Appunti
di
Probabilità e Teoria della Stima

Antonello Giannitrapani, Andrea Garulli
Master E^2C
Versione 1.0

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 Teoria della stima **35**

2.1 Stima parametrica 35
2.1.1 Formulazione del problema. 35
2.1.2 Proprietà di uno stimatore 37
2.1.3 Stima non polarizzata a minima varianza 42
2.2 Limite di Cramér-Rao 45
2.3 Stima a massima verosimiglianza 47
2.3.1 Rumore additivo 52
2.4 Problemi di stima lineare 53
2.5 Stima Bayesiana 56
2.5.1 Stima a minimo errore quadratico medio 56
2.5.2 Stima lineare a minimo errore quadratico medio 59
2.6 Esercizi 64

Elenco delle figure

1.1 Decomposizione di una funzione distribuzione di probabilità nelle sue componenti continue e discrete. 9
1.2 Interpretazione delle funzioni densità e distribuzione di probabilità. 11
1.3 Densità di probabilità uniforme $U[a,\ b]$. 12
1.4 Densità di probabilità gaussiana a media nulla e varianza unitaria. 20
1.5 Dominio di integrazione della (1.20) per ζ fissato. 23
1.6 La $\mathrm{P}(A \cap B)$ è pari alla probabilità che $(\boldsymbol{x}, \boldsymbol{y})$ appartenga alla zona in grigio. 27

2.1 Densità di probabilità di uno stimatore corretto e di uno polarizzato. 38
2.2 Densità di probabilità di uno stimatore consistente. 40

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 $\{\Omega,\ \mathcal{A},\ \mathrm{P}(\cdot)\}$, 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) $\forall A_1, A_2 \in \mathcal{A} \Rightarrow A_1 \cup A_2 \in \mathcal{A}$;

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

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

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

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

   (a) $\mathrm{P}(A) \geq 0, \quad \forall A \in \mathcal{A}$;

   (b) $\mathrm{P}(\Omega) = 1$;

   (c) $\forall A_1, A_2 \in \mathcal{A} : \ A_1 \cap A_2 = \emptyset \Rightarrow \mathrm{P}(A_1 \cup A_2) = \mathrm{P}(A_1) + \mathrm{P}(A_2)$.

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 $\mathrm{P}(\cdot)$ 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 $\overline{A} \in \mathcal{A}$. Ora, per le proprietà 2a-2b occorre che:

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

Ma, direttamente dalla definizione di insieme complementare, segue che $A \cup \overline{A} = \Omega$ e $A \cap \overline{A} = \emptyset$. 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 $\Omega = \{\omega_0, \omega_1, \ldots, \omega_{36}\}$, uno in corrispondenza di ciascun settore in cui può fermarsi la biglia:

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

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:

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

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

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

ovviamente:

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

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

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

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

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

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:

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

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:

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

△

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

- $\forall A \in \mathcal{A} \Rightarrow \boxed{\mathrm{P}\left(\overline{A}\right) = 1 - \mathrm{P}(A).}$
  Infatti, dalla 3b, $\mathrm{P}(\Omega) = 1$. Ma $\Omega = A \cup \overline{A}$, per cui:

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

  in virtù della 3c.

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

- $\boxed{\mathrm{P}(\emptyset) = 0.}$
  Infatti, poiché $\overline{\Omega} = \emptyset$, si ha:

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

- $\forall A_1, A_2 \in \mathcal{A} : \boxed{A_1 \subseteq A_2 \Rightarrow \mathrm{P}(A_1) \leq \mathrm{P}(A_2).}$
  Infatti, si osservi che, poiché $A_1 \subseteq A_2$ è sempre possibile scrivere

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

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

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

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

- $\forall A_1, A_2 \in \mathcal{A} \Rightarrow \boxed{\mathrm{P}(A_1 \cup A_2) = \mathrm{P}(A_1) + \mathrm{P}(A_2) - \mathrm{P}(A_1 \cap A_2).}$
  Innanzitutto, si osservi come è sempre possibile scrivere:

$$A_1 \cup A_2 = A_1 \cup \left(A_2 \cap \overline{A}_1\right).$$

con $A_1 \cap \left(A_2 \cap \overline{A}_1\right) = \emptyset$. Inoltre, vale l'identità:

$$A_2 \cap \overline{A}_1 = \left(A_2 \cap \overline{A}_1\right) \cup \emptyset = \left(A_2 \cap \overline{A}_1\right) \cup \left(A_2 \cap \overline{A}_2\right) = A_2 \cap \left(\overline{A}_1 \cup \overline{A}_2\right).$$

Quindi, utilizzando la proprietà 3c:

$$\begin{aligned} \mathrm{P}(A_1 \cup A_2) &= \mathrm{P}\left(A_1 \cup \left(A_2 \cap \left(\overline{A}_1 \cup \overline{A}_2\right)\right)\right) \\ &= \mathrm{P}(A_1) + \mathrm{P}\left(A_2 \cap \left(\overline{A}_1 \cup \overline{A}_2\right)\right). \end{aligned} \tag{1.3}$$

Ora, notando che:

$$\begin{aligned} A_2 \cap \left(\overline{A}_1 \cup \overline{A}_2\right) &= \overline{\overline{A_2 \cap \left(\overline{A}_1 \cup \overline{A}_2\right)}} = \overline{\overline{A}_2 \cup \overline{\left(\overline{A}_1 \cup \overline{A}_2\right)}} \\ &= \overline{\overline{A}_2 \cup (A_1 \cap A_2)}, \end{aligned}$$

si ricava:

$$\begin{aligned} \mathrm{P}\left(A_2 \cap \left(\overline{A}_1 \cup \overline{A}_2\right)\right) &= \mathrm{P}\left(\overline{\overline{A}_2 \cup (A_1 \cap A_2)}\right) = 1 - \mathrm{P}\left(\overline{A}_2 \cup (A_1 \cap A_2)\right) \\ &\overset{[1]}{=} 1 - \left(\mathrm{P}\left(\overline{A}_2\right) + \mathrm{P}(A_1 \cap A_2)\right) \\ &= 1 - (1 - \mathrm{P}(A_2) + \mathrm{P}(A_1 \cap A_2)) \\ &= \mathrm{P}(A_2) - \mathrm{P}(A_1 \cap A_2), \end{aligned} \tag{1.4}$$

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 $\{\Omega,\ \mathcal{A},\ \mathrm{P}(\cdot)\}$. Due eventi $A, B \in \mathcal{A}$ si dicono *statisticamente indipendenti* se:

$$\mathrm{P}(A \cap B) = \mathrm{P}(A) \cdot \mathrm{P}(B).$$

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

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

$$\mathrm{P}(A|B) \stackrel{\triangle}{=} \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}, \tag{1.5}$$

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

$\mathrm{P}(A|B)$ 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:

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

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 $\mathrm{P}(A_2) = \frac{1}{6}$.

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:

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

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

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

Dalla Definizione 1.3:

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

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

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

△

Finora, abbiamio 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 $\mathcal{A}$ affinchè 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*.

[2]In 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:

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

1.2 Variabili aleatorie, distribuzioni e densità di probabilità

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

$$\boldsymbol{x} : \Omega \to \mathbb{R}\,.$$

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

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

$$F_{\boldsymbol{x}}\left(\xi\right) \stackrel{\triangle}{=} \mathrm{P}\left(\boldsymbol{x} \leq \xi\right),$$

dove $\mathrm{P}\left(\cdot\right)$ denota la funzione di probabilità associata all'insieme Ω.

La funzione di distribuzione della probabilità gode delle seguenti proprietà:

1. $0 \leq F_{\boldsymbol{x}}\left(\xi\right) \leq 1$;
2. $\lim\limits_{\xi \to \infty} F_{\boldsymbol{x}}\left(\xi\right) = 1, \quad \lim\limits_{\xi \to -\infty} F_{\boldsymbol{x}}\left(\xi\right) = 0$;
3. $F_{\boldsymbol{x}}\left(\xi\right)$ 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):

$$F_{\boldsymbol{x}}\left(\xi\right) = F^c_{\boldsymbol{x}}(\xi) + F^d_{\boldsymbol{x}}(\xi),$$

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

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

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 $F_{\boldsymbol{x}}(\xi)$, la distribuzione è continua a destra, cioè $\lim_{\xi \to \xi_i^+} F_{\boldsymbol{x}}(\xi) = F_{\boldsymbol{x}}(\xi_i)$.

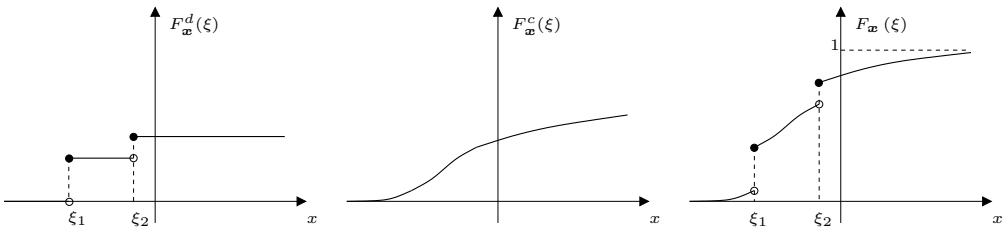


Figura 1.1: Decomposizione di una funzione distribuzione di probabilità nelle sue componenti continue e discrete.

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

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

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

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

Direttamente dalla Definizione 1.6 discende che:

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

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:

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

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

$$\mathrm{P}\left(a \leq \boldsymbol{x} \leq b\right) = \int_a^b f_{\boldsymbol{x}}(\xi)\,\mathrm{d}x. \tag{1.6}$$

La relazione precedente fornisce un'interpretazione della funzione di densità di probabilità (v. Figura 1.2). La $f_{\boldsymbol{x}}(\xi)$, infatti, consente di calcolare la probabilità (corrispondente all'area della zona in grigio in Figura 1.2) con la quale una v.a. $\boldsymbol{x}$ 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 $F_{\boldsymbol{x}}(b) - F_{\boldsymbol{x}}(a)$. Inoltre, supponendo che la densità di probabilità non contenga delle delta di Dirac, la $\mathrm{P}\left(\boldsymbol{x} = a\right) = 0$, per qualunque valore $a \in \mathbb{R}$. Infatti, il dominio di integrazione della $f_{\boldsymbol{x}}(\xi)$ 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 $f_{\boldsymbol{x}}(\xi)$ *non* corrisponde alla probabilità che $\boldsymbol{x} = \xi$. Essa indica solamente in quali "zone" dell'asse reale è più o meno probabile che la $\boldsymbol{x}$ si manifesti.

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

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

è detta densità *gaussiana*.

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

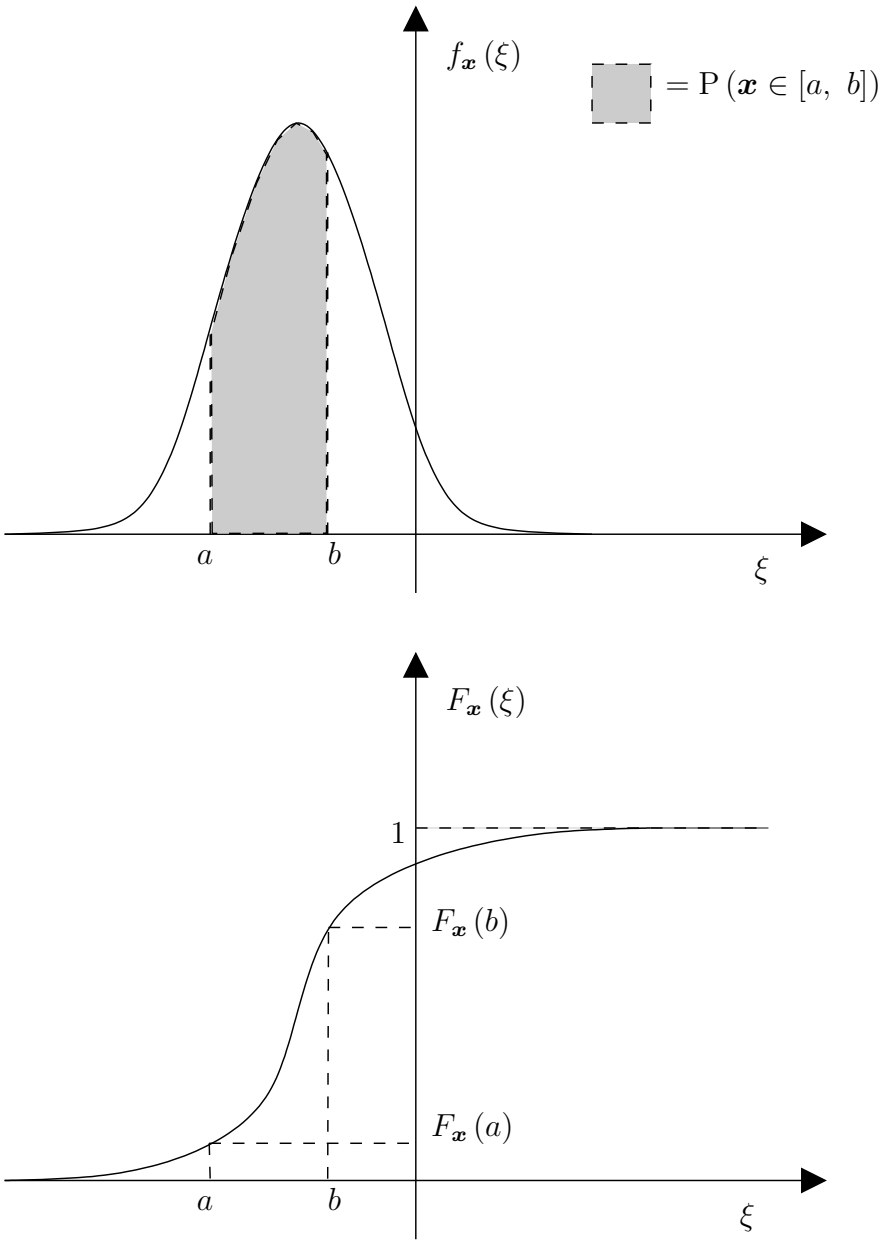


Figura 1.2: Interpretazione delle funzioni densità e distribuzione di probabilità.

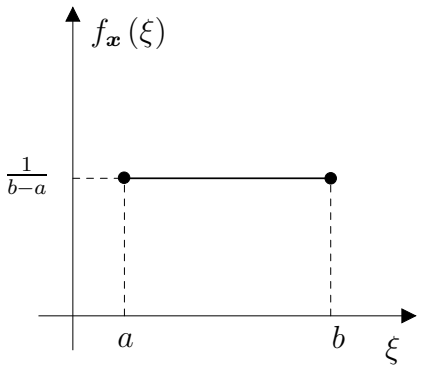


Figura 1.3: Densità di probabilità uniforme $U[a,\ b]$.

Esempio 1.4. Distribuzione uniforme. Una funzione di densità di probabilità del tipo:

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

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

Definizione 1.7. Una v.a. $\boldsymbol{x}$ 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:

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

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

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

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

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

1.3 Distribuzioni bivariate e multivariate

Siano $\boldsymbol{x}$ e $\boldsymbol{y}$ due variabili aleatorie.

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

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

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

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

Direttamente dalla Definizione 1.9 discende che:

$$F_{\boldsymbol{x},\boldsymbol{y}}(\xi,\eta) = \int_{-\infty}^{\eta}\int_{-\infty}^{\xi} f_{\boldsymbol{x},\boldsymbol{y}}(\xi',\eta')\,\mathrm{d}\xi'\mathrm{d}\eta'. \tag{1.8}$$

Una distribuzione di probabilità congiunta soddisfa le seguenti proprietà:

1. $F_{\boldsymbol{x},\boldsymbol{y}}(\xi,\eta)$ è monotona non decrescente in ξ e η;
2. $\lim\limits_{\substack{\xi\to-\infty\\ \eta\to-\infty}} F_{\boldsymbol{x},\boldsymbol{y}}(\xi,\eta) = 0, \qquad \lim\limits_{\substack{\xi\to+\infty\\ \eta\to+\infty}} F_{\boldsymbol{x},\boldsymbol{y}}(\xi,\eta) = 1.$

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

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

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

$$\mathrm{P}\left((\boldsymbol{x},\boldsymbol{y})\in A\right)=\int\int_{A} f_{\boldsymbol{x},\boldsymbol{y}}\left(\xi,\eta\right)\mathrm{d}\xi\mathrm{d}\eta, \tag{1.9}$$

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 $f_{\boldsymbol{x},\boldsymbol{y}}\left(\xi,\eta\right)$ è sempre possibile calcolare le singole densità marginali:

$$f_{\boldsymbol{x}}\left(\xi\right)=\int_{-\infty}^{\infty} f_{\boldsymbol{x},\boldsymbol{y}}\left(\xi,\eta\right)\mathrm{d}\eta; \tag{1.10}$$

$$f_{\boldsymbol{y}}\left(\eta\right)=\int_{-\infty}^{\infty} f_{\boldsymbol{x},\boldsymbol{y}}\left(\xi,\eta\right)\mathrm{d}\xi. \tag{1.11}$$

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

$$f_{\boldsymbol{x},\boldsymbol{y}}\left(\xi,\eta\right)=f_{\boldsymbol{x}}\left(\xi\right)f_{\boldsymbol{y}}\left(\eta\right).$$

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

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

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

Esempio 1.6. La funzione:

$$f_{\boldsymbol{x},\boldsymbol{y}}\left(\xi,\eta\right)=\begin{cases}\dfrac{6}{7}(\xi+\eta)^2 & 0\leq\xi\leq 1,\ \ 0\leq\eta\leq 1\\ 0 & \text{altrove}\end{cases}$$

rappresenta una densità di probabilità bivariata. △

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

$$\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^T,$$

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

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

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

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

$$F_{\boldsymbol{x}}(\xi) = F_{\boldsymbol{x_1},\ldots,\boldsymbol{x_n}}(\xi_1, \ldots, \xi_n) \stackrel{\triangle}{=} \mathrm{P}(\boldsymbol{x}_1 \leq \xi_1, \ldots, \boldsymbol{x}_n \leq \xi_n).$$

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

$$f_{\boldsymbol{x}}(\xi) = f_{\boldsymbol{x_1},\ldots,\boldsymbol{x_n}}(\xi_1, \ldots, \xi_n) \stackrel{\triangle}{=} \frac{\partial^n F_{\boldsymbol{x_1},\ldots,\boldsymbol{x_n}}(\xi_1, \ldots, \xi_n)}{\partial \xi_1 \ldots \partial \xi_n}.$$

Direttamente dalla Definizione 1.12 discende che:

$$F_{\boldsymbol{x_1},\ldots,\boldsymbol{x_n}}(\xi_1, \ldots, \xi_n) = \int_{-\infty}^{\xi_n} \ldots \int_{-\infty}^{\xi_1} f_{\boldsymbol{x_1},\ldots,\boldsymbol{x_n}}(\xi'_1, \ldots, \xi'_n)\, \mathrm{d}\xi'_1 \ldots \mathrm{d}\xi'_n. \quad (1.12)$$

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 dimesioni risultando chiare dal contesto.

1.4 Media e varianza

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

$$\mathbf{E}[\boldsymbol{x}] \stackrel{\triangle}{=} \int_{-\infty}^{+\infty} \xi f_{\boldsymbol{x}}(\xi)\,\mathrm{d}\xi = m_{\boldsymbol{x}}. \tag{1.13}$$

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

$$\mathbf{E}[\alpha\boldsymbol{x} + \beta\boldsymbol{y}] = \alpha\mathbf{E}[\boldsymbol{x}] + \beta\mathbf{E}[\boldsymbol{y}] \qquad \forall\alpha,\beta \in \mathbb{R}\,. \tag{1.14}$$

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

$$\mathbf{E}[g(\boldsymbol{x})] \stackrel{\triangle}{=} \int_{-\infty}^{+\infty} g(\xi) f_{\boldsymbol{x}}(\xi)\,\mathrm{d}\xi.$$

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

Particolare interesse riveste il momento centrale del secondo ordine.

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

$$\sigma^2_{\boldsymbol{x}} \stackrel{\triangle}{=} \mathbf{E}\left[(\boldsymbol{x} - m_{\boldsymbol{x}})^2\right] = \int_{-\infty}^{+\infty} (\xi - m_{\boldsymbol{x}})^2 f_{\boldsymbol{x}}(\xi)\,\mathrm{d}\xi. \tag{1.15}$$

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

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

Definizione 1.15. Si definisce *correlazione* o (*varianza incrociata*) $\sigma_{\boldsymbol{x},\boldsymbol{y}}$ delle v.a. $\boldsymbol{x}, \boldsymbol{y}$ la quantità:

$$\sigma_{\boldsymbol{x},\boldsymbol{y}} \stackrel{\triangle}{=} \mathbf{E}[(\boldsymbol{x} - m_{\boldsymbol{x}})(\boldsymbol{y} - m_{\boldsymbol{y}})].$$

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

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

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

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

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* $\rho_{\boldsymbol{x},\boldsymbol{y}}$ delle v.a. $\boldsymbol{x}, \boldsymbol{y}$ la quantità:

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

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

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

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

Se $\boldsymbol{x}$ ed $\boldsymbol{y}$ 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 $\boldsymbol{x}$ e $\boldsymbol{y}$ siano due v.a. *vettoriali*, le Definizioni 1.14 e 1.15 vanno modificate come segue.

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

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

Definizione 1.19. Si definisce *matrice di correlazione* o (*cross-covarianza*) $P_{\boldsymbol{x}\boldsymbol{y}}$ delle v.a. $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ la quantità:

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

Si noti che, nel caso in cui $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, i valori medi sono anch'essi vettori (colonna) di dimensione n: $m_{\boldsymbol{x}}, m_{\boldsymbol{y}} \in \mathbb{R}^n$, mentre le matrici $P_{\boldsymbol{x}}, P_{\boldsymbol{x}\boldsymbol{y}} \in \mathbb{R}^{n \times n}$.

1.5 Variabili aleatorie gaussiane

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

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

con $m_{\boldsymbol{x}} \in \mathbb{R}^n$ e $P_{\boldsymbol{x}}$ matrice quadrata $n \times n$, simmetrica e definita positiva [3]: $P_{\boldsymbol{x}} = P_{\boldsymbol{x}}^T > 0$. Spesso, per indicare una v.a. avente la d.d.p. precedente si scrive $\boldsymbol{x} \sim N(m_{\boldsymbol{x}}, P_{\boldsymbol{x}})$.

Si noti che, nel caso $n = 2$, posto $P_{\boldsymbol{x}} = \begin{bmatrix} \sigma_{\mathbf{1}}^2 & \rho\sigma_{\mathbf{1}}\sigma_{\mathbf{2}} \\ \rho\sigma_{\mathbf{1}}\sigma_{\mathbf{2}} & \sigma_{\mathbf{2}}^2 \end{bmatrix}$ si ottiene la distribuzione bivariata vista nell'Esempio 1.5.

Le v.a. gaussiane godono delle seguenti proprietà.

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

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

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

4. $\boldsymbol{x}$ e $\boldsymbol{y}$ incorrelate e gaussiane $\Rightarrow \boldsymbol{x}$ e $\boldsymbol{y}$ 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).

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.

[3] Una 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.

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

$$\boldsymbol{Z}_n = \frac{\sum_{i=1}^n \boldsymbol{x}_i - \sum_{i=1}^n m_{\boldsymbol{i}}}{\sqrt{\sum_{i=1}^n \sigma_{\boldsymbol{i}}^2}}.$$

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

$$\lim_{n\to\infty} F_{\boldsymbol{Z_n}}(\eta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\eta} e^{-\frac{t^2}{2}} \mathrm{d}t.$$

△

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

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

Per cui, si usa dire che la v.a. $\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$ risulta essere *asintoticamente* distribuita come $N(m_{\boldsymbol{x}}, \sigma_{\boldsymbol{x}}^2/n)$. 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 $\boldsymbol{x}$ con valor medio $m_{\boldsymbol{x}} = 0$ e varianza $\sigma_{\boldsymbol{x}}^2 = 1$ (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.

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

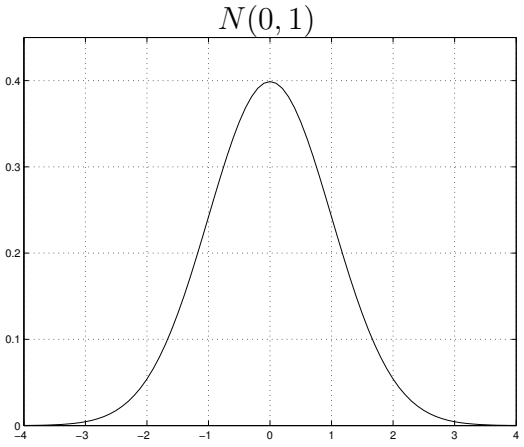


Figura 1.4: Densità di probabilità gaussiana a media nulla e varianza unitaria.

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

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

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 $\eta = (\xi - m_{\boldsymbol{x}})/\sigma_{\boldsymbol{x}}$. La funzione erf$(\cdot)$ (detta "funzione errore") è definita come[4]:

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

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

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

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

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

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

In MATLAB®, vedi comandi: `normpdf` (calcolare il valore di una d.d.p. gaussiana), `erf` (calcolare il valore di erf(x)).

1.6 Funzioni di una o più variabili aleatorie

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

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

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

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

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

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

$$\eta = g(\xi), \tag{1.18}$$

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

$$f_{\boldsymbol{y}}(\bar{\eta}) \stackrel{\triangle}{=} 0.$$

$\triangle$

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

Sia $f_{\boldsymbol{G}}(\gamma)$ 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 $\boldsymbol{R}$ può assumere solo valori compresi fra 9Ω e 11Ω:

$$f_{\boldsymbol{G}}(\gamma) = \begin{cases} \bigstar & \text{se} \quad \dfrac{1}{11} \leq \gamma \leq \dfrac{1}{9} \\ 0 & \text{se} \quad \gamma < \dfrac{1}{11} \quad \text{o} \quad \gamma > \dfrac{1}{9} \end{cases}$$

Procediamo al calcolo di $\bigstar$, 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

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

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

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

Quindi

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

$\triangle$

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

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

$$F_{\boldsymbol{z}}(\zeta) = \mathrm{P}(\boldsymbol{z} \leq \zeta) = \mathrm{P}(\boldsymbol{x} + \boldsymbol{y} \leq \zeta). \tag{1.19}$$

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

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

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

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

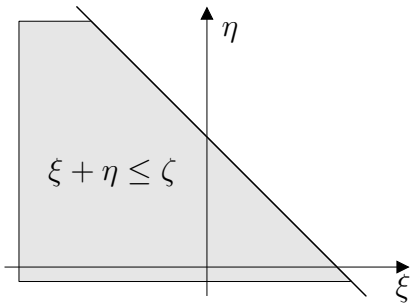


Figura 1.5: Dominio di integrazione della (1.20) per ζ fissato.

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

$$f_{\boldsymbol{z}}(\zeta) = \int_{-\infty}^{+\infty} f_{\boldsymbol{x}}(\zeta - \eta)\, f_{\boldsymbol{y}}(\eta)\, \mathrm{d}\eta = f_{\boldsymbol{x}}(\xi) * f_{\boldsymbol{y}}(\eta). \tag{1.22}$$

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

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

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

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

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

$$\eta = g(\xi) \tag{1.24}$$

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

$$f_{\boldsymbol{y}}(\bar{\eta}) \stackrel{\triangle}{=} 0.$$

$\triangle$

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

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

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

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

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

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:

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

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

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

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

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

Utilizzando l'identità:

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

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

$$\begin{aligned} -\frac{1}{2}(A^{-1}\eta - m_{\boldsymbol{x}})^T P_{\boldsymbol{x}}^{-1}(A^{-1}\eta - m_{\boldsymbol{x}}) &= -\frac{1}{2}\left[A^{-1}(\eta - Am_{\boldsymbol{x}})\right]^T P_{\boldsymbol{x}}^{-1} A^{-1}(\eta - Am_{\boldsymbol{x}}) \\ &\overset{[1]}{=} -\frac{1}{2}(\eta - Am_{\boldsymbol{x}})^T A^{-T} P_{\boldsymbol{x}}^{-1} A^{-1}(\eta - Am_{\boldsymbol{x}}) \\ &\overset{[2]}{=} -\frac{1}{2}(\eta - m_{\boldsymbol{y}})^T A^{-T} P_{\boldsymbol{x}}^{-1} A^{-1}(\eta - m_{\boldsymbol{y}}) \\ &\overset{[3]}{=} -\frac{1}{2}(\eta - m_{\boldsymbol{y}})^T P_{\boldsymbol{y}}^{-1}(\eta - m_{\boldsymbol{y}}), \end{aligned} \tag{1.28}$$

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:

$$m_{\boldsymbol{y}} = Am_{\boldsymbol{x}}, \tag{1.29}$$

mentre la [3] deriva dall'assegnazione:

$$P_{\boldsymbol{y}} = AP_{\boldsymbol{x}}A^T, \tag{1.30}$$

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:

$$\begin{aligned}(\det P_{\boldsymbol{y}})^{1/2} &= (\det A \det P_{\boldsymbol{x}} \det A^T)^{1/2} \\ &\overset{[1]}{=} \left((\det A)^2 \det P_{\boldsymbol{x}}\right)^{1/2} \\ &= |\det A| \, (\det P_{\boldsymbol{x}})^{1/2}\end{aligned} \tag{1.31}$$

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

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

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

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. $\boldsymbol{x}$ e $\boldsymbol{y}$ con densità di probabilità congiunta $f_{\boldsymbol{x},\boldsymbol{y}}(\xi, \eta)$. Supponendo che la variabile aleatoria $\boldsymbol{y}$ abbia assunto un certo valore η, come si modifica la densità di probabilità della $\boldsymbol{x}$? A tale scopo, si considerino i due eventi (v. Figura 1.6):

$$A = \{\boldsymbol{x} \leq \xi\}, \qquad B = \{\eta \leq \boldsymbol{y} \leq \eta + \mathrm{d}\eta\}.$$

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

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

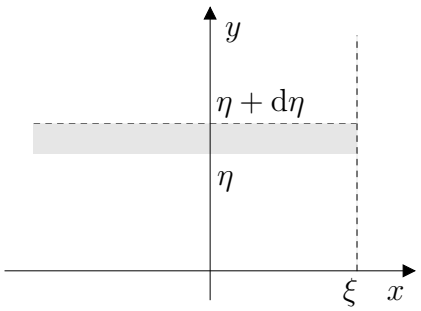


Figura 1.6: La $\mathrm{P}(A \cap B)$ è pari alla probabilità che $(\boldsymbol{x}, \boldsymbol{y})$ appartenga alla zona in grigio.

Lasciando tendere $\mathrm{d}\eta \to 0$:

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

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

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

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

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

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

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

$$f_{\boldsymbol{x}|\boldsymbol{y}}(\xi|\eta) = f_{\boldsymbol{x}}(\xi). \tag{1.37}$$

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

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

$$f_{\boldsymbol{x}}(\xi) = \int_{-\infty}^{+\infty} f_{\boldsymbol{x}|\boldsymbol{y}}(\xi|\eta)\, f_{\boldsymbol{y}}(\eta)\, \mathrm{d}\eta. \tag{1.38}$$

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

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

$$m_{\boldsymbol{x}|\boldsymbol{y}} = \int_{-\infty}^{+\infty} \xi f_{\boldsymbol{x}|\boldsymbol{y}}(\xi|\eta)\, \mathrm{d}\xi \tag{1.39}$$

$$\sigma^2_{\boldsymbol{x}|\boldsymbol{y}} = \int_{-\infty}^{+\infty} (\xi - m_{\boldsymbol{x}|\boldsymbol{y}})^2 f_{\boldsymbol{x}|\boldsymbol{y}}(\xi|\eta)\, \mathrm{d}\xi \tag{1.40}$$

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. $\boldsymbol{y}$. Per cui, ad esempio, il simbolo $m_{\boldsymbol{x}|\boldsymbol{y}}$ va inteso nel senso di $\mathbf{E}[\boldsymbol{x}|\boldsymbol{y} = \eta]$. Per brevità, la dipendenza da η, sarà spesso omessa. Nel caso in cui la $\boldsymbol{x}$ sia una v.a. vettoriale, la (1.40) va opportunamente modificata sostituendo $(\cdot)^2$ con $(\cdot)(\cdot)^T$ (v. Definizione 1.18), ottenendo cosí la matrice di covarianza condizionata $P_{\boldsymbol{x}|\boldsymbol{y}}$.

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

$$\mathbf{E}\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{bmatrix} = \begin{bmatrix}0\\ 0\end{bmatrix}, \qquad R = \mathbf{E}\begin{bmatrix}\boldsymbol{x}_1^{\mathbf{2}} & \boldsymbol{x}_1\boldsymbol{x}_2\\ \boldsymbol{x}_1\boldsymbol{x}_2 & \boldsymbol{x}_2^2\end{bmatrix} = \begin{bmatrix}\sigma_{\mathbf{1}}^2 & r\sigma_{\mathbf{1}}\sigma_{\mathbf{2}}\\ r\sigma_{\mathbf{1}}\sigma_{\mathbf{2}} & \sigma_{\mathbf{2}}^2\end{bmatrix},$$

con $|r| < 1$.

Si consideri il cambio di variabile $\tilde{\boldsymbol{x}}_1 = \boldsymbol{x}_1 - \frac{r\sigma_{\mathbf{1}}\sigma_{\mathbf{2}}}{\sigma_{\mathbf{2}}^2}\boldsymbol{x}_2$. Allora:

$$\begin{bmatrix} \tilde{\boldsymbol{x}}_1 \\ \boldsymbol{x}_2 \end{bmatrix} = \begin{bmatrix} 1 & -\frac{r\sigma_{\mathbf{1}}\sigma_{\mathbf{2}}}{\sigma_{\mathbf{2}}^2} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix}.$$

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

$$\begin{aligned} \mathbf{E}\left[\tilde{\boldsymbol{x}}_1\boldsymbol{x}_2\right] &= \mathbf{E}\left[\left(\boldsymbol{x}_1 - \frac{r\sigma_{\mathbf{1}}\sigma_{\mathbf{2}}}{\sigma_{\mathbf{2}}^2}\boldsymbol{x}_2\right)\boldsymbol{x}_2\right] \\ &\overset{[1]}{=} \mathbf{E}\left[\boldsymbol{x}_1\boldsymbol{x}_2\right] - \frac{r\sigma_{\mathbf{1}}\sigma_{\mathbf{2}}}{\sigma_{\mathbf{2}}^2}\mathbf{E}\left[\boldsymbol{x}_2\boldsymbol{x}_2\right] \\ &\overset{[2]}{=} r\sigma_{\mathbf{1}}\sigma_{\mathbf{2}} - \frac{r\sigma_{\mathbf{1}}\sigma_{\mathbf{2}}}{\sigma_{\mathbf{2}}^2}\sigma_{\mathbf{2}}^2 = 0, \end{aligned} \tag{1.41}$$

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. $\tilde{\boldsymbol{x}}_1$ e $\boldsymbol{x}_2$ 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:

$$f_{\tilde{\boldsymbol{x}}_{\mathbf{1}},\boldsymbol{x}_{\mathbf{2}}}\left(\tilde{x}_1, x_2\right) = f_{\tilde{\boldsymbol{x}}_{\mathbf{1}}}\left(\tilde{x}_1\right) f_{\boldsymbol{x}_{\mathbf{2}}}\left(x_2\right). \tag{1.42}$$

Dal Teorema 2 (v. anche Esempio 1.8):

$$f_{\tilde{\boldsymbol{x}}_{\mathbf{1}},\boldsymbol{x}_{\mathbf{2}}}\left(\tilde{x}_1, x_2\right) = \frac{f_{\boldsymbol{x}_{\mathbf{1}},\boldsymbol{x}_{\mathbf{2}}}\left(\tilde{x}_1 + \frac{r\sigma_{\mathbf{1}}\sigma_{\mathbf{2}}}{\sigma_{\mathbf{2}}^2}x_2, x_2\right)}{1}, \tag{1.43}$$

dal momento che:

$$\det \begin{bmatrix} 1 & -\frac{r\sigma_{\mathbf{1}}\sigma_{\mathbf{2}}}{\sigma_{\mathbf{2}}^2} \\ 0 & 1 \end{bmatrix} = 1.$$

Allora:

$$\begin{aligned} f_{\boldsymbol{x}_{\mathbf{1}}|\boldsymbol{x}_{\mathbf{2}}}\left(x_1|x_2\right) &\overset{[1]}{=} \frac{f_{\boldsymbol{x}_{\mathbf{1}},\boldsymbol{x}_{\mathbf{2}}}\left(x_1, x_2\right)}{f_{\boldsymbol{x}_{\mathbf{2}}}\left(x_2\right)} \overset{[2]}{=} \frac{f_{\tilde{\boldsymbol{x}}_{\mathbf{1}}\boldsymbol{x}_{\mathbf{2}}}\left(x_1 - \frac{r\sigma_{\mathbf{1}}\sigma_{\mathbf{2}}}{\sigma_{\mathbf{2}}^2}x_2, x_2\right)}{f_{\boldsymbol{x}_{\mathbf{2}}}\left(x_2\right)} \\ &\overset{[3]}{=} f_{\tilde{\boldsymbol{x}}_{\mathbf{1}}}\left(x_1 - \frac{r\sigma_{\mathbf{1}}\sigma_{\mathbf{2}}}{\sigma_{\mathbf{2}}^2}x_2\right), \end{aligned} \tag{1.44}$$

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

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

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

e

$$\begin{aligned}\sigma^2_{\tilde{\boldsymbol{x}}_1} &= \mathbf{E}\left[\tilde{\boldsymbol{x}}_1^2\right] = \mathbf{E}\left[\left(\boldsymbol{x}_1 - \frac{r\sigma_1\sigma_2}{\sigma_2^2}\boldsymbol{x}_2\right)^2\right] \\ &\overset{[1]}{=} \mathbf{E}\left[\boldsymbol{x}_1{}^2\right] + \frac{r\sigma_1^2\sigma_2^2}{\sigma_2^4}\mathbf{E}\left[\boldsymbol{x}_2{}^2\right] - 2\frac{r\sigma_1\sigma_2}{\sigma_2^2}\mathbf{E}\left[\boldsymbol{x}_1\boldsymbol{x}_2\right] \\ &\overset{[2]}{=} \sigma_1^2 + r^2\sigma_1^2 - 2r^2\sigma_1^2 = \sigma_1^2(1-r^2).\end{aligned}$$

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

$$f_{\tilde{\boldsymbol{x}}_1}\left(\tilde{x}_1\right) = \frac{1}{\sqrt{2\pi}\sigma_{\tilde{\boldsymbol{x}}_1}} e^{-\frac{\tilde{x}_1^2}{2\sigma^2_{\tilde{\boldsymbol{x}}_1}}}.$$

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

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

Quindi, la d.d.p di $\boldsymbol{x}_1$ condizionata a $\boldsymbol{x}_2$ 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:

$$\begin{aligned}m_{\boldsymbol{x}_1|\boldsymbol{x}_2} &= \frac{r\sigma_1\sigma_2}{\sigma_2^2}x_2; \\ \sigma^2_{\boldsymbol{x}_1|\boldsymbol{x}_2} &= \sigma_1^2\left(1-r^2\right).\end{aligned}$$

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

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

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

$$\mathbf{E}[\boldsymbol{x}] = \begin{bmatrix} m_{\mathbf{1}} \\ m_{\mathbf{2}} \end{bmatrix}, \qquad P_{\boldsymbol{x}} = \begin{bmatrix} R_1 & R_{12} \\ R_{12}^T & R_2 \end{bmatrix} \geq 0,$$

con $R_2 > 0$. Allora:

$$m_{\boldsymbol{x_1}|\boldsymbol{x_2}} = m_1 + R_{12}R_2^{-1}(x_2 - m_2); \tag{1.46}$$

$$P_{\boldsymbol{x}_1|\boldsymbol{x}_2} = R_1 - R_{12}R_2^{-1}R_{12}^T. \tag{1.47}$$

△

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. $\boldsymbol{x}_2$ "sposta" il valor medio di $\boldsymbol{x}_1$ e ne riduce l'incertezza. Infatti, poiché vale $R_1 - R_{12}R_2^{-1}R_{12}^T \geq 0$ (regola di Schur), allora $P_{\boldsymbol{x}_1|\boldsymbol{x}_2} = R_1 - R_{12}R_2^{-1}R_{12}^T \leq R_1$, cioè la varianza della v.a. $\boldsymbol{x}_1$ dopo l'osservazione del valore assunto da $\boldsymbol{x}_2$ è minore di (o tutt'al più uguale a) quella iniziale. Infine, $P_{\boldsymbol{x}_1|\boldsymbol{x}_2} = R_1$ 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. Sia $\boldsymbol{x}$ una v.a. uniformemente distribuita nell' intervallo $[-\frac{T}{2},\ \frac{T}{2}]$. Dopo aver calcolato il valor medio $m_{\boldsymbol{x}}$ e la varianza $\sigma_{\boldsymbol{x}}^2$ di $\boldsymbol{x}$, si utilizzi il comando `rand` di MATLAB® per generare un vettore `x` contenente `N` valori estratti dalla precedente distribuzione, con $T = 6$.

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

```
m_x = sum(x)/N
```

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

1.2. Dopo aver verificato che la $f_{\boldsymbol{x},\boldsymbol{y}}(\xi,\eta)$ 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 $[\boldsymbol{x},\ \boldsymbol{y}]^T$.

1.3. Sia $\boldsymbol{x} = [\boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \boldsymbol{x}_3]^T$ una v.a. in $\mathbb{R}^3$, con densità di probabilità:

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

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

1.4. Sia $\boldsymbol{x}$ una v.a. scalare gaussiana:

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

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

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

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

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

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

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

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

1.7. Una v.a. $\boldsymbol{x}$ avente d.d.p.

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

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

1.8. Sia la tensione $\boldsymbol{V}$, 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 $\boldsymbol{W} = \frac{1}{R}\boldsymbol{V}^2$ dissipata dalla resistenza.

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

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

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

per:

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

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

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

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

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

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

Capitolo 2

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 2.1-2.4, mentre il paragrafo 2.5 è dedicato all'approccio Bayesiano.

2.1 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, $\boldsymbol{y}_1, \dots \boldsymbol{y}_n$.

2.1.1 Formulazione del problema.

Siano:

- $\theta \in \Theta \subseteq \mathbb{R}^p$, un vettore incognito di *parametri*;
- $\boldsymbol{y} = (\boldsymbol{y}_1, \dots \boldsymbol{y}_n)^T \in \mathcal{Y} \subseteq \mathbb{R}^n$ un vettore aleatorio di *osservazioni* (o *misure*);
- $F_{\boldsymbol{y}}^{\theta}(y), f_{\boldsymbol{y}}^{\theta}(y)$ le funzioni di distribuzione e densità di probabilità del vettore delle osservazioni $\boldsymbol{y}$, 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 $\boldsymbol{y}$, prende il nome di *spazio delle osservazioni*. Si assume che la distribuzione $F_{\boldsymbol{y}}^{\theta}(y)$ (o, equivalentemente, la densità $f_{\boldsymbol{y}}^{\theta}(y)$) 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 $\boldsymbol{y} \in \mathcal{Y}$."*

Definizione 2.1. Uno *stimatore* T del parametro incognito θ è una funzione che mappa lo spazio delle osservazioni nello spazio dei parametri:

$$T : \mathcal{Y} \to \Theta.$$

Il valore $\hat{\theta} = T(y)$, assunto dallo stimatore in corrispondenza dell'osservazione $\boldsymbol{y} = y$, 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 $\boldsymbol{y}$, si stima θ con la quantità $\hat{\theta} = T(y)$.

Si noti che $\hat{\theta}$ può essere vista come una realizzazione della variabile aleatoria $T(\boldsymbol{y})$; in altre parole, poiché $T(\boldsymbol{y})$ è una funzione della v.a. $\boldsymbol{y}$, la stima $\hat{\theta}$ è essa stessa una variabile aleatoria.

2.1.2 Proprietà di uno stimatore

Ovviamente, in base alla Definizione 2.1, 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 $\hat{\theta} = T(y)$ 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 2.2. Uno stimatore $T(\boldsymbol{y})$ del parametro θ si dice *corretto* (o, *non polarizzato*) se

$$\mathbf{E}^{\theta}\left[T(\boldsymbol{y})\right] = \theta, \quad \forall \theta \in \Theta. \tag{2.1}$$

Nella precedente definizione si è evidenziata la dipendenza del valore atteso di $T(\boldsymbol{y})$ dal parametro incognito θ, dal momento che la d.d.p. di $\boldsymbol{y}$ dipende, per ipotesi, da θ stesso.

La condizione (2.1) 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 2.1).

Esempio 2.1. Siano $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ variabili aleatorie identicamente distribuite (i.d.) con valor medio m. La quantità:

$$\overline{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i \tag{2.2}$$

prende il nome di *media campionaria*. È immediato verificare che $\overline{\boldsymbol{y}}$ è uno stimatore non polarizzato di m. Infatti, per la linearità dell'operatore di

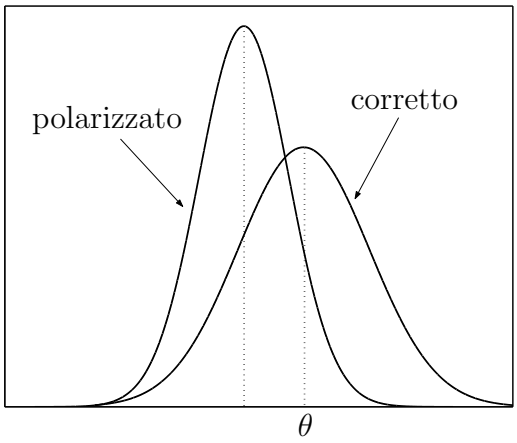


Figura 2.1: Densità di probabilità di uno stimatore corretto e di uno polarizzato.

valore atteso, si ha:

$$\mathbf{E}\left[\overline{\boldsymbol{y}}\right] = \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[\boldsymbol{y}_i\right] = \frac{1}{n}\sum_{i=1}^{n} m = m.$$

△

Esempio 2.2. Siano $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ variabili aleatorie scalari, indipendenti, identicamente distribuite (i.i.d.) con valor medio m e varianza σ^2. La quantità:

$$\hat{\sigma}^2_{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - \overline{\boldsymbol{y}})^2$$

è uno stimatore polarizzato della varianza σ^2. Infatti, dalla (2.2), si ha:

$$\begin{aligned}
\mathbf{E}\left[\hat{\sigma}^2_{\boldsymbol{y}}\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[\left(\boldsymbol{y}_i - \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{y}_j\right)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n^2}\mathbf{E}\left[\left(n\boldsymbol{y}_i - \sum_{j=1}^{n}\boldsymbol{y}_j\right)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n^2}\mathbf{E}\left[\left(n(\boldsymbol{y}_i - m) - \sum_{j=1}^{n}(\boldsymbol{y}_j - m)\right)^2\right].
\end{aligned}$$

Ma:

$$\mathbf{E}\left[\left(n(\boldsymbol{y}_i - m) - \sum_{j=1}^{n}(\boldsymbol{y}_j - m)\right)^2\right] = n^2\mathbf{E}\left[(\boldsymbol{y}_i - m)^2\right]$$
$$- 2n\mathbf{E}\left[(\boldsymbol{y}_i - m)\sum_{j=1}^{n}(\boldsymbol{y}_j - m)\right] + \mathbf{E}\left[\left(\sum_{j=1}^{n}(\boldsymbol{y}_j - m)\right)^2\right]$$
$$= n^2\sigma^2 - 2n\sigma^2 + n\sigma^2$$
$$= n(n-1)\sigma^2$$

poiché, per l'ipotesi di indipendenza, $\mathbf{E}\left[(\boldsymbol{y}_i - m)(\boldsymbol{y}_j - m)\right] = 0$ se $i \neq j$. Quindi:

$$\mathbf{E}\left[\hat{\sigma}^2_{\boldsymbol{y}}\right] = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n^2}n(n-1)\sigma^2 = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

△

Esempio 2.3. Siano $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ variabili aleatorie scalari, indipendenti, identicamente distribuite (i.i.d.) con valor medio m e varianza σ^2. La quantità:

$$S^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(\boldsymbol{y}_i - \overline{\boldsymbol{y}})^2$$

prende il nome di *varianza campionaria*. È immediato verificare che S^2 è uno stimatore non polarizzato della varianza σ^2. Infatti, notando che:

$$S^2 = \frac{n}{n-1}\hat{\sigma}^2_{\boldsymbol{y}},$$

segue immediatamente che:

$$\mathbf{E}\left[S^2\right] = \frac{n}{n-1}\mathbf{E}\left[\hat{\sigma}^2_{\boldsymbol{y}}\right] = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2.$$

△

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 2.3. Sia $\{\boldsymbol{y}_i\}_{i=1}^{\infty}$ una successione di v.a. La successione di stimatori $T_n = T(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ si dice *consistente* per il parametro θ se T_n converge a θ in probabilità, per ogni θ. Cioè, se:

$$\lim_{n\to\infty} \mathrm{P}\left(\|T_n - \theta\| > \varepsilon\right) = 0, \quad \forall \varepsilon > 0, \forall \theta \in \Theta.$$

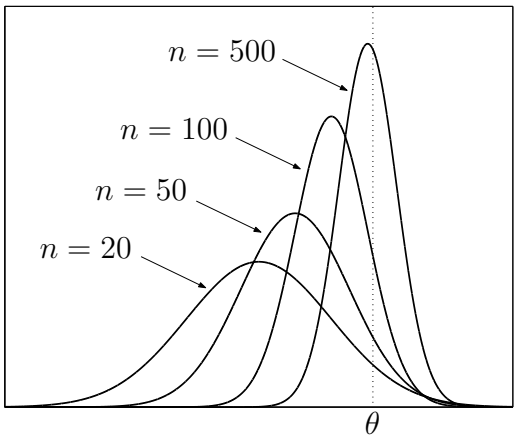


Figura 2.2: Densità di probabilità di uno stimatore consistente.

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 2.2). Il seguente Teorema fornisce una condizione sufficiente per la consistenza di una sequenza di stimatori.

Teorema 4. *Sia T_n una sequenza di stimatori corretti del parametro scalare θ:*

$$\mathbf{E}\left[T_n\right] = \theta, \quad \forall n, \quad \forall \theta \in \Theta.$$

Se:

$$\lim_{n\to\infty} \mathbf{E}\left[(T_n - \theta)^2\right] = 0,$$

allora la sequenza T_n è consistente.

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 2.4. Siano $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ variabili aleatorie indipendenti e identicamente distribuite (i.i.d.) con valor medio m e varianza σ^2. Nell'Esempio 2.1 si è visto come la media campionaria:

$$\overline{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i$$

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:

$$\begin{aligned}\mathrm{Var}(\overline{\boldsymbol{y}}) = \mathbf{E}\left[(\overline{\boldsymbol{y}} - m)^2\right] &= \mathbf{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i - m\right)^2\right] \\ &= \frac{1}{n^2}\mathbf{E}\left[\left(\sum_{i=1}^{n}(\boldsymbol{y}_i - m)\right)^2\right] = \frac{\sigma^2}{n}\end{aligned}$$

dal momento che le v.a. $\boldsymbol{y}_i$ sono indipendenti. Quindi:

$$\mathrm{Var}(\overline{\boldsymbol{y}}) = \frac{\sigma^2}{n} \to 0 \quad \text{per } n \to \infty$$

e, in virtù del Teorema 4, la media campionaria $\overline{\boldsymbol{y}}$ risulta essere uno stimatore consistente per il valor medio m. △

Errore quadratico medio

Una misura della bontà della stima fornita da uno stimatore è *l'errore quadratico medio.*

Definizione 2.4. Si definisce *errore quadratico medio* la quantità:

$$\mathbf{E}^{\theta}\left[(T(\boldsymbol{y})-\theta)^2\right]$$

Si osservi che, se uno stimatore è corretto, allora l'errore quadratico medio coincide con la varianza della stima.

Definizione 2.5. 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:

$$\mathbf{E}^{\theta}\left[(T_1(\boldsymbol{y})-\theta)^2\right] \leq \mathbf{E}^{\theta}\left[(T_2(\boldsymbol{y})-\theta)^2\right], \quad \forall\theta\in\Theta$$

La precedente definizione, in maniera del tutto naturale, afferma che fra due stimatori è preferibile utilizzare quello con errore quadratico medio minore.

2.1.3 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 2.6. Uno stimatore corretto $T^*(\cdot)$ del parametro scalare θ è detto *a minima varianza* (o UMVUE, Uniformly Minimum Variance Unbiased Estimator) se:

$$\mathbf{E}^{\theta}\left[(T^*(\boldsymbol{y})-\theta)^2\right] \leq \mathbf{E}^{\theta}\left[(T(\boldsymbol{y})-\theta)^2\right], \quad \forall\theta\in\Theta \tag{2.3}$$

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:

$$T(\boldsymbol{y}) = \sum_{i=1}^{n} a_i \boldsymbol{y}_i, \tag{2.4}$$

con $a_i \in \mathbb{R}$.

Definizione 2.7. Uno stimatore lineare corretto $T^*(\cdot)$ del parametro scalare θ viene detto $BLUE$ (Best Linear Unbiased Estimator) se:

$$\mathbf{E}^{\theta}\left[(T^*(\boldsymbol{y}) - \theta)^2\right] \leq \mathbf{E}^{\theta}\left[(T(\boldsymbol{y}) - \theta)^2\right], \quad \forall \theta \in \Theta$$

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 2.5. Siano $\boldsymbol{y}_i$ variabili aleatorie indipendenti, con media m e varianza $\sigma_{\boldsymbol{i}}^2$, $i = 1, \ldots, n$. Si desidera calcolare lo stimatore $BLUE$ di m. Dal momento che lo stimatore deve essere lineare esso avrà la forma descritta nell'eq. (2.4). Affinché $T(\cdot)$ sia non polarizzato, occorre che sia:

$$\mathbf{E}^{\theta}\left[T(\boldsymbol{y})\right] = \mathbf{E}^{\theta}\left[\sum_{i=1}^{n} a_i \boldsymbol{y}_i\right] = \sum_{i=1}^{n} a_i \mathbf{E}^{\theta}\left[\boldsymbol{y}_i\right] = m \sum_{i=1}^{n} a_i = m$$

Per cui, occorre che sia:

$$\sum_{i=1}^{n} a_i = 1 \tag{2.5}$$

Ora, tra tutti gli stimatori della forma (2.4), con i coefficienti a_i che verificano la (2.5), occorre determinare quello a minima varianza. Dal momento che le osservazioni $\boldsymbol{y}_i$ sono indipendenti, la varianza di $T(\boldsymbol{y})$ vale:

$$\mathbf{E}^{\theta}\left[(T(\boldsymbol{y}) - m)^2\right] = \mathbf{E}^{\theta}\left[\left(\sum_{i=1}^{n} a_i \boldsymbol{y}_i - m\right)^2\right] = \sum_{i=1}^{n} a_i^2 \sigma_{\boldsymbol{i}}^2.$$

Per cui, il calcolo dello stimatore BLUE si riduce a risolvere il seguente problema di ottimizzazione vincolata:

$$\min_{a_i} \sum_{i=1}^{n} a_i^2 \sigma_{\boldsymbol{i}}^2$$

s.t.

$$\sum_{i=1}^{n} a_i = 1$$

Si costruisca la Lagrangiana:

$$\mathcal{L}(a_1, \ldots, a_n, \lambda) = \sum_{i=1}^{n} a_i^2 \sigma_{\boldsymbol{i}}^2 + \lambda \left(\sum_{i=1}^{n} a_i - 1 \right)$$

e si imponga che:

$$\frac{\partial \mathcal{L}(a_1, \ldots, a_n, \lambda)}{\partial a_i} = 0, \quad i = 1, \ldots, n \tag{2.6}$$

$$\frac{\partial \mathcal{L}(a_1, \ldots, a_n, \lambda)}{\partial \lambda} = 0. \tag{2.7}$$

Dalla (2.7) si riottiene esattamente il vincolo (2.5), mentre la (2.6) implica che:

$$2 a_i \sigma_{\boldsymbol{i}}^2 + \lambda = 0, \quad i = 1, \ldots, n$$

da cui:

$$\lambda = -\frac{1}{\sum_{i=1}^{n} \frac{1}{2\sigma_{\boldsymbol{i}}^2}} \tag{2.8}$$

$$a_i = \frac{\frac{1}{\sigma_{\boldsymbol{i}}^2}}{\sum_{j=1}^{n} \frac{1}{\sigma_{\boldsymbol{j}}^2}}, \quad i = 1, \ldots, n \tag{2.9}$$

Quindi, lo stimatore BLUE del valor medio m è dato da:

$$\hat{m}_{BLUE} = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_{\boldsymbol{i}}^2}} \sum_{i=1}^{n} \frac{1}{\sigma_{\boldsymbol{i}}^2} \boldsymbol{y}_i \tag{2.10}$$

Si noti che, nel caso in cui tutte le misure abbiano la stessa varianza $\sigma_{\boldsymbol{i}}^2 = \sigma^2$, $\hat{m}_{BLUE}$ si riduce alla media campionaria $\overline{\boldsymbol{y}}$. Questa osservazione consente una immediata interpretazione dello stimatore $BLUE$ appena calcolato. Esso è una generalizzazione della media campionaria, nel caso in cui le misure $\boldsymbol{y}_i$ abbiano diversa accuratezza (ossia, diversa varianza $\sigma_{\boldsymbol{i}}^2$). 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, $\sigma_{\boldsymbol{i}}^2 \to \infty$. Questo corrisponde ad avere una misura $\boldsymbol{y}_i$ totalmente inaffidabile. Allora, il peso $\frac{1}{\sigma_{\boldsymbol{i}}^2}$ che la $\boldsymbol{y}_i$ avrà nella stima $\hat{m}_{BLUE}$ tende a zero. D'altro canto, si supponga di avere una misura $\boldsymbol{y}_i$ infinitamente precisa ($\sigma_{\boldsymbol{i}}^2 \to 0$). In questo caso, il peso $\frac{1}{\sigma_{\boldsymbol{i}}^2}$ dell'osservazione corrispondente è predominante rispetto a tutti gli altri, e lo stimatore BLUE diventa $\hat{m}_{BLUE} \simeq \boldsymbol{y}_i$, cioè circa uguale alla misura infinitamente accurata. △

2.2 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 5. *Sia $T(\cdot)$ uno stimatore non polarizzato del parametro scalare θ, 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:*

$$\mathbf{E}^{\theta}\left[(T(\boldsymbol{y}) - \theta)^2\right] \geq \left[I_n(\theta)\right]^{-1}, \tag{2.11}$$

dove:

$$I_n(\theta) = \mathbf{E}^{\theta}\left[\left(\frac{\partial \ln f_{\boldsymbol{y}}^{\theta}(y)}{\partial \theta}\right)^2\right] \tag{2.12}$$

è detta quantità di informazione di Fisher.

Nel caso in cui θ sia una grandezza vettoriale, la (2.11) diventa:

$$\mathbf{E}^{\theta}\left[(T(\boldsymbol{y}) - \theta)\,(T(\boldsymbol{y}) - \theta)^T\right] \geq \left[I_n(\theta)\right]^{-1},$$

dove la disuguaglianza va intesa in senso matriciale e la $I_n(\theta)$ denota la *matrice di informazione di Fisher*:

$$I_n(\theta) = \mathbf{E}^\theta \left[\left(\frac{\partial \ln f_{\boldsymbol{y}}^\theta(y)}{\partial \theta} \right) \left(\frac{\partial \ln f_{\boldsymbol{y}}^\theta(y)}{\partial \theta} \right)^T \right].$$

La matrice:

$$\mathbf{E}^\theta \left[(T(\boldsymbol{y}) - \theta) (T(\boldsymbol{y}) - \theta)^T \right]$$

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 2.8. Uno stimatore non polarizzato $T(\cdot)$ si dice *efficiente* se la sua varianza raggiunge il limite di Cramér-Rao:

$$\mathbf{E}^\theta \left[(T(\boldsymbol{y}) - \theta)^2 \right] = [I_n(\theta)]^{-1}.$$

Uno stimatore efficiente è caratterizzato dall'avere la piú piccola varianza possibile (fra tutti gli stimatori non polarizzati).

Nel caso di osservazioni $\boldsymbol{y}_i$ 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 2.6. Siano $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ variabili aleatorie indipendenti e identicamente distribuite (i.i.d.) con valor medio $m_{\boldsymbol{y}}$ e varianza $\sigma_{\boldsymbol{y}}^2$. Si è già visto (v. Esempio 2.1) come la media campionaria:

$$\overline{\boldsymbol{y}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}_i$$

sia uno stimatore non polarizzato della media, nonchè consistente (v. Esempio 2.4). Dal Teorema 5, e poiché le osservazioni sono i.i.d., discende che:

$$\mathbf{E}^{\theta}\left[(\overline{\boldsymbol{y}} - m_{\boldsymbol{y}})^2\right] = \frac{\sigma_{\boldsymbol{y}}^2}{n} \geq \left[I_n(\theta)\right]^{-1} = \frac{\left[I_1(\theta)\right]^{-1}}{n}.$$

Supponendo che le $\boldsymbol{y}_i$ abbiano una distribuzione gaussiana:

$$f_{\boldsymbol{y_i}}\left(y_i\right) = \frac{1}{\sqrt{2\pi}\sigma_{\boldsymbol{y}}} e^{-\frac{(y_i - m_{\boldsymbol{y}})^2}{2\sigma_{\boldsymbol{y}}^2}},$$

procediamo con il calcolo della quantità di informazione scalare:

$$I_1(\theta) = \mathbf{E}^{\theta}\left[\left(\frac{\partial \ln f_{\boldsymbol{y_1}}^{\theta}\left(y_1\right)}{\partial\theta}\right)^2\right].$$

Nell'esempio in esame, il parametro incognito da stimare è il valor medio: $\theta = m$. Per cui:

$$\frac{\partial \ln f_{\boldsymbol{y_1}}^{\theta}\left(y_1\right)}{\partial\theta} = \frac{\partial}{\partial m}\left(\ln\frac{1}{\sqrt{2\pi}\sigma_{\boldsymbol{y}}} - \frac{(y_1 - m)^2}{2\sigma_{\boldsymbol{y}}^2}\right)\Bigg|_{m=m_{\boldsymbol{y}}} = \frac{y - m_{\boldsymbol{y}}}{\sigma_{\boldsymbol{y}}^2},$$

e di conseguenza:

$$I_1(\theta) = \mathbf{E}^{\theta}\left[\frac{(y - m_{\boldsymbol{y}})^2}{\sigma_{\boldsymbol{y}}^4}\right] = \frac{1}{\sigma_{\boldsymbol{y}}^2}.$$

Il limite di Cramér-Rao, in questo caso, assume il valore:

$$\left[I_n(\theta)\right]^{-1} = \frac{\left[I_1(\theta)\right]^{-1}}{n} = \frac{\sigma_{\boldsymbol{y}}^2}{n},$$

che coincide proprio con la varianza dello stimatore $\overline{\boldsymbol{y}}$. Per cui, *nel caso di osservazioni i.i.d. gaussiane, la media campionaria è uno stimatore efficiente.* △

2.3 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 2.9. Sia $\boldsymbol{y}$ un vettore di osservazioni, avente densità di probabilità $f_{\boldsymbol{y}}^{\theta}(y)$, dipendente da un parametro incognito $\theta \in \Theta$. Si definisce *verosimiglianza* (o *likelihood*) la funzione:

$$L(\theta|y) = f_{\boldsymbol{y}}^{\theta}(y)\,.$$

Si noti che, una volta osservata per la v.a. $\boldsymbol{y}$ 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 $f_{\boldsymbol{y}}^{\theta}(y)$ 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 2.10. 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 2.1. Supponendo che la $f_{\boldsymbol{y}}^{\theta}(y)$ sia una funzione differenziabile di $\theta = (\theta_1, \ldots, \theta_p) \in \Theta \subseteq \mathbb{R}^p$, con Θ insieme aperto, se $\hat{\theta}$ è un punto di massimo per $L(\theta|y)$, esso deve soddisfare le equazioni:

$$\left.\frac{\partial L(\theta|y)}{\partial \theta_i}\right|_{\theta=\hat{\theta}} = 0, \quad i = 1, \ldots, p \tag{2.13}$$

o, equivalentemente,

$$\left.\frac{\partial \ln L(\theta|y)}{\partial \theta_i}\right|_{\theta=\hat{\theta}} = 0, \quad i = 1, \ldots, p. \tag{2.14}$$

Nella pratica, ogni $\hat{\theta}$ soluzione delle equazioni (2.13) (o delle (2.14)) 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. (2.13) 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 se Θ non è un insieme aperto, in quest'ultimo caso non è detto che valgano le (2.13) o (2.14)).

Esempio 2.7. Siano $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ variabili aleatorie gaussiane, indipendenti con valor medio $m_{\boldsymbol{y}}$ incognito e varianza $\sigma^2_{\boldsymbol{y}}$ nota. Si desidera calcolare la stima a massima verosimiglianza della media.

Dal momento che le misure sono indipendenti, la verosimiglianza vale:

$$L(\theta|y) = f^{\theta}_{\boldsymbol{y}}(y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_{\boldsymbol{y}}} e^{-\frac{(y_i-m)^2}{2\sigma^2_{\boldsymbol{y}}}}$$

In questo caso conviene massimizzare la log-likelihood, che ha la forma:

$$\begin{aligned} \ln L(\theta|y) &= \sum_{i=1}^{n} \left(\ln \frac{1}{\sqrt{2\pi}\sigma_{\boldsymbol{y}}} - \frac{(y_i-m)^2}{2\sigma^2_{\boldsymbol{y}}} \right) \\ &= n \ln \frac{1}{\sqrt{2\pi}\sigma_{\boldsymbol{y}}} - \sum_{i=1}^{n} \frac{(y_i-m)^2}{2\sigma^2_{\boldsymbol{y}}}. \end{aligned}$$

Imponendo la condizione (2.14), si ottiene:

$$\frac{\partial \ln L(\theta|y)}{\partial \theta} = \frac{\partial}{\partial m} \left(n \ln \frac{1}{\sqrt{2\pi}\sigma_{\boldsymbol{y}}} - \sum_{i=1}^{n} \frac{(y_i-m)^2}{2\sigma^2_{\boldsymbol{y}}} \right) \Bigg|_{m=\hat{m}_{ML}} = 0,$$

da cui:

$$\sum_{i=1}^{n} \frac{y_i - \hat{m}_{ML}}{\sigma^2_{\boldsymbol{y}}} = 0,$$

ossia:

$$\hat{m}_{ML} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i.$$

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 2.6) △

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 6. *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 2.8. Siano $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ variabili aleatorie gaussiane, indipendenti con valor medio $m_{\boldsymbol{y}}$ e varianza $\sigma^2_{\boldsymbol{y}}$, entrambi incogniti. Si desidera calcolare la stima a massima verosimiglianza della media e della varianza.

Analogamente a quanto derivato nell'Esempio 2.7, 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 (2.14) diventa:

$$\frac{\partial \ln L(\theta|y)}{\partial \theta_1} = \frac{\partial}{\partial m}\left(n \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_{i=1}^{n} \frac{(y_i - m)^2}{2\sigma^2} \right)\Bigg|_{(m=\hat{m}_{ML}, \sigma^2=\hat{\sigma}^2_{ML})} = 0,$$

$$\frac{\partial \ln L(\theta|y)}{\partial \theta_2} = \frac{\partial}{\partial \sigma^2}\left(n \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_{i=1}^{n} \frac{(y_i - m)^2}{2\sigma^2_{\boldsymbol{y}}} \right)\Bigg|_{(m=\hat{m}_{ML}, \sigma^2=\hat{\sigma}^2_{ML})} = 0.$$

Differenziando rispetto a m e a σ^2 si ottiene:

$$\sum_{i=1}^{n} \frac{y_i - \hat{m}_{ML}}{\hat{\sigma}^2_{ML}} = 0$$
$$-\frac{n}{2\sigma^2_{ML}} + \frac{1}{2\sigma^4_{ML}} \sum_{i=1}^{n} (y_i - \hat{m}_{ML})^2 = 0,$$

da cui:

$$\hat{m}_{ML} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}_i$$
$$\sigma^2_{ML} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{y}_i - m_{\boldsymbol{y}})^2.$$

Sebbene $\mathbf{E}^{\theta}[\hat{m}_{ML}] = m_{\boldsymbol{y}}$ (vedi Esempio 2.1), risulta $\mathbf{E}^{\theta}[\sigma^2_{ML}] = \frac{n-1}{n}\sigma^2_{\boldsymbol{y}}$ (vedi Esempio 2.2). Per cui, in questo caso, lo stimatore di massima verosimiglianza è polarizzato, e quindi non è efficiente. In virtù del Teorema 6, si può concludere che non esiste alcuno stimatore efficiente per il parametro $\theta = (m, \sigma^2)^T$. △

L'esempio precedente mostra come lo stimatore di massima verosimiglianza possa anche essere polarizzato. Tuttavia, oltre alle motivazioni fornite dal Teorema 6, 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:

- asintoticamente corretto;
- consistente;
- asintoticamente efficiente;
- asintoticamente normale.

2.3.1 Rumore additivo

Spesso, capita di dover stimare un parametro incognito θ sulla base di n misure $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^T$ corrotte da rumore additivo. Formalmente, sia:

$$U(\cdot) : \Theta \subseteq \mathbb{R}^p \to \mathbb{R}^n$$

una funzione deterministica del parametro incognito θ. L'obiettivo è stimare θ a partire dalle osservazioni:

$$\boldsymbol{y} = U(\theta) + \boldsymbol{\varepsilon}$$

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 $f_{\boldsymbol{\varepsilon}}(\varepsilon)$.

Sotto queste ipotesi, la verosimiglianza vale:

$$L(\theta|y) = f_{\boldsymbol{y}}^{\theta}(y) = f_{\boldsymbol{\varepsilon}}(y - U(\theta)).$$

Nel caso notevole in cui il rumore di misura abbia una distribuzione gaussiana:

$$f_{\boldsymbol{\varepsilon}}(\varepsilon) = \frac{1}{(2\pi)^{n/2}(\det \Sigma_{\boldsymbol{\varepsilon}})^{1/2}} e^{-\frac{1}{2}\varepsilon^T \Sigma_{\boldsymbol{\varepsilon}}^{-1} \varepsilon}$$

con matrice di covarianza $\Sigma_{\boldsymbol{\varepsilon}}$ nota, la log-likelihood assume la seguente forma:

$$\ln L(\theta|y) = K - \frac{1}{2}(y - U(\theta))^T \Sigma_{\boldsymbol{\varepsilon}}^{-1}(y - U(\theta)),$$

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:

$$\begin{aligned} \hat{\theta}_{ML} &= \arg\max_{\theta} \ln L(\theta|y) \\ &= \arg\min_{\theta} (y - U(\theta))^T \Sigma_{\boldsymbol{\varepsilon}}^{-1}(y - U(\theta)). \end{aligned} \tag{2.15}$$

Dal momento che $U(\cdot)$ è, in generale, una funzione non lineare di θ, la soluzione può essere calcolata tramite metodi numerici.

2.4 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:

$$\boldsymbol{y} = U\theta + \boldsymbol{\varepsilon}, \tag{2.16}$$

in cui U è una matrice $n \times p$.

Definizione 2.11. Sia $\boldsymbol{y}$ un vettore di v.a. legate al parametro incognito θ dalla relazione (2.16). Lo stimatore:

$$T_{LS}(\boldsymbol{y}) = (U^T U)^{-1} U^T \boldsymbol{y} \tag{2.17}$$

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:

$$\frac{\partial}{\partial\theta} \|y - U\theta\|^2 \Big|_{\theta=\hat{\theta}_{LS}} = 2\hat{\theta}_{LS}^T U^T U - 2y^T U = 0,$$

poiché $\frac{\partial x^T A x}{\partial x} = 2x^T A$ e $\frac{\partial Ax}{\partial x} = A$. 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 (2.17).

Sotto l'ipotesi di rumore a media nulla, $\mathbf{E}[\varepsilon] = 0$, lo stimatore ai minimi quadrati è corretto:

$$\begin{aligned}\mathbf{E}^\theta\left[\hat{\theta}_{LS}\right] &= \mathbf{E}^\theta\left[(U^TU)^{-1}U^T\boldsymbol{y}\right] = \mathbf{E}^\theta\left[(U^TU)^{-1}U^T(U\theta+\boldsymbol{\varepsilon})\right] \\ &= \mathbf{E}^\theta\left[\theta + (U^TU)^{-1}U^T\boldsymbol{\varepsilon}\right] = \theta.\end{aligned}$$

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 $\boldsymbol{y}_i$ contribuiscono in egual misura al valore finale della stima.

Definizione 2.12. Sia $\boldsymbol{y}$ un vettore di v.a. aleatorie legate al parametro incognito θ dalla relazione (2.16). Sia $\Sigma_{\boldsymbol{\varepsilon}}$ la matrice di covarianza della v.a. $\boldsymbol{\varepsilon}$. Lo stimatore:

$$T_{GM}(\boldsymbol{y}) = (U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U)^{-1}U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}\boldsymbol{y} \tag{2.18}$$

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:

$$\hat{\theta}_{GM} = \arg\min_\theta (y - U\theta)^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}(y - U\theta).$$

Sotto l'ipotesi di rumore a media nulla, $\mathbf{E}[\varepsilon] = 0$, lo stimatore di Gauss-Markov è corretto:

$$\begin{aligned}\mathbf{E}^\theta\left[\hat{\theta}_{GM}\right] &= \mathbf{E}^\theta\left[(U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U)^{-1}U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}\boldsymbol{y}\right] = \mathbf{E}^\theta\left[(U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U)^{-1}U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}(U\theta+\boldsymbol{\varepsilon})\right] \\ &= \mathbf{E}^\theta\left[\theta + (U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U)^{-1}U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}\boldsymbol{\varepsilon}\right] = \theta.\end{aligned}$$

La varianza di tale stimatore vale:

$$\begin{aligned}
\mathbf{E}^\theta\left[(\hat{\theta}_{GM}-\theta)(\hat{\theta}_{GM}-\theta)^T\right] &= \mathbf{E}^\theta\left[\left((U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U)^{-1}U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}\boldsymbol{y}-\theta\right)(\cdot)^T\right] \\
&= \mathbf{E}^\theta\left[\left((U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U)^{-1}U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}(U\theta+\boldsymbol{\varepsilon})-\theta\right)(\cdot)^T\right] \\
&= \mathbf{E}^\theta\left[\left((U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U)^{-1}U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}\boldsymbol{\varepsilon}\right)(\cdot)^T\right] \\
&= (U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U)^{-1}U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}\Sigma_{\boldsymbol{\varepsilon}}\Sigma_{\boldsymbol{\varepsilon}}^{-1}U(U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U)^{-1} \\
&= (U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U)^{-1}U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U(U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U)^{-1} \\
&= (U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U)^{-1}.
\end{aligned}$$

Si noti come la stima di Gauss-Markov richieda la conoscenza della matrice di covarianza $\Sigma_{\boldsymbol{\varepsilon}}$ del rumore di misura. Utilizzando tale informazione, le osservazioni vengono pesate in maniera inversamente proporzionale alla loro incertezza.

È possibile verificare che gli stimatori calcolati negli Esempi 2.1 e 2.5 corrispondono proprio allo stimatore del valor medio ai minimi quadrati e a quello di Gauss-Markov, rispettivamente, nel caso di osservazioni indipendenti (vedi Esercizio 2.1).

Nel paragrafo precedente si è visto che, nel caso in cui il rumore di misura $\boldsymbol{\varepsilon}$ sia gaussiano, lo stimatore di massima verosimiglianza si calcola risolvendo il problema di ottimizzazione (2.15). Nel caso in cui le osservazioni siano una funzione lineare di θ, come nell'eq. (2.16), tale problema diventa:

$$\hat{\theta}_{ML} = \arg\min_{\theta}(y-U\theta)^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}(y-U\theta). \tag{2.19}$$

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 $\sigma_{\boldsymbol{\varepsilon}}^2$:

$$\boldsymbol{\varepsilon} \sim N(0, \sigma_{\boldsymbol{\varepsilon}}^2 I)$$

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. gaussia-*

ne, lo stimatore a massima verosimiglianza coincide con quello ai minimi quadrati.

2.5 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 $\boldsymbol{x} \in \mathbb{R}^m$, sulla base di osservazioni della variabile aleatoria $\boldsymbol{y} \in \mathbb{R}^n$, conoscendo la densità di probabilità congiunta $f_{\boldsymbol{x},\boldsymbol{y}}(x, y)$.

Occorre, quindi, trovare uno stimatore $\hat{\boldsymbol{x}} = T(\boldsymbol{y})$, dove:

$$T(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$$

Analogamente a quanto fatto per la stima parametrica, per valutare la qualità della stima è necessario definire un opportuno criterio.

Definizione 2.13. Si definisce *funzionale di rischio di Bayes* la quantità:

$$J_r = \mathbf{E}\left[d(\boldsymbol{x}, T(\boldsymbol{y}))\right] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} d(x, T(y)) f_{\boldsymbol{x},\boldsymbol{y}}(x, y)\, \mathrm{d}x\mathrm{d}y.$$

La grandezza $d(\boldsymbol{x}, T(\boldsymbol{y}))$ denota la "distanza" tra la v.a. incognita $\boldsymbol{x}$ e la sua stima $T(\boldsymbol{y})$. Dal momento che tale quantità è a sua volta una variabile aleatoria, ci si propone di trovare lo stimatore $T^*(\cdot)$ che minimizza $d(\boldsymbol{x}, T(\boldsymbol{y}))$ "in media":

$$T^*(\cdot) = \arg\min_{T(\cdot)} J_r.$$

2.5.1 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*:

$$d(\boldsymbol{x}, T(\boldsymbol{y})) = \left\|\boldsymbol{x} - T(\boldsymbol{y})\right\|^2.$$

Definizione 2.14. Si definsce *stimatore a minimo errore quadratico medio (MEQM)* lo stimatore $\hat{\boldsymbol{x}}_{MEQM} = T^*(\cdot)$, dove:

$$T^*(\cdot) = \arg\min_{T(\cdot)} \mathbf{E}\left[\|\boldsymbol{x} - T(\boldsymbol{y})\|^2\right].$$

Si noti che il valore atteso nella precedente definizione viene calcolato rispetto ad entrambi le v.a. $\boldsymbol{x}$ e $\boldsymbol{y}$. Per cui è necessario conoscere la d.d.p. congiunta $f_{\boldsymbol{x},\boldsymbol{y}}(x, y)$.

Teorema 7. *Sia $\boldsymbol{x}$ una v.a. incognita e sia $\boldsymbol{y}$ un vettore di osservazioni. Allora, la stima a minimo errore quadratico medio $\hat{\boldsymbol{x}}_{MEQM}$ di $\boldsymbol{x}$ corrisponde al valore atteso condizionato:*

$$\hat{\boldsymbol{x}}_{MEQM} = \mathbf{E}\left[\boldsymbol{x}|\boldsymbol{y}\right].$$

Il risultato precedente afferma che, qualunque sia il legame fra la v.a. incognita $\boldsymbol{x}$ e le osservazioni $\boldsymbol{y}$, la stima che minimizza l'errore quadratico medio è data dal valore atteso di $\boldsymbol{x}$ condizionato all'aver osservato la realizzazione y per la v.a. $\boldsymbol{y}$:

$$\hat{\boldsymbol{x}}_{MEQM} = \int_{-\infty}^{+\infty} x f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)\mathrm{d}x. \tag{2.20}$$

La stima MEQM presenta altre caratteristiche che la rendono attraente. In particolare, indicata con Q la matrice:

$$Q(\boldsymbol{x}, T(\boldsymbol{y})) = \mathbf{E}\left[(\boldsymbol{x} - T(\boldsymbol{y}))(\boldsymbol{x} - T(\boldsymbol{y}))^T\right],$$

si dimostra che:

- $\hat{\boldsymbol{x}}_{MEQM}$ è lo stimatore che minimizza (in senso matriciale) $Q(\boldsymbol{x}, T(\boldsymbol{y}))$, ovvero:

  $$Q(\boldsymbol{x}, \hat{\boldsymbol{x}}_{MEQM}) \leq Q(\boldsymbol{x}, T(\boldsymbol{y})), \quad \forall T(\boldsymbol{y})$$

  dove la disuguaglianza matriciale $A \leq B$ va intesa nel senso che la matrice $B - A$ è semidefinita positiva;

- $\hat{\boldsymbol{x}}_{MEQM}$ minimizza ogni funzione scalare monotona crescente di $Q(\boldsymbol{x}, T(\boldsymbol{y}))$, 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 $f_{\boldsymbol{x},\boldsymbol{y}}(x,y)$ e il calcolo dell'integrale (2.20).

Esempio 2.9. Si considerino due v.a. $\boldsymbol{x}$ e $\boldsymbol{y}$, la cui d.d.p. congiunta vale:

$$f_{\boldsymbol{x},\boldsymbol{y}}(x,y) = \begin{cases} -\frac{3}{2}x^2 + 2xy & \text{se } 0 \le x \le 1, \quad 1 \le y \le 2 \\ 0 & \text{altrimenti} \end{cases}$$

Si desidera determinare la stima a minimo errore quadratico medio di $\boldsymbol{x}$ basata su un'osservazione di $\boldsymbol{y}$.

Dal Teorema 7, sappiamo che:

$$\hat{\boldsymbol{x}}_{MEQM} = \int_{-\infty}^{+\infty} x f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)\mathrm{d}x.$$

Occorre innanzitutto calcolare:

$$f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) = \frac{f_{\boldsymbol{x},\boldsymbol{y}}(x,y)}{f_{\boldsymbol{y}}(y)}.$$

La densità marginale di $\boldsymbol{y}$ può essere ricavata dalla d.d.p. congiunta:

$$\begin{aligned} f_{\boldsymbol{y}}(y) &= \int_0^1 -\frac{3}{2}x^2 + 2xy\mathrm{d}x \\ &= -\frac{x^3}{2} + yx^2\Big|_{x=0}^{x=1} = y - \frac{1}{2}. \end{aligned}$$

Quindi, la d.d.p. condizionata vale:

$$f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) = \begin{cases} \frac{-\frac{3}{2}x^2+2xy}{y-\frac{1}{2}} & \text{se } 0 \le x \le 1, \quad 1 \le y \le 2 \\ 0 & \text{altrimenti} \end{cases}$$

È possibile, a questo punto, calcolare la stima a minimo errore quadratico medio:

$$\begin{aligned}\hat{\boldsymbol{x}}_{MEQM} &= \int_0^1 x \frac{-\frac{3}{2}x^2 + 2xy}{y - \frac{1}{2}} \mathrm{d}x \\ &= \frac{1}{y - \frac{1}{2}} \left(-\frac{3}{8}x^4 + \frac{2}{3}x^3 y \right) \Big|_{x=0}^{x=1} = \frac{\frac{2}{3}y - \frac{3}{8}}{y - \frac{1}{2}}.\end{aligned}$$

△

2.5.2 Stima lineare a minimo errore quadratico medio

Restringiamo la nostra attenzione alla classe di stimatori lineari:

$$T(\boldsymbol{y}) = A\boldsymbol{y} + b \tag{2.21}$$

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 (2.21) si desidera calcolare quello che minimizza l'errore quadratico medio.

Definizione 2.15. Si definsce *stimatore lineare a minimo errore quadratico medio (LMEQM)* lo stimatore $\hat{\boldsymbol{x}}_{LMEQM} = A^*\boldsymbol{y} + b^*$, dove:

$$A^*, b^* = \arg\min_{A,b} \mathbf{E}\left[\|\boldsymbol{x} - A\boldsymbol{y} - b\|^2 \right]. \tag{2.22}$$

Teorema 8. *Sia $\boldsymbol{x}$ una v.a incognita e sia $\boldsymbol{y}$ un vettore di osservazioni, tali che*

$$\mathbf{E}[\boldsymbol{x}] = m_{\boldsymbol{x}}, \qquad \mathbf{E}[\boldsymbol{y}] = m_{\boldsymbol{y}}$$

$$\mathbf{E}\left[\begin{pmatrix} \boldsymbol{x} - m_{\boldsymbol{x}} \\ \boldsymbol{y} - m_{\boldsymbol{y}} \end{pmatrix} \begin{pmatrix} \boldsymbol{x} - m_{\boldsymbol{x}} \\ \boldsymbol{y} - m_{\boldsymbol{y}} \end{pmatrix}^T \right] = \begin{pmatrix} R_{\boldsymbol{x}} & R_{\boldsymbol{xy}} \\ R_{\boldsymbol{xy}}^T & R_{\boldsymbol{y}} \end{pmatrix}$$

Allora, la stima lineare a minimo errore quadratico medio $\hat{\boldsymbol{x}}_{LMEQM}$ di $\boldsymbol{x}$ vale:

$$\hat{\boldsymbol{x}}_{LMEQM} = m_{\boldsymbol{x}} + R_{\boldsymbol{xy}} R_{\boldsymbol{y}}^{-1} (\boldsymbol{y} - m_{\boldsymbol{y}}),$$

ossia:

$$A^* = R_{\boldsymbol{xy}} R_{\boldsymbol{y}}^{-1},$$
$$b^* = m_{\boldsymbol{x}} - R_{\boldsymbol{xy}} R_{\boldsymbol{y}}^{-1} m_{\boldsymbol{y}}.$$

Dimostrazione

Innanzitutto si osservi che il costo da minimizzare equivale alla traccia della matrice $\mathbf{E}\left[(\boldsymbol{x} - A\boldsymbol{y} - b)(\boldsymbol{x} - A\boldsymbol{y} - b)^T\right]$:

$$\mathbf{E}\left[\|\boldsymbol{x} - A\boldsymbol{y} - b\|^2\right] = \operatorname{tr}\left(\mathbf{E}\left[(\boldsymbol{x} - A\boldsymbol{y} - b)(\boldsymbol{x} - A\boldsymbol{y} - b)^T\right]\right).$$

Inoltre, poiché la traccia è una funzione monotona crescente, se si è in grado di determinare A^*, b^* tali che:

$$\mathbf{E}\left[(\boldsymbol{x} - A^*\boldsymbol{y} - b^*)(\boldsymbol{x} - A^*\boldsymbol{y} - b^*)^T\right] \leq \mathbf{E}\left[(\boldsymbol{x} - A\boldsymbol{y} - b)(\boldsymbol{x} - A\boldsymbol{y} - b)^T\right], \ \forall A, b \tag{2.23}$$

allora sicuramente A^*, b^* sono soluzione del problema (2.22). Per cui, indicando con $\tilde{\boldsymbol{x}} = \boldsymbol{x} - A\boldsymbol{y} - b$ l'errore di stima, si calcoli:

$$\begin{aligned}
\mathbf{E}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T\right] &= \mathbf{E}\left[(\boldsymbol{x} - m_{\boldsymbol{x}} - A(\boldsymbol{y} - m_{\boldsymbol{y}}) + m_{\boldsymbol{x}} - Am_{\boldsymbol{y}} - b)\right. \\
&\qquad \left. \times (\boldsymbol{x} - m_{\boldsymbol{x}} - A(\boldsymbol{y} - m_{\boldsymbol{y}}) + m_{\boldsymbol{x}} - Am_{\boldsymbol{y}} - b)^T\right] \\
&= R_{\boldsymbol{x}} + AR_{\boldsymbol{y}}A^T - R_{\boldsymbol{xy}}A^T - AR_{\boldsymbol{yx}} \\
&\qquad + (m_{\boldsymbol{x}} - Am_{\boldsymbol{y}} - b)(m_{\boldsymbol{x}} - Am_{\boldsymbol{y}} - b)^T \\
&= R_{\boldsymbol{x}} + AR_{\boldsymbol{y}}A^T - R_{\boldsymbol{xy}}A^T - AR_{\boldsymbol{xy}}^T + R_{\boldsymbol{xy}}R_{\boldsymbol{y}}^{-1}R_{\boldsymbol{xy}}^T - R_{\boldsymbol{xy}}R_{\boldsymbol{y}}^{-1}R_{\boldsymbol{xy}}^T \\
&\qquad + (m_{\boldsymbol{x}} - Am_{\boldsymbol{y}} - b)(m_{\boldsymbol{x}} - Am_{\boldsymbol{y}} - b)^T \\
&= R_{\boldsymbol{x}} - R_{\boldsymbol{xy}}R_{\boldsymbol{y}}^{-1}R_{\boldsymbol{xy}}^T + \left(R_{\boldsymbol{xy}}R_{\boldsymbol{y}}^{-1} - A\right)R_{\boldsymbol{y}}\left(R_{\boldsymbol{xy}}R_{\boldsymbol{y}}^{-1} - A\right)^T \\
&\qquad + (m_{\boldsymbol{x}} - Am_{\boldsymbol{y}} - b)(m_{\boldsymbol{x}} - Am_{\boldsymbol{y}} - b)^T.
\end{aligned}$$

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 (2.23):

$$A^* = R_{\boldsymbol{xy}}R_{\boldsymbol{y}}^{-1};$$
$$b^* = m_{\boldsymbol{x}} - Am_{\boldsymbol{y}} = m_{\boldsymbol{x}} - R_{\boldsymbol{xy}}R_{\boldsymbol{y}}^{-1}m_{\boldsymbol{y}}.$$

□

La stima LMEQM è non polarizzata, in quanto il valore atteso dell'errore di stima è zero:

$$\begin{aligned}\mathbf{E}[\tilde{\boldsymbol{x}}] &= \mathbf{E}[\boldsymbol{x} - \hat{\boldsymbol{x}}_{LMEQM}] = m_{\boldsymbol{x}} - \mathbf{E}\left[m_{\boldsymbol{x}} + R_{\boldsymbol{xy}}R_{\boldsymbol{y}}^{-1}(\boldsymbol{y} - m_{\boldsymbol{y}})\right] \\ &= m_{\boldsymbol{x}} - m_{\boldsymbol{x}} + R_{\boldsymbol{xy}}R_{\boldsymbol{y}}^{-1}\mathbf{E}[\boldsymbol{y} - m_{\boldsymbol{y}}] = 0.\end{aligned}$$

La varianza dell'errore di stima vale:

$$\begin{aligned}\mathbf{E}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T\right] &= \mathbf{E}\left[(\boldsymbol{x} - m_{\boldsymbol{x}} - R_{\boldsymbol{xy}}R_{\boldsymbol{y}}^{-1}(\boldsymbol{y} - m_{\boldsymbol{y}}))(\boldsymbol{x} - m_{\boldsymbol{x}} - R_{\boldsymbol{xy}}R_{\boldsymbol{y}}^{-1}(\boldsymbol{y} - m_{\boldsymbol{y}}))^T\right] \\ &= R_{\boldsymbol{x}} + R_{\boldsymbol{xy}}R_{\boldsymbol{y}}^{-1}R_{\boldsymbol{y}}R_{\boldsymbol{y}}^{-1}R_{\boldsymbol{xy}}^T - R_{\boldsymbol{xy}}R_{\boldsymbol{y}}^{-1}R_{\boldsymbol{xy}}^T - R_{\boldsymbol{xy}}R_{\boldsymbol{y}}^{-1}R_{\boldsymbol{xy}}^T \\ &= R_{\boldsymbol{x}} - R_{\boldsymbol{xy}}R_{\boldsymbol{y}}^{-1}R_{\boldsymbol{xy}}^T.\end{aligned}$$

È bene notare che, interpretando $R_{\boldsymbol{x}}$ come l'incertezza *a priori* sulla v.a. $\boldsymbol{x}$, $R_{\boldsymbol{x}} - R_{\boldsymbol{xy}}R_{\boldsymbol{y}}^{-1}R_{\boldsymbol{xy}}^T$ rappresenta la nuova incertezza su $\boldsymbol{x}$ dopo aver effettuato la misura $\boldsymbol{y}$. Dal momento che la matrice $R_{\boldsymbol{xy}}R_{\boldsymbol{y}}^{-1}R_{\boldsymbol{xy}}^T$ è sempre semidefinita positiva, l'effetto delle misure è quello di ridurre l'incertezza su $\boldsymbol{x}$. Inoltre, tale riduzione è tanto maggiore quanto più è "grande" $R_{\boldsymbol{xy}}$, cioè quanto più la misura $\boldsymbol{y}$ è legata alla v.a. incognita $\boldsymbol{x}$.

Si noti che, ai fini del calcolo della stima lineare a minimo errore quadratico medio, non è necessario conoscere la densità di probabilità congiunta $f_{\boldsymbol{x},\boldsymbol{y}}(x, y)$, ma solo le statistiche del primo e secondo ordine: $m_{\boldsymbol{x}}, m_{\boldsymbol{y}}, R_{\boldsymbol{x}}, R_{\boldsymbol{y}}$, $R_{\boldsymbol{xy}}$.

Inoltre, l'errore di stima $\tilde{\boldsymbol{x}}$ è scorrelato dalle osservazioni $\boldsymbol{y}$:

$$\begin{aligned}\mathbf{E}\left[\tilde{\boldsymbol{x}}\boldsymbol{y}^T\right] &= \mathbf{E}\left[\left(\boldsymbol{x} - m_{\boldsymbol{x}} - R_{\boldsymbol{xy}}R_{\boldsymbol{y}}^{-1}(\boldsymbol{y} - m_{\boldsymbol{y}})\right)\boldsymbol{y}^T\right] \\ &= R_{\boldsymbol{xy}} - R_{\boldsymbol{xy}}R_{\boldsymbol{y}}^{-1}R_{\boldsymbol{y}} = 0.\end{aligned}$$

Infine, in virtù del Teorema 3 del Capitolo 1, nel caso in cui le v.a. $\boldsymbol{x}, \boldsymbol{y}$ siano congiuntamente gaussiane, con media e matrice di covarianza definite come nel Teorema 8, il valore atteso di $\boldsymbol{x}$ condizionato a $\boldsymbol{y}$ vale proprio:

$$\mathbf{E}[\boldsymbol{x}|\boldsymbol{y}] = m_{\boldsymbol{x}} + R_{\boldsymbol{xy}}R_{\boldsymbol{y}}^{-1}(\boldsymbol{y} - m_{\boldsymbol{y}}).$$

Quindi si può concludere che, *se le v.a.* $\boldsymbol{x}, \boldsymbol{y}$ *sono congiuntamente gaussiane, la stima a minimo errore quadratico medio coincide con la stima lineare a minimo errore quadratico medio:* $\hat{\boldsymbol{x}}_{MEQM} = \hat{\boldsymbol{x}}_{LMEQM}$. In altre parole, nel caso gaussiano la stima MEQM è una funzione lineare della v.a. osservata $\boldsymbol{y}$.

Esempio 2.10. Siano $\boldsymbol{y}_1, \boldsymbol{y}_2$ due osservazioni rumorose della v.a. scalare $\boldsymbol{x}$ avente media $m_{\boldsymbol{x}}$ e varianza $\sigma^2_{\boldsymbol{x}}$:

$$\boldsymbol{y}_1 = \boldsymbol{x} + \boldsymbol{\varepsilon}_1$$
$$\boldsymbol{y}_2 = \boldsymbol{x} + \boldsymbol{\varepsilon}_2$$

Siano $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2$ due v.a. indipendenti, a media nulla e varianza $\sigma^2_{\mathbf{1}}, \sigma^2_{\mathbf{2}}$, rispettivamente. Sotto l'ipotesi che $\boldsymbol{x}$ e $\boldsymbol{\varepsilon}_i$ siano indipendenti, si desidera calcolare la stima lineare a minimo errore quadratico medio.

Definendo le v.a. vettoriali $\boldsymbol{y} = (\boldsymbol{y}_1\ \boldsymbol{y}_2)^T$ e $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1\ \boldsymbol{\varepsilon}_2)^T$, l'equazione delle misure può essere riscritta nella forma matriciale:

$$\boldsymbol{y} = U\boldsymbol{x} + \boldsymbol{\varepsilon},$$

dove $U = (1\ 1)^T$.

Innanzitutto si calcoli il valor medio di $\boldsymbol{y}$:

$$\mathbf{E}[\boldsymbol{y}] = \mathbf{E}[U\boldsymbol{x} + \boldsymbol{\varepsilon}] = Um_{\boldsymbol{x}}$$

Per determinare la sitma $\hat{\boldsymbol{x}}_{LMEQM}$ occorre calcolare le matrici di covarianza $R_{\boldsymbol{xy}}, R_{\boldsymbol{y}}$:

$$R_{\boldsymbol{xy}} = \mathbf{E}\left[(\boldsymbol{x} - m_{\boldsymbol{x}})\left(U\big(x - m_{\boldsymbol{x}}\big) + \boldsymbol{\varepsilon}\right)^T\right] = \sigma^2_{\boldsymbol{x}} U^T,$$

poiché $\boldsymbol{x}$ e $\boldsymbol{\varepsilon}$ sono incorrelate.

$$\begin{aligned} R_{\boldsymbol{y}} &= \mathbf{E}\left[\left(U\big(x - m_{\boldsymbol{x}}\big) + \boldsymbol{\varepsilon}\right)\left(U\big(x - m_{\boldsymbol{x}}\big) + \boldsymbol{\varepsilon}\right)^T\right] \\ &= U\sigma^2_{\boldsymbol{x}} U^T + R_{\boldsymbol{\varepsilon}}, \end{aligned}$$

dove:

$$R_{\boldsymbol{\varepsilon}} = \begin{pmatrix} \sigma^2_{\mathbf{1}} & 0 \\ 0 & \sigma^2_{\mathbf{1}} \end{pmatrix}$$

denota la matrice di covarianza della v.a. $\boldsymbol{\varepsilon}$. Infine, si calcoli l'inversa della matrice di covarianza delle misure:

$$\begin{aligned} R_{\boldsymbol{y}}^{-1} &= \left[\sigma_{\boldsymbol{x}}^2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \begin{pmatrix} \sigma_{\mathbf{1}}^2 & 0 \\ 0 & \sigma_{\mathbf{2}}^2 \end{pmatrix}\right]^{-1} \\ &= \begin{pmatrix} \sigma_{\boldsymbol{x}}^2 + \sigma_{\mathbf{1}}^2 & \sigma_{\boldsymbol{x}}^2 \\ \sigma_{\boldsymbol{x}}^2 & \sigma_{\boldsymbol{x}}^2 + \sigma_{\mathbf{2}}^2 \end{pmatrix}^{-1} \\ &= \frac{1}{\sigma_{\boldsymbol{x}}^2(\sigma_{\mathbf{1}}^2 + \sigma_{\mathbf{2}}^2) + \sigma_{\mathbf{1}}^2\sigma_{\mathbf{2}}^2} \begin{pmatrix} \sigma_{\boldsymbol{x}}^2 + \sigma_{\mathbf{2}}^2 & -\sigma_{\boldsymbol{x}}^2 \\ -\sigma_{\boldsymbol{x}}^2 & \sigma_{\boldsymbol{x}}^2 + \sigma_{\mathbf{1}}^2 \end{pmatrix}. \end{aligned}$$

Quindi, la stima lineare a minimo errore quadratico medio vale:

$$\begin{aligned} \hat{\boldsymbol{x}}_{LMEQM} &= m_{\boldsymbol{x}} + R_{\boldsymbol{xy}} R_{\boldsymbol{y}}^{-1}(\boldsymbol{y} - U m_{\boldsymbol{x}}) \\ &= m_{\boldsymbol{x}} + \sigma_{\boldsymbol{x}}^2 U^T R_{\boldsymbol{y}}^{-1}(\boldsymbol{y} - U m_{\boldsymbol{x}}) \\ &= m_{\boldsymbol{x}} + \frac{\sigma_{\boldsymbol{x}}^2}{\sigma_{\boldsymbol{x}}^2(\sigma_{\mathbf{1}}^2 + \sigma_{\mathbf{2}}^2) + \sigma_{\mathbf{1}}^2\sigma_{\mathbf{2}}^2} (1\ 1) \begin{pmatrix} \sigma_{\boldsymbol{x}}^2 + \sigma_{\mathbf{2}}^2 & -\sigma_{\boldsymbol{x}}^2 \\ -\sigma_{\boldsymbol{x}}^2 & \sigma_{\boldsymbol{x}}^2 + \sigma_{\mathbf{1}}^2 \end{pmatrix} \begin{pmatrix} \boldsymbol{y}_1 - m_{\boldsymbol{x}} \\ \boldsymbol{y}_2 - m_{\boldsymbol{x}} \end{pmatrix} \\ &= m_{\boldsymbol{x}} + \frac{1}{\sigma_{\mathbf{1}}^2 + \sigma_{\mathbf{2}}^2 + \frac{\sigma_{\mathbf{1}}^2\sigma_{\mathbf{2}}^2}{\sigma_{\boldsymbol{x}}^2}} (\sigma_{\mathbf{2}}^2\ \sigma_{\mathbf{1}}^2) \begin{pmatrix} \boldsymbol{y}_1 - m_{\boldsymbol{x}} \\ \boldsymbol{y}_2 - m_{\boldsymbol{x}} \end{pmatrix} \\ &= m_{\boldsymbol{x}} + \frac{\sigma_{\mathbf{2}}^2 \boldsymbol{y}_1 + \sigma_{\mathbf{1}}^2 \boldsymbol{y}_2 - m_{\boldsymbol{x}}(\sigma_{\mathbf{1}}^2 + \sigma_{\mathbf{2}}^2)}{\sigma_{\mathbf{1}}^2 + \sigma_{\mathbf{2}}^2 + \frac{\sigma_{\mathbf{1}}^2\sigma_{\mathbf{2}}^2}{\sigma_{\boldsymbol{x}}^2}} \\ &= \frac{\frac{m_{\boldsymbol{x}}\sigma_{\mathbf{1}}^2\sigma_{\mathbf{2}}^2}{\sigma_{\boldsymbol{x}}^2} + \sigma_{\mathbf{2}}^2 \boldsymbol{y}_1 + \sigma_{\mathbf{1}}^2 \boldsymbol{y}_2}{\sigma_{\mathbf{1}}^2 + \sigma_{\mathbf{2}}^2 + \frac{\sigma_{\mathbf{1}}^2\sigma_{\mathbf{2}}^2}{\sigma_{\boldsymbol{x}}^2}} = \frac{\frac{m_{\boldsymbol{x}}}{\sigma_{\boldsymbol{x}}^2} + \frac{1}{\sigma_{\mathbf{1}}^2}\boldsymbol{y}_1 + \frac{1}{\sigma_{\mathbf{2}}^2}\boldsymbol{y}_2}{\frac{1}{\sigma_{\boldsymbol{x}}^2} + \frac{\sigma_{\mathbf{1}}^2 + \sigma_{\mathbf{2}}^2}{\sigma_{\mathbf{1}}^2\sigma_{\mathbf{2}}^2}} \\ &= \frac{\frac{m_{\boldsymbol{x}}}{\sigma_{\boldsymbol{x}}^2} + \frac{1}{\sigma_{\mathbf{1}}^2}\boldsymbol{y}_1 + \frac{1}{\sigma_{\mathbf{2}}^2}\boldsymbol{y}_2}{\frac{1}{\sigma_{\boldsymbol{x}}^2} + \frac{1}{\sigma_{\mathbf{1}}^2} + \frac{1}{\sigma_{\mathbf{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. $\boldsymbol{x}$ (vale a dire il valor medio $m_{\boldsymbol{x}}$ e la varianza $\sigma_{\boldsymbol{x}}^2$), viene utilizzata come una misura aggiuntiva. In particolare, è interessante notare che se $\sigma_x^2 \to +\infty$ (ovvero l'informazione a priori su $\boldsymbol{x}$ è del tutto inattendibile), la stima $\hat{\boldsymbol{x}}_{LMEQM}$ tende a coincidere con la stima di Gauss-Markov della media m_x.

2.6 Esercizi

2.1. Verificare che, negli Esempi 2.1 e 2.5, assumendo che le variabili $\boldsymbol{y}_i$ siano indipendenti, gli stimatori del valor medio $\overline{\boldsymbol{y}}$ in (2.2) e $\hat{m}_{BLUE}$ in (2.10) corrispondono rispettivamente allo stimatore ai minimi quadrati e a quello di Gauss-Markov.

2.2. Siano $\boldsymbol{d}_1$, $\boldsymbol{d}_2$ 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 $\boldsymbol{d}_1$, $\boldsymbol{d}_2$. Determinare la stima di massima verosimiglianza di θ.

2.3. Siano $\boldsymbol{d}_1$, $\boldsymbol{d}_2$ variabili aleatorie gaussiane indipendenti, tali che:

$$\mathbf{E}\left[\boldsymbol{d}_1\right] = m, \quad \mathbf{E}\left[\boldsymbol{d}_2\right] = 3m, \quad \mathbf{E}\left[(\boldsymbol{d}_1 - m)^2\right] = 2, \quad \mathbf{E}\left[(\boldsymbol{d}_2 - 3m)^2\right] = 4$$

Siano δ_1, δ_2 le osservazioni disponibili di $\boldsymbol{d}_1$, $\boldsymbol{d}_2$. 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*?).

2.4. Sulla grandezza incognita x sono effettuate due misure:

$$\boldsymbol{y}_1 = x + \boldsymbol{d}_1$$
$$\boldsymbol{y}_2 = 2x + \boldsymbol{d}_2$$

dove $\boldsymbol{d}_1$ e $\boldsymbol{d}_2$ 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. Dire se la stima ottenuta è non polarizzata.

2.5. Si consideri per $\theta \in [-2, 2]$ la funzione definita da

$$f^\theta(x) = \begin{cases} \theta x + 1 - \dfrac{\theta}{2} & \text{se } x \in [0, 1] \\ 0 & \text{altrimenti} \end{cases}$$

a. Mostrare che per ogni $\theta \in [-2, 2]$ f^θ è una densità di probabilità.

b. Sia $\boldsymbol{y}$ una variabile aleatoria di densità f^θ. Calcolare in funzione di θ media e varianza di $\boldsymbol{y}$.

c. Sia $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ un campione di dimensione n della densità f^θ e si ponga

$$T(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = 12 \left(\frac{1}{n} \sum_{k=1}^{n} \boldsymbol{y}_k - \frac{1}{2} \right)$$

Mostrare che T è uno stimatore non polarizzato di θ.

2.6. Siano a e b due grandezze incognite, relativamente alle quali sono disponibili tre diverse misure:

$$\begin{aligned} \boldsymbol{y}_1 &= a + \boldsymbol{v}_1 \\ \boldsymbol{y}_2 &= b + \boldsymbol{v}_2 \\ \boldsymbol{y}_3 &= a + b + \boldsymbol{v}_3 \end{aligned}$$

dove $\boldsymbol{v}_i$, $i = 1, 2, 3$, sono variabili aleatorie indipendenti, a media nulla. Assumendo $\mathbf{E}[\boldsymbol{v}_1^2] = \mathbf{E}[\boldsymbol{v}_3^2] = 1$ e $\mathbf{E}[\boldsymbol{v}_2^2] = \frac{1}{2}$, 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 $\boldsymbol{y}_3$ non fosse disponibile. Come si modifica la varianza dell'errore di stima?

2.7. Si considerino due variabili aleatorie $\boldsymbol{x}$ e $\boldsymbol{y}$, la cui d.d.p. congiunta è

$$f_{\boldsymbol{x},\boldsymbol{y}}(x,y) = \begin{cases} -\frac{3}{2}x^2 + 2xy & 0 \le x \le 1,\ 1 \le y \le 2 \\ 0 & \text{altrimenti} \end{cases}$$

Si vuole determinare la stima $\hat{\boldsymbol{x}}_{LMEQM}$ di $\boldsymbol{x}$, basata su una osservazione di $\boldsymbol{y}$.

Riportare in un grafico l'andamento della stima $\hat{\boldsymbol{x}}_{LMEQM}$ calcolata sopra e della stima $\hat{\boldsymbol{x}}_{MEQM}$ calcolata nell'Esempio 2.9, al variare di y (valore osservato della v.a. $\boldsymbol{y}$). Calcolare inoltre il valore atteso delle stime trovate e confrontarlo con il valor medio a priori $\mathbf{E}[\boldsymbol{x}]$.

2.8. Siano $\boldsymbol{x}$ e $\boldsymbol{y}$ due variabili aleatorie con densità di probabilità congiunta

$$f_{\boldsymbol{x},\boldsymbol{y}}(x,y) = \begin{cases} \frac{1}{12}(x+y)e^{-y} & 0 \le x \le 4,\quad y \ge 0 \\ 0 & \text{altrimenti} \end{cases}$$

Si assuma di disporre di una osservazione y della v.a. $\boldsymbol{y}$.

a. Determinare le stime $\hat{\boldsymbol{x}}_{MEQM}$ e $\hat{\boldsymbol{x}}_{LMEQM}$ di $\boldsymbol{x}$, e plottarle in funzione del valore osservato y della v.a. $\boldsymbol{y}$.

b. Calcolare l'errore quadratico medio delle stime ottenute al punto *a* [Suggerimento: utilizzare MATLAB per il calcolo degli integrali].

2.9. Sia X una grandezza incognita di cui è disponibile una misura

$$\boldsymbol{y} = \ln\left(\frac{1}{X}\right) + \boldsymbol{v}$$

con $\boldsymbol{v}$ v.a., la cui d.d.p. è pari a $f_{\boldsymbol{v}}(v) = \begin{cases} e^{-v} & v \ge 0 \\ 0 & v < 0 \end{cases}$.

a) Determinare la stima a massima verosimiglianza di X.

b) Assumendo ora che X sia una v.a., con pdf

$$f_X(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{altrimenti} \end{cases},$$

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.

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.

Rohatgi, V. K. and A. K. Md. E. Saleh (2001). *An introduction to probability and statistics - 2nd edition*. Wiley Interscience.

Zappa, G. (1999). *Appunti del corso di "Modelllistica e Identificazione"*. Università di Siena.