

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 luoghi 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)

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

$$P = M \cdot g$$

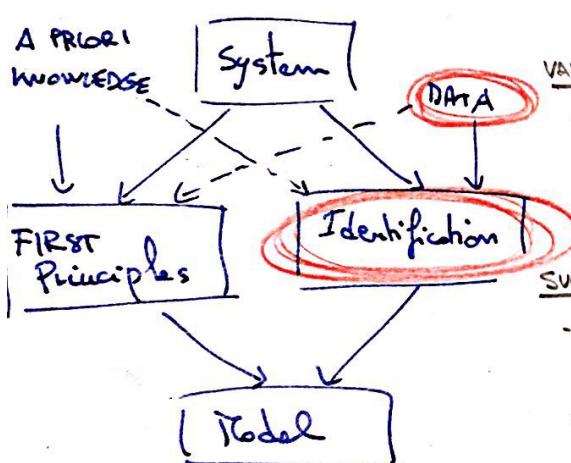
$$V = R \cdot I$$

conoscenza
A priori
presente

Due approcci fondamentali:

a) WHITE BOX MODELING: - approccio basato su leggi e principi base delle FISICA e MATEMATICA

$$\text{es. modello di un condensatore } I(t) = C \cdot \frac{dV(t)}{dt}$$



VANTAGGI

- conoscenza del significato delle variabili: (C: capacità di condensare)
- generalizzabile: se conosce C, il modello vale anche per le altre cause che relazione CAUSALE tra C ed I(t)

Svantaggi

- richiede conoscenza avanzata delle leggi del problema specifico
- Reitti, 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

- prescinde 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 a determini che AI SOLI individui uno 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 ad 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

Lsistemi statici: la sola conoscenza delle variabili u è sufficiente a calcolare $y \Rightarrow V(t) = R \cdot I(t)$

Lsistemi 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)



SISTEMI STATICI

STIMA PARAMETRICA → modelli matematici che descrivono e stima le
variabili statistiche ↓

- PROPRIETA' STIMATORI

- ESTRUCTURA PARALELICA
- PROCESO STOCÁSTICO
- PROPUESTA ESTIMATORIA

deterministic (um numeros/lettre)

- Shiva di posizioni di una popolazione: $y = \frac{1}{2} \ln P$
 - PREZIONE (o un po' più)

- Approssimazione lineare a y : lineare - regressione lineare
- $y = \beta^T x + \epsilon$
- No Assumption on PDF
- Assumption on error
- Massima Verosimiglianza \rightarrow Max Likelihood

- Massima Versus Infrastruttura -> gli Assunzioni PDF.

- Laser
 - Laser
 $y = g_x^T + \epsilon$
 - Regression Line

- Regresión Logística \Rightarrow si ASUMOON PDF

SISTEMI DINAMICI

- Regressão Logística

Caso 2) O monobife casula

-Shima Boyce

CONCETTI DI MACHINE LEARNING

- INTRO
 - PLAS/VARIANCE
 - REGULATION

(3)

 - VARIATION

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

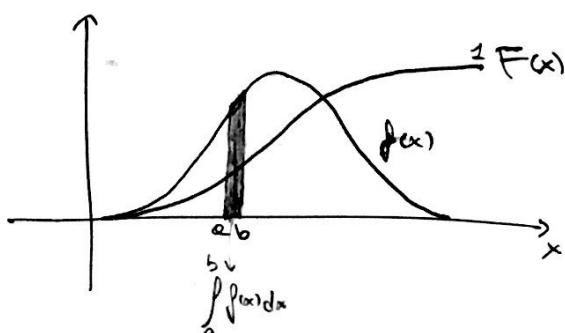
$$P(V \in [a, b]) = \int_a^b f(x) dx$$

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

- 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
 $f(x_i) \propto P(V=x_i)$

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

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

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

$$\forall \beta \in \mathbb{R}$$

- 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 $F(x_1, x_2, x_3, \dots, x_d) = P(V_1 \leq x_1, V_2 \leq x_2, \dots, V_d \leq x_d)$

$$= \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_d} f(x_1, x_2, \dots, x_d) dx_1 dx_2 \dots dx_d$$

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:

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

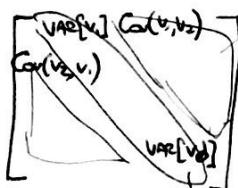
L'insieme $\{x \in \mathbb{R}^d \mid x \geq 0\}$ per numeri reali

L'insieme $\{M \in \mathbb{R}^{d \times d} \mid M \text{ simmetrica e semidefinita positiva}\}$ per numeri reali

L'autosvalore > 0 tranne $\{0\} = 0$

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

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



$$d \times d$$

\rightarrow VARIANZE DI v_1, v_2, \dots, v_d

- COVARIANZE TDI

$$v_1 \in \mathbb{R}, v_2 \in \mathbb{R}, v_3 \in \mathbb{R}$$

- SIMMETRICITÀ: 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

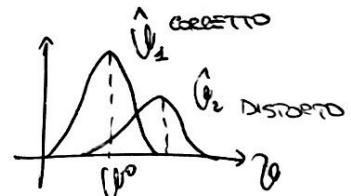
Es modo de fare MAD $\hat{\theta}_1 = T(D(s_1, \theta^*))$

Misura il peso degli studenti $\hat{\theta}_1 = T(D(s_1, \theta^*))$
L'intero s può misurare solo 5 studenti $D = \{y(1) - y(5)\} \Rightarrow \hat{\theta}_2 = T(D(s_2, \theta^*))$

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

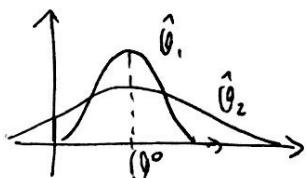


- Un stimatore si dice asintoticamente corretto se e solo se: $\lim_{N \rightarrow \infty} E[\hat{\theta}] = \theta^*$

- è una proprietà più debole.

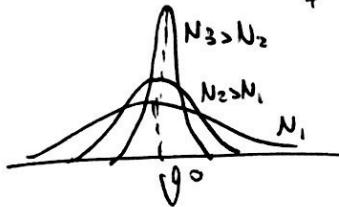
PRIMA DEF. CONSISTENTE

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, $s_{N-1}^2 = \frac{\sum_{i=1}^N (y_i - \hat{\mu})^2}{N-1}$.

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

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

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

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

$$= \frac{N}{N-1} \left(s_{N-1}^2 + E[y]^2 \right) = \frac{N}{N-1} \left((N-1) \frac{\text{Var}[y]}{N} + E[y]^2 \right) = \text{Var}_{\text{eff}}[y] = \boxed{\sigma^2}$$

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 infatti se un} ^{è grande 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 variabili x_1, x_2, \dots, x_d

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

N persone misurate

Definiamo: vettori:

$$\begin{aligned} \boldsymbol{\beta} &= \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix} & \boldsymbol{\varphi}(i) &= \begin{bmatrix} 1 \\ x_1(i) \\ x_2(i) \\ \vdots \\ x_d(i) \end{bmatrix} & \Rightarrow & \boxed{\begin{aligned} y(i) &= \boldsymbol{\varphi}(i)^T \cdot \boldsymbol{\beta} + e(i), \quad i=1 \dots N \\ & \vdots \quad \vdots \quad \vdots \end{aligned}} \end{aligned}$$

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

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

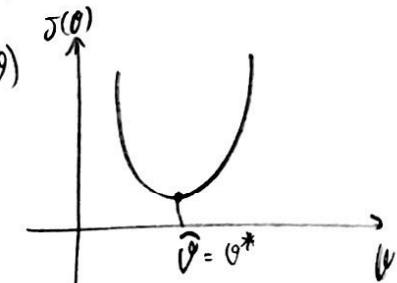
$$\Rightarrow \left[\sum_{i=1}^N \varphi_{(i)} \varphi_{(i)}^\top \right] \theta = \sum_{i=1}^N \varphi_{(i)} y_{(i)} \Rightarrow \boxed{\hat{\theta} = \left[\sum_{i=1}^N \varphi_{(i)} \varphi_{(i)}^\top \right]^{-1} \left[\sum_{i=1}^N \varphi_{(i)} y_{(i)} \right]}$$

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

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

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix}_{d+1}$$

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

↳ ogni colonna è un rappresentante / features

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

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

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

$$\nabla_\theta (J(\theta)) = (A + A^\top)\theta \quad \perp (Y^\top Y - 2\theta^\top X^\top Y + \theta^\top X^\top X\theta)$$

$$\nabla_\theta (J(\theta)) = b \quad \text{con} \quad b = \frac{(X\theta)^\top}{d+1} = \frac{X^\top X\theta}{d+1}$$

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

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

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

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

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

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

STIMA A MASSIMA VEROSSIGLIANZA

Ottieni presenti fior od ora diversi tipi di stimatori,

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

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

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

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

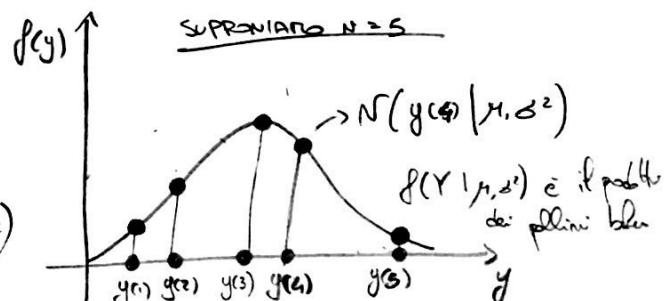
Ottieni presenti stimatori PARAMETRICI, aendo rappresentati i dati tramite un modello parametrico (es. modello lineare)

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

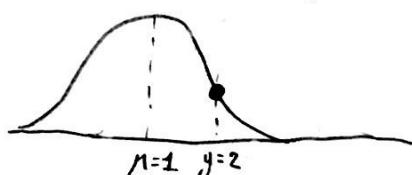
L' se im funz. della y, è una pdf N-dimensionale

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

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

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

NUMERO NOTO
FUNZIONE DI
y

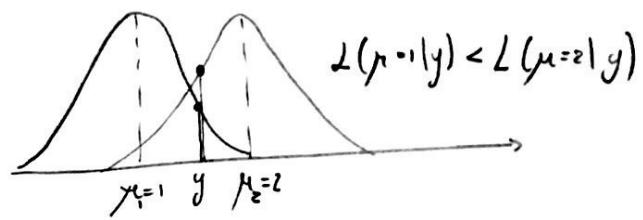


$$\Rightarrow L(\mu, \sigma^2 | y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{(y-\mu)^2}{\sigma^2}\right)}$$

(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 $L(\theta | y)$

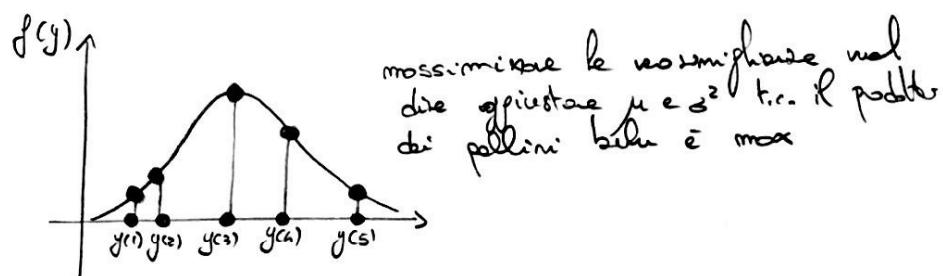
L'ad esempio, $\theta = \mu$ (σ^2 noto) $\Rightarrow \mu = 2$ è più verosimile di $\mu = 1$ perché $f(y | \mu = 1) < f(y | \mu = 2)$

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 per 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 di 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'indebolire 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 le stime max verosimiglianza di $\theta = [\mu, \sigma^2]$

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

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

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

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

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

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

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

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

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

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 distorta!

↓ 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 i dati osservati sono:

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

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

La densità 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:

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

Trovare 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

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

$$\begin{aligned} \mathcal{L}(\pi | Y) &= \prod_{i=1}^N \pi^{y(i)} \cdot (1-\pi)^{1-y(i)} = \pi^{\sum_{i=1}^N y(i)} \cdot (1-\pi)^{\sum_{i=1}^N (1-y(i))} \\ &= \pi^{\text{# success}} \cdot (1-\pi)^{\text{# failure}} \end{aligned}$$

Colours to be by - season, please

$$\begin{aligned}
 \ln L &= \ln \left[\pi \sum_{i=1}^N y_{(i)} \cdot (1-\pi) \sum_{i=1}^N (1-y_{(i)}) \right] = \ln \pi \sum_{i=1}^N y_{(i)} + \ln (1-\pi) \sum_{i=1}^N (1-y_{(i)}) \\
 &= \underbrace{\sum_{i=1}^N y_{(i)}}_{\text{no. of successes}} \cdot \ln \pi + \underbrace{\sum_{i=1}^N (1-y_{(i)})}_{\text{no. of failures}} \ln (1-\pi) \quad \left\{ \begin{array}{l} = \gamma \ln \pi + (N-\gamma) \ln (1-\pi) \\ \sum_{i=1}^N 1 - \sum_{i=1}^N y_{(i)} = N - \gamma \end{array} \right. \\
 &\text{w.r.t. } \pi
 \end{aligned}$$

Tower R Wossman

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

$$\Rightarrow \bar{y} - \bar{y}\bar{x} - \pi N + \bar{y}\bar{x} = 0 \Rightarrow \boxed{\bar{x} = \frac{\bar{y}}{N} = \frac{1}{N} \sum_{i=1}^n y(i)}$$

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 continuo 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 posso 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'non ha senso sommare un errore continuo (ϵ_R) ad una variabile y che può assumere soli valori come 0 e 1, e non 0,98 o 1,01.

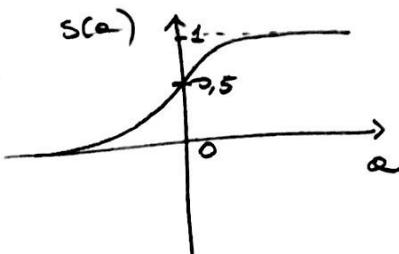
L'il modello potrebbe prevedere anche valori <0 o >1! Non c'è niente che "limite" l'uscita \hat{y} tra 0 ed 1.



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 $D = \{(\varphi(1), y_{(1)}), (\varphi(2), y_{(2)}), \dots, (\varphi(N), y_{(N)})\}$
 $\varphi(i) \in \mathbb{R}^{d_{\text{var}}}$, dove $y_{(i)} \in \{0, 1\}$, $i=1, \dots, N$, iid

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

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

Calcoliamo la Verosimiglianza dei dati

$$P(y_{(i)}=1 | \varphi_{(i)}) = \frac{1}{1+e^{-(\varphi_{(i)}^T \theta)}} \equiv \pi_{(i)}$$

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

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

dipende dai
parametri θ !!

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

i vari parametri
sono questi

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

Interpretazione della funzione di costo

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

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

CASE $y=1$

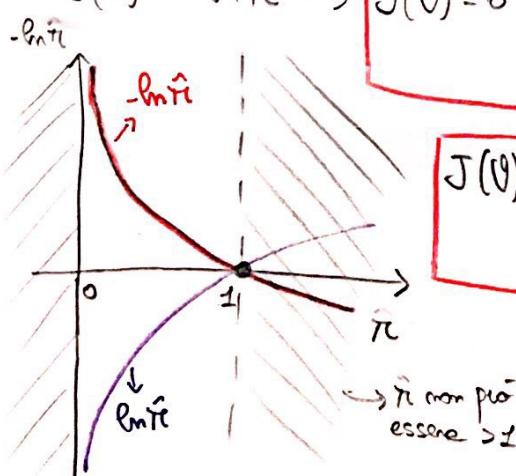
Costo \Rightarrow se predico giusto

$$J(\theta) = -\ln \pi \Rightarrow J(\theta) = 0 \quad \text{SE } y=1 \quad \& \quad \pi = 1$$

Cuttura l'intuizione che se $y=1$, non
è predire una bassa probabilità che $y=1$,
ma è predire una bassa probabilità che $y=0$,

ovvero predico $\text{re}\ll 1$ ($P(y=1|\varphi) \ll 1$)
Allora commetto un grande sbaglio e
 $J(\theta) \Rightarrow +\infty$ (perdono molto)

Io voglio minimizzare questo sbaglio!



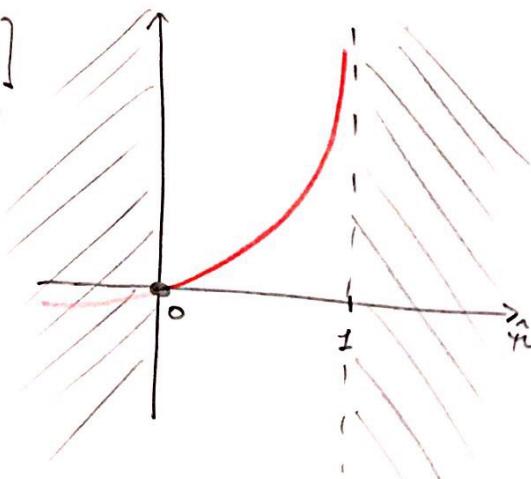
π non può essere > 1

(21)

CASO $y=0$

$$-\ln[1-\pi]$$

$$J(\theta) = -\ln[1-\pi]$$



$$J(\theta) = 0 \quad \text{SE} \quad p = 0 \\ \frac{d}{d\theta} J(\theta) = 0$$

$$J(\theta) = +\infty \quad \text{SE} \quad y = 0 \\ \frac{d}{d\theta} J(\theta) = +\infty$$

Se $y=0$ non si predice con alte probabilità che $y=1$, ovvero predice $\pi \gg 0$ ($P(y=1|p) \gg 0$) allora sbaglieri molto e $J(\theta) \rightarrow +\infty$.

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

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

uguale formazione

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

$$\begin{aligned} \frac{ds(\varphi^\top \theta)}{\theta} &= \frac{d}{d\theta} \left[\frac{1}{1+e^{-\varphi^\top \theta}} \right] = \frac{d}{d\theta} \left[(1+e^{-\varphi^\top \theta})^{-1} \right] = \underset{dx_1}{\varphi_1} \cdot \underset{dx_1}{(-1)} \underset{dx_1}{(1+e^{-\varphi^\top \theta})^{-2}} \underset{dx_1}{(e^{-\varphi^\top \theta})} \\ &= -\varphi \cdot (1+e^{-\varphi^\top \theta})^{-2} (e^{-\varphi^\top \theta}) = \underset{\text{stesso passaggio}}{\underset{\text{Prima}}{=}} \underset{dx_1}{\varphi} \cdot \underset{dx_1}{s(\varphi^\top \theta)} \underset{dx_1}{[1-s(\varphi^\top \theta)]} \\ &= \boxed{\varphi \cdot M \cdot (1-\pi)} \end{aligned}$$

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

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

$$\begin{aligned} \nabla J(\theta) &= -\sum_{i=1}^N \left(y_{(i)} \frac{\pi'_{(i)}}{\pi_{(i)}} + (1-y_{(i)}) \frac{-\pi'_{(i)}}{1-\pi_{(i)}} \right) = \\ &= -\sum_{i=1}^N \left(y_{(i)} \cdot \frac{e^{\theta^T x_{(i)}} \cdot \pi_{(i)} [1-\pi_{(i)}]}{\pi_{(i)}^2} + (1-y_{(i)}) \frac{-e^{\theta^T x_{(i)}} \pi_{(i)} [1-\pi_{(i)}]}{1-\pi_{(i)}} \right) \\ &= \sum_{i=1}^N \left(-y_{(i)} e^{\theta^T x_{(i)}} [1-\pi_{(i)}] - (1-y_{(i)}) (e^{\theta^T x_{(i)}} \cdot \pi_{(i)}) \right) \\ &= \sum_{i=1}^N \left(e^{\theta^T x_{(i)}} \left[-y_{(i)} + y_{(i)} \pi_{(i)} - y_{(i)} \pi_{(i)} + \pi_{(i)} \right] \right) \\ &= \sum_{i=1}^N \left(e^{\theta^T x_{(i)}} (\pi_{(i)} - y_{(i)}) \right) \\ &= \sum_{i=1}^N e^{\theta^T x_{(i)}} (\pi_{(i)} - y_{(i)}) \end{aligned}$$

Osservazione

Le derivate $\sum_{i=1}^N e^{\theta^T x_{(i)}} (\pi_{(i)} - y_{(i)}) = 0$ sono un sistema di $|\theta|$ equazioni non lineari in θ

↳ Non è immediato sapere se esiste una soluzione unica. Però si può provare a risolvere il sistema.

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

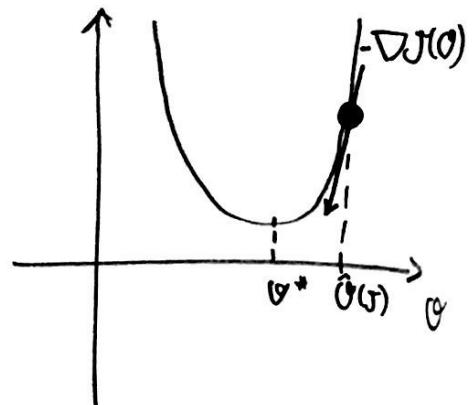
L'ottimizzazione è quindi solita utilizzare algoritmi iterativi di ottimizzazione. Uno di questi è il GRADIENT DESCENT:

↓
il valore nuovo dei parametri all'iterazione $j+1$ è:

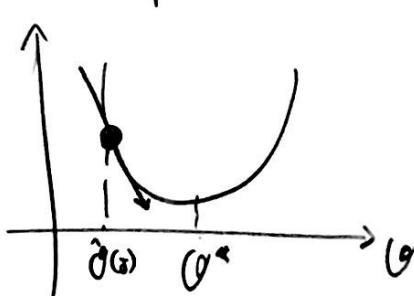
$\hat{\theta} \leftarrow \hat{\theta} - \alpha \nabla J(\theta)$ (dove α è la learning rate)

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

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



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



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
0	0,06	0,24
1	0,28	0,42

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

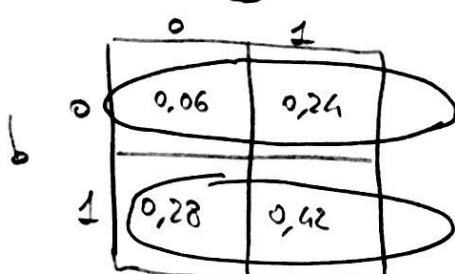
probabilità che si verifichi sia a che b , contemporaneamente

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

In questo 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



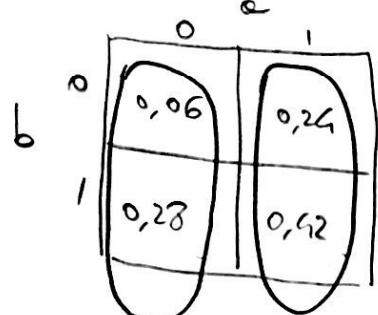
$$P(b=0) = 0,3$$

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

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

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

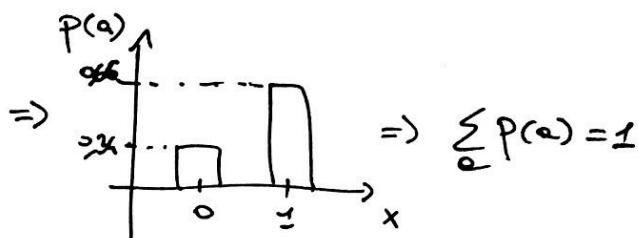
$$\Rightarrow \begin{array}{c} 0,3 \\ 0,7 \\ \hline 1 \end{array} \quad P(b=0, a=1) = P(b=0)$$



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

$$0,34$$

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



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

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

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

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

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

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

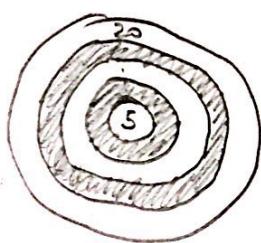
Oltre lo stesso modo possiamo calcolare $p(b|a)$.

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

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

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 sui possibili valori di θ , ovvero $P(\theta)$
- 2) L'informazione di potere i dati sui possibili valori di θ , ovvero le likelihood $P(Y|\theta)$

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

\Downarrow

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 in. Come posso dire su che distribuzione sarà $P(\theta|Y)$?
 - 1) In genere, nulla. Soltanto in casi fortunati la $P(\theta|Y)$ è in una forma funzionale nota
 - 2) Questo avviene se, ad esempio, $P(\theta)$ è Gaussiana e $P(Y|\theta)$ è Gaussiana. Allora $P(\theta|Y)$ sarà Gaussiana
 - 3) Un altro problema è che $P(Y)$ è un integrale da potremmo non sapere come risolvere.

\Downarrow

Per far fronte a questi problemi, si usano metodi numerici e di campionamento che evitano i calcoli analitici. 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)

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

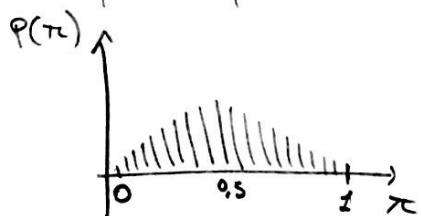
- Se $y=1 \Rightarrow P(Y=1|\pi) = \pi$

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

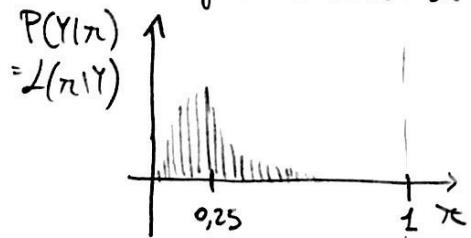
$$L(\pi|Y) = \prod_{i=1}^N \pi^{y^{(i)}} \cdot (1-\pi)^{1-y^{(i)}}$$

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

- Supponiamo una prior di questo tipo



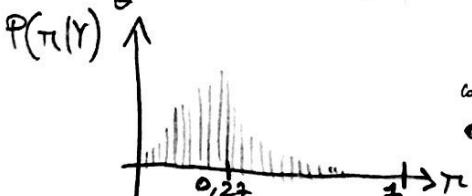
- Supponiamo di aver osservato $y=10$ successi su $N=40$ lanci. La likelihood ha la forma:



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

In questo caso $\hat{\pi} = \frac{10}{40} = 0,25$
 $P(Y|\theta) = \prod_{i=1}^N \pi^{y^{(i)}} \cdot (1-\pi)^{1-y^{(i)}}$ per ogni valore di θ

- Per calcolare la posterior faccio il quoziente di $P(Y|\theta)$ e $P(\theta)$ per ogni valore di θ e dividere per $P(Y) = \sum_{\theta} P(Y|\theta) P(\theta)$, che somma su ogni valore di θ (della griglia)



La NDA è un compromesso tra 0,25 e 0,5
 DIETRO \Rightarrow

(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 (*)

302

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 costante d'è risolvibile.

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

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

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:

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

- VARIANZA:

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

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

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

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

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 modificata 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 elevato 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 $E[y] = \mu_y \neq 0$

$E[y] = \mu_y \neq 0$

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

$$\lambda_{\theta y} \in \mathbb{R}^{m_y \times 1}$$

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

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

Oltre 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 FILTRO DI KALMAN, 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 $E[y]$, $\frac{\partial \log}{\partial \theta}$, $\frac{\partial y}{\partial \theta}$

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

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

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

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

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

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

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

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

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

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

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

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 •

Tutt'attorno a 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 ormoni in un individuo misurata ogni giorno alle stesse ore

Che problema vogliono risolvere?

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

$$\boxed{y(N+1|N)}$$

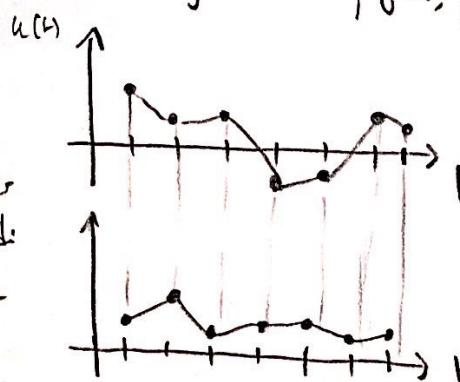
predico il valore a $t=N+1$ noti i dati fin al tempo N

• SISTEMI INGRESSO/USCITA •

Ottimare 2 insiemni di dati, uno di ingressi ed uno di uscite

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

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



Note

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

Esempio

INGRESSO (CURE)

comete

disostruzione medicale

mm di pioggia

(10)

USCITA (EFFETTI)

Coppe

Concentrazione ormoni

Concentrazione di

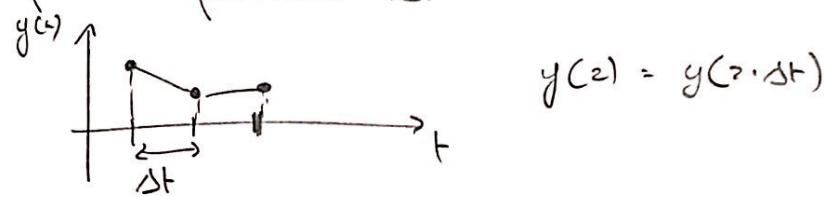
Che tipo di problemi vogliono risolvere?

1) PREDIZIONE: come prevedere

2) CONTROLLO: determinare le relazioni $u \rightarrow y$, in modo da progettare un controllore che determini $u(t)$

Osservazione

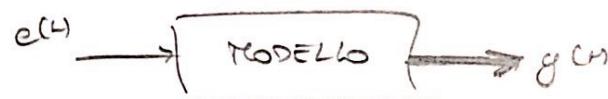
Proveremo con segnali e sistemi a tempo discreto. I segnali sono quindi composti con tempi di campionamento Δt



Come impostiamo il problema?

1 SERIE TEMPORALE

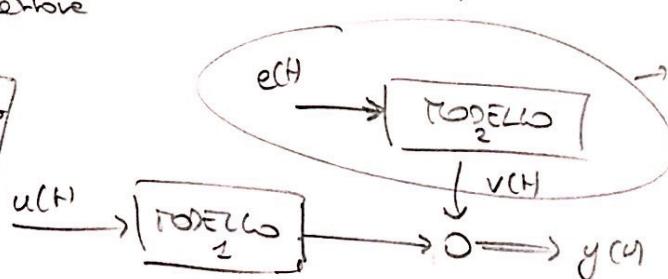
Modelliamo la serie temp. $y(t)$ come l'uscita di un sistema con imposto "rumore" non misurabile



SISTEMI I/O

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

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



- modello noto da $u(t)$ se riesce a spiegare dell'uscita $y(t)$
- rumore di misura
- errori di modello

Osservazione

Come nel caso dei sistemi statici, la y sarà affetta da rumore. In quel caso dovrà 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 composti da un segnale che evolue nel tempo

61

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

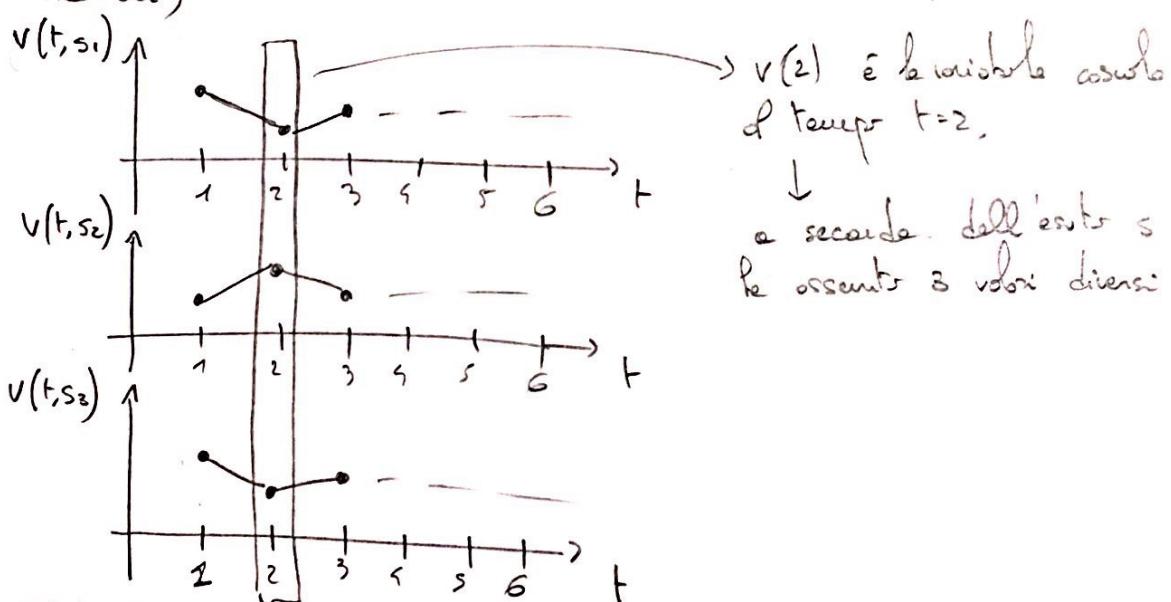
PROCESSI STOCASTICI

(INFINITO)

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, ottengo un'altra serie di valori
- Si può pensare ad un PS come ad un segnale (anche se il PS può generare diverse realizzazioni)



Note

Spesso ometteremo la dipendenza da s , indicando $v(1, \bar{s}), v(2, \bar{s}), \dots, v(3, \bar{s})$ con $v(1), v(2), v(3)$

Più facile interpretarono: dhi gci come variabili casuali per gestire l'incertezza delle loro misure

Adesso interpretiamo la serie di dhi $y(t)$ come realizzazione finita di un processo stocastico. Il segnale $y(t)$ è un segnale qualunque

\rightarrow STAZIONARIO

42

• 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

• COVARIANZA

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

- 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 :

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

La teoria di sviluppo si basa su un particolare tipo di ps.

• PROCESSI STOCASTICI STAZIONARI (PSS) +

Definizioni

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

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

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

$$2) \gamma(t_1, t_2) = \gamma(t_3, t_4) \quad \text{se} \quad |t_2 - t_1| = |t_4 - t_3| = n \quad \Rightarrow \begin{aligned} &\text{la covarianza dipende solo} \\ &\text{dal} \Delta t \text{ e non dai valori} \\ &\text{specifici di } t_1, t_2, t_3, t_4 \end{aligned}$$

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

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

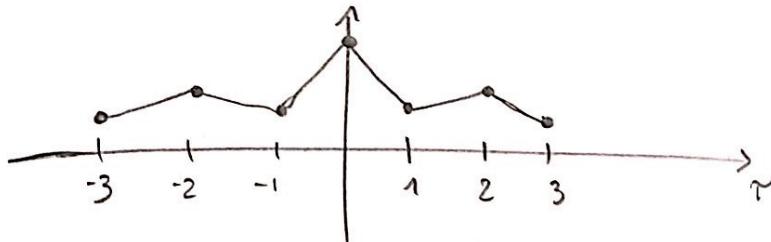
• è costante $\forall t$

PROPRIETÀ DELLA FUNZIONE 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 hanno le stesse valori attesi m e la stessa covarianza $\gamma(\tau) \quad \forall \tau$

Nota

Durante il corso studiavamo pss

CASO PARTICOLARE DI PSS: RUMORE BIANCO (WHITE NOISE)

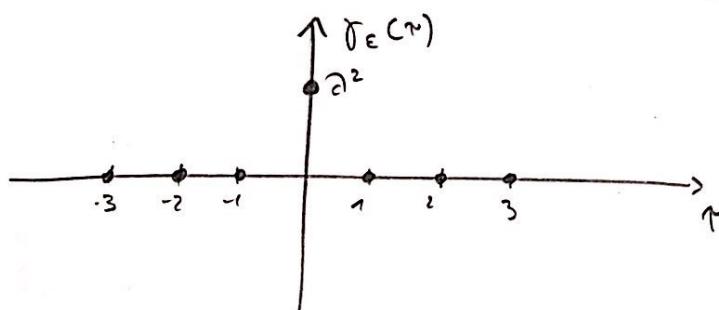
Definizione

Un pss $e(t)$ è detto ruido 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

(44)

Note

Ma è determinante la distribuzione delle singole vc. $\epsilon(t)$. Possono essere Gaussiane, uniformi, ... In particolare si indica con WGN un rumore bianco Gaussiano.

Note 2

Consideriamo pss a medie nulle, infatti non contiene le caratteristiche spettrali. I dati da studiare segnali nel tempo, se possono rappresentare in frequenza. RAPPRESENTAZIONE SPETTRALE di un pss. Nella tabella il pss → una quantità che nel tempo è la $f(\nu)$.

Sia $y(t)$ un pss. Si definisce densità spettrale di potenza la trasformata di Fourier a tempo discreto delle funz. di covarianza $R_y(\nu)$:

$$\text{DTFT} \quad \boxed{R_y(u) = \sum_{\nu=-\infty}^{+\infty} R_y(\nu) e^{-ju\nu}}$$

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

Note

$R_y(u)$ esiste solo per pos tali che $R_y(\nu) \rightarrow 0$ per $\nu \rightarrow +\infty$. Studieremo così in cui questo vale sempre.

Proprietà di $R_y(u)$:

1) $R_y(u)$ è una funzione reale delle variabili reali u :
 $\text{Im}(R_y(u)) = 0 \quad \forall u \in \mathbb{R}$

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

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

4) $R_y(u)$ è una funzione periodica di periodo 2π :

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

Note

Come conseguenza di 4) si può trarre la funzione sull'arco $[-\pi, \pi]$

Si può ricavare $r(\tau)$ partendo da $R_y(u)$ tramite l'antitrasformato:

$$\boxed{r(\tau) = \frac{1}{2\pi} \int_{-\pi}^{\pi} R_y(u) e^{j u \tau} du}$$

Si mette da:

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

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

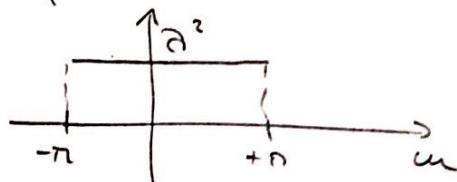
DENSITÀ SPECTRALE DI POTENZA DI UN SUONO PIANO

Sia $c(t) \sim \text{WN}(0, \sigma^2)$. Nel tempo è un segnale impredicibile

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

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

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 impredicibile

RAPPRESENTAZIONE DINAMICA DI UN PSS

Obbiamo rappresentare un pss sia nel tempo con $f(t)$ sia nelle frequenze con $F(\omega)$. Queste rappresentazioni sono però "statiche" \downarrow rappresentano il pss nella sua integrità

Per risolvere il problema della predizione, è necessario avere od una rappresentazione dinamica, che metta 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 $e(t) \sim \text{WN}(0, \sigma^2)$. Abbiamo visto che $e(t)$ ha un spettro costante



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

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



$v(t)$ è STAZIONARIO
perché convoluzioni lineari di

posso

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

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

fattura di
convoluzione

risposta di un sistema
dinamico caso AS STAB
ad un impulso $e(t)$

coefficients risposte all'impulso

è il movimento forzato,
(il movimento libero è o essendo AS STAB)

Il sistema dinamico ha w_i come risposta impulsiva e:

$$W(z) = \sum_{i=0}^{+\infty} w_i z^{-i} \quad \Rightarrow \quad V(t) = W(z) e(t)$$

come funzione di trasferimento

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

$W(z)$ fa funzione dello
spettro di $e(t)$ da e-
costante

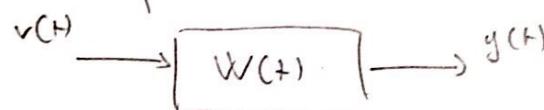
57

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

↓
È pss che si ottengono filtrando un rumore bianco tramite un filtre AS. Stab sono detti processi a spettro rispondente (rispondente fissa).

Teorema

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

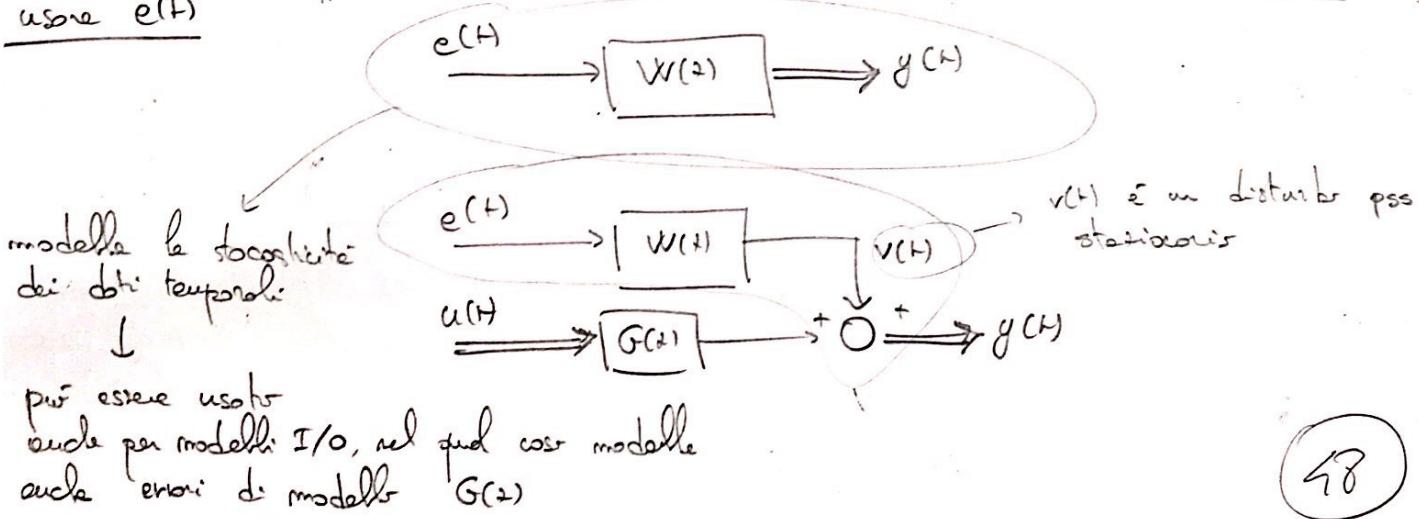


Condizione necessaria e sufficiente offinale: A condizione iniziale, se regime $y(t)$ sia un pss è che:

- 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 autostacione stabile, alimentata da un pss, è un pss

Riprendendo l'approssimazione iniziale di modellazione, abbiamo de la serie uscire $e(t)$



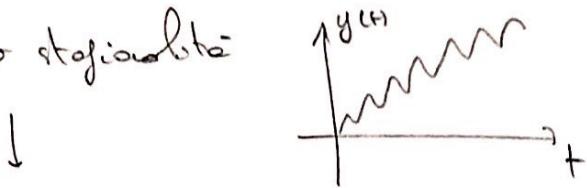
48

Note

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

Note

Nelle modellizzazioni di serie stocastiche $\xrightarrow{\text{eCH}} \boxed{w(t)} \Rightarrow y(t)$ ci potrebbero essere trend o stagionalità



Bisogna quindi prime rimuovere le trend che stagionalità 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 depolitizzazione permette di semplificare il calcolo di $\gamma(\tau)$ nel caso in cui un processo stocastico $v(t)$ 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 le stesse funzionali di convoluzioni (e stesse caratteristiche spettrali)



non si ha da mescolare generalità nello studiare processo e media nulla

FAMIGLIE DI MODELLI A SPETTO 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:

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

$$\Rightarrow \frac{y(t)}{e(t)} = \frac{z^m c_0 + z^{m-1} c_1 + \dots + c_m}{z^m}$$

m poli in $z=0$

$$\xrightarrow{e(t)} \boxed{C(z)} \rightarrow y(t)$$

I processi MA sono sempre stazionari

Calcolo dei parametri caratteristici

• Valore atteso

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

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

(50)

- $y(n)$, supponiamo $E[y^{(L)}] = 0$ per deplorabilità

$$\begin{aligned}
 & -\gamma_e(0) \cdot E \left[(y(t) - m(t))^2 \right] = E \left[y(u^2) \right] = E \left[(c_0 e^{ct} + c_1 e^{c(t-1)} + \dots + c_m e^{c(t-m)})^2 \right] = \\
 & = E \left[\underbrace{c_0^2 e^{ct^2} + c_1^2 e^{c(t-1)^2} + \dots + c_m^2 e^{c(t-m)^2}}_{\text{quadrati}} + \underbrace{2c_0c_1 e^{ct}e^{c(t-1)} + \dots + 2c_{m-1}c_m e^{c(t-m-1)}e^{c(t-m)}}_{\text{dappi prodotti}} \right] \\
 & = c_0^2 E[e^{cu^2}] + c_1^2 E[e^{c(t-1)^2}] + \dots + c_m^2 E[e^{c(t-m)^2}] \\
 & = c_0^2 \gamma_e(0) + c_1^2 \gamma_e(0) + \dots + c_m^2 \gamma_e(0) = \boxed{\gamma_e^2 \cdot \sum_{i=0}^m c_i^2}
 \end{aligned}$$

$$\begin{aligned}
 - \sigma^2 &= E \left[(y(t) - m(t)) (y(t-1) - m(t)) \right] = E \left[y(t) y(t-1) \right] = \\
 &= E \left[\left(c_0 e(t) + c_1 e(t-1) + \dots + c_m e(t-m) \right) \left(c_0 e(t-1) + c_1 e(t-2) + \dots + c_{m-1} e(t-m+1) \right) \right] \\
 &= \overline{c_0^2} E \left[e(t-1)^2 \right] + c_0 c_1 E \left[e(t-1) e(t-2) \right] + \dots + c_{m-1} c_m E \left[e(t-m)^2 \right] \\
 &= \overline{c^2} \cdot (c_0 + c_1 + \dots + c_{m-1}) \cdot \overline{e^2}
 \end{aligned}$$

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

$$-\gamma(m) = \partial^2 (\cos m)$$

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

Um processo MA(m) depende só de m e é direta imprevisível (ou seja, não pode ser previsível).

Un modo per sapere se una serie temporale è MA è quello di guardare se le sue $y(t)$ ve a zero dopo un certo t_0

Note

$$\text{If process } \tilde{\eta}(t) = c_0 \eta(t) + c_1 \eta(t-1) + \dots + c_m \eta(t-m) \quad \text{com} \quad \tilde{c}_i = \alpha \cdot c_i$$

$$\eta(t) \sim \text{wn}(0, \tilde{\sigma}^2) \quad \tilde{\sigma}^2 = \frac{\lambda^2}{n^2}$$

le stesse cose other e furse l'origine del processore

$y(t) = c_0 e^{(t)} + c_1 e^{(t-1)} + \dots + c_m e^{(t-m)}$ $\sim N(0, \sigma^2)$. Per entro queste ^{sarà} ipotesi, si prenderà $c_0 = 1$

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

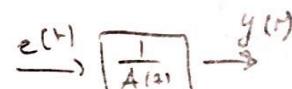
L'uscita di un AR(m) è la combinazione lineare degli ultimi m "vecchi" valori del processo stesso; più l'ingresso $e(t)$ all'istante

Forma operatoriale:

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

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

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



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

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

- m ZERI NELL'ORIGINE

- m POLI: quindi non è sempre stabile

Il sistema è stabile se tutti i poli sono in modulus < 1

← bisogna fare attenzione alle stime sui parametri

Calculari 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)

- $y(t) \rightarrow$ calcolo per processo AR(1) perché è complesso

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

$$\Rightarrow y(t)[1 - \alpha z^{-1}] = e(t) \Rightarrow \underbrace{y(t) = \frac{1}{1 - \alpha z^{-1}} e(t)}_{A(z)}$$

$y(t)$ 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 $|z| < 1$

Supponiamo $A(z)$ AS. STAB e che il processo $y(t)$ sia determinato (medie nulle)

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

$\Rightarrow y(t-1)$ è inserviziato con $e(t)$
perché dipende da $e(t-1)$ e da
 $y(t-2)$ → per ricorsione anche
 $y(t-2)$ è $\perp \!\!\! \perp e(t)$ 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:

$$\begin{cases} \gamma(0) = \frac{\alpha^2}{1-\alpha^2} \\ \gamma(r) = \alpha \cdot \gamma(r-1) \quad r > 0 \end{cases}$$

EQUAZIONI DI YULE-WALKER
PER UN AR(1)
esistono anche per un AR(m)

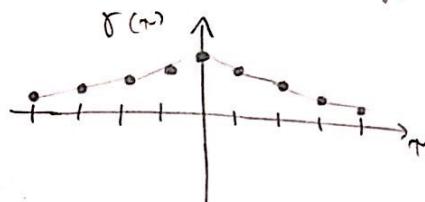
Osservazione

Dato che $|\alpha| < 1$ (obbligo supposto stazionario per poter calcolare $\gamma(r)$) si ha:

$$|\gamma(r+1)| < |\gamma(r)|$$

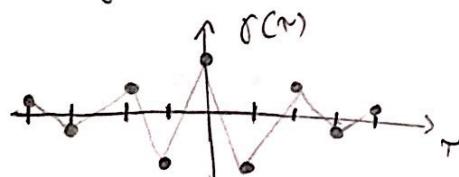
e dato che $\alpha \neq \pm 1$, $\gamma(0)$ esiste finito. Inoltre si ha che:

- Il processo $y(t) = \alpha y(t-1) + e(t)$ con $0 < \alpha < 1$ ha $\gamma(r) > 0 \forall r$, e sono finite le realizzazioni che con raggiungono mai lo 0.



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

- Il processo $y^{(u)} = \alpha y(t-1) + e(t)$ con $-1 < \alpha < 0$ ha le realizzazioni $\gamma(r)$ che cambiano segno od ogni r e decrescono in valore assoluto



Le realizzazioni del pss $y(t)$ ormai quindi (in media) cambiano di segno regolarmente, creando segnali con dei comportamenti in alte frequenze

Apprendimento

Osserviamo visto che, per un modello $MA(m)$, $\gamma(r) = 0$ per $r > m$. Per gli $AR(m)$ possono ottenere un comportamento simile sotto le FUNZIONI DI AUTOCORRELAZIONE PARZIALE (PACF) $\gamma_{par}(r)$, 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 è un $MA(m)$ plotando $\gamma(r)$
- 2) " " " " $AR(m)$ " $\gamma_{par}(r)$
- 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 $e(t) \sim N(0, \sigma^2)$, è detto di tipo $ARMA(m, n)$ se:

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

- m : L'ORDINE DEL MODELLO AR
- $\alpha_1, \alpha_2, \dots, \alpha_m$: COEFFICIENTI DEL MODELLO AR
- n : L'ORDINE DEL PROCESSO MA
- c_1, c_2, \dots, c_n : COEFFICIENTI DEL PROCESSO MA

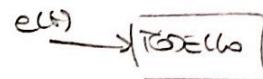
Notiamo che $ARMA(0, m) = MA(m)$ e che $ARMA(m, 0) = AR(m)$. Possono in forma generale:

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

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

MODELLO PER SISTEMI INPUT / OUTPUT

ARMAX (Autoregressive Trailing Average eXogenous)



Un processo $y(t)$, percorre:

attraverso di un ritardo bianco $e(t) \sim WN(\mu, \sigma^2)$ e da un imposto esterno $u(t)$ (independente dal processo), è detto ARMAX $(m, n, k+p)$ se:

$$y(t) = a_1 y(t-1) + a_2 y(t-2) + \dots + a_m y(t-m) + \text{PARTE AR}(m)$$

$$+ e(t) + c_1 e(t-1) + \dots + c_m e(t-m) + \text{PARTE MA}(m)$$

$$+ b_0 u(t-k) + b_1 u(t-k-1) + \dots + b_p u(t-k-p) \quad \text{PARTE X}(k+p)$$

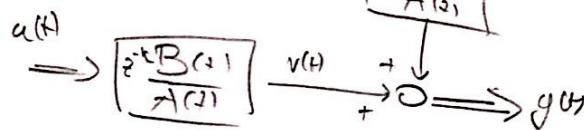
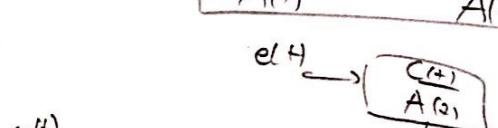
- m : ORDINE PROCESSO AR
- a_1, a_2, \dots, a_m : COEFFICIENTI PROCESSO AR
- m : ORDINE PROCESSO MA
- c_1, c_2, \dots, c_m : COEFFICIENTI PROCESSO MA
- p : ORDINE VARIABILE ESOGENA
- $p+1$: COEFFICIENTI VARIABILE ESOGENA
- k : RITARDO PURO TRA INGRESSO ED USCITA

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

$$+ u(t-k) \left[b_0 + b_1 z^{-1} + b_2 z^{-2} + \dots + b_p z^{-p} \right]$$

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

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



(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 c'è un processo stocastico)

In genere il processo ARTAX non è stazionario, ma le sue componenti stocastiche (ARTA) danno esecuzione. La fonte di non stazionarietà è una componente deterministica (nota), quindi le possiamo modellare

In genere le componenti di un'oscillazione e le tendenze (es trend, stagionalità) o le modelli (nel caso di ingresso esogeno noto "si modella da sé") in quanto

Esempio

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

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

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

Poli di $A(z)$

$$\frac{1 - \frac{1}{3} z^{-1}}{z} = 0 \Rightarrow z^{-1} = 3 \Rightarrow z = \frac{1}{3} < 1 \Rightarrow \begin{array}{l} - A(z) \text{ AS. STAB} \\ - u(t) \text{ STAZIONARIO} \end{array} \Rightarrow y(t) \text{ è PSS.}$$

Quale è il contributo di $u(t)$ sull'uscita $y(t)$? Teorema delle risposte impulsive: dato un ingresso sinusoidale $u(t) = Q \cdot \cos(\omega t + \phi)$ la uscita:

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

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

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

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

(57)

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

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

- $\tilde{y}(t) \Rightarrow$ Depolarizing $y(t)$ ed $e(t)$. $\left\{ \begin{array}{l} \tilde{y}(t) = y(t) - \frac{3}{2} \\ \tilde{e}(t) = e(t) - 1 \end{array} \right.$
 Scriv il processo come:

$$g(t) = \frac{1}{3} g(t_{-}) + c(t) + z$$

$$\tilde{y}^{(t)} + \frac{\alpha}{2} = \frac{1}{3} \left[\hat{y}^{(t-1)} + \frac{\alpha}{2} \right] + \tilde{e}^{(t)} + \epsilon_1 + \epsilon_2$$

$$\tilde{y}(t) = \frac{1}{\alpha} \tilde{y}(t-1) + \frac{\alpha^3 - 1}{\alpha^2} z_1 + z_2 + \tilde{e}(t)$$

$$\tilde{y}^{(t)} = \frac{1}{3} \tilde{y}^{(t-1)} + \tilde{\epsilon}^{(t)}$$

AR(1) = media mala

$$\frac{3 - 8 + 6}{2} = 0$$

$$\tilde{e}(H \sim \text{WN}(0, 1))$$

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

YULE WALKER AR(C)

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

Teorema

Dati un processo stocastico de seconda ARMA(m, n), esso può essere scritto come
 $Y_t = \phi_0 + \sum_{j=1}^m \phi_j Y_{t-j} + \sum_{j=1}^n \theta_j \epsilon_{t-j}$

ES ARCI)

$$y(t) = \omega y(t-1) + e(t) \quad e(t) \sim \text{WN}(\mu, \sigma^2)$$

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

$$= \sum_{k=0}^{+\infty} (az^{-1})^k \cdot e(t) = \sum_{k=0}^{+\infty} a^k \cdot e(t-k)$$

CALCOLO DELLO SPETTORE DI UN PSS A PARTIRE DAL SUO MODELLO

Note: usiamo i termini DENSITÀ SPECTRALE DI POTENZA e SPETTO in modo equivalente.

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

$$y(t) = F(z) v(t) \quad v(t) \xrightarrow{F(z)} y(t)$$

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

$$\boxed{\Gamma_y(u) = |F(e^{ju})|^2 \cdot \Gamma_v(u)}$$

- $\Gamma_y(u)$: spettro dell'uscita
- $|F(e^{ju})|^2$: modulo al quadrato della risposta in frequenza del filtro
- $\Gamma_v(u)$: spettro dell'ingresso

Se l'ingresso è un white noise $e(t) \sim \text{wn}(0, \gamma^2)$: $\boxed{\Gamma_y(u) = |F(e^{ju})|^2 \cdot \gamma^2}$

Es

Consideriamo un processo $T(A(z))$, calcolare $\Gamma_y(u)$:

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

1) Usando la definizione

$$\begin{aligned} \gamma^2(\text{GC}_1) &= \Gamma(u) = \sum_{\tau=-\infty}^{+\infty} \Gamma(\tau) e^{-ju\tau} \quad \Rightarrow \quad \gamma^2 \sum_{i=0}^m c_i^2 \\ &= \gamma(-1) e^{-ju(-1)} + \gamma(0) e^{-ju0} + \gamma(1) e^{-ju(1)} \\ &= 1^2 (1+c) e^{ju} + (1^2 + c^2) \cdot 1 + c e^{-ju} = c [e^{ju} + e^{-ju}] + c^2 + 1 \\ &= \boxed{2c \cdot \cos(u) + c^2 + 1} \end{aligned}$$

2) Usando il teorema

$$y(t) = (1+ce^{-\tau}) e(t) = C(z) e(t)$$

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

$$\begin{aligned} \Gamma_y &= |C(e^{ju})|^2 \cdot \Gamma_e(u) = |1+ce^{-ju}|^2 \cdot 1 = (1+ce^{-ju})(1+ce^{ju}) \\ &= 1 + c^2 (e^{ju} \cdot e^{-ju}) + c(e^{ju} + e^{-ju}) = \boxed{1 + c^2 + 2c \cdot \cos(u)} \end{aligned}$$

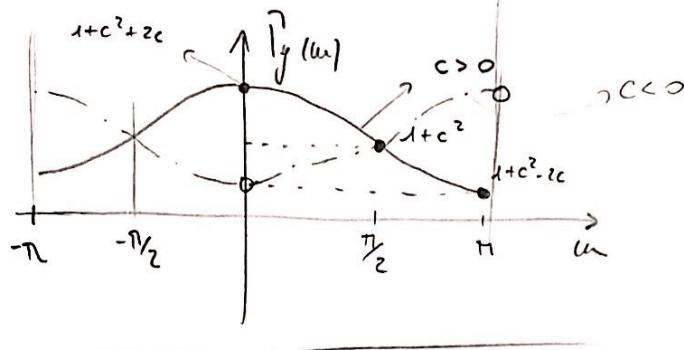
(55)

Tracciare le spalle per punti

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

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

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

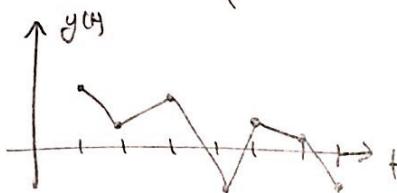


& PREDIZIONE &

Problema: dato una sequenza di dati (ad esempio serie temporale)
 $\{y^{(1)}, y^{(2)}, \dots, y^{(N)}\}$

Vogliamo IDENTIFICARE un modello ARMA

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



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

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

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

Approccio predittivo: un modello è buono se è capace di predire in
 processo

FILTO PASSA-TUTTO

È un filtro di ordine 2 con le seguenti forme:

$$T(z) = \frac{1}{2} \cdot \left(\frac{z+\alpha}{z+\frac{1}{\alpha}} \right) \quad \alpha \neq 0, \alpha \in \mathbb{R}$$

Il segnale è opposto al polo

Ricordando il calcolo dello spettro delle fdt, si ha che:

$$\Gamma_y(\omega) = |T(e^{j\omega})|^2 \cdot \Gamma_e(\omega)$$

$$\begin{aligned} -|T(e^{j\omega})|^2 &= \left(\frac{1}{2} \cdot \frac{e^{j\omega} + \alpha}{e^{j\omega} + \frac{1}{\alpha}} \right) \left(\frac{1}{2} \cdot \frac{e^{-j\omega} + \alpha}{e^{-j\omega} + \frac{1}{\alpha}} \right) \\ &= \frac{1}{2^2} \cdot \frac{(e^{j\omega} + \alpha)(e^{-j\omega} + \alpha)}{\left(e^{j\omega} + \frac{1}{\alpha} \right) \left(e^{-j\omega} + \frac{1}{\alpha} \right)} = \frac{1}{2^2} \cdot \frac{1 + \alpha^2 + 2\alpha(e^{j\omega} + e^{-j\omega})}{1 + \frac{1}{\alpha^2} + \frac{1}{\alpha}(e^{j\omega} + e^{-j\omega})} \\ &= \frac{1}{2^2} \cdot \frac{1 + \alpha^2 + 2\alpha \cos \omega}{\cancel{\alpha^2 + 1 + 2\alpha \cos \omega}} = \boxed{1 \pm} \end{aligned}$$

Quindi:

$$\Gamma_y(\omega) = |T(e^{j\omega})|^2 \cdot \Gamma_e(\omega) = \Gamma_e(\omega)$$

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

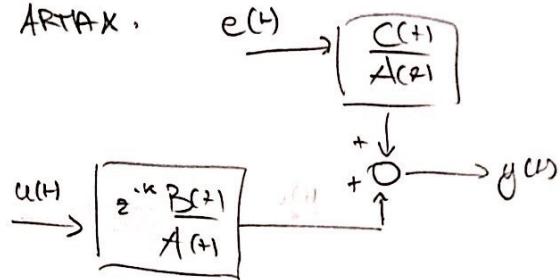


Il segnale in ingresso e quello in uscita del filtro passatutto sono EQUIVALENTE

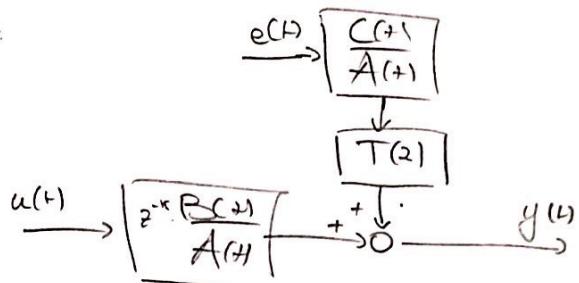
Note

Il filtro passatutto non modifica il modulo ma introduce una distorsione di fase, ritorcendo il segnale in ingresso. Ingresso e uscita NON SONO IDENTICI, ma sono SPEETRALMENTE EQUIVALENTI

Consideriamo un ARMAX.



questo risulterà equivalente a:



Osservazione

Il filtro passatutto è un "oggetto matematico". Non possiamo inserire tra $u(t)$ e $y(t)$, oltrimenti comprenderebbe la relazione infase/uscita, che è data da un "oggetto fisico", reale.

FORMA CANONICA

Con l'introduzione del filtro passatutto $T(z)$, otteniamo visto che il processo $y(t) = \frac{C(z+1)}{A(z+1)} e(t)$ ed il processo $y(t) = \frac{C(z+1)}{A(z+1)} T(z) e(t)$ sono EQUIVALENTI dal punto di vista spettrale

esistono altre rappresentazioni equivalenti?

Consideriamo questi 5 processi ARMA:

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

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

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

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

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

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

Calcoliamo gli spettri dei processi:

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

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

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

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

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



Tutti e 5 i processi sono spettri equivalenti. Un processo ARMA avrà infinite rappresentazioni equivalenti. Le cause di unicità sono:

- rettangoli puri (processo 2)
- fattori moltiplicativi che si conciliano (processo 3)
- coefficienti moltiplicativi su funzioni di trasferimento e rumore si compensano (processo 4)
- più/zeri "reciprocii" (processo 5)

TEOREMA DELLA FATTORIZZAZIONE SPECTRALE

Dato un processo a spettro risonante, 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) $C(z)$ ed $A(z)$ hanno relazioni interne di condiz. unitarie (poli e zeri $|z| < 1$)

Es

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

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

IL PROBLEMA DELLA PREDICTION

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

Informazioni disponibili

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

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

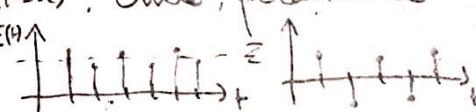
Es 2

$$\hat{y}(t+1|t) = \frac{2y(t) + \frac{1}{2}y(t-1) + \frac{1}{2}y(t-2)}{3} \quad \begin{array}{l} \text{diamo più peso a valori} \\ \text{più recenti} \end{array}$$

Potrà fare meglio? Una delle proprietà del miglior è che il predittore sia corretto, ovvero $E[\epsilon(t)] = E[y(t) - \hat{y}(t|t-1)] = 0$

• PREDITTORE OTTIMO •

Un predittore è ottimo se:

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

2) $\text{Var}[\epsilon(t)]$ MINIMA

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

- $\epsilon(t)$ parte impredicabile al tempo $t-k$
- $\hat{y}(t|t-k)$ parte del processo che è predicibile al tempo $t-k$

• PREDITTORE AD UN PASSO DI PROCESSI MA •

Supponiamo un processo MA(m) in forma canonica:

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

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

(65)

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 un fin d' tempo $t-1$ - E carattere: $E[\epsilon(t)] = 0$

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

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

- Ma è possibile trovare un predittore con $V[\epsilon(t)]$ minore, infatti non poss'essere vero che $V[\epsilon(t)]$

↓

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

Questo, tuttavia, è un predittore ottimo "del rumore". Vogliamo un predittore ottimo "di dati"

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

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

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

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

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

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

Osservazione

Il predittore è ricorsivo, bisogna dire quante volte $\hat{y}(1|0)$. Di solito si usa la media dei campioni (se la media del processo è nota, si usa quella)

Se il predittore è AS-STAB., l'effetto dell'inizializzazione scompare, / 65
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 predittore. Non basta per dire che è ottimo.

PREDITTORE K PASSI MA(m)

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

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

\Downarrow

$$\underbrace{E(t)}_{\perp} \quad \underbrace{\hat{y}(t|t-k)}$$

Il predittore è ottimo in quanto le condizioni necessarie sono soddisfatte ed $E(t)$ è a varianza minima

$$\text{Var}[E(t)] = \text{Var}[y(t)]$$

Osservazione

$$\begin{aligned} E_1(t) &= y(t) - \hat{y}(t|t-1) \Rightarrow \text{Var}[E_1(t)] = \text{Var}[\epsilon(t)] = \sigma^2 \\ E_2(t) &= y(t) - \hat{y}(t|t-2) \Rightarrow \text{Var}[E_2(t)] = \text{Var}[\epsilon(t) + c_1 \epsilon(t-1)] = (1 + c_1^2) \sigma^2 > \sigma^2 \\ &\vdots \\ E_{m+1}(t) &= y(t) - \hat{y}(t|t-m) \Rightarrow \text{Var}[E_{m+1}(t)] = \text{Var}[y(t) + c_1 \epsilon(t-1) + \dots + c_m \epsilon(t-m)] = \text{Var}[y(t)] \end{aligned}$$

La varianza di $E(t)$ aumenta con l'orizzonte di predizione, fino a diventare uguale alla varianza del processo. Il predittore $\hat{y}(t|t-m)$ sarà il predittore buono, ovvero la media del processo $\hat{y}(t|t-m) = E[y(t)] = 0 \Rightarrow$ un predittore 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 causale

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

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

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

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

Ci ricordiamo che possiamo esprimere $A(z)A(\bar{z})$ come $\tilde{R}(z)$

\downarrow
Non è però fattibile, perché la parte predittiva contienebbe ∞ 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. Ottieniamo 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 le passi di lunga divisione (possare essere ∞), ottieniamo le informazioni per una previsione a k passi

Es:

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

Lunga divisione di $k=2$ passi:

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

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

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

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

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

Quindi il predittore del rumore è:

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

L'errore di predizione è:

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

Osservazioni

- $\hat{y}(t|t-k)$ dipende dal rimanente flusso di tempo $t-k$
- $E[\hat{y}(t|t-k) \cdot e(t)] = 0$, infatti $E[\hat{y}(t|t-k) \cdot e(t)] = E\left[\left(\frac{\tilde{R}(t)}{A(t)} e(t-k)\right)(E(t) e(t))\right] = 0$
- Non è possibile trarre un predittore con le due popolazioni precedenti, con $V_{\text{er}}[e(t)]$ inferiore

↓

$$\hat{y}(t|t-k) = \frac{\tilde{R}(t)}{A(t)} e(t-k) \text{ è } \underline{\text{ottimo}}$$

Predditore dei dati

$$y(t) = \frac{C(t)}{A(t)} e(t) \Rightarrow e(t) = \frac{A(t)}{C(t)} y(t)$$

$$\hat{y}(t|t-k) = \frac{\tilde{R}(t)}{A(t)} e(t-k) = \frac{\tilde{R}(t) e^{-k}}{A(t)} e(t) = \frac{\tilde{R}(t) e^{-k}}{A(t)} \frac{A(t)}{C(t)} y(t) = \boxed{\frac{\tilde{R}(t)}{C(t)} y(t-k)}$$

Osservazione

$$E(t) = y(t) - \hat{y}(t|t-k) = E(t) e(t) \quad \text{P' erme di predizione è un processo MA}(k-1)$$

CASO k=1

$$\begin{array}{c} C(t) \\ -A(t) \\ \hline 1 \\ \hline C(t) - A(t) \\ R(t) \end{array}, \quad E(t)$$

$$E(t) = \pm \Rightarrow E(t) = E(t) e(t) = e(t)$$

$$R(t) = C(t) - A(t)$$

$$\boxed{\hat{y}(t|t-1) = \frac{C(t) - A(t)}{C(t)} y(t)}$$

PREDIZIONE
DATI
ATTUALI

$$\boxed{\hat{y}(t|t-1) = \frac{C(t) - A(t)}{A(t)} e(t)}$$

PREDIZIONE
DATI
PREVISTI

$$\boxed{E(t) = y(t) - \hat{y}(t|t-1) = e(t)}$$

Osservazione

Gli stimatori trovati sono corretti in quanto $E[E(t)] = 0$

QUALITÀ DEL PREDITTORE

Possedere valutare le qualità di un predittore mettendo a confronto le variazioni dell'errore di predizione con le variazioni delle predizioni bivariate.

$$\text{Error Signal Ratio} \leftarrow ESR = \frac{\text{Var} [y(t) - \hat{y}(t|t-\kappa)]}{\text{Var} [y(t)]} = \frac{\text{Var} [\epsilon_t(t)]}{\text{Var} [y(t)]}$$

vogliamo de
swa più
piccolo possibile

PREDITTORE OTTIMO DI UN PROCESSO ARMAX

Sia $y(t)$ un processo ARMAX(m, m, p), $\frac{C(+)}{A(+)}$ è in forma canonica

$$y(t) = \frac{B(+)}{A(+)} u(t-\kappa) + \frac{C(+)}{A(+)} e(t) \quad e(t) \sim \text{wn} (\mathbb{E}, \sigma^2)$$

$\frac{B(+)}{A(+)}$ è il modello del sistema, non si può mettere in forma canonica.

Si vuol fare una predizione a ke passi, immettere l'impresso nell'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

$$\hat{y}(t|t-\kappa) = \frac{B(+)}{A(+)} u(t-\kappa) + \frac{\tilde{R}(+)}{A(+)} e(t-\kappa)$$

$$E(t) = E(+ e(t))$$

Osservazioni

- $\hat{y}(t|t-\kappa)$ dipende dal $u(t)$ e dall'impresso fin' al tempo $t-\kappa$
- $E[\epsilon(t)] = 0$ CERTO
- $E[\hat{y}(t|t-\kappa) \cdot E(t)] = 0$ (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'impresso $u(t-\kappa)$ è completamente noto e non introduce incertezza

Preditore dei dati

$$y(t) = \frac{B(+) u(t-\kappa)}{A(\kappa)} + \frac{C(+)}{A(\kappa)} e(t) \Rightarrow e(t) = \frac{A(\kappa)}{C(+)} y(t) - \frac{B(+) u(t-\kappa)}{C(+)}$$

$$\hat{y}(t|\kappa) = \frac{B(+) u(t-\kappa)}{A(\kappa)} + \frac{R(+)}{A(\kappa)} e(t)$$

Faccendo i passaggi si ottiene:

$$\left| \begin{array}{l} \hat{y}(t|\kappa) = \frac{\tilde{R}(+)}{C(+)} y(t-\kappa) + \frac{B(+) E(+)}{C(+)} u(t-\kappa) \end{array} \right|$$

CASO $\kappa=2$

$$\begin{aligned} E(+) &= 1 \\ R(+) &= C(+) - A(+) \end{aligned}$$

$$\Rightarrow \left| \begin{array}{l} \hat{y}(t|\kappa=2) = \frac{C(+) - A(+)}{C(+)} y(t) + \frac{B(+)}{C(+)} u(t-1) \\ E(t) = E(2) e(t) = e(t) \end{array} \right|$$

Osservazione

Il predittore ARMAX ha varianza di $E(t)$ data solo dalla parte stocastica. Si può calcolare la bari del predittore 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 la varianza dell'errore di predizione

Il ritardo puro è $\kappa=2$, quindi le sussi calcolare un predittore a 2 passi

- Il processo è in forma causale? \Rightarrow Solo la parte stocastica può essere modificata!



Per il predittore, parte espressa e parte nascosta danno ovvero l' stesso denominatore

$$y(4) = \frac{(2+6z^{-1})(1+\frac{1}{2}z^{-1})}{1+\frac{1}{2}z^{-1}} u(t-2) + \frac{1}{1+\frac{1}{2}z^{-1}} e(t) \quad \begin{array}{l} \text{B}(2) \\ \text{C}(2) \\ \text{e}(t) \sim \text{wn}\left(0, \frac{\sigma^2}{8}\right) \end{array}$$

$$A(\omega) = -1 + \frac{1}{2} \omega^2$$

$$B_2(2) = -2 \left(1 + 3z^{-1} \right) \left(1 + \frac{1}{2} z^{-1} \right)$$

Il deunistro come è stato fatto dopo la coronazione di $\frac{C+1}{A+1}$

- Preditore è $k=2$ possi \Rightarrow 2 possi di lunga divisione fra $\frac{C+1}{A+1}$

$$\begin{array}{c}
 \text{Diagram showing partial fraction decomposition:} \\
 \frac{-1 - \frac{1}{2}z^{-1}}{-\frac{1}{2}z^{-1}} = \frac{1 + \frac{1}{2}z^{-1}}{1 - \frac{1}{2}z^{-1}} + \frac{\frac{1}{2}z^{-1} + \frac{1}{2}z^{-2}}{1 - \frac{1}{2}z^{-1}}
 \end{array}$$

$$\hat{y}(t|t-\kappa) = \frac{\tilde{R}(+)}{CC^2} y(t-\kappa) + \frac{P(+)}{CC+1} E(+) \cdot u(t-\kappa)$$

$$\hat{y}(t|t-2) = \frac{\frac{1}{z}}{1} y(t-2) + \frac{2(1+3z^{-1})(1+\frac{1}{2}z^{-1})\left(1-\frac{1}{2}z^{-1}\right)}{1} u(t-2)$$

$$= \frac{1}{5} g(t-2) + (2+6t^2) \left(1 - \frac{1}{5} t^{-2}\right) u(t-2)$$

$$= \frac{1}{5} y(t-2) + 2u(t-2) + 6u(t-3) - \frac{1}{2} u(t-4) - \frac{3}{2} u(t-5) \quad \text{"prime" di } t-2$$

$$\text{Var}[\epsilon(t)] = E[\epsilon(t)^2] = E[(E(t)\epsilon(t))^2] = E\left[\left(\left(1 - \frac{1}{2}\gamma^{-1}\right)e(t)\right)^2\right] = E\left[\left(e(t) - \frac{1}{2}\gamma e(t-\gamma)\right)^2\right]$$

$$= \left(1 + \frac{1}{4}\right)\text{Var}[e(t)] + \frac{\gamma}{4} \cdot \frac{4}{3} = \boxed{\frac{5}{3}}$$

Poiché se ne per un AR(1) confrontare $\text{Var}[\epsilon(t)]$ con $\text{Var}[v(t)]$, perché $v(t)$ potrebbe essere un processo stocastico?

Si confronta con $\text{Var}\left[\frac{C(t)}{A(t)} e(t)\right]$, ovvero con la parte stocastica del processo

$$\bullet \text{Var}\left[\frac{1}{1 + \frac{1}{2}\gamma^{-1}} e(t)\right] = \text{Var}\left[-\frac{1}{2}v(t-1) + e(t)\right] = E\left[\left(-\frac{1}{2}v(t-1) + e(t)\right)^2\right] = +\frac{1}{2}\text{Var}[v(t)] + \text{Var}[e(t)]$$

$$\begin{aligned} v(t) &= -\frac{1}{2}v(t-1) + e(t) \\ \Rightarrow \text{Var}[v(t)] &= \frac{1}{4}\text{Var}[v(t)] + \text{Var}[e(t)] \Rightarrow \frac{3}{4}\text{Var}[v(t)] = \frac{5}{3} \\ &\Rightarrow \text{Var}[v(t)] = \frac{16}{27} \end{aligned}$$

$$\text{ESR} = \frac{\text{Var}[e(t)]}{\text{Var}[v(t)]} = \frac{\frac{5}{3}}{\frac{16}{27}} = \frac{5 \cdot 3}{16} = 0,9375$$

IDENTIFICAZIONE

Le analisi sistematiche si basano sulle conoscenze dei propositi del modello dinamico.



La disciplina dell'identificazione consiste nello stimare il sistema ignoto.

Dipendendo da un set di dati spaziali-temporali input-output $u(t)$ 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 da dei parametri $\theta \in \mathbb{R}^d$, da stimare. Tipi di modelli:

- Tempo discreto / continuo

- Sistema lineare / non lineare

⇒ ARMAX sono le categorie di modelli più complete da utilizzare

- Tempo invariante / variante

- Diminuzionali / statici

b) Scelta degli ordini del modello

③ SCELTA DELLA CIFRA DI MERITO

È una funzione $J_N(\theta) : \mathbb{R}^d \rightarrow \mathbb{R}^+$ che definisce le qualità della stima di $\theta \in \mathbb{R}^d$.

L'obiettivo da usare è l'obiettivo PNM (Prediction Error Minimization), mimimizza la varianza dell'errore di predizione ad un passo.

$$J(\theta) = E \left[(y(t) - \hat{y}(t-1, \theta))^2 \right]$$

23

Dato che le soli 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 ALTRA DI KERNO

Vi sono diversi casi:

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 ha 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 T osservazioni 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: stimare dai dati le seguenti quantità teoriche

$$m = E[y(t)] \rightarrow \hat{m}_N(y(1), y(2), \dots, y(N))$$

$$\sigma(m) = E[(y(t)-m)(y(t+r)-m)] \rightarrow \hat{\sigma}_N(r)(y(1), y(2), \dots, y(N))$$

Per stimare 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

$$\begin{aligned}\hat{m}_N &= \frac{1}{N} \sum_{t=1}^N y(t, s) \Rightarrow E_s[\hat{m}_N] = \frac{1}{N} E_s\left[\sum_{t=1}^N y(t, s)\right] = \frac{1}{N} \sum_{t=1}^N E_s[y(t, s)] \\ &= \frac{1}{N} \sum_{t=1}^N m(t) = \frac{1}{N} \cdot N \cdot m = \boxed{m} \quad \text{corretto}\end{aligned}$$

Consistenza

Generalità dei teoremi:

Teorema

\hat{m}_N è consistente sse $\gamma(\tau) \rightarrow 0$ per $|\tau| \rightarrow +\infty$

Teorema

Dato un processo ARMA, si ha che $\gamma(\tau) \rightarrow 0$ per $|\tau| \rightarrow +\infty$

FUNZIONE DI COVARIANZA CAMPIONARIA

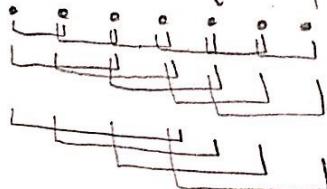
Supponiamo un processso stocastico nullo. Ricordando che $\gamma(\tau) = E[y(t) \cdot y(t+\tau)]$, un possibile stimatore può essere:

$$\hat{\gamma}_N(\tau) = \frac{1}{N-|\tau|} \sum_{t=1}^{N-|\tau|} y(t) \cdot y(t+|\tau|)$$

$$|\tau| \leq N-1$$

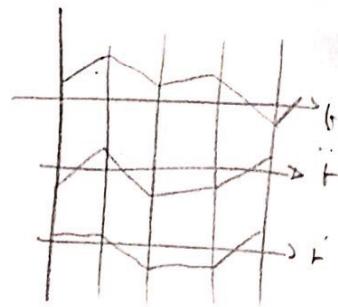
Osservazioni

- Più τ è grande, meno campioni possiamo usare. Quindi $\hat{\gamma}_N(\tau)$ è ormai se $N \gg |\tau|$.



Correttezza

$$\begin{aligned}E_s[\hat{\gamma}_N(\tau)] &= E_s\left[\frac{1}{N-|\tau|} \sum_{t=1}^{N-|\tau|} y(t) \cdot y(t+|\tau|)\right] = \frac{1}{N-|\tau|} \sum_{t=1}^{N-|\tau|} E_s[y(t) \cdot y(t+|\tau|)] \\ &= \frac{1}{N-|\tau|} \cdot (N-|\tau|) \cdot \gamma(\tau) = \boxed{\gamma(\tau)} \quad \text{CORRETTO!}\end{aligned}$$



INTRODUZIONE

Un primo modo per identificare un processo è studiare le caratteristiche fondamentali.



STIMA DI MEDIA,
COVARIANZA, SPECTRO

(75)-A

Consistenza

Teorema

$\hat{g}_N(\gamma)$ è consistente se $f(\gamma) \rightarrow 0$ per $|\gamma| \rightarrow +\infty$

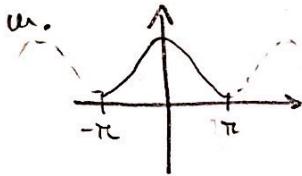
VARIANTE NON CORRETTA

$$\hat{g}'_N(\gamma) = \frac{1}{N} \cdot \sum_{t=1}^{N-1|\gamma|} y(t) \cdot y(t+1|\gamma|) \quad \text{non è corretto, ma è asintoticamente corretto} \quad N \rightarrow +\infty$$

Questo risultato è comunque utilizzabile perché anche nella variante corretta si suppone che $N \gg M$

DENSITÀ SPECTRALE CAMPIONARIA

Si tratta di un passo medio nullo. Sappiamo che $\tilde{f}(u) = \sum_{\gamma=-\infty}^{\gamma=\infty} g(\gamma) e^{-j\gamma u}$ (è DTFT - Discrete Time Fourier Transform delle $y(n)$). $\tilde{f}(u)$ è una funzione continua di u .



Osservazione

La pulsazione $u=\pi$ corrisponde alle frequenze di Nyquist del sistema \Rightarrow Es se $f_s = 20 \text{ Hz}$, $\pi = \frac{f_s}{2} = 50 \text{ Hz}$

Una possibile stima potrebbe essere $\hat{f}_N(u) = \sum_{\gamma=-\infty}^{+\infty} \hat{g}_N(\gamma) e^{-j\gamma u}$. Ricordando però che:

$$\hat{g}_N(\gamma) = \frac{1}{N-1|\gamma|} \cdot \sum_{t=1}^{N-1|\gamma|} y(t) y(t+1|\gamma|) \quad |\gamma| \leq N-1$$

notiamo che non possemo vere valori di $\hat{g}_N(\gamma)$ per $|\gamma| \geq N$. Quindi osserviamo:

$$\left. \begin{aligned} \hat{f}_N(u) &= \sum_{\gamma=-N+1}^{N-1} \hat{g}_N(\gamma) \cdot e^{-j\gamma u} \\ &\quad \end{aligned} \right\}$$

Osservazione

$\hat{f}(u)$ contiene due osservazioni:

- 1) Usiamo $\hat{g}_N(\gamma)$ al posto di $g(\gamma)$
- 2) Osserviamo truncata le sommatorie a $\pm (N-1)$

Osservazione

$\hat{f}(u)$ è continua. Nelle pratiche, abbiamo discretizzato l'intervallo $[0, \pi]$

Corretto

$$E_s[\hat{\Gamma}_N(u)] = E_s\left[\sum_{\tau=-N+1}^{N-1} \hat{f}_N(\tau) e^{-j\omega\tau}\right] \cdot \underbrace{\sum_{\tau=-N+1}^{N-1} E_s[\hat{f}_N(\tau)] e^{-j\omega\tau}}_{\delta(\tau)}$$

$$= \sum_{\tau=-N+1}^{N-1} \delta(\tau) e^{-j\omega\tau} \neq \Gamma(u) \text{ perché gli indici delle sommatorie sono diversi!}$$

NON CORRETTO

Se però $N \rightarrow \infty$, abbiamo che:

$$E_s[\hat{\Gamma}_N(u)] \xrightarrow{N \rightarrow \infty} \Gamma(u) \Rightarrow \hat{\Gamma}_n \text{ è } \underline{\text{ASINTOTICAMENTE CORRETTO}}$$

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 incomunicato con l'errore di stima ad una frequenza u_2

↓
difficile ridurre le misure delle stime

STIMATORE ALTERNATIVO

Usiamo la variante non corretta $\hat{f}'_N(\tau)$ invece di $\hat{f}_N(\tau)$. Abbiamo che:

$$\hat{\Gamma}'_N(u) := \sum_{\tau=-N+1}^{N-1} \hat{f}'_N(\tau) e^{-j\omega\tau} = \sum_{\tau=-N+1}^{N-1} \left[\frac{1}{N} \sum_{s=1}^{N-1} y(s) y(s+\tau) \right] e^{-j\omega\tau}$$

$$\Rightarrow \boxed{r=s+\tau} \Rightarrow \boxed{r=s} \quad \begin{array}{l} \bullet \tau = -N+1 \Rightarrow r=1, s=N \\ \bullet \tau = N-1 \Rightarrow r=N, s=1 \end{array}$$

$$\begin{aligned} &= \frac{1}{N} \sum_{r=1}^N \sum_{s=1}^N y(r) y(s) e^{-j\omega(r-s)} = \frac{1}{N} \sum_{r=1}^N y(r) e^{-j\omega r} \cdot \sum_{s=1}^N y(s) e^{j\omega s} \\ &= \boxed{\frac{1}{N} \cdot \left| \sum_{t=1}^N y(t) e^{-j\omega t} \right|^2} \end{aligned}$$

È il modulo del vettore ritornato dalla DFT (Discrete Fourier Transform), che porta lo spettro in un range discreto

Questo stimatore è meno corretto del precedente ma si calcola velocemente (FFT) senza passare per $\hat{f}_N(u)$

di Frequenze u

REGOLARIZZAZIONE DEGLI STIMA DELLO SPECTRO

Lo stimatore dello spettro ci gode di buone proprietà. Le stime non sono quindi buone.

↓
Un metodo per migliorare le stime è il seguente:

Optimale di dividere gli N dati del processo misurato in M parti

L calcoliamo $\hat{\Gamma}_{N/H}^{(i)}(u)$ per ciascuna parte i , $i = 1, \dots, M$

L calcoliamo la stima finale ($\hat{\Gamma}_N(u) = \frac{1}{M} \cdot \sum_{i=1}^M \hat{\Gamma}_{N/H}^{(i)}(u)$)

Si dimostra che:

$$E\left[\left(\hat{\Gamma}_N(u) - \Gamma(u)\right)^2\right] = \frac{1}{M} E\left[\left(\hat{\Gamma}_N(u) - \Gamma(u)\right)^2\right]$$

obbiamo cioè ridotto
 di $\frac{1}{M}$ la varianza dell'errore
 di stima

Osservazione

Le scelte di M rappresenta un trade-off. Infatti se regoliamo troppo (M grande), avrò uno stimatore meno corretto (perché useremo dati per le stime, e lo stimatore è solo asintoticamente corretto)

↓
Lo stesso trade-off esiste quando otteniamo visto le regolarizzazioni di funzionali.

$$J(\theta) = \frac{1}{N} \sum_{i=1}^N (y(i) - \varphi \hat{\Gamma}^{(i)}(\theta))^2 + \lambda \sum_{j=1}^d \theta_j^2$$

λ GRANDE \Rightarrow +BIOSS
 - VARIANCE

(B)
77

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 retardo pura $n=1$, le domande di generalità? No, perché, ad esempio, se K fosse = 2, identificheremmo $b_0 = 0$

Cifra di merito

È la varianza condizionata dell'errore di predizione

$$J_N(\theta) = \frac{1}{N} \sum_{t=1}^N (y(t) - \hat{y}(t|t-1, \theta))^2$$

Preditore:

$$\begin{aligned} \hat{y}(t|t-1) &= \frac{B(z)}{A(z)} u(t-1) + \frac{1}{A(z)} y(t) \\ &\quad \text{②} \quad \text{①} \rightarrow C(z) \\ &= (b_0 + b_1 z^{-1} + \dots + b_p z^{-p}) u(t-1) + (-a_1 z^{-1} - a_2 z^{-2} - \dots - a_m z^{-m}) y(t) \\ &= b_0 u(t-1) + b_1 u(t-2) + \dots + b_p u(t-p-1) - a_1 y(t-1) - a_2 y(t-2) - \dots - a_m y(t-m) \end{aligned}$$

VETTORE DEI PARAMETRI

$$\theta = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \\ b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} \in \mathbb{R}^{m+p+1} \times d$$

VETTORE DELLE OSSERVAZIONI (dei dati)

$$\varphi(t) = \begin{bmatrix} -y(t-1) \\ -y(t-2) \\ \vdots \\ -y(t-m) \\ u(t-1) \\ u(t-2) \\ \vdots \\ u(t-p-1) \end{bmatrix} \in \mathbb{R}^{m+p+1} \times d$$

$$\hat{y}(t|t-1) = \varphi(t)^T \cdot \theta$$

$$J_N(\theta) = \frac{1}{N} \sum_{t=1}^N (y(t) - \varphi(t)^T \cdot \theta)^2 \Rightarrow \boxed{\text{SOMMA A MINIMI QUADRATI}}$$

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

Osservazione

In generale per un ARx(m, p+1) è più 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: $\hat{y}(t|t-1) = \frac{B(z)}{C(z)} u(t-1) + \frac{C(z) - A(z)}{C(z)} y(t)$

$$\begin{aligned} \text{Cifre di misura, obiettivo } h &= \max(m, p+1) \\ &= 1 \end{aligned} \quad \begin{aligned} J_{10}(\theta) &= \frac{1}{10-1} \sum_{t=11}^{10} (y(t) - \hat{y}(t|t-1))^2 = \frac{1}{9} \sum_{t=2}^{10} (y(t) - bu(t-1) + \alpha y(t-1))^2 \end{aligned}$$

der partire da t=2 se w
non riesce a calcolare u(t-1), y(t-1)

reflessione
 $x_1(t)$ $x_2(t)$

$$\begin{cases} \frac{dJ_{10}(\theta)}{da} = \frac{2}{9} \sum_{t=2}^{10} (y(t) - bu(t-1) + \alpha y(t-1)) y(t-1) = 0 \\ \frac{dJ_{10}(\theta)}{db} = \frac{2}{9} \sum_{t=2}^{10} (y(t) - bu(t-1) + \alpha y(t-1)) (-u(t-1)) = 0 \end{cases} \quad \boxed{\text{OPERAZIONE}}$$

$$\Rightarrow \begin{bmatrix} \sum_{t=2}^{10} y(t-1)^2 & -\sum_{t=2}^{10} u(t-1)y(t-1) \\ -\sum_{t=2}^{10} u(t-1)y(t-1) & \sum_{t=2}^{10} u(t-1)^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} -\sum_{t=2}^{10} y(t)y(t-1) \\ \sum_{t=2}^{10} y(t)u(t-1) \end{bmatrix}$$

$$\begin{bmatrix} \hat{a}_{10} \\ \hat{b}_{10} \end{bmatrix} = \begin{bmatrix} \sum_{t=2}^{10} y(t-1)^2 & -\sum_{t=2}^{10} u(t-1)y(t-1) \\ -\sum_{t=2}^{10} u(t-1)y(t-1) & \sum_{t=2}^{10} u(t-1)^2 \end{bmatrix}^{-1} \cdot \begin{bmatrix} -\sum_{t=2}^{10} y(t)y(t-1) \\ \sum_{t=2}^{10} y(t)u(t-1) \end{bmatrix}$$

Se si considera il problema RER per modelli AR(1) può essere formulata del punto di vista matriciale:

MATRICE DEI DEGRASSI

$$\Phi = \begin{bmatrix} y(t-1) & u(t-1) \\ -y(1) & u(1) \\ -y(2) & u(2) \\ | & | \\ -y(s) & u(s) \end{bmatrix}_{(N-s) \times d}$$

$$\hat{\theta}_N = \left(\Phi^T \Phi \right)^{-1} \Phi^T Y$$

VETTORE OUTPUT

$$Y = \begin{bmatrix} y(2) \\ y(3) \\ | \\ y(s) \end{bmatrix}_{s \times 1}$$

$$Y = \Phi \theta \Rightarrow g(t|t-1) = -\alpha y(t-1) + b u(t-1)$$

$$y(2) = -\alpha y(1) + b u(1)$$

$$y(3) = -\alpha y(2) + b u(2)$$

$$y(s) = -\alpha y(s-1) + b u(s)$$

Esempio

Si supponga 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): $y(t) = \alpha y(t-1) + e(t)$ $e(t) \sim \text{unif}(0, \sigma^2)$
Usando il modello identificato, si calcoli $\hat{\theta}(6|5)$ e σ^2

Note

Se si vede comparsa $\hat{m}_s = \frac{1}{5} \sum_{t=1}^5 y(t) = -0,15$ con \hat{e} nullo. L'esercizio però ci dice di considerare media nulla, ulteriormente ovviamente la nostra definizione di fine di serie un predittore corretto, $E[\text{ECA}] = 0$

Calcoliamo il predittore

$$y(t) = \frac{1}{1-\alpha^2} e(t) \quad (\text{Supponendo } |\alpha| < 1) \Rightarrow \hat{y}(t|t-1) = \frac{(C+1)-AC+1}{C+1} y(t) = \frac{1+\alpha^{t-1}}{1} y(t) = \alpha y(t-1)$$

$$J_N(\theta) = \frac{1}{N-1} \sum_{t=2}^N (y(t) - \alpha y(t-1))^2$$

$$= \frac{1}{4} \left[(y(2) - \alpha y(1))^2 + (y(3) - \alpha y(2))^2 + (y(4) - \alpha y(3))^2 + (y(5) - \alpha y(4))^2 \right]$$

$$= \frac{1}{4} \left[\left(0 - \alpha \frac{1}{2} \right)^2 + (-1 - \alpha \cdot 0)^2 + \left(-\frac{1}{2} + \alpha \cdot 1 \right)^2 + \left(\frac{1}{4} + \alpha \cdot \frac{1}{2} \right)^2 \right]$$

(80)

$$= \frac{1}{4} \left[\frac{1}{4} \alpha^2 + \cancel{\left(1 + \frac{1}{4} \right)} + \alpha^2 \cdot \alpha + \cancel{\left(\frac{1}{16} \right)} + \frac{1}{4} \alpha^2 + \frac{1}{4} \alpha \right]$$

$$= \frac{1}{4} \left[\frac{16+4+1}{16} + \frac{-4\alpha + \alpha}{4} + \frac{\alpha^2 + \alpha^2 + 4\alpha^2}{4} \right] = \frac{1}{4} \left[\frac{21}{16} - \frac{3}{4}\alpha + \frac{3}{2}\alpha^2 \right]$$

Minimizzazione

$$\frac{dJ_5(\alpha)}{d\alpha} = 3\alpha - \frac{3}{4} = 0 \Rightarrow \boxed{\hat{\alpha}_s = \frac{1}{4}}$$

Modello identificato

$$\boxed{y(t) = \frac{1}{1 - \frac{1}{4}z^{-1}} e(t) \quad e(t) \sim \text{WN}(0, \sigma^2)}$$

Osservazione

Se avessimo ottenuto $|\hat{\alpha}_s| > 1$, allora potrei usare un filtro passatutto

$$\sigma^2 = \text{Var}[e(t)] = \text{Var}[E(t)] \approx J_5(\hat{\alpha}_s) = \frac{1}{4} \left[\frac{21}{16} - \frac{3}{4} \cdot \frac{1}{4} + \frac{3}{2} \cdot \left(\frac{1}{4}\right)^2 \right]$$

Clausura $\hat{y}(6|5)$

$$\hat{y}(t|t-1) = \frac{1}{4} y(t-1) \Rightarrow \hat{y}(6|5) = \frac{1}{4} y(5) = \frac{1}{4} \cdot \frac{1}{4} = \boxed{\frac{1}{16}}$$

IDENTIFICAZIONE 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 VVN Gaussiano, l'approccio per estensione al metodo ML

Dati disponibili

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

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

Generico modello ARMAX ($m, n, p+1$)

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

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \dots + a_m z^{-m}$$

$$B(z) = b_0 + b_1 z^{-1} + b_2 z^{-2} + \dots + b_p z^{-p}$$

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

(81)

Vettore dei parametri

$$\boldsymbol{\theta} = \begin{bmatrix} \alpha_1 \\ | \\ Q_m \\ b_0 \\ | \\ b_p \\ c_1 \\ | \\ C_m \end{bmatrix} \in \mathbb{R}^{(m+m+p+1) \times 1}$$

Approssimazione produttiva

$$\hat{\boldsymbol{\theta}}_N = \underset{\boldsymbol{\theta}}{\operatorname{arg\min}} J_N(\boldsymbol{\theta})$$

$$J_N(\boldsymbol{\theta}) = \frac{1}{N} \sum_{t=1}^N E(t, \boldsymbol{\theta})^2$$

Errore di predizione ad un passo

$$\text{Se } k=1 \Rightarrow E(z) = 1 \Rightarrow E(t) = e(t)$$

$$e(t) = E(t, \boldsymbol{\theta}) = \frac{A(z)}{C(z)} y(t) - \frac{B(z)}{C(z)} u(t-1)$$

$$J_N(\boldsymbol{\theta}) = \frac{1}{N} \sum_{t=1}^N \left(\frac{A(z)}{C(z)} y(t) - \frac{B(z)}{C(z)} u(t-1) \right)^2$$

PROBLEMA

A causa di $C(z)$ di denominazione, la cifra di merito non è quadratica in $\boldsymbol{\theta}$ (e le, in genere, più minimi locali)

Si usano quindi metodi iterativi per la minimizzazione

già visto con Logistic Regression!

PROBLEMA MINIMI LOCALI

- Si gestisce così:
- Scegliendo N inizializzazioni diverse, otengo N soluzioni
 - Se le N soluzioni sono uguali, posso pensare (non sono certi) di aver raggiunto il minimo globale d. $J_N(\boldsymbol{\theta})$
 - Se sono diverse, considerar quella che mi fa dtr $J_N(\boldsymbol{\theta})$ MINORE

METODO DI NEWTON

Idea: sviluppo in serie di Taylor troncato al secondo ordine di $J_N(\boldsymbol{\theta})$ nell'intorno di $(\boldsymbol{\theta}^i)$ (NOTO)

$$J_N(\boldsymbol{\theta}) \approx V(\boldsymbol{\theta}) = J_N(\boldsymbol{\theta}^i) + \underbrace{\left. \frac{dJ(\boldsymbol{\theta})}{d\boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^i}}_{\text{GRADIENTE}} \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}^i) + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}^i)^T \underbrace{\left. \frac{d^2 J(\boldsymbol{\theta})}{d\boldsymbol{\theta}^2} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^i}}_{\text{MESSIANA}}$$

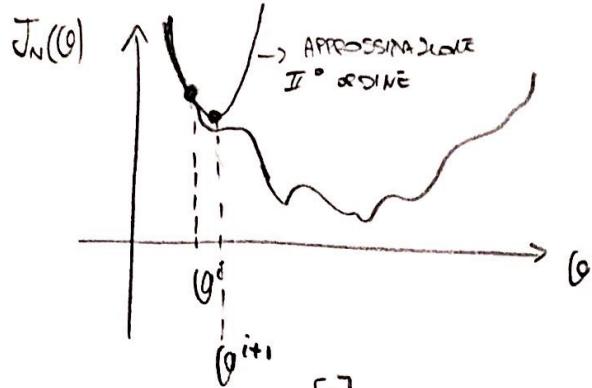
IPOTESI
GRADIENTE è un
VETTORE RIGA



MESSIANA
[] []

82

Disegno a θ^{i+1} il minimo di $V(\theta)$ ottenuto a $\theta^i \Rightarrow$ trov il minimo della parabola



Minimo di $V(\theta)$

$$\left[\frac{dV(\theta)}{d\theta} \right]_{\theta=\theta^i} = 0 \Rightarrow \left[\frac{dJ_N(\theta)}{d\theta} \Big|_{\theta=\theta^i} \right]^T + \frac{1}{2} \cdot 2 \cdot \left. \frac{d^2 J(\theta)}{d\theta^2} \right|_{\theta=\theta^i} \cdot (\theta - \theta^i) = 0$$

$$[\square] = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Rightarrow \boxed{\theta^{i+1} = \theta^i - \left[\frac{d^2 J_N(\theta)}{d\theta^2} \Big|_{\theta=\theta^i} \right]^{-1} \cdot \left[\frac{dJ_N(\theta)}{d\theta} \Big|_{\theta=\theta^i} \right]^T}$$

È simile al gradient descent se si mette costante



Il metodo funziona se l'Hessiana è **SEMI-DEFINITA POSITIVA**,
altrimenti vedi nelle direzioni seguenti

$$\left[\frac{d^2 J_N(\theta)}{d\theta^2} \Big|_{\theta=\theta^i} \right] = \alpha > 0$$

Calcoliamo i componenti dei gradiente.

- $\frac{d^2 J_N(\theta)}{d\theta^2} \Big|_{\theta=\theta^i}$ HESSIANO

- $\frac{dJ_N(\theta)}{d\theta} \Big|_{\theta=\theta^i}$ GRADIENTE

Calcol di $\frac{dJ_N(\theta)}{d\theta}$

Ricordiamo che: $J_N(\theta) = \frac{1}{N} \sum_{t=1}^N \epsilon(t)^2 \Rightarrow$

$$\left[\frac{dJ_N(\theta)}{d\theta} \right] = \frac{2}{N} \sum_{t=1}^N \epsilon(t) \cdot \frac{d\epsilon(t)}{d\theta}$$

- Calcolare di $\frac{d^2 J_N(\theta)}{d\theta^2}$ [1] [→] ⇒ definire le derivate I^o, regole delle derivazioni del prodotto

$$\frac{d^2 J_N(\theta)}{d\theta^2} \underset{[\equiv]}{=} \frac{2}{N} \sum_{t=1}^N \left[\frac{dE(t)}{d\theta} \right]^T \left[\frac{dE(t)}{d\theta} \right] + \frac{2}{N} \sum_{t=1}^N E(t) \frac{d^2 E(t)}{d\theta^2}$$

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 interessano i termini definiti positivo → la procedura è sicuramente di MINIMIZZAZIONE

Osserviamo quindi:

$$[\boxed{1}] \quad \theta^{(i+1)} = \theta^{(i)} - \left[\frac{2}{N} \sum_{t=1}^N \left(\frac{dE(t)^{(i)}}{d\theta} \right)^T \left(\frac{dE(t)^{(i)}}{d\theta} \right) \right]^{-1} \cdot \left[\frac{2}{N} \sum_{t=1}^N E(t)^{(i)} \cdot \frac{dE(t)^{(i)}}{d\theta} \right]$$

per garantire l'invertibilità, si offre un termine

Note + dI piccolo, matrice identità

La notazione "picce $c(i)$ " indica che stiamo voltando θ in $\theta^{(i)}$ (noto)

- Calcolare di $\frac{dE(t)}{d\theta}$

$$E(t) = e(t) = \frac{A(z)}{C(z)} y(t) - \frac{B(z)}{C(z)} u(t-1)$$

$$E(t) = \frac{1+a_1 z^{-1} + \dots + a_m z^{-m}}{1+c_1 z^{-1} + \dots + c_m z^{-m}} y(t) - \frac{b_0 + b_1 z^{-1} + \dots + b_p z^{-p}}{1+c_1 z^{-1} + \dots + c_m z^{-m}} u(t-1)$$

$$\theta = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \\ b_0 \\ \vdots \\ b_p \\ c_1 \\ \vdots \\ c_m \end{bmatrix} \in \mathbb{R}^{(m+p+m+1) \times 1}$$

Derivate di $E(t)$ rispetto a a_1, a_2, \dots, a_m :

$$\bullet \frac{dE(t)}{da_1} = \frac{z^{-1}}{C(z)} y(t) = \frac{1}{C(z)} y(t-1) = \alpha(t-1)$$

$$\boxed{\alpha(t) = \frac{1}{C(z)} y(t)}$$

$$\bullet \frac{dE(t)}{da_2} = \frac{z^{-2}}{C(z)} y(t) = \frac{1}{C(z)} y(t-2) = \alpha(t-2)$$

$$\bullet \frac{dE(t)}{da_m} = \frac{z^{-m}}{C(z)} y(t) = \frac{1}{C(z)} y(t-m) = \alpha(t-m)$$

Derivate di $E(t)$ rispetto a b_0, b_1, \dots, b_p

$$\bullet \frac{dE(t)}{db_0} = -\frac{1}{C(z)} u(t-1) = \beta(t-1)$$

$$\boxed{\beta(t) = -\frac{1}{C(z)} u(t)}$$

$$\bullet \frac{dE(t)}{db_1} = -\frac{z^{-1}}{C(z)} u(t-1) = \beta(t-2)$$

$$\bullet \frac{dE(t)}{db_p} = -\frac{z^{-p}}{C(z)} u(t-1) = \beta(t-p-1)$$

Derivate di $E(t)$ rispetto a c_1, c_2, \dots, c_m

$$E(t) = \frac{A(z)}{C(z)} y(t) - \frac{B(z)}{C(z)} u(t-1) \Rightarrow (1 + c_1 z^{-1} + \dots + c_m z^{-m}) E(t) = A(z) y(t) - B(z) u(t-1)$$

$$\Rightarrow d[(1 + c_1 z^{-1} + \dots + c_m z^{-m}) \cdot E(t)] = d[A(z) y(t) - B(z) u(t-1)] \rightarrow 0, \text{ non dipende da } c_i$$

derivabile

$$\Rightarrow z^{-1} E(t) + C(z) \frac{dE(t)}{dc_i} = 0 \Rightarrow \bullet \frac{dE(t)}{dc_i} = -\frac{1}{C(z)} E(t-1) = \gamma(t-1)$$

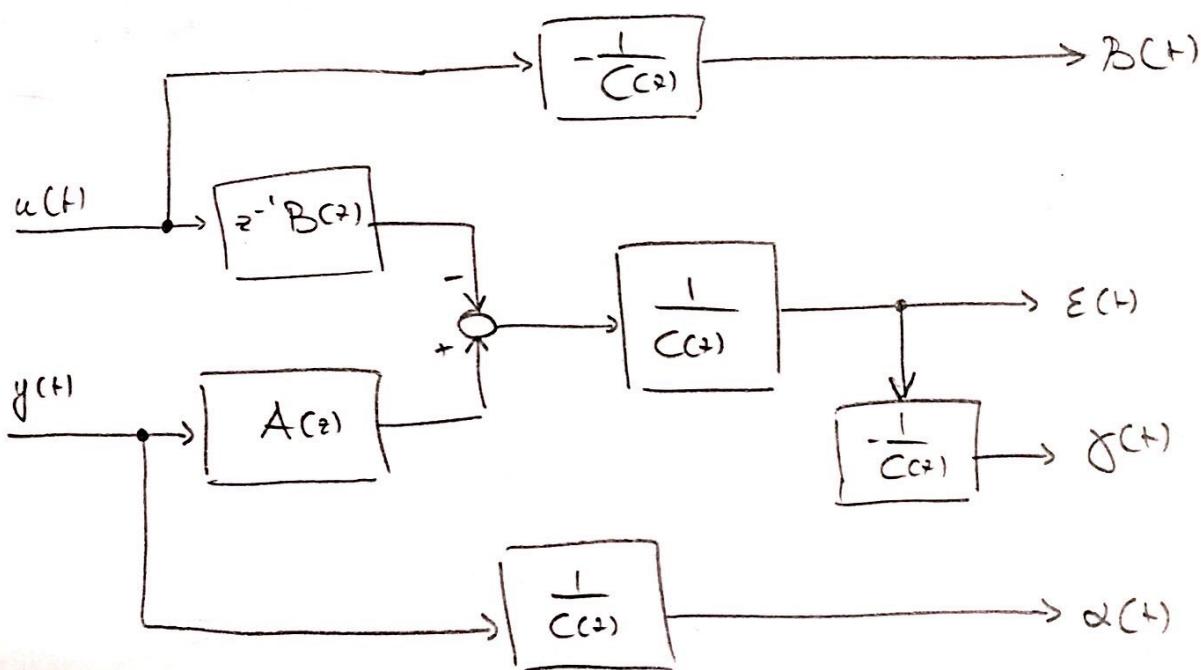
$$\bullet \frac{dE(t)}{dc_m} = -\frac{1}{C(z)} E(t-m) = \gamma(t-m)$$

$$\boxed{\gamma(t) = -\frac{1}{C(z)} E(t)}$$

Riassumendo, dobbiamo che il vettore delle domande è:

$$\frac{dE(t)}{d\theta} = \begin{bmatrix} \alpha(t-1) \\ \vdots \\ \alpha(t-m) \\ \beta(t-1) \\ \vdots \\ \beta(t-p) \\ \gamma(t-1) \\ \vdots \\ \gamma(t-m) \end{bmatrix} \in \mathbb{R}^{(m+m+p+1) \times 1} \quad | t = 1, \dots, N$$

Osserviamo quindi creare 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 $C(z)^{-1}$ sia AS-STAB: se non lo è dobbiamo rispettare i poli instabili nel calcolo anteriori (metodo di RADER).

8 IDENTIFICAZIONE: ANALISI E COMPLEMENTI

ANALISI ASINTOTICA METODO PEM

Ipotizziamo di avere N dati $\{y_{(1)}, y_{(2)}, \dots, y_{(N)}\}, \{u_{(1)}, u_{(2)}, \dots, u_{(N)}\}$

la stima PEM trova $\hat{\theta}_N$ da minimizzare: $J_N(\theta) = \frac{1}{N} \sum_{t=1}^N e(t, \theta)^2$

Come facciamo a sapere che questa stima è (almeno asymptoticamente) buona?
Considerando $N \rightarrow \infty$, abbiamo che:

$$J_N(\theta) \xrightarrow[N \rightarrow \infty]{} J(\theta) = E[e(t, \theta)^2]$$

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 $S \in M(\theta)$

S : SISTEMA VERO

$M(\theta)$: CLASSE DI MODELLI M CON PARAMETRI θ

$$\downarrow$$

$$SM(\theta^\circ) = S$$

θ° : VETTORE VERO DI PARAMETRI

Ci chiediamo se θ° appartiene all'insieme Δ dei minimi globali di $J(\theta)$

Dimostrazione

Consideriamo un modello $M(\theta)$ e scriviamo l'errore di predizione:

$$e(t) = y(t) - \hat{y}(t|t-1, \theta) \Rightarrow e(t) - \hat{y}(t|t-1, \theta^\circ) = \underbrace{y(t) - \hat{y}(t|t-1, \theta^\circ)}_{\text{RUMORE BIANCO}} - \hat{y}(t|t-1, \theta^\circ)$$

$$\Rightarrow e(t) = e(t) - \hat{y}(t|t-1, \theta) + \hat{y}(t|t-1, \theta^\circ)$$

RUMORE BIANCO $e(t)$ che sfiora il sistema (è l'errore di predizione ad un passo)

Applichiamo $E[(\cdot)^2]$ ad entrambi i membri:

$$E[e(t)^2] = E\left[\left(e(t) + \left(\hat{y}(t|t-1, \theta^\circ) - \hat{y}(t|t-1, \theta)\right)\right)^2\right] =$$

considerare gli ultimi 2 termini
nella loro interezza

$$\downarrow J(\theta) = E[e(t)^2] + E\left[\left(\hat{y}(t|t-1, \theta^\circ) - \hat{y}(t|t-1, \theta)\right)^2\right] + 2E\left[e(t)\left(\hat{y}(t|t-1, \theta^\circ) - \hat{y}(t|t-1, \theta)\right)\right]$$

REMARK: $E[e(t)] = 0$ perché usiamo predizioni corrette

predittori incostanti
con $e(t)$

87

$$J(\theta) = \lambda^2 + E\left[\left(\hat{g}(t|t-1; \theta^*) - \hat{g}(t|t-1; \theta)\right)^2\right] + 0$$

≥ 0 , si annulla per $\theta = \theta^*$

$$\Rightarrow J(\theta) \geq \lambda^2 = J(\theta^*), \forall \theta$$

↓

$J(\theta) \geq J(\theta^*) \quad \forall \theta$

Conclusione

Se $S \in M(\theta)$, il metodo PEA è in grado di garantire che il modello stima è quello veri, (nel caso in cui $N \rightarrow \infty$)

L PEA è ASINTOTICAMENTE CORRETTO

L PEA è ASINTOTICAMENTE CONSISTENTE (si minimizza le variazioni dell'errore)

Se la classe di modelli è sbagliata, i metodi PEA non convergeranno sui parametri veri

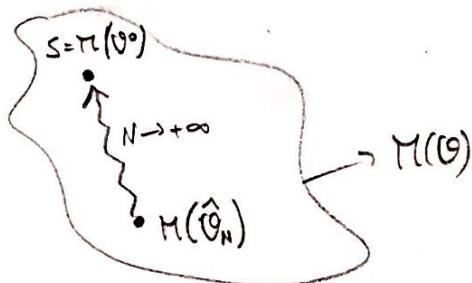
Osservazione

Se $S \in M(\theta)$, in corrispondenza di θ^* si ha che $E(t; \theta^*) = e(t)$ non

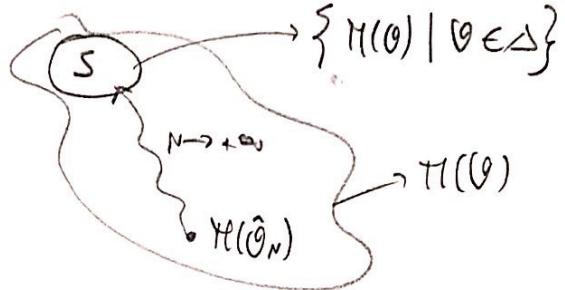
↓
Si può fare un test di banchese per verificare che il modello identificato sia quello vero.

Quando identifichiamo un modello, possono capitare diverse situazioni:

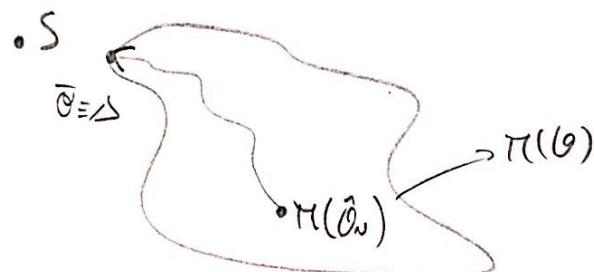
- 1) $S \in M(\theta)$ e $\Delta = \bar{\theta} \equiv \theta^*$. La famiglia di modelli scelta è quella del sistema vero. $J(\theta)$ ha un minimo che è θ^* . Osserva ormai che $\hat{\theta}_N \xrightarrow[N \rightarrow +\infty]{} \theta^*$



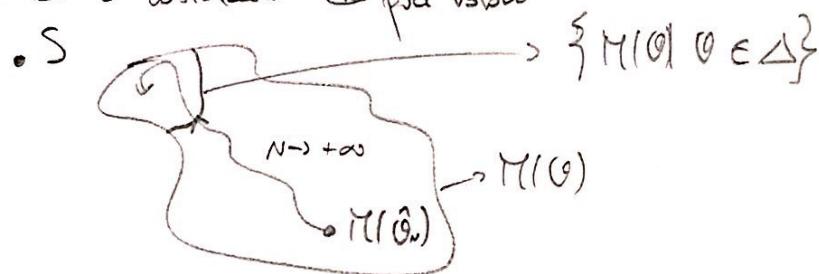
- 2) $S \in \mathcal{H}(\theta)$ ma Δ è costituito da più voci. Non è un problema tanto l'identificazione trova un set di parametri con le stesse voci delle cifre di merito, sono equivalenti del punto di vista predittivo



- 3) $S \notin \mathcal{H}(\theta)$ e $\Delta = \bar{\Theta}$. Si ottiene il modello $H(\bar{\theta})$ miglior approssimante di S nella classe di modelli $\mathcal{H}(\theta)$ scelta



- 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. Ammettiamo che otteniamo le stime (nel caso ARX($m, p+1$))

$$y(t) = \frac{B(z)}{A(z)} u(t-1) + \frac{1}{A(z)} e(t) \quad e(t) \sim \text{N}(0, \sigma^2) \quad B(z) = b_0 + b_1 z^{-1} + \dots + b_p z^{-p}$$

$$\hat{\theta}_N = \left[\sum_{t=1}^N \varphi(H) \varphi(H)^T \right]^{-1} \cdot \left[\sum_{t=1}^N \varphi(H) y(t) \right] = (\bar{\Phi}^T \bar{\Phi})^{-1} \bar{\Phi}^T y$$

$$A(z) = 1 + a_1 z^{-1} + \dots + a_m z^{-m}$$

PROBLEMA DI IDENTIFICABILITÀ

Quando $\hat{\theta}_N$ esiste ed è unica? \Leftrightarrow Quando $\sum_{t=1}^N \varphi(H) \varphi(H)^T$ è invertibile?

$$S(N) = \sum_{t=1}^N \varphi(H) \varphi(H)^T \Rightarrow \hat{\theta}_N = S(N)^{-1} \cdot \left[\sum_{t=1}^N \varphi(H) y(t) \right]$$

$$R(N) = \frac{1}{N} S(N) = \frac{1}{N} \sum_{t=1}^N \varphi(H) \varphi(H)^T \Rightarrow \hat{\theta}_N = R(N)^{-1} \left[\frac{1}{N} \sum_{t=1}^N \varphi(H) y(t) \right]$$

$R(N)$ è ≥ 0 in quanto prodotto di un vettore per se stesso. Offiché $\hat{\theta}_N$ esista esistono positivi

e sia unica occorre per che $R(N) > 0$ (DEF. POS) e quindi abbia tutti gli

Consideriamo il caso ASINTOTICO $N \rightarrow +\infty \Rightarrow R(N) \xrightarrow[N \rightarrow \infty]{d} \bar{R}$

Perc un ARX($m, p+1$), \bar{R} è una matrice quadrata $\underbrace{(m+p+1)}_{d} \times \underbrace{(m+p+1)}_{d}$ t.c:

$$\bar{R} = \begin{bmatrix} \bar{R}_y & & -\bar{R}_{yu} \\ & \ddots & \\ -\bar{R}_{uy} & & \bar{R}_u \end{bmatrix} \in \mathbb{R}^{(m+p+1) \times (m+p+1)}$$

$$\bullet \text{ARX}(1,1) \Rightarrow \varphi(H) = \begin{bmatrix} -y(t-1) \\ u(t-1) \end{bmatrix} \Rightarrow R(N) = \frac{1}{N} \sum_{t=1}^N \varphi(H) \cdot \varphi(H)^T = \frac{1}{N} \sum_{t=1}^N \begin{bmatrix} -y(t-1) \\ u(t-1) \end{bmatrix} \begin{bmatrix} -y(t-1) & u(t-1) \end{bmatrix}$$

$$= \frac{1}{N} \sum_{t=1}^N \begin{bmatrix} -y(t-1)^2 & -y(t-1)u(t-1) \\ -u(t-1)y(t-1) & u(t-1)^2 \end{bmatrix}$$

Se $N \rightarrow +\infty$ dobbiamo che la stima della covariante è ASINT. CORRETTA

$$\bar{R} = \begin{bmatrix} E[y(t)^2] & -E[u(t-1)y(t-1)] \\ -E[u(t-1)y(t-1)] & E[u(t-1)^2] \end{bmatrix}$$

IPOTESI
 $u(t)$ e $y(t)$ sono PSS

$$= \begin{bmatrix} \gamma_y(0) & \bar{R}_{y\bar{y}} \\ -\bar{R}_{y\bar{u}} & \gamma_u(0) \end{bmatrix} \rightarrow \begin{array}{l} \bar{R}_{y\bar{y}} \\ -\bar{R}_{y\bar{u}} \\ \bar{R}_{\bar{u}\bar{u}} \end{array}$$

\bar{R} è la matrice di VARIANZE-COVARIANZE

In generale:

$$\bar{R}_y = \begin{bmatrix} \gamma_y(0) & \gamma_y(1) & \cdots & \gamma_y(m-1) \\ \gamma_y(1) & \gamma_y(0) & \gamma_y(1) & \vdots \\ \vdots & \vdots & \ddots & \gamma_y(0) \\ \gamma_y(m-1) & \gamma_y(m-1) & \cdots & \gamma_y(0) \end{bmatrix}$$

- Matrice $m \times m$
- Matrice covariante di ordine m di $y(t)$
- Toeplitz

$$\bar{R}_u = \begin{bmatrix} \gamma_u(0) & \gamma_u(1) & \cdots & \gamma_u(p) \\ \gamma_u(1) & \gamma_u(0) & \gamma_u(1) & \vdots \\ \vdots & \vdots & \ddots & \gamma_u(0) \\ \gamma_u(p) & \gamma_u(p) & \cdots & \gamma_u(0) \end{bmatrix}$$

- Matrice $(p+1) \times (p+1)$
- Matrice covariante ordine p di $u(t)$
- Toeplitz

$$\bar{R}_{yu} = \begin{bmatrix} \gamma_{yu}(0) & \gamma_{yu}(1) & \cdots & \gamma_{yu}(p) \\ \gamma_{yu}(1) & \gamma_{yu}(0) & \gamma_{yu}(1) & \vdots \\ \vdots & \vdots & \ddots & \gamma_{yu}(0) \\ \gamma_{yu}(p) & \gamma_{yu}(p) & \cdots & \gamma_{yu}(0) \end{bmatrix}$$

- Matrice rettangolare $m \times (p+1)$
- Matrice covariante $u(t)$ e $y(t)$
- $\bar{R}_{yu} = \bar{R}_{yu}^T$

Vogliamo una condizione per l'invertibilità di \bar{R}

Lemme di Schur

Дата матрица H nella forma $H = \begin{bmatrix} F & K \\ K^T & H \end{bmatrix}$, con F e H симметрическими

condizione necessaria e sufficiente affinché $H \geq 0$ è che: - $H \geq 0$

$$- F - KH^{-1}K^T \geq 0$$

$$\bar{R} = \begin{bmatrix} -R_y & -R_{yu} \\ -R_{uy} & \bar{R}_u \end{bmatrix} \Rightarrow \boxed{\begin{array}{l} \text{CONDIZIONE NECESSARIA} \\ \bar{R}_u \geq 0 \end{array} \text{ per invertire } \bar{R} \text{ è che}}$$

la seconda condizione ($F - KH^{-1}K^T \geq 0$) è difficile da risolvere. Cerchiamo di impostare almeno la prima

La condizione $\bar{R}_u \geq 0$ riguarda solo l'impulso $u(t)$, che progettiamo noi!

Sia:

$$\bar{R}_u^{(i)} = \begin{bmatrix} \delta_{u(0)} & \delta_{u(1)} & \cdots & \delta_{u(i-1)} \\ \delta_{u(1)} & \ddots & & \delta_{u(i-2)} \\ \vdots & & \ddots & \delta_{u(i-1)} \\ \delta_{u(i-1)} & & & \delta_{u(0)} \end{bmatrix}$$

la matrice di covarianza di $u(t)$, d'ordine i

Il segnale $u(t)$ è detto PERSISTENTEMENTE ECITANTE DI ORDINE m se:

- $\bar{R}_u^{(1)} \geq 0, \bar{R}_u^{(2)} \geq 0, \dots, \bar{R}_u^{(m)} \geq 0$

- $\bar{R}_u^{(m+1)} \geq 0, \bar{R}_u^{(m+2)} \geq 0, \dots \geq 0$

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 ARX($m, p+1$) è che il segnale $u(t)$, usato per produrre i dati, sia "persistente mente eccitante" d'ordine più ad almeno $p+1$.

Osservazione

Consideriamo $u(t) \sim WN(0, \sigma^2)$. Abbiamo che:

$$\bar{R}_u^{(1)} = \begin{bmatrix} \sigma^2 & 0 & 0 & -0 \\ 0 & \sigma^2 & \sigma^2 & \sigma^2 \\ 1 & \sigma^2 & \sigma^2 & \sigma^2 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix} = \sigma^2 \cdot I^{(4)}$$

- 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)
 - Le eccitazioni con $u(t) \sim WN$
 - Il modello usato è ARX(3,3)
- Dato che il modello è SOVAPARAMETRIZZATO
esistono infinite soluzioni (tutte le cancellazioni pol.-zer.)
- \downarrow

SCEGLI COMPLICATITÀ DEL MODELLO

- Affinché un modello sia univocamente identificabile occorre avere:
- 1) IDENTIFICABILITÀ "STRUTTURALE": il modello non deve essere sovraparametrizzato rispetto al sistema
 - 2) IDENTIFICABILITÀ "SPERIMENTALE": i dati devono contenere sufficiente informazione

VALUTAZIONE DELL'INCERTITUDINE

Le analisi fatte in precedenza si basavano sull'ipotesi che $N \rightarrow +\infty$. Nella realtà otteniamo N finiti.

HIPOTESI

- $\Theta \in \mathcal{H}(\Theta) \Rightarrow \Theta^0 \in \Delta$
- Se $\Delta \ni \bar{\Theta} \Rightarrow \bar{\Theta} = \Theta^0$

La varianza dello stimatore è:

$$\text{Var}[\hat{\Theta}_N] = \frac{1}{N} \bar{\sigma}^2 \cdot \bar{M}^{-1}$$

Stimiamo $\hat{\Theta}_N$ con N dati.

$$\hat{\Theta}_N = \underset{\Theta}{\operatorname{arg \min}} J_N(\Theta) = \frac{1}{N} \sum_{t=1}^N \epsilon(t, \Theta)^2$$

→ E' UNA VARIABILE CASUALE!

Dalle ipotesi abbiamo che $E[\hat{\Theta}_N] = \Theta^0$

$$\bullet \bar{\sigma}^2 = \text{Var}[\epsilon(t)] = \text{Var}\left[g(t) \cdot \hat{g}(t|t_{-1}, \Theta^0)\right]$$

$$\bullet \bar{M} = E\left[\left(\frac{d\epsilon(t, \Theta)}{d\Theta}\Big|_{\Theta=\Theta^0}\right)^T \cdot \left(\frac{d\epsilon(t, \Theta)}{d\Theta}\Big|_{\Theta=\Theta^0}\right)\right]$$

Come stimiamo in pratica $\bar{\sigma}^2$ e \bar{M} ?

$$\boxed{\bar{\sigma}^2 = E[\epsilon(t, \Theta^0)] \approx E[\epsilon(t, \hat{\Theta}_N)] \approx \frac{1}{N} \sum_{t=1}^N \left(g(t) - \hat{g}(t|t_{-1}, \hat{\Theta}_N)\right)^2 = \boxed{J_N(\hat{\Theta}_N)}}$$

$$\boxed{\bar{M} \approx \hat{M} = \frac{1}{N} \sum_{t=1}^N \left[\left(\frac{d\epsilon(t, \Theta)}{d\Theta}\Big|_{\Theta=\hat{\Theta}_N} \right)^T \cdot \left(\frac{d\epsilon(t, \Theta)}{d\Theta}\Big|_{\Theta=\hat{\Theta}_N} \right) \right]}$$

Interpretazione di \bar{M}

$$\begin{aligned} &\text{Ricordiamo che } J(\Theta) : E[\epsilon(t, \Theta)^2] \Rightarrow \frac{dJ(\Theta)}{d\Theta} = E\left[2\epsilon(t, \Theta) \cdot \frac{d\epsilon(t, \Theta)}{d\Theta}\right] \\ &\Rightarrow \frac{d^2J(\Theta)}{d\Theta^2} = E\left[2 \frac{d\epsilon(t, \Theta)}{d\Theta}^T \cdot \frac{d\epsilon(t, \Theta)}{d\Theta} + 2\epsilon(t, \Theta) \frac{d^2\epsilon(t, \Theta)}{d\Theta^2}\right] \end{aligned}$$

Se $\Theta = \Theta^0 \Rightarrow \epsilon(t, \Theta) = e(t) \Rightarrow \frac{d^2\epsilon(t, \Theta)}{d\Theta^2}$ è funzione dell'errore di predizione
e dipende da $e(t_{-1}), e(t_{-2}, \dots)$

↓ quindi $\frac{d\epsilon(t)}{d\Theta}$

$$\frac{d}{d\Theta} \left[g(t) - \hat{g}(t|t_{-1}, \Theta) \right] \Rightarrow \text{dipende da } e(t_{-1}, \dots)$$

il predittore ↓

costruito con $e(t_{-1}, \dots)$

39

$$e(t) = g(t) - \hat{g}(t|t_{-1}) = e(t)$$

↓

per termine $E \left[2\epsilon(t) \frac{d^2\epsilon(t)}{d\theta^2} \right] = E \left[2\epsilon(t) \cdot \left(\frac{d^2\epsilon(t)}{d\theta^2} \right) \right] = 0$

→ dipende da $\epsilon(t)$, — INCORRETTI

Osserviamo che:

$$\left. \frac{d^2 f(\theta)}{d\theta^2} \right|_{\theta=0^\circ} = 2E \left[\left(\left. \frac{d\epsilon(t, \theta)}{d\theta} \right|_{\theta=0^\circ} \right)^T \cdot \left(\left. \frac{d\epsilon(t, \theta)}{d\theta} \right|_{\theta=0^\circ} \right) \right] = 2\bar{H}$$

↓

$$\boxed{\bar{H} = \frac{1}{2} \left. \frac{d^2 J(\theta)}{d\theta^2} \right|_{\theta=0^\circ}}$$

- \bar{H} è metà dell'Hessiano delle cifre di morita volontate nell'ottimo θ^*

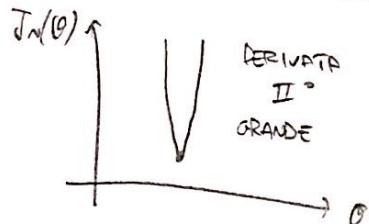
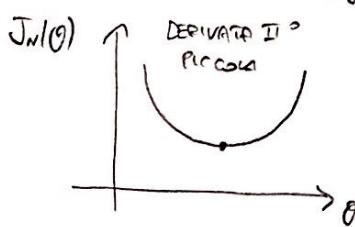
Conclusioni

$$Var[\hat{\theta}_N] = \frac{1}{N} \cdot \sigma^2 \cdot \bar{H}^{-1} = \frac{1}{N} \cdot \sigma^2 \left(\frac{1}{2} \left. \frac{d^2 J(\theta)}{d\theta^2} \right|_{\theta=0^\circ} \right)^{-1}$$

- $N \uparrow \Rightarrow Var[\hat{\theta}_N] \downarrow$

- $\sigma^2 \Rightarrow Var[\hat{\theta}_N] \uparrow$

- $\bar{H} \uparrow \Rightarrow Var[\hat{\theta}_N] \downarrow \Rightarrow$ Si vuole avere Hessiano grande nel punto di minimo
(grande variazione nell'intorno del minimo)



SCELTA DELLA COMPLESSITÀ DEL MODELLO

Scelta una classe di modelli $M(\theta)$, trovare $\hat{\theta}_N$ nel seguente modo:

$$\hat{\theta}_N = \underset{\theta}{\operatorname{argmin}} J_N(\theta) \quad J_N(\theta) = \frac{1}{N} \sum_{t=1}^N (y(t) - \hat{y}(t|t-1))^2$$

- Se $M(\theta)$ è un ARX \Rightarrow esiste una forma esplicita per $\hat{\theta}_N$
- Se $M(\theta)$ è un ARMAX \Rightarrow va risolti iterativamente un problema di ottimizzazione

[d] dimensione di θ .

Nel caso generale ARMAX abbiamo $d = m + n + p + 1$

Problema: scelta dell'ordine d del modello

Osservazione

In realtà dobbiamo scegliere 3 parametri, con 1. Per semplicità, si fissa $m = n = q$, gestendo quindi un solo parametro, d

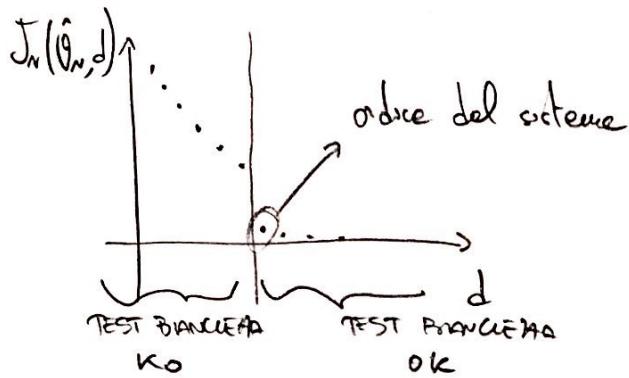
Sappiamo già che non è corretto scegliere l'ordine del modello con minor $J_N(\hat{\theta}_N)$ \Rightarrow OVERTFITTING

Vediamo quindi 3 metodi, da sì basare sull'ordine:

- N dati
- ordine del modello $d = 1, 2, 3, \dots$
- $J_N(\hat{\theta}_N, d)$ è il valore della funzione da minimo volutato all'"ottimo" (ovvero se ne trova 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 $d = 1, 2, 3, 4, \dots$

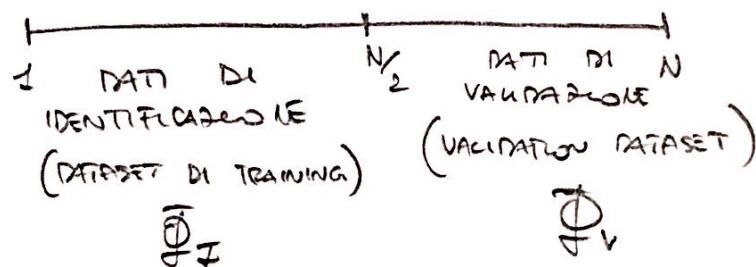


Un modo questo metodi funziona anche:

- Permette di vedere le "discontinuità"
- Il test di bianchetta va da una discontinuità rotta (da un "taglio" di valori)

METODO 2: CROSS-VALIDAZIONE

Dividere i dati in 2 sottoinsiemi



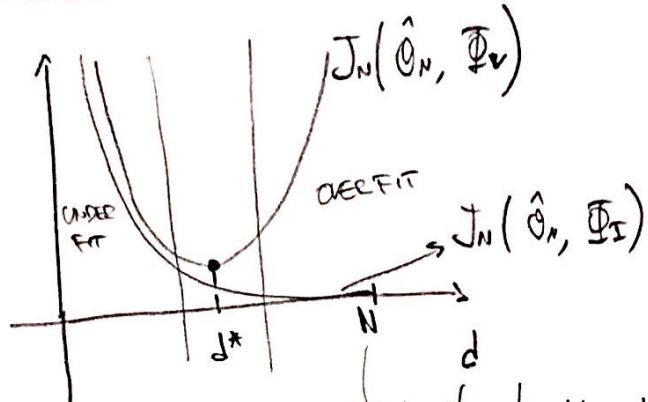
- Fissare un valore di d

- Trovare $\hat{\theta}_N$ minimizzando $J_N(\theta, \mathcal{D}_I)$
- Calcolare $J_N(\hat{\theta}_N, \mathcal{D}_I)$ e $J_N(\hat{\theta}_N, \mathcal{D}_V)$
- Ripetere per $d = 1, 2, 3, 4, \dots$, out-of-sample error

N.B.

Quell'errore da $J_N(\theta, \mathcal{D}_I)$ e $J_N(\theta, \mathcal{D}_V)$ è dipendente da d

Le differenze rispetto al caso statico è che NON POSSIAMO MISURARE I DATI, perché adesso le due dipendenze dal tempo



quando $d = N$, interpolo i dati perfettamente

La cross-validation è la procedura migliore, però richiede molti dati

METODO 3: FORMULE PER LA STIMA DELLA COMPLESSITÀ OTTIMA

Permettono 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 minimizzar

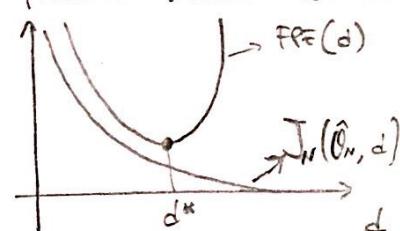
Sono state ricavate tecnicamente per ARX, ma in pratica si usano anche per ARMA

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)

$$MDL(d) = \ln(N) \cdot \frac{d}{N} + \ln[J_N(\hat{\theta}_N, d)]$$

$$\cdot d \uparrow \Rightarrow \ln(N) \cdot \frac{d}{N} \uparrow$$

$$\cdot d \uparrow \Rightarrow \ln[J_N(\hat{\theta}_N, d)] \downarrow$$

CONFRONTO FPE vs. AIC

I criteri sono simili. Se si calcola il logaritmo di FPE si ottiene:

$$\begin{aligned} \ln[FPE(d)] &= \ln\left[\frac{N+d}{N-d} \cdot J_N(\hat{\theta}_N, d)\right] = \ln\left[\frac{1+\frac{d}{N}}{1-\frac{d}{N}} \cdot J_N(\hat{\theta}_N, d)\right] = \\ &= \ln\left[\frac{1+\frac{d}{N}}{1-\frac{d}{N}}\right] + \ln[J_N(\hat{\theta}_N, d)] = \\ &= \ln\left[1 + \frac{d}{N}\right] - \ln\left[1 - \frac{d}{N}\right] + \ln[J_N(\hat{\theta}_N, d)] \end{aligned}$$

Picchiamo che $\ln(1+x) \approx x$ quando $x \approx 0$; inoltre $\frac{d}{N} \approx 0$ 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, $\boxed{\text{se } d \ll N} \Rightarrow \boxed{\ln[FPE(d)] = AIC}$ come per le z-score

Il minimo di $f(x)$ è anche il minimo di $\ln[f(x)]$, quindi i metodi sono equivalenti.

CONFRONTO AIC vs MDL

$$AIC(d) = \boxed{-2 \cdot \frac{d}{N}} + \ln[J_N(\hat{\theta}_N)] \quad \Leftrightarrow \quad MDL(d) = \ln(N) \cdot \frac{d}{N} + \ln[J_N(\hat{\theta}_N)]$$

, condizioni solo
questi termini

Se $\ln(N) > 2$, ovvero osservi 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 o 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 overfittivo

—————
FINE