

Dispense
di
Identificazione e Analisi dei Dati

Antonello Giannitrapani, Andrea Garulli

Versione 1.1

26 gennaio 2007

Indice

1	Teoria della probabilità	1
1.1	Spazi di probabilità	1
1.2	Variabili aleatorie, distribuzioni e densità	8
1.3	Distribuzioni bivariate e multivariate	13
1.4	Media e varianza	16
1.5	Variabili aleatorie gaussiane	18
1.6	Funzioni di una o più variabili aleatorie	21
1.7	Distribuzioni condizionali	26
1.8	Esercizi	31
2	Processi stocastici	37
2.1	Definizioni e proprietà	37
2.2	Esempi di processi stocastici	45
2.2.1	Processi stocastici puramente deterministici	45
2.2.2	Rumore bianco	47
2.2.3	Processo di Wiener	48
2.2.4	Processi esponenzialmente correlati	49
2.3	Analisi in frequenza	54
2.4	Sistemi lineari stocastici	62
2.5	Modelli lineari di processi stocastici	66
2.5.1	Processi MA	67
2.5.2	Processi AR	72
2.5.3	Processi ARMA	75

2.6	Processi ergodici	79
2.7	Esercizi	79
3	Teoria della stima	83
3.1	Stima parametrica	83
3.1.1	Formulazione del problema.	83
3.1.2	Proprietà di uno stimatore	85
3.1.3	Stima non polarizzata a minima varianza	90
3.2	Limite di Cramér-Rao	93
3.3	Stima a massima verosimiglianza	96
3.3.1	Rumore additivo	100
3.4	Problemi di stima lineare	101
3.5	Stima Bayesiana	107
3.5.1	Stima a minimo errore quadratico medio	107
3.5.2	Stima lineare a minimo errore quadratico medio	110
3.6	Esercizi	115
4	Filtraggio ottimo	119
4.1	Formulazione del problema	119
4.2	Filtro di Wiener FIR	122
4.3	Predittore di Wiener FIR	127
4.4	Fattorizzazione spettrale	129
4.5	Predittore di Wiener	134
4.6	Esercizi	141
5	Identificazione parametrica	147
5.1	Introduzione	147
5.2	Modelli lineari tempo-invarianti	152
5.2.1	Modelli ARX	153
5.2.2	Modelli ARMAX	154
5.2.3	Modelli OE	154
5.2.4	Modelli BJ	155

5.3	Criterio di scelta del modello	156
5.3.1	Scelta del funzionale di costo	161
5.4	Selezione del modello ottimo	162
5.4.1	Regressione lineare e minimi quadrati	162
5.4.2	Soluzioni numeriche del problema di ottimizzazione . .	165
6	Valutazione della qualità dei modelli	171
6.1	Confronto tra classi di modelli	171
6.1.1	Conoscenza a priori	173
6.1.2	Analisi dei dati	173
6.1.3	Confronto tra diverse strutture	174
6.1.4	Validazione del modello: analisi dei residui	179
6.2	Scelta dell'ingresso	181

Capitolo 1

Concetti fondamentali di teoria della probabilità

In questo capitolo vengono richiamate alcune nozioni fondamentali della teoria della probabilità.

1.1 Spazi di probabilità

Definizione 1.1. Si definisce *spazio di probabilità* la tripla $\{\Omega, \mathcal{A}, P(\cdot)\}$, in cui:

1. $\Omega = \{\omega_1, \omega_2, \dots\}$ è un insieme, eventualmente infinito, detto *spazio degli eventi elementari*;
2. $\mathcal{A} = \{A_1, A_2, \dots\}$ è una famiglia di sottoinsiemi A_i di Ω , detti *eventi*, chiusa rispetto alle operazioni di unione, intersezione e passaggio al complementare:¹

$$(a) \quad \forall A_1, A_2 \in \mathcal{A} \Rightarrow A_1 \cup A_2 \in \mathcal{A};$$

$$(b) \quad \forall A_1, A_2 \in \mathcal{A} \Rightarrow A_1 \cap A_2 \in \mathcal{A};$$

$$(c) \quad \forall A \in \mathcal{A} \Rightarrow \bar{A} \in \mathcal{A};$$

¹Una famiglia che gode di tali proprietà è detta *algebra*

3. $P(\cdot)$ è una funzione reale definita su \mathcal{A} che gode delle seguenti proprietà:

- (a) $P(A) \geq 0, \quad \forall A \in \mathcal{A};$
- (b) $P(\Omega) = 1;$
- (c) $\forall A_1, A_2 \in \mathcal{A}: A_1 \cap A_2 = \emptyset \Rightarrow P(A_1 \cup A_2) = P(A_1) + P(A_2).$

Per definire uno spazio di probabilità occorre specificare:

- lo spazio Ω degli eventi elementari ω_i , che rappresentano tutti i possibili esiti del fenomeno incerto che si intende studiare;
- la famiglia \mathcal{A} di eventi A_i , che rappresentano insiemi di interesse degli esiti del fenomeno incerto in esame;
- una funzione di probabilità $P(\cdot)$ definita su \mathcal{A} .

Si noti che, dalla definizione di \mathcal{A} , se $A \in \mathcal{A}$, allora, per la proprietà di chiusura 2c, anche il suo complementare $\bar{A} \in \mathcal{A}$. Ora, per le proprietà 2a-2b occorre che:

$$A \cup \bar{A} \in \mathcal{A} \quad \text{e} \quad A \cap \bar{A} \in \mathcal{A}.$$

Ma, direttamente dalla definizione di insieme complementare, segue che $A \cup \bar{A} = \Omega$ e $A \cap \bar{A} = \emptyset$. Per cui vale sempre che:

$$\Omega \in \mathcal{A}; \tag{1.1}$$

$$\emptyset \in \mathcal{A}. \tag{1.2}$$

Esempio 1.1. Si consideri il gioco della roulette. Una biglia viene lanciata intorno ad piatto posto in rotazione, suddiviso in 37 settori, numerati da 0 a 36. Vince chi indovina il settore nel quale si fermerà la biglia.

In questo caso, si hanno 37 eventi elementari $\Omega = \{\omega_0, \omega_1, \dots, \omega_{36}\}$, uno in corrispondenza di ciascun settore in cui può fermarsi la biglia:

$$\omega_i = \{ \text{la biglia si è fermata nel settore } i\text{-esimo} \}, \quad i = 0, \dots, 36.$$

Il gioco prevede anche che possano essere effettuate scommesse multiple, cioè scommettere contemporaneamente su più di un settore. Ad esempio, è

possibile scommettere che la biglia si fermi in un qualunque settore contrassegnato da un numero pari (o, viceversa, dispari). Se fossimo interessati solo a queste due puntate, la famiglia \mathcal{A} sarebbe composta dai seguenti eventi.

Siano:

$$\begin{aligned} P &= \{ \text{la biglia si è fermata in un settore } \textit{pari} \} \\ &= \{\omega_2, \omega_4, \dots, \omega_{36}\}; \\ D &= \{ \text{la biglia si è fermata in un settore } \textit{dispari} \} \\ &= \{\omega_1, \omega_3, \dots, \omega_{35}\}. \end{aligned}$$

Costruiamo la famiglia \mathcal{A} . In virtù di quanto osservato in precedenza (eq. (1.1)-(1.2)):

$$\mathcal{A} = \{\Omega, \emptyset, \dots\};$$

ovviamente:

$$\mathcal{A} = \{\Omega, \emptyset, P, D, \dots\};$$

per la proprietà 2a, \mathcal{A} deve contenere anche $P \cup D$:

$$\mathcal{A} = \{\Omega, \emptyset, P, D, P \cup D, \dots\};$$

per la proprietà 2c, \mathcal{A} deve contenere anche $\overline{P \cup D} = \{ \text{la biglia si è fermata nel settore } 0 \} = \{\omega_0\}$:

$$\mathcal{A} = \{\Omega, \emptyset, P, D, P \cup D, \{\omega_0\}, \dots\};$$

infine, per la proprietà 2b, \mathcal{A} conterrà anche $P \cup \{\omega_0\}$ e $D \cup \{\omega_0\}$:

$$\mathcal{A} = \{\Omega, \emptyset, P, D, P \cup D, \{\omega_0\}, P \cup \{\omega_0\}, D \cup \{\omega_0\}\}.$$

È immediato verificare che la famiglia \mathcal{A} così costruita verifica tutte le proprietà di chiusura 2a-2c.

Infine, per completare la definizione dello spazio di probabilità, dobbiamo specificare una funzione di probabilità. Dobbiamo, cioè, assegnare a ciascuno degli eventi contenuti in \mathcal{A} , un numero reale, in modo che le proprietà 3a-3c siano rispettate. Supponendo che il dispositivo non sia truccato e che, quindi, ciascun settore sia equiprobabile possiamo usare la seguente regola:

$$P(A_i) = \frac{\text{n.ro di esiti favorevoli}}{\text{n.ro di esiti possibili}} = \frac{|A_i|}{|\Omega|},$$

in cui $|X|$ denota la cardinalità dell'insieme X . Nel nostro caso, essendo il numero di tutti possibili esiti $|\Omega| = 37$ e $|P| = |D| = 18$ avremo:

$$\begin{aligned} P(\Omega) &= 1; \\ P(\emptyset) &= 0; \\ P(P) &= P(D) = \frac{18}{37}; \\ P(P \cup D) &= \frac{36}{37}; \\ P(\omega_0) &= \frac{1}{37}; \\ P(P \cup \omega_0) &= P(D \cup \omega_0) = \frac{19}{37}. \end{aligned}$$

△

Dalle proprietà 3a-3c che caratterizzano una funzione di probabilità, è possibile ricavare alcune interessanti conseguenze.

- $\forall A \in \mathcal{A} \Rightarrow \boxed{P(\overline{A}) = 1 - P(A)}.$

Infatti, dalla 3b, $P(\Omega) = 1$. Ma $\Omega = A \cup \overline{A}$, per cui:

$$1 = P(\Omega) = P(A \cup \overline{A}) = P(A) + P(\overline{A}),$$

in virtù della 3c.

- $\forall A \in \mathcal{A} \Rightarrow \boxed{0 \leq P(A) \leq 1}.$ Tale relazione segue immediatamente dalla proprietà 3a e dalla relazione precedente.

- $\boxed{P(\emptyset) = 0}.$

Infatti, poiché $\overline{\Omega} = \emptyset$, si ha:

$$P(\emptyset) = P(\overline{\Omega}) = 1 - P(\Omega) = 0.$$

- $\forall A_1, A_2 \in \mathcal{A} : \boxed{A_1 \subseteq A_2 \Rightarrow P(A_1) \leq P(A_2)}.$

Infatti, si osservi che, poiché $A_1 \subseteq A_2$ è sempre possibile scrivere

$$A_2 = A_1 \cup (A_2 \cap \overline{A_1}),$$

con $A_1 \cap (A_2 \cap \overline{A_1}) = \emptyset$. Per cui, sfruttando la proprietà 3c, si ha:

$$P(A_2) = P(A_1 \cup (A_2 \cap \overline{A_1})) = P(A_1) + P(A_2 \cap \overline{A_1}).$$

La tesi segue osservando che $P(A_2 \cap \overline{A_1}) \geq 0$ per la 3a.

- $\forall A_1, A_2 \in \mathcal{A} \Rightarrow \boxed{P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)}.$

Innanzitutto, si osservi come è sempre possibile scrivere:

$$A_1 \cup A_2 = A_1 \cup (A_2 \cap \overline{A_1}).$$

con $A_1 \cap (A_2 \cap \overline{A_1}) = \emptyset$. Inoltre, vale l'identità:

$$A_2 \cap \overline{A_1} = (A_2 \cap \overline{A_1}) \cup \emptyset = (A_2 \cap \overline{A_1}) \cup (A_2 \cap \overline{A_2}) = A_2 \cap (\overline{A_1} \cup \overline{A_2}).$$

Quindi, utilizzando la proprietà 3c:

$$\begin{aligned} P(A_1 \cup A_2) &= P(A_1 \cup (A_2 \cap (\overline{A_1} \cup \overline{A_2}))) \\ &= P(A_1) + P(A_2 \cap (\overline{A_1} \cup \overline{A_2})). \end{aligned} \quad (1.3)$$

Ora, notando che:

$$\begin{aligned} A_2 \cap (\overline{A_1} \cup \overline{A_2}) &= \overline{\overline{A_2 \cap (\overline{A_1} \cup \overline{A_2})}} = \overline{\overline{A_2} \cup \overline{(\overline{A_1} \cup \overline{A_2})}} \\ &= \overline{\overline{A_2} \cup (A_1 \cap A_2)}, \end{aligned}$$

si ricava:

$$\begin{aligned} P(A_2 \cap (\overline{A_1} \cup \overline{A_2})) &= P(\overline{\overline{A_2} \cup (A_1 \cap A_2)}) = 1 - P(\overline{A_2} \cup (A_1 \cap A_2)) \\ &\stackrel{[1]}{=} 1 - (P(\overline{A_2}) + P(A_1 \cap A_2)) \\ &= 1 - (1 - P(A_2) + P(A_1 \cap A_2)) \\ &= P(A_2) - P(A_1 \cap A_2), \end{aligned} \quad (1.4)$$

dal momento che $\overline{A_2}$ e $(A_1 \cap A_2)$ sono disgiunti (uguaglianza [1]). Infine, sostituendo la (1.4) nella (1.3), segue la tesi.

Si noti come la funzione di probabilità definita nell'Esempio 1.1 rispetti tutte le precedenti proprietà.

Definizione 1.2. Sia dato lo spazio di probabilità $\{\Omega, \mathcal{A}, P(\cdot)\}$. Due eventi $A, B \in \mathcal{A}$ si dicono *statisticamente indipendenti* se:

$$P(A \cap B) = P(A) \cdot P(B).$$

Se due eventi A e B sono statisticamente indipendenti, la probabilità con la quale essi si verificano contemporaneamente ($P(A \cap B)$) coincide col prodotto delle probabilità con le quali si verificano i singoli eventi ($P(A) \cdot P(B)$).

Definizione 1.3. Sia dato lo spazio di probabilità $\{\Omega, \mathcal{A}, P(\cdot)\}$. Dati due eventi A e B , si definisce *probabilità condizionata di A all'evento B* :

$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)}, \quad (1.5)$$

se l'evento B è tale che $P(B) \neq 0$, altrimenti $P(A|B)$ è indefinita.

$P(A|B)$ indica la nuova probabilità con cui si verifica l'evento A *una volta noto che* si sia già verificato l'evento B .

Esempio 1.2. Si consideri il lancio di un dado. Se il dado non è truccato, tutti i risultati da uno a sei sono equiprobabili. Per cui, indicato con A_i l'evento corrispondente al manifestarsi della faccia del dado contrassegnata dal numero i , si ha:

$$P(A_i) = \frac{1}{6}, \quad i = 1, \dots, 6.$$

Non disponendo di nessun'altra informazione la probabilità con cui ci si attende che il risultato del lancio sia, ad esempio, il numero due è $P(A_2) = \frac{1}{6}$.

Supponiamo, ora, che il dado venga lanciato ed un osservatore dell'esperimento ci informi del fatto che il risultato è un numero pari. Qual è, con questa ulteriore informazione, la probabilità che il risultato sia due? Poiché il risultato è pari, i soli valori ammissibili sono 2,4,6, ciascuno con probabilità $\frac{1}{3}$. Per cui ci aspettiamo che sia:

$$P(A_2 | \text{il risultato è pari}) = \frac{1}{3}.$$

Infatti, si indichi con A_p l'evento corrispondente al manifestarsi di una faccia del dado contrassegnata da un numero pari. Ovviamente, sarà:

$$P(A_p) = \frac{1}{2}.$$

Dalla Definizione 1.3:

$$P(A_2|A_p) = \frac{P(A_2 \cap A_p)}{P(A_p)}.$$

Ma, $A_2 \cap A_p = A_2$, per cui:

$$P(A_2|A_p) = \frac{P(A_2)}{P(A_p)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}.$$

△

Finora, abbiamo incontrato solo spazi di probabilità con un insieme di eventi elementari Ω costituito da un numero finito elementi. Anche in casi semplici come questi, può risultare alquanto laborioso definire uno spazio di probabilità. Inoltre, le funzioni di probabilità, essendo definite su insiemi, non sono facili da trattare. Le cose si complicano ulteriormente nel caso in cui Ω sia un insieme di cardinalità infinita (numerabile o non numerabile). In particolare, si consideri il caso notevole in cui $\Omega = \mathbb{R}$. Come definire la famiglia \mathcal{A} in modo che le proprietà di chiusura siano rispettate? Come definire una funzione di probabilità su \mathcal{A} ?

Il seguente risultato fornisce la risposta alla prima domanda.

Si denoti con $A_\alpha \in \mathbb{R}$ la semiretta:

$$A_\alpha = \{x \in \mathbb{R} | x \leq \alpha\}, \quad \forall \alpha \in \mathbb{R}.$$

Si noti che non è possibile scegliere semplicemente $\mathcal{A} = \{A_\alpha\}$, $\forall \alpha \in \mathbb{R}$ poiché essa non gode delle proprietà di chiusura 2a-2c. Si dimostra che, se si cerca di completare \mathcal{A} affinché valgano le 2a-2c, aggiungendo altri eventi che non siano semirette, si ottiene un'algebra costituita da tutti i sottoinsiemi di \mathbb{R} .² L'algebra così generata dalla semirette A_α è detta *algebra di Borel*.

²In realtà, è possibile costruire degli insiemi "patologici" che non appartengono alla famiglia \mathcal{A} così generata. D'altro canto, tutti i sottoinsiemi di \mathbb{R} di un qualche interesse pratico appartengono ad \mathcal{A} .

Infine, è possibile dimostrare che per assegnare una probabilità a ciascun insieme appartenente all'algebra di Borel, è sufficiente assegnare una probabilità a ciascuna semiretta A_α . Vale a dire basta definire:

$$P(A_\alpha) = P(\{x \in \mathbb{R} | x \leq \alpha\}), \quad \forall \alpha \in \mathbb{R}.$$

1.2 Variabili aleatorie, distribuzioni e densità di probabilità

Definizione 1.4. Una *variabile aleatoria* \mathbf{x} (brevemente v.a.) è una funzione dallo spazio degli eventi elementari Ω in \mathbb{R} :

$$\mathbf{x} : \Omega \rightarrow \mathbb{R}.$$

Il valore ξ assunto dalla variabile aleatoria \mathbf{x} , in corrispondenza dell'evento ω , prende il nome di *realizzazione* della v.a. \mathbf{x} .

Definizione 1.5. Per ciascuna v.a. \mathbf{x} , è possibile definire la *funzione di distribuzione della probabilità* $F_{\mathbf{x}}(\xi)$ indicante la probabilità dell'evento $\{\mathbf{x} \leq \xi\}$:

$$F_{\mathbf{x}}(\xi) \triangleq P(\mathbf{x} \leq \xi),$$

dove $P(\cdot)$ denota la funzione di probabilità associata all'insieme Ω .

La funzione di distribuzione della probabilità gode delle seguenti proprietà:

1. $0 \leq F_{\mathbf{x}}(\xi) \leq 1$;
2. $\lim_{\xi \rightarrow \infty} F_{\mathbf{x}}(\xi) = 1, \quad \lim_{\xi \rightarrow -\infty} F_{\mathbf{x}}(\xi) = 0$;
3. $F_{\mathbf{x}}(\xi)$ è monotona non decrescente.

Si noti che la distribuzione di probabilità di una v.a. non deve necessariamente essere una funzione continua. È sempre possibile, però, esprimerla come la somma di una componente continua ed una discreta (v. Figura 1.1):

$$F_{\mathbf{x}}(\xi) = F_{\mathbf{x}}^c(\xi) + F_{\mathbf{x}}^d(\xi),$$

in cui la parte discreta è una somma di gradini (opportunamente scalati e traslati):

$$F_{\mathbf{x}}^d(\xi) = \sum_{i=1}^m \gamma_i u(\xi - \xi_i),$$

dove

$$u(\xi) = \begin{cases} 1 & \text{se } \xi \geq 0 \\ 0 & \text{se } \xi < 0 \end{cases}.$$

In ciascun punto di discontinuità ξ_i della $F_{\mathbf{x}}(\xi)$, la distribuzione è continua a destra, cioè $\lim_{\xi \rightarrow \xi_i^+} F_{\mathbf{x}}(\xi) = F_{\mathbf{x}}(\xi_i)$.

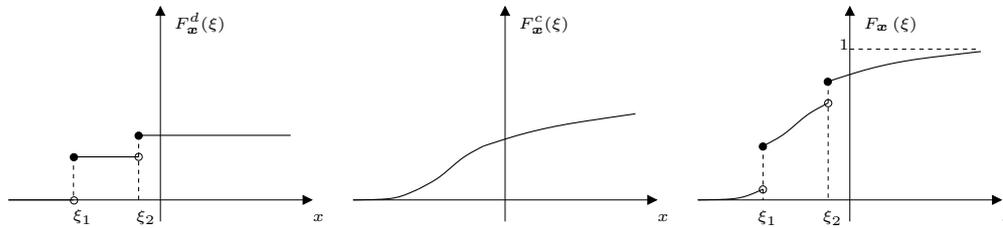


Figura 1.1: Decomposizione di una funzione distribuzione di probabilità nelle sue componenti continue e discrete.

Definizione 1.6. Per ciascuna v.a. \mathbf{x} , è possibile definire la *funzione di densità della probabilità* (brevemente d.d.p.) $f_{\mathbf{x}}(\xi)$ come la derivata della funzione di distribuzione della probabilità:

$$f_{\mathbf{x}}(\xi) \triangleq \frac{d}{d\xi} F_{\mathbf{x}}(\xi).$$

Affinché una funzione $f_{\mathbf{x}}(\xi)$ possa essere una densità di probabilità, essa deve soddisfare le seguenti proprietà:

1. $f_{\mathbf{x}}(\xi) \geq 0$;
2. $\int_{-\infty}^{\infty} f_{\mathbf{x}}(\xi) d\xi = 1$.

Direttamente dalla Definizione 1.6 discende che:

$$F_{\mathbf{x}}(\xi) = \int_{-\infty}^{\xi} f_{\mathbf{x}}(\tau) d\tau.$$

Nel caso in cui la distribuzione della probabilità contenga anche una componente discreta, la corrispondente funzione di densità della probabilità conterrà delle delta di Dirac (opportunamente scalate e traslate), in quanto:

$$\frac{d}{d\xi} F_{\mathbf{x}}^d(\xi) = \sum_{i=1}^m \gamma_i \delta(\xi - \xi_i).$$

Se $f_{\mathbf{x}}(\xi)$ rappresenta la densità di probabilità della v.a. \mathbf{x} , allora vale la seguente proprietà:

$$P(a \leq \mathbf{x} \leq b) = \int_a^b f_{\mathbf{x}}(\xi) d\xi. \quad (1.6)$$

La relazione precedente fornisce un'interpretazione della funzione di densità di probabilità (v. Figura 1.2). La $f_{\mathbf{x}}(\xi)$, infatti, consente di calcolare la probabilità (corrispondente all'area della zona in grigio in Figura 1.2) con la quale una v.a. \mathbf{x} assume valori appartenenti ad un generico intervallo $[a, b]$ (o, più in generale, ad un insieme). Dalla relazione che lega le funzioni di densità e distribuzione, è facile convincersi che tale probabilità è anche uguale alla differenza $F_{\mathbf{x}}(b) - F_{\mathbf{x}}(a)$. Inoltre, supponendo che la densità di probabilità non contenga delle delta di Dirac, la $P(\mathbf{x} = a) = 0$, per qualunque valore $a \in \mathbb{R}$. Infatti, il dominio di integrazione della $f_{\mathbf{x}}(\xi)$ ha, in questo caso, misura nulla. Graficamente, l'area della zona in grigio in Figura 1.2 tende a zero, quando b tende ad a . Quindi, la $f_{\mathbf{x}}(\xi)$ non corrisponde alla probabilità che $\mathbf{x} = \xi$. Essa indica solamente in quali "zone" dell'asse reale è più o meno probabile che la \mathbf{x} si manifesti.

Esempio 1.3. Variabile aleatoria gaussiana. Una funzione di densità di probabilità del tipo:

$$f_{\mathbf{x}}(\xi) = \frac{1}{\sqrt{2\pi}\sigma_{\mathbf{x}}} e^{-\frac{(\xi - m_{\mathbf{x}})^2}{2\sigma_{\mathbf{x}}^2}} \quad \sigma_{\mathbf{x}} > 0, m_{\mathbf{x}} \in \mathbb{R}, \quad (1.7)$$

è detta densità *gaussiana*.

Spesso, per indicare il fatto che una v.a. \mathbf{x} ha una densità di probabilità gaussiana come la (1.7), si scrive $\mathbf{x} \sim N(m_{\mathbf{x}}, \sigma_{\mathbf{x}}^2)$ (e la stessa v.a. è detta gaussiana o "normale"). \triangle

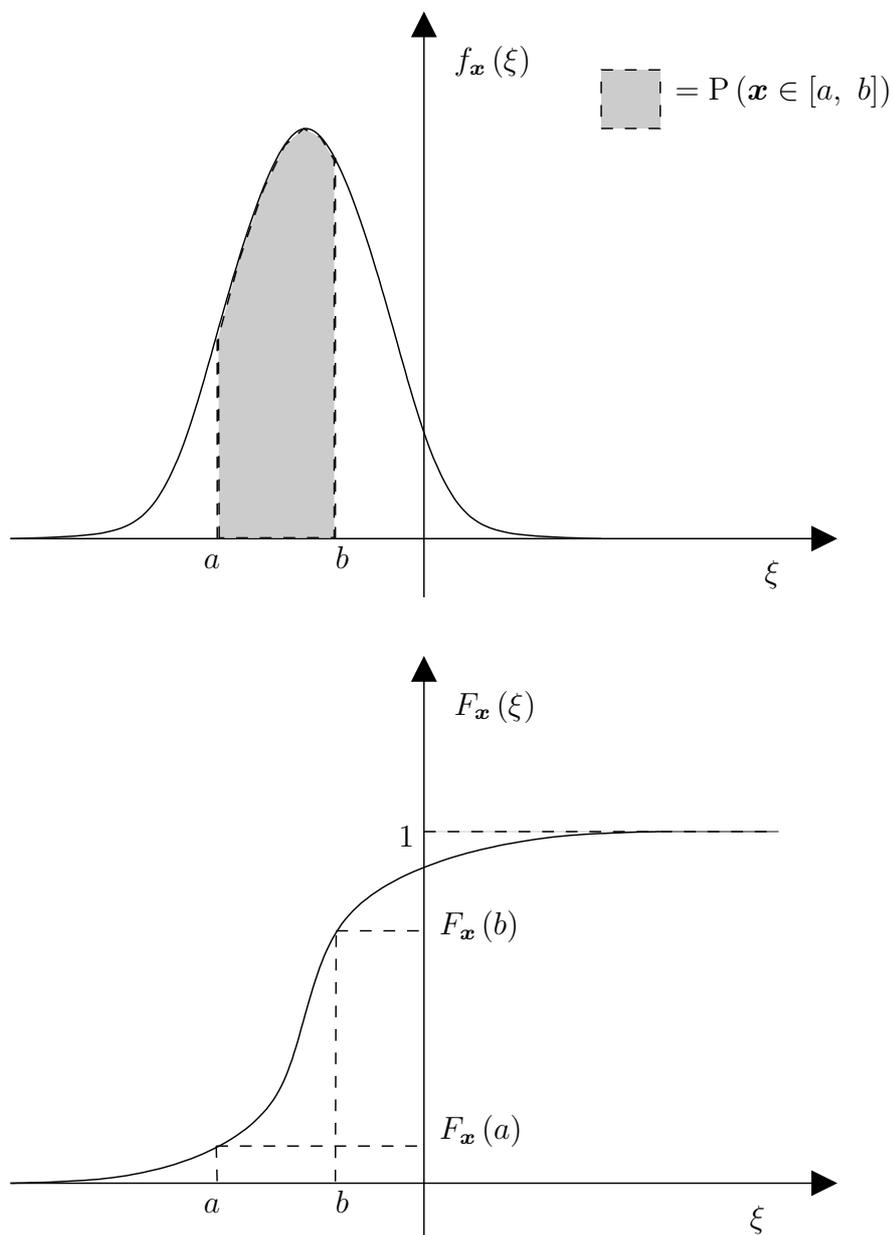


Figura 1.2: Interpretazione delle funzioni densità e distribuzione di probabilità.

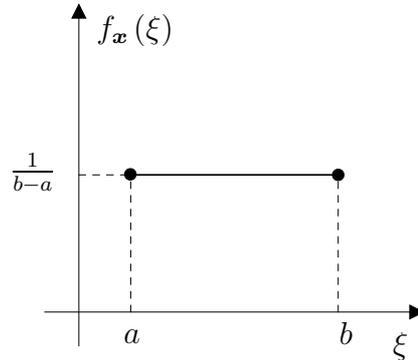


Figura 1.3: Densità di probabilità uniforme $U[a, b]$.

Esempio 1.4. Variabile aleatoria distribuita uniformemente. Una funzione di densità di probabilità del tipo:

$$f_{\mathbf{x}}(\xi) = \begin{cases} 0 & \text{se } \xi < a \\ \frac{1}{b-a} & \text{se } a \leq \xi \leq b \\ 0 & \text{se } \xi > b \end{cases}$$

è detta densità *uniforme* (v. Figura 1.3). Una v.a. \mathbf{x} avente tale densità di probabilità è detta *uniformemente distribuita* nell'intervallo $[a, b]$ e ciò si indica con $\mathbf{x} \sim U[a, b]$. △

Definizione 1.7. Una v.a. \mathbf{x} si dice *discreta* se essa può assumere un insieme numerabile $\{\xi_1, \xi_2, \dots\}$ (non necessariamente finito) di valori. In questo caso la sua densità di probabilità è data da una sommatoria (eventualmente infinita) di delta di Dirac:

$$f_{\mathbf{x}}(\xi) = \sum_{i=1}^{\infty} p_i \delta(\xi - i),$$

con $\sum_{i=1}^{\infty} p_i = 1$

Si consideri, ad esempio, una v.a. \mathbf{x} avente d.d.p.:

$$f_{\mathbf{x}}(\xi) = \frac{1}{6} \sum_{i=1}^6 \delta(\xi - i).$$

L'esito di quale esperimento può rappresentare la v.a. \mathbf{x} ?

1.3 Distribuzioni bivariate e multivariate

Siano \mathbf{x} e \mathbf{y} due variabili aleatorie.

Definizione 1.8. Si definisce funzione di *distribuzione di probabilità congiunta* (o *bivariata*) $F_{\mathbf{x},\mathbf{y}}(\xi, \eta)$ la probabilità dell'evento congiunto $\{\mathbf{x} \leq \xi, \mathbf{y} \leq \eta\}$:

$$F_{\mathbf{x},\mathbf{y}}(\xi, \eta) \triangleq P(\mathbf{x} \leq \xi, \mathbf{y} \leq \eta).$$

Definizione 1.9. Si definisce funzione di *densità di probabilità congiunta* (o *bivariata*) $f_{\mathbf{x},\mathbf{y}}(\xi, \eta)$ la derivata seconda mista della distribuzione di probabilità congiunta:

$$f_{\mathbf{x},\mathbf{y}}(\xi, \eta) \triangleq \frac{\partial^2 F_{\mathbf{x},\mathbf{y}}(\xi, \eta)}{\partial \xi \partial \eta}.$$

Direttamente dalla Definizione 1.9 discende che:

$$F_{\mathbf{x},\mathbf{y}}(\xi, \eta) = \int_{-\infty}^{\eta} \int_{-\infty}^{\xi} f_{\mathbf{x},\mathbf{y}}(\tau_1, \tau_2) d\tau_1 d\tau_2. \quad (1.8)$$

Una distribuzione di probabilità congiunta soddisfa le seguenti proprietà:

1. $F_{\mathbf{x},\mathbf{y}}(\xi, \eta)$ è monotona non decrescente in ξ e η ;
2. $\lim_{\substack{\xi \rightarrow -\infty \\ \eta \rightarrow -\infty}} F_{\mathbf{x},\mathbf{y}}(\xi, \eta) = 0, \quad \lim_{\substack{\xi \rightarrow +\infty \\ \eta \rightarrow +\infty}} F_{\mathbf{x},\mathbf{y}}(\xi, \eta) = 1.$

Inoltre, affinché una funzione $f_{\mathbf{x},\mathbf{y}}(\xi, \eta)$ possa essere una densità di probabilità congiunta, essa deve soddisfare le seguenti proprietà:

1. $f_{\mathbf{x},\mathbf{y}}(\xi, \eta) \geq 0$;
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{\mathbf{x},\mathbf{y}}(\xi, \eta) d\xi d\eta = 1.$

Analogamente alla (1.6), una densità di probabilità congiunta soddisfa la seguente proprietà:

$$P((\mathbf{x}, \mathbf{y}) \in A) = \int \int_A f_{\mathbf{x}, \mathbf{y}}(\xi, \eta) d\xi d\eta, \quad (1.9)$$

in cui A denota un generico insieme in \mathbb{R}^2 .

Si noti che la densità (o, equivalentemente, la distribuzione) congiunta caratterizza completamente il comportamento probabilistico delle due v.a. Infatti, nota la $f_{\mathbf{x}, \mathbf{y}}(\xi, \eta)$ è sempre possibile calcolare le singole densità marginali:

$$f_{\mathbf{x}}(\xi) = \int_{-\infty}^{\infty} f_{\mathbf{x}, \mathbf{y}}(\xi, \eta) d\eta; \quad (1.10)$$

$$f_{\mathbf{y}}(\eta) = \int_{-\infty}^{\infty} f_{\mathbf{x}, \mathbf{y}}(\xi, \eta) d\xi. \quad (1.11)$$

Definizione 1.10. Due v.a. \mathbf{x} ed \mathbf{y} si dicono *statisticamente indipendenti* se la loro densità di probabilità congiunta è uguale al prodotto delle rispettive densità di probabilità marginali:

$$f_{\mathbf{x}, \mathbf{y}}(\xi, \eta) = f_{\mathbf{x}}(\xi) f_{\mathbf{y}}(\eta).$$

Esempio 1.5. Una funzione di densità di probabilità bivariata del tipo:

$$f_{\mathbf{x}, \mathbf{y}}(\xi, \eta) = \frac{1}{2\pi\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)} \left\{ \frac{(\xi-m_{\mathbf{x}})^2}{\sigma_{\mathbf{x}}^2} - \frac{2\rho(\xi-m_{\mathbf{x}})(\eta-m_{\mathbf{y}})}{\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}} + \frac{(\eta-m_{\mathbf{y}})^2}{\sigma_{\mathbf{y}}^2} \right\}},$$

con $\sigma_{\mathbf{x}} > 0$, $\sigma_{\mathbf{y}} > 0$, $|\rho| < 1$, $m_{\mathbf{x}}, m_{\mathbf{y}} \in \mathbb{R}$, è detta densità *gaussiana bivariata*. △

Esempio 1.6. La funzione:

$$f_{\mathbf{x}, \mathbf{y}}(\xi, \eta) = \begin{cases} \frac{6}{7}(\xi + \eta)^2 & 0 \leq \xi \leq 1, \quad 0 \leq \eta \leq 1 \\ 0 & \text{altrove} \end{cases}$$

rappresenta una densità di probabilità bivariata. △

Le nozioni di distribuzione e densità bivariate possono essere generalizzate al caso di n variabili aleatorie, $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$. Per compattezza di notazione, si indichi con \mathbf{x} la variabile aleatoria *vettoriale*:

$$\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)^T,$$

e con ξ un generico vettore di \mathbb{R}^n :

$$\xi = (\xi_1, \xi_2, \dots, \xi_n)^T,$$

in cui $(\cdot)^T$ denota l'operatore di trasposizione.

Definizione 1.11. Si definisce funzione di *distribuzione di probabilità congiunta* (o *multivariata*) $F_{\mathbf{x}}(\xi)$ delle v.a. $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, la probabilità dell'evento congiunto $\{\mathbf{x}_1 \leq \xi_1, \mathbf{x}_2 \leq \xi_2 \dots \mathbf{x}_n \leq \xi_n\}$:

$$F_{\mathbf{x}}(\xi) = F_{\mathbf{x}_1, \dots, \mathbf{x}_n}(\xi_1, \dots, \xi_n) \triangleq P(\mathbf{x}_1 \leq \xi_1, \dots, \mathbf{x}_n \leq \xi_n).$$

Definizione 1.12. Si definisce funzione di *densità di probabilità congiunta* (o *multivariata*) $f_{\mathbf{x}}(\xi)$ delle v.a. $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, la derivata n -esima mista della distribuzione di probabilità congiunta:

$$f_{\mathbf{x}}(\xi) = f_{\mathbf{x}_1, \dots, \mathbf{x}_n}(\xi_1, \dots, \xi_n) \triangleq \frac{\partial^n F_{\mathbf{x}_1, \dots, \mathbf{x}_n}(\xi_1, \dots, \xi_n)}{\partial \xi_1 \dots \partial \xi_n}.$$

Direttamente dalla Definizione 1.12 discende che:

$$F_{\mathbf{x}_1, \dots, \mathbf{x}_n}(\xi_1, \dots, \xi_n) = \int_{-\infty}^{\xi_n} \dots \int_{-\infty}^{\xi_1} f_{\mathbf{x}_1, \dots, \mathbf{x}_n}(\tau_1, \dots, \tau_n) d\tau_1 \dots d\tau_n. \quad (1.12)$$

Tutte le proprietà illustrate in precedenza per le distribuzioni e le densità bivariate valgono allo stesso modo per le distribuzioni e le densità multivariate, così come le relazioni esistenti fra distribuzioni e densità bivariate e le corrispondenti funzioni marginali.

Notazione. D'ora in avanti, si utilizzerà lo stesso stile tipografico per denotare indifferentemente v.a. scalari o vettoriali, le dimensioni risultando chiare dal contesto.

1.4 Media e varianza

Definizione 1.13. Si definisce *valor medio* (o *valore atteso*) $m_{\mathbf{x}}$ della v.a. \mathbf{x} la quantità:

$$\mathbf{E}[\mathbf{x}] \triangleq \int_{-\infty}^{+\infty} \xi f_{\mathbf{x}}(\xi) d\xi = m_{\mathbf{x}}. \quad (1.13)$$

Il simbolo $\mathbf{E}[\cdot]$ denota l'*operatore di valore atteso*. Esso gode della seguente proprietà di *linearità*:

$$\mathbf{E}[\alpha\mathbf{x} + \beta\mathbf{y}] = \alpha\mathbf{E}[\mathbf{x}] + \beta\mathbf{E}[\mathbf{y}] \quad \forall \alpha, \beta \in \mathbb{R}. \quad (1.14)$$

Sia $g(\mathbf{x})$ una funzione deterministica di una generica v.a. \mathbf{x} . Per definizione:

$$\mathbf{E}[g(\mathbf{x})] \triangleq \int_{-\infty}^{+\infty} g(\xi) f_{\mathbf{x}}(\xi) d\xi.$$

Nel caso in cui $g(\mathbf{x}) = \mathbf{x}^p$, con p numero intero positivo, $\mathbf{E}[\mathbf{x}^p]$ si chiama *momento di ordine p* . Analogamente, $\mathbf{E}[(\mathbf{x} - m_{\mathbf{x}})^p]$ è detto *momento centrale di ordine p* .

Particolare interesse riveste il momento centrale del secondo ordine.

Definizione 1.14. Si definisce *varianza* $\sigma_{\mathbf{x}}^2$ della v.a. \mathbf{x} la quantità:

$$\sigma_{\mathbf{x}}^2 \triangleq \mathbf{E}[(\mathbf{x} - m_{\mathbf{x}})^2] = \int_{-\infty}^{+\infty} (\xi - m_{\mathbf{x}})^2 f_{\mathbf{x}}(\xi) d\xi. \quad (1.15)$$

Si noti che dalla Definizione 1.14 discende che $\sigma_{\mathbf{x}}^2 \geq 0$, in quanto la funzione integranda nella (1.15) è sempre maggiore o uguale a zero. Il fatto che $\sigma_{\mathbf{x}}^2 = 0$ implica che la v.a. \mathbf{x} può assumere un solo valore (che coincide con $m_{\mathbf{x}}$) e risulta essere, quindi, una grandezza deterministica.

Quando la varianza $\sigma_{\mathbf{x}}^2$ di una v.a. \mathbf{x} esiste, la sua radice quadrata $\sigma_{\mathbf{x}}$ viene chiamata *deviazione standard*.

Definizione 1.15. Si definisce *covarianza* o *varianza incrociata* $\sigma_{\mathbf{x},\mathbf{y}}$ delle v.a. \mathbf{x}, \mathbf{y} la quantità:

$$\sigma_{\mathbf{x},\mathbf{y}} \triangleq \mathbf{E}[(\mathbf{x} - m_{\mathbf{x}})(\mathbf{y} - m_{\mathbf{y}})].$$

È opportuno notare che la Definizione 1.15 coinvolge il valore atteso di una funzione deterministica $g(\mathbf{x}, \mathbf{y})$ di due v.a. \mathbf{x}, \mathbf{y} . In questo caso, nel calcolo di $\mathbf{E}[g(\mathbf{x}, \mathbf{y})]$ occorre usare la densità di probabilità congiunta delle v.a. \mathbf{x}, \mathbf{y} :

$$\mathbf{E}[g(\mathbf{x}, \mathbf{y})] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(\xi, \eta) f_{\mathbf{x}, \mathbf{y}}(\xi, \eta) d\xi d\eta.$$

Definizione 1.16. Due v.a. \mathbf{x}, \mathbf{y} si dicono *incorrelate* se:

$$\mathbf{E}[\mathbf{x}\mathbf{y}] = \mathbf{E}[\mathbf{x}] \mathbf{E}[\mathbf{y}].$$

È immediato verificare che, se due v.a. sono incorrelate, la loro varianza incrociata è uguale a zero.

Definizione 1.17. Si definisce *indice (o coefficiente) di correlazione* $\rho_{\mathbf{x}, \mathbf{y}}$ delle v.a. \mathbf{x}, \mathbf{y} la quantità:

$$\rho_{\mathbf{x}, \mathbf{y}} \triangleq \frac{\mathbf{E}[(\mathbf{x} - m_{\mathbf{x}})(\mathbf{y} - m_{\mathbf{y}})]}{\sqrt{\mathbf{E}[(\mathbf{x} - m_{\mathbf{x}})^2] \mathbf{E}[(\mathbf{y} - m_{\mathbf{y}})^2]}} = \frac{\sigma_{\mathbf{x}, \mathbf{y}}}{\sigma_{\mathbf{x}} \sigma_{\mathbf{y}}}.$$

Il coefficiente di correlazione è sempre compreso tra -1 ed 1:

$$-1 \leq \rho_{\mathbf{x}, \mathbf{y}} \leq 1.$$

Se le v.a. \mathbf{x} e \mathbf{y} sono incorrelate $\rho_{\mathbf{x}, \mathbf{y}} = 0$.

Se \mathbf{x} ed \mathbf{y} sono due v.a. indipendenti allora esse sono anche incorrelate. L'implicazione inversa non è vera, in generale. Vale a dire, due v.a. possono essere incorrelate ma non indipendenti.

Nel caso in cui \mathbf{x} e \mathbf{y} siano due v.a. *vettoriali*, le Definizioni 1.14 e 1.15 vanno modificate come segue.

Definizione 1.18. Si definisce *matrice di covarianza* $P_{\mathbf{x}}$ della v.a. $\mathbf{x} \in \mathbb{R}^n$ la quantità:

$$P_{\mathbf{x}} \triangleq \mathbf{E}[(\mathbf{x} - m_{\mathbf{x}})(\mathbf{x} - m_{\mathbf{x}})^T].$$

Definizione 1.19. Si definisce *matrice di cross-covarianza* o *matrice di varianza incrociata* $P_{\mathbf{x}\mathbf{y}}$ delle v.a. $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$ la quantità:

$$P_{\mathbf{x}\mathbf{y}} \triangleq \mathbf{E}[(\mathbf{x} - m_{\mathbf{x}})(\mathbf{y} - m_{\mathbf{y}})^T].$$

Si noti che i valori medi sono vettori (colonna) di dimensione $m_{\mathbf{x}} \in \mathbb{R}^n$, $m_{\mathbf{y}} \in \mathbb{R}^m$, mentre le covarianze sono matrici di dimensione $P_{\mathbf{x}} \in \mathbb{R}^{n \times n}$, $P_{\mathbf{xy}} \in \mathbb{R}^{n \times m}$.

1.5 Variabili aleatorie gaussiane

Una v.a. vettoriale $\mathbf{x} \in \mathbb{R}^n$ si dice gaussiana (o normale) se la sua densità di probabilità ha la forma:

$$f_{\mathbf{x}}(\xi) = \frac{1}{(2\pi)^{n/2}(\det P_{\mathbf{x}})^{1/2}} e^{-\frac{1}{2}(\xi - m_{\mathbf{x}})^T P_{\mathbf{x}}^{-1}(\xi - m_{\mathbf{x}})}, \quad (1.16)$$

con $m_{\mathbf{x}} \in \mathbb{R}^n$ e $P_{\mathbf{x}}$ matrice quadrata $n \times n$, simmetrica e definita positiva³: $P_{\mathbf{x}} = P_{\mathbf{x}}^T > 0$. Spesso, per indicare una v.a. avente la d.d.p. precedente si scrive $\mathbf{x} \sim N(m_{\mathbf{x}}, P_{\mathbf{x}})$.

Si noti che, nel caso $n = 2$, posto $P_{\mathbf{x}} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$ si ottiene la distribuzione bivariata vista nell'Esempio 1.5.

Le v.a. gaussiane godono delle seguenti proprietà.

1. $\mathbf{E}[\mathbf{x}] = m_{\mathbf{x}}$;
2. $\mathbf{E}[(\mathbf{x} - m_{\mathbf{x}})(\mathbf{x} - m_{\mathbf{x}})^T] = P_{\mathbf{x}}$;
3. \mathbf{x} gaussiana $\Rightarrow \mathbf{y} = A\mathbf{x} + b$ gaussiana, con:

$$m_{\mathbf{y}} = Am_{\mathbf{x}} + b,$$

$$P_{\mathbf{y}} = AP_{\mathbf{x}}A^T;$$

4. \mathbf{x} e \mathbf{y} incorrelate e gaussiane $\Rightarrow \mathbf{x}$ e \mathbf{y} indipendenti.

Qualunque combinazione lineare affine di v.a. congiuntamente gaussiane è ancora una v.a. gaussiana (Proprietà 3). Inoltre, per due v.a. gaussiane l'incorrelazione coincide con l'indipendenza statistica (Proprietà 4).

³Una matrice simmetrica P è detta definita positiva (d.p.) se $x^T P x > 0, \forall x \neq 0$. Una matrice d.p. è sempre invertibile.

Le v.a. gaussiane rivestono un ruolo importante nella teoria della stima, in quanto consentono il calcolo di stime ottime in maniera agevole.

Il seguente Teorema fornisce una giustificazione teorica al largo uso che viene fatto delle v.a. gaussiane nella pratica.

Teorema del limite centrale (Lyapunov). *Siano $\mathbf{x}_1, \dots, \mathbf{x}_n$ v.a. indipendenti, con distribuzione di probabilità arbitraria. Siano $\mathbf{E}[\mathbf{x}_i] = m_i$ e $\mathbf{E}[(\mathbf{x}_i - m_i)^2] = \sigma_i^2 < +\infty$. Si consideri la v.a. normalizzata:*

$$\mathbf{Z}_n = \frac{\sum_{i=1}^n \mathbf{x}_i - \sum_{i=1}^n m_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}.$$

Sotto opportune condizioni (ipotesi molto generali, vedi Rohatgi and Saleh (2001)), vale

$$\lim_{n \rightarrow \infty} F_{\mathbf{Z}_n}(\eta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\eta} e^{-\frac{t^2}{2}} dt.$$

△

Osservazione 1.1. Nel caso in cui le v.a. \mathbf{x}_i abbiano la stessa media e varianza: $m_i = m_{\mathbf{x}}$ e $\sigma_i^2 = \sigma_{\mathbf{x}}^2$, $i = 1, \dots, n$ la v.a. \mathbf{Z}_n assume la forma:

$$\mathbf{Z}_n = \frac{\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i - m_{\mathbf{x}}}{\sigma_{\mathbf{x}} / \sqrt{n}}.$$

Per cui, si usa dire che la v.a. $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$ risulta essere *asintoticamente* distribuita come $N(m_{\mathbf{x}}, \sigma_{\mathbf{x}}^2/n)$. Il Teorema del limite centrale afferma quindi che la distribuzione della somma di un numero sufficientemente grande di v.a. indipendenti può essere ben approssimata da una gaussiana. Quando un fenomeno incerto è dovuto ad un elevato numero di cause indipendenti, quindi, è ragionevole assumere che la sua distribuzione sia gaussiana.

L'andamento della densità di probabilità di una v.a. gaussiana scalare \mathbf{x} con valor medio $m_{\mathbf{x}} = 0$ e varianza $\sigma_{\mathbf{x}}^2 = 1$ (v. equazione (1.7)) è rappresentato in Figura 1.4. Il grafico ha una tipica forma a “campana”, assume il suo valore massimo in corrispondenza del valor medio ed è simmetrico rispetto ad esso.

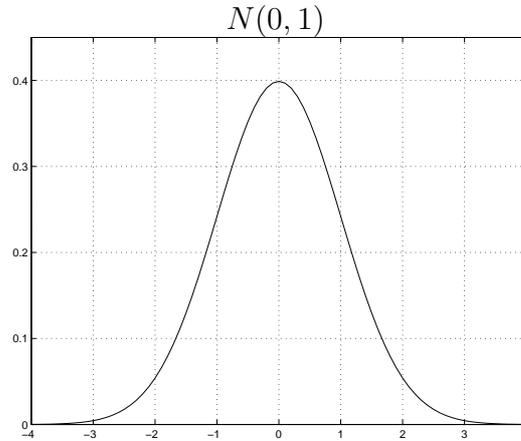


Figura 1.4: Densità di probabilità gaussiana a media nulla e varianza unitaria.

Qual è la probabilità con la quale una v.a. $\mathbf{x} \sim N(m_{\mathbf{x}}, \sigma_{\mathbf{x}}^2)$ assume valori compresi in un intervallo centrato nel suo valor medio e di semiampiezza pari alla sua deviazione standard?

Ricordando il significato di funzione di densità di probabilità (v. equazione (1.6)), si ha:

$$\begin{aligned}
 P(m_{\mathbf{x}} - \sigma_{\mathbf{x}} \leq \mathbf{x} \leq m_{\mathbf{x}} + \sigma_{\mathbf{x}}) &\triangleq \int_{m_{\mathbf{x}} - \sigma_{\mathbf{x}}}^{m_{\mathbf{x}} + \sigma_{\mathbf{x}}} \frac{1}{\sqrt{2\pi}\sigma_{\mathbf{x}}} e^{-\frac{(\xi - m_{\mathbf{x}})^2}{2\sigma_{\mathbf{x}}^2}} d\xi \\
 &\stackrel{[1]}{=} 2 \int_{m_{\mathbf{x}}}^{m_{\mathbf{x}} + \sigma_{\mathbf{x}}} \frac{1}{\sqrt{2\pi}\sigma_{\mathbf{x}}} e^{-\frac{(\xi - m_{\mathbf{x}})^2}{2\sigma_{\mathbf{x}}^2}} d\xi \\
 &\stackrel{[2]}{=} 2 \int_0^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{\eta^2}{2}} d\eta \\
 &\triangleq \operatorname{erf}\left(\frac{1}{\sqrt{2}}\right) = 0.6827.
 \end{aligned}$$

L'uguaglianza [1] è dovuta alla simmetria della d.d.p. gaussiana rispetto al suo valor medio, mentre la [2] si ottiene col cambio di variabile $\eta = (\xi - m_{\mathbf{x}})/\sigma_{\mathbf{x}}$. La funzione $\operatorname{erf}(\cdot)$ (detta “funzione errore”) è definita come⁴:

$$\operatorname{erf}(\xi) = \frac{2}{\sqrt{\pi}} \int_0^{\xi} e^{-t^2} dt.$$

⁴Questa definizione corrisponde alla funzione `erf` implementata in MATLAB[®]; in altri testi si trovano definizioni leggermente diverse della funzione errore.

È possibile verificare che $\lim_{\xi \rightarrow +\infty} \text{erf}(\xi) = 1$.

Dunque, abbiamo visto che se \mathbf{x} è una v.a. gaussiana, essa assume valori compresi in un intervallo (centrato in $m_{\mathbf{x}}$) di semiampiezza pari a $\sigma_{\mathbf{x}}$ con una probabilità di poco superiore al 68%. Analogamente, se consideriamo intervalli di semiampiezza pari a $2\sigma_{\mathbf{x}}$ e $3\sigma_{\mathbf{x}}$, la probabilità aumenta rispettivamente al 95.45% e 99.73%:

$$P(\mathbf{x} \in [m_{\mathbf{x}} - 2\sigma_{\mathbf{x}}, m_{\mathbf{x}} + 2\sigma_{\mathbf{x}}]) = 0.9545;$$

$$P(\mathbf{x} \in [m_{\mathbf{x}} - 3\sigma_{\mathbf{x}}, m_{\mathbf{x}} + 3\sigma_{\mathbf{x}}]) = 0.9973.$$

In MATLAB[®], vedi comandi: `normpdf` (calcolare il valore di una d.d.p. gaussiana), `erf` (calcolare il valore di $\text{erf}(x)$).

1.6 Funzioni di una o più variabili aleatorie

Sia \mathbf{x} una v.a. scalare avente d.d.p. $f_{\mathbf{x}}(\xi)$ nota. Si consideri la v.a. $\mathbf{y} = g(\mathbf{x})$, ottenuta trasformando la v.a. \mathbf{x} attraverso una *funzione deterministica* $g(\cdot)$. Quanto vale la densità di probabilità $f_{\mathbf{y}}(\eta)$ della nuova v.a. \mathbf{y} ? Il seguente Teorema fornisce la risposta al precedente quesito.

Teorema 1. *Sia \mathbf{x} una v.a. scalare con d.d.p. $f_{\mathbf{x}}(\xi)$. La densità di probabilità della v.a. $\mathbf{y} = g(\mathbf{x})$ vale:*

$$f_{\mathbf{y}}(\eta) = \sum_{i=1}^m \frac{f_{\mathbf{x}}(\xi_i)}{|g'(\xi_i)|}, \quad (1.17)$$

dove $g'(\xi_i)$ indica la derivata di g rispetto ad ξ :

$$g'(\xi_i) = \left. \frac{dg(\xi)}{d\xi} \right|_{\xi=\xi_i}$$

calcolata nei punti ξ_i tali che $g(\xi_1) = g(\xi_2) = \dots = g(\xi_m) = \eta$. La derivata $g'(\xi)$ va valutata in tutti i punti ξ_i soluzioni dell'equazione:

$$\eta = g(\xi), \quad (1.18)$$

per y fissato. In tutti gli $\bar{\eta}$ per cui l'equazione (1.18) non ammette soluzione si ha:

$$f_{\mathbf{y}}(\bar{\eta}) \triangleq 0.$$

△

Esempio 1.7. Si supponga di poter rappresentare la resistenza di un resistore mediante una v.a. \mathbf{R} uniformemente distribuita in $[9\Omega, 11\Omega]$. Si desidera calcolare la densità di probabilità della conduttanza $\mathbf{G} = \frac{1}{\mathbf{R}}$ del resistore.

Sia $f_{\mathbf{G}}(\gamma)$ la d.d.p. incognita. Ovviamente, se $\gamma > \frac{1}{9}$ o $\gamma < \frac{1}{11}$ la d.d.p. è nulla, in quanto la resistenza \mathbf{R} può assumere solo valori compresi fra 9Ω e 11Ω :

$$f_{\mathbf{G}}(\gamma) = \begin{cases} \star & \text{se } \frac{1}{11} \leq \gamma \leq \frac{1}{9} \\ 0 & \text{se } \gamma < \frac{1}{11} \text{ o } \gamma > \frac{1}{9} \end{cases}$$

Procediamo al calcolo di \star , utilizzando il Teorema 1. Nel caso in esame, la funzione $g(\cdot)$ vale $g(\xi) = \frac{1}{\xi}$. Quindi, fissato $\gamma \in [\frac{1}{11}, \frac{1}{9}]$, l'equazione (1.18) $\gamma = g(\xi)$ ammette un'unica soluzione $\xi = \frac{1}{\gamma}$. Derivando, si ha

$$g'(\frac{1}{\gamma}) = \left. \frac{dg(\xi)}{d\xi} \right|_{\xi=\frac{1}{\gamma}} = -\frac{1}{\xi^2} \Big|_{\xi=\frac{1}{\gamma}} = -\gamma^2.$$

Poiché $\gamma \in [\frac{1}{11}, \frac{1}{9}]$, sarà $\xi \in [9, 11]$, per cui $f_{\mathbf{R}}(x) = \frac{1}{2}$. Allora:

$$\star = \frac{f_{\mathbf{R}}(\frac{1}{\gamma})}{\left|g'(\frac{1}{\gamma})\right|} = \frac{1}{2\gamma^2}.$$

Quindi

$$f_{\mathbf{G}}(\gamma) = \begin{cases} \frac{1}{2\gamma^2} & \text{se } \frac{1}{11} \leq \gamma \leq \frac{1}{9} \\ 0 & \text{se } \gamma < \frac{1}{11} \text{ o } \gamma > \frac{1}{9} \end{cases}$$

△

Consideriamo, ora, funzioni deterministiche di più v.a. Ad esempio, supponiamo di voler calcolare la d.d.p. della somma di due v.a. $\mathbf{z} = \mathbf{x} + \mathbf{y}$. Nota la $f_{\mathbf{x},\mathbf{y}}(\xi, \eta)$, si desidera calcolare la $f_{\mathbf{z}}(\zeta)$.

Ricordando la definizione di funzione di distribuzione di probabilità, fissato $\zeta \in \mathbb{R}$ si ha:

$$F_{\mathbf{z}}(\zeta) = P(\mathbf{z} \leq \zeta) = P(\mathbf{x} + \mathbf{y} \leq \zeta). \quad (1.19)$$

Sfruttando la Proprietà (1.9), la (1.19) diventa (v. Figura 1.5):

$$F_{\mathbf{z}}(\zeta) = \int_{-\infty}^{+\infty} \int_{-\infty}^{\zeta-\eta} f_{\mathbf{x},\mathbf{y}}(\xi, \eta) d\xi d\eta. \quad (1.20)$$

Infine, derivando la (1.20) rispetto a ζ si ottiene proprio la $f_{\mathbf{z}}(\zeta)$:

$$f_{\mathbf{z}}(\zeta) = \int_{-\infty}^{+\infty} f_{\mathbf{x},\mathbf{y}}(\zeta - \eta, \eta) d\eta. \quad (1.21)$$

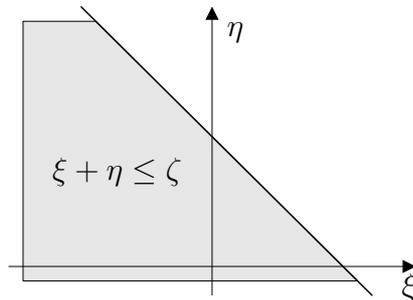


Figura 1.5: Dominio di integrazione della (1.20) per ζ fissato.

Nel caso (notevole) in cui le v.a. \mathbf{x} e \mathbf{y} siano statisticamente indipendenti (cioè $f_{\mathbf{x},\mathbf{y}}(\xi, \eta) = f_{\mathbf{x}}(\xi) f_{\mathbf{y}}(\eta)$), la (1.21) diventa:

$$f_{\mathbf{z}}(\zeta) = \int_{-\infty}^{+\infty} f_{\mathbf{x}}(\zeta - \eta) f_{\mathbf{y}}(\eta) d\eta = f_{\mathbf{x}} * f_{\mathbf{y}}. \quad (1.22)$$

Il risultato (1.22) stabilisce che la densità di probabilità della somma di due v.a. indipendenti è data dalla convoluzione delle singole densità di probabilità.

Vediamo ora come si generalizza il Teorema 1 al caso di funzioni di variabili aleatorie vettoriali.

Teorema 2. Sia $\mathbf{x} \in \mathbb{R}^n$ una v.a. con d.d.p. $f_{\mathbf{x}}(\xi)$ e sia $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ una funzione deterministica. La densità di probabilità della v.a. $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{y} = g(\mathbf{x})$ vale:

$$f_{\mathbf{y}}(\eta) = \sum_{i=1}^m \frac{f_{\mathbf{x}}(\xi_i)}{|J(\xi_i)|}, \quad (1.23)$$

dove $J(\xi_i)$ indica il determinante dello Jacobiano di g rispetto a ξ :

$$J(\xi_i) = \det \left. \frac{\partial g(\xi)}{\partial \xi} \right|_{\xi=\xi_i}$$

calcolato nei punti ξ_i tali che $g(\xi_1) = g(\xi_2) = \dots = g(\xi_m) = \eta$. Lo Jacobiano $J(\xi_i)$ va valutato in tutti i punti ξ_i soluzioni dell'equazione:

$$\eta = g(\xi) \quad (1.24)$$

per y fissato. In tutti gli $\bar{\eta}$ per cui l'equazione (1.24) non ammette soluzione si ha

$$f_{\mathbf{y}}(\bar{\eta}) \triangleq 0.$$

△

Esempio 1.8. Vogliamo mostrare che una combinazione lineare di v.a. congiuntamente gaussiane è ancora una v.a. gaussiana (v. Proprietà 3 a pag. 18).

Per questo, sia $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^T$ una v.a. vettoriale gaussiana con media $m_{\mathbf{x}}$ e matrice di covarianza $P_{\mathbf{x}}$. Consideriamo la nuova v.a. $\mathbf{y} \in \mathbb{R}^n$ ottenuta dalla \mathbf{x} di partenza attraverso una trasformazione lineare invertibile:

$$\mathbf{y} = A\mathbf{x},$$

dove A denota una generica matrice $n \times n$ invertibile.

Per il calcolo di $f_{\mathbf{y}}(\eta)$ utilizziamo il risultato del Teorema 2. Nel caso in esame, la funzione $g(\cdot)$ vale:

$$g(\mathbf{x}) = A\mathbf{x}.$$

Fissato $\eta \in \mathbb{R}^n$, l'equazione (1.24) diventa:

$$\eta = A\xi,$$

e, dal momento che la matrice A è invertibile, ammette un'unica soluzione $\xi = A^{-1}\eta$. Inoltre, lo Jacobiano della $g(\xi)$ è costante (non dipende dal particolare valore ξ) e vale:

$$\frac{\partial g(\xi)}{\partial \xi} = A.$$

Utilizzando la (1.23), si ottiene:

$$f_{\mathbf{y}}(\eta) = \frac{f_{\mathbf{x}}(A^{-1}\eta)}{|\det A|}. \quad (1.25)$$

Poiché \mathbf{x} è congiuntamente gaussiana, la sua d.d.p. vale (v. equazione (1.16)):

$$f_{\mathbf{x}}(\xi) = \frac{1}{(2\pi)^{n/2}(\det P_{\mathbf{x}})^{1/2}} e^{-\frac{1}{2}(\xi - m_{\mathbf{x}})^T P_{\mathbf{x}}^{-1}(\xi - m_{\mathbf{x}})}. \quad (1.26)$$

Sostituendo $\xi = A^{-1}\eta$ nella (1.26), la (1.25) diventa:

$$f_{\mathbf{y}}(\eta) = \frac{1}{(2\pi)^{n/2}(\det P_{\mathbf{x}})^{1/2} |\det A|} e^{-\frac{1}{2}(A^{-1}\eta - m_{\mathbf{x}})^T P_{\mathbf{x}}^{-1}(A^{-1}\eta - m_{\mathbf{x}})}. \quad (1.27)$$

Utilizzando l'identità:

$$A^{-1}\eta - m_{\mathbf{x}} = A^{-1}(\eta - Am_{\mathbf{x}}),$$

l'argomento della funzione esponenziale nella (1.27) può essere riscritto come:

$$\begin{aligned} -\frac{1}{2}(A^{-1}\eta - m_{\mathbf{x}})^T P_{\mathbf{x}}^{-1}(A^{-1}\eta - m_{\mathbf{x}}) &= -\frac{1}{2} [A^{-1}(\eta - Am_{\mathbf{x}})]^T P_{\mathbf{x}}^{-1} A^{-1}(\eta - Am_{\mathbf{x}}) \\ &\stackrel{[1]}{=} -\frac{1}{2}(\eta - Am_{\mathbf{x}})^T A^{-T} P_{\mathbf{x}}^{-1} A^{-1}(\eta - Am_{\mathbf{x}}) \\ &\stackrel{[2]}{=} -\frac{1}{2}(\eta - m_{\mathbf{y}})^T A^{-T} P_{\mathbf{x}}^{-1} A^{-1}(\eta - m_{\mathbf{y}}) \\ &\stackrel{[3]}{=} -\frac{1}{2}(\eta - m_{\mathbf{y}})^T P_{\mathbf{y}}^{-1}(\eta - m_{\mathbf{y}}), \end{aligned} \quad (1.28)$$

dove con A^{-T} si è indicata la trasposta dell'inversa $(A^{-1})^T$. L'uguaglianza [1] discende dalla proprietà dell'operatore di trasposizione $(BC)^T = C^T B^T$ (con B e C matrici di dimensioni opportune), la [2] si ottiene ponendo:

$$m_{\mathbf{y}} = Am_{\mathbf{x}}, \quad (1.29)$$

mentre la [3] deriva dall'assegnazione:

$$P_{\mathbf{y}} = AP_{\mathbf{x}}A^T, \quad (1.30)$$

e ricordando che $(BCD)^{-1} = D^{-1}C^{-1}B^{-1}$ (se B, C, D sono matrici quadrate invertibili di dimensioni opportune). Inoltre, notiamo che:

$$\begin{aligned} (\det P_{\mathbf{y}})^{1/2} &= (\det A \det P_{\mathbf{x}} \det A^T)^{1/2} \\ &\stackrel{[1]}{=} ((\det A)^2 \det P_{\mathbf{x}})^{1/2} \\ &= |\det A| (\det P_{\mathbf{x}})^{1/2} \end{aligned} \quad (1.31)$$

dal momento che $\det A = \det A^T$ (uguaglianza [1]). Infine, sostituendo le (1.28)-(1.31) nella (1.27) si ha:

$$f_{\mathbf{y}}(\eta) = \frac{1}{(2\pi)^{n/2} (\det P_{\mathbf{y}})^{1/2}} e^{-\frac{1}{2}(\eta - m_{\mathbf{y}})^T P_{\mathbf{y}}^{-1} (\eta - m_{\mathbf{y}})}. \quad (1.32)$$

Quindi, la v.a. \mathbf{y} è a sua volta gaussiana, con valor medio $m_{\mathbf{y}}$ e matrice di covarianza $P_{\mathbf{y}}$ date dalle (1.29)-(1.30). \triangle

1.7 Distribuzioni condizionali

Vogliamo estendere il concetto di probabilità condizionata fra eventi (v. Definizione 1.3) alle variabili aleatorie. Per questo, siano date due v.a. \mathbf{x} e \mathbf{y} con densità di probabilità congiunta $f_{\mathbf{x},\mathbf{y}}(\xi, \eta)$. Supponendo che la variabile aleatoria \mathbf{y} abbia assunto un certo valore η , come si modifica la densità di probabilità della \mathbf{x} ? A tale scopo, si considerino i due eventi (v. Figura 1.6):

$$A = \{\mathbf{x} \leq \xi\}, \quad B = \{\eta \leq \mathbf{y} \leq \eta + d\eta\}.$$

Ricordando il significato della funzione di distribuzione (v. Definizione 1.8), dalla (1.5) si ha:

$$\begin{aligned} P(A|B) &= \frac{F_{\mathbf{x},\mathbf{y}}(\xi, \eta + d\eta) - F_{\mathbf{x},\mathbf{y}}(\xi, \eta)}{F_{\mathbf{y}}(\eta + d\eta) - F_{\mathbf{y}}(\eta)} \\ &= [\text{dividendo numeratore e denominatore per } d\eta] \\ &= \frac{(F_{\mathbf{x},\mathbf{y}}(\xi, \eta + d\eta) - F_{\mathbf{x},\mathbf{y}}(\xi, \eta)) / d\eta}{(F_{\mathbf{y}}(\eta + d\eta) - F_{\mathbf{y}}(\eta)) / d\eta}. \end{aligned} \quad (1.33)$$

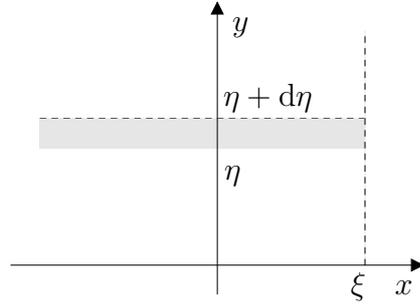


Figura 1.6: La $P(A \cap B)$ è pari alla probabilità che (\mathbf{x}, \mathbf{y}) appartenga alla zona in grigio.

Lasciando tendere $d\eta \rightarrow 0$:

$$\begin{aligned}
 P(A|B) &= \frac{\frac{\partial}{\partial \eta} F_{\mathbf{x}, \mathbf{y}}(\xi, \eta)}{f_{\mathbf{y}}(\eta)} \\
 &\stackrel{[1]}{=} \frac{\int_{-\infty}^{\xi} f_{\mathbf{x}, \mathbf{y}}(\tau, \eta) d\tau}{f_{\mathbf{y}}(\eta)} \\
 &\stackrel{[2]}{=} \int_{-\infty}^{\xi} \frac{f_{\mathbf{x}, \mathbf{y}}(\tau, \eta)}{f_{\mathbf{y}}(\eta)} d\tau,
 \end{aligned} \tag{1.34}$$

ottenuta derivando rispetto a η la relazione (1.8) (uguaglianza [1]) e portando $f_{\mathbf{y}}(\eta)$ dentro l'integrale (uguaglianza [2]). Si noti che, quando $d\eta \rightarrow 0$, l'evento B diviene $B = \{\mathbf{y} = \eta\}$. Allora, la $P(A|B)$ non è altro che la probabilità che $\mathbf{x} \leq \xi$, noto che $\mathbf{y} = \eta$. Per cui, dalla definizione di funzione di distribuzione (Definizione 1.5), possiamo scrivere $P(A|B) = F_{\mathbf{x}}(\xi|\mathbf{y} = \eta)$. Quindi, dalla (1.34):

$$F_{\mathbf{x}}(\xi|\mathbf{y} = \eta) = \int_{-\infty}^{\xi} \frac{f_{\mathbf{x}, \mathbf{y}}(\tau, \eta)}{f_{\mathbf{y}}(\eta)} d\tau. \tag{1.35}$$

Derivando ambo i membri della (1.35) rispetto a ξ , otteniamo la *densità di probabilità di \mathbf{x} condizionata a \mathbf{y}* :

$$f_{\mathbf{x}|\mathbf{y}}(\xi|\eta) = \frac{f_{\mathbf{x}, \mathbf{y}}(\xi, \eta)}{f_{\mathbf{y}}(\eta)}. \tag{1.36}$$

La (1.36) è anche nota come *formula di Bayes*, e descrive come si modifica la densità di probabilità di una v.a. \mathbf{x} quando si sia osservato l'esito η per il fenomeno incerto rappresentato dalla v.a. \mathbf{y} .

Se \mathbf{x} e \mathbf{y} sono due v.a. indipendenti, allora:

$$f_{\mathbf{x}|\mathbf{y}}(\xi|\eta) = f_{\mathbf{x}}(\xi). \quad (1.37)$$

Cioè, il fatto che la v.a. \mathbf{y} abbia assunto un certo valore non modifica la densità di probabilità *a priori* della \mathbf{x} . Alcune volte, l'indipendenza statistica fra due v.a. viene definita proprio con la relazione (1.37).

Si noti che, dalla formula di Bayes $f_{\mathbf{x},\mathbf{y}}(\xi, \eta) = f_{\mathbf{x}|\mathbf{y}}(\xi|\eta) f_{\mathbf{y}}(\eta)$. Per cui, dalla (1.10) deriva immediatamente:

$$f_{\mathbf{x}}(\xi) = \int_{-\infty}^{+\infty} f_{\mathbf{x}|\mathbf{y}}(\xi|\eta) f_{\mathbf{y}}(\eta) d\eta. \quad (1.38)$$

La (1.38) rappresenta l'estensione del *Teorema della probabilità totale* al caso di variabili aleatorie continue.

Supponiamo, ora, di aver osservato $\mathbf{y} = \eta$, per una certa v.a. \mathbf{y} . Sia nota la $f_{\mathbf{x},\mathbf{y}}(\xi, \eta)$. Quanto valgono il valor atteso e la varianza di \mathbf{x} ? Per tale calcolo, applichiamo le definizioni (1.13)-(1.15), avendo cura di sostituire la d.d.p. *a priori* $f_{\mathbf{x}}(\xi)$ con la d.d.p. condizionata $f_{\mathbf{x}|\mathbf{y}}(\xi|\eta)$:

$$m_{\mathbf{x}|\mathbf{y}} = \int_{-\infty}^{+\infty} \xi f_{\mathbf{x}|\mathbf{y}}(\xi|\eta) d\xi \quad (1.39)$$

$$\sigma_{\mathbf{x}|\mathbf{y}}^2 = \int_{-\infty}^{+\infty} (\xi - m_{\mathbf{x}|\mathbf{y}})^2 f_{\mathbf{x}|\mathbf{y}}(\xi|\eta) d\xi \quad (1.40)$$

Le (1.39)-(1.40) prendono il nome di *valor medio condizionato* (o *media condizionata*) e *varianza condizionata*, rispettivamente. Si noti che sia la media che la varianza condizionata sono funzioni del particolare valore η assunto dalla v.a. \mathbf{y} . Per cui, ad esempio, il simbolo $m_{\mathbf{x}|\mathbf{y}}$ va inteso nel senso di $\mathbf{E}[\mathbf{x}|\mathbf{y} = \eta]$. Per brevità, la dipendenza da η , sarà spesso omessa. Nel caso in cui la \mathbf{x} sia una v.a. vettoriale, la (1.40) va opportunamente modificata sostituendo $(\cdot)^2$ con $(\cdot)(\cdot)^T$ (v. Definizione 1.18), ottenendo così la matrice di covarianza condizionata $P_{\mathbf{x}|\mathbf{y}}$.

Esempio 1.9. Siano \mathbf{x}_1 e \mathbf{x}_2 due v.a. congiuntamente gaussiane con media nulla e matrice di covarianza R :

$$\mathbf{E} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad R = \mathbf{E} \begin{bmatrix} \mathbf{x}_1^2 & \mathbf{x}_1 \mathbf{x}_2 \\ \mathbf{x}_1 \mathbf{x}_2 & \mathbf{x}_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{bmatrix},$$

con $|\rho| < 1$.

Si consideri il cambio di variabile $\tilde{\mathbf{x}}_1 = \mathbf{x}_1 - \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}\mathbf{x}_2$. Allora:

$$\begin{bmatrix} \tilde{\mathbf{x}}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} 1 & -\frac{\rho\sigma_1\sigma_2}{\sigma_2^2} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}.$$

Dal momento che $\mathbf{x}_1, \mathbf{x}_2$ sono congiuntamente gaussiane, anche $\tilde{\mathbf{x}}_1$ e \mathbf{x}_2 sono congiuntamente gaussiane (v. Esempio 1.8). Calcoliamo:

$$\begin{aligned} \mathbf{E}[\tilde{\mathbf{x}}_1\mathbf{x}_2] &= \mathbf{E}\left[\left(\mathbf{x}_1 - \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}\mathbf{x}_2\right)\mathbf{x}_2\right] \\ &\stackrel{[1]}{=} \mathbf{E}[\mathbf{x}_1\mathbf{x}_2] - \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}\mathbf{E}[\mathbf{x}_2\mathbf{x}_2] \\ &\stackrel{[2]}{=} \rho\sigma_1\sigma_2 - \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}\sigma_2^2 = 0, \end{aligned} \quad (1.41)$$

in cui si è utilizzata (uguaglianza [1]) la proprietà di linearità dell'operatore di valore atteso (v. equazione (1.14)) e la definizione della matrice di covarianza R (uguaglianza [2]). Quindi le v.a. $\tilde{\mathbf{x}}_1$ e \mathbf{x}_2 sono incorrelate (v. Definizione 1.16). Ma, poichè sono anche congiuntamente gaussiane, questo equivale a dire che esse sono indipendenti (v. Proprietà 4 a pag. 18). Allora, dalla Definizione 1.10 di indipendenza:

$$f_{\tilde{\mathbf{x}}_1, \mathbf{x}_2}(\tilde{x}_1, x_2) = f_{\tilde{\mathbf{x}}_1}(\tilde{x}_1) f_{\mathbf{x}_2}(x_2). \quad (1.42)$$

Dal Teorema 2 (v. anche Esempio 1.8):

$$f_{\tilde{\mathbf{x}}_1, \mathbf{x}_2}(\tilde{x}_1, x_2) = \frac{f_{\mathbf{x}_1, \mathbf{x}_2}\left(\tilde{x}_1 + \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}x_2, x_2\right)}{1}, \quad (1.43)$$

dal momento che:

$$\det \begin{bmatrix} 1 & -\frac{\rho\sigma_1\sigma_2}{\sigma_2^2} \\ 0 & 1 \end{bmatrix} = 1.$$

Allora:

$$\begin{aligned} f_{\mathbf{x}_1|\mathbf{x}_2}(x_1|x_2) &\stackrel{[1]}{=} \frac{f_{\mathbf{x}_1, \mathbf{x}_2}(x_1, x_2)}{f_{\mathbf{x}_2}(x_2)} \stackrel{[2]}{=} \frac{f_{\tilde{\mathbf{x}}_1, \mathbf{x}_2}\left(x_1 - \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}x_2, x_2\right)}{f_{\mathbf{x}_2}(x_2)} \\ &\stackrel{[3]}{=} f_{\tilde{\mathbf{x}}_1}\left(x_1 - \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}x_2\right), \end{aligned} \quad (1.44)$$

in cui [1] è la formula di Bayes, [2] deriva dalla (1.43) ponendo $x_1 = \tilde{x}_1 + \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}x_2$ e la [3] si ottiene sfruttando l'indipendenza (1.42).

Dal momento che $\tilde{\mathbf{x}}_1$ è combinazione lineare di v.a. congiuntamente gaussiane, la sua d.d.p è ancora gaussiana, con:

$$m_{\tilde{\mathbf{x}}_1} = \mathbf{E}[\tilde{\mathbf{x}}_1] = \mathbf{E}[\mathbf{x}_1] - \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}\mathbf{E}[\mathbf{x}_2] = 0,$$

e

$$\begin{aligned}\sigma_{\tilde{\mathbf{x}}_1}^2 &= \mathbf{E}[\tilde{\mathbf{x}}_1^2] = \mathbf{E}\left[\left(\mathbf{x}_1 - \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}\mathbf{x}_2\right)^2\right] \\ &\stackrel{[1]}{=} \mathbf{E}[\mathbf{x}_1^2] + \frac{\rho\sigma_1^2\sigma_2^2}{\sigma_2^4}\mathbf{E}[\mathbf{x}_2^2] - 2\frac{\rho\sigma_1\sigma_2}{\sigma_2^2}\mathbf{E}[\mathbf{x}_1\mathbf{x}_2] \\ &\stackrel{[2]}{=} \sigma_1^2 + \rho^2\sigma_1^2 - 2\rho^2\sigma_1^2 = \sigma_1^2(1 - \rho^2).\end{aligned}$$

La [1] è dovuta alla linearità dell'operatore di valore atteso, e la [2] si ottiene notando che $\mathbf{E}[\mathbf{x}_1^2] = \sigma_1^2$, $\mathbf{E}[\mathbf{x}_2^2] = \sigma_2^2$ e $\mathbf{E}[\mathbf{x}_1\mathbf{x}_2] = \rho\sigma_1^2\sigma_2^2$ (v la matrice di covarianza R). Quindi:

$$f_{\tilde{\mathbf{x}}_1}(\tilde{x}_1) = \frac{1}{\sqrt{2\pi\sigma_{\tilde{\mathbf{x}}_1}}} e^{-\frac{\tilde{x}_1^2}{2\sigma_{\tilde{\mathbf{x}}_1}^2}}.$$

La (1.44) ci dice che la $f_{\mathbf{x}_1|\mathbf{x}_2}(x_1|x_2)$ è proprio uguale alla $f_{\tilde{\mathbf{x}}_1}(\tilde{x}_1)$, calcolata in $\tilde{x}_1 = x_1 - \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}x_2$. Per cui:

$$f_{\mathbf{x}_1|\mathbf{x}_2}(x_1|x_2) = \frac{1}{\sqrt{2\pi(1-\rho^2)\sigma_1}} e^{-\frac{\left(x_1 - \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}x_2\right)^2}{2\sigma_1^2(1-\rho^2)}}. \quad (1.45)$$

Quindi, la d.d.p di \mathbf{x}_1 condizionata a \mathbf{x}_2 è ancora gaussiana. Inoltre, confrontando la (1.45) con l'espressione di una generica v.a. gaussiana (v. equazione (1.7)), si verifica immediatamente che:

$$\begin{aligned}m_{\mathbf{x}_1|\mathbf{x}_2} &= \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}x_2; \\ \sigma_{\mathbf{x}_1|\mathbf{x}_2}^2 &= \sigma_1^2(1 - \rho^2).\end{aligned}$$

Si noti che la precedente espressione di $m_{\mathbf{x}_1|\mathbf{x}_2}$ vale solo nel caso in cui \mathbf{x}_1 e \mathbf{x}_2 siano entrambe a media nulla. \triangle

Il risultato ottenuto nell'esempio precedente può essere generalizzato al caso vettoriale e media non nulla.

Teorema 3. *Siano \mathbf{x}_1 e \mathbf{x}_2 due v.a. vettoriali congiuntamente gaussiane. Posto $\mathbf{x} = [\mathbf{x}_1 \ \mathbf{x}_2]^T$, siano:*

$$\mathbf{E}[\mathbf{x}] = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \quad P_{\mathbf{x}} = \begin{bmatrix} R_1 & R_{12} \\ R_{12}^T & R_2 \end{bmatrix} \geq 0,$$

con $R_2 > 0$. Allora:

$$m_{\mathbf{x}_1|\mathbf{x}_2} = m_1 + R_{12}R_2^{-1}(x_2 - m_2); \quad (1.46)$$

$$P_{\mathbf{x}_1|\mathbf{x}_2} = R_1 - R_{12}R_2^{-1}R_{12}^T. \quad (1.47)$$

△

Si noti che la condizione $R_2 > 0$ serve a garantire l'invertibilità della R_2 . L'osservazione del valore x_2 per la v.a. \mathbf{x}_2 “sposta” il valor medio di \mathbf{x}_1 e ne riduce l'incertezza. Infatti, poiché vale $R_1 - R_{12}R_2^{-1}R_{12}^T \geq 0$ (regola di Schur), allora $P_{\mathbf{x}_1|\mathbf{x}_2} = R_1 - R_{12}R_2^{-1}R_{12}^T \leq R_1$, cioè la varianza della v.a. \mathbf{x}_1 dopo l'osservazione del valore assunto da \mathbf{x}_2 è minore di (o tutt'al più uguale a) quella iniziale. Infine, $P_{\mathbf{x}_1|\mathbf{x}_2} = R_1$ solo nel caso in cui $R_{12} = 0$, cioè quando le due v.a. sono incorrelate e, poichè gaussiane, indipendenti.

1.8 Esercizi

1.1. Le capre e la Ferrari. Tre porte chiuse, etichettate con le lettere A, B e C, nascondono due capre e una Ferrari. Le porte sono sorvegliate da un severo custode che sa dove è nascosta la Ferrari. Si supponga di scegliere la porta A. A quel punto il custode apre la porta B da cui esce una capra, e vi propone di scegliere nuovamente. Da un punto di vista probabilistico, conviene mantenere la scelta iniziale A oppure cambiare e optare per la C? [Suggerimento: usare la probabilità condizionata!]

1.2. L'ago di Buffon. Un ago lungo 1 cm viene lanciato in maniera casuale su un foglio a righe, in cui la distanza tra le righe è esattamente 1 cm. Calcolare la probabilità che l'ago intersechi una delle righe.

1.3. Sia \mathbf{x} una v.a. uniformemente distribuita nell'intervallo $[-\frac{T}{2}, \frac{T}{2}]$. Dopo aver calcolato il valor medio $m_{\mathbf{x}}$ e la varianza $\sigma_{\mathbf{x}}^2$ di \mathbf{x} , si utilizzi il comando `rand` di MATLAB[®] per generare un vettore \mathbf{x} contenente N valori estratti dalla precedente distribuzione, con $T = 6$.

Si confronti il valor medio teorico $m_{\mathbf{x}}$ con la media aritmetica `m_y` del vettore \mathbf{x} :

$$\text{m_x} = \text{sum}(\mathbf{x})/N$$

al variare di $N = 10, 100, 1000$.

1.4. Dopo aver verificato che la $f_{\mathbf{x},\mathbf{y}}(\xi, \eta)$ dell'Esempio 1.6 sia effettivamente una densità di probabilità congiunta (vale a dire che essa soddisfi le Proprietà 1 e 2 di pag. 13), si calcoli il valor medio e la matrice di covarianza del vettore $[\mathbf{x}, \mathbf{y}]^T$.

1.5. Sia $\mathbf{x} = [x_1, x_2, x_3]^T$ una v.a. in \mathbb{R}^3 , con densità di probabilità:

$$f_{\mathbf{x}}(x_1, x_2, x_3) = \begin{cases} \frac{1}{2}x_1 + \frac{3}{2}x_2^2 + x_3^3 & \text{se } 0 \leq x_i \leq 1, \quad i = 1, 2, 3 \\ 0 & \text{altrimenti} \end{cases}$$

Dopo aver verificato che $f_{\mathbf{x}}(x_1, x_2, x_3)$ rappresenta una d.d.p. si calcoli il valor medio e la matrice di covarianza di \mathbf{x} .

1.6. Sia \mathbf{x} una v.a. scalare gaussiana:

$$f_{\mathbf{x}}(\xi) = \frac{1}{\sqrt{2\pi}\sigma_{\mathbf{x}}} e^{-\frac{(\xi-m_{\mathbf{x}})^2}{2\sigma_{\mathbf{x}}^2}} \quad \sigma_{\mathbf{x}} > 0, \quad m_{\mathbf{x}} \in \mathbb{R}.$$

- Verificare che il valor medio e la varianza di \mathbf{x} valgono $m_{\mathbf{x}}$ e $\sigma_{\mathbf{x}}^2$, rispettivamente, sapendo che:

$$\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi}.$$

- Si utilizzi il comando `randn` di MATLAB[®] per generare un vettore \mathbf{x} contenente N valori estratti dalla precedente distribuzione, con $m_{\mathbf{x}} = 0$ e $\sigma_{\mathbf{x}}^2 = 1$ e si confronti il valor medio teorico $m_{\mathbf{x}}$ con la media aritmetica $\mathbf{m_x}$ del vettore \mathbf{x} :

$$\mathbf{m_x} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}(i)$$

al variare di $N = 1, 2, \dots, 1000$.

1.7. Dimostrare che se \mathbf{x} e \mathbf{y} sono due v.a. gaussiane ed incorrelate, allora esse sono anche indipendenti.

1.8. Sia \mathbf{x} una v.a. gaussiana scalare, con valor medio $m_{\mathbf{x}}$ e varianza $\sigma_{\mathbf{x}}^2$. Verificare che la v.a. $\mathbf{y} = a\mathbf{x} + b$, con $a, b \in \mathbb{R}$, è ancora una v.a. gaussiana, con valor medio $m_{\mathbf{y}} = am_{\mathbf{x}} + b$ e varianza $\sigma_{\mathbf{y}}^2 = a^2\sigma_{\mathbf{x}}^2$.

1.9. Una v.a. \mathbf{x} avente d.d.p.

$$f_{\mathbf{x}}(x) = \begin{cases} \lambda e^{-\lambda x} & \text{se } x \geq 0 \\ 0 & \text{se } x < 0 \end{cases}$$

con $\lambda > 0$, si dice *esponenzialmente distribuita*. Verificare che il valor medio $m_{\mathbf{x}} = \frac{1}{\lambda}$ e la varianza $\sigma_{\mathbf{x}}^2 = \frac{1}{\lambda^2}$.

1.10. Sia la tensione \mathbf{V} , ai capi di una resistenza $R = 10\Omega$, assimilabile ad una v.a. uniformemente distribuita nell'intervallo $[5V, 10V]$. Calcolare la densità di probabilità della potenza $\mathbf{W} = \frac{1}{R}\mathbf{V}^2$ dissipata dalla resistenza.

1.11. Siano \mathbf{x} e \mathbf{y} due v.a. aventi densità di probabilità $f_{\mathbf{x},\mathbf{y}}(x, y)$. Calcolare la densità di probabilità della v.a. $\mathbf{z} = \mathbf{x} - \mathbf{y}$.

1.12. Si consideri un veicolo la cui posizione è descritta nel piano cartesiano dalle coordinate x, y . Si associno alle coordinate due variabili aleatorie \mathbf{x} e \mathbf{y} e si assuma che esse siano indipendenti e uniformemente distribuite nell'intervallo $[0, 1]$. Supponendo di disporre di un radar centrato nell'origine degli

assi, esso fornirà la posizione del veicolo nelle coordinate polari ρ e θ :

$$\rho = \sqrt{x^2 + y^2}, \quad \theta = \begin{cases} \arctan \frac{y}{x} & \text{se } x \geq 0 \\ \arctan \frac{y}{x} + \pi & \text{se } x < 0 \end{cases}.$$

Determinare la funzione di densità di probabilità congiunta delle variabili aleatorie ρ e θ .

Ripetere l'esercizio assumendo che \mathbf{x} e \mathbf{y} siano congiuntamente gaussiane, indipendenti, a media nulla e varianza unitaria.

1.13. Siano \mathbf{x}_i , $i = 1, \dots, n$ variabili aleatorie indipendenti e uniformemente distribuite nell'intervallo $[-\frac{T}{2}, \frac{T}{2}]$. Calcolare la densità di probabilità della v.a.

$$\mathbf{y}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i,$$

per:

1. $n = 2$;
2. $n = 3$.

Confrontare le $f_{\mathbf{y}_n}(\mathbf{y})$ ottenute con la densità gaussiana di media $m = 0$ e varianza $\sigma^2 = \frac{1}{n} \frac{T^2}{12}$.

1.14. Si supponga di voler misurare una grandezza incognita, assimilabile ad una v.a \mathbf{x} , mediante un sensore. Sia

$$\mathbf{y} = \mathbf{x} + \mathbf{v}$$

l'equazione della misura, in cui la v.a. \mathbf{y} rappresenta la lettura fornita dallo strumento, mentre la v.a. \mathbf{v} modella il rumore additivo introdotto dallo strumento. Sotto le ipotesi che \mathbf{x} e \mathbf{v} siano v.a. gaussiane, indipendenti, $\mathbf{x} \sim N(m_{\mathbf{x}}, \sigma_{\mathbf{x}}^2)$, $\mathbf{v} \sim N(0, \sigma_{\mathbf{v}}^2)$, calcolare il valor medio $m_{\mathbf{x}|\mathbf{y}}$ e la varianza $\sigma_{\mathbf{x}|\mathbf{y}}^2$ condizionati. Vale a dire, calcolare la media e la varianza della v.a. \mathbf{x} se la lettura rilevata sensore vale $\mathbf{y} = y$.

Inoltre, confrontare media e varianza condizionate con i corrispondenti valori $m_{\mathbf{x}}$ e $\sigma_{\mathbf{x}}^2$ *a priori* (cioè, prima di effettuare la misura). Interpretare il risultato ottenuto in termini del rapporto segnale/rumore, definito come $SNR = \frac{\sigma_{\mathbf{x}}^2}{\sigma_{\mathbf{v}}^2}$.

Capitolo 2

Processi stocastici

In questo capitolo verranno introdotti i processi stocastici e gli strumenti matematici necessari per il loro studio.

2.1 Definizioni e proprietà

Definizione 2.1. Un *processo stocastico* (brevemente p.s.) è una sequenza temporale di variabili aleatorie $x(t)$, con $t \in \mathcal{T}$, dove \mathcal{T} rappresenta l'asse dei tempi. Nel caso in cui \mathcal{T} sia un insieme numerabile $\{t_1, t_2, \dots, t_k, \dots\}$, il processo stocastico è detto *tempo-discreto*. Se $\mathcal{T} = \mathbb{R}^+$, il processo stocastico è detto *tempo-continuo*. Nel seguito, verranno presi in considerazione i p.s. tempo-discreto.

Ricordando la definizione di variabile aleatoria, possiamo vedere un processo stocastico come una funzione $x(t, \omega)$ che associa ad ogni coppia $(t, \omega) \in \mathcal{T} \times \Omega$ un numero reale ¹:

$$x : \mathcal{T} \times \Omega \rightarrow \mathbb{R}.$$

Fissato un istante temporale $\bar{t} \in \mathcal{T}$, il processo stocastico $x(\bar{t})$ si riduce ad una variabile aleatoria. D'altro canto, fissato l'evento ω (ad esempio,

¹Per semplicità di notazione, la dipendenza di x dall'evento ω verrà omessa nel seguito.

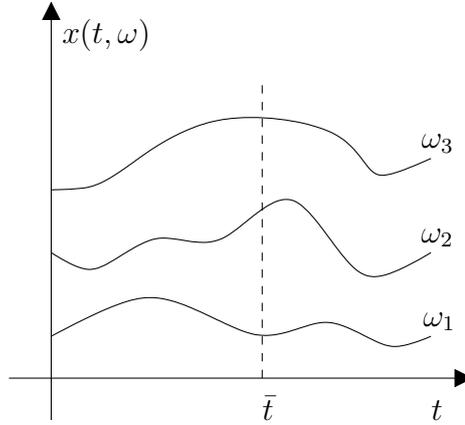


Figura 2.1: Diverse realizzazioni di un processo stocastico $x(t)$ al variare dell'evento ω .

effettuando un esperimento), il processo stocastico $x(t)$ diventa una funzione di t (detta *realizzazione* del processo stocastico), vedi Figura 2.1.

Un processo stocastico è completamente caratterizzato in senso statistico se è nota la probabilità:

$$P(x(t_1) \leq x_1, x(t_2) \leq x_2, \dots, x(t_k) \leq x_k), \quad \forall t_1, \dots, t_k \quad \forall x_1, \dots, x_k \quad \forall k.$$

Analogamente a quanto fatto con le variabili aleatorie, è possibile definire le funzioni di distribuzione e di densità della probabilità per ogni $k = 1, 2, \dots$ (dette *statistiche di ordine k*):

$$F_{\mathbf{x}}(x_1, \dots, x_k; t_1, \dots, t_k), \quad f_{\mathbf{x}}(x_1, \dots, x_k; t_1, \dots, t_k).$$

Nella pratica, le funzioni di distribuzione (o, equivalentemente, le densità) non sono note per ogni $k \in \mathbb{N}$. Per cui, spesso, ci si limita a considerare solo le statistiche del primo e secondo ordine. Siano

$$\begin{cases} F_{\mathbf{x}}(x; t) = P(x(t) \leq x), \\ f_{\mathbf{x}}(x; t) = \frac{\partial}{\partial x} F_{\mathbf{x}}(x; t), \end{cases}$$

e

$$\begin{cases} F_{\mathbf{x}_1, \mathbf{x}_2}(x_1, x_2; t_1, t_2) = P(x(t_1) \leq x_1, x(t_2) \leq x_2), \\ f_{\mathbf{x}_1, \mathbf{x}_2}(x_1, x_2; t_1, t_2) = \frac{\partial^2}{\partial x_1 \partial x_2} F_{\mathbf{x}_1, \mathbf{x}_2}(x_1, x_2; t_1, t_2), \end{cases}$$

le statistiche del primo e del secondo ordine, rispettivamente. Dal momento che fissati t_1 e t_2 , $x(t_1)$ e $x(t_2)$ sono variabili aleatorie, si ha (v. equazione (1.10)):

$$f_{\mathbf{x}}(x_1; t_1) = \int_{-\infty}^{+\infty} f_{\mathbf{x}_1, \mathbf{x}_2}(x_1, x_2; t_1, t_2) dx_2.$$

È possibile definire media e covarianza di un processo stocastico nel seguente modo.

Definizione 2.2. Si definisce *media* (o *valore atteso*) $m_{\mathbf{x}}(t)$ del p.s. $x(t)$ la quantità:

$$m_{\mathbf{x}}(t) \triangleq \mathbf{E}[x(t)] \triangleq \int_{-\infty}^{+\infty} x f_{\mathbf{x}}(x; t) dx. \quad (2.1)$$

Definizione 2.3. Si definisce *funzione di covarianza* (o *autocovarianza*) $R_{\mathbf{x}}(t, s)$ di un p.s. $x(t)$ la quantità:

$$\begin{aligned} R_{\mathbf{x}}(t, s) &\triangleq \mathbf{E}[(x(t) - m_{\mathbf{x}}(t))(x(s) - m_{\mathbf{x}}(s))^T] \\ &\triangleq \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x_1 - m_{\mathbf{x}}(t))(x_2 - m_{\mathbf{x}}(s))^T f_{\mathbf{x}_1, \mathbf{x}_2}(x_1, x_2; t, s) dx_1 dx_2. \end{aligned} \quad (2.2)$$

Nella definizione precedente, la notazione $(\cdot)^T$ è necessaria nel caso in cui $x(t)$ sia un p.s. vettoriale. Se $x(t) \in \mathbb{R}^n$, allora la media è un vettore di dimensione n , e la funzione di covarianza $R_{\mathbf{x}}(t, s)$ è una matrice quadrata $n \times n$.

Definizione 2.4. Si definisce *funzione di covarianza incrociata* (o *cross-covarianza*) $R_{\mathbf{x}, \mathbf{y}}(t, s)$ di due p.s. $x(t)$, $y(t)$ la quantità:

$$\begin{aligned} R_{\mathbf{x}, \mathbf{y}}(t, s) &\triangleq \mathbf{E}[(x(t) - m_{\mathbf{x}}(t))(y(s) - m_{\mathbf{y}}(s))^T] \\ &\triangleq \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - m_{\mathbf{x}}(t))(y - m_{\mathbf{y}}(s))^T f_{\mathbf{x}, \mathbf{y}}(x, y; t, s) dx dy. \end{aligned} \quad (2.3)$$

Si noti che sia la media che la covarianza di un p.s. sono, in generale, funzioni di uno (media) o due (covarianza) indici temporali. Ad esempio, la media di un p.s. può assumere valori diversi a seconda dell'istante t in cui viene calcolata.

Esiste un'importante classe di p.s. per i quali l'osservazione precedente non è più vera.

Definizione 2.5. Un processo stocastico $x(t)$ si dice *stazionario in senso forte* (o in *senso stretto*) se le statistiche di ogni ordine di $x(t)$ e $x(t + \tau)$ sono identiche, ovvero se:

$$P(x(t_1 + \tau) \leq x_1, \dots, x(t_k + \tau) \leq x_k)$$

non dipende da τ , $\forall t_1, \dots, t_k$, $\forall x_1, \dots, x_k$ e $\forall k$.

Quindi, la distribuzione di un p.s. stazionario in senso forte è invariante rispetto a traslazioni nel tempo. La stazionarietà forte di un p.s. è una proprietà molto stringente e di difficile verifica (occorre calcolare le statistiche di ogni ordine). Per questo, spesso, ci si accontenta di una versione più debole.

Definizione 2.6. Un processo stocastico $x(t)$ si dice *stazionario in senso debole* (o in *senso lato*) se i momenti del primo e del secondo ordine di $x(t)$ e $x(t + \tau)$ sono identici, ovvero se:

$$\begin{aligned} m_{\mathbf{x}}(t) &= m_{\mathbf{x}}(t + \tau) \\ R_{\mathbf{x}}(t, s) &= R_{\mathbf{x}}(t + \tau, s + \tau) \end{aligned}$$

per ogni τ .

Quindi, la media e la funzione di covarianza di un p.s. stazionario in senso debole sono invarianti rispetto a traslazioni nel tempo.

Poichè la Definizione 2.6 vale per ogni τ , segue immediatamente che un p.s. $x(t)$ è stazionario in senso debole se e solo se:

$$m_{\mathbf{x}}(t) = m_{\mathbf{x}} \tag{2.4}$$

$$R_{\mathbf{x}}(t, s) = R_{\mathbf{x}}(t - s) \tag{2.5}$$

In altre parole, un p.s. è stazionario in senso debole se la sua media è costante (equazione (2.4)) e la sua funzione di covarianza non dipende dai due istanti

temporali in cui viene calcolata ma solo dalla loro differenza (equazione (2.5)). Per questo motivo, spesso la funzione di covarianza di un p.s. stazionario viene indicata come funzione di un solo indice temporale $\tau = t - s$:

$$R_{\mathbf{x}}(\tau) = \mathbf{E} [(x(t + \tau) - m_{\mathbf{x}})(x(t) - m_{\mathbf{x}})^T]. \quad (2.6)$$

È possibile dimostrare che la stazionarietà forte implica la stazionarietà debole, mentre il contrario non è in generale vero. Vale a dire, un p.s. può essere stazionario in senso debole ma non esserlo in senso forte. D'ora in avanti, per stazionarietà si intenderà sempre stazionarietà in senso debole, salvo ove espressamente specificato.

Definizione 2.7. Due processi stocastici $x(t)$ e $y(t)$ si dicono *congiuntamente stazionari in senso debole* se ciascuno di essi è stazionario in senso debole e la funzione di cross-covarianza è invariante rispetto a traslazioni nel tempo:

$$R_{\mathbf{xy}}(t, s) = R_{\mathbf{xy}}(t + \tau, s + \tau), \quad \forall \tau.$$

Analogamente a quanto fatto per la funzione di covarianza di un processo stazionario, si è soliti esprimere la cross-covarianza di due processi congiuntamente stazionari come funzione di un solo indice temporale:

$$R_{\mathbf{xy}}(\tau) = \mathbf{E} [(x(t + \tau) - m_{\mathbf{x}})(y(t) - m_{\mathbf{y}})^T]. \quad (2.7)$$

Sia $x(t)$ un p.s. tempo-discreto, stazionario. Allora la sua funzione di covarianza $R_{\mathbf{x}}(\tau)$ gode delle seguenti proprietà:

1. $R_{\mathbf{x}}(\tau) = R_{\mathbf{x}}^T(-\tau)$;
2. $R_{\mathbf{x}}(0)$ ha elementi positivi sulla diagonale;
3. la matrice di covarianza:

$$P(m) = \begin{bmatrix} R_{\mathbf{x}}(0) & R_{\mathbf{x}}(1) & \dots & R_{\mathbf{x}}(m-1) \\ R_{\mathbf{x}}(-1) & R_{\mathbf{x}}(0) & \dots & R_{\mathbf{x}}(m-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_{\mathbf{x}}(1-m) & R_{\mathbf{x}}(2-m) & \dots & R_{\mathbf{x}}(0) \end{bmatrix}$$

è simmetrica e semidefinita positiva, $P(m) \geq 0, \forall m$;

4. sia $y(t)$ un p.s. congiuntamente stazionario con $x(t)$; la funzione di cross-covarianza $R_{\mathbf{x}y}(\tau)$ soddisfa:

$$R_{\mathbf{x}y}(\tau) = R_{\mathbf{y}x}^T(-\tau).$$

Dimostrazione

Senza perdita di generalità, sia $m_{\mathbf{x}} = 0$.

1. Dalla (2.6) segue:

$$\begin{aligned} R_{\mathbf{x}}(\tau) &= \mathbf{E} [x(t+\tau)x(t)^T] \stackrel{[1]}{=} \mathbf{E} [x(s)x(s-\tau)^T] \\ &\stackrel{[2]}{=} \mathbf{E} [(x(s-\tau)x(s)^T)^T] = R_{\mathbf{x}}^T(-\tau), \end{aligned}$$

in cui [1] si ottiene ponendo $s = t + \tau$ e [2] dalla proprietà dell'operatore di trasposizione $ab^T = (ba^T)^T$, con a, b generici vettori reali.

2. Sia $x(t) = [x_1(t), \dots, x_n(t)]^T \in \mathbb{R}^n$. Allora, gli elementi $r_{ii}(0)$, $i = 1, \dots, n$ sulla diagonale di $R_{\mathbf{x}}(0)$ valgono $r_{ii}(0) = \mathbf{E} [x_i(t)^2] > 0$.
3. Innanzitutto, si noti che è possibile scrivere:

$$P(m) = \mathbf{E} \left[\begin{pmatrix} x(t-1) \\ \vdots \\ x(t-m) \end{pmatrix} \begin{pmatrix} x(t-1)^T, \dots, x(t-m)^T \end{pmatrix} \right].$$

Sia $v = (v_1^T, \dots, v_m^T)^T$, con v_i generici vettori delle stesse dimensioni di $x(t)$. Allora:

$$\begin{aligned} v^T P(m) v &= (v_1^T, \dots, v_m^T) \mathbf{E} \left[\begin{pmatrix} x(t-1) \\ \vdots \\ x(t-m) \end{pmatrix} \begin{pmatrix} x(t-1)^T, \dots, x(t-m)^T \end{pmatrix} \right] \begin{pmatrix} v_1 \\ \vdots \\ v_m \end{pmatrix} \\ &= \mathbf{E} \left[\left(\sum_{k=1}^m v_k^T x(t-k) \right) \left(\sum_{k=1}^m x(t-k)^T v_k \right) \right] \\ &= \mathbf{E} \left[\left(\sum_{k=1}^m v_k^T x(t-k) \right)^2 \right] \geq 0 \end{aligned}$$

Poiché l'espressione precedente vale per ogni v , segue che $P(m) \geq 0$.

4. Vedi la dimostrazione del punto 1.

□

Nel caso in cui $x(t)$ sia un p.s. scalare, tempo-discreto, le proprietà precedenti possono essere interpretate nel seguente modo. La Proprietà 1 diventa:

$$r_{\mathbf{x}}(\tau) = r_{\mathbf{x}}(-\tau), \quad (2.8)$$

vale a dire, la funzione di covarianza di un p.s. scalare, stazionario è una funzione simmetrica rispetto alla variabile τ .

Inoltre, si consideri la combinazione lineare $\alpha_1 x(t + \tau) + \alpha_2 x(t)$, con $\alpha_1, \alpha_2 \in \mathbb{R}$, non entrambi nulli e $\tau \in \mathbb{N}$ (senza perdita di generalità, sia $m_{\mathbf{x}} = 0$). Allora:

$$\begin{aligned} 0 &\leq \mathbf{E} [(\alpha_1 x(t + \tau) + \alpha_2 x(t))^2] \\ &= \mathbf{E} [\alpha_1^2 x(t + \tau)^2 + 2\alpha_1 \alpha_2 x(t + \tau)x(t) + \alpha_2^2 x(t)^2] \\ &= \alpha_1^2 \mathbf{E} [x(t + \tau)^2] + 2\alpha_1 \alpha_2 \mathbf{E} [x(t + \tau)x(t)] + \alpha_2^2 \mathbf{E} [x(t)^2] \\ &= \alpha_1^2 r_{\mathbf{x}}(0) + 2\alpha_1 \alpha_2 r_{\mathbf{x}}(\tau) + \alpha_2^2 r_{\mathbf{x}}(0) = (\alpha_1^2 + \alpha_2^2)r_{\mathbf{x}}(0) + 2\alpha_1 \alpha_2 r_{\mathbf{x}}(\tau). \end{aligned}$$

Quindi, $(\alpha_1^2 + \alpha_2^2)r_{\mathbf{x}}(0) + 2\alpha_1 \alpha_2 r_{\mathbf{x}}(\tau) \geq 0$. È immediato verificare che tale disuguaglianza può essere riscritta nella forma:

$$\begin{bmatrix} \alpha_1 & \alpha_2 \end{bmatrix} \begin{bmatrix} r_{\mathbf{x}}(0) & r_{\mathbf{x}}(\tau) \\ r_{\mathbf{x}}(\tau) & r_{\mathbf{x}}(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \geq 0. \quad (2.9)$$

Poiché la (2.9) vale per ogni $\alpha_1, \alpha_2 \in \mathbb{R}$, vuol dire che la matrice $\begin{bmatrix} r_{\mathbf{x}}(0) & r_{\mathbf{x}}(\tau) \\ r_{\mathbf{x}}(\tau) & r_{\mathbf{x}}(0) \end{bmatrix}$ è semidefinita positiva. Perciò, il suo determinante sarà maggiore o tutt'al più uguale a zero:

$$\det \begin{bmatrix} r_{\mathbf{x}}(0) & r_{\mathbf{x}}(\tau) \\ r_{\mathbf{x}}(\tau) & r_{\mathbf{x}}(0) \end{bmatrix} = r_{\mathbf{x}}(0)^2 - r_{\mathbf{x}}(\tau)^2 \geq 0.$$

Dal momento che $r_{\mathbf{x}}(0) > 0$, la relazione precedente diventa:

$$|r_{\mathbf{x}}(\tau)| \leq r_{\mathbf{x}}(0), \quad \forall \tau \in \mathbb{N}.$$

Quindi, la funzione di covarianza di un p.s. scalare è sempre compresa nella “striscia” $[-r_{\mathbf{x}}(0), r_{\mathbf{x}}(0)]$ ed assume il suo valor massimo in corrispondenza di $\tau = 0$ (vedi Figura 2.2).

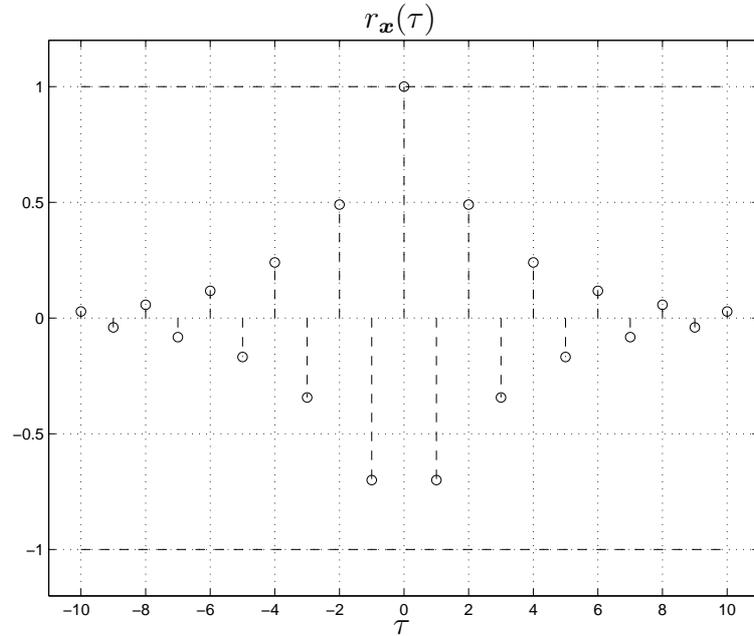


Figura 2.2: Una tipica funzione di covarianza di un processo stocastico scalare, tempo-discreto.

Si noti che $R_{\mathbf{x}}(0)$ rappresenta proprio la varianza (o matrice di covarianza, nel caso vettoriale) del p.s. stazionario $x(t)$.

Definizione 2.8. Un processo stocastico si dice *gaussiano* se le sue statistiche:

$$F_{\mathbf{x}}(x_1, \dots, x_k; t_1, \dots, t_k)$$

sono gaussiane, per ogni k . Un p.s. gaussiano è caratterizzato completamente da un punto di vista statistico qualora siano note la media e la funzione di covarianza. Inoltre, per i p.s. gaussiani la stazionarietà forte e quella debole coincidono.

Definizione 2.9. Un processo stocastico *bianco* (detto anche *rumore bianco*) è una sequenza di variabili aleatorie indipendenti. Se le v.a. sono anche

identicamente distribuite, si parla di processo a variabili *i.i.d.* (indipendenti e identicamente distribuite).

2.2 Esempi di processi stocastici

In questo paragrafo vengono presentati alcuni esempi di processi stocastici notevoli.

2.2.1 Processi stocastici puramente deterministici

Sono processi stocastici del tipo:

$$x(t) = \sum_{i=1}^m \mathbf{x}_i g_i(t),$$

in cui \mathbf{x}_i sono variabili aleatorie e $g_i(\cdot)$ funzioni deterministiche del tempo. Tali processi sono totalmente predicibili nel futuro, in quanto, una volta note le realizzazioni delle v.a. \mathbf{x}_i , il p.s. $x(t)$ è noto $\forall t \in \mathcal{T}$.

Seguono alcuni esempi di p.s. puramente deterministici.

- $x(t) = \mathbf{x}, \quad \forall t \in \mathcal{T},$

in cui \mathbf{x} rappresenta una variabile aleatoria avente valor medio $m_{\mathbf{x}}$ e varianza $\sigma_{\mathbf{x}}^2$. La media del p.s. $x(t)$ vale ovviamente:

$$\mathbf{E}[x(t)] = \mathbf{E}[\mathbf{x}] = m_{\mathbf{x}},$$

mentre la funzione di covarianza:

$$R_{\mathbf{x}}(t, s) = \mathbf{E}[(x(t) - m_{\mathbf{x}})(x(s) - m_{\mathbf{x}})] = \mathbf{E}[(\mathbf{x} - m_{\mathbf{x}})^2] = \sigma_{\mathbf{x}}^2.$$

Quindi, la $R_{\mathbf{x}}(t, s)$ non dipende da t ed s , ma è costante:

$$R_{\mathbf{x}}(\tau) = \sigma_{\mathbf{x}}^2, \quad \forall \tau.$$

Evidentemente, il p.s. $x(t)$ è stazionario in senso debole.

- $x(t) = A \cos(\omega t + \varphi)$,

con φ v.a. uniformemente distribuita nell'intervallo $[0, 2\pi]$. La media del p.s. $x(t)$ vale:

$$\mathbf{E}[x(t)] = \frac{1}{2\pi} \int_0^{2\pi} A \cos(\omega t + \varphi) d\varphi = 0,$$

mentre la funzione di covarianza:

$$R_{\mathbf{x}}(t, s) = \mathbf{E}[x(t)x(s)] = \frac{1}{2\pi} \int_0^{2\pi} A^2 \cos(\omega t + \varphi) \cos(\omega s + \varphi) d\varphi.$$

Poiché $\cos(\alpha) \cos(\beta) = \frac{1}{2} [\cos(\alpha + \beta) + \cos(\alpha - \beta)]$ (formule di Werner), la $R_{\mathbf{x}}(t, s)$ diventa:

$$\begin{aligned} R_{\mathbf{x}}(t, s) &= \frac{A^2}{4\pi} \int_0^{2\pi} [\cos(\omega t + \omega s + 2\varphi) + \cos(\omega t - \omega s)] d\varphi \\ &= \frac{A^2}{4\pi} 2\pi \cos[\omega(t - s)] = \frac{A^2}{2} \cos(\omega\tau) \triangleq R_{\mathbf{x}}(\tau). \end{aligned}$$

Quindi, il p.s. $x(t)$ è stazionario in senso debole.

- $x(t) = \mathbf{A} \sin(\omega t)$,

con $\mathbf{A} \sim f_{\mathbf{A}}(a)$ v.a. con media $m_{\mathbf{A}} = 0$ e varianza $\sigma_{\mathbf{A}}^2$. La media del p.s. $x(t)$ vale:

$$\mathbf{E}[x(t)] = \int_{-\infty}^{+\infty} a \sin(\omega t) f_{\mathbf{A}}(a) da = \sin(\omega t) \int_{-\infty}^{+\infty} a f_{\mathbf{A}}(a) da = \sin(\omega t) m_{\mathbf{A}} = 0,$$

mentre la funzione di covarianza:

$$\begin{aligned} R_{\mathbf{x}}(t, s) &= \mathbf{E}[x(t)x(s)] = \int_{-\infty}^{+\infty} a^2 \sin(\omega t) \sin(\omega s) f_{\mathbf{A}}(a) da \\ &= \sin(\omega t) \sin(\omega s) \int_{-\infty}^{+\infty} a^2 f_{\mathbf{A}}(a) da = \sigma_{\mathbf{A}}^2 \sin(\omega t) \sin(\omega s) \\ &= \frac{\sigma_{\mathbf{A}}^2}{2} [\cos(\omega(t - s)) - \cos(\omega(t + s))], \end{aligned}$$

in cui l'ultima uguaglianza è dovuta alla formula di Werner $\sin(\alpha) \sin(\beta) = \frac{1}{2} [\cos(\alpha - \beta) - \cos(\alpha + \beta)]$. Quindi, il p.s. $x(t)$ non è stazionario in senso debole.

L'ultimo esempio mostra come un processo puramente deterministico non sia necessariamente stazionario.

2.2.2 Rumore bianco

Abbiamo già definito come rumore bianco $x(t)$ una sequenza indipendente di variabili aleatorie. Quindi $x(t)$ è tale per cui $x(t_1)$ e $x(t_2)$ sono v.a. indipendenti se $t_1 \neq t_2$. Un siffatto processo è totalmente imprevedibile nel futuro.

Spesso, si estende il concetto di rumore bianco a tutti quei processi composti da sequenze incorrelate di variabili aleatorie:

$$\mathbf{E}[(x(t_1) - m_{\mathbf{x}}(t_1))(x(t_2) - m_{\mathbf{x}}(t_2))] = 0, \quad \forall t_1 \neq t_2.$$

Supponendo che le $x(t_i)$ siano *i.i.d.* con media nulla e varianza $\sigma_{\mathbf{x}}^2$, si ha:

$$m_{\mathbf{x}}(t) = \mathbf{E}[x(t)] = 0,$$

e

$$R_{\mathbf{x}}(t, s) = \begin{cases} \mathbf{E}[x(t)^2] = \sigma_{\mathbf{x}}^2 & \text{se } t = s \\ 0 & \text{se } t \neq s \end{cases} = \sigma_{\mathbf{x}}^2 \delta(t - s),$$

in cui $\delta(\cdot)$ denota la delta di Dirac nel caso di p.s. tempo-continui e l'impulso unitario nel caso di p.s. tempo-discreto:

$$\delta(\tau) = \begin{cases} 1 & \text{se } \tau = 0 \\ 0 & \text{se } \tau \neq 0 \end{cases}. \quad (2.10)$$

Per cui, posto $\tau = t - s$, si ha:

$$R_{\mathbf{x}}(\tau) = \sigma_{\mathbf{x}}^2 \delta(\tau). \quad (2.11)$$

Un esempio di p.s. bianco tempo-discreto è rappresentato dalla trasmissione binaria. Sia $x(t) \in \{-1, 1\}$, con $t \in \mathbb{N}$ e $x(t_1)$, $x(t_2)$ indipendenti se $t_1 \neq t_2$. Supponendo entrambi i valori $-1, 1$ equiprobabili, la statistica del primo ordine di $x(t)$ è data da:

$$f_{\mathbf{x}}(x; t) = \frac{1}{2} \delta(x - 1) + \frac{1}{2} \delta(x + 1).$$

La media del p.s. $x(t)$ vale:

$$\mathbf{E}[x(t)] = \int_{-\infty}^{+\infty} x \left(\frac{1}{2} \delta(x - 1) + \frac{1}{2} \delta(x + 1) \right) dx = \frac{1}{2} - \frac{1}{2} = 0,$$

in cui si è sfruttata la proprietà della delta di Dirac:

$$\int_{-\infty}^{+\infty} g(x)\delta(x - \alpha)dx = g(\alpha).$$

La varianza del p.s. $x(t)$ vale:

$$R_{\mathbf{x}}(t, t) = \mathbf{E}[x(t)^2] = \int_{-\infty}^{+\infty} x^2 \left(\frac{1}{2}\delta(x - 1) + \frac{1}{2}\delta(x + 1) \right) dx = \frac{1}{2} + \frac{1}{2} = 1.$$

Poiché $x(t_1)$ e $x(t_2)$ sono indipendenti se $t_1 \neq t_2$, esse sono anche incorrelate, per cui:

$$R_{\mathbf{x}}(t_1, t_2) = \mathbf{E}[x(t_1)x(t_2)] = 0.$$

Riassumendo, la funzione di covarianza di $x(t)$ vale:

$$R_{\mathbf{x}}(\tau) = \begin{cases} 1 & \text{se } \tau = 0 \\ 0 & \text{se } \tau \neq 0 \end{cases}$$

I processi stocastici bianchi giocheranno un ruolo fondamentale nel seguito della trattazione, poiché si vedrà che un'ampia classe di p.s. stazionari può essere ottenuta da un generico rumore bianco attraverso un opportuno filtro lineare.

Notaione. Spesso, per indicare che un p.s. $e(t)$ è bianco, con valor medio m_e e varianza σ_e^2 , si scrive:

$$e(t) \sim WN(m_e, \sigma_e^2).$$

Se, inoltre, $e(t)$ ha una distribuzione Gaussiana, $\forall t$, si scrive:

$$e(t) \sim WGN(m_e, \sigma_e^2).$$

2.2.3 Processo di Wiener

Un processo di Wiener $w(t)$, detto anche *moto browniano*, ha le seguenti caratteristiche:

- $w(0) = 0$;

- $\mathbf{E}[w(t)] = 0$;
- $R_w(t_1, t_2) = \begin{cases} \alpha t_2 & \text{se } t_1 \geq t_2 \\ \alpha t_1 & \text{se } t_1 \leq t_2 \end{cases}$
- per t fissato, $w(t) \sim N(0, \alpha t)$.

Si tratta, quindi, di un processo gaussiano non stazionario. Si noti che la varianza di $w(t)$ vale $R_w(t, t) = \alpha t$ e quindi la “dispersione” delle traiettorie cresce nel tempo.

Fissati $t_1 > t_2 > t_3$, vale la seguente proprietà:

$$\begin{aligned} \mathbf{E}[(w(t_1) - w(t_2))(w(t_2) - w(t_3))] &= R_w(t_1, t_2) - R_w(t_1, t_3) \\ &\quad - R_w(t_2, t_2) + R_w(t_2, t_3) \\ &= \alpha t_2 - \alpha t_3 - \alpha t_2 + \alpha t_3 = 0. \end{aligned}$$

Per questo, un processo di Wiener viene detto processo a incrementi indipendenti. Tale processo può essere ottenuto integrando un processo a variabili indipendenti.

2.2.4 Processi esponenzialmente correlati

Sono processi stocastici $x(t)$ tempo-discreto, stazionari, la cui funzione di covarianza ha la forma:

$$R_x(\tau) = \sigma_x^2 a^{|\tau|}, \quad \tau \in \mathbb{Z}, \quad (2.12)$$

con $|a| < 1$. La varianza del p.s. $x(t)$ vale $R_x(0) = \sigma_x^2$. Si noti che:

- se $a \rightarrow 1$, $x(t)$ tende ad un processo puramente deterministico;
- se $a \rightarrow 0$, $x(t)$ tende ad un processo a variabili incorrelate.

La funzione di covarianza rappresentata in Figura 2.2 è proprio del tipo (2.12), con $\sigma_x^2 = 1$ ed $a = -0.7$.

Fissato un generico istante temporale t , si considerino le v.a. $\mathbf{x}(t)$ e $\mathbf{x}(t+1)$. Dal momento che il p.s. $x(t)$ è stazionario, la media e la varianza delle due v.a. sono identiche:

$$\begin{aligned}\mathbf{E}[\mathbf{x}(t)] &= \mathbf{E}[\mathbf{x}(t+1)] = m_{\mathbf{x}} \\ \mathbf{E}[(\mathbf{x}(t) - m_{\mathbf{x}})^2] &= \mathbf{E}[(\mathbf{x}(t+1) - m_{\mathbf{x}})^2] = \sigma_{\mathbf{x}}^2\end{aligned}$$

Calcoliamo, ora, la covarianza incrociata tra $\mathbf{x}(t+1)$ e $\mathbf{x}(t)$:

$$\mathbf{E}[(\mathbf{x}(t+1) - m_{\mathbf{x}})(\mathbf{x}(t) - m_{\mathbf{x}})] \triangleq R_{\mathbf{x}}(1) = \sigma_{\mathbf{x}}^2 a,$$

in cui l'ultima uguaglianza si ottiene dalla (2.12) con $\tau = 1$. Allora, il coefficiente di correlazione tra $\mathbf{x}(t+1)$ e $\mathbf{x}(t)$ vale (v. Definizione 1.17):

$$\rho = \frac{\sigma_{\mathbf{x}}^2 a}{\sigma_{\mathbf{x}}^2} = a.$$

Quindi, il parametro a nella (2.12) rappresenta proprio l'indice di correlazione fra due v.a. corrispondenti al p.s. $x(t)$ valutato in due istanti temporali consecutivi. Per cui, è possibile descrivere “qualitativamente” l'andamento di $x(t)$ come segue:

- se $a > 0$, “si conserva il segno” di $x(t) - m_{\mathbf{x}}$;
- se $a < 0$, “cambia il segno” di $x(t) - m_{\mathbf{x}}$.

La dicitura “si conserva il segno” va intesa in senso probabilistico. Vale a dire che, se $a > 0$ e $x(t) > m_{\mathbf{x}}$, la probabilità che sia anche $x(t+1) > m_{\mathbf{x}}$ è maggiore di quella associata all'evento complementare $x(t+1) < m_{\mathbf{x}}$. Tanto più a è vicino ad uno, tanto maggiore sarà la probabilità che $x(t+1) > m_{\mathbf{x}}$. Conderazioni del tutto analoghe valgono per $a < 0$ e/o $x(t) < m_{\mathbf{x}}$.

In Figura 2.3 sono riportate due realizzazioni di p.s. esponenzialmente correlati, entrambi con valor medio pari a 10, ma con coefficienti di correlazione di segno opposto. È evidente come, quando $a = 0.9$, il p.s. $x(t)$ “cambi di segno” (rispetto al proprio valor medio) molto meno spesso di quanto non faccia quando $a = -0.9$.

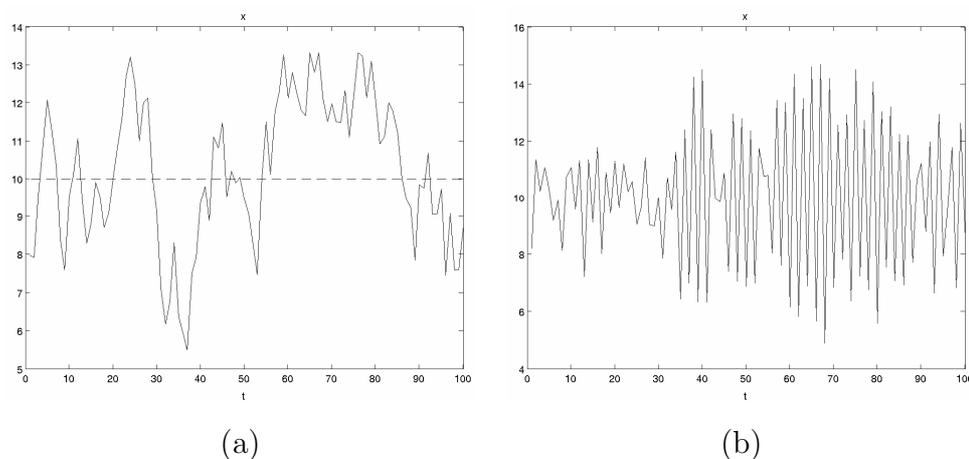


Figura 2.3: Realizzazioni di processi stocastici esponenzialmente correlati: $a = 0.9$ (a) e $a = -0.9$ (b).

I processi stocastici esponenzialmente correlati si ottengono *asintoticamente* come uscite di sistemi dinamici lineari, tempo-invarianti (LTI), con in ingresso un rumore bianco.

Esempio 2.1. Si consideri il p.s. $y(t)$ soluzione dell'equazione alle differenze:

$$y(t+1) = ay(t) + e(t), \quad |a| < 1,$$

dove $e(t)$ rappresenta un p.s. bianco, a media nulla e varianza σ_e^2 . Possiamo pensare che $y(t)$ sia l'uscita di un sistema LTI alimentato dal rumore $e(t)$ (vedi Figura 2.4). Vogliamo calcolare la media $m_y(t)$ e la funzione di

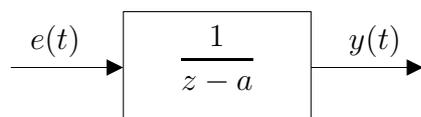


Figura 2.4.

covarianza $R_y(t+\tau, t)$, per poi verificare se il p.s. è stazionario. Dalla teoria dei sistemi, sappiamo che l'uscita di un sistema LTI è data dalla convoluzione della risposta impulsiva con l'ingresso. Per cui, nell'esempio in esame:

$$y(t) = \sum_{i=0}^{\infty} a^i e(t-i-1).$$

Si noti che, poiché $|a| < 1$, la serie $\sum_{i=0}^{\infty} a^i$ converge. Allora:

1. $m_{\mathbf{y}}(t) = \mathbf{E} \left[\sum_{i=0}^{\infty} a^i e(t-i-1) \right]$. Sfruttando la proprietà di linearità dell'operatore di valore atteso, e ricordando che $e(t)$ ha media nulla, otteniamo:

$$m_{\mathbf{y}}(t) = \sum_{i=0}^{\infty} a^i \mathbf{E} [e(t-i-1)] = 0.$$

2. $R_{\mathbf{y}}(t+\tau, t) = \mathbf{E} \left[\left(\sum_{i=0}^{\infty} a^i e(t+\tau-i-1) \right) \left(\sum_{j=0}^{\infty} a^j e(t-j-1) \right) \right]$. Ancora una volta, per la linearità dell'operatore di valore atteso, otteniamo:

$$\begin{aligned} R_{\mathbf{y}}(t+\tau, t) &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} a^i a^j \mathbf{E} [e(t+\tau-i-1)e(t-j-1)] \\ &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} a^{i+j} R_e(\tau-i+j), \end{aligned} \tag{2.13}$$

in cui l'ultima uguaglianza deriva dalla definizione di funzione di covarianza. Ma, siccome $e(t)$ è bianco, la sua funzione di covarianza vale:

$$R_e(\nu) = \begin{cases} \sigma_e^2 & \text{se } \nu = 0 \\ 0 & \text{se } \nu \neq 0 \end{cases}$$

Per cui, nella seconda sommatoria della (2.13) rimangono solo i termini con indici i, j tali che $\tau - i + j = 0$, ossia $j = i - \tau$. Inoltre, dal momento che l'indice j assume solo valori non negativi, nel caso in cui $\tau \geq 0$, affinché possa essere $j = i - \tau$, occorre che $i \geq \tau$. Da queste

considerazioni segue che:

$$\begin{aligned}
 R_{\mathbf{y}}(t + \tau, t) &= \begin{cases} \sum_{i=\tau}^{\infty} a^{2i-\tau} \sigma_e^2 & \text{se } \tau \geq 0 \\ \sum_{i=0}^{\infty} a^{2i-\tau} \sigma_e^2 & \text{se } \tau < 0 \end{cases} = \begin{cases} a^\tau \sigma_e^2 \sum_{i=\tau}^{\infty} a^{2i-2\tau} & \text{se } \tau \geq 0 \\ a^{-\tau} \sigma_e^2 \sum_{i=0}^{\infty} a^{2i} & \text{se } \tau < 0 \end{cases} \\
 &\stackrel{[1]}{=} \begin{cases} a^\tau \sigma_e^2 \sum_{k=0}^{\infty} (a^2)^k & \text{se } \tau \geq 0 \\ a^{-\tau} \sigma_e^2 \sum_{i=0}^{\infty} (a^2)^i & \text{se } \tau < 0 \end{cases} \stackrel{[2]}{=} \begin{cases} \frac{\sigma_e^2}{1-a^2} a^\tau & \text{se } \tau \geq 0 \\ \frac{\sigma_e^2}{1-a^2} a^{-\tau} & \text{se } \tau < 0 \end{cases} \\
 &= \frac{\sigma_e^2}{1-a^2} a^{|\tau|}
 \end{aligned}$$

L'uguaglianza [1] si ottiene dal cambiamento di indice $k = i - \tau$, mentre la [2] discende dalla somma della serie $\sum_{i=0}^{\infty} p^i = \frac{1}{1-p}$, se $|p| < 1$.

Riassumendo:

- $m_{\mathbf{y}}(t) = 0$;
- $R_{\mathbf{y}}(t + \tau, t) = \frac{\sigma_e^2}{1-a^2} a^{|\tau|}$;

per cui, $y(t)$ è stazionario in senso debole. △

Osservazione 2.1. È opportuno notare nell'analisi precedente si è considerata la condizione iniziale del sistema all'istante $t = -\infty$. Questo equivale a considerare l'uscita a regime del sistema LTI, cioè a transitorio esaurito. Ciò è giustificato dal fatto che il sistema che ha generato $y(t)$ è asintoticamente stabile, per cui, dopo un tempo sufficientemente lungo, l'influenza della condizione iniziale sull'uscita diventa trascurabile. Se si desidera analizzare anche il transitorio (questo equivale a considerare la condizione iniziale all'istante $t = 0$), l'uscita $y(t)$ è un p.s. solo *asintoticamente* stazionario:

- $\lim_{t \rightarrow \infty} m_{\mathbf{y}}(t)$ è costante;
- $\lim_{t \rightarrow \infty} R_{\mathbf{y}}(t + \tau, t)$ dipende solo da τ .

2.3 Analisi in frequenza

Sia $x(t)$ un processo stocastico tempo-discreto, stazionario, con funzione di covarianza $R_{\mathbf{x}}(\tau)$. Introduciamo alcuni utili strumenti per l'analisi dei processi stocastici nel dominio della frequenza.

Definizione 2.10. Si definisce *spettro* del processo stocastico stazionario $x(t)$ la funzione di variabile complessa $z \in \mathbb{C}$:

$$\phi_x(z) = \sum_{\tau=-\infty}^{\infty} R_{\mathbf{x}}(\tau) z^{-\tau}.$$

Definizione 2.11. Si definisce *densità spettrale di potenza*, del processo stocastico stazionario $x(t)$, lo spettro di $x(t)$ calcolato lungo la circonferenza unitaria:

$$\phi_x(e^{j\omega}) = \sum_{\tau=-\infty}^{\infty} R_{\mathbf{x}}(\tau) e^{-j\omega\tau}.$$

Si noti che, dalla definizione, segue che la densità spettrale di potenza è una funzione periodica, di periodo 2π , della variabile reale ω . Per questo, ci si limita a considerarla solo per $\omega \in [-\pi, \pi]$.

Siano $x(t)$ e $y(t)$ due processi stocastici congiuntamente stazionari e sia $R_{\mathbf{xy}}(\tau)$ la loro funzione di cross-covarianza.

Definizione 2.12. Si definisce *spettro incrociato* dei processi stocastici $x(t)$ e $y(t)$ la funzione di variabile complessa $z \in \mathbb{C}$:

$$\phi_{xy}(z) = \sum_{\tau=-\infty}^{\infty} R_{\mathbf{xy}}(\tau) z^{-\tau}.$$

Lo spettro $\phi_x(z)$ di un p.s. $x(t)$ è la *trasformata z bilaterale* della sua funzione di covarianza.

Noto lo spettro di un p.s. $x(t)$, è possibile risalire alla sua funzione di covarianza $R_{\mathbf{x}}(\tau)$? Nel seguito, \oint indica l'integrale di una funzione di variabile complessa, calcolato lungo la circonferenza unitaria. Allora, calcoliamo:

$$\frac{1}{2\pi j} \oint \phi_x(z) z^{k-1} dz. \quad (2.14)$$

Poiché l'integrale va valutato lungo la circonferenza unitaria, possiamo porre $z = e^{j\omega}$, con $\omega \in [-\pi, \pi]$. Da cui, dato che $dz = je^{j\omega}d\omega$, la 2.14 diventa:

$$\begin{aligned}
\frac{1}{2\pi j} \oint \phi_x(z)z^{k-1}dz &= \frac{1}{2\pi j} \int_{-\pi}^{\pi} \phi_x(e^{j\omega})e^{j\omega(k-1)}je^{j\omega}d\omega \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} \phi_x(e^{j\omega})e^{j\omega k}d\omega \\
[\text{dalla Def 2.11}] &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{\tau=-\infty}^{\infty} R_{\mathbf{x}}(\tau)e^{-j\omega\tau}e^{j\omega k}d\omega \\
&= \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} R_{\mathbf{x}}(\tau) \int_{-\pi}^{\pi} e^{j\omega(k-\tau)}d\omega \\
&\stackrel{[1]}{=} \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} R_{\mathbf{x}}(\tau)2\pi\delta_{k-\tau} \\
&= R_{\mathbf{x}}(k).
\end{aligned} \tag{2.15}$$

Si noti che la [1] dipende dal fatto che, $n \in \mathbb{N}$:

$$\int_{-\pi}^{\pi} e^{j\omega n}d\omega = \begin{cases} 2\pi & \text{se } n = 0 \\ 0 & \text{se } n \neq 0 \end{cases}$$

mentre δ_k indica la delta di Kronecker:

$$\delta_k = \begin{cases} 1 & \text{se } k = 0 \\ 0 & \text{se } k \neq 0 \end{cases} \tag{2.16}$$

Quindi, dallo spettro è sempre possibile calcolare la funzione di covarianza:

$$R_{\mathbf{x}}(\tau) = \frac{1}{2\pi j} \oint \phi_x(z)z^{\tau-1}dz, \quad \tau \in \mathbb{Z}.$$

Equivalentemente, la funzione di covarianza può essere ricavata anche a partire dalla densità spettrale di potenza (vedi seconda uguaglianza in (2.15)):

$$R_{\mathbf{x}}(\tau) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \phi_x(e^{j\omega})e^{j\omega\tau}d\omega, \quad \tau \in \mathbb{Z}.$$

Nel caso in cui $\tau = 0$, la relazione precedente diventa:

$$R_{\mathbf{x}}(0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \phi_x(e^{j\omega})d\omega. \tag{2.17}$$

La (2.17) fornisce una giustificazione della denominazione “densità spettrale di potenza” per la $\phi_x(e^{j\omega})$. Infatti, ricordando che $R_{\mathbf{x}}(0)$ corrisponde proprio alla varianza del p.s. $x(t)$ (e quindi alla potenza del segnale $x(t)$), la $\phi_x(e^{j\omega})$ descrive esattamente come si distribuisce la potenza di $x(t)$ alle varie frequenze.

Lo spettro e la densità spettrale godono delle seguenti proprietà:

1. $\phi_x(z) = \phi_x^T(z^{-1})$;
2. $\phi_x(e^{j\omega}) = \phi_x^T(e^{-j\omega})$;
3. $\phi_x(e^{j\omega}) \geq 0$;
4. $\phi_{xy}(z) = \phi_{yx}^T(z^{-1})$.

Dimostrazione

1. Direttamente dalla definizione di spettro, si ha:

$$\begin{aligned} \phi_x(z) &= \sum_{k=-\infty}^{\infty} R_{\mathbf{x}}(k)z^{-k} \stackrel{[1]}{=} \sum_{k=-\infty}^{\infty} R_{\mathbf{x}}^T(-k)z^{-k} = [\text{ponendo } \tau = -k] \\ &= \left(\sum_{\tau=-\infty}^{\infty} R_{\mathbf{x}}(\tau) (z^{-1})^{-\tau} \right)^T = \phi_x^T(z^{-1}), \end{aligned}$$

in cui la [1] deriva dalla Proprietà 1 a pag. 41 della funzione di covarianza.

2. Si ottiene dalla precedente, ponendo $z = e^{j\omega}$.
3. Limitiamoci al caso scalare. Supponiamo inoltre (per semplicità), che esistano due costanti reali $C > 0$ e $\alpha \in (0, 1)$ tali che $|r_{\mathbf{x}}(\tau)| \leq C\alpha^{|\tau|}$. Dalla Proprietà 3 di pag. 41 della funzione di covarianza, si ha che la matrice:

$$P(N) = \begin{bmatrix} r_{\mathbf{x}}(0) & r_{\mathbf{x}}(1) & \dots & r_{\mathbf{x}}(N-1) \\ r_{\mathbf{x}}(-1) & r_{\mathbf{x}}(0) & \dots & r_{\mathbf{x}}(N-2) \\ \vdots & \vdots & \ddots & \vdots \\ r_{\mathbf{x}}(-N+1) & r_{\mathbf{x}}(-N+2) & \dots & r_{\mathbf{x}}(0) \end{bmatrix}$$

è semidefinita positiva. Questo significa che, preso un qualunque vettore *complesso* $v \in \mathbb{C}^N$, vale $v^*P(N)v \geq 0$, in cui v^* indica il trasposto del vettore coniugato di v :

$$v^* = [\bar{v}_1, \bar{v}_2, \dots, \bar{v}_N].$$

Per cui, posto $v = [1, z^{-1}, \dots, z^{-N+1}]^T$, con $z = e^{j\omega}$, segue:

$$\begin{aligned} 0 &\leq \frac{1}{N} \left(1, \overline{z^{-1}}, \dots, \overline{z^{-N+1}}\right) P(N) \begin{pmatrix} 1 \\ z^{-1} \\ \vdots \\ z^{-N+1} \end{pmatrix} \\ &\stackrel{[1]}{=} \frac{1}{N} (1, z, \dots, z^{N-1}) \begin{pmatrix} r_{\mathbf{x}}(0) + r_{\mathbf{x}}(1)z^{-1} + \dots + r_{\mathbf{x}}(N-1)z^{-N+1} \\ r_{\mathbf{x}}(-1) + r_{\mathbf{x}}(0)z^{-1} + \dots + r_{\mathbf{x}}(N-2)z^{-N+1} \\ \vdots \\ r_{\mathbf{x}}(-N+1) + r_{\mathbf{x}}(-N+2)z^{-1} + \dots + r_{\mathbf{x}}(0)z^{-N+1} \end{pmatrix} \\ &= \frac{1}{N} \left[Nr(0) + \sum_{k=1}^N (N-k)(r(k)z^{-k} + r(-k)z^k) \right] \\ &= \frac{1}{N} \left[Nr(0) + \sum_{k=1}^N N(r(k)z^{-k} + r(-k)z^k) - \sum_{k=1}^N k(r(k)z^{-k} + r(-k)z^k) \right] \\ &\stackrel{[2]}{=} \sum_{k=-N}^N r(k)z^{-k} - \frac{1}{N} \sum_{k=-N}^N |k| r(k)z^{-k}, \end{aligned} \tag{2.18}$$

in cui la [1] discende dalla proprietà dei numeri complessi: $\overline{e^{-j\omega k}} = e^{j\omega k}$, mentre la [2] deriva dalla simmetria della funzione di covarianza (v. equazione (2.8)). Consideriamo il modulo dell'ultima sommatoria nella disuguaglianza precedente. Dalla disuguaglianza triangolare, segue:

$$\left| \frac{1}{N} \sum_{k=-N}^N |k| r(k)z^{-k} \right| \leq \frac{1}{N} \sum_{k=-N}^N ||k| r(k)z^{-k}| = \frac{1}{N} \sum_{k=-N}^N |k| |r(k)| |z^{-k}|$$

Ricordando che la relazione precedente vale per $z = e^{j\omega}$, e dal momento

che $|e^{j\omega}| = 1$, otteniamo:

$$\begin{aligned} \left| \frac{1}{N} \sum_{k=-N}^N |k| r(k) e^{-j\omega k} \right| &\leq \frac{1}{N} \sum_{k=-N}^N |k| |r(k)| \\ &\leq \frac{1}{N} \sum_{k=-N}^N |k| C\alpha^{|k|} = \frac{2C}{N} \sum_{k=1}^N k\alpha^k \end{aligned} \quad (2.19)$$

poiché, per ipotesi, $|r_{\mathbf{x}}(\tau)| \leq C\alpha^{|\tau|}$. Vogliamo mostrare che la quantità a destra (e di conseguenza il lato sinistro, essendo esso sempre positivo o al più nullo) della (2.19) tende a zero quando N tende all'infinito. Per fare ciò calcoliamo:

$$\begin{aligned} (1-\alpha)^2 \sum_{k=1}^N k\alpha^k &= \sum_{k=1}^N k\alpha^k - 2\alpha \sum_{k=1}^N k\alpha^k + \alpha^2 \sum_{k=1}^N k\alpha^k \\ &= \alpha + 2\alpha^2 - 2\alpha^2 + \sum_{k=3}^N [k - 2(k-1) + k - 2]\alpha^k \\ &\quad - 2N\alpha^{N+1} + (N-1)\alpha^{N+1} + N\alpha^{N+2} \\ &= \alpha - (N+1)\alpha^{N+1} + N\alpha^{N+2}. \end{aligned}$$

Dividendo per $(1-\alpha)^2$ ambo i membri della precedente uguaglianza, otteniamo:

$$\sum_{k=1}^N k\alpha^k = \frac{\alpha - (N+1)\alpha^{N+1} + N\alpha^{N+2}}{(1-\alpha)^2}.$$

Ricordando che, per ipotesi, $0 < \alpha < 1$, per N che tende all'infinito:

$$\sum_{k=1}^N k\alpha^k \rightarrow \frac{\alpha}{(1-\alpha)^2}$$

e quindi:

$$\frac{2C}{N} \sum_{k=1}^N k\alpha^k \rightarrow 0.$$

Concludendo, valutando la (2.18) per $z = e^{j\omega}$ e lasciando tendere N all'infinito otteniamo proprio la tesi:

$$\phi_x(e^{j\omega}) = \sum_{k=-\infty}^{\infty} r_{\mathbf{x}}(k) \geq 0.$$

4. Direttamente dalla definizione di spettro incrociato, si ha:

$$\begin{aligned}\phi_{xy}(z) &= \sum_{k=-\infty}^{\infty} R_{\mathbf{x}\mathbf{y}}(k) z^{-k} \stackrel{[1]}{=} \sum_{k=-\infty}^{\infty} R_{\mathbf{y}\mathbf{x}}^T(-k) z^{-k} = [\text{ponendo } \tau = -k] \\ &= \left(\sum_{\tau=-\infty}^{\infty} R_{\mathbf{y}\mathbf{x}}(\tau) (z^{-1})^{-\tau} \right)^T = \phi_{yx}^T(z^{-1}),\end{aligned}$$

in cui la [1] deriva dalla Proprietà 4 a pag. 42 della funzione di cross-covarianza.

□

Nel caso in cui $x(t)$ sia un processo stocastico scalare, il suo spettro soddisfa la relazione:

$$\phi_x(z) = \phi_x(z^{-1}), \quad (2.20)$$

mentre la densità spettrale è una funzione simmetrica di ω e sempre positiva (o al più nulla):

$$\phi_x(e^{j\omega}) = \phi_x(e^{-j\omega}), \quad (2.21)$$

$$\phi_x(e^{j\omega}) \geq 0, \quad \forall \omega. \quad (2.22)$$

Ricordando che la densità spettrale è anche una funzione periodica di periodo 2π , la proprietà di simmetria consente di limitarci a studiarla nell'intervallo $[0, \pi]$.

Esempio 2.2. Abbiamo visto che se $x(t)$ è un processo stocastico bianco, la sua funzione di covarianza vale:

$$R_{\mathbf{x}}(\tau) = \begin{cases} \sigma_{\mathbf{x}}^2 & \text{se } \tau = 0 \\ 0 & \text{se } \tau \neq 0 \end{cases}$$

Per cui, direttamente dalle definizioni di spettro e densità spettrale otteniamo:

$$\phi_x(z) = \sigma_{\mathbf{x}}^2, \quad \forall z \in \mathbb{C},$$

e

$$\phi_x(e^{j\omega}) = \sigma_{\mathbf{x}}^2, \quad \forall \omega \in [-\pi, \pi].$$

Quindi, la densità spettrale di un p.s. bianco è costante. Ciò vuol dire che la potenza del segnale è distribuita in egual misura a tutte le frequenze. Questo spiega la denominazione “bianco” per siffatti processi, in analogia con la luce bianca (che contiene tutti i “colori”, ossia tutte le frequenze, in egual misura). \triangle

Esempio 2.3. Si consideri il processo stocastico $x(t)$, esponenzialmente correlato, con funzione di covarianza:

$$R_{\mathbf{x}}(\tau) = \sigma_{\mathbf{x}}^2 a^{|\tau|}, \quad \tau \in \mathbb{Z}, \quad (2.23)$$

con $|a| < 1$. Desideriamo calcolare lo spettro e la densità spettrale di $x(t)$. Dalla definizione:

$$\begin{aligned} \phi_x(z) &= \sum_{\tau=-\infty}^{\infty} R_{\mathbf{x}}(\tau) z^{-\tau} = \sum_{\tau=-\infty}^{\infty} \sigma_{\mathbf{x}}^2 a^{|\tau|} z^{-\tau} \\ &= \sigma_{\mathbf{x}}^2 \left(\sum_{\tau=-\infty}^{-1} a^{-\tau} z^{-\tau} + \sum_{\tau=0}^{\infty} a^{\tau} z^{-\tau} \right) \\ &= \sigma_{\mathbf{x}}^2 \left(\sum_{k=1}^{\infty} a^k z^k + \sum_{\tau=0}^{\infty} a^{\tau} z^{-\tau} \right) \\ &= \sigma_{\mathbf{x}}^2 \left(\sum_{k=0}^{\infty} (az)^k + \sum_{\tau=0}^{\infty} (az^{-1})^{\tau} - 1 \right). \end{aligned}$$

Poichè la serie $\sum_{k=0}^{\infty} p^k$ converge al valore $\frac{1}{1-p}$, se $|p| < 1$, affinchè entrambe le serie precedenti convergano occorre sia:

$$\begin{aligned} |az| &< 1, \\ |az^{-1}| &< 1. \end{aligned}$$

Le precedenti disuguaglianze definiscono una regione del piano ove lo spettro converge:

$$|a| < |z| < \frac{1}{|a|}.$$

In Figura 2.5 è rappresentata la regione del piano complesso entro cui le serie convergono. Si può notare come essa sia una corona circolare centrata nell'origine e (dal momento che $|a| < 1$) contenente sempre la circonferenza unitaria.

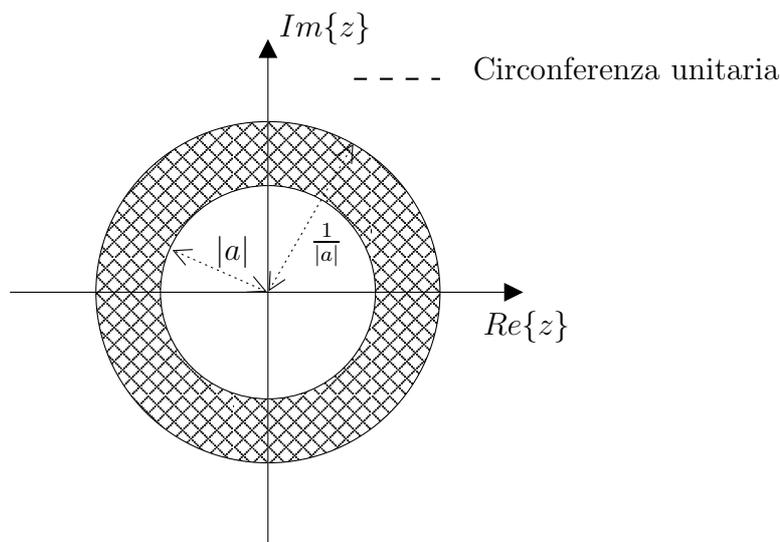


Figura 2.5: Regione di convergenza dello spettro $\phi_x(z)$.

Nella regione in cui lo spettro converge, esso vale:

$$\begin{aligned}
 \phi_x(z) &= \sigma_x^2 \left(\frac{1}{1-az} + \frac{1}{1-az^{-1}} - 1 \right) \\
 &= \frac{\sigma_x^2(1-az^{-1} + 1-az - 1 - a^2 + az + az^{-1})}{(1-az^{-1})(1-az)} \\
 &= \frac{\sigma_x^2(1-a^2)}{(1-az^{-1})(1-az)}.
 \end{aligned} \tag{2.24}$$

Poichè la regione di convergenza contiene sempre la circonferenza unitaria, è possibile calcolare la densità spettrale valutando l'espressione precedente per $z = e^{j\omega}$:

$$\begin{aligned}
 \phi_x(e^{j\omega}) &= \frac{\sigma_x^2(1-a^2)}{(1-ae^{-j\omega})(1-ae^{j\omega})} = \frac{\sigma_x^2(1-a^2)}{1+a^2-a(e^{j\omega}+e^{-j\omega})} \\
 &= \frac{\sigma_x^2(1-a^2)}{1+a^2-2a\cos\omega},
 \end{aligned} \tag{2.25}$$

in cui, nell'ultima uguaglianza si è usata la formula di Eulero $\cos \omega = \frac{e^{j\omega} + e^{-j\omega}}{2}$.

Si noti come lo spettro e la densità rispettino le proprietà 1-3 di pag. 56.

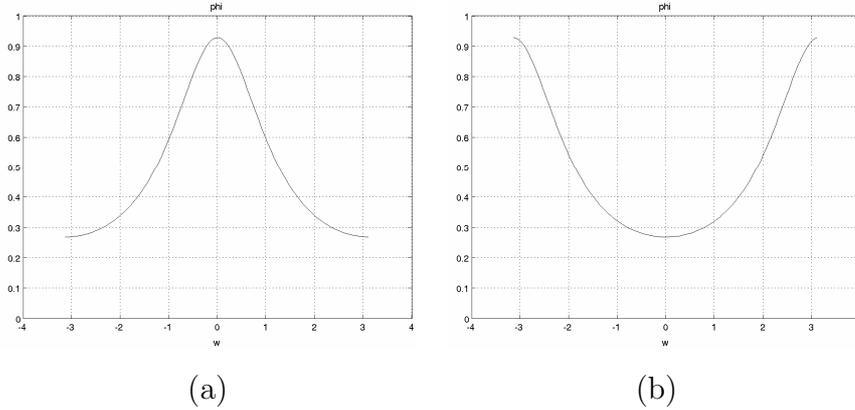


Figura 2.6: Densità spettrale di un processo esponenzialmente correlato con indice di correlazione $a = 0.4$ (a) e $a = -0.4$ (b).

In Figura 2.6 sono riportati i grafici della densità (2.25) per $a = 0.4$ e per $a = -0.4$. Si osservi come, quando il coefficiente di correlazione è positivo, la potenza del segnale sia concentrata a basse frequenze, mentre il contrario accade per valori negativi di a . Questo vuol dire che, per valori positivi di a , il processo, in media, presenta “minori oscillazioni”, rispetto al caso con a negativo (vedi Paragrafo 2.2.4). \triangle

2.4 Sistemi lineari stocastici

In questo paragrafo ci proponiamo di studiare cosa accade quando un processo stocastico viene filtrato attraverso un sistema lineare tempo-invariante. In particolare, concentreremo la nostra attenzione alla rappresentazione ingresso/uscita di sistemi LTI *asintoticamente stabili*, alimentati da p.s. *stazionari*.

Con riferimento alla Figura 2.7, sia:

$$G(z) = \sum_{k=0}^{\infty} g_k z^{-k} \quad (2.26)$$

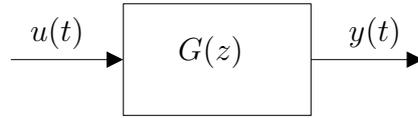


Figura 2.7: Il p.s. $y(t)$ è l'uscita di un sistema LTI avente in ingresso il p.s. $u(t)$.

la matrice di trasferimento di un sistema LTI.² Vale a dire, $G(z)$ è la trasformata zeta (monolaterale) della risposta impulsiva $\{g_k\}_{k=0}^{\infty}$ del sistema in esame. Ne consegue che l'uscita vale:

$$y(t) = \sum_{k=0}^{\infty} g_k u(t-k), \quad (2.27)$$

in cui $u(t)$ è il segnale posto in ingresso al sistema.

Nel caso in cui l'ingresso sia un processo stocastico, anche l'uscita sarà un processo stocastico. La relazione precedente lega le singole realizzazioni dell'uscita alle singole realizzazioni dell'ingresso. Sarebbe più utile ottenere una descrizione delle proprietà statistiche del p.s. $y(t)$ a partire da quelle dell'ingresso $u(t)$. Il seguente teorema fornisce la relazione cercata.

Teorema 4. *Sia $G(z)$ la matrice di trasferimento di un sistema LTI asintoticamente stabile. Sia $u(t)$ un processo stocastico stazionario con media m_u e spettro $\phi_u(z)$, posto in ingresso al sistema. Allora, l'uscita $y(t)$ è un processo stocastico asintoticamente stazionario. Indicati con m_y e $\phi_y(z)$ la media e lo spettro di $y(t)$, si ha infatti:*

1. $m_y = G(1)m_u$;
2. $\phi_y(z) = G(z)\phi_u(z)G^T(z^{-1})$.

Inoltre, $y(t)$ e $u(t)$ sono congiuntamente stazionari con:

$$3. R_{yu}(\tau) = \sum_{k=0}^{\infty} g_k R_u(\tau-k), \quad R_y(\tau) = \sum_{k=0}^{\infty} R_{yu}(\tau+k) g_k^T;$$

²In generale, se $u \in \mathbb{R}^m$ e $y \in \mathbb{R}^p$, allora $G(z) \in \mathbb{R}^{p \times m}$. Nel caso in cui $m = p = 1$, $G(z) \in \mathbb{R}$ prende il nome di funzione di trasferimento.

$$4. \phi_{yu}(z) = G(z)\phi_u(z), \quad \phi_y(z) = \phi_{yu}(z)G^T(z^{-1}).$$

Dimostrazione

1. Utilizzando la (2.27) e sfruttando la linearità dell'operatore di valore atteso, otteniamo:

$$m_{\mathbf{y}} = \mathbf{E}[y(t)] = \sum_{k=0}^{\infty} g_k \mathbf{E}[u(t-k)] \stackrel{[1]}{=} \left(\sum_{k=0}^{\infty} g_k \right) m_{\mathbf{u}} \stackrel{[2]}{=} G(1)m_{\mathbf{u}}.$$

La [1] discende dalla stazionarietà del p.s. $u(t)$, mentre la [2] deriva dalla definizione di funzione di trasferimento (v. eq. (2.26)), e sfrutta l'ipotesi di asintotica stabilità del sistema LTI.

2. Supponiamo, senza perdita di generalità, che $m_{\mathbf{u}} = m_{\mathbf{y}} = 0$ (altrimenti quanto segue vale ugualmente considerando i segnali $y(t) - m_{\mathbf{y}}$ e $u(t) - m_{\mathbf{u}}$). Innanzitutto, verifichiamo che il p.s. di uscita $y(t)$ è asintoticamente stazionario. Abbiamo già visto che (a regime) la sua media è costante. Calcoliamo la funzione di covarianza a partire dalla definizione e usando la (2.27):

$$\begin{aligned} R_{\mathbf{y}}(t+\tau, t) &= \mathbf{E} \left[\left(\sum_{k=0}^{\infty} g_k u(t+\tau-k) \right) \left(\sum_{l=0}^{\infty} g_l u(t-l) \right)^T \right] \\ &= \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} g_k R_{\mathbf{u}}(\tau-k+l) g_l^T. \end{aligned}$$

Quindi, $R_{\mathbf{y}}(t+\tau, t) = R_{\mathbf{y}}(\tau)$, per cui $y(t)$ è asintoticamente stazionario, nel senso dell'Osservazione 2.1. Dall'espressione di $R_{\mathbf{y}}(\tau)$ è possibile calcolare lo spettro di $y(t)$:

$$\phi_y(z) = \sum_{\tau=-\infty}^{\infty} R_{\mathbf{y}}(\tau) z^{-\tau} = \sum_{\tau=-\infty}^{\infty} \left(\sum_{k=0}^{\infty} \sum_{l=0}^{\infty} g_k R_{\mathbf{u}}(\tau-k+l) g_l^T \right) z^{-\tau}.$$

Notando che è possibile riscrivere $z^{-\tau} = z^{-(\tau-k+l)} z^{-k+l}$, l'equazione

precedente diventa:

$$\begin{aligned}
\phi_y(z) &= \sum_{\tau=-\infty}^{\infty} \left(\sum_{k=0}^{\infty} \sum_{l=0}^{\infty} g_k R_{\mathbf{u}}(\tau - k + l) g_l^T \right) z^{-(\tau-k+l)} z^{-k+l} \\
&= [\text{ponendo } m = \tau - k + l] \\
&= \sum_{m=-\infty}^{\infty} \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} g_k R_{\mathbf{u}}(m) g_l^T z^{-m} z^{-k+l} \\
&= \sum_{k=0}^{\infty} g_k z^{-k} \sum_{m=-\infty}^{\infty} R_{\mathbf{u}}(m) z^{-m} \sum_{l=0}^{\infty} g_l^T (z^{-1})^{-l} \\
&= G(z) \phi_u(z) G^T(z^{-1}),
\end{aligned}$$

in cui l'ultima uguaglianza deriva dalla (2.26) e dalla Definizione 2.10 di spettro.

3. Verifichiamo che $y(t)$ ed $u(t)$ sono congiuntamente stazionari:

$$\begin{aligned}
R_{\mathbf{y}\mathbf{u}}(t + \tau, t) &= \mathbf{E} \left[\left(\sum_{k=0}^{\infty} g_k u(t + \tau - k) \right) u^T(t) \right] \\
&= \sum_{k=0}^{\infty} g_k \mathbf{E} [u(t + \tau - k) u^T(t)] \\
&= \sum_{k=0}^{\infty} g_k R_{\mathbf{u}}(\tau - k) = R_{\mathbf{y}\mathbf{u}}(\tau).
\end{aligned}$$

4. Dall'espressione precedente di $R_{\mathbf{y}\mathbf{u}}(\tau)$ discende immediatamente che:

$$\begin{aligned}
\phi_{\mathbf{y}\mathbf{u}}(z) &= \sum_{\tau=-\infty}^{\infty} \left(\sum_{k=0}^{\infty} g_k R_{\mathbf{u}}(\tau - k) \right) z^{-\tau} \\
&= \sum_{k=0}^{\infty} g_k z^{-k} \left(\sum_{\tau=-\infty}^{\infty} R_{\mathbf{u}}(\tau - k) z^{-(\tau-k)} \right) \\
&= \sum_{k=0}^{\infty} g_k z^{-k} \phi_u(z) = G(z) \phi_u(z).
\end{aligned}$$

Le altre relazioni presenti ai punti 3 e 4 si dimostrano in maniera del tutto analoga. \square

Alcune conseguenze notevoli del precedente teorema:

$$\phi_y(e^{j\omega}) = G(e^{j\omega})\phi_u(e^{j\omega})G^T(e^{-j\omega}) \quad (2.28)$$

$$\phi_{uy}(z) = \phi_u^T(z^{-1})G^T(z^{-1}) \quad (2.29)$$

e nel *caso scalare*:

$$\phi_y(z) = G(z)G(z^{-1})\phi_u(z) \quad (2.30)$$

$$\phi_y(e^{j\omega}) = |G(e^{j\omega})|^2 \phi_u(e^{j\omega}) \quad (2.31)$$

$$\phi_{uy}(z) = G(z^{-1})\phi_u(z) \quad (2.32)$$

Prendiamo in esame, ora, il caso particolare in cui il segnale di ingresso $u(t)$ sia un rumore bianco $e(t)$, con media nulla e matrice di covarianza R_e :

$$m_e = 0 \quad (2.33)$$

$$R_e(\tau) = \begin{cases} R_e & \text{se } \tau = 0 \\ 0 & \text{se } \tau \neq 0 \end{cases} \quad (2.34)$$

Indicando con $y(t)$ l'uscita del sistema, si ha (v. eq. (2.28)):

$$\phi_y(e^{j\omega}) = G(e^{j\omega})R_eG^T(e^{-j\omega}).$$

Antitrasformando secondo Fourier (vedi eq. (2.17)) si ha la seguente versione del Teorema di Parseval:

$$R_y(0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \phi_y(e^{j\omega})d\omega = \frac{1}{2\pi} \int_{-\pi}^{\pi} G(e^{j\omega})R_eG^T(e^{-j\omega})d\omega, \quad (2.35)$$

che nel caso scalare diventa:

$$R_y(0) = \sigma_e^2 \frac{1}{2\pi} \int_{-\pi}^{\pi} |G(e^{j\omega})|^2 d\omega. \quad (2.36)$$

2.5 Modelli lineari di processi stocastici

In questo paragrafo, concentreremo la nostra attenzione sui processi stocastici scalari, ottenuti filtrando linearmente rumore bianco. A tale scopo, sia $e(t)$

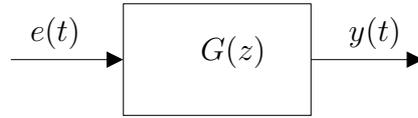


Figura 2.8: Classificazione di p.s. ottenuti filtrando rumore bianco.

un p.s. scalare, bianco, con media nulla e varianza σ_e^2 , posto in ingresso ad un sistema LTI. Si denoti con $y(t)$ l'uscita di tale sistema (v. Figura 2.8). Si è soliti classificare tali processi stocastici in funzione della struttura della funzione di trasferimento $G(z)$.

2.5.1 Processi MA

Si consideri il p.s. scalare $y(t)$, soluzione dell'equazione alle differenze:

$$y(t) = c_0 e(t) + c_1 e(t-1) + \dots + c_m e(t-m). \quad (2.37)$$

Un p.s. generato in questo modo è detto processo *a media mobile* di ordine m (o *moving average*, MA(m)), poichè il segnale $y(t)$ è la media pesata degli ultimi $m+1$ campioni del rumore bianco di ingresso $e(t)$.

Dal momento che $e(t)$ ha media nulla, anche $y(t)$ avrà valor medio uguale a zero:

$$m_y = \mathbf{E}[y(t)] = c_0 \mathbf{E}[e(t)] + c_1 \mathbf{E}[e(t-1)] + \dots + c_m \mathbf{E}[e(t-m)] = 0.$$

Calcoliamo la varianza di $y(t)$:

$$\begin{aligned} \sigma_y^2 &= \mathbf{E}[y(t)^2] = \mathbf{E}[(c_0 e(t) + c_1 e(t-1) + \dots + c_m e(t-m))^2] \\ &= (c_0^2 + c_1^2 + \dots + c_m^2) \sigma_e^2 \end{aligned}$$

in cui l'ultima uguaglianza deriva dalla proprietà del rumore bianco:

$$\mathbf{E}[e(t_1)e(t_2)] = \begin{cases} \sigma_e^2 & \text{se } t_1 = t_2 \\ 0 & \text{se } t_1 \neq t_2 \end{cases}. \quad (2.38)$$

Infine, calcoliamo la funzione di covarianza per un generico $\tau > 0$:

$$\begin{aligned} R_{\mathbf{y}}(\tau) &= \mathbf{E}[y(t+\tau)y(t)] \\ &= \mathbf{E}[(c_0e(t+\tau) + c_1e(t+\tau-1) + \dots + c_me(t+\tau-m)) \times \\ &\quad \times (c_0e(t) + c_1e(t-1) + \dots + c_me(t-m))]. \end{aligned}$$

Ancora una volta, dato che $e(t)$ è un rumore bianco, gli unici prodotti che

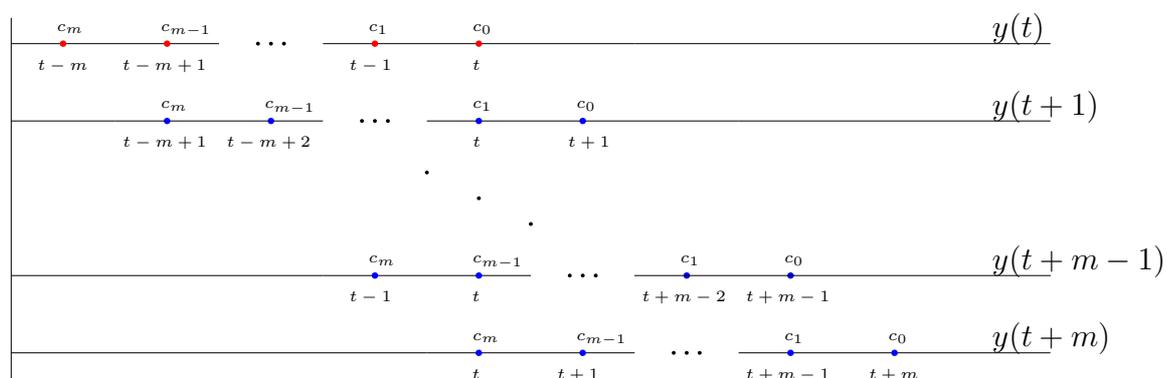


Figura 2.9: Calcolo della funzione di covarianza di un processo $MA(m)$.

danno luogo ad un valore atteso non nullo, sono quelli che coinvolgono il segnale e preso allo stesso istante: $e(\bar{t})e(\bar{t})$. Quindi, con riferimento alla Figura 2.9, si ha:

τ	$R_{\mathbf{y}}(\tau)$
1	$(c_1c_0 + c_2c_1 + \dots + c_mc_{m-1})\sigma_e^2$
2	$(c_2c_0 + c_3c_1 + \dots + c_mc_{m-2})\sigma_e^2$
\vdots	\vdots
$m-1$	$(c_{m-1}c_0 + c_mc_1)\sigma_e^2$
m	$c_mc_0\sigma_e^2$

Infine, se $\tau > m$, la funzione di covarianza è identicamente uguale a zero.

Interpretando z^{-1} come l'operatore di ritardo unitario³, tale che

$$z^{-1}x(t) = x(t - 1),$$

è immediato ricavare la funzione di trasferimento del filtro che genera il p.s. $y(t)$ a partire dall'ingresso $e(t)$. Infatti, la (2.37) può essere riscritta come:

$$y(t) = C(z)e(t), \quad \text{con } C(z) = c_0 + c_1z^{-1} + \dots + c_mz^{-m}.$$

Per cui, moltiplicando e dividendo il polinomio $C(z)$ per z^m , la funzione di trasferimento che lega $e(t)$ a $y(t)$ corrisponde a:

$$G(z) = C(z) = \frac{c_0z^m + c_1z^{m-1} + \dots + c_m}{z^m},$$

e possiede tutti i poli nell'origine. Il sistema rappresentato dalla funzione di trasferimento $G(z)$, quindi, è sempre asintoticamente stabile. Per cui, in virtù del Teorema 4, $y(t)$ è asintoticamente stazionario.

Riassumendo, un p.s. $y(t)$ a media mobile di ordine m , MA(m) presenta le seguenti caratteristiche:

- è un p.s. asintoticamente stazionario;
- ha valor medio nullo, $m_y = 0$ (se $e(t)$ ha media nulla);
- a regime, la sua funzione di covarianza vale:

$$R_y(\tau) = \begin{cases} \left(\sum_{k=|\tau|}^m c_k c_{k-|\tau|} \right) \sigma_e^2 & \text{se } |\tau| \leq m \\ 0 & \text{se } |\tau| > m \end{cases}$$

cioè, le v.a. $y(t_1)$ e $y(t_2)$ sono rigorosamente incorrelate, se distano più di m istanti di tempo, $|t_1 - t_2| > m$.

³Si tratta in realtà di un abuso di notazione, in quanto così facendo z denota sia l'operatore di anticipo, $z x(t) = x(t+1)$, sia la variabile complessa del dominio trasformato in cui sono definiti gli spettri dei processi stocastici di interesse, vedi Definizioni 2.10 e 2.12. Il significato sarà tuttavia chiaro di volta in volta, a seconda del contesto considerato.

Esempio 2.4. Si consideri il p.s. MA(1):

$$y(t) = e(t) + \frac{1}{2}e(t-1)$$

con $e(t)$ rumore bianco, a media nulla e varianza $\sigma_e^2 = 1$.

Dal momento che $e(t)$ è a media nulla, anche $m_y = 0$.

Calcoliamo la varianza di $y(t)$:

$$\begin{aligned} R_y(0) &= \mathbf{E} [y(t)^2] = \mathbf{E} \left[\left(e(t) + \frac{1}{2}e(t-1) \right)^2 \right] \\ &\stackrel{[1]}{=} \mathbf{E} [e(t)^2] + \frac{1}{4}\mathbf{E} [e(t-1)^2] = \left(1 + \frac{1}{4}\right)\sigma_e^2 \\ &= \frac{5}{4}, \end{aligned}$$

in cui la [1] deriva dalla proprietà del rumore bianco (2.38). La funzione di covarianza di $y(t)$ vale:

$$\begin{aligned} R_y(1) &= \mathbf{E} [y(t+1)y(t)] \\ &= \mathbf{E} \left[\left(e(t+1) + \frac{1}{2}e(t) \right) \left(e(t) + \frac{1}{2}e(t-1) \right) \right] \\ &= \frac{1}{2}\sigma_e^2 = \frac{1}{2}, \end{aligned}$$

$$\begin{aligned} R_y(2) &= \mathbf{E} [y(t+2)y(t)] \\ &= \mathbf{E} \left[\left(e(t+2) + \frac{1}{2}e(t+1) \right) \left(e(t) + \frac{1}{2}e(t-1) \right) \right] \\ &= 0 \end{aligned}$$

e lo stesso accade per $R_y(\tau)$, $\tau > 2$ Riassumendo:

$$R_y(\tau) = \begin{cases} \frac{5}{4} & \text{se } |\tau| = 0 \\ \frac{1}{2} & \text{se } |\tau| = 1 \\ 0 & \text{se } |\tau| > 1 \end{cases}.$$

Nel caso di p.s. MA, è immediato calcolare lo spettro e la densità spettrale, direttamente dalle Definizioni 2.10-2.11:

$$\phi_y(z) = \frac{5}{4} + \frac{1}{2}z^{-1} + \frac{1}{2}z,$$

$$\phi_y(e^{j\omega}) = \frac{5}{4} + \frac{1}{2}e^{-j\omega} + \frac{1}{2}e^{j\omega} \quad (2.39)$$

$$= \frac{5}{4} + \frac{e^{j\omega} + e^{-j\omega}}{2} \quad (2.40)$$

$$= \frac{5}{4} + \cos \omega, \quad (2.41)$$

△

Osservazione 2.2. Si noti che si perviene agli stessi risultati, applicando il Teorema 4. Infatti, essendo la funzione di trasferimento da $e(t)$ a $y(t)$:

$$G(z) = \frac{z + \frac{1}{2}}{z}$$

e lo spettro e la densità spettrale del segnale di ingresso $e(t)$:

$$\phi_e(z) = \phi_e(e^{j\omega}) = 1,$$

lo spettro e la densità spettrale del segnale di uscita $y(t)$ risultano:

$$\phi_y(z) = G(z)G(z^{-1}) = 1 + \frac{1}{4} + \frac{1}{2}z^{-1} + \frac{1}{2}z \quad (2.42)$$

$$= \frac{5}{4} + \frac{1}{2}z^{-1} + \frac{1}{2}z, \quad (2.43)$$

$$\phi_y(e^{j\omega}) = |G(e^{j\omega})|^2 = \left| \frac{e^{j\omega} + \frac{1}{2}}{e^{j\omega}} \right|^2 \quad (2.44)$$

$$= \frac{|e^{j\omega} + \frac{1}{2}|^2}{|e^{j\omega}|^2} \stackrel{[1]}{=} \left| \frac{1}{2} + \cos \omega + j \sin \omega \right|^2 \quad (2.45)$$

$$= \left(\frac{1}{2} + \cos \omega \right)^2 + \sin^2 \omega \quad (2.46)$$

$$= \frac{1}{4} + \cos^2 \omega + \cos \omega + \sin^2 \omega \quad (2.47)$$

$$= \frac{5}{4} + \cos \omega, \quad (2.48)$$

in cui si sono sfruttate in [1] le relazioni $|e^{j\omega}| = 1$ e $e^{j\omega} = \cos \omega + j \sin \omega$.

2.5.2 Processi AR

Si consideri il p.s. scalare $y(t)$, soluzione dell'equazione alle differenze:

$$y(t) + a_1y(t-1) + \dots + a_ny(t-n) = e(t). \quad (2.49)$$

Un p.s. generato in questo modo è detto processo *auto-regressivo* di ordine n (o *auto regressive*, AR(n)).

Ancora una volta, interpretando z^{-1} come l'operatore di ritardo unitario, è immediato ricavare la funzione di trasferimento del filtro che genera il p.s. $y(t)$ a partire dall'ingresso $e(t)$. Infatti, la (2.49) può essere riscritta come:

$$A(z)y(t) = e(t), \quad \text{con } A(z) = 1 + a_1z^{-1} + \dots + a_nz^{-n}.$$

Per cui, moltiplicando e dividendo il polinomio $A(z)$ per z^n , la funzione di trasferimento che lega $e(t)$ a $y(t)$ corrisponde a:

$$G(z) = \frac{1}{A(z)} = \frac{z^n}{z^n + a_1z^{n-1} + \dots + a_n}.$$

Se le radici del polinomio $A(z)$ hanno modulo strettamente minore di uno, il sistema rappresentato dalla funzione di trasferimento $G(z)$ è asintoticamente stabile. Per cui, in virtù del Teorema 4, $y(t)$ è un p.s. asintoticamente stazionario.⁴

Si consideri, ora, la risposta impulsiva $\{g_k\}_{k=0}^{\infty}$ del sistema LTI che genera il p.s. $y(t)$ a partire da $e(t)$. Come è noto, essa è l'antitrasformata Z (monolatera) della funzione di trasferimento $G(z)$ (vedi eq. (2.26)):

$$g_k = Z^{-1}\{G(z)\}.$$

L'uscita, quindi, varrà (vedi eq. (2.27)):

$$y(t) = \sum_{m=0}^{\infty} g_m e(t-m). \quad (2.50)$$

⁴Nella discussione che segue, considereremo il valore a regime del segnale $y(t)$, cioè a transitorio esaurito (v. Osservazione 2.1).

La precedente relazione mostra come un p.s. auto-regressivo possa sempre essere visto come un p.s. a media mobile con infiniti coefficienti.

Segue che:

$$\begin{aligned}\mathbf{E}[y(t)] &= \left(\sum_{k=0}^{\infty} g_k \right) m_e, \\ \mathbf{E}[y(t)^2] &= \left(\sum_{k=0}^{\infty} g_k^2 \right) \sigma_e^2, \\ \mathbf{E}[y(t+\tau)y(t)] &= \left(\sum_{k=0}^{\infty} g_{k+\tau}g_k \right) \sigma_e^2,\end{aligned}$$

in cui, oltre alla linearità dell'operatore di valore atteso, si è sfruttata la proprietà di bianchezza del segnale $e(t)$ (vedi eq. (2.38)). Si noti che le sommatorie precedenti convergono se e solo se la $G(z)$ è asintoticamente stabile, cioè se e solo se il polinomio $A(z)$ possiede tutte le radici all'interno del cerchio unitario.

Esempio 2.5. Si consideri il p.s. AR(1):

$$y(t) - \frac{1}{2}y(t-1) = e(t)$$

con $e(t)$ rumore bianco, a media nulla e varianza $\sigma_e^2 = 1$.

Si è già calcolato (vedi Esempio 2.1 di pag. 51) che la funzione di covarianza di un generico processo AR(1):

$$y(t) = ay(t-1) + e(t)$$

vale:

$$R_{\mathbf{y}}(\tau) = \frac{\sigma_e^2}{1-a^2} a^{|\tau|}.$$

Per cui, nell'esempio in esame, $a = 0.5$ e:

$$R_{\mathbf{y}}(\tau) = \frac{4}{3} 0.5^{|\tau|}.$$

Inoltre, nell'Esempio 2.3, abbiamo calcolato lo spettro e la densità spettrale per un generico p.s. AR(1). In questo caso ($a = 0.5$ e $\sigma_e^2 = \frac{4}{3}$, vedi eq. (2.23)) si ha:

$$\phi_x(z) = \frac{1}{\left(1 - \frac{1}{2}z^{-1}\right) \left(1 - \frac{1}{2}z\right)}$$

e

$$\phi_x(e^{j\omega}) = \frac{1}{\frac{5}{4} - \cos \omega}.$$

△

Esaminando l'Esempio 2.1, si può notare come il calcolo dello spettro (o della densità spettrale) a partire dalla definizione possa essere alquanto laborioso, anche in un caso semplice come quello in esame. Per questo può essere conveniente utilizzare il Teorema 4. Infatti, le relazioni (2.30)-(2.31) consentono agevolmente di ricavare $\phi_y(z)$ e $\phi_y(e^{j\omega})$, senza dover necessariamente passare attraverso il calcolo della funzione di covarianza o il calcolo delle somme infinite che appaiono nelle definizioni di spettro e densità.

Esiste un metodo algoritmico, che va sotto il nome di *Equazioni di Yule-Walker*, che consente di calcolare in maniera ricorsiva il valore della funzione di covarianza di un p.s. AR di ordine arbitrario. Consideriamo un p.s. AR(n) scalare; per $\tau \geq 0$, si ottiene dalla (2.49):

$$\mathbf{E} [y(t - \tau)(y(t) + a_1 y(t - 1) + \cdots + a_n y(t - n))] = \mathbf{E} [y(t - \tau)e(t)]. \quad (2.51)$$

Consideriamo il primo membro della precedente uguaglianza. Svolgendo il prodotto e ricordando che $R_{\mathbf{y}}(\tau) = R_{\mathbf{y}}(-\tau)$, è possibile riscriverlo come:

$$R_{\mathbf{y}}(\tau) + a_1 R_{\mathbf{y}}(\tau - 1) + \cdots + a_n R_{\mathbf{y}}(\tau - n)$$

Si noti che il p.s. y ad un generico istante t dipende da tutti e soli i valori assunti dal rumore bianco e fino all'istante t , ma non dai valori futuri di e (eq. (2.50)). Inoltre, poichè $e(t)$ è bianco, vale la proprietà (2.38). Quindi, ricordando che si sta considerando $\tau \geq 0$, il secondo membro della (2.51), vale:

$$\mathbf{E} [y(t - \tau)e(t)] = \begin{cases} \sigma_e^2 & \text{se } \tau = 0 \\ 0 & \text{se } \tau > 0 \end{cases}$$

Per cui, la (2.51) diventa:

$$R_{\mathbf{y}}(\tau) + a_1 R_{\mathbf{y}}(\tau - 1) + \cdots + a_n R_{\mathbf{y}}(\tau - n) = \begin{cases} \sigma_e^2 & \text{se } \tau = 0 \\ 0 & \text{se } \tau > 0 \end{cases} \quad (2.52)$$

Riscriviamo la relazione precedente (ancora, utilizzando la proprietà $R_{\mathbf{y}}(\tau) = R_{\mathbf{y}}(-\tau)$) per $\tau = 0, 1, \dots, n$:

$$\begin{aligned} R_{\mathbf{y}}(0) + a_1 R_{\mathbf{y}}(1) + \dots + a_n R_{\mathbf{y}}(n) &= \sigma_e^2 \\ R_{\mathbf{y}}(1) + a_1 R_{\mathbf{y}}(0) + \dots + a_n R_{\mathbf{y}}(n-1) &= 0 \\ &\vdots \\ R_{\mathbf{y}}(n) + a_1 R_{\mathbf{y}}(n-1) + \dots + a_n R_{\mathbf{y}}(0) &= 0 \end{aligned}$$

o, in forma matriciale,

$$\begin{pmatrix} 1 & a_1 & a_2 & \dots & a_{n-1} & a_n \\ a_1 & 1+a_2 & a_3 & \dots & a_n & 0 \\ a_2 & a_1+a_3 & 1+a_4 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n-1} & a_{n-2}+a_n & a_{n-3} & \dots & 1 & 0 \\ a_n & a_{n-1} & a_{n-2} & \dots & a_1 & 1 \end{pmatrix} \begin{pmatrix} R_{\mathbf{y}}(0) \\ R_{\mathbf{y}}(1) \\ R_{\mathbf{y}}(2) \\ \vdots \\ R_{\mathbf{y}}(n-1) \\ R_{\mathbf{y}}(n) \end{pmatrix} = \begin{pmatrix} \sigma_e^2 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$$

Risolvendo il precedente sistema algebrico, si ricavano $R_{\mathbf{y}}(0), R_{\mathbf{y}}(1), \dots, R_{\mathbf{y}}(n)$ e, in seguito, dalla (2.52) si può calcolare:

$$R_{\mathbf{y}}(\tau) = -a_1 R_{\mathbf{y}}(\tau-1) - \dots - a_n R_{\mathbf{y}}(\tau-n)$$

per $\tau > n$. Un notevole vantaggio che offrono le equazioni di Yule-Walker ai fini del calcolo della funzione di covarianza, è rappresentato dalla possibilità di implementare tale metodo mediante un semplice programma di calcolo.

2.5.3 Processi ARMA

Si consideri il p.s. scalare $y(t)$, soluzione dell'equazione alle differenze:

$$y(t) + a_1 y(t-1) + \dots + a_n y(t-n) = c_0 e(t) + c_1 e(t-1) + \dots + c_m e(t-m). \quad (2.53)$$

Un p.s. generato in questo modo è detto processo *auto-regressivo a media mobile* di ordine (n, m) (o *auto regressive moving average*, ARMA(n, m)), poiché l'uscita contiene una componente auto-regressiva ed una a media mobile.

In termini di funzioni di trasferimento, la (2.53) può essere riscritta come:

$$A(z)y(t) = C(z)e(t),$$

con:

$$A(z) = 1 + a_1z^{-1} + \dots + a_nz^{-n} \quad \text{e} \quad C(z) = c_0 + c_1z^{-1} + \dots + c_mz^{-m}.$$

Ancora una volta, dividendo i polinomi $A(z)$ e $C(z)$ per z^n , la funzione di trasferimento che lega $e(t)$ a $y(t)$ corrisponde a:

$$G(z) = \frac{C(z)}{A(z)} = \frac{c_0z^n + c_1z^{n-1} + \dots + c_mz^{n-m}}{z^n + a_1z^{n-1} + \dots + a_n}.$$

Se le radici del polinomio $A(z)$ hanno modulo strettamente minore di uno, il sistema rappresentato dalla funzione di trasferimento $G(z)$ è asintoticamente stabile. Per cui, in virtù del Teorema 4, $y(t)$ è asintoticamente stazionario.

Esempio 2.6. Si consideri il p.s. ARMA(1,1):

$$y(t) - 0.9y(t-1) = e(t) + 0.5e(t-1), \quad (2.54)$$

con $e(t)$ rumore bianco, a media nulla e varianza $\sigma_e^2 = 1$.

Poiché l'unica radice di $A(z) = 1 - 0.9z^{-1}$ è $p_1 = 0.9$ e $|p_1| < 1$, il p.s. $y(t)$ è asintoticamente stazionario. Inoltre, dal Teorema 4:

$$m_y = G(1)m_e = 0.$$

Calcoliamo la funzione di covarianza $R_y(\tau)$.

$$\begin{aligned} R_y(0) &= \mathbf{E} [y(t)^2] \stackrel{[1]}{=} \mathbf{E} [y(t) (0.9y(t-1) + e(t) + 0.5e(t-1))] \\ &= 0.9R_y(1) + R_{ye}(0) + 0.5R_{ye}(1) \end{aligned} \quad (2.55)$$

in cui la [1] discende dalla (2.54). Analogamente:

$$\begin{aligned} R_y(1) &= \mathbf{E} [y(t)y(t-1)] = \mathbf{E} [(0.9y(t-1) + e(t) + 0.5e(t-1))y(t-1)] \\ &= 0.9R_y(0) + R_{ye}(-1) + 0.5R_{ye}(0). \end{aligned}$$

Poiché il p.s. y all'istante t dipende solo dai valori assunti da e fino all'istante t e non da quelli futuri (dall'eq. (2.54) e poiché e è bianco) si ha:

$$R_{ye}(\tau) = 0, \quad \text{se } \tau < 0.$$

Per cui:

$$R_{\mathbf{y}}(1) = 0.9R_{\mathbf{y}}(0) + 0.5R_{ye}(0) \quad (2.56)$$

Occorre calcolare $R_{ye}(0)$ e $R_{ye}(1)$:

$$\begin{aligned} R_{ye}(0) &= \mathbf{E}[y(t)e(t)] = \mathbf{E}[(0.9y(t-1) + e(t) + 0.5e(t-1))e(t)] \\ &= \sigma_e^2 = 1, \end{aligned}$$

$$\begin{aligned} R_{ye}(1) &= \mathbf{E}[y(t)e(t-1)] = \mathbf{E}[(0.9y(t-1) + e(t) + 0.5e(t-1))e(t-1)] \\ &= 0.9R_{ye}(0) + 0.5\sigma_e^2 = 1.4, \end{aligned}$$

dal momento che $R_{ye}(-1) = 0$ e $R_e(1) = 0$ poiché $e(t)$ è bianco. Sostituendo le espressioni precedenti nelle (2.55)-(2.56), otteniamo il sistema algebrico:

$$\begin{cases} R_{\mathbf{y}}(0) - 0.9R_{\mathbf{y}}(1) = 1.7 \\ -0.9R_{\mathbf{y}}(0) + R_{\mathbf{y}}(1) = 0.5 \end{cases}$$

che risolto dà:

$$R_{\mathbf{y}}(0) = 11.32 \quad \text{e} \quad R_{\mathbf{y}}(1) = 10.68.$$

Infine, se $\tau \geq 2$:

$$\begin{aligned} R_{\mathbf{y}}(\tau) &= \mathbf{E}[y(t)y(t-\tau)] = \mathbf{E}[(0.9y(t-1) + e(t) + 0.5e(t-1))y(t-\tau)] \\ &= 0.9R_{\mathbf{y}}(\tau-1), \end{aligned}$$

poiché $R_{ye}(l) = 0$, se $l < 0$. Concludendo:

$$R_{\mathbf{y}}(\tau) = \begin{cases} 11.32 & \text{se } \tau = 0 \\ 10.68 \cdot 0.9^{|\tau|-1} & \text{se } |\tau| \geq 1 \end{cases}$$

Per calcolare lo spettro e la densità spettrale conviene, come già accennato in precedenza, utilizzare i risultati forniti dal Teorema 4. Per cui:

$$\phi_y(z) = G(z)G(z^{-1})\phi_e(z) = \frac{(1 + 0.5z^{-1})(1 + 0.5z)}{(1 - 0.9z^{-1})(1 - 0.9z)}$$

e

$$\phi_y(e^{j\omega}) = |G(e^{j\omega})|^2 \phi_e(e^{j\omega}) = \left| \frac{(1 + 0.5e^{j\omega})}{(1 - 0.9e^{j\omega})} \right|^2 = \frac{1.25 + \cos \omega}{1.81 - 1.8 \cos \omega}.$$

△

Analogamente a quanto visto nel caso AR, è possibile modificare leggermente le *Equazioni di Yule-Walker* per calcolare ricorsivamente il valore della funzione di covarianza di un generico processo ARMA. Consideriamo, quindi, un p.s. ARMA(n,m) scalare. Per $\tau \geq 0$, dalla (2.53) si ha:

$$\begin{aligned} \mathbf{E} [y(t - \tau)(y(t) + a_1y(t - 1) + \dots + a_ny(t - n))] = \\ \mathbf{E} [y(t - \tau)(c_0e(t) + c_1e(t - 1) + \dots + c_me(t - m))]. \end{aligned} \quad (2.57)$$

Svolgendo i prodotti, si ottiene:

$$R_{\mathbf{y}}(\tau) + a_1R_{\mathbf{y}}(\tau - 1) + \dots + a_nR_{\mathbf{y}}(\tau - n) = c_0R_{\mathbf{ey}}(\tau) + c_1R_{\mathbf{ey}}(\tau - 1) + \dots + c_mR_{\mathbf{ey}}(\tau - m). \quad (2.58)$$

In questo caso, così come è stato fatto nell'Esempio 2.6, occorre procedere prima al calcolo della funzione di covarianza incrociata $R_{\mathbf{ey}}(\tau)$. Dalla (2.53) si ha:

$$\begin{aligned} \mathbf{E} [e(t - \tau)(y(t) + a_1y(t - 1) + \dots + a_ny(t - n))] = \\ \mathbf{E} [e(t - \tau)(c_0e(t) + c_1e(t - 1) + \dots + c_me(t - m))]. \end{aligned} \quad (2.59)$$

Svolgendo i prodotti, si ottiene:

$$R_{\mathbf{ey}}(-\tau) + a_1R_{\mathbf{ey}}(-\tau + 1) + \dots + a_nR_{\mathbf{ey}}(-\tau + n) = \sigma_e^2(\delta_\tau c_0 + \delta_{\tau-1}c_1 + \dots + \delta_{\tau-m}c_m), \quad (2.60)$$

dove δ_τ rappresenta la delta di Kronecker (v. eq. (2.16)).

Ora, osservando che $R_{\mathbf{ey}}(\tau) = 0$ se $\tau > 0$ (vedi considerazioni nell'Esempio 2.6), l'equazione (2.60) valutata, nell'ordine, per $\tau = 0, 1, \dots, m$ consente di calcolare $R_{\mathbf{ey}}(\tau)$ per $\tau = 0, -1, \dots, -m$. A questo punto, tramite la (2.58), è possibile calcolare la $R_{\mathbf{y}}(\tau)$.

2.6 Processi ergodici

Supponiamo di osservare una singola realizzazione, $\bar{x}(t)$, di un processo stocastico stazionario, $x(t)$. Quali informazioni circa le proprietà statistiche di $x(t)$ (ad es. media e funzione di covarianza) è possibile trarre?

Un p.s. stazionario si dice *ergodico* se le sue statistiche possono essere determinate a partire da una singola realizzazione, o, equivalentemente, se le “medie temporali” (calcolate su una singola realizzazione) coincidono con le “medie d’insieme” (i valori attesi).

Per giungere ad una definizione rigorosa di ergodicità, occorre innanzitutto specificare cosa si intende per medie temporali. Limitandoci ai momenti del primo e secondo ordine, definiamo, a partire da una singola realizzazione $\bar{x}(t)$, la *media campionaria*:

$$\bar{m}_{\mathbf{x}}^N = \frac{1}{N} \sum_{t=1}^N \bar{x}(t) \quad (2.61)$$

e la *covarianza campionaria*:

$$\bar{R}_x^N(\tau) = \frac{1}{N-\tau} \sum_{t=1}^{N-\tau} (\bar{x}(t+\tau) - \bar{m}_{\mathbf{x}})(\bar{x}(t) - \bar{m}_{\mathbf{x}})^T \quad (2.62)$$

Definizione 2.13. Un p.s. stazionario $x(t)$ si dice *ergodico* se:

$$\lim_{N \rightarrow \infty} \bar{m}_{\mathbf{x}}^N = \mathbf{E}[x(t)] = m_{\mathbf{x}} \quad (2.63)$$

$$\lim_{N \rightarrow \infty} \bar{R}_x^N(\tau) = \mathbf{E}[(x(t+\tau) - m_{\mathbf{x}})(x(t) - m_{\mathbf{x}})^T] = R_{\mathbf{x}}(\tau) \quad (2.64)$$

2.7 Esercizi

2.1. Si consideri il p.s. $z(t)$ ottenuto dalla somma di due p.s. $x(t)$, $y(t)$:

$$z(t) = x(t) + y(t),$$

con

$$\begin{aligned}\mathbf{E}[x(t)] &= m_{\mathbf{x}}; \\ \mathbf{E}[y(t)] &= m_{\mathbf{y}}; \\ \mathbf{E}[(x(t) - m_{\mathbf{x}})(x(t) - m_{\mathbf{x}})^T] &= R_{\mathbf{x}}(\tau); \\ \mathbf{E}[(y(t) - m_{\mathbf{y}})(y(t) - m_{\mathbf{y}})^T] &= R_{\mathbf{y}}(\tau); \\ \mathbf{E}[(x(t) - m_{\mathbf{x}})(y(t) - m_{\mathbf{y}})^T] &= R_{\mathbf{xy}}(\tau).\end{aligned}$$

Si verifichi che:

- $R_{\mathbf{z}}(\tau) = R_{\mathbf{x}}(\tau) + R_{\mathbf{y}}(\tau) + R_{\mathbf{xy}}(\tau) + R_{\mathbf{yx}}(\tau);$
- $\phi_{\mathbf{z}}(z) = \phi_{\mathbf{x}}(z) + \phi_{\mathbf{y}}(z) + \phi_{\mathbf{xy}}(z) + \phi_{\mathbf{yx}}(z).$

Cosa accade se $x(t)$ e $y(t)$ sono incorrelati?

2.2. Dire quali tra le seguenti funzioni $R_i(\tau)$, $i = 1, \dots, 3$ può essere interpretata come la funzione di covarianza di un processo stocastico stazionario, giustificando la risposta. Dire di che tipo di processo potrebbe trattarsi e quanto vale la sua varianza.

$$\begin{aligned}R_1(0) &= -1 & R_2(0) &= 1 & R_3(0) &= 1 \\ R_1(\pm 1) &= 0.5 & R_2(\pm 1) &= 0.5 & R_3(1) &= 0.5 \\ R_1(\tau) &= 0 \quad |\tau| > 1 & R_2(\tau) &= 0, \quad |\tau| > 1 & R_3(-1) &= 0.7 \\ & & & & R_3(\tau) &= 0, \quad |\tau| > 1\end{aligned}$$

2.3. Si consideri il processo MA(2):

$$y(t) = e(t) + 2e(t-1) + e(t-2), \quad e \sim WN(0, 1).$$

Dopo aver calcolato il valore atteso $E[y(t)]$ e la funzione di covarianza $R_y(t_1, t_2)$, verificare che il processo $y(t)$ è stazionario (in senso debole).

Calcolare, inoltre, la funzione di trasferimento da $e(t)$ a $y(t)$, la densità spettrale $\varphi_y(e^{j\omega})$ e rappresentare graficamente l'andamento di quest'ultima.

2.4. Si consideri lo schema a blocchi rappresentato in Fig. 2.10, dove:

$$G(z) = \frac{z}{z - \frac{1}{4}}$$

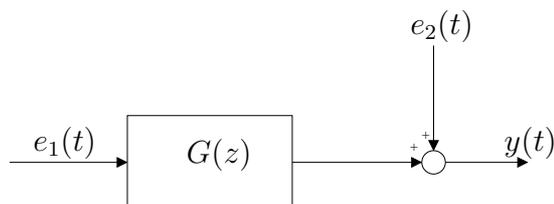


Figura 2.10.

ed e_1 è un rumore bianco di varianza unitaria, $e_1 \sim WN(0, 1)$.

Calcolare il valore atteso e la densità spettrale del processo $y(\cdot)$ nei seguenti casi:

- $e_2(\cdot) = 0, \forall t$;
- $e_2(\cdot) \sim WN(0, 1)$ indipendente da $e_1(\cdot)$;
- $e_2(y) = e_1(t), \forall t$.

2.5. Si consideri il seguente processo stocastico tempo discreto:

$$y(t) = 0.3y(t-1) + e(t) + 0.6e(t-1)$$

dove $e(t)$ è un rumore bianco con media nulla e varianza $\sigma_e^2 = 2$.

- Calcolare la funzione di trasferimento $G(z)$ da $e(t)$ a $y(t)$.
- Spiegare perché $y(t)$ è un processo stocastico asintoticamente stazionario.
- Calcolare il valore atteso $E[y(t)]$ e la funzione di covarianza $R_y(\tau)$ del processo $y(t)$ a regime.
- Calcolare lo spettro $\Phi_y(z)$ e la densità spettrale $\Phi_y(e^{j\omega})$ del processo $y(t)$.

2.6. Si consideri un processo stocastico stazionario $y(t)$ con funzione di covarianza $R_y(\tau)$. La funzione di covarianza normalizzata ($r_y(\tau) = R_y(\tau)/R_y(0)$) è rappresentata in Fig. 2.11.

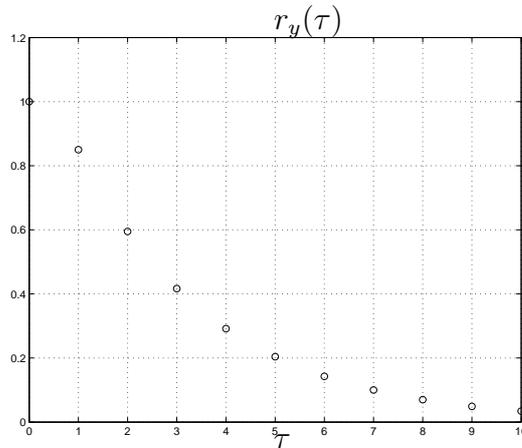


Figura 2.11.

1. Stabilire quale, fra i seguenti modelli, ha generato il processo $y(\cdot)$:

a) $y(t) = e(t) + 0.5e(t-1) + 1.8e(t-2)$

b) $y(t) = 0.7y(t-1) + e(t) + e(t-1)$

c) $y(t) = -0.7y(t-1) + e(t) + e(t-1)$

dove $e \sim WN(0, 1)$.

2. Rappresentare graficamente l'andamento della densità spettrale $\varphi_y(e^{j\omega})$

2.7. Si consideri il processo stocastico tempo discreto $y(t)$ generato come rappresentato in Figura 2.12, dove l'ingresso $u(t)$ è un processo stocastico tale che $u(t) = e(t) + 2e(t-1)$ e $G(z) = \frac{1}{1 + 0.5z^{-1}}$.

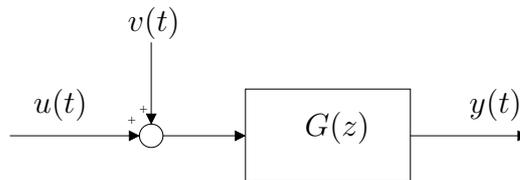


Figura 2.12.

Supponendo che $e(t)$ e $v(t)$ siano due rumori bianchi gaussiani scorrelati a media nulla e varianza $\sigma_e^2 = 1$ e $\sigma_v^2 = 2$, rispettivamente, determinare la densità spettrale $\phi_y(e^{j\omega})$ del processo $y(t)$ e tracciarne il grafico.

Capitolo 3

Teoria della stima

In questo capitolo vengono presentati alcuni elementi di base della *teoria della stima*. In un problema di stima, l'obiettivo è quello di determinare il valore di una grandezza incognita, sulla base di informazioni relative ad altre grandezze ad essa legate.

In funzione del tipo di informazione a priori disponibile sulla grandezza incognita da stimare, si distinguono due possibili scenari:

- Stima parametrica;
- Stima Bayesiana.

Alla stima parametrica sono dedicati i paragrafi 3.1-3.4, mentre il paragrafo 3.5 è dedicato all'approccio Bayesiano.

3.1 Stima parametrica

In un contesto di stima parametrica, ci si propone di stimare il valore di una grandezza deterministica θ sulla base di un certo numero di osservazioni, $\mathbf{y}_1, \dots, \mathbf{y}_n$.

3.1.1 Formulazione del problema.

Siano:

- $\theta \in \Theta \subseteq \mathbb{R}^p$, un vettore incognito di *parametri*;
- $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)^T \in \mathcal{Y} \subseteq \mathbb{R}^n$ un vettore aleatorio di *osservazioni* (o *misure*);
- $F_{\mathbf{y}}^{\theta}(y), f_{\mathbf{y}}^{\theta}(y)$ le funzioni di distribuzione e densità di probabilità del vettore delle osservazioni \mathbf{y} , dipendenti dal vettore incognito θ .

L'insieme Θ , in cui assume valori il vettore dei parametri θ , prende il nome di *spazio dei parametri*. Esso rappresenta tutta l'informazione a priori disponibile sui possibili valori del vettore θ . Nel caso limite in cui nulla sia noto circa i valori ammissibili dei parametri, $\Theta = \mathbb{R}^p$.

L'insieme \mathcal{Y} , in cui assume valori il vettore delle osservazioni \mathbf{y} , prende il nome di *spazio delle osservazioni*. Si assume che la distribuzione $F_{\mathbf{y}}^{\theta}(y)$ (o, equivalentemente, la densità $f_{\mathbf{y}}^{\theta}(y)$) delle osservazioni abbia *forma funzionale nota*, ma dipenda da un certo numero di parametri incogniti (il vettore θ).

Terminologia. D'ora in avanti, con il termine *parametro* si indicherà la grandezza incognita da stimare θ , anche nel caso in cui essa sia una grandezza vettoriale.

Un problema di stima parametrica può essere formulato nel seguente modo.

“Stimare il parametro incognito $\theta \in \Theta$, sulla base di un’osservazione y della variabile aleatoria $\mathbf{y} \in \mathcal{Y}$.”

Definizione 3.1. Uno *stimatore* T del parametro incognito θ è una funzione che mappa lo spazio delle osservazioni nello spazio dei parametri:

$$T : \mathcal{Y} \rightarrow \Theta.$$

Il valore $\hat{\theta} = T(y)$, assunto dallo stimatore in corrispondenza dell'osservazione $\mathbf{y} = y$, prende il nome di *stima* di θ .

Uno stimatore $T(\cdot)$, quindi, definisce una regola in base alla quale, se si osserva la realizzazione y per il vettore delle misure \mathbf{y} , si stima θ con la quantità $\hat{\theta} = T(y)$.

Si noti che $\hat{\theta}$ può essere vista come una realizzazione della variabile aleatoria $T(\mathbf{y})$; in altre parole, poiché $T(\mathbf{y})$ è una funzione della v.a. \mathbf{y} , la stima $\hat{\theta}$ è essa stessa una variabile aleatoria.

3.1.2 Proprietà di uno stimatore

Ovviamente, in base alla Definizione 3.1, la classe dei possibili stimatori è infinita. Per caratterizzare, quindi, la “bontà” di uno stimatore è utile introdurre alcune proprietà che lo caratterizzano.

Correttezza

Innanzitutto, sarebbe desiderabile che il valore $\hat{\theta} = T(\mathbf{y})$ fornito da uno stimatore $T(\cdot)$ coincidesse con il valore vero del parametro, θ . Dal momento che la stima $\hat{\theta}$ è una v.a., è ragionevole richiedere che questo accada in media.

Definizione 3.2. Uno stimatore $T(\mathbf{y})$ del parametro θ si dice *corretto* (o, *non polarizzato*) se

$$\mathbf{E}^\theta [T(\mathbf{y})] = \theta, \quad \forall \theta \in \Theta. \quad (3.1)$$

Nella precedente definizione si è utilizzata la notazione $\mathbf{E}^\theta [\cdot]$, in cui si evidenzia la dipendenza del valore atteso di $T(\mathbf{y})$ dal parametro incognito θ , dal momento che la d.d.p. di \mathbf{y} dipende, per ipotesi, da θ stesso.

La condizione (3.1) di non polarizzazione assicura che, in media, lo stimatore T non ha errori sistematici. Vale a dire, T non sovrastima né sottostima il parametro θ , in media (v. Figura 3.1).

Esempio 3.1. Siano $\mathbf{y}_1, \dots, \mathbf{y}_n$ variabili aleatorie identicamente distribuite (i.d.) con valor medio m . La quantità:

$$\bar{\mathbf{y}} = \frac{1}{n} \sum_{i=1}^n \mathbf{y}_i \quad (3.2)$$

prende il nome di *media campionaria*. È immediato verificare che $\bar{\mathbf{y}}$ è uno stimatore non polarizzato di m . Infatti, per la linearità dell'operatore di

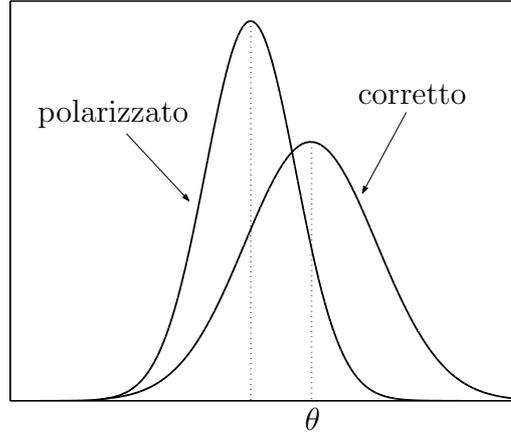


Figura 3.1: Densità di probabilità di uno stimatore corretto e di uno polarizzato.

valore atteso, si ha:

$$\mathbf{E}[\bar{\mathbf{y}}] = \mathbf{E}\left[\frac{1}{n} \sum_{i=1}^n \mathbf{y}_i\right] = \frac{1}{n} \sum_{i=1}^n \mathbf{E}[\mathbf{y}_i] = \frac{1}{n} \sum_{i=1}^n m = m.$$

△

Esempio 3.2. Siano $\mathbf{y}_1, \dots, \mathbf{y}_n$ variabili aleatorie scalari, indipendenti, identicamente distribuite (i.i.d.) con valor medio m e varianza σ^2 . La quantità:

$$\hat{\sigma}_{\mathbf{y}}^2 = \frac{1}{n} \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})^2$$

è uno stimatore polarizzato della varianza σ^2 . Infatti, dalla (3.2), si ha:

$$\begin{aligned} \mathbf{E}[\hat{\sigma}_{\mathbf{y}}^2] &= \frac{1}{n} \sum_{i=1}^n \mathbf{E}\left[\left(\mathbf{y}_i - \frac{1}{n} \sum_{j=1}^n \mathbf{y}_j\right)^2\right] \\ &= \frac{1}{n} \sum_{i=1}^n \frac{1}{n^2} \mathbf{E}\left[\left(n\mathbf{y}_i - \sum_{j=1}^n \mathbf{y}_j\right)^2\right] \\ &= \frac{1}{n} \sum_{i=1}^n \frac{1}{n^2} \mathbf{E}\left[\left(n(\mathbf{y}_i - m) - \sum_{j=1}^n (\mathbf{y}_j - m)\right)^2\right]. \end{aligned}$$

Ma:

$$\begin{aligned} \mathbf{E} \left[\left(n(\mathbf{y}_i - m) - \sum_{j=1}^n (\mathbf{y}_j - m) \right)^2 \right] &= n^2 \mathbf{E} [(\mathbf{y}_i - m)^2] \\ &\quad - 2n \mathbf{E} \left[(\mathbf{y}_i - m) \sum_{j=1}^n (\mathbf{y}_j - m) \right] + \mathbf{E} \left[\left(\sum_{j=1}^n (\mathbf{y}_j - m) \right)^2 \right] \\ &= n^2 \sigma^2 - 2n \sigma^2 + n \sigma^2 \\ &= n(n-1) \sigma^2 \end{aligned}$$

poiché, per l'ipotesi di indipendenza, $\mathbf{E} [(\mathbf{y}_i - m)(\mathbf{y}_j - m)] = 0$ se $i \neq j$.

Quindi:

$$\mathbf{E} [\hat{\sigma}_{\mathbf{y}}^2] = \frac{1}{n} \sum_{i=1}^n \frac{1}{n^2} n(n-1) \sigma^2 = \frac{n-1}{n} \sigma^2 \neq \sigma^2.$$

△

Esempio 3.3. Siano $\mathbf{y}_1, \dots, \mathbf{y}_n$ variabili aleatorie scalari, indipendenti, identicamente distribuite (i.i.d.) con valor medio m e varianza σ^2 . La quantità:

$$S^2 = \frac{1}{(n-1)} \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})^2$$

prende il nome di *varianza campionaria*. È immediato verificare che S^2 è uno stimatore non polarizzato della varianza σ^2 . Infatti, notando che:

$$S^2 = \frac{n}{n-1} \hat{\sigma}_{\mathbf{y}}^2,$$

segue immediatamente che:

$$\mathbf{E} [S^2] = \frac{n}{n-1} \mathbf{E} [\hat{\sigma}_{\mathbf{y}}^2] = \frac{n}{n-1} \frac{n-1}{n} \sigma^2 = \sigma^2.$$

△

Si noti che, se $T(\cdot)$ è uno stimatore non polarizzato di θ , allora $g(T(\cdot))$ non è uno stimatore non polarizzato di $g(\theta)$, a meno che $g(\cdot)$ non sia una funzione lineare.

Consistenza

Un'altra proprietà desiderabile di uno stimatore è che questo fornisca una stima che “si avvicina sempre più” al vettore vero di parametri θ , al crescere del numero di misure.

Definizione 3.3. Sia $\{\mathbf{y}_i\}_{i=1}^{\infty}$ una successione di v.a. La successione di stimatori $T_n = T(\mathbf{y}_1, \dots, \mathbf{y}_n)$ si dice *consistente* per il parametro θ se T_n converge a θ in probabilità, per ogni θ . Cioè, se:

$$\lim_{n \rightarrow \infty} \mathbf{P}(\|T_n - \theta\| \geq \varepsilon) = 0, \quad \forall \varepsilon > 0, \forall \theta \in \Theta.$$

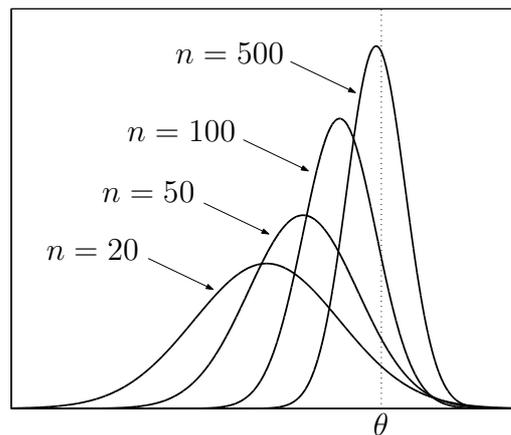


Figura 3.2: Densità di probabilità di uno stimatore consistente.

Si noti che la consistenza è una proprietà asintotica di uno stimatore. Essa garantisce che, al tendere all'infinito del numero di misure, la probabilità che la stima fornita sia diversa dal valore vero tende a zero (v. Figura 3.2). Il seguente Teorema fornisce una condizione sufficiente per la consistenza di una sequenza di stimatori.

Teorema 5. Sia T_n una sequenza di stimatori corretti del parametro scalare θ :

$$\mathbf{E}[T_n] = \theta, \quad \forall n, \quad \forall \theta \in \Theta.$$

Se:

$$\lim_{n \rightarrow \infty} \mathbf{E} [(T_n - \theta)^2] = 0,$$

allora la sequenza T_n è consistente.

Dimostrazione

Per una generica v.a. \mathbf{x} , vale la *diseguaglianza di Chebishev*

$$P(|\mathbf{x} - m_{\mathbf{x}}| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} \mathbf{E} [(\mathbf{x} - m_{\mathbf{x}})^2].$$

Perciò si ha

$$\lim_{n \rightarrow \infty} P(\|T_n - \theta\| \geq \varepsilon) \leq \lim_{n \rightarrow \infty} \frac{1}{\varepsilon^2} \mathbf{E} [(T_n - \theta)^2],$$

da cui il risultato segue immediatamente. \square

Quindi, affinché una sequenza di stimatori non polarizzati sia consistente è sufficiente che la varianza delle stime tenda a zero, al crescere del numero di misure.

Esempio 3.4. Siano $\mathbf{y}_1, \dots, \mathbf{y}_n$ variabili aleatorie indipendenti e identicamente distribuite (i.i.d.) con valor medio m e varianza σ^2 . Nell'Esempio 3.1 si è visto come la media campionaria:

$$\bar{\mathbf{y}} = \frac{1}{n} \sum_{i=1}^n \mathbf{y}_i$$

sia uno stimatore non polarizzato della media m . Mostriamo, ora, che essa è anche uno stimatore consistente per m . Infatti, la varianza della stima vale:

$$\begin{aligned} \text{Var}(\bar{\mathbf{y}}) &= \mathbf{E} [(\bar{\mathbf{y}} - m)^2] = \mathbf{E} \left[\left(\frac{1}{n} \sum_{i=1}^n \mathbf{y}_i - m \right)^2 \right] \\ &= \frac{1}{n^2} \mathbf{E} \left[\left(\sum_{i=1}^n (\mathbf{y}_i - m) \right)^2 \right] = \frac{\sigma^2}{n} \end{aligned}$$

dal momento che le v.a. \mathbf{y}_i sono indipendenti. Quindi:

$$\text{Var}(\bar{\mathbf{y}}) = \frac{\sigma^2}{n} \rightarrow 0 \quad \text{per } n \rightarrow \infty$$

e, in virtù del Teorema 5, la media campionaria $\bar{\mathbf{y}}$ risulta essere uno stimatore consistente per il valor medio m . \triangle

Errore quadratico medio

Una misura della bontà della stima fornita da uno stimatore è *l'errore quadratico medio*. Per semplificare la trattazione, nel seguito sarà considerato il caso di un parametro scalare ($\theta \in \mathbb{R}$). L'estensione al caso di un vettore di parametri è comunque immediata.

Definizione 3.4. Si definisce *errore quadratico medio* la quantità:

$$\mathbf{E}^\theta [(T(\mathbf{y}) - \theta)^2]$$

Si osservi che, se uno stimatore è corretto, allora l'errore quadratico medio coincide con la varianza della stima.

Definizione 3.5. Siano $T_1(\cdot)$ e $T_2(\cdot)$ due stimatori del parametro θ . Allora, $T_1(\cdot)$ si dice *preferibile* a $T_2(\cdot)$ se:

$$\mathbf{E}^\theta [(T_1(\mathbf{y}) - \theta)^2] \leq \mathbf{E}^\theta [(T_2(\mathbf{y}) - \theta)^2], \quad \forall \theta \in \Theta$$

La precedente definizione, in maniera del tutto naturale, afferma che fra due stimatori è preferibile utilizzare quello con errore quadratico medio minore.

3.1.3 Stima non polarizzata a minima varianza

Nel seguito, restringendo la nostra attenzione alla classe degli stimatori corretti, andremo alla ricerca di quello, se esiste, con minima varianza.

Definizione 3.6. Uno stimatore corretto $T^*(\cdot)$ del parametro scalare θ è detto *a minima varianza* (o UMVUE, Uniformly Minimum Variance Unbiased Estimator) se:

$$\mathbf{E}^\theta [(T^*(\mathbf{y}) - \theta)^2] \leq \mathbf{E}^\theta [(T(\mathbf{y}) - \theta)^2], \quad \forall \theta \in \Theta \quad (3.3)$$

per ogni stimatore corretto $T(\cdot)$ di θ .

Si noti che, affinché uno stimatore sia UMVUE, occorre che:

- sia non polarizzato (*Unbiased*);
- abbia la varianza minore, fra tutti gli stimatori non polarizzati (*Minimum Variance*);
- la precedente condizione valga per qualunque valore del parametro θ (*Uniformly*).

Sfortunatamente, esistono numerosi problemi per cui non esiste lo stimatore *UMVUE*. Per tale motivo, spesso si ricorre a classi più ristrette di stimatori. Si consideri la classe degli stimatori *lineari*, ossia del tipo:

$$T(\mathbf{y}) = \sum_{i=1}^n a_i \mathbf{y}_i, \quad (3.4)$$

con $a_i \in \mathbb{R}$.

Definizione 3.7. Uno stimatore lineare corretto $T^*(\cdot)$ del parametro scalare θ viene detto *BLUE* (Best Linear Unbiased Estimator) se:

$$\mathbf{E}^\theta [(T^*(\mathbf{y}) - \theta)^2] \leq \mathbf{E}^\theta [(T(\mathbf{y}) - \theta)^2], \quad \forall \theta \in \Theta$$

per ogni stimatore lineare corretto $T(\cdot)$ di θ .

A differenza dello stimatore *UMVUE*, lo stimatore *BLUE* ha una forma semplice e fissata (si tratta solo di determinare i coefficienti ottimi a_i), per cui il suo calcolo risulta molto più agevole.

Esempio 3.5. Siano \mathbf{y}_i variabili aleatorie indipendenti, con media m e varianza σ_i^2 , $i = 1, \dots, n$. Si desidera calcolare lo stimatore *BLUE* di m . Dal momento che lo stimatore deve essere lineare esso avrà la forma descritta nell'eq. (3.4). Affinché $T(\cdot)$ sia non polarizzato, occorre che sia:

$$\mathbf{E}^\theta [T(\mathbf{y})] = \mathbf{E}^\theta \left[\sum_{i=1}^n a_i \mathbf{y}_i \right] = \sum_{i=1}^n a_i \mathbf{E}^\theta [\mathbf{y}_i] = m \sum_{i=1}^n a_i = m$$

Per cui, occorre che sia:

$$\sum_{i=1}^n a_i = 1 \quad (3.5)$$

Ora, tra tutti gli stimatori della forma (3.4), con i coefficienti a_i che verificano la (3.5), occorre determinare quello a minima varianza. Dal momento che le osservazioni \mathbf{y}_i sono indipendenti, la varianza di $T(\mathbf{y})$ vale:

$$\mathbf{E}^\theta [(T(\mathbf{y}) - m)^2] = \mathbf{E}^\theta \left[\left(\sum_{i=1}^n a_i \mathbf{y}_i - m \right)^2 \right] = \sum_{i=1}^n a_i^2 \sigma_i^2.$$

Per cui, il calcolo dello stimatore BLUE si riduce a risolvere il seguente problema di ottimizzazione vincolata:

$$\begin{aligned} \min_{a_i} \quad & \sum_{i=1}^n a_i^2 \sigma_i^2 \\ \text{s.t.} \quad & \\ & \sum_{i=1}^n a_i = 1 \end{aligned}$$

Si costruisca la Lagrangiana:

$$\mathcal{L}(a_1, \dots, a_n, \lambda) = \sum_{i=1}^n a_i^2 \sigma_i^2 + \lambda \left(\sum_{i=1}^n a_i - 1 \right)$$

e si imponga che:

$$\frac{\partial \mathcal{L}(a_1, \dots, a_n, \lambda)}{\partial a_i} = 0, \quad i = 1, \dots, n \quad (3.6)$$

$$\frac{\partial \mathcal{L}(a_1, \dots, a_n, \lambda)}{\partial \lambda} = 0. \quad (3.7)$$

Dalla (3.7) si riottiene esattamente il vincolo (3.5), mentre la (3.6) implica che:

$$2a_i \sigma_i^2 + \lambda = 0, \quad i = 1, \dots, n$$

da cui:

$$\lambda = -\frac{1}{\sum_{i=1}^n \frac{1}{2\sigma_i^2}} \quad (3.8)$$

$$a_i = \frac{\frac{1}{\sigma_i^2}}{\sum_{j=1}^n \frac{1}{\sigma_j^2}}, \quad i = 1, \dots, n \quad (3.9)$$

Quindi, lo stimatore BLUE del valor medio m è dato da:

$$\hat{m}_{BLUE} = \frac{1}{\sum_{i=1}^n \frac{1}{\sigma_i^2}} \sum_{i=1}^n \frac{1}{\sigma_i^2} \mathbf{y}_i \quad (3.10)$$

Si noti che, nel caso in cui tutte le misure abbiano la stessa varianza $\sigma_i^2 = \sigma^2$, \hat{m}_{BLUE} si riduce alla media campionaria $\bar{\mathbf{y}}$. Questa osservazione consente una immediata interpretazione dello stimatore *BLUE* appena calcolato. Esso è una generalizzazione della media campionaria, nel caso in cui le misure \mathbf{y}_i abbiano diversa accuratezza (ossia, diversa varianza σ_i^2). In questo caso, lo stimatore BLUE non è altro che una media delle osservazioni, pesata in funzione della diversa precisione di ciascuna di esse. Infatti, si supponga che, per un certo i , $\sigma_i^2 \rightarrow \infty$. Questo corrisponde ad avere una misura \mathbf{y}_i totalmente inaffidabile. Allora, il peso $\frac{1}{\sigma_i^2}$ che la \mathbf{y}_i avrà nella stima \hat{m}_{BLUE} tende a zero. D'altro canto, si supponga di avere una misura \mathbf{y}_i infinitamente precisa ($\sigma_i^2 \rightarrow 0$). In questo caso, il peso $\frac{1}{\sigma_i^2}$ dell'osservazione corrispondente è predominante rispetto a tutti gli altri, e lo stimatore BLUE diventa $\hat{m}_{BLUE} \simeq \mathbf{y}_i$, cioè circa uguale alla misura infinitamente accurata. \triangle

3.2 Limite di Cramér-Rao

In questo paragrafo verrà presentato un importante risultato, che stabilisce un limite inferiore per la varianza di ogni stimatore corretto del parametro θ .

Teorema 6. *Sia $T(\cdot)$ uno stimatore non polarizzato del parametro scalare θ basato sulle osservazioni \mathbf{y} delle v.a. $\mathbf{y} \in \mathcal{Y}$, e si supponga che lo spazio delle osservazioni \mathcal{Y} sia indipendente da θ . Allora, sotto alcune ipotesi di regolarità (v. (Rohatgi and Saleh, 2001)), vale:*

$$\mathbf{E}^\theta [(T(\mathbf{y}) - \theta)^2] \geq [I_n(\theta)]^{-1}, \quad (3.11)$$

dove:

$$I_n(\theta) = \mathbf{E}^\theta \left[\left(\frac{\partial \ln f_{\mathbf{y}}^\theta(y)}{\partial \theta} \right)^2 \right] \quad (3.12)$$

è detta quantità di informazione di Fisher. Inoltre, se le v.a. osservate $\mathbf{y}_1, \dots, \mathbf{y}_n$ sono indipendenti e identicamente distribuite, con densità di probabilità $f_{\mathbf{y}_1}^\theta(y_1)$, si ha

$$I_n(\theta) = n I_1(\theta).$$

Nel caso in cui θ sia una grandezza vettoriale, la (3.11) diventa:

$$\mathbf{E}^\theta \left[(T(\mathbf{y}) - \theta) (T(\mathbf{y}) - \theta)^T \right] \geq [I_n(\theta)]^{-1},$$

dove la disuguaglianza va intesa in senso matriciale e la $I_n(\theta)$ denota la matrice di informazione di Fisher:

$$I_n(\theta) = \mathbf{E}^\theta \left[\left(\frac{\partial \ln f_{\mathbf{y}}^\theta(y)}{\partial \theta} \right) \left(\frac{\partial \ln f_{\mathbf{y}}^\theta(y)}{\partial \theta} \right)^T \right].$$

La matrice:

$$\mathbf{E}^\theta \left[(T(\mathbf{y}) - \theta) (T(\mathbf{y}) - \theta)^T \right]$$

è la matrice di covarianza dello stimatore.

Il Teorema precedente garantisce che non esiste alcuno stimatore corretto con varianza minore di $I_n(\theta)$. Si noti che la quantità $I_n(\theta)$ dipende, in generale, dal valore del parametro incognito θ (dal momento che la derivata parziale va valutata proprio in corrispondenza del valore reale di θ), che è sconosciuto all'utente. Ciononostante, il risultato di Cramér-Rao consente di definire il concetto di *efficienza* di uno stimatore.

Definizione 3.8. Uno stimatore non polarizzato $T(\cdot)$ si dice *efficiente* se la sua varianza raggiunge il limite di Cramér-Rao:

$$\mathbf{E}^\theta \left[(T(\mathbf{y}) - \theta)^2 \right] = [I_n(\theta)]^{-1}.$$

Uno stimatore efficiente è caratterizzato dall'aver la più piccola varianza possibile (fra tutti gli stimatori non polarizzati).

Nel caso di osservazioni \mathbf{y}_i indipendenti e identicamente distribuite, è possibile verificare che:

$$I_n(\theta) = nI_1(\theta),$$

dove $I_1(\theta)$ indica la quantità di informazione di Fisher nel caso di una sola misura. Quindi, fissato θ , il limite di Cramér-Rao migliora come $\frac{1}{n}$, all'aumentare del numero n di osservazioni.

Esempio 3.6. Siano $\mathbf{y}_1, \dots, \mathbf{y}_n$ variabili aleatorie indipendenti e identicamente distribuite (i.i.d.) con valor medio $m_{\mathbf{y}}$ e varianza $\sigma_{\mathbf{y}}^2$. Si è già visto (v. Esempio 3.1) come la media campionaria:

$$\bar{\mathbf{y}} = \frac{1}{n} \sum_{i=1}^n \mathbf{y}_i$$

sia uno stimatore non polarizzato della media, nonchè consistente (v. Esempio 3.4). Dal Teorema 6, e poiché le osservazioni sono i.i.d., discende che:

$$\mathbf{E}^\theta [(\bar{\mathbf{y}} - m_{\mathbf{y}})^2] = \frac{\sigma_{\mathbf{y}}^2}{n} \geq [I_n(\theta)]^{-1} = \frac{[I_1(\theta)]^{-1}}{n}.$$

Supponendo che le \mathbf{y}_i abbiano una distribuzione gaussiana:

$$f_{\mathbf{y}_i}(y_i) = \frac{1}{\sqrt{2\pi}\sigma_{\mathbf{y}}} e^{-\frac{(y_i - m_{\mathbf{y}})^2}{2\sigma_{\mathbf{y}}^2}},$$

procediamo con il calcolo della quantità di informazione scalare:

$$I_1(\theta) = \mathbf{E}^\theta \left[\left(\frac{\partial \ln f_{\mathbf{y}_1}^\theta(y_1)}{\partial \theta} \right)^2 \right].$$

Nell'esempio in esame, il parametro incognito da stimare è il valor medio: $\theta = m$. Per cui:

$$\frac{\partial \ln f_{\mathbf{y}_1}^\theta(y_1)}{\partial \theta} = \frac{\partial}{\partial m} \left(\ln \frac{1}{\sqrt{2\pi}\sigma_{\mathbf{y}}} - \frac{(y_1 - m)^2}{2\sigma_{\mathbf{y}}^2} \right) \Big|_{m=m_{\mathbf{y}}} = \frac{y - m_{\mathbf{y}}}{\sigma_{\mathbf{y}}^2},$$

e di conseguenza:

$$I_1(\theta) = \mathbf{E}^\theta \left[\frac{(y - m_{\mathbf{y}})^2}{\sigma_{\mathbf{y}}^4} \right] = \frac{1}{\sigma_{\mathbf{y}}^2}.$$

Il limite di Cramér-Rao, in questo caso, assume il valore:

$$[I_n(\theta)]^{-1} = \frac{[I_1(\theta)]^{-1}}{n} = \frac{\sigma_{\mathbf{y}}^2}{n},$$

che coincide proprio con la varianza dello stimatore $\bar{\mathbf{y}}$. Per cui, *nel caso di osservazioni i.i.d. gaussiane, la media campionaria è uno stimatore efficiente.* \triangle

3.3 Stima a massima verosimiglianza

In generale, non sempre esiste uno stimatore efficiente per un dato parametro incognito. Nell'esempio precedente, si è mostrato come il limite di Cramér-Rao consenta di verificare se un dato stimatore è efficiente o meno. Resta il problema di trovare degli stimatori candidati opportuni e, in caso nessuno di essi sia efficiente, concludere con certezza che, per il problema in esame, non esistono stimatori efficienti.

Definizione 3.9. Sia \mathbf{y} un vettore di osservazioni, avente densità di probabilità $f_{\mathbf{y}}^{\theta}(y)$, dipendente da un parametro incognito $\theta \in \Theta$. Si definisce *verosimiglianza* (o *likelihood*) la funzione:

$$L(\theta|y) = f_{\mathbf{y}}^{\theta}(y).$$

Si noti che, una volta osservata per la v.a. \mathbf{y} la particolare realizzazione y (cioè, una volta raccolte le misure), la verosimiglianza è una funzione del solo parametro incognito θ .

Una stima ragionevole del parametro incognito θ è quella che massimizza la probabilità dell'evento osservato. Infatti, ricordando il significato di densità di probabilità, massimizzare la $f_{\mathbf{y}}^{\theta}(y)$ rispetto a θ significa far sì che l'intervallo infinitesimo $[y, y + \delta y]$, definito a partire dalla misura osservata y , abbia la massima probabilità fra tutti gli intervalli di ampiezza δy .

Definizione 3.10. Si definisce *stimatore di massima verosimiglianza* (o *maximum likelihood estimator*, MLE) del parametro incognito θ , lo stimatore:

$$T_{ML}(y) = \arg \max_{\theta \in \Theta} L(\theta|y).$$

Alcune volte, per agevolare il calcolo, può essere conveniente massimizzare, al posto della funzione di verosimiglianza, la cosiddetta *log-likelihood*:

$$\ln L(\theta|y).$$

Si noti che ciò è giustificato dal fatto che, essendo il logaritmo una funzione monotona crescente, $L(\theta|y)$ e $\ln L(\theta|y)$ hanno gli stessi punti di massimo.

Osservazione 3.1. Supponendo che la $f_{\mathbf{y}}^{\theta}(y)$ sia una funzione differenziabile di $\theta = (\theta_1, \dots, \theta_p) \in \Theta \subseteq \mathbb{R}^p$, con Θ insieme aperto, se $\hat{\theta}$ è un punto di massimo per $L(\theta|y)$, esso deve soddisfare le equazioni:

$$\left. \frac{\partial L(\theta|y)}{\partial \theta_i} \right|_{\theta=\hat{\theta}} = 0, \quad i = 1, \dots, p \quad (3.13)$$

o, equivalentemente,

$$\left. \frac{\partial \ln L(\theta|y)}{\partial \theta_i} \right|_{\theta=\hat{\theta}} = 0, \quad i = 1, \dots, p. \quad (3.14)$$

Nella pratica, ogni $\hat{\theta}$ soluzione delle equazioni (3.13) (o delle (3.14)) viene chiamata stima di massima verosimiglianza. È bene osservare che spesso, anche nel caso di un parametro scalare, possono sorgere alcune complicazioni, in quanto l'eq. (3.13) può ammettere più di una soluzione, oppure la funzione di verosimiglianza può non essere differenziabile in tutto Θ o, infine, il suo punto di massimo può coincidere con un punto sulla frontiera di Θ (questo nel caso in cui Θ non sia un insieme aperto).

Esempio 3.7. Siano $\mathbf{y}_1, \dots, \mathbf{y}_n$ variabili aleatorie gaussiane, indipendenti con valor medio $m_{\mathbf{y}}$ incognito e varianza $\sigma_{\mathbf{y}}^2$ nota. Si desidera calcolare la stima a massima verosimiglianza della media.

Dal momento che le misure sono indipendenti, la verosimiglianza vale:

$$L(\theta|y) = f_{\mathbf{y}}^{\theta}(y) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_{\mathbf{y}}}} e^{-\frac{(y_i - m)^2}{2\sigma_{\mathbf{y}}^2}}$$

In questo caso conviene massimizzare la log-likelihood, che ha la forma:

$$\begin{aligned}\ln L(\theta|y) &= \sum_{i=1}^n \left(\ln \frac{1}{\sqrt{2\pi}\sigma_{\mathbf{y}}} - \frac{(y_i - m)^2}{2\sigma_{\mathbf{y}}^2} \right) \\ &= n \ln \frac{1}{\sqrt{2\pi}\sigma_{\mathbf{y}}} - \sum_{i=1}^n \frac{(y_i - m)^2}{2\sigma_{\mathbf{y}}^2}.\end{aligned}$$

Imponendo la condizione (3.14), si ottiene:

$$\frac{\partial \ln L(\theta|y)}{\partial \theta} = \frac{\partial}{\partial m} \left(n \ln \frac{1}{\sqrt{2\pi}\sigma_{\mathbf{y}}} - \sum_{i=1}^n \frac{(y_i - m)^2}{2\sigma_{\mathbf{y}}^2} \right) \Big|_{m=\hat{m}_{ML}} = 0,$$

da cui:

$$\sum_{i=1}^n \frac{y_i - \hat{m}_{ML}}{\sigma_{\mathbf{y}}^2} = 0,$$

ossia:

$$\hat{m}_{ML} = \frac{1}{n} \sum_{i=1}^n \mathbf{y}_i.$$

Quindi, in questo caso, lo stimatore di massima verosimiglianza coincide con la media campionaria. Dal momento che le osservazioni sono i.i.d. gaussiane, questo stimatore è anche efficiente (vedi Esempio 3.6) \triangle

Il risultato dell'Esempio precedente non è affatto casuale. Il seguente teorema giustifica l'importanza rivestita dalla stima a massima verosimiglianza, nell'ambito della teoria della stima.

Teorema 7. *Sotto le ipotesi di validità del limite di Cramér-Rao, se esiste uno stimatore $T^*(\cdot)$ efficiente, allora esso coincide con lo stimatore di massima verosimiglianza.*

Quindi, se si desidera ricercare uno stimatore efficiente per un parametro incognito θ , l'unico candidato possibile è lo stimatore di massima verosimiglianza.

Esempio 3.8. Siano $\mathbf{y}_1, \dots, \mathbf{y}_n$ variabili aleatorie gaussiane, indipendenti con valor medio $m_{\mathbf{y}}$ e varianza $\sigma_{\mathbf{y}}^2$, entrambi incogniti. Si desidera calcolare la stima a massima verosimiglianza della media e della varianza.

Analogamente a quanto derivato nell'Esempio 3.7, la log-likelihood risulta essere:

$$\ln L(\theta|y) = n \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_{i=1}^n \frac{(y_i - m)^2}{2\sigma^2}.$$

Si osservi che, ora, il parametro incognito da stimare è il vettore $\theta = (m, \sigma^2)^T$, per cui la condizione (3.14) diventa:

$$\begin{aligned} \frac{\partial \ln L(\theta|y)}{\partial \theta_1} &= \frac{\partial}{\partial m} \left(n \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_{i=1}^n \frac{(y_i - m)^2}{2\sigma^2} \right) \Bigg|_{(m=\hat{m}_{ML}, \sigma^2=\hat{\sigma}_{ML}^2)} = 0, \\ \frac{\partial \ln L(\theta|y)}{\partial \theta_2} &= \frac{\partial}{\partial \sigma^2} \left(n \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_{i=1}^n \frac{(y_i - m)^2}{2\sigma^2} \right) \Bigg|_{(m=\hat{m}_{ML}, \sigma^2=\hat{\sigma}_{ML}^2)} = 0. \end{aligned}$$

Differenziando rispetto a m e a σ^2 si ottiene:

$$\begin{aligned} \sum_{i=1}^n \frac{y_i - \hat{m}_{ML}}{\hat{\sigma}_{ML}^2} &= 0 \\ -\frac{n}{2\hat{\sigma}_{ML}^2} + \frac{1}{2\hat{\sigma}_{ML}^4} \sum_{i=1}^n (y_i - \hat{m}_{ML})^2 &= 0, \end{aligned}$$

da cui:

$$\begin{aligned} \hat{m}_{ML} &= \frac{1}{n} \sum_{i=1}^n \mathbf{y}_i \\ \hat{\sigma}_{ML}^2 &= \frac{1}{n} \sum_{i=1}^n (\mathbf{y}_i - \hat{m}_{ML})^2. \end{aligned}$$

Sebbene $\mathbf{E}^\theta [\hat{m}_{ML}] = m_{\mathbf{y}}$ (vedi Esempio 3.1), risulta $\mathbf{E}^\theta [\hat{\sigma}_{ML}^2] = \frac{n-1}{n} \sigma_{\mathbf{y}}^2$ (vedi Esempio 3.2). Per cui, in questo caso, lo stimatore di massima verosimiglianza è polarizzato, e quindi non è efficiente. In virtù del Teorema 7, si può concludere che non esiste alcuno stimatore efficiente per il parametro $\theta = (m, \sigma^2)^T$. \triangle

L'esempio precedente mostra come lo stimatore di massima verosimiglianza possa anche essere polarizzato. Tuttavia, oltre alle motivazioni fornite dal Teorema 7, esistono anche altre ragioni che rendono tale stimatore uno dei più usati nella pratica. Infatti, esso esibisce un buon comportamento asintotico (cioè all'aumentare del numero di misure) essendo, sotto opportune ipotesi tecniche:

- asintoticamente corretto;
- consistente;
- asintoticamente efficiente;
- asintoticamente normale.

3.3.1 Rumore additivo

Un classe di problemi di stima molto comuni è quella in cui si richiede di stimare un parametro incognito θ sulla base di n misure $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)^T$ corrotte da rumore additivo. Formalmente, sia:

$$U(\cdot) : \Theta \subseteq \mathbb{R}^p \rightarrow \mathbb{R}^n$$

una funzione deterministica del parametro incognito θ . L'obiettivo è stimare θ a partire dalle osservazioni:

$$\mathbf{y} = U(\theta) + \boldsymbol{\varepsilon}$$

dove $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ rappresenta il rumore di misura, modellato come un vettore di v.a. (generalmente a media nulla), avente densità di probabilità $f_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon})$.

Sotto queste ipotesi, la verosimiglianza vale:

$$L(\theta|y) = f_{\mathbf{y}}^{\theta}(y) = f_{\boldsymbol{\varepsilon}}(y - U(\theta)).$$

Nel caso notevole in cui il rumore di misura abbia una distribuzione gaussiana:

$$f_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}) = \frac{1}{(2\pi)^{n/2}(\det \Sigma_{\boldsymbol{\varepsilon}})^{1/2}} e^{-\frac{1}{2}\boldsymbol{\varepsilon}^T \Sigma_{\boldsymbol{\varepsilon}}^{-1} \boldsymbol{\varepsilon}}$$

con matrice di covarianza Σ_ε nota, la log-likelihood assume la seguente forma:

$$\ln L(\theta|y) = K - \frac{1}{2}(y - U(\theta))^T \Sigma_\varepsilon^{-1} (y - U(\theta)),$$

in cui K denota una quantità che non dipende da θ . A questo punto il calcolo dello stimatore di massima verosimiglianza si riduce al seguente problema di ottimizzazione:

$$\begin{aligned} \hat{\theta}_{ML} &= \arg \max_{\theta} \ln L(\theta|y) \\ &= \arg \min_{\theta} (y - U(\theta))^T \Sigma_\varepsilon^{-1} (y - U(\theta)). \end{aligned} \quad (3.15)$$

Dal momento che $U(\cdot)$ è, in generale, una funzione non lineare di θ , la soluzione può essere calcolata tramite metodi numerici.

3.4 Problemi di stima lineare

Un caso particolarmente interessante si ha quando la funzione che lega il parametro incognito θ alle osservazioni è lineare. In questo caso, l'equazione delle misure assume la forma:

$$\mathbf{y} = U\theta + \varepsilon, \quad (3.16)$$

in cui U è una matrice $n \times p$. Nel seguito si assumerà sempre $\text{rank}(U) = p$, cioè che il numero di misure linearmente indipendenti non sia inferiore al numero di parametri da stimare (ovvero, che il problema di stima sia ben posto).

Definizione 3.11. Sia \mathbf{y} un vettore di v.a. legate al parametro incognito θ dalla relazione (3.16). Lo stimatore:

$$T_{LS}(\mathbf{y}) = (U^T U)^{-1} U^T \mathbf{y} \quad (3.17)$$

è detto *stimatore ai minimi quadrati* (o *Least Square Estimator*, LSE) del parametro θ .

La denominazione di tale stimatore è giustificata dal fatto che esso minimizza la somma degli scarti quadratici:

$$\hat{\theta}_{LS} = \arg \min_{\theta} \|y - U\theta\|^2.$$

Infatti:

$$\|y - U\theta\|^2 = (y - U\theta)^T (y - U\theta) = y^T y + \theta^T U^T U \theta - 2y^T U \theta.$$

Imponendo che la derivata rispetto a θ si annulli, si ha:

$$\left. \frac{\partial}{\partial \theta} \|y - U\theta\|^2 \right|_{\theta = \hat{\theta}_{LS}} = 2\hat{\theta}_{LS}^T U^T U - 2y^T U = 0,$$

poiché $\frac{\partial x^T A x}{\partial x} = 2x^T A$ e $\frac{\partial A x}{\partial x} = A$. Risolvendo rispetto a $\hat{\theta}_{LS}^T$ si ottiene:

$$\hat{\theta}_{LS}^T = y^T U (U^T U)^{-1}.$$

Trasponendo l'uguaglianza precedente e osservando che la matrice $(U^T U)$ è simmetrica, segue esattamente la (3.17).

Sotto l'ipotesi di rumore a media nulla, $\mathbf{E}[\boldsymbol{\varepsilon}] = 0$, lo stimatore ai minimi quadrati è corretto:

$$\begin{aligned} \mathbf{E}^\theta [\hat{\theta}_{LS}] &= \mathbf{E}^\theta [(U^T U)^{-1} U^T \mathbf{y}] = \mathbf{E}^\theta [(U^T U)^{-1} U^T (U\theta + \boldsymbol{\varepsilon})] \\ &= \mathbf{E}^\theta [\theta + (U^T U)^{-1} U^T \boldsymbol{\varepsilon}] = \theta. \end{aligned}$$

Si noti come la stima ai minimi quadrati non necessiti di alcun tipo di informazione riguardo alle caratteristiche del rumore $\boldsymbol{\varepsilon}$. In particolare, ciò implica che tutte le osservazioni \mathbf{y}_i contribuiscono in egual misura al valore finale della stima.

Definizione 3.12. Sia \mathbf{y} un vettore di v.a. aleatorie legate al parametro incognito θ dalla relazione (3.16). Sia $\Sigma_\boldsymbol{\varepsilon}$ la matrice di covarianza del vettore dei rumori $\boldsymbol{\varepsilon}$. Lo stimatore:

$$T_{GM}(\mathbf{y}) = (U^T \Sigma_\boldsymbol{\varepsilon}^{-1} U)^{-1} U^T \Sigma_\boldsymbol{\varepsilon}^{-1} \mathbf{y} \quad (3.18)$$

è detto *stimatore di Gauss-Markov* (o stimatore ai minimi quadrati pesati *Weighted Least Square Estimator*, WLSE) del parametro θ .

Analogamente a quanto mostrato per lo stimatore ai minimi quadrati, è facile verificare che esso minimizza la somma *pesata* degli scarti quadratici medi:

$$\hat{\theta}_{GM} = \arg \min_{\theta} (y - U\theta)^T \Sigma_{\varepsilon}^{-1} (y - U\theta).$$

Si noti come la stima di Gauss-Markov richieda la conoscenza della matrice di covarianza Σ_{ε} del rumore di misura. Utilizzando tale informazione, le osservazioni vengono pesate in maniera inversamente proporzionale alla loro incertezza.

Sotto l'ipotesi di rumore a media nulla, $\mathbf{E}[\varepsilon] = 0$, anche lo stimatore di Gauss-Markov è corretto:

$$\begin{aligned} \mathbf{E}^{\theta} [\hat{\theta}_{GM}] &= \mathbf{E}^{\theta} [(U^T \Sigma_{\varepsilon}^{-1} U)^{-1} U^T \Sigma_{\varepsilon}^{-1} \mathbf{y}] = \mathbf{E}^{\theta} [(U^T \Sigma_{\varepsilon}^{-1} U)^{-1} U^T \Sigma_{\varepsilon}^{-1} (U\theta + \varepsilon)] \\ &= \mathbf{E}^{\theta} [\theta + (U^T \Sigma_{\varepsilon}^{-1} U)^{-1} U^T \Sigma_{\varepsilon}^{-1} \varepsilon] = \theta. \end{aligned}$$

Nel caso in cui il vettore dei rumori ε abbia media non nulla ma nota, $m_{\varepsilon} = \mathbf{E}[\varepsilon]$, le espressioni degli stimatori ai minimi quadrati e di Gauss-Markov restano quelle delle equazioni (3.17) e (3.18), a patto di sostituire alle misure \mathbf{y} le misure “corrette” $\mathbf{y} - m_{\varepsilon}$. Infatti, se si definisce il nuovo vettore di variabili aleatorie $\tilde{\varepsilon} = \varepsilon - m_{\varepsilon}$, l'equazione (3.16) può essere riscritta come

$$\mathbf{y} - m_{\varepsilon} = U\theta + \tilde{\varepsilon},$$

ed essendo chiaramente $\mathbf{E}[\tilde{\varepsilon}] = 0$, tutta la trattazione precedente può essere ripetuta, sostituendo \mathbf{y} con $\mathbf{y} - m_{\varepsilon}$.

Una importante proprietà dello stimatore di Gauss-Markov è quella di essere lo stimatore a minima varianza tra tutti gli stimatori lineari corretti, cioè lo stimatore BLUE (vedi Definizione 3.7). Vale infatti il seguente risultato.

Teorema 8. *Sia \mathbf{y} un vettore di v.a. aleatorie legate al parametro incognito θ dalla relazione (3.16). Sia Σ_{ε} la matrice di covarianza del vettore dei rumori ε . Allora lo stimatore BLUE coincide con lo stimatore di Gauss-Markov*

(3.18). La corrispondente varianza dell'errore di stima è pari a

$$\mathbf{E} \left[(\hat{\theta}_{GM} - \theta)(\hat{\theta}_{GM} - \theta)^T \right] = (U^T \Sigma_{\boldsymbol{\varepsilon}}^{-1} U)^{-1}. \quad (3.19)$$

Nel caso in cui sia $\Sigma_{\boldsymbol{\varepsilon}} = \sigma_{\boldsymbol{\varepsilon}}^2 I_n$ (con I_n matrice identità di dimensione n), ovvero le variabili $\boldsymbol{\varepsilon}$ siano tra loro scorrelate e abbiano la stessa varianza $\sigma_{\boldsymbol{\varepsilon}}^2$, lo stimatore BLUE coincide con lo stimatore ai minimi quadrati (3.17).

Dimostrazione

Poiché si considera la classe degli stimatori lineari corretti, si ha $T(\mathbf{y}) = A\mathbf{y}$, e $\mathbf{E}[A\mathbf{y}] = A\mathbf{E}[\mathbf{y}] = AU\theta$, per cui è necessario imporre il vincolo $AU = I_p$ per garantire che lo stimatore sia non polarizzato.

Dovendo determinare lo stimatore a minima varianza¹, occorre minimizzare

$$\begin{aligned} \mathbf{E} [(A\mathbf{y} - \theta)(A\mathbf{y} - \theta)^T] &= \mathbf{E} [(AU\theta + A\boldsymbol{\varepsilon} - \theta)(\cdot)^T] \\ &= \mathbf{E} [A\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T A^T] \\ &= A\Sigma_{\boldsymbol{\varepsilon}} A^T \end{aligned}$$

dove nella seconda eguaglianza si è sfruttato il vincolo $AU = I_p$. Ne consegue che lo stimatore BLUE si ottiene risolvendo il seguente problema di ottimizzazione vincolata

$$\begin{aligned} A_{BLUE} &= \arg \min_A A\Sigma_{\boldsymbol{\varepsilon}} A^T \\ &\text{s.t.} \\ &AU = I_p \end{aligned} \quad (3.20)$$

e ponendo quindi $T(\mathbf{y}) = A_{BLUE} \mathbf{y}$.

Essendo il vincolo $AU = I_p$ lineare nella variabile matriciale A , è possibile parametrizzare tutte le matrici ammissibili A mediante la varietà lineare affine

$$A = (U^T \Sigma_{\boldsymbol{\varepsilon}}^{-1} U)^{-1} U^T \Sigma_{\boldsymbol{\varepsilon}}^{-1} + M \quad (3.21)$$

con $M \in \mathbb{R}^{p \times n}$ tale che $MU = 0$. Si può facilmente verificare che tutte le matrici A definite da (3.21) soddisfano il vincolo $AU = I_p$. Sarà quindi

¹Nel caso in cui θ sia un vettore, la varianza dell'errore di stima $\theta - \hat{\theta}$ è una matrice, per cui la minimizzazione va intesa nel senso matriciale (cioè $P > Q$ significa che $P - Q$ è definita positiva).

sufficiente trovare quella che minimizza la quantità $A\Sigma_\epsilon A^T$. Sostituendo l'espressione (3.21) si ha:

$$\begin{aligned} A\Sigma_\epsilon A^T &= (U^T \Sigma_\epsilon^{-1} U)^{-1} U^T \Sigma_\epsilon^{-1} \Sigma_\epsilon \Sigma_\epsilon^{-1} U (U^T \Sigma_\epsilon^{-1} U)^{-1} \\ &\quad + (U^T \Sigma_\epsilon^{-1} U)^{-1} U^T \Sigma_\epsilon^{-1} \Sigma_\epsilon M^T \\ &\quad + M \Sigma_\epsilon \Sigma_\epsilon^{-1} U (U^T \Sigma_\epsilon^{-1} U)^{-1} + M \Sigma_\epsilon M^T \\ &= (U^T \Sigma_\epsilon^{-1} U)^{-1} + M \Sigma_\epsilon M^T \\ &\geq (U^T \Sigma_\epsilon^{-1} U)^{-1} \end{aligned}$$

dove la seconda eguaglianza è dovuta a $MU = 0$, mentre la diseguaglianza finale sfrutta il fatto che Σ_ϵ è una matrice definita positiva e quindi $M\Sigma_\epsilon M^T$ è semidefinita positiva. Poiché l'espressione $(U^T \Sigma_\epsilon^{-1} U)^{-1}$ non dipende da M , si può concludere che la soluzione del problema (3.20) si ottiene ponendo $M = 0$ nella (3.21), cioè scegliendo $A_{BLUE} = (U^T \Sigma_\epsilon^{-1} U)^{-1} U^T \Sigma_\epsilon^{-1}$ e quindi lo stimatore BLUE coincide con quello di Gauss-Markov. L'espressione della varianza dell'errore di stima (3.19) si ottiene immediatamente dall'espressione di $A\Sigma_\epsilon A^T$ quando $M = 0$.

Infine, se $\Sigma_\epsilon = \sigma_\epsilon^2 I_n$ si ha immediatamente $A_{BLUE} = (U^T U)^{-1} U^T$ (qualunque sia il valore di σ^2) e quindi lo stimatore BLUE coincide con quello ai minimi quadrati. \square

Nella sezione 3.3.1 si è visto che, nel caso in cui il rumore di misura ϵ sia gaussiano, lo stimatore di massima verosimiglianza si calcola risolvendo il problema di ottimizzazione (3.15). Nel caso in cui le osservazioni siano una funzione lineare di θ , come nell'eq. (3.16), tale problema diventa:

$$\hat{\theta}_{ML} = \arg \min_{\theta} (y - U\theta)^T \Sigma_\epsilon^{-1} (y - U\theta). \quad (3.22)$$

Ma, come osservato in precedenza, questo è proprio lo stimatore di Gauss-Markov. Per cui, *nel caso di osservazioni lineari, corrotte da rumore additivo gaussiano, lo stimatore a massima verosimiglianza coincide con quello di Gauss-Markov.*

Se, inoltre, le misure sono indipendenti ed hanno la stessa varianza σ_ϵ^2 :

$$\epsilon \sim N(0, \sigma_\epsilon^2 I_n)$$

è facile verificare che lo stimatore di Gauss-Markov coincide con lo stimatore ai minimi quadrati. Per cui, *nel caso di osservazioni lineari, i.i.d. gaussiane, lo stimatore a massima verosimiglianza coincide con quello ai minimi quadrati.*

La tabella seguente riassume le varie proprietà degli stimatori di Gauss-Markov e ai minimi quadrati, a seconda delle ipotesi che si fanno sul rumore ε .

Ipotesi su ε	Proprietà stimatore GM	Proprietà stimatore LS
nessuna	$\arg \min_{\theta} (y - U\theta)^T \Sigma_{\varepsilon}^{-1} (y - U\theta)$ con Σ_{ε} matrice data	$\arg \min_{\theta} \ y - U\theta\ ^2$
$\mathbf{E}[\varepsilon]$ nota	corretto	corretto
$\mathbf{E}[\varepsilon] = m_{\varepsilon}$ $\mathbf{E}[(\varepsilon - m_{\varepsilon})(\varepsilon - m_{\varepsilon})^T] = \Sigma_{\varepsilon}$	stimatore BLUE	stimatore BLUE se $\Sigma_{\varepsilon} = \sigma_{\varepsilon}^2 I_n$
$\varepsilon \sim N(m_{\varepsilon}, \Sigma_{\varepsilon})$	stimatore MLE	stimatore MLE se $\Sigma_{\varepsilon} = \sigma_{\varepsilon}^2 I_n$

Tabella 3.1: Proprietà degli stimatori di Gauss-Markov e ai minimi quadrati.

Esempio 3.9. Su di un parametro incognito θ sono disponibili n misure

$$y_i = \theta + v_i, \quad i = 1, \dots, n$$

dove i v_i sono realizzazioni di n variabili aleatorie \mathbf{v}_i , tra loro indipendenti, a media nulla e di varianza pari a σ_i^2 .

È immediato verificare che le misure y_i risultano essere realizzazioni di variabili aleatorie \mathbf{y}_i , le quali risultano avere media θ e varianza σ_i^2 . Il problema della stima del parametro θ può pertanto essere ricondotto alla stima della media di n variabili aleatorie (si vedano in proposito gli Esempi 3.1 e 3.5 e l'Esercizio 3.1).

3.5 Stima Bayesiana

In un contesto di stima Bayesiana, la grandezza che si desidera stimare non è deterministica, ma può essere modellata come una variabile aleatoria. In particolare, si vuole stimare la v.a. incognita $\mathbf{x} \in \mathbb{R}^m$, sulla base di osservazioni della variabile aleatoria $\mathbf{y} \in \mathbb{R}^n$, conoscendo la densità di probabilità congiunta $f_{\mathbf{x},\mathbf{y}}(x, y)$.

Occorre, quindi, trovare uno stimatore $\hat{\mathbf{x}} = T(\mathbf{y})$, dove:

$$T(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^m$$

Analogamente a quanto fatto per la stima parametrica, per valutare la qualità della stima è necessario definire un opportuno criterio.

Definizione 3.13. Si definisce *funzionale di rischio di Bayes* la quantità:

$$J_r = \mathbf{E} [d(\mathbf{x}, T(\mathbf{y}))] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} d(x, T(y)) f_{\mathbf{x},\mathbf{y}}(x, y) dx dy.$$

La grandezza $d(\mathbf{x}, T(\mathbf{y}))$ denota la “distanza” tra la v.a. incognita \mathbf{x} e la sua stima $T(\mathbf{y})$. Dal momento che tale quantità è a sua volta una variabile aleatoria, ci si propone di trovare lo stimatore $T^*(\cdot)$ che minimizza $d(\mathbf{x}, T(\mathbf{y}))$ “in media”:

$$T^*(\cdot) = \arg \min_{T(\cdot)} J_r.$$

3.5.1 Stima a minimo errore quadratico medio

Una scelta naturale per misurare la distanza fra una v.a. incognita e la sua stima è *l'errore quadratico*:

$$d(\mathbf{x}, T(\mathbf{y})) = \|\mathbf{x} - T(\mathbf{y})\|^2.$$

Definizione 3.14. Si definisce *stimatore a minimo errore quadratico medio (MEQM)* lo stimatore $\hat{\mathbf{x}}_{MEQM} = T^*(\cdot)$, dove:

$$T^*(\cdot) = \arg \min_{T(\cdot)} \mathbf{E} [\|\mathbf{x} - T(\mathbf{y})\|^2].$$

Si noti che il valore atteso nella precedente definizione viene calcolato rispetto ad entrambi le v.a. \mathbf{x} e \mathbf{y} . Per cui è necessario conoscere la d.d.p. congiunta $f_{\mathbf{x},\mathbf{y}}(x, y)$.

Teorema 9. *Sia \mathbf{x} una v.a. incognita e sia \mathbf{y} un vettore di osservazioni. Allora, la stima a minimo errore quadratico medio $\hat{\mathbf{x}}_{MEQM}$ di \mathbf{x} corrisponde al valore atteso condizionato:*

$$\hat{\mathbf{x}}_{MEQM} = \mathbf{E}[\mathbf{x}|\mathbf{y}].$$

Il risultato precedente afferma che, qualunque sia il legame fra la v.a. incognita \mathbf{x} e le osservazioni \mathbf{y} , la stima che minimizza l'errore quadratico medio è data dal valore atteso di \mathbf{x} condizionato all'aver osservato la realizzazione \mathbf{y} per la v.a. \mathbf{y} :

$$\hat{\mathbf{x}}_{MEQM} = \int_{-\infty}^{+\infty} \mathbf{x} f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) d\mathbf{x}. \quad (3.23)$$

La stima MEQM presenta altre caratteristiche che la rendono attraente. In particolare, indicata con Q la matrice:

$$Q(\mathbf{x}, T(\mathbf{y})) = \mathbf{E}[(\mathbf{x} - T(\mathbf{y}))(\mathbf{x} - T(\mathbf{y}))^T],$$

si dimostra che:

- $\hat{\mathbf{x}}_{MEQM}$ è lo stimatore che minimizza (in senso matriciale) $Q(\mathbf{x}, T(\mathbf{y}))$, ovvero:

$$Q(\mathbf{x}, \hat{\mathbf{x}}_{MEQM}) \leq Q(\mathbf{x}, T(\mathbf{y})), \quad \forall T(\mathbf{y})$$

dove la disuguaglianza matriciale $A \leq B$ va intesa nel senso che la matrice $B - A$ è semidefinita positiva;

- $\hat{\mathbf{x}}_{MEQM}$ minimizza ogni funzione scalare monotona crescente di $Q(\mathbf{x}, T(\mathbf{y}))$, come ad esempio:
 - la traccia di Q , che corrisponde alla stima MEQM;
 - la traccia di WQ , per ogni matrice $W > 0$, che corrisponde alla stima a minimo errore quadratico medio *pesato*.

Nonostante l'apparente semplicità della stima MEQM, in pratica il suo calcolo può risultare notevolmente complesso, in quanto richiede la conoscenza della d.d.p. congiunta $f_{\mathbf{x},\mathbf{y}}(x, y)$ e il calcolo dell'integrale (3.23).

Esempio 3.10. Si considerino due v.a. \mathbf{x} e \mathbf{y} , la cui d.d.p. congiunta vale:

$$f_{\mathbf{x},\mathbf{y}}(x, y) = \begin{cases} -\frac{3}{2}x^2 + 2xy & \text{se } 0 \leq x \leq 1, \quad 1 \leq y \leq 2 \\ 0 & \text{altrimenti} \end{cases}$$

Si desidera determinare la stima a minimo errore quadratico medio di \mathbf{x} basata su un'osservazione di \mathbf{y} .

Dal Teorema 9, sappiamo che:

$$\hat{\mathbf{x}}_{MEQM} = \int_{-\infty}^{+\infty} x f_{\mathbf{x}|\mathbf{y}}(x|y) dx.$$

Occorre innanzitutto calcolare:

$$f_{\mathbf{x}|\mathbf{y}}(x|y) = \frac{f_{\mathbf{x},\mathbf{y}}(x, y)}{f_{\mathbf{y}}(y)}.$$

La densità marginale di \mathbf{y} può essere ricavata dalla d.d.p. congiunta:

$$\begin{aligned} f_{\mathbf{y}}(y) &= \int_0^1 -\frac{3}{2}x^2 + 2xy dx \\ &= -\frac{x^3}{2} + yx^2 \Big|_{x=0}^{x=1} = y - \frac{1}{2}. \end{aligned}$$

Quindi, la d.d.p. condizionata vale:

$$f_{\mathbf{x}|\mathbf{y}}(x|y) = \begin{cases} \frac{-\frac{3}{2}x^2 + 2xy}{y - \frac{1}{2}} & \text{se } 0 \leq x \leq 1, \quad 1 \leq y \leq 2 \\ 0 & \text{altrimenti} \end{cases}$$

È possibile, a questo punto, calcolare la stima a minimo errore quadratico medio:

$$\begin{aligned} \hat{\mathbf{x}}_{MEQM} &= \int_0^1 x \frac{-\frac{3}{2}x^2 + 2xy}{y - \frac{1}{2}} dx \\ &= \frac{1}{y - \frac{1}{2}} \left(-\frac{3}{8}x^4 + \frac{2}{3}x^3y \right) \Big|_{x=0}^{x=1} = \frac{\frac{2}{3}y - \frac{3}{8}}{y - \frac{1}{2}}. \end{aligned}$$

△

3.5.2 Stima lineare a minimo errore quadratico medio

Restringiamo la nostra attenzione alla classe di stimatori lineari:

$$T(\mathbf{y}) = A\mathbf{y} + b \quad (3.24)$$

in cui la matrice $A \in \mathbb{R}^{m \times n}$ e il vettore $b \in \mathbb{R}^m$ rappresentano i coefficienti dello stimatore da determinare.

Tra tutti gli stimatori della forma (3.24) si desidera calcolare quello che minimizza l'errore quadratico medio.

Definizione 3.15. Si definisce *stimatore lineare a minimo errore quadratico medio (LMEQM)* lo stimatore $\hat{\mathbf{x}}_{LMEQM} = A^*\mathbf{y} + b^*$, dove:

$$A^*, b^* = \arg \min_{A, b} \mathbf{E} [\|\mathbf{x} - A\mathbf{y} - b\|^2]. \quad (3.25)$$

Teorema 10. Sia \mathbf{x} una v.a. incognita e sia \mathbf{y} un vettore di osservazioni, tali che

$$\begin{aligned} \mathbf{E}[\mathbf{x}] &= m_{\mathbf{x}}, & \mathbf{E}[\mathbf{y}] &= m_{\mathbf{y}} \\ \mathbf{E} \left[\begin{pmatrix} \mathbf{x} - m_{\mathbf{x}} \\ \mathbf{y} - m_{\mathbf{y}} \end{pmatrix} \begin{pmatrix} \mathbf{x} - m_{\mathbf{x}} \\ \mathbf{y} - m_{\mathbf{y}} \end{pmatrix}^T \right] &= \begin{pmatrix} R_{\mathbf{x}} & R_{\mathbf{x}\mathbf{y}} \\ R_{\mathbf{x}\mathbf{y}}^T & R_{\mathbf{y}} \end{pmatrix} \end{aligned}$$

Allora, la stima lineare a minimo errore quadratico medio $\hat{\mathbf{x}}_{LMEQM}$ di \mathbf{x} vale:

$$\hat{\mathbf{x}}_{LMEQM} = m_{\mathbf{x}} + R_{\mathbf{x}\mathbf{y}}R_{\mathbf{y}}^{-1}(\mathbf{y} - m_{\mathbf{y}}),$$

ossia:

$$\begin{aligned} A^* &= R_{\mathbf{x}\mathbf{y}}R_{\mathbf{y}}^{-1}, \\ b^* &= m_{\mathbf{x}} - R_{\mathbf{x}\mathbf{y}}R_{\mathbf{y}}^{-1}m_{\mathbf{y}}. \end{aligned}$$

Dimostrazione

Innanzitutto si osservi che il costo da minimizzare equivale alla traccia della matrice $\mathbf{E} [(\mathbf{x} - A\mathbf{y} - b)(\mathbf{x} - A\mathbf{y} - b)^T]$:

$$\mathbf{E} [\|\mathbf{x} - A\mathbf{y} - b\|^2] = \text{tr} (\mathbf{E} [(\mathbf{x} - A\mathbf{y} - b)(\mathbf{x} - A\mathbf{y} - b)^T]).$$

Inoltre, poiché la traccia è una funzione monotona crescente, se si è in grado di determinare A^*, b^* tali che:

$$\mathbf{E} [(\mathbf{x} - A^*\mathbf{y} - b^*)(\mathbf{x} - A^*\mathbf{y} - b^*)^T] \leq \mathbf{E} [(\mathbf{x} - A\mathbf{y} - b)(\mathbf{x} - A\mathbf{y} - b)^T], \forall A, b \quad (3.26)$$

allora sicuramente A^*, b^* sono soluzione del problema (3.25). Per cui, indicando con $\tilde{\mathbf{x}} = \mathbf{x} - A\mathbf{y} - b$ l'errore di stima, si calcoli:

$$\begin{aligned} \mathbf{E} [\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T] &= \mathbf{E} [(\mathbf{x} - m_{\mathbf{x}} - A(\mathbf{y} - m_{\mathbf{y}}) + m_{\mathbf{x}} - Am_{\mathbf{y}} - b) \\ &\quad \times (\mathbf{x} - m_{\mathbf{x}} - A(\mathbf{y} - m_{\mathbf{y}}) + m_{\mathbf{x}} - Am_{\mathbf{y}} - b)^T] \\ &= R_{\mathbf{x}} + AR_{\mathbf{y}}A^T - R_{\mathbf{x}\mathbf{y}}A^T - AR_{\mathbf{y}\mathbf{x}} \\ &\quad + (m_{\mathbf{x}} - Am_{\mathbf{y}} - b)(m_{\mathbf{x}} - Am_{\mathbf{y}} - b)^T \\ &= R_{\mathbf{x}} + AR_{\mathbf{y}}A^T - R_{\mathbf{x}\mathbf{y}}A^T - AR_{\mathbf{x}\mathbf{y}}^T + R_{\mathbf{x}\mathbf{y}}R_{\mathbf{y}}^{-1}R_{\mathbf{x}\mathbf{y}}^T - R_{\mathbf{x}\mathbf{y}}R_{\mathbf{y}}^{-1}R_{\mathbf{x}\mathbf{y}}^T \\ &\quad + (m_{\mathbf{x}} - Am_{\mathbf{y}} - b)(m_{\mathbf{x}} - Am_{\mathbf{y}} - b)^T \\ &= R_{\mathbf{x}} - R_{\mathbf{x}\mathbf{y}}R_{\mathbf{y}}^{-1}R_{\mathbf{x}\mathbf{y}}^T + (R_{\mathbf{x}\mathbf{y}}R_{\mathbf{y}}^{-1} - A)R_{\mathbf{y}}(R_{\mathbf{x}\mathbf{y}}R_{\mathbf{y}}^{-1} - A)^T \\ &\quad + (m_{\mathbf{x}} - Am_{\mathbf{y}} - b)(m_{\mathbf{x}} - Am_{\mathbf{y}} - b)^T. \end{aligned} \quad (3.27)$$

Si osservi che tutti i termini della precedente espressioni sono delle matrici semidefinite positive. Per cui, prendendo A^*, b^* tali che le ultime due matrici siano identicamente uguali a zero, si ottiene sicuramente la soluzione del problema (3.26):

$$\begin{aligned} A^* &= R_{\mathbf{x}\mathbf{y}}R_{\mathbf{y}}^{-1}; \\ b^* &= m_{\mathbf{x}} - Am_{\mathbf{y}} = m_{\mathbf{x}} - R_{\mathbf{x}\mathbf{y}}R_{\mathbf{y}}^{-1}m_{\mathbf{y}}. \end{aligned}$$

□

La stima LMEQM è non polarizzata, in quanto il valore atteso dell'errore di stima è zero:

$$\begin{aligned} \mathbf{E} [\tilde{\mathbf{x}}] &= \mathbf{E} [\mathbf{x} - \hat{\mathbf{x}}_{LMEQM}] = m_{\mathbf{x}} - \mathbf{E} [m_{\mathbf{x}} + R_{\mathbf{x}\mathbf{y}}R_{\mathbf{y}}^{-1}(\mathbf{y} - m_{\mathbf{y}})] \\ &= m_{\mathbf{x}} - m_{\mathbf{x}} + R_{\mathbf{x}\mathbf{y}}R_{\mathbf{y}}^{-1}\mathbf{E} [\mathbf{y} - m_{\mathbf{y}}] = 0. \end{aligned}$$

Ponendo $A = A^*$ e $b = b^*$ nell'espressione (3.27), si può concludere che la varianza dell'errore di stima per lo stimatore LMEQM vale:

$$\mathbf{E} [\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T] = R_{\mathbf{x}} - R_{\mathbf{x}\mathbf{y}}R_{\mathbf{y}}^{-1}R_{\mathbf{x}\mathbf{y}}^T.$$

È bene notare che, interpretando $R_{\mathbf{x}}$ come l'incertezza *a priori* sulla v.a. \mathbf{x} , $R_{\mathbf{x}} - R_{\mathbf{x}\mathbf{y}}R_{\mathbf{y}}^{-1}R_{\mathbf{x}\mathbf{y}}^T$ rappresenta la nuova incertezza su \mathbf{x} dopo aver effettuato la misura \mathbf{y} . Dal momento che la matrice $R_{\mathbf{x}\mathbf{y}}R_{\mathbf{y}}^{-1}R_{\mathbf{x}\mathbf{y}}^T$ è sempre semidefinita positiva, l'effetto delle misure è quello di ridurre l'incertezza su \mathbf{x} . Inoltre, tale riduzione è tanto maggiore quanto più è "grande" $R_{\mathbf{x}\mathbf{y}}$, cioè quanto più la misura \mathbf{y} è legata alla v.a. incognita \mathbf{x} .

Si noti che, ai fini del calcolo della stima lineare a minimo errore quadratico medio, non è necessario conoscere la densità di probabilità congiunta $f_{\mathbf{x},\mathbf{y}}(x, y)$, ma solo le statistiche del primo e secondo ordine: $m_{\mathbf{x}}, m_{\mathbf{y}}, R_{\mathbf{x}}, R_{\mathbf{y}}, R_{\mathbf{x}\mathbf{y}}$.

Inoltre, l'errore di stima $\tilde{\mathbf{x}}$ è scorrelato dalle osservazioni \mathbf{y} :

$$\begin{aligned} \mathbf{E} [\tilde{\mathbf{x}}\mathbf{y}^T] &= \mathbf{E} [(\mathbf{x} - m_{\mathbf{x}} - R_{\mathbf{x}\mathbf{y}}R_{\mathbf{y}}^{-1}(\mathbf{y} - m_{\mathbf{y}}))\mathbf{y}^T] \\ &= R_{\mathbf{x}\mathbf{y}} - R_{\mathbf{x}\mathbf{y}}R_{\mathbf{y}}^{-1}R_{\mathbf{y}} = 0. \end{aligned} \quad (3.28)$$

Il precedente risultato è comunemente indicato come *principio di ortogonalità*. Inoltre è possibile dimostrare che se uno stimatore lineare rispetta il principio di ortogonalità allora esso è lo stimatore LMEQM.

Infine, in virtù del Teorema 3 del Capitolo 1, nel caso in cui le v.a. \mathbf{x}, \mathbf{y} siano congiuntamente gaussiane, con media e matrice di covarianza definite come nel Teorema 10, il valore atteso di \mathbf{x} condizionato a \mathbf{y} vale proprio:

$$\mathbf{E} [\mathbf{x}|\mathbf{y}] = m_{\mathbf{x}} + R_{\mathbf{x}\mathbf{y}}R_{\mathbf{y}}^{-1}(\mathbf{y} - m_{\mathbf{y}}).$$

Quindi si può concludere che, se le v.a. \mathbf{x}, \mathbf{y} sono congiuntamente gaussiane, la stima a minimo errore quadratico medio coincide con la stima lineare a minimo errore quadratico medio: $\hat{\mathbf{x}}_{\text{MEQM}} = \hat{\mathbf{x}}_{\text{LMEQM}}$. In altre parole, nel caso gaussiano la stima MEQM è una funzione lineare della v.a. osservata \mathbf{y} .

Esempio 3.11. Siano $\mathbf{y}_1, \mathbf{y}_2$ due osservazioni rumorose della v.a. scalare \mathbf{x} avente media $m_{\mathbf{x}}$ e varianza $\sigma_{\mathbf{x}}^2$:

$$\mathbf{y}_1 = \mathbf{x} + \boldsymbol{\varepsilon}_1$$

$$\mathbf{y}_2 = \mathbf{x} + \boldsymbol{\varepsilon}_2$$

Siano $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2$ due v.a. indipendenti, a media nulla e varianza σ_1^2, σ_2^2 , rispettivamente. Sotto l'ipotesi che \mathbf{x} e $\boldsymbol{\varepsilon}_i$ siano indipendenti, si desidera calcolare la stima lineare a minimo errore quadratico medio.

Definendo le v.a. vettoriali $\mathbf{y} = (\mathbf{y}_1 \ \mathbf{y}_2)^T$ e $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1 \ \boldsymbol{\varepsilon}_2)^T$, l'equazione delle misure può essere riscritta nella forma matriciale:

$$\mathbf{y} = U\mathbf{x} + \boldsymbol{\varepsilon},$$

dove $U = (1 \ 1)^T$.

Innanzitutto si calcoli il valor medio di \mathbf{y} :

$$\mathbf{E}[\mathbf{y}] = \mathbf{E}[U\mathbf{x} + \boldsymbol{\varepsilon}] = Um_{\mathbf{x}}$$

Per determinare la stima $\hat{\mathbf{x}}_{LMEQM}$ occorre calcolare le matrici di covarianza $R_{\mathbf{x}\mathbf{y}}, R_{\mathbf{y}}$:

$$R_{\mathbf{x}\mathbf{y}} = \mathbf{E} \left[(\mathbf{x} - m_{\mathbf{x}}) (U(\mathbf{x} - m_{\mathbf{x}}) + \boldsymbol{\varepsilon})^T \right] = \sigma_{\mathbf{x}}^2 U^T,$$

poiché \mathbf{x} e $\boldsymbol{\varepsilon}$ sono incorrelate.

$$\begin{aligned} R_{\mathbf{y}} &= \mathbf{E} \left[(U(\mathbf{x} - m_{\mathbf{x}}) + \boldsymbol{\varepsilon}) (U(\mathbf{x} - m_{\mathbf{x}}) + \boldsymbol{\varepsilon})^T \right] \\ &= U\sigma_{\mathbf{x}}^2 U^T + R_{\boldsymbol{\varepsilon}}, \end{aligned}$$

dove:

$$R_{\boldsymbol{\varepsilon}} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

denota la matrice di covarianza della v.a. $\boldsymbol{\varepsilon}$. Infine, si calcoli l'inversa della

matrice di covarianza delle misure:

$$\begin{aligned}
 R_{\mathbf{y}}^{-1} &= \left[\sigma_{\mathbf{x}}^2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \right]^{-1} \\
 &= \begin{pmatrix} \sigma_{\mathbf{x}}^2 + \sigma_1^2 & \sigma_{\mathbf{x}}^2 \\ \sigma_{\mathbf{x}}^2 & \sigma_{\mathbf{x}}^2 + \sigma_2^2 \end{pmatrix}^{-1} \\
 &= \frac{1}{\sigma_{\mathbf{x}}^2(\sigma_1^2 + \sigma_2^2) + \sigma_1^2\sigma_2^2} \begin{pmatrix} \sigma_{\mathbf{x}}^2 + \sigma_2^2 & -\sigma_{\mathbf{x}}^2 \\ -\sigma_{\mathbf{x}}^2 & \sigma_{\mathbf{x}}^2 + \sigma_1^2 \end{pmatrix}.
 \end{aligned}$$

Quindi, la stima lineare a minimo errore quadratico medio vale:

$$\begin{aligned}
 \hat{\mathbf{x}}_{LMEQM} &= m_{\mathbf{x}} + R_{\mathbf{x}\mathbf{y}}R_{\mathbf{y}}^{-1}(\mathbf{y} - Um_{\mathbf{x}}) \\
 &= m_{\mathbf{x}} + \sigma_{\mathbf{x}}^2 U^T R_{\mathbf{y}}^{-1}(\mathbf{y} - Um_{\mathbf{x}}) \\
 &= m_{\mathbf{x}} + \frac{\sigma_{\mathbf{x}}^2}{\sigma_{\mathbf{x}}^2(\sigma_1^2 + \sigma_2^2) + \sigma_1^2\sigma_2^2} (1 \ 1) \begin{pmatrix} \sigma_{\mathbf{x}}^2 + \sigma_2^2 & -\sigma_{\mathbf{x}}^2 \\ -\sigma_{\mathbf{x}}^2 & \sigma_{\mathbf{x}}^2 + \sigma_1^2 \end{pmatrix} \begin{pmatrix} \mathbf{y}_1 - m_{\mathbf{x}} \\ \mathbf{y}_2 - m_{\mathbf{x}} \end{pmatrix} \\
 &= m_{\mathbf{x}} + \frac{1}{\sigma_1^2 + \sigma_2^2 + \frac{\sigma_1^2\sigma_2^2}{\sigma_{\mathbf{x}}^2}} (\sigma_2^2 \ \sigma_1^2) \begin{pmatrix} \mathbf{y}_1 - m_{\mathbf{x}} \\ \mathbf{y}_2 - m_{\mathbf{x}} \end{pmatrix} \\
 &= m_{\mathbf{x}} + \frac{\sigma_2^2\mathbf{y}_1 + \sigma_1^2\mathbf{y}_2 - m_{\mathbf{x}}(\sigma_1^2 + \sigma_2^2)}{\sigma_1^2 + \sigma_2^2 + \frac{\sigma_1^2\sigma_2^2}{\sigma_{\mathbf{x}}^2}} \\
 &= \frac{\frac{m_{\mathbf{x}}\sigma_1^2\sigma_2^2}{\sigma_{\mathbf{x}}^2} + \sigma_2^2\mathbf{y}_1 + \sigma_1^2\mathbf{y}_2}{\sigma_1^2 + \sigma_2^2 + \frac{\sigma_1^2\sigma_2^2}{\sigma_{\mathbf{x}}^2}} = \frac{\frac{m_{\mathbf{x}}}{\sigma_{\mathbf{x}}^2} + \frac{1}{\sigma_1^2}\mathbf{y}_1 + \frac{1}{\sigma_2^2}\mathbf{y}_2}{\frac{1}{\sigma_{\mathbf{x}}^2} + \frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2\sigma_2^2}} \\
 &= \frac{\frac{m_{\mathbf{x}}}{\sigma_{\mathbf{x}}^2} + \frac{1}{\sigma_1^2}\mathbf{y}_1 + \frac{1}{\sigma_2^2}\mathbf{y}_2}{\frac{1}{\sigma_{\mathbf{x}}^2} + \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}.
 \end{aligned}$$

Si noti come ciascuna misura abbia un peso inversamente proporzionale alla varianza del rumore che la corrompe. Inoltre, l'informazione a priori disponibile sulla v.a. \mathbf{x} (vale a dire il valor medio $m_{\mathbf{x}}$ e la varianza $\sigma_{\mathbf{x}}^2$), viene utilizzata come una misura aggiuntiva. In particolare, è interessante notare che se $\sigma_{\mathbf{x}}^2 \rightarrow +\infty$ (ovvero l'informazione a priori su \mathbf{x} è del tutto inattendibile), la stima $\hat{\mathbf{x}}_{LMEQM}$ tende a coincidere con la stima di Gauss-Markov della media $m_{\mathbf{x}}$. △

3.6 Esercizi

3.1. Verificare che nel problema considerato nell'Esempio 3.9, gli stimatori ai minimi quadrati e di Gauss-Markov del parametro incognito θ coincidono rispettivamente con $\bar{\mathbf{y}}$ in (3.2) e $\hat{\mathbf{m}}_{BLUE}$ in (3.10).

3.2. Siano $\mathbf{d}_1, \mathbf{d}_2$ variabili aleatorie indipendenti identicamente distribuite, con densità di probabilità data da:

$$f(\delta) = \begin{cases} \theta e^{-\theta\delta} & \text{se } \delta \geq 0 \\ 0 & \text{se } \delta < 0 \end{cases}$$

Siano δ_1, δ_2 le osservazioni disponibili di $\mathbf{d}_1, \mathbf{d}_2$. Determinare la stima di massima verosimiglianza di θ .

3.3. Siano $\mathbf{d}_1, \mathbf{d}_2$ variabili aleatorie gaussiane indipendenti, tali che:

$$\mathbf{E}[\mathbf{d}_1] = m, \quad \mathbf{E}[\mathbf{d}_2] = 3m, \quad \mathbf{E}[(\mathbf{d}_1 - m)^2] = 2, \quad \mathbf{E}[(\mathbf{d}_2 - 3m)^2] = 4$$

Siano δ_1, δ_2 le osservazioni disponibili di $\mathbf{d}_1, \mathbf{d}_2$. Determinare:

- a. la stima lineare a minima varianza di m tra gli stimatori non polarizzati;
- b. la varianza di tale stima;
- c. la stima di massima verosimiglianza (c'è differenza con la stima al punto a?).

3.4. Sulla grandezza incognita x sono effettuate due misure:

$$\mathbf{y}_1 = x + \mathbf{d}_1$$

$$\mathbf{y}_2 = 2x + \mathbf{d}_2$$

dove \mathbf{d}_1 e \mathbf{d}_2 sono disturbi indipendenti rappresentati da variabili aleatorie con densità di probabilità

$$f(\delta) = \begin{cases} \lambda e^{-\lambda\delta} & \text{se } \delta \geq 0 \\ 0 & \text{se } \delta < 0 \end{cases}$$

- a. Determinare la stima di massima verosimiglianza di x .
- b. Determinare se la stima ottenuta è corretta.

3.5. Si consideri per $\theta \in [-2, 2]$ la funzione definita da

$$f^\theta(x) = \begin{cases} \theta x + 1 - \frac{\theta}{2} & \text{se } x \in [0, 1] \\ 0 & \text{altrimenti} \end{cases}$$

- a. Mostrare che per ogni $\theta \in [-2, 2]$, f^θ è una densità di probabilità.
- b. Sia \mathbf{y} una variabile aleatoria di densità f^θ . Calcolare in funzione di θ media e varianza di \mathbf{y} .
- c. Calcolare la stima di massima verosimiglianza di θ basata su una osservazione y della v.a. \mathbf{y} .
- d. Sia $\mathbf{y}_1, \dots, \mathbf{y}_n$ un campione di dimensione n della densità f^θ e si ponga

$$T(\mathbf{y}_1, \dots, \mathbf{y}_n) = 12 \left(\frac{1}{n} \sum_{k=1}^n \mathbf{y}_k - \frac{1}{2} \right)$$

Mostrare che T è uno stimatore non polarizzato di θ .

- e. Determinare la varianza dell'errore di stima dello stimatore $T(\cdot)$ definito al punto d., nel caso $n = 1$. Calcolare inoltre la quantità di informazione di Fisher $I_1(\theta)$ e mostrare che la disuguaglianza (3.11) è soddisfatta.

3.6. Siano a e b due grandezze incognite, relativamente alle quali sono disponibili tre diverse misure:

$$\begin{aligned} \mathbf{y}_1 &= a + \mathbf{v}_1 \\ \mathbf{y}_2 &= b + \mathbf{v}_2 \\ \mathbf{y}_3 &= a + b + \mathbf{v}_3 \end{aligned}$$

dove \mathbf{v}_i , $i = 1, 2, 3$, sono variabili aleatorie indipendenti, a media nulla.

Assumendo $\mathbf{E}[\mathbf{v}_1^2] = \mathbf{E}[\mathbf{v}_3^2] = 1$ e $\mathbf{E}[\mathbf{v}_2^2] = \frac{1}{2}$, determinare:

- a) la stima ai minimi quadrati di a e b ;
- b) la stima di Gauss-Markov di a e b ;
- c) la varianza degli errori di stima, ovvero $\mathbf{E} \left[(a - \hat{a})^2 + (b - \hat{b})^2 \right]$, per le stime calcolate ai punti a) e b).

Confrontare inoltre le stime ottenute con quelle che si avrebbero se la misura \mathbf{y}_3 non fosse disponibile. Come si modifica la varianza dell'errore di stima?

3.7. Si considerino due variabili aleatorie \mathbf{x} e \mathbf{y} , la cui d.d.p. congiunta è

$$f_{\mathbf{x},\mathbf{y}}(x, y) = \begin{cases} -\frac{3}{2}x^2 + 2xy & 0 \leq x \leq 1, 1 \leq y \leq 2 \\ 0 & \text{altrimenti} \end{cases}$$

Si vuole determinare la stima $\hat{\mathbf{x}}_{LMEQM}$ di \mathbf{x} , basata su una osservazione di \mathbf{y} .

Riportare in un grafico l'andamento della stima $\hat{\mathbf{x}}_{LMEQM}$ calcolata sopra e della stima $\hat{\mathbf{x}}_{MEQM}$ calcolata nell'Esempio 3.10, al variare di y (valore osservato della v.a. \mathbf{y}). Calcolare inoltre il valore atteso delle stime trovate e confrontarlo con il valor medio a priori $\mathbf{E}[\mathbf{x}]$.

3.8. Siano \mathbf{x} e \mathbf{y} due variabili aleatorie con densità di probabilità congiunta

$$f_{\mathbf{x},\mathbf{y}}(x, y) = \begin{cases} \frac{1}{12}(x + y)e^{-y} & 0 \leq x \leq 4, y \geq 0 \\ 0 & \text{altrimenti} \end{cases}$$

Si assuma di disporre di una osservazione y della v.a. \mathbf{y} .

- a. Determinare le stime $\hat{\mathbf{x}}_{MEQM}$ e $\hat{\mathbf{x}}_{LMEQM}$ di \mathbf{x} , e plottarle in funzione del valore osservato y della v.a. \mathbf{y} .
- b. Calcolare l'errore quadratico medio delle stime ottenute al punto a [Suggerimento: utilizzare MATLAB per il calcolo degli integrali].

3.9. Sia X una grandezza incognita di cui è disponibile una misura

$$\mathbf{y} = \ln \left(\frac{1}{X} \right) + \mathbf{v}$$

con \mathbf{v} v.a., la cui d.d.p. è pari a $f_{\mathbf{v}}(v) = \begin{cases} e^{-v} & v \geq 0 \\ 0 & v < 0 \end{cases}$.

- a) Determinare la stima a massima verosimiglianza di X . Valutare la correttezza dello stimatore ottenuto. È possibile determinare uno stimatore corretto di X ?
- b) Assumendo ora che X sia una v.a. indipendente da \mathbf{v} , con densità di probabilità

$$f_X(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{altrimenti} \end{cases},$$

determinare le stime MEQM e LMEQM di X .

- c) Riportare in un grafico l'andamento delle stime calcolate ai punti a) e b), al variare di y .

Capitolo 4

Filtraggio ottimo

In questo capitolo vengono introdotti i concetti fondamentali della teoria del filtraggio ottimo, con particolare attenzione rivolta al *filtraggio alla Wiener* che, nelle sue diverse forme, è ampiamente utilizzato in svariate applicazioni di controllo o signal-processing.

4.1 Formulazione del problema

Dati due processi stocastici $s(t)$ e $y(t)$, il problema del filtraggio consiste nel determinare una stima di $s(k+l)$ sulla base delle osservazioni $Y_{k-N}^k = \{y(k), y(k-1), \dots, y(k-N)\}$. In questo contesto, uno stimatore di $s(k+l)$ viene chiamato *filtro* e la stima viene comunemente indicata con $\hat{s}(k+l|Y_{k-N}^k)$.

Fra tutti i filtri possibili, ci si propone di trovare quello “ottimo”, rispetto ad un opportuno indice di costo legato all’errore di stima:

$$\tilde{s}(k+l|Y_{k-N}^k) = s(k+l) - \hat{s}(k+l|Y_{k-N}^k). \quad (4.1)$$

Lo schema di Figura 4.1 mostra come viene generato l’errore di stima; il blocco Σ rappresenta il filtro. Come si è visto nel capitolo precedente, una possibile scelta è quella di determinare il filtro in modo da minimizzare l’errore quadratico medio:

$$EQM = \mathbf{E} \left[\left\{ \tilde{s}(k+l|Y_{k-N}^k) \right\}^2 \right]. \quad (4.2)$$

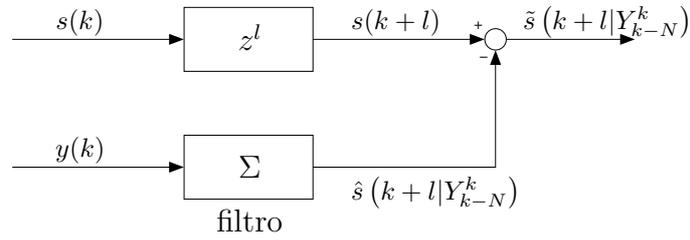


Figura 4.1: Problema del filtraggio lineare.

Dal momento che il calcolo dello stimatore a minimo errore quadratico medio (MEQM) risulta essere, in generale, troppo complesso, solitamente si restringe la classe dei filtri all'interno della quale ricercare quello ottimo. Imponendo che il filtro sia un sistema dinamico lineare, il problema di filtraggio ottimo diventa quello di trovare lo stimatore lineare che minimizza l'errore quadratico medio (LMEQM):

$$\min_{\Sigma \text{ lineare}} \mathbf{E} \left[\left\{ \tilde{s}(k+l|Y_{k-N}^k) \right\}^2 \right].$$

Esempio 4.1. Nel caso in cui il processo delle osservazioni $y(k)$ coincida con il processo $s(k)$ da stimare

$$y(k) = s(k)$$

ed $l > 0$, si parla di *predizione pura*. In questo caso si desidera stimare il valore che il processo osservato assumerà in un certo istante futuro. Applicazioni tipiche di tale problema riguardano la predizione di serie storiche in vari settori (economia, biologia, ecc.). \triangle

Esempio 4.2. Di frequente il processo $s(k)$ da stimare non è accessibile direttamente, ma ne è disponibile solo una misura rumorosa. Supponendo un modello di rumore additivo, le osservazioni assumono la forma

$$y(k) = s(k) + v(k),$$

in cui il rumore $v(k)$ si suppone scorrelato dal segnale di interesse $s(k)$. A seconda del valore assunto da l , si distinguono tre differenti problemi:

- $l > 0 \Rightarrow$ predizione;
- $l = 0 \Rightarrow$ filtraggio;
- $l < 0 \Rightarrow$ smoothing.

In generale, si parla di predizione quando si desidera stimare il valore di $s(k)$ in istanti successivi rispetto agli istanti delle osservazioni disponibili. Se per stimare $s(k)$ si utilizzano anche le osservazioni all'istante k , allora si parla di filtraggio vero e proprio. Infine, nel caso in cui si utilizzino anche osservazioni successive all'istante di interesse, si parla di smoothing. \triangle

Esempio 4.3. Si consideri il sistema LTI in forma di stato

$$\begin{aligned}x(k+1) &= Ax(k) + Gw(k) \\ y(k) &= Cx(k) + v(k)\end{aligned}$$

dove $w(k)$ e $v(k)$ sono processi stocastici di cui sono note alcune caratteristiche (ad es., media e covarianza). Un tipico problema di filtraggio consiste nello stimare lo stato $x(k)$ a partire dalle uscite $y(k)$. \triangle

Esempio 4.4. Quando si desidera stimare il valore dell'ingresso $u(k)$ di un sistema LTI, individuato dalla funzione di trasferimento $G(z)$, a partire dalle uscite rumorose

$$y(k) = G(z)u(k) + v(k)$$

si parla di problema di deconvoluzione ottima. Un tipico esempio è la ricostruzione di un segnale $u(k)$ trasmesso attraverso un canale rumoroso modellato tramite la f.d.t. $G(z)$, a partire dal segnale $y(k)$ ricevuto. \triangle

A seconda delle ipotesi sui processi stocastici $s(k)$ ed $y(k)$, si ottengono diverse soluzioni al problema del filtraggio ottimo. Di seguito, viene illustrata la teoria del filtraggio alla Wiener, che parte dall'assunzione di stazionarietà

per i processi stocastici $s(k)$ ed $y(k)$. Come vedremo, in questo contesto si utilizza una descrizione di tipo ingresso-uscita mediante funzioni di trasferimento. In primo luogo, verrà esaminato il caso di filtri a risposta finita (FIR) per i quali la soluzione del problema di filtraggio viene calcolata in termini di risposta impulsiva del filtro. Successivamente, verrà esaminato il caso più generale di filtri lineari tempo-invarianti (LTI), in cui il filtro ottimo LMEQM viene calcolato nel dominio della frequenza.

4.2 Filtro di Wiener FIR

Siano $s(k)$ ed $y(k)$ due processi stocastici *congiuntamente stazionari*. Senza perdita di generalità, supponiamo che essi siano a media nulla:

$$\mathbf{E}[s(k)] = \mathbf{E}[y(k)] = 0,$$

e siano $R_s(\tau)$ e $R_y(\tau)$ le rispettive funzioni di covarianza:

$$R_s(\tau) = \mathbf{E}[s(k + \tau)s(k)],$$

$$R_y(\tau) = \mathbf{E}[y(k + \tau)y(k)].$$

Infine, si indichi con $R_{sy}(\tau)$ la funzione di covarianza incrociata:

$$R_{sy}(\tau) = \mathbf{E}[s(k + \tau)y(k)].$$

Un generico filtro lineare si dice a *risposta impulsiva finita* (FIR) se la sua risposta impulsiva $\{g_i\}_{i=0}^{+\infty}$ è diversa da zero solo per un numero finito di campioni. Di conseguenza la funzione di trasferimento $G(z)$ del filtro può essere scritta come:

$$G(z) = \sum_{i=0}^N g_i z^{-i},$$

in cui l'intero N rappresenta l'ordine del filtro.

Riprendendo la formulazione presentata nel paragrafo precedente, l'obiettivo è quello di determinare il filtro lineare FIR che risolve il problema di filtraggio ottimo nel caso in cui $l = 0$ (*filtro di Wiener FIR*). Vale a dire, si desidera stimare il valore assunto dal processo stocastico $s(k)$

utilizzando le ultime $N + 1$ misure disponibili del processo $y(k)$, $Y_{k-N}^k = \{y(k), y(k-1), \dots, y(k-N)\}$.

Al fine di calcolare il filtro LMEQM FIR, si indichi con:

$$\hat{s}(k) = \sum_{i=0}^N g_i y(k-i) \quad (4.3)$$

l'uscita di un generico filtro FIR di ordine N alimentato dal processo delle osservazioni $y(k)$. Per calcolare il filtro di Wiener FIR, occorre determinare i valori dei coefficienti g_i che minimizzano l'errore quadratico medio (4.2):

$$\begin{aligned} EQM &= \mathbf{E} [(s(k) - \hat{s}(k))^2] \\ &= \mathbf{E} \left[\left(s(k) - \sum_{i=0}^N g_i y(k-i) \right) \left(s(k) - \sum_{i=0}^N g_i y(k-i) \right) \right] \\ &= \mathbf{E} [s^2(k)] - 2 \sum_{i=0}^N g_i \mathbf{E} [s(k)y(k-i)] \\ &\quad + \sum_{i=0}^N \sum_{j=0}^N g_i g_j \mathbf{E} [y(k-i)y(k-j)] \\ &= R_s(0) - 2 \sum_{i=0}^N g_i R_{s\mathbf{y}}(i) + \sum_{i=0}^N \sum_{j=0}^N g_i g_j R_{\mathbf{y}}(j-i) \end{aligned}$$

Imponendo che la derivata rispetto a ciascun g_i si annulli in corrispondenza dei coefficienti ottimi g_i^* , si ottiene:

$$-2R_{s\mathbf{y}}(i) + 2 \sum_{j=0}^N g_j^* R_{\mathbf{y}}(j-i) = 0, \quad i = 0, \dots, N,$$

da cui:

$$R_{s\mathbf{y}}(i) = \sum_{j=0}^N g_j^* R_{\mathbf{y}}(j-i), \quad i = 0, \dots, N. \quad (4.4)$$

Le (4.4) sono chiamate equazioni di Wiener-Hopf e rappresentano un sistema lineare di $N + 1$ equazioni in $N + 1$ incognite. Tale sistema può essere riscritto

in forma matriciale come:

$$\begin{pmatrix} R_{\mathbf{sy}}(0) \\ R_{\mathbf{sy}}(1) \\ \vdots \\ R_{\mathbf{sy}}(N) \end{pmatrix} = \begin{pmatrix} R_{\mathbf{y}}(0) & R_{\mathbf{y}}(1) & \dots & R_{\mathbf{y}}(N) \\ R_{\mathbf{y}}(-1) & R_{\mathbf{y}}(0) & \dots & R_{\mathbf{y}}(N-1) \\ \vdots & \vdots & \ddots & \vdots \\ R_{\mathbf{y}}(-N) & R_{\mathbf{y}}(-N+1) & \dots & R_{\mathbf{y}}(0) \end{pmatrix} \begin{pmatrix} g_0^* \\ g_1^* \\ \vdots \\ g_N^* \end{pmatrix}. \quad (4.5)$$

Si noti che la matrice delle covarianze di y che compare in (4.5) ha la proprietà che ogni diagonale è composta da elementi tutti uguali. Una matrice di questo tipo è detta matrice di Toeplitz. Nel paragrafo 2.1 si è visto che la matrice delle covarianze di un processo stocastico y è simmetrica e semidefinita positiva. Nel caso in cui essa risulti definita positiva, l'unica soluzione del sistema lineare (4.5) risulta essere:

$$\begin{pmatrix} g_0^* \\ g_1^* \\ \vdots \\ g_N^* \end{pmatrix} = \begin{pmatrix} R_{\mathbf{y}}(0) & R_{\mathbf{y}}(1) & \dots & R_{\mathbf{y}}(N) \\ R_{\mathbf{y}}(-1) & R_{\mathbf{y}}(0) & \dots & R_{\mathbf{y}}(N-1) \\ \vdots & \vdots & \ddots & \vdots \\ R_{\mathbf{y}}(-N) & R_{\mathbf{y}}(-N+1) & \dots & R_{\mathbf{y}}(0) \end{pmatrix}^{-1} \begin{pmatrix} R_{\mathbf{sy}}(0) \\ R_{\mathbf{sy}}(1) \\ \vdots \\ R_{\mathbf{sy}}(N) \end{pmatrix} \quad (4.6)$$

L'equazione (4.6) fornisce il valore dei coefficienti del filtro di Wiener FIR, la cui funzione di trasferimento è dunque:

$$G(z) = \sum_{i=0}^N g_i^* z^{-i}.$$

Dal principio di ortogonalità dello stimatore LMEQM (vedi (3.28)) discende che:

$$\mathbf{E} [\tilde{s}(k)y(k-i)] = 0, \quad i = 0, \dots, N. \quad (4.7)$$

In effetti, è possibile ricavare i coefficienti g_i^* anche a partire dalla (4.7) (vedi Esercizio 4.3).

Una volta derivata l'espressione del filtro, si può calcolare l'errore qua-

dratico medio come segue:

$$\begin{aligned}
 EQM^* &= \mathbf{E} [\tilde{s}^2(k)] = \mathbf{E} \left[\tilde{s}(k) \left(s(k) - \sum_{i=0}^N g_i^* y(k-i) \right) \right] \\
 &= \mathbf{E} [\tilde{s}(k)s(k)] - \sum_{i=0}^N g_i^* \mathbf{E} [\tilde{s}(k)y(k-i)] \\
 &\stackrel{[1]}{=} \mathbf{E} [\tilde{s}(k)s(k)] \stackrel{[2]}{=} \mathbf{E} [s^2(k)] - \sum_{i=0}^N g_i^* \mathbf{E} [s(k)y(k-i)] \\
 &= R_s(0) - \sum_{i=0}^N g_i^* R_{sy}(i),
 \end{aligned} \tag{4.8}$$

in cui la [1] deriva dal principio di ortogonalità (4.7) e la [2] direttamente dalla definizione di $\tilde{s}(k)$. Per valutare le prestazioni del filtro di Wiener FIR, un tipico modo di procedere consiste nel confrontarle con quelle del cosiddetto filtro “naif”:

$$\hat{s}(k) = y(k),$$

in cui si utilizza direttamente l’ultima misura $y(k)$ come stima $\hat{s}(k)$ (in pratica, senza procedere ad alcun filtraggio). In questo caso l’errore quadratico medio vale:

$$EQM_0 = \mathbf{E} [(s(k) - y(k))^2] = R_s(0) - 2R_{sy}(0) + R_y(0).$$

Un indicatore utilizzato di frequente per valutare le prestazioni del filtro di Wiener FIR è la

$$\text{Riduzione di EQM} = 10 \log_{10} \left(\frac{EQM_0}{EQM^*} \right) \text{ dB.} \tag{4.9}$$

Esempio 4.5. Si supponga di utilizzare un filtro FIR di ordine $N = 2$ per stimare il valore assunto dal p.s. $s(k)$, sulla base delle osservazioni:

$$y(k) = s(k) + v(k),$$

dove il rumore di misura $v(k)$ è bianco, a media nulla e varianza $\sigma_v^2 = 2$, scorrelato da $s(\cdot)$. Inoltre, sia:

$$R_s(\tau) = 0.95^{|\tau|}$$

la funzione di covarianza di $s(k)$, che si suppone essere a media nulla.

Si indichi con:

$$G(z) = g_0^* + g_1^* z^{-1} + g_2^* z^{-2}$$

la funzione di trasferimento del filtro. Occorre determinare i coefficienti g_0^* , g_1^* e g_2^* che minimizzano l'EQM. In accordo con la formula (4.6), si procede con il calcolo delle funzioni di covarianza $R_{\mathbf{y}}(\tau)$ e $R_{\mathbf{s}\mathbf{y}}(\tau)$.

$$\begin{aligned} R_{\mathbf{y}}(\tau) &= \mathbf{E}[y(k+\tau)y(k)] = \mathbf{E}[(s(k+\tau) + v(k+\tau))(s(k) + v(k))] \\ &= R_{\mathbf{s}}(\tau) + R_{\mathbf{v}}(\tau) \end{aligned}$$

dal momento che $s(\cdot)$ e $v(\cdot)$ sono scorrelati.

$$R_{\mathbf{s}\mathbf{y}}(\tau) = \mathbf{E}[s(k+\tau)(y(k))] = \mathbf{E}[s(k+\tau)(s(k) + v(k))] = R_{\mathbf{s}}(\tau)$$

Ne segue che

$$\begin{aligned} \begin{pmatrix} g_0^* \\ g_1^* \\ g_2^* \end{pmatrix} &= \begin{pmatrix} R_{\mathbf{s}}(0) + R_{\mathbf{v}}(0) & R_{\mathbf{s}}(1) + R_{\mathbf{v}}(1) & R_{\mathbf{s}}(2) + R_{\mathbf{v}}(2) \\ R_{\mathbf{s}}(-1) + R_{\mathbf{v}}(-1) & R_{\mathbf{s}}(0) + R_{\mathbf{v}}(0) & R_{\mathbf{s}}(1) + R_{\mathbf{v}}(1) \\ R_{\mathbf{s}}(-2) + R_{\mathbf{v}}(-2) & R_{\mathbf{s}}(-1) + R_{\mathbf{v}}(-1) & R_{\mathbf{s}}(0) + R_{\mathbf{v}}(0) \end{pmatrix}^{-1} \begin{pmatrix} R_{\mathbf{s}}(0) \\ R_{\mathbf{s}}(1) \\ R_{\mathbf{s}}(2) \end{pmatrix} \\ &\stackrel{[1]}{=} \begin{pmatrix} 3 & 0.95 & 0.95^2 \\ 0.95 & 3 & 0.95 \\ 0.95^2 & 0.95 & 3 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 0.95 \\ 0.95^2 \end{pmatrix} = \begin{pmatrix} 0.2203 \\ 0.1919 \\ 0.1738 \end{pmatrix} \end{aligned}$$

dove la [1] discende da:

$$R_{\mathbf{v}}(\tau) = \begin{cases} 2 & \text{se } \tau = 0 \\ 0 & \text{se } \tau \neq 0 \end{cases}$$

In questo caso l'errore quadratico medio vale:

$$EQM^* = R_{\mathbf{s}}(0) - \sum_{i=0}^N g_i^* R_{\mathbf{s}\mathbf{y}}(i) = 0.4406$$

mentre in assenza di filtraggio si avrebbe:

$$EQM_0 = R_{\mathbf{s}}(0) - 2R_{\mathbf{s}\mathbf{y}}(0) + R_{\mathbf{y}}(0) = 1 - 2 + 1 + 2 = 2$$

ottenendo, quindi, una

$$\text{Riduzione di EQM} = 10 \log_{10} \left(\frac{2}{0.4406} \right) = 6.57 \text{ dB.}$$

△

4.3 Predittore di Wiener FIR

In questo paragrafo vedremo come progettare un filtro FIR per predire il valore assunto da un processo stocastico $s(\cdot)$ all'istante $k+l$, con $l > 0$, sulla base delle osservazioni Y_{k-N}^k . Analogamente a quanto fatto in precedenza, all'interno della classe dei filtri lineari a risposta impulsiva finita, si ricerca quello che minimizza l'errore quadratico medio di predizione. Si denoti con:

$$\hat{s}(k+l) = \sum_{i=0}^N g_i y(k-i)$$

la predizione di $s(k+l)$ ottenuta filtrando le osservazioni $y(\cdot)$ fino all'istante k , mediante un filtro FIR di ordine N , rappresentato dalla funzione di trasferimento:

$$G(z) = \sum_{i=0}^N g_i z^{-i}.$$

Al fine di calcolare i coefficienti g_i^* che minimizzano l'EQM, si consideri l'errore di predizione:

$$\tilde{s}(k+l) = s(k+l) - \hat{s}(k+l).$$

Sotto l'ipotesi di stazionarietà dei p.s. $s(k)$ e $y(k)$, imponendo la condizione di ortogonalità:

$$\mathbf{E}[\tilde{s}(k+l)y(k-i)] = 0, \quad i = 0, \dots, N$$

si ottiene:

$$R_{s\mathbf{y}}(l+i) = \sum_{j=0}^N g_j^* R_{\mathbf{y}}(j-i), \quad i = 0, \dots, N$$

da cui:

$$\begin{pmatrix} g_0^* \\ g_1^* \\ \vdots \\ g_N^* \end{pmatrix} = \begin{pmatrix} R_{\mathbf{y}}(0) & R_{\mathbf{y}}(1) & \dots & R_{\mathbf{y}}(N) \\ R_{\mathbf{y}}(-1) & R_{\mathbf{y}}(0) & \dots & R_{\mathbf{y}}(N-1) \\ \vdots & \vdots & \ddots & \vdots \\ R_{\mathbf{y}}(-N) & R_{\mathbf{y}}(-N+1) & \dots & R_{\mathbf{y}}(0) \end{pmatrix}^{-1} \begin{pmatrix} R_{s\mathbf{y}}(l) \\ R_{s\mathbf{y}}(l+1) \\ \vdots \\ R_{s\mathbf{y}}(l+N) \end{pmatrix} \quad (4.10)$$

Il filtro così ottenuto va sotto il nome di *predittore di Wiener FIR*, e rappresenta la predizione lineare l passi in avanti che minimizza l'EQM di predizione. Si osservi che la differenza fra la (4.10) e la (4.6) consiste in una traslazione di l passi in avanti della covarianza incrociata $R_{s\mathbf{y}}(\tau)$.

In maniera del tutto simile a quanto fatto per il filtro di Wiener FIR, l'EQM di predizione può essere calcolato come:

$$EQM = R_s(0) - \sum_{i=0}^N g_i^* R_{s\mathbf{y}}(l+i). \quad (4.11)$$

Esempio 4.6. Si consideri il problema di predizione pura:

$$y(k) = s(k)$$

con:

$$R_s(\tau) = 0.95^{|\tau|}.$$

Si desidera predire il valore di $s(k+1)$ utilizzando un filtro FIR di ordine $N=2$. Allora:

$$R_{\mathbf{y}}(\tau) = R_{s\mathbf{y}}(\tau) = R_s(\tau)$$

da cui:

$$\begin{pmatrix} g_0^* \\ g_1^* \\ g_2^* \end{pmatrix} = \begin{pmatrix} 1 & 0.95 & 0.95^2 \\ 0.95 & 1 & 0.95 \\ 0.95^2 & 0.95 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0.95 \\ 0.95^2 \\ 0.95^3 \end{pmatrix} = \begin{pmatrix} 0.95 \\ 0 \\ 0 \end{pmatrix}$$

Quindi, il predittore di Wiener FIR a un passo in avanti vale:

$$\hat{s}(k+1) = 0.95s(k)$$

vale a dire, utilizza solo la misura più recente. Questa è una caratteristica dei processi esponenzialmente correlati. In generale si può dimostrare che se $s(k)$ è esponenzialmente correlato, il predittore di Wiener FIR ad l passi in avanti assume la forma:

$$\hat{s}(k+l) = \frac{R_s(l)}{R_s(0)} s(k)$$

qualunque sia l'ordine scelto N .

△

Da quanto visto negli ultimi due paragrafi, il calcolo del filtro di Wiener FIR (o, equivalentemente, il predittore FIR) può diventare eccessivamente oneroso quando si desidera utilizzare un elevato numero di misure N . In questi casi, infatti, occorre risolvere un sistema lineare di ordine elevato. Una possibile alternativa, che consente di considerare tutte le misure disponibili ($N = \infty$) ai fini della stima, consiste nell'utilizzare una descrizione del filtro in termini di funzione di trasferimento $G(z)$, anziché di risposta impulsiva $\{g_i\}$. Per questo, nel prossimo paragrafo saranno presentati alcuni risultati sulla fattorizzazione spettrale, che consentiranno di ricavare la f.d.t. del predittore LMEQM basato su tutte le misure $Y_{-\infty}^k$.

4.4 Fattorizzazione spettrale

Nel paragrafo 2.4 si è visto come si modifica lo spettro di un processo stocastico stazionario quando viene filtrato mediante un sistema lineare asintoticamente stabile. Ora ci proponiamo di risolvere il problema inverso. Dato un p.s. stazionario $y(k)$, caratterizzato da uno spettro razionale $\phi_y(z)$, si vuole trovare una funzione di trasferimento $H(z)$ razionale ed una varianza σ_e^2 , tale che

$$\phi_y(z) = H(z)H(z^{-1})\sigma_e^2.$$

Tale problema va sotto il nome di fattorizzazione spettrale e, se non si impongono altri vincoli, la soluzione non è unica, come mostrano i seguenti esempi.

Esempio 4.7. Sia $H_1(z)$ soluzione del problema di fattorizzazione spettrale con varianza σ_1^2 . Vale a dire:

$$\phi_y(z) = H_1(z)H_1(z^{-1})\sigma_1^2.$$

Allora, anche:

$$H_2(z) = kH_1(z), \quad k \neq 0$$

è soluzione con varianza $\sigma_2^2 = \frac{\sigma_1^2}{k^2}$. Infatti:

$$H_2(z)H_2(z^{-1})\sigma_2^2 = kH_1(z)kH_1(z^{-1})\frac{\sigma_1^2}{k^2} = H_1(z)H_1(z^{-1})\sigma_1^2 = \phi_y(z).$$

△

Esempio 4.8. Sia $H_1(z)$ soluzione del problema di fattorizzazione spettrale con varianza σ_1^2 e si supponga che $H_1(z)$ abbia uno zero in $z = \alpha$:

$$\phi_y(z) = H_1(z)H_1(z^{-1})\sigma_1^2 = \tilde{H}(z)(z - \alpha)\tilde{H}(z^{-1})(z^{-1} - \alpha)\sigma_1^2.$$

Allora, anche:

$$H_2(z) = \tilde{H}(z)\left(z - \frac{1}{\alpha}\right)$$

è soluzione con varianza $\sigma_2^2 = \alpha^2\sigma_1^2$. Infatti:

$$\begin{aligned} H_2(z)H_2(z^{-1})\sigma_2^2 &= \tilde{H}(z)\left(z - \frac{1}{\alpha}\right)\tilde{H}(z^{-1})\left(z^{-1} - \frac{1}{\alpha}\right)\alpha^2\sigma_1^2 \\ &= \tilde{H}(z)\frac{z}{\alpha}(\alpha - z^{-1})\tilde{H}(z^{-1})\frac{z^{-1}}{\alpha}(\alpha - z)\alpha^2\sigma_1^2 \\ &= \tilde{H}(z)(z - \alpha)\tilde{H}(z^{-1})(z^{-1} - \alpha)\sigma_1^2 = \phi_y(z) \end{aligned}$$

Quindi, se una f.d.t. con uno zero in α è soluzione con varianza σ^2 , anche la f.d.t. che si ottiene sostituendo lo zero in α col suo reciproco $\frac{1}{\alpha}$ è soluzione con varianza $\alpha^2\sigma^2$. Un discorso analogo vale nel caso in cui si scambi un polo col suo reciproco. △

Al fine di garantire l'unicità della soluzione del problema di fattorizzazione occorre quindi imporre alcuni vincoli sulla funzione di trasferimento $H(z)$.

Teorema 11. *Dato uno spettro razionale*

$$\phi_y(z) = \frac{\sum_{k=-m}^m \beta_k z^{-k}}{\sum_{h=-n}^n \alpha_h z^{-h}}$$

che non possiede poli sulla circonferenza unitaria, la fattorizzazione spettrale

$$\phi_y(z) = H(z)H(z^{-1})\sigma_e^2,$$

con

$$H(z) = \frac{B(z)}{A(z)}$$

esiste ed è unica se si impongono le seguenti condizioni:

1. *i polinomi $A(z)$ e $B(z)$ sono coprimi e monici, ovvero hanno forma*

$$A(z) = 1 + \alpha_1 z^{-1} + \dots + \alpha_n z^{-n}$$

$$B(z) = 1 + \beta_1 z^{-1} + \dots + \beta_m z^{-m}$$

e non hanno zeri in comune;

2. *gli zeri di $A(z)$ sono interni alla circonferenza unitaria e quelli di $B(z)$ sono interni o sulla circonferenza unitaria.*

La funzione di trasferimento $H(z)$ che soddisfa i precedenti requisiti viene chiamata *fattore spettrale canonico* di $\phi_y(z)$. Si osservi che il requisito 2) garantisce che $H(z)$ sia asintoticamente stabile. In questo modo, il p.s. $y(k)$, caratterizzato da uno spettro $\phi_y(z)$, può essere sempre visto come l'uscita di un sistema lineare avente f.d.t. $H(z)$, alimentato da un rumore bianco $e(k)$ con varianza σ_e^2 .

Esempio 4.9. Sia:

$$\phi_y(z) = \frac{1}{1.25 - 0.5z - 0.5z^{-1}}$$

lo spettro del p.s. $y(k)$. Si desidera calcolare il fattore spettrale canonico

$$H(z) = \frac{B(z)}{A(z)}$$

e la varianza σ_e^2 . Chiaramente, $B(z) = 1$. Ai fini del calcolo di $A(z)$ si osservi che il denominatore di $\phi_y(z)$ può essere riscritto come:

$$1.25 - 0.5z - 0.5z^{-1} = (1 - 0.5z^{-1})(1 - 0.5z),$$

per cui:

$$A(z) = 1 - 0.5z^{-1}.$$

Concludendo, il fattore spettrale canonico vale:

$$H(z) = \frac{1}{1 - 0.5z^{-1}},$$

con la varianza $\sigma_e^2 = 1$.

△

Esempio 4.10. Si consideri il processo stocastico tempo discreto $y(t)$ generato come rappresentato in Figura 4.2, dove l'ingresso $u(t)$ è un processo stocastico tale che $u(t) = w(t) + 2w(t-1)$ e $G(z) = \frac{1}{1 + 0.5z^{-1}}$. Suppo-

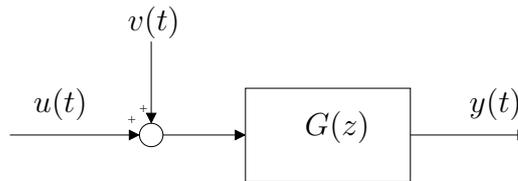


Figura 4.2.

nendo che $w(t)$ e $v(t)$ siano due rumori bianchi gaussiani scorrelati a media nulla e varianza $\sigma_w^2 = 1$ e $\sigma_v^2 = 2$, si desidera determinare il fattore spettrale canonico di $\phi_y(z)$.

Indicando con $\phi_x(z)$ lo spettro del processo stocastico

$$x(t) = u(t) + v(t),$$

è possibile calcolare lo spettro di $y(t)$ come:

$$\phi_y(z) = G(z)G(z^{-1})\phi_x(z).$$

Dal momento che i rumori $w(t)$ e $v(t)$ sono tra di loro scorrelati, si ha che:

$$\phi_x(z) = \phi_u(z) + \phi_v(z).$$

Ovviamente, essendo $v(t)$ bianco, il suo spettro vale:

$$\phi_v(z) = \sigma_v^2 = 2,$$

mentre lo spettro di $u(t)$ (processo stocastico MA(1)) vale:

$$\phi_u(z) = 2z + 5 + 2z^{-1}.$$

Allora:

$$\begin{aligned} \phi_y(z) &= G(z)G(z^{-1})(\phi_u(z) + \phi_v(z)) \\ &= \frac{2z + 7 + 2z^{-1}}{(1 + 0.5z^{-1})(1 + 0.5z)}. \end{aligned}$$

Sia

$$H(z) = \frac{B(z)}{A(z)}$$

il fattore spettrale canonico di $\phi_y(z)$, con varianza σ_e^2 da determinare. Il denominatore è già fattorizzato, per cui:

$$A(z) = 1 + 0.5z^{-1},$$

mentre, per calcolare il numeratore e la varianza σ_e^2 , ipotizziamo per $B(z)$ la seguente forma:

$$B(z) = 1 + bz^{-1},$$

con b coefficiente da determinare. Imponendo l'uguaglianza dei polinomi

$$B(z)B(z^{-1})\sigma_e^2 = (1 + bz^{-1})(1 + bz)\sigma_e^2 = 2z + 7 + 2z^{-1}$$

si ottiene il seguente sistema:

$$\begin{cases} \sigma_e^2 b = 2 \\ \sigma_e^2 (1 + b^2) = 7 \end{cases}$$

da cui

$$\begin{cases} \sigma_e^2 = \frac{2}{b} \\ 2b^2 - 7b + 2 = 0 \end{cases}$$

La seconda equazione ammette due soluzioni:

$$b_1 = 0.3139, \quad b_2 = 3.1861,$$

ciascuna il reciproco dell'altra.¹ Dal momento che b corrisponde proprio allo zero di $B(z)$, scegliamo la radice interna al cerchio unitario:

$$b = 0.3139,$$

affinché sia verificato il requisito 2) del fattore spettrale canonico. A questo punto:

$$\sigma_e^2 = \frac{2}{b} = 6.3715$$

e

$$H(z) = \frac{1 + 0.3139z^{-1}}{1 + 0.5z^{-1}}.$$

△

4.5 Predittore di Wiener

In questo paragrafo vedremo come risolvere il problema di predizione pura (vedi Esempio 4.6), qualora si vogliano utilizzare tutte le osservazioni disponibili fino all'istante in cui si effettua la predizione. L'obiettivo è stimare il valore assunto dal p.s. $y(\cdot)$ all'istante $k+l$, con $l > 0$, sulla base delle osservazioni $Y_{-\infty}^k = \{y(k), y(k-1), y(k-2), \dots\}$. In questo contesto, la stima di $\hat{y}(k+l|Y_{-\infty}^k)$ viene indicata più brevemente con $\hat{y}(k+l|k)$ (ovvero, stima di $y(k+l)$ basata su misure fino all'istante k).

Al solito, si desidera calcolare lo stimatore lineare di $y(k+l)$ a minimo errore quadratico medio. A differenza di quanto fatto nel paragrafo 4.6,

¹Si noti che questo accade sempre, per le proprietà degli spettri.

quando il numero di osservazioni è infinito, è preferibile lavorare nel dominio della frequenza, per cui ci si propone di calcolare direttamente la funzione di trasferimento $G(z)$ del predittore:

$$\hat{y}(k+l|k) = G(z)y(k)$$

che minimizza l'EQM, anziché la sua risposta impulsiva.

Sia:

$$H(z) = \frac{C(z)}{A(z)}$$

il fattore spettrale canonico di $\phi_y(z)$, con varianza σ_e^2 . Questo significa che denotando con $e(k)$ un processo stocastico bianco con varianza σ_e^2 ,

$$y(k) = H(z)e(k). \quad (4.12)$$

È sempre possibile riscrivere $H(z)$ come:

$$\frac{C(z)}{A(z)} = Q(z) + \frac{R(z)}{A(z)}z^{-l}$$

in cui $Q(z)$ rappresenta il quoziente e $R(z)z^{-l}$ il resto della divisione $\frac{C}{A}$. Da un punto di vista operativo, questo vuol dire procedere con la divisione lunga tra i polinomi (in z^{-1}) $C(z)$ ed $A(z)$ finché nel resto non compaia il fattore comune z^{-l} in tutti i termini del polinomio (e questo accade dopo *al più* l passi di divisione). Inoltre, è possibile verificare che il grado del polinomio $R(z)$ è minore di quello di $A(z)$:

$$\deg(R) < \deg(A),$$

mentre il grado del polinomio $Q(z)$ è minore di l :

$$\deg(Q) < l.$$

Sia:

$$Q(z) = q_0 + q_1z^{-1} + \dots + q_{l-1}z^{-(l-1)}.$$

Allora, dalla (4.12), segue:

$$\begin{aligned}
 y(k+l) &= H(z)e(k+l) = \frac{C(z)}{A(z)}e(k+l) \\
 &= \left(Q(z) + \frac{R(z)}{A(z)}z^{-l} \right) e(k+l) \\
 &= Q(z)e(k+l) + \frac{R(z)}{A(z)}e(k).
 \end{aligned} \tag{4.13}$$

Inoltre, per le proprietà del fattore spettrale canonico, $H(z)$ è asintoticamente stabile e, assumendo che $H(z)$ non abbia zeri sulla circonferenza unitaria, anche la sua inversa $H^{-1}(z)$ lo è. Perciò è possibile invertire la relazione (4.12), ottenendo:

$$e(k) = H^{-1}(z)y(k) = \frac{A(z)}{C(z)}y(k).$$

Sostituendo l'equazione precedente nella (4.13), si ottiene:

$$y(k+l) = Q(z)e(k+l) + \frac{R(z)}{C(z)}y(k). \tag{4.14}$$

Si noti che il primo termine a destra della precedente equazione corrisponde a:

$$Q(z)e(k+l) = q_0e(k+l) + q_1e(k+l-1) + \dots + q_{l-1}e(k+1)$$

e dipende dal valore assunto dal rumore $e(\cdot)$ in istanti successivi a quello di osservazione k . Inoltre, per la proprietà del rumore bianco, i valori $e(k+i)$, $i = 1, \dots, l$ sono scorrelati da tutte le osservazioni $Y_{-\infty}^k = \{y(k), y(k-1), y(k-2), \dots\}$, per cui sono imprevedibili. La miglior stima possibile di $e(k+i)$, quindi, corrisponde al suo valor medio

$$\hat{e}(k+i) = \mathbf{E}[e(k+i)] = 0.$$

Si può quindi concludere che la stima LMEQM di $y(k+l)$, sulla base delle misure disponibili fino all'istante k , è costituita dal secondo termine a destra nella (4.14), ovvero si ha:

$$\hat{y}(k+l|k) = \frac{R(z)}{C(z)}y(k). \tag{4.15}$$

Al fine di calcolare l'errore quadratico medio di predizione, sottraendo la (4.15) dalla (4.14) si ottiene:

$$\begin{aligned}\tilde{y}(k+l) &= y(k+l) - \hat{y}(k+l|k) = Q(z)e(k+l) \\ &= q_0e(k+l) + q_1e(k+l-1) + \cdots + q_{l-1}e(k+1) \\ &= \sum_{i=0}^{l-1} q_i e(k+l-i).\end{aligned}$$

Quindi:

$$\begin{aligned}EQM^* &= \mathbf{E} [\tilde{y}(k+l)^2] = \mathbf{E} \left[\left(\sum_{i=0}^{l-1} q_i e(k+l-i) \right)^2 \right] \\ &= \sigma_e^2 \sum_{i=0}^{l-1} q_i^2,\end{aligned}\tag{4.16}$$

dal momento che, essendo $e(\cdot)$ un rumore bianco, il valore atteso dei prodotti misti è nullo.

Esempio 4.11. Si consideri il processo stocastico ARMA(1,1), generato nel seguente modo:

$$y(k) - \frac{1}{2}y(k-1) = w(k) - 4w(k-1),$$

dove $w(k)$ è un rumore bianco a media nulla, con varianza $\sigma_w^2 = 2$.

Si desidera calcolare il predittore ottimo a 2 passi in avanti $\hat{y}(k+2|k)$. La funzione di trasferimento $W(z)$ del sistema che genera $y(k)$ a partire da $w(k)$ vale:

$$W(z) = \frac{1 - 4z^{-1}}{1 - \frac{1}{2}z^{-1}},$$

per cui lo spettro di $y(k)$ risulta:

$$\phi_y(z) = W(z)W(z^{-1})\sigma_w^2 = \frac{(1 - 4z^{-1})(1 - 4z)}{(1 - \frac{1}{2}z^{-1})(1 - \frac{1}{2}z)} \cdot 2$$

Ai fini del calcolo del fattore spettrale canonico

$$H(z) = \frac{C(z)}{A(z)}$$

osserviamo che, dalla fattorizzazione del denominatore, discende immediatamente

$$A(z) = 1 - \frac{1}{2}z^{-1}.$$

Il polinomio in z^{-1} del numeratore di $\phi_y(z)$, invece, presenta uno zero instabile ($z = 4$). Per calcolare $C(z)$ possiamo sfruttare l'osservazione fatta nell'Esempio 4.8, vale a dire possiamo scambiare lo zero instabile con il suo reciproco, a patto di modificare la varianza $\sigma_e^2 = 16 \cdot \sigma_w^2$. Questo significa che lo spettro può essere riscritto come:

$$\phi_y(z) = \frac{(1 - \frac{1}{4}z^{-1})(1 - \frac{1}{4}z)}{(1 - 0.5z^{-1})(1 - 0.5z)} \cdot 32$$

da cui:

$$C(z) = 1 - \frac{1}{4}z^{-1}$$

e $\sigma_e^2 = 32$. Il predittore ottimo a 2 passi in avanti vale:

$$\hat{y}(k+2|k) = \frac{R(z)}{C(z)}y(k),$$

dove

$$\frac{C(z)}{A(z)} = Q(z) + \frac{R(z)}{A(z)}z^{-2}.$$

Effettuiamo la divisione lunga $\frac{C(z)}{A(z)}$ finché nel resto non compare z^{-2} a fattore comune:

$$\begin{array}{r} 1 - \frac{1}{4}z^{-1} \\ 1 - \frac{1}{2}z^{-1} \\ \hline \frac{1}{4}z^{-1} \\ \frac{1}{4}z^{-1} - \frac{1}{8}z^{-2} \\ \hline \frac{1}{8}z^{-2} \end{array} \quad \begin{array}{r} \boxed{1 - \frac{1}{2}z^1} \\ 1 + \frac{1}{4}z^{-1} \end{array}$$

Quindi:

$$R(z) = \frac{1}{8} \quad Q(z) = 1 + \frac{1}{4}z^{-1}$$

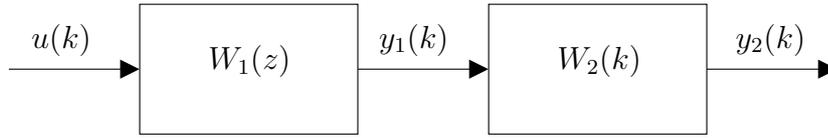


Figura 4.3.

e il predittore LMEQM a 2 passi in avanti assume la forma:

$$\hat{y}(k+2|k) = \frac{\frac{1}{8}}{1 - \frac{1}{4}z^{-1}} y(k).$$

Si noti che, nel dominio del tempo, l'equazione del predittore equivale ad una ricorsione che coinvolge il valore predetto dal filtro all'istante precedente:

$$\hat{y}(k+2|k) = \frac{1}{4}\hat{y}(k+1|k-1) + \frac{1}{8}y(k).$$

L'EQM del predittore appena calcolato vale:

$$EQM^* = (q_0^2 + q_1^2)\sigma_e^2 = \left(1 + \frac{1}{16}\right) 32 = 34.$$

Esempio 4.12. Si consideri il sistema rappresentato in Figura 4.12, in cui $u(k)$ è un processo stocastico che soddisfa l'equazione:

$$u(k) = w(k) + \frac{1}{5}w(k-1),$$

con $w(k)$ processo stocastico bianco, a media nulla e di varianza $\sigma_w^2 = 4$.

Assumendo:

$$W_1(z) = \frac{1}{1 + \frac{1}{4}z^{-1}}, \quad W_2(z) = \frac{1 + \frac{3}{10}z^{-1}}{1 + \frac{1}{4}z^{-1}}$$

si desidera calcolare l'equazione del predittore ottimo lineare MEQM a due passi in avanti $\hat{y}_2(k+2|k)$.

Innanzitutto, calcoliamo lo spettro di y_2 . Sia:

$$W(z) = W_1(z)W_2(z) = \frac{(1 + \frac{3}{10}z^{-1})}{(1 + \frac{1}{4}z^{-1})^2}$$

la funzione di trasferimento dall'ingresso $u(k)$ all'uscita $y_2(k)$. Allora:

$$\phi_{y_2}(z) = W(z)W(z^{-1})\phi_u(z),$$

dove

$$\phi_u(z) = (1 + \frac{1}{5}z^{-1})(1 + \frac{1}{5}z).$$

Per cui:

$$\begin{aligned} \phi_{y_2}(z) &= \frac{(1 + \frac{1}{5}z^{-1})(1 + \frac{3}{10}z^{-1})}{(1 + \frac{1}{4}z^{-1})^2} \frac{(1 + \frac{1}{5}z)(1 + \frac{3}{10}z)}{(1 + \frac{1}{4}z)^2} \cdot 4 \\ &\triangleq \frac{C(z)C(z^{-1})}{A(z)A(z^{-1})}\sigma_e^2. \end{aligned}$$

Osservando lo spettro appena calcolato, è possibile notare che i polinomi $C(z)$ ed $A(z)$ soddisfano le condizioni del Teorema 11, per cui il fattore spettrale canonico risulta essere:

$$H(z) = \frac{C(z)}{A(z)} = \frac{(1 + \frac{1}{5}z^{-1})(1 + \frac{3}{10}z^{-1})}{(1 + \frac{1}{4}z^{-1})^2} = \frac{1 + \frac{1}{2}z^{-1} + \frac{3}{50}z^{-2}}{1 + \frac{1}{2}z^{-1} + \frac{1}{16}z^{-2}}$$

con varianza $\sigma_e^2 = 4$. Per calcolare i coefficienti della funzione di trasferimento del predittore, effettuiamo la divisione lunga $\frac{C(z)}{A(z)}$:

$$\begin{array}{r} 1 + \frac{1}{2}z^{-1} + \frac{3}{50}z^{-2} \quad \left| \begin{array}{l} 1 + \frac{1}{2}z^{-1} + \frac{1}{16}z^{-2} \\ 1 + \frac{1}{2}z^{-1} + \frac{1}{16}z^{-2} \\ \hline -\frac{1}{400}z^{-2} \end{array} \right. \\ \hline \end{array}$$

da cui:

$$R(z) = -\frac{1}{400} \quad Q(z) = 1.$$

Si osservi che, sebbene si stia cercando il predittore a 2 passi in avanti, in questo caso è bastato un solo passo di divisione perché comparisse il fattore comune z^{-2} . In generale, con un orizzonte di predizione pari ad l , occorre effettuare al più l passi di divisione. La funzione di trasferimento del predittore è quindi:

$$G(z) = \frac{R(z)}{C(z)} = \frac{-\frac{1}{400}}{1 + \frac{1}{2}z^{-1} + \frac{3}{50}z^{-2}},$$

che corrisponde alla seguente equazione ricorsiva nel dominio del tempo:

$$\hat{y}_2(k+2|k) = -\frac{1}{2}\hat{y}_2(k+1|k-1) - \frac{3}{50}\hat{y}_2(k|k-2) - \frac{1}{400}y(k).$$

Infine, l'errore quadratico medio di predizione vale:

$$EQM^* = q_0^2 \sigma_e^2 = 4.$$

4.6 Esercizi

4.1. Utilizzando MATLAB[®], ripetere l'esercizio presentato nell'Esempio 4.5 al variare dell'ordine del filtro FIR, $N = 1, \dots, 20$, e plottare l' EQM^* e la riduzione di EQM in funzione di N . Per la costruzione dell'equazioni di Wiener-Hopf, si usi il comando `toeplitz` di MATLAB[®].

4.2. Si consideri il problema di predizione pura:

$$y(k) = s(k),$$

e sia:

$$R_s(\tau) = 0.5\delta(\tau) + 0.9^{|\tau|} \cos\left(\frac{\pi\tau}{8}\right)$$

la matrice di covarianza del processo stocastico $s(k)$. Calcolare i coefficienti del predittore di Wiener FIR di ordine $N = 2$ a due passi in avanti.

4.3. Ricavare l'espressione dei coefficienti g_i^* del filtro di Wiener FIR di ordine N imponendo le condizioni di ortogonalità (4.7).

4.4. Si consideri il sistema rappresentato in Figura 4.4, in cui $e(t)$ è un processo stocastico bianco di varianza $\sigma_e^2 = 16$, e K è un guadagno costante reale.

- a) Determinare per quali valori di $K \in \mathbb{R}$ il processo stocastico $y(t)$ è asintoticamente stazionario.
- b) Assumendo $K = -\frac{1}{3}$, calcolare la funzione di covarianza $R_y(\tau)$ del processo $y(t)$, per ogni intero τ .

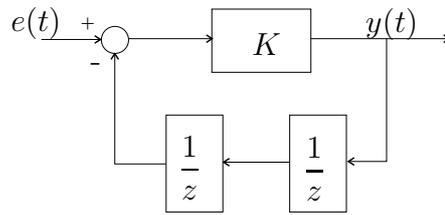


Figura 4.4.

- c) Assumendo $K = \frac{1}{5}$, determinare l'equazione del predittore ottimo lineare MEQM a due passi in avanti $\hat{y}(t+2|t)$ per il processo stocastico $y(t)$, ed il relativo errore quadratico medio di predizione $EQM = \mathbf{E} [\{y(t+2) - \hat{y}(t+2|t)\}^2]$.
- d) Determinare per quali valori di K , l'errore quadratico medio di predizione a quattro passi in avanti, $\mathbf{E} [\{y(t+4) - \hat{y}(t+4|t)\}^2]$, ottenuto mediante il predittore ottimo lineare MEQM, risulta essere inferiore a 60.

4.5. Si consideri il sistema rappresentato in Figura 4.5, in cui $u(t)$ è un processo stocastico che soddisfa l'equazione

$$u(t) = 0.6u(t-1) + e(t),$$

con $e(t)$ processo stocastico bianco di varianza $\sigma_e^2 = 1$, e

$$G(z) = 1 + 4z^{-1}.$$

- a) Assumendo $v(t) = 0, \forall t$, determinare l'equazione del predittore ottimo lineare MEQM a due passi in avanti $\hat{y}(t+2|t)$ per il processo stocastico $y(t)$, ed il relativo errore quadratico medio di predizione $EQM = \mathbf{E} [\{y(t+2) - \hat{y}(t+2|t)\}^2]$.
- b) Ripetere l'esercizio al punto a), assumendo $v(t)$ processo stocastico bianco di varianza $\sigma_v^2 = 10$, indipendente da $e(t)$.
- c) Si assuma ora che $y(t)$ sia un processo stocastico di tipo AR(1), di equazione

$$y(t) = 0.6y(t-1) + w(t),$$

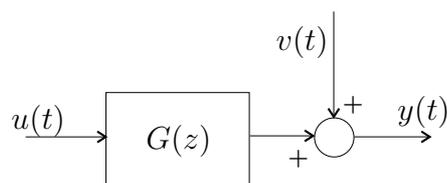


Figura 4.5.

con $w(t)$ processo stocastico bianco di varianza $\sigma_w^2 = 19$, indipendente da $e(t)$. Calcolare lo spettro $\Phi_v(z)$ di $v(t)$ e indicare un possibile modello del processo stocastico $v(t)$.

4.6. Si consideri il sistema rappresentato in Figura 4.6, in cui $e(t)$ è un



Figura 4.6.

processo stocastico bianco, a media nulla e varianza $\sigma_e^2 = 3$, e $P(z) = 1 - 2z^{-1}$.

- Assumendo $G(z) = 1$, determinare l'equazione del predittore ottimo lineare MEQM un passo in avanti $\hat{y}(t+1|t)$ per il processo stocastico $y(t)$, ed il relativo errore quadratico medio di predizione $EQM = \mathbf{E}[\{y(t+1) - \hat{y}(t+1|t)\}^2]$.
- Assumendo $G(z) = K$, con K costante reale positiva, determinare per quali valori di K l'EQM definito al punto a) risulta non superiore a 10.
- Assumendo $G(z) = \frac{1}{5 - z^{-1}}$, ripetere l'esercizio del punto a) precedente.

4.7. Si consideri il sistema rappresentato in Figura 4.7, in cui $e_1(t)$ e $e_2(t)$ sono processi stocastici stazionari, bianchi, tra loro indipendenti, di varianza rispettivamente pari a $\sigma_1^2 = 2$ e $\sigma_2^2 = 1$, e

$$G_1(z) = 1 - \frac{1}{3}z^{-1} \quad , \quad G_2(z) = 1 + 2z^{-1} .$$

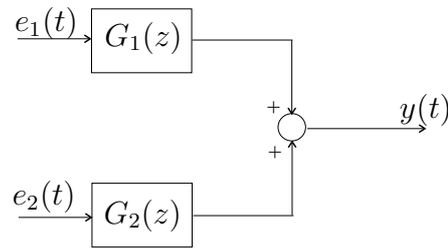


Figura 4.7.

- a) Determinare lo spettro $\Phi_y(z)$ del processo stocastico $y(t)$.
- b) Determinare l'equazione del predittore ottimo lineare MEQM un passo in avanti $\hat{y}(t+1|t)$ per il processo stocastico $y(t)$, ed il relativo errore quadratico medio di predizione $EQM = \mathbf{E} [\{y(t+1) - \hat{y}(t+1|t)\}^2]$.
- c) Si consideri il predittore "naif" $\tilde{y}(t+1|t) = y(t)$. Calcolare l'errore quadratico medio di predizione $EQM_{naif} = \mathbf{E} [\{y(t+1) - \tilde{y}(t+1|t)\}^2]$ e confrontarlo con EQM calcolato al punto b). Motivare il risultato ottenuto.
- d) Sia $r(t)$ il processo stocastico generato come uscita di un sistema lineare di funzione di trasferimento $W(z)$, avente per ingresso il processo $y(t)$. Determinare la funzione di trasferimento $W(z)$ in modo tale che $r(t)$ risulti essere un processo stazionario bianco, di varianza pari a $\frac{1}{4}$.

4.8. Si consideri il sistema rappresentato in Figura 4.7, in cui $e_1(t)$ e $e_2(t)$ sono processi stocastici stazionari, bianchi, tra loro indipendenti, di varianza rispettivamente pari a $\sigma_1^2 = 6$ e $\sigma_2^2 = 2$, e

$$G_1(z) = \frac{1}{1 - \frac{1}{2}z^{-1}} \quad , \quad G_2(z) = \frac{1}{1 - pz^{-1}} \quad .$$

- a) Determinare per quali valori di $p \in \mathbb{R}$, il processo stocastico $y(t)$ è asintoticamente stazionario.
- b) Assumendo $p = \frac{1}{3}$, determinare l'equazione del predittore ottimo lineare MEQM a due passi in avanti $\hat{y}(t+2|t)$ per il processo stocasti-

co $y(t)$, ed il relativo errore quadratico medio di predizione $EQM = \mathbf{E}\mathbf{E}[\{y(t+2) - \hat{y}(t+2|t)\}^2]$.

- c) Assumendo $p = \frac{1}{2}$, determinare il valore massimo che può assumere la varianza σ_2^2 del processo $e_2(t)$, affinché l'EQM del predittore ottimo lineare MEQM a due passi in avanti $\hat{y}(t+2|t)$ non sia superiore a 20.

4.9. Si consideri il sistema rappresentato in Figura 4.8, in cui $u(t)$ e $v(t)$ sono

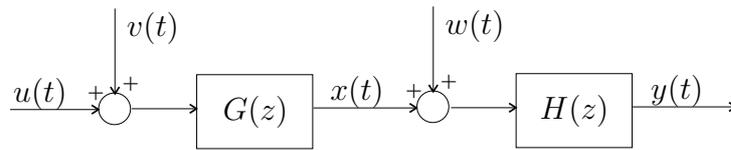


Figura 4.8.

processi stocastici bianchi, tra loro indipendenti, di varianza unitaria, e

$$G(z) = \frac{1 + \frac{1}{2}z^{-1}}{1 - \frac{1}{4}z^{-1}}, \quad H(z) = \frac{1 + 2z^{-1}}{1 - \frac{1}{3}z^{-1}}.$$

- a) Determinare l'equazione del predittore ottimo lineare MEQM a un passo in avanti $\hat{x}(t+1|t)$ per il processo stocastico $x(t)$, ed il relativo errore quadratico medio di predizione $EQM = \mathbf{E}[\{x(t+1) - \hat{x}(t+1|t)\}^2]$.
- b) Assumendo $w(t) \equiv 0$, determinare l'equazione del predittore ottimo lineare MEQM a due passi in avanti $\hat{y}(t+2|t)$ per il processo stocastico $y(t)$, ed il relativo errore quadratico medio di predizione $EQM = \mathbf{E}[\{y(t+2) - \hat{y}(t+2|t)\}^2]$.
- c) Determinare lo spettro $\Phi_y(z)$ del processo $y(t)$ nel caso in cui:
- 1) $w(t)$ sia un processo stocastico bianco, indipendente da $u(t)$ e $v(t)$, di varianza pari a $\sigma_w^2 = 9$;
 - 2) $w(t) \equiv v(t), \forall t$.

Capitolo 5

Identificazione parametrica di sistemi dinamici

In questo capitolo viene affrontato il problema dell'identificazione di modelli matematici per la descrizione di sistemi dinamici. In particolare, l'attenzione si concentrerà esclusivamente sull'identificazione di modelli parametrici per sistemi lineari tempo-invarianti. Per una trattazione completa si rimanda al libro (Ljung, 1999).

5.1 Introduzione

Con il termine *identificazione* si designa l'insieme di tecniche che consentono di costruire modelli matematici di sistemi dinamici, a partire da misure sperimentali (dati).

Gli elementi principali del problema sono:

- un set di dati; ad esempio, una sequenza di N campioni dell'ingresso $u(t)$ e dell'uscita $y(t)$ di un sistema dinamico, ovvero

$$\mathcal{Z}^N = \{u(1), y(1), \dots, u(N), y(N)\}; \quad (5.1)$$

- una classe di modelli candidati, di *struttura* prefissata, all'interno della

quale scegliere il modello identificato (ad esempio: funzioni di trasferimento di ordine fissato);

- un criterio di selezione del modello “ottimo”;
- le procedure di validazione del modello.

Il ruolo di ciascun elemento viene illustrato brevemente nel seguito.

Set di dati

I *dati* sono l'insieme delle misure effettuate sul sistema, in base alle quali si desidera stimare un modello del sistema stesso. In questo corso, verranno considerati soltanto insiemi di dati ingresso-uscita nel dominio del tempo, del tipo \mathcal{Z}^N in (5.1).

In molti casi è possibile scegliere l'ingresso da fornire al sistema, o, più in generale, le condizioni operative del sistema stesso. Come vedremo in seguito, ciò può influenzare significativamente la qualità del modello identificato. Altri aspetti importanti sono la durata dell'esperimento (cioè la lunghezza N del set di dati), il tempo di campionamento, ecc. Tutto ciò fa parte del “progetto dell'esperimento di identificazione”, che costituisce la prima fase dell'intera procedura. Purtroppo, non è sempre possibile scegliere l'ingresso e le condizioni operative sono fissate (ad esempio, quando il sistema in esame è inserito in un anello di controllo e gli ingressi sono selezionati dal controllore).

Classe di modelli

La scelta della classe di modelli rappresenta la parte più delicata della procedura di identificazione, poiché condiziona pesantemente il risultato finale. In generale, un modello può essere visto come una mappa dai dati alle uscite:

$$\hat{y}(t; \theta) = g(\theta, \mathcal{Z}^{t-1})$$

dove $\hat{y}(t; \theta)$ rappresenta l'uscita all'istante t predetta dal modello, \mathcal{Z}^{t-1} sono i dati raccolti fino all'istante $t-1$ e θ è un insieme di elementi da determinare che definiscono lo specifico modello all'interno della classe.

La classe di modelli sarà dunque individuata da

$$\mathcal{M} = \{g(\theta, \cdot) : \theta \in \Theta \subset \mathcal{S}\}$$

dove Θ è l'insieme che definisce la conoscenza a priori sul modello e \mathcal{S} è lo spazio degli elementi θ .

È possibile individuare due macrotipologie di modelli:

- modelli *parametrici*, in cui θ è un vettore di parametri, di dimensione finita (ad esempio, i coefficienti dei polinomi di una funzione di trasferimento razionale $G(z)$);
- modelli *non parametrici*, in cui θ è una funzione che descrive alcune caratteristiche fondamentali del modello; tipici esempi sono la risposta impulsiva o al gradino (nel dominio del tempo), o la risposta in frequenza (nel dominio della frequenza) di un sistema lineare tempo-invariante.

Nell'ambito dell'identificazione parametrica si possono distinguere modelli in cui i parametri derivano da leggi fisiche o conoscenze di altro tipo (modelli a “scatola grigia”), e modelli in cui i parametri sono usati per rappresentare il comportamento ingresso-uscita del sistema, senza alcun particolare significato fisico (modelli a “scatola nera”). Inoltre, i modelli possono essere classificati in base al grado di complessità (lineari o non lineari, di ordine ridotto o elevato, tempo-invarianti o tempo-varianti, a parametri concentrati o distribuiti, ecc.). In generale, un requisito fondamentale per una classe di modelli è la capacità di rappresentare i fenomeni di interesse nell'ambito considerato (modello dipendente dallo scopo).

In questo corso, verranno considerati modelli lineari tempo-invarianti (cioè descritti da funzioni di trasferimento razionali).

Selezione del modello “ottimo”

Una volta raccolti i dati e fissata la classe di modelli \mathcal{M} , il problema dell'identificazione consiste nel determinare l'elemento θ^* che minimizzi un certo funzionale di costo $J(\theta)$:

$$\hat{\theta}^* = \arg \min_{\theta \in \Theta} J(\theta)$$

nell'ambito della classe di modelli scelta. Si hanno così problemi di *ottimizzazione* (tanto più complessi quanto più ricca è la classe di modelli e più complicato è il funzionale $J(\theta)$).

Validazione del modello

Una volta che il modello $g(\theta^*, \cdot)$ è stato identificato, occorre *validarlo*, ossia verificare se esso è sufficientemente buono per gli scopi prefissi, e come si relaziona ai dati disponibili e alla conoscenza a priori. Esistono varie procedure generali per la validazione di un modello. L'importanza attribuita all'esito di ciascuna di esse è tuttavia fortemente dipendente dall'utilizzo che si intende fare del modello una volta che lo si è identificato (ad esempio, un modello impiegato per il progetto di un sistema di controllo avrà requisiti diversi da quelli di un modello utilizzato per fare una predizione dell'uscita del sistema).

Un modello può non essere validato per diverse ragioni:

- non convergenza dell'algoritmo di ottimizzazione (problema numerico);
- errata scelta del criterio $J(\theta)$;
- classe di modelli non appropriata;
- set di dati non sufficientemente informativo.

Come già detto, spesso il punto cruciale è proprio la scelta della classe di modelli (ad esempio, per modelli lineari tempo-invarianti, la scelta dell'ordine del modello). Nella pratica, la procedura di identificazione viene tipicamente

iterata fino al soddisfacimento dei criteri di validazione, modificando ogni volta uno o più elementi del problema (vedi Figura 5.1).

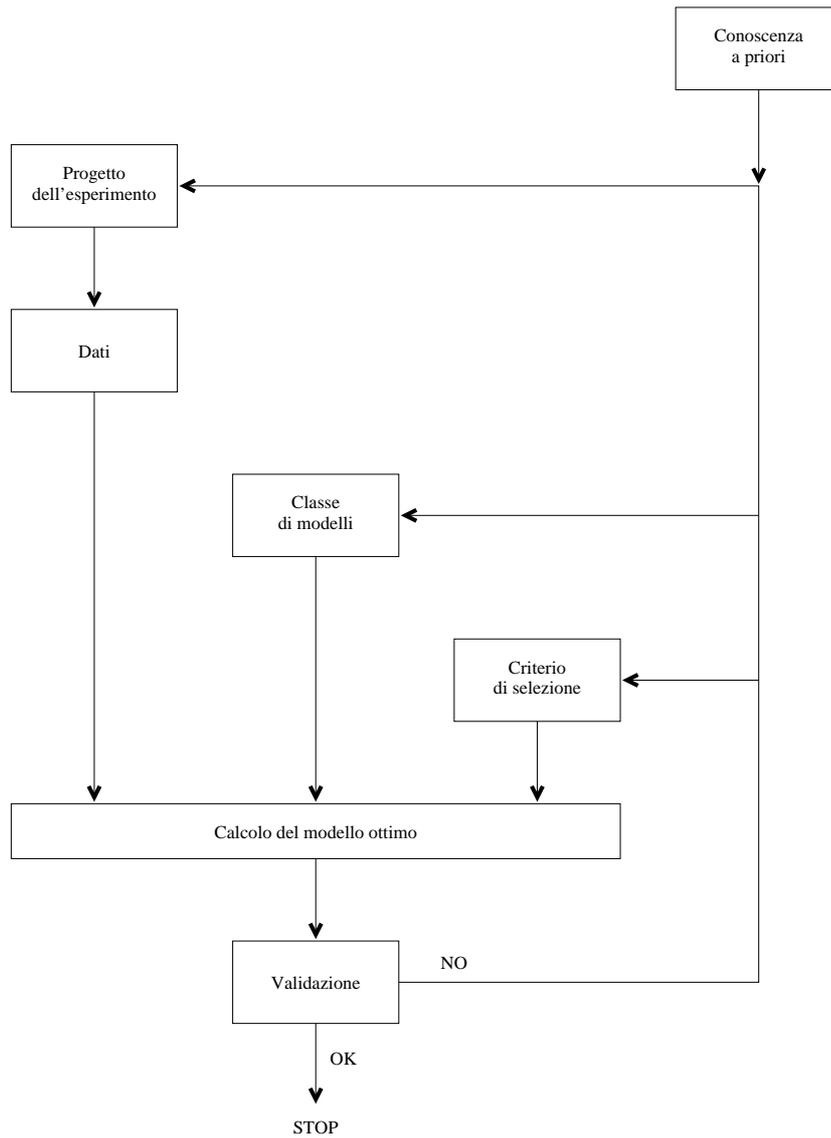


Figura 5.1. Procedura di identificazione.

5.2 Modelli lineari tempo-invarianti

Si parla di *identificazione parametrica* quando la classe di modelli può essere rappresentata tramite una struttura che permette di descrivere la dinamica ingresso-uscita del sistema e quella dell'errore di identificazione con un numero finito di valori numerici, detti appunto *parametri*. Una classe di modelli molto diffusa è quella dei modelli lineari tempo-invarianti (LTI):

$$y(t) = G(z)u(t) + H(z)e(t) \quad (5.2)$$

dove $u(t)$ rappresenta l'ingresso deterministico, $e(t)$ è un processo stocastico bianco, $G(z)$ e $H(z)$ sono opportune funzioni di trasferimento¹. Il termine $G(z)u(t)$ rappresenta la componente deterministica del modello e descrive la relazione tra l'ingresso $u(t)$ e l'uscita $y(t)$. Il termine $H(z)e(t)$ modella la componente stocastica presente nei dati $y(t)$ (dovuta, ad esempio, a errori di misura o a fenomeni non modellati dalla componente deterministica).

Il modello (5.2) è completamente specificato se sono note le risposte impulsive $\{g(k)\}_{k=0}^{\infty}$, $\{h(k)\}_{k=0}^{\infty}$ delle funzioni di trasferimento $G(z)$ e $H(z)$:

$$G(z) = \sum_{k=0}^{\infty} g(k)z^{-k}, \quad H(z) = \sum_{k=0}^{\infty} h(k)z^{-k}$$

e la densità di probabilità $f_e(e)$ del processo $e(t)$. Una scelta alternativa consiste nell'individuare un numero finito di parametri θ per descrivere il modello. In quest'ottica, la classe dei modelli LTI comprende tutti i modelli della forma

$$y(t) = G(z, \theta)u(t) + H(z, \theta)e(t)$$

dove $e(t)$ è un processo bianco con densità di probabilità $f_e(e, \theta)$, e $\theta \in \Theta \subset \mathbb{R}^d$ rappresenta il vettore dei parametri. Spesso si individua il processo $e(t)$ mediante due soli parametri, la media (di solito nulla) e la varianza σ_e^2 . Se il rumore è Gaussiano, ciò equivale a specificare la densità $f_e(e)$.

¹Anche in questo capitolo, con un abuso di notazione, si indica con z sia la variabile complessa argomento delle funzioni di trasferimento, sia l'operatore di shift temporale tale che, per una generica sequenza tempo-discreto $u(k)$, $z u(k) = u(k+1)$, $z^{-1} u(k) = u(k-1)$

Il modo più semplice di parametrizzare il modello è quello di scegliere θ come vettore dei coefficienti dei polinomi al numeratore e al denominatore delle funzioni razionali $G(z)$ e $H(z)$. Nel seguito si presentano le strutture di modelli LTI più diffuse.

5.2.1 Modelli ARX

Si supponga che il processo stocastico $y(t)$ sia generato dalla regressione lineare:

$$y(t) + a_1 y(t-1) + \dots + a_{n_a} y(t-n_a) = b_1 u(t-1) + \dots + b_{n_b} u(t-n_b) + e(t) \quad (5.3)$$

Se si definiscono i polinomi

$$A(z) = 1 + a_1 z^{-1} + \dots + a_{n_a} z^{-n_a}, \quad B(z) = b_1 z^{-1} + \dots + b_{n_b} z^{-n_b} \quad (5.4)$$

le funzioni di trasferimento del modello LTI sono

$$G(z, \theta) = \frac{B(z)}{A(z)}, \quad H(z, \theta) = \frac{1}{A(z)}.$$

In questo caso una scelta naturale per il vettore dei parametri è

$$\theta = [a_1 \dots a_{n_a} \ b_1 \dots b_{n_b}]'$$

e $d = n_a + n_b$. Un modello che rientra in questa struttura si dice di tipo ARX (Auto Regressive with eXogenous input). Si noti che da un punto di vista fisico si fa un'ipotesi forte: il rumore $e(t)$ viene filtrato da una funzione di trasferimento avente lo stesso denominatore della funzione di trasferimento $G(z)$ del canale deterministico.

Nel caso particolare in cui $n_a = 0$, si parla di modello FIR (Finite Impulse Response), poiché la funzione di trasferimento $G(z)$ vale

$$G(z) = b_1 z^{-1} + \dots + b_{n_b} z^{-n_b}$$

e quindi $g(k) = 0$, per ogni $k > n_b$.

5.2.2 Modelli ARMAX

Una struttura che fornisce più gradi di libertà (rispetto alla classe ARX) nella descrizione del rumore è la cosiddetta ARMAX (Auto Regressive Moving Average with eXogenous input). In questo caso si ipotizza un meccanismo di generazione di $y(t)$ secondo la regressione:

$$y(t) + a_1y(t-1) + \dots + a_{n_a}y(t-n_a) = b_1u(t-1) + \dots + b_{n_b}u(t-n_b) \\ + e(t) + c_1e(t-1) + \dots + c_{n_c}e(t-n_c)$$

Se si definisce il polinomio

$$C(z) = 1 + c_1z^{-1} + \dots + c_{n_c}z^{-n_c}$$

le funzioni di trasferimento valgono

$$G(z, \theta) = \frac{B(z)}{A(z)}, \quad H(z, \theta) = \frac{C(z)}{A(z)},$$

in cui $A(z)$ e $B(z)$ sono gli stessi polinomi definiti in (5.4). Il vettore dei parametri per un modello ARMAX è definito da

$$\theta = [a_1 \dots a_{n_a} \ b_1 \dots b_{n_b} \ c_1 \dots c_{n_c}]'$$

e $d = n_a + n_b + n_c$. Si noti che si suppone che i polinomi $A(z)$ e $C(z)$ siano monici, e l'entità della componente stocastica $H(z)e(t)$ è modellata mediante la varianza σ_e^2 del processo $e(t)$.

5.2.3 Modelli OE

Anche nel caso dei modelli ARMAX, le funzioni di trasferimento $G(z)$ e $H(z)$ hanno lo stesso denominatore $A(z)$. Se questa ipotesi non è realistica, si deve ricorrere ad altre classi di modelli. Ad esempio, si può assumere che $y(t)$ sia generato secondo le equazioni:

$$y(t) = w(t) + e(t) \\ w(t) + f_1w(t-1) + \dots + f_{n_f}w(t-n_f) = b_1u(t-1) + \dots + b_{n_b}u(t-n_b).$$

Se si definisce il polinomio

$$F(z) = 1 + f_1 z^{-1} + \dots + f_{n_f} z^{-n_f}$$

le funzioni di trasferimento del modello LTI valgono

$$G(z, \theta) = \frac{B(z)}{F(z)}, \quad H(z, \theta) = 1.$$

Un modello di questo tipo viene chiamato OE (Output Error), poiché assume che il rumore bianco $e(t)$ agisca (in maniera additiva) direttamente sull'uscita. In questo caso il vettore dei parametri è

$$\theta = [b_1 \dots b_{n_b} f_1 \dots f_{n_f}]'$$

ed ha dimensione pari a $d = n_b + n_f$.

5.2.4 Modelli BJ

Un ulteriore grado di libertà si può ottenere ipotizzando che le funzioni di trasferimento $G(z)$ e $H(z)$ siano completamente indipendenti, ossia

$$G(z, \theta) = \frac{B(z)}{F(z)}, \quad H(z, \theta) = \frac{C(z)}{D(z)},$$

dove

$$D(z) = 1 + d_1 z^{-1} + \dots + d_{n_d} z^{-n_d}$$

e $B(z)$, $C(z)$, $F(z)$ sono gli stessi polinomi definiti in precedenza. Modelli che posseggono questa struttura vengono detti BJ (Box-Jenkins) In questo caso il vettore dei parametri vale

$$\theta = [b_1 \dots b_{n_b} f_1 \dots f_{n_f} c_1 \dots c_{n_c} d_1 \dots d_{n_d}]'$$

ed ha dimensione $d = n_b + n_c + n_d + n_f$.

Le strutture di modelli presentate in precedenza possono essere viste come casi particolari di una famiglia generale di modelli dalla forma

$$A(z) = \frac{B(z)}{F(z)}u(t) + \frac{C(z)}{D(z)}e(t).$$

Un ulteriore parametro da fissare è il ritardo r tra l'ingresso $u(t)$ e l'uscita $y(t)$. Se questo dato è noto a priori, è importante incorporare questa informazione nella struttura, che quindi diventa

$$A(z) = \frac{B(z)}{F(z)} z^{-r} u(t) + \frac{C(z)}{D(z)} e(t).$$

Da un punto di vista pratico, è possibile includere il ritardo nel polinomio $B(z)$, fissando a zero i primi r coefficienti. Infatti:

$$B(z)z^{-r} = b_1 z^{-r-1} + \dots + b_{n_b} z^{-r-n_b}.$$

5.3 Criterio di scelta del modello

Una volta scelta la classe di modelli, occorre definire un criterio per selezionare il miglior modello all'interno di essa, sulla base dei dati \mathcal{Z}^N . L'idea di base è quella di confrontare le uscite $\{y(t)\}$ del sistema incognito con quelle $\{y_m(t)\}$ fornite dal modello, a parità di ingresso $\{u(t)\}$. Analizziamo più in dettaglio l'uscita y_m del modello. Essa è costituita, come detto, da una parte deterministica y_d e da una parte stocastica y_s :

$$y_m(t) = \underbrace{G(z, \theta)u(t)}_{y_d(t)} + \underbrace{H(z, \theta)e(t)}_{y_s(t)}.$$

Naturalmente $e(t)$ è ignoto e occorre fare una "previsione" di ciò che sarà $y(t)$ (e in particolare $y_s(t)$) in base al modello scelto e ai dati disponibili. Una scelta naturale è quella di far coincidere $y_m(t)$ con la *predizione* del valore assunto dall'uscita all'istante t sulla base dei dati \mathcal{Z}^{t-1} raccolti fino all'istante precedente, ovvero:

$$y_m(t) = \hat{y}(t|t-1).$$

Se ad esempio si vuole scegliere $y_m(t)$ in base alla teoria della stima Bayesiana a minimo errore quadratico medio, è noto che:

$$y_m(t) = \mathbf{E} [y(t)|\mathcal{Z}^{t-1}],$$

il cui calcolo richiede però la conoscenza della densità di probabilità del processo $y(t)$.

In generale, dato un modello $\mathcal{M}(\theta)$, occorre costruire un predittore dell'uscita del modello $\hat{y}(t|t-1)$, per poter formare l'errore di predizione:

$$\varepsilon(t, \theta) = y(t) - \hat{y}(t|t-1).$$

I metodi di identificazione parametrica basati sull'errore di predizione minimizzano tipicamente un funzionale di costo

$$J(\varepsilon(t, \theta)) = V_N(\theta, \mathcal{Z}^N)$$

dove l'espressione a destra dell'uguale evidenzia la dipendenza del funzionale sia dai parametri incogniti θ che dai dati disponibili \mathcal{Z}^N . Il valore ottimo dei parametri sarà quindi dato da:

$$\hat{\theta}_N^* = \hat{\theta}_N(\mathcal{Z}^N) = \arg \min_{\theta \in \Theta} V_N(\theta, \mathcal{Z}^N). \quad (5.5)$$

Nel seguito, vediamo come si può ricavare semplicemente la predizione lineare a minimo errore quadratico medio (LMEQM) $\hat{y}(t|t-1)$, per la classe di modelli LTI esaminati nella sezione 5.2:

$$y(t) = G(z)u(t) + H(z)e(t) = G(z)u(t) + v(t)$$

dove si è posto $v(t) = H(z)e(t)$. Essendo:

$$H(z) = 1 + \sum_{k=1}^{\infty} h(k)z^{-k},$$

possiamo riscrivere

$$v(t) = e(t) + \sum_{k=1}^{\infty} h(k)e(t-k).$$

Se si desidera stimare $v(t)$ sulla base dei dati fino all'istante $t-1$, si ha:

$$\begin{aligned} \hat{v}(t|t-1) &= \sum_{k=1}^{\infty} h(k)e(t-k) \\ &= (H(z) - 1)e(t) \\ &= \frac{H(z) - 1}{H(z)}v(t) \\ &= (1 - H^{-1}(z))v(t) \end{aligned}$$

sotto l'ipotesi che $H(z)$ sia invertibile, cioè che non abbia zeri instabili. Ne segue che:

$$\begin{aligned} \boxed{\hat{y}(t|t-1)} &= G(z)u(t) + (1 - H^{-1}(z))v(t) \\ &= G(z)u(t) + (1 - H^{-1}(z))(y(t) - G(z)u(t)) \\ &= \boxed{(1 - H^{-1}(z))y(t) + H^{-1}(z)G(z)u(t)}. \end{aligned} \quad (5.6)$$

Quindi, l'errore di predizione vale:

$$\begin{aligned} \boxed{\varepsilon(t, \theta)} &= y(t) - \hat{y}(t|t-1) = H^{-1}(z)y(t) - H^{-1}(z)G(z)u(t) \\ &= \boxed{H^{-1}(z)(y(t) - G(z)u(t))}. \end{aligned} \quad (5.7)$$

Osservazione 5.1. Si noti che si poteva giungere alla stessa espressione per $\hat{v}(t|t-1)$ (e quindi per $\hat{y}(t|t-1)$) applicando la teoria della predizione LME-QM di processi stocastici stazionari. In tal senso, l'ipotesi che $H(z)$ sia invertibile significa semplicemente che $H(z)$ è il fattore canonico spettrale della parte stocastica $v(t)$, e quindi non comporta alcuna perdita di generalità.

Nel seguito verrà analizzato l'errore di predizione per alcune classi di modelli notevoli.

ARX

Nel caso di un modello ARX si è visto che le funzioni di trasferimento valgono:

$$G(z, \theta) = \frac{B(z)}{A(z)}, \quad H(z, \theta) = \frac{1}{A(z)},$$

per cui si ha:

$$H^{-1}(z) = A(z), \quad H^{-1}(z)G(z) = B(z).$$

Quindi l'errore di predizione assume la forma:

$$\varepsilon(t, \theta) = A(z)y(t) - B(z)u(t).$$

Una proprietà importante dei modelli ARX è che l'errore di predizione è una funzione lineare dei parametri. Infatti, ricordando la definizione del vettore dei parametri θ e definendo il *regressore*

$$\varphi(t) = [-y(t-1) \ \dots \ -y(t-n_a) \ u(t-1) \ \dots \ u(t-n_b)]',$$

l'equazione (5.3) può essere riscritta come:

$$y(t) = \varphi'(t)\theta + e(t).$$

Di conseguenza la predizione $\hat{y}(t|t-1)$ vale

$$\hat{y}(t|t-1) = \varphi'(t)\theta$$

e l'errore di predizione diventa:

$$\varepsilon(t, \theta) = y(t) - \varphi'(t)\theta.$$

ARMAX

Nel caso di un modello ARMAX le funzioni di trasferimento valgono:

$$G(z, \theta) = \frac{B(z)}{A(z)}, \quad H(z, \theta) = \frac{C(z)}{A(z)},$$

per cui si ha:

$$H^{-1}(z) = \frac{A(z)}{C(z)}, \quad H^{-1}(z)G(z) = \frac{B(z)}{C(z)}.$$

In accordo all'equazione (5.6), la predizione dell'uscita vale:

$$\hat{y}(t|t-1) = \left(1 - \frac{A(z)}{C(z)}\right) y(t) + \frac{B(z)}{C(z)} u(t)$$

ossia

$$C(z)\hat{y}(t|t-1) = B(z)u(t) + (C(z) - A(z))y(t).$$

Sommando il termine $(1 - C(z))\hat{y}(t|t-1)$ ad ambo i membri della precedente equazione otteniamo:

$$\hat{y}(t|t-1) = B(z)u(t) + (1 - A(z))y(t) + (1 - C(z))(\hat{y}(t|t-1) - y(t)). \quad (5.8)$$

Ricordando la definizione dei polinomi A , B e C , la (5.8) corrisponde alla ricorsione:

$$\begin{aligned} \hat{y}(t|t-1) &= -a_1 y(t-1) - \dots - a_{n_a} y(t-n_a) + b_1 u(t-1) + \dots + b_{n_b} u(t-n_b) \\ &\quad + c_1 \varepsilon(t-1, \theta) + \dots + c_{n_c} \varepsilon(t-n_c, \theta) \\ &= \varphi'(t, \theta)\theta \end{aligned} \quad (5.9)$$

ove si è definito lo *pseudo-regressore* $\varphi(t, \theta)$ come:

$$\varphi(t, \theta) = [-y(t-1) \cdots -y(t-n_a) \ u(t-1) \cdots u(t-n_b) \ \varepsilon(t-1, \theta) \cdots \varepsilon(t-n_c, \theta)]'$$

A questo punto, l'errore di predizione assume la forma:

$$\varepsilon(t, \theta) = y(t) - \varphi'(t, \theta)\theta.$$

Si noti che nel caso ARMAX lo pseudo-regressore dipende anch'esso dal vettore dei parametri θ , per cui si dice che in questo caso l'errore si ottiene tramite una regressione *pseudo-lineare* (o, in altre parole, $\varepsilon(t, \theta)$ è una funzione *non lineare* di θ).

OE

Nel caso di un modello OE le funzioni di trasferimento valgono:

$$G(z, \theta) = \frac{B(z)}{F(z)}, \quad H(z, \theta) = 1,$$

per cui si ha:

$$H^{-1}(z) = 1, \quad H^{-1}(z)G(z) = \frac{B(z)}{F(z)}.$$

Allora, la predizione di $y(t)$ è data da:

$$\hat{y}(t|t-1) = (1-1)y(t) + \frac{B(z)}{F(z)}u(t),$$

Ponendo

$$w(t, \theta) \triangleq \frac{B(z)}{F(z)}u(t),$$

si ha

$$w(t, \theta) = -f_1 w(t-1, \theta) - \cdots - f_{n_f} w(t-n_f, \theta) + b_1 u(t-1) + \cdots + b_{n_b} u(t-n_b)$$

per cui la predizione soddisfa la regressione pseudo-lineare:

$$\begin{aligned} \hat{y}(t|t-1) &= -f_1 w(t-1, \theta) - \cdots - f_{n_f} w(t-n_f, \theta) + b_1 u(t-1) + \cdots + b_{n_b} u(t-n_b) \\ &= \varphi'(t, \theta)\theta. \end{aligned}$$

(5.10)

In questo caso si è definito lo *pseudo-regressore* come:

$$\varphi(t, \theta) = [-w(t-1, \theta) - \dots - w(t-n_f, \theta) \ u(t-1) \dots u(t-n_b)]'.$$

Perciò l'errore di predizione vale:

$$\varepsilon(t, \theta) = y(t) - \varphi'(t, \theta)\theta.$$

ed è *non lineare* in θ .

Si noti che $w(t-k, \theta) = \hat{y}(t-k|t-k-1)$ e quindi lo pseudo-regressore può essere calcolato utilizzando i campioni precedenti della predizione.

Osservazione 5.2. Si osservi che in generale, la predizione $\hat{y}(t|t-1)$ è esatta se i dati a disposizione all'istante $t-1$ sono infiniti, ovvero $\{u(k), y(k)\}_{k=-\infty}^{t-1}$. Ovviamente, si può approssimare $\hat{y}(t|t-1)$ considerando $u(t)$ e $y(t)$ nulli prima di un assegnato istante iniziale (ad es., per $t < 0$), ma in questo caso la predizione è soltanto subottima. Fa eccezione in tal senso il caso dei modelli ARX in cui la predizione dipende solo da n_a campioni passati dell'uscita e da n_b campioni passati dell'ingresso.

5.3.1 Scelta del funzionale di costo

Come già detto, il vettore dei parametri si ottiene minimizzando un funzionale di costo $V_N(\theta, \mathcal{Z}^N)$. Una scelta molto generale è del tipo:

$$V_N(\theta, \mathcal{Z}^N) = \frac{1}{N} \sum_{t=1}^N \ell(L(z)\varepsilon(t, \theta))$$

dove

- $\ell(\cdot)$ è una funzione scalare, tipicamente positiva. Una scelta standard è

$$\ell(\varepsilon) = \frac{1}{2}\varepsilon^2$$

che corrisponde a calcolare la stima di θ ai minimi quadrati.

- $L(z)$ è un filtro dell'errore di predizione, che consente di evidenziare frequenze di particolare interesse, ridurre l'influenza di disturbi ad alta frequenza, ecc. Per modelli LTI, l'utilizzo di $L(z)$ equivale a filtrare i dati:

$$\begin{aligned} L(z)\varepsilon(t, \theta) &= H^{-1}(z) \underbrace{(L(z)y(t))}_{y_f(t)} - G(z) \underbrace{L(z)u(t)}_{u_f(t)} \\ &= H^{-1}(z)L(z)(y(t) - G(z)u(t)) \end{aligned}$$

per cui l'effetto del filtraggio tramite $L(z)$ equivale di fatto a modificare la funzione di trasferimento del canale stocastico del modello da $H(z)$ a $\frac{H(z)}{L(z)}$.

5.4 Selezione del modello ottimo

In questo paragrafo viene affrontato il problema del calcolo del vettore dei parametri "ottimo" $\hat{\theta}_N^*$, ossia il problema di ottimizzazione (5.5), una volta scelta una classe di modelli e definito un funzionale di costo. Iniziamo col caso più semplice e per il quale esiste una soluzione in forma chiusa.

5.4.1 Regressione lineare e minimi quadrati

Si è visto che per un modello ARX l'errore di predizione vale

$$\varepsilon(t, \theta) = y(t) - \varphi'(t)\theta.$$

Se si sceglie il funzionale di costo corrispondente a $\ell(\varepsilon) = \frac{1}{2}\varepsilon^2$ e $L(z) = 1$:

$$V_N(\theta, \mathcal{Z}^N) = \frac{1}{2N} \sum_{t=1}^N (y(t) - \varphi'(t)\theta)^2$$

si ha il criterio dei minimi quadrati

$$\hat{\theta}_N^* = \arg \min_{\theta} V_N(\theta, \mathcal{Z}^N). \quad (5.11)$$

Si noti che il funzionale di costo vale:

$$V_N(\theta, \mathcal{Z}^N) = \frac{1}{2N} \left(\sum_{t=1}^N y(t)^2 - 2 \sum_{t=1}^N \theta' \varphi(t) y(t) + \sum_{t=1}^N \theta' \varphi(t) \varphi'(t) \theta \right) \quad (5.12)$$

Per il calcolo del vettore dei parametri ottimo, imponiamo che si annulli il gradiente di $V_N(\theta, \mathcal{Z}^N)$:

$$\left. \frac{d}{d\theta} V_N(\theta, \mathcal{Z}^N) = \frac{1}{2N} \sum_{t=1}^N (-2\varphi'(t)y(t) + 2\varphi(t)\varphi'(t)\theta) \right|_{\theta=\hat{\theta}_N^*} = 0,$$

da cui si ottengono le cosiddette *equazioni normali*:

$$\left(\frac{1}{N} \sum_{t=1}^N \varphi(t)\varphi'(t) \right) \theta = \frac{1}{N} \sum_{t=1}^N \varphi(t)y(t). \quad (5.13)$$

Nel caso in cui la matrice

$$R(N) \triangleq \frac{1}{N} \sum_{t=1}^N \varphi(t)\varphi'(t)$$

sia invertibile, le equazioni normali (5.13) ammettono un'unica soluzione, nota come *stima parametrica ai minimi quadrati*:

$$\hat{\theta}_N^{LS} = \left(\frac{1}{N} \sum_{t=1}^N \varphi(t)\varphi'(t) \right)^{-1} \frac{1}{N} \sum_{t=1}^N \varphi(t)y(t). \quad (5.14)$$

La matrice $R(N)$ riveste un ruolo fondamentale nell'ambito della stima dei parametri di un modello ARX. Di seguito sono discusse alcune sue proprietà.

- $R(N)$ è l'Hessiano di $V_N(\theta, \mathcal{Z}^N)$, ovvero $\frac{d^2}{d\theta^2} V_N(\theta, \mathcal{Z}^N)$. Dal momento che il funzionale di costo è una somma di quadrati, allora sicuramente $R(N) \geq 0$. In particolare, se $R(N) > 0$ allora $R(N)$ è invertibile e la soluzione di (5.11) è unica, altrimenti il problema di ottimizzazione ammette infinite soluzioni (condizione di *non identificabilità*).
- $R(N)$ è la somma di matrici diadiche $\varphi(t)\varphi'(t)$ di rango 1. Ne segue che $\text{rank}R(N) \leq N$. Se $d = n_a + n_b$ è il numero di parametri da stimare, affinché $R(N)$ sia invertibile occorre che $N \geq d$ (ovvero che il

numero dei dati ingresso-uscita non sia inferiore al numero dei parametri da stimare). Si noti che questa è una condizione necessaria ma non sufficiente.

- Si ponga $P(N) = NR(N)$. La matrice $P(N)$ può essere calcolata tramite la ricorrenza:

$$P(t) = P(t-1) + \varphi(t)\varphi'(t)$$

e naturalmente man mano che si aggiungono termini $\varphi(t)\varphi'(t)$ è più probabile che $R(N)$ diventi invertibile. Questo dipende ovviamente dalla scelta dell'ingresso (ad esempio, si provi a calcolare $P(t)$ nel caso di un modello FIR di ordine 2 con un ingresso costante... che cosa succede?).

- Gli elementi di $R(N)$ sono del tipo $\frac{1}{N} \sum_{t=1}^N y(t-i)y(t-j)$, $\frac{1}{N} \sum_{t=1}^N y(t-i)u(t-j)$ e $\frac{1}{N} \sum_{t=1}^N u(t-i)u(t-j)$, cioè sono stime campionarie delle matrici di covarianza $R_y(\tau)$, $R_u(\tau)$ e $R_y u(\tau)$.

Vogliamo ora esaminare la consistenza della stima ai minimi quadrati. Si supponga che i dati siano generati da una regressione lineare con vettore dei parametri θ_0 :

$$y(t) = \varphi'(t)\theta_0 + v_0(t).$$

Per ora non facciamo alcuna ipotesi sul rumore v_0 . Allora

$$\begin{aligned} \hat{\theta}_N^{LS} &= \left(\frac{1}{N} \sum_{t=1}^N \varphi(t)\varphi'(t) \right)^{-1} \frac{1}{N} \sum_{t=1}^N \varphi(t)(\varphi'(t)\theta_0 + v_0(t)) \\ &= R(N)^{-1} \left(R(N)\theta_0 + \frac{1}{N} \sum_{t=1}^N \varphi(t)v_0(t) \right) \\ &= \theta_0 + R(N)^{-1} \frac{1}{N} \sum_{t=1}^N \varphi(t)v_0(t) \end{aligned}$$

Cosa accade all'errore di stima quando il numero di dati N tende all'infinito? Supponiamo che $\varphi(t)$ e $v_0(t)$ siano stazionari e per loro valga il teorema di

ergodicità. Allora:

$$\lim_{N \rightarrow \infty} (\hat{\theta}_N^{LS} - \theta_0) = \lim_{N \rightarrow \infty} R(N)^{-1} \frac{1}{N} \sum_{t=1}^N \varphi(t) v_0(t) \triangleq (R^*)^{-1} f^*$$

in cui si è posto

$$R^* = \lim_{N \rightarrow \infty} R(N) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N \mathbf{E}[\varphi(t) \varphi'(t)],$$

$$f^* = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N \varphi(t) v_0(t) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N \mathbf{E}[\varphi(t) v_0(t)].$$

Per cui $\hat{\theta}_N^{LS} \rightarrow \theta_0$ quando $N \rightarrow \infty$, ossia la stima ai minimi quadrati è convergente, se:

- i) R^* è non singolare (dipende dalla scelta dell'ingresso);
- ii) $f^* = 0$ (dipende dalle caratteristiche di v_0). Ciò accade ad esempio se:
 - $v_0(t)$ è un rumore bianco a media nulla; oppure
 - l'ingresso $u(t)$ è indipendente da v_0 e $n_a = 0$ (modello FIR).

Se $n_a > 0$ e il rumore v_0 non è bianco, allora in generale $f^* \neq 0$, per cui la stima LS non è consistente.

5.4.2 Soluzioni numeriche del problema di ottimizzazione

Nel caso generale, il minimo del funzionale di costo

$$V_N(\theta, \mathcal{Z}^N) = \frac{1}{N} \sum_{t=1}^N \ell(\varepsilon(t, \theta))$$

non può essere calcolato analiticamente. Inoltre, il funzionale sarà in generale una funzione non convessa dei parametri, per cui possono esserci minimi locali. In questo caso si ricorre a metodi numerici che calcolano la soluzione mediante una procedura *iterativa*, secondo lo schema:

$$\hat{\theta}^{(i+1)} = \hat{\theta}^{(i)} + \alpha^{(i)} f^{(i)}.$$

Il termine $f^{(i)}$ rappresenta la *direzione di ricerca* ed è scelta in base alle informazioni sul funzionale $V_N(\theta, \mathcal{Z}^N)$ disponibili all' i -esimo passo dell'iterazione. Il termine $\alpha^{(i)}$ è determinato in modo da avere un decremento nel funzionale, ovvero in modo che $V_N(\hat{\theta}^{(i+1)}, \mathcal{Z}^N) < V_N(\hat{\theta}^{(i)}, \mathcal{Z}^N)$.

Si distinguono tre classi di metodi iterativi a seconda di come viene scelta la direzione $f^{(i)}$:

1. usando solo valori di $V_N(\hat{\theta}^{(i)}, \mathcal{Z}^N)$;
2. usando valori di $V_N(\hat{\theta}^{(i)}, \mathcal{Z}^N)$ e del gradiente $V'_N(\hat{\theta}^{(i)}, \mathcal{Z}^N)$;
3. usando valori di $V_N(\hat{\theta}^{(i)}, \mathcal{Z}^N)$, di $V'_N(\hat{\theta}^{(i)}, \mathcal{Z}^N)$ e dell'Hessiano $V''_N(\hat{\theta}^{(i)}, \mathcal{Z}^N)$.

Un tipico esempio di algoritmo appartenente al terzo gruppo è l'algoritmo di Newton, in cui

$$f^{(i)} = - \left[V''_N(\hat{\theta}^{(i)}, \mathcal{Z}^N) \right]^{-1} V'_N(\hat{\theta}^{(i)}, \mathcal{Z}^N)$$

e che garantisce la convergenza in un passo (per $\alpha = 1$) nel caso in cui il funzionale sia quadratico e convesso in θ .

Se la funzione peso $\ell(\cdot)$ è scelta in modo che:

$$\ell(\varepsilon(t, \theta)) = \frac{1}{2} \varepsilon^2(t, \theta)$$

si ha il problema dei *minimi quadrati non lineari*:

$$\min_{\theta} \frac{1}{2N} \sum_{t=1}^N \varepsilon^2(t, \theta).$$

In questo caso, il gradiente del funzionale di costo vale:

$$V'_N(\theta, \mathcal{Z}^N) = -\frac{1}{N} \sum_{t=1}^N \psi(t, \theta) \varepsilon(t, \theta) \quad (5.15)$$

dove $\psi(t, \theta) = -\frac{d}{d\theta} \varepsilon(t, \theta)$. Una famiglia classica di algoritmi iterativi per il calcolo del minimo ha la seguente struttura:

$$\hat{\theta}^{(i+1)} = \hat{\theta}^{(i)} - \mu_N^{(i)} \left[R_N^{(i)} \right]^{-1} V'_N(\hat{\theta}^{(i)}, \mathcal{Z}^N),$$

dove

- $R_N^{(i)}$ è una matrice che modifica la direzione di ricerca;
- $\mu_N^{(i)}$ rappresenta l'entità del passo di ricerca, scelto in modo che $V_N(\hat{\theta}^{(i+1)}, \mathcal{Z}^N) < V_N(\hat{\theta}^{(i)}, \mathcal{Z}^N)$.

Nel caso in cui $R_N^{(i)}$ corrisponde alla matrice identità si ottiene il metodo classico del gradiente, che risulta essere abbastanza inefficiente in prossimità del punto di minimo. Migliori prestazioni sono fornite dal metodo di Newton che, come detto, corrisponde a scegliere:

$$R_N^{(i)} = V_N''(\hat{\theta}^{(i)}, \mathcal{Z}^N).$$

Si noti che dalla (5.15) segue che:

$$V_N''(\theta, \mathcal{Z}^N) = \frac{1}{N} \sum_{t=1}^N \psi(t, \theta) \psi'(t, \theta) - \frac{1}{N} \sum_{t=1}^N \left(\frac{d}{d\theta} \psi(t, \theta) \right) \varepsilon(t, \theta),$$

dove $\frac{d}{d\theta} \psi(t, \theta)$ è l'Hessiano di $\varepsilon(t, \theta)$ (matrice $n \times n$).

Il calcolo di $\frac{d}{d\theta} \psi(t, \theta)$ può essere costoso. In molti casi esso può essere evitato poiché è necessario avere una buona approssimazione di $V_N''(\theta, \mathcal{Z}^N)$ solo nei pressi del minimo (lontano dal minimo il costo difficilmente può essere approssimato da una funzione quadratica e quindi metodi del gradiente e di Newton hanno prestazioni simili). Ma se il minimo si ha in corrispondenza di un θ_0 per cui l'errore di predizione $\varepsilon(t, \theta_0)$ è un processo bianco, allora

$$\mathbf{E} \left[\left(\frac{d}{d\theta} \psi(t, \theta_0) \right) \varepsilon(t, \theta_0) \right] = 0$$

e quindi:

$$V_N''(\theta, \mathcal{Z}^N) \simeq \frac{1}{N} \sum_{t=1}^N \psi(t, \theta) \psi'(t, \theta).$$

La precedente approssimazione di $V_N''(\theta, \mathcal{Z}^N)$ dà luogo al cosiddetto metodo di Gauss-Newton, che corrisponde alla scelta:

$$R_N^{(i)} = \frac{1}{N} \sum_{t=1}^N \psi(t, \hat{\theta}^{(i)}) \psi'(t, \hat{\theta}^{(i)})$$

Nel caso in cui la matrice $R_N^{(i)}$ sia vicina ad essere singolare, si utilizzeranno tecniche di regolarizzazione, quale ad esempio quella di Levenberg-Marquardt, in cui si pone:

$$R_N^{(i)} = \frac{1}{N} \sum_{t=1}^N \psi(t, \hat{\theta}^{(i)}) \psi'(t, \hat{\theta}^{(i)}) + \lambda I$$

dove λ è uno scalare usato per controllare la convergenza (in questo caso $\mu_N^{(i)} = 1$). Si osservi che quando $\lambda \rightarrow 0$ si ottiene il metodo di Gauss-Newton, mentre per $\lambda \rightarrow \infty$ si ottiene il metodo del gradiente.

I metodi sopra esaminati richiedono il calcolo di

$$\psi(t, \theta) = -\frac{d}{d\theta} \varepsilon(t, \theta) = \frac{d}{d\theta} \hat{y}(t|t-1, \theta).$$

Nel caso di modelli ARMAX si è visto che:

$$C(z) \hat{y}(t|t-1) = B(z)u(t) + (C(z) - A(z))y(t)$$

da cui:

$$\begin{aligned} C(z) \frac{\partial}{\partial a_k} \hat{y}(t|t-1) &= -y(t-k), & k = 1, \dots, n_a \\ C(z) \frac{\partial}{\partial b_k} \hat{y}(t|t-1) &= u(t-k), & k = 1, \dots, n_b \\ C(z) \frac{\partial}{\partial c_k} \hat{y}(t|t-1) + z^{-k} \hat{y}(t|t-1) &= y(t-k), & k = 1, \dots, n_c \end{aligned}$$

Le tre precedenti famiglie di equazioni possono essere espresse sinteticamente come

$$C(z) \frac{d}{d\theta} \hat{y}(t|t-1) = \varphi(t, \theta),$$

ovvero

$$C(z) \psi(t, \theta) = \varphi(t, \theta).$$

Quindi $\psi(t, \theta)$ si ottiene filtrando lo pseudo-regressore $\varphi(t, \theta)$ mediante il filtro $\frac{1}{C(z)}$.

Analogamente a quanto appena visto per i modelli ARMAX, nel caso di modelli OE il predittore soddisfa:

$$F(z) \hat{y}(t|t-1) = B(z)u(t),$$

da cui

$$F(z) \frac{\partial}{\partial f_k} \hat{y}(t|t-1) + z^{-k} \hat{y}(t|t-1) = 0, \quad k = 1, \dots, n_f$$

$$C(z) \frac{\partial}{\partial b_k} \hat{y}(t|t-1) = u(t-k), \quad k = 1, \dots, n_b$$

ovvero

$$F(z)\psi(t, \theta) = \varphi(t, \theta).$$

Ovviamente, la procedura di filtraggio dello pseudo-regressore necessita di una condizione iniziale e non è detto che la scelta $\varphi(0, \theta) = 0$ sia quella migliore. Una possibilità alternativa consiste nel porre:

$$\varphi(0, \theta) = \eta$$

ed includere η nel vettore θ dei parametri da stimare.

Capitolo 6

Valutazione della qualità dei modelli

Come evidenziato nella procedura di identificazione descritta nel paragrafo 5.1, una volta che un modello è stato identificato, è necessario *validarlo*, cioè verificare se esso è coerente con le informazioni disponibili sul sistema (conoscenza a priori, dati, ecc.) e in ultima analisi se può essere effettivamente impiegato per gli scopi per i quali è stato costruito. La qualità di un modello matematico dipenderà quindi anche dalla specifica applicazione considerata e dall'uso che si intende fare del modello stesso. Ad esempio, un modello utilizzato per progettare un sistema di controllo in retroazione dovrà avere requisiti differenti da quelli di un modello impiegato per predire l'andamento futuro dell'uscita del sistema in certe condizioni operative.

Ciò detto, esistono comunque alcune tecniche generali per la valutazione della qualità dei modelli, che possono essere applicate a prescindere dal contesto applicativo. Alcune di esse sono brevemente descritte in questo capitolo.

6.1 Confronto tra classi di modelli

Nello schema generale della procedura di identificazione illustrato in Figura 5.1, si può vedere che uno degli elementi da riconsiderare in fase di validazione

di un modello è la scelta della classe di modelli. Ovviamente, qualora siano disponibili informazioni a priori sul sistema da identificare, tale conoscenza deve essere utilizzata nella scelta della classe. In particolare, occorre scegliere:

- la tipologia di modelli (ad es., lineare o non lineare, ingresso/uscita o variabili di stato, basati su leggi fisiche o black-box, ecc.);
- la struttura del modello (ad es., nel caso di modelli lineari tempo-invarianti: ARX, ARMAX, OE, ecc.)
- il numero di parametri da stimare.

Vi sono diversi aspetti da tener presente quando si effettua la scelta della classe di modelli. In generale, la qualità di un modello può essere misurata quantitativamente mediante un opportuno indice J . Tipicamente, questi indici sono costituiti da due contributi:

$$J = J_B + J_V$$

dove J_B rappresenta un termine di “*bias*”, dipendente dalla classe di modelli scelta ma non dai dati, mentre J_V è un termine di “*varianza*” che dipende dai dati e dal numero di parametri da stimare. In sintesi, il termine J_B tiene conto della “distanza” esistente tra la classe di modelli scelta e il sistema che ha generato i dati, mentre il termine J_V quantifica la varianza dell’errore di stima parametrica $\theta - \hat{\theta}_N$, la quale tenderà a crescere col numero d di parametri da stimare (a parità del numero N di dati disponibili). Ne consegue che un modello ricco e flessibile (con struttura complessa e molti parametri da stimare) darà luogo ad un basso errore di bias ma ad un’elevata varianza; viceversa, un modello semplice privilegia J_V rispetto a J_B .

Un altro aspetto di cui tener conto riguarda il carico computazionale necessario per il calcolo del modello: un modello molto complicato e/o un funzionale di costo $V_N(\theta, \mathcal{Z}^N)$ di forma complessa possono rendere difficile il calcolo della stima del vettore dei parametri ottimo $\hat{\theta}_N$ (costo computazionale elevato). Infine, come detto, è importante considerare anche l’uso che si prevede fare

del modello: modelli complessi possono essere difficili da simulare, il progetto del sistema di controllo può diventare molto complicato, ecc.

In definitiva, la scelta finale della classe di modelli sarà il risultato di un compromesso tra tutti gli aspetti fin qui considerati. La valutazione di tali aspetti può essere effettuata mediante diverse tecniche e considerazioni, che comprendono:

- considerazioni basate sulla conoscenza a priori del sistema;
- analisi preliminare dei dati (a priori, prima del calcolo della stima);
- confronti tra diverse strutture;
- validazione del modello (a posteriori, sulla base dei dati).

6.1.1 Conoscenza a priori

Le considerazioni basate sulla conoscenza a priori sono molto legate allo specifico problema che si sta affrontando. Occorre scegliere se privilegiare gli aspetti fisici (meno parametri, modelli più legati al sistema in esame) o selezionare una struttura black-box (più semplice dal punto di vista algoritmico). Occorre anche decidere se considerare o meno la presenza di eventuali non-linearità (spesso note a priori almeno approssimativamente), o effettuare trasformazioni sui dati. Comunque, è buona norma partire sempre da strutture molto semplici (principio di *parsimonia*).

6.1.2 Analisi dei dati

Si può cercare di stimare l'ordine di un sistema lineare (e quindi della classe di modelli da usare) in diversi modi:

- utilizzando una stima non-parametrica in frequenza, ad esempio tramite analisi spettrale dei dati, per individuare picchi di risonanza, banda passante, pendenza ad alte frequenze, ecc.;

- valutando se è opportuno inserire nel modello la dipendenza da una nuova variabile $y(t - n - 1)$ o da un possibile disturbo misurato $w(t)$, analizzandone la correlazione con $y(t)$ (se questa è trascurabile non è opportuno aumentare l'ordine del modello);
- valutando il numero di condizionamento della matrice di informazione $\frac{1}{N} \sum_{t=1}^N \psi(t, \theta) \psi'(t, \theta)$ durante il calcolo della stima mediante metodi di Gauss-Newton (vedi paragrafo 5.4.2): se tale matrice è vicina ad essere singolare, significa che, in qualche direzione dello spazio dei regressori, $\psi(t, \theta)$ è circa nullo e quindi c'è una ridondanza nel vettore dei parametri ed è opportuno ridurre l'ordine del modello.

6.1.3 Confronto tra diverse strutture

In generale, la bontà di un modello identificato si valuta in base alla sua capacità di “riprodurre” i dati generati dal sistema incognito. Ma cosa si intende esattamente per “riprodurre i dati”? Una possibile cifra di merito è data dalla capacità predittiva k -passi in avanti del modello. Due casi notevoli sono:

$$k = 1 \quad \Longrightarrow \quad \hat{y}(t|t-1) = (1 - H^{-1}(z, \hat{\theta}_N))y(t) + H^{-1}(z, \hat{\theta}_N)G(z, \hat{\theta}_N)u(t)$$

$$k = \infty \quad \Longrightarrow \quad \hat{y}_s(t) = G(z, \hat{\theta}_N)u(t)$$

Quando $k = \infty$ si parla di *simulazione*, in quanto l'uscita del modello $\hat{y}_s(t)$ viene calcolata simulando il modello identificato $G(z, \hat{\theta}_N)$ con in ingresso il segnale $u(t)$. È chiaro che la capacità predittiva in simulazione è un requisito molto più stringente di quella in predizione, poiché non si utilizza la conoscenza dei valori passati dell'uscita $y(t)$.

Una prima possibile misura della qualità del modello è data dal valore del funzionale di costo in corrispondenza del vettore dei parametri ottimo:

$$V_N(\hat{\theta}_N, \mathcal{Z}^N) = \frac{1}{N} \sum_{t=1}^N (y(t) - \hat{y}(t|t-1))^2$$

Ovviamente, minore è tale valore e maggiore è la capacità del modello di riprodurre i dati. Tuttavia, è chiaro che aumentando la complessità del

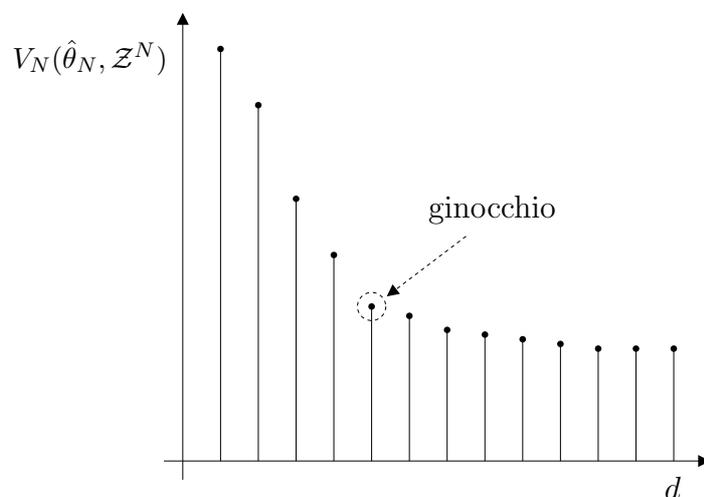


Figura 6.1. Valore del minimo del funzionale di costo al variare del numero di parametri.

modello (e quindi il numero $d = \dim(\theta)$ di parametri da stimare) il costo $V_N(\hat{\theta}_N, \mathcal{Z}^N)$ diminuisce sempre. Consideriamo infatti due classi di modelli di complessità crescente

$$\mathcal{M}_1 \subset \mathcal{M}_2,$$

ovvero il caso in cui $\Theta_1 \subseteq \mathbb{R}^{n_1}$, $\Theta_2 \subseteq \mathbb{R}^{n_2}$, con $n_2 > n_1$, e Θ_1 si può ottenere mediante una opportuna restrizione di Θ_2 a \mathbb{R}^{n_1} (mettendo a zero i parametri eccedenti). Si ha quindi:

$$V_N(\hat{\theta}_N^{(1)}, \mathcal{Z}^N) \geq V_N(\hat{\theta}_N^{(2)}, \mathcal{Z}^N)$$

dove $\hat{\theta}_N^{(1)}$ e $\hat{\theta}_N^{(2)}$ sono rispettivamente i parametri ottimi per le classi \mathcal{M}_1 e \mathcal{M}_2 . L'andamento qualitativo del costo al variare del numero di parametri d sarà del tipo riportato in Figura 6.1. Fino ad un certo valore di d la diminuzione del costo è significativa perché dovuta ad un modello più adatto a spiegare i dati. Viceversa, per valori più alti il costo diminuisce solo perché i nuovi parametri consentono di riprodurre anche la realizzazione del rumore presente sui dati (fenomeno detto “*overfit*”). Ovviamente questo contributo è del tutto inutile ai fini dell'identificazione del sistema dinamico. Il ginocchio della curva può essere usato come metodo euristico per individuare l'ordine

migliore per la classe di modelli scelta.

Un approccio più quantitativo consiste nello scegliere il numero di parametri ottimo sulla base di opportuni criteri numerici. L'idea di base consiste nell'associare una penalità a ciascun parametro, in modo che il costo finale aumenti al crescere dell'ordine del modello in presenza di overfit. In particolare, all'interno di una data classe di modelli \mathcal{M} , l'ordine ottimo d^* sarà quello che minimizza il funzionale

$$\bar{\mathcal{J}}(\mathcal{M}) = V_N(\hat{\theta}_N, \mathcal{Z}^N)(1 + U(d)).$$

A seconda della forma della funzione $U(\theta)$ si ottengono diversi criteri, tra i quali i più diffusi sono:

- **AIC**, criterio di *informazione di Akaike*, dove

$$U(\theta) = \frac{2d}{N};$$

- **FPE**, criterio *final prediction error*, dove

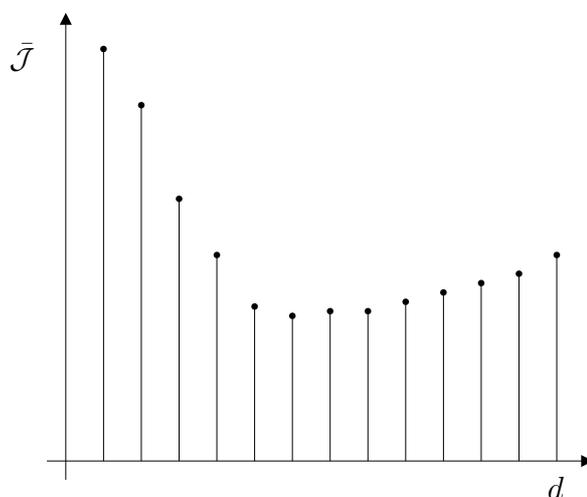
$$U(\theta) = \frac{2d}{N - d};$$

- **MDL**, criterio *minimum description length*, dove

$$U(\theta) = \frac{\log N}{N}d.$$

Per valori “piccoli” di d , il costo $\bar{\mathcal{J}}$ è decrescente (poiché al crescere di d decresce $V_N(\hat{\theta}_N, \mathcal{Z}^N)$). Al contrario, all'aumentare del numero di parametri d il costo diventa crescente poiché $V_N(\hat{\theta}_N, \mathcal{Z}^N)$ diviene pressochè costante e il termine $U(\theta)$ prevale (vedi Figura 6.2).

Le considerazioni fin qui esposte hanno sempre riguardato un confronto tra diverse strutture, basato sullo stesso funzionale di costo $V_N(\hat{\theta}_N, \mathcal{Z}^N)$ che viene minimizzato per ottenere il modello ottimo all'interno della classe. Naturalmente, è possibile impiegare funzionali diversi per la stima dei parametri e per il confronto tra modelli. Un tipico esempio è dato dall'utilizzo

Figura 6.2. Andamento del costo \bar{J} .

per quest'ultimo scopo dell'errore di simulazione

$$J_s = \frac{1}{N} \sum_{t=1}^N (y(t) - \hat{y}_s(t))^2 = \frac{1}{N} \sum_{t=1}^N \left(y(t) - G(z, \hat{\theta}_N) u(t) \right)^2. \quad (6.1)$$

Spesso, nel confronto tra diverse classi di modelli, si preferisce usare l'indicatore di FIT percentuale, definito come

$$FIT = \left\{ 1 - \frac{J_s}{\frac{1}{N} \sum_{t=1}^N (y(t) - \bar{y})^2} \right\} \times 100$$

dove $\bar{y} = \frac{1}{N} \sum_{t=1}^N y(t)$ è la media campionaria dei dati. Un FIT del 100% denota una perfetta riproduzione in simulazione dei dati da parte del modello (e corrisponde quindi al caso ideale in cui i dati sono stati generati proprio dalla funzione di trasferimento identificata $G(z, \hat{\theta}_N)$).

È bene richiamare che la richiesta di un FIT elevato in simulazione (ovvero di un errore J_s piccolo) è un requisito più stringente rispetto alla minimizzazione di $V_N(\hat{\theta}_N, \mathcal{Z}^N)$ e costituisce quindi un test più probante per la valutazione della qualità del modello identificato. Fa eccezione in tal senso la classe di modelli OE, per i quali si ha $\hat{y}_s(t) = \hat{y}(t|t-1)$.

Confronto su dati diversi

Una tipica misura della qualità del modello identificato è data dalla capacità di riprodurre set di dati differenti da quello usato per la stima dei parametri. Infatti, un modello non viene identificato per riprodurre i dati \mathcal{Z}^N utilizzati nella procedura di identificazione, ma perché sia rappresentativo del comportamento del sistema in diverse condizioni di funzionamento (ad esempio, per un segnale di ingresso diverso da quello presente nei dati \mathcal{Z}^N).

Se si desidera confrontare modelli di struttura differente, utilizzando dati diversi da quelli usati per identificare i modelli stessi, si può procedere nel modo seguente. In primo luogo, è necessario dividere i dati di cui si dispone in due sottoinsiemi:

- $\mathcal{Z}_1^N = \{u_1(1), y_1(1), \dots, u_1(N), y_1(N)\}$: *dati di identificazione*;
- $\mathcal{Z}_2^M = \{u_2(1), y_2(1), \dots, u_2(M), y_2(M)\}$: *dati di validazione*.

I dati di identificazione vengono quindi utilizzati per la stima dei parametri θ , secondo la tecnica di identificazione parametrica prescelta (ad esempio, una di quelle descritte nel Capitolo 5). Per evidenziare la dipendenza dai dati di identificazione, si denoti il vettore dei parametri ottenuto come $\hat{\theta}_N(\mathcal{Z}_1^N)$. Infine, si utilizzi come criterio quantitativo per la valutazione della qualità del modello un funzionale basato sui dati di validazione \mathcal{Z}_2^M . Il modo più diretto consiste naturalmente nell'utilizzare l'errore di predizione a un passo, ovvero

$$V_N(\hat{\theta}_N(\mathcal{Z}_1^N), \mathcal{Z}_2^M) = \frac{1}{M} \sum_{t=1}^M (y_2(t) - \hat{y}_2(t|t-1))^2$$

dove il pedice 2 nell'espressione di destra indica che i dati utilizzati per il calcolo del funzionale sono quelli di validazione \mathcal{Z}_2^M . Si noti in particolare che la predizione $\hat{y}_2(t|t-1)$ viene calcolata utilizzando il secondo set di dati, in base al modello $\hat{\theta}_N(\mathcal{Z}_1^N)$, ovvero

$$\hat{y}_2(t|t-1) = \left[1 - H^{-1}(z, \hat{\theta}_N(\mathcal{Z}_1^N)) \right] y_2(t) + H^{-1}(z, \hat{\theta}_N(\mathcal{Z}_1^N)) G(z, \hat{\theta}_N(\mathcal{Z}_1^N)) u_2(t).$$

Chiaramente, possono essere utilizzati funzionali diversi, come ad esempio J_s definito in (6.1).

Nel caso in cui si disponga di un solo set di dati, è comunque buona norma suddividerlo in due sottoinsiemi utilizzandone uno per l'identificazione e uno per la validazione. Nel caso in cui i dati a disposizione non siano molti, si tenderà a preservare il sottoinsieme più numeroso per l'identificazione, al fine di avere una stima sufficientemente accurata dei modelli.

6.1.4 Validazione del modello: analisi dei residui

Un ruolo importante nel contesto della validazione dei modelli è svolto dagli errori di predizione (detti anche *residui di predizione*):

$$\varepsilon(t, \hat{\theta}_N) = y(t) - \hat{y}(t|t-1).$$

Informalmente, i residui rappresentano quella parte dei dati che il modello non è in grado di riprodurre, per cui valutare le loro proprietà statistiche può fornire una utile indicazione sulla qualità del modello. In particolare, si considerano le seguenti grandezze:

$$\hat{R}_{\varepsilon u}^N(\tau) = \frac{1}{N-\tau} \sum_{t=1}^{N-\tau} \varepsilon(t+\tau)u(t), \quad -M \leq \tau \leq M \quad (6.2)$$

$$\hat{R}_{\varepsilon}^N(\tau) = \frac{1}{N-\tau} \sum_{t=1}^{N-\tau} \varepsilon(t+\tau)\varepsilon(t), \quad 1 \leq \tau \leq M \quad (6.3)$$

dove M è un intero maggiore di 1 e tipicamente $M \ll N - \tau$. Ovviamente, $\hat{R}_{\varepsilon u}^N(\tau)$ e $\hat{R}_{\varepsilon}^N(\tau)$ rappresentano rispettivamente una stima della funzione di covarianza incrociata fra ε ed u , e della funzione di covarianza dei residui ε . Intuitivamente, entrambe le precedenti quantità devono essere “piccole”, per i seguenti motivi.

- Il residuo di predizione non deve dipendere dal particolare set di dati impiegato. In particolare, non deve essere correlato con l'ingresso, altrimenti la qualità del modello potrebbe cambiare con ingressi differenti. Per questo occorre che i valori $\hat{R}_{\varepsilon u}^N(\tau)$ siano piccoli. Un altro modo per giustificare tale requisito è il seguente. Se c'è dipendenza tra

ε e u , allora c'è un contributo sull'uscita y che deriva dall'ingresso u e che non è stato spiegato dal modello identificato. Quindi il modello può essere migliorato.

- Se i residui di predizione sono tra loro correlati, allora una parte di $\varepsilon(t)$ avrebbe potuto essere prevista sulla base dei dati \mathcal{Z}^{t-1} disponibili fino all'istante precedente. Anche in questo caso quindi il modello può essere migliorato.

Da un punto di vista teorico, una giustificazione di quanto detto in precedenza si ha se si fa l'ipotesi ideale che i dati siano generati da un modello che appartiene effettivamente alla classe dei modelli LTI considerata. In tal caso si ha che i dati soddisfano effettivamente l'equazione

$$y(t) = G_o(z)u(t) + H_o(z)e(t).$$

per qualche $G_o(z)$, $H_o(z)$. Nel caso in cui l'algoritmo di identificazione riesca effettivamente a stimare i parametri esatti del modello in esame¹, utilizzando la (5.7) l'errore di predizione sarà dato da

$$\begin{aligned} \varepsilon(t) &= H_o^{-1}(z) (y(t) - G_o(z)u(t)) \\ &H_o^{-1}(z) (G_o(z)u(t) + H_o(z)e(t) - G_o(z)u(t)) = e(t). \end{aligned}$$

In altre parole, nel caso di perfetta identificazione, l'errore di predizione risulta essere un processo stocastico bianco, e quindi $R_{\varepsilon}(\tau) = 0$ per $\tau \neq 0$. Analogamente si può concludere che nel caso ideale $\varepsilon(t)$ deve essere scorrelato dall'ingresso $u(t)$ (nel caso in cui $u(t)$ sia un processo stocastico, è sufficiente assumere che sia scorrelato da $e(t)$), e quindi $R_{\varepsilon u}(\tau) = 0, \forall \tau$.

Dal momento che le (6.2)-(6.3) sono solo delle stime delle reali funzioni di covarianza, non è ragionevole richiedere che esse siano esattamente zero. Per questo ci si accontenta che siano sufficientemente piccole. Come quantificare queste considerazioni? Esistono dei test statistici che consentono di stabilire

¹Ciò ovviamente è possibile in generale, sotto opportune ipotesi, solo nel caso in cui si faccia tendere all'infinito il numero dei dati N .

(con un livello di confidenza fissato) se una sequenza di valori può essere interpretata come una realizzazione di un rumore bianco. Consideriamo il caso della \hat{R}_ε^N (ma una trattazione del tutto analoga può essere fatta anche per la $\hat{R}_{\varepsilon u}^N$). Supponiamo che i residui $\{\varepsilon(t)\}_{t=1}^N$ siano effettivamente bianchi, con varianza σ_ε^2 . Allora, si dimostra che asintoticamente (vale a dire al tendere di N all'infinito) il vettore:

$$E_M = \sqrt{N} \begin{bmatrix} \hat{R}_\varepsilon^N(1) \\ \hat{R}_\varepsilon^N(2) \\ \vdots \\ \hat{R}_\varepsilon^N(M) \end{bmatrix}$$

ha una distribuzione Gaussiana, a media nulla e matrice di covarianza pari a $\sigma_\varepsilon^4 I$. Quindi, un possibile test di bianchezza su $\{\varepsilon(t)\}_{t=1}^N$, può essere fatto verificando che per ogni $\tau \neq 0$, la covarianza stimata $\hat{R}_\varepsilon^N(\tau)$ appartenga ad un opportuno intervallo di confidenza $[-\gamma_\varepsilon, \gamma_\varepsilon]$. L'ampiezza dell'intervallo si determina a partire dal livello di confidenza desiderato. Ad esempio, un test al 99% (ossia, accettando una probabilità di commettere un errore pari a 1%) la soglia vale

$$\gamma_\varepsilon = \frac{\alpha}{\sqrt{N}} R_\varepsilon^N(0)$$

in cui α è tale per cui $P(|\mathbf{x}| < \alpha) = 0.99$, con \mathbf{x} v.a. normale. Per calcolare α in MATLAB[®], è possibile utilizzare la funzione `erfinv`:

$$\alpha = \sqrt{2} \text{erfinv}(0.99) = 2.58.$$

Per l'analisi dei residui di predizione, si veda anche il comando `resid` del *System Identification Toolbox* di MATLAB (Ljung, 2005).

6.2 Scelta dell'ingresso

In alcuni casi è possibile scegliere l'ingresso da applicare ad un sistema per raccogliere dati ai fini dell'identificazione di un modello. Poichè tale scelta influenza il risultato della procedura di identificazione, è opportuno tener presente alcuni aspetti.

- Le caratteristiche asintotiche del vettore dei parametri stimato non dipendono dalla forma d'onda dell'ingresso, ma solo dal suo contenuto in frequenza (spettro). In particolare, maggiore è la potenza dell'ingresso (e quindi maggiore il rapporto segnale/rumore) e minore sarà la varianza della stima.
- L'ingresso deve avere ampiezza limitata.
- L'ingresso deve avere un contenuto in frequenza elevato nella banda di interesse del sistema. L'idea è quella di individuare la banda del sistema e scegliere un ingresso con spettro piatto in tale banda.

È chiaro che vi sono esigenze contrastanti. Lo scopo è quello di ottenere la massima potenza dell'ingresso (almeno nella banda di interesse), nell'ambito dei limiti di ampiezza imposti. Di seguito riportiamo alcuni segnali di ingresso tipici per esperimenti di identificazione.

Rumore bianco. Ha spettro piatto e se opportunamente filtrato consente di ottenere qualunque spettro (almeno in linea di principio). Il vincolo sull'ampiezza massima \bar{u} può essere rispettato scartando i campioni con ampiezza maggiore di tale soglia. Nel caso di un rumore bianco Gaussiano, se si sceglie la deviazione standard σ in modo che $3\sigma = \bar{u}$ questo procedimento comporta solo piccole distorsioni dello spettro.

Random Binary Sequence (RBS). È un segnale casuale che può assumere solo due valori. Ad esempio, può essere generato prendendo il segno di un rumore bianco Gaussiano (eventualmente scalato, per ottenere i livelli desiderati).

Pseudo-Random Binary Sequence (PRBS). È un segnale periodico, deterministico, con caratteristiche simili a quelle del rumore bianco. Se indichiamo con M il periodo, lo spettro è costituito da M impulsi di pari ampiezza. Rispetto al rumore bianco, le proprietà di scorrelazione non valgono solo asintoticamente ma per ogni set di lunghezza pari ad un multiplo intero del periodo.

Somme di sinusoidi. In questo caso il segnale di ingresso ha la forma:

$$u(t) = \sum_{k=1}^{k=d} A_k \sin(\omega_k t + \phi_k).$$

Anche questo segnale ha uno spettro costituito da un insieme di impulsi a frequenza ω_k . Variando opportunamente i parametri d , ω_k , ϕ_k è possibile concentrare la potenza del segnale a frequenza ben precise.

Chirp. Un segnale di questo tipo è una sinusoidale la cui frequenza varia con continuità all'interno di un intervallo $[\omega_1, \omega_2]$, in un certo periodo temporale $0 \leq t \leq M$:

$$u(t) = A \cos \left(\omega_1 t + (\omega_2 - \omega_1) \frac{t^2}{2M} \right).$$

Alcuni dei precedenti ingressi sono periodici, altri no. È opportuno chiedersi quali siano i pro ed i contro dell'impegno di tali segnali. Un segnale periodico di periodo M ha lo svantaggio di eccitare solo M frequenze distinte. D'altro canto si possono individuare almeno due benefici derivanti dall'uso di ingressi periodici. Innanzitutto, è possibile mediare l'uscita tra i vari periodi e lavorare solo con un periodo di dati "medi". Questa operazione ha l'effetto di aumentare il rapporto segnale/rumore di un fattore pari al numero dei periodi mediati. Inoltre, si può stimare il rumore presente sull'uscita, sfruttando il fatto che la differenza tra le uscite nei vari periodi è dovuta unicamente al rumore (una volta esauriti gli effetti del transitorio).

Per le problematiche legate alla scelta dell'ingresso, si veda anche il comando `idinput` del *System Identification Toolbox* di MATLAB (Ljung, 2005).

Bibliografia

Bittanti, S. and M. Campi (1995). *Raccolta di problemi di Identificazione, Filtraggio, Controllo Predittivo*. Pitagora.

Lewis, F. L. (1986). *Optimal Estimation*. John Wiley & Sons.

Ljung, L. (1999). *System Identification : Theory for the User*. 2nd ed.. Prentice Hall.

Ljung, L. (2005). *System Identification Toolbox - Version 6.1*. The Mathworks.

Rohatgi, V. K. and A. K. Md. E. Saleh (2001). *An introduction to probability and statistics*. 2nd ed.. Wiley Interscience.

Zappa, G. (1999). *Appunti del corso di "Modellistica e Identificazione"*. Facoltà di Ingegneria, Università di Siena.