbf{x};\theta)$ come

$$p(\mathbf{x};\theta) = g(T(\mathbf{x}),\theta)h(\mathbf{x})$$

dove g è una funzione dipendente da $\mathbf{x}$ soltanto tramite la funzione $T(\mathbf{x})$ e h è una funzione dipendente solo da $\mathbf{x}$.

Esempio - DC level in WGN

Dimostriamo che in questo caso è possibile effettuare una fattorizzazione, assumendo che σ^2 sia noto. Riscriviamo l'esponente della PDF come

$$\sum_{n=0}^{N-1}(x[n]-A)^2 = \sum_{n=0}^{N-1}x^2[n] - 2A\sum_{n=0}^{N-1}x[n] + NA^2$$

in modo che la PDF si possa scrivere come

$$p(\mathbf{x};A) = \underbrace{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left[-\frac{1}{2\sigma^2}\left(NA^2 - 2A\sum_{n=0}^{N-1}x[n]\right)\right]}_{g(T(\mathbf{x}),A)}\underbrace{\exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right]}_{h(\mathbf{x})}$$

$T(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]$ è una statistica sufficiente per A.

Anche $T'(\mathbf{x}) = 2\sum_{n=0}^{N-1} x[n]$ è una statistica sufficiente per A

Qualsiasi funzione *biettiva* di $\sum_{n=0}^{N-1} x[n]$ è una statistica sufficiente.

Esempio - Potenza di WGN

Consideriamo ancora l'esempio precedente, stavolta con σ^2 sconosciuto e $A = 0$

Avremo

$$p(\mathbf{x}; \sigma^2) = \underbrace{\frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1} x^2[n]\right]}_{g(T(\mathbf{x}), A)} \cdot \underbrace{[1]}_{h(\mathbf{x})} .$$

Esempio - Fase di una sinusoide

Stimare la fase di una sinusoide in un rumore WGN

$$x[n] = A\cos(2\pi f_0 n + \phi) + w[n] \quad n = 0, 1, ...N-1.$$

L'ampiezza A della sinusoide e la frequenza f_0 sono note, così come la varianza σ^2. La likelihood function è

$$p(\mathbf{x};\phi) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}[x[n] - A\cos(2\pi f_0 n + \phi)]^2\right\}.$$

Si può espandere l'esponente come

$$\sum_{n=0}^{N-1} x^2[n] - 2A\sum_{n=0}^{N-1} x[n]\cos(2\pi f_0 n + \phi) + \sum_{n=0}^{N-1} A^2\cos^2(2\pi f_0 n + \phi).$$

Sfruttando la proprietà

$$\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$$

si ottiene

$$\sum_{n=0}^{N-1} x^2[n] - 2A\left(\sum_{n=0}^{N-1} x[n]\cos(2\pi f_0 n+\phi)\right)\cos\phi$$

$$+2A\left(\sum_{n=0}^{N-1} x[n]\sin(2\pi f_0 n+\phi)\right)\sin\phi + \sum_{n=0}^{N-1} A^2\cos^2(2\pi f_0 n+\phi).$$

Nessuna fattorizzazione del tipo del teorema di Neyman-Fisher $\Longrightarrow$

non esiste una *sola* statistica sufficiente.

Possiamo però fattorizzare la PDF come

$$p(\mathbf{x};\phi) = g(T_1(\mathbf{x}), T_2(\mathbf{x}), \phi)\cdot h(\mathbf{x})$$

in cui

$$g(T_1(\mathbf{x}), T_2(\mathbf{x}), \phi) = \\ = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{n=0}^{N-1} A^2\cos^2(2\pi f_0 n + \phi) + \right.\right. \\ \left.\left. -2AT_1(\mathbf{x})\cos\phi + 2AT_2(\mathbf{x})\sin\phi\right]\right\}.$$

e

$$h(\mathbf{x}) = \exp\left[\frac{1}{(2\pi\sigma^2)}\sum_{n=0}^{N-1} x^2[n]\right]$$

dove si è posto $T_1(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]\cos 2\pi f_0 n$ e
$T_2(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]\sin 2\pi f_0 n$

Esiste una generalizzazione del teorema, secondo cui $T_1(\mathbf{x})$ e $T_2(\mathbf{x})$ sono *congiuntamente* una statistica sufficiente per ϕ

Teorema di Rao-Blackwell-Lehmann-Scheffe

Se $\bar{\theta}$ è uno stimatore non polarizzato di θ e $T(\mathbf{x})$ è una statistica sufficiente per θ, allora $\hat{\theta} = E(\bar{\theta}|T(\mathbf{x}))$ è

- uno stimatore valido per θ (non dipendente da θ)
- non polarizzato
- uno stimatore con varianza minore o uguale a $\bar{\theta}$, per ogni θ.

Inoltre, se la statistica $T(\mathbf{x})$ sufficiente è completa allora $\hat{\theta}$ risulta essere lo stimatore MVU.

Una statistica si dice completa se esiste una sola funzione g della statistica che sia non polarizzata:

$$E[g(T(\mathbf{x}))] = \theta \ \ \forall\theta$$

o equivalentemente, data v(T)=g(T)-h(T),
$\int_{-\infty}^{\infty} v(T)p(T;\theta)dT = 0 \ \forall\theta$, è soddisfatta solo per $v(T) = 0, \ \forall T$

Procedura per trovare lo stimatore MVU

- Applicare il teorema di Neymann-Fisher e trovare una *singola* statistica $T(\mathbf{x})$ sufficiente per θ
- Se $T(\mathbf{x})$ è anche completo, procedere, altrimenti stop
- Trovare una funzione g in modo tale che $\hat{\theta} = g(T(\mathbf{x}))$ sia uno stimatore non polarizzato

$$\Longrightarrow \hat{\theta} \text{ è lo stimatore MVU.}$$

In generale, l'alternativa all'ultimo passo è calcolare lo stimatore come $\hat{\theta} = E(\bar{\theta}|T(\mathbf{x}))$ (difficile applicazione)

Esempio - DC level in WGN

Sappiamo che $\hat{A} = \bar{x}$ è lo stimatore MVU, in quanto raggiunge il limite CRLB.

Applichiamo comunque il teorema RBLS che può essere usato anche quando non esista uno stimatore efficiente, e dunque quando non sia attuabile il metodo CRLB.

Esistono due strade per trovare $\hat{A}$, lo stimatore MVU: entrambe sono basate sulla statistica sufficiente $T(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]$.

i. Trovare un qualsiasi stimatore non polarizzato di A, ad esempio $\check{A} = x[0]$, e determinare $\hat{A} = E(\check{A} \mid T)$. La media dev'essere effettuata rispetto a $p(\check{A} \mid T)$

ii. Trovare una funzione g tale che $\hat{A} = g(T)$ sia uno stimatore non polarizzato di A

Per quanto riguarda il primo metodo possiamo assumere che lo stimatore sia $\check{A} = x[0]$ e determinare $\hat{A} = E(x[0] \mid \sum_{n=0}^{N-1} x[n])$. Per un vettore $[x\ y]^T$, realizzazione di una distribuzione Gaussiana, avente come media un vettore $\mu = [E(x)E(y)]^T$ e matrice di covarianza

$$\mathcal{C} = \begin{bmatrix} \text{var}(x) & \text{cov}(x,y) \\ \text{cov}(x,y) & \text{var}(y) \end{bmatrix}$$

si può mostrare che

$$E(x \mid y) = \int_{-\infty}^{\infty} x p(x \mid y) dx$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} x \frac{p(x \mid y)}{p(y)} dx \\
&= E(x) + \frac{\mathrm{cov}(x,y)}{\mathrm{var}(y)}(y - E(y)). \qquad (4)
\end{aligned}
$$

Applicando questo risultato a $x = x[0]$ e $y = \sum_{n=0}^{N-1} x[n]$, si ha

$$
\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x[0] \\ \sum_{n=0}^{N-1} x[n] \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 1 & \dots & 1 \end{bmatrix}}_{\mathbf{L}} \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix}.
$$

Segue che la PDF di $[x\ y]^T$ è $\mathcal{N}(\boldsymbol{\mu}, \mathcal{C})$ in quanto rappresenta una

trasformazione lineare di un vettore Gaussiano, in cui

$$\boldsymbol{\mu} = \mathbf{L}E(\mathbf{x}) = \mathbf{L}A \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} A \\ NA \end{bmatrix}$$

$$\mathcal{C} = \sigma^2 \mathbf{L}\mathbf{L}^T = \sigma^2 \begin{bmatrix} 1 & 1 \\ 1 & N \end{bmatrix}$$

In definitiva lo stimatore MVU è dato da

$$\begin{aligned}\hat{A} &= E(x \mid y) = A + \frac{\sigma^2}{N\sigma^2}\left(\sum_{n=0}^{N-1} x[n] - NA\right) \\ &= \frac{1}{N}\sum_{n=0}^{N-1} x[n].\end{aligned}$$

Come anticipato, questo approccio, che richiede il calcolo della media condizionata, è generalmente difficile da trattare matematicamente

Legame fra l'unicità di g e la completezza della statistica

Il metodo prevede di trovare una qualche funzione g in modo che

$$\hat{A} = g\left(\sum_{n=0}^{N-1} x[n]\right)$$

sia uno stimatore non polarizzato di A: possiamo scegliere $g(x) \overset{\forall T}{=} x/N$, che ci porta ad ottenere

$$\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$$

come stimatore MVU.

Sappiamo che se $g(x)$ è unica abbiamo $E[g(\sum_{n=0}^{N-1} x[n])] = A$ e

$T(\mathbf{x})$ risulta completo. Tuttavia, supponiamo che esista un'altra funzione h tale che $E[h(\sum_{n=0}^{N-1} x[n])] = A$; questo vorrebbe dire che

$$E[g(T) - h(T)] = A - A = 0 \ \forall A.$$

Possiamo formalizzare il problema in maniera del tutto equivalente, imponendo che

$$\int_{-\infty}^{\infty} v(T)p(T;\theta)dT = f(A) \, \forall \theta$$

dove $v(T) = g(T) - h(T)$ con
$E[g(\sum_{n=0}^{N-1} x[n])] = E[h(\sum_{n=0}^{N-1} x[n])] = A$. Per dimostrare che $T(\mathbf{x})$ è completo, è sufficiente allora dimostrare che

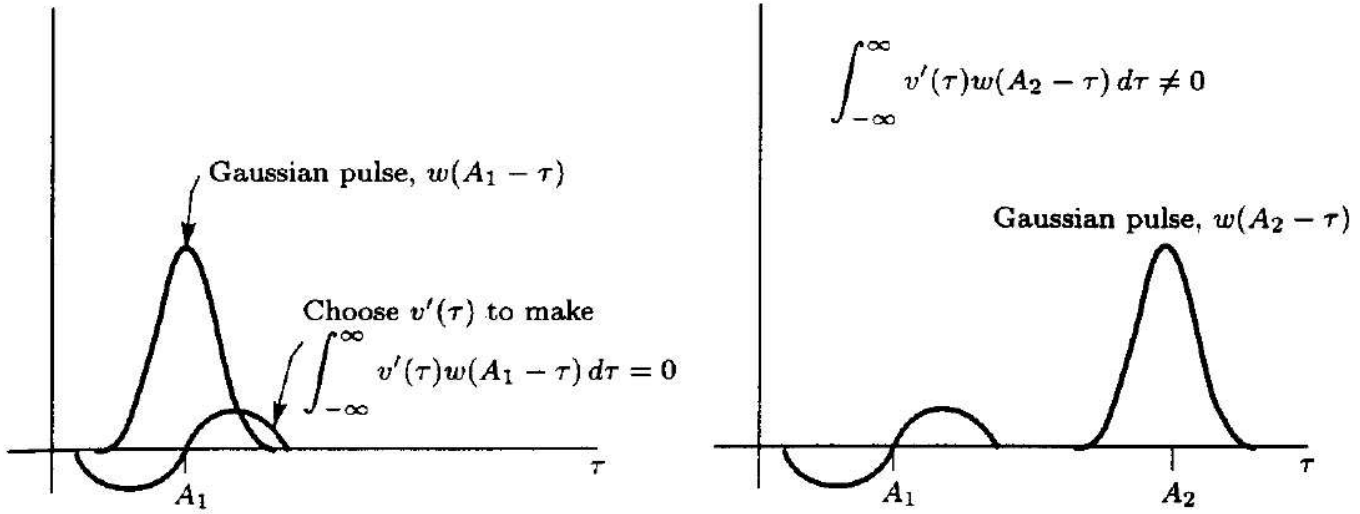
$$f(A) \overset{\forall A}{=} 0 \iff v(T) \overset{\forall T}{=} 0.$$

Dal fatto che $T \sim \mathcal{N}(NA, N\sigma^2)$, imponiamo $f(A) \overset{\forall A}{=} 0$, ottenendo

$$\int_{-\infty}^{\infty} v(T) \frac{1}{\sqrt{2\pi N\sigma^2}} \exp\left[-\frac{1}{2N\sigma^2}(T - NA)^2\right] dT = 0 \ \forall A.$$

Imponendo $\tau = T/N$ e $v'(\tau) = v(N\tau)$,

$$\int_{-\infty}^{\infty} v'(\tau) \frac{N}{\sqrt{2\pi N\sigma^2}} \exp\left[-\frac{N}{2\sigma^2}(A - \tau)^2\right] d\tau = 0 \ \forall A$$



in cui si può riconoscere la funzione di convoluzione fra $v'(\tau)$ con l'impulso Gaussiano $w(\tau)$

Affinché il risultato sia nullo, è necessario che $v'(\tau)$ sia identicamente nullo per tutti i valori di A. Un segnale è sempre nullo *se* e *solo se* la sua trasformata di Fourier è identicamente nulla; possiamo utilizzare questa proprietà per fissare la condizione

$$V'(f)W(f) = 0 \ \forall f$$

dove $V'(f) = \mathcal{F}[v'(\tau)]$ e $W(f)$ è la trasformata di Fourier dell'impulso Gaussiano. D'altra parte $W(f)$ è ancora Gaussiana e positiva per ogni valore di f, per cui la condizione è soddisfatta solo se $V'(f) = 0$. In altre parole dev'essere $v'(\tau) = 0 \, \forall \tau$, ovvero

$g = h$. Essendo la funzione g unica, la statistica $T(\mathbf{x})$ è completa e

$$\hat{A} = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$$

è lo stimatore MVU.

Esempio - Statistica sufficiente non-completa

Consideriamo il problema di stima di A per il dato

$$x[0] = A + w[0]$$

dove $w[0] \sim \mathcal{U}[-\frac{1}{2}, \frac{1}{2}]$

$x[0]$ statistica sufficiente (unico dato disponibile)

$x[0]$ stimatore non polarizzato

$g(x[0]) = x[0]$ è una statistica completa?

Supponiamo che esista un'altra funzione h con la proprietà di non polarizzazione $h(x[0]) = A$ e proviamo a dimostrare che $h = g$

Sia dunque $v(T) = g(T) - h(T)$, ed esaminiamo le possibili soluzioni per v dell'equazione

$$\int_{-\infty}^{\infty} v(T)p(\mathbf{x}; A)d\mathbf{x} = 0 \,\forall A.$$

$\mathbf{x} = x[0] = T \quad \Rightarrow$

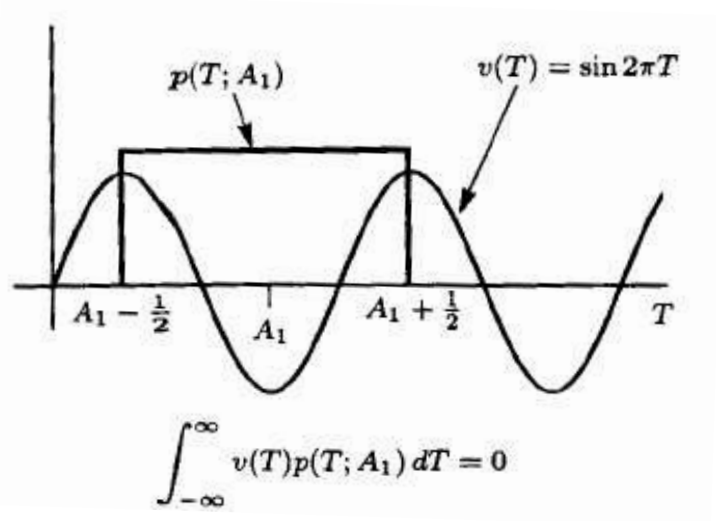
$$\int_{-\infty}^{\infty} v(T)p(T; A)dT = 0 \,\forall A$$

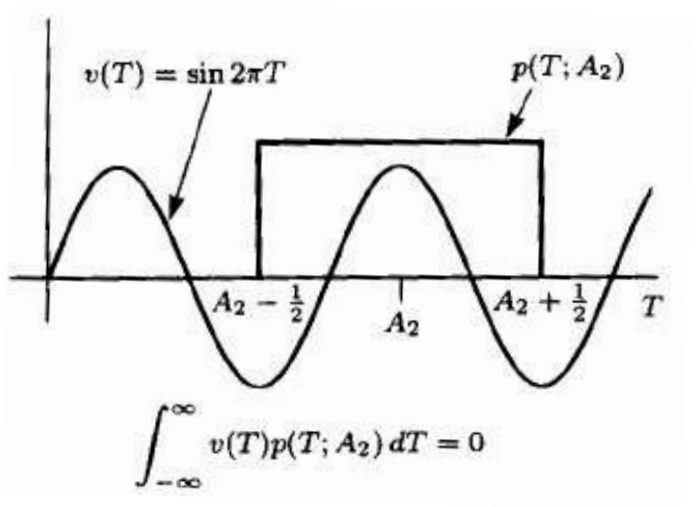
$$p(T; A) = \begin{cases} 1 & A - \frac{1}{2} \leq T \leq A + \frac{1}{2} \\ 0 & altrimenti \end{cases}$$

Dunque la condizione si riduce a

$$\int_{A-\frac{1}{2}}^{A+\frac{1}{2}} v(T)dT = 0 \,\forall A$$

Possiamo scegliere come funzione non nulla $v(T) = \sin 2\pi T$, che soddisfa questa condizione come illustrato in figura


$p(T; A_1)$
$v(T) = \sin 2\pi T$
$A_1 - \frac{1}{2}$
A_1
$A_1 + \frac{1}{2}$
T
$\int_{-\infty}^{\infty} v(T)p(T; A_1)\, dT = 0$


$v(T) = \sin 2\pi T$
$p(T; A_2)$
$A_2 - \frac{1}{2}$
A_2
$A_2 + \frac{1}{2}$
T
$\int_{-\infty}^{\infty} v(T)p(T; A_2)\, dT = 0$

La soluzione diviene

$$v(T) = g(T) - h(T) = \sin 2\pi T$$

ovvero

$$h(T) = T - \sin 2\pi T.$$

In definitiva, lo stimatore

$$\hat{A} = x[0] - \sin 2\pi x[0]$$

è basato su una statistica sufficiente per A e non è polarizzato.

Abbiamo trovato un altro stimatore non polarizzato con la stessa statistica sufficiente, pertanto possiamo concludere che la statistica non è completa; non è possibile applicare il teorema RBLS per trovare lo stimatore MVU.

Esempio - Media di un rumore bianco uniformemente distribuito

Supponiamo di osservare l'insieme di dati

$$x[n] = w[n] \quad n = 0, 1, ..., N-1,$$

in cui $w[n]$ siano i campioni di un rumore IID (indipendente e identicamente distribuito), con PDF $\mathcal{U}[0, \beta]$ con $\beta > 0$

Trovare lo stimatore MVU per la media $\theta = \beta/2$

Non è applicabile il teorema CRLB in quanto la distribuzione non soddisfa le condizioni di regolarità

Sembrerebbe abbastanza naturale scegliere per la media lo stimatore

$$\hat{\theta} = \frac{1}{N} \sum x[n]$$

La varianza è

$$\begin{aligned} \text{var}(\hat{\theta}) &= \frac{1}{N}\text{var}(x[n]) \\ &= \frac{\beta^2}{12N} \end{aligned}$$

Per determinare se questo sia lo stimatore MVU seguiremo una differente procedura. Definiamo il gradino unitario come

$$u(x) = \begin{cases} 1 & \text{per } x > 0 \\ 0 & \text{per } x < 0. \end{cases}$$

Possiamo riscrivere

$$p(x[n], \theta) = \frac{1}{\beta}[u(x([n]) - u(x[n] - \beta)],$$

in cui $\beta = 2\theta$, e pertanto la PDF dei dati è

$$p(\mathbf{x}, \theta) = \frac{1}{\beta^n} \prod_{n=0}^{N-1} [u(x([n]) - u(x[n] - \beta)].$$

Questa PDF è non nulla solo se $0 < x[n] < \beta$ per ogni $x[n]$,

pertanto si può riscrivere

$$p(\mathbf{x}, \theta) = \begin{cases} \frac{1}{\beta^n} & 0 < x[n] < \beta \quad n = 0, 1, ..., N-1 \\ 0 & \text{altrimenti} \end{cases}.$$

In alternativa possiamo scrivere

$$p(\mathbf{x}, \theta) = \begin{cases} \frac{1}{\beta^n} & \max x[n] < \beta, \min x[n] > 0 \\ 0 & \text{altrimenti} \end{cases}$$

per cui

$$p(\mathbf{x},\theta) = \underbrace{\frac{1}{\beta^n} u(\beta - \max x[n])}_{g(T(\mathbf{x}),\theta)} \underbrace{u(\min x[n])}_{h(\mathbf{x})}$$

Per il teorema di Neyman-Fisher $T(\mathbf{x}) = \max x[n]$ è una statistica sufficiente per θ, ed inoltre, omettendo la prova, possiamo dire che essa è anche una statistica completa.

Determiniamo il valore atteso di $T = \max x[n]$, calcolando dapprima la sua funzione di distribuzione cumulativa, tenendo presente che le variabili casuali sono IID:

$$\Pr\{T \leq \xi\} \quad = \quad \Pr\{x[0] \leq \xi, x[1] \leq \xi, ..., x[N-1] \leq \xi\}$$

$$= \prod_{n=0}^{N-1} \mathsf{Pr}\{x[n] \leq \xi\}$$
$$= \mathsf{Pr}\{x[n] \leq \xi\}^N.$$

Effettuando la derivata per avere la PDF

$$p_T(\xi) = \frac{d\,\mathsf{Pr}\{T \leq \xi\}}{d\xi}$$
$$= N\,\mathsf{Pr}\{x[n] \leq \xi\}^{N-1} \frac{d\,\mathsf{Pr}\{x[n] \leq \xi\}}{d\xi}.$$

Ma $\dfrac{d\,\Pr\{x[n] \leq \xi\}}{d\xi}$ è la PDF di $x[n]$

$$p_{x[n]}(\xi, \theta) = \begin{cases} \dfrac{1}{\beta} & 0 < \xi < \beta \\ 0 & \text{altrimenti} \end{cases}$$

Integrando otteniamo

$$\Pr\{x[n] \leq \xi\} = \begin{cases} 0 & \xi < 0 \\ \dfrac{\xi}{\beta} & 0 < \xi < \beta \\ 1 & \xi > \beta \end{cases}$$

che in definitiva produce

$$p_T(\xi) = \begin{cases} 0 & \xi < 0 \\ N \left(\frac{\xi}{\beta}\right)^{N-1} \frac{1}{\beta} & 0 < \xi < \beta \\ 0 & \xi > \beta \end{cases}$$

L'operazione di media produce

$$\begin{aligned} E(T) &= \int_{-\infty}^{\infty} \xi p_T(\xi) d\xi \\ &= \int_0^{\beta} \xi N \left(\frac{\xi}{\beta}\right)^{N-1} \frac{1}{\beta} d\xi \\ &= \frac{N}{N+1} \beta \end{aligned}$$

$$= \frac{2N}{N+1}\theta$$

Affinché questo valore sia non polarizzato, dobbiamo considerare il fattore moltiplicativo e definire $\hat{\theta} = [(N+1)/2N]T$, cioè

$$\hat{\theta} = \frac{N+1}{2N} \max_n x[n]$$

che è lo stimatore MVU.

Contrariamente a quanto l'intuito può suggerire, per un rumore uniformemente distribuito, la media campione non rappresenta lo stimatore MVU; si invita a verificare quanto appena trovato con una simulazione in ambiente MATLAB

Teorema: Neymann-Fisher Factorization (Caso Vettoriale)

Se risulta possibile fattorizzare la PDF $p(\mathbf{x};\boldsymbol{\theta})$ come

$$p(\mathbf{x};\boldsymbol{\theta}) = g(\mathbf{T}(\mathbf{x}),\boldsymbol{\theta})h(\mathbf{x})$$

dove g è una funzione dipendente da $\mathbf{x}$ attraverso $\mathbf{T}(\mathbf{x})$, statistica $r \times 1$, e da $\boldsymbol{\theta}$, e h è una funzione dipendente solo da $\mathbf{x}$, allora $\mathbf{T}(\mathbf{x})$ è una statistica sufficiente per $\boldsymbol{\theta}$.

Viceversa, se $\mathbf{T}(\mathbf{x})$ è una statistica sufficiente per $\boldsymbol{\theta}$, allora la PDF può essere fattorizzata

Se $\check{\boldsymbol{\theta}}$ risulta essere uno stimatore non polarizzato per $\boldsymbol{\theta}$ e $\mathbf{T}(\boldsymbol{\theta})$ è una statistica sufficiente $r \times 1$ per $\boldsymbol{\theta}$, allora $\hat{\boldsymbol{\theta}} = \mathbf{E}(\check{\boldsymbol{\theta}} \mid \mathbf{T}(\boldsymbol{\theta}))$ è

- uno stimatore valido per $\boldsymbol{\theta}$ (non dipendente da $\boldsymbol{\theta}$)
- non polarizzato
- uno stimatore con varianza minore o uguale a $\check{\boldsymbol{\theta}}$ (ogni elemento di $\hat{\boldsymbol{\theta}}$ ha varianza minore o uguale)

Inoltre, se la statistica sufficiente è anche completa, allora $\hat{\boldsymbol{\theta}}$ è lo stimatore MVU.

Problemi

- mancanza di un modello appropriato per la PDF
- anche quando la PDF è nota, l'applicazione dei metodi studiati finora non garantisce di trovare lo stimatore MVU

stimatore subottimo di cui possiamo valutare la varianza

approccio semplificativo: vincolo di **linearità dello stimatore rispetto ai dati osservati e calcolo dello stimatore MVU** $\longrightarrow$ ***Best Linear Unbiased Estimator (BLUE)***

valutabile semplicemente dalla conoscenza dei momenti di primo e secondo ordine della PDF

Definizione e Vincoli per il BLUE

Osservazione $\{x[0], x[1], ..., x[N-1]\}$

pdf $p(\mathbf{x}, \theta)$ dipendente da un parametro θ sconosciuto.

L'approccio BLUE vincola lo stimatore ad essere lineare, ovvero il problema consiste nel trovare delle costanti a_n tali che

$$\hat{\theta} = \sum a_n x[n]$$

Fra questi stimatori, il BLUE è definito come lo stimatore a minima varianza e non polarizzato

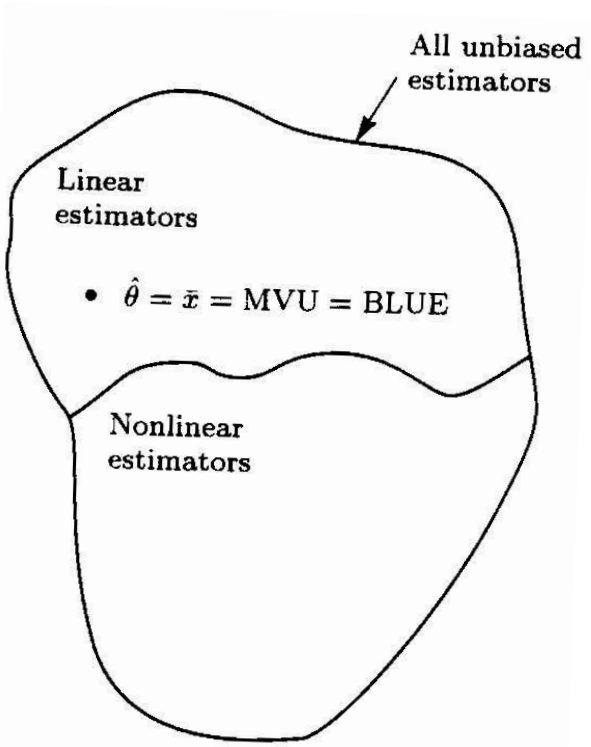
Il BLUE può essere ottimo o subottimo: è ottimo solo quando lo stimatore MVU ha in effetti un comportamento lineare

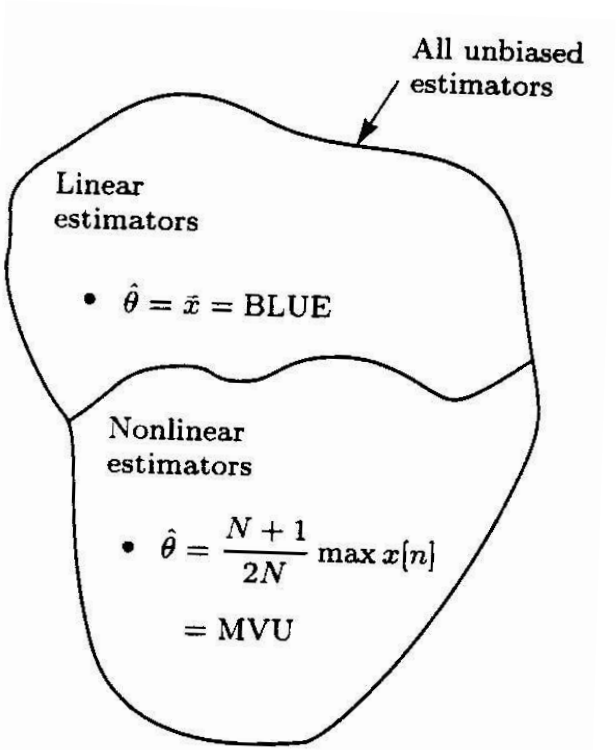
Il BLUE è ottimo nel caso del livello DC in WGN

$$\hat{\theta} = \bar{x} = \sum \frac{1}{N} x[n]$$

Il BLUE è subottimo nel caso di stima del livello di continua in rumore uniforme, in cui lo stimatore MVU è

$$\hat{\theta} = \frac{N+1}{2N} \max_n x[n]$$


All unbiased estimators
Linear estimators
• $\hat{\theta} = \bar{x}$ = MVU = BLUE
Nonlinear estimators


All unbiased estimators
Linear estimators
• $\hat{\theta} = \bar{x}$ = BLUE
Nonlinear estimators
• $\hat{\theta} = \frac{N+1}{2N} \max x[n]$
= MVU

Esistono problemi di stima per i quali il BLUE risulta completamente inappropriato, come ad esempio nel problema di stima della potenza di un processo WGN, per il quale lo stimatore MVU è

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=0}^{N-1} x^2[n]$$

che è evidentemente non lineare

Se vincoliamo lo stimatore ad essere lineare, in modo che

$$\hat{\sigma}^2 = \sum a_n x[n],$$

il valore atteso dello stimatore diventa

$$E(\hat{\sigma}^2) = \sum a_n E(x[n]) = 0,$$

in quanto $E(x[n]) = 0$ per tutti i valori di n

Non possiamo trovare uno stimatore lineare che sia non polarizzato.

Tuttavia, nonostante il BLUE non sia adatto al tipo di problema affrontato, utilizzando una trasformazione sui dati del tipo $y[n] = x^2[n]$, si produce uno stimatore efficace.

Vincolo di non polarizzazione per lo stimatore lineare $\hat{\theta}$

$$E(\hat{\theta}) = \sum a_n E(x[n]) = \theta$$

Si calcolano i coefficienti a_n in modo tale da minimizzare la varianza

Oltre al vincolo sul primo ordine della distribuzione di $\hat{\theta}$ si impone anche un vincolo sulla varianza - momento di secondo ordine:

$$\text{var}(\hat{\theta}) = E\left[\left(\sum a_n x[n] - E\left(\sum a_n x[n]\right)\right)^2\right]$$

Usando il vincolo $E(\hat{\theta}) = \theta$ e ponendo $\mathbf{a} = [a_0 a_1 ... a_{N-1}]^T$ abbiamo

$$\text{var}(\hat{\theta}) \quad = \quad E\left[(\mathbf{a}^T\mathbf{x} - \mathbf{a}^T E(\mathbf{x}))^2\right]$$

$$
\begin{aligned}
&= E\left[(\mathbf{a}^T(\mathbf{x} - E(\mathbf{x})))^2\right] \\
&= E\left[\mathbf{a}^T(\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))^T\mathbf{a}\right] \\
&= \mathbf{a}^T\mathcal{C}\mathbf{a}
\end{aligned}
$$

dove $\mathcal{C}$ è la matrice di covarianza

Il vettore $\mathbf{a}$ dei pesi può essere trovato minimizzando $\mathbf{a}^T\mathcal{C}\mathbf{a}$ con il vincolo di non polarizzazione

$$E(\hat{\theta}) = \sum a_n E(x[n]) = \theta \quad \forall\theta$$

$$\Rightarrow \qquad E(x[n]) = s[n]\theta$$

con coefficienti $s[n]$ noti, altrimenti il vincolo non può essere soddisfatto $\forall\theta$

Dunque $\sum a_n E(x[n]) = \theta \Rightarrow \sum a_n s[n]\theta = \theta \Rightarrow \sum a_n s[n] = 1$, cioè

$$\mathbf{a}^T\mathbf{s} = 1$$

Per minimizzare la quantità $\text{var}(\hat{\theta}) = \mathbf{a}^T \mathcal{C} \mathbf{a}$ rispetto al vincolo $\mathbf{a}^T \mathbf{s} = 1$ usiamo il metodo dei moltiplicatori di Lagrange, con la funzione Lagrangiana $\mathcal{J}$

$$\mathcal{J} = \mathbf{a}^T \mathcal{C} \mathbf{a} + \lambda(\mathbf{a}^T \mathbf{s} - 1).$$

Il gradiente della funzione Lagrangiana rispetto ad $\mathbf{a}$ è

$$\frac{\partial \mathcal{J}}{\partial \mathbf{a}} = 2\mathcal{C}\mathbf{a} + \lambda \mathbf{s}.$$

Uguagliando a zero e risolvendo, si ottiene

$$\mathbf{a} = -\frac{\lambda}{2} \mathcal{C}^{-1} \mathbf{s}.$$

La costante λ, moltiplicatore di Lagrange, si trova imponendo l'equazione del vincolo

$$\mathbf{a}^T\mathbf{s} = -\frac{\lambda}{2}\mathbf{s}^T\mathcal{C}^{-1}\mathbf{s} = 1,$$

da cui si deduce

$$-\frac{\lambda}{2} = \frac{1}{\mathbf{s}^T\mathcal{C}^{-1}\mathbf{s}}$$

e pertanto le due condizioni - gradiente nullo e vincolo di non polarizzazione - sono soddisfatte per

$$\mathbf{a}_{\text{opt}} = \frac{\mathcal{C}^{-1}\mathbf{s}}{\mathbf{s}^T\mathcal{C}^{-1}\mathbf{s}}$$

Perciò l'espressione per il BLUE risulta

$$\hat{\theta} = \mathbf{a}_{\text{opt}}^T \mathbf{x} = \frac{\mathbf{s}^T \mathcal{C}^{-1} \mathbf{x}}{\mathbf{s}^T \mathcal{C}^{-1} \mathbf{s}}$$

L'espressione della varianza per lo stimatore BLUE che si trova in corrispondenza di $\mathbf{a}_{\text{opt}}$ è

$$\begin{aligned} \text{var}(\hat{\theta}) &= \mathbf{a}_{\text{opt}}^T \mathcal{C} \mathbf{a}_{\text{opt}} \\ &= \frac{\mathbf{s}^T \mathcal{C}^{-1} \mathcal{C} \mathcal{C}^{-1} \mathbf{s}}{(\mathbf{s}^T \mathcal{C}^{-1} \mathbf{s})^2} \\ &= \frac{1}{\mathbf{s}^T \mathcal{C}^{-1} \mathbf{s}} \end{aligned}$$

Si noti che, essendo $E(\mathbf{x}) = \theta \mathbf{s}$, il BLUE è non polarizzato:

$$\begin{aligned} E(\hat{\theta}) &= \frac{\mathbf{s}^T \mathcal{C}^{-1} E(\mathbf{x})}{\mathbf{s}^T \mathcal{C}^{-1} \mathbf{s}} \\ &= \frac{\mathbf{s}^T \mathcal{C}^{-1} \theta \mathbf{s}}{\mathbf{s}^T \mathcal{C}^{-1} \mathbf{s}} = \theta \end{aligned}$$

Risulta più chiaro a questo punto il motivo per cui il BLUE necessita solo della conoscenza delle stime di primo e secondo ordine. Più precisamente, è necessaria la conoscenza di

1. $\mathbf{s}$, che è equivalente, a meno di una costante, alla media
2. $\mathcal{C}$, matrice di covarianza

Esempio - DC level in white noise

Supponiamo di osservare un insieme di dati

$$x[n] = A + w[n] \quad n = 0, 1, ..., N-1$$

$w[n]$ rumore bianco con varianza σ^2

PDF sconosciuta (o non definita)

Stimare A

Siccome il rumore $w[n]$ non è necessariamente Gaussiano, i campioni possono essere statisticamente dipendenti anche se scorrelati

$$E(x[n]) = A \Longrightarrow \mathbf{s} = [1, 1, ..., 1].$$

Dunque il BLUE è

$$\begin{aligned} \hat{A} &= \frac{\mathbf{s}^T \frac{1}{\sigma^2}\mathbf{I}\mathbf{x}}{\mathbf{s}^T \frac{1}{\sigma^2}\mathbf{I}\mathbf{s}} \\ &= \frac{1}{N}\sum x[n] = \bar{x} \end{aligned}$$

ed ha varianza minima

$$\begin{aligned} \text{var}(\hat{A}) &= \frac{1}{\mathbf{s}^T \frac{1}{\sigma^2}\mathbf{s}} \\ &= \frac{\sigma^2}{N} \end{aligned}$$

Dunque la media dei campioni risulta essere proprio il BLUE, e qualora il rumore si possa assumere con distribuzione Gaussiana, il BLUE è anche lo stimatore MVU.

Esempio - DC level in uncorrelated zero-mean noise

Sia adesso $w[n]$ un rumore non correlato e a media nulla, con varianza σ_n^2, e $x[n] = A + w[n]$

Ancora $\mathbf{s} = [1, 1, ..., 1]$, e

$$\begin{aligned} \hat{A} &= \frac{\mathbf{s}^T \mathcal{C}^{-1} \mathbf{x}}{\mathbf{s}^T \mathcal{C}^{-1} \mathbf{s}} \\ \text{var}(\hat{A}) &= \frac{1}{\mathbf{s}^T \mathcal{C}^{-1} \mathbf{s}}. \end{aligned}$$

La matrice di covarianza è

$$\mathcal{C} = \begin{bmatrix} \sigma_0^2 & 0 & \dots & 0 \\ 0 & \sigma_1^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{N-1}^2 \end{bmatrix}$$

e la sua inversa è

$$\mathcal{C}^{-1} = \begin{bmatrix} \frac{1}{\sigma_0^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_1^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_{N-1}^2} \end{bmatrix}$$

Dunque

$$\hat{A} = \frac{\mathbf{s}^T \mathcal{C}^{-1} \mathbf{x}}{\mathbf{s}^T \mathcal{C}^{-1} \mathbf{s}} = \frac{\sum \frac{x[n]}{\sigma_n^2}}{\sum \frac{1}{\sigma_n^2}}$$

$$\text{var}(\hat{A}) = \frac{1}{\mathbf{s}^T \mathcal{C}^{-1} \mathbf{s}} = \frac{1}{\sum \frac{1}{\sigma_n^2}}$$

Il denominatore è il fattore di scala necessario affinché sia soddisfatta la condizione di non polarizzazione

La matrice $\mathcal{C}^{-1}$ nello stimatore BLUE ha l'effetto di una operazione di *pre-whitening* sui dati che vengono successivamente mediati

BLUE - Caso Vettoriale

Se il parametro da stimare è un vettore $p \times 1$, allora la condizione da soddisfare diventa

$$\hat{\theta}_i = \sum a_{in} x[n] \quad i = 1, 2, ..., p$$

In forma matriciale $\hat{\boldsymbol{\theta}} = \mathbf{A}\mathbf{x}$ con $\mathbf{A}$ matrice $p \times N$

Per avere $\hat{\boldsymbol{\theta}}$ non polarizzato, dovrà essere

$$E(\hat{\theta}_i) = \sum a_{in} E(x[n]) = \theta_i \quad i = 1, 2, ..., p$$

cioè

$$E(\hat{\boldsymbol{\theta}}) = \mathbf{A} E(\mathbf{x}) = \boldsymbol{\theta}$$

che insieme al vincolo di non polarizzazione, implica

$$E(\mathbf{x}) = \mathbf{H}\boldsymbol{\theta}$$

con $\mathbf{H}$ matrice $N \times p$ nota, che generalizza il ruolo del vettore $\mathbf{s}$ incontrato nel caso scalare

In definitiva deve risultare

$$\mathbf{AH} = \mathbf{I}$$

Se definiamo il vettore $\mathbf{a}_i = \left[a_{i0} a_{i1} \ldots a_{i(N-1)}\right]$, in modo da avere $\hat{\theta}_i = \mathbf{a}_i^T \mathbf{x}$, il vincolo può essere riscritto per ogni $\mathbf{a}_i$, con

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \vdots \\ \mathbf{a}_p^T \end{bmatrix}$$

e

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \ldots & \mathbf{h}_p \end{bmatrix}$$

$$\mathbf{a}_i^T \mathbf{h}_i = \delta_{ij} \quad i = 1, 2, ..., p \; ; \; j = 1, 2, ...p$$

Per analogia con il caso scalare, la varianza è

$$\text{var}(\hat{\theta}_i) = \mathbf{a}_i^T \mathcal{C} \mathbf{a}_i$$

Il BLUE si determina minimizzando var$(\hat{\theta}_i)$ rispetto al vincolo $\mathbf{a}_i^T \mathbf{h}_i = \delta_{ij}$, e ripetendo la minimizzazione per ogni indice i. Si può dimostrare che la minimizzazione si ha per

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathcal{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathcal{C}^{-1} \mathbf{x}$$

e la matrice di covarianza risulta

$$\mathcal{C}_{\hat{\boldsymbol{\theta}}} = (\mathbf{H}^T \mathcal{C}^{-1} \mathbf{H})^{-1}$$

Generalizzazione dello stimatore MVU che avevamo derivato per rumore gaussiano bianco ($\mathcal{C}_{\hat{\boldsymbol{\theta}}} = \sigma^2\mathbf{I}$)

Quando i dati sono realmente Gaussiani, allora il BLUE corrisponde allo stimatore MVU:

in questo caso infatti, vincolare lo stimatore ad essere lineare non porta al compromesso di uno stimatore subottimo, in quanto l'MVU appartiene proprio all'insieme degli stimatori lineari.

Teorema di Gauss-Markov

Se i dati osservati sono nella forma generale di un modello lineare

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

$\mathbf{H}$ matrice $N \times p$ nota, $\boldsymbol{\theta}$ vettore $p \times 1$ dei parametri, $\mathbf{w}$ vettore di rumore $N \times 1$, con media nulla e covarianza $\mathcal{C}$,

allora il BLUE di $\boldsymbol{\theta}$ è

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathcal{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathcal{C}^{-1} \mathbf{x}$$

la minima varianza di $\hat{\theta}_i$ è

$$\text{var}(\hat{\theta}_i) = \left[(\mathbf{H}^T \mathcal{C}^{-1} \mathbf{H})^{-1}\right]_{ii}$$

e la matrice di covarianza di $\hat{\boldsymbol{\theta}}$ è

$$\mathcal{C}_{\hat{\boldsymbol{\theta}}} = (\mathbf{H}^T \mathcal{C}^{-1} \mathbf{H})^{-1}$$

Stima a massima verosimiglianza (MLE)

Maximum Likelihood Estimation

- metodo alternativo all'MVU
- quando l'MVU non esiste
- quando l'MVU non può essere trovato
- il più utilizzato nella pratica
  - semplice implementazione anche nel caso di complicati problemi di stima
  - buone prestazioni se disponiamo di grandi quantità di dati

MLE - Stimatori Consistenti (caso scalare)

L'MLE per un parametro scalare θ è il valore di θ che massimizza la funzione di verosimiglianza $p(\mathbf{x};\theta)$ di tutti i valori possibili di θ per $\mathbf{x}$ fissato

Esempio - DC level in WGN

Supponiamo di osservare i dati

$$x[n] = A + w[n], \quad n = 0, 1, \ldots, N-1$$

con A livello di segnale sconosciuto (ma positivo, $A > 0$), e $w[n]$ rumore WGN con varianza A

Proviamo a cercare l'MVU tramite il teorema CRLB

$$\begin{aligned} p(\mathbf{x}; A) &= \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi A}} \exp\left[-\frac{1}{2A}(x[n] - A)^2\right] \\ &= \frac{1}{(2\pi A)^{\frac{N}{2}}} \exp\left[-\frac{1}{2A}\sum_{n=0}^{N-1}(x[n] - A)^2\right] \end{aligned}$$

Effettuando la derivata della *log-likelihood function* (considerando il logaritmo della PDF come una funzione di A), otteniamo

$$\begin{aligned}\frac{\partial \ln p(\mathbf{x}; A)}{\partial A} &= -\frac{N}{2A} + \frac{1}{A}\sum(x[n] - A) + \frac{1}{2A^2}\sum(x[n] - A)^2 \\ &\stackrel{?}{=} I(A)(\hat{A} - A)\end{aligned}$$

Non è facile esprimere la derivata nella forma richiesta dal teorema di Cramer-Rao; in effetti si può dimostrare che ciò non è possibile, e di conseguenza che non esiste uno stimatore efficiente. Si può inoltre dimostrare che il CRLB per questo tipo di problema è

$$\text{var}(\hat{A}) \geq \frac{A^2}{N(A + \frac{1}{2})}$$

e che procedendo nel calcolo di uno stimatore efficiente mediante il teorema della fattorizzazione di Neyman-Fisher, non si può né ottenere una statistica sufficiente che sia non polarizzata, né ottenere una media condizionata matematicamente maneggevole.

Esaurite tutte le possibilità di utilizzare un approccio che porti ad uno stimatore ottimo o subottimo, dobbiamo affidarci ad un metodo da cui si possa derivare uno stimatore *approssimativamente ottimo*, ovvero che diventi efficiente per $N \to \infty$. Questa condizione comporta che

$$N \to \infty \implies \begin{cases} E(\hat{A}) & \to & A \\ \text{var}(\hat{A}) & \to & \text{CRLB} \end{cases}$$

Uno stimatore $\hat{A}$ che soddisfi la condizione sulla media è detto

asintoticamente non polarizzato.

Se soddisfa entrambe le condizioni, viene detto *asintoticamente efficiente*

Ovviamente per un insieme di dati finiti tuttavia non possiamo utilizzare dei metodi efficaci per valutare le caratteristiche di ottimalità

Ritornando all'esempio, possiamo proporre lo stimatore

$$\hat{A} = -\frac{1}{2} + \sqrt{\frac{1}{N}\sum x^2[n] + \frac{1}{4}}$$

che risulta polarizzato in quanto

$$\begin{aligned} E(\hat{A}) &= E\left(-\frac{1}{2} + \sqrt{\frac{1}{N}\sum x^2[n] + \frac{1}{4}}\right) \\ &\neq -\frac{1}{2} + \sqrt{E\left(\frac{1}{N}\sum x^2[n]\right) + \frac{1}{4}} \quad \forall A \\ &= -\frac{1}{2} + \sqrt{A + A^2 + \frac{1}{4}} \\ &= A \end{aligned}$$

Tuttavia lo stimatore è ragionevole, in quanto, per la legge dei grandi numeri, quando N tende ad infinito

$$\frac{1}{N}\sum x^2[n] \rightarrow E(x^2[n]) = A + A^2$$

pertanto

$$\hat{A} \rightarrow A$$

Lo stimatore $\hat{A}$ è detto *consistente*

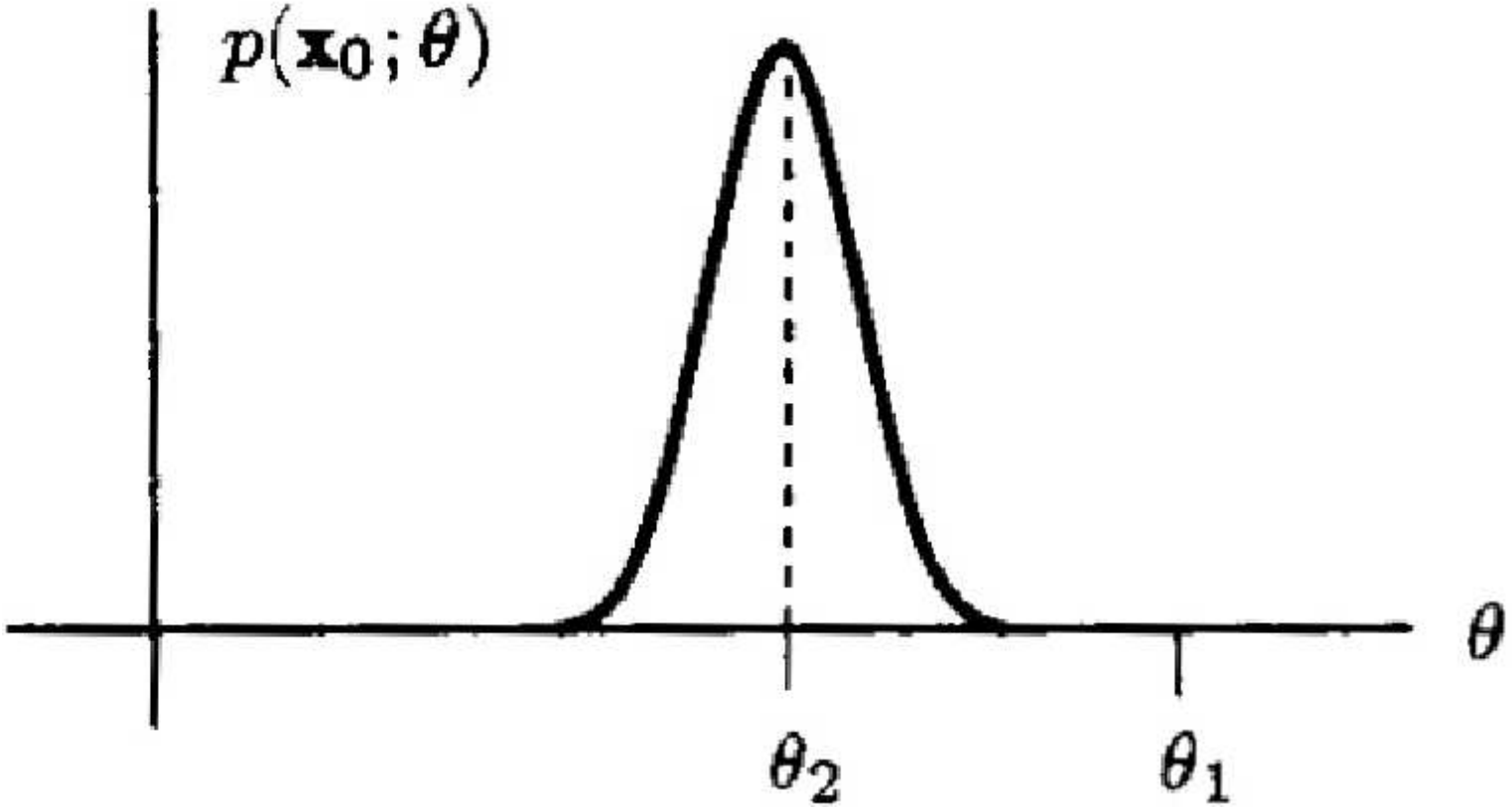
Si può inoltre dimostrare che è asintoticamente efficiente

Lo stimatore $\hat{A}$ proposto è proprio l'MLE

Trovare lo stimatore MLE

La stima a massima verosimiglianza per un parametro θ è il valore di θ che massimizza la funzione di verosimiglianza $p(\mathbf{x};\theta)$ di tutti i valori possibili di θ

Essendo $p(\mathbf{x};\theta)$ una funzione di $\mathbf{x}$, l'operazione di massimizzazione produce un $\hat{\theta}$ che è ancora una funzione di $\mathbf{x}$



In figura, la PDF è valutata per $\mathbf{x} = \mathbf{x}_0$. Il valore di $p(\mathbf{x} = \mathbf{x}_0; \theta)d\mathbf{x}$ per ogni valore di θ indica la probabilità di osservare $\mathbf{x}$ in una regione

di volume $d\mathbf{x}$ in $\mathcal{R}^N$, centrata attorno ad $\mathbf{x}_0$ per un dato valore di θ

Per $\theta = \theta_1$, la probabilità di osservare $\mathbf{x} = \mathbf{x}_0$ è trascurabile

E' invece altamente più probabile che $\theta = \theta_2$ sia il valore che produce la più alta probabilità di osservare $\mathbf{x} = \mathbf{x}_0$

$$p(\mathbf{x}; A) = \frac{1}{(2\pi A)^{\frac{N}{2}}} \exp\left[-\frac{1}{2A}\sum_{n=0}^{N-1}(x[n] - A)^2\right]$$

Derivando rispetto ad A otteniamo

$$\frac{\partial \ln p(\mathbf{x}; A)}{\partial A} = -\frac{N}{2A} + \frac{1}{A}\sum(x[n] - A) + \frac{1}{2A^2}\sum(x[n] - A)^2$$

che uguagliata a zero produce

$$A^2 + A - \frac{1}{N}\sum x^2[n] = 0$$

Le due soluzioni sono

$$\hat{A} = -\frac{1}{2} \pm \sqrt{\frac{1}{N}\sum x^2[n] + \frac{1}{4}},$$

e dovendo essere $A > 0$, scegliamo quella positiva.

Proprietà asintotiche dell'MLE

Se la PDF $p(\mathbf{x};\theta)$ dei dati $\mathbf{x}$ soddisfa la condizione di regolarità

$$E\left(\frac{\partial \ln p(\mathbf{x};A)}{\partial A}\right) = 0$$

$\Rightarrow$ stima MLE distribuita asintoticamente secondo la

$$\hat{\theta} \overset{a}{\sim} \mathcal{N}(\theta, I^{-1}(\theta))$$

dove $I(\theta)$ è la *Fisher Information* valutata per il valore vero del parametro sconosciuto.

Risultato importante che consente di valutare ed affermare l'ottimalità dell'MLE

Esempio - MLE per la stima della fase di una sinusoide

$$x[n] = A\cos(2\pi f_0 n + \phi) + w[n] \quad n = 0, 1, ...N-1$$

Note l'ampiezza A e la frequenza f_0 e con $w[n]$ WGN e varianza σ^2

Avevamo visto che esiste una statistica sufficiente congiunta

$$T_1(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]\cos 2\pi f_0 n$$

e

$$T_2(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]\sin 2\pi f_0 n$$

Ricerca della stima MLE tramite massimizzazione di

$$p(\mathbf{x};\phi) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}[x[n] - A\cos(2\pi f_0 n + \phi)]^2\right\}$$

o, equivalentemente, minimizzando

$$J(\phi) = \sum (x[n] - A\cos(2\pi f_0 n + \phi))^2.$$

Differenziando rispetto a ϕ

$$\frac{\partial J(\phi)}{\partial \phi} = 2\sum (x[n] - A\cos(2\pi f_0 n + \phi))A\sin(2\pi f_0 n + \phi)$$

che posto uguale a zero, restituisce

$$\sum x[n]\sin(2\pi f_0 n + \hat{\phi}) = A\sum \cos(2\pi f_0 n + \hat{\phi})\sin(2\pi f_0 n + \hat{\phi})$$

Per f_0 non troppo vicino a 0 e $1/2$, si ha

$$\frac{1}{N}\sum \cos(2\pi f_0 n + \hat{\phi}) \sin(2\pi f_0 n + \hat{\phi}) \approx 0.$$

Pertanto avremo

$$\sum x[n] \sin(2\pi f_0 n + \hat{\phi}) = 0,$$

che, sfruttando la proprietà

$$\sin(\alpha + \beta) = \sin\alpha \cos\beta + \cos\alpha \sin\beta,$$

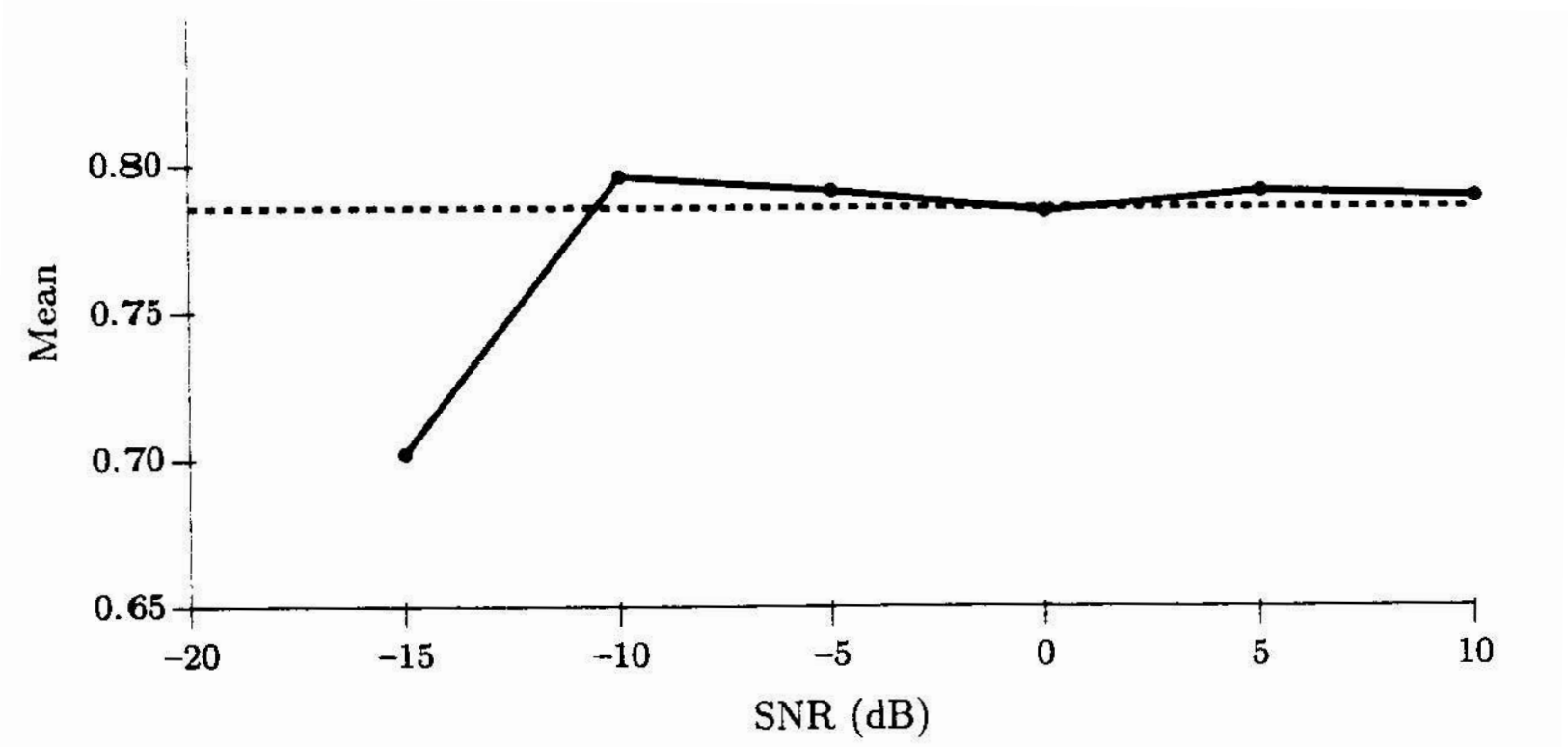
fornisce l'MLE della fase

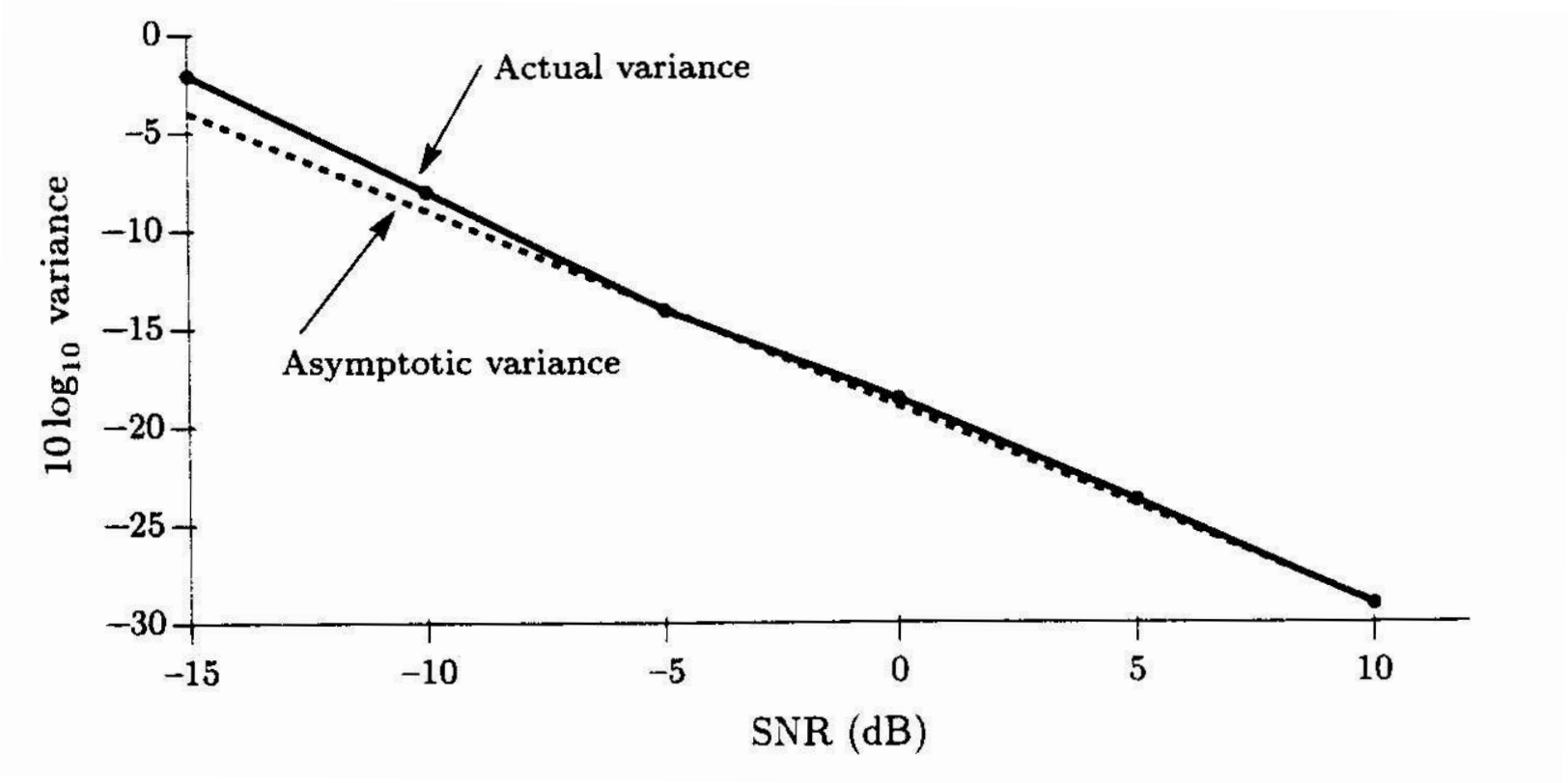
$$\hat{\phi} = -\arctan \frac{\sum x[n] \sin(2\pi f_0 n)}{\sum x[n] \cos(2\pi f_0 n)}$$

Data Record Length, N	Mean, $E(\hat{\phi})$	$N \times$ Variance, $N \operatorname{var}(\hat{\phi})$
20	0.732	0.0978
40	0.746	0.108
60	0.774	0.110
80	0.789	0.0990
Theoretical asymptotic value	$\phi = 0.785$	$1/\eta = 0.1$

$\eta = (A^2/2)/\sigma^2$ è il rapporto segnale rumore.

$$A = 1, \sigma^2 = 0.05, \phi = \pi/4, f_0 = 0.08$$


0.80
0.75
0.70
0.65
Mean
−20
−15
−10
−5
0
5
10
SNR (dB)



Varianza in funzione dell'SNR ($N = 80$)

Teorema: Proprietà di invarianza dell'MLE

L'MLE di un parametro $\alpha = g(\theta)$, dove la PDF $p(\mathbf{x}; \theta)$ è parametrizzata su θ, è dato da

$$\hat{\alpha} = g(\hat{\theta}),$$

in cui $\hat{\theta}$ è l'MLE per θ. L'MLE di $\hat{\theta}$ è ottenuto massimizzando $p(\mathbf{x}; \theta)$. Se g non è biettiva, allora $\hat{\alpha}$ massimizza la funzione modificata

$$\bar{p}(\mathbf{x}; \alpha) = \max_{\{\theta : \alpha = g(\theta)\}} p(\mathbf{x}; \theta).$$

Valutazione numerica dell'MLE

Ricerca numerica del massimo della funzione di verosimiglianza

Semplificazione del calcolo se la funzione $p(\mathbf{x};\theta)$ ammette solo valori di θ compresi in un intervallo $[a,b]$

Ricerca esaustiva (a griglia)

Problema per griglia fitta, se il range di θ non è un intervallo finito

$\Rightarrow$ Procedure iterative

- metodo di *Newton-Raphson*;
- metodo di *scoring*;
- metodo di *minimizzazione del valore atteso* (**Caso Vettoriale**)

In generale convergono se le condizioni iniziali dell’algoritmo sono abbastanza vicine al valore vero

La funzione da massimizzare non è nota a priori: la funzione di verosimiglianza cambia per ogni set di dati, pertanto è richiesto che la massimizzazione avvenga per una funzione *casuale*

Esempio - Metodo di Newton-Raphson - Metodo di Scoring

$$x[n] = r^n + w[n] \quad n = 0, 1, \ldots N-1$$

$w[n]$ rumore WGN con varianza σ^2

Stimare r, $r > 0$

Massimizzare

$$p(\mathbf{x}; r) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum(x[n] - r^n)^2\right]$$

o equivalentemente minimizzare

$$J(r) = \sum(x[n] - r^n)^2.$$

Differenziando $J(r)$ e ponendo la derivata uguale a zero, si ottiene

$$\sum (x[n] - r^n) n r^{n-1} = 0$$

equazione non lineare di r

1) Applichiamo il metodo di Newton-Raphson

Il metodo iterativo mira a massimizzare la *log-likelihood function* trovando gli zeri della derivata

$$\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta} = 0$$

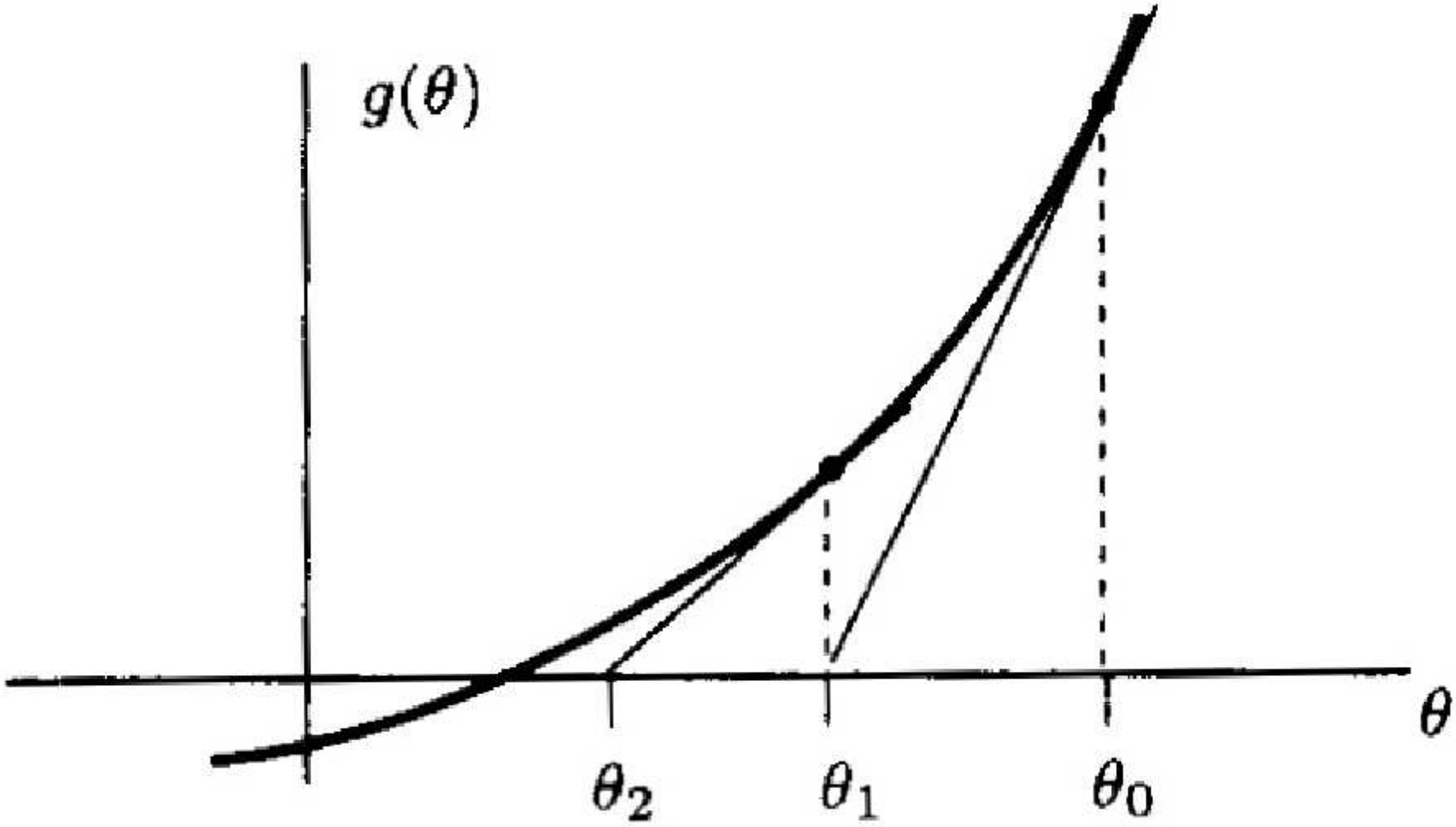
Sia

$$g(\theta) = \frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta},$$

e assumiamo di avere una ipotesi iniziale θ_0

Allora, se $g(\theta)$ è approssimativamente lineare nell'intorno di θ_0, possiamo approssimare la funzione con

$$g(\theta) \approx g(\theta_0) + \left.\frac{dg(\theta)}{d\theta}\right|_{\theta=\theta_0} (\theta - \theta_0)$$


$g(\theta)$
θ
θ_2
θ_1
θ_0

Metodo di Newton-Raphson per determinare gli zeri di una funzione

Risolviamo l'equazione rispetto a θ, e chiamiamo θ_1 lo zero trovato:

$$\theta_1 = \theta_0 - \frac{g(\theta_0)}{\left.\frac{dg(\theta)}{d\theta}\right|_{\theta=\theta_0}}.$$

Linearizziamo ancora la funzione g, ma stavolta nell'intorno di θ_1, e ripetiamo l'operazione eseguita al primo passo per trovare ancora lo zero della funzione linearizzata

L'iterazione al passo $k+1$ produce il valore θ_{k+1} dalla conoscenza

del valore θ_k, usando l'espressione

$$\theta_{k+1} = \theta_k - \frac{g(\theta_k)}{\left.\frac{dg(\theta)}{d\theta}\right|_{\theta=\theta_k}}.$$

Il metodo, se ben condizionato, condurrà allo zero della funzione $g(\theta)$, convergendo alla condizione $\theta_k = \theta_{k+1}$.

Possiamo anche riscrivere la condizione come

$$\theta_{k+1} = \theta_k - \left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta^2}\right]^{-1} \left.\frac{\partial \ln p(\mathbf{x};\theta)}{\partial\theta}\right|_{\theta=\theta_k}.$$

Due aspetti critici:

- l’iterazione può anche non convergere: derivata seconda della funzione di verosimiglianza al denominatore può produrre ampie fluttuazioni da iterazione a iterazione

- in caso di convergenza, il punto trovato potrebbe essere non un massimo assoluto, ma un massimo locale o addirittura un minimo locale. Per evitare questo tipo di problema tipicamente si scelgono più punti da cui far partire l’algoritmo

Applicando la condizione trovata all'esempio, si ottiene un metodo iterativo descritto da

$$r_{k+1} = r_k - \frac{\sum (x[n] - r_k^n) n r_k^{n-1}}{\sum n r_k^{n-2} [(n-1)x[n] - (2n-1) r_k^n]}$$

2) Applichiamo il metodo di *scoring*, che utilizza la *Fisher Information*

$$\left.\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta^2}\right|_{\theta=\theta_k} \approx -I(\theta_k)$$

valida per campioni IID e alti valori di N. Si ottiene con questa approssimazione

$$\begin{aligned}
\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta^2} &= \sum \frac{\partial^2 \ln p(x[n];\theta)}{\partial\theta^2} \\
&= N\frac{1}{N}\sum \frac{\partial^2 \ln p(x[n];\theta)}{\partial\theta^2} \\
&\approx NE\left[\frac{\partial^2 \ln p(x[n];\theta)}{\partial\theta^2}\right] \\
&= -Ni(\theta) = -I(\theta)
\end{aligned}$$

per la legge dei grandi numeri. La sostituzione della derivata seconda con il suo valore atteso migliora la stabilità dell'iterazione. Il metodo diventa

$$\theta_{k+1} = \theta_k + I^{-1}(\theta) \left. \frac{\partial \ln p(\mathbf{x}; \theta)}{\partial \theta} \right|_{\theta=\theta_k}$$

che è proprio la formulazione del metodo di *scoring*, che ha gli stessi problemi di convergenza del metodo di Newton-Raphson. Applicato al problema che stavamo considerando, abbiamo

$$I(\theta) = \frac{1}{\sigma^2} \sum n^2 r^{2n-2}$$

$$r_{k+1} = r_k - \frac{\sum (x[n] - r_k^n) n r_k^{n-1}}{\sum n^2 r_k^{2n-2}}$$

Algoritmo alternativo: EM

Si basa sul fatto che esistono insiemi di dati che permettono una stima dell'MLE molto più semplice rispetto all'insieme di dati originale: più precisamente, si suppone che esista un insieme di dati *completo*, e che esista una trasformazione invertibile g da un insieme incompleto di dati ad uno completo:

$$\mathbf{x} = g(y_1, y_2, \ldots, y_M) = g(\mathbf{y}).$$

L'obiettivo in questo caso è trovare l'MLE di θ non massimizzando la *log-likelihood function* $[\ln p_x(\mathbf{x}, \boldsymbol{\theta})]$, ma la $[\ln p_y(\mathbf{y}, \boldsymbol{\theta})]$ o equivalentemente, la

$$E_{\mathbf{x}|\mathbf{y}}[\ln p_y(\mathbf{y}, \boldsymbol{\theta})] = \int [\ln p_y(\mathbf{y}, \boldsymbol{\theta})][\ln p_y(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})] d\mathbf{y}$$

utilizzando nell'integrale l'informazione iniziale θ_k.

- Expectation: determinare la log-likelihood media del set di dati completo $\mathcal{U}(\boldsymbol{\theta}, \boldsymbol{\theta_k}) = \int [\ln p_y(\mathbf{y}, \boldsymbol{\theta})][\ln p_y(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})] d\mathbf{y}$

- Maximization:

$$\boldsymbol{\theta_{k+1}} = \arg \max_{\boldsymbol{\theta}} \mathcal{U}(\boldsymbol{\theta}, \boldsymbol{\theta_k})$$

Least Squares

Classe di stimatori che in generale non possiede proprietà di ottimalità, ma risulta efficace in molti casi di interesse pratico

Gauss, 1795. Metodo applicato allo studio dei moti planetari

Modello per il segnale in esame

Nessuna ipotesi probabilistica sui dati

Non ci sono criteri per stabilire l'ottimalità del metodo

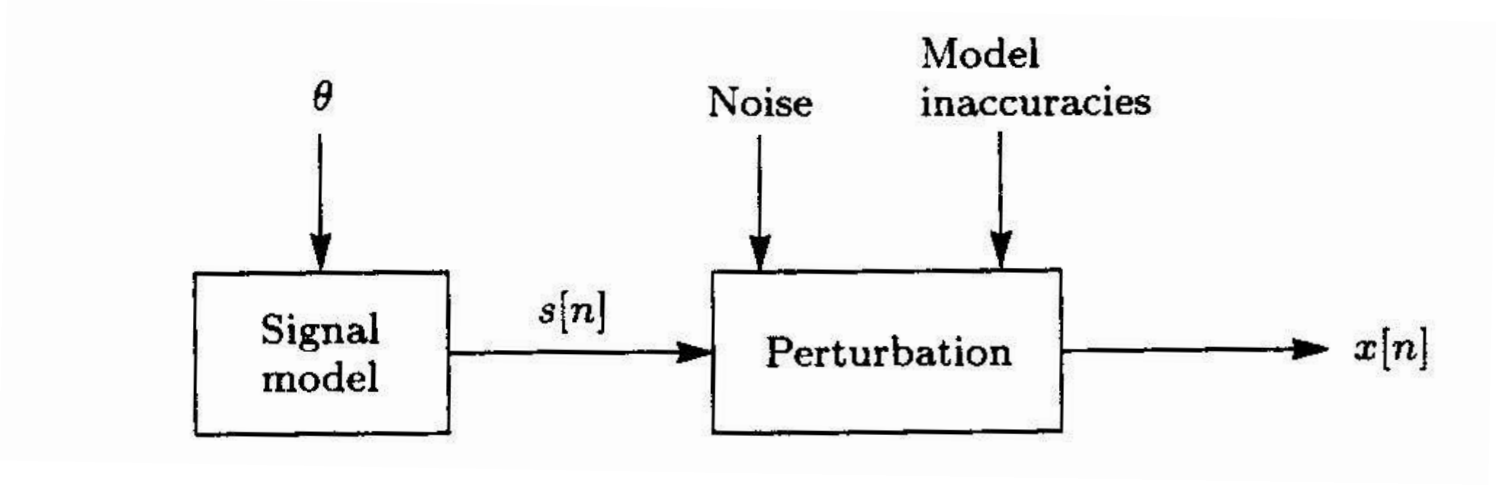
Non si possono valutare le prestazioni per la mancanza di specifiche assunzioni sulle proprietà statistiche dei dati

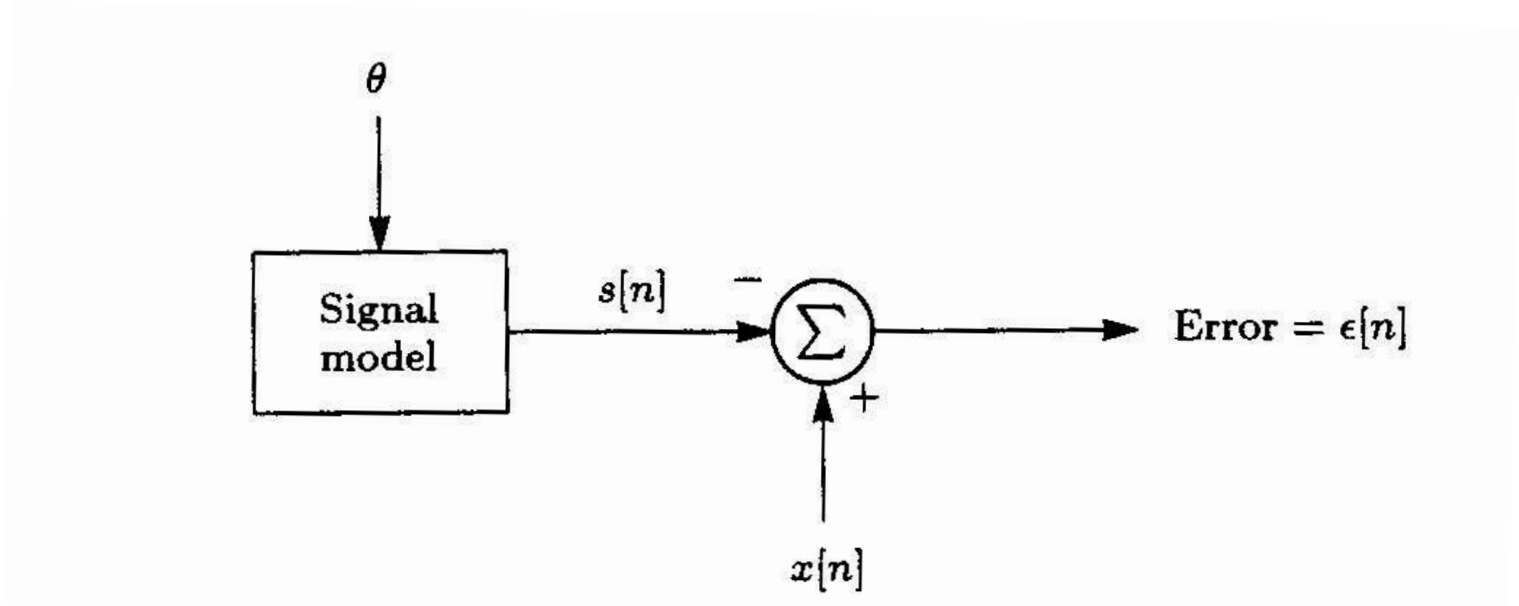
Semplice implementazione: minimizzare un errore ai minimi quadrati

Generalità sul problema dei minimi quadrati

I metodi descritti in precedenza utilizzano la varianza dello stimatore come una misura della qualità della stima effettuata, con l'obiettivo comune di minimizzare la differenza fra la stima e il valore vero (in media)

Nel ***metodo dei minimi quadrati***, si cerca di minimizzare la differenza quadratica fra i dati $x[n]$ acquisiti e il modello di segnale senza rumore


θ
Noise
Model inaccuracies
Signal model
s[n]
Perturbation
x[n]


θ
Signal model
s[n]
−
+
Error = ε[n]
x[n]

Segnale $s[n]$, deterministico, generato da modelli dipendenti da parametro θ sconosciuto

Osservazione $x[n]$ (nessuna assunzione probabilistica)

Errore LS

$$J(\theta) = \sum (x[n] - s[n])^2$$

Prestazioni dipendenti dalle proprietà del rumore e dalle caratteristiche del modello utilizzato

Tipologia del problema

1. $s[n]$ lineare in $\theta \longrightarrow$ comportamento quadratico per $J(\theta)$ e il problema si dice LLS, Linear Least Square Problem

2. problema non lineare, che richiede metodi esaustivi (*grid searches*), oppure metodi iterativi $\longrightarrow$ NLS

Esempio - Segnale sinusoidale

Sia $s[n] = A\cos 2\pi f_0 n$, nota f_0 ed A da stimare

$$J(A) = \sum (x[n] - A\cos 2\pi f_0 n)^2,$$

semplice, in quanto $J(A)$ è quadratico in A.

Se invece dobbiamo stimare f_0 invece di A, il problema è meno semplice, in quanto non lineare

Se è richiesta la stima di entrambi i parametri,

$$J(A, f_0) = \sum (x[n] - A\cos 2\pi f_0 n)^2$$

è quadratico in A ma non quadratico in f_0

$\Rightarrow$ *problema ai minimi quadrati separabili*

Minimi quadrati lineari - caso scalare

Per questo tipo di problema, dobbiamo assumere

$$s[n] = \theta h[n]$$

in cui $h[n]$ è una sequenza nota. Il criterio LS diventa

$$J(\theta) = \sum (x[n] - \theta h[n])^2.$$

La minimizzazione produce l'LSE

$$\hat{\theta} = \frac{\sum x[n]h[n]}{\sum h^2[n]} = \frac{\mathbf{x}^T\mathbf{h}}{\mathbf{h}^T\mathbf{h}}$$

$$
\begin{aligned}
J_{min} = J(\hat{\theta}) &= \sum (x[n] - \hat{\theta} h[n])(x[n] - \hat{\theta} h[n])) \\
&= \sum x[n](x[n] - \hat{\theta} h[n]) - \hat{\theta} \underbrace{\sum h[n](x[n] - \hat{\theta} h[n])}_{S=0} \\
&= \sum x^2[n] - \hat{\theta} \sum x[n]h[n]
\end{aligned}
$$

$$
J_{min} = \sum x^2[n] - \frac{\left(\sum x[n]h[n]\right)^2}{\sum h^2[n]}.
$$

che in definitiva può essere riscritta come

$$
J_{min} = E_x - \frac{\left(\sum x[n]h[n]\right)^2}{E_h}.
$$

Dunque in generale possiamo dire che

$$0 \leq J_{min} \leq E_x.$$

Minimi quadrati lineari - caso vettoriale

Segnale $\mathbf{s} = [s[0]s[1] \ldots s[N-1]]^T$ lineare nel parametro $\boldsymbol{\theta}$, vettore di dimensioni $p \times 1$

$$\mathbf{s} = \mathbf{H}\boldsymbol{\theta}$$

dove $\mathbf{H}$ è una matrice $N \times p$ nota ($N > p$) di rango p, detta *matrice di osservazione*. LSE minimizza

$$\begin{aligned} J(\boldsymbol{\theta}) &= \sum (x[n] - s[n])^2 \\ &= (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \end{aligned}$$

$$\begin{aligned} J(\boldsymbol{\theta}) &= \mathbf{x}^T\mathbf{x} - \mathbf{x}^T\mathbf{H}\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{x} + \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\theta} \\ &= \mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\mathbf{H}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\theta}. \end{aligned}$$

Il gradiente è

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -2\mathbf{H}^T\mathbf{x} + 2\mathbf{H}^T\mathbf{H}\boldsymbol{\theta},$$

che posto uguale a zero, restituisce come soluzione

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$$

L'equazione $\mathbf{H}^T\mathbf{H}\boldsymbol{\theta} = \mathbf{H}^T\mathbf{x}$ risolta in $\hat{\boldsymbol{\theta}}$ è denominata *equazione normale*; l'ipotesi che la matrice $\mathbf{H}$ abbia rango pari a p, assicura l'invertibilità di $(\mathbf{H}^T\mathbf{H})$

Stessa espressione dello stimatore BLUE, ma il BLUE richiede che siano verificate le ipotesi

$$E(\mathbf{x}) = \mathbf{H}\boldsymbol{\theta}$$

$$\mathcal{C}_{\mathbf{x}} = \sigma^2\mathbf{I}$$

$\mathbf{x}$ Gaussiano

Soluzione generale per il metodo dei minimi quadrati lineari (si ottiene per sostituzione)

$$J_{min} = \mathbf{x}^T(\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}})$$

Weighted LS

Introduciamo una matrice $\mathbf{W}$ di pesi, $N \times N$, simmetrica

$$J(\boldsymbol{\theta}) \quad = \quad (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{W} (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})$$

La matrice $\mathbf{W}$ viene introdotta con l'obiettivo di dare più importanza ai dati ritenuti più attendibili.

In questo caso l'LSE è

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{W} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{W} \mathbf{x}$$

e sostituendo, si ottiene l'errore minimo

$$J_{min} = \mathbf{x}^T (\mathbf{W} - \mathbf{W}\mathbf{H}(\mathbf{H}^T \mathbf{W} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{W}) \mathbf{x}$$

Interpretazione Geometrica

LS da un punto di vista geometrico

Modello di segnale $\mathbf{s} = \mathbf{H}\boldsymbol{\theta}$

$$\mathbf{s} = [\, \mathbf{h}_1 \, \mathbf{h}_2 \, \ldots \mathbf{h}_p \,] \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix} = \sum_{i=1}^{p} \theta_i \mathbf{h}_i$$

segnale visto come una combinazione lineare delle colonne $\mathbf{h}_i$

Esempio - Sovrapposizione di sinusoidi

$$s[n] = a\cos 2\pi f_0 n + b\sin 2\pi f_0 n, \quad n = 0, 1, \ldots, N-1$$

f_0 frequenza nota, $\boldsymbol{\theta} = [a\ b]^T$ vettore da stimare

$$\begin{bmatrix} s[0] \\ s[1] \\ \vdots \\ s[N-1] \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \cos 2\pi f_0 & \sin 2\pi f_0 \\ \vdots & \vdots \\ \cos 2\pi f_0(N-1) & \sin 2\pi f_0(N-1) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

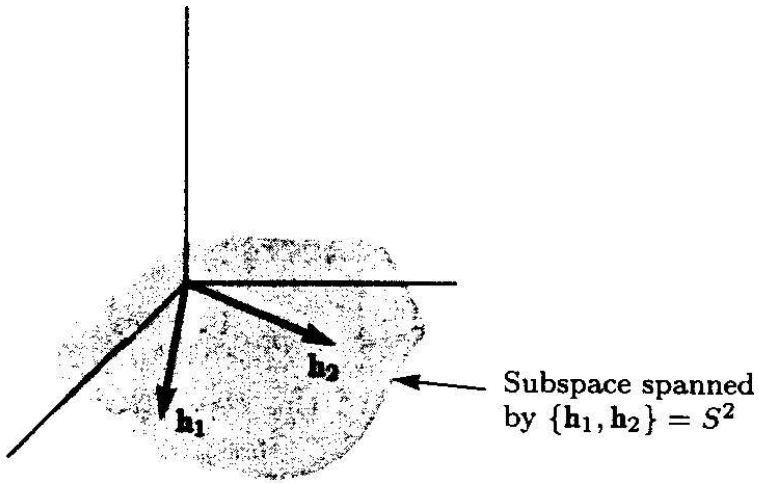
Colonne di $\mathbf{H}$ sono campioni di sequenze sinusoidali

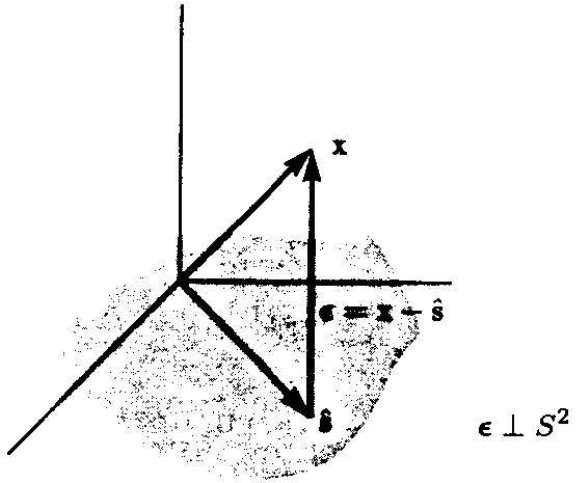
$$J(\boldsymbol{\theta}) = \| \mathbf{x} - \mathbf{H}\boldsymbol{\theta} \|^2 = \| \mathbf{x} - \sum_{i=1}^{p} \theta_i \mathbf{h}_i \|^2$$

Minimizzazione del quadrato della distanza fra i dati contenuti nel vettore $\mathbf{x}$ e il vettore $\sum_{i=1}^{p} \theta_i \mathbf{h}$, combinazione lineare delle colonne di $\mathbf{H}$

$\mathbf{H}$ ha rango p, pertanto genera uno spazio $p-$dimensionale, S^p sottospazio di $\mathcal{R}^n$

Esempio per $N = 3$ e $p = 2$





Tutte le possibili scelte per θ_1 e θ_2 producono dei segnali che appartengono al sottospazio

in generale $\mathbf{x}$ non appartiene ad S^2

il vettore $\hat{\mathbf{s}}$ nel sottospazio S^2 è il più vicino - in senso Euclideo - alla componente di $\mathbf{x}$ nel sottospazio

$\hat{\mathbf{s}}$ è la *proiezione ortogonale* di $\mathbf{x}$ sul sottospazio S^2

Condizione di ortogonalità: $(\mathbf{x} - \hat{\mathbf{s}}) \perp S^2$

Dunque

$$(\mathbf{x} - \hat{\mathbf{s}}) \perp \mathbf{h}_1$$

$$(\mathbf{x} - \hat{\mathbf{s}}) \perp \mathbf{h}_2$$

ovvero

$$(\mathbf{x} - \hat{\mathbf{s}})^T \mathbf{h}_1 = 0$$

$$(\mathbf{x} - \hat{\mathbf{s}})^T \mathbf{h}_2 = 0$$

Poiché $\hat{\mathbf{s}} = \theta_1 \mathbf{h}_1 + \theta_2 \mathbf{h}_2$, abbiamo

$$(\mathbf{x} - \theta_1 \mathbf{h}_1 - \theta_2 \mathbf{h}_2)^T \mathbf{h}_1 = 0$$

$$(\mathbf{x} - \theta_1 \mathbf{h}_1 - \theta_2 \mathbf{h}_2)^T \mathbf{h}_2 = 0$$

In forma matriciale

$$(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{h}_1 = 0$$

$$(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{h}_2 = 0.$$

Combinando le due equazioni si ottiene

$$(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{H} = \mathbf{0}^T$$

e in definitiva otteniamo l'LSE come

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x}.$$

Si noti che se $\boldsymbol{\epsilon} = \mathbf{x} - \mathbf{H}\boldsymbol{\theta}$ indica il vettore dell'errore, allora l'LSE può essere trovato dalla condizione

$$\boldsymbol{\epsilon}^T \mathbf{H} = \mathbf{0}^T.$$

principio di ortogonalità: errore ortogonale alle colonne di $\mathbf{H}$

L'errore rappresenta la parte di $\mathbf{x}$ che non può essere descritta dal modello assunto per il segnale

Il minimo errore LS è dato da

$$\| \mathbf{x} - \hat{\mathbf{s}} \|^2 = \| \mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}} \|^2 = (\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}})^T (\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}).$$

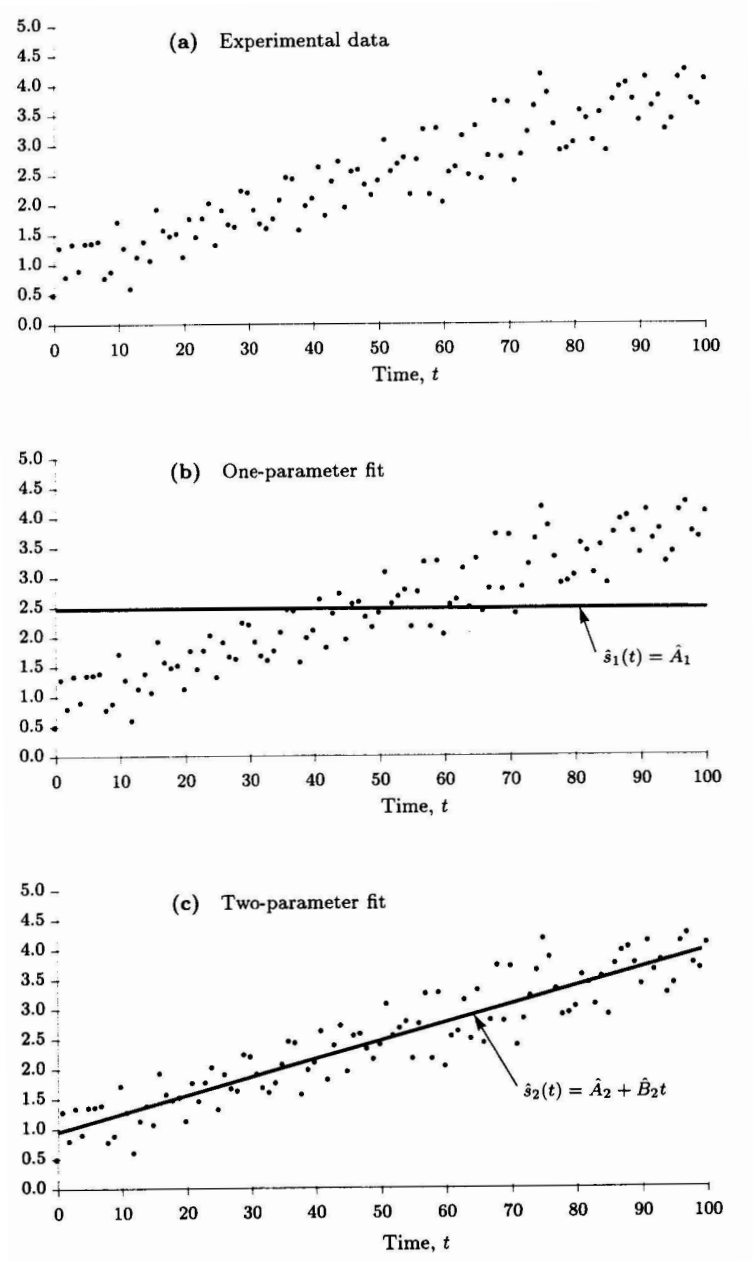
Soluzione LS interpretata come un problema di approssimazione di un vettore $\mathbf{x}$ in R^n con un altro vettore $\hat{\mathbf{s}}$, che è una combinazione lineare dei vettori $\{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_p\}$, in un sottospazio $p-$dimensionale di R^n

La soluzione è $\hat{\mathbf{s}}$ proiezione ortogonale di $\mathbf{x}$ nel sottospazio, cioè il vettore più vicino in senso Euclideo

Principali tecniche iterative LLS

- order recursive least squares;
- sequential least squares
- constrained least squares.

Modello per il segnale sconosciuto $\Rightarrow$ ipotesi adeguate


(a) Experimental data
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
10
20
30
40
50
60
70
80
90
100
Time, t
(b) One-parameter fit
$\hat{s}_1(t) = \hat{A}_1$
Time, t
(c) Two-parameter fit
$\hat{s}_2(t) = \hat{A}_2 + \hat{B}_2 t$
Time, t

Possiamo assumere come modello

$$s_1(t) = A$$

$$s_2(t) = A + Bt,$$

con $0 \leq t \leq T$

Campionamento

$$s_1(n) = A$$

$$s_2(n) = A + Bn.$$

Usando l'LSE con

$$\mathbf{H}_1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad \mathbf{H}_2 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & N-1 \end{bmatrix}$$

si ottiene lo stimatore

$$\hat{A} = \bar{x}$$

e

$$\hat{A}_2 = \frac{2(2N-1)}{N(N+1)} \sum x[n] - \frac{6}{N(N+1)} \sum nx[n]$$

$$\hat{B}_2 = -\frac{6}{N(N+1)} \sum x[n] + \frac{12}{N(N^2-1)} \sum nx[n]$$

$\hat{s}_1(t) = \hat{A}_1$ ed $\hat{s}_2(t) = \hat{A}_2 + \hat{B}_2 t$, in figura, con $T = 100$

Utilizzando due parametri, l'approssimazione migliora; l'errore LS minimo infatti deve decrescere all'aumentare dei parametri, fino a saturare, nel senso che ad un certo punto l'aggiunta di parametri non potrà più migliorare l'approssimazione, andando semplicemente ad approssimare l'errore dovuto al rumore come se fosse appartenente ai dati

Order recursive least squares

Riduce il calcolo dell'LSE aggiornandolo ad ogni passo

Il metodo permette di calcolare l'LSE basandosi su una matrice $\mathbf{H}$ di dimensioni $N \times (k+1)$ da una soluzione precedente basata su una matrice $\mathbf{H}$ di dimensioni $N \times k$

Per l'esempio precedente,supponiamo infatti di alterare l'intervallo di osservazione da $[0, N]$ a $[-M, M]$, facendo in modo che sia simmetrico. Adesso la matrice $\mathbf{H}_2$ presenta le colonne

ortogonalizzate,

$$\mathbf{H}_2 = \begin{bmatrix} 1 & -M \\ 1 & -(M-1) \\ \vdots & \vdots \\ 1 & M \end{bmatrix}.$$

Pertanto l'LSE può essere facilmente trovato per il fatto che

$$\mathbf{H}_2{}^T\mathbf{H}_2 = \begin{bmatrix} 2M+1 & 0 \\ 0 & \sum_{n=-M}^{M} n^2 \end{bmatrix}.$$

è una matrice diagonale. Le soluzioni sono

$$\hat{A}_1 = \frac{1}{2M+1} \sum_{n=-M}^{M} x[n]$$

$$\hat{A}_2 = \frac{1}{2M+1} \sum_{n=-M}^{M} x[n]$$

$$\hat{B}_2 = \frac{\sum_{n=-M}^{M} nx[n]}{\sum_{n=-M}^{M} n^2}.$$

In questo caso l'LSE di A non cambia se aggiungiamo dei parametri al modello

In termini geometrici possiamo dire che l'ortogonalità delle colonne della matrice ci permette di proiettare $\mathbf{x}$ lungo le due direzioni identificate dalle colonne separatamente e successivamente di *sovrapporre* i risultati (si noti la somiglianza con il principio di sovrapposizione degli effetti)

Ogni proiezione è indipendente dalle altre. In generale, le colonne della matrice $\mathbf{H}$ non sono ortogonali, ma possono essere sempre sostituite da un set di p vettori ortogonali (si ricordi infatti che la matrice ha rango p), tramite il processo di ortogonalizzazione di Gram-Schmidt. Nel caso $p = 2$, la colonna $\mathbf{h}_2$ viene proiettata nello spazio ortogonale ad $\mathbf{h}_1$; successivamente, essendo $\mathbf{h}_1$ ed $\mathbf{h}'_2$

ortogonali, la stima LS diventa

$$\hat{s} = \mathbf{h}_1\hat{\theta}_1 + \mathbf{h}_2'\hat{\theta}_2',$$

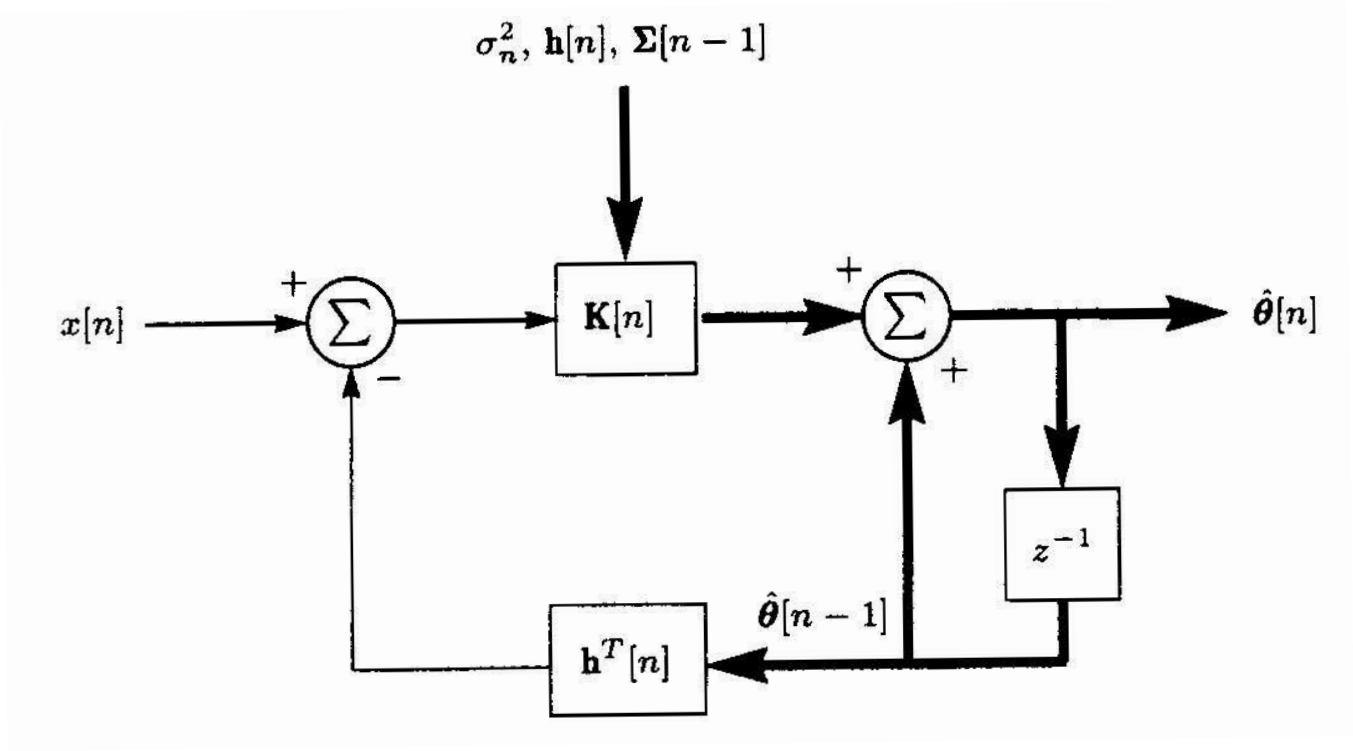
in cui $\mathbf{h}_1\hat{\theta}_1$ è anche l'LSE per il segnale, basato su $\mathbf{H} = \mathbf{h}_1$. questa procedura aggiunge ricorsivamente delle componenti al modello; si noti che il numero massimo di componenti che è possibile aggiungere corrisponde proprio a p, rango della matrice, oltre il quale non è più possibile aggiungere nessuna informazione aggiuntiva (o equivalentemente, non è possibile aggiungere una componente linearmente indipendente).

Sequential least squares

Nasce dalla necessità di analizzare ed elaborare dati in tempo reale, effettuando delle operazioni di stima senza attendere che il vettore di dati sia stato acquisito completamente

Assumendo di aver determinato l'LSE $\boldsymbol{\theta}$ a partire da $N-1$ campioni, il metodo permette di aggiornare il $\boldsymbol{\theta}$ dopo l'assunzione del campione $N-$esimo

Si aggiunge allo stimatore al passo $(N-1)-$esimo un termine correttivo, che deve decrescere al crescere di N in quanto lo stimatore 'al passo precedente' contiene sempre più campioni e deve pertanto avere più 'peso'

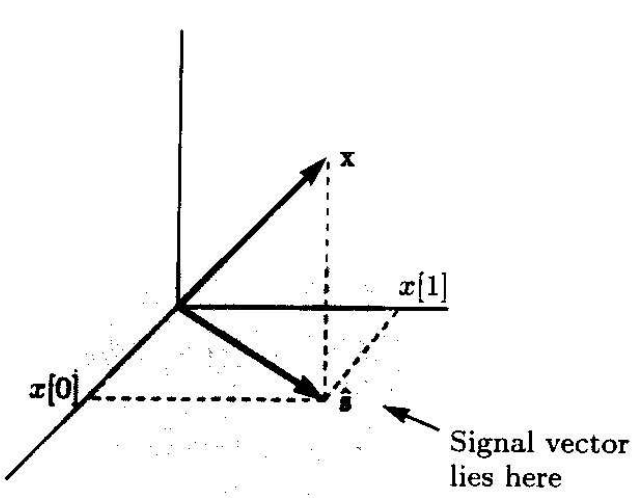

σ_n^2, $\mathbf{h}[n]$, $\mathbf{\Sigma}[n-1]$
$x[n]$
+
−
Σ
$\mathbf{K}[n]$
+
+
Σ
$\hat{\boldsymbol{\theta}}[n]$
z^{-1}
$\hat{\boldsymbol{\theta}}[n-1]$
$\mathbf{h}^T[n]$

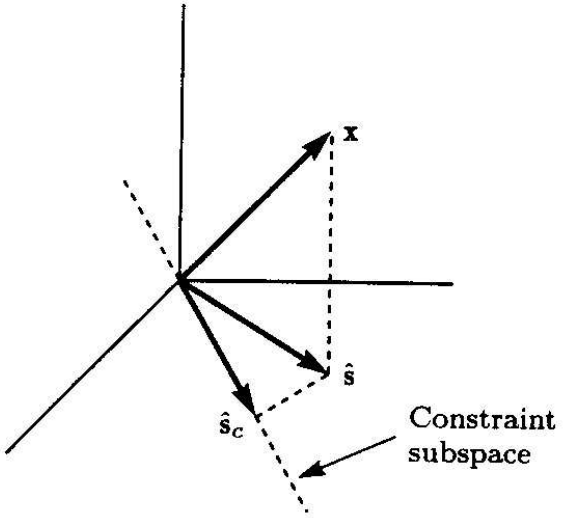
Hp) Stima vincolata ad una certa conoscenza priori sui dati

Il numero di parametri può essere ridotto e il problema affrontato con una tecnica iterativa nota come ***constrained least squares***

Nel caso di vincolo lineare - molto usuale nella pratica - questo approccio porta ad una soluzione in cui lo stimatore vincolato $\hat{\boldsymbol{\theta}}_c$ risulta una funzione lineare dello stimatore $\hat{\boldsymbol{\theta}}$ tramite le matrici $\mathbf{H}$ e la matrice $\mathbf{A}$ che esprime il vincolo

Interpretazione geometrica di questo tipo di soluzione: l'effetto delle combinazioni lineari prodotte dalle matrici $\mathbf{H}$ e $\mathbf{A}$ sullo stimatore $\hat{\boldsymbol{\theta}}$ sono delle rotazioni del vettore $\hat{\mathbf{s}}$, che in tal modo viene vincolato in un certo sottospazio identificato dal vincolo.


x
x[1]
x[0]
ŝ
Signal vector
lies here


x
ŝ
ŝc
Constraint
subspace

Problema LS non lineare

In generale non è possibile esprimere $\mathbf{s}(\boldsymbol{\theta})$ come una combinazione lineare $\mathbf{s} = \mathbf{H}\boldsymbol{\theta}$, ma è una funzione $N-$dimensionale non lineare di $\boldsymbol{\theta}$

Minimizzazione di J complicata, a volte impossibile

Problema di regressione non lineare

Minimizzazione basata su un approccio iterativo, ma se le dimensioni di $\boldsymbol{\theta}$ non superano $p = 5$, si preferisce spesso utilizzare dei criteri a griglia

Due possibilità per ridurre la complessità del problema:

1. trasformazioni di parametri

2. separazione dei parametri

1. si cerca una trasformazione biettiva di $\boldsymbol{\theta}$ che produca un modello lineare nel nuovo spazio

   Sia quindi

   $$\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta})$$

   in cui $\mathbf{g}$ è una funzione $p-$dimensionale di $\boldsymbol{\theta}$ invertibile. Se $\mathbf{g}$ è tale che

   $$\mathbf{s}(\boldsymbol{\theta}(\boldsymbol{\alpha})) = \mathbf{s}(\mathbf{g}^{-1}(\boldsymbol{\alpha})) = \mathbf{H}\boldsymbol{\alpha}$$

allora il modello sarà lineare in $\boldsymbol{\alpha}$.

Possiamo trovare l'LSE lineare di $\boldsymbol{\alpha}$ e dunque l'LSE non lineare di $\boldsymbol{\theta}$ da

$$\hat{\boldsymbol{\theta}} = \mathbf{g}^{-1}(\hat{\boldsymbol{\alpha}})$$

in cui

$$\hat{\boldsymbol{\alpha}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{g}^T\mathbf{x}.$$

2. Applicabile se il problema è separabile

Esempio - Digital Filter Design

Specifica sulla risposta impulsiva

$$\begin{aligned}\mathcal{H}(z) &= \frac{\mathcal{B}(z)}{\mathcal{A}(z)} \\ &= \frac{b[0] + b[1]z^{-1} + \ldots + b[q]z^{-q}}{a[0] + a[1]z^{-1} + \ldots + b[p]z^{-p}}.\end{aligned}$$

Se la risposta in frequenza desiderata è $H_d(f) = \mathcal{H}_d(\exp[j2\pi f])$, allora la risposta impulsiva è

$$h_d[n] = \mathcal{F}^{-1}\{H_d(f)\}$$

La risposta all'impulso è data da una equazione alle differenze ricorsiva, ottenuta dalla trasformata z inversa di $\mathcal{H}(z)$, nell'ipotesi $a[0] = 1$,

$$h[n] = \begin{cases} -\sum_{k=1}^{p} a[k]h[n-k] + \sum_{k=1}^{q} b[k]\delta[n-k] & n \geq 0 \\ 0 & n < 0. \end{cases}$$

La soluzione LS prevede che si scelgano degli $\{a[k], b[k]\}$ in modo tale da minimizzare

$$J = \sum (h_d[n] - h[n])^2$$

in cui N sia sufficientemente grande in modo da poter considerare $h[n]$ sostanzialmente nulla

Problema nonlineare

Ad esempio

$$\mathcal{H}(z) = \frac{b[0]}{1 + a[1]z^{-1}},$$

allora

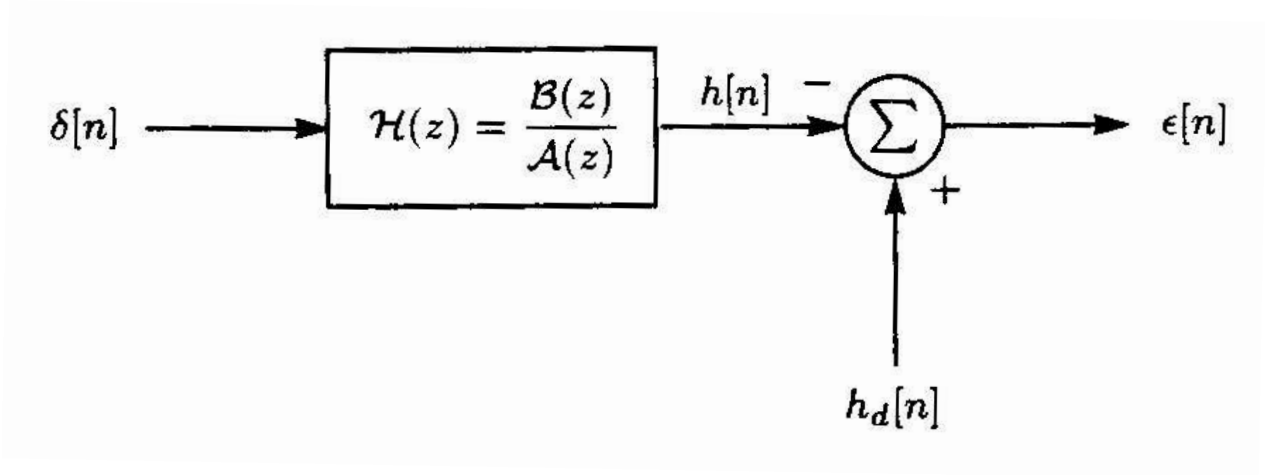
$$h[n] = \begin{cases} b[0](-a[1])^n & n \geq 0 \\ 0 & n < 0. \end{cases}$$

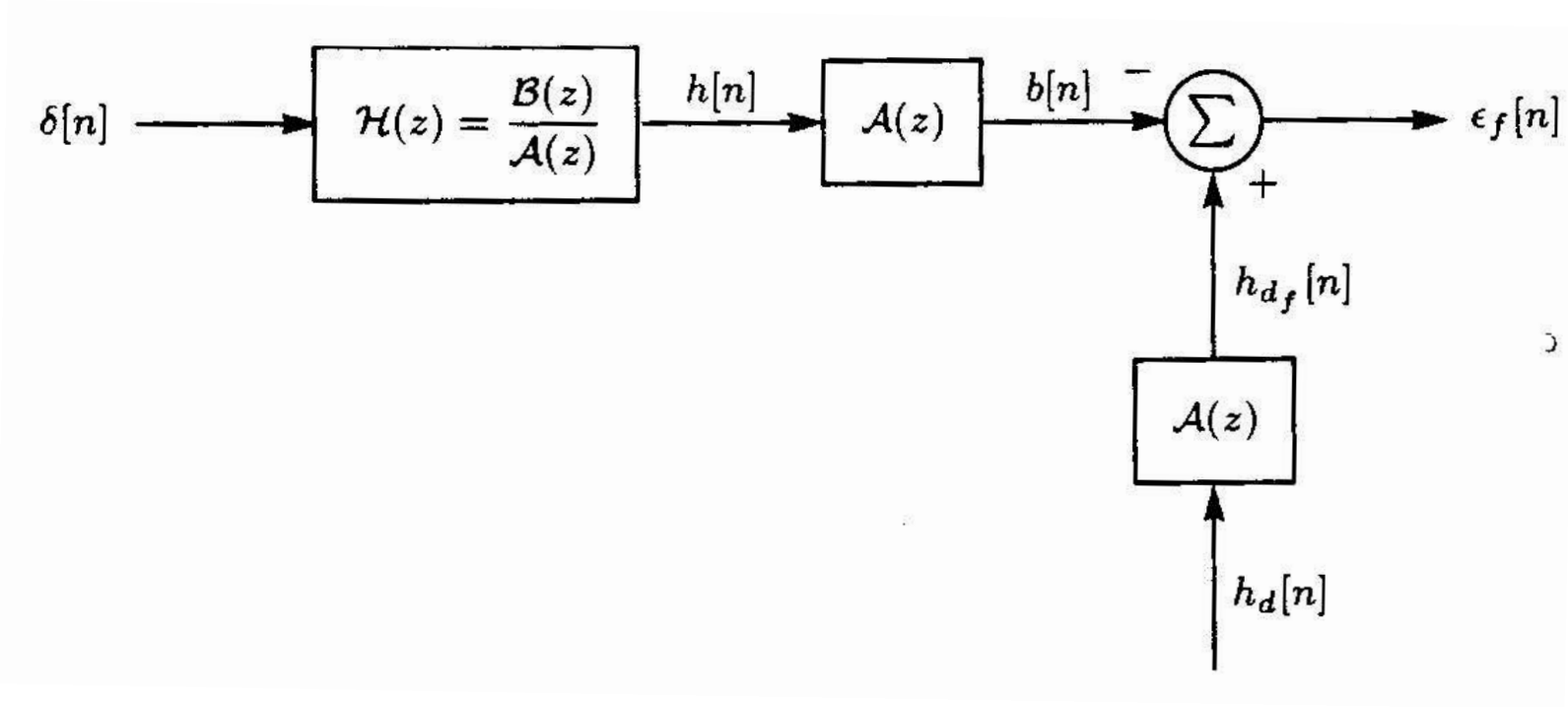
e

$$J = \sum (h_d[n] - b[0](-a[1])^n)^2,$$

che è non lineare in $a[1]$; infatti è proprio la presenza di $\mathcal{A}$ che fa in modo che l'errore LS sia non-quadratico

Per cercare di risolvere questo tipo di problema, possiamo filtrare $h_d[n]$ e $h[n]$ utilizzando il polinomio $\mathcal{A}$


$\delta[n]$
$\mathcal{H}(z) = \frac{\mathcal{B}(z)}{\mathcal{A}(z)}$
$h[n]$
$-$
$+$
$\epsilon[n]$
$h_d[n]$


$\delta[n]$
$\mathcal{H}(z) = \frac{\mathcal{B}(z)}{\mathcal{A}(z)}$
$h[n]$
$\mathcal{A}(z)$
$b[n]$
$-$
$+$
$\epsilon_f[n]$
$h_{d_f}[n]$
$\mathcal{A}(z)$
$h_d[n]$

A questo punto quindi possiamo minimizzare l'errore LS *filtrato*

$$J_f = \sum (h_{d_f}[n] - b[n])^2$$

in cui $h_{d_f}[n]$ è data da

$$h_{d_f}[n] = \sum_{k=0}^{p} a[k] h_d[n-k]$$

e $a[0] = 1$

Allora l'errore LS filtrato diventa

$$J_f = \sum \left(\sum_{k=0}^{p} a[k] h_d[n-k] - b[n] \right)^2,$$

che è una funzione quadratica degli $a[k]$ e $b[k]$

In alternativa

$$J_f = \sum \left[h_d[n] - \left(\sum_{k=1}^{p} a[k] h_d[n-k] - b[n] \right) \right]^2 .$$

Nel minimizzare la precedente espressione rispetto ai coefficienti del filtro, si noti che $b[n] = 0$ per $n > q$

Risultato

$$\begin{aligned} J_f(\mathbf{a}, \mathbf{b}) &= \sum_{n=0}^{q} \left[h_d[n] - \left(\sum_{k=1}^{p} a[k] h_d[n-k] - b[n] \right) \right]^2 \\ &+ \sum_{n=q+1}^{N-1} \left[h_d[n] - \left(\sum_{k=1}^{p} a[k] h_d[n-k] \right) \right]^2 \end{aligned}$$

La prima sommatoria può essere minimizzata (annullata) ponendo

$$h_d[n] + \sum_{k=1}^{p} a[k] h_d[n-k] = b[n].$$

Dunque in forma matriciale l'LSE dei coefficienti del numeratore è

$$\hat{\mathbf{b}} = \mathbf{h} + \mathbf{H}_0 \hat{\mathbf{a}}$$

in cui

$$\hat{\mathbf{h}} = [\ h_d[0]\ h_d[0]\ \ldots\ h_d[q]\ ]^T$$

$$\mathbf{H}_0 = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ h_d[0] & 0 & \cdots & 0 \\ h_d[1] & h_d[0] & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ h_d[q-1] & h_d[q-2] & \cdots & h_d[q-p] \end{bmatrix}.$$

Le dimensioni del vettore $\mathbf{H}$ sono $(q+1) \times 1$, mentre la matrice $\mathbf{H}_0$ ha le dimensioni $(q+1) \times p$. Per trovare l'LSE dei coefficienti del denominatore dobbiamo minimizzare

$$\begin{aligned} J(\mathbf{a}, \hat{\mathbf{b}}) &= \sum_{n=q+1}^{N-1} \left[h_d[n] - \left(\sum_{k=1}^{p} a[k] h_d[n-k] \right) \right]^2 \\ &= (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \end{aligned}$$

in cui $\mathbf{a} = \boldsymbol{\theta}$,

$$\mathbf{x} = [\, h_d[q+1]\, h_d[q-1]\, \ldots\, h_d[q-p+1]\,]^T$$

e

$$\mathbf{H}_0 = \begin{bmatrix} h_d[q] & h_d[q-1] & \cdots & h_d[q-p+1] \\ h_d[q+1] & h_d[q] & \cdots & h_d[q-p+2] \\ \vdots & \vdots & \vdots & \vdots \\ h_d[N-2] & h_d[N-3] & \cdots & h_d[N-1-p] \end{bmatrix}.$$

L'LSE per questo metodo, comunemente chiamato metodo di *Prony*, si trova come

$$\hat{\mathbf{a}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}.$$

Esempio - Stima AR di parametri per un modello ARMA

Andiamo adesso a descrivere un metodo per la stima dei parametri AR di un modello ARMA, autoregressive moving average. Un modello ARMA per un processo casuale WSS (Wide Sense Stationary) assume una distribuzione di potenza

$$P_{xx}(f) = \frac{\sigma_u^2 |B(f)|^2}{|A(f)|^2}$$

in cui

$$B(f) = 1 + \sum_{k=1}^{q} b[k] \exp(-j2\pi f k)$$

$$A(f) = 1 + \sum_{k=1}^{p} a[k] \exp(-j2\pi f k).$$

I $b[k]$ vengono chiamati *coefficienti MA*, mentre gli $a[k]$ vengono chiamati *coefficienti AR*. Il processo è ottenuto eccitando un filtro causale la cui risposta in frequenza sia $B(f)/A(f)$, con un rumore bianco di varianza σ_u^2. Estendiamo la PSD al piano z e calcoliamo l'inversa della trasformata z:

$$\mathcal{P}_{xx}(z) = \frac{\sigma_u^2 \mathcal{B}(z)\mathcal{B}(z^{-1})}{\mathcal{A}(z)\mathcal{A}(z^{-1})}$$

La PSD ovviamente può essere ottenuta valutando $\mathcal{P}$ sul circolo unitario del piano z.

Effettuiamo la trasformata inversa della $\mathcal{A}(z)\mathcal{P}_{xx}(z)$:

$$\mathcal{Z}^{-1}\{\mathcal{A}(z)\mathcal{P}_{xx}(z)\} = \mathcal{Z}^{-1}\left\{\sigma^2 \mathcal{B}(z)\frac{\mathcal{B}(z^{-1})}{\mathcal{A}(z^{-1})}\right\}.$$

Essendo il filtro causale, per $n < 0$

$$h[n] = \mathcal{Z}^{-1} \frac{\mathcal{B}(z)}{\mathcal{A}(z)} = 0$$

e per $n > 0$

$$h[-n] = \mathcal{Z}^{-1} \frac{\mathcal{B}(z^{-1})}{\mathcal{A}(z^{-1})} = 0.$$

Allora

$$\begin{aligned} \mathcal{Z}^{-1} \left\{ \sigma^2 \mathcal{B}(z) \frac{\mathcal{B}(z^{-1})}{\mathcal{A}(z^{-1})} \right\} &= \sigma^2 b[n] \star h[-n] \\ &= 0 \quad \text{per } n > q. \end{aligned}$$

Continuando abbiamo

$$\mathcal{Z}^{-1} \left\{ \mathcal{A}(z) \mathcal{P}_{xx}(z) \right\} = \mathcal{Z}^{-1} \left\{ \sigma^2 \mathcal{B}(z) \frac{\mathcal{B}(z^{-1})}{\mathcal{A}(z^{-1})} \right\}$$

$$= \quad 0 \quad \text{per } n > q.$$

In definitiva, possiamo scrivere una equazione alle differenze come

$$\sum_{k=0}^{p} a[k] r_{xx}[n-k] = 0 \quad \text{per } n > q,$$

in cui $a[0] = 1$. Questa equazione è denominata *equazione di Yule-Walker modificata*, ed è identica all'equazione di Yule-Walker ad eccezione del fatto che in quel caso $q = 0$. In pratica, viene stimato

$$\hat{r}_{xx}[k] = \frac{1}{N} \sum_{n=0}^{N-1-|k|} x[n] x[n+|k|],$$

assumendo di avere i valori reali di $x[n]$ per $n = 0, 1, \ldots, N-1$. Sostituendo l'ultima espressione

$$\sum_{k=0}^{p} a[k]\hat{r}_{xx}[n-k] = \epsilon[n], \quad n > q$$

in cui $\epsilon[n]$ indica l'errore dovuto all'effetto dell'errore di stima. Il modello diventa

$$\hat{r}_{xx}[n] = -\sum_{k=1}^{p} a[k]\hat{r}_{xx}[n-k] + \epsilon[n] \quad n > q$$

che può essere interpretata come una funzione lineare dei coefficienti AR. L'LSE per la stima degli a[k] minimizza

$$J = \sum_{n=q+1}^{M} \left[\hat{r}_{xx}[n] - \left(-\sum_{k=1}^{p} a[k]\hat{r}_{xx}[n-k]\right)\right]^2$$

$$= \quad (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})$$

con

$$\mathbf{x} = \begin{bmatrix} \hat{r}_{xx}[q+1] \\ \hat{r}_{xx}[q+2] \\ \vdots \\ \hat{r}_{xx}[M] \end{bmatrix}$$

$$\boldsymbol{\theta} = \begin{bmatrix} a[1] \\ a[2] \\ \vdots \\ a[p] \end{bmatrix}$$

$$\mathbf{H} = \begin{bmatrix} \hat{r}_{xx}[q] & \hat{r}_{xx}[q-1] & \dots & \hat{r}_{xx}[q-p+1] \\ \hat{r}_{xx}[q+1] & \hat{r}_{xx}[q] & \dots & \hat{r}_{xx}[q-p+2] \\ \vdots & \vdots & \vdots & \vdots \\ \hat{r}_{xx}[M-1] & \hat{r}_{xx}[M-2] & \dots & \hat{r}_{xx}[M-p] \end{bmatrix}.$$

L'LSE si ottiene come $\boldsymbol{\theta} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$, detta *equazione di Yule-Walker modificata ai minimi quadrati*.