

TEORIA DELLA STIMA

Michele TARAGNA

Dipartimento di Automatica e Informatica

Politecnico di Torino

`michele.taragna@polito.it`



Corso di Laurea Specialistica in Ingegneria Informatica

Modulo di “**Analisi, Stima e Previsione di Dati**”

Problema della stima

Il problema della stima riguarda la valutazione empirica di una grandezza incerta, come di un parametro caratteristico ignoto o di un segnale remoto, sulla base di osservazioni e misure sperimentali rilevate nello studio del fenomeno.

Un problema di stima presuppone sempre un'adeguata descrizione matematica (*modello*) del fenomeno:

- nella statistica classica, di solito si affrontano problemi relativi a modelli statici, caratterizzati da legami istantanei (o algebrici) fra le grandezze in gioco;
- in questo corso, si presentano metodi di stima per fenomeni che sono descritti adeguatamente da *modelli dinamici a tempo discreto*, caratterizzati da legami fra le variabili in gioco rappresentabili nella loro evoluzione temporale mediante equazioni alle differenze (per semplicità si assume cioè che il tempo sia discreto).

Problema della stima

$\vartheta(t)$: grandezza *reale* da stimare, scalare o vettoriale, costante o variabile nel tempo;

$d(t)$: dati disponibili, rilevati a certi N istanti di tempo t_1, t_2, \dots, t_N ;

$T = \{t_1, t_2, \dots, t_N\}$: insieme degli istanti di osservazione, distribuiti con cadenza regolare (in tal caso, $T = \{1, 2, \dots, N\}$) o non uniforme;

$d = \{d(t_1), d(t_2), \dots, d(t_N)\}$: insieme delle osservazioni.

Uno **stimatore** (o **algoritmo di stima**) è una *funzione* $f(\cdot)$ che associa ai dati un valore della grandezza da stimare:

$$\vartheta(t) = f(d)$$

Per **stima** si intende il particolare *valore* assunto dallo stimatore in corrispondenza ai particolari dati osservati.

Classificazione dei problemi di stima

- 1) $\vartheta(t)$ è costante nel tempo \Rightarrow problema di **identificazione parametrica**:
 - lo stimatore si indica con $\hat{\vartheta}$ oppure con $\hat{\vartheta}_T$;
 - il valore vero della grandezza incognita si indica con ϑ^o ;
- 2) $\vartheta(t)$ è funzione del tempo:
 - lo stimatore si indica con $\hat{\vartheta}(t|T)$, oppure con $\hat{\vartheta}(t|N)$ se gli istanti di osservazione sono distribuiti con cadenza regolare;
 - a seconda della posizione di t rispetto all'ultimo istante di osservazione t_N :
 - 2.a) se $t > t_N \Rightarrow$ problema di **predizione**;
 - 2.b) se $t = t_N \Rightarrow$ problema di **filtraggio**;
 - 2.c) se $t_1 < t < t_N \Rightarrow$ problema di **regolarizzazione** (o di **interpolazione** o di **smoothing**).

Esempio di problema di predizione: analisi delle serie temporali

Data una sequenza di osservazioni (serie temporale o storica) di una variabile y :

$$y(1), y(2), \dots, y(t)$$

si vuole valutare il valore successivo $y(t + 1)$ della stessa variabile



occorre trovare un buon **predittore** $\hat{y}(t + 1|t)$, cioè una funzione dei dati disponibili che fornisca una valutazione il più possibile accurata del valore successivo:

$$\hat{y}(t + 1|t) = f(y(t), y(t - 1), \dots, y(1)) \cong y(t + 1)$$

Un predittore si dice *lineare* se è una funzione lineare dei dati:

$$\hat{y}(t + 1|t) = a_1(t)y(t) + a_2(t)y(t - 1) + \dots + a_t(t)y(1) = \sum_{k=1}^t a_k(t)y(t - k + 1)$$

Un predittore lineare è *a memoria finita* n se è funzione lineare solo degli ultimi n dati:

$$\hat{y}(t+1|t) = a_1(t)y(t) + a_2(t)y(t-1) + \dots + a_n(t)y(t-n+1) = \sum_{k=1}^n a_k(t)y(t-k+1)$$

Se i parametri $a_i(t)$ sono costanti nel tempo, il predittore è anche *tempo-invariante*:

$$\hat{y}(t+1|t) = a_1y(t) + a_2y(t-1) + \dots + a_ny(t-n+1) = \sum_{k=1}^n a_ky(t-k+1)$$

ed è individuato dal vettore di parametri costanti nel tempo

$$\vartheta = [a_1 \quad a_2 \quad \dots \quad a_n]^T \in \mathbb{R}^n$$



il problema di predizione è ricondotto ad un problema di identificazione parametrica.

Problemi:

- come valutare la bontà del predittore?
- come determinare il predittore migliore?

Se il modello predittivo del fenomeno è lineare, tempo-invariante, a memoria finita n molto minore del numero di dati misurati fino all'istante t , è possibile valutarne la capacità predittiva sui dati già noti $y(i)$, $i = 1, 2, \dots, t$, nel modo seguente:

- ad ogni istante $i \geq n$, si calcola la predizione $\hat{y}(i+1|i)$ del valore successivo:
$$\hat{y}(i+1|i) = a_1 y(i) + a_2 y(i-1) + \dots + a_n y(i-n+1) = \sum_{k=1}^n a_k y(i-k+1)$$

e se ne valuta l'errore di predizione $\varepsilon(i+1)$ rispetto al dato $y(i+1)$:

$$\varepsilon(i+1) = y(i+1) - \hat{y}(i+1|i)$$

- il modello descritto da ϑ è un buon modello predittivo quando l'errore ε risulta "piccolo" su tutto l'arco dei dati disponibili \Rightarrow si introduce la cifra di merito:

$$J(\vartheta) = \sum_{k=n+1}^t \varepsilon(k)^2$$

- il miglior predittore è quello che minimizza J e che ha come valore dei parametri:

$$\vartheta^* = \arg \min_{\vartheta \in \mathbb{R}^n} J(\vartheta)$$

Questione fondamentale: il predittore che minimizza J è per forza un “buon” modello?

La bontà del predittore è legata al fatto che l’andamento temporale della sequenza degli errori di predizione $\varepsilon(\cdot)$ presenti le seguenti caratteristiche:

- deve essere a valor medio nullo, cioè non deve presentare un errore sistematico;
- deve essere “del tutto casuale”, cioè non deve contenere elementi di regolarità.

In termini probabilistici, il tutto equivale a richiedere che l’andamento dell’errore $\varepsilon(\cdot)$ sia quello di un **rumore bianco** (White Noise, WN), ossia di una sequenza di variabili casuali indipendenti con valor medio nullo e varianza costante σ^2 :

$$\varepsilon(\cdot) = WN(0, \sigma^2)$$



Un predittore è un “buon” modello se $\varepsilon(\cdot)$ ha le caratteristiche probabilistiche di un rumore bianco.

Il problema di predizione si può ricondurre così allo studio di un **sistema stocastico**, cioè di un sistema dinamico alimentato da segnali descritti probabilisticamente; infatti:

$$\begin{cases} \hat{y}(t|t-1) = a_1 y(t-1) + a_2 y(t-2) + \dots + a_n y(t-n) \\ \varepsilon(t) = y(t) - \hat{y}(t|t-1) \end{cases} \Rightarrow$$

$$y(t) = \hat{y}(t|t-1) + \varepsilon(t) = a_1 y(t-1) + a_2 y(t-2) + \dots + a_n y(t-n) + \varepsilon(t)$$

rappresenta un sistema dinamico a tempo discreto lineare tempo-invariante con uscita $y(t)$ e ingresso $\varepsilon(t)$

⇓

\mathcal{Z} -trasformando e ricordando che $\mathcal{Z}[y(t-k)] = z^{-k}Y(z)$, essendo z^{-1} l'operatore di ritardo unitario:

$$Y(z) = a_1 z^{-1}Y(z) + a_2 z^{-2}Y(z) + \dots + a_n z^{-n}Y(z) + \varepsilon(z)$$

⇓

$$G(z) = \frac{Y(z)}{\varepsilon(z)} = \frac{1}{1 - a_1 z^{-1} - a_2 z^{-2} - \dots - a_n z^{-n}} = \frac{z^n}{z^n - a_1 z^{n-1} - a_2 z^{n-2} - \dots - a_n}$$

rappresenta la funzione di trasferimento del sistema dinamico LTI \Rightarrow per essere un "buon" modello, il suo ingresso $\varepsilon(\cdot)$ deve possedere le caratteristiche probabilistiche di un rumore bianco.

Tipologie di descrizione dei dati

- Le informazioni effettivamente disponibili sono sempre:
 - limitate $\Rightarrow N$ è necessariamente finito;
 - affette da incertezza di diversa natura (es. rumore di misura).
- L'incertezza presente nei dati può essere descritta:
 - in termini probabilistici \Rightarrow si parla di **stima statistica** o **classica**;
 - in termini insiemistici, come appartenente ad un certo insieme (*set*) limitato \Rightarrow si parla di **stima Set Membership** o **Unknown But Bounded (UBB)**.

Esperimento casuale e sorgente casuale di dati

S : **spazio degli esiti**, cioè l'insieme dei possibili esiti s dell'esperimento casuale;

\mathcal{F} : **spazio degli eventi di interesse**, cioè l'insieme delle combinazioni di interesse in cui gli esiti in S possono essere raggruppati;

$P(\cdot)$: funzione **probabilità** definita in \mathcal{F} che associa ad ogni evento in \mathcal{F} un numero reale compreso fra 0 e 1.

$\mathcal{E} = (S, \mathcal{F}, P(\cdot))$: **esperimento casuale**

Esempio: lancio un dado a sei facce e guardo se esce un numero pari o dispari \Rightarrow

- $S = \{1, 2, 3, 4, 5, 6\}$ è l'insieme delle 6 facce del dado;
- $\mathcal{F} = \{A, B, S, \emptyset\}$, essendo $A = \{2, 4, 6\}$ e $B = \{1, 3, 5\}$ gli eventi di interesse, ossia gli insiemi dei numeri pari e dispari;
- $P(A) = P(B) = 1/2$ (se il dado non è truccato), $P(S) = 1$, $P(\emptyset) = 0$.

Una **variabile casuale** (o aleatoria) sull'esperimento \mathcal{E} è una variabile v i cui valori dipendono dall'esito s di \mathcal{E} attraverso un'opportuna funzione $\varphi(\cdot) : S \rightarrow V$, essendo V l'insieme dei valori assunti da v :

$$v = \varphi(s)$$

Esempio: si può definire la variabile casuale dipendente dall'esito del lancio del dado a sei facce

$$v = \varphi(s) = \begin{cases} +1 & \text{se } s \in A = \{2, 4, 6\} \\ -1 & \text{se } s \in B = \{1, 3, 5\} \end{cases}$$

Una **sorgente casuale di dati** genera dati che, oltre che dal processo in esame caratterizzato dal valore vero incognito ϑ^o della grandezza da stimare, sono anche funzioni di una variabile casuale; in particolare, all'istante t , il dato $d(t)$ dipende dalla variabile aleatoria $v(t)$.

Descrizione probabilistica dei dati

Nel contesto *probabilistico* (o *classico* o *statistico*), si assume che i dati d siano generati da una sorgente casuale di dati \mathcal{S} , influenzata:

- dall'esito s di un esperimento casuale \mathcal{E}
- dal valore "vero" ϑ^o della grandezza incognita da stimare

$$d = d(s, \vartheta^o)$$



i dati d sono variabili casuali, in quanto funzioni dell'esito s



una descrizione probabilistica completa dei dati è costituita dalla conoscenza della sua

- **distribuzione di probabilità** $F(q) = Prob \{d(s, \vartheta^o) \leq q\}$ oppure della sua

- **densità di probabilità** $f(q) = \frac{dF(q)}{dq}$

Caratteristiche degli stimatori

Una sorgente casuale di dati \mathcal{S} , influenzata dall'esito s di un esperimento casuale \mathcal{E} e dal valore "vero" ϑ^o della grandezza incognita da stimare, genera i dati d :

$$d = d(s, \vartheta^o)$$



i dati d sono variabili casuali, in quanto funzioni dell'esito s



anche lo stimatore $f(\cdot)$ e la stima $\hat{\vartheta}$ sono variabili casuali, in quanto funzioni di d :

$$\hat{\vartheta} = f(d) = f(d(s, \vartheta^o))$$



la bontà di $f(\cdot)$ e di $\hat{\vartheta}$ dipendono dalle loro caratteristiche probabilistiche.

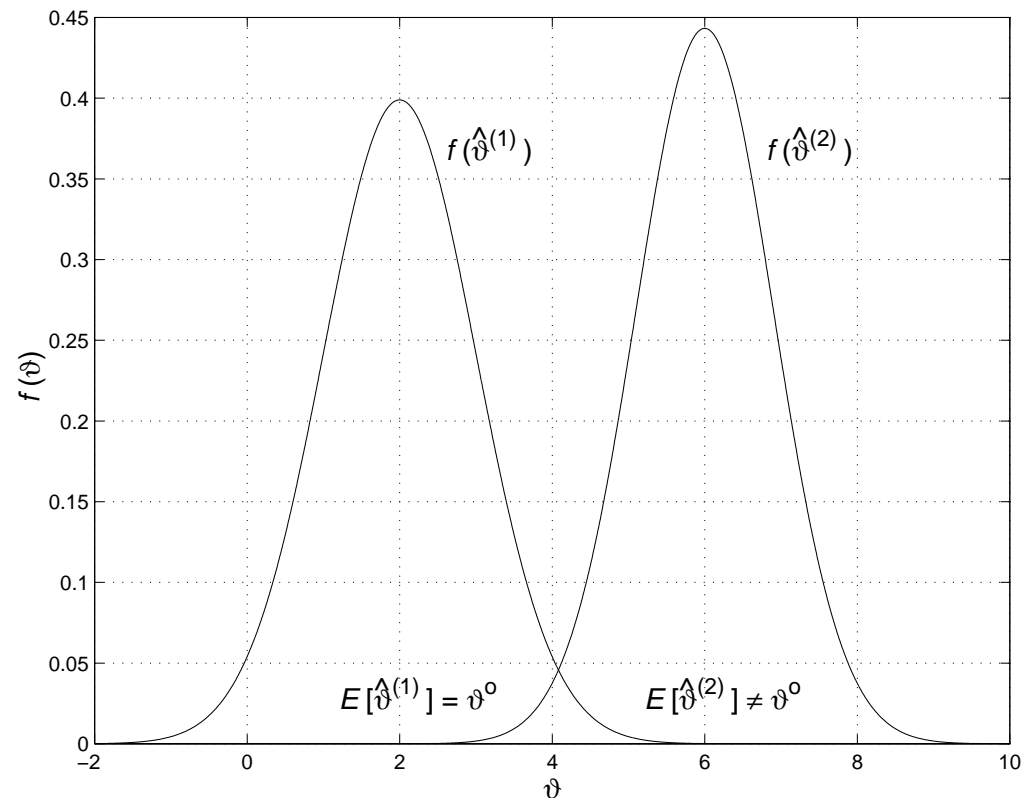
Caratteristiche probabilistiche degli stimatori

- Non polarizzazione (per evitare di introdurre un errore sistematico di stima)
- Minima varianza (una minore dispersione intorno al valore medio garantisce una probabilità più elevata di ottenere valori vicini al valore “vero” ϑ^0)
- Caratteristiche asintotiche (per $N \rightarrow \infty$):
 - convergenza in media quadratica
 - convergenza quasi-certa
 - consistenza

Caratteristiche probabilistiche degli stimatori

Uno stimatore si dice **corretto** (o **non polarizzato**, o **non deviato**) se

$$E[\hat{\vartheta}] = \vartheta^o$$

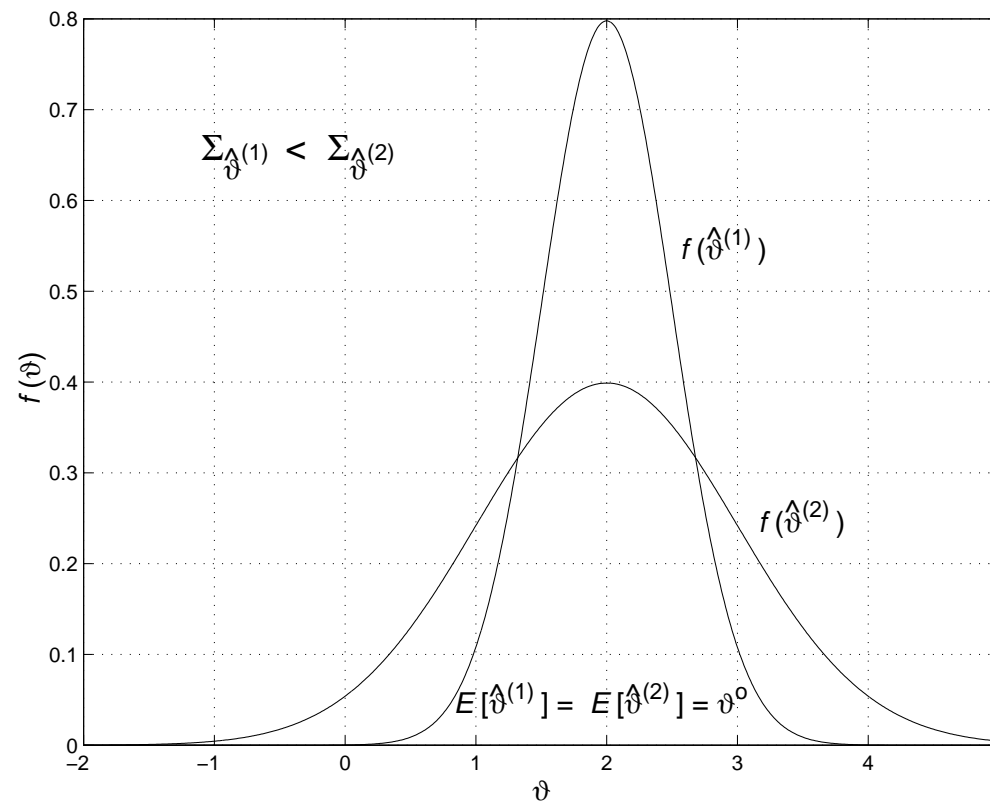


Uno stimatore corretto non introduce cioè un errore sistematico di stima.

Caratteristiche probabilistiche degli stimatori

Uno stimatore corretto $\hat{\vartheta}^{(1)}$ è a **minima varianza** (o **efficiente**) se

$$\text{Var}[\hat{\vartheta}^{(1)}] \leq \text{Var}[\hat{\vartheta}^{(2)}], \quad \forall \hat{\vartheta}^{(2)} \neq \hat{\vartheta}^{(1)}$$



Minore dispersione intorno al valore medio \Rightarrow maggiore probabilità di avvicinare ϑ^0 .

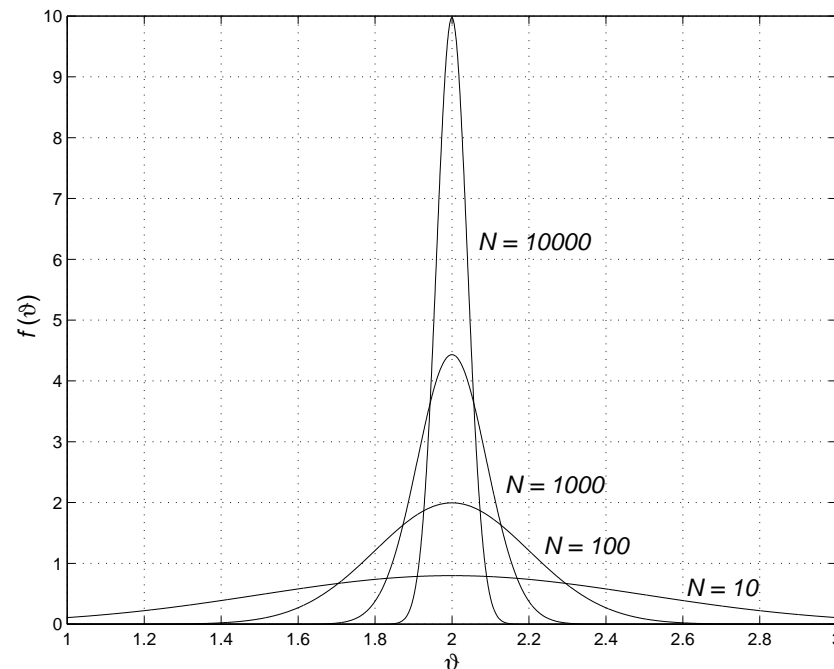
Caratteristiche asintotiche degli stimatori

Uno stimatore corretto **converge in media quadratica a ϑ^o** , cioè $\lim_{N \rightarrow \infty} \hat{\vartheta}_N = \vartheta^o$, se

$$\lim_{N \rightarrow \infty} E \left[\|\hat{\vartheta}_N - \vartheta^o\|^2 \right] = 0$$

dove $\|x\| = \sqrt{\sum_{i=1}^n x_i^2}$, $\forall x \in \mathbb{R}^n$, è la norma euclidea.

Uno stimatore corretto tale che $\lim_{N \rightarrow \infty} Var \left[\hat{\vartheta}_N \right] = 0$ converge in media quadratica:



Convergenza certa e quasi-certa, consistenza

Uno stimatore è funzione dell'esito s di un esperimento casuale, oltre che di ϑ^o :

$$\hat{\vartheta} = f(d) = f(d(s, \vartheta^o)) \Rightarrow \hat{\vartheta} = \hat{\vartheta}(s, \vartheta^o)$$

Se si considera un particolare esito $\bar{s} \in S$ e si valuta la sequenza delle stime $\hat{\vartheta}_N(\bar{s}, \vartheta^o)$ al crescere di N , si ottiene una successione numerica $\hat{\vartheta}_1(\bar{s}, \vartheta^o), \hat{\vartheta}_2(\bar{s}, \vartheta^o), \dots$, che può convergere a ϑ^o per qualche \bar{s} , e non convergere per altri \bar{s} .

Sia A l'insieme degli esiti \bar{s} per cui si ha la convergenza a ϑ^o :

- se $A \equiv S$, si parla di **convergenza certa**, avendo luogo $\forall \bar{s} \in S$;
- se $A \subset S$, considerando A come evento si può definire la probabilità $P(A)$;
se A è tale che $P(A) = 1$, si dice che $\hat{\vartheta}_N$ converge a ϑ^o *con probabilità 1*:

$$\lim_{N \rightarrow \infty} \hat{\vartheta}_N = \vartheta^o \quad q.c.$$

e si parla di **convergenza quasi-certa** (*q.c.*) \Rightarrow l'algoritmo è detto **consistente**.

Esempio

Problema: N dati scalari d_i con lo stesso valore medio $E [d_i] = \vartheta^o$, con varianze $Var [d_i]$ eventualmente diverse ma limitate ($\exists \sigma \in \mathbb{R}_+ : Var [d_i] \leq \sigma^2 < \infty, \forall i$); i dati sono tra loro incorrelati, cioè:

$$E [\{d_i - E [d_i]\} \{d_j - E [d_j]\}] = 0, \quad \forall i \neq j$$

Stimatore #1 a media campionaria:

$$\hat{\vartheta}_N = \frac{1}{N} \sum_{i=1}^N d_i$$

- è uno stimatore corretto:

$$E [\hat{\vartheta}_N] = E \left[\frac{1}{N} \sum_{i=1}^N d_i \right] = \frac{1}{N} \sum_{i=1}^N E [d_i] = \frac{1}{N} \sum_{i=1}^N \vartheta^o = \vartheta^o$$

- converge in media quadratica:

$$\begin{aligned} \text{Var} [\hat{\vartheta}_N] &= E \left[\left(\hat{\vartheta}_N - E [\hat{\vartheta}_N] \right)^2 \right] = E \left[\left(\frac{1}{N} \sum_{i=1}^N d_i - \vartheta^o \right)^2 \right] = \\ &= E \left[\left(\frac{1}{N} \sum_{i=1}^N d_i - \frac{1}{N} \sum_{i=1}^N \vartheta^o \right)^2 \right] = E \left[\left(\frac{1}{N} \sum_{i=1}^N (d_i - \vartheta^o) \right)^2 \right] = \\ &= E \left[\frac{1}{N^2} \left(\sum_{i=1}^N (d_i - \vartheta^o) \right)^2 \right] = \frac{1}{N^2} E \left[\left(\sum_{i=1}^N (d_i - \vartheta^o) \right)^2 \right] = \\ &= \frac{1}{N^2} E \left[\sum_{i=1}^N (d_i - \vartheta^o)^2 + \sum_{i=1}^N (d_i - \vartheta^o) \sum_{j=1, j \neq i}^N (d_j - \vartheta^o) \right] = \\ &= \frac{1}{N^2} \left\{ \sum_{i=1}^N E \left[(d_i - \vartheta^o)^2 \right] + \sum_{i=1}^N E \left[(d_i - \vartheta^o) \sum_{j=1, j \neq i}^N (d_j - \vartheta^o) \right] \right\} = \\ &= \frac{1}{N^2} \sum_{i=1}^N \text{Var} [d_i] \leq \frac{1}{N^2} \sum_{i=1}^N \sigma^2 = \sigma^2 / N \end{aligned}$$

$$\Downarrow$$

$$\lim_{N \rightarrow \infty} \text{Var} [\hat{\vartheta}_N] \leq \lim_{N \rightarrow \infty} \frac{\sigma^2}{N} = 0$$

$$\Downarrow$$

l'algoritmo converge in media quadratica, essendo corretto e con $\lim_{N \rightarrow \infty} \text{Var} [\hat{\vartheta}_N] = 0$.

Stimatore #2:

$$\hat{\vartheta}_N = d_j$$

- è uno stimatore corretto:

$$E [\hat{\vartheta}_N] = E [d_j] = \vartheta^o$$

- non converge in media quadratica:

$$\text{Var} [\hat{\vartheta}_N] = E \left[\left(\hat{\vartheta}_N - E [\hat{\vartheta}_N] \right)^2 \right] = E \left[(d_j - \vartheta^o)^2 \right] = \text{Var} [d_j] \leq \sigma^2$$

e quindi non varia col numero N di dati



l'incertezza nella stima è costante, non diminuisce al crescere del numero di dati.

Stimatore #3 a media campionaria pesata:

$$\hat{\vartheta}_N = \sum_{i=1}^N \alpha_i d_i$$

- è uno stimatore corretto se e solo se $\sum_{i=1}^N \alpha_i = 1$, poiché

$$E[\hat{\vartheta}_N] = E\left[\sum_{i=1}^N \alpha_i d_i\right] = \sum_{i=1}^N \alpha_i E[d_i] = \vartheta^o \sum_{i=1}^N \alpha_i = \vartheta^o \Leftrightarrow \sum_{i=1}^N \alpha_i = 1$$

Nota: l'algoritmo #1 corrisponde al caso $\alpha_i = \frac{1}{N}$, $\forall i$;

l'algoritmo #2 corrisponde al caso $\alpha_j = 1$ e $\alpha_i = 0$, $\forall i \neq j$

- si dimostra che lo stimatore corretto a minima varianza è quello avente pesi

$$\alpha_i = \frac{\alpha}{Var[d_i]}, \quad \alpha = \left[\sum_{i=1}^N \frac{1}{Var[d_i]} \right]^{-1}$$

intuitivamente, i dati più incerti sono ritenuti meno affidabili \Rightarrow hanno peso minore

- la varianza dello stimatore corretto a minima varianza vale

$$\begin{aligned} \text{Var} [\hat{\vartheta}_N] &= E \left[\left(\hat{\vartheta}_N - E [\hat{\vartheta}_N] \right)^2 \right] = E \left[\left(\sum_{i=1}^N \alpha_i d_i - \vartheta^o \right)^2 \right] = \\ &= E \left[\left(\sum_{i=1}^N \alpha_i d_i - \sum_{i=1}^N \alpha_i \vartheta^o \right)^2 \right] = E \left[\left(\sum_{i=1}^N \alpha_i (d_i - \vartheta^o) \right)^2 \right] = \\ &= E \left[\sum_{i=1}^N \alpha_i^2 (d_i - \vartheta^o)^2 + \sum_{i=1}^N \alpha_i (d_i - \vartheta^o) \sum_{j=1, j \neq i}^N \alpha_j (d_j - \vartheta^o) \right] = \\ &= \sum_{i=1}^N \alpha_i^2 E \left[(d_i - \vartheta^o)^2 \right] + \sum_{i=1}^N \alpha_i E \left[(d_i - \vartheta^o) \sum_{j=1, j \neq i}^N \alpha_j (d_j - \vartheta^o) \right] = \\ &= \sum_{i=1}^N \alpha_i^2 \text{Var} [d_i] = \sum_{i=1}^N \frac{\alpha^2}{\text{Var}[d_i]^2} \text{Var}[d_i] = \alpha^2 \sum_{i=1}^N \frac{1}{\text{Var}[d_i]} = \\ &= \alpha = \left[\sum_{i=1}^N \frac{1}{\text{Var}[d_i]} \right]^{-1} \leq \left[\sum_{i=1}^N \frac{1}{\sigma^2} \right]^{-1} = \frac{\sigma^2}{N} \end{aligned}$$

- l'algoritmo corretto a minima varianza converge in media quadratica, poiché

$$\lim_{N \rightarrow \infty} \text{Var} [\hat{\vartheta}_N] \leq \lim_{N \rightarrow \infty} \frac{\sigma^2}{N} = 0$$

Disuguaglianza di Cramér-Rao

La precisione della stima ha limiti intrinseci, dovuti solo alla sorgente casuale di dati: la varianza di un qualsiasi stimatore non può infatti scendere sotto un certo valore, poiché i dati sono soggetti a disturbi e la corrispondente incertezza si riflette in una incertezza strutturale nella stima, non eliminabile cambiando il tipo di stimatore:

- nel caso scalare $\vartheta \in \mathbb{R}$, la **disuguaglianza di Cramér-Rao** afferma che, per ogni stimatore corretto $\hat{\vartheta}$,

$$\text{Var} \left[\hat{\vartheta} \right] \geq m^{-1}$$

essendo m la **quantità di informazione di Fisher** definita come

$$m = E \left[\left\{ \frac{\partial}{\partial \vartheta} \ln f(d^{(\vartheta)}, \vartheta) \right\}^2 \right]_{\vartheta=\vartheta^o} = -E \left[\frac{\partial^2}{\partial \vartheta^2} \ln f(d^{(\vartheta)}, \vartheta) \right]_{\vartheta=\vartheta^o} \geq 0$$

$d^{(\vartheta)} \in \mathbb{R}^N$ sono i dati generati dalla sorgente casuale di dati per un generico valore ϑ del parametro ignoto, non necessariamente il valore “vero” ϑ^o ; $f(q, \vartheta)$, $q \in \mathbb{R}^N$, è la densità di probabilità di $d^{(\vartheta)}$;

- nel caso vettoriale $\vartheta \in \mathbb{R}^n$, la **disuguaglianza di Cramér-Rao** afferma che, per ogni stimatore corretto $\hat{\vartheta}$,

$$\text{Var} \left[\hat{\vartheta} \right] \geq M^{-1}$$

se M è invertibile, essendo M la **matrice di informazione di Fisher**

$$M = [m_{ij}] \in \mathbb{R}^{n \times n}$$

$$m_{ij} = -E \left[\frac{\partial^2}{\partial \vartheta_i \partial \vartheta_j} \ln f(d^{(\vartheta)}, \vartheta) \right]_{\vartheta = \vartheta^0}, \quad \forall i, j = 1, 2, \dots, n;$$

da tale disuguaglianza risulta quindi che

$$\text{Var} \left[\hat{\vartheta}_i \right] \geq [M^{-1}]_{ii}, \quad \forall i = 1, 2, \dots, n$$

Uno stimatore corretto è **efficiente** se è a minima varianza, cioè se la sua varianza raggiunge il minimo valore teorico stabilito dalla disuguaglianza di Cramér-Rao:

$$\text{Var} \left[\hat{\vartheta} \right] = m^{-1} \quad \text{oppure} \quad \text{Var} \left[\hat{\vartheta} \right] = M^{-1}$$

Metodo di stima ai Minimi Quadrati

Problema: date le misure di $n + 1$ variabili reali $y(t), u_1(t), \dots, u_n(t)$ su un certo arco temporale (ad esempio, per $t = 1, 2, \dots, N$), determinare se possibile i valori di n parametri reali $\vartheta_1, \vartheta_2, \dots, \vartheta_n$ per cui valga il legame (**regressione lineare**)

$$y(t) = \vartheta_1 u_1(t) + \dots + \vartheta_n u_n(t)$$

In termini matriciali, definendo i vettori

$$\vartheta = \begin{bmatrix} \vartheta_1 \\ \vdots \\ \vartheta_n \end{bmatrix} \in \mathbb{R}^n, \quad \varphi(t) = \begin{bmatrix} u_1(t) \\ \vdots \\ u_n(t) \end{bmatrix} \in \mathbb{R}^n \quad \Rightarrow \quad y(t) = \varphi(t)^T \vartheta$$

Nei problemi reali, tale legame non vale mai \Rightarrow c'è un errore $\varepsilon(t) = y(t) - \varphi(t)^T \vartheta$

\Downarrow

definendo $J(\vartheta) = \sum_{t=1}^N \varepsilon(t)^2$, il problema si risolve trovando $\vartheta^* = \arg \min_{\vartheta \in \mathbb{R}^n} J(\vartheta)$.

Per determinare il minimo della cifra di merito, occorre richiedere che

$$\frac{dJ(\vartheta)}{d\vartheta} = \left[\frac{dJ(\vartheta)}{d\vartheta_1} \quad \dots \quad \frac{dJ(\vartheta)}{d\vartheta_n} \right] = 0 \quad \Leftrightarrow$$

$$\frac{dJ(\vartheta)}{d\vartheta_i} = \frac{d}{d\vartheta_i} \left[\sum_{t=1}^N \varepsilon(t)^2 \right] = \sum_{t=1}^N \frac{d}{d\vartheta_i} \left[\varepsilon(t)^2 \right] = \sum_{t=1}^N \frac{d}{d\vartheta_i} \left[\left(y(t) - \varphi(t)^T \vartheta \right)^2 \right] =$$

$$= -2 \sum_{t=1}^N \left(y(t) - \varphi(t)^T \vartheta \right) u_i(t) = 0, \quad i = 1, 2, \dots, n \quad \Leftrightarrow$$

$$\frac{dJ(\vartheta)}{d\vartheta} = -2 \sum_{t=1}^N \left(y(t) - \varphi(t)^T \vartheta \right) \varphi(t)^T = 0 \quad \Leftrightarrow$$

$$\sum_{t=1}^N \left(y(t) \varphi(t)^T - \varphi(t)^T \vartheta \varphi(t)^T \right) = \sum_{t=1}^N y(t) \varphi(t)^T - \sum_{t=1}^N \varphi(t)^T \vartheta \varphi(t)^T = 0 \quad \Leftrightarrow$$

$$\sum_{t=1}^N \varphi(t)^T \vartheta \varphi(t)^T = \sum_{t=1}^N y(t) \varphi(t)^T \quad \Leftrightarrow$$

$$\sum_{t=1}^N \left[\varphi(t) \varphi(t)^T \right] \vartheta = \sum_{t=1}^N \varphi(t) y(t)$$

La relazione

$$\sum_{t=1}^N \left[\varphi(t) \varphi(t)^T \right] \vartheta = \sum_{t=1}^N \varphi(t) y(t)$$

costituisce un sistema di n equazioni scalari nelle n incognite scalari $\vartheta_1, \vartheta_2, \dots, \vartheta_n$

ed è detto sistema delle **equazioni normali**:

- se la matrice $\sum_{t=1}^N \varphi(t) \varphi(t)^T$ è invertibile ($\Leftrightarrow \det \sum_{t=1}^N \varphi(t) \varphi(t)^T \neq 0$, detta *condizione di identificabilità*), il sistema delle equazioni normali ammette come *unica* soluzione la **stima ai minimi quadrati (Least Squares, LS)**:

$$\hat{\vartheta} = \left[\sum_{t=1}^N \varphi(t) \varphi(t)^T \right]^{-1} \left[\sum_{t=1}^N \varphi(t) y(t) \right]$$

- se invece $\sum_{t=1}^N \varphi(t) \varphi(t)^T$ è singolare, si dimostra che le equazioni normali hanno infinite soluzioni, a causa della loro particolare struttura.

La condizione di stazionarietà $\frac{dJ(\vartheta)}{d\vartheta} = 0$ non garantisce che $\hat{\vartheta}$ sia effettivamente un minimo di $J(\vartheta) \Rightarrow$ occorre considerare la matrice

$$\begin{aligned}\frac{d^2 J(\vartheta)}{d\vartheta^2} &= \frac{d}{d\vartheta} \left[\frac{dJ(\vartheta)}{d\vartheta} \right]^T = \frac{d}{d\vartheta} \left[-2 \sum_{t=1}^N \left(y(t) - \varphi(t)^T \vartheta \right) \varphi(t)^T \right]^T = \\ &= \frac{d}{d\vartheta} \left[-2 \sum_{t=1}^N \left(y(t) \varphi(t)^T - \vartheta^T \varphi(t) \varphi(t)^T \right)^T \right] = \\ &= \frac{d}{d\vartheta} \left[-2 \sum_{t=1}^N y(t) \varphi(t) + 2 \sum_{t=1}^N \varphi(t) \varphi(t)^T \vartheta \right] = \\ &= 2 \sum_{t=1}^N \frac{d}{d\vartheta} \varphi(t) \varphi(t)^T \vartheta = 2 \sum_{t=1}^N \varphi(t) \varphi(t)^T\end{aligned}$$

che è semidefinita positiva, in quanto $\forall x \in \mathbb{R}^n$

$$x^T \frac{d^2 J(\vartheta)}{d\vartheta^2} x = x^T 2 \sum_{t=1}^N \varphi(t) \varphi(t)^T x = 2 \sum_{t=1}^N x^T \varphi(t) \varphi(t)^T x = 2 \sum_{t=1}^N \left(x^T \varphi(t) \right)^2 \geq 0$$



$\hat{\vartheta}$ è sicuramente un minimo (locale o globale) di $J(\vartheta)$.

Lo sviluppo in serie di Taylor di $J(\vartheta)$ nell'intorno di $\hat{\vartheta}$ permette di capire se $\hat{\vartheta}$ è un minimo locale o globale:

$$J(\vartheta) = J(\hat{\vartheta}) + \frac{dJ(\vartheta)}{d\vartheta} \Big|_{\hat{\vartheta}} (\vartheta - \hat{\vartheta}) + \frac{1}{2} (\vartheta - \hat{\vartheta})^T \frac{d^2J(\vartheta)}{d\vartheta^2} \Big|_{\hat{\vartheta}} (\vartheta - \hat{\vartheta}) + \dots = J(\hat{\vartheta}) + \frac{1}{2} (\vartheta - \hat{\vartheta})^T \frac{d^2J(\vartheta)}{d\vartheta^2} \Big|_{\hat{\vartheta}} (\vartheta - \hat{\vartheta})$$

essendo nulle sia $\frac{dJ(\vartheta)}{d\vartheta} \Big|_{\vartheta=\hat{\vartheta}}$ ($\hat{\vartheta}$ è un minimo) sia le derivate di $J(\vartheta)$ di ordine superiore al secondo ($J(\vartheta)$ è una funzione quadratica di ϑ)

$$J(\vartheta) - J(\hat{\vartheta}) = \frac{1}{2} (\vartheta - \hat{\vartheta})^T \frac{d^2J(\vartheta)}{d\vartheta^2} \Big|_{\hat{\vartheta}} (\vartheta - \hat{\vartheta}), \quad \frac{d^2J(\vartheta)}{d\vartheta^2} \Big|_{\hat{\vartheta}} = 2 \sum_{t=1}^N \varphi(t) \varphi(t)^T,$$

è una forma quadratica semidefinita positiva, essendo $\frac{d^2J(\vartheta)}{d\vartheta^2} \Big|_{\hat{\vartheta}}$ semidefinita positiva:

- se $\sum_{t=1}^N \varphi(t) \varphi(t)^T$ è invertibile $\Rightarrow \frac{d^2J(\vartheta)}{d\vartheta^2} \Big|_{\hat{\vartheta}}$ è definita positiva \Rightarrow la forma quadratica è definita positiva ed è un paraboloide con un unico minimo $\Rightarrow \hat{\vartheta}$ è il minimo globale di $J(\vartheta)$;
- se $\sum_{t=1}^N \varphi(t) \varphi(t)^T$ è singolare \Rightarrow la forma quadratica è semidefinita positiva e presenta infiniti punti di minimo locale, allineati su una retta tangente a $J(\vartheta)$.

I risultati ottenuti possono essere riscritti in forma matriciale compatta introducendo:

$$\Phi = \begin{bmatrix} \varphi(1)^T \\ \vdots \\ \varphi(N)^T \end{bmatrix} = \begin{bmatrix} u_1(1) & \dots & u_n(1) \\ \vdots & & \vdots \\ u_1(N) & \dots & u_n(N) \end{bmatrix} \in \mathbb{R}^{N \times n}, \quad y = \begin{bmatrix} y(1) \\ \vdots \\ y(N) \end{bmatrix} \in \mathbb{R}^N$$

$$\Downarrow$$

$$y(t) = \varphi(t)^T \vartheta, \quad t = 1, 2, \dots, N \quad \Leftrightarrow \quad \boxed{\mathbf{y} = \Phi \vartheta}$$

$$\Downarrow$$

$$\sum_{t=1}^N \varphi(t) \varphi(t)^T = \Phi^T \Phi, \quad \sum_{t=1}^N \varphi(t) y(t) = \Phi^T y$$

$$\Downarrow$$

il sistema delle equazioni normali diventa:

$$\Phi^T \Phi \vartheta = \Phi^T y$$

che, se $\Phi^T \Phi$ è invertibile (*condizione di identificabilità*), ha come unica soluzione la stima ai minimi quadrati:

$$\boxed{\hat{\vartheta} = [\Phi^T \Phi]^{-1} \Phi^T y}$$

Caratteristiche probabilistiche dello stimatore ai minimi quadrati

Ipotesi:

- vale la condizione di identificabilità: $\exists [\Phi^T \Phi]^{-1}$;
- la sorgente casuale dei dati ha la struttura

$$y(t) = \varphi(t)^T \vartheta^o + v(t), \quad t = 1, 2, \dots, N$$

dove $v(t)$ è un disturbo casuale a valor medio nullo \Rightarrow

si suppone che il legame fra y e u_1, u_2, \dots, u_n sia effettivamente lineare \Rightarrow

esiste effettivamente un valore “vero” ϑ^o della grandezza incognita;

in forma matriciale compatta, si ha:

$$y = \Phi \vartheta^o + v$$

dove $v = \begin{bmatrix} v(1) \\ \vdots \\ v(N) \end{bmatrix} \in \mathbb{R}^N$ è una variabile casuale vettoriale avente $E[v] = \mathbf{0}$;

Sotto tali ipotesi, lo stimatore ai minimi quadrati diventa:

$$\begin{aligned}\hat{\vartheta} &= [\Phi^T \Phi]^{-1} \Phi^T y = [\Phi^T \Phi]^{-1} \Phi^T (\Phi \vartheta^o + v) = \\ &= [\Phi^T \Phi]^{-1} \Phi^T \Phi \vartheta^o + [\Phi^T \Phi]^{-1} \Phi^T v = \vartheta^o + [\Phi^T \Phi]^{-1} \Phi^T v\end{aligned}$$

ed ha le seguenti caratteristiche probabilistiche:

- è **corretto**, avendo valor medio $E[\hat{\vartheta}] = \vartheta^o$

$$\begin{aligned}E[\hat{\vartheta}] &= E\left[[\Phi^T \Phi]^{-1} \Phi^T y\right] = [\Phi^T \Phi]^{-1} \Phi^T E[y] = [\Phi^T \Phi]^{-1} \Phi^T E[\Phi \vartheta^o + v] = \\ &= [\Phi^T \Phi]^{-1} \Phi^T (\Phi \vartheta^o + E[v]) = [\Phi^T \Phi]^{-1} \Phi^T \Phi \vartheta^o = \vartheta^o\end{aligned}$$

- se v è un vettore di variabili casuali a valor medio nullo fra loro incorrelate e con la stessa varianza σ_v^2 ($Var[v] = E[vv^T] = \sigma_v^2 I_N$), come nel caso di disturbo $v(\cdot)$ costituito da un rumore bianco $WN(0, \sigma_v^2) \Rightarrow Var[\hat{\vartheta}] = \sigma_v^2 [\Phi^T \Phi]^{-1}$

$$\begin{aligned}Var[\hat{\vartheta}] &= E\left[(\hat{\vartheta} - E[\hat{\vartheta}])(\hat{\vartheta} - E[\hat{\vartheta}])^T\right] = E\left[(\hat{\vartheta} - \vartheta^o)(\hat{\vartheta} - \vartheta^o)^T\right] = \\ &= E\left[\left([\Phi^T \Phi]^{-1} \Phi^T v\right)\left([\Phi^T \Phi]^{-1} \Phi^T v\right)^T\right] = E\left[[\Phi^T \Phi]^{-1} \Phi^T v v^T \Phi [\Phi^T \Phi]^{-1}\right] = \\ &= [\Phi^T \Phi]^{-1} \Phi^T E[vv^T] \Phi [\Phi^T \Phi]^{-1} = [\Phi^T \Phi]^{-1} \Phi^T \sigma_v^2 I_N \Phi [\Phi^T \Phi]^{-1} = \\ &= \sigma_v^2 [\Phi^T \Phi]^{-1} \Phi^T \Phi [\Phi^T \Phi]^{-1} = \sigma_v^2 [\Phi^T \Phi]^{-1}\end{aligned}$$

- Di solito la varianza σ_v^2 del disturbo v non è nota \Rightarrow sotto le stesse precedenti ipotesi, una stima “ragionevole” corretta $\hat{\sigma}_v^2$ (tale che $E[\hat{\sigma}_v^2] = \sigma_v^2$) può essere ricavata a partire direttamente dai dati come

$$\hat{\sigma}_v^2 = \frac{J(\hat{\vartheta})}{N - n}$$

con N = numero delle misure, n = numero dei parametri incogniti di ϑ ,

$$\begin{aligned} J(\hat{\vartheta}) &= \sum_{t=1}^N \varepsilon(t)^2 \Big|_{\vartheta=\hat{\vartheta}} = \sum_{t=1}^N \left[y(t) - \varphi(t)^T \hat{\vartheta} \right]^2 = [y - \Phi \hat{\vartheta}]^T [y - \Phi \hat{\vartheta}] = \\ &= \left((I_N - \Phi [\Phi^T \Phi]^{-1} \Phi^T) y \right)^T (I_N - \Phi [\Phi^T \Phi]^{-1} \Phi^T) y = \\ &= y^T (I_N - \Phi [\Phi^T \Phi]^{-1} \Phi^T) (I_N - \Phi [\Phi^T \Phi]^{-1} \Phi^T) y = \\ &= y^T (I_N - 2\Phi [\Phi^T \Phi]^{-1} \Phi^T + \Phi [\Phi^T \Phi]^{-1} \Phi^T \Phi [\Phi^T \Phi]^{-1} \Phi^T) y = \\ &= y^T (I_N - \Phi [\Phi^T \Phi]^{-1} \Phi^T) y \end{aligned}$$

\Downarrow

$$\text{Var}[\hat{\vartheta}] = \sigma_v^2 [\Phi^T \Phi]^{-1} \cong \hat{\sigma}_v^2 [\Phi^T \Phi]^{-1}$$

Metodo di stima ai Minimi Quadrati Ponderati

Con il metodo di stima ai minimi quadrati si attribuisce identico rilievo a tutti gli errori, in quanto si minimizza la cifra di merito

$$J_{LS}(\vartheta) = \sum_{t=1}^N \varepsilon(t)^2, \quad \text{con} \quad \varepsilon(t) = y(t) - \varphi(t)^T \vartheta, \quad t = 1, 2, \dots, N.$$

Se però alcune misure sono più accurate di altre, si possono assegnare pesi diversi ai vari errori definendo la cifra di merito

$$J_{WLS}(\vartheta) = \sum_{t=1}^N q(t) \varepsilon(t)^2 = \varepsilon^T Q \varepsilon$$

dove $q(t)$ sono i coefficienti di ponderazione (o *pesi*) per $t = 1, 2, \dots, N$,

$$Q = \text{diag}(q(t)) = \begin{bmatrix} q(1) & 0 & \dots & 0 \\ 0 & q(2) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & q(N) \end{bmatrix} \in \mathbb{R}^{N \times N}, \quad \varepsilon = \begin{bmatrix} \varepsilon(1) \\ \vdots \\ \varepsilon(N) \end{bmatrix} \in \mathbb{R}^N.$$

La **stima ai minimi quadrati ponderati (Weighted Least Squares, WLS)** minimizza la cifra di merito $J_{WLS}(\vartheta)$:

$$\hat{\vartheta}_{WLS} = [\Phi^T Q \Phi]^{-1} \Phi^T Q y$$

Se il disturbo v è un vettore di variabili casuali a valor medio nullo fra loro incorrelate e con varianza Σ_v , lo stimatore $\hat{\vartheta}_{WLS}$ ha le seguenti caratteristiche probabilistiche:

- è **corretto**, avendo valor medio $E[\hat{\vartheta}_{WLS}] = \vartheta^o$

$$\begin{aligned} E[\hat{\vartheta}_{WLS}] &= E\left[[\Phi^T Q \Phi]^{-1} \Phi^T Q y\right] = [\Phi^T Q \Phi]^{-1} \Phi^T Q E[y] = [\Phi^T Q \Phi]^{-1} \Phi^T Q E[\Phi \vartheta^o + v] = \\ &= [\Phi^T Q \Phi]^{-1} \Phi^T Q (\Phi \vartheta^o + E[v]) = [\Phi^T Q \Phi]^{-1} \Phi^T Q \Phi \vartheta^o = \vartheta^o \end{aligned}$$

- la sua varianza vale

$$\begin{aligned} Var[\hat{\vartheta}_{WLS}] &= E[(\hat{\vartheta}_{WLS} - E[\hat{\vartheta}_{WLS}])(\hat{\vartheta}_{WLS} - E[\hat{\vartheta}_{WLS}])^T] = \\ &= E[(\hat{\vartheta}_{WLS} - \vartheta^o)(\hat{\vartheta}_{WLS} - \vartheta^o)^T] = E\left[[\Phi^T Q \Phi]^{-1} \Phi^T Q v ([\Phi^T Q \Phi]^{-1} \Phi^T Q v)^T\right] = \\ &= E\left[[\Phi^T Q \Phi]^{-1} \Phi^T Q v v^T Q^T \Phi [\Phi^T Q \Phi]^{-1}\right] = \\ &= [\Phi^T Q \Phi]^{-1} \Phi^T Q E[v v^T] Q \Phi [\Phi^T Q \Phi]^{-1} = [\Phi^T Q \Phi]^{-1} \Phi^T Q \Sigma_v Q \Phi [\Phi^T Q \Phi]^{-1} \end{aligned}$$

e quindi dipende dalla varianza del disturbo Σ_v ;

- si dimostra che la miglior scelta di Q che minimizza $Var[\hat{\vartheta}_{WLS}]$ è data da

$$Q^* = \arg \min_{Q=\text{diag}(q(t)) \in \mathbb{R}^{n \times n}} Var[\hat{\vartheta}_{WLS}] = \Sigma_v^{-1}$$

e si parla in tal caso di **stima di Gauss-Markov**:

$$\hat{\vartheta}_{GM} = [\Phi^T \Sigma_v^{-1} \Phi]^{-1} \Phi^T \Sigma_v^{-1} \mathbf{y}$$

che ha come varianza

$$\begin{aligned} Var[\hat{\vartheta}_{GM}] &= [\Phi^T Q \Phi]^{-1} \Phi^T Q \Sigma_v Q \Phi [\Phi^T Q \Phi]^{-1} = \\ &= [\Phi^T \Sigma_v^{-1} \Phi]^{-1} \Phi^T \Sigma_v^{-1} \Sigma_v Q \Phi [\Phi^T \Sigma_v^{-1} \Phi]^{-1} \\ &= [\Phi^T \Sigma_v^{-1} \Phi]^{-1} ; \end{aligned}$$

se in particolare risulta $\Sigma_v = \sigma_v^2 I_N \Rightarrow$

$$\hat{\vartheta}_{GM} = \left[\Phi^T \frac{1}{\sigma_v^2} I_N \Phi \right]^{-1} \Phi^T \frac{1}{\sigma_v^2} I_N \mathbf{y} = [\Phi^T \Phi]^{-1} \Phi^T \mathbf{y} = \hat{\vartheta}_{LS}$$

Stimatori a Massima Verosimiglianza

I dati effettivi sono generati da una sorgente casuale, influenzata dall'esito s di un esperimento casuale e dal valore "vero" ϑ^o della grandezza incognita da stimare. Se però si considera un generico valore ϑ della grandezza da stimare, i dati possono essere visti come funzione sia di ϑ sia dell'esito $s \Rightarrow$

si indicano con $d^{(\vartheta)}(s)$, la cui d.d.p. $f(q, \vartheta)$ è anch'essa funzione di ϑ .

Sia δ una particolare osservazione dei dati, quella relativa ad un certo esito \bar{s} dello esperimento casuale:

$$\delta = d^{(\vartheta)}(\bar{s})$$

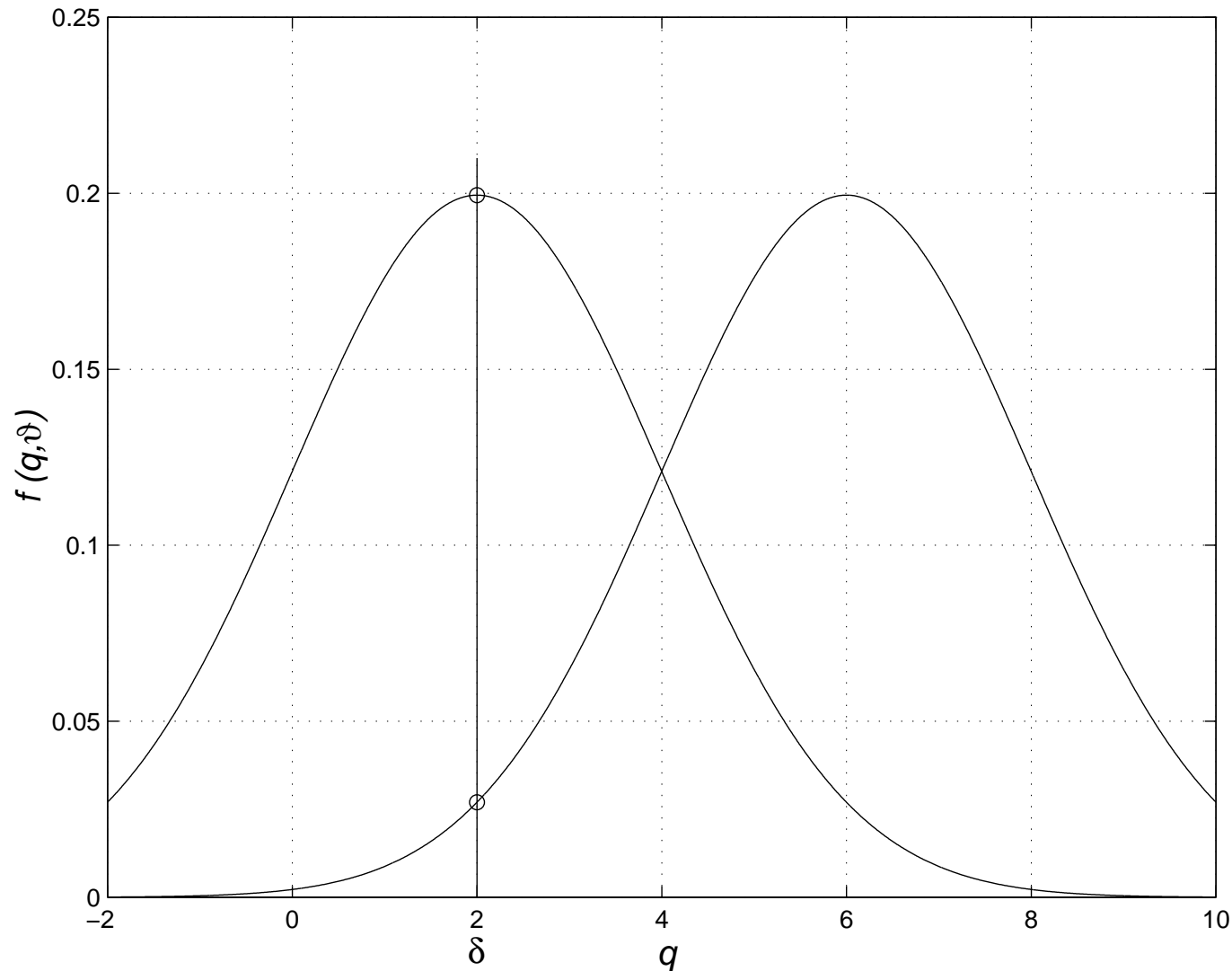
Si considera allora la **funzione di verosimiglianza** data dalla d.d.p. dei dati in δ :

$$L(\vartheta) = f(q, \vartheta)|_{q=\delta}$$

Si definisce **stima a massima verosimiglianza (Maximum Likelihood, ML)**

$$\hat{\vartheta}_{ML} = \arg \max_{\vartheta \in \mathbb{R}^n} L(\vartheta)$$

Esempio (ϑ^0 scalare, da stimare con un'unica misura avente d.d.p. gaussiana):



al variare del valore di ϑ scelto, la gaussiana trasla \Rightarrow cambia $L(\vartheta) = f(q, \vartheta)|_{q=\delta}$

Proprietà della stima di massima verosimiglianza

$\hat{\vartheta}_{ML}$ è una stima:

- asintoticamente corretta: $E \left(\hat{\vartheta}_{ML} \right) \xrightarrow{N \rightarrow \infty} \vartheta^0$
- asintoticamente efficiente: $\Sigma_{\hat{\vartheta}_{ML}} \leq \Sigma_{\hat{\vartheta}} \quad \forall \hat{\vartheta} \text{ se } N \rightarrow \infty$
- consistente: $\lim_{N \rightarrow \infty} \Sigma_{\hat{\vartheta}_{ML}} = 0$
- asintoticamente gaussiana (per $N \rightarrow \infty$)

Esempio: si supponga che la sorgente casuale dei dati abbia la struttura

$$y(t) = \psi(t, \vartheta^o) + v(t), \quad t = 1, 2, \dots, N \quad \Leftrightarrow \quad y = \Psi(\vartheta^o) + v$$

dove $\psi(t, \vartheta^o)$ è una generica funzione *non lineare* di ϑ^o ed il disturbo v è un vettore di variabili casuali gaussiane a valor medio nullo e con varianza Σ_v , con d.d.p.

$$f(q) = \mathcal{N}(0, \Sigma_v) = \frac{1}{\sqrt{(2\pi)^N \det \Sigma_v}} \exp\left(-\frac{1}{2} q^T \Sigma_v^{-1} q\right)$$

Poiché $v = y - \Psi(\vartheta^o) \Rightarrow$ la d.d.p. dei dati generati da una sorgente casuale in cui si considera un generico ϑ in luogo di ϑ^o è data da

$$f(q, \vartheta) = \frac{1}{\sqrt{(2\pi)^N \det \Sigma_v}} \exp\left(-\frac{1}{2} [q - \Psi(\vartheta)]^T \Sigma_v^{-1} [q - \Psi(\vartheta)]\right)$$

\Downarrow

$$L(\vartheta) = f(q, \vartheta)|_{q=\delta} = \frac{1}{\sqrt{(2\pi)^N \det \Sigma_v}} \exp\left(-\frac{1}{2} [\delta - \Psi(\vartheta)]^T \Sigma_v^{-1} [\delta - \Psi(\vartheta)]\right)$$



$f(q, \vartheta)|_{q=\delta}$ è una funzione esponenziale in ϑ



$$\hat{\vartheta}_{ML} = \arg \max_{\vartheta \in \mathbb{R}^n} L(\vartheta) = \arg \min_{\vartheta \in \mathbb{R}^n} \underbrace{\left\{ [\delta - \Psi(\vartheta)]^T \Sigma_v^{-1} [\delta - \Psi(\vartheta)] \right\}}_{R(\vartheta)}$$

Problema: bisogna trovare il minimo globale di $R(\vartheta)$ al variare di ϑ , che però può presentare minimi locali se $\Psi(\vartheta)$ è una generica funzione non lineare dei parametri; gli usuali algoritmi di ottimizzazione non lineare non garantiscono di arrivare sempre al minimo globale.

Caso particolare: $\Psi(\vartheta) =$ funzione *lineare* dei parametri $= \Phi\vartheta$

⇓

$$R(\vartheta) \text{ è quadratica in } \vartheta : R(\vartheta) = [\delta - \Phi\vartheta]^T \Sigma_v^{-1} [\delta - \Phi\vartheta]$$

⇓

esiste un solo minimo di $R(\vartheta)$

⇓

$$\begin{aligned} \hat{\vartheta}_{ML} &= (\Phi^T \Sigma_v^{-1} \Phi)^{-1} \Phi^T \Sigma_v^{-1} \delta = \text{stima di Gauss-Markov} = \hat{\vartheta}_{GM} = \\ &= \text{stima dei minimi quadrati pesati con la covarianza degli errori} \end{aligned}$$

Nel caso $\Sigma_v = \sigma_v^2 I_N$ (errori *i.i.d.*, cioè indipendenti identicamente distribuiti):

$$\hat{\vartheta}_{ML} = \hat{\vartheta}_{GM} = (\Phi^T \Phi)^{-1} \Phi^T \delta = \text{stima dei minimi quadrati}$$

Proprietà della stima di Gauss-Markov

$\hat{\vartheta}_{GM}$ è una stima:

- corretta: $E \left(\hat{\vartheta}_{GM} \right) = \vartheta^o$
- efficiente: $\Sigma_{\hat{\vartheta}_{GM}} \leq \Sigma_{\hat{\vartheta}} \quad \forall \hat{\vartheta}$
- consistente: $\lim_{N \rightarrow \infty} \Sigma_{\hat{\vartheta}_{GM}} = 0$
- gaussiana

Metodo di stima di Bayes

Il metodo di Bayes prevede l'utilizzo sia di dati sperimentali sia di informazioni *a priori* sull'incognita del problema di stima che, se ben sfruttate, possono migliorare la stima e compensare eventuali errori di carattere casuale nei dati:

- l'incognita \mathcal{V} è vista come variabile o vettore casuale, con una certa d.d.p. a priori (ossia in assenza di ogni dato) avente un certo andamento, un certo valor medio e una certa varianza



una possibile stima è il valor medio e la varianza costituisce l'incertezza a priori;

- man mano che giungono dati sperimentali, la d.d.p. di \mathcal{V} si aggiorna alla luce delle nuove informazioni, con il valor medio che cambia rispetto a quello a priori e con la varianza che dovrebbe diminuire grazie alle informazioni apportate dai dati.

Si ipotizza quindi che esista un esperimento casuale congiunto $\mathcal{E} = \mathcal{E}_1 \times \mathcal{E}_2$ avente come esito congiunto la coppia di esiti singoli $s = (s_1, s_2)$:

- l'incognita ϑ è generata da una prima sorgente casuale \mathcal{S}_1 in base all'esito s_1 di un primo esperimento casuale $\mathcal{E}_1 \Rightarrow \vartheta = \vartheta(s_1)$;
- i dati d sono generati da una seconda sorgente casuale \mathcal{S}_2 , influenzata
 - dall'esito s_2 di un secondo esperimento casuale \mathcal{E}_2
 - dal valore $\vartheta(s_1)$ dell'incognita da stimare

$$d = d(s_2, \vartheta(s_1))$$

Un generico stimatore è una funzione dei dati $\hat{\vartheta} = h(d)$ ed è tanto migliore quanto più la stima $\hat{\vartheta}$ è vicina all'incognita da stimare



considerando come cifra di merito

$$J(h(\cdot)) = E[\|\vartheta - h(d)\|^2]$$

lo stimatore ottimo sarà quella particolare funzione $h^*(\cdot)$ tale che

$$E[\|\vartheta - h^*(d)\|^2] \leq E[\|\vartheta - h(d)\|^2], \quad \forall h(\cdot)$$

Si dimostra che tale stimatore ottimo esiste ed è dato da:

$$h^*(x) = E[\vartheta | d = x]$$

dove x è il generico valore che possono assumere i dati d .

Lo **stimatore di Bayes** (o **stimatore a valor medio condizionato**) è la funzione

$$\hat{\vartheta} = E[\vartheta | d]$$

mentre la **stima di Bayes** (o **stima a valor medio condizionato**) è il valore numerico

$$\hat{\vartheta} = E[\vartheta | d = \delta]$$

dove δ è il particolare valore assunto dai dati d in corrispondenza di un certo esito.

Stimatore di Bayes nel caso gaussiano

Ipotesi: sia i dati d sia l'incognita ϑ sono variabili casuali scalari, gaussiane sia singolarmente sia congiuntamente: $\begin{bmatrix} d \\ \vartheta \end{bmatrix} \sim G \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_{d\vartheta} = \begin{bmatrix} \sigma_{dd} & \sigma_{d\vartheta} \\ \sigma_{\vartheta d} & \sigma_{\vartheta\vartheta} \end{bmatrix} \right)$

per cui la loro d.d.p. congiunta è

$$f(d, \vartheta) = C \exp \left\{ -\frac{1}{2} [d \quad \vartheta] \Sigma_{d\vartheta}^{-1} [d \quad \vartheta]^T \right\}, \quad C : \text{opportuna costante}$$

Poiché

$$\det \Sigma_{d\vartheta} = \det \begin{bmatrix} \sigma_{dd} & \sigma_{d\vartheta} \\ \sigma_{\vartheta d} & \sigma_{\vartheta\vartheta} \end{bmatrix} = \sigma_{dd}\sigma_{\vartheta\vartheta} - \sigma_{d\vartheta}^2 = \sigma_{dd} \left(\sigma_{\vartheta\vartheta} - \frac{\sigma_{d\vartheta}^2}{\sigma_{dd}} \right) = \sigma_{dd}\sigma^2,$$

$$\text{dove } \sigma^2 = \sigma_{\vartheta\vartheta} - \sigma_{d\vartheta}^2 / \sigma_{dd}$$

⇓

$$\Sigma_{d\vartheta}^{-1} = \frac{1}{\det \Sigma_{d\vartheta}} \begin{bmatrix} \sigma_{\vartheta\vartheta} & -\sigma_{d\vartheta} \\ -\sigma_{\vartheta d} & \sigma_{dd} \end{bmatrix} = \frac{1}{\sigma^2} \begin{bmatrix} \sigma_{\vartheta\vartheta}/\sigma_{dd} & -\sigma_{d\vartheta}/\sigma_{dd} \\ -\sigma_{\vartheta d}/\sigma_{dd} & 1 \end{bmatrix}$$

$$\begin{aligned}
 f(d, \vartheta) &= C \exp \left\{ -\frac{1}{2\sigma^2} [d \quad \vartheta] \begin{bmatrix} \sigma_{\vartheta\vartheta}/\sigma_{dd} & -\sigma_{d\vartheta}/\sigma_{dd} \\ -\sigma_{\vartheta d}/\sigma_{dd} & 1 \end{bmatrix} \begin{bmatrix} d \\ \vartheta \end{bmatrix} \right\} = \\
 &= C \exp \left\{ -\frac{1}{2\sigma^2} [d \quad \vartheta] \begin{bmatrix} \sigma_{\vartheta\vartheta}/\sigma_{dd} d - \sigma_{d\vartheta}/\sigma_{dd} \vartheta \\ -\sigma_{\vartheta d}/\sigma_{dd} d + \vartheta \end{bmatrix} \right\} = \\
 &= C \exp \left\{ -\frac{1}{2\sigma^2} \left(\frac{\sigma_{\vartheta\vartheta}}{\sigma_{dd}} d^2 - 2 \frac{\sigma_{d\vartheta}}{\sigma_{dd}} d\vartheta + \vartheta^2 \right) \right\}
 \end{aligned}$$

La d.d.p. dei dati d è invece

$$f(d) = C' \exp \left\{ -\frac{d^2}{2\sigma_{dd}} \right\}, \quad C' : \text{opportuna costante}$$

⇓

la d.d.p. dell'incognita ϑ condizionata dai dati d vale:

$$\begin{aligned}
 f(\vartheta|d) &= \frac{f(d, \vartheta)}{f(d)} = \frac{C}{C'} \exp \left\{ -\frac{1}{2\sigma^2} \left(\frac{\sigma_{\vartheta\vartheta}}{\sigma_{dd}} d^2 - 2 \frac{\sigma_{d\vartheta}}{\sigma_{dd}} d\vartheta + \vartheta^2 \right) + \frac{d^2}{2\sigma_{dd}} \right\} = \\
 &= C'' \exp \left\{ -\frac{1}{2\sigma^2} \left[\frac{\sigma_{d\vartheta}^2}{\sigma_{dd}^2} d^2 - 2 \frac{\sigma_{d\vartheta}}{\sigma_{dd}} d\vartheta + \vartheta^2 \right] \right\} = C'' \exp \left\{ -\frac{1}{2\sigma^2} \left[\vartheta - \frac{\sigma_{d\vartheta}}{\sigma_{dd}} d \right]^2 \right\}
 \end{aligned}$$

$$f(\vartheta|d) = C'' \exp \left\{ -\frac{1}{2\sigma^2} \left[\vartheta - \frac{\sigma_{d\vartheta}}{\sigma_{dd}} d \right]^2 \right\} \sim G \left(\frac{\sigma_{d\vartheta}}{\sigma_{dd}} d, \sigma^2 \right)$$

Lo stimatore di Bayes è la funzione

$$\hat{\vartheta} = E[\vartheta|d] = \frac{\sigma_{d\vartheta}}{\sigma_{dd}} d$$

mentre la stima di Bayes corrispondente ad un particolare dato δ è il valore numerico

$$\hat{\vartheta} = E[\vartheta|d = \delta] = \frac{\sigma_{d\vartheta}}{\sigma_{dd}} \delta$$

Poiché $E[d] = E[\vartheta] = 0 \Rightarrow$

$$E[\hat{\vartheta}] = E\left[\frac{\sigma_{d\vartheta}}{\sigma_{dd}} d\right] = \frac{\sigma_{d\vartheta}}{\sigma_{dd}} E[d] = 0$$

$$Var[\hat{\vartheta}] = E[(\hat{\vartheta} - E[\hat{\vartheta}])^2] = E[\hat{\vartheta}^2] = E\left[\frac{\sigma_{d\vartheta}^2}{\sigma_{dd}^2} d^2\right] = \frac{\sigma_{d\vartheta}^2}{\sigma_{dd}^2} E[d^2] = \frac{\sigma_{d\vartheta}^2}{\sigma_{dd}^2}$$

$$\begin{aligned} Var[\vartheta - \hat{\vartheta}] &= E[(\vartheta - \hat{\vartheta})^2] = E\left[\left(\vartheta - \frac{\sigma_{d\vartheta}}{\sigma_{dd}} d\right)^2\right] = E\left[\vartheta^2 - 2\frac{\sigma_{d\vartheta}}{\sigma_{dd}} d\vartheta + \frac{\sigma_{d\vartheta}^2}{\sigma_{dd}^2} d^2\right] = \\ &= E[\vartheta^2] - 2\frac{\sigma_{d\vartheta}}{\sigma_{dd}} E[d\vartheta] + \frac{\sigma_{d\vartheta}^2}{\sigma_{dd}^2} E[d^2] = \sigma_{\vartheta\vartheta} - 2\frac{\sigma_{d\vartheta}}{\sigma_{dd}} \sigma_{d\vartheta} + \frac{\sigma_{d\vartheta}^2}{\sigma_{dd}^2} \sigma_{dd} = \\ &= \sigma_{\vartheta\vartheta} - 2\frac{\sigma_{d\vartheta}^2}{\sigma_{dd}} + \frac{\sigma_{d\vartheta}^2}{\sigma_{dd}} = \sigma_{\vartheta\vartheta} - \frac{\sigma_{d\vartheta}^2}{\sigma_{dd}} = \sigma^2 \end{aligned}$$

Stimatore lineare di Bayes

Ipotesi: sia i dati d sia l'incognita ϑ sono variabili casuali scalari a valor medio nullo

e matrice di varianza $\Sigma_{d\vartheta} = \begin{bmatrix} \sigma_{dd} & \sigma_{d\vartheta} \\ \sigma_{\vartheta d} & \sigma_{\vartheta\vartheta} \end{bmatrix}$.

Si desidera stimare ϑ mediante uno stimatore *lineare*, con struttura

$$\hat{\vartheta} = \alpha d + \beta$$

con α, β parametri reali stimati minimizzando

$$J = E[(\vartheta - \hat{\vartheta})^2] = E[(\vartheta - \alpha d - \beta)^2]$$

\Downarrow

$$\begin{aligned} \frac{\partial J}{\partial \alpha} &= \frac{\partial}{\partial \alpha} E[(\vartheta - \alpha d - \beta)^2] = E\left[\frac{\partial}{\partial \alpha} (\vartheta - \alpha d - \beta)^2\right] = E[-2d(\vartheta - \alpha d - \beta)] = \\ &= -2E[d\vartheta] + 2\alpha E[d^2] + 2\beta E[d] = -2\sigma_{d\vartheta} + 2\alpha\sigma_{dd} = 0 \end{aligned}$$

$$\begin{aligned} \frac{\partial J}{\partial \beta} &= \frac{\partial}{\partial \beta} E[(\vartheta - \alpha d - \beta)^2] = E\left[\frac{\partial}{\partial \beta} (\vartheta - \alpha d - \beta)^2\right] = E[-2(\vartheta - \alpha d - \beta)] = \\ &= -2E[\vartheta] + 2\alpha E[d] + 2\beta = 2\beta = 0 \end{aligned}$$

\Downarrow

$$\begin{cases} \alpha = \sigma_{d\vartheta} / \sigma_{dd} \\ \beta = 0 \end{cases}$$