UNIVERSITÀ DEGLI STUDI DI BERGAMO

Dipartimento di Ingegneria Gestionale, dell'Informazione e della Produzione

IDENTIFICAZIONE DEI MODELLI E ANALISI DEI DATI (IMAD)

Corso di Laurea Magistrale in INGEGNERIA INFORMATICA

Lezione 7: Fondamenti di stima Bayesiana

SPEAKER
Prof. Mirko Mazzoleni

PLACE
Università degli Studi di Bergamo

Syllabus

Parte I: sistemi statici

1. **Richiami di statistica**
2. **Teoria della stima**
   - 2.1 Proprietà degli stimatori
3. **Stima a minimi quadrati**
   - 3.1 Stima di modelli lineari
   - 3.2 Algoritmo del gradient descent
4. **Stima a massima verosimiglianza**
   - 4.1 Proprietà della stima
   - 4.2 Stima di modelli lineari
5. **Regressione logistica**
   - 5.1 Stima di un modello di regressione logistica
6. **Fondamenti di machine learning**
   - 6.1 Bias-Variance tradeoff
   - 6.2 Overfitting
   - 6.3 Regolarizzazione
   - 6.4 Validazione
7. **Cenni di stima Bayesiana**
   - 7.1 Probabilità congiunte, marginali e condizionate
   - 7.2 Connessione con Filtro di Kalman

IMAD

Parte I: sistemi statici

Stima parametrica $\hat{\theta}$

- **θ deterministico**
  - ***NO assunzioni su ddp dei dati***
    - ✓ Stima parametri popolazione
    - ✓ Stima modello lineare: minimi quadrati
  - ***SI assunzioni su ddp dei dati***
    - ✓ Stima massima verosimiglianza parametri popolazione
    - ✓ Stima modello lineare: massima verosimiglianza
    - ✓ Regressione logistica
- **θ variabile casuale**
  - ***SI assunzioni su ddp dei dati***
    - ✓ Stima Bayesiana

Machine learning

Parte II: sistemi dinamici

Stima parametrica $\hat{\theta}$

- **θ deterministico**
  - ***NO assunzioni su ddp dei dati***
    - ✓ Modelli lineari di pss
    - ✓ Predizione
    - ✓ Identificazione
    - ✓ Persistente eccitazione
    - ✓ Analisi asintotica metodi PEM
    - ✓ Analisi incertezza stima (numero dati finito)
    - ✓ Valutazione del modello

Outline

1. Probabilità congiunte, condizionate, marginali
2. Introduzione alla stima Bayesiana
3. Stima ottima
4. Stima ottima lineare

Outline

1. **Probabilità congiunte, condizionate, marginali**
2. Introduzione alla stima Bayesiana
3. Stima ottima
4. Stima ottima lineare

Probabilità congiunte, marginali, condizionate

Supponiamo di avere **due variabili casuali discrete** e **binarie** a e b. Definiamo:

Distribuzione di probabilità congiunta

$P(a,b)$	$a=0$	$a=1$
$b=0$	0.06	0.24
$b=1$	0.28	0.42

$P(a=1,b=0)=0.24$

$P(a=0,b=1)=0.28$

$P(a,b)$: probabilità che sia a che b assumino un valore specifico

$$\sum_{a=0}^{1}\sum_{b=0}^{1} p(a,b) = 1$$

$$P(a,b) = P(b,a)$$

Probabilità congiunte, marginali, condizionate

Distribuzione di probabilità marginale

La **distribuzione marginale** è la distribuzione di probabilità di un **sottoinsieme** di variabili casuali

Nel nostro esempio, siccome abbiamo **due variabili casuali** a e b, avremo **due marginali**, ovvero $P(a)$ e $P(b)$. Se avessimo tre v.c discrete a, b, c avremmo le marginali $P(a), P(b), P(c), P(a, b), P(a, c), P(b, c)$

Nel caso di v.c. discrete, la distribuzione marginale è ottenuta «marginando» (ovvero, **sommando**) rispetto alle variabili che **non sono di interesse**. Nel caso di v.c. continue, si deve integrare anziché sommare

Probabilità congiunte, marginali, condizionate

Proviamo a calcolare la distribuzione marginale $P(b)$ partendo dalla distribuzione congiunta $P(a,b)$

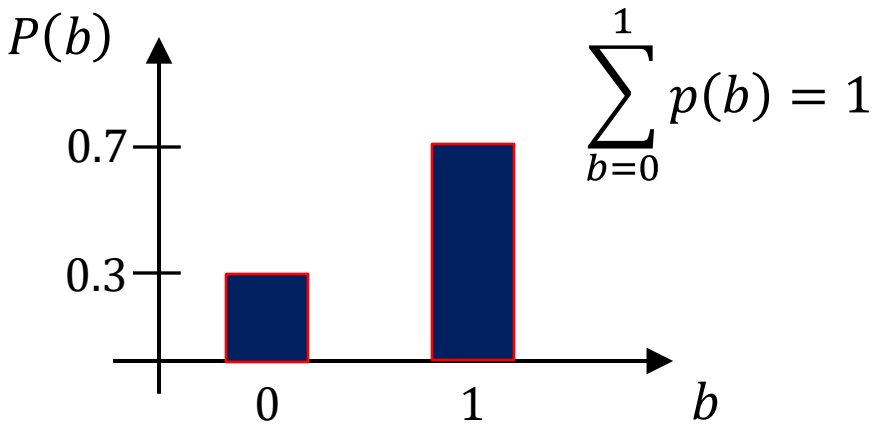
Non mi interessa che valore abbia a, l'importante è che $b = 0$

$P(a,b)$

b \ a	0	1
0	0.06	0.24
1	0.28	0.42

$$P(b=0) = P(a=0, b=0) + P(a=1, b=0) = 0.3$$

$$P(b=1) = 0.7$$



Probabilità congiunte, marginali, condizionate

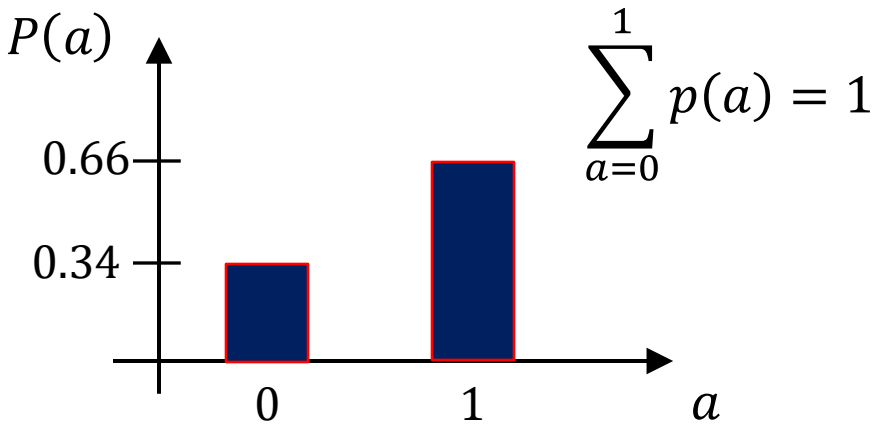
Proviamo a calcolare la distribuzione marginale $P(a)$ partendo dalla distribuzione congiunta $P(a,b)$

$P(a,b)$	$a=0$	$a=1$
$b=0$	0.06	0.24
$b=1$	0.28	0.42

$$P(a=0) = P(a=0,b=0) + P(a=0,b=1) = 0.34$$

$$P(a=1) = P(a=1,b=0) + P(a=1,b=1) = 0.66$$

$P(a=0) = 0.34 \qquad P(a=1) = 0.66$



Probabilità congiunte, marginali, condizionate

Distribuzione di probabilità condizionata

La **distribuzione condizionata** indica come la probabilità si **ridistribuisce** dato che si restringe la popolazione ad un particolare sottoinsieme

Esempio: siano date N persone, dove N_A è il numero di persone con capelli lunghi e N_B è il numero di persone che ascoltano i Black Sabbath. Definiamo gli eventi A e B come:

A: persone con capelli lunghi ⇨ $$P(A) = \frac{N_A}{N} = \frac{\#\text{ persone con capelli lunghi}}{\#\text{ totale di persone}}$$

B: persone che ascoltano i Black Sabbath ⇨ $$P(B) = \frac{N_B}{N} = \frac{\#\text{ persone che ascoltano i Black Sabbath}}{\#\text{ totale di persone}}$$

Probabilità congiunte, marginali, condizionate

Consideriamo **solo la popolazione che ascolta i Black Sabbath**, con $N_B < N$ persone

La probabilità che una persona **scelta a caso da questa popolazione abbia i capelli lunghi** è

$$P(A|B) = \frac{N_{AB}}{N_B} = \frac{\#\text{ persone con capelli lunghi e che ascoltano i Sabbath}}{\#\text{ persone che ascoltano i Sabbath}}$$

Abbiamo ristretto la popolazione da N a N_B, e quindi la **probabilità si è ridistribuita.** Prima avevamo $P(A)$, adesso abbiamo $P(A|B)$

$P(A|B)$ è chiamata **probabilità condizionata** (condizionata al fatto che le persone ascoltino i Black Sabbath)

Probabilità congiunte, marginali, condizionate

La probabilità di selezionare una persona con capelli lunghi che ascolti **anche** i Black Sabbath è la **probabilità congiunta** $P(A,B)$

$$P(A,B) = \frac{N_{AB}}{N} = \frac{\#\text{ persone con capelli lunghi e che ascoltano i Sabbath}}{\#\text{ totale di persone}}$$

Posso quindi esprimere $P(A|B)$ come

$$P(A|B) = \frac{N_{AB}/N}{N_B/N} = \frac{P(A,B)}{P(B)}$$

$P(B)$ è una marginale. E' la probabilità che una persona ascolti i Black Sabbath, indipendentemente dalla lunghezza dei capelli

Probabilità congiunte, marginali, condizionate

Dall'esempio precedente abbiamo visto che

$$P(A|B) = \frac{P(A,B)}{P(B)} \quad \Rightarrow \quad P(A,B) = P(A|B)P(B)$$

Osservazioni

- La probabilità che accada sia A che B è la probabilità che si verifichi B per la probabilità che si verifichi A dato che B si è verificato. **Attenzione:** non c'è per forza una causalità temporale

- $P(A,B) = P(A)P(B)$ solo se $P(A|B) = P(A)$. Questo vuol dire che A e B sono eventi **indipendenti**, ovvero il verificarsi di B non modifica le probabilità di verificarsi di A

Teorema di Bayes

Esempio:

A: lancio un dado ed esce «4»

B: lancio una moneta ed esce «TESTA»

⇨ Anche se la moneta fosse uscita «CROCE», il dado ha la stessa probabilità di risultare in un «4»

Sappiamo che $P(A,B) = P(B,A)$. Inoltre $P(B,A) = P(B|A)P(A)$, e di conseguenza

$$P(A|B)P(B) = P(B|A)P(A)$$

⇩

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

TEOREMA DI BAYES

Teorema di Bayes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

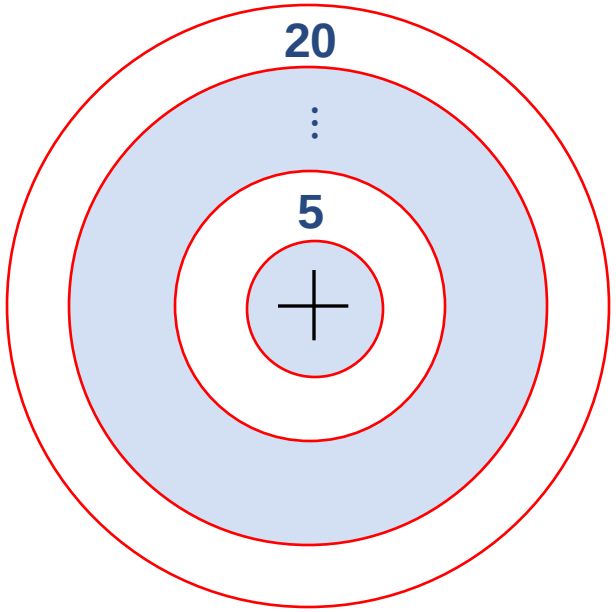
Osservazioni

- Il teorema di Bayes permette di **ridistribuire la probabilità**: prima conoscevamo $P(A)$, adesso conosco $P(A|B)$. La probabilità di A è cambiata in seguito all'informazione portata da B

- La distribuzione marginale $P(B) = \sum_A P(A,B) = \sum_A P(A|B)P(B)$ appare come un fattore di normalizzazione

Esempio: probabilità condizionata come ridistribuzione

Consideriamo un bersaglio da freccette con 20 cerchi. Supponiamo che un lanciatore abbia uguale probabilità di prendere ognuno dei 20 cerchi. **Qual è la probabilità che colpisca il cerchio #5?**



$$P(\#5) = \frac{1}{20}$$

Dopo un lancio, un amico gli dice che **non ha preso il cerchio #7**. Qual è ora la probabilità che abbia preso il cerchio #5?

Esempio: probabilità condizionata come ridistribuzione

Dato che sicuramente non ha preso il #7, la probabilità di aver preso il #5 è

$$P(\#5|\ \mathbf{NOT}\ \#7) = \frac{1}{19}$$

poiché, dopo, l'esclusione del cerchio #7, rimangono solo 19 cerchi «prendibili»

Il condizionamento a «**NOT** #7» significa che certi «stati» sono ora **inaccessibili**, e di conseguenza la probabilità si deve **ridistribuire** su quelli accessibili

$$P(\#5|\ \mathbf{NOT}\ \#7) = \frac{P(\#5, \mathbf{NOT}\ \#7)}{P(\mathbf{NOT}\ \#7)} = \frac{P(\#5) \cdot P(\mathbf{NOT}\ \#7 \mid \#5)}{P(\mathbf{NOT}\ \#7)} = \frac{\frac{1}{20} \cdot 1}{\frac{19}{20}} = \boxed{\frac{1}{19}}$$

Probabilità congiunte, marginali, condizionate

Riprendiamo l'esempio iniziale e proviamo a calcolare la distribuzione $P(a|b)$

$P(a|b)$

b \ a	0	1
0	0.2	0.8
1	0.4	0.6

$$P(a=1|b=0) = \frac{P(a=1,b=0)}{p(b=0)} = \frac{0.24}{0.3} = 0.8$$

$$P(a=0|b=0) = \frac{P(a=0,b=0)}{p(b=0)} = \frac{0.06}{0.3} = 0.2$$

$$P(a=1|b=1) = \frac{P(a=1,b=1)}{p(b=1)} = \frac{0.42}{0.7} = 0.6$$

$$P(a=0|b=1) = \frac{P(a=0,b=1)}{p(b=1)} = \frac{0.28}{0.7} = 0.4$$

Outline

1. Probabilità congiunte, condizionate, marginali

2. **Introduzione alla stima Bayesiana**

3. Stima ottima

4. Stima ottima lineare

Introduzione alla stima Bayesiana

Abbiamo finora considerato il vettore di parametri ignoto $\boldsymbol{\theta} \in \mathbb{R}^{d\times 1}$ come una **variabile deterministica**. Spesso però, ancora prima di collezionare i dati, abbiamo delle **informazioni** (o supposizioni) sui possibili valori che potrebbe assumere $\boldsymbol{\theta}$

Esempi:

1. Stima della concentrazione di una sostanza nell'aria: si ha un'idea dell'ordine di grandezza, per esempio in base a studi precedenti

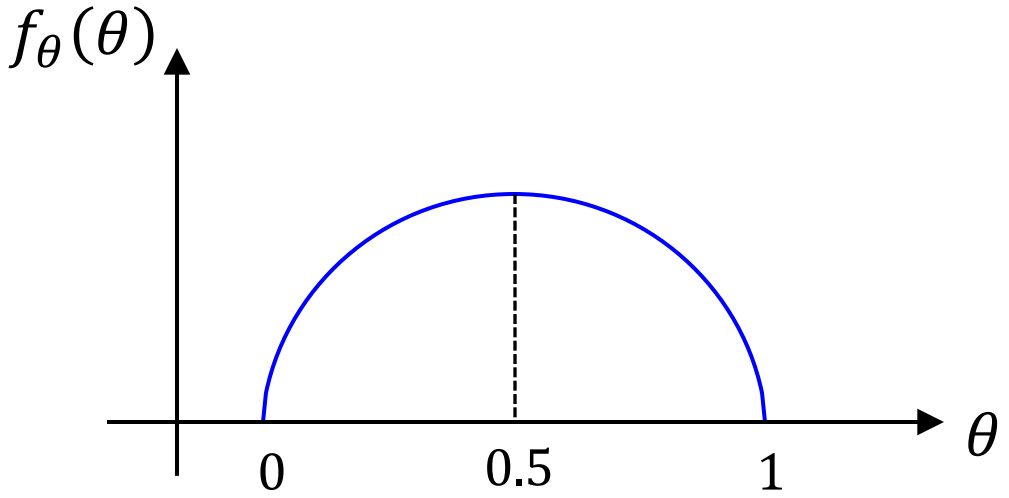
2. Stima della probabilità che una moneta risulti «TESTA» dopo un lancio: so già che il valore sarà intorno a 0.5, se suppongo non sia truccata

Introduzione alla stima Bayesiana

Ha quindi senso considerare $\boldsymbol{\theta}$ come una **variabile casuale:** in questo modo, posso specificare una distribuzione di probabilità per $\boldsymbol{\theta}$, per **descriverne i valori** (e la probabilità che $\boldsymbol{\theta}$ li assuma) che **io credo** che possa assumere

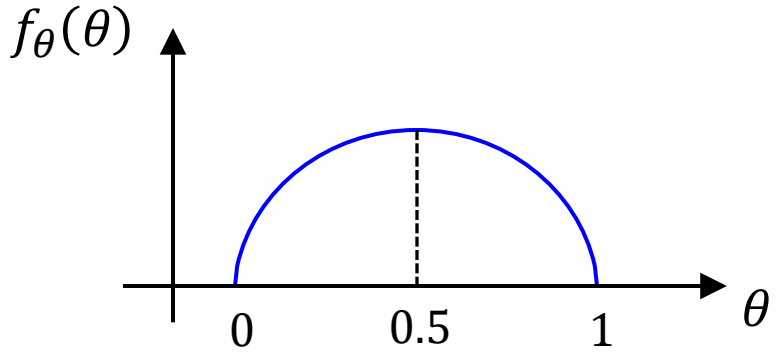
- assegno **maggior probabilità** ai valori che **io credo** siano più probabili che $\boldsymbol{\theta}$ **possa assumere**, e minor probabilità ai valori che **io credo** non possa assumere

Esempio: Sia θ la probabilità che una moneta risulta in «TESTA». Una possibile distribuzione (continua) $f_\theta(\theta)$ per θ, se suppongo che la moneta non sia truccata, è:



Introduzione alla stima Bayesiana

Osservazioni



- $f_\theta(\theta)$ ha dominio $[0,1]$ poiché θ, modellando una probabilità, deve stare tra 0 e 1

- Siccome suppongo che la moneta non è truccata, $\theta = 0.5$ sarà il valore che io suppongo sia più probabile, mentre $\theta \approx 0$ o $\theta \approx 1$ saranno poco probabili

- Data $f_\theta(\theta)$, abbiamo già una stima del valore di θ ancora prima di aver osservato i dati **(STIMA A-PRIORI)**. Ad esempio (ma non per forza) posso prendere come **valore puntuale** per la stima di θ il suo valore atteso. **L'incertezza sulla stima** sarà allora la varianza di θ **(INCERTEZZA A-PRIORI)**

Introduzione alla stima Bayesiana

Con l'osservazione dei dati, ci si aspetta che:

1. La stima puntuale di $\boldsymbol{\theta}$ **cambi**
2. L'incertezza sulla stima **decresca** (ho più informazioni!)

Abbiamo quindi due elementi che portano informazione:

1. La distribuzione a-priori $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ sui possibili valori di $\boldsymbol{\theta}$
2. L'informazione che portano i dati sui possibili valori di $\boldsymbol{\theta}$, ovvero la likelihood $f_{Y|\boldsymbol{\theta}}(Y|\boldsymbol{\theta})$

Quello che veramente ci interessa è sapere **quanto può valere $\boldsymbol{\theta}$ dato che ho osservato i dati,** ovvero la distribuzione $f_{\boldsymbol{\theta}|Y}(\boldsymbol{\theta}|Y)$

Distribuzione a-posteriori

Usando il teorema di Bayes possiamo unire i due elementi di informazione:

LIKELIHOOD PRIOR

$$f_{\boldsymbol{\theta}|Y}(\boldsymbol{\theta}|Y) = \frac{f_{Y|\boldsymbol{\theta}}(Y|\boldsymbol{\theta}) \cdot f_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{f_Y(Y)}$$

POSTERIOR MARGINAL LIKELIHOOD

Osservazioni

- $f_{\boldsymbol{\theta}|Y}(\boldsymbol{\theta}|Y)$ è una **distribuzione a-posteriori di possibili valori** di $\boldsymbol{\theta}$. Le probabilità di questi valori, rispetto a $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, sono state **riallocate** dall'aver osservato i dati Y
- Nel caso in cui $f_{Y|\boldsymbol{\theta}}(Y|\boldsymbol{\theta})$ e $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ sono pdf continue, allora $f_Y(Y) = \int_{-\infty}^{+\infty} f_{Y|\boldsymbol{\theta}}(Y|\boldsymbol{\theta})\, f_{\boldsymbol{\theta}}(\boldsymbol{\theta})\, d\boldsymbol{\theta}$

Distribuzione a-posteriori

Conosciamo la forma funzionale di $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ e $f_{Y|\boldsymbol{\theta}}(Y|\boldsymbol{\theta})$ poiché derivano dalle nostre assunzioni sui dati Y e sui parametri $\boldsymbol{\theta}$. Posso dire qualcosa su $f_{\boldsymbol{\theta}|Y}(\boldsymbol{\theta}|Y)$?

- In generale, **non posso dire nulla**. Solo in alcuni casi fortunati ho che $f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|Y)$ ha un'espressione analitica nota
- Un altro problema è che $f_Y(Y)$, nel caso di dati intesi come v.c. continue, è un integrale che potremmo **non sapere come risolvere**. In questo caso si usano tecniche numeriche note come **Markov Chain Monte Carlo (MCMC)**
- Un caso fortunato avviene, per esempio ma non solo, se $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ è **Gaussiana** e anche $f_{Y|\boldsymbol{\theta}}(Y|\boldsymbol{\theta})$ è **Gaussiana**. Allora, anche $f_{\boldsymbol{\theta}|Y}(\boldsymbol{\theta}|Y)$ è **Gaussiana**

Distribuzione a-posteriori

Quando la **posterior** $f_{\boldsymbol{\theta}|Y}(\boldsymbol{\theta}|Y)$ ha la stessa forma della **prior** $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ (e.g. sono entrambe delle Gaussiane) allora la **likelihood** e la **prior** si dicono **coniugate**

Un modo (computazionalmente oneroso ma semplice) per calcolare la posterior $f_{\boldsymbol{\theta}|Y}(\boldsymbol{\theta}|Y)$ è quello di **discretizzare** il range di valori del parametro $\boldsymbol{\theta}$ tramite una griglia di valori

- In questo modo valuto $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ e $f_{Y|\boldsymbol{\theta}}(Y|\boldsymbol{\theta})$ solo in quei valori di $\boldsymbol{\theta}$ all'interno della griglia
- Questo metodo va bene se $\boldsymbol{\theta}$ consiste di un paio di parametri. Altrimenti, diventa troppo oneroso ed è meglio ricorrere ad MCMC (a meno che non esista un'espressione analitica nota per la posterior)

Esempio: stima probabilità che la moneta esca testa

Vogliamo stimare la probabilità $\theta \equiv \pi$ che la moneta risulti in «TESTA». Supponiamo di lanciare una moneta $N = 10$ volte, e di osservare $N_s = 7$ «TESTA» $(y = 1)$ e $N - N_s = 3$ «CROCE» $(y = 0)$. I dati $\mathcal{D}$ sono (l'ordine non importa essendo i dati i.i.d. per ipotesi):

$$\underset{10 \times 1}{Y} = [1\ 1\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 0]^{\mathsf{T}}$$

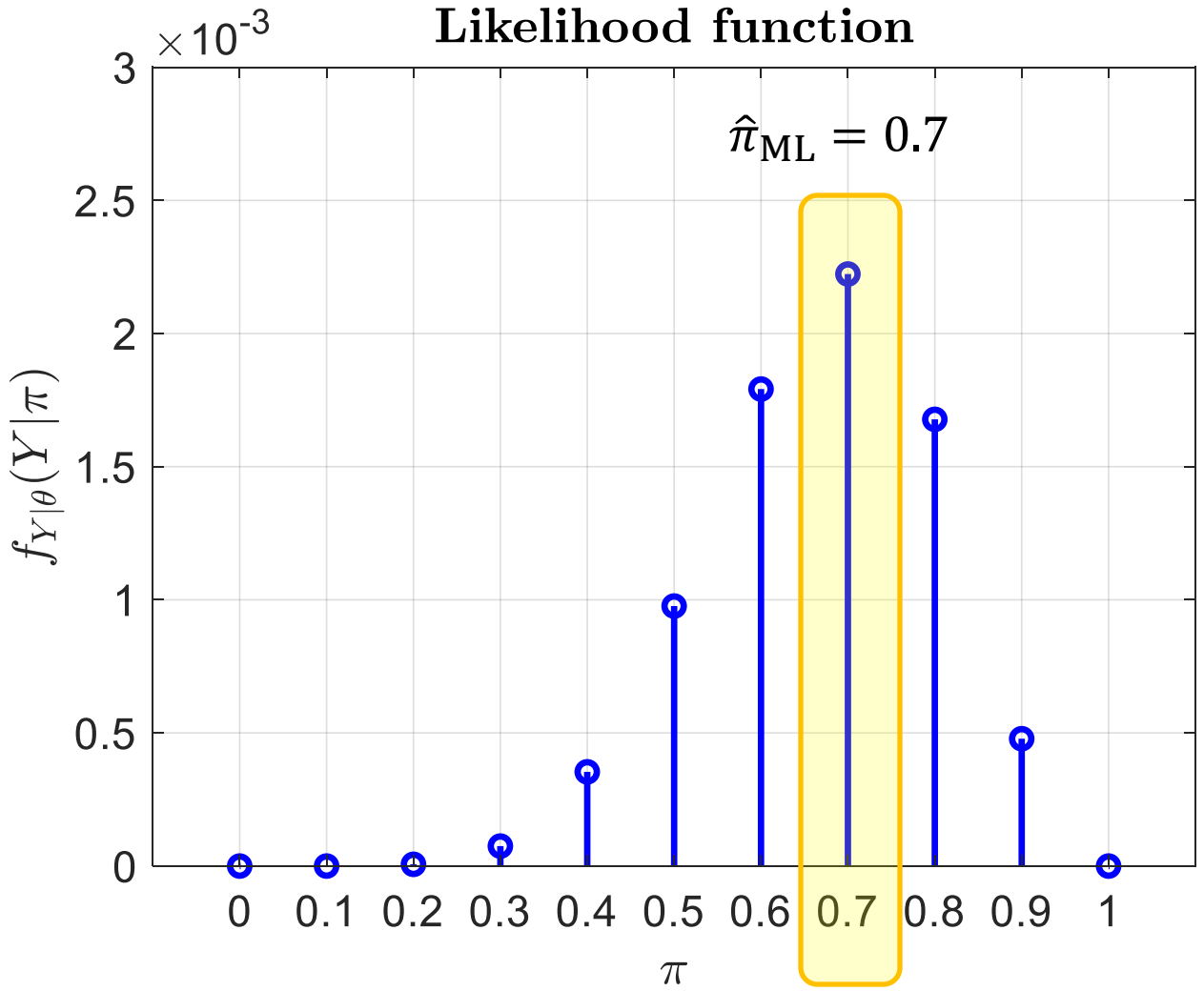
Modelliamo i dati come realizzazioni i.i.d. di una v.c. avente distribuzione di Bernoulli:

$$y(i) \sim \text{Bernoulli}(\pi), \quad \text{i.i.d.} \quad \Longrightarrow \quad f_y(y(i)|\pi) = \pi^{y(i)} \cdot (1-\pi)^{(1-y(i))}$$

Likelihood: $$f_{Y|\theta}(Y|\pi) = \prod_{i=1}^{N} \pi^{y(i)} \cdot (1-\pi)^{(1-y(i))} = \pi^{\overbrace{\Sigma_{i=1}^{N} y(i)}^{\#\text{ successi}}} \cdot (1-\pi)^{\overbrace{\Sigma_{i=1}^{N} 1-y(i)}^{\#\text{ insuccessi}}}$$

Esempio: stima probabilità che la moneta esca testa

Se facessimo una **stima a massima verosimiglianza,** prenderemmo come stima il valore $\hat{\pi}_{\mathrm{ML}}$ che **massimizza la verosimiglianza,** ovvero $\hat{\pi}_{\mathrm{ML}} = N_s/N = 0.7$

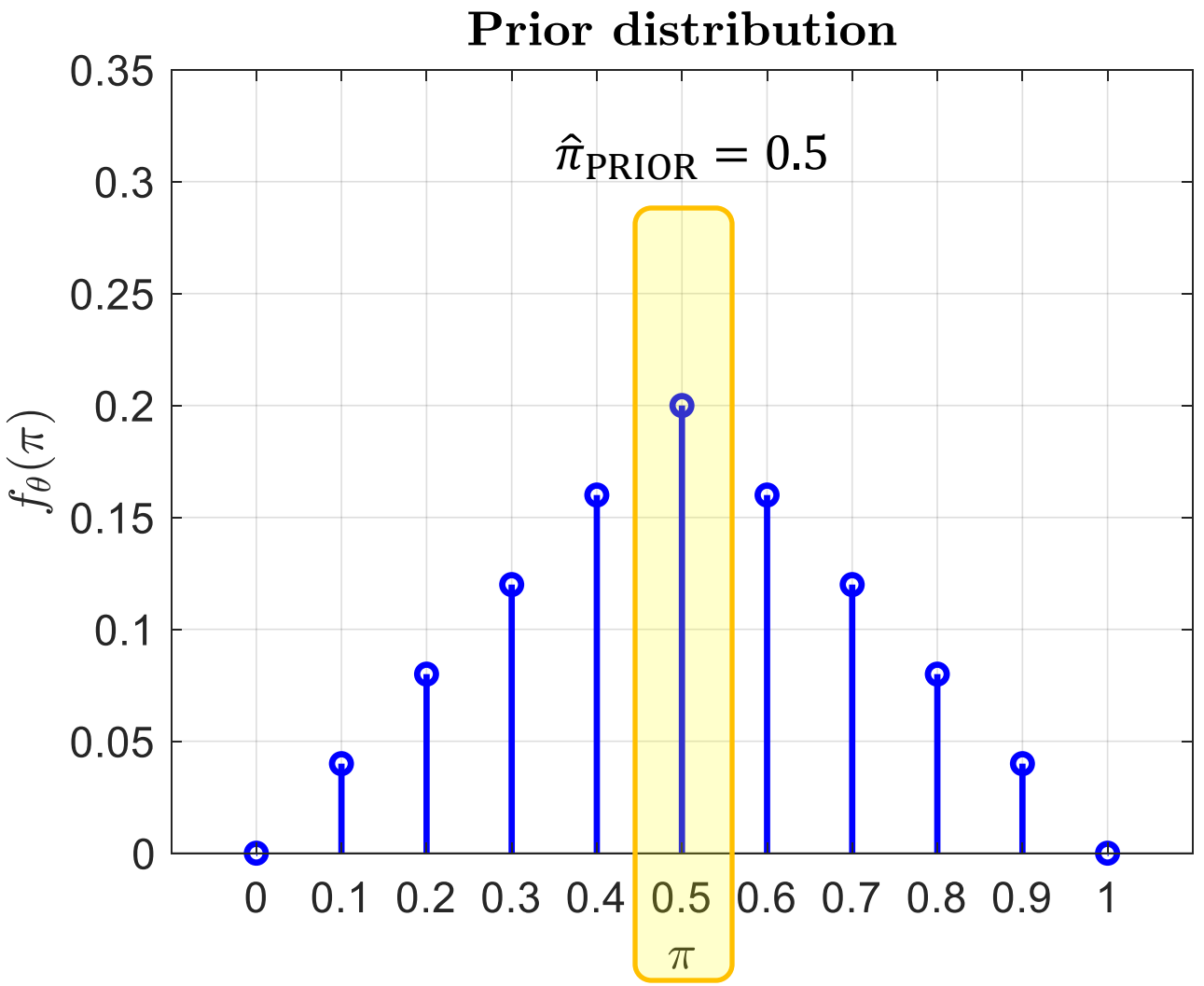


Esempio: stima probabilità che la moneta esca testa

Supponiamo però di avere una **buona confidenza** che la **moneta non sia truccata.** Potremmo esprimere questa nostra informazione a-priori tramite una distribuzione $f_\theta(\pi)$

In questa «rappresentazione della nostra credenza», diamo più probabilità al fatto che $\pi = 0.5$.

Possiamo prendere come stima di π il valore $\hat{\pi}_{\text{PRIOR}} = 0.5$

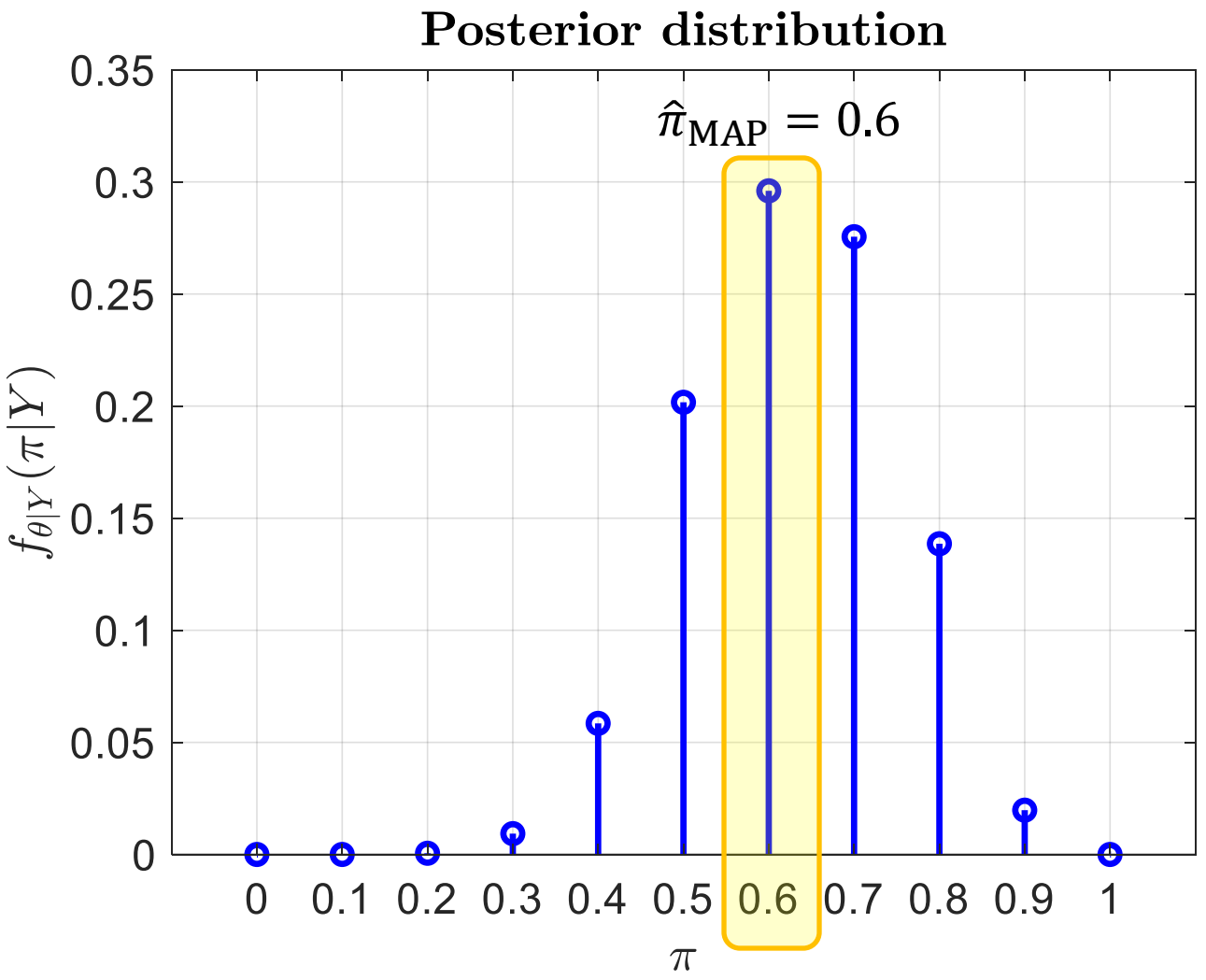


Esempio: stima probabilità che la moneta esca testa

Unendo le informazioni di prior e di likelihood ottengo una distribuzione di valori di π che è un **compromesso** tra la prior e la likelihood

In questo senso, la procedura di stima Bayesiana **«regolarizza»** la stima di π

Il valore di $\hat{\pi}_{\mathrm{MAP}}$ che massimizza la posterior è chiamato **stima MAP** (Maximum A Posteriori)



Esempio: stima probabilità che la moneta esca testa

LIKELIHOOD PRIOR

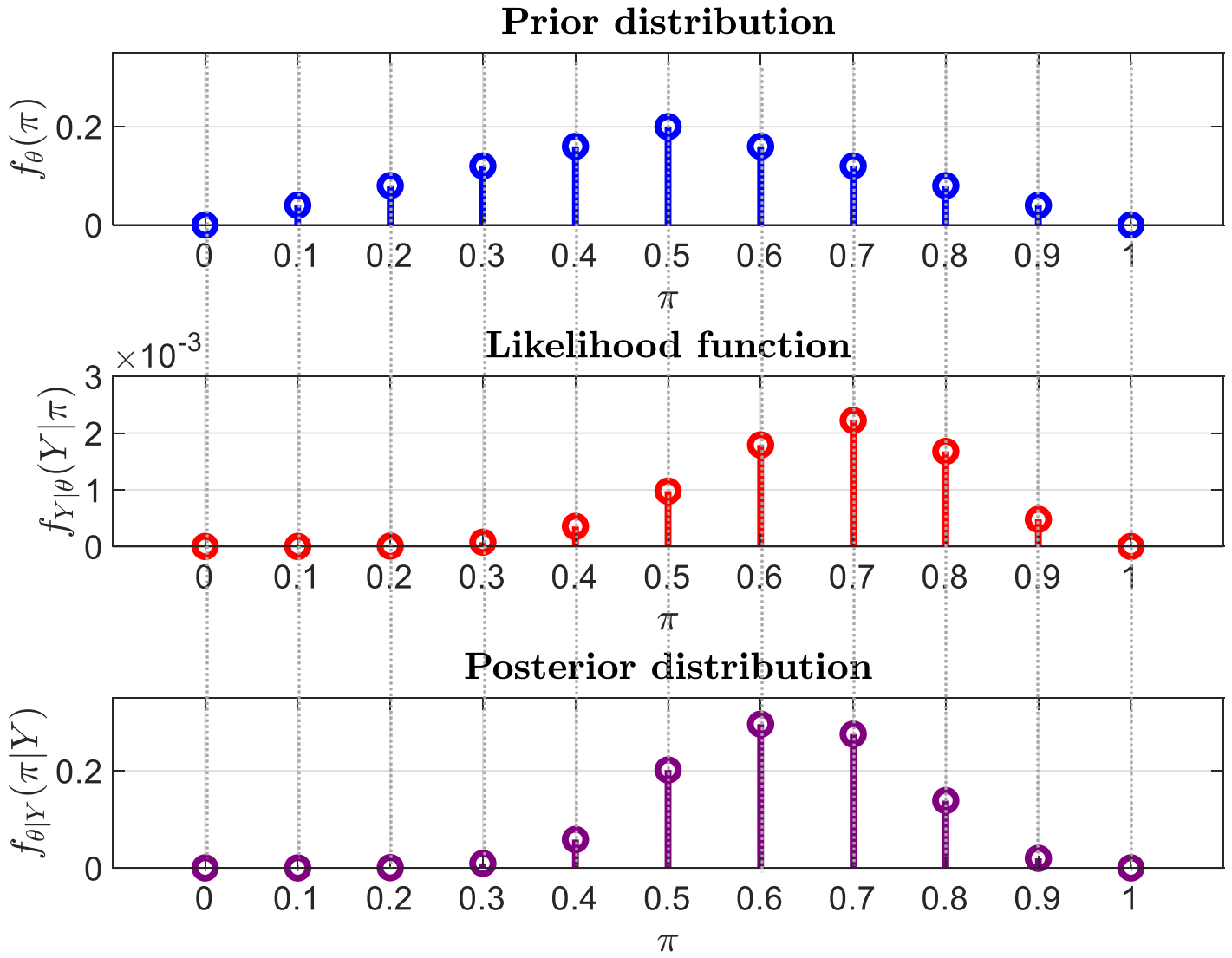
$$f_{\theta|Y}(\pi|Y) = \frac{f_{Y|\theta}(Y|\pi) \cdot f_{\theta}(\pi)}{f_Y(Y)}$$

POSTERIOR

MARGINAL LIKELIHOOD

$$f_Y(Y) = \sum_{\pi} f_{Y|\theta}(Y|\pi) \cdot f_{\theta}(\pi)$$

$$= 9.683 \cdot 10^{-4}$$



Outline

1. Probabilità congiunte, condizionate, marginali
2. Introduzione alla stima Bayesiana
3. **Stima ottima**
4. Stima ottima lineare

Stima ottima

Supponiamo di avere la **posterior** $f_{\boldsymbol{\theta}|Y}(\boldsymbol{\theta}|Y)$. Abbiamo quindi una distribuzione di valori dei parametri ignoti $\boldsymbol{\theta}$. Spesso però ci serve un **valore solo, puntuale**. Abbiamo varie scelte:

- **Stima MAP:** $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}|Y}(\boldsymbol{\theta}|Y)$
- **Valore atteso a posteriori:** $\widehat{\boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{\theta}}\left[f_{\boldsymbol{\theta}|Y}(\boldsymbol{\theta}|Y)\right] \equiv \mathbb{E}[\boldsymbol{\theta}|Y]$, ovvero il valore atteso della posterior
- **Altre quantità**, come la mediana, ecc…

Ricordiamo che in generale indichiamo uno **stimatore** come una funzione $T(\quad)$ dei dati $\mathcal{D}$:

$$\widehat{\boldsymbol{\theta}} = T(\mathcal{D})$$

Stima ottima

Consideriamo il caso θ **scalare** per semplicità. Vorremmo che la variabile casuale $\hat{\theta}$ fosse «vicina» alla variabile casuale θ. Per quantificare questa «distanza», usiamo il concetto di **Mean Squared Error (MSE)** già visto in precedenza (si veda Lezione 02)

$$\mathrm{MSE} \equiv \mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right] = \mathbb{E}[(T(\mathcal{D}) - \theta)^2]$$

Lo **stimatore ottimo di Bayes** è quella funzione $T^{\mathrm{opt}}(\quad)$ tale che:

$$\mathbb{E}[(T^{\mathrm{opt}}(\mathcal{D}) - \theta)^2] < \mathbb{E}[(T(\mathcal{D}) - \theta)^2], \quad \forall T(\mathcal{D})$$

cioè che **minimizza il MSE**

Stima ottima

Si dimostra che

$$T^{\text{opt}}(Y) = \mathbb{E}[\boldsymbol{\theta}|\mathcal{D} = Y]$$

Ovvero, lo **stimatore che minimizza il MSE è il valore atteso condizionato** (al fatto che i dati $\mathcal{D}$ abbiano assunto i valori in Y)

Nota

Nel caso in cui $\boldsymbol{\theta}$ sia un **vettore di parametri**, il calcolo del MSE si modifica come segue

$$\text{MSE} \equiv \text{tr}\left\{\mathbb{E}\left[\underset{d\times 1}{(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})}\underset{1\times d}{(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^{\mathsf{T}}}\right]\right\} = \mathbb{E}\left[\underset{1\times d}{(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^{\mathsf{T}}}\underset{d\times 1}{(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})}\right] = \mathbb{E}\left[\underset{1\times 1}{\|(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\|_2^2}\right]$$

Stima ottima: il caso Gaussiano

Supponiamo ora di avere un dato interpretato come realizzazione di una variabile casuale Gaussiana $y \sim \mathcal{N}(0, \lambda_{yy}^2)$, e che anche il parametro ignoto (scalare per comodità) sia Gaussiano $\theta \sim \mathcal{N}(0, \lambda_{\theta\theta}^2)$.

$$\underbrace{\begin{bmatrix} y \\ \theta \end{bmatrix}}_{\boldsymbol{z}} \sim \mathcal{N}\left(\underbrace{\begin{bmatrix} 0 \\ 0 \end{bmatrix}}_{\boldsymbol{\mu}}, \underbrace{\begin{bmatrix} \lambda_{yy}^2 & \lambda_{y\theta} \\ \lambda_{\theta y} & \lambda_{\theta\theta}^2 \end{bmatrix}}_{\Sigma} \right)$$

2×1 2×1 2×2

La loro pdf **congiunta** $f_{y\theta}(y, \theta)$ è ancora **Gaussiana**

$$f_{y\theta}(y, \theta) = \frac{1}{\sqrt{(2\pi)^2 \det \Sigma}} \exp\left(-\frac{1}{2} (\boldsymbol{z} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{z} - \boldsymbol{\mu}) \right)$$

Al quadrato perché ho 2 variabili

Stima ottima: il caso Gaussiano

La pdf dei dati $f_y(y)$ è: $$f_y(y) = \frac{1}{\sqrt{2\pi\,\lambda_{yy}^2}} \exp\left(-\frac{1}{2\lambda_{yy}^2}(y-0)^2\right)$$

Si dimostra che la **posterior** $f_{\theta|y}(\theta|y) = f_{y\theta}(y,\theta)/\, f_y(y)$ è ancora **Gaussiana** con:

- **Valore atteso:** $$\mu_{\theta|y} = \frac{\lambda_{\theta y}}{\lambda_{yy}^2} \cdot y$$

- **Varianza:** $$\lambda_{\theta|y}^2 = \lambda_{\theta\theta}^2 - \frac{\lambda_{\theta y}^2}{\lambda_{yy}^2}$$

- Se $\lambda_{\theta y} = 0$, ovvero se y **non porta informazioni** su θ, la stima di θ rimane quella a priori
- Notiamo che $\frac{\lambda_{\theta y}^2}{\lambda_{yy}^2} > 0$. Quindi, **l'incertezza a posteriori è minore** di quella a priori
- Se λ_{yy}^2 è **grande**, la varianza **diminuisce di poco**, perché i dati sono molto incerti

Stima ottima: il caso Gaussiano

Avendo osservato il valore $y(1)$ di y, lo stima ottenuta dallo **stimatore ottimo Bayesiano nel caso Gaussiano** sarà:

$$\hat{\theta}_{\text{opt}} = \mathbb{E}[\theta | y = y(1)] = \frac{\lambda_{\theta y}}{\lambda_{yy}^2} \cdot y(1)$$

Outline

1. Probabilità congiunte, condizionate, marginali
2. Introduzione alla stima Bayesiana
3. Stima ottima
4. **Stima ottima lineare**

Stima ottima lineare

Non è sempre detto che y e θ siano congiuntamente Gaussiane. Vogliamo quindi trovare uno stimatore che **non faccia ipotesi sulla ddp congiunta** di y e θ

Supponiamo y e θ due *variabili casuali scalari* con valore atteso nullo e varianza λ_{yy}^2 e $\lambda_{\theta\theta}^2$ rispettivamente

- $\mathbb{E}[y] = 0$
- $\mathbb{E}[\theta] = 0$
- $\mathbb{E}[y^2] = \lambda_{yy}^2$
- $\mathbb{E}[\theta^2] = \lambda_{\theta\theta}$
- $\mathbb{E}[\theta y] = \lambda_{\theta y}$

Vogliamo stimare θ tramite uno **stimatore lineare**, del tipo:

$$\hat{\theta}^{\text{lin}} = \alpha \cdot y + \beta, \qquad \alpha, \beta \in \mathbb{R}$$

Stima ottima lineare

Per trovare α e β, **minimizziamo la funzione di costo** data dal Mean Square Error

$$\mathrm{MSE} \equiv J(\alpha,\beta) = \mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right] = \mathbb{E}[(\alpha \cdot y + \beta - \theta)^2]$$

Calcoliamo il gradiente e poniamolo uguale a zero (non verifichiamo sia un minimo):

$$\frac{\partial J(\alpha,\beta)}{\partial \alpha} = 0 \quad \Rightarrow \quad 2 \cdot \mathbb{E}[(\alpha \cdot y + \beta - \theta) \cdot y] = 0 \quad \Rightarrow \quad \mathbb{E}[\alpha y^2] + \mathbb{E}[\beta y] - \mathbb{E}[\theta y] = 0$$

$$\Rightarrow \quad \alpha \cdot \lambda_{yy}^2 + \beta \cdot 0 - \lambda_{\theta y} = 0 \quad \Rightarrow \quad \alpha \cdot \lambda_{yy}^2 = \lambda_{\theta y}$$

$$\Rightarrow \quad \boxed{\alpha = \lambda_{\theta y} / \lambda_{yy}^2}$$

Stima ottima lineare

$$\frac{\partial J(\alpha,\beta)}{\partial \beta} = 0 \quad \Rightarrow \quad 2 \cdot \mathbb{E}[(\alpha \cdot y + \beta - \theta) \cdot 1] = 0 \quad \Rightarrow \quad \mathbb{E}[\alpha y] + \mathbb{E}[\beta] - \mathbb{E}[\theta] = 0$$

$$\Rightarrow \quad \alpha \cdot 0 + \beta - 0 = 0 \quad \Rightarrow \quad \boxed{\beta = 0}$$

$$\begin{cases} \dfrac{\partial J(\alpha,\beta)}{\partial \alpha} = 0 \\ \dfrac{\partial J(\alpha,\beta)}{\partial \beta} = 0 \end{cases} \Rightarrow \boxed{\begin{cases} \alpha = \lambda_{\theta y}/\lambda_{yy}^2 \\ \beta = 0 \end{cases}}$$

Stima ottima lineare

Lo **stimatore lineare ottimo** è quindi dato da

$$\hat{\theta}_{\text{opt}}^{\text{lin}} = \hat{\alpha} \cdot y + \hat{\beta} \quad = \frac{\lambda_{\theta y}}{\lambda_{yy}^2} \cdot y$$

Coincide con lo stimatore ottimo di Bayes per il caso Gaussiano!

La varianza della stima si ricava **essere uguale al caso Gaussiano:**

$$\text{Var}\left[\hat{\theta}_{\text{opt}}^{\text{lin}} - \theta\right] = \lambda_{\theta\theta}^2 - \frac{\lambda_{\theta y}^2}{\lambda_{yy}^2}$$

Stima ottima lineare

Osservazioni

- Lo stimatore ottimo lineare non **fa nessuna ipotesi su che tipo di distribuzione** hanno y e θ. Assume solo che siano v.c. con una certa media e una certa varianza

- Potrebbe dunque esserci uno **stimatore migliore** (nel senso che ha MSE minore) **rispetto a quello lineare ottimo**

- Se però y e θ sono **congiuntamente Gaussiani**, allora **non esiste nessuno stimatore migliore** di quello lineare ottimo

Stima ottima lineare

Generalizzazione 1: valore atteso non nullo, y e θ scalari

Se:
- $\mathbb{E}[y] = \mu_y \neq 0$
- $\mathbb{E}[\theta] = \mu_\theta \neq 0$

$$\Rightarrow \quad \hat{\theta}^{\text{lin}}_{\text{opt}} = \mu_\theta + \frac{\lambda_{\theta y}}{\lambda^2_{yy}} \cdot (y - \mu_y) \qquad \text{Var}[\hat{\theta}^{\text{lin}}_{\text{opt}} - \theta] = \lambda^2_{\theta\theta} - \frac{\lambda^2_{\theta y}}{\lambda^2_{yy}}$$

Generalizzazione 2: $Y \in \mathbb{R}^{N\times 1}$ e $\boldsymbol{\theta} \in \mathbb{R}^{d\times 1}$ vettoriali

Se:
- $\underset{N\times 1}{\mathbb{E}[Y]} = \underset{N\times 1}{\boldsymbol{\mu}_Y} \neq \mathbf{0}$
- $\underset{d\times 1}{\mathbb{E}[\boldsymbol{\theta}]} = \underset{d\times 1}{\boldsymbol{\mu}_{\boldsymbol{\theta}}} \neq \mathbf{0}$

$$\underset{d\times 1}{\text{Var}\begin{bmatrix} Y \\ \boldsymbol{\theta} \end{bmatrix}} = \begin{bmatrix} \underset{N\times N}{\Lambda_{YY}} & \underset{N\times d}{\lambda_{Y\boldsymbol{\theta}}} \\ \underset{d\times N}{\Lambda_{\boldsymbol{\theta}Y}} & \underset{d\times d}{\Lambda_{\boldsymbol{\theta}\boldsymbol{\theta}}} \end{bmatrix}$$

$$\underset{d\times 1}{\hat{\boldsymbol{\theta}}^{\text{lin}}_{\text{opt}}} = \underset{d\times 1}{\mu_{\boldsymbol{\theta}}} + \underset{d\times N}{\Lambda_{\boldsymbol{\theta}Y}} \cdot \underset{N\times N}{\Lambda^{-1}_{YY}} \cdot \underset{N\times 1}{(Y - \boldsymbol{\mu}_Y)} \qquad \underset{d\times d}{\text{Var}[\hat{\boldsymbol{\theta}}^{\text{lin}}_{\text{opt}} - \boldsymbol{\theta}]} = \underset{d\times d}{\Lambda_{\boldsymbol{\theta}\boldsymbol{\theta}}} - \underset{d\times N}{\Lambda_{\boldsymbol{\theta}Y}} \cdot \underset{N\times N}{\Lambda^{-1}_{YY}} \cdot \underset{N\times d}{\Lambda_{Y\boldsymbol{\theta}}}$$

Connessione con il Filtro di Kalman

Le formule appena viste ammettono una **forma ricorsiva:** appena arriva un dato osservato nuovo, si aggiorna la stima corrente senza considerare nuovamente tutti i dati

Queste espressioni ricorsive dello stimatore lineare ottimo sono alla base del **Filtro di Kalman**, un algoritmo che ha l'obiettivo di **stimare lo stato $\boldsymbol{x}(t)$ di un sistema dinamico**

- lo stato $\boldsymbol{x}(t)$ e l'uscita $y(t)$ del sistema dinamico lineare sono visti come variabili casuali
- si vuole **stimare lo stato $\boldsymbol{x}(t)$**, visto come l'incognita $\boldsymbol{\theta}$, sulla base dello **stato stimato al tempo precedente (stima a priori)** e sui dati che man mano arrivano dalle **misure dei sensori $y(t)$ (dati osservati)**

UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione