

Statistica e Modelli Stocastici II

A.A. 2021/2022

Corso in modalità duale

- Prof. Francesco Finazzi francesco.finazzi@unibg.it
- Dott. Frank Massoda frank.massodatchoussi@unibg.it



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

ILIAS

- Piattaforma e-learning
- Link di accesso: <https://ilias.unibg.it>
- Utilizzare pulsante **IPD login** e inserire le proprie credenziali UNIBG
- Corso: **Repository -> Risorse -> Corsi Attivi -> Statistica -> Statistica e Modelli Stocastici - AA 21/22**
- Materiale abilitato in lettura con l'avanzare del corso
- Forum sempre aperto per commentare argomenti del corso ed esercitazioni
- Tutte le comunicazioni agli studenti avverranno tramite ILIAS (no MS Teams salvo problemi della piattaforma ILIAS)



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Struttura del corso

- Corso in modalità duale
- Piattaforme utilizzate:
 - Microsoft Teams (lezioni ed esercitazioni)
 - ILIAS (materiale del corso, forum, test, comunicazioni agli studenti)
 - MATLAB (esercitazioni e lavori di gruppo)
- Calendario
 - Lunedì 14-15:30 (lezione)
 - Martedì 14-15:30 (lezione)
 - Martedì 16-17:30 (lab, dall'1 marzo)
- Possibile scambio tra esercitazioni e lezioni in base all'avanzamento del programma



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Modalità esame – Prove intermedie

- Oggetto di valutazione
 - Lavoro di gruppo
 - Teoria: tramite test ILIAS
- Lavoro di gruppo
 - Analisi data set scelto dal gruppo
 - Analisi basata sugli argomenti visti a teoria
- Gruppi
 - Composti da 3 o 4 studenti
 - Sono ammessi gruppi da 1 (ma non da 2)
 - E' possibile ma non obbligatorio ricostituire gli stessi gruppi di SMS I



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Modalità esame – Prove intermedie

- Scelta del data set
 - Il data set è scelto dal gruppo
 - Se più gruppi scelgono lo stesso data set, questo deve essere ritagliato su range temporali e/o spaziali differenti
 - Idealmente il data set deve includere serie storiche multiple
 - Cosa è una serie storica?
 - Caso tipico: serie storiche indicizzate nello spazio
 - Serie storiche concorrenti (stesso periodo temporale)
 - Più semplice se si tratta di serie storiche ad intervallo regolare
 - Eventualmente con dati mancanti
 - Eventualmente ad alta frequenza (non necessario)
 - Possibilmente con una o più serie storiche che dipendono da tutte le altre



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Modalità esame – Prove intermedie

- Valutazione
 - Due valutazioni del lavoro di gruppo (intermedia e finale)
 - Due test ILIAS con domande a risposta multipla su argomenti teorici ([da remoto](#))
 - Eventuale orale
- Composizione del voto
 - Media pesata: $V = 0.3G_1 + 0.3G_2 + 0.2T_1 + 0.2T_2$
 - G_1 e G_2 sono le valutazioni dei lavori di gruppo, T_1 e T_2 sono le valutazioni dei test
 - Tutte le valutazioni sono in 33esimi
 - round(V) > 30 implica la lode
 - Se $V \geq 18$, è possibile sostenere un orale facoltativo per migliorare (o peggiorare) il voto V
 - Se $V < 18$ oppure $(T_1 + T_2)/2 < 18$, l'orale è obbligatorio ([da svolgere in corrispondenza degli orali degli appelli](#))
 - In caso di orale obbligatorio **insufficiente** è necessario presentarsi ad uno degli appelli scritti



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Modalità esame – Appello scritto

- Scritto classico con domande aperte sugli argomenti teorici **e l'uso di Matlab** (no sviluppo codice)
- L'esame si considera superato se il voto dello scritto è ≥ 18
- Orale facoltativo
- Tipicamente l'orale si svolge entro 4 giorni dallo scritto
- **Per partecipare all'appello scritto è necessario prenotarsi tramite sportello internet di UNIBG**



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Quale modalità scegliere?

- Prove intermedie
 - Applico la teoria all'analisi di un dataset
 - Lavoro in gruppo (imparo dagli altri)
 - Sono obbligato a rimanere al passo con il corso
 - Stesura di 2 report + 2 test a risposta multipla (**impegno continuo**)
 - Consegnà del secondo report intorno al 12/6. In caso di insufficienza non riesco a partecipare al primo appello di giugno. (Perché?)
 - Lavoro in gruppo (coordinamento)
- Appello scritto
 - Studio e mi presento all'appello (rapido)
 - Devo aver capito quello che ho studiato
 - Devo saper scrivere in italiano
 - Devo saper argomentare anche con formule ed equazioni



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Verbalizzazione voto (sufficiente)

- Prove intermedie
 - Per verbalizzare un voto sufficiente ottenuto nelle prove intermedie è necessario iscriversi ad uno degli appelli del corso tramite sportello internet.
 - **Se ci si presenta allo scritto dell'appello, il voto ottenuto nelle prove intermedie viene perso, anche in caso di ritiro durante lo scritto.**
- Appello scritto
 - La verbalizzazione di un voto sufficiente ottenuto nell'appello scritto è automatica in quanto già iscritti tramite sportello internet. Ovviamente è possibile rinunciare al voto ottenuto.
- Consiglio:
 - Se svolgete le prove intermedie iscrivetevi all'appello di giugno non appena l'appello risulta prenotabile. **Questo nonostante il voto finale delle prove intermedie sarà disponibile dopo la data dello scritto del primo appello.**



Cronoprogramma

- Primo test online teorico (da remoto)
- Consegnna primo elaborato
- **Valutazione primo elaborato**
- Secondo test online teorico (da remoto)
- Consegnna secondo elaborato
- Scritto del primo appello
- **Valutazione secondo elaborato, voto globale prove intermedie, elenco studenti con orale obbligatorio**
- Orale del primo appello (e orali obbligatori da prove intermedie)
- Verbalizzazione voti primo appello (incluse prove intermedie)
- Scritto del secondo appello
- Orale del secondo appello (e orali obbligatori da prove intermedie)
- Verbalizzazione voti secondo appello (incluse prove intermedie)



Introduzione

- SMS II è la prosecuzione di SMS I
 - Viene dato per scontato che gli argomenti di SMS I siano acquisiti: distribuzioni di prob., intervalli di confidenza, test d'ipotesi, regressione
- Corso orientato all'analisi dei dati nell'ambito dell'inferenza statistica
- Preso un data set, l'obiettivo è estrarre informazione utile dai dati
- Per poter prendere decisioni ragionate in condizioni di incertezza
- Non esiste un (solo) libro di SMS II. Il materiale di riferimento è composto da
 - Slide del corso
 - Libri di testo (biblioteca)
 - Articoli scientifici
- Lettura consigliata: Shumway & Stoffer. *Time Series Analysis and Its Applications*. 4th edition.



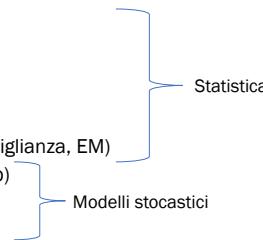
Introduzione

- Differenza fondamentale SMS I e II
 - SMS I introduce l'inferenza statistica per **dati IID**
 - SMS II rilassa questa ipotesi ed affronta il caso di
 - Dati correlati (tipicamente nel tempo)
 - Errori correlati
 - Cosa implica la correlazione? Minore informazione.
- Dipendenza tra n variabili
 - Da variabili casuali univariate a multivariate



Introduzione

- Argomenti del corso
 - Normale multivariata
 - Modelli di regressione avanzata
 - Regressione funzionale
 - Tecniche di validazione e cross-validazione
 - Metodi di stima (dei momenti, massima verosimiglianza, EM)
 - Simulazione stocastica (Monte Carlo e bootstrap)
 - Processi stazionari a tempo discreto
 - Modelli state-space e filtro di Kalman



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Introduzione

- Competenze acquisite
 - Modellazione di data set con n variabili
 - Regressione multipla
 - Validazione e scelta dei modelli
 - Analisi di serie storiche
 - Previsioni nel tempo
 - Inferenza statistica sull'output di modelli/algoritmi complessi



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Introduzione

- Lavori di tesi
- Progetto Earthquake Network (www.sismo.app)
 - Sviluppo Android
 - Sviluppo iOS
 - Analisi dati sismici spazio-temporali
 - Deep learning
 - Analisi dati di mobilità/lockdown



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Statistica e Modelli Stocastici 2017-18

Lezione 1

Normale Multivariata

1. Distribuzioni multivariate
2. Momenti
3. I dati multivariati
4. La distribuzione normale multivariata

Distribuzioni multivariate

Consideriamo il vettore aleatorio

$$x = (x_1, \dots, x_k)'$$

La sua densità di probabilità (*pdf*) congiunta è

$$f_{x_1, \dots, x_k}(t) = f_{x_1, \dots, x_k}(t_1, \dots, t_k) \geq 0 \quad \forall t \in R^k$$

tale che

$$\int_{R^k} f(t) dt = 1$$

Marginali

Le distribuzioni delle v.c. marginali x_j si deducono dalla distribuzione congiunta f_{x_1, \dots, x_k} tramite integrazione:

$$f_{x_j}(t_j) = \int_{R^{k-1}} f_{x_1, \dots, x_k}(t) \prod_{i \neq j} dt_i$$

Condizionate

Indichiamo con a e b due sottoinsiemi di $(1, \dots, k)$. La distribuzione condizionata di $x_a | x_b$ è data da

$$f_{x_a|x_b}(t_a | t_b) = \frac{f_{x_{a|b}}([t_a, t_b])}{f_{x_b}(t_b)}$$

per $f_{x_b}(t_b) > 0$

Momenti misti

Medie

Il vettore aleatorio x ha vettore delle medie (se esistono)

$$E(x) = \mu = (\mu_1, \dots, \mu_k)'$$

dove

$$\mu_j = E(x_i) = \int_R t f_{x_j}(t) dt$$

Momenti misti bivariati

$$\mu_{ij} = E(x_i x_j) = \int_{R^2} t_i t_j f_{x_i x_j}(t_i, t_j) dt_i dt_j$$

Matrice di varianze-covarianze

Il vettore aleatorio x ha matrice di varianze-covarianze $k \times k$ dimensionale (se esiste):

$$V(x) = \Sigma = (\sigma_{ij})_{i,j=1,\dots,k} = \begin{pmatrix} \sigma_{11} & \dots & \sigma_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_{k1} & \dots & \sigma_{kk} \end{pmatrix}$$

NB:

$$V(x) = E(x - \mu)(x - \mu)' = \Sigma = (\sigma_{ij})$$

$$\text{Covarianze : } \sigma_{ij} = \sigma_{ji} = \text{Cov}(x_i, x_j) = E(x_i - \mu_i)(x_j - \mu_j)$$

$$\text{Varianze : } \sigma_{ii} = \sigma_i^2 = \text{Var}(x_i) = E(x_i - \mu_i)^2$$

Proprietà di Σ

La matrice Σ

1. è simmetrica

$$\sigma_{ij} = \sigma_{ji}$$

2. è semidefinita positiva:

$$\Sigma \geq 0$$

$$\det(\Sigma) \geq 0$$

3. se $\Sigma > 0$ allora

a. esiste Σ^{-1}

b. $x' \Sigma^{-1} x > 0 \forall x \neq 0 \in R^k$

4. $\Sigma > 0$ sse autovalori tutti > 0

5. $\Sigma \geq 0$ sse autovalori tutti ≥ 0

Esempi con matlab

```
SIGMA1=[3.0 1.0 0.5; 1.0 1.0 2.0; 0.5 2.0 5.0]
```

```
SIGMA2=[3.0 2.0 1.5; 2.0 5.0 2.0; 1.5 2.0 4.0];
```

Sono matrici di var-cov ben definite?

```
Autovalore_minimo = min(eig(SIGMA1))
```

Matrice di correlazione

$$R = (\rho_{ij})_{i,j=1,\dots,k} = \begin{pmatrix} 1 & \dots & \rho_{1k} \\ \vdots & \ddots & \vdots \\ \rho_{k1} & \dots & 1 \end{pmatrix}$$

Legame fra matrice di Correlazione e matrice di varianze-covarianze:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

$$\sigma_{ij} = \sigma_i \sigma_j \rho_{ij}$$

Esempio matlab

```
% medie
```

```
mu = [5,80];
```

```
% deviazioni standard
```

```
stdY = [10 3]';
```

```
% matrice di correlazione
```

```
R = [1 0.5; 0.5 1];
```

```
% matrice di varianze-covarianze
```

```
SIGMA = stdY.*stdY'.*R;
```

```
% è ben definita ?
```

```
determinante = det(SIGMA)
```

```
Autovalore_minimo = min(eig(SIGMA))
```

SMS2_L01 - $N_k(\mu, \Sigma)$

p.10

è ben definita ?

```

simmetria = isequal(SIGMA, SIGMA')
determinante = det(SIGMA)
Autovalore_minimo = min(eig(SIGMA))

```

SMS2_L01 - $N_k(\mu, \Sigma)$

p.12

Esercizio AR(1)Implementare il calcolo della matrice di varianze-covarianze $n \times n$ data da

$$\Sigma = \sigma^2 (\alpha^{i-j})_{i,j=1,\dots,n}$$

per $n = 100, \sigma^2 = 1.5, \alpha = 0.8$.NB: è la matrice di varianze-covarianze di un modello *AR(1)*

```

n=100;
sigma=sqrt(1.5);
alpha=0.8;
idx=[1:n]';
SIGMA = sigma^2 * alpha.^abs(idx-idx');

```

SMS2_L01 - $N_k(\mu, \Sigma)$

p.11

Dati multivariatiI dati multivariati sono tipicamente rappresentati da una matrice $n \times k$:

$$X = \begin{pmatrix} x_{11} & \dots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nk} \end{pmatrix} = \begin{pmatrix} \mathbf{x}'_1 \\ \vdots \\ \mathbf{x}'_n \end{pmatrix}$$

relativi ad n individui, campioni istanti temporali, punti nello spazio etc. (unità statistiche).Le medie campionarie di X sono date dal vettore \bar{x} che ha elementi

$$\bar{x}_j = \frac{1}{n} \sum_{r=1}^n x_{rj}$$

SMS2_L01 - $N_k(\mu, \Sigma)$

p.13

Momenti secondi campionari

La matrice di varianze-covarianze campionarie di X è data da:

$$S = (s_{ij})_{i,j=1,\dots,k} = \left(\frac{1}{n} \sum_{r=1}^n (x_{ri} - \bar{x}_i)(x_{rj} - \bar{x}_j) \right)_{i,j=1,\dots,k}$$

La matrice di correlazione campionaria di X è data da:

$$R = \left(\frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} \right)_{i,j=1,\dots,k}$$

SMS2_L01 - $N_k(\mu, \Sigma)$

p.14

Esempio matlab Sterilizzatrici

```
load('Lisa_Lab_34col.mat')
x = A.TChambInt;
y = A.PChamb;
scatter(x,y,'.')
xbar = mean([x,y])
r_tp=corr(x,y)
R=corr([x,y])
S=cov([x,y])

x=double(A(:,[9:15]));
S=cov(X)
```

SMS2_L01 - $N_k(\mu, \Sigma)$

p.15

min(eig(S))

SMS2_L01 - $N_k(\mu, \Sigma)$

p.16

Distribuzione della matrice X

Tipicamente il vettore k -dimensionale \mathbf{x}_i è una realizzazione del vettore aleatorio x con distribuzione $f_x : \mathbf{x}_i \equiv f_x$

1. IID: Spesso $\mathbf{x}_1, \dots, \mathbf{x}_n$ sono iid f_x , cioè

$$f_{\mathbf{x}_1, \dots, \mathbf{x}_n}(\mathbf{t}_1, \dots, \mathbf{t}_n) = f_x(\mathbf{t}_1) \times \dots \times f_x(\mathbf{t}_n)$$

2. Stazionarietà: Altre volte $\mathbf{x}_i \equiv f_x$ ma le righe di X non sono indipendenti
3. Nonstazionarietà: Altre volte ancora ogni riga ha una distribuzione diversa e le righe sono o meno indipendenti

SMS2_L01 - $N_k(\mu, \Sigma)$

p.17

Normale multivariata

Diremo che x ha distribuzione normale multivariata $N_k(\mu, \Sigma)$, con $\Sigma > 0$, se

$$f_{\mu, \Sigma}(t) = K \exp\left(-\frac{1}{2} (t - \mu)' \Sigma^{-1} (t - \mu)\right) \quad \forall t \in R^k$$

con

$$K^{-1} = |\Sigma|^{1/2} (2\pi)^{k/2}$$

Interpretazione

$$P(x \geq t) \propto f_{\mu, \Sigma}(t)$$

SMS2_L01 - $N_k(\mu, \Sigma)$

p.18

Esempio 1 (opz): la Normale doppia N_2

Consideriamo la distribuzione di (x, y) dove le marginali sono normali

$$x \equiv N(\mu_x, \sigma_x^2)$$

$$y \equiv N(\mu_y, \sigma_y^2)$$

e la (x, y) ha distribuzione congiunta normale con coefficiente di correlazione ρ

$$f_{x,y}(u, v) = \\ = K \times \exp\left(-\frac{1}{2(1-\rho^2)} \left(\left(\frac{u-\mu_x}{\sigma_x}\right)^2 + \left(\frac{v-\mu_y}{\sigma_y}\right)^2 - 2\rho \left(\frac{u-\mu_x}{\sigma_x}\right) \left(\frac{v-\mu_y}{\sigma_y}\right) \right)\right)$$

con

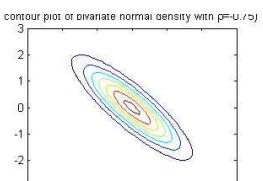
$$K^{-1} = 2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}$$

SMS2_L01 - $N_k(\mu, \Sigma)$

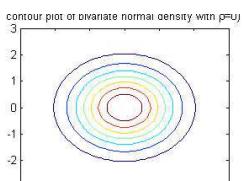
p.19

Esempio 2. Grafico 3D della campana gaussiana

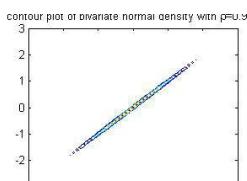
```
contour(X, Y, f)
surf(X, Y, f)
```



$N_2(\rho = -0.75)$



$N_2(\rho = 0)$



$N_2(\rho = +0.99)$

Curve di livello di $f(u, v; \rho)$

SMS2_L01 - $N_k(\mu, \Sigma)$

p.20

Esempio 2. codice

```
x=[-3:.1:3]'; y=x';
rho = -0.5;
R = rho.^ (1-eye(2));
[X Y] = meshgrid(x,y);
f = mvnpdf([X(:) Y(:)], [0,0], R);
f = reshape(f, size(x));
```

SMS2_L01 - $N_k(\mu, \Sigma)$

p.21

Note:

1. Se $x \in N_p$ le distribuzioni marginali univariate e k -variate ($k < p$) sono normali.
a. Non vale il viceversa.
2. Se $x \in N_p$ le distribuzioni condizionate sono normali.
3. $\Sigma \geq 0$ è semidefinita positiva.
4. Se $|\Sigma| = 0$ la distribuzione normale è degenera.
5. Se $|\Sigma| = 0$ c'è multicollinearità.
6. $Rango(\Sigma) \leq k$ e suo significato statistico.
7. $Rango(\Sigma) = Rango(R)$.

SMS2_L01 - $N_k(\mu, \Sigma)$

p.22

Distribuzioni Condizionate**Caso Generale:** Consideriamo la partizione del vettore k -dimensionale x :

$$x = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix},$$

dove

$$\mathbf{x}_i \text{ è } k_i - \text{dim}, \quad i = 1, 2 \quad \text{e} \quad k_1 + k_2 = k$$

ed in corrispondenza

$$Ex = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad \text{con} \quad \boldsymbol{\mu}_i \text{ } k_i - \text{dim}$$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

con $\Sigma_{12} = \Sigma_{21}'$.SMS2_L01 - $N_k(\mu, \Sigma)$

p.23

Naturalmente, per $i = 1, 2$,

$$\mathbf{x}_i \text{ è } N_{k_i}(\boldsymbol{\mu}_i, \Sigma_{ii}).$$

Inoltre la distribuzione condizionata è

$$(x_1 | x_2 = \mathbf{x}) \text{ è } N_{k_1}(\boldsymbol{\mu}_{1|2}(\mathbf{x}), \Sigma_{1|2})$$

dove l'iperpiano di regressione è

$$\begin{aligned} \boldsymbol{\mu}_{1|2}(\mathbf{x}) &= E(x_1 | x_2 = \mathbf{x}) = \boldsymbol{\mu}_1 + A(\mathbf{x} - \boldsymbol{\mu}_2), \\ A &= \Sigma_{12}\Sigma_{22}^{-1} \end{aligned}$$

con $\mathbf{x} \in \mathbf{R}^{k_2}$. La varianza condizionata è costante

$$\Sigma_{1|2} = Var(x_1 | x_2) = \Sigma_{11} - A\Sigma_{21}$$

SMS2_L01 - $N_k(\mu, \Sigma)$

p.24

Previsione GaussianaNoto \mathbf{x}_2 vogliamo prevedere \mathbf{x}_1 . Questo può essere fatto considerando $\mathbf{x}_1 | \mathbf{x}_2$:**Previsione**

$$\boldsymbol{\mu}_{1|2}(\mathbf{x}_2) = Moda(\mathbf{x}_1 | \mathbf{x}_2) = E(\mathbf{x}_1 | \mathbf{x}_2) = \boldsymbol{\mu}_1 + A(\mathbf{x} - \boldsymbol{\mu}_2)$$

Incertezza

$$\Sigma_{1|2} = Var(\mathbf{x}_1 | \mathbf{x}_2) = \Sigma_{11} - A\Sigma_{21}$$

Intervallo di previsioneNel caso scalare, ipotizzando noti i parametri e posto $\sigma_{1|2}^2 = \Sigma_{1|2}$:

$$\boldsymbol{\mu}_{1|2}(\mathbf{x}_2) \pm z_{\alpha/2} \sigma_{1|2}$$

SMS2_L01 - $N_k(\mu, \Sigma)$

p.25

Note

1. Linearità vs nonlinearità:

Il previsore è lineare

$$\mu_{1|2}(\mathbf{x}_2) = \boldsymbol{\mu}_1 + A(\mathbf{x} - \boldsymbol{\mu}_2)$$

2. Omoschedasticità vs eteroschedasticità:

La varianza dell'errore di previsione è costante:

$$\Sigma_{1|2} = \text{Var}(\mathbf{x}_1|\mathbf{x}_2) = \Sigma_{11} - A\Sigma_{21} \quad \forall \mathbf{x}_2$$

3. $x_2 \in N_{k_2}$ e $x_1|x_2 \in N_{k_1}(\mu_{1|2}(\mathbf{x}), \Sigma_{1|2})$ allora $(\begin{smallmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{smallmatrix}) \in N_k$.

Caso Bivariato

$$k_1 = k_2 = 1$$

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \quad R = \begin{pmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{pmatrix}$$

Media condizionata

$$\mu_{1|2}(x) = \mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(x - \mu_2) = \alpha + \beta x$$

$$\sigma_{1|2}^2 = \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2} = \sigma_1^2(1 - \rho_{12}^2)$$

Rappresentazione regressiva (vale anche in generale)

$$x_1 = \alpha + \beta x_2 + \varepsilon \quad \text{con} \quad \varepsilon = N(0, \sigma_{1|2}^2).$$

Singolarità

$$|\Sigma| = 0 \Leftrightarrow \rho_{12} = 1$$

\Updownarrow

$$x_1 = \alpha + \beta x_2 \quad q.c.$$

Infatti, in questo caso:

$$\sigma_{1|2}^2 = 0.$$

Quanto visto vale anche più in generale per $k > 2$, ($k_1 = 1$ per semplicità):

$$|\Sigma| = 0$$

\Updownarrow

$$x_1 = \alpha + \beta' \mathbf{x}_2 \quad q.c.$$

Trasformate Lineari

Se $\mathbf{x} \in N_k(\boldsymbol{\mu}, \Sigma)$ e A è una matrice $p \times k$, $p \leq k$, allora

$$y = Ax \in N_p(A\boldsymbol{\mu}, A\Sigma A')$$

Se $\Sigma > 0$ esiste (almeno) una matrice A tale che

$$\Sigma^{-1} = A * A'$$

$$A\Sigma A' = I \quad \text{e} \quad A^{-1}\Sigma^{-1}A^{-T} = I.$$

Indichiamo una di queste matrici con

$$\Sigma^{-1/2} = A$$

e, per ovvi motivi,

$$\Sigma^{1/2} = A^{-1}.$$

Proprietà

Se $x \in N_k(0, I)$ allora

$$y = \Sigma^{1/2}x \quad \text{è} \quad N_k(0, \Sigma)$$

Se $x \in N_k(0, \Sigma)$ allora

$$y = \Sigma^{-1/2}x \quad \text{è} \quad N_k(0, I)$$

NB:

$$\Sigma^{1/2} * (\Sigma^{1/2})' = \Sigma$$

per esempio

$$\Sigma^{1/2} = \text{chol}(\Sigma)'$$

SMS2_L01 - $N_k(\mu, \Sigma)$

p.30

Forme quadratiche

Se $\Sigma > 0$, allora

$$X^2 = (x - \mu)' \Sigma^{-1} (x - \mu) \equiv \chi_k^2$$

Più in generale se $x \in N_k(0, I)$

$$x' A x \equiv \chi_g^2$$

dove i gradi di libertà g dipendono dal rango di A .

SMS2_L01 - $N_k(\mu, \Sigma)$

p.31

Matlab: Generazione di Numeri Normali Multivariati

```
n=1000; k=2; z =randn(n, 2);
rho = 0.8; SIGMA = ([1 rho; rho 1])
sqrtSIGMA = SIGMA^0.5
chk_sqrt = sqrtSIGMA*sqrtSIGMA
X=z*sqrtSIGMA;
plot(X(:,1),X(:,2),'.');
corr(X)
cholsigma = chol(SIGMA)
chk_chol = cholsigma' * cholsigma
X=z*chol(SIGMA);
corr(X)
X=mvnrnd(zeros(1,k), SIGMA, n);
```

SMS2_L01 - $N_k(\mu, \Sigma)$

p.32

Esempio AR(1)

Se

$$y = (y_1, \dots, y_k)'$$

con

$$y_i = \alpha y_{i-1} + \varepsilon_i$$

per $i = 2, \dots, k$ ed $\varepsilon_i \equiv NID(0, \sigma^2)$ e $y_1 \equiv N\left(0, \frac{\sigma^2}{1-\alpha^2}\right)$ allora

$$y \equiv N_k(0, \Sigma)$$

dove la matrice di varianze covarianze è data da

$$\Sigma = \frac{\sigma^2}{1-\alpha^2} (\alpha^{|i-j|})_{i,j=1,\dots,k}$$

SMS2_L01 - $N_k(\mu, \Sigma)$

p.33

Esercizi

1. Simulare y AR(1) Gaussiano definito sopra tramite:
 - a. l'eq.ne ricorsiva $y_t = \alpha y_{t-1} + \varepsilon_t$
 - b. usando il comando *filter*
 - c. usando *mvnrnd* e la matrice $\Sigma = \sigma(\alpha^{i-j})_{i,j=1,\dots,n}$
2. Previsione Gaussiana
 - a. Con riferimento all'AR(1) di cui sopra fare la previsione di y_k quando $\mu = 0$:

$$\hat{y}_k = E(y_k | \mathbf{y}_{1:k-1} = \mathbf{x}) = \Sigma_{12} \Sigma_{22}^{-1} \mathbf{x}$$

- b. verificare con matlab che

$$\hat{y}_k = \alpha y_{k-1}$$

SMS2_L01 - $N_k(\mu, \Sigma)$

p.34

Modelli stocastici

Lezione 01-B

Analisi e test di Normalità

Test Monte Carlo

Test ed Ellissoidi sulla normale multivariata

L01B - Test di Normalità

p.1

Forma della distribuzione

Supponiamo y iid con distribuzione f che ammette momenti finiti fino all'ottavo.**ASIMMETRIA**Lo skew della distribuzione di y è dato da

$$sk = E\left(\frac{y - \mu}{\sigma}\right)^3 = \frac{\bar{\mu}_3}{\bar{\mu}_2^{3/2}}$$

dove $\bar{\mu}_j$ è il momento centrato di ordine j .**Matlab*****skewness (X)***

L01B - Test di Normalità

p.2

CURTOSILa curtosi della distribuzione di y è data da

$$k = E\left(\frac{y - \mu}{\sigma}\right)^4 = \frac{\bar{\mu}_4}{\bar{\mu}_2^2}$$

NB: alcuni testi usano $k = \frac{\bar{\mu}_4}{\bar{\mu}_2^2} - 3$.**Matlab*****kurtosis (X)***

L01B - Test di Normalità

p.3

Test di Normalità sui momenti (Jarque-Bera)

PREMESSA: Se $y \equiv NID(\mu, \sigma^2)$ allora $sk = 0$ e $k = 3$.

Sia y_t iid ed \bar{m}_j il momento centrato campionario di ordine j . Allora lo skew campionario è dato da

$$\hat{sk} = \frac{\bar{m}_3}{\bar{m}_2^{3/2}}$$

e la curtosi campionaria è data da

$$\hat{k} = \frac{\bar{m}_4}{\bar{m}_2^2}$$

Matlab

```
[h,p,jbstat,critval] = jbtest(x,alpha,mctol);
```

Distribuzione asintotica

Se $y_1, \dots, y_n \equiv NID(\mu, \sigma^2)$ per $n \rightarrow \infty$,

$$S_n = \sqrt{n} \begin{pmatrix} \hat{sk} \\ \hat{k} - 3 \end{pmatrix} \cong N \left(0, \Sigma_S = \begin{pmatrix} 6 & 0 \\ 0 & 24 \end{pmatrix} \right)$$

e analogamente a quanto visto sopra, per $n \rightarrow \infty$,

$$JB_n = S_n' \Sigma_S^{-1} S_n \cong \chi_2^2$$

NB:

$$JB_n = S_n' \Sigma_S^{-1} S_n = \frac{n}{6} (\hat{sk})^2 + \frac{n}{24} (\hat{k} - 3)^2$$

Test di Jarque e Bera

Si rifiuta l'ipotesi di normalità

$$H_0 : P(y_t < y) = \Phi\left(\frac{y - \mu}{\sigma}\right), \text{ con } \mu \text{ e } \sigma \text{ ignoti}$$

se si hanno valori alti di

$$JB_n = \frac{n}{6} (\hat{sk})^2 + \frac{n}{24} (\hat{k} - 3)^2$$

- I Valori critici asintotici di JB_n sono dati da $\chi_{2,\alpha}^2$
- I Valori critici di JB_n per numerosità piccole si calcolano via simulazione (vedi oltre: TEST MONTE CARLO).
- Il test è consistente per violazioni di

$$H'_0 : k = 3, sk = 0$$

Esercizio Matlab

1. Scrivere il codice per eseguire il test usando anche
 - skewness() e kurtosis().
 - Provarlo su dati simulati dalla distribuzione normale, lognormale e χ^2 .
2. Studiare l'approssimazione χ^2 al variare di n confrontando
 - la significatività nominale basata sul χ^2
 - e quella effettiva via Metodo Monte Carlo.
3. **Dati temperature (L07):** testare la normalità dei dati osservati e dei residui del modello AR(1) stimato.

Approfondimenti - dimostrazione:

Si osserva che, se $y_1, \dots, y_n \sim NID(0, 1)$, allora

$$S_n = \sqrt{n} \begin{pmatrix} m_3 \\ \bar{m}_4 - 3 \end{pmatrix} \cong N\left(0, \Sigma_S = \begin{pmatrix} 15 & 0 \\ 0 & 96 \end{pmatrix}\right)$$

inoltre

$$JB = S' \Sigma_S^{-1} S \cong \chi^2_2$$

ove, ricordando che

$$\mu_k = EZ^k = \frac{k!}{(k/2)! 2^{k/2}}$$

è facile vedere che

$$Var(\bar{m}_4) \cong \frac{96}{n} = \frac{\mu_8 - \mu_4^2}{n} \text{ e } Var(m_3) \cong \frac{15}{n} = \frac{\mu_6 - \mu_3^2}{n}$$

tuttavia i conti per sk e \hat{k} sono più complessi.

Approfondimenti - Calcolo di $Var(\hat{k})$

Abbiamo che

$$Var(\bar{m}_4) \cong Var(m_4)$$

in quanto togliere la media non è rilevante perché

$$\sqrt{n} \bar{m}_j \cong \sqrt{n} m_j, \quad j = 2, 4$$

Tuttavia il denominatore di \hat{k} è importante, perché

$$Em_4 = 3.$$

da cui si verifica facilmente che

$$nVar(\hat{k}) \cong nVar(m_4 - 3m_2^2) = 24$$

Approfondimenti - Calcolo di $Var(sk)$

Togliere la media è rilevante nel calcolo di \bar{m}_3 e sk per cui si ha

$$nVar(sk) \cong nVar(\bar{m}_3) \cong 6 = Var(m_3) - 9$$

infatti, sviluppando il cubo,

$$\sqrt{n} \bar{m}_3 = \sqrt{n} m_3 + 2\sqrt{n} m_1^3 - 3\sqrt{n} m_2 m_1$$

ora, il 2° addendo è trascurabile perché

$$(\sqrt{n} m_1)^3 \rightarrow N(0, \tau_1^2)^3$$

mentre per il 3° addendo

$$\sqrt{n} m_1 \times m_2 \rightarrow \mu_2 N(0, \tau_1^2).$$

Perciò

$$\sqrt{n} \bar{m}_3 \cong \sqrt{n} m_3 - 3\sqrt{n} m_1$$

e vale la var di cui sopra.

Test Monte Carlo

Quando la distribuzione della statistica test è ignota sotto H_0 ma si è capaci di simulare i dati sotto H_0 si può stimare il p-value ($\hat{\alpha}$) ed il valore critico col metodo Monte Carlo.

Sviluppiamo il concetto usando l'esempio del test Jarque-Bera (JB). Il metodo usato va bene per qualsiasi test che rifiuta H_0 per valori alti.

Sia

$$\mathbf{y} = (y_1, \dots, y_n)$$

un campione casuale dalla distribuzione sotto H_0 , nel caso di JB usiamo $N(0, 1)$.

Sia

$$jb = JB(\mathbf{y})$$

la statistica test di Jargue-Bera.

Test Monte Carlo (segue)

I test Monte Carlo si sviluppano in 3+1 step, come segue

STEP 1 - Simulazione dati sotto H_0

Simuliamo m replicazioni iid di \mathbf{y} :

$$Y^* = \begin{pmatrix} \mathbf{y}_1^* \\ \vdots \\ \mathbf{y}_m^* \end{pmatrix} = \begin{pmatrix} y_{11}^* & \dots & y_{1n}^* \\ \vdots & & \vdots \\ y_{m1}^* & \dots & y_{mn}^* \end{pmatrix}$$

STEP 2 - Simulazione test

e calcoliamo le statistiche test simulate:

$$\mathbf{j}\mathbf{b}^* = (jb_1^*, \dots, jb_m^*)'$$

dove

$$jb_j^* = jb(\mathbf{y}_j^*).$$

Test Monte Carlo (segue)

STEP 3 - P value/valore critico

Il p -value dei nostri dati è dato dall'azione delle simulazioni che superano la statistica test calcolata:

$$\hat{\alpha} \approx \hat{\alpha}^* = \frac{1}{m} \sum_{j=1}^m I(jb_j^* > jb_{calc})$$

La precisione di tale approssimazione è data dalla ben nota formula del 1° modulo:

$$std(\hat{\alpha}^*) = \sqrt{\frac{\hat{\alpha}^*(1 - \hat{\alpha}^*)}{m}}.$$

Il valore critico simulato per il livello di significatività α è dato dal quantile di jb^* :

$$jb_{crit} = jb_{1-\alpha}^*$$

Test Monte Carlo (segue)

STEP 4 - Potenza

Si vuole calcolare la potenza in relazione ad una specifica alternativa H_1 , fissato il rischio di primo tipo α e ottenuto il valore critico come sopra.

A tal fine è sufficiente fare gli step 1 e 2 di cui sopra ma generando i dati sotto H_1 che danno luogo alla matrice $m \times n$ Y^{H_1} .

Se indichiamo con $jb_j^{H_1}$ le corrispondenti statistiche test simulate sotto H_1 ,

$$jb_j^{H_1} = jb(\mathbf{y}_j^{H_1}),$$

abbiamo che la potenza è data da

$$\pi \cong \pi^* = \frac{1}{m} \sum_{j=1}^m I(jb_j^{H_1} > jb_{crit})$$

Esempio- Test JB via MC

Implementazione test JB via Monte Carlo e calcolo potenza.

NB: si suppone di avere dei dati che hanno fornito un valore $jb_{calc} = 1.2$.

Step 1:

```
n=50; m=1000; alfa=.05; jb_calc=1.2;
X=randn(m,n);
```

Step 2:

```
jb = skewness(X') .^2 * n/6 + (kurtosis(X') -3) .^2 * n/24;
```

Step 3: Valore critico:

```
jb_crit = prctile(jb,100*(1-alfa));
figure; subplot(1,2,1); histogram(jb); title('JB|H0')
hold on; plot(jb_crit,0,'r *')
```

Step 3: P-value

```
pval_MC=mean(jb>jb_calc)
std_pval = sqrt(pval_MC*(1-pval_MC) / m)
```

Step 4: Potenza per H1=χ²

```
df=5;
Y = chi2rnd(df,m,n);
jb1 = skewness(Y') .^2 * n/6 + (kurtosis(Y') -3) .^2 * n/24;
potenza_MC = mean(jb1>jb_crit)
subplot(1,2,2); histogram(jb1); title('JB|H1')
hold on; plot(jb_crit,0,'r *')
```

Test di adattamento χ²

Vogliamo testare l'ipotesi che i dati y_t provengano dalla distribuzione F_0 nota:

$$H_0 : P(y_t \leq y) = F_0(y)$$

Idea base:

- Confrontare l'istogramma osservato con quello teorico (histfit)
- Ovvero, raggruppati i dati nelle k classi C_1, \dots, C_k , confrontare le frequenze osservate

$$O_j = \sum_{t=1}^n I(y_t \in C_j)$$

con quelle teoriche

$$E_j = nP(y_t \in C_j | F_0)$$

Test di adattamento χ²

Perciò si rigetta H_0 se la statistica test di Pearson

$$X^2 = \sum_{j=1}^k \frac{(O_j - E_j)^2}{E_j}$$

ha valori alti.

Distribuzione di X^2

Se F_0 è nota nella forma e nei parametri, allora, per $n \rightarrow \infty$,

$$X^2 \cong \chi_{k-1}^2$$

Se F_0 contiene un vettore m -dimensionale di parametri incogniti θ , stimati a ML, allora

$$X^2 \cong \chi_{k-1-m}^2$$

Osservazioni

Il test può essere usato per qualsiasi tipo di distribuzione F_0 , anche discreta.

In questo caso le classi C_j , spesso coincidono con i valori della y .

Problemi

1. Approssimazione scarsa per n piccolo:
 - a. metodo Montecarlo se θ è noto
 - b. se θ è ignoto si può fare qualcosa ?
2. Scarsa efficienza (cfr test su CDF)
3. Scelta di k
4. Le classi C_j devono essere abbastanza numerose: $E_j \geq 5$

Esercizio Matlab

1. Adattare il codice e l'analisi dell'Esempio **Test JB via MC** al test χ^2 .
2. Scrivere il codice per eseguire il test usando anche
 - `chi2gof()`.
 - Provarlo su dati simulati dalla distribuzione normale, lognormale e χ^2_{10} .
3. Studiare l'approssimazione χ^2 al variare di n confrontando
 - la significatività nominale basata sul χ^2
 - e quella effettiva via Montecarlo.
4. **Dati temperature (L07):** testare la normalità dei dati osservati e dei residui del modello AR(1) stimato.

Test sulla CDF

Consideriamo ora dati con distribuzione continua e l'ipotesi sulla cdf:

$$H_0 : F = F_0.$$

Sia \hat{F} la funzione di ripartizione empirica di y_1, \dots, y_n

$$\hat{F}(x) = \frac{1}{n+1} \sum_{t=1}^n I(y_t \leq x)$$

dove $I(\quad)$ è la funzione indicatrice.

Per testare H_0 , si confronta \hat{F} con F_0 . Abbiamo

1. un approccio grafico (test informale) ed
2. test che controllano gli errori di primo e secondo tipo.

Carte di Probabilità Normali

Quando $F_0 = \Phi$, l'approccio grafico confronta visivamente $\hat{F}(y_i)$ con $\Phi\left(\frac{y_i - \bar{y}}{s}\right)$, più in particolare, si confronta y_i con z_i dove:

$$y_i \approx z_i := s\Phi^{-1}\left(\hat{F}(y_i)\right) + \bar{y}.$$

NB: si possono considerare anche altre distribuzioni per H_0 con le dovute modifiche.

Cosa fa normplot??

Simulare $n = 100$ osservazioni dalla $N(\mu = 1000, \sigma^2 = 49)$ e graficare la carta di probabilità normali (`normplot()`).

Soluzione

```
y=normrnd(1000,7, 100,1);
n=length(y)
Fhat=([1:n]-0.5)/n;
z=std(y)*norminv(Fhat)+mean(y);
subplot(1,2,1)
plot(sort(y),z); hold on; plot(sort(y),sort(y))
subplot(1,2,2)
normplot(y)
```

L01B - Test di Normalità

p.24

Test di Kolmogorov (opzionale)

Se la distribuzione F_0 di y_t non ha parametri incogniti sotto H_0 allora l'ipotesi

$$H_0 : P(y_t \leq y) = F_0(y)$$

può essere testata rifiutando H_0 per valori alti della statistica

$$D_n = \max_{t=1,\dots,n} |\hat{F}(y_t) - F_0(y_t)|$$

Osservazioni

1. la distribuzione asintotica di D_n
 - a. è nota e i suoi valori critici sono tabulati
 - b. non dipende da F_0 (neppure per n piccolo)
2. Se F_0 dipende da parametri ignoti, sostituirli con MLE modifica la distribuzione di D :
 - a. non è nota neanche a livello asintotico
 - b. dipende da F_0
 - c. Si devono calcolare valori critici e/o i p-value per simulazione.

Test di Lilliefors

Il test di Kolmogorov per la verifica di ipotesi di normalità in presenza di parametri ignoti:

$$H_0 : P(y_t \leq y) = \Phi\left(\frac{y - \mu}{\sigma}\right), \mu \text{ e } \sigma \text{ ignoti}$$

rifiuta H_0 quando si hanno valori alti della statistica

$$D = \max_{t=1,\dots,n} \left| \hat{F}(y_t) - \Phi\left(\frac{y_t - \bar{y}}{s_y}\right) \right|$$

i valori critici ed il p-value sono tabulati o si calcolano per simulazione.

Matlab

```
[h,p,dstat,critval] =
lillietest(x,'Alpha',alpha,'MCTol',mctol);
```

L01B - Test di Normalità

p.26

Esercizio Matlab

1. Adattare il codice e l'analisi dell'Esempio **Test JB via MC** al test di Lilliefors.
2. Scrivere il codice per eseguire il test usando anche
 - `lillietest()`.
 - Provarlo su dati simulati dalla distribuzione normale, lognormale e χ^2_{10} .
3. Confrontare il p-value tabulato con quello via Montecarlo.
4. **Dati temperature (L07):**
 - a. testare la normalità dei dati osservati e dei residui del modello AR(1) stimato.
 - b. Confrontare i risultati dei diversi test di normalità fatti sui dati.
5. Confrontare per via Monte Carlo i test χ^2 , di Lilliefors ed il test *JB* in relazione ad approssimazione asintotica e potenza.

L01B - Test di Normalità

p.27

Trasformate "normalizzanti"

Per migliorare l'approssimazione con la normale si usano delle trasformate:

1. Radice quadrata $z_t = \sqrt{y_t - c}$

2. log $z_t = \log(y_t - c)$

3. log return:

$$z_t = \log\left(\frac{y_t}{y_{t-1}}\right)$$

4. Box-Cox

$$z_t(\lambda) = \frac{y_t^\lambda - 1}{\lambda}.$$

dove λ può essere scelto ottimizzando la verosimiglianza dei dati.

Test sulla normale multivariata opzionale

Supponiamo di avere un campione $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ iid dal vettore casuale $\mathbf{x}' = (x_1, \dots, x_k)$ con distribuzione $N_k(\boldsymbol{\mu}, \Sigma)$.

1. Stima di $\boldsymbol{\mu}$: $\bar{\mathbf{x}} = \frac{1}{n} \sum_{t=1}^n \mathbf{x}_t$

2. Stima di Σ :

$$S = \left(\frac{1}{n-1} \sum_t (\mathbf{x}_t - \bar{\mathbf{x}})(\mathbf{x}_t - \bar{\mathbf{x}})' \right)$$

Interessa l'ipotesi sulla media

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$$

test T^2

Varianza nota:

$$n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \Sigma^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \text{ è } \chi_k^2$$

Varianza ignota (opzionale):

$$T^2 = \frac{n-k}{k} \frac{n}{n-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' S^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \text{ è } F_{k,n-k}$$

Test di simmetria (opzionale)

$$H_0 : \mu_1 = \dots = \mu_k.$$

Usando una matrice *contrasto* C , $(k-1) \times k$, di rango $k-1$, con

$$C\mathbf{1}_{k,1} = 0_{k,1},$$

si trasformano i dati

$$\mathbf{y}_t = C\mathbf{x}_t$$

e si applica poi il test T^2 al problema trasformato: sotto H_0

$$T^2 = \frac{n-k}{k-1} \frac{n}{n-1} \bar{\mathbf{y}}' S_y^{-1} \bar{\mathbf{y}}$$

è $F_{k-1,n-k}$ che risulta C -invariante.

Costruzione di C :

1. $x_i - x_1$, $i = 2, \dots, k$
2. $x_i - x_{i-1}$, $i = 2, \dots, k$

Ellissoidi di confidenza

Se θ è k -variato e

$$\hat{\theta} \cong N(\theta, \Sigma)$$

allora, ricordando le proprietà delle forme quadratiche sulla normale, un ellissoide di confidenza di livello $1 - \alpha$ è dato da

$$(\hat{\theta} - \theta)' \Sigma^{-1} (\hat{\theta} - \theta) < \chi_{k,\alpha}^2$$

Esempio sulla media della normale

l'ellissoide di confidenza per la media μ della normale multivariata a livello di confidenza $1 - \alpha$ è dato da

$$n(\bar{\mathbf{x}} - \mu_0)' \Sigma^{-1} (\bar{\mathbf{x}} - \mu_0) < \chi_k^2$$

Regressione Multipla

Consideriamo il Modello Lineare

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon \\ &= \boldsymbol{\beta}' \mathbf{x} + \varepsilon \end{aligned}$$

dove

1. $\mathbf{x} = (x_1, \dots, x_k)'$ variabili esplicative o regressori ed $\mathbf{x} = (1, x')'$;
2. ε è $N(0, \sigma^2)$, σ^2 parametro di disturbo per $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)'$

Esempi

1. Retta di regressione

$$y = \beta_0 + \beta_1 x + \varepsilon$$

2. Modelli senza intercetta, $\beta_0 = 0$

3. Modelli polinomiali: ponendo $x_i = x^i$, $i = 1, \dots, k$

$$y = \beta_0 + \beta_1 x + \dots + \beta_k x^k + \varepsilon$$

Stima a Minimi Quadrati

Date n osservazioni $(x_1, y_1), \dots, (x_n, y_n)$ si ha la forma matriciale,

$$Y = X\beta + \epsilon$$

data da

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

dove

- $Y = (y_1, \dots, y_n)'$ è $n \times 1$
- X è $n \times (k+1)$, ha come i -esima riga $\mathbf{x}_i' = (1, x_{i1}, \dots, x_{ik})$
- ϵ è $n \times 1$

L'espressione di $\hat{\beta}$ si trova dalle $k+1$ condizioni del prim'ordine:

$$\frac{\partial Q}{\partial \beta_j} = -2 \sum x_{ij}'(y_i - \beta' \mathbf{x}_i) = 0$$

o, in forma matriciale,

$$\frac{\partial Q}{\partial \beta} = -2X'(Y - X\beta) = \mathbf{0}.$$

Si ha così il sistema detto delle **eq^{ui} normali**

$$X'X\hat{\beta} = X'Y$$

che ha, appunto, soluzione

$$\hat{\beta}_{LS} = (X'X)^{-1}X'Y.$$

Usando i minimi quadrati

$$Q(\beta) = \sum (y_i - \beta' \mathbf{x}_i)^2 = (Y - X\beta)'(Y - X\beta)$$

definiamo stima **Least Squares (LS)** la soluzione dei minimi quadrati:

$$\hat{\beta}_{LS} = \hat{\beta} = \arg \min Q(\beta).$$

che ha forma esplicita data da

$$\hat{\beta}_{LS} = (X'X)^{-1}X'Y.$$

Esistenza

La condizione $\det(X'X) > 0$ è sempre soddisfatta a meno che una o più colonne della matrice X non sia una combinazione lineare delle altre. Supponiamo, per esempio, che l'ultima colonna sia una tale combinazione:

$$x_{ik} = \sum_{j=1}^{k-1} a_j x_{ij}$$

allora l'osservazione della corrispondente variabile esplicativa x_k non porta informazioni aggiuntive rispetto alle altre per il sistema che si sta studiando e va eliminata dal modello.

Problemi

1. dato $\hat{y} = x\beta + \varepsilon$, con $x \in R^1$ mostrare che $\hat{\beta} = \frac{\sum x_t y_t}{\sum x_t^2}$
2. dato $\hat{y} = \beta_0 + \beta_1 x_1$
 - a. calcolare $X'X$ ed $(X'X)^{-1}$.
 - b. dedurre le formule note per $\hat{\beta}_0$ e $\hat{\beta}_1$
3. dato $f(x; \beta) = \beta_0 + \beta_1 x + \dots + \beta_k x^k$ con $\beta = (\beta_0, \dots, \beta_k)'$: studiare $\hat{\beta}$.
4. Posto $\hat{\Sigma} = (\hat{\sigma}_{ij})$ dimostrare che

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^k \hat{\beta}_j \bar{x}_j$$

e

$$\begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = \hat{\Sigma}^{-1} \begin{pmatrix} \hat{\sigma}(x_1, y) \\ \vdots \\ \hat{\sigma}(x_k, y) \end{pmatrix}$$

Proprietà dei residui**Media nulla:**

$$\sum_{t=1}^n e_t = 0$$

Incorrelazione con i regressori:

$$\sum_{t=1}^n x_{ti} e_t = 0 \quad \forall i = 1, \dots, k$$

Valori interpolati e residuiNoto $\hat{\beta}$ si possono calcolare i **valori interpolati**:

$$\hat{y}_t = \hat{\beta}' x_t$$

che hanno media uguale a quella dei dati

$$\frac{1}{n} \sum \hat{y}_t = \bar{y}.$$

Inoltre si possono calcolare i **residui** della regressione

$$e_t = y_t - \hat{y}_t$$

Proprietà della stima LS**1 - Nondistorsione**

$$E(\hat{\beta}) = \beta$$

2 - Matrice di varianze-covarianze

$$Var(\hat{\beta}) = \sigma_\varepsilon^2 (X'X)^{-1}$$

- $\det(X'X) \geq 0 \Rightarrow$ stime scadenti.

- Posto

$$V = diag((X'X)^{-1})$$

si ha per $j = 0, \dots, k$

$$Var(\hat{\beta}_j) = \sigma_\varepsilon^2 V_{jj}$$

3 - Normalità

• piccoli campioni $\varepsilon \text{ iid } N(0, \sigma^2) \Rightarrow \hat{\beta} \text{ è } N_{k+1}(\beta, \sigma^2(X'X)^{-1})$

• grandi campioni: se

$$\frac{1}{n}X'X \rightarrow \Gamma > 0$$

allora, se $\varepsilon \text{ iid } (0, \sigma^2)$ con $E(\varepsilon^4) < \infty$, vale il seg. TLC:

$$\hat{\beta} \cong N\left(\beta, \frac{\sigma^2}{n}\Gamma^{-1}\right).$$

4 - Multicollinearità

Se $\det(X'X) = 0$ ovvero $X'X$ è singolare i regressori si dicono multicollineari e i dati non hanno informazione sufficiente per il modello che si vuole stimare.

In pratica è rilevante il caso quasi singolare: $\det(X'X) \approx 0$

Esercizio

Scrivere la matrice Γ per i seguenti modelli di regressione:

1. Retta per l'origine: $y = \beta x + \varepsilon$ quando

- a. $E(x) = 0$
- b. $E(x) \neq 0$

2. Retta con intercetta: $y = \alpha + \beta x + \varepsilon$

3. Piano di regressione ($k = 2$): $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

5 - La matrice Γ

Poiché

$$\lim_{n \rightarrow \infty} \frac{1}{n}X'X = \Gamma > 0$$

La matrice

$$\Gamma = (\gamma_{i,j})_{i,j=1,\dots,k+1}$$

è simmetrica e composta di momenti di ordine 0,1,2. Nel caso di modello con l'intercetta vale:

$$\gamma_{ij} = E(x_{i-1}x_{j-1})$$

dove si è posto $x_0 = 1$.

Nel caso senza intercetta, con $E(x_j) = 0$, Γ è la matrice di varianze covarianze di (x_1, \dots, x_k)

Scomposizione della Varianza

In modo analogo alla retta di regressione abbiamo:

Devianza totale

$$D_{tot} = \sum(y_t - \bar{y})^2 \approx \sigma^2 \chi^2_{n-1}$$

Devianza residua

$$D_{res} = \sum(y_t - \hat{y}_t)^2 \approx \sigma^2 \chi^2_{n-k-1}$$

Devianza spiegata

$$D_{sp} = \sum(\hat{y}_t - \bar{y})^2 = D_{tot} - D_{res} \approx \sigma^2 \chi^2_k$$

Adattamento

Varianza Residua, stima di σ^2_ε :

$$s^2 = \frac{1}{n - k - 1} D_{res}.$$

Coefficiente di Determinazione Multipla

$$R^2 = 1 - \frac{D_{res}}{D_{tot}}$$

sotto $H_0 : \beta = 0$, per n grande, nR^2 ha distribuzione approssimata di tipo χ^2_k .

Coefficiente di Correlazione Multipla

$$R = +\sqrt{R^2}$$

Tuttavia quando n non è grande si possono avere R^2 alti come solo effetto di interpolazione. Procediamo perciò a formulare dei test per l'analisi del modello trovato dai minimi quadrati.

Tabella di ANOVA

sorgente	DF	SS	MS	F	p –value
regressione	k	D_{sp}	$MS_{sp} = \frac{D_{sp}}{k}$	$\frac{MS_{sp}}{s^2}$	$P(F_{k,n-k-1} > F)$
errori	$n - k - 1$	D_{res}	$s^2 = \frac{D_{res}}{n-k-1}$		
totale	$n - 1$	D_{tot}	S_y^2		

Analisi della Varianza e Test F

Il modello è significativo ?

Interessa valutare la significatività del modello nel suo insieme:

$$H_0 : \beta_1 = \dots = \beta_k = 0$$

A tal fine usiamo la statistica

$$F = \frac{D_{sp}/k}{D_{res}/(n - k - 1)}.$$

In ipotesi di normalità, sotto H_0 la statistica F ha distribuzione **F di Snedecor** con k ed $n - k - 1$ gradi di libertà

$$F \sim F_{k,n-k-1}$$

Test t sui coefficienti

Interessa valutare la significatività dei singoli coefficienti β_j :

$$H_{0j} : \beta_j = 0$$

si usa la statistica t :

$$t = \frac{\hat{\beta}_j}{s \sqrt{v_{j+1}}}$$

dove

$$v = \text{diag}((X'X)^{-1})$$

e t ha distribuzione t di Student con $n - k - 1$ gradi di libertà.

Intervalli di Confidenza nella regressione

IC sui coefficienti

$$\hat{\beta}_j - t_{n-k-1, \frac{\alpha}{2}} s \sqrt{V_{j+1}} \leq \beta_j \leq \hat{\beta}_j + t_{n-k-1, \frac{\alpha}{2}} s \sqrt{V_{j+1}}$$

dove $t_{n-k-1, \frac{\alpha}{2}}$ è il valore critico della t di Student con $n - k - 1$ gradi di libertà.

Grandi campioni

$$\hat{\beta}_j - z_{\frac{\alpha}{2}} s \sqrt{V_{j+1}} \leq \beta_j \leq \hat{\beta}_j + z_{\frac{\alpha}{2}} s \sqrt{V_{j+1}}$$

IC sulle previsioni

Interessa l'IC per

$$y = \beta' \mathbf{x} + \epsilon$$

in corrispondenza ad una \mathbf{x} non osservata.

La sua stima è

$$\hat{y} = \hat{\beta}' \mathbf{x}$$

osservando che

$$\hat{\beta} = \beta + (X'X)^{-1} X' \epsilon,$$

si ottiene la varianza dell'errore di previsione:

$$Var(y - \hat{y}) = \sigma^2 [1 + \mathbf{x}'(X'X)^{-1} \mathbf{x}]$$

IC sulla superficie attesa

Interessa l'IC per

$$\mu_y(\mathbf{x}) = E(y|\mathbf{x})$$

in corrispondenza ad \mathbf{x} non osservato. La sua stima LS è

$$\hat{y}_y(\mathbf{x}) = \hat{\beta}' \mathbf{x}$$

con varianza

$$Var(\hat{\beta}' \mathbf{x}) = \sigma^2 \mathbf{x}'(X'X)^{-1} \mathbf{x}$$

Perciò l'IC per $\mu_y(\mathbf{x})$ è

$$\hat{\beta}' \mathbf{x} \mp t_{n-k-1, \frac{\alpha}{2}} s \sqrt{\mathbf{x}'(X'X)^{-1} \mathbf{x}}$$

dove $t_{n-k-1, \frac{\alpha}{2}}$ è il valore critico della t di Student con $n - k - 1$ gradi di libertà.

Perciò l'IC per \hat{y} è

$$\hat{\beta}' \mathbf{x} \mp t_{n-k-1, \frac{\alpha}{2}} s \sqrt{1 + \mathbf{x}'(X'X)^{-1} \mathbf{x}}$$

dove $t_{n-k-1, \frac{\alpha}{2}}$ è il valore critico della t di Student con $n - k - 1$ gradi di libertà.

Bibliografia

1. D.C. Montgomery, G.C. Runger & N.F. Hubele (2000): **Engineering Statistics**, Second Edition, John Wiley&Sons, Inc. New York.
2. D.C. Montgomery, G.C. Runger & N.F. Hubele (2004): **Statistica per Ingegneria**, Egea
3. S. Ross (2003) **Probabilità e Statistica per l'Ingegneria e le Scienze**, Apogeo

L02B: Metodi avanzati di regressione

Statistica e Modelli Stocastici

Alessandro Fassó

16 aprile 2012

Argomenti trattati

- ① Influenza della correlazione sulla regressione lineare
- ② Minimi quadrati generalizzati
- ③ Minimi quadrati ponderati
- ④ Modelli localmente lineari e localmente polinomiali
- ⑤ Spline
- ⑥ Reti neurali

Date n osservazioni $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ dove \mathbf{x} è $k - \text{dim}$, il modello di regressione lineare è

$$y_t = \mathbf{x}'_t \beta + \epsilon_t \quad t = 1, \dots, n$$

o, in forma matriciale,

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{\epsilon}$$

Usando i minimi quadrati ordinari (*OLS* o *LS*)abbiamo la forma esplicita della stima di β :

$$\begin{aligned}\beta &= (X'X)^{-1} X'Y \\ &= \left(\sum_t \mathbf{x}_t \mathbf{x}'_t \right)^{-1} \sum_t \mathbf{x}_t y_t\end{aligned}$$

- Esempi di LS nei modelli dinamici:

Modelli ARX : *LLJ example 1.4 e 1.5.*

Nondistorsione asintotica

Sotto ampie ipotesi su y_t ed \mathbf{x}_t si ha che, per $n \rightarrow \infty$,

$$\begin{aligned}\frac{1}{n} X' \mathbf{e} &= \frac{1}{n} \sum_t \mathbf{x}_t e_t \rightarrow \delta = E(\mathbf{x}\mathbf{e}) \\ \frac{1}{n} X' X &= \frac{1}{n} \sum_t \mathbf{x}_t \mathbf{x}'_t \rightarrow \Sigma > 0\end{aligned}$$

e, quindi,

$$\begin{aligned}\hat{\beta} &= \beta + \left(\frac{X'X}{n} \right)^{-1} \frac{X'\mathbf{e}}{n} \\ &\rightarrow \beta + \Sigma^{-1} \delta\end{aligned}$$

Nondistorsione

Sostituendo

$$Y = \mathbf{X}\beta + \mathbf{e}$$

nella formula di $\hat{\beta}$ si ottiene:

$$\begin{aligned}\hat{\beta} &= \beta + (X'X)^{-1} X' \mathbf{e} \\ &= \beta + \left(\sum_t \mathbf{x}_t \mathbf{x}'_t \right)^{-1} \sum_t \mathbf{x}_t e_t\end{aligned}$$

Quando le \mathbf{x} sono deterministiche la condizione di nondistorsione

$$E(\hat{\beta}) = \beta$$

è soddisfatta per ogni n se

$$E(e_t) = 0.$$

- Perciò per la nondistorsione asintotica (e la consistenza) è necessario che

$$\delta = 0$$

- Condizioni sufficienti sono:

- \mathbf{x}_t è non-stocastica
- errore e_t è iid.

1. Influenza della correlazione sulla regressione lineare | Proprietà dei Minimi quadrati ordinari

- Perciò per la nondistorsione asintotica (e la consistenza) è necessario che
$$\delta = 0$$
- Condizioni sufficienti sono:
 - x_t è non-stocastica
 - errore e_t è iid.

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿

Statistica e Modelli Stocastici (II modulo) | L02B: Metodi avanzati di regressione | 16 aprile 2012 | 7 / 39

1. Influenza della correlazione sulla regressione lineare | Minimi quadrati con errori correlati

Conseguenze

Per valutare la distorsione calcoliamo $\delta = E(\mathbf{x}\mathbf{e})$:

$$\begin{aligned}\delta &= E(y_{t-1}(\zeta_t + \zeta_{t-1})) = \\ &= E[(\beta y_{t-2} + \zeta_{t-1} + \zeta_{t-2})(\zeta_t + \zeta_{t-1})] = \sigma_{\zeta}^2\end{aligned}$$

e la stima LS di β è in questo caso distorta anche per $n \rightarrow \infty$:

$$\hat{\beta} \rightarrow \beta + \frac{\sigma_{\zeta}^2}{\sigma_y^2}$$

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿

Statistica e Modelli Stocastici (II modulo) | L02B: Metodi avanzati di regressione | 16 aprile 2012 | 9 / 39

1. Influenza della correlazione sulla regressione lineare | Minimi quadrati con errori correlati

Contro esempio

Consideriamo il caso di autoregressione con errori autocorrelati:

$$\begin{aligned}\mathbf{x}_t &= y_{t-1} \\ \mathbf{e}_t &= \zeta_t + \zeta_{t-1} \\ \zeta_t &\equiv \text{IID } (0, \sigma_{\zeta}^2)\end{aligned}$$

cioè y_t obbedisce all'eq^{ne} ricorsiva

$$y_t = \beta y_{t-1} + \zeta_t + \zeta_{t-1}$$

e supponiamo $|\beta| < 1$.

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿

Statistica e Modelli Stocastici (II modulo) | L02B: Metodi avanzati di regressione | 16 aprile 2012 | 8 / 39

1. Influenza della correlazione sulla regressione lineare | Esercizio Matlab

Esercizio Matlab

Usando la funzione `filter()` simulare $n = 1000$ osservazioni da

$$y_t = \beta y_{t-1} + \zeta_t + \zeta_{t-1}$$

con

$$\begin{aligned}\beta &= 0.9 \\ \sigma_{\zeta}^2 &= 1\end{aligned}$$

e valutare la sovrastima di β via OLS.

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿

Statistica e Modelli Stocastici (II modulo) | L02B: Metodi avanzati di regressione | 16 aprile 2012 | 10 / 39

Se gli errori e_t non sono *iid* usare le stime *OLS* può dare risultati scadenti.

- In talune situazioni saranno sbagliate le varianze e gli IC degli stimatori.
- In altre le stime stesse saranno (anche asintoticamente) distorte

In queste situazioni è importante conoscere la struttura della dipendenza delle y_t per ottenere delle stime (almeno asintoticamente):

- - ▶ nondistorte
 - ▶ con varianze appropriate
 - ▶ ottimali

Le tecniche di stima per dati correlati si riconducono alla massima verosimiglianza (*MLE*) ai minimi quadrati generalizzati o ad altre tecniche (*IV*) che ovviano agli inconvenienti visti.

GLS

- Nel modello di regressione

$$Y = X\beta + \epsilon$$

si è fatta l'ipotesi che gli errori fossero sferici cioè che la matrice di varianze covarianze fosse

$$\text{Var}(\epsilon) = E(\epsilon\epsilon') = \sigma^2 I_n$$

dove I_n è la matrice identica $n - \text{dimensionale}$.

- Più in generale se $\delta = 0$ ma

$$\text{Var}(\epsilon) = V > 0$$

la stima LSE, pur consistente, è subottimale ed ha distribuzione

$$\text{Var}(\hat{\beta}) = (X'X)^{-1} X'VX (X'X)^{-1}$$

Perciò l'eteroschedasticità e la correlazione degli errori influenza gli IC e i test sul modello.

GLS

- Nel modello di regressione

$$Y = X\beta + \epsilon$$

si è fatta l'ipotesi che gli errori fossero sferici cioè che la matrice di varianze covarianze fosse

$$\text{Var}(\epsilon) = E(\epsilon\epsilon') = \sigma^2 I_n$$

dove I_n è la matrice identica $n - \text{dimensionale}$.

- Più in generale se $\delta = 0$ ma

$$\text{Var}(\epsilon) = V > 0$$

la stima LSE, pur consistente, è subottimale ed ha distribuzione

$$\text{Var}(\hat{\beta}) = (X'X)^{-1} X'VX (X'X)^{-1}$$

Perciò l'eteroschedasticità e la correlazione degli errori influenza gli IC e i test sul modello.

Stima in presenza di errori correlati

Si può dimostrare che la stima ottimale (*BLUE*) di β è data dallo stimatore *GLS*:

$$\tilde{\beta} = (X'V^{-1}X)^{-1} X'V^{-1}Y$$

Distribuzione dei GLS

In ipotesi di normalità degli errori, o per $n \rightarrow \infty$, si ha

$$\tilde{\beta} \equiv N\left(\beta, (X'V^{-1}X)^{-1}\right)$$

e tutte le formule per gli IC si modificano in corrispondenza.

- Se gli errori sono incorrelati ma hanno varianze diverse:

$$Var(\mathbf{e})^{-1} = V^{-1} = diag(\omega_1, \dots, \omega_n) = \Omega$$

si ottiene:

$$\begin{aligned}\beta &= (X'\Omega X)^{-1} X'\Omega Y \\ &= \left(\sum_t \mathbf{x}_t \mathbf{x}_t' \omega_t \right)^{-1} \sum_t \mathbf{x}_t y_t \omega_t\end{aligned}$$

- Ciò equivale a cercare l'ottimo di

$$\sum (y_t - \beta' x_t)^2 \omega_t$$

Ossia i valori con varianza maggiore hanno un peso minore

- Se gli errori sono incorrelati ma hanno varianze diverse:

$$Var(\mathbf{e})^{-1} = V^{-1} = diag(\omega_1, \dots, \omega_n) = \Omega$$

si ottiene:

$$\begin{aligned}\beta &= (X'\Omega X)^{-1} X'\Omega Y \\ &= \left(\sum_t \mathbf{x}_t \mathbf{x}_t' \omega_t \right)^{-1} \sum_t \mathbf{x}_t y_t \omega_t\end{aligned}$$

- Ciò equivale a cercare l'ottimo di

$$\sum (y_t - \beta' x_t)^2 \omega_t$$

Ossia i valori con varianza maggiore hanno un peso minore

Modello localmente lineare

Consideriamo uno sviluppo in serie di $m(x)$ vicino ad x_0 :

$$\begin{aligned}m(x) &\cong m(x_0) + m'(x_0)(x - x_0) \\ &= \beta_0(x_0) + \beta_1(x_0)x\end{aligned}$$

La stima di $\beta(x_0)$ viene fatta dando peso diverso alle n osservazioni in relazione alla distanza da x_0 :

$$\hat{\beta}(x_0) = \underset{t=1}{argmin} \sum_{t=1}^n (y_t - \beta(x_0)' x_t)^2 K_\lambda(x_t - x_0)$$

Kernel

Il Kernel è basato sul nucleo K :

$$K_\lambda(x) = \frac{1}{\lambda} K\left(\frac{x}{\lambda}\right)$$

che può avere forma diversa:

- ➊ Rettangolare: $K(x) = \frac{1}{2}$ per $x \in (-1, 1)$, $K(x) = 0$ altrimenti
- ➋ Gaussiano: $K(x) = \phi(x)$ per $x \in R$
- ➌ Quadratico (Epanechnikov): $K(x) = \frac{3}{4}(1 - x^2)$ per $x \in (-1, 1)$, $K(x) = 0$ altrimenti
- ➍ ...

Kernel

Il Kernel è basato sul nucleo K :

$$K_\lambda(x) = \frac{1}{\lambda} K\left(\frac{x}{\lambda}\right)$$

che può avere forma diversa:

- ➊ Rettangolare: $K(x) = \frac{1}{2}$ per $x \in (-1, 1)$, $K(x) = 0$ altrimenti
- ➋ Gaussiano: $K(x) = \phi(x)$ per $x \in R$
- ➌ Quadratico (Epanechnikov): $K(x) = \frac{3}{4}(1 - x^2)$ per $x \in (-1, 1)$, $K(x) = 0$ altrimenti
- ➍ ...

Kernel

Il Kernel è basato sul nucleo K :

$$K_\lambda(x) = \frac{1}{\lambda} K\left(\frac{x}{\lambda}\right)$$

che può avere forma diversa:

- ➊ Rettangolare: $K(x) = \frac{1}{2}$ per $x \in (-1, 1)$, $K(x) = 0$ altrimenti
- ➋ Gaussiano: $K(x) = \phi(x)$ per $x \in R$
- ➌ Quadratico (Epanechnikov): $K(x) = \frac{3}{4}(1 - x^2)$ per $x \in (-1, 1)$, $K(x) = 0$ altrimenti
- ➍ ...

Kernel

Il Kernel è basato sul nucleo K :

$$K_\lambda(x) = \frac{1}{\lambda} K\left(\frac{x}{\lambda}\right)$$

che può avere forma diversa:

- ➊ Rettangolare: $K(x) = \frac{1}{2}$ per $x \in (-1, 1)$, $K(x) = 0$ altrimenti
- ➋ Gaussiano: $K(x) = \phi(x)$ per $x \in R$
- ➌ Quadratico (Epanechnikov): $K(x) = \frac{3}{4}(1 - x^2)$ per $x \in (-1, 1)$, $K(x) = 0$ altrimenti
- ➍ ...

Parametro di lisciamento

λ è il parametro che determina quante osservazioni entrano nella stima di $\beta(x_0)$ e influenza in modo inverso bias e varianza:

- λ grande: $V(\hat{m}_\lambda)$ piccola $|E(m - \hat{m}_\lambda)|$ grande
- λ piccolo: $V(\hat{m}_\lambda)$ grande $|E(m - \hat{m}_\lambda)|$ piccolo

Parametro di lisciamento

λ è il parametro che determina quante osservazioni entrano nella stima di $\beta(x_0)$ e influenza in modo inverso bias e varianza:

- λ grande: $V(\hat{m}_\lambda)$ piccola $|E(m - \hat{m}_\lambda)|$ grande
- λ piccolo: $V(\hat{m}_\lambda)$ grande $|E(m - \hat{m}_\lambda)|$ piccolo

Il parametro λ può essere stimato tramite

- Dataset di validazione
- Tecniche di crossvalidazione:

$$\hat{\lambda} = \arg \min_{\lambda} \widehat{MSE}(\lambda)$$

dove

$$\widehat{MSE}(\lambda) = \frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{m}_{(-i)}(x|\lambda) \right)^2$$

e $\hat{m}_{(-i)}(x|\lambda)$ è la stima di y_i basata sul modello stimato senza la i -esima osservazione.

- La varianza penalizzata:

$$AIC(\lambda) = \log(\hat{\sigma}_e^2) + 1 + \frac{2(\text{tr}(A+1))}{n - \text{tr}(A) - 2}$$

dove la matrice di lisciamento A ($n \times n$) è tale che $\hat{y} = AY$ e generalizza la hat matrix della regressione $\mathcal{A} = X(X'X)^{-1}X'$, mentre $\text{tr}(A)$ generalizza il numero di parametri per il caso nonparametrico.

Il parametro λ può essere stimato tramite

- Dataset di validazione
- Tecniche di crossvalidazione:

$$\hat{\lambda} = \arg \min_{\lambda} \widehat{MSE}(\lambda)$$

dove

$$\widehat{MSE}(\lambda) = \frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{m}_{(-i)}(x|\lambda) \right)^2$$

e $\hat{m}_{(-i)}(x|\lambda)$ è la stima di y_i basata sul modello stimato senza la i -esima osservazione.

- La varianza penalizzata:

$$AIC(\lambda) = \log(\hat{\sigma}_e^2) + 1 + \frac{2(\text{tr}(A+1))}{n - \text{tr}(A) - 2}$$

dove la matrice di lisciamento A ($n \times n$) è tale che $\hat{y} = AY$ e generalizza la hat matrix della regressione $\mathcal{A} = X(X'X)^{-1}X'$, mentre $\text{tr}(A)$ generalizza il numero di parametri per il caso nonparametrico.

Il parametro λ può essere stimato tramite

- Dataset di validazione
- Tecniche di crossvalidazione:

$$\hat{\lambda} = \arg \min_{\lambda} \widehat{MSE}(\lambda)$$

dove

$$\widehat{MSE}(\lambda) = \frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{m}_{(-i)}(x|\lambda) \right)^2$$

e $\hat{m}_{(-i)}(x|\lambda)$ è la stima di y_i basata sul modello stimato senza la i -esima osservazione.

- La varianza penalizzata:

$$AIC(\lambda) = \log(\hat{\sigma}_e^2) + 1 + \frac{2(\text{tr}(A+1))}{n - \text{tr}(A) - 2}$$

dove la matrice di lasciamento A ($n \times n$) è tale che $\hat{y} = AY$ e generalizza la hat matrix della regressione $\mathcal{A} = X(X'X)^{-1}X'$, mentre $\text{tr}(A)$ generalizza il numero di parametri per il caso nonparametrico.

- Estensione a polinomi locali come sviluppo in serie di ordine superiore:

$$m(x) \cong m(x_0) + m'(x_0)(x - x_0) + \frac{1}{2}m''(x_0)(x - x_0)^2$$

$$y = \sum_{j=0}^p \beta_j(x_0)(x - x_0)^j + e$$

- Casi importanti

- $p = 0$: Kernel smoother

- $p = 1, 2$:

- Estensione a polinomi locali come sviluppo in serie di ordine superiore:

$$m(x) \cong m(x_0) + m'(x_0)(x - x_0) + \frac{1}{2}m''(x_0)(x - x_0)^2$$

$$y = \sum_{j=0}^p \beta_j(x_0)(x - x_0)^j + e$$

- Casi importanti

- $p = 0$: Kernel smoother

- $p = 1, 2$:

4. Polinomi locali Polinomi locali

- Estensione a polinomi locali come sviluppo in serie di ordine superiore:

$$m(x) \cong m(x_0) + m'(x_0)(x - x_0) + \frac{1}{2}m''(x_0)(x - x_0)^2$$

$$y = \sum_{j=0}^p \beta_j(x_0)(x - x_0)^j + e$$

- Casi importanti
- $p = 0$: Kernel smoother
- $p = 1, 2$:

Statistica e Modelli Stocastici (II modulo) L02B: Metodi avanzati di regressione 16 aprile 2012 20 / 39

Outline

- Introduzione
- Funzioni di Basis
- Regression spline
- Smoothing spline
- Dati Funzionali
- Regressione lineare funzionale per risposta scalare
- Conclusione

Frank Massoda Spline

Regressione Non Parametrica e Spline

Frank Massoda

6 aprile 2020

Frank Massoda Spline

Introduzione

Lo scopo principale della regressione non parametrica è la modellazione delle effetti anche non lineari di variabili indipendenti sulla variabile dipendente.

In effetti la regressione lineare multipla pur avendo vantaggi in termini di interpretazione e di inferenza, soffre di importanti limiti per quanto riguarda la sua capacità previsiva. Questo è particolarmente vero in quanto si assume un modello lineare per i dati, introducendo così facendo un errore di bias. Al fine di avere miglior capacità previsiva, si rilassa l'assunzione di linearità e si introduce tramite estensioni di modelli lineari, metodi di regressione più flessibili cercando quindi di mantenere più interpretazioni possibili e permettendo di catturare anche relazioni non lineari.

Frank Massoda Spline

Funzioni di basis

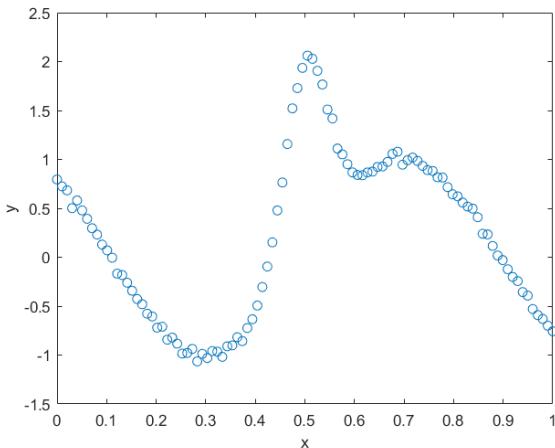
In generale si assume che, la relazione tra variabili dipendente e quelle indipendenti viene modellata con una funzione di covariate più un errore additivo . Per semplicità se si considera una sola covariata abbiamo che

$$y_i = f(x_i) + \epsilon_i$$

assunzioni diversi sulla funzione f porta a possibilità diverse.

Frank Massoda Spline

Funzioni di basis (con't)



Frank Massoda Spline

Funzioni di basis (con't)

Nella regressione lineare

$$f(x_i) = \beta_0 + \beta_1 x_i$$

Se i dati presentano curvatura : regressione polinomiale

$$f(x_i) = \beta_0 + x_i \beta_1 + x_i^2 \beta_2 + \dots + x_i^m \beta_m$$

dove m è il grado del polinomio però di ordine $m+1$.

In generale se si vuole catturare trend più complessi nei dati la funzione f può essere rappresentata come combinazioni lineari di funzioni di basis

$$f(x_i) = \sum_{j=1}^K c_j \phi(x_i)$$

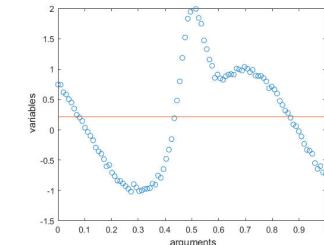
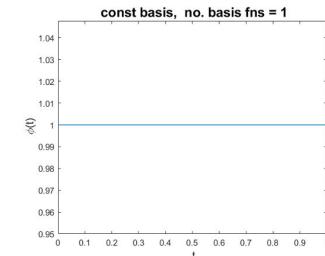
dove $\phi(x_i)$ è una funzione base di f , e c_j un coefficiente di tale base.

Frank Massoda Spline

Funzioni di basis (con't) : Esempio

Basi Monomiali:

costante : $\phi(x) = 1$ package: create_constant_basis(rangevalue)

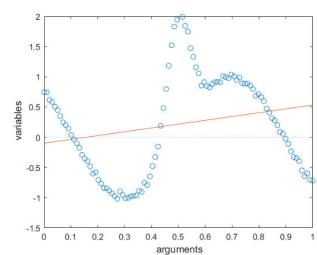
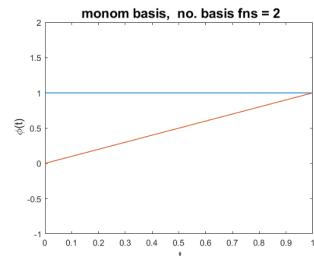


Frank Massoda Spline

Funzioni di basis (con't) : Esempio

Ordine 2 : $\phi(x) = (1, x)$

package: create_monomial_basis(rangevalue, 2)

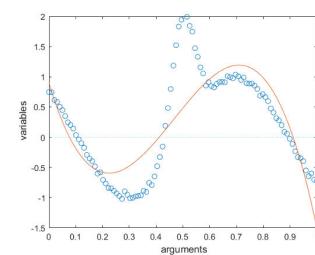
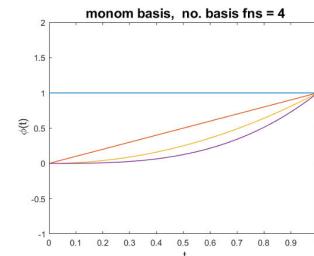


Frank Massoda Spline

Funzioni di basis (con't) : Esempio

Ordine 4 : $\phi(x) = (1, x, x^2, x^3)$

package: create_monomial_basis(rangevalue, 4)

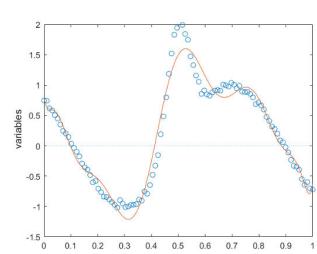
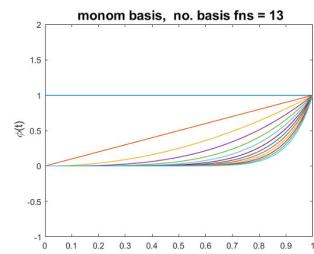


Frank Massoda Spline

Funzioni di basis (con't) : Esempio

Ordine 14: $\phi(x) = (1, x, \dots, x^{13})$

package: create_monomial_basis(rangevalue, 14)



Frank Massoda Spline

Funzioni di basis (con't) : Esempio

. Uno svantaggio delle basi polinomiali è che le funzioni di base sono "non locali", il che significa che il valore stimato di y a un dato valore $x = x_0$ dipende fortemente dai valori dei dati con x lontani da x_0 . Perciò nelle applicazioni reali le funzioni di base monomiali vengono usati raramente, ma possono servire di benchmark per funzioni di basi più potenti come le Spline ed in particolare le B-spline, o le funzione di Fourier.

Frank Massoda Spline

Funzioni di basis (con't) : Esempio

Funzioni di base di fourier: Essi vengono maggiormente usati quando la funzione f si ripete per un certo intervallo T , come per esempio trend saisonali in serie temporali.

Vengono espressi come funzioni di seno e coseno con frequenza crescente nel modo seguente:

$$1, \sin(wt), \cos(wt), \sin(2wt), \cos(2wt), \dots, \sin(mwt), \cos(mwt)$$

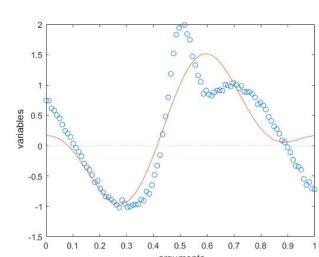
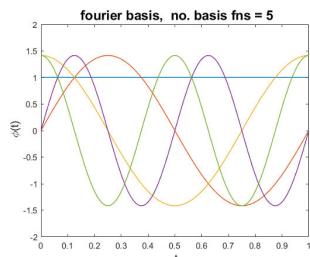
dove $w = \frac{2\pi}{T}$ rappresenta la frequanza di oscillazioni della prima coppia sen/cos.

se la serie contiene ogni elemento di ciascuna coppia di seno/coseno allora il numero di funzioni di base è dato da $K = 2m + 1$

Frank Massoda Spline

Funzioni di basis (con't) : Esempio

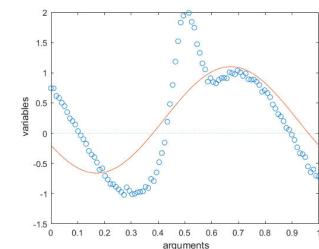
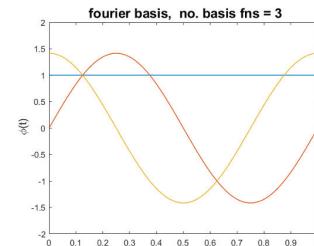
$m = 2$: $\phi(x) = (1, \sin(wx), \cos(wx), \sin(2wx), \cos(2wx))$
package: `create_fourier_basis(rangeval, 5)`



Frank Massoda Spline

Funzioni di basis (con't) : Esempio

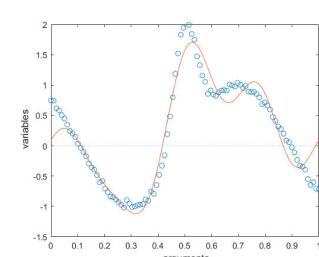
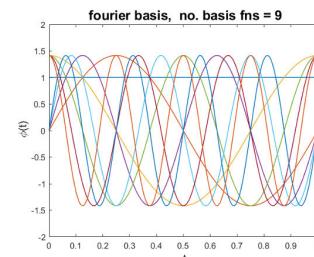
$m = 1$: $\phi(x) = (1, \sin(wx), \cos(wx))$
package: `create_fourier_basis(rangeval, 3)`



Frank Massoda Spline

Funzioni di basis (con't) : Esempio

$m = 4$: $\phi(x) = (1, \sin(wx), \cos(wx), \dots, \sin(4wx), \cos(4wx))$
package: `create_fourier_basis(rangeval, 9)`



Frank Massoda Spline

Funzioni di basis (con't) : Esempio

Le funzioni di basi di fourier hanno buone proprietà computazionali specialmente se le osservazioni sono equidistanti. Sono il "go to" per descrivere dati periodici come il ciclo meteorologico annuale. Ma tanti funzioni non sono periodiche perciò bisogna di un' altro tipo di funzioni base.

Frank Massoda Spline

Funzioni di basis (con't) : Esempio

Il numero e la posizione dei knots così come il grado del polinomio in ogni intervallo hanno un'influenza sulla stima della funzione f . Il Cubic B-Spline(ovvero B-spline di ordine 4) è il default che si viene maggiormente usato in quanto soddisfa le condizioni imposte precedentemente. Un numero di knots più(o meno) alto porta ad una funzione di f più(o meno) flessibile.

Esistono quindi vari metodi per scegliere la posizione dei knots: *equidistanti knots*, *knots basati su quantili*, *knots basati su osservazione del scatter plot*.

Per quanto riguarda il numero di knots, possiamo usare un'approccio di penalizzazione, simile alla regressione ridge

Frank Massoda Spline

Funzioni di basis (con't) : Esempio

Splines

Sono costruiti suddividendo l'intervallo delle osservazioni in sotto-intervalli con punti di suddivisione chiamati *knots*. In ogni dei sotto-intervalli la funzione di spline è una polinomiale di ordine fisso m (*grado + 1*).

Tuttavia in ogni sotto-intervallo le funzioni polinomiali sono costrette ad essere *unite* ai knots, ed essere $m-2$ volte continuamente differenziabile.

Le *B-Spline* sono una categoria particolare di spline per affrontare queste constraints

Frank Massoda Spline

Funzioni di basis (con't) : Esempio

Proprietà delle B-spline

Le *B-spline* formano una base locale: Ogni funzione di base è positiva solo in al più m intervalli adiacenti, questo porta ad una computazione più efficiente anche per miliaia di funzioni di base.

Scomposizione unitaria: Per ogni punto appartenente nel dominio delle covariate abbiamo che $\sum_{j=1}^K \phi_j(x) = 1$

Frank Massoda Spline

Funzioni di basis (con't) : Esempio

Proprietà delle B-spline

Le B-spline formano una base locale: Ogni funzione di base è positiva solo in al più m intervalli adiacenti, questo porta ad una computazione più efficiente anche per migliaia di funzioni di base.

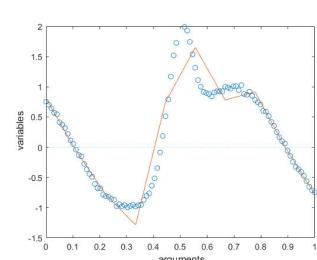
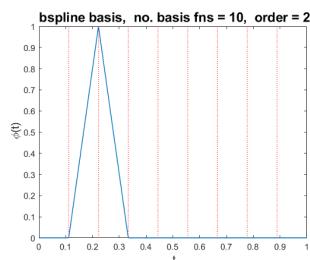
Scomposizione unitaria: Per ogni punto appartenente nel dominio delle covariate abbiamo che $\sum_{j=1}^K \phi_j(x) = 1$

il numero di basis K = order + numero di knots interiori, dove numero di knots interiori = lunghezza(knots) - 2

Frank Massoda Spline

Funzioni di basis (con't) : Esempio

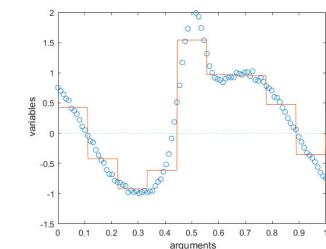
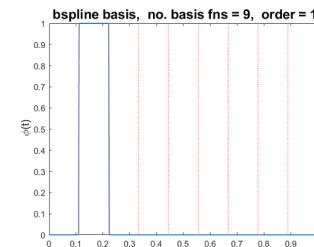
B-spline di ordine 2 con 10 funzione di base e 10 knots equidistanti



Frank Massoda Spline

Funzioni di basis (con't) : Esempio

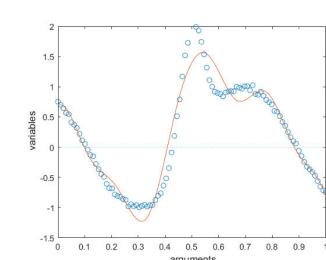
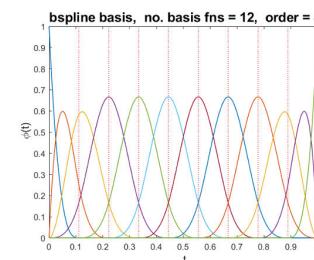
B-spline di ordine 1 con 9 funzioni di base e 10 knots equidistanti
package: `create_bspline_basis(rangeval, nbasis, norder, knots)`



Frank Massoda Spline

Funzioni di basis (con't) : Esempio

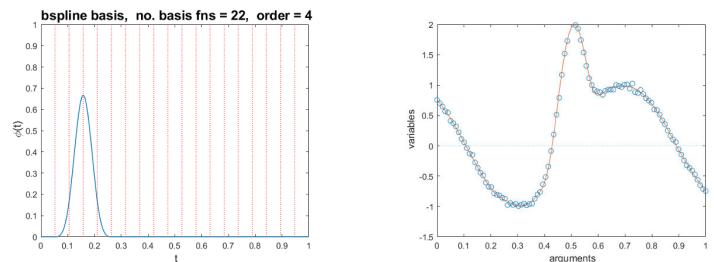
B-spline di ordine 4 con 12 funzioni di basi e 10 knots equidistanti



Frank Massoda Spline

Funzioni di basis (con't) : Esempio

B-spline di ordine 4 con 22 funzioni di basi e 20 knots equidistanti



si può quindi notare come il numero di funzioni base influenzia la stima della funzione f e si riesce a catturare il punto at 0.5 cosa non accaduta con le basi monomiale e anche di fourier.
Però se si usa tanti funzioni di basi c'è possibilità di avere overfitting, mentre se se ne sceglie poco c'è possibilità di avere underfitting. Quindi il numero di funzioni base introduce un trade-off tra bias e varianza

Frank Massoda Spline

Regressione con spline (con't)

la somma residua dei quadrati RSS viene quindi minimizzata usando il metodo dei minimi quadrati. perciò la stima dei coefficienti delle funzioni di base è data da:

$$\hat{\mathbf{c}} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}$$

abbiamo quindi che

$$\hat{\mathbf{y}} = \hat{\mathbf{f}}(\mathbf{x}) = (\hat{\mathbf{c}}^T \Phi^T)^T = \Phi \hat{\mathbf{c}}$$

Come nella regressione lineare si può avere la *matrice di smoothing* che mette in relazioni il vettore delle osservazioni e quelle delle stime

$$\hat{\mathbf{y}} = \Phi (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}$$

$\mathbf{S} = \Phi (\Phi^T \Phi)^{-1} \Phi^T$ è la matrice di *smoothing*, con $\text{trace}(\mathbf{S}) = K$ il numero di funzioni di base, perciò il numero di coefficienti, ovvero il numero di parametri nel modello.

Frank Massoda Spline

Regressione con spline

Come nella regressione lineare per n osservazioni, si vuole minimizzare la somma residua dei quadrati RSS

$$RSS = \sum_{i=1}^n (y_i - f(x_i))^2 = \sum_{i=1}^n (y_i - \sum_{j=1}^K c_j \phi_j(x_i))^2$$

In forma matriciale si può scrivere:

$$RSS = (\mathbf{y} - \Phi \mathbf{c})^T (\mathbf{y} - \Phi \mathbf{c})$$

dove \mathbf{y} è un vettore $nx1$, Φ è una matrice nxK e \mathbf{c} è il vettore dei coefficienti delle funzioni base di dimensione $kx1$

Frank Massoda Spline

Regressione con spline (con't)

Dato che il numero di basi introduce un trade-off tra bias e varianza si può usare la cross-validation per scegliere il numero di funzione base che minimizza il quadrato della somma dei residui. Se assumiamo omoschedasticità degli errori, ovvero che errori hanno varianza costante, una stima non distorta della varianza degli errori è

$$\hat{\sigma}^2 = \frac{RSS}{n - \text{numerodibasi}}$$

Sapendo che

$$\hat{\mathbf{c}} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}$$

possiamo scrivere $\hat{\mathbf{c}} = C_{map} * \mathbf{y}$, dove $C_{map} = (\Phi^T \Phi)^{-1} \Phi^T$

Frank Massoda Spline

Regressione con spline (con't)

Dalla regressione lineare multipla si può scrivere

$$\text{cov}(\hat{\mathbf{c}}) = \sigma^2 C_{map} C_{map}^T = \sigma^2 (\Phi^T \Phi)^{-1}$$

. sostituendo $\hat{\sigma}^2$ nell'equazione precedente si ha che la stima della matrice di varianza-covarianza del vettore dei coefficienti stimati delle funzioni di base è:

$$\widehat{\text{cov}(\hat{\mathbf{c}})} = \hat{\sigma}^2 C_{map} C_{map}^T = \hat{\sigma}^2 (\Phi^T \Phi)^{-1}$$

La stima della matrice di varianza-covarianza dei valori stimati sarà quindi uguale a:

$$\widehat{\text{cov}(\hat{\mathbf{y}})} = \hat{\sigma}^2 \mathbf{S}$$

Frank Massoda Spline

Smoothing spine

In generale vogliamo che il quarato della somma dei residui *RSS* sia piccolo, questo potrebbe accadere se si usa funzione f che interpola tutte le osservazioni. Tale interpolazione porterà renderà la funzione f molto flessibile conducendo così ad overfitting. Quello che si vuole è che l'*RSS* sia piccolo e che f sia smooth.

Una soluzione a tale scopo sarebbe di *penalizzare l'RSS* con l'integrale del quadrato della derivata seconda di f , ovvero

$$RSS + \lambda \int [f''(x)]^2 dx$$

dove λ è il parametro di smoothing. La funzione f che minimizza l'equazione precedente è detta *smoothing spline*.

Frank Massoda Spline

Regressione con spline (con't)

Se assumiamo inoltre che gli errori siano normali allora l'intervallo di confidenza con livello α per un punto fisso nel dominio delle covariate è dato da:

$$\hat{\mathbf{y}} \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{cov}(\hat{\mathbf{y}})}}$$

dove $z_{1-\frac{\alpha}{2}}$ è l' $1 - \frac{\alpha}{2}$ -quantile della distribuzione normale standard. Nota bene : Questo intervallo di confidenza è valida solo per ogni punto nel dominio delle covariate ma non tutti punti presi simultaneamente. Questo intervallo ignora il bias della funzione stimata ma dà un'impressione di quanto bene sia stata stimata la funzione f .

Frank Massoda Spline

Smoothing spine (con't)

Nota: la derivata seconda di una funzione, misura la quantità con la quale la pendenza di una funzione cambia. Ad esempio una linea (una funzione perfettamente smooth) ha pendenza costante perciò derivata seconda nulla.

Prendendo quindi analogia con la linea, si può dire che più la derivata seconda di una funzione è piccola in valore assoluto più la funzione è smooth, mentre più essa è grande in valore assoluto più è ruvida. La derivata seconda è quindi una misura della ruvidità di una funzione.

Prendendo l'integrale come una somma $\lambda \int [f''(x)]^2 dx$ è una misura della curvatura totale della funzione f .

Frank Massoda Spline

Smoothing spine (con't)

$\lambda - > 0$ il termine di penalizzazione non ha effetto.

$\lambda - > \infty$ la funzione sarà perfettamente smooth e quindi una linea retta.

Per valori intermedi di λ , la funzione f cercherà di approssimare i dati di osservazione e allo stesso tempo di essere smooth.

Si può dimostrare che tale funzione f che minimizza l'equazione precedente appartiene alla famiglia delle cubic spline naturali, ovvero spline con knots ad ogni osservazione e $m - 2$ derivate continue, per lo più lineare negli intervalli che si trovano ai bordi ovvero a $[a, k_2]$ e $[k_{n-1}, b]$ dove k_2 è il primo knots interiori e k_{n-1} è l'ultimo knots interiori, a e b sono i bordi del dominio delle covariate.

Frank Massoda Spline

Smoothing spline (con't)

A differenza della regressione con spline in cui il numero dei gradi di libertà df usati per approssimare la funzione f è esattamente il numero delle funzioni basi responsabili del trade-off tra bias e varianza. Diversamente nelle smoothing spline è che è responsabile di tale trade-off tra bias e varianza e abbiamo che

$$df(\lambda) = \text{trace}(\mathbf{S}).$$

se i dati sono periodici si può usare come penalizzazione il *quadrato di un cceleratore armonico* definito da

$$\int (\omega^2 f'(x) + f'''(x))^2$$

Notare che le derivate si possono esprimere come operatori lineari differenziali, e se si usano anche funzioni di base, la matrice di penalizzazione \mathbf{R} può essere semplice da calcolare per casi semplici, o essere risolto con metodi numerici per casi più complessi.

Frank Massoda Spline

Smoothing spine (con't)

facendo analogia alla regressione lineare con penalizzazione di tipo ridge abbiamo che

$$\hat{\mathbf{c}} = (\Phi^T \Phi + \lambda \mathbf{R})^{-1} \Phi^T \mathbf{y}$$

dove \mathbf{R} è la matrice di penalizzazione.

I valori stimati sono espressi da

$$\hat{\mathbf{y}} = \Phi (\Phi^T \Phi + \lambda \mathbf{R})^{-1} \Phi^T \mathbf{y}$$

con $\mathbf{S} = \Phi (\Phi^T \Phi + \lambda \mathbf{R})^{-1} \Phi^T$

Frank Massoda Spline

Smoothing spine (con't)

Essendo la regressione con spline e le smoothing spline dei *metodi di smoothing lineari* è opportuno usare il metodo di cross-validation per scegliere valori che influenzano il trade-off tra bias e varianza con una cross-validation, ovvero con

$$CV = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{f}(x_i)}{1 - s_{ii}} \right)^2$$

dove s_{ii} sono i valori diagonali della matrice \mathbf{S} , ma per data set di ampiezza grandi, il calcolo della matrice \mathbf{S} e dei suoi elementi diagonali potrebbe risultare molto complesso perciò un'alternativa a la cross-validation è la *cross-validation generalizzata* in cui si sostituisce gli s_{ii} con la loro media, ottenendo così:

$$GCV = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{f}(x_i)}{1 - \frac{\text{trace}(\mathbf{S})}{n}} \right)^2$$

Frank Massoda Spline

Dati Funzionali

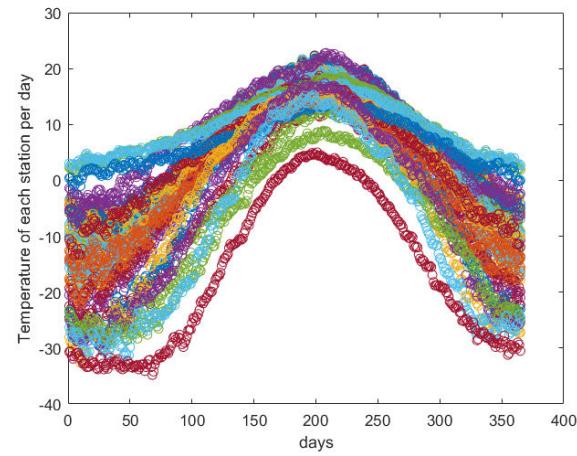
Di solito i dati che vengono analizzati sono dati puntiformi, ovvero ogni osservazione rappresenta un singolo elemento, al contrario i dati funzionali sono dati in cui la singola osservazione rappresenta funzioni(o curve), superficie o qualsiasi altra cosa che varia su un dominio continuo. Di solito il dominio potrebbe essere il tempo, ma anche una locazioni spaziali.

I dati funzionali sono quindi dati che descrivono un processo che cambia in maniera liscia(smooth) e continuo sul suo dominio. Possono includere anche misure ripetute dello stesso processo, quindi ci sono riplichezioni.

Possono essere visti come realizzazioni di un processo stocastico

Frank Massoda Spline

Dati Funzionali (con't)



Frank Massoda Spline

Dati Funzionali (con't)

Caratteristiche:

I dati sono misure di un processo smooth nel suo dominio

Di solito non si fa alcun assunzione parametriche del modello

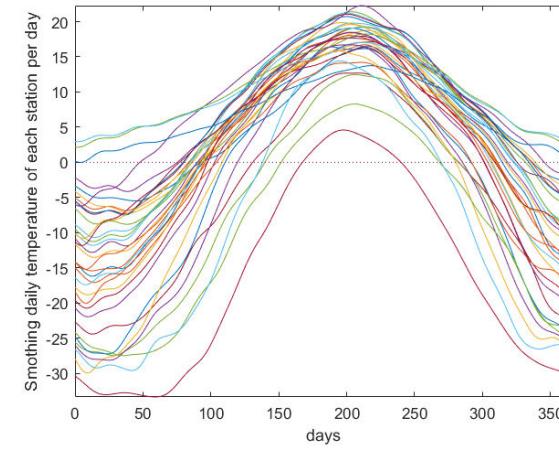
Più misure dello stesso processo

I dati hanno una risoluzione alta e rumore basso e se così non fosse possiamo sempre essere in grado di ricavare il processo smooth che ha generato i dati.

Nota: Nell'analizzare dati funzionali, ogni replicazioni di funzioni viene considerato come una sola osservazione.

Frank Massoda Spline

Dati Funzionali (con't)



Frank Massoda Spline

Dati Funzionali (con't)

Statistica descrittiva funzionale: le statistiche descrittive viste in modulo 1, in particolare , media, varianza e covarianza si applicano anche in questo contesto. Se abbiamo N funzioni(o curve) campionarie da fare addestrare ai dati, le statistiche descrittive sono definite nel modo seguente.

Media

$$\bar{f}(t) = \frac{\sum_i f_i(t)}{N}$$

Varianza

$$var[f(t)] = \frac{\sum_i f_i(t) - \bar{f}(t)^2}{N - 1}$$

Covarianza

$$cov[f(t_1, t_2)] = \frac{(\sum_i f_i(t_1) - \bar{f}(t_1)) \sum_i (f_i(t_2) - \bar{f}(t_2))}{N - 1}$$

Correlazione

$$cor[f(t_1, t_2)] = \frac{cov[f(t_1, t_2)]}{\sqrt{var[f(t_1)]} \sqrt{var[f(t_2)]}}$$

Frank Massoda

Spline

Regressione lineare funzionale per risposta scalare (con't)

Se sostituissimo il vettore di covariate di ogni osservazioni (x_{i1}, \dots, x_{ip}) con una funzione $x_i(t)$, sorge che coefficiente β deve esser anche a sua volta una funzione di tempo in quanto deve poter tener traccia del effetto funzionale delle covariate sulla variabile dipendente.Possiamo quindi scrivere che

$$y_i = \beta_0 + \int \beta(t)x_i(t)dt + \epsilon_i$$

dato che il coefficiente β varia su un dominio continuo, è quindi necessario prendere l'integrale(che rappresenta la somma su intervallo) per mappare le covariate funzionali ad una variabile dipendente scalare.

Frank Massoda

Spline

Regressione lineare funzionale per risposta scalare

Nella regressione lineare multipla abbiamo che

$$y_i = \beta_0 + \sum \beta_j x_{ij} + \epsilon_i$$

vogliamo estendere questo approccio al caso in cui le covariate siano funzionali, e la variabile dipendente scalare.

supponiamo di avere n osservazioni, e che il dominio(covariate) sia tutto un anno con $n < 365$, sorge quindi che applicando il modello di regressione lineare, avremo quindi infiniti insiemi di soluzioni ognuna dei quali dà una predizione perfetta dei dati. Al fine di avere una soluzione unica bisogna $n > p$ con p il numero di covariate.

Frank Massoda Spline

Regressione lineare funzionale per risposta scalare (con't)

Supponendo di avere omoschedasticità, bisogna quindi trovare il $\beta(t)$ che minimizza la somma quadrata dei residui, ovvero

$$\beta(t) = \underset{\beta}{\operatorname{argmin}} \sum \left(y_i - (\beta_0 + \int \beta(t)x_i(t)dt) \right)^2$$

Ma abbiamo solo n osservazioni per determinare $\beta(t)$ che è di dimensione infinita. Come nel caso precedente esiste sempre un insieme di soluzioni tale che $\epsilon_i = 0$ portando così ad overfitting. Anche se espressimo le covariate con funzioni di basi c'è sempre il rischio di avere $n < p$.

Frank Massoda Spline

Regressione lineare funzionale per risposta scalare (con't)

Per risolvere tale problema possiamo :

- . Esprimere $\beta(t)$ come combinazioni lineari di funzioni di basi di dimensione ridotta k tale che $n > k$.

$$\beta(t) = \mathbf{c}^T \phi(t)$$

dove \mathbf{c}^T è un vettore k -dimensionale.

- . Usare un metodo di regolarizzazione con β espressa come combinazione lineari di funzioni base di dimensione alta, ma con penalizzazione, ovvero

$$RSS(\lambda) = \sum \left(y_i - (\beta_0 + \int \beta(t)x_i(t)dt) \right)^2 + \lambda \int \beta''(t)^2 dt$$

che assomiglia molto alle smoothing spline.

Frank Massoda

Spline

Regressione lineare funzionale per risposta scalare (con't)

Notare che si possono includere più covariate funzionali ed anche scalare nel modello rigressione funzionali, avendo così.

$$y_i = \beta_0 + \mathbf{z}_i^T \alpha + \sum_{j=1}^p \int x_{ij}(t)\beta_j(t)dt + \epsilon_i$$

dove \mathbf{z}_i^T è il vettore delle covariate, una penalizzazione può essere usata per ogni $\beta_j(t)$

Frank Massoda

Spline

Regressione lineare funzionale per risposta scalare

possiamo quindi scrivere che

$$y_i = \beta_0 + \int (\phi(t))\mathbf{c}^T x_i(t)dt + \epsilon_i = \beta_0 + \mathbf{x}_i \mathbf{c} + \epsilon_i$$

in forma matriciale avremmo quindi che

$$\mathbf{y} = \mathbf{X}[\beta_0, \mathbf{c}^T] + \epsilon$$

applicando quindi il metodo dei minimi quadrati si ha

$$[\widehat{\beta}_0, \widehat{\mathbf{c}}^T] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

usando la regolarizzazione abbiamo

$$[\widehat{\beta}_0, \widehat{\mathbf{c}}^T] = (\mathbf{X}^T \mathbf{X} + \lambda R)^{-1} \mathbf{X}^T \mathbf{y}$$

Frank Massoda

Spline

Conclusione

Si può dire che esistono vari metodi, per gestire non linearità nei dati, ma questi metodi possono essere visti come estensione del modello classico lineare.

Frank Massoda

Spline

Modelli stocastici

Capitolo 03

Metodi di stima

L03 - Stima

1

Paradigma B

$$X_1, \dots, X_n \equiv f_{n,\theta}(x_1, \dots, x_n)$$

θ identifica univocamente f_n .

NB: Cade l'ipotesi di indipendenza.

Paradigma C

$$X_1, \dots, X_n \text{ iid } F$$

$$\theta = \theta(F)$$

Stima o Stimatore

$$\tilde{\theta} = \tilde{\theta}(X_1, \dots, X_n) = \tilde{\theta}_n$$

NB: include buona parte della stima nonparametrica.

L03 - Stima

3

Richiami sul concetto di stima dei parametri

Il corso riguarda prevalentemente la stima parametrica. Sono tuttavia di interesse anche argomenti di stima nonparametrica come i polinomi locali, le spline, la stima kernel della densità e argomenti collegati.

θ è parametro scalare o k -dimensionale

$\Theta \in \mathbb{R}^k$ spazio dei parametri. Es: retta.

Paradigma A

$$X_1, \dots, X_n \text{ iid } f_\theta(x)$$

θ identifica univocamente f .

NB: è il caso standard della statistica classica

L03 - Stima

2

Normalità asintotica

Per valutare l'incertezza di una stima, calcolare IC ed effettuare test su θ è importante conoscere almeno approssimativamente la distribuzione dello stimatore.

La normalità asintotica ($n \rightarrow \infty$) di $\tilde{\theta}_n$ può essere espressa in 3 modi strettamente legati:

$$\sqrt{n} \frac{\tilde{\theta}_n - a_n}{b_n} \rightarrow N(0, 1) \text{ per } n \rightarrow \infty$$

$$\sqrt{n} (\tilde{\theta}_n - a_n) \rightarrow N(0, b^2) \text{ per } n \rightarrow \infty$$

$$\tilde{\theta}_n \rightarrow N\left(a, \frac{1}{n} b^2\right)$$

L03 - Stima

4

Notazione

ModStoc. aa1920

$$\frac{d}{dx}f(x^*) = \left. \frac{d}{dx}f(x) \right|_{x=x^*}$$

se θ è un vettore $k - \text{dim}$:

$\frac{\partial}{\partial \theta} f_\theta(x)$ è il vettore delle derivate parziali rispetto a $\theta_1, \dots, \theta_k$

$$E_\theta(h(X)) = \int_{-\infty}^{+\infty} h(x)f_\theta(x)dx$$

$$E_{\theta^0}(f_\theta(X)) = \int_{-\infty}^{+\infty} f_\theta(x)f_{\theta^0}(x)dx$$

L03 - Stima

5

NOTE

- θ può essere scalare o vettoriale
- Non serve f in forma esplicita
- I momenti misti di (X, Y) sono definiti da $\mu_{ij} = E(X^i Y^j)$.
- Similmente sono definiti i momenti misti centrati $\bar{\mu}_{ij}$.

L03 - Stima

7

Metodo dei Momenti

Posto

$$\mu_j(\theta) = E_\theta(X^j) \quad \text{e} \quad \boldsymbol{\mu} = (\mu_1, \dots, \mu_k)'$$

se $\exists h$:

$$\theta = h(\boldsymbol{\mu})$$

allora, posto

$$m_j = \frac{1}{n} \sum X_i^j, \quad \bar{m}_j = \frac{1}{n} \sum (X_i - m_1)^j$$

ed

$$\mathbf{m} = (m_1, \dots, m_k)$$

la stima dei momenti di θ è data da

$$\hat{\theta}_{MM} = h(\mathbf{m})$$

L03 - Stima

6

ModStoc. aa1920

Esempio MM1 (media)

$$\theta = E(X) \quad \Rightarrow \quad h(\mu) = \mu_1 \quad \Rightarrow \quad \hat{\theta}_{MM} = \bar{X} = m_1$$

Esempio MM2 (sqm)

$$\theta = \sqrt{Var(X)} \quad \Rightarrow \quad h(\mu) = \sqrt{\mu_2 - \mu_1^2} \quad \Rightarrow \quad \hat{\theta}_{MM} = \sqrt{m_2 - m_1^2} = \sqrt{\bar{m}_2}$$

Esempio MM3 (correlazione)

$$\theta = \rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad \Rightarrow \quad h(\mu) = \frac{\bar{\mu}_{11}}{\sqrt{\bar{\mu}_{20}\bar{\mu}_{02}}} \quad \Rightarrow \quad \hat{\theta}_{MM} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

L03 - Stima

8

Esempio MM4 - $R(0, \theta)$

Sia X_1, \dots, X_n iid $R(0, \theta)$.

da

$$E(X) = \frac{\theta}{2}$$

abbiamo

$$\theta = h(\mu_1) = 2E(X).$$

Perciò

$$\hat{\theta}_{MM} = 2\bar{X}.$$

Normalità Asintotica di MM

Nelle ipotesi aggiuntive:

1. X_1, \dots, X_n iid $E|X^{2k}| < \infty$
2. h sviluppabile in serie vicino a μ con $E|h''| < \infty$

$\check{\theta}_n$ è asintoticamente normale $n \rightarrow \infty$:

$$\check{\theta} \rightarrow N\left(\theta, \frac{1}{n} J V J'\right)$$

dove J è la matrice $p \times k$ Jacobiana di h

$$J = \frac{\partial}{\partial \mu'} h(\mu)$$

e V è la matrice $k \times k$ di varianze covarianze di

$$\mathbf{X} = (X, X^2, \dots, X^k),$$

data da

$$V = V(\mathbf{X}) = E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'.$$

Teoria asintotica MM

Ipotesi

1. X_1, \dots, X_n iid $E|X^k| < \infty$
2. $\mu = (\mu_1, \dots, \mu_k)'$
3. $\mathbf{m}^{(n)} = (m_1, \dots, m_k)'$
4. h è continua in μ

Consistenza

Per $n \rightarrow \infty$

$$\begin{aligned} m_j^{(n)} &\rightarrow \mu_j, j = 1, \dots, k \\ \Rightarrow h(\mathbf{m}^{(n)}) &\rightarrow h(\mu) = \theta. \end{aligned}$$

Traccia dimostrazione (caso scalare)

Premessa: Dal TLC si ha

$$E(X^{2k}) < \infty \quad \Rightarrow \quad m_j^{(n)} \rightarrow N\left(\mu_j, \frac{1}{n}(\mu_{2j} - \mu_j^2)\right)$$

inoltre

$$\mathbf{m}^{(n)} \rightarrow N\left(\mu, \frac{1}{n} V\right)$$

Si sviluppa in serie al prim'ordine $\check{\theta}$:

$$\check{\theta} = h(m) = \theta + (m - \mu)h'(\mu)' + o(m - \mu)$$

e, dalla premessa, si ha

$$\sqrt{n}(h(m^{(n)}) - h(\theta)) \rightarrow N\left(0, V \times (h')^2\right).$$

Esempio MM1 (segue)

Se $\text{Var}(X) < \infty$, allora

$$\bar{x} = m_1 \rightarrow N\left(\mu, \frac{\sigma^2}{n}\right)$$

Esempio MM2 (segue)

Se X_1, \dots, X_n sono iid $N(\mu_1 = 0, \sigma^2 = \mu_2 - \mu_1^2)$, allora la stima MM di σ è:

$$\check{\sigma} = \sqrt{m_2 - m_1^2} \rightarrow N\left(\sigma, \frac{\sigma^2}{2n}\right)$$

Traccia (NB: per semplificare i conti si usa l'ipotesi di cui sopra: $\mu_1 = 0$):

$$V = \text{Var}((X, X^2)) = \text{diag}(\sigma^2, 2\sigma^4)$$

e

$$J' = \left(\frac{-\mu}{\sigma}, \frac{1}{2\sigma} \right) = \left(0, \frac{1}{2\sigma} \right)$$

Matlab su Esempio MM2 (segue)

Illustrare tramite simulazione Monte Carlo la consistenza e la normalità asintotica dello stimatore $\check{\sigma}$ di cui sopra.

Conclusioni MM

Il metodo dei MM da spesso buoni stimatrori e ha il vantaggio della semplicità.

La normalità asintotica consente di fare IC e test in modo semplice

Limiti di MM

- Non sempre è unico
- Cosa si può dire a riguardo di

Ottimalità ???

Metodo della Massima Verosimiglianza

Funzione di verosimiglianza

Paradigma A

Sia $\theta \in R^1$ e X_1, \dots, X_n iid $f_\theta(x)$

$$L(\theta|\mathbf{x}) = \prod_{i=1}^n f_\theta(x_i)$$

Paradigma B

Sia $\theta \in R^k$ e $X_1, \dots, X_n \equiv f_{\theta,n}(\mathbf{x})$

$$L(\theta; \mathbf{x}) = f_{\theta,n}(\mathbf{x})$$

dove

$$f_{\theta,n}(\mathbf{x}) = f_\theta(x_1, \dots, x_n) = f_{1,\theta}(x_1)f_{2,\theta}(x_2|x_1)\dots f_{n,\theta}(x_n|x_1, \dots, x_{n-1})$$

Esempio MLE1 (media)

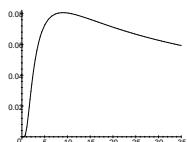
Supponiamo X_i iid $N(\theta, \sigma^2)$, allora, applicando il paradigma A, abbiamo:

$$\begin{aligned} L(\theta) &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2} \sum_i (x_i - \theta)^2\right) \\ &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left[-\frac{1}{2\sigma^2} \left(n(\theta - \bar{x})^2 + \sum_i (x_i - \bar{x})^2\right)\right] \\ &= \frac{K_x}{(\sigma\sqrt{2\pi})^n} \exp\left[-\frac{1}{2\sigma^2/n} (\theta - \bar{x})^2\right] \end{aligned}$$

Esempio MLE2 (varianza)

Posto X_i iid $N(0, \theta = \sigma^2)$, allora, applicando il paradigma A, abbiamo:

$$L(\theta) = \frac{1}{(\sqrt{2\pi\theta})^n} \exp\left(-\frac{\sum x_i^2}{2\theta}\right)$$



$$\hat{\theta} = \frac{1}{n} \sum x_i^2$$

Esercizio: scrivere la verosimiglianza nel caso generico in cui $E(X) = \mu$ non è necessariamente zero.

Esempio MLE3 (AR(1))

Sia $x_i = \theta x_{i-1} + \varepsilon_i$ $t = 1, \dots, n$ con $x_0 = 0$ ed ε_t NID($0, \sigma^2$), allora, applicando il paradigma B, abbiamo:

$$L(\theta|\mathbf{x}) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2} \sum_i (x_i - \theta x_{i-1})^2\right)$$

Infatti

$$x_i|x_{i-1} \equiv N(\theta x_{i-1}, \sigma^2)$$

Stima di Massima Verosimiglianza (MLE)

La MLE di θ è definita da

$$\hat{\theta} = \arg \max_{\theta} L(\theta; x)$$

Se $\theta = \theta^*$ è il "vero" parametro, la MLE $\hat{\theta}$ è una stima di θ^* .

Posto

$$l(\theta) = \log(L(\theta)) = \sum \log(f_{\theta}(X_i))$$

allora

$$\hat{\theta} = \arg \max_{\theta} L(\theta; x) = \arg \min_{\theta} (-l(\theta))$$

Equazione di verosimiglianza

Quando $f_{\theta}(x)$ è "REGOLARE" (cfr oltre) $\hat{\theta}$ è soluzione di

$$\frac{\partial}{\partial \theta} \log L(\theta) = 0$$

se $\theta \in R^k$ si ha un sistema di k equazioni (lineari/non-lineari).

Informazione di Fisher

Una grandezza importante nel seguito è la cosiddetta informazione di Fisher (di una osservazione) che entro ipotesi di regolarità (cfr oltre) è data, indifferentemente, da uno dei 2 segg. valori attesi

$$i_1(\theta) = E_{\theta} \left[\left(\frac{\partial}{\partial \theta} \log f \right)^2 \right] = E_{\theta} \left(-\frac{\partial^2}{\partial \theta^2} \log f \right)$$

$$0 < i_1(\theta) < \infty$$

Nel caso multiparametrico, $\theta = (\theta_1, \dots, \theta_k)$, l'informazione di Fisher $i_1(\theta)$ è una matrice $k \times k$, nonsingolare.

Nel caso di n osservazioni iid l'informazione dell'intero campione è

$$i_n(\theta) = E_{\theta} \left[\left(\frac{\partial}{\partial \theta} \log L_n \right)^2 \right] = n \times i_1(\theta)$$

Esempio MLE1 (segue)

Siano $X_1, \dots, X_n NID(\theta, \sigma^2)$, dimostrare che l'Informazione di Fisher per θ è

$$i_n(\theta) = \frac{n}{\sigma^2}$$

Nota: In questo caso particolare $i(\theta) = i$ non dipende da θ .

Focus: confrontare con teoria primo modulo.

Esempio MLE3 (segue)

Sia $x_i = \theta x_{i-1} + \varepsilon_i$ $i = 1, \dots, n$ con $x_0 = 0$, $|\theta| < 1$ ed $\varepsilon_i \sim NID(0, \sigma^2)$.

Allora (MLE condizionata)

$$\hat{\theta} = \frac{\sum_{i=2}^n x_i x_{i-1}}{\sum_{i=2}^n x_{i-1}^2}$$

inoltre

$$i_n(\theta) = \frac{\sum_{i=2}^n E(x_{i-1}^2)}{\sigma^2} = \frac{(n-1)\sigma_x^2}{\sigma_\varepsilon^2} \cong \frac{n-1}{1-\theta^2}$$

Infatti

$$x_i | x_{i-1} \equiv N(\theta x_{i-1}, \sigma^2)$$

inoltre, vedremo più avanti che, per i grande,

$$\sigma_x^2 \cong \frac{\sigma_\varepsilon^2}{1-\theta^2}$$

Esempio MLE4 (t di Student)

Obiettivo di questo esercizio è mostrare che

1. non sempre ha senso usare la distribuzione normale
2. ogni modello ha una sua tecnica di stima
3. MLE è spesso superiore per n grande
4. Si può usare il metodo MC/Bootstrap per valutare l'incertezza degli estimatori

Tema specifico

Si considerano le variazioni del cambio fra il dollaro e la vecchia lira.

Si vuole stimare media e varianza, fuori dalle ipotesi di normalità, usando una distribuzione con code alte come la t di Student.

Si confronta inoltre la stima MLE con quella tradizionale.

Svolgimento introduttivo

```
load cambio_dollaro_lira.mat
y=diff(t_cambio_dollaro_lira.cambio);
plot(y)
histogram(y)
kurtosis(y)
jbtest(y)
```

Modello specifico

Vista l'alta curtosì usiamo il modello t di Student di locazione e scala dato da

$$y = \sigma t_v + \mu$$

con media e varianza date da

$$E(y) = \mu$$

$$Var(y) = \sigma^2 \frac{v}{v-2}$$

Stimiamo $\theta = (\mu, \sigma, v)$ via MLE:

```
t_hat=fitdist(y,'tLocationScale')
```

Domande

1. l'adattamento con la t di Student è meglio che con la normale?
2. quant'è la varianza delle stime e quali sono gli IC?
3. Questi stimatori di locazione e scala sono diversi da quelli che conosciamo:
 $\hat{\mu} = \bar{x}$ e $\hat{\sigma}^2 = \frac{\sum (x_i - \bar{x})^2}{n-2} = S^2$?
4. Se sono diversi, quali sono migliori?

NB: Si veda il codice cambio_dollaro_lira.m

Proprietà Asintotiche MLE

Le proprietà asintotiche di *MLE* ne fanno lo strumento di stima preferito in numerose situazioni. Specialmente quando n non è piccolo.

In particolare andiamo a considerare le proprietà di:

1. Consistenza
2. Normalità asintotica
3. Ottimalità asintotica

Consideriamo dapprima il caso *iid*

Consistenza

Sotto ampie ipotesi di regolarità su f

$$\hat{\theta}_n \rightarrow \theta^0 \text{ q.c.} \quad \text{per } n \rightarrow \infty$$

Dimostrazione consistenza (opzionale)**Premessa**

Per comprendere questo risultato ricordiamo la seg. diseguaglianza di Jensen:

Diseguaglianza di Jensen

Se g ha la concavità rivolta verso il basso allora

$$E(g(X)) < g(E(X)).$$

Dimostrazione della Consistenza (segue)

Più che una dimostrazione formale (cfr Wald o Serfling) vediamo un'argomentazione euristica.

Da Jensen segue che, $\forall \theta \neq \theta^0$,

$$E_{\theta^0} \left(\log \frac{f_\theta}{f_{\theta^0}} \right) < \log E_{\theta^0} \left(\frac{f_\theta}{f_{\theta^0}} \right) = 0$$

perciò,

$$E_{\theta^0} \log f_\theta < E_{\theta^0} \log f_{\theta^0}.$$

e quindi

$$\theta^0 = \arg \max_{\theta} E_{\theta^0} \log f_\theta.$$

Inoltre da SLLN $\frac{1}{n} \log L_n(\theta) \rightarrow E_{\theta^0} \log f_\theta$ q.c perciò

$$\hat{\theta}_n = \arg \max \hat{E}_{\theta^0} \log f_\theta \cong \theta^0.$$

Normalità Asintotica MLE

Sotto ampie ipotesi (cfr oltre),
per $n \rightarrow \infty$ si ha il seguente limite in distribuzione

$$\sqrt{n} (\hat{\theta} - \theta^0) \rightarrow N \left(0, \frac{1}{i_1(\theta^0)} \right)$$

Traccia dimostrazione TLC

Consideriamo lo sviluppo in serie di $l(\hat{\theta})'$ intorno a θ^0

$$0 = l(\hat{\theta})' = l(\theta^0)' + (\hat{\theta} - \theta^0) l(\theta^*)''.$$

Ora vale il seg. TLC

$$\frac{1}{n} l(\theta^0)' = \frac{1}{n} \sum \frac{\partial}{\partial \theta} \log f_{\theta^0}(X_i) \rightarrow N \left(0, \frac{i(\theta^0)}{n} \right)$$

inoltre

$$\frac{1}{n} l(\theta^*)'' \rightarrow -i(\theta^0)$$

perciò

$$\sqrt{n} (\hat{\theta} - \theta^0) \cong -\sqrt{n} \frac{l(\theta^0)'}{l(\theta^0)''} \cong \sqrt{n} \frac{\frac{1}{n} l(\theta^0)'}{i(\theta^0)} \rightarrow N \left(0, \frac{1}{i(\theta^0)} \right)$$

Ipotesi di regolarità

Discutiamo, dapprima in ambito *IID*, le ipotesi di regolarità **sufficienti** affinchè valga il TLC e le approssimazioni di cui sopra (e.g. Serfling):

A1 Dominio di X è θ -indipendente

A2 $\frac{\partial^j}{\partial \theta^j} \log f_\theta(x) \exists \forall x, j = 1, 2, 3;$

A3 $\left| \frac{\partial^j}{\partial \theta^j} \log f_\theta(x) \right| \leq g_j(x) \forall \theta, j = 1, 2$ con

$$\int_{-\infty}^{+\infty} g_j(x) dx < \infty$$

A4 $\left| \frac{\partial^3}{\partial \theta^3} \log f_\theta(x) \right| \leq H(x)$ con $E_\theta H(X) < \infty$ per $\theta \cong \theta^0$;

A5 $0 < -E_\theta \left(\frac{\partial^2}{\partial \theta^2} \log f_\theta(X) \right) = i(\theta) < \infty$, per $\theta \cong \theta^0$.

Dalle assunzioni segue che

- dalla 2: $\log L$ è sviluppabile
- da 1-3: $\partial \int = \int \partial$ due volte
- da 3 e 4: Trascurabilità (in *m.g.*) del resto dello sviluppo
- da 5: la Var asintotica di $\hat{\theta}$, $nVar(\hat{\theta})$, è finita e non-degenera.

ModStoc. aa1920

L03 - Stima

37

-A2,3,4- Assunzioni di regolarità di f

Le prime 4 sono assunzioni "tecniche" sulla distribuzione di probabilità e giocano un ruolo di condizioni "sufficienti".

In particolare la 3 e la 4 possono essere rilassate in comuni casi particolari.

Contro-eSEMPIO

Si noti per esempio che la 4 non è soddisfatta da casi standard come

$$X_i \text{ iid } N(0, \theta)$$

in cui

$$\frac{\partial^3}{\partial \theta^3} \log f_\theta(x) = -\frac{1}{\theta^3} + \frac{3x^2}{\theta^4} \rightarrow \infty \text{ per } \theta \rightarrow 0$$

ma, come ben noto,

$$\hat{\theta} = \frac{1}{n} \sum X_i^2 \rightarrow N\left(\theta, \frac{2\theta^2}{n}\right).$$

L03 - Stima

39

-A1- Dominio θ -indipendente

Se il dominio (o supporto) della X dipende da θ i risultati asintotici non valgono in modo automatico.

Esempio - $R(0, \theta)$

$$X_1, \dots, X_n \text{ iid } R(0, \theta)$$

$$\hat{\theta}_{MLE} = \max(X_1, \dots, X_n)$$

Esercizio per casa

Studiare e confrontare nondistorsione ed efficienza di $\hat{\theta}_{MLE}$ e $\hat{\theta}_{MM} = 2\bar{X}$ tramite metodo Monte Carlo. Per $\hat{\theta}_{MM}$ si suggerisce di confrontare i risultati Monte Carlo con quelli ottenuti per via analitica.

L03 - Stima

38

ModStoc. aa1920

-A5- Assunzione di nonsingolarità

L'ipotesi 5 sulla finitezza e non-nullità dell'informazione $i(\theta)$ è il punto più importante per il modellista.

Infatti questa assunzione può essere influenzata dalla particolare parametrizzazione e come abbiamo visto determina la varianza delle stime

$$Var(\hat{\theta}) \cong i_n(\theta)^{-1}$$

Si nota che nel caso multiparametrico in cui θ è k -dim allora $i(\theta)$ è una matrice $k \times k$ e con l'assunzione 5 è finita e non singolare.

L03 - Stima

40

-A0- Indipendenza

L'ipotesi *iid* non è irrinunciabile, valgono CLT per serie storiche e processi stocastici dove le ipotesi di regolarità vengono espresse in termini di distribuzioni condizionate

$$l_i(\theta) = \log f_\theta(x_i | x_1, \dots, x_{i-1})$$

e l'indipendenza viene sostituita dalla più debole indipendenza asintotica (mixing condition) del tipo

$$\text{cov}(l_j, l_{j+m}) \rightarrow 0 \text{ per } m \rightarrow \infty$$

e

$$\frac{1}{n} \sum l_i'' \rightarrow I > 0.$$

Efficienza Asintotica MLE

Per comprendere le proprietà di ottimalità di *MLE* utilizziamo l'informazione di Fisher e la disuguaglianza di Rao-Cramer.

Disuguaglianza di Rao-Cramer

In ipotesi di regolarità, un qualsiasi stimatore nondistorto

$$Et_n = \theta$$

è tale per cui

$$\text{Var}(t_n) \geq \frac{1}{i_n(\theta)}$$

- Tale limite è detto *MVB*

- Segue che uno stimatore t_n^* è "ottimale" se raggiunge il *MVB* :

$$\text{Var}(t_n^*) = i_n(\theta)^{-1}$$

MLE è asintoticamente MVB

Abbiamo visto che

$$n\text{Var}(\hat{\theta}) \rightarrow \frac{1}{i_1(\theta^\circ)}.$$

L'ottimalità è perciò "asintotica" nel senso che, per $n \rightarrow \infty$, $\hat{\theta}$ è nondistorto e la sua varianza raggiunge il limite inferiore di Rao-Kramer.

**Soluzione numerica
equazione di verosimiglianza**

L'Algoritmo è basato sullo sviluppo del prim'ordine di $I(\theta)' = \frac{\partial}{\partial \theta} \log L$ o, equivalentemente, su un sviluppo del second'ordine di $\log L$.

Per semplicità consideriamo θ scalare:

$$0 = I(\hat{\theta})' \cong I(\theta)' + (\hat{\theta} - \theta)I(\theta)''$$

da cui si ha il passo

$$\theta_1 = \theta_0 - \frac{I(\theta_0)'}{I(\theta_0)''}.$$

Posto

$$\delta_i = \theta_i - \theta_{i-1}$$

si ha la seguente procedura iterativa.

Algoritmo

Si scelgono i valori iniziali θ_0 , poi, per $i = 1, 2, \dots$, si iterano i segg. passi:

1. Si pone

$$\delta_i = -\frac{l(\theta_{i-1})'}{l(\theta_{i-1})''}$$

2. Si sceglie λ in modo che

$$l(\theta_{i-1} + \lambda \delta_i) > l(\theta_{i-1})$$

3. Si pone

$$\theta_i = \theta_{i-1} + \lambda \delta_i$$

Il loop si arresta quando, scelti ϵ e $\tau > 0$ piccoli,

$$|\theta_i - \theta_{i-1}| < \epsilon(|\theta_{i-1}| + \tau)$$

e

$$|l(\theta_i) - l(\theta_{i-1})| < \epsilon(l(\theta_{i-1}) + \tau)$$

NB: a convergenza avvenuta

$$l(\hat{\theta})' = 0$$

inoltre

$$-\frac{1}{n} l(\hat{\theta})'' = -\frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log(f(x_i; \hat{\theta})) \approx \frac{1}{n} \sum_{i=1}^n \left(\frac{\partial}{\partial \theta} \log(f(x_i; \hat{\theta})) \right)^2$$

sono stime consistenti di $i_1(\hat{\theta})$.

Perciò

$$Var(\hat{\theta}_n) \cong \frac{i_1(\hat{\theta})^{-1}}{n} \cong (l_n(\hat{\theta}_n)')^{-1}.$$

Caso vettoriale

Quando abbiamo k parametri

- θ è un vettore colonna $k - \text{dim}$,
- $l' = \frac{\partial}{\partial \theta} \log L_n$ è un vettore colonna $k - \text{dim}$,
- $l'' = \frac{\partial^2}{\partial \theta \partial \theta'} \log L_n$ è una matrice $k \times k$ e
- l'incremento δ è un vettore $k - \text{dim}$ e lo step 1 dell'algoritmo è dato da

$$\delta = (l(\theta)')^{-1} l(\theta)'.$$

Scelta dei valori iniziali

- informazioni a priori
- modelli collegati più semplici
- metodo dei momenti
- minimi quadrati.

Esempi

Nel capitolo sulle serie storiche useremo questi algoritmi per la stima dei modelli ARMA (vedi anche Shumway-Stoffer paragrafi: 3.1-3.3 e 3.6).

Algoritmo EM

In presenza di dati mancanti scrivere la verosimiglianza delle osservazioni effettive può essere un problema.

Supponiamo che la matrice dei dati "completa" Z sia nella forma

$$Z = (X, Y)$$

dove X sono dati mancanti e Y sono le osservazioni vere e proprie (dette parziali o incomplete).

Esempio

- X_t è lo stato del sistema al tempo t
- Y_t è l'osservazione affetta da errore, e.g.

$$Y_t = X_t + \varepsilon_t$$

L'algoritmo EM è utile quando:

1. la verosimiglianza completa

$$L_\theta(Z) = L_\theta(X, Y)$$

ha una forma analitica semplice da scrivere e da massimizzare per cui, se X fosse noto, sarebbe facile avere

$$\hat{\theta}_{MLE}(Z).$$

2. Inoltre la verosimiglianza delle osservazioni effettive, data da

$$L_\theta = L_\theta(Y) = \int L_\theta(Z)dX$$

$$= E_X(L_\theta(Z|X)),$$

è difficile da scrivere come funzione di θ e/o da massimizzare.

3. Infine la log-verosimiglianza condizionata

$$L_\theta(Z|Y)$$

è "facilmente trattabile".

In particolare la quantità

$$Q(\theta|\theta') = E_{\theta'}(\log L_\theta(Z)|Y)$$

vista come funzione di θ per fissato θ' è facilmente calcolabile e massimizzabile.

Si ragiona quindi iterativamente e l'algoritmo *EM* è dato dall'iterazione di coppie di Expectation e Maximization. In particolare l' i -esima iterazione ha il passo "E" dato da

E-step

$$Q(\theta|\theta_i) = E_{\theta_i}(\log L_\theta(Z)|Y),$$

ed il passo M è dato da:

M-step

$$\hat{\theta}^{(i+1)} = \arg \max_{\theta} Q(\theta|\theta^{(i)}).$$

Monotonicità

L'aspetto interessante è che, entro ipotesi di identificabilità, ad ogni passo, la "vera log-verosimiglianza", cioè quella delle osservazioni incomplete, cresce:

$$\log L_{\theta_{i+1}}(Y) > \log L_{\theta_i}(Y).$$

Perciò

$$\hat{\theta}^{(i)} \rightarrow \hat{\theta}_{MLE}(Y) = \arg \max_{\theta} L_{\theta_i}(Y). \text{ per } i \rightarrow \infty$$

Dimostrazione monotonicità (opzionale)

Osservando che

$$L(Z|Y) = \frac{L(Z)}{L(Y)}$$

abbiamo che

$$\log L(Y) = \log L(Z) - \log L(Z|Y)$$

Facciamo ora il valore atteso di $\log L(Y)$ rispetto alla distribuzione a posteriori $L(Z|Y)$ per un fissato θ^0 :

$$\begin{aligned} \log L_\theta(Y) &= E_{\theta^0}(\log L_\theta(Y)|Y) \\ &= E_{\theta^0}(\log L_\theta(Z)|Y) - E_{\theta^0}(\log L_\theta(Z|Y)|Y) \\ &= Q(\theta|\theta^0) - H(\theta|\theta^0) \end{aligned}$$

Dalla dimostrazione della consistenza di *MLE* sappiamo che:

$$\theta^0 = \arg \max_{\theta} H(\theta|\theta^0)$$

Pertanto, posto

$$\theta' = \arg \max_{\theta} Q(\theta|\theta^0),$$

abbiamo che

$$\begin{aligned} \log L_{\theta'}(Y) &= Q(\theta'|\theta^0) - H(\theta'|\theta^0) \\ &\geq Q(\theta^0|\theta^0) - H(\theta^0|\theta^0) \\ &= \log L_{\theta^0}(Y) \end{aligned}$$

con diseguaglianza stretta per $\theta' \neq \theta^0$ se H e Q sono (localmente) unimodali.

Nota n.1 - modelli gerarchici

Se il problema è separabile , cioè:

$$\theta = (\pi, \psi)$$

e

$$L_\theta(Z) = L\pi(X)L\psi(Y|X)$$

allora l'ottimizzazione di Q è semplificata dal fatto che

$$\begin{aligned} Q(\theta|\theta^0) &= E_{\theta^0}(\log L_\theta(Z)|Y) \\ &= E(L_\psi(Y|X)|Y) + E(L_\pi(X)|Y) \\ &= Q_1(\psi|\theta^0) + Q_2(\pi|\theta^0) \end{aligned}$$

Nota n.2 - forma semplificata

In numerose situazioni applicative dove essendo noto X si stima θ tramite lo stimatore in forma esplicita

$$\hat{\theta}_{MLE} = \hat{\theta}_{MLE}(g(X))$$

si ha che l'iterazione dell'EM è data da

$$\hat{\theta}^{(i)} = \hat{\theta}_{MLE}(E_{\hat{\theta}^{(i-1)}}(g(X)|Y))$$

perciò l'EM si riduce al calcolo ripetuto della media a posteriori

$$E_{\hat{\theta}^{(i-1)}}(g(X)|Y)$$

Esempio: EM per misture

Sia

$$Y|X \equiv N(\mu_X, \sigma^2)$$

ed

$$X \equiv Bin(n = 1, \pi = \pi_1).$$

Dedurre il passo i -esimo dell'algoritmo EM per stimare i parametri $\theta = (\pi_1, \mu_0, \mu_1, \sigma_0^2, \sigma_1^2)$ utilizzando un campione

$$\mathbf{Y} = (Y_1, \dots, Y_n)$$

dalla marginale Y .

A tal fine indichiamo con $\phi(x)$ la pdf della $N(0, 1)$ e poniamo $\pi_0 = 1 - \pi_1$.

Soluzione

Osserviamo innanzitutto che la verosimiglianza dei dati Y è basata sulla mistura di due gaussiane:

$$f(y|\theta) = \pi_1 \frac{1}{\sigma_1} \phi\left(\frac{y - \mu_1}{\sigma_1}\right) + (1 - \pi_1) \frac{1}{\sigma_0} \phi\left(\frac{y - \mu_0}{\sigma_0}\right)$$

e non è facile da massimizzare.

Osserviamo inoltre che il problema è separabile nella parametrizzazione estesa $\theta = (\pi, \psi)$, con $\psi = (\mu_0, \mu_1, \sigma_0^2, \sigma_1^2)$

$$\log L_\theta(Z) = \log L_\pi(X) + \log L_\psi(Y|X)$$

Perciò ci concentriamo su $\log L_\psi(Y|X)$.

Naturalmente, quando X è noto, abbiamo

$$-2\sigma^2 \log L_\psi(\mathbf{Y}|\mathbf{X}) = \sum_{t=1}^n (y_t - \mu_0)^2 I(x_t = 0) + \sum_{t=1}^n (y_t - \mu_1)^2 I(x_t = 1)$$

In tal caso, per $x = 0, 1$, abbiamo

$$\hat{\mu}_x = \frac{1}{n_x} \sum_{t=1}^n y_t I(x_t = x) = \bar{y}_x$$

dove $n_x = \sum_{t=1}^n I(x_t = x)$.

Osserviamo che, per $x = 0, 1$, si ha

$$P_\theta(X = x|Y = y) = \frac{\phi(\frac{y-\mu_x}{\sigma})\pi_x}{\phi(\frac{y-\mu_0}{\sigma})\pi_0 + \phi(\frac{y-\mu_1}{\sigma})\pi_1}$$

Posto dunque, per $x = 0, 1$,

$$\pi_x(y, \theta) = P_\theta(X = x|Y = y)$$

abbiamo che

$$\begin{aligned} -2Q(\theta|\theta^{(i-1)}) &= -2E(\log L(Y|X)|Y) \\ &= \frac{1}{\sigma_0^2} \sum_{t=1}^n (y_t - \mu_0)^2 \pi_0(y_t, \theta^{(i-1)}) + \frac{1}{\sigma_1^2} \sum_{t=1}^n (y_t - \mu_1)^2 \pi_1(y_t, \theta^{(i-1)}) \end{aligned}$$

e l' i -esima iterazione dell'algoritmo EM è data dalla seguente soluzione WLS:

$$\mu_x^{(i)} = \frac{\sum_{t=1}^n y_t \pi_x(y_t, \theta^{(i-1)})}{\sum_{t=1}^n \pi_x(y_t, \theta^{(i-1)})}, \text{ per } x = 0, 1$$

Inoltre è facile verificare che

$$\sigma_x^{2(i)} = \frac{\sum_{t=1}^n (y_t - \mu_x^{(i)})^2 \pi_x(y_t, \theta^{(i-1)})}{\sum_{t=1}^n \pi_x(y_t, \theta^{(i-1)})}, \text{ per } x = 0, 1$$

e

$$\pi_1^{(i)} = \frac{\sum_{t=1}^n \pi_1(y_t, \theta^{(i-1)})}{n}.$$

NB:

Con riferimento alla forma semplificata della nota n.2, se X è noto allora

$$\hat{\mu}_x(g_x(X)) = \frac{\sum_{t=1}^n y_t I(x_t = x)}{\sum_{t=1}^n I(x_t = x)}, \text{ per } x = 0, 1$$

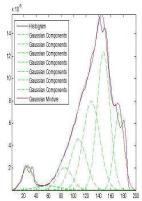
in questo caso posto

$$g_x(X) = (I(x_t = x), t = 1, \dots, n), \text{ per } x = 0, 1$$

la formula dell'EM della pagina precedente deriva da

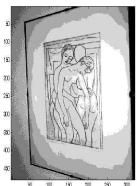
$$\pi_x(y_t, \theta^{(i-1)}) = E_{\theta^{(i-1)}} I(x_t = x), \text{ per } x = 0, 1$$

Applicazione alle immagini



L03 - Stima

ModStoc. aa1920



65

Modello nello spazio degli stati

Vedi LEZ-07

L03 - Stima

ModStoc. aa1920

Stima Bayesiana

In taluni contesti lo stesso parametro θ è considerato una VC, poniamo con densità

$$g(\theta).$$

In queste situazioni le informazioni legate alle osservazioni "sperimentali" su θ riassunte dalla verosimiglianza

$$L(\theta) = f(x_1, \dots, x_n | \theta)$$

vengono integrate con le informazioni "a priori" su θ date, appunto, da $g(\theta)$ dando luogo, tramite il teorema di Bayes, alla cosiddetta "distribuzione a posteriori"

$$h(\theta | x_1, \dots, x_n) = \frac{g(\theta)L(\theta)}{\int_{-\infty}^{+\infty} L(\theta)g(\theta)d\theta}.$$

L03 - Stima

67

ModStoc. aa1920

Stima MAP

L'equivalente di MLE è ora la stima di massima probabilità a posteriori (MAP), cioè la moda di h oppure la media di h .

L03 - Stima

ModStoc. aa1920

68

Esempio

Se

$$x_i \text{ iid } N(\theta, \sigma^2)$$

e

$$\theta \equiv N(\mu_0, \sigma_0^2)$$

allora si può mostrare che (cfr Cifarelli e Muliere, es. 36.1)

$$(\theta|x_1, \dots, x_n) \equiv N\left(\frac{\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}\right)$$

Aspetti di calcolo

In talune semplici situazioni h ha forma esplicita nota ed il calcolo della sua moda o della sua media è semplice.

In altre situazioni h non è nota in forma esplicita e il calcolo del denominatore della formula di Bayes, dato da

$$\int_{-\infty}^{+\infty} L(\theta)g(\theta)d\theta,$$

può diventare lo scoglio computazionale di questo approccio.

Bibliografia

1. Massima verosimiglianza:

- Ross (2003) Probabilità e Statistica per l'ingegneria e le scienze, Apogeo. Cap 7.2.
- Azzalini A. (2004) Inferenza statistica, Springer, §§ 2.1, 2.2 e 3 fino a 3.3.4

2. Algoritmo EM e DLS

- Shumway R.H. and Stoffer D.S. (2006) *Time series analysis and its applications with R examples, 2nd edn*. Springer, Capitolo 6, in particolare §63
- APPROFONDIMENTI: McLachlan G.J., Krishnan T. (1997) The EM Algorithm and Extensions, Wiley

3. Stima bayesiana:

Ross (2003) Probabilità e Statistica per l'ingegneria e le scienze, Apogeo. Cap 7.8, pp269-274. (si consiglia di guardare qualche problema, pp 274 e segg.)

Statistica e Modelli Stocastici Modulo 2 - L04

Simulazione Stocastica: Monte Carlo e Bootstrap

Alessandro Fassò

aa 2019/20

Introduzione

Si considerano qui alcuni metodi di simulazione di fenomeni stocastici.

Il metodo Monte Carlo in particolare è utilizzato per stimare l'incertezza di un *sistema complesso* quando è nota la sua dinamica.

(In particolare avete già visto i test Monte Carlo)

Introduzione

Si considerano qui alcuni metodi di simulazione di fenomeni stocastici.

Il metodo Monte Carlo in particolare è utilizzato per stimare l'incertezza di un *sistema complesso* quando è nota la sua dinamica.

(In particolare avete già visto i test Monte Carlo)

Il metodo Bootstrap invece è una tecnica di simulazione condizionata che prevede l'uso di osservazioni reali come base delle simulazioni.

Introduzione

Entrambi questi metodi trovano utilizzo all'interno dell'inferenza statistica per valutare l'incertezza delle stime o per eseguire dei test.

Ma trovano anche utilizzo nell'applicazione della statistica, per esempio nell'analisi del rischio.

Origine del metodo Monte Carlo

Calcolo approssimato di un integrale

Si vuole calcolare

$$A = \int_0^1 h(u) du$$

dove la funzione h ha una forma complessa o è nota solo per via numerica per cui è difficile/impossibile calcolare la sua primitiva ma h è facilmente computabile per ogni μ .

Origine del metodo Monte Carlo

Posso usare un campione simulato di m variabili casuali $R(0, 1)$:

U_1, \dots, U_m iid $R(0, 1)$

e stimare l'integrale A tramite

$$\bar{h}_m = \frac{1}{m} \sum_{j=1}^m h(U_j)$$

Origine del metodo Monte Carlo

Dalla teoria dell'inferenza statistica abbiamo che lo stimatore è corretto:

$$E(\bar{h}_m) = E(h) = A$$

E' il metodo "migliore" ?

Monte Carlo in generale

Ci interessa la distribuzione $G(z)$ di

$$z = h(y)$$

dove:

- h è nota solo per via numerica (computabile) ma non (facilmente) conoscibile per via analitica (es: computer models);
 - y è un vettore stocastico $k - \dim$ con distribuzione $F(y)$;
 - la funzione di ripartizione F è nota e siamo capaci di simulare replicazioni di y

Monte Carlo in generale

Ci interessa la distribuzione $G(z)$ di

$$z = h(y)$$

dove:

- h è nota solo per via numerica (computabile) ma non (facilmente) conoscibile per via analitica (es: computer models);
 - y è un vettore stocastico $k - \dim$ con distribuzione $F(y)$;
 - la funzione di ripartizione F è nota e siamo capaci di simulare replicazioni di y

Monte Carlo in generale

Ci interessa la distribuzione $G(z)$ di

$$z = h(y)$$

dove:

- h è nota solo per via numerica (computabile) ma non (facilmente) conoscibile per via analitica (es: computer models);
 - y è un vettore stocastico $k - \dim$ con distribuzione $F(y)$;
 - la funzione di ripartizione F è nota e siamo capaci di simulare replicazioni di y

Monte Carlo in generale

La distribuzione $G(z)$ si può calcolare per via analitica solo in casi molto semplici.

Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis oooooo Bootstrap oooo B non parametrico ooooooo IC ooo Regressione ooooo

Monte Carlo in generale

La distribuzione $G(z)$ si può calcolare per via analitica solo in casi molto semplici.

L'idea è allora quella di

simulare y un grande numero m di volte

e ripetere il calcolo della corrispondente z :

$$z_i = h(y_i) \quad i = 1, \dots, m.$$

In questo modo z_1, \dots, z_m è un *grande campione* da $G()$.

9 / 41

Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis oooooo Bootstrap oooo B non parametrico ooooooo IC ooo Regressione ooooo

Monte Carlo in generale

Si può quindi usare z_1, \dots, z_m per avere informazioni su G tramite:

- la funzione di ripartizione empirica delle z

$$\hat{G}(t) = \frac{1}{m} \sum_{i=1}^m I(z_i \leq t)$$

con $I()$ funzione indicatrice

- la funzione di densità (istogramma, Kernel smoothing, etc.)
- la media: $\bar{z} = \frac{1}{m} \sum_{i=1}^m z_i$
- la varianza: $S_z^2 = \frac{1}{m-1} \sum_{i=1}^m (z_i - \bar{z})^2$
- e così via ...

10 / 41

Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis oooooo Bootstrap oooo B non parametrico ooooooo IC ooo Regressione ooooo

Monte Carlo in generale

Si può quindi usare z_1, \dots, z_m per avere informazioni su G tramite:

- la funzione di ripartizione empirica delle z

$$\hat{G}(t) = \frac{1}{m} \sum_{i=1}^m I(z_i \leq t)$$

con $I()$ funzione indicatrice

- la funzione di densità (istogramma, Kernel smoothing, etc.)
- la media: $\bar{z} = \frac{1}{m} \sum_{i=1}^m z_i$
- la varianza: $S_z^2 = \frac{1}{m-1} \sum_{i=1}^m (z_i - \bar{z})^2$
- e così via ...

10 / 41

Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis oooooo Bootstrap oooo B non parametrico ooooooo IC ooo Regressione ooooo

Monte Carlo in generale

Si può quindi usare z_1, \dots, z_m per avere informazioni su G tramite:

- la funzione di ripartizione empirica delle z

$$\hat{G}(t) = \frac{1}{m} \sum_{i=1}^m I(z_i \leq t)$$

con $I()$ funzione indicatrice

- la funzione di densità (istogramma, Kernel smoothing, etc.)
- la media: $\bar{z} = \frac{1}{m} \sum_{i=1}^m z_i$
- la varianza: $S_z^2 = \frac{1}{m-1} \sum_{i=1}^m (z_i - \bar{z})^2$
- e così via ...

10 / 41

Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis oooooo Bootstrap oooo B non parametrico ooooooo IC ooo Regressione ooooo

Test Monte Carlo

Esempio Matlab - correlazione

Sia

$$Z \equiv N_2(0, \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})$$

Si vuole testare l'ipotesi nulla

$$H_0 : \rho = 0.5$$

contro l'alternativa unidirezionale destra sulla base del coefficiente di correlazione campionario di $n=10$ osservazioni Z_1, \dots, Z_n .

Calcolare il valore critico approssimato usando $m=1000$ simulazioni MC per un livello di significatività nominale $\alpha = 1\%$.

12 / 41

Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis oooooo Bootstrap oooo B non parametrico ooooooo IC ooo Regressione ooooo

Test Monte Carlo

Esempio Matlab - correlazione

```
n=3;
m=4;
alfa=0.01;
Z=nan(n,2,m);
r=nan(m,1);
for j=1:m
    Z(:,:,j)=mvnrnd(zeros(2,1),[1,.5;.5,1],n);
    c = corrcoef(Z(:,:,j));
    r(j) = c(1,2);
end
rcrit = prctile(r,100*(1-alfa))
```

13 / 41

Origine oooo Generalità ooo Test ooo **Stima ooo** Risk Analysis oooooo Bootstrap oooo B non parametrico ooooooo IC ooo Regressione ooooo

Valutazione degli estimatori via Monte Carlo

Interessa ora il caso in cui la funzione $h()$ è lo stimatore di un parametro θ :

$$\hat{\theta} = \hat{\theta}(\mathbf{y}) = h(\mathbf{y})$$

e \mathbf{y} è un campione $\mathbf{y} = (y_1, \dots, y_n)$ con distribuzione $F(y; \xi^o)$.

Indichiamo con

$$\mathbf{y}_j^* = (y_{j,1}^*, \dots, y_{j,n}^*)$$

il campione $n - dim$ simulato tramite un generatore di numeri casuali con distribuzione $F(y; \xi^o)$.

Indichiamo con

$$\hat{\theta}_j^* = \hat{\theta}(y_j^*)$$

la stima basata su \mathbf{y}_j^* .

14 / 41

Origine oooo Generalità ooo Test ooo **Stima ooo** Risk Analysis oooooo Bootstrap oooo B non parametrico ooooooo IC ooo Regressione ooooo

Valutazione degli estimatori via Monte Carlo

Interessa ora il caso in cui la funzione $h()$ è lo stimatore di un parametro θ :

$$\hat{\theta} = \hat{\theta}(\mathbf{y}) = h(\mathbf{y})$$

e \mathbf{y} è un campione $\mathbf{y} = (y_1, \dots, y_n)$ con distribuzione $F(y; \xi^o)$.

Indichiamo con

$$\mathbf{y}_j^* = (y_{j,1}^*, \dots, y_{j,n}^*)$$

il campione $n - dim$ simulato tramite un generatore di numeri casuali con distribuzione $F(y; \xi^o)$.

Indichiamo con

$$\hat{\theta}_j^* = \hat{\theta}(y_j^*)$$

la stima basata su \mathbf{y}_j^* .

14 / 41

Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis ooooo Bootstrap oooo B non parametrico ooooooo IC ooo Regressione ooooo

Valutazione degli estimatori via Monte Carlo

Esempio Matlab - correlazione

Come al Capitolo precedente, sia $Z \equiv N_2(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})$
 Si considera la stima di $\rho = 0.5$ data dal coefficiente di correlazione campionario sulla base di n=10 osservazioni.
 Stimare bias e standard error della stima usando m=1000 simulazioni MC.

```
bias_r=mean(r)-.5  
se_r = sqrt(mse(r-.5))
```

16 / 41

Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis ooooo Bootstrap oooo B non parametrico ooooooo IC ooo Regressione ooooo

Valutazione degli estimatori via Monte Carlo

Esempio Matlab - correlazione

Come al Capitolo precedente, sia $Z \equiv N_2(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})$
 Si considera la stima di $\rho = 0.5$ data dal coefficiente di correlazione campionario sulla base di n=10 osservazioni.
 Stimare bias e standard error della stima usando m=1000 simulazioni MC.

```
bias_r=mean(r)-.5  
se_r = sqrt(mse(r-.5))
```

16 / 41

Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis ooooo Bootstrap oooo B non parametrico ooooooo IC ooo Regressione ooooo

Risk analysis

Interessa la "perdita" z funzione delle "condizioni ambientali"
 y :

$$z = h(y)$$

dove $h()$ è nota e di y si conosce la distribuzione.
 In particolare interessa la probabilità di una "perdita molto alta"

$$P(z > soglia) = 1 - G_z(soglia)$$

- z = perdita finanziaria
- z = n° decessi in una pandemia
- z = intensità scossa sismica
- ...

Esempi:

17 / 41

Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis ooooo Bootstrap oooo B non parametrico ooooooo IC ooo Regressione ooooo

Risk analysis

Interessa la "perdita" z funzione delle "condizioni ambientali"
 y :

$$z = h(y)$$

dove $h()$ è nota e di y si conosce la distribuzione.
 In particolare interessa la probabilità di una "perdita molto alta"

$$P(z > soglia) = 1 - G_z(soglia)$$

- z = perdita finanziaria
- z = n° decessi in una pandemia
- z = intensità scossa sismica
- ...

Esempi:

17 / 41

Risk analysis

Esempio Matlab

Simulatore modello AR-Poisson

```

function [y, lam] = simARP(omega, alpha, beta, n)
y = nan(n, 1);
lam = nan(n, 1);
y(1)=1;
lam(1)=0.5;
for t=2:n
    lam(t) = exp(omega + alpha*log(y(t-1)+1) +...
                 beta*log(lam(t-1)));
    y(t) = poissrnd(lam(t));
end

```

Origine Generalità Test Stima Risk Analysis Bootstrap B non parametrico IC Regressione

Risk analysis

Esempio Matlab

Simulazione MC - modello AR-Poisson

```

omega=.105; alpha=0.1; beta=0.89
t_star=100;
soglia=1000
m=100
n=t_star+10;
Y = nan(n,m);
parfor b=1:m
    Y(:,b) = simARP(omega,alpha,beta,n);
end

```

Risk analysis

Esempio Matlab

Simulazione MC - modello AR-Poisson

```

z = (Y(t_star,:)>soglia)';
Pz = mean(z);
disp(num2str(Pz,'pz=%5.3f'))
figure; plot(Y);
hold on; plot(repmat(soglia,n,1)); axis tight
title('Superamento soglia AR-Poisson')

```

Origine Generalità Test Stima Risk Analysis Bootstrap B non parametrico IC Regressione

Il metodo Bootstrap



November 12, 2018;

The International Prize in Statistics: Bradley Efron, professor of statistics and data science at Stanford University, for developing the "bootstrap," in 1977.

• • •

200,000 citations in more than 200 journals worldwide, in computer science, engineering, immunology, mathematics, biochemistry, medicine, physics and astronomy and the social sciences.

Dal metodo MC al Bootstrap

Se la distribuzione di y non è nota è possibile fare simulazioni Monte Carlo?

La risposta è no.

Tuttavia se si hanno a disposizione dei dati osservati
 $y = (y_1, \dots, y_n)$,
una strada percorribile per capire la variabilità di stime o test
consiste nel costruire delle simulazioni basate sulla
distribuzione empirica del **particolare vettore osservato**:

$$y = (y_1, \dots, y_n)$$

Inizialmente supponiamo che y sia *iid*, poi estenderemo.

II Bootstrap Parametrico

Quando

- 1 la distribuzione delle componenti di y , $F(y_j|\xi)$ è nota nella forma
 - 2 ma il parametro ξ è ignoto
 - 3 e si hanno a disposizione dei dati osservati $y = (y_1, \dots, y_n)$

↓

si può stimare $\hat{\xi}$ tramite $\hat{\xi}(y)$ e poi applicare il metodo Monte Carlo con $\xi = \hat{\xi}$.

II Bootstrap Nonparametrico

Se F non è nota, l'idea è allora quella di stimarla tramite la funzione di ripartizione empirica del campione:

$$\hat{F}_n(x) = \frac{\sum I(y_t \leq x)}{n}$$

e applicare poi il metodo Monte Carlo a tale distribuzione.

II Bootstrap Nonparametrico

Se F non è nota, l'idea è allora quella di stimarla tramite la funzione di ripartizione empirica del campione:

$$\hat{F}_n(x) = \frac{\sum I(y_t \leq x)}{n}$$

e applicare poi il metodo Monte Carlo a tale distribuzione.

Nella prossima slide vediamo che questo equivale a ri-campionare i dati osservati (y_1, \dots, y_n).

Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis oooooo Bootstrap oooo B non parametrico oooooo IC ooo Regressione ooooo

II Bootstrap Nonparametrico

Indichiamo il j – esimo campione simulato , con

$$\mathbf{y}_j^* = (y_{j,1}^*, \dots, y_{j,N}^*)$$

dove $N \leq n$ e $j = 1, \dots, m$.

Ciascun elemento di y_j^* viene estratto a caso (con rimessa) dal campione originario.

In pratica si ottiene generando un numero casuale equidistribuito sugli interi da 1 ad n .

Indicato con r_t tale numero casuale, si pone

$$y_{jt}^* = y_{r_t} \quad t = 1, \dots, N, j = 1, \dots, m.$$



 28 / 41

Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis oooooo Bootstrap oooo B non parametrico oooooo IC ooo Regressione ooooo

II Bootstrap Nonparametrico

Indichiamo il j – esimo campione simulato , con

$$\mathbf{y}_j^* = (y_{j,1}^*, \dots, y_{j,N}^*)$$

dove $N \leq n$ e $j = 1, \dots, m$.

Ciascun elemento di y_j^* viene estratto a caso (con rimessa) dal campione originario.

In pratica si ottiene generando un numero casuale equidistribuito sugli interi da 1 ad n .

Indicato con r_t tale numero casuale, si pone

$$y_{jt}^* = y_{r_t} \quad t = 1, \dots, N, j = 1, \dots, m.$$



 28 / 41

Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis oooooo Bootstrap oooo B non parametrico oooooo IC ooo Regressione ooooo

II Bootstrap Nonparametrico

Indichiamo il j – esimo campione simulato , con

$$\mathbf{y}_j^* = (y_{j,1}^*, \dots, y_{j,N}^*)$$

dove $N \leq n$ e $j = 1, \dots, m$.

Ciascun elemento di y_j^* viene estratto a caso (con rimessa) dal campione originario.

In pratica si ottiene generando un numero casuale equidistribuito sugli interi da 1 ad n .

Indicato con r_t tale numero casuale, si pone

$$y_{jt}^* = y_{r_t} \quad t = 1, \dots, N, j = 1, \dots, m.$$



 28 / 41

Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis oooooo Bootstrap oooo B non parametrico oooooo IC ooo Regressione ooooo

II Bootstrap Nonparametrico

Indichiamo il j – esimo campione simulato , con

$$\mathbf{y}_j^* = (y_{j,1}^*, \dots, y_{j,N}^*)$$

dove $N \leq n$ e $j = 1, \dots, m$.

Ciascun elemento di y_j^* viene estratto a caso (con rimessa) dal campione originario.

In pratica si ottiene generando un numero casuale equidistribuito sugli interi da 1 ad n .

Indicato con r_t tale numero casuale, si pone

$$y_{jt}^* = y_{r_t} \quad t = 1, \dots, N, j = 1, \dots, m.$$



 28 / 41

II Bootstrap Nonparametrico

Si ha così una stima

$$\hat{\theta}_j^* = \hat{\theta}(\mathbf{y}_j^*).$$

Iterando m volte come nel metodo MC si arriva ad una valutazione della distribuzione e delle proprietà dello stimatore $\hat{\theta}_N$.

Proprietà asintotiche

Ci interessa la distribuzione bootstrap di $\hat{\theta}_N$ al crescere di n ed m , per fissato N .

Tale distribuzione è data da

$$\hat{G}_{n,m}^N(t) \cong G^N(t) = P\left(\hat{\theta}_N \leq t\right)$$

Proprietà asintotiche

Ci interessa la distribuzione bootstrap di $\hat{\theta}_N$ al crescere di n ed m , per fissato N .

Tale distribuzione è data da

$$\hat{G}_{n,m}^N(t) \cong G^N(t) = P(\hat{\theta}_N \leq t)$$

che calcoliamo come

$$\hat{G}_{n,m}^N(t) = \frac{1}{m} \sum_{i=1}^m I\left(\hat{\theta}_{N,j}^* \leq t\right)$$

Poiché le simulazioni sono iid, per la legge forte dei grandi numeri abbiamo che

$$\lim_{m \rightarrow \infty} \hat{G}_{n,m}^N(t) = G_n^N(t) = P(\hat{\theta}_{N,j}^* \leq t | y \equiv \hat{F}_n)$$

Inoltre, applicando ancora la legge forte dei grandi numeri ma al campione di partenza, abbiamo che

$$\lim_{n \rightarrow \infty} \hat{F}_n(x) = F(x)$$

Perciò si conclude che

$$\lim_{n,m \rightarrow \infty} \hat{G}_{n,m}(t) = G^N(t).$$

Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis oooooo Bootstrap oooo B non parametrico oooooo● IC ooo Regressione ooooo

Esempio con Matlab

Consideriamo il caso semplice della stima della media nei due casi:

- caso parametrico: assumiamo che

$$y_i \text{ IID } N(\theta = \mu, \sigma^2) \text{ con } \mu, \sigma^2 \text{ ignote}$$

- caso non parametrico: F_y ignota

32 / 41

Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis oooooo Bootstrap oooo B non parametrico oooooo● IC ooo Regressione ooooo

Esempio con Matlab

```
n=5; m=300;
y=[3.75 4 -1 7 0.33]'; % dati osservati
N=7;

ystar_par=normrnd(mean(y),std(y),N,m);
mhu_star_par = mean(ystar_par)

ystar_nonpar=y(unidrnd(n,N,m));
mhu_star_nonpar = mean(ystar_nonpar)
```

33 / 41

Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis oooooo Bootstrap oooo B non parametrico oooooo● IC ●oo Regressione ooooo

IC Bootstrap

Un semplice intervallo di confidenza bootstrap a livello $1 - \alpha$ approssimato per θ si può basare sui percentili di

$$\hat{\theta}_1^* \leq \dots \leq \hat{\theta}_m^*$$

ed avrà la forma:

$$\hat{\theta}_{[\frac{\alpha}{2}m]}^* \div \hat{\theta}_{[(1-\frac{\alpha}{2})m]}^*$$

dove $[x]$ è la parte intera di x .

34 / 41

Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis oooooo Bootstrap oooo B non parametrico oooooo● IC ●oo Regressione ooooo

IC Bootstrap studentizzato

Se è disponibile una stima della varianza di $\hat{\theta}$, diciamo $s^2(\hat{\theta})$, un IC migliore si ottiene utilizzando le replicazioni bootstrap studentizzate

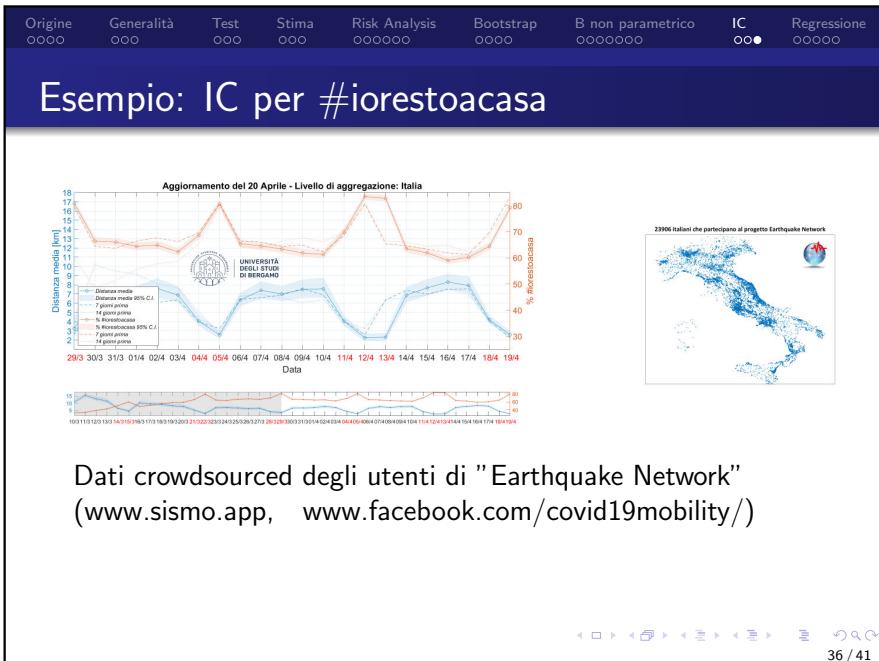
$$t_j^* = \frac{\hat{\theta}_j^* - \hat{\theta}}{s(\hat{\theta}_j^*)}$$

che vengono poi ordinate $t_1^* \leq \dots \leq t_m^*$.

L'intervallo di confidenza studentizzato è così dato da

$$\hat{\theta} + t_{[\frac{\alpha}{2}m]}^* s(\hat{\theta}) \div \hat{\theta} + t_{[(1-\frac{\alpha}{2})m]}^* s(\hat{\theta}).$$

35 / 41



Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis oooooo Bootstrap oooo B non parametrico ooooooo IC ooo Regressione ooooo

MC e regressione

Consideriamo la simulazione di un modello di regressione

$$y_t = \beta x_t + \varepsilon_t$$

dove ε_t iid $N(0, \sigma^2)$, i parametri $\xi = (\beta, \sigma^2)$ sono noti e le x fissate.

Una simulazione Monte Carlo per $t = 1, \dots, n$ si ottiene con il comando

```
Y=X*beta+randn(mhu,sigma,n,1);
```

oppure

```
Y=mvnrnd(X*beta,sigma2*eye(10));
```

37 / 41

Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis oooooo Bootstrap oooo B non parametrico ooooooo IC ooo Regressione ooooo

Bootstrap e regressione

Consideriamo la simulazione dello stesso modello

$$y_t = \beta x_t + \varepsilon_t$$

dove però

- ε_t iid $F(0, \sigma^2)$,
- i parametri $\xi = (\beta, \sigma^2)$ sono IGNOTI
- Si hanno a disposizione i dati osservati (x_t, y_t) per $t = 1, \dots, n$.

38 / 41

Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis oooooo Bootstrap oooo B non parametrico ooooooo IC ooo Regressione ooooo

Bootstrap e regressione

CASO 1

F è nota nella forma ma non nei parametri
⇒ **bootstrap parametrico**:

- si stima $\hat{\xi} = (\hat{\beta}, \hat{\sigma}^2)$, usando i dati osservati $(x_1, y_1), \dots, (x_n, y_n)$;
- si simulano gli errori $\varepsilon_{j,t}^*$ IID $F(0, \hat{\sigma}^2)$ e
- si simula $y_j^* = (y_{j,1}^*, \dots, y_{j,N}^*)$ tramite:

$$y_{j,t}^* = \hat{\beta} x_t + \varepsilon_{j,t}^*$$

39 / 41

Origine oooo Generalità ooo Test ooo Stima ooo Risk Analysis oooooo Bootstrap oooo B non parametrico oooooo IC ooo Regressione oooo●

Bootstrap e regressione

CASO 3

F NON È NOTA nella forma e si "crede poco" alla struttura lineare $X\beta$.

⇒ **Bootstrap non parametrico:**

$$(x_{j,1}^*, y_{j,1}^*), \dots, (x_{j,n}^*, y_{j,n}^*) \text{ IID } (x_1, y_1), \dots, (x_n, y_n)$$

41 / 41

Processi Stocastici oo oooooo oooo GP ooo Markovianità ooo Previsione ooooo Stazionarietà ooo AR oooooo ooooooooooooo

Modelli stocastici

Capitolo 05-A

Processi stocastici a memoria breve

Alessandro Fassò
aa 2019/20

1 / 54

Processi Stocastici oo oooooo oooo GP ooo Markovianità ooo Previsione ooooo Stazionarietà ooo AR oooooo ooooooooooooo

Contenuti

- 1 Processi Stocastici**
 - Introduzione
 - Esempi: traiettorie/serie storiche
 - Distribuzioni di un PS
- 2 GP**
- 3 Markovianità**
- 4 Previsione**
- 5 Stazionarietà**
 - Momenti del 2° ordine
 - Esempi
- 6 AR**
 - AR(p)

2 / 54

Processi Stocastici ● oooooo oooo GP ooo Markovianità ooo Previsione ooooo Stazionarietà ooo AR oooooo ooooooooooooo

Introduzione

Processi Stocastici

Un processo stocastico è un modello probabilistico dato da una famiglia di variabili casuali

$$y_t, \quad \text{con } t \in \mathcal{T} = \{0, \pm 1, \pm 2, \dots\}$$

che è caratterizzata dall'insieme delle sue distribuzioni congiunte

$$f_{y_{t_1}, y_{t_2}, \dots, y_{t_k}}(z_1, z_2, \dots, z_k)$$

per ogni $t_j \in \mathcal{T}, j = 1, \dots, k$ e $k = 1, 2, \dots$

3 / 54

Processi Stocastici

Processi stocastici a tempo discreto:

$\mathcal{T} = \{0, \pm 1, \pm 2, \dots\}$ oppure $\mathcal{T} = \{1, 2, \dots\}$

Processi stocastici a tempo continuo:

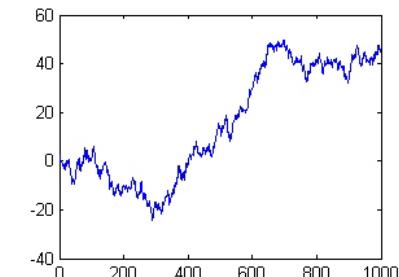
$\mathcal{T} \in \mathbb{R}$ oppure $\mathcal{T} \in \mathbb{R}^+$

Processi stocastici spaziali (campi stocastici):

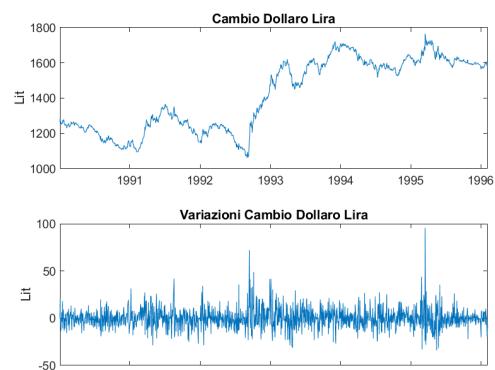
$$\mathcal{T} \in \mathbb{R}^d, \text{ con } d \geq 1$$

Dati simulati da Modello Random Walk

$$\begin{aligned} y_t &= y_{t-1} + \varepsilon_t \\ \varepsilon_t &\equiv NID(0, \sigma^2) \end{aligned}$$

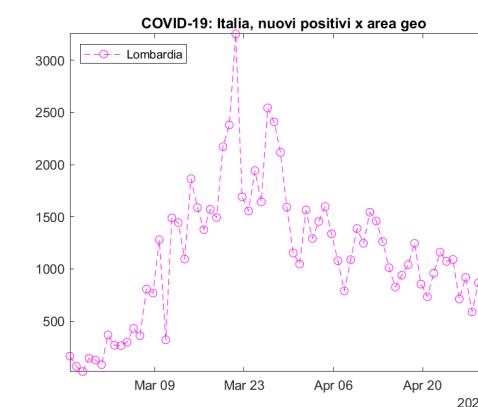


Cambio Dollaro Lira



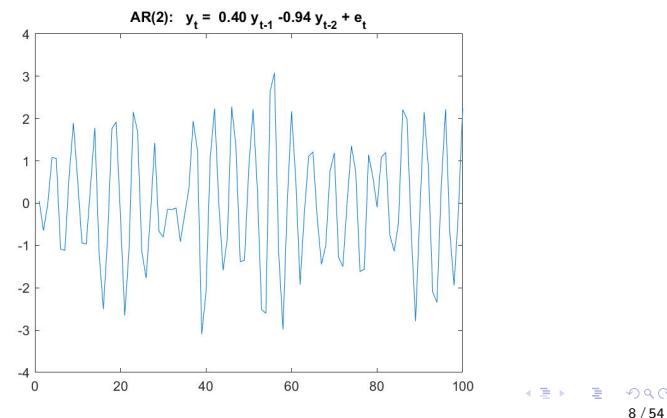
COVID-19

Lombardia, Nuovi positivi



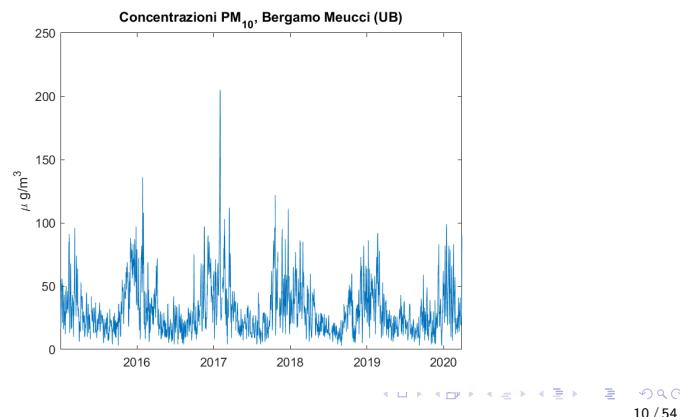
Dati simulati da AR(2)

Caso con radici complesse coniugate



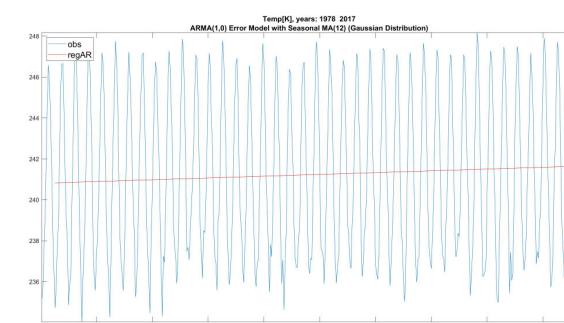
Qualità dell'aria

PM₁₀ a Bergamo Meucci (UB)



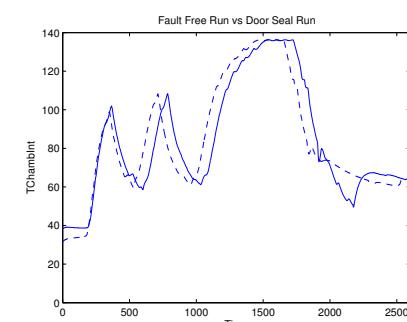
Global warming

Medie mensili EU a 850hpa (dati IGRA)



Dati di processo in alta frequenza

Sterilizzazione strumenti medicali



Esempio IID

$$y_t \text{ IID } N(0, \sigma^2)$$

\Downarrow

$$f_{y_{t_1}, y_{t_2}, \dots, y_{t_k}}(z_1, z_2, \dots, z_k) = \prod_{j=1}^k \frac{1}{\sigma} \phi(z_j/\sigma)$$

NB: è il caso di processo "banale" della statistica classica in cui non vi è né dinamica temporale né prevedibilità.

Esercizio AR(1)

Il processo stocastico $AR(1)$:

$$\begin{aligned} y_t &= \alpha y_{t-1} + \varepsilon_t \quad t = 1, 2, \dots \\ \varepsilon_t &\equiv NID(0, \sigma^2) \\ |\alpha| &< 1 \end{aligned}$$

con valore iniziale $y_0 = 0$ ha distribuzione

$$f_{y_{t_1}, \dots, y_{t_k}}(z_1, \dots, z_k) = \prod_{j=1}^k \frac{1}{\sigma} \phi((z_j - \alpha z_{t-j})/\sigma).$$

Esempio trend lineare + IID

$$y_t = \beta t + \varepsilon_t, \quad \varepsilon_t \text{ IID } N(0, \sigma^2)$$

\Downarrow

$$f_{y_{t_1}, \dots, y_{t_k}}(z_1, \dots, z_k) = \prod_{j=1}^k \frac{1}{\sigma} \phi((z_j - \beta t_j)/\sigma)$$

Processi Gaussiani

Un PS si dice gaussiano se tutte le sue distribuzioni sono gaussiane. Posto $\mathbf{t} = (t_1, \dots, t_k)$ vale:

$$(y_{t_1}, \dots, y_{t_k}) \equiv N_k(\mu_t, \Gamma_t)$$

dove Γ_t è la matrice di varianze-covarianze di $(y_{t_1}, \dots, y_{t_k})$

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo oooo	o●o	ooo	ooooo	oo ooo oooo	oooo oooooooooooo

Processi condizionatamente Gaussiani

Sono PS in cui "l'equazione di misura" è gaussiana ma la dinamica può essere non lineare.

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo oooo	o●o	ooo	ooooo	oo ooo oooo	oooo oooooooooooo

Processi condizionatamente Gaussiani

Sono PS in cui "l'equazione di misura" è gaussiana ma la dinamica può essere non lineare.

Sia

$$\mathbf{S}_{t-1} = y_1, \dots, y_{t-1}$$

la storia antecedente il tempo t .

$$(y_t | \mathbf{S}_{t-1}) \equiv N(\mu(\mathbf{S}_{t-1}), \sigma^2(\mathbf{S}_{t-1}))$$

- $\mu()$: Dinamica Lineare/Nonlineare
 - regressione lineare con errori correlati
 - Spline, polinomi locali, reti neurali, ...
- $\sigma^2()$: Omoschedasticità/Eteroschedasticità

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo oooo	o●o	ooo	ooooo	oo ooo oooo	oooo oooooooooooo

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo oooo	●oo	ooo	ooooo	oo ooo oooo	oooo oooooooooooo

Processi Markoviani

Nota il presente y_t , allora il futuro $y_{t+j}, j \geq 1$ non dipende dal passato \mathbf{S}_{t-1} :

$$f(y_{t+j} | \mathbf{S}_{t-1}) = f(y_t | y_{t-1})$$

Allora

$$f(y_t, y_{t-1}, y_{t-2}, \dots | y_0) = \prod_{j=1}^t f(y_j | y_{j-1})$$

Processi Stocastici oo ooooooo oooo	GP ooo	Markovianità ●oo	Previsione ooooo	Stazionarietà oo ooo oooooo	AR oooo oooooooooooo
--	-----------	---------------------	---------------------	--------------------------------------	----------------------------

Processi Markoviani

Nota il presente y_t , allora il futuro $y_{t+j}, j \geq 1$ non dipende dal passato \mathbf{S}_{t-1} :

$$f(y_{t+j}|\mathbf{S}_{t-1}) = f(y_t|y_{t-1})$$

Allora

$$f(y_t, y_{t-1}, y_{t-2}, \dots | y_0) = \prod_{j=1}^t f(y_j | y_{j-1})$$

Esempio AR(1)

$$\begin{aligned} y_t &= \alpha y_{t-1} + \varepsilon_t, \quad \varepsilon_t \equiv NID(0, \sigma^2) \\ &\Downarrow \\ (y_t | y_{t-1}, y_{t-2}, \dots) &\equiv y_t | y_{t-1} \equiv N(\alpha y_{t-1}, \sigma_\varepsilon^2) \end{aligned}$$

Processi Stocastici oo ooooooo oooo	GP ooo	Markovianità ●oo	Previsione ooooo	Stazionarietà oo ooo oooooo	AR oooo oooooooooooo
--	-----------	---------------------	---------------------	--------------------------------------	----------------------------

Processi Markoviani

Esercizio

Fare l'esempio di un processo che è markoviano ma non è un AR(1)

Processi Stocastici oo ooooooo oooo	GP ooo	Markovianità ●oo	Previsione ooooo	Stazionarietà oo ooo oooooo	AR oooo oooooooooooo
--	-----------	---------------------	---------------------	--------------------------------------	----------------------------

Processi Stocastici oo ooooooo oooo	GP ooo	Markovianità ●oo	Previsione ●oooo	Stazionarietà oo ooo oooooo	AR oooo oooooooooooo
--	-----------	---------------------	---------------------	--------------------------------------	----------------------------

Teoria della Previsione

Sia \mathbf{S}_t l'insieme delle informazioni (storia) fino a t .

Un previsore a k passi di y_{t+k} è una funzione di \mathbf{S}_t :

$$\hat{y}_{t+k}(k) = f_k(y_t, y_{t-1}, \dots) = f_k(\mathbf{S}_t)$$

Processi Stocastici oo oooooo oooo	GP ooo	Markovianità ooo	Previsione ●oooo	Stazionarietà oo ooo oooooo	AR ooooo oooooooooooo
---	-----------	---------------------	---------------------	--------------------------------------	-----------------------------

Teoria della Previsione

Sia \mathbf{S}_t l'insieme delle informazioni (storia) fino a t .

Un previsore a k passi di y_{t+k} è una funzione di \mathbf{S}_t :

$$\hat{y}_{t+k}(k) = f_k(y_t, y_{t-1}, \dots) = f_k(\mathbf{S}_t)$$

Nel ns. caso:

$$\mathbf{S}_t = (y_t, y_{t-1}, \dots).$$

Processi Stocastici oo oooooo oooo	GP ooo	Markovianità ooo	Previsione ●oooo	Stazionarietà oo ooo oooooo	AR ooooo oooooooooooo
---	-----------	---------------------	---------------------	--------------------------------------	-----------------------------

Teoria della Previsione

Se $Ey_t^2 < \infty$ allora la previsione ottima (a minima varianza) è

$$\hat{y}_{t+k}(k) = E(y_{t+k}|\mathbf{S}_t)$$

In particolare è importante la **previsione ad un passo**

$$\hat{y}_{t+1} = \hat{y}_{t+1}(1) = E(y_{t+1}|\mathbf{S}_t)$$

Processi Stocastici oo oooooo oooo	GP ooo	Markovianità ooo	Previsione ●oooo	Stazionarietà oo ooo oooooo	AR ooooo oooooooooooo
---	-----------	---------------------	---------------------	--------------------------------------	-----------------------------

Teoria della Previsione

Se $Ey_t^2 < \infty$ allora la previsione ottima (a minima varianza) è

$$\hat{y}_{t+k}(k) = E(y_{t+k}|\mathbf{S}_t)$$

Processi Stocastici oo oooooo oooo	GP ooo	Markovianità ooo	Previsione ●oooo	Stazionarietà oo ooo oooooo	AR ooooo oooooooooooo
---	-----------	---------------------	---------------------	--------------------------------------	-----------------------------

Teoria della Previsione

Se $Ey_t^2 < \infty$ allora la previsione ottima (a minima varianza) è

$$\hat{y}_{t+k}(k) = E(y_{t+k}|\mathbf{S}_t)$$

In particolare è importante la **previsione ad un passo**

$$\hat{y}_{t+1} = \hat{y}_{t+1}(1) = E(y_{t+1}|\mathbf{S}_t)$$

Processi Stocastici oo oooooo oooo	GP ooo	Markovianità ooo	Previsione ●oooo	Stazionarietà oo ooo oooooo	AR ooooo oooooooooooo
---	-----------	---------------------	---------------------	--------------------------------------	-----------------------------

Teoria della Previsione

Se $Ey_t^2 < \infty$ allora la previsione ottima (a minima varianza) è

$$\hat{y}_{t+k}(k) = E(y_{t+k}|\mathbf{S}_t)$$

In particolare è importante la **previsione ad un passo**

$$\hat{y}_{t+1} = \hat{y}_{t+1}(1) = E(y_{t+1}|\mathbf{S}_t)$$

Linearità

Un previsore si dice lineare se è una funzione lineare del passato:

$$\hat{y}_{t+1} = E(y_{t+1}|\mathbf{S}_t) = \sum_{j=1}^{\infty} \xi_j y_{t-j}$$

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo oooo	ooo	ooo	ooo•oo	oo ooo oooo	oooo oooooooooooo

Esempio AR(1)

- Previsione a un passo

$$\hat{y}_{t+1} = \alpha y_t$$

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo oooo	ooo	ooo	ooo•oo	oo ooo oooo	oooo oooooooooooo

Esempio AR(1)

- Previsione a un passo

$$\hat{y}_{t+1} = \alpha y_t$$

- Previsione a k passi

$$\hat{y}_{t+1}(k) = \alpha^k y_t$$

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo oooo	ooo	ooo	ooo•oo	oo ooo oooo	oooo oooooooooooo

Esempio AR(1)

- Previsione a un passo

$$\hat{y}_{t+1} = \alpha y_t$$

- Previsione a k passi

$$\hat{y}_{t+1}(k) = \alpha^k y_t$$

- Memoria breve:

$$|\alpha| < 1 \quad \Rightarrow \quad \hat{y}_t(k) \rightarrow 0 \quad per \quad k \rightarrow \infty$$

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo oooo	ooo	ooo	ooo•oo	oo ooo oooo	oooo oooooooooooo

L'errore di previsione

$$e_t = \hat{y}_t - y_t$$

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo oooooo oooo	ooo	ooo	ooo•oo	oo ooo oooooo	oooo oooooooooooo

L'errore di previsione

$$e_t = \hat{y}_t - y_t$$

- ha media nulla: $E(e_t) = 0$

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo oooooo oooo	ooo	ooo	ooo•oo	oo ooo oooooo	oooo oooooooooooo

L'errore di previsione

$$e_t = \hat{y}_t - y_t$$

- ha media nulla: $E(e_t) = 0$
- è incorrelato (ortogonale) col passato

$$E(e_t y_{t-j}) = 0, \quad j = 1, 2, \dots$$

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo oooooo oooo	ooo	ooo	ooo•oo	oo ooo oooooo	oooo oooooooooooo

L'errore di previsione

$$e_t = \hat{y}_t - y_t$$

- ha media nulla: $E(e_t) = 0$
- è incorrelato (ortogonale) col passato

$$E(e_t y_{t-j}) = 0, \quad j = 1, 2, \dots$$

- La sua varianza determina la precisione della previsione:

$$\sigma_{e_t}^2 = E(e_t^2)$$

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo oooooo oooo	ooo	ooo	ooo•oo	oo ooo oooooo	oooo oooooooooooo

L'errore di previsione

$$e_t = \hat{y}_t - y_t$$

- ha media nulla: $E(e_t) = 0$
- è incorrelato (ortogonale) col passato

$$E(e_t y_{t-j}) = 0, \quad j = 1, 2, \dots$$

- La sua varianza determina la precisione della previsione:

$$\sigma_{e_t}^2 = E(e_t^2)$$

- Supporremo nel seguito omoschedasticità:

$$\sigma^2 = \sigma_{e_t}^2 = E(e_t^2 | \mathbf{S}_{t-1})$$

Processi Stocastici GP Markovianità Previsione Stazionarietà AR

```

graph LR
    A[Processi Stocastici] --> B[GP]
    B --> C[Markovianità]
    C --> D[Previsione]
    D --> E[Stazionarietà]
    E --> F[AR]
  
```

26 / 54

Processi Stocastici GP Markovianità Previsione Stazionarietà AR

```

graph LR
    A[Processi Stocastici] --> B[GP]
    B --> C[Markovianità]
    C --> D[Previsione]
    D --> E[Stazionarietà]
    E --> F[AR]
  
```

27 / 54

Processi Stazionari

Un processo stocastico si dice **stazionario in senso forte** se le sue distribuzioni non dipendono dall'origine dei tempi:

$$f_{y_{t_1}, y_{t_2}, \dots, y_{t_k}}(z_1, z_2, \dots, z_k) = f_{y_{t_1+h}, y_{t_2+h}, \dots, y_{t_k+h}}(z_1, z_2, \dots, z_k) \text{ per ogni } h$$

Processi Stocastici GP Markovianità Previsione Stazionarietà AR

```

graph LR
    A[Processi Stocastici] --> B[GP]
    B --> C[Markovianità]
    C --> D[Previsione]
    D --> E[Stazionarietà]
    E --> F[AR]
  
```

27 / 54

Processi Stazionari

Un processo stocastico si dice **stazionario in senso forte** se le sue distribuzioni non dipendono dall'origine dei tempi:

$$f_{y_{t_1}, y_{t_2}, \dots, y_{t_k}}(z_1, z_2, \dots, z_k) = f_{y_{t_1+h}, y_{t_2+h}, \dots, y_{t_k+h}}(z_1, z_2, \dots, z_k) \text{ per ogni } h$$

Si dice **stazionario in senso debole** se i momenti del primo e del second'ordine non dipendono da t

$$E(y_t) = \mu_y \quad \text{per ogni } t$$

$$\text{Var}(y_t) = \sigma_y^2 \quad \text{per ogni } t$$

$$E(y_t y_{t+h}) = \gamma(h) \quad \text{per ogni } t$$

27 / 54

Processi Stocastici GP Markovianità Previsione Stazionarietà AR

```

graph LR
    A[Processi Stocastici] --> B[GP]
    B --> C[Markovianità]
    C --> D[Previsione]
    D --> E[Stazionarietà]
    E --> F[AR]
  
```

28 / 54

Violazioni tipiche della stazionarietà

Vediamo alcuni esempi di non-stazionarietà.

- La media $\mu_t = E(y_t)$ non è stazionaria:
 - 1 trend lineare $\mu_t = \alpha + \beta t$
 - 2 trend non-lineari $\mu_t = \mu(t) = \text{trigonometrica/esponenziale/polinomiale/spline} \dots$
 - 3 regressione $\mu_t = x_t^\top \beta$
- non stazionarietà in varianza (eteroschedasticità)
- non stazionarietà in covarianza

Autocovarianza

L'autocovarianza di un PS che ha varianza finita è data da:

$$\gamma_t(h) = \text{Cov}(y_t, y_{t+h}) = E(y_t - \mu_t)(y_{t+h} - \mu_{t+h})$$

Autocorrelazione

L'autocorrelazione del PS v_t è data da

$$\rho_t(h) = \rho(y_t, y_{t+h}) = \frac{\text{Cov}(y_t, y_{t+h})}{\sigma_{y_t} \sigma_{y_{t+h}}}, \quad h = 0, \pm 1, \pm 2, \dots$$

Se il PS è stazionario allora

$$\rho_t(h) = \rho(|h|) = \frac{\gamma(h)}{\gamma(0)}$$

Autocovarianza

L'autocovarianza di un PS che ha varianza finita è data da:

$$\gamma_t(h) = \text{Cov}(y_t, y_{t+h}) = E(y_t - \mu_t)(y_{t+h} - \mu_{t+h})$$

Se il PS è stazionario allora l'autocovarianza dipende solo dalla distanza $|h|$:

$$\gamma_t(h) = \gamma(|h|) = E(y_t y_{t+h}) - \mu_y^2$$

Autocorrelazione campionaria

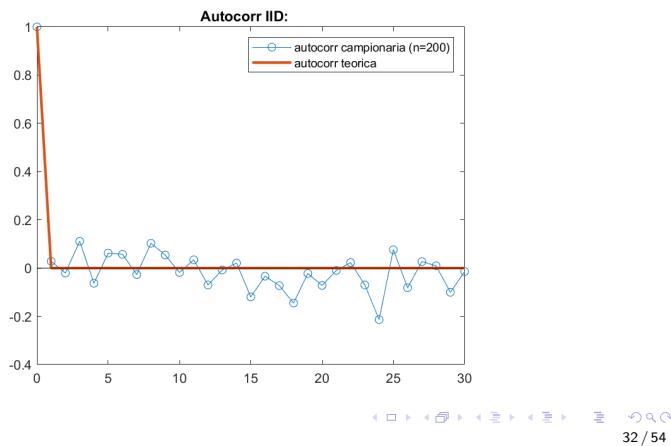
L'autocovarianza campionaria è data da

$$c(h) = \frac{1}{n} \sum_{t=1}^{n-h} (y_t - \bar{y})(y_{t+h} - \bar{y})$$

L'autocorrelazione campionaria è data da

$$r(h) = \frac{c(h)}{c(0)}$$

Autocorrelazione IID



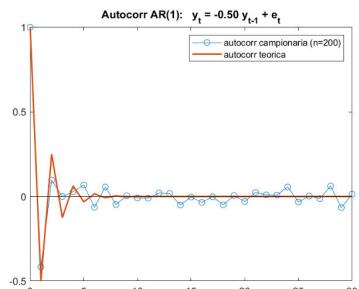
Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo	ooo	ooo	oooo	oo	oooo
oooooooo				ooo	oooooooooooo
oooo				ooo•ooo	oooooooooooo
Esempi					

Autocorrelazione AR(1)

$$y_t = \alpha y_{t-1} + \varepsilon_t$$

$$|\alpha| < 1$$

$$\Downarrow$$



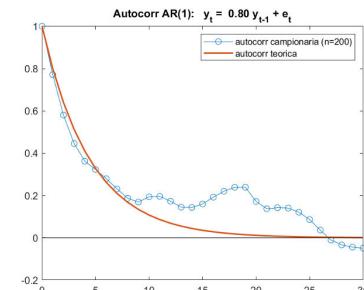
24 / 54

Autocorrelazione AR(1)

$$y_t = \alpha y_{t-1} + \varepsilon_t$$

$$|\alpha| < 1$$

$$\rho(h) = \alpha^h$$



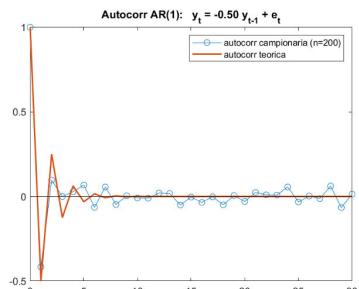
Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo	ooo	ooo	oooo	oo	oooo
oooooooo				ooo	oooooooooooo
oooo				ooo•ooo	
Esempi					

Autocorrelazione AR(1)

$$y_t = \alpha y_{t-1} + \varepsilon_t$$

$$|\alpha| < 1$$

$$\Downarrow$$

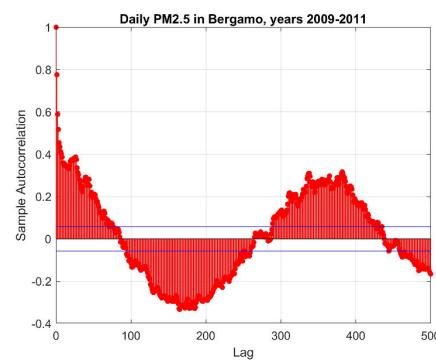


24 / 54

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo	ooo	ooo	oooo	oo	oooo
oooooooo				ooo	oooooooooooo
ooo				ooo●oo	
Esempi					

Dati giornalieri

PM2.5 a Bergamo



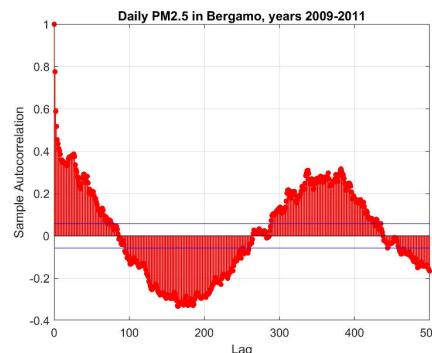
A set of small, light-blue navigation icons typically found in Beamer presentations, including symbols for back, forward, search, and table of contents.

A set of small, light-blue navigation icons typically found in Beamer presentations, including symbols for back, forward, search, and table of contents.

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo oooo	ooo	ooo	ooooo	oo oooo oooo	oooo oooooooooooo
Esempi					

Dati giornalieri

PM2.5 a Bergamo



Cosa sono le due rette orizzontali blu?

35 / 54

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo oooo	ooo	ooo	ooooo	oo oooo oooo	oooo oooooooooooo
Esempi					

Distribuzione dell'autocorrelazione campionaria

y_t IID



$$\sqrt{nr}(h) \cong N(0, 1)$$

36 / 54

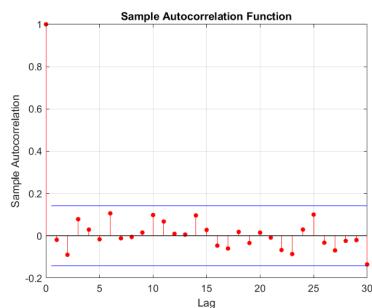
Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo oooo	ooo	ooo	ooooo	oo oooo oooo	oooo oooooooooooo
Esempi					

Distribuzione dell'autocorrelazione campionaria

y_t IID



$$\sqrt{nr}(h) \cong N(0, 1)$$



Correlogramma di $n=200$ osservazioni
NID

36 / 54

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo oooo	ooo	ooo	ooooo	oo oooo oooo	oooo oooooooooooo
Esempi					

37 / 54

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo oooo	ooo	ooo	ooooo	ooo oooo oooo	●ooo oooooooooooo

Modelli ARMA

I modelli AutoRegressivi (AR) a Media Mobile (MA) sono dei modelli lineari che descrivono un'ampia casistica di funzioni di autocorrelazione caratterizzate da memoria breve.

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo oooo	ooo	ooo	ooooo	ooo oooo oooo	●ooo oooooooooooo

Modelli ARMA

I modelli AutoRegressivi (AR) a Media Mobile (MA) sono dei modelli lineari che descrivono un'ampia casistica di funzioni di autocorrelazione caratterizzate da memoria breve.

Per memoria breve si intende che la correlazione tende a zero velocemente:

$$\rho(h) \approx \lambda^h$$

con $|\lambda| < 1$

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo oooo	ooo	ooo	ooooo	ooo oooo oooo	●ooo oooooooooooo

Modelli ARMA

I modelli AutoRegressivi (AR) a Media Mobile (MA) sono dei modelli lineari che descrivono un'ampia casistica di funzioni di autocorrelazione caratterizzate da memoria breve.

Per memoria breve si intende che la correlazione tende a zero velocemente:

$$\rho(h) \approx \lambda^h$$

con $|\lambda| < 1$ Iniziamo ora lo studio della parte AR.

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo oooo	ooo	ooo	ooooo	ooo oooo oooo	●●oo oooooooooooo

Studio dettagliato di AR(1)

Caratteristiche del modello autoregressivo di ordine uno:

$$y_t = \alpha y_{t-1} + \varepsilon_t, \quad t = 0, \pm 1, \pm 2, \dots$$

$$\varepsilon_t \equiv NID(0, \sigma^2)$$

$$|\alpha| < 1$$

1 Rappresentazione MA(∞): $y_t = \sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j}$

Studio dettagliato di AR(1)

Caratteristiche del modello autoregressivo di ordine uno:

$$\begin{aligned} y_t &= \alpha y_{t-1} + \varepsilon_t, \quad t = 0, \pm 1, \pm 2, \dots \\ \varepsilon_t &\equiv NID(0, \sigma^2) \\ |\alpha| &< 1 \end{aligned}$$

- 1** Rappresentazione MA(∞): $y_t = \sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j}$

2 Media: $E(y_t) = 0$

Studio dettagliato di AR(1)

Caratteristiche del modello autoregressivo di ordine uno:

$$\begin{aligned} y_t &= \alpha y_{t-1} + \varepsilon_t, \quad t = 0, \pm 1, \pm 2, \dots \\ \varepsilon_t &\equiv NID(0, \sigma^2) \\ |\alpha| &< 1 \end{aligned}$$

- 1 Rappresentazione MA(∞): $y_t = \sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j}$
 - 2 Media: $E(y_t) = 0$
 - 3 Varianza: $Var(y_t) = \sigma^2 / (1 - \alpha^2)$

Studio dettagliato di AR(1)

Caratteristiche del modello autoregressivo di ordine uno:

$$y_t = \alpha y_{t-1} + \varepsilon_t, \quad t = 0, \pm 1, \pm 2, \dots$$

$$\varepsilon_t \equiv NID(0, \sigma^2)$$

$$|\alpha| < 1$$

- 1 Rappresentazione MA(∞): $y_t = \sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j}$
 - 2 Media: $E(y_t) = 0$
 - 3 Varianza: $Var(y_t) = \sigma^2 / (1 - \alpha^2)$
 - 4 Autocovarianza: $\gamma(h) = \alpha^h \times \sigma^2 / (1 - \alpha^2)$

Studio dettagliato di AR(1)

Caratteristiche del modello autoregressivo di ordine uno:

$$\begin{aligned} y_t &= \alpha y_{t-1} + \varepsilon_t, \quad t = 0, \pm 1, \pm 2, \dots \\ \varepsilon_t &\equiv NID(0, \sigma^2) \\ |\alpha| &< 1 \end{aligned}$$

- 1 Rappresentazione MA(∞): $y_t = \sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j}$
 - 2 Media: $E(y_t) = 0$
 - 3 Varianza: $Var(y_t) = \sigma^2 / (1 - \alpha^2)$
 - 4 Autocovarianza: $\gamma(h) = \alpha^h \times \sigma^2 / (1 - \alpha^2)$
 - 5 Autocorrelazione: $\rho(h) = \alpha^h$

Processi Stocastici oo ooooooo oooo	GP ooo	Markovianità ooo	Previsione ooooo	Stazionarietà oo oooooo	AR oooo ●oooooooooooo
AR(p)					

Modelli AR(p)

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \dots + a_p y_{t-p} + \varepsilon_t$$

$\varepsilon_t \text{ iid } (0, \sigma^2)$

$$t = 0, \pm 1, \pm 2, \dots$$

Problema: è markoviano ?

Se $h > p$

$$E(y_t y_{t-h} | y_{t-1}, \dots, y_{t-p}) = 0$$

Viene detto Markoviano di ordine p

Processi Stocastici oo ooooooo oooo	GP ooo	Markovianità ooo	Previsione ooooo	Stazionarietà oo oooooo	AR oooo ○●oooooooooooo
AR(p)					

Autocorrelazione parziale

L'autocorrelazione parziale del PS stazionario y_t è data da

$$\bar{\rho}(h) = \rho(y_t, y_{t+h} | y_{t+1}, \dots, y_{t+h-1}) \quad h = 0, \pm 1, \pm 2, \dots$$

Processi Stocastici oo ooooooo oooo	GP ooo	Markovianità ooo	Previsione ooooo	Stazionarietà oo oooooo	AR oooo ●oooooooooooo
AR(p)					

Autocorrelazione parziale

L'autocorrelazione parziale del PS stazionario y_t è data da

$$\bar{\rho}(h) = \rho(y_t, y_{t+h} | y_{t+1}, \dots, y_{t+h-1}) \quad h = 0, \pm 1, \pm 2, \dots$$

Se y_t è un modello AR(p) allora, per $h > p$, si ha

$$\bar{\rho}(h) = 0.$$

Processi Stocastici oo ooooooo oooo	GP ooo	Markovianità ooo	Previsione ooooo	Stazionarietà oo oooooo	AR oooo ○○●oooooooooooo
AR(p)					

Forma polinomiale di un AR

Operatore ritardo

Consideriamo dapprima B l'operatore di ritardo:

$$By_t = y_{t-1},$$

Processi Stocastici oo ooooooo ooo	GP ooo	Markovianità ooo	Previsione ooooo	Stazionarietà oo oooo	AR oooo oooooooooooo
AR(p)					

Forma polinomiale di un AR

Operatore ritardo

Consideriamo dapprima B l'operatore di ritardo:

$$By_t = y_{t-1},$$

allora

$$B^0 y_t = y_t$$

$$B(By_t) = B^2 y_t = y_{t-2}$$

$$B^k y_t = y_{t-k}$$

Processi Stocastici oo ooooooo ooo	GP ooo	Markovianità ooo	Previsione ooooo	Stazionarietà oo oooo	AR oooo oooooooooooo
AR(p)					

Forma polinomiale di un AR

Operatore AR

Riscriviamo il modello AR(p):

$$y_t - a_1 y_{t-1} - a_2 y_{t-2} - \dots - a_p y_{t-p} = \varepsilon_t$$

allora

$$a(B) y_t = \varepsilon_t$$

dove l'operatore AR è dato da:

$$a(B) = 1 - a_1 B - \dots - a_p B^p$$

Processi Stocastici oo ooooooo ooo	GP ooo	Markovianità ooo	Previsione ooooo	Stazionarietà oo oooo	AR oooo oooooooooooo
AR(p)					

Esempio AR(1)

Per il modello AR(1) abbiamo le seguenti identità

$$y_t - \alpha y_{t-1} = \varepsilon_t$$

$$a(B)y_t = \varepsilon_t$$

$$(1 - \alpha B)y_t = \varepsilon_t$$

Processi Stocastici oo ooooooo ooo	GP ooo	Markovianità ooo	Previsione ooooo	Stazionarietà oo oooo	AR oooo oooooooooooo
AR(p)					

Esempio AR(1)

Per il modello AR(1) abbiamo le seguenti identità

$$y_t - \alpha y_{t-1} = \varepsilon_t$$

$$a(B)y_t = \varepsilon_t$$

$$(1 - \alpha B)y_t = \varepsilon_t$$

In particolare per $\alpha = 1$ abbiamo il Random Walk in cui

$$y_t - y_{t-1} = (1 - B)y_t = \varepsilon_t$$

Stazionarietà di un AR(p)

Sia $z \in C$ la variabile complessa. Il PS y_t dato da un modello AR(p) è stazionario sse le p soluzioni¹ dell'equazione caratteristica

$$a\left(\frac{1}{z}\right) = 0$$

sono interne al cerchio unitario $|z| < 1$.

¹Teorema fondamentale dell'algebra e dintorni

Stazionarietà di un AR(p)

Sia $z \in C$ la variabile complessa. Il PS y_t dato da un modello AR(p) è stazionario sse le p soluzioni¹ dell'equazione caratteristica

$$a\left(\frac{1}{z}\right) = 0$$

sono interne al cerchio unitario $|z| < 1$.

Ovvero sse

$$a\left(\frac{1}{z}\right) = 1 - a_1 \frac{1}{z} - \dots - a_p \frac{1}{z^p} \neq 0 \text{ per } |z| \geq 1$$

¹Teorema fondamentale dell'algebra e dintorni

Esempio AR(1)

Per il modello AR(1) abbiamo

$$a\left(\frac{1}{z}\right) = 1 - \frac{\alpha}{z}$$

e la soluzione dell'eq^{ne} caratteristica è

$$z = \alpha$$

Collegamento AR(1) - AR(p)

Se $\lambda_1, \dots, \lambda_p$ sono le soluzioni dell'eq^{ne} caratteristica, allora l'operatore AR(p) si può riscrivere² come prodotto di operatori AR(1)

$$a\left(\frac{1}{z}\right) = \prod_{j=1}^p \left(1 - \frac{\lambda_j}{z}\right)$$

così i ragionamenti che abbiamo fatto sugli AR(1) si estendono al caso generale.

²vedi ancora Teorema fondamentale dell'algebra e dintorni

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo oooo	ooo	ooo	ooooo	ooo ooooo oooo	oooo oooooooooooo

AR(p)

Forma MA(∞)

Se y_t è un AR(p) stazionario allora esistono i coefficienti $\omega_j, j = 1, \dots, \infty$, tali che:

$$y_t = \frac{1}{a(B)} \varepsilon_t = \sum_{j=1}^{\infty} \omega_j \varepsilon_{t-j} + \varepsilon_t$$

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo oooo	ooo	ooo	ooooo	ooo ooooo oooo	oooo oooooooooooo

AR(p)

Forma MA(∞)

Se y_t è un AR(p) stazionario allora esistono i coefficienti $\omega_j, j = 1, \dots, \infty$, tali che:

$$y_t = \frac{1}{a(B)} \varepsilon_t = \sum_{j=1}^{\infty} \omega_j \varepsilon_{t-j} + \varepsilon_t$$

Caso AR(1): $\omega_j = \alpha^j$

$$y_t = \sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j}$$

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo oooo	ooo	ooo	ooooo	ooo ooooo oooo	oooo oooooooooooo

AR(p)

Esempio AR(2)

Il modello

$$y_t = 1.7y_{t-1} - 0.72y_{t-2} + \varepsilon_t, \quad \varepsilon_t \equiv NID(0, \sigma^2)$$

ha operatore polinomiale

$$a(B) = 1 - 1.7B + 0.72B^2.$$

Le soluzioni dell'equazione caratteristica $a(\frac{1}{z}) = 0$ sono

$$z_1 = 0.9 \text{ e } z_2 = 0.8.$$

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo oooo	ooo	ooo	ooooo	ooo ooooo oooo	oooo oooooooooooo

AR(p)

Esempio AR(2)

Il modello

$$y_t = 1.7y_{t-1} - 0.72y_{t-2} + \varepsilon_t, \quad \varepsilon_t \equiv NID(0, \sigma^2)$$

ha operatore polinomiale

$$a(B) = 1 - 1.7B + 0.72B^2.$$

Le soluzioni dell'equazione caratteristica $a(\frac{1}{z}) = 0$ sono

$$z_1 = 0.9 \text{ e } z_2 = 0.8.$$

Poiché $|z_{1,2}| < 1$ il processo generato dal modello di cui sopra è **stazionario**.

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo ooo	ooo	ooo	ooooo	ooo ooo oooo	oooo oooooooooooo●oo

AR(p)

... segue Esempio AR(2) - matlab

Il modello di cui sopra con $\sigma^2 = 0.5$ è implementato da:

```
>>m_AR2 = regARIMA('AR',[1.7, -.72], 'Variance', .5,  
'Intercept', 0) 3
```

e le sue radici sono date da

```
>>r = roots([1, -1.7, .72]) 4
```

con modulo

```
>>abs(r)
```

³Ovvero: arima(...)

⁴Ovvero: roots([1, -cell2mat(m_AR2.AR)])

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo ooo	ooo	ooo	ooooo	ooo ooo oooo	oooo oooooooooooo●oo

AR(p)

... segue Esempio AR(2) - simulazione

Simulazione di $m=10$ replicazioni di una serie storica lunga $n=100$ dell'AR(2) Gaussiano di cui sopra

Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo ooo	ooo	ooo	ooooo	ooo ooo oooo	oooo oooooooooooo●oo

AR(p)

... segue Esempio AR(2) - simulazione

Simulazione di $m=10$ replicazioni di una serie storica lunga $n=100$ dell'AR(2) Gaussiano di cui sopra

```
>> y_star = simulate(m_AR2,n,'NumPaths',m);
```

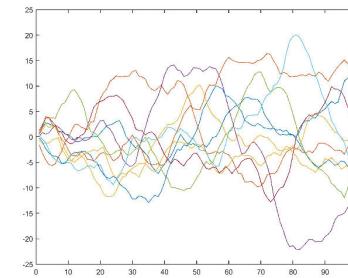
Processi Stocastici	GP	Markovianità	Previsione	Stazionarietà	AR
oo ooooooo ooo	ooo	ooo	ooooo	ooo ooo oooo	oooo oooooooooooo●oo

AR(p)

... segue Esempio AR(2) - simulazione

Simulazione di $m=10$ replicazioni di una serie storica lunga $n=100$ dell'AR(2) Gaussiano di cui sopra

```
>> y_star = simulate(m_AR2,n,'NumPaths',m);
```



Processi Stocastici oo oooooo ooo	GP ooo	Markovianità ooo	Previsione oooo	Stazionarietà oo ooo oooo	AR oooo oooooooooooo●
AR(p)					

... segue: Esercizio per casa

Esaminare secondo la logica vista sopra i modelli definiti dai seguenti operatori AR:

$$a(B) = 1 - 1B + 0.34B^2$$

$$a(B) = 1 - 0.4B + 0.9425B^2$$

$$a(B) = 1 - 1.2B + 1B^2$$

Processi Stocastici oo oooooo ooo	GP ooo	Markovianità ooo	Previsione oooo	Stazionarietà oo ooo oooo	AR oooo oooooooooooo●
AR(p)					

... segue: Esercizio per casa

Esaminare secondo la logica vista sopra i modelli definiti dai seguenti operatori AR:

$$a(B) = 1 - 1B + 0.34B^2$$

$$a(B) = 1 - 0.4B + 0.9425B^2$$

$$a(B) = 1 - 1.2B + 1B^2$$

Esaminare anche autocorrelazione e autocorrelazione parziale!!!

MA+ARMA oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooooo
-------------------	---------------------------	------------------	-----------------	----------------------	---------------	-----------------------------

Modelli stocastici Capitolo 05-B Processi stocastici a memoria breve

Alessandro Fassò

aa 2019/20

MA+ARMA oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooooo
-------------------	---------------------------	------------------	-----------------	----------------------	---------------	-----------------------------

Contenuti

- 1 MA+ARMA
 - MA(1)
 - ARMA
- 2 ARIMA+MLE
 - MLE
- 3 regARIMA
- 4 AIC/BIC
- 5 DIAG
- 6 VAL
- 7 Bootstrap

MA+ARMA oo ooo•ooo ARMA	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooo	VAL ooooo	Bootstrap oooooooooooo
----------------------------------	---------------------------	------------------	-----------------	------------------	--------------	---------------------------

Stazionarietà

La stazionarietà di un ARMA è determinata dalla componente AR come già visto per i modelli AR(p): le soluzioni dell'equazione caratteristica dell'operatore AR devono essere interne al cerchio unitario:

$$a\left(\frac{1}{z}\right) = 0 \Rightarrow |z| < 1.$$

In questo caso vale la forma MA(∞):

$$y_t = \frac{c(B)}{a(B)}\varepsilon_t = \sum_{j=1}^{\infty} \omega_j \varepsilon_{t-j} + \varepsilon_t$$

7 / 56

MA+ARMA oo ooo•ooo ARMA	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooo	VAL ooooo	Bootstrap oooooooooooo
----------------------------------	---------------------------	------------------	-----------------	------------------	--------------	---------------------------

Invertibilità

Un modello ARMA si dice invertibile sse le soluzioni dell'equazione caratteristica dell'operatore MA sono interne al cerchio unitario:

$$c\left(\frac{1}{z}\right) = 0 \Rightarrow |z| < 1.$$

In questo caso vale la forma AR(∞):

$$\varepsilon_t = \frac{a(B)}{c(B)}y_t = y_t - \sum_{j=1}^{\infty} \psi_j y_{t-j}$$

8 / 56

MA+ARMA oo ooo•ooo ARMA	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooo	VAL ooooo	Bootstrap oooooooooooo
----------------------------------	---------------------------	------------------	-----------------	------------------	--------------	---------------------------

Previsione AR

La previsione a un passo di un modello AR(p) è data da:

$$\hat{y}_t = \hat{y}_t(1) = E(y_t | \mathbf{S}_{t-1}) = a_1 y_{t-1} + \dots + a_p y_{t-p}$$

Proprietà

1 L'errore di previsione

$$e_t = y_t - \hat{y}_t = \varepsilon_t$$

è anche l'**Innovazione** del PS.

2 Linearità del previsore

- 3 Omoschedasticità: la previsione ha sempre la stessa incertezza

9 / 56

MA+ARMA oo ooo•ooo ARMA	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooo	VAL ooooo	Bootstrap oooooooooooo
----------------------------------	---------------------------	------------------	-----------------	------------------	--------------	---------------------------

Previsione AR a più passi

Per i modelli lineari in generale, e in particolare per gli ARMA, la previsione a più passi è ricorsiva.

Se l'orizzonte previsivo è $k = 2$, abbiamo:

$$\hat{y}_t(2) = a_1 \hat{y}_{t-1} + a_2 y_{t-2} + \dots + a_p y_{t-p}.$$

Più in generale

$$\hat{y}_t(k) = a_1 \hat{y}_{t-1}(k-2) + a_2 \hat{y}_{t-2}(k-1) + \dots + a_p \hat{y}_{t-p}(k-p).$$

NB: Se il PS è non lineare questo approccio può divergere.

10 / 56

MA+ARMA oo oooooo• ARMA	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooooooooooo
----------------------------------	---------------------------	------------------	-----------------	----------------------	---------------	-----------------------------------

Previsione ARMA

Nel caso di ARMA invertibile la previsione a un passo è data dalla seguente ricorsione

$$\hat{y}_t = a_1 y_{t-1} + \dots + a_p y_{t-p} + c_1 e_{t-1} + \dots + c_q e_{t-q}$$

$$e_t = y_t - \hat{y}_t$$

con valori iniziali

$$e_0, \dots, e_{1-q} = 0.$$

Affinché la ricorsione converga per $t \rightarrow \infty$, occorre che l'operatore MA sia invertibile. Allora

$$e_t = y_t - \hat{y}_t \cong y_t - \sum_{j=1}^{\infty} \psi_j y_{t-j} = \varepsilon_t$$

11 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooooooooooo
-------------------------	---------------------------	------------------	-----------------	----------------------	---------------	-----------------------------------

Test per radici unitarie

Test di Dickey-Fuller

$$y_t = \alpha y_{t-1} + \varepsilon_t$$

$$H_0 : \alpha = 1, \quad H_1 : \alpha < 1$$

per $n \rightarrow \infty$: $n(\hat{\alpha}_{OLS} - 1) \rightarrow v.c.\text{ nota sotto } H_0$

dove

$$\hat{\alpha}_{OLS} = \frac{\sum_{t=2}^n y_t y_{t-1}}{\sum_{t=2}^n y_{t-1}^2}$$

13 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooooooooooo
-------------------------	---------------------------	------------------	-----------------	----------------------	---------------	-----------------------------------

Radici Unitarie e modelli ARIMA

Se l'equazione caratteristica dell'operatore AR si annulla per $z = 1$:

$$a(1) = 0,$$

allora il modello ARMA è non stazionario.

In questi casi si fattorizza il termine con la radice unitaria e si ottiene un modello ARIMA che ha la forma:

$$a(B)(1 - B)y_t = c(B)\varepsilon_t.$$

In sostanza si usano le differenze prime:

$$(1 - B)y_t = y_t - y_{t-1} \equiv \text{ARMA stazionario}$$

12 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooooooooooo
-------------------------	---------------------------	------------------	-----------------	----------------------	---------------	-----------------------------------

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooooooooooo
-------------------------	---------------------------	------------------	-----------------	----------------------	---------------	-----------------------------------

COVID-19 - Introduzione caso di studio

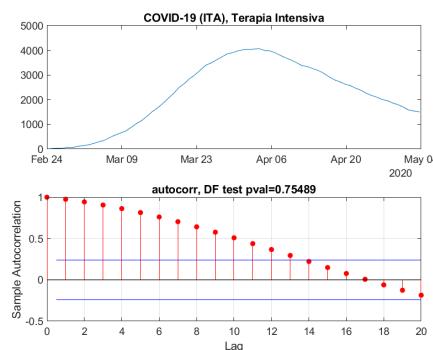
Consideriamo la modellazione ARIMA del numero di ricoverati in terapia intensiva per COVID-19 in Italia.

Obiettivo è mostrare come fare previsione ad un giorno con il modello stimato. Inclusa l'incertezza della previsione.

File dati Covid19_dpc_Italia.mat, aggiornamento 4 maggio 2020, fonte Dipartimento Protezione Civile.

14 / 56

COVID-19: stazionarietà

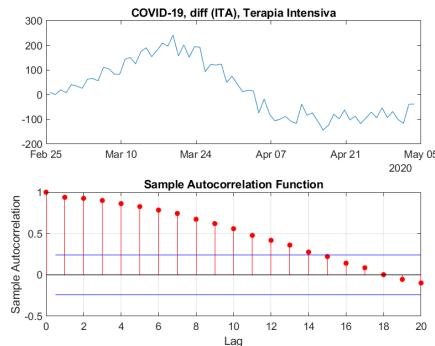


- Forte non stazionarietà
- Test di Dickey-Fuller non significativo.
- \Rightarrow ARIMA(p,1,q)
- Matlab:
`>>autocorr(y)`
`>>[h,p]=adftest(y)`

15 / 56

MLE

COVID-19: differenze prime



- Differenze prime con forte autocorrelazione
- \Rightarrow ARIMA(2,1,2)^a
- Matlab:
`>>m0=arima(2,1,2);`

^aModello non ottimizzato

17 / 56

MLE

Stima MLE (estimate)

La stima di massima verosimiglianza dei modelli ARIMA riguarda il vettore di parametri

$$\theta = (a_1, \dots, a_p, c_1, \dots, c_q, \sigma_\varepsilon^2)^t$$

allora la stima del modello è data da:

$$\hat{\theta}_{MLE} = \arg \max(\ln L_n(\theta))$$

Proprietà asintotiche: per n grande abbiamo:

$$(\hat{\theta} - \theta_0) \approx N(0, V_n)$$

dove

$$V_n = \left(\frac{\partial^2}{\partial \theta \partial \theta'} \ln L_n(\theta) \right)^{-1}$$

16 / 56

MLE

COVID-19: MLE

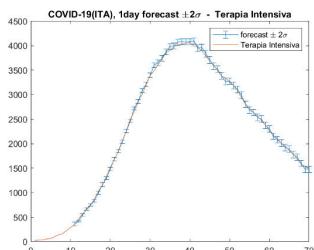
```
>> [mhat,Vhat,logL,info]=estimate(m0,y);
>> summarize(mhat)
ARIMA(2,1,2) Model (Gaussian Distribution)
Effective Sample Size: 70
Number of Estimated Parameters: 6
LogLikelihood: -337.834, AIC: 687.668, BIC: 701.159
```

	Value	StandardError	TStatistic	PValue
Constant	0.49642	0.42075	1.1799	0.23806
AR1	1.9441	0.049857	38.993	0
AR2	-0.95507	0.050112	-19.059	0
MA1	-1.7044	0.08357	-20.394	0
MA2	0.81942	0.084449	9.7031	0
Variance	910.98	162.84	5.5945	2.21e-08

18 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo●	regARIMA oooo	AIC/BIC oooo	DIAG oooooooo	VAL ooooo	Bootstrap oooooooooooo
MLE						

COVID-19: Previsione



Stima e previsione iterativa ($t=11:n$)

```
>>mhat{t-1}=estimate(m0,y(1:t-1));
>>[yhat(t),ymse(t)] = ...
    forecast(mhat{t-1},y(1:t-1));
>>errorbar(yhat,2*sqrt(ymse))
```

Forecast performance: rmse=38.6, rrmse=2%

19 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo●	regARIMA oooo	AIC/BIC oooo	DIAG oooooooo	VAL ooooo	Bootstrap oooooooooooo
MLE						

20 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA ●ooo	AIC/BIC oooo	DIAG oooooooo	VAL ooooo	Bootstrap oooooooooooo

Regressione con errori ARIMA

Consideriamo il modello

$$y_t = m(x_t) + e_t \\ a(B)e_t = c(B)\varepsilon_t.$$

Casi importanti sono:

- Regressione lineare con errori correlati:

$$m(x) = \beta'x$$

- Modelli nonlineari con errori correlati. Per esempio

$$m(x) = \text{spline}(x)$$

21 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA ●ooo	AIC/BIC oooo	DIAG oooooooo	VAL ooooo	Bootstrap oooooooooooo

OSSERVAZIONE

Se $a\left(\frac{1}{z}\right) = 0$ ha solo radici interne al cerchio unitario, il modello di regressione lineare a errori ARMA può essere scritto come:

$$y_t = \beta'x_t + \frac{c(B)}{a(B)}\varepsilon_t$$

22 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA ooo•o	AIC/BIC oooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooooooooooo
-------------------------	---------------------------	-------------------	-----------------	----------------------	---------------	-----------------------------------

Esempio: trend lineare su dati giornalieri

$$y_t = 2 + \frac{1.5}{365} t + \frac{\varepsilon_t}{1 - .75B + .5B^2}, \quad \varepsilon_t \text{ iid } t(5)$$

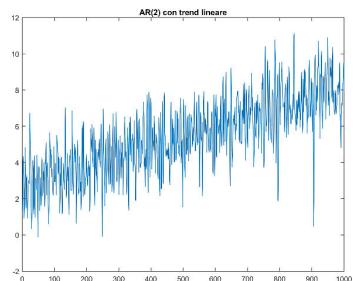


Figura: Trend lineare con errori AR(2)

23 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA ooo•o	AIC/BIC oooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooooooooooo
-------------------------	---------------------------	-------------------	-----------------	----------------------	---------------	-----------------------------------

Esempio: trend lineare su dati giornalieri

In Matlab il modello

$$y_t = 2 + \frac{1.5}{365} t + \frac{\varepsilon_t}{1 - .75B + .5B^2}, \quad \varepsilon_t \text{ iid } t(5)$$

si implementa in classe regARIMA:

```
>>m=regARIMA('Intercept',2,'Beta',1./365,'AR',[.75 -.5]);
>>e_distr.Name='t';
>>e_distr.DoF=5;
>>m.Distribution=e_distr;
```

24 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA ooo•o	AIC/BIC oooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooooooooooo
-------------------------	---------------------------	-------------------	-----------------	----------------------	---------------	-----------------------------------

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA ooo•o	AIC/BIC ●ooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooooooooooo
-------------------------	---------------------------	-------------------	-----------------	----------------------	---------------	-----------------------------------

Determinazione di (p,q)

Penalizzazione per la complessità:

- Criterio BIC $-2\log L + n_{par} \log(n)$
- Criterio AIC $-2\log L + 2n_{par}$

Matlab:

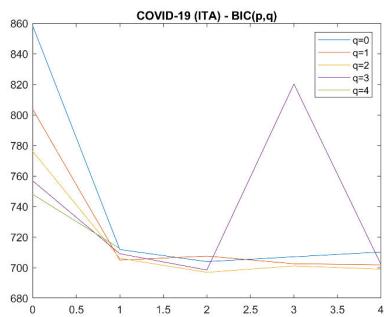
```
[m_hat,V_hat,logL,info] = estimate(m0,y);
n = length(y);
npar = sum(any(V_hat))-1
[aic,bic] = aicbic(logL,npar,n)
```

26 / 56

25 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC ooo●	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooooooooooo
-------------------------	---------------------------	------------------	-----------------	----------------------	---------------	-----------------------------------

COVID-19: scelta modello



- Minimizzazione BIC:
ARIMA(2,1,2)
- Altri modelli simili
- Validazione esterna: ricerca modelli che minimizza RMSE di previsione

27 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC ooo●	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooooooooooo
-------------------------	---------------------------	------------------	-----------------	----------------------	---------------	-----------------------------------

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC ooo●	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooooooooooo
-------------------------	---------------------------	------------------	-----------------	----------------------	---------------	-----------------------------------

Esercizio aic-bic per AR

- 1 Simulare $n = 750$ osservazioni di un modello AR(2) con

$$a(B) = 1 - .75B + .5B^2$$

$$\text{e } \sigma^2 = 0.001;$$
- 2 calcolare e plottare aic e bic per $p = 0, \dots, 4$;
- 3 individuare l'ordine ottimale secondo i due criteri aic e bic .

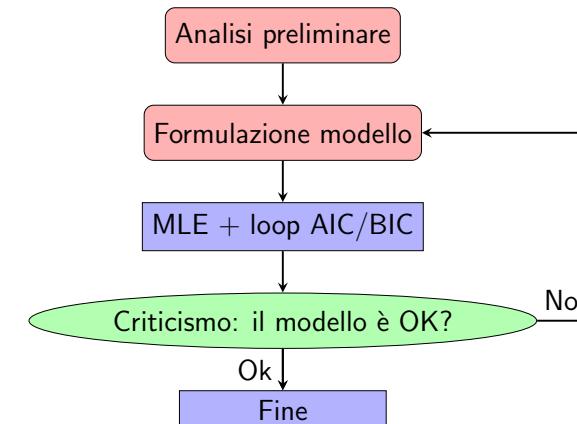
28 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC ooo●	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooooooooooo
-------------------------	---------------------------	------------------	-----------------	----------------------	---------------	-----------------------------------

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC ooo	DIAG ●oooooooooooo	VAL oooooo	Bootstrap oooooooooooooooooooo
-------------------------	---------------------------	------------------	----------------	-----------------------	---------------	-----------------------------------

Diagnostica e validazione

Flowchart modellazione



30 / 56

29 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG o•oooooooo	VAL oooo	Bootstrap oooooooooooo
-------------------------	---------------------------	------------------	-----------------	--------------------	-------------	---------------------------

Diagnostica e validazione

La fase del "criticismo" si applica al modello individuato dalla procedura automatica e consiste nell'analizzare:

- la significatività statistica e sostanziale dei coefficienti individuati;
- le proprietà del modello (stazionarietà, linearità etc)
- simulazione di traiettorie e bootstrap
- le proprietà dei residui: **Diagnostica**;
- la capacità previsiva: **Validazione**.

Trattiamo nel seguito la **Diagnostica** dei residui, la **Validazione** del modello ed il **Bootstrap**.

31 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG o•oooooooo	VAL oooo	Bootstrap oooooooooooo
-------------------------	---------------------------	------------------	-----------------	--------------------	-------------	---------------------------

Diagnostica dei residui

La diagnostica dei residui valuta se le assunzioni fatte sui residui (e sul modello) siano soddisfatte e, in caso contrario, dà indicazioni sulle modifiche da apportare al modello.

Ci si chiede se i residui si comportino come una realizzazione di VC IID, stanti i vincoli legati alla procedura di ottimizzazione:

- L'autocorrelazione è zero?
- La varianza è costante?
- I residui sono IID?
- I residui sono gaussiani?
- Ci sono outliers?

La diagnostica dei residui in genere si svolge nel set di stima (learning) ma può essere anche fatta nel set di validazione.

32 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG o•oooooooo	VAL oooo	Bootstrap oooooooooooo
-------------------------	---------------------------	------------------	-----------------	--------------------	-------------	---------------------------

Test di Lijung-Box

Il test di Box-Pierce/Ljung-Box verifica l'ipotesi nulla che il modello sia correttamente specificato e, quindi, che i residui corrispondano a delle innovazioni IID

$$Q_h^2(e) = n \sum_{j=1}^h r_e(j)^2 \frac{n+2}{n-j} \approx \begin{cases} \chi^2_{h-(p^*+q^*)} & \text{Learning set} \\ \chi^2_h & \text{Testing set} \end{cases}$$

Dove p^* e q^* sono il numero di coefficienti stimati della parte AR ed MA.

Nell'analisi preliminare il test LB può essere applicato ai dati stessi, y_t , per testare l'ipotesi che siano IID. In questo caso non vi è perdita di gradi di libertà.

33 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG o•oooooooo	VAL oooo	Bootstrap oooooooooooo
-------------------------	---------------------------	------------------	-----------------	--------------------	-------------	---------------------------

Omoschedasticità e outliers

Si vuole capire se

$$\text{var}(\varepsilon_t) = \sigma^2 = \text{cost}$$

oppure:

- dipende da t
- dipende dal passato.

Nel seguito consideriamo solo il primo punto.

L'identificazione degli outliers, nel caso più semplice, è basata sui valori estremi dei residui.

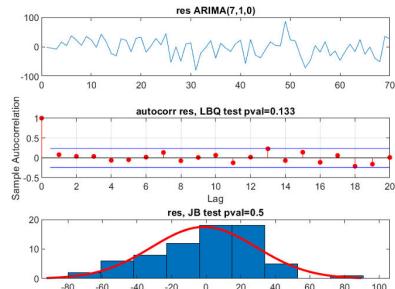
34 / 56



COVID-19: diagnostica residui

- Autocorrelazione e non normalità trascurabili
 - Matlab:
`>>res=infer(mhat,y);
>>[h,pval]=lbqtest(res,^a);
>>[h,pval]=jbtest(res);`
 - Omoschedasticità?

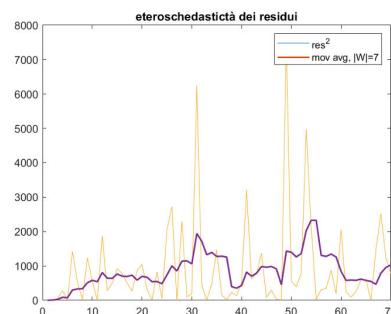
^a'Lags',20,'DoF',20-npar



COVID-19: diagnostica residui 2

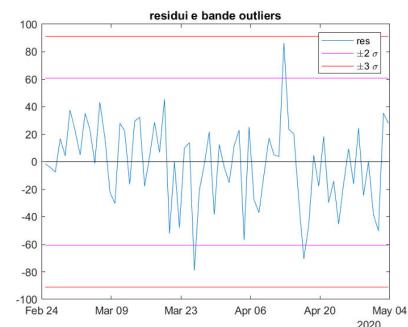
- Eteroschedasticità moderata:
 - fase iniziale a variabilità ridotta (conteggi)
 - effetto mitigato da previsione iterativa
 - Matlab:
`>>plot(movavg(res2, 3))`

^a'simple'.7



COVID-19: diagnostica residui 3

- Non si evidenziano particolari outliers.
 - Ove opportuno è possibile usare varianti di questo approccio
 - Matlab:
`>>isoutlier(res)`
offre varie tecniche.



Validazione e capacità previsiva

La validazione è prevalentemente orientata a valutare la capacità previsiva del modello individuato. Il significato assume sfumature diverse in relazione all'obiettivo del modello stesso:

- 1 Descrittivo
- 2 Previsivo
- 3 Qualità del dato
- 4 Monitoraggio e/o sorveglianza

Poiché individueremo più di un criterio, l'eventuale ottimizzazione dovrà, appunto, essere multi-criterio.

39 / 56

Validazione interna ed esterna

Nella **validazione interna** la capacità previsiva viene valutata sugli stessi dati usati per la stima. Questo approccio può essere esposto al rischio di sovra-ottimismo e overfitting.

Nella **validazione esterna** chiamata cross-validation (CV) in senso lato si usano dati diversi per stima e valutazione della capacità previsiva:

- CV classica (k-fold): la randomizzazione crea problemi per dati equispaziati nel tempo.
- CV Leave-one-out
- CV bootstrap
- CV incrementale (cfr previsione COVID-19)

40 / 56

Capacità previsiva

Indici di capacità previsiva per $y = \hat{y} + e$

	Assoluti	Relativi	(Pseudo) Normalizzati
Bias	$\frac{-1}{n} \sum e$	$\frac{-1}{n} \sum \frac{e}{y}$	
Corr			$r(y, \hat{y})$
MSE	$\frac{1}{n} \sum e^2$		$R^2 = 1 - \frac{MSE}{MSE_0}$
RMSE	$\sqrt{\frac{1}{n} \sum e^2}$	$\sqrt{\frac{1}{n} \sum (\frac{e}{y})^2}$	
MAE	$\frac{1}{n} \sum e $	$\frac{1}{n} \sum \frac{e}{y} $	$1 - MAE/MAE_0$

$$MSE_0 \leq \sigma_y^2, MAE_0 \leq MAD$$

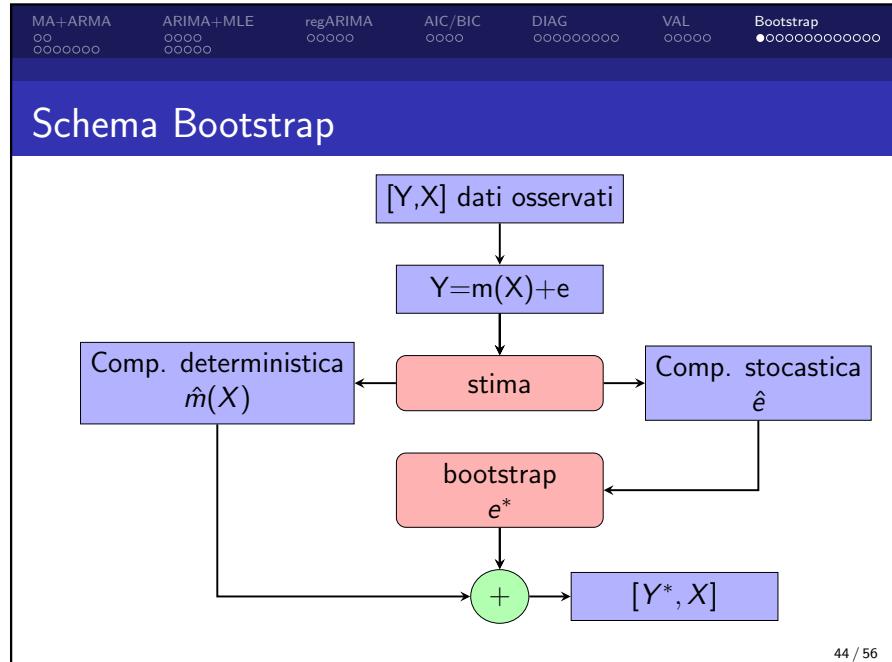
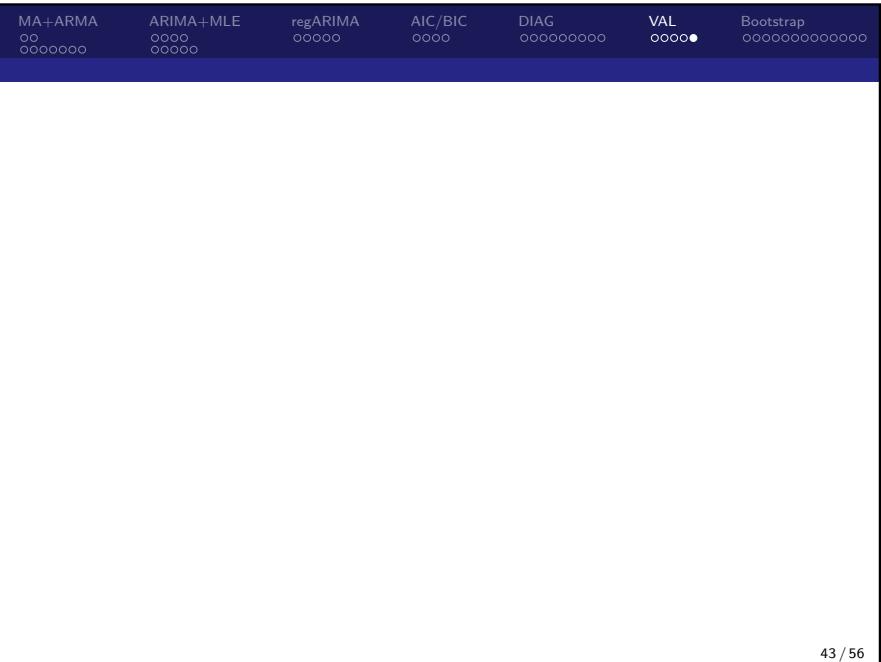
41 / 56

COVID-19 Performance analysis

Valutazione delle prestazioni del modello rispetto al modello "oggi come ieri" (RW).

	bias	rmse	rrmse
ARIMA	8.73	38.6	0.02
ARIMA-Ante picco	12.91		
ARIMA-Post picco	5.07		
RW	-21.38	110.7	0.12
RW-Ante picco	-108.3		
RW-Post picco	82.32		

42 / 56



MA+ARMA ARIMA+MLE regARIMA AIC/BIC DIAG VAL Bootstrap

```
oo          oooo      oooooo      oooo      ooooooo    ooooooo
ooooooo   oooooo   oooooo      oooo      ooooooo  ooooooo
```

Bootstrap per serie storiche

Vogliamo simulare un PS **stazionario** di cui però non conosciamo le distribuzioni.
 Abbiamo però a disposizione una serie storica

$$Y = (y_1, \dots, y_n)$$

che proviene dal PS di cui sopra.

Vogliamo simulare le serie

$$Y_j^* = (y_1^*, \dots, y_N^*)$$

per $j = 1, \dots, m$ che riproducano le proprietà di correlazione e distribuzione del PS.

45 / 56

MA+ARMA ARIMA+MLE regARIMA AIC/BIC DIAG VAL Bootstrap

```
oo          oooo      oooooo      oooo      ooooooo    ooooooo
ooooooo   oooooo   oooooo      oooo      ooooooo  ooooooo
```

Bootstrap parametrico

Consideriamo l'esempio della simulazione di un AR(1) gaussiano.
 Se i parametri $\xi = (\alpha, \sigma^2)$ fossero noti allora potremmo simulare Y_j^* tramite la ricorsione:

$$y_t^* = \alpha y_{t-1}^* + \varepsilon_t^*$$

$$\varepsilon_t^* \equiv NID(0, \sigma^2)$$

Bootstrap parametrico: Si stima $\hat{\xi}_{MLE} = (\hat{\alpha}, \hat{\sigma}^2)$ usando i dati Y e si simula Y_j^* tramite:

$$y_t^* = \hat{\alpha} y_{t-1}^* + \varepsilon_t^*$$

$$\varepsilon_t^* \equiv NID(0, \hat{\sigma}^2)$$

46 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooo●oooooooooooo
-------------------------	---------------------------	------------------	-----------------	----------------------	---------------	--------------------------------

Bootstrap semi-parametrico lineare

Si rinuncia alla normalità delle innovazioni, perciò stimato ξ e i residui $e = (e_1, \dots, e_n)$, si simula Y_j^* tramite

$$y_t^* = \hat{\alpha} y_{t-1}^* + e_t^*$$

$$e_t^* = \text{bootstrap } IID \text{ da } (e_1, \dots, e_n)$$

47 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooo●oooooooooooo
-------------------------	---------------------------	------------------	-----------------	----------------------	---------------	--------------------------------

Bootstrap semi-parametrico nonlineare

Si rinuncia all'ipotesi iid per gli errori, ammettendo che abbiano una struttura di dipendenza nonlineare di carattere generale all'interno della stazionarietà.

Perciò stimato ξ e i residui $e = (e_1, \dots, e_n)$, si simula Y_j^* tramite

$$y_t^* = \hat{\alpha} y_{t-1}^* + e_t^*$$

$$e_t^* = \text{bootstrap a blocchi da } (e_1, \dots, e_n)$$

dove il bootstrap non-parametrico a blocchi è descritto nel lucido successivo.

48 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooo●oooooooooooo
-------------------------	---------------------------	------------------	-----------------	----------------------	---------------	--------------------------------

Bootstrap stazionario a blocchi

Nel Block-stationary-bootstrap l'idea è quella di costruire il campione simulato $Y_j^* = (y_{j1}^*, \dots, y_{jN}^*)$ come unione di un certo numero k_j di blocchi di lunghezza variabile estratti dai dati osservati $y = (y_1, \dots, y_n)$, con $n \leq N$.

In pratica si definisce

$$Y_j^* = (Y_{j1}^*, \dots, Y_{jk_j}^*)$$

dove ciascun blocco

$$Y_{ji}^* = (y_{ti}, \dots, y_{ti+l_i})$$

è scelto con inizio casuale $t_i = \text{Uniforme}(1, n)$ e lunghezza casuale l_i .

49 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooo●oooooooooooo
-------------------------	---------------------------	------------------	-----------------	----------------------	---------------	--------------------------------

NB: occorre verificare che

$$t_i + l_i \leq n$$

Inoltre k_j è tale che

$$\sum_i^{k_j-1} (l_i + 1) = a_j < N$$

infine l'ultimo blocco è troncato: $l_{k_j} = N - a_j$ in modo da garantire la dimensione corretta a y_j^* .

50 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooo●oooo
-------------------------	---------------------------	------------------	-----------------	----------------------	---------------	--------------------------------

Esempi:

La lunghezza dei blocchi può essere:

- 1 $I_i = \text{Poisson}(L)$
- 2 $I_i = \text{geom}(1/L)$
- 3 $I_i = \text{Uniforme}(1, 2L) \Rightarrow$ crea anomalie sul correlogramma

51 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooo●oooo
-------------------------	---------------------------	------------------	-----------------	----------------------	---------------	--------------------------------

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooo●oooo
-------------------------	---------------------------	------------------	-----------------	----------------------	---------------	--------------------------------

Bootstrap e stazionarietà

Finora abbiamo assunto che y_t fosse stazionario.

- Se y_t è un ARIMA(p,1,q) tecnicamente è possibile eseguire uno dei metodi visti ma occorre tener conto che:
 - le traiettorie simulate hanno proprietà statistiche analoghe a quelle di y_t
 - non sono "verosimili" nel senso che le differenze $y_t - y_t^*$ divergono.
- Se y_t è un regARIMA(p,0,q) del tipo

$$y_{j,t} = m_t + ARMA_t$$

Allora si possono fare delle simulazioni condizionate a \hat{m}_t

52 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooo●oooo
-------------------------	---------------------------	------------------	-----------------	----------------------	---------------	--------------------------------

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooo●oooo
-------------------------	---------------------------	------------------	-----------------	----------------------	---------------	--------------------------------

Matlab

- bootstrap parametrico (ε_t Gaussiana o t di Student)


```
>> Ystar = simulate(mhat,...  
n, 'NumPaths', B, 'X', X);2
```
- bootstrap semiparametrico


```
>> [Ystar, idx] = bootstrp(B,...  
@(resstar) filter(mhat,resstar,'X',X),...  
res);3
```

²mhat: modello stimato, n: lunghezza serie, B replicazioni bootstrap, X: regressori (se presenti)

³res: residui modello stimato

53 / 56

54 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooooo
-------------------------	---------------------------	------------------	-----------------	----------------------	---------------	-----------------------------

Matlab

la function `stationaryBB()`⁴ esegue il bootstrap a blocchi.
Nell'esempio `ese_2018_block_bootstrap.m` si ricampiona un modello

$$y_{j,t}^* = \hat{m}_t + e_{j,t}^*$$

dove m_t è modellato tramite una spline mentre e_t^* è ottenuto col bootstrap a blocchi applicato ai residui della spline:

$$e_t = y_t - \hat{m}_t.$$

⁴NB: non è una funzione standard di Matlab ma è disponibile su ILIAS

55 / 56

MA+ARMA oo oooooo	ARIMA+MLE oooo oooo	regARIMA oooo	AIC/BIC oooo	DIAG oooooooooooo	VAL oooooo	Bootstrap oooooooooooooo
-------------------------	---------------------------	------------------	-----------------	----------------------	---------------	-----------------------------

56 / 56

SMS2. aa1920

Modelli stocastici

Lezione 05-B

Inferenza sulla media per dati correlati

SMS2. aa1920

Premessa

Il Teorema Limite Centrale per dati iid

$$y_1, \dots, y_n \text{ iid}(\mu, \sigma^2) \Rightarrow \frac{\bar{y}_n - \mu}{\sigma/\sqrt{n}} \cong N(0, 1)$$

Aspetti chiave

1. $E(\bar{y}_n) = \mu \quad (\rightarrow)$
2. $Var(\bar{y}_n) = \frac{\sigma^2}{n} \quad (\rightarrow)$
3. $N(\dots) \text{ per } n \rightarrow \infty$

Esempi Positivi (Classici):

SMS2. aa1920

1. $m = 1000$ simulazioni di $y_1, \dots, y_{10} \text{ NID}(\mu, \sigma^2)$

$$\Rightarrow \bar{y}_n \equiv N\left(\mu, \frac{\sigma^2}{n}\right)$$

La normalità vale $\forall n$ in modo esatto

2. $m = 1000$ simulazioni di $n = 10 \exp(\lambda = 1)$

$$\Rightarrow \bar{y}_n \cong N\left(\mu, \frac{\sigma^2}{n}\right)$$

I 3 aspetti chiave sono soddisfatti, in particolare media e varianza sono esatte

L05B: iid or not iid ?

3

Ruolo delle Ipotesi IID

SMS2. aa1920

Andiamo a considerare separatamente il ruolo delle ipotesi

1. Uguale distribuzione
2. Indipendenza

Vogliamo capire cosa succede al noto risultato

$$Var(\bar{y}_n) = \frac{\sigma^2}{n}.$$

quando le y_t sono correlate.

L05B: iid or not iid ?

4

Varianza della media per dati correlati

Posto, per semplicità, $\mu = 0$, abbiamo

$$\begin{aligned} Var(\bar{y}_n) &= E\left(\frac{1}{n} \sum_t y_t\right)^2 = \frac{1}{n^2} \sum_{s,t=1}^n E(y_s y_t) = \\ &= \frac{1}{n^2} \sum_{s,t=1}^n \gamma(y_s y_t) = \end{aligned}$$

dove γ indica l'autocovarianza.

L05B: iid or not iid ?

5

Nel caso di processi stazionari, abbiamo che

$$\rho(y_s, y_t) = \rho(|s - t|) = \frac{\gamma(|s - t|)}{\gamma(0)}$$

da cui segue che

$$\begin{aligned} Var(\bar{y}_n) &= \frac{1}{n^2} \sum_{s,t=1}^n \gamma(s - t) = \frac{1}{n} \sum_{h=-n+1}^{n-1} \frac{n-h}{n} \gamma(h) \\ &= \frac{\sigma^2}{n} \left(1 + 2 \sum_{h=1}^{n-1} \frac{n-h}{n} \rho(h) \right) \cong \frac{\sigma^2}{n} \left(1 + 2 \sum_{h=1}^{n-1} \rho(h) \right) \end{aligned}$$

L05B: iid or not iid ?

6

Traccia dimostrazione

Il limite della slide precedente

$$nVar(\bar{y}_n) = \sum_{h=-n}^n \frac{n-h}{n} \gamma(h) \rightarrow \sum_{h=-\infty}^{+\infty} \gamma(h)$$

è facilmente dedotto nel caso in cui $\gamma(h)$ tende a zero con velocità tale per cui $\lim_{n \rightarrow \infty} \sum_{h=-n}^n h|\gamma(h)| < \infty$. Infatti allora

$$\sum_{h=-n}^n \frac{n-h}{n} \gamma(h) = \sum_{h=-n}^n \gamma(h) - \frac{1}{n} \sum_{h=-n}^n h\gamma(h) \rightarrow \sum_{h=-\infty}^{+\infty} \gamma(h)$$

L05B: iid or not iid ?

7

Numero effettivo di osservazioni

Uguagliando la formula trovata con quella per osservazioni IID troviamo il cosiddetto numero effettivo di osservazioni.

Infatti da

$$Var(\bar{y}_n) = \frac{\sigma_y^2}{n} \left(1 + 2 \sum_{h=1}^{n-1} \frac{n-h}{n} \rho(h) \right) = \frac{\sigma_y^2}{n_e}$$

otteniamo

$$n_e = n \left(1 + 2 \sum_{h=1}^{n-1} \frac{n-h}{n} \rho(h) \right)^{-1} \cong n \left(1 + 2 \sum_{h=1}^{n-1} \rho(h) \right)^{-1}$$

L05B: iid or not iid ?

8

Osservazioni

1. Nel caso IID, ovviamente,

$$\rho(h) = 0 \text{ per } h \neq 0$$

e vale il risultato standard: $Var(\bar{y}_n) = \frac{\sigma^2}{n}$.

2. Quando y_t è stazionario il comportamento asintotico di

$$Var(\bar{y}_n)$$

dipende dal comportamento dell'autocorrelazione $\rho(h)$ per $|h| \rightarrow \infty$

3. In particolare per i processi a memoria breve, l'autocovarianza è assolutamente sommabile:

$$\sum_{h=-\infty}^{+\infty} |\rho(h)| < \infty$$

e, allora, per $n \rightarrow \infty$,

$$nVar(\bar{y}_n) = \sum_{h=-n}^n \frac{n-h}{n} \gamma(h) \rightarrow \sum_{h=-\infty}^{+\infty} \gamma(h) = \tau^2 < \infty$$

L05B: iid or not iid ?

9

4. Nel caso AR(1) si ha che il numero effettivo di osservazioni è:

$$n_e = n \frac{1-\alpha}{1+\alpha}$$

(per la dimostrazione si veda più sotto).

5. I processi in cui le covarianze sono "assolutamente-sommabili" si dicono a memoria breve e viceversa a "memoria lunga".

6. Il caso a memoria lunga di tipo frazionario è studiato in L06-LRD.

7. Fra i processi a memoria breve, particolare importanza hanno i processi in cui

$$\rho(h) \rightarrow 0$$

con velocità esponenziale. Per esempio questo accade per i processi di tipo ARMA.

L05B: iid or not iid ?

10

Esempio (Media Mobile)

$$\begin{aligned}y_t - \mu &= \varepsilon_t + c\varepsilon_{t-1} \\ \varepsilon_t &\equiv NID(0, \sigma_\varepsilon^2)\end{aligned}$$

Ne consegue:

$$\begin{aligned}E(y_t) &= \mu \\ \gamma_{t,t} &= Var(y_t) = (1+c^2)\sigma_\varepsilon^2 =: \sigma_y^2 \\ \gamma_{t,t\pm 1} &= E(\varepsilon_t + c\varepsilon_{t-1})(\varepsilon_{t-1} + c\varepsilon_{t-2}) = c\sigma_\varepsilon^2 \\ \gamma_{t,t\pm h} &= 0 \quad \forall h > 1\end{aligned}$$

NB: Un processo $MA(q)$ si dice a memoria finita perché $\gamma(h) = 0$ per $|h| > q$.

L05B: iid or not iid ?

11

Per calcolare $Var(\bar{y}) = \frac{1}{n^2} \sum_{s,t} \gamma_{s,t}$ osserviamo che

$$\frac{1}{n} \sum_{s \neq t} \gamma_{s,t} = \frac{1}{n} \sum_{s=t\pm 1} \gamma_{s,t} = 2 \frac{n-1}{n} c \sigma_\varepsilon^2$$

da cui

$$\begin{aligned}Var(\bar{y}_n) &= (1+c^2) \frac{\sigma_\varepsilon^2}{n} + \frac{n-1}{n^2} 2c \sigma_\varepsilon^2 = \frac{\sigma_\varepsilon^2}{n} \left(1 + c^2 + 2c - \frac{2c}{n}\right) \\ &= \frac{\sigma_\varepsilon^2}{n} \left[(1+c^2) + 2c \left(1 - \frac{1}{n}\right)\right] \\ &= \frac{\sigma_y^2}{n} \left(1 + \frac{2c}{1+c^2} \left(1 - \frac{1}{n}\right)\right)\end{aligned}$$

L05B: iid or not iid ?

12

Pertanto, per n grande, trascurando il termine $\frac{1}{n^2}$, abbiamo

$$Var(\bar{y}_n) \cong \frac{\sigma_y^2}{n} \left(1 + \frac{2c}{1+c^2}\right) \begin{cases} > \frac{\sigma_y^2}{n} & \text{se } c > 0 \\ = \frac{\sigma_y^2}{n} & \text{se } c = 0 \\ < \frac{\sigma_y^2}{n} & \text{se } c < 0 \end{cases}$$

In particolare se $c = 1$

$$Var(\bar{y}_n) \cong 2 \frac{\sigma_y^2}{n} \quad \text{per ogni } c$$

In questo caso gli IC basati sulla formula IID sono distorti con una dimensione sottostimata di circa il 30%, cioè

$$0.30 \cong 1 - \frac{1}{\sqrt{2}}$$

e hanno, in questo caso, una confidenza effettiva $(1 - \alpha^*)$ assai più bassa di quella nominale $(1 - \alpha)$.

L05B: iid or not iid ?

13

Esempio (Autoregressione AR(1))

$$\begin{aligned}y_t - \mu &= \alpha(y_{t-1} - \mu) + \varepsilon_t \text{ con } |\alpha| < 1 \\ \varepsilon_t &\equiv NID(0, \sigma_\varepsilon^2) \\ y_0 &\equiv N\left(\mu, \frac{\sigma_\varepsilon^2}{1-\alpha^2}\right)\end{aligned}$$

Ne conseguono le note proprietà dei modelli AR(1):

$$y_t - \mu = \sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j} \equiv MA(\infty) \quad \text{con} \quad E(y_t) = \mu$$

$$\gamma_{t,t} = Var(y_t) = E\left(\sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j}\right)^2 = \frac{\sigma_\varepsilon^2}{1-\alpha^2} = \sigma_y^2$$

$$\gamma(h) = \gamma_{t,t+h} = E\left(\sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j} \sum_{i=0}^{\infty} \alpha^i \varepsilon_{t+h-i}\right) = \alpha^{|h|} \frac{\sigma_\varepsilon^2}{1-\alpha^2}$$

L05B: iid or not iid ?

14

Per calcolare $Var(\bar{y}_n)$ interessa la matrice di varianze covarianze di

$$Y = (y_1, \dots, y_n)'$$

cioè

$$\Gamma = (\gamma_{i,j})_{i,j=1,\dots,n} = \gamma(|i-j|) = \frac{\sigma_\varepsilon^2}{1-\alpha^2} (\alpha^{|i-j|})_{i,j=1,\dots,n}$$

perciò ricordando il teorema in apertura:

$$\begin{aligned} Var(\bar{y}_n) &\cong \frac{1}{n} \sum_{h=-\infty}^{+\infty} \gamma(h) = \frac{\sigma_y^2}{n} \sum_{h=-\infty}^{+\infty} \alpha^{|h|} \\ &= \frac{\sigma_y^2}{n} \frac{1+\alpha}{1-\alpha} \\ n_e &= n \frac{1-\alpha}{1+\alpha} \end{aligned}$$

L05B: iid or not iid ?

15

Caso generale

Se indichiamo

$$Y = (y_1, \dots, y_n)$$

la media campionaria può essere scritta come

$$\bar{y} = \frac{1}{n} \mathbf{1}'_n Y$$

dove $\mathbf{1}_n$ è un vettore $n \times 1$ composto di uno.

Se Y ha matrice di varianze-covarianze

$$\Sigma = V(Y)$$

allora

$$V(\bar{y}) = \frac{1}{n^2} \mathbf{1}'_n \Sigma \mathbf{1}_n$$

NB: questa formula vale anche nel caso non stazionario.

16

Esercizi

- Calcolare α^* per l'IC sulla MA(1).
- Simulazione:** CLT_cntresempio_MA.m
- Matlab:** costruire una funzione per calcolare $Var(\bar{y})$ quando y_t è AR(1) o più in generale quando $Y = (y_1, \dots, y_n)$ ha matrice di varianze covarianze Γ .
- Matlab:** costruire una funzione per calcolare $Var(\bar{y})$ a partire da autocorr.
- Matlab - Esempio MA(1) :** Confrontare le formule per $Var(\bar{y})$ viste sopra con la loro valutazione Monte Carlo
- Matlab - Esempio AR(1) :** Confrontare le formule per $Var(\bar{y})$ viste sopra con la loro valutazione Monte Carlo

L05B: iid or not iid ?

17

Conclusioni

- Quanto sopra vale, mutatis mutandis, anche per la regressione e i minimi quadrati.
- Il contenuto informativo (sulla media) dipende anche dalla dipendenza presente fra le osservazioni.
- E' importante conoscere la struttura di dipendenza per usare le misure di incertezza appropriate.
- Nel caso di $y = \mu + ARIMA$ l'inferenza corretta su μ è fatta stimando il corrispondente modello in classe **regARIMA** con **Intercept**= μ .

L05B: iid or not iid ?

18

Modelli stocastici

Lezione 07

Modelli nello spazio degli stati

Modello nello spazio degli stati

Questi modelli detti **state-space models** sono anche noti come **Kalman Filter models** o **Dynamic Linear System models (DLS)**.

Consideriamo un sistema il cui stato è definito dal vettore $m - \text{dim}$, indicato con x_t , e di cui si dispongono le osservazioni $k - \text{dim}$, indicate con y_t .

Equazione di stato

La dinamica del fenomeno è data da un modello markoviano

$$x_t = Ax_{t-1} + \omega_t$$

con innovazione $\omega_t \sim NID(0, \Sigma_\Omega)$ e valori iniziali $x_0 \sim N(\mu_0, \Sigma_0)$.

NB:

1. $\Sigma_\Omega = Q = \mathbf{B}\mathbf{B}'$
dove \mathbf{B} è una matrice $m \times q$.
2. Per la stazionarietà (stabilità) di x_t occorre che gli autovalori di A siano minori di 1 in valore assoluto.

Equazione di misura

Lo stato del sistema x_t non è direttamente osservabile e le osservazioni sono legate agli stati dall'equazione:

$$y_t = Cx_t + \varepsilon_t$$

con $\varepsilon_t \sim NID(0, \Sigma_\varepsilon = DD' = R)$.

Covariate u_t

Il modello di cui sopra è facilmente estendibile al caso con input osservati u_t

$$y_t - \beta'u_t = Cx_t + \varepsilon_t$$

Correlazione fra le due equazioni

Per ora assumiamo

$$E(\varepsilon\omega') = 0$$

Inferenza sullo stato

Noto il modello $(A, B, C, D, \mu_0, \Sigma_0)$ può interessare:

1. Stima o smoothing (ex post)

$$\hat{x}_t^n = E(x_t | y_1, \dots, y_n)$$

2. Filtro (on-line)

$$\hat{x}_t^t = E(x_t | y_1, \dots, y_t)$$

3. Previsione (ex ante)

$$\hat{x}_t^{t-1} = E(x_t | y_1, \dots, y_{t-1})$$

Le formule ricorsive per questi conti prendono il nome da **Kalman** e sono interpretabili come medie condizionate per **processi Gaussiani**.

Esistono estensioni per processi **non Gaussiani** e **non lineari**.

Esempio 1: Local level

Consideriamo il modello random walk con errore di misura:

$$y_t = x_t + \varepsilon_t$$

$$x_t = x_{t-1} + \omega_t$$

con $x_0 \equiv N(\mu_0, \Sigma_0)$. Inoltre $Var(\omega_t) = \sigma_\omega^2$ e $Var(\varepsilon_t) = \sigma_\varepsilon^2$.

Se si usa $\theta = (\sigma_\omega^2, \sigma_\varepsilon^2)$ come vettore dei parametri da stimare si possono avere problemi di identificabilità e spesso si preferisce fissarne uno. Per esempio σ_ε^2 può essere noto se ε è un errore di misura.

In questo caso

$$\hat{\mu}_t = E(x_t | y_1, \dots, y_n)$$

è un classico "smoother" dei dati.

Matlab: l'oggetto in classe `ssm` è dato da `A=1, B=sigma_omega, C=1, D=sigma_epsilon :` `ese_2018_local_level.m`

Esempio 2: Stima della velocità di una particella

Sia y_t la posizione osservata di una particella in un canale unidimensionale a istanti di tempo regolari $t = 0, \Delta t, 2\Delta t, \dots$. Interessa stimare la velocità

$$v_t = \frac{\partial y}{\partial t}$$

In prima approssimazione, la velocità è stimata da

$$\frac{\partial y}{\partial t} \approx \frac{y_t - y_{t-1}}{\Delta t}$$

Tuttavia se y_t è affetto da errore di misura

$$y_t = x_{1t} + \varepsilon_{1t}$$

allora la stima di cui sopra è inattendibile perché amplifica gli errori di un fattore $\frac{\sigma_{\varepsilon_1}}{\Delta t}$.

Esempio 2 (segue): Impostazione spazio degli stati

Eq.ne di stato per posizione e velocità $x_t = (p_t, v_t)'$:

$$p_t = p_{t-1} + \Delta t v_t + \omega_{1t}$$

$$v_t = v_{t-1} + \omega_{2t}$$

$$x_0 \equiv N(\mu_0, \Sigma_0)$$

Eq.ne di misura

$$y_t = p_t + \varepsilon_t$$

con

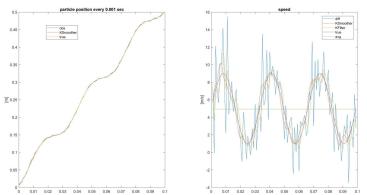
$$Var(\omega_t) = diag(\sigma_{\omega_1}^2, \sigma_{\omega_2}^2) \quad \text{e} \quad Var(\varepsilon_t) = \sigma_\varepsilon^2.$$

Esempio 2 (segue): soluzione in matlab

In classe `ssm` abbiamo

$$A = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix} \text{ e } B = \text{diag}(\sigma_{\omega_1}, \sigma_{\omega_2})$$

$C = 1$ e $D = \sigma_{\varepsilon}$.



Matlab: ese_2018_particle_speed.m

Modelli tempo varianti

Le matrici del modello variano con t :

$$y_t = C_t \mu_t + \varepsilon_t$$

$$x_t = A_t x_{t-1} + \omega_t$$

con

$$\text{Var}(\omega_t) = Q_t \quad \text{e} \quad \text{Var}(\varepsilon_t) = R_t$$

In questo caso A_t , C_t , Q_t e R_t sono matrice note oppure son funzioni di un vettore di parametri da stimare:

$$(A_t, Q_t, R_t, \mu_0, \Sigma_0) = (A_t(\theta), Q_t(\theta), R_t(\theta), \mu_0(\theta), \Sigma_0(\theta))$$

Esempio 3: Regressione dinamica

Abbiamo un modello di regressione

$$y_t = \alpha_t + \beta_t x_t + \varepsilon_t$$

dove α_t e β_t variano con t come un random walk:

$$\begin{pmatrix} \alpha_t \\ \beta_t \end{pmatrix} = \begin{pmatrix} \alpha_{t-1} \\ \beta_{t-1} \end{pmatrix} + \begin{pmatrix} \omega_{1,t} \\ \omega_{2,t} \end{pmatrix}$$

con

$$\text{Var}(\omega_t) = Q \quad \text{e} \quad \text{Var}(\varepsilon_t) = R$$

Allora $\theta = (Q, R)$ e i coefficienti della regressione sono stimati dal Kalman smoother

$$\hat{\alpha}_t = E(\alpha_t | y_1, \dots, y_n) \quad \text{e} \quad \hat{\beta}_t = E(\beta_t | y_1, \dots, y_n)$$

Matlab: ese_2020_StateSpace_QA.m + tVarPar.m (incluso)

Stima EM per State Space

(opzionale)

Supposta nota la matrice di osservazione C , si vuole stimare il set di parametri

$$\Theta = (A, Q, R, \mu_0, \Sigma_0)$$

La log-verosimiglianza completa di

$$(\mathbf{X}, \mathbf{Y}) = x_0, (x_1, y_1), \dots, (x_n, y_n)$$

è data da:

$$\begin{aligned} -2 \log L(\Theta | \mathbf{X}, \mathbf{Y}) &= \log |\Sigma_0| + (x_0 - \mu_0)' \Sigma_0^{-1} (x_0 - \mu_0) + \\ & n \log |Q| + \sum_{t=1}^n (x_t - Ax_{t-1})' Q^{-1} (x_t - Ax_{t-1}) + \\ & n \log |R| + \sum_{t=1}^n (y_t - Cx_t)' R^{-1} (y_t - Cx_t) \end{aligned}$$

Indichiamo i momenti a posteriori (Kalman smoother) con:

$$x_t^n = E(x_t | \mathbf{Y})$$

$$V_t^n = V(x_t | \mathbf{Y})$$

$$V_{t,t-1}^n = Cov(x_t, x_{t-1} | \mathbf{Y})$$

$$P_t = E(x_t x_t' | \mathbf{Y}) = V_t^n + x_t^n (x_t^n)'$$

$$P_{t,t-1} = E(x_t x_{t-1}' | \mathbf{Y}) = V_{t,t-1}^n + x_t^n (x_{t-1}^n)'$$

Lo step i -esimo dell'algoritmo EM è dato da:

$$\begin{aligned} A^i &= \left(\sum_{t=1}^n P_{t,t-1} \right) \left(\sum_{t=1}^n P_{t-1} \right)^{-1} \\ Q^i &= \frac{1}{n} \left(\sum_{t=1}^n P_t - A^i \left(\sum_{t=1}^n P_{t,t-1} \right)' \right)' \\ R^i &= \frac{1}{n} \sum_{t=1}^n ((y_t - Cx_t^n)(y_t - Cx_t^n)' + A V_t^n A') \\ \mu_0^i &= x_0^n, \quad \mathbf{e} \quad \Sigma_0^i = V_0^n \end{aligned}$$

Traccia dimostrazione

- supporre note anche le x_t
- usare x ed y scalari
- per A ricordare i minimi quadrati
- per Q ricordare che $V_{res} = V(x_t) - \hat{A}V(x_{t-1})$
- Poiché le x non sono note, osservare che l'EM equivale ad usare, iterativamente, al posto di x_t , $x_t x_{t-1}'$ etc. le loro medie a posteriori $E(x_t | Y), \dots$

Filtro di Kalman

Supponiamo note le matrici

$$A, C, Q, R, \mu_0, \Sigma_0$$

e (per semplicità) costanti nel tempo.

Ci poniamo il problema di "**stimare ricorsivamente**" lo stato x_t partendo dalle osservazioni y_t in ipotesi di normalità.

$$\mathbf{Y}_t = (y_1, \dots, y_t)$$

Definiamo, per $t \leq \tau$, i seguenti momenti condizionati:

$$x_t^\tau = E(x_t | \mathbf{Y}_\tau)$$

$$V_t^\tau = \text{Var}(x_t | \mathbf{Y}_\tau)$$

$$V_{t,t-1}^\tau = \text{Cov}(x_t, x_{t-1} | \mathbf{Y}_\tau)$$

Allora il filtro è dato dalle seguenti ricorsione per $t = 1, \dots, n$:

1. Previsione

$$x_t^{t-1} = Ax_{t-1}^{t-1}$$

2. Varianza della previsione

$$V_t^{t-1} = AV_{t-1}^{t-1}A' + Q$$

3. Guadagno

$$K_t = V_t^{t-1}C'(CV_t^{t-1}C' + R)^{-1}$$

4. Stima

$$x_t^t = x_t^{t-1} + K_t(y_t - Cx_t^{t-1})$$

5. Varianza della stima

$$V_t^t = (I - K_t C)V_t^{t-1}$$

Kalman smoother

Lo smoother infine è dato dalla ricorsione backword per $t = n, n-1, \dots, 1$:

$$J_{t-1} = V_{t-1}^{t-1}A'(V_t^{t-1})^{-1}$$

$$x_{t-1}^n = x_{t-1}^{t-1} + J_{t-1}(x_t^n - Ax_{t-1}^{t-1})$$

$$V_{t-1}^n = V_{t-1}^{t-1} + J_{t-1}(V_t^n - V_t^{t-1})J_{t-1}'$$

$$P_t = V_t^n + x_t^n(x_t^n)'$$

$$P_{t,t-1} = V_{t,t-1}^n + x_t^n(x_{t-1}^n)'$$

$$V_{t-1,t-2}^n = V_{t-1}^{t-1}J_{t-2} + J_{t-1}(V_{t,t-1}^n - AV_{t-1}^{t-1})J_{t-2}'$$

con valori iniziali

$$V_{n,n-1}^n = (I - K_n C)AV_{n-1}^{n-1}$$

Perciò possiamo usare x_{t-1}^n, P_t e $P_{t,t-1}$ nell'algoritmo EM.

Kalman Filter e normalità

In ipotesi di normalità degli errori ε e delle innovazioni ω , costruendo iterativamente le y_t e le x_t per $t = 1, 2, \dots, n$, condizionatamente ai valori iniziali ed alle covariate, si deduce che

$$(X, Y) = x_0, (x_1, y_1), \dots, (x_n, y_n)$$

ha la distribuzione congiunta gaussiana che abbiamo visto sopra.

Il calcolo di

$$x_t^t = E(x_t | \mathbf{Y}_t)$$

equivale quindi al calcolo del piano di regressione di una normale multivariata.

A tal fine ricordiamo che, per la normale multivariata

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \equiv N_{k_1+k_2} \left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right),$$

il piano di regressione è dato da

$$E(\mathbf{x}_1 | \mathbf{x}_2 = \mathbf{x}) = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2),$$

con varianza condizionata

$$Var(\mathbf{x}_1 | \mathbf{x}_2) = \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Arrivati all'istante $t - 1$, a priori rispetto ad y_t , la miglior stima di x_t (o previsione) è

$$x_t^{t-1} = E(x_t | \mathbf{Y}_{t-1}).$$

Poichè

$$x_t = Ax_{t-1} + \omega_t$$

sostituendo, abbiamo che

$$x_t^{t-1} = E(Ax_{t-1} + \omega_t | \mathbf{Y}_{t-1}) = AE(x_{t-1} | \mathbf{Y}_{t-1}) = Ax_{t-1}^{t-1}$$

Similmente la varianza di $(x_t | \mathbf{Y}_{t-1})$, cioè dell'errore di previsione, è data da

$$\begin{aligned} V_t^{t-1} &= E((x_t - x_t^{t-1})(x_t - x_t^{t-1})' | \mathbf{Y}_{t-1}) = \\ &= E((Ax_{t-1} + \omega_t - x_t^{t-1})(Ax_{t-1} + \omega_t - x_t^{t-1})' | \mathbf{Y}_{t-1}) = \\ &= E(A(x_{t-1} - x_{t-1}^{t-1})(x_{t-1} - x_{t-1}^{t-1})' A' | \mathbf{Y}_{t-1}) + Q = \\ &= AE((x_{t-1} - x_{t-1}^{t-1})(x_{t-1} - x_{t-1}^{t-1})' | \mathbf{Y}_{t-1})A' \\ &= AV_{t-1}^{t-1}A' + Q \end{aligned}$$

perciò x_t ha distribuzione condizionata gaussiana, data da

$$(x_t | \mathbf{Y}_{t-1}) \equiv N(Ax_{t-1}^{t-1}, AV_{t-1}^{t-1}A' + Q).$$

Dopo aver osservato y_t , interessa la stima a posteriori di x_t , cioè:

$$x_t^t = E(x_t | \mathbf{Y}_t) = E(x_t^{t-1} | y_t, \mathbf{Y}_{t-1})$$

Il calcolo di questo valore atteso si riconduce ad una regressione lineare come spiegato sotto.

A tal fine, la stima a priori di y_t (o previsione) è :

$$y_t^{t-1} = Cx_t^{t-1}$$

ed il suo errore

$$e_t = y_t - y_t^{t-1} = C(x_t - x_t^{t-1}) + \varepsilon_t.$$

Ricordando che l'errore ε_t è incorrelato da x_t , abbiamo:

$$\Sigma_{12} = Cov(x_t, y_t | \mathbf{Y}_{t-1}) = E((x_t - x_t^{t-1}) e_t' | \mathbf{Y}_{t-1}) = V_t^{t-1} C'$$

Inoltre

$$\Sigma_{22} = V(e_t | \mathbf{Y}_{t-1}) = CV_t^{t-1} C' + R$$

perciò

$$\Sigma_{12}\Sigma_{22}^{-1} = K_t = V_t^{t-1} C' (CV_t^{t-1} C' + R)^{-1}$$

da cui si conclude

$$x_t^t = x_t^{t-1} + K_t e_t$$

Similmente dalla formula della varianza residua della normale

$$\Sigma_{12} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

si ottiene

$$\begin{aligned} V_t^t &= V_t^{t-1} - K_t C V_t^{t-1} \\ &= (I - K_t C) V_t^{t-1} \end{aligned}$$

Valori iniziale per EM

Caso scalare con $C = 1$.

Si osserva che l'autocovarianza delle osservazioni è

$$\gamma_y(h) = E(y_t y_{t-h}) = E((x_t + \varepsilon_t)(x_{t-h} + \varepsilon_{t-h})) = \begin{cases} \frac{\varrho}{1-a^2} + R & h = 0 \\ \gamma_x(h) = a^h \frac{\varrho}{1-a^2} & h > 0 \end{cases}$$

perciò, usando il metodo dei momenti, abbiamo:

$$\hat{a}_0 = \frac{\hat{\gamma}_y(2)}{\hat{\gamma}_y(1)}$$

$$\hat{Q}_0 = \hat{\gamma}_y(1) \frac{(1 - \hat{a}_0^2)}{\hat{a}_0} \quad \hat{R}_0 = \hat{\gamma}_y(0) - \frac{\hat{Q}_0}{1 - \hat{a}_0^2}$$

Caso generale

Ora per $h \geq 1$ si ha che la matrice di autocovarianza di un $VAR(1)$ è data dalle equazioni di Yule e Walker (cfr Lutkepohl)

$$\Gamma_x(h) = A\Gamma_x(h-1)$$

inoltre, similmente al caso scalare:

$$vec(\Gamma_x(0)) = (I - A \otimes A)^{-1} vec(Q)$$

Ora però

$$\Gamma_y(h) = CT\Gamma_x(h)C'$$

se $C = 1$ allora i valori iniziali sono facilmente calcolati da

$$\hat{A}_0 = \hat{\Gamma}_y(2)\hat{\Gamma}_y(1)^{-1}$$

NB: risultati esplicativi si hanno anche se C è quadrata e invertibile.

Osservazioni

Ottimalità del filtro:

In ipotesi di normalità x_t^t è la media condizionata. Se cadono queste ipotesi vale tuttavia ancora che x_t^t minimizza l'MSE fra tutti gli stimatori lineari.

Robustezza:

In presenza di dati fortemente non normali e/o nonlineari è bene usare opportune varianti del filtro (EKF, particle filter, etc).

Altre osservazioni:

- I risultati ottenuti sono condizionati ai valori iniziali x_0 e Σ_0
- Le matrice coinvolte sono spesso funzioni di un vettore θ di parametri $A(\theta), B(\theta), C(\theta), \dots$
in tal caso θ può essere stimato tramite *MLE* o *PEM*.
- L'estensione del filtro a modelli tempo varianti A_t, \dots etc. è immediata.

Bibliografia

1. Massima verosimiglianza:

- a. Ross (2003) Probabilità e Statistica per l'ingegneria e le scienze, Apogeo. Cap 7.2.
- b. Azzalini A. (2004) Inferenza statistica, Springer, §§ 2.1,2.2 e 3 fino a 3.3.4

2. Algoritmo EM e modelli State Space

- a. Shumway R.H. and Stoffer D.S. (2006) *Time series analysis and its applications with R examples, 2nd edn*. Springer, Capitolo 6, in particolare §63
- b. **APPROFONDIMENTI:** McLachlan G.J., Krishnan T. (1997) *The EM Algorithm and Extensions*, Wiley

3. Stima bayesiana:

- Ross (2003) Probabilità e Statistica per l'ingegneria e le scienze, Apogeo. Cap 7.8, pp269-274. (si consiglia di guardare qualche **problema**, pp 274 e segg.)