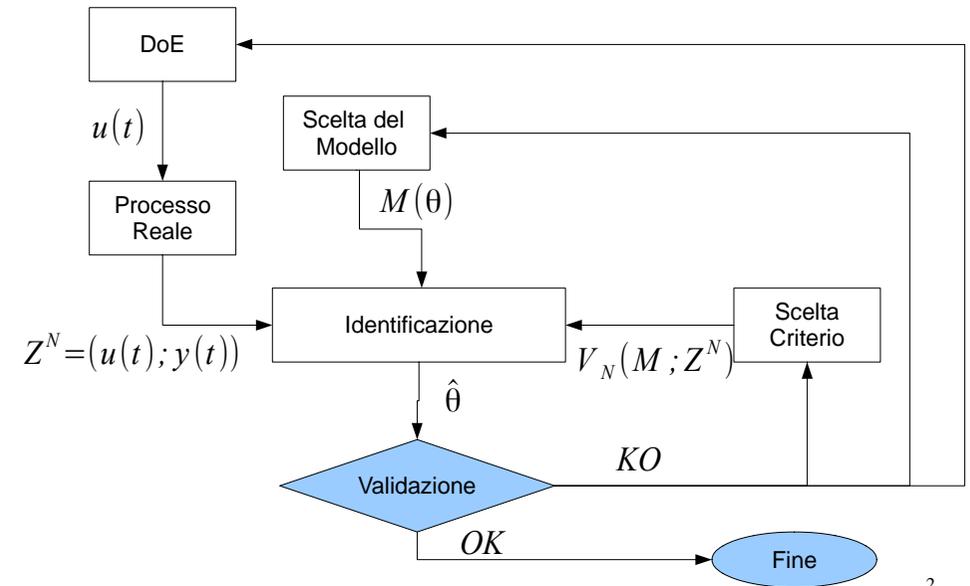


Validazione

Principali criteri

- Problema della validazione
- Expected fit
- Cross Validazione
- Test di bianchezza (cambio dei segni)

Processo di identificazione



2

Validazione - possibili approcci

La validazione può avvenire secondo due approcci

- **Assoluto:** ho un modello le prestazioni sono ottimali
- **Per confronto:** definire il migliore fra due modelli
 - di famiglie diverse avente stesso ordine
 - della stessa famiglia, con ordine diverso
- Entrambe le soluzioni si basano su indici e test.

3

Problema di validazione - I

• Problema di scelta della struttura

Dati 2 modelli di diversa famiglia ma lo stesso numero di parametri p e identificati dallo stesso data set Z^N ;

$$M_1(\hat{\theta}_1(Z^N)) \quad M_2(\hat{\theta}_2(Z^N)) \quad \hat{\theta}_1, \hat{\theta}_2 \in R^p$$

determinare il modello "migliore"

• Esempio:

- Segnale di ingresso: $u(t) \sim WGN(10; 2)$
- Famiglie: $ARX(na, nb); OE(nb, nf) \quad na + nb = nf + nb$
- Strutture $M_1(\theta_1): ARX(2,1) \quad M_2(\theta_2): OE(2,1)$
- Modelli da scegliere: $M_1(\hat{\theta}_1(Z^N)) \quad M_2(\hat{\theta}_2(Z^N))$

4

Problema di validazione - II

- **Problema di scelta dell'ordine:**

Dati 2 modelli della stessa famiglia con diverso ordine identificati a partire dallo stesso data set Z^N ;

$$M_1(\hat{\theta}_1(Z^N)) \quad M_2(\hat{\theta}_2(Z^N)) \quad \text{con } M_1 \subset M_2$$

determinare il modello "migliore"

- **Esempio:**

- Segnale di ingresso: $u(t) \sim WGN(10; 2)$
- Famiglia di modelli $ARX(na, nb)$
- Ordini $M_1(\theta_1): ARX(2,1)$ $M_2(\theta_2): ARX(2,2)$
- Modelli da scegliere $M_1(\hat{\theta}_1(Z^N))$ $M_2(\hat{\theta}_2(Z^N))$

5

Indice di valutazione - Expected fit

- **Idea:** scelgo il modello che minimizza l'errore di predizione.

$$SSR = \sum_{t=1}^N (y(t) - \hat{y}(t|\hat{\theta}))^2$$

- Con segnali a media nulla, si può usare un indice normalizzato

$$R^2 = 1 - \frac{SSR(\hat{\theta}, Z^N)}{\sum_{i=1}^N y^2(t)}$$

- **Osservazione:** I dati a media nulla si hanno imponendo

$$u_n(t) = u(t) - \frac{1}{N} \sum_{i=1}^N u(i) \quad y_n(t) = y(t) - \frac{1}{N} \sum_{i=1}^N y(i)$$

- **Interpretazione:** l'indice R^2 indica la percentuale di dati "spiegati dal modello".

6

Expected Fit - scelta della struttura

Esempio

- Sistema $y(t) = [0.02q^{-1} + 0.3q^{-2} - 0.2q^{-3} + 0.01q^{-4}]y(t) + q^{-1}u(t) + e(t)$
- DoE: $u(t) \sim WGN(10,1)$ $e(t) \sim WGN(1;1)$

- Dati normalizzati:

$$u_n(t) = u(t) - \frac{1}{N} \sum_{i=1}^N u(i) \quad y_n(t) = y(t) - \frac{1}{N} \sum_{i=1}^N y(i)$$

- Modelli: $M_1: ARX(2,1)$ $M_2: ARX(1,2)$
- Expected fit

$$R^2(M_1) = 0.9502 \quad R^2(M_2) = 0.9054$$

7

Expected Fit - interpretazione

- Sistema $FIR(3)$

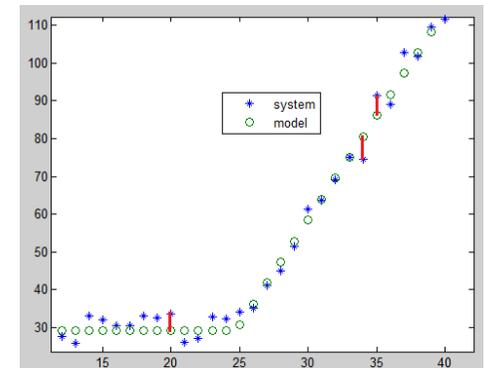
$$y(t) = [2q^{-1} + 3q^{-2} + 2q^{-3}]u(t) + e(t) \quad e(t) \sim WGN(1;1)$$

- DoE: $u(t) = \begin{cases} 0 & t < 25 \\ t - 25 & t \geq 25 \end{cases}$

- Modello identificato

$$\hat{y}(t) = [1.92q^{-1} + 5.01q^{-2}]u(t)$$

- In rosso si vede come l'errore di predizione sia quanto manca alla predizione per "raggiungere" l'uscita del sistema reale



8

Expected Fit - scelta dell'ordine

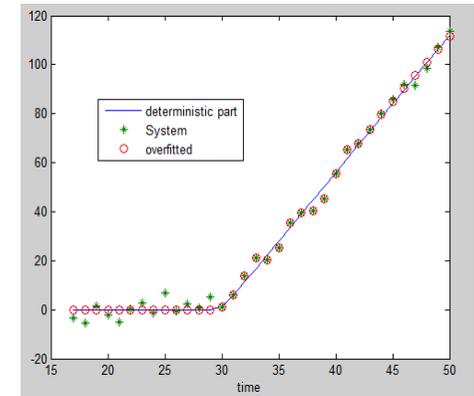
- **Fatto:** si dimostra che $E[R^2]$ cresce con l'ordine del modello identificato
- **Conseguenza** il modello di ordine maggiore è il migliore
- Esempio
 - Sistema $y(t)=[0.02q^{-1}+0.3q^{-2}-0.2q^{-3}+0.01q^{-4}]y(t)+q^{-1}u(t)+e(t)$
 - DoE: $u(t)\sim WGN(10,1)$ $e(t)\sim WGN(1;1)$
 - Modelli: $M_i: ARX(i,1)$ $i=1..10$
 - Expected fit (100 iterazioni)

N	1	2	3	4	5	6	7	8	9	10
SSR(%)	84.01	95.02	98.37	99.17	99.19	99.20	99.21	99.22	99.23	99.24

9

Overfit - esempio

- Sistema $FIR(3)$ $y(t)=[2q^{-1}+3q^{-2}+2q^{-3}]u(t)+e(t)$ $e(t)\sim WGN(1;1)$
- DoE: $u(t)=\begin{cases} 0 & t < 25 \\ t-25 & t \geq 25 \end{cases}$
- Modello identificato $FIR(10)$
- Si vede come il modello tenda ad inseguire l'errore invece che la componente determinista del segnale come auspicabile.



10

Cross validazione - Idea

- **Fatto:** $E[SSR(\hat{\theta}, Z^N)]$ diminuisce al crescere dell'ordine.
- **Fatto:** identificazione e validazione usano lo stesso criterio
- **Congettura:** il modello identificato descrive anche l'errore [fenomeno di overfitting]
- **Fatto:** $SSR(\hat{\theta}, Z^N)$ è legato anche ai dati.
- **Idea:** Svincolo l'identificazione dalla validazione se non nel criterio nei dati creando due set distinti
 - per l'identificazione $Z_i^{N_1}$
 - per la validazione $Z_v^{N_2}$
- **Osservazione:** tipicamente $N/4 \leq N_2 \leq N/2$

11

Cross validazione - definizione

1. divido i dati in due parti tali che

$$Z^N = Z_i^{N_1} \cup Z_v^{N_2} \quad N_1 + N_2 = N$$

2. identifico i modelli usando solo $Z_i^{N_1}$

$$M_1(\hat{\theta}_1(Z_i^{N_1})) \quad M_2(\hat{\theta}_2(Z_i^{N_1}))$$

3. valuto i vari SSR usando $Z_v^{N_2}$

$$SSR_1(\hat{\theta}_1, Z_v^{N_2}) = \sum_{t=N_1+1}^N (y(t) - \hat{y}_1(t|\hat{\theta}_1))^2$$

$$SSR_2(\hat{\theta}_2, Z_v^{N_2}) = \sum_{t=N_1+1}^N (y(t) - \hat{y}_2(t|\hat{\theta}_2))^2$$

4. Scelgo il modello con SSR minore

12

Analisi dei residui - Idea

- **Fatto:** Se il modello è corretto l'errore di predizione è bianco
- **Idea:** calcolo l'errore di predizione e verifico se questo è bianco se lo è il modello è accettato altrimenti proseguo la ricerca.
- Esistono diversi test "di bianchezza"
- **Nota:** Alcuni autori definiscono questa approccio e le relative tecniche come "Analisi dei Residui".

13

Test del cambio dei segni - I

- Ipotesi:
 $w(t) \sim WN \quad P(w(t) > 0) = 0.5 \quad P(w(t) < 0) = 0.5$
- Definisco:

$$x_N = \sum_{i=1}^{N-1} s_i \quad s_i = \begin{cases} 1 & w(i)w(i+1) < 0 \\ 0 & w(i)w(i+1) > 0 \end{cases}$$
- Si dimostra che

$$x_N \underset{N \rightarrow \infty}{\sim} N \left(\frac{N}{2}, \frac{N}{4} \right)$$
- Pertanto si ha che

$$\frac{x_N - \frac{N}{2}}{\frac{\sqrt{N}}{2}} \underset{N \rightarrow \infty}{\sim} Z$$

14

Test del cambio dei segni II

- Dato un segnale $w(t)$
- calcolato $x_N = \sum_{i=1}^{N-1} s_i \quad s_i = \begin{cases} 1 & w(i)w(i+1) < 0 \\ 0 & w(i)w(i+1) > 0 \end{cases}$
- Fissato:
 - il livello di significatività α (tipicamente $\alpha = 0.05$)
 - trovo z_α tale che $P(|Z| \leq z_\alpha) = 0.95$.
- Adotto la regola:

$$\left| \frac{x_N - \frac{N}{2}}{\frac{\sqrt{N}}{2}} \right| < z_\alpha \rightarrow w(t) \sim WN \quad \left| \frac{x_N - \frac{N}{2}}{\frac{\sqrt{N}}{2}} \right| \geq z_\alpha \rightarrow w(t) \text{ colorato}$$

15

Utilizzo pratico

- Non vi è una regola precisa
- Possibile approccio 1
 - Scelta della struttura (metodo validazione 2)
 - Scelgo l'ordine (metodo validazione 2)
- Possibile approccio 2
 - Scelta della struttura (metodo validazione 1)
 - Scelgo l'ordine (metodo validazione 2)
- Possibile approccio 3
 - Testo tutti i modelli insieme
 - Combino i metodi
(il modello ottimale è quello con ssr minore fra quelli a residuo bianco)

16