

APPUNTI DEL CORSO

IDENTIFICAZIONE DEI MODELLI E ANALISI DEI DATI

A.A. 2017/2018 - UNIVERSITY OF BERGAMO

PARTE I: SISTEMI STATICI

AUTORE: MIRKO MAZZOLENI



L'uso e la distribuzione di questi appunti è consentita previa citazione dell'autore e della fonte originari

Corsi di IDENTIFICAZIONE DEI MODELLI & ANALISI DEI DATI

I) IDENTIFICAZIONE DEI MODELLI

Modello: Descrizione matematica di un fenomeno o di un sistema

trovare il legame tra queste
grandezze e descrivere
matematicamente

L'economia: relazione tra reddito ed educazione

L'sociale: relazione tra luogo di abitazione e criminalità

L'fisico: relazione tra massa e peso di una persona

Sistema: Mecanismo o struttura che trasforma input (cause) in output (effetti)

$$u \rightarrow [S] \rightarrow y$$

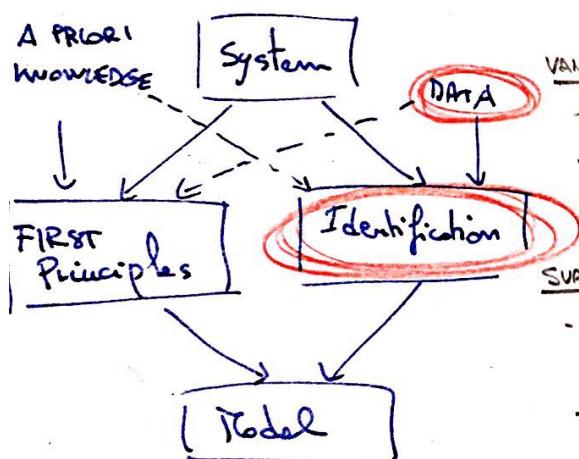
$$P = M \cdot g$$

$$V = R \cdot I$$

conoscenza
A priori
presente

Due approcci fondamentali:

a) WHITE BOX MODELING: - approccio basato su leggi e principi base delle FISICA e MATEMATICA
L'es. modello di un condensatore $I(t) = C \cdot \frac{dV(t)}{dt}$



VANTAGGI
 - conoscenza del significato delle variabili: (C: capacità di condensare)
 - generalizzabile: se conosce C, il modello vale anche per le altre
 L'ipotesi conosce che relazione CAUSALE tra C ed I(t)

Svantaggi
 - richiede conoscenza avanzata delle leggi del problema specifico
 L'ente, costi alti
 - per sistemi complessi la scrittura di molte equazioni diventa impossibile
 - limitato ai campi in cui esistono leggi causali

b) BLACK BOX MODELING: - approccio basato su DATI Sperimentali

VANTAGGI
 - presiedono dal particolare tipo di problema, limitandosi a caratterizzare il legame tra le variabili $y = f(u)$
 - veloci da costruire

Svantaggi
 - non interpretabili fisicamente
 - non generali, dipendono dal tipo di dati acquisiti. Per ogni modifica del sistema, bisogna ripetere l'esperimento



①

2) ANALISI DEI DATI

- Determinare le caratteristiche statistiche dei dati e delle variabili misurate
 - L'essi infatti sono affetti da RUMORE ed INCERTITUDINE
 - L'media L'correlazione tra variabili.
 - L'varianza L'distribuzione probabilistica
- Individuare una STRUTTURA, delle regolarità (se ci sono)
 - L'i dati presentano dei "PATTERN" riconoscibili o sono RANDOM?
 - L'osservazioni ACCENNANO ad alcuni che AI SOLI individuano pattern

STATISTICA
DESCRITTIVA

LE PROCEDURE 1) ed 2)

1) e 2) sono strettamente interconnesse:

- i) Spesso l'analisi preliminare dei dati dà indicazioni sul modello migliore per descriverli
- ii) Tecniche di analisi dei dati sono usate per descrivere la tendenza del modello
- iii) Una rappresentazione probabilistica dei dati fa leggi od un modello probabilistico capace di gestire l'incertezza
 - L'è sia sulla misura
 - L'è sia sulla conoscenza della realtà
 - L'quantifica quello che non si

DECLINERETE le due procedure sia per sistemi statici che per sist. dinamici

L'SISTEMI STATICI: le sole conoscenze delle variabili u è sufficiente a calcolare $y \Rightarrow V(t) = R \cdot I(t)$

L'SISTEMI DINAMICI: bisogna sapere le condizioni iniziali: $\frac{dV(t)}{dt} = \frac{1}{C} \cdot I(t)$
per conoscere $V(t)$ deve sapere $V(t_0)$

L'equazione di STATO: $\begin{cases} \dot{x}_1(t) = \frac{1}{C} \cdot u(t) \\ y(t) = x_1(t) \end{cases}$

L'Fondamentale di automotrice !! \Rightarrow le G(s) era DATA.

Come trovare?: -WHITE BOX
-BLACK BOX

$G(z)$

(8)

1

INAD

SISTEMI DINAMICI

STIMA PARAMETRICA

- PROCESSI STOCASTICI
- PROPRIETÀ STIMATORI

CASO 1) o deterministico

- Shiva di parametri da una popolazione: $y = \mu + \sigma z$ (no assunzioni PDF)
- Modello lineare di processi stocastici razionali
- PREDIZIONE (od un passo)
- IDENTIFICAZIONE
 - Filtrato dei minimi quadrati (ARX)
 - " " " " Massima verosimiglianza (MLV)
- PERSISTENTE ECCITAZIONE
- ANALISI ASINTOTICA PER
- ANALISI INCERTITUDINE STIMA (N punti)

- Massima Verosimiglianza → SI ASSUNZIONE PDF
- Caso y
L'caso $y = g^T x + \epsilon$
- Regressione lineare
- Regressione logistica ⇒ SI ASSUNZIONE PDF
- CASO 2) o variabile casuale
 - Shiva Bayesiana

CONCETTI DI MACHINE LEARNING

- INTRO
- RISOLVIMENTO
- REGULARIZATION

- INTRO
- VARIANZA
- REGULARIZATION

(c)

RICHIATI DI STATISTICA

- Una variabile casuale V è una variabile definita a partire dall'osito S di un esperimento casuale \rightarrow Es. L'esperimento è il lancio di un monete. A seconda che cosa teste o croce, V assume un valore
L'indichiamo con v.c. come $V(S)$
L'el valore ossunto da V è segnato di un particolare esito S è $V(S)$

Se V può assumere diversi valori, come li descriv? \Rightarrow Assegnare una probabilità che ogni esito occorra \Rightarrow questo influenza sulla probabilità dei valori che V può assumere.

- Se V assume valori DISCRETI (V è una variabile casuale discreta)
 - L'funzione di probabilità di massa $p(x) = P(V=x)$ associa ad ogni valore x di V una probabilità

Indichiamo con x_i : valori di V . Se V può assumere m diversi valori, allora $\sum_{i=1}^m p(x_i) = 1$

Esempio n. 20

$$\begin{array}{ll} x_1=1 & p(x_1) = P(V=x_1) = P(V=1) = \frac{1}{6} \\ x_2=2 & p(x_2) = P(V=2) = \frac{1}{6} \\ | & | \\ x_6=6 & p(x_6) = P(V=6) = \frac{1}{6} \\ m=6 & \end{array} \quad \sum_{i=1}^6 p(x_i) = 6 \cdot \frac{1}{6} = 1$$

- Se V assume valori CONTINUI (V è una v.c. continua)

L'funzione di densità di probabilità $f(x)$
(pdf)

~~- $P(v=x)$~~ non ha senso
perché se i valori possano essere infiniti e possibili valori:
la prob. di un valore sarebbe $\frac{1}{\infty} = 0$

Es. V è l'ottanza di un uomo adulto
b) non ha senso chiedersi la probabilità che un uomo sia alto ESATTAMENTE 1,7235142... metri

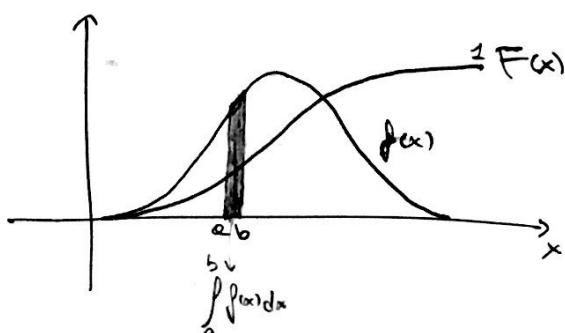
$$P(V \in [a, b]) = \int_a^b f(x) dx$$

$$\begin{aligned} &f(x) \geq 0 \\ &\int_{-\infty}^{+\infty} f(x) dx = 1 \end{aligned}$$

- Funzione di densità cumulata (cdf) o distribuzione di probabilità:

$$F(z) = \int_{-\infty}^z f(x) dx = P(X \leq z)$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



A FINI PRATICI X È DISCRETO
 $f(x_i) \propto P(V=x_i)$

①

- Il VALORE ATTESO di una v.c. continua è:

$$E[v] = \int_{-\infty}^{+\infty} x \cdot f(x) dx$$

Somma pesata dei valori x da v per osservare. I pesi sono le prob. $f(x)$. Peso di ogni valore per le sue probabilità di manifestarsi.

- LINEARITÀ: $E[\alpha v_1 + \beta v_2 + \gamma] = \alpha E[v_1] + \beta E[v_2] + \gamma \quad \forall \alpha, \beta, \gamma \in \mathbb{R}$

- La varianza di una v.c. continua è:

$$\text{Var}[v] = \int_{-\infty}^{+\infty} (x - E[v])^2 \cdot f(x) dx$$

$$= E[(x - E[x])^2]$$

- di quanto i valori x si scostano dalla loro media
- se più volte, v assume valori molto vicini fra loro

Osservazione

- $\text{Var}[v] \geq 0$. Se $\text{Var}[v] = 0$, la variabile v è deterministica (assume sempre un solo valore)

- Deviazione Standard: $\sigma[v] = \sqrt{\text{Var}[v]}$

$$\begin{aligned} \text{Var}[v] &= E[(v - E[v])^2] = E[v^2 - 2E[v]v + E[v]^2] = E[v^2] - 2E[v \cdot E[v]] + E[E[v]^2] \\ &= E[v^2] - 2E[v] \cdot E[v] + E[v]^2 = [E[v^2] - E[E[v]]]^2 \end{aligned}$$

$$\text{Var}[\alpha \cdot v_1 + \beta] = \alpha^2 \cdot \text{Var}[v_1] \quad \forall \alpha \in \mathbb{R} \quad \forall \beta \in \mathbb{R}$$

- Date due v.c. v_1 e v_2 si definisce il coefficiente di correlazione come:

$$\rho = \frac{E[(v_1 - E[v_1]) \cdot (v_2 - E[v_2])]}{\sigma[v_1] \cdot \sigma[v_2]}$$

- ρ indica il grado di dipendenza lineare tra v_1 e v_2 . Infatti se $v_2 = \alpha v_1 + \beta \Rightarrow \rho = 1$

- Se $\rho = 0$ le due variabili si dicono sconelte

- Date v_1 e v_2 si definisce covarianza la varianza come

$$\text{Cov}(v_1, v_2) = E[(v_1 - E[v_1]) \cdot (v_2 - E[v_2])]$$

e quindi:

$$\rho = \frac{\text{Cov}(v_1, v_2)}{\sigma[v_1] \cdot \sigma[v_2]}$$

- v_1 e v_2 sono sconelte se $\text{Cov}(v_1, v_2) = 0$

- Le precedenti definizioni si possono estendere al caso di vettore di variabili casuali $\bar{v} = [v_1, v_2, \dots, v_d]^T$

- distribuzioni di probabilità

$$\begin{aligned} F(x_1, x_2, x_3, \dots, x_d) &= P(v_1 \leq x_1, v_2 \leq x_2, \dots, v_d \leq x_d) \\ &= \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_d} f(x_1, x_2, \dots, x_d) dx_1 dx_2 \dots dx_d \end{aligned}$$

pdf congiunta

- valore atteso è un vettore colonna di d componenti

$$E[\bar{v}] = [E[v_1], E[v_2], \dots, E[v_d]]^T \in \mathbb{R}^{d \times 1}$$

- la varianza è una matrice $d \times d$ semidefinita positiva e simmetrica:

$$x \in \mathbb{R}^{d \times 1}$$

L'insieme $\{x \in \mathbb{R}^d \mid x \geq 0\}$ per numeri reali

L'insieme $\{x \in \mathbb{R}^d \mid x^T M x \geq 0\}$ è simmetrico se M è simmetrica

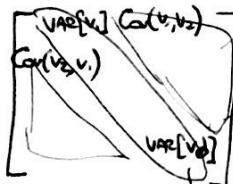
$$M \in \mathbb{R}^{d \times d}$$

$$x \in \mathbb{R}^d$$

L'autosvalore > 0 tranne $\lambda_1 = 0$

$$\text{Var}[\bar{v}] = \int_{\mathbb{R}^d} ((x - E[\bar{v}]) (x - E[\bar{v}])^T f(x) dx$$

$$x \in \mathbb{R}^d$$



$$d \times d$$

VARIANZE DI v_1, v_2, \dots, v_d

- COVARIANZE TDI

$$v_i \in \mathbb{R}, v_1 \in \mathbb{R}_3,$$

- SIMMETRICI: la covariante tra v_1 e v_2 è la stessa che tra v_2 e v_1 .

MATRICE DI VARIANZE - COVARIANZE

- Due variabili casuali v_1 e v_2 con funzione di probabilità composta f si dicono indipendenti se e solo se:

$$f(v_1, v_2) = f(v_1) \cdot f(v_2)$$

Teorema

Se v_1 e v_2 sono indipendenti, allora sono scorrutte

avremo μ e σ^2 determinati
come sono le caratteristiche dei miei dati

Es la densità di probabilità di y ha una forma gaussiana che dipende da $(P_f)_1$

V^2
 $d=2$

STIMA $\hat{\theta}$

Observe che ci concentriamo sulla **STIMA PARAMETRICA**. Vogliamo quindi trovare il parametro θ^* che le garantisce i dati $D = \{y(1), \dots, y(n)\}$

Interpretazione: dati come v.c. per definire le loro incertezze, $D = D(\bar{s}, \theta^*)$

l'incertezza i dati sono
volutamente associati alle
misurazioni

che associa ai dati un valore del parametro da stimare

L'incertezza delle misurazioni esiste $\bar{s} \Rightarrow D = D(\bar{s}, \theta^*)$

Un STIMATORE è una funzione $T(D(s, \theta^*))$. La STIMA è il risultato di un stimatore $\hat{\theta} = T(D(\bar{s}, \theta^*))$ → poiché il risultato di T dipende dall'intero s (da cui dipende i dati), allora lo stimatore è una variabile casuale dipendente da s

Es modo de fare MAD $\hat{\theta}_1 = T(D(s_1, \theta^*))$

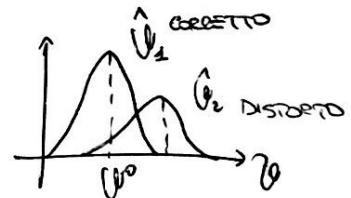
Misura il peso degli studenti $\hat{\theta}_2 = T(D(s_2, \theta^*))$

L'idea è misurare solo 5 studenti $D = \{y(1), \dots, y(5)\} \Rightarrow \hat{\theta}_2 = T(D(s_2, \theta^*))$

Ha senso quindi calcolare valore atteso e varianza di questa variabile casuale: in base a queste, valuteremo la bontà di un stimatore.

PROPRIETÀ DI UNO STIMATORE

- Un stimatore si dice corretto se e solo se: $E[\hat{\theta}] = \theta^*$
L'altro criterio è un errore sistematico di stima

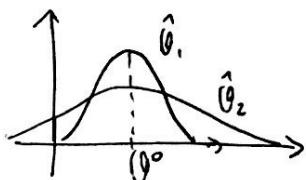


- Un stimatore si dice asintoticamente corretto se e solo se: $\lim_{N \rightarrow \infty} E[\hat{\theta}] = \theta^*$

- è una proprietà più debole.

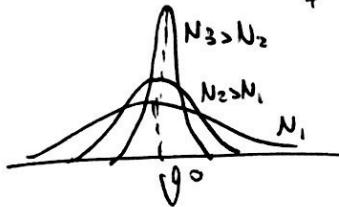
PRIMA DEF. CONSISTENTE

Se due stimatori sono entrambi corretti, qual è il migliore? Quello è minima varianza



$\hat{\theta}_1$ ha una maggiore probabilità di ritrovare una stima vicina al valore vero!

- Un stimatore si definisce consistente se: $\lim_{n \rightarrow \infty} \text{Var}[\hat{\theta}] = 0$



- La consistenza grafica che all'aumentare del numero dei campioni ha qualità delle stime aumentate

- Se $\hat{\theta}$ è corretto si ha che $\text{Var}[\hat{\theta}] = E[(\hat{\theta} - E[\hat{\theta}])^2] = E[(\hat{\theta} - \theta^*)^2] = \text{Var}[E^2]$

$$E = \hat{\theta} - \theta^*$$

- ERRORE DI STIMA

- Un stimatore $\hat{\theta}$ si dice ottimale se la sua varianza è la più piccola per una serie N di dati $D = \{y_1, y_2, \dots, y_N\}$

ES STIMATORE MEDIA

Siano $D = \{y_1, \dots, y_N\}$ variabili casuali con media μ . Il stimatore media campionaria

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N y_i \text{ è corretto. Infatti } E[\hat{\mu}] = E\left[\frac{1}{N} \sum_{i=1}^N y_i\right] = \frac{1}{N} \sum_{i=1}^N E[y_i] = \frac{1}{N} \cdot N \cdot \mu = \mu$$

Si dimostra che è consistente, $\text{Var}[\hat{\mu}] = \frac{\text{Var}[y]}{N} = \frac{\sigma^2}{N}$

$$\text{ES STIMATORE VARIANZA e media} = \text{Var}\left[\frac{1}{N} \sum_{i=1}^N y_i\right] = \frac{1}{N^2} \sum_{i=1}^N \text{Var}[y_i] = \frac{1}{N^2} \cdot N \cdot \sigma^2 = \frac{\sigma^2}{N} = \frac{\text{Var}[y]}{N}$$

$D = \{y_1, \dots, y_N\}$ con varianza σ^2 ? Il stimatore s_{N-1}^2 è corretto, $s_{N-1}^2 = \frac{\sum_{i=1}^N (y_i - \hat{\mu})^2}{N-1}$.

$$s_{N-1}^2 = E\left[\frac{1}{N-1} \sum_{i=1}^N (y_i - \hat{\mu})^2\right] = E\left[\frac{1}{N-1} \sum_{i=1}^N (y_i^2 + \hat{\mu}^2 - 2y_i\hat{\mu})\right] \xrightarrow{\text{spesso}} E\left[\frac{1}{N-1} \left(\sum_{i=1}^N y_i^2 + N\hat{\mu}^2 - 2\hat{\mu} \left(\sum_{i=1}^N y_i\right)\right)\right]$$

$$= E\left[\frac{1}{N-1} \left(\sum_{i=1}^N y_i^2 + N\hat{\mu}^2 - 2\hat{\mu} \cdot N\hat{\mu}\right)\right] = E\left[\frac{1}{N-1} \left(\sum_{i=1}^N y_i^2 - N\hat{\mu}^2\right)\right] = \cancel{\left[\frac{1}{N-1} \left(\sum_{i=1}^N y_i^2 - N\hat{\mu}^2\right)\right]}$$

$$\text{Var}[s_{N-1}^2] = E[s_{N-1}^2] - E[y]^2 = \frac{1}{N-1} \left(\sum_{i=1}^N E[y_i^2] - N \cdot E[\hat{\mu}^2] \right) = \frac{1}{N-1} \left(N \cdot E[y^2] - N \cdot E[\hat{\mu}^2] \right) = \frac{N}{N-1} \left(E[y^2] - E[\hat{\mu}^2] \right)$$

$$= \frac{N}{N-1} \left(\cancel{\text{Var}[y]} + E[y]^2 - \cancel{\text{Var}[\hat{\mu}]} - E[\hat{\mu}]^2 \right) = \frac{N}{N-1} \left(\cancel{\text{Var}[y]} + \mu^2 - \cancel{\text{Var}[\hat{\mu}]} - \mu^2 \right)$$

$$= \frac{N}{N-1} \left(\cancel{\text{Var}[y]} - \frac{\text{Var}[y]}{N} \right) = \frac{N}{N-1} \left(\frac{(N-1)}{N} \cancel{\text{Var}[y]} \right) = \text{Var}_{\text{CORRETTO}}[y] = \boxed{\sigma^2}$$

CORRETTO!

LIMITE DI CRAMER-RAO (CRAMER-RAO BOUND)

Stabilisce un limite inferiore per la varianza di un qualsiasi stimatore

non povero essere più preciso di un altro vettore

↳ questo perché i dati sono offetti da un errore di misura che non posso rimuovere con le mie stime

Nel caso di stimatori connotati abbiamo che: $\text{Var}[\hat{\theta}] \geq m^{-1}$ m: ^{quantità di Fisher} informazione di

↳ se $\hat{\theta}$ è un vettore: $\text{Var}[\hat{\theta}] - M^{-1} \geq 0$ ^{di d} ^{di d} ^{↳ la differenza è semi-definita positiva}

- Un stimatore ~~è~~ si dice efficiente se $\text{Var}[\hat{\theta}] = m^{-1}$
- Un stimatore si dice asintoticamente efficiente se $\lim_{N \rightarrow \infty} \text{Var}[\hat{\theta}] = m^{-1}$ ^{è infinito grande} ^{è grande se la varianza è piccola}

STIMA DI PESOLO LINEARE:

STIMA A MINIMI QUADRATI (LEAST SQUARES)

Obbiamo trovare descritto i dati $y(1), \dots, y(N)$ in termini della loro media e varianza, dando degli stimatori per queste due quantità $\hat{\mu} = \frac{1}{N} \sum_{i=1}^N y(i)$, $S_{yy}^2 = \frac{1}{N-1} \sum_{i=1}^N (y(i) - \hat{\mu})^2$

Supponiamo ora di voler descrivere i dati D tramite una relazione lineare:

i dati obbiamo questa struttura, la impostare ^{suppongo che} $y(i) = \beta_0 + \beta_1 x_1(i) + \beta_2 x_2(i) + \dots + \beta_d x_d(i) + e(i)$ ^{errore}

dove x_1, x_2, \dots, x_d sono variabili di cui si dispongono misure. Questo modello prende il nome di regressione lineare

regressori o feature

Es

$$y = \text{Peso } [kg]$$

$x_1 = \text{altezza } [m] \Rightarrow$ variabile numerica (c'è ordinamento) e quantificare le distanze

$x_2 = \text{sessu } [M/F] \Rightarrow$ variabile categiriale (non c'è ordinamento)

Vogliamo esprimere il peso in funzione delle varie x_1, x_2, \dots, x_d

$$x_d = \text{luogo di nascita } [\equiv]$$

N persone misurate

Definiamo: vettori:

$$\begin{aligned} \vartheta &= \begin{bmatrix} \vartheta_0 \\ \vartheta_1 \\ \vdots \\ \vartheta_d \end{bmatrix} & \varphi(i) &= \begin{bmatrix} 1 \\ x_1(i) \\ x_2(i) \\ \vdots \\ x_d(i) \end{bmatrix} & \Rightarrow & \boxed{\begin{aligned} y(i) &= \varphi(i)^T \cdot \vartheta + e(i), i=1 \dots N \\ & \vdots \quad \vdots \quad \vdots \end{aligned}} \end{aligned}$$

Approfondimento

Possiamo misurare le stesse entità con diversi tipi di variabili. Ad esempio, i partecipanti ad una maratona possono essere rappresentati come:

- 1) Tempo impiegato per raffinare il traguardo \Rightarrow VARIABILE METRICA (NUMERICA)
- 2) Posizione di arrivo (primo, secondo, terzo, ...) \Rightarrow VARIABILE ORDINALE
- 3) Nome del team di appartenenza \Rightarrow VARIABILE NOMINALE (CATEGORICA)

VARIABILE METRICA (METRIC VARIABLE)

- Descrivono una quantità (es. tempo, altezza, temperatura, peso)
- È definito un ordinamento (si dice quale valore è "più grande" di un altro)
- È definita una distanza (si "quanto" un numero è più grande di un altro)

Un'altra specie di variabile metrica è una VARIABILE CONTIGUA (COUNT VARIABLE)

L'esprime il numero di eventi accorsi (nel tempo o nello spazio)

L'Es. numero di macchie transitate al cosello in un'ora

VARIABILE ORDINALE

- Descrivono oggetti sui quali la scorsa impone un ordine
- Non ha senso chiedersi "di quanto" un valore è più grande di un altro
- Es. Posizionamento in una corsa (primo, secondo, ...)

L'ordine di confidenza su un oggetto ($0 = \text{non confidante}$; $1 = \text{poor confidante}$, $2 = \text{moderately confidante}$)

VARIABILE CATEGORICA

- Descrivono delle categorie di appartenenza
- Non ha senso impostare un ordinamento
- " " " chiedersi di quanto un valore è più grande di un altro
- Es.

L'sessu: (M/F)

L'affiliazione politica (REPUBBLICANA/DEMOCRATICA)

L'colore degli occhi (BLU, VERDE, MARRONE)

VEDI RETRO

8.5

PERCHÉ A INTERESSA IL TIPO DI VARIABILE?

Ci interessa perché, a seconda del tipo di variabile, utilizziamo un modello appropriato per quel tipo.

Es.

- 1) VARIABILE METRICA: $y \sim N(\mu, \sigma^2)$
- 2) VARIABILE COUNT: $y \sim \text{Poisson}(\lambda)$ λ : # eventi nell'unità di tempo
- 3) VARIABILE CATEGORICA
DICOTOMICA: $y \sim \text{Bernoulli}(\pi)$ π : probabilità che $y=1$
 $(y=0, y=1)$

Esistono modelli più complessi per dati ordinati.

Consigli

Uno degli step iniziali per sviluppare un modello dei dati è DETERMINARE LA TIPLOGIA DELLE VARIABILI IN gioco

Il metodo dei minimi quadrati minimizza la somma quadratica tra i dati ed il modello

$$\text{GRADIENTE} \quad J(\theta) = \frac{1}{N} \sum_{i=1}^N (y_{(i)} - \varphi_{(i)}^\top \theta)^2 = \sum_{i=1}^N e_i^2$$

Vogliamo il $\hat{\theta}$ che minimizza queste quantità

$$\nabla J(\theta) = \frac{dJ(\theta)}{d\theta} = 0 \Rightarrow \frac{1}{N} \sum_{i=1}^N \varphi_{(i)} \cdot (y_{(i)} - \varphi_{(i)}^\top \theta) = 0 \Rightarrow \sum_{i=1}^N \varphi_{(i)} y_{(i)} - \sum_{i=1}^N \varphi_{(i)} \varphi_{(i)}^\top \theta = 0$$

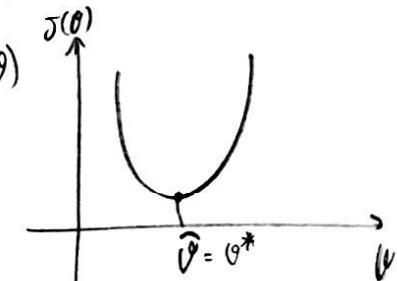
$$\Rightarrow \left[\sum_{i=1}^N \varphi_{(i)} \varphi_{(i)}^\top \right] \theta = \sum_{i=1}^N \varphi_{(i)} y_{(i)} \Rightarrow \boxed{\hat{\theta} = \left[\sum_{i=1}^N \varphi_{(i)} \varphi_{(i)}^\top \right]^{-1} \left[\sum_{i=1}^N \varphi_{(i)} y_{(i)} \right]}$$

Osservazioni:

- Se $\det \left[\sum_{i=1}^N \varphi_{(i)} \varphi_{(i)}^\top \right] \neq 0$ la soluzione è unica!

- - - - - $= 0$, \exists INFINITE SOLUTION

- Dato che il modello è lineare e le funzioni di costi quadratica, essa ha una forma quadratica di $J(\theta)$.
Si dimostra che $\hat{\theta}$ è MINIMO GLOBALE di $J(\theta)$



MINIMI QUADRATI - NOTAZIONE MATRICIALE

$$X = \begin{bmatrix} 1 & x_1(1) & x_2(1) & \cdots & x_d(1) \\ 1 & x_1(2) & x_2(2) & \cdots & x_d(2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1(N) & x_2(N) & \cdots & x_d(N) \end{bmatrix}_{N \times d}$$

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix}_{d+1}$$

$$Y = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix}_{N \times 1}$$

$$E = \begin{bmatrix} e(1) \\ e(2) \\ \vdots \\ e(N) \end{bmatrix}_{N \times 1}$$

↳ ogni colonna è un repatente / features

$$\|X\| = \sqrt{x_1^2 + x_2^2 + \dots + x_d^2}$$

$$X = \begin{bmatrix} \varphi_{(1)}^\top \\ \varphi_{(2)}^\top \\ \vdots \\ \varphi_{(N)}^\top \end{bmatrix}_{d \times N}$$

$$Y = X\theta + E \Rightarrow J(\theta) = \frac{1}{N} \|Y - X\theta\|^2 = \frac{1}{N} (Y - X\theta)^\top (Y - X\theta) =$$

$$\text{scrivendo} \quad \frac{1}{N} (Y^\top Y - Y^\top X\theta) - (\theta^\top X^\top Y) + (\theta^\top X^\top X\theta)$$

$$\nabla_\theta (J(\theta)) = (A + A^\top)\theta \quad \frac{1}{N} (Y^\top Y - 2\theta^\top X^\top Y + \theta^\top X^\top X\theta) \quad \nabla_\theta (J(\theta)) = b \quad \frac{(X\theta)^\top}{d+1} = \theta^\top X^\top$$

$$\nabla J(\theta) = 0$$

$$\Rightarrow \frac{1}{N} \left(-2X^\top Y + 2X^\top X\theta \right) = 0 \Rightarrow \boxed{\hat{\theta} = (X^\top X)^{-1} X^\top Y}$$

$$\nabla_\theta (-) = (X^\top X + (X^\top X)^\top) / N = 2X^\top X\theta$$

Come si computa lo stimatore a tenere conto degli errori (modello lineare) nel caso in cui il sistema vero sia effettivamente lineare?

$$y(i) = \varphi(i)^T \theta^* + v(i) \quad (\theta^*: \text{valore vero dei parametri})$$

- Supponendo $v(i)$ un rumore casuale di valori nello $E[v(i)] = 0$

$$\downarrow \quad E[\hat{\theta}] = \theta^* \quad \text{CORRETTO}$$

- Supponendo inoltre che i rumori siano indipendenti: $E[v(i)v(j)] = 0 \quad \forall i \neq j$
e variano $\sigma^2 \rightarrow \text{Var}[v(i)] = \sigma^2$

$$\downarrow \quad \text{Var}[\hat{\theta}] = \sigma^2 \cdot \left[\sum_{i=1}^N \varphi(i) \varphi(i)^T \right]^{-1} \quad \text{CONSISTENTE}$$

Es θ^* scalare

$$\begin{aligned} S: \quad & y(i) = x(i)\theta^* + v(i) \\ T: \quad & y(i) = x(i)\theta + e(i) \Rightarrow J(\theta) = \frac{1}{N} \sum_{i=1}^N (y(i) - x(i)\theta)^2 \\ & \frac{dJ(\theta)}{d\theta} = 0 \rightarrow -\frac{2}{N} \sum_{i=1}^N (y(i) - x(i)\theta) x(i) = 0 \\ & \Rightarrow \sum_{i=1}^N (y(i)x(i) - x(i)^2\theta) = 0 \Rightarrow \sum_{i=1}^N y(i)x(i) - \sum_{i=1}^N x(i)^2\theta = 0 \\ & \Rightarrow \hat{\theta} = \frac{\sum_{i=1}^N y(i)x(i)}{\sum_{i=1}^N x(i)^2} \\ E[\hat{\theta}] = & E \left[\frac{\sum_{i=1}^N y(i)x(i)}{\sum_{i=1}^N x(i)^2} \right] = \frac{\sum_{i=1}^N E[y(i)x(i)]}{\sum_{i=1}^N x(i)^2} = \frac{\sum_{i=1}^N E[x(i)\theta^* + v(i)x(i)]}{\sum_{i=1}^N x(i)^2} \\ = & \frac{\sum_{i=1}^N (x(i)\theta^* + 0)x(i)}{\sum_{i=1}^N x(i)^2} = \boxed{\theta^*} \end{aligned}$$

$$\text{Var}[\hat{\theta}] = \frac{\sigma^2}{\sum_{i=1}^N x(i)^2}$$

ESTIMA A MASSIMA VEROSSIGLIANZA

Ottieni presenti fior od ora diversi tipi di stimatori,

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N y(i) \quad \text{media campionaria} \Rightarrow \hat{\mu} = \mu \in \mathbb{R}$$

$$LS^2 = \frac{1}{N-1} \sum (y(i) - \hat{\mu})^2 \quad \text{varianza campionaria} \Rightarrow \hat{\sigma}^2 = \sigma^2 \in \mathbb{R}$$

$E(i) \sim d(0, \lambda^2)$

$$\begin{aligned} & \text{L'STIMA MIN. QUADRATI } y(i) = \theta_0 + \theta_1 x_1(i) + \theta_2 x_2(i) + \dots + \theta_d x_{d-1}(i) + \epsilon(i) \\ & \Rightarrow \hat{\theta} = [\hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_d]^T \in \mathbb{R}^{d+1} \end{aligned}$$

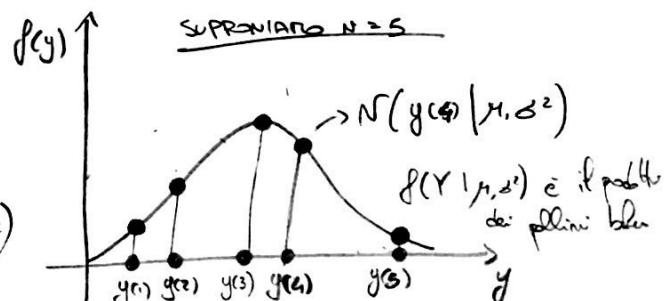
Ottieni presenti stimatori PARAMETRICI, aendo rappresentati i dati tramite un modello parametrico (es. modello lineare)

L'Non ottieni mai fatti assunzioni sulle pdf dei dati $D = \{y(1), \dots, y(N)\}$

Il metodo della MASSIMA VEROSSIGLIANZA è una procedura di stima che, dato un modello probabilistico, stima i suoi parametri in modo che siano il più possibile consistenti con i dati osservati

Supponiamo di avere $\mathbf{Y} = [y(1), \dots, y(N)]^T$: N osservazioni della variabile scobie y

L' $y(i) \sim N(\mu, \sigma^2)$ i.i.d.



Se pdf del vettore dati è:

$$f(y(1), y(2), \dots, y(N) | \mu, \sigma^2) = \prod_{i=1}^N N(y(i) | \mu, \sigma^2)$$

- È la prob. che si realisi il vettore di dati osservato

L'siccome $y(i) \sim d$, la prob. di ottenere $y(1)$ AND $y(2)$ AND ... è il prodotto delle varie pdf delle singole voci

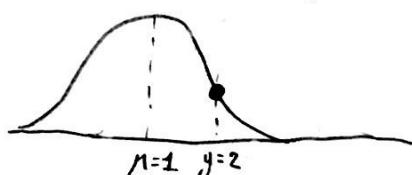
L' se im funz. della y, è una pdf N-dimensionale

però se il valore reale delle $y(i)$ \Rightarrow se conoscere anche μ e σ^2 , poss. calcolare il valore osservato dalla pdf

- Quand. queste funz. è vista im funz. di μ e σ^2 (conoscendo le Y), allora prob. vero prende il nome di VEROSIGLIANZA (LIKELIHOOD)

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{(y-\mu)}{\sigma}\right)^2} = f(y|\mu, \sigma^2)$$

NUMERO NOTO
FUNZIONE DI
y

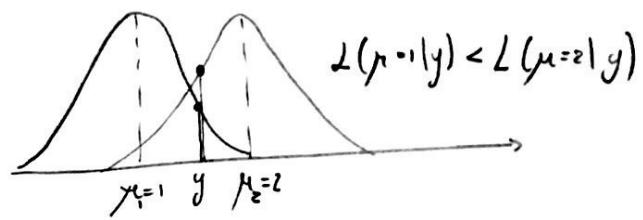


$$\Rightarrow L(\mu, \sigma^2 | y) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\left(\frac{(y-\mu)}{\sigma}\right)^2}$$

(11)

$$L(\mu, \sigma^2 | y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \left(\frac{y-\mu}{\sigma} \right)^2}$$

supponendo σ^2 noto $\rightarrow L(\mu | y)$



Lo stimatore massima verosimiglianza è quel valore del parametro θ che massimizza $L(\theta | y)$

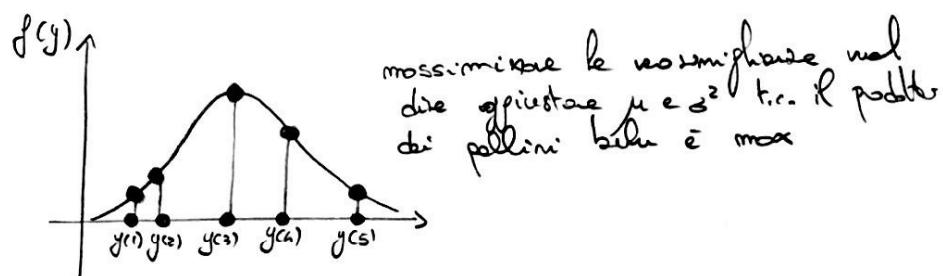
L'ad esempio, $\theta = \mu$ (σ^2 noto) $\Rightarrow \mu=2$ è più verosimile di $\mu=1$ perché $f(y|\mu=1) < f(y|\mu=2)$

In questo caso, lo stima più verosimile sarà $\mu=y$



Quindi, nel caso di più osservazioni ^{iid} di y , $Y = [y^{(1)}, \dots, y^{(N)}]^T$, dare massimizzare $f(y^{(1)}, \dots, y^{(N)} | \mu, \sigma^2) = L(\mu, \sigma^2 | Y) = \prod_{i=1}^N N(y^{(i)} | \mu, \sigma^2)$

SUPPONIAMO $N=5$



$$\hat{\theta}_m = \begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix} = \arg \max_{\theta} L(\theta | Y) = \arg \max_{\theta} \prod_{i=1}^N N(y^{(i)} | \theta)$$

In genere posso attribuire ai dati qualsiasi pdf $f(Y|\theta)$



$$\boxed{\hat{\theta}_m = \arg \max_{\theta} L(\theta | Y) = \arg \max_{\theta} \prod_{i=1}^N f(y^{(i)} | \theta)}$$

Spesso, anche massimizzando $L(\theta | Y)$, si massimizza il suo logaritmo naturale

L'atto de il logaritmo ^{di L(theta)} è una funzione monotona crescente, ha lo stesso massimo di $L(\theta)$

L'è efficiente del punto di vista implementativo, perché evita l'indebolimento del prodotto di piccole probabilità (addizionandone le somme delle log-probabilità)

$$\boxed{\hat{\theta}_m = \arg \max_{\theta} \ln [L(\theta | Y)]}$$

Soltanente queste stime possono essere effettuate con metodi numerici iterativi



Si dà così si può fare analiticamente (Gaussiano, ...)

ENTRATA DI PARAMETRI DI UNA POPOLAZIONE:

Esempio: Supponiamo che siano dati i punti delle popolazione delle $y_{i=1, \dots, N}$

Siamo $y(i) \sim N(\mu, \sigma^2)$ i.i.d. \Rightarrow trovare la stima max verosimiglianza di $\theta = [\mu, \sigma^2]$

$$f(y(i)|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \left(\frac{y(i)-\mu}{\sigma} \right)^2} \Rightarrow \text{i.i.d.} \Rightarrow L(\underbrace{\mu, \sigma^2}_{\theta} | y(1), \dots, y(N)) = \prod_{i=1}^N f(y(i)|\mu, \sigma^2)$$

$$L(\theta|Y) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \left(\frac{y(i)-\mu}{\sigma} \right)^2} \xrightarrow{\ln} \ln[L(\theta|Y)] = \sum_{i=1}^N \ln \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \left(\frac{y(i)-\mu}{\sigma} \right)^2} \right]$$

$$= \sum_{i=1}^N \left(\ln \left[\frac{1}{\sqrt{2\pi\sigma^2}} \right] + \ln \left[e^{-\frac{1}{2} \left(\frac{y(i)-\mu}{\sigma} \right)^2} \right] \right) = \sum_{i=1}^N \ln \left[\frac{1}{\sqrt{2\pi\sigma^2}} \right] + \sum_{i=1}^N \ln \left[e^{-\frac{1}{2} \left(\frac{y(i)-\mu}{\sigma} \right)^2} \right]$$

$$= N \cdot \ln \left[\frac{1}{\sqrt{2\pi\sigma^2}} \right] + \sum_{i=1}^N -\frac{1}{2} \left(\frac{y(i)-\mu}{\sigma} \right)^2 \ln e = N \cdot \ln \left[2\pi\sigma^2 \right]^{\frac{1}{2}} - \frac{1}{2} \sum_{i=1}^N \left(\frac{y(i)-\mu}{\sigma} \right)^2 =$$

$$-\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \sum_{i=1}^N \left(\frac{y(i)-\mu}{\sigma} \right)^2 = \boxed{-\frac{N}{2} \ln 2\pi - \frac{N}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^N (y(i)-\mu)^2}$$

$$\hat{\theta} = \text{argmax}_{\theta} L(\theta|Y) \quad \Rightarrow \quad \begin{cases} \frac{\partial L(\mu, \sigma^2|Y)}{\partial \mu} = 0 \\ \frac{\partial L(\mu, \sigma^2|Y)}{\partial \sigma^2} = 0 \end{cases} \Rightarrow \begin{cases} \frac{1}{2\sigma^2} \sum_{i=1}^N (y(i)-\mu) = 0 \\ -\frac{N}{2} \cdot \frac{1}{\sigma^2} - \frac{1}{2} \sum_{i=1}^N (y(i)-\mu)^2 \cdot \left(-\frac{1}{\sigma^4} \right) = 0 \end{cases}$$

$$\frac{1}{x} \Rightarrow \frac{dx}{dx} = -x^{-2} = -(\sigma^2)^{-2}$$

$$\begin{cases} \frac{1}{2\sigma^2} \sum_{i=1}^N (y(i)-\mu) = 0 \Rightarrow \sum_{i=1}^N (y(i)-\mu) = 0 \Rightarrow \sum_{i=1}^N y(i) - \sum_{i=1}^N \mu = 0 \Rightarrow \sum_{i=1}^N y(i) - N\mu = 0 \\ -\frac{N}{2} \cdot \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^N (y(i)-\mu)^2 = 0 \end{cases}$$

sostituire $\hat{\mu} \rightarrow$

$$\frac{1}{2\sigma^2} \sum_{i=1}^N (y(i)-\hat{\mu})^2 = \frac{N}{2} \cdot \frac{1}{\sigma^2}$$

CORRETTO! $\Rightarrow \hat{\mu} = \frac{1}{N} \sum y(i)$

MEDIA
CAMPIONARIA

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{N} \sum (y(i)-\hat{\mu})^2$$

VARIANZA
CAMPIONARIA

DISTORTO!!

Se l'immagine è mossa non si può più essere distinta!

↓ Il grande per, esser pote di buone proprietà

PROPRIETÀ STIMA MASSIMA VEROSIMILANZA

1) Assintoticamente corretta: $\lim_{N \rightarrow +\infty} E[\hat{\theta}_n] = \theta^*$ Es. STIMATORE VARIANZA $\hat{\sigma}_{\text{re}}^2 = \frac{1}{N} \sum (y_i - \bar{y})^2$

2) Consistente: più N grande, + stime precise quando $N \rightarrow \infty$ dunque per $N = \infty$ per $N \rightarrow \infty$ non cambia

3) Asintoticamente efficiente: $\lim_{N \rightarrow +\infty} \text{Var}[\hat{\theta}_n] = H^{-1}$ H : matrice di informazione di Fisher

4) Quantitativamente normale: $\hat{\theta}_n \sim N(\theta^*, \frac{1}{H})$ se $N \rightarrow +\infty$

L' $\hat{\theta}_n$ è centrato sul valore vero e ha varianza più ~~piccola~~ inverso dell'informazione di Fisher

Esempio 1 - con numeri

Sia $y^{(i)} \sim N(\mu, \sigma^2 = 1)$, $i = 1, 2$, i.d. Calcolare la stima di μ nel caso in cui:

$$y^{(1)} = 4 \quad y^{(2)} = 6$$

$$f(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2} \left(\frac{(y-\mu)^2}{\sigma^2} \right)} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (y-\mu)^2}$$

La densità im corrispondente delle due osservazioni è:

~~$$f(y^{(1)}, y^{(2)} | \mu) = f(y^{(1)} | \mu, \sigma^2 = 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (4-\mu)^2}$$~~
~~$$f(y^{(2)} | \mu) = f(y^{(2)} | \mu, \sigma^2 = 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (6-\mu)^2}$$~~

La pdf condizionata (i.e.) è:

$$f(y^{(1)}=4, y^{(2)}=6 | \mu, \sigma^2 = 1) = \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (4-\mu)^2} \right) \cdot \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (6-\mu)^2} \right)$$

↓
È FUNZIONE SOLO DI μ !

Interpretando $\mathcal{L}(\mu | y_{(1)}=4, y_{(2)}=6 | \mu, \sigma^2=1)$ come funzione di μ , otteniamo
la VEROSIMILITUDINE

$$\mathcal{L}(\mu | \underbrace{y_{(1)}=4, y_{(2)}=6}_{\Theta}, Y = [y_{(1)}, y_{(2)}]) = \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(4-\mu)^2} \right) \cdot \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(6-\mu)^2} \right)$$

$$\hat{\mu} = \underset{\mu}{\operatorname{arg\ max}} \mathcal{L}(\mu | y_{(1)}=4, y_{(2)}=6)$$

Calcolare la log-likelihood:

$$\begin{aligned} \ln[\mathcal{L}] &= \ln \left[\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(4-\mu)^2} \right) \cdot \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(6-\mu)^2} \right) \right] \\ &= \ln \left[\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(4-\mu)^2} \right] + \ln \left[\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(6-\mu)^2} \right] = \\ &= \ln \frac{1}{\sqrt{2\pi}} + \ln \left[e^{-\frac{1}{2}(4-\mu)^2} \right] + \ln \frac{1}{\sqrt{2\pi}} + \ln \left[e^{-\frac{1}{2}(6-\mu)^2} \right] \\ &= 2 \cdot \ln \frac{1}{\sqrt{2\pi}} - \frac{1}{2}(4-\mu)^2 \ln e - \frac{1}{2}(6-\mu)^2 \ln e \\ &= \underline{2 \ln \frac{1}{\sqrt{2\pi}} - \frac{1}{2}(4-\mu)^2 - \frac{1}{2}(6-\mu)^2} \end{aligned}$$

Trovarne il massimo:

$$\begin{aligned} \frac{\partial \ln[\mathcal{L}]}{\partial \mu} = 0 &\Rightarrow \frac{2}{2}(4-\mu) + \frac{2}{2}(6-\mu) = 0 \Rightarrow \frac{4+6}{2} = 2\mu \\ &\Rightarrow \boxed{\hat{\mu} = \frac{4+6}{2} = 5} \end{aligned}$$

MEDIA
CAMPIONARIA!

E_s

~~•~~ Colabă b. dimostrare moștenirea verosimilității sol. cor. în cui N obi îl
parcurg de ac. distribuția de Bernoulli ca probabilitate

$$P(y| \pi) = \pi^y \cdot (1-\pi)^{1-y} \quad y=0,1 \quad \text{hence we make the ad esce test}$$

$$\rightarrow L(\pi | Y) = \prod_{i=1}^N \pi^{y(i)} \cdot (1-\pi)^{1-y(i)} = \pi^{\sum_{i=1}^N y(i)} \cdot (1-\pi)^{\sum_{i=1}^N (1-y(i))}$$

Colours to be by - season, please

$$\begin{aligned}
 \ln L &= \ln \left[\pi \sum_{i=1}^N y_{(i)} \cdot (1-\pi) \sum_{i=1}^N (1-y_{(i)}) \right] = \ln \pi + \sum_{i=1}^N y_{(i)} + \ln (1-\pi) \\
 &= \left(\sum_{i=1}^N y_{(i)} \right) \ln \pi + \left(\sum_{i=1}^N (1-y_{(i)}) \right) \ln (1-\pi) \quad \left\{ \begin{array}{l} \gamma \ln \pi + (N-\gamma) \ln (1-\pi) \\ \text{with } \gamma \text{ masses} \end{array} \right. \\
 &\quad \sum_{i=1}^N 1 - \sum_{i=1}^N y_{(i)} = N - \gamma
 \end{aligned}$$

Now the measures

$$\frac{\partial \ln[L]}{\partial \pi} = 0 \Rightarrow \frac{\gamma}{\pi} - \frac{(N-\gamma)}{1-\pi} = 0 \Rightarrow \frac{(\pi-\gamma) - \pi(N-\gamma)}{\pi(1-\pi)} = 0$$

$$\Rightarrow \gamma - \pi f - \pi N + g \kappa = 0 \Rightarrow \boxed{\bar{c} = \frac{\gamma}{N} = \frac{1}{N} \sum_{i=1}^n y(i)}$$

MEDIA
CAMPIONADA

Osservazioni

e le di successi

La distribuzione di Bernoulli $p(y|\pi)$ è una distribuzione discreta. Infatti π è fisso ad un valore ed il dbr y è la variabile che assume solo 2 valori discetti: 0 e 1

La Likelihood $L(\pi | Y) = \pi^y \cdot (1-\pi)^{n-y}$ è una funzione continua del parametro π che è continua tra $[0, 1]$. Non è una distribuzione perché non integra a 1

Osservazione

È importante notare che m massimizzazione le log-likelihood equivalenti e m minimizzazione la meno log-likelihood

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} \ln [L(\theta | Y)] \\ = \underset{\theta}{\operatorname{argmin}} -\ln [L(\theta | Y)]$$

In questo modo, abbiamo un problema di minimizzazione come con la regressione lineare, dove minimizziamo (tramite il metodo dei minimi quadrati):

$$J(\theta) = \frac{1}{N} \sum_{i=1}^N (y(i) - \varphi(i)^T \theta)^2$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} J(\theta)$$

SINTA

MASSIMAZIONE VEROSIMILANZA DI MODELLI LINEARI

Come nel caso in cui non vi erano osservazioni sulla pdf dei dati, dobbiamo cercare degli stimatori $\hat{\theta}$ per descrivere i dati con dei parametri delle loro popolazioni.

→ possiamo usare il metodo ML anche nel caso in cui vogliamo descrivere i dati attraverso un modello lineare.

$$y(i) = \theta_0 + \theta_1 x_1(i) + \theta_2 x_2(i) + \dots + \theta_d x_d(i) + e(i)$$

$$= \varphi(i)^T \theta + e(i) \quad e(i) \sim N(0, \lambda^2) \text{ (i.i.d.)}, \quad e(i) \perp \theta$$

$$\varphi(i) = \begin{bmatrix} 1 & x_1(i) & x_2(i) & \dots & x_d(i) \end{bmatrix}^T$$

$$\boxed{y(i) \sim N(\varphi(i)^T \theta, \lambda^2)}$$

La modellazione è espressa come funzione lineare dei regressori!

La probabilità di osservare i dati misurati è data dalla probabilità condizionata delle $y^{(i)}$:

$$f(\underbrace{y^{(1)}, \dots, y^{(N)}}_Y | X, \theta, \lambda^2) = \prod_{i=1}^N f(y^{(i)} | \varphi^{(i)}, \theta, \lambda^2) =$$

$$X = \begin{bmatrix} \varphi^{(1)^T} \\ \varphi^{(2)^T} \\ \vdots \\ \varphi^{(N)^T} \end{bmatrix}_{N \times d} = \prod_{i=1}^N N(\varphi^{(i)^T} \theta, \lambda^2) =$$

$$= \prod_{i=1}^N \frac{1}{\sqrt{2\pi\lambda^2}} e^{-\frac{1}{2} \left(\frac{y^{(i)} - \varphi^{(i)^T} \theta}{\lambda} \right)^2} = L(\theta, \lambda^2 | Y, X)$$

Supponiamo λ^2 noto per semplicità.

L'è verosimiglianza è funzione del sol vettore dei coefficienti $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_{d-1} \end{bmatrix}$

Calcolo la log-verosimiglianza

$$\ln[L(\theta | X, Y)] = \ln \left[\prod_{i=1}^N \frac{1}{\sqrt{2\pi\lambda^2}} e^{-\frac{1}{2} \left(\frac{y^{(i)} - \varphi^{(i)^T} \theta}{\lambda} \right)^2} \right]$$

$$= \sum_{i=1}^N \ln \left[\frac{1}{\sqrt{2\pi\lambda^2}} \cdot e^{-\frac{1}{2} \left(\frac{y^{(i)} - \varphi^{(i)^T} \theta}{\lambda} \right)^2} \right] = \sum_{i=1}^N \left(\ln \frac{1}{\sqrt{2\pi\lambda^2}} + \ln \left[e^{-\frac{1}{2} \left(\frac{y^{(i)} - \varphi^{(i)^T} \theta}{\lambda} \right)^2} \right] \right)$$

$$= \sum_{i=1}^N \ln \frac{1}{\sqrt{2\pi\lambda^2}} + \sum_{i=1}^N \ln \left[e^{-\frac{1}{2} \left(\frac{y^{(i)} - \varphi^{(i)^T} \theta}{\lambda} \right)^2} \right] = N \cdot \ln(2\pi\lambda^2)^{-\frac{1}{2}} + \sum_{i=1}^N -\frac{1}{2} \left(\frac{y^{(i)} - \varphi^{(i)^T} \theta}{\lambda} \right)^2$$

$$= -\frac{1}{2} N \cdot \ln 2\pi\lambda^2 - \frac{1}{2\lambda^2} \sum_{i=1}^N (y^{(i)} - \varphi^{(i)^T} \theta)^2 = \boxed{-\frac{N}{2} \ln 2\pi\lambda^2 - \frac{N}{2} \ln \lambda^2 - \frac{1}{2\lambda^2} \sum_{i=1}^N (y^{(i)} - \varphi^{(i)^T} \theta)^2}$$

Calcolare il massimo di $\ln[L(\theta | X, Y)]$ è uguale a calcolare il minimo di $-\ln[L(\theta | X, Y)]$

$$-\ln[L(\theta | X, Y)] = +\frac{N}{2} \ln 2\pi + \frac{N}{2} \ln \lambda^2 + \frac{1}{2\lambda^2} \sum_{i=1}^N (y^{(i)} - \varphi^{(i)^T} \theta)^2$$

NON DIPENDONO DA θ

$$\Rightarrow \boxed{\hat{\theta}_{ML} = \underset{\theta}{\operatorname{arg\,min}} \frac{1}{2\lambda^2} \sum_{i=1}^N (y^{(i)} - \varphi^{(i)^T} \theta)^2}$$

Osservazione

Le stime ML così ottenuta ha lo stesso minimo (è equivalente) alle stime ottenute con i minimi quadrati (in assenza di osservazioni problematiche).

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{arg\min}} \frac{1}{2N} \sum_{i=1}^N (y_{(i)} - \phi_{(i)}^T \theta)^2$$

$$\hat{\theta}_{LS} = \underset{\theta}{\operatorname{arg\min}} \frac{1}{N} \sum_{i=1}^N (y_{(i)} - \phi_{(i)}^T \theta)^2$$

\Rightarrow scelta per una costante (che esse sia $\frac{1}{2N}$ o $\frac{1}{N}$) non cambia il minimo delle funz. di cost.



Le stime ML del modello $y_{(i)} = \phi_{(i)}^T \theta + \text{e}_i$, dove $e_i \sim N(0, \sigma^2)$ iid, è equivalente alle stime LS.

↳ queste osservazioni di modello sono origine al modello di

REGRESSIONE LINEARE

Osservazione

Combinando le ipotesi sulla distribuzione del rumore, si ottengono le funzioni di costi e quindi altri algoritmi, che modellano i dati in modo diverso delle regressioni lineare.

* REGRESSIONE LOGISTICA *

Il procedimento delle regressione lineare modellizza dati metrici attraverso un modello lineare, tramite l'ausilio di regressori (features).



Un problema frequente è la modellizzazione di dati CATEGORICI DISCONTINUI, in cui y assume valori 0 o 1. \Rightarrow Esempi:

- predire se una persona in un studio demografico sia maschio o femmina in base a pose e età
- predire cosa voterà una persona fra due candidati in base al reddito
- predire se un giocatore di baseball colpirà la pallina in base al suo ruolo

In questi casi, NON HA SENSO utilizzare il modello lineare $y_{(i)} = \phi_{(i)}^T \theta + \text{e}_i$.



L'errore non somma un errore continuo (ER) ad una variabile y che può assumere soli valori come 0 e 1, e non 0,98 o 1,01.

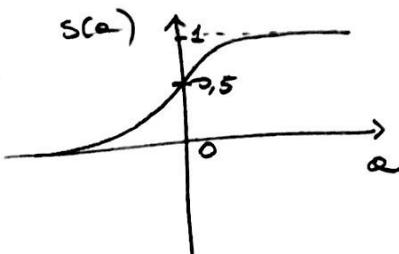
L'errore potrebbe prevedere anche valori <0 o >1! Non c'è niente che "limite" l'uscita \hat{y} tra 0 ed 1.



quello che si fa è utilizzare la **FUNZIONE LOGISTICA (SIGMOIDE)**

(19)

$$s(a) = \frac{1}{1+e^{-a}} = \frac{e^a}{1+e^a}$$



- se $a \gg 0 \Rightarrow s(a) \approx 1$
- se $a \ll 0 \Rightarrow s(a) \approx 0$

L'obiettivo di questo modello è modellare la probabilità che $y=1$ tramite un modello lineare

Probabilità \Downarrow

$$\rightarrow P(y=1 | \varphi) = s(\varphi^T \cdot \psi) = \frac{1}{1+e^{-(\varphi^T \cdot \psi)}}$$

l'output di $s(\varphi^T \cdot \psi)$ è interpretato come una probabilità

- se $\varphi^T \cdot \psi \gg 0 \Rightarrow P(y=1 | \varphi) \approx 1$
- se $\varphi^T \cdot \psi \ll 0 \Rightarrow P(y=1 | \varphi) \approx 0$

REGRESSIONE LINEARE

$$\mu = \varphi^T \cdot \psi = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_d x_d$$

$$y \sim N(\mu, \sigma^2)$$

REGRESSIONE LOGISTICA

$$\pi = s(\varphi^T \cdot \psi) = s(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_d x_d)$$

$$y \sim \text{Bernoulli}(\pi)$$

Sia la regressione lineare da la regressione logistica per parte dei cosiddetti GLM (Generalized Linear Model) in cui:

L'idea dietro di un modello lineare è usata per modellare un parametrazione di "tendenza centrale" delle distribuzioni dei dati

Il termine "fittazione" indica che il modello è un'ipotesi della distribuzione dei dati: il valore medio dei dati μ non è sempre la media! I dati y sono modellati tramite una distribuzione di probabilità in cui c'è il parametro μ

In generale: fittazione generica

$$\pi = f(\theta_0 + x_1 \theta_1 + x_2 \theta_2 + \dots + x_d \theta_d) = f(\varphi^T \cdot \psi)$$

$$y \sim \text{pdf}(\pi, [\text{altri parametri}])$$

REGRESSIONE LINEARE

$$\begin{aligned} \pi &= f(\varphi^T \cdot \psi) = \varphi^T \cdot \psi \quad (\text{fittazione}) \\ \pi &= \mu \Rightarrow \mu = \varphi^T \cdot \psi \\ y &\sim N(\mu, \sigma^2) \end{aligned}$$

REG. LOGISTICA

$$\begin{aligned} \pi &= f(\varphi^T \cdot \psi) = s(\varphi^T \cdot \psi) \quad (\text{fittazione logistica}) \\ \pi &= \pi \Rightarrow \pi = s(\varphi^T \cdot \psi) \\ y &\sim \text{Bernoulli}(\pi) \end{aligned}$$

STIMA MAXIMUM LIKELIHOOD DI UN MODELLO DI REGRESSIONE LOGISTICA

Sia dato un dataset $D = \{(\varphi(1), y_{(1)}), (\varphi(2), y_{(2)}), \dots, (\varphi(N), y_{(N)})\}$ con $\varphi \in \mathbb{R}^{d_x}$, dove $y_i \in \{0, 1\}$, $i=1, \dots, N$, iid

Stimare un modello di regressione logistica $P(y=1 | p) = \frac{1}{1+e^{-(p^T \varphi)}} = \hat{\pi}$

Interpretazione: dati come $y \sim \text{Bernoulli}(\hat{\pi})$

Calcoliamo la Verosimiglianza dei dati

$$P(y_{(i)}=1 | p_{(i)}) = \frac{1}{1+e^{-(p_{(i)}^T \varphi)}} = \hat{\pi}_{(i)}$$

$L(\hat{\pi} | Y) = \prod_{i=1}^N \hat{\pi}_{(i)}^{y_{(i)}} \cdot (1-\hat{\pi}_{(i)})^{1-y_{(i)}}$ \Rightarrow calcola la verosimiglianza \rightarrow funzione da ottimizzare da minimizzazione

$$Y = \begin{bmatrix} y_{(1)} \\ \vdots \\ y_{(N)} \end{bmatrix}$$

dipende dai parametri θ !!

$$L(\hat{\pi} | Y) = L(\theta | Y)$$

i veri parametri sono questi

$$\begin{aligned} -\ln[L(\hat{\pi} | Y)] &= -\ln \left[\prod_{i=1}^N \hat{\pi}_{(i)}^{y_{(i)}} (1-\hat{\pi}_{(i)})^{1-y_{(i)}} \right] = \\ &= -\sum_{i=1}^N \ln \left[\hat{\pi}_{(i)}^{y_{(i)}} (1-\hat{\pi}_{(i)})^{1-y_{(i)}} \right] = -\sum_{i=1}^N \left(\ln \hat{\pi}_{(i)}^{y_{(i)}} + \ln [1-\hat{\pi}_{(i)}]^{1-y_{(i)}} \right) = \\ &= \boxed{-\sum_{i=1}^N \left(y_{(i)} \ln \hat{\pi}_{(i)} + (1-y_{(i)}) \ln [1-\hat{\pi}_{(i)}] \right)} = J(\theta) \end{aligned}$$

Interpretazione delle funzioni di costo

Supponiamo di avere un solo dato $D = \{(\varphi, y)\}$:

$$J(\theta) = \begin{cases} -\ln \hat{\pi} & \text{se } y=1 \\ -\ln [1-\hat{\pi}] & \text{se } y=0 \end{cases}$$

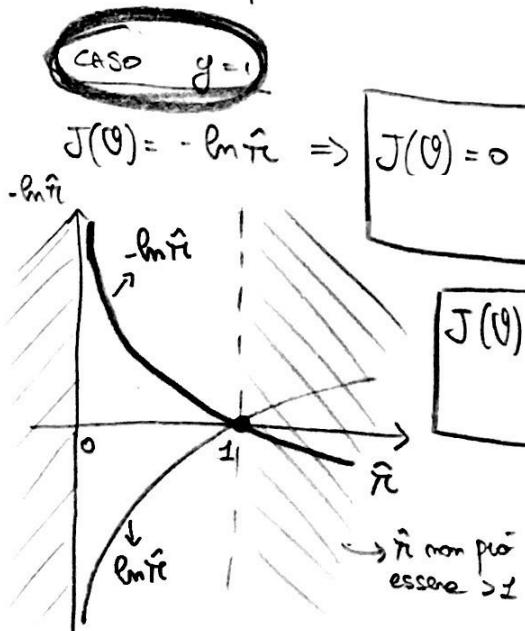
CASO $y=1$

Costo \Rightarrow se predice giusto

$$J(\theta) = -\ln \hat{\pi} \Rightarrow \boxed{J(\theta) = 0 \text{ SE } y=1 \text{ e } \hat{\pi} = 1}$$

Cottura l'interpretazione che se $y=1$, ma si predice una bassa probabilità che $y=1$, ovvero predice $\hat{\pi} \ll 1$ ($P(y=1 | p) \ll 1$) allora commette un grande sbaglio e $J(\theta) \Rightarrow +\infty$ (penalizza molto)

ci vogliono minimizzare questo sbaglio!

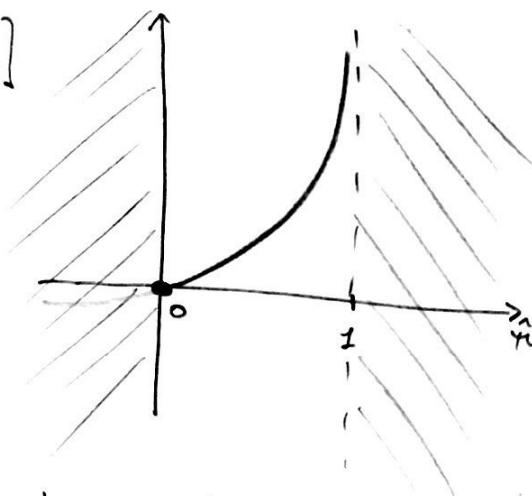


(21)

CASO $y=0$

$$-\ln[1-\hat{y}]$$

$$J(\theta) = -\ln[1-\hat{y}]$$



$$\boxed{J(\theta) = 0 \text{ SE } y=0 \\ \frac{\partial}{\partial \theta} J(\theta) = 0}$$

$$\boxed{J(\theta) = +\infty \text{ SE } y=0 \\ \frac{\partial}{\partial \theta} J(\theta) = 0 \\ \hat{y}=1}$$

Se $y=0$ ma si predice con alta probabilità che $y=1$, ovvero $\hat{y} \gg 0$ ($P(y=1|\theta) \gg 0$) allora sbaglihi molto e $J(\theta) \rightarrow +\infty$

CALCOLO DEL MINIMO

Calcoliamo il gradiente di $J(\theta)$ rispetto al vettore di parametri $\theta \in \mathbb{R}^d$

Per prima cosa, calcoliamo la derivata di $s(a) = \frac{1}{1+e^{-a}}$

$$\begin{aligned} \frac{ds(a)}{a} &= \frac{d}{da} \left[\frac{1}{1+e^{-a}} \right] = \frac{d}{da} \left[(1+e^{-a})^{-1} \right] = -(1+e^{-a})^{-2} \cdot (e^{-a})(-1) = -(1+e^{-a})^{-2}(-e^{-a}) \\ &= \frac{-e^{-a}}{(1+e^{-a})^2} = \frac{e^{-a}}{(1+e^{-a})^2} = \frac{1}{(1+e^{-a})} \cdot \frac{e^{-a}}{(1+e^{-a})} = \frac{1}{(1+e^{-a})} \cdot \frac{(1+e^{-a})^{-1}}{1+e^{-a}} \\ &= \underbrace{\frac{1}{1+e^{-a}}}_{s(a)} \cdot \left(\underbrace{\frac{1+e^{-a}}{1+e^{-a}}}_{1} - \underbrace{\frac{1}{1+e^{-a}}}_{s(a)} \right) = \boxed{s(a) \cdot [1-s(a)]} \end{aligned}$$

(questa formulazione)

Nel caso in cui $a = \varphi^\top \theta \Rightarrow s(a) = s(\varphi^\top \theta) = \frac{1}{1+e^{-\varphi^\top \theta}}$

$$\begin{aligned} \frac{ds(\varphi^\top \theta)}{\theta} &= \frac{d}{d\theta} \left[\frac{1}{1+e^{-\varphi^\top \theta}} \right] = \frac{d}{d\theta} \left[(1+e^{-\varphi^\top \theta})^{-1} \right] = \underset{dx_1}{\varphi_1} \cdot \underset{dx_1}{(-1)} \left(1+e^{-\varphi^\top \theta} \right)^{-2} \left(e^{-\varphi^\top \theta} \right) \\ &= -\varphi \cdot \left(1+e^{-\varphi^\top \theta} \right)^{-2} \left(e^{-\varphi^\top \theta} \right) = \underset{\text{stessa ragione}}{\underset{\text{Prima}}{\underset{dx_1}{\varphi}}} \cdot \underset{dx_1}{\left[\frac{1}{1+e^{-\varphi^\top \theta}} \right]} \left[1 - \underset{dx_1}{\left[\frac{1}{1+e^{-\varphi^\top \theta}} \right]} \right] \\ &= \boxed{\varphi \cdot \hat{y} \cdot (1-\hat{y})} \end{aligned}$$

Bisogna ora calcolare il gradiente della $J(\theta)$

$$J(\theta) = -\sum_{i=1}^N \left(y_{(i)} \ln \hat{p}_{(i)} + (1-y_{(i)}) \ln [1-\hat{p}_{(i)}] \right) \quad \hat{p}_{(i)} = \frac{1}{1+e^{-\theta^T x_{(i)}}}$$

$$\begin{aligned} \nabla J(\theta) &= -\sum_{i=1}^N \left(y_{(i)} \frac{\hat{p}'_{(i)}}{\hat{p}_{(i)}} + (1-y_{(i)}) \frac{-\hat{p}'_{(i)}}{1-\hat{p}_{(i)}} \right) = \\ &= -\sum_{i=1}^N \left(y_{(i)} \cdot \frac{\varphi_{(i)} \cdot \hat{p}_{(i)} [1-\hat{p}_{(i)}]}{\hat{p}_{(i)}} + (1-y_{(i)}) \frac{-\varphi_{(i)} \hat{p}_{(i)} [1-\hat{p}_{(i)}]}{1-\hat{p}_{(i)}} \right) \\ &= \sum_{i=1}^N \left(-y_{(i)} \varphi_{(i)} [1-\hat{p}_{(i)}] - (1-y_{(i)}) (\varphi_{(i)} \cdot \hat{p}_{(i)}) \right) \\ &= \sum_{i=1}^N \left(\varphi_{(i)} [-y_{(i)} + y_{(i)} \hat{p}_{(i)}] + \varphi_{(i)} [\hat{p}_{(i)} - y_{(i)} \hat{p}_{(i)}] \right) \\ &= \sum_{i=1}^N \left(\varphi_{(i)} [-y_{(i)} + y_{(i)} \hat{p}_{(i)} - y_{(i)} \hat{p}_{(i)} + \hat{p}_{(i)}] \right) \\ &= \sum_{i=1}^N \underbrace{\varphi_{(i)} (\hat{p}_{(i)} - y_{(i)})} \end{aligned}$$

Osservazione

Le derivate $\sum_{i=1}^N \varphi_{(i)} (\hat{p}_{(i)} - y_{(i)}) = 0$ sono un sistema di 1d equazioni non lineari in θ

↳ Non si mettono in schiera in forma diuse come per la regressione lineare \Rightarrow per via della non linearità della sigmoida

↳ si dimostra però che $J(\theta)$ è convessa, quindi ha un unico minimo

L'ottimizzazione è quindi solita utilizzare algoritmi iterativi di ottimizzazione.
Uno di questi è il GRADIENT DESCENT:

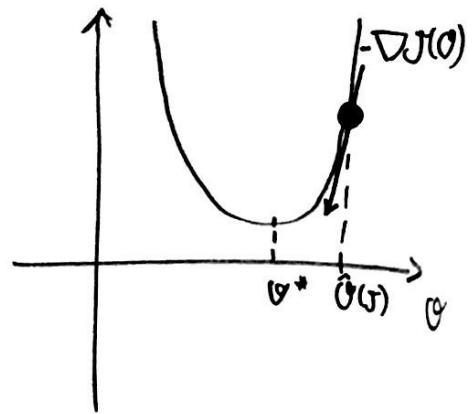
↳ il valore minore dei parametri all'iterazione $j+1$ è:

↳ α è la LEARNING RATE (determina il passo con cui abbassa il valore)

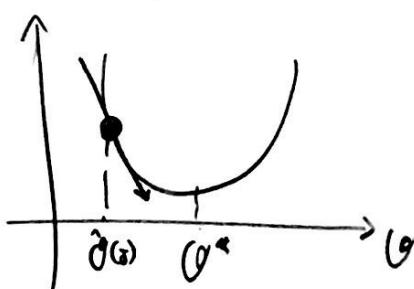
↳ $\hat{\theta}(0)$ è inizializzato RANDOM

$$\hat{\theta}(j+1) = \hat{\theta}(j) - \alpha \nabla J(\theta) \Big|_{\theta=\hat{\theta}(j)}$$

$$\hat{\theta}(\mathbf{J+1}) = \hat{\theta}(\mathbf{J}) - \alpha \nabla J(\theta) \quad |_{\theta = \hat{\theta}(\mathbf{J})}$$



- se $\nabla J(\theta) \Big|_{\theta=\hat{\theta}(\mathbf{J})} > 0 \Rightarrow \hat{\theta}(\mathbf{J+1}) < \hat{\theta}(\mathbf{J})$



STIMA BAYESIANA (BAYESIAN INFERENCE)

PROBABILITÀ CONGIUNTE, CONDIZIONATE, MARGINALI

Supponiamo di avere 2 variabili casuali a e b , discrete bimode, con le seguenti distribuzioni di probabilità congiunta:

DISTRIBUZIONE CONGIUNTA

$P(a, b)$

| | | a | |
|-----|-----|------|------|
| | | 0 | 1 |
| b | 0 | 0,06 | 0,24 |
| | 1 | 0,28 | 0,42 |

$$\sum_{a,b} P(a,b) = 1$$

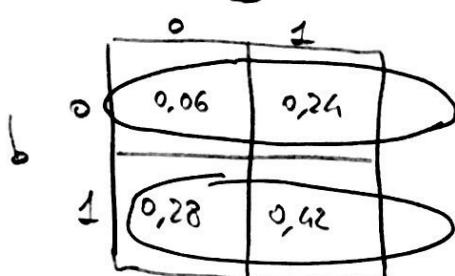
probabilità che si verifichi sia a che b , contemporaneamente

Le distribuzioni MARGINALI sono le distribuzioni di probabilità di un sottoinsieme di variabili casuali.

In nel nostro caso, dato che abbiamo 2 variabili casuali, vi saranno 2 prob. marginali, ovvero $p(a)$ e $p(b)$

- È ottenuta "marginalmente", ovvero sommando, rispetto alle variabili che non sono di interesse

DISTRIBUZIONE TISSAGNATE



$$P(b=0) = 0,3$$

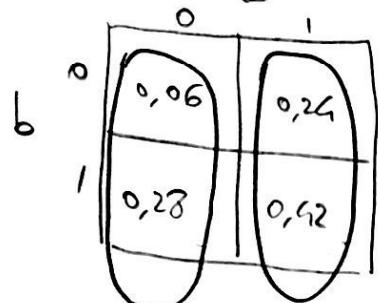
$$P(b=1) = 0,7$$

non mi interessa se $a=0$ o $a=1$,
→ interessate solo che $b=0$. Quindi la probabilità di $b=0$ è la somma delle probabilità quando

$$P(b) \quad b=0 \Rightarrow P(b=0, a=0) +$$

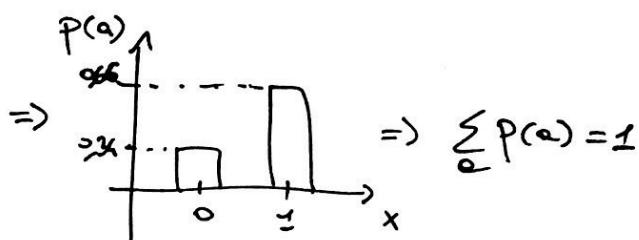


$$P(b=1) = P(b=1, a=0) + P(b=1, a=1) \Rightarrow \sum_b P(b) = 1$$



$$P(a=0) \quad P(a=1) = 0,66$$

11
0,34



$$\begin{aligned} P(a=0) &= P(a=0, b=0) + P(a=0, b=1) \\ P(a=1) &= P(a=1, b=0) + P(a=1, b=1) \end{aligned}$$

La distribuzione condizionata indica come le probabilità si redistribuiscono dato che si restringono le probabilità ad un particolare sottoinsieme.

Es

Siano date N persone, dove N_A è il numero di persone con capelli lunghi e N_B è il numero di persone di sesso femminile. Siano gli eventi:

A = persone con capelli lunghi

B = persone di sesso femminile

$$P(A) = \frac{N_A}{N} = \frac{\# \text{ di persone con capelli lunghi}}{\# \text{ totali di persone}}$$

$$P(B) = \frac{N_B}{N} = \frac{\# \text{ di donne}}{\# \text{ totali di persone}}$$

Consideriamo la sola popolazione femminile:

La probabilità di una persona scelta a caso da queste persone avere i capelli lunghi è $\frac{N_{AB}}{N_B}$, dove N_{AB} è il numero di donne con capelli lunghi.

L'questa probabilità è detta probabilità condizionata (al fatto che le persone sia di sesso femminile)

$$P(A|B) = \frac{N_{AB}}{N_B}$$

La popolazione considerata è N_B , non N

- Se probabilità di selezione tra donne con capelli lunghi è $P(A, B) = \frac{N_{AB}}{N}$
 $= \frac{\text{numero di donne con capelli lunghi}}{\text{totale di persone}}$



- Posso esprimere $P(A|B)$ come: $P(A|B) = \frac{N_{AB}/N}{N_B/N} = \frac{P(A, B)}{P(B)}$



Quindi: $P(A|B) = \frac{P(A, B)}{P(B)} \Rightarrow \boxed{P(A, B) = P(A|B)P(B)}$

- Osservazione
- Se probabilità che accade sia A che B è la probabilità che si verifichino B moltiplicata per la probabilità che si verifichi A dato che B si è verificato
 - $P(A, B) = P(A) \cdot P(B)$ se e solo se $P(A|B) = P(A)$. Questo vuol dire che A e B sono indipendenti, ovvero il verificarsi di B non modifica la probabilità che A si verifichi

Es: A: lancio un dadi ed esca 4
 B: lancio una moneta ed esca TESTA \Rightarrow anche se uscire croce, il dadi ha la stessa probabilità (1/6) di risultare impari 4
 $\hookrightarrow P(A, B) = P(A) \cdot P(B)$

- Supponiamo che $P(A|B) = P(B|A)$. Quindi: $P(B, A) = P(B|A)P(A)$, e di conseguenza:

$$P(A|B)P(B) = (P(B|A)P(A)) \Rightarrow P(B|A) = \boxed{\frac{P(A|B)P(B)}{P(A)}}$$

Osservazione

TEOREMA DI BAYES

- Il teorema di Bayes permette di ridistribuire le probabilità: prima conoscevo $P(B)$, adesso $P(B|A) \Rightarrow$ la probabilità di B è cambiata in seguito alla conoscenza di A

- $P(A) = \sum_B P(A|B)P(B)$ è la margionale di A, ovvero sommo rispettivamente i valori di B

(26)

Riprendendo l'esempio delle tabelle; calcoliamo la distribuzione $p(a|b)$

| | | |
|---|-----|-----|
| | 0 | 1 |
| 0 | 0,2 | 0,8 |
| 1 | 0,4 | 0,6 |

$$p(a=1|b=0) = \frac{p(a=1, b=0)}{p(b=0)} = \frac{0,24}{0,3} = 0,8$$

$$p(a=1|b=1) = \frac{p(a=1, b=1)}{p(b=1)} = \frac{0,42}{0,7} = 0,6 \quad p(a|b) = \frac{p(a, b)}{p(b)}$$

$$p(a=0|b=1) = \frac{p(a=0, b=1)}{p(b=1)} = \frac{0,28}{0,7} = 0,4$$

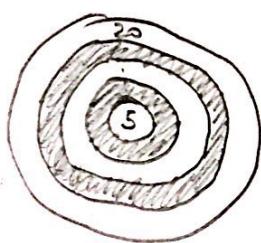
Oltre lo stesso modo possiamo calcolare $p(b|a)$.

$$p(b=1) = p(b=1|a=0)p(a=0) + p(b=1|a=1)p(a=1)$$

$$p(b=1, a=0) + p(b=1, a=1)$$

Esempio : Interpretazione delle probabilità condizionate come ridistribuzione delle probabilità

Supponiamo di tirare bersagliando una freccette contro un bersaglio con 20 cerchi concentrici



Le probabilità di beccare un cerchio qualsiasi, senza vedere, supponiamo sia $\frac{1}{20}$ (ogni cerchio è equiprobabile)

- Qual è la probabilità di aver beccato il cerchio numero 5?

$$P(\text{cerchio } \# 5) = \frac{1}{20}$$

Supponiamo che io abbia già dato che non ha preso il cerchio $\# 20$.

- Qual è allora la probabilità di aver beccato il cerchio $\# 5$?

Dato che non ho sicuramente preso il $\# 20$, la probabilità di aver preso il $\# 5$ è $P(\# 5 | \text{NOT } \# 20) = \frac{1}{19}$, perché fra i 19 valori possibili, escludendo escluso $\# 20$

Le probabilità si è quindi ridistribuita sui 19 esiti restanti sui 20 esiti

$$P(\# 5 | \text{NOT } \# 20) = \frac{P(\# 5, \text{NOT } \# 20)}{P(\text{NOT } \# 20)} = \frac{P(\# 5)}{P(\text{NOT } \# 20)} = \frac{\frac{1}{20}}{\frac{19}{20}} = \boxed{\frac{1}{19}}$$

(22)

INTRODUZIONE ALLA STIMA BAIEZIANA

Ottaviani finora considera il parametro ignoto θ come una variabile deterministica. Specie, però, ottaviani delle informazioni, delle credenze, sui possibili valori che potrebbe avere θ .

↳ Esempio:

↳ stima della concentrazione di anidride solforosa nell'aria: si ha un'idea dell'ordine di grandezza, in base anche a studi precedenti.

↳ stima della probabilità di una moneta non truccata risulti TESTA dopo un buco: si sa che non potrà essere 0,1 o 0,8 ma sarà attorno agli 0,5

Ha quindi senso considerare θ come una variabile casuale anche come variabile deterministica.

↳ In questo modo possiamo specificare una distribuzione di probabilità per θ (dato che è una v.c.), assegnando una probabilità maggiore a valori di θ di in credo sono più verosimili di θ assunto, e minor probabilità a valori di θ di in credo non potranno osservare.

Es.

Sia θ la probabilità di il lato di una moneta non truccata risulti in TESTA. Una possibile distribuzione per θ è: $P(\theta)$

Osservazioni

- $P(\theta)$ ha dominio $[0,1]$, perché θ , modello di una probabilità, deve stare tra 0 ed 1
- Dato che la moneta è non truccata, $\theta=0,5$ sarà il valore più probabile di θ , e $\theta=0$ o $\theta=1$ sono praticamente impossibili (la probabilità che θ sia 0 o 1 è vicina a 0)
- Dato che distribuzione su θ , ottaviani già una stima di θ (STIMA A PRIORI). Ad esempio possono prendere come valore pentuto per la stima di θ il valore ottenuto di $P(\theta)$. L'incertezza sulla stima sarà allora la varianza di $P(\theta)$ (INCERTITUDINE A PRIORI)
- Con l'arrivo di dati osservati, ci si aspetta che:
 - 1) Il valore ottenuto cambia;
 - 2) L'incertezza decresca (cioè più informazioni!)

Obbiamo quindi: due elementi di potere informazione:

- 1) La distribuzione ^{A PRIORI} sui possibili valori di θ , ovvero $P(\theta)$
- 2) L'informazione del potere: i dati sui possibili valori di θ , ovvero le likelihood $P(Y|\theta)$

Quello che vogliamo veramente è sapere quanto θ dà le osservazioni dati $P(\theta|Y)$

↓

Usando il Teorema di Bayes possova avere i due elementi di informazione:

$$P(\theta|Y) = \frac{P(Y|\theta) P(\theta)}{P(Y)}$$

- $P(\theta)$: PRIOR

- $P(Y|\theta)$: LIKELIHOOD

- $P(Y)$: MARGINAL LIKELIHOOD

- $P(\theta|Y)$: POSTERIOR

Osservazione

- $P(\theta|Y)$ è una distribuzione di possibili valori di θ , le cui probabilità sono modificate (riallocate, ridistribuite), rispetto a $P(\theta)$, dall'aver osservato i dati Y
- Nel caso in cui $P(Y|\theta)$ e $P(\theta)$ siano pdf continue (es. Gaussiane), allora $P(Y)$ sarà: $P(Y) = \int P(Y|\theta) P(\theta) d\theta$
- Considerare le forme funzionali di $P(\theta)$ e $P(Y|\theta)$ perché le imposte: come posso dire su che distribuzione sono $P(\theta|Y)$?
 - 1) In genere, nella. Solitamente la forma $P(\theta|Y)$ è in una forma funzionale nota.
 - 2) Questo avviene se, ad esempio, $P(\theta)$ è Gaussiana e $P(Y|\theta)$ è Gaussiana. Allora anche $P(\theta|Y)$ sarà Gaussiana.
 - 3) Un altro problema è che $P(Y)$ è un integrale da potremmo non sapere come risolvere.

↓

Per far fronte a questi problemi, si usano metodi numerici e di campionamento che evitano il calcolo analitico. Questi metodi si chiamano MARKOV CHAIN MONTE CARLO (MCMC)

Un modo per calcolare $P(\theta|Y)$ da cui si basa nei sul calcolo analitico nei suoi metodi MCMC è quello di discretizzare il range di valori del parametro θ tramite una griglia di valori.

→ Valori $P(Y|\theta)$ e $P(\theta)$ solo in quei valori di θ

Esempio

Stimare le probabilità che il lancio di una moneta risulti in TESTA. Supponiamo di lanciare una moneta N volte. Osserviamo i dati $y^{(i)}$:

$$y^{(i)} = \begin{cases} 1 & \text{se TESTA} \\ 0 & \text{se CROCE} \end{cases} \quad i = 1, \dots, N$$

Modellizziamo i dati, categorici e dicotomici, con una distribuzione di Bernoulli.

$y^{(i)} \sim \text{Bernoulli}(\pi)$, iid., π : Prob. TESTA (parametro ignoto)

$$P(Y|\pi) = \pi^y \cdot (1-\pi)^{1-y}$$

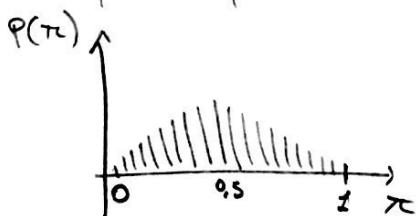
- Se $y=1 \Rightarrow P(Y=1|\pi) = \pi$

- se $y=0 \Rightarrow P(Y=0|\pi) = 1-\pi$

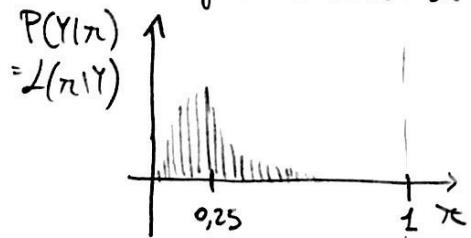
$$L(\pi|Y) = \prod_{i=1}^N \pi^{y^{(i)}} \cdot (1-\pi)^{1-y^{(i)}}$$

$$\stackrel{\downarrow}{\text{O}} \left[y^{(1)}, \dots, y^{(N)} \right] \stackrel{i=1}{=} \pi^{\sum_{i=1}^N y^{(i)}} \cdot (1-\pi)^{\sum_{i=1}^N 1-y^{(i)}} = \pi^{\text{#successi}} \cdot (1-\pi)^{\text{#fallimenti}}$$

- Supponiamo una prior di questo tipo



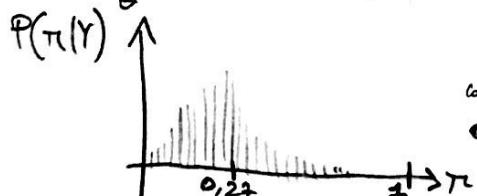
- Supponiamo di aver osservato $y=10$ successi su $N=40$ lanci. La likelihood ha la forma:



Il valore più probabile del parametro π è la stima ML. Nel caso di likelihood Bernoulli, br che $\hat{\pi} = \frac{1}{N} \sum_{i=1}^N y^{(i)}$, ovvero la % di successi.

In questo caso $\hat{\pi} = \frac{10}{40} = 0,25$
e dividere per $P(Y) = \sum_{\theta} P(Y|\theta) P(\theta)$, che somma su ogni valore di θ (della griglia)

- Per calcolare la posterior faccio il quoziente di $P(Y|\theta)$ e $P(\theta)$ per ogni valore di θ



La NDA è un compromesso tra
0,25 e 0,5

D'INTRO
⇒

(30)

- L'opposi^ta a qualche cosa c'è percepibile nel caso in cui θ sia un vettore con molte componenti.



Il PC ci impiegherebbe troppo a fare tutte le combinazioni di parametri

- La soluzioⁿe è o usare prior e likelihood tali che le posteriori siano forme ^{note} che si può ricavare analiticamente (se le posteriori hanno stesse forme delle priori, prior e likelihood si dicono CONVOLUTE)

L'opzione usare metodi MCMC

Supponiamo di avere $P(\Theta|Y)$. Ossiamo una distribuzione di valori del parametro Θ ignoto. Ci sono poi un valore solo, un valore puntuale

de valore puntuale per la nostra stima $\hat{\Theta}$?

Ci sono varie possibilità:

1) $\hat{\Theta} = \text{argmax}_{\Theta} P(\Theta|Y)$, ovvero prendo il valore Θ da cui la risposta esca più probabile

Questa stima è nota come stima MAXIMUM A POSTERIORI (MAP)

2) $\hat{\Theta} = E[P(\Theta|Y)] = E[\Theta|Y]$, la MEDIA delle distribuzioni a posteriori

3) Altre quantità come la MEDIANA, ecc.

Ricordiamo che in generale individuiamo un stimatore come una funzione T dei dati D :

$$\hat{\Theta} = T(D)$$

Vogliamo che la variabile casuale $\hat{\Theta}$ sia vicina alla variabile casuale Θ . Usiamo quindi la funzione di costo:

$$J(T(\cdot)) = E[\|\Theta - T(D)\|^2] \quad (*) \quad \text{MEAN SQUARED ERROR}$$

La stima ottima di Bayes è quella funzione $T^*(\cdot)$ tale che:

$$E[\|\Theta - T^*(D)\|^2] \leq E[\|\Theta - T(D)\|^2] \quad \forall T(\cdot)$$

cioè che minimizza la cifra di merito rispetto a $T(\cdot)$

Si dimostra che $T^*(Y) = E[\Theta | D=Y]$, ovvero il valore ottenuto dalla

distribuzione $P(\Theta|Y)$, cioè il valore ottenuto condizionato al fatto che i dati D abbiano assunto valore Y

↓

Considereremo quindi $E[\Theta|Y]$ come stima puntuale di $\hat{\Theta}$, soprattutto in che senso essa è una stima ottima

nel senso che

mimimizza (*)

Supponiamo ora che sia dato che il parametro θ sia una v.r. Gaussiana, quindi la loro pdf conjunta è Gaussiana.

↓
Supponiamo per semplicità di avere un dato scelto y e che o sia scelte, tali che $E[y] = 0$ e $E[\theta] = 0$

Vogliamo calcolare $P(\theta|y)$. Essendo θ e y completamente Gaussiane si ha che:

$$\begin{bmatrix} y \\ \theta \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \lambda_{yy} & \lambda_{y\theta} \\ \lambda_{\theta y} & \lambda_{\theta\theta} \end{bmatrix}\right)$$

μ : vettore medie
 Σ : matrice varianza-covarianza

al qualsiasi perché
è coniugata d'è risolvibile.

La pdf coniugata $P(\theta, y)$ ha quindi la forma:

$$P(\theta, y) = \frac{1}{\sqrt{(2\pi)^3 \det \Sigma}} e^{-\frac{1}{2} (z - \mu)^T \Sigma^{-1} (z - \mu)}$$

La pdf dei dati è $P(y) = \frac{1}{\sqrt{2\pi \lambda_{yy}}} e^{-\frac{1}{2\lambda_{yy}} (y - 0)^2}$

Si dimostra che $P(\theta|y) = \frac{P(\theta, y)}{P(y)}$ è una Gaussiana, $P(\theta|y) = N(\mu_{\theta|y}, \lambda_{\theta|y})$, con:

- VALORE ATTESO:

$$\boxed{\mu_{\theta|y} = \frac{\lambda_{\theta y}}{\lambda_{yy}} \cdot y}$$

- VARIANZA:

$$\boxed{\lambda_{\theta|y} = \lambda_{\theta\theta} - \frac{\lambda_{\theta y}^2}{\lambda_{yy}}}$$

Il valore $\frac{\lambda_{\theta y}}{\lambda_{yy}}$ è > 0 . Quindi l'incertezza è posteriore $\lambda_{\theta|y}$ è MINORE di quella a priori

Ora se osserviamo il dato $y = y^{(1)}$, lo stimatore ottimale Bayes sarebbe:

$$\boxed{\hat{\theta} = E[\theta | y = y^{(1)}] = \frac{\lambda_{\theta y}}{\lambda_{yy}} y^{(1)}}$$

Si può calcolare le varianze dell'errore di stima, ovvero:

$$\text{Var}[\theta - \hat{\theta}] = E\left[\left((\theta - \hat{\theta}) - E[\theta - \hat{\theta}]\right)^2\right]$$

$$\hookrightarrow E[\theta - \hat{\theta}] = E[\theta] - E[\hat{\theta}] = 0 - E\left[\frac{\partial \theta}{\partial y} y\right] = 0 - 0 \\ = 0 \text{ per ipotesi}$$

$$\Rightarrow \text{Var}[\theta - \hat{\theta}] = E[(\theta - \hat{\theta})^2] = E\left[\left(\theta - \frac{\partial \theta}{\partial y} y\right)^2\right] = E\left[\theta^2 - 2\frac{\partial \theta}{\partial y} \theta y + \frac{\partial^2 \theta}{\partial y^2} y^2\right] \\ = E[\theta^2] - 2\frac{\partial \theta}{\partial y} E[\theta y] + \frac{\partial^2 \theta}{\partial y^2} E[y^2] \\ = \lambda_{\theta\theta} - 2\frac{\partial \theta}{\partial y} \cdot \lambda_{\theta y} + \frac{\partial^2 \theta}{\partial y^2} \cdot \lambda_{yy} = \boxed{\lambda_{\theta\theta} - \frac{\partial^2 \theta}{\partial y^2}}$$

VARIANZA
DELLA STIMA

STIMA LINEARE

Non è sempre il caso che θ e y siano confiamente Gaussiane. Vogliamo quindi trovare un stimatore di non forte dipendenza sulle pdf compiute di θ ed y .

Supponiamo θ e y r.c. solni con valore atteso nullo e varianza $\lambda_{\theta\theta}$ e λ_{yy} rispettivamente.

$$E[\theta] = 0 \quad E[y] = 0 \quad E[\theta^2] = \lambda_{\theta\theta} \quad E[y^2] = \lambda_{yy} \quad E[\theta y] = \lambda_{\theta y}$$

Vogliamo stimare θ dato y tenendo un stimatore buono, t.c.:

$$\hat{\theta} = \alpha y + \beta \quad \alpha, \beta \in \mathbb{R} \text{ parametri reali}$$

Per trovare α e β , impostare la cifra da minima da mimimizzare
è la varianza dell'errore di stima

$$J(\alpha, \beta) = E\left[\left(\theta - \hat{\theta}\right)^2\right] = E\left[\left(\theta - \alpha y - \beta\right)^2\right]$$

$$\bullet \frac{\partial J(\alpha, \beta)}{\partial \alpha} = 0 \Rightarrow 2 \cdot E\left[\left(\theta - \alpha y - \beta\right) \cdot (-y)\right] = 2 \left(E[-\theta y] + E[\alpha y^2] + E[\beta y]\right) = \\ = 2 \left(-\lambda_{\theta y} + \alpha \lambda_{yy} + \beta \cdot 0\right) = 2(-\lambda_{\theta y} + \alpha \lambda_{yy}) = 0$$

$$\bullet \frac{\partial J(\alpha, \beta)}{\partial \beta} = 0 \Rightarrow 2E[(\theta - \alpha y - \beta) \cdot (-1)] = 2E[-(\theta + \alpha y + \beta)] = \\ = 2(E[\theta] + \alpha E[y] + E[\beta]) = 2\beta = 0$$

$$\begin{cases} 2(-\alpha y + \beta y) = 0 \\ 2\beta = 0 \end{cases} \quad \begin{cases} \alpha = \frac{\partial \theta}{\partial y} \\ \beta = 0 \end{cases}$$

Lo stimatore lineare ottimo è dato quindi da:

$$\hat{\theta} = \alpha y + \beta = \frac{\partial \theta}{\partial y} \cdot y + 0 = \frac{\partial \theta}{\partial y} \cdot y$$

CONCIDE CON LO STIMATORE DI PATES !!
NEL CASO GAUSSIANO

La varianza dell'errore di stima si ricava essere:

$$\text{Var}[\theta - \hat{\theta}] = \sigma_{\theta\theta} - \frac{\partial \theta}{\partial y}^2$$

COME PATES NEL CASO GAUSSIANO !
↓
l'incertezza dell'errore di stima è minore
rispetto a quella a priori

Osservazione

Lo stimatore lineare con le medesime ipotesi sulla distribuzione congiunta delle variabili. Infatti gli basta conoscere $\partial \theta / \partial y$ e σ_{yy} .



Potrebbe dunque esserci un stimatore migliore di quello lineare ottimo, cioè con varianza dell'errore di stima minore



Se però incognita è data somma distribuzione congiuntamente gaussiana, non esiste stimatore migliore di quello lineare ottimo

Osservazioni

- 1) Se $\partial \theta / \partial y = 0$, cioè θ e y sono incorrrelati, ovvero il dato y non porta informazioni su θ , allora le stime a priori con viene modificate dal dato. Infatti: $\hat{\theta} = 0$ se $\partial \theta / \partial y = 0$, e $\text{Var}[\theta - \hat{\theta}] = \text{Var}[\theta] = \sigma_{\theta\theta}$
- 2) A parità di $\partial \theta / \partial y$, più elevata è σ_{yy} , e più piccola sarà la diminuzione di $\text{Var}[\theta - \hat{\theta}]$ causata dal dato y . Un valore elevato di σ_{yy} significa che il dato y è affetto da elevata incertezza (quindi porta poca informazione)

GENERALIZZAZIONE 1: valore ottenuto nullo, θ e y scarsi

Se $E[\theta] = \mu_\theta \neq 0 \Rightarrow$ lo stimatore di Bayes nel caso gaussiano e lo stimatore lineare ottimale sono:

$$\hat{\theta} = \mu_\theta + \frac{\lambda_{\theta y}}{\lambda_{yy}} (y - \mu_y)$$
$$\text{Var}[\theta - \hat{\theta}] = \lambda_{\theta\theta} - \frac{\lambda_{\theta y}^2}{\lambda_{yy}}$$

GENERALIZZAZIONE 2: y e θ sono vettoriali, $y \in \mathbb{R}^{m_y \times 1}$, $\theta \in \mathbb{R}^{m_\theta \times 1}$

Se $E[y] = \mu_y \neq 0$
 $E[\theta] = \mu_\theta \neq 0$

$$\text{Var}\begin{bmatrix} y \\ \theta \end{bmatrix} = \begin{bmatrix} \Lambda_{yy} & \Lambda_{y\theta} \\ \Lambda_{\theta y} & \Lambda_{\theta\theta} \end{bmatrix} \quad \text{con } \Lambda_{y\theta} = \Lambda_{\theta y}^T$$

$$\mu_\theta \in \mathbb{R}^{m_\theta \times 1}$$
$$\mu_y \in \mathbb{R}^{m_y \times 1}$$

$$\Lambda_{y\theta} \in \mathbb{R}^{m_y \times m_\theta}, \quad \Lambda_{\theta\theta} \in \mathbb{R}^{m_\theta \times m_\theta}$$
$$\text{Var}\begin{bmatrix} y \\ \theta \end{bmatrix} \in \mathbb{R}^{(m_y + m_\theta) \times (m_y + m_\theta)}$$

Oltora lo stimatore di Bayes nel caso gaussiano e lo stimatore lineare ottimale sono dati da:

$$\hat{\theta} = \mu_\theta + \Lambda_{\theta y} \Lambda_{yy}^{-1} (y - \mu_y)$$
$$\text{Var}[\theta - \hat{\theta}] = \Lambda_{\theta\theta} - \Lambda_{\theta y} \Lambda_{yy}^{-1} \Lambda_{y\theta}$$

Note

Le formule appena viste omaggiano alle forme ricorsive: si effettua cioè la stima $\hat{\theta}$ con l'ausilio di nuovi dati, portando della stima precedente



Queste equazioni ricorsive saranno alla base del FILTO DE HANNAN, in cui lo stato $x(t)$ e l'uscita $y(t)$ sono visti come variabili casuali, e si vuol stimare lo stato $x(t)$ (l'incognita) dato l'osservazione dei dati $y(t)$

STIMA PARISIANA DEL VALORE ATTESO DI VARIABILE GAUSSIANA

Siano $y_i \sim N(\theta, \sigma_{yy}^2)$, i.i.d., ignoto. Si vuole stimare il parametro θ tramite stima Bayesiana. Supponiamo $N=1$

Si definisce quindi una priori sul parametro θ , ovvero $P(\theta)$. Osserviamo che il parametro ignoto in questo caso è $\theta = E[y]$, ovvero il valore atteso di y .

Imponiamo una priori Gaussiana su θ , ovvero $P(\theta) = N(\mu_\theta, \sigma_{\theta\theta}^2)$

Un modo per descrivere i dati y è: $y(i) = \theta + e(i)$ con $e(i) \sim N(0, \sigma_{ee}^2)$, i.i.d., $e(i) \perp \theta$, ovvero, i dati hanno media data dal valore di θ e disturbi dati da $e(i)$

possiamo definire la likelihood $P(y|\theta) = N(\theta, \sigma_{ee}^2)$, in cui θ è la variabile indipendente.

Siamo quindi nel tipico caso di inferenza bayesiana in cui ho un prior su un parametro, $P(\theta)$, e la mia likelihood in funzione di quel parametro, $P(y|\theta)$. Possiamo calcolare la posterior come:

$$P(\theta|y) = \frac{P(y|\theta) \cdot P(\theta)}{P(y)} \quad \rightarrow \text{sicché è Gaussiana!}$$

dove $P(y) = \int_{-\infty}^{+\infty} P(y|\theta) P(\theta) d\theta$, e usare come $\hat{\theta}$ il valore atteso condizionato di $P(\theta|y)$.

Osserviamo poi che, dato che $P(y|\theta)$ e $P(\theta)$ sono Gaussiane, allora anche $P(y|\theta)$ è Gaussiana.

Quindi useremo le formule parziali per le distribuzioni condizionate nel caso Gaussiano, e useremo $\hat{\theta} = E[\theta|y]$ (che coincide con la stima da buone ottime).

La stima ottima è quindi:

$$\hat{\theta} = \mu_\theta + \frac{\partial \log}{\partial \theta} (y - E[y])$$

Dovendo calcolare $E[y]$, $\partial \log$, $\partial \theta$

$$\bullet E[g] = E[\theta + e] = E[\theta] + E[e] = \mu_\theta + 0 = \boxed{\mu_\theta}$$

$$\bullet \gamma_{\theta y} = E[(\theta - \mu_\theta) \cdot (y - \mu_y)] = E[\theta y - \theta \mu_y - \mu_\theta y + \mu_\theta^2]$$

$$= E[\theta y] - E[\theta \mu_y] - E[\mu_\theta y] + E[\mu_\theta^2]$$

$$= E[\theta(\theta + e)] - \cancel{\mu_\theta \mu_\theta} - \cancel{\mu_\theta \mu_\theta} + \cancel{\mu_\theta^2}$$

$$= E[\theta^2] + E[ee] - \mu_\theta^2 = E[\theta^2] - E[\theta]^2 = \text{Var}[\theta] = \boxed{\lambda_{\theta\theta}}$$

$$\bullet \gamma_{yy} = E[(y - E[y])^2] = E[(y - \mu_y)^2] = E[y^2 - 2y\mu_y + \mu_y^2]$$

$$= E[y^2] - 2\mu_y E[y] + E[\mu_y^2]$$

$$= E[(\theta + e)^2] - 2\mu_\theta \mu_y + \mu_y^2$$

$$= E[\theta^2 + 2\theta e + e^2] - \mu_y^2 = E[\theta^2] + 2E[\theta e] + E[e^2] - \mu_y^2$$

$$= \underbrace{E[\theta^2] - E[\theta]^2}_{\text{Var}[\theta]} + E[e^2]$$

$$= \text{Var}[\theta] + \text{Var}[e] = \boxed{\lambda_{\theta\theta} + \lambda_{ee}}$$

Quindi:

$$\hat{\theta} = \mu_\theta + \frac{\lambda_{\theta y}}{\lambda_{yy}} (y - E[y]) = \mu_\theta + \frac{\lambda_{\theta\theta}}{\lambda_{\theta\theta} + \lambda_{ee}} (y - \mu_y)$$

$$= \mu_\theta + \frac{\lambda_{\theta\theta}}{\lambda_{\theta\theta} + \lambda_{ee}} y - \frac{\lambda_{\theta\theta}}{\lambda_{\theta\theta} + \lambda_{ee}} \mu_y = \frac{\mu_\theta (\lambda_{ee} + \lambda_{\theta\theta}) + \lambda_{\theta\theta} y - \lambda_{\theta\theta} \mu_y}{\lambda_{\theta\theta} + \lambda_{ee}}$$

$$= \boxed{\frac{\lambda_{ee}}{\lambda_{\theta\theta} + \lambda_{ee}} \mu_\theta + \frac{\lambda_{\theta\theta}}{\lambda_{\theta\theta} + \lambda_{ee}} y}$$

È IL VALORE ATTESO DELLA
POSTERIORI $P(\theta | y)$

↓
la distribuzione a posteriori delle
media le queste valori ottenuti

Osservazioni

- La ~~stima~~ stima a posteriori del valore osservato è una via di mettere tra le stime a priori μ_0 e le stime date dal dato, ovvero il valore y
- Nel caso in cui osserviamo N dati, ha che:

$$\hat{\theta} = \frac{\lambda_{\text{rec}}}{N \cdot \lambda_{\text{rec}} + \lambda_{\text{pri}}} \mu_0 + \frac{N \cdot \lambda_{\text{rec}}}{N \cdot \lambda_{\text{rec}} + \lambda_{\text{pri}}} \hat{\mu}_{\text{ML}}$$

STIMA ML della media
 di una gaussiana
 $\mu_{\text{ML}} = \frac{1}{N} \sum_{i=1}^N y_i$

L se $N \rightarrow \infty$, allora $\hat{\theta} = \hat{\mu}_{\text{ML}} \Rightarrow$ for un sacco di evidenze!!

L se $\lambda_{\text{rec}} \gg N \lambda_{\text{pri}}$, allora i dati hanno molta incertezza e non puoi combinarre le stime a priori

* PARTE II: SISTEMI DINAMICI *

Tuttavia ci sono 2 problemi:

- 1) Analisi e modellistica di serie temporali
- 2) Analisi e modellistica di sistemi I/O

* SERIE TEMPORALI *

Immagine di dati nel tempo $D = \{y(1), y(2), \dots, y(N)\}$. Indichiamo ogni dato con $y(t)$, anziché $y(i)$ che denotava dati statici

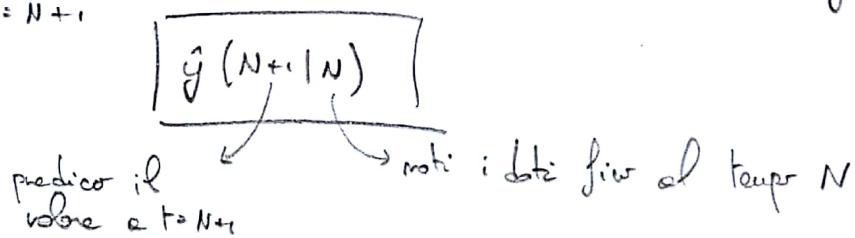


Esempi

- Valori di un titolo azionario
- mm di piogge caduti in una settimana
- concentrazione di un ormone in un individuo, misurata ogni giorno alla stessa ora

Che problema vogliono risolvere?

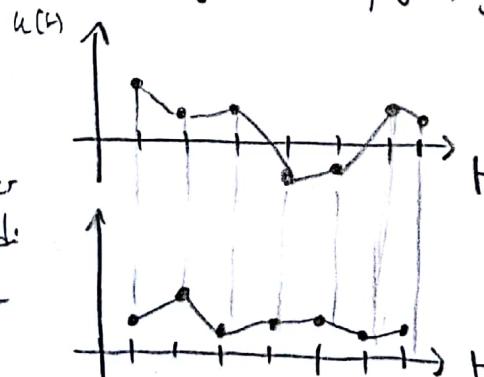
PREDICTION: noti i dati da $t=1$ a $t=N$, prevedere il valore di y al tempo $t=N+1$



* SISTEMI INGRESSO/USCITA *

Ottiamo 2 insiemni di dati, uno di ingressi ed un di uscite

$$\{u(1), u(2), \dots, u(N)\} \quad \{y(1), y(2), \dots, y(N)\}$$



Note

La presenza di un ingresso può ridurre l'incertezza di previsione dell'uscita

Esempio

INGRESSO (CURVA)

comune
disfoglio medicinale
mm di pioggia

(40)

USCITA (FETTE)

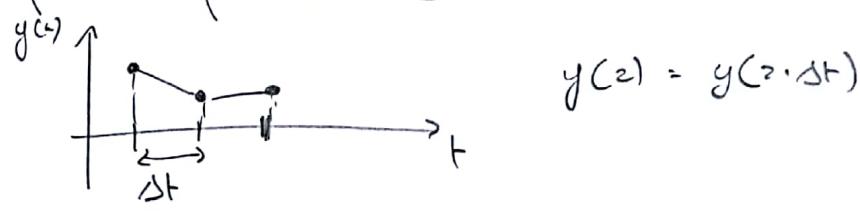
Coppia
concentrazione
ormone
concentrazione di

Che tipi di problemi vogliono risolvere?

- 1) PREDIZIONE: come prevedere
- 2) CONTROLLO: determinare le relazioni $a \rightarrow y$, in modo da progettare un controllore che determini $u(t)$

Osservazione

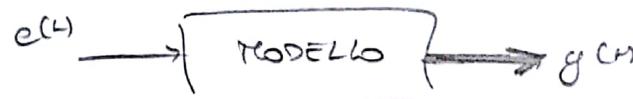
Provare con segnali e sistemi a tempo discreto. I segnali sono quindi compresi con tempi di campionamento Δt



Come impostare il problema?

T SERIE TEMPORALE

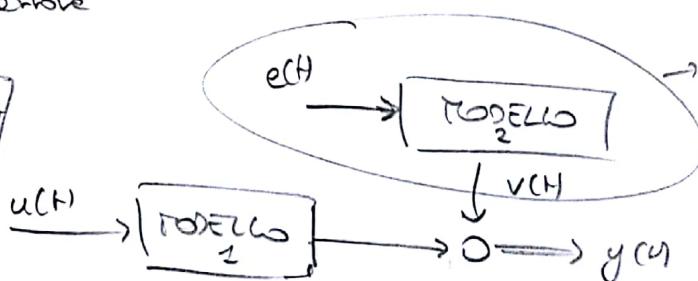
Modellizzare la serie temporale $y(t)$ come l'uscita di un sistema con impulsi non misurabili



SISTEMI I/O

Modellizzare l'uscita come somma di una componente deterministica e una componente di errore

Note: $e(t)$ è un impulso standard noto come white noise



modelli suelli che
u(t) non riconosce e
spiegare dell'uscita
y(t)
- rumore di misura
- errori di modelli

Osservazione

Come nel caso dei sistemi statici, le y sono offerte da rumore. In quel caso ossiamo modellato i dati y come delle variabili casuali (faccendo ipotesi sulle loro distribuzioni di probabilità)

↓

In questo caso però i dati non sono indipendenti, ma sono composta da un segnale che evolue nel tempo

Q1

Non obbligatori più quindi osservazioni di variabili casuali simple, ma osservazioni una successione di v.c. nel tempo \Rightarrow processi stocastici

PROCESSI STOCASTICI

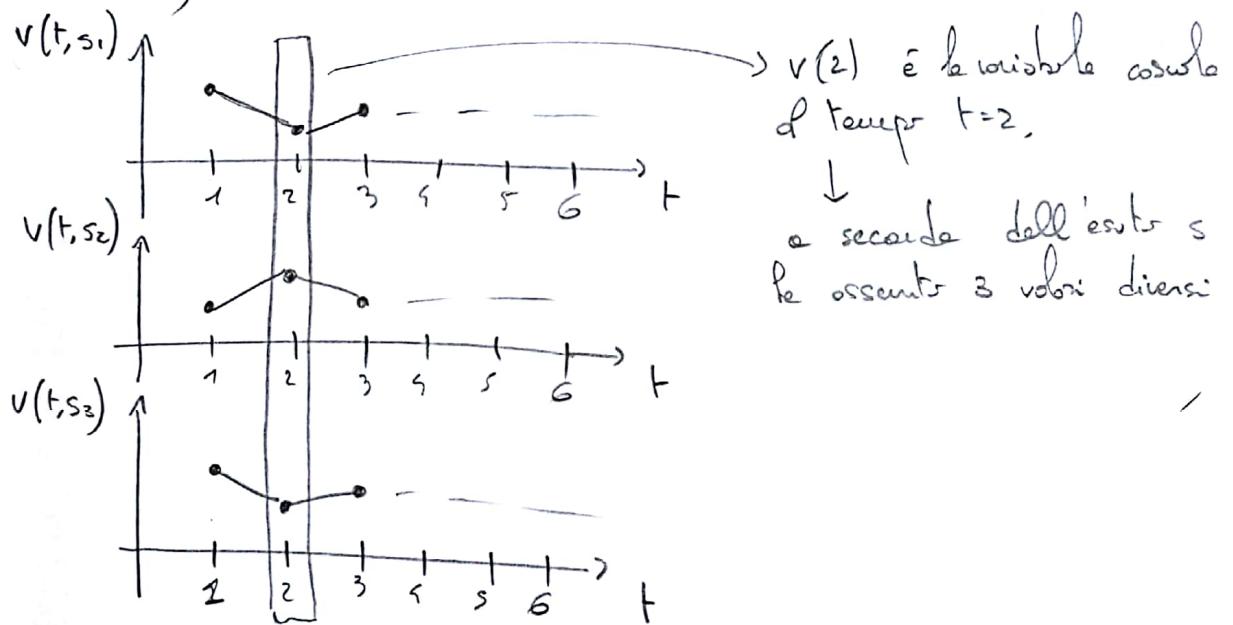
(INFINITA)

Un processo stocastico a tempo discreto è una successione di v.c. definite a partire dello stesso esperimento casuale s e ordinate secondo un indice temporale t

$$v(1, s), v(2, s), v(3, s), \dots, v(t, s)$$

- Fissato l'esito $s = \bar{s}$, si ottiene una REALIZZAZIONE del processo stocastico. Se cambia l'esito, ottengono un'altra serie di valori

- Si può pensare ad un PS come ad un segnale (anche se il PS può avere diverse realizzazioni)



Note

Spostare omettendo le dipendenze da s , indicando $v(1, \bar{s}), v(2, \bar{s}), \dots, v(3, \bar{s})$ con $v(1), v(2), v(3)$

Prima interpretazione: dhi gci come variabili casuali per gestire l'incertezza delle loro misure

Adesso interpretiamo le serie di dhi g(t) come realizzazione limite di un processo stocastico. Il segnale $u(t)$ è un segnale qualsunque \rightarrow STAZIONARIO

Dato un processo stocastico $v(t, s)$ si definiscono:

- **VALORE ATTECO**

$$m(t) = E[v(t, s)]$$

- è il valore atteso della variabile casuale $v(t, s)$ al tempo t

- Il valore atteso è rispetto a tutti gli esiti s

- **CORRIVANZA**

$$\gamma(t_1, t_2) = E[(v(t_1) - m(t_1)) \cdot (v(t_2) - m(t_2))]$$

- è la covarianza tra la variabile v al tempo t_1 e al tempo t_2

Nel caso in cui $t_1 = t_2 = t$, otteniamo la varianza al tempo t :

$$\gamma(t, t) = E[(v(t) - m(t))^2]$$

Le teorie che svilupperà si basano su un particolare tipo di ps.

PROCESSI STOCASTICI STAZIONARI (PSS)

Definizione

- Un processo stocastico si dice STAZIONARIO IN SENSO TOTALE se e solo se, $\forall n \in \mathbb{N}$, scelti t_1, t_2, \dots, t_m , il comportamento delle m-uple $v(t_1 + \tau), v(t_2 + \tau), \dots, v(t_m + \tau)$ è lo stesso di quello delle m-uple $v(t_1), v(t_2), \dots, v(t_m)$

↓

le caratteristiche probabilistiche delle m-uple $v(t_1), v(t_2), \dots, v(t_m)$ sono uguali a quelle delle m-uple $v(t_1 + \tau), v(t_2 + \tau), \dots, v(t_m + \tau)$

- Un processo stocastico si dice STAZIONARIO IN SENSO DEROLLE se:

- 1) $m(t) = m \quad \forall t$

- 2) $\gamma(t_1, t_2) = \gamma(t_3, t_4) \quad \text{se} \quad |t_2 - t_1| = |t_4 - t_3| = \tau$
 - la covarianza dipende solo dal tempo τ e non dai valori specifici di t_1, t_2, t_3, t_4

Dato che la covarianza dipende solo da τ , si usa:

$$\gamma(\tau) = E[(v(t) - m) \cdot (v(t + \tau) - m)]$$

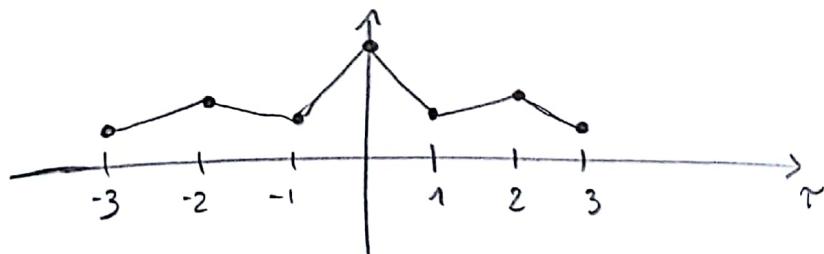
- è costante $\forall t$

PROPRIETÀ DELLE FUNZIONI DI COVARIANZA DI UN PSS

- 1) $\gamma(0) = E[(v(t) - m)^2] \geq 0$ VARIANZA DEL PROCESSO

- 2) $|\gamma(\tau)| \leq \gamma(0) \quad \forall \tau$ (la funzione è limitata)

- 3) $\gamma(\tau) = \gamma(-\tau)$ (è una funzione pari)



Definizione

Due processi stoc. stat. $v_1(t)$ e $v_2(t)$ si dicono equivalenti se loro le stesse volte ottengono lo stesso valore m e la stessa covariante $\gamma(\tau)$ $\forall \tau$

Note

Durante il corso studiare un pss

CASO PARTICOLARE DI PSS: RUMORE BIANCO (WHITE NOISE)

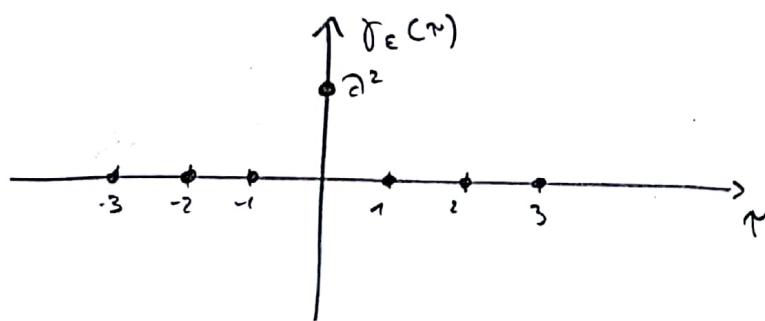
Definizione

Un pss $e(t)$ è detto RUMORE BIANCO, e lo si indica come $e(t) \sim WN(\mu, \sigma^2)$ se:

- 1) $E[e(t)] = \mu$

- 2) $\gamma(0) = E[(e(t) - \mu)^2] = \sigma^2 \quad \forall t$

- 3) $\gamma(\tau) = E[(e(t) - \mu)(e(t+\tau) - \mu)] = 0 \quad \forall t, \forall \tau \neq 0$



Il wn varia in modo imprevedibile da un istante all'altro

Note

Ma è determinata la distribuzione delle svolte v.c. $\epsilon(t)$. Possiamo essere Gaussiane, uniformi, ... In particolare si indica con WGN un rumore bianco Gaussiano

Note 2

Consideriamo pss a media nulla, infatti non cambia la caratteristica spettrale
 → dts de studiamo segnali nel tempo, si possono rappresentare in frequenza
RAPPRESENTAZIONE SPETTRALE DI UN PSS → se studiare il pss → una quantità che nel tempo è la $y(t)$, vogliamo studiare il pss → una quantità di b

Sia $y(t)$ un pss. Si definisce densità spettrale di potenza la trasformata di Fourier a tempo discreto delle funzoni di corrispondente $\tilde{Y}(u)$:

$$\tilde{Y}_y(u) = \sum_{v=-\infty}^{+\infty} Y_y(v) e^{-j2\pi u v}$$

- come le varie frequenze di $y(t)$ contribuiscono alla varianza di $y(t)$
- come l'energia del segnale si distribuisce alle varie frequenze

Note

$\tilde{Y}_y(u)$ esiste solo per pss tali che $Y_y(v) \rightarrow 0$ per $v \rightarrow +\infty$. Studieremo così in cui questo vale sempre

Proprietà di $\tilde{Y}_y(u)$:

1) $\tilde{Y}_y(u)$ è una funzione reale delle variabili reali u

$$\text{Im}(\tilde{Y}_y(u)) = 0 \quad \forall u \in \mathbb{R}$$

2) $\tilde{Y}_y(u)$ è una funzione positiva: $\tilde{Y}_y(u) \geq 0 \quad \forall u \in \mathbb{R}$

3) $\tilde{Y}_y(u)$ è una funzione pari: $\tilde{Y}_y(u) = \tilde{Y}_y(-u) \quad \forall u \in \mathbb{R}$

4) $\tilde{Y}_y(u)$ è una funzione periodica di periodo 2π :

$$\tilde{Y}_y(u) = \tilde{Y}_y(u + k \cdot 2\pi) \quad \forall u \in \mathbb{R}, \quad k \in \mathbb{Z}$$

Note

Come conseguenza di 4) si può tracciare la funzione solo tra $[-\pi, \pi]$

Si può ricavare $y(t)$ partendo da $\tilde{Y}_y(u)$ tramite l'antitrasformata:

$$y(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \tilde{Y}_y(u) e^{+ju\pi} du$$

Si mette de:

$$P_y(0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} P_y(\omega) e^{j\omega \cdot 0} d\omega = \frac{1}{2\pi} \int_{-\pi}^{\pi} P_y(\omega) d\omega \rightarrow \text{area sotto } P_y(\omega)$$

Quindi, la varianza è l'area sotto della densità spettrale di potenza, a meno del fattore 2π .

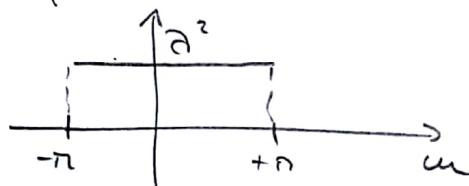
DENSITÀ SPETTRALE DI POTENZA DI UN RUOTONE BIANCO

Sia $c(t) = \alpha \cos(\omega t + \theta)$. Nel tempo è un segnale imprevedibile

Sappiamo che: $P_e(\omega) = \begin{cases} \alpha^2 & \text{se } \omega = 0 \\ 0 & \text{se } \omega \neq 0 \end{cases}$

Quindi: $P_e(\omega) = \sum_{\tau=-\infty}^{+\infty} g_e(\tau) e^{-j\omega\tau} = \alpha^2 \cdot e^{-j\omega \cdot 0} = \boxed{\alpha^2}$

La densità spettrale di potenza del rumore bianco è una costante



Questo vuol dire che tutte le frequenze contribuiscono in egual misura alla visibilità del segnale, non vi sono frequenze predominate

quindi \downarrow imprevedibile

RAPPRESENTAZIONE DINAMICA DI UN PSS

Obbligatori rappresentare un PSS sia nel tempo con $f(t)$ sia nelle frequenze con $P(\omega)$. Queste rappresentazioni sono però "statiche"

rappresentare \downarrow il pss nelle sue integrità

Per risolvere il problema della predizione, è necessario avere un rappresentazione dinamica, che mette in luce come il futuro dipende dal passato

\downarrow
In che modo è possibile esprimere un pss in forma dinamica?

66

Consideriamo un rumore bianco $e(t) \sim \text{WN}(0, \sigma^2)$. Osserviamo che $e(t)$ ha uno spettro costante

Un pss $v(t)$ può quindi essere rappresentato pesando opportunamente differenti campioni di $e(t)$ ad istanti diversi di tempo

$$v(t) = w_0 e(t) + w_1 e(t-1) + w_2 e(t-2) + \dots$$

Inoltre:

$$v(t-1) = w_0 e(t-1) + w_1 e(t-2) + \dots$$

$$v(t-2) = w_0 e(t-2) + w_1 e(t-3) + \dots$$

$$\gamma(0) = E[v(t)^2] = (w_0^2 + w_1^2 + w_2^2 + \dots) \sigma^2$$

$$\gamma(1) = E[v(t)v(t-1)] = E[w_0 w_1 e(t-1)^2 + w_1 w_2 e(t-2)^2 + \dots] = (w_0 w_1 + w_1 w_2 + \dots) \sigma^2$$

$$\gamma(2) = E[v(t)v(t-2)] = E[w_0 w_2 e(t-1)^2 + w_1 w_3 e(t-2)^2 + \dots] = (w_0 w_2 + w_1 w_3 + \dots) \sigma^2$$

Quindi, combinando i pesi w_i , posso ottenere funzioni di covarianza (e quindi densità spettrali di potenza) arbitrarie

quindi posso ottenere pss arbitrarie (infatti le $f(x)$ sono state ottenute usando le proprietà del $e(t)$ che è un pss)

$v(t)$ è STAZIONARIO
perché caratteristiche fissate di:

$$v(t) = w_0 e(t) + w_1 e(t-1) + \dots$$

$$= \sum_{i=0}^{+\infty} w_i e(t-i)$$

forse di convoluzione

risposta di un sistema dinamico casuale AS-STAB

ad un impulso $e(t)$

è il movimento forzato,
(il movimento libero è o essendo ds.)

Il sistema dinamico ha w_i come risposta impulsiva e:

$$W(z) = \sum_{i=0}^{+\infty} w_i z^{-i}$$

$$\Rightarrow V(t) = W(z) e(t)$$

come funzione di trasferimento

$e(t)$ viene filtrato da $W(z)$ costante

$W(z)$ fa scomparsa dello spettro di $e(t)$ da e

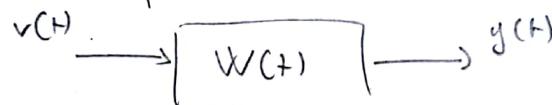
47

Studieremo il caso in cui $W(z)$ sia un filtre razionale fissa, ovvero $W(z) = \frac{C(z)}{A(z)}$ (filter digitale)

Il pss che si ottengono filtrando un rumore bianco tramite un filtre AS. STAB sui detti processi è spettro razionale (razionale fissa).

Teorema

Dato un processo stocastico $y(t)$, uscita di regime di un filtro $W(z)$, alimentato da un processo stocastico $v(t)$



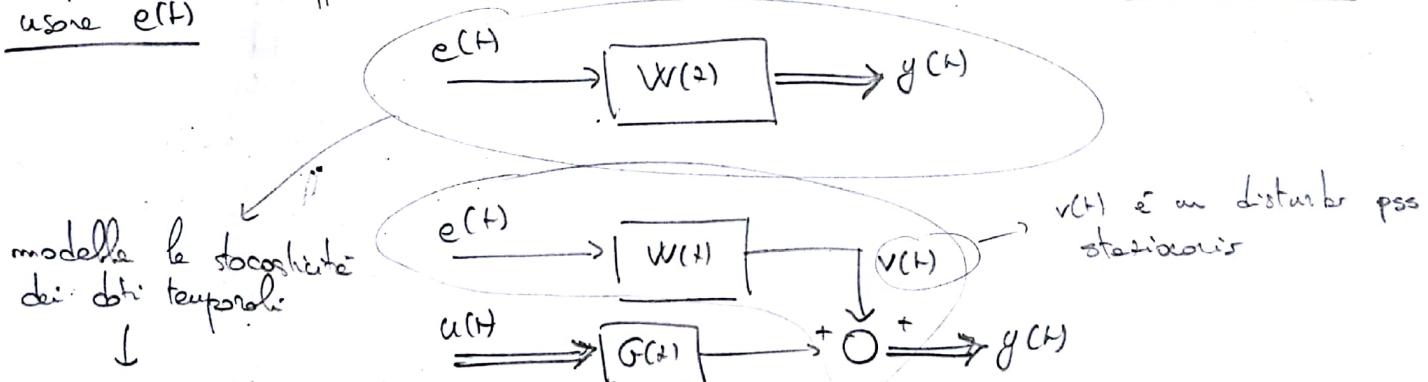
Condizione necessaria e sufficiente offinale, & condizione iniziale, a regime $y(t)$ sia un pss è che:

1) $v(t)$ sia pss

2) $W(z)$ sia AS. STAB \Rightarrow se $W(z) = \frac{C(z)}{A(z)} \Rightarrow$ radici di $A(z) | | < 1$

Ovvero, l'uscita di regime di un filtro esistenziale stabile, alimentata da un pss, è un pss

Riprendendo l'approccio iniziale di modellazione, abbiamo che la serie usore $e(t)$



può essere usato anche per modelli I/O, nel qual caso modellare anche errori di modello $G(z)$

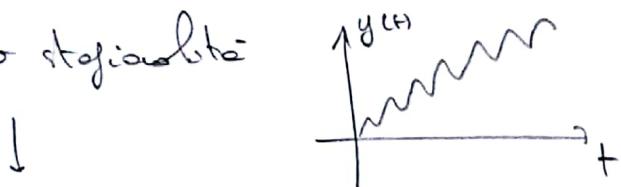
48

Nota

$G(z)$ è un sistema fisico, reale. $W(z)$ ed $e(t)$ non esistono: sono solo un metodo per modellare aff che $G(z)$ non riesce a modellare le stocasticità della serie di dati.

Note

Nelle modellizzazioni di serie stocastiche $\xrightarrow{\text{OCH}} \boxed{W(t)} \Rightarrow y(t)$ ci potrebbe essere trend o stazionalità



Bisogna quindi prime rimuovere il trend che stazionalità per ottenere un processo stazionario



Un'altra operazione è rimuovere le medie, in modo da semplificare il calcolo dei metodi di identificazione dei modelli

DEPOLARIZZAZIONE

La depolarizzazione permette di semplificare il calcolo di $\gamma(\tau)$ nel caso in cui un processo stocastico $\overset{\text{stat.}}{\sim} v(t)$ abbia media $m_v \neq 0$

$$\gamma(\tau) = E \left[(v(t) - m_v)(v(t+\tau) - m_v) \right]$$

Se ovessimo $m_v = 0$

$$\gamma(\tau) = E \left[v(t)v(t+\tau) \right]$$

Definiamo quindi $\tilde{v}(t) = v(t) - m_v$

$$- E[\tilde{v}(t)] = E[v(t) - m_v] = E[v(t)] - m_v = m_v - m_v = 0$$

$$- \tilde{\gamma}(\tau) = E \left[\tilde{v}(t)\tilde{v}(t+\tau) \right] = E \left[(v(t) - m_v)(v(t+\tau) - m_v) \right] = \gamma(\tau)$$

Quindi $v(t)$ e $\tilde{v}(t)$ hanno la stessa funzione di covariance (e stesse caratteristiche spettrali)



non si deve mettere generalità nello studiare p.s. e media nulla

FAMIGLIE DI MODELLI A SPECTRO RAZIONALE

MODELLI PER SERIE TEMPORALI

PROCESSI MA (Moving Average)

Un processo $y(t)$, generato a partire dal rumore bianco $e(t)$, è detto di tipo MA(m) se:

$$y(t) = c_0 e(t) + c_1 e(t-1) + c_2 e(t-2) + \dots + c_m e(t-m) = \sum_{i=0}^m c_i e(t-i)$$

- c_0, c_1, \dots, c_m : COEFFICIENTI DEL MODELLO MA

- m : ORDINE DEL MODELLO

- MA(m): IL MODELLO MA

L'uscita di un MA(m) è la combinazione lineare degli ultimi $m+1$ valori del segnale im impulso $e(t)$

Ricordando che $z^{-1}x(t) = x(t-1)$, possiamo scrivere $y(t)$ come:

$$\begin{aligned} y(t) &= c_0 e(t) + c_1 e(t) z^{-1} + c_2 e(t) z^{-2} + \dots + c_m e(t) z^{-m} \\ &= (c_0 + c_1 z^{-1} + c_2 z^{-2} + \dots + c_m z^{-m}) \cdot e(t) = \boxed{C(z) e(t)} \\ \Rightarrow \frac{y(t)}{e(t)} &= \frac{z^m c_0 + z^{m-1} c_1 + \dots + c_m}{z^m} \end{aligned}$$

$$\text{m poli in } z=0 \quad \leftarrow \qquad \qquad \qquad \overset{e(t)}{\rightarrow} \boxed{C(z)} \longrightarrow y(t)$$

I processi MA sono sempre stazionari

Calcolo dei parametri caratteristici

• Valore atteso

$$\begin{aligned} m(t) &= E[y(t)] = E[c_0 e(t) + c_1 e(t-1) + \dots + c_m e(t-m)] = c_0 E[e(t)] + c_1 E[e(t)] \\ &\quad + \dots + c_m E[e(t-m)] \\ &= c_0 \mu + c_1 \mu + \dots + c_m \mu \\ &= \boxed{\mu \sum_{i=0}^m c_i} \end{aligned}$$

$$\Rightarrow \text{se } e(t) \sim wN(0, \sigma^2) \Rightarrow \boxed{E[y(t)] = 0}$$

• $\gamma(\tau)$, supponiamo $E[y(t)] = 0$ per definizione

$$\begin{aligned}
 -\gamma(0) \cdot E[(v(t) - m(t))^2] &= E[v(t)^2] = E[(c_0 e^{ct} + c_1 e^{ct-1} + \dots + c_m e^{ct-m})^2] = \\
 &= E[\underbrace{c_0^2 e^{ct^2} + c_1^2 e^{ct-1^2} + \dots + c_m^2 e^{ct-m^2}}_{\text{quadrati}} + \underbrace{2c_0c_1 e^{ct}e^{ct-1} + \dots +}_{2c_{m-1}c_m e^{ct-m}e^{ct-m}}] \\
 &= c_0^2 E[e^{ct^2}] + c_1^2 E[e^{ct-1^2}] + \dots + c_m^2 E[e^{ct-m^2}] \\
 &= c_0^2 \gamma_E(0) + c_1^2 \gamma_E(0) + \dots + c_m^2 \gamma_E(0) = \boxed{\gamma^2 \cdot \sum_{i=0}^m c_i^2}
 \end{aligned}$$

$$\begin{aligned}
 -\gamma(1) &= E[(v(t) - m(t))(v(t-1) - m(t-1))] = E[v(t)v(t-1)] = \\
 &= E[(c_0 e^{ct} + c_1 e^{ct-1} + \dots + c_m e^{ct-m})(c_0 e^{ct-1} + c_1 e^{ct-2} + \dots + c_{m-1} e^{ct-m-1})] \\
 &= \underbrace{c_0 c_1 E[e^{ct-1^2}] + c_1 c_2 E[e^{ct-2^2}] + \dots + c_{m-1} c_m E[e^{ct-m^2}]}_{\gamma^2 \cdot (c_0 c_1 + c_1 c_2 + \dots + c_{m-1} c_m)} \\
 &= \boxed{\gamma^2 \cdot (c_0 c_1 + c_1 c_2 + \dots + c_{m-1} c_m)}
 \end{aligned}$$

$$-\gamma(2) = \gamma^2 (c_0 c_1 + c_1 c_2 + \dots + c_{m-2} c_m)$$

$$-\gamma(m) = \gamma^2 (c_0 c_m)$$

$$-\gamma(\tau) \text{ t.c. } \tau > m \Rightarrow \boxed{\gamma(\tau) = 0}$$

Un processo MA(m) dipende solo dagli m valori primi. Soprattutto imprevedibile (non borsellabile)

In modo analogo se una serie temporale è MA è quello di grandezza se le sue $\gamma(\tau)$ va a zero dopo un certo τ

Note

Il processo $\tilde{y}(t) = \tilde{c}_0 \eta(t) + \tilde{c}_1 \eta(t-1) + \dots + \tilde{c}_m \eta(t-m)$ con $\tilde{c}_i = \alpha \cdot c_i$:
 $\eta(t) \sim WN(0, \tilde{\sigma}^2)$ $\tilde{\sigma}^2 = \frac{\sigma^2}{\alpha^2}$

Le stesse cose si fanno al contrario del processo

$y(t) = c_0 e^{ct} + c_1 e^{ct-1} + \dots + c_m e^{ct-m}$ $e(t) \sim WN(0, \sigma^2)$. Per avere questa ^{sotto-}parametrizzazione, si pone di solito $c_0 = 1$

(51)

PROCESSI AR (Autoregressive)

Un processo $y(t)$, generato a partire da $e(t) \sim WN(\mu, \sigma^2)$, è detto di tipo AR(m) se:

$$y(t) = \alpha_1 y(t-1) + \alpha_2 y(t-2) + \dots + \alpha_m y(t-m) + e(t) = \sum_{i=1}^m \alpha_i y(t-i) + e(t)$$

- $\alpha_1, \alpha_2, \dots, \alpha_m$: COEFFICIENTI DEL MODELLO AR
- m: ORDINE DEL MODELLO

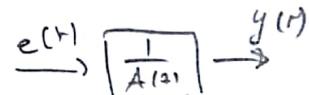
L'uscita di un AR(m) è la combinazione lineare degli ultimi m "vecchi" valori del processo stesso, più l'infusso $e(t)$ all'istante istante.

Forma operatoriale:

$$y(t) = \alpha_1 y(t-1) + \alpha_2 y(t-2) + \dots + \alpha_m y(t-m) + e(t)$$

$$y(t) = \alpha_1 y(t) z^{-1} + \alpha_2 y(t) z^{-2} + \dots + \alpha_m y(t) z^{-m} + e(t)$$

$$y(t) \left[1 - \alpha_1 z^{-1} - \alpha_2 z^{-2} - \dots - \alpha_m z^{-m} \right] = e(t)$$



$$\frac{y(t)}{e(t)} = \frac{1}{1 - \alpha_1 z^{-1} - \alpha_2 z^{-2} - \dots - \alpha_m z^{-m}} = \frac{1}{A(z)} \Rightarrow \boxed{y(t) = \frac{1}{A(z)} e(t)}$$

$$\Rightarrow \frac{y(t)}{e(t)} = \frac{z^m}{z^m - \alpha_1 z^{m-1} - \alpha_2 z^{m-2} - \dots - \alpha_m}$$

- m ZERI NELL'ORIGINE

- m POLI: quindi non è sempre stabile

è stabile se tutti i poli sono in modulo < 1

bisogna fare attenzione alle stime dei parametri

Globi parametri caratteristici (nel caso in cui AR(m) è pss)

• Valore atteso

$$m = E[y(t)] = E[\alpha_1 y(t-1) + \alpha_2 y(t-2) + \dots + \alpha_m y(t-m) + e(t)]$$

$$= \alpha_1 E[y(t-1)] + \alpha_2 E[y(t-2)] + \dots + \alpha_m E[y(t-m)] + E[e(t)]$$

$$E[y(t)] = (\alpha_1 + \alpha_2 + \dots + \alpha_m) E[y(t)] + \mu$$

$$(1 - \alpha_1 - \alpha_2 - \dots - \alpha_m) E[y(t)] = \mu$$

$$\Rightarrow \boxed{E[y(t)] = \frac{\mu}{1 - \alpha_1 - \alpha_2 - \dots - \alpha_m}}$$

- Se $E[e(t)] = 0$
 $\Rightarrow E[y(t)] = \mu$

52

- $\gamma(n) \rightarrow$ calcolo per processo AR(1) perché è complesso

Dato il processo AR(1): $y(t) = \alpha y(t-1) + e(t)$ $e(t) \sim \text{wnn}(\mu, \sigma^2)$

$$\Rightarrow y(t)[1 - \alpha z^{-1}] = e(t) \Rightarrow \boxed{y(t) \sim \frac{1}{1 - \alpha z^{-1}} e(t)}$$

$y(t)$ è stazionario se il polo ha modulo < 1

$$1 - \alpha z^{-1} = 0 \Rightarrow z - \alpha = 0 \Rightarrow \text{polo in } z = \alpha$$

$A(z)$ AS. STAB. se $|\alpha| < 1$

Supponiamo $A(z)$ AS. STAB e che il processo $y(t)$ sia deponzitario (media nulla)

$$\begin{aligned} \bullet \gamma(0) &= E[y(t)^2] = E[(\alpha y(t-1) + e(t))^2] = E[\alpha^2 y(t-1)^2 + e(t)^2 + 2\alpha y(t-1)e(t)] \\ &= \alpha^2 E[y(t-1)^2] + E[e(t)^2] + 2\alpha E[y(t-1)e(t)] \\ &= \alpha^2 \gamma(0) + \sigma^2 + 0 \end{aligned}$$

$$\Rightarrow \gamma(0) = \alpha^2 \gamma(0) + \sigma^2 \Rightarrow \boxed{\gamma(0) = \frac{\sigma^2}{1 - \alpha^2}}$$

$y(t-1)$ è incostante con $e(t)$
perché dipende da $e(t-1)$ e da
 $y(t-2)$ → per ricorsione anche
 $y(t-2)$ è $\perp \alpha e(t)$ e
così via

$$\begin{aligned} \bullet \gamma(1) &= E[y(t)y(t-1)] = E[(\alpha y(t-1) + e(t)) \cdot y(t-1)] = E[\alpha y(t-1)^2 + y(t-1)e(t)] \\ &= \alpha E[y(t-1)^2] + E[y(t-1)e(t)] = \alpha \gamma(0) \Rightarrow \boxed{\gamma(1) = \alpha \cdot \gamma(0)} \end{aligned}$$

$$\begin{aligned} \bullet \gamma(2) &= E[y(t)y(t-2)] = E[(\alpha y(t-1) + e(t)) y(t-2)] = E[\alpha y(t-1)y(t-2) + y(t-2)e(t)] \\ &= \alpha E[y(t-1)y(t-2)] + E[y(t-2)e(t)] = \alpha \gamma(1) \Rightarrow \boxed{\gamma(2) = \alpha \cdot \gamma(1)} \end{aligned}$$



Generalità:

$$\left\{ \begin{array}{l} f(0) = \frac{\alpha^2}{1-\alpha^2} \\ f(n) = \alpha \cdot f(n-1) \quad n > 0 \end{array} \right.$$

EQUAZIONI DI YULE-WALKER
PER UN AR(1)

esistono anche per un AR(m)

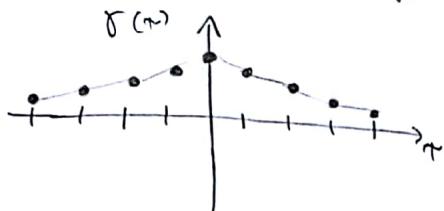
Osservazione

Dato che $|\alpha| < 1$ (obbligo supposto stazionario per piani colabbi $f(n)$) si ha

$$|r(n+1)| < |\delta(n)|$$

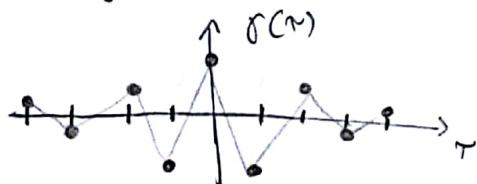
e dato che $\alpha \neq \pm 1$, $f(0)$ esiste finito. Inoltre si ha che:

- Il processo $y(t) = \alpha y(t-1) + e(t)$ con $0 < \alpha < 1$ ha $f(n) > 0 \forall n$, e sono decrescenti da un raffigurazione mai b. 0.



Le realizzazioni del pss $y(t)$ variano lentamente perché i dati sono molto correlati fra loro e la componenti di segnale (in media). Ci si aspetta una realizzazione con componenti a basse frequenze

- Il processo $y(t) = \alpha y(t-1) + e(t)$ con $-1 < \alpha < 0$ ha una funzione $f(n)$ che curva segno ad ogni n e decrescente in valore assoluto



Le realizzazioni del pss $y(t)$ assumono quindi segni (in media) cambia di segno ripetutamente, creando segnali con dei comportamenti in alte frequenze

Approfondimenti

Osserviamo visto che, per un modello MA(m), $\gamma(r) = 0$ per $r > m$. Per gli AR(m) possiamo ottenere un comportamento simile con la FUNZIONE DI AUTOCORRELAZIONE PARZIALE (PACF) $\gamma_{\text{PAR}}(r)$, che si annulla (nel caso di un AR(m)) per $r > m$.

Nell'analisi delle serie temporali in pratica, si fanno questi passaggi:

1) Controlla se c'è un MA(m) plotando $\gamma(r)$

2) $\sim \sim \sim$ AR(m) $\sim \sim \sim \gamma_{\text{PAR}}(r)$

3) Se nessuna delle due si annulla, mi servono altri modelli.

Un altro tipo di modelli per risolvere il punto 3) è il seguente:

MODELLO ARMA (AutoRegressive Moving Average)

Un processo $y(t)$, generato a partire da un rumore bianco $e(t) \sim \text{WN}(0, \sigma^2)$, è detto di tipo ARMA(m, n) se:

$$y(t) = \alpha_1 y(t-1) + \alpha_2 y(t-2) + \dots + \alpha_m y(t-m) + \text{PARTE AR}(n) \\ + e(t) + c_1 e(t-1) + \dots + c_n e(t-n) \quad \text{PARTE MA}(n)$$

- m : ~~ordine~~ L'ORDINE DEL MODELLO AR

- $\alpha_1, \alpha_2, \dots, \alpha_m$: COEFFICIENTI DEL TODELLO AR

- n : ~~ordine~~ L'ORDINE DEL PROCESSO MA

- c_1, c_2, \dots, c_n : COEFFICIENTI DEL TODELLO MA

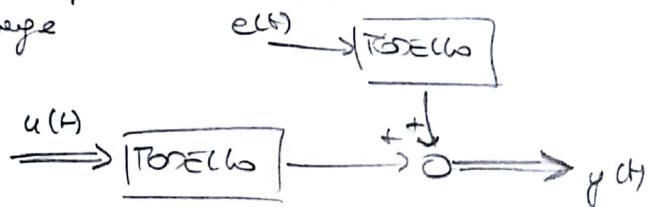
Notiamo che $\text{ARMA}(0, m) = \text{MA}(m)$ e che $\text{ARTA}(m, 0) = \text{AR}(m)$. Possendo in forma spettrale:

$$y(t) \left[1 - \alpha_1 z^{-1} - \alpha_2 z^{-2} - \dots - \alpha_m z^{-m} \right] = \left(c_1 z^{-1} + c_2 z^{-2} + \dots + c_n z^{-n} \right) e(t) \\ \Rightarrow y(t) = \frac{1 + c_1 z^{-1} + c_2 z^{-2} + \dots + c_n z^{-n}}{1 - \alpha_1 z^{-1} - \alpha_2 z^{-2} - \dots - \alpha_m z^{-m}} e(t) = \frac{\frac{c}{A}(t)}{A(t)}$$

- $\frac{C(z)}{A(z)}$ è stabile se $A(z)$ ha radici $|z| < 1$

MODELLO PER SISTEMI INPUT / OUTPUT z

ARMAX (Autoregressive Moving Average exogenous)



Un processo $y(t)$, pensato a

fatilità da un rumore bianco $e(t) \sim WN(\mu, \sigma^2)$ e da un ingresso esogeno $u(t)$ (independente del processo), è detto ARMAX $(m, m, n+p)$ se:

$$\begin{aligned} y(t) = & a_1 y(t-1) + a_2 y(t-2) + \dots + a_m y(t-m) + \text{PARTE AR}(m) \\ & + e(t) + c_1 e(t-1) + \dots + c_m e(t-m) + \text{PARTE MA}(m) \\ & + b_0 u(t-n) + b_1 u(t-n-1) + \dots + b_p u(t-n-p) \quad \text{PARTE X}(n+p) \end{aligned}$$

- m : ORDINE PROCESSO AR

- a_1, a_2, \dots, a_m : COEFFICIENTI PROCESSO AR

- n : ORDINE PROCESSO MA

- c_1, c_2, \dots, c_m : COEFFICIENTI PROCESSO MA

- p : ORDINE VARIABILE ES-GENA

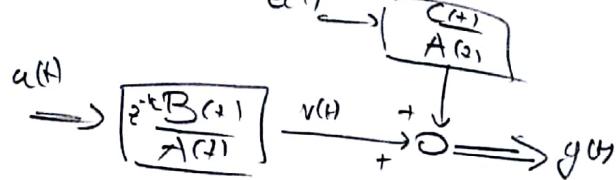
- $p+1$: COEFFICIENTI VARIABILE ES-GENA

- n : RITARDO PURO TRA INGRESSO ED OSCILTA

$$y(t) \left[1 - a_1 z^{-1} - a_2 z^{-2} - \dots - a_m z^{-m} \right] = e(t) \left[1 + c_1 z^{-1} + c_2 z^{-2} + \dots + c_m z^{-m} \right] + u(t-n) \left[b_0 + b_1 z^{-1} + b_2 z^{-2} + \dots + b_p z^{-p} \right]$$

$$y(t) A(z) = e(t) C(z) + u(t-n) B(z)$$

$$\boxed{y(t) = \frac{C(z)}{A(z)} e(t) + \frac{B(z)}{A(z)} u(t-n)} = \boxed{\frac{C(z)}{A(z)} e(t) + \frac{B(z)}{A(z)} z^{-n} u(t)}$$



(56)

Osservazione

Dato che l'ingresso $u(t)$ influenza il processo, non si può dire se $y(t)$ sia stazionario o meno. $y(t)$ è stazionario se:

- le radici di $A(z)$ sono all'interno della circonferenza di raggio unitario
- $u(t)$ è stazionario (quindi costante, dato che non è un processo stocastico)

In generale il processo ARTAX non è stazionario, ma la sua componente stocastica (ARTA) deve esserlo. La fonte di non stazionarietà è una componente deterministica (nota), quindi la posso modellare

In generale, le componenti di un' stazionarietà o le tasse (es trend, stagionali) o le mode (nel senso di ingresso esprimere noto "si modella da solo" in questo modo)

Es

Dire se il seguente processo è stazionario e calcolarne media e funzione di covarianza

$$y(t) = \frac{1}{3} y(t-1) + e(t) + 2 \quad e(t) \sim \text{wnn}(z, z)$$

$$y(t) \left[1 - \frac{1}{3} z^{-1} \right] = e(t) + 2 \Rightarrow y(t) = \frac{1}{1 - \frac{1}{3} z^{-1}} e(t) + \frac{2}{1 - \frac{1}{3} z^{-1}} \cdot 2 \rightarrow u(t)$$

Poli di $A(z)$

$$\frac{1 - \frac{1}{3} z^{-1}}{z} = 0 \Rightarrow z^{-1} = 3 \Rightarrow z = \frac{1}{3} < 1 \Rightarrow -A(z) \text{ AS. STAB}$$

$$-u(t) \text{ STAZIONARIO} \Rightarrow$$

$y(t)$ è PSS.

Quel è il contributo di $u(t)$ sull'uscita $y(t)$? Teorema della risposta in frequenza: dato un ingresso sinusoidale $u(t) = a \cdot \cos(ut + \phi)$ la uscita

$$w(t) = |G(e^{j\omega})| \cdot a \cdot \cos(ut + \phi + \angle G(e^{j\omega})) \quad G(z) = \frac{B(z)}{A(z)}$$

Dato che $u(t)$ costante, la frequenza $\omega = 0$, quindi:

$$w(t) = |G(e^{j0})| \cdot a = \left| \frac{1}{1 - \frac{1}{3} z^0} \right| \cdot 2 = \frac{1}{1 - \frac{1}{3}} \cdot 2 = \frac{1}{\frac{2}{3}} \cdot 2 = \frac{3}{2} \cdot 2 = 3$$

L'effetto di $u(t)$ è
sostituire la media
del PSS

(57)

$$- E[y] = my = E\left[\frac{1}{3}y(t-1) + e(t) + z\right] = \frac{1}{3}my + 1 + z \Rightarrow \left(1 - \frac{1}{3}\right)my = 3$$

$$\Rightarrow \frac{2}{3}my = 3 \Rightarrow \boxed{my = \frac{9}{2}}$$

- $\tilde{y}(t) \Rightarrow$ Depolarizzare $y(t)$ ed $e(t)$.

Sarà il processo come:

$$y(t) = \frac{1}{3}y(t-1) + e(t) + z$$

$$\tilde{y}(t) + \frac{z}{2} = \frac{1}{3}\left[\tilde{y}(t-1) + \frac{z}{2}\right] + \tilde{e}(t) + 1 + z$$

$$\tilde{y}(t) = \frac{1}{3}\tilde{y}(t-1) + \underbrace{\frac{8}{6} - \frac{z}{2}}_{\text{media nulla}} + 3 + \tilde{e}(t)$$

$$\boxed{\tilde{y}(t) = \frac{1}{3}\tilde{y}(t-1) + \tilde{e}(t)}$$

$\xrightarrow{\text{AR(1) è media nulla}}$

$$\frac{3 - 3 + 6}{2} = 0$$

$$\tilde{f}^{(\tau)} = f(\tau) = \left(\frac{1}{3}\right)^{\tau} \cdot \frac{1}{1 - \frac{1}{3}} = \frac{3}{8} \cdot \frac{1}{3^{\tau}} = \boxed{\frac{3^{2-\tau}}{8}}$$

YULE WALKER

$$\alpha^{\tau} \cdot \frac{\alpha^2}{1 - \alpha^2} = \alpha^{\tau} \cdot f(0)$$

Teorema

Dato un processo stocastico stazionario ARMA(m, n), esso può essere scritto come MA(∞)

Ese AR(1)

$$y(t) = \alpha y(t-1) + e(t) \quad e(t) \sim \mathcal{WN}(0, \sigma^2)$$

$$y(t) = \frac{1}{1 - \alpha z^{-1}} e(t) \quad \text{limite serie geometrica di ragione } \alpha z^{-1}$$

$$= \sum_{k=0}^{+\infty} (\alpha z^{-1})^k \cdot e(t) = \sum_{k=0}^{+\infty} \alpha^k \cdot e(t-k) \quad \text{MA}(\infty)$$

CALCOLO DELLO SPETTRO DI UN PSS A PARTIRE DAL SUO MODELLO

Note: useremo i termini DENSITÀ SPECTRALE DI POTENZA e SPETTRO in modo equivalente

Se il pss $y(t)$ è rappresentabile come uscita di refime di un filtro AS. STABILE dimentato da un pss $v(t)$:

$$y(t) = F(z) v(t)$$



è possibile calcolare lo spettro di $y(t)$ come:

$$\boxed{\bar{P}_y(\omega) = |F(e^{j\omega})|^2 \cdot \bar{P}_v(\omega)}$$

- $\bar{P}_y(\omega)$: SPETTRO DELL' USCITA
- $|F(e^{j\omega})|^2$: MODULO AL QUADRATO DELLA RISPOSTA IN FREQUENZA DEL FILTRO
- $\bar{P}_v(\omega)$: SPETTRO DELL' INGRESSO

Se l'ingresso è un white noise $e(t) \sim \text{wn}(0, \sigma^2)$: $\boxed{\bar{P}_y(\omega) = |F(e^{j\omega})|^2 \cdot \sigma^2}$

Ese

Consideriamo un processo MA(1), calcolare $\bar{P}_y(\omega)$:

$$y(t) = e(t) + c e(t-1) \quad e(t) \sim \text{wn}(0, 1)$$

1) Usando la definizione

$$\begin{aligned} \bar{P}_y(\omega) &= \sum_{r=-\infty}^{+\infty} \bar{P}(r) e^{-j\omega r} \rightarrow \sigma^2 \sum_{i=0}^m c_i^2 \\ \bar{P}_e(\omega) &= \bar{P}(0) = 1 \\ &= (\bar{P}(-1) e^{-j\omega(-1)} + \bar{P}(0) e^{-j\omega 0} + \bar{P}(1) e^{-j\omega 1}) \\ &= 1^2 (1 + c) e^{j\omega} + (1^2 + c^2) \cdot 1 + c e^{-j\omega} = c \left[e^{j\omega} + e^{-j\omega} \right] + c^2 + 1 \\ &= 2c \cos \omega + c^2 + 1 \end{aligned}$$

2) Usando il teorema

$$y(t) = (1 + c z^{-1}) e(t) = C(z) e(t)$$

- $C(z)$ AS. STAB (poli nell'origine)
- $e(t)$ pss $\Rightarrow y(t)$ pss

$$\begin{aligned} \bar{P}_y &= |C(e^{j\omega})|^2 \cdot \bar{P}_e(\omega) = \left| 1 + c e^{-j\omega} \right|^2 \cdot 1 = (1 + c e^{-j\omega})(1 + c e^{j\omega}) \\ &= 1 + c^2 (e^{j\omega} \cdot e^{-j\omega}) + c (e^{j\omega} + e^{-j\omega}) = \boxed{1 + c^2 + 2c \cos \omega} \end{aligned}$$

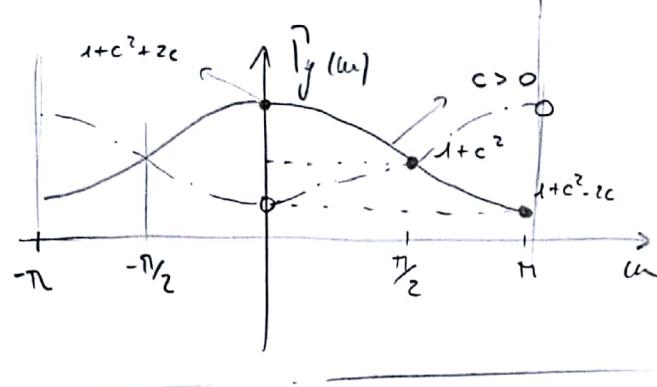
(58)

Tracciare lo spettro per punti

$$\tilde{P}_y(0) = 1 + c^2 + 2c \cos(0) = 1 + 2c + c^2 = (1+c)^2$$

$$\tilde{P}_y(\pi/2) = 1 + c^2 + 2c \cdot \cos\left(\frac{\pi}{2}\right) = 1 + c^2$$

$$\tilde{P}_y(\pi) = 1 + c^2 + 2c \cdot \cos(\pi) = (1-c)^2$$



& PREDIZIONE &

Problema: data una sequenza di dati (ad esempio serie temporale)

$$\{y^{(1)}, y^{(2)}, \dots, y^{(N)}\}$$

Vogliamo identificare un modello ARMA

$$y^{(t)} = \frac{C^{(+)}}{A^{(+)}} e^{(t)} \quad e^{(t)} \sim WN(0, \sigma^2)$$

Per stimare le incognite (coefficienti di $C^{(+)}$ ed $A^{(+)}$, eventualmente anche σ^2) seguiremo questi passi:

- Calcolare il PREDIOTTO DEL MODELLO DATI DATI $\hat{y}(t|t-1) \rightarrow$ dipende delle incognite θ

- Minimizzare $J(\theta) = \sum_{i=1}^N (y^{(i)} - \hat{y}(t|t-1, \theta))^2$ VARIANZA CAMPIONARIA ERRORE DI PREDIOTTO

Approssimazione: un modello è buono se è capace di predire in
processo

FILTO PASSA-TUTTO

È un filtro di ordine 2 con le seguenti forme:

$$T(z) = \frac{1}{\alpha} \cdot \left(\frac{z+\alpha}{z+\frac{1}{\alpha}} \right) \quad \alpha \neq 0, \alpha \in \mathbb{R}$$

→ la zera è opposta al polo

Ricordando il teorema delle fattorizzazioni spettrale, si ha che:

$$\Gamma_y(u) = |T(e^{ju})|^2 \cdot \Gamma_e(u)$$

$$\begin{aligned} -|T(e^{ju})|^2 &= \left(\frac{1}{\alpha} \cdot \frac{e^{ju} + \alpha}{e^{ju} + \frac{1}{\alpha}} \right) \left(\frac{1}{\alpha} \cdot \frac{e^{-ju} + \alpha}{e^{-ju} + \frac{1}{\alpha}} \right) \\ &= \frac{1}{\alpha^2} \frac{(e^{ju} + \alpha)(e^{-ju} + \alpha)}{\left(e^{ju} + \frac{1}{\alpha} \right) \left(e^{-ju} + \frac{1}{\alpha} \right)} = \frac{1}{\alpha^2} \cdot \frac{1 + \alpha^2 + 2\alpha(e^{ju} + e^{-ju})}{1 + \frac{1}{\alpha^2} + \frac{1}{\alpha}(e^{ju} - e^{-ju})} \\ &= \frac{1}{\alpha^2} \cdot \frac{1 + \alpha^2 + 2\alpha \cos u}{\frac{\alpha^2 + 1 + 2\alpha \cos u}{\alpha^2}} = 1 \end{aligned}$$

Quindi:

$$\Gamma_y(u) = |T(e^{ju})|^2 \cdot \Gamma_e(u) = \Gamma_e(u)$$

Il filtro passatutto non distorce lo spettro del segnale da lui alimentato

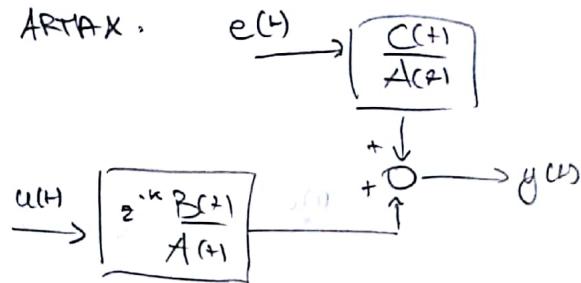


Il segnale in ingresso e quello in uscita di filtro passatutto sono EQUIVALENTI

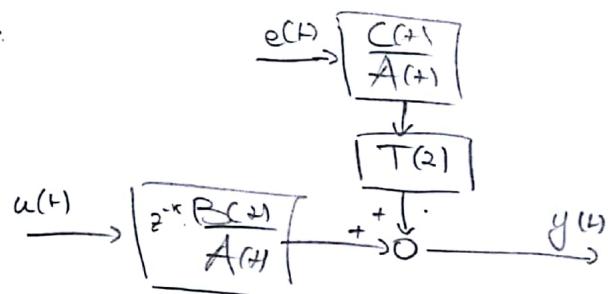
Note

Il filtro passatutto non modifica il modulo ma introduce una distorsione di fase, ritardando il segnale in ingresso. L'ingresso e uscita NON SONO IDENTICI, ma sono SPECTRALMENTE EQUIVALENTI

Considerando un ARMAX:



questo risulterà equivalente a:



Osservazione

Il filtro passatutto è un "oggetto matematico". Non posso inserire tra $u(t)$ e $y(t)$, oltremodo cambierei le relazioni ingresso/uscita, che è data da un "oggetto fisico", reale.

FORMA CANONICA

Con l'introduzione del filtro passatutto $T(z)$, obiettivo visto che il processo $y(t) = \frac{C(z)}{A(z)} e(t)$ ed il processo $y(t) = \frac{C(z)}{A(z)} T(z) e(t)$ siano EQUIVALENTI dal punto di vista spettrale

esistono altre rappresentazioni equivalenti?

Consideriamo questi 5 processi ARMA:

$$1) y_1(t) = \frac{z + \frac{1}{2}}{z - \frac{1}{3}} e(t) \quad e(t) \sim \text{wn}(0, 1)$$

$$4) y_4(t) = \frac{2z + 1}{z - \frac{1}{3}} e(t) \quad e(t) \sim \text{wn}(0, \frac{1}{9})$$

$$2) y_2(t) = \frac{z + \frac{1}{2}}{z - \frac{1}{3}} e(t-1) \quad e(t) \sim \text{wn}(0, 1)$$

$$5) y_5(t) = \frac{z + 2}{z - \frac{1}{3}} e(t) \quad e(t) \sim \text{wn}(0, \frac{1}{4})$$

$$3) y_3(t) = \frac{z^2 - \frac{1}{4}}{z^2 + \frac{1}{6} - \frac{5}{6}z} e(t) \quad e(t) \sim \text{wn}(0, 1)$$

Si osserva che $m_{y_1} = m_{y_2} = \dots = m_{y_5} = 0$

Calcoliamo gli spettri dei processi:

$$1) \hat{P}_{y_1}(\omega) = \left| \frac{z + \frac{1}{2}}{z - \frac{1}{3}} \right|^2 \cdot \hat{P}_e(\omega) = \left| \frac{e^{j\omega} + \frac{1}{2}}{e^{j\omega} - \frac{1}{3}} \right|^2 \cdot 1$$

$$2) \hat{P}_{y_2}(\omega) = \left| \frac{z + \frac{1}{2}}{z - \frac{1}{3}} \cdot z^{-2} \right|^2 \cdot \hat{P}_e(\omega) = \left| \frac{e^{j\omega} + \frac{1}{2}}{e^{j\omega} - \frac{1}{3}} \right|^2 \cdot (e^{-2j\omega})^2 \cdot 1 \\ = \hat{P}_{y_1}(\omega) \cdot (e^{-2j\omega}) \cdot (e^{2j\omega}) = \hat{P}_{y_1}(\omega)$$

$$3) \hat{P}_{y_3}(\omega) = \left| \frac{z^2 - \frac{1}{4}}{z^2 + \frac{1}{6} - \frac{5}{6}z} \right|^2 \cdot \hat{P}_e(\omega) = \left| \frac{\left(z - \frac{1}{2}\right)\left(z + \frac{1}{2}\right)}{\left(z - \frac{1}{2}\right)\left(z - \frac{1}{3}\right)} \right|^2 \cdot 1 = \hat{P}_{y_1}(\omega)$$

$$4) \hat{P}_{y_4}(\omega) = \left| \frac{2z + 1}{z - \frac{1}{3}} \right|^2 \cdot \hat{P}_e(\omega) = \left| 2 \cdot \frac{z + \frac{1}{2}}{z - \frac{1}{3}} \right|^2 \cdot \frac{1}{4} = 4 \cdot \left| \frac{e^{j\omega} + \frac{1}{2}}{e^{j\omega} - \frac{1}{3}} \right|^2 \cdot \frac{1}{4}$$

$$5) \hat{P}_{y_5}(\omega) = \left| \frac{z + 2}{z - \frac{1}{3}} \right|^2 \cdot \hat{P}_e(\omega) = \left| \frac{z + 2}{z - \frac{1}{3}} \cdot 2 \cdot \frac{z + \frac{1}{2}}{z + 2} \right|^2 \cdot \frac{1}{4} \xrightarrow{\text{PASSA TUTTO } T(2)} \\ = 4 \cdot \frac{1}{4} \left| \frac{e^{j\omega} + \frac{1}{2}}{e^{j\omega} - \frac{1}{3}} \right|^2 = \hat{P}_{y_1}(\omega)$$



Tutti e 5 i processi sono quindi equivalenti. Un processo ARMA ammette infinite rappresentazioni equivalenti. Le cause di univocità sono:

- ritardi puri (processo 2)
- fattori moltiplicativi che si cancellano (processo 3)
- coefficienti moltiplicativi su funzione di trasferimento e rumore si compensano (processo 4)
- poli/zeri "reciprocii" (processo 5)

TEOREMA DELLA FATTORIATTAZIONE SPECTRALE

Dato un processo a spettro razionale, esiste una ed una sola rappresentazione del processo come uscita di un sistema dinamico alimentato da un rumore bianco tale che:

- 1) $C(z)$ ed $A(z)$ hanno stesso grado (grado relativo nullo)
- 2) $C(z)$ ed $A(z)$ sono coprimi (non hanno fattori in comune)
- 3) $C(z)$ ed $A(z)$ sono monici (il coefficiente del termine di grado max è 1)
- 4) (A) ed $A(z)$ hanno radici interne al cerchio unitario (poli esterni $|z| < 1$)

Ese

$$y(t) = \frac{z+2}{z-\frac{1}{3}} e(t-2) \quad e(t) \sim \text{WN}(0, 1)$$

$$\begin{aligned} y(t) &= \frac{z+2}{z-\frac{1}{3}} \cdot \left(2 \cdot \frac{z+\frac{1}{2}}{z+2} \right) e(t-2) = \frac{z+\frac{1}{2}}{z-\frac{1}{3}} \cdot 2e(t-2) \quad \eta(t) \sim \text{WN}(0, 4) \\ \Rightarrow \boxed{y(t)} &= \frac{1 + \frac{1}{2} z^{-1}}{1 - \frac{1}{3} z^{-1}} \eta(t) \quad \eta(t) \sim \text{WN}(0, 4) \end{aligned}$$

IL PROBLEMA DELLA PREDIZIONE

Stimare il dato al tempo $t+k$ conoscendo i dati fino al tempo t . Indichiamo il predittore con le notazioni $\hat{y}(t+k|t)$, oppure $\hat{y}(t+k|t-k)$

Informazioni disponibili

- Dati $y(t), y(t-1), \dots, y(t-N)$
- Vecchie predizioni: $\hat{y}(t+k-1|t-1), \hat{y}(t+k-2|t-2), \dots$
- Modelli $\frac{C(z)}{A(z)}$

Vogliamo trovare il predittore ottimo dei dati. Esistono molti modi per calcolare il predittore dei dati:

Es 1

$$\hat{y}(t+1|t) = \frac{y(t) + y(t-1) + y(t-2)}{3} \quad \text{media valori passati}$$

Es 2

$$\hat{y}(t+1|t) = \frac{2y(t) + \frac{1}{2}y(t-1) + \frac{1}{2}y(t-2)}{3} \quad \begin{array}{l} \text{dai cui più pesa ai valori} \\ \text{più recenti} \end{array}$$

Potrà fare meglio? Una delle proprietà del predittore è che il predittore sia corretto, ovvero

$$E[\varepsilon(t)] = E[y(t) - \hat{y}(t|t-1)] = 0$$

errore di predizione

• PREDITTORE OTTIMO •

Un predittore è ottimo se:

- 1) $E[\hat{y}(t|t-k), \varepsilon(t)] = 0$, dove $\varepsilon(t) = y(t) - \hat{y}(t|t-k)$. Ovvero, predittore ed errore di predizione devono essere scostanti. tutta l'informazione è stata utilizzata dal predittore
- 2) $\text{Var}[\varepsilon(t)]$ MINIMA

Possiamo quindi scomporre $y(t)$ come: $y(t) = \hat{y}(t|t-k) + \varepsilon(t)$ (con:

- $\varepsilon(t)$ parte imprevedibile al tempo $t-k$
- $\hat{y}(t|t-k)$ parte del processo che è prevedibile al tempo $t-k$

• PREDITTORE VAD CON PASSO DI PROCESSI MA •

Supponiamo un processo MA(m) in forma canonica:

$$y(t) = e(t) + c_1 e(t-1) + c_2 e(t-2) + \dots + c_m e(t-m) \quad e(t) \sim \mathcal{N}(0, \sigma^2)$$

$$y(t) = \underbrace{e(t)}_{\text{parte imprevedibile}} + \underbrace{c_1 e(t-1) + \dots + c_m e(t-m)}_{\text{parte prevedibile al tempo } t-1}$$

Un possibile predittore potrebbe quindi essere:

$$\hat{y}(t|t-1) = c_1 e(t-1) + c_2 e(t-2) + \dots + c_m e(t-m)$$

Osserviamo che:

- $\hat{y}(t|t-1)$ dipende dal unico al tempo $t-1$ - È corretto: $E[\epsilon(t)] = 0$

- $E[\hat{y}(t|t-1) \epsilon(t)] = 0$, infatti $\epsilon(t) = y(t) - \hat{y}(t|t-1) = e(t)$

$$\Rightarrow E[\hat{y}(t|t-1) \epsilon(t)] = E[(c_1 e(t-1) + c_2 e(t-2) + \dots + c_m e(t-m)) \cdot \epsilon(t)] = 0$$

- Non è possibile trovare un predittore con $\text{Var}[\epsilon(t)]$ minore, infatti non possiamo avere un errore di predizione di $\text{Var}[\epsilon(t)]$

$$\boxed{\hat{y}(t|t-1) = c_1 e(t-1) + \dots + c_m e(t-m)} \text{ è ottimo}$$

Questo, tuttavia, è un predittore ottimo "dal remore". Vogliamo un predittore ottimo "di fatti"

$$y(t) = (1 + c_1 z^{-1} + c_2 z^{-2} + \dots + c_m z^{-m}) e(t) \Rightarrow e(t) = \frac{1}{1 + c_1 z^{-1} + c_2 z^{-2} + \dots + c_m z^{-m}} y(t)$$

$$\hat{y}(t|t-1) = (c_1 z^{-1} + c_2 z^{-2} + \dots + c_m z^{-m}) e(t)$$

$$= \frac{c_1 z^{-1} + c_2 z^{-2} + \dots + c_m z^{-m}}{1 + c_1 z^{-1} + c_2 z^{-2} + \dots + c_m z^{-m}} y(t) \Rightarrow$$

$$\hat{y}(t|t-1) \left\{ 1 + c_1 z^{-1} + \dots + c_m z^{-m} \right\} = (c_1 + c_2 z^{-1} + \dots + c_m z^{-m-1}) y(t-1)$$

si nota come $C(z)$
deve essere AS. STAR, quindi è necessaria la
fase causale

$$\Rightarrow \boxed{\hat{y}(t|t-1) = -c_1 \hat{y}(t-1|t-2) - c_2 \hat{y}(t-2|t-3) - \dots - c_m \hat{y}(t-m|t-m-1) + \text{vecchie predizioni}}$$

$$c_1 y(t-1) + c_2 y(t-2) + \dots + c_m y(t-m) \quad \text{DATI FINO A } t-1$$

Osservazione

Il predittore è ricorsivo, bisogna dire quanto vale $\hat{y}(1|0)$. Di solito si usa la media dei campioni

Se il predittore è AS. STAR, l'effetto dell'inizializzazione svanisce dopo un certo periodo di tempo

Osservazione

La proprietà $E[\hat{y}(t|t-k) \epsilon(t)] = 0$, dove $\epsilon(t) = y(t) - \hat{y}(t|t-k)$ è l'errore di predizione, è una condizione necessaria ma non sufficiente per l'ottimalità del preditore.

PREDITTORE K PASSI MA(m)

Generico processo MA(m) in forma canonica $e(t) \sim \text{wn}(0, \sigma^2)$

$$y(t) = \underbrace{c_0 + c_1 e(t-1) + \dots + c_{k-1} e(t-k+1)}_{\text{PARTE IMPREDICIBILE}} + \underbrace{c_k e(t-k) + \dots + c_m e(t-m)}_{\text{PARTE PREDICIBILE}}$$

\downarrow \downarrow
 $\epsilon(t)$ $\hat{y}(t|t-k)$

Il preditore è ottimo in quanto le condizioni necessarie sono soddisfatte ed $\epsilon(t)$ è a varianza minima

Osservazione

$$\epsilon_1(t) = y(t) - \hat{y}(t|t-1) \Rightarrow \text{Var}[\epsilon_1(t)] = \text{Var}[e(t)] = \sigma^2$$

$$\epsilon_2(t) = y(t) - \hat{y}(t|t-2) \Rightarrow \text{Var}[\epsilon_2(t)] = \text{Var}[e(t) + c_1 e(t-1)] = (1 + c_1^2) \sigma^2$$

$$\vdots$$

$$\epsilon_m(t) = y(t) - \hat{y}(t|t-m) \Rightarrow \text{Var}[\epsilon_m(t)] = \text{Var}[e(t) + c_1 e(t-1) + \dots + c_m e(t-m)] = \text{Var}[y(t)]$$

La varianza di $\epsilon(t)$ aumenta con l'aumento di predizione, fino a diventare uguale alla varianza del processo. Il preditore $\hat{y}(t|t-m)$ sarà il preditore buono, ovvero la media del processo $\hat{y}(t|t-m) = E[y(t)] = 0 \Rightarrow$ un preditore non può mai avere varianza dell'errore di predizione maggiore della varianza del processo

PREDITTORE DI PROCESSO ARMA

Sia dato un processo ARMA(m, n) in forma canonica

$$y(t) = \frac{C(z^{-1})}{A(z^{-1})} e(t) \quad e(t) \sim \text{wn}(0, \sigma^2)$$

$$C(z^{-1}) = 1 + c_1 z^{-1} + c_2 z^{-2} + \dots + c_m z^{-m}$$

$$A(z^{-1}) = 1 - a_1 z^{-1} - a_2 z^{-2} - \dots - a_n z^{-n}$$

Non è chiaro come scomporre parte nota e parte impredicibile, perché $y(t), y(t-1), \dots$ dipendono dai passi precedenti di $e(t)$

Ci ricordiamo che possiamo esprimere $A(z)A(z^{-1})$ come $\tilde{A}(z)$

↓
Non è però fattibile, perché le poste predittive contenrebbero os elementi

Si esprime quindi $\frac{C(z)}{A(z)}$ come un quoziente $E(z)$ più un resto $R(z) = z^{-k} \tilde{R}(z)$
effettuando una lunga divisione. Otteniamo quindi:

$$C(z) = E(z) \cdot A(z) + R(z) \Rightarrow \frac{C(z)}{A(z)} = E(z) + \frac{R(z)}{A(z)} = E(z) + z^{-k} \frac{\tilde{R}(z)}{A(z)}$$

Effettuando la posta di lunga divisione (posta essere ∞), otteniamo le
informazioni per una revisione a k passi

Es

$$y(z) = \frac{1 + \frac{1}{2} z^{-1}}{1 + \frac{1}{3} z^{-1}} e(z) \quad e(z) \sim \mathcal{WN}(0, \sigma^2) \quad k=2$$

Lunga divisione di $k=2$ passi:

$$\begin{array}{c} C(z) \\ \downarrow \\ \left(\begin{array}{c} 1 + \frac{1}{2} z^{-1} \\ -1 - \frac{1}{3} z^{-1} \\ \hline 1 \end{array} \right) \end{array} \quad \begin{array}{c} A(z) \\ \downarrow \\ \left(\begin{array}{c} 1 + \frac{1}{3} z^{-1} \\ 1 + \frac{1}{6} z^{-1} \\ \hline \end{array} \right) \end{array}$$
$$E(z) = \frac{1}{6} z^{-1}$$
$$R(z) = z^{-k} \tilde{R}(z) = z^{-2} \cdot \left(\frac{-1}{18} \right)$$

Sostituendo in $y(z) = \frac{C(z)}{A(z)} e(z)$, otteniamo

$$y(z) = E(z) e(z) + \frac{\tilde{R}(z)}{A(z)} e(z-k)$$

1) $E(z) e(z)$ è IMPREDICIBILE al tempo $t-k$, dipende solo da $e(z), e(z-1), \dots, e(z-k+1)$

2) $\frac{\tilde{R}(z)}{A(z)} e(z-k)$ è COMPLETAMENTE NOTO, dipende da $e(z-k), e(z-k-1), e(z-k-2), \dots$

Quindi il predittore del rumore è:

$$\hat{g}(t|t-k) = \frac{\tilde{R}(z)}{A(z)} e(z-k)$$

L'errore di predizione è:

$$E(z) = y(z) - \hat{g}(t|t-k) = E(z) e(z)$$

Osservazioni

- $\hat{y}(t|t-k)$ dipende dal numero fini di tempo $t-k$
- $E[\hat{y}(t|t-k) \cdot e(t)] = 0$, infatti $E[\hat{y}(t|t-k) e(t)] = E\left[\left(\frac{\tilde{R}(k)}{A(k)} e(t-k)\right)(E(t) e(t))\right] = 0$
- Non è possibile trovare un predittore con le due proprietà precedenti, com
 $\text{Var}[e(t)]$ infinito

↓

$$\hat{y}(t|t-k) = \frac{\tilde{R}(k)}{A(k)} e(t-k) \text{ è ottimo}$$

Preditore dei dati

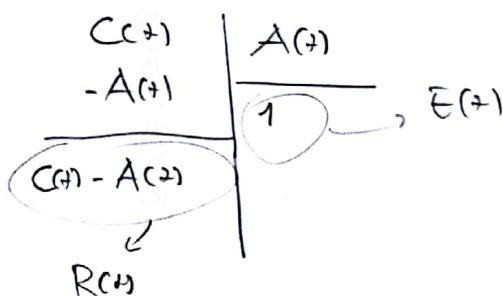
$$y(t) = \frac{C(t)}{A(t)} e(t) \Rightarrow e(t) = \frac{A(t)}{C(t)} y(t)$$

$$\hat{y}(t|t-k) = \frac{\tilde{R}(k)}{A(k)} e(t-k) = \frac{\tilde{R}(k) e^{-k}}{A(k)} e(t) = \frac{\tilde{R}(k) e^{-k}}{A(k)} \frac{A(t)}{C(t)} y(t) = \boxed{\frac{\tilde{R}(k)}{C(t)} y(t-k)}$$

Osservazione

$$e(t) = y(t) - \hat{y}(t|t-k) = E(t) e(t) \quad \text{L'errore di predizione è un processo MA}(k-1)$$

CASO k=1



$$E(t) = \pm$$

$$R(t) = C(t) - A(t)$$

$$\hat{y}(t|t-1) = \frac{C(t) - A(t)}{C^2} y(t)$$

PREDITTORE
DATA
DATA

$$\hat{y}(t|t-1) = \frac{C(t) - A(t)}{A(t)} e(t)$$

PREDITTORE
DATA
PREDICTION

$$E(t) = y(t) - \hat{y}(t|t-1) = e(t)$$

Osservazione

Gli stimatori trovati sono corretti in quanto $E[e(t)] = 0$

QUALITÀ DEL PREDITTORE

Possiamo valutare le qualità di un predittore mettendo a confronto le varianze dell'errore di predizione con le varianze delle predizioni bivariate.

$$\text{Error Signal Ratio} \leftarrow ESR = \frac{\text{Var} [y(t) - \hat{y}(t|t-\kappa)]}{\text{Var} [y(t)]} = \frac{\text{Var} [E_u(t)]}{\text{Var} [y(t)]}$$

oggi non
sia più
possibile

IL PREDITTORE OTTIMO DI UN PROCESSO ARMAX o

Sia $y(t)$ un processo ARMAX(m, m, p), $\frac{C(+)}{A(+)}$ è in forma canonica

$$y(t) = \frac{B(+)}{A(+)} u(t-\kappa) + \frac{C(+)}{A(+)} e(t) \quad e(t) \sim \text{wn} (0, \sigma^2)$$

$\frac{B(+)}{A(+)}$ è il modello del sistema, non si può mettere in forma canonica.

Si vuol fare una predizione a k passi, in modo che l'impasso influisci l'uscita. Utilizziamo la lunga divisione per scoprire $\frac{C(+)}{A(+)}$.

$$y(t) = \underbrace{\frac{B(+)}{A(+)} u(t-\kappa)}_{\text{PARTE PREDICIBILE A } t-\kappa} + \underbrace{\frac{\tilde{R}(+)}{A(+)} e(t-\kappa) + \underbrace{E(+)}_{\text{PARTE IMPREDICIBILE}} e(t)}$$

Predittore del rumore

$$\hat{y}(t|t-\kappa) = \frac{B(+)}{A(+)} u(t-\kappa) + \frac{\tilde{R}(+)}{A(+)} e(t-\kappa)$$

$$E(t) = E(+) e(t)$$

Osserviamo che

- $\hat{y}(t|t-\kappa)$ dipende dal $u(t)$ e dall'impasso fin dal tempo $t-\kappa$
- $E[\epsilon(t)] = 0$ CORRETTO
- $E[\hat{y}(t|t-\kappa) \cdot E(t)] = 0$ (consideriamo $u(t) \perp e(t)$)
- Si dimostra che $\hat{y}(t|t-\kappa)$ è ottimo
- $E(t)$ è identico al caso ARMA: questo perché l'impasso $u(t-\kappa)$ è completamente NOTO e non introduce incertezza

Predittore dei dati

$$y(t) = \frac{B(+)}{A(2)} u(t-\kappa) + \frac{C(+)}{A(2)} e(t) \Rightarrow e(t) = \left(\frac{A(2)}{C(+)} y(t) - \frac{B(+)}{C(+)} u(t-\kappa) \right)$$

$$\hat{y}(t|\kappa) = \frac{B(+)}{A(2)} u(t-\kappa) + \frac{R(+)}{A(2)} e(t)$$

Facendo i passaggi si ottiene:

$$\left| \hat{y}(t|\kappa) = \frac{\tilde{R}(+)}{C(+)} y(t-\kappa) + \frac{B(+)}{C(+)} e(t-\kappa) \right|$$

CASO $\kappa=1$

$$\begin{aligned} E(+) &= 1 \\ R(+) &= C(+) - A(+) \end{aligned}$$

\Rightarrow

$$\left| \begin{aligned} \hat{y}(t|t-1) &= \frac{C(+) - A(+)}{C(+)} y(t) + \frac{B(+)}{C(+)} u(t-1) \\ E(t) &= E(+) e(t) = e(t) \end{aligned} \right|$$

Osservazione

Il predittore ARMAX ha varianza di $E(t)$ data solo dalla parte ARMA. Si può calcolare la bala' delle predizioni come:

$$ESR = \frac{\text{Var}[E(t)]}{\text{Var}\left[\frac{C(+)}{A(+)} e(t)\right]}$$

Es

Dato il processo $y(t) = (2 + 6z^{-1}) u(t-2) + \frac{2}{3 + \frac{3}{2} z^{-1}} \eta(t-1)$ $\eta(t) \sim WN(0, 1)$

Calcolare il predittore dei dati e le varianze dell'errore

Il ritardo per $\kappa=2$, quindi la sara' calcolare un predittore a 2 passi

- Il processo è in forma causale? \Rightarrow Solo la parte stocistica può essere modificata!



$$\begin{aligned}
 y(t) &= (2 + 6z^{-1}) \cdot u(t-2) + \frac{z^{-1}}{1 + \frac{1}{2}z^{-1}} \cdot \frac{2}{3} \eta(t) \\
 &= (2 + 6z^{-1}) u(t-2) + \frac{1}{1 + \frac{1}{2}z^{-1}} \left(\frac{2}{3} \eta(t) \right) \rightarrow e(t) \sim \text{WN}\left(0, \frac{4}{3}\right) \\
 &= (2 + 6z^{-1}) u(t-2) + \frac{1}{1 + \frac{1}{2}z^{-1}} e(t) \rightarrow C(z) \\
 &= (2 + 6z^{-1}) u(t-2) + \frac{1}{1 + \frac{1}{2}z^{-1}} e(t) \rightarrow A(z)
 \end{aligned}$$

Per il predittore, parte espressa e parte restante devono avere lo stesso denominatore

$$A(2) : \quad y(4) = \frac{(2+6z^{-1})(1+\frac{1}{2}z^{-1})}{1+\frac{1}{2}z^{-1}} u(t-2) + \frac{1}{1+\frac{1}{2}z^{-1}} e(t)$$

$$A(\omega) = -1 + \frac{1}{2} \omega^2 \quad C(\omega) = 2$$

$$P_0(z) = -z \left(1 + 3z^{-1}\right) \left(1 + \frac{1}{z} z^{-1}\right)$$

If democratic norms in states suffer deterioration be committed to the $\frac{C+1}{A+1}$

- Predittore è $k=2$ possi \Rightarrow 2 possi di lunga divisione fra $\frac{(C+1)}{A(Q)}$

$$\begin{array}{c}
 \text{Diagram showing partial fraction decomposition:} \\
 \frac{-1 - \frac{1}{2}z^{-1}}{-\frac{1}{2}z^{-1}} = 1 + \frac{1}{2}z^{-1} \rightarrow A(z) \\
 + \frac{\frac{1}{2}z^{-1} + \frac{1}{2}z^{-2}}{z^{-2}} = 1 - \frac{1}{2}z^{-1} \rightarrow E(z) \\
 + \frac{\frac{1}{2}z^{-2}}{z^{-2}} = R(z) = z^{-2} \tilde{R}(z)
 \end{array}$$

$$\hat{y}(t|t-\kappa) = \frac{\tilde{R}(+)}{C(+)} y(t-\kappa) + \frac{P(+|+) E(+)}{C(+)} \cdot u(t-\kappa)$$

↓

$$\hat{y}(t|t-2) = \frac{\frac{1}{\zeta}}{1-\zeta} y(t-2) + \frac{2(1+3z^{-1})(1+\frac{1}{z}z^{-1})(1-\frac{1}{z}z^{-1})}{z} u(t-2)$$

$$\begin{aligned} \text{Var}[\epsilon(t)] &= E[\epsilon(t)^2] = E[(E(t)\epsilon(t))^2] = E\left[\left(1 - \frac{1}{2}\sigma^2\right)E(t)\epsilon(t)\right]^2 = E\left[\left(E(t) - \frac{1}{2}\sigma^2\right)^2\right] \\ &= \left(1 + \frac{1}{4}\right)\text{Var}[\epsilon(t)] = \frac{5}{4} \cdot \frac{4}{3} = \boxed{\frac{5}{3}} \end{aligned}$$

Non ha senso per un AR(1) confrontare $\text{Var}[\epsilon(t)]$ con $\text{Var}[y(t)]$, perché $y(t)$ potrebbe essere un valore
si confronta con $\text{Var}\left[\frac{C(t)}{A(t)}\epsilon(t)\right]$, ovvero con la parte stocastica del processo

$$\cdot \text{Var}\left[\frac{1}{1 + \frac{1}{2}\sigma^2}\epsilon(t)\right] = \text{Var}\left[-\frac{1}{2}v(t-1) + \epsilon(t)\right] = E\left[\left(-\frac{1}{2}v(t-1) + \epsilon(t)\right)^2\right] = +\frac{1}{2}\text{Var}[v(t)] + \text{Var}[\epsilon(t)]$$

$$V(t) \Rightarrow \text{Var}[v(t)] = \frac{1}{2}\text{Var}[v(t)] + \text{Var}[\epsilon(t)] \Rightarrow \frac{3}{2}\text{Var}[v(t)] = \frac{5}{3} \Rightarrow \text{Var}[v(t)] = \frac{16}{27}$$

$$\text{ESR} = \frac{\text{Var}[\epsilon(t)]}{\text{Var}[v(t)]} = \frac{\frac{5}{3}}{\frac{16}{27}} = \frac{5 \cdot 3}{16} = 0.9375$$

IDENTIFICAZIONE

Le analisi riute finora si basavano sulle conoscenze dei parametri del modello dinamico.

↓

La disciplina dell'identificazione consiste nello stimare il sistema ignoto.

Dipendendo da un set di dati per i quali input - uscita $u(t)$ e $y(t)$, si vuole ricavare il sistema $G(s)$ tale che:

$$u(t) \xrightarrow{G(s)} y(t)$$

I possibili problemi di identificazione sono:

① RACCOLTA DATI Sperimentali

L' Scelta del tipo di impulso $u(t)$, in modo da massimizzare l'informazione contenuta nei dati

L' Numero di dati da acquisire

② SCELTA DELLA FAMIGLIA DI MODELLI

a) Scelta del tipo di modello $M(\theta)$ da usare, il quale dipende dai parametri: $\theta \in \Theta^{\text{discrete}}$, da stimare. Tipi di modelli:

- Tempo discreto / continuo
- Sistema lineare / non lineare \Rightarrow ARMAX sono le categorie di modelli più complete da vedere
- Tempo invariante / variante
- Dinamico / statico

b) Scelta degli ordini del modello

③ SCELTA DELL'ALTRA DI VERITO

È una funzione $J_n(\theta) : \mathbb{R}^{d_\theta} \rightarrow \mathbb{R}^+$ che definisce la qualità della stima di $\theta \in \Theta^{\text{discrete}}$.
L'obiettivo da usare sarà l'obiettivo PES (Prediction Error Minimization), mimimizzare la varianza dell'errore di predizione ad un passo.

$$J(\theta) = E \left[(y(t) - \hat{y}(t-1, \theta))^2 \right]$$

73

Dato che le sol N dati, si usa la variante campionaria:

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

↓ è un stimatore
CORRETTO!

④ MINIMIZZAZIONE ATTRA DI MERITO

Vi sono diversi così:

a) $J_N(\theta)$ è quadratica: è possibile trovare il minimo $\hat{\theta}_N$ in forma esplicita (minimi quadrati)

b) $J_N(\theta)$ non quadratica ma con la minimi locali: metodi iterativi
 - gradienti
 - Newton \Rightarrow convergenza verso l'unico minimo
 - Quasi-Newton

c) $J_N(\theta)$ non quadratica e con minimi locali: metodi iterativi, tentando di evitare minimi locali

- AR/ARX $\rightarrow J_N(\theta)$ quadratica
- ARMAX, ARMA, MA $\rightarrow J_N(\theta)$ non quadratica + minimi locali

⑤ VALIDAZIONE DEL MODELLO $M(\hat{\theta}_N)$

← STIMA CAMPIONARIA DI MEDIA E FUNZIONE COVARIANZA →

Sia $y(t)$ un pss e supponiamo di aver misurato ne realizzazioni di $y(t)$.
 $\{y(1), y(2), \dots, y(N)\}$. Intendiamo $y(1) = y(1, \bar{s})$, $y(2) = y(2, \bar{s}), \dots$

Problema: trovare dai dati le seguenti quantità teoriche

$$m = E[y(t)] \rightarrow \hat{m}_N(y(1), y(2), \dots, y(N))$$

$$\sigma^2 = E[(y(t) - m)(y(t+\tau) - m)] \rightarrow \hat{\sigma}^2_N(\tau)(y(1), y(2), \dots, y(N))$$

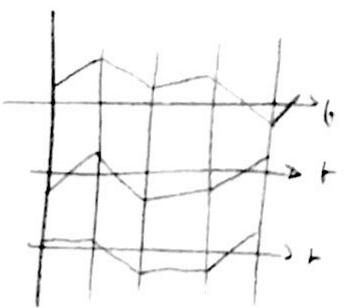
Per trovare ci serve perché non possiamo calcolare il valore atteso dato che non conosciamo la distribuzione dei dati

MEDIA CAMPIONARIA

Un possibile stimatore è $\hat{m}_N = \frac{1}{N} \sum_{t=1}^N y(t)$

Verifica correttezza

$$\begin{aligned}\hat{m}_N &= \frac{1}{N} \sum_{t=1}^N y(t) \Rightarrow E_s[\hat{m}_N] = \frac{1}{N} E_s\left[\sum_{t=1}^N y(t)\right] = \frac{1}{N} \sum_{t=1}^N E_s[y(t)] \\ &= \frac{1}{N} \sum_{t=1}^N m(t) = \frac{1}{N} \cdot N \cdot m = \boxed{m} \quad \text{corretto}\end{aligned}$$



Consistenza

Quantità dei teoremi:

Teoremi

\hat{m}_N è consistente sse $\gamma(\tau) \rightarrow 0$ per $|\tau| \rightarrow +\infty$

Teorema

Dato un processo AR(1), si ha che $\gamma(\tau) \rightarrow 0$ per $|\tau| \rightarrow +\infty$

& FUNZIONE DI COVARIANZA CAMPIONARIA

Ottimizziamo un passo a media nulla. Ricordando che $\gamma(\tau) = E[y(t) \cdot y(t+\tau)]$, un possibile stimatore può essere:

$$\hat{\gamma}_N(\tau) = \frac{1}{N-|\tau|} \sum_{t=1}^{N-|\tau|} y(t) \cdot y(t+|\tau|) \quad \boxed{|\tau| \leq N-1}$$

Osservazioni

- Più τ è grande, meno campioni passa utile. Quindi $\hat{\gamma}_N(\tau)$ è ottimo se $N \gg |\tau|$.



Correttezza

$$\begin{aligned}E_s[\hat{\gamma}_N(\tau)] &= E_s\left[\frac{1}{N-|\tau|} \sum_{t=1}^{N-|\tau|} y(t) \cdot y(t+|\tau|)\right] = \frac{1}{N-|\tau|} \sum_{t=1}^{N-|\tau|} E_s[y(t) y(t+|\tau|)] \\ &= \frac{1}{N-|\tau|} \cdot (N-|\tau|) \cdot \gamma(\tau) = \boxed{\gamma(\tau)} \quad \text{CORRETTO!}\end{aligned}$$

INTRODUZIONE

Un altro modo per identificare un processo è attraverso le caratteristiche fondamentali

↓

STIMA DI MEDIA,
COVARIANZA, SPECTRO

RECUPERAZIONE DELLA STIMA DELLO SPECTRO

Lo stimatore dello spettro non gode di buone proprietà. Le stime non sono quindi buone.

↓
Un metodo per migliorare le stime è il seguente:

L'ottimale di dividere gli N dati del processo misurato in M parti

L'calcoliamo $\hat{\Gamma}_{N/M}^{(i)}(\omega)$ per ciascuna parte i , $i=1, \dots, M$

L'calcoliamo la stima finale ($\hat{\Gamma}_N(\omega) = \frac{1}{M} \cdot \sum_{i=1}^M \hat{\Gamma}_{N/M}^{(i)}(\omega)$)

Si dimostra che:

$$E\left[\left(\hat{\Gamma}_N(\omega) - \Gamma(\omega)\right)^2\right] = \frac{1}{M} E\left[\left(\hat{\Gamma}_N(\omega) - \Gamma(\omega)\right)^2\right] \quad \begin{array}{l} \text{obtienere cioè ridotto} \\ \text{di } \frac{1}{M} \text{ la varianza dell'errore} \\ \text{di stima} \end{array}$$

Osservazione

Le scelte di M rappresenta un trade-off. Infatti se raggruppi troppi (M grande), avrai uno stimatore meno corretto (perché usi meno dati per le stime, e lo stimatore è solo asintoticamente corretto)

↓
Per ottenere trade-off esistono due modi: obiettivo visto la risoluzione di funzioni dei costi

$$J(\theta) = \frac{1}{N} \sum_{i=1}^N (y(i) - \hat{\Gamma}^{(i)}(\theta))^2 + \lambda \sum_{i=1}^M \theta_i^2 \quad \bullet \lambda \text{ GRANDE} \Rightarrow \begin{array}{l} + \text{BIAS} \\ - \text{VARIANCE} \end{array}$$



RIPRESA IDENTIFICAZIONE

Consistenza

Teorema

$\hat{f}_N(\gamma)$ è consistente sse $f(\gamma) \rightarrow 0$ per $|\gamma| \rightarrow +\infty$

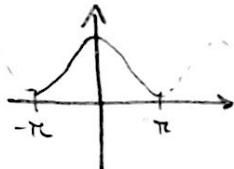
VARIANTE NON CORRETTA

$\hat{f}'_N(\gamma) = \frac{1}{N} \cdot \sum_{t=1}^{N-1|\gamma|} y(t) \cdot y(t+1|\gamma|)$ non è corretto, ma è asintoticamente corretto
 $N \rightarrow +\infty$

Questa chiave è comunque utilizzabile perché anche nella variante corretta si suppone che $N \gg M$

DENSITÀ SPECTRALE CAMPIONARIA

Opzioni: un passo a media nulla. Sappiamo che $\tilde{f}(u) = \sum_{\gamma=-\infty}^{\gamma=\infty} f(\gamma) e^{-j\gamma u}$ (è DTFT - Discrete Time Fourier Transform della $y(n)$). $\tilde{f}(u)$ è una funzione continua di u .



Osservazione

La pulsazione $u=\pi$ corrisponde alla frequenza di Nyquist del sistema \Rightarrow Es se $f_s = 20 \text{ Hz}$, $\pi = \frac{f_s}{2} = 50 \text{ Hz}$

Una possibile chiave potrebbe essere $\hat{f}_N(u) = \sum_{\gamma=-\infty}^{+\infty} \hat{f}_N(\gamma) e^{-j\gamma u}$. Ricordando però che:

$$\hat{f}_N(\gamma) = \frac{1}{N-1|\gamma|} \cdot \sum_{t=1}^{N-1|\gamma|} y(t) y(t+1|\gamma|) \quad (|\gamma| \leq N-1)$$

notiamo che non posso avere valori di $\hat{f}_N(\gamma)$ per $|\gamma| \geq N$. Quindi oppure:

$$\left| \begin{array}{l} \hat{f}_N(u) = \sum_{\gamma=-N+1}^{N-1} \hat{f}_N(\gamma) \cdot e^{-j\gamma u} \\ \gamma = -(N-1) \end{array} \right.$$

Osservazione

$\tilde{f}(u)$ contiene due opposizioni:

- 1) Usare $\hat{f}_N(\gamma)$ al posto di $f(\gamma)$
- 2) Obbligare finito la sommatoria a $\pm(N-1)$

Osservazione

$\tilde{f}(u)$ è continua. Nelle pratiche, dobbiamo discretizzare l'intervallo $[0, \pi]$

Corretto

$$E_s[\hat{\Gamma}_N(u)] = E_s\left[\sum_{\tau=-N+1}^{N-1} \hat{g}_N(\tau) e^{-j\omega\tau}\right] = \sum_{\tau=-N+1}^{N-1} \underbrace{E_s[\hat{g}_N(\tau)] e^{-j\omega\tau}}_{g(\tau)} \\ = \sum_{\tau=-N+1}^{N-1} g(\tau) e^{-j\omega\tau} \neq \Gamma(u) \quad \text{perché gli indici delle somme sono diversi!} \quad \text{NON CORRETTO}$$

Se però $N \rightarrow \infty$, abbiamo che:

$$E_s[\hat{\Gamma}_N(u)] \xrightarrow{N \rightarrow \infty} \Gamma(u) \Rightarrow \hat{\Gamma}_N \text{ è ASINTOTICAMENTE CORRETTO}$$

Consistenza

$$\text{Si dimostra che: } \lim_{N \rightarrow +\infty} E_s\left[\left(\hat{\Gamma}_N(u) - \Gamma(u)\right)^2\right] = \Gamma(u)^2 \geq 0$$

$\hat{\Gamma}_N(u)$ NON È CONSISTENTE

Inoltre abbiamo che:

$$\lim_{N \rightarrow +\infty} E_s\left[\left(\hat{\Gamma}_N(u_1) - \Gamma(u_1)\right) \cdot \left(\hat{\Gamma}_N(u_2) - \Gamma(u_2)\right)\right] = 0 \quad \forall u_1, u_2, u_1 \neq u_2$$

Ora l'errore di stima ad una frequenza u_1 è inconfondibile con l'errore di stima ad una frequenza u_2 .

↓
difficile ridurre la varianza della stima

STIMATORE ALTERNATIVO

Usiamo la variante NON CORRETTA $\hat{g}'_N(\tau)$ invece di $\hat{g}_N(\tau)$. Abbiamo che:

$$\hat{\Gamma}'_N(u) := \sum_{\tau=-N+1}^{N-1} \hat{g}'_N(\tau) e^{-j\omega\tau} = \sum_{\tau=-N+1}^{N-1} \left[\frac{1}{N} \sum_{s=1}^{N-1} y(s) y(s+\tau) \right] e^{-j\omega\tau}$$

$$\Rightarrow \boxed{r=s+\tau} \Rightarrow \boxed{\tau=r-s} \quad \begin{array}{l} \bullet \tau=-N+1 \Rightarrow r=1, s=-N \\ \bullet \tau=N-1 \Rightarrow r=N, s=1 \end{array}$$

$$= \frac{1}{N} \sum_{r=1}^N \sum_{s=1}^{N-1} y(r) y(s) e^{-j\omega(r-s)} = \frac{1}{N} \sum_{r=1}^N y(r) e^{-j\omega r} \cdot \sum_{s=1}^N y(s) e^{j\omega s}$$

$$= \frac{1}{N} \cdot \left| \sum_{r=1}^N y(r) e^{-j\omega r} \right|^2$$

È il modul^o del volo ritornato dalla DFT (Discrete Fourier Transform), che voluta lo spettro in un range discreto di Frequenze u

Questo stimatore è meno corretto del precedente ma si calcola velocemente (FFT) senza passare per $\hat{g}_N(\tau)$

IDENTIFICAZIONE DI MODELLI ARX

Dati disponibili: $\{y(1), y(2), \dots, y(N)\}$ e $\{u(1), u(2), \dots, u(N)\}$

Consideriamo un generico modello ARX($m, p+1$):

$$y(t) = \frac{B(z)}{A(z)} u(t-1) + \frac{1}{A(z)} e(t) \quad e(t) \sim WN(0, \sigma^2)$$

$$B(z) = b_0 + b_1 z^{-1} + \dots + b_p z^{-p} \quad C(z) = 1$$

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \dots + a_m z^{-m}$$

Osservazione

$C(z) = 1$ poiché non esiste la parte MA. Fissando il rettangolo per $n=1$ bediamo di generalità? No, perché, ad esempio, se n fosse = 2, identificheremmo $b_0 = 0$.

Cifra di merito

È la varianza composta dell'errore di predizione

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

Preditore:

$$\begin{aligned} \hat{y}(t|t-1) &= \frac{B(z)}{A(z)} u(t-1) + \frac{1}{A(z)} y(t) \\ &\quad \xrightarrow{\text{d}} C(z) \\ &= (b_0 + b_1 z^{-1} + \dots + b_p z^{-p}) u(t-1) + (-a_1 z^{-1} - a_2 z^{-2} - \dots - a_m z^{-m}) y(t) \\ &= b_0 u(t-1) + b_1 u(t-2) + \dots + b_p u(t-p-1) - a_1 y(t-1) - a_2 y(t-2) - \dots - a_m y(t-m) \end{aligned}$$

VETTORE DEI PARAMETRI

$$\theta = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \\ b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} \in \mathbb{R}^{m+p+1} \times 1 \quad d = m+p+1$$

VETTORE DELLE OSSERVAZIONI (dei dati)

$$\varphi(t) = \begin{bmatrix} -y(t-1) \\ -y(t-2) \\ \vdots \\ -y(t-m) \\ u(t-1) \\ u(t-2) \\ \vdots \\ u(t-p-1) \end{bmatrix} \in \mathbb{R}^{m+p+1} \times 1$$

$$\hat{y}(t|t-1) = \varphi(t)^T \cdot \theta$$

$$J_N(\theta) = \frac{1}{N} \sum_{t=1}^N (y(t) - \varphi(t)^T \cdot \theta)^2 \Rightarrow \boxed{\text{SOMMA A MINIMI QUADRATI}}$$

(78)

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

Osservazione

In generale per un ARX $(m, p+1)$ è qui corretto scrivere

$$J_N(\theta) = \frac{1}{N-h} \sum_{t=h+1}^N (y(t) - \varphi(t)^T \theta)^2 \quad \text{con } h = \max(m, p+1)$$

Es

Supponiamo di avere $N=10$ dati I/O $\{y(1), y(2), \dots, y(10)\}$

Stimare un modello ARX del tipo: $\{u(1), u(2), \dots, u(10)\}$

$$y(t) = \frac{b}{1+\alpha z^{-1}} u(t-1) + \frac{1}{1+\alpha z^{-1}} e(t) \quad e(t) \sim \text{unif}(0, 1)$$

È un ARX(1, 1). Il predittore è: $\hat{y}(t|t-1) = \frac{b(2)}{C(2)} u(t-1) + \frac{C(1) - A(2)}{C(2)} y(t)$

$$\begin{aligned} \text{Cifre di merito, ottienere } h &= \max(m, p+1) \\ &= 1 \end{aligned} \quad \begin{aligned} &= b u(t-1) - \alpha y(t-1) \quad \theta = \begin{bmatrix} \alpha \\ b \end{bmatrix} \in \mathbb{R}^{2 \times 1} \\ &\text{dove partire da } t=2 \text{ se non} \\ &\text{conosci } u(t-1), y(t-1) \quad \begin{array}{l} \text{reflessione} \\ x_1(t) \\ x_2(t) \end{array} \end{aligned}$$

$$J_{10}(\theta) = \frac{1}{10-1} \sum_{t=11}^{10} (y(t) - \hat{y}(t|t-1))^2 = \frac{1}{9} \sum_{t=2}^{10} (y(t) - b u(t-1) + \alpha y(t-1))^2$$

$$\left\{ \begin{array}{l} \frac{d J_{10}(\theta)}{d \alpha} = \frac{2}{9} \sum_{t=2}^{10} (y(t) - b u(t-1) + \alpha y(t-1)) y(t-1) = 0 \\ \frac{d J_{10}(\theta)}{d b} = \frac{2}{9} \sum_{t=2}^{10} (y(t) - b u(t-1) + \alpha y(t-1)) (-u(t-1)) = 0 \end{array} \right. \quad \text{OPTEIMIZZAZIONE}$$

$$\Rightarrow \begin{bmatrix} \sum_{t=2}^{10} y(t-1)^2 & - \sum_{t=2}^{10} u(t-1) y(t-1) \\ - \sum_{t=2}^{10} u(t-1) y(t-1) & + \sum_{t=2}^{10} u(t-1)^2 \end{bmatrix} \begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} - \sum_{t=2}^{10} y(t) y(t-1) \\ \sum_{t=2}^{10} y(t) u(t-1) \end{bmatrix}$$

$$\begin{bmatrix} \hat{\alpha}_{10} \\ \hat{b}_{10} \end{bmatrix} = \begin{bmatrix} \sum_{t=2}^{10} y(t-1)^2 & - \sum_{t=2}^{10} u(t-1) y(t-1) \\ - \sum_{t=2}^{10} u(t-1) y(t-1) & + \sum_{t=2}^{10} u(t-1)^2 \end{bmatrix}^{-1} \cdot \begin{bmatrix} - \sum_{t=2}^{10} y(t) y(t-1) \\ \sum_{t=2}^{10} y(t) u(t-1) \end{bmatrix}$$

78

la soluzione di problemi REN per modelli AR&R può essere formulata dal punto di vista matriciale.

MATRICE DEI DEGRADATORI

$$\Phi = \begin{bmatrix} y(t-1) & u(t-1) \\ -y(1) & u(1) \\ -y(2) & u(2) \\ | & | \\ -y(s) & u(s) \end{bmatrix}_{(N-h) \times d}$$

$(N-1) \times d$
 $s \times 2$

VETTORE OUTPUT

$$Y = \begin{bmatrix} y(2) \\ y(3) \\ | \\ y(s) \end{bmatrix}_{s \times 1}$$

$$Y = \Phi Y \quad g(t|t-1) = -a y(t-1) + b u(t-1)$$

↓

$$y(2) = -a y(1) + b u(1)$$

$$y(3) = -a y(2) + b u(2)$$

$$y(s) = -a y(s-1) + b u(s)$$

$$\hat{\Theta}_N = \left(\Phi^T \Phi \right)^{-1} \Phi^T Y$$

Es

Si suppone di avere 5 dati da una serie temporale $y(t)$ a media nulla.

$$y(1) = \frac{1}{2} \quad y(2) = 0 \quad y(3) = -1 \quad y(4) = -\frac{1}{2} \quad y(5) = +\frac{1}{4}$$

Si identifichi un modello AR(1): $y(t) = \alpha y(t-1) + e(t)$ $e(t) \sim \text{wn}(0, \sigma^2)$

Usando il modello identificato, si calcoli $\hat{g}(6|5)$ e $\hat{\sigma}^2$

Note

Se vede comparsa $\hat{m}_s = \frac{1}{5} \sum_{t=1}^5 y(t) = -0,15$ con \hat{e} nulla. L'esercizio però ci dice di considerare media nulla, altrimenti avremmo fatto deposizionale al fine di avere un predittore corretto, $E[\text{ECA}] = 0$

Calcoliamo il predittore

$$y(t) = \frac{1}{1-\alpha^2} e(t) \quad (\text{Supponiamo } |\alpha| < 1) \quad \Rightarrow \hat{y}(t|t-1) = \frac{(C+1)-AC+1}{C+1} y(t)$$

$$J_N(\theta) = \frac{1}{N-1} \sum_{t=2}^N (y(t) - \alpha y(t-1))^2 = \frac{1+\alpha^{t-1}}{1} y(t) = \boxed{\alpha y(t-1)}$$

$$= \frac{1}{4} \left[(y(2) - \alpha y(1))^2 + (y(3) - \alpha y(2))^2 + (y(4) - \alpha y(3))^2 + (y(5) - \alpha y(4))^2 \right]$$

$$= \frac{1}{4} \left[\left(0 - \alpha \frac{1}{2} \right)^2 + \left(-1 - \alpha \cdot 0 \right)^2 + \left(-\frac{1}{2} + \alpha \cdot 1 \right)^2 + \left(\frac{1}{4} + \alpha \cdot \frac{1}{2} \right)^2 \right]$$

80

$$= \frac{1}{4} \left[\frac{1}{4} \alpha^2 + 1 + \frac{1}{4} + \alpha^2 \cdot \alpha + \frac{1}{16} + \frac{1}{4} \alpha^2 + \frac{1}{4} \alpha \right]$$

$$= \frac{1}{4} \left[\frac{16+4+1}{16} + \frac{-4\alpha + \alpha}{4} + \frac{\alpha^2 + \alpha^2 + 4\alpha^2}{4} \right] = \frac{1}{4} \left[\frac{21}{16} - \frac{3}{4} \alpha + \frac{3}{2} \alpha^2 \right]$$

Minimizzazione

$$\frac{d J_5(\hat{\alpha})}{d \alpha} = 3\alpha - \frac{3}{4} = 0 \Rightarrow \boxed{\hat{\alpha}_5 = \frac{1}{4}}$$

Modello identificabile

$$\left| \begin{array}{l} y(t) = \frac{1}{1 - \frac{1}{4} z^{-1}} e(t) \\ e(t) \sim \mathcal{N}(0, \sigma^2) \end{array} \right.$$

Osservazione

Se osserviamo ottenuto $|\hat{\alpha}_5| > 1$, ovvero potuto usare un filtro passatutto

$$\sigma^2 = \text{Var}[e(t)] = \text{Var}[E(t)] \approx J_5(\hat{\alpha}_5) = \frac{1}{4} \left[\frac{21}{16} - \frac{3}{4} \cdot \frac{1}{4} + \frac{3}{2} \cdot \left(\frac{1}{4} \right)^2 \right]$$

Calcoliamo $\hat{y}(615)$

$$\hat{y}(t|t-1) = \frac{1}{4} y(t-1) \Rightarrow \hat{y}(615) = \frac{1}{4} y(5) = \frac{1}{4} \cdot \frac{1}{4} = \boxed{\frac{1}{16}}$$

↪ IDENTIFICABILITÀ PER DI MODELLI ARMAX ↪ (Metodo delle massime verosimiglianze)

Questo metodo si dice così perché si dimostra che, nel caso in cui $e(t)$ sia un VN Gaussiano, l'approccio per estremale del metodo ML

Dati disponibili

$$\{u(1), u(2), \dots, u(N)\}$$

$$\{y^{(1)}, y^{(2)}, \dots, y^{(n)}\}$$

Generico modello ARMAX $(m, n, p+1)$

$$y(t) = \frac{B(z^{-1})}{A(z^{-1})} u(t-1) + \frac{C(z^{-1})}{A(z^{-1})} e(t) \quad e(t) \sim \mathcal{N}(0, \sigma^2)$$

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \dots + a_m z^{-m}$$

$$B(z) = b_0 + b_1 z^{-1} + b_2 z^{-2} + \dots + b_p z^{-p}$$

$$C(z) = 1 + c_1 z^{-1} + c_2 z^{-2} + \dots + c_n z^{-n}$$

Vettore dei parametri

$$\theta = \begin{bmatrix} \alpha_1 \\ | \\ \alpha_m \\ b_0 \\ b_p \\ c_1 \\ | \\ c_m \end{bmatrix} \in \mathbb{R}^{(m+m+p+1) \times 1}$$

Approssimazione

$$\hat{\theta}_N = \arg \min_{\theta} J_N(\theta)$$

$$J_N(\theta) = \frac{1}{N} \sum_{t=1}^N \epsilon(t, \theta)^2$$

Errore di predizione ad un passo

$$\text{Se } k=1 \Rightarrow E(z) = 1 \Rightarrow E(t) = e(t)$$

$$e(t) = E(t, \theta) = \frac{A(z)}{C(z)} y(t) - \frac{B(z)}{C(z)} u(t-1)$$

$$J_N(\theta) = \frac{1}{N} \sum_{t=1}^N \left(\frac{A(z)}{C(z)} y(t) - \frac{B(z)}{C(z)} u(t-1) \right)^2$$

PROBLEMA

A causa di $C(z)$ al denominatore, la cifra di mercato non è quadratica in θ (e le, in genere, più minimi locali)

Si usano quindi metodi iterativi per la minimizzazione

già visto con la Logistic Regression!

PROBLEMA MINIMI LOCALI

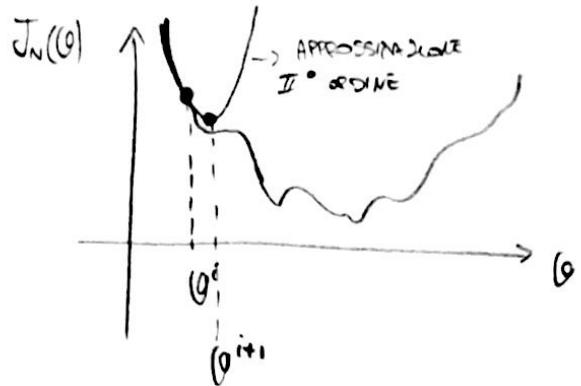
- Si gestisce così:
- Scegliere M inizializzazioni diverse, ottenere M soluzioni
 - Se le M soluzioni sono uguali, possiamo pensare (non sono certi) di aver raggiunto il minimo globale di $J_N(\theta)$
 - Se sono diverse, considerare quella che mi fa ottenere $J_N(\theta)$ minore

METODO DI NEWTON

Idea: sviluppare in serie di Taylor troncate al secondo ordine di $J_N(\theta)$ nell'intorno di θ^i (noto)

$$J_N(\theta) \approx V(\theta) = J_N(\theta^i) + \underbrace{\left. \frac{dJ(\theta)}{d\theta} \right|_{\theta=\theta^i}}_{\text{GRADIENTE}} \cdot (\theta - \theta^i) + \frac{1}{2} (\theta - \theta^i)^T \underbrace{\left. \frac{d^2 J(\theta)}{d\theta^2} \right|_{\theta=\theta^i}}_{\text{MESSIANA}} (\theta - \theta^i)$$

Ossegnare a θ^{i+1} il minimo di $V(\theta)$ ottenuto a $\theta^i \Rightarrow$ trova il minimo della parabola



Minimo di $V(\theta)$

$$\frac{dV(\theta)}{d\theta} = 0 \Rightarrow \left. \frac{dJ(\theta)}{d\theta} \right|_{\theta=\theta^i} + \frac{1}{2} \cdot 2 \cdot \left. \frac{d^2J(\theta)}{d\theta^2} \right|_{\theta=\theta^i} \cdot (\theta - \theta^i) = 0$$

$$\Rightarrow \left. \theta^{i+1} = \theta^i - \left[\frac{d^2J_N(\theta)}{d\theta^2} \right]^{-1} \cdot \left[\frac{dJ_N(\theta)}{d\theta} \right] \right|_{\theta=\theta^i}$$

È simile al gradient descent se si montre costante

$$\left. \frac{d^2J_N(\theta)}{d\theta^2} \right|_{\theta=\theta^i} = \alpha > 0$$

Il metodo funziona se l'Hessiana è DEFINITA POSITIVA,
altrimenti va nelle direzioni sbagliate

Calcoliamo in dettaglio due parti:

- $\bullet \left. \frac{d^2J_N(\theta)}{d\theta^2} \right|_{\theta=\theta^i}$ HESSIANO

- $\bullet \left. \frac{dJ_N(\theta)}{d\theta} \right|_{\theta=\theta^i}$ GRADIENTE

• Calcolo di $\frac{dJ_N(\theta)}{d\theta}$

Ricordiamo che: $J_N(\theta) = \frac{1}{N} \sum_{t=1}^N \epsilon(t)^2 \Rightarrow \left. \frac{dJ_N(\theta)}{d\theta} = \frac{2}{N} \sum_{t=1}^N \epsilon(t) \cdot \frac{d\epsilon(t)}{d\theta} \right|$

- Calcolo di $\frac{d^2 J_N(\theta)}{d\theta^2}$ =) deriva la derivata I^o, regole della derivazione del prodotto

$$\frac{d^2 J_N(\theta)}{d\theta^2} = \frac{2}{N} \sum_{t=1}^N \frac{dE(t)}{d\theta} \cdot \frac{dE(t)^T}{d\theta} + \frac{2}{N} \sum_{t=1}^N E(t) \frac{d^2 E(t)}{d\theta^2}$$

Si ignora questi termini, approssimando così l'Hessian (metodi Quasi-Newton). Le valutazioni sono:

- 1) Se siamo vicini all'ottimo, $E(t)$ è piccolo e il termine conta poco
- 2) Possiamo evitare di calcolare $\frac{d^2 E(t)}{d\theta^2}$
- 3) Ci assicuriamo un Hessian DEFINITO POSITIVO la procedura è sicuramente di MINIMIZZAZIONE

Osserviamo quindi:

$$\theta^{i+1} = \theta^i - \left[\frac{2}{N} \sum_{t=1}^N \frac{dE(t)^{(i)}}{d\theta} \cdot \left(\frac{dE(t)^{(i)}}{d\theta} \right)^T \right]^{-1} \cdot \left[\frac{2}{N} \sum_{t=1}^N E(t)^{(i)} \cdot \frac{dE(t)^{(i)}}{d\theta} \right]$$

per garantire l'invertibilità;
si oppone al termine

Note + $dE(t)$ piccolo, matrice identità

La notazione "picce (i)" indica che stiamo volutamente θ in θ^i (noi)

- Calcolo di $\frac{dE(t)}{d\theta}$

$$E(t) = e(t) = \frac{A(t)}{C(t)} y(t) - \frac{B(t)}{A(t)} u(t-1)$$

$$E(t) = \frac{1+a_1 z^{-1} + \dots + a_m z^{-m}}{1+c_1 z^{-1} + \dots + c_m z^{-m}} y(t) - \frac{b_0 + b_1 z^{-1} + \dots + b_p z^{-p}}{1+c_1 z^{-1} + \dots + c_m z^{-m}} u(t-1)$$

$$\theta = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \\ b_0 \\ \vdots \\ b_p \\ c_1 \\ \vdots \\ c_m \end{bmatrix} \in \mathbb{R}^{(m+p+m+1) \times 1}$$

Derivate di $E(t)$ rispetto a a_1, a_2, \dots, a_m :

- $\frac{dE(t)}{da_1} = \frac{z^{-1}}{C(z)} y(t) = \frac{1}{C(z)} y(t-1) = \alpha(t-1)$

$$\boxed{\alpha(t) = \frac{1}{C(z)} y(t)}$$

- $\frac{dE(t)}{da_2} = \frac{z^{-2}}{C(z)} y(t) = \frac{1}{C(z)} y(t-2) = \alpha(t-2)$

- $\frac{dE(t)}{da_m} = \frac{z^{-m}}{C(z)} y(t) = \frac{1}{C(z)} y(t-m) = \alpha(t-m)$

Derivate di $E(t)$ rispetto a b_0, b_1, \dots, b_p

- $\frac{dE(t)}{db_0} = -\frac{1}{C(z)} u(t-1) = \beta(t-1)$

$$\boxed{\beta(t) = -\frac{1}{C(z)} u(t)}$$

- $\frac{dE(t)}{db_1} = -\frac{z^{-1}}{C(z)} u(t-1) = \beta(t-2)$

- $\frac{dE(t)}{db_p} = -\frac{z^{-p}}{C(z)} u(t-1) = \beta(t-p-1)$

Derivate di $E(t)$ rispetto a c_1, c_2, \dots, c_m

$$E(t) = \frac{A(z)}{C(z)} y(t) - \frac{B(z)}{C(z)} u(t-1) \Rightarrow (1 + c_1 z^{-1} + \dots + c_m z^{-m}) E(t) = A(z) y(t) - B(z) u(t-1)$$

$$\Rightarrow \frac{d}{dc_i} \left[(1 + c_1 z^{-1} + \dots + c_m z^{-m}) \cdot E(t) \right] = \frac{d}{dc_i} \left[A(z) y(t) - B(z) u(t-1) \right]$$

derivabile

$$\Rightarrow z^{-1} E(t) + C(z) \frac{dE(t)}{dc_i} = 0 \Rightarrow \bullet \frac{dE(t)}{dc_i} = -\frac{1}{C(z)} E(t-1) = f(t-1)$$

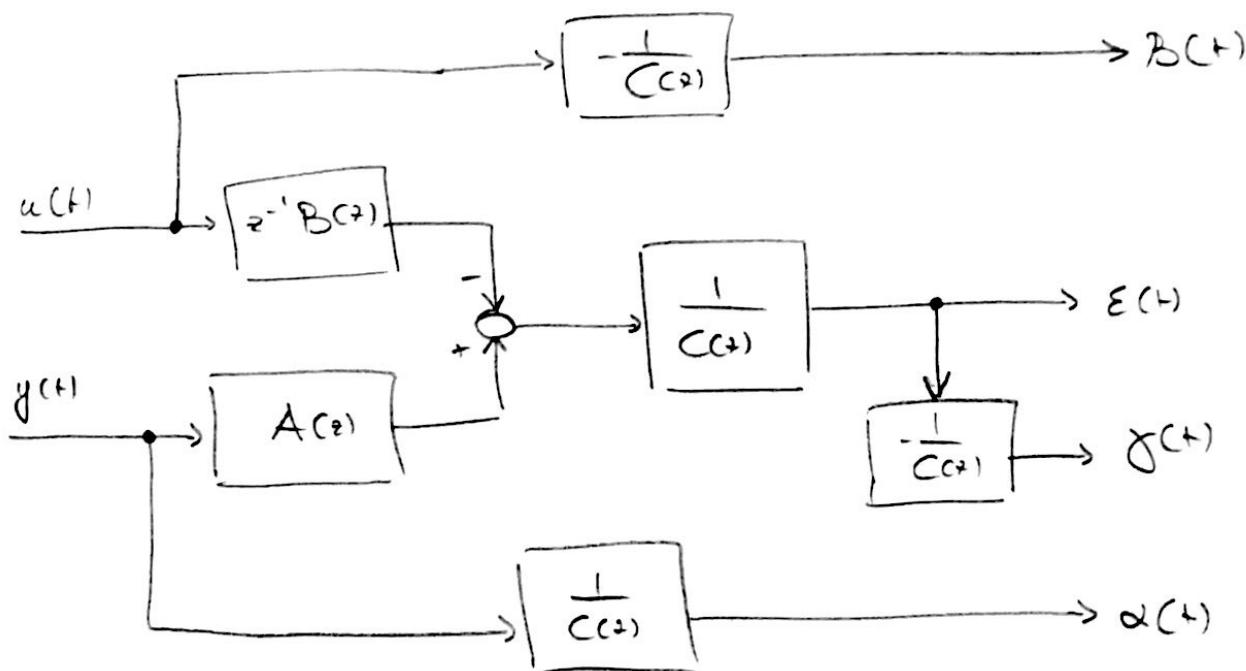
- $\frac{dE(t)}{dc_m} = -\frac{1}{C(z)} E(t-m) = f(t-m)$

$$\boxed{f(t) = -\frac{1}{C(z)} E(t)}$$

Riassumendo, dobbiamo che il vettore delle variazioni è:

$$\frac{dE(t)}{d\theta} = \begin{bmatrix} \alpha(t-1) \\ | \\ \alpha(t-m) \\ B(t-1) \\ | \\ B(t-1-p) \\ \gamma(t-1) \\ | \\ \gamma(t-m) \end{bmatrix} \in \mathbb{R}^{(m+m+p+1) \times 1} \quad t = 1, \dots, N$$

Ovviamente quindi creando un FILTRO che, dati $u(t)$, $y(t)$ e θ^i , permette di calcolare il necessario per dare θ^{i+1} :



Osservazione

Dobbiamo verificare che $C(z+1)^{-1}$ sia AS-STAB: se non lo è dobbiamo ripetere i passi instabili nel corso anteriori (metodo di RADER).

8 IDENTIFICAZIONE: ANALISI E COMPLEMENTI

ANALISI ASINTOTICA METODO PES

Ipotizziamo di avere N dati $\{y(1), y(2), \dots, y(N)\}, \{u(1), u(2), \dots, u(N)\}$
 La stima PES trova $\hat{\theta}_N$ che minimizza: $J_N(\theta) = \frac{1}{N} \sum_{t=1}^N \epsilon(t, \theta)^2$

Come facciamo a sapere che questa stima è (oltre assintoticamente) buona?
 Considerando $N \rightarrow \infty$, otteniamo che:

$$J_N(\theta) \xrightarrow[N \rightarrow \infty]{} J(\theta) = E[\epsilon(t, \theta)^2]$$

L'insieme dei punti di minimo globale di $J(\theta)$ è $\Delta = \left\{ \bar{\theta} \mid J(\theta) \geq J(\bar{\theta}), \forall \theta \right\}$

Osservazioni

- Caso particolare: $\Delta = \bar{\theta}$ ($J(\theta)$ ha un unico minimo globale)

- Dato che $J_N(\theta) \xrightarrow[N \rightarrow \infty]{} J(\theta)$, ci aspettiamo che $\hat{\theta}_N \xrightarrow[N \rightarrow \infty]{} \Delta$

Supponiamo che $S \in M(\theta)$.
 ↓
 $\exists M(\theta^*) = S$ θ^* : vettore verso di parametri

S: SISTEMA VERSO
 $M(\theta)$: CLASSE DI MODELLI M CON PARAMETRI θ

Dimostrazione

Consideriamo un modello $M(\theta)$ e scriviamo l'errore di predizione:

$$\epsilon(t) = y(t) - \hat{y}(t|t-1, \theta) \Rightarrow \epsilon(t) - \hat{y}(t|t-1, \theta^*) = \underbrace{y(t) - \hat{y}(t|t-1, \theta)}_{\text{RUMORE BIANCO}} - \underbrace{\hat{y}(t|t-1, \theta^*)}_{\text{PENSO DI PREDIRE AD UN PASSO}}$$

$$\Rightarrow \epsilon(t) = e(t) - \hat{y}(t|t-1, \theta) + \hat{y}(t|t-1, \theta^*)$$

Applichiamo $E[(\cdot)^2]$ ad entrambi i membri:

$$E[\epsilon(t)^2] = E\left[\left(e(t) + (\hat{y}(t|t-1, \theta^*) - \hat{y}(t|t-1, \theta))\right)^2\right] = \text{consider gli ultimi 2 termini nella lista intera}$$

$$\begin{aligned} J(\theta) &= E[e(t)^2] + E\left[\left(\hat{y}(t|t-1, \theta^*) - \hat{y}(t|t-1, \theta)\right)^2\right] + 2E\left[e(t) \cdot (\hat{y}(t|t-1, \theta^*) - \hat{y}(t|t-1, \theta))\right] \\ &\quad \text{preditori incosistenti con } e(t) \end{aligned}$$

$$J(\theta) = \lambda^2 + E\left[\left(\hat{y}(t|t-1, \theta^*) - \hat{y}(t|t-1, \theta)\right)^2\right] + 0$$

≥ 0 , si annulla per $\theta = \theta^*$

$$\Rightarrow J(\theta) \geq \lambda^2 = J(\theta^*), \forall \theta$$

$$\boxed{J(\theta) \geq J(\theta^*) \quad \forall \theta}$$

Conclusione

Se $S \in M(\theta)$, il metodo PEA è in grado di garantire che il modello studi è quello vero, (nel caso in cui $N \rightarrow \infty$)

L'PEA è ASINTOTICAMENTE CORRETTO

L'PEA è ASINTOTICAMENTE CONSISTENTE (si minimizza la varianza dell'errore)

Se le classi di modelli è sbagliata, i metodi PEA non convergeranno sui parametri veri

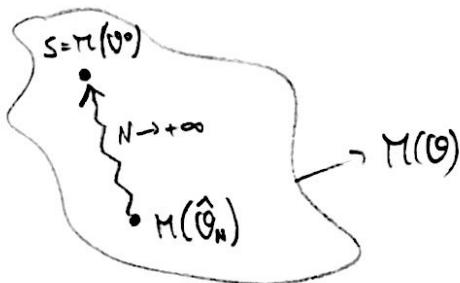
Osservazione

Se $S \in M(\theta)$, in corrispondenza di θ^* si ha che $E(\cdot, \theta^*) = e(t) \sim \mathcal{N}(0, \sigma^2)$

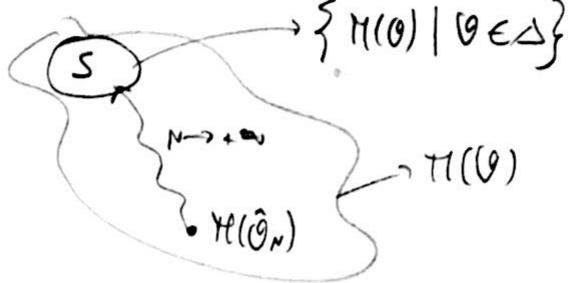
Si può fare un test di bontà per verificare che il modello identificato sia quello vero.

Quando identifichiamo un modello, possono capitare diverse situazioni:

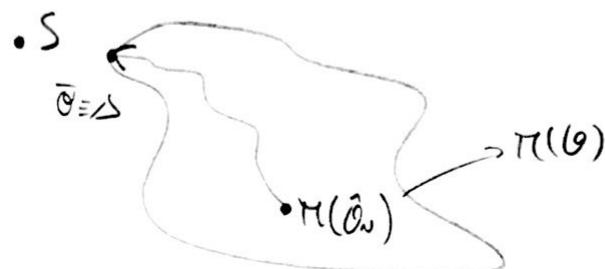
- i) $S \in M(\theta)$ e $\Delta = \bar{\theta} \equiv \theta^*$. La famiglia di modelli scelta è quella del sistema vero. $J(\theta)$ ha un minimo che è θ^* . Oltre ovviamente che $\hat{\theta}_N \xrightarrow[N \rightarrow \infty]{} \theta^*$



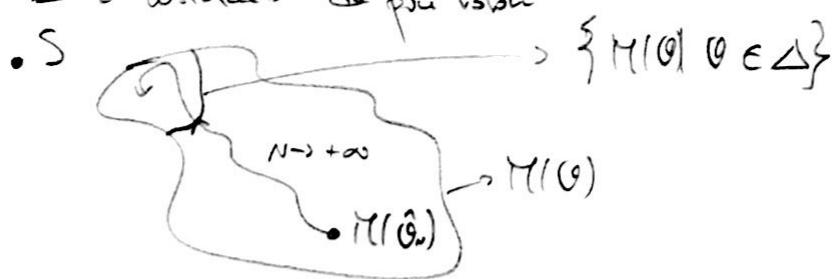
- 2) $S \in \mathcal{H}(\Theta)$ ma Δ è costituito da più voci. Non è un problema tanto l'identificazione tra un set di parametri con le stesse voci delle cifre di merito, sono equivalenti del punto di vista predittivo



- 3) $S \notin \mathcal{H}(\Theta)$ e $\Delta = \bar{\Theta}$. Si ottiene il modello $H(\bar{\theta})$ miglior approssimante di S nelle classi di modelli $\mathcal{H}(\theta)$ scelte



- 4) $S \notin \mathcal{H}(\Theta)$ e Δ è costituito da più voci



IDENTIFICABILITÀ DEI MODELLI

Obbiamo visto che i metodi PEM sono ASINTOTICAMENTE CORRETTI. Anzitutto come ottieniamo le stime (nel caso ARX($m, p+1$))

$$y(t) = \frac{B(z)}{A(z)} u(t-1) + \frac{1}{A(z)} e(t) \quad e(t) \sim \text{distrn}(0, \sigma^2) \quad B(z) = b_0 + b_1 z^{-1} + \dots + b_p z^{-p}$$

$$\hat{\theta}_N = \left[\sum_{t=1}^N \varphi(t) \varphi(t)^T \right]^{-1} \cdot \left[\sum_{t=1}^N \varphi(t) y(t) \right] = (\Phi^T \Phi)^{-1} \Phi^T \gamma$$

PROBLEMA DI IDENTIFICABILITÀ

Quando $\hat{\theta}_N$ esiste ed è unica? \Leftrightarrow Quando $\sum_{t=1}^N \varphi(t) \varphi(t)^T$ è invertibile?

$$S(N) = \sum_{t=1}^N \varphi(t) \varphi(t)^T \Rightarrow \hat{\theta}_N = S(N)^{-1} \cdot \left[\sum_{t=1}^N \varphi(t) y(t) \right]$$

$$R(N) = \frac{1}{N} S(N) = \frac{1}{N} \sum_{t=1}^N \varphi(t) \varphi(t)^T \Rightarrow \hat{\theta}_N = R(N)^{-1} \left[\frac{1}{N} \sum_{t=1}^N \varphi(t) y(t) \right]$$

$R(N)$ è ≥ 0 in quanto prodotto di un vettore per se stesso. Offrirebbe $\hat{\theta}_N$ esiste solo se tutte le componenti sono positive.

Consideriamo il caso ASINTOTICO $N \rightarrow +\infty \Rightarrow R(N) \xrightarrow[N \rightarrow \infty]{\delta} \bar{R}$

Per un ARX($m, p+1$), \bar{R} è una matrice quadrata $(m+p+1) \times (m+p+1)$ t.c.:

$$\bar{R} = \begin{bmatrix} \bar{R}_y & \cdots & -\bar{R}_{yu} \\ \vdots & \ddots & \vdots \\ -\bar{R}_{uy} & \cdots & \bar{R}_u \end{bmatrix} \in \mathbb{R}^{(m+p+1) \times (m+p+1)}$$

$$\bullet \text{ARX}(1,1) \Rightarrow \varphi(t) = \begin{bmatrix} -y(t-1) \\ u(t-1) \end{bmatrix} \Rightarrow R(N) = \frac{1}{N} \sum_{t=1}^N \varphi(t) \varphi(t)^T = \frac{1}{N} \sum_{t=1}^N \begin{bmatrix} -y(t-1) \\ u(t-1) \end{bmatrix} \begin{bmatrix} -y(t-1) & u(t-1) \end{bmatrix}$$

$$= \frac{1}{N} \sum_{t=1}^N \begin{bmatrix} -y(t-1)^2 & -y(t-1)u(t-1) \\ -u(t-1)y(t-1) & u(t-1)^2 \end{bmatrix}$$

Se $N \rightarrow +\infty$ obbiamo (dato da l'stimatore della covarienza è ASINT. CORRETTO)

$$\bar{R} = \begin{bmatrix} E[y(t-i)^2] & -E[u(t-i)y(t-i)] \\ -E[u(t-i)y(t-i)] & E[u(t-i)^2] \end{bmatrix}$$

IPOTESI
 $u(t)$ e $y(t)$ sono PSS

$$= \begin{bmatrix} \gamma_y(0) & -\bar{R}_{yu} \\ -\bar{R}_{uy}(0) & \gamma_u(0) \end{bmatrix}$$

$\rightarrow \bar{R}_y$
 $\rightarrow -\bar{R}_{yu}$
 $\rightarrow -\bar{R}_{uy}$
 $\rightarrow \bar{R}_u$

\bar{R} è la matrice di VARIANZE-COVARIANZE

caso generale:

$$\bar{R}_y = \begin{bmatrix} \gamma_y(0) & \gamma_y(1) & \cdots & \gamma_y(m-1) \\ \gamma_y(1) & \gamma_y(0) & \gamma_y(1) & \\ | & \gamma_y(1) & & \\ \gamma_y(m-1) & & \gamma_y(0) & \end{bmatrix}$$

- Matrice $m \times m$
- Matrice covarianza di ordine $m-1$ di $y(t)$
- Toeplitz

$$\bar{R}_u = \begin{bmatrix} \gamma_u(0) & \gamma_u(1) & \cdots & \gamma_u(p) \\ \gamma_u(1) & \gamma_u(0) & \gamma_u(1) & \\ | & \gamma_u(1) & & \\ \gamma_u(p) & & \gamma_u(0) & \end{bmatrix}$$

- Matrice $(p+1) \times (p+1)$
- Matrice covarianza ordine p di $u(t)$
- Toeplitz

$$\bar{R}_{yu} = \begin{bmatrix} \gamma_{yu}(0) & \gamma_{yu}(1) & \cdots & \gamma_{yu}(p) \\ \gamma_{yu}(1) & \gamma_{yu}(0) & \gamma_{yu}(1) & \\ | & \gamma_{yu}(1) & & \\ \gamma_{yu}(m-1) & & \gamma_{yu}(0) & \end{bmatrix}$$

- Matrice rettangolare $m \times (p+1)$
- Matrice covarianza $u(t)$ e $y(t)$
- $\bar{R}_{uy} = \bar{R}_{yu}^T$

Vogliamo una condizione per l'invertibilità di \bar{R}

Lemme di Schur

Dato una matrice H nella forma $H = \begin{bmatrix} F & K \\ K^T & H \end{bmatrix}$, con F e H simmetriche
condizione necessaria e sufficiente affinché $H > 0$ è che: - $H > 0$
- $F - KH^T K^T > 0$

$$R = \begin{bmatrix} -R_y & -R_{yu} \\ -R_{uy}^T & R_u \end{bmatrix} \Rightarrow \boxed{\begin{array}{l} \text{CONDIZIONE NECESSARIA} \\ R_u > 0 \end{array} \text{ per invertire } R \text{ è che}}$$

le seconda condizione ($F - KH^T K^T > 0$) è difficile da
dutone. Cerciamo d'imporsi obnre la prima

La condizione $R_u > 0$ riguarda solo l'impulsore $u(t)$, che progettiamo noi!
Sia:

$$\bar{R}_u^{(i)} = \begin{bmatrix} \delta u(0) & \delta u(1) & \cdots & \delta u(i-1) \\ \delta u(1) & \ddots & & \delta u(i-2) \\ \vdots & & \ddots & \\ \delta u(i-1) & & & \delta u(0) \end{bmatrix}$$

la matrice d'osservazione di $u(t)$, d'ordine i :

Il segnale $u(t)$ è detto PERSISTENTEMENTE ERITANTE DI ORDINE m se:

- $\bar{R}_u^{(1)} > 0, \bar{R}_u^{(2)} > 0, \dots, \bar{R}_u^{(m)} > 0$
- $\bar{R}_u^{(m+1)} > 0, \bar{R}_u^{(m+2)} > 0, \dots \geq 0$

ove m è l'ordine massimo di
 $\bar{R}_u^{(i)}$ per cui questa matrice è
invertibile



Condizione NECESSARIA per l'identificabilità di un modello ARK(m, p_+) è
che il segnale $u(t)$, usato per produrre i dati, sia "persistente eritante"
di ordine più ad almeno p_+ .

Osservazione

Consideriamo $u(t) \sim WN(0, \gamma^2)$. Abbiamo che:

$$\bar{R}_u^{(1)} = \begin{bmatrix} \gamma^2 & 0 & 0 & -0 \\ 0 & \gamma^2 & & \\ 1 & & \gamma^2 & \\ 0 & & & \gamma^2 \end{bmatrix} = \gamma^2 I^{(4)}$$

- Il WN è un segnale persistentemente eccitante di ordine ∞
- Se usiamo un WN per identificare il sistema, siamo certi che è un segnale sufficientemente ricco di informazioni per poter identificare il sistema
- Il WN eccita tutte le frequenze, avendo un spettro piatto

Osservazione

La condizione vista è solo NECESSARIA. Anche con $u(t) \sim WN$ la \bar{R} potrebbe non essere invertibile.

Esempio

- Il sistema vero è ARX(1,1)
 - Lo stimiamo con $u(t) \sim WN$
 - Il modello usato è ARX(3,3)
- Dato che il modello è SOVRAPARAMETRIZZATO esistono ∞ soluzioni (tutte le cancellazioni pol.-zer.)

SCEGLTA COMPLICATITÀ DEL MODELLO

Affinché un modello sia univocamente identificabile occorre avere:

- 1) IDENTIFICABILITÀ "STRUTTURALE": il modello non deve essere sovrapparametrizzato rispetto al sistema
- 2) IDENTIFICABILITÀ "SPERIMENTALE": i dati devono contenere sufficiente informazione

VALUTAZIONE DELL'INCERTEZZA

Le analisi fatte in precedenza si basavano sull'ipotesi che $N \rightarrow +\infty$. Nelle realtà otteniamo N finiti.

IPOTESI

- $S \in M(\theta) \Rightarrow \theta^* \in \Delta$
- Se $\Delta \ni \bar{\theta} \Rightarrow \bar{\theta} = \theta^*$

La varianza dello stimatore è:

$$\text{Var}[\hat{\theta}_N] = \frac{1}{N} \lambda^2 \cdot \bar{M}'$$

Stimiamo $\hat{\theta}_N$ con N dati.

$$\hat{\theta}_N \underset{\theta}{\underset{\text{arg min}}{\circlearrowleft}} J_N(\theta) = \frac{1}{N} \sum_{t=1}^N \epsilon(t, \theta)^2$$

→ E' UNA VARIANZA CASUALE!

Dalle ipotesi abbiamo che $E[\hat{\theta}_N] = \theta^*$

$$\bullet \lambda^2 = \text{Var}[\epsilon(t)] = \text{Var}\left[\underbrace{y(t)}_{\text{Var}[\epsilon(t, \theta^*)]} \cdot \hat{g}(t|t-1, \theta^*) \right]$$

$$\bullet \bar{M} = E\left[\left(\left. \frac{d\epsilon(t, \theta)}{d\theta} \right|_{\theta=\theta^*} \right) \cdot \left(\left. \frac{d\epsilon(t, \theta)}{d\theta} \right|_{\theta=\theta^*} \right)^T \right]$$

Come stimiamo in pratica λ^2 e \bar{M} ?

$$\boxed{\lambda^2 = E[\epsilon(t, \theta^*)] \approx E[\epsilon(t, \hat{\theta}_N)] \approx \frac{1}{N} \sum_{t=1}^N (y(t) - \hat{g}(t|t-1, \hat{\theta}_N))^2 = \boxed{J_N(\hat{\theta}_N)}}$$

$$\bar{M} \approx \hat{M} = \frac{1}{N} \sum_{t=1}^N \left[\left(\left. \frac{d\epsilon(t, \theta)}{d\theta} \right|_{\theta=\hat{\theta}_N} \right) \cdot \left(\left. \frac{d\epsilon(t, \theta)}{d\theta} \right|_{\theta=\hat{\theta}_N} \right)^T \right]$$

Interpretazione di \bar{M}

$$\begin{aligned} &\text{Ricordiamo che } J(\theta) : E[\epsilon(t, \theta)^2] \Rightarrow \frac{dJ(\theta)}{d\theta} = E\left[2\epsilon(t, \theta) \frac{d\epsilon(t, \theta)}{d\theta} \right] \\ &\Rightarrow \frac{d^2 J(\theta)}{d\theta^2} = E\left[2 \frac{d\epsilon(t, \theta)}{d\theta} \cdot \frac{d\epsilon(t, \theta)}{d\theta}^T + 2\epsilon(t, \theta) \frac{d^2 \epsilon(t, \theta)}{d\theta^2} \right] \end{aligned}$$

Se $\theta = \theta^* \Rightarrow \epsilon(t, \theta) = e(t) \Rightarrow \frac{d^2 \epsilon(t, \theta)}{d\theta^2}$ è funzione dell'errore di predizione
e dipende da $e(t-1), e(t-2)$

↓ quindi $\frac{d\epsilon(t)}{d\theta}$:

$$\frac{d(y(t) - \hat{g}(t|t-1, \theta))}{d\theta} \rightarrow \text{dipende da } e(t-1)$$

↓ predittore
controllare con $e(t-1)$

$$E(t) = y(t) - \hat{g}(t|t-1) = e(t)$$

34

↓ dipende da $e(t-1)$, -

per termine $E \left[2e(t) \frac{d^2 E(t)}{d\theta^2} \right] = E \left[2e(t) \cdot \frac{d^2 E(t)}{d\theta^2} \right] = 0$

Osserviamo che:

$$\left. \frac{d^2 J(\theta)}{d\theta^2} \right|_{\theta=0^\circ} = 2E \left[\left(\left. \frac{d E(t, \theta)}{d\theta} \right|_{\theta=0^\circ} \right) \cdot \left(\left. \frac{d E(t, \theta)}{d\theta} \right|_{\theta=0^\circ} \right)^T \right] = 2\bar{H}$$

$$\downarrow$$

$$\bar{H} = \frac{1}{2} \left. \frac{d^2 J(\theta)}{d\theta^2} \right|_{\theta=0^\circ}$$

- \bar{H} è metà dell'Hessiano delle cifre di monitor volutamente nell'ottimo θ^*

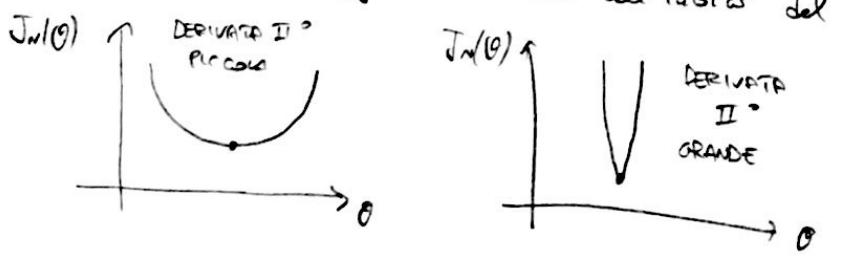
Conclusione

$$Var[\hat{\theta}_N] = \frac{1}{N} \cdot \sigma^2 \cdot \bar{H}^{-1} = \frac{1}{N} \cdot \sigma^2 \cdot \frac{1}{2} \left. \frac{d^2 J(\theta)}{d\theta^2} \right|_{\theta=0^\circ}$$

- $N \uparrow \Rightarrow Var[\hat{\theta}_N] \downarrow$

- $\sigma^2 \Rightarrow Var[\hat{\theta}_N] \uparrow$

- $\bar{H} \uparrow \Rightarrow Var[\hat{\theta}_N] \downarrow \Rightarrow$ Si vuole avere Hessiano grande nel punto di minimo
(grande varianza all'interno del minimo)



SCELTA DELLA COMPLESSITÀ DEL MODELLO

Scelta una classe di modelli $H(\theta)$, trovare $\hat{\theta}_N$ nel seguente modo:

$$\hat{\theta}_N = \underset{\theta}{\operatorname{arg\,min}} J_N(\theta) \quad J_N(\theta) = \frac{1}{N} \sum_{t=1}^N (y(t) - \hat{y}(t|t-1))^2$$

- Se $H(\theta)$ è un ARX \Rightarrow esiste una forma esplicita per $\hat{\theta}_N$
- Se $H(\theta)$ è un ARMAX \Rightarrow va risolti iterativamente un problema di ottimizzazione

[d] dimensione di θ .

Nel caso generale ARMAX abbiamo $d = m + n + p + 1$

\downarrow
Problema: scelta dell'ordine d del modello

Osservazione

In realtà dobbiamo scegliere 3 parametri, con 1° per semplicità, si fissa $m = n = p$, gestendo quindi un solo parametro, d

Sappiamo già che non è corretto scegliere l'ordine del modello con minor

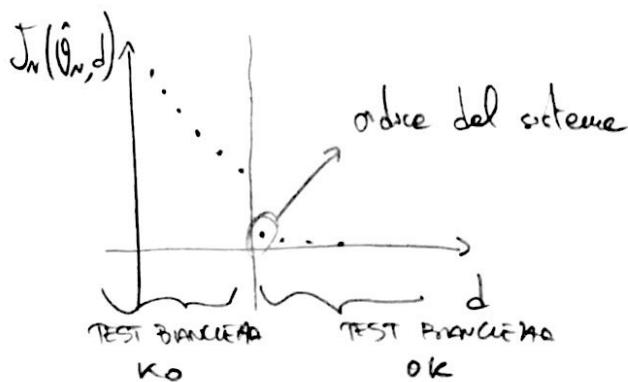
$J_N(\hat{\theta}_N) \Rightarrow \underline{\text{OVERFITTING}}$

Vediamo quindi 3 metodi, che si basano sull'area:

- N dati
- ordine del modello $d = 1, 2, 3, 4, \dots$
- $J_N(\hat{\theta}_N, d)$ è il valore della area da minima volutaria all'“ottimo” (ovvero dopo aver trovato il miglior modello di dimensione d)

METODO 1 : TEST DI BIANCHETTA

- Procedure.
- Fissare un valore di d
 - Trovare $\hat{\theta}_N$
 - calcolare $J_N(\hat{\theta}_N, d)$
 - Test di bianchetta su $\epsilon(t, \hat{\theta}_N) = y(t) - \hat{y}(t|t-1, \hat{\theta}_N)$
 - Ripetere per $d = 1, 2, 3, 4, \dots$

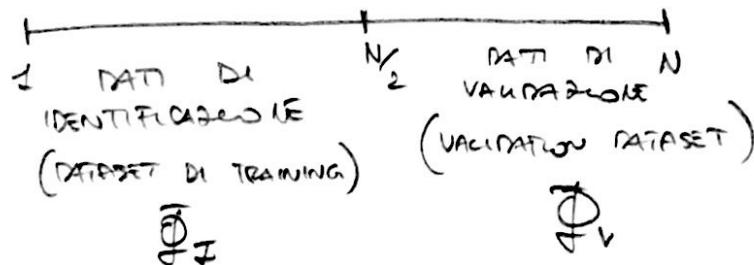


Un缺点 questo metodo funziona anche

- Permette le "discontinuità"
- Il test di bianchetta ha dei punti di discontinuità retta (da un "rayo" di valori)

METODO 2: CROSS-VACCINAZIONE

Dividere i dati in 2 sottoinsiemi

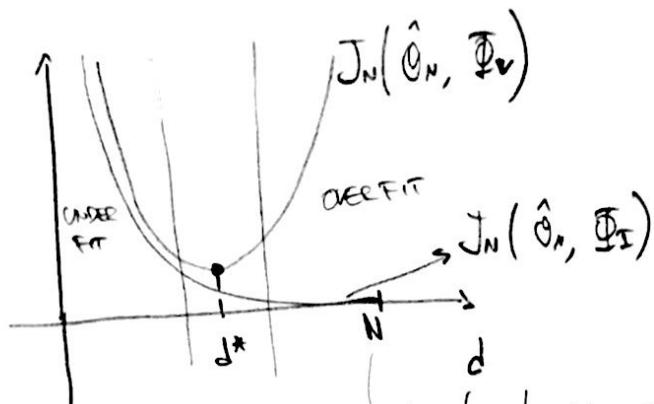


- Fissare un valore di d
- Trovare $\hat{\theta}_N$ minimizzando $J_N(\theta, D_I)$
- Calcolare $J_N(\hat{\theta}_N, D_I)$ e $J_N(\hat{\theta}_N, D_V)$
- Ripetere per $d = 1, 2, 3, 4, \dots$

Note

Quell'errore è $J_N(\theta, D_I)$ e
 $J_N(\theta, D_V)$ è dipendente da d

Le differenze rispetto al caso statico è che non possiamo mischiare i dati,
perché adesso lo sono dipendente dal tempo



quando $d = N$, interpola i dati perfettamente

La cross-validation è la procedura migliore, però richiede molti dati

METODO 3: FORMULE PER LA STIMA DELLA COMPLICATEZZA OTTIMA

Potremmo di stimare out-of-sample error senza usare dati diversi rispetto ai dati di identificazione (di train) \Rightarrow modificare la cifra di merito e lo minimizzare

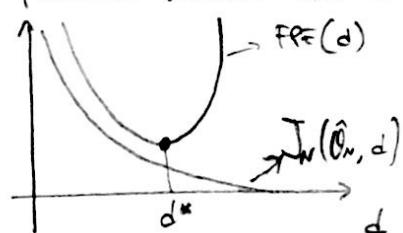
Sono state ricavate tecnicamente per ARX, ma in pratica si usano anche per ARMA

FINAL PREDICTION ERROR (FPE)

$$FPE(d) = \frac{N+d}{N-d} \cdot J_N(\hat{\theta}_N, d)$$

- $d \uparrow \Rightarrow \frac{N+d}{N-d} \uparrow$
- $d \uparrow \Rightarrow J_N(\hat{\theta}_N, d) \downarrow$

Il metodo produce modelli con d grandi



AKAIKE INFORMATION CRITERION (AIC)

$$AIC(d) = 2 \cdot \frac{d}{N} + \ln [J_N(\hat{\theta}_N, d)]$$

- $d \uparrow \Rightarrow 2 \cdot \frac{d}{N} \uparrow$
- $d \uparrow \Rightarrow \ln [J_N(\hat{\theta}_N, d)] \downarrow$

• MINIMUM DESCRIPTION LENGTH (MDL)

$$MDL(d) = \ln(N) \cdot \frac{d}{N} + \ln[J_N(\hat{\theta}_N, d)]$$

$$\bullet d \uparrow \Rightarrow \ln(N) \cdot \frac{d}{N} \uparrow$$

$$\bullet d \uparrow \Rightarrow \ln[J_N(\hat{\theta}_N, d)] \downarrow$$

CONFRONTO FPE vs AIC

I criteri sono simili. Se si calcola il logaritmo di FPE si ottiene:

$$\begin{aligned} \ln[FPE(d)] &= \ln\left[\frac{N+d}{N-d} \cdot J_N(\hat{\theta}_N, d)\right] = \ln\left[\frac{1+\frac{d}{N}}{1-\frac{d}{N}} \cdot J_N(\hat{\theta}_N, d)\right] = \\ &= \ln\left[\frac{1+\frac{d}{N}}{1-\frac{d}{N}}\right] + \ln[J_N(\hat{\theta}_N, d)] = \\ &= \ln\left[1 + \frac{d}{N}\right] - \ln\left[1 - \frac{d}{N}\right] + \ln[J_N(\hat{\theta}_N, d)] \end{aligned}$$

Ricordiamo che $\ln(1+x) \approx x$ quando $x \gg 0$; inoltre $\frac{d}{N} \gg 0$ per avere overfitting.

$$\begin{aligned} \ln[FPE(d)] &\approx \frac{d}{N} - \left(-\frac{d}{N}\right) + \ln[J_N(\hat{\theta}_N, d)] \\ &\approx 2 \frac{d}{N} + \ln[J_N(\hat{\theta}_N, d)] = \boxed{AIC(d)} \end{aligned}$$

Quindi, se $d \ll N$ $\Rightarrow \boxed{\ln[FPE(d)] = AIC}$ come per le z-score

Il minimo di $f(x)$ è anche il minimo di $\ln[\text{Pca}]$, quindi i metodi sono equivalenti

CONFRONTO AIC vs MDL

$$AIC(d) = 2 \cdot \frac{d}{N} + \ln[J_N(\hat{\theta}_N)] \quad \Leftrightarrow \quad MDL(d) = \ln(N) \cdot \frac{d}{N} + \ln[J_N(\hat{\theta}_N)]$$

> considerare solo
quegli termini

Se $\ln(N) > 2$, ovvero osservare più di 8 dati, MDL penalizza di più, e quindi preferisce di usare modelli più parsimoniosi.

HDL quindi diminuisce il rischio di fare overfitting (e soprattutto un
maggiore rischio di fare underfitting)



- Nel caso in cui la scelta delle famiglie di modelli è sicura, è possibile usare AIC = FPE, in quanto il rischio di fare overfitting è minore
- Nel caso in cui non si conosce nulla del sistema, è meglio usare HDL, essendo più robusto e overfittig

—————
FINE