

Sistemi Intelligenti Stimatori e sistemi lineari - III

Alberto Borghese

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



Overview

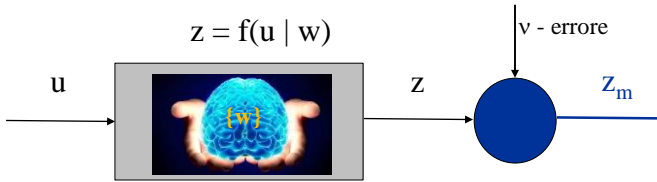


Stima alla massima verosimiglianza

Varianza della stima nei sistemi lineari



I problemi associati ai Modelli



u – causa $\Rightarrow z$ (effetto); z_m – effetto (misurato con errore)

Control / Classification / Prediction: determine $\{z\}$ from $\{u\}, \{w\}$ – utilizzo forward

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

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



Probabilità di un certo insieme di misure

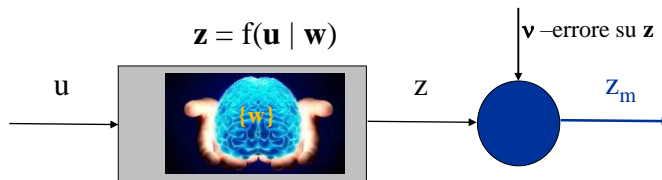
$z = f(u | w)$ misura $\{z_i\}$ in corrispondenza di $\{u_i\}$. $\{z_i\}$ è ottenuto come uscita del modello, tramite i parametri $\{w_j\}$

Z è misurato con un errore (statistico). Avrò che: $z_{i,m} = z_i + v_i$

Se le **misure sono indipendenti** posso scrivere che la probabilità di ottenere le misure: $z_{1,m}, z_{2,m}, z_{3,m}, \dots$ È la probabilità congiunta:

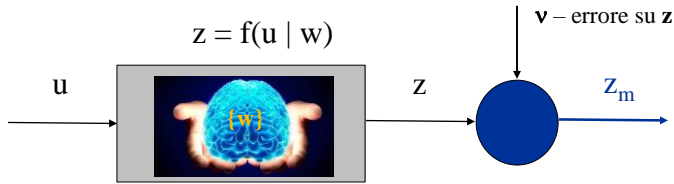
$$p(z_{1,m}, z_{2,m}, z_{3,m}) = \prod_i p(z_{i,m})$$

(cf. dadi nel caso discreto)





Dipendenza delle misure



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

$$L = p(z_{1m}, z_{2m}, z_{3m}, \dots, z_{Nm}) = \prod_i P(z_{im} | u_i, w) =$$

Scrivo la probabilità esplicitamente come condizionata al valore di u e di w .

La probabilità congiunta è chiamata funzione di Likelihood (verosimiglianza), $L(\cdot)$.

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



Funzione di verosimiglianza

- Siano date N variabili casuali **indipendenti** $\{z_1, \dots, z_N\}$? ... Quale è la **probabilità di misurare il vettore** $\{z_{1m}, \dots, z_{Nm}\}$?

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

- La probabilità congiunta è il prodotto delle probabilità semplici (**misure indipendenti tra loro**).
- Questa è la **Funzione di verosimiglianza** o **funzione di Likelihood**, $L(\cdot)$



Stima alla massima verosimiglianza



- Troviamo i parametri $\{w\}$ tali per cui è massima la probabilità di misurare il vettore di misure:
 $\mathbf{z}_m = \{z_{im}, i=1 \dots N\}$.
- **Stima alla massima verosimiglianza.**
- **Massimizziamo** $L=L(\mathbf{z}_m | u, w)$ rispetto a w

$$L(z_{1m}, z_{2m}, z_{3m}, \dots, z_{Nm} | (w; u_1, u_2, u_3, \dots, u_{Nm})) =$$

$$= p(z_{1m} | w; u_1) p(z_{2m} | w; u_2) p(z_{3m} | w; u_3) \dots p(z_{Nm} | w; u_N)$$



Forma alternativa della L(.)



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

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

Scriviamo il **negativo del logaritmo** della verosimiglianza:

$$-\ln(L(.)) = -\ln \prod_{i=1}^N p(z_{im}) = f(z_{1m}, z_{2m}, z_{3m}, \dots, z_{Nm} | (m, q); u_1, u_2, u_3, \dots, u_N)$$

Dato che $\ln(\cdot)$ è una funzione strettamente monotona, non cambia la posizione dei minimi e massimi di $L(\cdot)$.

Il segno «-» conduce alla minimizzazione invece che alla massimizzazione.

Trasformiamo la produttoria in sommatoria:

$$-\ln(L(.)) = -\ln \prod_{i=1}^N p(z_{im}) = -\sum_{i=1}^N \ln(p(z_{im}))$$

$\ln(abc) = \ln(a) + \ln(b) + \ln(c)$



Modello lineare ed errore Gaussiano



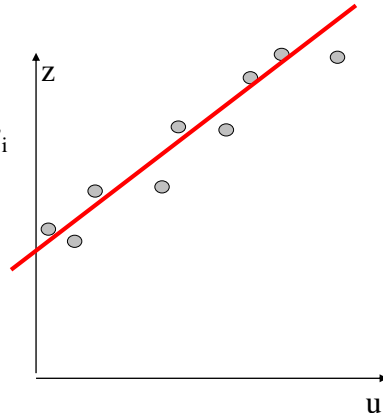
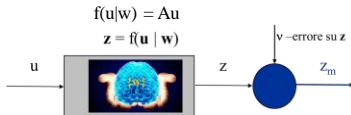
- $z = Au$
- Equazione di una retta: $z = mu + q$ (la matrice A è un vettore che contiene m e q , gli input sono $\{u\}$)
- Scriviamo prima di tutto la densità di probabilità di ottenere z_{im} per ciascun dato, per errore, v_i , Gaussiano a media nulla:

$$p(v) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{z_m - z}{\sigma}\right)^2}$$

$$z_{im} - z_i = v_i \quad \iff \quad z_{im} = (m u + q) + v_i$$

$$p(z_{im} | m, q; u_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{z_{im} - (mu_i + q)}{\sigma}\right)^2}$$

dove m e q non sono i parametri del modello.



A.A. 2023-2024

9/50

<http://borghese.di.unimi.it>



Likelihood



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

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

Scriviamo il **negativo del logaritmo** della verosimiglianza:

$$-\ln(L(\cdot)) = -\ln \left[\prod_{i=1}^N p(z_{im}) \right] = f(z_{1m}, z_{2m}, z_{3m}, \dots, z_{Nm} | (m, q); u_1, u_2, u_3, \dots, u_N)$$

Trasformiamo la produttoria in sommatoria:

$$-\ln(L(\cdot)) = -\ln \left[\prod_{i=1}^N p(z_{im}) \right] = -\sum_{i=1}^N \ln(p(z_{im}))$$

Specifichiamo $p(\cdot)$ per un **modello lineare ed errore Gaussiano a media nulla**:

$$\begin{aligned} -\ln(L(\cdot)) &= -\ln \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{z_{1m} - (mu_1 + q)}{\sigma}\right)^2} \right) - \ln \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{z_{1m} - (mu_1 + q)}{\sigma}\right)^2} \right) + \dots = \\ &= -\sum_i \ln \left(\left(\frac{1}{\sqrt{2\pi}\sigma} \right) e^{-\frac{1}{2}\left(\frac{z_{im} - (mu_i + q)}{\sigma}\right)^2} \right) \end{aligned}$$

A.A. 2023-2024

10/50

<http://borghese.di.unimi.it>



Stima a massima verosimiglianza



Occorre massimizzare $L(\cdot)$ rispetto ai parametri w , (m, q) nel caso della retta.

Abbiamo considerato il negativo del logaritmo:

$$-\ln(L(\cdot)) = -\ln \prod_{i=1}^N p(z_{im}) = f(z_{1m}, z_{2m}, \dots, z_{Nm}; m, q; u_1, u_2, \dots, u_N) := -\sum_i \ln \left(\left(\frac{1}{\sqrt{2\pi}\sigma} \right) e^{-\frac{1}{2} \left(\frac{z_{im} - (mu_i + q)}{\sigma} \right)^2} \right)$$

Dato che consideriamo il negativo, occorre minimizzare il negativo della verosimiglianza, $L(\cdot)$ rispetto a (m, q) .

Dato che $\ln(\cdot)$ è una funzione strettamente monotona, non cambia la posizione dei minimi e massimi di $L(\cdot)$.

Per calcolare il minimo di $-\ln(L(\cdot))$ occorre calcolare il punto di stazionarietà \Rightarrow porre le derivate parziali = 0.



Stima a massima verosimiglianza di m - I



$$-\ln(L(\cdot)) = -\sum_i \ln \left(\left(\frac{1}{\sqrt{2\pi}\sigma} \right) e^{-\frac{1}{2} \left(\frac{z_{im} - (mu_i + q)}{\sigma} \right)^2} \right)$$

□ Poniamo a zero la derivata prima rispetto a m :

$$\frac{\partial f(z_{1m}, z_{2m}, z_{3m}, \dots, z_{Nm} | (m, q); u_1, u_2, u_3, \dots, u_N)}{\partial m} = 0$$

$$\frac{\partial f(z_{1m}, z_{2m}, z_{3m}, \dots, z_{Nm} | (m, q); u_1, u_2, u_3, \dots, u_N)}{\partial m} = \frac{\partial \left(-\sum_{i=1}^N \ln \left(\left(\frac{1}{\sqrt{2\pi}\sigma} \right) e^{-\frac{1}{2} \left(\frac{z_{im} - (mu_i + q)}{\sigma} \right)^2} \right) \right)}{\partial m} =$$

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

$$\ln(ab) = \ln(a) + \ln(b)$$



Stima a massima verosimiglianza di m - II



$$\frac{\partial f(z_{1m}, z_{2m}, z_{3m}, \dots, z_{Nm} | (m, q); u_1, u_2, u_3, \dots, u_N)}{\partial m} = \frac{\partial \left(-\sum_{i=1}^N \ln \left(\frac{1}{\sqrt{2\pi}\sigma} \right) \right)}{\partial m} + \frac{\partial \left(+\sum_{i=1}^N \frac{1}{2} \left(\frac{z_{im} - (mu_i + q)}{\sigma} \right)^2 \right)}{\partial m} = 0$$

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

$$\frac{\partial \left(-\sum_{i=1}^N \ln \left(\frac{1}{\sqrt{2\pi}\sigma} \right) \right)}{\partial m} + \frac{1}{2\sigma^2} \sum_{i=1}^N \left(\frac{\partial \left((z_{im} - (mu_i + q))^2 \right)}{\partial m} \right) = 0$$

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



Stima a massima verosimiglianza di m - III



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

Per tutte le coppie di $\{u_i, z_{im}\}$:

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

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

1ª equazione



Stima a massima verosimiglianza di q - I

$$-\ln(L(\cdot)) = -\ln \prod_{i=1}^N p(z_{im}) = f(z_{1m}, z_{2m}, z_{3m}, \dots, z_{Nm} | (m, q); u_1, u_2, u_3, \dots, u_N)$$

- minimizziamo $-\ln(L(\cdot))$ ponendo a zero la derivata prime rispetto a m :

$$\frac{\partial f(z_{1m}, z_{2m}, z_{3m}, \dots, z_{Nm} | (m, q); u_1, u_2, u_3, \dots, u_N)}{\partial q} = 0$$

$$\frac{\partial f(z_{1m}, z_{2m}, z_{3m}, \dots, z_{Nm} | (m, q); u_1, u_2, u_3, \dots, u_N)}{\partial q} = \frac{\partial \left(-\sum_{i=1}^N \ln \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \left(\frac{z_{im} - (mu_i + q)}{\sigma} \right)^2} \right) \right)}{\partial q} =$$

$$\ln(ab) = \ln(a) + \ln(b)$$

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



Stima a massima verosimiglianza di q - II

$$\frac{\partial f(z_{1m}, z_{2m}, z_{3m}, \dots, z_{Nm} | (m, q); u_1, u_2, u_3, \dots, u_N)}{\partial q} = \frac{\partial \left(-\sum_{i=1}^N \ln \left(\frac{1}{\sqrt{2\pi}\sigma} \right) \right)}{\partial q} + \frac{\partial \left(+\sum_{i=1}^N \frac{1}{2} \left(\frac{z_{im} - (mu_i + q)}{\sigma} \right)^2 \right)}{\partial q} = 0$$

$$\frac{\partial \left(-\sum_{i=1}^N \ln \left(\frac{1}{\sqrt{2\pi}\sigma} \right) \right)}{\partial q} + \frac{1}{2\sigma^2} \frac{\partial \left(\sum_{i=1}^N (z_{im} - (mu_i + q))^2 \right)}{\partial q} = \frac{\partial \left(-\sum_{i=1}^N \ln \left(\frac{1}{\sqrt{2\pi}\sigma} \right) \right)}{\partial q} + \frac{1}{2\sigma^2} \sum_{i=1}^N \left(\frac{\partial \left((z_{im} - (mu_i + q))^2 \right)}{\partial q} \right) = 0$$

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



Stima a massima verosimiglianza di q - III



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

Per tutte le coppie di $\{u_i, z_{im}\}$:

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

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

2^a equazione



Valori m e q che massimizzano la verosimiglianza



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

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

Per ogni punto 2 equazioni lineari in 2 incognite

Per tanti punti: $Ax = b$

$$A = \begin{bmatrix} \sum_{i=1}^N (u_i^2) & [\sum_{i=1}^N (u_i)] \\ [\sum_{i=1}^N (u_i)] & N \end{bmatrix}$$

$$x = \begin{bmatrix} m \\ q \end{bmatrix}$$

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



Problema lineare



Torniamo al nostro problema lineare. Misuro N punti $\{z_{im}\}$, e ho una retta come modello:
 $\{z_{im} = z_i + v = m u_i + q\}$

Scrivo in forma matriciale le N equazioni che legano le misure $\{z_{im}\}$ alle ascisse $\{u_i\}$

$$A = \begin{bmatrix} u_1 & 1 \\ \dots & \dots \\ u_N & 1 \end{bmatrix} \begin{bmatrix} m \\ q \end{bmatrix} \quad \mathbf{b} = \begin{matrix} z_{im} \\ \begin{bmatrix} z_1 \\ \dots \\ z_N \end{bmatrix} \end{matrix} + \begin{matrix} v_1 \\ \dots \\ v_N \end{matrix}$$

Matrice A : $N \times 2$

$\mathbf{Ax} = \mathbf{b} + \text{Errore}$



Soluzione del sistema lineare



$$A^T A = \begin{bmatrix} u_1 & \dots & u_N \\ 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} u_1 & 1 \\ \dots & \dots \\ u_N & 1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^N (u_i^2) & \sum_{i=1}^N (u_i) \\ \sum_{i=1}^N (u_i) & N \end{bmatrix}$$

$$A^T \mathbf{b} = \begin{bmatrix} u_1 & \dots & u_N \\ 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} z_{1m} \\ \dots \\ z_{Nm} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^N (z_{im} u_i) \\ \sum_{i=1}^N (z_{im}) \end{bmatrix}$$

Equazioni normali: $A^T A \mathbf{x} = A^T \mathbf{b}$

sono le stesse ottenute dalla massima verosimiglianza!!



Soluzione come problema di ottimizzazione



Funzione costo: $(Ax - b)^2 = \sum_k v_k^2 = \|Ax - b\|^2$

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.

$$\min_x \sum_k v_k^2 = \min_x (Ax - b)^2$$

$$A^T A x = A^T b$$

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

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

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 stima a massima verosimiglianza - I



$$z = 2u + 1 \quad m = 2; q = 1 \text{ non noti}$$

Misuro con errore = 0 e ottengo:

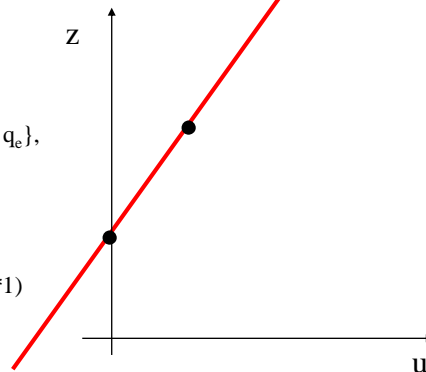
$$\begin{array}{ll} u_1 = 1; z_1 = 3 & z_{1m} = 3 \\ u_2 = 0; z_2 = 1 & z_{1m} = 1 \end{array}$$

Quanto varrà la stima a massima verosimiglianza. $\{m_e, q_e\}$, di m e q ?

$$\begin{aligned} \left[\sum_{i=1}^N (u_i^2) \right] \cdot m + \left[\sum_{i=1}^N (u_i) \right] \cdot q &= \left[\sum_{i=1}^N (z_i \cdot u_i) \right] \\ \Rightarrow (1 \cdot 1 + 0 \cdot 0) m_e + (1 + 0) q_e &= (1 \cdot 3 + 0 \cdot 1) \end{aligned}$$

$$\begin{aligned} \left[\sum_{i=1}^N (u_i) \right] \cdot m + \left[\sum_{i=1}^N (1) \right] \cdot q &= \left[\sum_{i=1}^N (z_i) \right] \\ \Rightarrow (1 + 0) m_e + 2 q_e &= (3 + 1) \end{aligned}$$

$$m_e + q_e = 3 \quad m_e + 2q_e = 4 \quad \Rightarrow \text{per sottrazione} \quad q_e = 1; m_e = 2 \quad \text{retta ideale}$$





Esempio stima a massima verosimiglianza - II



$$z = 2u + 1$$

$m = 2; q = 1$ non noti

Misuro con errore e ottengo:

$$u_1 = 1; z_1 = 3$$

$$z_{1m} = 2.8$$

$$u_2 = 0; z_2 = 1$$

$$z_{2m} = 1.2$$

Quanto varrà la stima a massima verosimiglianza. $\{m_e, q_e\}$, di m e q ?

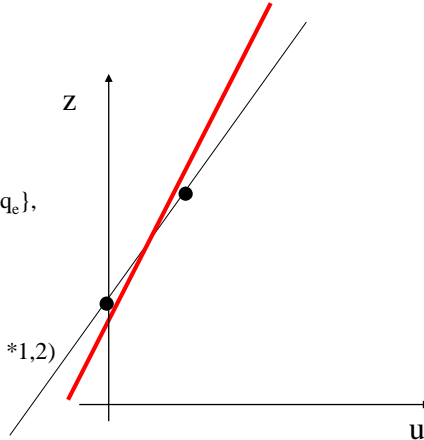
$$\left[\sum_{i=1}^N (u_i^2) \right] \cdot m + \left[\sum_{i=1}^N (u_i) \right] \cdot q = \left[\sum_{i=1}^N (z_i \cdot u_i) \right]$$

$$\Rightarrow (1 \cdot 1 + 0 \cdot 0) m_e + (1 + 0) q_e = (1 \cdot 2.8 + 0 \cdot 1.2)$$

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

$$\Rightarrow (1 + 0) m_e + 2 q_e = (2.8 + 1.2)$$

$$m_e + q_e = 2.8 \quad m_e + 2q_e = 4 \quad \Rightarrow \text{per sottrazione} \quad q_e = 1.2; m_e = 1.6 \quad \text{diversi dai valori reali}$$



Esempio - Caso 2D (2 parametri)



$N = 20$ punti

m reale = 1

q reale = 2

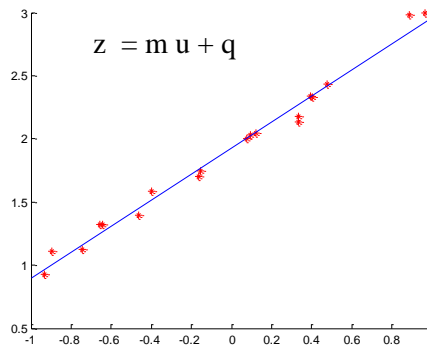
Errore non polarizzato:

$\mu = 0$

$\sigma_o^2 = 0.01$

m stimato = 0.9931

q stimato = 1.0106



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



Stima a massima verosimiglianza e minimi quadrati



$$A^T A x = A^T b$$

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

La soluzione a massima verosimiglianza, quando l'errore è **Gaussiano a media nulla**, e le **misure sono indipendenti**, 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 quando si massimizza $\log(L(\cdot))$.



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



Stima alla massima verosimiglianza

Varianza della stima nei sistemi lineari



Valutazione della bontà della stima



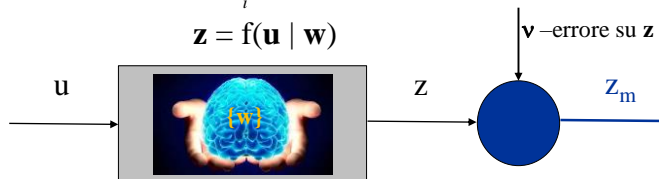
$z = f(u | w)$ misura $\{z_i\}$ in corrispondenza di $\{u_i\}$. $\{z_i\}$ è ottenuto come uscita del modello, tramite i parametri $\{w_j\}$

Z è misurato con un errore (statistico). Avrò che: $z_{i,m} = z_i + v_i$

Se le **misure sono indipendenti** posso scrivere che la probabilità di ottenere le misure: $z_{1m}, z_{2m}, z_{3m}, \dots$ È la probabilità congiunta:

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

$$z = f(u | w)$$



Quanto sono buoni i valori di w calcolati?



Valutazione della bontà della stima



$$x = (A^T A)^{-1} A^T b \iff \min_x \sum_k v_k^2 = \min_x (Ax - b)^2$$

Errore di misura, v , Gaussiano a media nulla $N(0, \sigma_0^2)$

$$\langle v_k \rangle = 0$$

e con una certa varianza nota: σ_0^2 .

La varianza (campionaria) dell'errore di misura \rightarrow Varianza dell'errore per $N \rightarrow \infty$

$$\lim_{N \rightarrow \infty} \frac{\sum_{i=1}^N v_i^2}{N} = \sigma_0^2$$



Valutazione della bontà della stima del singolo parametro, x



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

$$x = CA^T b$$

Chiamiamo varianza dell'errore di misura Δz e Δv le variabili casuali associate rispettivamente all'errore sulle incognite e all'errore di misura, rispettivamente.

Si suppone errore a media nulla e Gaussianamente distribuito.

$$(x + \Delta x) = C A^T (b + \Delta v)$$



$$x = CA^T b \quad \Delta x = C A^T \Delta v$$

$$E[\Delta v] = E[\Delta x] = 0$$

C è la matrice di covarianza



Impostazione del calcolo della correlazione tra i parametri



$$\Delta x = C A' \Delta v$$

Abbiamo M parametri

Vogliamo individuare la correlazione tra due parametri i e j. Devo quindi determinare la loro covarianza:

$$\langle \Delta x_i, \Delta x_j \rangle$$

e la loro variazione (varianza):

$$\langle \Delta x_i, \Delta x_i \rangle = \langle \Delta x_i^2 \rangle$$

$$\text{Covarianza di } x: \langle \Delta x_i, \Delta x_j \rangle = \begin{bmatrix} \Delta x_1^2 & \Delta x_1 \Delta x_2 & \dots & \Delta x_1 \Delta x_M \\ \Delta x_2 \Delta x_1 & \Delta x_2^2 & \dots & \Delta x_2 \Delta x_M \\ & & \dots & \\ \Delta x_M \Delta x_1 & \Delta x_2 \Delta x_M & & \Delta x_M^2 \end{bmatrix}$$



Matrice di covarianza



Date N variabili casuali: $x = [x_1, x_2, \dots, x_N]$ si può misurare la correlazione tra coppie di variabili. E' comodo rappresentare la correlazione tra variabili casuali in un'unica matrice detta **matrice di covarianza** come:

$$C = \begin{bmatrix} \sigma_{x_1 x_1} & \sigma_{x_1 x_2} & \dots & \sigma_{x_1 x_N} \\ \sigma_{x_2 x_1} & \sigma_{x_2 x_2} & \dots & \sigma_{x_2 x_N} \\ \cdot & \cdot & \cdot & \cdot \\ \sigma_{x_N x_1} & \sigma_{x_N x_2} & \dots & \sigma_{x_N x_N} \end{bmatrix}$$

$$\text{Varianza: } \sigma_{x_i x_i} = \sigma^2_{x_i}$$

N parametri

$$\text{Covarianza: } \sigma_{x_i x_j} = \sigma_{x_j x_i} \quad i \neq j$$

$(N-1)^2/2$ parametri



Impostazione del calcolo della correlazione tra i parametri



$\Delta x = C A' \Delta v$
Abbiamo M parametri

$$\begin{bmatrix} \Delta x_1^2 & \Delta x_1 \Delta x_2 & \dots & \Delta x_1 \Delta x_M \\ \Delta x_2 \Delta x_1 & \Delta x_2^2 & \dots & \Delta x_2 \Delta x_M \\ & & \dots & \\ \Delta x_M \Delta x_1 & \Delta x_2 \Delta x_M & & \Delta x_M^2 \end{bmatrix}$$

$$\Delta x = C A' \Delta v \quad \Rightarrow \quad \Delta x' = \Delta v' A (C)' \quad \Rightarrow \quad \Delta x \Delta x' = C A' \Delta v \Delta v' A C'$$

Applicando l'operatore di media, si ottiene:

$$\langle \Delta x \Delta x' \rangle = C A' \langle \Delta v \Delta v' \rangle A C'$$

Dato che v sono i residui, e sono indipendenti, e tutte i punti di controllo hanno lo stesso tipo di errore di misura, si avrà che $\langle \Delta v \Delta v' \rangle = I \sigma_0^2$.



Incertezza sulla stima dei parametri



$$\langle \Delta x \Delta x' \rangle = C A' I A C' \sigma_0^2 = C' \sigma_0^2$$

$$\langle \Delta x' \Delta x \rangle = C \sigma_0^2$$

Segue che:

$$\sigma^2(\Delta x_i) = c_{ii} \sigma_0^2.$$

Varianza sulla stima del parametro, x_i .

$$\sigma^2(\Delta x_{ij}) = c_{ij} \sigma_0^2.$$

Covarianza sulla stima dei parametri x_i, x_j

Incetzza su z -> incetzza sui parametri stimati, x



Esempio: 1 parametro

$$\sigma^2(\Delta x_{ij}) = c_{ij} \sigma_0^2$$

$$\langle \Delta x^T \Delta x \rangle = C \sigma_0^2$$

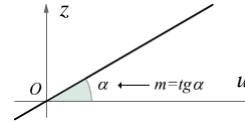
Considero la retta per l'origine, e come incognita x , m (problema di identificazione)

$u \ m = z + \text{errore}$

$A \ x = b + \text{errore}$

Determino la pendenza m della retta

Calcolo m ai minimi quadrati.



Quanto è sensibile questa stima? Cosa succede se, per effetto dell'errore quando invece di misurare z , misuro $z + v$?

$$C = (A^T A)^{-1} \Rightarrow A_{|x|} = u \Rightarrow C = (u^T u)^{-1}$$

La varianza di m varierà in modo inversamente proporzionale a u^2 . L'errore viene cioè moltiplicato per $1/u^2$.

$$\sigma^2(m) = c_m \sigma_0^2 = \sigma_0^2 / u^2$$

Tanto più il punto con cui stimiamo m si trova lontano dall'origine, tanto minore è la varianza della stima di m .



Visione geometrica (1 parametro, m)

Considero la retta per l'origine:

$u \ m = z + \text{errore}$

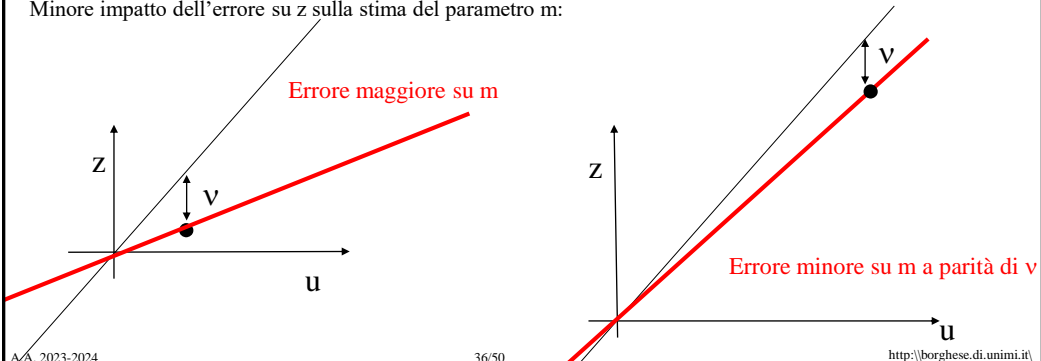
$A \ x = b + \text{errore}$

Determino la pendenza m della retta e calcolo l'affidabilità della stima:

$$\sigma^2(m) = c_m \sigma_0^2 = \sigma_0^2 / u^2$$

Tanto più prendo i punti lontani dall'origine ($|u|$ grande), tanto meglio riesco a stimare m .

Minore impatto dell'errore su z sulla stima del parametro m :





La covarianza: momenti di 2 variabili statistiche



$$\text{Covarianza} = E[(x - \mu_x)(y - \mu_y)]$$

$$\text{Varianza} = E[(x - \mu_x)(x - \mu_x)]$$

Per due variabili indipendenti, la covarianza = 0, non variano assieme (covariano)

$$C = \begin{bmatrix} \sigma_x^2 & \sigma_x \sigma_y \\ \sigma_y \sigma_x & \sigma_y^2 \end{bmatrix}$$

```
>> x = randn(N,1);  
>> y = randn(N,1);  
>> temp = x.*y;  
>> covarianza = mean(temp)
```



Misura di correlazione su 2 parametri



Misura la inter-dipendenza tra 2 variabili statistiche:

$$-1 \leq \frac{\sigma_{xy}}{\sigma_x \sigma_y} = c = \lim_{N \rightarrow \infty} \frac{\sum_k (x_k - \mu_x)(y_k - \mu_y)}{\sqrt{\sum_k (x_k - \mu_x)^2} \sqrt{\sum_k (y_k - \mu_y)^2}} \leq +1$$

```
>> x = randn(N,1);  
>> y1 = randn(N,1);  
>> y2 = x;  
>> temp1 = x.*y1;  
>> temp2 = x.*y2;  
>> covarianza1 = mean(temp1) % Uncorrelated variables (c -> 1)  
>> covarianza2 = mean(temp2) % Correlated variables (c = 0)
```



Sommario



Matrici e Sistemi lineari

Esempio di sistema linearizzato

Soluzione di un sistema lineare

Analisi dell'affidabilità della stima

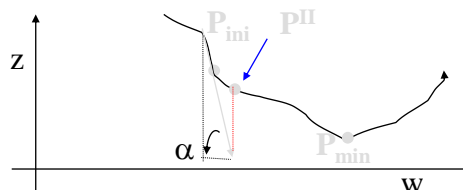
Determinazione dei parametri di un modello non-lineare



Minimizzazione tramite derivata (metodo del primo ordine): 1 variabile



Tecnica della derivata, applicata alla minimizzazione (ricerca del punto di minimo) applicata a funzioni non-lineari $h(\cdot)$ di **una variabile**, u , e di **un parametro**, w : $z = h(u | w)$.



La derivata, mi dà due informazioni:

- 1) In quale direzione di w , la funzione decresce.
- 2) Quanto rapidamente decresce.

E' una descrizione locale dell'andamento della funzione.

Definisco uno spostamento arbitrario lungo la pendenza: maggiore la pendenza maggiore lo spostamento.

$dw \propto -f'(u;w)$ dati u, w . La derivata viene calcolata rispetto a w .

Occorre un'inizializzazione.

Metodo iterativo.



Estensione al caso di più variabili



$z = f(\mathbf{u}; \mathbf{w})$ Occorre sostituire la derivate con il gradiente (derivate direzionali).

Per funzioni di più variabili, $z = f(\mathbf{u}; \mathbf{w})$, la linearizzazione nell'intorno di $\mathbf{P}_0(\mathbf{u}_0, z_0)$, si può scrivere come:

$$f(\mathbf{u}, \mathbf{w}) = f(\mathbf{u}_0, \mathbf{w}) + \sum_i \left. \frac{\partial f(\mathbf{u}_i, \mathbf{w})}{\partial u_i} \right|_{\mathbf{u}_0} du_i$$

E' un'approssimazione del primo ordine.



Esempio di applicazione tecnica del gradiente per funzioni di 1 variabile - I



Supponiamo che il modello da noi considerato sia semplice: $z = a u^2$

Vogliamo determinare a . La funzione è lineare in a (problema di identificazione).

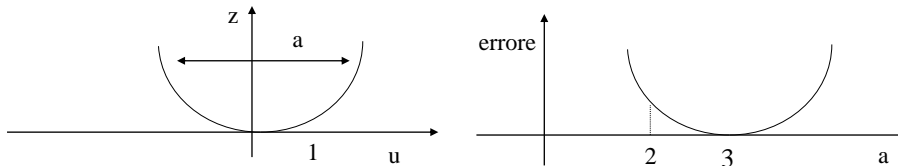
Nel nostro sistema: z è il termine noto, a l'incognita e u il coefficiente noto.

Misuriamo un punto sulla parabola: $u = 1, z = 3$.

Vogliamo modificare il parametro a in modo che la parabola passi per $P(u, z) = P(1, 3)$.

La funzione costo da minimizzare sarà: errore = $h(a, u, z) = (z - a u^2)^2$

La soluzione è $a = 3$





Esempio di applicazione tecnica del gradiente per funzioni di 1 variabile - II



Supponiamo che il modello da noi considerato sia semplice: $z = a u^2$

Vogliamo determinare a . La funzione è lineare in a (problema di identificazione).

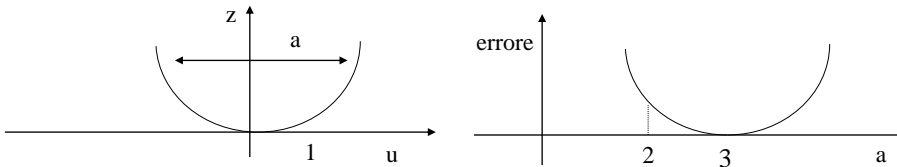
Nel nostro sistema: z è il termine noto, a l'incognita e u il coefficiente noto.

Vogliamo modificare il parametro a in modo che la parabola passi per $P(u,z) = P(1,3)$.

La funzione costo da minimizzare sarà: errore = $h(a, u, z) = (z - a u^2)^2$

Partiamo da $a_{ini} = 2$.

Utilizziamo il metodo del gradiente:



$$\text{errore} = (3 - 2 \cdot 1)^2 = 1$$

$$\text{Calcoliamo la derivata di } h(a, u, z) \text{ rispetto ad } a \rightarrow h'(a, u, z) = -2(z - a u^2) u^2$$



Minimizzazione - underdamping



Misuriamo un punto sulla parabola: $u = 1, z = 3$.

Consideriamo $\alpha = 1$

$$\text{Calcoliamo la derivata di } h(\cdot) \rightarrow h'(\cdot) = -2(z - a u^2) u^2$$

Utilizziamo il metodo del gradiente:

Passo 1:

Calcoliamo l'incremento da dare al parametro a :

$$da = -[-2(3 - 2 \cdot 1) \cdot 1] = -[-6 + 4] = 2 \qquad a^1 = 2 + 2 = 4$$

Passo 2:

Calcoliamo l'incremento da dare al parametro a :

$$da = -[-2(3 - 4 \cdot 1) \cdot 1] = -[-6 + 8] = -2 \qquad a^2 = 4 - 2 = 2$$

Oscillazioni!!!

Mi sposto troppo velocemente da una parte all'altra del minimo.



Minimizzazione -2 passi



Consideriamo $\alpha = 0.4$

Calcoliamo la derivata di $k(\cdot) \rightarrow k'(\cdot) = -2(z - a u^2) u^2$

Utilizziamo il metodo del gradiente:

Passo 1:

Calcoliamo l'incremento da dare al parametro a:

$$da = -0.4 [-2(3 - 2 \cdot 1) \cdot 1] = -[-6 + 4] = 0.8$$

$$a' = 2 + 0.8 = 2.8$$

Passo 2:

Calcoliamo l'incremento da dare al parametro a:

$$da = -0.4 [-2(3 - 2.8 \cdot 1) \cdot 1] = -[-6 + 5.6] = 0.16 \quad a'' = 2.8 + 0.16 = 2.96$$

Converge ad $a = 3$.

Posso correre il rischio di spostarmi troppo lentamente

Scheduling del parametro α



Minimizzazione di funzioni di più variabili



$\min(f(\mathbf{x}, \mathbf{w}))$ funzione costo oppure errore, \mathbf{w} vettore.

Modifico il valore dei pesi di una quantità proporzionale alla pendenza della funzione costo rispetto a quel parametro.

La pendenza è una direzione nello spazio, non è più solamente destra / sinistra.

Devo calcolare la derivata spaziale = **gradiente** della funzione costo, $f(\cdot)$.

Estensione della tecnica del gradiente a più variabili.

$$d\mathbf{w} = -\alpha \nabla f(\mathbf{x}; \mathbf{w}), \text{ dato } \mathbf{P}, \mathbf{W}.$$

Serve un' **approssimazione iniziale** per i pesi $\mathbf{W}_{ini} = \{w_j\}_{ini}$.



Evoluzione dei metodi del primo ordine



- α è un parametro critico. Se è troppo piccolo convergenza molto lenta, se è troppo grande overshooting.
- Ottimizzazione di α . Ad ogni passo viene calcolato α ottimale, per cui la funzione è decrescente (line search).



Metodo di Gauss-Newton



- L'idea:

Inizializzazione:

- Inizializzo i parametri ad un valore iniziale.

Iterazioni:

- 1) Linearizzazione delle equazioni sul valore corrente dei parametri.
- 2) Stima dell'aggiornamento dei parametri nel modello linearizzato ai minimi quadrati (soluzione ottimale, minimo del problema linearizzato).
- 3) Correzione dei parametri.

Può essere pesante perchè richiede l'inversione della matrice di covarianza. Spesso si preferiscono utilizzare metodi di ottimizzazione del primo ordine.



In pratica



$\mathbf{z} = f(\mathbf{u}, \mathbf{w})$ \mathbf{u}, \mathbf{z} vettori di N ed M elementi rispettivamente

$\mathbf{z}_0 = f(\mathbf{u}_0, \mathbf{w})$ $\mathbf{u}_0, \mathbf{z}_0$ valore iniziale

Iterazione di (nella prima iterazione $k = 0$):

□ $\mathbf{dz}_k = (\sum \delta f(\mathbf{u}, \mathbf{w}) / \mathbf{du})_{\mathbf{u}_k} \mathbf{du}$ $(\sum \delta f(\mathbf{u}, \mathbf{w}) / \mathbf{dw})_{\mathbf{u}_k}$ **are numbers!**

Si ottiene un sistema lineare

- Viene risolto come $\mathbf{du} = (\mathbf{A}\mathbf{A}^T)^{-1} \mathbf{A}^T \mathbf{dz}_k$
- **Si aggiorna il valore di \mathbf{u} come $\mathbf{u}_{k+1} = \mathbf{u}_k + \mathbf{du}_k$**
- **Si aggiorna il valore di \mathbf{z} come $f(\mathbf{u}_{k+1}, \mathbf{w})$**

Fino a convergenza



Sommario



Matrici e Sistemi lineari

Esempio di sistema linearizzato

Soluzione di un sistema lineare

Analisi dell'affidabilità della stima

Determinazione dei parametri di un modello non-lineare