

Sistemi di Controllo Digitale

Identificazione di Modelli Dinamici

Contenuto delle Lezioni

- Il problema della stima e della predizione
- Teoria della stima e caratteristiche degli stimatori
- Stima ai minimi quadrati
- Scelta della complessità dei modelli
- *Il toolbox di identificazione di Matlab*

25/01/2012

2

Il Problema della Stima e della Predizione

Il problema della stima nasce quando si vogliono determinare uno o più **parametri incogniti a partire da osservazioni sperimentali**.



Nella gran maggioranza dei casi i **parametri incogniti sono costanti** $\vartheta(t) = \vartheta$

$T = \{t_1, t_2, \dots, t_N\}$ insieme degli istanti di osservazione

- In generale non è detto che i t_i debbano avere cadenza regolare
- Se si ha accesso alla generazione dei t_i conviene "addensarli" dove le osservazioni sperimentali sono più significative

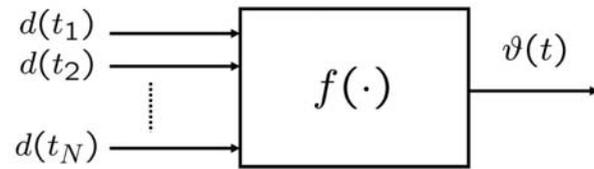
25/01/2012

3

25/01/2012

4

Stimatore



Lo stimatore e' quindi una **funzione deterministica** che a partire dai dati osservati in ingresso produce in uscita i parametri incogniti.

Stima di parametri costanti

- Se $v(t) = \bar{v} = \text{cost}$ si parla di **stima** o **identificazione parametrica**.
- La stima prodotta dallo stimatore si indica con \hat{v} o con \hat{v}_T qualora si voglia evidenziare l'insieme degli istanti di osservazione.
- Il valore "vero" del parametro incognito si indica con v^o

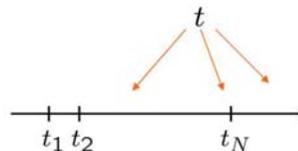
Stima di parametri varianti nel tempo

- La stima prodotta dallo stimatore si indica con $\hat{v}(t|T)$ o semplicemente con $\hat{v}(t|N)$ qualora si possa prendere $T = \{1, 2, \dots, N\}$
- A seconda di come si ponga l'istante t rispetto a t_N si hanno i tre casi:

$t > t_N$: problema di **predizione**

$t = t_N$: problema di **filtraggio**

$t < t_N$: problema di **regolarizzazione (smoothing)**



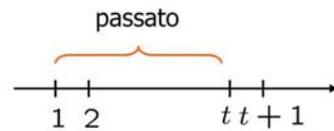
Problema della predizione

Si tratta di un problema fondamentale nell'ambito **dell'identificazione dei sistemi dinamici**.

- Per fissare le idee consideriamo il caso delle *serie temporali*.
- Si ha una sequenza di osservazioni $y(1), y(2), \dots, y(N)$ di una variabile $y(\cdot)$
- Si vuole stimare $y(t + 1)$
- Si vuole quindi determinare un **predittore**

$$\hat{y}(t + 1|t) = f[y(t), y(t - 1), \dots, y(1)]$$

- Il predittore esprime quindi una **stima** $\hat{y}(t+1|t)$ di $y(t+1)$ in funzione di N valori assunti da $y(\cdot)$ nel **passato**.



- Un predittore si dice **lineare** se

$$\hat{y}(t+1|t) = a_1(t)y(t) + \dots + a_t(t)y(1)$$

- Un predittore si dice **a memoria finita** (ovvero utilizza una memoria limitata del passato) se

$$\hat{y}(t+1|t) = a_1(t)y(t) + \dots + a_n(t)y(t-n+1)$$

- Un predittore si dice **lineare tempo-invariante** se

$$\hat{y}(t+1|t) = a_1 y(t) + \dots + a_n y(t-n+1)$$

dove i parametri a_1, \dots, a_n sono **costanti**

- Si definisce il vettore dei parametri $\vartheta^\top = [a_1 \dots a_n]$

Determinare un buon predittore significa determinare un opportuno vettore ϑ tale che la predizione $\hat{y}(t+1|t)$ sia il piu' possibile accurata.

Piu' precisamente:

- Si consideri un predittore **lineare tempo-invariante a memoria finita**

$$\hat{y}(t+1|t) = a_1 y(t) + \dots + a_n y(t-n+1)$$

dove n e' piccolo rispetto al numero di dati misurati fino all'istante t

- La capacita' predittiva del predittore puo' quindi essere valutata rispetto ai dati gia' noti $y(i), i = 1, \dots, t$:

- si calcola

$$\hat{y}(i+1|i) = a_1 y(i) + \dots + a_n y(i-n+1), \forall i > n$$

- si valuta **l'errore di predizione**

$$\varepsilon(i+1) = y(i+1) - \hat{y}(i+1|i), \forall i > n$$

Il vettore $\vartheta^\top = [a_1 \dots a_n]$ e' "buono" se ε e' piccolo su tutti i dati disponibili

- definiamo allora una cifra di merito:

$$J(\vartheta) = \sum_{i=n+1}^t \varepsilon(i)^2$$

- quindi

$$\vartheta^\circ = \arg \min_{\vartheta} J(\vartheta)$$

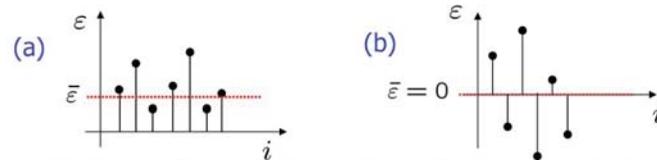
La determinazione di ϑ° e' quindi ricondotta alla soluzione di un problema di ottimizzazione.

Osservazioni

E' molto importante chiarire che cosa si intende per ε "piccolo"



la minimizzazione di $J(\vartheta)$ non e' di per se' un criterio soddisfacente



- Caso (a): non e' soddisfacente perche' l'errore medio $\bar{\varepsilon}$ e' non nullo errore sistematico
- Caso (b): non e' soddisfacente perche' nonostante l'errore medio $\bar{\varepsilon}$ sia nullo, la sequenza e' alternata, quindi ad ogni passo si sa gia' che segno avra' l'errore al passo successivo

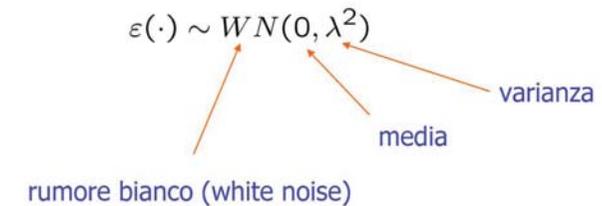
Il predittore non ha acquisito tutta l'informazione

25/01/2012

13

Quindi:

La situazione ideale sarebbe avere un errore ε di valor medio il piu' piccolo possibile e che abbia una natura il piu' possibile **imprevedibile**.



25/01/2012

14

Predittore come sistema dinamico

$$\hat{y}(t|t-1) = a_1 y(t-1) + \dots + a_n y(t-n)$$

$$\varepsilon(t) = y(t) - \hat{y}(t|t-1) \quad \text{green arrow} \quad y(t) = \varepsilon(t) + \hat{y}(t|t-1)$$

$$\text{green arrow} \quad y(t) = a_1 y(t-1) + \dots + a_n y(t-n) + \varepsilon(t)$$

$$\text{green arrow} \quad y(t) = (a_1 z^{-1} + \dots + a_n z^{-n}) y(t) + \varepsilon(t)$$

$$\text{green arrow} \quad A(z) y(t) = \varepsilon(t) \quad \text{con} \quad A(z) = 1 - a_1 z^{-1} - \dots - a_n z^{-n}$$

$$\text{green arrow} \quad y(t) = \frac{1}{A(z)} \varepsilon(t)$$

25/01/2012

15

Teoria della Stima e Caratteristiche degli Stimatori

25/01/2012

16

Generalita`

- In generale abbiamo:

$$d = d(s, \vartheta^{\circ})$$

dove

- d sono i dati osservati
- ϑ° e` la quantita` da stimare
- s e` l'esito dell'esperimento casuale

- Lo stimatore e` una funzione:

$$\hat{\vartheta} = f [d(s, \vartheta^{\circ})]$$



Lo stimatore e` variabile aleatoria in quanto il suo valore dipende dall'esito dell'esperimento casuale s

25/01/2012

17

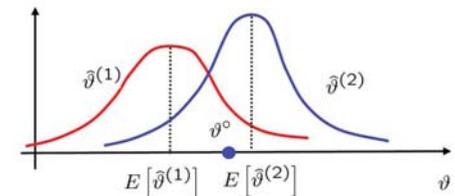
Polarizzazione

- In generale lo stimatore $\hat{\vartheta} = f [d(s, \vartheta^{\circ})]$ si dice **non polarizzato** se

$$E(\hat{\vartheta}) = \vartheta^{\circ}$$

- La non polarizzazione di uno stimatore e` evidentemente una proprieta` che bisogna cercare di assicurare.

Nel caso in figura gli stimatori sono ambedue polarizzati ma lo stimatore $\hat{\vartheta}^{(2)}$ ha una polarizzazione minore.



25/01/2012

18

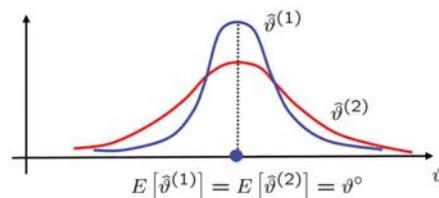
Minima varianza

- La non polarizzazione (o correttezza) non e` l'unico criterio con cui valutare la qualita` di uno stimatore.

Nel caso in figura gli stimatori sono ambedue non polarizzati.

Pero`

$$\text{var} [\hat{\vartheta}^{(1)}] \ll \text{var} [\hat{\vartheta}^{(2)}]$$



- Quindi lo stimatore $\hat{\vartheta}^{(1)}$ ha probabilita` maggiore di produrre valori della stima vicini al valore vero ϑ° rispetto allo stimatore $\hat{\vartheta}^{(2)}$

- Si vuole quindi che la varianza dello stimatore sia la piu` piccola possibile.

25/01/2012

19

- In generale, a parita` di caratteristiche di polarizzazione, diremo che lo stimatore $\hat{\vartheta}^{(1)}$ e` migliore dello stimatore $\hat{\vartheta}^{(2)}$ se

$$\text{var} [\hat{\vartheta}^{(1)}] \leq \text{var} [\hat{\vartheta}^{(2)}]$$

ovvero se la matrice (ϑ puo` essere un vettore)

$$\text{var} [\hat{\vartheta}^{(2)}] - \text{var} [\hat{\vartheta}^{(1)}] \geq 0$$

- Ricordiamo che $A \geq 0 \implies \det A \geq 0, \lambda_i \geq 0, a_{ii} \geq 0$

Quindi

$$\text{var} [\hat{\vartheta}^{(2)}] - \text{var} [\hat{\vartheta}^{(1)}] \geq 0 \implies \text{var} [\hat{\vartheta}_i^{(2)}] \geq \text{var} [\hat{\vartheta}_i^{(1)}]$$

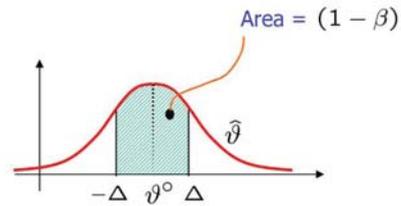
dove $\hat{\vartheta}_i^{(1)}, \hat{\vartheta}_i^{(2)}$ rappresentano le componenti i -esime dei vettori $\hat{\vartheta}^{(1)}, \hat{\vartheta}^{(2)}$

25/01/2012

20

Confidenza della stima

Consideriamo uno stimatore $\hat{\vartheta}$:

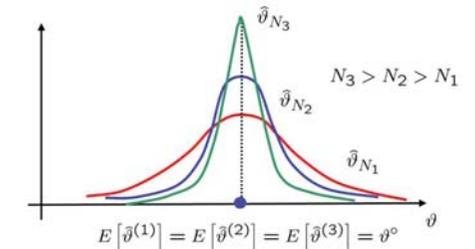


Si dice che la stima $\hat{\vartheta}$ è entro l'intervallo $(-\Delta, \Delta)$ attorno a ϑ^0 con confidenza $(1 - \beta) \cdot 100\%$

Caratteristiche asintotiche

- Se il numero N di dati a disposizione cresce nel tempo
 - ↳ aumenta l'informazione a disposizione per effettuare la stima
 - ↳ diminuisce l'incertezza
- Da questo punto di vista uno stimatore $\hat{\vartheta}_N$ è buono se

$$\lim_{N \rightarrow \infty} \text{var} [\hat{\vartheta}_N] = 0$$



- Un'altra caratterizzazione della bontà di uno stimatore $\hat{\vartheta}_N$ in cui la stima viene determinata in base ad un numero N di dati crescente nel tempo è

$$\lim_{N \rightarrow \infty} E \left[\|\hat{\vartheta}_N - \vartheta^0\|^2 \right] = 0 \quad (*)$$

Se vale la (*) si dice che la stima $\hat{\vartheta}_N$ converge in media quadratica a ϑ^0

- Osserviamo che $\hat{\vartheta}_N$ è un vettore casuale, ϑ^0 è un vettore costante e $\|\hat{\vartheta}_N - \vartheta^0\|^2$ è una variabile aleatoria scalare per cui l'operazione "valore atteso" è perfettamente legittima

Convergenza quasi certa

- Ricordiamo che lo stimatore basato su N dati è

$$\hat{\vartheta}_N(s, \vartheta^0) = f[d(s, \vartheta^0)]$$

- Fissato $\bar{s} \in S$, si avrà una sequenza

$$\hat{\vartheta}_1(\bar{s}, \vartheta^0), \hat{\vartheta}_2(\bar{s}, \vartheta^0), \dots, \hat{\vartheta}_N(\bar{s}, \vartheta^0), \dots$$

- Potrebbe allora succedere che:

$$\bar{s} \in S \quad \longrightarrow \quad \lim_{N \rightarrow \infty} \hat{\vartheta}_N(\bar{s}, \vartheta^0) = \vartheta^0$$

$$\bar{s} \in S \quad \longrightarrow \quad \lim_{N \rightarrow \infty} \hat{\vartheta}_N(\bar{s}, \vartheta^0) \neq \vartheta^0$$

- Introduciamo l'insieme di esiti

$$A \subset S, \quad A = \left\{ s \in S : \lim_{N \rightarrow \infty} \hat{\vartheta}_N(s, \vartheta^\circ) = \vartheta^\circ \right\}$$

- Se $A = S$ \rightarrow **Convergenza certa**
- Se $A \subset S$ e $P(A) = 1$ \rightarrow **Convergenza quasi-certa**

Notiamo che se la misura dell'insieme $S \setminus A$ è nulla cioè implica $P(A) = 1$ e quindi convergenza quasi-certa

- Evidentemente $A = S \rightarrow P(A) = 1$
 \rightarrow **Convergenza certa** \rightarrow **Convergenza quasi-certa**
- Uno stimatore per cui si abbia convergenza quasi-certa si dice **consistente**

25/01/2012

25

Esempio 1

- Si considerino N dati scalari $d(1), d(2), \dots, d(N)$ tali che

$$E[d(i)] = \vartheta^\circ, \quad i = 1, \dots, N$$

- Si supponga che i dati siano mutuamente scorrelati, cioè`

$$E\{[d(i) - \vartheta^\circ][d(j) - \vartheta^\circ]\} = 0, \quad \forall i \neq j$$

- Si consideri lo stimatore

$$\hat{\vartheta}_N = \frac{1}{N} \sum_{i=1}^N d(i) \quad \text{Stimatore a media campionaria}$$

25/01/2012

26

- Polarizzazione:

$$E[\hat{\vartheta}_N] = E\left\{ \frac{1}{N} \sum_{i=1}^N d(i) \right\} = \frac{1}{N} \sum_{i=1}^N E[d(i)] = \frac{1}{N} \sum_{i=1}^N \vartheta^\circ = \vartheta^\circ$$

\rightarrow lo stimatore non è polarizzato

- Varianza:

$$\begin{aligned} \text{var}(\hat{\vartheta}_N) &= E\left\{ [\hat{\vartheta}_N - E(\hat{\vartheta}_N)]^2 \right\} = E\left\{ \left[\frac{1}{N} \sum_{i=1}^N d(i) - \frac{1}{N} \sum_{i=1}^N \vartheta^\circ \right]^2 \right\} \\ &= E\left\{ \frac{1}{N^2} \left[\sum_{i=1}^N d(i) - \sum_{i=1}^N \vartheta^\circ \right]^2 \right\} = \frac{1}{N^2} \sum_{i=1}^N E\{[d(i) - \vartheta^\circ]^2\} \\ &= \frac{1}{N^2} \sum_{i=1}^N \text{var}[d(i)] \quad (\text{i termini "misti" sono nulli per l'ipotesi di dati mutuamente scorrelati}) \end{aligned}$$

Se $\text{var}[d(i)] \leq \sigma, \quad i = 1, \dots, N$

$\rightarrow \lim_{N \rightarrow \infty} \text{var}(\hat{\vartheta}_N) \leq \lim_{N \rightarrow \infty} \frac{\sigma}{N} = 0 \rightarrow$ lo stimatore converge in media quadratica

25/01/2012

27

Esempio 2

- Si considerino N dati scalari $d(1), d(2), \dots, d(N)$ tali che

$$E[d(i)] = \vartheta^\circ, \quad i = 1, \dots, N$$

- Si supponga che i dati siano mutuamente scorrelati, cioè`

$$E\{[d(i) - \vartheta^\circ][d(j) - \vartheta^\circ]\} = 0, \quad \forall i \neq j$$

- Si consideri lo stimatore

$$\hat{\vartheta}_N = \sum_{i=1}^N \alpha(i) d(i)$$

25/01/2012

28

- Polarizzazione:

$$E[\hat{\vartheta}_N] = E\left\{\sum_{i=1}^N \alpha(i) d(i)\right\} = \sum_{i=1}^N \alpha(i) E[d(i)] = \vartheta^0 \sum_{i=1}^N \alpha(i)$$

↳ lo stimatore non è polarizzato ↔ $\sum_{i=1}^N \alpha(i) = 1$ (*)

N.B. Nel caso precedente $\alpha(i) = \frac{1}{N}$ per cui (*) è soddisfatta

La condizione (*) è un vincolo da soddisfare affinché lo stimatore non sia polarizzato e caratterizza una classe di infiniti stimatori possibili

- Determiniamo lo stimatore migliore tra quelli non polarizzati (quindi che soddisfano il vincolo (*)) andando a scegliere quello di minima varianza

$$\begin{cases} \min \text{var}(\hat{\vartheta}_N) = \min \sum_{i=1}^N [\alpha(i)]^2 \text{var}[d(i)] \\ 1 - \sum_{i=1}^N \alpha(i) = 0 \end{cases}$$

dati scorrelati

Usando il metodo dei moltiplicatori di Lagrange

$$J(\hat{\vartheta}) = \sum_{i=1}^N [\alpha(i)]^2 \text{var}[d(i)] + \lambda \left(1 - \sum_{i=1}^N \alpha(i)\right)$$

↳ $\frac{\partial J}{\partial \alpha(i)} = 0$ ↔ $2\alpha(i)\text{var}[d(i)] - \lambda = 0$ ↔ $\alpha(i) = \frac{\lambda}{2\text{var}[d(i)]}$

- Imponendo il vincolo di non polarizzazione (*)

$$\sum_{i=1}^N \alpha(i) = 1 \quad \leftrightarrow \quad \frac{\lambda}{2} \sum_{i=1}^N \frac{1}{\text{var}[d(i)]} = 1 \quad \leftrightarrow \quad \lambda = \frac{2}{\sum_{i=1}^N \frac{1}{\text{var}[d(i)]}}$$

↳ $\alpha(i) = \frac{1}{\text{var}[d(i)]} \alpha$ con $\alpha = \frac{1}{\sum_{i=1}^N \frac{1}{\text{var}[d(i)]}}$

Quindi $\alpha(i)$ viene scelto inversamente proporzionale alla varianza del dato $\text{var}[d(i)]$; più è grande e meno peso gli viene associato.

- Calcoliamo ora la varianza dello stimatore:

$$\begin{aligned} \text{var}(\hat{\vartheta}_N) &= E\left\{[\hat{\vartheta}_N - E(\hat{\vartheta}_N)]^2\right\} = E\left\{\left[\sum_{i=1}^N \alpha(i) d(i) - \vartheta^0 \sum_{i=1}^N \alpha(i)\right]^2\right\} \\ &= E\left\{\left[\sum_{i=1}^N \alpha(i) [d(i) - \vartheta^0]\right]^2\right\} = \sum_{i=1}^N \alpha(i)^2 E\{[d(i) - \vartheta^0]^2\} \\ &= \sum_{i=1}^N \alpha(i)^2 \text{var}[d(i)] = \alpha^2 \sum_{i=1}^N \frac{1}{\text{var}[d(i)]} = \frac{1}{\sum_{i=1}^N \frac{1}{\text{var}[d(i)]}} \end{aligned}$$

Se $\text{var}[d(i)] \leq \sigma$, $i = 1, \dots, N$

↳ $\lim_{N \rightarrow \infty} \text{var}(\hat{\vartheta}_N) \leq \lim_{N \rightarrow \infty} \frac{\sigma}{N} = 0$ → lo stimatore converge in media quadratica

Generalizzazione

- Nel caso in cui le quantità da stimare siano **varianti nel tempo** e` necessario modificare gli indici di bontà degli stimatori
- Si indichi con $\hat{\vartheta}(t|t-1)$ la stima di $\vartheta^\circ(t)$ in base ai dati fino all'istante $t-1$
- Evidentemente, siccome $\vartheta^\circ(t)$ varia nel tempo, non e` sensato parlare di convergenza asintotica rispetto alla disponibilita` di dati in quanto i dati nel passato possono non essere piu` significativi.

- Un criterio tipico e`

$$E \left[\left\| \hat{\vartheta}(t|t-1) - \vartheta^\circ(t) \right\|^2 \right] \leq c$$

dove c e` una costante ragionevolmente piccola

- **Non si chiede convergenza ma si chiede "non divergenza"**

25/01/2012

33

Stima ai Minimi Quadrati (LSM)

25/01/2012

34

Regressione lineare

- Si tratta del tipico contesto in cui si usa lo stimatore ai minimi quadrati (MQ)
- Si hanno $q+1$ variabili $y(t), u_1(t), \dots, u_q(t)$ sull'arco temporale $t = 1, 2, \dots, N$
- Si vogliono calcolare (se possibile) q parametri $\vartheta_1, \vartheta_2, \dots, \vartheta_q$ tali che

$$y(t) = \vartheta_1 u_1(t) + \dots + \vartheta_q u_q(t), \quad t = 1, \dots, N \quad (*)$$

(*) viene definita come la regressione lineare della variabile $y(t)$ sulle variabili $u_1(t), \dots, u_q(t)$

25/01/2012

35

- Il problema puo` essere formulato in maniera vettoriale

$$\vartheta = \begin{bmatrix} \vartheta_1 \\ \vdots \\ \vartheta_q \end{bmatrix} \quad \varphi(t) = \begin{bmatrix} u_1(t) \\ \vdots \\ u_q(t) \end{bmatrix}$$

 $y(t) = \varphi(t)^\top \vartheta$

- Nei problemi reali in realta` si avra` un certo errore

$$\varepsilon(t) = y(t) - \varphi(t)^\top \vartheta$$

25/01/2012

36

- Obiettivo del problema delle regressione lineare sara` quindi quello di minimizzare l'errore $\varepsilon(t)$ determinando un vettore ϑ^o per cui questo minimo venga raggiunto
- Si definisce la **funzione di costo quadratica**

$$J(\vartheta) = \sum_{t=1}^N [\varepsilon(t)]^2 = \sum_{t=1}^N [y(t) - \varphi(t)^\top \vartheta]^2$$

↳ $\vartheta^o = \arg \min_{\vartheta} J(\vartheta)$ **Stimatore ai minimi quadrati**

- Indicando con ϑ_i la componente i-esima del vettore ϑ

↳
$$\frac{\partial J}{\partial \vartheta_i} = \frac{\partial}{\partial \vartheta_i} \left\{ \sum_{t=1}^N [y(t) - \varphi(t)^\top \vartheta]^2 \right\}$$

$$= -2 \sum_{t=1}^N [y(t) - \varphi(t)^\top \vartheta] \varphi_i(t), \quad i = 1, 2, \dots, q$$

ed osservando che
$$\frac{\partial J}{\partial \vartheta} = \left[\frac{\partial J}{\partial \vartheta_1} \quad \frac{\partial J}{\partial \vartheta_2} \quad \dots \quad \frac{\partial J}{\partial \vartheta_q} \right]$$

↳
$$\frac{\partial J}{\partial \vartheta} = -2 \sum_{t=1}^N [y(t) - \varphi(t)^\top \vartheta] \varphi(t)^\top$$

- Imponendo $\frac{\partial J}{\partial \vartheta} = [0 \ 0 \ \dots \ 0]$

↳
$$-2 \sum_{t=1}^N [y(t) - \varphi(t)^\top \vartheta] \varphi(t)^\top = [0 \ 0 \ \dots \ 0]$$

↳
$$\sum_{t=1}^N y(t) \varphi(t)^\top = \sum_{t=1}^N \varphi(t)^\top \vartheta \varphi(t)^\top$$

convertendo l'uguaglianza tra vettori riga in un'uguaglianza tra vettori colonna si puo` mettere in evidenza il vettore ϑ :

$$\sum_{t=1}^N \varphi(t) y(t) = \left[\sum_{t=1}^N \varphi(t) \varphi(t)^\top \right] \vartheta$$

Equazioni normali dei minimi quadrati (q eqz. in q incognite)

- Se $\sum_{t=1}^N \varphi(t) \varphi(t)^\top$ e` non singolare

↳
$$\hat{\vartheta}_N = \left[\sum_{t=1}^N \varphi(t) \varphi(t)^\top \right]^{-1} \sum_{t=1}^N \varphi(t) y(t)$$

Formula dei minimi quadrati

Interpretazione geometrica delle eq. normali

Poniamo

$$\varepsilon_{\vartheta}^N = \begin{bmatrix} \varepsilon_{\vartheta}(1) \\ \vdots \\ \varepsilon_{\vartheta}(N) \end{bmatrix} \quad y^N = \begin{bmatrix} y(1) \\ \vdots \\ y(N) \end{bmatrix} \quad \Phi_N = \begin{bmatrix} \varphi(1)^\top \\ \vdots \\ \varphi(N)^\top \end{bmatrix}$$

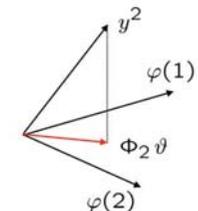
Scriviamo quindi

$$J(\vartheta) = \sum_{t=1}^N [y(t) - \varphi(t)^\top \vartheta]^2 = \|y^N - \Phi_N \vartheta\|^2$$

Evidentemente $\|y^N - \Phi_N \vartheta\|$ e`

minima quando $y^N - \Phi_N \vartheta$ e`

ortogonale a $\Phi_N \vartheta$



- Verifichiamo che $\hat{\vartheta}_N$ sia un **minimo** valutando la definitezza della matrice simmetrica

$$\left[\frac{d^2 J}{d\vartheta^2} \right]_{i,j} = \frac{\partial^2 J}{\partial \vartheta_i \partial \vartheta_j}, \quad i, j = 1, \dots, q$$

si ha

$$\left(\frac{\partial J}{\partial \vartheta} \right)^\top = 2 \left\{ \left[\sum_{t=1}^N \varphi(t) \varphi(t)^\top \right] \vartheta - \sum_{t=1}^N \varphi(t) y(t) \right\}$$

$$\rightarrow \frac{d^2 J}{d\vartheta^2} = 2 \left[\sum_{t=1}^N \varphi(t) \varphi(t)^\top \right]$$

evidentemente si tratta di una matrice simmetrica e semidefinita positiva

$$\rightarrow \hat{\vartheta}_N \text{ e' un minimo locale di } J(\vartheta)$$

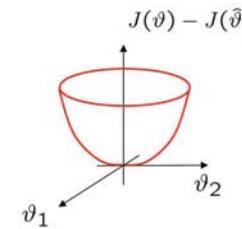
25/01/2012

41

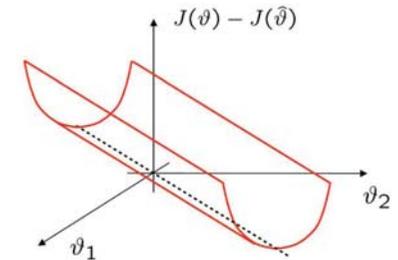
- Considerando quindi la forma quadratica

$$J(\vartheta) - J(\hat{\vartheta}) = \frac{1}{2} (\vartheta - \hat{\vartheta})^\top \frac{d^2 J}{d\vartheta^2} \Big|_{\hat{\vartheta}} (\vartheta - \hat{\vartheta})$$

si hanno i due casi possibili:



$$\det \left[\sum_{t=1}^N \varphi(t) \varphi(t)^\top \right] \neq 0$$



$$\det \left[\sum_{t=1}^N \varphi(t) \varphi(t)^\top \right] = 0$$

25/01/2012

42

- Quindi:

$$\circ \text{ Se } \det \left[\sum_{t=1}^N \varphi(t) \varphi(t)^\top \right] \neq 0 \rightarrow \hat{\vartheta}_N \text{ unico minimo globale}$$

$$\circ \text{ Se } \det \left[\sum_{t=1}^N \varphi(t) \varphi(t)^\top \right] = 0 \rightarrow \hat{\vartheta}_N \text{ e' uno degli infiniti minimi globali}$$

- La condizione $\det \left[\sum_{t=1}^N \varphi(t) \varphi(t)^\top \right] \neq 0$ e' detta

e' detta **condizione di identificabilita'**

25/01/2012

43

Caratteristiche probabilistiche dello stimatore MQ

- Supponiamo che la condizione di identificabilita' sia verificata:

$$\det \left[\sum_{t=1}^N \varphi(t) \varphi(t)^\top \right] \neq 0$$

per cui

$$\hat{\vartheta}_N = \left[\sum_{t=1}^N \varphi(t) \varphi(t)^\top \right]^{-1} \sum_{t=1}^N \varphi(t) y(t)$$

- Ipotesi:** $y(t) = \varphi(t)^\top \vartheta^\circ + \xi(t)$

dove il processo e' scorrelato da $u(\cdot)$ e $E[\xi(t)] = 0$

Si sta quindi ipotizzando che il legame vero tra $y(t)$ e $u_1(t), \dots, u_q(t)$ sia proprio lineare + rumore scorrelato e a media nulla

25/01/2012

44

Polarizzazione

$$\begin{aligned}\hat{\vartheta}_N &= \left[\sum_{t=1}^N \varphi(t) \varphi(t)^\top \right]^{-1} \sum_{t=1}^N \varphi(t) y(t) \\ &= \left[\sum_{t=1}^N \varphi(t) \varphi(t)^\top \right]^{-1} \sum_{t=1}^N \varphi(t) [\varphi(t)^\top \vartheta^0 + \xi(t)] \\ &= \vartheta^0 + \left[\sum_{t=1}^N \varphi(t) \varphi(t)^\top \right]^{-1} \sum_{t=1}^N \varphi(t) \xi(t)\end{aligned}$$

$$\hookrightarrow \hat{\vartheta}_N - \vartheta^0 = \left[\sum_{t=1}^N \varphi(t) \varphi(t)^\top \right]^{-1} \sum_{t=1}^N \varphi(t) \xi(t)$$

$$\hookrightarrow E(\hat{\vartheta}_N - \vartheta^0) = \left[\sum_{t=1}^N \varphi(t) \varphi(t)^\top \right]^{-1} \sum_{t=1}^N \varphi(t) E[\xi(t)] = 0$$

$$\hookrightarrow E(\hat{\vartheta}_N) = \vartheta^0 \quad \text{Stimatore MQ e' non polarizzato}$$

25/01/2012

45

OSSERVAZIONE IMPORTANTE

Nell'analisi di polarizzazione dello stimatore ai MQ abbiamo considerato il vettore dei regressori $\varphi(t)$ come **noto e fissato**. Se invece l'analisi fosse stata effettuata considerando il vettore $\varphi(s, t)$ come **vettore casuale** (e quindi funzione dell'esito s di un esperimento casuale) allora si otterrebbe che lo stimatore ai MQ e' polarizzato per qualunque valore finito di

$$E[\hat{\vartheta}_N - \vartheta^0] = E \left\{ \left[\sum_{t=1}^N \varphi(t) \varphi(t)^\top \right]^{-1} \sum_{t=1}^N \varphi(t) \xi(t) \right\}$$

25/01/2012

46

Varianza

Ulteriore ipotesi: $\xi(t) \sim WN(0, \lambda^2)$

Introduciamo la matrice simmetrica $S(N) = \sum_{t=1}^N \varphi(t) \varphi(t)^\top$

$$\begin{aligned}\hookrightarrow \text{var}(\hat{\vartheta}_N) &= E[(\hat{\vartheta}_N - \vartheta^0)(\hat{\vartheta}_N - \vartheta^0)^\top] \\ &= E \left\{ \left[S(N)^{-1} \sum_{t=1}^N \varphi(t) \xi(t) \right] \left[S(N)^{-1} \sum_{s=1}^N \varphi(s) \xi(s) \right]^\top \right\} \\ &= E \left\{ \left[S(N)^{-1} \sum_{t=1}^N \varphi(t) \xi(t) \right] \left[\sum_{s=1}^N \xi(s) \varphi(s)^\top S(N)^{-1} \right] \right\} \\ &= S(N)^{-1} E \left[\sum_{t=1}^N \varphi(t) \xi(t) \sum_{s=1}^N \xi(s) \varphi(s)^\top \right] S(N)^{-1}\end{aligned}$$

$S(N)$ simmetrica

25/01/2012

47

Nel prodotto $\sum_{t=1}^N \varphi(t) \xi(t) \sum_{s=1}^N \xi(s) \varphi(s)^\top$ si hanno due tipi di termini:

- $\varphi(t) \xi(t)^2 \varphi(t)^\top$ se $t = s$
- $\varphi(t) \xi(t) \xi(s) \varphi(s)^\top$ se $t \neq s$

Ma $\xi(t) \sim WN(0, \lambda^2) \rightarrow E[\xi(t)\xi(s)] = \begin{cases} \lambda^2 & \text{se } t = s \\ 0 & \text{se } t \neq s \end{cases}$

$$\hookrightarrow E \left[\sum_{t=1}^N \varphi(t) \xi(t) \sum_{s=1}^N \xi(s) \varphi(s)^\top \right] = \sum_{t=1}^N \lambda^2 \varphi(t) \varphi(t)^\top = \lambda^2 S(N)$$

$$\hookrightarrow \text{var}(\hat{\vartheta}_N) = S(N)^{-1} \lambda^2 S(N) S(N)^{-1} = \lambda^2 S(N)^{-1}$$

25/01/2012

48

Interpretazione

Supponiamo che ϑ° sia scalare per cui anche $\varphi(t)$ e $\xi(t)$ sono scalari

$$\rightarrow y(t) = \varphi(t) \vartheta^\circ + \xi(t) = u(t) \vartheta^\circ + \xi(t)$$

da cui

$$\hat{\vartheta}_N = \left[\sum_{t=1}^N \varphi(t) \varphi(t)^\top \right]^{-1} \sum_{t=1}^N \varphi(t) y(t) = \frac{\frac{1}{N} \sum_{t=1}^N u(t) y(t)}{\frac{1}{N} \sum_{t=1}^N u(t)^2}$$

Ma:

- $\frac{1}{N} \sum_{t=1}^N u(t) y(t)$ stima campionaria della cross-correlazione $E[u(t) y(t)]$
- $\frac{1}{N} \sum_{t=1}^N u(t)^2$ stima campionaria di $E[u(t)^2]$ (varianza se $E(u) = 0$)

25/01/2012

49

Inoltre

$$\text{var}(\hat{\vartheta}_N) = \lambda^2 S(N)^{-1} = \frac{1}{N} \frac{\lambda^2}{\frac{1}{N} \sum_{t=1}^N u(t)^2}$$

Pertanto:

- $\text{var}(\hat{\vartheta}_N)$ cresce al crescere di λ^2 \rightarrow L'incertezza sulla stima aumenta all'aumentare dell'incertezza sui dati
- a parità di N e λ^2 , $\text{var}(\hat{\vartheta}_N)$ diminuisce all'aumentare della varianza campionaria di u il che è intuitivamente corretto in quanto il rumore viene "sopraffatto" dal segnale contenente informazione utile

25/01/2012

50

- $\frac{\lambda^2}{\frac{1}{N} \sum_{t=1}^N u(t)^2}$ è una sorta di rapporto rumore/segnale

- Se la varianza di u è limitata $\rightarrow \lim_{N \rightarrow \infty} \text{var}(\hat{\vartheta}_N) = 0$ e siccome lo stimatore non è polarizzato si ha:

$$\lim_{N \rightarrow \infty} E \left(\|\hat{\vartheta}_N - \vartheta^\circ\|^2 \right) = 0$$

ovvero lo stimatore MQ converge in media quadratica

25/01/2012

51

Inoltre si può scrivere

$$\begin{aligned} \hat{\vartheta}_N &= \frac{1}{\sum_{t=1}^N u(t)^2} \sum_{t=1}^N u(t) [u(t) \vartheta^\circ + \xi(t)] \\ &= \vartheta^\circ + \frac{\frac{1}{N} \sum_{t=1}^N u(t) \xi(t)}{\frac{1}{N} \sum_{t=1}^N u(t)^2} \xrightarrow{N \rightarrow \infty} \vartheta^\circ + \frac{E[u(t) \xi(t)]}{E[u(t)^2]} \end{aligned}$$

- Se u è deterministico

$$\rightarrow \vartheta^\circ + \frac{E[u(t) \xi(t)]}{E[u(t)^2]} = \vartheta^\circ + u(t) \frac{E[\xi(t)]}{E[u(t)^2]} = \vartheta^\circ$$

- Se u è stocastico ma scorrelato da ξ

$$\rightarrow \vartheta^\circ + \frac{E[u(t) \xi(t)]}{E[u(t)^2]} = \vartheta^\circ + \frac{E[u(t)] E[\xi(t)]}{E[u(t)^2]} = \vartheta^\circ$$

da cui si evince tra l'altro l'importanza di avere i dati scorrelati dal rumore

25/01/2012

52

Scelta della Complessità dei Modelli

Premessa: verifica di bianchezza (test di Anderson)

- A procedura di identificazione ultimata bisogna verificare che l'errore di predizione sia assimilabile il più possibile ad un **processo bianco**.
- Sia dato un processo stazionario $\varepsilon(\cdot)$ a media nulla e si consideri la funzione di covarianza campionaria:

$$\hat{\gamma}(\tau) = \frac{1}{N} \sum_{t=1}^{N-\tau} \varepsilon(t) \varepsilon(t + \tau)$$

dove N è l'ampiezza della finestra temporale considerata.

- Nel test di Anderson si utilizza la funzione di covarianza campionaria **normalizzata**:

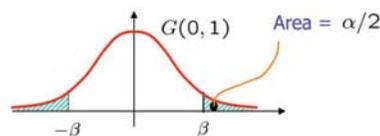
$$\hat{\rho}(\tau) = \frac{\hat{\gamma}(\tau)}{\hat{\gamma}(0)}$$

- Si dimostra che se $\varepsilon(\cdot)$ è bianco $\rightarrow \sqrt{N} \hat{\rho}(\tau) \sim \text{As}G(0, 1)$ e che $\hat{\rho}(i)$ è asintoticamente scorrelato da $\hat{\rho}(j)$, $i \neq j$

- Il test di Anderson si effettua così:

- Si fissa un livello di confidenza $0 < \alpha < 1$ (per esempio $\alpha = 0.01$)

- Si determina β tale che le code della Gaussiana $G(0, 1)$ negli intervalli $(-\infty, -\beta)$ e (β, ∞) abbiano area $\alpha/2$



- Si considera un certo numero M di valutazioni di $\hat{\rho}(\tau)$: $\hat{\rho}(0), \hat{\rho}(1), \hat{\rho}(2), \dots, \hat{\rho}(M)$

- Si considera l'intervallo $(-\beta/\sqrt{N}, \beta/\sqrt{N})$ e si valuta il numero n di campioni di $\hat{\rho}(\tau)$ per cui $\hat{\rho}(\tau) \notin (-\beta/\sqrt{N}, \beta/\sqrt{N})$

- Se $\frac{n}{M} < \alpha \rightarrow \varepsilon(\cdot)$ bianco

Complessità dei modelli

- Caratterizziamo la complessità del modello (a famiglia fissata) col numero n dei suoi parametri
- Consideriamo il criterio quadratico

$$J(\vartheta) = \frac{1}{N} \sum_{i=1}^N [\varepsilon(t)]^2$$

dove ϑ è il vettore dei parametri incogniti, $n = \dim(\vartheta)$ e $\varepsilon(t)$ è l'errore di predizione all'istante t : $\varepsilon(t) = y(t) - \hat{y}(t | t-1)$

- Si consideri $\hat{\vartheta}_N = \arg \min_{\vartheta} J(\vartheta)$
- $J(\hat{\vartheta}_N)$ può essere considerato un **indice di aderenza del modello ai dati**
- Tuttavia, **a parità di realizzazione dei dati**, $J(\hat{\vartheta}_N)$ diminuisce al crescere della complessità $n \rightarrow J(\hat{\vartheta}_N)$ non è utile di per sé per determinare la complessità ottima del modello

Esempio

Consideriamo il processo (sistema vero):

$$S: \quad y(t) = 1.2y(t-1) - 0.32y(t-2) + u(t-1) + 0.5u(t-2) + e(t)$$

$$e(\cdot) \sim WN(0, 1), \quad u(\cdot) \sim WN(0, 4), \quad e(\cdot), u(\cdot) \text{ scorrelati}$$

Consideriamo poi la famiglia di modelli ARX(n,n):

$$\mathcal{M}(\vartheta): \quad y(t) = a_1 y(t-1) + \dots + a_n y(t-n) + \xi(t)$$

$$+ b_1 u(t-1) + \dots + b_n u(t-n) + \xi(t)$$

ed identifichiamo i modelli nei casi $n = 1, 2, 3$ su una finestra di 2000 dati, ovvero $\{u(t), y(t)\}_{t=1, \dots, 2000}$

25/01/2012

57

ARX(1,1)	$\hat{a} = 0.932 \quad (0.6\%)$ $\hat{b} = 0.975 \quad (2.3\%)$ $J = 3.864$ T.And. 5% : 7		
ARX(2,2)	$\hat{a}_1 = 1.204 \quad (1\%)$ $\hat{b}_1 = 0.984 \quad (1\%)$ $J = 0.998$ T.And. 5% : 0 (OK)	$\hat{a}_2 = -0.32 \quad (3\%)$ $\hat{b}_2 = 0.485 \quad (3\%)$	
ARX(3,3)	$\hat{a}_1 = 1.194 \quad (2\%)$ $\hat{b}_1 = 0.984 \quad (1\%)$ $J = 0.997$ T.And. 5% : 0 (OK)	$\hat{a}_2 = -0.299 \quad (10\%)$ $\hat{b}_2 = 0.494 \quad (5\%)$	$\hat{a}_3 = -0.019 \quad (68\%)$ $\hat{b}_3 = -0.016 \quad (120\%)$

25/01/2012

58

- Osserviamo che $J(\hat{\vartheta}_{2000})$ decresce al crescere di n
- Il test di Anderson fornisce risultati che migliorano al crescere di n
- Per $n \geq 3$ la stima dei parametri \hat{a}_n e \hat{b}_n e' molto piccola e l'incertezza sulla stima dei coefficienti e' molto grande indicando chiaramente una **sovraparametrizzazione (modello troppo complesso rispetto ai dati a disposizione)**

 ARX(2,2) e' il modello corretto

25/01/2012

59

Considerazioni:

- In generale il test di A. puo' non essere soddisfatto anche per grandi valori di n nel qual caso non e' possibile pervenire ad una scelta chiara ed univoca come nell'esempio.
- Il fatto che, a parita' di realizzazione dei dati, $J(\hat{\vartheta}_N)$ diminuisca al crescere della complessita' n - di fatto impedendo di utilizzare $J(\hat{\vartheta}_N)$ per determinare la complessita' ottima del modello - e' conseguenza di un **errore concettuale**:

utilizzare gli stessi dati per identificare il modello e per validarlo

 $J(\hat{\vartheta}_N)$ non puo' costituire un indicatore oggettivo per valutare la bonta' del modello identificato

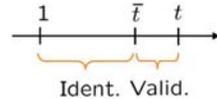
E' necessario validare il modello su dati diversi da quelli utilizzati per identificarlo

25/01/2012

60

Cross-validazione

- Si supponga di disporre di un numero N di dati sufficientemente elevato
- Si usa **una parte dei dati** per validare il modello identificato con i dati restanti



Quindi:

$$J_{CV}(\vartheta) = \frac{1}{N - \bar{t}} \sum_{k=\bar{t}}^N [\varepsilon(k)]^2$$

ed ora si valuta n per cui $J_{CV}(\vartheta)$ è minimo

- Ora, **a parità di realizzazione dei dati**, $J_{CV}(\hat{\vartheta}_{\bar{t}})$ non è monotona decrescente al crescere della complessità n → $J_{CV}(\hat{\vartheta}_{\bar{t}})$ può essere utilizzato come criterio oggettivo per determinare la complessità ottima del modello
- La procedura di CV è però onerosa e spesso non si dispone di un numero sufficiente di dati per poterla applicare correttamente.

25/01/2012

61

Final prediction error (FPE)

- Costruiamo ora un criterio in cui si cerchi di valutare la bontà del modello rispetto alle diverse possibili realizzazioni dei dati:

$$\bar{J}(\vartheta) = E \{ [y(t, s) - \hat{y}(t, s, \vartheta)]^2 \}$$

dove s rappresenta l'esito dell'esperimento casuale con cui vengono estratti i dati

- Quindi $\bar{J}(\vartheta)$ esprime l'**aderenza media** del modello la cui parametrizzazione è ϑ a tutte le possibili sequenze di dati
- Come sempre $\hat{\vartheta}_N = \arg \min_{\vartheta} J(\vartheta)$
- in cui la minimizzazione è effettuata in corrispondenza di una specifica sequenza di dati. Al variare dell'esperimento casuale si ha $\hat{\vartheta}_N = \hat{\vartheta}_N(s)$
- Mediando ancora si definisce

$$FPE = E \{ \bar{J}[\hat{\vartheta}_N(s)] \}$$

e la complessità ottima è quella per cui FPE risulta minimizzato

25/01/2012

62

Valutiamo FPE in un caso particolare:

$$\mathcal{S}: AR(n) \quad \mathcal{M}: AR(n)$$

Quindi:

$$\mathcal{S}: y(t, s) = \varphi(t, s)^T \vartheta^\circ + \xi(t) \quad \xi(\cdot) \sim WN(0, \lambda^2)$$

$$\widehat{\mathcal{M}}(\vartheta): \hat{y}(t, s) = \varphi(t, s)^T \vartheta$$

Ma $\varphi(t, s)$ e $\xi(t)$ sono scorrelati da cui

$$\bar{J}(\vartheta) = E \{ [y(t, s) - \hat{y}(t, s, \vartheta)]^2 \} = E \{ [\varphi(t, s)^T (\vartheta^\circ - \vartheta) + \xi(t)]^2 \}$$

$$= (\vartheta^\circ - \vartheta)^T E [\varphi(t, s) \varphi(t, s)^T] (\vartheta^\circ - \vartheta) + \lambda^2$$

Ponendo $\bar{R} = E [\varphi(t, s) \varphi(t, s)^T]$ si ha

$$\bar{J}(\vartheta) = (\vartheta^\circ - \vartheta)^T \bar{R} (\vartheta^\circ - \vartheta) + \lambda^2$$

per cui, dalla definizione consegue che

$$FPE = E \{ \bar{J}[\hat{\vartheta}_N(s)] \} = E \{ [\vartheta^\circ - \hat{\vartheta}_N(s)]^T \bar{R} [\vartheta^\circ - \hat{\vartheta}_N(s)] + \lambda^2 \}$$

25/01/2012

63

D'altra parte, per N sufficientemente elevato:

$$\text{var} [\vartheta^\circ - \hat{\vartheta}_N(s)] \sim \frac{\lambda^2}{N} \bar{R}^{-1}$$

Ponendo ora $\nu = \vartheta^\circ - \hat{\vartheta}_N(s)$ si ha

$$\text{var}(\nu) = \frac{\lambda^2}{N} \bar{R}^{-1} \quad \rightarrow \quad \bar{R} = \text{var}(\nu)^{-1} \frac{\lambda^2}{N}$$

e quindi

$$FPE = E (\nu^T \bar{R} \nu) + \lambda^2 = E [\nu^T \text{var}(\nu)^{-1} \nu] \frac{\lambda^2}{N} + \lambda^2$$

Ma $\nu^T \text{var}(\nu)^{-1} \nu$ è uno scalare per cui coincide con la sua traccia:

$$\nu^T \text{var}(\nu)^{-1} \nu = \text{tr} [\nu^T \text{var}(\nu)^{-1} \nu]$$

Infine $\text{tr}(AB) = \text{tr}(BA)$ purché AB e BA abbiano senso.

25/01/2012

64

Pertanto:

$$\begin{aligned}
 E \left[\nu^\top \text{var}(\nu)^{-1} \nu \right] &= E \left\{ \text{tr} \left[\nu^\top \text{var}(\nu)^{-1} \nu \right] \right\} \\
 &= E \left\{ \text{tr} \left[\text{var}(\nu)^{-1} \nu \nu^\top \right] \right\} \\
 &= \text{tr} \left\{ E \left[\text{var}(\nu)^{-1} \nu \nu^\top \right] \right\} \\
 &= \text{tr} \left[\text{var}(\nu)^{-1} E \left(\nu \nu^\top \right) \right] \\
 &= \text{tr} \left[\text{var}(\nu)^{-1} \text{var}(\nu) \right] \\
 &= \text{tr} (I) = n
 \end{aligned}$$

e quindi

$$\text{FPE} = \frac{n}{N} \lambda^2 + \lambda^2$$

Si può dimostrare che per N suff. grande, una stima di λ^2 è data da

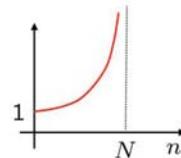
$$\hat{\lambda}^2 = \frac{1}{N-n} \sum_{t=1}^N [\varepsilon(t)]^2 = \frac{N}{N-n} \frac{1}{N} \sum_{t=1}^N [\varepsilon(t)]^2 = \frac{N}{N-n} J(\hat{\vartheta}_N)^{(n)}$$

dove $J(\hat{\vartheta}_N)^{(n)}$ rappresenta il costo puntuale sui dati in corrispondenza del modello di complessità n

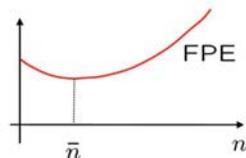
$$\text{FPE} = \frac{N+n}{N-n} J(\hat{\vartheta}_N)^{(n)}$$

Considerazioni:

- La funzione $\frac{N+n}{N-n}$ ha l'andamento



mentre la funzione $J(\hat{\vartheta}_N)^{(n)}$ è monotona decrescente rispetto a n per cui, per un valore fissato di N , l'andamento tipico del FPE è:



La complessità ottima rispetto al criterio FPE è \bar{n}

- La formula per FPE vale anche per le altre tipologie di modelli pur di ridefinire opportunamente n . Per esempio nel caso ARX si pone $n = n_a + n_b$ mentre nel caso ARMAX si pone $n = n_a + n_b + n_c$

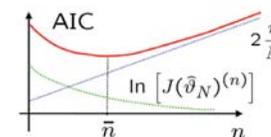
Akaike information criterion (AIC)

- Si tratta di un criterio di scelta della complessità del modello di tipo **statistico**. Viene determinato minimizzando la **distanza di Kullback** tra la densità di probabilità dei dati osservati e quella che produrrebbe il modello in esame dove si definisce distanza di Kullback la quantità

$$E \left(\ln \frac{p_{\text{vera}}}{p_{\text{modello}}} \right)$$

- Si dimostra

$$\text{AIC} = 2 \frac{n}{N} + \ln [J(\hat{\vartheta}_N)^{(n)}]$$



La complessità ottima rispetto al criterio FPE è \bar{n}

Osserviamo che la velocità di crescita della retta $2 \frac{n}{N}$ diminuisce all'aumentare di N → AIC tende a privilegiare modelli di ordine minore quando i dati a disposizione sono pochi.

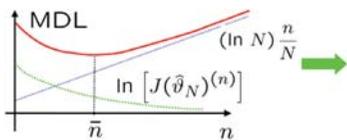
Minimum description length (MDL)

- Si tratta di un criterio di scelta della complessita` del modello che discende da considerazioni della teoria dell'informazione:

per un certo insieme di dati, la complessita` ottima di un modello e` quella per cui il modello si descrive col minimo numero di bit.

- Tenendo conto del fatto che la crescita della dimensione del vettore di parametri (e quindi della complessita` della sua descrizione) e` compensata dalla diminuzione (in media) del numero di bit che descrivono l'errore di predizione, si dimostra che

$$\text{MDL} = (\ln N) \frac{n}{N} + \ln [J(\hat{\vartheta}_N)^{(n)}]$$



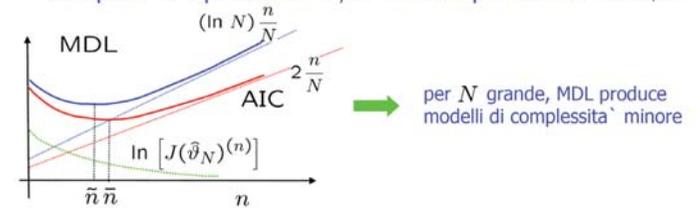
La complessita` ottima rispetto al criterio FPE e` \bar{n}

Confronto tra FPE, AIC e MDL

- Per N elevato, FPE e AIC tendono a produrre gli stessi risultati:

$$\begin{aligned} \ln \text{FPE} &= \ln \left[\frac{N+n}{N-n} J(\hat{\vartheta}_N)^{(n)} \right] = \ln \left[\frac{1+n/N}{1-n/N} J(\hat{\vartheta}_N)^{(n)} \right] \\ &= \ln (1+n/N) - \ln (1-n/N) + \ln [J(\hat{\vartheta}_N)^{(n)}] \\ &\simeq 2 \frac{n}{N} + \ln [J(\hat{\vartheta}_N)^{(n)}] = \text{AIC} \end{aligned}$$

- AIC e MDL hanno struttura simile e differiscono per la costante che moltiplica n : per AIC e` $2/N$ mentre per MDL e` $\ln N/N$



- In generale non e` detto che i criteri abbiano un unico minimo