

# Sistemi Intelligenti Stimatori e identificazione - II

Alberto Borghese

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



## Overview



### Modelli

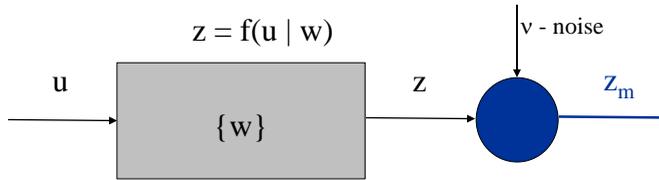
Sistemi lineari

Densità di probabilità

Massima versosimiglianza



# Modello (predittivo)



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

## Regresione predittiva

### LA CORSA DEL LISTINO CINESE

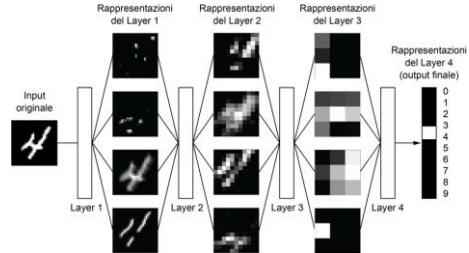
Andamento dell'indice delle A-share della Borsa di Shanghai



A.A. 2020-2021

3/55

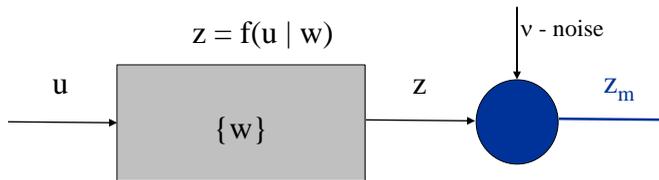
## Classificazione



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



# Modello (predittivo)



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



Realizzazione del modello

utilizzo

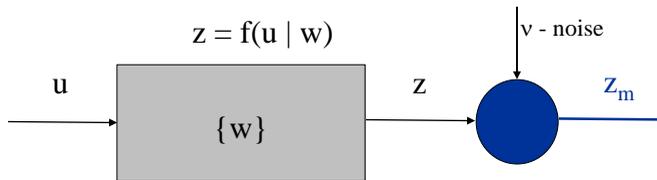
A.A. 2020-2021

4/55

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



## I problemi associate 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  $\{u\}, \{z_m\}$  - Learning**



## Overview

Modelli

**Sistemi lineari**

Densità di probabilità

Massima versosimiglianza



# Sistema lineare



$$a_{11}x_1 + a_{12}x_2 + \dots + a_{1N}x_N = b_1$$

$$a_{21}x_1 + a_{22}x_2 + \dots + a_{2N}x_N = b_2$$

$\{a_{ij}\}$  – coefficienti in numero  $N \times M$   
 $\{x_j\}$  – incognite,  $N$   
 $\{b_j\}$  – termini noti,  $M$

$$a_{M1}x_1 + a_{M2}x_2 + \dots + a_{MN}x_N = b_M$$

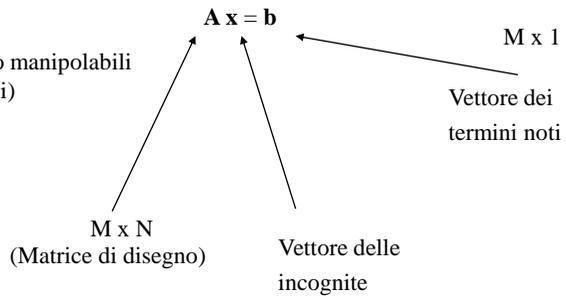
I sistemi lineari sono interessanti perchè sono manipolabili con operazioni semplici (algebra delle matrici)

### Esempio:

$$3x_1 + 2x_2 + \dots + 4x_N = 5$$

$$4x_1 - 2x_2 + \dots + 0.5x_N = 3$$

$$2x_1 + 3x_2 + \dots - 3x_N = -1$$



# Sistema lineare e modelli



$$a_{11}x_1 + a_{12}x_2 + \dots + a_{1N}x_N = b_1$$

$$a_{21}x_1 + a_{22}x_2 + \dots + a_{2N}x_N = b_2$$

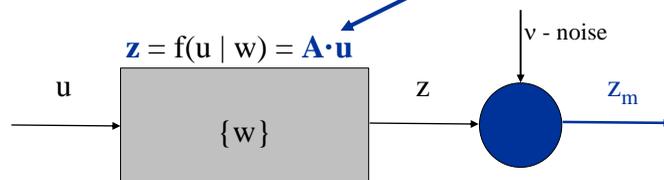
$\{a_{ij}\}$  – coefficienti in numero  $N \times M$   
 $\{x_j\}$  – incognite,  $N$   
 $\{b_j\}$  – termini noti,  $M$

$$a_{M1}x_1 + a_{M2}x_2 + \dots + a_{MN}x_N = b_M$$

### Inverse problem. Determine $\{u\}$ from $\{w\}, \{z\}$ .

- I termini noti  $b_i$  sono i valori misurati in uscita,  $\{z_m\}$ .
- Le incognite  $x$  sono gli ingressi  $\{u\}$  che hanno protetto le  $\{z_m\}$ .
- I coefficienti  $a_{ij}$  sono i parametri del modello,  $\{w\}$

Modello lineare





## Matrici



Insieme di valori organizzati per righe e colonne

$$A = [a_{i,j}]$$

$$A^T = [a_{j,i}]$$

$$\alpha A = [\alpha a_{i,j}]$$

$$C = A + B = [a_{i,j} + b_{i,j}]$$

$$C = AB = [c_{i,j}] \quad \alpha = \text{cost} \quad \text{dove } [c_{i,j}] = \sum_{k=1}^n a_{i,k} b_{k,j}$$

Prodotto degli elementi di una riga per gli elementi di una colonna.

$$\text{Se } A (n \times m) \rightarrow B (m \times p) \rightarrow C (n \times p)$$

La somma è associativa e commutativa  $(A + B) + C = A + (B + C)$ .

Il prodotto è associativo rispetto alla somma ma non gode della proprietà commutativa:

$$(A+B)C = AC + BC.$$

$$AB \neq BA$$



## Altre proprietà delle matrici



Una matrice  $W = \{w_{ij}\}$  si dice diagonale se  $w_{ij} = \begin{cases} w_{ii} & \text{per } i = j \\ 0 & \text{altrimenti} \end{cases}$

Matrice identità.  $I = \{a_{ij}\} : \begin{cases} 1 & \text{per } i = j \\ 0 & \text{altrimenti} \end{cases} \quad A \cdot I = I \cdot A = A$

$$(A B C)^T = C^T B^T A^T$$

Una matrice  $U$ , si dice ortonormale se  $U^T U = I \rightarrow U^{-1} = U^T$

### Condizione di ortonormalità:

Il determinante è  $\pm 1$ .

La somma dei prodotti di due righe o di due colonne è  $0$ .

La somma dei quadrati degli elementi su righe e colonne  $= 1$

Esempio notevole: **matrice di rotazione (cambio di sistema di riferimento)**.



## Rango di una matrice

Data una matrice  $A$  di ordine  $n$  ( $n \times n$ ),

una matrice  $A$   $n \times n$  ha rango  $m < n$  se e solo se esiste un suo minore di ordine  $m$  non nullo (determinante  $\neq 0$ ) mentre sono nulli tutti i minori di ordine  $m + 1$ .

Una matrice  $A$   $n \times n$  ha rango  $n$  (rango pieno) se e solo se il suo determinante è diverso da 0

**Rango** di una matrice  $M \times N$  è la dimensione massima di tutte le matrici quadrate estraibili da  $A$  e con determinante non nullo. Il rango è massimo quando non è inferiore alla dimensione minima della matrice.



## Matrice inversa

Viene definita per matrici **quadrate** ( $N \times N$ ):

$$A^{-1}A = I$$

Esiste ed è unica se  $\det(A) \neq 0$ .

*Somma dei prodotti degli elementi di una riga o colonna per il loro complemento algebrico (formula di Leibniz).*

$$Ax = b \rightarrow A^{-1}Ax = A^{-1}b \rightarrow Ix = A^{-1}b \rightarrow \boxed{x = A^{-1}b}$$



## Sistema lineare: soluzione robusta (SVD)

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

$$\mathbf{A} = \mathbf{U}^T \mathbf{W} \mathbf{V}$$

Ortonormale  $M \times N$

Diagonale ( $N \times N$ )

Ortonormale  $N \times N$

Se  $N = M$

$$\mathbf{x} = \mathbf{V}^T \mathbf{W}^{-1} \mathbf{U}^T \mathbf{b}$$

$$\mathbf{A}^{-1} = (\mathbf{U}^T \mathbf{W} \mathbf{V})^{-1} = \mathbf{V}^T \mathbf{W}^{-1} \mathbf{U}$$

$\mathbf{W}^{-1}$  è diagonale.  $w_{ii}^{-1} = 1/w_{ii}$   
 $w_{ii}$  sono detti valori singolari.

$$\mathbf{A} \mathbf{x} = \mathbf{b} \rightarrow \mathbf{x} = \mathbf{A}^{-1} \mathbf{b} = \mathbf{V}^T \mathbf{W}^{-1} \mathbf{U}^T \mathbf{b}$$



## Condizionamento di una matrice

La matrice inversa esiste ed è unica se  $\det(\mathbf{A}) \neq 0$ .

$$\mathbf{A}^{-1} = (\mathbf{U}^T \mathbf{W} \mathbf{V})^{-1} = \mathbf{V}^T \mathbf{W}^{-1} \mathbf{U}$$

$$\det(\mathbf{A}) = \det(\mathbf{U}^T) \det(\mathbf{W}) \det(\mathbf{V}) = 1 \cdot \left(\prod_{i=1}^N w_{ii}\right) \cdot 1$$

**Numero di condizionamento di una matrice:** rapporto tra il valore singolare maggiore e minore ( $w_{11} / w_{nn}$ ) - cf. Funzione cond in Matlab).

È una misura di **sensibilità della soluzione** di un sistema lineare a variazioni nei dati.

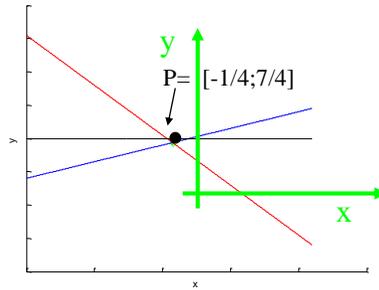


## Esempio di soluzione di un sistema lineare



$$y = x + 2$$

$$y = -3x + 1$$



$$1 x_1 - 1 x_2 = -2$$

$$-3 x_1 - 1 x_2 = -1$$

$$y = x_2$$

$$x = x_1$$

Risolvero per sostituzione:  $x_1 = -2 + 1 x_2$ .

$$-3(-2 + x_2) - x_2 = -1 \quad \rightarrow \quad x_2 = 7/4$$

$$x_1 - 1/4 = 2 \quad \rightarrow \quad x_1 = -1/4$$


## Rette e sistemi lineari



Scrivo il sistema lineare:  $Ax = b$

$$A = \begin{bmatrix} 1 & -1 \\ -3 & -1 \end{bmatrix} \quad b = \begin{bmatrix} -2 \\ -1 \end{bmatrix} \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$y = x + 2$$

$$y = -3x + 1$$

$$y = x_1$$

$$x = x_2$$

$$1 x_1 - 1 x_2 = -2$$

$$-3 x_1 - 1 x_2 = -1$$

$X$  è una soluzione se soddisfa **tutte** le equazioni del sistema stesso.



## Soluzione di sistemi lineari quadrati



$$x = A^{-1} b$$

Condizione di esistenza dell'inversa è  $\det(A) \neq 0$

Il sistema ammette 1 ed 1 sola soluzione se  $\det(A) \neq 0$

Altrimenti: **nessuna** o **infinite** soluzioni



## Risoluzione di un sistema 2x2



$$a_{11}x_1 + a_{12}x_2 = b_1$$

$$a_{21}x_1 + a_{22}x_2 = b_2$$

$$y = Ax$$

$$x = A^{-1} y$$

$$A^{-1} = \frac{1}{\det(A)} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

$$\det(A) = a_{11} * a_{22} - a_{12} * a_{21}$$



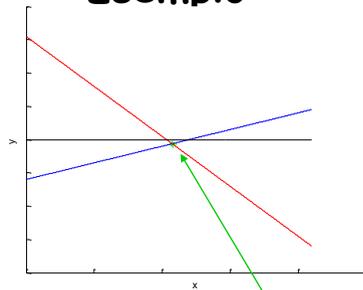
## Esempio

$$y = x + 2$$

$$y = -3x + 1$$

$$A = \begin{bmatrix} 1 & -1 \\ -3 & -1 \end{bmatrix} \quad b = \begin{bmatrix} -2 \\ -1 \end{bmatrix}$$

$$\det(A) = 1(-1) - (-1)(-3) = -1 - 3 = -4 \neq 0$$



$$1 x_1 - 1 x_2 = -2$$

$$-3 x_1 - 1 x_2 = -1$$

$$x_1 = x$$

$$x_2 = y$$

**Rango di A è pieno**

$$P = [-1/4 \quad 7/4]$$

$$A^{-1}$$

$$x = A^{-1} b = -\frac{1}{4} \begin{bmatrix} -1 & +1 \\ +3 & +1 \end{bmatrix} \begin{bmatrix} -2 \\ -1 \end{bmatrix}$$

$$x_1 = -1/4$$

$$x_2 = 7/4$$



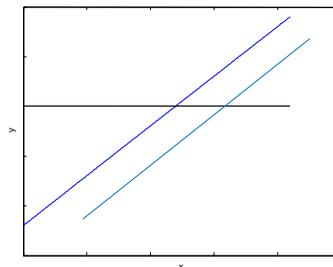
## Esempio di soluzione non univoca ( $\det(A) = 0$ )

$$y = x + 2$$

$$2y = 2x + 3$$

$$A = \begin{bmatrix} 1 & -1 \\ 2 & -2 \end{bmatrix} \quad b = \begin{bmatrix} -2 \\ -3 \end{bmatrix}$$

Non esistono soluzioni



$$1 x_1 - 1 x_2 = -2$$

$$2 x_1 - 2 x_2 = -3$$

$$x_1 = x$$

$$x_2 = y$$

$\det(A) = 1(-2) - (-1)(2) = -2 + 2 = 0$  La soluzione non esiste o  $\infty$  soluzioni.

$$y = x + 2$$

$$2y = 2x + 4$$

La soluzione, non è unica: tutti i punti della retta soddisfano contemporaneamente le 2 equazioni. In questo caso  $\infty$  soluzioni: rette sovrapposte.



## Sistema $M \times N$ , $M > N$



$$a_{11}x_1 + a_{12}x_2 + \dots + a_{1N}x_N = b_1$$

$$A x = b$$

$$a_{21}x_1 + a_{22}x_2 + \dots + a_{2N}x_N = b_2$$

$A$  è  $M \times N$ ,  $M > N$ , non è una matrice quadrata.

.....

$N$  equazioni sono sufficienti per determinare la soluzione.

$$a_{M1}x_1 + a_{M2}x_2 + \dots + a_{MN}x_N = b_M$$

Ho delle equazioni di troppo, devono essere correlate (combinare linearmente), perché il sistema ammetta soluzione.

### Esempio:

$$3x_1 + 2x_2 + \dots + 4x_N = 5$$

$$4x_1 - 2x_2 + \dots + 0.5x_N = 3$$

.....

1, nessuna,  $\infty$  soluzioni.

$$2x_1 + 3x_2 + \dots - 3x_N = -1$$

Posso sempre calcolare la soluzione in forma matriciale.



## Sistemi lineari con $m > n$



$A$  è rettangolare: numero di righe maggiore del numero di colonne

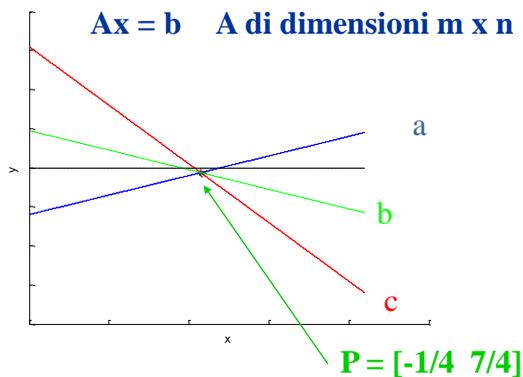
$$y = x + 2$$

$$y = -3x + 1$$

$$y = -x + 3/2$$

Una delle 3 righe di  $A$  è combinazione lineare delle altre. Risolvo per sostituzione

$$A = \begin{bmatrix} 1 & -1 \\ -3 & -1 \\ -1 & -1 \end{bmatrix} \quad b = \begin{bmatrix} -2 \\ -1 \\ -1.5 \end{bmatrix}$$



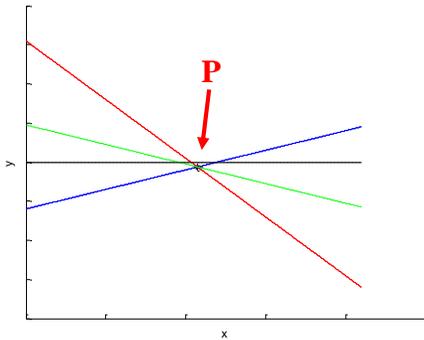
Esiste un'equazione "di troppo"

**Nessuna, 1 o  $\infty$  soluzioni**

**Rango di  $A$  è pieno  $\rightarrow$  1 soluzione**



## Relazione tra le equazioni (combinazione lineare)



$$\alpha_1 (y - x - 2) +$$

$$\alpha_2 (y + 3x - 1) =$$

$$(y + x - 3/2)$$

In questo caso:

$$\alpha_1 = -1/2$$

$$\alpha_2 = -1/2$$

Tutte le rette per la soluzione P possono essere descritte come un fascio (di rette).

Un fascio di rette è univocamente identificato da due rette (che si incontrino in un punto).

La terza equazione è combinazione lineare delle prime due.



## Sistema lineare: soluzione algebrica



Caso generale:

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



$(\mathbf{A}^T \mathbf{A})$  gioca il ruolo di  $\mathbf{A}$  quadrata.

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

Quale criterio viene soddisfatto da  $\mathbf{x}$ ?

$\mathbf{C} = (\mathbf{A}^T * \mathbf{A})^{-1}$  è la matrice di **covarianza** (matrice quadrata  $n \times n$ )



## Sistemi lineari con $m > n$

$$\begin{aligned} y &= x + 2 & x_1 - x_2 &= -2 \\ y &= -3x + 1 & -3x_1 - x_2 &= -1 \\ y &= -x + 3/2 & -x_1 - x_2 &= -3/2 \end{aligned}$$

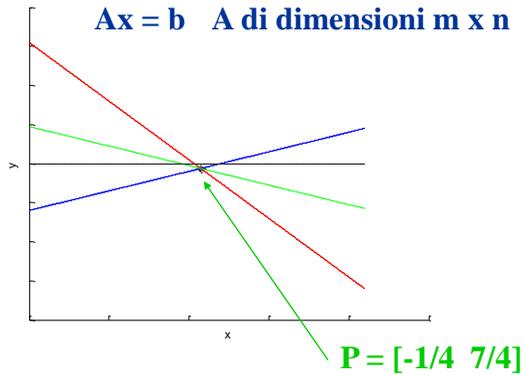
$$A = \begin{bmatrix} 1 & -1 \\ -3 & -1 \\ -1 & -1 \end{bmatrix} \quad b = \begin{bmatrix} -2 \\ -1 \\ -1.5 \end{bmatrix}$$

$$A^T * A = \begin{bmatrix} 11 & 3 \\ 3 & 3 \end{bmatrix} \quad \det = 24$$

$$C = (A^T A)^{-1} = \begin{bmatrix} 0.1250 & -0.1250 \\ -0.1250 & 0.4583 \end{bmatrix}$$

$$x = C * A^T * b \quad P = [-0.25 \quad +1.75]$$

**intersezione**

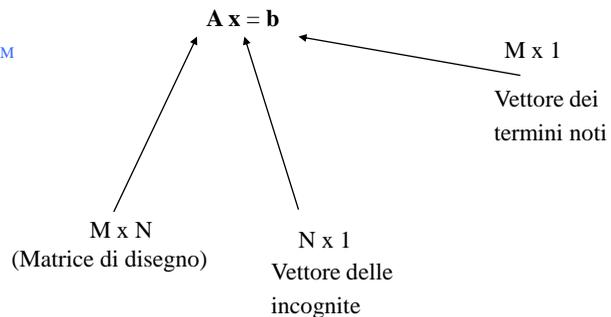


## Sistema lineare con errore sul termine $b_i$

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1N}x_N &\neq b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2N}x_N &\neq b_2 \end{aligned}$$

.....

$$a_{M1}x_1 + a_{M2}x_2 + \dots + a_{MN}x_N \neq b_M$$



Esiste una soluzione? Qual è il valore di  $x$  che posso calcolare?



## Riformulazione del problema con errore



$$a_{11}x_1 + a_{12}x_2 + \dots a_{1N}x_N = b_1 + v_1$$

$$a_{21}x_1 + a_{22}x_2 + \dots a_{2N}x_N = b_2 + v_2$$

.....

$$a_{M1}x_1 + a_{M2}x_2 + \dots a_{MN}x_N = b_M + v_M$$

Errore di modello (sistematico, randomico, additivo,  $v$ ).  $M \times 1 \Rightarrow \{v_i\} = \text{residuo}$

$$\mathbf{A} \mathbf{x} = \mathbf{b} + \mathbf{N}$$

$M \times N$   
(Matrice di disegno)

$N \times 1$   
Vettore delle incognite

$M \times 1$   
Vettore dei termini noti

Esiste una soluzione? Qual è il valore di  $x$  che posso calcolare?

A.A. 2020-2021

27/55

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



## Significato del problema con errore - I



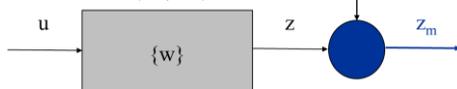
$$a_{11}x_1 + a_{12}x_2 + \dots a_{1N}x_N = b_1 + v_1$$

$$a_{21}x_1 + a_{22}x_2 + \dots a_{2N}x_N = b_2 + v_2$$

.....

$$a_{M1}x_1 + a_{M2}x_2 + \dots a_{MN}x_N = b_M + v_M$$

$$\mathbf{z} = \mathbf{f}(\mathbf{u} | \mathbf{w}) = \mathbf{A} \cdot \mathbf{u}$$



$v_i$  residuo

$$\mathbf{A} \mathbf{x} = \mathbf{b} + \mathbf{N}$$

$M \times N$   
(Matrice di disegno)

$N \times 1$   
Vettore delle incognite

$M \times 1$   
Vettore dei termini noti

### Problema inverso:

- $U$  ha il ruolo di  $X$  (incognite)
- $f(\cdot)$  ha il ruolo di matrice  $A$
- $z$  ha il ruolo di  $b$
- $z$  misurata,  $z_m$ , ha il ruolo di  $b + v$

A.A. 2020-2021

28/55

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



## Significato del problema con errore - II

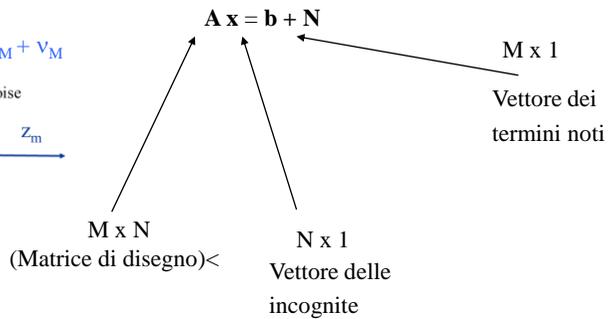
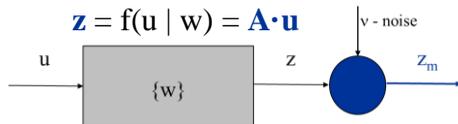


$$a_{11}x_1 + a_{12}x_2 + \dots + a_{1N}x_N = b_1 + v_1$$

$$a_{21}x_1 + a_{22}x_2 + \dots + a_{2N}x_N = b_2 + v_2$$

.....

$$a_{M1}x_1 + a_{M2}x_2 + \dots + a_{MN}x_N = b_M + v_M$$



**Problema inverso (learning):**

- *W* ha il ruolo di *X* (incognite)
- $f(\cdot)$  ha il ruolo di matrice *A*
- $z$  ha il ruolo di  $b$
- $z$  misurata,  $z_m$ , ha il ruolo di  $b + v$

A.A. 2020-2021

29/55

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



## Soluzione come problema di ottimizzazione



$$Ax = b + N$$

$$\text{Funzione costo: } \sum_k v_k^2 = \|Ax - b\|^2$$

Assegno un costo al fatto che la soluzione  $x$ , non soddisfi tutte le equazioni, la somma dei residui associati ad ogni equazione viene minimizzata.

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

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

$$A^T Ax = A^T b$$

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

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.

A.A. 2020-2021

30/55

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



## Sistemi lineari con $m > n$

$$\begin{aligned}x_1 - x_2 &= -2 \\ -3x_1 - x_2 &= -1 \\ -x_1 - x_2 &= -3/2\end{aligned}$$

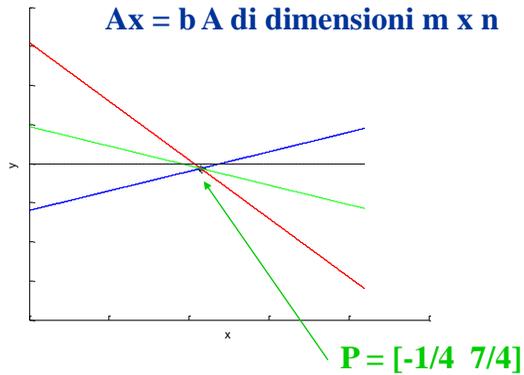
$$A = \begin{bmatrix} 1 & -1 \\ -3 & -1 \\ -1 & -1 \end{bmatrix} \quad b = \begin{bmatrix} -2 \\ -1 \\ -1.5 \end{bmatrix}$$

$$A^T * A = \begin{bmatrix} 11 & 3 \\ 3 & 3 \end{bmatrix} \quad \det = 24$$

$$C = (A^T A)^{-1} = \begin{bmatrix} 0.1250 & -0.1250 \\ -0.1250 & 0.4583 \end{bmatrix}$$

$$x = C * A^T * b \quad P = [-0.25 \quad +1.75]$$

**intersezione**



$$\|Ax - b\| = 0$$

A.A. 2020-2021

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



## Sistemi lineari con $m > n$ - non esiste soluzione (matematica)

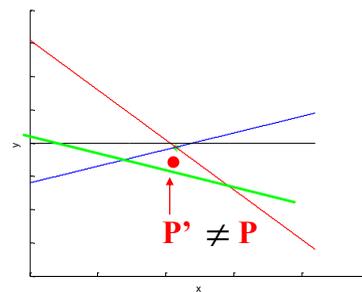
$$\begin{aligned}x_1 - x_2 &= -2 + 0 \\ 3x_1 + x_2 &= +1 - 0.5 = +0.5 \\ -x_1 - x_2 &= -3/2 + 0 = -3/2\end{aligned}$$

$$A = \begin{bmatrix} 1 & -1 \\ 3 & 1 \\ -1 & -1 \end{bmatrix} \quad b = \begin{bmatrix} -2 \\ 0.5 \\ -1.5 \end{bmatrix}$$

$$A^T * A = \begin{bmatrix} 11 & 3 \\ 3 & 3 \end{bmatrix} \quad \det = 24$$

$$C = (A^T A)^{-1} = \begin{bmatrix} 0.1250 & -0.1250 \\ -0.1250 & 0.4583 \end{bmatrix}$$

$$AX = b \quad A \text{ di dimensioni } m \times n$$



$$\sum_k v_k^2 = \|Ax - b\|^2 = 0.2041241$$

$$x = C * A^T * b \quad P' = [-0.375 \quad +1.70833]$$

**No intersezione**

$$\|Ax - b\| \neq 0$$

A.A. 2020-2021

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



## Soluzione ai minimi quadrati (least squares solution)



$$\sum_k v_k^2 = \|Ax - b\|^2 = \sum_k \|A_{k,*}x - b_k\|^2 = \text{Sommo per tutte le righe}$$

$$[(A_{11}x_1 + A_{12}x_2) - b_1]^2 + [(A_{21}x_1 + A_{22}x_2) - b_2]^2 + [(A_{31}x_1 + A_{32}x_2) - b_3]^2$$

**P' = [-0.375 +1.70833]**  $x_1 - x_2 = -2$        $-0.375 - 1.70833 + 2 = v_1 = -0.083333$   
**No intersezione**       $3x_1 + x_2 = +1/2$        $-1.125 + 1.70833 - 0.5 = v_2 = 0.083333$   
                                   $-x_1 - x_2 = -3/2$        $+0.375 - 1.70833 + 1.5 = v_3 = 0.16666$

Lo scarto misura la somma quadratic delle distanze (verticali) tra il punto che rappresenta la soluzione e le 3 rette.



## Sistema lineare: soluzione robusta



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

Numero di condizionamento varia circa con la norma di  $(A^T A)$ .

*Soluzione tramite Singular Value Decomposition*

Numero di condizionamento varia circa con  $A$ .

$$A x = b$$

$$U W V^T x = b$$

$$x = V^T W^{-1} U^T b$$

Ortonormale  $M \times N$

Diagonale ( $N \times N$ )

Ortonormale  $N \times N$

- La matrice  $C = (A^T A)^{-1}$  non viene formata.
- $W^{-1}$  contiene i reciproci degli elementi di  $W$ .

$W^{-1}$  è diagonale.  $w_{ii}^{-1} = 1/w_{ii}$



# Overview



Modelli

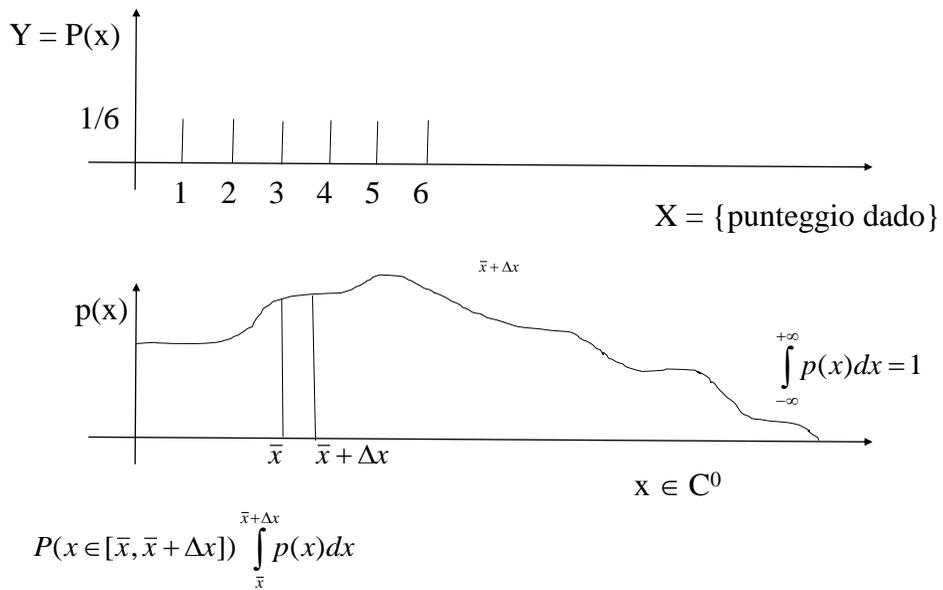
Sistemi lineari

Distribuzione di probabilità

Massima verosimiglianza



## La probabilità nel caso continuo





# Definizione di p(x)

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

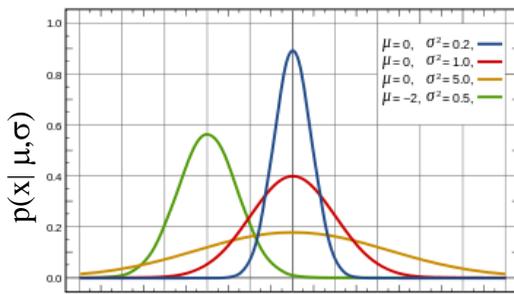
Caso continuo: i valori che X può assumere sono infiniti. Devo trovare un modo per definirne la probabilità. Descrizione **analitica** mediante la funzione densità di probabilità. Si considera la probabilità che x cada in un certo intervallo.

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

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



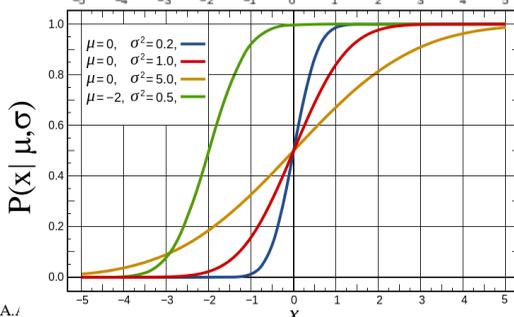
# Distribuzioni notevoli: la Gaussiana



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

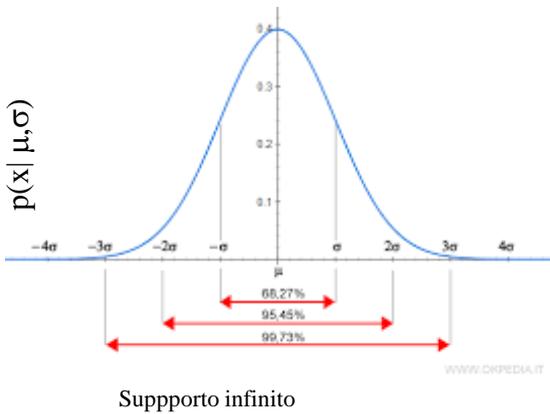
D = dimensione, in questo caso D = 1

Il valore con probabilità più elevato è il valore medio  $\mu$ .





## Concentrazione dei valori



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

D = dimensione, in questo caso D = 1

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

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

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



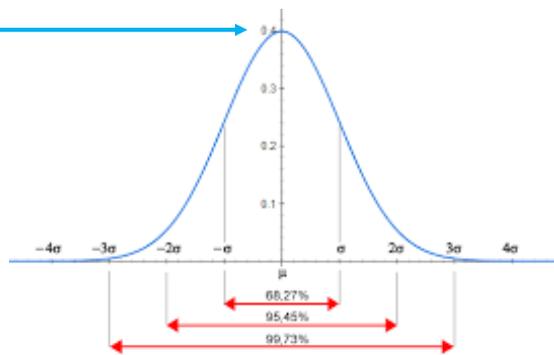
## I momenti di una variabile statistica



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

Valore atteso (Expected value) di X = media distribuzione

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





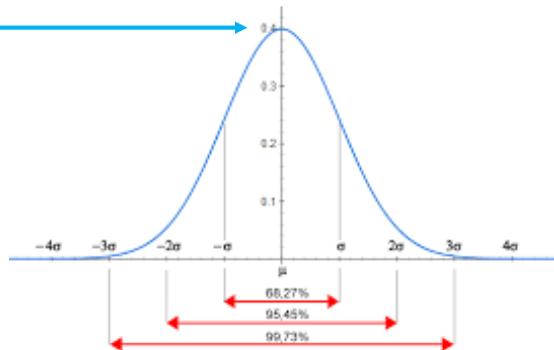
# I momenti di una variabile statistica



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

Valore atteso (Expected value) di X = media distribuzione

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



A.A. 2020-2021

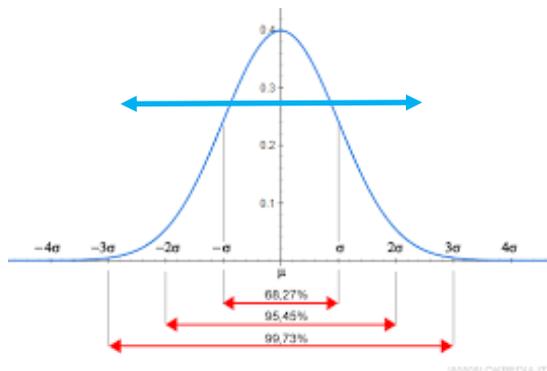
WWW.OKPEDIA.IT mi.it\



# I momenti di una variabile statistica



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



A.A. 2020-2021

42/55

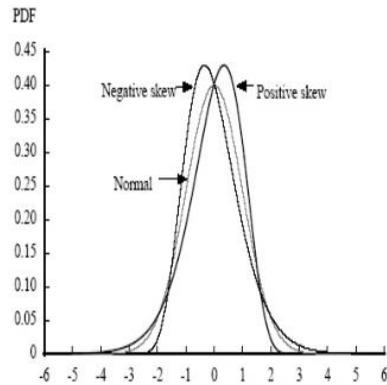
http:\borghese.di.unimi.it\



# I momenti di una variabile statistica

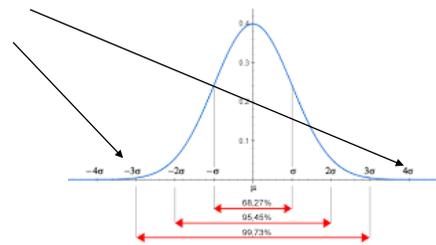
Asimmetria

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



Kurtosi – peso delle code di p(x)

$$E[(X - \mu)^4] = \int_{-\infty}^{+\infty} (x - \mu)^4 p(x)$$



## Overview



Modelli

Sistemi lineari

Distribuzione di probabilità

**Massima verosimiglianza**



## Probabilità di un certo insieme di misure

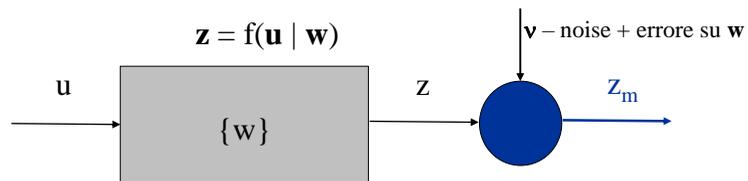


$z = f(u | w)$  misura  $\{z_i\}$  in corrispondenza di  $\{u_i\}$ .  $\{z_i\}$  è ottenuto come uscita del modello, tramite i parametri  $\{w_j\}$

Avrò che:  $f(u_i, w) = z_{i,m} = z_i + v_i$

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

$$p(z_{1m}, z_{2m}, z_{3m}) = \prod_i p(z_{im}) \quad (\text{cf. dadi nel caso discreto})$$



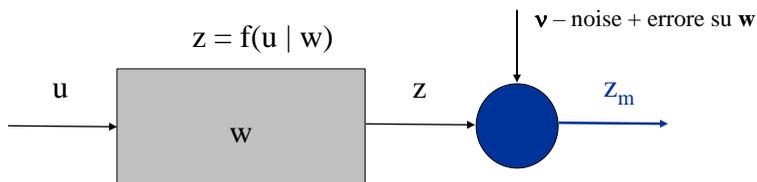
A.A. 2020-2021

45/55

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



## Dipendenza delle misure



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

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

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

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

A.A. 2020-2021

46/55

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



# Esempio: fitting di una retta



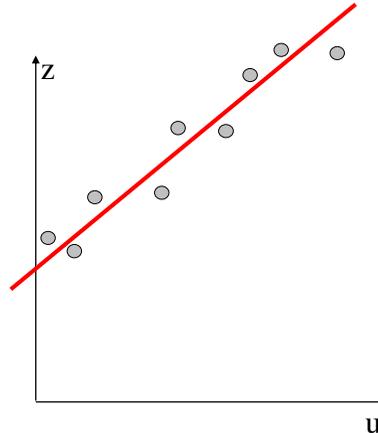
Vogliamo stimare i parametri di una retta:  $z = f(u | w) = m u + q$ , con  $m$  e  $q$  incogniti:  $W = \{m, q\}$

**La retta è un modello lineare.**

Abbiamo a disposizione  $N$  misure rumorose effettuate.

Supponiamo che le  $z_i$  siano affette da errore **Gaussiano a media nulla.**

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



$z_{im} = z_i + v_i = (m u_i + q) + v_i$  dove  $v_i$  è l'errore di misura, **Gaussiano a media nulla.**



# Stima dei parametri del modello

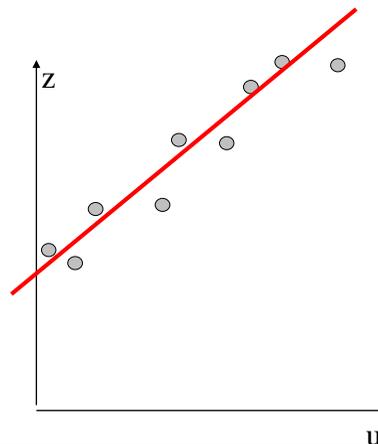


Per ogni punto, dovrebbe valere  $z_i = m u_i + q$ .

Ma c'è l'errore di misura, misuriamo in realtà  $z_i + v_i = z_{im}$

$$v_i = z_{im} - (m u_i + q)$$

**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 indipendenti**... 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



- In questo caso le  $z$  sono legate alla variabile indipendente  $u$  da  $f(u, w)$ .  
Nel caso della retta  $f(\cdot) = mu + q$
- 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**  $L=L(z | u, w)$  rispetto a  $w$

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

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



## Osservazioni

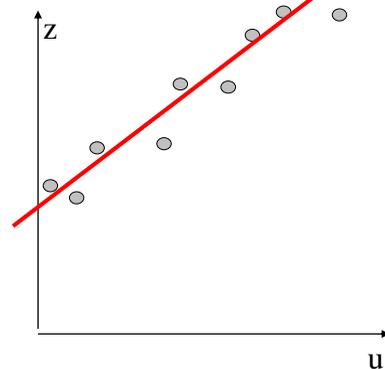
$$L(z_{1m}, z_{2m}, z_{3m}, \dots, z_{Nm} | (w; u_1, u_2, u_3, \dots, u_{Nm})) = \\ = p(z_{1m} | w; u_1) p(z_{2m} | w; u_2) p(z_{3m} | w; u_3) \dots p(z_{Nm} | w; u_N)$$

- Più in generale, le variabili possono avere un residuo,  $v$ , descritto da densità di probabilità diverse.
- La relazione tra ingresso e uscita è la stessa per tutte le misure, ed è rappresentata dal modello.
- La forma del modello dipende da un insieme di parametri,  $w$ .
- Massimizzando la funzione di verosimiglianza rispetto a tali parametri se ne effettua la stima in modo tale che il vettore osservato  $z_m = \{z_{im}\}_{i=1 \dots N}$  sia massimamente probabile (massima verosimiglianza) ovvero i valori prodotti dal modello siano il più vicino possibile ai valori misurati.



## Esempio della retta

- La funzione di verosimiglianza dipende dai parametri che definiscono le densità di probabilità delle variabili casuali che entrano nella verosimiglianza.
- Modificando il valore dei parametri adatto la funzione  $f(\cdot)$  in modo che l'uscita  $z$  in corrispondenza degli input  $u$  sia la più vicina a  $z_m$ .
- Nel caso della retta ruoto e traslo la retta in modo tale che si avvicini "il più possibile" ai punti misurati.





# Stima alla massima verosimiglianza per modello lineare

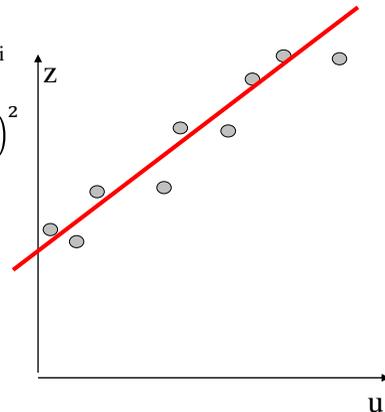


- Equazione di una retta:  $z = mu + q$
- Scriviamo prima di tutto la densità di probabilità di ottenere  $z_{im}$  per ciascun dato, per errore,  $v_i$ , Gaussiano a media nulla:

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

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

dove  $m$  e  $q$  non sono note.



# Stima a massima verosimiglianza



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

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

Devo trovare un modo efficiente per massimizzare la funzione di verosimiglianza, o likelihood,  $L(\cdot)$  rispetto ai parametric  $w$  che determinano la forma del modello.



# Overview



Modelli

Sistemi lineari

Distribuzione di probabilità

**Massima verosimiglianza**