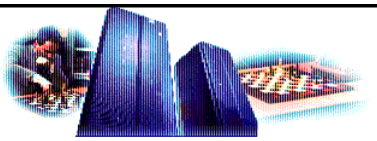
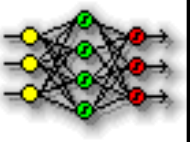




Sistemi Intelligenti
Stimatori e sistemi lineari - III

Alberto Borghese

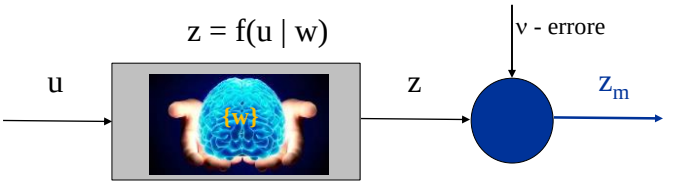
Università degli Studi di Milano
Laboratory of Applied Intelligent Systems (AIS-Lab)
Dipartimento di Informatica
borghese@di.unimi.it

Overview

Stima alla massima verosimiglianza

Varianza della stima nei sistemi lineari

I problemi associate ai Modelli



u – causa => 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} - ***Learning***

Probabilità di un certo insieme di misure

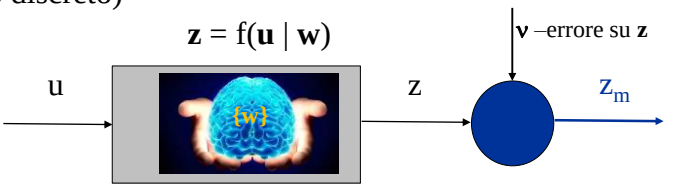
z = f(u | w) misuro $\{z_i\}$ in corrispondenza di $\{u_i\}$.$\{z_i\}$ è ottenuto come uscita del modello, tramite i parametri $\{w_j\}$

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

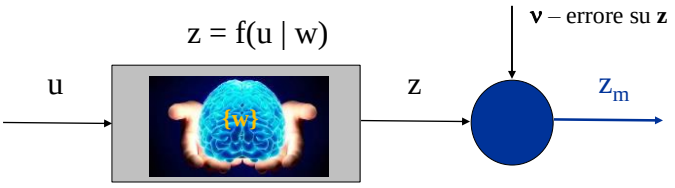
Se le **misure sono indipendenti** posso scrivere che la probabilità di ottenere le misure: z_{1m}, z_{2m}, z_{3m} ….. È la probabilità congiunta:

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

(cf. dadi nel caso discreto)



Dipendenza delle misure



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

$$L = p(z_{1m}, z_{2m}, z_{3m}, \ldots z_{Nm}) = \prod_i p(z_{im} \mid u_i, w) = \prod_i^{z_i} p(f(u_i; w) \mid u_i, w)$$

Scrivo la probabilità esplicitamente come condizionata al valore di u e di w.
La probabilità congiunta è chiamata funzione di Likelihood (verosimiglianza), L(.).

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

Nel caso dei problemi inversi voglio trovare i valori $\{u_i\}$ che massimizzano la probabilità congiunta.

Funzione di verosimiglianza

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

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

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

Stima alla massima verosimiglianza

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

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

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

Forma alternativa della L(.)

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

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

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

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

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

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, 1})
- Scriviamo prima di tutto la densità di probabilità di ottenere z_{im} per ciascun dato, per errore, ν_i, Gaussiano a media nulla:

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

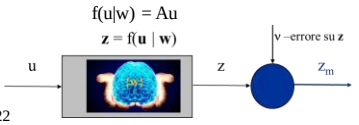
z

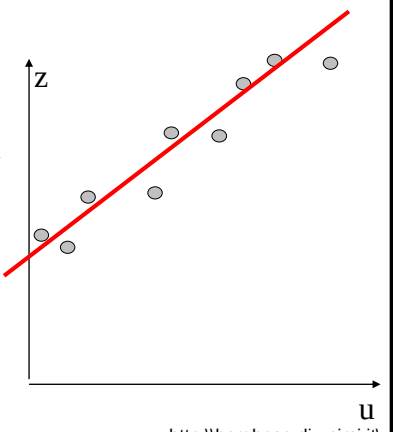
$z_{im} - z_i = \nu_i$ <== $z_{im} = (m\,u + q) + \nu_i$

ν_i

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

dove m e q non sono i parametri del modello.





Likelihood

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

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

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

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

Trasformiamo la produttoria in sommatoria:

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

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

$$\text{-ln(L(.))} = -ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) e^{-\frac{1}{2}\left(\frac{z_{1m} - (mu_1 + q)^2}{\sigma^2}\right)} - ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) e^{-\frac{1}{2}\left(\frac{z_{2m} - (mu_2 + q)^2}{\sigma^2}\right)} + \cdots =$$

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

Stima a massima verosimiglianza

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

Abbiamo considerato il negativo del logaritmo:

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

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

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

Per calcolare il minimo di –ln(L(.)) occorre calcolare le derivate parziali e porle = 0.

Stima a massima verosimiglianza di m

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

- minimizziamo –ln(L(.)) ponendo a zero la derivata prime rispetto a m:

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

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

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

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

Stima a massima verosimiglianza di m

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

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

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

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

Stima a massima verosimiglianza di m

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

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

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

1ª equazione

Stima a massima verosimiglianza di q

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

- minimizziamo –ln(L(.)) ponendo a zero la derivata prime rispetto a m:

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

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

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

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

Stima a massima verosimiglianza di q

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

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

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

Stima a massima verosimiglianza di q

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

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

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

2ª equazione

Valori m e q che massimizzano la verosimiglianza

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

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

2 equazioni **lineari** in **2 incognite**

Ax = b

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

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

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

Problema lineare

Torniamo al nostro problema lineare. Ho N misure sulla retta:
$\{z_{im} = m\, u_i + q + \nu\}$

Scrivo in forma matriciale:

$$A = \begin{bmatrix} u_1 & 1 \\ \cdots & \\ u_N & 1 \end{bmatrix} \begin{bmatrix} m \\ q \end{bmatrix} \qquad b = \overbrace{\begin{bmatrix} b_1 \\ \cdots \\ b_N \end{bmatrix}}^{z_{im}} + \begin{bmatrix} v_1 \\ \cdots \\ v_N \end{bmatrix}$$

Matrice A: N x 2

Ax = b + Errore

Soluzione ai minimi quadrati

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

$$A^Tb = \begin{bmatrix} u_1 & u_N \\ 1 & 1 \end{bmatrix} \begin{bmatrix} z_{1m} \\ \cdots \\ z_{Nm} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N}\left(z_{im}u_i\right) \\ \sum_{i=1}^{N}\left(z_{im}\right) \end{bmatrix}$$

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.

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

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

$$\mathbf{A^TAx = A^Tb}$$

$$\mathbf{x = (A^TA)^{-1}A^Tb}$$

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

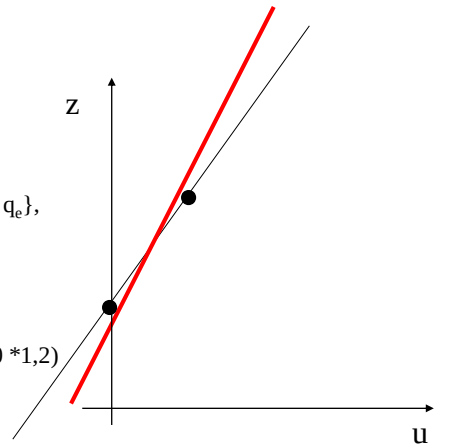
$z = 2\,u + 1$ **m = 2; q = 1 non noti**

Misuro con errore e ottengo:

$u_1 = 1$; $z_1 = 3$ $\quad z_{1m} = 2.8$

$u_2 = 0$; $z_2 = 1$ $\quad z_{1m} = 1.2$

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



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

$\Rightarrow (1*1 + 0*0)\ m_e + (1 + 0)\ q_e = (1 * 2{,}8 + 0 * 1{,}2)$

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

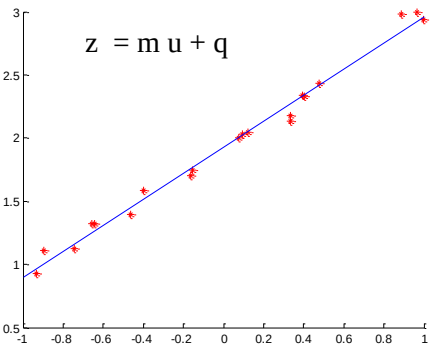
$\Rightarrow (1 + 0)\ m_e + 2\ q_e = (2{,}8 + 1{,}2)$

$m_e + q_e = 2{,}8$ $\quad m_e + 2q_e = 4$ $\quad \Rightarrow$ per sottrazione $\quad \mathbf{q_e = 1{,}2;\ m_e = 1{,}6}$

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 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(.)).

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

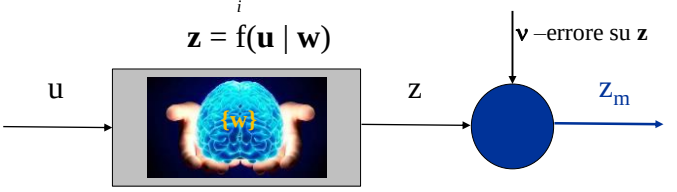
Valutazione della bontà della stima

$z = f(u \mid w)$ misuro $\{z_i\}$ 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: $z_{i,m} = z_i + \nu_i$

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

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



Quanto sono buoni i valori di w calcolati?

Valutazione della bontà della stima

$$x = (A'*A)^{-1}A'*b \quad <==> \quad \min_x \sum_k \nu_k^2 = \min_x (\mathbf{Ax} - \mathbf{b})^2$$

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

$$<\nu_k> = 0$$

e con una certa varianza nota: σ_0^2.

La varianza dell'errore di misura (campionaria) → Varianza dell'errore

$$\lim_{N\to\infty} \frac{\sum_{i=1}^{N} \nu_i^2}{N} = \sigma_0^2$$

Valutazione della bontà della stima del singolo parametro, x

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

$x = CA'^* b$

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.

$$(x + \Delta x) = C\,A'\,(b + \Delta v)$$

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

$$\mathbf{x = CA'b} \qquad \Delta x = C\,A'^*\,\Delta v$$

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 covariazione:

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

e la loro variazione (varianza):

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

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: $x = [x_1, x_2, \ldots x_N]$ 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:

$$C = \begin{bmatrix} \sigma_{x_1x_1} & \sigma_{x_1x_2} & . & \sigma_{x_1x_N} \\ \sigma_{x_2x_1} & \sigma_{x_2x_2} & . & \sigma_{x_2x_N} \\ . & . & . & . \\ \sigma_{x_Nx_1} & \sigma_{x_Nx_2} & . & \sigma_{x_Nx_N} \end{bmatrix}$$

Varianza: $\sigma_{x_1x_1} = \sigma^2{}_{x_1}$ N parametri

Covarianza: $\sigma_{x_ix_j} = \sigma_{x_jx_i}$ $i \neq j$ $(N-1)^2/2$ parametri

Impostazione del calcolo della correlazione tra i parametri

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

Abbiamo M parametri

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

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

Applicando l'operatore di media, si ottiene:

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

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 $<\Delta v \Delta v'> = I\sigma_0^2$.

Incertezza sulla stima dei parametri

$\langle \Delta x\, \Delta x' \rangle = C A'\, I\, A\, C' \sigma_0^2 = C'\, \sigma_0^2$ $\quad$ $\boxed{\langle \Delta x'\, \Delta x \rangle = C\, \sigma_0^2}$

Segue che:

$\sigma^2(\Delta x_i) = c_{ii}\ \sigma_0^2$. Varianza sulla stima del parametro, x_i.

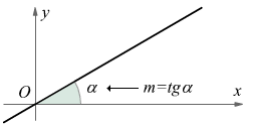
$\sigma^2(\Delta x_{ij}) = c_{ij}\ \sigma_0^2$. Covarianza sulla stima dei parametri x_i, x_j

Incertezza su z -> incertezza sui parametri stimati, x

Esempio: 1 parametro

$\sigma^2(\Delta x_{ij}) = c_{ij}\ \sigma_0^2$ $\quad$ $\boxed{\langle \Delta x'\, \Delta x \rangle = C\, \sigma_0^2}$

Considero la retta per l'origine:
u m = z + errore
A x = b + 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?

$C = (A'*A)^{-1}$ => $A_{1x1} = u$ => $C = (u * u)^{-1}$

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

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

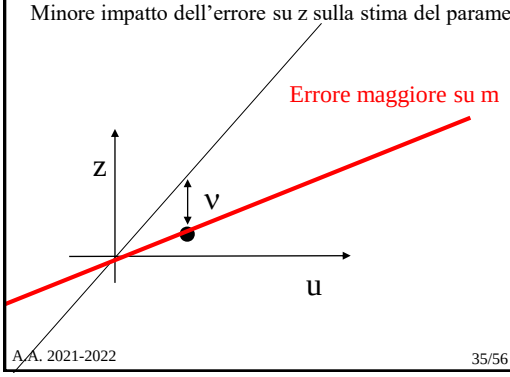
Visione geometrica (1 parametro, m)

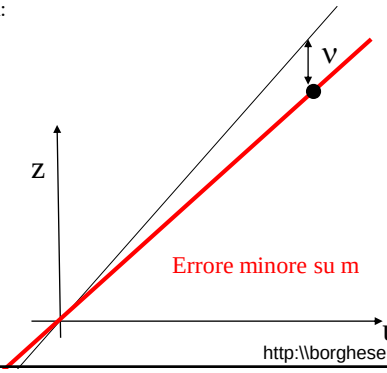
Considero la retta per l'origine:
u m = z + errore
A x = b + errore

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

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

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:

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

```
>> 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)
```


Correlazione tra i parametri

$<\Delta x \Delta x'> = C A' I A C' \sigma_0^2 = C' \sigma_0^2$ $\quad$ $<\Delta x' \Delta x> = C \sigma_0^2$

Da cui si giustifica il nome di matrice di covarianza per C.

Segue che: $\sigma^2(x_{ij}) = c_{ij} \sigma_0^2$ Varianza sulla stima del parametro.

$$-1 \leq r_{ij} = \frac{\langle \Delta x_i \Delta x_j \rangle}{\sqrt{\langle \Delta x_i \rangle \langle \Delta x_j \rangle}} = \frac{c_{ij}}{\sqrt{c_i c_j}} \leq +1$$

Indice di correlazione tra il parametro i ed il parametro j
(empiricamente si scartano parametri quando la correlazione è superiore al 95%)

Vanno rapportati alle dimensioni dei parametri coinvolti.

Caso 2D

N = 20 punti $\sigma_o^2 = 0.1$
m reale = 1 q reale = 2

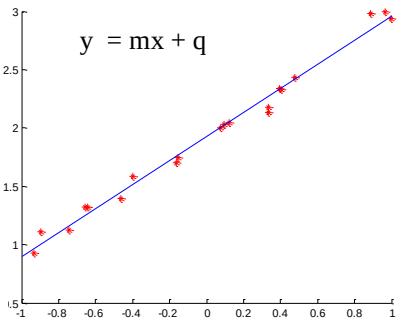
C =
0.0494 0.0087
0.0087 0.0515

m stimato = 0.9931
q stimato = 2.0106

Altra realizzazione del rumore:

C=
0.1702 0.0124
0.0124 0.0509

m stimato = 0.9937
q stimato = 1.9522



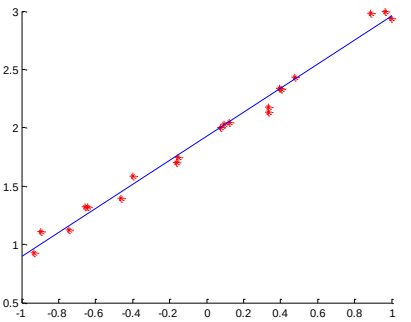
$\sigma^2(m) = c_{11}\, \sigma_0^2 = 0.1702 * 0.1 = 0.017$
$\sigma^2(q) = c_{22}\, \sigma_0^2 = 0.0515 * 0.1 = 0.005$

Caso 2D – less points

N = 10 punti $\sigma_o^2 = 0.1$
m reale = 1 q reale = 2

C =
0.5927 -0.0030
-0.0030 0.1000
m_stimato =
1.0081
q_stimato =
1.9616

C =
0.2514 -0.0360
-0.0360 0.1051
m_stimato =
1.0012
q_stimato =
1.9107



y = mx + q

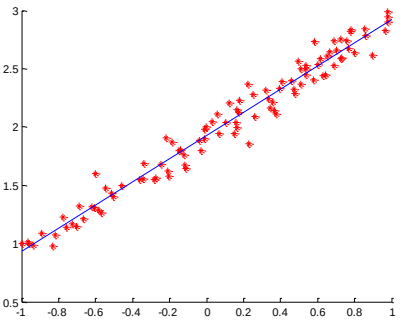
Diminuisce la confidenza nella stima

Caso 2D – more points

N = 100 punti $\sigma_o^2 = 0.1$
m reale = 1 q reale = 2

C =
0.0327 -0.0034
-0.0034 0.0103

m_stimato =
0.9942
q_stimato =
1.9978



y = mx + q

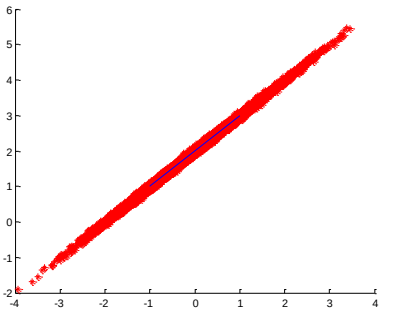
Aumenta la confidenza nella stima

Caso 2D – even more points

N = 10000 punti $\sigma_o^2 = 0.01$
m reale = 1 q reale = 2

C =
0.0000993 0.0000004
0.0000004 0.0001000

m_stimato =
1.000039
q_stimato =
2.0012



y = mx + q

Aumenta la confidenza nella stima

Sommario

Analisi dell'affidabilità della stima

Metodo del gradiente

Linearizzazione e metodo di Gauss-Newton

Probabilità di un certo insieme di misure

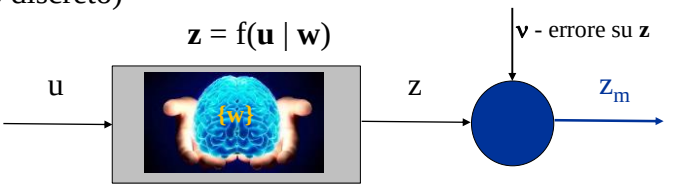
$z = f(u \mid w)$ misuro $\{z_i\}$ in corrispondenza di $\{u_i\}$.$\{z_i\}$ è ottenuto come uscita del modello, tramite i parametri $\{w_j\}$

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

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

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

(cf. dadi nel caso discreto)

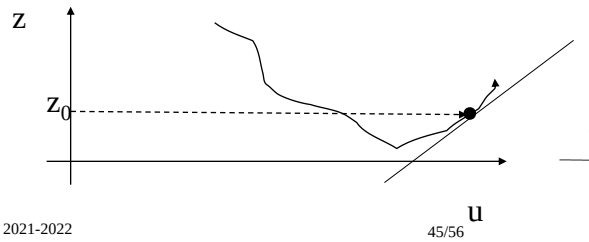


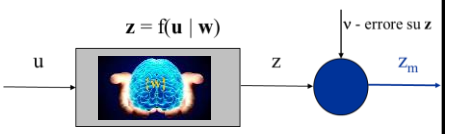
Linearizzazione

z = f(u|w) viene linearizzata utilizzando il differenziale (retta tangente):

$$dz = f(u_0\,|w) + \left.\frac{\partial f(u|w)}{du}\right|_{u_o} \mathrm{du} = z_0 + \left.\frac{\partial f(u|w)}{du}\right|_{u_o} \mathrm{du}$$

Si può vedere come sviluppo di Taylor arrestato al 1° ordine. E' un'equazione lineare.





Linearizzazione nel caso di più variabili

z = f(**u**|w) viene linearizzata utilizzando il differenziale (piano tangente).
Occorre sostituire la derivate con il gradient (derivate direzionale):

Per funzioni di più variabili, z= f(**u;w**), la linearizzazione nell'intorno di $\mathbf{P_0}$ ($\mathbf{u_0}$,z_0), si può scrivere come:

$$f(\mathbf{u}, w) = f(\mathbf{u_0}, w) + \sum_i \left.\frac{\partial f(u_i,w)}{du_i}\right|_{u_o} \mathrm{du_i}$$

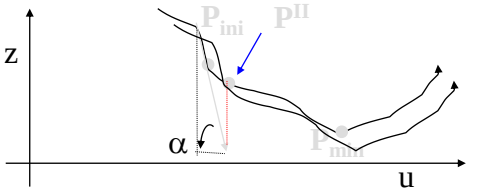
E' un'equazione lineare che descrive il comportamento della funzione f(.) nell'intorno del punto **Po** nello spazio n+1 dimensionale con n dimensionalità di **u**.

$$z = f(\mathbf{u}, w) = z_0 + \sum_i k_i\, \mathrm{du_i}$$

Fissato un valore di z desiderato possiamo scrivere: $(z\text{-}z_0) = J\,du$

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

Tecnica del gradiente applicata alla minimizzazione di funzioni non-lineari di **una variabile, u,** e di **un parametro, w**: z = f (**u** | **w**).



La derivata, mi dà due informazioni:

1) In quale direzione di **u**, la funzione decresce.
2) Quanto rapidamente decresce.

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

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

Occorre un'inizializzazione.

Metodo iterativo.

Esempio di applicazione tecnica del gradiente per funzioni di 1 variabile

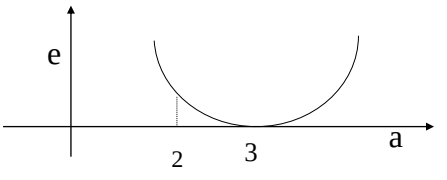
Supponiamo che il modello da noi considerato sia semplice: $z = a\,u^2$

Vogliamo determinare a. La funzione è lineare in a.

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 a in modo che la parabola passi per P(u,z) = P(1,3).
La funzione costo da minimizzare sarà: $e = k(a.\,u,\ z) = (z - au^2)^2$
La soluzione è a = 3

Partiamo da $a_{ini} = 2$.
$e = (3 – 2*1)*1 = 1$



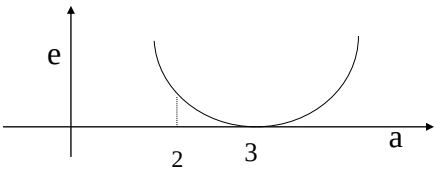
Utilizziamo il metodo del gradiente:
Calcoliamo la derivata di k(a,u, z) rispetto ad a ➔ $k'(a.u.z) = -2\,(z – a\,u^2)\,u^2$

Esempio di applicazione tecnica del gradiente per funzioni di 1 variabile

Supponiamo che il modello da noi considerato sia semplice: $z = a\,u^2$
Vogliamo determinare a. La funzione è lineare in a.
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 a in modo che la parabola passi per P(u,z) = P(1,3).
La funzione costo da minimizzare sarà: $e = k(a.\,u,\ z) = (z - au^2)^2$
La soluzione è a = 3

Partiamo da $a_{ini} = 2$.
$e = (3 – 2*1)*1 = 1$



Utilizziamo il metodo del gradiente:
Calcoliamo la derivata di k(a,u, z) rispetto ad a ➔ $k'(a.u.z) = -2\,(z – a\,u^2)\,u^2$

Minimizzazione – underdamping

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

Consideriamo $\alpha = 1$
Calcoliamo la derivata di k(.) ➔ $k'(.) = -2\,(z – a\,u^2)\,u^2$

Utilizziamo il metodo del gradiente:

Passo 1:
Calcoliamo l'incremento da dare al parametro a:
da =-[-2 (3 – 2 1) 1] = -[-6 +4] = 2 a' = 2 + 2 = 4

Passo 2:
Calcoliamo l'incremento da dare al parametro a:
da = -[-2 (3 - 4 1) 1] = -[-6 +8] = -2 a'' = 4 – 2 = 2
Oscillazioni!!!

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

Minimizzazione -2 passi

Consideriamo α = **0.4**
Calcoliamo la derivata di k(.) ➔ k'(.) = -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 1) 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 1) 1] = -[-6 + 5.6] = 0.16 a'' = 2.8 + 0.16 = 2.96
Converge ad a = 3.

Posso correre il rischio di spostarmi troppo lentamente

Minimizzazione di funzioni di più variabili

min(f(**x**, **w**)) funzione costo od errore, **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(.).
Estensione della tecnica del gradiente a più variabili.

$dw = - \alpha \nabla f(x;w)$, dato P, W.

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

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 a 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

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

$\mathbf{z_0} = f(\mathbf{u_0},\mathbf{w})$ u_0, z_0 valore iniziale

Iterazione di (nella prima iterazione k = 0):

- $\mathbf{dz_k} = (\Sigma \delta f(\mathbf{u},\mathbf{w})/d\mathbf{u})_{\mathbf{uk}}\ \mathbf{du}$ $(\Sigma \delta f(\mathbf{u},\mathbf{w})/d\mathbf{w})_{\mathbf{uk}}$ **are numbers!**
  Si ottiene un sistema lineare
- Viene risolto come $\mathbf{du} = (AA^T)^{-1}A^T\mathbf{dz_k}$
- **Si aggiorna il valore di u come $\mathbf{u_{k+1}} = \mathbf{u_k} + \mathbf{du_k}$**
- **Si aggiorna il valore di z come $\mathbf{f(u_{k+1},w)}$**

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