

RIASSUNTO IMAD

RICHIAMI DI STATISTICA

PROPRIETA' DELLE VARIABILI CASUALI

DEF: Una variabile casuale v è una variabile definita a partire dall'esito s di un esperimento casuale. Dato che v può assumere diversi valori, assegno una probabilità che ogni esito accada (distribuzione di probabilità). $V(s) \rightarrow$ DIPENDE DALL'ESITO DELL'ESPIMENTO

1. Se v assume valori discreti (variabile casuale discreta)

Funzione di probabilità di massa (pmf): $p(x) = p(v = x) \rightarrow$ associa ad ogni valore di x una probabilità v .

$$\sum_{i=1}^m p(x_i) = 1$$

Figura 1 $x(i)$ sono i valori di v e m sono i diversi valori che v può assumere

2. Se v assume valori continui (variabile casuale continua)

Funzione di densità di probabilità (pdf): $f_v(x)$. Dato che v può assumere infiniti valori, la probabilità che v assuma esattamente un valore specifico è zero. $P(v = x) = 0$

Se la variabile v assume tutti valori equiprobabili (come per il lancio di un dado), la probabilità che v assuma un valore specifico sarebbe $1/\infty = 0$

La pdf $f_v(x)$ definisce la probabilità che v appartenga ad un intervallo di valori $[a,b]$. $P(v \in [a,b]) = \int_a^b f_v(x) dx$

La funzione di densità cumulata (cdf) è $F_v(z)$

I PESI SONO LE PROBABILITA'

Il valore atteso di una variabile casuale v è la somma pesata dei valori x che v può assumere. Il valore atteso gode della proprietà di linearità (valore atteso di una combinazione lineare pesata = somma dei valori attesi moltiplicati per il peso). L'operatore valore atteso $E_s[v]$ considera tutti i possibili esiti s della variabile casuale v .

La varianza di una variabile casuale v è quanto i valori di x si discostano dalla loro media

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

DIPENDENZA LINEARE

Figura 2 coefficiente di correlazione

(SE $\rho=0$ ALLORA E' POSSIBILE DIPENDENZA NON LINEARE)

$$\text{VAR}(v) = \mathbb{E}[(v - \mathbb{E}[v])^2] = \mathbb{E}[v^2] - 2\mathbb{E}[v]^2 + \mathbb{E}[v]^2 = \mathbb{E}[v^2] - \mathbb{E}[v]^2$$

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

Figura 3 covarianza

Vettore di variabili casuali: assumiamo che v sia una v.c. continua

Funzione di densità cumulata (cdf): $F_v(z_1, \dots, z_d)$

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

Figura 4 valore atteso

$$\begin{aligned} \text{Var}[\mathbf{v}] &= \int_{\mathbb{R}^d} (\mathbf{x} - \mathbb{E}[\mathbf{v}]) (\mathbf{x} - \mathbb{E}[\mathbf{v}])^\top f_{\mathbf{v}}(\mathbf{x}) d\mathbf{x} \\ &= \begin{bmatrix} \text{Var}[v_1] & \cdots & \text{Cov}[v_1, v_d] \\ \vdots & \ddots & \vdots \\ \text{Cov}[v_d, v_1] & \cdots & \text{Var}[v_d] \end{bmatrix} \end{aligned}$$

Figura 5 varianza di una matrice semidefinita positiva e simmetrica

$$f_{v_1, v_2}(x_1, x_2) = f_{v_1}(x_1) \cdot f_{v_2}(x_2)$$

Figura 6 due variabili casuali con funzione di probabilità congiunta indipendenti

Se sono indipendenti, allora sono anche scorrelate (non vale il viceversa).

STIMA E STIMATORI

Per gestire l'incertezza dei dati, li interpretiamo come variabili casuali. Vogliamo quindi stimare il vettore di parametri θ^0 che ha generato i dati $D = \{y(1), \dots, y(N)\}$. CASO MONETA: IN θ C'È LA PROBABILITÀ CHE ESCA TRO

Quando i dati D dipendono sia dall'esito s , sia dai parametri θ^0 allora $D(s, \theta^0)$. I dati osservati dipendono da uno specifico esito \bar{s} : $D = D(\bar{s}, \theta^0)$.

Uno stimatore è una funzione di variabili casuali. La stima è il risultato di uno stimatore su una specifica realizzazione dei dati. Lo stimatore è una variabile casuale che dipende da s .

La bontà di uno stimatore si giudica dalle caratteristiche della sua distribuzione di probabilità. Uno stimatore $\hat{\theta}$ si dice corretto se $E[\hat{\theta}] = \theta^0$, dove θ^0 è il valore del parametro.

$$\text{bias} = E[\hat{\theta}] - \theta^0 \quad \text{STIMATORE DISTORTO}$$

$$\lim_{N \rightarrow +\infty} E[\hat{\theta}] = \theta^0$$

Figura 7 stimatore asintoticamente corretto

$$\lim_{N \rightarrow +\infty} P(|\hat{\theta} - \theta^0| \geq \varepsilon) = 0$$

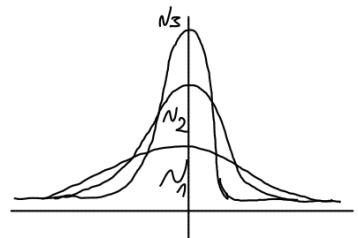


Figura 8 stimatore consistente

$$N \rightarrow +\infty$$

Al crescere di N la stima diventa sempre più precisa.

Valutare la bontà di uno stimatore per N finito: se due estimatori sono entrambi corretti, il migliore è quello a minima varianza.



Limite di Cramer-Rao: dato uno stimatore corretto, non possiamo rendere la sua varianza più piccola di una certa quantità. La quantità m (M caso vettoriale) è detta quantità (o matrice) di informazione di Fisher. Avrò sempre un certo livello di incertezza sui dati che uso per fare la stima, i dati non saranno mai "informativi al 100%".

Uno stimatore si dice efficiente se $Var[\hat{\theta}] = 1/m$

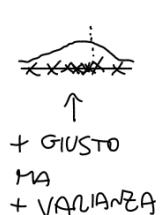


Uno stimatore si dice asintoticamente efficiente se $\lim_{n \rightarrow +\infty} Var[\hat{\theta}] = 1/m$ $\lim_{m \rightarrow +\infty} Var[\hat{\theta}] = 1/m$

Uno stimatore $\hat{\theta}^m$ corretto si dice a minima varianza se $Var[\hat{\theta}^m] \leq Var[\hat{\theta}]$

Se $\hat{\theta}$ è efficiente, allora è a minima varianza (non vale il viceversa).

Serve un indicatore globale che consideri sia il bias che la varianza. Uso come criterio l'errore quadratico medio



Proprietà

$$MSE = E[(\hat{\theta} - \theta^0)^2]$$

• Caso θ^0 scalare

$$MSE = \text{bias}[\hat{\theta}]^2 + \text{Var}[\hat{\theta}]$$

REGOLARIZZAZIONE

BIAS-VARIANCE dilemma

REGRESSIONE LINEARE

STIMA AI MINIMI QUADRATI

$$y(i) = \theta_0 + \theta_1 \varphi_1(i) + \dots + \theta_{d-1} \varphi_{d-1}(i)$$

Figura 9 descrizione dei dati tramite una relazione lineare

$$\begin{aligned}
 y(i) &= \underbrace{\theta_0 + \theta_1 \varphi_1(i) + \dots + \theta_{d-1} \varphi_{d-1}(i)}_{\substack{i\text{-esima} \\ \text{osservazione}}} + \epsilon(i) = \sum_{j=0}^{d-1} \theta_j \varphi_j(i) + \epsilon(i) \\
 &= \varphi^T(i) \theta + \epsilon(i) \quad \begin{array}{l} \cdot \quad \varphi_0 = 1 \\ \cdot \quad \varphi = [\varphi_0, \varphi_1, \dots, \varphi_{d-1}]^T \in \mathbb{R}^{d \times 1} \quad \text{VETTORE FEATURES} \\ \cdot \quad \theta = [\theta_0, \theta_1, \dots, \theta_{d-1}]^T \in \mathbb{R}^{d \times 1} \quad \text{VETTORE PARAMETRI} \end{array} \\
 &\qquad \qquad \qquad \downarrow \text{ERRORE}
 \end{aligned}$$

Figura 10 relazione tra le variabili di input e una variabile di output usando un modello lineare

FUNZIONE DI COSTO

Regressione lineare: modello lineare + metodo di stima di somma ai minimi quadrati.

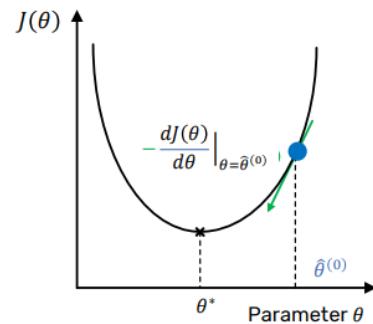
Il metodo di regressione lineare **stima i parametri θ minimizzando l'errore quadratico tra output osservati e stimati dal modello lineare.**

Poiché il modello è lineare nei parametri e la misura dell'errore è quadratico, la funzione di costo è convessa (ammette un minimo unico globale). Nel caso della regressione lineare, il minimo può anche essere trovato in forma chiusa.

GRADIENT DESCENT

È un metodo iterativo per minimizzare le funzioni differenziabili. Considero prima il caso scalare: dato un valore iniziale, la stima è

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} - \alpha \cdot \frac{\partial J(\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}^{(k)}}$$



e questa nuova stima è più vicina al valore ottimale θ^* .

Nel caso multivariabile dobbiamo sostituire la derivata con il vettore gradiente. [VEDERE SLIDE](#)

Quando si hanno più regressori è utile normalizzare i valori: calcolo la media per ogni regressore, calcolo la varianza per ogni regressore e infine sottraggo la media e divido per la deviazione standard

$$\hat{\mu}_j = \frac{1}{N} \sum_{i=1}^N \varphi_j(i) \quad \hat{\sigma}_j^2 = \frac{1}{N} \sum_{i=1}^N (x_j(i) - \mu_j)^2 \quad \varphi_j(i) = \frac{\varphi_j(i) - \hat{\mu}_j}{\sqrt{\hat{\sigma}_j^2}} \quad j = 1, \dots, d-1$$

PROPRIETA' DELLO STIMATORE AI MINIMI QUADRATI

Sia $\epsilon(i)$ una variabile casuale a media nulla, con una certa varianza λ^2 , allora lo stimatore è corretto e se i rumori sono incorrelati, allora lo stimatore è consistente.

STIMA A MASSIMA VEROVEROSIMIGLIANZA

STIMA A MASSIMA VEROVEROSIMIGLIANZA

Gli estimatori visti fino ad ora sono parametrici, ovvero stimano dei parametri; non abbiamo fai fatto assunzioni sulla distribuzione dei dati D.

Il metodo della massima verosimiglianza è una procedura di stima che, dato un modello probabilistico, stima i suoi parametri in modo tale che siano più coerenti con i dati osservati.

Supponiamo di avere a disposizione N osservazioni Y, la PDF congiunta dei dati indica la probabilità che si realizzi il vettore di dati osservato. Siccome le $y(i)$ sono i.i.d. (indipendenti e identicamente distribuite), la probabilità di osservare $y(1)$ AND ... AND $y(N)$ è il prodotto della PDF delle singole variabili.

Io conosco il valore di Y, se conoscessi anche μ e σ potrei calcolare la probabilità di aver osservato Y; siccome non li conosco, sono proprio i due parametri che voglio stimare.

Quando la PDF congiunta è vista in funzione dei parametri μ e σ , è chiamata funzione di likelihood.

La stima a massima verosimiglianza è quindi quel valore del parametro θ che massimizza la verosimiglianza $L(\theta/Y)$.

- Variabili non note** **Parametri NOTI**
- Se $f_Y(Y|\mu, \sigma^2)$ è funzione dei dati Y : **pdf multivariabile**
- Dati NOTI** **Variabili non note**
- Se $f_Y(Y|\mu, \sigma^2)$ è funzione dei parametri μ e σ^2 : **likelihood** $\mathcal{L}(\mu, \sigma^2|Y)$

Spesso, anziché massimizzare $L(\theta|Y)$, si massimizza il suo logaritmo naturale. Dato che il logaritmo è una funzione monotona crescente, $\ln L(\theta|Y)$ ha lo stesso massimo di $L(\theta|Y)$. Usare il logaritmo è efficiente dal punto di vista implementativo. L'ottimizzazione è effettuata con metodi iterativi.

Lo stimatore a massima verosimiglianza è asintoticamente corretto, consistente (più N è grande, più la stima è precisa), asintoticamente efficiente e asintoticamente normale.

STIMA A MASSIMA VEROVEROSIMIGLIANZA DI MODELLI LINEARI

Questo metodo può essere usato anche per stimare modelli lineari. Bisogna impostare un modello probabilistico alle osservazioni $y(i)$.

$$f_Y(y(1), y(2), \dots, y(N)|X, \boldsymbol{\theta}, \lambda^2) \stackrel{i.i.d.}{=} \prod_{i=1}^N f_Y(y(i)|\boldsymbol{\phi}(i), \boldsymbol{\theta}, \lambda^2) = \mathcal{L}(\boldsymbol{\theta}|Y, X, \lambda^2)$$

Figura 11 distribuzione congiunta dei dati

Dopodichè si calcola la log-verosimiglianza.

La stima a massima verosimiglianza del modello lineare è equivalente alla stima ai minimi quadrati.

REGRESSIONE LOGISTICA

IL PROBLEMA DELLA CLASSIFICAZIONE

Nel modello di regressione lineare, la variabile di risposta è quantitativa. In molte situazioni invece è qualitativa, questa tipologia di variabile assume valori in un insieme non ordinato.

Dati metrici: descrivono una quantità, è definito un ordine, è definita una distanza.

Dati categorici: descrivono categorie di appartenenza, non ha senso ordinarli, non ha senso calcolare distanze

Il processo di stima di output categorici è chiamato classificazione. Spesso però siamo più interessati a stimare le probabilità che φ (insieme di regressori) appartenga a ciascuna categoria in C . Se si vuole ottenere una classificazione, la categoria più probabile viene scelta come classe per l'osservazione φ .

PERCHE' NON USARE LA REGRESSIONE LINEARE?

In generale, non esiste un modo naturale per convertire una variabile di risposta qualitativa con più di due livelli (categorie) in una risposta quantitativa che sia adatta alla regressione lineare.

Tuttavia, se usiamo la regressione lineare, alcune delle nostre stime potrebbero essere al di fuori dell'intervallo [0;1]. Non ha senso perché non c'è nulla che satura l'uscita tra 0 e 1 (funzione logistica → sigmoide)

REGRESSIONE LOGISTICA: FORMULAZIONE DEL PROBLEMA

Vogliamo stimare la probabilità che le osservazioni φ appartenagno ad una di due classi.

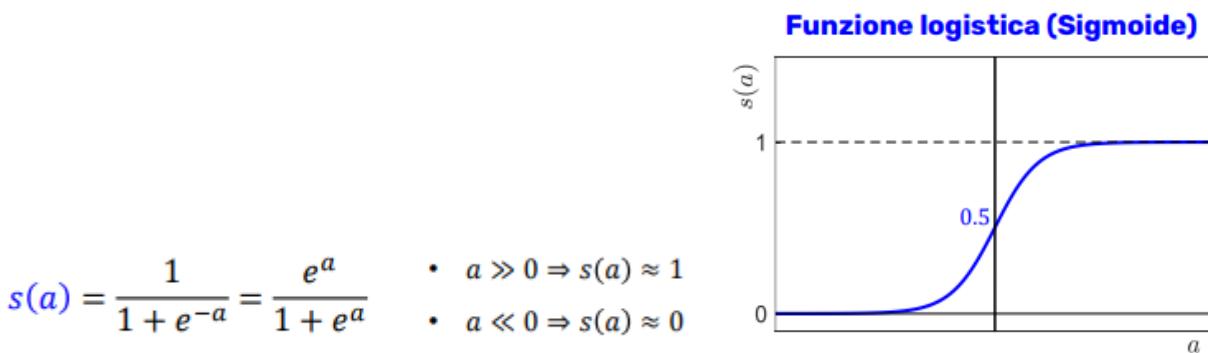


Figura 12 la formula $s(a)$ è la funzione logistica

Il modello di regressione logistica modella la probabilità che $y = 1$ tramite un modello lineare:

$$P(y = 1 | \boldsymbol{\varphi}) = s(a) = s(\boldsymbol{\varphi}^\top \boldsymbol{\theta}) = \frac{1}{1 + e^{-\boldsymbol{\varphi}^\top \boldsymbol{\theta}}}$$

L'output è interpretato come una probabilità:

- $\boldsymbol{\varphi}^\top \boldsymbol{\theta} \gg 0$ allora φ è classificato nella classe <<1>>
- $\boldsymbol{\varphi}^\top \boldsymbol{\theta} \ll 0$ allora φ è classificato nella classe <<0>>

La regressione lineare e logistica fanno parte di una categoria di modelli più generale detti Generalized Linear Model (GLM).

REGRESSIONE LOGISTICA: FUNZIONE COSTO

Vogliamo modellare i dati tramire una regressione logistica $P(y = 1 | \boldsymbol{\varphi}) = \frac{1}{1 + e^{-\boldsymbol{\varphi}^\top \boldsymbol{\theta}}} \equiv \pi$

Interpretiamo i dati come relazioni di una distribuzione di Bernoulli $y \sim \text{Bernoulli}(\pi)$.

Inizialmente, si calcola la meno-log-likelihood, successivamente si calcola il gradiente e infine si effettua un'ottimizzazione per trovare il minimo di $J(\theta)$ (costo).

Assumiamo che ci sia un solo dato:

- Caso $y = 1$
 - $J(\theta) \sim 0$ se $y = 1$ e $\pi \sim 1 \rightarrow$ decisione corretta
 - $J(\theta) \sim +\infty$ se $y = 1$ e $\pi \sim 0 \rightarrow$ decisione sbagliata
- Caso $y = 0$
 - $J(\theta) \sim 0$ se $y = 0$ e $\pi \sim 0 \rightarrow$ decisione corretta
 - $J(\theta) \sim +\infty$ se $y = 0$ e $\pi \sim 1 \rightarrow$ decisione sbagliata

RIASSUNTO

Il modello di regressione logistica viene utilizzato per la classificazione e non per la regressione. Possiamo classificare un dato in una particolare classe se la probabilità per quella classe è superiore ad una soglia (solitamente 0,5).

Il confine di classificazione che viene generato dalla regressione logistica è lineare.

FONDAMENTI DI MACHINE LEARNING

INTRODUZIONE AL MACHINE LEARNING E ALLA DATA SCIENCE

Il machine learning ha senso di essere applicato quando:

1. Esiste un pattern nei dati
2. Non possiamo descriverlo matematicamente
3. Abbiamo dati su di esso (l'unico vero vincolo)

Il machine learning mi consente di predire: se studio tre gatti, quando vedo il quarto lo riconosco.

Data science conduce solo degli studi

PROBLEMI SUPERVISIONATI E NON SUPERVISIONATI

I componenti del learning sono:

- Input: $\varphi \in R^{d \times 1}$
- Output: $y \rightarrow$ la decisione che dobbiamo prendere
- Funzione target: $f: R^{d \times 1} \rightarrow y \rightarrow$ ignota, si deve stimare
- Dati $D = \{(\varphi(1), y(1)), \dots, (\varphi(n), y(n))\}$: ogni vettore delle features è costituito da diverse informazioni utilizzate per prevedere la variabile di output

- Ipotesi scelta: $g: R^{d \times 1} \rightarrow y, g \in M \rightarrow g$ è un'approssimazione di f

M è chiamato spazio delle ipotesi. Insieme all'algoritmo di learning forma il modello di learning.

Nell'apprendimento supervisionato, la risposta corretta y è nota e bisogna prevedere y da un set di inputs.

Esempi di questa tipologia di apprendimento sono la regressione lineare e la regressione logistica; in entrambi i casi l'obiettivo è quello di stimare la funzione $f()$ usando i dati D . La funzione f viene cercata, dall'algoritmo di learning, nello spazio delle ipotesi M . Vogliamo trovare una funzione $h \in M$ che approssimi bene f , non solo sui dati D a disposizione, ma sull'intero dominio $R^{d \times 1}$ di f .

Nell'apprendimento non supervisionato anziché (input, output) abbiamo (input, ?), infatti non c'è una funzione f da apprendere e si vogliono esplorare le proprietà di φ . C'è una rappresentazione ad alto livello dell'input e gli elementi nello stesso cluster hanno proprietà simili.

Nell'apprendimento per reinforzo anziché (input, output) abbiamo (input, output, ricompensa); l'algoritmo cerca di imparare quale azione intraprendere al fine di massimizzare la ricompensa.

Nell'apprendimento supervisionato, dobbiamo definire una misura di errore o di costo. Le misure di errore puntuali $\ell(\varphi; \theta)$ sono basate su un singolo punto φ (alcuni esempi sono l'errore quadratico e l'errore binario).

Le misure di errore globali considerano tutte le N osservazioni. È importante distinguere tra errore in-sample ed errore out-of-sample:

- Errore in-sample: errore che il modello fa sugli N dati osservati a disposizione, che sono stati usati per stimarlo.
- Errore out-of-sample: errore che il modello fa sull'intero dominio di f (quindi anche dati che non ho osservato).

FEASIBILITY OF LEARNING

Non è possibile conoscere con certezza come sarà il comportamento delle funzioni f su punti che non ho osservato.

Focalizziamoci sul supervised learning, classificazione binaria

- Problema: stimare una funzione ignota f
- Risposta: impossibile, la funzione f può assumere qualsiasi valore al di fuori del dati che abbiamo a disposizione ?

Tramite la similitudine delle biglie pescate da un'urna possiamo dire che:

- \hat{p} è l'errore sample in-sample $E_{in}(h)$
- p è l'errore out-of-sample $E_{out}(h)$

Nel caso del learning di modelli, ciò che ci interessa veramente stimare è E_{out} , non E_{in} , in quanto E_{in} non è un buon indicatore della bontà del modello.

In uno scenario reale, h non è fissata a priori. L'algoritmo di learning è usato per scandagliare lo spazio delle ipotesi M , al fine di trovare la migliore h appartenente ad M che approssima bene i dati osservati. Con tante ipotesi in M , c'è un rischio maggiore di trovare una funzione g che fa bene sui dati osservati solo per fortuna. Esiste quindi un tradeoff tra approssimazione e generalizzazione, infatti si vuole un buon modello sui dati misurati e un buon modello sui dati non visti.

La quantità $E_{\text{out}}(g) - E_{\text{in}}(g)$ è chiamata errore di generalizzazione.

L'obiettivo finale è avere un piccolo E_{out} : buona approssimazione di f out-of-sample. Se lo spazio delle ipotesi M è più complesso allora si hanno migliori possibilità di approssimare f in-sample. Se lo spazio delle ipotesi M è meno complesso allora si hanno migliori possibilità di generalizzare f out-of-sample. Il caso ideale sarebbe avere uno spazio delle ipotesi M che contiene solo la funzione f : $M = \{f\}$

Esiste una terapia della generalizzazione che studia i casi in cui è probabile generalizzare. Il concetto fondamentale è che il learning è fattibile in modo probabilistico. Se si riesce ad affrontare il tradeoff approssimazione-generalizzazione, possiamo dire con alta probabilità che l'errore di generalizzazione è piccolo.

Un modo per studiare questo tradeoff è valutare i concetti di bias e varianza di un modello di learning:

- Quanto bene M può approssimare f -> Bias
- Quanto bene riusciamo a scegliere una buona h appartenente ad M , usando i dati -> Varianza

BIAS-VARIANCE TRADEOFF

Supponiamo di osservare i dati senza rumore

$$E_{\text{out}}(g^{(\mathcal{D})}) = \mathbb{E}_{\varphi} \left[\left(g^{(\mathcal{D})}(\varphi) - f(\varphi) \right)^2 \right] \quad \bar{g}(\varphi) = \mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\varphi)]$$

errore out-of-sample

ipotesi media

L'errore out-of-sample atteso è indipendente dalla particolare realizzazione dei dati utilizzati per stimare $g^{(\mathcal{D})}$.

L'ipotesi media può essere interpretata come l'ipotesi che deriva dall'usare K dataset e costruendola come

$$\bar{g}(\varphi) \approx \frac{1}{N} \sum_{k=1}^K g^{(\mathcal{D}_k)}(\varphi) \quad \text{e quindi:} \quad \mathbb{E}_{\mathcal{D}}[E_{\text{out}}(g^{(\mathcal{D})})] = \text{bias}^2 + \text{var}$$

Il termine di bias² ($(\bar{g}(\varphi) - f(\varphi))^2$) misura quanto il modello è lontano dalla funzione target f . Infatti \bar{g} apprende da un numero illimitato di datasets e quindi è limitata solo dai limiti di M .

Il termine varianza $E_D[(g^{(D)}(\varphi) - \bar{g}(\varphi))^2]$ misura quanto $g^{(D)}$ si disperde da \bar{g} .

- Set di modelli molto piccolo: visto che esiste una sola ipotesi, e quindi \bar{g} e $g^{(D)}$ saranno uguali e $\text{var} = 0$, allora il bias dipenderà da quanto bene questa singola ipotesi si avvicina al target f e quindi ci aspettiamo un bias grande
- Set di modelli molto grande: diversi set di dati porteranno a diverse ipotesi $g^{(D)}$ e il bias ~ 0 poiché in media $g^{(D)}$ è vicina ad f . La varianza allora è grande.

La complessità del modello deve seguire il numero di dati, non la complessità della funzione di target-

LEARNING CURVES

Le learning curves sono uno strumento grafico per capire se il modello di learning soffre di problemi di bias o varianza. L'idea è di rappresentare, al variare del numero di dati N , l'errore out-of-sample atteso e l'errore in-sample atteso. In pratica, le curve vengono calcolate usando un solo dataset.

Il bias può essere presente quando l'errore atteso è piuttosto elevato e E_{in} è simile a E_{out} ; quando è presente il bias, è improbabile che ottenere più dati aiuti. La varianza può essere rappresentata quando c'è tanto divario tra E_{in} e E_{out} ; quando è presente la varianza, è probabile che ottenere più dati sia d'aiuto.

OVERFITTING

Una causa di overfitting è il rumore stocastico sui dati in uscita y . Si parla di overfitting quando diminuire E_{in} porta ad un aumento di E_{out} . L'overfitting porta a una cattiva generalizzazione; un modello può mostrare una cattiva generalizzazione anche se non overfittato.

Supponiamo che vi sia un rumore stocastico η con media zero e varianza σ^2 che affligge le misure. L'errore stocastico σ^2 non può essere portato a zero e il rumore stocastico causa overfitting.

REGOLARIZZAZIONE

La regolarizzazione è la prima linea di difesa contro l'overfitting. I modelli più complessi sono più inclini all'overfitting, perché più potenti. I modelli semplici mostrano meno varianza a causa della loro espressività limitata. Tuttavia, se ci atteniamo solo a modelli semplici, potremmo non ottenere un'approssimazione soddisfacente della funzione target f .

L'idea è quella che, oltre a minimizzare il funzionale di costo, va minimizzata anche la complessità del modello.

Al posto di E_{in} , minimizziamo un errore aumentato E_{aug} (questo conduce ad un modello migliore):

$$E_{\text{aug}}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^N \left(y(i) - h(\boldsymbol{\varphi}(i); \boldsymbol{\theta}) \right)^2 + \lambda_{\text{reg}}$$

Quanto male il modello fitta i dati

λ_{reg} pesa l'importanza di minimizzare $E_{\text{in}} \equiv J(\theta)$ rispetto a minimizzare $\Omega(\theta)$.

Se regolarizzo troppo, imparerà la funzione più semplice possibile.

$$E_{\text{out}}(\boldsymbol{\theta}) = E_{\text{in}}(\boldsymbol{\theta}) + \tilde{\Omega}(\boldsymbol{\theta}) \quad E_{\text{aug}}(\boldsymbol{\theta}) = E_{\text{in}}(\boldsymbol{\theta}) + \lambda \Omega(\boldsymbol{\theta})$$

Somma di due contributi

L'errore E_{aug} è migliore rispetto a E_{in} come proxy per E_{out} .

L'ideale sarebbe avere un'espressione E_{out} da minimizzare, così sarebbe possibile minimizzare direttamente l'errore out-of-sample invece di quello in-sample.

La regolarizzazione aiuta nello stimare la quantità $\Omega(\theta)$, che, sommata ad E_{in} , fornisce E_{aug} , il quale è stima di E_{out} .

Esistono diversi tipi di regolarizzazione:

- Regolarizzazione L2 chiamata penalità Ridge (la penalità ridge tende a ridurre tutti i coefficienti a un valore inferiore)
- Regolarizzazione L1 chiamata penalità Lasso (la penalità lasso tende a portare più coefficienti esattamente a zero).

Gli effetti della regolarizzazione possono essere osservati nei termini di bias e varianza:

- La regolarizzazione aumenta di poco il bias al fine di ridurre considerevolmente la varianza
- Porta ad avere ipotesi più smooth, riducendo il rischio di overfitting
- λ deve essere scelto in modo specifico. Solitamente si usa una procedura come la validazione o la cross-validation

VALIDAZIONE, CROSS-VALIDAZIONE E FORMULE DI COMPLESSITÀ OTTIMA

L'errore out-of-sample può essere visto come $E_{\text{out}}(\theta) = E_{\text{in}}(\theta) + \text{penalità per la complessità del modello}$: la validazione stima $E_{\text{out}}(\theta)$ e la regolarizzazione stima la penalità per la complessità del modello.

L'idea della validazione è di stimare E_{out} , utilizzando un dataset diverso rispetto a quello usato per la stima del modello. Regolarizzazione e validazione sono tecniche che possono essere usate insieme, la prima aiuta a stimare un modello che può generalizzare meglio e la seconda fornisce una stima dell'errore out-of-sample del modello stimato. Entrambe sono fondamentali anche dell'identificazione dei sistemi dinamici.

L'obiettivo del set di validazione è quello di stimare le performance out-of-sample. La procedura che si segue è:

1. Rimuovo un subset di dati dai dati formali
2. Stimo il modello sulla parte di dati rimanente
3. Valuto le performance del modello sul subset di dati che ho rimosso al punto 1.
4. Ri-alleno il modello su tutti i dati

$$\underbrace{N_{val} \text{ dati: } \textcolor{red}{\textbf{validazione}}}_{\mathcal{D}_{val}} \quad \underbrace{N - N_{val} \text{ dati: } \textcolor{blue}{\textbf{training}}}_{\mathcal{D}_{train}}$$

N_{val} piccolo: stima di E_{out} non buona. N_{val} grande: possibilità di imparare un modello non buono.

Le recedute di validazione possono essere utilizzate per due scopi: valutare le performance del modello e stimare e scegliere il modello migliore da un insieme di diversi modelli.

Se abbiamo N_m set di modelli tra cui imparare un modello:

- Stimo $g(\bar{m})$ usando D_{train} per ogni set di modelli
- Valuto $g(\bar{m})$ usando D_{val}
- Seleziono il modello $m = m^*$ con l'errore $E(\bar{m})$ più basso

Problema: se uso D_{val} tante volte, allora il dataset di validazione D_{val} non fornisce più una buona stima dell'errore out-of-sample E_{out} .

Intuizione: usare D_{val} per compiere delle scelte fa sì che tali scelte siano diendenti dai particolari valori dei dati contenuti in D_{val} .

Soluzione: c'è bisogno di un terzo dataset. Il dataset di test sul quale si calcolerà l'errore di test E_{test} .

Finora abbiamo tre stime dell'errore E_{out} :

- Training set: totalmente contaminato
- Validation set: un po' contaminato
- Test set: totalmente <>pulito>>

La divisione del dataset in tre parti è fattibile se i dati a disposizione sono molti. In teoria vorremmo che:

$$E_{\text{out}}(g) \approx E_{\text{out}}(g^-) \approx \overbrace{E_{\text{val}}(g^-)}^{\begin{array}{l} (N_{\text{val}} \text{ piccolo}) \\ (N_{\text{val}} \text{ grande}) \end{array}}$$

N_{val} grande: in questo caso, il valore di E_{val} calcolato usando g^- sarebbe simile al valore di E_{out} ottenuto da g^- .

L'obiettivo di E_{val} è proprio di stimare E_{out} .

N_{val} piccolo: in questo caso, il valore di E_{out} ottenuto da g^- sarebbe simile al valore di E_{out} ottenuto da g . Questo è il valore che mi interessa ma che non posso calcolare direttamente.

La cross-validation permette di <>avere N_{val} sia grande che piccolo>>.

$$E_{\text{cv}} = \frac{1}{N} \sum_{i=1}^N \ell(i)$$

Errore di cross validazione

Stimo N modelli usando $N - 1$ dati, e li valido usando N stime dell'errore di validazione.

La cross-validation con $N_{\text{val}} = 1$ ha gli svantaggi che è computazionalmente costosa e la stima dell'errore E_{cv} ha una varianza elevata. È possibile riservare più punti per la validazione suddividendo il training set in <>folds>>.

La K-fold cross-validation richiede N/N_{val} sessioni di train, ognuna con $N - N_{\text{val}}$ dati. Un buon compromesso è usare $K = 10$: 10 – fold cross validation: $N_{\text{val}} = N/10$ (attenzione a non ridurre troppo il training set).

Le formule di complessità ottima permettono di stimare l'errore out-of-sample E_{out} utilizzando solo il dataset train. Per questo motivo, si usano quando ho troppi pochi dati. L'idea è simile alla regolarizzazione: modificare la funzione di costo dell'errore in-sample E_{in} , aggiungendo un termine additivo che penalizza la complessità del modello.

Supponiamo di avere un modello con d parametri. Indichiamo la stima dei parametri, ottenuta con N dati, con $\hat{\theta}_N$. La stima ottenuta minimizzando la funzione costo $J(\theta; d)$ dove esplicitiamo la dipendenza del costo dal numero di parametri d .

$$\text{AIC}(d) = 2 \cdot \frac{d}{N} + \ln[J(\hat{\theta}_N; d)] \quad \text{MDL}(d) = \ln[N] \cdot \frac{d}{N} + \ln[J(\hat{\theta}_N; d)]$$

Akaike Information Criterion

Minimum Description Length

La differenza consiste solo nel termine 2 e $\ln[N]$. Se $\ln[N] > 2$, la formula MDL suggerisce di usare modelli più parsimoniosi.

Se disponiamo di molti dati, il modo migliore per valutare le performance e selezionare il modello è dividere il dataset in 3 parti (training, validazione e set); altrimenti usiamo la cross-validation. Se i dati sono davvero pochi, possiamo utilizzare formule per la scelta della complessità del modello ottimale che utilizzano solo il training set (AIC e MDL).

FONDAMENTI DI STIMA BAYESIANA

PROBABILITÀ CONGIUNTE, CONDIZIONATE, MARGINALI

Supponiamo di avere due variabili casuali discrete e binarie a e b . La distribuzione di probabilità congiunta $P(a,b)$ è la probabilità che sia a che b assumono un valore specifico

$$\sum_{a=0}^1 \sum_{b=0}^1 p(a,b) = 1 \quad P(a,b) = P(b,a)$$

La distribuzione di probabilità marginare è la distribuzione di probabilità di un sottoinsieme di variabili casuali. Nel nostro esempio, siccome abbiamo due v.c. a e b , avremo due marginali ($P(a)$ e $P(b)$).

Nel caso di v.c. discrete, la distribuzione marginale è ottenuta sommando rispetto alle variabili che non sono di interesse. Nel caso di v.c. continue, si deve integrare anziché sommare.

La distribuzione condizionata indica come la probabilità di ridistribuisce dato che si restringe la popolazione ad un particolare sottoinsieme.

La probabilità che accada sia A che B è la probabilità che si verifichi B per la probabilità che si verifichi A dato che B si è verificato. $P(A,B) = P(A)P(B)$ solo se $P(A|B) = P(A)$: questo vuol dire che A e B solo eventi indipendenti, ovvero il verificarsi di B non modifica le probabilità di verificarsi di A.

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

Sappiamo che $P(A,B) = P(B,A)$. Inoltre $P(B,A) = P(B/A)P(A)$ e quindi $P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)}$

Il teorema di Bayes permette di ridistribuire la probabilità: prima conoscevamo $P(A)$, adesso conosco $P(A|B)$. La probabilità di A è cambiata in seguito all'informazione portata da B.

INTRODUZIONE ALLA STIMA BAYESIANA

Finora abbiamo considerato il vettore di parametri incognito θ come una variabile deterministica. Spesso però, ancora prima di collezionare i dati, abbiamo delle informazioni sui possibili valori che potrebbe assumere θ

Ha quindi senso considerare θ come una variabile casuale: posso specificare una distribuzione di probabilità θ per descrivere i valori che io credo possa assumere. Assegno maggior probabilità ai valori che io credo siamo più probabili che θ possa assumere e minor probabilità ai valori che io credo non possa assumere.

$f_\theta(\theta)$ ha dominio $[0,1]$ poiché θ , modellando una probabilità, deve stare tra 0 e 1.

$\theta = 0,5$ sarà il valore che io suppongo sia più possibile, $\theta \sim 0$ o $\theta \sim 1$ saranno poco probabili. Data $P(\theta)$, abbiamo già una stima del valore di θ ancora prima di aver osservato i dati (stima a priori).

Con l'osservazione dei dati ci si aspetta che:

- La stima puntuale di θ cambi
- L'incertezza sulla stima decresca

Abbiamo quindi due elementi che portano informazione:

1. La distribuzione a priori $f_\theta(\theta)$ sui possibili valori di θ
2. L'informazione che portano i dati sui possibili valori di θ (likelihood)

Quello che ci interessa è sapere quanto può valere θ dato che ho osservato i dati.

Usando il teorema di Bayes possiamo unire i due elementi di informazione.

$$f_{\theta|Y}(\theta|Y) = \frac{\text{LIKELIHOOD} \quad \text{PRIOR}}{\text{POSTERIOR}} = \frac{f_{Y|\theta}(Y|\theta) \cdot f_\theta(\theta)}{f_Y(Y) \quad \text{MARGINAL LIKELIHOOD}}$$

$f_{\theta|Y}(\theta|Y)$ è una distribuzione a posteriori di possibili valori di θ . Le probabilità di questi valori sono riallocate dall'aver osservati i dati Y .

In generale non posso dire nulla su $f_{\theta|Y}(\theta|Y)$. Un altro problema è che $f_Y(Y)$, nel caso di dati intesi come v.c. continue è un integrale che potremmo non sapere come risolvere. In questo caso si usano tecniche numeriche note come Markov Chain Monte Carlo. Un caso fortunato avviene se $f_\theta(\theta)$ è gaussiana e anche $f_{Y|\theta}(Y|\theta)$ è gaussiana. Allora, anche $f_{\theta|Y}(\theta|Y)$ è gaussiana.

Quando la posterior $f_{\theta|Y}(\theta|Y)$ ha la stessa forma della prior $f_\theta(\theta)$ allora la likelihood e la prior si dicono congiunte. Un modo per calcolare la posterior è quello di discretizzare il range di valori del parametro θ tramite una griglia di valori: in questo modo valuto $f_\theta(\theta)$ e $f_{Y|\theta}(Y|\theta)$ sono in quei valori di θ all'interno della griglia. Questo metodo va bene se θ consiste di un paio di parametri.

STIMA OTTIMA

Supponiamo di avere la posterior: spesso però ci serve un valore solo di θ , puntuale. In questi casi usiamo la stima MAP, il valore atteso a posteriori o altre quantità come la mediana.

Consideriamo θ scalare. Vorremmo che la variabile casuale $\hat{\theta}$ fosse vicina alla variabile casuale θ . Per quantificare questa distanza, usiamo il concetto di Mean Squared Error (MSE).

Lo stima ore ottimo di Bayes è quella funzione $T^{\text{opt}}()$ tale che:

$$\mathbb{E}[(T^{\text{opt}}(\mathcal{D}) - \theta)^2] < \mathbb{E}[(T(\mathcal{D}) - \theta)^2], \quad \forall T(\mathcal{D})$$

Cioè che minimizza il MSE.

Lo stimatore che minimizza il MSE è il valore atteso condizionato. Nel caso in cui θ sia un vettore di parametri, il calcolo del MSE si modifica come segue:

$$\text{MSE} \equiv \text{tr} \left\{ \mathbb{E}_{d \times 1} \left[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T \right] \right\} = \mathbb{E}_{1 \times d} \left[(\hat{\theta} - \theta)^T (\hat{\theta} - \theta) \right] = \mathbb{E}_{d \times 1} \left[\|(\hat{\theta} - \theta)\|_2^2 \right]$$

Supponiamo ora di avere un dato interpretato come realizzazione di una variabile casuale Gaussiana $y \sim N(0, \lambda_{yy}^2)$, e che anche il parametro ignoto sia Gaussiana. La loro pdf congiunta è ancora Gaussiana.

Si dimostra che la posterior è ancora Gaussiana con il valore atteso e la varianza.

Stima ottenuta dallo stimatore ottimo Bayesiano nel caso Gaussiano: $\hat{\theta}_{\text{opt}} = \mathbb{E}[\theta | y = y(1)] = \frac{\lambda_{\theta y}}{\lambda_{yy}^2} \cdot y(1)$

STIMA OTTIMA LINEARE

Non è sempre detto che y e θ siano congiuntamente Gaussiane. Vogliamo quindi trovare uno stimatore che non faccia ipotesi sulla ddp congiunta di y e θ . Supponiamo y e θ due variabili casuali scalari, vogliamo stimare θ tramite uno stimatore lineare $\hat{\theta}^{\text{lin}} = \alpha * y + \beta$

Per trovare α e β , minimizziamo la funzione costo data dal $\text{MSE} \equiv J(\alpha, \beta) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}[(\alpha * y + \beta - \theta)^2]$,
calcoliamo il gradiente e poniamolo uguale a zero. Otteniamo: $\alpha = \frac{\lambda_{\theta y}}{\lambda_{yy}^2}$ e $\beta = 0$

Lo stimatore lineare ottimo è dato da $\hat{\theta}_{\text{opt}}^{\text{lin}} = \frac{\lambda_{\theta y}}{\lambda_{yy}^2} * y + \beta = \frac{\lambda_{\theta y}}{\lambda_{yy}^2} * y$ e coincide con lo stimatore ottimo di Bayes per il caso gaussiano; anche la varianza è uguale al caso gaussiano.

Lo stimatore ottimo lineare non fa nessuna ipotesi su che tipo di distribuzione hanno y e θ , assieme solo che siano v.c. con una certa media e una certa varianza. Potrebbe dunque esserci uno stimatore migliore rispetto a quello lineare ottimo. Se però y e θ sono congiuntamente gaussiani, allora non esiste nessuno stimatore migliore di quello lineare ottimo.

- Y e θ scalari e valore atteso non nullo: $\hat{\theta}_{\text{opt}}^{\text{lin}} = \mu_{\theta} + \frac{\lambda_{\theta Y}}{\lambda_{yy}^2} \cdot (y - \mu_y)$ $\text{Var}[\hat{\theta}_{\text{opt}}^{\text{lin}} - \theta] = \lambda_{\theta\theta}^2 - \frac{\lambda_{\theta Y}^2}{\lambda_{yy}^2}$
- Y e θ vettoriali: $\hat{\theta}_{\text{opt}}^{\text{lin}} = \mu_{\theta} + \Lambda_{\theta Y} \cdot \Lambda_{YY}^{-1} \cdot (Y - \mu_Y)$ $\text{Var}[\hat{\theta}_{\text{opt}}^{\text{lin}} - \theta] = \Lambda_{\theta\theta} - \Lambda_{\theta Y} \cdot \Lambda_{YY}^{-1} \cdot \Lambda_{Y\theta}$

Le formule ammettono una forma ricorsiva e sono alla base del filtro di Kalman: un algoritmo che ha l'obiettivo di stimare lo stato $x(t)$ di un sistema dinamico

- Lo stato $x(t)$ e l'uscita $y(t)$ del sistema dinamico lineare sono visti come v.c. (perché affette da un disturbo)
- Si vuole stimare lo stato $x(t)$, visto come l'incognita θ , sulla base dello stato stimato al tempo precedente (stima a priori) e sui dati che man mano arrivato dalle misure dei sensori $y(t)$ (dati osservati).

PROCESSI STOCASTICI

INTRODUZIONE ALLA STIMA DI MODELLI DINAMICI

Due tipologie di problemi: analisi e modellistica di serie temporali e di sistemi ingresso/uscita

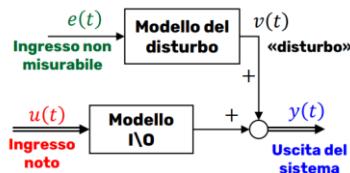
Una serie temporale è un insieme di dati D indicizzati nel tempo (y_2 temporalmente viene dopo y_1). Indichiamo ogni dato con $y(t)$.

Modelleremo $y(t)$ come l'uscita di un sistema dinamico con ingresso remoto non misurabile $e(t)$.



I sistemi ingresso/uscita processano un segnale di input $u(t)$ per generare un segnale di uscita $y(t)$.

Modelleremo $y(t)$ come il contributo di una componente esogena nota $u(t)$ ed una componente di disturbo $v(t)$ ignota.



I modelli che considereremo saranno modelli di sistemi dinamici lineari tempo invarianti e discreti.

Il termine disturbo è utilizzato per modellare differenti fenomeni: rumore di misura, disturbi di processo, effetto di segnali esogeni di input non misurabili, effetti di linearizzazioni del sistema-

Quindi $v(t)$ modella tutto ciò che il modello lineare I/O non riesce a spiegare. La cosa difficile è separare l'effetto che $u(t)$ ha su $y(t)$ rispetto a quello che $e(t)$ ha su $y(t)$ (perché io misuro solo u e y).

Vogliamo risolvere due problemi:

1. Predizione di uscite a istanti futuri $t + k$ in base alle informazioni attualmente a disposizione al tempo t . Predizione ad un passo: $\hat{y}(t + 1 | t)$
2. Identificazione (obiettivo finale, stima) dei modelli descritti, in modo da poter catturare le relazioni tra gli ingressi e l'uscita del sistema che genera i dati.

Lavoreremo con segnali a tempo discreto e assumeremo che le uscite $y(t)$ siano affette dal disturbo $v(t)$, che può essere visto come rumore che sporca la vera misura ($y(0)$) calcolata subito dopo il blocco del modello I/O dell'uscita.

Nel caso di sistemi statici, per gestire questa incertezza sulla misura dei dati, avevamo interpretato i dati come v.c.; nel caso di sistemi dinamici però i dati sono campionati da un segnale che evolve nel tempo: non abbiamo più osservazioni di v.c. singole, ma osserviamo una successione di v.c. nel tempo -> processi stocastici.

Seguiremo tre fasi per risolvere il problema:

1. Definizione delle classi di modelli M di sistemi dinamici: ci concentreremo su modelli di sistemi dinamici lineari, espressi da funzioni di trasferimento razionali fratte. I parametri ignoti sono i coefficienti dei polinomi
2. Predizione: data una particolare classe di modello, quale è il predittore ottimo?
3. Identificazione: come stimo il valore dei parametri?

PROCESSI STOCASTICI

Un processo stocastico $v(t,s)$ a tempo discreto è una successione infinita di v.c., definite a partire dallo stesso esperimento casuale s e ordinate secondo un indice temporale t .

Fissato un esito $s = \bar{s}$, si ottiene una realizzazione, ovvero una serie di valori deterministici (tutte le v.c. hanno assunto un valore). Fissato un istante temporale $t = \bar{t}$, si ottiene la v.c. al tempo \bar{t} . Fissati $s = \bar{s}$ e $t = \bar{t}$, si ottiene un numero $v(\bar{t}, \bar{s})$.

Un processo stocastico è come una funzione che restituisce funzioni nel tempo (tutte le v.c. assumono un valore e quindi si genera un segnale); è completamente caratterizzato dal punto di vista probabilistico se, per ogni n-upla di v.c., è nota la distribuzione di probabilità congiunta di queste variabili (ogni v.c. a tempi diversi ha ddp diverse).

I processi stocastici sono utili se vogliamo analizzare fenomeni che non possiamo o non vogliamo descrivere deterministicamente.

Dato un processo stocastico si definiscono:

- Valore atteso (momento del primo ordine): è una funzione che rappresenta il valore atteso della v.c. ad ogni istante di tempo t. E' una media dei valori che tutte le realizzazioni del processo assumono in quel momento. $m_v(t) \equiv \mathbb{E}_s[v(t,s)]$
- Funzione di autocorrelazione (momento del secondo ordine): permette di capire i valori che il processo assume ad un istante t_1 rispetto a quelli che assume ad un istante t_2 -> correlazione a due istanti diversi del medesimo processo. $R_{vv}(t_1, t_2) \equiv \mathbb{E}_s[v(t_1, s)v(t_2, s)]$
Se $R_{vv}(t_1, t_2) > 0$ -> $v(t_1)$ e $v(t_2)$ hanno lo stesso segno, in media
Se $R_{vv}(t_1, t_2) < 0$ -> $v(t_1)$ e $v(t_2)$ hanno segno diverso, in media
- Funzione di autocovarianza: è la covarianza tra $v(t_1, s)$ e $v(t_2, s)$. Devo togliere il valore atteso ad ogni v.c. $\gamma_{vv}(t_1, t_2) \equiv \mathbb{E}_s[(v(t_1, s) - m_v(t_1)) \cdot (v(t_2, s) - m_v(t_2))]$

Consideriamo due processi stocastici $v(t,s)$ e $x(t,s)$: possiamo definire una finzione di corss-correlazione e una di cross-covarianza, che cnsideri l'interazione tra i due proessi stocastici (cross: tra me e altro, auto: tra me e me):

Funzione di cross-correlazione	Funzione di cross-covarianza
$R_{vx}(t_1, t_2) \equiv \mathbb{E}_s[v(t_1, s) \cdot x(t_2, s)] = R_{xv}(t_2, t_1)$	$\gamma_{vx}(t_1, t_2) \equiv \mathbb{E}_s[(v(t_1, s) - m_v(t_1)) \cdot (x(t_2, s) - m_x(t_2))] = \gamma_{xv}(t_2, t_1)$

Due processi si dicono incorrelati se la funzione di cross-covarianza è uguale a 0

PROCESSI STOCASTICI STAZIONARI (PSS)

Un processo stocastico si dice stazionario in senso forte se $\forall n \in N$, scelti t_n istanti di tempo, le caratteristiche probabilistiche della n-upla $v(t_1), \dots, v(t_n)$ sono uguali a quelle della n-upla $v(t_1 + \tau), \dots, v(t_n + \tau)$. -> impone che la distribuzione congiunta non cambi mai

Un processo stocastico si dice stazionario in senso debole se $m_v(t) = m$ (per ogni t) e se $\gamma_{vv}(t_1, t_2) = \gamma_{vv}(t_3, t_4)$ nel caso in cui $|t_4 - t_3| = |t_2 - t_1| = \tau$.

Se un processo è stazionario in senso forte allora lo è anche in senso debole.

Se volessimo, potremmo decomporre un processo stazionario in diverse componenti: trend + stagionalità (andamento oscillatorio periodico) + processo stocastico stazionario

Due pss si dicono equivalenti se hanno lo stesso valore atteso m e la stessa funzione di autocovarianza $\gamma(\tau)$.

Dalla definizione di autocovarianza $\gamma_{vv}(\tau) = \mathbb{E}_s[(v(t, s) - m) \cdot (v(t + \tau, s) - m)]$ si deducono le proprietà:

1. $\gamma_{vv}(0) = \mathbb{E}[v(t) - m]^2 \geq 0$ -> varianza del processo ($\tau = 0$, non varia nel tempo come la media)
2. $|\text{autocovarianza}| \leq \gamma_{vv}(0)$ funzione limitata
3. $\gamma_{vv}(\tau) = \gamma_{vv}(-\tau)$ -> funzione pari

Un pss è detto rumore bianco se $E[e(t)] = 0$, $\gamma_{ee}(0) = \lambda^2$ e se l'autocovarianza è zero per ogni $t, \tau \neq 0$. Siccome non vi è correlazione tra il valore ad un istante t ed un valore all'istante $t + \tau$, il rumore bianco è un processo stocastico le cui realizzazioni variano in modo impredicibile da un istante all'altro.

MOMENTI TEMPORALI ED ERGODICITA'

La media temporale su orizzonte finito è quella in cui considero un numero finito di istanti temporali

$$\langle v(t,s) \rangle_N \equiv \frac{1}{N} \sum_{t=0}^{N-1} v(t,s)$$

Media temporale

$$\langle v(t,s) \rangle \equiv \lim_{N \rightarrow +\infty} \langle v(t,s) \rangle_N$$

Autocorrelazione temporale

$$\langle v(t,s) \cdot v(t+\tau,s) \rangle \equiv \lim_{N \rightarrow +\infty} \sum_{t=0}^{N-1} v(t,s) \cdot v(t+\tau,s)$$

La quantità $\langle v(t,s) \rangle_N$ è una v.c. perché dipende all'esito s , mentre $\langle v(t,s) \rangle$ è un limite di v.c.

Teo: se $v(t,s)$ è un pss e $|E_s[v(t,s)]| < +\infty$ (se la media esiste ed è finita), allora il limite $\langle v(t,s) \rangle$ converge quasi certamente. Le conseguenze del teorema sono che:

1. $E_s[\langle v(t,s) \rangle] = E_s[v(t,s)] = m$ (per stimare m dai dati devo avere più realizzazioni del processo stocastico, con le v.c. statiche non è un problema, ma nel caso di pss spesso ho solo una relaizzazione finita della serie temporale perché ho un solo segnale d'ingresso e uno di uscita).
2. $E_s[\langle v(t,s) * v(t+\tau,s) \rangle] = R_{vv}(\tau)$.

Ovvero si dimostra che $\langle v(t,s) \rangle$ e $\langle v(t,s) * v(t+\tau,s) \rangle$ sono stimatori corretti del valore atteso e della funzione di autocorrelazione di $v(t,s)$.

Il processo stocastico $v(t,s)$ è detto ergodico se $v(t,s)$ è stazionario e se per $N \rightarrow +\infty$ i momenti temporali convergono quasi certamente ai rispettivi momenti di insieme. Il processo è detto ergodico nella media se

$$\lim_{N \rightarrow +\infty} \langle v(t,s) \rangle_N = m \quad \text{q. c.}$$

Teo: sia $v(t,s)$ un pss in senso debole. Se la varianza esiste finita e la funzione di autocovarianza tende a zero, il processo è ergodico nella media.

Teo: sia $v(t,s)$ stazionario e gaussiano. Se la varianza esiste finita e la funzione di autocovarianza tende a zero, il processo è ergodico.

Con l'ergodicità riesco a fare stime dei processi stocastici anche se ho una sola realizzazione. Se un processo stocastico è ergodico, allora ogni singola realizzazione è rappresentativa di tutte le possibili realizzazioni. Per essere rappresentativa, la realizzazione deve dimenticare i valori iniziali ed esplorare tutto il dominio del processo.

Però, per sapere se un processo è ergodico, devo conoscere $\gamma_{vv}(\tau)$, però, a meno che non abbia già informazioni su $\gamma_{vv}(\tau)$, devo stimarla dai dati. Per stimarla dai dati però il processo deve essere ergodico. Spesso quindi non posso fare altro che ipotizzare l'ergodicità.

TRASFORMATA Z E TRASFORMATA DI FOURIER

La trasformata Zeta bilatera di un segnale deterministico $g(t)$ è definita come ($g(t)$ è una funzione reale di variabile intera e $G(z)$ è una funzione complessa di variabile complessa).

$$Z[g(t)] = G(z) \equiv \sum_{t=-\infty}^{+\infty} g(t) \cdot z^{-t}, \quad z \in \mathbb{C}$$

È lineare, z è un operatore di anticipo unitario e z^{-1} è un operatore di ritardo unitario.

La convoluzione tra due segnali discreti è $y(t) = \sum_{i=-\infty}^{+\infty} g(i)u(t-i) = \sum_{i=-\infty}^{+\infty} g(t-i)u(i)$ e la trasformata Z della convoluzione è il prodotto delle trasformate Z: $Y(z) = G(z)U(z)$.

Sia $u(t)$ un segnale a tempo discreto, deterministico, assolutamente sommabile; allora si definisce trasformata di fourier a tempo discreto (DTFT):

$$\mathcal{F}[u(t)] \equiv \sum_{t=-\infty}^{+\infty} u(t) \cdot e^{-j\omega t}$$

La trasformata di fourier a tempo discreto è la restrizione di $U(z)$ alla circonferenza di raggio unitario.

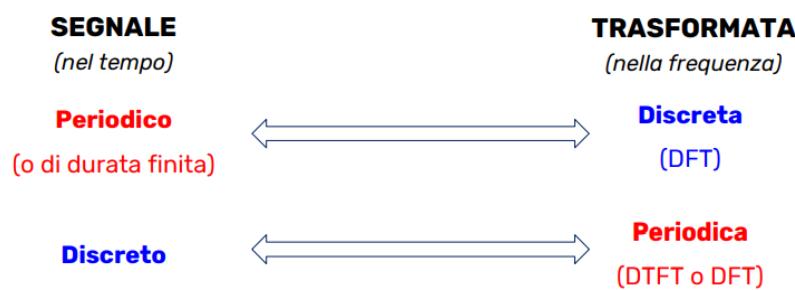
Semprerebbe che la trasformata di fourier contenga meno informazioni rispetto alla trasformata Z. In realtà, è possibile ricostruire completamente $u(t)$ partendo da $U(e^{j\omega})$. La DTFT è inoltre periodica e il complesso coniugato di $X(e^{j\omega})$ si trova cambiando il segno all'angolo ω .

Consideriamo un segnale discreto $u(t)$ di durata finita, definiamo la trasformata di fourier discreta (DFT):

$$\tilde{U}(k) \equiv \sum_{t=0}^{N-1} u(t) \cdot e^{-j \cdot t \cdot k \phi} \quad \begin{aligned} &\bullet \quad \phi = \frac{2\pi}{N} \\ &\bullet \quad k = 0, \dots, N-1 \end{aligned}$$

È una funzione complessa della variabile intera k che considera un segnale discreto di durata finita.

La DFT può essere vista come un campionamento in frequenza della DTFT. Esiste una DFT inversa tale che è possibile ricostruire $u(t)$ partendo da $\tilde{U}(k)$. Quindi la DFT non perde alcuna informazione. La risoluzione della DFT, chiamata anche frequency bin è data da bin = frequenza di campionamento / N. Dato che la DFT è simmetrica, solo $N/2$ dati portano informazione. La DFT è sia discreta che periodica:



DENSITÀ SPETTRALE DI POTENZA

Sia $v(t,s)$ un pss (il valore atteso di un pss sposta la media ma non influenza se la caratterizzazione del processo è smooth o altro, molto spesso infatti studio pss a media nulla). Sia il valore atteso che la funzione di autocovarianza sono caratterizzazioni nel tempo; è però possibile caratterizzare un pss nella frequenza. L'evoluzione delle realizzazioni di un pss è prettamente caratterizzata dalla funzione di autocovarianza. Per questo motivo, spesso si studiano pss depurati dalla loro media \rightarrow anziché $\gamma_{vv}(\tau)$, considero le sue trasformate.

Def: dati un pss, si definisce densità spettrale di potenza $\Gamma_{vv}(\omega)$. Come la DTFT di $\gamma_{vv}(\tau)$:

$$\Gamma_{vv}(\omega) \equiv \mathcal{F}[\gamma_{vv}(\tau)] = \sum_{\tau=-\infty}^{+\infty} \gamma_{vv}(\tau) \cdot e^{-j\omega\tau} . \text{ Data } \Phi_{vv}(z), \text{ si ha che } \Gamma_{vv}(\omega) = \Phi_{vv}(e^{j\omega}),$$

Conoscere $\Gamma_{vv}(\omega)$ o $\Phi_{vv}(z)$, è equivalente: posso risalire a $\gamma_{vv}(\tau)$ con l'antitrasformata. Affinché $\Gamma_{vv}(\omega)$ converga, $\gamma_{vv}(\tau)$ deve tendere a zero in modo sufficientemente rapido (i pss infatti non devono avere una memoria lunga, il passato molto lontano non influisce sul presente).

Proprietà di $\Gamma_{vv}(\omega)$: reale, positiva, pari e periodica di periodo 2π

A tempo discreto, la più grande pulsazione osservabile è quella di una cosinusoida che cambia valore ad ogni istante di tempo t (ogni istante temporale campiona un dato). Tra l'istante t e l'istante $t + 1$, trascorre un tempo di campionamento T_s . Il più piccolo periodo osservabile è quindi $T = 2T_s = 2/f_s$. La pulsazione più grande osservabile corrisponde quindi a: $\omega = \frac{2\pi}{T} = \frac{\pi}{T_s} = \pi * f_s$ (Teorema del campionamento).

ω è una pulsazione normalizzata rispetto alla frequenza di campionamento f_s . L'interpretazione è che π corrisponde a $f_s/2$.

È possibile risalire a $\gamma_{vv}(\tau)$ tramite l'antitrasformata

$$\gamma_{vv}(\tau) = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \Gamma_{vv}(\omega) \cdot e^{j\omega\tau} d\omega$$

Inoltre, è possibile espiare la varianza del processo stazionario come l'area sorpresa alla densità spettrale di potenza (dalla formula di prima tolgo 2π e non è $\gamma_{vv}(\tau)$, ma $\gamma_{vv}(0)$).

Sia $e(t) \sim WN(0, \lambda^2)$. Sappiamo che nel tempo è un segnale impredicibile (non ho frequenze che contribuiscono più delle altre) dato che: $\gamma_{ee}(\tau) = \begin{cases} 0 & \text{se } \tau \neq 0 \\ \lambda^2 & \text{se } \tau = 0 \end{cases} \Rightarrow \Gamma_{ee}(\omega) = \sum_{\tau=-\infty}^{+\infty} \gamma_{ee}(\tau) * e^{-j\omega\tau} = \lambda^2 e \Phi(z) = \lambda^2$

La densità spettrale di potenza del rumore bianco è quindi costante. Non vi sono frequenze dominanti: tutte contribuiscono in modo uguale alla variabilità del segnale.

Dati due processi stazionari, definiamo la densità di potenza cross-spettrale come:

$$\Gamma_{vx}(\omega) \equiv \mathcal{F}[\gamma_{vx}(\tau)]$$

$$\Phi_{vx}(z) \equiv Z[\gamma_{vx}(\tau)]$$

STIMA SPETTRALE (prima forma di modellizzazione dei dati)

Ipotizzo che $v(t)$ sia ergodico, che $E[v(t)] = m_v = 0$ e che di N dati disponibili ho solo una realizzazione del processo $v(t)$.

Possiamo stimare il valore atteso m_v di un pss ergodico $v(t)$ come $\widehat{m}_v = \frac{1}{N} \sum_{t=0}^{N-1} v(t)$ (media temporale).

Supponiamo ora che $v(t)$ sia un pss a media nulla (altrimenti avrei dovuto sottrarla). Dato che $\gamma_{vv}(\tau) = E_s[v(t)v(t + \tau)]$, l'autocovarianza sarà: $\widehat{\gamma}_{vv}(\tau) = \frac{1}{N - |\tau|} \sum_{t=0}^{N-|\tau|-1} v(t)v(t + |\tau|), \quad |\tau| < N$

Per $\tau = 0$, stimo la varianza del processo. Uso $|\tau|$ perché la stima è analoga sia per $\tau > 0$ che per $\tau < 0$, data la simmetria di $\gamma_{vv}(\tau)$. Più τ è grande, meno dati posso usare per la stima.

Se $v(t)$ è gaussiano, $\widehat{\gamma}_{vv}(\tau)$ è lo stimatore a massima verosimiglianza. Inoltre, lo stimatore è corretto e consistente. Però, per $\tau \sim N$, si ha che $\text{var}[\widehat{\gamma}_{vv}(\tau)]$ è grande perché ci sono pochi addendi e quindi possiamo pensare ad uno stimatore alternativo (distorto ma asintoticamente corretto). Per τ fissato è consistente:

$$\widehat{\gamma}'_{vv}(\tau) = \frac{1}{N} \sum_{t=0}^{N-|\tau|-1} v(t)v(t + |\tau|), \quad |\tau| < N$$

Valore atteso di $\widehat{\gamma}'_{vv}(\tau)$, ovvero $E[\widehat{\gamma}'_{vv}(\tau)] = \frac{N-|\tau|}{N} \gamma_{vv}(\tau)$. Supponiamo di voler calcolare la stima per: $\tau = N-3, \tau = N-2, \tau = N-1$. Per $\tau \sim N$, il valore atteso dello stimaore viene schiacciato verso il basso, cosa che non succedeva con lo stimatore corretto. Lo stimatore non corretto peggiora il bias ma riduce la varianza.

Si definisce periodogramma il seguente stimatore della densità spettrale $I_N(\omega) \equiv \sum_{\tau=-(N-1)}^{N-1} \widehat{\gamma}'_{vv}(\tau) \cdot e^{-j\omega\tau}$ di potenza ($I_N(\omega)$ è una funzione reale, continua e 2π -periodica):

$I_N(\omega)$ è proporzionale al modulo del quadrato della DTFT: $I_N(\omega) = \frac{1}{N} |V(e^{j\omega})|^2$. Per segnali di durata finita, la DFT è un campionamento della DTFT. Per cui posso calcolare: $\mathcal{I}_N(k) = \frac{1}{N} |V(e^{j \cdot k \cdot 2\pi/N})|^2$.

Lo stimatore $I_N(\omega)$ non è corretto, ma è asintoticamente corretto (non lo sarebbe stato neanche se avessi usato $\widehat{\gamma}_{vv}(\tau)$ al posto di $\widehat{\gamma}'_{vv}(\tau)$). Siccome la varianza dello stimatore non decresce al crescere di N , allora lo stimatore non è neanche consistente. Per $N \rightarrow +\infty$, $I_N(\omega_1)$ e $I_N(\omega_2)$ tendono a diventare incorrelati, quindi il periodogramma è una funzione poco continua.

Un metodo per migliorare la stima è quello di regolarizzare la stima facendo la media di diversi periodogrammi.

Metodo di Bartlett: N dati a disposizione. Dividvo questi dati in $K = N/M$ parti (M è la lunghezza di ogni porzione di dati), calcolo il periodogramma $I_{M,K}^{[i]}(\omega)$ per ciascuna parte i , infine faccio una media dei periodogrammi e ottengo una stima: $\bar{I}_{M,K}(\omega) = \frac{1}{K} \sum_{i=1}^K I_{M,K}^{[i]}(\omega)$

Se $\gamma_{vv}(\tau) \rightarrow$ in modo sufficientemente rapido, i K periodogrammi sono circa indipendenti. Il bias [$\bar{I}_{M,K}(\omega)$] è maggiore rispetto a quello di $I_N(\omega)$, quindi ho una maggiore perdita di risoluzione in frequenza. Se so che $\Gamma_{vv}(\omega)$ ha picchi molto stretti, devo usare M grande in modo da avere abbastanza risoluzione in frequenza.

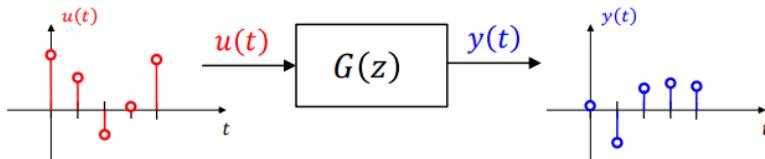
SISTEMI DINAMICI LTI DISCRETI DETERMINISTICI

Un sistema dinamico può essere rappresentato in spazio di stato oppure in forma ingresso/uscita; l'obiettivo è quello di stimare la FDT.

Def: un sistema dinamico è LTI se la sua uscita $y(t)$ può essere espressa tramite la convoluzione dell'input $u(t)$ e della risposta all'impulso $g(t)$ del sistema (se $g(t) = 0$ per $t < 0$, questa ipotesi di causalità può solo influenzare l'uscita ad istanti $s \geq t$):

$$y(t) = \sum_{i=-\infty}^{\infty} g(t-i)u(i) = \sum_{j=0}^{\infty} g(j)u(t-j)$$

La FDT $G(z)$ descrive la relazione tra il segnale di ingresso $u(t)$ e il segnale di uscita $y(t)$, quando $x(0) = 0$.



Gli zeri sono le radici del numeratore e i poli sono le radici del denominatore.

Def: un sistema dinamico LTI a tempo discreto si dice asintoticamente stabile se $|poli| < 1$.

Conseguenza dell'asintotica stabilità: la risposta all'impulso tende esponenzialmente a 0 per $t \rightarrow +\infty$:

$\lim_{t \rightarrow +\infty} g(t) = 0$. Il guadagno del sistema è la sommatoria dei valori della risposta all'impulso:

$\sum_{t=0}^{+\infty} g(t) = G(1)$. Se applico $u(t) = \text{scalino}(t)$ e il sistema è asintoticamente stabile, allora $\lim_{t \rightarrow +\infty} y(t) = \mu$.

Consideriamo un'onda sinusoidale campionata con periodo di campionamento T_s . I valori campionati sono:

$$s(t) = A \cdot \sin(2\pi f_0 \cdot T_s \cdot t + \varphi)$$

Ampiezza
 Frequenza
 Fase

La frequenza di Nyquist è: $f_{Nyq} = \frac{f_s}{2} = \frac{1}{2T_s}$. Per poter campionare correttamente, bisogna avere $f_s = 1/T_s$ sufficientemente alta. La frequenza sinusoidale deve rispettare il criterio di Nyquist.

Sia $G(z)$ la FDT di un sistema asintoticamente stabile. Consideriamo un input sinusoidale del tipo $u(t) = A * \sin(2\pi f_s t * f + \varphi)$. Il segnale di output sarà: $y(t) = \hat{y}(t) + \bar{A} * \sin(2\pi f_s t * f + \bar{\varphi})$ tale che:

Transitorio

Aampiezza uscita

Fase dell'uscita

$$\lim_{t \rightarrow \infty} \tilde{y}(t) = 0$$

$$\bar{A} = A \cdot |G(e^{j \cdot 2\pi T_s \cdot f})|$$

$$\bar{\varphi} = \varphi + \angle G(e^{j \cdot 2\pi T_s \cdot f})$$

Valutando $G(z)$ in $z = e^{jwT_s}$ si ottiene la risposta in frequenza del sistema, con modulo $|G(e^{jw})|$ e fase $\angle G(e^{jw})$.

SISTEMI DINAMICI LTI DISCRETI STOCASTICI

Supponiamo che $u(t)$ sia un processo stazionario in senso debole, con media m_u e autocovarianza $\gamma_{vv}(\tau)$, e $G(z)$ una FDT razionale fratta, asintoticamente stabile al guadagno μ .

Valore atteso: $E[y(t)] = \sum_{i=0}^{+\infty} g(i)E[u(t-i)] = G(1) * m_u = \mu * m_u$ (non dipende da t).

Autocovarianza $(m_u = 0)$: $y(t) = \sum_{i=0}^{+\infty} g(i)u(t-i) \rightarrow y(t+\tau) = \sum_{i=0}^{+\infty} g(i)u(t-i+\tau) \rightarrow u(t)y(t+\tau) = \sum_{i=0}^{+\infty} u(t) * g(i)u(t-i+\tau) \rightarrow \gamma_{uy}(t, t+\tau) = \sum_{i=0}^{+\infty} g(i)\gamma_{uu}(t, t-i+\tau)$

$$\underline{\gamma_{uy}(\tau)} = \sum_{i=0}^{+\infty} g(i)\gamma_{uu}(\tau - i) \quad \Gamma_{uy}(\omega) = G(e^{j\omega})\Gamma_{uu}(\omega)$$

Inoltre: $y(t)y(t+\tau) = \sum_{i=0}^{+\infty} y(t) * g(i)u(t-i+\tau) \rightarrow \gamma_{yy}(t, t+\tau) = \sum_{i=0}^{+\infty} g(i)\gamma_{yu}(t, t-i+\tau)$

$$\underline{\gamma_{yy}(\tau)} = \sum_{i=0}^{+\infty} g(i)\gamma_{yu}(\tau - i) \quad \Gamma_{yy}(\omega) = G(e^{j\omega})\Gamma_{yu}(\omega)$$

Teo: sia $u(t)$ un pss che alimenta un sistema dinamico asintoticamente stabile, allora anche $y(t)$ è un pss.

Nella pratica, $u(t)$ viene applicato da $t = 0$ e non da $t = -\infty$, per cui $y(t)$ sarà stazionario dopo un transitorio. Questa è una condizione necessaria e sufficiente. A regime, per ogni condizione iniziale, $y(t)$ è un pss se $u(t)$ è un pss e se $G(z)$ è asintoticamente stabile.

Teo:	$\Gamma_{yy}(\omega) = G(e^{j\omega}) ^2 \cdot \Gamma_{uu}(\omega)$	$\Phi_{yy}(z) = G(z)G(z^{-1}) \cdot \Phi_{uu}(z)$
------	----------------------------------------------------------------------	---------------------------------------------------

Dimo:

Possiamo interpretare un pss $y(t)$ come l'uscita di un sistema dinamico $G(z)$ asintoticamente stabile alimentato da rumore bianco, tale che $\Gamma_{yy}(\omega) = |G(e^{j\omega})|^2$

Ne segue che, data $G(z)$ asintoticamente stabile, è possibile esprimere un qualunque pss $y(t)$ come combinazione lineare di infiniti campioni di rumore bianco. Se conosco $\Gamma_{yy}(\omega)$ e se riesco a trovare $G(z)$ asintoticamente stabile e causale tale che $\Gamma_{yy}(\omega) = |G(e^{j\omega})|^2$, posso anche simulare diverse realizzazioni del processo $y(t)$.

Quindi $e(t)$ dello schema iniziale sarà proprio il rumore bianco e l'obiettivo è stimare $H(z)$ (modello del disturbo) e $G(z)$ (modello I/O).

La depolarizzazione consiste nel rimuovere il valore atteso m ad un pss $v(t)$, semplifica il calcolo dell'autocovarianza.

FAMIGLIE DI MODELLI STOCASTICI

FAMIGLIE DI MODELLI A SPETTRO RAZIONALE

I processi stocastici che si ottengono filtrando un rumore bianco tramite un filtro asintoticamente stabile $H(z) = C(z)/A(z)$ sono detti processi a spettro razionale, dove $C(z)$ e $A(z)$ sono polinomi a coefficienti reali nella variabile z .

SERIE TEMPORALI: MODELLI MA, AR, ARMA

MODELLI MA: MOVING AVERAGE

Def: un processo stocastico, generato a partire dal rumore bianco $e(t) \sim WN(\mu, \lambda^2)$, è detto di tipo MA(n_c) se:

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

L'uscita di un modello MA(n_c) è combinazione lineare degli ultimi $n_c + 1$ valori del rumore bianco in ingresso.

Possiamo scrivere il processo come:

$$\begin{aligned} y(t) &= c_0 e(t) + c_1 e(t-1) + c_2 e(t-2) + \dots + c_{n_c} e(t-n_c) \\ &= c_0 e(t) + c_1 z^{-1} e(t) + c_2 z^{-2} e(t) + \dots + c_{n_c} z^{-n_c} e(t) \\ &= \underbrace{[c_0 + c_1 z^{-1} + c_2 z^{-2} + \dots + c_{n_c} z^{-n_c}]}_{C(z)} \cdot e(t) \end{aligned}$$

$y(t) = C(z)e(t)$

Valore atteso

$$m_y = \mathbb{E}[y(t)] = \mathbb{E}[c_0 e(t) + c_1 e(t-1) + \dots + c_{n_c} e(t-n_c)]$$

$$(\text{se } e(t) \sim \text{WN}(0, \lambda^2), \text{ allora } = c_0 \mathbb{E}[e(t)] + c_1 \mathbb{E}[e(t-1)] + \dots + c_{n_c} \mathbb{E}[e(t-n_c)])$$

$\mathbb{E}[y(t)] = 0$:

$$= c_0 \mu + c_1 \mu + \dots + c_{n_c} \mu = \mu \cdot \sum_{i=0}^{n_c} c_i \quad \boxed{\text{Non dipende dal tempo } t}$$

Supponiamo $\mathbb{E}[y(t)] = 0$:

$$\begin{aligned} \gamma_{yy}(0) &= \mathbb{E}[(y(t) - m_y)^2] = \mathbb{E}[(y(t))^2] = \mathbb{E}\left[\left(c_0 e(t) + c_1 e(t-1) + \dots + c_{n_c} e(t-n_c)\right)^2\right] \\ &= \mathbb{E}\left[c_0^2 e(t)^2 + c_1^2 e(t-1)^2 + \dots + c_{n_c}^2 e(t-n_c)^2 + \right. \\ &\quad \left. + 2c_0 c_1 e(t)e(t-1) + \dots + 2c_{n_c-1} c_{n_c} e(t-n_c+1)e(t-n_c)\right] = c_0^2 \mathbb{E}[e(t)^2] + \dots + c_{n_c} \mathbb{E}[e(t-n_c)^2] \\ &= c_0^2 \gamma_{ee}(0) + c_1^2 \gamma_{ee}(0) + \dots + c_{n_c}^2 \gamma_{ee}(0) = \lambda^2 \cdot \sum_{i=0}^{n_c} c_i^2 \quad \boxed{\text{Non dipende dal tempo } t} \end{aligned}$$

$$\begin{aligned} \gamma_{yy}(1) &= \mathbb{E}[(y(t) - m_y)(y(t-1) - m_y)] = \mathbb{E}[y(t)y(t-1)] = \\ &= \mathbb{E}\left[\left(c_0 e(t) + c_1 e(t-1) + \dots + c_{n_c} e(t-n_c)\right) \cdot \left(c_0 e(t-1) + c_1 e(t-2) + \dots + c_{n_c} e(t-n_c-1)\right)\right] \\ &= c_0 c_1 \mathbb{E}[e(t-1)^2] + c_1 c_2 \mathbb{E}[e(t-2)^2] + \dots + c_{n_c-1} c_n \mathbb{E}[e(t-n_c-1)^2] \\ &= \lambda^2 \cdot (c_0 c_1 + c_1 c_2 + \dots + c_{n_c-1} c_n) \end{aligned}$$

$$\begin{aligned} \gamma(2) &= \lambda^2 \cdot (c_0 c_2 + c_1 c_3 + \dots + c_{n_c-2} c_{n_c}) & \gamma(n_c) &= \lambda^2 \cdot (c_0 c_{n_c}) \\ \gamma(\tau) &= 0 \text{ se } \tau > n_c & \text{Un processo MA}(n_c) \text{ dipende solo dagli } n_c \text{ valori precedenti al tempo corrente} \end{aligned}$$

Un modo per capire se una serie temporale può essere modellata tramite un MA(n_c) è quello di guardare se la funzione di autocovarianza va a zero dopo n_c lags.

MODELLI AR: AUTOREGRESSIVE

Def: un processo stocastico $y(t)$ è detto di tipo AR(n_a) se:

$$y(t) = a_1 y(t-1) + a_2 y(t-2) + \dots + a_{n_a} y(t-n_a) + e(t) = \sum_{i=1}^{n_a} a_i y(t-i) + e(t)$$

L'uscita di un modello AR(n_a) è combinazione lineare degli ultimi n_a valori del processo stesso e del rumore bianco in ingresso.

$$\begin{aligned} y(t) &= a_1 y(t-1) + a_2 y(t-2) + \dots + a_{n_a} y(t-n_a) + e(t) \\ &= a_1 z^{-1} y(t) + a_2 z^{-2} y(t) + \dots + a_{n_a} z^{-n_a} y(t) \\ y(t) \underbrace{[1 - a_1 z^{-1} + a_2 z^{-2} + \dots + a_{n_a} z^{-n_a}]}_{A(z)} &= e(t) \quad \Rightarrow \quad y(t) = \frac{1}{A(z)} e(t) \end{aligned}$$

Un processo AR(n_a) è stazionario sse $1/A(z)$ è asintoticamente stabile.

Valore atteso:

$$m_y = \mathbb{E}[y(t)] = \mathbb{E}[a_1 y(t-1) + a_2 y(t-2) + \dots + a_{n_a} y(t-n_a) + e(t)]$$

$$= a_1 \mathbb{E}[y(t-1)] + \dots + a_{n_a} \mathbb{E}[y(t-n_a)] + \mathbb{E}[e(t)]$$

$$= a_1 \mathbb{E}[y(t)] + \dots + a_{n_a} \mathbb{E}[y(t)] + \mu \quad \Rightarrow \quad (1 - a_1 - \dots - a_{n_a}) \mathbb{E}[y(t)] = \mu$$

$$\Rightarrow \boxed{\mathbb{E}[y(t)] = \frac{\mu}{1 - a_1 - \dots - a_{n_a}}}$$

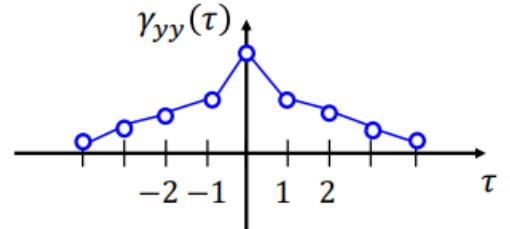
Consideriamo AR(1) del tipo $y(t) = a_1y(t-1) + e(t)$. Supponiamo che il processo sia asintoticamente stabile e a media nulla:

$$\begin{aligned}\gamma_{yy}(0) &= \mathbb{E}[y(t)^2] = \mathbb{E}[(a_1y(t-1) + e(t))^2] = \mathbb{E}[a_1^2y(t-1)^2 + e(t)^2 + 2a_1y(t-1)e(t)] \\ &= a_1^2\mathbb{E}[y(t-1)^2] + \mathbb{E}[e(t)^2] + 2\mathbb{E}[y(t-1)e(t)] \quad \text{y}(t-1) \text{ dipende solo da } e(t-1), e(t-2), \dots \\ &= a_1^2\gamma_{yy}(0) + \lambda^2 + 0 \quad \Rightarrow \quad \gamma_{yy}(0)[1 - a_1^2] = \lambda^2 \quad \Rightarrow \quad \boxed{\gamma_{yy}(0) = \frac{\lambda^2}{1 - a_1^2}} \\ \gamma_{yy}(1) &= \mathbb{E}[y(t)y(t-1)] = \mathbb{E}[(a_1y(t-1) + e(t)) \cdot y(t-1)] = \mathbb{E}[a_1y(t-1)^2 + y(t-1)e(t)] \\ &= a_1\mathbb{E}[y(t-1)^2] + \mathbb{E}[y(t-1)e(t)] = a_1\gamma_{yy}(0) \quad \Rightarrow \quad \boxed{\gamma_{yy}(1) = a_1\gamma_{yy}(0)} \\ \gamma_{yy}(2) &= \mathbb{E}[y(t)y(t-2)] = \mathbb{E}[(a_1y(t-1) + e(t)) \cdot y(t-2)] \\ &= \mathbb{E}[a_1y(t-1)y(t-2) + y(t-2)e(t)] = a_1\mathbb{E}[y(t-1)y(t-2)] + \mathbb{E}[y(t-2)e(t)] \\ &= a_1\gamma_{yy}(1) \quad \Rightarrow \quad \boxed{\gamma_{yy}(2) = a_1\gamma_{yy}(1)}\end{aligned}$$

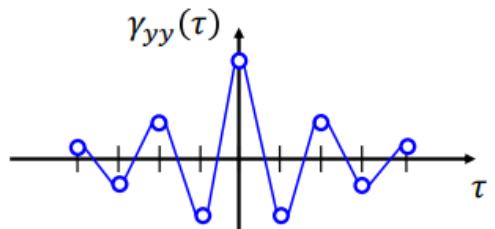
Generalizzando: $\begin{cases} \gamma_{yy}(0) = \frac{\lambda^2}{1 - a_1^2} \text{ se } \tau = 0 \\ \gamma_{yy}(\tau) = a_1 * \gamma(\tau-1) \text{ se } \tau > 0 \end{cases}$

Equazioni di Yule-Walke per un AR(1)

Il processo $\bar{y}(t) = a_1\bar{y}(t-1) + e(t)$ con $0 < a_1 < 1$ ha funzione di autocovarianza $\gamma_{yy}(\tau) > 0$ e sarà decrescente, senza raggiungere mai lo zero. Le realizzazioni del processo «variano lentamente» e sono «smooth», poiché le variabili casuali sono correlate positivamente fra loro. In media, le realizzazioni «non cambiano segno» da un istante al successivo. Le componenti a bassa frequenza dominano nella densità spettrale di potenza.



Il processo $\bar{y}(t) = a_1\bar{y}(t-1) + e(t)$ con $-1 < a_1 < 0$ ha funzione di autocovarianza che cambia segno ad ogni τ , in modo alternato. Le realizzazioni del processo «variano velocemente» e sono «nervose», poiché le variabili casuali sono correlate negativamente fra loro. In media, le realizzazioni «cambiano segno» da un istante al successivo. Le componenti ad alta frequenza dominano nella densità spettrale di potenza.



MODELLO ARMA: AUTOREGRESSIVE MOVING AVERAGE

Def: $y(t) = a_1y(t-1) + a_2y(t-2) + \dots + a_n y(t-n_a) + \dots + e(t) + c_1e(t-1) + c_2e(t-2) + \dots + c_n e(t-n_c)$ Parte AR(n_a) Parte MA(n_c)

Notiamo che ARMA(0, n_c) = MA(n_c) e ARMA(n_a , 0) = AR(n_a).

$$y(t)[1 - a_1z^{-1} - a_2z^{-2} - \dots - a_n z^{-n_a}] = [1 + c_1z^{-1} + c_2z^{-2} + \dots + c_n z^{-n_c}]e(t)$$

$$y(t) = \frac{1 + c_1z^{-1} + c_2z^{-2} + \dots + c_n z^{-n_c}}{1 - a_1z^{-1} - a_2z^{-2} - \dots - a_n z^{-n_a}} e(t) \quad \Rightarrow \quad \boxed{y(t) = \frac{C(z)}{A(z)} e(t)}$$

$Y(t)$ è stazionario sse $C(z)/A(z)$ è asintoticamente stabile.

Teo: dato un pss ARMA(n_a, n_c), esso può essere scritto come un MA(inf).

Supponiamo di avere un AR(1) del tipo $y(t) = a_1y(t-1) + e(t)$. $y(t) = 1 / (1 - az^{-1})$ e $e(t) = \sum_{i=0}^{inf} (az^{-1})^i e(t-i) = \sum_{i=0}^{inf} a^i * e(t-1) \rightarrow$ MA(inf).

SISTEMI INGRESSO/USCITA: MODELLI ARX E ARMAX

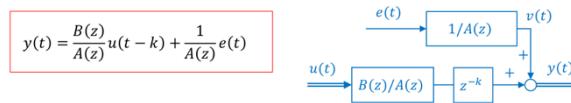
MODELLI ARX: AR WITH EXOGENOUS INPUT

Def ARX (n_a, n_b, k): $y(t) = a_1y(t-1) + a_2y(t-2) + \dots + a_{n_a}y(t-n_a) + e(t)$ Parte AR(n_a)
 $+ b_0u(t-k) + b_1u(t-k-1) + \dots + b_{n_b}u(t-k-n_b)$ Parte X(n_b)

Il termine k è il ritardo puro tra ingresso $u(t)$ e uscita $y(t)$.

$$y(t)[1 - a_1z^{-1} - \dots - a_{n_a}z^{-n_a}] = [b_0z^{-k} + b_1z^{-k-1} + \dots + b_{n_b}z^{-k-n_b}]u(t) + e(t)$$

$$y(t) = \frac{b_0 + b_1z^{-1} + \dots + b_{n_b}z^{-n_b}}{1 - a_1z^{-1} - \dots - a_{n_a}z^{-n_a}} u(t-k) + \frac{1}{1 - a_1z^{-1} - \dots - a_{n_a}z^{-n_a}} e(t)$$



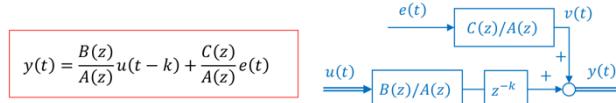
MODELLI ARMAX: ARMA WITH EXOGENOUS INPUT

Def ARMAX(n_a, n_c, n_b, k): $y(t) = a_1y(t-1) + a_2y(t-2) + \dots + a_{n_a}y(t-n_a) +$ Parte AR(n_a)
 $+ b_0u(t-k) + b_1u(t-k-1) + \dots + b_{n_b}u(t-k-n_b)$ Parte X(n_b)
 $+ c_1e(t-1) + c_2e(t-2) + \dots + c_{n_c}e(t-n_c)$ Parte MA(n_c)

$$y(t)[1 - a_1z^{-1} - \dots - a_{n_a}z^{-n_a}] = [b_0z^{-k} + b_1z^{-k-1} + \dots + b_{n_b}z^{-k-n_b}]u(t) +$$

$$+ [1 + c_1z^{-1} + c_2z^{-2} + \dots + c_{n_c}z^{-n_c}]e(t)$$

$$y(t) = \frac{b_0 + b_1z^{-1} + \dots + b_{n_b}z^{-n_b}}{1 - a_1z^{-1} - \dots - a_{n_a}z^{-n_a}} u(t-k) + \frac{1 + c_1z^{-1} + \dots + c_{n_c}z^{-n_c}}{1 - a_1z^{-1} - \dots - a_{n_a}z^{-n_a}} e(t)$$



SISTEMI INGRESSO/USCITA: MODELLI FIR, OE, BJ

MODELLI FIR: FINITE IMPULSE RESPONSE

$$y(t) = b_0u(t-k) + b_1u(t-k-1) + \dots + b_{n_b}u(t-k-n_b) = \sum_{i=0}^{n_b} b_i \cdot u(t-k-i) + e(t)$$

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

L'uscita di un modello FIR(n_b) dipende solo da valori passati dell'ingresso $u(t)$ e del rumore bianco $e(t)$.

MODELLI OE: OUTPUT ERROR

$$y(t) = \frac{B(z)}{F(z)} u(t-k) + e(t)$$

Simile al modello ARX ma, a differenza di quest'ultimo, suppone che il rumore entri solo dopo che l'uscita $y(t)$ è stata generata.

MODELLI BJ: BOX-JENKINS

$$y(t) = \frac{B(z)}{F(z)} u(t - k) + \frac{C(z)}{D(z)} e(t)$$

indipendente.

Questi modelli hanno polinomi diversi al denominatore, per cui parte esogena e parte stocastica sono parametrizzate in modo

PREDIZIONE

PREDIZIONE, FILTRAGGIO E SMOOTHING

Siano $y(\cdot)$ e $x(\cdot)$ due processi stocastici stazionari con $y(\cdot)$ osservabile. Un problema interessante è quello di ottenere una stima di $x(t)$ nei seguenti casi:

- $y(t) = x(t)$. Misuro il processo $x(t)$ che mi interessa stimare
- $y(t) = x(t) + e(t)$ con $e(t)$ rumore bianco.

Vogliamo ottenere una stima $x(t|t_{\text{info}})$, basata sulla conoscenza di $y(s)$ per valori di:

- $s < t_{\text{info}}$ -> predizione
 - $\hat{x}(t | t - k)$ -> predizione a k passi. Obiettivo: stimare il valore di $x(t)$ a istanti futuri.
- $s = t_{\text{info}}$ -> filtraggio
 - $\hat{x}(t | t)$ -> filtraggio, ha senso solo se $x(t) \neq y(t)$. Obiettivo: ottenere una stima di $x(t)$ all'istante corrente
- $s > t_{\text{info}}$ -> smoothing
 - $\hat{x}(t | t + k)$ -> smoothing, ha senso solo se $x(t) \neq y(t)$. Obiettivo: ottenere una stima di $x(t)$ all'istante passato.

Noi studieremo il problema della predizione per $x(t) = y(t)$, ovvero ci interesserà trovare una stima $\hat{y}(t | t - k)$ di $y(t)$ al tempo t , a e di a disposizione i dati fino al tempo $t - k$. Dato che il predittore $\hat{y}(t | t - k)$ si basa su valori passati di $y(t)$, sarà anch'esso un processo stocastico. L'errore di predizione $\varepsilon_k(t)$ è un processo stocastico definito come:

$$\varepsilon_k(t) = y(t) - \hat{y}(t | t - k)$$

Vogliamo preveditori lineari ottimi con errore di predizione a MSE minimo. Un modello ed buono se è il grado di predire bene i dati.

Dato che lavoriamo con pss, le scritture $\hat{y}(t | t - k)$ e $\hat{y}(t + k | t)$ sono equivalenti, nel senso che la forma del predittore ottimo è la stessa.

SCOMPOSIZIONE DI WOLD

Un processo stazionario $y(t)$ si dice completamente predicibile se esistono coefficienti a_i con $i = 1, 2, \dots$ tali che:

$$y(t) = \sum_{i=1}^{+\infty} a_i y(t-i)$$

Se conosco i coefficienti posso prevedere senza errore i valori futuri di

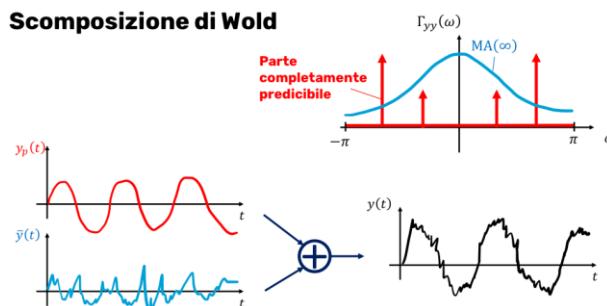
$y(\cdot)$ senza errori, a partire da valori passati. Tali processi sono l'opposto del rumore bianco, che è completamente impredicibile.

Un processo stazionario $y(t)$ è completamente predicibile sse

$$\Gamma_{yy}(\omega) = \sum_i \alpha_i \delta(\omega - \omega_i)$$

La densità spettrale di potenza di un processo completamente predicibile è una combinazione lineare di delta di Dirac. Il rumore bianco, in contrasto, ha una densità spettrale di potenza costante.

Ogni processo stocastico stazionario $y(t)$ può essere scritto come: $y(t) = \bar{y}(t) + y_p(t)$ con $y_p(t)$ processo stocastico stazionario completamente predicibile, $\bar{y}(t)$ parte puramente stocastica e le due y sono incorrelate.



Nella pratica questo risultato di fornisce una linea guida per stimare modelli di serie temporali: faccio una stima spettrale per riconoscere eventuali righe, stimo le componenti sinusoidali $y_p(t)$ ottenendo $\hat{y}_p(t)$, ottengo la componente puramente stocastica come $\bar{y}(t) = y(t) - \hat{y}_p(t)$, risolvo il problema della predizione stocastica $\bar{y}(t)$ ottenendo $\hat{y}(t | t-k)$ e infine ottengo la predizione finale come $\hat{y}(t | t-k) = \hat{y}_p(t) - \hat{y}(t | t-k)$.

Anche eventuali componenti di non stazionarietà come stagionalità o trend devono essere stimate e rimosse dai dati per ottenere solo la parte stocastica del processo. Un trend può anche essere un valore costante, tale valore può essere visto come la componente a frequenza zero, che viene rimossa dal processo con la procedura vista precedentemente.

Estrarre le righe dal periodogramma non è facile perché:

1. Gli stimatori basati sul periodogramma non sono molto buoni
2. Risonanze nella densità spettrale di potenza potrebbero essere dovute non solo alla presenza di delta di Dirac stimate male, ma anche a poli.

FILTRO PASSA – TUTTO E FORMA CANONICA

Il filtro passa-tutto è un filtro di ordine 1 definito come

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

Il filtro passa-tutto non modifica il modulo delle frequenze nella densità spettrale di potenza dell'ingresso.

Il processo $y(t)$ in uscita al passa-tutto è spettralmente equivalente al processo $v(t)$ in ingresso al passa tutto.

I due processi $y(t)$ e $v(t)$ non sono identici poiché il passa-tutto introduce uno sfasamento.

Abbiamo detto che vogliamo risolvere il problema della predizione per processi a spettro razionale. Il problema della fattorizzazione spettrale consiste nel trovare tutte le coppie $\{H(z), \lambda^2\}$ tali che $\phi_{zz}(z) = \lambda^2 * H(z)H(z^{-1})$. Per processi a spettro razionale, esistono infiniti fattori spettrali. Ai fini della predizione ottima, ci servirà un fattore spettrale particolare, detto canonico.

Teorema della fattorizzazione spettrale: dato un processo stocastico a spettro razionale, esiste un solo fattore spettrale $\{\tilde{H}(z), \tilde{\lambda}^2\}$, detto fattore spettrale canonico, dove $\tilde{H}(z) = C(z)/A(z)$, tale che: $C(z)$ e $A(z)$ hanno lo stesso grado, sono coprimi, monici e hanno radici interne al cerchio unitario.

PREDITTORE OTTIMO

Predizione: stimare il dato al tempo t avendo a disposizione dati fino al tempo $t - k$. Indichiamo il predittore come $\hat{y}(t | t - k)$ o $\hat{y}(t + k | t)$.

Informazioni disponibili: dati, vecchie predizioni e modella della parte stocastica del processo $C(z)/A(z)$.

Ipotesi di lavoro: supponiamo $y(t)$ un pss puramente stocastico, depurato da componenti predicibili e il modello $C(z)/A(z)$ in forma canonica.

Un predittore lineare è ottimo se

1. $E[\varepsilon_k(t)] = E[y(t) - \hat{y}(t | t - k)] \rightarrow$ il valore atteso dell'errore di predizione è nullo
2. $E[\hat{y}(t | t - k) * \varepsilon_k(t)] = 0 \rightarrow$ il predittore e l'errore di predizione sono incorrelati
3. $var[\varepsilon_k(t)^2]$ minima.

Avendo definito l'errore di predizione come $\varepsilon_k(t) = y(t) - \hat{y}(t | t - k)$, possiamo scomporre il processo $y(t)$ come $y(t) = \hat{y}(t | t - k) + \varepsilon_k(t)$ dove:

- $\hat{y}(t | t - k)$ è la parte predicibile al tempo $t - k$
- $\varepsilon_k(t)$ è la parte impredicibile al tempo $t - k$

PREDITTORE OTTIMO PER PROCESSI MA

Consideriamo un processo MA(n_c) in forma canonica:

$$y(t) = \underbrace{e(t)}_{\text{Parte impredicibile al tempo } t} + \underbrace{c_1 e(t-1) + c_2 e(t-2) + \cdots + c_{n_c} e(t-n_c)}_{\text{Parte predicibile al tempo } t-1}, \quad e(t) \sim WN(0, \lambda^2)$$

Un possibile predittore potrebbe quindi essere dato dalla parte predicibile al tempo $t - k$:

$$\hat{y}(t|t-1) = c_1 e(t-1) + c_2 e(t-2) + \cdots + c_{n_c} e(t-n_c)$$

- $\hat{y}(t|t-1)$ è corretto, infatti $E[y(t)] = E[\hat{y}(t|t-1)] = 0$
- $\hat{y}(t|t-1)$ dipende dal WN fino al tempo $t-1$
- $E[\hat{y}(t|t-1) * \varepsilon_1(t)] = 0$, infatti $\varepsilon_1(t) = y(t) - \hat{y}(t|t-1) = e(t)$ e quindi $E[\hat{y}(t|t-1) * \varepsilon_1(t)] = E[c_1 e(t-1) + \cdots + c_{n_c} e(t-n_c) * e(t)] = 0$
- Non è possibile trovare un predittore con $\text{var}[\varepsilon_1(t)] < \text{var}[e(t)]$

Tuttavia, l'espressione di $\hat{y}(t|t-1)$ dipende dal rumore $e(t)$ e non dai dati $y(t)$.

$$\hat{y}(t|t-1) = \frac{c_1 z^{-1} + c_2 z^{-2} + \cdots + c_{n_c} z^{-n_c}}{1 + c_1 z^{-1} + c_2 z^{-2} + \cdots + c_{n_c} z^{-n_c}} y(t)$$

Passando in forma canonica ricorsiva si ottiene

$$\begin{aligned} \hat{y}(t|t-1) = & -c_1 \hat{y}(t-1|t-2) - \cdots - c_{n_c} \hat{y}(t-n_c|t-1-n_c) + \quad \text{Predizioni passate} \\ & + c_1 y(t-1) + \cdots + c_{n_c} y(t-n_c) \quad \text{Dati passati} \end{aligned}$$

Consideriamo un processo MA(n_c) in forma canonica

$$y(t) = \underbrace{e(t) + c_1 e(t-1) + \cdots + c_{k-1} e(t-k+1)}_{\text{Parte impredicibile al tempo } t-k} + \underbrace{c_k e(t-k) + \cdots + c_{n_c} e(t-n_c)}_{\text{Parte predicibile al tempo } t-k}$$

Si dimostra che il predittore ottimo dal rumore è dato dalla parte predicibile, ovvero

$$\hat{y}(t|t-k) = c_k e(t-k) + \cdots + c_{n_c} e(t-n_c)$$

La varianza di $\varepsilon_k(t)$ aumenta con l'orizzonte di predizione, fino a diventare uguale alla varianza del processo $y(t)$. Il predittore $\hat{y}(t|t-n_c-1)$ sarà il predittore banale, di solito la media del processo.

PREDITTORE OTTIMO PER PROCESSI ARMA

Sia dato un processo ARMA in forma canonica, non è immediatamente chiaro come scomporre la parte predicibile da quella impredicibile.

Idea: si ottiene $C(z)/A(z)$ come un quoziente $E(z)$ più un resto $R(z) = z^{-k} \tilde{R}(z)$:

$$\begin{array}{ccc} \text{Quoziente} & & \text{Resto} \\ C(z) = E(z) A(z) + R(z) & \Rightarrow & \frac{C(z)}{A(z)} = E(z) + \frac{R(z)}{A(z)} = E(z) + \frac{z^{-k} \tilde{R}(z)}{A(z)} \\ \text{Didivendo} & \text{Divisore} & \end{array}$$

Sostituendo l'espressione $C(z)$ in $y(t) = \frac{C(z)}{A(z)e(t)}$ otteniamo:

$$y(t) = \underbrace{E(z)e(t)}_{\text{Parte impredicibile al tempo } t} + \underbrace{\frac{\tilde{R}(z)}{A(z)}e(t-k)}_{\text{Parte predicibile al tempo } t-k}$$

Il predittore ottimo del rumore è: $\hat{y}(t|t-k) = \frac{\tilde{R}(z)}{A(z)}e(t-k)$

Calcoliamo il predittore ottimo dei dati tramite il filtro sbiancante:

$$y(t) = \frac{C(z)}{A(z)}e(t) \Rightarrow e(t) = \underbrace{\frac{A(z)}{C(z)}y(t)}_{\text{FILTRO SBIANCANTE}}$$

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

Il predittore ottimo del rumore è: $\hat{y}(t|t-k) = \frac{\tilde{R}(z)}{C(z)}y(t-k)$ mentre l'errore di predizione corrispondente è: $\varepsilon_k(t) = y(t) - \hat{y}(t|t-k) = E(z)e(t)$

Caso particolare: predizione ad un passo $k=1$:



Osserviamo che:

- $\hat{y}(t|t-k)$ è corretto e dipende dal WN fino al tempo $t-k$
- $E[\hat{y}(t|t-k) * \varepsilon_k(t)] = 0$
- Si dimostra che non è possibile trovare un predittore con $\text{Var}[\varepsilon_k(t)]$ minore

Ne segue quindi che $\hat{y}(t|t-k)$ è il predittore lineare ottimo.

Possiamo valutare la qualità del predittore mettendo a confronto la varianza dell'errore di predizione ottenuto con la varianza dell'errore di predizione di un predittore banale:

$$\text{ESR} = \frac{\text{Var}[y(t) - \hat{y}(t|t-k)]}{\text{Var}[y(t) - 0]} = \frac{\text{Var}[\varepsilon_k(t)]}{\text{Var}[y(t)]}$$

Il valore $1 - ESR$ ci fornisce la percentuale di varianza del processo che è stata <<catturata>> dal predittore e l'ESR varia tra 0 e 1: un valore di ESR inferiore indica un predittore migliore.

PREDITTORE OTTIMO PER PROCESSI ARMAX

Sia dato un processo ARMAX con $C(z)/A(z)$ in forma canonica. In questo caso, è sensato fare una previsione a k passi, in modo che l'ingresso riesca ad influenzare l'uscita. Quindi, <<confondiamo>> i k passi di previsione con i k passi di ritardo puro tra ingresso e uscita. Applichiamo k passi di lunga divisione per scomporre $C(z)/A(z)$:

$$y(t) = \underbrace{\frac{B(z)}{A(z)} u(t-k) + \frac{\tilde{R}(z)}{A(z)} e(t-k)}_{\text{Parte predicibile al tempo } t-k} + \underbrace{E(z)e(t)}_{\text{Parte impredicibile al tempo } t}$$

Il predittore ottimo dal rumore è:

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

Calcoliamo il predittore ottimo dai dati:

$$y(t) = \frac{B(z)}{A(z)} u(t-k) + \frac{C(z)}{A(z)} e(t) \quad \Rightarrow \quad e(t) = \underbrace{\frac{A(z)}{C(z)} y(t) - \frac{B(z)}{C(z)} u(t-k)}_{\text{FILTRO SBIANCANTE}}$$

$$\hat{y}(t|t-k) = \frac{B(z)}{A(z)} u(t-k) + \frac{\tilde{R}(z)}{A(z)} e(t-k) \quad \Rightarrow \quad \hat{y}(t|t-k) = \frac{\tilde{R}(z)}{C(z)} y(t-k) + \frac{B(z)E(z)}{C(z)} u(t-k)$$

Il predittore ottimo dei dati è:

$$\hat{y}(t|t-k) = \frac{\tilde{R}(z)}{C(z)} y(t-k) + \frac{B(z)E(z)}{C(z)} u(t-k)$$

L'errore di previsione corrispondente è:

$$\varepsilon_k(t) = E(z)e(t)$$

Caso particolare: previsione ad un passo $k = 1$. Il predittore dai dati a un passo è:

$$\hat{y}(t|t-1) = \frac{C(z) - A(z)}{C(z)} y(t) + \frac{B(z)}{C(z)} u(t-1)$$

mentre l'errore di previsione ad un passo è:

$$\varepsilon_1(t) = E(z)e(t) = e(t)$$

Osserviamo che:

- $\hat{y}(t|t-k)$ è corretto e dipende dal WN e da $u(t)$ fino al tempo $t-k$
- $E[\hat{y}(t|t-k) * \varepsilon_k(t)] = 0$
- Si dimostra che non è possibile trovare un predittore con $\text{Var}[\varepsilon_k(t)]$ minore

Ne segue quindi che $\hat{y}(t|t-k)$ è il predittore lineare ottimo.

La varianza di $\varepsilon_k(t)$ è data solo dalla parte ARMA, in quanto unica parte stocastica del modello. La bontà del predittore si può calcolare come

$$\text{ESR} = \frac{\text{Var}[\varepsilon_k(t)]}{\text{Var}\left[\frac{C(z)}{A(z)} e(t)\right]}$$

PREDITTORE OTTIMO AD UN PASSO PER SISTEMI INGRESSO/USCITA

Nel caso di sistemi LTI SISO ingresso/uscita, usiamo un modello $M(\theta)$ della seguente forma:

$$M(\theta): y(t) = G(z, \theta)u(t) + H(z, \theta)e(t), \quad e(t) \sim \text{WN}(0, \lambda^2)$$

Notiamo che il filtro sbiancante si ottiene come: $e(t) = H^{-1}(z, \theta)[y(t) - G(z, \theta)u(t)]$

Sostituendo l'espressione del filtro sbaincante che produce $e(t)$ nel secondo termine otteniamo:

$$\begin{aligned} y(t) &= G(z, \theta)u(t) + [H(z, \theta) - 1]H^{-1}(z, \theta)[y(t) - G(z, \theta)u(t)] + e(t) \\ &= H^{-1}(z, \theta)G(z, \theta)u(t) + [1 - H^{-1}(z, \theta)]y(t) + e(t) \end{aligned}$$

Dato che $H(z, \theta)$ è in forma canonica, anche $H^{-1}(z, \theta)$ è in forma canonica: $\frac{1}{H(z, \theta)} = 1 + d_1 z^{-1} + d_2 z^{-2} + \dots$

Supponendo che $G(z, \theta)$ sia strettamente propria, si ha che $H^{-1}(z, \theta)G(z, \theta)u(t) + [1 - H^{-1}(z, \theta)]y(t)$

dipende solo da $H(z, \theta)$, $G(z, \theta)$ e dai dati $u(t-1), u(t-2), \dots$ e $y(t-1), y(t-2), \dots$ Questa quantità è quindi completamente predicibile al tempo $t-1$.

Il predittore ottimo ad un passo è: $\hat{M}(\theta): \hat{y}(t|t-1; \theta) = H^{-1}(z, \theta)G(z, \theta)u(t) + [1 - H^{-1}(z, \theta)]y(t)$

L'errore di predizione ad un passo è: $\varepsilon_1(t; \theta) = H^{-1}(z, \theta)[y(t) - G(z, \theta)u(t)]$

Possiamo osservare che sostituendo l'equazione del sistema che genera i dati $y(t) = G_0(z)u(t) + H_0(z)e(t)$ all'interno dell'errore di predizione a un passo si ha:

$$\begin{aligned} \varepsilon_1(t; \theta) &= H^{-1}(z, \theta)[y(t) - G(z, \theta)u(t)] \\ &= H^{-1}(z, \theta)[G_0(z)u(t) + H_0(z)e(t) - G(z, \theta)u(t)] \end{aligned}$$

Se il sistema vero appartiene alla famiglia di modelli scelta, otteniamo che $\varepsilon_1(t; \theta^0) = e(t)$. Quindi il valore θ^0 è l'unico che rende $\varepsilon_1(t; \theta^0) = e(t)$ e minimizza la varianza dell'errore di predizione a un passo. Ne segue che $\varepsilon_1(t)$ è un buon indicatore della bontà di un modello dinamico.

CONFRONTO CON IL PREDITTORE DI KALMAN

La teoria della predizione vista fino ad ora è nota come predizione alla Kolmogorov-Wiener. La teoria di Kalman è più generale però la teoria KW ci fornisce la base per lo sviluppo di metodi di identificazione intuitivi ed efficaci.

IDENTIFICAZIONE: CONCETTI FONDAMENTALI

INTRODUZIONE ALL'IDENTIFICAZIONE DEI MODELLI DINAMICI

I dati possono essere raccolti o dal funzionamento nominale del sistema, oppure tramite esperimenti progettati ad-hoc, in modo da ottenere specifiche informazioni. Ci sono diverse famiglie di modelli: lineare/non lineare, tempo invariante/tempo variante e discreto/continuo. Avendo a disposizione le misure e la famiglia di modelli scelta, è necessario decidere come confrontare il modello con i dati. Questo si traduce nella definizione di una funzione di costo da minimizzare. In tutti questi aspetti, una conoscenza a priori può essere d'aiuto. Un modello potrebbe essere buono o meno a seconda dell'applicazione per il quale verrà usato.

Ipotesi di lavoro 1: i dati sono generati da un sistema LTI SISO con uscita rumorosa (ho un disturbo v che corrompe l'uscita). I parametri da stimare solo i coefficienti del numeratore e del denominatore di $G_0(z)$. Il disturbo $v(t)$ modella tutto quello che non riesco a modellare con $G(z)$.

Ipotesi di lavoro 2: il disturbo $v(t)$ è modellizzabile come un processo stocastico stazionario a spettro razionale, indipendente da $u(t)$. In questo caso vogliamo sia stimare i coefficienti del numeratore e denominatore di $G_0(z)$ sia quelli di $H_0(Z)$ (e anche stimare λ^2).

Caso particolare: non c'è ingresso $u(t)$, ovvero sto trattando una serie temporale. In pratica, misuro solo l'uscita alimentata dal rumore bianco. Vogliamo sia stimare i coefficienti del numeratore e denominatore di $H_0(z)$ e λ^2 . Posso poi calcolare $\Gamma_{yy}(\omega)$, questo approccio prende il nome di stima spettrale parametrica.

Il modello più generale che usiamo per stimare un sistema dinamico è dato da:

$$y(t) = G(z, \theta)u(t) + H(z, \theta)e(t), \quad e(t) \sim WN(0, \lambda^2)$$

$H(z, \theta)$ è il fattore spettrale canonico e $G(z, \theta)$ è la funzione di trasferimento.

Proprietà di $G(z, \theta)$:

- $G(z, \theta)$ è strettamente propria, per ipotesi, ovvero il grado del numeratore è minore del grado del denominatore. Questo fa sì che vi sia un ritardo puro k diverso da 0 tra ingresso e uscita.
- $G(z, \theta)$ può avere zeri fuori dal cerchio o numeratore e denominatore non monici.
- $G(z, \theta)$ rappresenta un sistema fisico, $H(z, \theta)$ e $e(t)$ non esistono nella realtà.

METODI A MINIMIZZAZIONE DELL'ERRORE DI PREDIZIONE (PEM)

Una volta definita la classe di modelli, potrei stimare i coefficienti θ usando la stima a massima verosimiglianza o la stima Bayesiana. Dovrei però fare ipotesi sulla distribuzione dei dati. Oppure potrei usare un approccio basato sulla minimizzazione di una somma di residui al quadrato.

Caso <<semplice>>: non mi interessa stimare un modello per $v(t)$. Abbiamo a disposizione serie temporali di ingresso e uscita. Il valore $y_\theta(t)$ è la simulazione del sistema $G(z, \theta)$ a fronte dell'ingresso $u(t)$. La stima a minimi quadrati si trova minimizzando l'errore di simulazione $\epsilon_\theta(t)$ è $\hat{\theta}_{LS} = \arg \min \sum_{t=1}^N \epsilon_\theta(t)^2$

Caso <<più difficile>>: oltre a stimare $G_0(z)$, voglio stimare anche un modello per $v(t)$. Anche se $v(t,s)$ è un processo stocastico, nella pratica una volta che i dati sono stati collezionati, è già avvenuta una scelta dell'esito $s = \bar{s}$ che ha generato quei dati osservati $y(t, s = \bar{s})$. Quindi la quantità $y(t, s = \bar{s})$ è un vettore di numeri perché frutto di una particolare realizzazione $v(t, s = \bar{s})$. Il modello invece ha come uscita $y_\theta(t, s)$ che, se non fisso un esito, è un processo stocastico. Abbiamo quindi il dilemma che non possiamo confrontare un vettore di numeri con un processo stocastico. Se conoscessi il valore dell'esito $s = \bar{s}$, allora potrei simulare l'uscita del mio modello con quell'esito e far sì che $y_\theta(t, s = \bar{s})$ sia un vettore di numeri. Tuttavia, non conosco l'esito \bar{s} .

Idea: considero come residuo $\epsilon_\theta(t)$ da minimizzare l'errore di predizione a un passo $\varepsilon_1(t; \theta)$. Definiamo quindi la stima ottenuta: $\hat{\theta}_N = \arg \min_{\theta \in \Theta} J_N(\theta)$ $J_N(\theta) = \frac{1}{N} \sum_{t=1}^N \varepsilon_1(t; \theta)^2$
E' possibile stimare la varianza λ^2 di $e(t)$ come: $\hat{\lambda}^2 = J_N(\hat{\theta}_N) = \frac{1}{N} \sum_{t=1}^N \varepsilon_1(t; \hat{\theta}_N)^2$

Abbiamo già in parte visto che l'errore di predizione a un passo gode di interessanti proprietà, che ci permettono di distinguere θ^0 (valore vero dei parametri) da un θ qualsiasi:

- Dato θ e i dati, è sempre possibile calcolare $\varepsilon_1(t, \theta^0) = e(t)$.
- Se $\exists \theta = \theta^0$ tale che $G_0(z) = G(z, \theta^0)$ e $H_0(z) = H(z, \theta^0)$, abbiamo che $\varepsilon_1(t, \theta^0) = e(t)$, ovvero ci permette di capire se il modello è buono.
- $\varepsilon_1(t, \theta^0) \neq e(t)$ per qualsiasi $\theta \neq \theta^0$.
- θ^0 minimizza la varianza dell'errore di predizione a un passo.

I metodi di stima basati sulla minimizzazione dell'errore di predizione prendono il nome di Prediction Error Methods (PEM). Se ipotizzo che il sistema appartenga alla classe di modelli scelti, e $e(t) \sim WN$ Gaussiano, lo stimatore PEM è circa uguale allo stimatore a massima verosimiglianza. La differenza sta in come i due approcci trattano l'inizializzazione del predittore. Se i dati sono molti, non c'è differenza.

IDENTIFICAZIONE PEM DI MODELLI ARX

Consideriamo un modello ARX con N dati a disposizione. Abbiamo supposto che $k = 1$. Fissando un ritardo unitario non perdiamo di generalità.

Il modello in forma di predittore è dato da:

$$\begin{aligned} \widehat{\mathcal{M}}(\theta): \hat{y}(t|t-1; \theta) &= H^{-1}(z, \theta)G(z, \theta)u(t) + [1 - H^{-1}(z, \theta)]y(t) \\ &= B(z, \theta)u(t-1) + [1 - A(z, \theta)]y(t) \end{aligned}$$

Per trovare la stima PEM minimiamo la funzione di costo:

$$J_N(\boldsymbol{\theta}) = \frac{1}{N} \sum_{t=1}^N \varepsilon_1(t; \boldsymbol{\theta})^2 = \frac{1}{N} \sum_{t=1}^N (y(t) - \hat{y}(t|t-1; \boldsymbol{\theta}))^2 = \frac{1}{N} \sum_{t=1}^N (y(t) - \underbrace{\boldsymbol{\varphi}^\top(t) \boldsymbol{\theta}}_{1 \times d} \underbrace{\boldsymbol{\theta}}_{d \times 1})^2$$

La soluzione è analoga alla stima dei minimi quadrati di un modello lineare statico: $\hat{\boldsymbol{\theta}}_N = \left[\sum_{t=1}^N \boldsymbol{\varphi}(t) \boldsymbol{\varphi}^\top(t) \right]^{-1} \cdot \left[\sum_{t=1}^N \boldsymbol{\varphi}(t) y(t) \right]$

IDENTIFICAZIONE PEM DI MODELLI ARMAX

Consideriamo un modello ARMAX con N dati a disposizione. Calcoliamo l'espressione dell'errore di predizione a un passo. Si ha che $E(z) = 1$, e quindi $\varepsilon_1(t) = e(t)$. Esprimendo $e(t)$ in funzione di $u(t)$ e $y(t)$:

$$\varepsilon_1(t; \boldsymbol{\theta}) = e(t) = \frac{A(z, \boldsymbol{\theta})}{C(z, \boldsymbol{\theta})} y(t) - \frac{B(z, \boldsymbol{\theta})}{C(z, \boldsymbol{\theta})} u(t-1)$$

Utilizzando l'approccio predittivo otteniamo:

$$J_N(\boldsymbol{\theta}) = \frac{1}{N} \sum_{t=1}^N \varepsilon_1(t; \boldsymbol{\theta})^2 = \frac{1}{N} \sum_{t=1}^N \left[\frac{A(z, \boldsymbol{\theta})}{C(z, \boldsymbol{\theta})} y(t) - \frac{B(z, \boldsymbol{\theta})}{C(z, \boldsymbol{\theta})} u(t-1) \right]^2$$

e dato che $C(z, \boldsymbol{\theta})$ è al denominatore, questa funzione di costo non è più convessa e quindi avrà dei minimi locali.

Data un'inizializzazione $\hat{\boldsymbol{\theta}}^{(0)}$ all'iterazione 0:

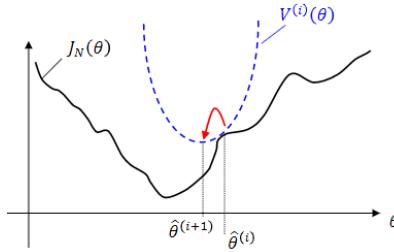
- Scegliamo N_{init} inizializzazioni $\hat{\boldsymbol{\theta}}^{(0)}$ diverse, ottenendo N_{init} soluzioni
- Se le N_{init} soluzioni sono uguali, posso pensare di aver raggiunto il minimo globale di $J_N(\boldsymbol{\theta})$
- Se le N_{init} soluzioni sono diverse, considero quella che mi ha dato $J_N(\boldsymbol{\theta})$ minore.

METODO DI NEWTON

Sviluppo in serie di Taylor troncata al 2° ordine di $J_N(\boldsymbol{\theta})$, nell'intorno della stima all'iterazione i-esima $\hat{\boldsymbol{\theta}}^{(i)}$

$$V^{(i)}(\boldsymbol{\theta}) = J_N(\hat{\boldsymbol{\theta}}^{(i)}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{(i)})^\top \underbrace{\frac{dJ_N(\boldsymbol{\theta})}{d\boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(i)}}}_{\text{Gradiente}} + \frac{1}{2} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{(i)})^\top \underbrace{\frac{d^2J_N(\boldsymbol{\theta})}{d\boldsymbol{\theta}^2} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(i)}}}_{\text{Matrice Hessiana}} \cdot (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{(i)})$$

Una volta ottenuta l'approssimazione $V^{(i)}(\boldsymbol{\theta})$ si calcola $\hat{\boldsymbol{\theta}}^{(i+1)}$ come il minimo di $V^{(i)}(\boldsymbol{\theta})$.



Troviamo l'espressione esplicita per $\hat{\boldsymbol{\theta}}^{(i+1)}$ imponendo $\frac{dV^{(i)}(\boldsymbol{\theta})}{d\boldsymbol{\theta}} = 0$

$$\frac{dV^{(i)}(\boldsymbol{\theta})}{d\boldsymbol{\theta}} = \underbrace{\frac{dJ_N(\boldsymbol{\theta})}{d\boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(i)}}}_{d \times 1} + \frac{1}{2} \cdot 2 \underbrace{\frac{d^2J_N(\boldsymbol{\theta})}{d\boldsymbol{\theta}^2} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(i)}}}_{d \times d} \cdot (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{(i)}) \stackrel{!}{=} \mathbf{0} \quad \Rightarrow \text{Ricavo il minimo e lo chiamo } \hat{\boldsymbol{\theta}}^{(i+1)}$$

Regola di update per il metodo di newton:

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

Dobbiamo quindi calcolare

$$\frac{d J_N(\boldsymbol{\theta})}{d\boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(i)}} \quad \text{Gradiente di } J_N(\boldsymbol{\theta})$$

$$\frac{d^2 J_N(\boldsymbol{\theta})}{d\boldsymbol{\theta}^2} \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(i)}} \quad \text{Hessiano di } J_N(\boldsymbol{\theta})$$

Calcolo di $\frac{d J_N(\boldsymbol{\theta})}{d\boldsymbol{\theta}}$

$$\frac{d J_N(\boldsymbol{\theta})}{d\boldsymbol{\theta}} = \frac{d}{d\boldsymbol{\theta}} \cdot \frac{1}{N} \sum_{t=1}^N \varepsilon_1(t; \boldsymbol{\theta})^2 = \frac{1}{N} \sum_{t=1}^N \frac{d}{d\boldsymbol{\theta}} \varepsilon_1(t; \boldsymbol{\theta})^2 = \frac{2}{N} \sum_{t=1}^N \varepsilon_1(t; \boldsymbol{\theta}) \cdot \frac{d\varepsilon_1(t; \boldsymbol{\theta})}{d\boldsymbol{\theta}}$$

Calcolo di $\frac{d^2 J_N(\boldsymbol{\theta})}{d\boldsymbol{\theta}^2}$

$$\frac{d^2 J_N(\boldsymbol{\theta})}{d\boldsymbol{\theta}^2} = \frac{d}{d\boldsymbol{\theta}} \frac{d J_N(\boldsymbol{\theta})}{d\boldsymbol{\theta}} = \frac{d}{d\boldsymbol{\theta}} \left[\frac{2}{N} \sum_{t=1}^N \varepsilon_1(t; \boldsymbol{\theta}) \cdot \frac{d\varepsilon_1(t; \boldsymbol{\theta})}{d\boldsymbol{\theta}} \right]$$

$$= \frac{2}{N} \sum_{t=1}^N \frac{d\varepsilon_1(t; \boldsymbol{\theta})}{d\boldsymbol{\theta}} \cdot \frac{d\varepsilon_1(t; \boldsymbol{\theta})^\top}{d\boldsymbol{\theta}} + \frac{2}{N} \sum_{t=1}^N \varepsilon_1(t; \boldsymbol{\theta}) \cdot \frac{d^2\varepsilon_1(t; \boldsymbol{\theta})}{d\boldsymbol{\theta}^2}$$

$$\frac{d(vu)}{dx} = v \cdot \frac{du}{dx} + v \frac{dv}{dx} u^\top$$

Nel nostro caso:
 $v = \varepsilon_1(t; \boldsymbol{\theta})$
 $u = \frac{d\varepsilon_1(t; \boldsymbol{\theta})}{d\boldsymbol{\theta}}$
 $x = \boldsymbol{\theta}$

Ignoriamo il secondo termine, approssimando l'hessiana, dato che:

- Se siamo vicini all'ottimo, $\varepsilon_1(t, \boldsymbol{\theta})$ è <<piccolo>> e il termine conta poco.
- Possiamo evitare di calcolare $\frac{d^2\varepsilon_1(t, \boldsymbol{\theta})}{d\boldsymbol{\theta}^2}$
- Ci assicuriamo una Hessiana semi definita positiva (equivale a dire che la parabola è rivolta sempre verso l'alto, il minimo della parabola porta al minimo della funzione costo)

L'aggiornamento da $\widehat{\boldsymbol{\theta}}^{(i)}$ a $\widehat{\boldsymbol{\theta}}^{(i+1)}$, in generale, può essere fatto con tre categorie di metodi:

- Metodo del gradiente
 - Semplice e robusto
 - Può essere molto lento quando ci avviciniamo al minimo
- Metodo di Newton
 - Converge velocemente
 - Computazionalmente più complesso
 - Rischio di instabilità se l'Hessiana è definita negativa
- Metodi <<quasi Newtoniani>>
 - Più semplice del metodo di Newton
 - Più veloce del metodo del gradiente
 - Non c'è rischio di allontanarsi dal minimo
 - Non è veloce come il metodo di Newton

I metodi <<quasi Newtoniani>> sono molto usati e differiscono fra loro nel modo in cui viene fatta l'approssimazione. Per garantire l'invertibilità di $O^{-1} \geq 0$, si aggiunge un termine positivo, molto piccolo, di regolarizzazione: $\frac{d^2 J_N(\theta)}{d\theta^2} \approx \frac{2}{N} \sum_{t=1}^N \frac{d\varepsilon_1(t; \theta)}{d\theta} \cdot \frac{d\varepsilon_1(t; \theta)^\top}{d\theta} + \delta I_d$

Dopo aver introdotto l'approssimazione dell'Hessiana, la regola di update diventa:

$$\widehat{\theta}^{(i+1)} = \widehat{\theta}^{(i)} - \left[\frac{2}{N} \sum_{t=1}^N \frac{d\varepsilon_1(t; \theta)}{d\theta} \Big|_{\theta=\widehat{\theta}^{(i)}} \cdot \frac{d\varepsilon_1(t; \theta)^\top}{d\theta} \Big|_{\theta=\widehat{\theta}^{(i)}} \right]^{-1} \cdot \left[\frac{2}{N} \sum_{t=1}^N \varepsilon_1(t; \widehat{\theta}^{(i)}) \cdot \frac{d\varepsilon_1(t; \theta)}{d\theta} \Big|_{\theta=\widehat{\theta}^{(i)}} \right]_{d \times d}$$

Calcoliamo $\frac{d\varepsilon_1(t; \theta)}{d\theta}$

Ricordiamo che $\varepsilon_1(t; \theta) = \varepsilon(t) = \frac{A(z, \theta)}{C(z, \theta)} y(t) - \frac{B(z, \theta)}{C(z, \theta)} u(t-1)$

$$\varepsilon_1(t; \theta) = \frac{1 - a_1 z^{-1} - \dots - a_{n_a} z^{-n_a}}{1 + c_1 z^{-1} + \dots + c_{n_c} z^{-n_c}} y(t) - \frac{b_0 + b_1 z^{-1} + \dots + b_{n_b} z^{-n_b}}{1 + c_1 z^{-1} + \dots + c_{n_c} z^{-n_c}} u(t-1)$$

$$\theta = [a_1 \dots a_{n_a} \ b_0 \ b_1 \dots b_{n_b} \ c_1 \dots c_{n_c}]^\top$$

Derivate di $\varepsilon_1(t; \theta)$ rispetto a a_1, a_2, \dots, a_{n_a}

$$\frac{d\varepsilon_1(t)}{da_1} = -\frac{z^{-1}}{C(z)} y(t) = \alpha(t-1)$$

$$\frac{d\varepsilon_1(t)}{da_2} = -\frac{z^{-2}}{C(z)} y(t) = \alpha(t-2)$$

$$\alpha(t) \equiv -\frac{1}{C(z)} y(t)$$

:

$$\frac{d\varepsilon_1(t)}{da_{n_a}} = -\frac{z^{-n_a}}{C(z)} y(t) = \alpha(t-n_a)$$

Derivate di $\varepsilon_1(t; \theta)$ rispetto a b_0, b_1, \dots, b_{n_b}

$$\frac{d\varepsilon_1(t)}{db_0} = -\frac{1}{C(z)} u(t-1) = \beta(t-1)$$

$$\frac{d\varepsilon_1(t)}{db_1} = -\frac{z^{-1}}{C(z)} u(t-1) = \beta(t-2)$$

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

Derivate di $\varepsilon_1(t; \theta)$ rispetto a c_1, c_2, \dots, c_{n_c}

$$\varepsilon_1(t) = \frac{A(z)}{C(z)} y(t) - \frac{B(z)}{C(z)} u(t-1)$$



$$(1 + c_1 z^{-1} + c_2 z^{-2} + \dots + c_{n_c} z^{-n_c}) \varepsilon_1(t) = A(z) y(t) - B(z) u(t-1)$$

Non dipende da c_1

$$\frac{d[(1 + c_1 z^{-1} + c_2 z^{-2} + \dots + c_{n_c} z^{-n_c}) \cdot \varepsilon_1(t)]}{dc_1} = \frac{d[A(z)y(t) - B(z)u(t-1)]}{dc_1}$$

$$\frac{d[(1 + c_1 z^{-1} + c_2 z^{-2} + \dots + c_{n_c} z^{-n_c}) \cdot \varepsilon_1(t)]}{dc_1} = 0 \quad \text{Devo fare la derivata del prodotto}$$

Derivate di $\varepsilon_1(t; \theta)$ rispetto a c_1, c_2, \dots, c_{n_c}

$$\frac{d[(1 + c_1 z^{-1} + c_2 z^{-2} + \dots + c_{n_c} z^{-n_c}) \cdot \varepsilon_1(t)]}{dc_1} = 0$$

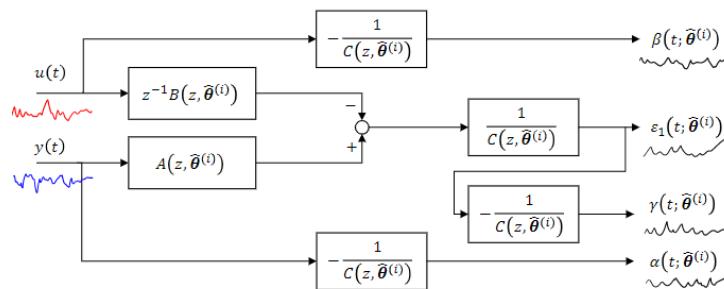
$$y(t) \equiv -\frac{1}{C(z)} \cdot \varepsilon_1(t)$$

$$z^{-1} \cdot \varepsilon_1(t) + C(z) \frac{d\varepsilon_1(t)}{dc_1} = 0 \quad \Rightarrow \quad \frac{d\varepsilon_1(t)}{dc_1} = -\frac{1}{C(z)} \cdot \varepsilon_1(t-1) = \gamma(t-1)$$

:

$$\frac{d\varepsilon_1(t)}{dc_{n_c}} = -\frac{1}{C(z)} \cdot \varepsilon_1(t-n_c) = \gamma(t-n_c)$$

È possibile definire in modo elegante il calcolo del gradiente tramite una serie di filtri per i segnali di ingresso e uscita. Abbiamo il seguente schema di filtraggio dei segnali per trovare il gradiente:



Riassunto dell'implementazione dell'algoritmo di Newton per modelli ARMAX:

1. Calcolare i polinomi $A(z, \hat{\theta}^{(i)})$, $B(z, \hat{\theta}^{(i)})$, $C(z, \hat{\theta}^{(i)})$ all'iterazione i -esima.
2. Calcolare i segnali $\varepsilon_1(z, \hat{\theta}^{(i)})$, $\alpha(z, \hat{\theta}^{(i)})$, $\beta(z, \hat{\theta}^{(i)})$, $\gamma(z, \hat{\theta}^{(i)})$ filtrando i dati u, y disponibili
3. Costruire il vettore gradiente $\frac{d\varepsilon_1(t, \theta)}{d\theta} |_{\theta = \hat{\theta}^{(i)}}$ coi segnali ricavati al passo 2
4. Aggiornare la stima dei parametri tramite la regola di update

Prima di filtrare attraverso $1/C(z, \hat{\theta}^{(i)})$, dobbiamo controllare che $C(z, \hat{\theta}^{(i)})$ abbia radici interne al cerchio unitario. Se non è il caso, possiamo utilizzare un filtro passa tutto per rendere $1/C(z, \hat{\theta}^{(i)})$ as. stabile.

IDENTIFICAZIONE – ANALISI E COMPLEMENTI

ANALISI ASINTOTICA DEI METODI PEM

Nella lezione precedente abbiamo visto come ottenere una stima $\hat{\theta}_N$ data una singola sequenza di N dati, utilizzano l'approccio PEM. Abbiamo esplicitato l'esito $s = \bar{s}$ (da cui dipende la y) per indicare che lavoriamo con delle sequenze di numeri. Idea: data una famiglia di modelli, la stima $\hat{\theta}_N$ ottenuta minimizzando la funzione di costo è:

$$J_N(\boldsymbol{\theta}) = \frac{1}{N} \sum_{t=1}^N \varepsilon_1(t; \boldsymbol{\theta})^2$$

Problema: la stima $\hat{\theta}_N$ ci fornisce un buon modello? Dobbiamo considerare tutte le possibili realizzazioni di sequenze di dati.

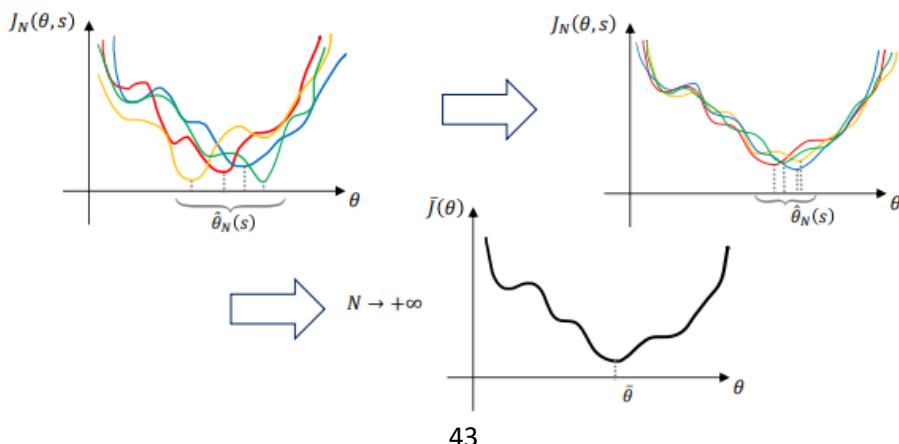
Ipotesi di lavoro: sia l'ingresso, sia l'uscita sono pss ed ergodici (implica che tutte le fdt siano as. stabili).

I dati misurati saranno una realizzazione dei processi in corrispondenza di un particolare esito \bar{s}

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

La funzione di costo dipende anch'essa dall'esito \bar{s} poiché utilizza i dati misurati: $J_N(\boldsymbol{\theta}, \bar{s}) = \frac{1}{N} \sum_{t=1}^N \varepsilon_1(t; \boldsymbol{\theta}, \bar{s})^2$

da cui otterrò la stima $\hat{\theta}_N(\bar{s})$. In generale la stima $\hat{\theta}_N(s)$ è una variabile casuale perché il suo valore dipende dai dati, i quali dipendono dall'esito s .



Grazie all'ipotesi di ergodicità i momenti temporali convergono ai rispettivi momenti d'insieme → studio il caso asintotico. Quindi:

$$J_N(\boldsymbol{\theta}, s) = \frac{1}{N} \sum_{t=1}^N \varepsilon_1(t; \boldsymbol{\theta}, s)^2 \xrightarrow[N \rightarrow +\infty]{} \bar{J}(\boldsymbol{\theta}) \equiv \mathbb{E}_s[\varepsilon_1(t, \boldsymbol{\theta})^2]$$

cioè, le curve $J_N(\theta, s)$ convergono ad un'unica (deterministica) curva $\bar{J}(\theta)$. Definiamo l'insieme dei punti di minimo globale di $\bar{J}(\theta)$ come: $\Delta_{\theta} = \{\bar{\theta} \mid \bar{J}(\theta) \geq \bar{J}(\bar{\theta}), \forall \theta\}$

Caso particolare: $\bar{J}(\theta)$ ha un unico minimo globale.

Teorema: sotto le ipotesi correnti, man mano che il numero di dati N tende all'infinito, si ha che

$$J_N(\theta, s) \rightarrow \bar{J}(\theta) \quad \hat{\vartheta}_N \rightarrow \Delta_{\theta} \quad \text{per } n \rightarrow \infty. \quad \text{Ne segue che } \Delta_{\theta} = \bar{\theta}, \text{ allora } \hat{\vartheta}_N \rightarrow \bar{\theta}$$

Il risultato dell'identificazione PEM è lo stesso, indipendentemente dalle realizzazioni misurate dei processi $u(t), y(t)$ purchè il numero di dati N sia abbastanza grande (tende a infinito).

Idea: per studiare le proprietà della stima, studiamo le sue caratteristiche asintotiche, ovvero studiamo il modello stimato asintotico $M(\bar{\theta})$ oppure l'insieme dei modelli stimati asintotici. Se le proprietà di $M(\bar{\theta})$ sono buone, posso pensare che lo siano anche quelle di $M(\hat{\vartheta}_N)$ (modello con un numero finito di dati), fintanto che N è grande (grande rispetto al numero di dati da stimare).

Ipotesi di lavoro aggiuntiva: assumiamo che $S \in M(\theta)$, ovvero che esista $\theta^0 \in \Theta$ tale che $S = M(\theta^0)$ (sistema: modello valutato in θ^0). Domanda: il vettore <>vero<> dei parametri θ^0 appartiene all'insieme Δ_{θ} dei minimi globali della cifra di costo $\bar{J}(\theta)$? Ciò è equivalente a chiedersi se $\hat{\vartheta}_N$ tende asintoticamente a θ^0 .

Se ciò fosse vero, vorrebbe dire che i metodi PEM sono in grado di trovare la parametrizzazione <>vera<> del modello. Dimostriamo che, sotto le ipotesi fatte, θ^0 appartiene sempre a Δ_{θ} (DOMANDA ESAME!!!!!!)

Supponiamo che i dati siano generati dal sistema S, tale che

$$y(t) = \hat{y}(t|t-1; \theta^0) + e(t), \quad e(t) \sim WN(0, \lambda^2) \quad \bullet \quad \text{Vera quando } \hat{y} \text{ è il predittore del sistema}$$

Consideriamo un generico modello $M(\theta)$ (al quale corrisponde un suo predittore che dipende da θ), per il quale

$$y(t) = \hat{y}(t|t-1; \theta) + \varepsilon_1(t; \theta) \quad \begin{array}{l} \text{Non è detto che } \varepsilon_1(t; \theta) \\ \text{sia bianco...} \end{array} \quad \rightarrow \text{ solo se il predittore è costruito con la struttura e i parametri del sistema vero.}$$

L'errore di predizione ad un passo commesso dal modello $M(\theta)$ è dunque

$$\varepsilon_1(t; \theta) = y(t) - \hat{y}(t|t-1; \theta) \rightarrow \text{riscrivo l'espressione sopra.}$$

Aggiungiamo e togliamo $\hat{y}(t|t-1; \theta^0)$ del sistema vero, ovvero il predittore del sistema S che genera i dati

$$\varepsilon_1(t; \theta) = \underbrace{y(t) - \hat{y}(t|t-1; \theta^0)}_{\text{Errore di predizione «ottimo»}} + \hat{y}(t|t-1; \theta^0) - \hat{y}(t|t-1; \theta)$$

Errore di predizione «ottimo»

$$\varepsilon_1(t; \theta^0) = e(t)$$

- Errore di predizione a un passo quando costruito con i dati veri

$$\varepsilon_1(t; \theta) = e(t) + \hat{y}(t|t-1; \theta^0) - \hat{y}(t|t-1; \theta)$$

Calcoliamo la varianza dell'errore di predizione:

$$\begin{aligned} \mathbb{E}[\varepsilon_1(t; \theta)^2] &= \mathbb{E}[(e(t) + \hat{y}(t|t-1; \theta^0) - \hat{y}(t|t-1; \theta))^2] \\ \Rightarrow \bar{J}(\theta) &= \mathbb{E}[e(t)^2] + \mathbb{E}[(\hat{y}(t|t-1; \theta^0) - \hat{y}(t|t-1; \theta))^2] \\ &\quad + 2\mathbb{E}[e(t) \cdot (\hat{y}(t|t-1; \theta^0) - \hat{y}(t|t-1; \theta))] \end{aligned}$$

Le quantità $\hat{y}(t|t-1; \theta^0)$ e $\hat{y}(t|t-1; \theta)$ sono predittori, e quindi dipendono solo dai dati a tempi passati. Per cui, sono incorrelati con $e(t)$.

$$\begin{aligned} \bar{J}(\theta) &= \mathbb{E}[e(t)^2] + \mathbb{E}[(\hat{y}(t|t-1; \theta^0) - \hat{y}(t|t-1; \theta))^2] \\ \Rightarrow \bar{J}(\theta) &= \lambda^2 + \underbrace{\mathbb{E}[(\hat{y}(t|t-1; \theta^0) - \hat{y}(t|t-1; \theta))^2]}_{\substack{\text{È una varianza, quindi una quantità } \geq 0. \text{ In particolare, si annulla solo per } \theta = \theta^0}} \\ \Rightarrow \bar{J}(\theta) &\geq \lambda^2 = \bar{J}(\theta^0), \quad \forall \theta \end{aligned}$$

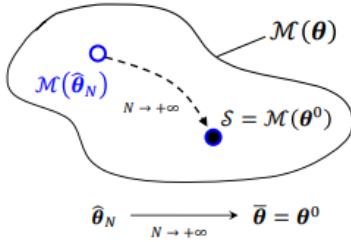
$$\boxed{\bar{J}(\theta) \geq \bar{J}(\theta^0), \quad \forall \theta}$$
 θ^0 è un minimo di $\bar{J}(\theta)$

Conclusione: se $S \in M(\theta)$ e $u(t), y(t)$ sono pss ergodici, allora, per $n \rightarrow \inf$, un metodo PEM garantisce che il modello stimato è quello <>vero>>.

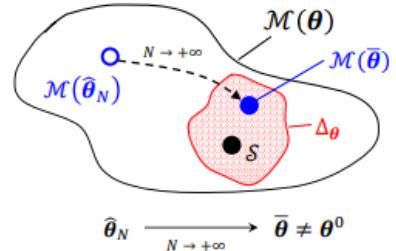
Se $S \notin M(\theta)$ (il sistema non appartiene all'insieme di modelli scelti), allora i metodi PEM non garantiscono di stimare correttamente tutte le componenti del sistema S. Se $S \in M(\theta)$, allora in corrispondenza di θ^0 si ha che $\varepsilon_1(t; \theta^0) = e(t) \sim WN$. Quindi, possiamo verificare a posteriori se il modello identificato è quello vero facendo un test di bianchezza sui residui $\varepsilon_1(t; \hat{\theta}^N)$.

Quando identifichiamo un modello $M(\theta)$, possono capitarcisi quattro casi possibili:

1) $S \in \mathcal{M}(\theta)$ e $\Delta_\theta = \bar{\theta}$, allora $\bar{\theta} = \theta^0$

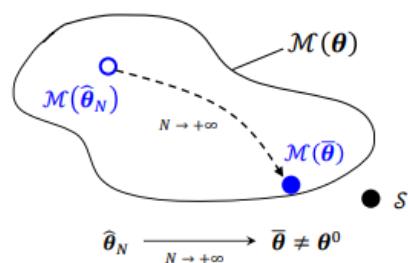


2) $S \in \mathcal{M}(\theta)$ e Δ_θ contiene più valori



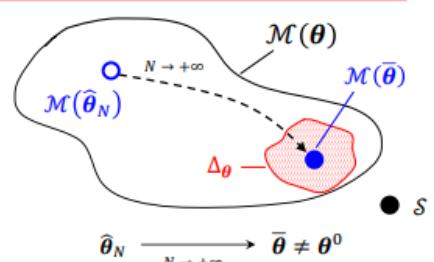
Ma $\mathcal{M}(\bar{\theta})$ ha la **stessa capacità** di $\mathcal{M}(\theta^0)$ nello spiegare i dati

3) $S \notin \mathcal{M}(\theta)$ e $\Delta_\theta = \bar{\theta}$, allora $\bar{\theta} \neq \theta^0$



$\mathcal{M}(\bar{\theta})$ è la **miglior approssimazione** di S nella famiglia di modelli $\mathcal{M}(\theta)$

4) $S \notin \mathcal{M}(\theta)$ e Δ_θ contiene più valori



$\mathcal{M}(\bar{\theta})$ con $\bar{\theta} \in \Delta_\theta$ sono i **migliori approssimanti (equivalenti)** di S nella famiglia di modelli $\mathcal{M}(\theta)$

IDENTIFICABILITA' DEI MODELLI E PERSISTENTE ECCITAZIONE

L'analisi asintotica vista precedentemente ci dice che, se $S \in \mathcal{M}(\theta)$, allora i metodi PEM stimano asintoticamente il modello vero $S = M(\theta^0)$ o un insieme equivalente di modelli.

Questa situazione, in cui troviamo un modello all'interno di $\{M(\theta) | \theta \in \Delta_\theta\}$, porta ad una legittima domanda: in quali condizioni il sistema S può essere identificato univocamente dai dati? Vogliamo un modello unico.

1. Identificabilità strutturale: il modello $M(\theta)$ non deve essere sovraparametrizzato rispetto al sistema $S \rightarrow$ quando non è troppo complesso rispetto al sistema vero
2. Identificabilità sperimentale: i dati $\{u(t), y(t)\}_{t=1}^N$ devono contenere sufficiente informazione.

Il problema di non identificabilità più critico è quello sperimentale: se non abbiamo sufficiente informazione nei dati, non possiamo fare nulla.

La non identificabilità strutturale è, invece, facilmente risolvibile riducendo l'ordine di modello

Investighiamo il problema di identificabilità sperimentale, considerando per semplicità l'identificazione di un modello ARX($n_a, n_b, 1$), avendo N dati. Sappiamo che la stima può essere ottenuta tramite il metodo dei minimi quadrati:

somma in forma chiusa \Rightarrow

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

Problema di identificabilità:

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

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

Le matrici $S(N)$ e $R(N)$ sono semidefinite positive in quanto prodotto di vettore per sé stesso. Affinché $\hat{\theta}^N$ esista e sia unico, è però necessario che $S(N) > 0$ o $R(N) > 0$, cioè che $\det(R(N)) > 0$.

Analizziamo la matrice $R(N)$ per $N \rightarrow \infty$. Consideriamo come punto di partenza un ARX(1,0,1): $y(t) = a_1 y(t-1) + b_0 u(t-1) + e(t)$ con $e(t) \sim WN(0, \lambda^2)$ dove $y(t)$ e $u(t)$ sono pss ergodici a media nulla

$$y(t) = a_1 y(t-1) + b_0 u(t-1) + e(t), \quad e(t) \sim WN(0, \lambda^2) \quad \theta = \begin{bmatrix} a \\ b \end{bmatrix} \quad \varphi(t) = \begin{bmatrix} y(t-1) \\ u(t-1) \end{bmatrix}$$

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

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

Notiamo che $R(N)$ contiene «somme temporali»

Grazie all'ipotesi di ergodicità, abbiamo che $R(N) \rightarrow \bar{R}$ con

$$\bar{R} = \begin{bmatrix} \gamma_{yy}(0) & \gamma_{yu}(0) \\ \gamma_{uy}(0) & \gamma_{uu}(0) \end{bmatrix}$$

La matrice \bar{R} è la matrice di autocovarianza del processo congiunto $\{y(t), u(t)\}$.

Idea: trovare le condizioni per cui \bar{R} è invertibile. Quando queste condizioni valgono allora possiamo supporre con ragionevole certezza che, per N grande, anche $R(N)$ è invertibile.

In generale per un modello ARX abbiamo che la matrice \bar{R} può essere divisa in 4 sotto-matrici

$$\bar{R} = \begin{bmatrix} \bar{R}_{yy} & \bar{R}_{yu} \\ \bar{R}_{uy} & \bar{R}_{uu} \end{bmatrix}$$

$n_a \times n_a$ $n_a \times (n_b + 1)$
 $(n_b + 1) \times n_a$ $(n_b + 1) \times (n_b + 1)$

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

Data una matrice nella forma $M = \begin{bmatrix} F & K \\ K^T & H \end{bmatrix}$, con F e H simmetriche. Condizione necessaria e sufficiente per l'invertibilità di M è che valgano.

- $H > 0$
- $F - KH^{-1}K^T > 0$

Ricordando che $\bar{R} = \begin{bmatrix} \bar{R}_{yy} & \bar{R}_{yu} \\ \bar{R}_{uy} & \bar{R}_{uu} \end{bmatrix}$ → condizione necessaria per l'invertibilità di \bar{R} è che $\bar{R}_{uu} > 0$

La condizione (solo necessaria) sulla matrice \bar{R}_{uu} è interessante perché riguarda solo il segnale d'ingresso $u(t)$, che tipicamente progettiamo noi. Possiamo quindi tenere conto di questa condizione in fase di progettazione dell'esperimento, e scegliere il segnale di eccitazione più opportuno al fine di ottenere dati informativi.

Def di persistente eccitazione: definiamo la matrice $\bar{R}_{uu}^{(i)}$ di autocovarianza di $u(t)$ di ordine i come

$$\bar{R}_{uu}^{(i)} = \begin{bmatrix} \gamma_{uu}(0) & \gamma_{uu}(1) & \dots & \gamma_{uu}(i-1) \\ \gamma_{uu}(1) & \gamma_{uu}(0) & \dots & \gamma_{uu}(i-2) \\ \dots & \dots & \dots & \dots \\ \gamma_{uu}(i-1) & \gamma_{uu}(i-2) & \dots & \gamma_{uu}(0) \end{bmatrix}_{i \times i}$$

Il segnale $u(t)$ è detto persistentemente eccitante di ordine n se

- $\bar{R}_{uu}^{(1)} > 0, \bar{R}_{uu}^{(2)} > 0, \dots, \bar{R}_{uu}^{(n)} > 0$
- $\bar{R}_{uu}^{(n+1)} \geq 0, \bar{R}_{uu}^{(n+2)} \geq 0, \dots, \bar{R}_{uu}^{(n+)} \geq 0$

Ovvero n è il massimo ordine per cui $\bar{R}_{uu}^{(i)}$ è invertibile.

Se un segnale $u(t)$ è persistentemente eccitante di ordine n , allora è anche persistentemente eccitante di ordine $n-1$. Ribadiamo che la condizione vista è solamente necessaria: anche se $\bar{R}_{uu}^{(i)} > 0$, la \bar{R} potrebbe comunque non essere invertibile per ragioni di non identificabilità strutturale. Il concetto di persistente eccitazione che abbiamo visto è stato esemplificato per la stima di modelli ARX, ma avere un segnale eccitante è importante in ogni caso si voglia identificare un modello dinamico.

Nella pratica, però, è impossibile generare un rumore bianco «perfetto»: al più, le sequenze di numeri saranno pseudo-casuali, e non casuali. A causa di limiti dell'elettronica, il segnale generato e trasmesso agli attuatori sarà «filtrato passa-basso», per cui non si avrà uno spettro «perfettamente piatto». Inoltre, talvolta non si vuole sollecitare troppo gli attuatori ad alta frequenza per non rovinarli. Inoltre, l'ampiezza del rumore bianco non è «limitata». Talvolta, è necessario garantire che l'attuatore non saturi l'ingresso, al fine di non introdurre nonlinearità nell'esperimento e nei dati misurati.

Il segnale di tipo PRBS è un segnale deterministico, periodico, a tempo discreto, che commuta tra due livelli. L'utente deve definire i due livelli $[-\bar{u}; +\bar{u}]$, il periodo e l'intervallo di clock.

Il periodo viene posto uguale al numero di dati N che si vuole collezionare, e l'intervallo di clock a un tempo di campionamento (che è il minimo possibile).

La funzione di autocovarianza di un max.length PRBS può essere espressa come

$$\gamma_{uu}(\tau) = \frac{1}{M} \sum_{t=1}^M (u(t) - m_u)(u(t + \tau) - m_u) = \begin{cases} \bar{u}^2 \left(1 - \frac{1}{T} \right) & \text{se } \tau = 0, \pm M, \pm 2M, \dots \\ -\frac{\bar{u}^2}{T} \left(1 + \frac{1}{T} \right) & \text{se } \tau \neq 0, \pm M, \pm 2M, \dots \end{cases}$$

m_u è la media del PRBS, che non è esattamente zero

Quando $T \rightarrow \inf$ il PRBS approssima un rumore bianco.

Il segnale multiseno (composta da più seni) è un segnale periodico, definito come una media pesata di sinusoidi, con frequenze multiple della risoluzione in frequenza della DFT $f_0 = f_s / N$.

$$u(t) = \sum_{k=0}^F A_k \cdot \cos(2\pi \cdot k f_0 \cdot t + \phi_k)$$

Il numero F di componenti in frequenza deve soddisfare il teorema del campionamento. Gli sfasamenti ϕ_k sono in generale scelti in modo casuale, e si possono anche ottimizzare per minimizzare il valore di picco del segnale.

Molto spesso le ampiezze A_k vengono scelte ad un valore costante nella banda di frequenze di interesse, e 0 altrove.

Quando l'ingresso $u(t)$ da un multiseno, anche il segnale di uscita $y(t)$ di un sistema LTI è un multiseno. Di solito si scartano i primi periodi del segnale multiseno generato per rimuovere gli effetti del transitorio. Quando si progetta un multiseno, si può fissare la risoluzione in frequenza desiderata e la massima frequenza eccitata, per calcolare automaticamente la lunghezza $N \cdot P$ del segnale.

VALUTAZIONE DELL'INCERTEZZA DELLA STIMA PEM

Abbiamo visto che, quando $S \in \mathcal{M}(\theta)$ e se l'ingresso è sufficientemente eccitante da rendere i dati informativi, i metodi PEM portano a stimare il valore vero dei parametri. Questo risultato vale però per $N \rightarrow \inf$ (abbiamo quindi visto proprietà asintotiche).

Ipotesi di lavoro:

- $S \in M(\theta)$, per cui $\theta^0 \in \Delta_\theta$
- $\Delta_\theta \in \bar{\theta}$, ovvero esiste un solo punto di minimo globale. Quindi $\theta^0 = \bar{\theta}$

Ipotizziamo di avere un numero finito di dati e di stimare $\hat{\theta}_N = \arg \min \frac{1}{n} \sum_{t=1}^N \varepsilon_1(t, \theta)^2 \rightarrow$ minimizzazione della varianza campionaria dell'errore di predizione a un passo

Ricordiamo che $\hat{\theta}_N$ è una variabile casuale in quanto i dati provengono da realizzazioni di processi stocastici.
Vogliamo calcolare l'incertezza di stima parametrica $\text{var}[\hat{\theta}_N] = E[(\hat{\theta}_N - \theta^0) * (\hat{\theta}_N - \theta^0)^\top]$. Si dimostra che:

$$\text{Var}[\hat{\theta}_N] \equiv \bar{P}_{\theta} = \frac{1}{N} \lambda^2 \cdot \bar{R}_{\theta}^{-1}$$

- $\bar{R}_{\theta} = E[\psi(t; \theta^0) \psi(t; \theta^0)^\top]$
- $\psi(t; \theta^0) = -\frac{d}{d\theta} \varepsilon_1(t; \theta) \Big|_{\theta=\theta^0}$
- $\lambda^2 = \text{Var}[e(t)]$

Tali quantità dipendono da θ^0 . Nella pratica, si approssimano come

$$\hat{\lambda}^2 = \frac{1}{N} \sum_{t=1}^N (y(t) - \hat{y}(t|t-1; \hat{\theta}_N))^2 = J(\hat{\theta}_N)$$

$$\hat{R}_{\theta} = \frac{1}{N} \sum_{t=1}^N \psi(t; \hat{\theta}) \cdot \psi(t; \hat{\theta})^\top$$

Ricordiamo che $\bar{J}(\theta) = E[\varepsilon_1(t, \theta)^2]$. Riprendendo quanto visto per la stima ARMAX:

$$\frac{d\bar{J}(\theta)}{d\theta} = E \left[2\varepsilon_1(t; \theta) \cdot \frac{d\varepsilon_1(t; \theta)}{d\theta} \right] \quad \frac{d^2\bar{J}(\theta)}{d\theta^2} = E \left[2 \frac{d\varepsilon_1(t; \theta)}{d\theta} \cdot \frac{d\varepsilon_1(t; \theta)^\top}{d\theta} + 2\varepsilon_1(t; \theta) \cdot \frac{d^2\varepsilon_1(t; \theta)}{d\theta^2} \right]$$

La derivata seconda di $\varepsilon_1(t, \theta)$ è funzione dell'errore di predizione e pertanto dipende dai valori passati.

Notiamo però che se $\theta = \theta^0$, allora $\varepsilon_1(t, \theta) = e(t)$. Quindi:

$$\frac{d^2\bar{J}(\theta)}{d\theta^2} \Big|_{\theta=\theta^0} = E \left[2 \frac{d\varepsilon_1(t; \theta)}{d\theta} \Big|_{\theta=\theta^0} \cdot \frac{d\varepsilon_1(t; \theta)^\top}{d\theta} \Big|_{\theta=\theta^0} \right] = 2 \cdot E[\psi(t; \theta^0) \psi(t; \theta^0)^\top] = 2 \cdot \bar{R}_{\theta}$$



$$\bar{R}_{\theta} = \frac{1}{2} \cdot \frac{d^2\bar{J}(\theta)}{d\theta^2} \Big|_{\theta=\theta^0}$$

\bar{R}_{θ} è la metà dell'Hessiana della funzione di costo valutata nell'ottimo

Notiamo che la varianza dell'errore di predizione di stima nei parametri decresce all'aumentare di N. la varianza dell'errore di stima dei parametri aumenta all'aumentare di λ^2 . La varianza dell'errore di stima dei parametri diminuisce all'aumentare della derivata seconda della funzione di costo all'ottimo.

Inoltre, più grande è la potenza del segnale di ingresso u(t), più piccola è la matrice di varianza delle stime \bar{P}_{θ} , questo perché \bar{P}_{θ} è proporzionale all'inverso della potenza del vettore di segnali

$$\Psi(t; \theta) = -\frac{d\varepsilon_1(t, \theta)}{d\theta}$$

e questo vettore di segnali è più potente tanto più u(t) è potente.

Stima ARX: la stima ottenuta tramite l'algoritmo dei minimi quadrati è: $\hat{y}(t|t-1; \theta) = \varphi^T(t)\theta$

L'errore di predizione a un passo è:

$$\varepsilon_1(t; \theta) = y(t) - \hat{y}(t|t-1; \theta) = y(t) - \varphi^T(t)\theta$$

Quindi $\psi(t; \theta) = -\frac{d}{d\theta} \varepsilon_1(t; \theta) = \varphi(t)$ e di conseguenza

$$\bar{P}_\theta = \frac{1}{N} \lambda^2 \cdot \bar{R}_\theta^{-1} = \frac{1}{N} \lambda^2 \cdot \mathbb{E}[\psi(t; \theta^0) \psi(t; \theta^0)^T]^{-1} = \frac{1}{N} \lambda^2 \cdot \mathbb{E}[\varphi(t; \theta^0) \varphi(t; \theta^0)^T]^{-1}$$

Usando la stima campionaria di \bar{P}_θ abbiamo che:

$$\begin{aligned} \bar{P}_\theta &= \frac{1}{N} \lambda^2 \cdot \mathbb{E}[\varphi(t; \theta^0) \varphi(t; \theta^0)^T]^{-1} \approx \frac{1}{N} \hat{\lambda}^2 \cdot \left[\frac{1}{N} \sum_{t=1}^N \varphi(t; \hat{\theta}) \cdot \varphi(t; \hat{\theta}_N)^T \right]^{-1} \\ &= \hat{\lambda}^2 \cdot \left[\sum_{t=1}^N \varphi(t; \hat{\theta}) \cdot \varphi(t; \hat{\theta}_N)^T \right]^{-1} = \hat{\lambda}^2 \cdot S(N)^{-1} \end{aligned}$$

Le proprietà probabilistiche della stima PEM di modelli ARX sono uguali a quelle della stima a minimi quadrati di modelli lineari «statici».

Se $S \in M(\theta)$, la distribuzione $\hat{\theta}_N$ ottenuta tramite stima PEM converge asintoticamente ad una Gaussiana $\hat{\theta}_N \sim N(\theta^0, \bar{P}_\theta)$. Questa reazione con la stima di \bar{P}_θ al posto di \bar{P}_θ può essere usata nella pratica per calcolare intervalli di confidenza sulla stima $\hat{\theta}_N$, e valutare così l'affidabilità della stima di un certo parametro.

Tali intervalli di confidenza ci dicono la probabilità p_θ che l'intervallo di confidenza contenga il vettore vero dei parametri: $\mathcal{C}_\theta = \{\theta \mid (\theta - \hat{\theta}_N)^T \cdot \bar{P}_\theta^{-1} \cdot (\theta - \hat{\theta}_N) \leq \alpha\}$

Dato che $\hat{\theta}_N$ è una variabile casuale, anche i modelli identificati $G(z, \hat{\theta}_N)$ e $H(z, \hat{\theta}_N)$ lo saranno. In molte situazioni è probabilmente più di interesse analizzare la varianza della stima delle funzioni di trasferimento $G(z, \hat{\theta}_N)$ e $H(z, \hat{\theta}_N)$, piuttosto che la varianza delle stime dei parametri $\hat{\theta}_N$.

Assumendo sempre che se $S \in M(\theta)$ e che $u(t)$ sia incorrelato a $e(t)$, l'espressione può essere approssimata come: $\text{Var}[G(e^{j\omega}, \hat{\theta}_N)] \approx \frac{n}{N} \cdot \frac{\Gamma_{vv}(\omega)}{\Gamma_{uu}(\omega)}$

Notiamo che possiamo fare «input shaping» dell'ingresso $u(t)$ per «favorire» la stima in una certa banda di frequenze piuttosto che in altre.

La varianza della stima della funzione di trasferimento del rumore $H(z, \hat{\theta}_N)$ è $\text{Var}[H(e^{j\omega}, \hat{\theta}_N)] \approx \frac{n}{N} \cdot \frac{\Gamma_{vv}(\omega)}{\lambda^2}$

dove λ^2 è la varianza del rumore $e(t)$ che alimenta $H_0(z)$.

IDENTIFICAZIONE – VALUTAZIONE DEL MODELLO

SCELTA DELLA STRUTTURA E COMPLESSITÀ DEL MODELLO

La scelta della famiglia di modelli $M(\theta)$ appropriata può essere scomposta in due diversi aspetti:

1. Scelta della struttura del modello: scelta della struttura delle funzioni di trasferimento $G(z, \theta)$ e $H(z, \theta)$
2. Scelta della complessità del modello: scelta degli ordini dei polinomi delle fdt

L'obiettivo è quello di trovare un buon modello ad un prezzo ragionevole. Nel caso generale, un modello è buono se ha poco bias e poca varianza: nel caso specifico, un modello deve essere buono per l'utilizzo che se ne deve fare.

Per prezzo ragionevole si intende quanto sforzo è necessario per identificare e utilizzare il modello: quanto tempo ci vuole per trovare la stima e quando un modello è di ordine ridotto.

Vi è un trade-off tra bontà e prezzo del modello. Due aspetti importanti:

- La complessità computazionale dei metodi di ottimizzazione non lineare e il vantaggio di usare schemi di regressione lineare
- L'abilità di modellare bene $G_0(z)$ anche se $H_0(z)$ non è modellato bene

In base alla fisica del processo che deve essere modellato, è possibile avere informazioni sul minimo ordine del modello necessario ed è possibile avere informazioni su come pre-processare i segnali a disposizione.

L'analisi non parametrica tramite ETFE può dare importanti informazioni sull'ordine del modello soprattutto per quanto riguarda la posizione di risonanze. Un modello dovrebbe limitarsi a modellare al massimo 3 decadi in frequenza:

- Per un modello che usa dati campionati ad alta frequenza, le dinamiche lente vengono viste come integratori
- Per un modello a bassa frequenza, le dinamiche veloci vengono viste come relazioni statiche
- Se necessario, costruire più modelli con dati campionati a frequenze diverse

Una volta stimati i parametri di un modello, guardiamo le loro deviazioni standard. Se è tale da includere lo zero, allora quel parametro potrebbe non essere significativo. Questa analisi permette di scegliere il ritardo puro k più opportuni, identificando diversi modelli con diversi valori di k , e scegliendo quello per cui tutti i coefficienti $B(z)$ sono significativi.

Per rendersi conto del ritardo del sistema e della sua linearità è possibile effettuare una risposta allo scalino, con diverse ampiezze di scalino.

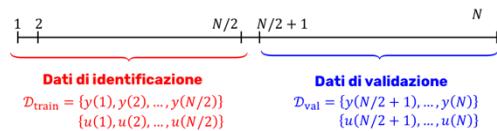
La bontà di un modello può essere valutata:

- Analizzando i residui, meglio con dati di validazione
- Confrontando l'uscita simulata o predetta con l'uscita misurata, su dati di validazione, e calcolando un indicatore FIT
- Rappresentando un grafico polo – zero con rispettive bande di confidenza, per vedere se vi sono cancellazioni
- Rappresentando diagrammi di Bode di diversi modelli e confrontandoli con la ETFE

VALIDAZIONE O FORMULE DI COMPLESSITÀ OER LA SCELTA DELLA COMPLESSITÀ

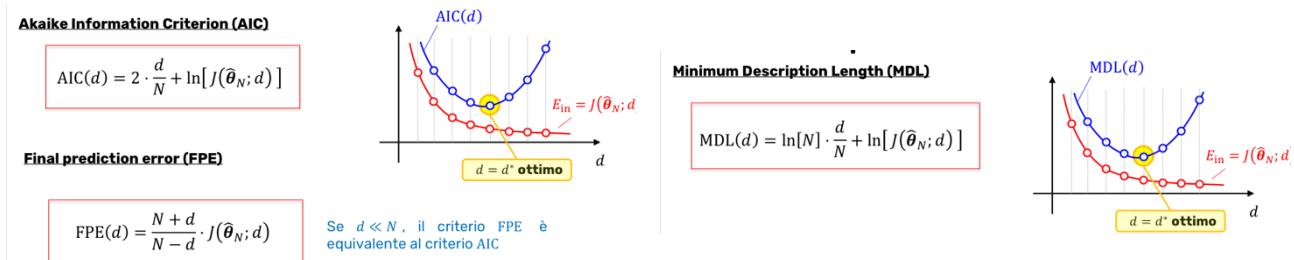
Fissata la struttura di una famiglia di modelli $M(\theta)$, dobbiamo poi scegliere la complessità del modello (numero di parametri). Un metodo semplice ma efficace consiste nell'identificare un insieme di modelli di diversa complessità utilizzando un dataset di identificazione, e confrontandone la bontà su un dataset di validazione (problema multidimensionale).

Il metodo della validazione è molto simile a quello visto per i sistemi statici. Supponiamo di avere N dati, e dividiamoli in 2 sotto sequenze:



Per ogni ordine $m = 1, \dots, M$, identifichiamo un modello minimizzando $J(\theta, \mathcal{D}_{\text{train}})$ e calcoliamo $J(\hat{\theta}_{N/2}, \mathcal{D}_{\text{val}})$ sui dati di validazione. Scegliamo l'ordine m^* che minimizza $J(\hat{\theta}_{N/2}, \mathcal{D}_{\text{val}})$.

A differenza del caso statico, con i sistemi dinamici non è possibile estrarre i dati di identificazione e di validazione in modo casuale dal dataset completo, perché romperei la causalità temporale dei dati. Anche in questo caso dinamico, la validazione è una procedura che da risultati molto buoni, ma richiede tanti dati.



ANALISI DEI RESIDUI

Ricordiamo che

$$\varepsilon_1(t; \theta) = \frac{1}{H(z, \theta)} [(G_0(z) - G(z, \theta))u(t) + H_0(z)e(t)]$$

$e(t) \sim WN(0, \lambda^2)$ è il rumore sul sistema vero

Sommiamo e sottraiamo $e(t)$:

$$\begin{aligned}\varepsilon_1(t; \boldsymbol{\theta}) &= \frac{1}{H(z, \boldsymbol{\theta})} [(G_0(z) - G(z, \boldsymbol{\theta}))u(t) + H_0(z)e(t)] - e(t) + e(t) \\ &= \frac{1}{H(z, \boldsymbol{\theta})} [(G_0(z) - G(z, \boldsymbol{\theta}))u(t) + (H_0(z) - H(z, \boldsymbol{\theta}))e(t)] + e(t)\end{aligned}$$

$$= \frac{G_0(z) - G(z, \boldsymbol{\theta})}{H(z, \boldsymbol{\theta})} u(t) + \frac{H_0(z) - H(z, \boldsymbol{\theta})}{H(z, \boldsymbol{\theta})} e(t) + e(t)$$

Se $\exists \boldsymbol{\theta}^0$ t.c. $G(\boldsymbol{\theta}^0) = G_0$
e $H(\boldsymbol{\theta}^0) = H_0$, allora
 $\varepsilon_1(t, \boldsymbol{\theta}^0) = e(t)$

La stima asintoti a quando $n \rightarrow \infty$ può quindi essere ottenuta come $\bar{\boldsymbol{\theta}} = \arg \min J(\boldsymbol{\theta}) = \arg \min E[\varepsilon_1(t; \boldsymbol{\theta})]^2$

In frequenza, $J(\boldsymbol{\theta})$ è esprimibile come

$$J(\boldsymbol{\theta}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{|G_0(e^{j\omega}) - G(e^{j\omega}, \boldsymbol{\theta})|^2 \cdot \Gamma_{uu}(\omega) + |H_0(e^{j\omega}) - H(e^{j\omega}, \boldsymbol{\theta})|^2 \cdot \lambda^2}{|H(e^{j\omega}, \boldsymbol{\theta})|^2} d\omega$$

L'espressione mette in risalto come la stima è ottenuta minimizzando l'errore di stima del modello I/O ($G_0 - G$) e del modello del rumore ($H_0 - H$), pesati per la densità spettrale del rispettivo segnale di ingresso. Inoltre, vi è una pesatura pari all'inverso del modello del rumore.

Con il prefiltraggio tramite filtro $L(z)$ dei dati abbiamo $u_F(t) = L(z)u(t)$ e $y_F(t) = L(z)y(t)$ e la funzione di costo asintotica diventa

$$J(\boldsymbol{\theta}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{|G_0(e^{j\omega}) - G(e^{j\omega}, \boldsymbol{\theta})|^2 \cdot \Gamma_{uu}(\omega) + |H_0(e^{j\omega}) - H(e^{j\omega}, \boldsymbol{\theta})|^2 \cdot \lambda^2}{|H(e^{j\omega}, \boldsymbol{\theta})|^2} \cdot |L(e^{j\omega})|^2 d\omega$$

Dopo aver selezionato un modello $M(\boldsymbol{\theta})$ e averne effettuato l'identificazione PEM, è possibile valutarne la struttura e la complessità tramite analisi dei residui.

Obiettivo: avendo la stima $\hat{\boldsymbol{\theta}}_N$ e i dati $\{u(t), y(t)\}_{t=1}^N$, determinare se $M(\boldsymbol{\theta})$ è tale che

- $S \in M(\boldsymbol{\theta}) \rightarrow$ caso migliore. Sia G che H del modello includono la parte G ed H del sistema
- $S \notin M(\boldsymbol{\theta})$ con $G_0(z) \in G(\boldsymbol{\theta}) \rightarrow$ parte esogena G del modello contiene quella del sistema. Molto interessante nella pratica in cui vogliamo che $G(z, \hat{\boldsymbol{\theta}}_N) \rightarrow G_0(z)$ anche se il modello dell'errore è sbagliato
- $S \notin M(\boldsymbol{\theta})$ con $G_0(z) \notin G(\boldsymbol{\theta}) \rightarrow$ sia G che H del modello non contengono G ed H del sistema

Consideriamo il caso asintotico $N \rightarrow +\infty$, in cui $\hat{\boldsymbol{\theta}}_N \rightarrow \bar{\boldsymbol{\theta}}$. Abbiamo che:

$$\varepsilon_1(t; \bar{\boldsymbol{\theta}}) = H^{-1}(z; \bar{\boldsymbol{\theta}}) (y(t) - G(z, \bar{\boldsymbol{\theta}})u(t)) = \frac{G_0(z) - G(z, \bar{\boldsymbol{\theta}})}{H(z, \bar{\boldsymbol{\theta}})} u(t) + \frac{H_0(z)}{H(z, \bar{\boldsymbol{\theta}})} e(t)$$

La scelta della struttura e della complessità del modello $M(\boldsymbol{\theta})$ può essere effettuata osservando:

- La funzione di autocovarianza dei residui: $\gamma_{\varepsilon\varepsilon}(\tau)$. Se il modello è perfetto, l'errore di predizione diventa un rumore bianco

- La funzione di cross-covarianza tra i residui ed il segnale di ingresso: $\gamma_{\varepsilon u}(\tau)$. Posso capire se ho stimato bene la parte G

Analizziamo tre situazioni:

- A) Situazione A (situazione ideale): $\gamma_{\varepsilon\varepsilon}(\tau) = \lambda^2 * \delta(\tau)$ e $\gamma_{\varepsilon u}(\tau) = 0 \forall \tau$

Supponiamo di osservare:

$$\gamma_{\varepsilon\varepsilon}(\tau) = \begin{cases} \lambda^2 & \text{se } \tau = 0 \\ 0 & \text{se } \tau \neq 0 \end{cases} \quad \gamma_{\varepsilon u}(\tau) = 0 \quad \forall \tau$$

Questa situazione accade quando

$$\varepsilon_1(t; \bar{\theta}) = \frac{G_0(z) - G(z, \bar{\theta})}{H(z, \bar{\theta})} u(t) + \frac{H_0(z)}{H(z, \bar{\theta})} e(t) = 0 \cdot u(t) + 1 \cdot e(t)$$

Ovvero se e solo se $G(z, \bar{\theta}) = G_0(z)$ e $H(z, \bar{\theta}) = H_0(z)$. Questo avviene se e solo se $S \in M(\theta)$

- B) Situazione B (bene G_0 ma non ho H): $\gamma_{\varepsilon\varepsilon}(\tau) \neq \lambda^2 * \delta(\tau)$ e $\gamma_{\varepsilon u}(\tau) = 0 \forall \tau$

Supponiamo di osservare

$$\gamma_{\varepsilon\varepsilon}(\tau) \neq \lambda^2 * \delta(\tau) \quad \gamma_{\varepsilon u}(\tau) = 0 \quad \forall \tau$$

Il residuo non si comporta come un WN, tuttavia è incorrelato con l'ingresso.

$$\varepsilon_1(t; \bar{\theta}) = \frac{G_0(z) - G(z, \bar{\theta})}{H(z, \bar{\theta})} u(t) + \frac{H_0(z)}{H(z, \bar{\theta})} e(t) = 0 \cdot u(t) + \underbrace{\frac{H_0(z)}{H(z, \bar{\theta})}}_{\neq 1} \cdot e(t)$$

Ovvero se e solo se $G(z, \bar{\theta}) = G_0(z)$ e $H(z, \bar{\theta}) \neq H_0(z)$.

Questa situazione avviene se e solo se $S \notin M(\theta)$ con $G_0(z) \in G(\theta)$ per $M(\theta)$ OE, BJ, FIR.

Infatti, se $M(\theta)$ è OE, BJ o FIR è possibile parametrizzare in modo indipendente $G(z, \eta)$ e $H(z, \xi)$, con $\theta = [\eta^T \ \xi^T]^T$

$$J(\theta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{|G_0(e^{j\omega}) - G(e^{j\omega}, \eta)|^2 \cdot \Gamma_{uu}(\omega) + |H_0(e^{j\omega}) - H(e^{j\omega}, \xi)|^2 \cdot \lambda^2}{|H(e^{j\omega}, \xi)|^2} \cdot d\omega$$

Il vettore $\bar{\theta}$ che minima questa cifra di merito è $\bar{\theta} = \begin{bmatrix} \bar{\eta} \\ \bar{\xi} \end{bmatrix} = \begin{bmatrix} \eta_0 \\ \xi_0 \end{bmatrix}$

Per cui abbiamo che $G(z, \bar{\eta}) = G_0(z)$ e $H(z, \bar{\xi}) \neq H_0(z)$

Il fatto di poter stimare bene $G_0(z)$ anche se non stimo bene $H_0(z)$, non accade se usiamo un modello ARX o ARMAX, che se $G_0(z) \in G(\theta)$. Infatti la funzione di costo è

$$\bar{J}(\theta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{|G_0(e^{j\omega}) - G(e^{j\omega}, \theta)|^2 \cdot \Gamma_{uu}(\omega) + |H_0(e^{j\omega}) - H(e^{j\omega}, \theta)|^2 \cdot \lambda^2}{|H(e^{j\omega}, \theta)|^2} \cdot d\omega$$

Per cui, anche se esistesse θ_0 tale che $G(z, \theta_0) = G_0(z)$, tale vettore minimizza solo il termine $|G_0(e^{j\omega}) - G(e^{j\omega}, \theta)|^2$, ma non $|H_0(e^{j\omega}) - H(e^{j\omega}, \theta)|^2$, poiché $H(z, \theta_0) \neq H_0(z)$. Ne consegue che $\bar{\theta} \neq \theta_0$ e quindi $G_0(z)$ non viene stimata in modo corretto anche se $G_0(z) \in G(\theta)$. Si può comunque arrivare ad una buona approssimazione aumentando l'ordine del modello.

C) Situazione C (non posso dire nulla su come ho stimato G ed H): $\gamma_{\varepsilon\varepsilon}(\tau) \neq \lambda^2 * \delta(\tau)$ e $\exists \tau \gamma_{\varepsilon u}(\tau) \neq 0$

Supponiamo di osservare

$$\gamma_{\varepsilon\varepsilon}(\tau) \neq \lambda^2 \cdot \delta(\tau)$$

$$\exists \tau \text{ t.c. } \gamma_{\varepsilon u}(\tau) \neq 0$$

Questa situazione accade quando

$$\varepsilon_1(t; \bar{\theta}) = \overbrace{\frac{G_0(z) - G(z, \bar{\theta})}{H(z, \bar{\theta})}}^{\neq 0} u(t) + \frac{H_0(z)}{H(z, \bar{\theta})} e(t)$$

Ovvero se e solo se $G(z, \bar{\theta}) \neq G_0(z)$

Questo avviene se e solo se

$$\left\{ \begin{array}{l} S \notin M(\theta) \text{ con } G_0(z) \in \mathcal{G}(\theta) \text{ per } M(\theta) \text{ ARX, ARMAX} \\ S \notin M(\theta) \text{ con } G_0(z) \notin \mathcal{G}(\theta) \end{array} \right.$$

Dove $\delta(\tau)$ è un delta di Dirac centrata in τ

I) $M(\theta)$ è OE, BJ o FIR

- **Situazione A:** $\gamma_{\varepsilon\varepsilon}(\tau) = \lambda^2 \cdot \delta(\tau)$ e $\gamma_{\varepsilon u}(\tau) = 0 \forall \tau \rightarrow S \in M(\theta)$
- **Situazione B:** $\gamma_{\varepsilon\varepsilon}(\tau) \neq \lambda^2 \cdot \delta(\tau)$ e $\gamma_{\varepsilon u}(\tau) = 0 \forall \tau \rightarrow S \notin M(\theta) \text{ con } G_0(z) \in \mathcal{G}(\theta)$
- **Situazione C:** $\gamma_{\varepsilon\varepsilon}(\tau) \neq \lambda^2 \cdot \delta(\tau)$ e $\exists \tau \text{ t.c. } \gamma_{\varepsilon u}(\tau) \neq 0 \rightarrow S \notin M(\theta) \text{ con } G_0(z) \notin \mathcal{G}(\theta)$

II) $M(\theta)$ è ARX, ARMAX

- **Situazione A:** $\gamma_{\varepsilon\varepsilon}(\tau) = \lambda^2 \cdot \delta(\tau)$ e $\gamma_{\varepsilon u}(\tau) = 0 \forall \tau \rightarrow S \in M(\theta)$
- **Situazione C:** $\gamma_{\varepsilon\varepsilon}(\tau) \neq \lambda^2 \cdot \delta(\tau)$ e $\exists \tau \text{ t.c. } \gamma_{\varepsilon u}(\tau) \neq 0 \rightarrow S \notin M(\theta)$

	$N \rightarrow +\infty$		N finito	
	$\gamma_{\varepsilon\varepsilon}(\tau)$	$\gamma_{\varepsilon u}(\tau)$	$\hat{\gamma}_{\varepsilon\varepsilon}(\tau)$	$\hat{\gamma}_{\varepsilon u}(\tau)$
$S \in M(\theta)$	$0 \forall \tau \neq 0$	$0 \forall \tau$	«piccola» ∈ intervallo di confidenza	«piccola» ∈ intervallo di confidenza
$S \notin M(\theta)$ $G_0(z) \in \mathcal{G}(\theta)$	$\exists \tau \neq 0 \text{ t.c. } \gamma_{\varepsilon\varepsilon}(\tau) \neq 0$	$0 \forall \tau$	«grande» ∉ intervallo di confidenza	«piccola» ∈ intervallo di confidenza OE, BJ, FIR
$S \notin M(\theta)$ $G_0(z) \notin \mathcal{G}(\theta)$	$\exists \tau \neq 0 \text{ t.c. } \gamma_{\varepsilon\varepsilon}(\tau) \neq 0$	$\exists \tau \text{ t.c. } \gamma_{\varepsilon u}(\tau) \neq 0$	«grande» ∉ intervallo di confidenza	«grande» ∉ intervallo di confidenza

La procedura vista ora basta su un test statico dei residui serve a validare l'ipotesi che $S \in M(\theta)$ sulla base dei dati disponibili. Una validazione della struttura che si conclude con un successo non garantisce che $G(z, \hat{\theta}_N)$ e $H(z, \hat{\theta}_N)$ siano buone stime di $G_0(z)$ e $H_0(z)$. È necessario controllare anche la varianza delle stime.

ANALISI DELL'INCERTEZZA DELLA STIMA

Con l'analisi dell'incertezza della stima intendiamo:

1. Incertezza sulla stima dei parametri: può essere utilizzata per verificare la significatività statistica di un parametro, dove per significatività intendiamo quanto è probabile che il parametro vero sia effettivamente diverso da 0
2. Incertezza sulla stima delle FDT (costruite usando i parametri)

3. Incertezza sulla posizione dei poli e degli zeri (dipendono dai parametri)

L'incertezza sulla stima dei parametri può essere utilizzata per verificare la significatività statistica di un parametro, dove per significatività intendiamo quanto è probabile che il parametro vero sia effettivamente diverso da 0.