



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



IDENTIFICAZIONE DEI MODELLI E ANALISI DEI DATI (IMAD)

Setup dell'ambiente di sviluppo codice

**Corso di Laurea Magistrale in
INGEGNERIA INFORMATICA**

SPEAKER
Prof. Mirko Mazzoleni

PLACE
Università degli Studi di
Bergamo

Outline

1. Installare Matlab
2. Installare la distribuzione Anaconda Python
3. Git e Github



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

Outline

1. Installare Matlab

2. Installare la distribuzione Anaconda Python

3. Git e Github



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

Installare Matlab

Durante il corso, usaremos Matlab (principalmente) e Python (quando necessario)

Il codice per gli esercizi sarà disponibile, di volta in volta, su **MS Teams**

Installare Matlab seguendo le istruzioni date dall'università, usando l'account universitario



Outline

1. Installare Matlab

2. Installare la distribuzione Anaconda Python

3. Git e Github



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

Installare Python 3

Anaconda è un insieme di librerie Python (distribuzione) che sono garantite funzionare bene insieme <https://www.anaconda.com/distribution/#download-section>



Individual Edition

Your data science toolkit

With over 25 million users worldwide, the open-source Individual Edition (Distribution) is the easiest way to perform Python/R data science and machine learning on a single machine. Developed for solo practitioners, it is the toolkit that equips you to work with thousands of open-source packages and libraries.

Anaconda Individual Edition

[Download](#)

For Windows
Python 3.8 • 64-Bit Graphical Installer • 477 MB

[Get Additional Installers](#)

| |



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

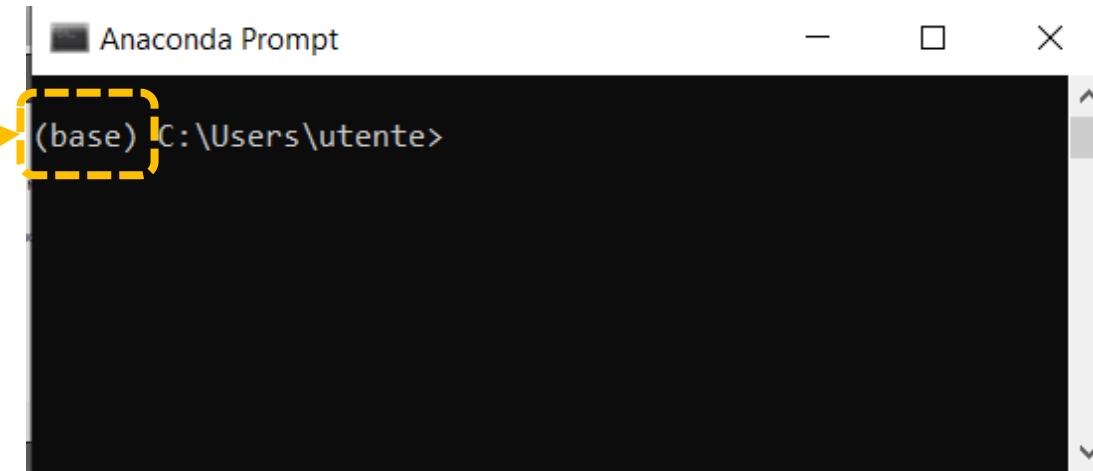
Installare Python 3

Anaconda installa molte cose:

- **Anaconda prompt:** questa è una shell dove si possono inserire comandi (es. gestire le librerie, avviare jupyter notebooks, gestire gli environments)

✓ La parola **(base)** rappresenta

l'ambiente attuale usato



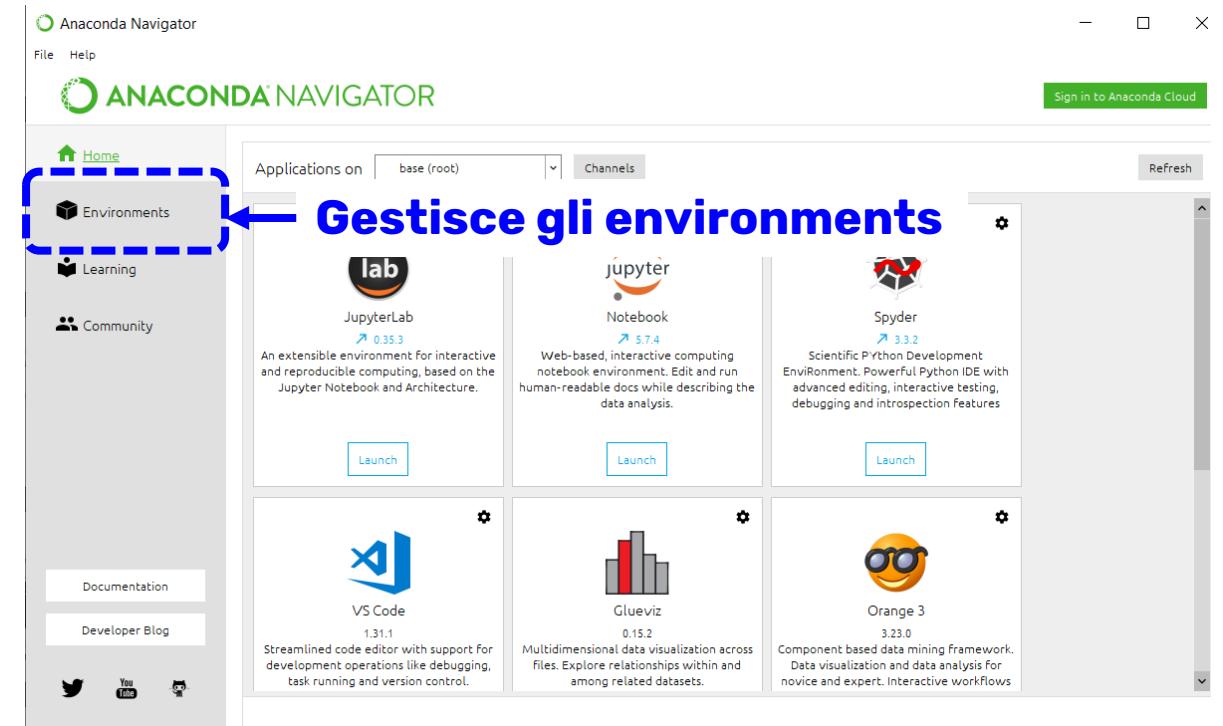
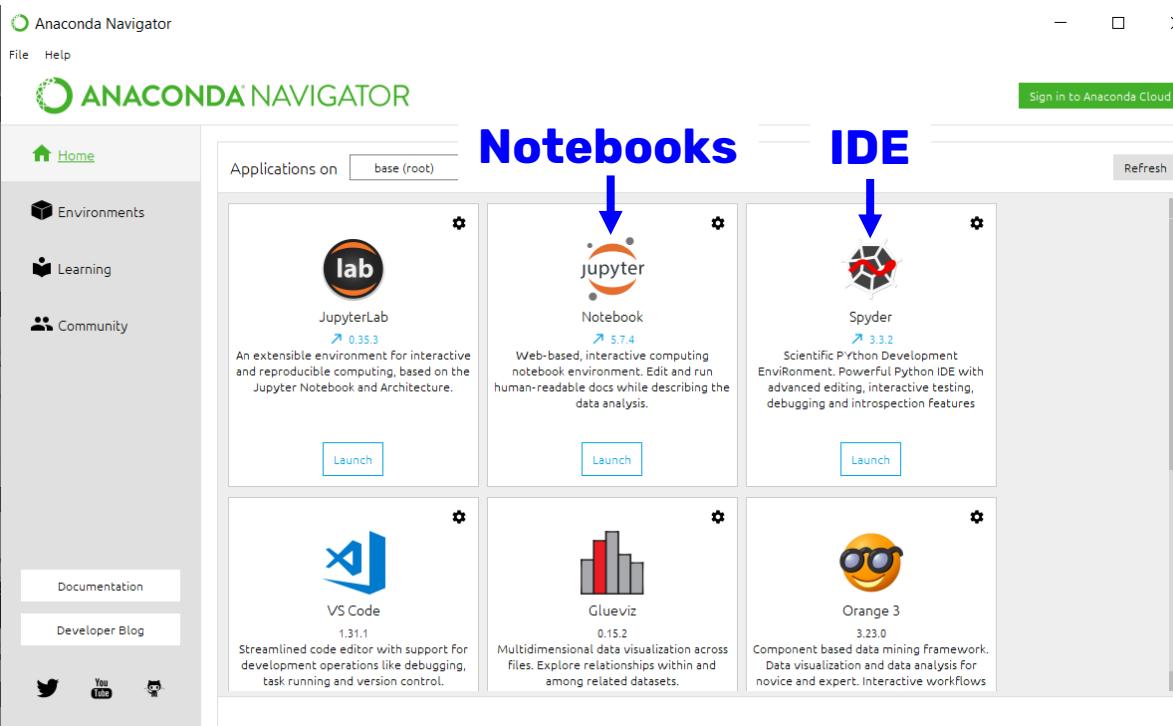
È utile attivare i ambienti appena creati, diversi per ogni progetto

Un ambiente è un insieme di librerie attualmente utilizzate (caricate). Si dovrebbero avere più ambienti, per evitare che l'aggiornamento di una libreria generi conflitti con altre librerie



Installare Python 3

- **Anaconda Navigator:** è una GUI in cui è possibile eseguire programmi che utilizzano Python (Notebook, IDE,...)



Comandi utili

(https://docs.conda.io/projects/conda/en/4.6.0/_downloads/52a95608c49671267e40c689e0bc00ca/conda-cheatsheet.pdf)

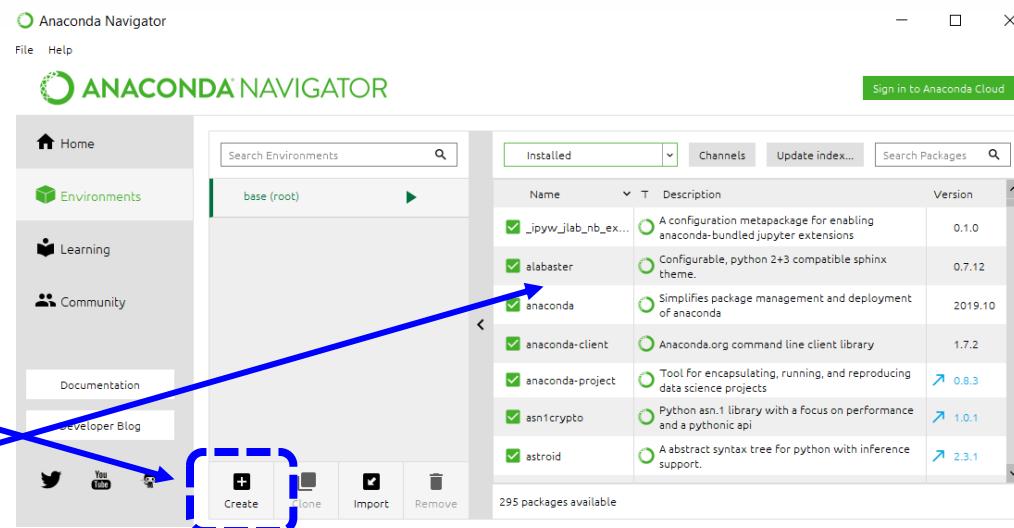
- **Create environment:** conda create --name [name_environment] python=3.7
- **Create environment from file:** conda env create -f [name_environment]
- **Delete environment:** conda env remove --name [name_environment]
- **Activate environment:** conda activate [name_environment]
 - ✓ WINDOWS: conda activate [name_environment]
 - ✓ LINUX, macOS: source conda activate [name_environment]
- **Install library:** conda install [name_library]
- **Get a list of all my environments (active environment is shown with *):** conda env list
- **List all packages and versions installed in active environment:** conda list



Comandi utili

Per creare un ambiente e lanciare un notebook:

1. Aprire Anaconda Navigator → Environments → Crea
2. Selezionare le librerie da installare nel nuovo environment
3. Home → selezionare il nuovo environment → installare Jupyter → lanciare Jupyter



Un modo alternativo per avviare Jupyter (una volta installato) è il seguente :

- Aprire Anaconda Prompt
- cd (change directory) nella directory in cui hai il notebook
- Activare l'environement conda activate [name_environment]
- Lanciare il comando jupyter notebook



Outline

1. Installare Matlab
2. Installare la distribuzione Anaconda Python

3. Git e Github



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

Git and Github

Git è un software che ti consente di eseguire il **controllo di versione** (tenere traccia delle modifiche apportate ai file ed eventualmente ripristina una versione precedente del file). Un insieme di file e cartelle di versionati è chiamato **repository** (o «repo»)

Molto utile se un team di sviluppatori lavora allo **stesso progetto** (ma è utile anche se si lavora da soli) Il software gestisce automaticamente gli aggiornamenti di ogni sviluppatore e unisce le modifiche in un unico file

Mi aspetto che tu usiate git (e, eventualmente, **Github**)

Git viene eseguito dalla riga di comando. Si può usare **Github Desktop**, un software intuitivo che usa git e github con una GUI



Git and Github

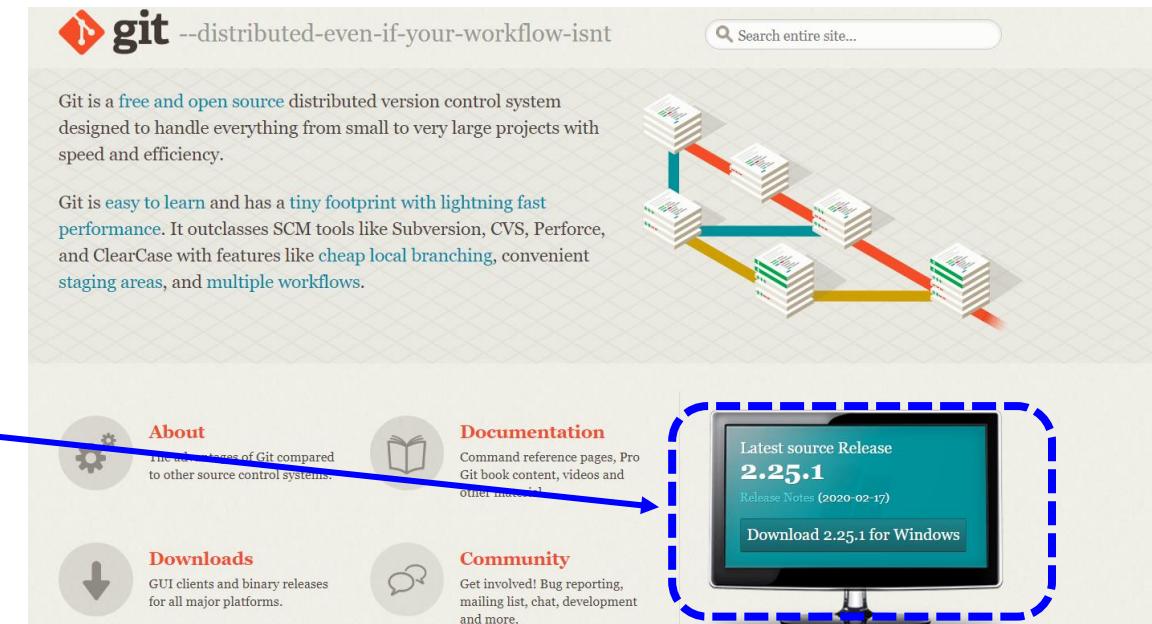
Come installare Git

- Andare a <https://git-scm.com/>
- Installare (usate i settings di default)

Terminologia di Git

Git permette di fare molte operazioni. I seguenti comandi sono utili (oppure usare Github Desktop)

- **Commit:** salva e archivia le modifiche ai file
- **Ripristina:** torna alla versione precedentemente impegnata
- **Push:** invia i file impegnati a un server (ad esempio l'account Github)
- **Pull:** recupera i file inviati da un server (ad esempio l'account Github)
- **Clona:** copia un repository da un server (ad esempio l'account Github)



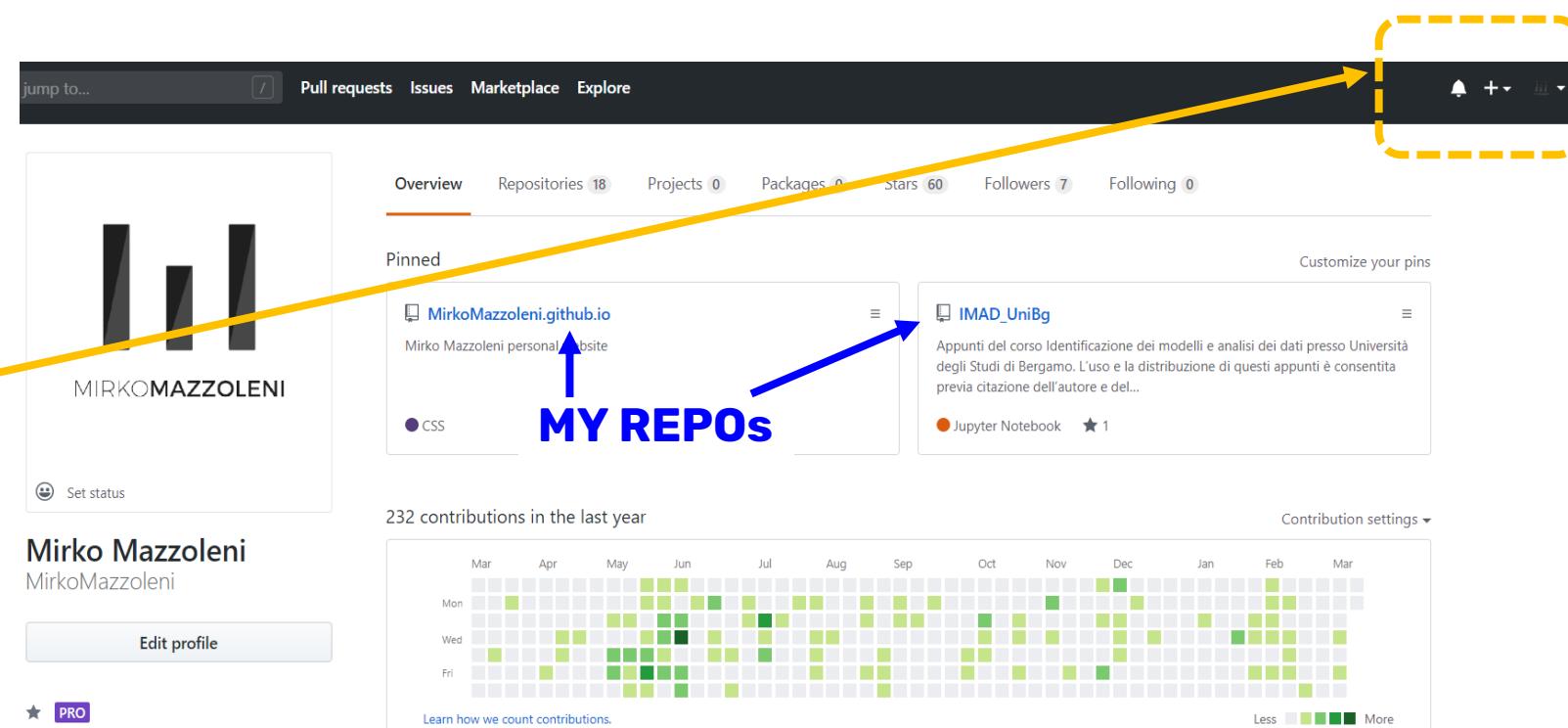
Git and Github

GitHub, Inc. è un servizio di hosting basato sul Web che fornisce hosting per il controllo della versione dello sviluppo software utilizzando Git. Consente il repository (pubblico) gratuito

Per studenti c'è anche **Github Education** che offre repository privati

Come setappare Github

- Andare a <https://github.com/>
- Creare un account
- Creare un nuovo repo
- Clonarlo su PC in locale
- Popolarlo
- Pushare gli updates



Git and Github

Inserire il nome del repository

Andare sul repo → Clone or download → Copiare il link

Questo link può essere usato in Github Desktop
per clonare il repository sul computer locale

Create a new repository

A repository contains all project files, including the revision history. Already have a project repository elsewhere?

[Import a repository.](#)

Owner

Repository name *

Great repository names are short and memorable. Need inspiration? How about [legendary-robot](#)?

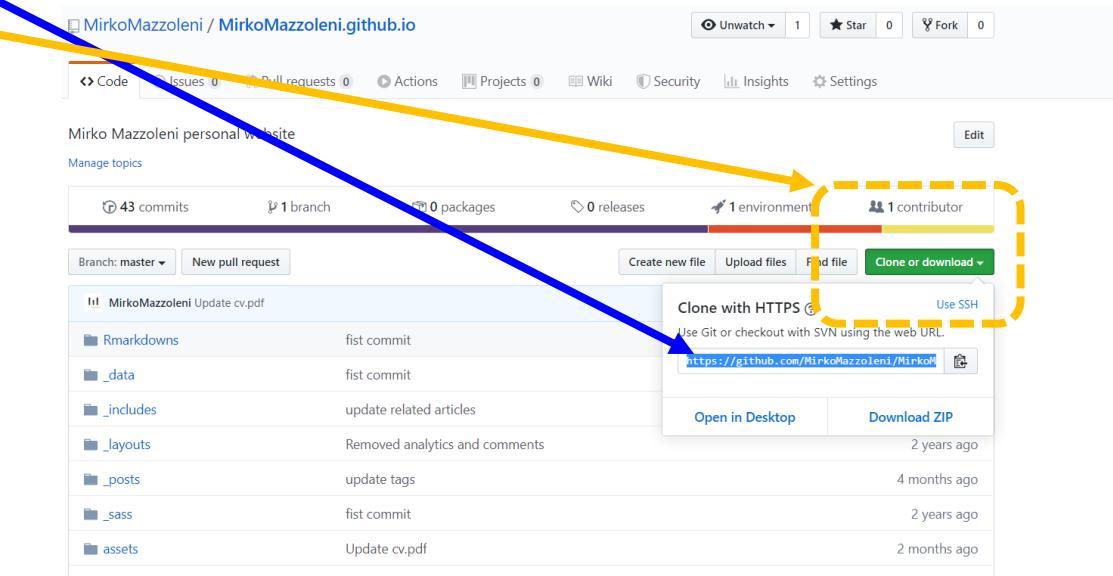
Description (optional)

Public

Anyone can see this repository. You choose who can commit.

Private

You choose who can see and commit to this repository.



MirkoMazzoleni / MirkoMazzoleni.github.io

Unwatch 1 Star 0 Fork 0

Code Issues 0 Pull requests 0 Actions Projects 0 Wiki Security Insights Settings

Mirko Mazzoleni personal website

Manage topics

43 commits 1 branch 0 packages 0 releases 1 environment 1 contributor

Branch: master New pull request

Clone with HTTPS Use SSH Use Git or checkout with SVN using the web URL.
<https://github.com/MirkoMazzoleni/MirkoMazzoleni.github.io>

Open in Desktop Download ZIP

File	Commit	Date
MirkoMazzoleni Update cv.pdf	fist commit	2 years ago
_Markdowns	fist commit	4 months ago
_data	update related articles	2 years ago
_includes	Removed analytics and comments	4 months ago
_layouts	update tags	2 years ago
_posts	fist commit	2 months ago
_sass	Update cv.pdf	2 months ago
assets		2 months ago



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

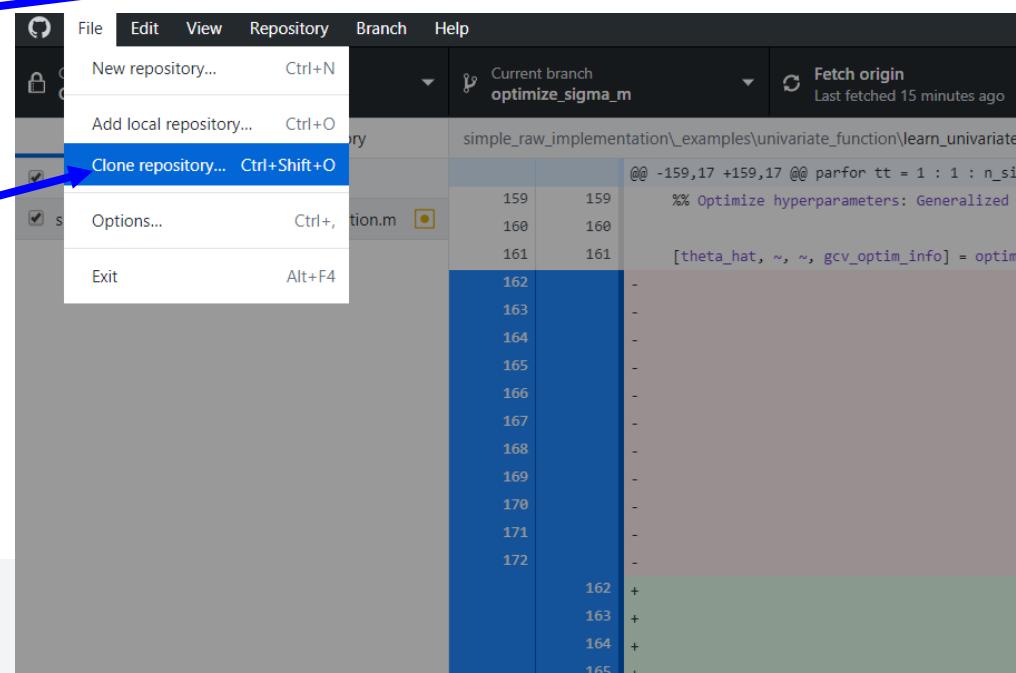
Github desktop

GitHub desktop è un software che semplifica l'uso di git con github (è necessario disporre di un account Github per usarlo)



Come setappare Github Desktop

- Andare a <https://desktop.github.com/>
- Scaricare e installare
- Clonare il repository





UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



IDENTIFICAZIONE DEI MODELLI E ANALISI DEI DATI (IMAD)

Lezione 1: Introduzione al corso

**Corso di Laurea Magistrale in
INGEGNERIA INFORMATICA**

SPEAKER
Prof. Mirko Mazzoleni

PLACE
Università degli Studi di
Bergamo

Chi sono

- **Name:** Mirko Mazzoleni
- **Attualmente:** Assistant Professor (RTD-B), **Laboratorio di Automatica (CAL UniBG)**
 - ✓ *Ricerca:* System identification, machine learning, fault diagnosis
 - ✓ *Didattica:* 1. Identificazione dei Modelli e Analisi dei Dati (IMAD) – LM Ingegneria Informatica
2. Adaptive learning, estimation and supervision of dynamical systems (ALES)
3. Gestione, analisi e rappresentazione dei dati (GARD) – LT Ingegneria Gestionale
- **Altro:** Co-fondatore AISENT srl startup <https://aisent.io/>  AISENT

• Contatti

- ✓ mirko.mazzoleni@unibg.it 
- ✓ <https://mirkomazzoleni.github.io/> 
- ✓ <http://cal.unibg.it/> **Sito laboratorio CAL UniBG** 
- ✓ <https://www.facebook.com/ControlAutomationLabUnibg/> 



Control and Automation Laboratory (CAL) @ University of Bergamo, Italy

Personale:

- **6 professori**
- **1-3 visiting professors**
- **6 studenti di dottorato**

- **Didattica** (control systems)
- **Ricerca** (fault diagnosis, identification, MPC)
- **Progetti industriali** (manufacturing, aerospace, packaging,...)



Outline

1. Presentazione del corso di Identificazione dei Modelli e Analisi dei Dati (IMAD)
2. Introduzione e motivazione
3. La stima di un modello dai dati: l'approccio supervisionato
4. Sistemi (e modelli) statici
5. Sistemi (e modelli) dinamici
6. Riassunto



Outline

- 1. Presentazione del corso di Identificazione dei Modelli e Analisi dei Dati**
2. Introduzione e motivazione
3. La stima di un modello dai dati: l'approccio supervisionato
4. Sistemi (e modelli) statici
5. Sistemi (e modelli) dinamici
6. Riassunto



Prerequisiti del corso

È **caldamente consigliato** avere delle buone basi delle seguenti materie

- Fondamenti di automatica \implies Il corso di IMAD è del ssd ING-INF\04 – AUTOMATICA !
- Algebra lineare
- Analisi 1 e Analisi 2
- Statistica

Come «rinfrescare» i prerequisiti?

- Fondamenti di automatica \implies Corso UniBg
- Algebra lineare \implies Corso UniBg, [corso Gilbert Strang @MIT](#) su YouTube
- Analisi 1 e Analisi 2 \implies Corso UniBg
- Statistica \implies Corso UniBg, prima parte del libro «*Doing Bayesian Data Analysis*»



Esame

- Esame **scritto** da **2 ore**
- **3 esercizi** numerici + **3 domande** aperte di teoria
- Vedere «**tema d'esame di esempio**» sul sito del corso

Come prepararsi all'esame?

- Seguire le lezioni e le esercitazioni
- Studiare la teoria
- Rifare le esercitazioni
- Fare esercitazioni aggiuntive



Bittanti Sergio, Campi Marco, **Raccolta di Problemi di Identificazione, Filtraggio, Controllo predittivo.**
Pitagora Editrice, Bologna (2013)



Obiettivi formativi del corso

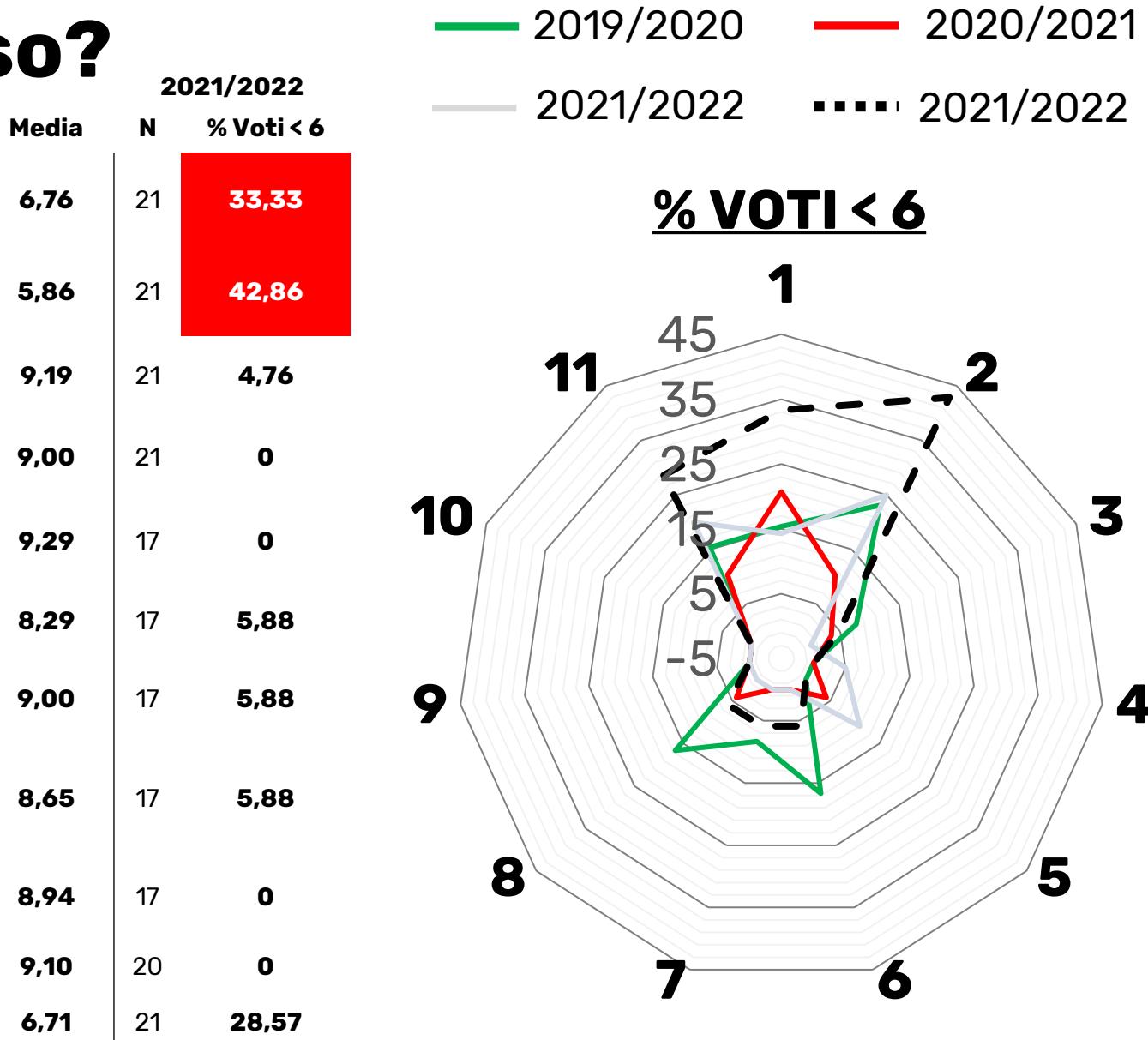
Alla fine del corso, dovrete essere in grado di:

- **Applicare** diverse metodologie di stima, a seconda della domanda a cui si vuol fornire una risposta tramite l'analisi del dato
- **Formulare** un problema di stima, individuando le variabili del problema (e.g. dati di input e output)
- **Stimare** un modello **statico** o **dinamico** dai dati, attraverso la risoluzione di un problema di ottimizzazione
- **Scegliere** il modello più opportuno per la tipologia di dati a disposizione
- **Valutare** la bontà del modello stimato dai dati



Cosa aspettarsi dal corso?

Le conoscenze preliminari possedute sono risultate sufficienti per la comprensione degli argomenti previsti nel programma d'esame?	
Il carico di studio dell'insegnamento è proporzionato ai crediti assegnati?	
Il materiale didattico (indicato e disponibile) è adeguato per lo studio della materia?	
Le modalità di esame sono state definite in modo chiaro?	
Gli orari di svolgimento di lezioni, esercitazioni e altre eventuali attività didattiche sono rispettati?	
Il docente stimola/motiva l'interesse verso la disciplina?	
Il docente espone gli argomenti in modo chiaro?	
Le attività didattiche integrative come esercitazioni, tutorati, laboratori, ecc. (non sono compresi gli addestramenti linguistici) ove esistenti, sono utili all'apprendimento della materia?	
L'insegnamento è stato svolto in maniera coerente con quanto dichiarato sul sito Web del corso di studio?	
Il docente è reperibile per chiarimenti e spiegazioni?	
E' interessato/a agli argomenti trattati nell'insegnamento?	



Materiali didattici

Materiali forniti dal docente

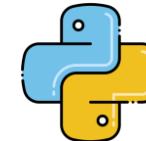
- Slide e appunti delle lezioni



- Pdf delle lezioni e delle esercitazioni



- Codice Matlab\Simulink o Python



Interazione e feedback

- Durante le lezioni ci saranno dei quiz
- Durante la settimana vi darò delle attività da fare e dei test a cui rispondere. Sono **facoltativi** ma vi aiutano a capire il grado di apprendimento prima dell'esame. Inoltre, contribuiranno a dare un **bonus di +3 punti** al voto finale

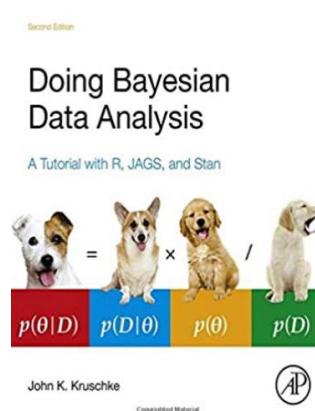
Useremo le **attività** di **MS Teams**



Materiali didattici

Libri consigliati

- G. James, D. Witten, T. Hastie, R. Tibshirani, **Introduzione all'apprendimento statistico con applicazioni in R**, Piccin (2020)
- John K. Kruschke, **Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan.** Academic Press (2014)



- Bittanti Sergio, **Teoria della predizione e del filtraggio**, Pitagora Editrice, Bologna (2003)



- Bittanti Sergio, **Identificazione dei Modelli e Sistemi Adattativi**, Pitagora Editrice, Bologna (2003)



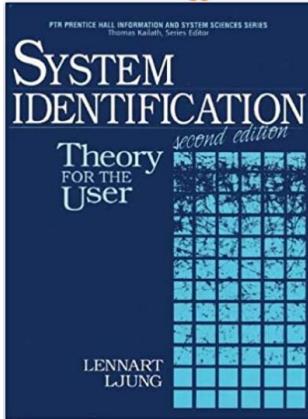
- Bittanti Sergio, Campi Marco, **Raccolta di Problemi di Identificazione, Filtraggio, Controllo predittivo**. Pitagora Editrice, Bologna (2013)



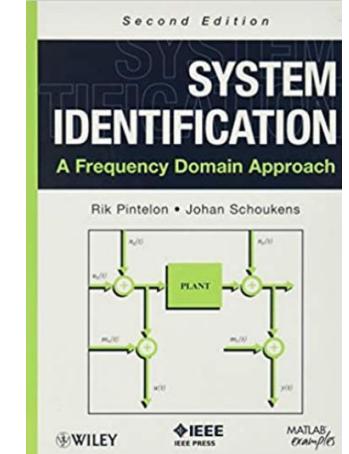
Materiali didattici

Libri avanzati di approfondimento

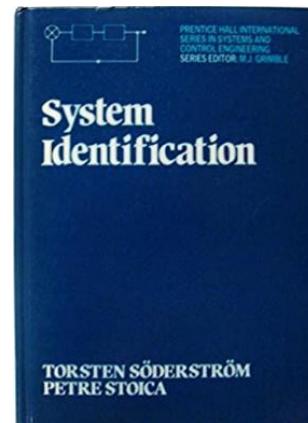
- Lennart Ljung, **System Identification: Theory for the User**, Pearson (1998)



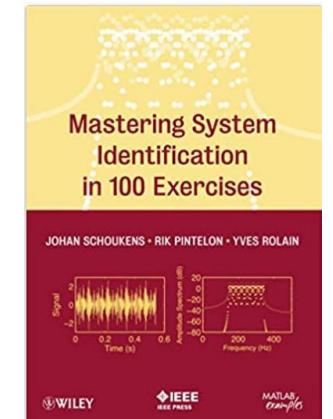
- Rik Pintelon, Johan Schoukens, **System Identification: A Frequency Domain Approach**, IEEE (2012)



- Torsten Soderstrom, Petre Stoica, **System identification**, Prentice Hall international (2001)



- Johan Schoukens, Rik Pintelon, Yves Rolain, **Mastering System Identification in 100 Exercises**, IEEE (2003)



Syllabus

Parte I: sistemi statici

1. Richiami di statistica

2.1 Proprietà degli stimatori

3. Stima a minimi quadrati

3.1 Stima di modelli lineari

3.2 Algoritmo del gradient descent

4. Stima a massima verosimiglianza

4.1 Proprietà della stima

4.2 Stima di modelli lineari

5. Regressione logistica

5.1 Stima di un modello di regressione logistica

6. Fondamenti di machine learning

6.1 Bias-Variance tradeoff

6.2 Overfitting

6.3 Regolarizzazione

6.4 Validazione

7. Cenni di stima Bayesiana

7.1 Probabilità congiunte, marginali e condizionate

7.2 Connessione con Filtro di Kalman



Syllabus

Parte II: sistemi dinamici

8. Processi stocastici

- 8.1 Processi stocastici stazionari (pss)
- 8.3 Rappresentazione spettrale di un pss
- 8.4 Stimatori campionari media\covarianza
- 8.5 Densità spettrale campionaria

9. Famiglie di modelli a spettro razionale

- 9.1 Modelli per serie temporali (MA, AR, ARMA)
- 9.2 Modelli per sistemi input/output (ARX, ARMAX)

10. Predizione

- 10.1 Filtro passa-tutto

10.2 Forma canonica

10.3 Teorema della fattorizzazione spettrale

10.4 Soluzione al problema della predizione

11. Identificazione

- 11.3 Identificazione di modelli ARX
- 11.4 Identificazione di modelli ARMAX
- 11.5 Metodo di Newton

12. Identificazione: analisi e complementi

- 12.1 Analisi asintotica metodi PEM
- 12.2 Identificabilità dei modelli
- 12.3 Valutazione dell'incertezza di stima

13. Identificazione: valutazione



Parte I: sistemi statici**Parte II: sistemi dinamici**Stima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

✓ Stima parametri popolazione

✓ Stima modello lineare: minimi quadrati

○ **SI assunzioni su ddp dei dati**

✓ Stima massima verosimiglianza parametri popolazione

✓ Stima modello lineare: massiva verosimiglianza

✓ Regressione logistica

• θ variabile casuale○ **SI assunzioni su ddp dei dati**

✓ Stima Bayesiana

Stima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

✓ Modelli lineari di pss

✓ Predizione

✓ Identificazione

✓ Persistente eccitazione

✓ Analisi asintotica metodi PEM

✓ Analisi incertezza stima (numero dati finito)

✓ Valutazione del modello

Machine learning

Outline

1. Presentazione del corso di Identificazione dei Modelli e Analisi dei Dati

2. Introduzione e motivazione

3. La stima di un modello dai dati: l'approccio supervisionato

4. Sistemi (e modelli) statici

5. Sistemi (e modelli) dinamici

6. Riassunto



Identificazione dei Modelli e Analisi dei Dati

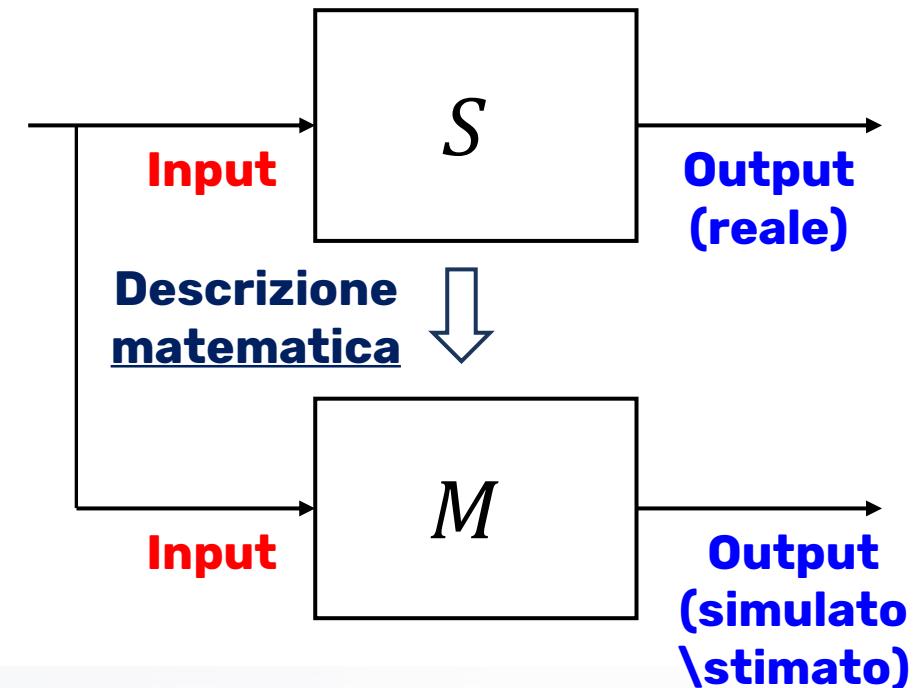
A) IDENTIFICAZIONE DEI MODELLI

In questo corso parleremo di **modelli matematici** per descrivere **fenomeni** o **sistemi**

- **Sistema:** meccanismo astratto che trasforma **inputs** (cause) in **outputs** (effetti)
- **Modello:** descrizione **matematica** di un sistema

Esempi di sistemi:

- **economici:** relazione tra reddito ed educazione
- **sociali:** relazione tra luogo di abitazione e criminalità
- **fisici:** relazione tra corrente e tensione



Identificazione dei Modelli e Analisi dei Dati

Vi sono tre approcci fondamentali per definire un modello M di un sistema S

1) Modellorazione white-box:** basato su **leggi** e **principi** della fisica o conoscenza a priori**

Esempio: sistema massa-molla-smorzatore

- Scrivo le **equazioni** in base alla **fisica** del sistema
- **Conosco/misuro** direttamente i parametri m, c, k

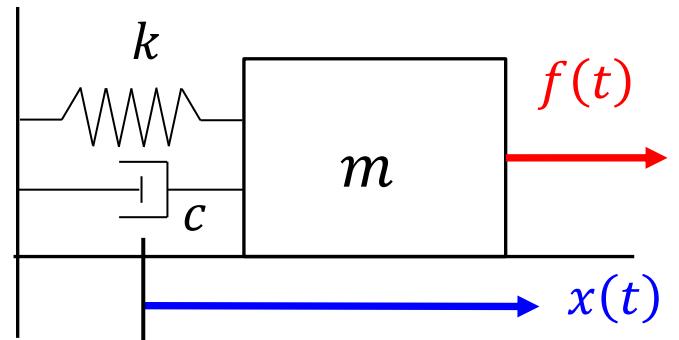
Vantaggi:

- Conoscenza del significato fisico delle variabili
- Modello generalizzabile

Svantaggi:

- Richiede di conoscere tutte le leggi e il valore dei parametri del problema specifico
 - Approccio che richiede tempo e costi
 - Non fattibile nel caso di sistemi complessi, con molti componenti

$$m\ddot{x}(t) = f(t) - c\dot{x}(t) - kx(t)$$



Identificazione dei Modelli e Analisi dei Dati

2) Modellazione black-box: basata su dati sperimentali

Esempio: sistema massa-molla-smorzatore

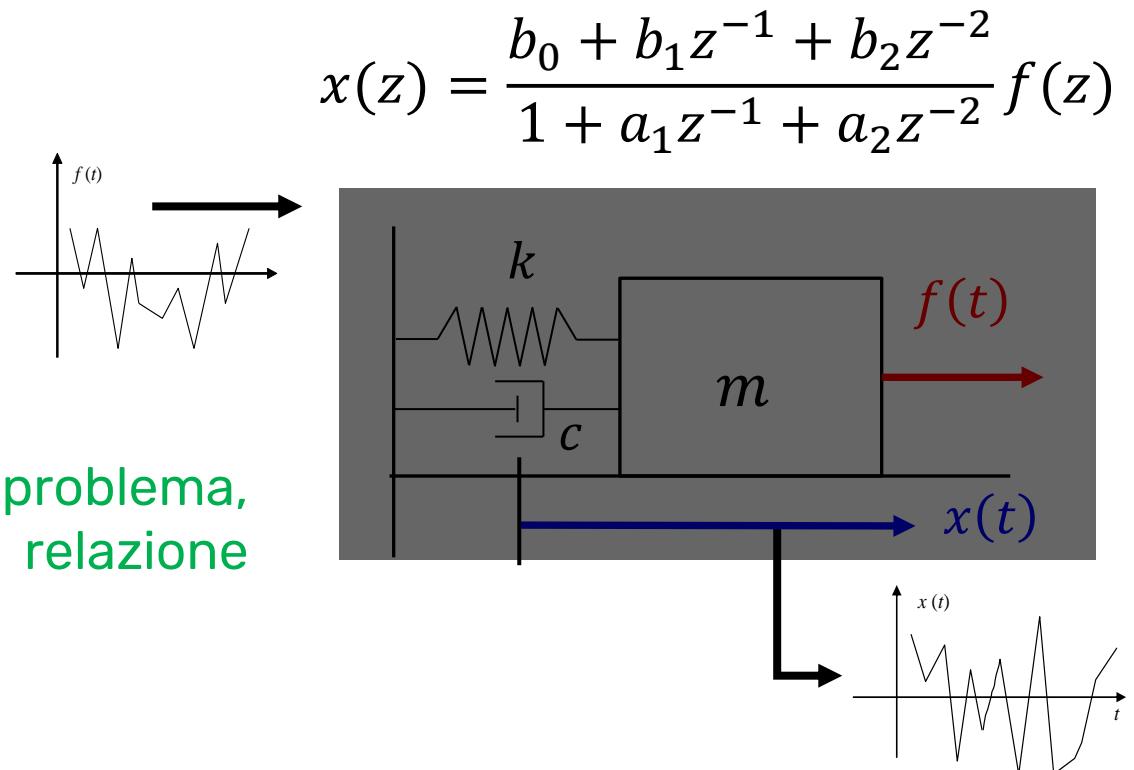
- Faccio un esperimento I/O
- Identifico (stimo) i parametri di un **modello (digitale)** generico di ordine adeguato

Vantaggi:

- Prescindono dal particolare problema, limitandosi a caratterizzare la relazione ingresso-uscita
- Veloci da costruire

Svantaggi:

- Non interpretabili fisicamente
- Non generali: se il sistema cambia, devo ripetere l'esperimento



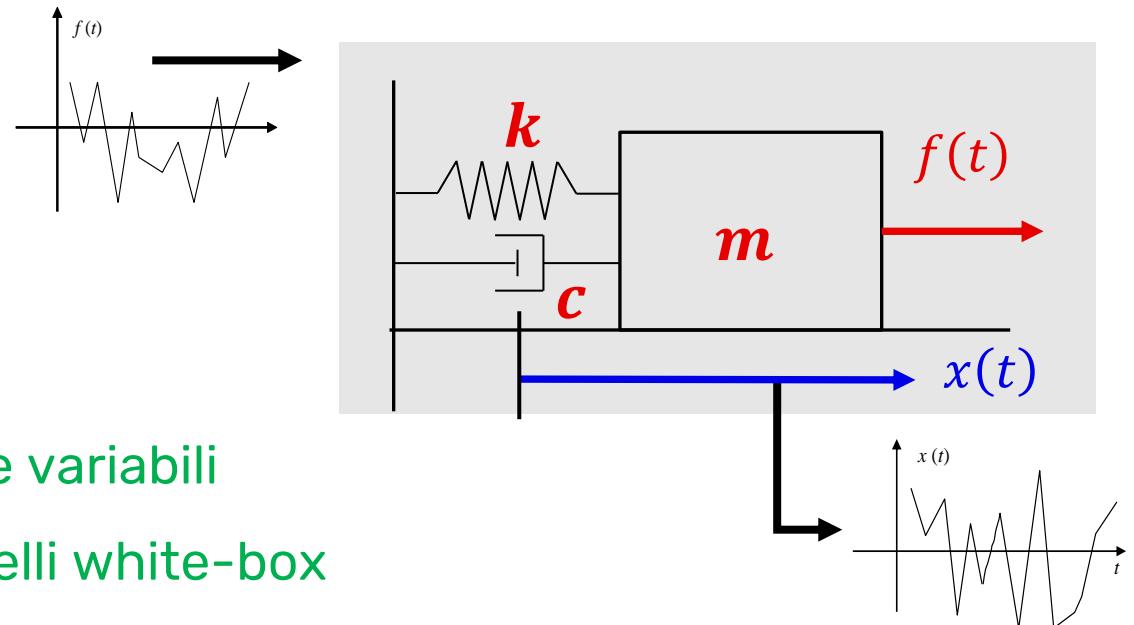
Identificazione dei Modelli e Analisi dei Dati

3) Modellazione gray-box: basata su dati sperimentali

Esempio: sistema massa-molla-smorzatore

- Conosco le equazioni del sistema
- Faccio un esperimento I/O
- Identifico (tutti o alcuni) i **parametri (fisici)** di un **modello fisico**

$$m\ddot{x}(t) = f(t) - c\dot{x}(t) - kx(t)$$



Vantaggi:

- Conoscenza del significato fisico delle variabili
- Più veloci da costruire rispetto a modelli white-box

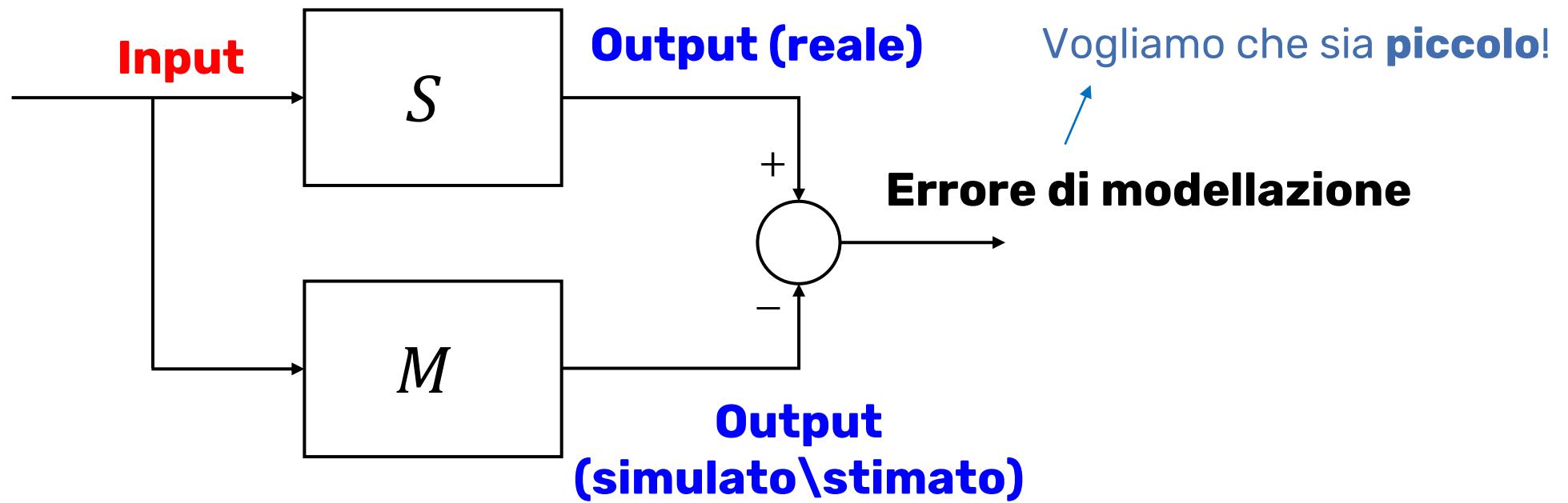
Svantaggi:

- Più lenti da costruire rispetto a modelli black-box



Identificazione dei Modelli e Analisi dei Dati

Come faccio a sapere se un modello è «buono»?



Se gli output **reali (misurati)** e **simulati dal modello (calcolati)** sono simili, il modello è in grado di replicare il fenomeno reale

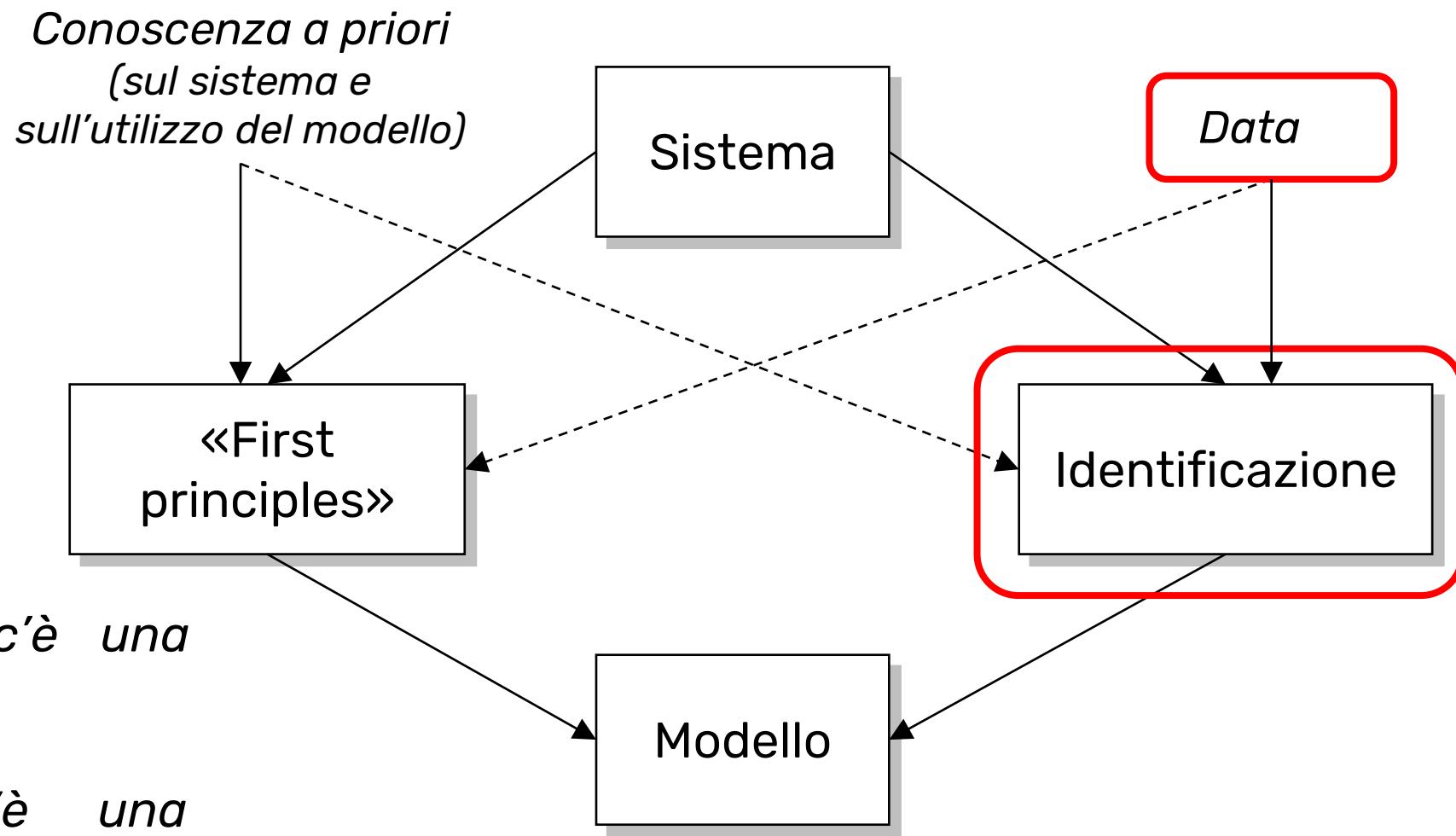


Identificazione dei Modelli e Analisi dei Dati

Questo corso si concentrerà sulla stima di modelli **black-box**

Considereremo:

- sistemi **statici** («non c'è una dipendenza dal tempo»)
- sistemi **dinamici** («c'è una dipendenza dal tempo»)



Identificazione dei Modelli e Analisi dei Dati

In conclusione, **Identificazione dei modelli** vuol dire **risolvere un problema di stima**

- in particolare, stima di un modello che descriva i dati

IDENTIFICAZIONE DEI MODELLI



PROBLEMA DI STIMA



Identificazione dei Modelli e Analisi dei Dati

B) ANALISI DEI DATI

Obiettivo 1: Determinare le caratteristiche statistiche dei dati e delle variabili misurate.

Essi sono affetti da **rumore** e **incertezza**

- Media
- Varianza
- Correlazione
- Distribuzione di probabilità



STATISTICA DESCRITTIVA

Obiettivo 2: Individuare delle **regolarità** (pattern) nei dati (se ci sono regolarità)

- I dati presentano dei «pattern» riconoscibili o sono random?
- Possiamo **allenare algoritmi** che, **da soli**, individuino questi pattern?



MACHINE LEARNING



Identificazione dei Modelli e Analisi dei Dati

Le tematiche di **identificazione dei modelli** e quelle di **analisi dei dati** sono collegate:

- L'analisi preliminare dei dati dà indicazioni sul modello migliore per descriverli
- Tecniche di analisi dei dati possono essere usate per descrivere la **bontà** del modello
- Una rappresentazione probabilistica dei dati dà luogo ad un **modello probabilistico** capace di **gestire l'incertezza**:
 - ✓ nelle misure
 - ✓ nella conoscenza della realtà (quanto «non conosco?»)

Declineremo le due procedure sia per **sistemi statici** che per **sistemi dinamici**



Outline

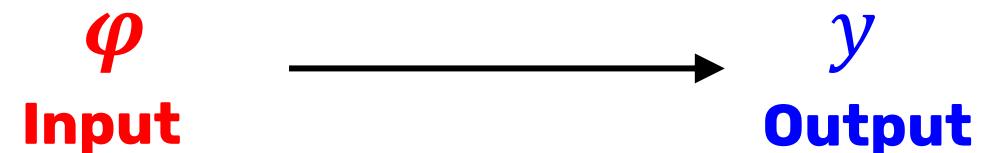
1. Presentazione del corso di Identificazione dei Modelli e Analisi dei Dati
2. Introduzione e motivazione
- 3. La stima di un modello dai dati: l'approccio supervisionato**
4. Sistemi (e modelli) statici
5. Sistemi (e modelli) dinamici
6. Riassunto



La stima di un modello dai dati

Le tecniche di stima (apprendimento, identificazione) di un **modello dai dati** possono essere (largamente) classificate in:

- **Apprendimento supervisionato** (supervised learning): stimare un (o più) **output** y sulla base di uno o più **input** φ



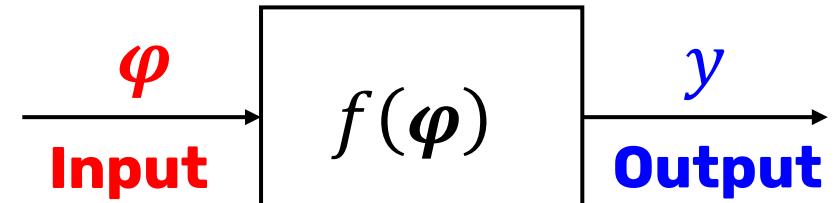
- **Apprendimento non supervisionato** (unsupervised learning): **non c'è l'output!**
L'obiettivo è scoprire relazioni e strutture nel solo input

In questo corso considereremo solo le **tecniche di apprendimento supervisionate** (regressione e classificazione)



La stima di un modello dai dati

L'obiettivo dell'apprendimento supervisionato è stimare (imparare, identificare) la funzione *ignota* $f(\varphi)$, che mappa gli **inputs** φ nell'**output** y , in modo che $y = f(\varphi)$



L'**input** è rappresentato da un vettore $\varphi = [\varphi_0 \ \varphi_1 \ \cdots \ \varphi_{d-1}] \in \mathbb{R}^{d \times 1}$, chiamato **vettore dei regressori** o delle **features**

- Ogni elemento $\varphi_0 \ \varphi_1 \ \cdots \ \varphi_{d-1}$ è chiamato **regressore** o **feature**

L'**output** y può essere

- un **numero** (output *continuo*), cioè $y \in \mathbb{R}$ \Rightarrow **Regessione**
- una **categoria** (output *discreto*), cioè $y \in \{"\text{Cat. 1}", \dots, "Cat. C"\}$ \Rightarrow **Classificazione**



Outline

1. Presentazione del corso di Identificazione dei Modelli e Analisi dei Dati
2. Introduzione e motivazione
3. La stima di un modello dai dati: l'approccio supervisionato
- 4. Sistemi (e modelli) statici**
5. Sistemi (e modelli) dinamici
6. Riassunto

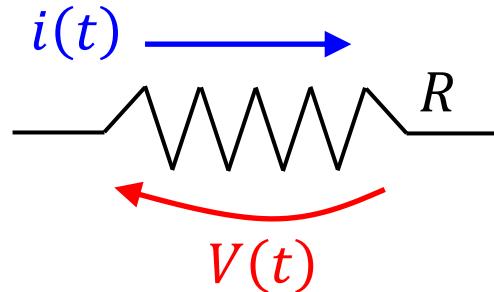


Sistemi statici

Con il termine **sistema statico** indichiamo quei sistemi per cui la sola conoscenza delle variabili di **input** è sufficiente a determinare il valore dell'**output**

Esempio: legge di Ohm per un resistore

$$i(t) = \frac{V(t)}{R}$$



L'uscita $i(t)$ all'istante t dipende solo dall'ingresso $V(t)$ al medesimo istante t

All'interno di questo corso, considereremo le tematiche di **MACHINE LEARNING** come quelle tecniche che permetto di stimare (apprendere) **sistemi statici**



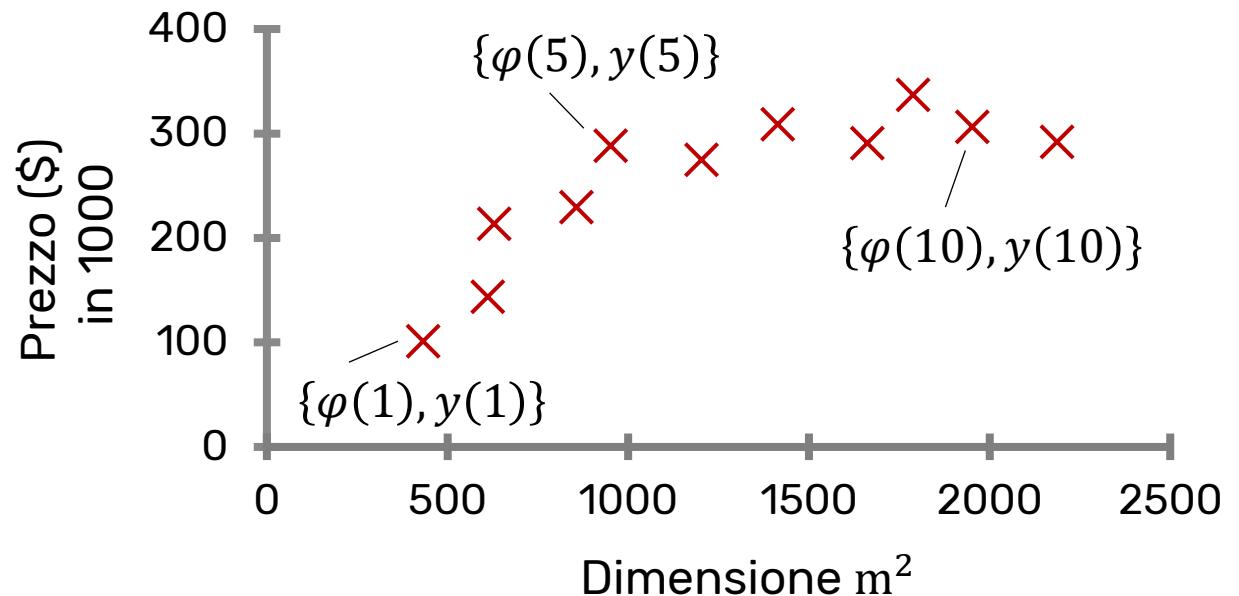
Esempio 1: stimare il prezzo delle case

Supponiamo di voler stimare il prezzo delle case nell'area di Boston

Vogliamo **imparare** (stimare) la relazione $y = f(\varphi)$ tra:

- φ : grandezza della casa in m^2 (regressore o feature)
- y : prezzo della casa (output)

Per poter fare questo, abbiamo bisogno di un **dataset** $\mathcal{D} = \{\varphi(i), y(i)\}_{i=1}^N$ di **osservazioni** dei valori sia di φ che di y



Esempio 1: stimare il prezzo delle case

Area (m ²)	# Cam. letto	Prezzo (1000\$)
523	1	115
645	1	150
708	2	210
1034	3	280
2290	4	355
2545	4	440

- **Obiettivo:** stimare prezzo case
- L'output y è **continuo**, $y \in \mathbb{R}$

REGRESSIONE

$$\varphi \in \mathbb{R}$$

$$\overbrace{\hspace{150pt}}$$

$$\varphi \in \mathbb{R}^{2 \times 1}$$

y → Imparare la relazione **DA Area A Prezzo**

y → Imparare la relazione **DA Area E # Camere da letto A Prezzo**



Esempio 2: image classification

Immagine	Output label
	Gatto
	Non gatto
	Gatto
	Non gatto

- **Obiettivo:** sviluppare un'applicativo per riconoscere se c'è un gatto nell'immagine
- Imparare il mapping **DA** un'immagine **A** una «classe di appartenenza»
- L'output y è una **categoria** (Gatto \ Non gatto)

CLASSIFICAZIONE



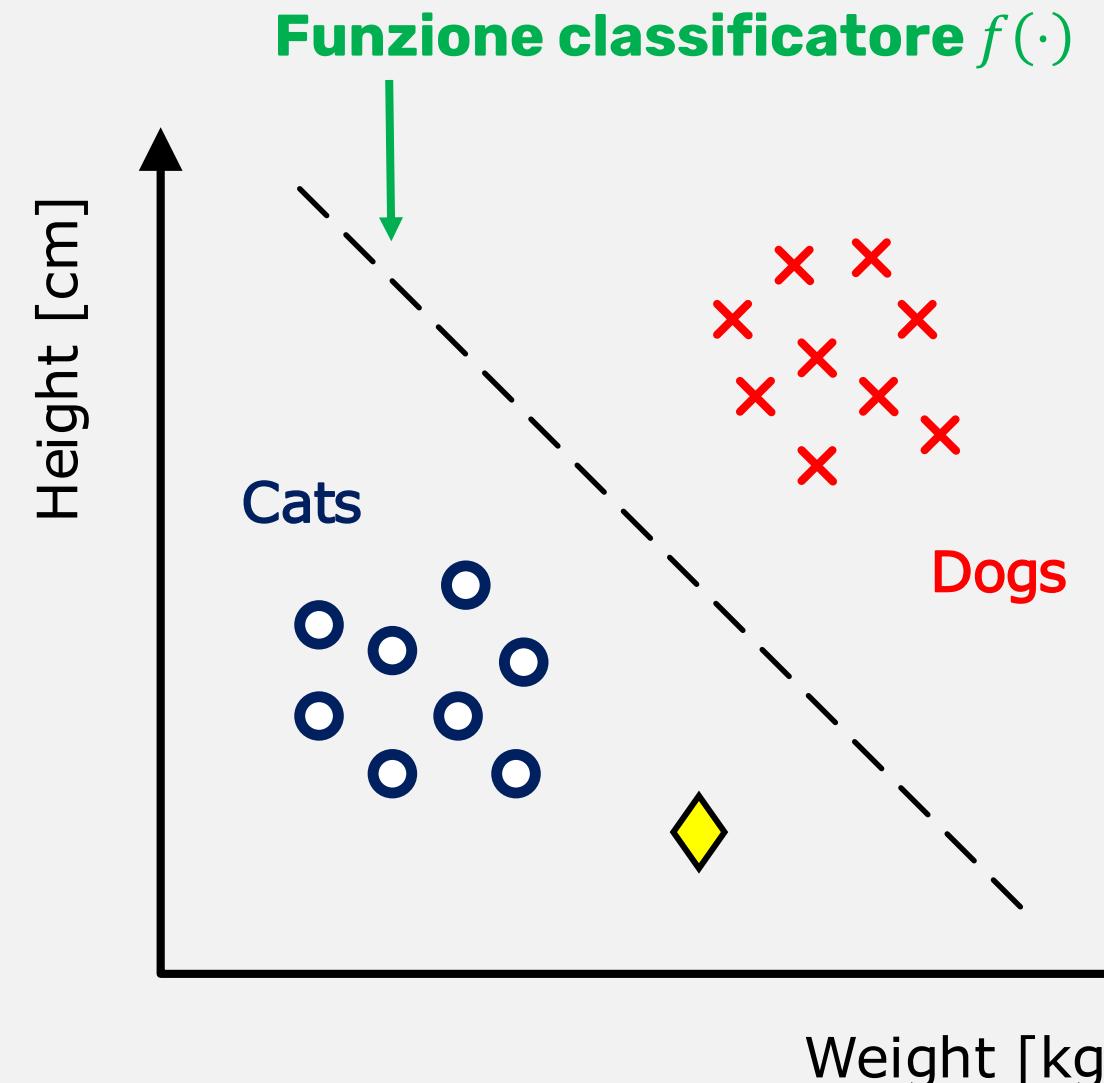
QUIZ!

Supponiamo di misurare il **peso** e **l'altezza** di alcuni cani e gatti

Vogliamo imparare la funzione $f(\cdot)$ che ci dica se $\varphi = [\varphi_1, \varphi_2]^T$ è un cane o un gatto

- φ_1 : peso
- φ_2 : altezza

DOMANDA: Il punto  come è classificato dal modello? _____



Outline

1. Presentazione del corso di Identificazione dei Modelli e Analisi dei Dati
2. Introduzione e motivazione
3. La stima di un modello dai dati: l'approccio supervisionato
4. Sistemi (e modelli) statici
- 5. Sistemi (e modelli) dinamici**
6. Riassunto



Sistemi dinamici

Con il termine **sistema dinamico** indichiamo quei sistemi per cui la sola conoscenza delle variabili di **input** (in un certo istante di tempo) **non è sufficiente** a determinare il valore dell'**output** al medesimo istante di tempo. Servono anche delle **condizioni iniziali**

I **modelli dinamici** consentono di descrivere **l'evoluzione futura** delle variabili coinvolte in funzione del loro **andamento passato** e delle **variabili esterne** (ingressi esogeni)

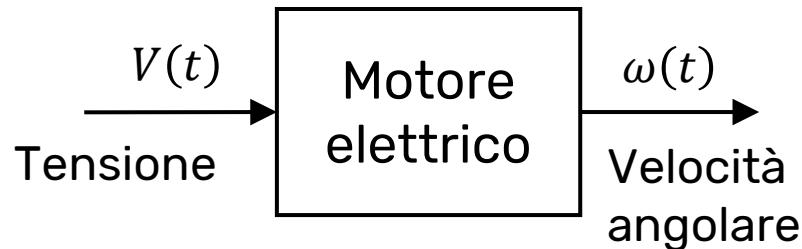
I sistemi dinamici coinvolgono il **tempo**: l'output $y(t)$ dipende da sè stesso a istanti passati $y(t - 1), y(t - 2), \dots y(t - n_a)$

Questa **dipendenza dal passato** conferisce al modello una **«memoria»** (cioè la dinamica), del comportamento passato, il quale influisce il comportamento presente

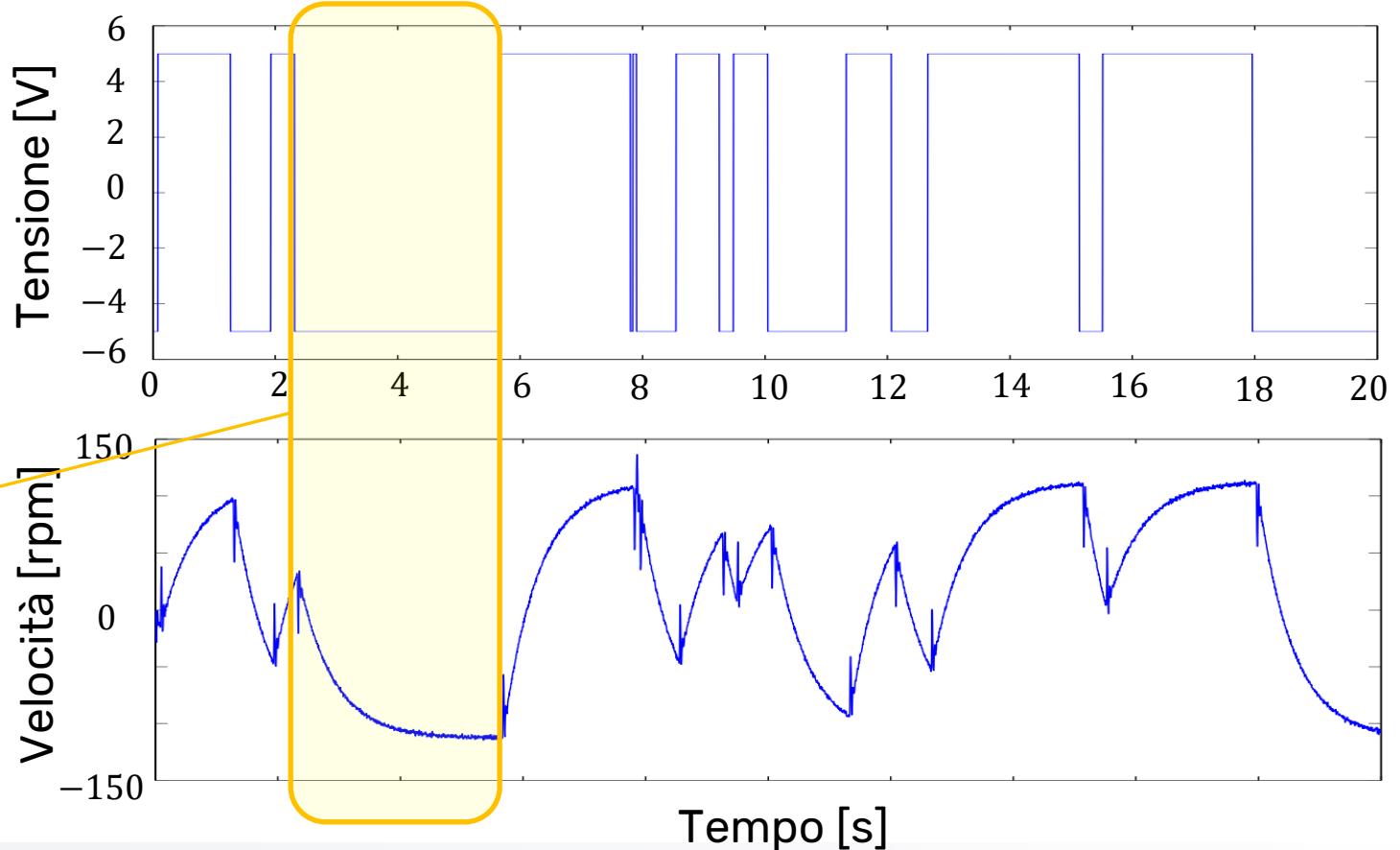


Sistemi dinamici

Questa **dipendenza dal passato** conferisce al modello una «**memoria**» (cioè la dinamica), del comportamento passato, il quale influisce il comportamento presente



Vediamo che il sistema è dinamico perché, anche se **l'input è fermo, l'output continua ad evolvere**

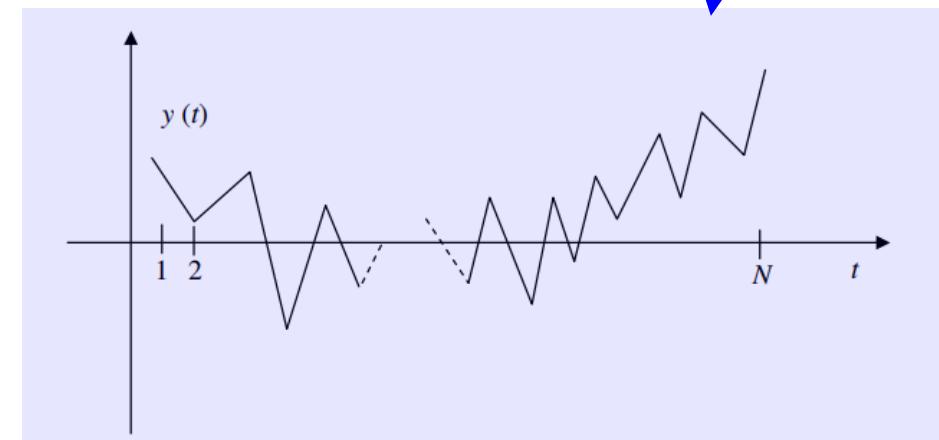
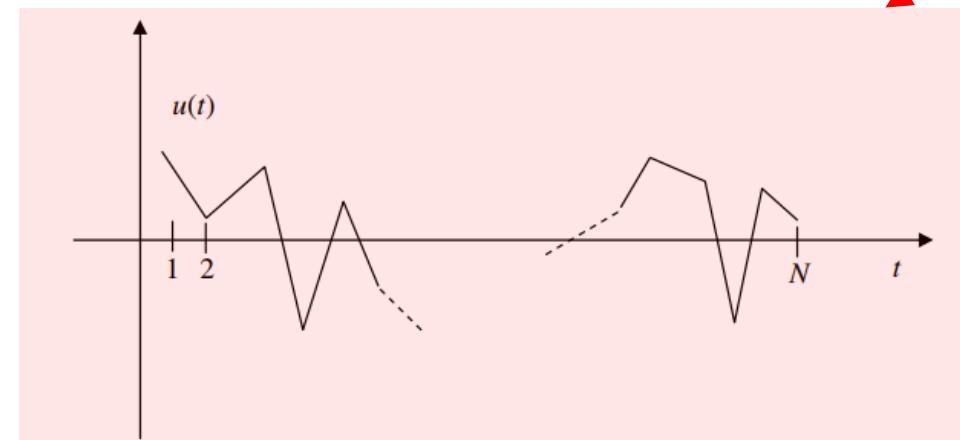
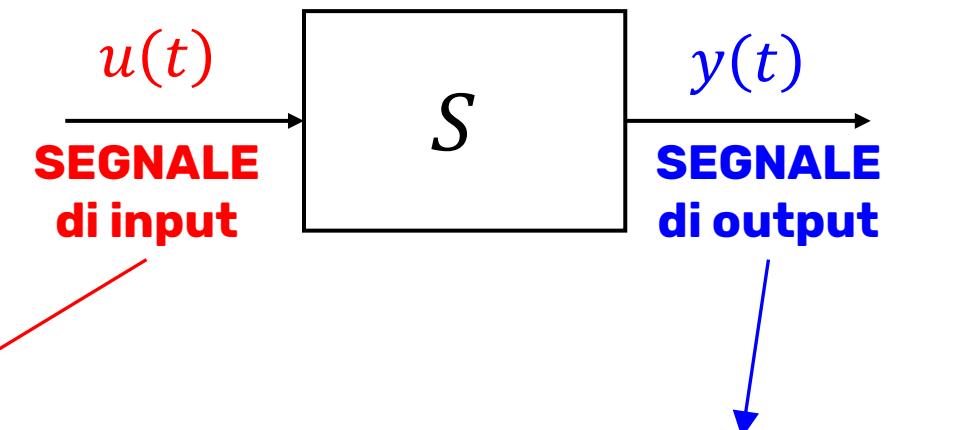


Sistemi dinamici

I sistemi dinamici, per la presenza della variabile tempo, vengono utilizzati per modellare le relazioni tra **segnali** di ingresso $u(t)$ e di uscita $y(t)$

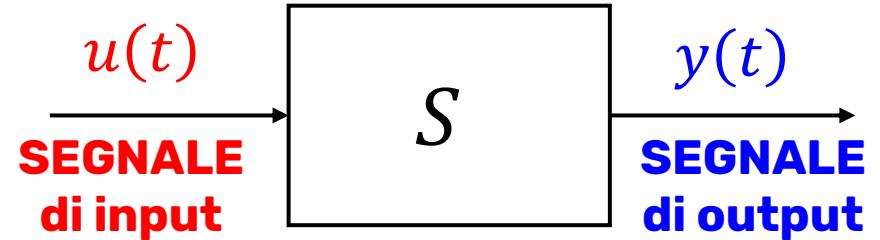
Due set di N dati sono collezionati, **campionando i segnali** a istanti temporali $t = 1, 2, \dots, N$

- Dati di input $\{u(1), u(2), \dots, u(N)\}$
- Dati di output $\{y(1), y(2), \dots, y(N)\}$



Sistemi dinamici

Esempi di sistemi dinamici e segnali di input \ output



Segnale di input

Segnale audio (prima della trasmisione)

Corrente

Quantità di un medicinale

Millimetri di pioggia

Segnale di output

Segnale audio (dopo la trasmissione)

Coppia motore

Concentrazione di un ormone

Concentrazione di un inquinante



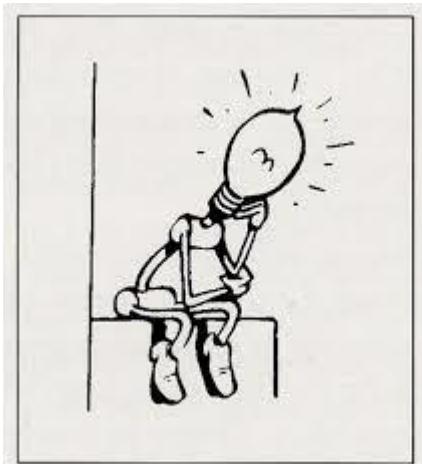
Sistemi dinamici

I sistemi dinamici possono essere definiti a **tempo continuo** o a **tempo discreto**

I **fenomeni naturali** e **fisici** sono intrinsecamente **continui**

- In questo caso, il sistema è descritto attraverso equazioni differenziali, del tipo

$$\frac{dy}{dt} = \dot{y}(t) = -2 \cdot y(t) + 3 \cdot u(t)$$



La derivata è la «rappresentazione matematica del comportamento futuro di una funzione». Questa nozione di «futuro» è esattamente ciò di cui abbiamo bisogno per rappresentare la memoria di un sistema dinamico a tempo continuo

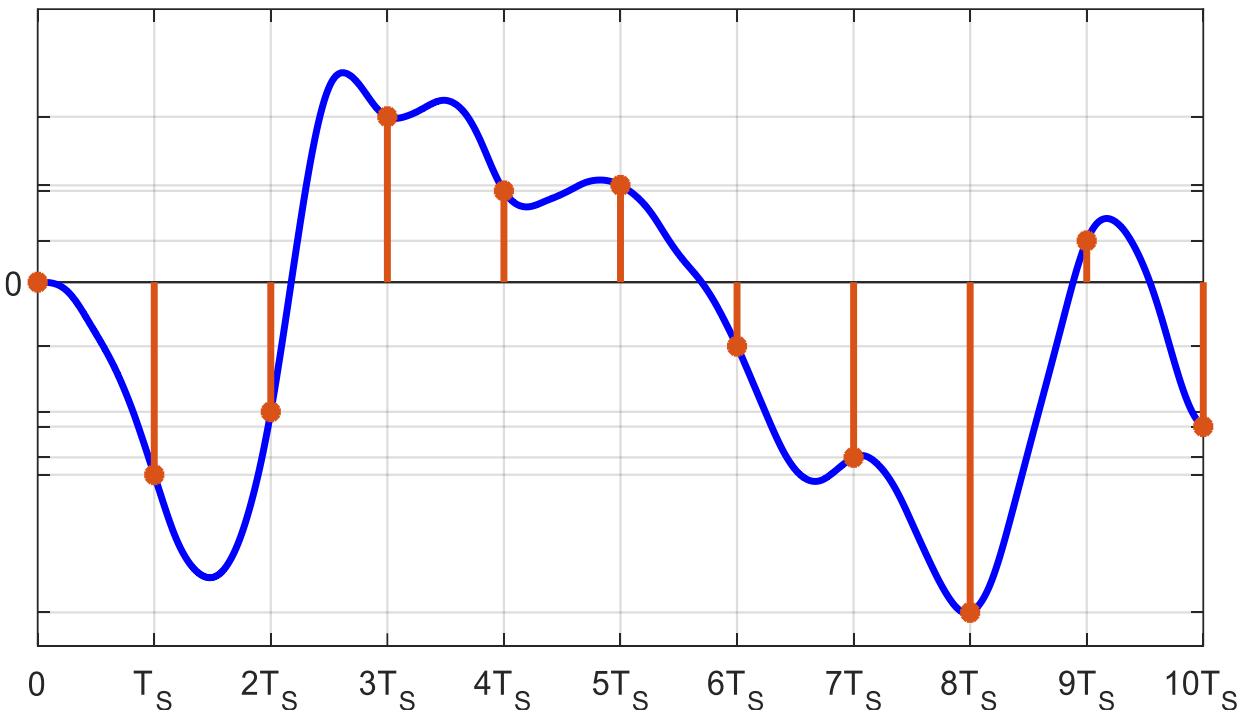


Sistemi dinamici

Tuttavia, il computer può gestire solo una quantità limitata di dati. Pertanto, i segnali devono essere **campionati** con un tempo di campionamento T_s , tale da memorizzare una quantità finita di dati a tempi discreti $t \cdot T_s$, con $t = 1, \dots, N$

$$y(t) = y(t \cdot T_s)$$

Nel seguito, lavoreremo solo con **sistemi discreti**. Per semplicità di notazione, useremo $y(t)$ per indicare $y(t \cdot T_s)$



Sistemi dinamici

L'evoluzione dei **segnali a tempo discreto** può essere descritta dai **sistemi a tempo discreto**:

- invece di un'equazione differenziale, abbiamo un'**equazione alle differenze**

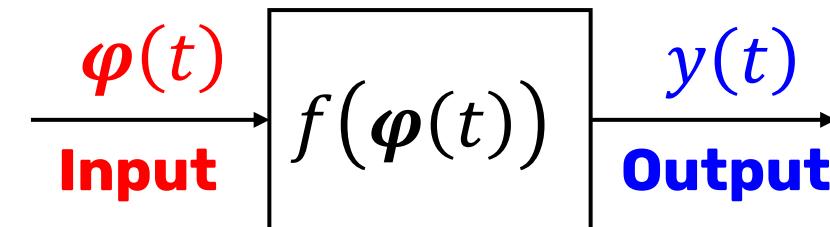
$$y(t) = -0.5 \cdot y(t - 1) + 3 \cdot u(t)$$

Con l'equazione alle differenze, è molto chiaro che $y(t)$ **dipende dai suoi valori precedenti** (e anche dall'input $u(t)$)



Modelli di sistemi dinamici: approccio

Allo scopo di identificare modelli di sistemi dinamici, formuleremo il problema di stima proprio come fatto per i sistemi statici



L'unica differenza risiede nella **definizione del vettore dei regressori**. Poiché l'uscita dipende dai segnali di ingresso e di uscita $u(t)$ e $y(t)$, il vettore dei regressori $\varphi(t)$ in un determinato momento t sarà simile a:

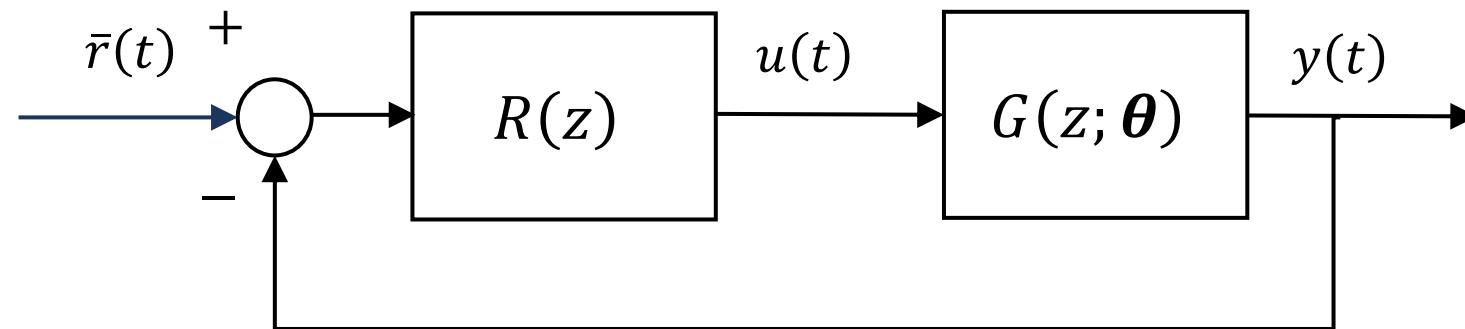
$$\varphi(t) = [y(t-1) \cdots y(t-n_a) \ u(t) \ \cdots u(t-n_b)]^\top$$



Modelli di sistemi dinamici: motivazione

Modelli di sistemi dinamici sono usati nell'ingegneria per:

- **Progettazione del controllo:** spesso è necessario conoscere la funzione di trasferimento $G(s)$ o $G(z)$ del sistema, al fine di tarare un controllore opportuno



Problema: chi ci dice quale è la $G(z; \theta)$ del sistema? Quanti poli\zeri ha? E che valore hanno? E quanto vale il guadagno? C'è un ritardo?



**IDENTIFICAZIONE DEI MODELLI
(dinamici)**

Fondamenti di automatica - 9 CFU

Controlli automatici - 6 CFU

Advanced and Multivariable control - 6 CFU



Modelli di sistemi dinamici: motivazione

Modelli di sistemi dinamici sono usati nell'ingegneria per:

- **Simulazione:** possiamo simulare, con un computer, la risposta (output) di un modello a determinati input. Osservando la risposta del modello, comprendiamo meglio il comportamento del sistema modellato



Problema: chi ci dice qual è il valore di certi parametri? E se come facciamo a modellare una relazione nonlineare ignota?

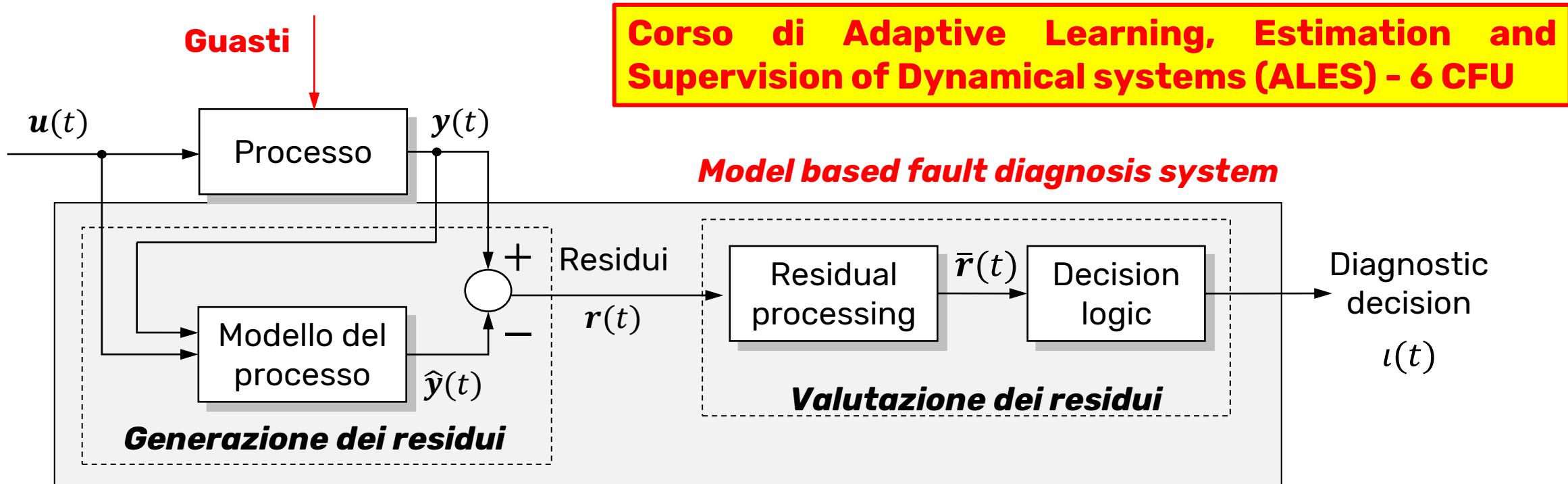


**IDENTIFICAZIONE DEI MODELLI
(dinamici)**

Modelli di sistemi dinamici: motivazione

Modelli di sistemi dinamici sono usati nell'ingegneria per:

- **Diagnosi dei guasti:** confrontando i segnali misurati con i segnati simulati dal modello, è possibile individuare se il sistema ha dei guasti (sugli attuatori, sensori, o sul processo)

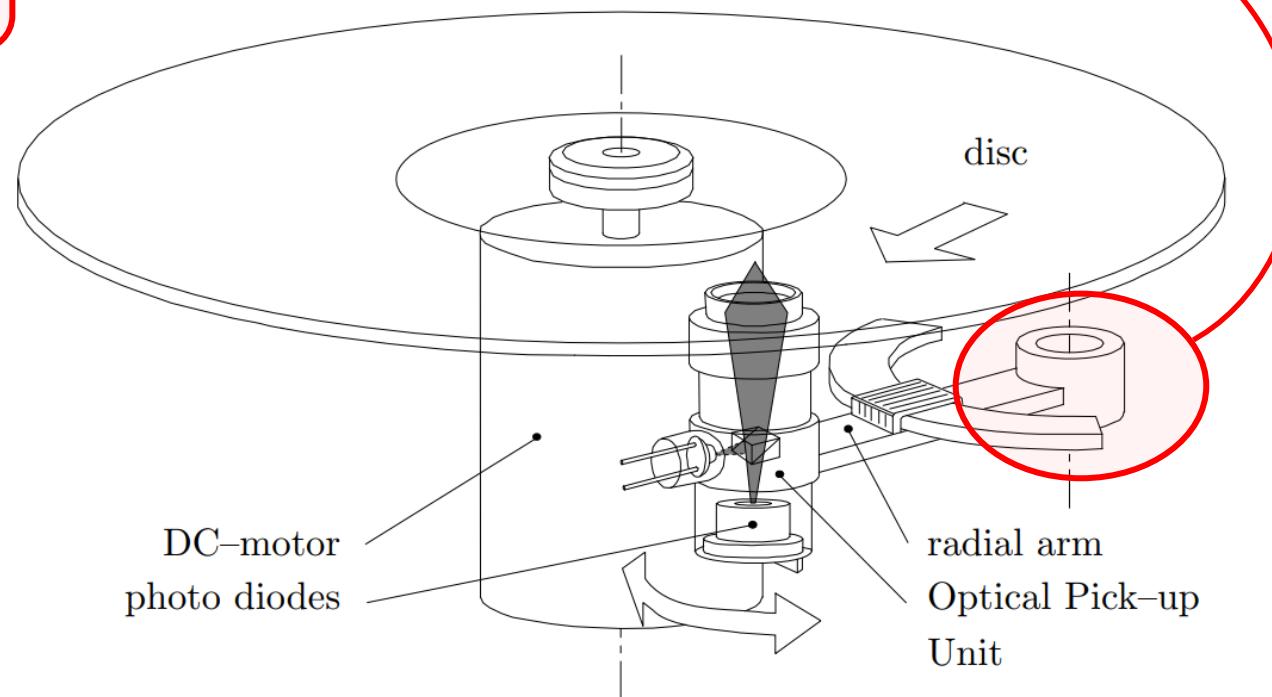


Esempio 1: controllo del lettore laser per lettore CD

Obiettivo: posizionare la testina laser sulla traccia corretta, tramite un braccio meccanico

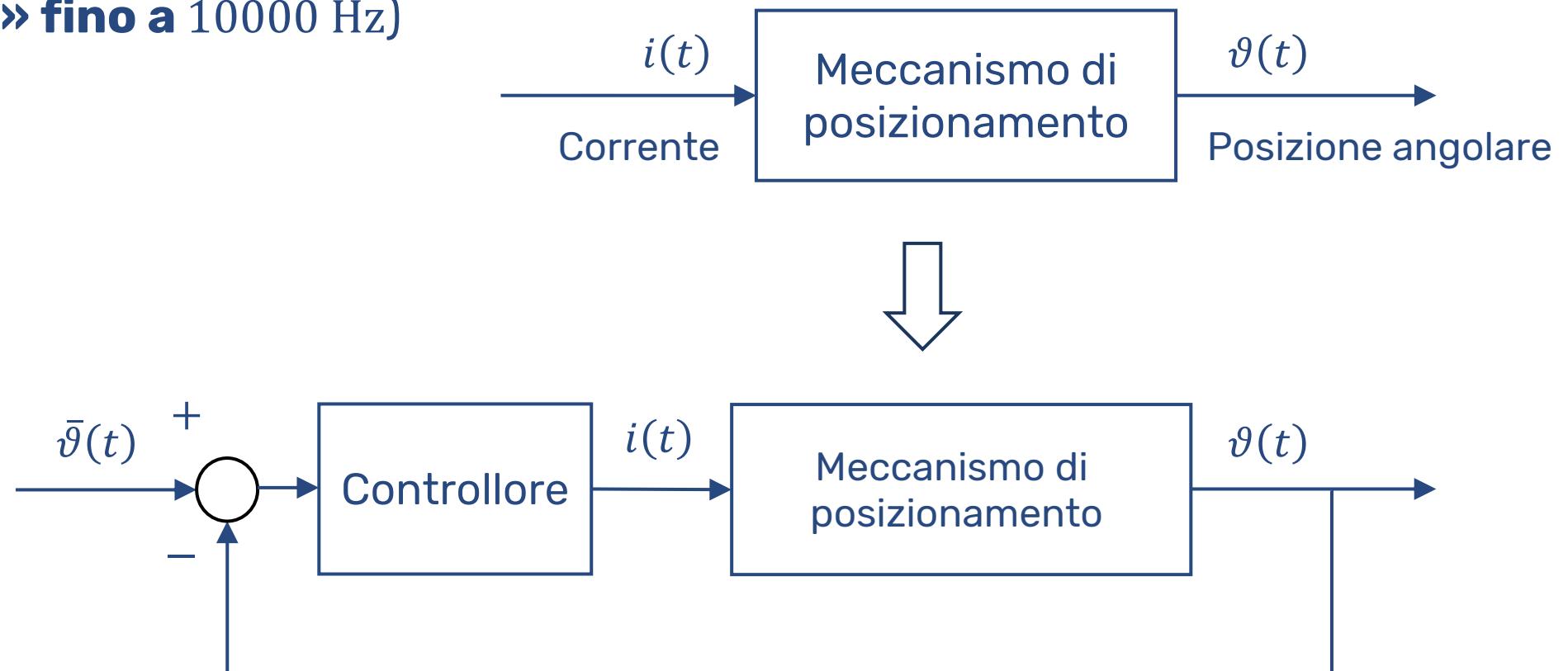
Per fare ciò, vogliamo **controllare la corrente** $i(t)$ al motore del braccio radiale

Questo motore posiziona una testina che emette una fonte laser. La posizione $\vartheta(t)$ del raggio laser è ottenuta tramite fotodiodi



Esempio 1: controllo del lettore laser per lettore CD

Al fine di raggiungere l'obiettivo, si vuole **progettare un controllore** per il sistema di posizionamento, in modo da avere una banda di controllo di 1000 Hz (devo avere un **modello «buono» fino a 10000 Hz**)



Esempio 1: controllo del lettore laser per lettore CD

Proviamo a fare un modello basandoci sulle leggi note della fisica (white-box)

Relazione tra corrente $i(t)$ e coppia $T_m(t)$ di un motore DC: $T_m(t) = k \cdot i(t)$

Trascurando gli attriti, la legge di Newton ci dice che: $J \cdot \ddot{\vartheta}(t) = T_m(t)$

Quindi otteniamo un **doppio integratore**:

$$\hat{G}(s) = \frac{\vartheta(s)}{i(s)} = \frac{k}{J s^2}$$

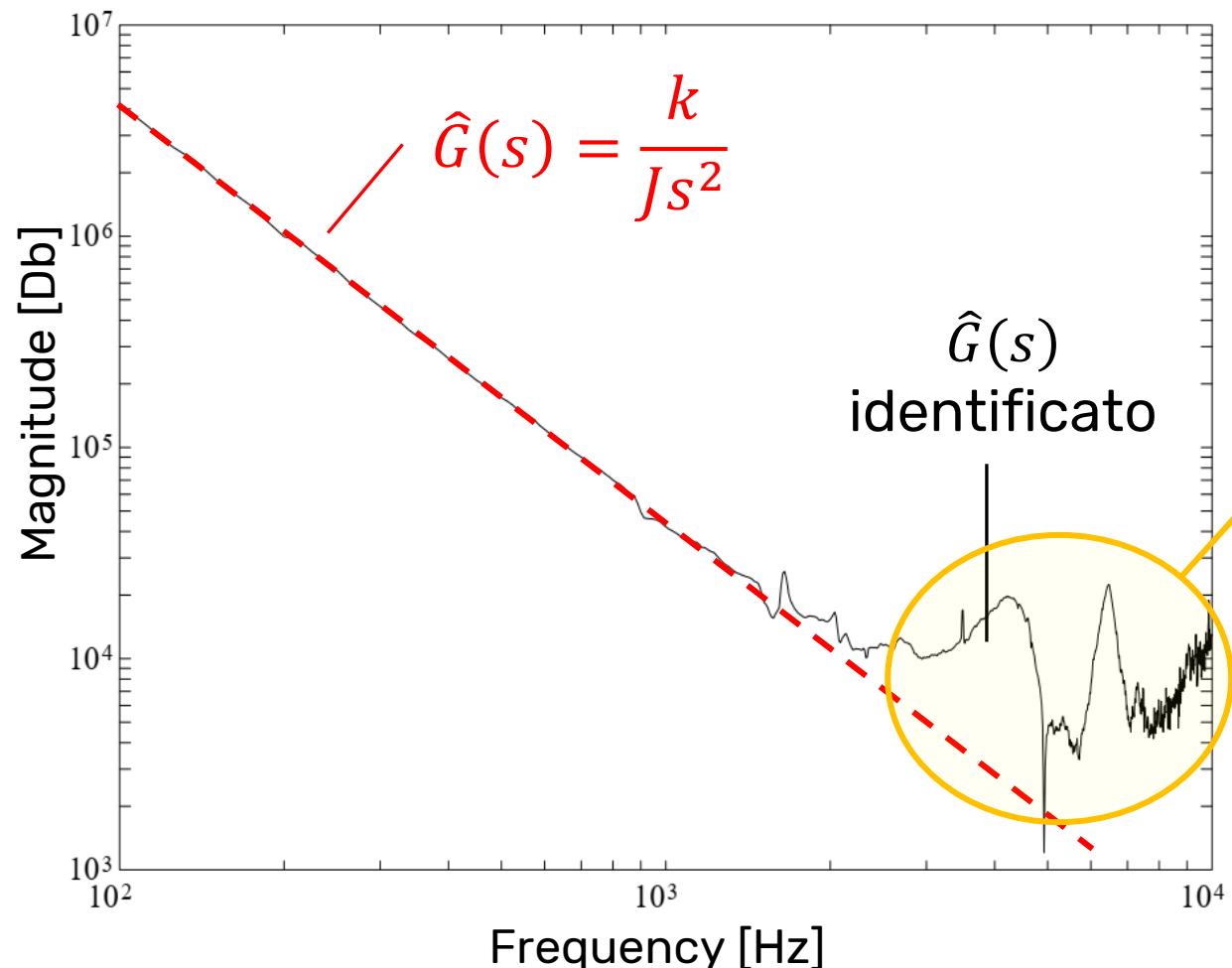
- J : inerzia
- k : costante elettrica del motore

Il controllore progettato con questo modello non riesce ad ottenere la banda desirata senza causare **vibrazioni indesiderate**  **Modello non (sufficientemente) corretto!**



Esempio 1: controllo del lettore laser per lettore CD

A seguito di un esperimento di **identificazione**, il seguente modello è stato identificato

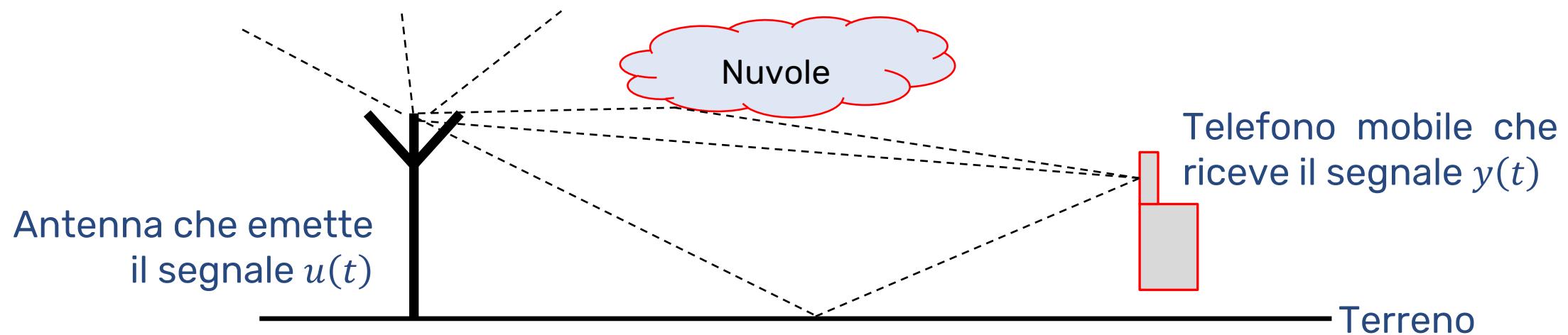


Si nota la presenza di **modi flessibili** che devono essere tenuti in considerazione durante la progettazione del controllore

Tali modi sono pressoché **impossibili da modellare tramite leggi fisiche**



Esempio 2: ricezione segnale nella telefonia mobile



Il segnale $y(t)$ che viene ricevuto è composto da **versioni ritardate del segnale emesso** $u(t)$ e da un **rumore** $v(t)$

$$y(t) = g_1 u(t - n_1) + g_2 u(t - n_2) + \dots + v(t)$$

$$= G_0(z)u(t) + v(t)$$

$$\bullet \quad G_0(z) = g_1 z^{-n_1} + \dots$$



Esempio 2: ricezione segnale nella telefonia mobile

$$y(t) = G_0(z)u(t) + v(t)$$

Obiettivo: ricostruire $u(t)$ partendo da $y(t)$

Se il modello fosse noto e non ci fosse rumore, potremmo calcolare $u(t)$ come

$$u(t) = \frac{1}{G_0(z)} y(t)$$

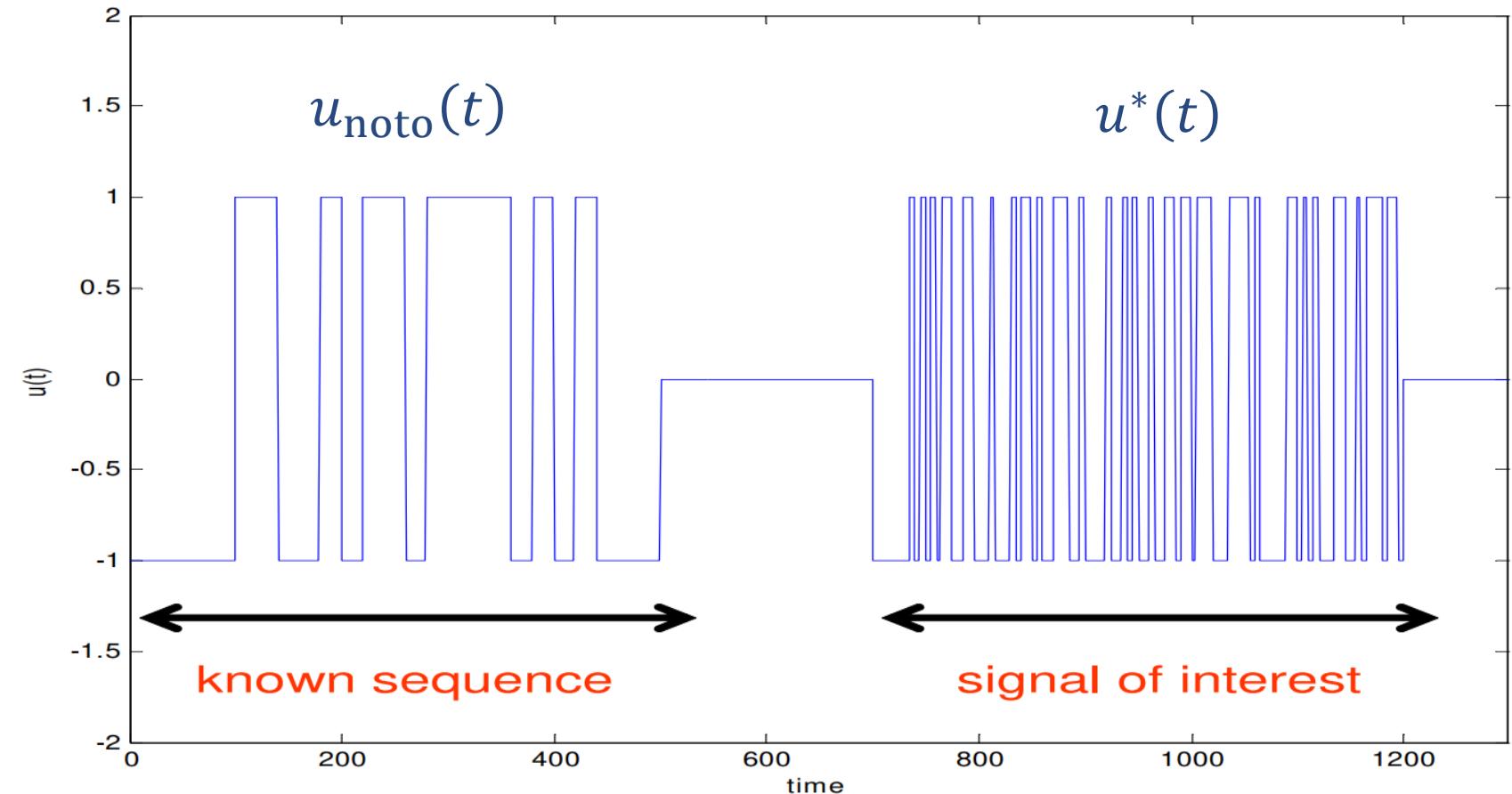
Problema: il modello $G_0(z)$ non è noto perché dipende dalla posizione del telefonino (che è mobile per definizione). Servirebbe un modello per ogni posizione del cellulare...

Soluzione: un **modello $\hat{G}(z)$ viene identificato ad ogni chiamata** dal software GSM



Esempio 2: ricezione segnale nella telefonia mobile

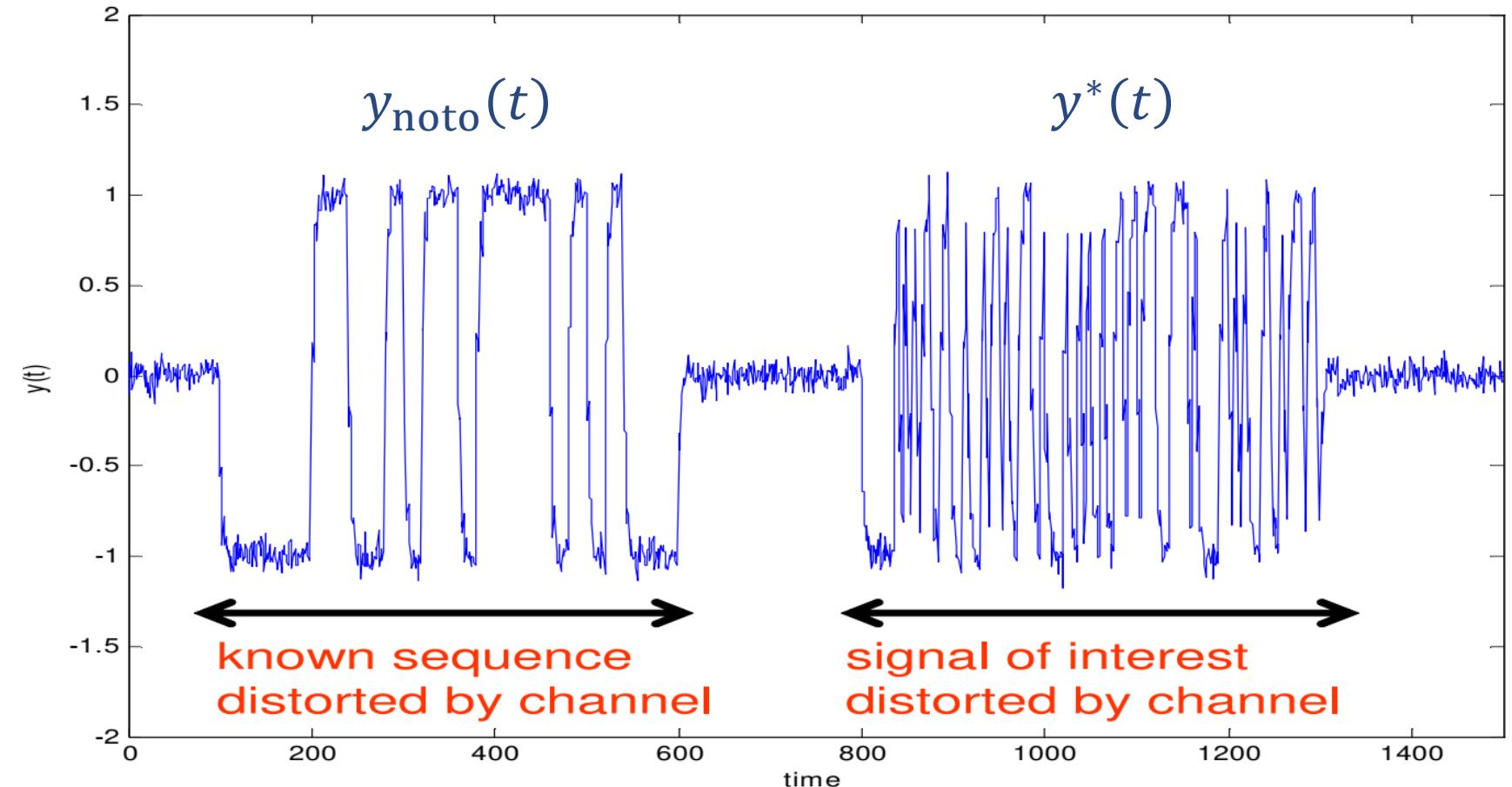
Quando $u(t)$ viene emesso, il **segnale di interesse** $u^*(t)$ è preceduto da un **segnale noto** $u_{\text{noto}}(t)$



Esempio 2: ricezione segnale nella telefonia mobile

In ricezione, sia $u^*(t)$ che $u_{\text{noto}}(t)$ vengono corrotti dal canale di trasmissione

- Segnale noto
ricevuto: $y_{\text{noto}}(t)$
- Segnale di interesse
ricevuto: $y^*(t)$



Esempio 2: ricezione segnale nella telefonia mobile

Poiché $u_{\text{noto}}(t)$ è un segnale noto, il software GSM usa i dati di $u_{\text{noto}}(t)$ e $y_{\text{noto}}(t)$ per **identificare il modello del canale** $\hat{G}(z)$

Dopodichè, il segnale di interesse $u^*(t)$ è **stimato** come:

$$\hat{u}^*(t) = \frac{1}{\hat{G}(z)} y^*(t)$$



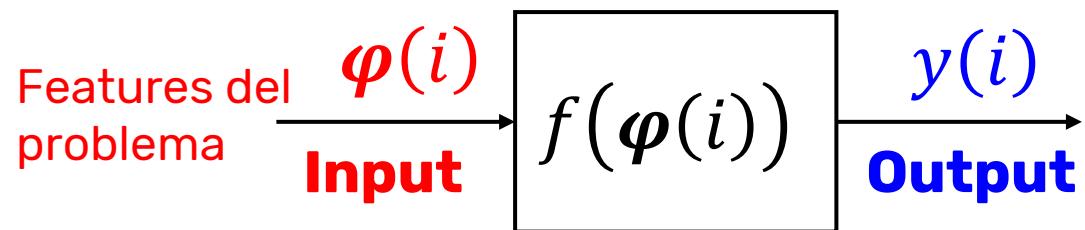
Outline

1. Presentazione del corso di Identificazione dei Modelli e Analisi dei Dati
2. Introduzione e motivazione
3. La stima di un modello dai dati: l'approccio supervisionato
4. Sistemi (e modelli) statici
5. Sistemi (e modelli) dinamici
- 6. Riassunto**

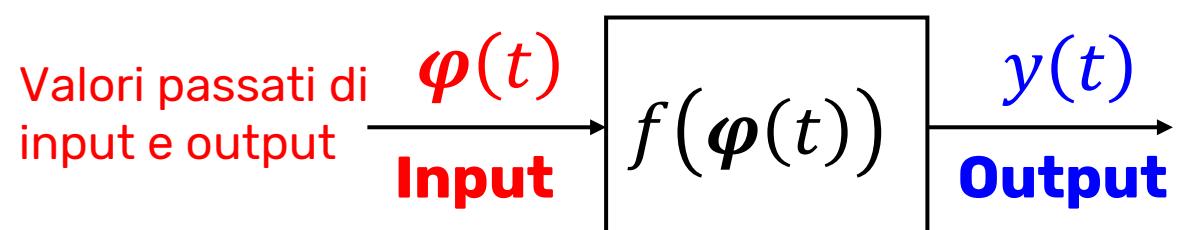


Riassunto di quello che impararemo a fare

Sistemi statici



Sistemi dinamici



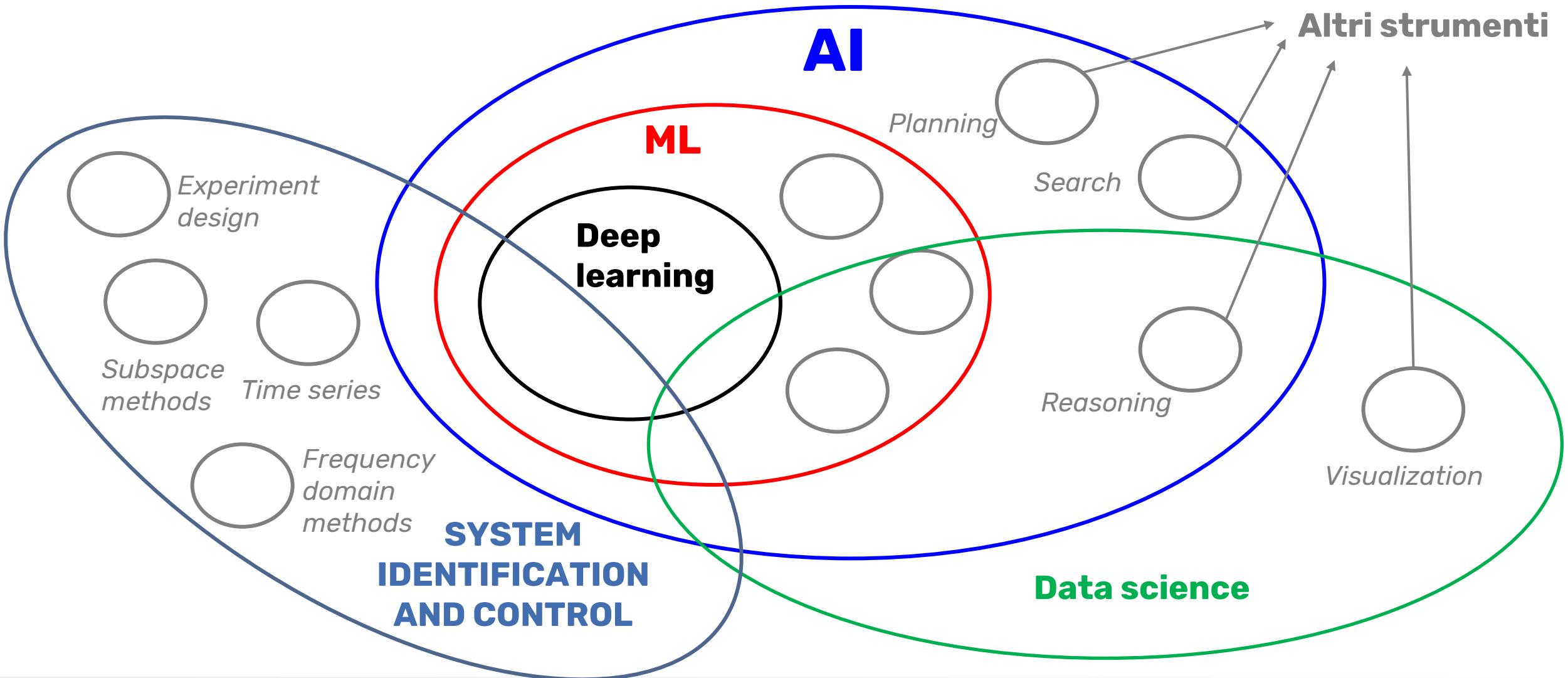
- Con i **sistemi statici**, indicizzeremo le osservazioni con la lettera i
- Con i **sistemi dinamici**, indicizzeremo le osservazioni con la lettera t

In ogni caso il nostro obiettivo sarà **stimare $f(\cdot)$ dai dati**

- Nel caso statico, parleremo di «**apprendimento**» (*model learning*)
- Nel caso dinamico, parleremo di «**identificazione**» (*system identification*)



ML, data science, AI and dynamical systems



QUIZ!

DOMANDA: La stima di sistemi dinamici dai dati (**system identification**) è un problema di:

- A. Apprendimento **supervisionato**: nello specifico, è un problema di **regressione**
- B. Apprendimento **supervisionato**: nello specifico, è un problema di **classificazione**
- C. Apprendimento **non supervisionato**





**UNIVERSITÀ
DEGLI STUDI
DI BERGAMO**

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



IDENTIFICAZIONE DEI MODELLI E ANALISI DEI DATI (IMAD)

Lezione 2: Richiami di statistica

**Corso di Laurea Magistrale in
INGEGNERIA INFORMATICA**

SPEAKER
Prof. Mirko Mazzoleni

PLACE
Università degli Studi di
Bergamo

Syllabus

Parte I: sistemi statici

1. Richiami di statistica

2. Teoria della stima

2.1 Proprietà degli stimatori

3. Stima a minimi quadrati

3.1 Stima di modelli lineari

3.2 Algoritmo del gradient descent

4. Stima a massima verosimiglianza

4.1 Proprietà della stima

4.2 Stima di modelli lineari

5. Regressione logistica

5.1 Stima di un modello di regressione logistica

6. Fondamenti di machine learning

6.1 Bias-Variance tradeoff

6.2 Overfitting

6.3 Regolarizzazione

6.4 Validazione

7. Cenni di stima Bayesiana

7.1 Probabilità congiunte, marginali e condizionate

7.2 Connessione con Filtro di Kalman



Parte I: sistemi statici**Parte II: sistemi dinamici**Stima parametrica $\hat{\theta}$

- θ deterministico
 - **NO assunzioni su ddp dei dati**
 - ✓ Stima parametri popolazione
 - ✓ Stima modello lineare: minimi quadrati
 - **SI assunzioni su ddp dei dati**
 - ✓ Stima massima verosimiglianza parametri popolazione
 - ✓ Stima modello lineare: massiva verosimiglianza
 - ✓ Regressione logistica
- θ variabile casuale
 - **SI assunzioni su ddp dei dati**
 - ✓ Stima Bayesiana

Machine learning

Outline

1. Definizione e proprietà delle variabili casuali: caso scalare
2. Definizione e proprietà delle variabili casuali: caso multivariabile
3. Stima e stimatori
4. Proprietà degli stimatori



Outline

- 1. Definizione e proprietà delle variabili casuali: caso scalare**
2. Definizione e proprietà delle variabili casuali: caso multivariabile
3. Stima e stimatori
4. Proprietà degli stimatori



Variabili casuali (random variables)

Intuizione: una **variabile casuale** ν è una variabile definita a partire dall'**esito** s di un **esperimento casuale**

Esempio: l'esperimento è il lancio di una moneta. A seconda se l'esito è $s = \text{testa}$ o $s = \text{croce}$, la variabile ν assume un valore diverso

$$\nu = \begin{cases} 1 & s = \text{testa} \\ 0 & s = \text{croce} \end{cases}$$

- Indichiamo una variabile casuale (v.c.) come $\nu(s)$
- Il valore assunto da una v.c. ν a seguito di un particolare esito \bar{s} è $\nu(\bar{s})$

Problema: dato che ν può assumere diversi valori (a seconda del valore assunto da s), come posso descriverli?



Assegno una probabilità che ogni esito accada. Questo influisce sulla probabilità che ν assuma i valori che può assumere (*distribuzione di probabilità*)



Variabili casuali (random variables)

Caso 1) v assume valori DISCRETI (v è una variabile casuale discreta)

- **Funzione di probabilità di massa (pmf)** $p(x) = P(v = x)$

Associa ad ogni valore x di v una probabilità

Indichiamo con x_i i valori di v . Se v può assumere m diversi valori, allora

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

Esempio: Esperimento «lancio di un dado»

$$m = 6 \quad x_1 = 1$$

$$x_2 = 2$$

⋮

$$x_6 = 6$$

$$p(x_1) = P(v = x_1) = P(v = 1) = 1/6$$

$$p(x_2) = P(v = x_2) = P(v = 2) = 1/6$$

⋮

$$p(x_6) = P(v = x_6) = P(v = 6) = 1/6$$



Variabili casuali (random variables)

Caso 2) v assume valori CONTINUI (v è una variabile casuale continua)

- **Funzione di densità di probabilità (pdf)** $f_v(x)$

In questo caso, dire $P(v = x)$ **non ha senso**. Infatti, dato che v può assumere **infiniti valori**, la probabilità che v assuma esattamente un valore specifico è praticamente zero!

$$\rightarrow P(v = x) = 0$$

Intuizione: se la variabile v (continua) assumesse valori tutti equiprobabili (come nel caso del dado), la probabilità che v assuma una valore specifico sarebbe $\frac{1}{\infty} = 0$

Esempio: sia v l'altezza di un uomo adulto. Non ha senso chiedersi la probabilità che un uomo sia alto **esattamente** 1,7425415478795121795387 metri



Variabili casuali (random variables)

La pdf $f_v(x)$ definisce la probabilità che v appartenga ad un **intervallo di valori** $[a, b]$

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

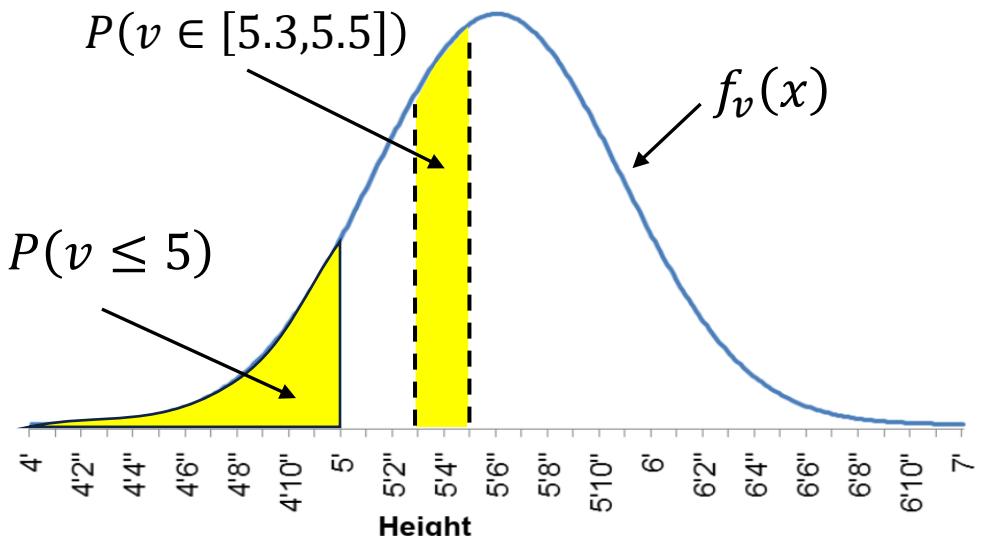
- $f_v(x) \geq 0$
- $\int_{-\infty}^{+\infty} f_v(x) dx = 1$

- **Funzione di densità cumulata (cdf)**

$$F_v(z)$$

$$F_v(z) = \int_{-\infty}^z f_v(x) dx = P(v \leq z)$$

$$F_v(5) = P(v \leq 5)$$



Valore atteso

Il **valore atteso** di una variabile casuale v è:

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

Somma pesata dei valori x che v può assumere. I pesi sono la probabilità di osservare il valore x

Il valore atteso gode della proprietà di **linearità**:

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

Nota: l'operatore valore atteso $\mathbb{E}_s[v]$ considera **tutti i possibili esiti** s della variabile casuale v . Di seguito, per semplicità, **renderemo implicita la dipendenza** da s , esplicitandola quando necessario



Varianza

La **varianza** di una variabile casuale v è:

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

- Quanto i valori x si discostano dalla loro media
- Se varianza piccola, v assume valori x molto vicini fra loro

Osservazioni

- $\text{Var}[v] \geq 0$. Se $\text{Var}[v] = 0$, la variabile v è deterministica (assume sempre un solo valore)
- Deviazione standard: $\sigma[v] = \sqrt{\text{Var}[v]}$
- $\boxed{\text{Var}[v]} = \mathbb{E}[(v - \mathbb{E}[v])^2] = \mathbb{E}[v^2 - 2\mathbb{E}[v]v + \mathbb{E}[v]^2] = \mathbb{E}[v^2] - 2\mathbb{E}[\mathbb{E}[v]v] + \mathbb{E}[\mathbb{E}[v]^2]$
 $= \mathbb{E}[v^2] - 2\mathbb{E}[v] \cdot \mathbb{E}[v] + \mathbb{E}[v]^2 \quad \boxed{= \mathbb{E}[v^2] - \mathbb{E}[v]^2}$
- $\text{Var}[\alpha \cdot v + \beta] = \alpha^2 \cdot \text{Var}[v] \quad \forall \alpha, \beta \in \mathbb{R}$



Correlazione

Date due variabili casuali v_1 e v_2 , si definisce il **coefficiente di correlazione** come:

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

- ρ indica il grado di **dipendenza lineare** tra v_1 e v_2 . Infatti, se $v_2 = \alpha v_1 + \beta$, si ha $\rho = 1$
- Se $\rho = 0$, le due variabili si dicono **scorrelate**



Covarianza

Date due variabili casuali v_1 e v_2 , si definisce la **covarianza** come:

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

E quindi

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

- Le variabili casuali v_1 e v_2 sono **scorrelate** se $\text{Cov}[v_1, v_2] = 0$



Outline

1. Definizione e proprietà delle variabili casuali: caso scalare
- 2. Definizione e proprietà delle variabili casuali: caso multivariabile**
3. Stima e stimatori
4. Proprietà degli stimatori



Variabili casuali, caso multivariabile

Le precedenti definizioni si possono estendere al caso di **vettore** di variabili casuali

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

Assumiamo che \boldsymbol{v} sia una v.c. continua

- **Funzione di densità cumulata (cdf)**

$$F_{\boldsymbol{v}}(z_1, z_2, \dots, z_d)$$

$$F_{\boldsymbol{v}}(z_1, z_2, \dots, z_d) = P(v_1 \leq z_1, v_2 \leq z_2, \dots, v_d \leq z_d)$$

$$= \int_{-\infty}^{z_1} \int_{-\infty}^{z_2} \cdots \int_{-\infty}^{z_d} f_{v_1, v_2, \dots, v_d}(x_1, x_2, \dots, x_d) dx_1 dx_2 \dots dx_d$$

Pdf congiunta



Variabili casuali, caso multivariabile

- Il **valore atteso** è un vettore colonna di d componenti

$$\mathbb{E}[\boldsymbol{\nu}] = \begin{bmatrix} \mathbb{E}[\nu_1], \mathbb{E}[\nu_2], \dots, \mathbb{E}[\nu_d] \end{bmatrix}^T \in \mathbb{R}^{d \times 1}$$

- La **varianza** è una matrice $d \times d$ **semidefinita positiva** e **simmetrica**

$$\text{Var}[\boldsymbol{\nu}] = \int_{\mathbb{R}^d} (\boldsymbol{x} - \mathbb{E}[\boldsymbol{\nu}]) (\boldsymbol{x} - \mathbb{E}[\boldsymbol{\nu}])^T f_{\boldsymbol{\nu}}(\boldsymbol{x}) d\boldsymbol{x}$$

$$= \begin{bmatrix} \text{Var}[\nu_1] & \cdots & \text{Cov}[\nu_1, \nu_d] \\ \vdots & \ddots & \vdots \\ \text{Cov}[\nu_d, \nu_1] & \cdots & \text{Var}[\nu_d] \end{bmatrix}$$

- «simile» al ≥ 0 per numeri reali
- Una matrice M reale e simmetrica è definita positiva se $\mathbf{z}^T M \mathbf{z} \geq 0 \forall \mathbf{z} \in \mathbb{R}$
- Autovalori di M sono ≥ 0



Indipendenza

Due variabili casuali v_1 e v_2 con funzione di probabilità congiunta $f_{v_1,v_2}(x_1, x_2)$ si dicono **indipendenti** se

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

- Se due variabili v_1 e v_2 sono **indipendenti**, allora sono anche scorrelate (non vale il viceversa in quanto potrebbero essere dipendenti in modo **non lineare**)

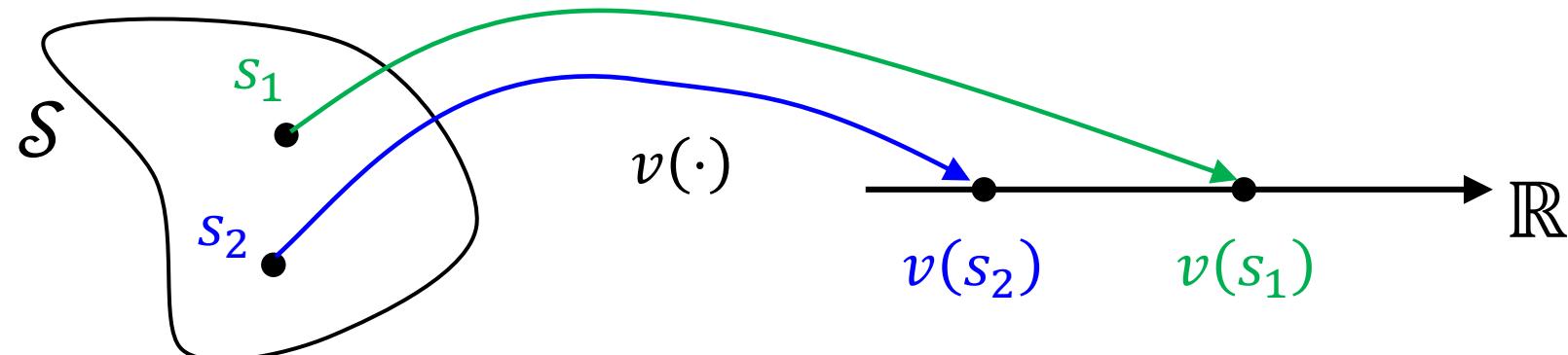


Variabili casuali: approfondimento

APPROFONDIMENTO

Una definizione più rigorosa di variabile casuale è quella di considerare una v.c. come una **funzione**, che, in funzione di un valore dell'esito s , ritorna un valore della v.c.

Definizione: una variabile casuale (scalare, reale) è una funzione definita sull'insieme degli esiti \mathcal{S} , che, ad ogni esito s_i , restituisce un numero reale $v(\cdot): \mathcal{S} \rightarrow \mathbb{R}$



Probabilità: approfondimento

APPROFONDIMENTO

Una definizione più rigorosa di probabilità include la definizione di un **insieme degli eventi**, ovvero di **combinazioni di esiti**

La probabilità è assegnata ad ogni **singolo evento**, e non all'esito (nel caso in cui gli eventi siano i singoli esiti, si ritorna alla nostra definizione intuitiva basata sugli esiti)

Esempio: Lancio di un dado. Supponiamo di essere interessati alla probabilità che esca un numero pari o un numero dispari

Definisco l'insieme degli eventi $\mathcal{P} = \{\{1,3,5\}, \{2,4,6\}\}$, i cui elementi (eventi) sono $\{1,3,5\}$ e $\{2,4,6\}$, ed assegno una probabilità ad ognuno di essi:

$$P(\{1,3,5\}) = 1/2$$

$$P(\{2,4,6\}) = 1/2$$



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

Outline

1. Definizione e proprietà delle variabili casuali: caso scalare
2. Definizione e proprietà delle variabili casuali: caso multivariabile
- 3. Stima e stimatori**
4. Proprietà degli stimatori



Teoria della stima

Per **gestire l'incertezza** presente nei dati (e.g. rumore di misura) **interpretiamo** i dati come variabili casuali. I **dati osservati** saranno i valori assunti dalle variabili casuali

In questo corso ci concentreremo sul problema della **stima parametrica**. Vogliamo quindi stimare il vettore di parametri θ^0 che ha generato i **dati** $\mathcal{D} = \{y(1), \dots, y(N)\}$

Esempio: Lancio di una moneta. Osserviamo $N = 8$ dati $\mathcal{D} = \{1,0,0,1,1,1,0,1\}$. In questo caso, il parametro di interesse θ^0 è la probabilità che esca testa

Quindi, i dati \mathcal{D} dipendono sia dall'esito s , sia dai parametri θ^0 $\rightarrow \mathcal{D}(s, \theta^0)$

I **dati osservati** dipendono da uno specifico esito \bar{s} $\rightarrow \mathcal{D} = \mathcal{D}(\bar{s}, \theta^0)$



Teoria della stima

Uno **stimatore** è una **funzione** $T(\mathcal{D}(s, \theta^0))$ dei dati (ovvero, una funzione di variabili casuali)

La **stima** è il risultato di uno stimatore su una specifica realizzazione dei dati $\mathcal{D}(\bar{s}, \theta^0)$

$$\hat{\theta} = T(\mathcal{D}(\bar{s}, \theta^0))$$

Osservazione

Poiché il risultato di $T(\)$ dipende dall'esito s (dal quale dipendono i dati), allora **lo stimatore è una variabile casuale** che dipende da s



Teoria della stima

Esempio: Supponiamo di voler **stimare l'altezza media** degli studenti e delle studentesse che seguono il corso di IMAD

Supponiamo di poter misurare solo $N = 10$ persone (se misurassimo tutti, non sarebbe più una stima, ma avremmo il valore vero del parametro «altezza media», cioè θ^0)

- **esito** s_1 : primi 10 studenti «estratti» $\rightarrow T(\mathcal{D}(s_1, \theta^0)) = \hat{\theta}_{(s_1)}$
- **esito** s_2 : altri 10 studenti «estratti» $\rightarrow T(\mathcal{D}(s_2, \theta^0)) = \hat{\theta}_{(s_2)} \neq \hat{\theta}_{(s_1)}$

La stima $\hat{\theta}_{(s)}$ fornita da $T(\quad)$ dipende da s . Quindi, lo **stimatore è una variabile casuale**



«Ha senso» parlare di **distribuzione di probabilità**,
valore atteso e **varianza** dello stimatore



Outline

1. Definizione e proprietà delle variabili casuali: caso scalare
2. Definizione e proprietà delle variabili casuali: caso multivariabile
3. Stima e stimatori
- 4. Proprietà degli stimatori**



Proprietà di uno stimatore

La «**bontà**» di uno stimatore non si giudica da una singola stima, ma dalle caratteristiche della sua distribuzione di probabilità

Correttezza (non polarizzazione, non deviato, unbiased)

Uno stimatore (scalare) $\hat{\theta}$ si dice **corretto**

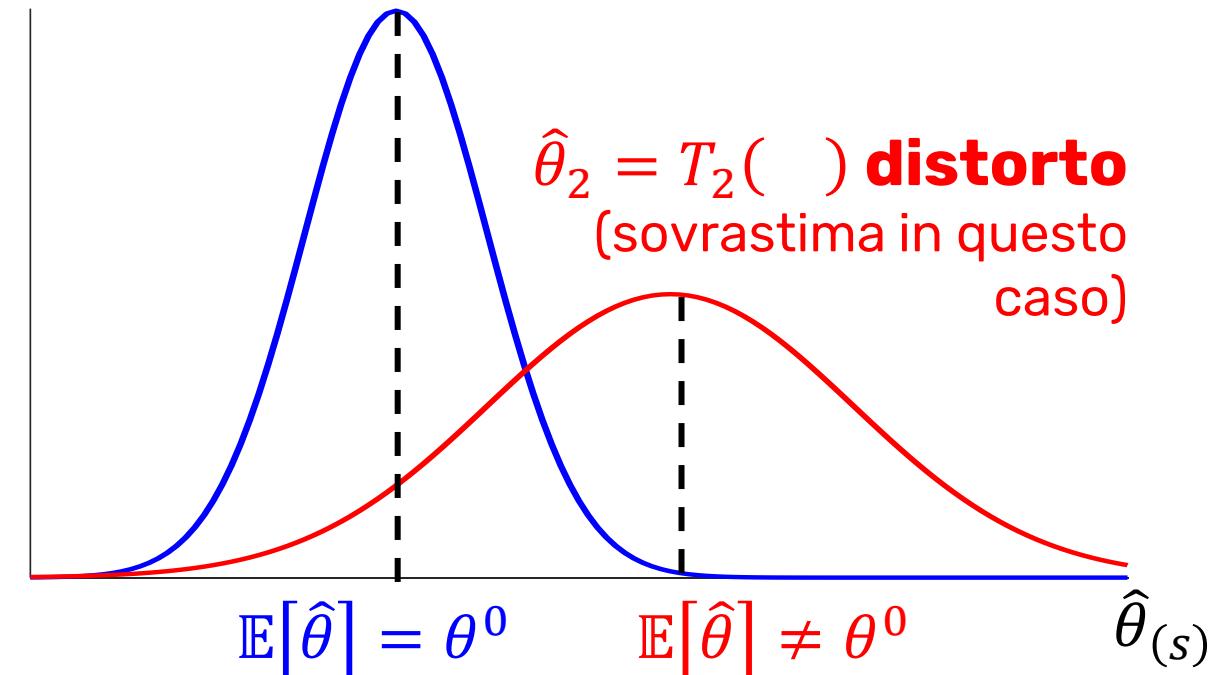
se $\mathbb{E}[\hat{\theta}] = \theta^0$, dove θ^0 è il valore vero del parametro

«In media» lo stimatore mi stima il valore vero del parametro

$$\text{bias} = \mathbb{E}[\hat{\theta}] - \theta^0$$

$\hat{\theta}_1 = T_1(\quad)$ **corretto**

$\hat{\theta}_2 = T_2(\quad)$ **distorto**
(sovraffima in questo caso)



Proprietà di uno stimatore

Correttezza asintotica

Uno stimatore (scalare) $\hat{\theta}$ si dice **asintoticamente corretto** se $\lim_{N \rightarrow +\infty} \mathbb{E}[\hat{\theta}] = \theta^0$

Proprietà più debole rispetto alla correttezza



Proprietà di uno stimatore

Consistenza

Uno stimatore (scalare) $\hat{\theta}$ si dice **consistente** se, per $N \rightarrow +\infty$, $\hat{\theta}$ **converge** a θ^0 in **probabilità**

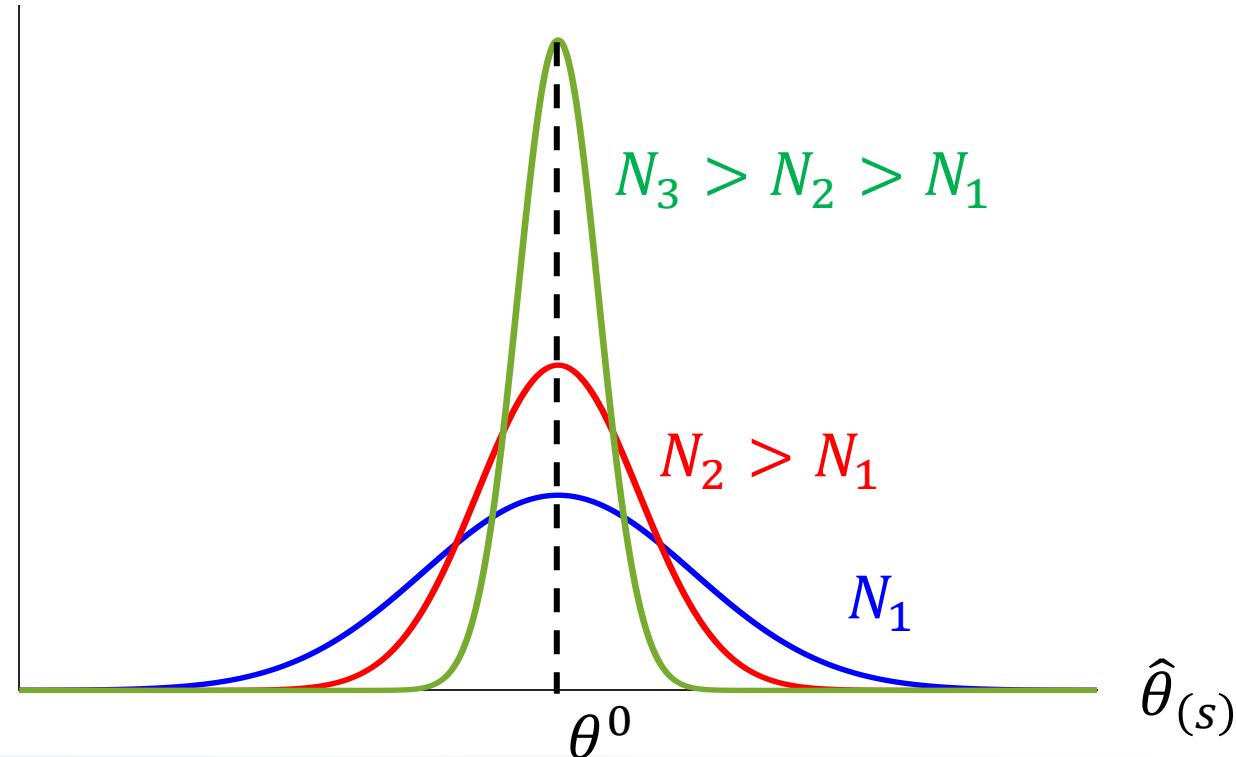
Al crescere di N , la stima diventa sempre più precisa (la probabilità di commettere un errore $\geq \varepsilon$ tende a 0)

Convergenza in media quadratica

$$\lim_{N \rightarrow +\infty} P(|\hat{\theta} - \theta^0|^2) = 0$$

Implica la convergenza in probabilità

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



Proprietà di uno stimatore

Cerchiamo di valutare la bontà di uno stimatore senza per forza far riferimento a proprietà asintotiche come la consistenza, quindi per N **finito**

Se due estimatori sono entrambi **corretti**, qual è il migliore? Quello a **minima varianza**



Quanto **piccola** può essere la varianza della stima?



Proprietà di uno stimatore

Limite di Cramer-Rao: Dato uno stimatore corretto $\hat{\theta}$, non possiamo rendere la sua varianza più piccola di una certa quantità

$$\text{Var}[\hat{\theta}] - M^{-1}$$

semidefinita positiva

Caso scalare $\text{Var}[\hat{\theta}] \geq 1/m$

Caso vettoriale

$$\text{Var}[\hat{\theta}] \geq M^{-1}$$

La quantità m (o M) è detta **quantità (matrice) di informazione di Fisher**

Intuizione: avrò sempre un certo livello di incertezza sui dati che uso per fare la stima, che non posso rimuovere. Quindi, i dati non saranno mai «informativi al 100%» proprio perché affetti da rumore. Esistono dei limiti «strutturali» alla stima



Proprietà di uno stimatore

Efficienza e efficienza asintotica

Uno stimatore (scalare) $\hat{\theta}$ si dice **efficiente** se $\text{Var}[\hat{\theta}] = 1/m$

Uno stimatore (scalare) $\hat{\theta}$ si dice **asintoticamente efficiente** se $\lim_{N \rightarrow +\infty} \text{Var}[\hat{\theta}] = 1/m$

Minima varianza

Uno stimatore (scalare) $\hat{\theta}^m$ corretto si dice **a minima varianza** se $\text{Var}[\hat{\theta}^m] \leq \text{Var}[\hat{\theta}]$, dove $\hat{\theta}$ è un qualsiasi stimatore corretto

- Se $\hat{\theta}$ è efficiente, allora è a minima varianza
- **Non vale il viceversa.** Ci sono casi in cui esistono stimatori a minima varianza che non sono efficienti. Questo accade quando non esistono stimatori che raggiungono il limite di Cramer-Rao

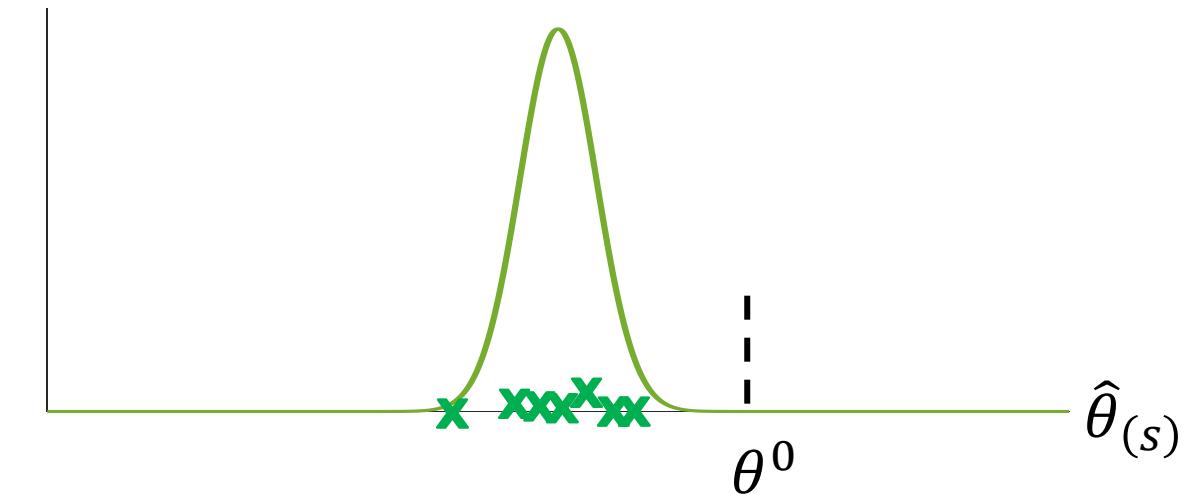
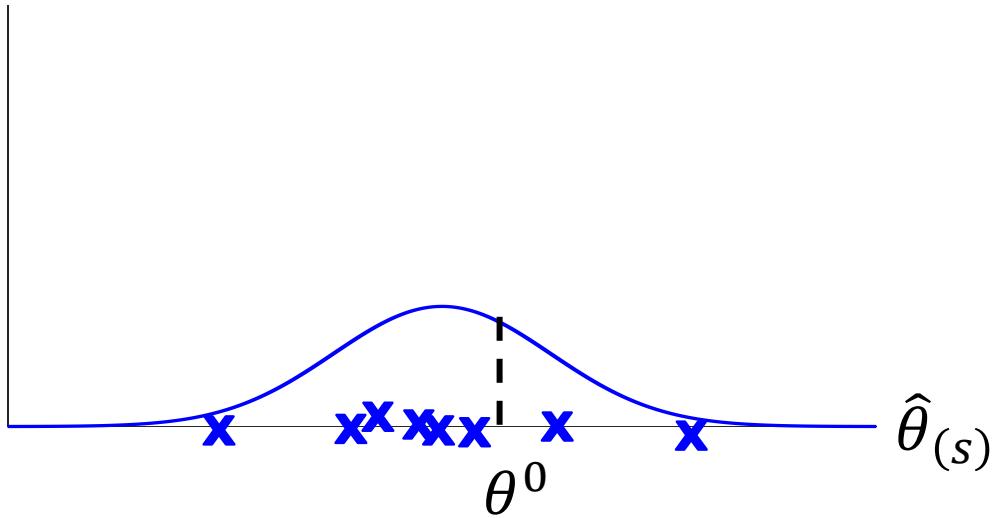


Proprietà di uno stimatore

Mean squared error

Per estimatori **non corretti** (distorti, polarizzati), la varianza, da sola, non è sufficiente come criterio di bontà

Varianza più piccola ma
stimatore **«peggiore»** In che senso?



Proprietà di uno stimatore

Abbiamo bisogno di un indicatore «globale», che consideri sia il bias sia la varianza

Idea: uso come criterio **l'errore quadratrico medio** (MSE – mean squared error)

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

- Caso θ^0 scalare

Proprietà

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

BIAS-VARIANCE dilemma

Questa proprietà tornerà utile quando vorremo stimare (identificare) modelli dai dati. In quel caso, il «soggetto» non sarà un parametro θ quanto piuttosto l'intero modello (che è di fatto uno stimatore di una funzione)



Esempio (stimatore della media)

Siano $\mathcal{D} = \{y(1), y(2), \dots, y(N)\}$ variabili casuali con media μ e varianza σ^2 . Lo **stimatore media campionaria $\hat{\mu}$** è **corretto** e **consistente**

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

In questo caso il parametro di interesse θ è la media μ della popolazione

Vogliamo dimostrare la correttezza, ovvero che $\mathbb{E}[\hat{\mu}] = \mu$

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N y(i)\right] = \frac{1}{N} \mathbb{E}\left[\sum_{i=1}^N y(i)\right] = \frac{1}{N} \mathbb{E}[y(1) + y(2) + \dots + y(N)] = \frac{1}{N} \cdot N \cdot \mu = \mu$$



Esempio (stimatore della varianza)

Siano $\mathcal{D} = \{y(1), y(2), \dots, y(N)\}$ variabili casuali con media μ e varianza σ^2 . Lo **stimatore varianza campionaria S_{N-1}^2** è **corretto**

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

Esercizio: dimostrare la correttezza di S_{N-1}^2 . Suggerimenti:

- Usare la proprietà di linearità del valore atteso
- Usare la proprietà della varianza tale che $\text{Var}[v] = \mathbb{E}[v^2] - \mathbb{E}[v]^2$





**UNIVERSITÀ
DEGLI STUDI
DI BERGAMO**

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



IDENTIFICAZIONE DEI MODELLI E ANALISI DEI DATI (IMAD)

Lezione 3: Regressione lineare

**Corso di Laurea Magistrale in
INGEGNERIA INFORMATICA**

SPEAKER
Prof. Mirko Mazzoleni

PLACE
Università degli Studi di
Bergamo

Syllabus

Parte I: sistemi statici

1. Richiami di statistica

2. Teoria della stima

2.1 Proprietà degli stimatori

3. Stima a minimi quadrati

3.1 Stima di modelli lineari

3.2 Algoritmo del gradient descent

4. Stima a massima verosimiglianza

4.1 Proprietà della stima

4.2 Stima di modelli lineari

5. Regressione logistica

5.1 Stima di un modello di regressione logistica

6. Fondamenti di machine learning

6.1 Bias-Variance tradeoff

6.2 Overfitting

6.3 Regolarizzazione

6.4 Validazione

7. Cenni di stima Bayesiana

7.1 Probabilità congiunte, marginali e condizionate

7.2 Connessione con Filtro di Kalman



Parte I: sistemi statici**Parte II: sistemi dinamici**Stima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

✓ Stima parametri popolazione

✓ Stima modello lineare: minimi quadrati

○ **SI assunzioni su ddp dei dati**

✓ Stima massima verosimiglianza parametri popolazione

✓ Stima modello lineare: massiva verosimiglianza

✓ Regressione logistica

• θ variabile casuale○ **SI assunzioni su ddp dei dati**

✓ Stima Bayesiana

Stima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

✓ Modelli lineari di pss

✓ Predizione

✓ Identificazione

✓ Persistente eccitazione

✓ Analisi asintotica metodi PEM

✓ Analisi incertezza stima (numero dati finito)

✓ Valutazione del modello

Machine learning

Outline

1. Stima a minimi quadrati
2. Funzione di costo
3. Gradient descent
4. Proprietà dello stimatore a minimi quadrati
5. Esercizi con codice



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

Outline

- 1. Stima a minimi quadrati**
2. Funzione di costo
3. Gradient descent
4. Proprietà dello stimatore a minimi quadrati
5. Esercizi con codice



Stima a minimi quadrati (least squares)

Abbiamo finora descritto i dati $\mathcal{D} = \{y(1), y(2), \dots, y(N)\}$ in termini della loro media e varianza, dando degli stimatori per queste quantità

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

Supponiamo ora di **voler descrivere** (cioè, assumiamo che i dati abbiamo questa struttura) i dati tramite una **relazione lineare**

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



Stima a minimi quadrati (least squares)

Obiettivo: Supponiamo di avere a disposizione N dati $\mathcal{D} = \{(\boldsymbol{\varphi}(1), y(1)), \dots, (\boldsymbol{\varphi}(N), y(N))\}$.

Trovare la relazione tra le variabili di input (regressori, features) $\boldsymbol{\varphi} \in \mathbb{R}^{(d-1) \times 1}$ e una variabile di output $y \in \mathbb{R}$, usando un **modello lineare**

$$\begin{aligned} y(i) &= \theta_0 + \theta_1 \varphi_1(i) + \cdots + \theta_{d-1} \varphi_{d-1}(i) + \epsilon(i) = \sum_{j=0}^{d-1} \theta_j \varphi_j(i) + \epsilon(i) \\ &= \underbrace{\boldsymbol{\varphi}^\top(i)}_{\substack{1 \times d \\ [\dots]}} \underbrace{\boldsymbol{\theta}}_{\substack{d \times 1 \\ [\vdots]}} + \epsilon(i) \quad \begin{aligned} &\bullet \quad \varphi_0 = 1 \\ &\bullet \quad \boldsymbol{\varphi} = [\varphi_0, \varphi_1, \dots, \varphi_{d-1}]^\top \in \mathbb{R}^{d \times 1} \\ &\bullet \quad \boldsymbol{\theta} = [\theta_0, \theta_1, \dots, \theta_{d-1}]^\top \in \mathbb{R}^{d \times 1} \end{aligned} \end{aligned}$$

- Il vettore $\boldsymbol{\theta} \in \mathbb{R}^{d \times 1}$ è il **vettore dei parametri**
- Il vettore $\boldsymbol{\varphi}(i) \in \mathbb{R}^{d \times 1}$ è il **vettore delle features** per la i -esima osservazione
- La quantità $\epsilon(i) \in \mathbb{R}$ è l'errore dovuto ad una non perfetta spiegazione di $y(i)$ tramite $\boldsymbol{\varphi}(i)$



Esempio (stimare il prezzo delle case)

Numero di osservazioni N

Area (feet ²)	# Camere da letto	# Piani	Età	Prezzo (1000\$)
2104	5	1	45	115
1416	3	2	40	150
1534	2	1	30	210
:	:	:	:	:

Singola feature φ_3

Variabile di output y

Singola osservazione (regressore\features vector) φ

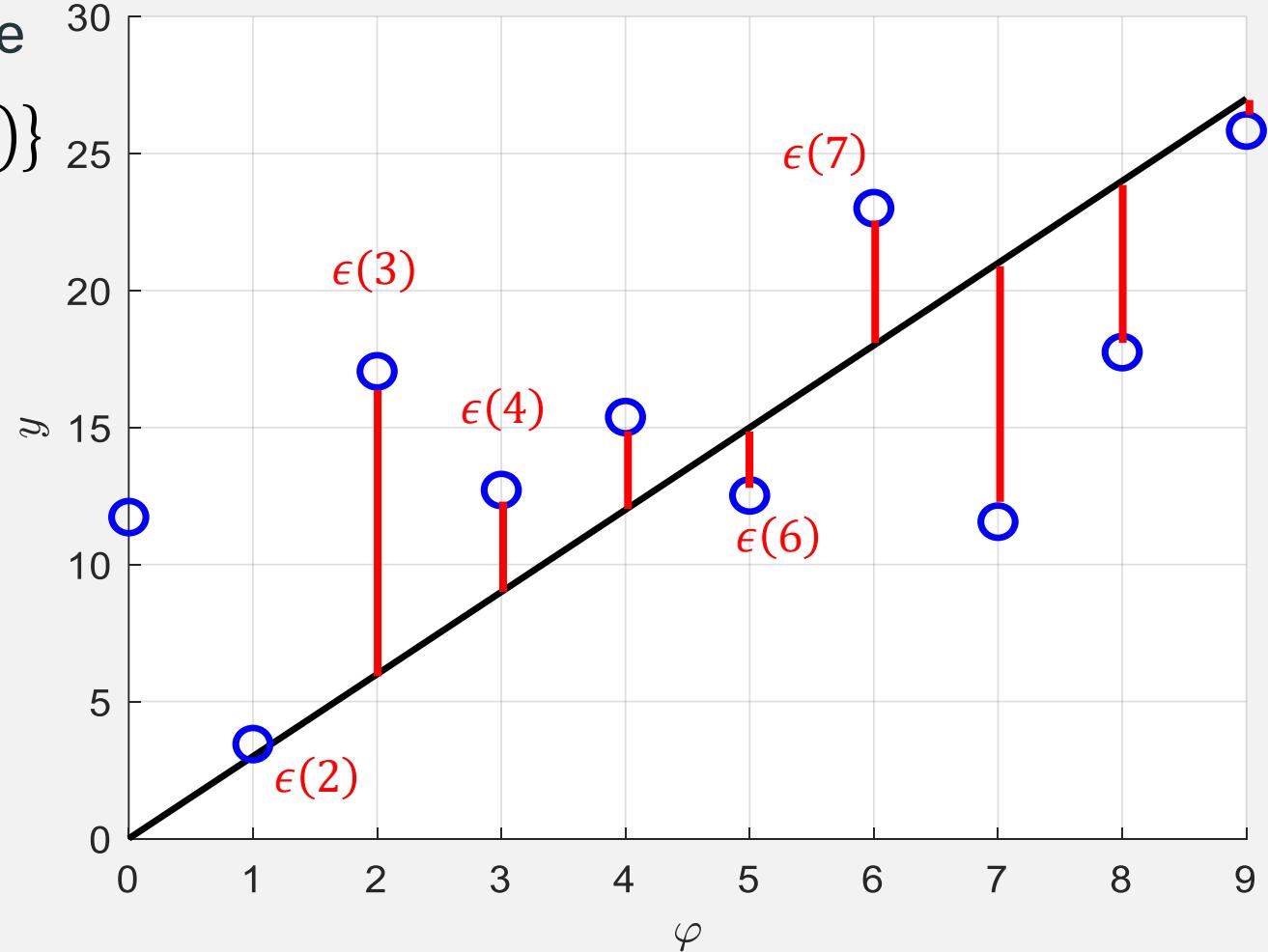
- Il numero delle righe è il numero di osservazioni N
- L'osservazione i -esima è il vettore $\varphi(i) = [\varphi_1(i) \ \varphi_2(i) \ \varphi_3(i) \ \varphi_4(i)]^\top \in \mathbb{R}^{4 \times 1}$
- Ogni regressore φ ha associata una risposta $y \in \mathbb{R}$ che vogliamo stimare



QUIZ!

Nel grafico seguente, quante osservazioni $\{(\varphi(1), y(1)), \dots, (\varphi(N), y(N))\}$ abbiamo?

- $N = 10$ osservazioni
- $N = 7$ osservazioni
- $N = 9$ osservazioni

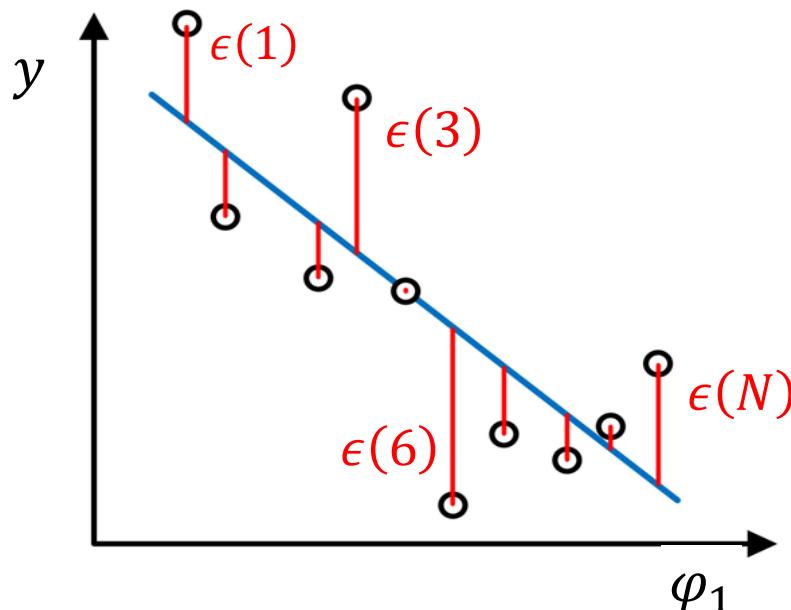


Interpretazione geometrica

Caso scalare (retta)

In questo caso c'è **un solo regressore** φ_1
e **due parametri** θ_0, θ_1

$$y(i) = \theta_0 + \theta_1 \varphi_1(i) + \epsilon(i)$$



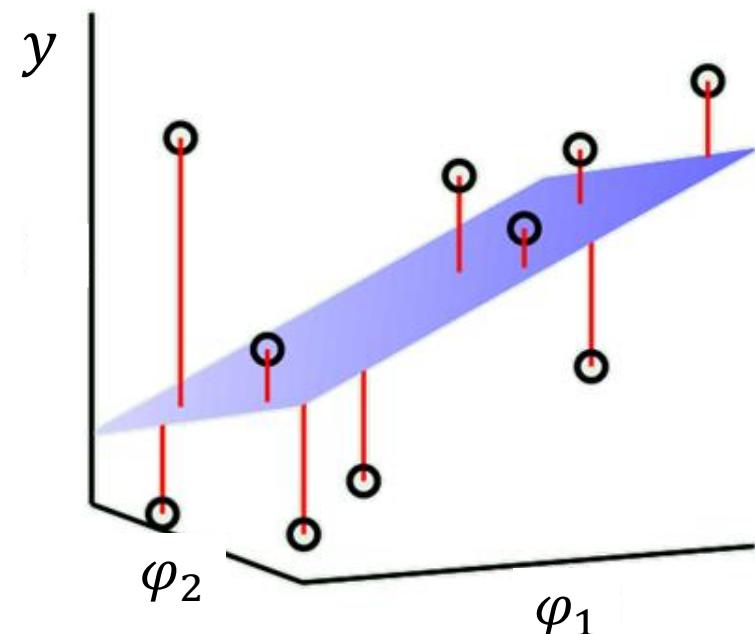
Esempio

- y : peso [kg]
- φ_1 : altezza [m]
- φ_2 : età

Caso con 2 regressori (piano)

In questo caso ci sono **due regressori** φ_1, φ_2 e **tre parametri** $\theta_0, \theta_1, \theta_2$

$$y(i) = \theta_0 + \theta_1 \varphi_1(i) + \theta_2 \varphi_2(i) + \epsilon(i)$$



Outline

1. Stima a minimi quadrati

2. Funzione di costo

3. Gradient descent

4. Proprietà dello stimatore a minimi quadrati

5. Esercizi con codice



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

Funzione di costo

Regressione lineare: modello lineare + minimi quadrati

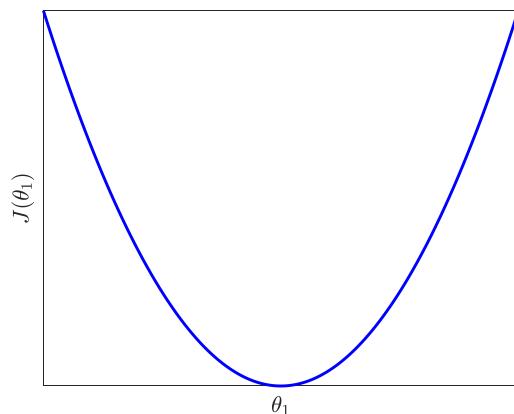
Il metodo della regressione lineare stima i parametri θ minimizzando l'errore quadratico tra **output osservati** e **stimati** dal modello lineare

$$\hat{\theta} = \arg \min_{\theta} J(\theta)$$

Funzione di costo
(cifra di merito)

**Caso scalare senza
intercetta, $\theta_0 = 0$**

$$y(i) = \theta_1 \varphi_1(i) + \epsilon(i)$$

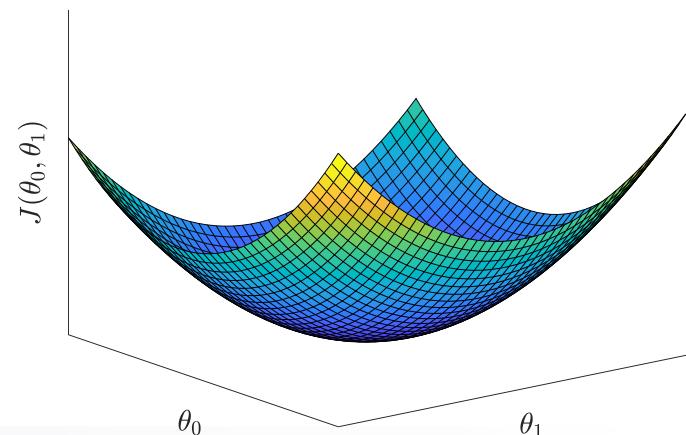


$$J(\theta) = \frac{1}{N} \sum_{i=1}^N (y(i) - \varphi^\top(i)\theta)^2 = \frac{1}{N} \sum_{i=1}^N \epsilon(i)^2$$

$y(i)$ $\varphi^\top(i)$ $\epsilon(i)$

**Caso scalare con
intercetta, $\theta_0 \neq 0$**

$$y(i) = \theta_0 + \theta_1 \varphi_1(i) + \epsilon(i)$$



Minimizzazione della funzione di costo

$$J(\theta) = \frac{1}{N} \sum_{i=1}^N (\textcolor{brown}{y}(i) - \boldsymbol{\varphi}(i)^\top \boldsymbol{\theta})^2$$

$$\nabla J(\boldsymbol{\theta}) = \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0} \Rightarrow \frac{2}{N} \sum_{i=1}^N \boldsymbol{\varphi}(i) \cdot (y(i) - \boldsymbol{\varphi}^\top(i) \boldsymbol{\theta}) = \mathbf{0} \Rightarrow \sum_{i=1}^N \boldsymbol{\varphi}(i)y(i) - \sum_{i=1}^N \boldsymbol{\varphi}(i)\boldsymbol{\varphi}^\top(i)\boldsymbol{\theta} = \mathbf{0}$$
$$\Rightarrow \left[\sum_{i=1}^N \boldsymbol{\varphi}(i)\boldsymbol{\varphi}^\top(i) \right] \boldsymbol{\theta} = \sum_{i=1}^N \boldsymbol{\varphi}(i)y(i)$$
$$\Rightarrow \widehat{\boldsymbol{\theta}} = \left[\sum_{i=1}^N \boldsymbol{\varphi}(i)\boldsymbol{\varphi}^\top(i) \right]^{-1} \cdot \left[\sum_{i=1}^N \boldsymbol{\varphi}(i)y(i) \right]$$

Poiché il modello è **lineare nei parametri** e la misura dell'errore è **quadratica**, la funzione di costo è **convessa** → ammette un **minimo unico** (globale)

Nel caso della regressione lineare, il minimo può anche essere trovato in **forma chiusa**



Funzione di costo: caso matriciale

Possiamo esprimere il problema della regressione lineare usando delle matrici

Vettore dei regressori $\varphi^\top(1)_{1 \times d}$

$$X = \begin{bmatrix} 1 & \varphi_1(1) & \varphi_2(1) & \cdots & \varphi_{d-1}(1) \\ 1 & \varphi_1(2) & \varphi_2(2) & & \varphi_{d-1}(2) \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & \varphi_1(N) & \varphi_2(N) & \cdots & \varphi_{d-1}(N) \end{bmatrix}_{N \times d}$$

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_{d-1} \end{bmatrix}_{d \times 1} \quad Y = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix}_{N \times 1} \quad E = \begin{bmatrix} \epsilon(1) \\ \epsilon(2) \\ \vdots \\ \epsilon(N) \end{bmatrix}_{N \times 1}$$

$$= \begin{bmatrix} \varphi^\top(1) \\ \varphi^\top(2) \\ \vdots \\ \varphi^\top(N) \end{bmatrix}_{N \times 1} \quad Y = X\theta^{d \times 1} + E \Rightarrow$$

$$J(\theta) = \frac{1}{N} \|Y - X\theta\|_2^2 = \frac{1}{N} (Y - X\theta)^{\top} (Y - X\theta)_{1 \times N}$$



Funzione di costo: caso matriciale

È utile ricordare queste proprietà di derivazione matriciale (https://en.wikipedia.org/wiki/Matrix_calculus)

$$\nabla_{\mathbf{x}}(\mathbf{x}^T \cdot A \cdot \mathbf{x}) = (A + A^T) \cdot \mathbf{x}$$

$1 \times d \quad d \times d \quad d \times 1$

$$\nabla_{\mathbf{x}}(\mathbf{x}^T \cdot \mathbf{b}) = \mathbf{b}$$

$1 \times d \quad d \times 1 \quad d \times 1$

$$\begin{aligned} J(\boldsymbol{\theta}) &= \frac{1}{N} (Y - X\boldsymbol{\theta})^T (Y - X\boldsymbol{\theta}) = \frac{1}{N} (Y^T Y - Y^T X \cdot \boldsymbol{\theta}^T - \boldsymbol{\theta}^T \cdot X^T Y + \boldsymbol{\theta}^T \cdot X^T X \cdot \boldsymbol{\theta}) \\ &= \frac{1}{N} (Y^T Y - 2 \cdot \boldsymbol{\theta}^T \cdot X^T Y + \boldsymbol{\theta}^T \cdot X^T X \cdot \boldsymbol{\theta}) \end{aligned}$$

$$\nabla J(\boldsymbol{\theta}) = \mathbf{0} \Rightarrow \frac{1}{N} (-2X^T Y + 2X^T X \boldsymbol{\theta}) = \mathbf{0} \Rightarrow$$

$$\widehat{\boldsymbol{\theta}} = (X^T X)^{-1} X^T Y$$

$d \times 1 \quad d \times d \quad d \times N \quad N \times 1$



Normal equations

$$\hat{\theta} = (X^T X)^{-1} X^T Y$$

Normal equations

Cosa succede se $X^T X$ **non è invertibile**?



Si usa la **pseudo-inversa**. In MatLab:

```
theta_hat = pinv(X' * X) * X * Y
```

- **Regressori ridondanti** (linearmente dipendenti)

- ✓ φ_1 = altezza in m
- ✓ φ_2 = altezza in feet

- Il metodo delle normal equation è **lento** se d è molto grande
 - ✓ Per risolvere questo problema, si usano **metodi iterativi** come il **gradient descent**



Outline

1. Stima a minimi quadrati
2. Funzione di costo
- 3. Gradient descent**
4. Proprietà dello stimatore a minimi quadrati
5. Esercizi con codice



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

Gradient descent

Il **gradient descent** è un metodo **iterativo** per minimizzare le funzioni differenziabili (ovvero funzioni in cui possiamo calcolare le derivate in ogni punto del dominio)

Consideriamo prima il **caso scalare** (abbiamo un solo parametro $\theta \in \mathbb{R}$ da stimare)

Dato un valore iniziale $\hat{\theta}^{(0)}$, la stima $\hat{\theta}^{(k+1)}$ del parametro θ all'iterazione $k + 1$ è:

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

$\alpha \in \mathbb{R}_{>0}$: learning rate



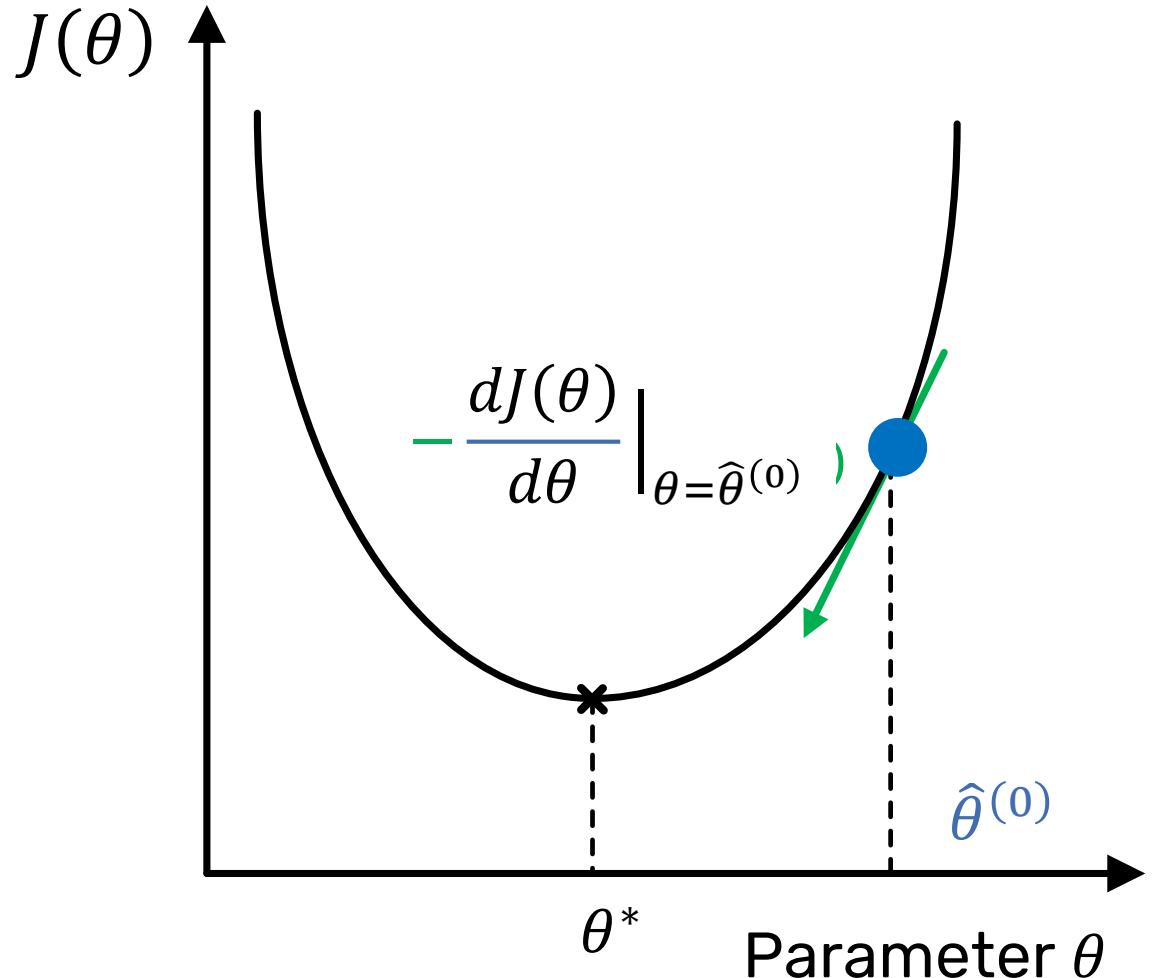
Gradient descent

Caso scalare $\theta \in \mathbb{R}$

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

$$\frac{dJ(\theta)}{d\theta} \Big|_{\theta=\hat{\theta}^{(k)}} > 0 \Rightarrow \hat{\theta}^{(k+1)} < \hat{\theta}^{(k)}$$

La nuova stima è più vicina al valore ottimale θ^*



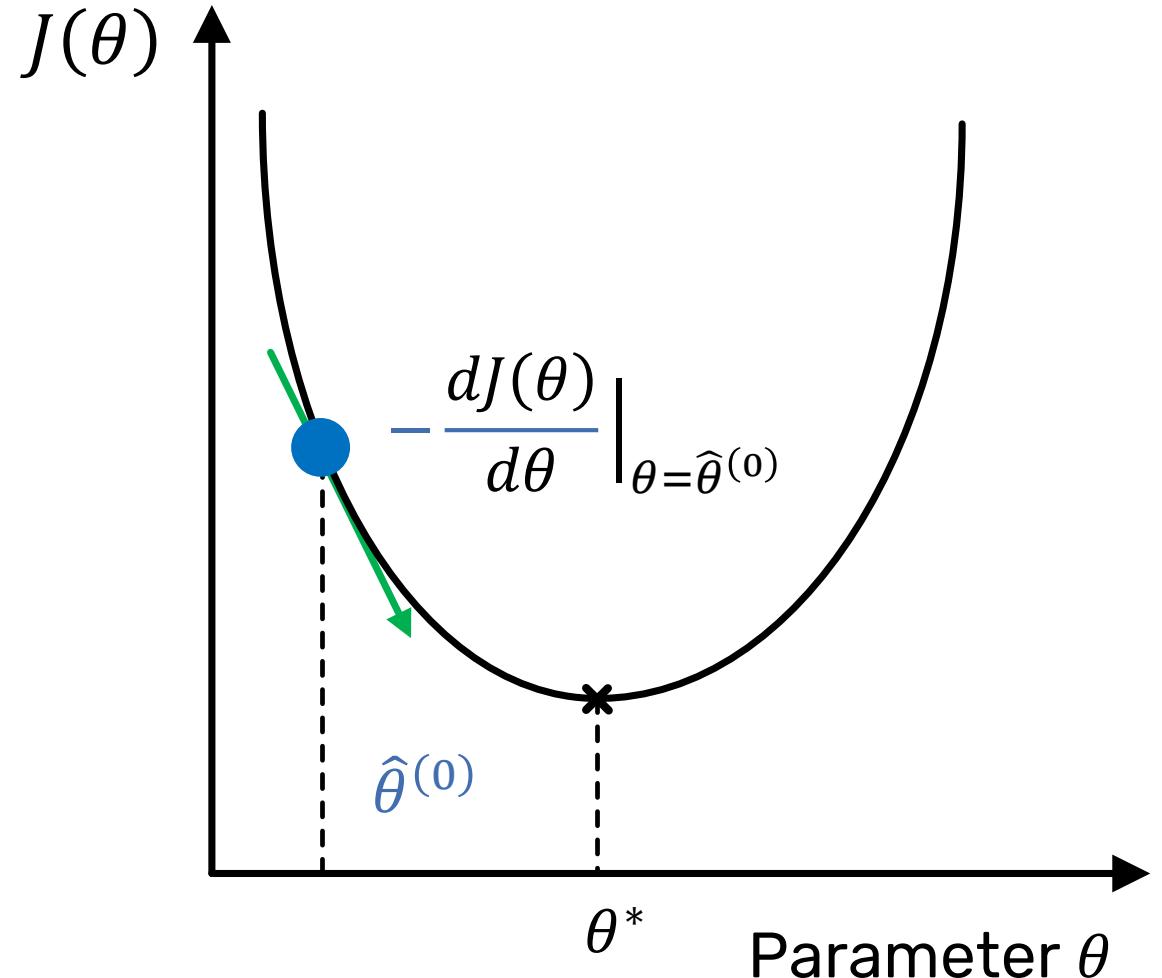
Gradient descent

Caso scalare $\theta \in \mathbb{R}$

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

$$\frac{dJ(\theta)}{d\theta} \Big|_{\theta=\hat{\theta}^{(k)}} < 0 \Rightarrow \hat{\theta}^{(k+1)} > \hat{\theta}^{(k)}$$

La nuova stima è più vicina al valore ottimale θ^*



Gradient descent

Nel caso generale **multivariabile** (i.e. stimare un vettore di parametri $\theta \in \mathbb{R}^{d \times 1}$), dobbiamo sostituire la derivata con il **vettore gradiente** $\nabla J(\theta) \in \mathbb{R}^{d \times 1}$

Dato un valore iniziale $\hat{\theta}^{(0)}$, la stima $\hat{\theta}^{(k+1)}$ del vettore di parametri θ all'iterazione $k + 1$ è:

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

$\alpha \in \mathbb{R}_{>0}$: learning rate



Gradient descent: trick computazionale

Quando sono presenti più regressori (caso **multivariabile**) è utile normalizzarne i valori, in modo che l'algoritmo del gradient descent «faccia meno fatica» a raggiungere il minimo

Calcolo la media per ogni regressore (che non sia quello dell'intercetta)

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

Calcolo la varianza per ogni regressore (che non sia quello dell'intercetta)

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

Sottraggo media e divido per deviazione standard

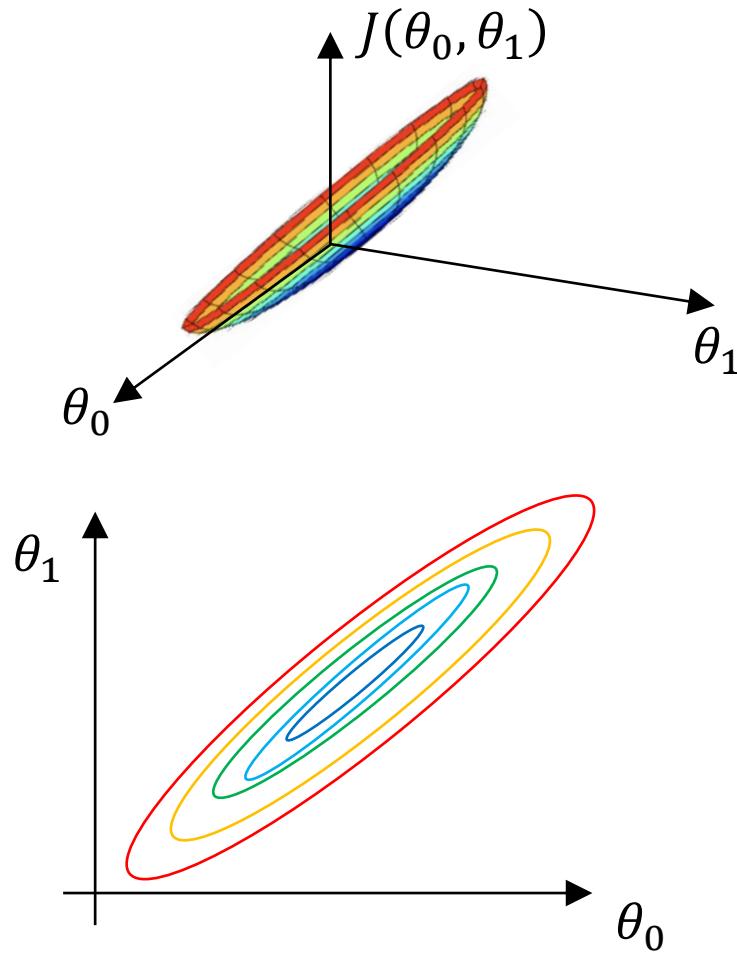
$$\varphi_j(i) = \frac{\varphi_j(i) - \hat{\mu}_j}{\sqrt{\hat{\sigma}_j^2}} \quad j = 1, \dots, d - 1$$

Normalizzare i nuovi dati usando LA STESSA MEDIA E LA STESSA VARIANZA calcolata sul dataset usato per stimare il modello

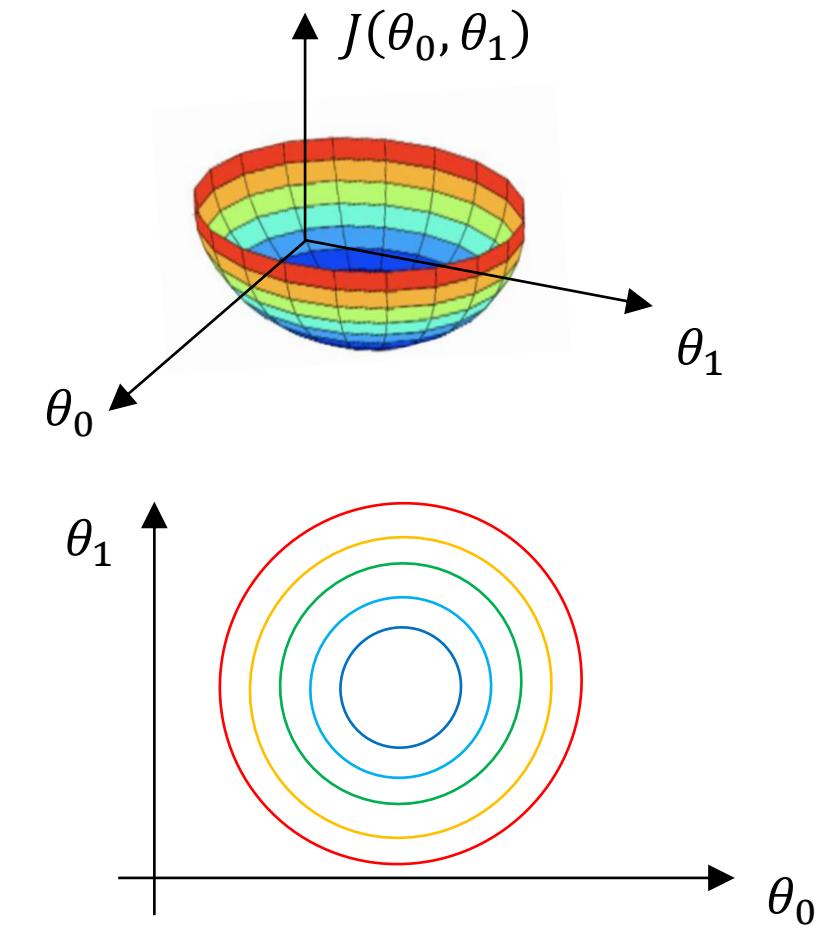


Gradient descent: trick computazionale

Regressori non normalizzati



Regressori normalizzati



Outline

1. Stima a minimi quadrati
2. Funzione di costo
3. Gradient descent
- 4. Proprietà dello stimatore a minimi quadrati**
5. Esercizi con codice



Proprietà dello stimatore a minimi quadrati

Dubbio legittimo: come si comporta lo **stimatore a minimi quadrati** di un modello lineare nel caso in cui il sistema vero (che genera i dati) sia effettivamente lineare?

$$y(i) = \boldsymbol{\varphi}^T(i)\boldsymbol{\theta}^0 + \epsilon(i)$$

Supponiamo che $\epsilon(i)$ sia una variabile casuale a media nulla, con una certa varianza λ^2

Nota: non stiamo assumendo nessuna specifica distribuzione di probabilità su $\epsilon(i)$

Proprietà dello stimatore a minimi quadrati (nel caso del sistema di cui sopra)

- Lo stimatore è **corretto**: $\mathbb{E}[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}^0$
- Supponendo inoltre che i rumori siano incorrelati $\mathbb{E}[\epsilon(i)\epsilon(j)] = 0, \forall i \neq j$, lo stimatore è **consistente**: $\text{Var}[\hat{\boldsymbol{\theta}}] = \lambda^2 \cdot (X^T X)^{-1} = \lambda^2 P(t)$



Outline

1. Stima a minimi quadrati
2. Funzione di costo
3. Gradient descent
4. Proprietà dello stimatore a minimi quadrati

5. Esercizi con codice



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

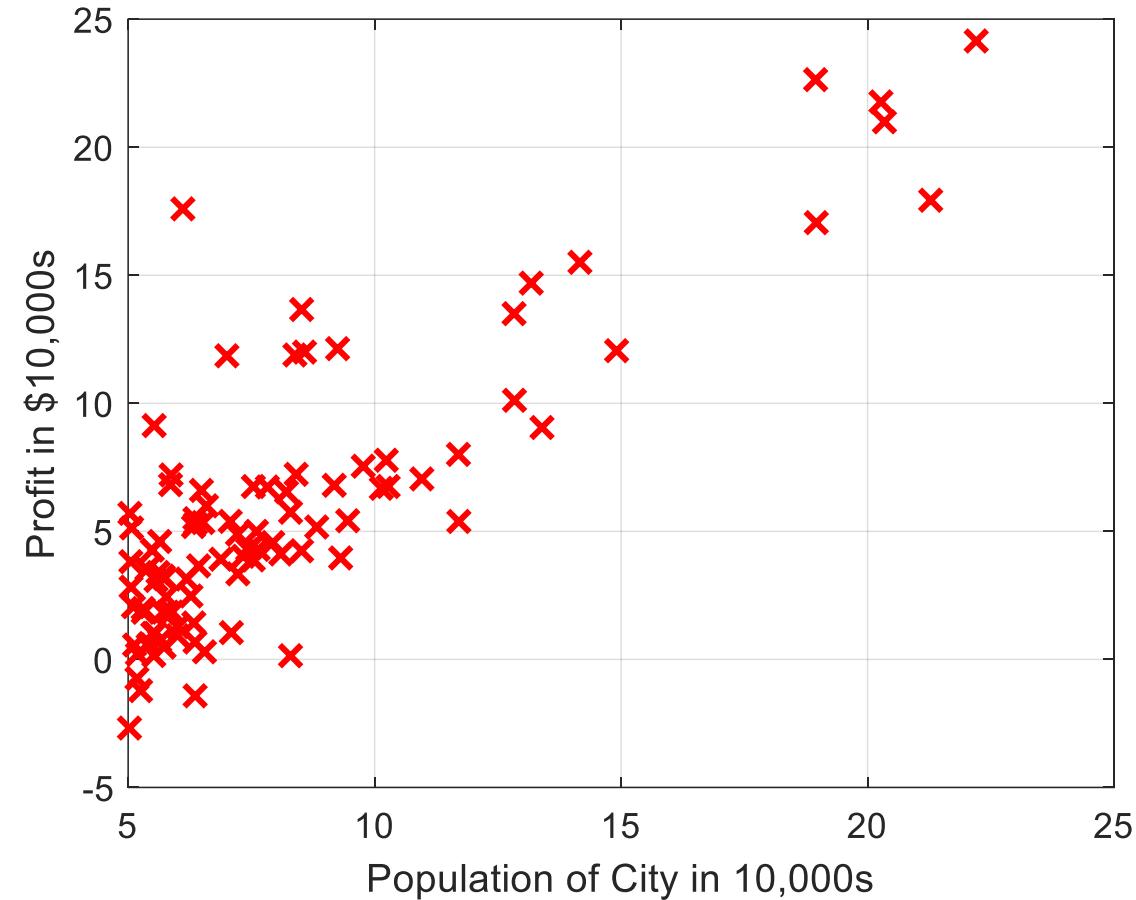
Esercizio 1: Stima dei profitti di un ristorante

Problema: il CEO di un franchising di ristoranti che sta valutando diverse città per l'apertura di un nuovo ristorante.

La catena ha già ristoranti in varie città e sono disponibili dati di profitti e di popolazione di questa città.

L'obiettivo è utilizzare questi dati per selezionare in quale città aprire la nuova attività

- Ogni città è descritta da:
 - ✓ φ_1 : Popolazione [in 10000 unità]
- ✓ L'output y è il profitto [in 10000\$]
- Il dataset consiste di $N = 97$ città con $\varphi_1(i)$, e $y(i)$, per $i = 1, \dots, N$



Esercizio 2: stima dei prezzi delle case

Vogliamo **stimare il prezzo** delle case a Portland, Oregon. L'output y è quindi il prezzo

- Ogni casa è descritta da:
 - ✓ φ_1 : Area [feet²]
 - ✓ φ_2 : Numero di camere da letto
- Il dataset consiste di $N = 47$ case con $\varphi_1(i), \varphi_2(i)$ e $y(i)$, per $i = 1, \dots, N$

$$y(i) = \boldsymbol{\varphi}^\top(i) \boldsymbol{\theta} + \epsilon(i) \quad \boldsymbol{\varphi}(i) = [1 \quad \varphi_1(i) \quad \varphi_2(i)]^\top$$

$$X = \begin{bmatrix} \boldsymbol{\varphi}^\top(1) \\ \boldsymbol{\varphi}^\top(2) \\ \vdots \\ \boldsymbol{\varphi}^\top(N) \end{bmatrix}_{47 \times 3} \quad \boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}_{3 \times 1} \quad Y = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(47) \end{bmatrix}_{47 \times 1}$$

```
% Read data from file
data = csvread('ex2data.txt');
X = data(:, 1:2); % Features
y = data(:, 3); % Price
N = length(y); % Number of data

% Add intercept term to X
X = [ones(N, 1) X];

% Calculate the parameters from the
% normal equation
theta_hat = pinv(X'*X)*X'*y;

% Estimate the price of a 1650 sq-
% ft, 3 br house
price_hat = [1 3 1650]*theta_hat;
```

Punto non visto durante la stima di $\boldsymbol{\theta}$



Calcolare e implementare il gradiente

Come calcoliamo il gradiente? Supponiamo che il nostro modello sia

$$y = \theta_0 + \theta_1 \cdot \varphi + \epsilon$$

$$J(\theta_0, \theta_1) = \frac{1}{N} \sum_{i=1}^N (y(i) - \theta_0 - \theta_1 \cdot \varphi(i))^2 \quad \rightarrow \quad \nabla J(\theta_0, \theta_1) = \begin{bmatrix} \frac{\partial J(\theta_0, \theta_1)}{\partial \theta_0} & \frac{\partial J(\theta_0, \theta_1)}{\partial \theta_1} \end{bmatrix}^\top$$

$$\frac{\partial J(\theta_0, \theta_1)}{\partial \theta_0} = \frac{2}{N} \sum_{i=1}^N (y(i) - \theta_0 - \theta_1 \cdot \varphi(i)) \cdot (-1) = -\frac{2}{N} X(:, 1)^\top \cdot (Y - X\theta)$$

$$\frac{\partial J(\theta_0, \theta_1)}{\partial \theta_1} = \frac{2}{N} \sum_{i=1}^N (y(i) - \theta_0 - \theta_1 \cdot \varphi(i)) \cdot (-\varphi(i)) = -\frac{2}{N} X(:, 2)^\top \cdot (Y - X\theta)$$



Calcolare e implementare il gradiente

In generale, se abbiamo **più di un regressore** (ovvero, un vettore $\varphi = [1 \ \varphi_1, \ \varphi_2, \dots, \varphi_{d-1}]^\top \in \mathbb{R}^{d \times 1}$) possiamo implementare il gradient descent come di seguito:

For {

$$\theta_0 = \theta_0 - \alpha \cdot \frac{2}{N} \sum_{i=1}^N (y(i) - \varphi^\top(i) \boldsymbol{\theta}) \cdot (-1)$$

$$\theta_1 = \theta_1 - \alpha \cdot \frac{2}{N} \sum_{i=1}^N (y(i) - \varphi^\top(i) \boldsymbol{\theta}) \cdot (-\varphi_1(i))$$

:

$$\theta_{d-1} = \theta_{d-1} - \alpha \cdot \frac{2}{N} \sum_{i=1}^N (y(i) - \varphi^\top(i) \boldsymbol{\theta}) \cdot (-\varphi_{d-1}(i))$$

}





**UNIVERSITÀ
DEGLI STUDI
DI BERGAMO**

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



IDENTIFICAZIONE DEI MODELLI E ANALISI DEI DATI (IMAD)

**Lezione 4: Stima a massima
verosimiglianza (maximum likelihood
estimation)**

**Corso di Laurea Magistrale in
INGEGNERIA INFORMATICA**

SPEAKER
Prof. Mirko Mazzoleni

PLACE
Università degli Studi di
Bergamo

Syllabus

Parte I: sistemi statici

1. Richiami di statistica

2. Teoria della stima

2.1 Proprietà degli stimatori

3. Stima a minimi quadrati

3.1 Stima di modelli lineari

3.2 Algoritmo del gradient descent

4. Stima a massima verosimiglianza

4.1 Proprietà della stima

4.2 Stima di modelli lineari

5. Regressione logistica

5.1 Stima di un modello di regressione logistica

6. Fondamenti di machine learning

6.1 Bias-Variance tradeoff

6.2 Overfitting

6.3 Regolarizzazione

6.4 Validazione

7. Cenni di stima Bayesiana

7.1 Probabilità congiunte, marginali e condizionate

7.2 Connessione con Filtro di Kalman



Parte I: sistemi statici**Parte II: sistemi dinamici**Stima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

✓ Stima parametri popolazione

✓ Stima modello lineare: minimi quadrati

○ **SI assunzioni su ddp dei dati**

✓ Stima massima verosimiglianza parametri popolazione

✓ Stima modello lineare: massiva verosimiglianza

✓ Regressione logistica

• θ variabile casuale○ **SI assunzioni su ddp dei dati**

✓ Stima Bayesiana

Stima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

✓ Modelli lineari di pss

✓ Predizione

✓ Identificazione

✓ Persistente eccitazione

✓ Analisi asintotica metodi PEM

✓ Analisi incertezza stima (numero dati finito)

✓ Valutazione del modello

Machine learning

Outline

1. Stima a massima verosimiglianza
2. Stima a massima verosimiglianza di parametri della popolazione
3. Stima a massima verosimiglianza di modelli lineari



Outline

- 1. Stima a massima verosimiglianza**
2. Stima a massima verosimiglianza di parametri della popolazione
3. Stima a massima verosimiglianza di modelli lineari



Stima a massima verosimiglianza

Abbiamo presentato finora diversi tipi di estimatori:

- **Media campionaria:** $\hat{\mu} = \frac{1}{N} \sum_{i=1}^N y(i) \quad \Rightarrow \quad \hat{\theta} = \mu \in \mathbb{R}$
- **Varianza campionaria:** $S_{N-1}^2 = \frac{1}{N-1} \cdot \sum_{i=1}^N (y(i) - \hat{\mu})^2 \quad \Rightarrow \quad \hat{\theta} = \sigma^2 \in \mathbb{R}$
- **Stimatore a minimi quadrati** $y(i) = \theta_0 + \theta_1 \varphi_1(i) + \cdots + \theta_{d-1} \varphi_{d-1}(i) + \epsilon(i)$
di un modello lineare:
 $\epsilon(i)$ indipendenti media nulla e varianza λ^2
 $\Rightarrow \hat{\theta} = [\hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_{d-1}] \in \mathbb{R}^{d \times 1}$



Stima a massima verosimiglianza

Gli stimatori presentati sono **parametrici**, nel senso che stimano dei parametri del sistema che ha generato i dati

- Nel fare ciò, non abbiamo **mai fatto assunzioni** sulla **distribuzione di probabilità** dei dati $\mathcal{D} = \{y(1), y(2), \dots, y(N)\}$

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



Stima a massima verosimiglianza

Supponiamo di avere a disposizione N osservazioni $Y = [y(1), y(2), \dots, y(N)]^\top$, dove

$$y(i) \sim \mathcal{N}(\mu, \sigma^2) \text{ i.i.d.} \quad \Rightarrow \quad f_y(y(i)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y(i) - \mu}{\sigma}\right)^2\right]$$

Probability density function

La **pdf congiunta** dei dati è $f_Y(y(1), y(2), \dots, y(N)|\mu, \sigma^2) = f_Y(Y|\mu, \sigma^2) = \prod_{i=1}^N \mathcal{N}(y(i)|\mu, \sigma^2)$

La pdf congiunta $f_Y(Y|\mu, \sigma^2)$ indica la **probabilità che si realizzi il vettore di dati osservato**

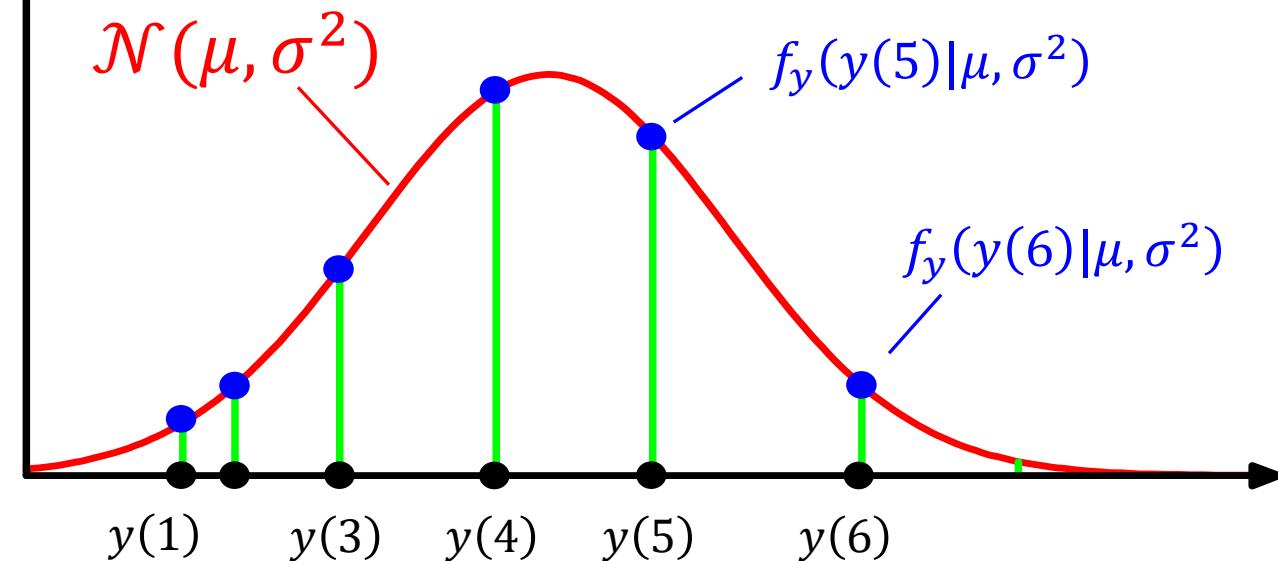
- Siccome le $y(i)$ sono i.i.d., la probabilità di osservare $y(1)$ AND $y(2)$ AND ... AND $y(N)$ è il **prodotto delle pdf** delle singole variabili



Esempio: calcolo pdf congiunta, parametri noti

Supponiamo di avere $N = 6$ dati $\mathcal{D} = \{y(1), y(2), \dots, y(6)\}$, $y(i) \sim \mathcal{N}(\mu, \sigma^2)$ i. i. d.

Il valore assunto dalla pdf congiunta $f_Y(Y|\mu, \sigma^2)$, con μ e σ^2 **noti**, valutata nei dati osservati \mathcal{D} , è il prodotto dei **pallini blu** ●



$$f_Y(Y|\mu, \sigma^2) = f_y(y(1)|\mu, \sigma^2) \cdot \\ \cdot f_y(y(2)|\mu, \sigma^2) \cdot \\ \vdots \\ \cdot f_y(y(6)|\mu, \sigma^2)$$



Stima a massima verosimiglianza

Se funzione dei dati Y , la pdf congiunta è una **distribuzione multivariabile**. Io però **conosco il valore di** Y , dato che ho osservato i dati

Se conoscessi anche μ e σ , potrei calcolare la probabilità di avere osservato Y . Però **non conosco** μ e σ ! E' proprio quello che voglio stimare!

Quando $f_Y(Y|\mu, \sigma^2)$ (la **pdf congiunta**) è vista in funzione dei parametri μ and σ , è chiamata funzione di **likelihood** $\mathcal{L}(\mu, \sigma^2 | Y)$

Cambia solo **l'interpretazione**, ma $f_Y(Y|\mu, \sigma^2)$ e $\mathcal{L}(\mu, \sigma^2 | Y)$ sono lo **stesso oggetto matematico**!



Stima a massima verosimiglianza

Riassunto:

Variabili non note **Parametri NOTI**

- Se $f_Y(Y | \mu, \sigma^2)$ è funzione dei dati Y : **pdf multivariabile**

Dati NOTI **Variabili non note**

- Se $f_Y(Y | \mu, \sigma^2)$ è funzione dei parametri μ e σ^2 : **likelihood** $\mathcal{L}(\mu, \sigma^2 | Y)$

Di solito si cambia la notazione di $f_Y(Y | \mu, \sigma^2)$ in $\mathcal{L}(\mu, \sigma^2 | Y)$ per rendere più chiaro chi è supposto noto («*a destra della barra |*») e chi non è noto («*a sinistra della barra |*»)



Stima a massima verosimiglianza

La stima a massima verosimiglianza è quel valore del parametro θ che **massimizza la verosimiglianza** $\mathcal{L}(\theta|Y)$

Esempio: supponiamo di avere **un solo dato**

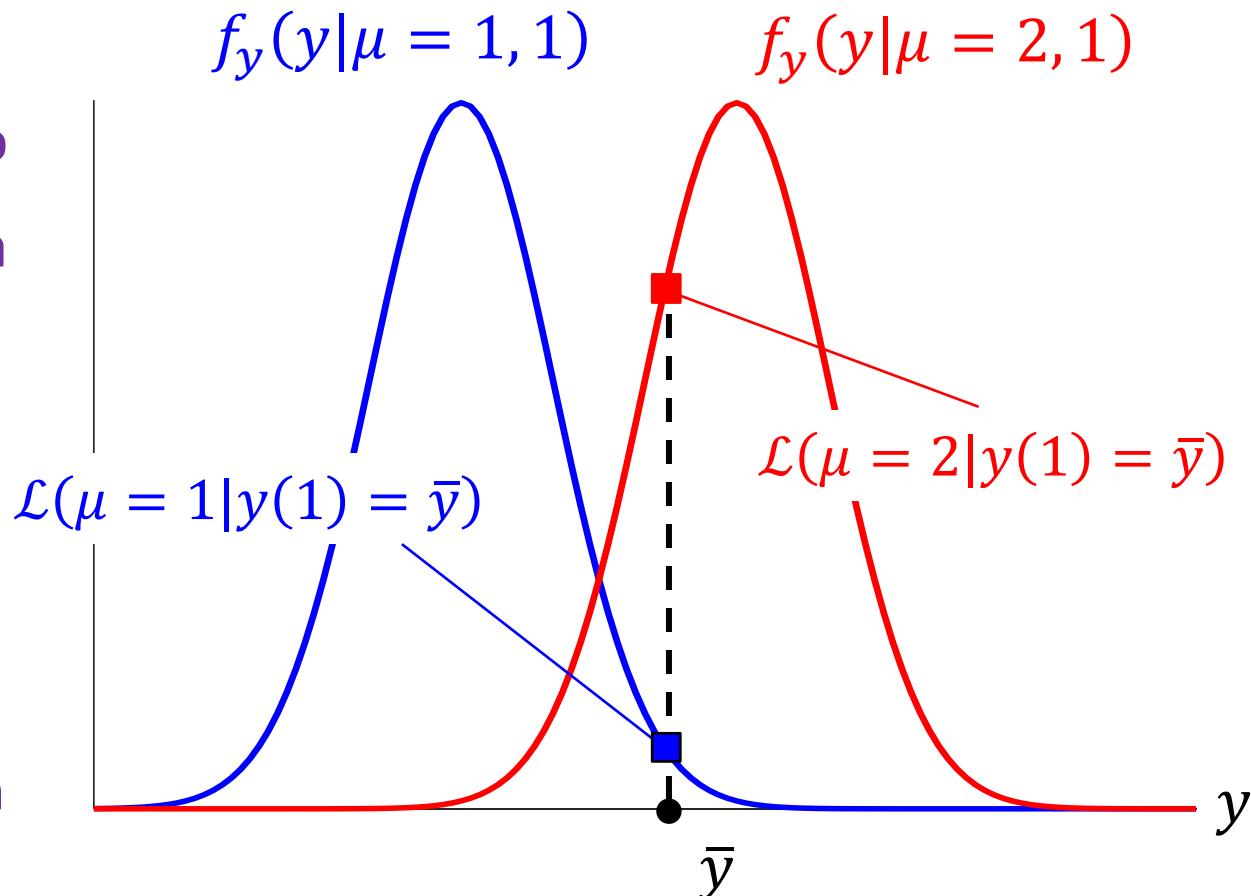
$y(1) \sim \mathcal{N}(\mu, \sigma^2 = 1)$, e che il suo valore sia

$y(1) = \bar{y}$. Il **parametro da stimare** è $\theta = \mu$

Notiamo che:

$$\mathcal{L}(\mu = 2|y(1) = \bar{y}) > \mathcal{L}(\mu = 1|y(1) = \bar{y})$$

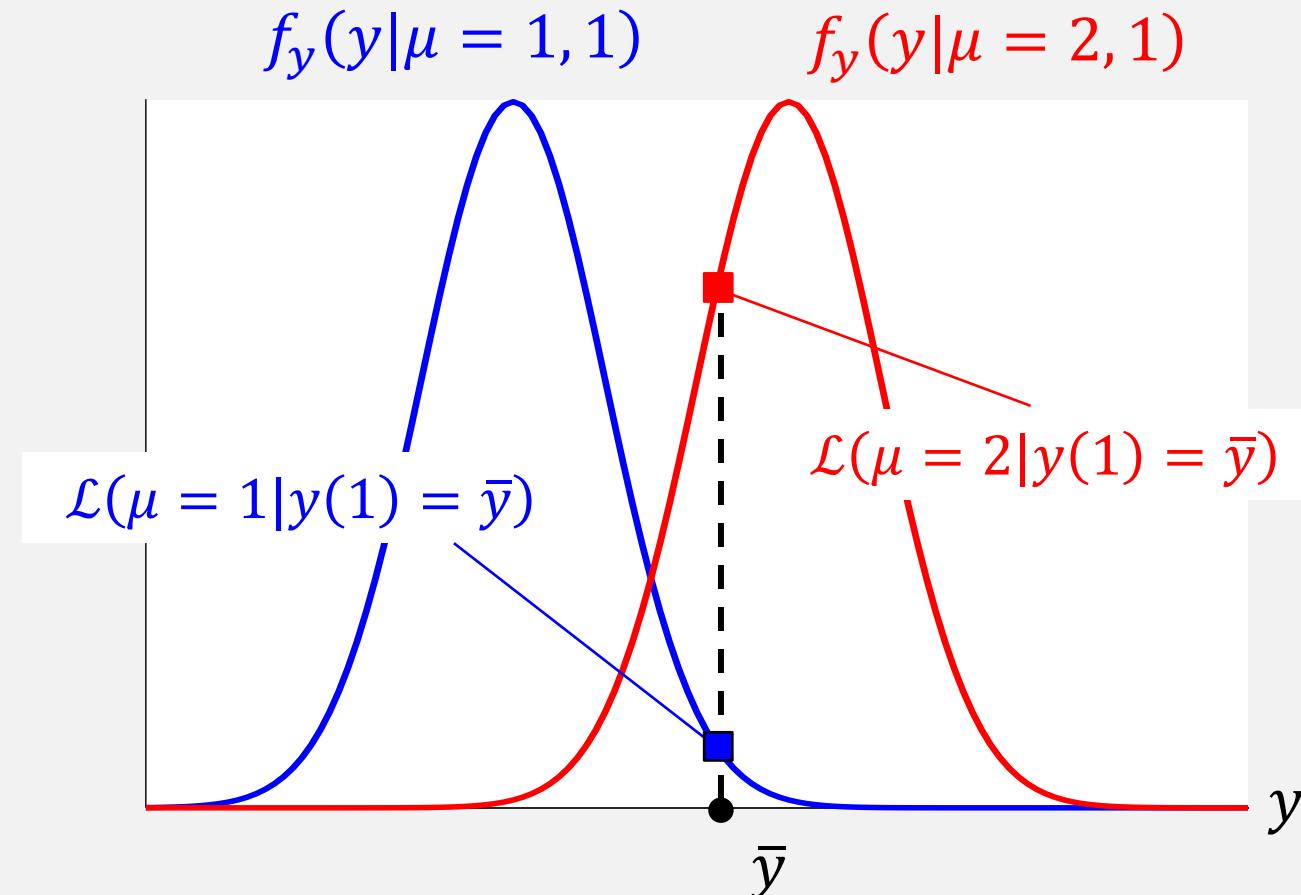
Per cui $\mu = 2$ è **più verosimile** di $\mu = 1$, in base a questo modello e questi dati



QUIZ!

In questo esempio, la **stima a massima verosimiglianza** è:

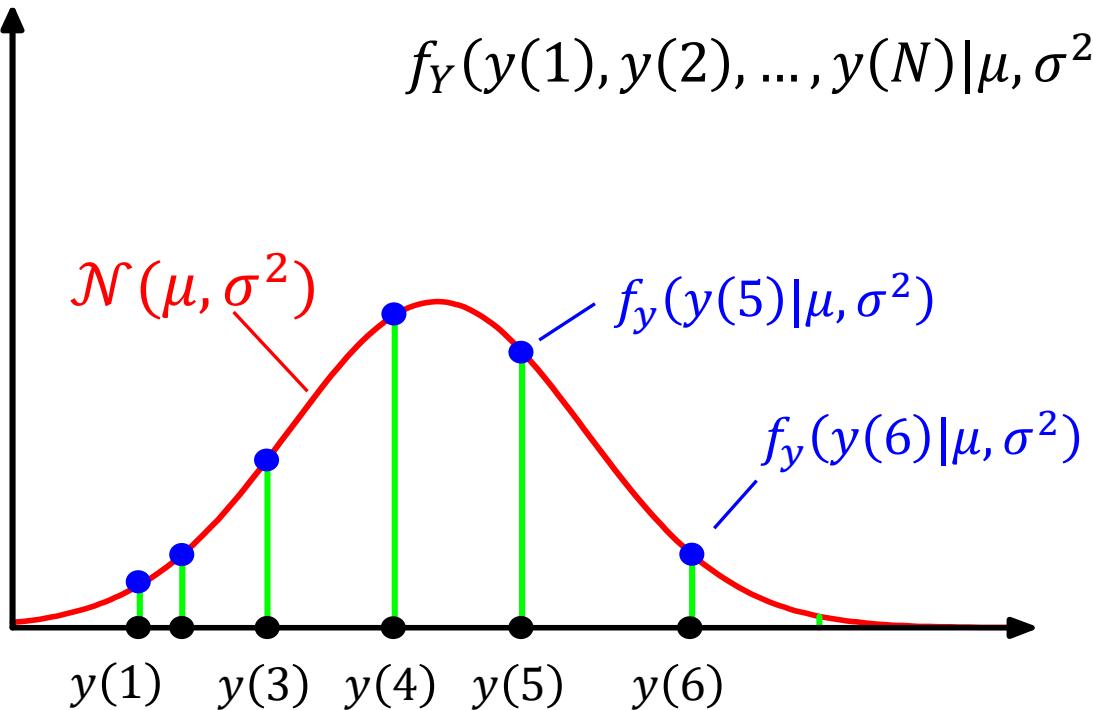
- $\hat{\mu} = 2\bar{y}$
- $\hat{\mu} = \bar{y}$
- $\hat{\mu} = 2$



Stima a massima verosimiglianza

L'esempio precedente considerava il caso in cui avevamo un solo dato osservato. Nel caso di **più osservazioni i.i.d.** di $y \sim \mathcal{N}(\mu, \sigma^2)$, ovvero $Y = [y(1), y(2), \dots, y(N)]^\top$, devo comunque **massimizzare la varosimiglianza**, cioè

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



Massimizzare la verosimiglianza significa «cambiare» i valori dei parametri μ e σ^2 tale che il **prodotto dei puntini blu** ● è massimizzato



Stima a massima verosimiglianza

La stima a massima verosimiglianza dell'esempio precedente può essere espressa come:

$$\hat{\theta}_{\text{ML}} = \begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix}_{2 \times 1} = \arg \max_{\theta} \mathcal{L}(\theta | Y) = \arg \max_{\theta} \prod_{i=1}^N \mathcal{N}(y(i) | \mu, \sigma^2)$$

In generale posso attribuire ai dati qualsiasi distribuzione di probabilità $d(\cdot)$, sia continua che discreta

$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta} \mathcal{L}(\theta | Y)_{d \times 1}$$



Stima a massima verosimiglianza

Spesso, anziché massimizzare $\mathcal{L}(\theta|Y)$, si massimizza il suo **logaritmo naturale**

- Dato che il logaritmo è una funzione monotona crescente, $\ln \mathcal{L}(\theta|Y)$ **ha lo stesso massimo** di $\mathcal{L}(\theta|Y)$
- 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)

$$\widehat{\boldsymbol{\theta}}_{\text{ML}} = \arg \max_{d \times 1} \ln \mathcal{L}(\boldsymbol{\theta}|Y)$$

A meno di casi particolari fortunati, l'ottimizzazione è effettuata con **metodi iterativi**



Stima a massima verosimiglianza: proprietà

Lo stimatore a massima verosimiglianza gode di **buone proprietà**. Infatti, esso è:

1. **Asintoticamente corretto:**

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

Lo stimatore può essere distorto. Per esempio lo stimatore a massima verosimiglianza della varianza di una popolazione Guassiana è distorto

2. **Consistente:** più N è grande, più la stima è precisa

3. **Asintoticamente efficiente:**

$$\lim_{N \rightarrow +\infty} \text{Var}[\hat{\theta}_{\text{ML}}] = M^{-1}$$

M : Matrice di informazione di Fisher

4. **Asintoticamente normale:**

$$\hat{\theta}_{\text{ML}} \sim \mathcal{N}(\theta^0, M^{-1}) \quad \text{per } N \rightarrow +\infty$$



Outline

1. Stima a massima verosimiglianza
- 2. Stima a massima verosimiglianza di parametri della popolazione**
3. Stima a massima verosimiglianza di modelli lineari



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

Stima ML di parametri della popolazione

Consideriamo il caso in cui vogliamo **stimare la media** μ di una popolazione di variabili casuali Gaussiane, supponendo di **conoscere la varianza** della distribuzione

Assumiamo di avere osservato **2 dati** $y(i) \sim \mathcal{N}(\mu, \sigma^2 = 1)$, $i = 1, 2$, i.i.d., tali che i valori osservati sono $y(1) = 4$, $y(2) = 6$

La forma della **pdf delle singole variabili** $y(i)$ è:

$$f_y(y(i)|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y(i) - \mu}{\sigma}\right)^2\right] = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(y(i) - \mu)^2\right]$$



Stima ML di parametri della popolazione

Il **valore assunto dalla pdf** in corrispondenza delle due osservazioni è:

$$y(1) = 4$$



$$f_y(y(1) = 4 | \mu, \sigma^2 = 1) = \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{1}{2}(4 - \mu)^2 \right]$$

$$y(2) = 6$$



$$f_y(y(2) = 6 | \mu, \sigma^2 = 1) = \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{1}{2}(6 - \mu)^2 \right]$$

La **pdf congiunta** è il prodotto delle due pdf singole (essendo i dati i.i.d.)

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



Stima ML di parametri della popolazione

La pdf congiunta è funzione solo di μ , poichè il **valore dei dati è noto**. Con questa interpretazione, la pdf congiunta è la **funzione di verosimiglianza** (likelihood function)

$$\mathcal{L}(\mu|y(1) = 4, y(2) = 6) = \left(\frac{1}{\sqrt{2\pi}} \exp \left[-\frac{1}{2}(4 - \mu)^2 \right] \right) \cdot \left(\frac{1}{\sqrt{2\pi}} \exp \left[-\frac{1}{2}(6 - \mu)^2 \right] \right)$$

La stima $\hat{\mu}_{ML}$ è valore di μ che **massimizza** la verosimiglianza

$$\hat{\mu}_{ML} = \arg \max_{\mu} \ln \mathcal{L}(\mu|y(1) = 4, y(2) = 6)$$



Stima ML di parametri della popolazione

È più conveniente massimizzare il logaritmo della verosimiglianza. Questa nuova funzione (la **log-verosimiglianza**) ha lo stesso massimo della verosimiglianza

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



Stima ML di parametri della popolazione

Massimizzando l'espressione ottenuta rispetto a μ otteniamo:

$$\frac{\partial \ln \mathcal{L}}{\partial \mu} = 0 \Rightarrow (4 - \mu) + (6 - \mu) = 0 \Rightarrow \hat{\mu}_{\text{ML}} = \frac{4 + 6}{2} = \boxed{5}$$

La **stima a massima verosimiglianza** del parametro μ per il modello Gaussiano è uguale allo stima ottenuta tramite lo **stimatore media campionaria!**

Questo risultato, seppur non generalizzabile, rende molto interpretabile ed intuitivo lo stimatore a massima verosimiglianza



Stima ML di parametri della popolazione

Osservazione: massimizzare la «log-verosimiglianza» equivale a minimizzare la «meno log-verosimiglianza»

$$\widehat{\boldsymbol{\theta}}_{\text{ML}} = \arg \max_{d \times 1} \ln[\mathcal{L}(\boldsymbol{\theta}|Y)] = \arg \min_{\boldsymbol{\theta}} -\ln[\mathcal{L}(\boldsymbol{\theta}|Y)]$$

Formulando il problema di stima a massima verosimiglianza in questo modo, abbiamo un problema di **minimizzazione** proprio come con lo stimatore a minimi quadrati!

$$\widehat{\boldsymbol{\theta}}_{\text{LS}} = \arg \min_{d \times 1} J(\boldsymbol{\theta})$$

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



Outline

1. Stima a massima verosimiglianza
2. Stima a massima verosimiglianza di parametri della popolazione
- 3. Stima a massima verosimiglianza di modelli lineari**



Stima ML di modelli lineari

Il metodo della massima verosimiglianza (ML) può essere usato anche per **stimare modelli lineari**. Quello che bisogna fare è imporre un **modello probabilistico** alle osservazioni $y(i)$

$$y(i) = \theta_0 + \theta_1 \varphi_1(i) + \cdots + \theta_{d-1} \varphi_{d-1}(i) + \epsilon(i) = \boldsymbol{\varphi}^\top(i) \boldsymbol{\theta} + \epsilon(i)$$

$1 \times d \quad d \times 1 \quad 1 \times 1$

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



$$y(i) \sim \mathcal{N}(\boldsymbol{\varphi}^\top(i) \boldsymbol{\theta}, \lambda^2) \text{ i.i.d.}$$

$1 \times d \quad d \times 1$

$$\boldsymbol{\varphi} = \begin{bmatrix} 1 \\ \varphi_1 \\ \vdots \\ \varphi_{d-1} \end{bmatrix}_{d \times 1} \quad \boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_{d-1} \end{bmatrix}_{d \times 1}$$

La media $\mu(i)$ di $y(i)$ è espressa come combinazione lineare dei regressori, $\mu(i) = \boldsymbol{\varphi}(i)^\top \boldsymbol{\theta}$!



Stima ML di modelli lineari

La **distribuzione congiunta** dei dati è:

$$\begin{aligned} f_Y(y(1), y(2), \dots, y(N) | X, \boldsymbol{\theta}, \lambda^2) &= \prod_{i=1}^N f_y(y(i) | \boldsymbol{\varphi}(i), \boldsymbol{\theta}, \lambda^2) \\ &= \prod_{i=1}^N \mathcal{N}(y(i) | \boldsymbol{\varphi}(i), \boldsymbol{\theta}, \lambda^2) \\ &= \prod_{i=1}^N \frac{1}{\sqrt{2\pi\lambda^2}} \exp \left[-\frac{1}{2} \left(\frac{y(i) - \boldsymbol{\varphi}^\top(i)\boldsymbol{\theta}}{\lambda} \right)^2 \right] \\ &= \mathcal{L}(\boldsymbol{\theta} | Y, X, \lambda^2) \quad \text{Supponiamo } \lambda^2 \text{ noto} \end{aligned}$$



Stima ML di modelli lineari

Calcoliamo la **log-verosimiglianza**

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



Stima ML di modelli lineari

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

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

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



Stima ML di modelli lineari

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

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

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

$$\widehat{\boldsymbol{\theta}}_{\text{ML}} = \arg \min_{\boldsymbol{\theta}} \frac{1}{2\lambda^2} \sum_{i=1}^N (y(i) - \boldsymbol{\varphi}^\top(i)\boldsymbol{\theta})^2$$

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



Stima ML di modelli lineari

Osservazione: cambiando ipotesi sulla distribuzione del rumore (e quindi dei dati), si ottengono **altre funzioni di costo** e **altri modelli**, che modellano i dati in modo diverso rispetto alla regressione lineare

Uno di questi altri modelli (che vedremo nella prossima lezione) è il modello di **regressione logistica**





**UNIVERSITÀ
DEGLI STUDI
DI BERGAMO**

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



IDENTIFICAZIONE DEI MODELLI E ANALISI DEI DATI (IMAD)

Lezione 5: Regressione logistica

Corso di Laurea Magistrale in
INGEGNERIA INFORMATICA

SPEAKER
Prof. Mirko Mazzoleni

PLACE
Università degli Studi di
Bergamo

Syllabus

Parte I: sistemi statici

1. Richiami di statistica

2.1 Proprietà degli stimatori

3. Stima a minimi quadrati

3.1 Stima di modelli lineari

3.2 Algoritmo del gradient descent

4. Stima a massima verosimiglianza

4.1 Proprietà della stima

4.2 Stima di modelli lineari

5. Regressione logistica

5.1 Stima di un modello di regressione logistica

6. Fondamenti di machine learning

6.1 Bias-Variance tradeoff

6.2 Overfitting

6.3 Regolarizzazione

6.4 Validazione

7. Cenni di stima Bayesiana

7.1 Probabilità congiunte, marginali e condizionate

7.2 Connessione con Filtro di Kalman



Parte I: sistemi statici**Parte II: sistemi dinamici**Stima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

✓ Stima parametri popolazione

✓ Stima modello lineare: minimi quadrati

○ **SI assunzioni su ddp dei dati**

✓ Stima massima verosimiglianza parametri popolazione

✓ Stima modello lineare: massiva verosimiglianza

✓ Regressione logistica

• θ variabile casuale○ **SI assunzioni su ddp dei dati**

✓ Stima Bayesiana

Stima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

✓ Modelli lineari di pss

✓ Predizione

✓ Identificazione

✓ Persistente eccitazione

✓ Analisi asintotica metodi PEM

✓ Analisi incertezza stima (numero dati finito)

✓ Valutazione del modello

Machine learning

Outline

1. Il problema della classificazione
2. Perché non usare la regressione lineare?
3. Regressione logistica: formulazione del problema
4. Regressione logistica: funzione di costo
5. Riassunto
6. Esercizi con codice



Outline

- 1. Il problema della classificazione**
2. Perché non usare la regressione lineare?
3. Regressione logistica: formulazione del problema
4. Regressione logistica: funzione di costo
5. Riassunto
6. Esercizi con codice



Il problema della classificazione

Il modello di **regressione lineare** discusso nella lezione precedente presuppone che la variabile di risposta sia **quantitativa (metrica)**

- in molte situazioni la variabile di risposta è invece **qualitativa (categorica)**

Le variabili qualitative assumono valori in un insieme non ordinato $\mathcal{C} = \{"\text{cat}_1", \dots, "cat_C"\}$,

come

- **eye color** $\in \{"\text{brown}", "\text{blue}", "\text{green}"\}$
- **email** $\in \{"\text{spam}", "\text{not spam}"\}$

Dati metrici

- Descrivono una quantità
- È definito un ordine
- È definita una distanza

Dati categorici

- Descrivono «categorie di appartenza»
- Non ha senso applicare un ordine
- Non ha senso calcolare le distanze



Il problema della classificazione

Il processo di stima di **output categorici**, utilizzando un insieme di regressori φ , è chiamato **classificazione**

Spesso però siamo più interessati a **stimare le probabilità** che φ appartenga a ciascuna categoria in \mathcal{C}

Se si vuol ottenere una classificazione, la **categoria più probabile** viene scelta come **classe** (categoria) per l'osservazione φ



Esempi di problemi di classificazione

- Una persona arriva al pronto soccorso con una **serie di sintomi** che potrebbero essere attribuiti a una delle **tre condizioni mediche**

Quale delle tre condizioni affligge il paziente?

- Un sistema bancario online gestisce delle **transazioni**, memorizzando l'indirizzo IP dell'utente, la cronologia delle transazioni passate e così via

La transazione è fraudolenta o no?

- Un biologo raccoglie dati su **sequenze di DNA** per un certo numero di pazienti **con e senza una determinata patologia**

Quali mutazioni genetiche causano una patologia e quali no?



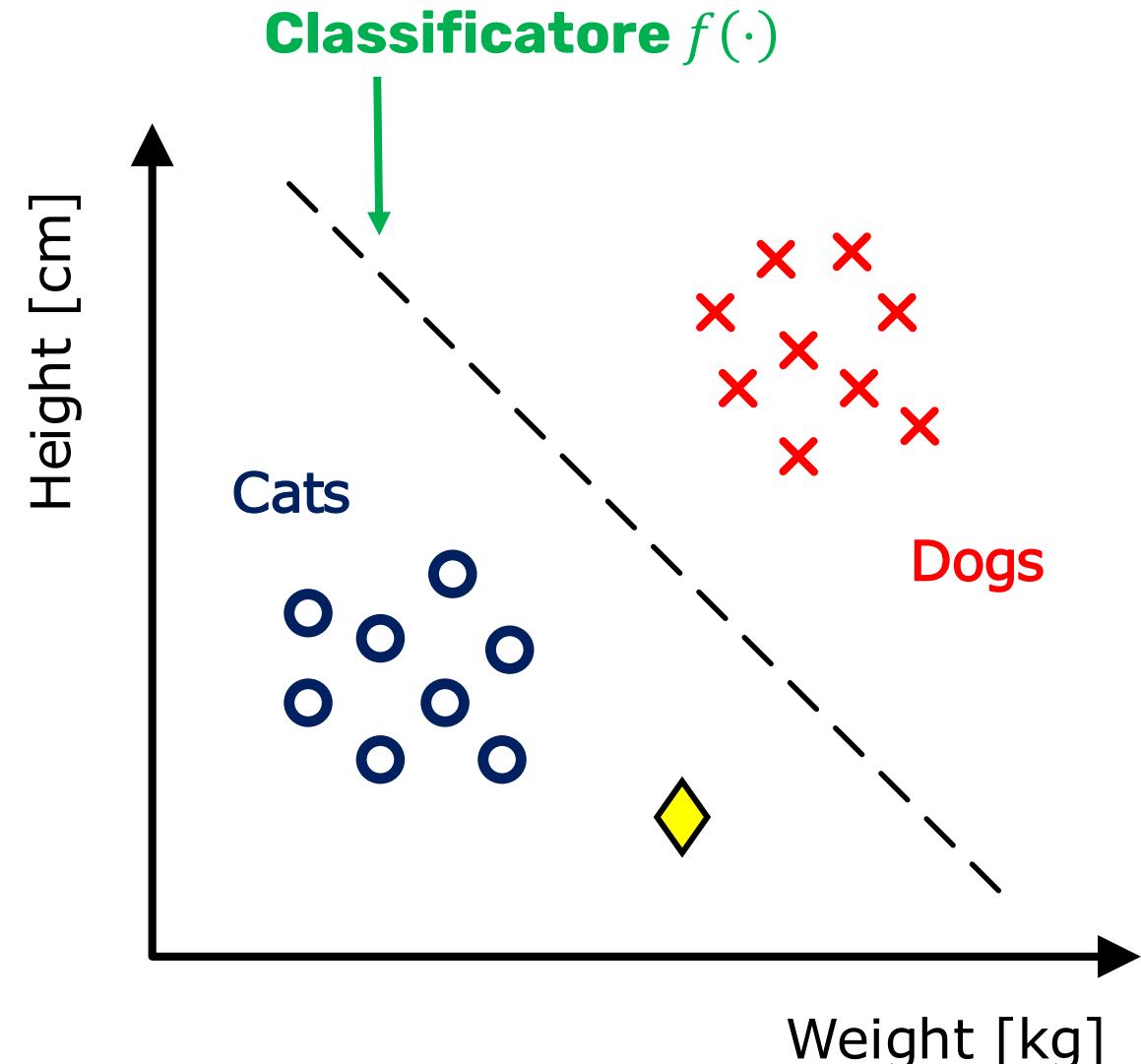
Esempio: cane vs. gatto

Supponiamo di misurare il **peso** e **l'altezza** di alcuni cani e gatti

Vogliamo imparare la funzione $f(\cdot)$ che ci dica se $\varphi = [\varphi_1, \varphi_2]^T$ è un cane o un gatto

- φ_1 : peso
- φ_2 : altezza

DOMANDA: Il punto  come è classificato dal modello? _____



QUIZ!

Consideriamo un'azienda che produce cancelli scorrevoli. I cancelli possono avere quattro pesi {300 kg, 400 kg, 500 kg, 600 kg}. Vogliamo rilevare il peso del cancello. Questo è un:

- Problema di regressione
- Problema di classificazione
- Sia un problema di classificazione che un problema di regressione



Outline

1. Il problema della classificazione
- 2. Perché non usare la regressione lineare?**
3. Regressione logistica: formulazione del problema
4. Regressione logistica: funzione di costo
5. Riassunto
6. Esercizi con codice



Perché non usare la regressione lineare?

Supponiamo di volere stimare la condizione di una paziente sulla base dei suoi sintomi. Ci sono tre possibilità: **stroke**, **drug overdose** and **epileptic seizure**

Potremmo considerare di codificare questi valori come una variabile **quantitativa**:

$$y = \begin{cases} 1 & \text{if } \textcolor{red}{\text{stroke}} \\ 2 & \text{if } \textcolor{red}{\text{drug overdose}} \\ 3 & \text{if } \textcolor{red}{\text{epileptic seizure}} \end{cases}$$

Tuttavia, stiamo implicitamente dicendo che la «differenza» tra **drug overdose** e **stroke** è la medesima che tra **epileptic seizure** e **drug overdose**, il che **non ha molto senso**



Perché non usare la regressione lineare?

Potremmo anche cambiare la codifica in:

$$y = \begin{cases} 1 & \text{if } \text{epileptic seizure} \\ 2 & \text{if } \text{stroke} \\ 3 & \text{if } \text{drug overdose} \end{cases}$$

Questo implicherebbe un **relazione totalmente differente** tra le tre condizioni

- ognuna di queste codifiche produrrebbe modelli lineari fondamentalmente diversi...
- ...che alla fine porterebbe a diverse stime per nuove osservazioni

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



Perché non usare la regressione lineare?

Con due livelli, la situazione è migliore. Ad esempio, forse ci sono solo due possibilità per le condizioni mediche del paziente: **stroke** e **drug overdose**

$$y = \begin{cases} 0 & \text{if } \text{stroke} \\ 1 & \text{if } \text{drug overdose} \end{cases}$$

Potremmo fare una regressione lineare e classificare come **drug overdose** se $\hat{y} > 0.5$ e **stroke** altrimenti, interpretando \hat{y} come una **probabilità di overdose**

Tuttavia, se usiamo la regressione lineare, alcune delle nostre stime potrebbero **essere al di fuori dell'intervallo [0, 1]**, il che non ha senso come probabilità. Non c'è nulla che "satura" l'uscita tra 0 e 1. → **Logistic function (Sigmoid)**



Outline

1. Il problema della classificazione
2. Perché non usare la regressione lineare?
- 3. Regressione logistica: formulazione del problema**
4. Regressione logistica: funzione di costo
5. Riassunto
6. Esercizi con codice



Regressione logistica: formulazione del problema

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

Definiamo la combinazione lineare:

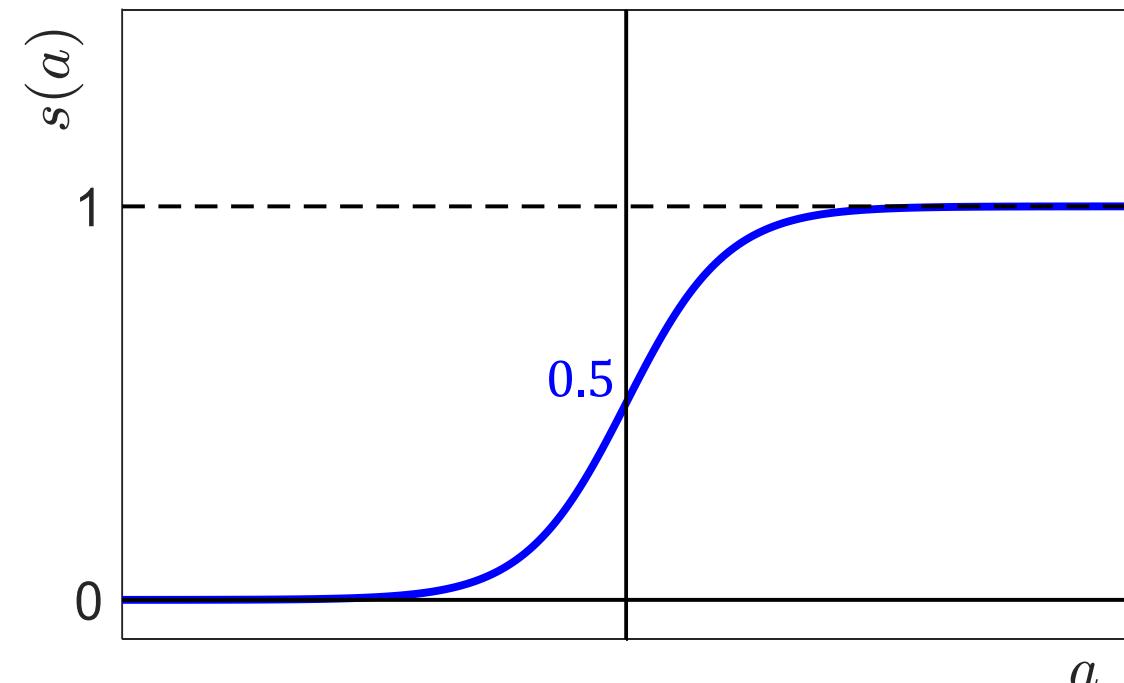
$$a = \sum_{j=0}^{d-1} \varphi_j \cdot \theta_j = \varphi^T \cdot \theta$$

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

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

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

Funzione logistica (Sigmoide)



Regressione logistica: formulazione del problema

In particolare, il modello di regressione logistica modella la probabilità che $y = 1$ **tramite un modello lineare**

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

L'output di $s(\boldsymbol{\varphi}^\top \boldsymbol{\theta})$ è **interpretato come una probabilità**

- $\boldsymbol{\varphi}^\top \boldsymbol{\theta} \gg 0 \Rightarrow s(\boldsymbol{\varphi}^\top \boldsymbol{\theta}) \gg 0.5 \Rightarrow P(y = 1|\boldsymbol{\varphi}) \approx 1 \rightarrow \boldsymbol{\varphi} \text{ è classificato nella classe «1»}$
- $\boldsymbol{\varphi}^\top \boldsymbol{\theta} \ll 0 \Rightarrow s(\boldsymbol{\varphi}^\top \boldsymbol{\theta}) \ll 0.5 \Rightarrow P(y = 1|\boldsymbol{\varphi}) \approx 0 \rightarrow \boldsymbol{\varphi} \text{ è classificato nella classe «0»}$



Regressione lineare e regressione logistica

La regressione lineare e la regressione logistica fanno parte di una categoria di modelli più generale detti **Generalized Linear Models (GLM)**

Regressione lineare

$$\mu = \boldsymbol{\varphi}^\top \boldsymbol{\theta} = \theta_0 + \theta_1 \varphi_1 + \cdots + \theta_{d-1} \varphi_{d-1}$$

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

Regressione logistica

$$\pi = s(\boldsymbol{\varphi}^\top \boldsymbol{\theta}) = s(\theta_0 + \theta_1 \varphi_1 + \cdots + \theta_{d-1} \varphi_{d-1})$$

Link function

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

Probabilità che $y = 1$

In questi modelli, un parametro di «tendenza centrale» di una distribuzione di probabilità è modellato tramite un modello lineare. I dati sono poi modellati come realizzazioni di questa distribuzione



Outline

1. Il problema della classificazione
2. Perché non usare la regressione lineare?
3. Regressione logistica: formulazione del problema
- 4. Regressione logistica: funzione di costo**
5. Riassunto
6. Esercizi con codice



Regressione logistica: funzione di costo

Supponiamo di avere a disposizione un dataset $\mathcal{D} = \{(\varphi(1), y(1)), \dots, (\varphi(N), y(N))\}$ dove $\varphi \in \mathbb{R}^{d \times 1}$ e $y(i) \in \{0, 1\}, i = 1, \dots, N$, i. i. d. → Notiamo che y è una **variabile categorica**

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

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

Procederemo nel modo seguente:

- Calcolo della meno-log-likelihood $J(\theta) = -\ln \mathcal{L}(\pi|\mathcal{D})$
- Calcolo del gradiente $\nabla_\theta J(\theta)$
- Ottimizzazione per trovare il minimo di $J(\theta)$



Regressione logistica: funzione di costo

Calcoliamo la verosimiglianza

$$\pi(i) \equiv P(y(i) = 1 | \varphi(i)) = \frac{1}{1 + e^{-\varphi^\top(i)\theta}}$$

$$\mathcal{L}(\pi|Y) = \prod_{i=1}^N \pi(i)^{y(i)} \cdot (1 - \pi(i))^{1-y(i)} \Rightarrow \text{Calcolo la meno-log-likelihood} \Rightarrow$$

$$-\ln[\mathcal{L}(\pi|Y)] = -\ln \left[\prod_{i=1}^N \pi(i)^{y(i)} \cdot (1 - \pi(i))^{1-y(i)} \right] = -\sum_{i=1}^N \ln [\pi(i)^{y(i)} \cdot (1 - \pi(i))^{1-y(i)}]$$

$$= -\sum_{i=1}^N \left(\ln[\pi(i)^{y(i)}] + \ln[(1 - \pi(i))^{1-y(i)}] \right)$$

$$= -\sum_{i=1}^N (y(i) \cdot \ln \pi(i) + (1 - y(i)) \cdot \ln[1 - \pi(i)])$$

$$\equiv J(\theta)$$



Interpretazione della funzione di costo

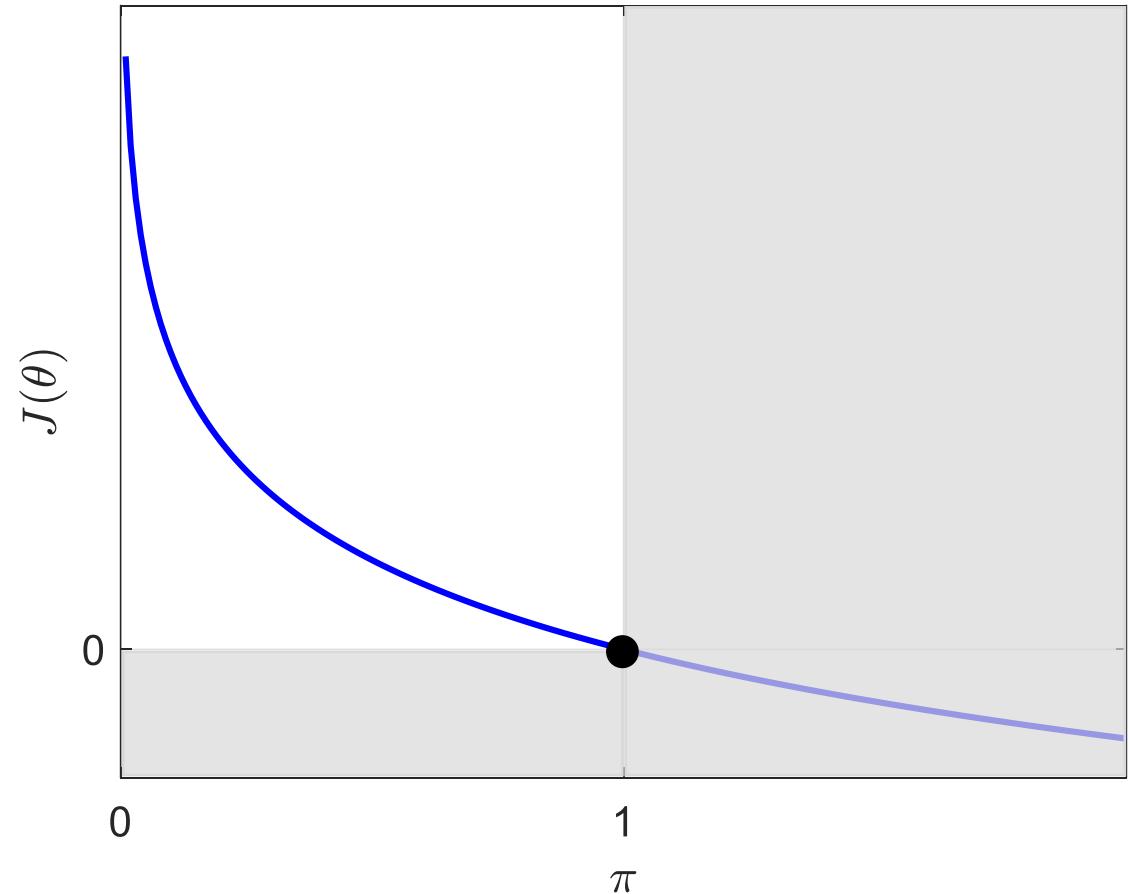
Assumiamo ci sia **un solo dato** $\mathcal{D} = \{(\varphi, y)\}$

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

Caso $y = 1$

$$J(\theta) = -\ln \pi$$

- $J(\theta) \approx 0$ se $y = 1$ e $\pi \approx 1$
- $J(\theta) \approx +\infty$ se $y = 1$ e $\pi \approx 0$



Interpretazione della funzione di costo

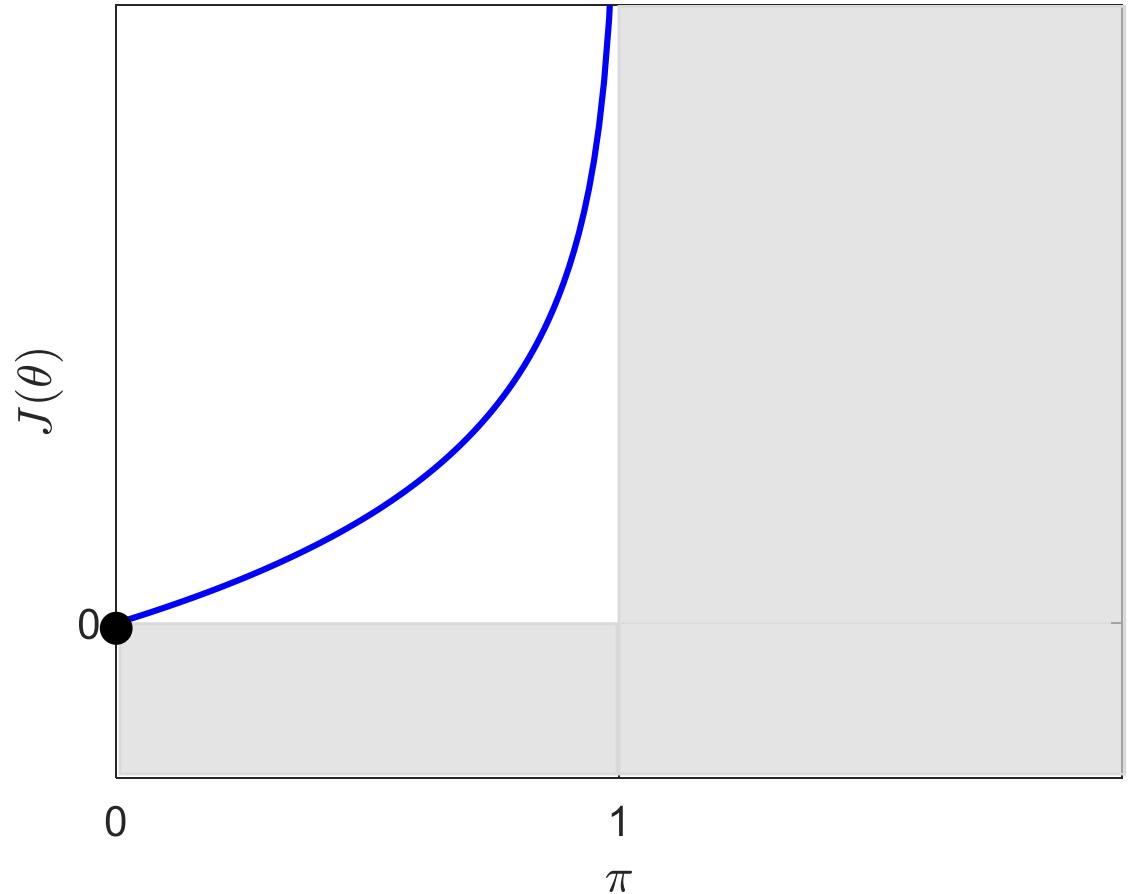
Assumiamo ci sia **un solo dato** $\mathcal{D} = \{(\varphi, y)\}$

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

Caso $y = 0$

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

- $J(\theta) \approx 0$ se $y = 0$ e $\pi \approx 0$
- $J(\theta) \approx +\infty$ se $y = 0$ e $\pi \approx 1$



QUIZ!

Nella funzione di costo della regressione logistica, dove sono i parametri θ che vogliamo stimare?

- Nei termini $y(i)$
- Nei termini \ln
- Nei termini $\pi(i)$

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



Calcolo del gradiente

Dobbiamo calcolare il gradiente di $J(\theta)$ rispetto a $\theta \in \mathbb{R}^{d \times 1}$. Per prima cosa, calcoliamo la

derivate di $s(a) = \frac{1}{1+e^{-a}}$ rispetto allo scalare $a \in \mathbb{R}$

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

Nel caso in cui $a = \varphi^\top \theta$, abbiamo

$$\nabla_{\theta} s(\varphi^\top \theta) = \varphi \cdot s(\varphi^\top \theta) \cdot [1 - s(\varphi^\top \theta)] = \varphi \cdot \pi \cdot [1 - \pi]$$

$d \times 1$ $d \times 1$ 1×1 1×1 $d \times 1$



Calcolo del gradiente

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

$$J(\theta) = - \sum_{i=1}^N \left(y(i) \ln \pi(i) + (1 - y(i)) \ln[1 - \pi(i)] \right)$$

$$\pi(i) = \frac{1}{1 + e^{-\varphi(i)^\top \theta}}$$

$$\nabla_{\theta} J(\theta) = - \sum_{i=1}^N \left(y(i) \frac{\pi'(i)}{\pi(i)} + (1 - y(i)) \frac{-\pi'(i)}{1 - \pi(i)} \right)$$

$$= - \sum_{i=1}^N \left(y(i) \frac{\varphi(i) \pi(i) [1 - \pi(i)]}{\pi(i)} + (1 - y(i)) \frac{-\varphi(i) \pi(i) [1 - \pi(i)]}{1 - \pi(i)} \right)$$



Calcolo del gradiente

$$= \sum_{i=1}^N (-y(i)\boldsymbol{\varphi}(i)[1 - \pi(i)] - (1 - y(i))(-\boldsymbol{\varphi}(i)\pi(i)))$$

$$= \sum_{i=1}^N (\boldsymbol{\varphi}(i) \cdot [-y(i) + y(i)\pi(i)] + \boldsymbol{\varphi}(i) \cdot [\pi(i) - y(i)\pi(i)])$$

$$= \sum_{i=1}^N (\boldsymbol{\varphi}(i) \cdot [-y(i) + y(i)\pi(i) - y(i)\pi(i) + \pi(i)])$$

Gradiente $\nabla_{\theta}J(\theta)$

$$= \sum_{i=1}^N \boldsymbol{\varphi}(i) \cdot (\pi(i) - y(i))$$



Outline

1. Il problema della classificazione
2. Perché non usare la regressione lineare?
3. Regressione logistica: formulazione del problema
4. Regressione logistica: funzione di costo
- 5. Riassunto**
6. Esercizi con codice



Riassunto

Il modello di regressione logistica, nonostante il suo nome, non viene utilizzato per la regressione, ma per la **classificazione**

Una volta che il modello stima la probabilità di una classe, possiamo classificare un dato 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 utilizzando il modello: $s(\boldsymbol{\varphi}^\top \boldsymbol{\theta}) = \frac{1}{1 + e^{-\boldsymbol{\varphi}^\top \boldsymbol{\theta}}}$

Il punto $\boldsymbol{\varphi}$ può quindi essere classificato alla classe $y = 1$ se $s(\boldsymbol{\varphi}^\top \boldsymbol{\theta}) \geq 0.5$



Riassunto

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

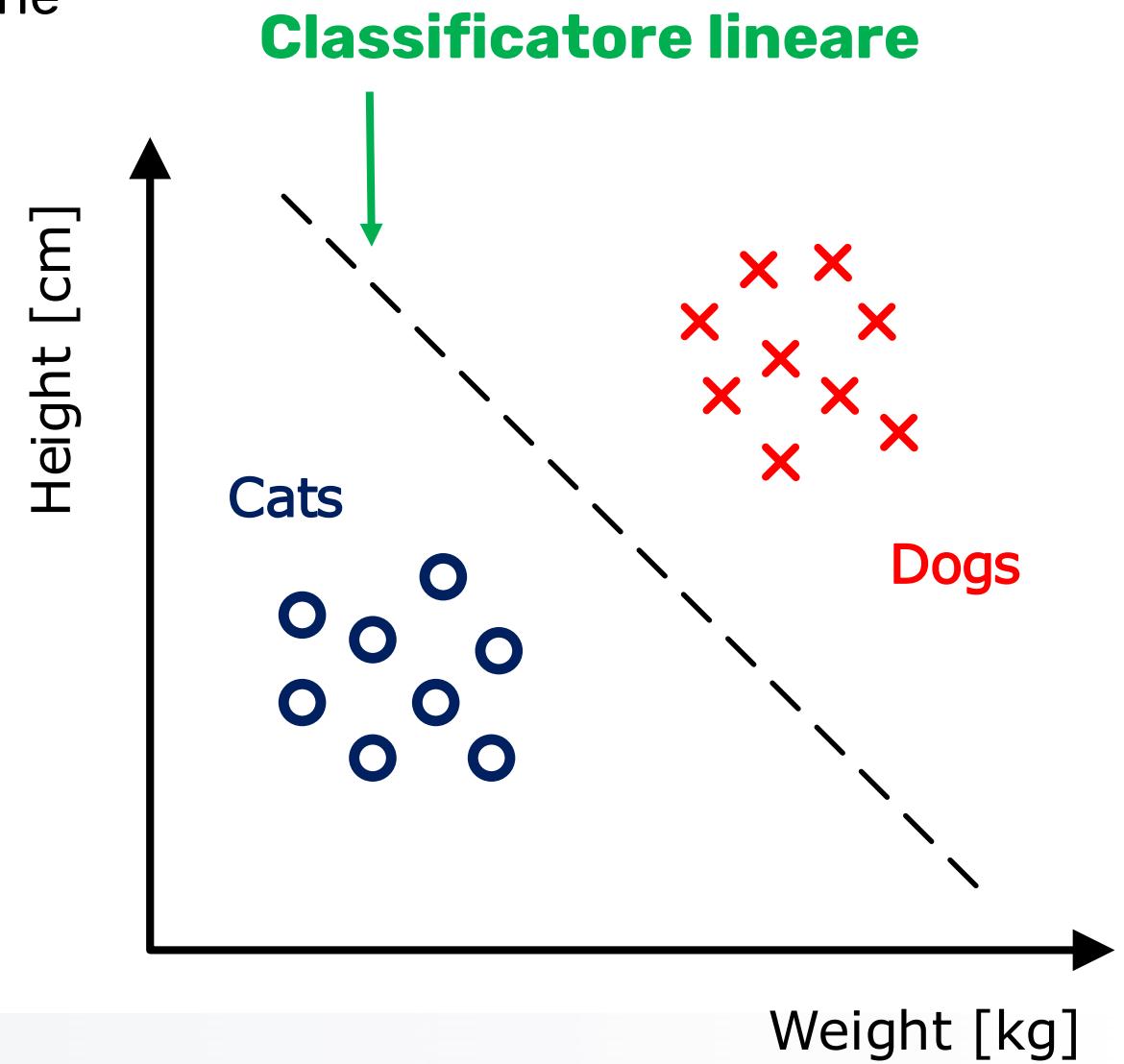
Infatti, classificare con la regola:

$$y = 1 \text{ if } s(\varphi^T \theta) \geq 0.5$$

è **equivalente** a dire

$$y = 1 \text{ if } \varphi^T \theta \geq 0$$

modello lineare



Outline

1. Il problema della classificazione
2. Perché non usare la regressione lineare?
3. Regressione logistica: formulazione del problema
4. Regressione logistica: funzione di costo
5. Riassunto
- 6. Esercizi con codice**



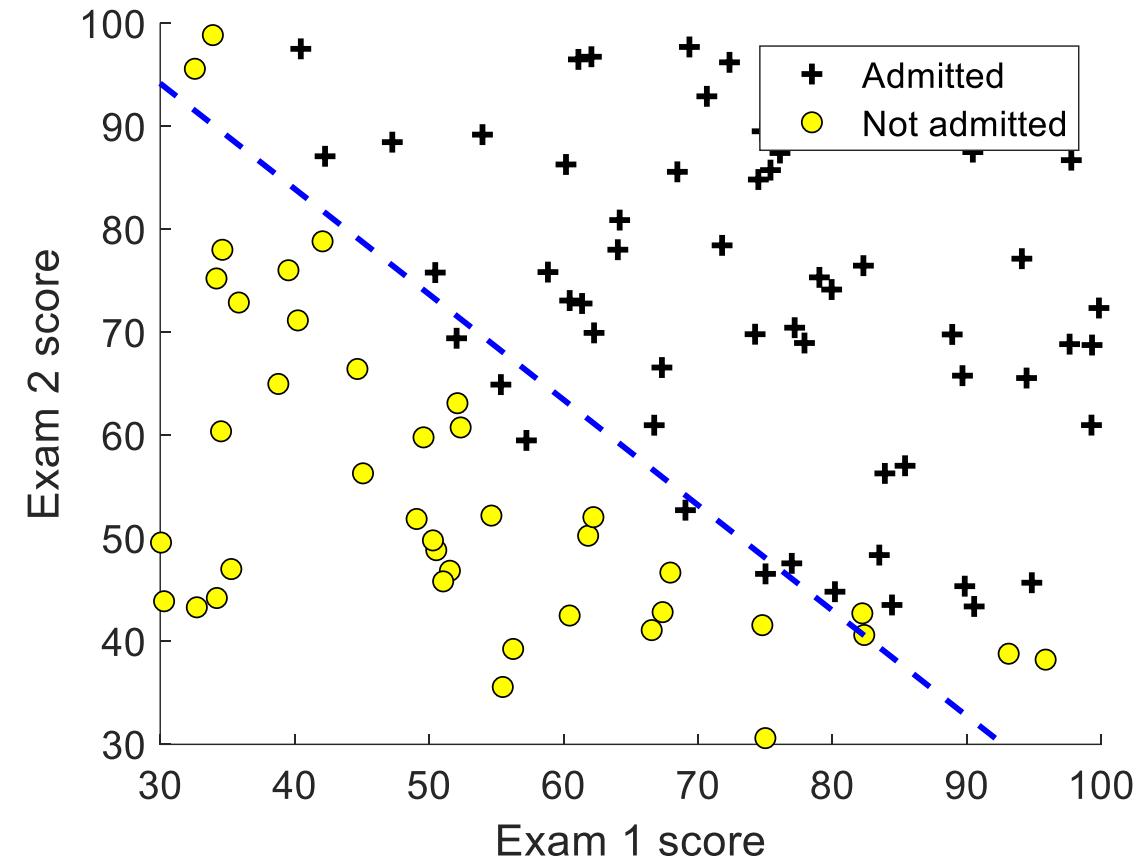
Esercizio: stima ammissione studenti

Vogliamo stimare la **probabilità di ammissione** $P(y = 1)$ di uno studente (o studentessa) all'università, visti i risultati di due esami (φ_1, φ_2), tramite una regressione logistica

Il dataset consiste di $N = 100$ studenti con $\varphi_1(i), \varphi_2(i)$ e $y(i) \in \{0,1\}$, per $i = 1, \dots, N$

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

- Matrice dei dati $X \in \mathbb{R}^{100 \times 3}$
- Vettore delle label $Y \in \mathbb{R}^{100 \times 1}$
- Vettore dei parametri $\boldsymbol{\theta} \in \mathbb{R}^{3 \times 1}$





**UNIVERSITÀ
DEGLI STUDI
DI BERGAMO**

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



IDENTIFICAZIONE DEI MODELLI E ANALISI DEI DATI (IMAD)

Lezione 6: Fondamenti di machine learning

Corso di Laurea Magistrale in
INGEGNERIA INFORMATICA

SPEAKER
Prof. Mirko Mazzoleni

PLACE
Università degli Studi di
Bergamo

Syllabus

Parte I: sistemi statici

1. Richiami di statistica

2.1 Proprietà degli stimatori

3. Stima a minimi quadrati

3.1 Stima di modelli lineari

3.2 Algoritmo del gradient descent

4. Stima a massima verosimiglianza

4.1 Proprietà della stima

4.2 Stima di modelli lineari

5. Regressione logistica

5.1 Stima di un modello di regressione logistica

6. Fondamenti di machine learning

6.1 Bias-Variance tradeoff

6.2 Overfitting

6.3 Regolarizzazione

6.4 Validazione

7. Cenni di stima Bayesiana

7.1 Probabilità congiunte, marginali e condizionate

7.2 Connessione con Filtro di Kalman



Parte I: sistemi staticiStima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

✓ Stima parametri popolazione

✓ Stima modello lineare: minimi quadrati

○ **SI assunzioni su ddp dei dati**

✓ Stima massima verosimiglianza parametri popolazione

✓ Stima modello lineare: massiva verosimiglianza

✓ Regressione logistica

• θ variabile casuale○ **SI assunzioni su ddp dei dati**

✓ Stima Bayesiana

Parte II: sistemi dinamiciStima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

✓ Modelli lineari di pss

✓ Predizione

✓ Identificazione

✓ Persistente eccitazione

✓ Analisi asintotica metodi PEM

✓ Analisi incertezza stima (numero dati finito)

✓ Valutazione del modello

Machine learning

Outline

1. Introduzione al machine learning e alla data science
2. Problemi supervisionati e non supervisionati
3. Feasibility of learning
4. Bias-variance tradeoff
5. Learning curves
6. Overfitting
7. Regolarizzazione
8. Validazione, cross-validation e formule di complessità ottima
9. Esercizi con codice



Outline

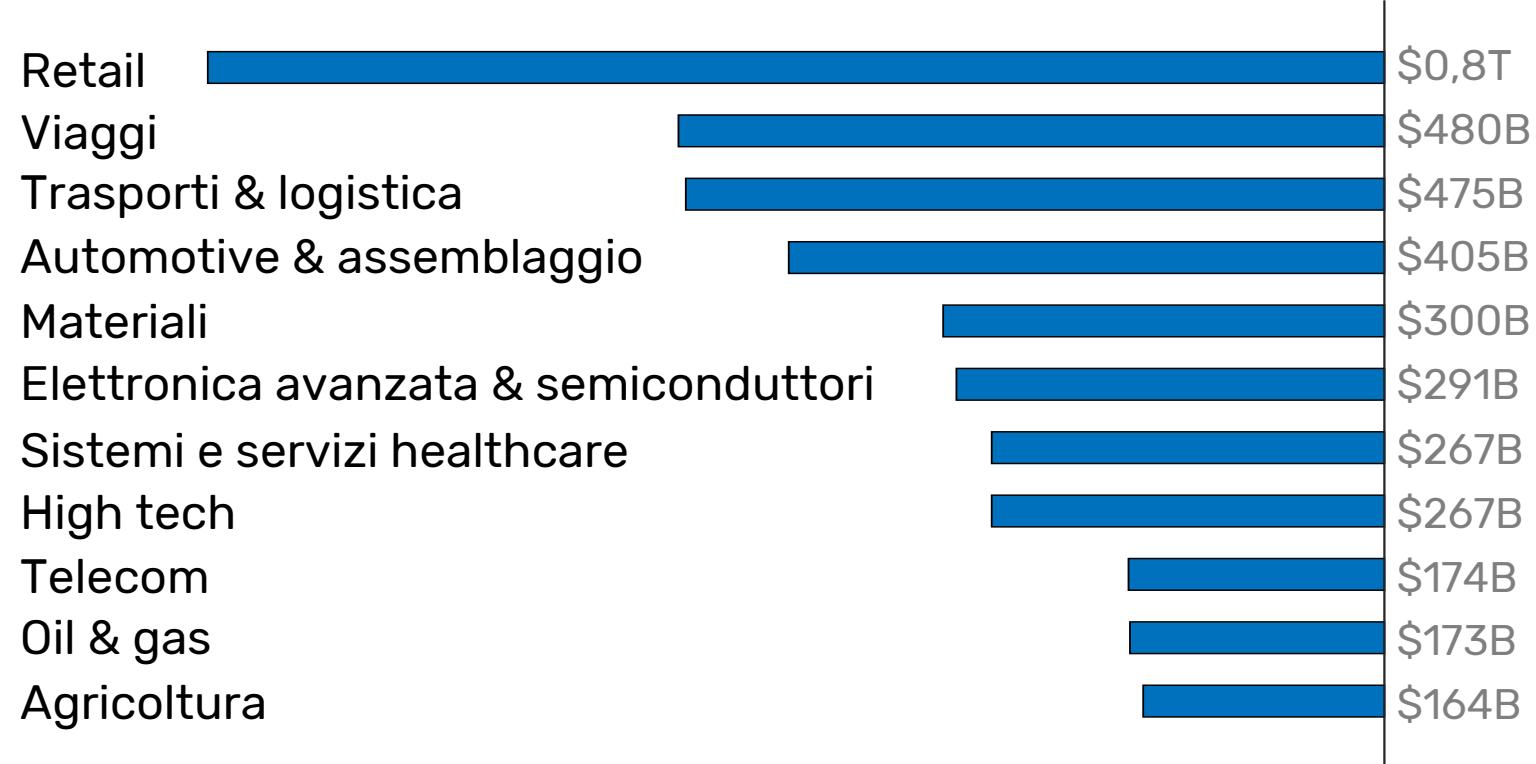
- 1. Introduzione al machine learning e alla data science**
2. Problemi supervisionati e non supervisionati
3. Feasibility of learning
4. Bias-variance tradeoff
5. Learning curves
6. Overfitting
7. Regolarizzazione
8. Validazione, cross-validation e formule di complessità ottima
9. Esercizi con codice



Introduzione al machine learning e alla data science

Valore creato dall'Artificial Intelligence (AI) entro il 2030

\$13
Trillions
1 trillion = 10^{12} dollari



- E' **difficile** trovare un settore industriale **che non beneficerà** dell'AI nel prossimo futuro



Perché proprio ora?

Una soluzione AI funziona quando abbiamo **dati** che riguardano il nostro business:

- E' possibile **collezionare dati** da svariati aspetti del business (operations, linea di produzione, supply-chain, parco clienti, campagne di marketing)
- L'informazione collezionata deve essere **analizzata** per ottenere **risultati azionabili**
- Una grande quantità di dati necessita di **specifiche infrastrutture** per essere gestita
- Una grande quantità di dati necessita di **potenza di calcolo** per essere analizzata
- Possiamo far sì che siano i computer a prendere decisioni, basandosi su **esempi**
- Nascita di **specifici lavori** e **specifici titoli di lavoro**



Esempi di learning

Ultimi 10 anni: straordinari passi avanti nelle applicazioni di **visione artificiale**



In che cosa consiste il learning

Il machine learning ha senso di essere applicato quando

1. Esiste un «pattern» nei dati
2. Non possiamo descriverlo matematicamente (non esiste una funzione analitica)

3. Abbiamo dati su di esso

I presupposti 1. e 2. non sono obbligatori:

- Se un pattern non esiste, non imparerò nulla
- Se posso descrivere un pattern matematicamente, presumibilmente non imparerò la migliore relazione
- Il vero vincolo è l'assunzione 3.



Machine learning vs. data science

Superficie della casa (feet ²)	# camere letto	# bagni	Rinnovata di recente	Prezzo (1000\$)
523	1	2	No	115
645	1	3	No	150
708	2	1	No	210
1034	3	3	Si	280

Output

Inputs

Machine learning

- Predire Output dato Input
- Software di AI operative (sito web \ app mobile)



Output: Codice e programma

Data science

- Le case con 3 bagni sono più costose di quelle con 2 bagni della medesima dimensione
- Le case rinnovate di recente costano 15% in più



Output: Presentazione



Organizzare i dati per imparare modelli dai dati

È importante specificare i **criteri di accettazione**: quanto il nostro modello è «buono»?

Esempio: si vuole progettare un sistema di visione per controllo qualità



NO DIFETTO



NO DIFETTO



DIFETTO

Obiettivo: individuare i difetti con
un'accuratezza del 95%



È necessario avere un dataset (spesso più di uno)
su cui valutare le performance del modello

Dati di validazione

Organizzare i dati per imparare modelli dai dati

Dati di Training\Identificazione



Dati di validazione



Output dell'algoritmo

NO DIFETTO

Valuto il modello



NO DIFETTO

NO DIFETTO

**66,7%
accuratezza**



Outline

1. Introduzione al machine learning e alla data science
- 2. Problemi supervisionati e non supervisionati**
3. Feasibility of learning
4. Bias-variance tradeoff
5. Learning curves
6. Overfitting
7. Regolarizzazione
8. Validazione, cross-validation e formule di complessità ottima
9. Esercizi con codice



I componenti del learning

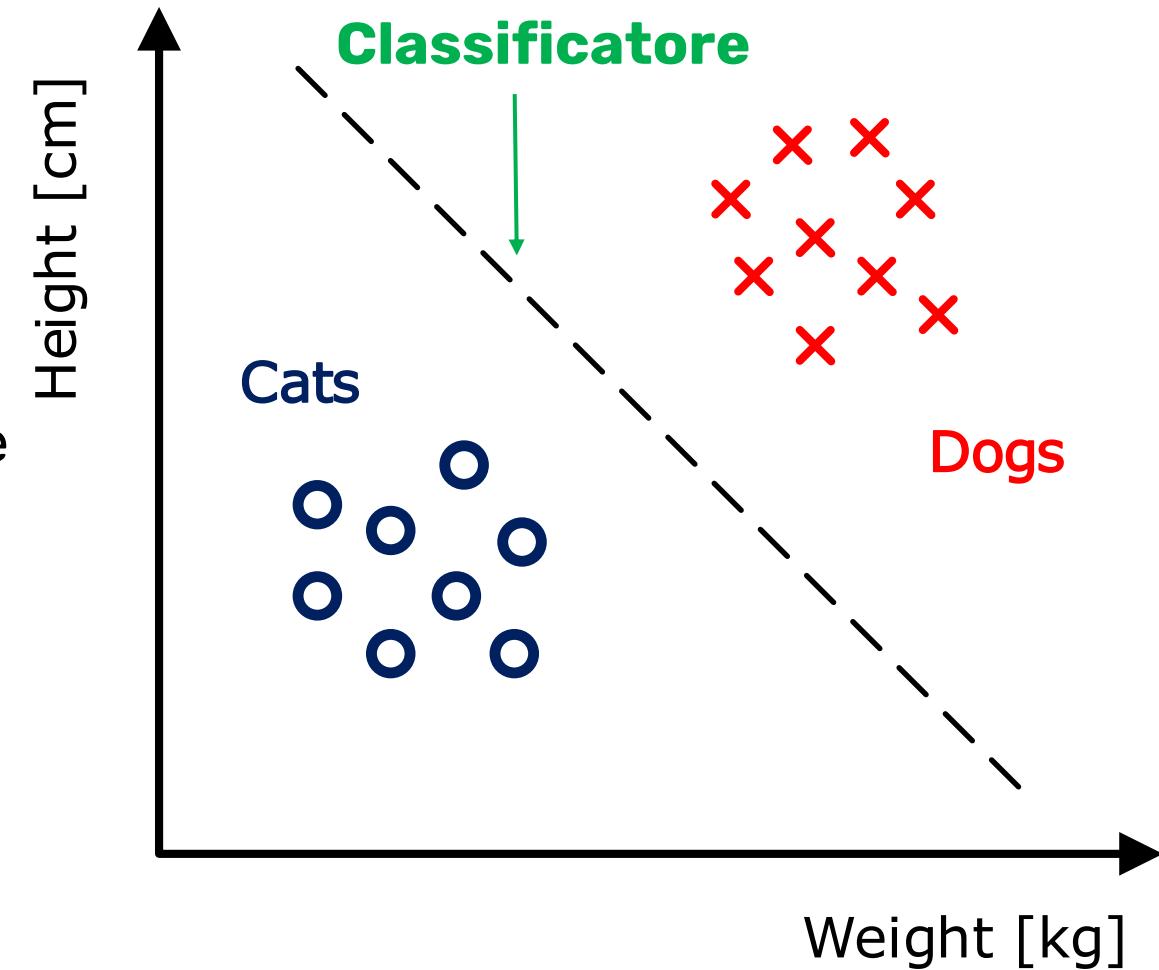
- Input: $\varphi \in \mathbb{R}^{d \times 1}$ (contenuto testuale della email) → ogni dimensione è un «attributo» della mail
 $d \times 1$
- Output: y (spam / non spam?) → la decisione che dobbiamo prendere
- Funzione target: $f: \mathbb{R}^{d \times 1} \rightarrow \mathcal{Y}$ (Formula ideale del filtro anti-spam) → ignota, si deve stimare
- Dati: $\mathcal{D} = \{(\varphi(1), y(1)), \dots, (\varphi(N), y(N))\}$ (record storico di emails)
 - ✓ Ogni **vettore delle features** (regressori) è costituito da diverse informazioni utilizzate per prevedere la variabile di output
- Ipotesi scelta: $g: \mathbb{R}^{d \times 1} \rightarrow \mathcal{Y}, g \in \mathcal{M}$ (formula che viene usata) → g è una **approssimazione** di f

\mathcal{M} è chiamato **spazio delle ipotesi** (o **set dei modelli**). Insieme all'**algoritmo di learning** forma il **modello di learning**



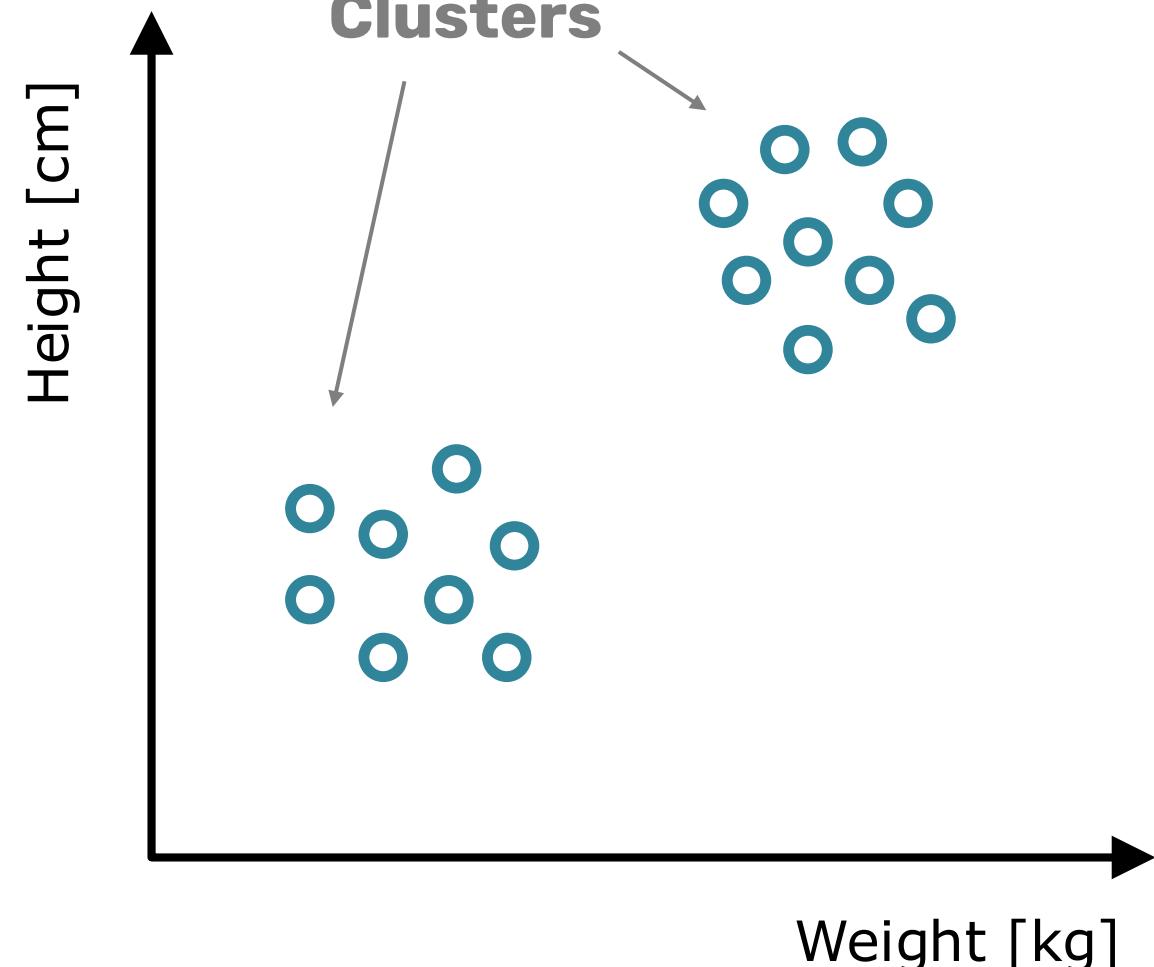
Apprendimento supervisionato (supervised learning)

- La «risposta corretta» (output label) y è nota
- Prevedere y da un set di inputs $\varphi \in \mathbb{R}^{d \times 1}$
- **Regressione:** prevedere una variabile continua $y \in \mathbb{R}$ (**valore reale**)
 1×1
- **Classification:** prevedere una variabile categorica $y \in \{1, 2, \dots, C\}$ (**classe\categoria**)



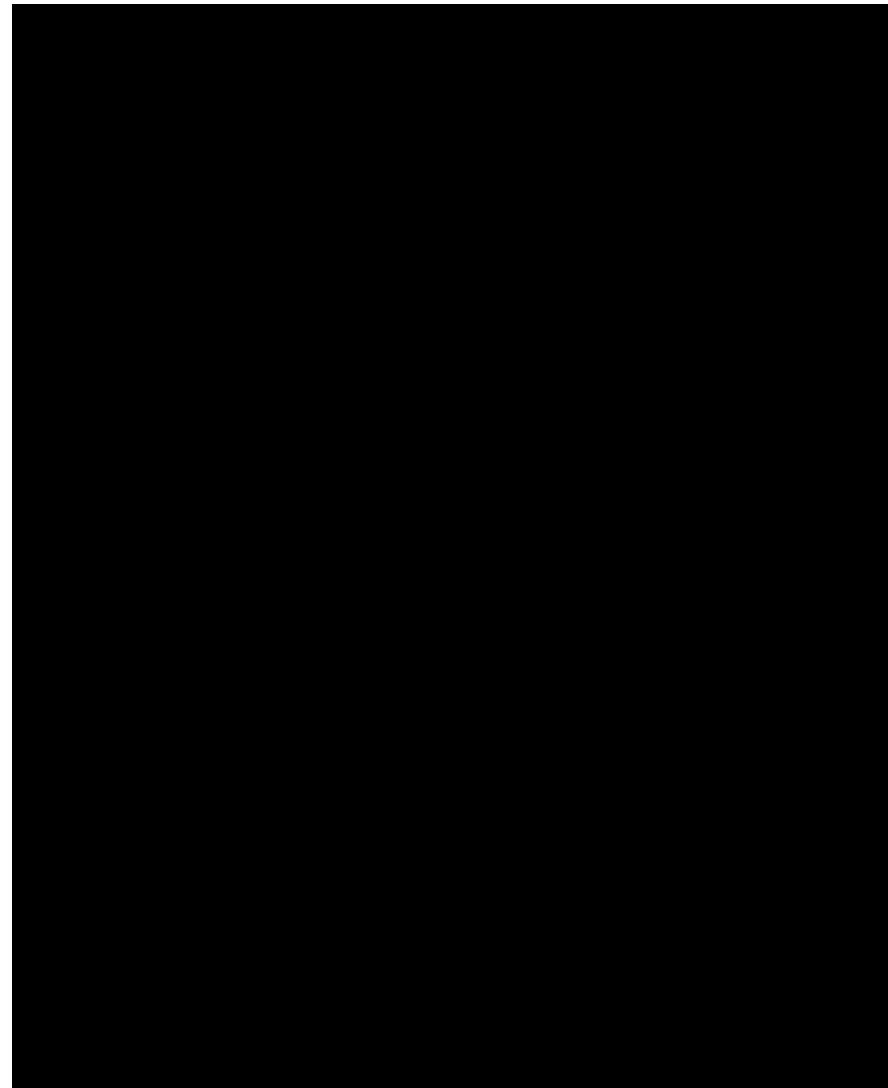
Apprendimento non supervisionato (unsupervised learning)

- Anzichè **(input, output)** abbiamo **(input, ?)**
- Non c'è una funzione f da apprendere
- Si vuol esplorare le proprietà di $\varphi \in \mathbb{R}^{d \times 1}$
 $d \times 1$
- Rappresentazione «ad alto livello» dell'input
- Gli elementi nello **stesso cluster** hanno proprietà simili



Apprendimento per rinforzo (reinforcement learning)

- Anzichè **(input, output)** abbiamo **(input, output, ricompensa)**
- L'algoritmo cerca di imparare quale azione intraprendere (policy), al fine di massimizzare la ricompensa
- Applicazioni in controllo, robotica, A\B testing (e.g. multi-armed bandits)



Esempi di problemi supervisionati e non

Supervisionati

- Filtro anti-spam Classificazione
- Approvazione crediti Classificazione
- Riconoscere oggetti in immagini Classificazione
- Prevedere i prezzi delle case Regressione
- Prevedere the stock market Regressione

Non supervisionati

- Segmentazione mercato Clustering
- Market basket analysis Co-occurrence grouping
- Modelli del linguaggio (word2vec) Similarity matching
- Analisi social networks Link prediction
Data reduction
- Low-order data representations

Supervisionato o non supervisionato

- Recommendation systems Similarity matching



Supervised learning: definizione del problema

Concentriamoci sul **problema dell'apprendimento supervisionato**. Abbiamo già visto due algoritmi di questo paradigma:

- **Regressione lineare:** $y(i) = f(\boldsymbol{\varphi}(i)) = \boldsymbol{\varphi}^\top(i)\boldsymbol{\theta}$ $\xrightarrow{1 \times d \quad d \times 1}$ **regressione**
- **Regressione logistica:** $y(i) = f(\boldsymbol{\varphi}(i)) = s(\boldsymbol{\varphi}^\top(i)\boldsymbol{\theta})$ $\xrightarrow{} \text{classificazione}$

In entrambi i casi, l'obiettivo era quello di stimare la funzione $f(\cdot)$ usando i dati \mathcal{D} :

- La funzione f viene cercata, dall'algoritmo di learning, nello spazio delle ipotesi \mathcal{M}
e.g. «lo spazio di tutte le funzioni lineari»
- 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



Supervised learning: definizione del problema

Cosa vuol dire che $h \approx f$?

- Dobbiamo definire una **misura di errore** o di **costo**. In precedenza, abbiamo già definito delle funzioni di costo $J(\theta)$ per la regressione lineare e logistica

Misure di errore puntuali

Le misure di errore puntuali $\ell(\varphi; \theta)$ sono basate su un singolo punto φ . Esempi sono:

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



Supervised learning: definizione del problema

Misure di errore globali

Queste misure considerano tutte le N osservazioni. È importante distinguere tra **errore in-sample** (errore di train) ed **errore out-of-sample** (errore di validazione o test)

Errore in-sample

Errore che il modello fa sugli N **dati osservati a disposizione**, che sono stati usati per stimarlo

$$E_{\text{in}}(h(\boldsymbol{\theta})) \equiv J(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^N \ell(f(\boldsymbol{\varphi}), h(\boldsymbol{\varphi}; \boldsymbol{\theta}))$$

Errore out-of-sample

Errore che il modello fa sull'intero dominio di f (quindi anche **dati che non ho osservato**)

$$E_{\text{out}}(h(\boldsymbol{\theta})) = \mathbb{E}_{\boldsymbol{\varphi}}[\ell(f(\boldsymbol{\varphi}), h(\boldsymbol{\varphi}; \boldsymbol{\theta}))]$$



Outline

1. Introduzione al machine learning e alla data science
2. Problemi supervisionati e non supervisionati
- 3. Feasibility of learning**
4. Bias-variance tradeoff
5. Learning curves
6. Overfitting
7. Regolarizzazione
8. Validazione, cross-validation e formule di complessità ottima
9. Esercizi con codice



QUIZ!

Domanda: Quale funzione ha generato i punti in figura?

Risposta: _____



QUIZ!

Domanda: Come vengono classificati i punti in figura?

● ○ ○ $f = ?$

● ○ ● ● ○ ● ● $f = 1$

○ ● ○ ○ ○ ○ ● ○ ○ ● $f = 0$

$f = 1$

$f = 0$

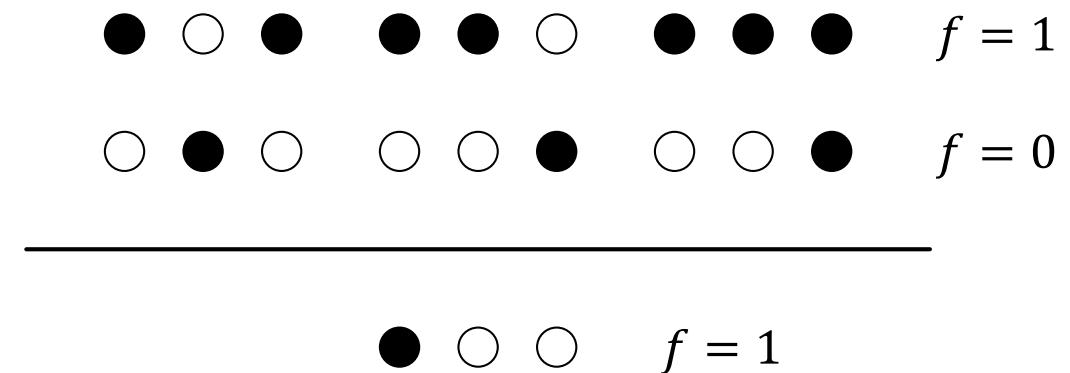
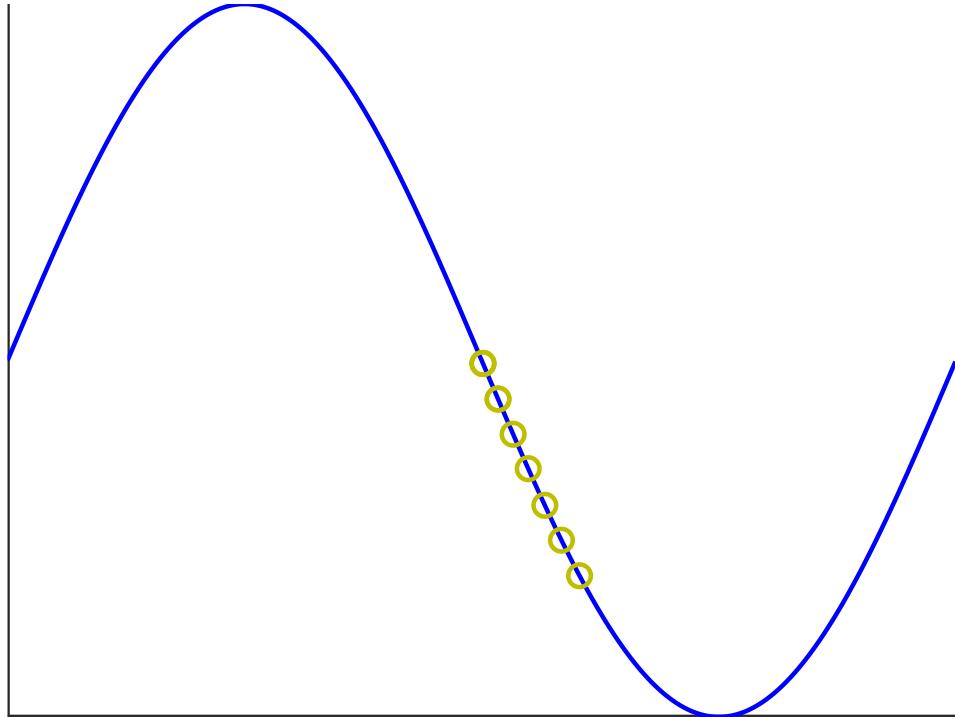


UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

Feasibility of learning

Non è possibile conoscere con certezza come sarà il comportamento della funzione f su punti che non ho osservato (problema dell'induzione di Hume)



Feasibility of learning

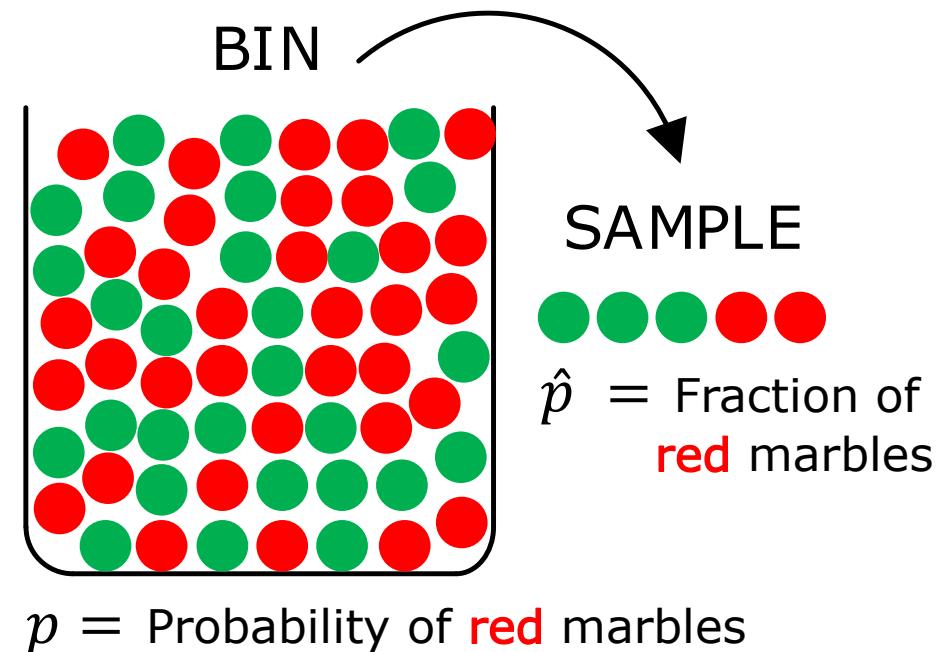
Focalizziamoci sul caso **supervised learning, classificazione binaria**

Problema: stimare una funzione ignota f

Risposta: Impossibile 😞. La funzione f può assumere qualsiasi valore al di fuori dei dati che abbiamo a disposizione

Consideriamo il seguente esempio

- Prendiamo un'urna con delle biglie **rosse** e **verdi**
- $\mathbb{P}[\text{prendere una biglia rossa}] = p$
- Il valore di p non è noto
- Estraiamo N biglie
- Frazione di biglie rosse nel campione estratto = \hat{p}



\hat{p} ci dice qualcosa riguardo a p ?

NO!

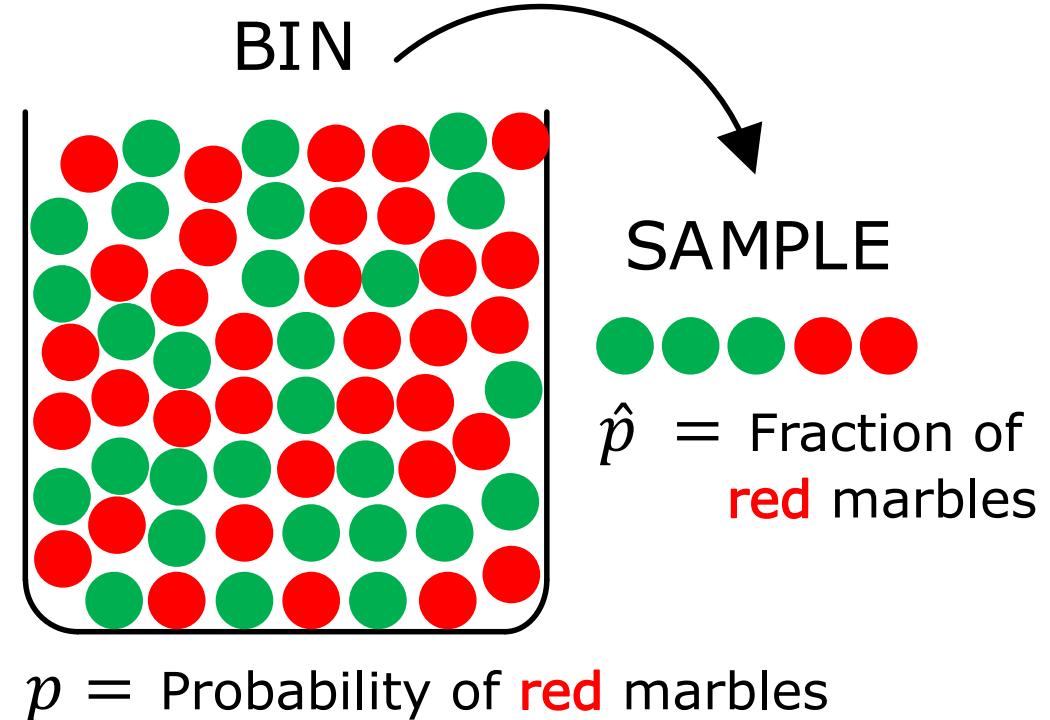
Il campione può essere per la maggior parte **verde** mentre il contenuto dell'urna potrebbe essere per la maggior parte **rosso**

POSSIBILE

SII!

Se estraggo tante biglie, il valore \hat{p} sarà «vicino» al valore di p

PROBABILE



Connessione con il learning di modelli dai dati

Urna: l'incognita è un **numero** p

Learning: l'incognita è una **funzione** $f: \mathbb{R}^{d \times 1} \rightarrow \mathcal{Y}$

Supponiamo che ogni biglia ● rappresenti un punto di input $\varphi \in \mathbb{R}^{d \times 1}$

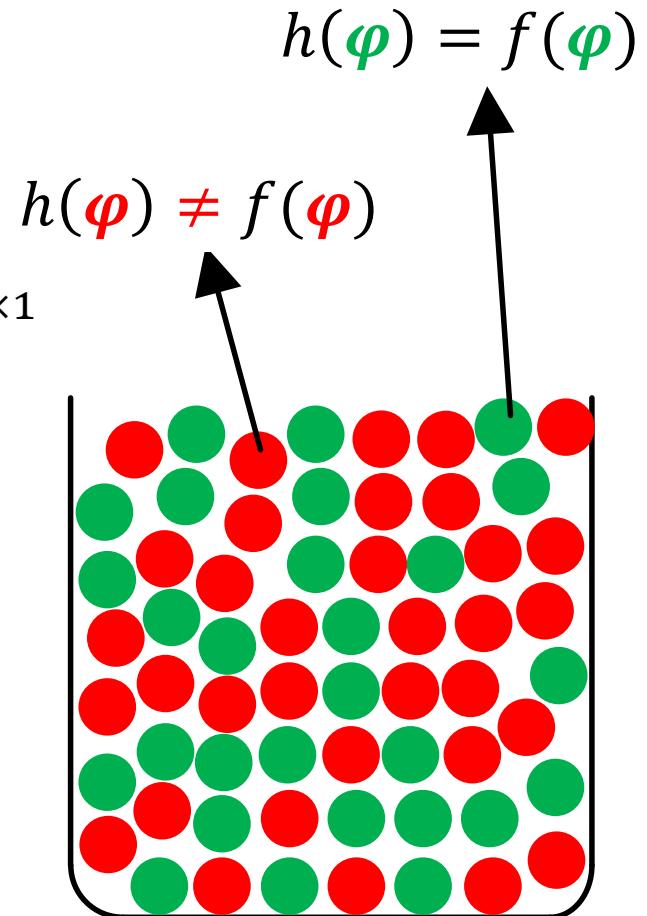
Per una **specifica ipotesi** $h \in \mathcal{M}$ ed un punto φ :

biglia verde ● \rightarrow **h classifica correttamente** $h(\varphi) = f(\varphi)$

biglia rossa ● \rightarrow **h sbaglia a classificare** $h(\varphi) \neq f(\varphi)$

Sia p che \hat{p} dipendono dalla particolare ipotesi h :

$\hat{p} \rightarrow$ errore sample in-sample $E_{\text{in}}(h)$
 $p \rightarrow$ errore out-of-sample $E_{\text{out}}(h)$



Connessione con il learning di modelli dai dati

Tramite la similitudine delle biglie e dell'urna abbiamo detto che:

$$\hat{p} \rightarrow \text{errore sample in-sample } E_{\text{in}}(h)$$
$$p \rightarrow \text{errore out-of-sample } E_{\text{out}}(h)$$

Nel caso delle biglie e dell'urna, ciò che ci interessava veramente stimare era p , non \hat{p}

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**



Connessione con il learning EFFETTIVO di modelli

In uno scenario di learning reale, la funzione h **non è fissata a-priori**

- *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'è un rischio maggiore di trovare una funzione g che «**fa bene**» sui dati osservati **solo per caso** → la funzione può spiegare benissimo i dati misurati ma fare malissimo su dati nuovi

Esiste quindi **tradeoff** tra **approssimazione** e **generalizzazione**. Si vuole:

- avere un buon modello sui dati **misurati** (training set)
- avere un buon modello su dati **non visti** (e quindi non usati per la stima del modello)

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



Approssimazione vs. Generalizzazione

L'obiettivo finale è avere un piccolo E_{out} : buona approssimazione di f out-of-sample

Spazio delle ipotesi \mathcal{M} PIÙ complesso



Migliori possibilità di **approssimare** f in-sample

Spazio delle ipotesi \mathcal{M} MENO complesso



Migliori possibilità di **generalizzare** f out-of-sample

Il caso ideale sarebbe avere uno spazio delle ipotesi \mathcal{M} che contiene solo la funzione f

$$\mathcal{M} = \{ f \}$$

Vincere un biglietto della
lotteria ☺



Approssimazione vs. Generalizzazione

L'esempio mostra:

- **Fit perfetto** sui dati di train

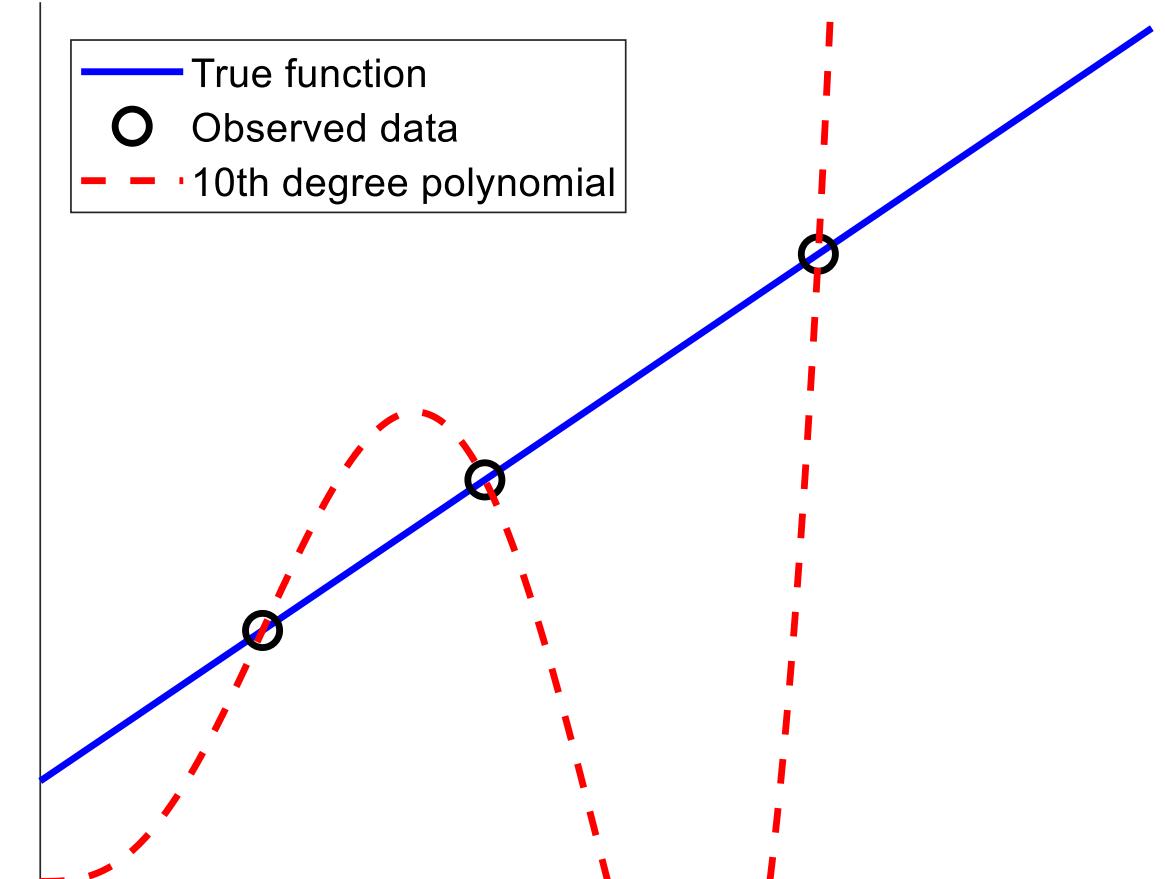


$$E_{\text{in}} = 0$$

- **Fit pessimo** on out of sample (test) data



$$E_{\text{out}} \text{ enorme}$$



Teoria della generalizzazione

Esiste una **teoria della generalizzazione** che studia i casi in cui è **probabile** generalizzare

- Il concetto da portare a casa è che **il learning è fattibile in modo probabilistico**
- Se siamo in grado di affrontare il tradeoff approssimazione-generalizzazione, possiamo **dire con alta probabilità che l'errore di generalizzazione è piccolo**

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

L'approccio **bias-varianza** decomponete E_{out} in:

1. Quanto bene \mathcal{M} può approssimare $f \rightarrow \text{Bias}$
2. Quanto bene riusciamo a scegliere una buona $h \in \mathcal{M}$, usando i dati $\rightarrow \text{Variance}$



Outline

1. Introduzione al machine learning e alla data science
2. Problemi supervisionati e non supervisionati
3. Feasibility of learning
- 4. Bias-variance tradeoff**
5. Learning curves
6. Overfitting
7. Regolarizzazione
8. Validazione, cross-validation e formule di complessità ottima
9. Esercizi con codice



Bias e varianza di un modello di learning

Supponiamo di osservare i **dati senza rumore**, cioè che $y = f(\boldsymbol{\varphi})$. L'errore out-of-sample può essere espresso come (rendendo esplicita la dipendenza di g da \mathcal{D})

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

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

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



Bias e varianza di un modello di learning

Concetriamoci su $\mathbb{E}_{\mathcal{D}} \left[\left(g^{(\mathcal{D})}(\boldsymbol{\varphi}) - f(\boldsymbol{\varphi}) \right)^2 \right]$. Definiamo **l'ipotesi «media»** $\bar{g}(\boldsymbol{\varphi}) = \mathbb{E}_{\mathcal{D}} [g^{(\mathcal{D})}(\boldsymbol{\varphi})]$

Questa «ipotesi media» può essere interpretata come l'ipotesi che deriva dall'usare K dataset $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_K$ e costruendola come $\bar{g}(\boldsymbol{\varphi}) \approx \frac{1}{K} \sum_{k=1}^K g^{(\mathcal{D}_k)}(\boldsymbol{\varphi})$

Abbiamo quindi:

questo è uno strumento concettuale e \bar{g} non ha bisogno di appartenere all'insieme delle ipotesi \mathcal{M}

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



Bias e varianza di un modello di learning

$$\mathbb{E}_{\mathcal{D}} \left[\left(g^{(\mathcal{D})}(\boldsymbol{\varphi}) - f(\boldsymbol{\varphi}) \right)^2 \right] = \underbrace{\mathbb{E}_{\mathcal{D}} \left[\left(g^{(\mathcal{D})}(\boldsymbol{\varphi}) - \bar{g}(\boldsymbol{\varphi}) \right)^2 \right]}_{\text{var}(\boldsymbol{\varphi})} + \underbrace{\left(\bar{g}(\boldsymbol{\varphi}) - f(\boldsymbol{\varphi}) \right)^2}_{\text{bias}^2(\boldsymbol{\varphi})}$$

Quindi;

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

Questo concetto è analogo al concetto di Mean Squared Error (MSE) per uno stimatore parametrico (Lezione 02)

$$= \text{bias}^2 + \text{var}$$



Bias e varianza di un modello di learning

Interpretazione

- Il termine di **bias**² $(\bar{g}(\varphi) - f(\varphi))^2$ misura quanto il nostro modello (cioè la nostra funzione stimata \bar{g}) è «lontano» dalla funzione target f

Infatti, \bar{g} ha il vantaggio di apprendere da un numero illimitato di datasets. Quindi, \bar{g} , nella capacità di approssimare f , è limitata solo dai «limiti» di \mathcal{M}

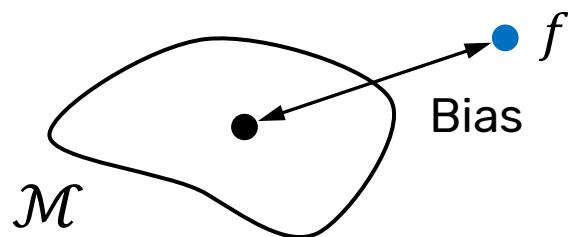
Esempio: sia \mathcal{M} l'insieme delle rette. Per quanti dati possa avere, una retta non potrà mai approssimare bene una curva...

- Il termine **varianza** $\mathbb{E}_{\mathcal{D}} \left[\left(g^{(\mathcal{D})}(\varphi) - \bar{g}(\varphi) \right)^2 \right]$ misura quanto $g^{(\mathcal{D})}$ si «disperde» da \bar{g} , e può essere pensata come quanto l'ipotesi finale $g^{(\mathcal{D})}$ differisca dall'ipotesi «migliore» (ovvero quella «media»)

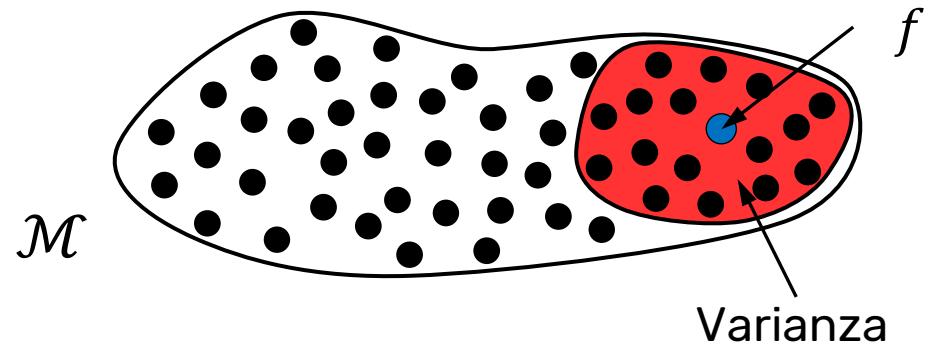


Bias e varianza di un modello di learning

$$\text{bias}^2 = (\bar{g}(\varphi) - f(\varphi))^2$$



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



Set di modelli (spazio delle ipotesi) **molto PICCOLO**. Poiché esiste una sola ipotesi, sia la funzione media \bar{g} che l'ipotesi finale $g^{(\mathcal{D})}$ saranno uguali, per qualsiasi set di dati. Quindi, **var** = 0. Il bias dipenderà esclusivamente da quanto bene questa singola ipotesi si avvicina al target f e, a meno che non siamo estremamente fortunati, **ci aspettiamo un grande bias**

Set di modelli (spazio delle ipotesi) **molto GRANDE**. La funzione targe sta in \mathcal{M} . Diversi set di dati porteranno a diverse ipotesi $g^{(\mathcal{D})}$ che concordano con set di dati a disposizione, e queste ipotesi sono sparse intorno alla regione rossa. Quindi, **bias** ≈ 0 poichè in media $g^{(\mathcal{D})}$ è vicina ad f . **La varianza è grande** (rappresentato euristicamente dalla dimensione della regione rossa)

Bias e varianza di un modello di learning

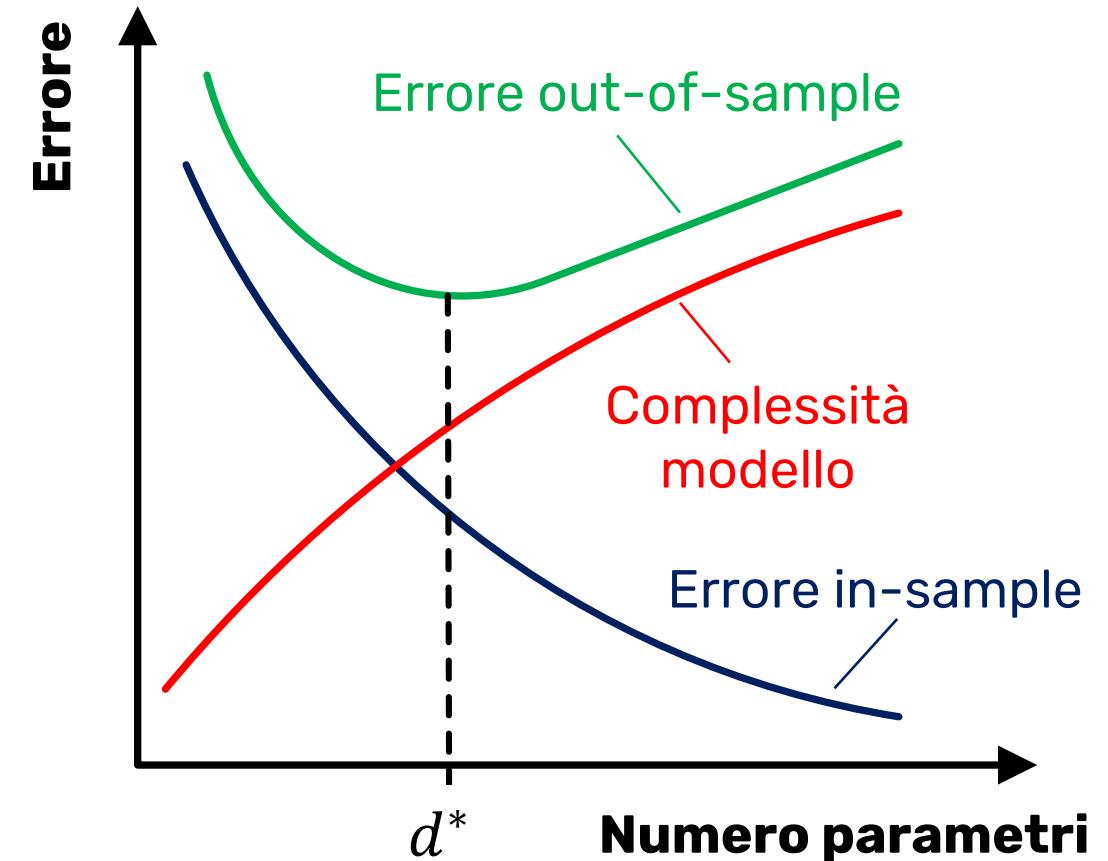
Regola euristica

Quanti punti N sono richiesti per assicurarsi un **buona probabilità di generalizzare?**

$$N \geq 10 \cdot \text{numero parametri modello}$$

Principio generale

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



Outline

1. Introduzione al machine learning e alla data science
2. Problemi supervisionati e non supervisionati
3. Feasibility of learning
4. Bias-variance tradeoff
- 5. Learning curves**
6. Overfitting
7. Regolarizzazione
8. Validazione, cross-validation e formule di complessità ottima
9. Esercizi con codice

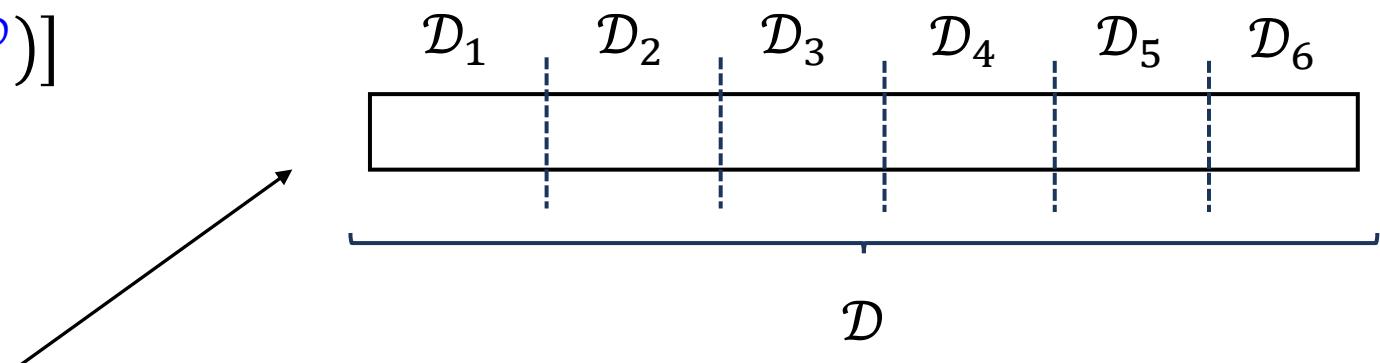


Learning curves

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:

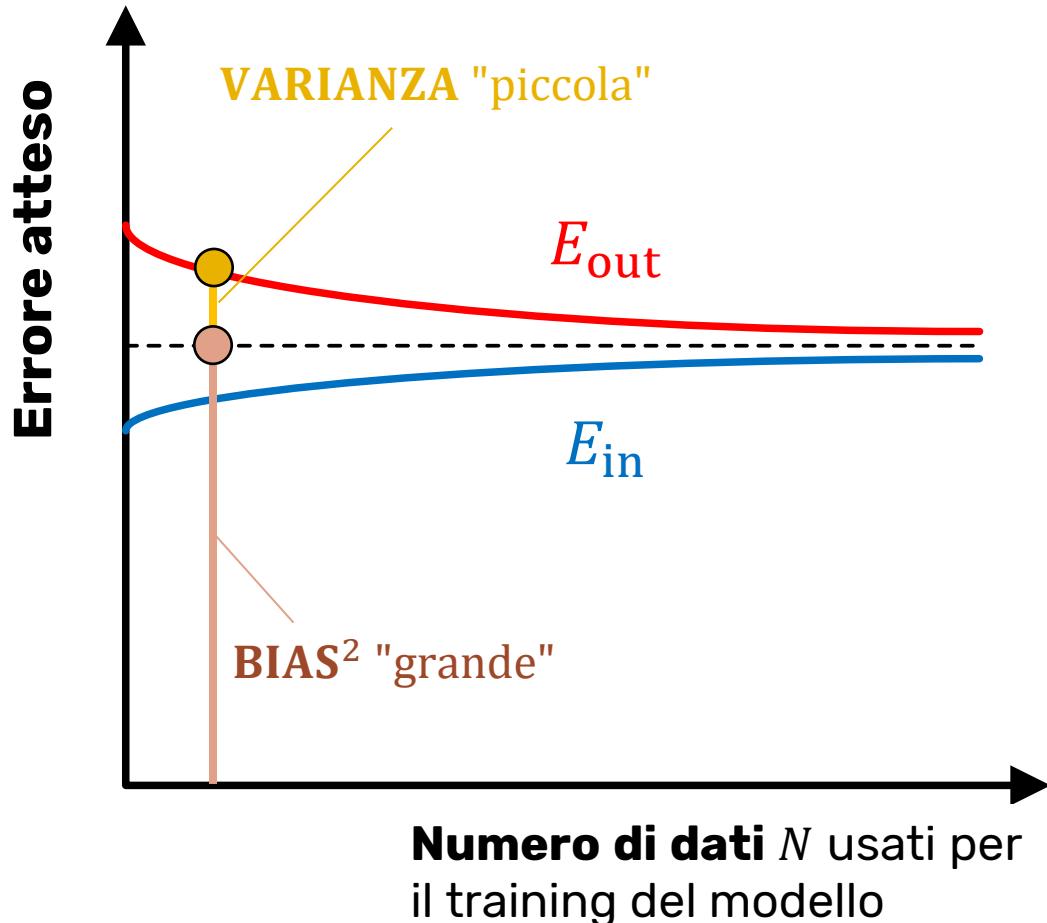
- l'errore out-of-sample **atteso** $\mathbb{E}_{\mathcal{D}}[E_{\text{out}}(g^{\mathcal{D}})]$
- l'errore in-sample **atteso** $\mathbb{E}_{\mathcal{D}}[E_{\text{in}}(g^{\mathcal{D}})]$



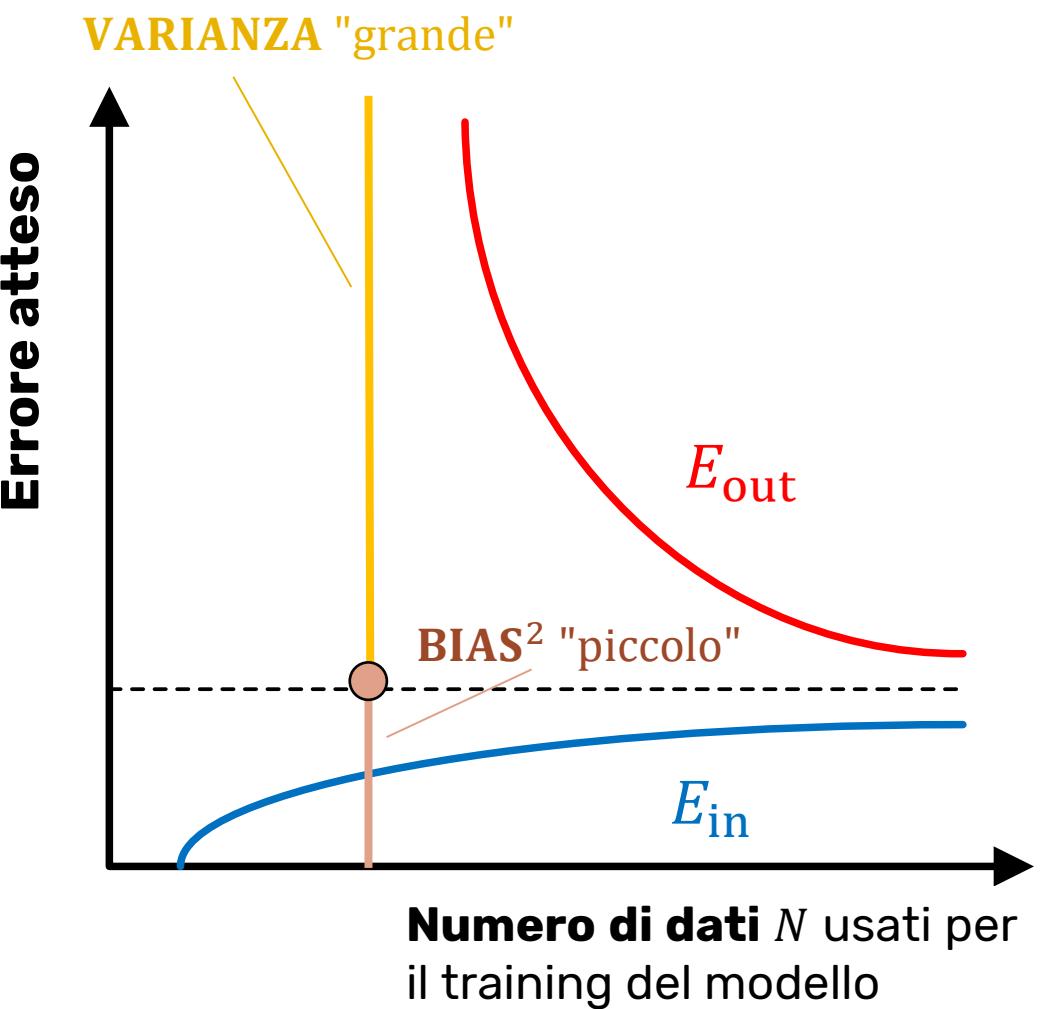
In pratica, le curve vengono calcolate **usando un solo dataset**, oppure dividendolo in più parti e prendendo la «curva media» risultante dai vari sub-datasets



Learning curves



Modello «semplice»



Modello «complesso»



Learning curves

Interpretazione

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

Risolvere un problema di bias

- Aggiungere features, per esempio combinazioni di features originarie
- Boosting

Risolvere un problema di varianza

- Usare meno features
- Acquisire più dati
- **Usare la regolarizzazione**
- Bagging



Outline

1. Introduzione al machine learning e alla data science
2. Problemi supervisionati e non supervisionati
3. Feasibility of learning
4. Bias-variance tradeoff
5. Learning curves
- 6. Overfitting**
7. Regolarizzazione
8. Validazione, cross-validation e formule di complessità ottima
9. Esercizi con codice



Overfitting



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

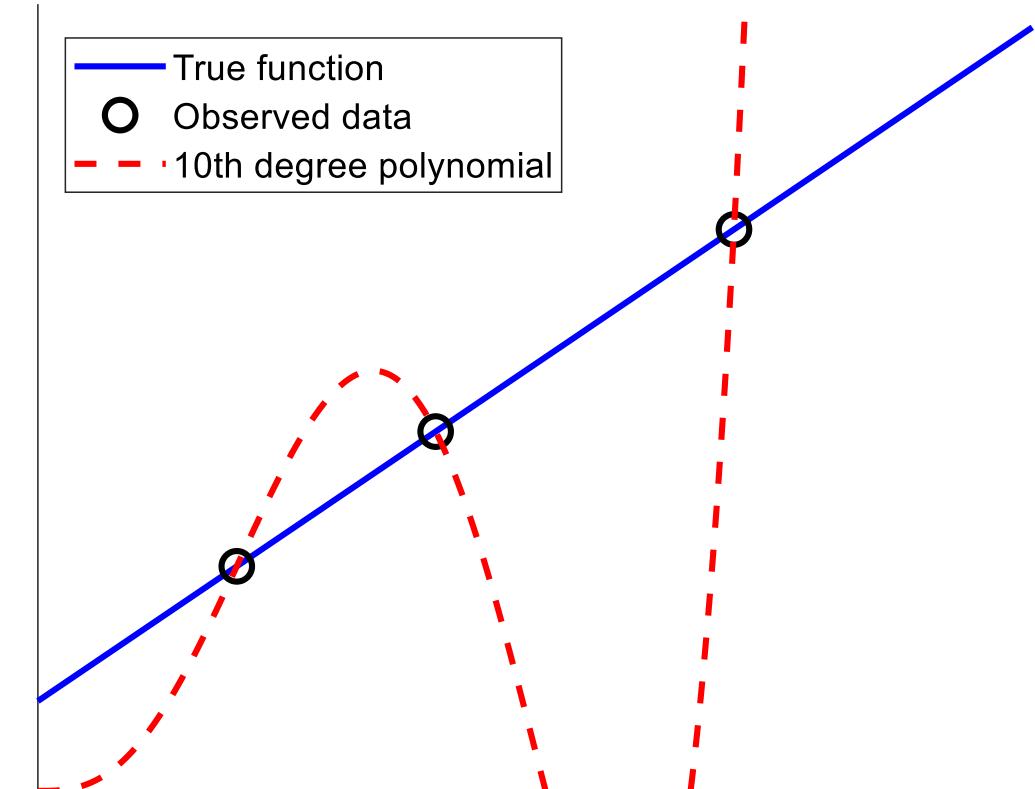
Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

Overfitting

Abbiamo già incontrato il fenomeno dell'**overfitting** quando abbiamo parlato del **tradeoff approssimazione-generalizzazione**

Abbiamo visto come dobbiamo **usare modelli più semplici se abbiamo pochi dati**, indipendentemente dalla complessità della funzione target

Introduciamo ora un'altra causa di overfitting: il **rumore stocastico sui dati** in uscita y



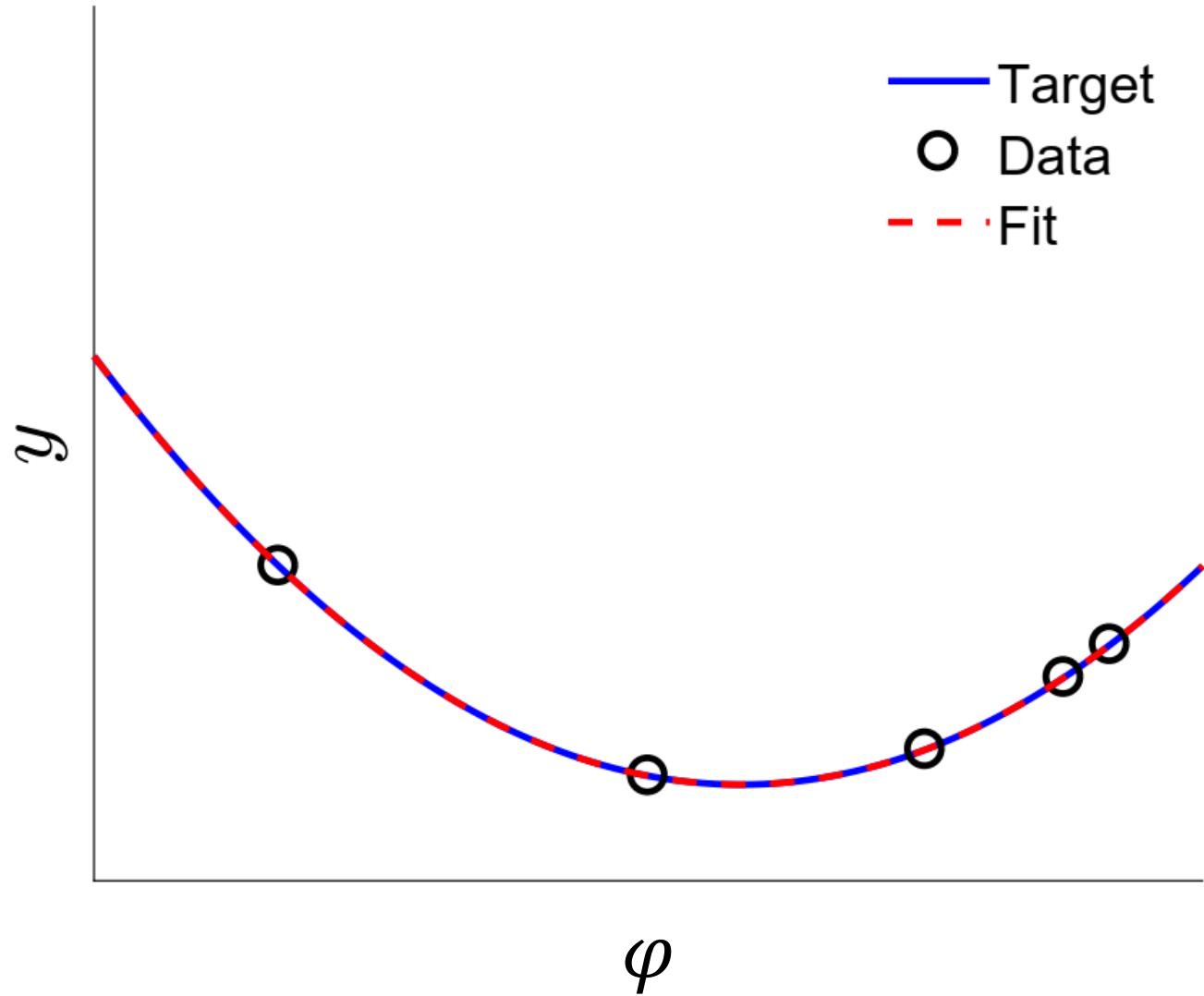
Esempio di overfitting

Consideriamo il seguente esempio:

- Funzione semplice da imparare
- $N = 5$ punti
- Modello: polinomio del 4° ordine

$$E_{\text{in}} = 0$$

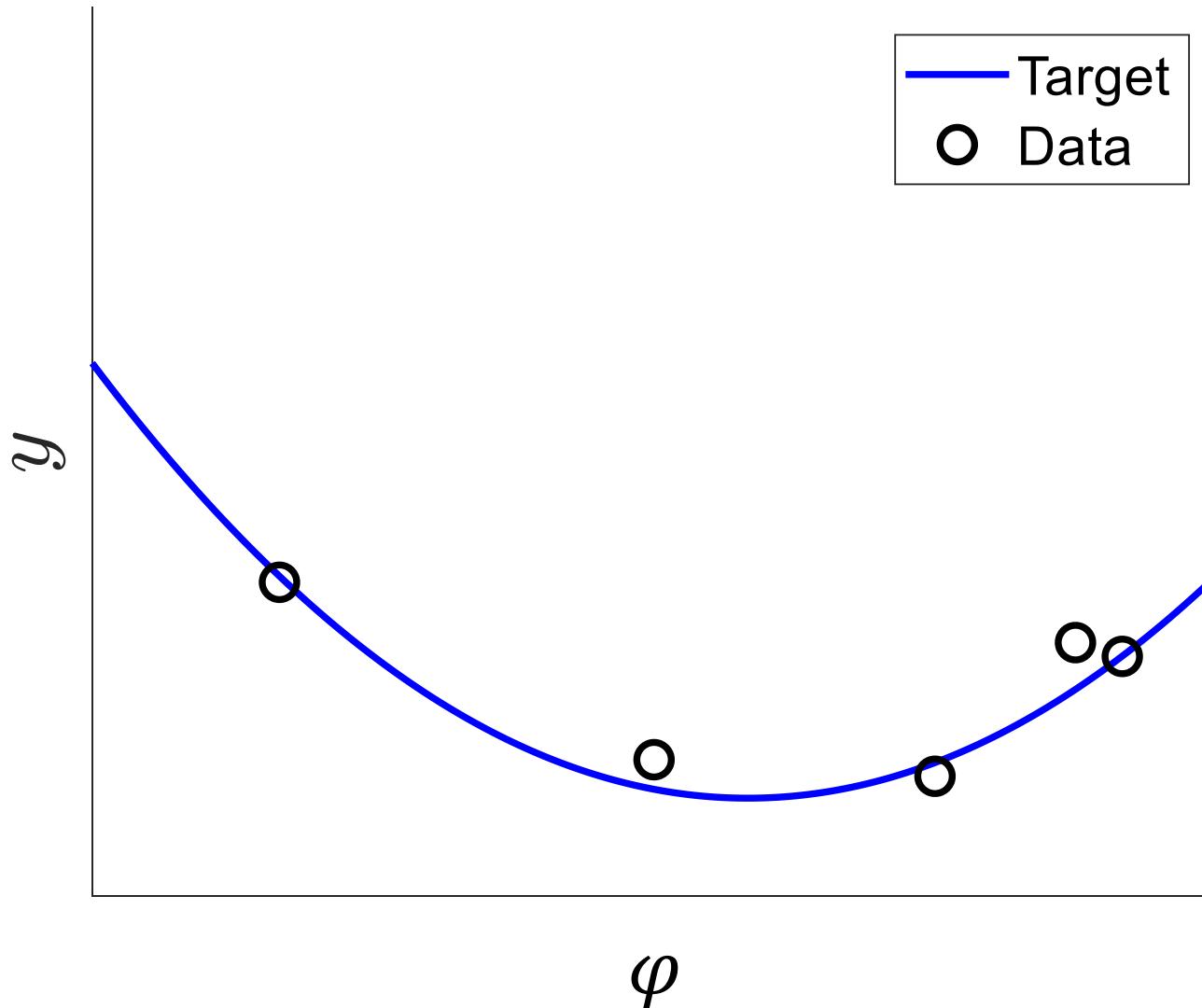
$$E_{\text{out}} = 0$$



Esempio di overfitting

Consideriamo il seguente esempio:

- Funzione semplice da imparare
- $N = 5$ punti **rumorosi**
- Modello: polinomio del 4° ordine



Esempio di overfitting

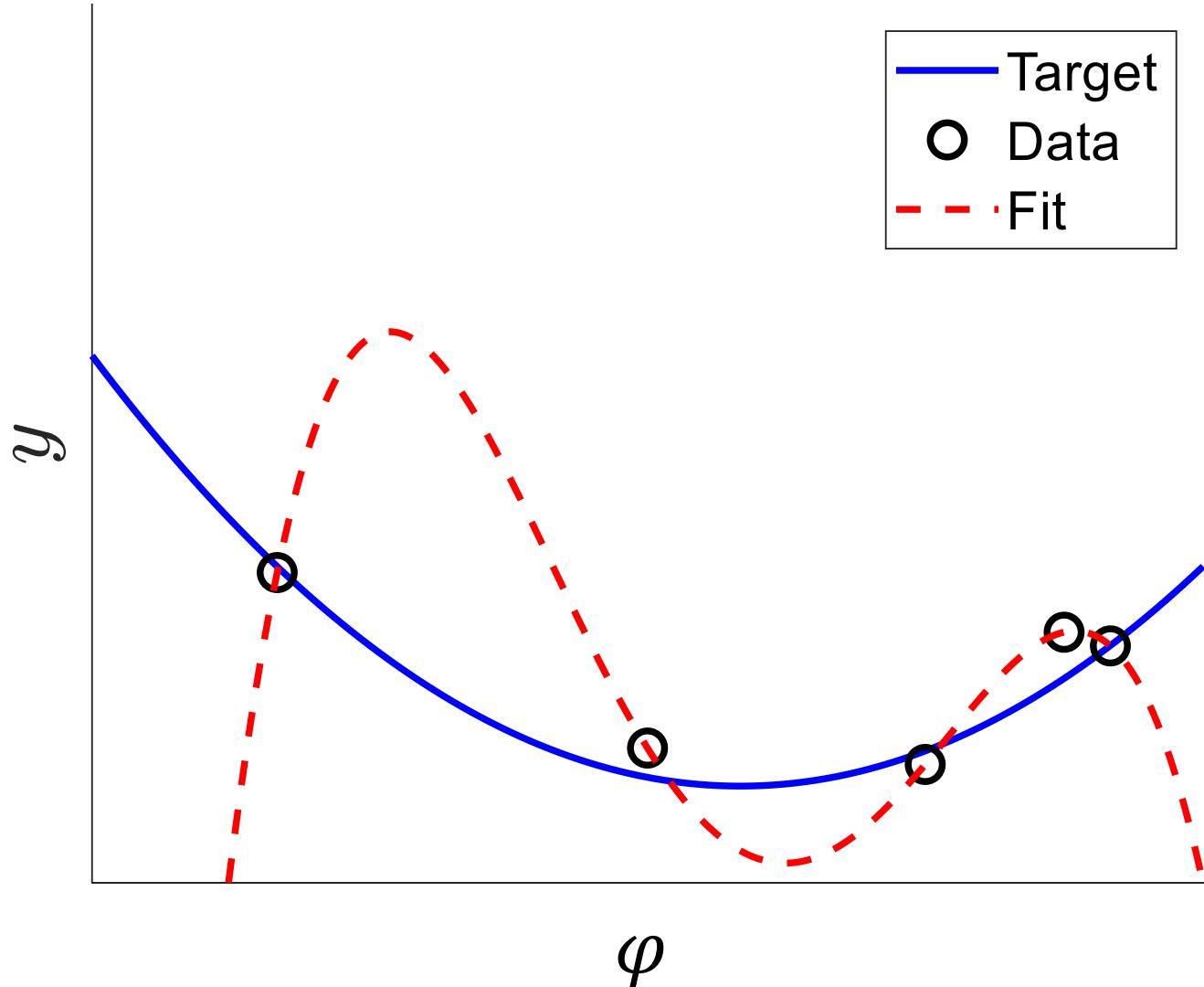
Consideriamo il seguente esempio:

- Funzione semplice da imparare

- $N = 5$ punti **rumorosi**

- Modello: polinomio del 4° ordine

$$E_{\text{in}} = 0 \quad E_{\text{out}} = \text{enorme}$$



Esempio: studente che deve apprendere dei concetti

Per comprendere in modo intuitivo il **fenomeno dell'overfitting**, consideriamo la seguente similitudine

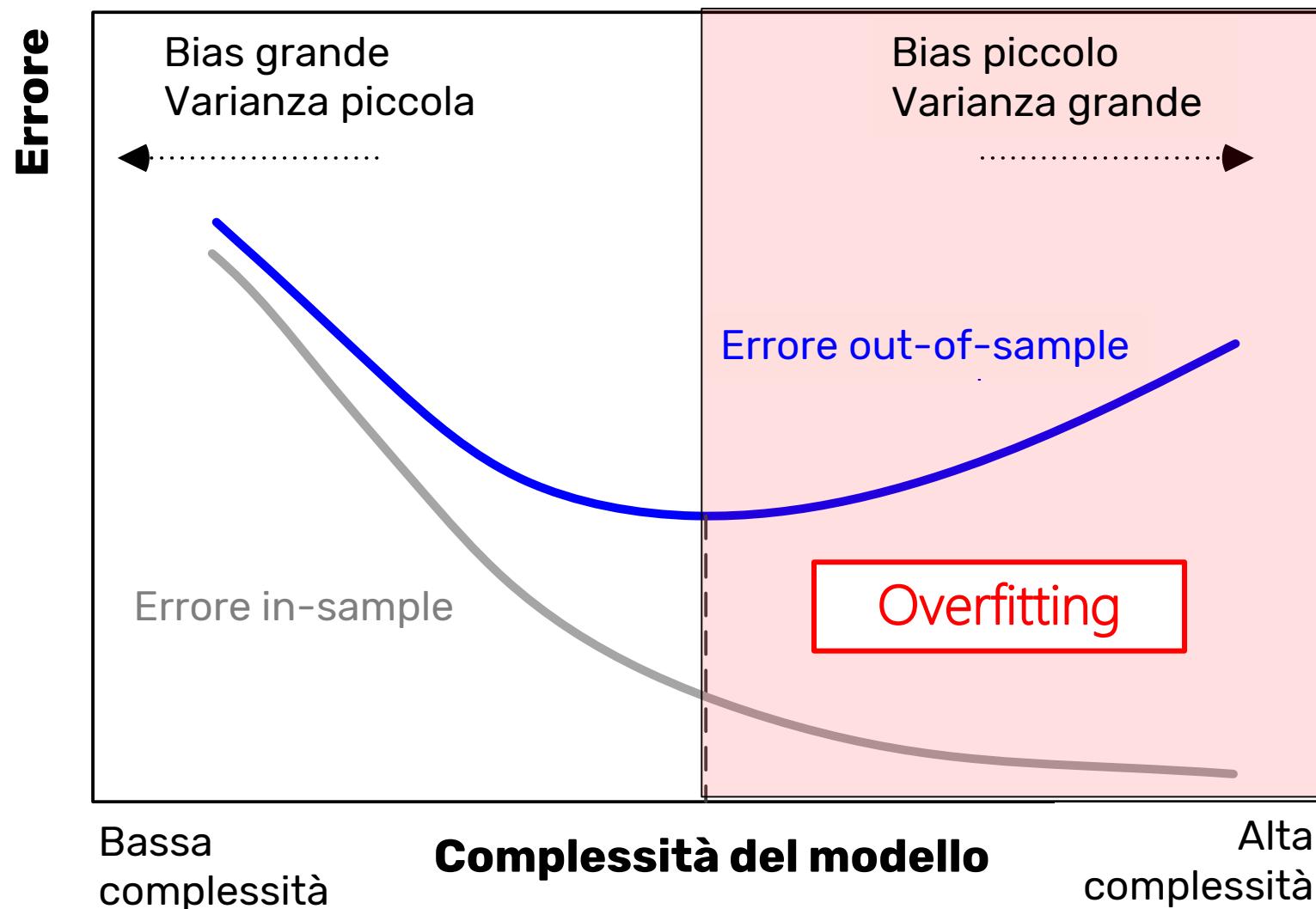
Il docente di un corso fornisce degli esercizi risolti al fine di insegnare a risolvere un problema. Gli esercizi d'esame devono per forza **essere diversi** da quelli forniti a lezione, altrimenti il docente non è in grado di capire se lo studente (o studentessa) ha solo **imparato a memoria** come risolvere gli esercizi o se ha **appreso veramente** i concetti

Nel primo caso (imparare a memoria) lo studente (o studentessa) **non ha veramente imparato**: quando si troverà di fronte un esercizio **simile (ma diverso)** non sarà in grado di risolverlo. Lo studente (o studentessa) ha **overfittato** l'esercizio visto a lezione, **senza averne generalizzato** i concetti e quindi il metodo risolutivo

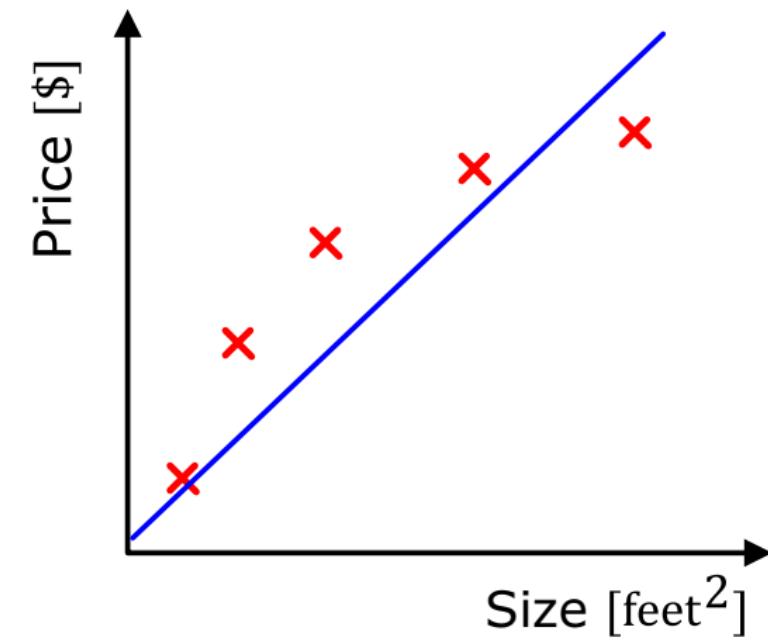


Overfitting vs. complessità del modello

- Si parla di overfitting quando diminuire E_{in} porta ad un aumento di E_{out}
- Principale fonte di non funzionamento dei modelli di learning
- L'overfitting porta a una cattiva generalizzazione
- Un modello può mostrare una cattiva generalizzazione anche se non overfittato

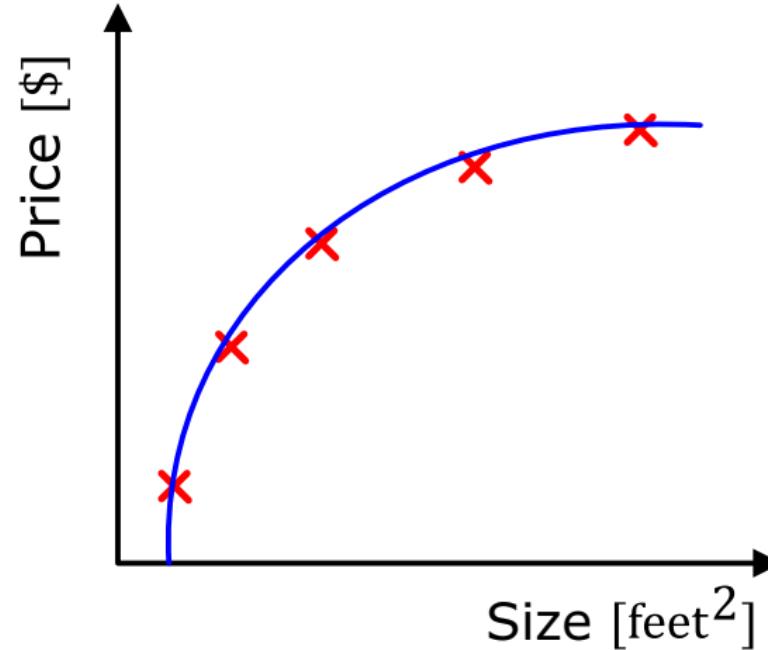


Overfitting vs. complessità del modello

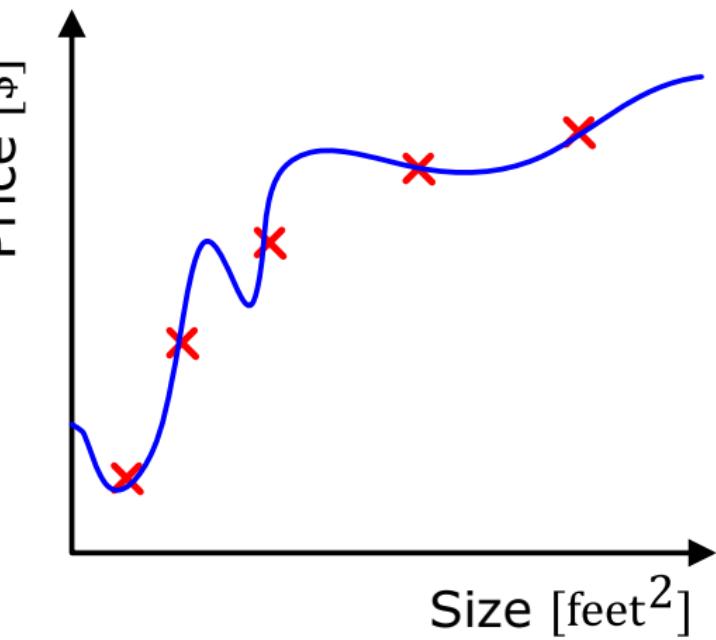


Underfit

Tanto bias



OK



Overfit

Tanta varianza



Tradeoff bias – varianza: rivisitazione

Supponiamo che via sia un **rumore stocastico** (una v.c.) η con media zero e varianza σ^2 che affligge le misure, tale che $y = f(\varphi) + \eta$

Anzichè $f(\varphi)$, osserviamo
 $y = f(\varphi) + \eta(\varphi)$

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

$$= \text{bias}^2 + \text{var} + \sigma^2$$

- L'errore stocastico σ^2 non può essere portato a zero
- Il rumore stocastico contribuisce alla varianza dell'ipotesi scelta, **causando overfitting**

→ **Errore «irriducibile»**

È un po' come quando parlavamo del limite di Cramer-Rao per gli stimatori...



Outline

1. Introduzione al machine learning e alla data science
2. Problemi supervisionati e non supervisionati
3. Feasibility of learning
4. Bias-variance tradeoff
5. Learning curves
6. Overfitting
- 7. Regolarizzazione**
8. Validazione, cross-validation e formule di complessità ottima
9. Esercizi con codice



Regolarizzazione

La **regolarizzazione** è la prima linea di difesa contro l'overfitting

Abbiamo visto che i **modelli più complessi** sono più inclini all'**overfitting**. Questo perché sono più «potenti» (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** ($\text{bias}^2 + \text{var} + \sigma^2$)

Tuttavia, se ci atteniamo solo a modelli semplici, potremmo non ottenere un'approssimazione soddisfacente della funzione target f

*Come possiamo conservare i vantaggi di **entrambi** i tipi di modello?*



Regolarizzazione

Idea: oltre che minimizzare il funzionale di costo di «fit del modello ai dati» $E_{\text{in}}(\theta) \equiv J(\theta)$, **minimizziamo anche la complessità del modello**

Al posto di $E_{\text{in}}(\theta)$, minimizziamo un **errore aumentato** $E_{\text{aug}}(\theta)$

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

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

Regolarizzatore: quanto il modello è «complesso»

Il termine λ_{reg} (iper-parametro) **pesa l'importanza** di minimizzare $E_{\text{in}} \equiv J(\theta)$ rispetto a minimizzare $\Omega(\theta)$



Esempio: regolarizzazione L_2

La regolarizzazione L_2 penalizza la somma del quadrato dei coefficienti $\theta \in \mathbb{R}^{d \times 1}$

$$E_{\text{aug}}(\theta) = \frac{1}{N} \sum_{i=1}^N \left(y(i) - h(\varphi(i); \theta) \right)^2 + \lambda_{\text{reg}} \cdot \sum_{j=0}^{d-1} (\theta_j)^2$$

- Se questa regolarizzazione L_2 viene applicata ad un problema di regressione lineare, il metodo viene chiamato **Ridge regression**
- L'intercetta θ_0 talvolta non si penalizza. In questo caso j partirebbe da 1
- Questo problema può anche essere visto come un problema di **ottimizzazione vincolata**



Esempio: regolarizzazione L_2

$$\text{minimize } E_{\text{in}}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^N \left(f(\boldsymbol{\varphi}(i)) - h(\boldsymbol{\varphi}(i); \boldsymbol{\theta}) \right)^2$$

$$\text{subject to } \boldsymbol{\theta}^\top \boldsymbol{\theta} \leq c$$

$1 \times d \quad d \times 1 \quad 1 \times 1$

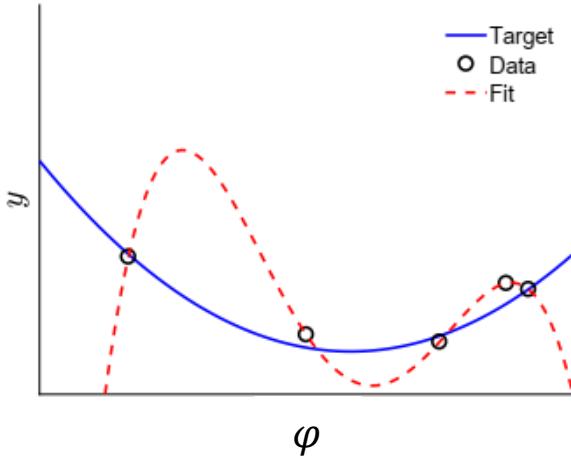
- Con questa interpretazione, stiamo esplicitamente vincolando i coefficienti $\boldsymbol{\theta}$ a **non assumere valori grandi**
- C'è una relazione tra c e λ_{reg} in modo tale che se $c \uparrow$, allora $\lambda \downarrow$
Infatti, c **più grande** significa che i **pesi** possono essere **maggiori**. Questo è uguale a impostare per un λ_{reg} **inferiore**, perché il termine di regolarizzazione sarà meno importante e quindi i pesi non verranno ridotti così tanto



Effetto dell'iperparametro di regolarizzazione λ

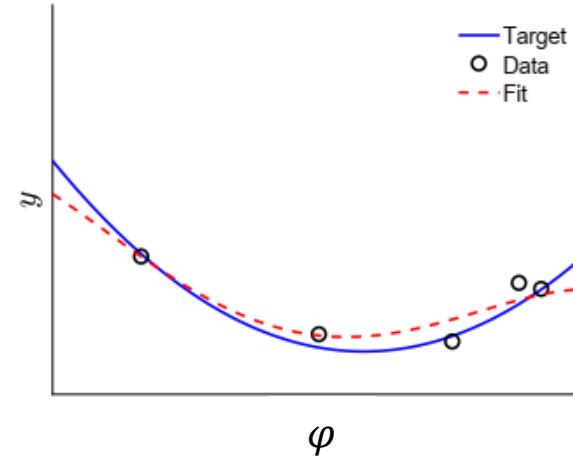
$$\lambda_{\text{reg}_1}$$

Target
○ Data
- - Fit



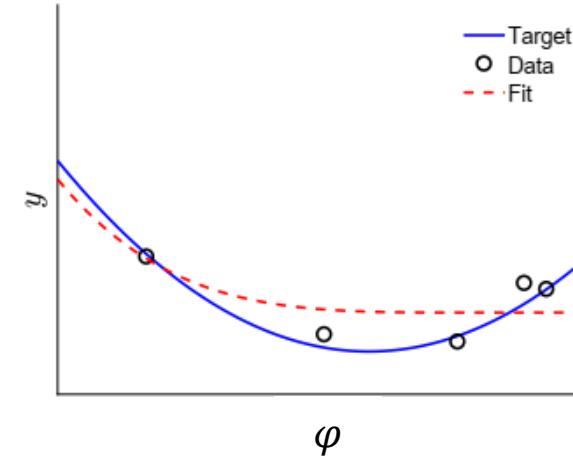
$$\lambda_{\text{reg}_2} > \lambda_{\text{reg}_1}$$

Target
○ Data
- - Fit



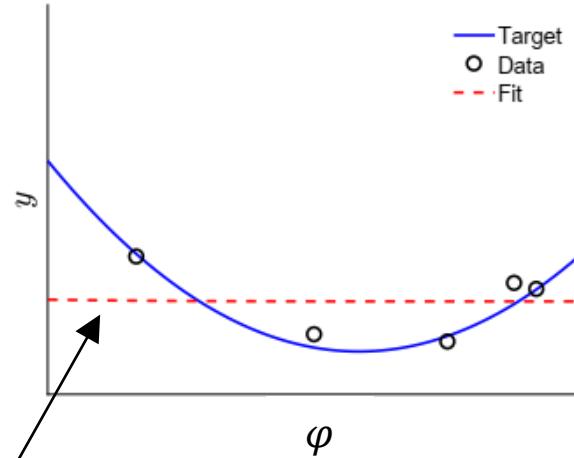
$$\lambda_{\text{reg}_3} > \lambda_{\text{reg}_2}$$

Target
○ Data
- - Fit



$$\lambda_{\text{reg}_4} > \lambda_{\text{reg}_3}$$

Target
○ Data
- - Fit



Overfit



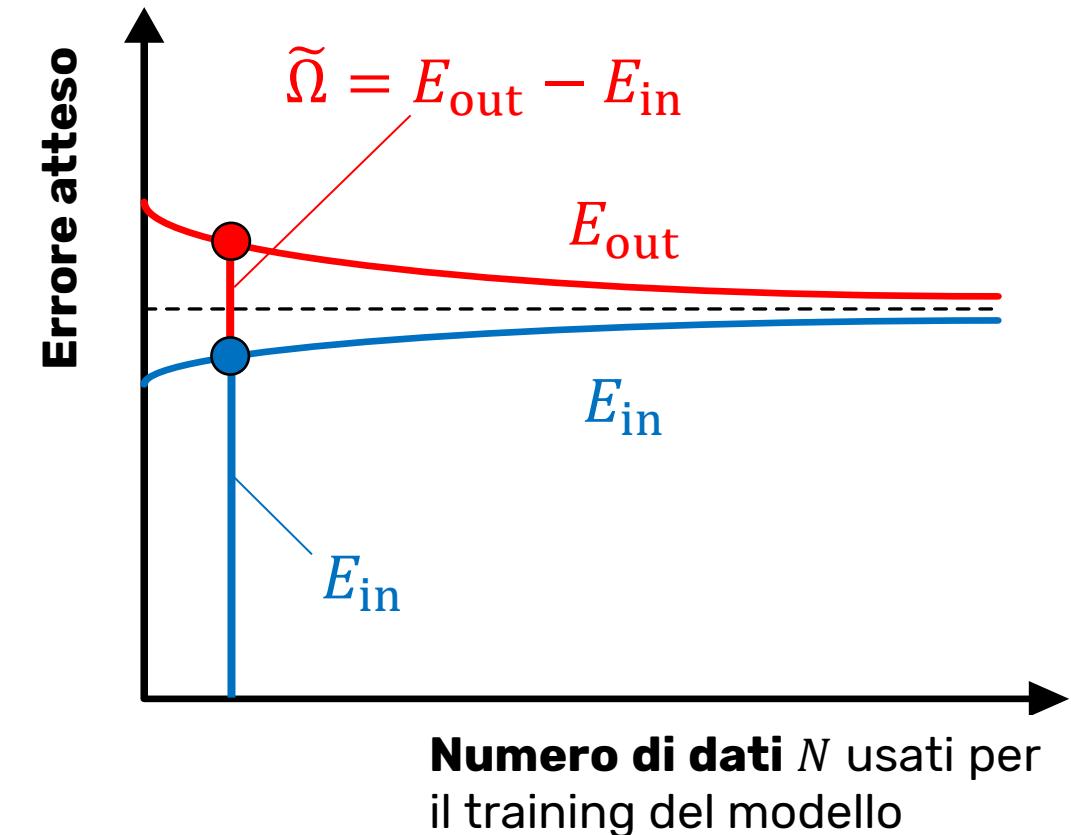
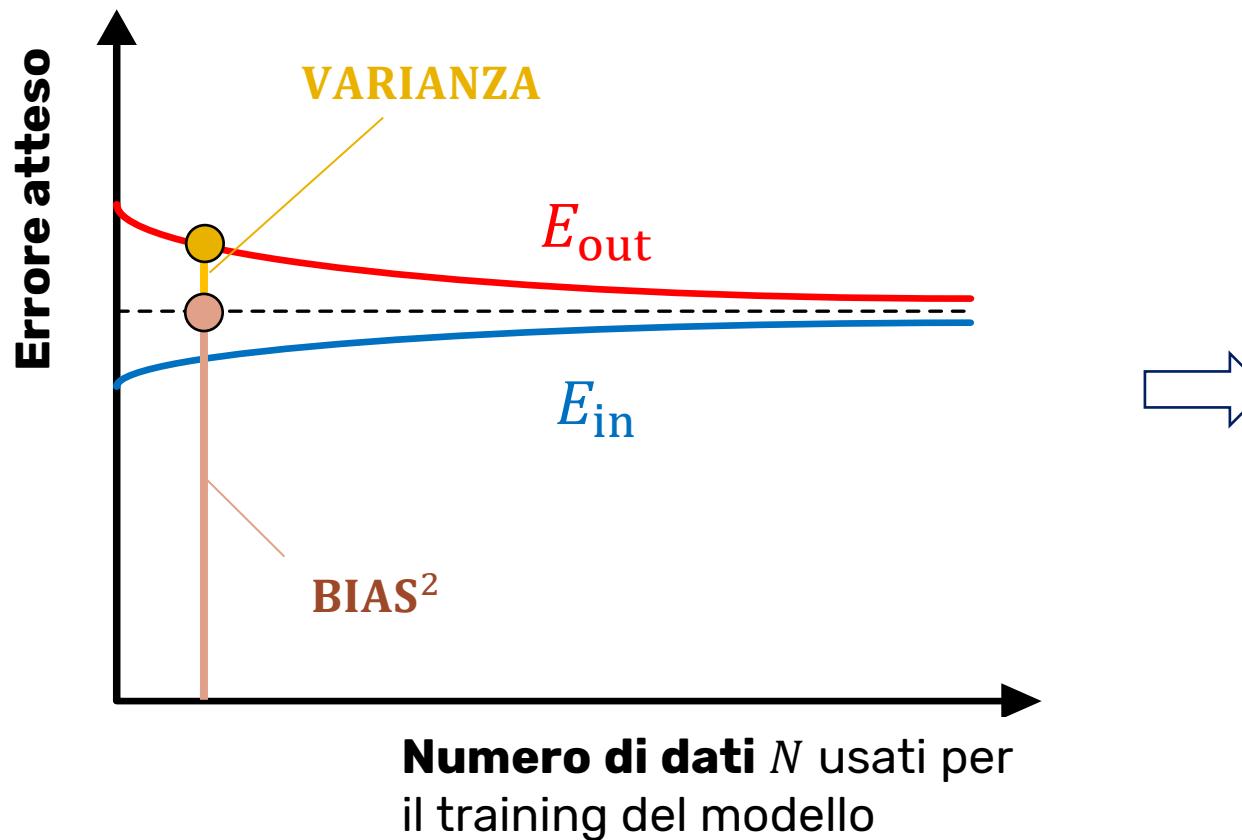
Underfit

Se regolarizzo troppo, imparerò la funzione più semplice possibile, ovvero una retta orizzontale (costante) con intercetta θ_0



Intuizione sull'importanza di E_{aug} rispetto a E_{in}

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)



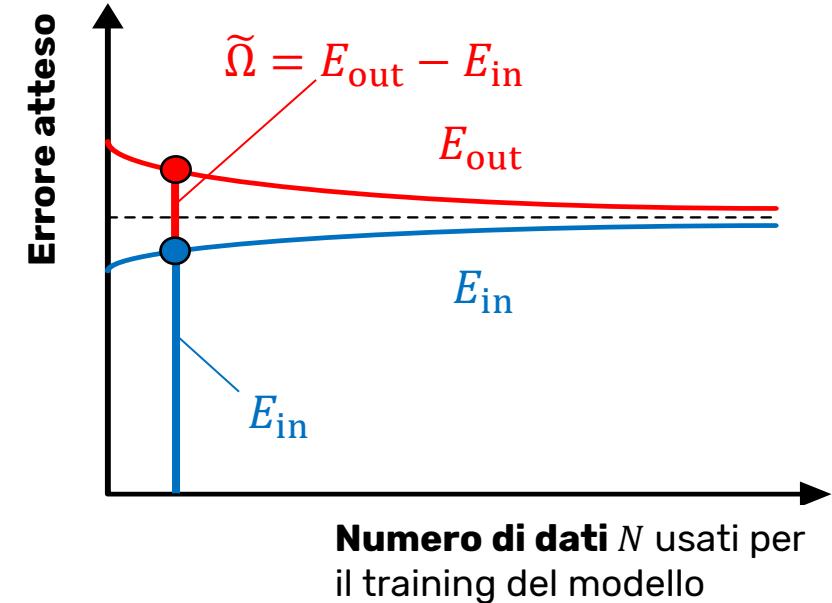
Intuizione sull'importanza di E_{aug} rispetto a E_{in}

Dal grafico precedente, oltre che tramite bias e varianza, possiamo interpretare E_{out} come la somma di due contributi:

$$E_{\text{out}}(\theta) = E_{\text{in}}(\theta) + \tilde{\Omega}(\theta)$$

Ricordando la definizione di E_{aug} abbiamo

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



L'errore E_{aug} è **migliore** rispetto ad E_{in} come proxy per E_{out}



Intuizione sull'importanza di E_{aug} rispetto a E_{in}

Il Santo Graal del machine learning sarebbe avere
un'espressione di E_{out} da minimizzare

- In questo modo, sarebbe possibile minimizzare direttamente l'errore out-of-sample invece di quello in-sample (o dell'errore aumentato)
- La **regolarizzazione** aiuta nello stimare la quantità $\Omega(\theta)$, che, sommata ad E_{in} , fornisce E_{aug} , il quale è una stima di E_{out}



Scelta del termine di regolarizzazione

Esistono diversi tipi di regolarizzazione. I più usati sono:

- **Regolarizzazione L_2** : chiamata anche penalità **Ridge** $\Omega(\theta) = \sum_{j=0}^{d-1} (\theta_j)^2$
- **Regolarizzazione L_1** : chiamata anche penalità **Lasso** $\Omega(\theta) = \sum_{j=0}^{d-1} |\theta_j|$
- **Regolarizzazione elastic-net**: $\Omega(\theta) = \beta \sum_{j=0}^{d-1} (\theta_j)^2 + (1 - \beta) \sum_{j=0}^{d-1} |\theta_j|$

La penalità **Ridge** tende a ridurre tutti i coefficienti a un **valore inferiore**

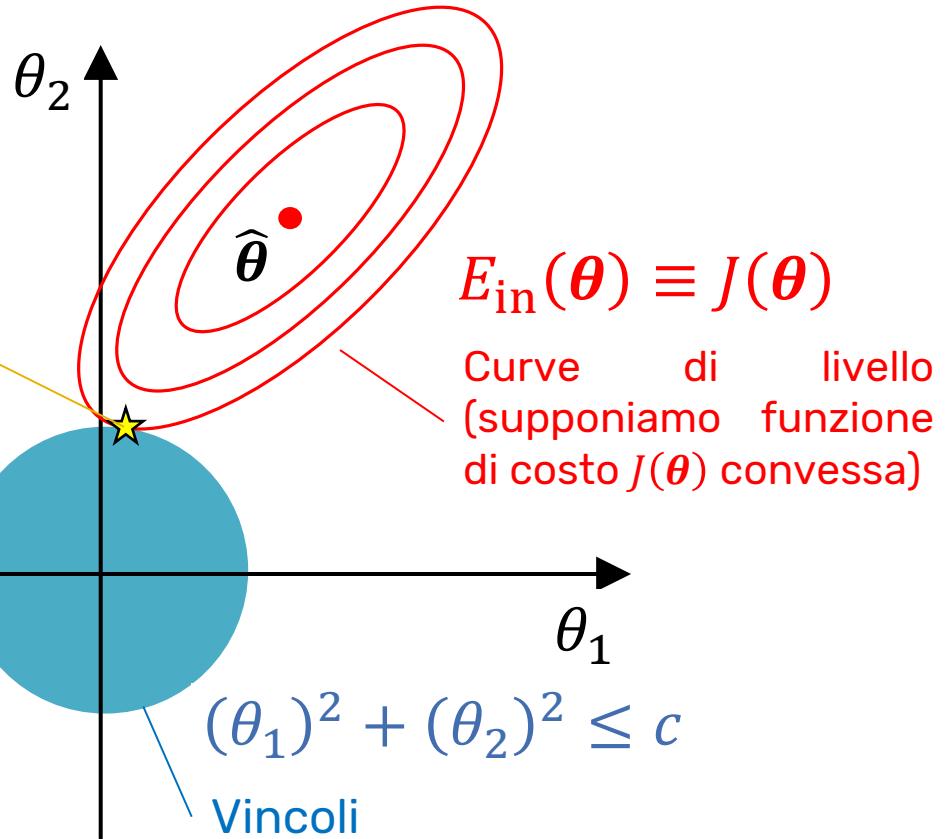
La penalità **Lasso** tende a portare più coefficienti **esattamente a zero**



Scelta del termine di regolarizzazione

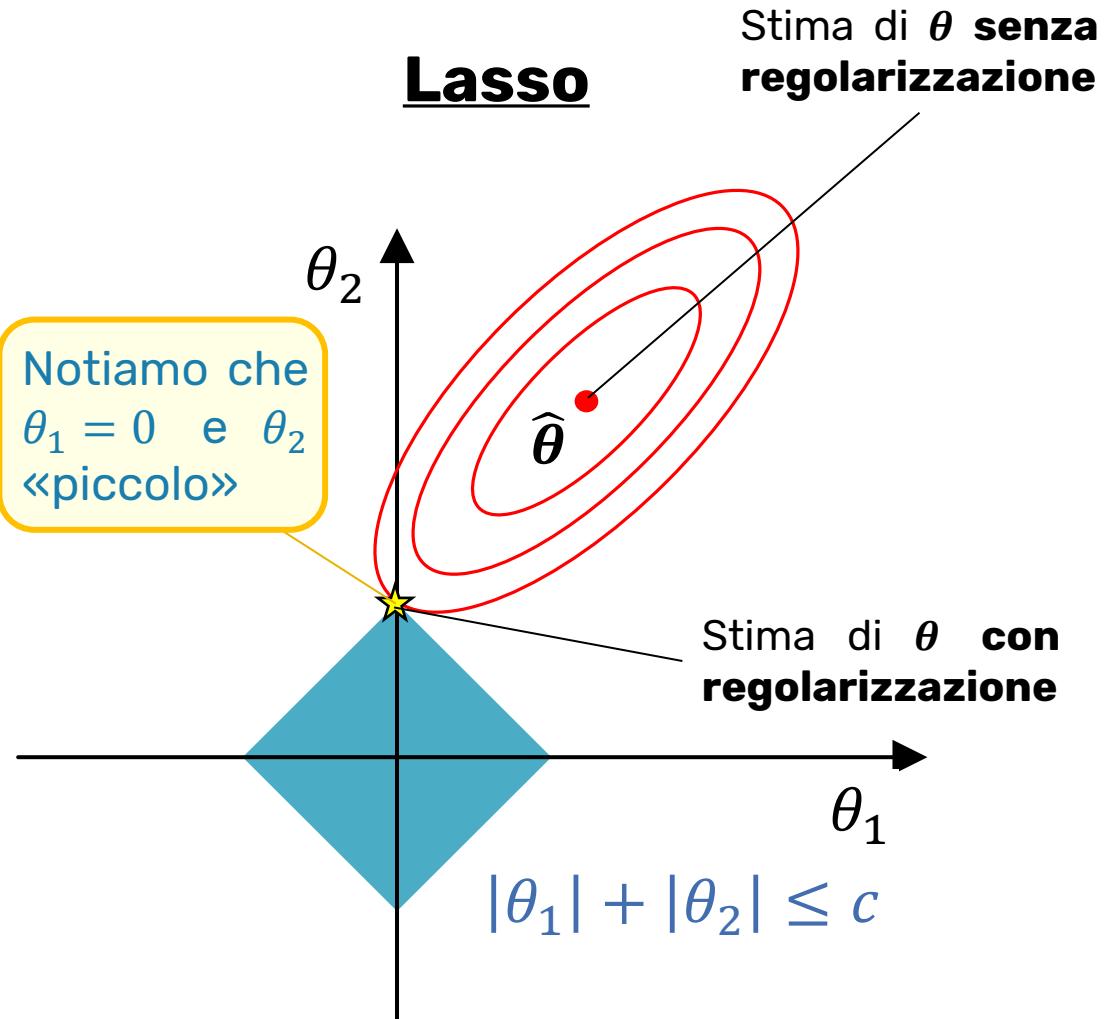
Ridge

Notiamo che θ_1 e θ_2 sono «piccoli»



Lasso

Notiamo che $\theta_1 = 0$ e θ_2 «piccolo»



Regolarizzazione e bias-varianza tradeoff

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

- La regolarizzazione **aumenta di poco il bias** (perché ottengo un modello più 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-validation**



Outline

1. Introduzione al machine learning e alla data science
2. Problemi supervisionati e non supervisionati
3. Feasibility of learning
4. Bias-variance tradeoff
5. Learning curves
6. Overfitting
7. Regolarizzazione
- 8. Validazione, cross-validation e formule di complessità ottima**
9. Esercizi con codice



Validazione

L'errore out-of-sample può essere visto come:

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

Regolarizzazione

$$E_{\text{out}}(\boldsymbol{\theta}) = E_{\text{in}}(\boldsymbol{\theta}) + \underbrace{\text{penalità per la complessità del modello}}_{\text{La REGOLARIZZAZIONE stima questa quantità}}$$

Validazione

$$\underbrace{E_{\text{out}}(\boldsymbol{\theta})}_{\text{La VALIDAZIONE stima questa quantità}} = E_{\text{in}}(\boldsymbol{\theta}) + \text{penalità per la complessità del modello}$$



Validazione

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 stimato

Nota: La regolarizzazione e la validazione non vengono usate solo nell'ambito machine learning! **Sono fondamentali anche nell'identificazione di sistemi dinamici!**



Set di validazione

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

1. **Rimuovo** un subset di dati dai dati totali → questo subset non è usato per il training (stima)
2. **Stimo** il modello sulla parte di dati rimanente → il modello sarà allenato su *meno* dati
3. **Valuto** le performance del modello sul subset di dati che ho rimosso al punto 1. → in questo modo ottengo una stima corretta (unbiased) dell'errore out-of-sample
4. **Ri-alleno** il modello su tutti i dati

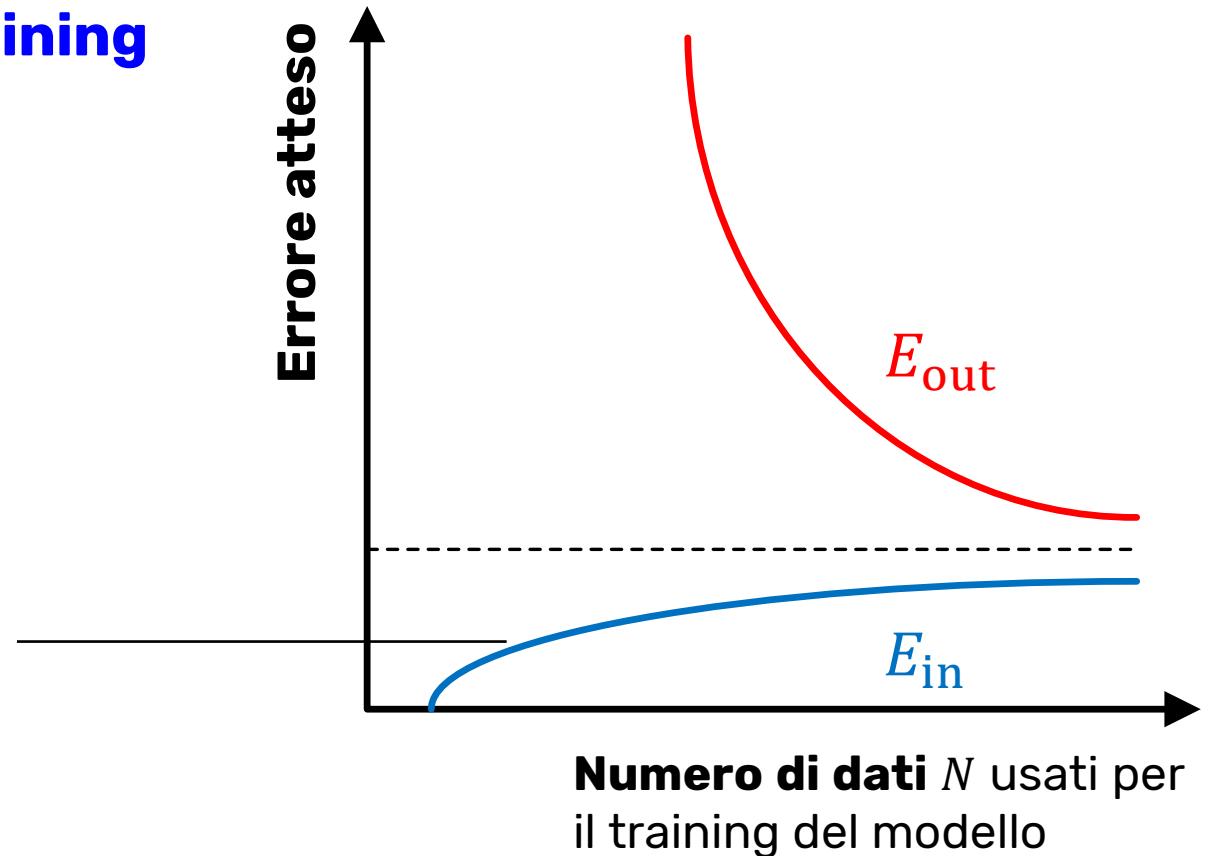


Set di validazione

Supponiamo di avere il dataset $\mathcal{D} = \{(\varphi(1), y(1)), \dots, (\varphi(N), y(N))\}$. Si procede come segue:

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

- N_{val} «**piccolo**»: stima di E_{out} non buona
- N_{val} «**grande**»: possibilità di imparare un modello non buono (guardare le learning curves)



Set di validazione

$$\mathcal{D} \rightarrow \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}}$$

$$\downarrow \quad \downarrow \quad \downarrow$$

$$N_{\text{val}} \quad N - N_{\text{val}} \quad N_{\text{val}}$$

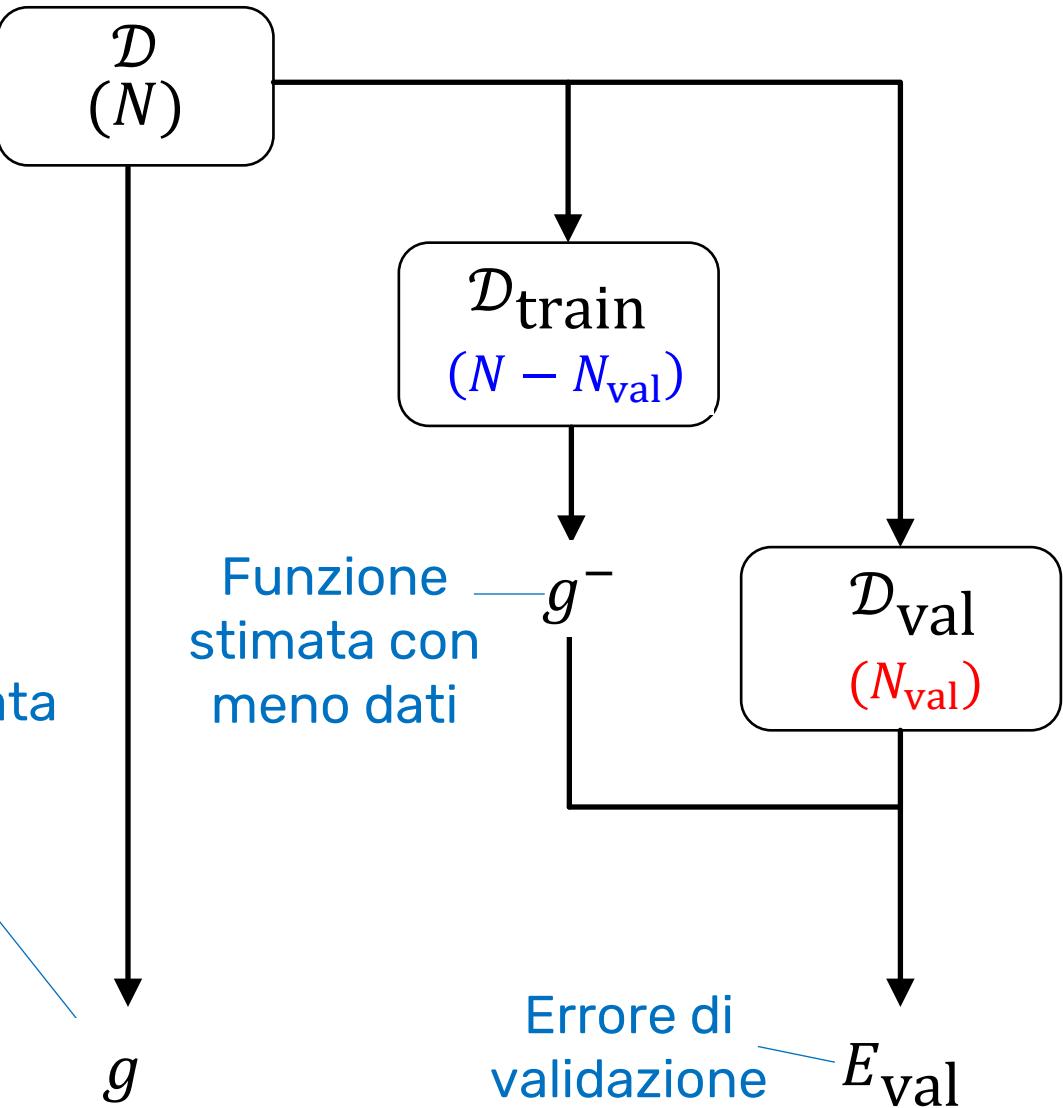
$$\mathcal{D} \Rightarrow g \quad \mathcal{D}_{\text{train}} \Rightarrow g^-$$

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

Rule of thumb

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

Funzione
(modello) stimata
su tutti i dati



Selezione del modello migliore usando validazione

Le procedure di validazione possono essere utilizzate per due scopi:

1. **Valutare le performance** del modello stimato (e.g. stimare E_{out})
2. **Scegliere il modello migliore** da un insieme di diversi modelli

Per esempio, la scelta del modello migliore include:

- scegliere tra un modello lineare e uno non lineare
- scegliere il numero di regressori da usare
- scegliere il valore del parametro di regolarizzazione λ_{reg}
- ...qualsiasi altra scelta che influisce sull'apprendimento del modello



Selezione del modello migliore usando validazione

Supponiamo di avere N_m **set di modelli**

$\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_{N_m}$ tra cui imparare un modello

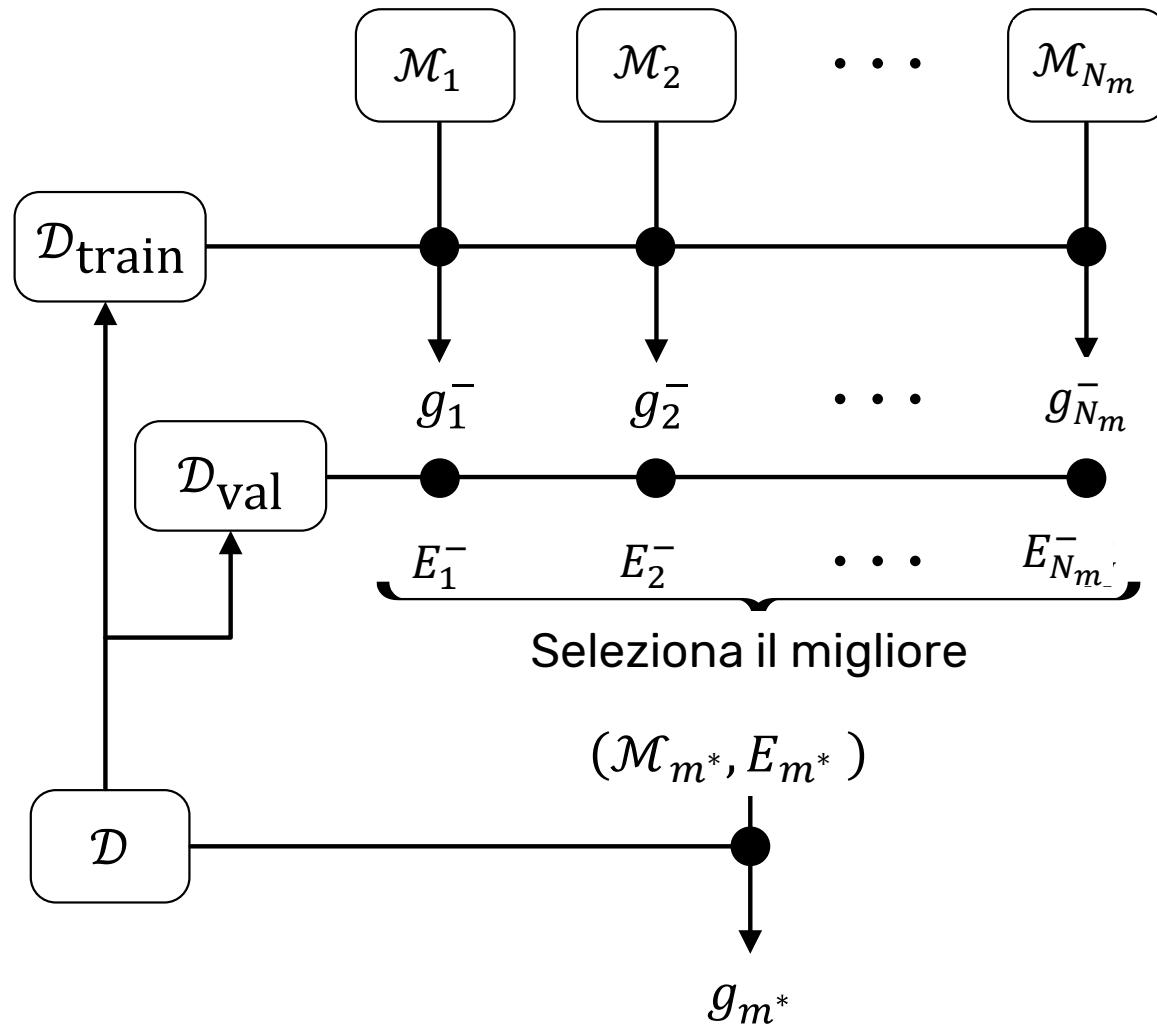
- **Stimo** $\bar{g_m}$ usando $\mathcal{D}_{\text{train}}$ per ogni set di modelli

- **Valuto** $\bar{g_m}$ usando \mathcal{D}_{val}

$$E_m^- = E_{\text{val}}(\bar{g_m}) \quad m = 1, \dots, N_m$$

- **Seleziono** il modello $m = m^*$ con l'errore

E_m^- più basso



Selezione del modello migliore usando validazione

Problema: se uso il dataset di validazione \mathcal{D}_{val} «**tante volte**» per compiere delle scelte (e.g. scegliere tra modelli diversi), allora il dataset di validazione \mathcal{D}_{val} **non fornisce più una buona stima dell'errore out-of sample** E_{out}

Intuzione: usare \mathcal{D}_{val} per compiere delle scelte su quale modello usare fa sì che **taли сcelte siano dipendenti dai particolari valori dei dati** contenuti in \mathcal{D}_{val} . Chi mi garantisce che con dati diversi avrei compiuto le medesime scelte?

Quello che sta succedendo è che stiamo **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}

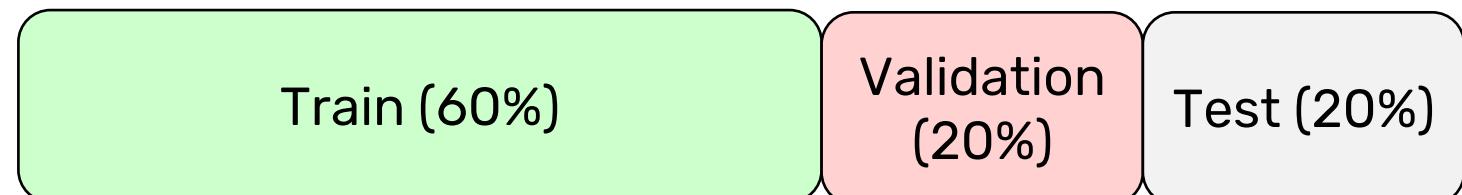


«Contaminazione» dei dataset

Abbiamo finora ottenuto **tre stime** dell'errore E_{out}

Contaminazione: bias ottimistico nello stimare E_{out} (e.g. dire che E_{out} è più piccolo di quanto è in realtà)

- **Training set:** totalmente contaminato
- **Validation set:** un po' contaminato
- **Test set:** totalmente «pulito»



Cross-validation

La divisione del dataset in **tre parti** (train, validation, test) è fattibile se i dati a disposizione sono molti (dove «molti» dipende dal problema...si guardino le learning curves)

In teoria, vorremmo che:

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

(N_{val} piccolo) (N_{val} grande)

È l'unico che posso calcolare!

- N_{val} **grande**: in questo caso, il valore di E_{val} calcolato usando g^- sarebbe simile al valore di E_{out} ottenuto da 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, il valore di E_{out} ottenuto g^- sarebbe simile al valore di E_{out} ottenuto da g (la funzione stimata su tutti i dati), poichè uso tanti dati $N - N_{\text{val}}$ per il train di g^- . Questo è il valore che mi interessa ma che non posso calcolare direttamente



Cross-validation

La **cross-validation** permette di «avere N_{val} sia grande che piccolo»

Leave-one-out cross-validation

Usiamo $N - 1$ dati per il training e $N_{\text{val}} = 1$ dato per la validazione

- Dato rimosso dai dati usati usati per il train e usato per la validazione

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

dove \mathcal{D}_i è il dataset di training senza il dato i -esimo.

La funzione (il modello) imparata usando \mathcal{D}_i è g_i^-



Cross-validation

L'errore di validazione sul punto «rimosso» $\varphi(i)$ è $\ell(i) = E_{\text{val}}(g_i^-) = \ell(y(i), g_i^-(\varphi(i)))$

E' possibile definire l'**errore di cross-validation** E_{cv} come:

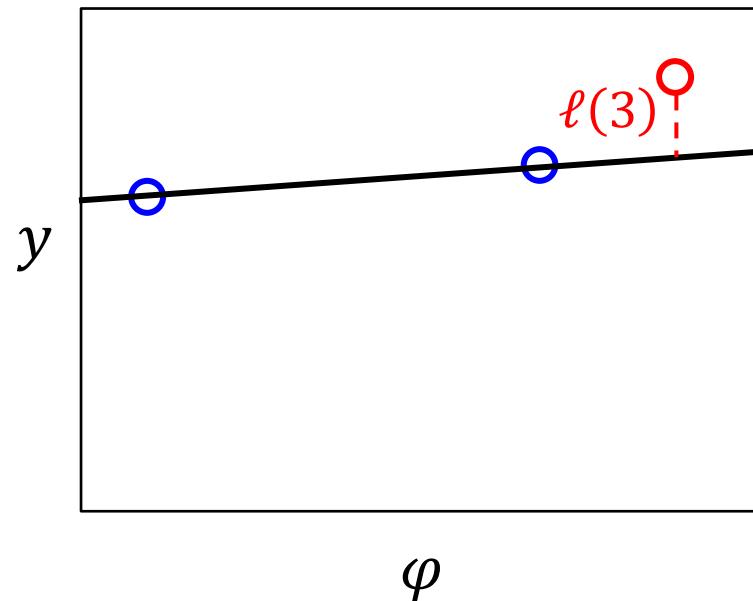
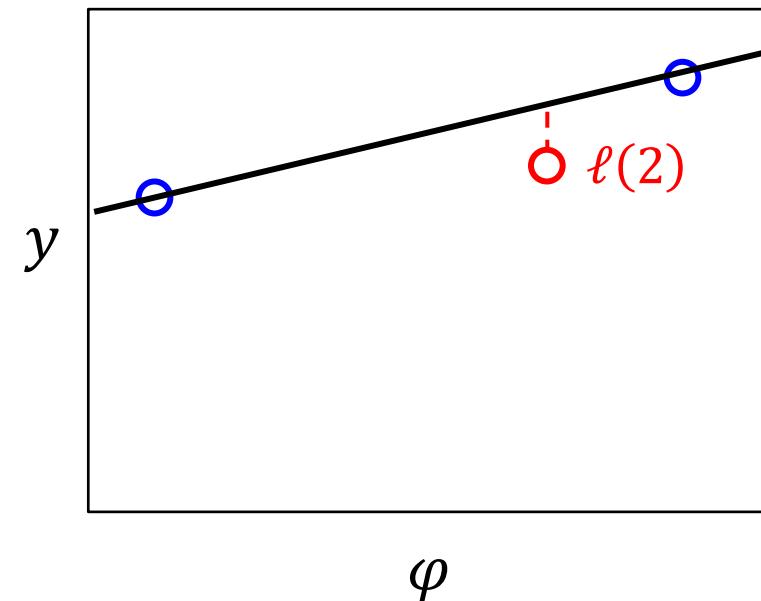
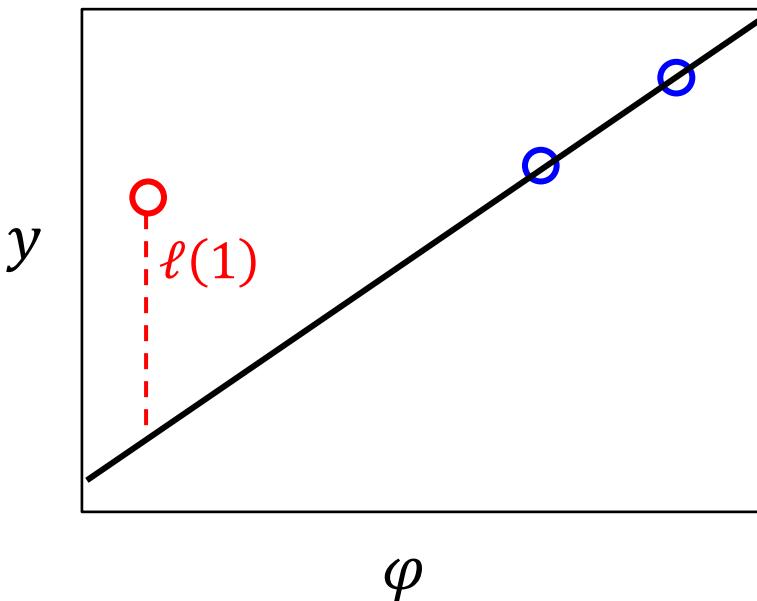
$$E_{\text{cv}} = \frac{1}{N} \sum_{i=1}^N \ell(i)$$

- In questo modo, stimo N modelli usando $N - 1$ dati, e li valido usando N stime dell'errore di validazione (calcolate ognuna su $N_{\text{val}} = 1$ dati)
- È possibile anche calcolare la deviazione standard (campionaria) dei vari errori $\ell(i)$. Se è grande, vuol dire che il modello è molto sensibile ai dati sui quali viene allenato



Esempio di cross-validation

Esempio: supponiamo di voler imparare un modello lineare usando un regressore e $N = 3$ osservazioni, utilizzando $N_{\text{val}} = 1$ per fare la cross-validation



$$E_{\text{cv}} = \frac{1}{3}(\ell(1) + \ell(2) + \ell(3))$$



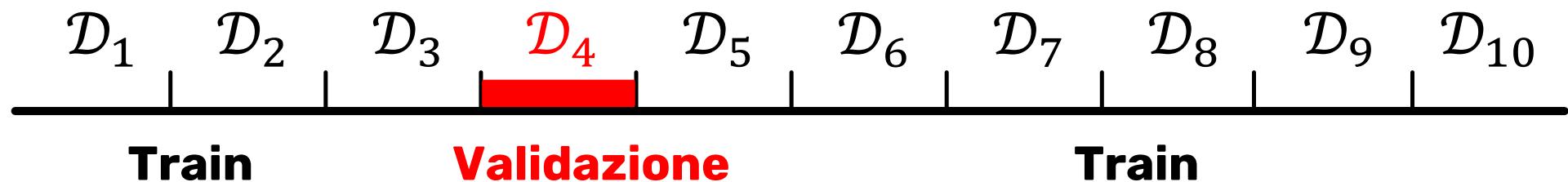
Cross-validation K -fold

La cross-validation con $N_{\text{val}} = 1$ (leave-one-out cross-validation) ha gli svantaggi che:

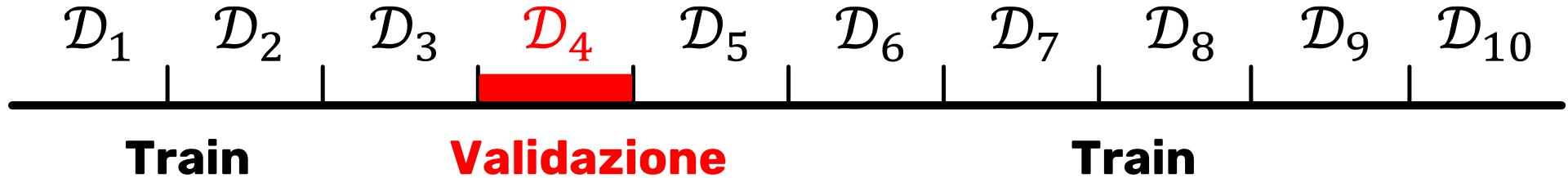
- È **computazionalmente costosa**. Nel caso volessimo usarla per scegliere tra M modelli, richiederebbe un totale di N sessioni di training per ciascuno degli M modelli
- La stima dell'errore E_{cv} ha una **varianza elevata**, poiché si basa su un solo dato

E' possibile riservare più punti per la validazione suddividendo il training set in «**folds**».

Per esempio, se $K = 10$ avremmo



Cross-validation K -fold



- La K -fold cross-validation richiede N/N_{val} sessioni di train, ognuna con $N - N_{\text{val}}$ dati
- Un buon **compromesso** è usare $K = 10$

$$10 - \textbf{fold cross validation: } N_{\text{val}} = \frac{N}{10}$$

- Attenzione a **non ridurre troppo il training set** (guardare le learning curves)



Esempio: modo corretto di usare la cross-validation

Consideriamo un problema di **classificazione** con un **tanti regressori** (features). Una strategia per costruire un modello potrebbe essere la seguente:

1. Trovare un **sottoinsieme di regressori** che mostrano una forte **correlazione** (univariata) con le label
2. **Usando questo sottoinsieme** di predittori, imparare un classificatore
3. Utilizzare la cross-validation per **stimare gli iperparametri** (e.g. model selection) e per **stimare l'errore out-of-sample**

Si tratta di una **corretta applicazione** della cross-validation?

NO!



Esempio: modo corretto di usare la cross-validation

I regressori selezionati per la stima del modello hanno un **vantaggio sleale**, in quanto sono **stati scelti sulla base di tutti i dati** (step 1)

Rimuovere dati per la cross-validation **dopo che i regressori sono stati selezionati** non imita correttamente l'applicazione del classificatore a un set di test completamente indipendente

I regressori (e quindi il modello) **hanno «già visto» i dati di validazione**

I dati utilizzati per la validazione sono stati già **utilizzati per effettuare una scelta** che ha coinvolto le label di output (questo **non è corretto**)



Esempio: modo corretto di usare la cross-validation

Quello che si sarebbe dovuto fare sarebbe stato di **includere la scelta del sottoinsieme dei regressori all'interno della procedura di cross-validation**, oppure usare la regolarizzazione per la stima del modello



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 **troppi pochi dati** per poter usare validazione o cross-validation

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**

Vedremo le seguenti formule \ criteri di complessità:

- **Akaike Information Criterion (AIC)**. Un indicatore **equivalente** è il Final Prediction Error (FPE) utilizzato per la stima di modelli dinamici
- **Minimum Description Length (MDL)**, derivante dal Bayesian Information Criterion (BIC)



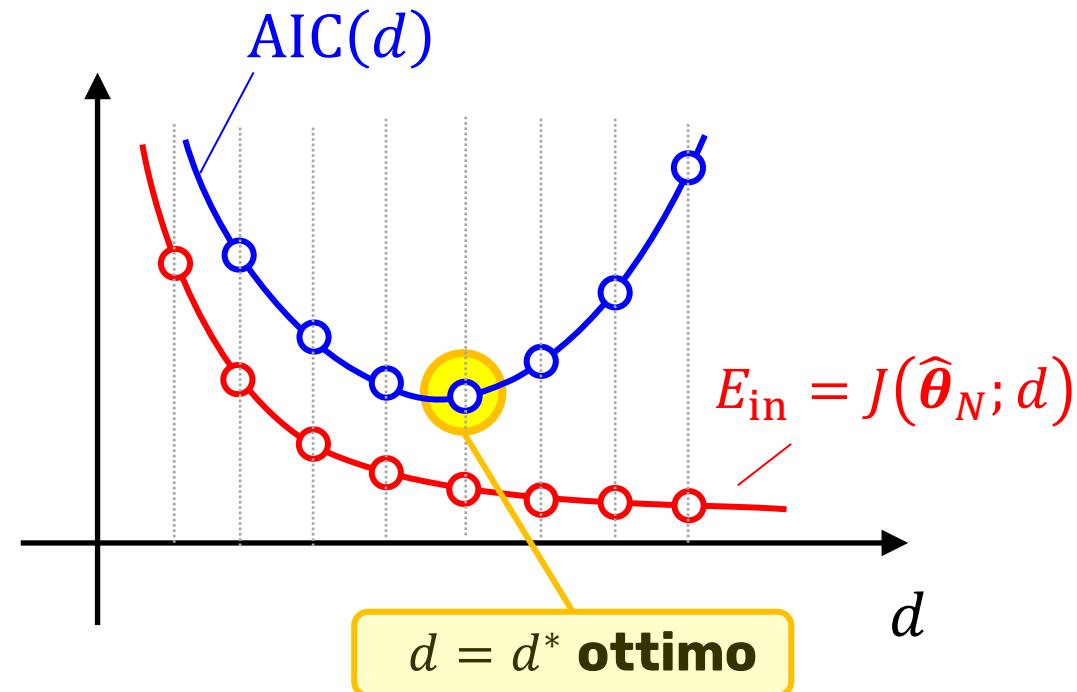
Formule di complessità ottima

Supponiamo di avere un modello con d parametri. Indichiamo la stima dei parametri, ottenuta con N dati, con $\hat{\theta}_N \in \mathbb{R}^{d \times 1}$. La stima è ottenuta mimizzando la funzione di costo $J(\theta; d)$ dove esplicitiamo la dipendenza del costo dal numero di parametri d

Akaike Information Criterion (AIC)

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

- $d \uparrow \Rightarrow \frac{d}{N} \uparrow$
- $d \uparrow \Rightarrow J(\hat{\theta}_N; d) \downarrow$

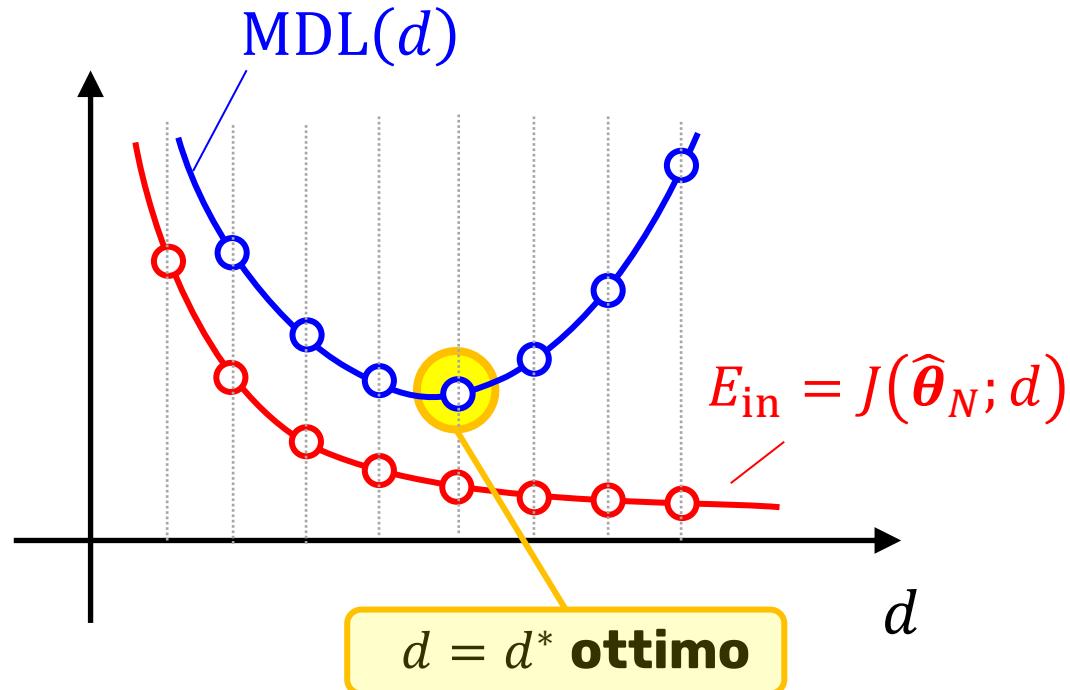


Formule di complessità ottima

Minimum Description Length (MDL)

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

- $d \uparrow \Rightarrow \ln[N] \cdot \frac{d}{N} \uparrow$
- $d \uparrow \Rightarrow J(\hat{\theta}_N; d) \downarrow$



Il criterio MDL di fatto si comporta come AIC. È però interessante confrontare più nel dettaglio AIC e MDL



Confronto tra AIC e MDL

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

Notiamo che la differenza consiste solo nei termini 2 e $\ln[N]$. Quindi, se $\ln[N] > 2$ (ovvero se abbiamo più di 8 dati), la formula **MDL suggerisce di usare modelli più parsimoniosi**

Nota: 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 **sovrestimare 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**



Riassunto della validazione

- Se disponiamo di **molti dati**, il modo migliore per valutare le performance e selezionare il modello è dividere il **dataset in 3 parti** (training, validazione e test)
- Altrimenti, usiamo **cross-validation**
- Se i dati sono **davvero pochi**, possiamo utilizzare formule per la scelta della complessità del modello ottimale che utilizzano **solo il training set**:
 - ✓ Akaike Information Criterion (**AIC**), Minimum Description Length (**MDL**)

Osservazione: le tecniche di validazione che abbiamo visto si possono usare anche nel caso di identificazione di **modelli dinamici**. Però, in questo caso validazione non può contenere dati estratti randomicamente, altrimenti romperei la cronologia temporale dei dati. Per cui, dovrò scegliere dati di validazione **cronologicamente contigui** ai dati di identificazione



Outline

1. Introduzione al machine learning e alla data science
2. Problemi supervisionati e non supervisionati
3. Feasibility of learning
4. Bias-variance tradeoff
5. Learning curves
6. Overfitting
7. Regolarizzazione
8. Validazione, cross-validation e formule di complessità ottima
- 9. Esercizi con codice**



Regressione lineare con regolarizzazione L_2

Consideriamo il modello di **regressione lineare**, con un termine di regolarizzazione L_2 (**Ridge regression**)

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

Si può dimostrare che la stima in forma chiusa dei parametri si ottiene come:

$$\widehat{\boldsymbol{\theta}}_{\text{reg}} = \left(X^\top X + \lambda_{\text{reg}} \cdot I_d \right)^{-1} X^\top Y$$



Ridge regression e gradient descent

Supponiamo di avere **solo 2 parametri** $\theta = [\theta_0, \theta_1]^\top \in \mathbb{R}^{2 \times 1}$ per semplicità

$$E_{\text{aug}}(\theta) \equiv J(\theta) = \frac{1}{N} \sum_{i=1}^N (y(i) - \theta_0 - \theta_1 \cdot \varphi_1(i))^2 + \lambda_{\text{reg}} \cdot \sum_{j=0}^{d-1} (\theta_j)^2$$

E' possibile implementare l'algoritmo del gradient descent come:

For {

$$\theta_0 = \theta_0 - \alpha \cdot 2 \left[\frac{1}{N} \sum_{i=1}^N (y(i) - \theta_0 - \theta_1 \cdot \varphi_1(i)) \cdot (-1) + \lambda_{\text{reg}} \cdot \theta_0 \right]$$

$$\theta_1 = \theta_1 - \alpha \cdot 2 \left[\frac{1}{N} \sum_{i=1}^N (y(i) - \theta_0 - \theta_1 \cdot \varphi_1(i)) \cdot (-\varphi_1(i)) + \lambda_{\text{reg}} \cdot \theta_1 \right]$$

}



Regressione logistica con regolarizzazione L_2

Consideriamo il modello di **regressione logistica**, con un termine di regolarizzazione L_2

$$E_{\text{aug}}(\boldsymbol{\theta}) \equiv J(\boldsymbol{\theta}) = \sum_{i=1}^N \left(y(i) \cdot \ln \pi(i; \boldsymbol{\theta}) + (1 - y(i)) \cdot \ln[1 - \pi(i; \boldsymbol{\theta})] \right) + \lambda_{\text{reg}} \cdot \sum_{j=0}^{d-1} (\theta_j)^2$$

E' possibile implementare l'algoritmo del gradient descent come:

For {

$$\theta_0 = \theta_0 - \alpha \cdot \sum_{i=1}^N 1 \cdot (\pi(i) - y(i)) + 2\lambda_{\text{reg}} \cdot \theta_0$$

⋮

$$\theta_{d-1} = \theta_{d-1} - \alpha \cdot \sum_{i=1}^N \varphi_{d-1}(i) \cdot (\pi(i) - y(i)) + 2\lambda_{\text{reg}} \cdot \theta_{d-1}$$

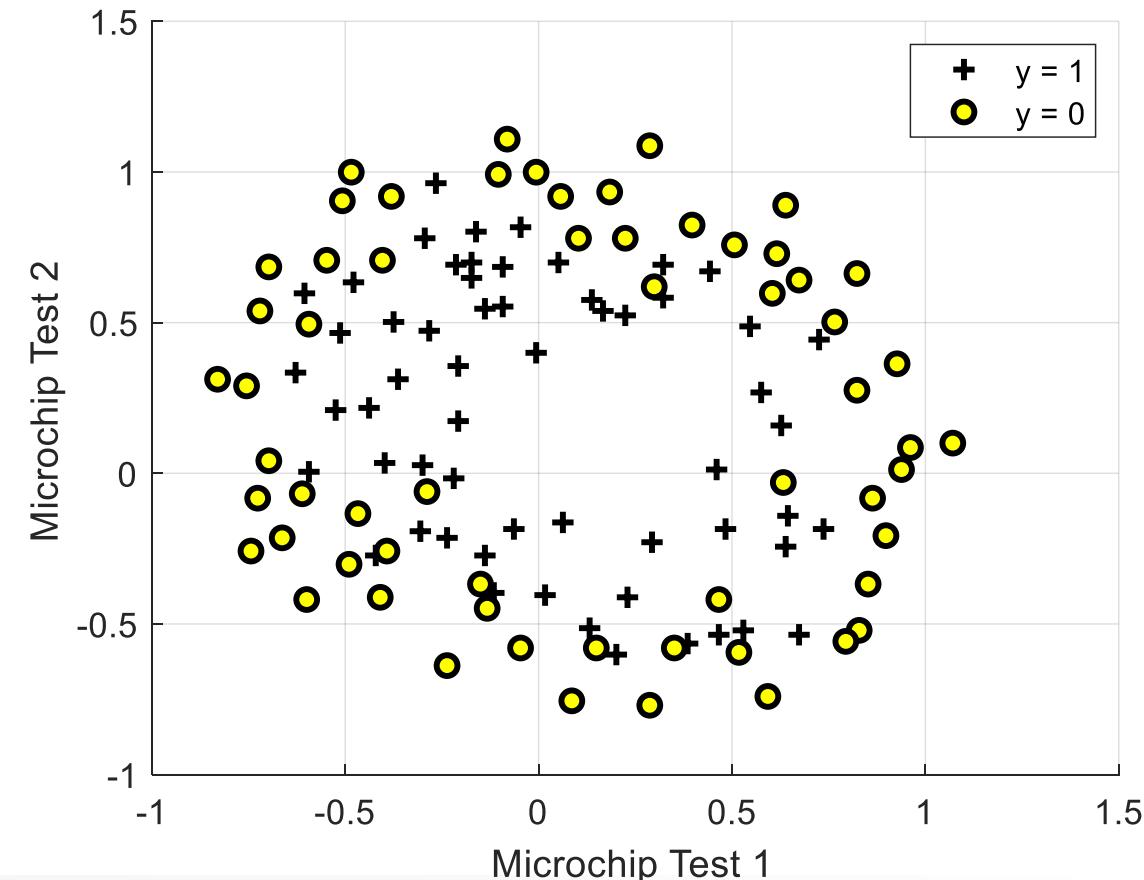
$$\begin{aligned} \pi(i) &\equiv \frac{1}{1 + e^{-\boldsymbol{\varphi}^\top(i)\boldsymbol{\theta}}} \\ &= P(y(i) = 1 | \boldsymbol{\varphi}(i)) \end{aligned}$$



Esercizio: classificare microchips difettosi

Vogliamo rilevare se un microchip è **difettoso** in base ai risultati di due test di qualità alla fine della linea di produzione, tramite un modello di **regressione logistica**

- Ogni microchip è descritto dalle seguenti features
 - ✓ φ_1 : Risultato del test 1
 - ✓ φ_2 : Risultato del test 2
- Il dataset è costituito da $N = 118$ microchips



Esercizio: classificare microchips difettosi

Per ottenere un **confine nonlineare** tramite il classificatore lineare, è possibile usare **features polinomiali**. Ad esempio, usando feature polinomiali di grado 2 otteniamo:

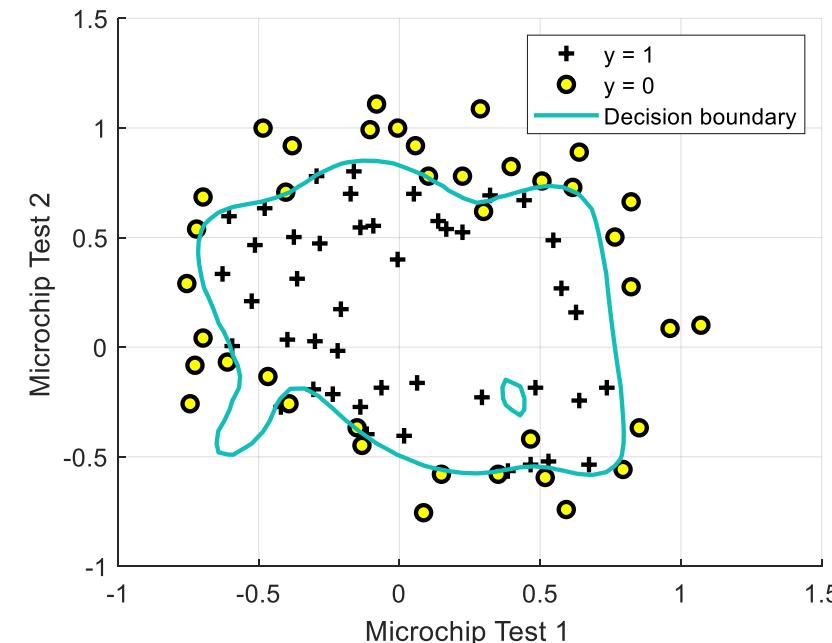
$$\varphi_1^2, \varphi_2^2, \dots, \varphi_{d-1}^2,$$

$$\varphi_1\varphi_2, \dots, \varphi_1\varphi_{d-1},$$

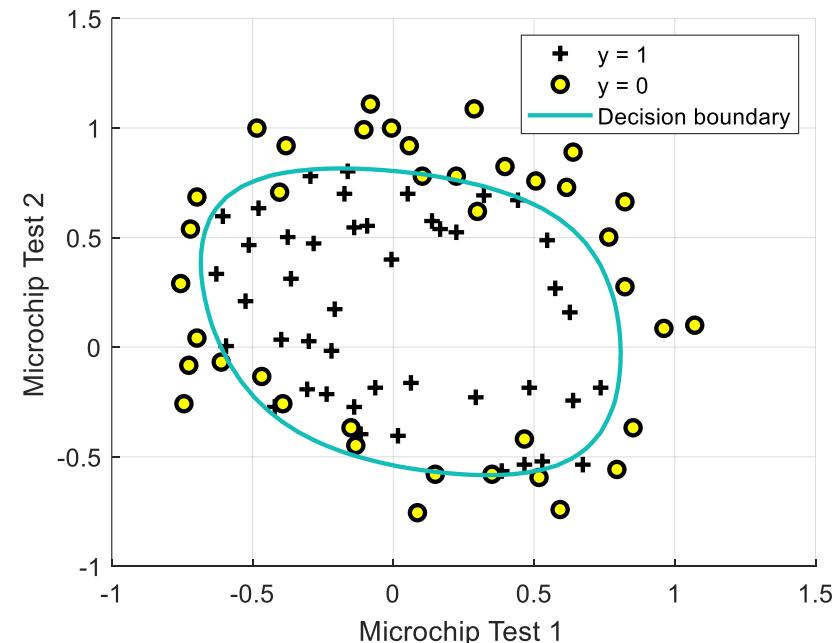
$$\varphi_1\varphi_2^2, \dots, \varphi_1\varphi_{d-1}^2,$$

$$\varphi_1^2\varphi_2, \dots, \varphi_1^2\varphi_{d-1}, \dots$$

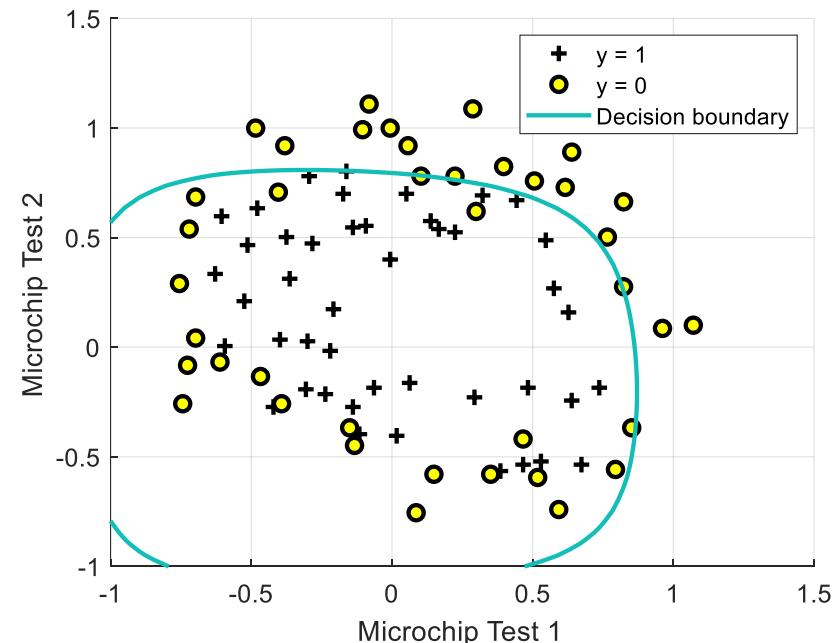
$$\lambda_{\text{reg}} = 0$$



$$\lambda_{\text{reg}} = 0.1$$

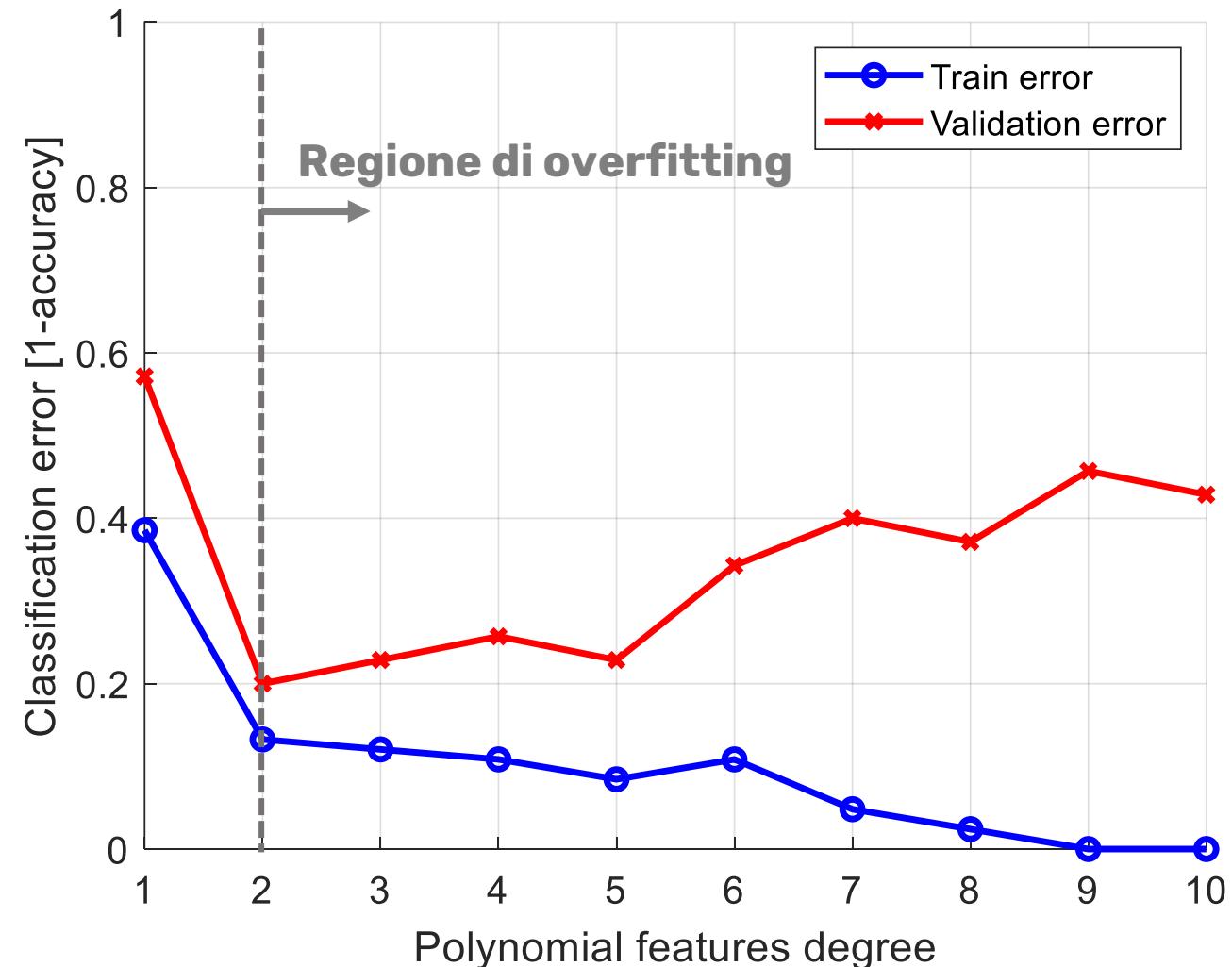


$$\lambda_{\text{reg}} = 10$$



Esercizio: classificare microchips difettosi

È possibile dividere i dati in training e validation set, in modo da **selezionare l'ordine ottimale** per le features polinomiali tramite **validazione**





**UNIVERSITÀ
DEGLI STUDI
DI BERGAMO**

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



IDENTIFICAZIONE DEI MODELLI E ANALISI DEI DATI (IMAD)

Lezione 7: Fondamenti di stima Bayesiana

Corso di Laurea Magistrale in
INGEGNERIA INFORMATICA

SPEAKER
Prof. Mirko Mazzoleni

PLACE
Università degli Studi di
Bergamo

Syllabus

Parte I: sistemi statici

1. Richiami di statistica

2.1 Proprietà degli stimatori

3. Stima a minimi quadrati

3.1 Stima di modelli lineari

3.2 Algoritmo del gradient descent

4. Stima a massima verosimiglianza

4.1 Proprietà della stima

4.2 Stima di modelli lineari

5. Regressione logistica

5.1 Stima di un modello di regressione logistica

6. Fondamenti di machine learning

6.1 Bias-Variance tradeoff

6.2 Overfitting

6.3 Regolarizzazione

6.4 Validazione

7. Cenni di stima Bayesiana

7.1 Probabilità congiunte, marginali e condizionate

7.2 Connessione con Filtro di Kalman



Parte I: sistemi staticiStima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

- ✓ Stima parametri popolazione
- ✓ Stima modello lineare: minimi quadrati

○ **SI assunzioni su ddp dei dati**

- ✓ Stima massima verosimiglianza parametri popolazione
- ✓ Stima modello lineare: massiva verosimiglianza
- ✓ Regressione logistica

• θ variabile casuale○ **SI assunzioni su ddp dei dati**

- ✓ Stima Bayesiana

Parte II: sistemi dinamiciStima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

- ✓ Modelli lineari di pss
- ✓ Predizione
- ✓ Identificazione
- ✓ Persistente eccitazione
- ✓ Analisi asintotica metodi PEM
- ✓ Analisi incertezza stima (numero dati finito)
- ✓ Valutazione del modello

Machine learning

Outline

1. Probabilità congiunte, condizionate, marginali
2. Introduzione alla stima Bayesiana
3. Stima ottima
4. Stima ottima lineare



Outline

1. Probabilità congiunte, condizionate, marginali

2. Introduzione alla stima Bayesiana

3. Stima ottima

4. Stima ottima lineare



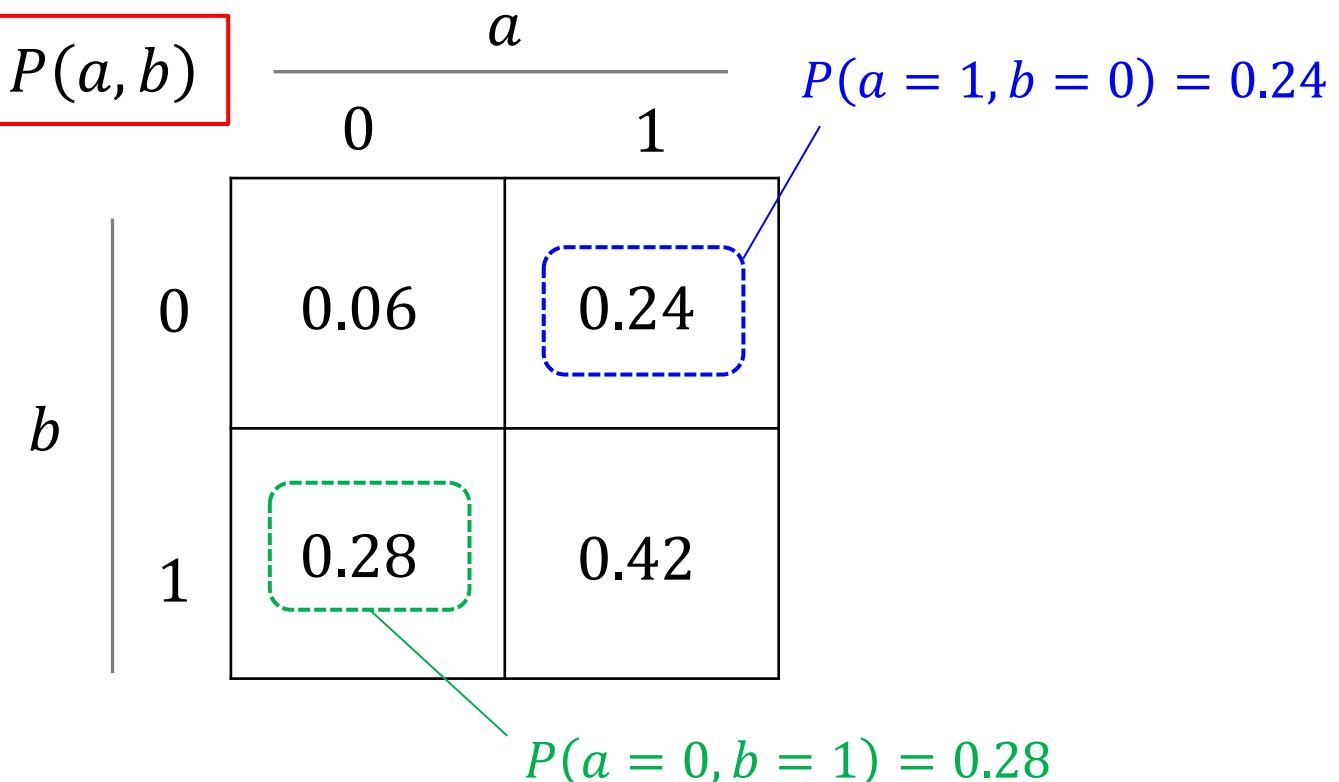
UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

Probabilità congiunte, marginali, condizionate

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 assumano un valore specifico

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

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



Probabilità congiunte, marginali, condizionate

Distribuzione di probabilità marginale

La **distribuzione marginale** è la distribuzione di probabilità di un **sottoinsieme** di variabili casuali

Nel nostro esempio, siccome abbiamo **due variabili casuali** a e b , avremo **due marginali**, ovvero $P(a)$ e $P(b)$. Se avessimo tre v.c discrete a, b, c avremmo le marginali $P(a), P(b), P(c), P(a, b), P(a, c), P(b, c)$

Nel caso di v.c. discrete, la distribuzione marginale è ottenuta «marginando» (ovvero, **sommmando**) rispetto alle variabili che **non sono di interesse**. Nel caso di v.c. continue, si deve integrare anziché sommare



Probabilità congiunte, marginali, condizionate

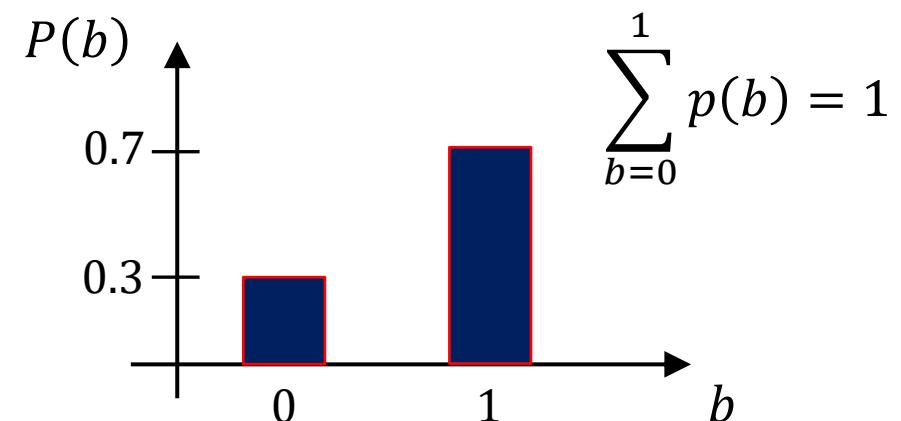
Proviamo a calcolare la distribuzione marginale $P(b)$ partendo dalla distribuzione congiunta $P(a, b)$

		a	
		0	1
b	0	0.06	0.24
	1	0.28	0.42

Non mi interessa che valore
abbia a , l'importante è che $b = 0$

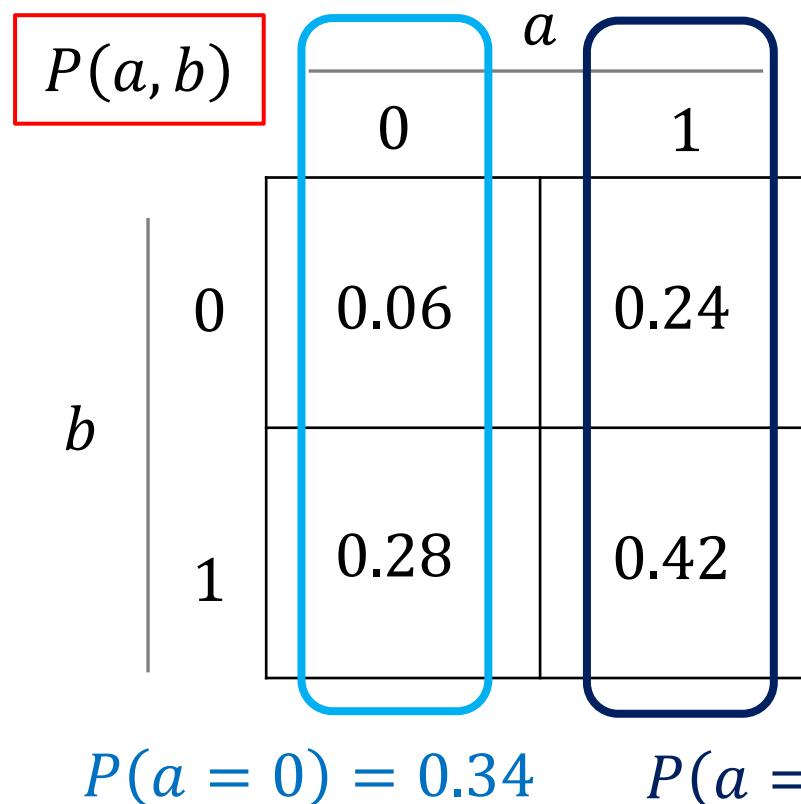
$$P(b = 0) = P(a = 0, b = 0) + P(a = 1, b = 0) = 0.3$$

$$P(b = 1) = 0.7$$



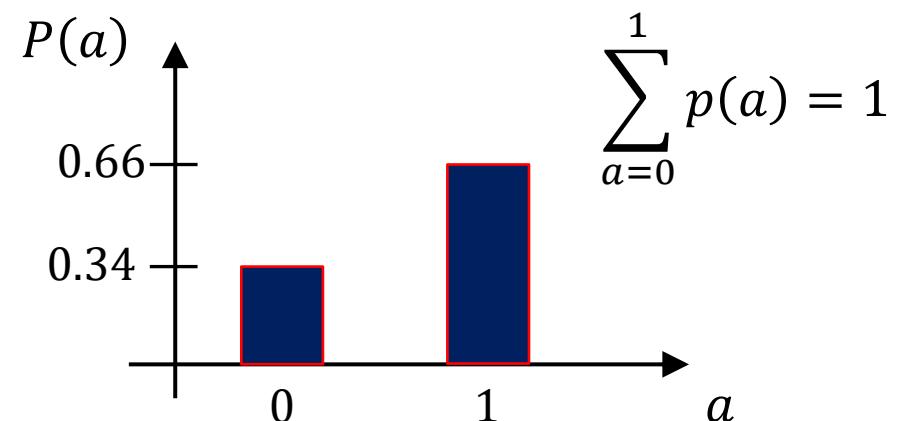
Probabilità congiunte, marginali, condizionate

Proviamo a calcolare la distribuzione marginale $P(a)$ partendo dalla distribuzione congiunta $P(a, b)$



$$P(a = 0) = P(a = 0, b = 0) + P(a = 0, b = 1) = 0.34$$

$$P(a = 1) = P(a = 1, b = 0) + P(a = 1, b = 1) = 0.66$$



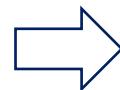
Probabilità congiunte, marginali, condizionate

Distribuzione di probabilità condizionata

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

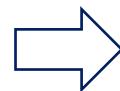
Esempio: siano date N persone, dove N_A è il numero di persone con capelli lunghi e N_B è il numero di persone che ascoltano i Black Sabbath. Definiamo gli eventi A e B come:

A : persone con capelli lunghi



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

B : persone che ascoltano i
Black Sabbath



$$P(B) = \frac{N_B}{N} = \frac{\# \text{ persone che ascoltano i Black Sabbath}}{\# \text{ totale di persone}}$$



Probabilità congiunte, marginali, condizionate

Consideriamo **solo la popolazione che ascolta i Black Sabbath**, con $N_B < N$ persone

La probabilità che una persona **scelta a caso da questa popolazione abbia i capelli lunghi** è

$$P(A|B) = \frac{N_{AB}}{N_B} = \frac{\text{\# persone con capelli lunghi e che ascoltano i Sabbath}}{\text{\# persone che ascoltano i Sabbath}}$$

Abbiamo ristretto la popolazione da N a N_B , e quindi la **probabilità si è ridistribuita**. Prima avevamo $P(A)$, adesso abbiamo $P(A|B)$

$P(A|B)$ è chiamata **probabilità condizionata** (condizionata al fatto che le persone ascoltino i Black Sabbath)



Probabilità congiunte, marginali, condizionate

La probabilità di selezionare una persona con capelli lunghi che ascolti **anche** i Black Sabbath è la **probabilità congiunta** $P(A, B)$

$$P(A, B) = \frac{N_{AB}}{N} = \frac{\# \text{ persone con capelli lunghi e che ascoltano i Sabbath}}{\# \text{ totale di persone}}$$

Posso quindi esprimere $P(A|B)$ come

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

$P(B)$ è una marginale. E' la probabilità che una persona ascolti i Black Sabbath, indipendentemente dalla lunghezza dei capelli



Probabilità congiunte, marginali, condizionate

Dall'esempio precedente abbiamo visto che

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

Osservazioni

- La probabilità che accada sia A che B è la probabilità che si verifichi B per la probabilità che si verifichi A dato che B si è verificato. **Attenzione:** non c'è per forza una causalità temporale
- $P(A, B) = P(A)P(B)$ solo se $P(A|B) = P(A)$. Questo vuol dire che A e B sono eventi **indipendenti**, ovvero il verificarsi di B non modifica le probabilità di verificarsi di A



Teorema di Bayes

Esempio:

A : lancio un dado ed esce «4»



Anche se la moneta fosse uscita «CROCE», il dado ha la stessa probabilità di risultare in un «4»

B : lancio una moneta ed esce «TESTA»

Sappiamo che $P(A, B) = P(B, A)$. Inoltre $P(B, A) = P(B|A)P(A)$, e di conseguenza

|

|

$$P(A|B)P(B) = P(B|A)P(A)$$



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

TEOREMA DI BAYES



Teorema di Bayes

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

Osservazioni

- Il teorema di Bayes permette di **ridistribuire la probabilità**: prima conoscevamo $P(A)$, adesso conosco $P(A|B)$. La probabilità di A è cambiata in seguito all'informazione portata da B
- La distribuzione marginale $P(B) = \sum_A P(A, B) = \sum_A P(A|B)P(B)$ appare come un fattore di normalizzazione

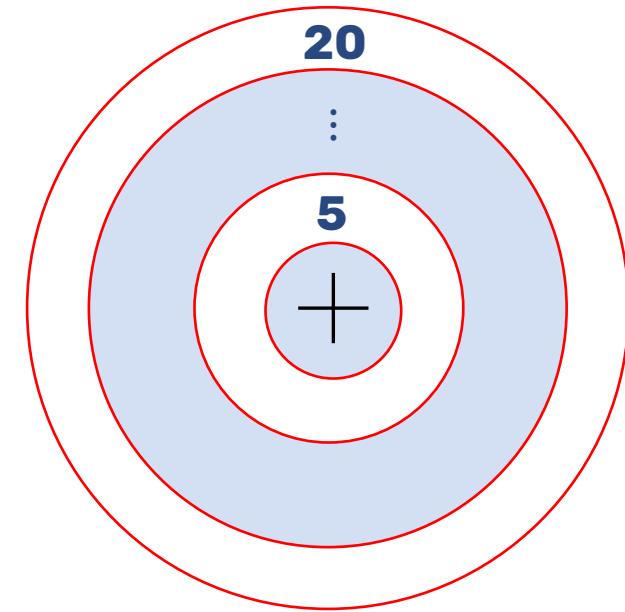


Esempio: probabilità condizionata come ridistribuzione

Consideriamo un bersaglio da frecce con 20 cerchi.

Supponiamo che un lanciatore abbia uguale probabilità di prendere ognuno dei 20 cerchi. **Qual è la probabilità che colpisca il cerchio #5?**

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



Dopo un lancio, un amico gli dice che **non ha preso il cerchio #7**. Qual è ora la probabilità che abbia preso il cerchio #5?



Esempio: probabilità condizionata come ridistribuzione

Dato che sicuramente non ha preso il #7, la probabilità di aver preso il #5 è

$$P(\#5 | \text{NOT } \#7) = \frac{1}{19}$$

poiché, dopo, l'esclusione del cerchio #7, rimangono solo 19 cerchi «prendibili»

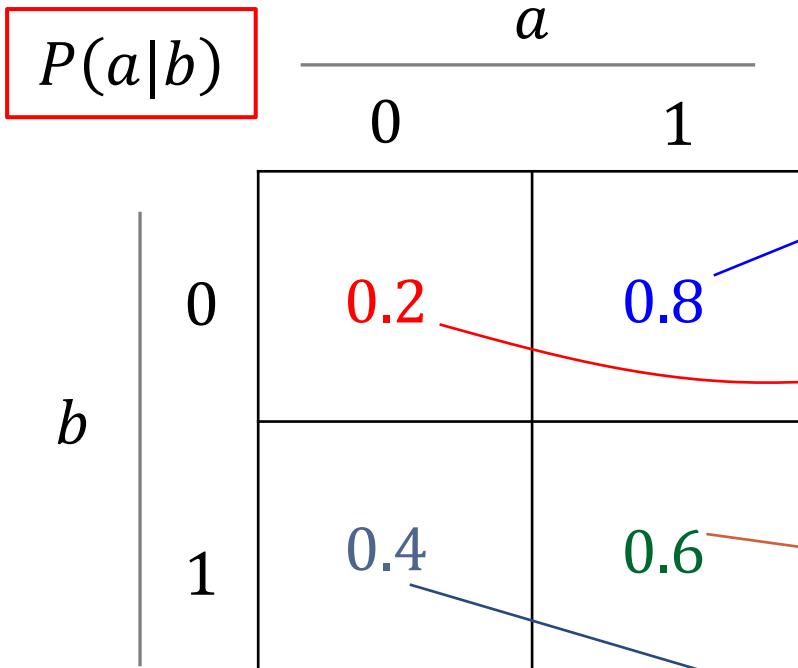
Il condizionamento a «NOT #7» significa che certi «stati» sono ora **inaccessibili**, e di conseguenza la probabilità si deve **ridistribuire** su quelli accessibili

$$P(\#5 | \text{NOT } \#7) = \frac{P(\#5, \text{NOT } \#7)}{P(\text{NOT } \#7)} = \frac{P(\#5) \cdot P(\text{NOT } \#7 | \#5)}{P(\text{NOT } \#7)} = \frac{\frac{1}{20} \cdot 1}{\frac{19}{20}} = \boxed{\frac{1}{19}}$$



Probabilità congiunte, marginali, condizionate

Riprendiamo l'esempio iniziale e proviamo a calcolare la distribuzione $P(a|b)$



$$P(a = 1|b = 0) = \frac{P(a = 1, b = 0)}{p(b = 0)} = \frac{0.24}{0.3} = 0.8$$

$$P(a = 0|b = 0) = \frac{P(a = 0, b = 0)}{p(b = 0)} = \frac{0.06}{0.3} = 0.2$$

$$P(a = 1|b = 1) = \frac{P(a = 1, b = 1)}{p(b = 1)} = \frac{0.42}{0.7} = 0.6$$

$$P(a = 0|b = 1) = \frac{P(a = 0, b = 1)}{p(b = 1)} = \frac{0.28}{0.7} = 0.4$$



Outline

1. Probabilità congiunte, condizionate, marginali

2. Introduzione alla stima Bayesiana

3. Stima ottima

4. Stima ottima lineare



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

Introduzione alla stima Bayesiana

Abbiamo finora considerato il vettore di parametri ignoto $\theta \in \mathbb{R}^{d \times 1}$ come una **variabile deterministica**. Spesso però, ancora prima di collezionare i dati, abbiamo delle **informazioni** (o supposizioni) sui possibili valori che potrebbe assumere θ

Esempi:

1. Stima della concentrazione di una sostanza nell'aria: si ha un'idea dell'ordine di grandezza, per esempio in base a studi precedenti
2. Stima della probabilità che una moneta risulti «TESTA» dopo un lancio: so già che il valore sarà intorno a 0.5, se suppongo non sia truccata

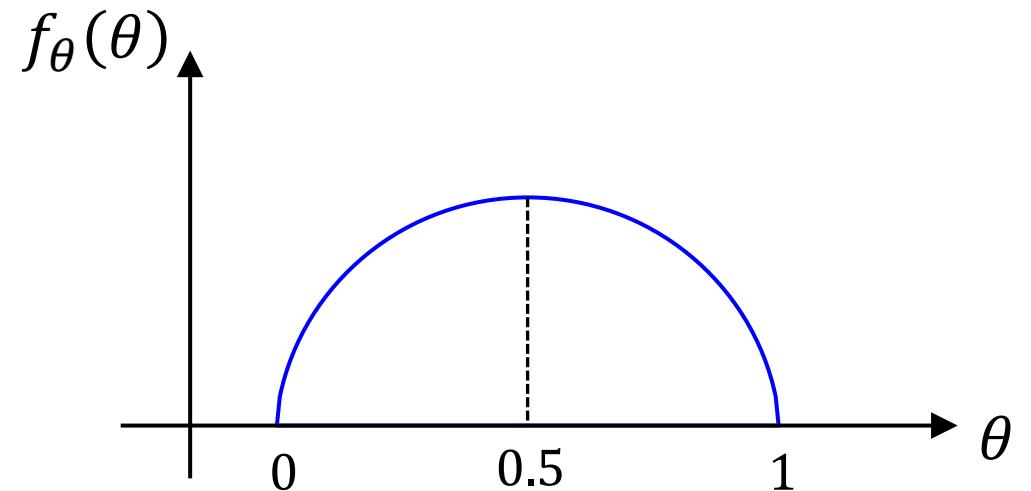


Introduzione alla stima Bayesiana

Ha quindi senso considerare θ come una **variabile casuale**: in questo modo, posso specificare una distribuzione di probabilità per θ , per **descriverne i valori** (e la probabilità che θ li assuma) che **io credo** che possa assumere

- assegno **maggior probabilità** ai valori che **io credo** siano più probabili che θ **possa assumere**, e minor probabilità ai valori che **io credo** non possa assumere

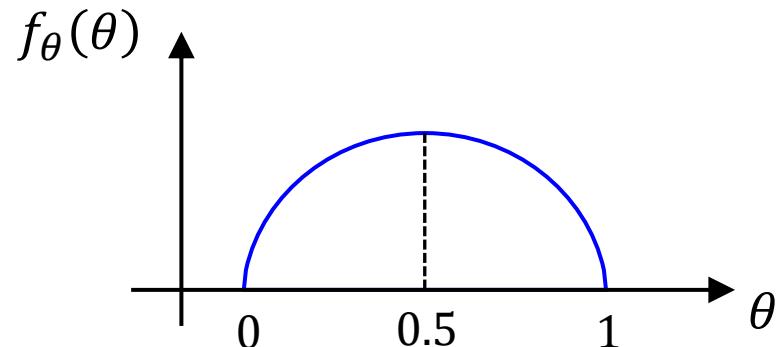
Esempio: Sia θ la probabilità che una moneta risulta in «TESTA». Una possibile distribuzione (continua) $f_\theta(\theta)$ per θ , se suppongo che la moneta non sia truccata, è:



Introduzione alla stima Bayesiana

Osservazioni

- $f_\theta(\theta)$ ha dominio $[0,1]$ poiché θ , modellando una probabilità, deve stare tra 0 e 1
- Siccome suppongo che la moneta non è truccata, $\theta = 0.5$ sarà il valore che io suppongo sia più probabile, mentre $\theta \approx 0$ o $\theta \approx 1$ saranno poco probabili
- Data $f_\theta(\theta)$, abbiamo già una stima del valore di θ ancora prima di aver osservato i dati **(STIMA A-PRIORI)**. Ad esempio (ma non per forza) posso prendere come **valore puntuale** per la stima di θ il suo valore atteso. **L'incertezza sulla stima** sarà allora la varianza di θ **(INCERTEZZA A-PRIORI)**



Introduzione alla stima Bayesiana

Con l'osservazione dei dati, ci si aspetta che:

1. La stima puntuale di θ **cambi**
2. L'incertezza sulla stima **decresca** (ho più informazioni!)

Abbiamo quindi due elementi che portano informazione:

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

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



Distribuzione a-posteriori

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

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

Osservazioni

- $f_{\theta|Y}(\theta|Y)$ è una **distribuzione a-posteriori di possibili valori** di θ . Le probabilità di questi valori, rispetto a $f_{\theta}(\theta)$, sono state **riallocate** dall'aver osservato i dati Y
- Nel caso in cui $f_{Y|\theta}(Y|\theta)$ e $f_{\theta}(\theta)$ sono pdf continue, allora $f_Y(Y) = \int_{-\infty}^{+\infty} f_{Y|\theta}(Y|\theta) f_{\theta}(\theta) d\theta$



Distribuzione a-posteriori

Conosciamo la forma funzionale di $f_{\theta}(\theta)$ e $f_{Y|\theta}(Y|\theta)$ poiché derivano dalle nostre assunzioni sui dati Y e sui parametri θ . Posso dire qualcosa su $f_{\theta|Y}(\theta|Y)$?

- In generale, **non posso dire nulla**. Solo in alcuni casi fortunati ho che $f_{\theta}(\theta|Y)$ ha un'espressione analitica nota
- Un altro problema è che $f_Y(Y)$, nel caso di dati intesi come v.c. continue, è un integrale che potremmo **non sapere come risolvere**. In questo caso si usano tecniche numeriche note come **Markov Chain Monte Carlo (MCMC)**
- Un caso fortunato avviene, per esempio ma non solo, se $f_{\theta}(\theta)$ è **Gaussiana** e anche $f_{Y|\theta}(Y|\theta)$ è **Gaussiana**. Allora, anche $f_{\theta|Y}(\theta|Y)$ è **Gaussiana**



Distribuzione a-posteriori

Quando la **posterior** $f_{\theta|Y}(\theta|Y)$ ha la stessa forma della **prior** $f_{\theta}(\theta)$ (e.g. sono entrambe delle Gaussiane) allora la **likelihood** e la **prior** si dicono **coniugate**

Un modo (computazionalmente oneroso ma semplice) per calcolare la posterior $f_{\theta|Y}(\theta|Y)$ è quello di **discretizzare** il range di valori del parametro θ tramite una griglia di valori

- In questo modo valuto $f_{\theta}(\theta)$ e $f_{Y|\theta}(Y|\theta)$ solo in quei valori di θ all'interno della griglia
- Questo metodo va bene se θ consiste di un paio di parametri. Altrimenti, diventa troppo oneroso ed è meglio ricorrere ad MCMC (a meno che non esista un'espressione analitica nota per la posterior)



Esempio: stima probabilità che la moneta esca testa

Vogliamo stimare la probabilità $\theta \equiv \pi$ che la moneta risulti in «TESTA». Supponiamo di lanciare una moneta $N = 10$ volte, e di osservare $N_s = 7$ «TESTA» ($y = 1$) e $N - N_s = 3$ «CROCE» ($y = 0$). I dati \mathcal{D} sono (l'ordine non importa essendo i dati i.i.d. per ipotesi):

$$Y = [1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 0 \ 0 \ 0]^{\top} \\ 10 \times 1$$

Modelliamo i dati come realizzazioni i.i.d. di una v.c. avente distribuzione di Bernoulli:

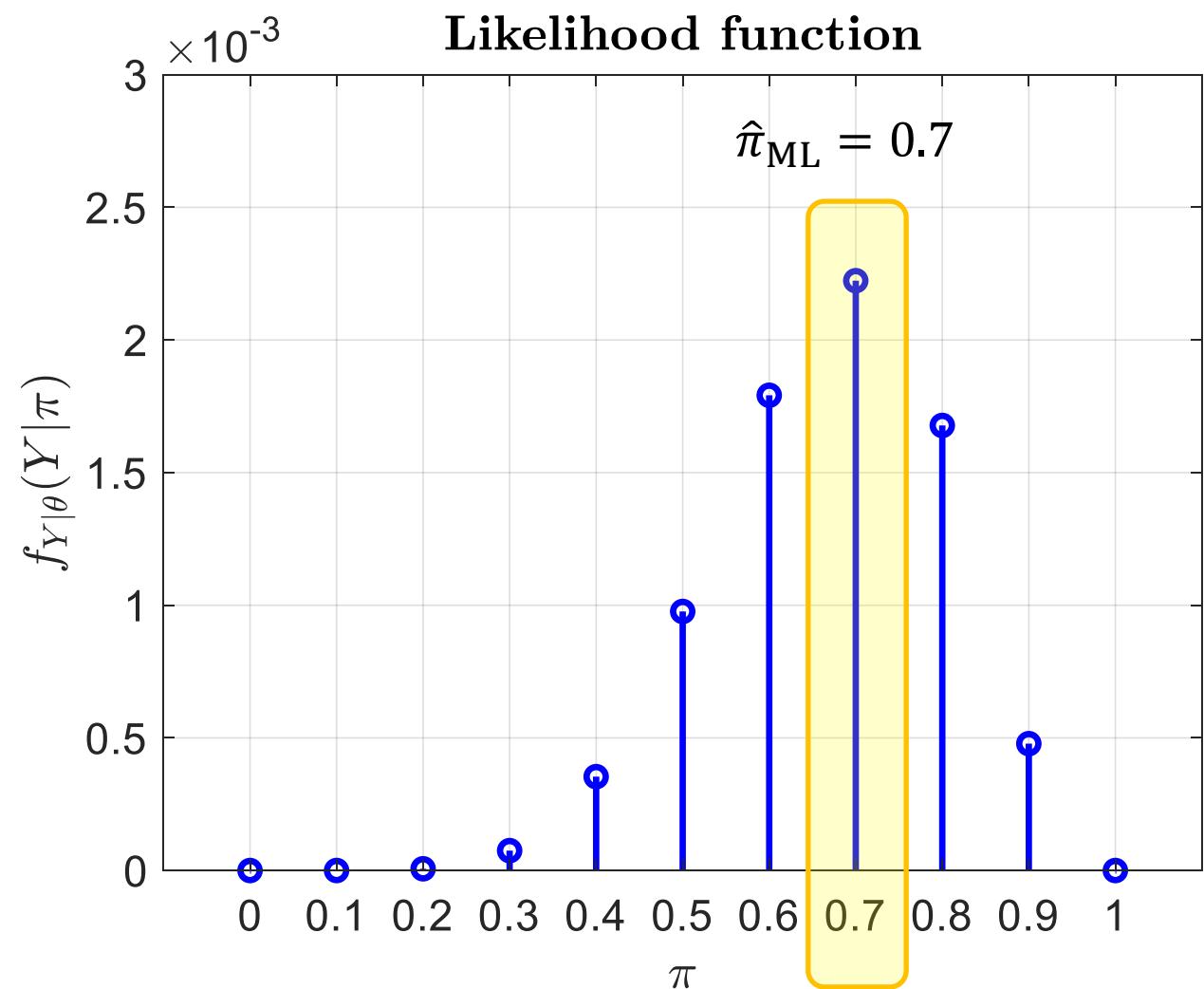
$$y(i) \sim \text{Bernoulli}(\pi), \quad \text{i. i. d.} \quad \implies f_y(y(i)|\pi) = \pi^{y(i)} \cdot (1 - \pi)^{(1-y(i))}$$

Likelihood: $f_{Y|\theta}(Y|\pi) = \prod_{i=1}^N \pi^{y(i)} \cdot (1 - \pi)^{(1-y(i))} = \overbrace{\pi^{\sum_{i=1}^N y(i)}}^{\# \text{ successi}} \cdot (1 - \pi)^{\sum_{i=1}^N 1-y(i)} \overbrace{(1-\pi)^{\sum_{i=1}^N 1-y(i)}}^{\# \text{ insuccessi}}$



Esempio: stima probabilità che la moneta esca testa

Se facessimo una **stima a massima verosimiglianza**, prenderemmo come stima il valore $\hat{\pi}_{ML}$ che **massimizza la verosimiglianza**, ovvero $\hat{\pi}_{ML} = N_s/N = 0.7$

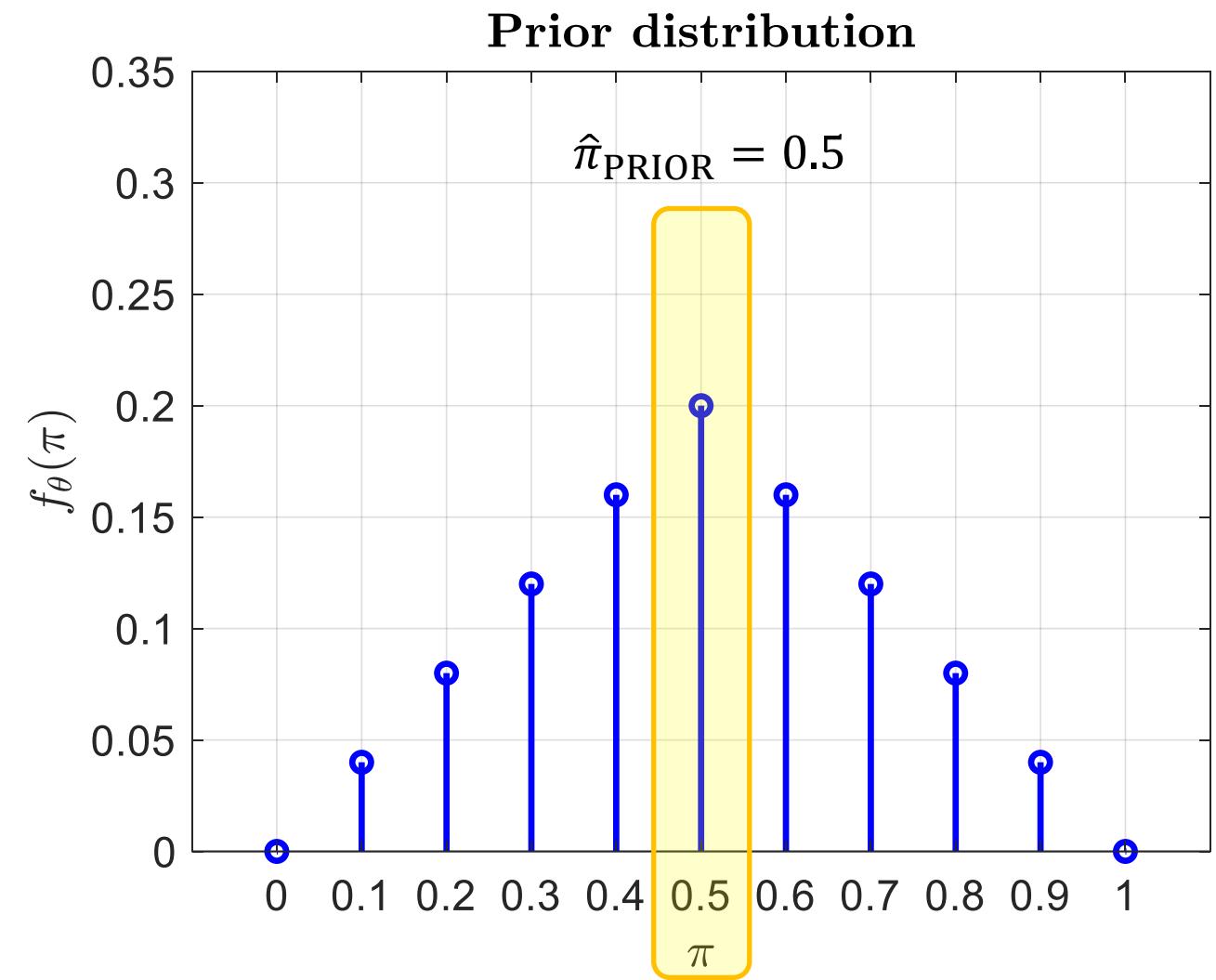


Esempio: stima probabilità che la moneta esca testa

Supponiamo però di avere una **buona confidenza** che la **moneta non sia truccata**. Potremmo esprimere questa nostra informazione a-priori tramite una distribuzione $f_\theta(\pi)$

In questa «rappresentazione della nostra credenza», diamo più probabilità al fatto che $\pi = 0.5$.

Possiamo prendere come stima di π il valore $\hat{\pi}_{\text{PRIOR}} = 0.5$

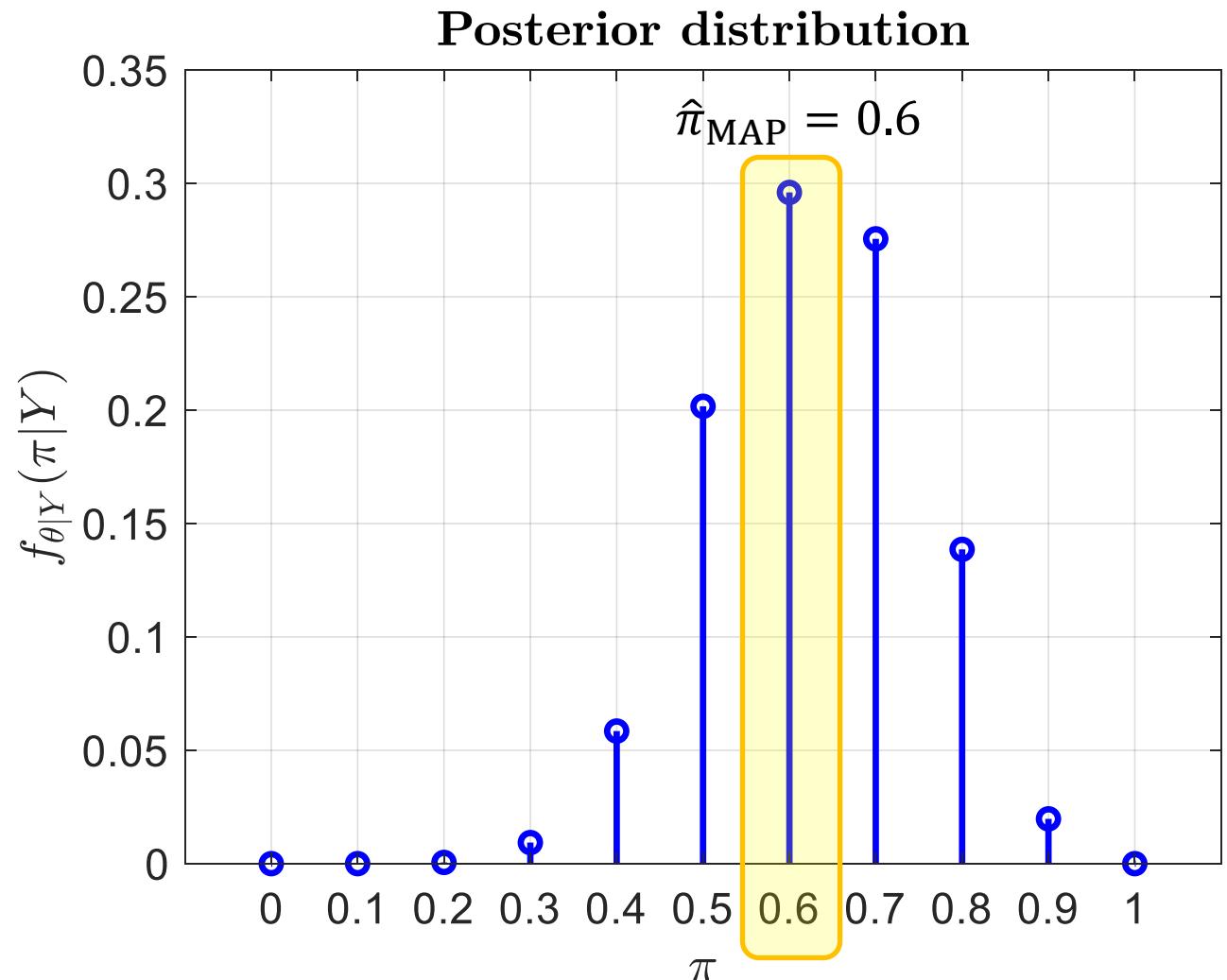


Esempio: stima probabilità che la moneta esca testa

Unendo le informazioni di prior e di likelihood ottengo una distribuzione di valori di π che è un **compromesso** tra la prior e la likelihood

In questo senso, la procedura di stima Bayesiana «**regolarizza**» la stima di π

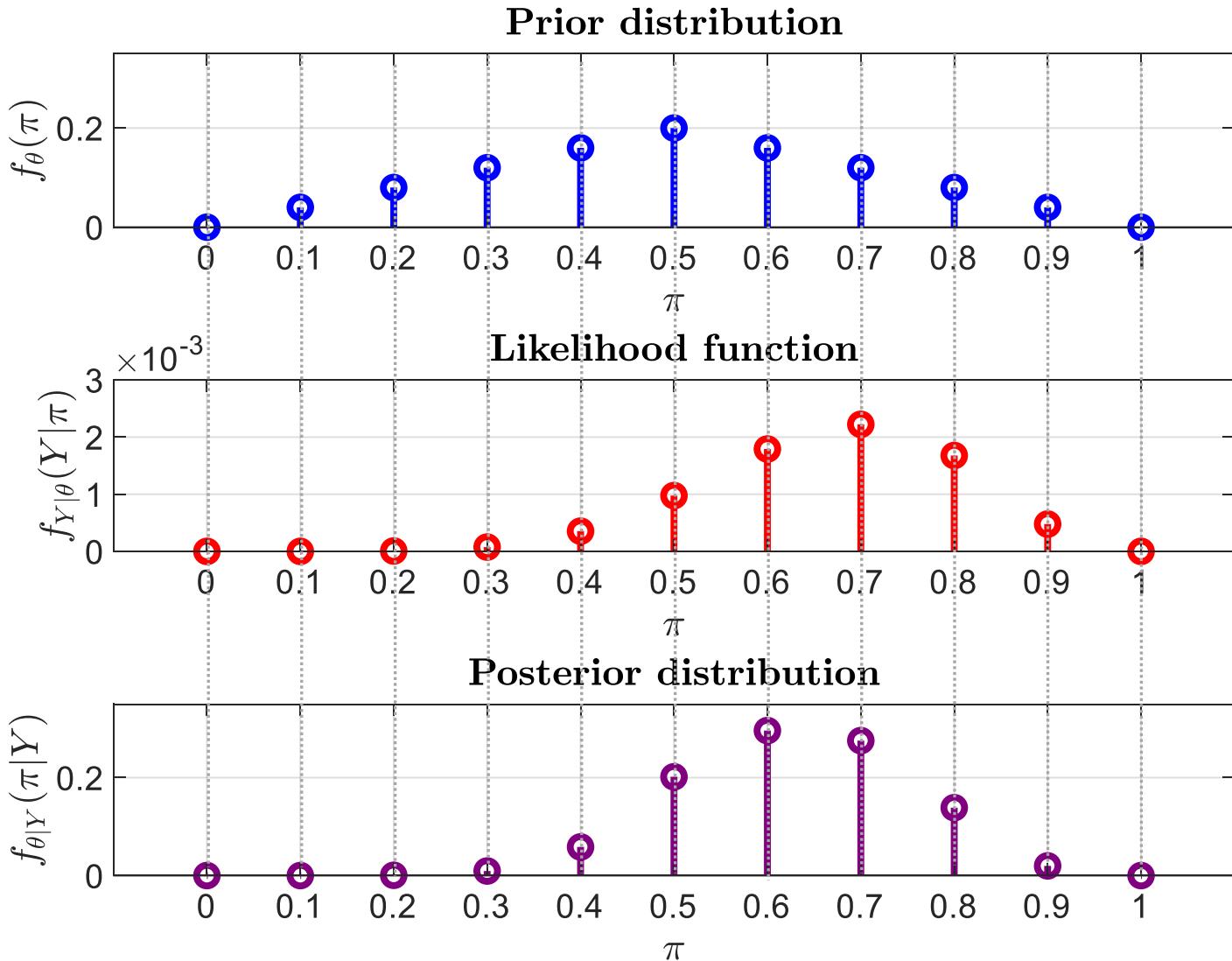
Il valore di $\hat{\pi}_{\text{MAP}}$ che massimizza la posterior è chiamato **stima MAP**
(Maximum A Posteriori)



Esempio: stima probabilità che la moneta esca testa

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

$$f_Y(Y) = \sum_{\pi} f_{Y|\theta}(Y|\pi) \cdot f_{\theta}(\pi)$$
$$= 9.683 \cdot 10^{-4}$$



Outline

1. Probabilità congiunte, condizionate, marginali
2. Introduzione alla stima Bayesiana
- 3. Stima ottima**
4. Stima ottima lineare



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

Stima ottima

Supponiamo di avere la **posterior** $f_{\theta|Y}(\theta|Y)$. Abbiamo quindi una distribuzione di valori dei parametri ignoti θ . Spesso però ci serve un **valore solo, puntuale**. Abbiamo varie scelte:

- **Stima MAP:** $\hat{\theta} = \arg \max_{\theta} f_{\theta|Y}(\theta|Y)$
- **Valore atteso a posteriori:** $\hat{\theta} = \mathbb{E}_{\theta}[f_{\theta|Y}(\theta|Y)] \equiv \mathbb{E}[\theta|Y]$, ovvero il valore atteso della posterior
- **Altre quantità**, come la mediana, ecc...

Ricordiamo che in generale indichiamo uno **stimatore** come una funzione $T(\cdot)$ dei dati \mathcal{D} :

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



Stima ottima

Consideriamo il caso θ **scalare** per semplicità. 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)** già visto in precedenza (si veda Lezione 02)

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

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

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

cioè che **minimizza il MSE**



Stima ottima

Si dimostra che

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

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

Nota

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

$$\text{MSE} \equiv \text{tr} \left\{ \mathbb{E} \left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \right] \right\} = \mathbb{E} \left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \right] = \mathbb{E} \left[\|(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})\|_2^2 \right]$$

$d \times 1$ $1 \times d$ $1 \times d$ $d \times 1$ 1×1



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 Guassiano $\theta \sim \mathcal{N}(0, \lambda_{\theta\theta}^2)$.

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

La loro pdf **congiunta** $f_{y\theta}(y, \theta)$ è ancora **Gaussiana**

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

Al quadrato perché ho 2 variabili



Stima ottima: il caso Gaussiano

La pdf dei dati $f_y(y)$ è:
$$f_y(y) = \frac{1}{\sqrt{2\pi \lambda_{yy}^2}} \exp\left(-\frac{1}{2\lambda_{yy}^2}(y - 0)^2\right)$$

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

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

• **Varianza:** $\lambda_{\theta|y}^2 = \lambda_{\theta\theta}^2 - \frac{\lambda_{\theta y}^2}{\lambda_{yy}^2}$

- Se $\lambda_{\theta y} = 0$, ovvero se y **non porta informazioni** su θ , la stima di θ rimane quella a priori
- Notiamo che $\frac{\lambda_{\theta y}^2}{\lambda_{yy}^2} > 0$. Quindi, **l'incertezza a posteriori è minore** di quella a priori
- Se λ_{yy}^2 è **grande**, la varianza **diminuisce di poco**, perché i dati sono molto incerti



Stima ottima: il caso Gaussiano

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

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



Outline

1. Probabilità congiunte, condizionate, marginali
2. Introduzione alla stima Bayesiana
3. Stima ottima
- 4. Stima ottima lineare**



Stima ottima lineare

Non è sempre detto che y e θ siano congiuntamente Gaussiane. Vogliamo quindi trovare uno stimatore che **non faccia ipotesi sulla ddp congiunta** di y e θ

Supponiamo y e θ due *variabili casuali scalari* con valore atteso nullo e varianza λ_{yy}^2 e $\lambda_{\theta\theta}^2$ rispettivamente

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

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

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



Stima ottima lineare

Per trovare α e β , **minimizziamo la funzione di costo** data dal Mean Square Error

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

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

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



Stima ottima lineare

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

$$\begin{cases} \frac{\partial J(\alpha, \beta)}{\partial \alpha} = 0 \\ \frac{\partial J(\alpha, \beta)}{\partial \beta} = 0 \end{cases} \Rightarrow \boxed{\begin{cases} \alpha = \lambda_{\theta y} / \lambda_{yy}^2 \\ \beta = 0 \end{cases}}$$



Stima ottima lineare

Lo **stimatore lineare ottimo** è quindi dato da

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

Coincide con lo stimatore ottimo di Bayes per il caso Gaussiano!

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

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



Stima ottima lineare

Osservazioni

- Lo stimatore ottimo lineare non **fa nessuna ipotesi su che tipo di distribuzione** hanno y e θ . Assume solo che siano v.c. con una certa media e una certa varianza
- Potrebbe dunque esserci uno **stimatore migliore** (nel senso che ha MSE minore) **rispetto a quello lineare ottimo**
- Se però y e θ sono **congiuntamente Gaussiani**, allora **non esiste nessuno stimatore migliore** di quello lineare ottimo



Stima ottima lineare

Generalizzazione 1: valore atteso non nullo, y e θ scalari

Se: • $\mathbb{E}[y] = \mu_y \neq 0$
• $\mathbb{E}[\theta] = \mu_\theta \neq 0$



$$\hat{\theta}_{\text{opt}}^{\text{lin}} = \mu_\theta + \frac{\lambda_{\theta y}}{\lambda_{yy}^2} \cdot (y - \mu_y)$$

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

Generalizzazione 2: $Y \in \mathbb{R}^{N \times 1}$ e $\theta \in \mathbb{R}^{d \times 1}$ vettoriali

Se: • $\mathbb{E}[Y] = \mu_Y \neq 0$
 $N \times 1$

• $\mathbb{E}[\theta] = \mu_\theta \neq 0$
 $d \times 1$

$$\text{Var} \begin{bmatrix} Y \\ \theta \end{bmatrix} = \begin{bmatrix} \Lambda_{YY} & \lambda_{Y\theta} \\ \Lambda_{\theta Y} & \Lambda_{\theta\theta} \end{bmatrix}$$

$$\hat{\theta}_{\text{opt}}^{\text{lin}} = \mu_\theta + \Lambda_{\theta Y} \cdot \Lambda_{YY}^{-1} \cdot (Y - \mu_Y)$$

$$\text{Var}[\hat{\theta}_{\text{opt}}^{\text{lin}} - \theta] = \Lambda_{\theta\theta} - \Lambda_{\theta Y} \cdot \Lambda_{YY}^{-1} \cdot \Lambda_{Y\theta}$$



Connessione con il Filtro di Kalman

Le formule appena viste ammettono una **forma ricorsiva**: appena arriva un dato osservato nuovo, si aggiorna la stima corrente senza considerare nuovamente tutti i dati

Queste espressioni ricorsive dello stimatore lineare ottimo sono alla base del **Filtro di Kalman**, un algoritmo che ha l'obiettivo di **stimare lo stato $x(t)$ di un sistema dinamico**

- lo stato $x(t)$ e l'uscita $y(t)$ del sistema dinamico lineare sono visti come variabili casuali
- si vuole **stimare lo stato $x(t)$** , visto come l'incognita θ , sulla base dello **stato stimato al tempo precedente (stima a priori)** e sui dati che man mano arrivano dalle **misure dei sensori $y(t)$ (dati osservati)**





**UNIVERSITÀ
DEGLI STUDI
DI BERGAMO**

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



IDENTIFICAZIONE DEI MODELLI E ANALISI DEI DATI (IMAD)

Lezione 8: Processi stocastici

**Corso di Laurea Magistrale in
INGEGNERIA INFORMATICA**

SPEAKER
Prof. Mirko Mazzoleni

PLACE
Università degli Studi di
Bergamo

Syllabus

Parte II: sistemi dinamici

8. Processi stocastici

- 8.1 Processi stocastici stazionari (pss)
- 8.3 Rappresentazione spettrale di un pss
- 8.4 Stimatori campionari media\covarianza
- 8.5 Densità spettrale campionaria

9. Famiglie di modelli a spettro razionale

- 9.1 Modelli per serie temporali (MA, AR, ARMA)
- 9.2 Modelli per sistemi input/output (ARX, ARMAX)

10. Predizione

- 10.1 Filtro passa-tutto

10.2 Forma canonica

10.3 Teorema della fattorizzazione spettrale

10.4 Soluzione al problema della predizione

11. Identificazione

- 11.3 Identificazione di modelli ARX
- 11.4 Identificazione di modelli ARMAX
- 11.5 Metodo di Newton

12. Identificazione: analisi e complementi

- 12.1 Analisi asintotica metodi PEM
- 12.2 Identificabilità dei modelli
- 12.3 Valutazione dell'incertezza di stima

13. Identificazione: valutazione



Parte I: sistemi staticiStima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

✓ Stima parametri popolazione

✓ Stima modello lineare: minimi quadrati

○ **SI assunzioni su ddp dei dati**

✓ Stima massima verosimiglianza parametri popolazione

✓ Stima modello lineare: massiva verosimiglianza

✓ Regressione logistica

• θ variabile casuale○ **SI assunzioni su ddp dei dati**

✓ Stima Bayesiana

Machine learning**Parte II: sistemi dinamici**Stima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

✓ Modelli lineari di pss

✓ Predizione

✓ Identificazione

✓ Persistente eccitazione

✓ Analisi asintotica metodi PEM

✓ Analisi incertezza stima (numero dati finito)

✓ Valutazione del modello

Outline

1. Introduzione alla stima di modelli dinamici
2. Processi stocastici
3. Processi stocastici stazionari
4. Momenti temporali ed ergodicità
5. Trasformata \mathcal{Z} e trasformata di Fourier
6. Densità spettrale di potenza
7. Stima spettrale
8. Sistemi dinamici lineari discreti deterministici
9. Sistemi dinamici lineari discreti stocastici



Outline

- 1. Introduzione alla stima di modelli dinamici**
2. Processi stocastici
3. Processi stocastici stazionari
4. Momenti temporali ed ergodicità
5. Trasformata \mathcal{Z} e trasformata di Fourier
6. Densità spettrale di potenza
7. Stima spettrale
8. Sistemi dinamici lineari discreti deterministici
9. Sistemi dinamici lineari discreti stocastici



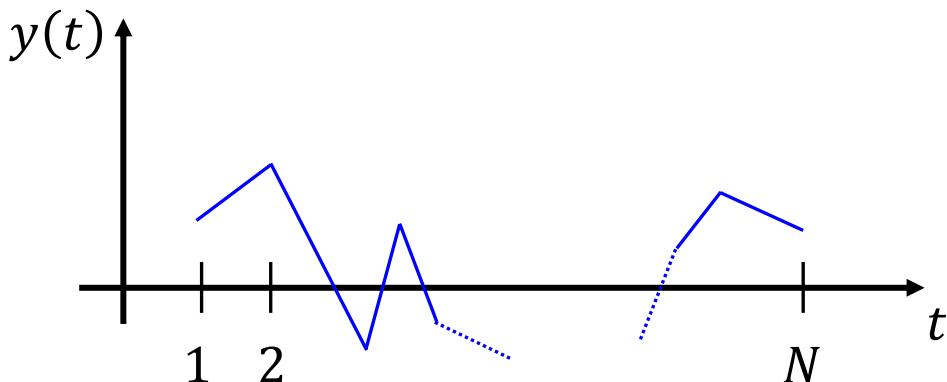
Introduzione alla stima di modelli dinamici

Tratteremo due tipi di problemi, collegati tra loro:

1. Analisi e modellistica di **serie temporali**
2. Analisi e modellistica di **sistemi ingresso\uscita**

SERIE TEMPORALI

Definizione: Una serie temporale (discreta) è un insieme di dati $\mathcal{D} = \{y(1), y(2), \dots, y(N)\}$ indicizzati nel tempo. Indichiamo ogni dato con $y(t)$, dove $t \in \mathbb{Z}$



Esempi:

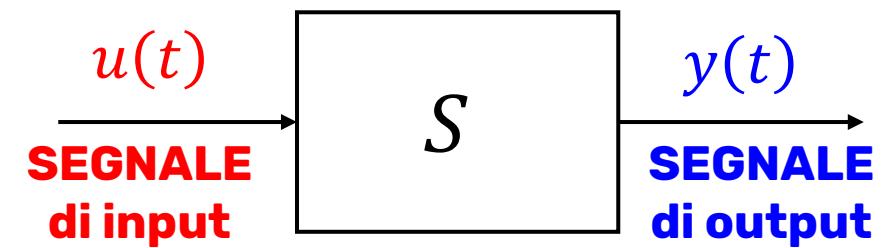
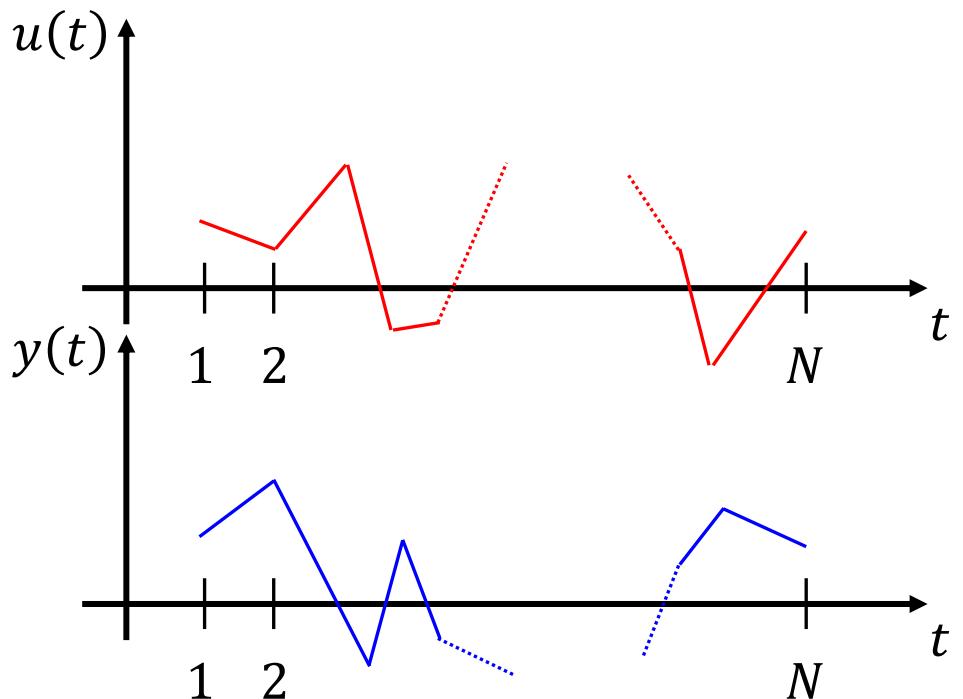
- Valori di un titolo azionario
- Mm di pioggia caduti in una settimana
- Velocità del vento
- Moti ondosi
- ...



Introduzione alla stima di modelli dinamici

SISTEMI INGRESSO\USCITA

I sistemi dinamici processano un segnale di input $u(t)$ per generare un segnale di uscita $y(t)$. Abbiamo dati di input $\{u(1), u(2), \dots, u(N)\}$ e dati di output $\{y(1), y(2), \dots, y(N)\}$



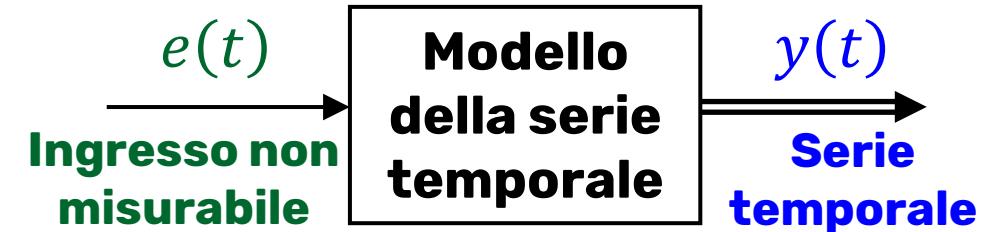
Esempi:

- Sistemi dinamici di varia natura: meccanici, economici, biologici...

Impostazione del problema

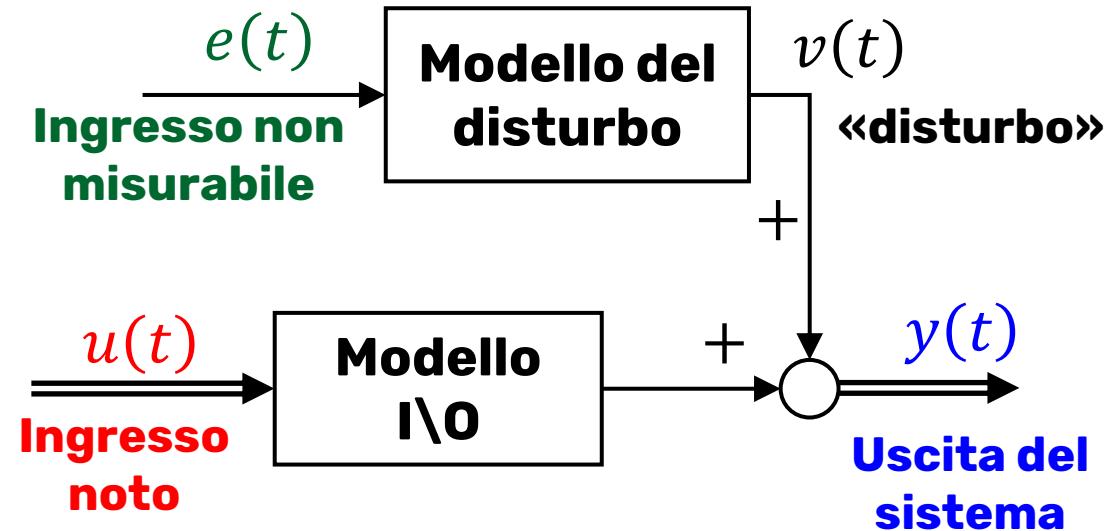
SERIE TEMPORALI

Modelleremo la serie temporale $y(t)$ come l'**uscita di un sistema dinamico** con **ingresso «remoto» non misurabile** $e(t)$



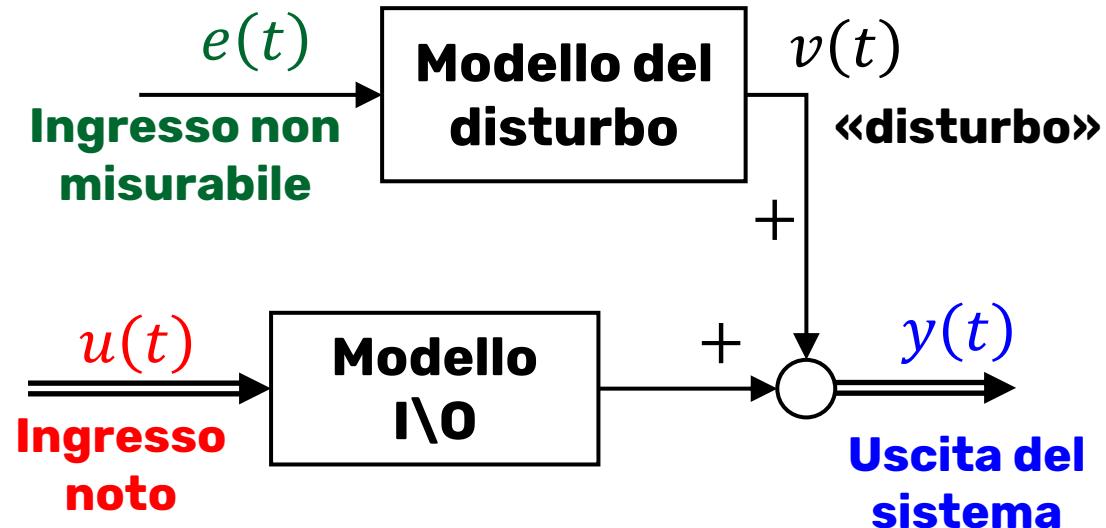
SISTEMI INGRESSO\USCITA

Modelleremo l'uscita $y(t)$ come il contributo di una **componente esogena nota** $u(t)$ ed una **componente di «disturbo»** $v(t)$ **ignota**



Impostazione del problema

I modelli che considereremo saranno modelli di **sistemi dinamici lineari tempo invarianti** (LTI) e **discreti**



Il termine di «disturbo» $v(t)$ è utilizzato per **modellare differenti fenomeni**:

- Rumore di misura
- Disturbi di processo
- Effetto di segnali esogeni di input non misurabili
- Effetti di linearizzazioni del sistema

In sostanza, $v(t)$ modella **tutto ciò che il modello lineare I\O non riesce a spiegare** per quanto riguarda la relazione tra i dati misurati di $u(t)$ e $y(t)$. La cosa difficile è separare l'effetto che $u(t)$ ha su $y(t)$ rispetto a quello che $e(t)$ ha su $y(t)$



Impostazione del problema

Che tipo di problemi vogliamo risolvere?

Sia nel caso di serie temporali che nel caso di sistemi I/O, vogliamo risolvere due problemi:

1. **Predizione** di uscite a istanti futuri $t + k$ in base alle informazioni attualmente a disposizione al tempo t . Indichiamo la predizione con $\hat{y}(t + 1|t)$
2. **Identificazione (stima)** dei modelli descritti, in modo da poter catturare le relazioni tra gli ingressi (noti ed ignoti) e l'uscita del sistema che genera i dati

Osservazioni

- Lavoreremo con **segnali e sistemi a tempo discreto**. I segnali sono campionati con periodo di campionamento T_s . Per semplicità di notazione, indicheremo il **dato al t -esimo istante di campionamento come $y(t)$, intendendo $y(t \cdot T_s)$**



Impostazione del problema

- Assumeremo che uscite $y(t)$ siano affette dal «disturbo» $v(t)$, che può essere visto come un **«rumore» che sporca la vera misura** dell'uscita

Nel caso di **sistemi statici**, per gestire questa incertezza sulla misura dei dati, avevamo interpretato i dati come delle **variabili casuali**

Nel caso di sistemi dinamici, però, i dati non sono indipendenti, ma sono campionati da un segnale che evolve nel tempo. Non abbiamo più osservazioni di v.c. singole, ma osserviamo **una successione di v.c. nel tempo**



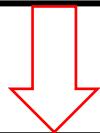
PROCESSI STOCASTICI



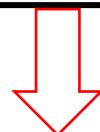
Gli step per la risoluzione del problema

Seguiremo tre fasi per risolvere il problema della **modellazione di sistemi dinamici**:

Definizione delle **classi di modelli** \mathcal{M} di sistemi dinamici



Predizione



Identificazione

Ci concentreremo su modelli di **sistemi dinamici lineari**, espressi da **funzioni di trasferimento razionali fratte**. I parametri ignoti sono i coefficienti dei polinomi al numeratore e denominatore

Data una particolare classe di modello, supponendo di conoscerne il valore dei parametri, qual è il **preditore ottimo**? Quanto vale la predizione ottima?

Come **stimo il valore dei parametri** del modello scelto per la modellazione dei dati?



Outline

1. Introduzione alla stima di modelli dinamici
- 2. Processi stocastici**
3. Processi stocastici stazionari
4. Momenti temporali ed ergodicità
5. Trasformata \mathcal{Z} e trasformata di Fourier
6. Densità spettrale di potenza
7. Stima spettrale
8. Sistemi dinamici lineari discreti deterministici
9. Sistemi dinamici lineari discreti stocastici



Processi stocastici

Definizione: Un **processo stocastico** $v(t, s)$ a *tempo discreto* è una **successione infinita di variabili casuali**, definite a partire dallo **stesso esperimento casuale** s e ordinate secondo un **indice temporale** $t \in \mathbb{N}$

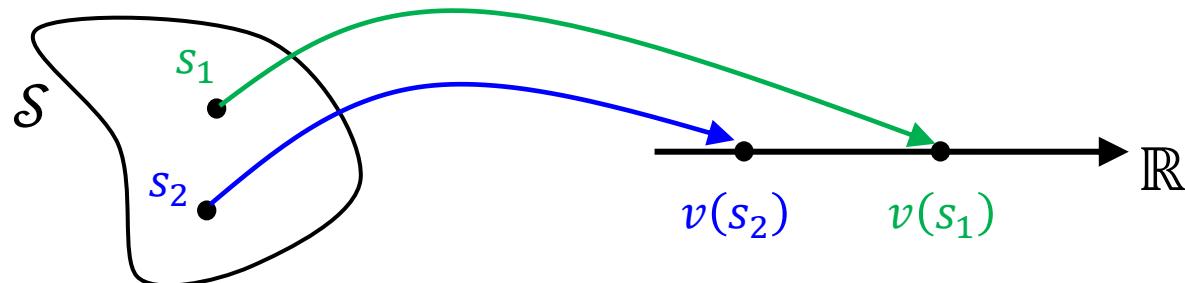
$$v(1, s), v(2, s), \dots, v(N, s) \quad N \in \mathbb{N}$$

- **Fissato un esito** $s = \bar{s}$, si ottiene una **realizzazione** $v(t, \bar{s})$ del processo stocastico, ovvero una serie di valori **deterministici** nel tempo (un segnale)
- **Fissato un istante temporale** $t = \bar{t}$, si ottiene la **variabile casuale** $v(\bar{t}, s)$, ovvero la variabile casuale al tempo \bar{t}
- **Fissati** $s = \bar{s}$ e $t = \bar{t}$, si ottiene un **numero** $v(\bar{t}, \bar{s})$

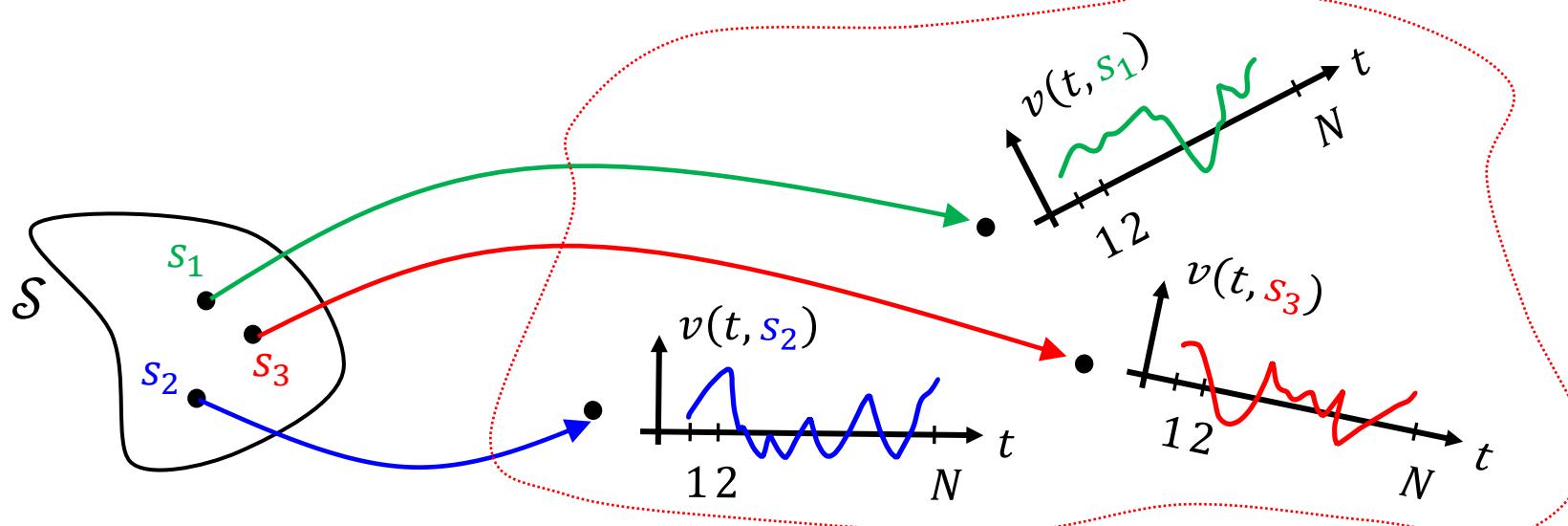


Processi stocastici

Come nel caso delle variabili casuali, è possibile pensare ad un processo stocastico come una **funzione**, che, anziché restituire numeri reali, **restituisce funzioni nel tempo**



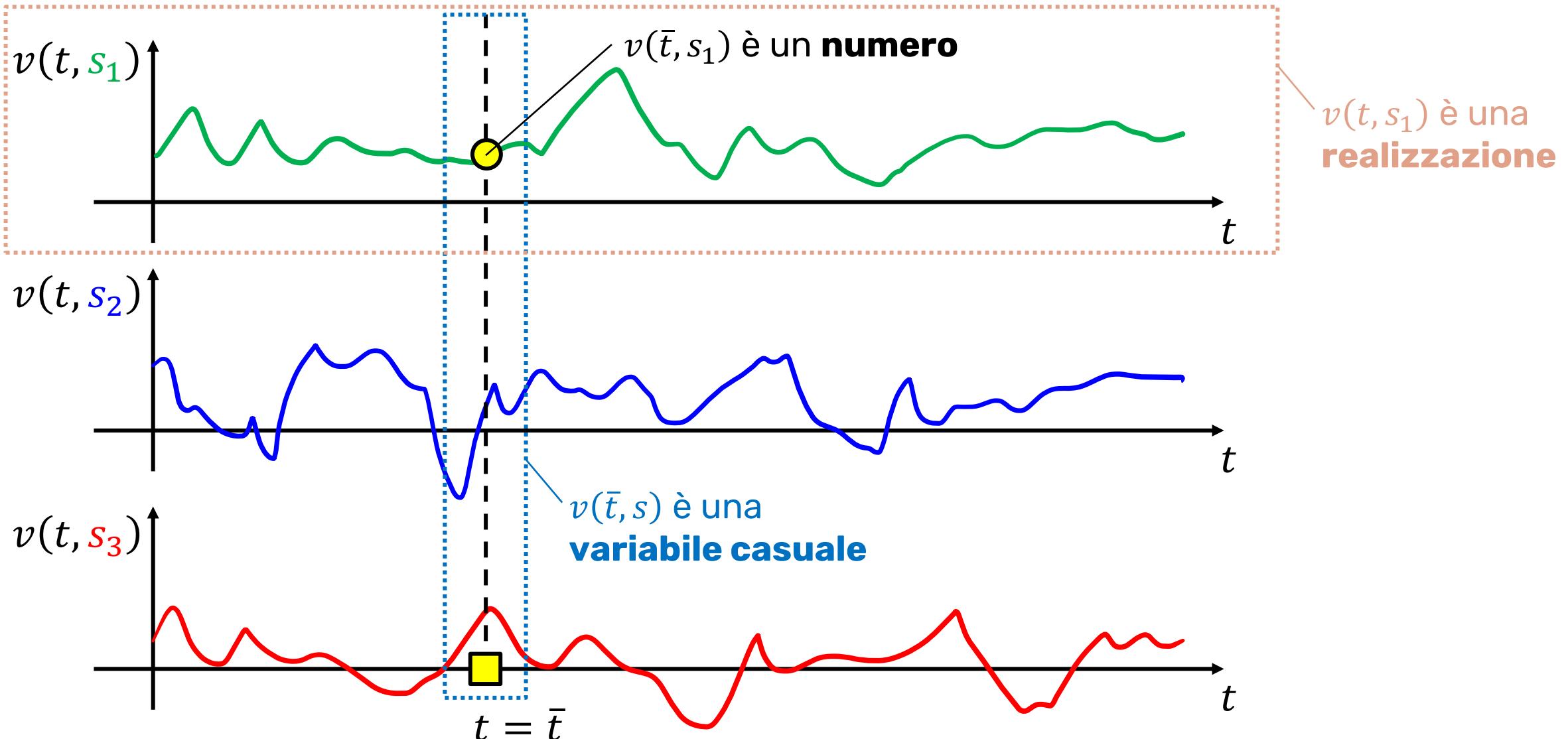
Variabile casuale



Processo stocastico



Processi stocastici



Esempio: random walk

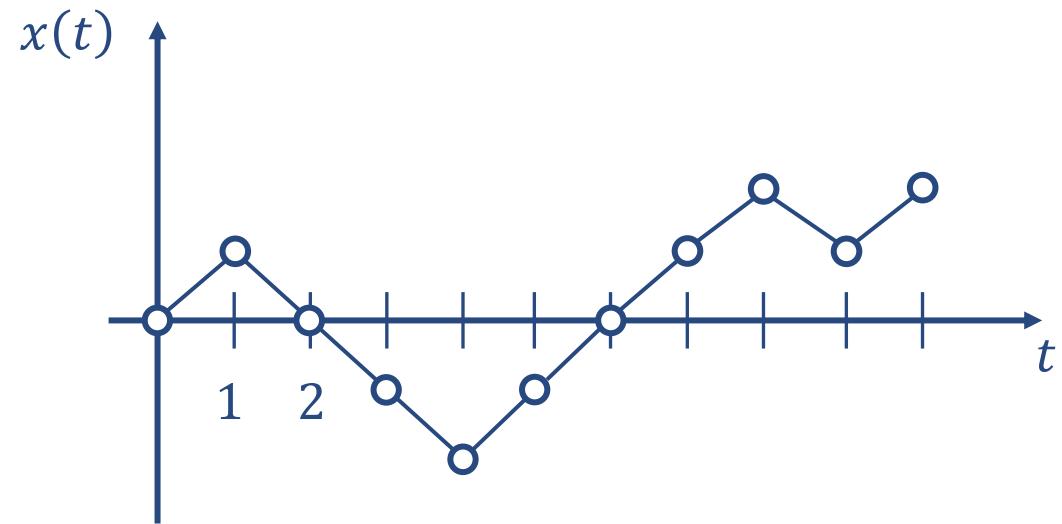
Consideriamo un esperimento costituito da una **sequenza di prove di Bernoulli** (che hanno esito «successo» con un certa probabilità π e «insuccesso» con probabilità $1 - \pi$)

$$x(t) = x(t - 1) + v(t)$$

$$v(t) = \begin{cases} 1 & s = \text{successo} \\ -1 & s = \text{insuccesso} \end{cases}$$

La variabile casuale $x(t)$ può essere pensata come l'andamento dei beni di un giocatore d'azzardo che gioca sempre la stessa posta a testa\croce

Questo processo è chiamato **random walk** ed ha molte applicazioni (può essere immaginato come una «camminata dell'ubriaco» che barcolla avanti e indietro)



Processi stocastici

L'utilizzo dei processi stocastici non si limita allo studio degli ubriachi! Sono utili ogni volta che vogliamo analizzare fenomeni che **non possiamo o non vogliamo** (per comodità) **descrivere deterministicamente**

Ciò capita, per esempio, quando è **troppo difficile descrivere la fisica** di un fenomeno

Esempio: Supponiamo di voler descrivere la **traiettoria di una palla di cannone**. Esso avrà una «traiettoria media» descrivibile con le leggi della cinematica, ma questa traiettoria sarà anche «sporcata» dal vento, dalla densità dell'aria, dalla dilatazione termina della canna del cannone...

Per cui, anziché cercare di descrivere tutto con delle leggi fisiche, si descrivere la traiettoria «media» in modo deterministico, e poi gli «scostamenti» vengono descritte tramite un processo casuale con certe proprietà



Processi stocastici

Un processo stocastico è **completamente caratterizzato** dal punto di vista probabilistico se, per ogni n -upla di variabili casuali $v(1), v(2), \dots, v(n)$, è nota la **distribuzione di probabilità congiunta** di queste variabili

- Posso immaginare (nel caso di processi discreti) come al dover conoscere la ddp congiunta di un **vettore di dimensione infinita di variabili casuali**

Con poche eccezioni (e.g. tutte le v.c. sono Gaussiane) questo è **bello ma impossibile**

Per cui, spesso ci si limita a considerare solo **valore atteso** e **funzione di covarianza** (o di correlazione) di un processo stocastico (caratterizzazione del 2° ordine)



Caratterizzazione del secondo ordine

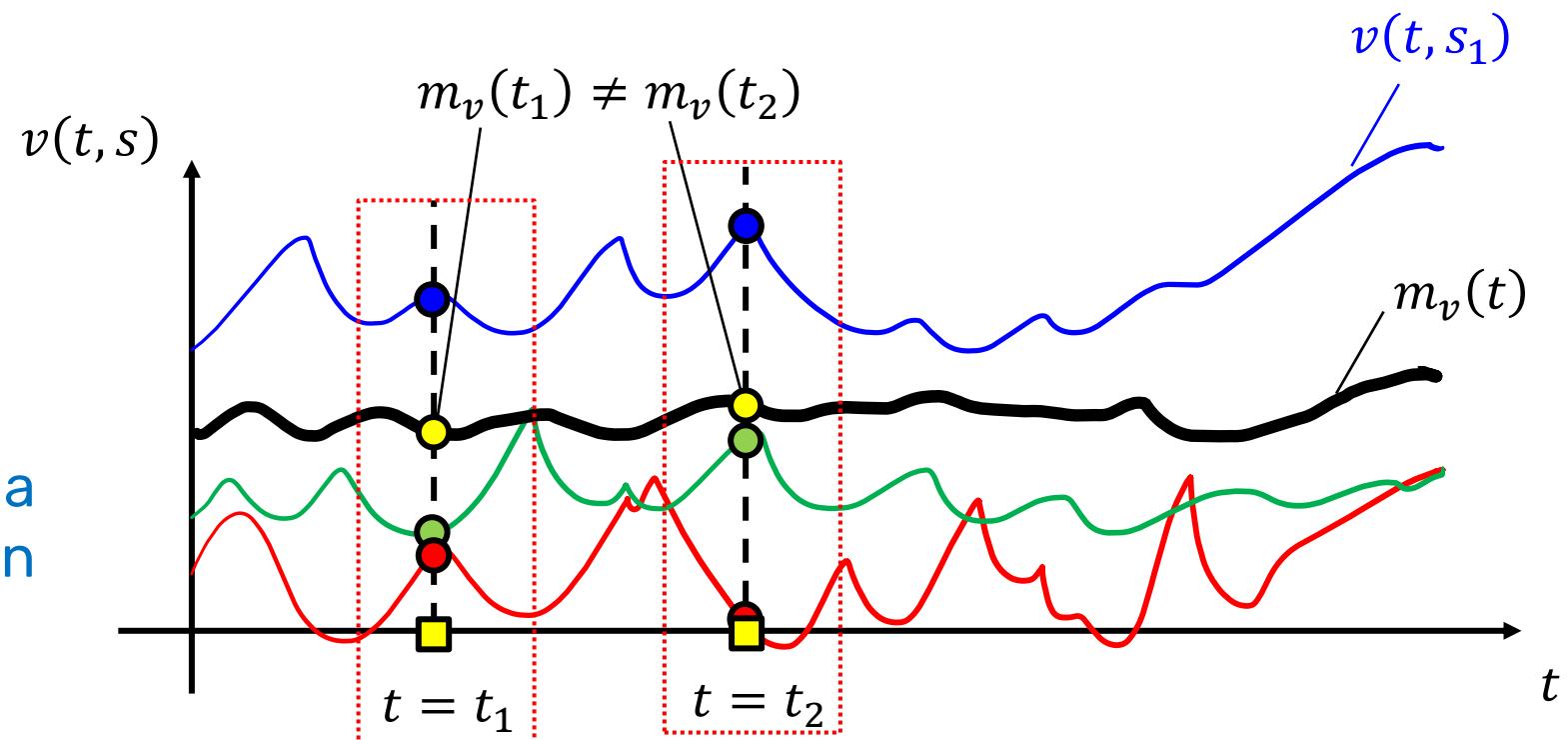
Dato un processo stocastico $v(t, s)$ si definiscono:

VALORE ATTESO (*momento del primo ordine*)

È una **funzione** che rappresenta il valore atteso della v.c. $v(t, s)$ al tempo t

$$m_v(t) \equiv \mathbb{E}_s[v(t, s)]$$

È la «**media in verticale**» (media di insieme) non quella in «**orizzontale**» (media temporale)

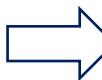


Caratterizzazione del secondo ordine

FUNZIONE DI AUTOCORRELAZIONE (*momento del secondo ordine*)

Permette di capire i valori che il processo assume ad un istante t_1 rispetto a quelli che assume ad un istante t_2

$$R_{vv}(t_1, t_2) \equiv \mathbb{E}_s[v(t_1, s)v(t_2, s)]$$

- $R_{vv}(t_1, t_2) > 0$  se $v(t_1, s) > 0$, allora, *in media*, $v(t_2, s) > 0$
se $v(t_1, s) < 0$, allora, *in media*, $v(t_2, s) < 0$
- $R_{vv}(t_1, t_2) < 0$  se $v(t_1, s) > 0$, allora, *in media*, $v(t_2, s) < 0$
se $v(t_1, s) < 0$, allora, *in media*, $v(t_2, s) > 0$

$v(t_1)$ e $v(t_2)$ **stesso segno**, in media

$v(t_1)$ e $v(t_2)$ **segno diverso**, in media



Caratterizzazione del secondo ordine

FUNZIONE DI AUTOCOVARIANZA

È la **covarianza** tra $v(t_1, s)$ e $v(t_2, s)$

$$\gamma_{vv}(t_1, t_2) \equiv \mathbb{E}_s[(v(t_1, s) - m_v(t_1)) \cdot (v(t_2, s) - m_v(t_2))]$$

- $\gamma_{vv}(t, t) = \text{Var}[v(t, s)]$: **varianza** del processo a tempo $t = t_1 = t_2$
- $\gamma_{vv}(t_1, t_2) = R_{vv}(t_1, t_2) - m_v(t_1) \cdot m_v(t_2)$
- Può essere vista come la funzione di autocorrelazione del **processo depolarizzato** $v(t, s) - m_v(t)$



Caratterizzazione del secondo ordine

FUNZIONE DI AUTOCOVARIANZA NORMALIZZATA

È una generalizzazione del coefficiente di correlazione tra $v(t_1, s)$ e $v(t_2, s)$

$$\rho_{vv}(t_1, t_2) \equiv \frac{\gamma_{vv}(t_1, t_2)}{\sqrt{\gamma_{vv}(t_1, t_1) \cdot \gamma_{vv}(t_2, t_2)}}$$

- $|\rho_{vv}(t_1, t_2)| \leq 1$
- È l'autocovarianza del **processo normalizzato**

$$\tilde{v}(t, s) = \frac{v(t, s) - m_v(t)}{\sqrt{\text{Var}[v(t, s)]}}$$



Processi stocastici congiunti

Consideriamo due processi stocastici $v(t, s)$ e $x(t, s)$. È possibile definire una funzione di cross-correlazione e cross-covarianza, che consideri l'interazione tra $v(\cdot)$ e $x(\cdot)$

FUNZIONE DI CROSS-CORRELAZIONE

$$R_{vx}(t_1, t_2) \equiv \mathbb{E}_s[v(t_1, s) \cdot x(t_2, s)] = R_{xv}(t_2, t_1)$$

FUNZIONE DI CROSS-COVARIANZA

$$\gamma_{vx}(t_1, t_2) \equiv \mathbb{E}_s[(v(t_1, s) - m_v(t)) \cdot (x(t_2, s) - m_x(t))] = \gamma_{xv}(t_2, t_1)$$



Processi stocastici congiunti

Due processi stocastici $v(t, s)$ e $x(t, s)$ si dicono **incorrelati** se

$$\gamma_{vx}(t_1, t_2) = 0, \quad \forall t_1, t_2$$

FUNZIONE DI CROSS-COVARIANZA NORMALIZZATA

$$\rho_{vx}(t_1, t_2) \equiv \frac{\gamma_{vx}(t_1, t_2)}{\sqrt{\gamma_{vv}(t_1, t_1) \cdot \gamma_{xx}(t_2, t_2)}}$$

Nota: spesso ometteremo da $v(t, s)$ la dipendenza dall'esito s , indicando $v(1, s), v(2, s), \dots$ con $v(1), v(2), \dots$



Outline

1. Introduzione alla stima di modelli dinamici
2. Processi stocastici
- 3. Processi stocastici stazionari**
4. Momenti temporali ed ergodicità
5. Trasformata \mathcal{Z} e trasformata di Fourier
6. Densità spettrale di potenza
7. Stima spettrale
8. Sistemi dinamici lineari discreti deterministici
9. Sistemi dinamici lineari discreti stocastici



Processi stocastici stazionari

Definizione: un processo stocastico $v(t)$ si dice **stazionario in senso forte** se $\forall n \in \mathbb{N}$, scelti t_1, t_2, \dots, t_n istanti di tempo, le caratteristiche probabilistiche della n -upla $v(t_1), v(t_2), \dots, v(t_n)$ sono uguali a quelle della n -upla $v(t_1 + \tau), v(t_2 + \tau), \dots, v(t_n + \tau)$, $\forall \tau \in \mathbb{N}$

Definizione: un processo stocastico $v(t)$ si dice **stazionario in senso debole** se:

- $m_v(t) = m, \quad \forall t$
- $\gamma_{vv}(t_1, t_2) = \gamma_{vv}(t_3, t_4)$ nel caso in cui $|t_4 - t_3| = |t_2 - t_1| = \tau$



L' autocovarianza dipende solo dal **LAG** τ e non dagli specifici valori di t_1, t_2, t_3, t_4



Processi stocastici stazionari

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

Nel corso **supporremo che i processi siano stazionari in senso debole**, non dovendo mai imporre specifiche distribuzioni di probabilità sulle variabili casuali che lo compongono

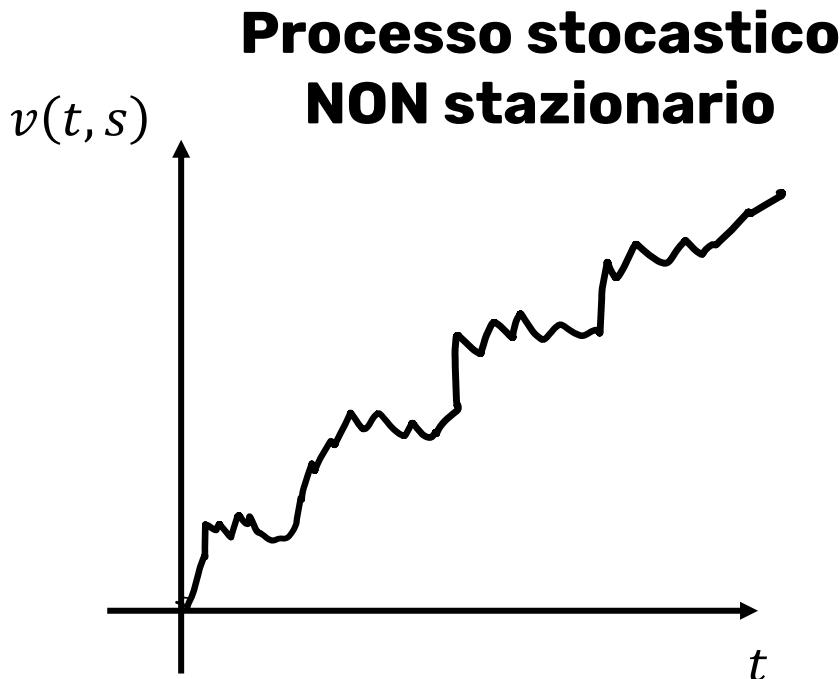
Dato che, nel caso di processi stocastici stazionari (pss), la **funzione di autocovarianza dipende solo dal lag τ** , si scrive:

$$\gamma_{vv}(\tau) = \mathbb{E}_s[(v(\textcolor{blue}{t}, s) - m) \cdot (v(\textcolor{blue}{t} + \tau, s) - m)]$$

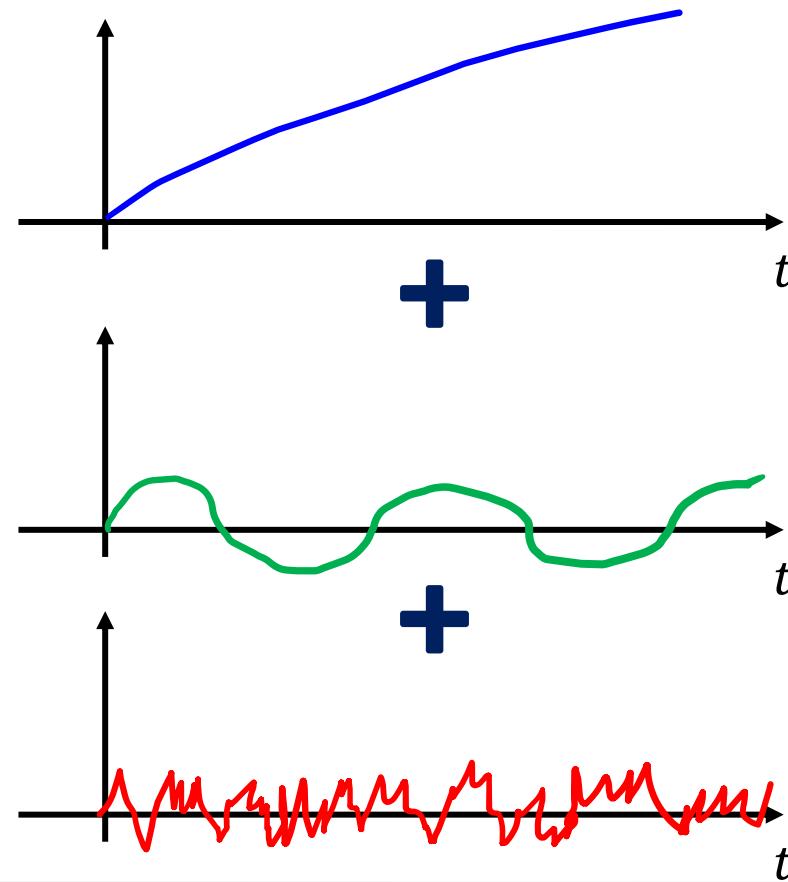


Processi stocastici stazionari

Lo studio dei processi stocastici stazionari **non è limitante**, in quanto potremmo **decomporre** un processo non stazionario in diverse componenti:



=



Processi stocastici stazionari

Definizione: due processi stocastici stazionari $\nu_1(t)$ e $\nu_2(t)$ si dicono **equivalenti** se hanno lo stesso valore atteso m e stessa funzione di autocovarianza $\gamma(\tau)$

Nota: durante il corso studieremo processi stocastici stazionari



Proprietà della funzione di autocovarianza di un pss

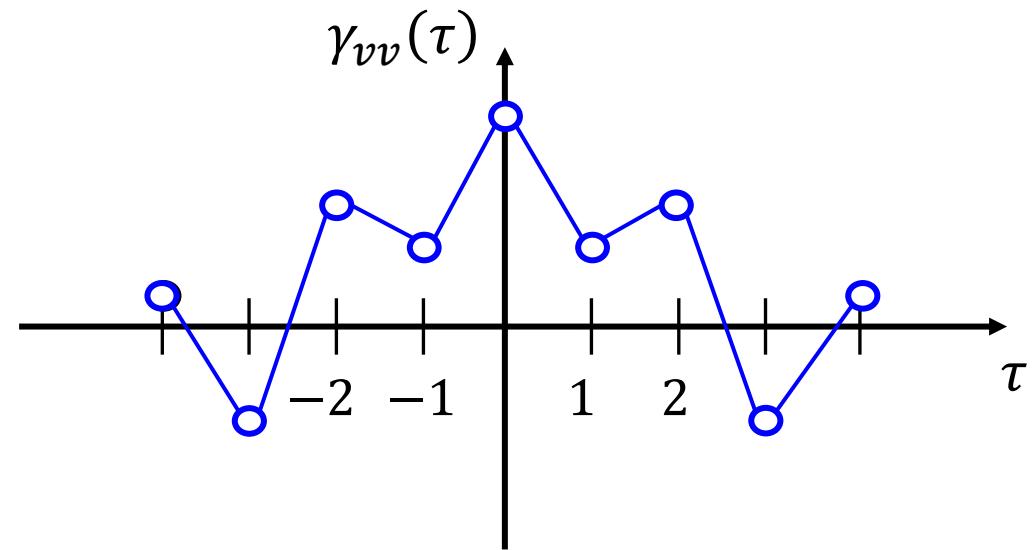
Dalla definizione di funzione di autocovarianza (e di autocorrelazione) di un processo stocastico stazionario, se ne deducono le seguenti proprietà:

1) $\gamma_{vv}(0) = \mathbb{E}[(v(t) - m)^2] \geq 0$

Varianza del processo

2) $|\gamma_{vv}(\tau)| \leq \gamma_{vv}(0), \forall \tau$ **Funzione limitata**

Il legame tra $v(t)$ e se stesso è più forte che tra $v(t)$ e $v(t + \tau)$, $\tau \neq 0$



3) $\boxed{\gamma_{vv}(\tau)} = \gamma_{vv}(t, t + \tau) \Rightarrow \bar{t} = t + \tau \Rightarrow \gamma_{vv}(\bar{t} - \tau, \bar{t}) = \gamma_{vv}(\bar{t}, \bar{t} - \tau) = \boxed{\gamma_{vv}(-\tau)}$

Funzione pari



Caso particolare di pss: rumore bianco (white noise)

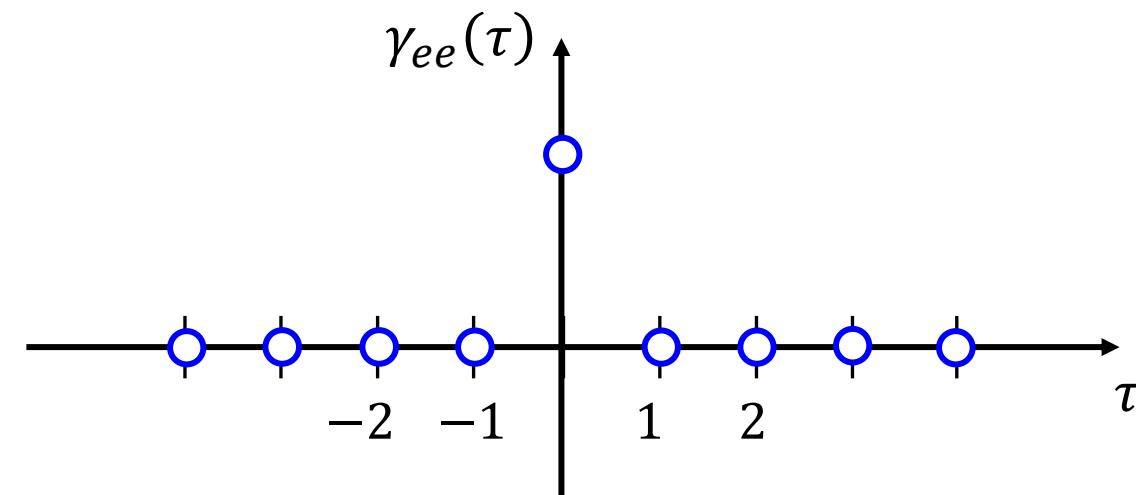
Definizione: Un pss $e(t) \sim WN(\mu, \lambda^2)$ è detto **rumore bianco** se:

$$1) \mathbb{E}[e(t)] = \mu$$

$$2) \gamma_{ee}(0) = \mathbb{E}[(e(t) - \mu)^2] = \lambda^2, \quad \forall t$$

$$3) \gamma_{ee}(\tau) = \mathbb{E}[(e(t) - \mu) \cdot (e(t + \tau) - \mu)] = 0, \quad \forall t, \forall \tau \neq 0$$

Siccome **non vi è correlazione** tra il valore ad un istante t ed un valore all'istante $t + \tau$, il **rumore bianco** (stazionario) è un processo stocastico le cui **realizzazioni variano in modo impredicibile** da un istante all'altro



Caso particolare di pss: rumore bianco (white noise)

Nota: Per i nostri fini, non è importante la distribuzione delle singole v.c. $e(t_1), e(t_2), \dots$ del processo rumore bianco

Nota: Spesso, considereremo processi stocastici stazionari a media nulla. Infatti, il valore della media del processo non modifica la sua funzione di autocovarianza (in altri termini, si dice che non modifica le «caratteristiche spettrali» del processo)



Outline

1. Introduzione alla stima di modelli dinamici
2. Processi stocastici
3. Processi stocastici stazionari
- 4. Momenti temporali ed ergodicità**
5. Trasformata \mathcal{Z} e trasformata di Fourier
6. Densità spettrale di potenza
7. Stima spettrale
8. Sistemi dinamici lineari discreti deterministici
9. Sistemi dinamici lineari discreti stocastici



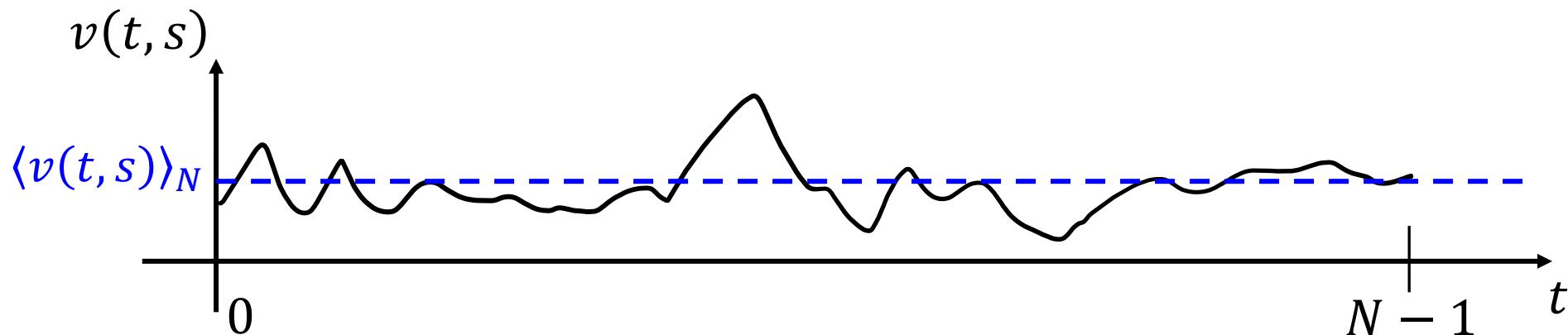
Momenti temporali

Supponiamo che $v(t, s)$ sia un **processo stazionario**. Definiamo:

MEDIA TEMPORALE SU ORIZZONTE FINITO

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

Sto considerando N campioni temporali di una realizzazione di un pss



Momenti temporali

MEDIA TEMPORALE

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

AUTOCORRELAZIONE TEMPORALE

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



Momenti temporali

Ciò che abbiamo presentato sono quantità simili a degli **stimatori campionari**

Osservazioni

1. La quantità $\langle v(t, s) \rangle_N$ è una **variabile casuale** perché dipende dall'esito s
2. La quantità $\langle v(t, s) \rangle$ è un **limite di variabili casuali**. Quando converge?

Teorema

Se

1. $v(t, s)$ è un pss
2. $|\mathbb{E}_s[v(t, s)]| < +\infty$ (ovvero se la media esiste finita)

Allora il limite $\langle v(t, s) \rangle$ **converge quasi certamente**



$P\left(\lim_{n \rightarrow +\infty} v_n = v\right) = 1$. La v.c. v_n converge alla v.c. v per $n \rightarrow +\infty$, e differiranno solo su eventi di probabilità nulla. È il tipo di convergenza più forte



Momenti temporali

Le conseguenze del teorema sono che:

- $\mathbb{E}_s[\langle v(t, s) \rangle] = \mathbb{E}_s[v(t, s)] = m$
- $\mathbb{E}_s[\langle v(t, s) \cdot v(t + \tau, s) \rangle] = R_{vv}(\tau)$

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



Momenti temporali -stima della media del processo

Idea: sia $v(t, s)$ un processo stocastico stazionario di cui voglio stimare il valore atteso m .

In teoria, mi servirebbero n realizzazioni $v(t, s_1), v(t, s_2), \dots, v(t, s_n)$ del processo $v(t, s)$.

Potrei poi stimare il valore atteso, scegliendo un istante \bar{t} , come:

$$\hat{m} = \frac{1}{n} \sum_{i=1}^n v(\bar{t}, s_i)$$

**Media in «verticale»
o media di insieme**

A differenza del caso di variabili casuali, in cui di solito ho tante osservazioni di questa variabile, nel caso di pss spesso ho **solo una realizzazione finita** della serie temporale (che interpreto come realizzazione di un pss)

Possiamo usare $\langle v(t, s) \rangle_N$ come stimatore di m ?

**Media in «orizzontale»
o media temporale**



Processi stocastici ergodici

Definizione: il processo stocastico $v(t, s)$ è detto **ergodico** se:

1. $v(t, s)$ è stazionario
2. Per $N \rightarrow +\infty$, i momenti temporali convergono quasi certamente ai rispettivi momenti di insieme

Definizione: il processo stazionario $v(t, s)$ è detto **ergodico nella media** (proprietà più debole) se:

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



Processi stocastici ergodici

Teorema Sia $v(t, s)$ un **pss in senso debole**. Allora, se

1. $|\gamma_{vv}(0)| < +\infty$ (la varianza esiste finita)
2. $\lim_{\tau \rightarrow +\infty} \gamma_{vv}(\tau) = 0$ (la funzione di autocovarianza tende a zero)

si ha che $v(t, s)$ è **ergodico nella media**

Teorema Sia $v(t, s)$ **stazionario e Gaussiano**. Allora, se

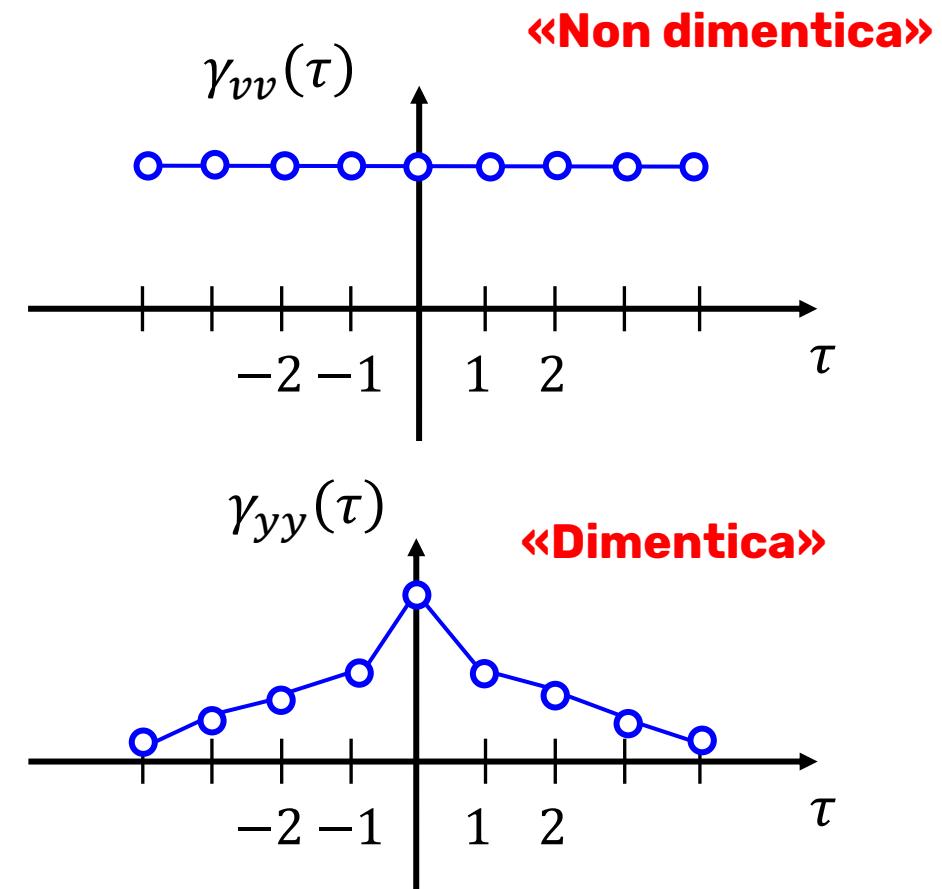
1. $|\gamma_{vv}(0)| < +\infty$ (la varianza esiste finita)
2. $\lim_{\tau \rightarrow +\infty} \gamma_{vv}(\tau) = 0$ (la funzione di autocovarianza tende a zero)

si ha che $v(t, s)$ è **ergodico**



Processi stocastici ergodici

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



Processi stocastici ergodici

Nella pratica, l'utilizzo dell'ergodicità è un «cane che si morde la coda»:

- Per sapere se un processo è **ergodico**, devo conoscere $\gamma_{vv}(\tau)$, si vedano le condizioni sufficienti dei teoremi
- Però, a meno che non abbia già informazioni su $\gamma_{vv}(\tau)$, devo **stimarla dai dati**. Per stimarla dai dati però, il processo deve essere ergodico

Se non ho informazioni precise sul meccanismo di generazione dati, spesso non posso fare altro che **ipotizzare l'ergodicità** senza poterla dimostrare, e procedere stimando le caratteristiche del processo tramite momenti temporali



Outline

1. Introduzione alla stima di modelli dinamici
2. Processi stocastici
3. Processi stocastici stazionari
4. Momenti temporali ed ergodicità
- 5. Trasformata Z e trasformata di Fourier**
6. Densità spettrale di potenza
7. Stima spettrale
8. Sistemi dinamici lineari discreti deterministici
9. Sistemi dinamici lineari discreti stocastici



Trasformata Z

Definizione: la **trasformata Zeta bilatera** di un segnale discreto deterministico $g(t)$, $t \in \mathbb{Z}$, è definita come

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

- $g(t)$ è una **funzione reale** di **variabile intera** $t \in \mathbb{Z}$ (funzione a tempo discreto)
- $G(z)$ è una **funzione complessa** di **variabile complessa** $z \in \mathbb{C}$

Riguardare la Lezione 29 di Fondamenti di Automatica (9 cfu)!



Trasformata Z

Proprietà della trasformata Zeta bilatera

- **Linearità:** $\mathcal{Z}[\alpha g(t) + \beta h(t)] = \alpha G(z) + \beta H(z), \forall \alpha, \beta \in \mathbb{R}$
- **Anticipo:** $\mathcal{Z}[g(t + 1)] = z \cdot G(z)$

Dimostrazione: $\sum_{t=-\infty}^{+\infty} g(t + 1)z^{-t}$

$$= \dots + g(-1)z^2 + g(0)z^1 + g(1) + g(2)z^{-1} + g(3)z^{-2} + \dots$$

$$= z \cdot [\dots + g(-1)z^1 + g(0) + g(1)z^{-1} + g(2)z^{-2} + g(3)z^{-3} + \dots] = z \cdot G(z)$$

- **Ritardo:** $\mathcal{Z}[g(t - 1)] = z^{-1} \cdot G(z)$

z : operatore di **anticipo unitario**

z^{-1} : operatore di **ritardo unitario**



Convoluzione di segnali discreti

Definizione: la **convoluzione** (discreta) tra due segnali discreti $g(t)$ e $u(t)$ è definita come

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

Proprietà: si dimostra che la **trasformata \mathcal{Z} della convoluzione** è il **prodotto delle trasformate \mathcal{Z}**

$$Y(z) = G(z)U(z)$$



Trasformata di Fourier a Tempo Discreto (DTFT)

Definizione: Sia $u(t)$ un segnale discreto, deterministico, assolutamente sommabile, ovvero tale che

$$\sum_{t=-\infty}^{+\infty} |u(t)| < +\infty$$

Allora, si definisce **trasformata di Fourier a tempo discreto (DTFT)** la quantità

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

- È una **funzione complessa** della **variabile reale** $\omega \in \mathbb{R}$. Quindi, $\mathcal{F}[u(t)]$ è una funzione continua poiché ω assume valori in \mathbb{R}



Trasformata di Fourier a Tempo Discreto (DTFT)

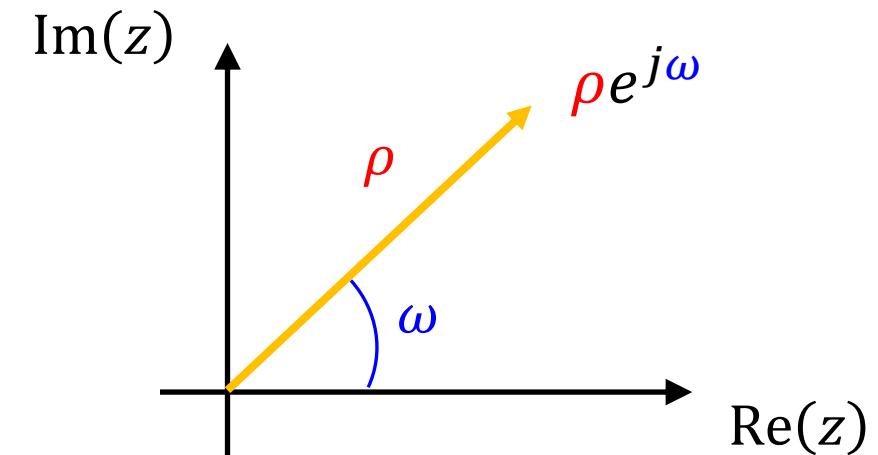
Proprietà della DTFT: la trasformata di Fourier a tempo discreto di un segnale discreto $u(t)$ si ottiene valutando la trasformata \mathcal{Z} di $u(t)$ in $z = e^{j\omega}$:

$$\mathcal{F}[u(t)] = U(e^{j\omega})$$

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

Interpretazione: la trasformata di Fourier a tempo discreto è la **restrizione** di $U(z)$ alla **circonferenza di raggio unitario, cioè i valori di z** che si possono scrivere come $e^{j\omega}$, ovvero i punti con modulo $\rho = 1$

$$\mathcal{F}[u(t)] = U(z) \Big|_{z=e^{j\omega}}$$

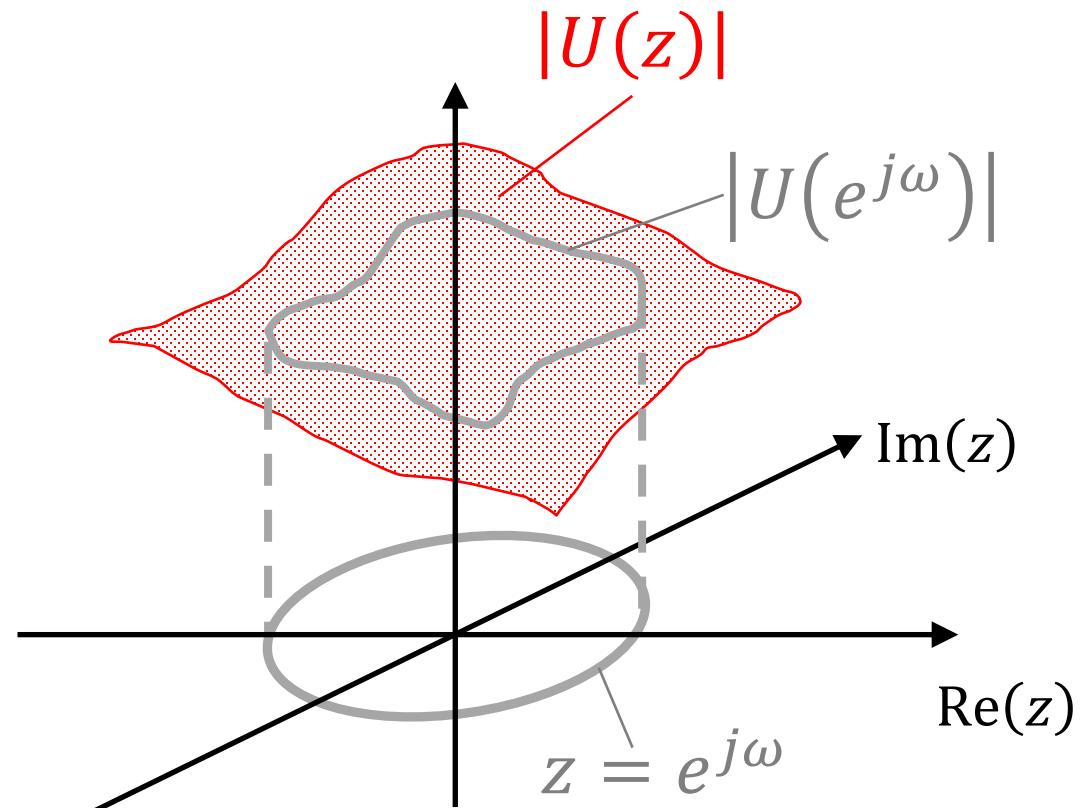


Trasformata di Fourier a Tempo Discreto (DTFT)

Interpretazione: la trasformata di Fourier a tempo discreto è la **restrizione** di $U(z)$ alla **circonferenza di raggio unitario**

$$\mathcal{F}[u(t)] = U(z) \Big|_{z=e^{j\omega}}$$

Rappresentiamo per semplicità solo il modulo di $U(z)$ e $U(e^{j\omega})$. Allora, $|U(z)|$ è una **superficie** nel piano complesso, mentre $|U(e^{j\omega})|$ è un «**percorso**» nel piano complesso



Trasformata di Fourier a Tempo Discreto (DTFT)

Sembrerebbe quindi che la trasformata di Fourier contenga meno informazioni rispetto alla trasformata \mathcal{Z} . In realtà, è possibile **ricostruire completamente** $u(t)$ partendo da $U(e^{j\omega})$

Altre proprietà della DTFT

- $X(e^{j(\omega+2k\pi)}) = X(e^{j\omega})$: 2π **periodica**, infatti quando «completo un giro» della circonferenza $e^{j\omega}$, ritorno al punto di partenza
- $\bar{X}(e^{j\omega}) = X(e^{-j\omega})$, il **complesso coniugato** $\bar{X}(e^{j\omega})$ del numero $X(e^{j\omega})$ si trova cambiando il segno dell'angolo ω



Tutta l'informazione è contenuta in $[0, \pi]$



Trasformata di Fourier Discreta (DFT)

Consideriamo un segnale discreto $u(t)$ di **durata finita**, definito su $t \in [0, N - 1] \subset \mathbb{N}$.

Definiamo la **trasformata di Fourier discreta (DFT)** come

$$\check{U}(k) \equiv \sum_{t=0}^{N-1} u(t) \cdot e^{-j \cdot t \cdot k \phi}$$

- $\phi = \frac{2\pi}{N}$
- $k = 0, \dots, N - 1$

- È una **funzione complessa** della **variabile intera** $k \in \mathbb{N}$. Quindi, $\check{U}(k)$ è una funzione discreta poiché k assume valori in \mathbb{N}
- Parto con $u(t)$ che è definito come un vettore di N numeri reali, e arrivo con $\check{U}(k)$ che è definita come un vettore di N numeri complessi



Trasformata di Fourier Discreta (DFT)

Proprietà della DFT

Consideriamo un segnale $u(t)$ tale che esso valga 0 per $t < 0$ e $t \geq N$. Allora, la **DFT** può essere vista come un «**campionamento**» della DTFT

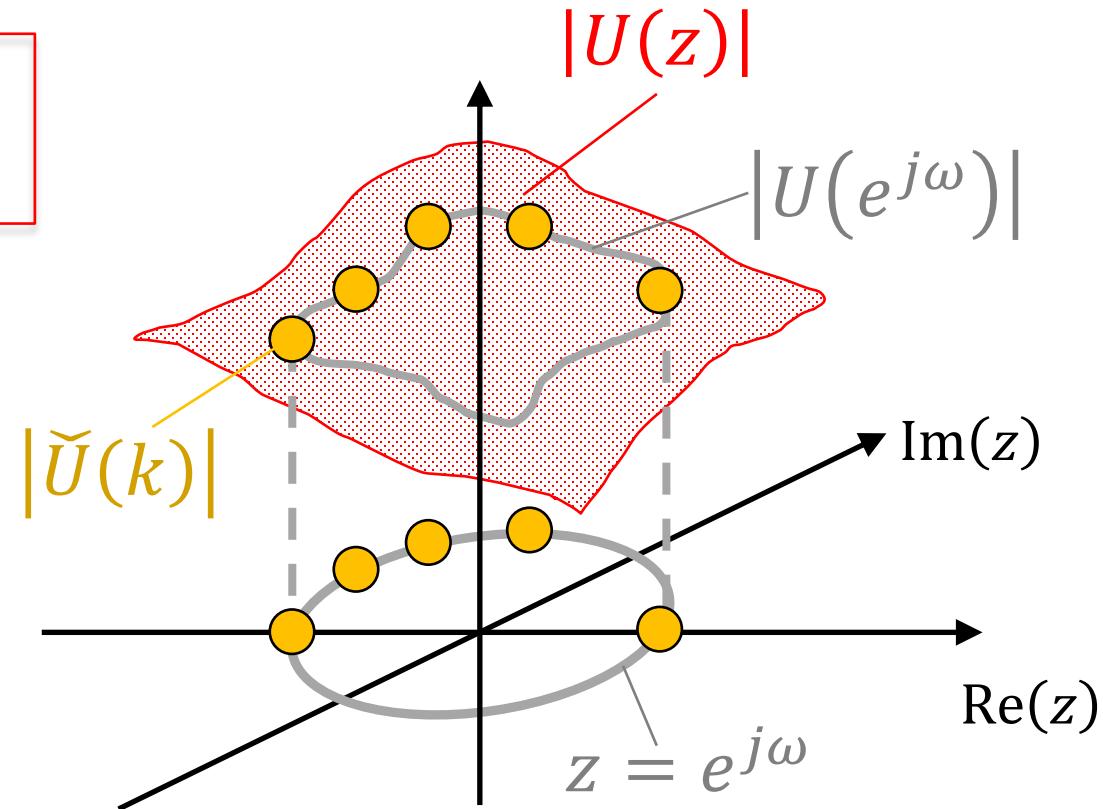
$$\check{U}(k) = U(e^{j \cdot k \cdot 2\pi/N})$$

Dimostrazione

$$U(e^{j\omega}) \equiv \sum_{t=-\infty}^{+\infty} u(t) \cdot e^{-j\omega t} = \sum_{t=0}^{N-1} u(t) \cdot e^{-j\omega t}$$

→ Se impongo $\omega = k \cdot \phi = k \cdot \frac{2\pi}{N}$ →

la proprietà è dimostrata



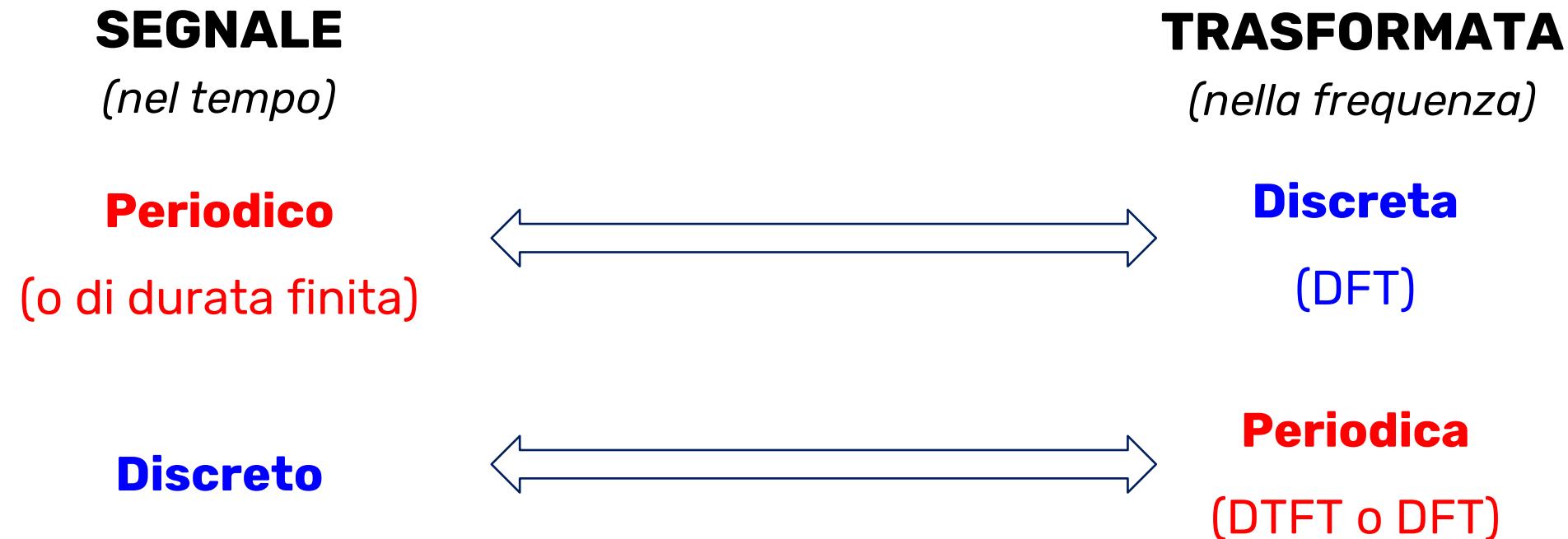
Trasformata di Fourier Discreta (DFT)

Proprietà della DFT

- Esiste una **DFT inversa (IDFT)** tale che è possibile **ricostruire** $u(t)$ partendo da $\tilde{U}(k)$.
Quindi, la **DFT non fa perdere alcuna informazione**
 - ✓ Per poter usare la DFT, abbiamo bisogno di **segnali di durata finita**. È anche vero però che nella pratica i nostri segnali saranno sempre di durata finita...
- La **risoluzione della DFT**, chiamata anche «**frequency bin**» è data da $\text{bin} = f_s/N$, dove f_s è la frequenza di campionamento
 - ✓ Dato che la **DFT è simmetrica**, solo $N/2$ dati portano informazione, la $N/2$ -esima frequenza $k = N/2$ rappresenta la frequenza di Nyquist $f_s/2$



Nota sui segnali e sulle loro trasformate di Fourier



La **DFT** è **sia discreta che periodica**. Quindi, presuppone che il segnale nel tempo sia discreto e anche periodico. Se il segnale non è periodico e applico la DFT, potrei avere problemi di **leakage** (ma non tratteremo questo problema)



Outline

1. Introduzione alla stima di modelli dinamici
2. Processi stocastici
3. Processi stocastici stazionari
4. Momenti temporali ed ergodicità
5. Trasformata \mathcal{Z} e trasformata di Fourier
- 6. Densità spettrale di potenza**
7. Stima spettrale
8. Sistemi dinamici lineari discreti deterministici
9. Sistemi dinamici lineari discreti stocastici



Densità spettrale di potenza di un pss

Sia $\nu(t, s)$ un **pss**. Abbiamo visto diversi modi per poterlo caratterizzare, come:

- Valore atteso $m = \mathbb{E}_s[\nu(t, s)]$
- Funzione di autocovarianza $\gamma_{\nu\nu}(\tau)$

Sia il valore atteso che la funzione di autocovarianza sono **caratterizzazioni «nel tempo»**. È però possibile, proprio come per i segnali deterministicici, caratterizzare un pss **«nella frequenza»** (ovvero, nel *dominio delle trasformate*)

L'evoluzione delle realizzazioni di un pss è prettamente caratterizzata dalla funzione di autocovarianza. Per questo motivo, *spesso si studiano pss «depurati» dalla loro media*

Idea: anziché $\gamma_{\nu\nu}(\tau)$, considero le sue **trasformate**



Densità spettrale di potenza di un pss

Definizione: 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 $\gamma_{vv}(\tau)$:

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

La trasformata \mathcal{Z} di $\gamma_{vv}(\tau)$ è:

$$\Phi_{vv}(z) \equiv \mathcal{Z}[\gamma(\tau)] = \sum_{\tau=-\infty}^{\tau=+\infty} \gamma_{vv}(\tau) \cdot z^{-\tau}$$

- Data $\Phi_{vv}(z)$, si ha che $\Gamma_{vv}(\omega) = \Phi_{vv}(e^{j\omega})$



Densità spettrale di potenza di un pss

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
- Come l'energia del processo si distribuisce alle varie frequenze



Densità spettrale di potenza di un pss

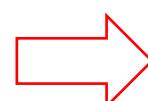
Osservazioni

- Conoscere $\Gamma_{vv}(\omega)$ o $\Phi_{vv}(z)$ è equivalente: posso risalire a $\gamma_{vv}(\tau)$ con l'antitrasformata
- Affinché $\Gamma_{vv}(\omega)$ converga, $\gamma_{vv}(\tau)$ devo tendere a zero in modo sufficientemente rapido.
Studieremo casi in cui questo vale sempre

Proprietà di $\Gamma_{vv}(\omega)$

1. Reale: dato che $\gamma_{vv}(\tau)$ è pari, i termini immaginari del tipo $\pm j\sin(\omega)$ si elidono

2. Positiva: $\Gamma_{vv}(\omega) \geq 0, \quad \forall \omega \in \mathbb{R}$



Ci basta valutare $\Gamma_{vv}(\omega)$ tra $[0, \pi]$

3. Pari: $\Gamma_{vv}(\omega) = \Gamma_{vv}(-\omega), \quad \forall \omega \in \mathbb{R}$

4. Periodica di periodo 2π : $\Gamma_{vv}(\omega) = \Gamma_{vv}(\omega + k \cdot 2\pi), \quad \forall \omega \in \mathbb{R}, \forall k \in \mathbb{Z}$



Densità spettrale di potenza di un pss

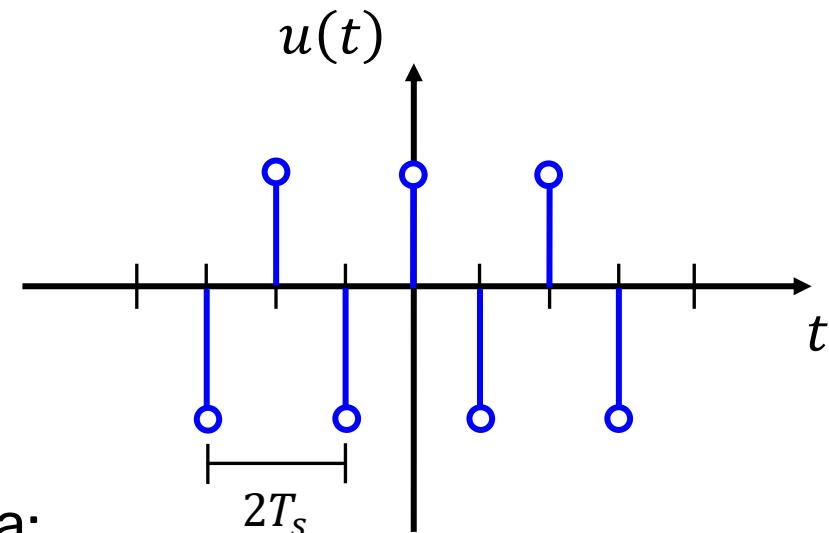
Osservazione

A tempo discreto, la **più grande pulsazione osservabile** è quella di una cosinusoide che cambia valore ad ogni istante di tempo t

Tra l'istante t l'istante $t + 1$, trascorre un tempo di campionamento T_s . Il più **piccolo periodo osservabile** è quindi $T = 2T_s = 2/f_s$ [s]

La **pulsazione [rad/s] osservabile più grande** corrisponde a:

$$\omega = \frac{2\pi}{T} = \frac{\pi}{T_s} = \pi \cdot f_s \text{ [rad/s]} \quad (\text{Teorema del campionamento})$$

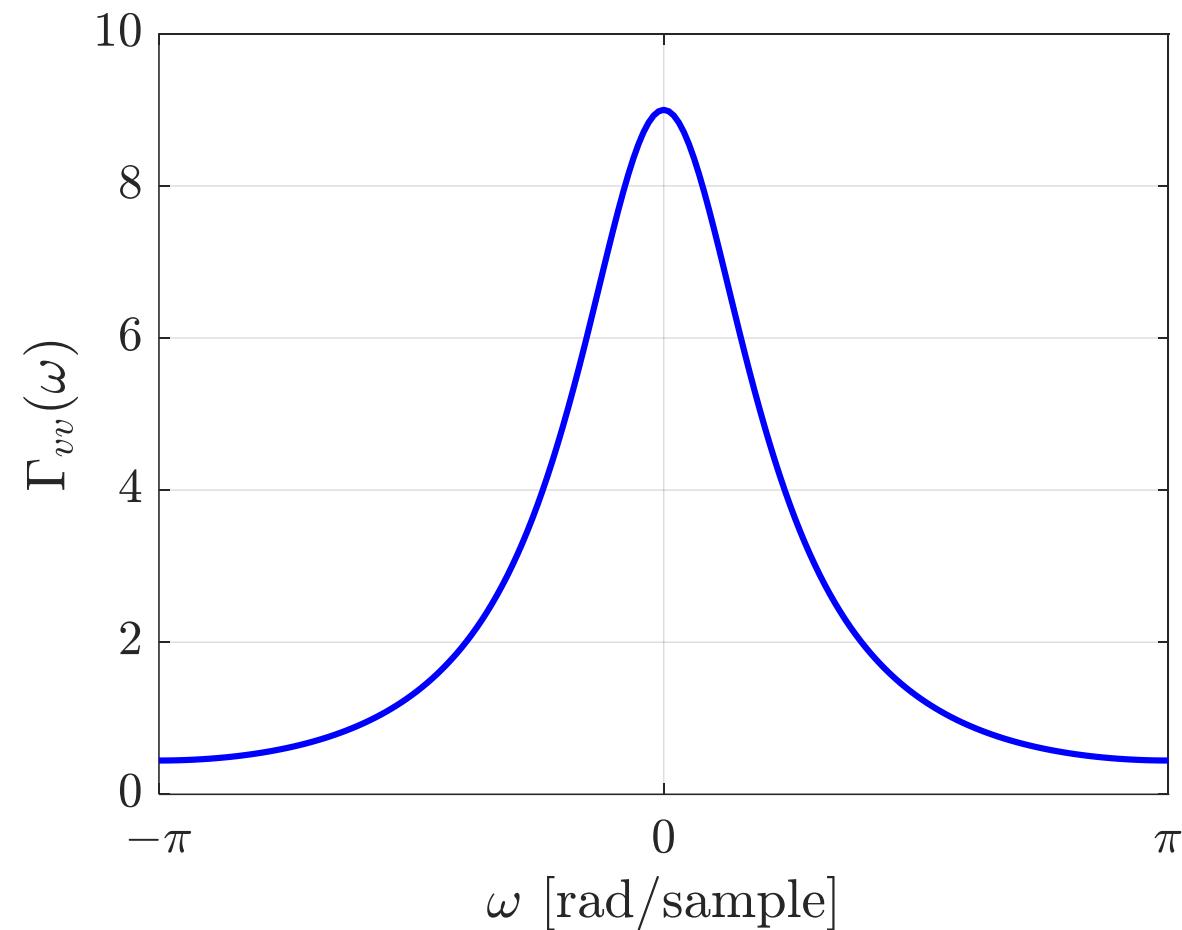


Densità spettrale di potenza di un pss

Quindi, la pulsazione ω , con la quale si rappresenta la densità spettrale di potenza, è una **«pulsazione normalizzata»** rispetto alla frequenza di campionamento f_s

Interpretazione: π [rad/s]corrisponde a $f_s/2$ [Hz]

Questa interpretazione deriva dalla definizione della DTFT, che considera «in modo implicito» il tempo di campionamento dei dati T_s



Densità spettrale di potenza di un pss

Altra proprietà di $\Gamma_{vv}(\omega)$

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

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

Si nota che è possibile esprimere la **varianza del processo** stazionario come **l'area sottesa** alla densità spettrale di potenza (a meno del fattore 2π)

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



Densità spettrale di potenza di un rumore bianco

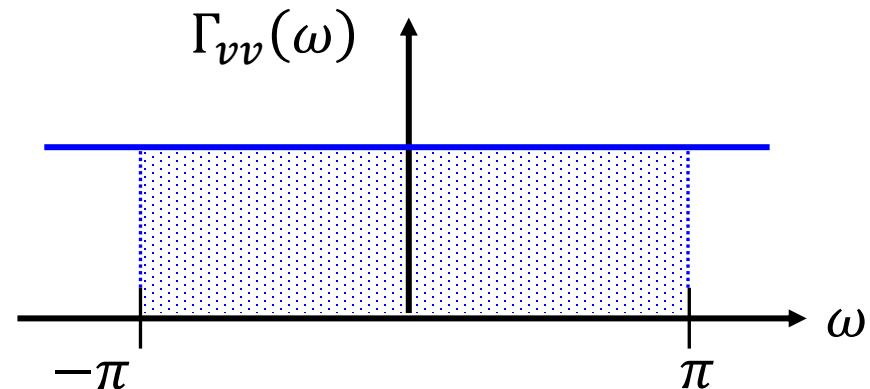
Sia $e(t) \sim WN(0, \lambda^2)$. Sappiamo che nel tempo è un segnale **impredicibile**, dato che:

$$\gamma_{ee}(\tau) = \begin{cases} 0 & \text{se } \tau \neq 0 \\ \lambda^2 & \text{se } \tau = 0 \end{cases} \quad \rightarrow \quad \Gamma_{ee}(\omega) = \sum_{\tau=-\infty}^{+\infty} \gamma_{ee}(\tau) \cdot e^{-j\omega\tau} = \lambda^2 \cdot e^{-j\omega 0} = \boxed{\lambda^2}$$
$$\Phi(z) = \lambda^2$$

La densità spettrale di potenza del rumore bianco è quindi **costante**. Tutte le frequenze hanno tutte la stessa potenza media

Non vi sono frequenze dominanti: tutte contribuiscono in modo uguale alla variabilità del segnale

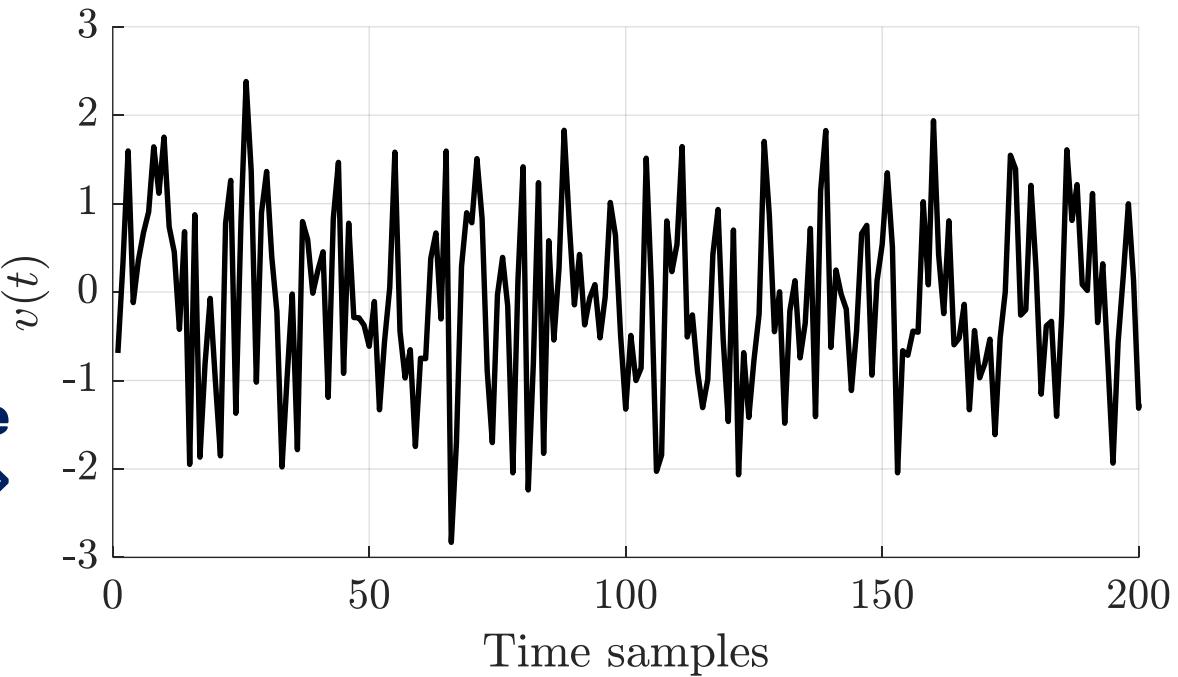
IMPREVIDIBILITÀ DEL PROCESSO



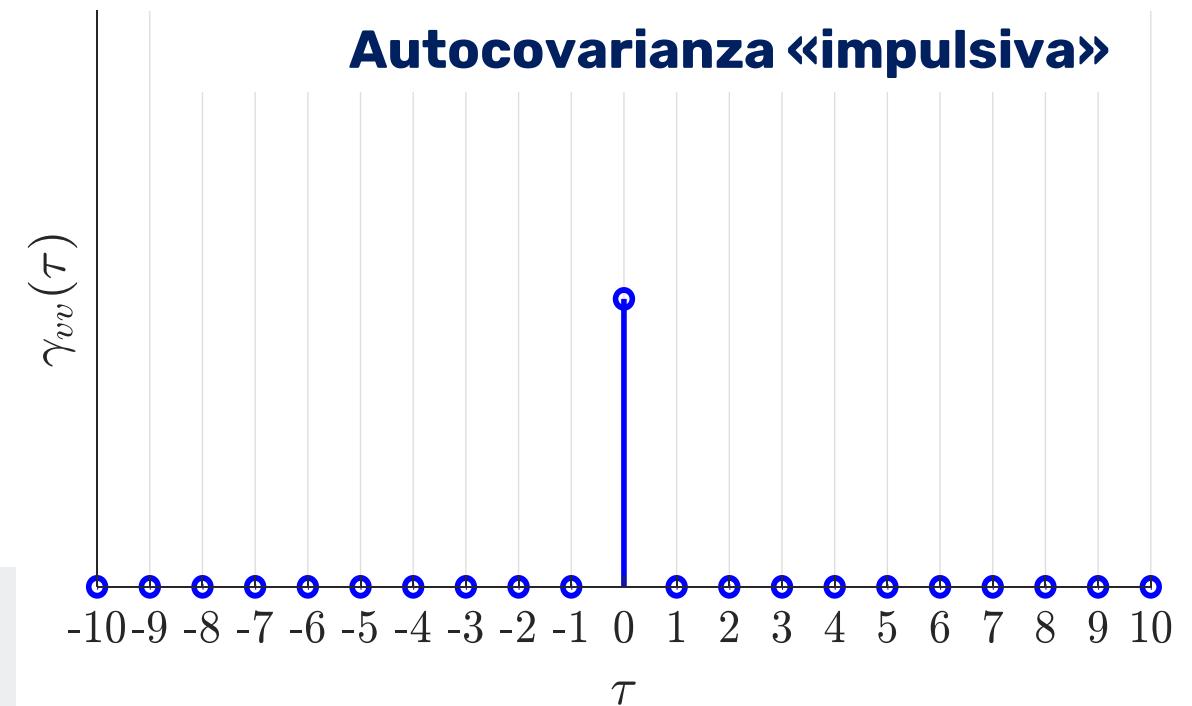
Esempio: rumore bianco

Pss a media zero

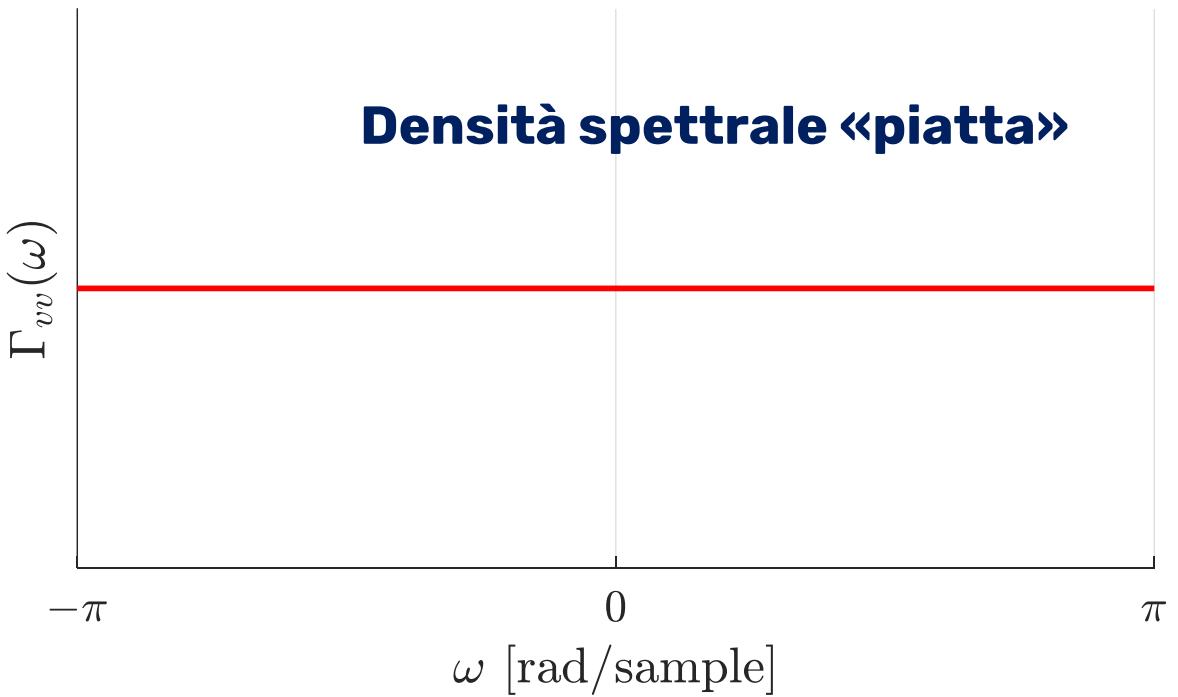
**Andamento temporale
«impredicibile»**



Autocovarianza «impulsiva»



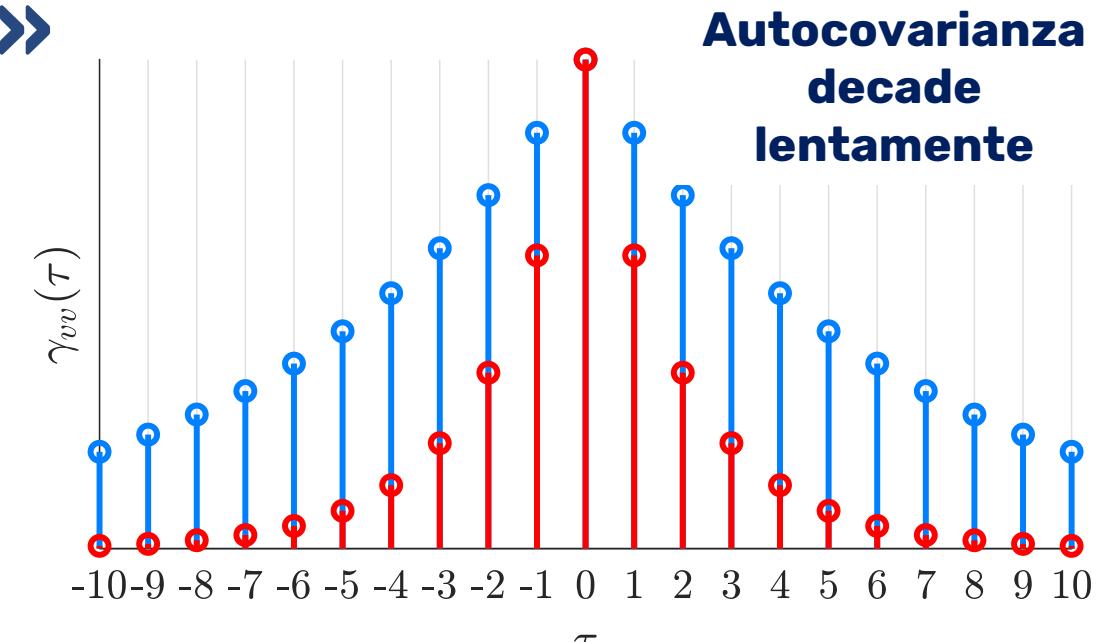
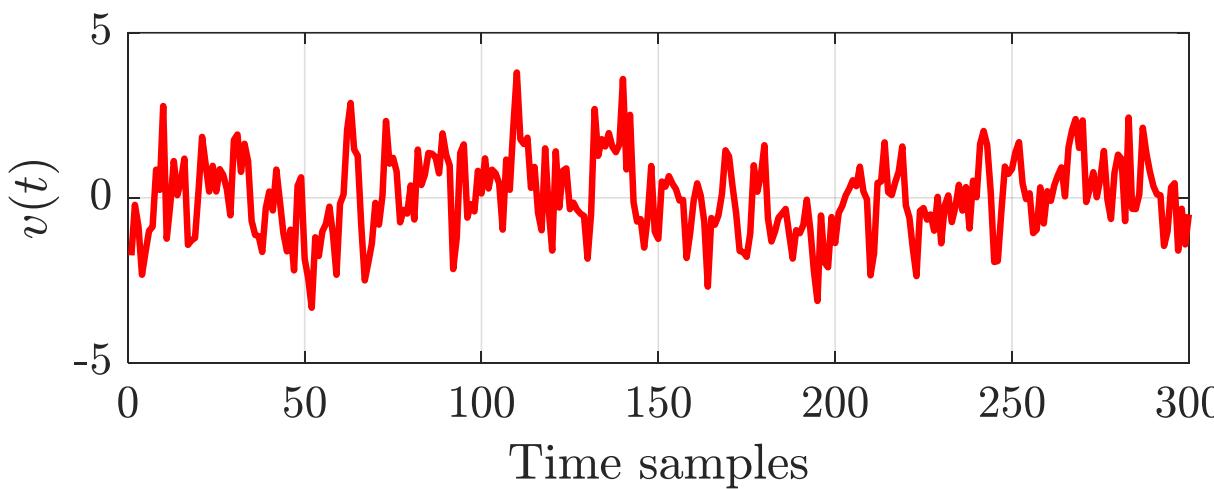
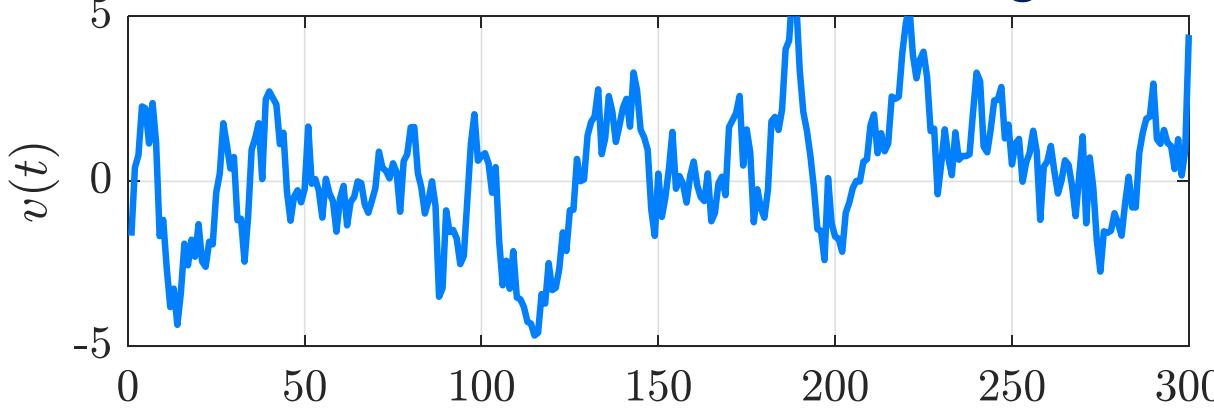
Densità spettrale «piatta»



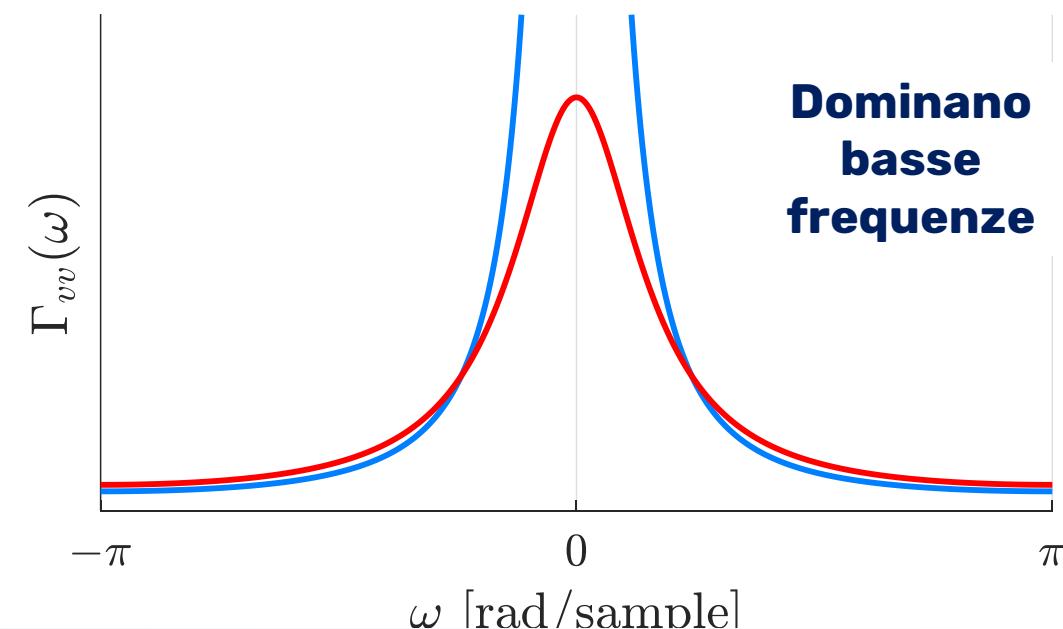
Esempio: processo «regolare»

Pss a media zero

Andamento regolare



Autocovarianza
decade
lentamente



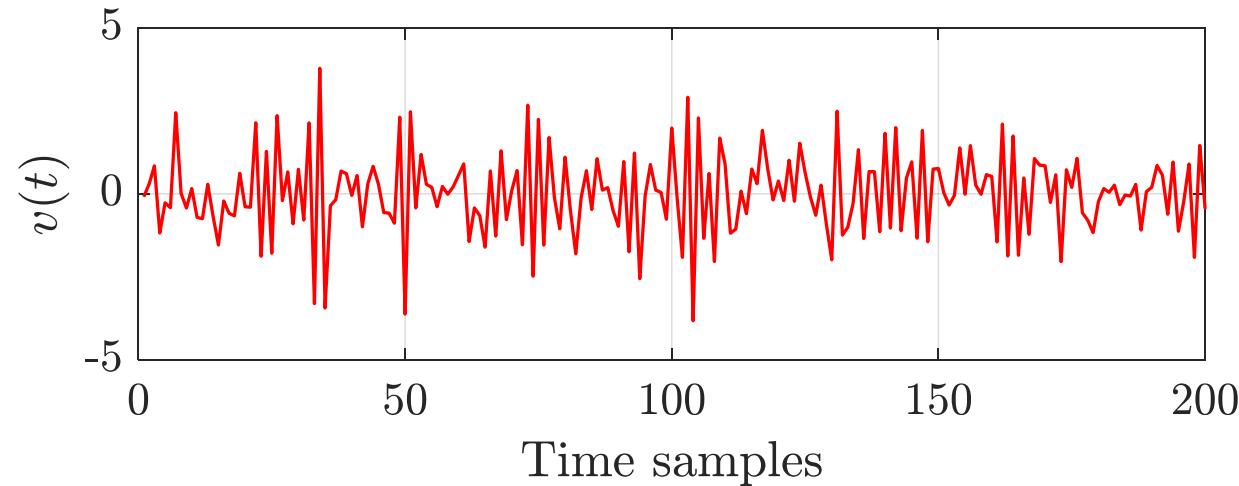
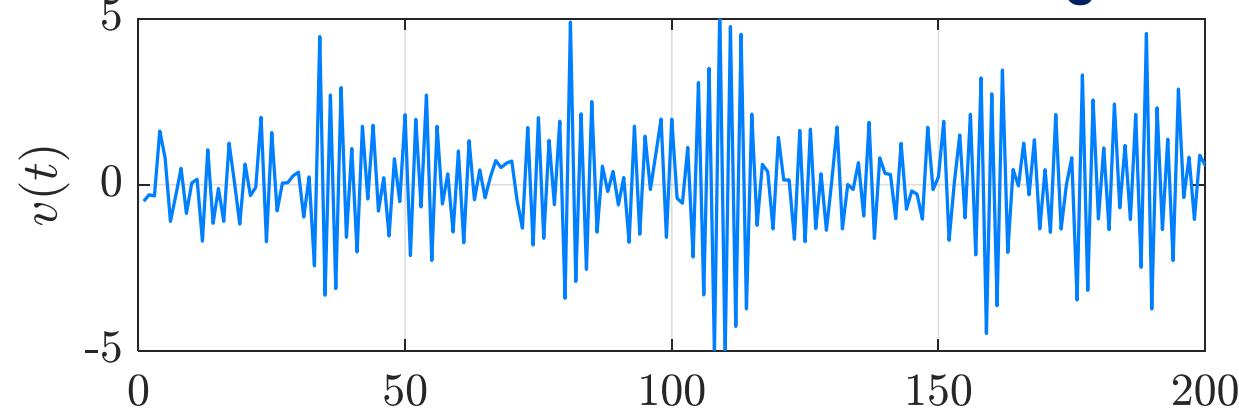
Dominano
basse
frequenze



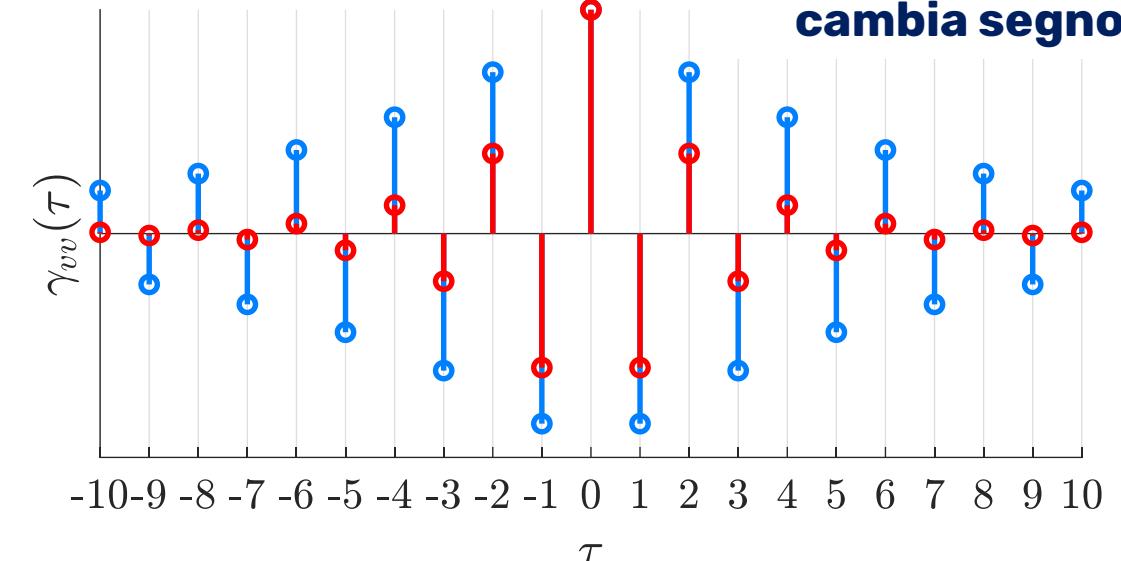
Esempio: processo «alternante»

Pss a media zero

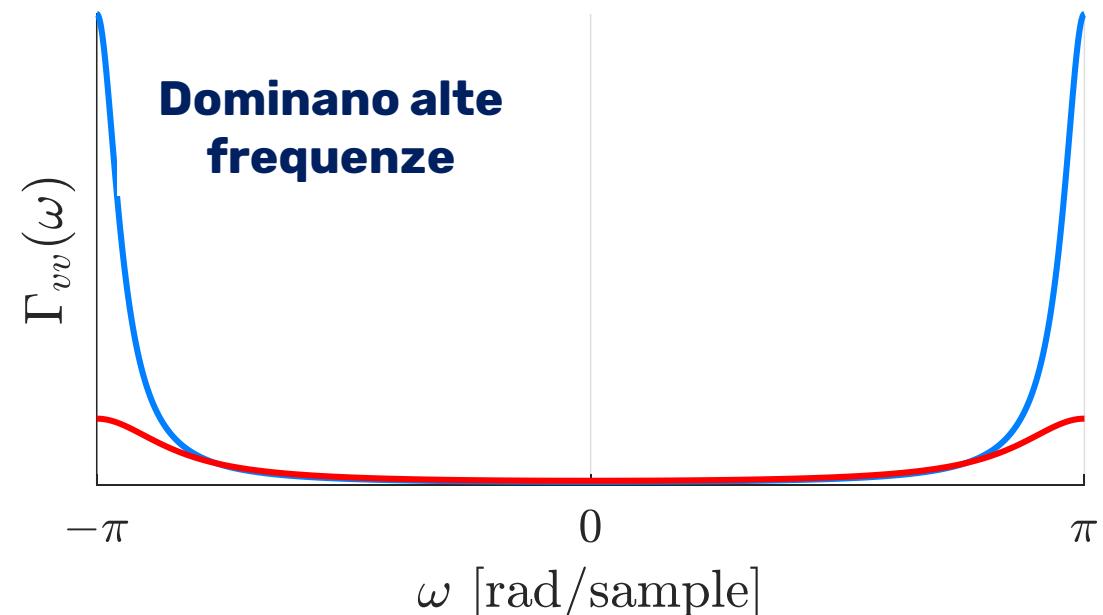
Andamento irregolare



Autocovarianza
cambia segno

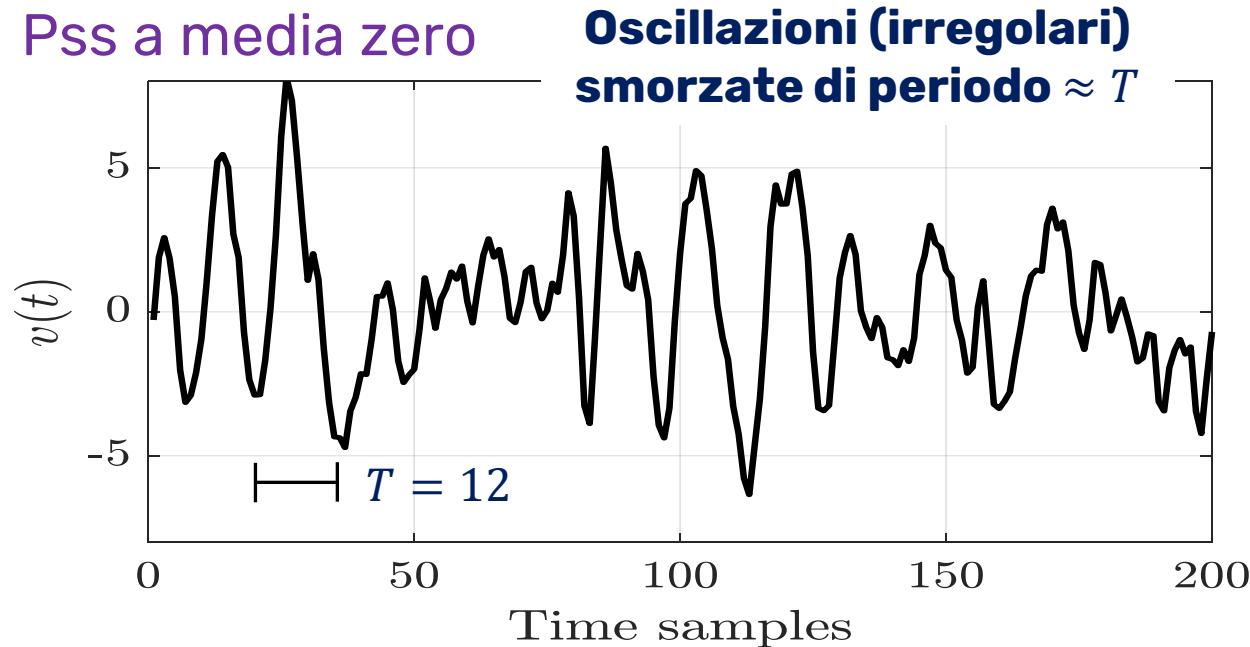


Dominano alte
frequenze

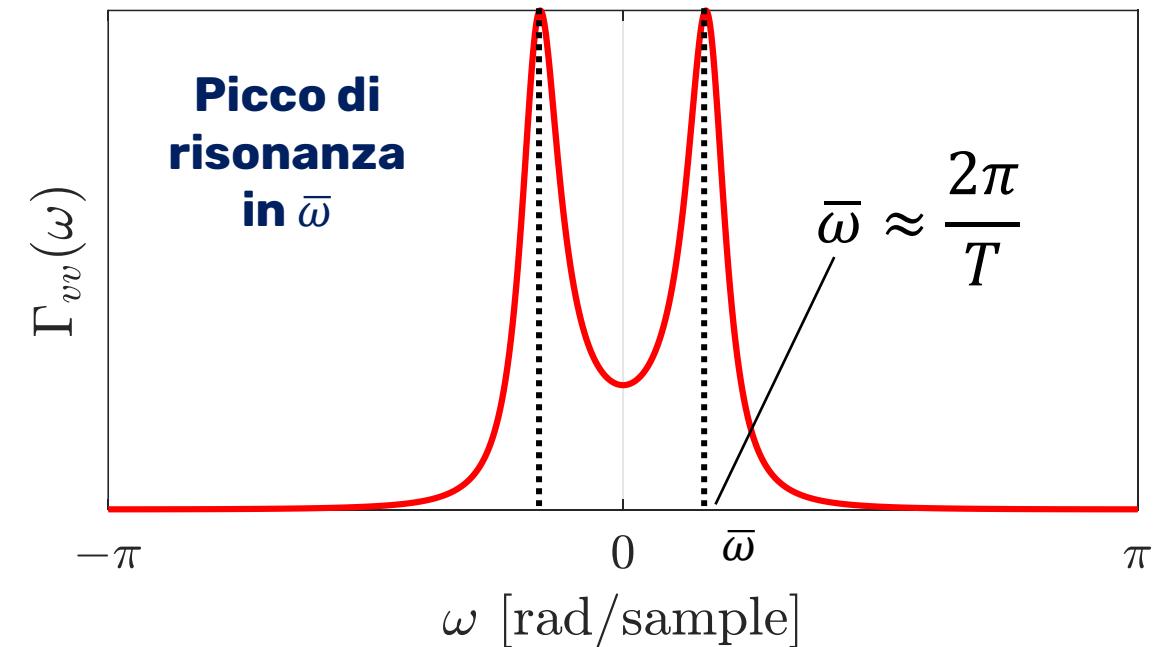
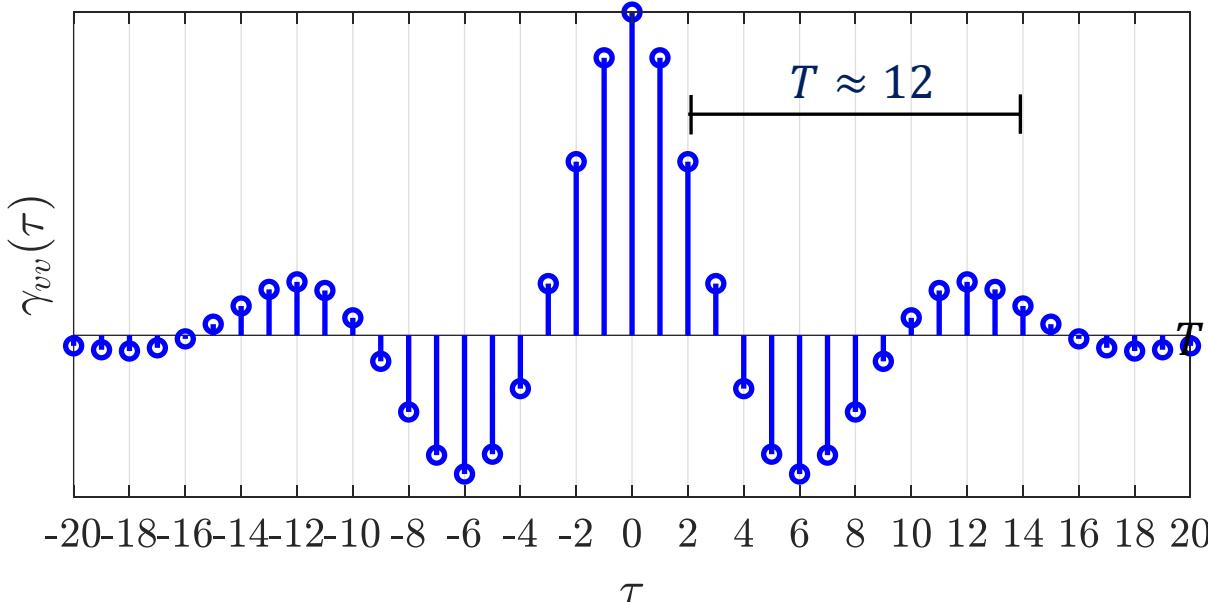


Esempio: processo con frequenza dominante

Pss a media zero



Oscillazioni (irregolari)
smorzate di periodo $\approx T$



Densità cross-spettrale

Dati due processi stocastici stazionari $v(t, s)$ e $x(t, s)$, definiamo la **densità di potenza cross-spettrale** (e la relativa trasformata \mathcal{Z}) come:

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

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

Proprietà

- $\gamma_{vx}(\tau) = \gamma_{xv}(-\tau)$
- $\Phi_{vx}(z) = \Phi_{xv}(z^{-1})$
- $\Gamma_{vx}(\omega) = \Gamma_{xv}(-\omega)$



Outline

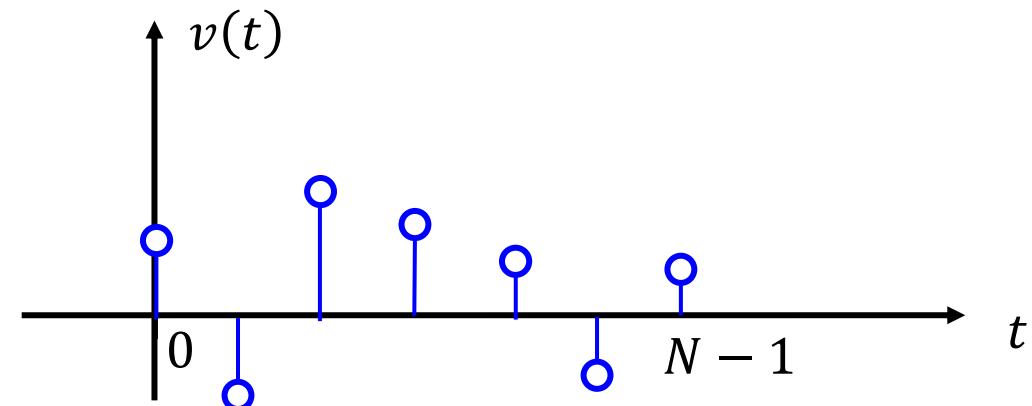
1. Introduzione alla stima di modelli dinamici
2. Processi stocastici
3. Processi stocastici stazionari
4. Momenti temporali ed ergodicità
5. Trasformata \mathcal{Z} e trasformata di Fourier
6. Densità spettrale di potenza
- 7. Stima spettrale**
8. Sistemi dinamici LTI discreti deterministici
9. Sistemi dinamici LTI discreti stocastici



Stima delle proprietà di un pss ergodico

Ipotesi:

- $v(t)$ processo stazionario **ergodico**
- $\mathbb{E}[v(t)] = m_v = 0$, infatti, se $m_v \neq 0$, posso stimare \hat{m}_v tramite una media temporale (grazie all'ergodicità) e analizzare $v(t) - \hat{m}_v$
- N dati disponibili: **una sola realizzazione** del processo $v(t)$, $0 \leq t \leq N - 1$



Stima delle proprietà di un pss ergodico

MEDIA (TEMPORALE) CAMPIONARIA

Abbiamo già visto questo stimatore parlando di ergodicità. Possiamo stimare il **valore atteso** m_ν di un pss ergodico $\nu(t)$ come

$$\hat{m}_\nu = \frac{1}{N} \sum_{t=0}^{N-1} \nu(t)$$

La **correttezza** dello stimatore si dimostra come nel caso di variabili casuali



Stima delle proprietà di un pss ergodico

FUNZIONE DI AUTOCOVARIANZA (TEMPORALE) CAMPIONARIA

Supponiamo che $v(t)$ sia un pss ergodico a media nulla.

Ricordandoci che $\gamma_{vv}(\tau) = \mathbb{E}_s[v(t)v(t + \tau)]$, possiamo stimare **l'autocovarianza** come

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

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



Stima delle proprietà di un pss ergodico

Osservazioni:

- Si dimostra che se $v(t)$ è Gaussiano, $\hat{\gamma}_{vv}(\tau)$ è lo stimatore a massima verosimiglianza
- $\mathbb{E}[\hat{\gamma}_{vv}(\tau)] = \gamma_{vv}(\tau)$, ovvero lo stimatore è **corretto**
- Per τ fissato, lo stimatore è **consistente**, sotto le ipotesi di ergodicità
- Per $\tau \approx N$, si ha che $\text{Var}[\hat{\gamma}_{vv}(\tau)]$ è grande perché ci sono pochi addendi

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



Stima delle proprietà di un pss ergodico

FUNZIONE DI AUTOCOVARIANZA (TEMPORALE) CAMPIONARIA – versione alternativa

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

Osservazioni:

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



Stima delle proprietà di un pss ergodico

Studiamo meglio il valore atteso di $\hat{\gamma}'_{vv}(\tau)$, ovvero $\mathbb{E}[\hat{\gamma}_{vv}(\tau)] = \frac{N - |\tau|}{N} \gamma_{vv}(\tau)$

Supponiamo di voler calcolare la stima per $\tau = N - 3, \tau = N - 2, \tau = N - 1$

$$\left\{ \begin{array}{l} \mathbb{E}[\hat{\gamma}'_{vv}(N - 3)] = \frac{3}{N} \gamma_{vv}(N - 3) \\ \\ \mathbb{E}[\hat{\gamma}'_{vv}(N - 2)] = \frac{2}{N} \gamma_{vv}(N - 2) \\ \\ \mathbb{E}[\hat{\gamma}'_{vv}(N - 1)] = \frac{1}{N} \gamma_{vv}(N - 1) \end{array} \right.$$

Per $\tau \approx N$, il valore atteso dello stimatore viene **«schiaffiato verso il basso»** (cosa che non succedeva con lo stimatore corretto $\hat{\gamma}_{vv}(\tau)$)

Lo stimatore non corretto $\hat{\gamma}'_{vv}(\tau)$ **peggiora il bias** ma **riduce la varianza** (le stime saranno «per più volte» vicine a valori piccoli. **Meglio tendere a zero** che dare i numeri del lotto. Inoltre, **per molti processi** si ha che $\lim_{\tau \rightarrow +\infty} \gamma_{vv}(\tau) = 0$



Densità spettrale campionaria

Sappiamo che la **densità spettrale di potenza** di un pss $v(t)$ è definita come

$\Gamma_{vv}(\omega) = \sum_{\tau=-\infty}^{\tau=+\infty} \gamma_{vv}(\tau) \cdot e^{-j\omega\tau}$. **Idea:** non conoscendo $\gamma_{vv}(\tau)$, uso $\hat{\gamma}_{vv}(\tau)$ oppure $\hat{\gamma}'_{vv}(\tau)$

Si definisce **periodogramma** il seguente **stimatore** della densità spettrale di potenza

$$I_N(\omega) \equiv \sum_{\tau=-(N-1)}^{N-1} \hat{\gamma}'_{vv}(\tau) \cdot e^{-j\omega\tau}$$

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



Densità spettrale campionaria

Proprietà: si dimostra come il periodogramma $I_N(\omega)$ è proporzionale al quadrato del modulo della DTFT $\mathcal{F}[v(t)]$ della realizzazione misurata del pss $v(t)$

$$I_N(\omega) = \frac{1}{N} |V(e^{j\omega})|^2$$

Per **segnali di durata finita** (ovvero tutti quelli che possiamo avere a disposizione in pratica), la **DFT** è un campionamento della **DTFT**. Per cui, «accontentandomi» di un campionamento del periodogramma in una griglia di frequenze, posso calcolare

$$\check{I}_N(k) = \frac{1}{N} |V(e^{j \cdot k \cdot 2\pi/N})|^2, \quad k = 0, 1, \dots, N - 1$$

In Matlab:
`abs(fft(v)) .^2`



Densità spettrale campionaria

Osservazioni

- Lo stimatore $I_N(\omega)$ **non è corretto**, ma è **asintoticamente corretto**. Infatti

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

Notiamo che non lo sarebbe stato neanche se avessi usato $\hat{\gamma}_{vv}(\tau)$ al posto di $\hat{\gamma}'_{vv}(\tau)$

- Si dimostra come $\text{Var}[I_N(\omega)] \approx \Gamma_{vv}^2(\omega)$. Per cui, la varianza dello stimatore non decresce al crescere di N . Lo stimatore **non è consistente**
- Per $N \rightarrow +\infty$, $I_N(\omega_1)$ 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»)



Stimatori della densità spettrale «smussati»

Lo stimatore dello spettro non gode di buone proprietà. Un metodo semplice ma efficace per migliorare la stima (riducendone la varianza a scapito del bias) è quello di **«regolarizzare»** la stima facendo la **media di diversi periodogrammi**

Metodo di Bartlett: Ipotizziamo di avere N dati a disposizione

- dividiamo questi dati in $K = N/M$ parti, dove M è la lunghezza di ogni porzione di dati
- calcoliamo il periodogramma $I_{M,K}^{[i]}(\omega)$ per ciascuna parte $i = 1, 2, \dots, K$
- facciamo la media dei periodogrammi, ottenendo la stima

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



Stimatori della densità spettrale «smussati»

Osservazioni

- Se $\gamma_{vv}(\tau) \rightarrow 0$ in modo sufficientemente rapido, i K periodogrammi sono **circa indipendenti**. In questo caso, si ha che $\text{Var}[\bar{I}_{M,K}(\omega)] = O\left(\frac{1}{K}\Gamma_{vv}^2(\omega)\right)$
- Il $\text{Bias}[\bar{I}_{M,K}(\omega)]$ è maggiore rispetto a quello di $I_N(\omega)$. Questo comporta una maggior **perdita di risoluzione in frequenza**
- Se so che $\Gamma_{vv}(\omega)$ ha **picchi molto stretti**, devo usare M **grande** in modo da avere abbastanza risoluzione in frequenza (un po' come avviene con la DFT)



Esempio: stima spettrale

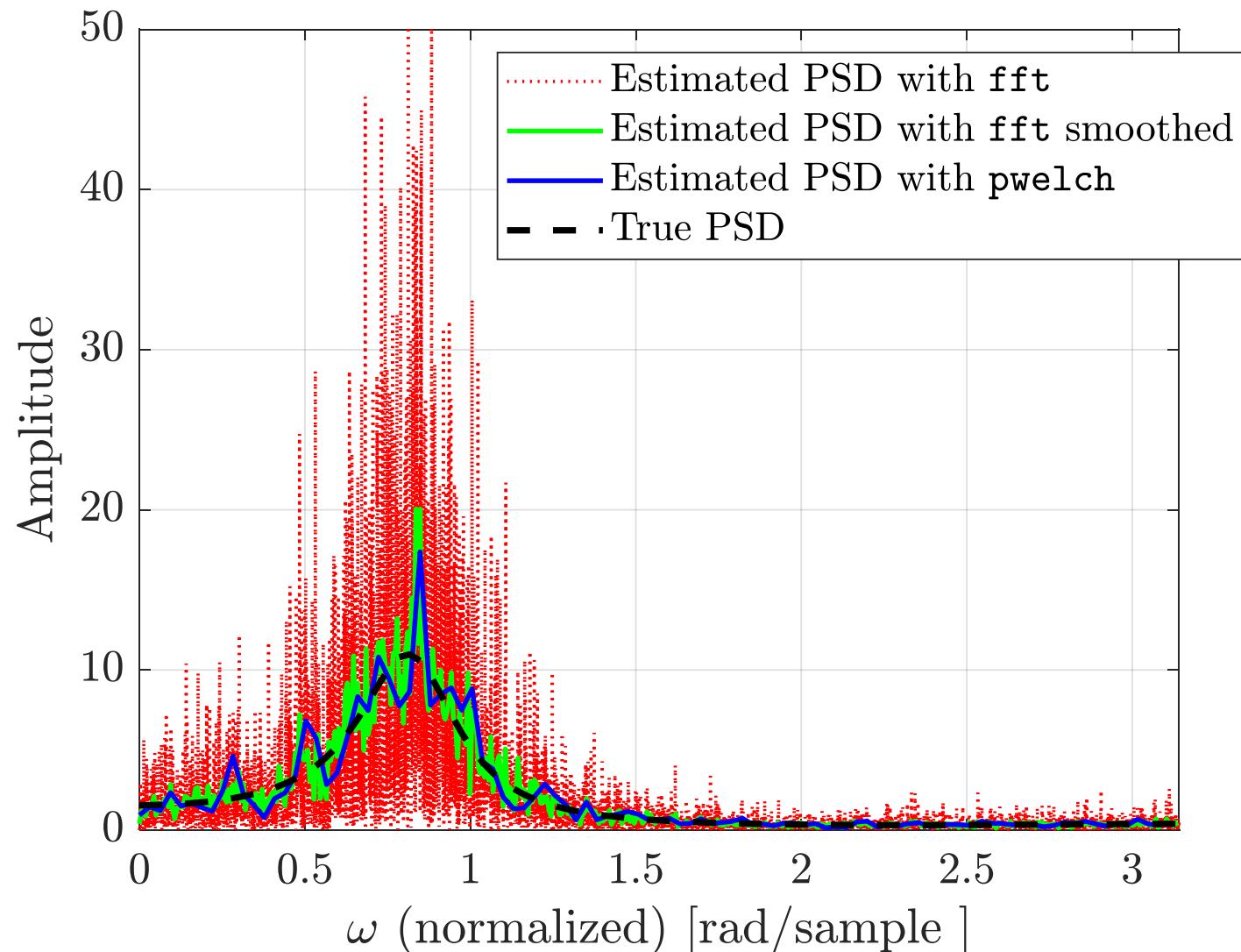
Consideriamo il seguente pss:

$$y(t) = 0.7y(t-1) + 0.2y(t-2) + 0.3y(t-3) + e(t)$$

$$e(t) \sim WN(0,1)$$

Con $T_s = 0.02$ s, $N = 10000$ dati

Dividiamo i dati in $K = 10$ folds per stimare la densità spettrale di potenza con il metodo di Bartlett



Outline

1. Introduzione alla stima di modelli dinamici
2. Processi stocastici
3. Processi stocastici stazionari
4. Momenti temporali ed ergodicità
5. Trasformata \mathcal{Z} e trasformata di Fourier
6. Densità spettrale di potenza
7. Stima spettrale
- 8. Sistemi dinamici LTI discreti deterministici**
9. Sistemi dinamici LTI discreti stocastici



Sistemi dinamici LTI discreti deterministici

L'obiettivo di questa seconda parte del corso è identificare (stimare) un modello di un **sistema dinamico**. Ci concentreremo su sistemi **Lineari Tempo Invarianti (LTI)** a tempo discreto, **Single Input Single Output (SISO)**

Un sistema dinamico può essere rappresentato in **spazio di stato** oppure in forma **ingresso\uscita** (funzione di trasferimento). Ci concentremo sulla rappresentazione ingresso\uscita: l'obiettivo è quindi quello di **stimare la funzione di trasferimento**

SPAZIO DI STATO

$$\begin{cases} \begin{matrix} x(t+1) = & A \cdot x(t) + & B \cdot u(t) \\ n \times 1 & n \times n & n \times 1 \quad 1 \times 1 \end{matrix} \\ \\ \begin{matrix} y(t) = & C \cdot x(t) + & D \cdot u(t) \\ 1 \times 1 & 1 \times n & 1 \times 1 \end{matrix} \end{cases}$$

FUNZIONE DI TRASFERIMENTO

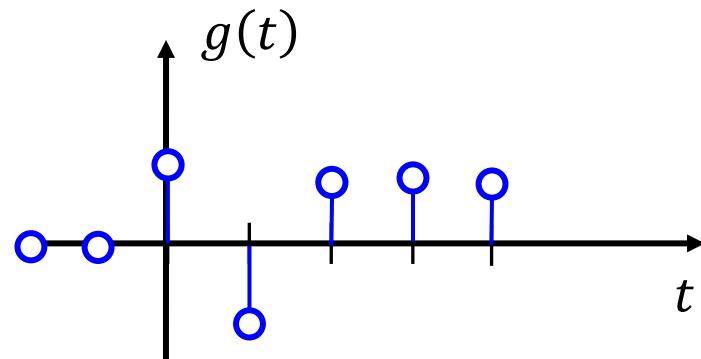
$$\frac{Y(z)}{U(z)} = G(z)$$



Sistemi dinamici LTI discreti deterministici

Definizione: un **sistema dinamico** (causale) è LTI se la sua uscita $y(t)$ può essere espressa tramite la **convoluzione** (discreta, causale) dell'input $u(t)$ e della **risposta all'impulso** $g(t)$ del sistema

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



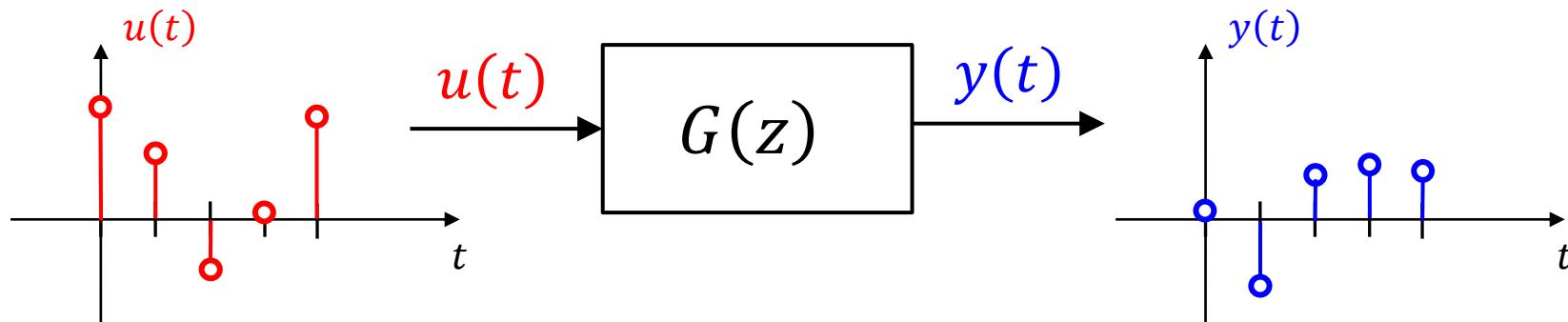
Facciamo l'ipotesi che $g(t) = 0$ per $t < 0$. Questa è un'ipotesi di **causalità**, che implica come l'ingresso $u(t)$ può solo influenzare l'uscita ad istanti $s \geq t$

Un primo modo per identificare un sistema dinamico è quello di applicare un impulso e stimare $g(t)$. Però, non è sempre possibile dare un ingresso impulsivo



Funzione di trasferimento

Consideriamo un sistema LTI SISO discreto. La **funzione di trasferimento** $G(z)$ descrive la relazione tra il segnale di ingresso $u(t)$ e il segnale di uscita $y(t)$, quando $x(0) = \mathbf{0}$



È possibile esprimere $G(z)$ come il rapporto tra la trasformata \mathcal{Z} di $u(t)$ e di $y(t)$, che equivale alla trasformata \mathcal{Z} della risposta all'impulso $g(t)$

$$G(z) = \sum_{t=0}^{+\infty} g(t) \cdot z^{-t} \quad \rightarrow \quad G(z) = \frac{\mathcal{Z}[y(t)]}{\mathcal{Z}[u(t)]} = \frac{Y(z)}{U(z)}$$

Quindi, $G(z)$ sarà il rapporto di due polinomi razionali in z



Funzione di trasferimento e forma ricorsiva

Supponiamo di avere la seguente funzione di trasferimento

$$G(z) = \frac{3z - 0.3}{z^2 - 0.3z - 0.1}$$

Possiamo scrivere

$$Y(z) = G(z)U(z) = \frac{3z - 0.3}{z^2 - 0.3z - 0.1} U(z) \quad \rightarrow \quad Y(z) = \frac{3z^{-1} - 0.3z^{-2}}{1 - 0.3z^{-1} - 0.1z^{-2}} U(z)$$

$$Y(z)[1 - 0.3z^{-1} - 0.1z^{-2}] = [3z^{-1} - 0.3z^{-2}]U(z) \quad \rightarrow$$

$$Y(z) - 0.3z^{-1} \cdot Y(z) - 0.1z^{-2} \cdot Y(z) = 3z^{-1} \cdot U(z) - 0.3z^{-2} \cdot U(z) \quad \rightarrow$$

Antitrasformando

$$y(t) = 0.3y(t-1) + 0.1y(t-2) + 3u(t-1) - 0.3u(t-2)$$



Funzione di trasferimento e forma ricorsiva

Nota: nel seguito, faremo uso di un piccolo abuso di notazione, scrivendo

$$y(t) = G(z)u(t) = \frac{3z - 0.3}{z^2 - 0.3z - 0.1} u(t)$$

Questo ci permetterà di «passare velocemente» dalla $G(z)$ alla rappresentazione ricorsiva

$$y(t) - 0.3z^{-1} \cdot y(t) - 0.1z^{-2} \cdot y(t) = 3z^{-1} \cdot u(t) - 0.3z^{-2} \cdot u(t)$$

$$y(t) = 0.3y(t-1) + 0.1y(t-2) + 3u(t-1) - 0.3u(t-2)$$



Rappresentazione dei sistemi LTI discreti

Riassumendo, possiamo rappresentare un sistema dinamico lineare LTI come

1) Spazio di stato

$$\begin{cases} x_1(t+1) = 0.1x_1(t) + 0.4x_2(t) + u(t) \\ x_2(t+1) = 0.3x_2(t) + 0.2x_1(t) \\ y(t) = 3x_1(t) + x_2(t) \end{cases}$$

$\xrightarrow{\mathcal{C}(zI_n - A)^{-1}B + D}$
 $\xleftarrow{\text{Realizzazione}}$

2) Funzione di trasferimento

$$G(z) = \frac{3z^{-1} - 0.3z^{-2}}{1 - 0.3z^{-1} - 0.1z^{-2}}$$

3) Forma ricorsiva (o di filtraggio)

$$y(t) = 0.3y(t-1) + 0.1y(t-2) + 3u(t-1) - 0.3u(t-2)$$

Lo spazio di stato è la rappresentazione più completa. La forma della funzione di trasferimento rappresenta solo gli stati che sono raggiungibili\osservabili dai segnali di ingresso\uscita, rispettivamente



Zeri e poli della funzione di trasferimento

I polinomi della funzione di trasferimento descrivono le proprietà del sistema dinamico

- **Zeri:** radici del numeratore
- **Poli:** radici del denominatore

$$G(z) = \frac{3z - 0.3}{z^2 - 0.3z - 0.1}$$

Diagram illustrating the transfer function $G(z)$. The numerator polynomial $3z - 0.3$ is highlighted with a blue dashed box and labeled "Zeri: radici del numeratore". The denominator polynomial $z^2 - 0.3z - 0.1$ is highlighted with a red dashed box and labeled "Poli: radici del denominatore". Arrows point from the definitions to their respective parts in the transfer function.

Definizione: Un sistema dinamico LTI a tempo discreto si dice **asintoticamente stabile** se i suoi **poli sono in modulo minore di 1**

$$z^2 - 0.3z - 0.1 \rightarrow \text{Poles: } z_1 = 0.5; z_2 = -0.2$$

$$|z_1| < 1 \text{ && } |z_2| < 1$$

**Sistema
asintoticamente
stabile**

La stabilità asintotica implica che l'output del sistema abbia un'«energia limitata», dato un input di «energia limitata»

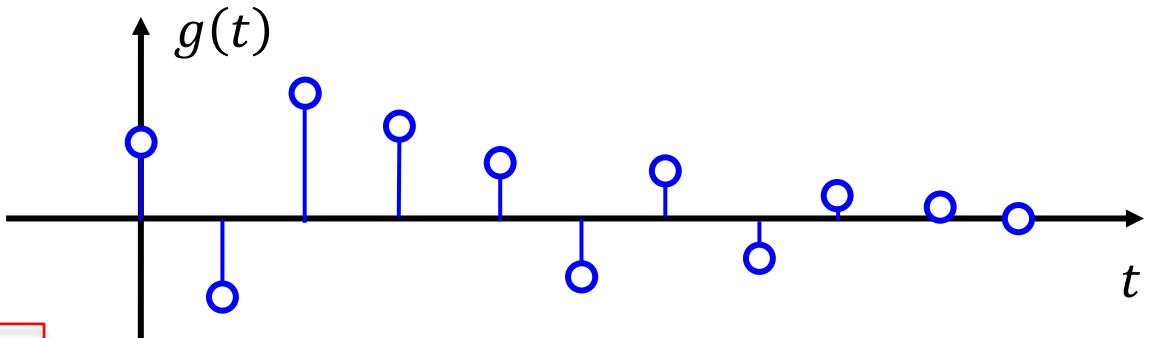
Se un sistema è in uno stato di equilibrio stabile, vi tornerà dopo una perturbazione



Guadagno

Una conseguenza della **asintotica stabilità** è che la **risposta all'impulso** tende **esponenzialmente a zero** per $t \rightarrow +\infty$

$$\lim_{t \rightarrow +\infty} g(t) = 0$$

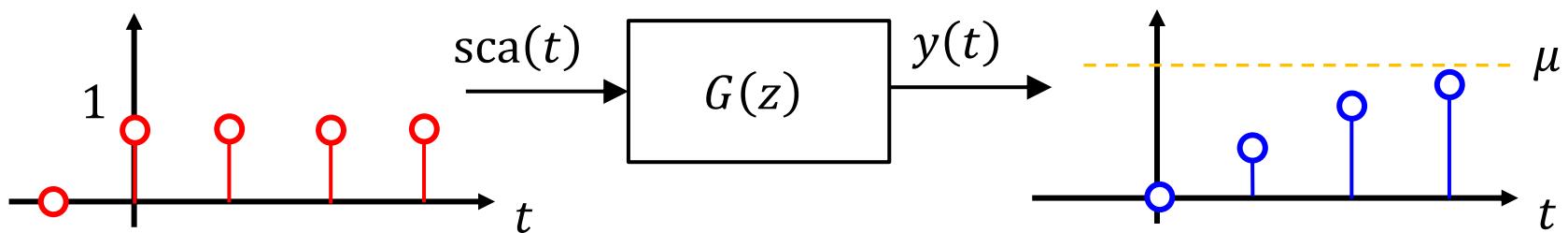


Guadagno del sistema: $\mu = \sum_{t=0}^{+\infty} g(t) = G(1)$

«Area» della risposta impulsiva

Proprietà: se applico $u(t) = \text{sca}(t)$, e il sistema è asintoticamente stabile ($\mu < \infty$), allora

$$\lim_{t \rightarrow +\infty} y(t) = \mu$$



Risposta in frequenza

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

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

Ampiezza Frequenza Fase

Con periodo di campionamento T_s , la **frequenza di Nyquist** è: $f_{Nyq} = \frac{f_s}{2} = \frac{1}{2 \cdot T_s}$

Per poter campionare correttamente è necessario utilizzare una frequenza di campionamento $f_s = 1/T_s$ «sufficientemente alta». La frequenza sinusoidale deve rispettare il **criterio di Nyquist (teorema di campionamento)**



Risposta in frequenza di sistemi LTI

Sia $G(z)$ la funzione di trasferimento di un sistema dinamico **asintoticamente stabile**.

Consideriamo un input sinusoidale del tipo $u(t) = A \cdot \sin(2\pi T_s t \cdot f + \varphi)$

Il segnale di output sarà: $y(t) = \tilde{y}(t) + \bar{A} \cdot \sin(2\pi T_s t \cdot f + \bar{\varphi})$

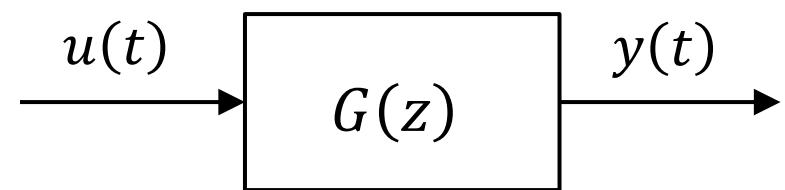
tale che:

Transitorio

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

**Effetto del guadagno
del sistema**

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



**Effetto dello sfasamento
indotto dal sistema**

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



Risposta in frequenza di sistemi LTI

Valutando $G(z)$ in $z = e^{j\omega T_s}$ si ottiene la **risposta in frequenza (FRF)** del sistema

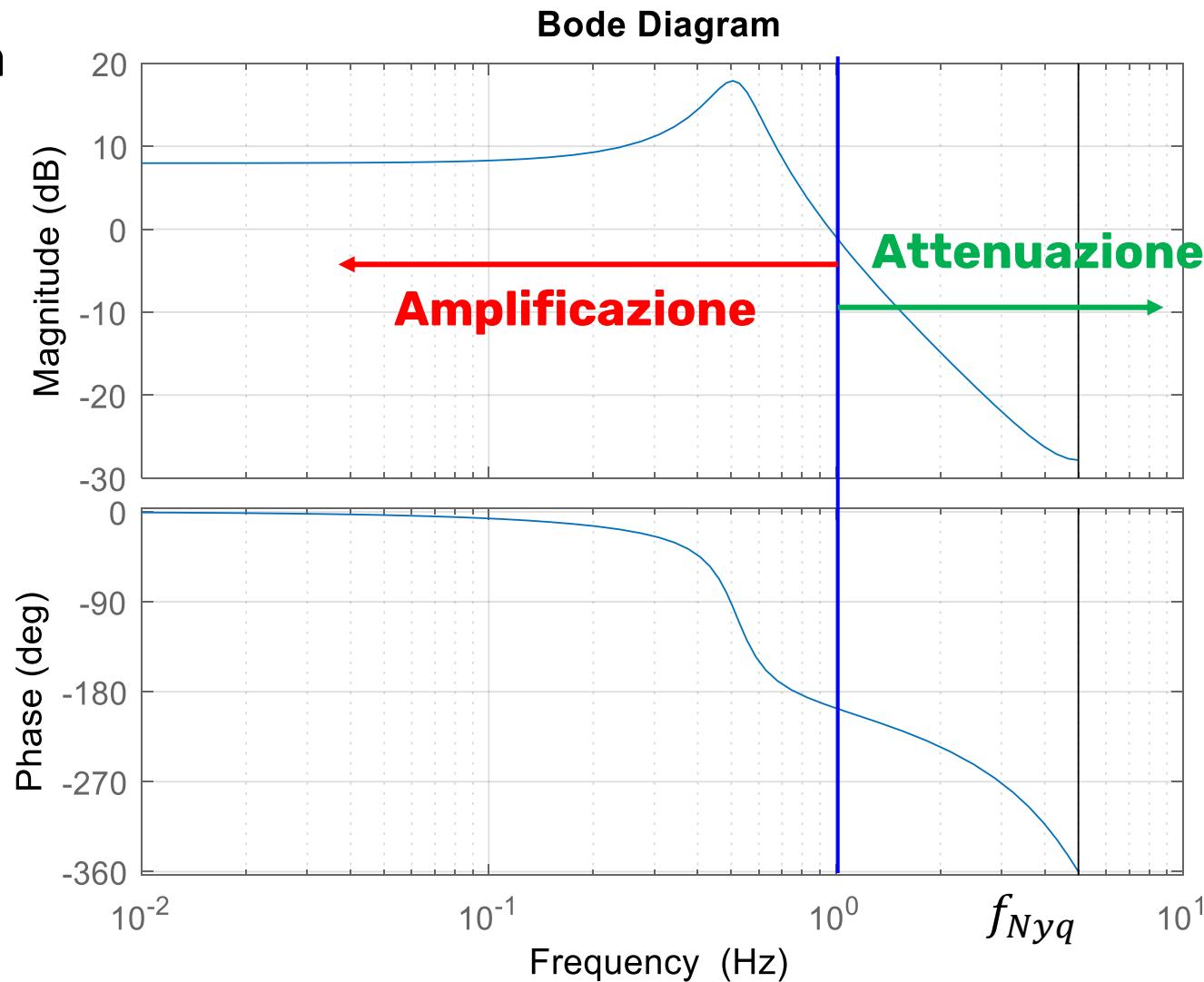
Modulo della FRF

$$|G(e^{j\omega})|$$

Rendiamo implicito
 T_s per semplicità

Fase della FRF

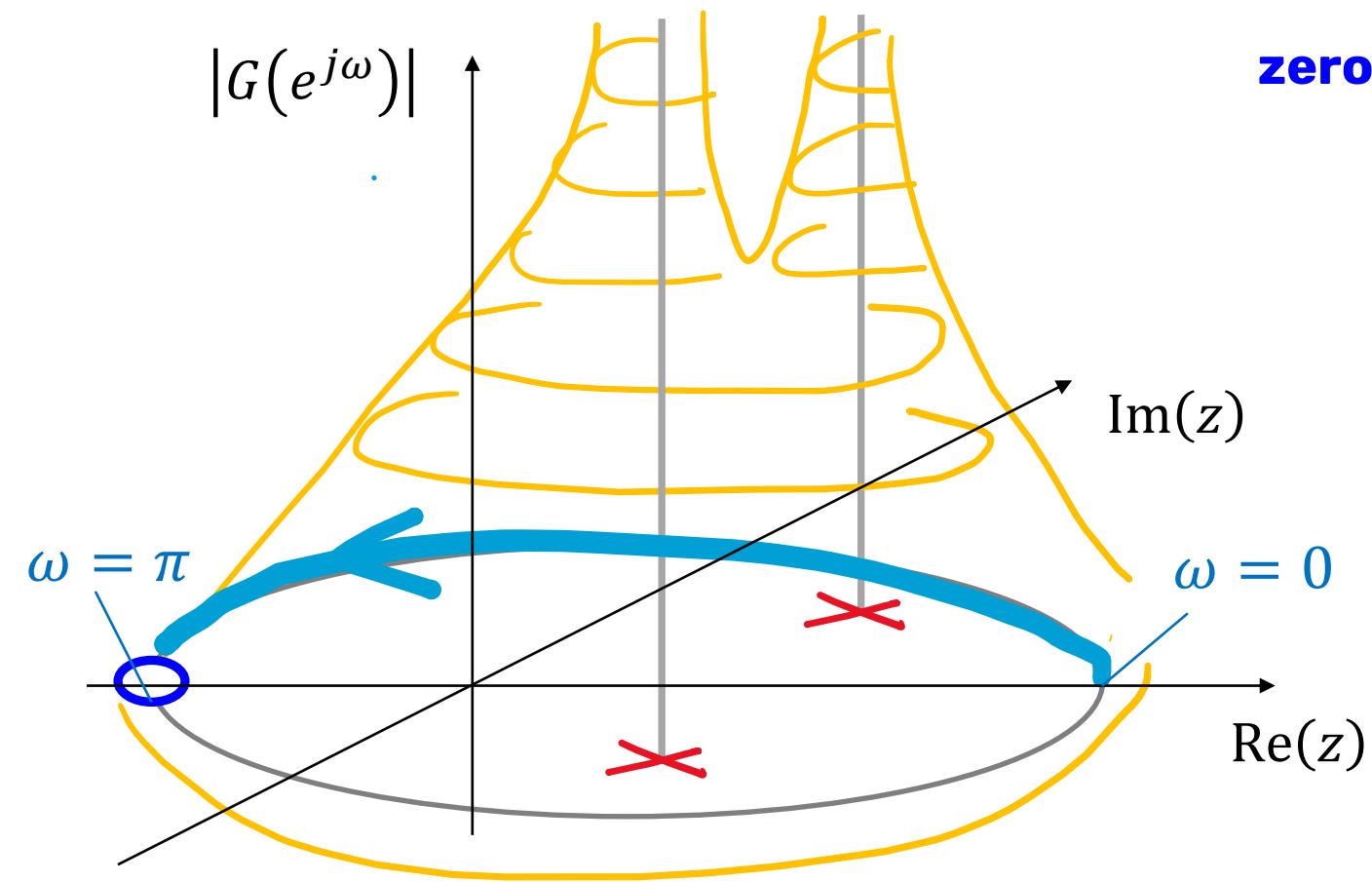
$$\angle G(e^{j\omega})$$



Risposta in frequenza di sistemi LTI

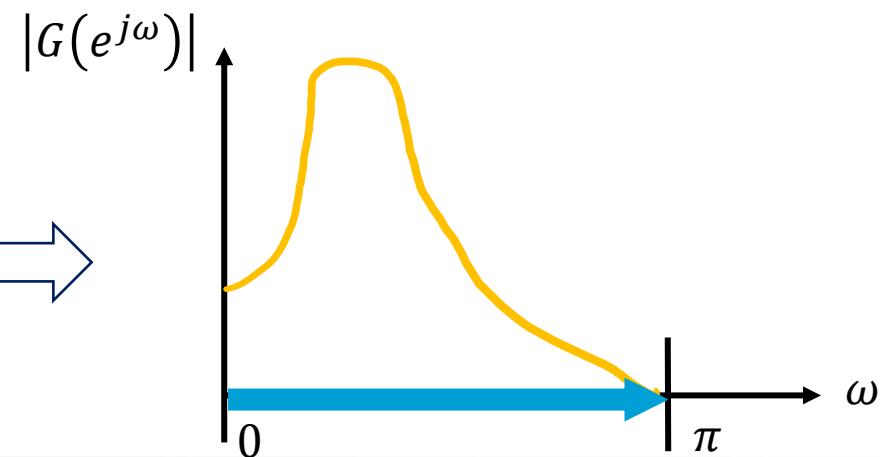
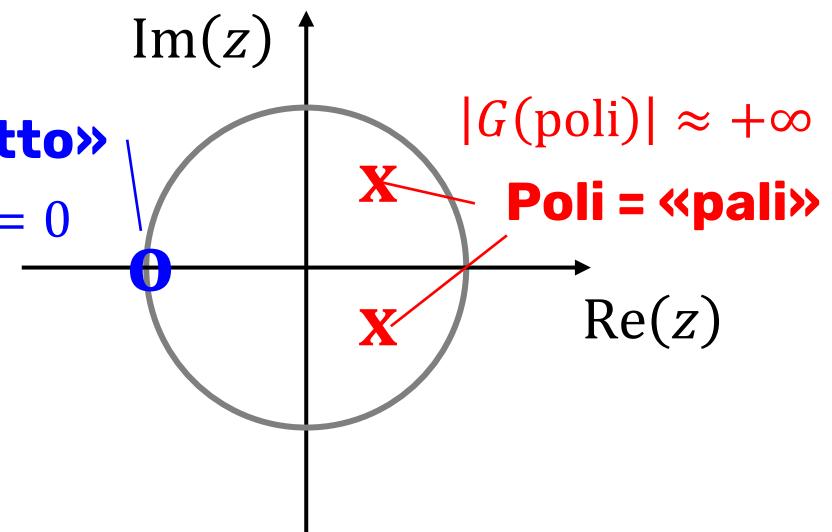
Se conosco la posizione dei **poli** e degli **zeri** posso farmi un'idea della forma di $|G(e^{j\omega})|$

Similitudine del «**tendone del circo**»



zero = «picchetto»

$$|G(\text{zero})| = 0$$



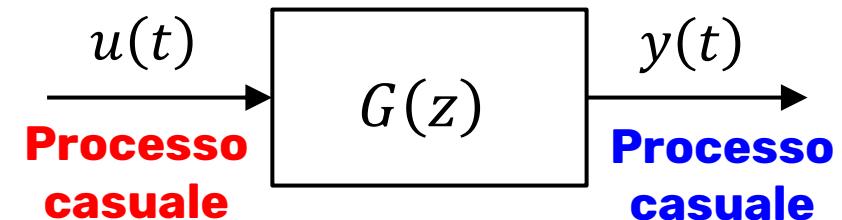
Outline

1. Introduzione alla stima di modelli dinamici
2. Processi stocastici
3. Processi stocastici stazionari
4. Momenti temporali ed ergodicità
5. Trasformata \mathcal{Z} e trasformata di Fourier
6. Densità spettrale di potenza
7. Stima spettrale
8. Sistemi dinamici LTI discreti deterministici
- 9. Sistemi dinamici LTI discreti stocastici**



Sistemi LTI discreti con ingressi stocastici

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



Quali sono le proprietà di $y(t)$?

VALORE ATTESO
$$\begin{aligned}\mathbb{E}[y(t)] &= \sum_{i=0}^{+\infty} g(i)\mathbb{E}[u(t-i)] = G(1) \cdot m_u \\ &= \mu \cdot m_u\end{aligned}$$

Il valore atteso di $y(t)$ non dipende da t !



Sistemi LTI discreti con ingressi stocastici

AUTOCOVARIANZA (per semplicità consideriamo $m_u = 0$, dato che l'espressione dell'autocovarianza non dipende dalla media del processo)

$$y(t) = \sum_{i=0}^{+\infty} g(i)u(t-i) \quad \Rightarrow \quad y(t+\tau) = \sum_{i=0}^{+\infty} g(i)u(t-i+\tau)$$

$\gamma_{uu}(\tau)$ non dipende da t
perché $u(t)$ è stazionario
per ipotesi

$$u(t)y(t+\tau) = \sum_{i=0}^{+\infty} u(t) \cdot g(i)u(t-i+\tau) \quad \Rightarrow \quad \text{applico } \mathbb{E}[\quad]$$

$$\gamma_{uy}(t, t+\tau) = \sum_{i=0}^{+\infty} g(i)\gamma_{uu}(t, t-i+\tau)$$



$\gamma_{uy}(\tau)$ non
dipende da t

$$\gamma_{uy}(\tau) = \sum_{i=0}^{+\infty} g(i)\gamma_{uu}(\tau - i)$$

$$\Gamma_{uy}(\omega) = G(e^{j\omega})\Gamma_{uu}(\omega)$$



Sistemi LTI discreti con ingressi stocastici

Analogamente a prima, si ricava che

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

applico $\mathbb{E}[\quad]$

$\gamma_{yu}(\tau)$ Non dipende da t , dato che neanche
 $\gamma_{uy}(\tau)$ dipende da t (si veda slide 24)

$$\gamma_{yy}(t, t + \tau) = \sum_{i=0}^{+\infty} g(i)\gamma_{yu}(t, t - i + \tau)$$

$$\gamma_{yy}(\tau) = \sum_{i=0}^{+\infty} g(i)\gamma_{yu}(\tau - i)$$

$$\Gamma_{yy}(\omega) = G(e^{j\omega})\Gamma_{yu}(\omega)$$

La funzione di autocovarianza di $y(t)$ non dipende da t !



Sistemi LTI discreti con ingressi stocastici

Teorema Sia $u(t)$ è un processo stocastico **stazionario** che alimenta un sistema dinamico **asintoticamente stabile**. Allora, anche $y(t)$ è un processo stocastico **stazionario**

Osservazioni

- Nella pratica, $u(t)$ viene applicato dall'istante $t = 0$ e non da $t = -\infty$, per cui $y(t)$ sarà stazionario **dopo un transitorio**

Questa è una condizione **necessaria e sufficiente**. A **regime**, per ogni condizione iniziale, $y(t)$ è un **pss** se valgono:

1. $u(t)$ è un pss
2. $G(z)$ è asintoticamente stabile



Densità spettrale di potenza dell'uscita

Teorema

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

$$\Phi_{yy}(z) = G(z)G(z^{-1}) \cdot \Phi_{uu}(z)$$

slide 69

slide 98

Dimostrazione $\Gamma_{yy}(\omega) = G(e^{j\omega})\Gamma_{yu}(\omega) = G(e^{j\omega})\Gamma_{uy}(-\omega) = G(e^{j\omega})G(e^{-j\omega}) \cdot \Gamma_{uu}(-\omega)$

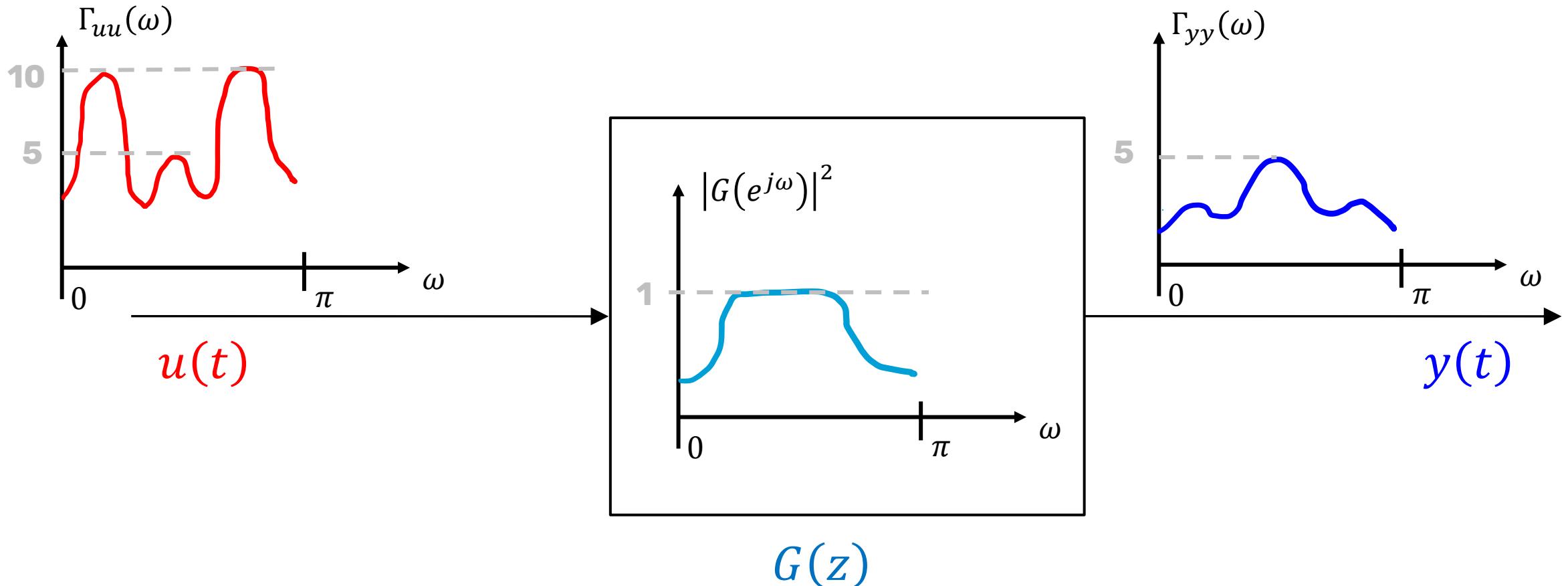
Complesso coniugato
 $= G(e^{j\omega})G(e^{j\omega})^* \cdot \Gamma_{uu}(\omega) = |G(e^{j\omega})|^2 \cdot \Gamma_{uu}(\omega)$
Funzione pari

$\Phi_{yy}(z) = G(z)\Phi_{yu}(z) = G(z)\Phi_{yu}(z^{-1}) = G(z)G(z^{-1}) \cdot \Phi_{uu}(z)$ Per la simmetria di $\gamma_{uu}(\tau)$



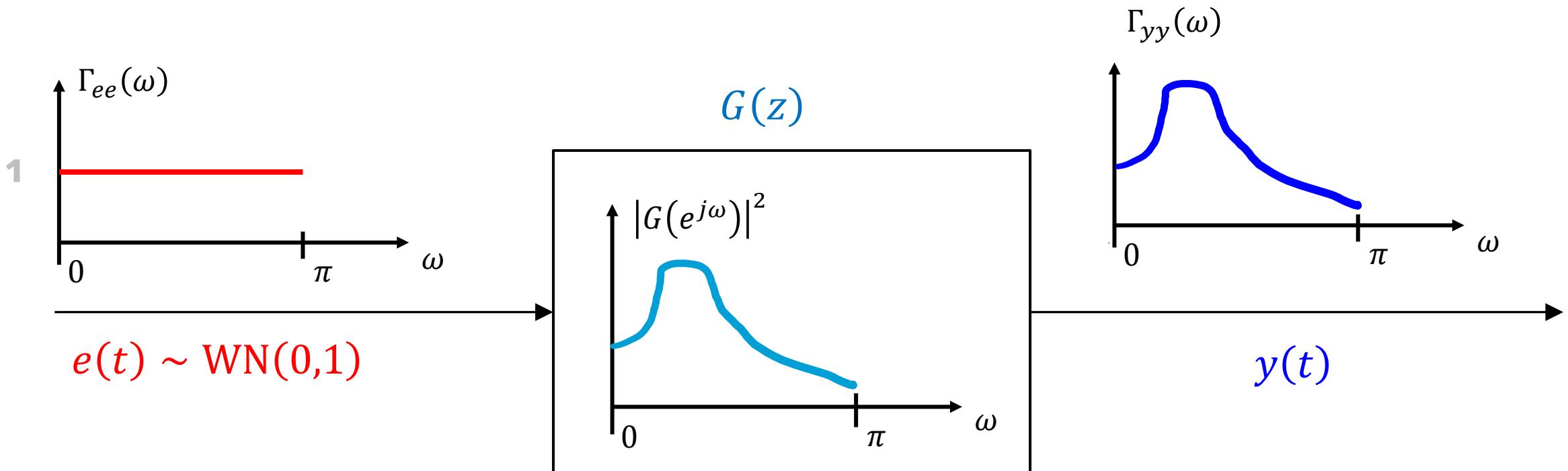
Densità spettrale di potenza dell'uscita

Possiamo dire $|G(e^{j\omega})|^2$ «modula» la **densità spettrale** di $u(t)$, ottenendo $y(t)$



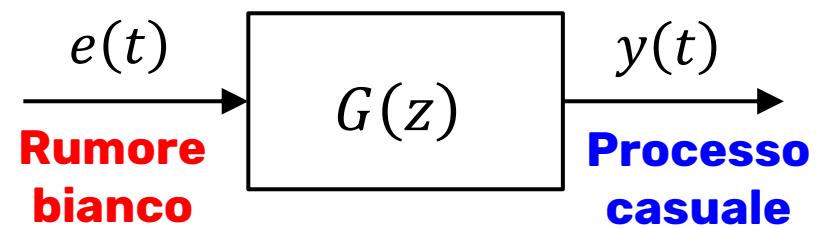
Rappresentazione dinamica di un pss

Il risultato precedente è molto importante! Infatti, ci dice che possiamo **interpretare** un processo stocastico stazionario $y(t)$ come **l'uscita di un sistema dinamico** $G(z)$ **asintoticamente stabile** alimentato da **rumore bianco**, tale che $\Gamma_{yy}(\omega) = |G(e^{j\omega})|^2$



Rappresentazione dinamica di un pss

Ne segue che, data $G(z)$ asintoticamente stabile, è **possibile esprimere un qualunque processo stocastico stazionario** $y(t)$ come **combinazione lineare di infiniti campioni di rumore bianco**



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

Vedremo nella lezione 9 che questo modello si chiama MA(∞)

$$= g(0)e(t) + g(1)e(t-1) + g(2)e(t-2) + \dots$$



Rappresentazione dinamica di un pss

Se conosco (oppure stimo) $\Gamma_{yy}(\omega)$, e se riesco a trovare $G(z)$ asintoticamente stabile e causale tale che $\Gamma_{yy}(\omega) = |G(e^{j\omega})|^2$, posso anche **simulare** diverse realizzazioni del processo $y(t)$, generando al computer delle sequenze di variabili casuali incorrelate, che fungono da rumore bianco $e(t)$

Esempio: simulare il vento

Se volessi simulare il vento, allora, dato lo spettro seguente, devo trovare una $G(z)$, «un tendone del circo» tale che abbia il profilo desiderato

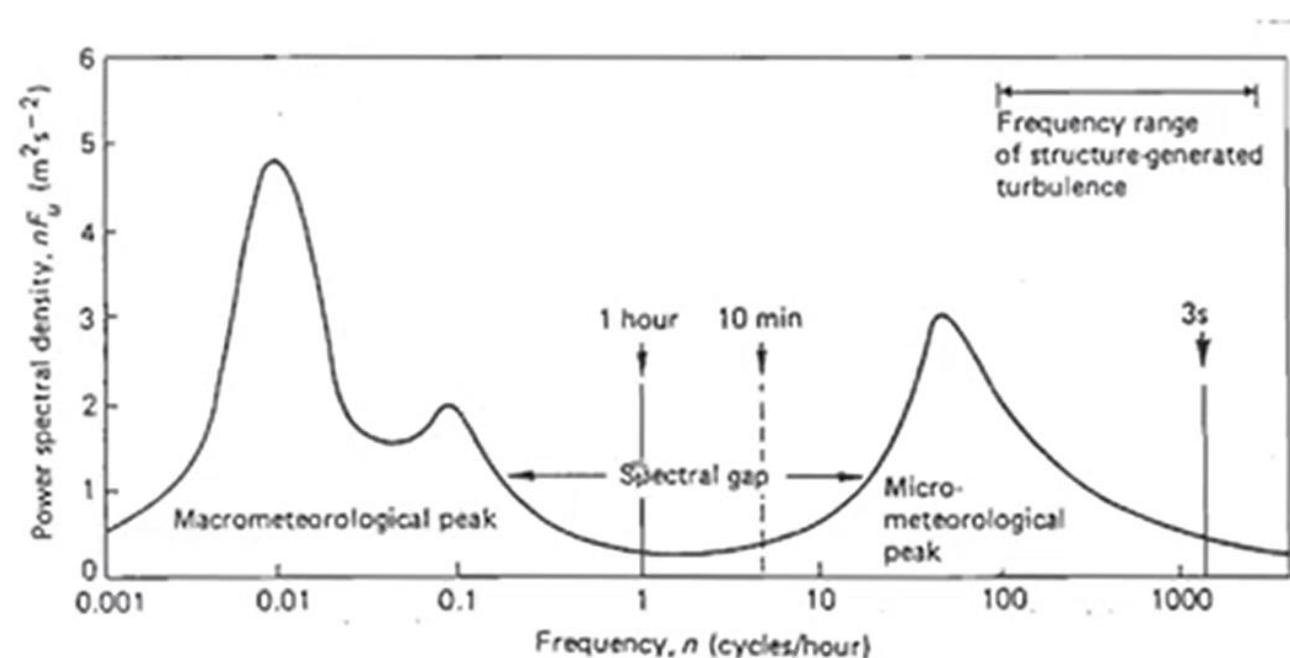
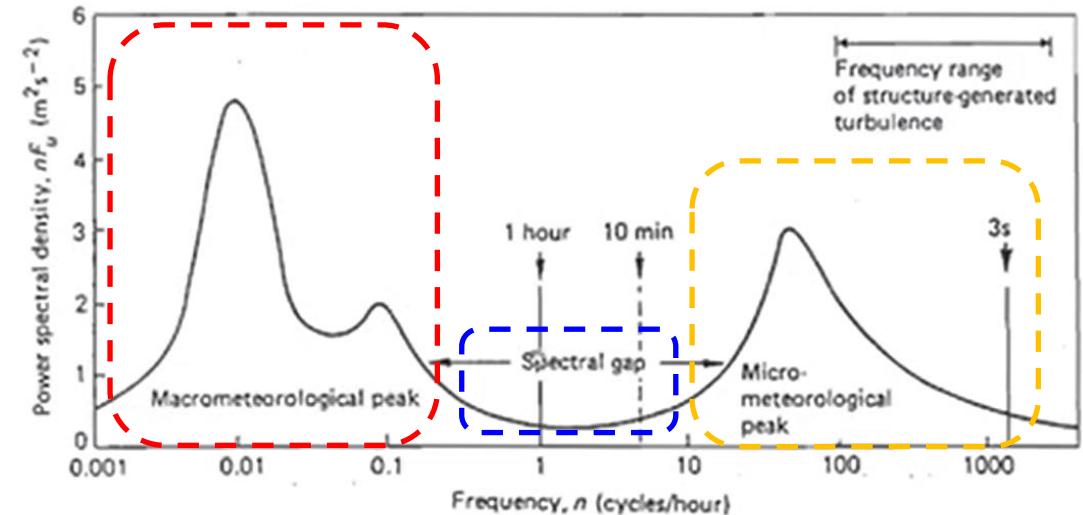
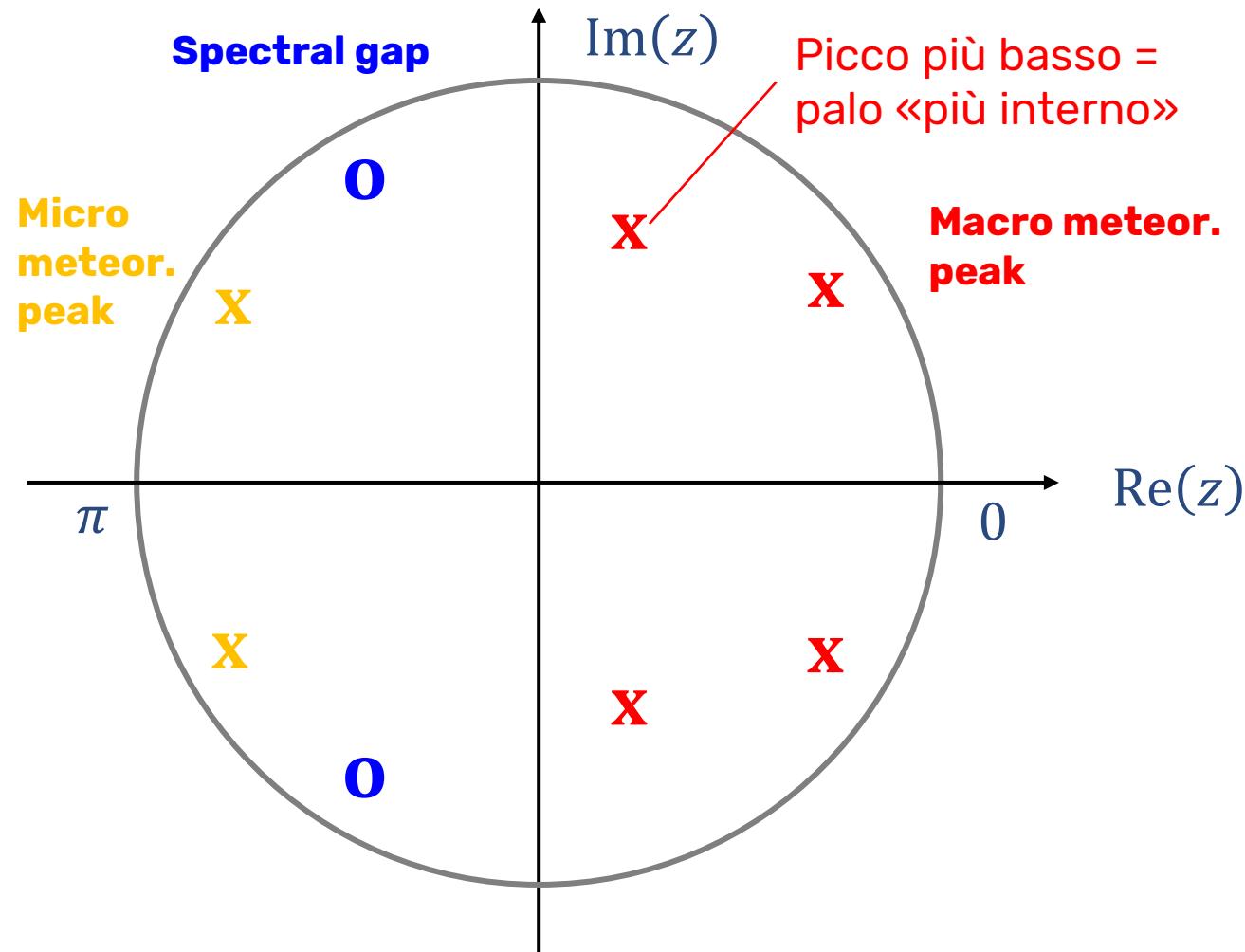


Fig. 1.2.1 Densità spettrale di potenza della velocità orizzontale del vento
(da Cook, 1985)



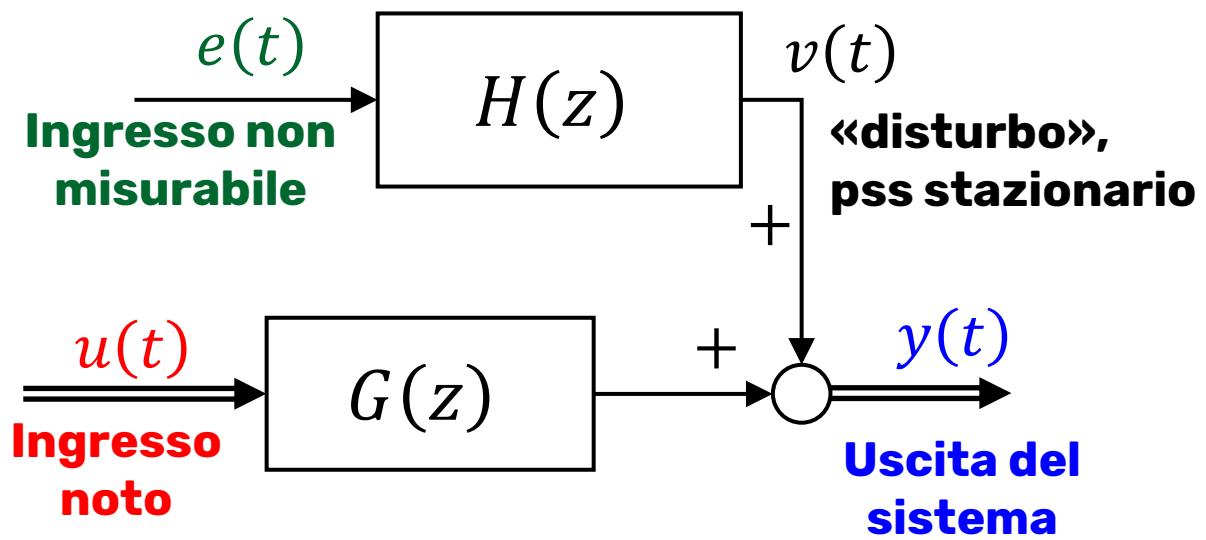
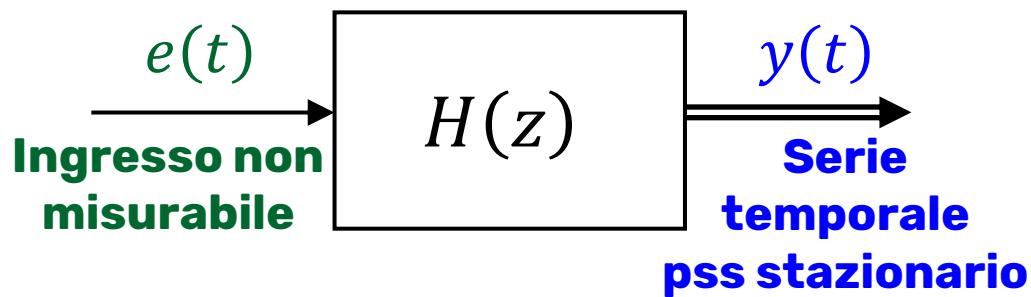
Esempio: simulare il vento



Per simulare il vento, ho bisogno di **almeno 3 coppie di poli** complessi coniugati, e quindi **l'ordine** della $G(z)$ deve essere **almeno 6**

Modellazione di serie temporali e sistemi dinamici

Riprendendo quanto detto a inizio lezione, abbiamo quindi che l'ingresso esogeno non misurabile sarà proprio in **rumore bianco** $e(t)$. L'obiettivo sarà ottenere una **stima delle funzioni di trasferimento** $H(z)$ e $G(z)$



Nota: Nel caso di sistemi dinamici, $G(z)$ rappresenta un **sistema dinamico fisico, reale**. $H(z)$ e $e(t)$ **non esistono fisicamente**: sono solo uno *strumento matematico* per modellare ciò che $G(z)$ non riesce a catturare della relazione tra $u(t)$ e $y(t)$

Depolarizzazione

La depolarizzazione consiste nel **rimuovere il valore atteso** m ad un processo stocastico stazionario $v(t)$. È utile per semplificare il calcolo della funzione di autocovarianza

$$\gamma_{vv}(\tau) = \mathbb{E}[(v(t) - m) \cdot (v(t + \tau) - m)]$$

Se avessimo $m = 0$, il calcolo diventerebbe $\gamma_{vv}(\tau) = \mathbb{E}[v(t) \cdot v(t + \tau)]$

Definiamo quindi $\tilde{v}(t) = v(t) - m$. Abbiamo che:

- $\mathbb{E}[\tilde{v}(t)] = \mathbb{E}[v(t) - m] = \mathbb{E}[v(t)] - m = m - m = 0$
- $\tilde{\gamma}_{vv}(\tau) = \mathbb{E}[\tilde{v}(t)\tilde{v}(t + \tau)] = \mathbb{E}[(v(t) - m) \cdot (v(t + \tau) - m)] = \gamma_{vv}(\tau)$

I processi $v(t)$ e $\tilde{v}(t)$ hanno la **stessa autocovarianza** (e quindi le stesse caratteristiche spettrali). Non **lede alcuna generalità studiare processi a media nulla**





**UNIVERSITÀ
DEGLI STUDI
DI BERGAMO**

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



IDENTIFICAZIONE DEI MODELLI E ANALISI DEI DATI (IMAD)

Lezione 9: Famiglie di modelli stocastici

Corso di Laurea Magistrale in
INGEGNERIA INFORMATICA

SPEAKER
Prof. Mirko Mazzoleni

PLACE
Università degli Studi di
Bergamo

Syllabus

Parte II: sistemi dinamici

8. Processi stocastici

- 8.1 Processi stocastici stazionari (pss)
- 8.3 Rappresentazione spettrale di un pss
- 8.4 Stimatori campionari media\covarianza
- 8.5 Densità spettrale campionaria

9. Famiglie di modelli a spettro razionale

- 9.1 Modelli per serie temporali (MA, AR, ARMA)
- 9.2 Modelli per sistemi input/output (ARX, ARMAX)

10. Predizione

- 10.1 Filtro passa-tutto

10.2 Forma canonica

10.3 Teorema della fattorizzazione spettrale

10.4 Soluzione al problema della predizione

11. Identificazione

- 11.3 Identificazione di modelli ARX
- 11.4 Identificazione di modelli ARMAX
- 11.5 Metodo di Newton

12. Identificazione: analisi e complementi

- 12.1 Analisi asintotica metodi PEM
- 12.2 Identificabilità dei modelli
- 12.3 Valutazione dell'incertezza di stima

13. Identificazione: valutazione



Parte I: sistemi staticiStima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

- ✓ Stima parametri popolazione
- ✓ Stima modello lineare: minimi quadrati

○ **SI assunzioni su ddp dei dati**

- ✓ Stima massima verosimiglianza parametri popolazione
- ✓ Stima modello lineare: massiva verosimiglianza
- ✓ Regressione logistica

• θ variabile casuale○ **SI assunzioni su ddp dei dati**

- ✓ Stima Bayesiana

Parte II: sistemi dinamiciStima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

- ✓ Modelli lineari di pss
- ✓ Predizione
- ✓ Identificazione
- ✓ Persistente eccitazione
- ✓ Analisi asintotica metodi PEM
- ✓ Analisi incertezza stima (numero dati finito)
- ✓ Valutazione del modello

Machine learning

Outline

1. Famiglie di modelli a spettro razionale
2. Serie temporali: modelli MA, AR, ARMA
3. Sistemi ingresso\uscita: modelli ARX e ARMAX
4. Sistemi ingresso\uscita: modelli FIR, OE, BJ
5. Densità spettrale di potenza: esempio di calcolo



Outline

- 1. Famiglie di modelli a spettro razionale**
2. Serie temporali: modelli MA, AR, ARMA
3. Sistemi ingresso\uscita: modelli ARX e ARMAX
4. Sistemi ingresso\uscita: modelli FIR, OE, BJ
5. Densità spettrale di potenza: esempio di calcolo



Gli step per la risoluzione del problema

Seguiremo tre fasi per risolvere il problema della **modellazione di sistemi dinamici**:

Definizione delle **classi di modelli** \mathcal{M} di sistemi dinamici

Ci concentreremo su modelli di **sistemi dinamici lineari**, espressi da **funzioni di trasferimento razionali fratte**. I parametri ignoti sono i coefficienti dei polinomi al numeratore e denominatore

Predizione

Data una particolare classe di modello, supponendo di conoscerne il valore dei parametri, qual è il **preditore ottimo**? Quanto vale la predizione ottima?

Identificazione

Come **stimo il valore dei parametri** del modello scelto per la modellazione dei dati?



Famiglie di modelli a spettro razionale

I processi stocastici che si ottengono filtrando un **rumore bianco** tramite un **filtro asintoticamente stabile** $H(z) = C(z)/A(z)$ sono detti **processi a spettro razionale**, dove $C(z)$ e $A(z)$ sono *polinomi a coefficienti reali nella variabile z* (oppure z^{-1})

Di seguito, vedremo sia modelli di processi stocastici per **serie temporali**:

- MA (Moving Average)
- AR (AutoRegressive)
- ARMA (AutoRegressive Moving Average)

sia modelli di processi stocastici per **sistemi dinamici** (quindi con ingresso $u(t)$ noto)

- ARX (AutoRegressive with eXogenous input)
- ARMAX
- OE (Output Error)
- BJ (Box-Jenkins)



Outline

1. Famiglie di modelli a spettro razionale
- 2. Serie temporali: modelli MA, AR, ARMA**
3. Sistemi ingresso\uscita: modelli ARX e ARMAX
4. Sistemi ingresso\uscita: modelli FIR, OE, BJ
5. Densità spettrale di potenza: esempio di calcolo



Modelli Moving Average (MA)

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

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

- c_0, c_1, \dots, c_{n_c} : coefficienti del modello MA(n_c)
- n_c : ordine del modello

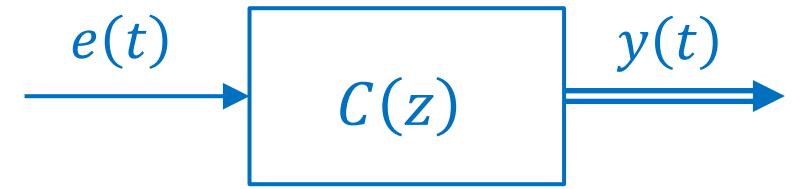
L'uscita di un modello MA(n_c) è **combinazione lineare** degli ultimi $n_c + 1$ valori del rumore bianco in ingresso



Modelli Moving Average (MA): funzione di trasferimento

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

$$y(t) = c_0e(t) + c_1e(t - 1) + c_2e(t - 2) + \cdots + c_{n_c}e(t - n_c)$$



$$= c_0e(t) + c_1z^{-1}e(t) + c_2z^{-2}e(t) + \cdots + c_{n_c}z^{-n_c}e(t)$$

$$= \underbrace{[c_0 + c_1z^{-1} + c_2z^{-2} + \cdots + c_{n_c}z^{-n_c}]}_{C(z)} \cdot e(t) \quad \Rightarrow \quad y(t) = C(z)e(t)$$

$$\frac{y(t)}{e(t)} = \frac{z^{n_c}c_0 + z^{n_c-1}c_1 + \cdots + c_{n_c}}{z^{n_c}} = C(z)$$

Osserviamo n_c poli in 0. Quindi, i processi MA(n_c) sono sempre stazionari



Modelli Moving Average (MA): proprietà

VALORE ATTESO

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

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

$$= c_0 \mu + c_1 \mu + \cdots + c_{n_c} \mu$$

$$= \mu \cdot \sum_{i=0}^{n_c} c_i$$

Non dipende dal tempo t



Se $e(t) \sim WN(0, \lambda^2)$, allora $\mathbb{E}[y(t)] = 0$



Modelli Moving Average (MA): proprietà

FUNZIONE DI AUTOCOVARIANZA

Per semplicità, supponiamo $\mathbb{E}[y(t)] = 0$, tramite depolarizzazione

$$\bullet \quad \gamma_{yy}(0) = \mathbb{E}[(y(t) - m_y)^2] = \mathbb{E}[(y(t))^2] = \mathbb{E}\left[\left(c_0 e(t) + c_1 e(t-1) + \dots + c_{n_c}^2 e(t-n_c)\right)^2\right]$$

$$= \mathbb{E} \left[\begin{aligned} & c_0^2 e(t)^2 + c_1^2 e(t-1)^2 + \dots + c_{n_c}^2 e(t-n_c)^2 + \\ & + 2c_0 c_1 e(t)e(t-1) + \dots \\ & + 2c_{n_c-1} c_{n_c} e(t-n_c+1)e(t-n_c) \end{aligned} \right] = c_0^2 \mathbb{E}[e(t)^2] + \dots + c_{n_c}^2 \mathbb{E}[e(t-n_c)^2]$$

$$= c_0^2 \gamma_{ee}(0) + c_1^2 \gamma_{ee}(0) + \dots + c_{n_c}^2 \gamma_{ee}(0)$$

$$= \lambda^2 \cdot \sum_{i=0}^{n_c} c_i^2$$

Non dipende dal tempo t



Modelli Moving Average (MA): proprietà

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

- $\gamma_{yy}(2) = \lambda^2 \cdot (c_0 c_2 + c_1 c_3 + \dots + c_{n_c-2} c_{n_c})$
- $\gamma_{yy}(n_c) = \lambda^2 \cdot (c_0 c_{n_c})$

- $\gamma_{yy}(\tau) = 0$ se $\tau > n_c$

Un processo MA(n_c) dipende solo dagli n_c valori precedenti al tempo corrente



Modelli Moving Average (MA): proprietà

Osservazioni

- Un modo per **capire se una serie temporale può essere modellata tramite un MA(n_c)** è quello di guardare se la sua **funzione di autocovarianza** (nella pratica, una stima di essa) **va a zero** dopo n_c lags
- Il processo $\tilde{y}(t) = \tilde{c}_0\tilde{e}(t) + \tilde{c}_1\tilde{e}(t - 1) + \tilde{c}_2\tilde{e}(t - 2) + \cdots + \tilde{c}_{n_c}\tilde{e}(t - n_c)$ con $\tilde{e}(t) \sim \text{WN}(0, \tilde{\lambda}^2)$, $\tilde{c}_i = \alpha \cdot c_i$, $\tilde{\lambda}^2 = \lambda^2/\alpha^2$, ha lo **stesso valore atteso e autocovarianza** del processo $y(t) = c_0e(t) + c_1e(t - 1) + c_2e(t - 2) + \cdots + c_{n_c}e(t - n_c)$

Per evitare questa **sovraparametrizzazione** del modello, spesso si fissa $c_0 = 1$



Modelli AutoRegressive (AR)

Definizione: Un processo stocastico $y(t)$, generato a partire dal rumore bianco $e(t) \sim WN(\mu, \lambda^2)$, è detto di tipo AR(n_a), se:

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

- a_1, \dots, a_{n_a} : coefficienti del modello AR(n_a)
- n_a : ordine del modello

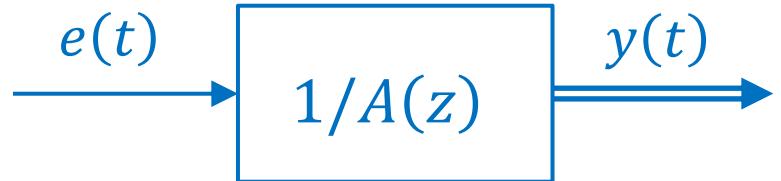
L'uscita di un modello AR(n_a) è **combinazione lineare** degli ultimi n_a valori del processo stesso e del rumore bianco in ingresso



Modelli AutoRegressive (AR): funzione di trasferimento

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

$$y(t) = a_1y(t - 1) + a_2y(t - 2) + \cdots + a_{n_a}y(t - n_a) + e(t)$$



$$= a_1z^{-1}y(t) + a_2z^{-2}y(t) + \cdots + a_{n_a}z^{-n_a}y(t)$$

$$y(t) \underbrace{[1 - a_1z^{-1} + a_2z^{-2} + \cdots + a_{n_a}z^{-n_a}]}_{A(z)} = e(t) \quad \Rightarrow$$

$$y(t) = \frac{1}{A(z)}e(t)$$

$$\frac{y(t)}{e(t)} = \frac{z^{n_a}}{z^{n_a} - a_1z^{n_a-1} - \cdots - a_{n_a}} = A(z)$$

- n_a zeri nell'origine
- n_a poli

Un processo $AR(n_a)$ è stazionario se e solo se $1/A(z)$ è **asintoticamente stabile**



Modelli AutoRegressive (AR): proprietà

VALORE ATTESO (*nel caso il processo sia stazionario*)

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

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

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



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



Modelli AutoRegressive (AR): proprietà

FUNZIONE DI AUTOCOVARIANZA (*nel caso il processo sia stazionario*)

Consideriamo processi AR(1) del tipo $y(t) = a_1y(t-1) + e(t)$, $e(t) \sim WN(0, \lambda^2)$. Supponiamo che il processo sia **asintoticamente stabile** (ovvero, $|a_1| < 1$), e a **media nulla**

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



Modelli AutoRegressive (AR): proprietà

- $\gamma_{yy}(1) = \mathbb{E}[y(t)y(t-1)] = \mathbb{E}[(a_1y(t-1) + e(t)) \cdot y(t-1)] = \mathbb{E}[a_1y(t-1)^2 + y(t-1)e(t)]$

$$= a_1 \mathbb{E}[y(t-1)^2] + \cancel{\mathbb{E}[y(t-1)e(t)]} = a_1 \gamma_{yy}(0) \quad \Rightarrow \quad \boxed{\gamma_{yy}(1) = a_1 \gamma_{yy}(0)}$$

- $\gamma_{yy}(2) = \mathbb{E}[y(t)y(t-2)] = \mathbb{E}[(a_1y(t-1) + e(t)) \cdot y(t-2)]$

$$= \mathbb{E}[a_1y(t-1)y(t-2) + y(t-2)e(t)] = a_1 \mathbb{E}[y(t-1)y(t-2)] + \cancel{\mathbb{E}[y(t-2)e(t)]}$$

$$= a_1 \gamma_{yy}(1) \quad \Rightarrow \quad \boxed{\gamma_{yy}(2) = a_1 \gamma_{yy}(1)}$$



Modelli AutoRegressive (AR): equazioni Yule-Walker

Generalizzando, si ha che, per un processo AR(1),

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

Equazioni di Yule-Walker

per un AR(1)

Esistono anche per AR(n_a)

Osservazioni

- Dato che abbiamo supposto un processo AR(1) stazionario, allora $|a_1| < 1$. Quindi

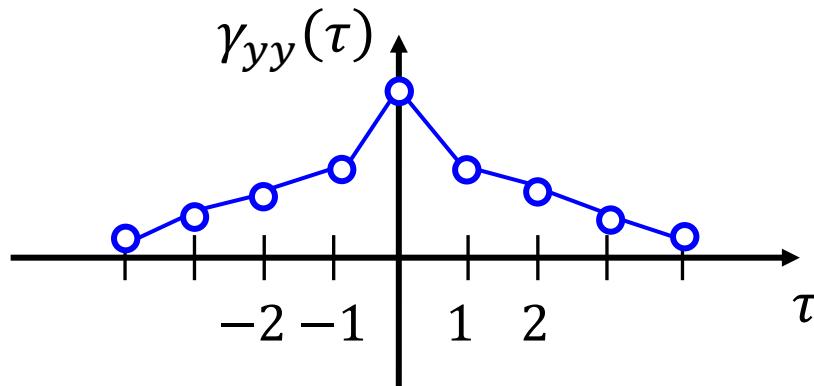
$$|\gamma_{yy}(\tau + 1)| < |\gamma_{yy}(\tau)|$$

e dato che $|a_1| \neq 1$, allora $\gamma_{yy}(0)$ esiste finito



Modelli AutoRegressive (AR)

- Il processo $\bar{y}(t) = a_1\bar{y}(t - 1) + e(t)$, $e(t) \sim WN(0, \lambda^2)$, con $0 < a_1 < 1$, ha funzione di **autocovarianza** $\gamma_{yy}(\tau) > 0 \ \forall \tau$, e sarà **decrescente** (*ma non raggiunge mai lo zero*)

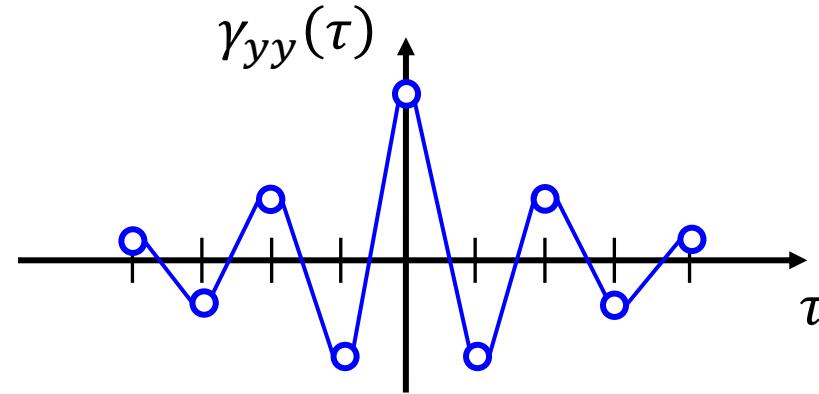


Le **realizzazioni** del processo «**variano lentamente**» e sono «smooth», poiché le variabili casuali sono **correlate positivamente** fra loro. In media, le realizzazioni «**non cambiano segno**» da un istante al successivo. Le componenti a **bassa frequenza** dominano nella densità spettrale di potenza



Modelli AutoRegressive (AR)

- Il processo $\bar{y}(t) = a_1\bar{y}(t - 1) + e(t)$, $e(t) \sim WN(0, \lambda^2)$, con $-1 < a_1 < 0$, ha funzione di **autocovarianza** che **cambia segno** ad ogni τ , in modo alternato (e decresce in valore assoluto senza raggiungere lo zero)



Le **realizzazioni** del processo «**variano velocemente**» e sono «nervose», poiché le variabili casuali sono **correlate negativamente** fra loro. In media, le realizzazioni «**cambiano segno**» da un istante al successivo. Le componenti ad **alta frequenza** dominano nella densità spettrale di potenza



Modelli AutoRegressive (AR)

Abbiamo visto che, per un process MA(n_c), $\gamma_{yy}(\tau) = 0 \forall \tau > n_c$. Per i processi AR(n_a), possiamo osservare un comportamento analogo guardando la **funzione di autocorrelazione parziale (PACF)** $\gamma_{yy}^{\text{PAR}}(\tau)$. La PACF è tale che $\gamma_{yy}^{\text{PAR}}(\tau) = 0 \forall \tau > n_a$

Nell'*analisi pratica di serie temporali*, si seguono questi passaggi:

1. Controllo se la serie temporale può essere modellata con un MA(n_c) guardando $\gamma_{yy}(\tau)$
2. Controllo se la serie temporale può essere modellata con un AR(n_a) guardando $\gamma_{yy}^{\text{PAR}}(\tau)$
3. Se nessuna delle due funzioni si annulla da un certo τ in poi, ho bisogno di **altri modelli**

Un'altra categoria di modelli per serie temporali sono gli ARMA(n_a, n_c)



Modelli AutoRegressive Moving Average (ARMA)

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

$$y(t) = a_1 y(t-1) + a_2 y(t-2) + \cdots + a_{n_a} y(t-n_a) + e(t) + c_1 e(t-1) + c_2 e(t-2) + \cdots + c_{n_c} e(t-n_c)$$

Parte AR(n_a)

Parte MA(n_c)

- a_1, \dots, a_{n_a} : coefficienti della parte AR(n_a)
- n_a : ordine della parte AR(n_a)
- c_0, c_1, \dots, c_{n_c} : coefficienti della parte MA(n_c)
- n_c : ordine della parte MA(n_c)

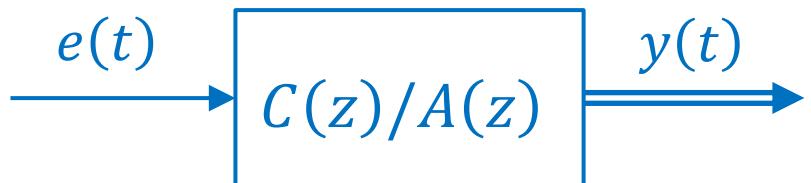
Notiamo che $ARMA(0, n_c) = MA(n_c)$ e $ARMA(n_a, 0) = AR(n_a)$



Modelli AutoRegressive Moving Average (ARMA)

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

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

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


Il processo $y(t)$ è stazionario se e solo se $C(z)/A(z)$ è asintoticamente stabile



Modelli AutoRegressive Moving Average (ARMA)

Teorema Dato un processo stocastico stazionario ARMA(n_a, n_c), esso può essere scritto come un MA(∞)

Esempio

Supponiamo di avere un AR(1) del tipo $y(t) = ay(t - 1) + e(t)$, $e(t) \sim WN(0, \lambda^2)$

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

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



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



Outline

1. Famiglie di modelli a spettro razionale
2. Serie temporali: modelli MA, AR, ARMA
- 3. Sistemi ingresso\uscita: modelli ARX e ARMAX**
4. Sistemi ingresso\uscita: modelli FIR, OE, BJ
5. Densità spettrale di potenza: esempio di calcolo



Modelli AR with eXogenous input (ARX)

Definizione: Un processo stocastico $y(t)$, generato a partire dal rumore bianco $e(t) \sim WN(\mu, \lambda^2)$, è detto di tipo ARX(n_a, n_b, k), se:

$$y(t) = a_1 y(t-1) + a_2 y(t-2) + \cdots + a_{n_a} y(t-n_a) + e(t) \quad \text{Parte AR}(n_a)$$

$$+ b_0 u(t-k) + b_1 u(t-k-1) + \cdots + b_{n_b} u(t-k-n_b) \quad \text{Parte X}(n_b)$$

- a_1, \dots, a_{n_a} : coefficienti della parte AR(n_a)
- n_a : ordine della parte AR(n_a)
- b_0, b_1, \dots, b_{n_b} : coefficienti della parte X(n_b)
- n_b : ordine della parte X(n_b)

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



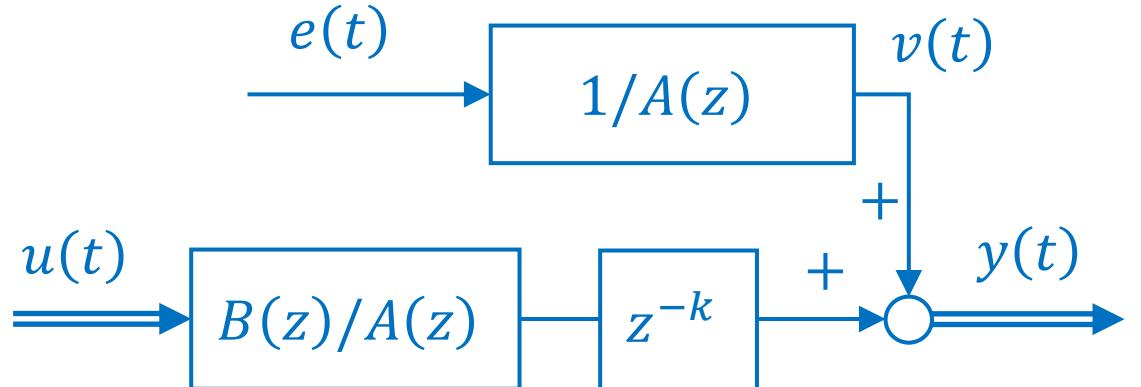
Modelli AR with eXogenous input (ARX)

La **funzione di trasferimento** di un ARX(n_a, n_b, k) risulta essere

$$y(t)[1 - a_1 z^{-1} - \dots - a_{n_a} z^{-n_a}] = [b_0 z^{-k} + b_1 z^{-k-1} + \dots + b_{n_b} z^{-k-n_b}]u(t) + e(t)$$

$$y(t) = \frac{b_0 + b_1 z^{-1} + \dots + b_{n_b} z^{-n_b}}{1 - a_1 z^{-1} - \dots - a_{n_a} z^{-n_a}} B(z) u(t - k) + \frac{1}{1 - a_1 z^{-1} - \dots - a_{n_a} z^{-n_a}} e(t)$$

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



Modelli ARMA with eXogenous input (ARMAX)

Definizione: Un processo stocastico $y(t)$, generato a partire dal rumore bianco $e(t) \sim WN(\mu, \lambda^2)$, è detto di tipo ARMAX(n_a, n_c, n_b, k), se:

$$y(t) = a_1 y(t-1) + a_2 y(t-2) + \cdots + a_{n_a} y(t-n_a) + \quad \text{Parte AR}(n_a)$$

$$+ b_0 u(t-k) + b_1 u(t-k-1) + \cdots + b_{n_b} u(t-k-n_b) \quad \text{Parte X}(n_b)$$

$$+ e(t) + c_1 e(t-1) + c_2 e(t-2) + \cdots + c_{n_c} e(t-n_c) \quad \text{Parte MA}(n_c)$$



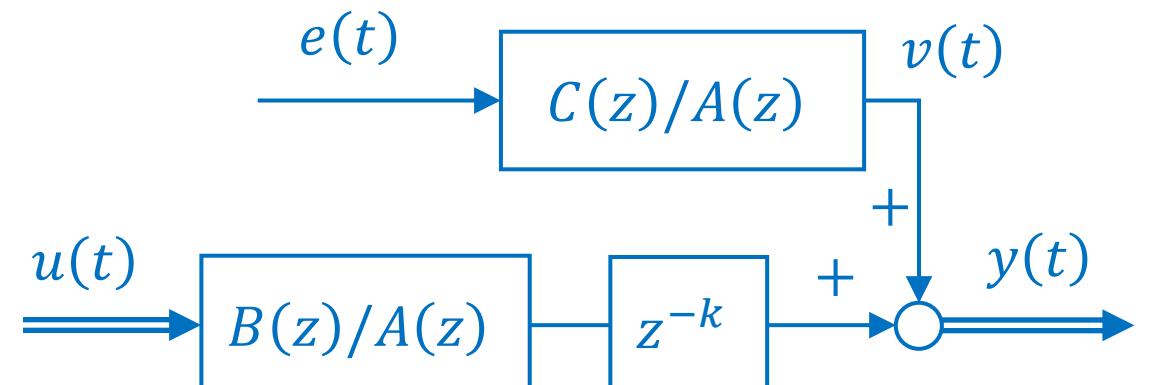
Modelli ARMA with eXogenous input (ARMAX)

La **funzione di trasferimento** di un ARMAX(n_a, n_c, n_b, k) risulta essere

$$y(t)[1 - a_1 z^{-1} + \dots + a_{n_a} z^{-n_a}] = [b_0 z^{-k} + b_1 z^{-k-1} + \dots + b_{n_b} z^{-k-n_b}]u(t) + \\ + [1 + c_1 z^{-1} + c_2 z^{-2} + \dots + c_{n_c} z^{-n_c}]e(t)$$

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

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



Outline

1. Famiglie di modelli a spettro razionale
2. Serie temporali: modelli MA, AR, ARMA
3. Sistemi ingresso\uscita: modelli ARX e ARMAX
- 4. Sistemi ingresso\uscita: modelli FIR, OE, BJ**
5. Densità spettrale di potenza: esempio di calcolo

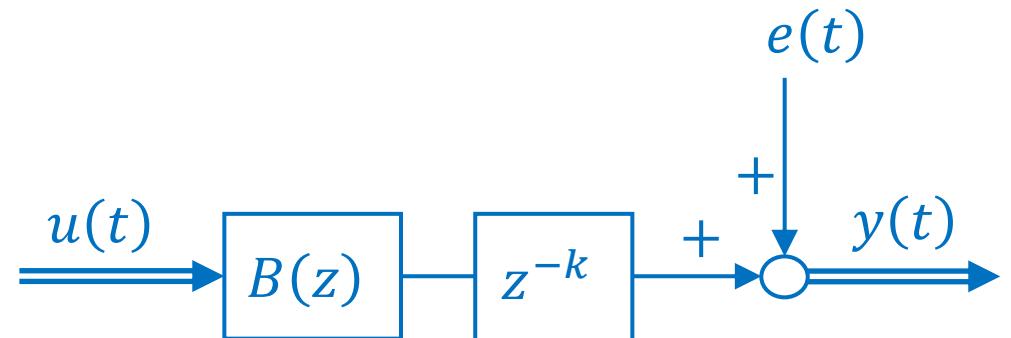


Modelli Finite Impulse Response (FIR)

Definizione: un modello FIR(n_b, k), con rumore bianco additivo $e(t) \sim WN(0, \lambda^2)$, è definito come

$$\begin{aligned}y(t) &= b_0 u(t - k) + b_1 u(t - k - 1) + \cdots + b_{n_b} u(t - k - n_b) = \sum_{i=0}^{n_b} b_i \cdot u(t - k - i) + e(t) \\&= B(z)u(t - k) + e(t)\end{aligned}$$

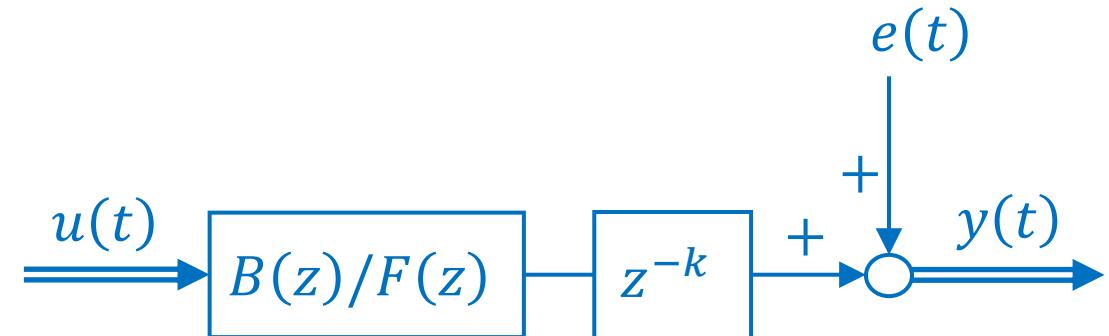
L'uscita di un modello FIR(n_b) dipende solo da
valori passati dell'ingresso $u(t)$ e dal rumore
bianco $e(t)$



Modelli Output Error (OE)

Definizione: un modello $\text{OE}(n_b, n_f, k)$, con rumore bianco additivo $e(t) \sim \text{WN}(0, \lambda^2)$, è definito come

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



Questo modello è simile al modello ARX(n_a, n_b, k), ma, a differenza di quest'ultimo, suppone che il **rumore entri solo dopo** che l'uscita «non rumorosa» è stata generata

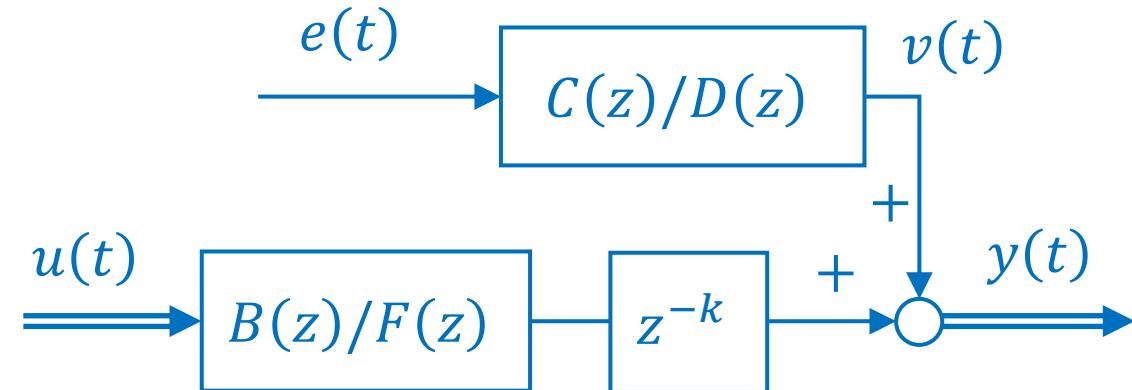
Il modello $\text{OE}(n_b, n_f, k)$ è uno dei più utilizzati per modellare processi con solo **errore di misura**



Modelli Box-Jenkins (BJ)

Definizione: Un processo stocastico $y(t)$, generato a partire dal rumore bianco $e(t) \sim WN(\mu, \lambda^2)$, è detto di tipo $BJ(n_c, n_b, n_d, n_f, k)$, se:

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



A differenza dei modelli ARMAX(n_a, n_c, n_b, k), questi modelli hanno **polinomi diversi al denominatore**, per cui parte esogena e parte stocastica sono **parametrizzate in modo indipendente**. Tale proprietà si ha anche coi modelli FIR(n_b, k) e OE(n_b, n_f, k)

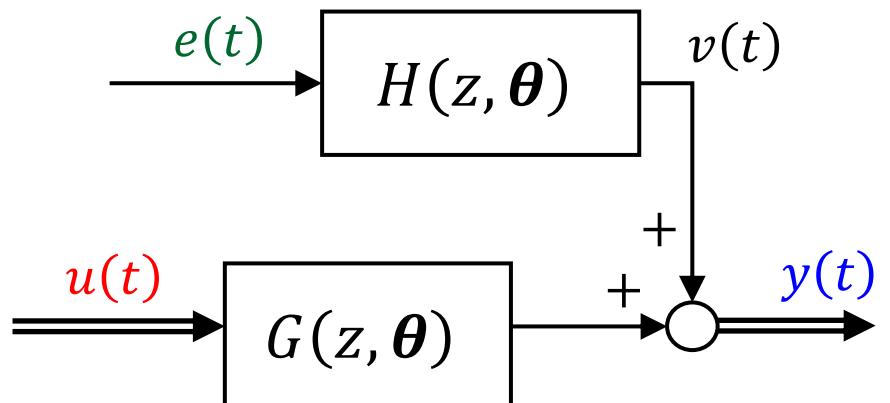


Famiglie di modelli per sistemi ingresso\uscita

Famiglia di modello	$G(z, \theta)$	$H(z, \theta)$
ARX	$\frac{B(z, \theta)}{A(z, \theta)} z^{-k}$	$\frac{1}{A(z, \theta)}$
ARMAX	$\frac{B(z, \theta)}{A(z, \theta)} z^{-k}$	$\frac{C(z, \theta)}{A(z, \theta)}$
OE	$\frac{B(z, \theta)}{F(z, \theta)} z^{-k}$	1
FIR	$B(z, \theta) z^{-k}$	1
BJ	$\frac{B(z, \theta)}{F(z, \theta)} z^{-k}$	$\frac{C(z, \theta)}{D(z, \theta)}$

I termini «famiglia», «classe» o «struttura» di modello sono usati come **sinonimi**

Il vettore θ rappresenta i **parametri del modello** (valore dei coefficienti dei polinomi)



Outline

1. Famiglie di modelli a spettro razionale
2. Serie temporali: modelli MA, AR, ARMA
3. Sistemi ingresso\uscita: modelli ARX e ARMAX
4. Sistemi ingresso\uscita: modelli FIR, OE, BJ
- 5. Densità spettrale di potenza: esempio di calcolo**



Densità spettrale di potenza: esempio di calcolo

Consideriamo un processo MA(1) del tipo

$$y(t) = c_0 e(t) + c_1 e(t-1), \quad c_0 = 1, e(t) \sim \text{WN}(0,1)$$

1) Calcolo della densità spettrale di potenza usando la definizione

$$\begin{aligned}\Gamma_{yy}(\omega) &= \sum_{\tau=-\infty}^{+\infty} \gamma_{yy}(\tau) e^{-j\omega\tau} = \gamma_{yy}(-1)e^{-j\omega(-1)} + \gamma_{yy}(0)e^{-j\omega(0)} + \gamma_{yy}(+1)e^{-j\omega(+1)} \\ &= \lambda^2 c_0 c_1 \cdot e^{j\omega} + \lambda^2(c_0^2 + c_1^2) \cdot 1 + \lambda^2 c_0 c_1 \cdot e^{-j\omega} = c_1 \cdot e^{j\omega} + (1 + c_1^2) + c_1 \cdot e^{-j\omega} \\ &= c_1 [e^{j\omega} + e^{-j\omega}] + c_1^2 + 1 = \boxed{2c_1 \cos \omega + c_1^2 + 1}\end{aligned}$$



Densità spettrale di potenza: esempio di calcolo

2) Calcolo della densità spettrale di potenza usando il modello

$$y(t) = e(t) + c_1 e(t-1) \quad \Rightarrow \quad y(t) = [1 + c_1 z^{-1}] e(t) \quad \Rightarrow \quad y(t) = C(z) e(t)$$

$$\begin{aligned}\Gamma_{yy}(\omega) &= |C(e^{j\omega})|^2 \cdot \Gamma_{ee}(\omega) = (1 + c_1 e^{j\omega}) \cdot (1 + c_1 e^{-j\omega}) \\ &= 1 + c_1 e^{-j\omega} + c_1 e^{j\omega} + c_1^2 (e^{-j\omega} \cdot e^{j\omega}) = 1 + c_1 [e^{j\omega} + e^{-j\omega}] + c_1^2 \\ &= 2c_1 \cos \omega + c_1^2 + 1\end{aligned}$$



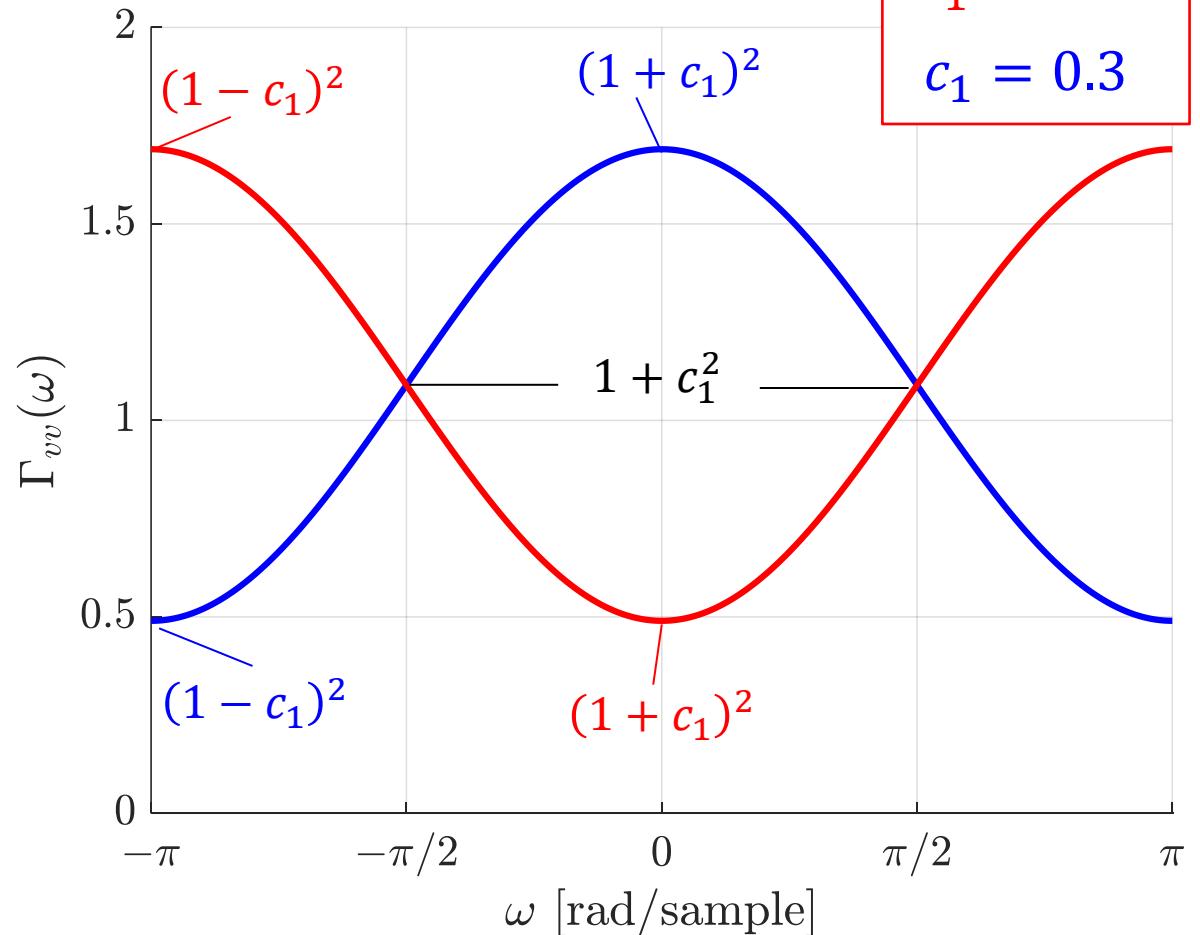
Densità spettrale di potenza: esempio di calcolo

Per **disegnare qualitativamente** lo spettro, lo valutiamo in 3 frequenze:

$$\begin{aligned}\Gamma_{yy}(0) &= 1 + c_1^2 + 2c_1 \cos 0 = 1 + c_1^2 + 2c_1 \\ &= (1 + c_1)^2\end{aligned}$$

$$\Gamma_{yy}\left(\frac{\pi}{2}\right) = 1 + c_1^2 + 2c_1 \cos \frac{\pi}{2} = 1 + c_1^2$$

$$\begin{aligned}\Gamma_{yy}(\pi) &= 1 + c_1^2 + 2c_1 \cos \pi = 1 + c_1^2 - 2c_1 \\ &= (1 - c_1)^2\end{aligned}$$

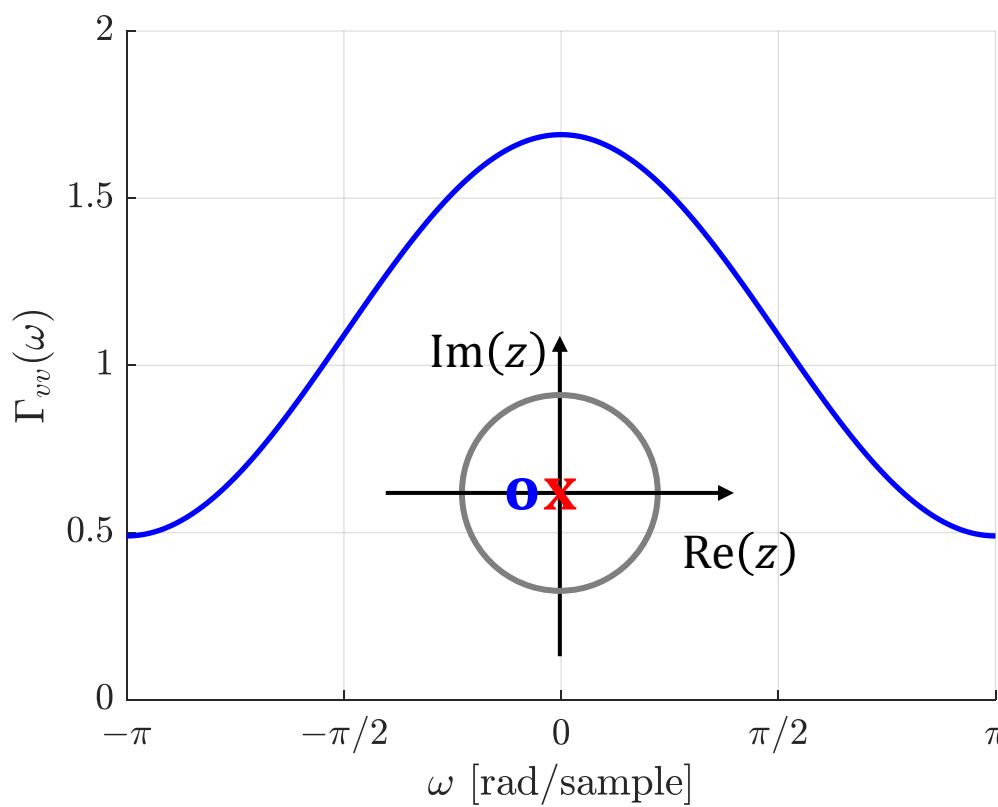


Densità spettrale di potenza: esempio di calcolo

$$y(t) = [1 + 0.3z^{-1}]e(t)$$

zero $z = -0.3$

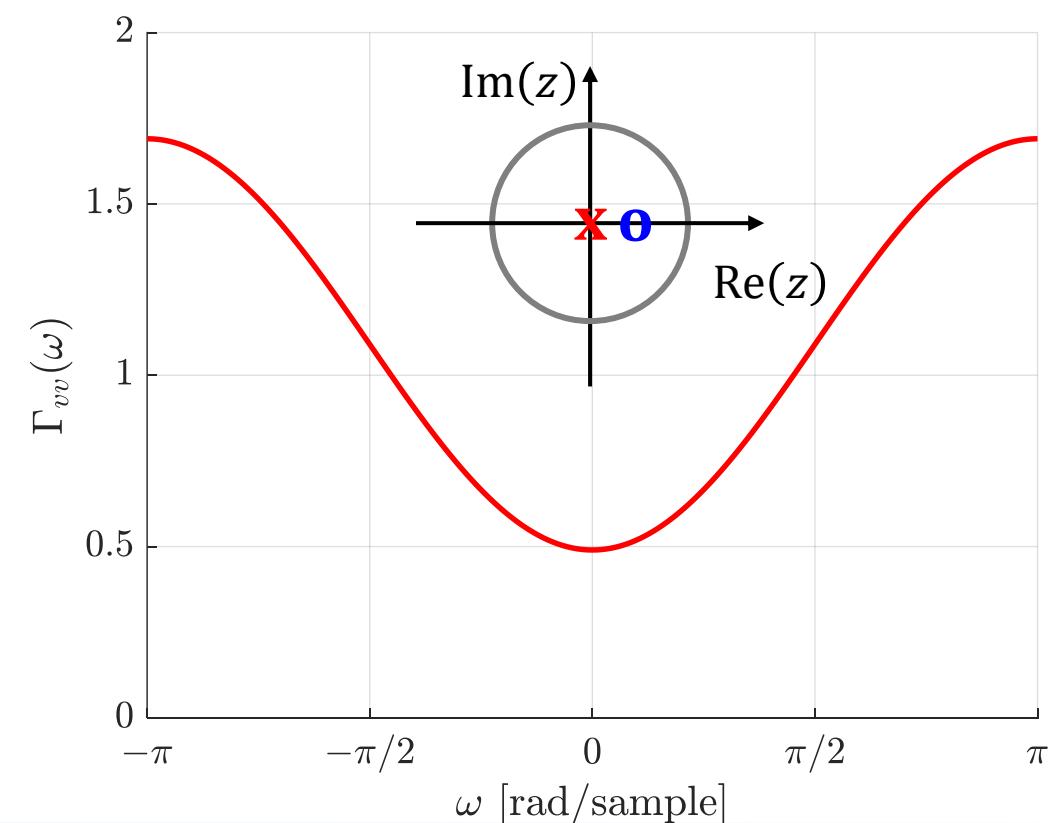
polo $z = 0$



$$y(t) = [1 - 0.3z^{-1}]e(t)$$

zero $z = +0.3$

polo $z = 0$





**UNIVERSITÀ
DEGLI STUDI
DI BERGAMO**

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



IDENTIFICAZIONE DEI MODELLI E ANALISI DEI DATI (IMAD)

Lezione 10: Predizione

**Corso di Laurea Magistrale in
INGEGNERIA INFORMATICA**

SPEAKER
Prof. Mirko Mazzoleni

PLACE
**Università degli Studi di
Bergamo**

Syllabus

Parte II: sistemi dinamici

8. Processi stocastici

- 8.1 Processi stocastici stazionari (pss)
- 8.3 Rappresentazione spettrale di un pss
- 8.4 Stimatori campionari media\covarianza
- 8.5 Densità spettrale campionaria

9. Famiglie di modelli a spettro razionale

- 9.1 Modelli per serie temporali (MA, AR, ARMA)
- 9.2 Modelli per sistemi input/output (ARX, ARMAX)

10. Predizione

- 10.1 Filtro passa-tutto

10.2 Forma canonica

10.3 Teorema della fattorizzazione spettrale

10.4 Soluzione al problema della predizione

11. Identificazione

- 11.3 Identificazione di modelli ARX
- 11.4 Identificazione di modelli ARMAX
- 11.5 Metodo di Newton

12. Identificazione: analisi e complementi

- 12.1 Analisi asintotica metodi PEM
- 12.2 Identificabilità dei modelli
- 12.3 Valutazione dell'incertezza di stima

13. Identificazione: valutazione



Parte I: sistemi statici**Parte II: sistemi dinamici**Stima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

- ✓ Stima parametri popolazione
- ✓ Stima modello lineare: minimi quadrati

○ **SI assunzioni su ddp dei dati**

- ✓ Stima massima verosimiglianza parametri popolazione
- ✓ Stima modello lineare: massiva verosimiglianza
- ✓ Regressione logistica

• θ variabile casuale○ **SI assunzioni su ddp dei dati**

- ✓ Stima Bayesiana

Stima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

- ✓ Modelli lineari di pss
- ✓ Predizione
- ✓ Identificazione
- ✓ Persistente eccitazione
- ✓ Analisi asintotica metodi PEM
- ✓ Analisi incertezza stima (numero dati finito)
- ✓ Valutazione del modello

Machine learning

Outline

1. Predizione, filtraggio e smoothing
2. Scomposizione di Wold
3. Filtro passa-tutto e forma canonica
4. Predittore ottimo
5. Predittore ottimo per processi MA
6. Predittore ottimo per processi ARMA
7. Predittore ottimo per processi ARMAX
8. Predittore ottimo ad un passo per sistemi ingresso\uscita
9. Confronto con il predittore di Kalman



Outline

- 1. Predizione, filtraggio e smoothing**
2. Scomposizione di Wold
3. Filtro passa-tutto e forma canonica
4. Predittore ottimo
5. Predittore ottimo per processi MA
6. Predittore ottimo per processi ARMA
7. Predittore ottimo per processi ARMAX
8. Predittore ottimo ad un passo per sistemi ingresso\uscita
9. Confronto con il predittore di Kalman



Gli step per la risoluzione del problema

Seguiremo tre fasi per risolvere il problema della **modellazione di sistemi dinamici**:

Definizione delle **classi di modelli** \mathcal{M} di sistemi dinamici

Ci concentreremo su modelli di **sistemi dinamici lineari**, espressi da **funzioni di trasferimento razionali fratte**. I parametri ignoti sono i coefficienti dei polinomi al numeratore e denominatore

Predizione

Data una particolare classe di modello, supponendo di conoscerne il valore dei parametri, qual è il **preditore ottimo**? Quanto vale la predizione ottima?

Identificazione

Come **stimo il valore dei parametri** del modello scelto per la modellazione dei dati?



Predizione, filtraggio e smoothing

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

- $y(t) = x(t)$ **Misuro il processo** $x(t)$ che mi interessa stimare
- $y(t) = x(t) + e(t)$, $e(t) \sim WN(0, \lambda^2)$ **Misuro una «versione rumorosa»** di $x(t)$

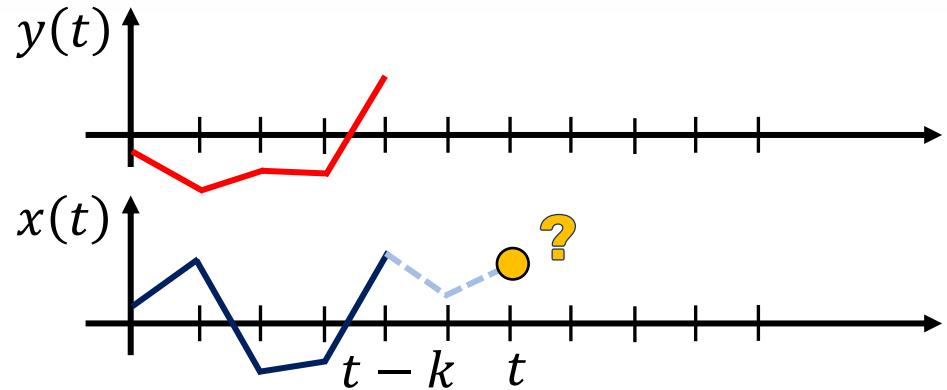
Vogliamo ottenere una stima $x(t|t_{\text{info}})$, basata sulla conoscenza di $y(s)$, per valori di:

- $s < t_{\text{info}}$ **predizione**
- $s = t_{\text{info}}$ **filtraggio**
- $s > t_{\text{info}}$ **smoothing**



Predizione a k passi

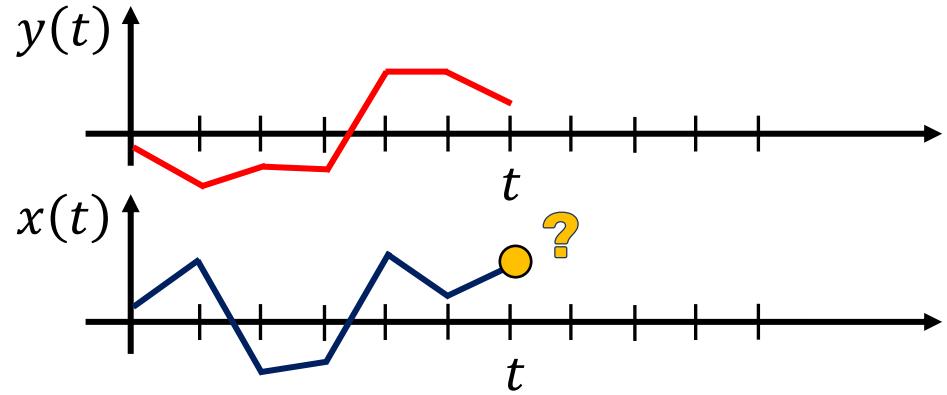
- $\hat{x}(t|t-k)$
Obiettivo: stimare il valore di $x(t)$ a istanti futuri



Filtraggio

ha senso solo se $x(t) \neq y(t)$

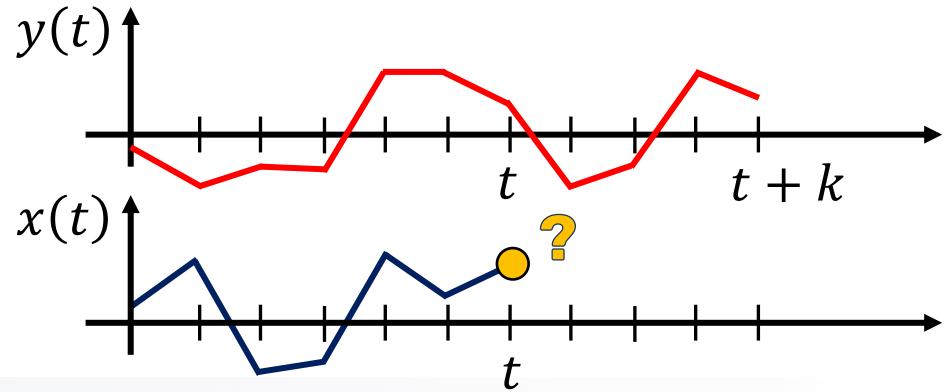
- $\hat{x}(t|t)$
Obiettivo: ottenere una stima di $x(t)$ all'istante corrente e «pulita dal rumore» (es. filtro di Kalman)



Smoothing

ha senso solo se $x(t) \neq y(t)$

- $\hat{x}(t|t+k)$
Obiettivo: ottenere una stima di $x(t)$ all'istante passato (es. ricostruzione traiettorie missili NASA)



Predizione ottima

Noi studieremo il **problema della predizione** per $x(t) = y(t)$, ovvero ci interesserà trovare una stima $\hat{y}(t|t - k)$ di $y(t)$ al tempo t , avendo a disposizione i dati fino al tempo $t - k$

Dato che il **predittore** $\hat{y}(t|t - k)$ si basa su valori passati di $y(t)$, sarà anch'esso un processo stocastico. L'**errore di predizione** $\varepsilon_k(t)$ è un processo stocastico definito come

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

Vogliamo predittori (stimatori) **lineari ottimi**, i.e. con errore di predizione a **MSE minimo**

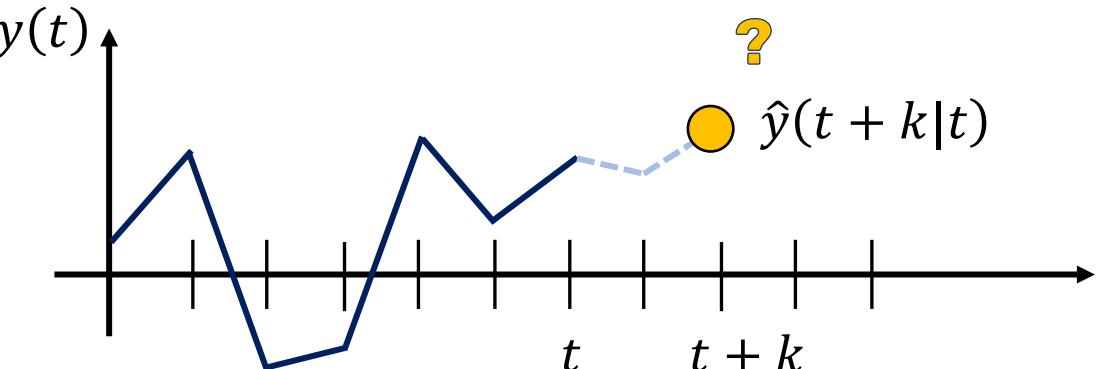
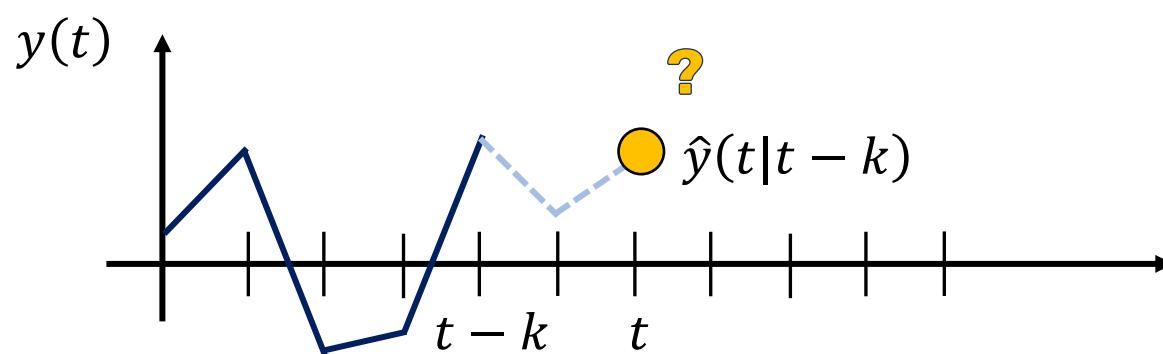
Approccio predittivo: un modello è **buono** se è in grado di **predire bene** i dati.
Più avanti, stimeremo i parametri di un modello dinamico *minimizzando la varianza dell'errore di predizione*



Predizione, filtraggio e smoothing

Note

- Studieremo il predittore ottimo per pss a spettro razionale delle famiglie **ARMA** e **ARMAX**. I predittori ottimi per le altre famiglie FIR, OE, BJ non verranno investigati
- Dato che lavoriamo con pss, le scritture $\hat{y}(t|t - k)$ e $\hat{y}(t + k|t)$ sono **equivalenti**, nel senso che la **forma del predittore ottimo è la stessa** (le predizioni potranno essere diverse)



Outline

1. Predizione, filtraggio e smoothing
- 2. Scomposizione di Wold**
3. Filtro passa-tutto e forma canonica
4. Predittore ottimo
5. Predittore ottimo per processi MA
6. Predittore ottimo per processi ARMA
7. Predittore ottimo per processi ARMAX
8. Predittore ottimo ad un passo per sistemi ingresso\uscita
9. Confronto con il predittore di Kalman



Processi stocastici completamente predicibili

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

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

Se conosco i coefficienti a_i posso **prevedere senza errore** i valori futuri di $y(\cdot)$ senza errori, a **partire dai valori passati**

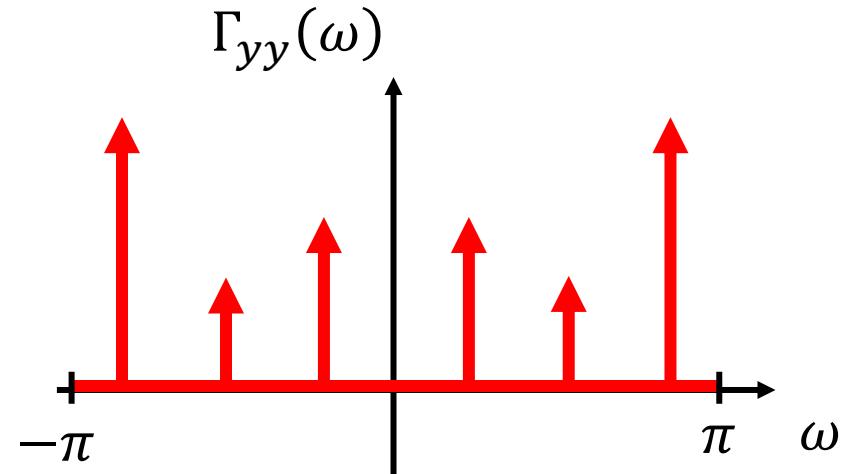
Tali processi sono «l'opposto» del **rumore bianco**, che è **completamente impredicibile**



Processi stocastici completamente predicibili

Proprietà: Un processo stazionario $y(t)$ è completamente predicibile **se e solo se**

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



La densità spettrale di potenza di un **processo completamente predicibile** è una combinazione lineare di **delta di Dirac**

Il **rumore bianco**, in contrasto, ha una densità spettrale di potenza **costante**



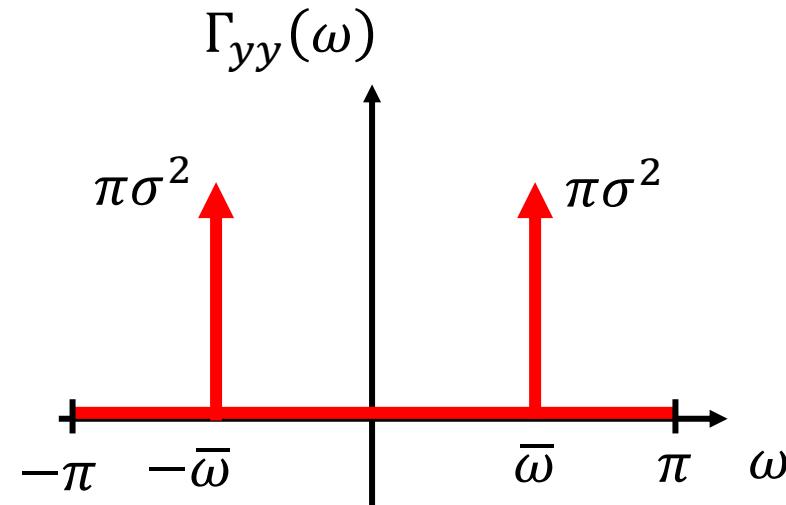
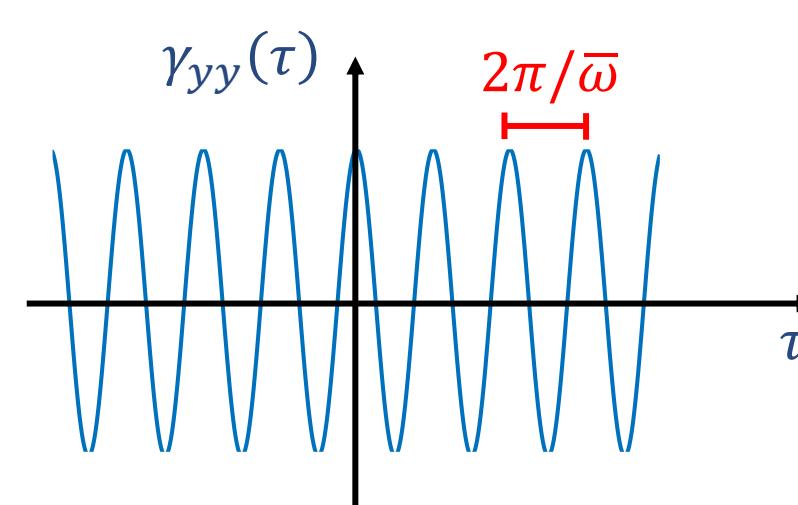
Processi stocastici completamente predibili

Esempio: Consideriamo il seguente processo stocastico

$$y(t) = v_1 \cos(\bar{\omega}t) + v_2 \sin(\bar{\omega}t)$$

Dove v_1 e v_2 sono variabili casuali **incorelate** con $\mathbb{E}[v_1] = \mathbb{E}[v_2]$, e $\text{Var}[v_1] = \text{Var}[v_2] = \sigma^2$

Si verifica che questo è un processo stazionario (non ergodico) con $\gamma_{yy}(\tau)$ **cosinusoidale**



Una volta che ho capito l'andamento sinusoidale del processo, **sono in grado di prevederlo** da lì all'infinito



Processi stocastici completamente predicibili

Osservazione

In generale, quando $\Gamma_{yy}(\omega)$ è «a righe», il processo $y(t)$ è una **combinazione lineare di seni e coseni** (che è perfettamente predicibile dai valori passati)

In altre parole, le realizzazioni del processo sono **periodiche**



Scomposizione di Wold

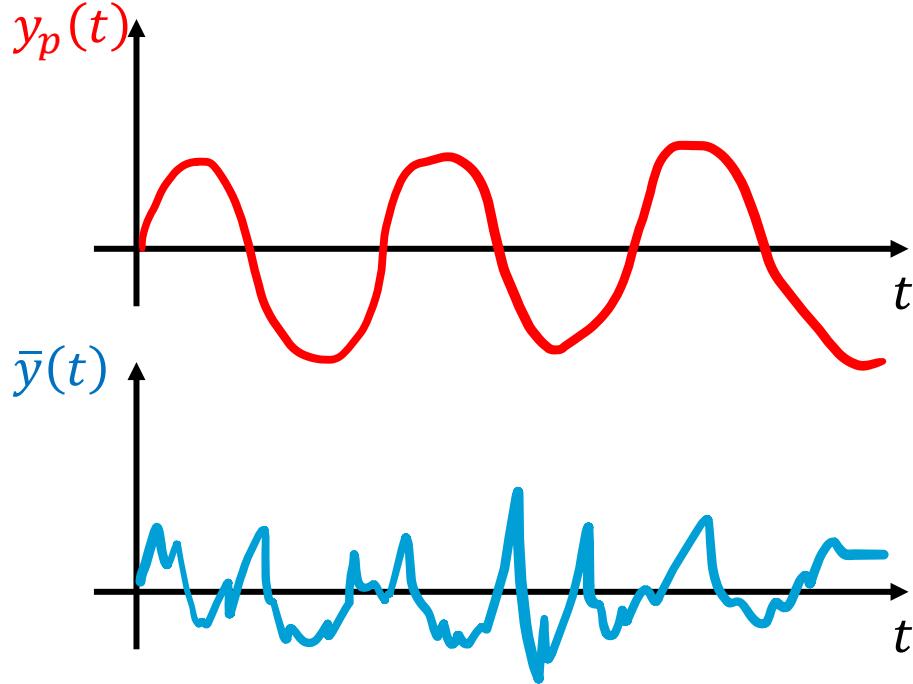
Ogni **processo stocastico stazionario** $y(t)$ può essere scritto come

$$y(t) = \bar{y}(t) + y_p(t)$$

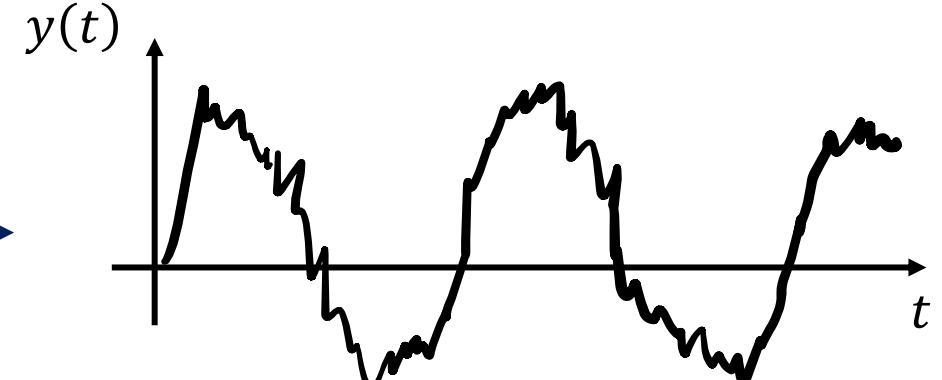
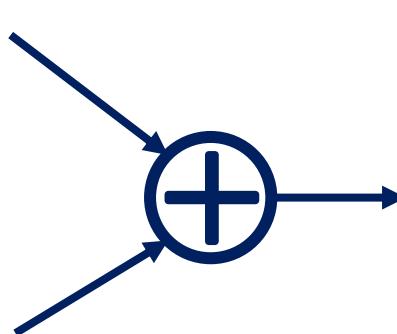
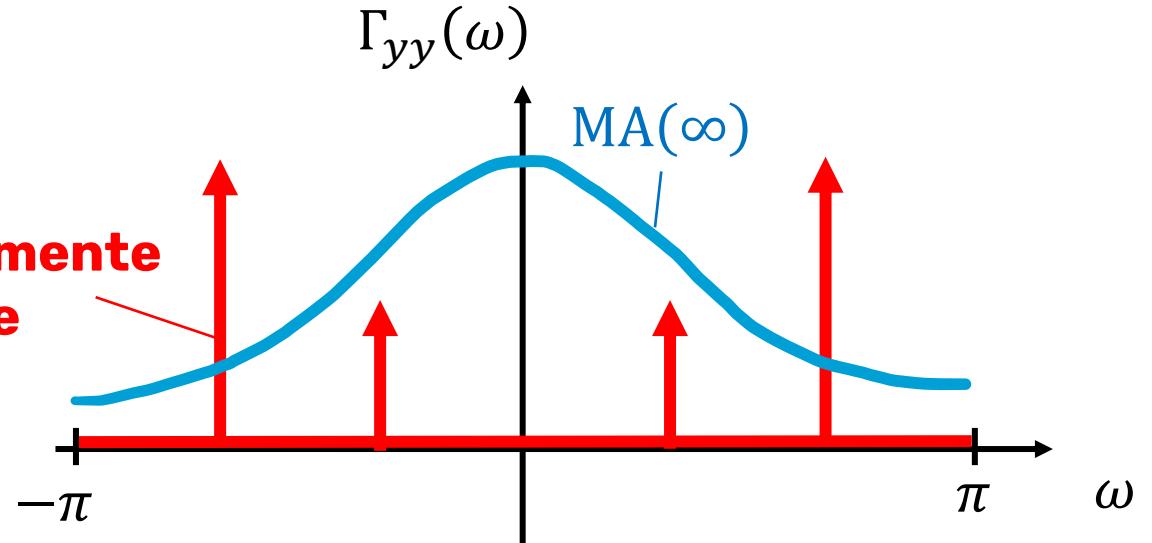
- $y_p(t)$: processo stocastico stazionario **completamente predicibile**
- $\bar{y}(t)$: parte **puramente stocastica**, tale che $\bar{y}(t) = \sum_{i=0}^{+\infty} c_i e(t-i)$, con $e(t) \sim \text{WN}(0, \lambda^2)$,
 $\sum_{i=0}^{+\infty} c_i^2 < \infty$
- $\bar{y}(t)$ e $y_p(t)$ sono **incorrelati**



Scomposizione di Wold



Parte
completamente
predicibile



Scomposizione di Wold nella pratica

Nella pratica, questo risultato ci fornisce una «linea guida» per **stimare modelli di serie temporali**

- Faccio una **stima spettrale** (es. periodogramma) per riconoscere eventuali «righe»
- **Stimo le componenti sinusoidali** $y_p(t)$ (minimi quadrati o «a mano»), ottenendo $\hat{y}_p(t)$
- Ottengo la **componente puramente stocastica** come $\bar{y}(t) = y(t) - \hat{y}_p(t)$
- Risolvo il problema della **predizione** per la parte stocastica $\bar{y}(t)$, ottenendo $\hat{\bar{y}}(t|t - k)$
- **Ottengo la predizione finale** come $\hat{y}(t|t - k) = \hat{y}_p(t) + \hat{\bar{y}}(t|t - k)$



Scomposizione di Wold nella pratica

Osservazioni

Anche eventuali **componenti di non stazionarietà** come **stagionalità** o **trend** devono essere stimate e rimosse dai dati per ottenere solo la parte stocastica del processo

- Esistono modelli di serie temporali più complessi (e.g. SARIMA) che cercano di **modellare la stagionalità**, anziché stimarla prima per poi sottrarla dai dati
- Un **trend** può anche essere un valore costante, e.g. il **valore atteso del processo**. Tale valore può essere visto come la **«componente a frequenza zero»**, che viene rimossa dal processo con la procedura vista precedentemente (per esempio stimando il valore atteso con una media temporale)



Scomposizione di Wold nella pratica

Ipotesi di lavoro

- Facciamo **l'ipotesi** che la parte puramente stocastica $\bar{y}(t)$, ovvero il processo MA(∞), possa essere ben approssimato da **processi a spettro razionale**. Abbiamo già visto che ciò è possibile usando un ARMA
- L'ipotesi non è restrittiva anche perché i coefficienti c_i diventano più piccoli col tempo, e quindi potrei usare anche un MA(n_c)

Nel seguito, **supporremo di lavorare con processi depurati dalle componenti non stazionarie e completamente predicibili** (da cui deriva che avranno media nulla)

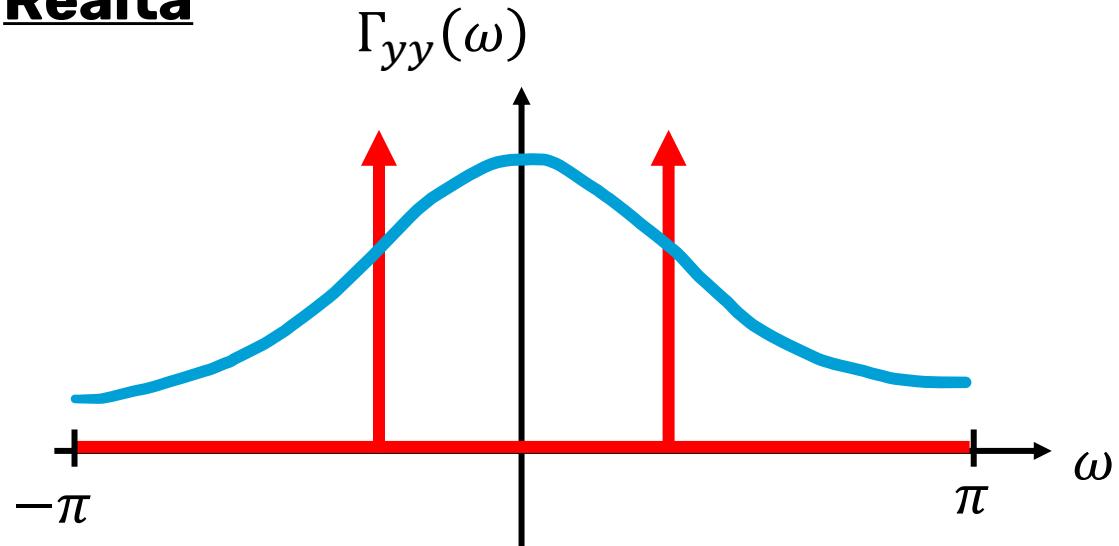


Scomposizione di Wold nella pratica

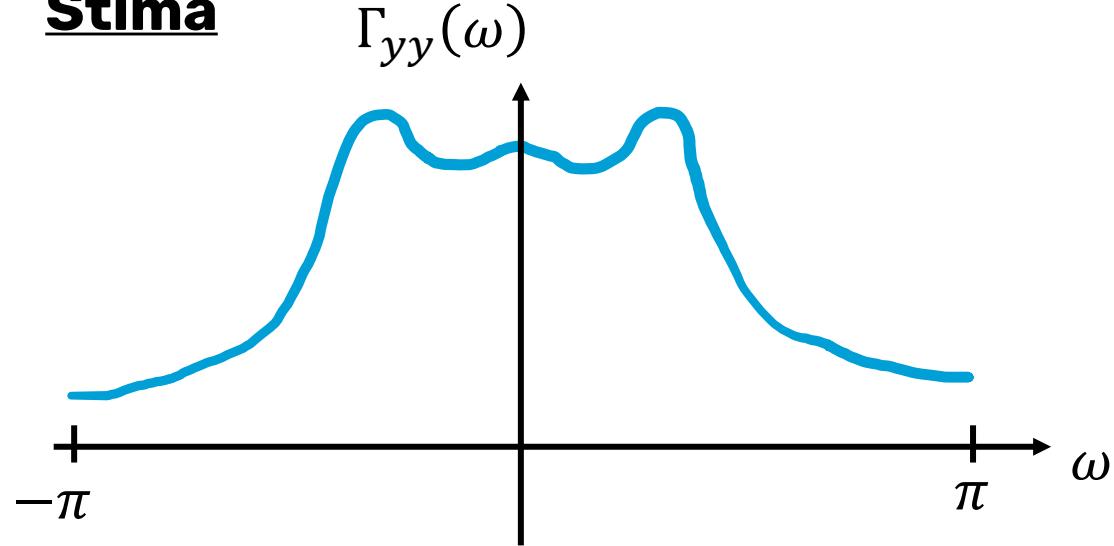
«Estrarre le righe» dal periodogramma **non è così facile**, per due motivi:

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

Realtà



Stima



Outline

1. Predizione, filtraggio e smoothing
2. Scomposizione di Wold
- 3. Filtro passa-tutto e forma canonica**
4. Predittore ottimo
5. Predittore ottimo per processi MA
6. Predittore ottimo per processi ARMA
7. Predittore ottimo per processi ARMAX
8. Predittore ottimo ad un passo per sistemi ingresso\uscita
9. Confronto con il predittore di Kalman



Filtro passa-tutto

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 è il reciproco del polo
- Il fattore moltiplicativo è come il polo

Proviamo a calcolare la **densità 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 $v(t)$

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



Filtro passa-tutto

$$|T(e^{j\omega})|^2 = \left(\frac{1}{a} \cdot \frac{e^{j\omega} + a}{e^{j\omega} + \frac{1}{a}} \right) \cdot \left(\frac{1}{a} \cdot \frac{e^{-j\omega} + a}{e^{-j\omega} + \frac{1}{a}} \right) = \frac{1}{a^2} \cdot \frac{(e^{j\omega} + a) \cdot (e^{-j\omega} + a)}{\left(e^{j\omega} + \frac{1}{a} \right) \cdot \left(e^{-j\omega} + \frac{1}{a} \right)}$$

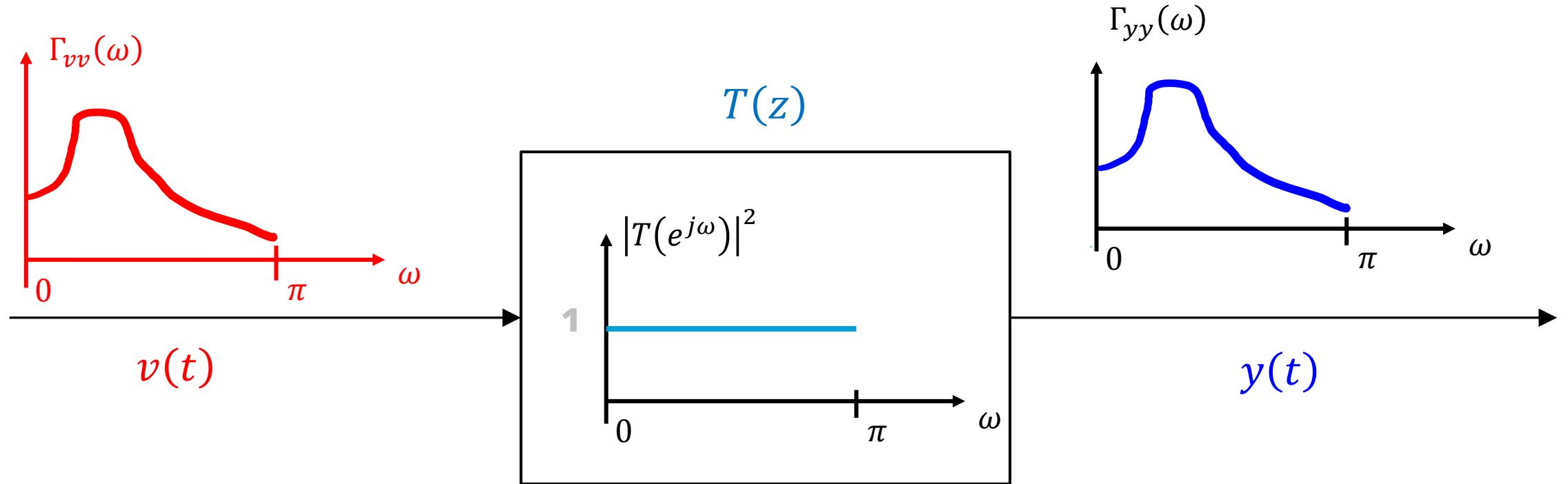
$$= \frac{1}{a^2} \cdot \frac{1 + a^2 + a(e^{j\omega} + e^{-j\omega})}{1 + \frac{1}{a^2} + \frac{1}{a}(e^{j\omega} + e^{-j\omega})} = \frac{1}{a^2} \cdot \frac{1 + a^2 + 2a \cos \omega}{a^2 + 1 + 2a \cos \omega} = 1$$

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

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



Filtro passa-tutto



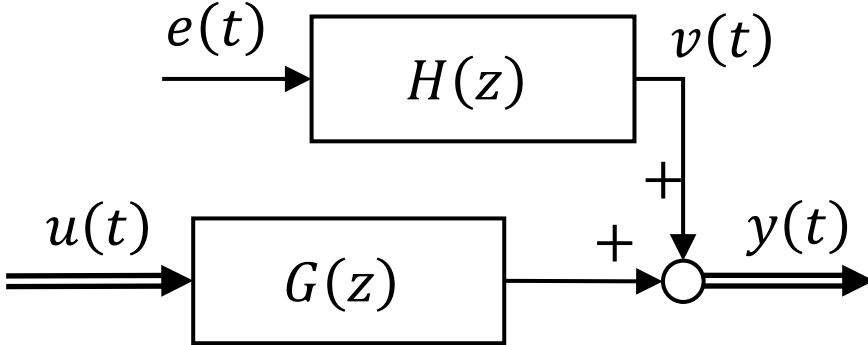
$$\Gamma_{yy}(\omega) = \Gamma_{vv}(\omega)$$



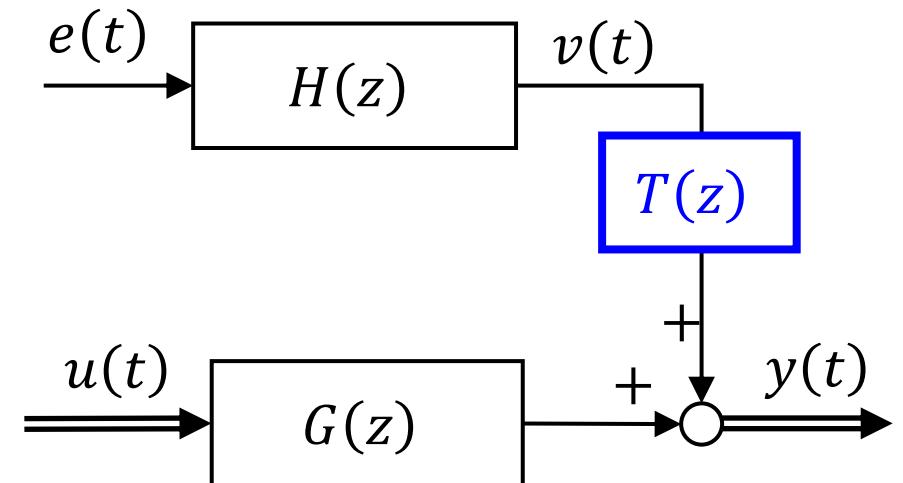
Filtro passa-tutto

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

I due processi $y(t)$ e $v(t)$ **non sono identici** poichè il passa-tutto introduce uno **sfasamento** (come tutti i filtri causali)



**Spettralmente
equivalente a**

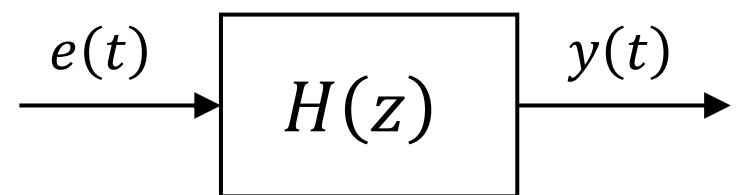


Fattorizzazione spettrale

Abbiamo detto che 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 rumore bianco $e(t) \sim \text{WN}(0, \lambda^2)$

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

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



$$H(z) = C(z)/A(z) \quad \begin{matrix} \textbf{fdt} \\ \textbf{razionale fratta} \end{matrix}$$

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



Forma canonica

Consideriamo questi 5 processi ARMA:

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

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

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

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

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



Forma canonica

Calcoliamo le densità spettrali di potenza dei processi $y_i(t)$

$$1) \quad \Gamma_{y_1 y_1}(\omega) = \left| \frac{e^{j\omega} + \frac{1}{2}}{e^{j\omega} - \frac{1}{3}} \right|^2 \cdot 1$$

$$2) \quad \Gamma_{y_2 y_2}(\omega) = \left| \frac{e^{j\omega} + \frac{1}{2}}{e^{j\omega} - \frac{1}{3}} \cdot e^{-2j\omega} \right|^2 \cdot 1 = \Gamma_{y_1 y_1}(\omega) \cdot |e^{-2j\omega}|^2 = \Gamma_{y_1 y_1}(\omega)$$

$$3) \quad \Gamma_{y_3 y_3}(\omega) = \left| \frac{e^{2j\omega} - \frac{1}{4}}{e^{2j\omega} - \frac{5}{6}e^{j\omega} + \frac{1}{6}} \right|^2 \cdot 1 = \left| \frac{\left(e^{j\omega} - \frac{1}{2} \right) \cdot \left(e^{j\omega} + \frac{1}{2} \right)}{\left(e^{j\omega} - \frac{1}{2} \right) \cdot \left(e^{j\omega} - \frac{1}{3} \right)} \right|^2 = \Gamma_{y_1 y_1}(\omega)$$



Forma canonica

Calcoliamo le densità spettrali di potenza dei processi $y_i(t)$

$$4) \quad \Gamma_{y_4y_4}(\omega) = \left| 2 \cdot \frac{e^{j\omega} + \frac{1}{2}}{e^{j\omega} - \frac{1}{3}} \right|^2 \cdot \frac{1}{4} = 4 \cdot \left| \frac{e^{j\omega} + \frac{1}{2}}{e^{j\omega} - \frac{1}{3}} \right|^2 \cdot \frac{1}{4} = \Gamma_{y_1y_1}(\omega)$$

$$5) \quad \Gamma_{y_5y_5}(\omega) = \left| \frac{z+2}{z-\frac{1}{3}} \right|^2 \cdot \frac{1}{4} = \left| \frac{z+2}{z-\frac{1}{3}} \cdot \frac{z+\frac{1}{2}}{z+\frac{1}{2}} \right|^2 \cdot \frac{1}{4} = \left| 2 \frac{z+\frac{1}{2}}{z-\frac{1}{3}} \right|^2 \cdot \frac{1}{4}$$
$$= \left| 2 \cdot \frac{e^{j\omega} + \frac{1}{2}}{e^{j\omega} - \frac{1}{3}} \right|^2 \cdot \frac{1}{4} = \Gamma_{y_1y_1}(\omega)$$

Uso il **filtro passa-tutto** per cancellare lo **zero fuori dal cerchio**



Forma canonica

Tutti e 5 i processi visti sono **equivalenti** (hanno la stessa densità spettrale di potenza).

Le **cause di non univocità** sono:

1. **Ritardi** puri (processo **2**)
2. **Fattori moltiplicativi** che si cancellano (processo **3**)
3. **Coefficienti moltiplicativi** che si compensano tra funzione di trasferimento e spettro dell'ingresso (processo **4**)
4. **Poli\zeri reciproci** (processo **5**)

La **forma canonica** di un processo stocastico stazionario a spettro razionale è univoca e ci permetterà di risolvere il problema della predizione



Teorema della fattorizzazione spettrale

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

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**



Esempio: calcolo della forma canonica

Consideriamo il seguente processo ARMA(1,1)

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

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



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



Outline

1. Predizione, filtraggio e smoothing
2. Scomposizione di Wold
3. Filtro passa-tutto e forma canonica
- 4. Predittore ottimo**
5. Predittore ottimo per processi MA
6. Predittore ottimo per processi ARMA
7. Predittore ottimo per processi ARMAX
8. Predittore ottimo ad un passo per sistemi ingresso\uscita
9. Confronto con il predittore di Kalman



Predittore ottimo

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

Informazioni disponibili

- **Dati** $y(1), y(2), \dots, y(N)$
- Vecchie **predizioni** $\hat{y}(t - 1|t - k - 1), \hat{y}(t - 2|t - k - 2), \dots$
- **Modello** della parte stocastica del processo $C(z)/A(z)$

Ipotesi di lavoro

- Supponiamo $y(t)$ un pss **puramente stocastico**, depurato da componenti predibili
- Modello $C(z)/A(z)$ **in forma canonica**



Predittore ottimo

Esistono diversi modi per definire un **predittore**, per esempio potremmo usare:

$$\hat{y}(t+1|t) = \frac{y(t) + y(t-1) + y(t-2)}{3}$$

Media di alcuni valori passati

$$\hat{y}(t+1|t) = \frac{2y(t) + \frac{1}{2}y(t-1) + \frac{1}{2}y(t-2)}{3}$$

Diamo più peso a valori più recenti

Vogliamo però trovare il **predittore lineare ottimo** dai dati, ovvero quello che minimizza il seguente criterio **Mean Squared Error (MSE)**

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



Predittore ottimo

Definizione: Un predittore (lineare) è **ottimo** se

1. $\mathbb{E}[\varepsilon_k(t)] = \mathbb{E}[y(t) - \hat{y}(t|t-k)] = 0$, i.e. il **valore atteso** dell'errore di predizione è **nullo**

Significa che il processo $y(t)$ e il predittore $\hat{y}(t|t-k)$ hanno lo **stesso valore atteso**

2. $\mathbb{E}[\hat{y}(t|t-k) \cdot \varepsilon_k(t)] = 0$, i.e. il predittore e l'errore di predizione sono **incorrelati**

Significa che **il predittore ha utilizzato tutta l'informazione disponibile**. Se $\hat{y}(t|t-k)$ e $\varepsilon_k(t)$ fossero correlati, significa che «c'è qualcosa» in $\varepsilon_k(t)$ che c'è anche in $\hat{y}(t|t-k)$. Ma allora questo qualcosa avrebbe dovuto stare in $\hat{y}(t|t-k)$ per «aiutarlo» nella previsione

3. $\text{Var}[\varepsilon_k(t)]$ **minima**



Predittore ottimo

Avendo definito l'**errore di predizione** come

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

possiamo **scomporre il processo** $y(t)$ come

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

dove:

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



Outline

1. Predizione, filtraggio e smoothing
2. Scomposizione di Wold
3. Filtro passa-tutto e forma canonica
4. Predittore ottimo
- 5. Predittore ottimo per processi MA**
6. Predittore ottimo per processi ARMA
7. Predittore ottimo per processi ARMAX
8. Predittore ottimo ad un passo per sistemi ingresso\uscita
9. Confronto con il predittore di Kalman



Predittore ottimo per processi MA

PREDITTORE AD UN PASSO

Consideriamo un processo MA(n_c) in **forma canonica**

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

Un possibile predittore potrebbe quindi essere dato dalla **parte predicibile** al tempo $t - 1$

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



Predittore ottimo per processi MA

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

Osservazioni

- $\hat{y}(t|t-1)$ è corretto, infatti $\mathbb{E}[y(t)] = \mathbb{E}[\hat{y}(t|t-1)] = 0$
- $\hat{y}(t|t-1)$ dipende dal WN fino al tempo $t-1$
- $\mathbb{E}[\hat{y}(t|t-1) \cdot \varepsilon_1(t)] = 0$, infatti $\varepsilon_1(t) = y(t) - \hat{y}(t|t-1) = e(t)$ e quindi

$$\mathbb{E}[\hat{y}(t|t-1) \cdot \varepsilon_1(t)] = \mathbb{E}[c_1 e(t-1) + c_2 e(t-2) + \cdots + c_{n_c} e(t-n_c) \cdot e(t)] = 0$$

- Non è possibile trovare un predittore con $\text{Var}[\varepsilon_1(t)] < \text{Var}[e(t)]$

Ne consegue che $\hat{y}(t|t-1)$ è il **predittore lineare ottimo**



Predittore ottimo per processi MA

Tuttavia, l'espressione di $\hat{y}(t|t - 1)$ **dipende dal rumore** $e(t)$, e non **dai dati** $y(t)$. Troviamo il «predittore dai dati» osservando che

$$y(t) = [1 + c_1z^{-1} + c_2z^{-2} + \dots + c_{n_c}z^{-n_c}] \cdot e(t) \quad \Rightarrow \quad e(t) = \underbrace{\frac{1}{1 + c_1z^{-1} + c_2z^{-2} + \dots + c_{n_c}z^{-n_c}} y(t)}$$

$$\hat{y}(t|t - 1) = c_1e(t - 1) + c_2e(t - 2) + \dots + c_{n_c}e(t - n_c)$$

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

FILTRO SBIANCANTE
(è stabile grazie alla forma canonica)



$$\hat{y}(t|t - 1) = \frac{c_1z^{-1} + c_2z^{-2} + \dots + c_{n_c}z^{-n_c}}{1 + c_1z^{-1} + c_2z^{-2} + \dots + c_{n_c}z^{-n_c}} y(t)$$

Predittore ottimo dai dati ad un passo per processi MA(n_c)



Predittore ottimo per processi MA

Passando in **forma ricorsiva** si ottiene

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

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

Osservazione

Quando il processo ha una componente MA, il predittore **è dinamico**. C'è bisogno di definire il valore della **condizione iniziale** $\hat{y}(1|0)$. Di solito si usa la media del processo (i.e. zero). Se il predittore è asintoticamente stabile, **l'effetto dell'inizializzazione svanisce** col tempo



Predittore ottimo per processi MA

PREDITTORE A k PASSI

Consideriamo un processo MA(n_c) in **forma canonica**, con $e(t) \sim WN(0, \lambda^2)$

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

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

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



Predittore ottimo per processi MA

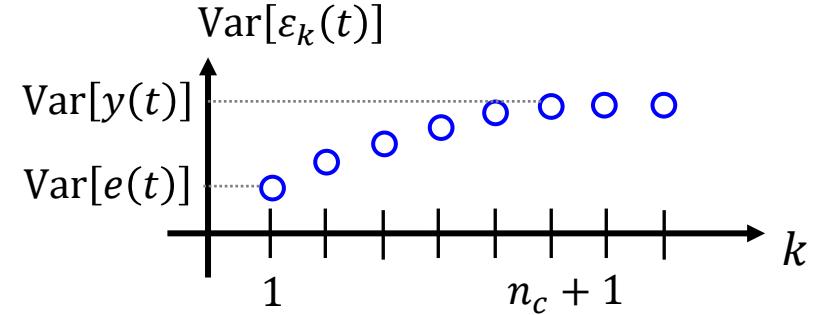
Osservazione

$$\varepsilon_1(t) = y(t) - \hat{y}(t|t-1) \quad \Rightarrow \quad \text{Var}[\varepsilon_1(t)] = \text{Var}[e(t)] = \lambda^2$$

$$\varepsilon_2(t) = y(t) - \hat{y}(t|t-2) \quad \Rightarrow \quad \text{Var}[\varepsilon_2(t)] = \text{Var}[e(t) + c_1 e(t-1)] = (1 + c_1^2)\lambda^2 > \lambda^2$$

:

$$\begin{aligned} \varepsilon_{n_c+1}(t) = y(t) - \hat{y}(t|t-n_c-1) &\Rightarrow \text{Var}[\varepsilon_{n_c+1}(t)] = \text{Var}[e(t) + c_1 e(t-1) + \dots + c_{n_c} e(t-n_c)] \\ &= \text{Var}[y(t)] \end{aligned}$$



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



Outline

1. Predizione, filtraggio e smoothing
2. Scomposizione di Wold
3. Filtro passa-tutto e forma canonica
4. Predittore ottimo
5. Predittore ottimo per processi MA
- 6. Predittore ottimo per processi ARMA**
7. Predittore ottimo per processi ARMAX
8. Predittore ottimo ad un passo per sistemi ingresso\uscita
9. Confronto con il predittore di Kalman



Predittore ottimo per processi ARMA

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

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

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

Problema: In questo caso, **non è immediatamente chiaro come scomporre** la parte impredicibile da quella predicibile, poiché $y(t)$ dipende anche da $y(t-1), y(t-2), \dots$ i quali dipendono da $e(t), e(t-1), \dots$



Predittore ottimo per processi ARMA

Idea: si esprime $C(z)/A(z)$ come un **quoziente** $E(z)$ più un **resto** $R(z) = z^{-k}\tilde{R}(z)$ effettuando una **lunga divisione** tra polinomi

Quoziente	Resto	
$C(z) = E(z) A(z) + R(z)$		\Rightarrow
Didivendo	Divisore	

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

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

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



Esempio: lunga divisione

Consideriamo il processo ARMA($n_a = 1, n_c = 1$) e facciamo $k = 2$ passi di lunga divisione

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

- $E(z)e(t) = e(t) + \frac{1}{6}e(t-1)$ è **impredicibile** al tempo $t-2$
- $\frac{\tilde{R}(z)}{A(z)}e(t-k) = -\frac{\frac{1}{18}}{A(z)}e(t-2)$ è **predicibile** al tempo $t-2$

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



Predittore ottimo per processi ARMA

Il **predittore ottimo dal rumore** è

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

Calcoliamo il **predittore ottimo dai dati** tramite il filtro sbiancante

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

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



Predittore ottimo per processi ARMA

Il **predittore ottimo dai dati** è

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

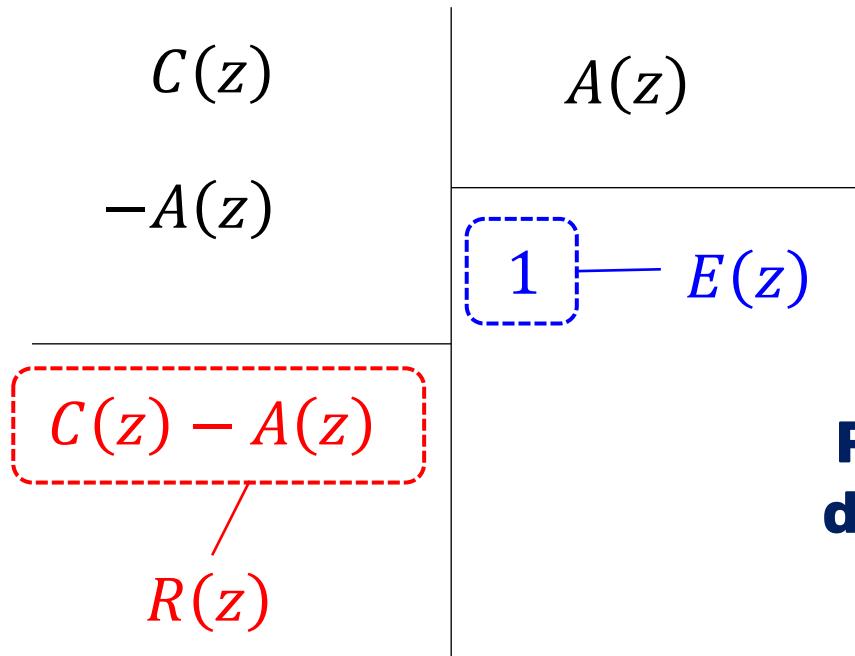
L'**errore di predizione** corrispondente è

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



Predittore ottimo per processi ARMA

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



$$E(z) = 1 \quad \Rightarrow \quad \varepsilon_1(t) = E(z)e(t) = e(t)$$

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

Predittore dai dati a un passo

Errore di predizione

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

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



Predittore ottimo per processi ARMA

Osservazioni

- $\hat{y}(t|t - k)$ è corretto, infatti $\mathbb{E}[y(t)] = \mathbb{E}[\hat{y}(t|t - k)] = 0$
- $\hat{y}(t|t - k)$ dipende dal WN fino al tempo $t - k$
- $\mathbb{E}[\hat{y}(t|t - k) \cdot \varepsilon_k(t)] = 0$, infatti $\mathbb{E}[\hat{y}(t|t - k) \cdot \varepsilon_k(t)] = \mathbb{E}\left[\left(\frac{\tilde{R}(z)}{A(z)} e(t - k)\right) \cdot (E(z)e(t))\right] = 0$
- Si dimostra che non è possibile trovare un predittore con $\text{Var}[\varepsilon_k(t)]$ minore

Ne consegue che $\hat{y}(t|t - k)$ è il **predittore lineare ottimo**



Qualità del predittore

Possiamo valutare la **qualità del predittore** mettendo a confronto la **varianza dell'errore di predizione** ottenuto con la varianza dell'errore di predizione di un **predittore banale** (che predice sempre la media processo, cioè sempre zero)

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

- Il valore $1 - \text{ESR}$ ci fornisce la **percentuale di varianza del processo che è stata «catturata» dal predittore**
- L'ESR varia tra 0 e 1. Un valore di ESR **inferiore** indica un predittore **migliore**



Outline

1. Predizione, filtraggio e smoothing
2. Scomposizione di Wold
3. Filtro passa-tutto e forma canonica
4. Predittore ottimo
5. Predittore ottimo per processi MA
6. Predittore ottimo per processi ARMA
- 7. Predittore ottimo per processi ARMAX**
8. Predittore ottimo ad un passo per sistemi ingresso\uscita
9. Confronto con il predittore di Kalman



Predittore ottimo per processi ARMAX

Sia dato un processo ARMAX(n_a, n_c, n_b, k), con $C(z)/A(z)$ in **forma canonica**

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

- $C(z) = 1 + c_1 z^{-1} + c_2 z^{-2} + \cdots + c_{n_c} z^{-n_c}$
- $A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - \cdots - a_{n_a} z^{-n_a}$
- $B(z) = b_0 + b_1 z^{-1} + \cdots + b_{n_b} z^{-n_b}$

In questo caso, è **sensato fare una previsione a k passi**, in modo che l'ingresso riesca ad influenzare l'uscita. Quindi, «confondiamo» i k passi di previsione con i k passi di ritardo puro tra ingresso e uscita



Predittore ottimo per processi ARMAX

Applichiamo k passi di lunga divisione per scomporre $C(z)/A(z)$

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

Il **predittore ottimo dal rumore** è

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



Predittore ottimo per processi ARMAX

Calcoliamo il **predittore ottimo dai dati**

FILTRO SBIANCANTE

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

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



Predittore ottimo per processi ARMAX

Il **predittore ottimo dai dati** è

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

L'**errore di predizione** corrispondente è

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



Predittore ottimo per processi ARMAX

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

$$E(z) = 1 \quad \Rightarrow \quad \varepsilon_1(t) = E(z)e(t) = e(t)$$

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

Predittore dai dati a un passo

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

Errore di predizione a un passo

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



Predittore ottimo per processi ARMAX

Osservazioni

- $\hat{y}(t|t - k)$ è corretto, infatti $\mathbb{E}[y(t)] = \mathbb{E}[\hat{y}(t|t - k)] = 0$
- $\hat{y}(t|t - k)$ dipende dal WN e da $u(t)$ fino al tempo $t - k$
- $\mathbb{E}[\hat{y}(t|t - k) \cdot \varepsilon_k(t)] = 0$. Nel caso in cui $u(t)$ fosse un processo stocastico, si assume ragionevolmente che $u(t) \perp e(t)$
- Non è possibile trovare un predittore con $\text{Var}[\varepsilon_k(t)]$ minore

Ne consegue che $\hat{y}(t|t - k)$ è il **predittore lineare ottimo**



Predittore ottimo per processi ARMAX

Osservazione

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

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



Esempio: forma canonica e calcolo del predittore

Sia dato il processo $y(t)$. Calcolare il predittore dai dati e la varianza dell'errore di predizione

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

Il ritardo puro è $k = 2$. Ha quindi senso calcolare un predittore per $k = 2$ passi in avanti

Il processo è in forma canonica?

$$y(t) = (2 + 6z^{-1})u(t - 2) + \frac{1}{1 + \frac{1}{2}z^{-1}} \cdot \frac{2}{3}e(t - 1) \xrightarrow{\text{red box}} \frac{2}{3}e(t - 1) \equiv \eta(t) \sim WN\left(0, \frac{4}{9}\right)$$



Esempio: forma canonica e calcolo del predittore

$$y(t) = (2 + 6z^{-1})u(t - 2) + \frac{1}{1 + \frac{1}{2}z^{-1}} \cdot \eta(t)$$

Per calcolare il predittore, la parte esogena e quella stocastica devono avere il medesimo polinomio $A(z)$ al denominatore

$$y(t) = \frac{(2 + 6z^{-1})\left(1 + \frac{1}{2}z^{-1}\right)}{1 + \frac{1}{2}z^{-1}} u(t - 2) + \frac{1}{1 + \frac{1}{2}z^{-1}} \cdot \eta(t)$$
$$\eta(t) \sim \text{WN}\left(0, \frac{4}{9}\right)$$

The equation is shown with three parts highlighted by colored dotted rectangles: $B(z)$ (red, top), $A(z)$ (green, bottom), and $C(z)$ (pink, right). A green arrow points from the $A(z)$ box to the denominator of the $C(z)$ term.



Esempio: forma canonica e calcolo del predittore

Notiamo che il denominatore comune è stato fatto dopo la canonizzazione di $C(z)/A(z)$. La fdt $B(z)/A(z)$ non ha bisogno di essere in forma canonica, in quanto non è la parte stocastica. Calcoliamo il **predittore** a $k = 2$ passi in avanti

$$\begin{array}{r} \boxed{1} \xrightarrow{\quad C(z) \quad} \\ -1 - \frac{1}{2}z^{-1} \\ \hline -\frac{1}{2}z^{-1} \\ \\ \frac{1}{2}z^{-1} + \frac{1}{4}z^{-2} \\ \hline R(z) \xrightarrow{\quad \boxed{+\frac{1}{4}z^{-2}} \quad} \end{array}$$

The diagram shows the partial fraction decomposition of the system. The first row contains the terms 1 and $-1 - \frac{1}{2}z^{-1}$, with 1 highlighted by a pink dashed box and $-1 - \frac{1}{2}z^{-1}$ highlighted by a green dashed box. The second row contains the term $-\frac{1}{2}z^{-1}$. The third row contains the terms $\frac{1}{2}z^{-1} + \frac{1}{4}z^{-2}$, with the entire sum highlighted by a red dashed box. The fourth row contains the term $R(z)$, with the constant term $+ \frac{1}{4}z^{-2}$ highlighted by a red dashed box.

$$E(z) = 1 - \frac{1}{2}z^{-1} \quad R(z) = z^{-k}\tilde{R}(z) = \frac{1}{4}z^{-2}$$

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

$$\hat{y}(t|t-2) = \frac{\frac{1}{4}}{1}y(t-2) + \frac{2(1+3z^{-1})\left(1+\frac{1}{2}z^{-1}\right)\left(1-\frac{1}{2}z^{-1}\right)}{1}u(t-2)$$



Esempio: forma canonica e calcolo del predittore

Esprimiamo il **predittore in forma ricorsiva**

$$\hat{y}(t|t-2) = \frac{1}{4}y(t-2) + 2(1+3z^{-1})\left(1+\frac{1}{2}z^{-1}\right)\left(1-\frac{1}{2}z^{-1}\right)u(t-2)$$

$$= \frac{1}{4}y(t-2) + 2(1+3z^{-1})\left(1-\frac{1}{4}z^{-2}\right)u(t-2)$$

$$\hat{y}(t|t-2) = \frac{1}{4}y(t-2) + 2u(t-2) + 6u(t-3) - \frac{1}{2}u(t-4) - \frac{3}{2}u(t-5)$$

Notiamo che nell'espressione del predittore non vi sono termini «prima» di $t-2$



Esempio: forma canonica e calcolo del predittore

Calcoliamo la **varianza dell'errore di predizione**

$$\begin{aligned}\text{Var}[\varepsilon_2(t)] &= \mathbb{E}[\varepsilon_2(t)^2] = \mathbb{E}\left[\left(E(z)\eta(t)\right)^2\right] = \mathbb{E}\left[\left(\left(1 - \frac{1}{2}z^{-1}\right)\eta(t)\right)^2\right] = \left[1 + \frac{1}{4}\right]\text{Var}[\eta(t)] \\ &= \frac{5}{4} \cdot \frac{4}{9} = \boxed{\frac{5}{9}}\end{aligned}$$

Calcoliamo la **bontà del predittore ottimo** rispetto a quella del **predittore banale**

$$\hat{y}(t|t-2) = \mathbb{E}[y(t)] = 0$$

$$\text{ESR} = \frac{\text{Var}[\varepsilon_2(t)]}{\text{Var}\left[\frac{C(z)}{A(z)}\eta(t)\right]}$$



Esempio: forma canonica e calcolo del predittore

$$\text{Var}\left[\frac{C(z)}{A(z)}\eta(t)\right] = \text{Var}\left[\frac{1}{1 + \frac{1}{2}z^{-1}}\eta(t)\right] = \text{Var}[\nu(t)] \quad \text{con} \quad \nu(t) = \frac{1}{1 + \frac{1}{2}z^{-1}}\eta(t)$$

$$\Rightarrow \nu(t) = -\frac{1}{2}\nu(t-1) + \eta(t)$$

$$\Rightarrow \text{Var}[\nu(t)] = \text{Var}\left[-\frac{1}{2}\nu(t-1) + \eta(t)\right] = \mathbb{E}\left[\left(-\frac{1}{2}\nu(t-1) + \eta(t)\right)^2\right]$$

$$= \mathbb{E}\left[\frac{1}{4}\nu(t-1)^2 + \eta(t)^2 - \nu(t-1)\eta(t)\right] = \frac{1}{4}\text{Var}[\nu(t)] + \text{Var}[\eta(t)]$$

$$\frac{3}{4}\text{Var}[\nu(t)] = \text{Var}[\eta(t)] \quad \Rightarrow$$

$$\boxed{\text{Var}[\nu(t)] = \frac{4}{3} \cdot \frac{4}{9} = \frac{16}{27}}$$



Esempio: forma canonica e calcolo del predittore

$$\text{ESR} = \frac{\text{Var}[\varepsilon_2(t)]}{\text{Var}\left[\frac{C(z)}{A(z)}\eta(t)\right]} = \frac{\frac{5}{9}}{\frac{16}{27}} = \frac{5}{9} \cdot \frac{27}{16} = \frac{15}{16} = 0.9375$$

Il predittore ottimo **ha ridotto l'incertezza con cui prevediamo «due passi in avanti»** circa del 7%.

Nota: «Predire il 7%» di un processo, anche se sembra poco, non significa sia inutile: dipende dal contesto applicativo



Outline

1. Predizione, filtraggio e smoothing
2. Scomposizione di Wold
3. Filtro passa-tutto e forma canonica
4. Predittore ottimo
5. Predittore ottimo per processi MA
6. Predittore ottimo per processi ARMA
7. Predittore ottimo per processi ARMAX
- 8. Predittore ottimo ad un passo per sistemi ingresso\uscita**
9. Confronto con il predittore di Kalman

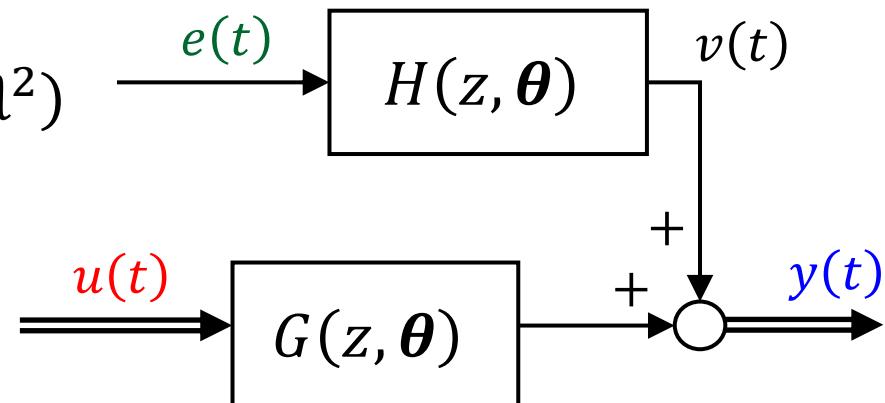


Predittore ottimo ad un passo per sistemi I\O

Abbiamo visto come, nel caso di sistemi dinamici **LTI SISO ingresso\uscita**, usiamo un modello $\mathcal{M}(\boldsymbol{\theta})$ della forma seguente

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

$$G(z, \boldsymbol{\theta}) = \sum_{i=1}^{\infty} g_i(\boldsymbol{\theta}) \cdot z^{-i} \quad H(z) = 1 + \sum_{i=1}^{\infty} h_i(\boldsymbol{\theta}) \cdot z^{-i}$$



Si suppone $H(z)$ **in forma canonica**

Notiamo che il **filtro sbiancante** si ottiene come

$$e(t) = H^{-1}(z, \boldsymbol{\theta})[y(t) - G(z, \boldsymbol{\theta})u(t)]$$



Predittore ottimo ad un passo per sistemi I\O

$$y(t) = G(z, \theta)u(t) + H(z, \theta)e(t) \quad \text{sommo e tolgo } e(t)$$

$$= G(z, \theta)u(t) + [H(z, \theta) - 1]e(t) + e(t)$$

Sostituendo l'espressione del **filtro sbiancante** che produce $e(t)$ nel secondo termine

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

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

Dato che $H(z, \theta)$ è in **forma canonica**, anche $H^{-1}(z, \theta)$ è in forma canonica, per cui

$$\frac{1}{H(z, \theta)} = 1 + d_1z^{-1} + d_2z^{-2} + \dots$$



Predittore ottimo ad un passo per sistemi I\O

Supponendo che $G(z, \theta)$ sia **strettamente propria** (i.e. almeno un passo di ritardo tra ingresso e uscita), si ha che la quantità

$$H^{-1}(z, \theta)G(z, \theta)u(t) + [1 - H^{-1}(z, \theta)]y(t)$$

dipende solo da $H(z, \theta)$, $G(z, \theta)$ e dai dati $u(t-1), u(t-2), \dots$ e $y(t-1), y(t-2), \dots$

Questa quantità è quindi **completamente predicibile** al tempo $t-1$



Predittore ottimo ad un passo per sistemi I\O

Il **predittore ottimo ad un passo** $\hat{\mathcal{M}}(\boldsymbol{\theta})$ per la **classe di modelli** $\mathcal{M}(\boldsymbol{\theta})$ è dato da

$$\hat{\mathcal{M}}(\boldsymbol{\theta}): \hat{y}(t|t-1; \boldsymbol{\theta}) = H^{-1}(z, \boldsymbol{\theta})G(z, \boldsymbol{\theta})u(t) + [1 - H^{-1}(z, \boldsymbol{\theta})]y(t)$$

L'**errore di predizione ad un passo** $\varepsilon_1(t; \boldsymbol{\theta}) = y(t) - \hat{y}(t|t-1; \boldsymbol{\theta})$ è

$$\varepsilon_1(t; \boldsymbol{\theta}) = H^{-1}(z, \boldsymbol{\theta})[y(t) - G(z, \boldsymbol{\theta})u(t)]$$



Predittore ottimo ad un passo per sistemi I\O

Osservazione

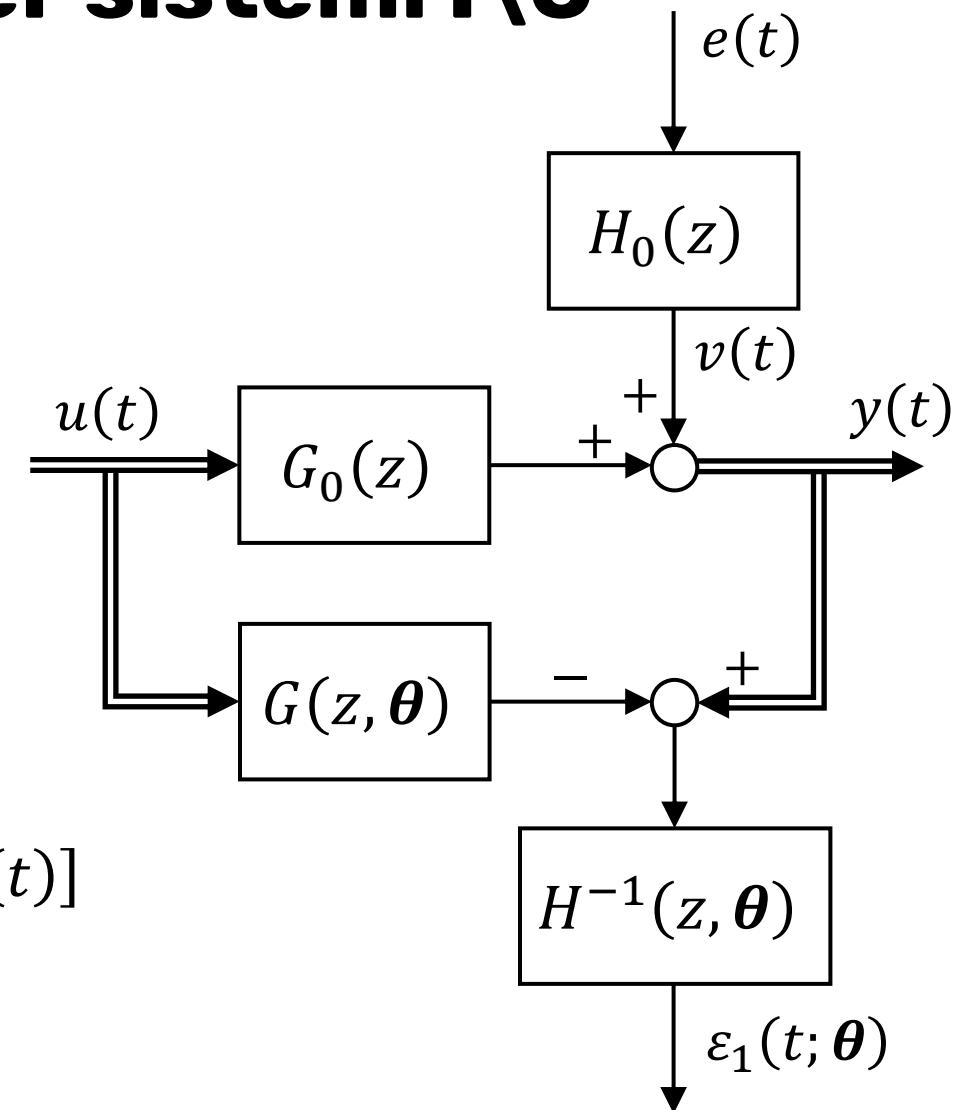
Sostituendo l'equazione del **sistema che genera i**

dati $y(t) = G_0(z)u(t) + H_0(z)e(t)$ all'interno dell'

errore di predizione a un passo $\varepsilon_1(t)$ si ha

$$\varepsilon_1(t; \theta) = H^{-1}(z, \theta)[y(t) - G(z, \theta)u(t)]$$

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



Predittore ottimo ad un passo per sistemi I\O

Se $\exists \theta = \theta^0$ t.c. $G_0(z) = G(z, \theta^0)$ e $H_0(z) = H(z, \theta^0)$, ovvero, **se il sistema vero appartiene alla famiglia di modelli scelta**, otteniamo che

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

Quindi, il valore θ^0 :

1. È **l'unico valore** che rende $\varepsilon_1(t; \theta^0) = e(t)$
2. **Minimizza la varianza** dell'errore di predizione a un passo

Ne consegue che $\varepsilon_1(t)$ è un buon indicatore della **bontà di un modello dinamico**.

Useremo questa proprietà per trovare un **criterio di identificazione** dei modelli dinamici



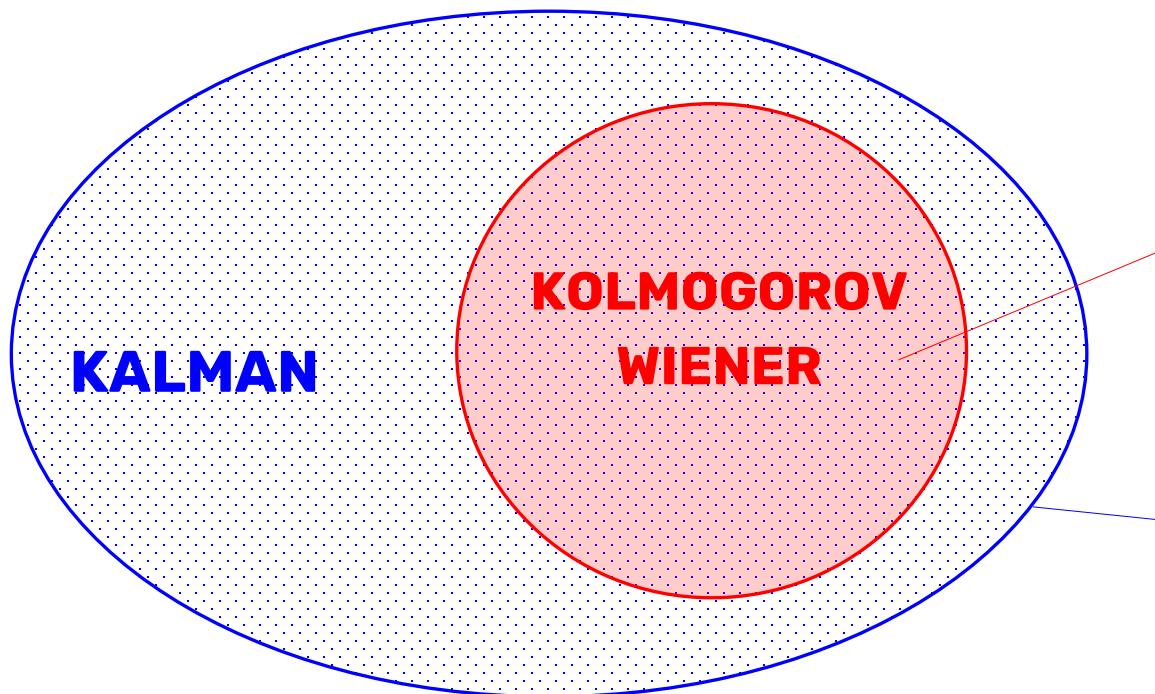
Outline

1. Predizione, filtraggio e smoothing
2. Scomposizione di Wold
3. Filtro passa-tutto e forma canonica
4. Predittore ottimo
5. Predittore ottimo per processi MA
6. Predittore ottimo per processi ARMA
7. Predittore ottimo per processi ARMAX
8. Predittore ottimo ad un passo per sistemi ingresso\uscita
- 9. Confronto con il predittore di Kalman**



Confronto con il predittore di Kalman

La teoria della predizione che abbiamo visto è anche nota come **predizione alla Kolmogorov-Wiener**. È interessante confrontarla con **la teoria della predizione di Kalman** per sistemi dinamici espressi in **spazio di stato**



Sistemi:

- Lineari
- Tempo Invarianti
- Asintoticamente stabili
- A regime

Sistemi:

- Non lineari
- Tempo varianti
- Non asintoticamente stabili
- In transitorio



Confronto con il predittore di Kalman

La teoria di Kalman è quindi **più generale** della teoria di Kolmogorov-Wiener. Però, la teoria KW ci fornisce la base per lo **sviluppo di metodi di identificazione** intuitivi ed efficaci

Nella prossima lezione, ci baseremo sulla teoria Kolmogorov-Wiener per definire metodi di identificazione dei modelli basati sulla **minimizzazione della varianza dell'errore di predizione**





**UNIVERSITÀ
DEGLI STUDI
DI BERGAMO**

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



IDENTIFICAZIONE DEI MODELLI E ANALISI DEI DATI (IMAD)

Lezione 11: Identificazione – concetti fondamentali

Corso di Laurea Magistrale in
INGEGNERIA INFORMATICA

SPEAKER
Prof. Mirko Mazzoleni

PLACE
Università degli Studi di
Bergamo

Syllabus

Parte II: sistemi dinamici

8. Processi stocastici

- 8.1 Processi stocastici stazionari (pss)
- 8.3 Rappresentazione spettrale di un pss
- 8.4 Stimatori campionari media\covarianza
- 8.5 Densità spettrale campionaria

9. Famiglie di modelli a spettro razionale

- 9.1 Modelli per serie temporali (MA, AR, ARMA)
- 9.2 Modelli per sistemi input/output (ARX, ARMAX)

10. Predizione

- 10.1 Filtro passa-tutto

10.2 Forma canonica

10.3 Teorema della fattorizzazione spettrale

10.4 Soluzione al problema della predizione

11. Identificazione

- 11.3 Identificazione di modelli ARX
- 11.4 Identificazione di modelli ARMAX
- 11.5 Metodo di Newton

12. Identificazione: analisi e complementi

- 12.1 Analisi asintotica metodi PEM
- 12.2 Identificabilità dei modelli
- 12.3 Valutazione dell'incertezza di stima

13. Identificazione: valutazione



Parte I: sistemi statici**Parte II: sistemi dinamici**Stima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

- ✓ Stima parametri popolazione
- ✓ Stima modello lineare: minimi quadrati

○ **SI assunzioni su ddp dei dati**

- ✓ Stima massima verosimiglianza parametri popolazione
- ✓ Stima modello lineare: massiva verosimiglianza
- ✓ Regressione logistica

• θ variabile casuale○ **SI assunzioni su ddp dei dati**

- ✓ Stima Bayesiana

Machine learningStima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

- ✓ Modelli lineari di pss
- ✓ Predizione
- ✓ Identificazione
- ✓ Persistente eccitazione
- ✓ Analisi asintotica metodi PEM
- ✓ Analisi incertezza stima (numero dati finito)
- ✓ Valutazione del modello

Outline

1. Introduzione all'identificazione dei modelli dinamici
2. Metodi a minimizzazione dell'errore di predizione (PEM)
3. Identificazione PEM di modelli ARX
4. Identificazione PEM di modelli ARMAX



Outline

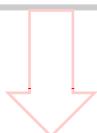
- 1. Introduzione all'identificazione dei modelli dinamici**
2. Metodi a minimizzazione dell'errore di predizione (PEM)
3. Identificazione PEM di modelli ARX
4. Identificazione PEM di modelli ARMAX



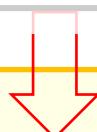
Gli step per la risoluzione del problema

Seguiremo tre fasi per risolvere il problema della **modellazione di sistemi dinamici**:

Definizione delle **classi di modelli** \mathcal{M} di sistemi dinamici



Predizione



Identificazione

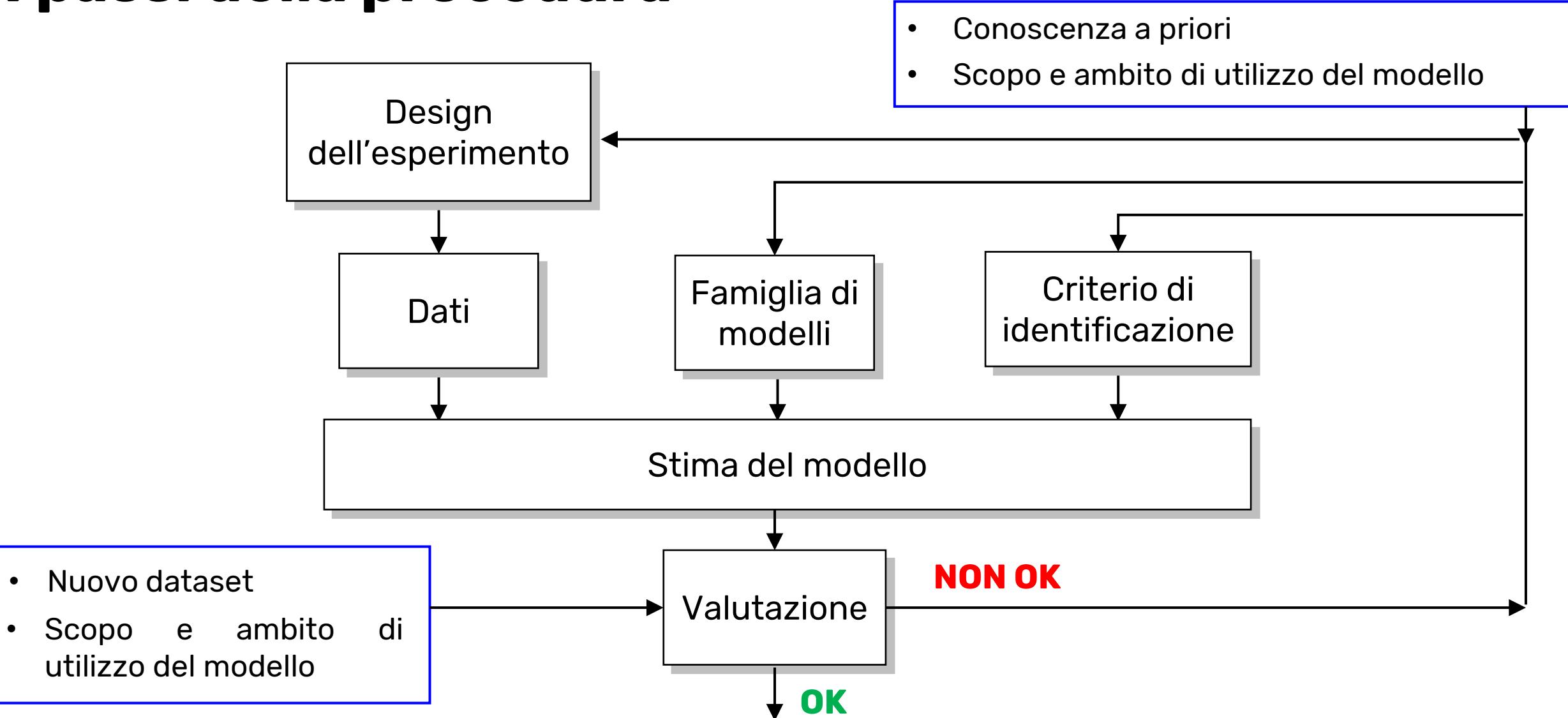
Ci concentreremo su modelli di **sistemi dinamici lineari**, espressi da **funzioni di trasferimento razionali fratte**. I parametri ignoti sono i coefficienti dei polinomi al numeratore e denominatore

Data una particolare classe di modello, supponendo di conoscerne il valore dei parametri, qual è il **preditore ottimo**? Quanto vale la predizione ottima?

Come **stimo il valore dei parametri** del modello scelto per la modellazione dei dati?



I passi della procedura



I passi della procedura di identificazione

Dati: i dati possono essere raccolti o dal funzionamento nominale del sistema, oppure tramite **esperimenti progettati ad-hoc**, in modo da ottenere **specifiche informazioni** (guadagno, risposte a scalino, risposta in frequenza,...)

Famiglia di modelli: è necessario scegliere quale tipo di modello usare per spiegare il fenomeno. Possibili scelte sono:

- **Lineare** \ non lineare
- **Tempo invariante** \ tempo variante
- **Discreto** \ continuo
- Altre proprietà (e.g. complessità del modello, struttura,...)



I passi della procedura di identificazione

Criterio di identificazione: avendo a disposizione le misure e la famiglia di modelli scelta, è necessario decidere **come confrontare il modello con i dati**. Ciò si traduce spesso nella definizione di una **funzione di costo** da minimizzare

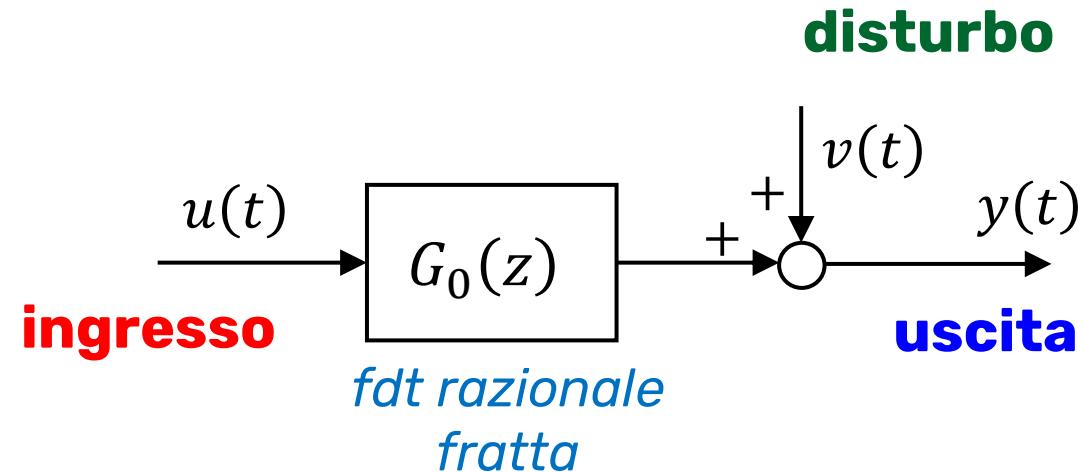
In tutti questi tre aspetti, una **conoscenza a priori** riguardo il sistema da identificare può essere di aiuto (es. la fisica mi dice che il modello deve avere un certo numero di poli\zeri)

Validazione del modello: in aggiunta a criteri di bontà «oggettivi», un modello potrebbe essere buono o meno a seconda dell'**applicazione** per il quale verrà usato. Per esempio, modelli «per il controllo» possono essere meno accurati di un «modello per la simulazione»



Identificazione dei sistemi dinamici

Ipotesi di lavoro 1: i dati sono generati da un **sistema LTI SISO con uscita rumorosa**

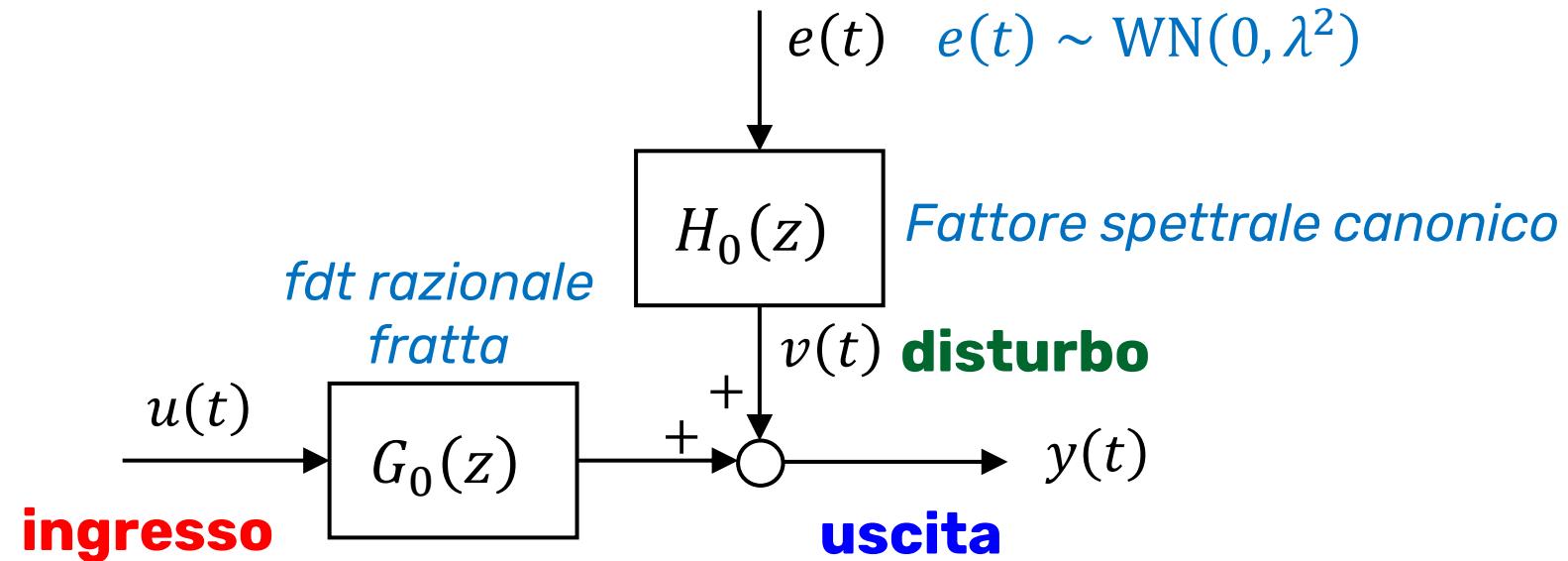


- I **parametri da stimare** sono i coefficienti del numeratore e del denominatore di $G_0(z)$
- Il **disturbo** $v(t)$ modella rumore di misura, ingressi non misurabili. Nel caso in cui il sistema vero non sia LTI, $v(t)$ modella anche gli scostamenti da questa assunzione



Identificazione dei sistemi dinamici

Ipotesi di lavoro 2: il disturbo $v(t)$ è modellizzabile come un **processo stocastico stazionario a spettro razionale**, indipendente da $u(t)$

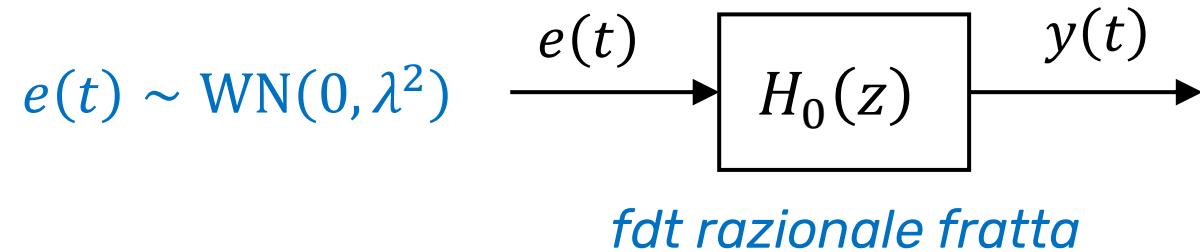


- In questo caso, vogliamo sia stimare i **coefficienti del numeratore e denominatore** di $G_0(z)$ sia quelli di $H_0(z)$ (e anche stimare λ^2)



Identificazione dei sistemi dinamici

Caso particolare: non c'è l'ingresso $u(t)$, ovvero sto trattando una **serie temporale**. In pratica, misuro solo l'uscita alimentata dal rumore bianco



Vogliamo sia stimare i **coeffienti del numeratore e denominatore** di $H_0(z)$ e λ^2 . Posso poi calcolare $\Gamma_{yy}(\omega) = |H_0(e^{j\omega})|^2 \cdot \lambda^2$

Questo approccio alla stima di $\Gamma_{yy}(\omega)$ prende il nome di **stima spettrale parametrica** (la **stima «nonparametrica»** è quella basata sul periodogramma)



Identificazione dei sistemi dinamici

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

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

- $H(z, \theta)$: **fattore spettrale canonico**, ovvero numeratore e denominatore monici, coprimi, di uguale grado, poli e zeri nel cerchio unitario
- $G(z, \theta)$: **funzione di trasferimento** che descrive l'effetto dell'ingresso $u(t)$, misurabile o noto, sull'uscita $y(t)$



Identificazione dei sistemi dinamici

Proprietà di $G(z, \theta)$

- Spesso si ipotizza che $G(z, \theta)$ sia **strettamente propria**, ovvero che il grado del numeratore è minore del grado del denominatore
 - ✓ Questo fa sì che vi sia un **ritardo puro** $k \neq 0$ tra ingresso e uscita
- $G(z, \theta)$ può avere **zeri fuori dal cerchio** o numeratore e denominatore **non monici**
- $G(z, \theta)$ rappresenta un **sistema fisico**, mentre $H(z, \theta)$ ed $e(t)$ **non esistono nella realtà** (sono solo costrutti matematici)



Outline

1. Introduzione all'identificazione dei modelli dinamici

2. Metodi a minimizzazione dell'errore di predizione (PEM)

3. Identificazione PEM di modelli ARX

4. Identificazione PEM di modelli ARMAX



L'approccio predittivo all'identificazione

Una volta definita la classe di modelli (per esempio ARMAX), potrei stimare i coefficienti θ usando la stima a **massima verosimiglianza** o la **stima Bayesiana**

In questi casi, dovrei però fare ipotesi sulla distribuzione dei dati (per esempio assumendo che $e(t)$ sia un **processo gaussiano**)

In alternativa, una via più **semplice** ed **intuitiva**, e che **non necessita di ipotesi ulteriori**, è quella di trovare un approccio basato sulla **minimizzazione di una somma di residui al quadrato** (come per il minimi quadrati)

L'approccio che seguiremo si basa su questa strada

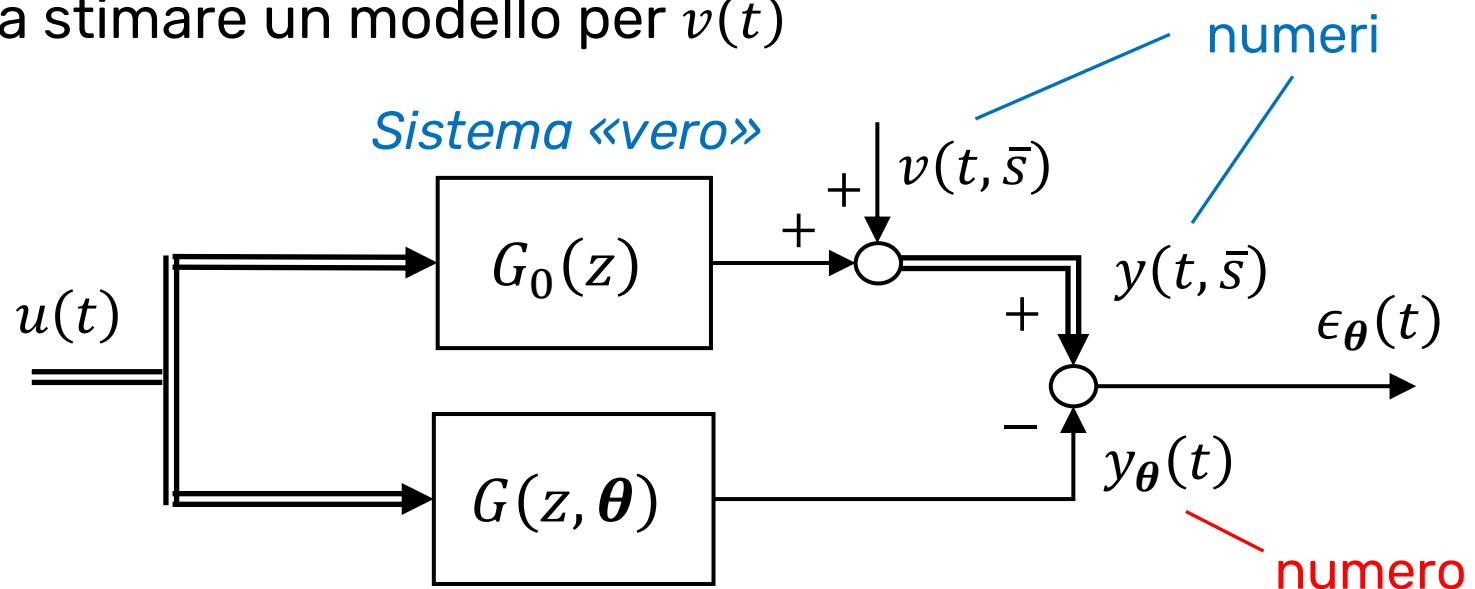


L'approccio predittivo all'identificazione

Caso «semplice»: non mi interessa stimare un modello per $v(t)$

Dati a disposizione

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



Il valore $y_\theta(t)$ è la **simulazione** del modello $G(z, \theta)$ a fronte dell'ingresso $u(t)$. La stima a minimi quadrati si trova minimizzando l'**errore di simulazione** $\epsilon_\theta(t)$

$$\hat{\theta}_{\text{LS}} = \arg \min_{\theta} \sum_{t=1}^N \epsilon_\theta(t)^2$$

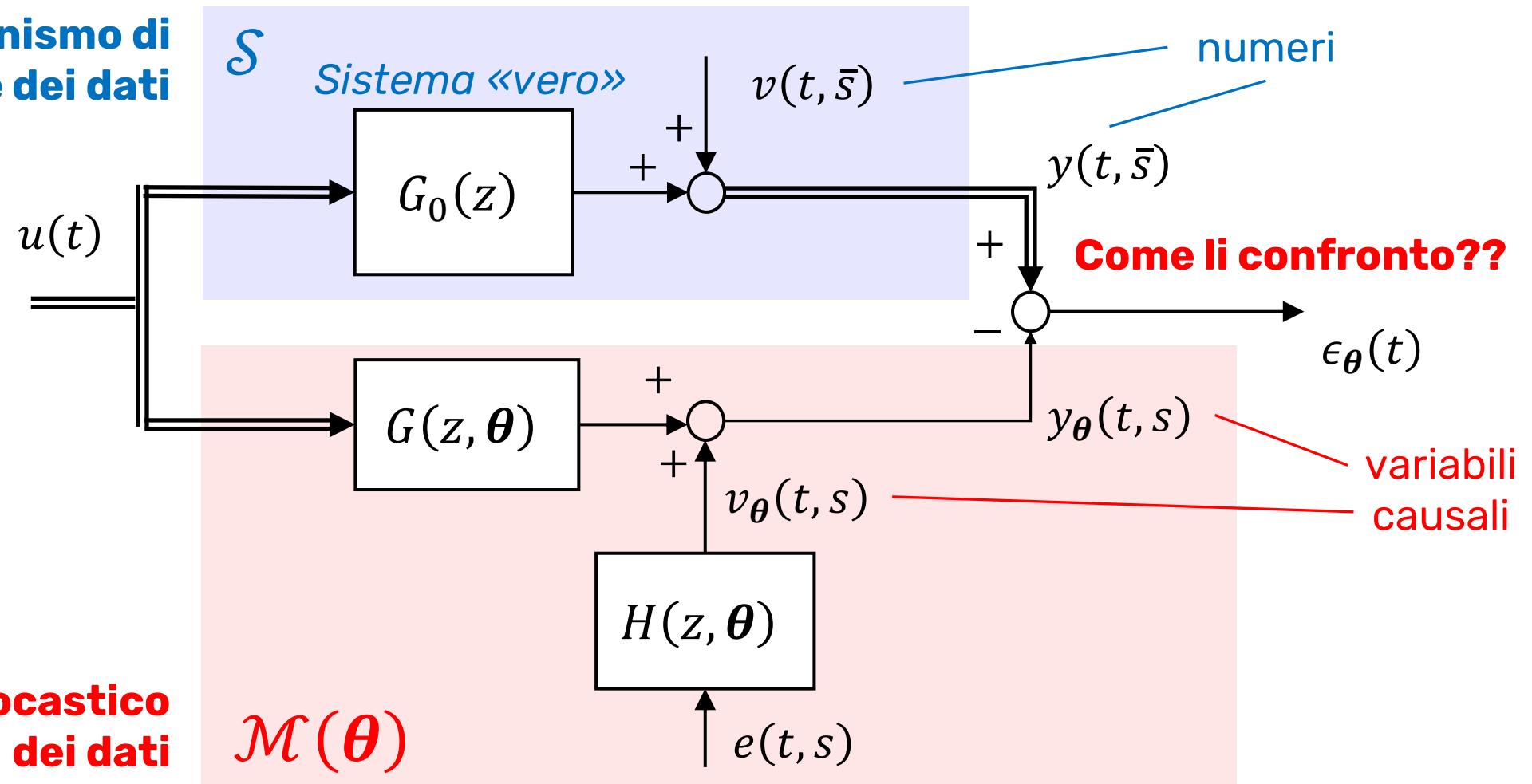
La soluzione è in forma chiusa solo se $y_\theta(t)$ è lineare dei parametri



L'approccio predittivo all'identificazione

Caso «più difficile»: oltre a stimare $G_0(z)$, voglio stimare anche un modello per $v(t)$

Meccanismo di generazione dei dati



L'approccio predittivo all'identificazione

Anche se $\nu(t, s)$ è un processo stocastico, nella pratica **una volta che i dati sono stati collezionati**, «è già avvenuta» una «scelta» dell'esito $s = \bar{s}$ che ha **generato quei dati osservati** $y(t, s = \bar{s})$. Quindi, la quantità $y(t, s = \bar{s})$ è un vettore di **numeri** perché frutto di una particolare realizzazione $\nu(t, s = \bar{s})$

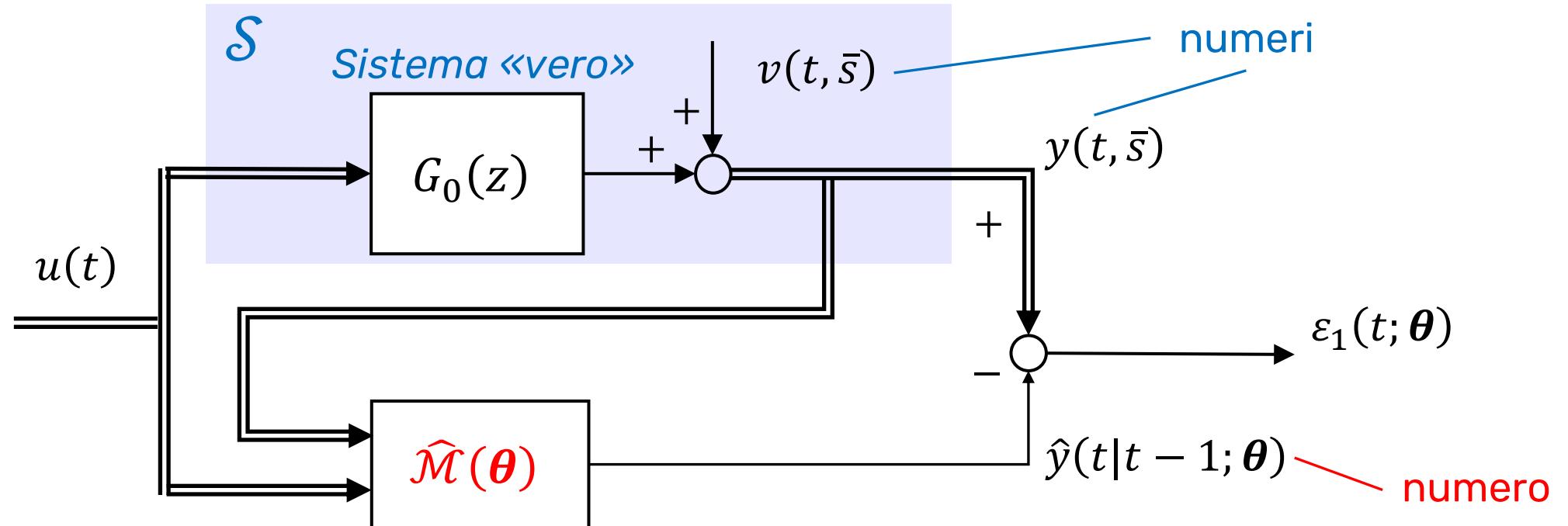
Il mio modello, invece, ha come uscita $y_\theta(t, s)$ che, se non fisso un esito, è un **processo stocastico**. Abbiamo quindi il dilemma che **non possiamo confrontare** un vettore di numeri con un processo stocastico

Se conoscessi il valore dell'esito $s = \bar{s}$, allora **potrei simulare** l'uscita del mio modello con quell'esito, e far si che $y_\theta(t, s = \bar{s})$ sia un vettore di numeri. Tuttavia, **non conosco l'esito** \bar{s}



L'approccio predittivo all'identificazione

Idea: considero come residuo $\epsilon_{\theta}(t)$ da minimizzare l'**errore di predizione a un passo** $\varepsilon_1(t; \theta)$



$\hat{\mathcal{M}}(\theta)$: **predittore ottimo ad un passo** associato al modello $\mathcal{M}(\theta)$



L'approccio predittivo all'identificazione

Definiamo quindi la stima ottenuta, considerando gli errori di predizione, come:

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

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

Θ è l'insieme dei valori ammissibili di $\boldsymbol{\theta}$

È possibile stimare la varianza λ^2 di $e(t)$ come:

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

Uno stimatore corretto richiederebbe di dividere per $N - d$, dove d è il numero di parametri



L'approccio predittivo all'identificazione

Osservazioni

- Θ : insieme dei **valori ammissibili** per θ . Per esempio, se $\mathcal{M}(\theta) = \text{ARMAX}$, voglio solo i θ tali per cui $C(z)/A(z)$ è canonico
- Per valori iniziali $t = 1, 2, \dots$ il predittore potrebbe non avere dati disponibili. Si usa quindi un' **inizializzazione «convenzionale»** (ipotizzo che i valori passati di $y(\cdot)$ siano nulli). L'inizializzazione non è un problema in quanto il **predittore è stabile**
- ✓ In alternativa, **scarto quei dati iniziali** che non hanno una predizione associata



L'approccio predittivo all'identificazione

- Abbiamo già in parte visto che l'errore di predizione a un passo $\varepsilon_1(t; \theta)$ gode di interessanti proprietà, che ci permettono di **distinguere θ^0** da un θ qualsiasi
 1. Dato θ e i dati $\{u(1), \dots, u(N)\}, \{y(1), \dots, y(N)\}$, è sempre possibile calcolare $\varepsilon_1(t; \theta)$
 2. Se $\exists \theta = \theta^0$ t.c. $G_0(z) = G(z, \theta^0)$ e $H_0(z) = H(z, \theta^0)$, abbiamo che $\varepsilon_1(t; \theta^0) = e(t)$, ovvero ci permette di capire se il modello è buono
 3. $\varepsilon_1(t; \theta) \neq e(t)$ per qualsiasi $\theta \neq \theta^0$ (supponendo di avere un *ingresso adeguato*)
 4. θ^0 minimizza la varianza dell'errore di predizione a un passo



L'approccio predittivo all'identificazione

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

Nota

Se ipotizzo che $\mathcal{S} = \mathcal{M}(\theta^0)$ e $e(t) \sim \text{WN Gaussiano}$, lo **stimatore PEM è circa uguale allo stimatore a massima verosimiglianza**

La differenza sta in come i due approcci trattano l'inizializzazione del predittore: la funzione di verosimiglianza «propriamente detta» sarebbe più difficile da trattare rispetto a quella «**condizionata**» ai valori iniziali, per la quale c'è l'equivalenza coi metodi PEM

Se i dati sono molti, non c'è differenza



Outline

1. Introduzione all'identificazione dei modelli dinamici
2. Metodi a minimizzazione dell'errore di predizione (PEM)
- 3. Identificazione PEM di modelli ARX**
4. Identificazione PEM di modelli ARMAX



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

Identificazione PEM di modelli ARX

Consideriamo un **modello** ARX($n_a, n_b, 1$) , e di avere a disposizione N dati $\{u(1), \dots, u(N)\}, \{y(1), \dots, y(N)\}$

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

- $B(z) = b_0 + b_1 z^{-1} + \dots + b_{n_b} z^{-n_b}$
- $A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - \dots - a_{n_a} z^{-n_a}$

Osservazioni

- $C(z) = 1$ poiché non c'è la parte MA
- Abbiamo supposto che $k = 1$. Fissando un ritardo unitario **non lediamo di generalità**.

Per esempio, se il ritardo vero fosse $k = 2$, stimeremmo $b_0 = 0$



Identificazione PEM di modelli ARX

Il modello in **forma di preditore** è dato da

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

$$= B(z, \boldsymbol{\theta})u(t-1) + [1 - A(z, \boldsymbol{\theta})]y(t)$$

$$= (b_0 + b_1 z^{-1} + \cdots + b_{n_b} z^{-n_b})u(t-1) + (a_1 z^{-1} + \cdots + a_{n_a} z^{-n_a})y(t)$$



$$\hat{y}(t|t-1; \boldsymbol{\theta}) = b_o u(t-1) + b_1 u(t-2) + \cdots + b_{n_b} u(t-n_b-1) + a_1 y(t-1) + \cdots + a_{n_a} y(t-n_a)$$



Identificazione PEM di modelli ARX

Definendo i vettori

$$\boldsymbol{\theta} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{n_a} \\ b_0 \\ b_1 \\ \vdots \\ b_{n_b} \end{bmatrix} \quad \begin{matrix} (n_a + n_b + 1) \times 1 \\ d \times 1 \end{matrix}$$
$$\boldsymbol{\varphi}(t) = \begin{bmatrix} y(t-1) \\ y(t-2) \\ \vdots \\ y(t-n_a) \\ u(t-1) \\ u(t-2) \\ \vdots \\ u(t-n_b-1) \end{bmatrix} \quad \begin{matrix} (n_a + n_b + 1) \times 1 \\ d \times 1 \end{matrix}$$

Possiamo scrivere

$$\mathcal{M}(\boldsymbol{\theta}): y(t) = \boldsymbol{\varphi}^\top(t) \boldsymbol{\theta} + e(t)$$

$$\hat{\mathcal{M}}(\boldsymbol{\theta}): \hat{y}(t|t-1; \boldsymbol{\theta}) = \boldsymbol{\varphi}^\top(t) \boldsymbol{\theta} \quad \Rightarrow \quad \text{Il predittore è \textbf{lineare nei parametri } } \boldsymbol{\theta}!$$



Identificazione PEM di modelli ARX

Per trovare la stima PEM minimizziamo la funzione di costo

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

La soluzione è analoga alla **stima a minimi quadrati** di un modello lineare statico!

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

Se $S(N) = \sum_{t=1}^N \boldsymbol{\varphi}(t) \boldsymbol{\varphi}^\top(t)$ è invertibile, allora la soluzione $\hat{\boldsymbol{\theta}}_N$ è **unica** in quanto la funzione di costo è convessa



Identificazione PEM di modelli ARX

Come per la regressione lineare, posso esprimere il modello ARX in forma matriciale

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

$$\theta = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{n_a} \\ b_0 \\ b_1 \\ \vdots \\ b_{n_b} \end{bmatrix}_{(n_a + n_b + 1) \times 1 \quad d \times 1}$$

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

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

Da cui segue che

$$Y = \Phi \theta + E$$

$$N \times 1 \quad N \times d \quad d \times 1 \quad N \times 1$$



$$\hat{\theta}_N = (\Phi^\top \Phi)^{-1} \Phi^\top Y$$

$$d \times 1 \quad d \times d \quad d \times N \quad N \times 1$$



Esempio: stima di un modello AR(1)

Si supponga di avere 5 dati da una serie temporale $y(t)$, stazionaria e ergodica, a media nulla

$$y(1) = \frac{1}{2} \quad y(2) = 0 \quad y(3) = -1 \quad y(4) = -\frac{1}{2} \quad y(5) = \frac{1}{4}$$

Identificare un modello AR(1) del tipo $y(t) = ay(t - 1) + e(t)$, $e(t) \sim WN(0, \lambda^2)$

Usando il modello identificato, si calcoli la **predizione** $\hat{y}(6|5)$ e la **varianza** del rumore $\hat{\lambda}^2$

Nota

La media campionaria $\hat{m} = \frac{1}{5} \sum_{t=1}^5 y(i) = -0,15$ non è nulla. Il problema però ci dice di considerare una media nulla. In caso contrario, avremmo dovuto depolarizzare il processo al fine di avere un predittore corretto, tale che $\mathbb{E}[\varepsilon_1(t)] = 0$



Esempio: stima di un modello AR(1)

Calcoliamo il predittore ad un passo

Supponiamo che il modello sia in forma canonica (ovvero che $|a| < 1$)

$$y(t) = \frac{1}{1 - az^{-1}} e(t) \quad \Rightarrow \quad \hat{y}(t|t-1; a) = \frac{C(z) - A(z)}{C(z)} y(t) = \frac{1 - 1 + az^{-1}}{1} y(t) = ay(t-1)$$

Calcoliamo la funzione di costo

Abbiamo due alternative: o inizializzo i valori mancanti del predittore (per esempio $\hat{y}(1|0)$) a zero, oppure parto da $t = 2$ fino a $t = 5$. Scegliamo questa seconda strada (per tanti dati il risultato non cambia)



Esempio: stima di un modello AR(1)

$$J_5(\theta) = \frac{1}{5-1} \sum_{t=2}^5 (y(t) - ay(t-1))^2$$

$$\begin{array}{lll} y(1) = \frac{1}{2} & y(2) = 0 & y(5) = \frac{1}{4} \\ y(3) = -1 & y(4) = -\frac{1}{2} & \end{array}$$

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

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

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



Esempio: stima di un modello AR(1)

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

Minimizziamo la funzione di costo

$$\frac{dJ_5(a)}{da} = 0 \quad \Rightarrow \quad 3a - \frac{3}{4} = 0 \quad \Rightarrow \quad \hat{a}_5 = \frac{1}{4}$$

Se avessimo ottenuto $|\hat{a}_5| > 1$, avremmo potuto usare un filtro passa-tutto per rendere il modello stabile



Esempio: stima di un modello AR(1)

Stimiamo la varianza del rumore

$$\hat{\lambda}^2 = J_5(\hat{a}_5) = \frac{1}{4} \left[\frac{21}{16} - \frac{3}{4} \cdot \frac{1}{4} + \frac{3}{2} \cdot \left(\frac{1}{4} \right)^2 \right] \approx \boxed{0.3}$$

Predizione a un passo $\hat{y}(6|5)$

Notiamo che λ^2 non è molto importante: infatti non mi serve per calcolare la predizione

$$\hat{y}(t|t-1; \hat{a}_5) = \hat{a}_5 y(t-1) \quad \Rightarrow \quad \hat{y}(6|5) = \frac{1}{4} y(5) = \boxed{\frac{1}{16}}$$



Esempio: stima di un modello AR(1)

Modello identificato

$$y(t) = \frac{1}{1 - \frac{1}{4}z^{-1}} e(t), \quad e(t) \sim WN(0, 0.3)$$



Outline

1. Introduzione all'identificazione dei modelli dinamici
2. Metodi a minimizzazione dell'errore di predizione (PEM)
3. Identificazione PEM di modelli ARX
- 4. Identificazione PEM di modelli ARMAX**



Identificazione PEM di modelli ARMAX

Consideriamo un **modello** ARMAX($n_a, n_c, n_b, 1$) , e di avere a disposizione N dati $\{u(1), \dots, u(N)\}, \{y(1), \dots, y(N)\}$

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

- $B(z) = b_0 + b_1 z^{-1} + \dots + b_{n_b} z^{-n_b}$
- $A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - \dots - a_{n_a} z^{-n_a}$
- $C(z) = 1 + c_1 z^{-1} + c_2 z^{-2} + \dots + c_{n_c} z^{-n_c}$

Il vettore dei parametri, in questo caso, è:

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

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



Identificazione PEM di modelli ARMAX

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 $y(t)$,

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

Potevamo ottenere la stessa cosa anche con l'espressione generica di $\varepsilon_1(t)$:

$$\varepsilon_1(t; \boldsymbol{\theta}) = H^{-1}(z, \boldsymbol{\theta})[y(t) - G(z, \boldsymbol{\theta})u(t)]$$

$$= \frac{A(z)}{C(z)} \left[y(t) - \frac{B(z)}{A(z)} u(t - 1) \right] = \frac{A(z)}{C(z)} y(t) - \frac{B(z)}{C(z)} u(t - 1)$$



Identificazione PEM di modelli ARMAX

Utilizziamo l'approccio predittivo

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

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

Osservazioni

- Dato che ho $C(z, \boldsymbol{\theta})$ al denominatore, questa funzione di costo **non è più convessa!** In generale, avrò dei **minimi locali**
- Per la risoluzione del problema di minimizzazione, devo utilizzare dei **metodi iterativi** (per esempio il **metodo del gradiente**)



Identificazione PEM di modelli ARMAX

Come gestire i minimi locali

Una strategia semplice è la seguente. Data una inizializzazione $\hat{\theta}^{(0)}$ all'iterazione 0:

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

Come alternativa (più efficiente) al metodo del gradiente, vedremo il metodo di **ottimizzazione iterativo** noto come **Metodo di Newton**. Questo metodo, oltre al gradiente, sfrutta anche l'informazione data dalla **matrice Hessiana**



Identificazione PEM di modelli ARMAX

METODO DI NEWTON

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

$J_N(\theta) \approx V(\theta) \longrightarrow$ La funzione $V(\theta)$ è un **paraboloide** (è facile calcolarne il minimo)

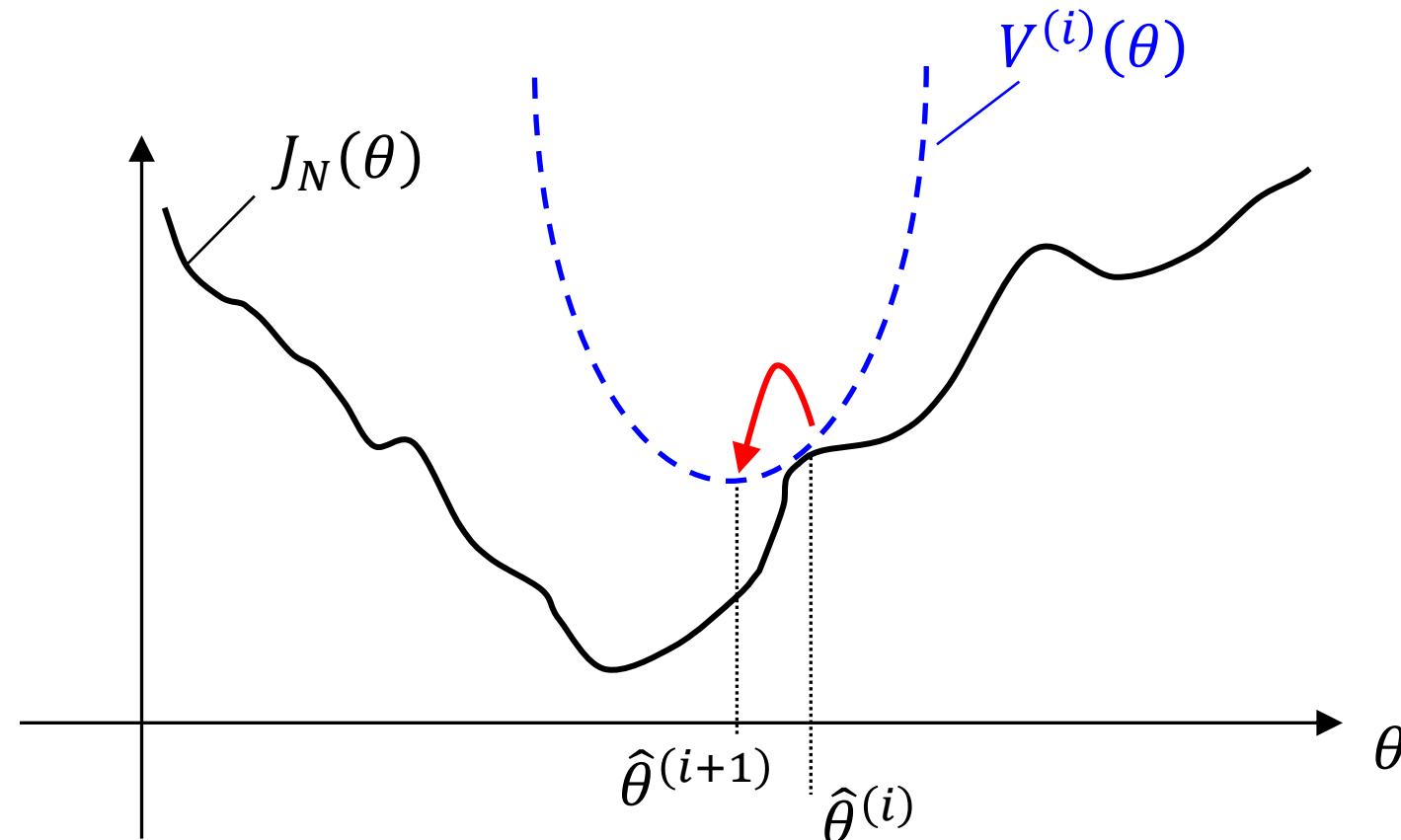
$$V^{(i)}(\theta) = J_N(\hat{\theta}^{(i)}) + (\theta - \hat{\theta}^{(i)})^\top \cdot \underbrace{\frac{dJ_N(\theta)}{d\theta} \Big|_{\theta=\hat{\theta}^{(i)}}}_{\text{Gradiente}} + \frac{1}{2} (\theta - \hat{\theta}^{(i)})^\top \cdot \underbrace{\frac{d^2J_N(\theta)}{d\theta^2} \Big|_{\theta=\hat{\theta}^{(i)}}}_{\text{Matrice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



Identificazione PEM di modelli ARMAX

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

Consideriamo un caso scalare per semplicità



Identificazione PEM di modelli ARMAX

$$\nabla_x(x^\top b) = b$$

$$\nabla_x(x^\top Ax) = (A + A^\top)x$$

Troviamo un'espressione esplicita per $\hat{\theta}^{(i+1)}$ imponendo $\frac{dV^{(i)}(\theta)}{d\theta} = \mathbf{0}_{d \times 1}$

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

Regola di update per il metodo di Newton



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

È simile al gradient descent se al posto dell'Hessiana metto la learning rate α



Identificazione PEM di modelli ARMAX

Dobbiamo quindi calcolare queste due quantità:

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

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

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

$$\begin{aligned} \frac{dJ_N(\boldsymbol{\theta})}{d\boldsymbol{\theta}} &= \frac{d}{d\boldsymbol{\theta}} \left[\frac{1}{N} \sum_{t=1}^N \varepsilon_1(t; \boldsymbol{\theta})^2 \right] = \frac{1}{N} \sum_{t=1}^N \frac{d}{d\boldsymbol{\theta}} \varepsilon_1(t; \boldsymbol{\theta})^2 \\ &\quad d \times 1 \end{aligned} \quad = \quad \boxed{\frac{2}{N} \sum_{t=1}^N \varepsilon_1(t; \boldsymbol{\theta}) \cdot \frac{d\varepsilon_1(t; \boldsymbol{\theta})}{d\boldsymbol{\theta}}} \quad d \times 1$$



Identificazione PEM di modelli ARMAX

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

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

Derivata del prodotto

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

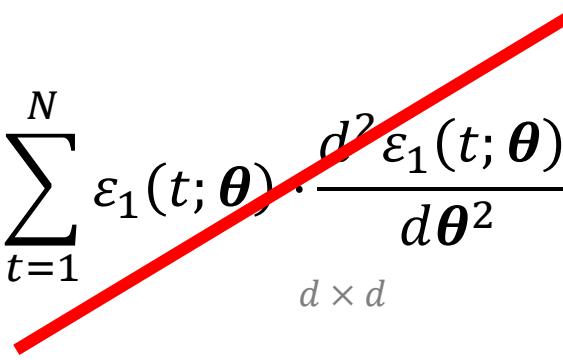
$$\frac{d(vu)}{dx} = v \cdot \frac{du}{dx} + \frac{dv}{dx} u^\top$$

Nel nostro caso:

$$v = \varepsilon_1(t; \theta)$$
$$u = \frac{d\varepsilon_1(t; \theta)}{d\theta}$$
$$x = \theta$$



Identificazione PEM di modelli ARMAX

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


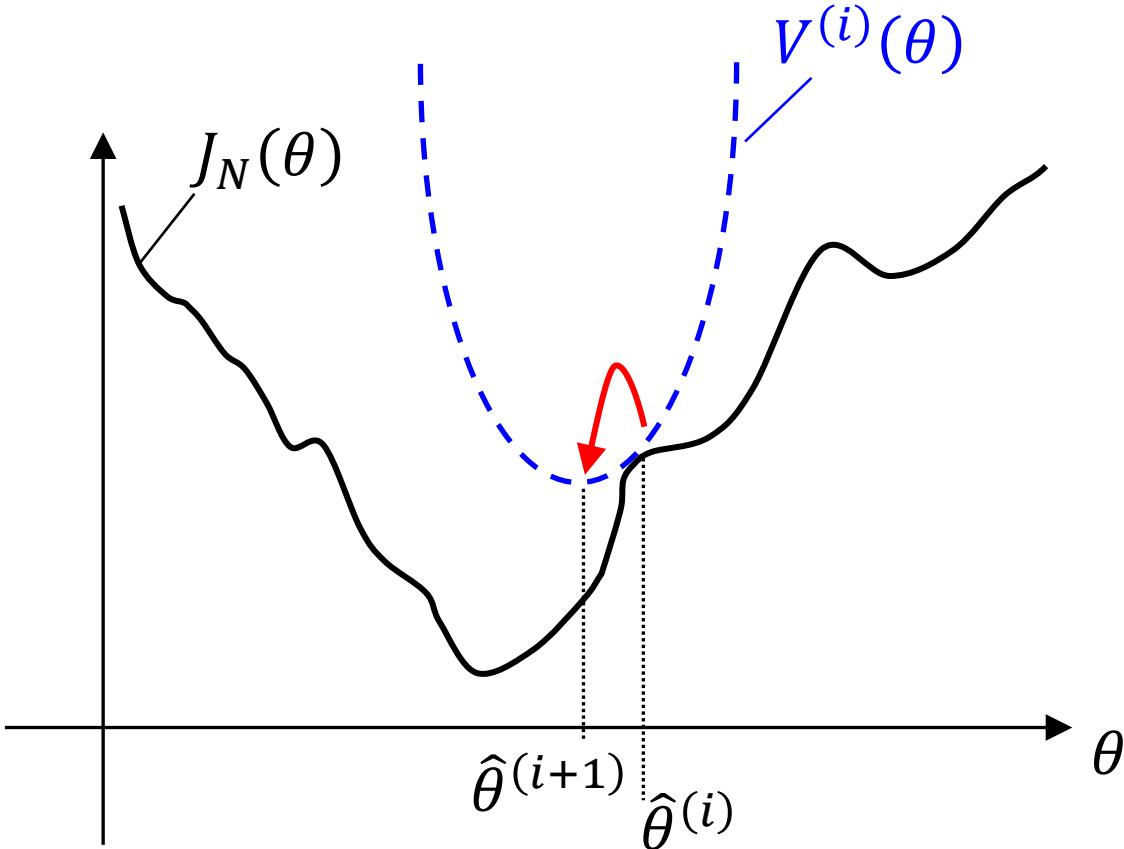
Ignoriamo il secondo termine, **approssimando l'Hessiana**, dato che:

- Se siamo **vicini all'ottimo**, $\varepsilon_1(t; \boldsymbol{\theta})$ è «**piccolo**» e il termine «conta poco»
- Possiamo **evitare di calcolare** $\frac{d^2\varepsilon_1(t; \boldsymbol{\theta})}{d\boldsymbol{\theta}^2}$
- Ci assicuriamo una Hessiana **semi-definita positiva**. In questo modo, la direzione dell'algoritmo è **sicuramente discendente** (concetto simile ad avere learning rate ≥ 0)

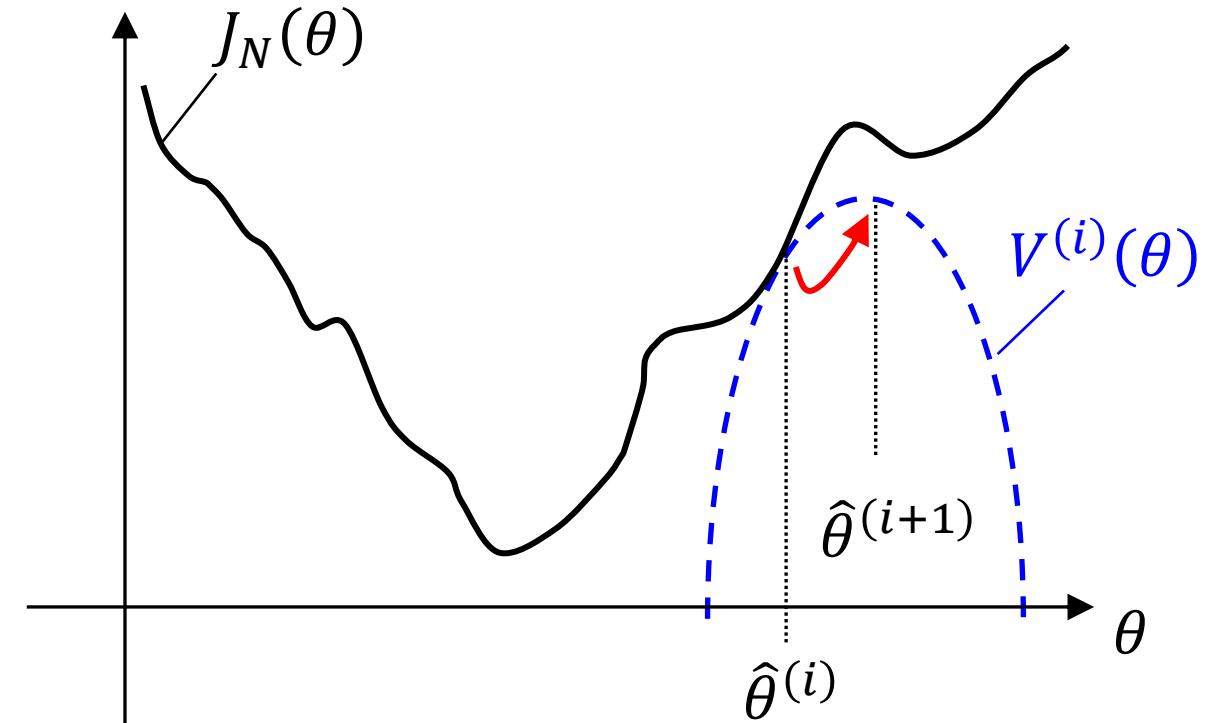


Identificazione PEM di modelli ARMAX

Hessiana > 0



Hessiana < 0



Identificazione PEM di modelli ARMAX

Osservazione

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

1. Metodo del **gradiente**
2. Metodo di **Newton**
3. Metodi di «**quasi-Newton**»

Metodo del gradiente

$$\hat{\theta}^{(i+1)} = \hat{\theta}^{(i)} - \alpha \cdot \left[\frac{dJ_N(t; \theta)}{d\theta} \Big|_{\theta=\hat{\theta}^{(i)}} \right]$$

- **Semplice e robusto** (va sempre nella direzione di discesa)
- Può essere **molto lento** quando ci avviciniamo al minimo (poiché il gradiente tende a 0)



Identificazione PEM di modelli ARMAX

Metodo di Newton

L'Hessiana «modula» il passo del gradiente

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

- Converge **velocemente**
- Computazionalmente più **complesso**
- Rischio di **instabilità** se l'Hessiana è definita negativa

Metodi «quasi Newtoniani»

O^{-1} è un'approssimazione dell'Hessiana semidefinita positiva o definita positiva

$$\hat{\theta}^{(i+1)} = \hat{\theta}^{(i)} - O^{-1} \cdot \left[\frac{dJ_N(t; \theta)}{d\theta} \Big|_{\theta=\hat{\theta}^{(i)}} \right]$$

- **Più semplice** del metodo di Newton
- **Più veloce** del metodo del gradiente
- **Non c'è rischio** di allontanarsi dal minimo
- **Non è veloce** come il metodo di Newton



Identificazione PEM di modelli ARMAX

I metodi «quasi Newtoniani» sono molto usati e differiscono fra loro nel **modo in cui viene fatta l'approssimazione** definita positiva dell'Hessiana

Per **garantire l'invertibilità** di $\theta^{-1} \geq 0$, si aggiunge un termine positivo, molto piccolo, di **«regolarizzazione»**

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



Identificazione PEM di modelli ARMAX

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

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

- Notiamo che i termini $2/N$ si possono semplificare



Identificazione PEM di modelli ARMAX

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

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

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

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

$d \times 1$



Identificazione PEM di modelli ARMAX

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

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

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

⋮

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

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



Identificazione PEM di modelli ARMAX

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

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

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

⋮

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

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



Identificazione PEM di modelli ARMAX

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

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



$$(1 + c_1z^{-1} + c_2z^{-2} + \dots + c_{n_c}z^{-n_c})\varepsilon_1(t) = A(z)y(t) - B(z)u(t-1)$$

$$\frac{d[(1 + c_1z^{-1} + c_2z^{-2} + \dots + c_{n_c}z^{-n_c}) \cdot \varepsilon_1(t)]}{dc_1} = \frac{d[A(z)y(t) - B(z)u(t-1)]}{dc_1}$$

Non dipende da c_1

$$\frac{d[(1 + c_1z^{-1} + c_2z^{-2} + \dots + c_{n_c}z^{-n_c}) \cdot \varepsilon_1(t)]}{dc_1} = 0$$

Devo fare la derivata del prodotto



Identificazione PEM di modelli ARMAX

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

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

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

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

:

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



Identificazione PEM di modelli ARMAX

Riassumendo, il **vettore gradiente** è:

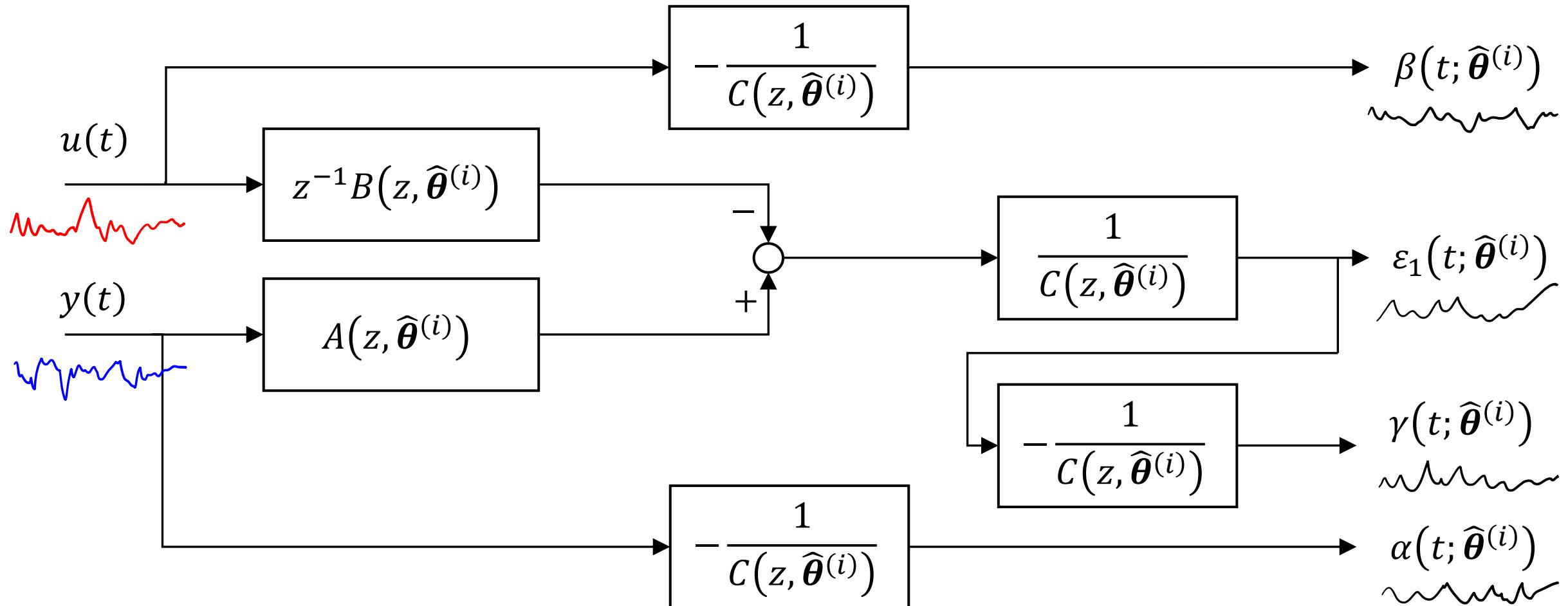
$$\frac{d\varepsilon_1(t)}{d\theta} = \begin{bmatrix} \alpha(t-1) \\ \vdots \\ \alpha(t-n_a) \\ \beta(t-1) \\ \vdots \\ \beta(t-n_b-1) \\ \gamma(t-1) \\ \gamma(t-n_c) \end{bmatrix} \quad t = 1, \dots, N$$

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



Identificazione PEM di modelli ARMAX

Abbiamo il seguente **schema di filtraggio dei segnali** per trovare il gradiente



Identificazione PEM di modelli ARMAX

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

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

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



Identificazione PEM di modelli ARMAX

Osservazione

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

Per le famiglie di modelli viste, vale che

- **ARX, FIR:** funzione di costo **convessa** (minimo globale)
- **ARMAX, BJ, OE:** funzione di costo **non convessa** (minimi locali)





**UNIVERSITÀ
DEGLI STUDI
DI BERGAMO**

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



IDENTIFICAZIONE DEI MODELLI E ANALISI DEI DATI (IMAD)

Lezione 12: Identificazione – analisi e complementi

**Corso di Laurea Magistrale in
INGEGNERIA INFORMATICA**

SPEAKER
Prof. Mirko Mazzoleni

PLACE
Università degli Studi di
Bergamo

Syllabus

Parte II: sistemi dinamici

8. Processi stocastici

- 8.1 Processi stocastici stazionari (pss)
- 8.3 Rappresentazione spettrale di un pss
- 8.4 Stimatori campionari media\covarianza
- 8.5 Densità spettrale campionaria

9. Famiglie di modelli a spettro razionale

- 9.1 Modelli per serie temporali (MA, AR, ARMA)
- 9.2 Modelli per sistemi input/output (ARX, ARMAX)

10. Predizione

- 10.1 Filtro passa-tutto

10.2 Forma canonica

10.3 Teorema della fattorizzazione spettrale

10.4 Soluzione al problema della predizione

11. Identificazione

- 11.3 Identificazione di modelli ARX
- 11.4 Identificazione di modelli ARMAX
- 11.5 Metodo di Newton

12. Identificazione: analisi e complementi

- 12.1 Analisi asintotica metodi PEM
- 12.2 Identificabilità dei modelli
- 12.3 Valutazione dell'incertezza di stima

13. Identificazione: valutazione



Parte I: sistemi staticiStima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

✓ Stima parametri popolazione

✓ Stima modello lineare: minimi quadrati

○ **SI assunzioni su ddp dei dati**

✓ Stima massima verosimiglianza parametri popolazione

✓ Stima modello lineare: massiva verosimiglianza

✓ Regressione logistica

• θ variabile casuale○ **SI assunzioni su ddp dei dati**

✓ Stima Bayesiana

Machine learning**Parte II: sistemi dinamici**Stima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

✓ Modelli lineari di pss

✓ Predizione

✓ Identificazione

✓ Persistente eccitazione

✓ Analisi asintotica metodi PEM

✓ Analisi incertezza stima (numero dati finito)

✓ Valutazione del modello

Outline

1. Analisi asintotica dei metodi PEM
2. Identificabilità dei modelli e persistente eccitazione
3. Valutazione dell'incertezza della stima PEM
4. Robustezza dei metodi PEM e prefiltraggio
5. Empirical Transfer Function Estimate (ETFE)



Outline

- 1. Analisi asintotica dei metodi PEM**
- Identificabilità dei modelli e persistente eccitazione
- Valutazione dell'incertezza della stima PEM
- Robustezza dei metodi PEM e prefiltraggio
- Empirical Transfer Function Estimate (ETFE)



Analisi asintotica dei metodi PEM

Nella lezione precedente, abbiamo visto come ottenere una stima $\hat{\theta}_N$ data una **singola sequenza** di N dati $\{u(1), \dots, u(N)\}, \{y(1, \bar{s}), \dots, y(N, \bar{s})\}$, utilizzando **l'approccio PEM**.

Abbiamo esplicitato l'esito $s = \bar{s}$ per indicare che lavoriamo con delle **sequenze di numeri**

L'idea dell'approccio predittivo è la seguente: data una **famiglia di modelli** $\{\mathcal{M}(\theta) | \theta \in \Theta\}$, la stima $\hat{\theta}_N$ è ottenuta **minimizzando la funzione di costo**

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

dove $\varepsilon_1(t; \boldsymbol{\theta}) = y(t) - \hat{y}(t|t-1; \boldsymbol{\theta})$ è l'**errore di predizione ad un passo**



Analisi asintotica dei metodi PEM

Problema: la stima $\hat{\theta}_N$, calcolata in questo modo, **ci fornisce un «buon» modello?**

Per poter rispondere a questa domanda, non possiamo limitarci ad una **specifica** realizzazione dei dati corrispondente ad un particolare esito \bar{s} : dobbiamo studiare quello che succede «in generale», ovvero considerando **tutte le possibili realizzazioni di sequenze di dati**

IPOTESI DI LAVORO

- Sia **l'ingresso** $u(t, s)$ sia **l'uscita** $y(t, s)$ sono processi stocastici **stazionari ed ergodici** (il che implica che tutte le funzioni di trasferimento sono **asintoticamente stabili**)



Analisi asintotica dei metodi PEM

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}

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

La **funzione di costo** dipende anch'essa dall'esito \bar{s} poiché utilizza i dati misurati

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

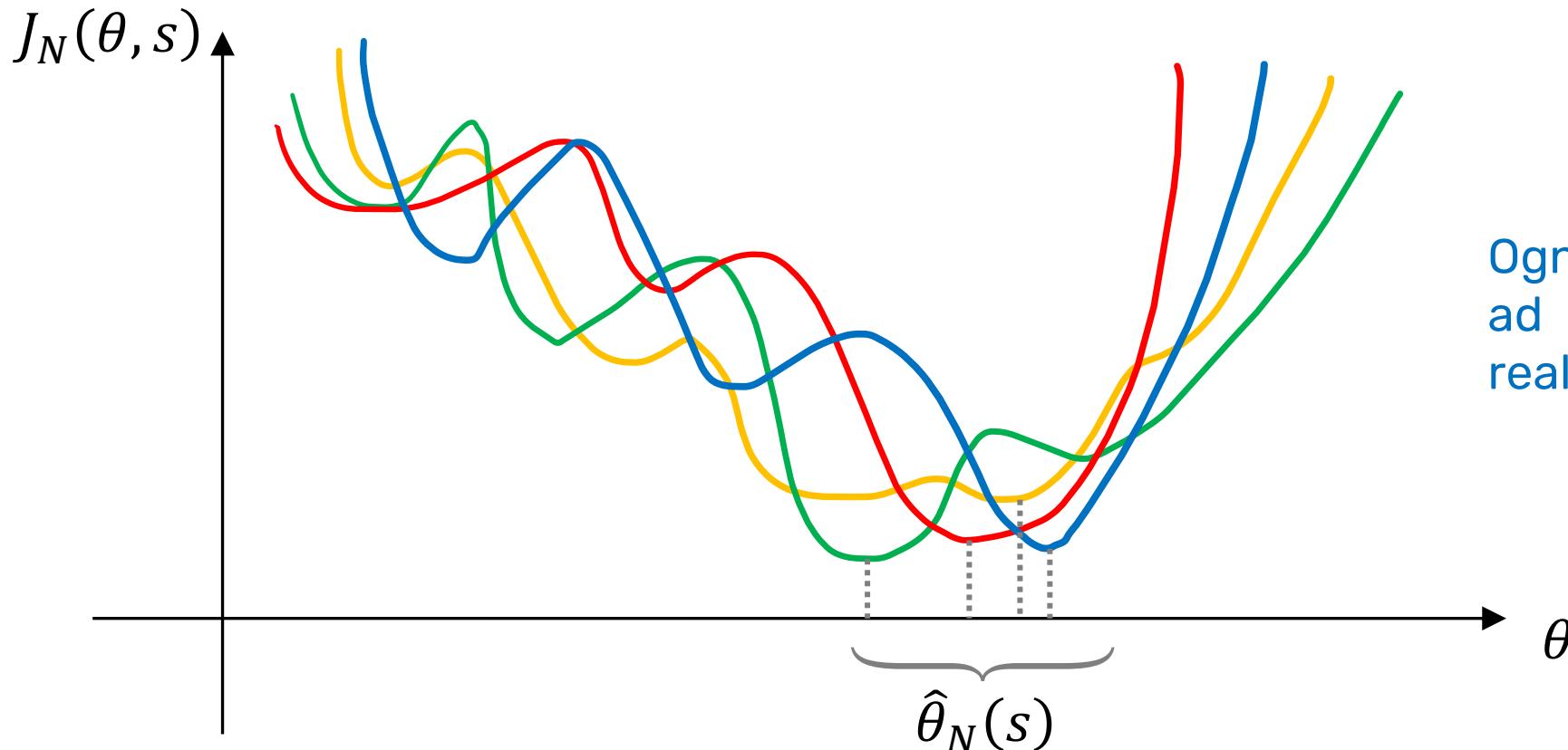
da cui otterrò la stima $\hat{\boldsymbol{\theta}}_N(\bar{s})$

Ne consegue che, in generale, la stima $\hat{\boldsymbol{\theta}}_N(s)$ è una **variabile casuale** perché il suo valore dipende dai dati, i quali dipendono dall'esito s

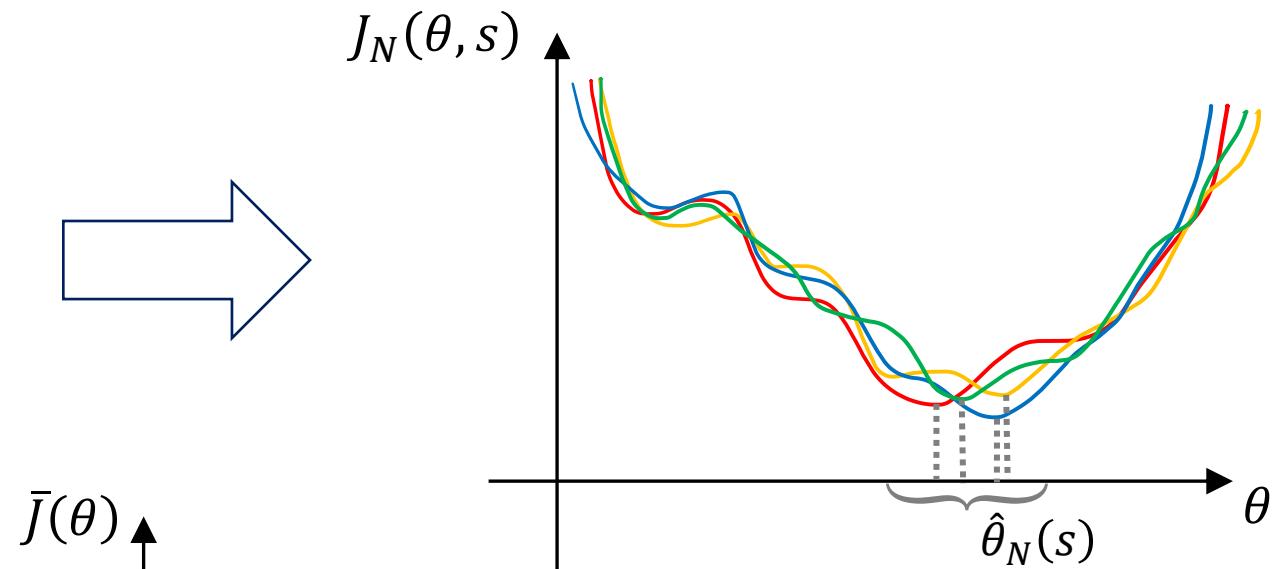
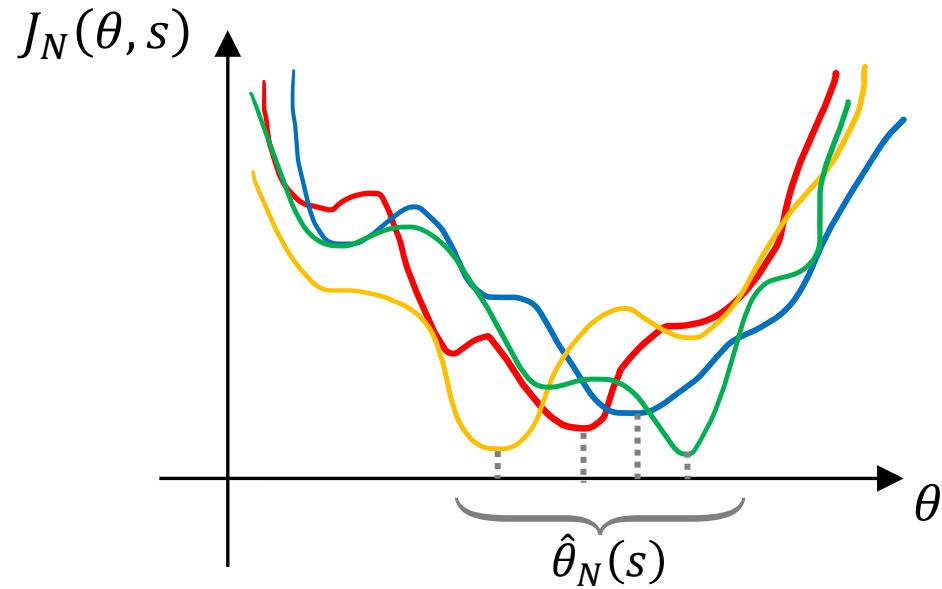


Analisi asintotica dei metodi PEM

La funzione di costo $J_N(\theta, s)$ dovrebbe essere interpretata come un **insieme di curve**, e la stima $\hat{\theta}_N(s)$ come un **insieme di punti** → difficile da descrivere per N finito!



Analisi asintotica dei metodi PEM



$N \rightarrow +\infty$



Analisi asintotica dei metodi PEM

Grazie all'**ipotesi di ergodicità** di $u(t,s)$ e $y(t,s)$, abbiamo che i momenti temporali convergono ai rispettivi momenti di insieme. Di conseguenza:

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

cioè, le curve $J_N(\boldsymbol{\theta}, s)$ convergono ad un'**unica (deterministica) curva** $\bar{J}(\boldsymbol{\theta})$

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

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



Analisi asintotica dei metodi PEM

Caso particolare: $\Delta_\theta = \bar{\theta}$, ovvero $\bar{J}(\theta)$ ha un **unico minimo globale**

Teorema

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

$$J_N(\theta, s) \xrightarrow[N \rightarrow +\infty]{} \bar{J}(\theta) \quad \hat{\theta}_N \xrightarrow[N \rightarrow +\infty]{} \Delta_\theta$$

Ne segue che se $\Delta_\theta = \bar{\theta}$, allora $\hat{\theta}_N \xrightarrow[N \rightarrow +\infty]{} \bar{\theta}$



Analisi asintotica dei metodi PEM

Osservazioni

- Il teorema dice che **il risultato dell'identificazione PEM è lo stesso, indipendentemente dalle realizzazioni misurate** dei processi $u(t), y(t)$ purché il numero di dati N sia **abbastanza grande**

Analizzare la singola curva $\bar{J}(\theta)$ è molto più facile che analizzare un insieme di curve!

Idea: per studiare le proprietà della stima, studiamo le sue **caratteristiche asintotiche**, ovvero studiamo il **modello stimato asintotico** $\mathcal{M}(\bar{\theta})$ oppure **l'insieme di modelli stimati asintotici** $\{\mathcal{M}(\theta) | \theta \in \Delta_\theta\}$

Se le proprietà di $\mathcal{M}(\bar{\theta})$ sono buone, posso pensare che lo siano anche quelle di $\mathcal{M}(\hat{\theta}_N)$, fintanto che N è grande



Analisi asintotica dei metodi PEM

IPOTESI DI LAVORO AGGIUNTIVA

Assumiamo che $\mathcal{S} \in \mathcal{M}(\theta)$, ovvero che esista $\theta^0 \in \Theta$ tale che $\mathcal{S} = \mathcal{M}(\theta^0)$

Domanda: il vettore «vero» dei parametri θ^0 appartiene all'insieme Δ_θ dei minimi globali della cifra di costo $\bar{J}(\theta)$? Ciò è equivalente a chiedersi se $\hat{\theta}_N$ tende asintoticamente a θ^0

Se ciò fosse vero, vorrebbe dire che **i metodi PEM sono in grado di trovare la parametrizzazione «vera» del modello**

Dimostriamo che, sotto le ipotesi fatte, θ^0 **appartiene sempre** a Δ_θ



Analisi asintotica dei metodi PEM

Dimostrazione

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

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

Consideriamo un **generico modello** $\mathcal{M}(\boldsymbol{\theta})$, per il quale

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

Non è detto che $\varepsilon_1(t; \boldsymbol{\theta})$
sia bianco...

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

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



Analisi asintotica dei metodi PEM

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

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

Errore di predizione «ottimo»

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

Pertanto

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

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

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



Analisi asintotica dei metodi PEM

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

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

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

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



Analisi asintotica dei metodi PEM

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

$$\rightarrow \bar{J}(\boldsymbol{\theta}) = \lambda^2 + \mathbb{E}\left[\left(\hat{y}(t|t-1; \boldsymbol{\theta}^0) - \hat{y}(t|t-1; \boldsymbol{\theta})\right)^2\right]$$

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

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

$$\boxed{\bar{J}(\boldsymbol{\theta}) \geq \bar{J}(\boldsymbol{\theta}^0), \quad \forall \boldsymbol{\theta}}$$

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



Analisi asintotica dei metodi PEM

Conclusione (fondamentale)

Se $\mathcal{S} \in \mathcal{M}(\theta)$ e $u(t), y(t)$ sono pss ergodici, allora, per $N \rightarrow +\infty$, **un metodo PEM garantisce che il modello stimato è quello «vero»**

Osservazioni

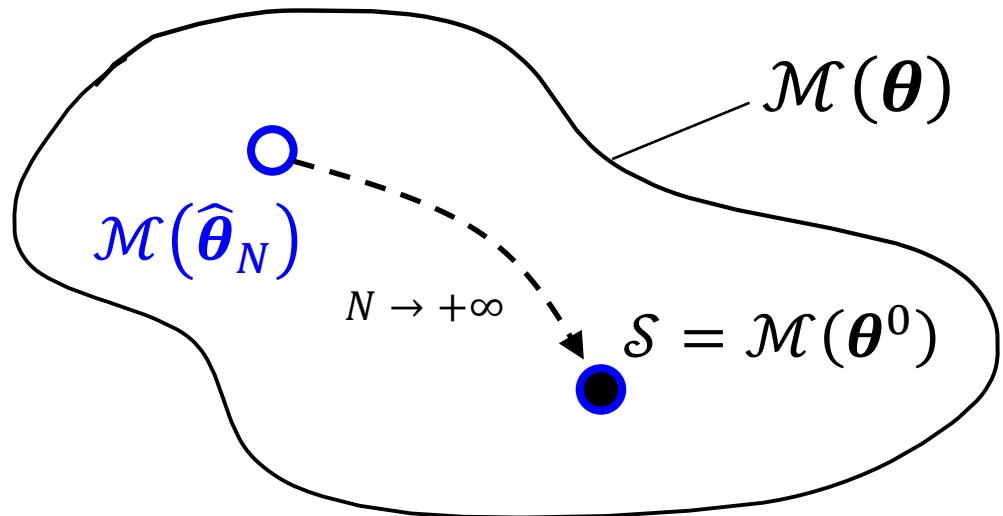
- Se $\mathcal{S} \notin \mathcal{M}(\theta)$, allora i metodi PEM **non garantiscono** di stimare correttamente **TUTTE** le componenti del sistema \mathcal{S} (i.e. modello I/O e modello del rumore)
- Se $\mathcal{S} \in \mathcal{M}(\theta)$, allora in corrispondenza di θ^0 si ha che $\varepsilon_1(t; \theta^0) = e(t) \sim WN$

Quindi, possiamo **verificare a posteriori** se il modello identificato è quello vero facendo un **test di bianchezza sui residui** $\varepsilon_1(t; \hat{\theta}_N)$



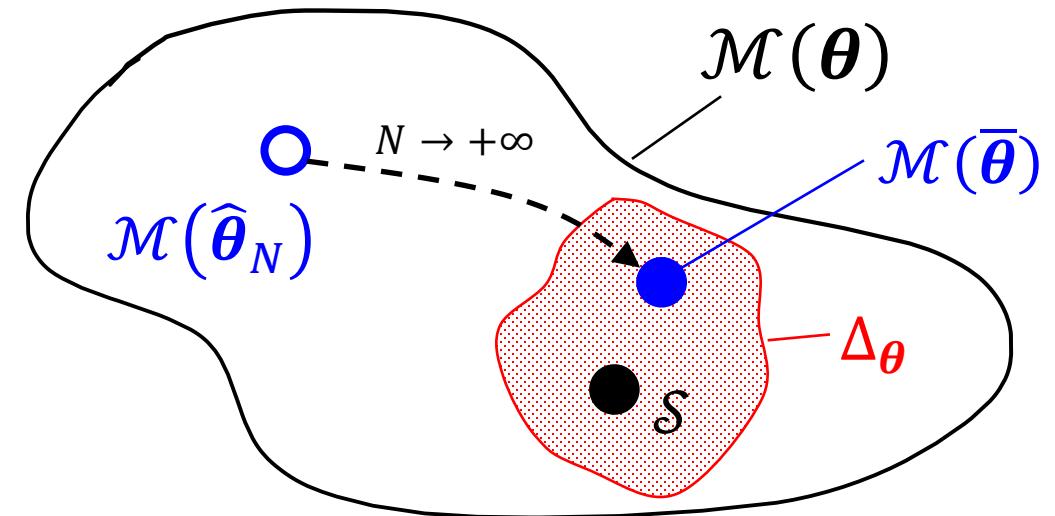
Quando identifichiamo un modello $\mathcal{M}(\theta)$, possono capitare quattro casi possibili:

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



$$\hat{\theta}_N \xrightarrow{N \rightarrow +\infty} \bar{\theta} = \theta^0$$

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

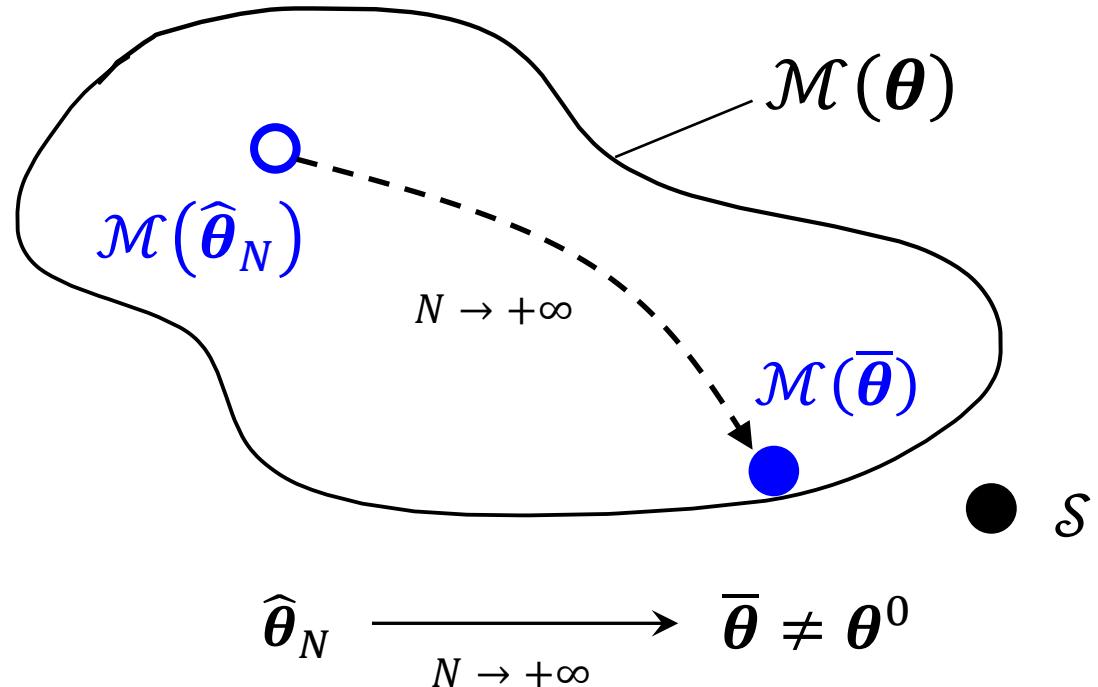


$$\hat{\theta}_N \xrightarrow{N \rightarrow +\infty} \bar{\theta} \neq \theta^0$$

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

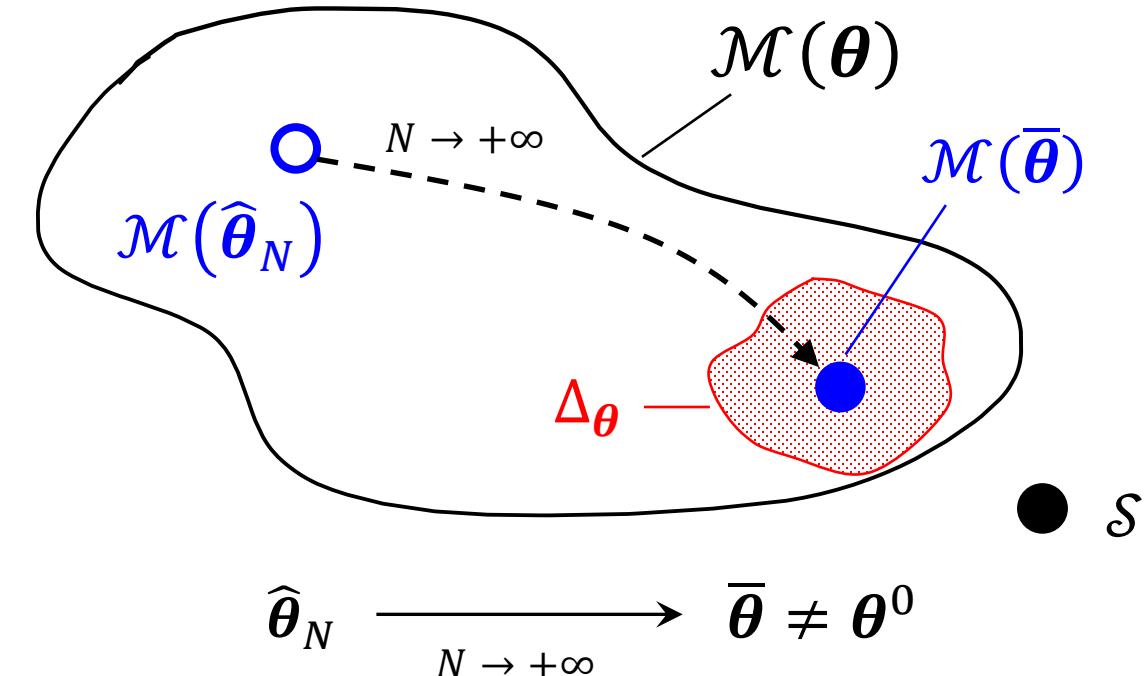


3) $\mathcal{S} \notin \mathcal{M}(\theta)$ e $\Delta_\theta = \bar{\theta}$, allora $\bar{\theta} \neq \theta^0$



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

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



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

Outline

1. Analisi asintotica dei metodi PEM
- 2. Identificabilità dei modelli e persistente eccitazione**
3. Valutazione dell'incertezza della stima PEM
4. Robustezza dei metodi PEM e prefiltraggio
5. Empirical Transfer Function Estimate (ETFE)



Identificabilità dei modelli

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

Questa seconda situazione, in cui troviamo un modello all'interno di $\{\mathcal{M}(\theta) | \theta \in \Delta_\theta\}$, porta ad un legittima domanda:

In quali condizioni il sistema \mathcal{S} può essere **identificato univocamente** dai dati?



Identificabilità dei modelli

Affinché un modello sia **univocamente identificabile** è necessario avere:

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

Il problema di non identificabilità **più critico è quello sperimentale**: se non abbiamo sufficiente informazione nei dati, non possiamo fare nulla (se non ripetere l'esperimento)

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



Esempio: problema di identificabilità strutturale

Supponiamo che i dati siano generati da un **sistema** (deterministico) del tipo

$$\mathcal{S}: \quad G_0(z) = \frac{z}{(z + 0.5)(z + 0.8)}$$

Per l'identificazione, usiamo un **modello** del tipo

$$\mathcal{M}(\boldsymbol{\theta}): \quad G(z) = \frac{z(b_1 z + b_2)}{(z + a_1)(z + a_2)(z + a_3)}$$
$$\boldsymbol{\theta} = [b_1 \ b_2 \ a_1 \ a_2 \ a_3]^\top$$

Notiamo che $\mathcal{S} \in \mathcal{M}(\boldsymbol{\theta})$. Infatti, posso ottenere $G_0(z)$ imponendo, per esempio, che $(b_1 z + b_2) = (z + a_3)$, in modo che il modello diventi

$$\mathcal{M}(\boldsymbol{\theta}): \quad G(z) = \frac{z}{(z + a_1)(z + a_2)}$$



Esempio: problema di identificabilità strutturale

Nonostante riusciamo a identificare un modello che è simile al sistema che genera i dati, tale modello identificato **non è unico in quanto** la condizione $(b_1z + b_2) = (z + a_3)$ è soddisfatta da un **numero infinito** di triplete di valori (b_1, b_2, a_3)

Questo accade perché **la struttura del modello è sovra-parametrizzata** rispetto alla struttura del sistema vero



Esempio: problema di identificabilità sperimentale

Supponiamo che i dati siano generati da un **sistema** del tipo

$$\mathcal{S}: \quad y(t) = \underbrace{\frac{b_0 z^{-1}}{1 + f_0 z^{-1}}}_{G_0(z)} u(t) + \underbrace{\frac{1}{1 + d_0 z^{-1}}}_{H_0(z)} e(t)$$

Consideriamo un ingresso $u(t) = 0 \ \forall t$, cioè un **ingresso costante a zero**, e un modello

$$\mathcal{M}(\theta): \quad y(t) = \underbrace{\frac{bz^{-1}}{1 + fz^{-1}}}_{G(z; \theta)} u(t) + \underbrace{\frac{1}{1 + dz^{-1}}}_{H(z; \theta)} \eta(t) \quad \theta = [b \ f \ d]^\top$$

Notiamo che $\mathcal{S} \in \mathcal{M}(\theta)$



Esempio: problema di identificabilità sperimentale

Calcoliamo l'errore di predizione a un passo

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

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

$$\varepsilon_1(t; \boldsymbol{\theta}) = \frac{1 + dz^{-1}}{1 + d_0 z^{-1}} e(t)$$

La quantità $\mathbb{E}[\varepsilon_1(t; \boldsymbol{\theta})^2]$ è **minimizzata** quando $\varepsilon_1(t; \bar{\boldsymbol{\theta}}) = e(t)$. Questo accade quando $H_0(z) = H(z, \bar{\boldsymbol{\theta}})$, ovvero quando

$$\bar{\boldsymbol{\theta}} = [b \ d_0 \ f]^\top, \quad \forall b, f \in \mathbb{R}$$

Abbiamo stimato il vero
parametro d_0 !



Esempio: problema di identificabilità sperimentale

Misurando l'uscita $y(t)$ quando l'ingresso $u(t) = 0$ ci consente di «misurare l'effetto prodotto dal rumore $e(t)$ sull'uscita». Ciò permette di **identificare correttamente il modello dell'errore** (sotto l'ipotesi che $H_0(z) \in H(z, \theta)$,隐式在於 $\mathcal{S} \in \mathcal{M}(\theta)$)

Però, questi dati **non ci consentono di identificare** i parametri di $G_0(z)$! Infatti, la funzione di costo è minimizzata indipendentemente da b ed f , che possono **assumere qualsiasi valore**. Ancora una volta, il modello non viene stimato in modo univoco

Diciamo quindi che i dati **non sono abbastanza informativi** per stimare in modo completo ed univoco tutto il sistema (cioè, sia $G_0(z)$ che $H_0(z)$)



Identificabilità sperimentale dei modelli ARX

Investighiamo il problema di **identificabilità sperimentale**, considerando per semplicità l'identificazione di un modello ARX($n_a, n_b, 1$), avendo N dati $\{u(1), \dots, u(N)\}, \{y(1), \dots, y(N)\}$

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

- $B(z) = b_0 + b_1 z^{-1} + \dots + b_{n_b} z^{-n_b}$
- $A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - \dots - a_{n_a} z^{-n_a}$

Sappiamo che la stima può essere ottenuta tramite il metodo dei minimi quadrati

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



Identificabilità sperimentale dei modelli ARX

Problema di identificabilità: quando $\hat{\theta}_N$ **esiste** ed è **unico?**



quando $\sum_{t=1}^N \varphi(t)\varphi^\top(t)$ è **invertibile?**

Definiamo:

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

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



Identificabilità sperimentale dei modelli ARX

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

$$\det_{d \times d}(R(N)) > 0$$

Analizziamo la matrice $R(N)$ per $N \rightarrow +\infty$. Consideriamo come punto di partenza un modello ARX(1, 0, 1), ovvero con $n_a = 1, n_b = 0, k = 1$

$$y(t) = a_1 y(t-1) + b_0 u(t-1) + e(t), \quad e(t) \sim \text{WN}(0, \lambda^2)$$

dove $y(t)$ e $u(t)$ sono **pss ergodici a media nulla**



Identificabilità sperimentale dei modelli ARX

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

$$\varphi(t)\varphi^\top(t) = \begin{bmatrix} y(t-1)^2 & y(t-1)u(t-1) \\ u(t-1)y(t-1) & u(t-1)^2 \end{bmatrix}_{2 \times 1 \quad 1 \times 2}$$

$$S(N) = \sum_{t=1}^N \varphi(t)\varphi^\top(t)$$

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

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



Identificabilità sperimentale dei modelli ARX

Grazie all'**ipotesi di ergodicità**, abbiamo che $R(N) \xrightarrow[N \rightarrow +\infty]{} \bar{R}$

Anche perché $y(\cdot)$ e $u(\cdot)$ hanno media nulla

$$\bar{R} = \begin{bmatrix} \gamma_{yy}(0) & \gamma_{yu}(0) \\ \gamma_{uy}(0) & \gamma_{uu}(0) \end{bmatrix}_{2 \times 2}$$

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

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



Identificabilità sperimentale dei modelli ARX

In generale, per un generico modello ARX($n_a, n_b, 1$) abbiamo che la matrice \bar{R} può essere divisa in quattro sotto-matrici

$$\bar{R} = \begin{bmatrix} \bar{R}_{yy} & \bar{R}_{yu} \\ \bar{R}_{uy} & \bar{R}_{uu} \end{bmatrix} \quad \begin{matrix} n_a \times n_a & n_a \times (n_b + 1) \\ (n_b + 1) \times n_a & (n_b + 1) \times (n_b + 1) \end{matrix}$$

$$\bar{R}_{yy} = \begin{bmatrix} \gamma_{yy}(0) & \gamma_{yy}(1) & \gamma_{yy}(2) & \cdots & \cdots & \gamma_{yy}(n_a - 1) \\ \gamma_{yy}(1) & \gamma_{yy}(0) & \gamma_{yy}(1) & \gamma_{yy}(2) & \cdots & \gamma_{yy}(n_a - 2) \\ \gamma_{yy}(2) & \gamma_{yy}(1) & \gamma_{yy}(0) & \gamma_{yy}(1) & \cdots & \gamma_{yy}(n_a - 3) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \gamma_{yy}(n_a - 1) & \cdots & \cdots & \cdots & \cdots & \gamma_{yy}(0) \end{bmatrix} \quad \begin{matrix} n_a \times n_a \end{matrix}$$

- Matrice autocovarianza di ordine $n_a - 1$ di $y(t)$
- Struttura Toeplitz
- Dimensioni $n_a \times n_a$



Identificabilità sperimentale dei modelli ARX

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

- Matrice autocovarianza di ordine n_b di $u(t)$
- Struttura Toeplitz
- Dimensioni $(n_b + 1) \times (n_b + 1)$

$$\bar{R}_{yu} = \begin{bmatrix} \gamma_{yu}(0) & \gamma_{yu}(1) & \gamma_{yu}(2) & \cdots & \cdots & \gamma_{yu}(n_a - 1) \\ \gamma_{yu}(1) & \gamma_{yu}(0) & \gamma_{yu}(1) & \gamma_{yu}(2) & \cdots & \gamma_{yu}(n_a - 2) \\ \gamma_{yu}(2) & \gamma_{yu}(1) & \gamma_{yu}(0) & \gamma_{yu}(1) & \cdots & \gamma_{yu}(n_a - 3) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \gamma_{yu}(n_a - 1) & \cdots & \cdots & \cdots & \cdots & \gamma_{yu}(0) \end{bmatrix}_{n_a \times (n_b + 1)}$$

- Matrice autocovarianza tra $y(\cdot)$ e $u(\cdot)$
- $\bar{R}_{yu} = \bar{R}_{uy}^\top$
- Dimensioni $n_a \times (n_b + 1)$

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



Identificabilità sperimentale dei modelli ARX

Lemma di Schur

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

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

Ricordando che

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

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



Identificabilità sperimentale dei modelli ARX

La condizione (**solo necessaria**) sulla matrice \bar{R}_{uu} è interessante perché riguarda **solo il segnale di ingresso** $u(t)$, che tipicamente **progettiamo noi!**

Possiamo quindi tenere conto di questa condizione in fase di progettazione dell'esperimento, e **scegliere il segnale di eccitazione** più opportuno al fine di ottenere **dati informativi**



Persistente eccitazione

Definizione (Persistente eccitazione)

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

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

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

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

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



Persistente eccitazione

Possiamo quindi dire che **condizione necessaria** per l'identificabilità sperimentale di un modello ARX($n_a, n_b, 1$) è che il segnale $u(t)$, usato per produrre i dati, sia **«persistentemente eccitante» di ordine pari ad almeno $n_b + 1$** (infatti, \bar{R}_{uu} ha dimensione $(n_b + 1) \times (n_b + 1)$)

Tale conclusione si può generalizzare nel modo seguente. Supponendo che $\mathcal{S} \in \mathcal{M}(\theta)$, definiamo il numero di parametri di $G(z, \theta)$ come n_g . Allora, la soluzione del problema di identificazione $\bar{\theta} = \arg \min_{\theta} \mathbb{E}[\varepsilon_1(t, \theta)^2]$ ha un'**unica soluzione** $\bar{\theta} = \theta^0$ se il segnale $u(t)$ che genera i dati è **persistente eccitante di ordine $\geq n_g$**



Persistente eccitazione

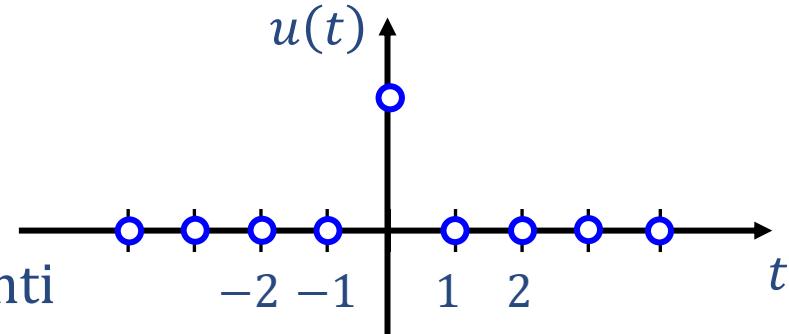
Osservazioni

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



Esempio: ingresso impulsivo

Consideriamo come ingresso $u(t) = \text{imp}(t) = \begin{cases} 1 & \text{se } t = 0 \\ 0 & \text{se altrimenti} \end{cases}$



La matrice $R_{uu}(N)$ assume la forma seguente:

$$R_{uu}(N) = \frac{1}{N} \begin{bmatrix} \sum_{t=1}^N u(t-1)^2 & \sum_{t=1}^N u(t)u(t-1) & \dots & \sum_{t=1}^N u(t)u(t-n_b) \\ \sum_{t=1}^N u(t)u(t-1) & \sum_{t=1}^N u(t-1)^2 & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \sum_{t=1}^N u(t)u(t-n_b) & \dots & \dots & \sum_{t=1}^N u(t-1)^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{N} & 0 & \dots & 0 \\ 0 & \frac{1}{N} & \dots & \dots \\ \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & \frac{1}{N} \end{bmatrix}$$



Esempio: ingresso impulsivo

Dato che $R_{uu}(N) = \begin{bmatrix} \frac{1}{N} & 0 & \cdots & 0 \\ 0 & \frac{1}{N} & \cdots & \cdots \\ \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & \frac{1}{N} \end{bmatrix}_{(n_b+1) \times (n_b+1)}$

allora $\lim_{N \rightarrow \infty} R_{uu}(N) = \bar{R}_{uu} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \cdots \\ \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & 0 \end{bmatrix}_{(n_b+1) \times (n_b+1)}$

Per cui l'impulso **non è persistentemente eccitante di nessun ordine**

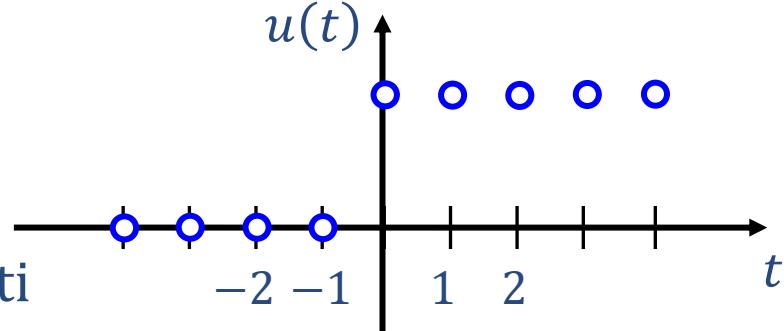
Questo non vuol dire che la risposta all'impulso non può essere usata per stimare modelli, anzi ci dà molte informazioni sul sistema $G_0(z)$. La nozione di persistente eccitazione serve però a garantire **stime consistenti** anche in presenza di **disturbi**

In questo caso, il problema è che «osservo solo una volta» ogni valore della risposta all'impulso, e quindi se quel valore è corrotto da rumore «non posso confrontarlo» con altri valori analoghi, e me lo devo tenere sporco dal rumore



Esempio: ingresso scalino

Consideriamo come ingresso $u(t) = \text{sca}(t) = \begin{cases} 1 & \text{se } t \geq 0 \\ 0 & \text{se } \text{altrimenti} \end{cases}$



La matrice $R_{uu}(N)$ assume la forma seguente:

$$R_{uu}(N) = \frac{1}{N} \begin{bmatrix} \sum_{t=1}^N u(t-1)^2 & \sum_{t=1}^N u(t)u(t-1) & \dots & \sum_{t=1}^N u(t)u(t-n_b) \\ \sum_{t=1}^N u(t)u(t-1) & \sum_{t=1}^N u(t-1)^2 & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \sum_{t=1}^N u(t)u(t-n_b) & \dots & \dots & \sum_{t=1}^N u(t-1)^2 \end{bmatrix} \xrightarrow{(n_b+1) \times (n_b+1)} \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & \dots \\ \dots & \dots & \dots & 1 \\ 1 & \dots & \dots & 1 \end{bmatrix}$$

È persistentemente eccitante di ordine 1



Esempio: ingresso impulso e scalino

Come informazione ulteriore, notiamo che **impulso** e **scalino caratterizzano completamente** un sistema dinamico lineare (a meno della sua condizione iniziale), grazie anche al periodo di **transitorio** della risposta a questi segnali

È utile ricordare che il framework PEM che stiamo studiando qui **non è adatto all'utilizzo di segnali con transitori** (infatti richiede che essi siano realizzazioni di pss)

Altri metodi, come quelli di **identificazione a sottospazi**, possono utilizzare segnali con transitorio



Esempio: ingresso rumore bianco

Consideriamo ora come ingresso $u(t) = \text{WN}(0, \lambda^2)$. In questo caso, la matrice $\bar{R}_{uu}^{(i)}$ risulta essere:

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

Quindi, 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»)



Persistente eccitazione nel dominio delle frequenze

La nozione di **persistent eccentricity** di un segnale di ingresso $u(t)$ può essere interpretata anche nel **dominio delle frequenze**

A tale fine, consideriamo un processo stocastico stazionario generato nel modo seguente:

$$y(t) = c_0 u(t) + c_1 u(t - 1) + c_2 u(t - 2) + \cdots + c_n u(t - n)$$

dove $u(t)$ è un **processo stazionario a media nulla** e con funzione di autocovarianza $\gamma_{uu}(\tau)$ ($u(t)$ non è per forza un WN)

Questo processo è anche chiamato **Generalized Moving Average (GMA)** data la sua somiglianza al processo MA



Persistente eccitazione nel dominio delle frequenze

Calcoliamo la varianza di $y(t)$

$$\begin{aligned}\text{Var}[y(t)] &= \mathbb{E}[y(t)^2] = c_0^2 \mathbb{E}[u(t)^2] + c_1^2 \mathbb{E}[u(t-1)^2] + \cdots + c_n^2 \mathbb{E}[u(t)^2] + \\ &+ c_0 c_1 \mathbb{E}[u(t)u(t-1)] + c_1 c_2 \mathbb{E}[u(t-1)u(t-2)] + \cdots \\ &+ c_0 c_2 \mathbb{E}[u(t)u(t-2)] + c_1 c_3 \mathbb{E}[u(t-1)u(t-2)] + \cdots\end{aligned}$$

Per cui

$$\text{Var}[y(t)] = (c_0^2 + c_1^2 + \cdots + c_n^2) \gamma_{uu}(0) + (c_0 c_1 + c_1 c_2 + \cdots) \gamma_{uu}(1) + (c_0 c_2 + c_1 c_3 + \cdots) \gamma_{uu}(2) + \cdots$$



Persistente eccitazione nel dominio delle frequenze

Definiamo ora la matrice $\bar{R}_{uu}^{(i)}$ di ordine $i = n + 1$

$$\bar{R}_{uu}^{(i)} = \begin{bmatrix} \gamma_{uu}(0) & \gamma_{uu}(1) & \cdots & \gamma_{uu}(i-1) \\ \gamma_{uu}(1) & \gamma_{uu}(0) & \cdots & \gamma_{uu}(i-2) \\ \dots & \dots & \dots & \dots \\ \gamma_{uu}(i-1) & \cdots & \cdots & \gamma_{uu}(0) \end{bmatrix}_{i \times i} \quad \xrightarrow{\hspace{1cm}} \quad \bar{R}_{uu}^{(n+1)} = \begin{bmatrix} \gamma_{uu}(0) & \gamma_{uu}(1) & \cdots & \gamma_{uu}(n) \\ \gamma_{uu}(1) & \gamma_{uu}(0) & \cdots & \gamma_{uu}(n-1) \\ \dots & \dots & \dots & \dots \\ \gamma_{uu}(n) & \gamma_{uu}(n-1) & \cdots & \gamma_{uu}(0) \end{bmatrix}_{(n+1) \times (n+1)}$$

Ponendo $\mathbf{c} = [c_0 \ c_1 \ c_2 \ \dots \ c_n]^\top \in \mathbb{R}^{(n+1) \times 1}$, possiamo scrivere che

$$\text{Var}[y(t)] = \mathbf{c}^\top \cdot \bar{R}_{uu}^{(n+1)} \cdot \mathbf{c}_{(n+1) \times (n+1) \quad (n+1) \times 1}$$



Esempio: autocovarianza di un processo GMA

Consideriamo il seguente processo GMA di ordine $n = 1$, dove $u(t)$ possa a media nulla

$$y(t) = c_0 u(t) + c_1 u(t - 1)$$

Calcoliamo la varianza di $y(t)$

$$\text{Var}[y(t)] = \mathbb{E}[y(t)^2] = c_0^2 \mathbb{E}[u(t)^2] + c_1^2 \mathbb{E}[u(t - 1)^2] + c_0 c_1 \mathbb{E}[u(t)u(t - 1)]$$

$$= c_0^2 \gamma_{uu}(0) + c_1^2 \gamma_{uu}(0) + c_0 c_1 \gamma(1) + c_1 c_0 \gamma(1)$$

La matrice di autocovarianze di $u(t)$ è: $\bar{R}_{uu}^{(1+1)} = \begin{bmatrix} \gamma_{uu}(0) & \gamma_{uu}(1) \\ \gamma_{uu}(1) & \gamma_{uu}(0) \end{bmatrix}$



Esempio: autocovarianza di un processo GMA

Definiamo $\mathbf{c} = [c_0 \ c_1]^\top \in \mathbb{R}^{2 \times 1}$

Per cui

$$\begin{aligned}\text{Var}[y(t)] &= \mathbf{c}^\top \cdot \bar{R}_{uu}^{(1+1)} \cdot \mathbf{c} = [c_0 \ c_1] \cdot \begin{bmatrix} \gamma_{uu}(0) & \gamma_{uu}(1) \\ \gamma_{uu}(1) & \gamma_{uu}(0) \end{bmatrix} \cdot \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} = \\ &= [c_0 \ c_1] \cdot \left(c_0 \begin{bmatrix} \gamma_{uu}(0) \\ \gamma_{uu}(1) \end{bmatrix} + c_1 \begin{bmatrix} \gamma_{uu}(1) \\ \gamma_{uu}(0) \end{bmatrix} \right) = [c_0 \ c_1] \cdot \begin{bmatrix} c_0 \cdot \gamma_{uu}(0) + c_1 \cdot \gamma_{uu}(1) \\ c_0 \cdot \gamma_{uu}(1) + c_1 \cdot \gamma_{uu}(0) \end{bmatrix} \\ &= c_0^2 \gamma_{uu}(0) + c_1^2 \gamma_{uu}(0) + c_0 c_1 \gamma_{uu}(1) + c_1 c_0 \gamma_{uu}(1)\end{aligned}$$



Persistente eccitazione nel dominio delle frequenze

Consideriamo quindi un generico vettore $\alpha = [\alpha_1 \ \alpha_2 \ \dots \ \alpha_n]^\top \in \mathbb{R}^{n \times 1}$, $\alpha \neq 0$, ed il segnale $\tilde{u}(t)$ generato filtrando il pss $u(t)$ a media nulla come

$$\tilde{u}(t) = \alpha_1 u(t-1) + \alpha_2 u(t-2) + \dots + \alpha_n u(t-n)$$

La funzione $H_\alpha(z)$ di trasferimento del filtro è

$$H_\alpha(z) = \frac{\alpha_1 z^{n-1} + \dots + \alpha_n}{z^n}$$

Possiamo quindi scrivere

$$\tilde{u}(t) = H_\alpha(z)u(t)$$



Persistente eccitazione nel dominio delle frequenze

Essendo $\tilde{u}(t)$ un processo GMA, la sua varianza è

$$\text{Var}[\tilde{u}(t)] = \boldsymbol{\alpha}^\top \cdot \bar{R}_{uu}^{(n)} \cdot \boldsymbol{\alpha}$$

È però possibile esprimere la varianza di $\tilde{u}(t)$ anche come

$$\text{Var}[\tilde{u}(t)] = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \Gamma_{\tilde{u}\tilde{u}}(\omega) d\omega = \frac{1}{2\pi} \int_{-\pi}^{+\pi} |H_\alpha(e^{j\omega})|^2 \cdot \Gamma_{uu}(\omega) d\omega$$

Da cui segue che

$$\boldsymbol{\alpha}^\top \cdot \bar{R}_{uu}^{(n)} \cdot \boldsymbol{\alpha} = \frac{1}{2\pi} \int_{-\pi}^{+\pi} |H_\alpha(e^{j\omega})| \cdot \Gamma_{uu}(\omega) d\omega$$



Persistente eccitazione nel dominio delle frequenze

$$\alpha^\top \cdot \bar{R}_{uu}^{(n)} \cdot \alpha = \frac{1}{2\pi} \int_{-\pi}^{+\pi} |H_\alpha(e^{j\omega})|^2 \cdot \Gamma_{uu}(\omega) d\omega$$

Il fatto che $\bar{R}_{uu}^{(n)} > 0$ (e quindi che $u(t)$ sia **persistentemente eccitante** di ordine n) equivale a dire che $\alpha^\top \cdot \bar{R}_{uu}^{(n)} \cdot \alpha > 0$. Quindi $\bar{R}_{uu}^{(n)}$ è **non singolare se e solo se**

$$\alpha^\top \bar{R}_{uu}^{(n)} \alpha = 0 \implies \alpha = 0$$

La precedente implicazione può essere riscritta come

$$\frac{1}{2\pi} \int_{-\pi}^{+\pi} |H_\alpha(e^{j\omega})|^2 \cdot \Gamma_{uu}(\omega) d\omega = 0 \implies H_\alpha(e^{j\omega}) = 0 \quad \forall \omega$$



Persistente eccitazione nel dominio delle frequenze

Quando succede che $\alpha^\top \cdot \bar{R}_{uu}^{(n)} \cdot \alpha = 0$ nonostante sia $\alpha \neq 0$? Notiamo che la funzione di trasferimento $H_\alpha(z)$ può avere al **massimo** $n - 1$ zeri

Quindi, $H_\alpha(z)$ può **annullarsi** se $z_k = e^{j\omega_k}$ è uno dei suoi **zeri**, con $k = 1, \dots, n - 1$

Però, se $\Gamma_{uu}(\omega) \neq 0$ per **almeno** n **pulsazioni distinte nell'intervallo** $\omega \in (-\pi, \pi]$, allora

$$\int_{-\pi}^{+\pi} |H_\alpha(e^{j\omega})| \cdot \Gamma_{uu}(\omega) d\omega \neq 0 \text{ e quindi } \alpha^\top \cdot \bar{R}_{uu}^{(n)} \cdot \alpha > 0$$

Questa interpretazione ci permette di dare una **definizione equivalente** di persistente eccitazione



Persistente eccitazione nel dominio delle frequenze

Definizione (Persistente eccitazione)

Un segnale $u(t)$, con spettro $\Gamma_{uu}(\omega)$, è **persistemente eccitante** di ordine n se e solo se, per tutti i filtri della forma

$$H_\alpha(z) = \alpha_1 z^{-1} + \alpha_2 z^{-2} + \cdots + \alpha_n z^{-n}$$

la relazione

$$\left| H_\alpha(e^{j\omega}) \right|^2 \cdot \Gamma_{uu}(\omega) = 0 \quad \forall \omega$$

implica che $H_\alpha(e^{j\omega}) = 0 \quad \forall \omega$



Persistente eccitazione nel dominio delle frequenze

Possiamo quindi concludere che:

1. Un **pss** $u(t)$ a **media nulla** (media non nulla) è **persistentemente eccitante di ordine n ($n + 1$)** se $\Gamma_{uu}(\omega) \neq 0$ in almeno n pulsazioni $\omega \in (-\pi, \pi]$ **distinte**

Esempio: il segnale $u(t) = \sin(\omega_0 t)$ è persistentemente eccitante di ordine $n = 2$ ($\Gamma_{uu}(\omega)$ ha contributi alle pulsazioni $\pm\omega_0$)

2. Un **processo ARMA** è **persistentemente eccitante di ogni ordine**

Infatti, $\Gamma_{yy}(\omega_k) = 0$ solo se $z_k = e^{j\omega_k}$ è uno degli zeri di $C(z)$, e vi sono altre infinite frequenze per cui la sua densità spettrale di potenza non si annulla

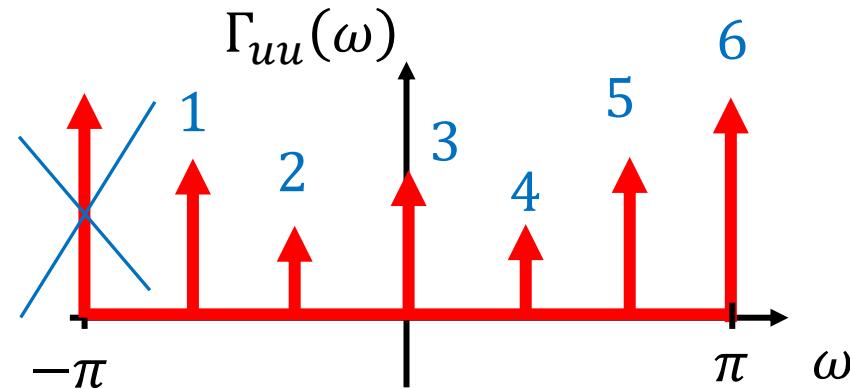


Persistente eccitazione nel dominio delle frequenze

Considerazioni ulteriori

- Un segnale $u(t)$, **periodico** di periodo T , è **al massimo** (potrebbe essere anche minore) persistentemente eccitante di ordine $n = T$

Esempio: consideriamo un pss completamente predicibile periodico di periodo T , con media non nulla



$\omega = -\pi$ e $\omega = \pi$ **sono lo stesso punto** e vanno contati una volta sola



Esempio: effetto dell'ingresso sulla stima

Consideriamo il seguente meccanismo di generazione dei dati \mathcal{S}

$$\mathcal{S}: y(t) = \frac{0.103 + 0.181z^{-1}}{1 - 1.991z^{-1} + 2.203z^{-2} - 1.841z^{-3} + 0.894z^{-4}} z^{-3} u(t) + e(t), \quad e(t) \sim WN(0, 0.5^2)$$

Notiamo che il sistema è del tipo **Output Error** OE(n_b, n_f, k), con $k = 3, n_b = 1, n_f = 4$

Utilizziamo una famiglia di modelli tali che $\mathcal{S} \in \mathcal{M}(\boldsymbol{\theta})$, ovvero

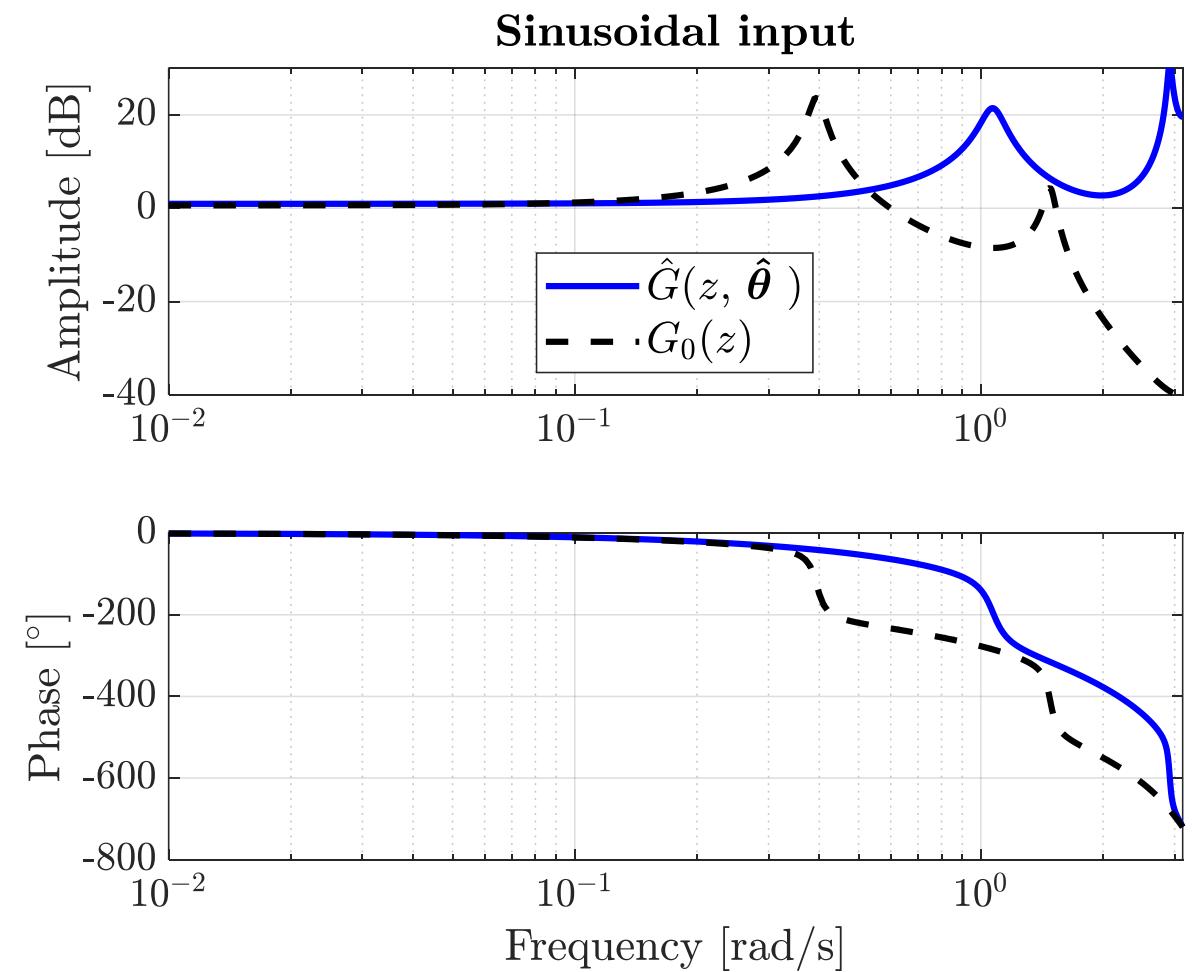
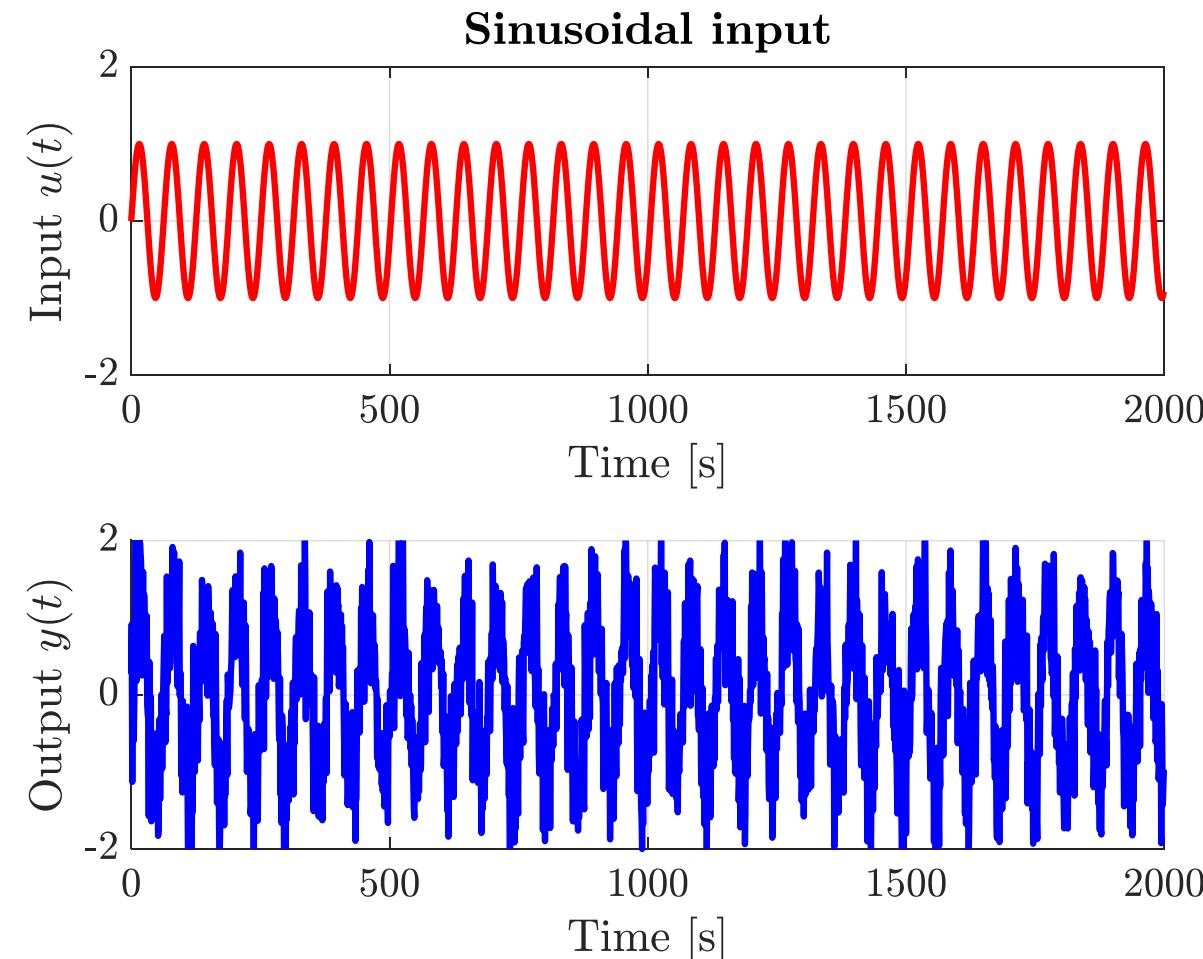
$$\mathcal{M}(\boldsymbol{\theta}): \left\{ G(z, \boldsymbol{\theta}) = \frac{b_0 + b_1 z^{-1}}{1 + f_1 z^{-1} + f_2 z^{-2} + f_3 z^{-3} + f_4 z^{-4}} z^{-3}; \quad H(z, \boldsymbol{\theta}) = 1 \right\}$$

I parametri sono $\boldsymbol{\theta} = [b_0 \ b_1 \ f_1 \ f_2 \ f_3 \ f_4]^\top \in \mathbb{R}^{6 \times 1}$. Il numero di parametri di $G(z, \boldsymbol{\theta})$ è $n_g = 6$



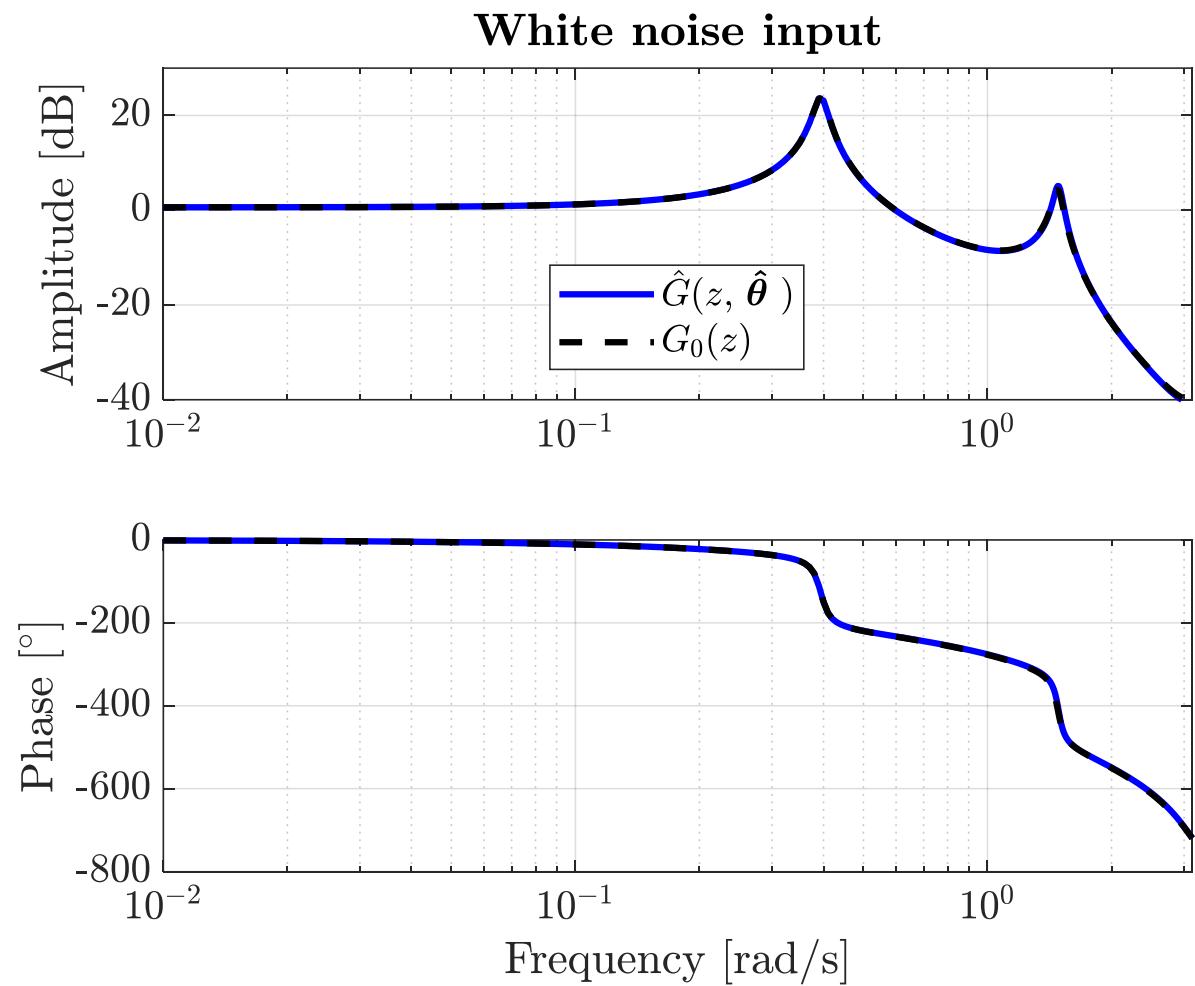
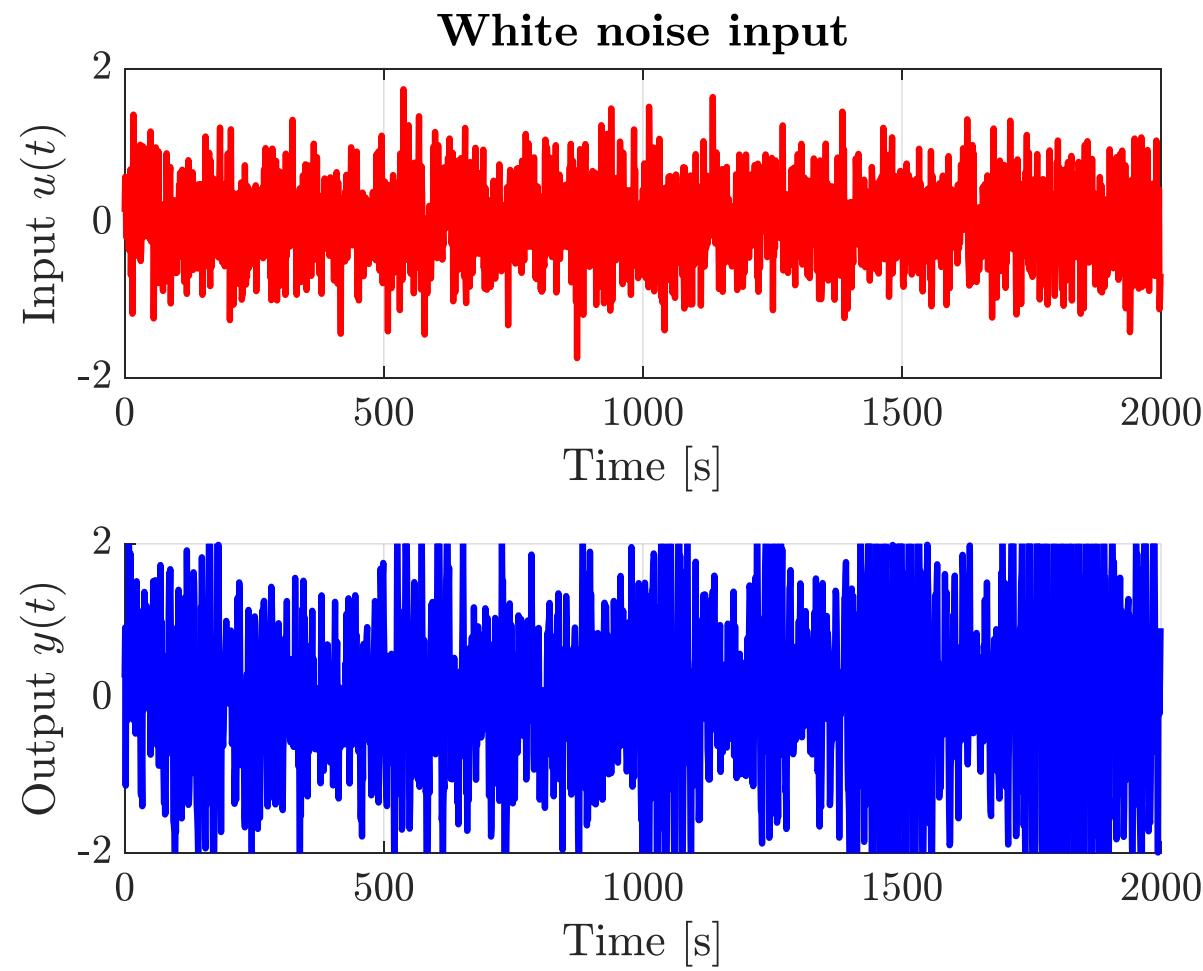
Esempio: effetto dell'ingresso sulla stima

Misuriamo $N = 2000$ dati con **ingresso sinusoidale** $u_{\sin}(t) = \sin(0.1t)$, **p.e. di ordine 2**



Esempio: effetto dell'ingresso sulla stima

Misuriamo $N = 2000$ dati con **ingresso rumore bianco** $u_{\text{wn}}(t) \sim \text{WN}(0, 0.5^2)$, p.e. di ordine ∞



Segnali di eccitazione ulteriori

Abbiamo visto che un **rumore bianco** è un ottimo segnale di eccitazione. Nella pratica, però, è **impossibile generare** un rumore bianco «perfetto»:

- al più, le sequenze di numeri saranno **pseudo-casuali**, e non casuali
- a causa di limiti dell'elettronica (e.g. capacità parassite), il segnale generato e trasmesso agli attuatori sarà **«filtrato passa-basso»**, per cui **non si avrà** uno spettro «perfettamente piatto». Inoltre, talvolta **non si vuole sollecitare troppo** gli attuatori ad alta frequenza per non rovinarli

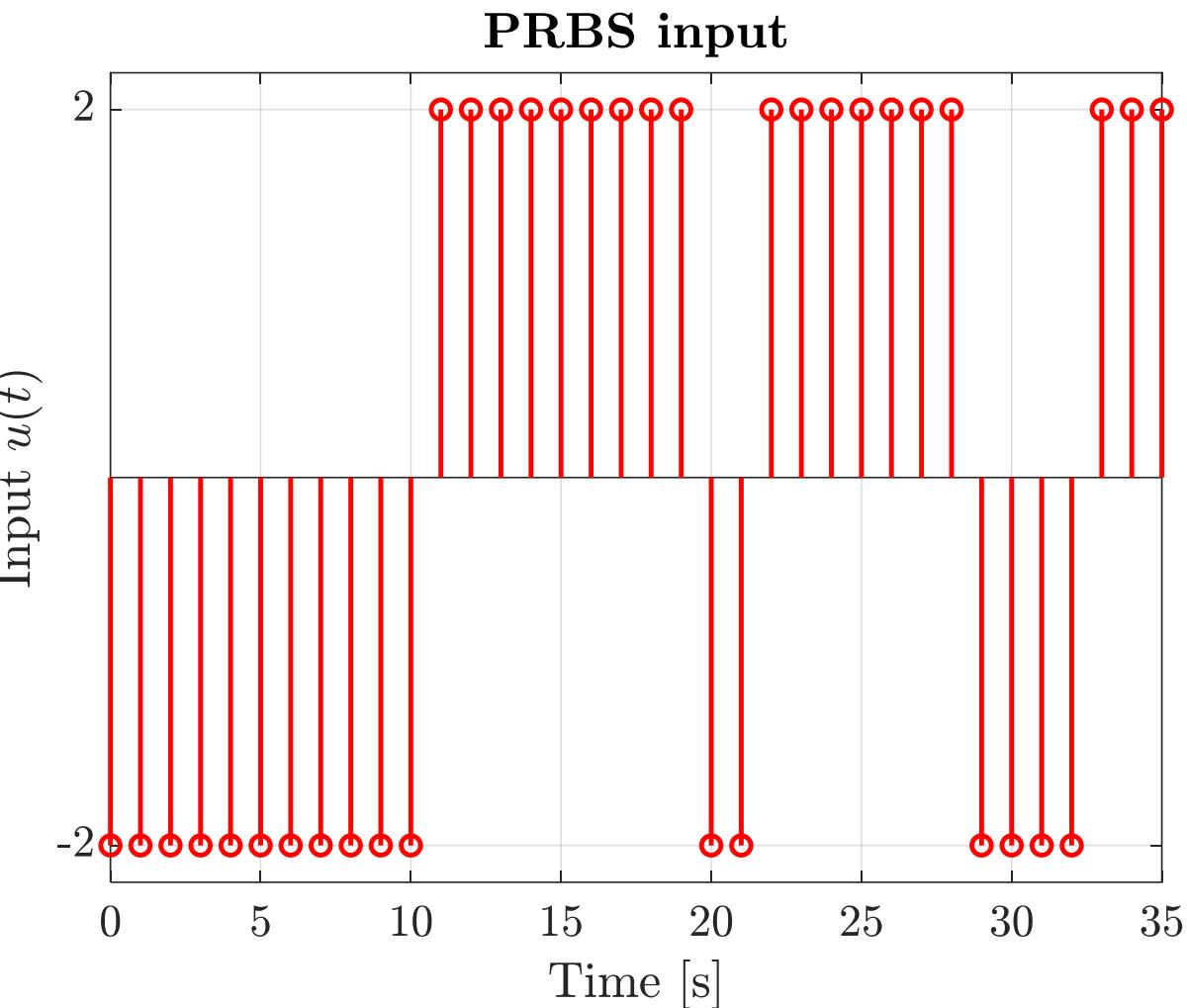
Inoltre, **l'ampiezza** del rumore bianco **non è «limitata»**. Talvolta, è necessario garantire che l'attuatore **non saturi** l'ingresso, al fine di **non introdurre nonlinearità** nell'esperimento e nei dati misurati



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** $[-\bar{u}, +\bar{u}]$, il **periodo** e **l'intervallo di clock** (il numero minimo di intervalli di tempo dopo i quali il segnale può cambiare livello)



Pseudo-Random Binary Signal (PRBS)

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

I tool per la generazione dei segnali (come Matlab) generano automaticamente dei **PRBS di lunghezza massima**, ovvero dei segnali il cui periodo è $T = 2^{n_{\text{prbs}}} - 1$, dove n_{prbs} è detto **ordine** del PRBS. È poi possibile **ripetere più volte** il segnale con tale periodo

Per cui, se settiamo $T = N$ e N non è esprimibile come $2^{n_{\text{prbs}}} - 1$, il periodo effettivamente generato non sarà di N ma del valore $2^{n_{\text{prbs}}} - 1$ più vicino a N

In matlab

```
idinput([N 1 3], 'prbs', Band, Range) Periodo di lunghezza  $N$  ripetuto 3 volte
```



Pseudo-Random Binary Signal (PRBS)

Tali «**maximum length PRBS**» hanno proprietà interessanti. Infatti, si dimostra come la **funzione di autocovarianza** di un max. length PRBS può essere espressa come

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

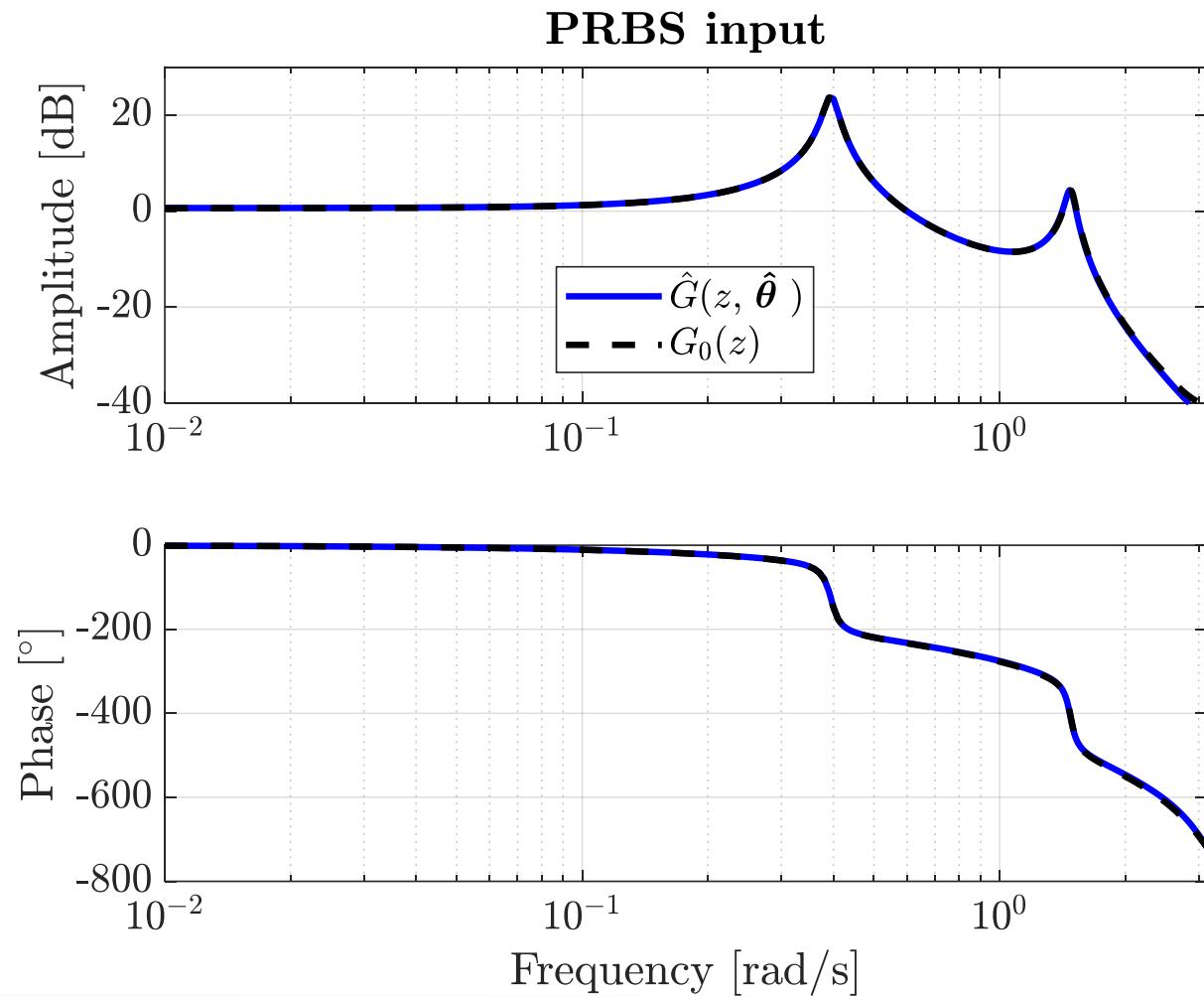
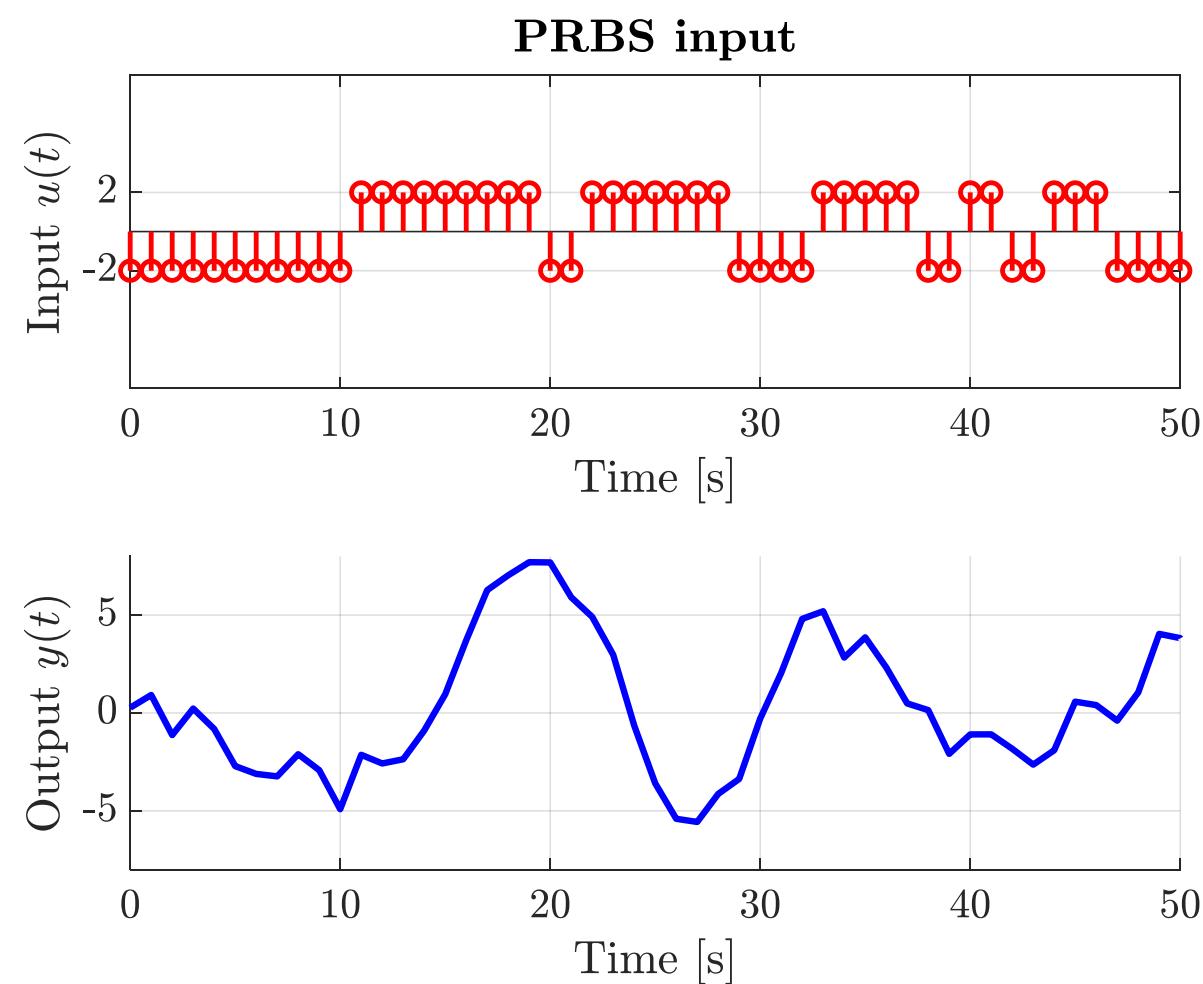
m_u è la media del PRBS, che non è esattamente zero

Notiamo che quando $T \rightarrow \infty$ il PRBS approssima un **rumore bianco**. Il PRBS è **persistentemente eccitante** di ordine T (e non può esserlo di più essendo periodico)



Esempio: effetto dell'ingresso sulla stima

Riprendiamo l'esempio precedente e misuriamo $N = 2000$ dati con **ingresso** PRBS($-2, +2$)



Multiseno

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

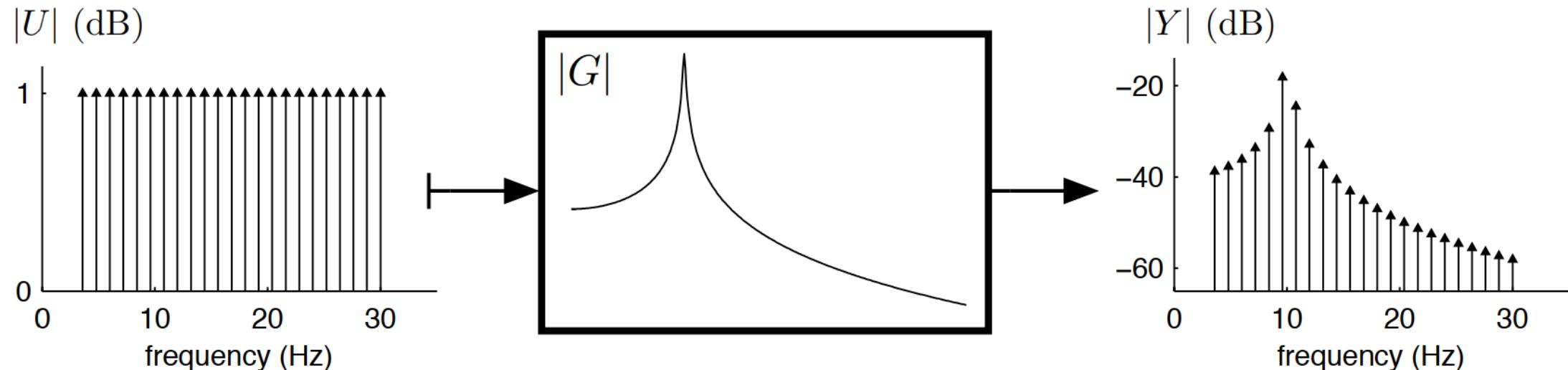
$$u(t) = \sum_{k=0}^F A_k \cdot \cos(2\pi \cdot kf_0 \cdot t + \phi_k)$$

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



Multiseno

Molto spesso le **ampiezze** A_k vengono scelte ad un **valore costante nella banda di frequenze di interesse**, e 0 altrove. Per esempio, se $A_k = 1 \ \forall k$, l'ampiezza dello spettro in frequenza dell' uscita $y(t)$ assume la forma dell'ampiezza della funzione di trasferimento alle frequenze eccitate



Multiseno

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

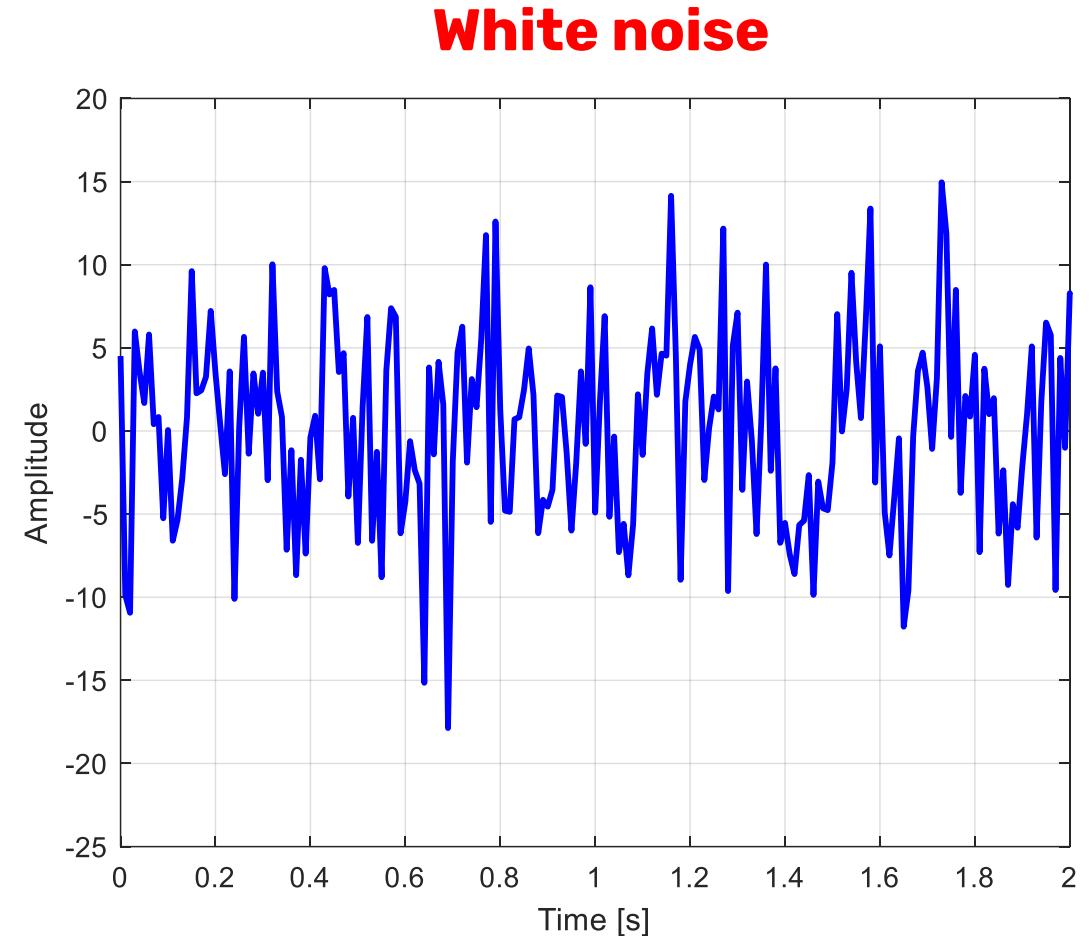
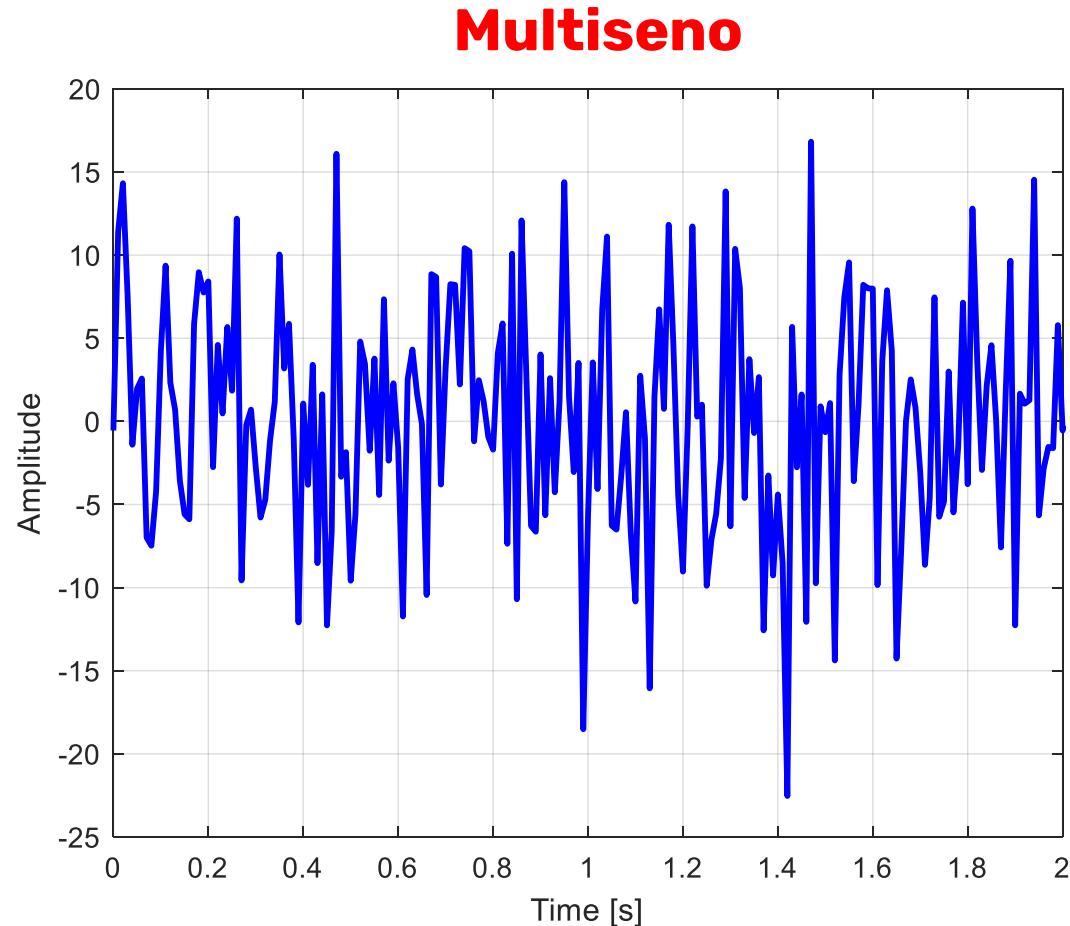
- Di solito si **scartano i primi periodi** del segnale multiseno generato

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, dove N = numero dati per periodo e P = numero di periodi del multiseno



Multiseno

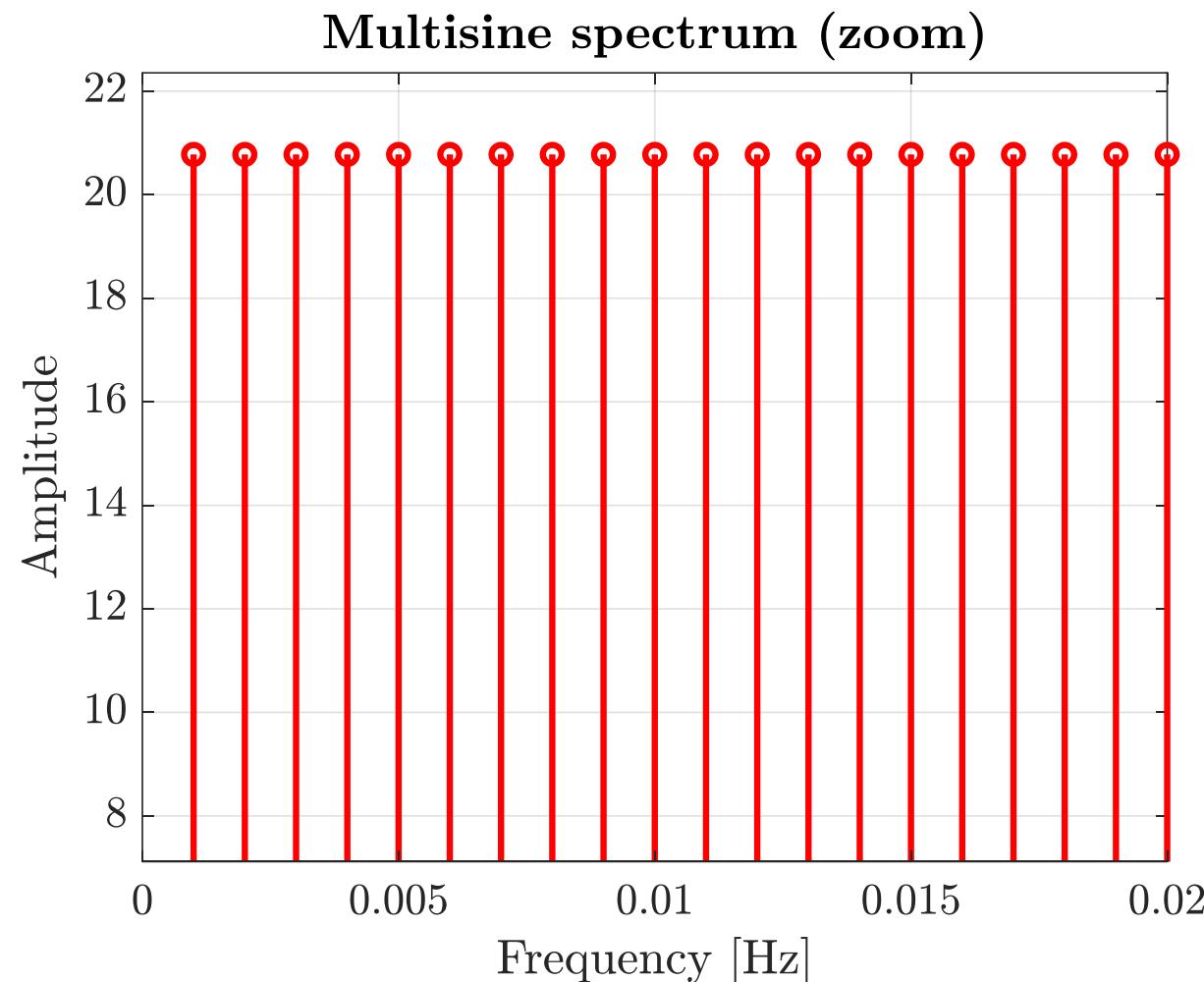
Visivamente, un multiseno è molto simile ad un rumore bianco



Esempio: multiseno

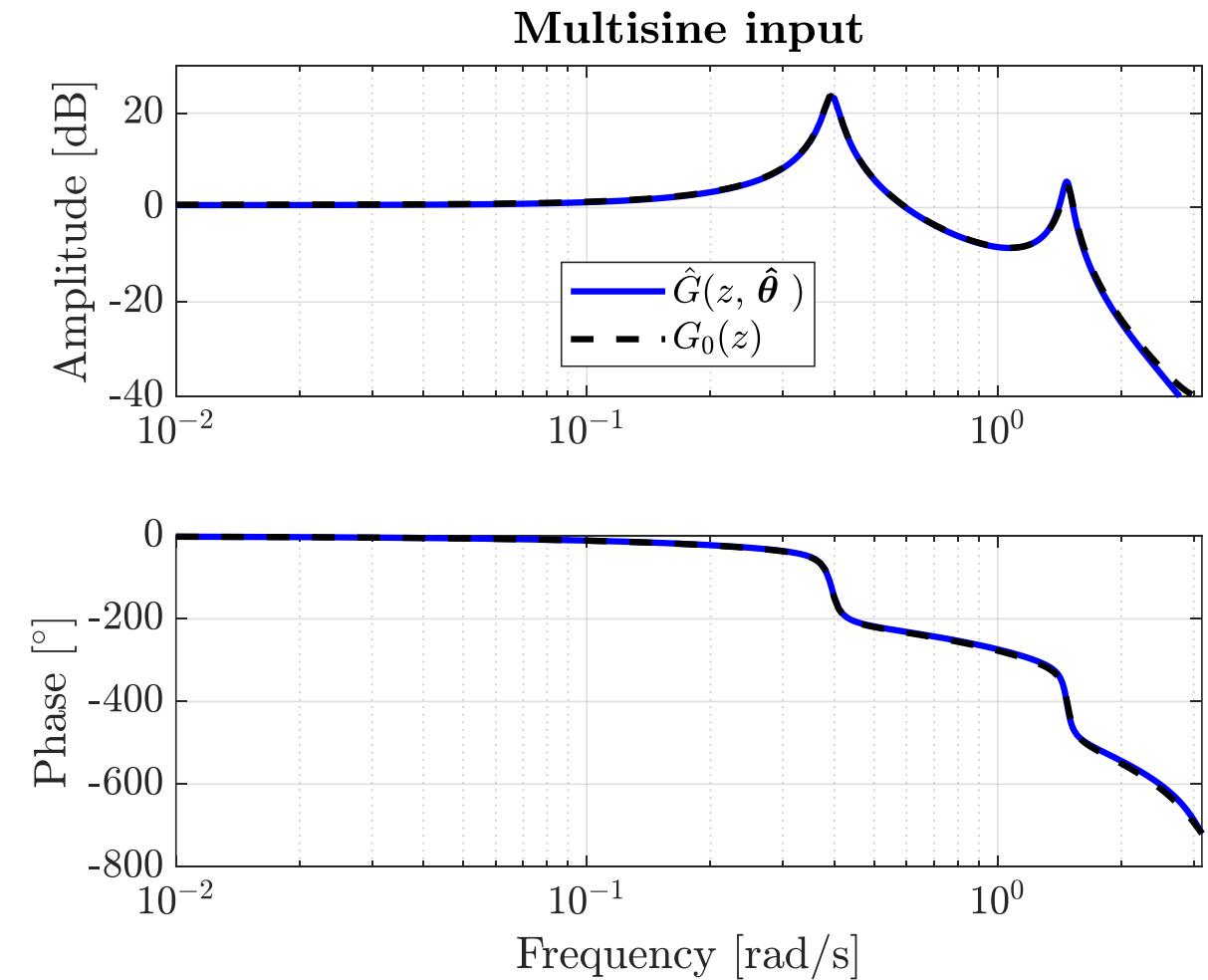
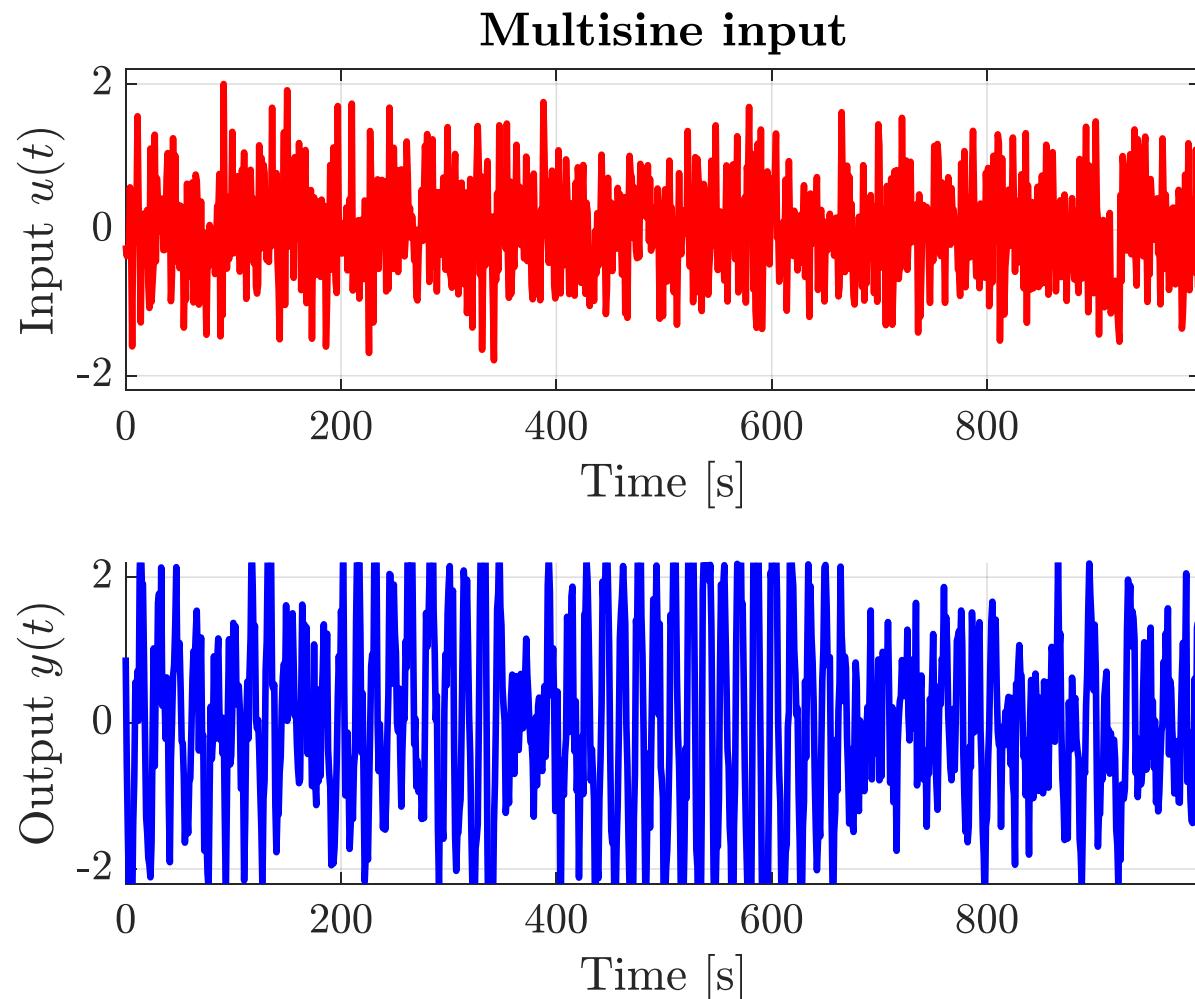
Progettiamo un multiseno con $N = 1000$ campioni per periodo, campionato a $f_s = 1 \text{ Hz}$

Misuriamo $P = 2$ periodi. Ogni periodo del multiseno è un segnale che eccita le frequenze da $1 \text{ bin} = f_s/N = 10^{-3} \text{ Hz}$ fino a 0.5 Hz , con una risoluzione di $1 \text{ bin} = f_s/N$ (la componente DC non è eccitata)



Esempio: multiseno

Riprendiamo l'esempio precedente e misuriamo $N = 2000$ dati con **ingresso multiseno**



Outline

1. Analisi asintotica dei metodi PEM
2. Identificabilità dei modelli e persistente eccitazione
- 3. Valutazione dell'incertezza della stima PEM**
4. Robustezza dei metodi PEM e prefiltraggio
5. Empirical Transfer Function Estimate (ETFE)



Varianza della stima dei parametri

Abbiamo visto che, quando $\mathcal{S} \in \mathcal{M}(\theta)$ e se l'ingresso è **sufficientemente eccitante** da rendere i dati informativi, i metodi PEM portano a stimare il valore vero dei parametri

Questo risultato vale però per $N \rightarrow \infty$. Cosa possiamo dire sulla **bontà della stima** PEM con un **numero finito** di dati?

IPOTESI DI LAVORO

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

Ipotizziamo di avere un **numero finito di dati** e di stimare $\hat{\theta}_N = \arg \min_{\theta \in \Theta} \frac{1}{N} \sum_{t=1}^N \varepsilon_1(t; \theta)^2$



Varianza della stima dei parametri

Ci ricordiamo che $\hat{\boldsymbol{\theta}}_N$ è una **variabile casuale** in quanto i dati provengono da realizzazioni di processi stocastici. Dalle ipotesi precedenti ($S \in \mathcal{M}(\boldsymbol{\theta})$ e $\boldsymbol{\theta}^0 = \bar{\boldsymbol{\theta}}$) abbiamo che $\mathbb{E}[\hat{\boldsymbol{\theta}}_N] = \boldsymbol{\theta}^0$

Vogliamo calcolare l'**incertezza di stima parametrica** $\text{Var}[\hat{\boldsymbol{\theta}}_N] = \mathbb{E}[(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^0) \cdot (\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^0)^T]$

Si dimostra che:

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

$d \times d$ $d \times d$

- $\bar{R}_{\boldsymbol{\theta}} = \mathbb{E}[\boldsymbol{\psi}(t; \boldsymbol{\theta}^0) \boldsymbol{\psi}(t; \boldsymbol{\theta}^0)^T]$
- $$\boldsymbol{\psi}(t; \boldsymbol{\theta}^0) = -\frac{d}{d\boldsymbol{\theta}} \varepsilon_1(t; \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0}$$
$$d \times 1$$
- $\lambda^2 = \text{Var}[e(t)]$



Varianza della stima dei parametri

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

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

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

Interpretazione di $\bar{P}_{\boldsymbol{\theta}}$

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

$$\frac{d\bar{J}(\boldsymbol{\theta})}{d\boldsymbol{\theta}} = \mathbb{E} \left[2\varepsilon_1(t; \boldsymbol{\theta}) \cdot \frac{d\varepsilon_1(t; \boldsymbol{\theta})}{d\boldsymbol{\theta}} \right] \quad d \times 1$$

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



Varianza della stima dei parametri

La quantità $\frac{d^2 \varepsilon_1(t; \theta)}{d\theta^2}$ è funzione dell'errore di predizione e pertanto dipende dai valori passati $e(t-1), e(t-2), \dots$. Notiamo però che se $\theta = \theta^0$, allora $\varepsilon_1(t; \theta) = e(t)$.

Ne consegue che il termine $\mathbb{E} \left[2\varepsilon_1(t; \theta) \cdot \frac{d^2 \varepsilon_1(t; \theta)}{d\theta^2} \right]$ si **annulla**, in quanto $\varepsilon_1(t)$ e $\frac{d^2 \varepsilon_1(t; \theta)}{d\theta^2}$ sono **incorrelati**. Quindi:

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



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

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



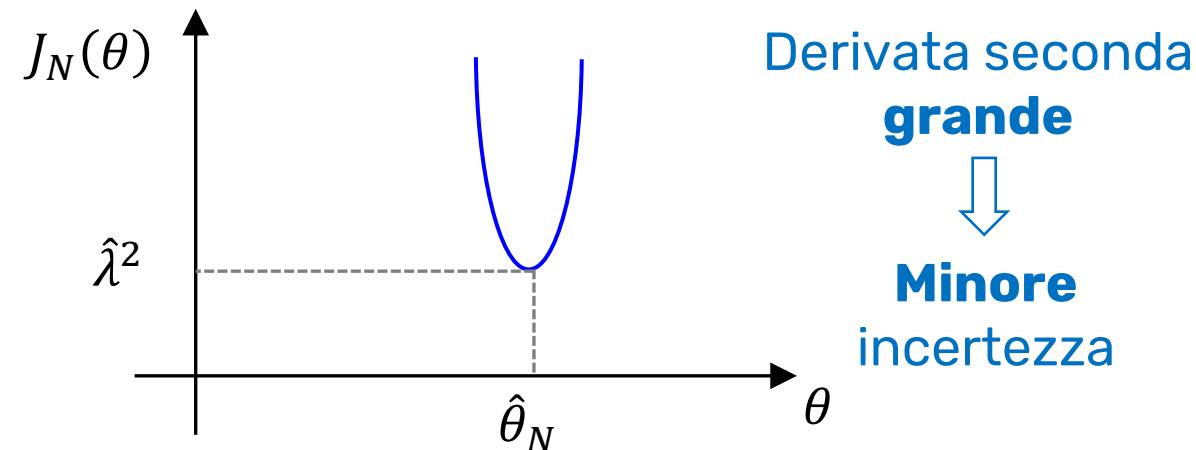
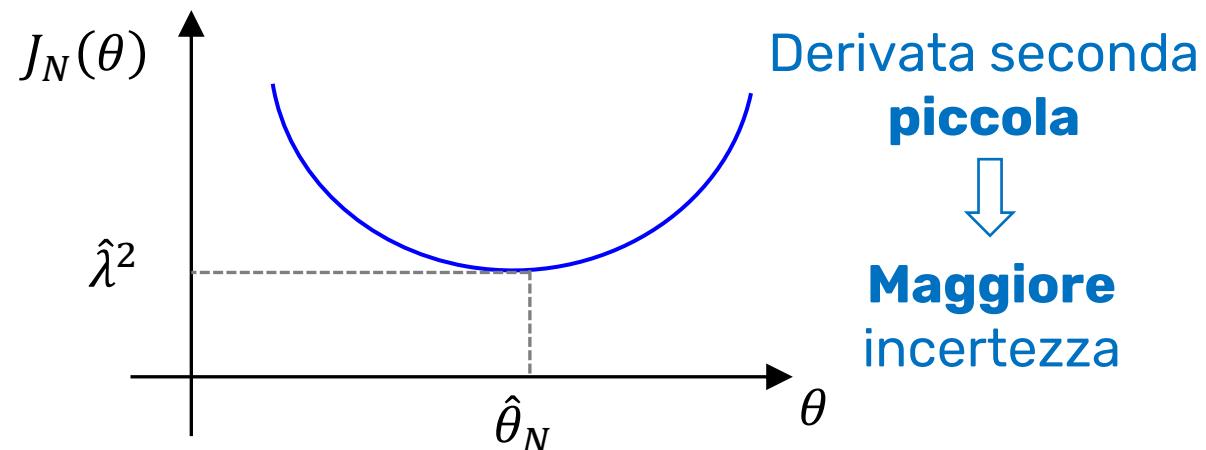
Varianza della stima dei parametri

Osservazione 1

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

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

- La varianza dell'errore di stima dei parametri decresce all'aumentare di N
- La varianza dell'errore di stima dei parametri aumenta all'aumentare di λ^2
- La varianza dell'errore di stima dei parametri diminuisce all'aumentare della derivata seconda (Hessiana) della funzione di costo all'ottimo



Varianza della stima dei parametri

Osservazione 2

Più **grande** è la «potenza» del segnale di ingresso $u(t)$, più **piccola** è la matrice di varianza delle stime \bar{P}_{θ}

Questo perché \bar{P}_{θ} è proporzionale all'inverso della «potenza» del vettore di segnali $\psi(t; \theta) = -\frac{d\varepsilon_1(t; \theta)}{d\theta}$, e questo vettore di segnali è più «potente» tanto più $u(t)$ è potente.

Infatti

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

Da cui si vede che $\varepsilon_1(t; \theta)$, e quindi anche $\psi(t; \theta)$, è proporzionale a $u(t)$



Varianza della stima dei parametri

Caso particolare: stima ARX

La stima è ottenuta tramite l'algoritmo dei minimi quadrati, con

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

$$\boldsymbol{\varphi}(t) = \begin{bmatrix} y(t-1) \\ y(t-2) \\ \vdots \\ y(t-n_a) \\ u(t-1) \\ u(t-2) \\ \vdots \\ u(t-n_b-1) \end{bmatrix}_{d \times 1}$$

L'errore di predizione a un passo è

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

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

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



Varianza della stima dei parametri

Caso particolare: stima ARX

Usando la stima campionaria \hat{P}_{θ} di \bar{P}_{θ} , abbiamo che

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

Le **proprietà probabilistiche** della stima PEM di modelli ARX sono uguali a quelle della **stima a minimi quadrati** di modelli lineari «statici» ([Lezione 3 – slide 25](#))



Distribuzione asintotica delle stime dei parametri

Se $\mathcal{S} \in \mathcal{M}(\theta)$, la distribuzione $\widehat{\theta}_N$ ottenuta tramite stima PEM converge asintoticamente ad una **Gaussiana**

$$\widehat{\theta}_N \sim \mathcal{N}(\boldsymbol{\theta}^0, \bar{P}_{\theta})$$

$d \times 1$

Nel caso in cui $\mathcal{S} \in \mathcal{M}(\theta)$, l'espressione di \bar{P}_{θ} è quella ricavata precedentemente. Altrimenti, assume una forma più complicata

La relazione $\widehat{\theta}_N \sim \mathcal{N}(\boldsymbol{\theta}^0, \hat{P}_{\theta})$, con \hat{P}_{θ} al posto di \bar{P}_{θ} , può essere **usata nella pratica** per calcolare **intervalli di confidenza** sulla stima $\widehat{\theta}_N$, e valutare così **l'affidabilità della stima** di un certo parametro



Distribuzione asintotica delle stime dei parametri

Tali **intervalli di confidenza** ci dicono la probabilità p_{θ} che l'intervallo di confidenza contenga il vettore vero dei parametri

$$\mathcal{C}_{\theta} = \left\{ \theta \mid (\theta - \hat{\theta}_N)^T \cdot \hat{P}_{\theta}^{-1} \cdot (\theta - \hat{\theta}_N) \leq \alpha \right\}$$

dove α è tale che $P(\chi^2(d) < \alpha) = p_{\theta}$. Di solito si usa $p_{\theta} = 0.95$

Il set di valori \mathcal{C}_{θ} è un **ellissoide** nello spazio dei parametri θ , centrato in $\hat{\theta}_N$. Notiamo che più \hat{P}_{θ} è grande, più grande sarà l'ellissoide



Varianza della stima delle funzioni di trasferimento

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

Per esempio, considerando $G(z, \theta)$, si può scrivere

$$\text{Var}[G(e^{j\omega}, \hat{\theta}_N)] \equiv \mathbb{E} \left[|G(e^{j\omega}, \hat{\theta}_N) - G(e^{j\omega}, \theta^0)|^2 \right]$$

Tale quantità può essere **stimata dai dati** usando \hat{P}_{θ} e $\hat{\theta}_N$



Varianza della stima delle funzioni di trasferimento

Assumendo sempre che $\mathcal{S} \in \mathcal{M}(\theta)$ e che $u(t) \perp e(t)$, e che inoltre $n \rightarrow +\infty$ (dove n è l'ordine di $G(z, \hat{\theta}_N)$, cioè il numero di variabili di stato), l'espressione si può approssimare come

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

in cui $\Gamma_{vv}(\omega)$ è la densità spettrale di potenza del disturbo $v(t) = H_0(z)e(t)$ e $\Gamma_{uu}(\omega)$ è la densità spettrale dell'ingresso $u(t)$

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



Varianza della stima delle funzioni di trasferimento

La varianza della stima della funzione di trasferimento del rumore $H(z, \hat{\theta}_N)$ è

$$\text{Var}[H(e^{j\omega}, \hat{\theta}_N)] \approx \frac{n}{N} \cdot \frac{\Gamma_{vv}(\omega)}{\lambda^2}$$

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



Outline

1. Analisi asintotica dei metodi PEM
2. Identificabilità dei modelli e persistente eccitazione
3. Valutazione dell'incertezza della stima PEM
- 4. Robustezza dei metodi PEM e prefiltraggio**
5. Empirical Transfer Function Estimate (ETFE)



Robustezza dei metodi PEM

Un obiettivo principale dell'identificazione è valutare la **mancata corrispondenza tra il modello e realtà**. Sappiamo che quando $\mathcal{S} \notin \mathcal{M}(\theta)$ **non possiamo** raggiungere la parametrizzazione vera sia di $G_0(z)$ che di $H_0(z)$. Come possiamo **caratterizzare l'errore di stima del modello** in questa situazione?

Consideriamo il seguente meccanismo di generazione dei dati

$$\mathcal{S}: y(t) = G_0(z)u(t) + H_0(z)e(t), \quad e(t) \sim \text{WN}(0, \lambda^2)$$

E il modello ARMAX

$$\mathcal{M}(\theta): y(t) = \frac{B(z, \theta)}{A(z, \theta)} u(t-1) + \frac{C(z, \theta)}{A(z, \theta)} \eta(t), \quad \eta(t) \sim \text{WN}(0, \lambda_\eta^2)$$

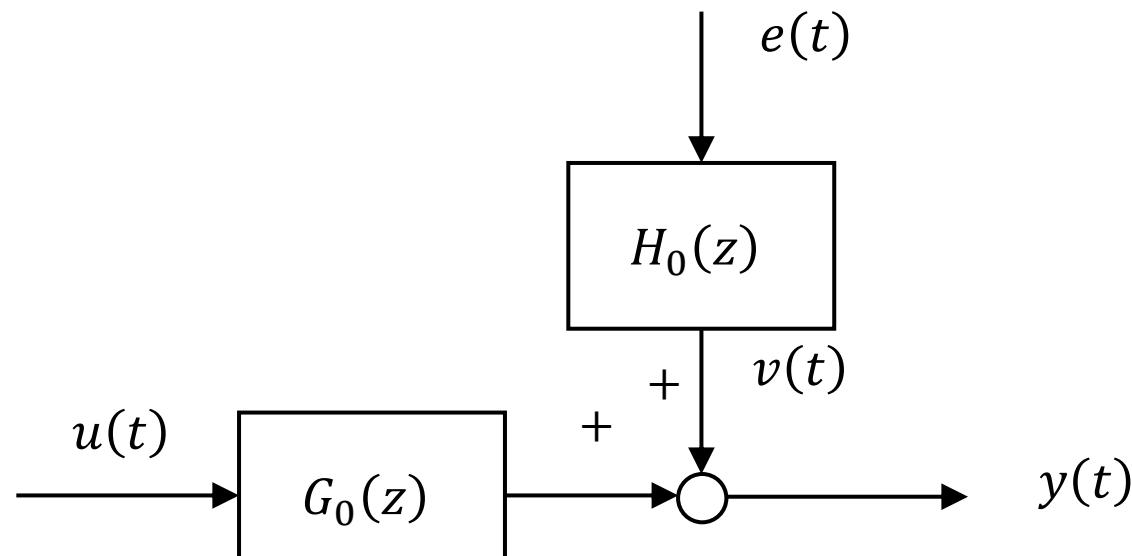


Robustezza dei metodi PEM

Problema: vogliamo analizzare la **robustezza** dei metodi PEM nel caso in cui $\mathcal{S} \notin \mathcal{M}(\theta)$. In particolare, caratterizzare come **l'errore di modellazione** di $G_0(z) \notin G(z, \theta)$ dipende dall'errore di predizione a un passo

IPOTESI DI LAVORO

1. Supponiamo che il **segnale disturbo** $v(t)$ sia **«piccolo»** rispetto all'uscita del blocco $G_0(z)$
2. Vogliamo **stimare** $G_0(z)$ nel caso in cui $G_0(z) \notin G(z, \theta)$



Robustezza dei metodi PEM

Definiamo l'**errore di modellazione** come

$$\Delta G(z, \theta) = G_0(z) - G(z, \theta)$$

Osservazione

Possiamo dividere l'errore di modellazione di $G_0(z)$ in due componenti:

- $\Delta G(z, \theta) = G_0(z) - G(z, \theta) = \underbrace{\left(G_0(z) - G(z, \bar{\theta}) \right)}_{\text{Bias di modello}} + \underbrace{\left(G(z, \bar{\theta}) - G(z, \theta) \right)}_{\text{Varianza di modello}}$

Se $\mathcal{S} \in \mathcal{M}(\theta)$, allora $G_0(z) - G(z, \bar{\theta}) = 0$. Il termine $G(z, \bar{\theta}) - G(z, \theta)$ va a zero asintoticamente



Robustezza dei metodi PEM

Ricaviamo l'espressione dell'errore $\eta(t)$ del modello ARMAX

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

$$= G(z, \boldsymbol{\theta})u(t) + H(z, \boldsymbol{\theta})\eta(t) \quad \Rightarrow \quad \boxed{\eta(t) = H^{-1}(z, \boldsymbol{\theta})[y(t) - G(z, \boldsymbol{\theta})u(t)]}$$

Sostituiamo l'espressione di $y(t) = G_0(z)u(t) + H_0(z)e(t)$

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



Robustezza dei metodi PEM

Notiamo però che $\eta(t)$ è anche l'errore di predizione ad un passo $\varepsilon_1(t; \boldsymbol{\theta})$ (abbiamo fatto gli stessi passaggi in [Lezione 10 - slide 75](#)), per cui possiamo scrivere

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

Questa espressione connette l'**errore di modellazione** con l'**errore di predizione**, mettendo in risalto che $\varepsilon_1(t; \boldsymbol{\theta})$ dipende linearmente da $u(t)$ e da $e(t)$.

In particolare, l'influenza di $u(t)$ su $\varepsilon_1(t; \boldsymbol{\theta})$ è dettata da $\Delta G(z, \boldsymbol{\theta})$



Robustezza dei metodi PEM

Il vettore di parametri $\boldsymbol{\theta}$ è stimato tramite un metodo PEM, ovvero minimizzando la funzione di costo $J_N(\boldsymbol{\theta})$

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

Asintoticamente, supponendo l'ergodicità di $u(t)$ e $y(t)$, abbiamo che

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

Possiamo **interpretare in frequenza** la funzione di costo $\bar{J}(\boldsymbol{\theta})$ come

$$\bar{J}(\boldsymbol{\theta}) = \mathbb{E}[\varepsilon_1(t, \boldsymbol{\theta})^2] = \gamma_{\varepsilon_1 \varepsilon_1}(0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Gamma_{\varepsilon_1 \varepsilon_1}(\omega) d\omega$$



Robustezza dei metodi PEM

Ipotizzando ora che $u(t) \perp e(t)$, che non è un'ipotesi restrittiva quando il sistema è in **anello aperto**, abbiamo che $\Gamma_{\varepsilon_1 \varepsilon_1}(\omega)$ è la **somma dei contributi** dati da $\Gamma_{uu}(\omega)$ e $\Gamma_{ee}(\omega)$

$$\Gamma_{\varepsilon_1 \varepsilon_1}(\omega) = \left| \frac{\Delta G(e^{j\omega}, \boldsymbol{\theta})}{H(e^{j\omega}, \boldsymbol{\theta})} \right|^2 \Gamma_{uu}(\omega) + \left| \frac{H_0(e^{j\omega})}{H(e^{j\omega}, \boldsymbol{\theta})} \right|^2 \Gamma_{ee}(\omega)$$

Per l'**ipotesi di lavoro 1** trascuriamo il secondo termine, ottenendo

$$\Gamma_{\varepsilon_1 \varepsilon_1}(\omega) \approx \left| \frac{\Delta G(e^{j\omega}, \boldsymbol{\theta})}{H(e^{j\omega}, \boldsymbol{\theta})} \right|^2 \Gamma_{uu}(\omega)$$

Sostituendo $\Gamma_{\varepsilon_1 \varepsilon_1}(\omega) \approx \left| \frac{\Delta G(e^{j\omega}, \boldsymbol{\theta})}{H(e^{j\omega}, \boldsymbol{\theta})} \right|^2 \Gamma_{uu}(\omega)$ nell'espressione $\bar{J}(\boldsymbol{\theta}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Gamma_{\varepsilon_1 \varepsilon_1}(\omega) d\omega$ si ha:



Robustezza dei metodi PEM

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

- $|\Delta G(e^{j\omega}, \boldsymbol{\theta})|^2$ rappresenta **l'errore di stima della risposta in frequenza**, per ogni pulsazione ω
- $g(\omega, \boldsymbol{\theta})$ è un **peso in frequenza**. Per esempio, $g(\omega, \boldsymbol{\theta}) = \text{costante } \forall \omega$, allora gli errori alle diverse frequenze sono pesati allo stesso modo. Oppure, se $g(\omega, \boldsymbol{\theta}) = 0$ se $\omega > \bar{\omega}$, allora solo gli errori nella banda di frequenze $[0, \bar{\omega}]$ contribuiranno alla funzione di costo



Robustezza dei metodi PEM

Se consideriamo un **rumore bianco** come ingresso, abbiamo che

$$g(\omega, \theta) = \frac{\Gamma_{uu}(\omega)}{|H(e^{j\omega}, \theta)|^2} = \frac{\lambda^2}{|H(e^{j\omega}, \theta)|^2}$$

ovvero la pesatura in frequenza dipende solo dal modello del rumore $H(e^{j\omega}, \theta)$. Se facciamo l'ulteriore ipotesi che il **modello del rumore sia una funzione fissata** $H^*(e^{j\omega})$:

$$g(\omega) = \frac{\lambda^2}{|H^*(e^{j\omega})|^2}$$

Notiamo che $g(\omega)$ non dipende più da θ

cioè **peso di più le frequenze dove l'entità del rumore è più bassa** (che vuole dire che do più importanza a dati «più affidabili»)



Robustezza dei metodi PEM

Osservazioni

- Se $G_0(z) \in G(z, \theta)$, allora $\bar{J}(\bar{\theta}) = 0$ e la funzione peso non riveste alcun ruolo
- Usando il **rumore bianco**, non posso «focalizzare gli sforzi» dell'identificazione in una determinata **banda di frequenze** di interesse (ad esempio, spesso per il controllo mi basta un modello buono solo fino alla ω_c)
- Potrei quindi scegliere un ingresso $u(t)$ diverso dal rumore bianco (o similari), e quindi avere un $\Gamma_{uu}(\omega)$ diverso, per pesare di più alcune frequenze a discapito di altre

Purtroppo, non è sempre possibile scegliere l'ingresso a piacere. È però comunque possibile influenzare la pesatura in frequenza tramite **prefiltraggio dei dati**



Prefiltraggio

Il **prefiltraggio** consiste nel **filtrare sia l'ingresso che l'uscita** tramite un filtro $L(z)$ **asintoticamente stabile**

$$u_F(t) = L(z)u(t)$$

$$y_F(t) = L(z)y(t)$$

Tale operazione **non modifica la relazione ingresso-uscita**, anche se modifica la densità spettrale di potenza del rumore. Infatti:

$$L(z)y(t) = G_0(z)L(z)u(t) + H_0(z)L(z)e(t)$$



Prefiltraggio

Filtrare l'ingresso $u(t)$ e l'uscita $y(t)$ equivale a filtrare l'errore di predizione $\varepsilon_1(t; \boldsymbol{\theta})$, ottenendo un **errore filtrato** $\varepsilon_F(t; \boldsymbol{\theta})$, infatti:

$$\varepsilon_F(t; \boldsymbol{\theta}) = L(z)\varepsilon_1(t; \boldsymbol{\theta}) = \frac{L(z)}{H(z, \boldsymbol{\theta})} [y(t) - G(z, \boldsymbol{\theta})u(t)] = \frac{1}{H(z, \boldsymbol{\theta})} [L(z)y(t) - G(z, \boldsymbol{\theta})L(z)u(t)]$$

La funzione di costo asintotica, in frequenza, diventa quindi:

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

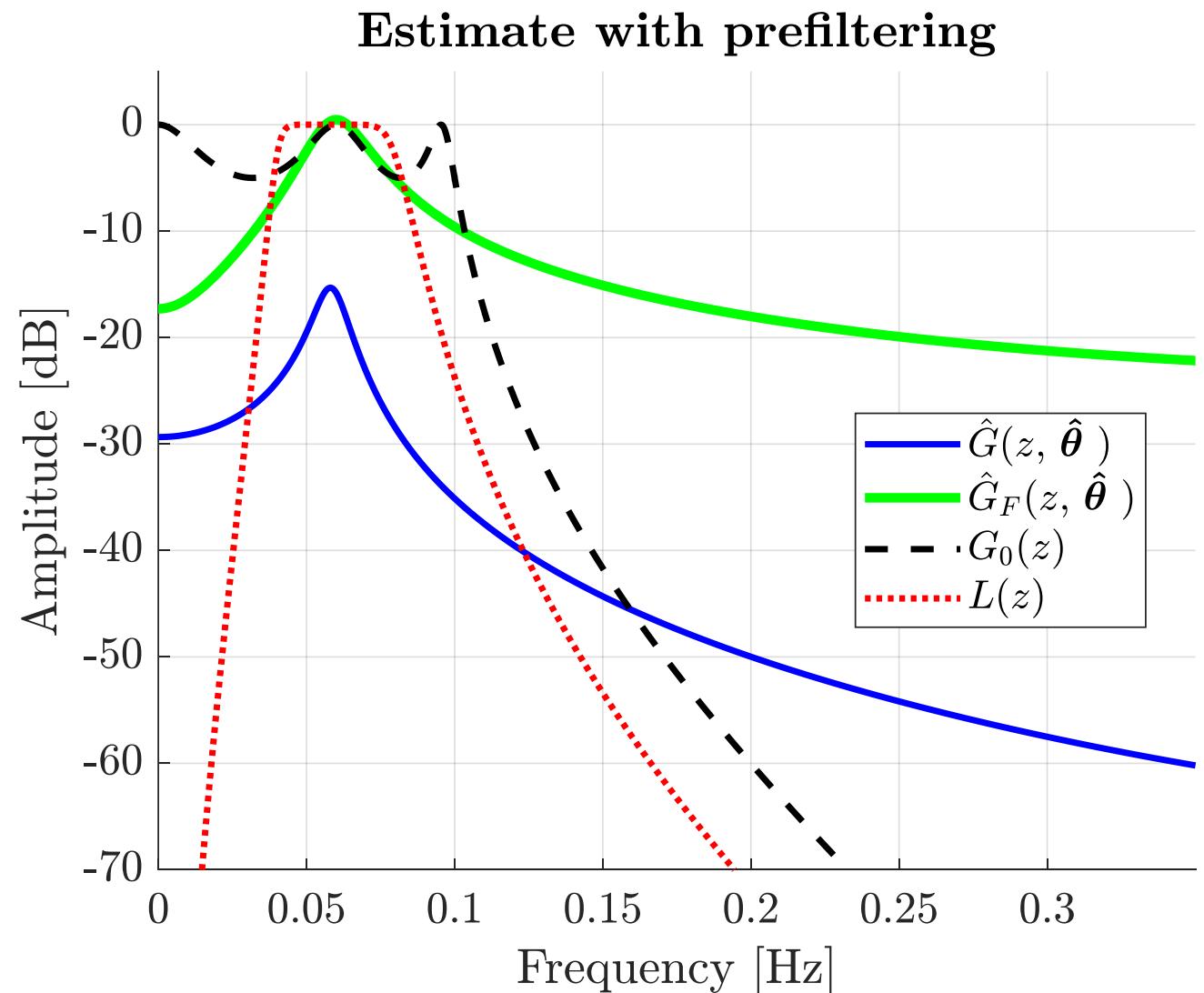


Esempio: prefiltraggio

Misuriamo $N = 5000$ dati da un sistema OE(5, 5, 0) con risposta in frequenza come in figura, e con $\lambda^2 = 0.01$

Identifichiamo un modello ARX(2, 2, 1), per il quale $S \notin \mathcal{M}(\theta)$

Progettiamo un filtro $L(z)$ che abbia un modulo alto nella banda di frequenze $[0.04, 0.08]$ Hz



Outline

1. Analisi asintotica dei metodi PEM
2. Identificabilità dei modelli e persistente eccitazione
3. Valutazione dell'incertezza della stima PEM
4. Robustezza dei metodi PEM e prefiltraggio
- 5. Empirical Transfer Function Estimate (ETFE)**



Multiseno e stima nonparametrica

Dato che il multiseno è un segnale **periodico**, è possibile utilizzare la DFT (tramite l'algoritmo FFT) per calcolare lo spettro di uno (o più) periodi del segnale, ottenendo uno spettro **senza leakage**

Questo viene molto comodo in quanto è possibile ottenere una **stima nonparametrica** iniziale della $G(z)$, **prima di effettuare la stima parametrica** con metodi PEM

Supponiamo di dividere i segnali $u(t)$ e $y(t)$, entrambi multiseni, in sotto-sequenze $u^{[p]}, y^{[p]}$ che contengono i diversi $p = 1, \dots, P$ periodi. Definiamo:

$$\breve{U}^{[p]} = \text{DFT}(u^{[p]}) \quad \text{DFT dell'ingresso}$$

$$\breve{Y}^{[p]} = \text{DFT}(y^{[p]}) \quad \text{DFT dell'uscita}$$



Empirical Transfer Function Estimate (ETFE)

Possiamo ottenere una **stima nonparametrica** della $G(z)$ nelle **griglia di frequenze** definite dalla DFT $k = \text{bin:bin:} f_s/2$, come

$$\hat{U}(k) = \frac{1}{P} \sum_{p=1}^P \check{U}^{[p]}(k)$$



$$\hat{Y}(k) = \frac{1}{P} \sum_{p=1}^P \check{Y}^{[p]}(k)$$

$$\hat{G}(k) = \frac{\hat{Y}(k)}{\hat{U}(k)}$$

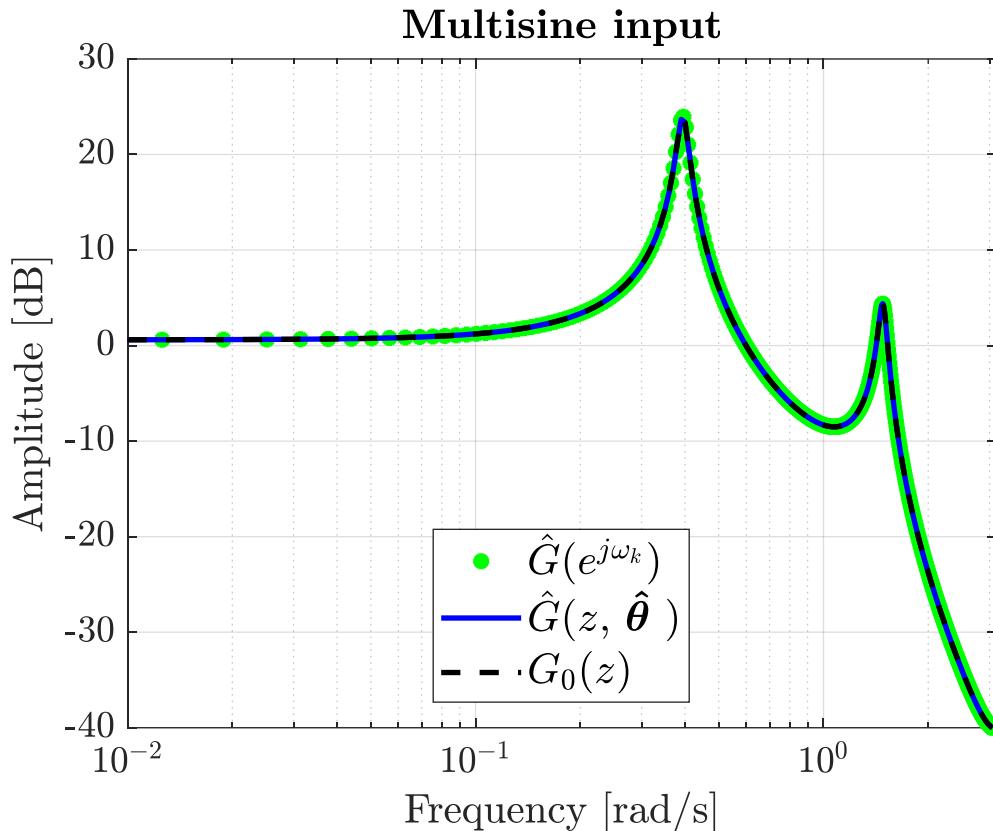
Empirical Transfer Function Estimate (ETFE)

$\hat{G}(k)$ è un **numero complesso** che rappresenta la **risposta in frequenza** del sistema alla frequenza k . È poi possibile calcolare **modulo** e **fase** del numero complesso a diverse frequenze k per tracciare i **diagrammi di Bode**

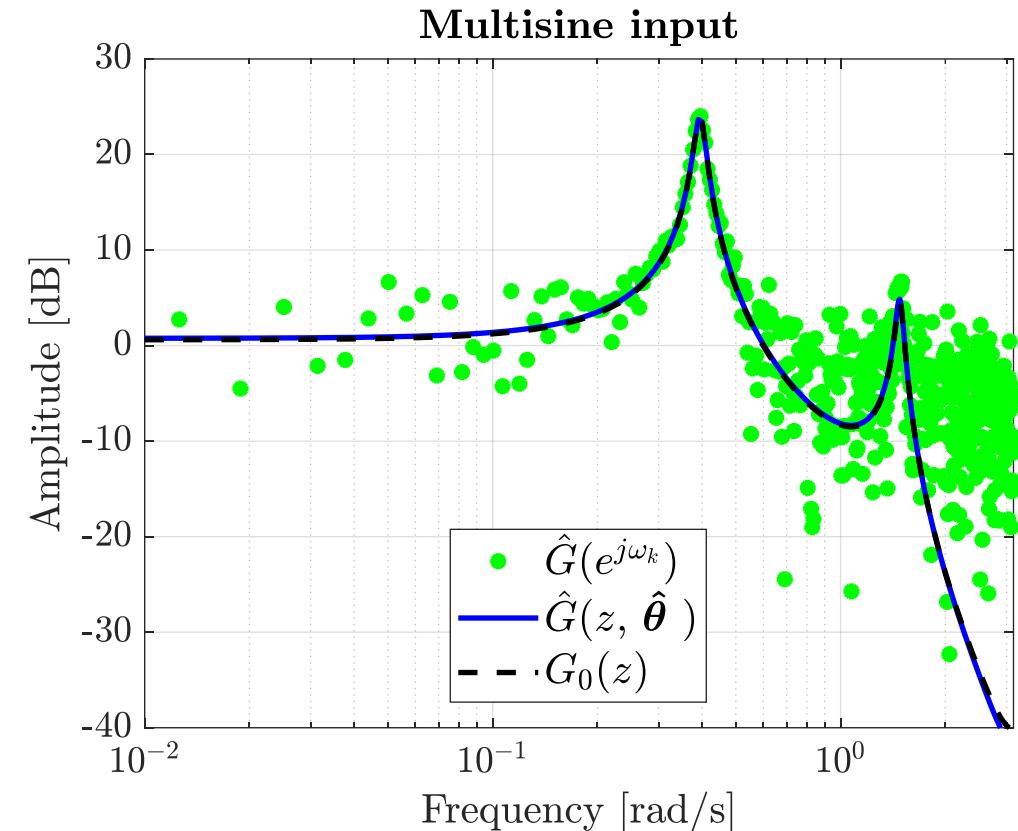


Esempio: ETFE tramite multiseno

Senza rumore sull'uscita



Con rumore sull'uscita



Senza rumore, la stima è perfetta perché **non c'è leakage** indotto dalla FFT, grazie alla **periodicità** di $u(t)$ e $y(t)$



Empirical Transfer Function Estimate (ETFE)

La **stima nonparametrica ETFE** è usata per ottenere una **prima indicazione** sulle caratteristiche del sistema, in particolar modo sulla $G(z)$. **Non permette** di ottenere informazioni sullo **spettro del disturbo**

La stima ottenuta ci dà inoltre un'indicazione sulla **banda di frequenza di interesse** del sistema, sia per l'**identificazione parametrica** (es. usando poi il **prefiltraggio**) sia per l'applicazione del modello per fini di **progettazione del controllo**

La ETFE può essere effettuata anche con **segnali non periodici** (In Matlab **etfe**): in questo caso, è utile usare delle **finestrature** (es. Hanning window) per ridurre il leakage (a discapito della risoluzione in frequenza), e delle procedure di **spectral averaging**



Stima nonparametrica tramite spectral analysis

In Matlab esiste anche il comando **spa** oppure **spafdr** che permette la stima della funzione di trasferimento $G_0(z)$ e $H_0(z)$ tramite **analisi spettrale**

Dato $y(t) = G_0(z)u(t) + v(t)$, con $v(t) \perp u(t)$, abbiamo che $\Gamma_{uy}(\omega) = G_0(e^{j\omega})\Gamma_{uu}(\omega)$

Slide 98
Lezione 08

E quindi una stima nonparametrica di $G_0(z)$ si può ottenere come

$$\hat{G}(e^{j\omega}) = \frac{\hat{\Gamma}_{uy}(\omega)}{\hat{\Gamma}_{uu}(\omega)}$$

Inoltre, dato che $\Gamma_{yy}(\omega) = |G_0(e^{j\omega})|^2 \cdot \Gamma_{uu}(\omega) + \Gamma_{vv}(\omega)$, una stima di $\Gamma_{vv}(\omega)$ è

$$\hat{\Gamma}_{vv}(\omega) = \hat{\Gamma}_{yy}(\omega) - \frac{|\hat{\Gamma}_{uy}(\omega)|^2}{\hat{\Gamma}_{uu}(\omega)}$$



Identificazione in frequenza

Una volta che si è ottenuta la ETFE, è anche possibile «fittare i punti in frequenza» anziché «fittare i dati nel tempo». Abbiamo quindi una vera e propria **identificazione parametrica in frequenza**, per esempio minimizzando

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta} \in \Theta} \frac{1}{F} \sum_{k=1}^F |\hat{G}(e^{j\omega_k}) - G(e^{j\omega_k}; \boldsymbol{\theta})|^2 W_k$$

W_k è un peso in frequenza

L'approccio è particolarmente utile quando:

- $u(t)$ è un **multiseno** e il numero di dati nel dominio del tempo è **grande**
- vi sono **dati da diversi esperimenti**, che devono essere usati assieme
- si vuole identificare modelli a **tempo continuo**





**UNIVERSITÀ
DEGLI STUDI
DI BERGAMO**

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione



IDENTIFICAZIONE DEI MODELLI E ANALISI DEI DATI (IMAD)

Lezione 13: Identificazione – valutazione del modello

Corso di Laurea Magistrale in
INGEGNERIA INFORMATICA

SPEAKER
Prof. Mirko Mazzoleni

PLACE
Università degli Studi di
Bergamo

Syllabus

Parte II: sistemi dinamici

8. Processi stocastici

- 8.1 Processi stocastici stazionari (pss)
- 8.3 Rappresentazione spettrale di un pss
- 8.4 Stimatori campionari media\covarianza
- 8.5 Densità spettrale campionaria

9. Famiglie di modelli a spettro razionale

- 9.1 Modelli per serie temporali (MA, AR, ARMA)
- 9.2 Modelli per sistemi input/output (ARX, ARMAX)

10. Predizione

- 10.1 Filtro passa-tutto

10.2 Forma canonica

10.3 Teorema della fattorizzazione spettrale

10.4 Soluzione al problema della predizione

11. Identificazione

- 11.3 Identificazione di modelli ARX
- 11.4 Identificazione di modelli ARMAX
- 11.5 Metodo di Newton

12. Identificazione: analisi e complementi

- 12.1 Analisi asintotica metodi PEM
- 12.2 Identificabilità dei modelli
- 12.3 Valutazione dell'incertezza di stima

13. Identificazione: valutazione



Parte I: sistemi staticiStima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

✓ Stima parametri popolazione

✓ Stima modello lineare: minimi quadrati

○ **SI assunzioni su ddp dei dati**

✓ Stima massima verosimiglianza parametri popolazione

✓ Stima modello lineare: massiva verosimiglianza

✓ Regressione logistica

• θ variabile casuale○ **SI assunzioni su ddp dei dati**

✓ Stima Bayesiana

Machine learning**Parte II: sistemi dinamici**Stima parametrica $\hat{\theta}$ • θ deterministico○ **NO assunzioni su ddp dei dati**

✓ Modelli lineari di pss

✓ Predizione

✓ Identificazione

✓ Persistente eccitazione

✓ Analisi asintotica metodi PEM

✓ Analisi incertezza stima (numero dati finito)

✓ Valutazione del modello

Outline

1. Scelta della struttura e complessità del modello
2. Validazione o formule di complessità per la scelta della complessità
3. Analisi dei residui
4. Analisi dell'incertezza della stima
5. Simulazione, predizione del modello identificato
6. Confronto con stima nonparametrica
7. Considerazioni pratiche

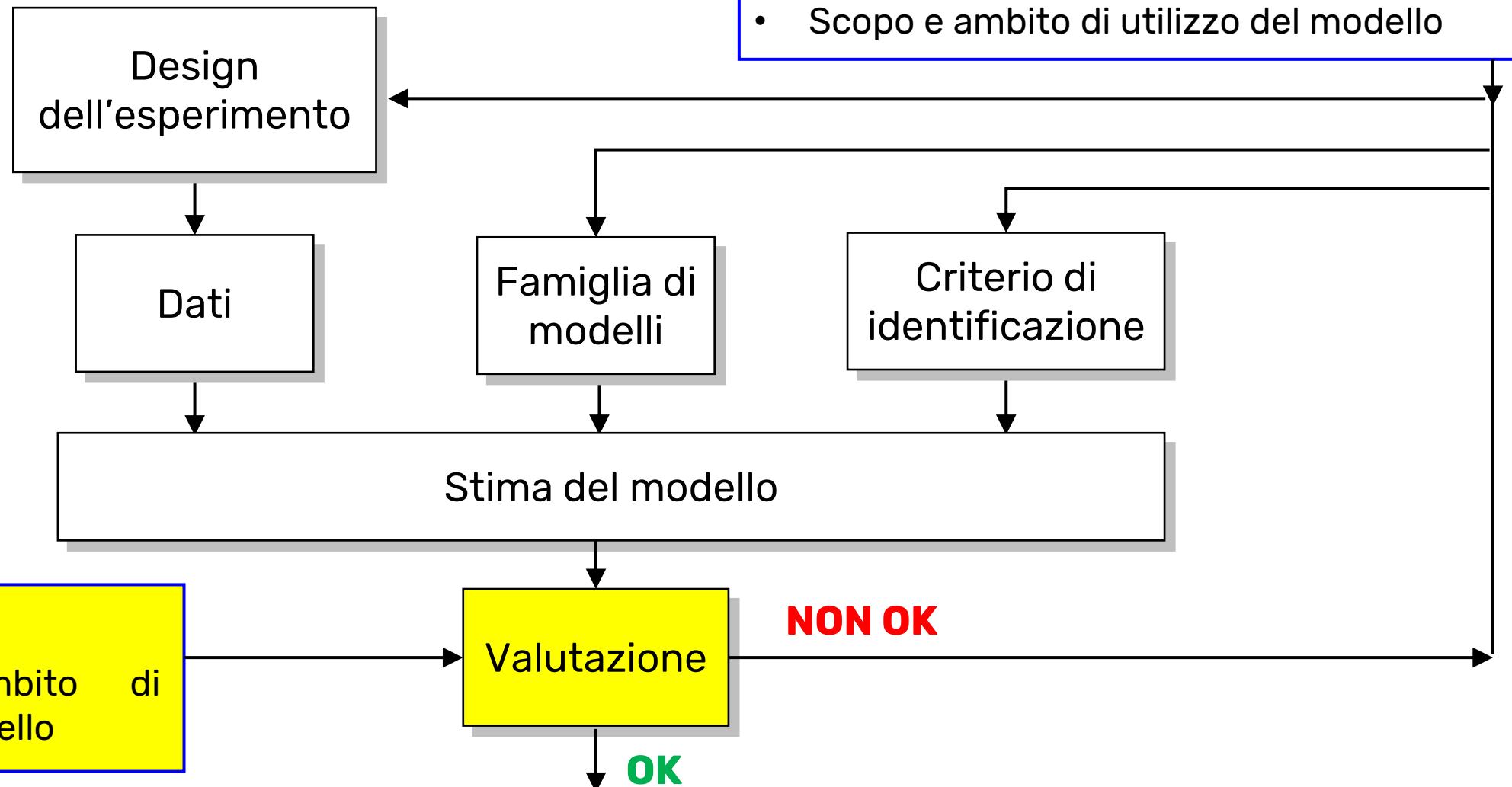


Outline

- 1. Scelta della struttura e complessità del modello**
2. Validazione o formule di complessità per la scelta della complessità
3. Analisi dei residui
4. Analisi dell'incertezza della stima
5. Simulazione, predizione del modello identificato
6. Confronto con stima nonparametrica
7. Considerazioni pratiche



I passi della procedura



Scelta struttura e complessità del modello

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

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

L'obiettivo è quello di trovare un «**buon modello**» ad un «**prezzo ragionevole**»

Nel caso **generale**, un modello è buono se ha poco bias e poca varianza: nel caso **specifico**, un modello deve essere buono per **l'utilizzo** che se ne deve fare



Scelta struttura e complessità del modello

Per «prezzo ragionevole» si intende quanto sforzo è necessario per **identificare** e **utilizzare** il modello:

- Quanto **tempo** ci vuole per trovare la stima e se la stima dipende da inizializzazioni
- Quanto un modello è di **ordine ridotto**, per poter essere utilizzato in ambito real-time

Vi è un tradeoff tra bontà e prezzo del modello. In particolare, due aspetti sono importanti:

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



Considerazioni generali

- In base alla **fisica** del processo che deve essere modellato, è possibile avere informazioni sul **minimo ordine del modello** necessario (e.g. un sistema massa-molla-smorzatore sarà almeno di ordine 2)
- In base alla **fisica** del processo che deve essere modellato, è possibile avere informazioni su come **pre-processare** i segnali a disposizione. Per esempio, potrebbe essere utile elevare al quadrato o fare il logaritmo di certi segnali e poi usare una regressione lineare con le variabili trasformate, stimando un ARX
- Se ho pochi dati, non posso usare un modello di ordine elevato, pena overfitting

Rule-of-thumb: $N \gg 10 \cdot d$



Considerazioni generali

- L'analisi **nonparametrica** tramite ETFE può dare importanti informazioni sull'ordine del modello soprattutto per quanto riguarda la posizione di **risonanze**
- Un modello dovrebbe limitarsi a modellare al massimo **3 decadì in frequenza**:
 - ✓ Per un modello che usa dati campionati ad **alta frequenza**, le dinamiche lente vengono viste come integratori
 - ✓ Per un modello a **bassa frequenza**, le dinamiche veloci vengono viste come relazioni statiche. In questo caso, si può introdurre un termine senza delay $b_o u(t)$
 - ✓ Se necessario, costruire più modelli con dati campionati a frequenze diverse



Considerazioni generali

- Una volta stimati i parametri di un modello, guardiamo le loro **deviazioni standard**. Se la deviazione standard è tale da **includere lo zero**, allora quel parametro potrebbe **non essere significativo**
 - ✓ Questa analisi permette di scegliere il **ritardo puro** k più opportuno, identificando diversi modelli (di solito ARX) con diversi valori di k , e scegliendo quello per cui tutti i coefficienti $B(z)$ sono significativi
- Per rendersi conto del **ritardo** del sistema e della sua **linearità** (e anche della dinamica dominante) è possibile effettuare una **risposta allo scalino**, con diverse ampiezze di scalino



Considerazioni generali

- La **bontà** di un modello può essere valutata:
 - ✓ Analizzando i residui (cioè gli errori di predizione a un passo), meglio con dati di **validazione**
 - ✓ Confrontando l'uscita **simulata** o **predetta** con l'uscita **misurata**, su dati di **validazione**, e calcolandone un indicatore di «**FIT**»
 - ✓ Rappresentando un grafico di **poli-zeri** con rispettive **bande di confidenza**, per vedere se vi sono cancellazioni (e quindi se si può semplificare il modello)
 - ✓ Rappresentando i **diagrammi di Bode** di diversi modelli e confrontandoli con la ETFE



Outline

1. Scelta della struttura e complessità del modello
- 2. Validazione o formule di complessità per la scelta della complessità**
3. Analisi dei residui
4. Analisi dell'incertezza della stima
5. Simulazione, predizione del modello identificato
6. Confronto con stima nonparametrica
7. Considerazioni pratiche



Validazione o formule di complessità

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

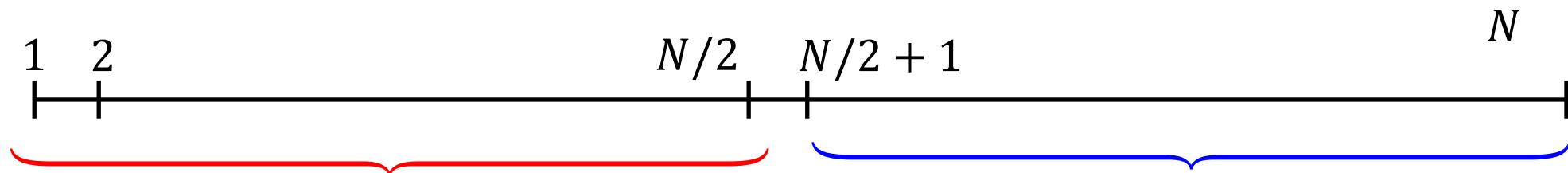
Un metodo semplice ma efficace consiste nell'identificare un insieme di modelli di diversa complessità utilizzando un dataset di **identificazione**, e confrontarne la bontà (e.g. calcolando il valore di $J(\hat{\theta}_N)$) su un dataset di **validazione**

Il problema è **multidimensionale**: per esempio, se usassimo un modello ARMAX dovremmo scegliere il valore di n_a, n_b, n_c . Per **semplicità**, si pone $n_a = n_b = n_c \equiv m$, facendo quindi variare solo il valore del parametro m . In questo caso avremmo $d = 3 \cdot m$



Validazione

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



Dati di identificazione

$$\begin{aligned}\mathcal{D}_{\text{train}} &= \{y(1), y(2), \dots, y(N/2)\} \\ &\quad \{u(1), u(2), \dots, u(N/2)\}\end{aligned}$$

Dati di validazione

$$\begin{aligned}\mathcal{D}_{\text{val}} &= \{y(N/2 + 1), \dots, y(N)\} \\ &\quad \{u(N/2 + 1), \dots, u(N)\}\end{aligned}$$

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



Validazione

A differenza del caso statico, con i sistemi dinamici non è possibile «estrarre» i dati di identificazione e di validazione in modo casuale dal dataset completo, perché **romperei la causalità temporale** dei dati!

Anche in questo caso dinamico, la validazione è una procedura che da risultati molto buoni, ma richiedere **tanti dati**

In alternativa, se i dati sono pochi si possono usare le **formule di complessità ottima** già viste in [Lezione 06](#)



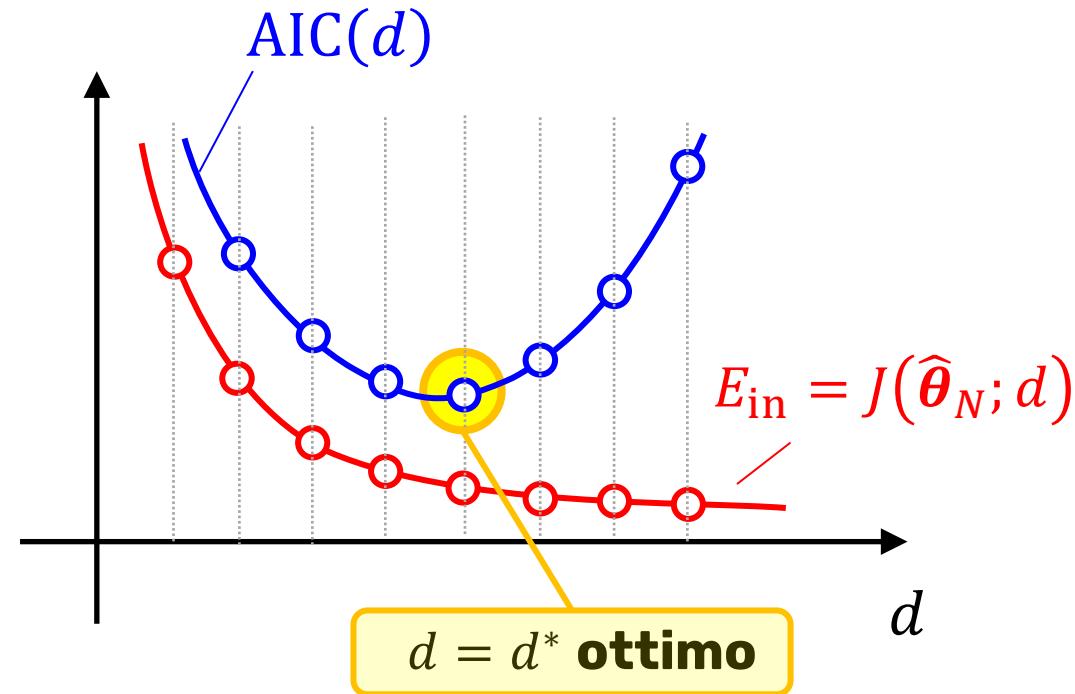
Formule di complessità

Akaike Information Criterion (AIC)

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

Final prediction error (FPE)

$$\text{FPE}(d) = \frac{N+d}{N-d} \cdot J(\hat{\theta}_N; d)$$



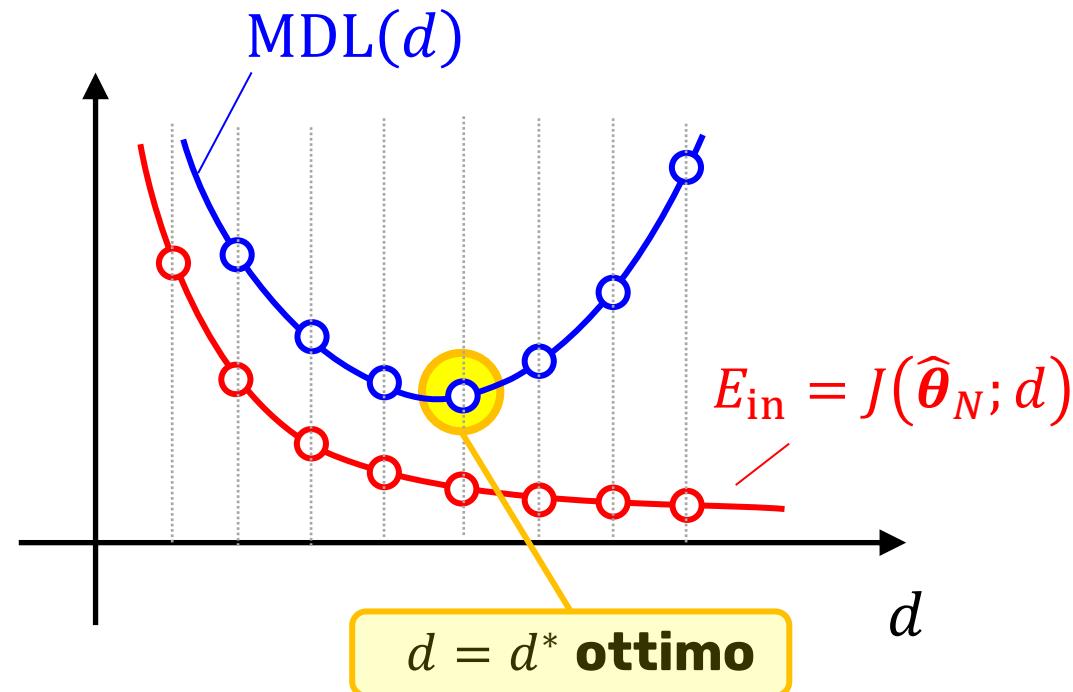
Se $d \ll N$, il criterio FPE è equivalente al criterio AIC



Formule di complessità

Minimum Description Length (MDL)

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



In Matlab:

- `v = arxstruc(data_ident, data_val, NN)` Calcola la funzione di costo per ARX di struttura diversa definita in `NN`
- `order = selstruc(v, 'AIC');` Seleziona l'ordine migliore del modello ARX usando il criterio selezionato



Outline

1. Scelta della struttura e complessità del modello
2. Validazione o formule di complessità per la scelta della complessità
- 3. Analisi dei residui**
4. Analisi dell'incertezza della stima
5. Simulazione, predizione del modello identificato
6. Confronto con stima nonparametrica
7. Considerazioni pratiche



Interpretazione in frequenza del costo PEM

Ricordiamo che (Lezione 10 - slide 75)

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

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

Sommiamo e sottraiamo $e(t)$

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

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

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

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



Interpretazione in frequenza del costo PEM

La stima asintotica può quindi essere ottenuta come $\bar{\theta} = \arg \min_{\theta \in \Theta} \bar{J}(\theta) = \arg \min_{\theta \in \Theta} \mathbb{E}[\varepsilon_1(t; \theta)^2]$

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

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

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



Interpretazione in frequenza del costo PEM

Con il **prefiltraggio** tramite filtro $L(z)$ dei dati

$$u_F(t) = L(z)u(t)$$

$$y_F(t) = L(z)y(t)$$

la funzione di costo asintotica diventa

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



Analisi dei residui

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

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

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

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

Il caso $\mathcal{S} \notin \mathcal{M}(\theta)$ con $G_0 \in \mathcal{G}(\theta)$ è di **molto interesse nella pratica**, in cui vogliamo che $G(z, \hat{\theta}_N) \rightarrow G_0(z)$ anche se il modello dell'errore è sbagliato



Analisi dei residui

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

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

con $e(t) \sim \text{WN}(0, \lambda^2)$

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

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



Analisi dei residui

Avendo definito l'obiettivo come in precedenza, possiamo incorrere in tre situazioni:

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

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

Studiamo le tre situazioni singolarmente



Analisi dei residui: Situazione A

Supponiamo di osservare:

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

Questa situazione accade quando

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

Ovvvero **se e solo se** $G(z, \bar{\theta}) = G_0(z)$ e $H(z, \bar{\theta}) = H_0(z)$

Questo avviene **se e solo se** $\mathcal{S} \in \mathcal{M}(\theta)$, come dimostrato nella [Lezione 12](#)



Analisi dei residui: Situazione B

Supponiamo di osservare:

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

$$\gamma_{\varepsilon u}(\tau) = 0 \quad \forall \tau$$

Questa situazione accade quando

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

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

Questo avviene **se e solo se** $\mathcal{S} \notin \mathcal{M}(\theta)$ con $G_0(z) \in \mathcal{G}(\theta)$ per $\mathcal{M}(\theta)$ **OE, BJ, FIR**



Analisi dei residui: Situazione B

Infatti, se $\mathcal{M}(\theta)$ è **OE**, **BJ** oppure **FIR**, è possibile **parametrizzare in modo indipendente** $G(z, \boldsymbol{\eta})$ e $H(z, \boldsymbol{\xi})$, con $\theta = [\boldsymbol{\eta}^\top \ \boldsymbol{\xi}^\top]^\top$

La funzione di costo PEM diventa

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

Il vettore $\bar{\theta}$ che minimizza questa cifra di merito è $\bar{\theta} = \begin{bmatrix} \bar{\boldsymbol{\eta}} \\ \bar{\boldsymbol{\xi}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\eta}_0 \\ \boldsymbol{\xi} \end{bmatrix}$

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



Analisi dei residui: Situazione B

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

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

Per cui, anche se esistesse θ_0 t.c. $G(z, \theta_0) = G_0(z)$, tale vettore minimizza solo il termine $|G_0(e^{j\omega}) - G(e^{j\omega}, \theta)|^2$, ma non $|H_0(e^{j\omega}) - H(e^{j\omega}, \theta)|^2$, poiché $H(z, \theta_0) \neq H_0(z)$

Ne consegue che $\bar{\theta} \neq \theta_0$ e quindi $G_0(z)$ **non viene stimata in modo corretto** anche se $G_0(z) \in \mathcal{G}(\theta)$. Si può comunque arrivare ad una buona approssimazione aumentando l'ordine del modello



Analisi dei residui: Situazione C

Supponiamo di osservare:

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

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

Questa situazione accade quando

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

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

Questo avviene **se e solo se**

$$\begin{cases} \mathcal{S} \notin \mathcal{M}(\theta) \text{ con } G_0(z) \in \mathcal{G}(\theta) \text{ per } \mathcal{M}(\theta) \text{ ARX, ARMAX} \\ \mathcal{S} \notin \mathcal{M}(\theta) \text{ con } G_0(z) \notin \mathcal{G}(\theta) \end{cases}$$



Conclusioni analisi residui

In base ai risultati nel caso asintotico, possiamo concludere che:

1) $\mathcal{M}(\theta)$ è OE, BJ o FIR

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



Conclusioni analisi residui

In base ai risultati nel caso asintotico, possiamo concludere che:

2) $\mathcal{M}(\theta)$ è ARX, ARMAX

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

Nella situazione C non riusciamo a capire se $G_0(z) \in \mathcal{G}(\theta)$ oppure $G_0(z) \notin \mathcal{G}(\theta)$



Conclusioni analisi residui

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



Conclusioni analisi residui

La procedura vista ora basata su un test statistico dei residui (nel caso di N finito) serve a validare l'ipotesi che $\mathcal{S} \in \mathcal{M}(\theta)$ sulla base dei **dati disponibili**

Una validazione della struttura che si conclude con un successo **non garantisce** che $G(z, \hat{\theta}_N)$ e $H(z, \hat{\theta}_N)$ siano **buone stime** di $G_0(z)$ e $H_0(z)$.

È necessario controllare anche la **varianza delle stime** (sia dei parametri sia delle funzioni di trasferimento)



Esempio: analisi residui

Applichiamo un ingresso $u(t)$ a scalino al sistema ignoto \mathcal{S} e osserviamo la risposta

$$\mathcal{S}: y(t) = \frac{B(z)}{F(z)} u(t-3) + \frac{C(z)}{D(z)} e(t) \quad e(t) \sim WN(0,0.09)$$

Il sistema \mathcal{S} è un Box-Jenkins con

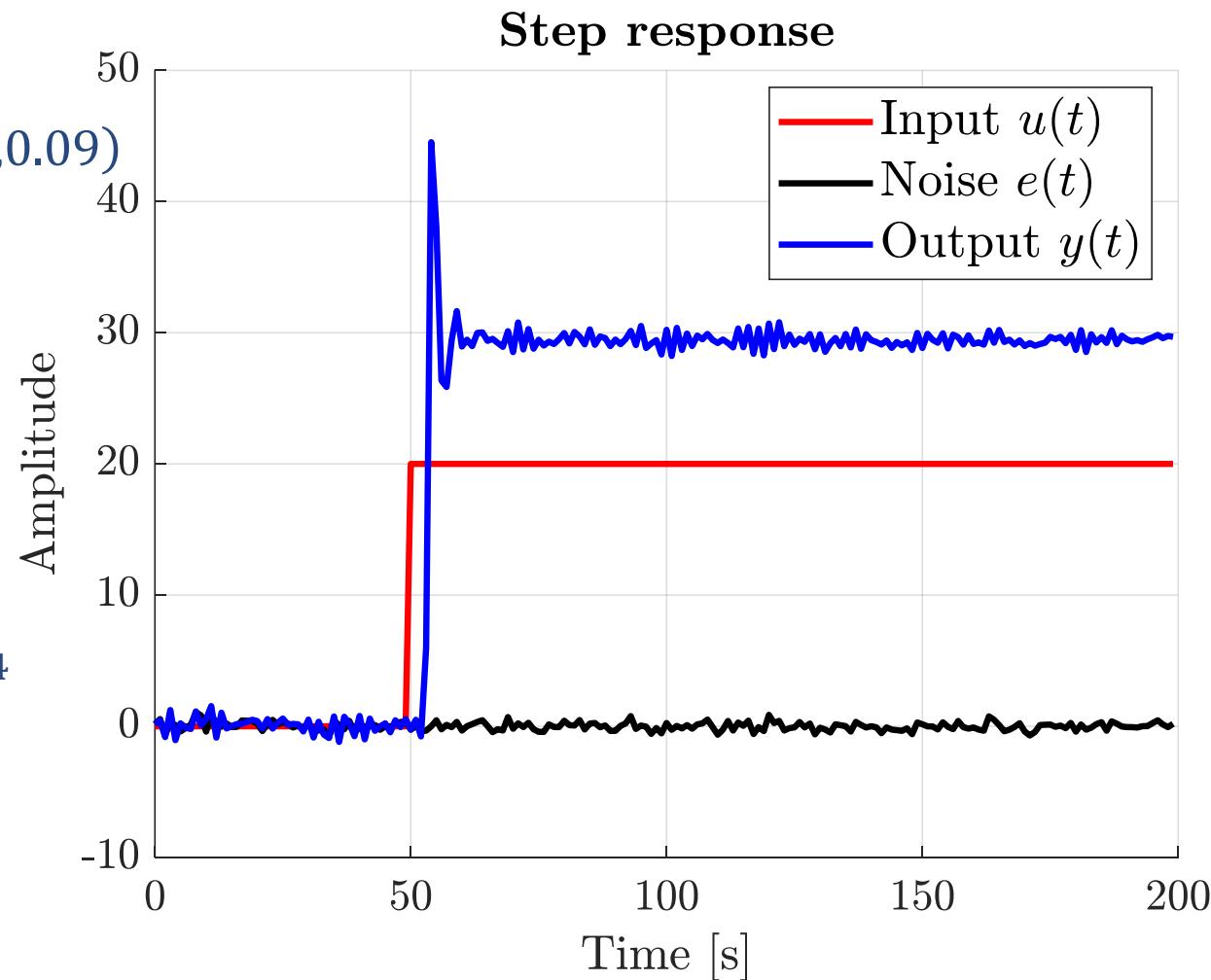
$$n_b = 2, n_c = 4, n_d = 4, n_f = 3, k = 3$$

$$B(z) = 0.3 + 2z^{-1} + 0.13z^{-2}$$

$$C(z) = 1 + 0.3z^{-1} + 0.72z^{-2} + 0.76z^{-3} + 0.05z^{-4}$$

$$D(z) = 1 + 0.9z^{-1} + 0.09z^{-2} + 0.1z^{-3} + 0.2z^{-4}$$

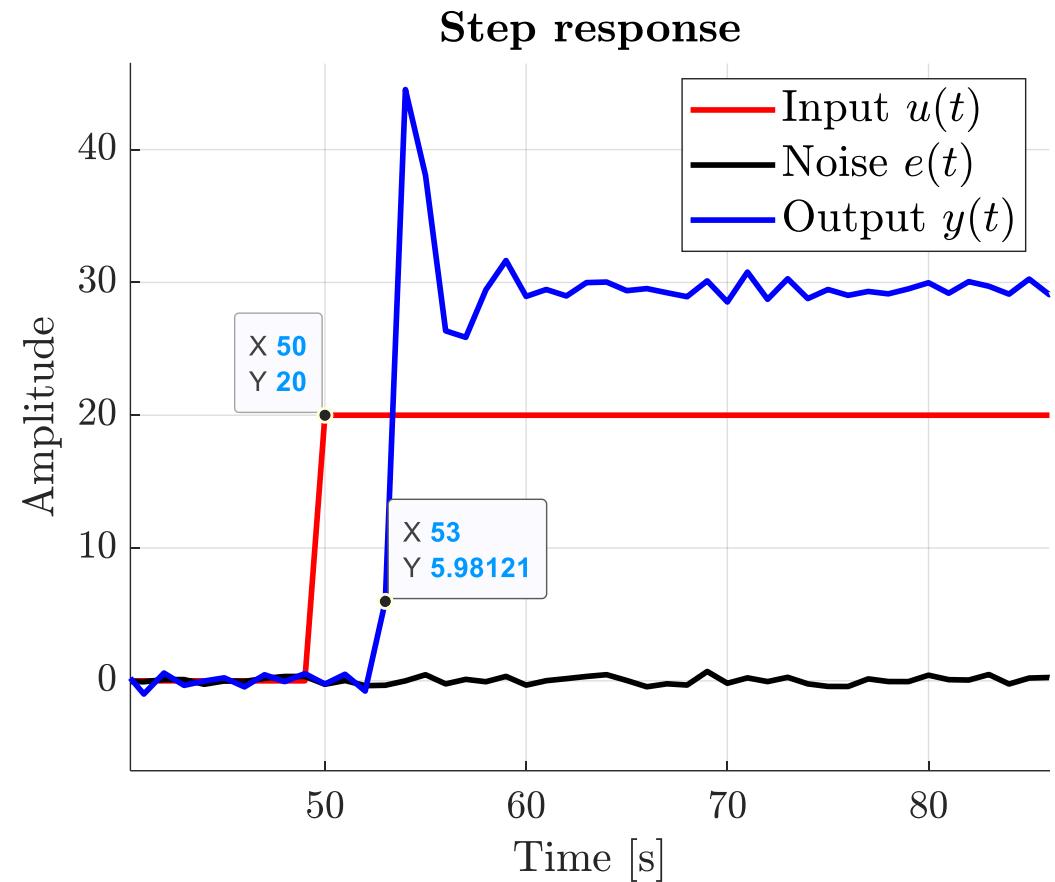
$$F(z) = 1 + 0.2z^{-1} + 0.3z^{-2} + 0.15z^{-3}$$



Esempio: analisi residui

Dalla risposta allo scalino osserviamo che il ritardo puro vale $k = 3$

Osserviamo inoltre che il sistema sembra essere «semplice», non avendo una risposta allo scalino particolarmente «strana»



Collezioniamo ora $N = 5000$ dati usando un ingresso $u(t) \sim WN(0, 100)$



Esempio: analisi residui

Primo tentativo

Scegliamo un modello BJ con

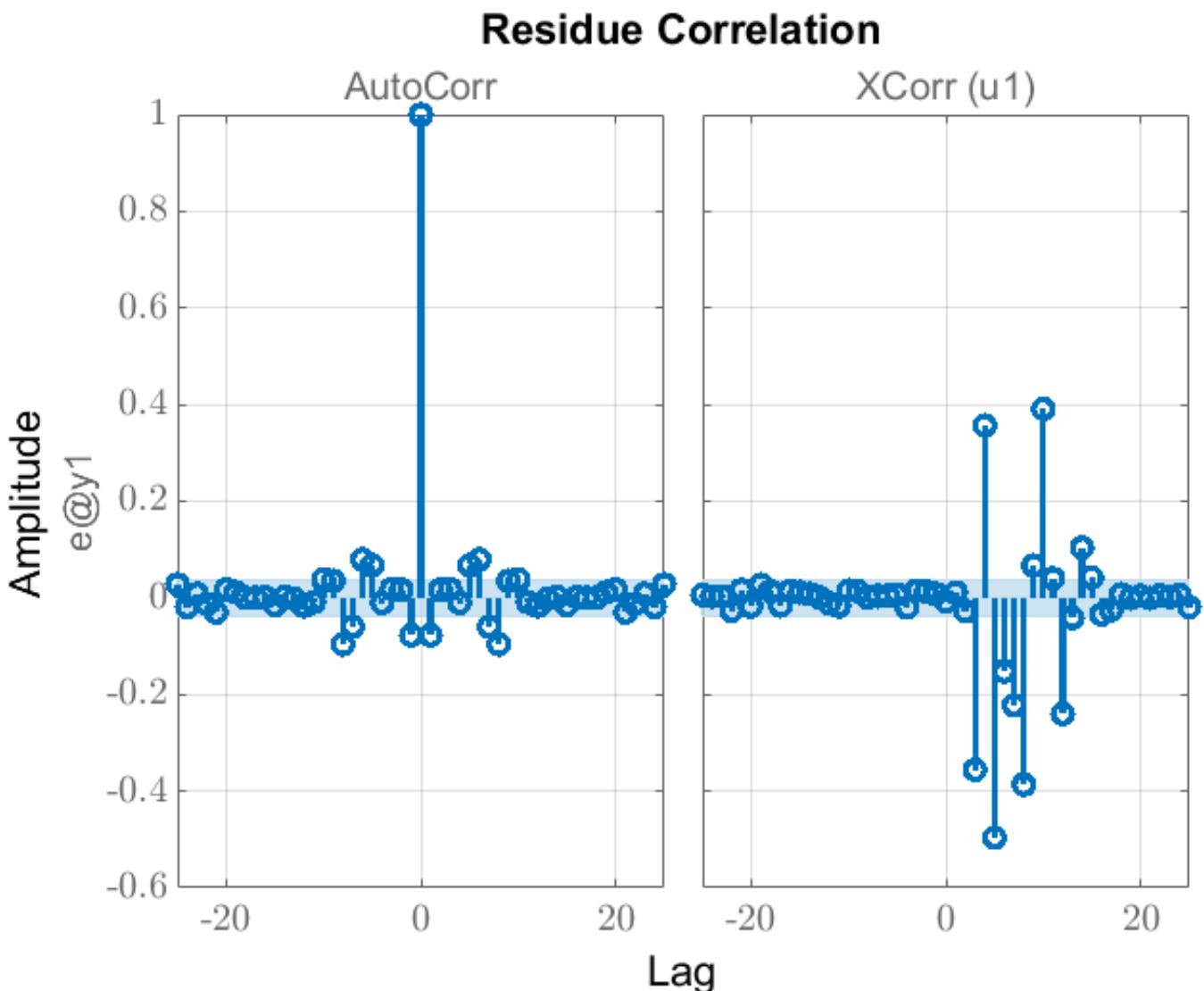
$$n_b = 1, n_c = 2, n_d = 2, n_f = 2, k = 3$$

Osserviamo che:

$$\gamma_{\varepsilon\varepsilon}(\tau) \neq \lambda^2 \cdot \delta(\tau) \quad \exists \tau \text{ t.c. } \gamma_{\varepsilon u}(\tau) \neq 0$$

Conclusione:

$$\mathcal{S} \notin \mathcal{M}(\theta) \text{ con } G_0(z) \notin \mathcal{G}(\theta)$$



Esempio: analisi residui

Secondo tentativo

Scegliamo un modello BJ con

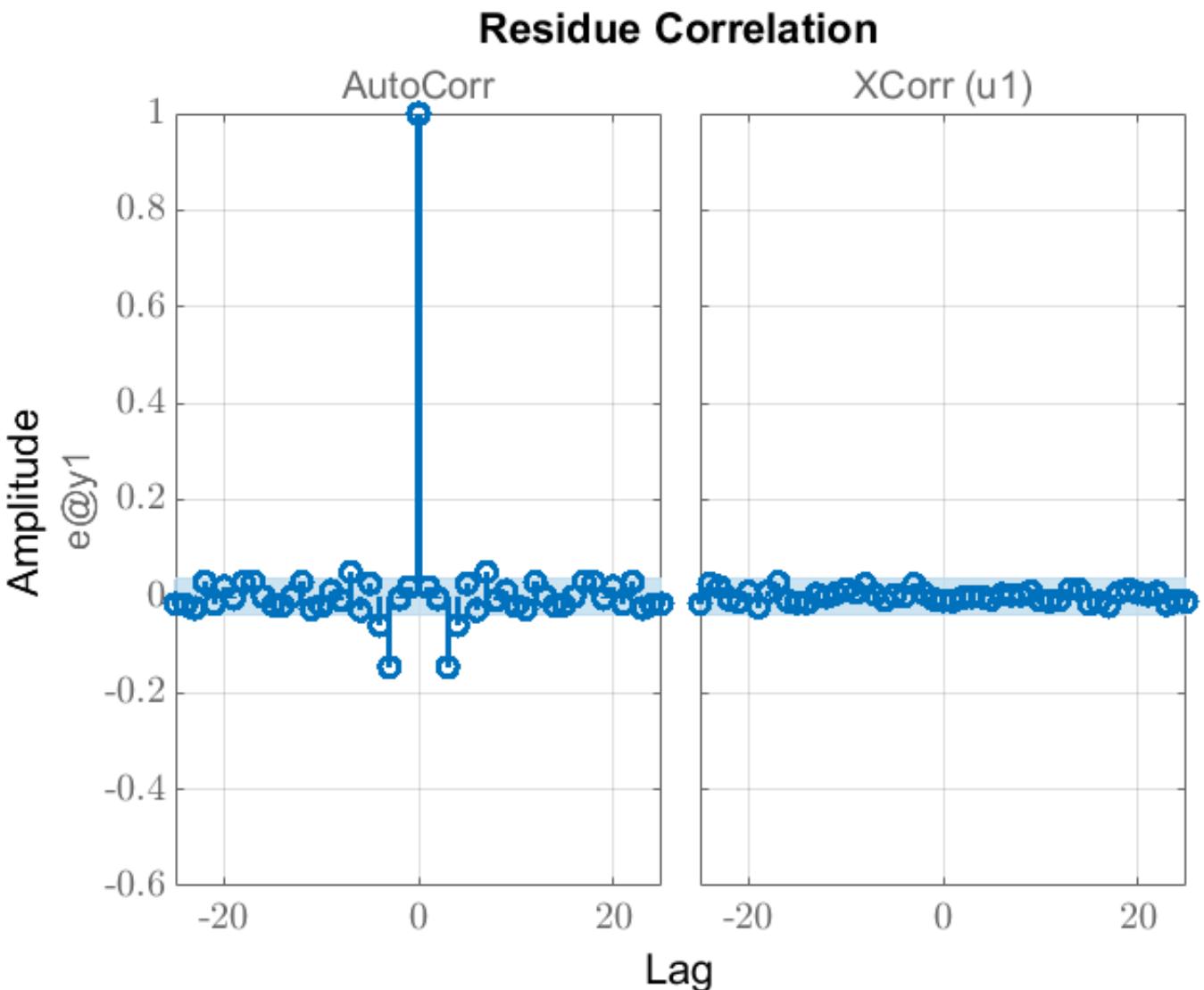
$$n_b = 2, n_c = 3, n_d = 3, n_f = 3, k = 3$$

Osserviamo che:

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

Conclusione:

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



Esempio: analisi residui

Terzo tentativo

Scegliamo un modello BJ con

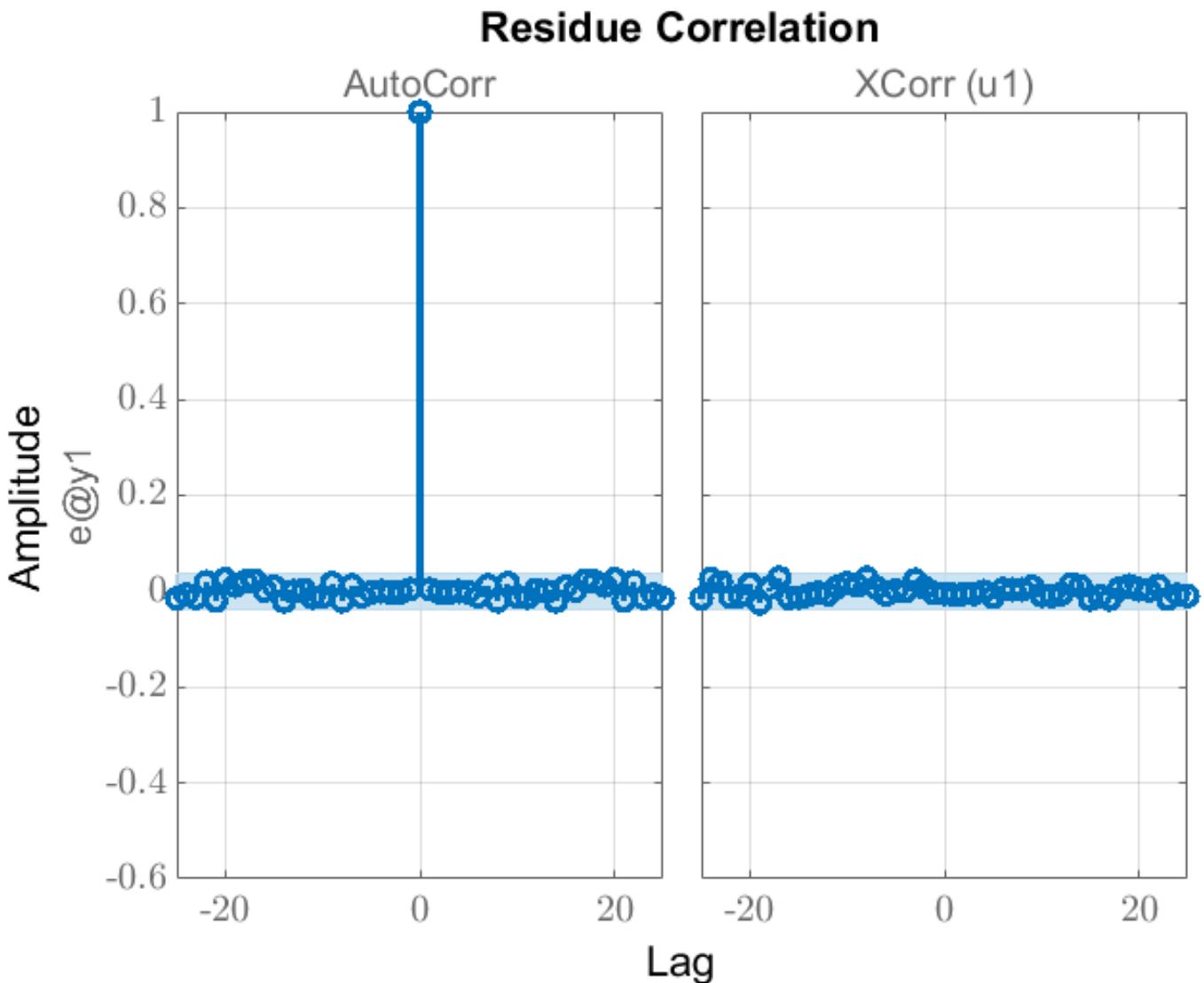
$$n_b = 2, n_c = 4, n_d = 4, n_f = 3, k = 3$$

Osserviamo che:

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

Conclusione:

$$\mathcal{S} \in \mathcal{M}(\theta) \text{ con } G_0(z) \in \mathcal{G}(\theta)$$



Esempio: identificazione trasmissione meccanica

Si consideri il sistema \mathcal{S} espresso tramite la famiglia di modelli **OE**

$$\mathcal{S}: y(t) = \frac{0.10276 + 0.18123z^{-1}}{1 - 1.99185z^{-1} + 2.20265z^{-2} - 1.84083z^{-3} + 0.89413z^{-4}} u(t-3) + e(t)$$

Identifichiamo il sistema con $u(t) \sim WN(0, \lambda^2)$, SNR = 15, $N = 5000$, usando:

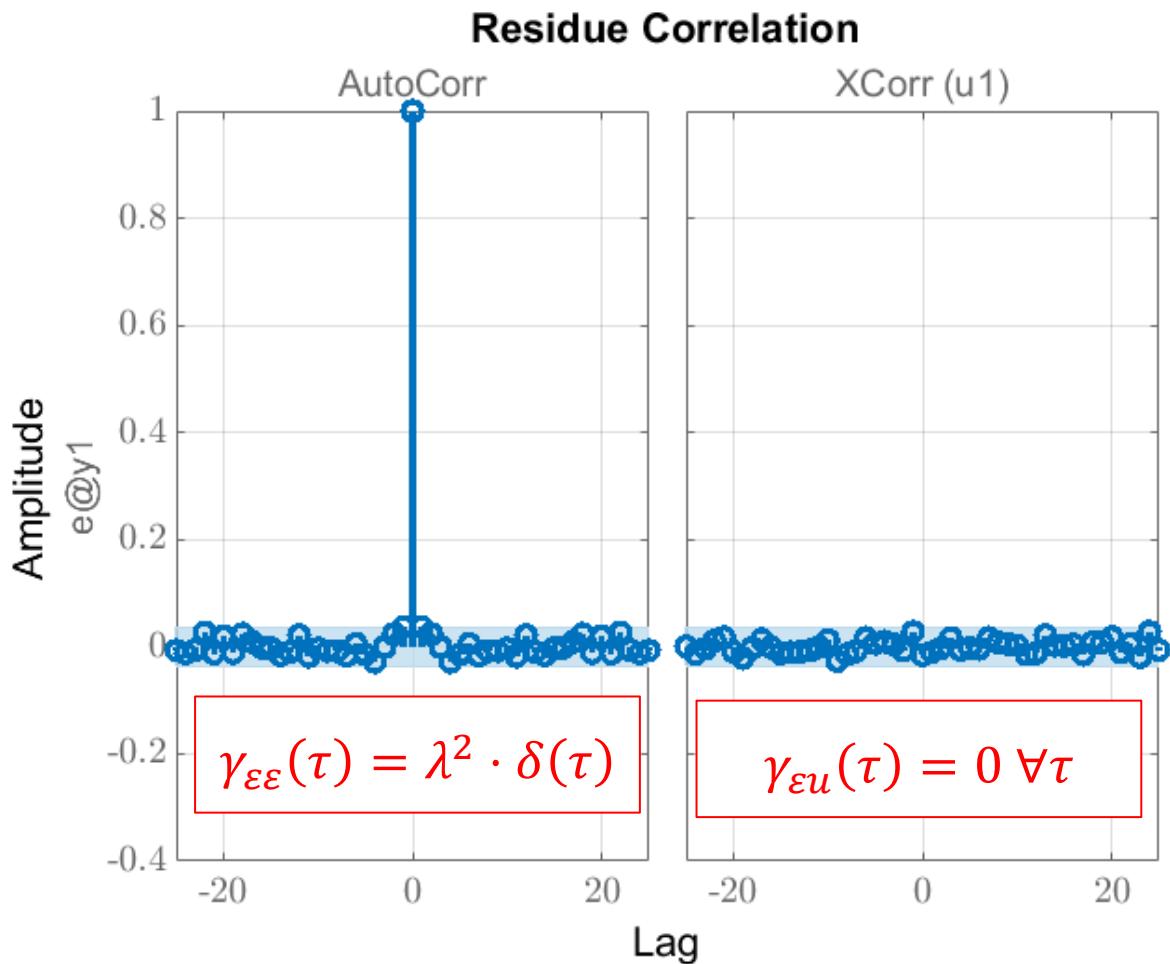
1. un modello **OE** di ordine esatto con $n_b = 1, n_f = 4, k = 3$
2. un modello **ARX** di ordine esatto con $n_b = 1, n_a = 4, k = 3$

Notiamo che $G_0(z) \in \mathcal{G}_{OE}(\theta)$ e $G_0(z) \in \mathcal{G}_{ARX}(\theta)$

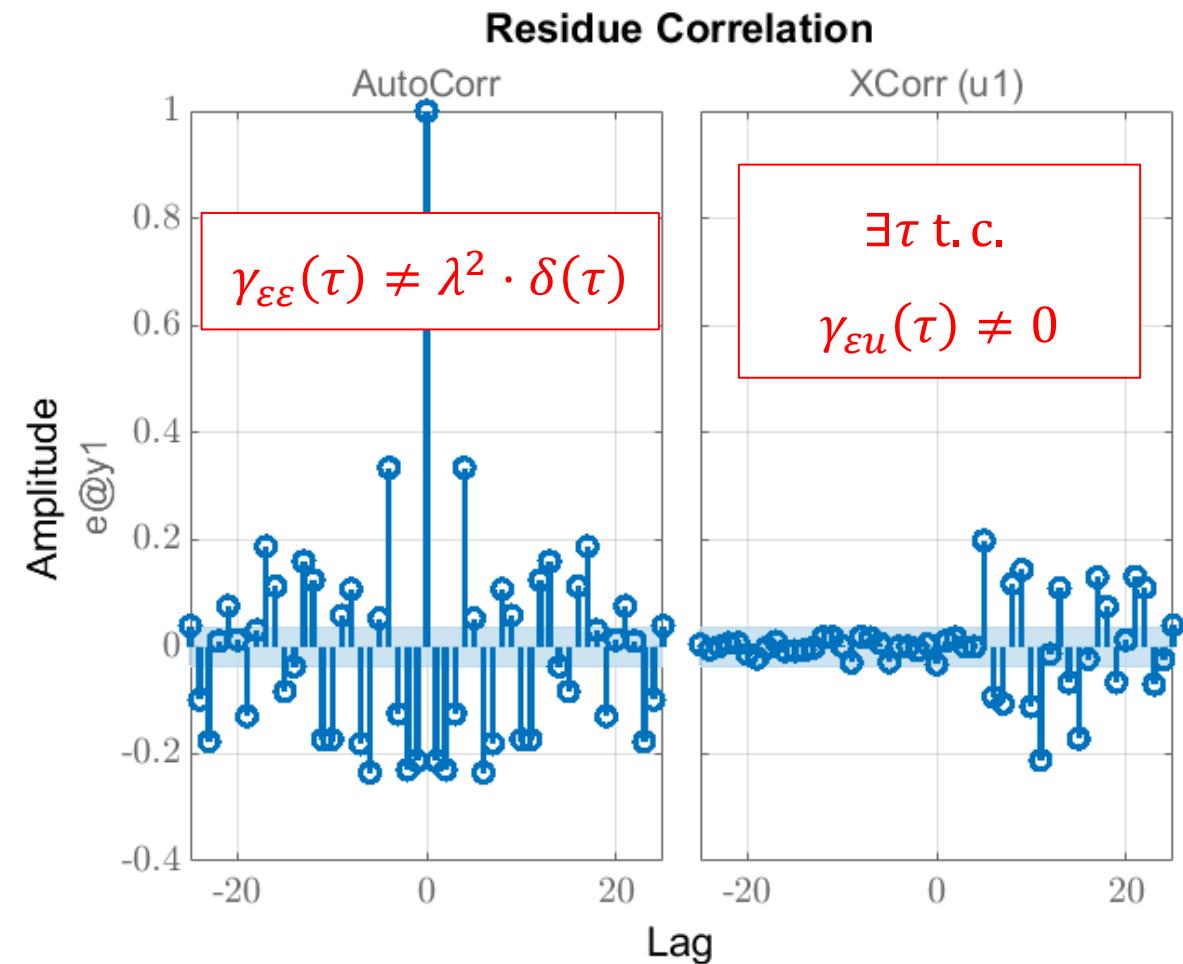


Esempio: identificazione trasmissione meccanica

Modello OE



Modello ARX

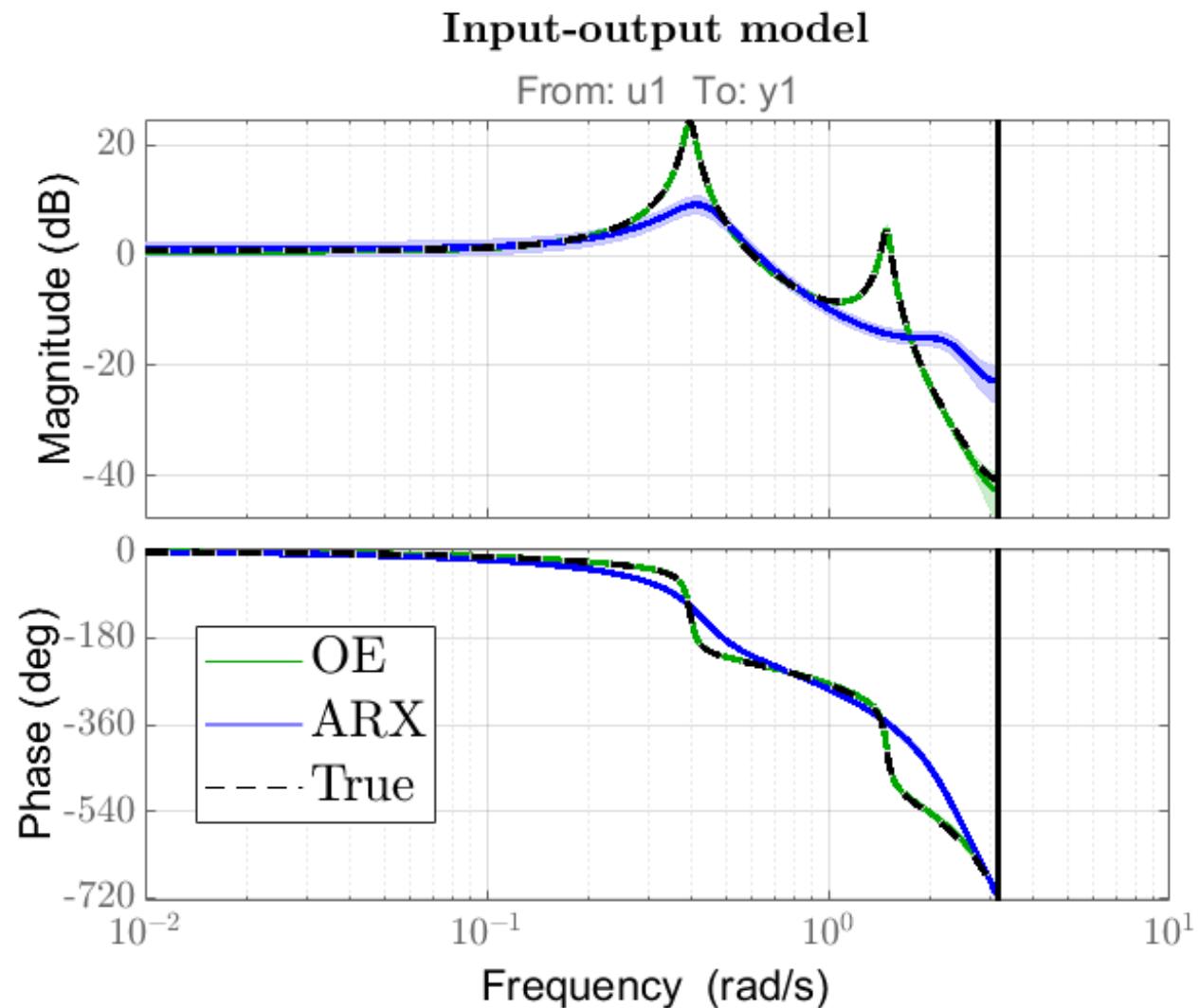


Esempio: identificazione trasmissione meccanica

Anche se $G_0(z) \in \mathcal{G}_{\text{ARX}}(\theta)$, il modello **ARX** non è in grado, **in media**, di stimare correttamente $G_0(z)$ (a meno di avere un SNR elevato)

Il modello **ARX** presenta un'incertezza di stima maggiore

Il modello **OE** di ordine esatto stima perfettamente il sistema vero **in media** (teoria PEM) e ha un'incertezza bassa



Esempio: identificazione trasmissione meccanica

Si consideri ora il sistema \mathcal{S} espresso tramite la famiglia di modelli **ARX**

$$\mathcal{S}: y(t) = \frac{0.10276 + 0.18123z^{-1}}{A(z)} u(t-3) + \frac{1}{A(z)} e(t)$$

$$A(z) = 1 - 1.99185z^{-1} + 2.20265z^{-2} - 1.84083z^{-3} + 0.89413z^{-4}$$

Identifichiamo il sistema con $u(t) \sim WN(0, \lambda^2)$, SNR = 15, $N = 5000$, usando:

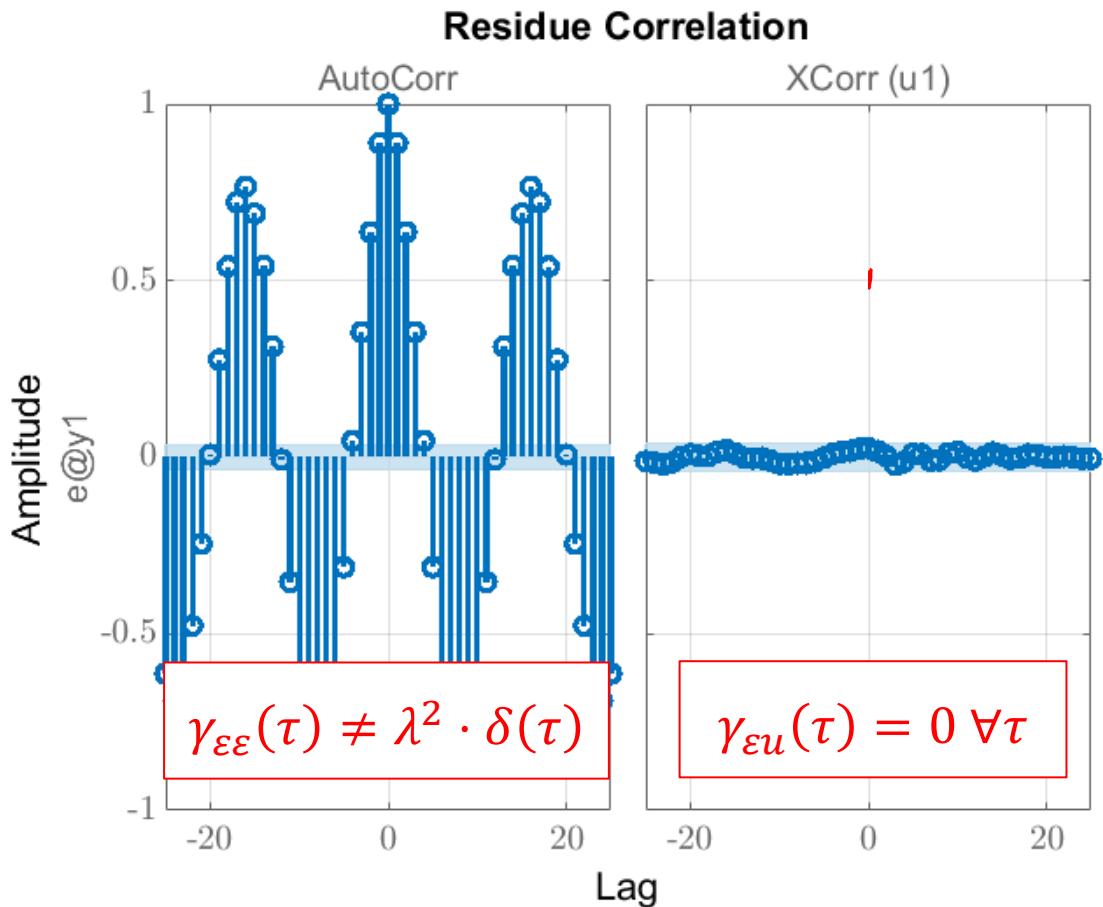
1. un modello **OE** di ordine esatto con $n_b = 1, n_f = 4, k = 3$
2. un modello **ARX** di ordine esatto con $n_b = 1, n_a = 4, k = 3$

Notiamo che $G_0(z) \in \mathcal{G}_{OE}(\theta)$ e $G_0(z) \in \mathcal{G}_{ARX}(\theta)$

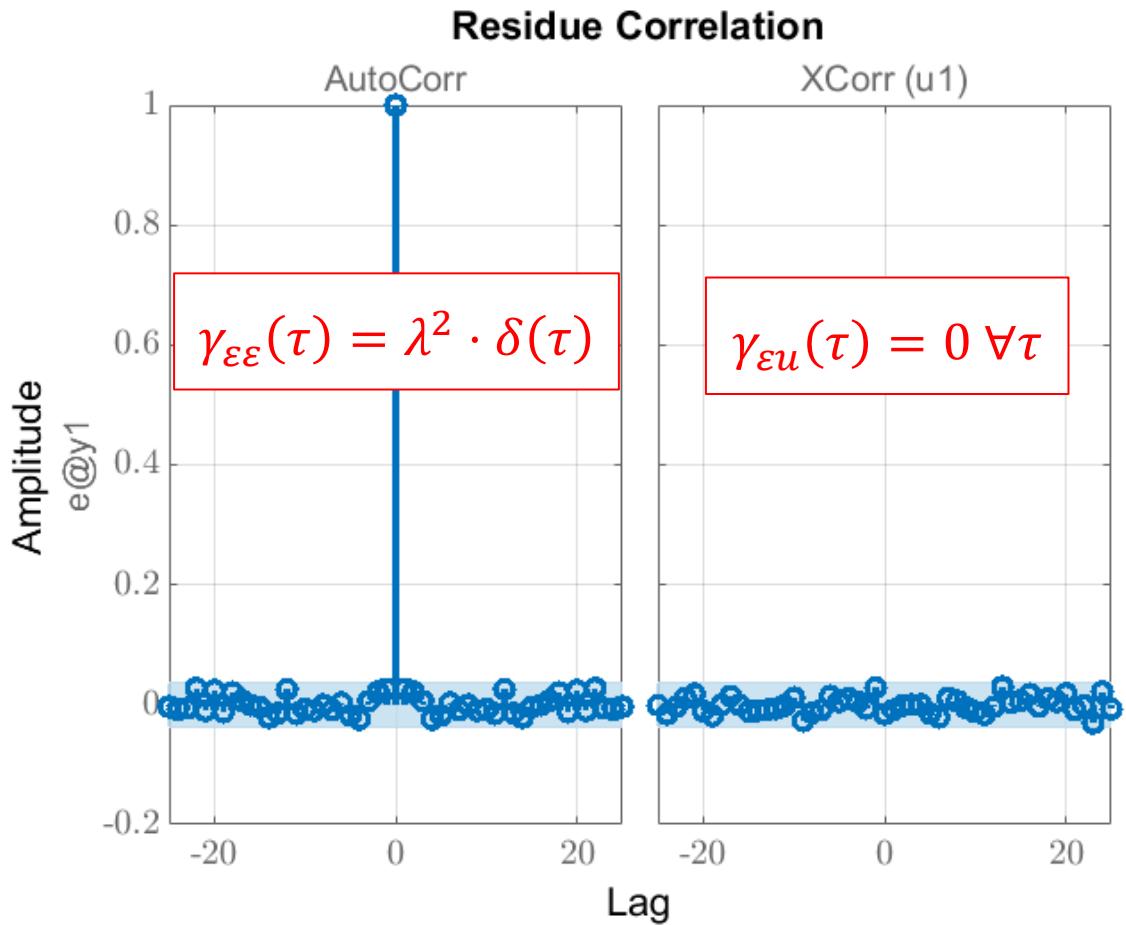


Esempio: identificazione trasmissione meccanica

Modello OE



Modello ARX



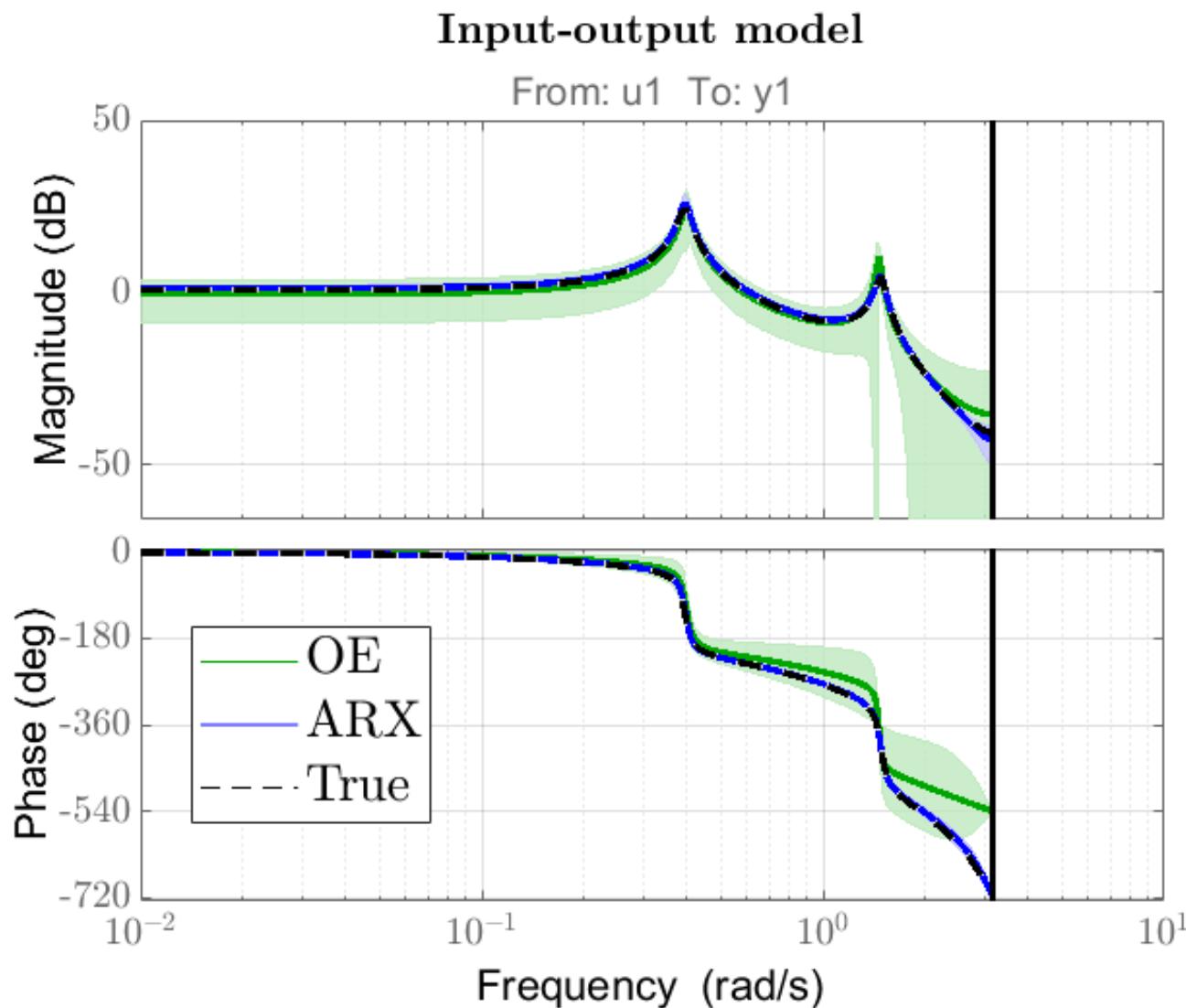
$\mathcal{S} \notin \mathcal{M}_{OE}(\theta)$ con $G_0(z) \in \mathcal{G}_{OE}(\theta)$

$\mathcal{S} \in \mathcal{M}_{ARX}(\theta)$

Esempio: identificazione trasmissione meccanica

Anche se $S \notin \mathcal{M}_{OE}(\theta)$, dato che $G_0(z) \in \mathcal{G}_{OE}(\theta)$, il modello **OE** riesce comunque, **in media**, a stimare correttamente $G_0(z)$

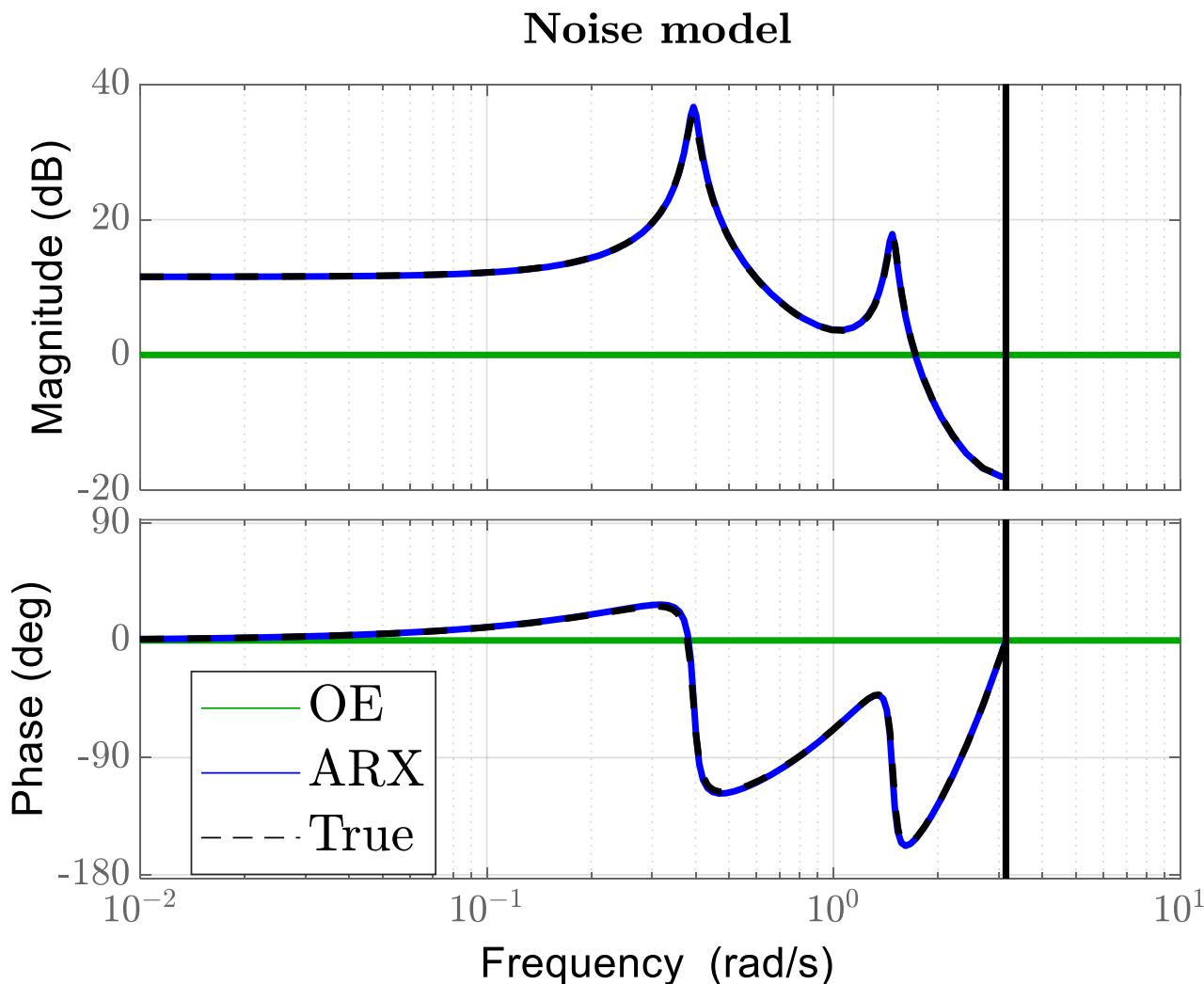
Il modello **ARX** ha **meno varianza** nella stima delle funzioni di trasferimento, dato che è il modello esatto del sistema



Esempio: identificazione trasmissione meccanica

Il modello **ARX** stima perfettamente anche la funzione di trasferimento del rumore

Il modello **OE** non modella il disturbo



Outline

1. Scelta della struttura e complessità del modello
2. Validazione o formule di complessità per la scelta della complessità
3. Analisi dei residui
- 4. Analisi dell'incertezza della stima**
5. Simulazione, predizione del modello identificato
6. Confronto con stima nonparametrica
7. Considerazioni pratiche



Analisi incertezza stima

Con l'analisi dell'incertezza della stima intendiamo:

1. Incertezza sulla stima dei **parametri**
2. Incertezza sulla stima delle **funzioni di trasferimento**
3. Incertezza sulla posizione dei **poli** e degli **zeri**

Incertezza sulla stima dei parametri

L'incertezza sulla stima dei parametri ([Lezione 12: slide 83](#)) può essere utilizzata per verificare la **significatività statistica** di un parametro, dove per significatività intendiamo quanto è probabile che il parametro vero sia effettivamente **diverso da 0**



Analisi incertezza stima sui parametri

Una volta identificato un modello in Matlab, è possibile usare il comando **present** per visualizzare le stime dei parametri e la loro deviazione standard

Esempio

Consideriamo il sistema \mathcal{S} identificato tramite un modello OE con $k = 3, n_b = 1, n_f = 4$ usando $N = 2000$ dati di ingresso $u_{\text{wn}}(t) \sim \text{WN}(0, 0.5^2)$

$$\mathcal{S}: y(t) = \frac{0.103 + 0.181z^{-1}}{1 - 1.991z^{-1} + 2.203z^{-2} - 1.841z^{-3} + 0.894z^{-4}} z^{-3} u(t) + e(t), \quad e(t) \sim \text{WN}(0, 0.5^2)$$



Esempio: analisi incertezza stima sui parametri

Utilizzando il comando `present` otteniamo

$$B(z) = 0.1066 \text{ (+/- 0.01186)} z^{-3} + 0.1798 \text{ (+/- 0.01288)} z^{-4}$$

$$F(z) = 1 - 1.981 \text{ (+/- 0.008776)} z^{-1} + 2.192 \text{ (+/- 0.01835)} z^{-2} - 1.845 \text{ (+/- 0.01647)} z^{-3} + 0.9007 \text{ (+/- 0.006833)} z^{-4}$$

Siccome **nessuna** deviazione standard è in grado di **annullare** il rispettivo parametro, abbiamo una ragionevole possibilità che **tutti** i parametri siano **significativi**



Esempio: analisi incertezza stima sui parametri

Questa analisi può essere usata per **stimare il ritardo puro k**

Ad esempio, possiamo stimare un ARX con $k = 0, n_b = 3, n_a = 4$ e ottenere

$$B(z) = 0.0437 \text{ } (+/- 0.04107) + 0.05822 \text{ } (+/- 0.04108) \text{ } z^{-1} + 0.1113 \text{ } (+/- 0.0411) \\ z^{-2} + 0.1218 \text{ } (+/- 0.04124) \text{ } z^{-3}$$

Il che ci ha un'indicazione che è meglio non includere i termini b_0 e $b_1 z^{-1}$

Osservazione

In genere è buona norma «sperimentare» con modelli ARX in quanto si stimano velocemente ed il minimo è unico, quindi la soluzione non dipende dall'inizializzazione



Analisi incertezza stima funzioni di trasferimento

Abbiamo visto nella [Lezione 12 – Slide 85](#) come, nel caso $S \in \mathcal{M}(\theta)$, l'espressione della varianza sulla stima del valore della **funzione di trasferimento** $G(z, \hat{\theta}_N)$ ad ogni frequenza $z = e^{j\omega}$, può essere approssimata come

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

Se $S \notin \mathcal{M}(\theta)$, esistono altre espressioni più complesse. Questa quantità si può calcolare anche per $H(z, \hat{\theta}_N)$

Quando $\sqrt{\text{Var}[G(e^{j\omega}, \hat{\theta}_N)]}$ può essere **considerata piccola?** La risposta dipende dall'**uso del modello**



Analisi incertezza stima funzioni di trasferimento

Per esempio, se vogliamo usare il modello per **progettare un controllo**, $\sqrt{\text{Var}[G(e^{j\omega}, \hat{\theta}_N)]}$ deve essere «piccola» fino alla banda di controllo ω_c . Una regola euristica è:

$$\sqrt{\text{Var}[G(e^{j\omega}, \hat{\theta}_N)]} < 0.1 |G(e^{j\omega}, \hat{\theta}_N)| \quad \text{per } \omega \leq \omega_c$$

Se $\sqrt{\text{Var}[G(e^{j\omega}, \hat{\theta}_N)]}$ è **troppo grande**, non possiamo garantire che $G(z, \hat{\theta}_N)$ sia una buona stima di $G_0(z)$



Analisi incertezza stima funzioni di trasferimento

Per ridurre la varianza di stima, si possono:

- Collezionare **più dati** N
- Aumentare il **valore della densità spettrale di potenza dell'ingresso** $\Gamma_{uu}(\omega)$ alle frequenze in cui $\sqrt{\text{Var}[G(e^{j\omega}, \hat{\theta}_N)]}$ è grande



Esempio: analisi incertezza funzioni di trasferimento

Si consideri il sistema \mathcal{S} espresso tramite la famiglia di modelli **ARX**

$$\mathcal{S}: y(t) = \frac{0.10276 + 0.18123z^{-1}}{A(z)} u(t-3) + \frac{1}{A(z)} e(t) \quad e(t) \sim \text{WN}(0, 0.01)$$

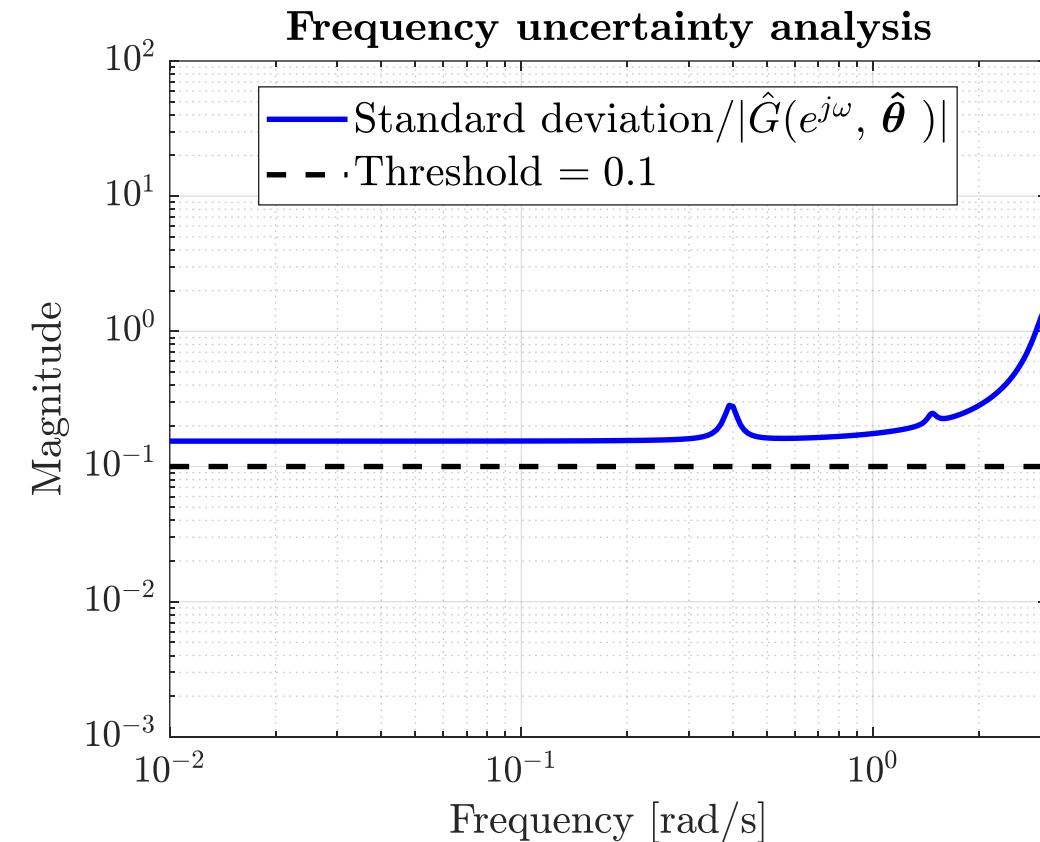
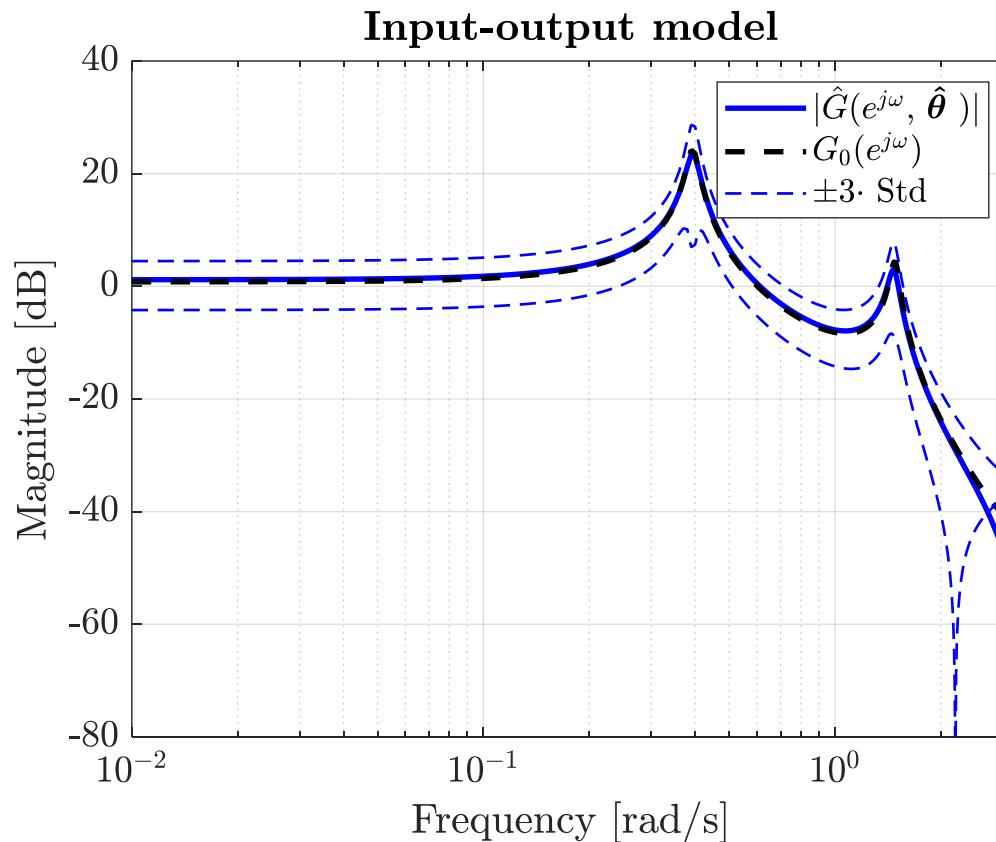
$$A(z) = 1 - 1.99185z^{-1} + 2.20265z^{-2} - 1.84083z^{-3} + 0.89413z^{-4}$$

Identifichiamo il sistema con un modello **ARX** di ordine esatto con $n_b = 1, n_a = 4, k = 3$ e un ingresso $u(t) \sim \text{WN}(0, 0.005)$, usando $N = 2000$ dati

Vogliamo identificare un modello per poi **progettare un controllo nella banda** $[0, 1]$ rad/s



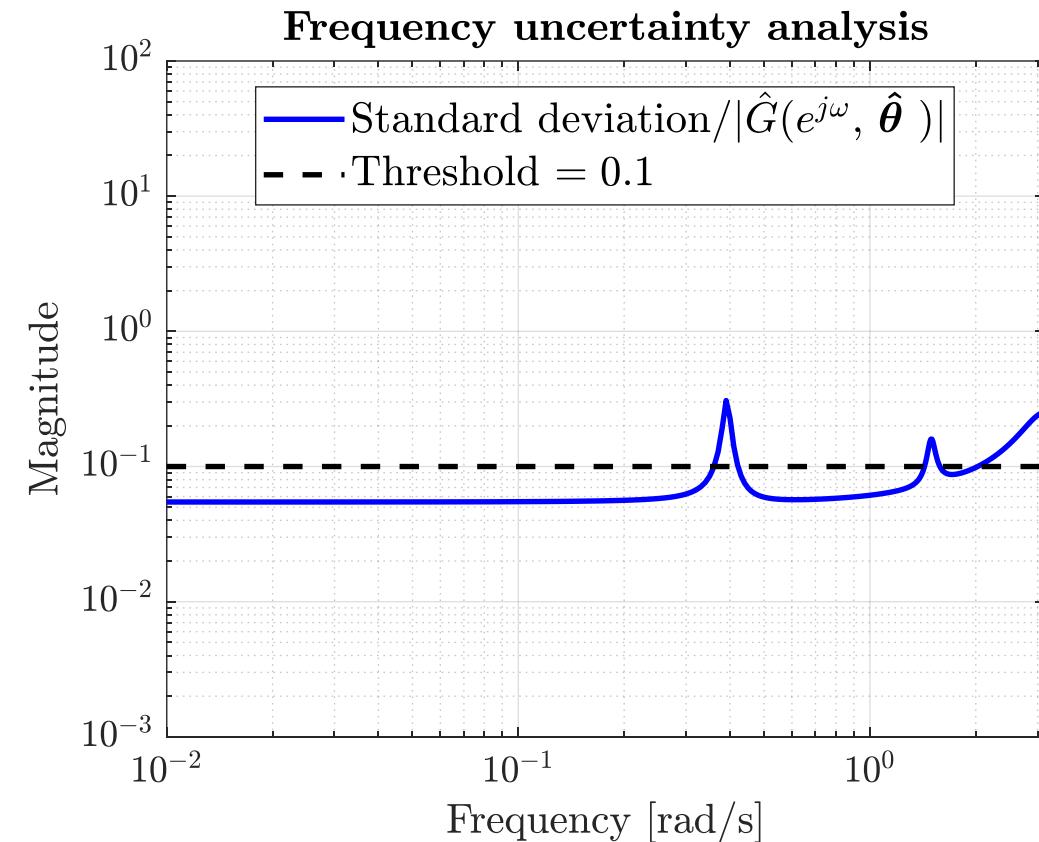
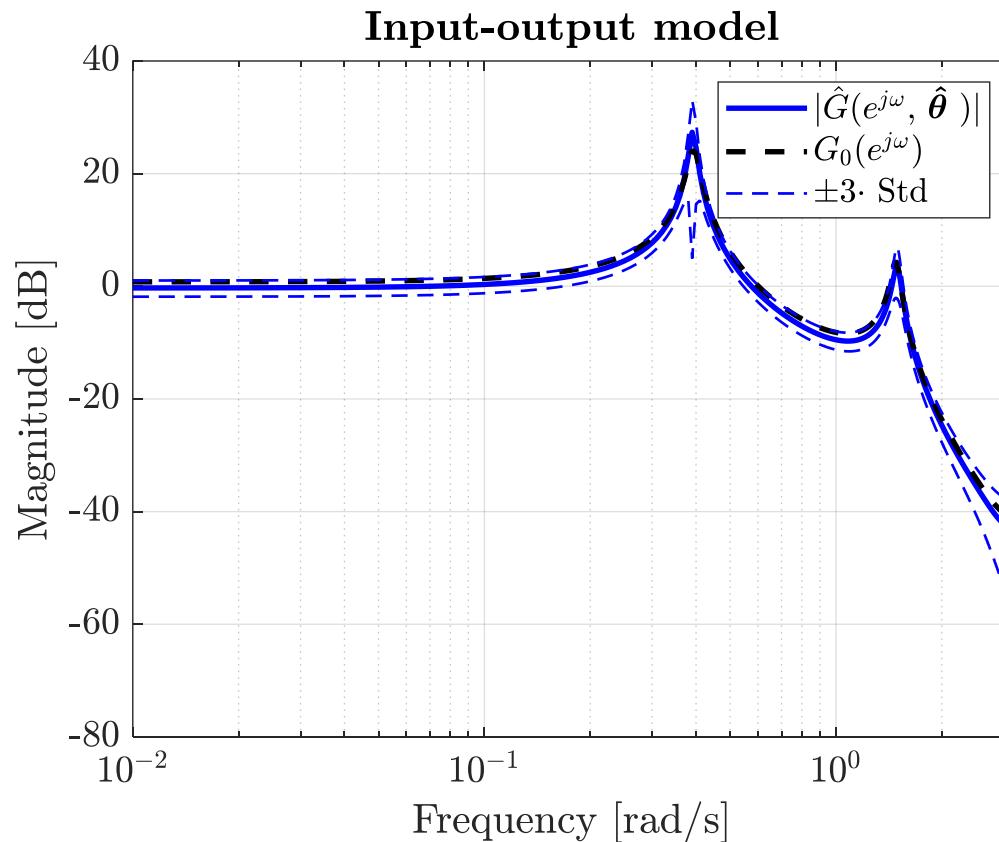
Esempio: analisi incertezza funzioni di trasferimento



Notiamo come la regola $\sqrt{\text{Var}[G(e^{j\omega}, \hat{\theta}_N)]} < 0.1|G(e^{j\omega}, \hat{\theta}_N)|$ non è soddisfatta. Proviamo a usare un **ingresso con più energia**, come $u(t) \sim \text{WN}(0, 0.05)$

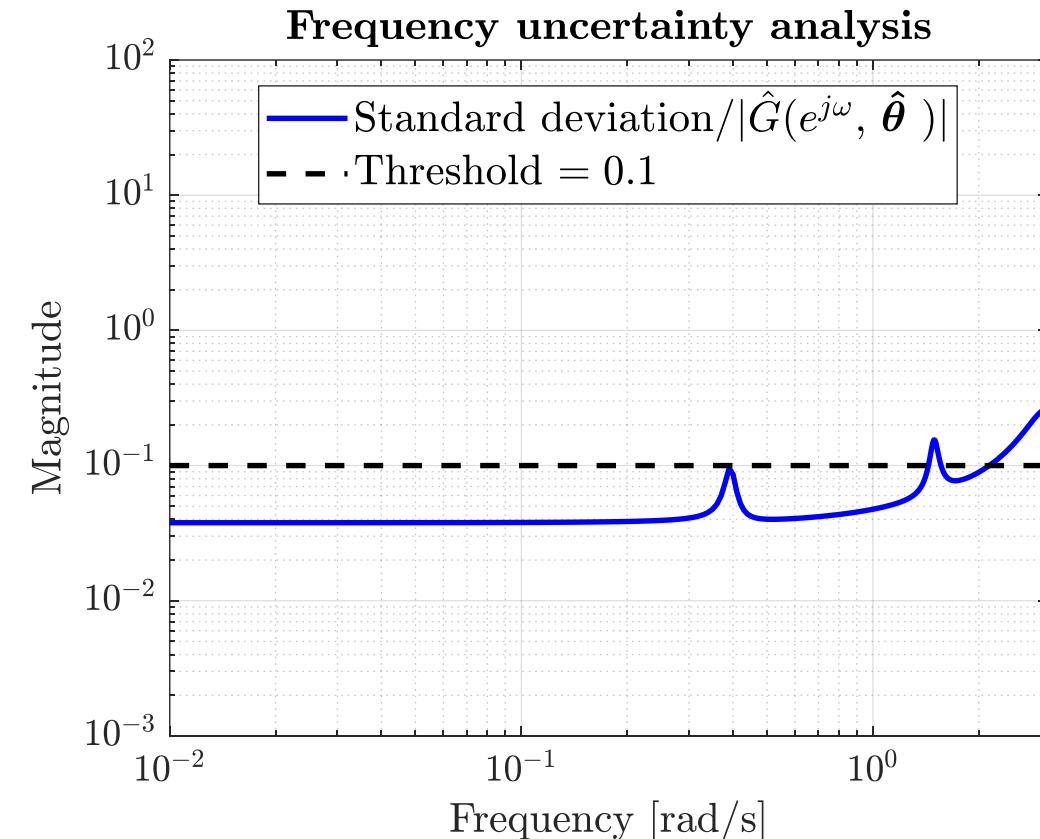
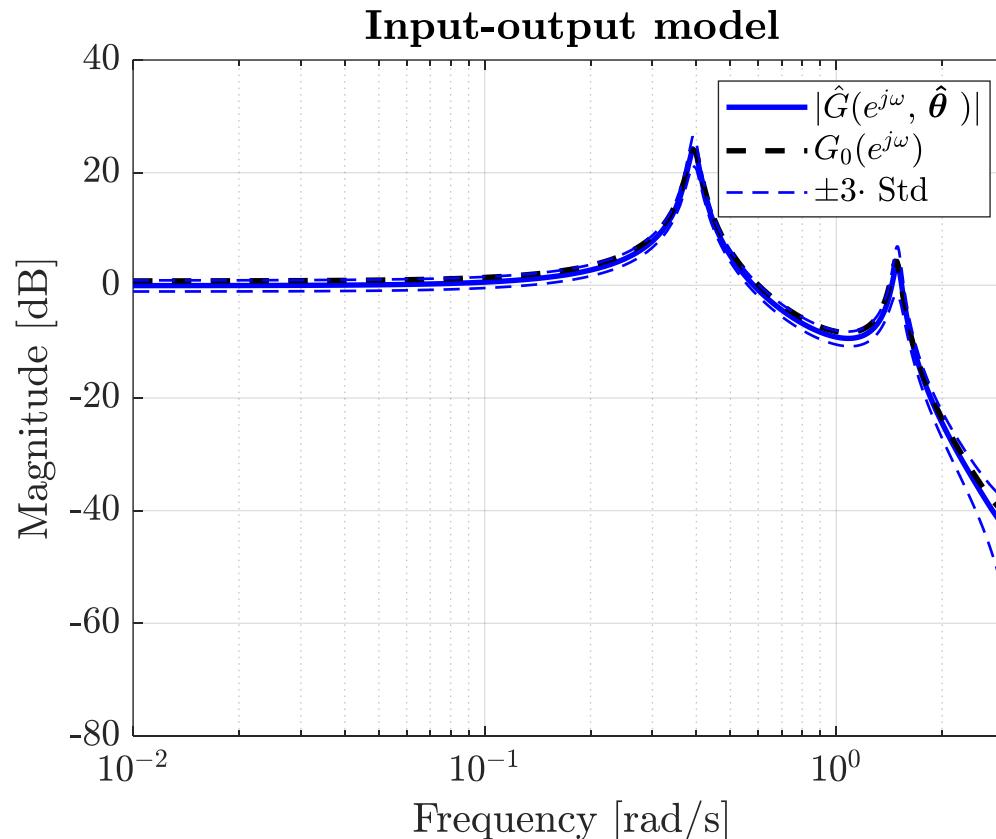


Esempio: analisi incertezza funzioni di trasferimento



Meglio, però nella banda $[0.3, 0.5]$ **c'è ancora troppa incertezza**. Proviamo a utilizzare un ingresso del tipo $u^*(t) = u(t) + 0.2\sin(0.3 \cdot t) + 0.2\sin(0.4 \cdot t)$, con $u(t) \sim \text{WN}(0, 0.05)$

Esempio: analisi incertezza funzioni di trasferimento



L'incertezza è stata ridotta anche nella prima risonanza



Analisi incertezza della stima di poli e zeri

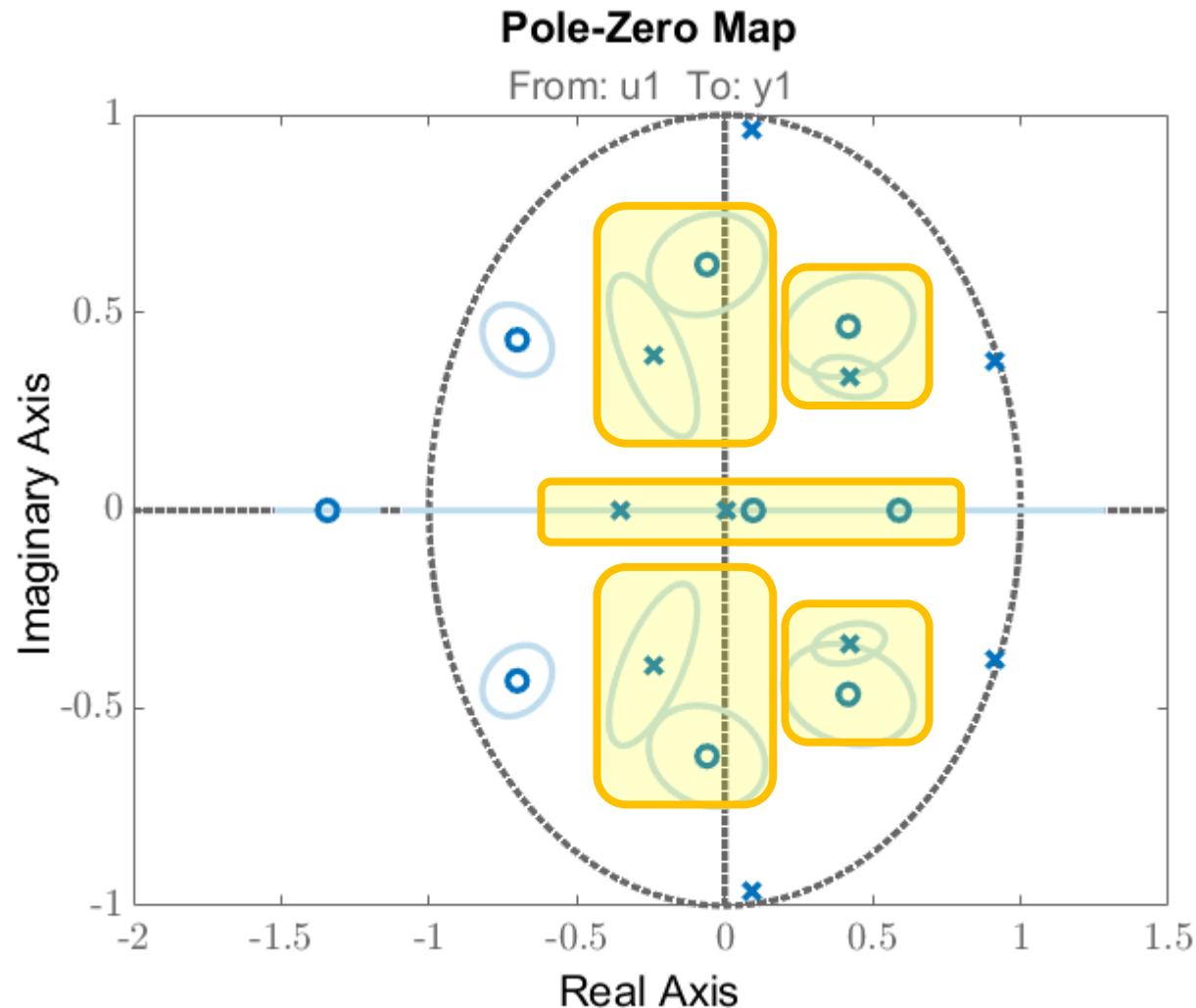
Una volta stimato il modello e analizzato i residui per capire se $S \in \mathcal{M}(\theta)$, è utile **rappresentare i poli e gli zeri** del modello stimato, per verificare la possibilità di **cancellazioni polo\zero** (e quindi evitare sovraparametrizzazioni)

L'idea è verificare 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



Analisi incertezza della stima di poli e zeri

Nel grafico seguente, vi sono 6 situazioni in cui poli e zeri possono potenzialmente essere semplificati e rimossi



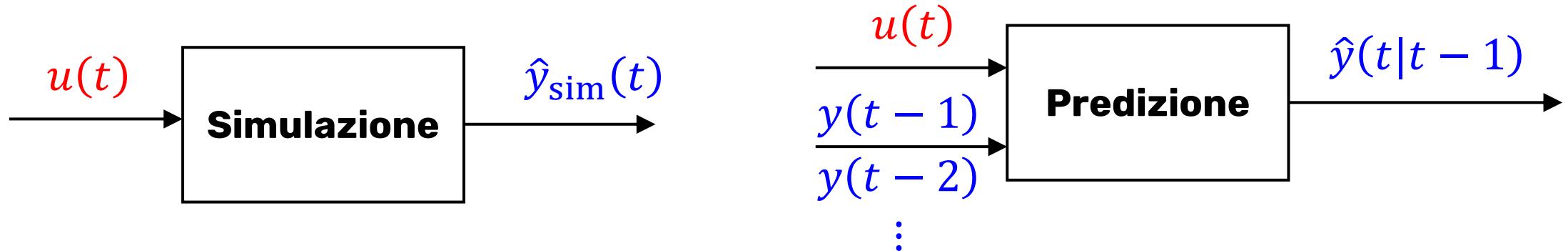
Outline

1. Scelta della struttura e complessità del modello
2. Validazione o formule di complessità per la scelta della complessità
3. Analisi dei residui
4. Analisi dell'incertezza della stima
- 5. Simulazione, predizione del modello identificato**
6. Confronto con stima nonparametrica
7. Considerazioni pratiche



Simulazione e predizione del modello

È possibile **simulare** o **predire** 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 è $e_{\text{sim}}(t) = y(t) - \hat{y}_{\text{sim}}(t) = y(t) - G(z, \hat{\theta}_N)u(t)$. Un modello buono non rende necessariamente $e_{\text{sim}}(t)$ piccolo, poiché la **simulazione non considera il modello del rumore**



Simulazione e predizione del modello

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

In Matlab si usa il metodo **compare**, che permette di confrontare la simulazione o la predizione a k passi di uno o più modelli rispetto ai dati

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

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

È poi possibile confrontare una **stima della densità spettrale di potenza** di $\hat{v}(t)$ con il **modello del rumore** identificato $H(z, \hat{\theta}_N)$



Esempio: Simulazione e predizione del modello

Consideriamo ancora il sistema ARX, e usiamo modelli OE e ARX di ordine esatto

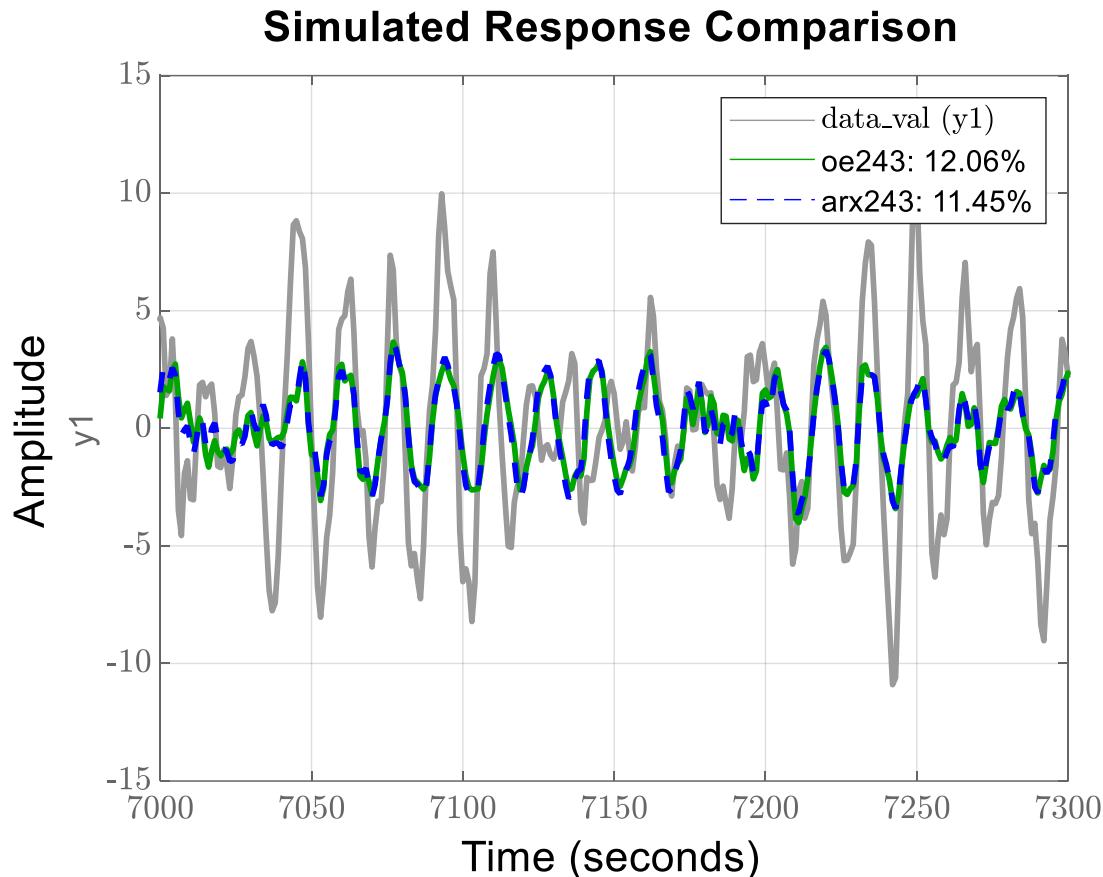
$$\mathcal{S}: y(t) = \frac{0.10276 + 0.18123z^{-1}}{A(z)} u(t-3) + \frac{1}{A(z)} e(t) \quad e(t) \sim \text{WN}(0, \lambda^2)$$

$$A(z) = 1 - 1.99185z^{-1} + 2.20265z^{-2} - 1.84083z^{-3} + 0.89413z^{-4}$$

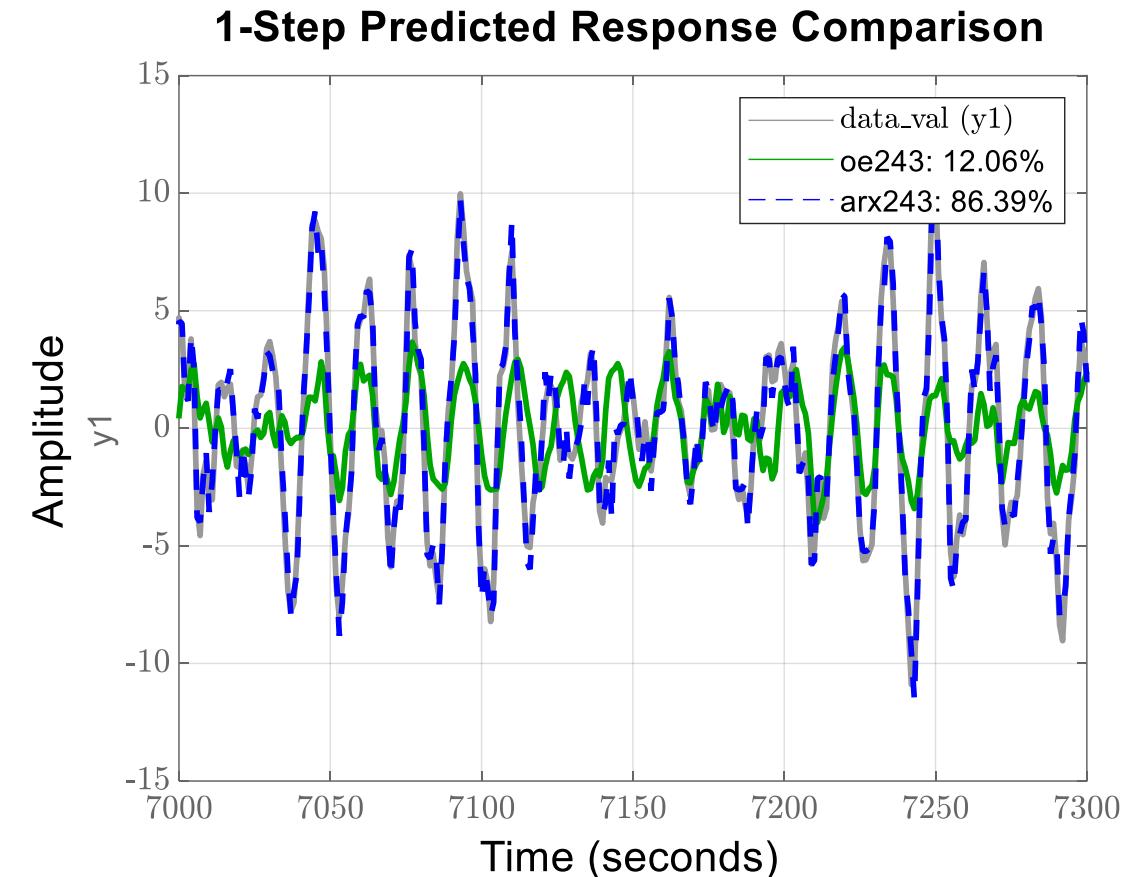
Confrontiamo sia la simulazione che la predizione dei due modelli identificati per scegliere il modello migliore



Esempio: Simulazione e predizione del modello



La risposta **simulata** dei due modelli è simile (entrambi stimano $G_0(z)$ molto bene). Ciò che non è spiegato è una stima del rumore $H_0(z)e(t)$



Il modello ARX **predice** molto meglio i dati, poiché modella correttamente il rumore $H_0(z)e(t)$

Outline

1. Scelta della struttura e complessità del modello
2. Validazione o formule di complessità per la scelta della complessità
3. Analisi dei residui
4. Analisi dell'incertezza della stima
5. Simulazione, predizione del modello identificato
- 6. Confronto con stima nonparametrica**
7. Considerazioni pratiche



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 stima spettrale ([Lezione 12](#)). In Matlab si usa **etfe**, **spa**, **spafdr**

Esempio. Consideriamo ancora il sistema ARX, e usiamo modelli OE e ARX di ordine esatto

$$\mathcal{S}: y(t) = \frac{0.10276 + 0.18123z^{-1}}{A(z)} u(t-3) + \frac{1}{A(z)} e(t) \quad e(t) \sim WN(0, \lambda^2)$$

$$A(z) = 1 - 1.99185z^{-1} + 2.20265z^{-2} - 1.84083z^{-3} + 0.89413z^{-4}$$



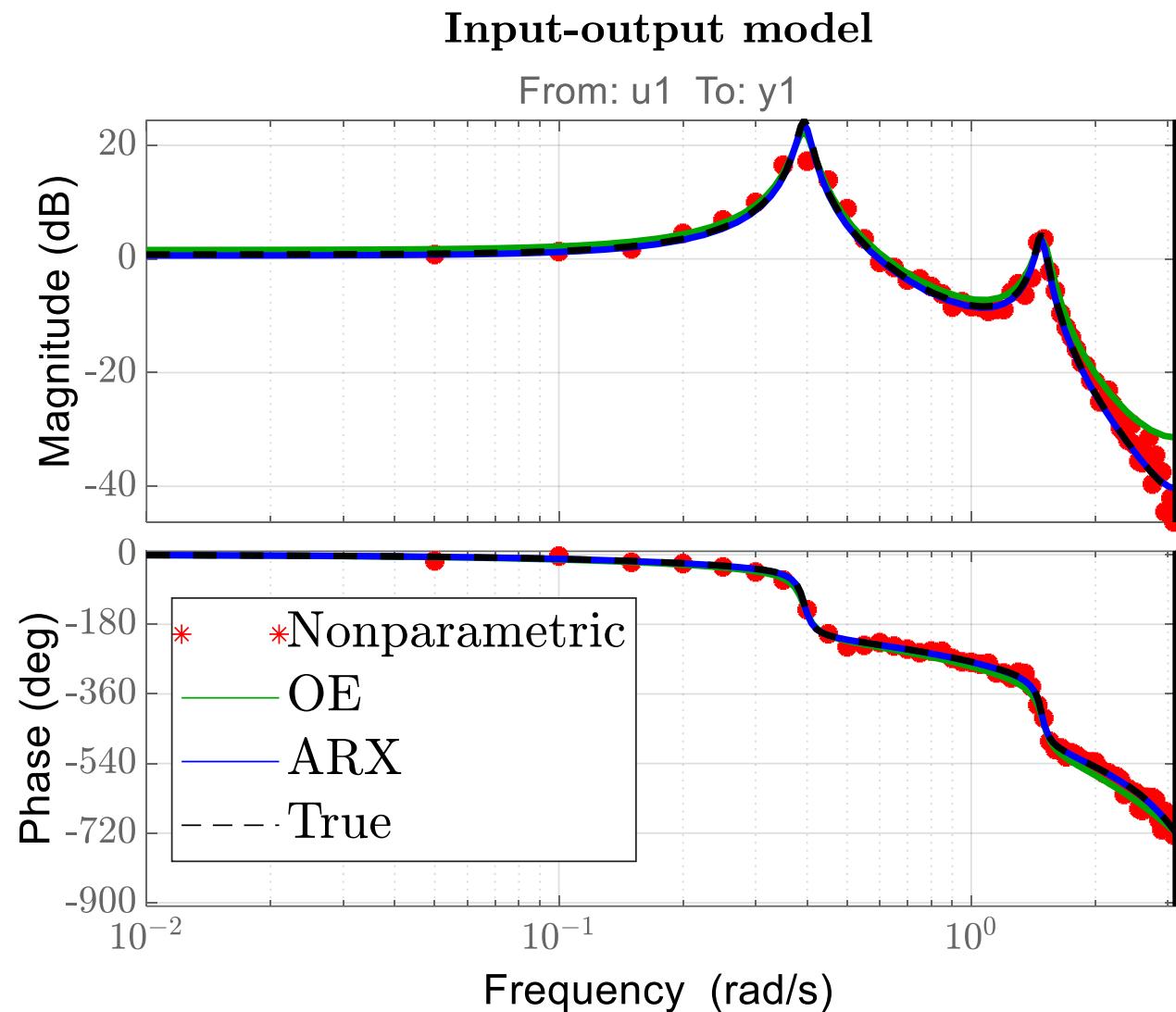
Esempio: confronto con stima nonparametrica

Entrambi i modelli OE e ARX stimano bene $G_0(z)$ e la stima nonparametrica (che prendiamo come proxy di $G_0(z)$ che in non sappiamo)

Simuliamo i modelli OE e ARX per calcolare una stima del rumore

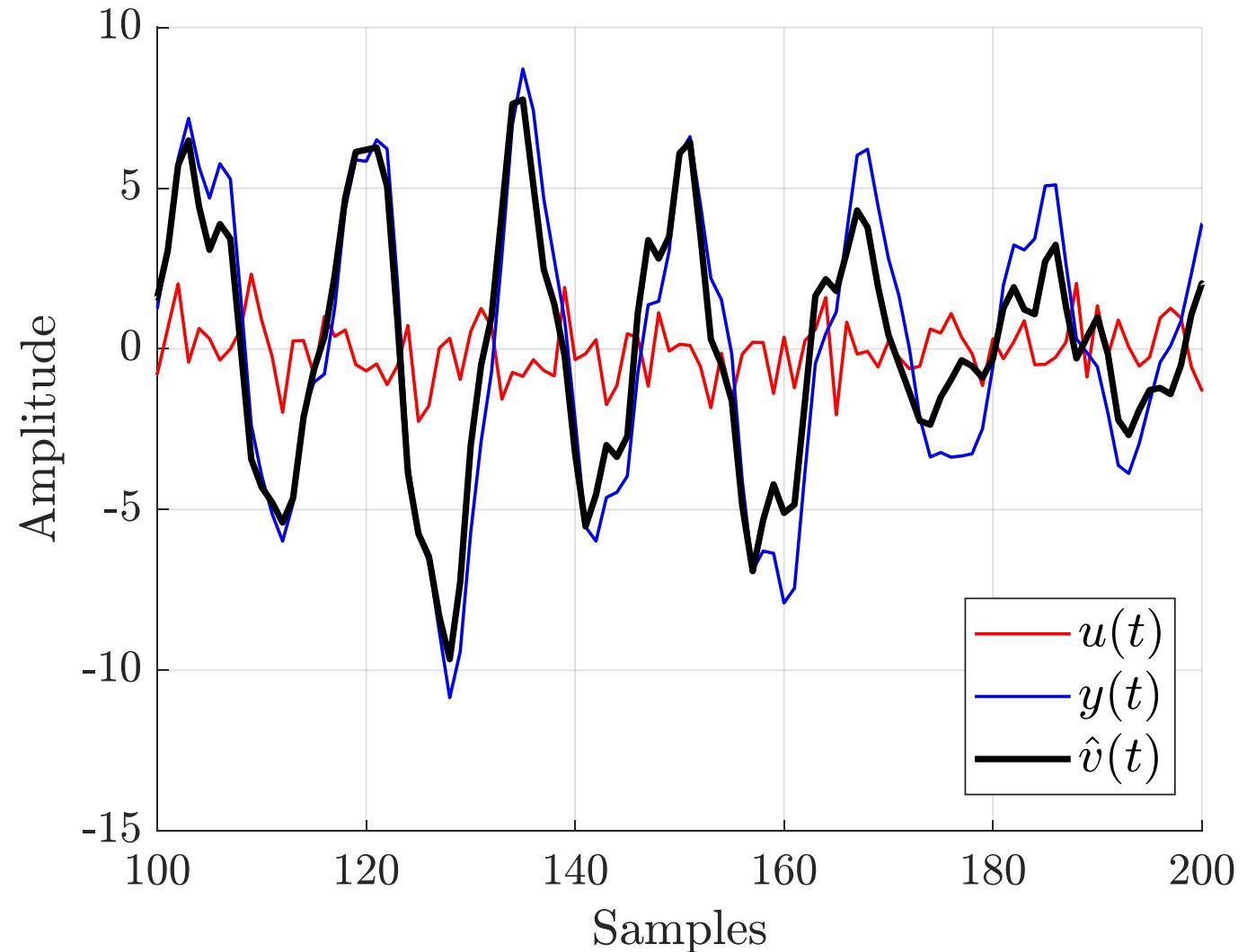
$$\hat{v}(t) = y(t) - \hat{y}_{\text{sim}}(t)$$

e verificare se il modello del rumore è opportuno

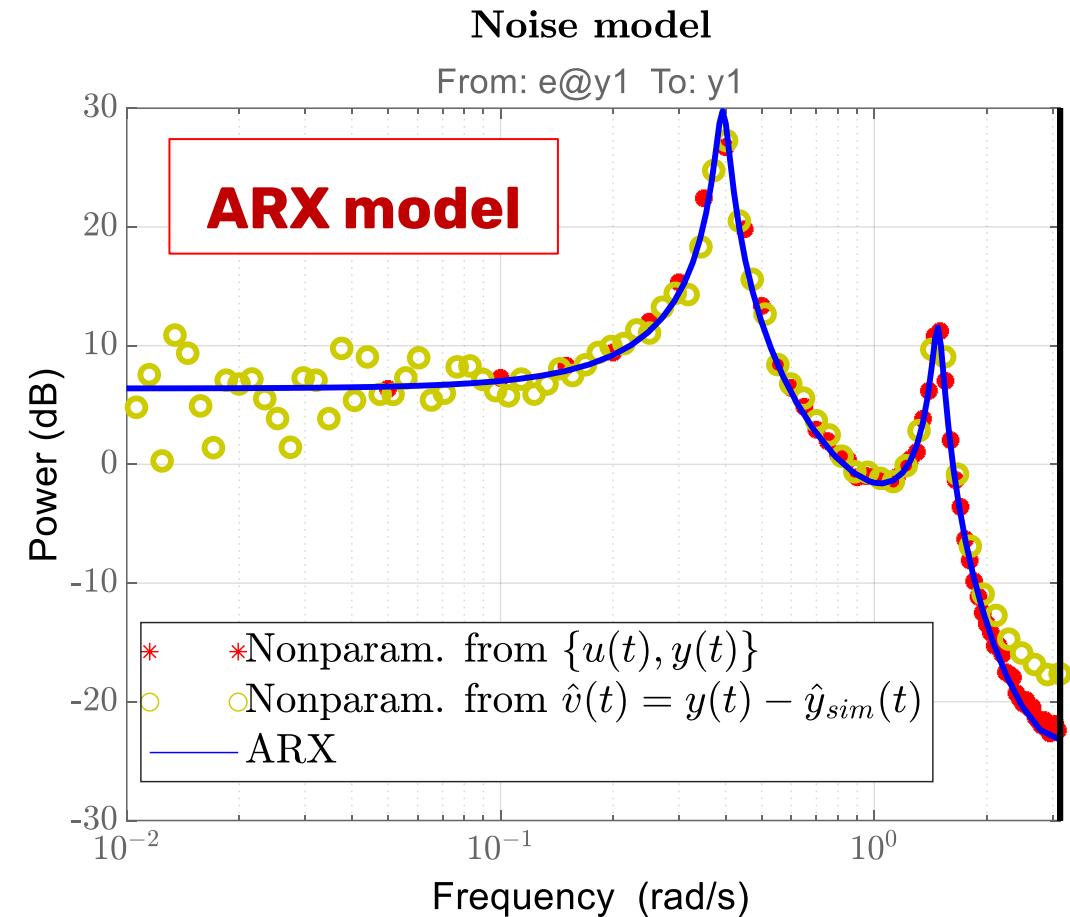
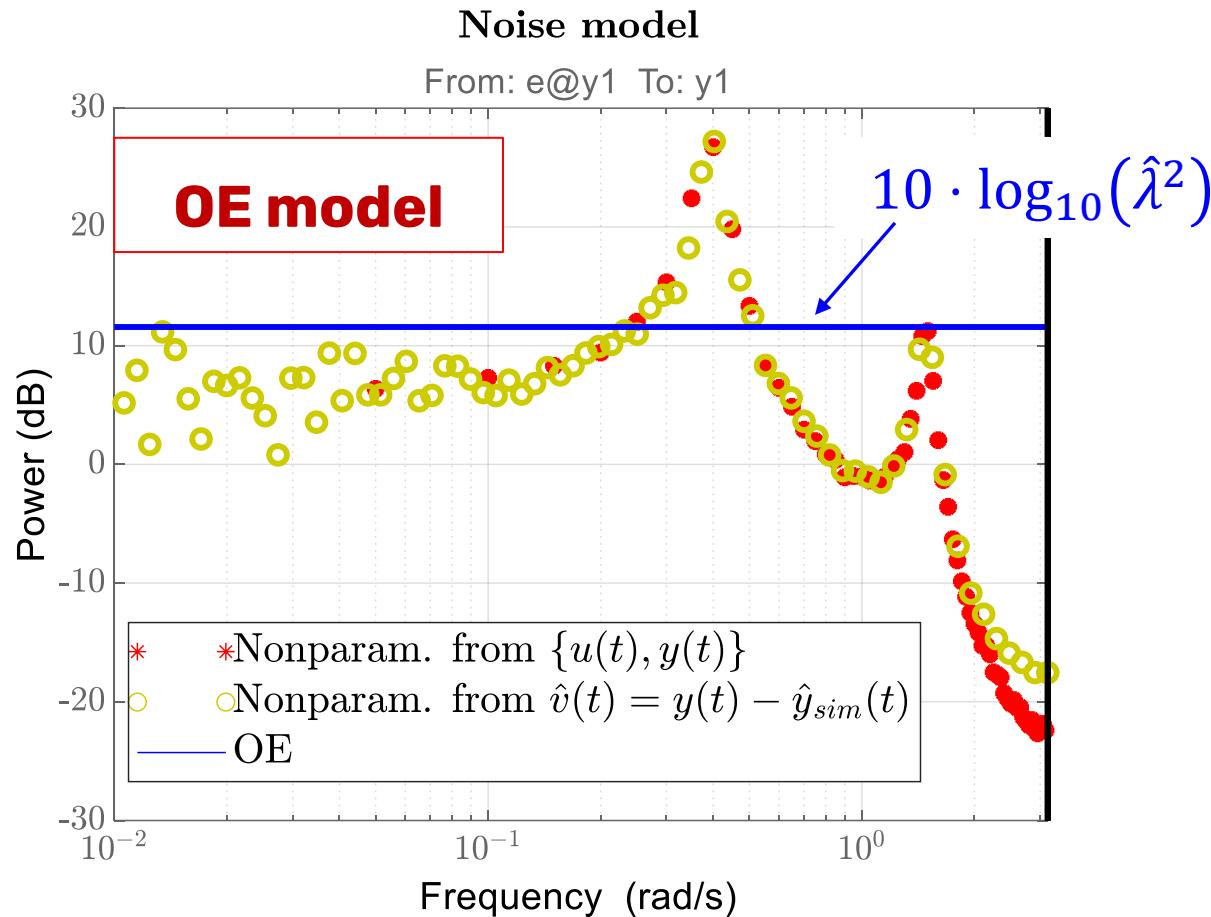


Esempio: confronto con stima nonparametrica

Vediamo che il disturbo $\hat{v}(t)$ non è
trascurabile rispetto a $y(t)$
misurato



Esempio: confronto con stima nonparametrica



Il modello del rumore parametrico **non è simile** al modello del rumore nonparametrico → **probabilmente il modello del rumore parametrico è SBAGLIATO**

Il modello del rumore parametrico **è simile** al modello del rumore nonparametrico → **probabilmente il modello del rumore parametrico è CORRETTO**

Outline

1. Scelta della struttura e complessità del modello
2. Validazione o formule di complessità per la scelta della complessità
3. Analisi dei residui
4. Analisi dell'incertezza della stima
5. Confronto con stima nonparametrica
6. Simulazione, predizione del modello identificato
- 7. Considerazioni pratiche**



Considerazioni pratiche

Per quanto riguarda la **frequenza di campionamento** con la quale acquisire i dati:

- Per la **ETFE**, meglio f_s **alta**, in quanto avrò più risoluzione in frequenza
- Per l'identificazione **PEM**, meglio f_s **non troppo alta**, poichè, quando f_s cresce, i poli del modello discreto tendono a 1, rendendolo non più asintoticamente stabile

Scelta tipica di f_s per **l'identificazione parametrica**:

$$10 \cdot \frac{\omega_c}{2\pi} \leq f_s \leq 30 \cdot \frac{\omega_c}{2\pi}$$

dove ω_c denota la **banda del sistema** (in rad/s), osservata tramite ETFE



Considerazioni pratiche

Dati con f_s **inferiore** possono essere ottenuti tramite:

- **ripetizione** dell'esperimento
- **ricampionamento** dei segnali (In Matlab **resample, decimate**)



Tipica procedura (iterativa) di identificazione

1. Cercare di capire la **fisica** del sistema e configurare il sistema di **acquisizione**
2. Ottenere una **stima nonparametrica** di $G_0(z)$ e $H_0(z)$
3. Sulla base della stima nonparametrica e, se possibile, di risposte allo scalino, scegliere la **frequenza di campionamento** per la stima parametrica
4. Identificare un modello lineare nei parametri, **ARX** o **FIR**, scegliendone l'ordine con *validazione* o *formule di complessità*. Controllare che $G_0(z)$ sia stimata bene tramite *analisi dei residui*
5. Controllare se vi siano *cancellazioni poli\zeri* e determinare l'ordine del modello. Usare questo ordine per stimare un modello **OE**. Validare il modello con *analisi residui*, *confronto con stima nonparametrica*, *analisi incertezza* e *simulazione*
6. Aggiungere un **modello del rumore**, per esempio usando una struttura **Box-Jenkins**, e validarla tramite *analisi residui*, *confronto con stima nonparametrica*, *analisi incertezza* e *simulazione*





UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

FINE DEL CORSO DI IMAD

Grazie!

Mirko Mazzoleni
mirko.mazzoleni@unibg.it