Scelta della complessita` dei modelli

Premessa: verifica di bianchezza (test di Anderson)

- A procedura di identificazione ultimata bisogna verificare che l'errore di predizione sia assimilabile il piu` 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 e` l'ampiezza della finestra temporale considerata.

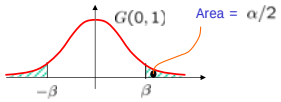
- Nel test di Anderson si utilizza la funzione di covarianza campionaria normalizzata:

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

- Si dimostra che se $\varepsilon(\cdot)$ e` bianco ➡ $\sqrt{N}\,\hat{\rho}(\tau) \sim \text{As}\,G(0,1)$

e che $\hat{\rho}(i)$ e` asintoticamente scorrelato da $\hat{\rho}(j),\ i \neq j$

- Il test di Anderson si effettua cosi`:
  - 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 $\hat{\rho}(\tau)$:

$\hat{\rho}(0), \hat{\rho}(1), \hat{\rho}(2), \ldots, \hat{\rho}(M)$

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

Complessita` dei modelli

- Caratterizziamo la complessita` 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 e` il vettore dei parametri incogniti, $n = \dim(\vartheta)$ e $\varepsilon(t)$ e` l'errore di predizione all'istante t : $\varepsilon(t) = y(t) - \hat{y}(t \mid t-1)$

- Si consideri $\hat{\vartheta}_N = \arg\min_{\vartheta} J(\vartheta)$
- $J(\hat{\vartheta}_N)$ puo` essere considerato un indice di aderenza del modello ai dati
- Tuttavia, a parita` di realizzazione dei dati, $J(\hat{\vartheta}_N)$ diminuisce al crescere della complessita` n ➡ $J(\hat{\vartheta}_N)$ non e` utile di per se` per determinare la complessita` ottima del modello

Esempio

Consideriamo il processo (sistema vero):

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

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

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

$$\mathcal{M}(\vartheta): \quad y(t) = a_1\,y(t-1) + \cdots + a_n\,y(t-n) + \xi(t)$$
$$+ b_1\,u(t-1) + \cdots + b_n\,u(t-n) + \xi(t)$$

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

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

- 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

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:

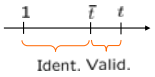
  utilizzare gli stessi dati per identificare il modello e per validarlo

  ↳ $J(\hat{\vartheta}_N)$ 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

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 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 $J_{CV}(\hat{\vartheta})$ e`minimo

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

Final prediction error (FPE)

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

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

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 e` ϑ a tutte le possibili sequenze di dati
- Come sempre $\hat{\vartheta}_N = \arg\min_{\vartheta} J(\vartheta)$

in cui la minimizzazione e`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

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

e la complessita` ottima e` quella per cui FPE risulta minimizzato

Valutiamo FPE in un caso particolare:

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

Quindi:

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

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

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

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

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

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

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

per cui, dalla definizione consegue che

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

D'altra parte, per N sufficientemente elevato:

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

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

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

e quindi

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

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

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

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

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 puo`dimostrare che per $N$ suff. grande, una stima di $\lambda^2$ e`data da

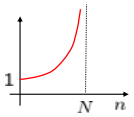
$$\bar{\lambda}^2 = \frac{1}{N-n}\sum_{t=1}^{N}[\varepsilon(t)]^2 = \frac{N}{N-n}\frac{1}{N}\sum_{t=1}^{N}[\varepsilon(t)]^2 = \frac{N}{N-n}J(\bar{\vartheta}_N)^{(n)}$$

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

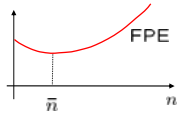
$$\text{FPE} = \frac{N+n}{N-n}J(\bar{\vartheta}_N)^{(n)}$$

Considerazioni:

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



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



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

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

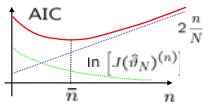
Akaike information criterion (AIC)

- Si tratta di un criterio di scelta della complessita`del modello di tipo statistico. Viene determinato minimizzando la distanza di Kullback tra la densita`di probabilita` dei dati osservati e quella che produrrebbe il modello in esame dove si definisce distanza di Kullback la quantita`

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

- Si dimostra

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



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

Osserviamo che la velocita`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.

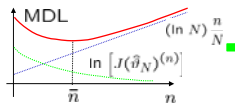
Minimum description length (MDL)

- 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

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



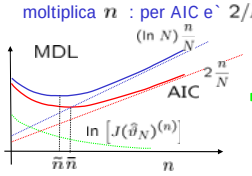
La complessita` ottima rispetto al criterio FPE e` $\bar{n}$

Confronto tra FPE, AIC e MDL

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

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

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



per N grande, MDL produce modelli di complessita` minore

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