Imad domande

Silviu Filote

July 2023

Contents

1 Lezione 4: Stima a massima verosimiglianza

Cos'è la massima verosimiglianza e calcolarla

Il metodo della massima verosimiglianza (MLE – Maximum Likelihood Estimation) 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 $y(i) \sim \mathcal{N}(\mu, \sigma^2)$ iid e $Y = [y(1), y(2), \ldots, y(N)]^T$.

La pdf congiunta dei dati o la probability density function:

Indica la probabilità che si realizzi il vettore di dati osservato

$$f_Y(y(1), y(2), \ldots, y(N)|\mu, \sigma^2) = f_Y(Y|\mu, \sigma^2) = \prod_{i=1}^{N} \mathcal{N}(y(i)|\mu, \sigma^2)$$
$$= f_y(y(1)|\mu, \sigma^2) \cdot \ldots \cdot f_y(y(N)|\mu, \sigma^2)$$

Quando la **pdf multivariabile** $f_Y(Y|\mu, \sigma^2)$ è vista in funzione dei parametri μ e σ è chiamata **funzione di likelihood** $\mathcal{L}_Y(\mu, \sigma^2|Y)$, cambia solo interpretazione, ma sono lo stesso oggetto matematico, dunque $\boldsymbol{\theta}_{ML} = \left[\mu, \sigma^2\right]$.

$$f(Y|\mu, \sigma^2) = \mathcal{L}(\mu, \sigma^2|Y) = \prod_{i=1}^{N} \mathcal{N}\left(y(i)|\mu, \sigma^2\right)$$

La stima a massima verosimigliana può essere espressa come:

$$\underset{2\times1}{\widehat{\boldsymbol{\theta}}_{ML}} = \begin{bmatrix} \widehat{\mu} \\ \widehat{\sigma}^2 \end{bmatrix} = arg \ \max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|Y) = arg \ \max_{\boldsymbol{\theta}} \prod_{i=1}^{N} \mathcal{N}\left(y(i)|\mu, \sigma^2\right)$$

In generale posso attribuire ai dati qualsiasi distribuziome di probabilitá $d(\cdot)$, sia continua che discreta:

$$\underset{d\times1}{\widehat{\boldsymbol{\theta}}_{ML}} = arg \ \max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|Y)$$

Lo stimatore a massima verosimiglianza gode di buone proprietá:

- **Asintoticamente corretto:** $\lim_{N\to+\infty} \mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{ML}\right] = \boldsymbol{\theta}^0$
- **Consistente:** più N é grande, più la stima é precisa
- **Asintoticamente efficiente:** $\lim_{N\to+\infty} Var\left[\widehat{\boldsymbol{\theta}}_{ML}\right] = M^{-1}$
- **Asintoticamente normale:** $\widehat{\boldsymbol{\theta}}_{ML} \sim \mathcal{N}(\boldsymbol{\theta}^0, M^{-1})$ per $N \to +\infty$

Note:

- Spesso, anziché massimizzare $\mathcal{L}(\boldsymbol{\theta}|Y)$, si massimizza il suo logaritmo naturale $ln \ \mathcal{L}(\boldsymbol{\theta}|Y)$
- Dato che il logaritmo è una funzione monotona crescente, $ln(\mathcal{L}(\boldsymbol{\theta}|Y))$ ha lo stesso massimo di $\mathcal{L}(\boldsymbol{\theta}|Y)$
- Usare il logaritmo è efficiente da un punto di vista implementativo, perchè evita possibili underflow dati dal prodotto di piccole probabilità (sostituendolo con la somma delle log-probabilità)
- A meno di casi particolari fortunati, l'ottimizzazione è effettuata con metodi iterativi

$$\underset{d\times1}{\widehat{\boldsymbol{\theta}}_{ML}} = arg \ \max_{\boldsymbol{\theta}} ln \ \mathcal{L}(\boldsymbol{\theta}|Y)$$

NB: massimizzare la log-verosimiglianza equivale a **minimizzare la meno log-verosimiglianza:**

$$\underset{d\times1}{\widehat{\boldsymbol{\theta}}_{ML}} = arg \ \max_{\boldsymbol{\theta}} ln \left[\mathcal{L}(\boldsymbol{\theta}|Y)\right] = arg \ \min_{\boldsymbol{\theta}} -ln \left[\mathcal{L}(\boldsymbol{\theta}|Y)\right]$$

Formulando il problema in questo modo abbiamo un problema di minimizzazione proprio come con lo ***stimatore a minimi quadrati****:* $\underset{d\times1}{\widehat{\boldsymbol{\theta}}_{LS}} = arg \ \min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$

Stima a massima verosimiglianza di un modello lineare

Imponiamo un modello probabilistico alle osservazioni $y(i)$:

$$y(i) = \theta_0 + \theta_1\varphi_1(i) + \cdots + \theta_{d-1}\varphi_{d-1}(i) + \epsilon(i) = \boldsymbol{\varphi}^T(i)\cdot\boldsymbol{\theta} + \epsilon(i) \qquad \textbf{\textit{modello lineare}}$$

dove $\boldsymbol{\varphi}$ é il vettore delle features e $\boldsymbol{\theta}$ é il vettore dei parametri

$$\underset{d\times 1}{\boldsymbol{\varphi}} = \begin{bmatrix} 1 \\ \varphi_1 \\ \vdots \\ \varphi_{d-1} \end{bmatrix} \qquad \underset{d\times 1}{\boldsymbol{\theta}} = \begin{bmatrix} 1 \\ \theta \\ \vdots \\ \theta_{d-1} \end{bmatrix}$$

In particolare, se assumiamo che $\epsilon \sim \mathcal{N}(0,\lambda^2)$ iid

$$\Rightarrow \quad y(i) \sim \mathcal{N}(\boldsymbol{\varphi}^T(i)\cdot\boldsymbol{\theta}, \lambda^2) \qquad iid$$

$$\mu(i) = \boldsymbol{\varphi}^T(i)\cdot\boldsymbol{\theta}$$

La media $\mu(i)$ di $y(t)$ é espressa come combinazione lineare dei regressori

La **distribuzione congiunta** dei dati è:

$$\begin{aligned} f_Y\big(y(1),y(2),\ldots,y(N)|X,\boldsymbol{\theta},\lambda^2\big) &\underset{iid}{=} \prod_{i=1}^{N} f_y\big(y(i)|\boldsymbol{\varphi}(i),\boldsymbol{\theta},\lambda^2\big) \\ &= \prod_{i=1}^{N} \mathcal{N}\big(y(i)|\boldsymbol{\varphi}(i),\boldsymbol{\theta},\lambda^2\big) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\lambda^2}}\cdot exp\bigg[-\frac{1}{2}\bigg(\frac{y(i)-\boldsymbol{\varphi}^T(i)\cdot\boldsymbol{\theta}}{\lambda}\bigg)^2\bigg] \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\lambda^2}}\cdot exp\bigg[-\frac{1}{2}\bigg(\frac{y(i)-\mu(i)}{\lambda}\bigg)^2\bigg] \\ &= \mathcal{L}(\boldsymbol{\theta}|Y,X,\lambda^2) \;\; \textit{supponiamo noto } \lambda^2 \end{aligned}$$

Calcolo la **log-verosimiglianza**:

$$\begin{aligned} ln\big[\mathcal{L}(\boldsymbol{\theta}|Y,X,\lambda^2)\big] &= ln\bigg[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\lambda^2}}\cdot exp\bigg[-\frac{1}{2}\bigg(\frac{y(i)-\boldsymbol{\varphi}^T(i)\cdot\boldsymbol{\theta}}{\lambda}\bigg)^2\bigg]\bigg] \\ &= \sum_{i=1}^{N} ln\bigg[\frac{1}{\sqrt{2\pi\lambda^2}}\cdot exp\bigg[-\frac{1}{2}\bigg(\frac{y(i)-\boldsymbol{\varphi}^T(i)\cdot\boldsymbol{\theta}}{\lambda}\bigg)^2\bigg]\bigg] \\ &= \sum_{i=1}^{N} ln\bigg(\frac{1}{\sqrt{2\pi\lambda^2}}\bigg) + \sum_{i=1}^{N} ln\bigg[exp\bigg[-\frac{1}{2}\bigg(\frac{y(i)-\boldsymbol{\varphi}^T(i)\cdot\boldsymbol{\theta}}{\lambda}\bigg)^2\bigg]\bigg] \\ &= N\cdot ln\bigg(\frac{1}{\sqrt{2\pi\lambda^2}}\bigg) + \sum_{i=1}^{N} -\frac{1}{2}\cdot\bigg(\frac{y(i)-\boldsymbol{\varphi}^T(i)\cdot\boldsymbol{\theta}}{\lambda}\bigg)^2 \\ &= N\cdot ln\big(2\pi\lambda^2\big)^{-\frac{1}{2}} - \frac{1}{2\lambda^2}\sum_{i=1}^{N}\big(y(i)-\boldsymbol{\varphi}^T(i)\cdot\boldsymbol{\theta}\big)^2 \\ &= -\frac{1}{2}N\cdot ln\big(2\pi\lambda^2\big) - \frac{1}{2\lambda^2}\sum_{i=1}^{N}\big(y(i)-\boldsymbol{\varphi}^T(i)\cdot\boldsymbol{\theta}\big)^2 \end{aligned}$$

Calcolare il massimo di $ln\big[\mathcal{L}(\boldsymbol{\theta}|Y,X,\lambda^2)\big]$ è equivalente a calcolare il minimo di $-ln\big[\mathcal{L}(\boldsymbol{\theta}|Y,X,\lambda^2)\big]$:

Siccome non dipende da $\boldsymbol{\theta}$, questo termine non contribuisce al calcolo del minimo

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

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

$$\underset{d\times 1}{\hat{\boldsymbol{\theta}}_{ML}} = arg\ \min_{\boldsymbol{\theta}} \frac{1}{2\lambda^2}\sum_{i=1}^{N}\big(y(i) - \boldsymbol{\varphi}^T(i)\cdot\boldsymbol{\theta}\big)^2$$

La stima ML del modello lineare $y(i) = \boldsymbol{\varphi}^T(i)\cdot\boldsymbol{\theta} + \epsilon(i)$, con $\epsilon \sim \mathcal{N}(0,\lambda^2)$ iid, é **equivalente alla stima a minimi quadrati** (che non aveva assunzioni sulla pdf dei dati).

Sia dato il modello lineare seguente:

$$y(i) = \boldsymbol{\varphi}(i)^T \cdot \boldsymbol{\theta} + \epsilon(i)$$

dove:

- $y(i) \in \mathbb{R}^{1\times 1}$ è la variabile risposta per l'osservazione i
- $\boldsymbol{\varphi}(i) \in \mathbb{R}^{d\times 1}$ il vettore delle features
- $\boldsymbol{\theta} \in \mathbb{R}^{d\times 1}$ è il vettore dei parametri ed $\epsilon(i) \approx N(0,\lambda^2)$
- $\epsilon(i) \in \mathbb{R}^{1\times 1}$ è un rumore gaussiano *iid* ed è indipendente da $\boldsymbol{\varphi}(i)$
- $y(i) \approx N(\boldsymbol{\varphi}(i)^T, \boldsymbol{\theta}, \lambda^2)$ dove $\boldsymbol{\varphi}(i)^T \cdot \boldsymbol{\theta} = \mu(i)$

Ricavare la stima a Massima Verosimiglianza del vettore dei parametri $\boldsymbol{\theta}$, supponendo di aver osservato N valori della variabile y.

Vedi domanda precedente.

---

2 Lezione 5: Regressione logistica

Descrivere il modello di regressione logistica indicando:

- la tipologia di dati che modella
- funzione di costo da minimizzare
- interpretazione della funzione di costo
- quali algoritmi si possono usare per la minimizzazione della funzione di costo

Tipologia di dati

Il modello di regressione logistica viene applicato quando la variabile di risposta risulta essere **qualitativa** e non quantitativa. I dati di tipo categorico descrivono una categoria di appartenenza, non hanno un ordinamento e non hanno alcuna metrica di distanza. Il processo di stima di output categorici, utilizzando un insieme di regressori $\boldsymbol{\varphi}$, è chiamato **classificazione**. Tramite classificazione la categoria più probabile viene scelta come **classe** (categoria) per l'osservazione $\boldsymbol{\varphi}$.

Supponiamo di avere a disposizione un dataset $\mathcal{D} = \big\{\{\boldsymbol{\varphi}(\mathbf{1}), y(1)\}, \dots, \{\boldsymbol{\varphi}(\boldsymbol{N}), y(N)\}\big\}$, dove $\boldsymbol{\varphi} \in \mathbb{R}^{d\times 1}$ e $y(i) \in \{0, 1\}$, con $i = 1, \dots, N$ iid (y variabile categorica).

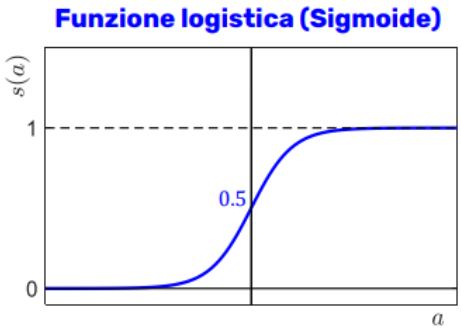
Obiettivo: Stimare la probabilità che le osservazioni $\boldsymbol{\varphi} \in \mathbb{R}^{d\times 1}$ appartengano ad una di due classi $y \in \{0, 1\}$

Definiamo la combinazione lineare:

$$a = \sum_{j=0}^{d-1} \varphi_j \cdot \theta_j = \boldsymbol{\varphi}^T \cdot \boldsymbol{\theta} \in \mathbb{R}$$

*La formula $s(a)$ é la **funzione logistica:***

$$s(a) = \frac{1}{1+e^{-a}} = \frac{e^a}{a+e^a} = \begin{cases} a \gg 0 & \Rightarrow \quad s(a) \approx 1 \\ a \ll 0 & \Rightarrow \quad s(a) \approx 0 \end{cases}$$



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

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

L'output di $s(a) = s(\boldsymbol{\varphi}^T\boldsymbol{\theta})$ é interpretato come una probabilitá:

$$\begin{cases} s(\boldsymbol{\varphi}^T\boldsymbol{\theta}) \gg 0.5 \Rightarrow P(y = 1|\boldsymbol{\varphi}) \approx 1 & \boldsymbol{\varphi} \text{ é classificato nella classe 1} \\ s(\boldsymbol{\varphi}^T\boldsymbol{\theta}) \ll 0.5 \Rightarrow P(y = 1|\boldsymbol{\varphi}) \approx 0 & \boldsymbol{\varphi} \text{ é classificato nella classe 0} \end{cases}$$

Interpretiamo i dati come realizzazioni di una distribuzione di Bernoulli:

$$y \sim Bernoulli(\pi) = \pi^y \cdot (1-\pi)^{1-y}$$

Una volta che il modello stima la probabilità di una classe, possiamo classificare un punto $\boldsymbol{\varphi}$ in una particolare classe se la probabilità per quella classe è **superiore a una soglia** (di solito 0.5).

La funzione che stiamo stimando è: $f(\boldsymbol{\varphi}) = P(y = 1|\boldsymbol{\varphi})$

La regressione logistica tenta di modellare f con: $s(\boldsymbol{\varphi}^T\boldsymbol{\theta}) = \frac{1}{1+e^{-\boldsymbol{\varphi}^T\boldsymbol{\theta}}}$

Il punto $\boldsymbol{\varphi}$ può quindi essere classificato alla classe $y = 1$ se:

$$s(\boldsymbol{\varphi}^{\boldsymbol{T}} \cdot \boldsymbol{\theta}) = \frac{1}{1+e^{-\boldsymbol{\varphi}^T\cdot\boldsymbol{\theta}}} \geqslant 0.5$$

Funzione di costo da minimizzare

La probabilità che $\boldsymbol{\varphi(i)} \in \mathbf{1}$, ossia $y(i) = 1$:

$$\pi(i) \equiv P\big(y(i) = 1|\boldsymbol{\varphi}(i)\big) = s(\boldsymbol{\varphi}(i)^T \cdot \boldsymbol{\theta}) = \frac{1}{1 + e^{-\boldsymbol{\varphi}(i)^T \cdot \boldsymbol{\theta}}}$$

Calcoliamo la **verosimiglianza**:

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

Calcoliamo la **funzione di costo**:

$$\begin{aligned}
-ln\big[\mathcal{L}(\pi|Y)\big] &= -ln\left[\prod_{i=1}^{N} \pi(i)^{y(i)} \cdot \big(1 - \pi(i)\big)^{1-y(i)}\right] \\
&= -\sum_{i=1}^{N} ln\left[\pi(i)^{y(i)} \cdot \big(1 - \pi(i)\big)^{1-y(i)}\right] \\
&= -\sum_{i=1}^{N} \left(ln\left[\pi(i)^{y(i)}\right] + ln\left[\big(1 - \pi(i)\big)^{1-y(i)}\right]\right) \\
&= -\sum_{i=1}^{N} \Big(y(i) \cdot ln\big[\pi(i)\big] + \big(1 - y(i)\big) \cdot ln\big[1 - \pi(i)\big]\Big) \equiv J(\boldsymbol{\theta})
\end{aligned}$$

Interpretazione della funzione di costo

Interpretazione della **funzione di costo**:
Assumiamo ci sia un solo dato $\mathcal{D} = \big\{\{\boldsymbol{\varphi}, y\}\big\}$

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

Caso $\boldsymbol{\varphi} \in y = 1$:

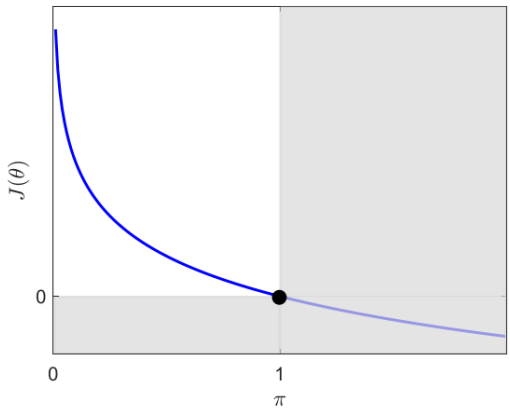
$$J(\boldsymbol{\theta}) = -ln\big[\pi\big]$$

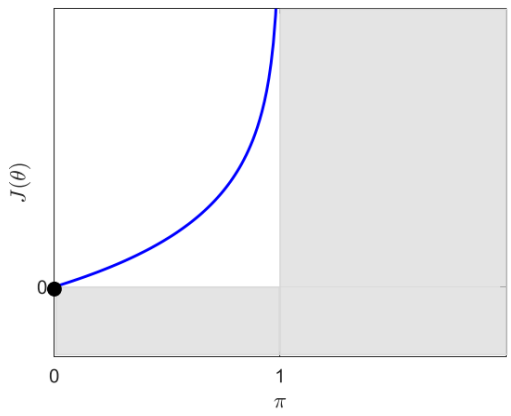
$$\begin{cases} J(\boldsymbol{\theta}) \approx 0, \ \pi \approx 1 \\ J(\boldsymbol{\theta}) \approx +\infty, \ \pi \approx 0 \end{cases}$$

Caso $\boldsymbol{\varphi} \in y = 0$:

$$J(\boldsymbol{\theta}) = -ln\big[1 - \pi\big]$$

$$\begin{cases} J(\boldsymbol{\theta}) \approx 0, \ \pi \approx 0 \\ J(\boldsymbol{\theta}) \approx +\infty, \ \pi \approx 1 \end{cases}$$





NB: π viene calcolato come la probabilitá che l'osservazione appartenza alla classe 1 ($\boldsymbol{\varphi(i)} \in y = 1$) infatti: $\pi = P(y(i) = 1|\boldsymbol{\varphi}(i))$. Allo stesso modo, la probabilitá che l'osservazione appartenza alla classe 2 ($\boldsymbol{\varphi(i)} \in y = 0$) é data da: $1 - \pi$.

Quali algoritmi si possono usare per la minimizzazione della funzione di costo

Per minimizzare la funzione di costo si può ad esempio utilizzare il metodo iterativo del calcolo del gradiente.

Supponendo di avere a disposizione N osservazioni *iid* distribuite come una distribuzione di Bernoulli:

$$p(y_i|\pi) \sim \pi^y \cdot (1-\pi)^{1-y}$$

- calcolare lo stimatore a massima verosimiglianza per π
- dimostrare che lo stimatore è corretto $\mathbb{E}\big[y_i\big] = \pi \ \forall i$
- dare un'interpretazione della formula ottenuta

Calcolare lo stimatore a massima verosimiglianza

Guardare passaggi completi sopra.

$$-ln\big[\mathcal{L}(\pi|Y)\big] = -\sum_{i=1}^{N}\Big(y(i)\cdot ln\big[\pi(i)\big] + \big(1-y(i)\big)\cdot ln\big[1-\pi(i)\big]\Big) \equiv J(\boldsymbol{\theta})$$

Dimostrazione correttezza

$$Pongo \quad Y = \sum_{i=1}^{N} y(i) \quad e \quad N - Y = \sum_{i=1}^{N}\big(1-y(i)\big)$$

$$J(\boldsymbol{\theta}) \equiv -\sum_{i=1}^{N}\Big(y(i)\cdot ln\big[\pi(i)\big] + \big(1-y(i)\big)\cdot ln\big[1-\pi(i)\big]\Big) =$$
$$= -Y\cdot ln(\pi) - ln(1-\pi)(N-Y)$$

$$\nabla_\pi J(\boldsymbol{\theta}) = 0$$

$$\nabla_\pi\big(-Y\cdot ln(\pi) - (N-Y)\cdot ln(1-\pi)\big) = 0$$
$$-\nabla_\pi\big(Y\cdot ln(\pi) + (N-Y)\cdot ln(1-\pi)\big) = 0$$
$$\nabla_\pi\big(Y\cdot ln(\pi) + (N-Y)\cdot ln(1-\pi)\big) = 0$$

$$\frac{Y}{\pi} - \frac{N-Y}{1-\pi} = 0$$
$$\frac{Y(1-\pi) - \pi(N-Y)}{\pi(1-\pi)} = 0$$
$$Y - \pi Y - \pi N + \pi Y = 0$$
$$Y - \pi N = 0$$

Lo stimatore risulta:

$$\hat{\pi} = \frac{Y}{N} = \frac{1}{N}\cdot\sum_{i=1}^{N} y(i)$$

Lo stimatore risulta essere corretto, poichè:

$$\mathbb{E}\big[\hat{\pi}\big] = \mathbb{E}\Big[\frac{1}{N}\cdot\sum_{i=1}^{N} y(i)\Big]$$

Interpretazione della funzione di costo

Guardare domande precedenti.

---

3 Lezione 6: Fondamenti di machine learning

Spiegare i concetti di:

- bias e varianza
- bias-variance tradeoff
- come capire se il modello soffre di uno o l'altro (learning curves)

bias e varianza

Insieme delle features (input):	$\boldsymbol{\varphi} \in \mathbb{R}^{d\times 1}$
Response variable (output categorico/scalare):	$y \in \mathbb{R}, y \in \{cat1, cat2, \ldots\}$
Funzione target (da stimare/ignota):	$f : \mathbb{R}^{d\times 1} \to \mathcal{Y}$
Ipotesi scelta (approssimazione di f):	$g : \mathbb{R}^{d\times 1} \to \mathcal{Y}$

Osservazioni:

- $\mathcal{M}$ è chiamato spazio delle ipotesi (o set dei modelli)
- $g \in \mathcal{M}$, é un'approssimazione di f
- Obiettivo é stimare f usando i dati $\mathcal{D}$
- La funzione f viene cercata, dall'algoritmo di learning, nello spazio delle ipotesi $\mathcal{M}$
- Vogliamo trovare una funzione $h \in \mathcal{M}$ che approssima bene f, non solo sui dati $\mathcal{D}$ a disposizione, ma sull'intero dominio $\mathbb{R}^{d\times 1}$ di f.
- $h \approx f \Rightarrow$ Dobbiamo definire una misura di errore o di costo. **Misure di errore puntuali**, definite su un singolo punto $\boldsymbol{\varphi}$. **Misure di errore globale,** definite considerando tutte le N osservazioni.

Misure di errore puntuali:

Errore quadratico:	$l\big(f(\boldsymbol{\varphi}), h(\boldsymbol{\varphi};\boldsymbol{\theta})\big) = \big(f(\boldsymbol{\varphi}) - h(\boldsymbol{\varphi};\boldsymbol{\theta})\big)^2$	*usata per regressione*
Errore binario:	$l(f(\boldsymbol{\varphi}), h(\boldsymbol{\varphi};\boldsymbol{\theta})) = \mathbb{I}\{f(\boldsymbol{\varphi}) \neq h(\boldsymbol{\varphi};\boldsymbol{\theta})\}$	*usata per classificazione*

Misure di errore globali:

Errore in-sample (errore di train):	$E_{in}(h(\boldsymbol{\theta})) \equiv J(\boldsymbol{\theta}) = \frac{1}{N} \cdot \sum_{i=1}^{N} l(f(\boldsymbol{\varphi}), h(\boldsymbol{\varphi};\boldsymbol{\theta}))$
Errore out-of-sample (errore di validazione):	$E_{out}(h(\boldsymbol{\theta})) = \mathbb{E}_{\boldsymbol{\varphi}}\big[l(f(\boldsymbol{\varphi}), h(\boldsymbol{\varphi};\boldsymbol{\theta}))\big]$

- Non è possibile conoscere con certezza come sarà il comportamento della funzione f su punti che non ho osservato (problema **dell'induzione di Hume**).
- 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.
- L'algoritmo di learning è usato per scandagliare lo spazio delle ipotesi $\mathcal{M}$, al fine di trovare la miglior $h \in \mathcal{M}$ che approssima bene i dati osservati → chiamiamo questa ipotesi g. Con tante ipotesi in $\mathcal{M}$, c'é un rischio maggiore di trovare una funzione g che spiega benissimo i dati misurati ma fa malissimo su dati nuovi
- Esiste quindi tradeoff tra **approssimazione** e **generalizzazione**. Si vuole: avere un buon modello sui dati misurati (training set) e avere un buon modello su dati non visti (e quindi non usati per la stima del modello)

Errore di generalizzazione:	$E_{out}(g) - E_{in}(g)$
Spazio delle ipotesi $\mathcal{M}$ piú complesso:	*Migliori possibilità di approssimare f in-sample*
Spazio delle ipotesi meno $\mathcal{M}$ complesso:	*Migliori possibilità di generalizzare f out-of-sample*

Un modo per studiare questo trade-off è valutare **i concetti di bias e varianza** di un modello di learning. L'approccio bias-varianza decompone E_{out} in:

1. Quanto bene $\mathcal{M}$ puó approssimare f → **bias**
2. Quanto bene riusciamo a scegliere una buona $h \in \mathcal{M}$, usando i dati → **varianza**

Sia $g = h \in \mathcal{M}$ la miglior ipotesi che approssima bene sui dati osservati e supponiamo di osservare i dati **senza rumore**, cioé che $y = f(\boldsymbol{\varphi})$. L'errore out-of-sample può essere espresso come (rendendo esplicita la dipendenza di g da $\mathcal{D}$)

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

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

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

Definiamo ***ipotesi media***: $\bar{g} = \mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(\boldsymbol{\varphi})]$
Ipotesi che deriva dall'usare K dataset $\mathcal{D}_1, .., \mathcal{D}_K$ e costruendola come

$$\bar{g}(\boldsymbol{\varphi}) \approx \frac{1}{K}\sum_{k=1}^{K} g^{\mathcal{D}_K}(\boldsymbol{\varphi})$$

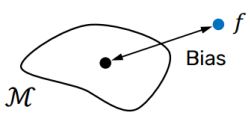
$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[\left(g^{\mathcal{D}}(\boldsymbol{\varphi}) - f(\boldsymbol{\varphi})\right)^2\right] &= \mathbb{E}_{\mathcal{D}}\left[\left(g^{\mathcal{D}}(\boldsymbol{\varphi}) - \bar{g}(\boldsymbol{\varphi}) + \bar{g}(\boldsymbol{\varphi}) - f(\boldsymbol{\varphi})\right)^2\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\left(g^{\mathcal{D}}(\boldsymbol{\varphi}) - \bar{g}(\boldsymbol{\varphi})\right)^2 + \left(\bar{g}(\boldsymbol{\varphi}) - f(\boldsymbol{\varphi})\right)^2 + 2\cdot\left(g^{\mathcal{D}}(\boldsymbol{\varphi}) - \bar{g}(\boldsymbol{\varphi})\right)\cdot\left(\bar{g}(\boldsymbol{\varphi}) - f(\boldsymbol{\varphi})\right)\right] \\
&= \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(g^{\mathcal{D}}(\boldsymbol{\varphi}) - \bar{g}(\boldsymbol{\varphi})\right)^2\right]}_{Var(\boldsymbol{\varphi})} + \underbrace{\left(\bar{g}(\boldsymbol{\varphi}) - f(\boldsymbol{\varphi})\right)^2}_{bias^2(\boldsymbol{\varphi})}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[E_{out}(g^{\mathcal{D}})\right] &= \mathbb{E}_{\boldsymbol{\varphi}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{\mathcal{D}}(\boldsymbol{\varphi}) - f(\boldsymbol{\varphi})\right)^2\right]\right] \\
&= \mathbb{E}_{\boldsymbol{\varphi}}\left[bias^2(\boldsymbol{\varphi}) + var(\boldsymbol{\varphi})\right] \\
&= bias^2 + var
\end{aligned}$$

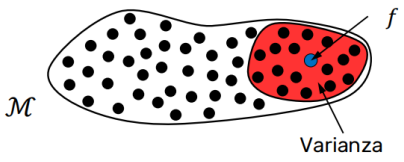
Interpretazione:

- Il termine $bias^2(\boldsymbol{\varphi}) = \left(\bar{g}(\boldsymbol{\varphi}) - f(\boldsymbol{\varphi})\right)^2$ misura quanto il nostro modello (cioè la nostra funzione stimata g) é lontano dall funzione target f. Infatti $\bar{g}$ ha il vantaggio di apprendere da un numero illimitato di datasets. Quindi $\bar{g}$, nella capacitá di approssimare f, é limitata solo dai limiti di $\mathcal{M}$.
- $var(\boldsymbol{\varphi}) = \mathbb{E}_{\mathcal{D}}\left[\left(g^{\mathcal{D}}(\boldsymbol{\varphi}) - \bar{g}(\boldsymbol{\varphi})\right)^2\right]$ misura quanto $g^{\mathcal{D}}$ si disperde da $\bar{g}$ e si puó vedere quanto l'ipotesi finale $g^{\mathcal{D}}$ differisca dall'ipotesi migliore ovvero quella media $\bar{g}$.

$\textbf{bias}^2 = \left(\bar{g}(\boldsymbol{\varphi}) - f(\boldsymbol{\varphi})\right)^2$



$\textbf{varianza} = \mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\boldsymbol{\varphi}) - \bar{g}(\boldsymbol{\varphi})\right)^2\right]$



Anziché osservare $y = f(\boldsymbol{\varphi})$, osserviamo $y = f(\boldsymbol{\varphi}) + \eta(\boldsymbol{\varphi})$ dove η é un rumore stocastico con media zero e varianza σ^2

$$\mathbb{E}_{\mathcal{D},\boldsymbol{\varphi},\eta}\left[\left(g^{\mathcal{D}}(\boldsymbol{\varphi}) - \big(f(\boldsymbol{\varphi}) + \eta(\boldsymbol{\varphi})\big)\right)^2\right] = bias^2 + var + \sigma^2$$

L'errore stocastico σ^2 non puó essere portato a zero e contribuisce alla varianza dell'ipotesi scelta, causando overfitting → **errore irriducibile**.

Bias-variance tradeoff

Il trade-off bias-varianza indica che ridurre il bias di un modello aumenta la sua varianza e viceversa. In altre parole, esiste una relazione inversa tra la capacità del modello di adattarsi ai dati di addestramento (approssimazione) e la sua capacità di generalizzare su nuovi dati. L'obiettivo è trovare il giusto equilibrio tra bias e varianza per ottenere un modello con una buona capacità di generalizzazione.

I principali fattori che influenzano questo aspetto sono:

- **Dimensione del set di addestramento:** un set di addestramento più grande risulta essere molto informativo e tende a ridurre la varianza. Al contrario, con un set di addestramento più piccolo, il modello può soffrire di una maggiore varianza, poiché è meno probabile che rappresenti l'intera distribuzione dei dati.
- **Complessità del modello:** un modello più complesso ha una maggiore capacità di adattarsi ai dati di addestramento, riducendo il bias. Tuttavia, un modello più complesso tende anche ad avere una maggiore varianza, poiché può adattarsi anche al rumore o alle peculiarità casuali dei dati.
- **Rumore nei dati:** la presenza di rumore nei dati può aumentare la varianza del modello. Se i dati contengono rumore o errori casuali, il modello può adattarsi anche a queste irregolarità, introducendo un'elevata varianza. Ridurre il rumore nei dati può ridurre la varianza del modello.
- **Selezione delle feature:** l'utilizzo di feature rilevanti e informative può aiutare a ridurre il bias, ma l'uso di feature non informative o irrilevanti può aumentare la varianza.
- **Tecniche di validazione:** aiutano a valutare il trade-off bias-varianza.

Un modello con:

- **alto bias e bassa varianza:** il modello ha una capacità limitata di adattarsi ai dati di addestramento ed è troppo semplice per catturare la complessità del problema → non é in grado di catturare i pattern nei dati
- **basso bias e alta varianza:** il modello ha una capacità elevata di adattarsi ai dati di addestramento ed è molto complesso, possibile overfitting con cattiva generalizzazione.
- **bias adeguato e varianza adeguata:** il modello ha una buona capacità di adattarsi ai dati di addestramento (approssimazione) e di generalizzare su nuovi dati.

modelli con basso bias sono più complessi → tendono ad overfittare e a inglobare errore
modelli con bassa varianza sono meno complessi → potrebbero non cogliere la relazione dei dati

Modello semplice:	**Modello complesso:**
Bias ↑	*Bias* ↓
Varianza ↓	*Varianza* ↑
Approssimazione e generalizzazione cattive	*Approssimazione positiva generalizzazione negativa*
All'aumentare di N rimane uguale	All'aumentare di N migliora

Il **trade-off tra approssimazione e generalizzazione** si traduce nel valutare i concetti di bias e varianza di un modello di learning. L'approccio bias-varianza decompone E_{out} in:

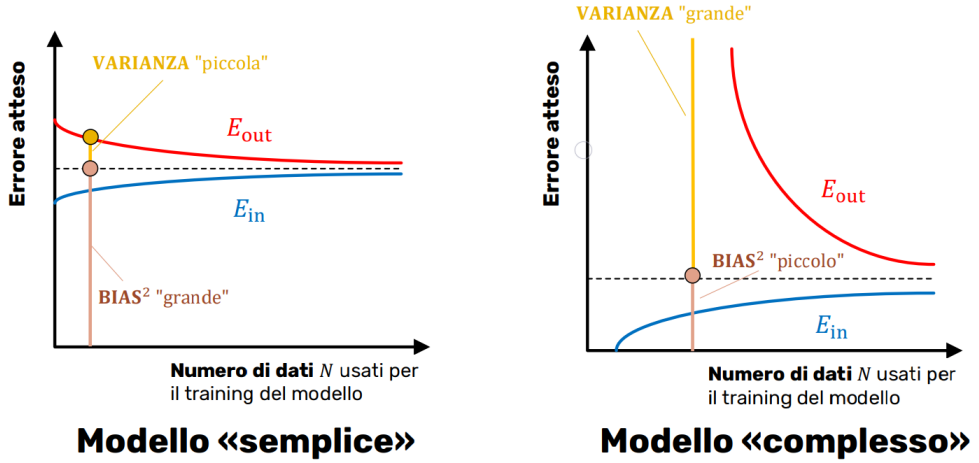
1. Quanto bene $\mathcal{M}$ (spazio delle ipotesi) puó approssimare f (funzione target) → **bias**
2. Quanto bene riusciamo a scegliere una buona $h \in \mathcal{M}$, usando i dati → **varianza**

Learning curves

Le learning curves sono uno strumento grafico per capire se un modello di learning soffre di **problemi di bias o varianza**. L'idea è di rappresentare, al variare del numero di dati N usati per stimare il modello:

- l'errore out-of-sample atteso → $\mathbb{E}_{\mathcal{D}}\left[E_{out}(g^{\mathcal{D}})\right]$
- l'errore in-sample atteso → $\mathbb{E}_{\mathcal{D}}\left[E_{in}(g^{\mathcal{D}})\right]$

Le curve vengono calcolate usando un solo dataset, oppure dividendolo in più parti e prendendo la **curva media** risultante dai vari sub-datasets.



- Il bias può essere presente quando l'errore atteso è piuttosto elevato e E_{in} in è simile a E_{out}, è improbabile che ottenere più dati aiuti.

  $\Rightarrow$ *Aggiungere features (per esempio combinazioni di features originarie) o boosting*

- La varianza può essere presente quando c'è un tanto divario tra E_{in} e E_{out}, è probabile che ottenere più dati sia d'aiuto

  $\Rightarrow$ *Usare meno features, acquisire più dati,* ***usare regolarizzazione****, bagging*

Discutere il metodo delle learning curves specificando:

- il motivo del loro utilizzo
- la tipologia di problemi diagnosticabili con esse
- dei possibili rimedi per questi problemi

Guardare la domanda precedente.

Overfitting, da cosa è causato, come si può risolvere e tipologie di regolarizzazione

Overfitting e cause di overfitting

L'overfitting è un problema comune nel machine learning ed è causato da un modello che si adatta troppo ai dati di addestramento, perdendo la capacità di generalizzare correttamente su nuovi dati.

- Un modello che fa overfitting presenta:

  $$Bias = basso \qquad E_{in} = bassso$$

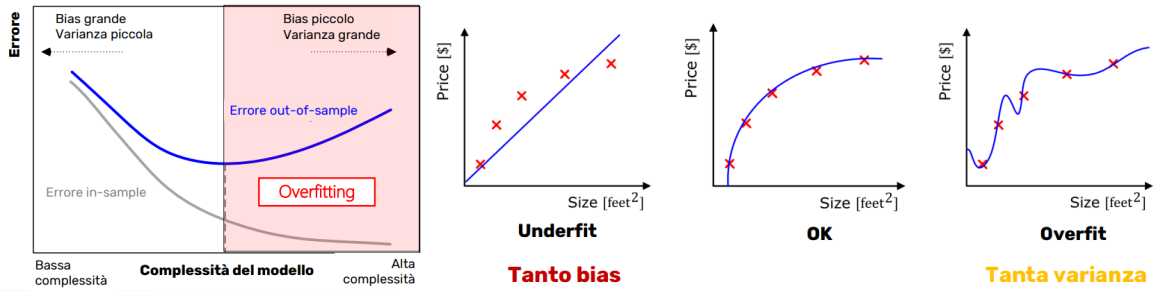
  $$Var = alta \qquad E_{out} = alto$$

  Solitamente questo comportamento é tipico dei modelli complessi che hanno un potere espressivo molto alto e tendono ad adattarsi al rumore. Nel caso di pochi dati é necessario usare modelli più semplici indipendentemente dalla complessità della funzione target.

- Una seconda causa di overfitting è il rumore stocastico η (supponiamo media 0 e var σ^2) che affligge le misure. Al posto di osservare $y = f(\boldsymbol{\varphi})$ osservo $y = f(\boldsymbol{\varphi}) + \eta(\boldsymbol{\varphi})$, dunque:

$$\mathbb{E}_{\mathcal{D},\boldsymbol{\varphi},\eta}\left[\left(g^{\mathcal{D}}(\boldsymbol{\varphi}) - \big(f(\boldsymbol{\varphi}) - \eta(\boldsymbol{\varphi})\big)\right)^2\right] = bias^2 + var + \sigma^2$$

  L'errore stocastico σ^2 non puó essere portate a zero e contribuisce alla varianza dell'ipotesi scelta, causando overfitting → **errore irriducibile**



Regolarizzazione

I modelli più complessi sono più inclini all'overfitting, questo perché sono molto espressivi e quindi possono adattarsi anche al rumore. I modelli semplici mostrano meno varianza a causa della loro espressività limitata. La riduzione della varianza del modello è spesso maggiore dell'aumento del suo bias, per cui, nel complesso, errore atteso complessivo diminuisce ($bias^2 + var + \sigma^2$).
Tuttavia, se ci atteniamo solo a modelli semplici, potremmo non ottenere un'approssimazione soddisfacente della funzione target f.

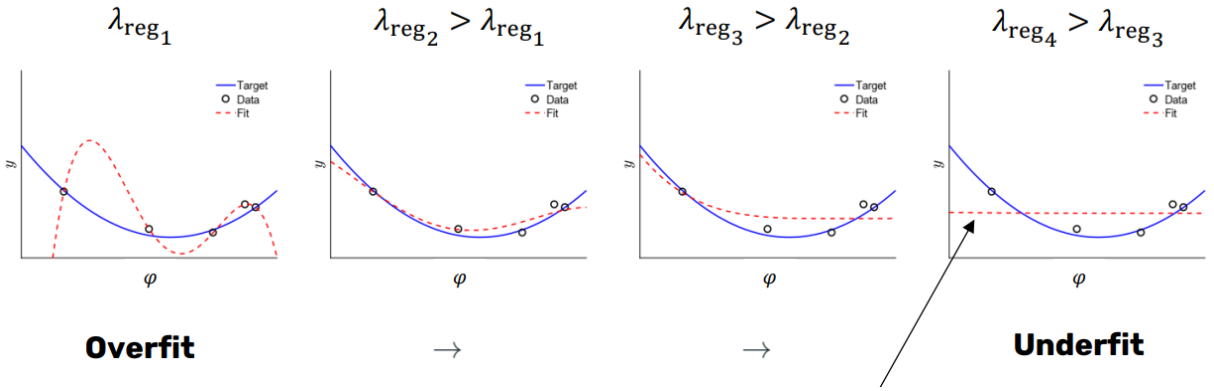
La regolarizzazione oltre che minimazzzare la funzione di costo $E_{in} \equiv J(\boldsymbol{\theta})$ minizza anche la complessitá del modello. Al posto di E_{in} minimizziamo dunque un **errore aumentato** $E_{aug}(\boldsymbol{\theta})$

$$E_{aug}(\boldsymbol{\theta}) = \underbrace{\frac{1}{N} \cdot \sum_{i=1}^{N} \big(y(i) - h(\boldsymbol{\varphi}(i);\boldsymbol{\theta})\big)^2}_{Quanto\ male\ il\ modello\ fitta\ i\ dati} + \lambda_{reg} \cdot \Omega(\boldsymbol{\theta})$$

$\Omega(\boldsymbol{\theta})$ → *Regolarizzatore: quanto il modello è complesso*

$h(\cdot)$ *è qualche funzione che rappresenta il nostro modello*

Il termine λ_{reg} (iper-parametro) **pesa l'importanza** di minimizzare $E_{in} \equiv J(\boldsymbol{\theta})$ rispetto a minimizzare $\Omega(\boldsymbol{\theta})$, ossia la complessitá del modello.



Se regolarizzo troppo, imparerò la funzione più semplice possibile, ovvero una retta orizzontale (costante) con intercetta θ_0.
Minimizzare E_{aug} rispetto ad E_{in} conduce ad un modello migliore (ovvero un modello con miglior capacità di generalizzare e quindi con E_{out} minore).

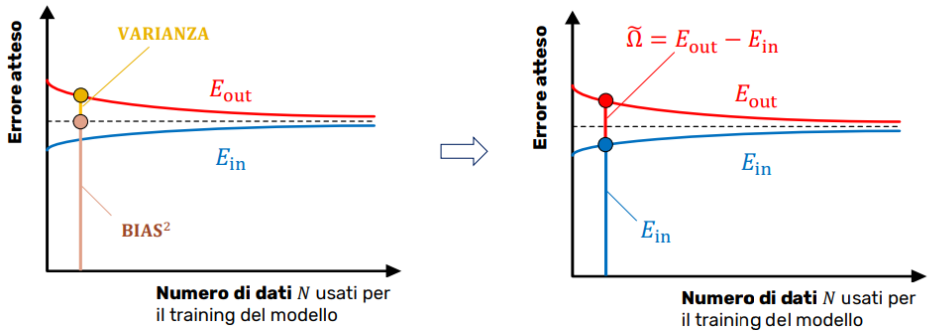
Possiamo interpretare E_{out} *come la somma di due contributi:*

$$E_{out}(\boldsymbol{\theta}) = E_{in}(\boldsymbol{\theta}) + \widetilde{\Omega}(\boldsymbol{\theta})$$

Ricordando la definizione di E_{aug} *abbiamo:*

$$E_{aug}(\boldsymbol{\theta}) = E_{in}(\boldsymbol{\theta}) + \lambda_{reg} \cdot \Omega(\boldsymbol{\theta})$$

$\Rightarrow$ *E_{aug} é migliore rispetto ad E_{in} come proxy per E_{out}*



La regolarizzazione aiuta nello stimare la quantità $\Omega(\boldsymbol{\theta})$, che sommata ad E_{in} fornisce E_{aug}, il quale é una stima di E_{out}. Esistono diversi tipi di regolarizzazione. I più usati sono:

Regolarizzazione L_1 $\rightarrow$ ***Lasso***

Tende a portare più coefficienti esattamente a zero

$$\Omega(\boldsymbol{\theta}) = \sum_{j=0}^{d-1} |\theta_j|$$

Regolarizzazione L_2 $\rightarrow$ ***Ridge***

Tende a ridurre tutti i coefficienti a un valore inferiore

$$\Omega(\boldsymbol{\theta}) = \sum_{j=0}^{d-1} (\theta_j)^2$$

Regolarizzazione elastic-net:

Combinazione di L_1 e L_2

$$\Omega(\boldsymbol{\theta}) = \beta \sum_{j=0}^{d-1} (\theta_j)^2 + (1-\beta) \sum_{j=0}^{d-1} |\theta_j|$$

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

- La regolarizzazione **aumenta di poco il bias** (perché ottengo un modello piú semplice) al fine di **ridurre considerevolmente la varianza** del modello di learning
- La regolarizzazione porta ad avere **ipotesi più smooth**, regolari, riducendo il rischio di overfitting
- L'iperparametro di regolarizzazione λ_{reg} deve essere scelto in modo specifico per ogni tipo di regolarizzatore. Solitamente si usa una procedure come la validazione o la cross-validazione

---

Cross-validazione

Validazione

$$E_{out}(\boldsymbol{\theta}) = E_{in}(\boldsymbol{\theta}) + \textit{ penalità per la complessità del modello}$$

La REGOLARIZZAZIONE stima questa quantità
La VALIDAZIONE stima questa quantità

L'idea delle procedure di validazione è quella di stimare E_{out} utilizzando un dataset diverso (validation set) rispetto a quello usato per la stima del modello (training\identification set). La regolarizzazione e la validazione sono due tecniche che possono (e devono) essere usate insieme:

- la regolarizzazione aiuta a stimare un modello che può generalizzare meglio
- la validazione fornisce una stima dell'errore out-of-sample del modello

Una procedura comune che si segue è:

1. **Rimuovo** un subset di dati dai dati totali → $\mathcal{D}_{val}$
2. **Stimo** il modello sulla parte di dati rimanente → $\mathcal{D}_{train}$
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

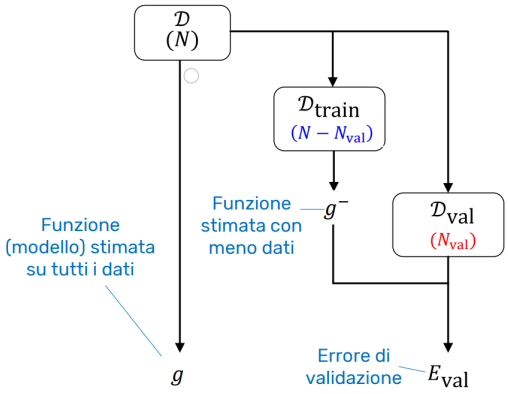
$$\begin{array}{ccccc} \mathcal{D} & \rightarrow & \mathcal{D}_{\text{train}} & \cup & \mathcal{D}_{\text{val}} \\ \downarrow & & \downarrow & & \downarrow \\ N_{\text{val}} & & N - N_{\text{val}} & & N_{\text{val}} \end{array}$$

$$\mathcal{D} \Rightarrow g \qquad \mathcal{D}_{\text{train}} \Rightarrow g^-$$

$$E_{\text{val}} = E_{\text{val}}(g^-)$$

Rule of thumb

$$N_{\text{val}} = \frac{N}{5}$$



Le procedure di validazione possono essere utilizzate per due scopi:

- Valutare le performance del modello stimato → stimare E_{out}
- Scegliere il modello migliore da un insieme di diversi modelli

Problema: se uso il dataset di validazione $\mathcal{D}_{val}$ tante volte per compiere delle scelte, allora il dataset di validazione $\mathcal{D}_{val}$ non fornisce piú una buona stima dell'errore out-of-sample E_{out}.

Intuizione: usare $\mathcal{D}_{val}$ per compiere delle scelte su quale modello usare fa si che tali scelte siano dipendenti dai particolari valori dei dati contenuti in $\mathcal{D}_{val}$. Chi mi garantisce che con dati diversi avrei compiuto le medesime scelte? *Stiamo dunque* ***overfittando il validation set.***

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

Contaminazione: si riferisce all'effetto negativo che può verificarsi quando i dati in un determinato set (ad esempio, il training set o il validation set) influenzano eccessivamente le decisioni prese durante lo sviluppo di un modello. La ***contaminazione*** nei tre set di dati (training, validation e test) si riferisce al livello di presenza di dati sovrapposti o influenze reciproche tra i set. Quando dici che un set è "contaminato", significa che contiene dati che sono stati usati in modo improprio in altre fasi dello sviluppo, potenzialmente portando a una stima ottimistica delle prestazioni del modello.

- **Training set (60% dei dati):** totalmente contaminato
- **Validation set (20% dei dati):** un pò contaminato
- **Test set (20% dei dati):** totalmente pulito

Cross-validazione

La divisione del dataset in tre parti (train, validation, test) è fattibile se i dati a disposizione sono molti.

in teoria vorremmo che:

$$\underbrace{E_{out}(g) \approx E_{out}(\bar{g})}_{N_{val}\ piccolo} \approx E_{val}(\bar{g})$$

$$E_{out}(g) \approx \underbrace{E_{out}(\bar{g}) \approx E_{val}(\bar{g})}_{N_{val}\ grande}$$

- $E_{val}(\bar{g})$ é l'unico che posso calcolare
- N_{val} **grande:** in questo caso $E_{val}(\bar{g}) \approx E_{out}(\bar{g})$ poichè uso tanti dati N_{val} per la validazione. Ricordiamoci che l'obiettivo di E_{val} è proprio quello di stimare E_{out}
- N_{val} **piccolo**: in questo caso $E_{out}(\bar{g}) \approx E_{out}(g)$, poiché uso tanti dati $N - N_{val}$ per il train di $\bar{g}$. Questo è il valore che mi interessa ma che non posso calcolare direttamente

La cross-validazione permette di avere N_{val} sia grande che piccolo.

Leave-one-out cross-validation

- Usiamo $N-1$ per il training e $N_{val}=1$ dato per la validazione, dove $\mathcal{D}_i$ é il dataset di training senza il dato i-esimo:

$$\mathcal{D}_i = \{(\boldsymbol{\varphi}(1), y(1)), \ldots, \cancel{(\boldsymbol{\varphi}(i), y(i))}, \ldots, (\boldsymbol{\varphi}(N), y(N)\}$$

$$\textbf{\textit{La fuzione imparata usando }} \mathcal{D}_i \textbf{\textit{ é}} \rightarrow g_i^-$$

- L'errore di validazione sul **singolo punto rimosso** $\boldsymbol{\varphi}(i)$ (usando l'errore puntuale) risulta essere:

$$l(i) = E_{val}(g_i^-) = l\Big(y(i), g_i^-\big(\boldsymbol{\varphi}(i)\big)\Big)$$

- E' possibile definire **l'errore di cross-validazione** E_{CV} come:

$$E_{CV} = \frac{1}{N}\sum_{i=1}^{N} l(i)$$

- Stimo N modelli usando $N-1$ dati, e li valido usando N stime dell'errore di validazione (con $N_{val}=1$). La cross-validazione $N_{val}=1$ **(leave-one-out cross-validation) risulta dunque essere computazionalmente costosa.** Nel caso volessimo usarla per scegliere tra M modelli, richiederebbe un totale di N sessioni di training per ciascuno degli M modelli
- La stima dell'errore E_{CV} ha una **varianza elevata**, poiché si basa su un solo dato

Cross-validazione K-fold

- E' possibile riservare più punti per la validazione suddividendo il training set in **folds**. Per esempio, se $k=10$ avremmo $\mathcal{D}_1, \ldots, \mathcal{D}_{10}$ datasets, di cui ad ogni iterazione utilizzo 9 datasets per il training e 1 per la validazione.
- La K-fold cross-validation richiede $\frac{N}{N_{val}}$ sessioni di training, ognuna con $N - N_{val}$ dati. Un buon compromesso è usare $K=10$ dove $N_{val} = \frac{N}{10}$
- Attenzione a non ridurre troppo il training set (guardare le learning curves)

Formule di complessità ottima

- Queste formule permettono di stimare l'errore out-of-sample E_{out} **utilizzando solo il dataset di train**. Per questo motivo, **si usano quando ho pochi dati** per poter usare validazione o cross-validazione.
- 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**.
- Indichiamo la **stima dei parametri**, ottenuta con N dati, con $\hat{\boldsymbol{\theta}}_N \in \mathbb{R}^{d\times 1}$.
- La stima è ottenuta mimizzando la funzione di costo $J(\boldsymbol{\theta}; d)$, dove esplicitiamo la dipendenza del costo dal numero di parametri $d \Rightarrow AIC, BIC \equiv J(\theta)_{aug}$ minori sono i migliori

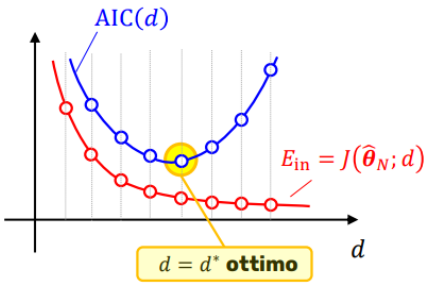
Akaike Information Criterion (AIC) ≡ FPE

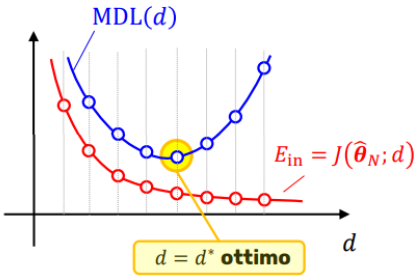
$$AIC(d) = 2 \cdot \frac{d}{N} + ln\big[J(\hat{\boldsymbol{\theta}}_N; d)\big]$$

Minimum Description Length (MDL) *derivante dal* BIC

$$MDL(d) = ln\big[N\big] \cdot \frac{d}{N} + ln\big[J(\hat{\boldsymbol{\theta}}_N; d)\big]$$

Se $ln[N] > 2$ (ovvero se abbiamo più di 8 dati), MDL suggerisce di usare modelli più parsimoniosi.





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

NB: Sotto l'assunzione che il meccanismo di generazione dei dati appartenga alla classe di modelli scelta, ***FPE e AIC*** hanno una probabilità non nulla di ***sovrastimare l'ordine del modello***, mentre ***MDL*** porta ad una ***stima asintoticamente corretta dell'ordine***. Dato che raramente l'assunzione è verificata, si preferisce usare ***AIC o FPE***, sovrastimando leggermente d.

Data la funzione di costo definita con termini vettoriali e matriciali e regolarizzata con un Ridge Regularizer dimostrare che la minimizzazione della funzione è quella che si ottiene usando i minimi quadrati ma con il regolarizzatore in più da considerare

Formule:

dove x é un vettore $1 \times d$, A/B é una matrice $d \times d$ e I é identitá

1	$\big\|\|A\|\big\|_2^2 = A^T A$
2	$\big(AB\big)^T = B^T A^T$
3	$\nabla_x \big(x^T A x\big) = \big(A + A^T\big) \cdot x$

Consideriamo il modello di regressione lineare con un termine di regolarizzazione L_2:

$$\begin{aligned}
E_{aug}(\boldsymbol{\theta}) \equiv J(\boldsymbol{\theta}) &= \frac{1}{N} \cdot \sum_{i=1}^{N} \big(y(i) - \boldsymbol{\varphi}^T(i)\boldsymbol{\theta}\big)^2 + \lambda_{reg} \sum_{j=0}^{d-1} (\theta_j)^2 \\
&= \frac{1}{N} \cdot \big\|\|Y - X \cdot \boldsymbol{\theta}\big\|\|_2^2 + \lambda_{reg} \cdot \big\|\|\boldsymbol{\theta}\big\|\|_2^2 \\
&\underset{1}{=} \frac{1}{N} \cdot \big(Y - X\boldsymbol{\theta}\big)^T \big(Y - X\boldsymbol{\theta}\big) + \lambda_{reg} \boldsymbol{\theta}^T \boldsymbol{\theta} \\
&\underset{2}{=} \frac{1}{N} \cdot \big(Y^T Y - Y^T X\boldsymbol{\theta} - \boldsymbol{\theta}^T X^T Y + \boldsymbol{\theta}^T X^T X\boldsymbol{\theta}\big) + \lambda_{reg} \boldsymbol{\theta}^T \boldsymbol{\theta} \\
&= \frac{1}{N} \cdot \big(Y^T Y - 2\boldsymbol{\theta}^T X^T Y + \boldsymbol{\theta}^T X^T X\boldsymbol{\theta}\big) + \lambda_{reg} \boldsymbol{\theta}^T \boldsymbol{\theta}
\end{aligned}$$

Il minimo risulta essere:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = 0 \Rightarrow \frac{1}{N} \cdot \big(-2X^T Y + 2X^T X\boldsymbol{\theta}\big) + 2\lambda_{reg}\boldsymbol{\theta} = 0 \\
\frac{1}{N} \cdot \big(-X^T Y + X^T X\boldsymbol{\theta}\big) + \lambda_{reg}\boldsymbol{\theta} = 0 \\
-X^T Y + X^T X\boldsymbol{\theta} + N\lambda_{reg} I_d \boldsymbol{\theta} = 0 \\
\big(X^T X + \lambda_{reg} I_d\big)\boldsymbol{\theta} = X^T Y
\end{aligned}$$

Risulta dunque:

$$\underset{d\times 1}{\hat{\boldsymbol{\theta}}_{reg}} = \big(\underset{d\times d}{X^T X} + \lambda_{reg} \underset{d\times d}{I_d}\big)^{-1} X^T Y$$

---

4 Lezione 7: Fondamenti di stima Bayesiana

Probabilità congiunta, marginale e condizionata. Stima bayesiana. Relazione stimatore lineare ottimo e caso in cui la distribuzione congiunta tra due variabili è gaussiana.

Supponiamo di avere due variabili casuali discrete e binarie a e b. Definiamo:

Distribuzione di probabilità congiunta:

$P(a, b)$ probabilità che sia a che b assumino un valore specifico.

La sommatoria di tutte le combinazioni specifiche di valori deve essere uguale a 1

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

$$P(a, b) = P(b, a)$$

Specifici valori:

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

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

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

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

Distribuzione di probabilità marginale:

La distribuzione marginale è la distribuzione di probabilità di un **sottoinsieme di variabili casuali**. Siccome abbiamo 2 variabili casuale (a e b), avremo due marginali $P(a)$ e $P(b)$. Se avessimo 3 v.c discrete a, b, c avremmo le marginali $P(a)$, $P(b)$, $P(c)$, $P(a, b)$, $P(a, c)$, $P(b, c)$.

$$n \text{ } var, \text{ } totale = n!$$

La distribuzione marginale è ottenuta marginando (**sommando**) rispetto alle **variabili che non sono di interesse**. Nel caso di **v.c. continue**, si deve **integrare** anziché sommare.

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

Distribuzione di probabilità condizionata:

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

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

$P(A, B) = P(A) \cdot P(B)$ se e solo se $P(A|B) = P(A)$, ossia A e B sono eventi **indipendeti**, ovvero il verificarsi di B non modifica le probabilità di verificarsi di A.

Temora di Bayes

$$P(A, B) = P(B, A)$$
$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$
$$Allora$$

$$Teorema \text{ } di \text{ } bayes$$
$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

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

Stima Bayesiana

- Abbiamo finora considerato il vettore di parametri ignoto $\boldsymbol{\theta} \in \mathbb{R}^{d\times 1}$ come una **variabile deterministica**. Spesso però, ancora prima di collezionare i dati, abbiamo delle informazioni (o supposizioni) sui possibili valori che potrebbe assumere $\boldsymbol{\theta}$.
- Ha quindi senso considerare $\boldsymbol{\theta}$ come una **variabile casuale**: in questo modo, posso specificare una distribuzione di probabilità per $\boldsymbol{\theta}$, **per descriverne i valori** che io credo che possa assumere e la probabilitá che $\boldsymbol{\theta}$ li assuma. Assegno **maggior probabilità** ai valori che **io credo** siano più probabili che $\boldsymbol{\theta}$ **possa assumere** e minor probabilità ai valori che io credo non possa assumere.
- La distribuzione a-priori $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ sui possibili valori di $\boldsymbol{\theta}$, ha dominio $[0, 1]$ poiché θ modellando una probabilitá deve stare tra 0 e 1. Data $f_\theta(\theta)$, abbiamo già una stima del valore di θ **ancora prima di aver osservato i dati (STIMA A PRIORI)**. **L'incertezza sulla stima** sarà allora la varianza di 0 **(INCERTEZZA A PRIORI)**.

Abbiamo quindi due elementi che portano informazione:

- La distribuzione a-priori $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ sui possibili valori di $\boldsymbol{\theta}$
- L'informazione che portano i dati sui possibili valori di $\boldsymbol{\theta}$, ovvero la likehood $f_{Y|\boldsymbol{\theta}}(Y|\boldsymbol{\theta})$

NB: Quello che veramente ci interessa è sapere q**uanto può valere $\boldsymbol{\theta}$ dato che ho osservato i dati**, ovvero la distribuzione $f_{\boldsymbol{\theta}|Y}(\boldsymbol{\theta}|Y)$

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

$$f_{\boldsymbol{\theta}|Y}(\boldsymbol{\theta}|Y) = \frac{f_{Y|\boldsymbol{\theta}}(Y|\boldsymbol{\theta}) \cdot f_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{f_Y(Y)}$$

LIKELIHOOD PRIOR — POSTERIOR — MARGINAL LIKELIHOOD

$f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$	*Distribuzione a-priori sui possibili valori di* $\boldsymbol{\theta}$
$f_Y(Y)$	*Distribuzione dei dati* Y
$f_{Y\|\boldsymbol{\theta}}(Y\|\boldsymbol{\theta})$	*Distribuzione dei dati* Y *dato che ho supposto a priori* $\boldsymbol{\theta}$
$f_{\boldsymbol{\theta}\|Y}(\boldsymbol{\theta}\|Y)$	*Distribuzione di* $\boldsymbol{\theta}$ *dato che ho osservato i dati* Y

Osservazioni:

- $f_{\boldsymbol{\theta}|Y}(\boldsymbol{\theta}|Y)$ **è una distribuzione a-posteriori di possibili valori di $\boldsymbol{\theta}$**. Le probabilità di questi valori, rispetto a $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ sono state riallocate dall'aver osservato i dati Y (ecco perché a posteriori)
- Nel caso in cui $f_{Y|\boldsymbol{\theta}}(Y|\boldsymbol{\theta})$ e $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ sono pdf continue allora:

$$f_Y(Y) = \int_{-\infty}^{+\infty} f_{Y|\boldsymbol{\theta}}(Y|\boldsymbol{\theta}) \cdot f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{-\infty}^{+\infty} f(Y, \boldsymbol{\theta}) d\boldsymbol{\theta} \quad \textit{altrimenti sommatoria}$$

- 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 possiamo utilizzare **Markov Chain Monte Carlo (MCMC)**
- In generale **non posso dire nulla sulla posterior**, solo in casi fortunati ha un'espressione analitica nota. Caso fortunato se $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ = ***gaussiana*** e anche $f_{Y|\boldsymbol{\theta}}(Y|\boldsymbol{\theta})$ = ***gaussiana*** allora anche $f_{\boldsymbol{\theta}|Y}(\boldsymbol{\theta}|Y)$ = ***gaussiana***
- Un modo (computazionalmente oneroso ma semplice) per calcolare la posterior $f_{\boldsymbol{\theta}|Y}(\boldsymbol{\theta}|Y)$ è quello di **discretizzare** il range di valori del parametro $\boldsymbol{\theta}$ tramite una griglia di valori.

Stima ottima

Indichiamo uno stimatore come una funzione $T(\ )$ dei dati $\mathcal{D}$:

$$\hat{\boldsymbol{\theta}} = T(\mathcal{D})$$

Consideriamo il caso 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)**

$$MSE \equiv \mathbb{E}\big[\big(\hat{\theta} - \theta\big)^2\big] = \mathbb{E}\big[\big(T(\mathcal{D}) - \theta\big)^2\big]$$

Lo **stimatore ottimo di Bayes** é quella funzione $T^{opt}(\ )$ che **minimizza il MSE**

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

Si dimostra che **lo stimatore che minimizza il MSE è il valore atteso condizionato** (al fatto che i dati $\mathcal{D}$ abbiano assunto i valori in Y)

$$\hat{\boldsymbol{\theta}} = T^{opt}(Y) = \mathbb{E}\big[\boldsymbol{\theta} | \mathcal{D} = Y\big]$$

Nel caso in cui $\boldsymbol{\theta}$ si un **vettore di parametri**, il calcolo del MSE si modifica come segue:

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

Stima ottima : il caso Gaussiano

Supponiamo ora di avere un dato interpretato come realizzazione di una variabile casuale Guassiana $y \sim \mathcal{N}(0, \lambda_{yy}^2)$, e che anche il parametro ignoto (scalare per comodità) sia Gaussiano $\theta \sim \mathcal{N}(0, \lambda_{\theta\theta}^2)$

$$\underbrace{\underset{2\times 1}{\begin{bmatrix} y \\ \theta \end{bmatrix}}}_{\boldsymbol{z}} \sim \mathcal{N}\left(\underbrace{\underset{2\times 1}{\begin{bmatrix} 0 \\ 0 \end{bmatrix}}}_{\boldsymbol{\mu}}, \underbrace{\underset{2\times 2}{\begin{bmatrix} \lambda_{yy}^2 & \lambda_{y\theta} \\ \lambda_{\theta y} & \lambda_{\theta\theta}^2 \end{bmatrix}}}_{\Sigma}\right)$$

Congiunta è gaussiana: $f_{y\theta}(y, \theta)$

$$f_{y\theta}(y,\theta) = \frac{1}{\sqrt{(2\pi)^2 \det \Sigma}} \exp\left(-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{z} - \boldsymbol{\mu})\right)$$

Pdf dei dati é gaussiana: $f_y(y)$

$$f_y(y) = \frac{1}{\sqrt{2\pi \cdot \lambda_{yy}^2}} \cdot \exp\left(-\frac{1}{2\lambda_{yy}^2}(y-0)^2\right)$$

Si dimostra che la ***posterior*** $f_{\theta|y}(\theta|y) = \frac{f_{y\theta}(y,\theta)}{f_y(y)}$ *è ancora* ***Gaussiana*** *con:*

Valore atteso: $\mu_{\theta|y} = \dfrac{\lambda_{\theta y}}{\lambda_{yy}^2} \cdot y$

Varianza: $\lambda_{\theta|y}^2 = \lambda_{\theta\theta}^2 - \dfrac{\lambda_{\theta y}^2}{\lambda_{yy}^2}$

Avendo osservato il valore $y(1)$ di y, la stima ottenuta dallo **stimatore ottimo Bayesiano nel caso Gaussiano** sará:

$$\hat{\theta}_{opt} = \mathbb{E}\big[\theta | y = y(1)\big] = \frac{\lambda_{\theta y}}{\lambda_{yy}^2} \cdot y(1)$$

Stima ottima lineare

Vogliamo quindi trovare uno **stimatore che non faccia ipotesi sulla ddp congiunta** di y e θ.
Supponiamo y e θ due variabili casuali scalari con valore atteso nullo e varianza λ_{yy}^2 e $\lambda_{\theta\theta}^2$ rispettivamente:

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

Vogliamo stimare θ tramite uno **stimatore lineare** del tipo:

$$\hat{\theta}^{lin} = \alpha \cdot y + \beta \qquad \alpha, \beta \in \mathbb{R}$$

Per trovare α e β, **minimizziamo la funzione di costo** data dal Mean square error:

$$MSE \equiv J(\alpha, \beta) = \mathbb{E}\big[\big(\hat{\theta} - \theta\big)^2\big] = \mathbb{E}\big[\big(\alpha \cdot y + \beta - \theta\big)^2\big]$$

Calcoliamo il gradiente e poniamolo uguale a zero (non verifichiamo sia un minimo):

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

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

Lo stimatore lineare ottimo **coincide con lo stimatore ottimo di Bayes per il caso Gaussiano:**

$$\hat{\theta}_{opt}^{lin} = \hat{\alpha} \cdot y + \hat{\beta} = \frac{\lambda_{\theta y}}{\lambda_{yy}^2} \cdot y$$

La varianza della stima si ricava essere uguale al caso Gaussiano:

$$Var[\hat{\theta}_{opt}^{lin} - \theta] = \lambda_{\theta\theta}^2 - \frac{\lambda_{\theta y}^2}{\lambda_{yy}^2}$$

---

5 Lezione 8: Processi stocastici

Proprietà della densità spettrale di potenza, stima dello spettro, proprietà statistiche, come migliorarne la stima

Densitá spettrale di potenza e proprietá

Definzione: Dato un processo stocastico stazionario (sia in senso debole che in senso forte), si definisce **densità spettrale di potenza** $\Gamma_{vv}(\omega)$ come la DTFT di $\gamma_{vv}(\tau)$

$$\Gamma_{vv}(\omega) \equiv \mathcal{F}\big[\gamma_{vv}(\tau)\big] = \sum_{\tau=-\infty}^{+\infty} \gamma_{vv}(\tau) \cdot e^{-j\omega\tau}$$

$$\textit{La trasformata } \mathcal{Z} \textit{ si } \gamma_{vv}(\tau) \textit{ é:} \quad \Phi_{vv}(z) \equiv \mathcal{Z}\big[\gamma_{vv}(\tau)\big] = \sum_{\tau=-\infty}^{+\infty} \gamma_{vv}(\tau) \cdot z^{-t}$$

Data $\Phi_{vv}(z)$ si ha che $\Gamma_{vv}(\omega) = \Phi_{vv}(e^{j\omega})$ (stesso concetto tra $\mathcal{Z}$ e $\mathcal{F}$).

Interpretazione: la densità spettrale di potenza ci dice come, in media, le componenti in frequenza delle varie realizzazioni del processo stocastico $v(t, s)$ contribuiscono alla sua varianza. O in parole piú povere **come l'energia del processo si distribuisce alle varie frequenze**.

Proprietá di $\Gamma_{vv}(\omega)$:

- **Reale:** dato che $\gamma_{vv}(\tau)$ é pari, i termini immaginari del tipo $\pm j \cdot sin(\omega)$ si elidono
- **Positiva:** $\Gamma_{vv}(\omega) \geqslant 0,\ \forall\omega \in \mathbb{R}$
- **Pari:** $\Gamma_{vv}(\omega) = \Gamma_{vv}(-\omega),\ \forall\omega \in \mathbb{R}$
- **Periodica di periodo 2π:** $\Gamma_{vv}(\omega) = \Gamma_{vv}(\omega + k \cdot 2\pi),\ \forall\omega \in \mathbb{R}, \forall k \in \mathbb{Z}$
- Affinché $\Gamma_{vv}(\omega)$ **converga**, $\gamma_{vv}(\tau)$ deve tendere a zero in modo sufficientemente rapido.
- Ci basta valutare $\Gamma_{vv}(\omega)$ tra $[0, \pi]$
- È possibile risalire a $\gamma_{vv}(\tau)$ tramite l'antitrasformata

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

- é possibile esprimere la **varianza del processo** stazionario come l'area sottesa alla densità spettrale di potenza

$$\gamma_{vv}(0) = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \Gamma_{vv}(\omega) d\tau$$

Stima dello spettro, proprietá statistiche, come migliorarne la stima

Vedere risposta successiva.

Descrivere come si può effettuare la stima dello spettro (densità spettrale di potenza) di un processo stocastico stazionario, partendo dai dati misurati. Discutere inoltre le proprietà dello stimatore. Cosa si può fare per migliorare la stima?

Definzione: Dato un processo stocastico stazionario (sia in senso debole che in senso forte), si definisce **densità spettrale di potenza** $\Gamma_{vv}(\omega)$ come la DTFT di $\gamma_{vv}(\tau)$

$$\Gamma_{vv}(\omega) \equiv \mathcal{F}\big[\gamma_{vv}(\tau)\big] = \sum_{\tau=-\infty}^{+\infty} \gamma_{vv}(\tau) \cdot e^{-j\omega\tau}$$

Autocovarianza (temporale) campionaria

$$\hat{\gamma}_{vv}(\tau) = \frac{1}{N-|\tau|} \sum_{t=0}^{N-|\tau|-1} v(t) \cdot v(t+|\tau|), \qquad |\tau| < N$$

- $\mathbb{E}\big[\hat{\gamma}_{vv}(\tau)\big] = \gamma_{vv}(\tau)$, ovvero lo stimatore é **corretto**
- Per τ fissato, lo stimatore è **consistente**, sotto le ipotesi di ergodicità
- Per $\tau \approx N$, si ha che $var\big[\hat{\gamma}_{vv}(\tau)\big]$ é grande perché ci sono pochi addendi

Per risolvere quest'ultimo problema, possiamo pensare ad uno stimatore alternativo (seppur non corretto)

Autocovarianza (temporale) campionaria - versione alternativa

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

- $\mathbb{E}\big[\hat{\gamma}'_{vv}(\tau)\big] = \frac{N-|\tau|}{N} \cdot \gamma_{vv}(\tau)$, ovvero lo stimatore è **distorto**, ma **asintoticamente corretto**
- Per τ fissato, lo stimatore è **consistente**, sotto le ipotesi di ergodicità

Densità spettrale campionaria

Non conoscendo $\gamma_{vv}(\tau)$, uso $\hat{\gamma}_{vv}(\tau)$ oppure $\hat{\gamma}'_{vv}(\tau)$. Si definisce **periodogramma** il seguente **stimatore** della densitá spettrale di potenza:

$$\textbf{Stimatore: periodogramma} \qquad I_N(\omega) = \sum_{\tau=-(N-1)}^{N-1} \hat{\gamma}'_{vv}(\tau) \cdot e^{-j\omega\tau}$$

- A differenza di $\Gamma_{vv}(\omega)$, $I_N(\omega)$ é definito solo da $\tau = -(N-1)$ a $\tau = N-1$
- Essendo la DTFT di $\hat{\gamma}_{vv}(\tau)$, $I_N(\omega)$ é una funzione **reale, continua,** 2π **periodica**

Proprietá dello stimatore

- Lo stimatore $I_N(\omega)$ **non é corretto**, ma é **asintoticamente corretto**. Notiamo che non lo sarebbe stato neanche se avessi usato $\hat{\gamma}_{vv}(\tau)$ al posto di $\hat{\gamma}'_{vv}(\tau)$, infatti:

$$\mathbb{E}\big[I_N(\omega)\big] = \sum_{\tau=-(N-1)}^{N-1} \underbrace{\mathbb{E}\big[\hat{\gamma}'_{vv}(\tau)\big]}_{} \cdot e^{-j\omega\tau} =$$

$$\mathbb{E}\big[I_N(\omega)\big] = \sum_{\tau=-(N-1)}^{N-1} \underbrace{\frac{N-|\tau|}{N} \cdot \gamma_{vv}(\tau)}_{} \cdot e^{-j\omega\tau} \neq \Gamma_{vv}(\tau)$$

- Si dimostra che come $Var[I_N(\omega)] \approx \Gamma^2_{vv}(\omega)$. Per cui, la varianza dello stimatore non decresce al crescere di N. Lo stimatore **non è consistente**
- Per $N \to +\infty$, $I_N(\omega_1)$ e $I_N(\omega_2)$ tendono a **diventare incorrelati**, $\forall \omega_1 \neq \omega_2$. Questo ci dà l'idea che il periodogramma sia una funzione **poco continua**, poiché la stima in una frequenza può non essere simile alla stima in una frequenza anche adiacente (una sorta di **rumore bianco in frequenza**).

Per migliorare la stima - Metodo di Bartlett

- Ipotizziamo di avere N dati a disposizione
- Dividiamo i dati in $K = \frac{N}{M}$ parti, dove M é la lunghezza di ogni porzione di dati
- Calcoliamo il periodogramma $I^{[i]}_{M,K}(\omega)$ per ciascuna parte $i = 1, 2, \ldots, K$

- facciamo la media dei periodogrammi, ottenendo la stima

$$\bar{I}_{M,K}(\omega) = \frac{1}{K} \sum_{i=1}^{K} I^{[i]}_{M,K}(\omega)$$

Osservazioni:

- se $\gamma_{vv}(\tau) \to 0$ in modo sufficientemente rapido, i K periodogrammi sono **circa indipendenti**. In questo caso, si ha che

$$Var\big[\bar{I}_{M,K}(\omega)\big] = O\big(\frac{1}{K} \cdot \Gamma^2_{vv}(\omega)\big)$$

- Il $Bias\big[\bar{I}_{M,K}(\omega)\big]$ è maggiore rispetto a quello di $I_N(\omega)$. Questo comporta una maggior **perdita di informazione in frequenza**
- Se so che $\Gamma_{vv}(\omega)$ ha **pichi molto stretti**, devo usare M **grande** in modo da avere abbastanza risoluzione in frequenza

---

6 Lezione 9: Famiglie di modelli stocastici

Teorema dei processi ARMA

Definizione: Un processo stocastico $y(t)$, generato a partire dal rumore bianco $e(t) \sim WN(\mu, \lambda^2)$, é detto di tipo $\boldsymbol{ARMA(n_a, n_c)}$, se:

$$\begin{aligned} y(t) = a_1 \cdot y(t-1) + a_2 \cdot y(t-2) + \cdots + a_{n_a} \cdot y(t-n_a) \qquad & \textbf{Parte } AR(n_a) \\ + e(t) + c_1 \cdot e(t-1) + c_2 \cdot e(t-2) + \cdots + c_{n_c} \cdot e(t-n_c) \qquad & \textbf{Parte } MA(n_c) \end{aligned}$$

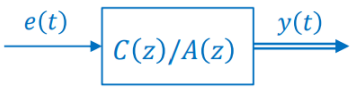
$a_1, \ldots, a_{n_a}$: *coeffcienti del modello* $AR(n_a)$ $\qquad$ n_a : *ordine del modello* $AR(n_a)$

$c_0, c_1, \ldots, c_{n_c}$: *coeffcienti del modello* $MA(n_c)$ $\qquad$ n_c : *ordine del modello* $MA(n_c)$

La **funzione di trasferimento** di un $ARMA(n_a, n_c)$ risulta essere:

$$y(t)\left[1 - a_1 z^{-1} - a_2 z^{-2} - \cdots - a_{n_a} z^{-n_a}\right] = \left[1 + c_1 z^{-1} + c_2 z^{-2} + \cdots + c_{n_c} z^{-n_c}\right] e(t)$$

$$y(t) = \frac{1 + c_1 z^{-1} + c_2 z^{-2} + \cdots + c_{n_c} z^{-n_c}}{1 - a_1 z^{-1} - a_2 z^{-2} - \cdots - a_{n_a} z^{-n_a}} e(t) \quad \Rightarrow \quad \boxed{y(t) = \frac{C(z)}{A(z)} e(t)}$$



Il processo $y(t)$ é stazionario se e solo se $\frac{C(z)}{A(z)}$ é asintoticamente stabile.

Teorema: Dato un processo stocastico stazionario $ARMA(n_a, n_c)$, esso puó essere descritto come un $MA(\infty)$.

Esempio: *Supponiamo di avere un AR(1) del tipo*

$$y(t) = ay(t-1) + e(t) \qquad e(t) \sim WN(0, \lambda^2)$$

$$y(t) = \frac{1}{1 - az^{-1}} e(t)$$

può essere visto come il **limite di una serie geometrica** di ragione az^{-1}

$$\Rightarrow \quad = \sum_{i=0}^{+\infty} (az^{-1})^i \cdot e(t) \quad = \sum_{i=0}^{+\infty} a^i \cdot e(t-i) \qquad \text{MA}(\infty)$$

7 Lezione 10: Predizione

Spiegare il concetto di filtro passa-tutto e le sue particolarità. Enunciare inoltre il teorema di fattorizzazione spettrale (forma canonica) e come il filtro passa tutto possa essere sfruttato per raggiungere la forma canonica.

Filtro passa-tutto

Definizione: il filtro passa-tutto è un filtro di ordine 1 definito come:

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

Lo zero é reciproco del polo

$$zero : z = -a \qquad polo : z = -\frac{1}{a}$$

Il fattore moltiplicativo é come il polo

Osservazioni:

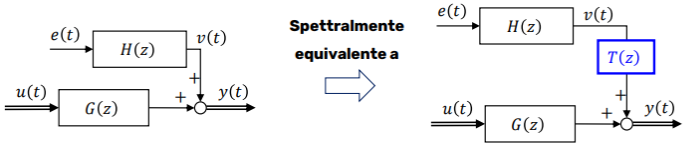
- La densitá spettrale di potenza $\Gamma_{yy}(\omega)$ di un processo $y(t)$ in uscita dal passa tutto $T(z)$ alimentato da un generico processo stazionario in ingresso $v(t)$ é uguale a:

  $$\Gamma_{yy}(\omega) = |T(e^{j\omega})|^2 \cdot \Gamma_{vv}(\omega)$$

  Il filtro passa-tutto non modifica il modulo delle frequenze nella densità spettrale di potenza dell'ingresso. Quindi, si ha che:

  $$\Gamma_{yy}(\omega) = \Gamma_{vv}(\omega)$$

- Il processo $y(t)$ in uscita al passa-tutto è **spettralmente equivalente** al processo $v(t)$ in ingresso al passatutto, tuttavia i due processi $y(t)$ e $v(t)$ **non sono identici** poiché il passatutto introduce uno **sfasamento**



Filtro passa-tutto e forma canonica

Vogliamo risolvere il problema della predizione per **processi a spettro razionale**, ovvero processi $y(t)$ generati in uscita da un sistema dinamico lineare asintoticamente stabile con funzione di trasferimento $H(z)$ razionale fratta alimentato da $e(t) \sim WN(0, \lambda^2)$.

Il problema della fattorizzazione spettrale consiste nel trovare tutte le coppie $\{H(z), \lambda^2\}$ tali che:

$$\Phi_{yy}(z) = \lambda^2 \cdot H(z) \cdot H(z^{-1})$$

Per processi a spettro razionale, esistono **infiniti fattori spettrali** $\{H(z), \lambda^2\}$. Ai fini della predizione ottima ci servirá un fattore spettrale detto **canonico**.

Il filtro viene applicato alla funzione di trasferimento $H(z)$ in modo tale da eliminare eventuali radici che non sono interne al cerchio unitario e rendere dunque il processo in forma canonica.

Teorema della fattorizzazione spettrale

Dato un processo stocastico stazionario a spettro razionale, esiste **un solo fattore spettrale** $\{\widetilde{H}(z), \widetilde{\lambda}^2\}$, detto **fattore spettrale canonico**, dove $\widetilde{H}(z) = \frac{C(z)}{A(z)}$ tale che:

1. $C(z)$ e $A(z)$ hanno lo **stesso grado** (grado relativo nullo)
2. $C(z)$ e $A(z)$ sono **coprimi** (non ci son fattori in comune)
3. $C(z)$ e $A(z)$ sono **monici** (il coefficiente del termine di grado massimo è 1)
4. $C(z)$ e $A(z)$ hanno **radici interne al cerchio unitario**

Si consideri un processo ARMA dato in rappresentazione canonica. Spiegare perché la varianza dell'errore di predizione a un passo coincide con la varianza del rumore bianco che genera il processo.

Sia dato un processo $ARMA(n_a, n_c)$ in forma canonica:

$$y(t) = H(z) \cdot e(t) = \frac{C(z)}{A(z)} \cdot e(t) \qquad e(t) \sim WN(0, \lambda^2)$$

$$C(z) = 1 + c_1 z^{-1} + c_2 z^{-2} + \cdots + c_{n_c} z^{-nc}$$
$$A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - \cdots - c_{n_a} z^{-na}$$

Per scomporre la parte impredicibile da quella predicibile viene impiegata la **lunga divisione**, in questo modo possiamo esprimere $C(z)/A(z)$ come:

$$\frac{C(z)}{A(z)} = E(z) + \frac{R(z)}{A(z)} = E(z) + \frac{z^{-k}\widetilde{R}(z)}{A(z)} \qquad \textit{con } k = \textit{passi di lunga divisione}$$

$$y(t) = \underbrace{E(z) \cdot e(t)}_{\textit{parte impred.}} + \underbrace{\frac{\widetilde{R}(z)}{A(z)} \cdot e(t-k)}_{\textit{parte pred}}$$

Filtro sbiancante:

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

Il predittore ottimo dal rumore

$$\widehat{y}(t|t-k) = \frac{\widetilde{R}(z)}{A(z)} \cdot e(t-k)$$

Il predittore ottimo dai dati

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

$$\widehat{y}(t|t-k) = \frac{\widetilde{R}(z)}{C(z)} \cdot y(t-k)$$

L'errore di predizione

$$\varepsilon_k(t) = y(t) - \widehat{y}(t|t-k) = E(z) \cdot e(t)$$

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

$$\bullet E(z) = 1 \qquad \bullet\ R(z) = C(z) - A(z)$$

$$\widehat{y}(t|t-1) = \frac{C(z) - A(z)}{C(z)} \cdot y(t)$$
$$\varepsilon_1(t) = E(z) \cdot e(t) = e(t)$$

Il ***predittore lineare ottimo*** *dai dati é quello che minimizza il seguente criterio* ***Mean Squared Error (MSE):***

$$var\big[\varepsilon_k(t)\big] = \mathbb{E}\big[\varepsilon_k(t)^2\big] = \mathbb{E}\big[\big(y(t) - \hat{y}(t|t-k)\big)^2\big]$$

Affiché il predittore sia ottimo occorre che:

- $\mathbb{E}\big[\varepsilon_k(t)\big] = \mathbb{E}\big[y(t) - \hat{y}(t|t-k)\big] = 0 \Rightarrow$ sia **Corretto**, ossia valore atteso nullo.
- $\mathbb{E}\big[\hat{y}(t|t-k) \cdot \varepsilon_k(t)\big] = 0 \Rightarrow$ il predittore e l'errore di predizione sono **incorrelati**
- $Var\big[\varepsilon_k(t)\big] = \mathbb{E}\big[\varepsilon_k(t)^2\big] = \mathbb{E}\big[\big(y(t) - \hat{y}(t|t-k)\big)^2\big] = \mathbb{E}\big[\big(E(z) \cdot e(t)\big)^2\big] = \mathbb{E}\big[e(t)^2\big]$ minima

8 Lezione 11: Identificazione: concetti fondamentali

Descrivere il metodo della massima verosimiglianza per la stima dei modelli ARMAX

I metodi di stima basati sulla minimizzazione dell'errore di predizione prendono il nome di **Prediction Error Methods (PEM)**.

Se ipotizzo che $\mathcal{S} = \mathcal{M}(\boldsymbol{\theta}^0)$ 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, ma se i dati sono molti, non c'é differenza.

Consideriamo un modello $ARMAX(n_a, n_c, n_b, k = 1)$, (dove $\boldsymbol{k}$ é il ritardo puro tra ingresso $u(t)$ ed uscita $y(t)$) e di avere a disposizione N dati $\{u(1), \dots, u(n)\}$ e $\{y(1), \dots, y(n)\}$

$$\mathcal{M}(\boldsymbol{\theta}): \quad y(t) = \frac{B(z,\boldsymbol{\theta})}{A(z,\boldsymbol{\theta})} \cdot u(t-1) + \frac{C(z,\boldsymbol{\theta})}{A(z,\boldsymbol{\theta})} \cdot e(t) \qquad e(t) \sim WN(0, \lambda^2)$$

$$B(z) = b_0 + b_1 z^{-1} + \cdots + b_{n_b} z^{-n_b}$$

$$A(z) = 1 - a_1 z^{-1} - \cdots - a_{n_a} z^{-n_a}$$

$$C(z) = 1 + c_1 z^{-1} + \cdots + c_{n_c} z^{-n_c}$$

Calcoliamo l'espressione dell'errore di predizione ad un passo. In questo caso, si ha che $E(z) = 1$, e quindi $\varepsilon_1(t) = e(t)$. Di conseguenza, esprimendo $e(t)$ in funzione di $u(t)$ e $y(t)$:

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

$$\textit{oppure tramite:} \qquad \varepsilon_1(t;\boldsymbol{\theta}) = H^{-1}(z,\boldsymbol{\theta}) \cdot \big[y(t) - G(z,\boldsymbol{\theta} \cdot u(t)\big]$$

Utilizziamo l'approccio predittivo:

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

$$\hat{\boldsymbol{\theta}} = arg \min_{\boldsymbol{\theta}} J_N(\boldsymbol{\theta})$$

Osservazioni:

- Dato che ho $C(z,\boldsymbol{\theta})$ al denominatore, questa funzione di costo non é piú convessa. In generale avró dei **minimi locali**
- Per la risoluzione del problema di minimizzazione, devo utilizzare dei **metodi iterativi** per esempio **il metodo del gradiente**, oppure un'alternativa (più efficiente) al metodo del gradiente il metodo di ottimizzazione iterativo noto come **Metodo di Newton**. Questo metodo, oltre al gradiente, sfrutta anche l'informazione data dalla **matrice Hessiana.**

Rappresentare la cifra di merito e lo schema a blocchi del filtraggio dell'algoritmo di identificazione dei processi ARMAX. Si descriva l'approssimazione dell'Hessiana utilizzata, indicando le motivazioni che inducono a questa approssimazione

Cifra di merito o funzione di costo

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

Osservazioni:

- Dato che ho $C(z,\boldsymbol{\theta})$ al denominatore, questa funzione di costo non é piú convessa. In generale avró dei **minimi locali**
- Per la risoluzione del problema di minimizzazione, devo utilizzare dei **metodi iterativi** per esempio **il metodo del gradiente**, oppure un'alternativa (più efficiente) al metodo del gradiente il metodo di ottimizzazione iterativo noto come **Metodo di Newton**. Questo metodo, oltre al gradiente, sfrutta anche l'informazione data dalla **matrice Hessiana.**

Metodo di Newton

Sviluppo in serie di Taylor troncata al 2° ordine di $J_N(\boldsymbol{\theta})$, nell'intorno della stima all'iterazione $i-esima$ $\widehat{\boldsymbol{\theta}}^{(i)}$ (approssimiamo la funzione di costo con il paraboloide).

$$J_N(\boldsymbol{\theta}) \approx V(\boldsymbol{\theta})$$

La funzione $V(\boldsymbol{\theta})$ é un ***paraboloide***
(é facile da calcolare il minimo)

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

Una volta ottenuta l'approssimazione $V^{(i)}(\boldsymbol{\theta})$, si calcola $\widehat{\boldsymbol{\theta}}^{(i+1)}$ come minimo di $V^{(i)}(\boldsymbol{\theta})$. Troviamo un'espressione esplicita per $\widehat{\boldsymbol{\theta}}^{(i+1)}$ imponendo:

$$\frac{dV^{(i)}(\boldsymbol{\theta})}{d\boldsymbol{\theta}} = \underset{d\times 1}{\mathbf{0}}$$

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

Da seguente formula ricaviamo ***la regola di update per il metodo di Newton:***

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

Gradiente di $J_N(\boldsymbol{\theta})$:

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

Hessiano di $J_N(\boldsymbol{\theta})$:

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

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

- Se siamo vicini all'ottimo, $\varepsilon_1(t;\boldsymbol{\theta})$ é 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**. In questo modo, la direzione dell'algoritmo è sicuramente discendente (concetto simile ad avere learning rate $\geqslant 0$)

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

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

$$\hat{\boldsymbol{\theta}}^{(i+1)} = \hat{\boldsymbol{\theta}}^{(i)} - \left[hessiana\right]^{-1} \cdot \left[gradiente\right]$$

Ricordando che:

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

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

$$\boldsymbol{\theta} = \begin{bmatrix} a_1 \dots a_{n_a} & b_0 b_1 \dots b_{n_b} & c_1 \dots c_{n_c} \end{bmatrix}^T$$

$$d \times 1 = n_a + n_b + 1 + n_c \times 1$$

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

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

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

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

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

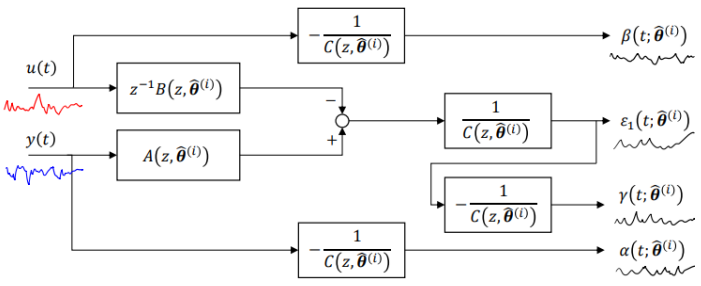
$$\frac{d[C(z)\cdot\varepsilon_1(t)]}{dc_1} = 0 \Rightarrow z^{-1}\varepsilon_1(t) + C(z)\frac{d\varepsilon_1(t)}{dc_1} = 0$$

$$\frac{d\varepsilon_1(t)}{dc_1} = -\frac{1}{C(z)}\varepsilon(t-1) = \gamma(t-1) \qquad \gamma(t) \equiv -\frac{1}{C(z)}\varepsilon_1(t)$$

Riassumendo, il vettore gradiente è:

$$\underset{d\times 1}{\frac{d\varepsilon_1(t)}{d\boldsymbol{\theta}}} = \begin{bmatrix} \alpha(t-1) \\ \vdots \\ \alpha(t-n_a) \\ \beta(t-1) \\ \vdots \\ \beta(t-n_b-1) \\ \gamma(t-1) \\ \vdots \\ \gamma(t-n_c) \end{bmatrix} \quad t = 1, \dots, N$$

È possibile definire in modo elegante il calcolo del gradiente tramite una serie di **filtraggi dei segnali di ingresso e uscita**



9 Lezione 12: Identificazione analisi e complementi

Si dimostri che, se il sistema da identificare appartiene alla classe dei modelli e il minimo globale della cifra di merito è unico, tale minimo globale corrisponde al sistema che genera i dati.

Ipotesi di lavoro: Sia l'ingresso $u(t,s)$ e sia l'uscita $y(t,s)$ sono processi stocastici stazionari ed ergodici. Di conseguenza, i dati misurati saranno una realizzazione dei processi $u(t,s)$ e $y(t,s)$ in corrispondenza di un particolare esito $\bar{s}$.

La funzione di costo dipende anch'essa dall'esito $\bar{s}$ poiché utilizza i dati misurati da cui ottengo la stima $\widehat{\boldsymbol{\theta}}_N(\bar{s})$, dunque la stima $\widehat{\boldsymbol{\theta}}_N(s)$ é una **variabile causuale**.

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

La funzione di costo $J_N(\boldsymbol{\theta}, s)$ dovrebbe essere interpretata come un **insieme di curve** le quali dipendono da s e la stima $\widehat{\boldsymbol{\theta}}_N(s)$ come un **un insieme di punti**.

Osservazione: grazie **all'ipotesi di ergodicitá** di $u(t,s)$ e $y(t,s)$ abbiamo che i momenti temporali convergono ai rispettivi momenti di insieme $\Rightarrow$ cioé le curve $J_N(\boldsymbol{\theta}, s)$ convergono ad **unica (deterministica) curva** $\bar{J}(\boldsymbol{\theta})$.

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

Definiamo inoltre **l'insieme dei punti di minimo globale** di $\bar{J}(\boldsymbol{\theta})$ come:

$$\Delta_{\boldsymbol{\theta}} = \{\bar{\boldsymbol{\theta}} \mid \bar{J}(\boldsymbol{\theta}) \geqslant \bar{J}(\bar{\boldsymbol{\theta}}), \quad \forall \boldsymbol{\theta}\}$$

Da dimostrare: se $\mathcal{S} \in \mathcal{M}(\boldsymbol{\theta})$ e $\Delta_{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}$, allora $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}^0$ dove $\mathcal{S} = \mathcal{M}(\boldsymbol{\theta}^0)$.

Ipotesi di lavoro:

- Assumiamo che $\mathcal{S} \in \mathcal{M}(\boldsymbol{\theta})$, ovvero che esista $\boldsymbol{\theta}^0 \in \boldsymbol{\Theta}$ tale che $\mathcal{S} = \mathcal{M}(\boldsymbol{\theta}^0)$.
- $\boldsymbol{\theta}^0 \in \Delta_{\boldsymbol{\theta}}$ di $\bar{J}(\boldsymbol{\theta})$? ossia la stima $\widehat{\boldsymbol{\theta}}_N$ tende asintoticamente a $\boldsymbol{\theta}^0$? I metodi PEM sono in grado di trovare la parametrizzazione vera del modello ?

Dimostrazione: dimostriamo che, sotto le ipotesi fatte $\boldsymbol{\theta}^0$ appartiene sempre a $\Delta_{\boldsymbol{\theta}}$

Supponiamo che i dati siano generati dal sistema $\mathcal{S}$, tale che:

$$y(t) = \hat{y}(t|t-1;\boldsymbol{\theta}^0) + e(t), \qquad e(t) \sim WN(0,\lambda^2)$$

Consideriamo un generico modello $\mathcal{M}(\boldsymbol{\theta})$,
senza assumere che $\varepsilon_1(t;\boldsymbol{\theta})$ sia bianco

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

L'errore di predizione ad un passo commesso dal modello $\mathcal{M}(\boldsymbol{\theta})$ é dunque:

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

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

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

$$\varepsilon_1(t;\boldsymbol{\theta}) = \underbrace{y(t) - \hat{y}(t|t-1;\boldsymbol{\theta}^0)}_{\substack{\text{Err. di predizione ottimo} \\ \varepsilon_1(t;\boldsymbol{\theta}^0)=e(t)}} + \hat{y}(t|t-1;\boldsymbol{\theta}^0) - \hat{y}(t|t-1;\boldsymbol{\theta})$$

$$\text{Errore di predizione ottimo} = y(t) - \hat{y}(t|t-1;\boldsymbol{\theta}^0) = \varepsilon_1(t;\boldsymbol{\theta}^0) = e(t)$$

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

Calcoliamo la varianza dell'errore di predizione (che è a media nulla poiché il predittore è corretto):

$$\mathbb{E}\Big[\varepsilon_1(t;\boldsymbol{\theta})^2\Big] = \mathbb{E}\Big[\Big(e(t) + \hat{y}(t|t-1;\boldsymbol{\theta}^0) - \hat{y}(t|t-1;\boldsymbol{\theta})\Big)^2\Big]$$

$$\Rightarrow \bar{J}(\boldsymbol{\theta}) = \mathbb{E}\Big[e(t)^2\Big] + \mathbb{E}\Big[\Big(\hat{y}(t|t-1;\boldsymbol{\theta}^0) - \hat{y}(t|t-1;\boldsymbol{\theta})\Big)^2\Big] + \underbrace{2\mathbb{E}\Big[e(t)\cdot\Big(\hat{y}(t|t-1;\boldsymbol{\theta}^0) - \hat{y}(t|t-1;\boldsymbol{\theta})\Big)\Big]}_{incorrelati}$$

$$= \lambda^2 + \underbrace{\mathbb{E}\Big[\Big(\hat{y}(t|t-1;\boldsymbol{\theta}^0) - \hat{y}(t|t-1;\boldsymbol{\theta})\Big)^2\Big]}_{varianza}$$

È una varianza, quindi una quantità $\geqslant 0$.
In particolare si annulla solo per $\boldsymbol{\theta} = \boldsymbol{\theta}^0$

$$\Rightarrow \bar{J}(\boldsymbol{\theta}) \geqslant \lambda^2 = \bar{J}(\boldsymbol{\theta}^0), \quad \forall \boldsymbol{\theta}$$

$$\bar{J}(\boldsymbol{\theta}) \geqslant \bar{J}(\boldsymbol{\theta}^0) \quad \forall \boldsymbol{\theta}$$

$$\boldsymbol{\theta}^0 \textbf{ é un minimo di } \bar{J}(\boldsymbol{\theta})$$

Conclusione (fondamentale): Se $\mathcal{S} \in \mathcal{M}(\boldsymbol{\theta})$ e $u(t), y(t)$ sono pss ergodici, allora, per $N \to +\infty$, *un metodo PEM garantisce che il modello stimato è quello vero* $\mathcal{S} = \mathcal{M}(\boldsymbol{\theta}^0)$ o un insieme equivalente di modelli $\{\mathcal{M}(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Delta_{\boldsymbol{\theta}}\}$ con la stessa capacitá nello spiegare i dati.

- Se $\mathcal{S} \in \mathcal{M}(\boldsymbol{\theta})$, allora in corrispondenza di $\boldsymbol{\theta}^0$ si ha che $\varepsilon_1(t;\boldsymbol{\theta}^0) = e(t) \sim WN$ basta fare test bianchezza sui residui $\varepsilon_1(t;\hat{\boldsymbol{\theta}}_N)$.

$\mathcal{S} \in \mathcal{M}(\boldsymbol{\theta})$ *e* $\Delta_{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}$, *allora* $N_{\to\infty}\ \hat{\boldsymbol{\theta}}_N \approx \bar{\boldsymbol{\theta}} = \boldsymbol{\theta}^0$	*modello perfetto*
$\mathcal{S} \in \mathcal{M}(\boldsymbol{\theta})$ *e* $\Delta_\theta = \{\boldsymbol{\theta_i}, ..\}$, *allora* $N_{\to\infty}\ \hat{\boldsymbol{\theta}}_N \approx \bar{\boldsymbol{\theta}} \neq \boldsymbol{\theta}^0$	$\mathcal{M}(\bar{\boldsymbol{\theta}}) \approx \mathcal{M}(\boldsymbol{\theta}^0)$

- Se $\mathcal{S} \notin \mathcal{M}(\boldsymbol{\theta})$, allora i metodi PEM non garantiscono di stimare correttamente TUTTE le componenti del sistema $\mathcal{S}$, ma il modello trovato $\mathcal{M}(\bar{\boldsymbol{\theta}})$ risulta essere la migliore approssimazione di $\mathcal{S}$ nella famiglia di modelli $\mathcal{M}(\boldsymbol{\theta})$

$\mathcal{S} \notin \mathcal{M}(\boldsymbol{\theta})$ *e* $\Delta_{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}$, *allora* $N_{\to\infty}\ \hat{\boldsymbol{\theta}}_N \approx \bar{\boldsymbol{\theta}} \neq \boldsymbol{\theta}^0$	$\mathcal{M}(\bar{\boldsymbol{\theta}}) \sim \mathcal{M}(\boldsymbol{\theta}^0)$
$\mathcal{S} \notin \mathcal{M}(\boldsymbol{\theta})$ *e* $\Delta_\theta = \{\boldsymbol{\theta_i}, ..\}$, *allora* $N_{\to\infty}\ \hat{\boldsymbol{\theta}}_N \approx \bar{\boldsymbol{\theta}} \neq \boldsymbol{\theta}^0$	$\mathcal{M}(\bar{\boldsymbol{\theta}}) \sim \mathcal{M}(\boldsymbol{\theta}^0)$

Identificabilità sperimentale e segnale persistentemente eccitante

Vedere risposta successiva.

Spiegare cosa si intende per identificabilità sperimentale e strutturale precisando la nozione di persistente eccitazione di un segnale

Identificabilità sperimentale e strutturale

L'analisi asintotica ci dice che se $\mathcal{S} \in \mathcal{M}(\boldsymbol{\theta})$ e $u(t), y(t)$ sono pss ergodici, allora, per $N \to +\infty$, un metodo PEM garantisce che il modello stimato è quello vero $\mathcal{S} = \mathcal{M}(\boldsymbol{\theta}^0)$ o un insieme equivalente di modelli $\{\mathcal{M}(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Delta_{\boldsymbol{\theta}}\}$.

Nel secondo caso dobbiamo capire in quali condizioni il sistema $\mathcal{S}$ può essere **identificato univocamente** dai dati. Affinché un modello sia univocamente identificabile è necessario avere:

- **Identificabilitá strutturale:** il modello $\mathcal{M}(\boldsymbol{\theta})$ non deve essere *sovraparametrizzato* rispetto al sistema $\mathcal{S}$.
- **Identificabilitá sperimentale:** i dati $\{u(t), y(t)\}_{t=1}^N$ devono contenere *sufficiente informazione*

ll 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 del modello.

Persistente eccitazione di un segnale

Vedere domande successive.

Dire cosa si intende per identificabilità di un processo ARX e descrivere i principali problemi di identificabilità

Identificabilità dei modelli ARX

Dato un modello $ARX(n_a, n_b, 1)$, avendo N dati $\{u(1), \dots, u(N)\}$, $\{y(1), \dots, y(N)\}$ risulta essere:

$$\mathcal{M}(\boldsymbol{\theta}): \quad y(t) = \frac{B(z,\boldsymbol{\theta})}{A(z,\boldsymbol{\theta})} \cdot u(t-1) + \frac{1}{A(z,\boldsymbol{\theta})} \cdot e(t) \qquad e(t) \sim WN(0, \lambda^2)$$

$$B(z) = b_0 + b_1 z^{-1} + \dots + b_{n_b} z^{-n_b}$$

$$A(z) = 1 + a_1 z^{-1} + \dots + a_{n_a} z^{-n_a}$$

La stima tramite il metodo dei minimi quadrati risulta essere:

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

Problema di identificabilitá: quando $\hat{\boldsymbol{\theta}}_N$ esiste ed é unico? $\Leftrightarrow$ quando $\sum_{t=1}^{N} \boldsymbol{\varphi}(t) \cdot \boldsymbol{\varphi}^T(t)$ é invertibile?

Definiamo:

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

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

Le matrici $S(N)$ e $R(N)$ sono **semidefinite positive** in quanto prodotto di un vettore per sé stesso. Affinché $\hat{\boldsymbol{\theta}}_N$ **esista** ed sia **unico**, é necessario che $S(N) \succ 0$ oppure $R(N) \succ 0$, ossia che:

$$\underset{d\times d}{det\big(R(N)\big)} > 0$$

Grazie all'ipotesi di ergodicità, abbiamo che $R(N) \underset{N\to+\infty}{\longrightarrow} \bar{R}$, dove la matrice $\bar{R}$ é la **matrice di autocovarianze** del processo congiunto $\{y(t), u(t)\}$. In generale, per un generico modello $ARX(n_a, n_b, 1)$ abbiamo che la matrice $\bar{R}$ puó essere divisa in quattro sotto-matrici:

$$\underset{(n_a+n_b+1)\times(n_a+n_b+1)}{\bar{R}} = \left[\begin{array}{c|c} \underset{n_a\times n_a}{\bar{R}_{yy}} & \underset{n_a\times(n_b+1)}{\bar{R}_{yu}} \\ \hline \underset{(n_b+1)\times n_a}{\bar{R}_{uy}} & \underset{(n_b+1)\times(n_b+1)}{\bar{R}_{uu}} \end{array}\right]$$

Lemma di Schur

Data una matrice M nella forma $M = \left[\begin{array}{c|c} F & K \\ \hline K^\top & H \end{array}\right]$, con F e K simmetriche. Condizione **necessaria e sufficiente** per l'invertibilità di M è che valgano:

- $H \succ 0$
- $F - KH^{-1}K^\top \succ 0$

Ricordando che

$$\bar{R} = \left[\begin{array}{c|c} \bar{R}_{yy} & \bar{R}_{yu} \\ \hline \bar{R}_{uy} & \bar{R}_{uu} \end{array}\right] \quad \Rightarrow$$

Condizione **necessaria** per l'invertibilità di $\bar{R}$ è che $\bar{R}_{uu} \succ 0$

La condizione (solo necessaria) sulla matrice $\bar{R}_{uu}$ riguarda quindi solo il segnale di ingresso $u(t)$ che 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

Identificabilità sperimentale e strutturale

Vedere risposte precedenti.

Discutere il concetto di identificabilità di un modello ARX, trattando approfonditamente la nozione di persistente eccitazione di un segnale nella stima di modelli dinamici e fare degli esempi di segnali persistentemente eccitanti

Identificabilità dei modelli ARX

Vedi risposte precedenti.

Persistente eccitazione nella stima di modelli dinamici

Definizione (Persistente eccitazione)

Definiamo la matrice $\bar{R}_{uu}^{(i)}$ di autocovarianza di $u(t)$ di ordine i come

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

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

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

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

Possiamo quindi dire che **condizione necessaria** per l'identificabilità sperimentale di un modello $ARX(n_a, n_b, 1)$, usato per produrre i dati, sia **persistemente eccitante di ordine pari ad almeno** $n_b + 1$ (Infatti $\bar{R}_{uu}$ ha dimensione $(n_b + 1) \times (n_b + 1)$).

Conclusione generale: supponendo quindi $\mathcal{S} \in \mathcal{M}(\boldsymbol{\theta})$, definiamo il numero di parametri di $G(z;\boldsymbol{\theta})$ come n_g. Allora la soluzione del problema di identificazione

$$\bar{\boldsymbol{\theta}} = arg \min_{\boldsymbol{\theta}} \mathbb{E}\left[\varepsilon_1(t,\boldsymbol{\theta})^2\right]$$

$$ha\ un'unica\ soluzione$$

$$\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}^0$$

$$se\ il\ segnale\ u(t)\ che\ genera\ i\ dati\ é$$
$$persistemente\ eccitante\ di\ ordine \geqslant n_g$$

$\Rightarrow$ Avere dunque un segnale eccitante è importante in ogni caso si voglia indentificare un modello dinamico.

Osservazioni:

- Se un segnale $u(t)$ è persistentemente eccitante di ordine n, allora è anche persistentemente eccitante di ordine $n-1$
- badiamo che la **condizione vista è solamente necessaria**: anche se $\overline{R}_{uu} \succ 0$, La $\overline{R}$ **potrebbe comunque non essere invertibile** per **ragioni di non identificabilità strutturale**, per le quali il minimo di $\overline{J}(\boldsymbol{\theta})$ non è unico
- 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

Esempi di segnali:

- **Rumore bianco:** $u(t) \sim WN(0, \lambda^2)$

$$\underset{i \times i}{\bar{R}_{uu}^{(i)}} = \begin{bmatrix} \gamma_{uu}(0) & \gamma_{uu}(1) & \cdots & \gamma_{uu}(i-1) \\ \gamma_{uu}(1) & \gamma_{uu}(0) & \cdots & \gamma_{uu}(i-2) \\ \cdots & \cdots & \cdots & \cdots \\ \gamma_{uu}(i-1) & \gamma_{uu}(i-2) & \cdots & \gamma_{uu}(0) \end{bmatrix} = \begin{bmatrix} \lambda^2 & 0 & 0 & \cdots & 0 \\ 0 & \lambda^2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 0 & \lambda^2 \end{bmatrix} = \lambda^2 \cdot I_i \succ 0$$

  Un **white noise** è un segnale persistentemente eccitante di ordine ∞. Se usiamo un WN per eccitare il sistema, i dati generati saranno molto informativi. Questo perché eccitiamo tutte le frequenze del sistema (con la stessa energia).

  un rumore bianco è un ottimo segnale di eccitazione. Nella pratica, però, **è impossibile generare un rumore bianco perfetto**:

  - le sequenze di numeri saranno pseudo-casuali, e non casuali
  - **a causa di limiti dell'elettronica** (capacitá parassite), il segnale generato e trasmesso agli attuatori sarà filtrato passa basso per cui non si avrà uno spettro perfettamente piatto
  - Inoltre, l'ampiezza del rumore bianco non è limitata. Talvolta, è necessario garantire che l'attuatore non saturi l'ingresso.

- **Pseudo-Random Binary Signal (PRBS)**
  - 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 $\left[-\bar{u}, +\bar{u}\right]$, il periodo e l'intervallo di clock.
  - Di solito il periodo viene posto uguale al numero di dati N che si vuole collezionare, e l'intervallo di clock a un tempo di campionamento
  - Quanto $T \to \infty$ il PRBS approssima un rumore bianco
  - Il PRBS è persistentemente eccitante di ordine T

- **Multiseno**
  - Il segnale multiseno è un segnale periodico, definito come una **media pesata di sinusoidi**, con frequenze multiple della risoluzione in frequenza della DFT $f_0 = \frac{f_s}{N}$

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

  - ll 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 (dopo un transitorio), come conseguenza del principio di sovrapposizione degli effetti $\Rightarrow$ di solito si scartano i primi periodi del segnale multiseno generato
  - 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

$$N = numero\ dati\ per\ periodo$$
$$P = numero\ di\ periodi\ del\ multiseno$$

---

10 Lezione 13: Identificazione - valutazione del modello

I metodi per scegliere la complessità del modello.

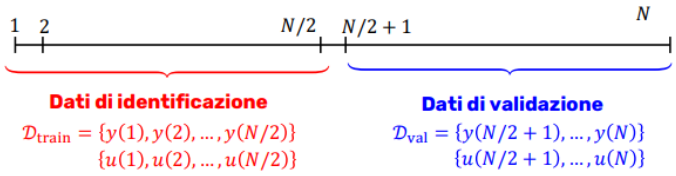
- Validazione e formule di complessità
- Analisi dei residui
- Analisi incertezza stima

Descrivere almeno due metodi alternativi per la selezione dell'ordine del modello dinamico nelle procedure di identificazione a minimizzazione dell'errore di predizione. Discutere quando conviene un metodo piuttosto che un altro.

Fissata la struttura di una famiglia di modelli $\mathcal{M}(\boldsymbol{\theta})$, dobbiamo scegliere la **complessitá del modello** (numero dei parametri).

Validazione

Un metodo semplice ma efficace consiste nell'identificare un insieme di modelli di diversa complessità utilizzando un dataset di **identificazione**, e confrontarne la bontá (esempio calcolo $J(\hat{\boldsymbol{\theta}}_N)$) su un dataset di **validazione**. Supponendo di avere N dati e li dividiamo in 2 sotto-sequenze:



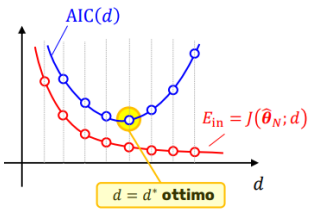
Per ogni ordine $m = 1, \dots, M$ si identifica un modello minimizzando $J(\boldsymbol{\theta}, \mathcal{D}_{train})$ e si calcola $J(\hat{\boldsymbol{\theta}}_{\frac{N}{2}}, \mathcal{D}_{val})$ sui dati di validazione infine si sceglie l'ordine m^* che minimizza $J(\hat{\boldsymbol{\theta}}_{\frac{N}{2}}, \mathcal{D}_{val})$.

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**, inoltre per utilizzare questo approccio sono neccessari molti dati.

Formule di complessità

In alternativa, **se i dati sono pochi** si possono usare le **formule di complessità** ottima: stimare d^*

- **Akaike Information Criterion** $AIC(d) = 2 \cdot \dfrac{d}{N} + ln\big[J(\hat{\boldsymbol{\theta}}_N; d)\big]$
- **Final prediction error** $FPE(d) = \dfrac{N+d}{N-d} \cdot J(\hat{\boldsymbol{\theta}}_N; d)$
- **Minimum Description Length** $MDL(d) = ln\big[N\big] \cdot \dfrac{d}{N} + ln\big[J(\hat{\boldsymbol{\theta}}_N; d)\big]$



Analisi residui identificazione modello dinamico per scelta della struttura e della complessità

Analisi dei residui

Date le seguenti ipotesi:

$$\mathcal{S}: \quad y(t) = G_0(z)u(t) + H_0(z)e(t)$$
$$\mathcal{M}(\boldsymbol{\theta}): \quad y(t) = G(z,\boldsymbol{\theta})u(t) + H(z,\boldsymbol{\theta})e(t)$$

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

Se $\exists \boldsymbol{\theta}^0$ t.c. $G(z,\boldsymbol{\theta}^0) = G_0(z)$ e $H(z,\boldsymbol{\theta}^0) = H_0(z)$, allora $\varepsilon_1(t;\boldsymbol{\theta}) = e(t)$, dove $e(t) \sim WN(0,\lambda^2)$.

La stima asintotica può quindi essere ottenuta come:

$$\bar{\boldsymbol{\theta}}_N = arg\ \min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \bar{J}(\boldsymbol{\theta}) = arg\ \min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{E}\big[\varepsilon_1(t;\boldsymbol{\theta})^2\big]$$

Dopo aver selezionato un modello $\mathcal{M}(\boldsymbol{\theta})$ e averne effettuato l'identificazione PEM, è possibile **validarne** (a-posteriori) la **struttura** e la **complessità** tramite analisi dei residui

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

- $\mathcal{S} \in \mathcal{M}(\boldsymbol{\theta})$
- $\mathcal{S} \notin \mathcal{M}(\boldsymbol{\theta})$ con $G_0(z) \in \mathcal{G}(\boldsymbol{\theta})$
- $\mathcal{S} \notin \mathcal{M}(\boldsymbol{\theta})$ con $G_0(z) \notin \mathcal{G}(\boldsymbol{\theta})$

Dove $\mathcal{G}$ è l'insieme dei modelli che descrivono la relazione input-output del sistema

Consideriamo il caso asintotico $N \to +\infty$,in cui $\hat{\boldsymbol{\theta}}_N \to \bar{\boldsymbol{\theta}}$, abbiamo che

$$\varepsilon_1(t;\bar{\boldsymbol{\theta}}) = H^{-1}(z;\bar{\boldsymbol{\theta}})\big(y(t) - G(z,\bar{\boldsymbol{\theta}})u(t)\big) = \frac{G_0(z) - G(z,\bar{\boldsymbol{\theta}})}{H(z,\bar{\boldsymbol{\theta}})}u(t) + \frac{H_0(z)}{H(z,\bar{\boldsymbol{\theta}})}e(t) \quad e(t) \sim WN(0,\lambda^2)$$

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

- la funzione di **autocovarianza dei residui** $\to \gamma_{\varepsilon\varepsilon}(\tau)$
- la funzione di **cross-covarianza trai i residui** ed il segnale di **ingresso** $\to \gamma_{\varepsilon u}(\tau)$

Sapendo che $\delta(\tau)$ è un delta di Dirac centrata in τ, possiamo incorrere in 3 situazioni:

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

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

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

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

- **Situazione B:** $\gamma_{\varepsilon\varepsilon}(\tau) \neq \lambda^2 \cdot \delta(\tau)$ e $\gamma_{\varepsilon u}(\tau) = 0\ \ \forall\tau$

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

Ovvero ***se e solo se*** $G(z,\bar{\boldsymbol{\theta}}) = G_0(z)$ *e* $H(z,\bar{\boldsymbol{\theta}}) \neq H_0(z)$
Questo avviene ***se e solo se*** $\mathcal{S} \notin \mathcal{M}(\boldsymbol{\theta})$ *con* $G_0(z) \in \mathcal{G}(\boldsymbol{\theta})$
per $\mathcal{M}(\boldsymbol{\theta})$ ***OE, BJ, FIR***

- **Situazione C:** $\gamma_{\varepsilon\varepsilon}(\tau) \neq \lambda^2 \cdot \delta(\tau)$ e $\exists\tau$ *t.c.* $\gamma_{\varepsilon u}(\tau) \neq 0$

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

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

Questo avviene **se e solo se** $\begin{cases} \mathcal{S} \notin \mathcal{M}(\boldsymbol{\theta}) \text{ con } G_0(z) \in \mathcal{G}(\boldsymbol{\theta}) \text{ per } \mathcal{M}(\boldsymbol{\theta}) \textbf{ ARX, ARMAX} \\ \mathcal{S} \notin \mathcal{M}(\boldsymbol{\theta}) \text{ con } G_0(z) \notin \mathcal{G}(\boldsymbol{\theta}) \end{cases}$

Considerando che $\mathcal{M}_1(\boldsymbol{\theta})$ è **OE, BJ o FIR** e $\mathcal{M}_2(\boldsymbol{\theta})$ è invece **ARX, ARMAX**.

A:	$\gamma_{\varepsilon\varepsilon}(\tau) = \lambda^2 \cdot \delta(\tau)$ e $\gamma_{\varepsilon u} = 0$ $\forall\tau$	$\mathcal{S} \in \mathcal{M}_1(\boldsymbol{\theta})$	$\mathcal{S} \in \mathcal{M}_2(\boldsymbol{\theta})$
B:	$\gamma_{\varepsilon\varepsilon}(\tau) \neq \lambda^2 \cdot \delta(\tau)$ e $\gamma_{\varepsilon u} = 0$ $\forall\tau$	$\mathcal{S} \notin \mathcal{M}_1(\boldsymbol{\theta}), G_0(z) \in \mathcal{G}(\boldsymbol{\theta})$	
C:	$\gamma_{\varepsilon\varepsilon}(\tau) \neq \lambda^2 \cdot \delta(\tau)$ e $\exists\tau$ $t.c.$ $\gamma_{\varepsilon u} \neq 0$	$\mathcal{S} \notin \mathcal{M}_1(\boldsymbol{\theta}), G_0(z) \notin \mathcal{G}(\boldsymbol{\theta})$	$\mathcal{S} \notin \mathcal{M}_2(\boldsymbol{\theta}), G_0(z) \notin$ o $\in \mathcal{G}(\boldsymbol{\theta})$?

- **Situazione A:** $\gamma_{\varepsilon\varepsilon}(\tau) = \lambda^2 \cdot \delta(\tau)$ e $\gamma_{\varepsilon u}(\tau) = 0\ \forall\tau$
- **Situazione B:** $\gamma_{\varepsilon\varepsilon}(\tau) \neq \lambda^2 \cdot \delta(\tau)$ e $\gamma_{\varepsilon u}(\tau) = 0\ \forall\tau$
- **Situazione C:** $\gamma_{\varepsilon\varepsilon}(\tau) \neq \lambda^2 \cdot \delta(\tau)$ e $\exists\tau$ t.c. $\gamma_{\varepsilon u}(\tau) \neq 0$

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

Una validazione della struttura che si conclude con un successo non **garantisce che** $G(z, \hat{\boldsymbol{\theta}}_N)$ e $H(z, \hat{\boldsymbol{\theta}}_N)$ siano buone stime di $G_0(z)$ e $H_0(z)$. È necessario controllare anche la varianza delle stime (sia dei parametri sia delle funzioni di trasferimento).

Descrivere tre modi per valutare la bontà della stima (identificazione) su un modello dinamico.

La bontà di un modello può essere valutata:

- Analizzando i residui (cioè gli errori di predizione a un passo), meglio con dati di validazione
- Analisi incertezza stima: parametri, funzione di trasferimento e posizione dei poli e degli zeri
- Confrontando l'uscita simulata o predetta con l'uscita misurata, su dati di validazione, e calcolandone un indicatore di **FIT**
- Confronto con stima nonparametrica

Analisi dei residui

Vedere precedentemente, nel caso diminuire la roba da scrivere.

Analisi incertezza stima

Con l'analisi dell'incertezza della stima intendiamo:

- **Incertezza sulla stima dei parametri:**
  Si verifica la significatività statistica di un parametro, dove per significatività intendiamo quanto è probabile che il parametro vero sia effettivamente diverso da 0.
- **Incertezza sulla stima delle funzioni di trasferimento**
  Nel caso $\mathcal{S} \in \mathcal{M}(\boldsymbol{\theta})$, l'espressione della varianza sulla stima del valore della funzione di trasferimento $G(z, \hat{\boldsymbol{\theta}}_N)$ ad ogni frequenza $z = e^{j\omega}$ e assumendo che $u(t) \perp e(t)$, puó essere approssimata come:

$$Var\left[G(e^{j\omega}, \hat{\boldsymbol{\theta}}_N)\right] \approx \frac{n}{N} \cdot \frac{\Gamma_{vv}(\omega)}{\Gamma_{uu}(\omega)}$$

- n : *n variabili di stato di G (ordine)*
- $\Gamma_{vv}(\omega)$: $v(t) = H_0(z)e(t)$ *disturbo*

  Se $\mathcal{S} \notin \mathcal{M}(\boldsymbol{\theta})$, esistono altre espressioni più complesse.

  Quando $\sqrt{Var\left[G(e^{j\omega}, \hat{\boldsymbol{\theta}}_N)\right]}$ puó essere considerata piccola? La risposta dipende dall'uso del modello. Se è considerata troppo grande, non possiamo garantire che $G(e^{j\omega}, \hat{\boldsymbol{\theta}}_N)$ sia una buona stima di $G_0(z)$.

- **Incertezza sulla posizione dei poli e degli zeri**
  Una volta stimato il modello e analizzato i residui per capire se $\mathcal{S} \in \mathcal{M}(\boldsymbol{\theta})$, è utile rappresentare i poli e gli zeri del modello stimato, per verificare la possibilità di **cancellazioni** polo\zero **(e quindi evitare sovraparametrizzazioni)**.

  Si verificano se gli ellissoidi di confidenza di poli e zeri si sovrappongono o sono molto vicini. In questo caso, è probabile che tali poli\zeri possano essere rimossi dal modello.

Simulazione e predizione del modello

Si simula o predice l'output del modello per confrontarlo con l'output misurato, a fronte del medesimo input. L'idea è che se la simulazione o predizione sono simili all'output misurato, allora il modello è buono.

L'errore di simulazione é $e_{sim}(t)$. Un modello buono non rende necessariamente $e_{sim}(t)$ piccolo, poiché la simulazione non considera il modello del rumore.

$$e_{sim} = y(t) - \widehat{y}_{sim}(t) = y(t) - G(z, \widehat{\boldsymbol{\theta}}_N)u(t)$$

La simulazione può solo fornire informazioni sull'accuratezza di $G(z, \widehat{\boldsymbol{\theta}}_N)$. Inoltre, la simulazione è una rappresentazione più realistica di come il modello si comporta a fronte di un ingresso noto $u(t)$.

La simulazione permette anche di stimare il rumore $v(t) = H_0(z) \cdot e(t)$ tramite:

$$\widehat{v}(t) = y(t) - \widehat{y}_{sim}(t)$$

Confronto con stima nonparametrica

La stima nonparametrica lascia parlare i dati. È quindi importante valutare se le funzioni di trasferimento del modello aderiscono alle loro stime nonparametriche. Stime nonparametriche delle funzioni $G_0(z)$ e $H_0(z)$ si possono ottenere con la ETFE o con la stima spettrale.
