Note relative a

- **test di bianchezza**
- **rimozione delle componenti deterministiche da una serie temporale**

a supporto del Progetto di Identificazione dei Modelli e Analisi dei Dati

Maria Prandini
Dipartimento di Elettronica e Informazione
Politecnico di Milano
e-mail: prandini@elet.polimi.it

Test di bianchezza

- test per decidere se un processo $X(\cdot)$ è bianco, sulla base di una sua realizzazione $x(1), x(2),\ldots, x(N)$

- nel contesto dell'identificazione, il test di bianchezza é utile per
  - verificare la bontà di un modello identificato mediante l'analisi del residuo, cioè dell'errore di predizione associato al modello identificato: se il residuo è un processo bianco (a media nulla), allora il modello identificato è una buona descrizione della serie temporale/sistema dinamico in esame
  - valutare sulla base di una realizzazione se, una volta rimosse le eventuali componenti deterministiche, la componente stocastica di una serie temporale è un processo bianco

- Idea: se X($\cdot$) è un processo bianco (per semplicità supponiamo con media nulla), allora
  $\widehat{\gamma}_N(X,n) = \frac{1}{N-|n|}\sum_{t=1}^{N-|n|} X(t)X(t+|n|),$
  è uno stimatore consistente in media quadratica della funzione di covarianza del processo X($\cdot$). Ciò significa che per n≠0

  $E[\widehat{\gamma}_N(X,n)] = 0, \, \forall N$
  $E[\widehat{\gamma}_N(X,n)^2] \to 0, \, N \to \infty$

  Ci si aspetta quindi che, per N sufficientemente elevato, la stima $\widehat{\gamma}_N(x,n)$ basata sui dati x=(x(1), x(2), …, x(N)) a disposizione assuma valori prossimi a zero per ogni n≠0
- Il test statistico di Anderson rende questa idea quantitativa
- Per eliminare l'influenza del fattore di scala, si considera lo stimatore della funzione di covarianza normalizzata $\widehat{\rho}_N(X,n) = \frac{\widehat{\gamma}_N(X,n)}{\widehat{\gamma}_N(X,0)}$

Test di bianchezza di Anderson

- E' un test statistico in cui l'ipotesi nulla è "il processo X(·) è bianco".

- Sulla base dell'ipotesi nulla si ricavano certe proprietà dello stimatore $\widehat{\rho}_N(X,n)$ legate al modo con cui tende a zero quando il numero di dati N tende all'infinito per n≠0.

- Se i dati x=(x(1), x(2), …, x(N)) sono tali per cui tali proprietà sono soddisfatte dalla stima $\widehat{\rho}_N(x,n)$, allora l'ipotesi nulla è accettata, altrimenti è rigettata e viene accettata l'ipotesi opposta.

Proprietà dello stimatore $\widehat{\rho}_N(X,n)$

Se il processo X(·) è bianco,

- $\sqrt{N}\widehat{\rho}_N(X,n) \sim G(0,1), N \to \infty$
- $E[\widehat{\rho}_N(X,n_1)\widehat{\rho}_N(X,n_2)] \to 0, N \to \infty$

per n, n_1, n_2>0, $n_1 \neq n_2$.

Ciò implica che per N sufficientemente elevato:

- la variabile casuale $\sqrt{N}\widehat{\rho}_N(X,n)$ è distribuita come una gaussiana con valore medio nullo e varianza unitaria (gaussiana standard), quindi la probabilità α che $\sqrt{N}\widehat{\rho}_N(X,n)$ sia maggiore in modulo di β è data da

  $\alpha = \int_{|y|>\beta} \frac{1}{\sqrt{2\pi}} exp(-y^2/2)$
- le variabili casuali $\sqrt{N}\widehat{\rho}_N(X,n_1)$ e $\sqrt{N}\widehat{\rho}_N(X,n_2)$ sono incorrelate

Il test di bianchezza viene riformulato come test di gaussianità

Test di bianchezza di Anderson

Si potrebbe basare il test di bianchezza sulla proprietà

$$P(|\sqrt{N}\widehat{\rho}(X,n)| > \beta) = \alpha$$

per un singolo n≠0, procedendo nel modo seguente:

1. si fissa α e si calcola il valore di β corrispondente
2. si calcola $\sqrt{N}\widehat{\rho}_N(x,n)$ sulla sequenza di dati x=(x(1), x(2), …, x(N)) a disposizione
3. si verifica se $\sqrt{N}\widehat{\rho}_N(x,n)$ ha modulo inferiore a β. In caso affermativo, si dice che "la sequenza passa il test di bianchezza con livello di significatività del 100α%" e si accetta l'ipotesi nulla; altrimenti si rigetta l'ipotesi nulla

Nel test di Anderson si considerano contemporaneamente le stime

$$\sqrt{N}\widehat{\rho}_N(x,n),\ n = 1, 2, \ldots, M$$

Osservazioni

- α ($0<\alpha<1$) è detto "parametro di significatività"

- per il test di bianchezza descritto:
  - la probabilità di concludere che $X(\cdot)$ non è bianco quando lo è vale α per $N\to\infty$. Diminuendo α si diminuisce tale probabilità, ma si aumenta quella di concludere che $X(\cdot)$ è bianco quando non lo è.
  - la probabilità di concludere che $X(\cdot)$ è bianco quando non lo è non è calcolabile in quanto non è nota la distribuzione di $\sqrt{N}\hat{\rho}_N(X,n)$ quando $X(\cdot)$ non è bianco.

Test di bianchezza di Anderson

1. si fissa il parametro di significatività α e si calcola il valore di β corrispondente

2. si calcolano $\sqrt{N}\widehat{\rho}_N(x,n),\ n=1,2,\ldots,M$ sulla sequenza di dati x=(x(1), x(2), …, x(N)) a disposizione

3. si conta il numero M_{out} di valori di n tali che $|\sqrt{N}\widehat{\rho}_N(x,n)| > \beta$

4. se $\frac{M_{out}}{M} \leq \alpha$, allora "la sequenza passa il test di bianchezza con livello di significatività del 100α%" e si accetta l'ipotesi nulla; altrimenti si rigetta l'ipotesi nulla

Osservazioni

- nel test di Anderson vengono utilizzate contemporaneamente tutte le stime $\sqrt{N}\widehat{\rho}_N(x,n), \; n = 1, 2, \ldots, M$

- è possibile dimostrare che la probabilità di concludere che X(·) non è bianco quando lo è vale α per $N \to \infty$ (si deve sfruttare l'incorrelazione tra gli stimatori associati a valori di n diversi)

- Valori tipici:

  α=0.05 (β=1.96)

  M tra 5 e N/4 (la stima della funzione di covarianza fatta su N dati è poco accurata per valori di n prossimi a N)

Rimozione di componenti deterministiche da una serie temporale

- Data la sequenza di dati y(1), y(2), …, y(N) relativi ad una serie temporale $Y(\cdot)$, prima di cercare di descrivere la serie temporale come un rumore bianco filtrato attraverso un sistema ARMA, AR o MA, è necessario verificare che tale descrizione sia adeguata. In particolare, $Y(\cdot)$ deve essere un processo stazionario, con funzione di autocorrelazione $\gamma(n)$ che tende a zero quando n tende all'infinito.

- Tali proprietà vengono a mancare quando è presente nel processo $Y(\cdot)$ una componente che è una funzione deterministica del tempo, non costante. In tal caso:
  - la componente deterministica deve essere stimata e rimossa, per poi identificare un modello ARMA, AR o MA del processo così ottenuto;
  - il modello finale per $Y(\cdot)$ sarà costituito dalla componente deterministica stimata e dal processo ARMA, AR o MA identificato

Decomposizione classica con modello additivo

Supponiamo che

$Y(t)=V(t)+p(t)+s(t)$

dove

- $V(\cdot)$ è un processo stazionario a media nulla e con funzione di autocorrelazione che tende a zero
- $p(\cdot)$ è una funzione lineare del tempo (trend lineare)
- $s(\cdot)$ è una funzione periodica del tempo (componente stagionale)

Note:

- una volta rimosso il contributo di $p(\cdot)$ e $s(\cdot)$, ha senso identificare il modello ARMA più opportuno che alimentato da un rumore bianco con valore medio nullo descrive il processo $V(\cdot)$
- se $Y(\cdot)$ fosse stazionario a media non nulla, $p(\cdot)$ sarebbe una costante pari al valore medio del processo
- data la presenza di $p(\cdot)$, si può assumere che $s(\cdot)$ abbia media temporale sul periodo nulla

Come ci si accorge della presenza di trend lineari o componenti stagionali?

La presenza di un trend lineari o di una componente stagionale può

- essere suggerita dal fenomeno fisico sottostante il processo in esame

  per esempio è noto che le temperature sulla superficie terrestre stanno aumentando progressivamente e che variano durante l'anno con una certa periodicità per effetto delle stagioni

- essere messa in luce rappresentando i dati $y(1), y(2), \dots, y(N)$ in funzione del tempo

  per esempio se $Y(t)=V(t)+p(t)$ e la componente stocastica $V(t)$ ha varianza piccola rispetto al contributo del trend lineare, allora si nota un andamento lineare dei dati con una perturbazione sovrapposta

- emergere a seguito dell'elaborazione di y= (y(1), y(2), …, y(N)):
  - la trasformata di Fourier di y può mettere in luce la presenza di una componente stagionale somma di sinusoidi
    - i picchi presenti nella trasformata di Fourier indicano le pulsazioni delle sinusoidi che contribuiscono alla componente stagionale
  - la stima $\widehat{\rho}_N(y,n)$ della funzione di autocorrelazione normalizzata ottenuta a partire dai dati y considerando $Y(\cdot)$ stazionario:
    - se è presente un trend lineare, tende a rimanere diversa da zero per un numero elevato di passi
    - se è presente una componente stagionale, mostra anch'essa una componente stagionale con lo stesso periodo
  - il modello identificato sulla base dei dati y considerando $Y(\cdot)$ stazionario
    - presenta poli vicini al cerchio di raggio unitario, alla pulsazione delle sinusoidi che contribuiscono alla componente stagionale o vicino a 1 se è presente un trend lineare

Rimozione del trend lineare

- Attraverso l'operatore derivata a tempo discreto:

  $\Delta Y(t) := Y(t) - Y(t-1)$

  Infatti, se $Y(t)=V(t)+ a_0+a_1t$, applicando tale operatore si ottiene $Z(t)=\Delta Y(t)=V(t)-V(t-1)+a_1$

  cioè un processo stazionario con media a_1, la quale può essere facilmente stimata facendo la media dei dati $z(1), z(2), \ldots, z(N)$ così calcolati.

  Si noti che:

  - una volta stimato un modello ARMA per il processo $Z(t)$ depolarizzato, il processo $Y(\cdot)$ può essere espresso come l'integrale a tempo discreto del processo ARMA $Z(t)$, cioè $Y(t)=1/(1-z^{-1})\, Z(t)$.
    $Y(\cdot)$ è un processo non stazionario "ARIMA".
  - se il trend fosse quadratico invece che lineare bisognerebbe applicare l'operatore derivata 2 volte

- Mediante la risoluzione di un problema di interpolazione:
  - si considera p(t)=a+bt e si determinano $\hat{a}, \hat{b}$ che minimizzano la norma del vettore degli errori (y(1)-p(1), y(2)-p(2), …, y(N)-p(N)):

    $$J(\theta) = \sum_{t=1}^{N} [y(t) - \phi(t)^T\theta]^2$$

    dove $\theta=[a\ b]^T$ e $\phi(t)=[1\ t]^T$
  - si rimuove il contributo di $\hat{p}(t) = \hat{a} + \hat{b}t$ dai dati y(1), y(2), …, y(N)

Si noti che:

- quello che si deve risolvere è un problema ai minimi quadrati, di facile risoluzione data la struttura lineare nei parametri di p(t)
- l'approccio è generalizzabile al caso di p(t) polinomio in t di ordine maggiore di 1 (trend polinomiale)

Rimozione della componente stagionale

- Attraverso un'operazione di media temporale:

  se T è il periodo della componente stagionale, indicato con K il numero di periodi all'interno della finestra temporale [1, N], si stima s(t) in un periodo come

$$\hat{s}(t) = \frac{1}{K} \sum_{h=0}^{K-1} y(t + hT),\ t = 1, 2, \ldots, T$$

  Si noti che:

  - si richiede sia noto il periodo della componente stagionale
  - se Y(t)=V(t)+s(t) con V(t) processo stazionario a valore medio nullo, allora la stima di s(t) così calcolata tende al valore vero quando K tende all'infinito:
    - » è necessario che K sia elevato per avere una stima accurata di s(t)
    - » se fosse presente una componente p(t) sarebbe necessario rimuoverla prima di effettuare la stima di s(t), questo a meno che p(t) non sia costante nel tempo

- Mediante la risoluzione di un problema di interpolazione:
  - se s(t) è la somma di un numero finito m di sinusoidi di pulsazione nota:

$$s(t) = \sum_{i=1}^{m} A_i sen(\omega_i t + \varphi_i)$$

    - si riparametrizza s(t) come

$$s(t) = \sum_{i=1}^{m} [a_i sen(\omega_i t) + b_i cos(\omega_i t)]$$

mediante la trasformazione $a_i = A_i cos(\varphi_i) \quad b_i = A_i sen(\varphi_i)$

    - si determinano $\widehat{a}_i, \widehat{b}_i,\ i = 1, 2, \ldots, m$ che minimizzano

$$J(\theta) = \sum_{t=1}^{N} [y(t) - \phi(t)^T \theta]^2$$

dove $\theta=[a_1\ b_1 \ldots a_m\ b_m]^T$ e $\phi(t)=[sen(\omega_1 t)\ \ cos(\omega_1 t) \ldots sen(\omega_m t)\ \ cos(\omega_m t)]^T$

    - si definiscono $\widehat{A}_i = \sqrt{\widehat{a}_i^2 + \widehat{b}_i^2}$ e $\widehat{\varphi}_i$ come l'angolo misurato in radianti tale che $cos(\widehat{\varphi}_i) = \widehat{a}_i / \widehat{A}_i \quad sen(\widehat{\varphi}_i) = \widehat{b}_i / \widehat{A}_i$

  - si rimuove il contributo di $\widehat{s}(t) = \sum_{i=1}^{m} \widehat{A}_i sen(\omega_i t + \widehat{\varphi}_i)$ da y(1), … , y(N)

Si noti che:

- quello che si deve risolvere è un problema ai minimi quadrati di facile risoluzione data la struttura lineare nei parametri di s(t)
- si richiede sia noto il periodo della componenti sinusoidali
- dato che s(t) è periodico, allora può essere sempre espresso come la somma di un numero infinito di sinusoidi (sviluppo in serie di Fourier):
    - se la serie è in effetti una somma finita, il metodo può essere applicato
    - se la serie ha infiniti termini, il metodo può essere usato per stimare il contributo di un numero finito di armoniche