



UNIVERSITÀ  
DEGLI STUDI  
DI BERGAMO

Dipartimento  
di Ingegneria Gestionale,  
dell'Informazione e della Produzione



# IDENTIFICAZIONE DEI MODELLI E ANALISI DEI DATI (IMAD)

**Lezione 4: Stima a massima  
verosimiglianza** (maximum likelihood  
estimation)

**Corso di Laurea Magistrale in  
INGEGNERIA INFORMATICA**

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 Regularizzazione

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

## Parte II: sistemi dinamici

### Stima parametrica $\hat{\theta}$

- $\theta$  deterministico

- ***NO assunzioni su pdf dei dati***

- ✓ Stima parametri popolazione
- ✓ Stima modello lineare: minimi quadrati

- ***SI assunzioni su pdf dei dati***

- ✓ Stima massiva verosimiglianza parametri popolazione
- ✓ Stima modello lineare: massiva verosimiglianza
- ✓ Regressione logistica

- $\theta$  variabile casuale

- ***SI assunzioni su pdf dei dati***

- ✓ Stima bayesiana

### Machine learning

### Stima parametrica $\hat{\theta}$

- $\theta$  deterministico

- ***NO assunzioni su pdf dei dati***

- ✓ Modelli lineari di pss
- ✓ Predizione
- ✓ Identificazione
- ✓ Persistente eccitazione
- ✓ Analisi asintotica metodi PEM
- ✓ Analisi incertezza stima (numero dati finito)
- ✓ Scelta complessità del modello



# Outline

1. Stima a massima verosimiglianza
2. Stima a massima verosimiglianza di parametri della popolazione
3. Stima a massima verosimiglianza di modelli lineari



# Outline

## 1. Stima a massima verosimiglianza

2. Stima a massima verosimiglianza di parametri della popolazione

3. Stima a massima verosimiglianza di modelli lineari



# Stima a massima verosimiglianza

Abbiamo presentato finora diversi tipi di stimatori:

- **Media campionaria:**  $\hat{\mu} = \frac{1}{N} \sum_{i=1}^N y(i) \Rightarrow \hat{\theta} = \mu \in \mathbb{R}$
- **Varianza campionaria:**  $S_{N-1}^2 = \frac{1}{N-1} \cdot \sum_{i=1}^N (y(i) - \hat{\mu})^2 \Rightarrow \hat{\theta} = \sigma^2 \in \mathbb{R}$
- **Stimatore a minimi quadrati**  $y(i) = \theta_0 + \theta_1 \varphi_1(i) + \dots + \theta_{d-1} \varphi_{d-1}(i) + \epsilon(i)$   
di un modello lineare:  
 $\epsilon(i)$  indipendenti media nulla e varianza  $\lambda^2$   
 $\Rightarrow \hat{\theta} = [\hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_{d-1}] \in \mathbb{R}^{d \times 1}$

# Stima a massima verosimiglianza

Gli stimatori presentati sono **parametrici**, nel senso che stimano dei parametri del sistema che ha generato i dati

- Nel fare ciò, non abbiamo **mai fatto assunzioni** sulla **distribuzione di probabilità** dei dati  $\mathcal{D} = \{y(1), y(2), \dots, y(N)\}$

Il metodo della **massima verosimiglianza** (MLE – Maximum Likelihood Estimation) è una procedura di stima che, **dato un modello probabilistico**, stima i suoi **parametri** in modo tale che siano **più coerenti con i dati** osservati

# Stima a massima verosimiglianza

Supponiamo di avere a disposizione  $N$  osservazioni  $Y = [y(1), y(2), \dots, y(N)]^T$ , dove

$$y(i) \sim \mathcal{N}(\mu, \sigma^2) \text{ i.i.d.} \quad \Rightarrow \quad f_y(y(i)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{1}{2} \left( \frac{y(i) - \mu}{\sigma} \right)^2 \right]$$

*Probability density function*

La **pdf congiunta** dei dati è  $f_Y(y(1), y(2), \dots, y(N) | \mu, \sigma^2) = f_Y(Y | \mu, \sigma^2) = \prod_{i=1}^N \mathcal{N}(y(i) | \mu, \sigma^2)$

La pdf congiunta  $f_Y(Y | \mu, \sigma^2)$  indica la **probabilità che si realizzi il vettore di dati osservato**

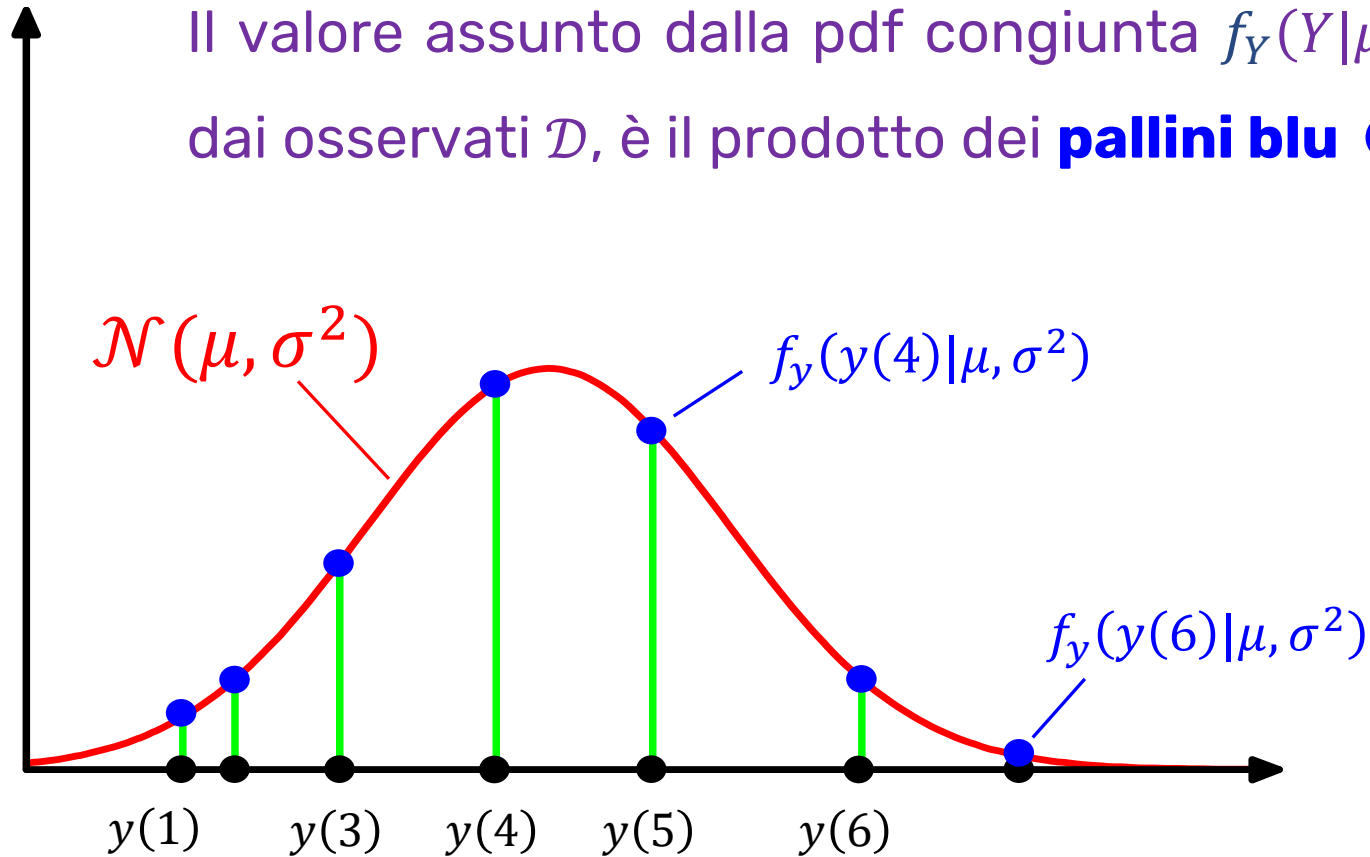
- Siccome le  $y(i)$  sono i.i.d., la probabilità di osservare  $y(1)$  AND  $y(2)$  AND ... AND  $y(N)$  è il **prodotto delle pdf** delle singole variabili



# Esempio: calcolo pdf congiunta, parametri noti

Supponiamo di avere  $N = 6$  dati  $\mathcal{D} = \{y(1), y(2), \dots, y(6)\}$ ,  $y(i) \sim \mathcal{N}(\mu, \sigma^2)$  i.i.d.

Il valore assunto dalla pdf congiunta  $f_Y(Y|\mu, \sigma^2)$ , con  $\mu$  e  $\sigma^2$  **noti**, valutata nei dai osservati  $\mathcal{D}$ , è il prodotto dei **pallini blu** ●



$$\begin{aligned} f_Y(Y|\mu, \sigma^2) &= f_y(y(1)|\mu, \sigma^2) \cdot \\ &\quad \cdot f_y(y(2)|\mu, \sigma^2) \cdot \\ &\quad \vdots \\ &\quad \cdot f_y(y(6)|\mu, \sigma^2) \end{aligned}$$

# Stima a massima verosimiglianza

Se funzione dei dati  $Y$ , la pdf congiunta è una **distribuzione multivariabile**. Io però **conosco il valore di**  $Y$ , dato che ho osservato i dati

Se conoscessi anche  $\mu$  e  $\sigma$ , potrei calcolare la probabilità di avere osservato  $Y$ . Però **non conosco**  $\mu$  e  $\sigma$ ! E' proprio quello che voglio stimare!

Quando  $f_Y(Y|\mu, \sigma^2)$  (la **pdf congiunta**) è vista in funzione dei parametri  $\mu$  and  $\sigma$ , è chiamata funzione di **likelihood**  $\mathcal{L}(\mu, \sigma^2|Y)$

Cambia solo **l'interpretazione**, ma  $f_Y(Y|\mu, \sigma^2)$  e  $\mathcal{L}(\mu, \sigma^2|Y)$  sono lo **stesso oggetto matematico!**

# Stima a massima verosimiglianza

## Riassunto:

**Variabili non note**

**Parametri NOTI**

- Se  $f_Y(\textcolor{red}{Y} | \mu, \sigma^2)$  è funzione dei dati  $Y$ : **pdf multivariabile**

**Dati NOTI**

**Variabili non note**

- Se  $f_Y(\textcolor{blue}{Y} | \mu, \textcolor{red}{\sigma}^2)$  è funzione dei parametri  $\mu$  e  $\sigma^2$ : **likelihood**  $\mathcal{L}(\mu, \textcolor{red}{\sigma}^2 | \textcolor{blue}{Y})$

Di solito si cambia la notazione di  $f_Y(Y|\mu, \sigma^2)$  in  $\mathcal{L}(\mu, \sigma^2|Y)$  per rendere più chiaro chi è supposto noto («a destra della barra |») e chi non è noto («a sinistra della barra |» )

# Stima a massima verosimiglianza

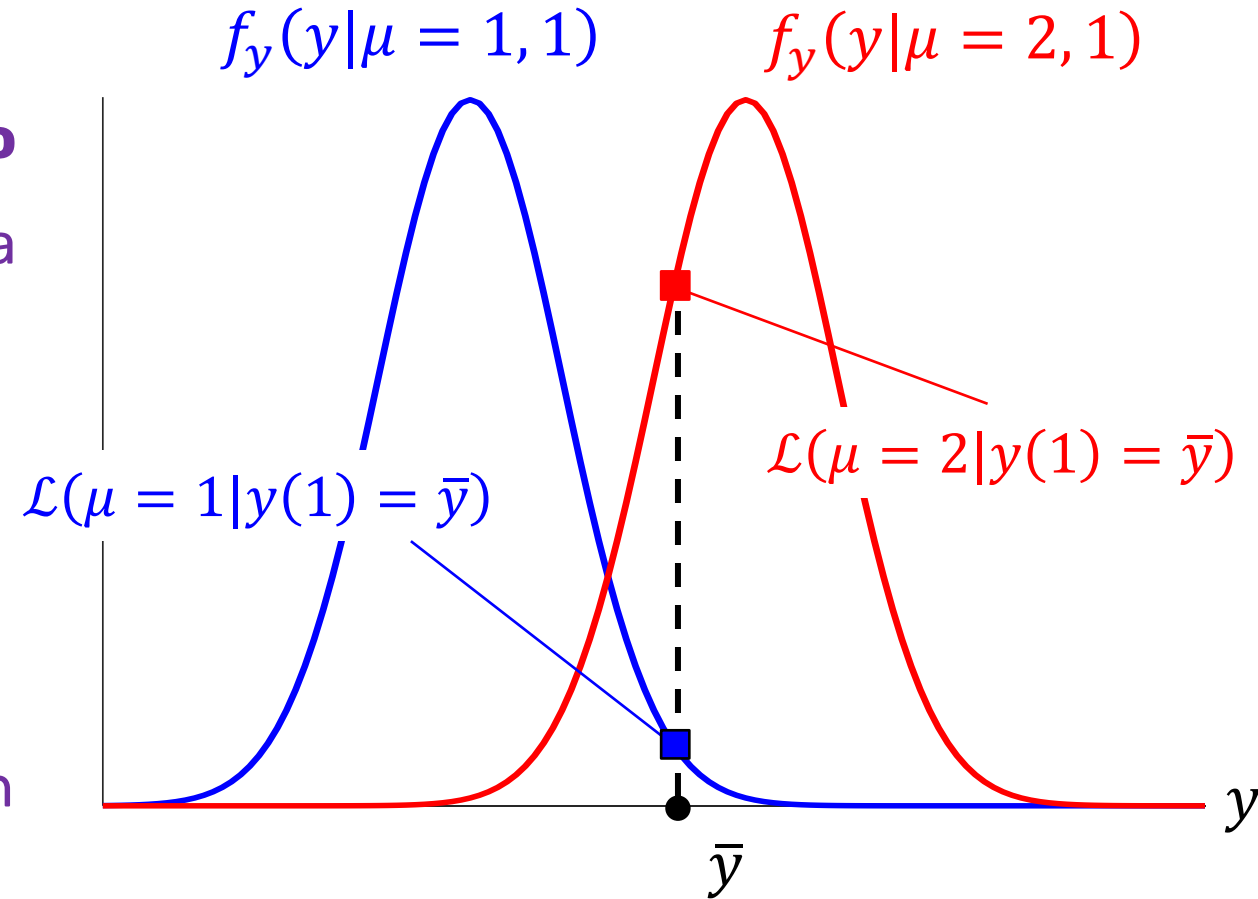
La stima a massima verosimiglianza è quel valore del parametro  $\theta$  che **massimizza la verosimiglianza**  $\mathcal{L}(\theta|Y)$

**Esempio:** supponiamo di avere **un solo dato**  $y(1) \sim \mathcal{N}(\mu, \sigma^2 = 1)$ , e che il suo valore sia  $y(1) = \bar{y}$ . Il **parametro da stimare** è  $\theta = \mu$

Notiamo che:

$$\mathcal{L}(\mu = 2|y(1) = \bar{y}) > \mathcal{L}(\mu = 1|y(1) = \bar{y})$$

Per cui  $\mu = 2$  è **più verosimile** di  $\mu = 1$ , in base a questo modello e questi dati



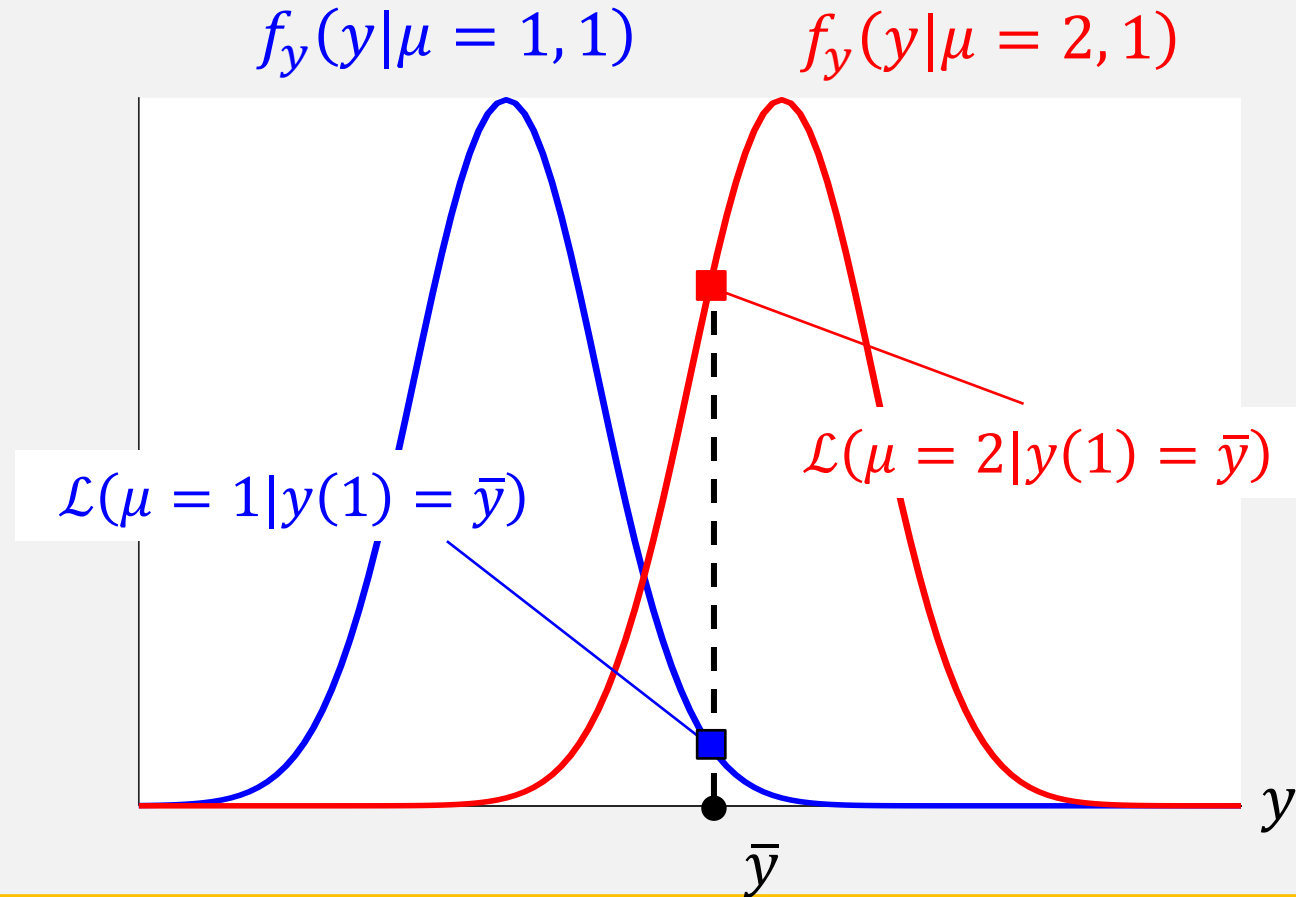
# QUIZ!

In questo esempio, la **stima a massima verosimiglianza** è:

☐  $\hat{\mu} = 2\bar{y}$

☐  $\hat{\mu} = \bar{y}$

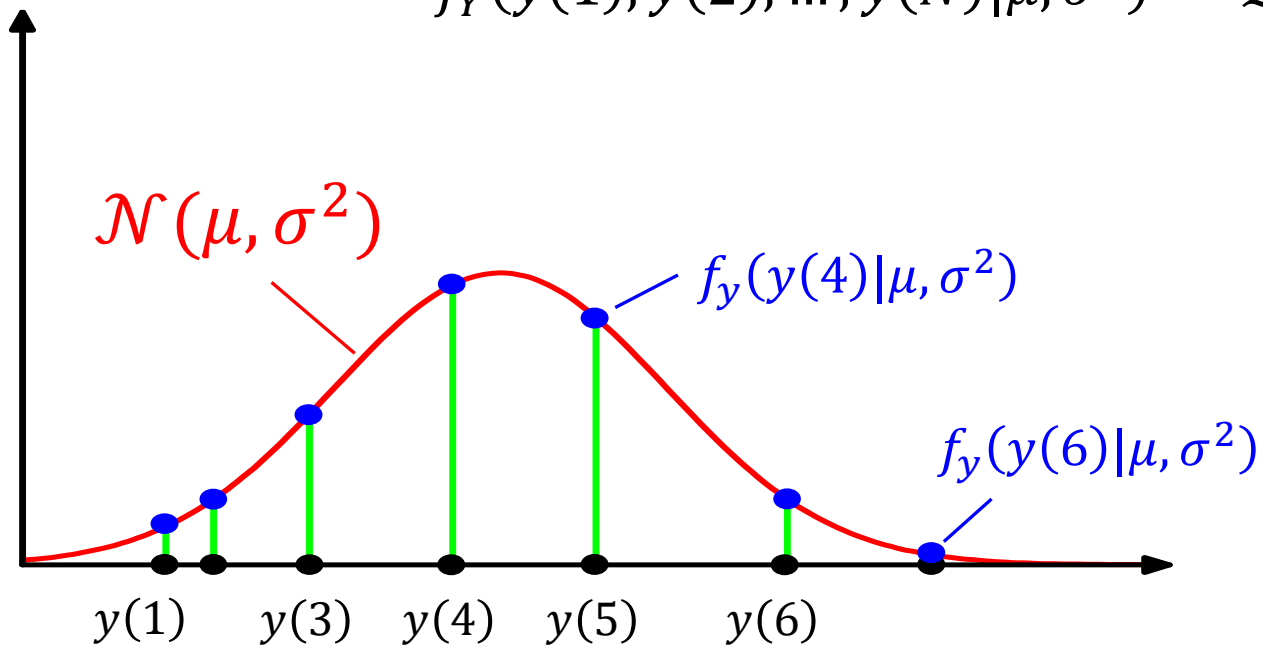
☐  $\hat{\mu} = 2$



# Stima a massima verosimiglianza

L'esempio precedente considerava il caso in cui avevamo un solo dato osservato. Nel caso di **più osservazioni i.i.d.** di  $y \sim \mathcal{N}(\mu, \sigma^2)$ , ovvero  $Y = [y(1), y(2), \dots, y(N)]^T$ , devo comunque **massimizzare la verosimiglianza**, cioè

$$f_Y(y(1), y(2), \dots, y(N) | \mu, \sigma^2) = \mathcal{L}(\mu, \sigma^2 | Y) = \prod_{i=1}^N \mathcal{N}(y(i) | \mu, \sigma^2)$$



**Massimizzare la verosimiglianza** significa «cambiare» i valori dei parametri  $\mu$  e  $\sigma^2$  tale che il **prodotto dei puntini blu ●** è **massimizzato**

# Stima a massima verosimiglianza

La stima a massima verosimiglianza dell'esempio precedente può essere espressa come:

$$\hat{\boldsymbol{\theta}}_{\text{ML}} = \begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix}_{2 \times 1} = \arg \max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|Y) = \arg \max_{\boldsymbol{\theta}} \prod_{i=1}^N \mathcal{N}(y(i)|\mu, \sigma^2)$$

In generale posso attribuire ai dati qualsiasi distribuzione di probabilità  $d(\cdot)$ , sia continua che discreta

$$\hat{\boldsymbol{\theta}}_{\text{ML}} = \arg \max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|Y)$$

$d \times 1$

# Stima a massima verosimiglianza

Spesso, anziché massimizzare  $\mathcal{L}(\theta|Y)$ , si massimizza il suo **logaritmo naturale**

- Dato che il logaritmo è una funzione monotona crescente,  $\ln \mathcal{L}(\theta|Y)$  **ha lo stesso massimo** di  $\mathcal{L}(\theta|Y)$
- Usare il logaritmo è efficiente da un **punto di vista implementativo**, perchè evita possibili underflow dati dal prodotto di piccole probabilità (sostituendolo con la somma delle log-probabilità)

$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta} \ln \mathcal{L}(\theta|Y)$$

$d \times 1$

A meno di casi particolari fortunati, l'ottimizzazione è effettuata con **metodi iterativi**



# Stima a massima verosimiglianza: proprietà

Lo stimatore a massima verosimiglianza gode di **buone proprietà**. Infatti, esso è:

**1. Asintoticamente corretto:**  $\lim_{N \rightarrow +\infty} \mathbb{E}[\hat{\theta}_{\text{ML}}] = \theta^0$

**Lo stimatore può essere distorto.** Per esempio lo stimatore a massima verosimiglianza della varianza di una popolazione Guassiana è distorto

**2. Consistente:** più  $N$  è grande, più la stima è precisa

**3. Asintoticamente efficiente:**  $\lim_{N \rightarrow +\infty} \text{Var}[\hat{\theta}_{\text{ML}}] = M^{-1}$   $M$ : Matrice di informazione di Fisher

**4. Asintoticamente normale:**  $\hat{\theta}_{\text{ML}} \sim \mathcal{N}(\theta^0, M^{-1})$  per  $N \rightarrow +\infty$

# Outline

1. Stima a massima verosimiglianza

**2. Stima a massima verosimiglianza di parametri della popolazione**

3. Stima a massima verosimiglianza di modelli lineari



# Stima ML di parametri della popolazione

Consideriamo il caso in cui vogliamo **stimare la media**  $\mu$  di una popolazione di variabili casuali Gaussiane, supponendo di **conoscere la varianza** della distribuzione

Assumiamo di avere osservato **2 dati**  $y(i) \sim \mathcal{N}(\mu, \sigma^2 = 1), i = 1, 2$ , i.i.d., tali che i valori osservati sono  $y(1) = 4, y(2) = 6$

La forma della **pdf delle singole variabili**  $y(i)$  è:

$$f_y(y(i)|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{1}{2} \left( \frac{y(i) - \mu}{\sigma} \right)^2 \right] = \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} (y(i) - \mu)^2 \right]$$

# Stima ML di parametri della popolazione

Il **valore assunto dalla pdf** in corrispondenza delle due osservazioni è:

$$y(1) = 4$$



$$f_y(y(1) = 4 | \mu, \sigma^2 = 1) = \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} (4 - \mu)^2 \right]$$

$$y(2) = 6$$



$$f_y(y(2) = 6 | \mu, \sigma^2 = 1) = \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} (6 - \mu)^2 \right]$$

La **distribuzione congiunta** è il prodotto delle due pdf singole (essendo i dati i.i.d.)

$$p(y(1), y(2) | \mu, \sigma^2 = 1) = \left( \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} (4 - \mu)^2 \right] \right) \cdot \left( \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} (6 - \mu)^2 \right] \right)$$

# Stima ML di parametri della popolazione

La pdf congiunta è funzione solo di  $\mu$ , poichè il **valore dei dati è noto**. Con questa interpretazione, la pdf congiunta è la **funzione di verosimiglianza** (likelihood function)

$$\mathcal{L}(\mu|y(1) = 4, y(2) = 6) = \left( \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} (4 - \mu)^2 \right] \right) \cdot \left( \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} (6 - \mu)^2 \right] \right)$$

La pdf congiunta è funzione solo di  $\mu$ , poichè il **valore dei dati è noto**. Con questa interpretazione, la pdf congiunta è la **funzione di verosimiglianza** (likelihood function)

La stima  $\hat{\mu}_{ML}$  è valore di che **massimizza** la verosimiglianza

$$\hat{\mu}_{ML} = \arg \max_{\mu} \ln \mathcal{L}(\mu|y(1) = 4, y(2) = 6)$$

# Stima ML di parametri della popolazione

È più conveniente massimizzare il logaritmo della verosimiglianza. Questa nuova funzione (la **log-verosimiglianza**) ha lo stesso massimo della verosimiglianza

$$\begin{aligned}\ln(\mathcal{L}) &= \ln \left[ \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} (4 - \mu)^2 \right) \cdot \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} (6 - \mu)^2 \right) \right] \\&= \ln \left[ \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} (4 - \mu)^2 \right) \right] + \ln \left[ \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} (6 - \mu)^2 \right) \right] \\&= \ln \frac{1}{\sqrt{2\pi}} + \ln \left[ \exp \left( -\frac{1}{2} (4 - \mu)^2 \right) \right] + \ln \frac{1}{\sqrt{2\pi}} + \ln \left[ \exp \left( -\frac{1}{2} (6 - \mu)^2 \right) \right] \\&= 2 \cdot \ln \frac{1}{\sqrt{2\pi}} - \frac{1}{2} (4 - \mu)^2 - \frac{1}{2} (6 - \mu)^2\end{aligned}$$

# Stima ML di parametri della popolazione

Massimizzando l'espressione ottenuta rispetto a  $\mu$  otteniamo:

$$\frac{\partial \ln \mathcal{L}}{\partial \mu} = 0 \Rightarrow (4 - \mu) + (6 - \mu) = 0 \Rightarrow \hat{\mu}_{ML} = \frac{4 + 6}{2} = 5$$

La **stima a massima verosimiglianza** del parametro  $\mu$  per il modello Gaussiano è uguale allo stima ottenuta tramite lo **stimatore media campionaria**!

Questo risultato, seppur non generalizzabile, rende molto interpretabile ed intuitivo lo stimatore a massima verosimiglianza

# Stima ML di parametri della popolazione

**Osservazione:** **massimizzare** la «log-verosimiglianza» equivale a **minimizzare** la «meno log-verosimiglianza»

$$\hat{\boldsymbol{\theta}}_{\text{ML}} = \arg \max_{\boldsymbol{\theta}} \ln[ \mathcal{L}(\boldsymbol{\theta}|Y) ] = \arg \min_{\boldsymbol{\theta}} -\ln[ \mathcal{L}(\boldsymbol{\theta}|Y) ]$$

Formulando il problema di stima a massima verosimiglianza in questo modo, abbiamo un problema di **minimizzazione** proprio come con lo stimatore a minimi quadrati!

$$\hat{\boldsymbol{\theta}}_{\text{LS}} = \arg \min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

$$J(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^N (y(i) - \boldsymbol{\varphi}^{\text{T}}(i)\boldsymbol{\theta})^2$$



# Outline

1. Stima a massima verosimiglianza
2. Stima a massima verosimiglianza di parametri della popolazione
- 3. Stima a massima verosimiglianza di modelli lineari**



# Stima ML di modelli lineari

Il metodo della massima verosimiglianza (ML) può essere usato anche per **stimare modelli lineari**. Quello che bisogna fare è imporre un **modello probabilistico** alle osservazioni  $y(i)$

$$y(i) = \theta_0 + \theta_1 \varphi_1(i) + \dots + \theta_{d-1} \varphi_{d-1}(i) + \epsilon(i) = \underset{1 \times d}{\boldsymbol{\varphi}^\top(i)} \underset{d \times 1}{\boldsymbol{\theta}} + \underset{1 \times 1}{\epsilon(i)}$$

In particolare, se **assumiamo** che  $\epsilon(i) \sim \mathcal{N}(0, \lambda^2)$  i.i.d.



$$y(i) \sim \mathcal{N}(\underset{1 \times d}{\boldsymbol{\varphi}^\top(i)} \underset{d \times 1}{\boldsymbol{\theta}}, \lambda^2) \text{ i.i.d.}$$

$$\underset{d \times 1}{\boldsymbol{\varphi}} = \begin{bmatrix} 1 \\ \varphi_1 \\ \vdots \\ \varphi_{d-1} \end{bmatrix} \quad \underset{d \times 1}{\boldsymbol{\theta}} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_{d-1} \end{bmatrix}$$

La media  $\mu(i)$  di  $y(i)$  è espressa come combinazione lineare dei regressori,  $\mu(i) = \boldsymbol{\varphi}(i)^\top \boldsymbol{\theta}$ !

# Stima ML di modelli lineari

La **distribuzione congiunta** dei dati è:

$$\begin{aligned} f_Y(y(1), y(2), \dots, y(N) | X, \boldsymbol{\theta}, \lambda^2) &\stackrel{\text{i.i.d.}}{=} \prod_{i=1}^N f_y(y(i) | \boldsymbol{\varphi}(i), \boldsymbol{\theta}, \lambda^2) \\ &= \prod_{i=1}^N \mathcal{N}(y(i) | \boldsymbol{\varphi}(i), \boldsymbol{\theta}, \lambda^2) \\ &= \prod_{i=1}^N \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{y(i) - \boldsymbol{\varphi}^\top(i) \boldsymbol{\theta}}{\lambda} \right)^2 \right] \\ &= \mathcal{L}(\boldsymbol{\theta} | Y, X, \lambda^2) \quad \text{Supponiamo } \lambda^2 \text{ noto} \end{aligned}$$

# Stima ML di modelli lineari

Calcoliamo la **log-verosimiglianza**

$$\begin{aligned}\ln[\mathcal{L}(\boldsymbol{\theta}|Y, X, \lambda^2)] &= \ln \left[ \prod_{i=1}^N \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{y(i) - \boldsymbol{\varphi}^\top(i)\boldsymbol{\theta}}{\lambda} \right)^2 \right] \right] \\ &= \sum_{i=1}^N \ln \left[ \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{y(i) - \boldsymbol{\varphi}^\top(i)\boldsymbol{\theta}}{\lambda} \right)^2 \right] \right] \\ &= \sum_{i=1}^N \ln \frac{1}{\sqrt{2\pi}} + \sum_{i=1}^N \ln \left[ \exp \left[ -\frac{1}{2} \left( \frac{y(i) - \boldsymbol{\varphi}^\top(i)\boldsymbol{\theta}}{\lambda} \right)^2 \right] \right]\end{aligned}$$

# Stima ML di modelli lineari

$$\ln[\mathcal{L}(\boldsymbol{\theta}|Y, X, \lambda^2)] = N \cdot \ln \frac{1}{\sqrt{2\pi}} + \sum_{i=1}^N -\frac{1}{2} \left( \frac{y(i) - \boldsymbol{\varphi}^\top(i)\boldsymbol{\theta}}{\lambda} \right)^2$$

$$= N \cdot \ln(2\pi)^{-\frac{1}{2}} - \frac{1}{2\lambda^2} \sum_{i=1}^N (y(i) - \boldsymbol{\varphi}^\top(i)\boldsymbol{\theta})^2$$

$$= -\frac{1}{2} N \cdot \ln 2\pi - \frac{1}{2\lambda^2} \sum_{i=1}^N (y(i) - \boldsymbol{\varphi}^\top(i)\boldsymbol{\theta})^2$$

# Stima ML di modelli lineari

**Calcolare il massimo** di  $\ln[\mathcal{L}(\boldsymbol{\theta}|Y, X, \lambda^2)]$  è equivalente a **calcolare il minimo** di  $-\ln[\mathcal{L}(\boldsymbol{\theta}|Y, X, \lambda^2)]$ , per cui:

Siccome non dipende da  $\boldsymbol{\theta}$ , questo termine non contribuisce al calcolo del minimo

$$-\ln[\mathcal{L}(\boldsymbol{\theta}|Y, X, \lambda^2)] = + \cancel{\frac{1}{2} N \ln 2\pi} + \frac{1}{2\lambda^2} \sum_{i=1}^N (y(i) - \boldsymbol{\varphi}^\top(i)\boldsymbol{\theta})^2$$

$$\hat{\boldsymbol{\theta}}_{\text{ML}} = \arg \min_{\substack{\boldsymbol{\theta} \\ d \times 1}} \frac{1}{2\lambda^2} \sum_{i=1}^N (y(i) - \underset{1 \times d}{\boldsymbol{\varphi}^\top(i)} \underset{d \times 1}{\boldsymbol{\theta}})^2$$

La stima ML del modello lineare  $y(i) - \boldsymbol{\varphi}^\top(i)\boldsymbol{\theta} + \epsilon(i)$ , con  $\epsilon(i) \sim \mathcal{N}(0, \lambda^2)$  i.i.d., è **equivalente alla stima a minimi quadrati** (che non aveva assunzioni sulla pdf dei dati)

# Stima ML di modelli lineari

**Osservazione:** cambiando ipotesi sulla distribuzione del rumore (e quindi dei dati), si ottengono **altre funzioni di costo** e **altri modelli**, che modellano i dati in modo diverso rispetto alla regressione lineare

Uno di questi altri modelli (che vedremo nella prossima lezione) è il modello di **regressione logistica**



**UNIVERSITÀ  
DEGLI STUDI  
DI BERGAMO**

Dipartimento  
di Ingegneria Gestionale,  
dell'Informazione e della Produzione