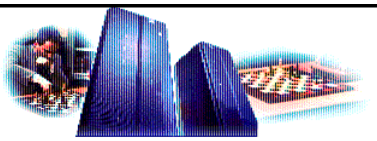
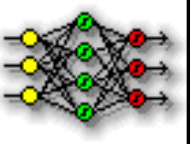




Sistemi Intelligenti
Stimatori e sistemi lineari - III

Alberto Borghese

Università degli Studi di Milano
Laboratory of Applied Intelligent Systems (AIS-Lab)
Dipartimento di Informatica
borghese@di.unimi.it

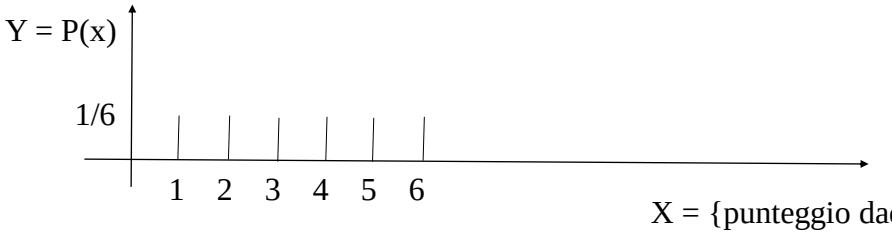
Overview

Densità di probabilità

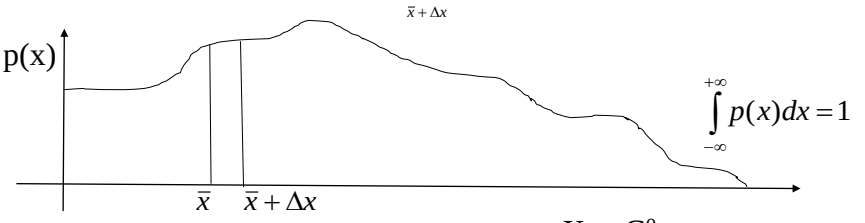
Funzione di verosimiglianza

Stima alla massima verosimiglianza

La probabilità nel caso continuo



X = {punteggio dado}



$X \in C^0$

$$P(x \in [\bar{x}, \bar{x}+\Delta x]) \int_{\bar{x}}^{\bar{x}+\Delta x} p(x)dx$$

Definizione di p(x)

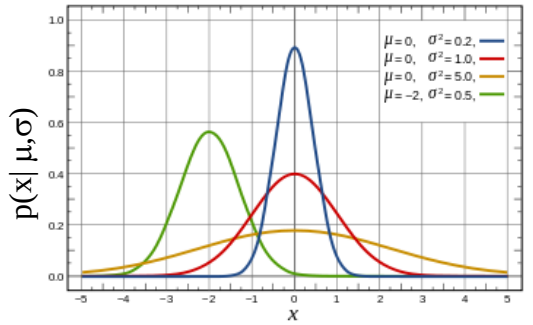
Caso discreto: prescrizione della probabilità per ognuno dei finiti valori che la variabile X può assumere: P(X).

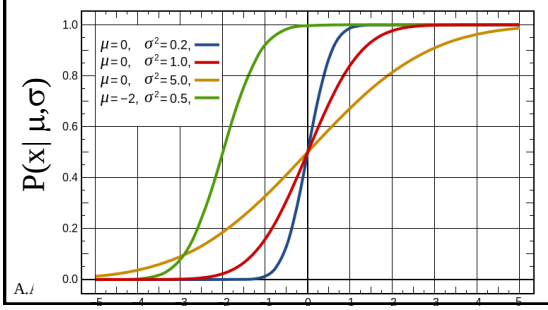
Caso continuo: i valori che X può assumere sono infiniti. Devo trovare un modo per definirne la probabilità. Descrizione **analitica** mediante la funzione densità di probabilità. Si considera la probabilità che x cada in un certo intervallo.

Valgono le stesse relazioni del caso discreto, dove alla somma si sostituisce l'integrale.

$$P(X = x \in [\bar{x}, \bar{x}+\Delta x]) \int_{\bar{x}}^{\bar{x}+\Delta x} \int_{-\infty}^{+\infty} p(x,y)dxdy$$

Distribuzioni notevoli: la Gaussiana





$$p(x \mid \mu,\sigma) = \frac{1}{\left(\sqrt{2\pi}\right)\Sigma^{D}} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\Sigma}\right)^{2}\right]$$

D = dimensione, in questo caso D = 1

$\Pr(|X-\mu| < \sigma) = 0.68268$

$\Pr(|X-\mu| < 2\sigma) = 0.95452$

$\Pr(|X-\mu| < 3\sigma) = 0.9973$

I momenti di una variabile statistica

$$\mu^{k}(X) = \int_{-\infty}^{+\infty} (x-a)^{k}\, p(x)dx$$ Momento rispetto ad a, solitamente alla media

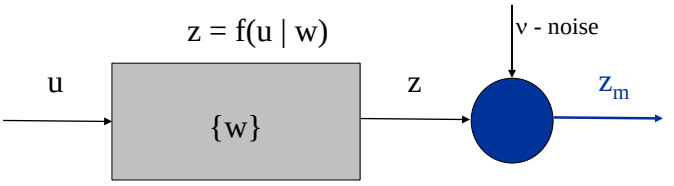
$$E[X] = \int_{-\infty}^{+\infty} (x-\mu)\, p(x)$$ Valore atteso (Expected value) di X = media distribuzione

$$E[(X-\mu)^{2}] = \int_{-\infty}^{+\infty} (x-\mu)^{2}\, p(x)$$ Varianza (σ^2)

$$E[(X-\mu)^{3}] = \int_{-\infty}^{+\infty} (x-\mu)^{3}\, p(x)$$ Asimmetria

$$E[(X-\mu)^{4}] = \int_{-\infty}^{+\infty} (x-\mu)^{4}\, p(x)$$ Kurtosi – peso delle code di p(x)

Modello



u – causa => z_m – effetto (misurato on errore)

Control / Classification / Prediction: determine {z} from {u},{w}

Inverse problem: determine cause {u} from $\{z_m\}$,{w}

Inverse problem: Identification: determine {w} from {u},$\{z_m\}$ - *Learning*

Overview

Densità di probabilità

Funzione di verosimiglianza

Stima alla massima verosimiglianza

Probabilità di un certo insieme di misure

$p(z_{1m}, z_{2m}, z_{3m}) = \prod_i p(z_{im})$

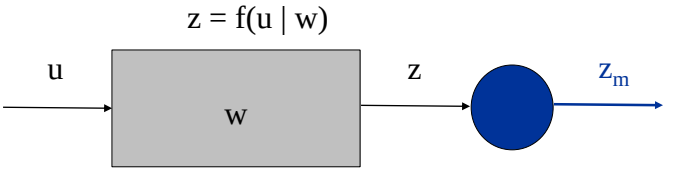
$z = f(u \mid w)$ misuro $\{z_i\}$ in corrispondenza di $\{u_i\}$ con i parametri $\{w_j\}$

Avrò che: $f(u_i, w) = z_{i,m} + z_i + v_i \Rightarrow f(u_i, w) - z_{im} = v_i$ the noise

Se le misure sono indipendenti posso scrivere che la probabilità di ottenere le misure z_{1m}, z_{2m}, z_{3m} ….. è:

$p(z_{1m}, z_{2m}, z_{3m}) = \prod_i p(z_{im})$

Dipendenza delle misure



Le misure dipendono dal valore dalle variabili in ingresso u e dai parametri w.

$p(z_{1m}, z_{2m}, z_{3m}) = \prod_i p(z_{im} \mid u_i, w) = \prod_i p(f(u_i; w) \mid u_{i,} w)$

Tanto più i parametri saranno corretti tanto maggiore sarà la probabilità di avere y_m dal modello.

Fitting di una retta

Vogliamo stimare i parametri di una retta: $z = f(u \mid w) = m\,u + q$, con m e q incogniti: $W = \{m, q\}$

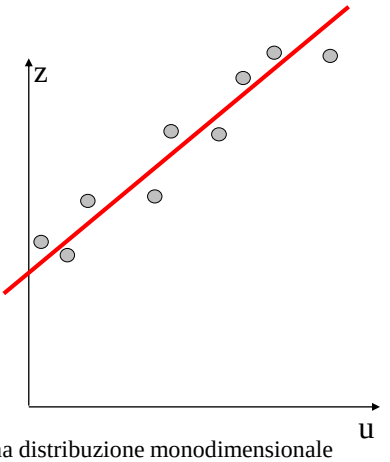
La retta è un modello lineare.

Abbiamo a disposizione N misure rumorose effettuate: z_{im} che sono funzione di u e w: $z_{im} = \{z_{im}; u_i, w\}$, con rumore su z_i

Supponiamo che le z_{im} siano affette da **errore Gaussiano a media nulla.** In pratica:

$z_{im} = z_i + \nu_i = m\,u_i + q + \nu_i$ dove ν_i è l'errore di misura.

$\nu_i = z_{im} - m\,u_i + q$ è Gaussiano a media nulla.



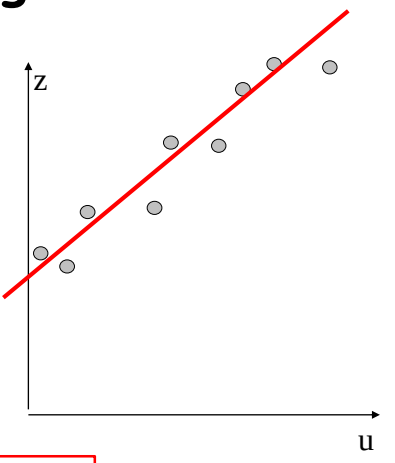
Possiamo anche scrivere che: $z_{im} = z_i + G(\mu, \sigma^2)$ indica una distribuzione monodimensionale gaussiana a media μ e varianza σ^2. Errore di misura: $G(0, \sigma^2)$

Stima ai minimi quadrati e verosimiglianza

Per ogni punto, dovrebbe valere $z_i = m\,u_i + q$.

Ma c'è l'errore di misura, misuriamo in realtà $z_i + \nu_i$.

Cerchiamo i parametri m e q che sono più verosimili.



Cosa vuol dire che sono più verosimili?
Quanto sono più verosimili?

Funzione di verosimiglianza

- Siano date **N variabili casuali <u>indipendenti</u>**… Quale è la **probabilità di misurare il vettore [z_{1m}, …, z_{Nm}]**?

$$p(z_{1m}, z_{2m}, ..., z_{Nm}) = p(z_{1m}) \cdot p(z_{2m}) \cdot ... p(z_{Nm}) = L(z_{1m}, z_{2m}, ..., z_{Nm})$$

- La probabilità congiunta è il prodotto delle probabilità semplici.
- Questa è la **Funzione di verosimiglianza** o **funzione di Likelihood**, L(.)
- In questo caso le z sono legate alle u da f(u,w).

Overview

Funzione di verosimiglianza

Stima alla massima verosimiglianza

Stima alla massima verosimiglianza

- Se **massimizziamo** L=L(**z** | u, w) rispetto a w
- troviamo i parametri w tali per cui è massima la probabilità di misurare il vettore di dati $\mathbf{z_m} = \{z_{im}, i=1\ldots N\}$.
- **Stima alla massima verosimiglianza**.
- Più in generale, le variabili possono avere densità di probabilità diverse, ciascuna descritta da un set di parametri. I parametri delle diverse densità di probabilità possono essere calcolati utilizzando l'approccio alla massima verosimiglianza…
- La funzione di verosimiglianza dipende dai parametri che definiscono le densità di probabilità delle variabili casuali che entrano nella verosimiglianza…
- Massimizzando la funzione di verosimiglianza rispetto a tali parametri se ne effettua la stima in modo tale che il vettore osservato $z = \{z_{im}\}$ i =1…N sia massimamente probabile (massima verosimiglianza).

Stima alla massima verosimiglianza per modello lineare

- Impostiamo il problema scrivendo la funzione di verosimiglianza e massimizzando tale funzione rispetto a m e q…
- Scriviamo prima di tutto la densità di probabilità di ottenere z_{im} per ciascun dato:

$$p(z_{im} \mid m, q, u_i) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{z_{im} - (m u_i + q)}{\sigma} \right)^2 \right]$$

Dove m e q non sono note. $z_{im} = z_i + \nu_i = m\,u + q + \nu_i$

Stima a massima verosimiglianza

Sapendo che le misure sono indipendenti, possiamo scrivere la probabilità di ottenere le N misure $\{z_{im}\}$: funzione di verosimiglianza.

Scriviamo il logaritmo negativo della verosimiglianza per **errore Gaussiano.**

$$p(z_{1m}, z_{2m}, ..., z_{Nm}) = p(z_{1m}) \cdot p(z_{2m}) \cdot ... p(z_{Nm}) = L(z_{1m}, z_{2m}, ..., z_{Nm}) = \prod_i p(z_{im})$$

$$-\ln(L(.) = -\ln \prod_{i=1}^{N} p(z_{im}) = f(z_{1m}, z_{2m}....z_{Nm}; m, q; u_1, u_2.....u_N) = -\sum_{i=1}^{N} \ln\left\{\frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{z_{im} - mu_i - q}{\sigma}\right)^2\right]\right\} =$$

$$-\sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \sum_{i=1}^{N}\left[-\frac{1}{2}\left(\frac{z_{im} - mu_i - q}{\sigma}\right)^2\right] =$$

$$= -\sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) + \frac{1}{2\sigma^2}\sum_{i=1}^{N}(z_{im} - mu_i - q)^2$$

Stima a massima verosimiglianza

- E massimizziamo L(.) ponendo a zero le derivate prime rispetto a m:

$$\frac{\partial f(y_{1m}, y_{2m}.....y_{Nm;}m, q; x_{1m}, x_{2m}.....x_{Nm})}{\partial m} = \frac{\partial}{\partial m}\left[-\sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) + \frac{1}{2\sigma^2}\sum_{i=1}^{N}(z_{im} - m \cdot u_i - q)^2\right] =$$

$$= 0 + \frac{1}{2\sigma^2}\sum_{i=1}^{N}(z_{im} - m \cdot u_i - q) \cdot 2 \cdot (-u_i) =$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{N}(z_{im} - m \cdot u_i - q) \cdot x_i = 0 \Rightarrow \sum_{i=1}^{N}(z_i - m \cdot u_i - q) \cdot u_i = 0 \Rightarrow$$

$$\left[\sum_{i=1}^{N}(z_{im} \cdot u_i)\right] - m \cdot \left[\sum_{i=1}^{N}(u_i^2)\right] - q \cdot \left[\sum_{i=1}^{N}(u_i)\right] = 0 \Rightarrow$$

$$m \cdot \left[\sum_{i=1}^{N}(u_i^2)\right] + q \cdot \left[\sum_{i=1}^{N}(u_i)\right] = \left[\sum_{i=1}^{N}(z_{im} \cdot u_i)\right]$$

1ª equazione

Stima a massima verosimiglianza

... e a q:

$$= 0 + \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(z_{im} - m\cdot u_i - q\right)\cdot 2\cdot(-1) =$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{N}\left(z_{im} - m\cdot u_i - q\right) = 0 \Rightarrow \sum_{i=1}^{N}\left(z_{im} - m\cdot u_i - q\right) = 0 \Rightarrow$$

$$\left[\sum_{i=1}^{N}\left(z_{im}\right)\right] - m\cdot\left[\sum_{i=1}^{N}\left(u_i\right)\right] - q\cdot\left[\sum_{i=1}^{N}(1)\right] = 0 \Rightarrow$$

$$m\cdot\left[\sum_{i=1}^{N}\left(u_i\right)\right] + q\cdot\left[\sum_{i=1}^{N}(1)\right] = \left[\sum_{i=1}^{N}\left(z_{im}\right)\right]$$

2ª equazione

Stima a massima verosimiglianza

$$\left[\sum_{i=1}^{N}\left(u_i^{\,2}\right)\right]\cdot m + \left[\sum_{i=1}^{N}\left(u_i\right)\right]\cdot q = \left[\sum_{i=1}^{N}\left(z_{im}\cdot u_i\right)\right]$$

1° equazione

$$\left[\sum_{i=1}^{N}\left(u_i\right)\right]\cdot m + \left[\sum_{i=1}^{N}(1)\right]\cdot q = \left[\sum_{i=1}^{N}\left(z_{im}\right)\right]$$

2° equazione

Le incognite, m e b, compaiono con esponente 1 => equazioni lineari in m e q
Potrei risolvere per sostituzione

Soluzione al problema di stima

Ottengo un sistema di due equazioni in due incognite:

Ax = b

Dove:

$$A = \begin{bmatrix} \sum_{i=1}^{N}(u_i^2) & \sum_{i=1}^{N}(u_i) \\ \sum_{i=1}^{N}(u_i) & N \end{bmatrix} \qquad b = \begin{bmatrix} \sum_{i=1}^{N}(z_{im}u_i) \\ \sum_{i=1}^{N}(z_{im}) \end{bmatrix}$$

x = [m q]'

Esempio

z = 2 u + 1 **m = 2; q = 1**

Misuro con errore e ottengo:
- $u_1 = 1$; $z_1 = 2,8$
- $u_2 = 0$; $z_2 = 1,2$

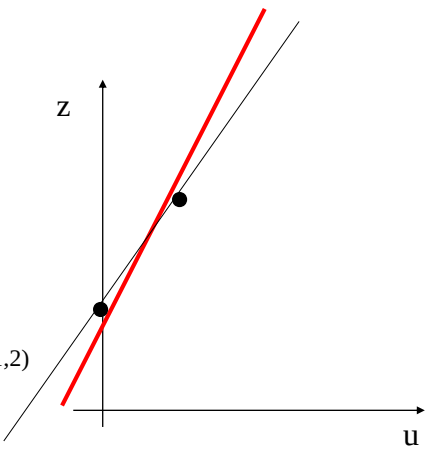
Quanto varranno le stime di m e q?

$$\left[\sum_{i=1}^{N}(u_i^2)\right]\cdot m + \left[\sum_{i=1}^{N}(u_i)\right]\cdot q = \left[\sum_{i=1}^{N}(z_i \cdot u_i)\right]$$

=> (1*1 + 0*0)m + (1 + 0) q = (1 * 2,8 + 0 *1,2)

$$\left[\sum_{i=1}^{N}(u_i)\right]\cdot m + \left[\sum_{i=1}^{N}(1)\right]\cdot q = \left[\sum_{i=1}^{N}(z_i)\right]$$

=> (1 + 0) m_e + 2 q_e = (2,8+1,2)



$m_e + q_e = 2,8$ $m_e+2q_e = 4$ => per sottrazione **$q_e = 1,2$; $m_e = 1,6$** NB 2,8 = 1,6*1 + 1,2

Problema lineare

Ho N misure indipendenti: $A_i = [u_i\ 1]$

$\{z_{im} = m\,u_i + q + \nu\}$

$A\,x = b$

$$A_{Mx2} = \begin{bmatrix} u_1 & 1 \\ . & . \\ u_M & 1 \end{bmatrix}$$

$$x_{2x1} = \begin{bmatrix} m \\ q \end{bmatrix} \qquad b_{Mx1} = \begin{bmatrix} z_1 \\ . \\ z_M \end{bmatrix}$$

Dal Sistema alla soluzione ai minimi quadrati

$$A^TA = \begin{bmatrix} u_1 & . & u_M \\ 1 & . & 1 \end{bmatrix} \begin{bmatrix} u_1 & 1 \\ . & . \\ u_M & 1 \end{bmatrix} = \begin{bmatrix} \left[\sum_{i=1}^{N}\left(u_i^{\,2}\right)\right] & \left[\sum_{i=1}^{N}\left(u_i\right)\right] \\ \left[\sum_{i=1}^{N}\left(u_i\right)\right] & N \end{bmatrix}$$

$$A^Tb = \begin{bmatrix} u_1 & . & u_M \\ 1 & . & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ . \\ z_M \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N}\left(z_i \cdot u_i\right) \\ \sum_{i=1}^{N}\left(z_i\right) \end{bmatrix}$$

Equazioni normali: $\mathbf{A^TAx = A^tb}$

sono le stesse ottenute dalla massima verosimiglianza

Soluzione come problema di ottimizzazione

Funzione costo: $(Ax - b)^2 = \sum_k \nu_k^{\,2} = \|Ax - b\|^2$

Assegno un costo al fatto che la soluzione x, non soddisfi tutte le equazioni, la somma dei residui associati ad ogni equazioni viene minimizzata. Geometricamente: viene trovato il punto a distanza (verticale) minima da tutte le rette.

$$\min_{\mathbf{x}} \sum_k \nu_k^{\,2} = \min_{\mathbf{x}} (\mathbf{Ax} - \mathbf{b})^2$$

$$\frac{d}{d\mathbf{x}} (\mathbf{Ax} - \mathbf{b})^2 = 2\mathbf{A}^{\mathbf{T}} (\mathbf{Ax} - \mathbf{b}) = 0$$

$$\mathbf{A^T A x = A^T b}$$

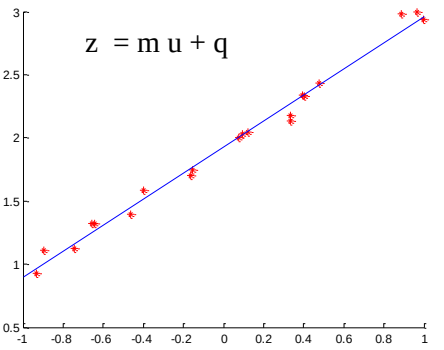
$$\mathbf{x = (A^T A)^{-1} A^T b}$$

NB le funzioni costo sono spesso quadratiche (problemi di minimizzazione convessi) perchè il costo cresce sia che il modello sovrastimi che sottostimi le misure.
Inoltre, le derivate calcolate per imporre le condizioni di stazionarietà (minimo), sono relativamente semplici.

Esempio - Caso 2D (2 parametri)

N = 20 punti
$\sigma_o^2 = 0.01$
m reale = 1
q reale = 2

m stimato = 0.9931
q stimato = 2.0106



Cosa vuol dire che {m,q} sono i più verosimili?
Quanto sono più verosimili?

Stima a massima verosimiglianza e minimi quadrati

$\mathbf{A^T A x = A^T b}$

$\mathbf{x = (A^T A)^{-1} A^T b}$

La soluzione a massima verosimiglianza, quando il rumore è Gaussiano a media nulla, coincide con la soluzione ai minimi quadrati del sistema lineare associato (la soluzione ai minimi quadrati è un caso particolare della stima alla massima verosimiglianza).

La soluzione è quella che minimizza lo scarto quadratico medio dei residui, ovverosia è a minima varianza.

La stima a massima verosimiglianza è un approccio generale, e si presta a p(x) di qualsiasi forma. La Gaussiana consente di ottenere una formulazione lineare del problema.

Stima ai minimi quadrati e verosimiglianza

- Nella soluzione ai minimi quadrati del sistema lineare Ax=b si definisce un vettore errore v = Ax - b;
- Nel caso di soluzione "perfetta" v = 0;
- Dal momento che abbiamo un numero di equazioni maggiore rispetto al numero di incognite, cerchiamo il vettore e a norma minima;
- In pratica cerchiamo x t.c. $v^T v = \Sigma_i v_i^2$ è minimo.

Relazione con la soluzione dei sistemi lineari

- Nel caso precedente le incognite erano (x_1,x_2) cioè le coordinate di un punto del piano.
- In questo caso le incognite sono (m,q) i parametri della retta.

- Nel caso precedente i dati erano i parametri di ogni retta (m_i, q_i / a_{1i}, a_{2i})
- In questo caso i dati sono i punti misurati sulla retta (u_i, z_i)

Giustificazione statistica

- **C'è un solo insieme vero dei parametri**, mentre ci **possono essere infiniti universi di dati per effetto dell'errore di misura**.

- La domanda quindi più corretta sarebbe: "Dato un certo insieme di parametri, qual'è la probabilità che questo insieme di dati sia estratto?" (più correttamente si parla di densità di probabilità?)

- Cioè, **per ogni insieme di parametri, calcoliamo la probabilità che i dati siano estratti. Ovverosia la likelihood (verosimiglianza) dei dati, dato un certo insieme di parametri.**

La stima ai minimi quadrati dei parametri è equivalente a determinare i parametri che massimizzano la funzione di **verosimiglianza** sotto l'ipotesi di errore **Gaussiano a media nulla**.

Overview

Densità di probabilità

Funzione di verosimiglianza

Stima alla massima verosimiglianza