IDENTIFICAZIONE DI SISTEMI DINAMICI

(DYNAMICAL SYSTEM IDENTIFICATION)

GIORGIO PICCI

Dipartimanto di Ingegneria dell'Informazione,
Università di Padova, Italy

Primo semestre 2013-2014

Motivations

System Identification is **automatic construction of mathematical models of dynamical systems from observed data**. Importance and usage has grown tremendously in the last decades. Progress of microelectronics and computer hardware and the dramatic increase of real-time computing power after the 1990's are leading to a shift of paradigms in the design of engineering systems. To cope with the growing complexity and the rising demand for sophistication and high performance, the design of modern control and communication systems has to be based on *quantitative models* of the signals and systems involved.

Application in diverse fields like: automatic control and communication systems, econometrics, geophysics, hydrology, structural testing in civil engineering, bioengineering, automotive science, ... model-based coding and recognition of audio and video signals.
Devices based on on-line signal identification are now quite common e.g. mobile phones, and part of commercially available communication systems.

Scope of System Identification

Construct a conceptual framework and algorithms for automatic model building from observed data.

Observed variables accessible to measurement, are ”inputs” or *exogenous* variables, denoted $(\mathbf{u})$, and ”outputs” or *explained* variables denoted $(\mathbf{y})$. Normally the variables can only be measured at discrete instants of time t and collected in a string of data called a *time series*.

Data are collected in one unrepeatable experiment. No preparation of the experiment is possible (i.e. we cannot choose the experimental conditions or the input function to the system at our will) we are forced to do the best with the data coming from one unrepeatable experiment.

There may be a variety of different reasons to build models. We shall be interested in model building for the purpose of **prediction and control**. This imposes a fundamental requirement that the identified model should

be useful for prediction or control of future; i.e. not yet observed, system variables.

Models can be *physical, black-box or grey-box*. We will deal with black-box identification.
The allowable models generally belong to a model class, selected on basis of flexibility and mathematical simplicity, say the class of finite-dimensional linear time-invariant systems of a given order and the identification problem is generally formulated as that of inferring a "best" mathematical model in the model class on the basis of the observed data. So the identification problem consists of three ingredients:
the data, the model class, the model selection criterion.
We shall discuss these three ingredients below.

Essential ingredients 1: The model class

In real systems, there are always many other variables besides the preselected inputs and outputs which influence the time evolution of the system. These variables represent the unavoidable interaction of the system with its environment. For this reason, even in the presence of a true causal relation between inputs and outputs there always are some *unpredictable* fluctuations of the values taken by the measured output $\mathbf{y}(t)$ which are not explainable in terms of past input (and/or output) history.

We cannot (and do not want to) take into account these variables explicitely in the model as some of them may be inaccessible to measurement and in any case this would lead to complicated models with too many variables. We need to work with models of small complexity and treat the unpredictable fluctuations in some simple *aggregate* manner.

Models (however accurate) are *always mathematical idealizations of nature*. No physical phenomenon, even if the experiments were conducted in an ideal interactions-free environment can be described **exactly** by a set of differential or difference equations and even more so if the equations are a priori restricted to be linear, finite-dimensional and time-invariant. So the observables, even in an ideal "disturbance-free" situation cannot be expected to obey *exactly* any linear time-invariant model.

A realistic formulation of the problem requires a satisfactory notion of *non-rigid*, i.e. *flexible or approximate*, notion of mathematical model of the observed data.
A model should be able to accept as legitimate, data sets (time series) which may possibly differ slightly from each another.

Imposing rigid ”exact” descriptions of the type $F(u,y)=0$ to experimental data has been criticized since the early beginnings of experimental science. Particularly illuminating is Gauss' general philosophical discussion in *Theoria motus corporum caelestium* sect. III, p. 236.

Example: there has been a widespread belief in the early years of control science that identification was merely a matter of solving (exactly) for h a linear convolution equation

$$y(t)=\sum_{t_0}^{t} h(t-\tau)u(\tau)$$

or, equivalently, by matching exactly pointwise harmonic response data with linear transfer function models. Results have always been extremely sensitive even to small perturbations in the data.

New incoming data tend to change the model drastically, which means that a model determined in this way has very poor predictive capabilities.

The reason is that data obey exactly rigid relations of this kind “with probability zero”. If in addition the model class is restricted to be finite-dimensional, which is what is necessary for control applications, imposing the integral equation model () on real data normally leads to disastrous results. This is by now very well-known and documented in the early literature, see e.g. [Phillips,Twomey,Hunt,Ekstrom]. In the language of numerical analysis, fitting rigid models to measured data invariably leads to very *ill-conditioned problems*.

We shall follow Gauss idea of describing data by a *distribution function*; i.e. work in a **probabilistic setting**. Models will then be probabilistic objects.

Other alternatives are possible, say using deterministic model classes consisting of a rigid “exact” model as a “nominal” object, plus an uncertainty ball around it. In this case, besides a nominal model, the identification procedure is required to provide at least bounds on the magnitude of the relative “uncertainty region” around the nominal model.

Here one should provide a mathematical description of how the dynamic uncertainty ball is distributed in the frequency domain, rather than, as more traditionally done, in the parameter space, about the nominal identified model.

Essential ingredients 2: The Data

We need to introduce a *probabilistic description of the data*. The data at our disposal at some fixed time instant represent only partial evidence about the behaviour of the system as we do not know the future continuation of the input and output time series. Yet,
all possible continuations of our present data must carry information about the same physical phenomenon we are about to model, and hence the possible continuations of the data cannot be "totally random" and must be related to what we have observed so far. So, data must have a "memory"; i.e. their own dynamics, and in order to discover models of systems, we have to first understand models of uncertain signals.

Data will be modeled as **stochastic processes** in fact discrete-time stochastic processes. Since the underlying phenomenon (system) that we want to describe is assumed to be time invariant the stochastic processes which model the observed data will be stationary.

Essential ingredients 3: The Model Selection Criterion

In these lectures we shall take the probabilistic point of view and model uncertainty with the apparatus of probability theory. In this framework **identification is essentially a problem of mathematical statistics**. General idea: *minimize criteria based on a notion of distance between the data and the model class*.

Trivial example: Least squares fitting. Note that in Gauss' work least squares come out as a solution method for optimally fitting a certain class of *density functions* to the observed data (maximum likelihood).

Nota Bene: the basic problem of identification is, much more than designing algorithms which fit models to observed data (the easy part), the quantification of the *uncertainty bounds* or the description of the *dynamic errors* which will be incurred when using the model with generic data. Any sensible identification method should provide some mathematical description of how uncertainty is distributed in time or frequency about the nominal identified model. In this respect statistics and probability offer an ideal framework. Describing a probability distribution is the same as modeling uncertainty.

A common critique

It has been argued that the abstract "urn model" of probability theory looks inadequate to deal with situations like the one we have envisaged, where there is just one unrepeatable experiment and there is really no sample space around from which the results of the experiment could possibly have been drawn.

The critique has the merit of criticizing large sectors of the literature where the statistical framework is often imposed dogmatically.

In our opinion however, the critique originates from a tendendency to confuse physical reality with mathematical modelling. In effect the urn model (i.e. the underlying probability space) is just a mathematical device which is *not required to have any physical interpretation* and could in principle be used to model things which, to be described deterministically, would require extremely complicated mathematical models with myriads of variables.

On the same grounds it could be questioned if there are in nature objects like differential or difference equations.

Outline of the Course

1. Classical statistical Background
2. Review of SISO stochastic models
3. Minimal Prediction Error Methods (PEM)
4. Asymptotics: Ergodicity and CLT
5. Validation and model structure estimation
6. Computational methods, recursive identification
7. Almost periodic signals and spectral estimation

FONDAMENTI DI STATISTICA

La teoria moderna della Probabilità è una teoria *assiomatica*.
Punto di partenza: uno *spazio degli esperimenti* Ω (ad esempio l'insieme di tutti i possibili "esiti" di una misura), una "σ-algebra" $\mathscr{A}$ di *eventi* osservabili (i sottoinsiemi "probabilizzabili" di Ω) e una *misura di probabilità* P, definita su $\mathscr{A}$, per cui valgano i noti assiomi.

Mentre è quasi sempre facile (e in ogni caso arbitrario) descrivere l'insieme dei possibili risultati di un esperimento con un insieme Ω e i relativi eventi che interessano per mezzo di una σ-algebra di sottoinsiemi di Ω (si pensi al lancio di un dado o alla misura della lunghezza di un tavolo) il processo attraverso cui si assegna P, ad eccezione di un numero limitatissimo di casi, non è a priori affatto ovvio.

Questo processo costituisce l'oggetto della statistica.

La statistica si occupa di assegnare probabilità sulla base dell'evidenza sperimentale. L'assegnazione di P a un dato spazio di esperimenti $\{\Omega, \mathscr{A}\}$ è un processo *induttivo* non oggettivabile a priori.

STIMA E VERIFICA D'IPOTESI

DATA una famiglia $\mathscr{P}$, o più famiglie *disgiunte* $\mathscr{P}_i$, $i = 1, \ldots, k$ (k *finito*), di possibili misure di probabilità P su $\{\Omega, \mathscr{A}\}$ e il risultato di un esperimento, $\bar{\omega}$, $\bar{\omega} \in \Omega$

Tradizionalmente si distinguono due tipi di problemi:

Problemi di stima: Sulla base del dato osservato $\bar{\omega}$, assegnare una probabilità ammissibile, i.e. un elemento $P = P(\bar{\omega}) \in \mathscr{P}$.

Problemi di verifica di ipotesi: Sulla base del dato osservato $\bar{\omega}$, assegnare P ad una delle classi $\mathscr{P}_i$ (i.e. "decidere" a quale classe $\mathscr{P}_i$ appartiene P).

ESEMPIO

Supponiamo di lanciare una moneta e sia p = probabilità **incognita** di avere "testa" (T) e $1-p$ = probabilità di avere "croce" (C). Si vogliono ricavare informazioni su p lanciando la moneta N volte consecutive.

Sia Ω = $\{$tutti i possibili esiti di N lanci successivi$\}$ tutte le successioni di N simboli "T" e "C" e $\mathscr{A}$ la famiglia di tutti i sottoinsiemi di Ω (che è, come è ben noto, un'algebra Booleana).
Descriviamo matematicamente il fatto che "ogni lancio non influenza l'esito del successivo" scegliendo a priori un classe di misure di probabilità $\mathscr{P} := \{P_p\}$ su $\{\Omega, \mathscr{A}\}$ definita per ogni evento elementare $\omega \in \Omega$ dalla

$$P_p(\{\omega\}) = p^{n(T)}\,(1-p)^{N-n(T)} \quad , \qquad 0 < p < 1 \quad ,$$

dove $n(T)$ è il numero di simboli "T" nella successione ω.

In questo caso la famiglia $\mathscr{P}$ è "parametrica", ovvero $\mathscr{P} := \left\{ P_p \,;\, 0 < p < 1 \right\}$.
Il problema della **stima di** P si riduce alla scelta di un valore "plausibile" di p in base all'osservazione dei risultati di N lanci successivi della moneta.

Problema di verifica di ipotesi: usare l'osservazione $\bar{\omega}$ per validare una convinzione a priori che si ha su p, ad esempio $p = 1/2$ (ovvero che T e C sono equiprobabili). In questo secondo caso bisogna decidere sulla base di $\bar{\omega}$ se P_p appartiene alla classe

$$\mathscr{P}_0 := \{P_{1/2}\} \quad ,$$

oppure se P_p sta in

$$\mathscr{P}_1 := \left\{P_p \,;\, p \neq 1/2\right\} \quad .$$

PROBLEMI PARAMETRICI

Si dicono *parametrici* quei problemi in cui $\mathscr{P}$ ha la forma

$$\mathscr{P} = \{P_\theta \,;\, \theta \in \Theta\} \quad ,$$

dove Θ è un sottoinsieme di uno spazio reale di dimensione finita, p, i.e. $\Theta \subseteq \mathbb{R}^p$.

Si parla allora di *stima del parametro* θ (che individua univocamente la misura di probabilità P) oppure di *verifica di ipotesi sul parametro* θ. In quest'ultimo caso si possono pensare assegnati k sottoinsiemi disgiunti (Θ_i, $i = 1, \ldots, k$) di Θ tali per cui $\mathscr{P}_i = \{P_\theta \mid \theta \in \Theta_i\}$, $i = 1, \ldots, k$. In corrispondenza, il problema di verifica di ipotesi diventa quello di decidere, in base ai dati osservati, a quale classe Θ_i appartiene θ.

Il problema della moneta considerato poco fa è appunto un problema parametrico. Qui Θ è l'intervallo $(0,1)$, $\Theta_0 = \{1/2\}$, $\Theta_1 = (0,1) - \{1/2\}$.

VARIABILI CASUALI

In questo corso ci occuperemo esclusivamente di probabilità *indotte da variabili casuali* o da famiglie (eventualmente infinite) di variabili casuali.
Sia $\mathbf{y} = [\mathbf{y}_1 \cdots \mathbf{y}_m]^\top$ una variabile aleatoria m-dimensionale definita su $\{\Omega, \mathscr{A}\}$ (cioè una funzione misurabile da Ω in $\mathbb{R}^m$). Se P è una probabilità definita su $\mathscr{A}$, ricordiamo che la *probabilità indotta da* $\mathbf{y}$, $P_\mathbf{y}$, è la misura di probabilità definita su $\{\mathbb{R}^m, \mathscr{B}^m\}$ ($\mathscr{B}^m :=$ σ-algebra di Borel di $\mathbb{R}^m$) ponendo

$$P_\mathbf{y}(E) := P\{\omega \mid \mathbf{y}(\omega) \in E\} = P\{\mathbf{y}^{-1}(E)\}$$

per ogni evento E di $\mathscr{B}^m$.
$P_\mathbf{y}$ è univocamente individuata dalla sua **funzione distribuzione di probabilità** (o di ripartizione) $F : \mathbb{R}^m \to [0,1]$

$$F(y) = F(y_1, \ldots, y_m) = P\{\omega \mid \mathbf{y}_1(\omega) \leq y_1, \ldots, \mathbf{y}_m(\omega) \leq y_m\}$$

Definiamo un nuovo spazio di probabilità

$$\tilde{\Omega} = \mathbb{R}^m \quad , \qquad \tilde{\mathscr{A}} = \mathscr{B}^m \quad , \qquad \tilde{P} = P_\mathbf{y} \quad ,$$

e la variabile

$$\tilde{\mathbf{y}} : \mathbb{R}^m \to \mathbb{R}^m \quad , \qquad \tilde{\mathbf{y}}_i(y_1, \ldots, y_m) := y_i \quad , \qquad i = 1, \ldots, m \quad ,$$

Allora *la variabile casuale* $\tilde{\mathbf{y}}$ *definita su* $\{\tilde{\Omega}, \tilde{\mathscr{A}}, P_{\mathbf{y}}\}$ *ha la stessa funzione di ripartizione* F *di* $\mathbf{y}$, ovvero

$$P_{\tilde{\mathbf{y}}} = P_{\mathbf{y}} \quad .$$

Pertanto $\tilde{\mathbf{y}}$ e $\mathbf{y}$ possono essere riguardate come la *stessa* variabile casuale.

Questa rappresentazione **sullo spazio dei valori campionari** di una variabile casuale è molto comoda perché permette di individuare $\mathbf{y}$ *assegnando solo la sua funzione di ripartizione*.
Così, quando si parla di una *variabile reale Gaussiana* di media μ e varianza σ^2 si intende la funzione identità

$$\mathbf{y} : \mathbb{R} \to \mathbb{R} \quad , \qquad \mathbf{y}(y) := y \quad ,$$

sullo spazio di probabilità $\{\Omega, \mathscr{A}, P\} = \{\mathbb{R}, \mathscr{B}, P_{\mathbf{y}}\}$ in cui

$$P_{\mathbf{y}}(E) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_E e^{-\frac{(y-\mu)^2}{2\sigma^2}} \, dy, \qquad E \in \mathscr{B}.$$

Lo spazio dei valori campionari è "tagliato su misura" per $\mathbf{y}$ e su di esso possono definirsi solo v.c. che sono *funzioni di* $\mathbf{y}$ (su questo spazio $\mathbf{y}$ è la funzione identità e ogni funzione su di esso può essere considerata dipendente da y attraverso l'identità, cioè $\mathbf{y}$). Su $\{\tilde{\Omega}, \tilde{\mathscr{A}}, P_{\mathbf{y}}\}$ non può, ad esempio, essere definita una v.c. (non costante) indipendente da $\mathbf{y}$.
D'ora in avanti: problemi di inferenza statistica in cui $\mathscr{P}$ (o $\{\mathscr{P}_i\}$) è una famiglia di *funzioni di ripartizione* su $\mathbb{R}^m$: $\mathscr{P} := \{F(\cdot)\}$.

Supporremo che la "forma" delle F sia nota a priori e che

$$\mathscr{P} = \left\{ F_\theta \mid \theta \in \Theta \right\} \quad .$$

Θ è il campo dei valori ammissibili dal parametro.
In questo schema i dati sperimentali $(\bar{\omega})$ sono determinazioni di una variabile aleatoria m-dimensionale $\mathbf{y}$ di distribuzione incognita F. Possiamo in generale immaginare che le componenti $\mathbf{y}_i$ di $\mathbf{y}$ rappresentino certe m grandezze fisiche simultaneamente misurabili (mutuamente interagenti) e di voler determinare una distribuzione di probabilità F che descriva in modo "plausibile" i risultati di un certo numero (sperabilmente grande) di

dati di misura che si hanno a disposizione. In genere supporremo di avere informazioni a priori sufficienti per scegliere a priori una certa famiglia di distribuzioni di probabilità che descrive i dati di misura.

CAMPIONI CASUALI

Come eseguire le misure in modo tale che esse diano la "massima informazione" sulla distribuzione (incognita) di $\mathbf{y}$?
Nel caso limite in cui le N misure fossero eseguite tutte esattamente nelle stesse condizioni sperimentali (cioè se le cause di errore accidentale fossero tutte *esattamente* le stesse nelle N prove) si avrebbe $y_1 = y_2 = \ldots = y_N$ e i dati relativi alla seconda, terza,..., N-sima misura sarebbero inutili.

Per questo motivo, è necessario cercare di predisporre l'esperimento in modo tale che le suddette cause di errore accidentale siano il più possibile tra loro diverse nelle diverse misure. Il modello probabilistico che vogliamo costruire (cioè F) deve descrivere proprio (gli effetti di) queste cause d'errore.

Definizione 1 *Sia $\mathscr{P} = \{F\}$ una famiglia di distribuzioni di probabilità su $\mathbb{R}^m$ e siano $\mathbf{y}_1, \ldots, \mathbf{y}_N$ vettori casuali m-dimensionali aventi la stessa distribuzione F e mutuamente indipendenti per ogni F nella classe $\mathscr{P}$. Si dice allora che $\mathbf{y}_1, \ldots, \mathbf{y}_N$ sono un "campione casuale" di numerosità N relativo alla classe $\mathscr{P}$ (o "estratto" da $\mathscr{P}$).*

Si può intuire che un campione casuale fornisce la "massima informazione" sulla distribuzione di probabilità incognita. Precisare meglio questa affermazione richiederebbe una lunga digressione e l'introduzione di concetti che vengono introdotti in altri contesti, per cui noi la lascieremo un pò nel vago.

Se F è un elemento di una famiglia parametrica $\{F_\theta\,;\,\theta \in \Theta\}$ la distribuzione congiunta del campione casuale si può scrivere allora, per ogni $\theta \in \Theta$, come:

$$F_\theta^N(y_1, \ldots, y_N) = F_\theta(y_1), \ldots, F_\theta(y_N) \quad , \qquad y_t \in \mathbb{R}^m \quad .$$

Esistono metodi opportuni per eseguire le misure in modo tale da avvicinarsi il più possibile alla situazione ideale del campione casuale. Di essi si occupa la *teoria dei campioni*.

Notiamo comunque fin da adesso che nelle situazioni che ci interessano, il "modo" in cui si eseguono le misure non è sotto il controllo dello statistico. Spesso i dati vengono forniti sotto forma di "serie storica" ed esiste una chiara evidenza che $(y_1, \ldots, y_t)$ "influenzano" il dato successivo y_{t+1}. In questi casi si è in presenza di fenomeni *dinamici* e l'ipotesi di campione casuale, su cui è basata larga parte della teoria statistica classica (che è una teoria essenzialmente statica) non è valida.

STATISTICHE

Definizione 2 *Sia* $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ *un campione estratto da una d.d.p.* F *incognita, appartenente a una famiglia parametrica* $\{F_\theta\,;\, \theta \in \Theta\}$. *Si chiama* statistica *una qualunque funzione (misurabile)* ϕ*, a valori vettoriali,*

$$\phi : \mathbb{R}^m \times \cdots \times \mathbb{R}^m \to \mathbb{R}^q \quad ,$$

che non dipende dal parametro θ.

(Si noti che il campione non è necessariamente casuale).

Una statistica può sempre essere interpretata come una *funzione del campione*. (In seguito scriveremo spesso $\phi(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ al posto di ϕ). Essa è pertanto una *variabile casuale* la cui distribuzione si può ricavare dalla F_θ^N attraverso le regole del calcolo delle probabilità. Esempi semplici ma molto importanti sono i seguenti*

*In statistica è spesso di grande interesse studiare le proprietà di una statistica al variare della numerosità campionaria N. Per questo motivo, quando servirà mettere in evidenza la dipendenza di una statistica ϕ da N useremo la notazione ϕ_N.

La *media campionaria*, $\bar{\mathbf{y}}_N$,

$$\bar{\mathbf{y}}_N = \frac{1}{N}\sum_{t=1}^{N} \mathbf{y}_t \quad ;$$

questa è una statistica a valori in $\mathbb{R}^m$. Se le $\{\mathbf{y}_t\}$ sono un campione casuale estratto da una d.d.p. incognita F_{θ_0} con $F_{\theta_0} \in \{F_\theta\,;\,\theta \in \Theta\}$ si ha $\mathbb{E}_0\bar{\mathbf{y}}_N = \mathbb{E}_0\mathbf{y}$, dove $\mathbb{E}_0$ denota l'operatore di media rispetto alla distribuzione F_{θ_0}. Anticipiamo il fatto notevole, che scende dalla legge dei grandi numeri (che verrà richiamata più avanti), che il limite

$$\lim_{N\to\infty} \frac{1}{N}\sum_{1}^{N} \mathbf{y}_t$$

esiste *con probabilità 1* e vale $\mathbb{E}_0\mathbf{y} = \int_{\mathbb{R}^m} y\, dF_{\theta_0}(y)$. In altre parole il limite

$$\lim_{N\to\infty} \frac{1}{N}\sum_{1}^{N} \mathbf{y}_t$$

esiste per "quasi tutti" i possibili risultati delle misure ripetute $\{y_1, y_2, \ldots,$

$y_t, \ldots\}$ ed è uguale proprio alla *media* $\mathbb{E}_0\mathbf{y}$ della distribuzione F_{θ_0}. Questo spiega l'origine del nome attribuito a $\bar{\mathbf{y}}_N$.

La *varianza campionaria*,

$$\hat{\mathbf{\Sigma}}_N := \frac{1}{N} \sum_{t=1}^{N} (\mathbf{y}_t - \bar{\mathbf{y}}_N) (\mathbf{y}_t - \bar{\mathbf{y}}_N)^\top$$

è una statistica a valori in $\mathbb{R}^{m \times m}$ (una matrice aleatoria).

La varianza campionaria di un campione casuale gode di proprietà asintotiche analoghe a $\bar{\mathbf{y}}_N$. In effetti se $\{\mathbf{y}_t\}$ è un campione casuale estratto da F_{θ_0}, il limite

$$\lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} (\mathbf{y}_t - \bar{\mathbf{y}}_N) (\mathbf{y}_t - \bar{\mathbf{y}}_N)'$$

esiste ancora con probabilità 1 (per "quasi tutte" le possibili successioni di risultati di misura $\{y_t\}$) e vale

$$\mathbb{E}_0(\mathbf{y} - \mathbb{E}_0\mathbf{y}) (\mathbf{y} - \mathbb{E}_0\mathbf{y})^\top \quad ,$$

che è proprio la matrice delle varianze del vettore $\mathbf{y}$ (oppure della d.d.p. F_{θ_0}).

Nel seguito useremo la notazione $\mathbf{y} \sim \{F_\theta\}$ per intendere che $\mathbf{y}$ è distribuita secondo una d.d.p. incognita appartenente alla famiglia parametrica $\{F_\theta\,;\, \theta \in \Theta\}$.

Esempio : *Sia $\mathbf{y} \sim \mathcal{N}(\mu, \Sigma)$ (la distribuzione normale di media $\mu \in \mathbb{R}^m$ e varianza $\Sigma \in \mathbb{R}^{m\times m}$) con $\mu = \theta$ incognito. Sia inoltre $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ un campione casuale estratto da $\mathcal{N}(\theta, \Sigma)$: allora*

$$V^2 := \frac{1}{N} \sum_1^N (\mathbf{y}_t - E\mathbf{y}) (\mathbf{y}_t - E\mathbf{y})^\top$$

dipende da $\theta = E\mathbf{y}$ e non è *una statistica.*

FISHER/BAYES

Nell' approccio "classico" o"Fisheriano" (da R.A. Fisher, uno dei padri fondatori della statistica) si postula che il parametro θ sia una grandezza deterministica costante, in linea di principio suscettibile di essere determinata esattamente (idealmente, ad esempio, mediante una serie infinita di esperimenti indipendenti). Questo punto di vista può essere accettabile se θ ha un significato strumentale, legato alla descrizione matematica del fenomeno che si vuol modellare (ad esempio la varianza di una distribuzione Gaussiana o i coefficienti di un'equazione differenziale o alle differenze). Benchè non sia a priori necessario, spesso nell'approccio classico si postula l'esistenza di un *valore vero*, θ_0, del parametro (ovviamente incognito), in corrispondenza al quale si ha una descrizione probabilistica "esatta" dei dati di misura da parte della d.d.p. F_{θ_0}, chiamata la distribuzione "vera" dei dati. L'esistenza del parametro vero, nonostante possa a volte essere un'utile idea per schematizzare le proprietà di certe procedure statistiche, è sempre da considerarsi un'ipotesi irrealistica.

L'approccio Fisheriano diventa un poco questionabile se il parametro descrive invece il valore di grandezze fisiche che si stanno misurando. Ogni grandezza fisica è infatti conoscibile solo in modo approssimato, sia a causa di inevitabili interazioni e disturbi dell'apparato di misura con l'ambiente esterno, sia perchè spesso all'aumentare della sensibilità del procedimento di misura (al limite ad esempio arrivando a scala atomica o subatomica oppure a livello di segnale nei sistemi di trasmissione e calcolo moderni), la nozione stessa di valore numerico da attribuire a variabili fisiche come lunghezza, peso, tensione, corrente etc, perde di significato.

Secondo la la *teoria Bayesiana* θ è da riguardarsi *sempre* come una variabile casuale (che denoteremo col simbolo $\mathbf{x}$) e la famiglia $\{F_\theta\,;\,\theta \in \Theta\}$ come una distribuzione di probabilità *condizionata*, dati i possibili valori che $\mathbf{x}$ potrebbe assumere. Quindi in questo contesto si pone

$$F_\theta(\cdot) \equiv F(\cdot \mid \mathbf{x} = \theta)\,.$$

Questa identificazione è sempre possibile (basta che $F_\theta(y)$, funzione delle due variabili y e θ, sia misurabile rispetto a θ). Rimane però aperta la

questione della conoscenza della distribuzione di $\mathbf{x}$, che viene chiamata *distribuzione a priori* del parametro. In molti problemi di misura tale distribuzione è approssimativamente nota e in questo caso il punto di vista Bayesiano permette di ridurre l'inferenza statistica su $\mathbf{x}$ a un puro problema di calcolo delle probabilità.

In altri casi la distribuzione a priori del parametro non è nota. Questo fatto può essere tradotto dicendo che l'informazione a priori disponibile per risolvere il problema di inferenza è *minore*. In queste situazioni è naturale seguire l'approccio Fisheriano. In ultima analisi i due approcci portano a impostazioni del problema di inferenza in presenza di *diversa conoscenza a priori*.

Come vedremo meglio nel seguito, l'approccio Fisheriano è in linea generale quello che riflette in modo più verosimile il tipo di informazione a priori che è disponibile nei problemi di modellistica cosiddetti a "scatola nera". In questi problemi si cerca di descrivere i dati osservati per mezzo di modelli probabilistici da scegliersi all'interno di famiglie parametriche che hanno

struttura assegnata a priori. Di norma i parametri non hanno in questo caso un significato fisico ed è naturale impostare il problema prescindendo da informazioni a priori che in pratica sono assai raramente disponibili.

In questo corso ci si riferirà di norma all'impostazione Fisheriana.

STIMA PARAMERICA

Sia $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ un campione estratto da una distribuzione (incognita) della famiglia $\{F_\theta\,;\, \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^p$.

Definizione 3 *Si chiama* **stimatore** *di* θ *una qualunque statistica* ϕ *a valori in* Θ*. Il valore assunto da* $\phi(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ *in corrispondenza ai valori campionari* $(y_1, \ldots, y_N)$ *di* $\mathbf{y}_1, \ldots, \mathbf{y}_N$*,*

$$\hat{\theta} = \phi(y_1, \ldots, y_N) \quad ,$$

si chiama **stima** *di* θ*, basata sui dati* $(y_1, \ldots, y_N)$*.*

Ovviamente si vorrebbe che le stime calcolate in base ai dati fossero "vicine" al valore vero, θ_0, del parametro. In particolare si vorrebbe che la media d'insieme delle stime ottenute in corrispondenza a varie serie di misure, $(y'_1, \ldots, y'_N)$, $(y''_1, \ldots, y''_N)$, ..., fosse proprio θ_0. Questa condizione si può esprimere scrivendo

$$\mathbb{E}_{\theta_0}\, \phi(\mathbf{y}_1, \ldots, \mathbf{y}_N) = \theta_0 \quad ,$$

dove $\mathbb{E}_{\theta_0}$ è l'operatore di media corrispondente alla distribuzione *vera*, $F^N_{\theta_0}$, del campione. Dato che θ_0 è incognito occorre chiedere che la valga per tutti i possibili valori del parametro. Si arriva così alla definizione seguente,

Definizione 4 *Uno stimatore ϕ si dice (uniformemente)* **corretto**[*] *se*

$$\mathbb{E}_\theta\, \phi(\mathbf{y}_1, \ldots, \mathbf{y}_N) = \theta \quad , \qquad \forall \theta \in \Theta \quad .$$

Un buono stimatore dovrebbe inoltre fornire valori molto concentrati attorno alla media, avere cioè una bassa dispersione. Naturalmente perchè l'idea di minima dispersione abbia senso occorre restringere a priori la classe degli stimatori ammissibili. Infatti se si prendesse lo stimatore $\phi = costante$ (deterministico, ad es. la funzione nulla) questo avrebbe ovviamente dispersione (o varianza) nulla.

Definizione 5 *Uno stimatore ϕ si dice (uniformemente) a* **minima varianza** *nella classe $\mathscr{C}$ se la varianza di ϕ*

$$\mathrm{var}_\theta(\phi) := \mathbb{E}_\theta (\phi - \mathbb{E}_\theta\, \phi)^\top\, (\phi - \mathbb{E}_\theta\, \phi)$$

[*]In inglese *unbiased*.

è la più piccola fra le varianze di tutti gli stimatori della classe $\mathscr{C}$, ovvero se

$$\mathrm{var}_\theta(\phi) \leq \mathrm{var}_\theta(\psi) \quad , \qquad \forall \psi \in \mathscr{C} \quad ,$$

per tutti i $\theta \in \Theta$.

Come abbiamo già visto, per evitare banalità occorre restringere la classe $\mathscr{C}$ a una opportuna sottoclasse di tutte le funzioni misurabili dei dati. Vedremo tra poco che se si prende per $\mathscr{C}$ la classe degli stimatori *corretti* di θ, non sono possibili situazioni degeneri del tipo appena visto. Questo scende da una celebre disuguaglianza, detta di Cramèr-Rao.

DISUGUAGLIANZA DI CRAMÈR-RAO

Sia $\mathbf{x}$ vettore aleatorio n-dimensionale con $\mathbf{x} \sim \{F_\theta\,;\, \theta \in \Theta\}$. ($\mathbf{x}$ potrebbe in particolare essere un campione casuale $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$, ma la disuguaglianza di Cramèr-Rao non richiede l'indipendenza delle componenti di $\mathbf{x}$).

A.1) F_θ ammette una densità $p(\cdot, \theta)$ derivabile (parzialmente) *due volte* rispetto a θ.

A.2) Per ogni statistica ϕ con $\mathbb{E}_\theta \phi < \infty$,

$$\frac{\partial}{\partial \theta_i} \int_{\mathbb{R}^n} \phi(x)\, p(x,\theta)\, dx = \int_{\mathbb{R}^n} \phi(x)\, \frac{\partial}{\partial \theta_i}\, p(x,\theta)\, dx, \qquad i = 1, \ldots, p,\ \forall\, \theta \in \Theta.$$

In particolare $\displaystyle \frac{\partial}{\partial \theta_i} \int_{\mathbb{R}^n} p(x,\theta)\, dx = \int_{\mathbb{R}^n} \frac{\partial}{\partial \theta_i}\, p(x,\theta)\, dx, \quad i = 1, \ldots p.$

A.3) $\displaystyle \frac{\partial^2}{\partial \theta_i\, \partial \theta_j} \int_{\mathbb{R}^n} p(x,\theta)\, dx = \int_{\mathbb{R}^n} \frac{\partial^2}{\partial \theta_i\, \partial \theta_j}\, p(x,\theta)\, dx \qquad i, j = 1, \ldots, p\ \forall\, \theta \in \Theta.$

Definizione 6 *Vettore aleatorio delle* **sensitività** *del modello parametrico* $\{p_\theta\, \theta \in \Theta\}$*:*

$$\mathbf{z}_\theta := \left[\frac{\partial \log p(\mathbf{x},\theta)}{\partial \theta_i}\right]_{i=1,\ldots,p} = \left[\frac{\frac{\partial\, p(\mathbf{x},\theta)}{\partial \theta_i}}{p(\mathbf{x},\theta)}\right]_{i=1,\ldots,p}$$

La **matrice di informazione di Fisher** $I(\theta)$*, del modello parametrico:*

$$I(\theta) := \mathbb{E}_\theta\left[\mathbf{z}_\theta \mathbf{z}_\theta^\top\right] = \left[\mathbb{E}_\theta\left(\frac{\partial \log p(\mathbf{x},\theta)}{\partial \theta_i}\cdot\frac{\partial \log p(\mathbf{x},\theta)}{\partial \theta_j}\right)\right]_{i,j=1,\ldots,p} \tag{1}$$

che si può anche scrivere come

$$I(\theta) = \left[-\mathbb{E}_\theta\,\frac{\partial^2 \log p(\mathbf{x},\theta)}{\partial \theta_i\,\partial \theta_j}\right]_{i,j=1,\ldots,p} . \tag{2}$$

ed è una matrice almeno semidefinita positiva.

L'equivalenza di (2) e (1) si ricava da $\int p(x,\theta)\,dx = 1$ (costante rispetto a θ) e derivando questa uguaglianza membro a membro si trova

$$\begin{aligned}
\int_{\mathbb{R}^n} \frac{\partial p(x,\theta)}{\partial \theta_i}\,dx &= 0 \quad , \qquad i = 1,\ldots,p \quad , \\
\int_{\mathbb{R}^n} \frac{\partial^2 p(x,\theta)}{\partial \theta_i\,\partial \theta_j}\,dx &= 0 \quad , \qquad i,j = 1,\ldots,p \quad .
\end{aligned}$$

Dalla prima segue immediatamente che $\mathbb{E}_\theta \frac{\partial \log p}{\partial \theta_i} = 0$ per tutti gli i e quindi

$$\mathbb{E}_\theta\, \mathbf{z}_\theta = 0$$

e pertanto $I(\theta)$ *è la varianza di* $\mathbf{z}_\theta$. La (2) scende allora subito dalla

$$-\frac{\partial^2 \log p}{\partial \theta_i\,\partial \theta_j} = \frac{\partial \log p}{\partial \theta_i}\,\frac{\partial \log p}{\partial \theta_j} - \frac{1}{p}\,\frac{\partial^2 p}{\partial \theta_i\,\partial \theta_j} \quad ,$$

usando la (3).

DISUGUAGLIANZA DI CRAMÈR-RAO

Teorema 1 (Disuguaglianza di Cramèr-Rao) *Sia g una funzione derivabile da Θ in $\mathbb{R}^q$ e ϕ uno stimatore corretto di $g(\theta)$. Sia $V(\theta)$ la matrice varianza di ϕ e $G(\theta)$ la matrice jacobiana di g*

$$G(\theta) = \left[\frac{\partial g_i(\theta)}{\partial \theta_j}\right]_{\substack{i=1,\ldots,q \\ j=1,\ldots,p}} .$$

Se la matrice di Fisher $I(\theta)$ è invertibile si ha

$$V(\theta) - G(\theta)\, I^{-1}(\theta)\, G^{\top}(\theta) \geq 0 \quad , \tag{3}$$

dove ≥ 0 significa che la matrice a primo membro è semidefinita positiva.

La dimostrazione è basata sulla formula per la varianza d'errore dello stimatore lineare Bayesiano $\hat{\phi}(\mathbf{x}) := \hat{\mathbb{E}}_{\theta}[\,\phi(\mathbf{x}) \mid \mathbf{z}_{\theta}\,]$ del vettore $\phi(\mathbf{x})$ dato il vettore $\mathbf{z}_{\theta}$

$$\mathrm{Var}_{\theta}\{\phi(\mathbf{x}) - \hat{\phi}(\mathbf{x})\} = \mathrm{Var}_{\theta}\{\phi(\mathbf{x})\} - \mathrm{Cov}_{\theta}\{\phi(\mathbf{x}), \mathbf{z}_{\theta}\}\mathrm{Var}_{\theta}\{\mathbf{z}_{\theta}\}^{-1}\mathrm{Cov}_{\theta}\{\phi(\mathbf{x}), \mathbf{z}_{\theta}\}^{\top}.$$

Dato che $\phi(\mathbf{x})$ è uno stimatore corretto di $g(\theta)$; i.e.

$$\int_{\mathbb{R}^n} \phi(x)\; p(x,\theta)\, dx = g(\theta) \quad , \qquad \forall \theta \in \Theta \quad ,$$

applicando la A.3) si ottiene

$$\mathbb{E}_\theta\, \phi(\mathbf{x})\, \mathbf{z}_\theta^j \;\; = \;\; \int_{\mathbb{R}^n} \phi(x)\, \frac{\partial p(x,\theta)}{\partial \theta_j} \cdot \frac{1}{p(x,\theta)} \cdot p(x,\theta)\, dx = \frac{\partial g(\theta)}{\partial \theta_j} \quad ,$$

$$j = 1, \ldots, p \quad ,$$

e quindi $\frac{\partial g(\theta)}{\partial \theta_j}$ è la j-sima colonna della matrice di covarianza di ϕ e $\mathbf{z}_\theta$,

$$\mathbb{E}_\theta\, \phi(\mathbf{x})\, \mathbf{z}_\theta^\top = \mathbb{E}_\theta\, \phi(\mathbf{x})\, [\mathbf{z}_\theta^1, \ldots, \mathbf{z}_\theta^p] \quad ,$$

ovvero

$$\mathbb{E}_\theta\, \phi\, \mathbf{z}_\theta^\top = G(\theta) \quad .$$

Per concludere basta allora notare che la matrice varianza del vettore aleatorio $\phi(\mathbf{x}) - G(\theta)\, I(\theta)^{-1}\, \mathbf{z}_\theta$ è semidefinita positiva.

CONSEGUENZE

Se ϕ è uno stimatore corretto di θ (cioè se g è l'identità) si ha $G(\theta) = I\ (p \times p)$ e pertanto la (3) diventa

$$V(\theta) - I(\theta)^{-1} \geq 0 \quad .$$

Notiamo che la varianza scalare $\mathrm{var}_\theta(\phi) = \sum_1^p \mathbb{E}_\theta(\phi_i - \theta_i)^2$ è proprio la traccia della matrice $V(\theta)$. Dato che

$$\mathrm{Tr}\, V(\theta) - \mathrm{Tr}\, I^{-1}(\theta) = \mathrm{Tr}\left[V(\theta) - I^{-1}(\theta)\right] \geq 0$$

(la traccia di una matrice è la somma degli autovalori e una matrice semidefinita positiva ha autovalori ≥ 0) si ricava che la varianza scalare di uno stimatore corretto del parametro θ non può essere inferiore al numero positivo $\mathrm{Tr}\, I(\theta)^{-1}$,

$$\mathrm{var}_\theta(\phi) \geq \mathrm{Tr}\left[I(\theta)^{-1}\right] \quad , \qquad \forall\, \theta \quad .$$

Esiste quindi un limite inferiore per la varianza di ogni stimatore *corretto, indipendente dal criterio di stima adottato*.

Non è detto che il limite inferiore di Cramèr-Rao sia la maggiorazione migliore possibile. Può benissimo darsi che uno stimatore abbia varianza *strettamente* più grande di $\mathrm{Tr}[I(\theta)^{-1}]$ e sia ugualmente lo stimatore (corretto!) a minima varianza.

ESEMPI

Supponiamo di avere un campione casuale di numerosità N estratto da una distribuzione Gaussiana $\mathcal{N}(\theta, \sigma^2)$ con σ^2 nota. Si ha a che fare con un vettore aleatorio $\mathbf{x} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ $(r = N)$ e

$$p(y_1, \ldots, y_N; \theta) = \prod_{t=1}^{N} p(y_t; \theta) \quad .$$

Dato che $\log p(y; \theta) = C - \frac{1}{2} \frac{(y-\theta)^2}{\sigma^2}$ si trova

$$\log p(y_1, \ldots, y_N; \theta) = N \times Const - \frac{1}{2} \sum_{t=1}^{N} \frac{(y_t - \theta)^2}{\sigma^2} \quad ,$$

$$\frac{d \log p}{d\theta} = \sum_{t=1}^{N} \frac{y_t - \theta}{\sigma^2} \quad ,$$

e, usando l'indipendenza delle variabili del campione, si verifica facilmente

che,

$$I(\theta) = \mathbb{E}_\theta \left[\frac{d \log p(\mathbf{y}; \theta)}{d\theta} \right]^2 = \frac{1}{\sigma^4} \cdot N\sigma^2 = \frac{N}{\sigma^2} \quad .$$

Consideriamo la media campionaria

$$\bar{\mathbf{y}}_N = \frac{1}{N} \sum_{t=1}^{N} \mathbf{y}_t$$

che è distribuita come $\mathcal{N}(\theta, \sigma^2/N)$. Ovviamente $\bar{\mathbf{y}}_N$ è uno stimatore corretto di θ e la sua varianza vale σ^2/N, uguale all'inversa della matrice di informazione di Fisher. La media campionaria è quindi il miglior stimatore possibile di θ (nel caso Gaussiano). Si dice che uno stimatore corretto per cui $V(\theta) = I(\theta)^{-1}$ è *efficiente*.

INTERPRETAZIONE DI $I(\theta)$

Cerchiamo di caratterizzare quantitativamente lo scostamento fra due variabili casuali $\mathbf{x}_1 \sim p(\cdot, \theta_1)$ e $\mathbf{x}_2 \sim p(\cdot, \theta_2)$ e avere in questo modo una misura della capacità che hanno le osservazioni di *discriminare* valori diversi del parametro θ.

Si definisce *pseudo-distanza di Kullback-Leibler* tra due densità f e p, il numero

$$K(f,p) := \int_{\mathbb{R}^n} [\log f - \log p]\, f(x)\, dx = \int_{\mathbb{R}^n} \log f/p\; f(x)\, dx = \mathbb{E}_f \log f/p; \quad (4)$$

È immediato verificare che $K(f,p) = 0$ solo nel caso in cui $f = p$. Inoltre dalla disuguaglianza di Jensen:

$$\int \log g(x)\, d\mu \leq \log\{\int g(x)\, d\mu\}$$

valida per ogni $g(x) > 0$ rispetto ad una arbitraria misura μ, si ricava

$$-K(f,p) = \int_{\mathbb{R}^n} \log \frac{p}{f} f dx \leq \log\{\int_{\mathbb{R}^n} \frac{p}{f} f\, dx\} = \log\{1\} = 0$$

e pertanto $K(f,p) \geq 0$.
Per questo motivo $K(f,p)$ può essere preso come misura della deviazione fra le due densità f e p; da notare però che $K(f,p)$ non è una vera metrica perché $K(p,f) \neq K(f,p)$ e non soddisfa la disuguaglianza triangolare. In teoria dell'informazione $K(f,p)$ viene chiamata *divergenza* e indicata col simbolo $D(f\|p)$. Per approfondire questi concetti si può vedere l'articolo in Wikipedia

`http://en.wikipedia.org/wiki/KullbackLeibler_divergence`

e la relativa bibliografia.

Supponiamo ora che p soddisfi alle stesse ipotesi di regolarità usate nella sezione **??** e poniamo $f \equiv p(\cdot,\theta_0)$ e $p \equiv p(\cdot,\theta)$, $\theta_0,\theta \in \Theta$. Chiameremo $K(\theta_0,\theta)$ la quantità $K(p(\cdot,\theta_0),p(\cdot,\theta))$. Ponendo $\theta = \theta_0 + \Delta\theta$ si ha

$$K(\theta_0,\theta) = K(\theta_0,\theta_0) + \left.\frac{\partial K}{\partial \theta}\right|_{\theta_0} \Delta\theta + \frac{1}{2}\Delta\theta^\top \left[\frac{\partial^2 K}{\partial\theta_i\,\partial\theta_j}\right]_{\theta_0} \Delta\theta + o(\|\Delta\theta\|^2)\,.$$

Notiamo subito che $K(\theta_0, \theta_0) = 0$ e inoltre

$$\frac{\partial K}{\partial \theta_i} = -\int_{\mathbb{R}^n} p(x, \theta_0) \, \frac{\partial \log p(x, \theta)}{\partial \theta_i} \, dx \quad ,$$

di modo che,

$$\left.\frac{\partial K}{\partial \theta_i}\right|_{\theta_0} = -\int_{\mathbb{R}^n} \left[\frac{\partial p(x, \theta)}{\partial \theta_i}\right]_{\theta_0} dx = 0$$

per tutti gli $i = 1, \ldots, p$.

Nello stesso modo si verifica poi che

$$\left.\frac{\partial^2 K}{\partial \theta_i \, \partial \theta_j}\right|_{\theta_0} = -\int_{\mathbb{R}^n} p(x, \theta_0) \, \left[\frac{\partial^2 \log p(x, \theta)}{\partial \theta_i \, \partial \theta_j}\right]_{\theta_0} dx = -\mathbb{E}_{\theta_0} \, \left[\frac{\partial^2 \log p(x, \theta)}{\partial \theta_i \, \partial \theta_j}\right]_{\theta_0}$$

e quindi il primo membro di questa relazione è l'elemento di posto (i, j) della matrice di Fisher $I(\theta_0)$. Quindi, per piccole variazioni del parametro θ, si ha

$$K(\theta_0, \theta) \cong \frac{1}{2} \Delta\theta^\top \, I(\theta_0) \, \Delta\theta \quad ; \tag{5}$$

MORALE

Per piccoli scostamenti $\Delta\theta$ del parametro dal valore di riferimento θ_0, la distanza (di Kullback) fra le due densità $p(\cdot, \theta)$ e $p(\cdot, \theta_0)$ è una forma quadratica la cui matrice peso è proprio la matrice di Fisher $I(\theta_0)$.

Nella prossima sezione vedremo una conseguenza notevole di questo fatto.

Problema :

Calcolare la distanza di Kullback-Leibler tra le due densità Gaussiane, $f \equiv \mathcal{N}(\mu\,,\,\sigma_0^2)$ e $p \equiv \mathcal{N}(\mu\,,\,\sigma^2)$.

Soluzione: Stante che $f \equiv \mathcal{N}(\mu\,,\,\sigma_0^2)$ e $p \equiv \mathcal{N}(\mu\,,\,\sigma^2)$, si ha

$$\log\frac{f}{p} = \frac{1}{2}\log\frac{\sigma^2}{\sigma_0^2} - \frac{1}{2}(\mathbf{y}-\mu)^2\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma^2}\right)$$

per cui la distanza di Kullback-Leibler tra le due densità è

$$\frac{1}{2}\log\frac{\sigma^2}{\sigma_0^2} - \frac{1}{2}\mathbb{E}_f(\mathbf{y}-\mu)^2\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma^2}\right) = \frac{1}{2}\left(\log\frac{\sigma^2}{\sigma_0^2} + \left(\frac{\sigma_0^2}{\sigma^2} - 1\right)\right)$$

Notare che $(x-1) - \log x \geq 0$ per $x > 0$ e uguale a zero solo se $x = 1$..

IDENTIFICABILITÀ

Supponiamo che $\Theta = \mathbb{R}^2$ e che F_θ dipenda da (θ_1, θ_2) solo attraverso il loro prodotto, ad esempio

$$F_\theta \sim \mathcal{N}(\theta_1\theta_2, \sigma^2)$$

Sia $\bar{\theta} = (\bar{\theta}_1, \bar{\theta}_2)'$ un valore fissato dal parametro. È evidente che $\hat{\theta} = \left(\alpha\bar{\theta}_1, \frac{1}{\alpha}\bar{\theta}_2\right)'$, $\alpha \neq 0$, è tale per cui $F_{\bar{\theta}}(x) = F_{\hat{\theta}}(x)$, $\forall x$
quindi qualunque campione estratto da F_θ, qualunque sia la sua numerosità, non sarà *mai* in grado di discriminare fra $\bar{\theta}$ e $\hat{\theta}$.

Definizione 7 *Due valori del parametro θ_1 e θ_2 in Θ si dicono "indistinguibili" se $F_{\theta_1}(x) = F_{\theta_2}(x)$, $\forall x \in \mathbb{R}^n$.*

È evidente che la relazione di indistinguibilità, che nel seguito indicheremo col simbolo "$\simeq$", è una relazione di equivalenza su Θ (essa è infatti simmetrica, riflessiva e transitiva). Come tale essa partiziona Θ in classi

di equivalenza $[\theta] := \{\theta' \mid \theta' \simeq \theta\}$ tali che $F_{\theta'} = F_{\theta''}$ se e solo se θ' e θ'' appartengono alla stessa classe $[\theta]$.

Definizione 8 *La famiglia $\{F_\theta\}$ (qualche volta, impropriamente, si dice che* il parametro $\theta \in \Theta$*) è* **globalmente identificabile** *se $\theta' \simeq \theta''$, o, equivalentemente, $F_{\theta'} = F_{\theta''}$, implica $\theta' = \theta''$ per tutti i θ', θ'' in Θ.*

Quindi, la famiglia parametrica $\{F_\theta\}$ (ovvero, il parametro θ), è globalmente identificabile se le classi di equivalenza si riducono a punti in Θ.
Per le applicazioni alla stima parametrica, la condizione di identificabilità globale è in generale troppo restrittiva ed in realtà è sufficiente una condizione di tipo locale.

Definizione 9 *La famiglia $\{F_\theta\,;\, \theta \in \Theta\}$ è* **localmente identificabile** *in θ_0, se esiste un intorno aperto di θ_0 che non contiene valori di θ indistinguibili da θ_0 (tranne θ_0 stesso).*

Problemi di identificabilità sorgono normalmente solo quando si ha a che fare con strutture parametriche abbastanza complesse e nella statistica

parametrica classica, questi concetti giocano un ruolo molto limitato. Invece nelle applicazioni moderne, ad esempio in econometria, nell'identificazione di sistemi dinamici specie se a più ingressi e più uscite, lo studio di identificabilità e la ricerca di parametrizzazioni identificabili costituiscono un problema fondamentale.
Esiste una notevole relazione tra identificabilità (locale) e non singolarità della matrice di Fisher. Questa relazione è messa in luce dal seguente teorema.

Teorema 2 (Rothenberg) *Siano valide le ipotesi A.1, A.2, A.3. Allora* θ_0 *è localmente identificabile se e solo se* $I(\theta_0)$ *è non singolare.*

[Cenno di prova] Come abbiamo visto, per piccoli scostamenti $\Delta\theta$ del parametro θ dal valore di riferimento θ_0, la distanza di Kullback fra le due densità $p(\cdot, \theta)$ e $p(\cdot, \theta_0)$ è la forma quadratica $\frac{1}{2}\Delta\theta' \, I(\theta_0) \, \Delta\theta$. Ne segue che in ogni intorno di θ_0 si possono avere valori del parametro $\theta \neq \theta_0$ per cui $p(\cdot, \theta) = p(\cdot, \theta_0)$ se e solo se $I(\theta_0)$ è singolare. □

Tornando al nostro esempio, si ha

$$I(\theta) = \mathbb{E}_\theta \begin{bmatrix} \frac{(\mathbf{x}-\theta_1\theta_2)^2}{\sigma^4}\theta_2^2 & \frac{(\mathbf{x}-\theta_1\theta_2)^2}{\sigma^4}\theta_1\theta_2 \\ \frac{(\mathbf{x}-\theta_1\theta_2)^2}{\sigma^4}\theta_1\theta_2 & \frac{(\mathbf{x}-\theta_1\theta_2)^2}{\sigma^4}\theta_1^2 \end{bmatrix} = \frac{1}{\sigma^2}\begin{bmatrix} \theta_2^2 & \theta_1\theta_2 \\ \theta_1\theta_2 & \theta_1^2 \end{bmatrix}.$$

Si vede che $\det I(\theta) = 0$, $\forall \theta \in \mathbb{R}^2$ e quindi θ non è mai identificabile.

MASSIMA VEROSIMIGLIANZA

Sia $\mathbf{x}$ un vettore aleatorio a valori in $\mathbb{R}^n$ (che, in particolare, ma non necessariamente, potrebbe essere un campione casuale di numerosità N) distribuito con densità incognita appartenente alla famiglia parametrica $\{p(\cdot,\theta)\,;\,\theta\in\Theta\}$. Sia x_0 un valore osservato di $\mathbf{x}$.

Definizione 10 *La* funzione di verosimiglianza *dell'osservazione* x_0 *è la funzione* $L(x_0,\cdot):\Theta\to R_+$ *(i reali non negativi) definita ponendo*

$$L(x_0,\theta) := p(x_0,\theta) \quad . \tag{6}$$

Il "principio della massima verosimiglianza", introdotto da Gauss nel 1856 e successivamente popolarizzato da R.A. Fisher, suggerisce di assumere come stima di θ, corrispondente all'osservazione x_0, il vettore $\hat{\theta}\in\Theta$ che massimizza $L(x_0,\cdot)$

$$L(x_0,\hat{\theta}) = \max_{\theta\in\Theta} L(x_0,\theta) \quad ;$$

supponendo implicitamente che il massimo esista. Il valore del parametro $\hat{\theta}$ è quindi quello che rende "a posteriori" più probabile l'osservazione x_0.

Seguendo questo procedimento, in esperimenti diversi, al variare del valore campionario osservato x_0, si possono ottenere corrispondenti valori di $\hat{\theta}$ in generale tra loro diversi. La corrispondenza $x_0 \mapsto \hat{\theta}$ definisce lo **stimatore di massima verosimiglianza** (M.V.), $\hat{\theta}(\mathbf{x})$, come la *funzione* che massimizza $L(\mathbf{x}, \cdot)$ rispetto a θ (assumendo ovviamente che un massimo esista $\forall x_0 \in \mathbb{R}^n$)

$$L\Big(\mathbf{x}, \hat{\boldsymbol{\theta}}(\mathbf{x})\Big) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\mathbf{x}, \boldsymbol{\theta}) \quad .$$

In teoria, $\hat{\theta}(\mathbf{x})$ si può calcolare massimizzando $p(\mathbf{x}, \theta)$ rispetto a θ, considerando $\mathbf{x}$ come un parametro libero.

Per fare i calcoli spesso conviene prendere il logaritmo di $L(x,\cdot)$ (che è una funzione monotona di L e quindi si massimizza per gli stessi valori di θ). La funzione (di θ)

$$\ell(\mathbf{x},\cdot) = \log L(\mathbf{x},\cdot) \tag{7}$$

si chiama funzione di "log-verosimiglianza".

In casi semplici, quando $p(\mathbf{x},\cdot)$ è derivabile rispetto a θ, $\hat{\theta}(\mathbf{x})$ si può calcolare esplicitamente risolvendo il sistema di p equazioni (*sistema di verosimiglianza*)

$$\frac{\partial \ell}{\partial \theta_k}(\mathbf{x},\theta) = 0 \quad , \qquad k = 1,\ldots,p \quad , \tag{8}$$

rispetto a θ e andando poi a vedere quale delle soluzioni fornisce un massimo assoluto di $\ell(\mathbf{x},\cdot)$. In genere però bisogna accontentarsi di procedimenti numerici per calcolare la singola stima $\hat{\theta}$, data x_0.

Esempio

Sia $\mathbf{y} \sim \mathcal{N}(\theta_1, \theta_2^2)$, scalare, e sia $\mathbf{x} = (\mathbf{y}_1,\ldots,\mathbf{y}_N)$ un campione casuale di numerosità N.

Sia $x = (y_1, \ldots, y_N)$ il risultato delle N osservazioni. Allora,

$$\begin{aligned} \ell(x,\theta) &= \log \left\{ \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\theta_2^2}} \exp -\frac{1}{2} \frac{(y_i-\theta_1)^2}{\theta_2^2} \right\} \\ &= -\frac{N}{2} \log 2\pi - \frac{N}{2} \log \theta_2^2 - \frac{1}{2} \sum_{1}^{N} \frac{(y_i-\theta_1)^2}{\theta_2^2} \quad , \end{aligned}$$

e le (8) diventano

$$\begin{aligned} \frac{\partial \ell}{\partial \theta_1} &= \frac{1}{\theta_2^2} \left(\sum_{1}^{N} y_i - N\theta_1 \right) = 0 \quad , \\ \frac{\partial \ell}{\partial \theta_2^2} &= -\frac{N}{2\theta_2^2} + \frac{1}{2\theta_2^4} \sum_{1}^{N} (y_i-\theta_1)^2 = 0 \quad . \end{aligned}$$

La prima è un'equazione nella sola θ_1 che dà

$$\hat{\theta}_1 = \frac{1}{N} \sum_{1}^{N} y_i = \bar{y}_N \quad . \tag{9}$$

Sostituendo questo valore nella seconda equazione, si trova subito

$$\hat{\theta}_2^2 = \frac{1}{N} \sum_{1}^{N} (y_i - \bar{y}_N)^2 = \hat{\sigma}_N^2 \quad .$$

ovvero, lo stimatore di massima verosimiglianza di θ_2^2 è la varianza campionaria. È facile verificare che queste soluzioni danno effettivamente un massimo assoluto di $\ell(x, \cdot)$. Si ha così

Proposizione 1 *Gli stimatori di M.V. basati su un campione casuale* $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ *estratto da* $\mathcal{N}(\theta_1, \theta_2^2)$ *sono la media e la varianza campionarie.*

Il risultato continua a valere pari pari nel caso multivariabile. Se $\mathbf{y} \sim \mathcal{N}(\theta, \Sigma)$, dove $\theta \in \mathbb{R}^p$ e $\Sigma \in \mathbb{R}^{p \times p}$ sono la media e la varianza incognite di $\mathbf{y}$, si dimostra che $\ell(\mathbf{x}, \theta, \Sigma)$ è massimizzata dallo stimatore $\phi = [\bar{\mathbf{y}}_N, \hat{\Sigma}_N^2]$ dove $\bar{\mathbf{y}}_N$ *ed* $, \hat{\Sigma}_N^2$ *sono la media e la varianza (matriciale) campionarie*. Per i dettagli della dimostrazione si veda [Soderstrom].

PROPRIETÀ DEGLI STIMATORI DI MASSIMA VEROSIMIGLIANZA

Gli stimatori di M.V. *non sono necessariamente corretti*. Per convincersene basta prendere l'esempio (scalare) appena considerato dello stimatore di M.V. di θ_2^2. Il fatto che $\hat{\sigma}_N^2$ non è corretto segue dalla seguente identità

$$N\hat{\sigma}_N^2(\mathbf{y}) = \sum_1^N (\mathbf{y}_i - \theta_1)^2 - N\,(\bar{\mathbf{y}}_N - \theta_1)^2 \quad . \tag{10}$$

che si dimostra partendo dalla

$$\begin{aligned}\sum_1^N (y_i - \theta_1)^2 &= \sum_1^N (y_i - \bar{y}_N + \bar{y}_N - \theta_1)^2 \\ &= \sum_1^N (y_i - \bar{y}_N)^2 + 2\sum_1^N (y_i - \bar{y}_N)\,(\bar{y}_N - \theta_1) + N(\bar{y}_N - \theta_1)^2 \quad .\end{aligned}$$

notando che la somma degli scarti dalla media campionaria, $\sum_1^N (y_i - \bar{y}_N)$, è necessariamente zero.

Calcolando il valore sperato E_θ dei due membri in (10) e tenendo conto del fatto che $\bar{\mathbf{y}} \sim \mathscr{N}(\theta_1, \theta_2^2/N)$ si trova

$$E_\theta \{N\hat{\sigma}_N^2\} = (N-1)\theta_2^2 \quad ,$$

per cui

$$E_\theta \, \hat{\sigma}_N^2 = \theta_2^2 \, \frac{N-1}{N} \quad . \tag{11}$$

C'è quindi un errore sistematico (*bias*) uguale a θ_2^2/N. La ragione di questo fatto risiede nel cosiddetto "principio di invarianza" della M.V..

Teorema 3 (Principio di invarianza) *Sia g una arbitraria funzione da Θ in Γ, dove Γ è un intervallo di $\mathbb{R}^k$ (k finito). Se $\hat{\theta}(\mathbf{x})$ è lo stimatore di M.V. di θ, allora $g(\hat{\theta}(\mathbf{x}))$ è lo stimatore di M.V. di $g(\theta)$.*

Una giustificazione intuitiva del principio di invarianza si può dare come segue. Supponiamo che g possieda un'inversa g^{-1} e definiamo

$$\tilde{\ell}(x, \gamma) = \ell\Big(x, g^{-1}(\gamma)\Big) = \ell(x, \theta)\Big|_{\theta = g^{-1}(\gamma)} \tag{12}$$

(questa è una riparametrizzazione della verosimiglianza $\ell(x,\cdot)$ dell'osservazione x).

Ora è facile convincersi che $\tilde{\ell}(x,\gamma)$ ha un massimo per $\gamma = \hat{\gamma}(x)$ se e solo se $\ell(x,\theta)$ ha un massimo (di uguale valore) in $\theta = \hat{\theta}(x)$ e i due punti di massimo sono legati fra loro dalla trasformazione $\theta = g^{-1}(\gamma)$, ovvero

$$\hat{\theta}(x) = g^{-1}\Big(\hat{\gamma}(x)\Big) \quad .$$

Ne segue che la stima di M.V. di γ è $\hat{\gamma}(x) = g(\hat{\theta}(x))$.

Si vede subito che se $\hat{\theta}$ è uno stimatore corretto di θ, $g(\hat{\theta})$ non può in generale essere uno stimatore corretto di $g(\theta)$ giacché E_θ e $g(\cdot)$ "non commutano", ovvero

$$E_\theta\, g\Big(\hat{\theta}(\mathbf{x})\Big) \neq g\Big(E_\theta\, \hat{\theta}(\mathbf{x})\Big) = g(\theta) \quad ,$$

a meno che g non sia una funzione *lineare*.

IL METODO DEI MOMENTI

Il *metodo dei momenti* è un metodo per costruire stimatori di certi parametri di una distribuzione di probabiltà nota uguagliando le espressioni dei primi momenti della distribuzione, ad esempio medie e varianze “teoriche”, a quelle dei corrispondenti momenti *campionari*; ad esempio alla media e alla varianza campionarie del campione.

Si vede subito che nel caso di distribuzioni Gaussiane *parametrizzate da media e varianza* (θ_1, θ_2^2), questo principio stipula che gli stimatori di (θ_1, θ_2^2) sono proprio la media e la varianza campionarie del campione. Però anche nel caso Gaussiano (θ_1, θ_2^2) potrebbero essere date come funzioni note di un altro parametro incognito diverso, ad esempio di un unico parametro reale nel caso scalare oppure un parametro vettoriale di dimensione più bassa di quella di (μ, Σ) nel caso vettoriale. In questo caso il metodo fornisce una procedura effettiva per il calcolo di stimatori del parametro da cui dipendono (μ, Σ).

Si dimostra che nel caso Gaussiano questo metodo di stima è sostanzialmente equivalente alla massima verosimiglianza . Questo però non è necessariamente il caso per distribuzioni non Gaussiane.

Esempio (da Wikipedia) Sia $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ un campione casuale estratto dalla distribuzione Gamma di parametro $\theta = [\,\alpha, \beta\,]$,

$$p_\theta(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}\,, \qquad x \geq 0$$

e zero per $x < 0$. Nei corsi di calcolo delle probabilità si mostra che la media di $\mathbf{y}_k$ vale

$$\mathbb{E}_\theta \mathbf{y}_k = \alpha\beta$$

mentre il momento secondo è

$$\mathbb{E}_\theta \mathbf{y}_k^2 = \beta^2 \alpha(\alpha+1)\,.$$

Gli stimatori di α e β col metodo dei momenti, si ottengono pertanto uguagliando i momenti teorici ai momenti campionari, ovvero risolvendo il sistema di

due equazioni:

$$\begin{cases} \alpha\beta & = & \bar{\mathbf{y}}_N = \frac{1}{N}\sum_{k=1}^{N} \mathbf{y}_k \\ \beta^2\alpha(\alpha+1) & = & \bar{\mathbf{m}}_N^2 := \frac{1}{N}\sum_{k=1}^{N} \mathbf{y}_k^2 \,. \end{cases}$$

ottenedo

$$\hat{\alpha} = \frac{\bar{\mathbf{y}}_N^2}{\bar{\mathbf{m}}_N^2 - \bar{\mathbf{y}}_N^2} \qquad \hat{\beta} = \frac{\bar{\mathbf{m}}_N^2 - \bar{\mathbf{y}}_N^2}{\bar{\mathbf{y}}_N} \,. \tag{13}$$

Notare che gli stimatori sono sempre funzioni dei momenti campionari. $\diamondsuit$

Come si vedrà più avanti, sotto ipotesi molto blande (ad esempio per un campione casuale) lo stimatore col metodo dei momenti è consistente ma in generale non è (neanche asintoticamente) efficiente.

MODELLI STATISTICI

Sia $\mathbf{y}$ un vettore aleatorio N-dimensionale di d.d.p. incognita appartenente alla famiglia parametrica $\{F_\theta\,;\,\theta \in \Theta\}$. Chiameremo *modello statistico* (o, equivalentemente, *modello probabilstico*) di $\mathbf{y}$ una rappresentazione del tipo

$$\mathbf{y} = f(\theta, \mathbf{w}) \quad ,$$

dove f è una funzione nota e $\mathbf{w}$ è un vettore aleatorio di struttura più semplice di quella di $\mathbf{y}$, di distribuzione di probabilità nota.

Un modello è da riguardarsi come una descrizione del fenomeno che genera le osservazioni. In molte applicazioni, $\mathbf{w}$ normalmente rappresenta il "rumore" ovvero le cause accidentali che rendono incerta la relazione fra il parametro θ che si vuole determinare e le misure $\mathbf{y}$ che si eseguono per arrivare alla sua conoscenza.

Nonostate la descrizione di $\mathbf{y}$ tramite un modello sia in teoria equivalente alla conoscenza di $\{F_\theta\,;\,\theta \in \Theta\}$, dato che si può pensare di ricavare, per

ogni θ, la distribuzione di probabilità di $\mathbf{y}$ a partire dalla distribuzione (nota) di $\mathbf{w}$ mediante le regole del calcolo delle probabilità, in ingegneria e nelle scienze applicate è molto più frequente (e spesso più intuitivo) descrivere i dati per mezzo di un modello statistico che non mediante una famiglia parametrica $\{F_{\theta}\}$.

IL MODELLO DI GAUSS

Si supponga di eseguire una serie di N misure, non necessariamente mediante lo stesso apparato sperimentale, su una certa p-pla di variabili (che si assume siano costanti nel tempo) non accessibili direttamente che modelleremo come un parametro p-dimensionale deterministico (ma incognito) θ.
Ammettiamo che l'incertezza sul risultato di ciascuna misura si possa esprimere come un errore additivo secondo uno schema del tipo

$$y_k = s_k(\theta) + w_k \quad , \qquad k = 1, \ldots, N \quad ,$$

dove $s_k(\theta)$ è la caratteristica "ideale" dello strumento di misura, funzione nota di θ e w_k è un termine d'errore. In molti processi di misura w_k è il risultato a livello macroscopico "aggregato" di molte cause d'errore accidentale "microscopiche" fra loro indipendenti. Le variabili d'errore accidentale microscopiche si suppongono mediamente piccole e si può quindi ragionevolmente assumere che esse (una volta normalizzate attraverso opportuni fattori di scala) si combinino linearmente (i.e. si sommino) per

produrre l'effetto macroscopico w_k. In questo contesto vale il teorema del limite centrale e w_k si può descrivere come la determinazione di una *variabile aleatoria Gaussiana* $\mathbf{w}_k$. Supponendo che vi sia assenza di errori sistematici, $\mathbf{w}_k$ può essere ipotizzata a *media nulla*.

Pensiamo allora y_k come la determinazione di una variabile casuale scalare $\mathbf{y}_k$ e raccogliamo gli N campioni $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ in un vettore colonna $\mathbf{y}$. Si può così scrivere sinteticamente

$$\mathbf{y} = s(\theta) + \mathbf{w} \quad ,$$

dove abbiamo introdotto i due vettori colonna

$$\begin{aligned} s(\theta) &= [s_1(\theta), \ldots, s_N(\theta)]' \quad , \\ \mathbf{w} &= [\mathbf{w}_1, \ldots, \mathbf{w}_N]' \, . \end{aligned}$$

Questo modello "della Teoria degli Errori" di Gauss è del tipo "misura" = "segnale" più "rumore" (Gaussiano) additivo ed è simile alla descrizione che si usa per i canali di comunicazione numerica o per misure fatte sequenzialmente nel tempo da sensori numerici nei sistemi di controllo e in miriadi di altre applicazioni ingegneristiche.

Nel modello () la matrice di covarianza del rumore

$$R := E\mathbf{w}\mathbf{w}^\top$$

è in generale solo parzialmente nota. In effetti nella pratica si possono dare situazioni estremamente diverse. La più semplice è quella di errori $\mathbf{w}_k$ *indipendenti e statisticamente identici*, in particolare tutti con la stessa varianza $r_{kk} = \sigma^2$, $k = 1, \ldots, N$. Conviene in questo caso introdurre nel modello come ulteriore parametro incognito la varianza del rumore scrivendo

$$\mathbf{y} = s(\theta) + \sigma\mathbf{w} \quad ,$$

dove $\mathbf{w} \sim \mathcal{N}(0, I)$.

L'altro caso estremo si presenta quando l'intera matrice varianza di $\mathbf{w}$ è incognita e va quindi considerata tra i parametri incogniti da stimare. I problemi di stima associati ad un modello di questo tipo sono però molto complicati. Nel seguito supporremo $\mathbf{w}$ Gaussiano e di covarianza parzialmente nota, della forma $\sigma^2 R$ con σ^2 *incognita* ed R *nota* e definita positiva.

Faremo inoltre l'ipotesi che $s(\theta)$ sia una funzione lineare del parametro θ, cioè

$$s(\theta) = S\theta \quad , \qquad S \in \mathbb{R}^{N\times p} \quad ,$$

con S matrice nota di dimensione $N \times p$. In questa sezione ci occuperemo della stima dei parametri nel *modello lineare*

$$\mathbf{y} = S\theta + \sigma\mathbf{w}\,. \tag{14}$$

Quando la varianza del rumore R è nota il modello (14) può essere facilmente normalizzato a uno in cui $\mathrm{Var}\{\mathbf{w}\}$ è l'identità.
In effetti, dato che R è simmetrica e definita positiva, essa ammette "radici quadrate" $R^{1/2} \in R^{N\times N}$ tali che $R = R^{1/2}(R^{1/2})^\top$, calcolabili ad esempio mediante una fattorizzazione di Cholesky. Il modello normalizzato si ottiene moltiplicando a sinistra entrambi i membri dell'equazione per $R^{-1/2}$ e ridefinendo opportunamente le variabili come

$$\bar{\mathbf{y}} := R^{-1/2}\mathbf{y}\,, \qquad \bar{S} := R^{-1/2}S\,, \qquad \bar{\mathbf{w}} := R^{-1/2}\mathbf{w}\,.$$

Sebbene questa normalizzazione faciliti i calcoli, in questo capitolo essa non verrà usata perchè nella soluzione dei problemi di stima legati al modello (14) la matrice R gioca un ruolo importante di "matrice peso" che ispira i procedimenti empirici di stima ai minimi quadrati che verranno esposti più avanti e sono importanti nelle applicazioni. Con la normalizzazione questo ruolo verrebbe completamente mascherato.

STIMA DI M.V. NEL MODELLO LINEARE-GAUSSIANO

Problema 1 *Trovare le stime di M.V. dei parametri $\theta \in \mathbb{R}^p$ e $\sigma^2 \in \mathbb{R}_+$ nel modello lineare*

$$\mathbf{y} = S\theta + \sigma\mathbf{w}$$

dove $S \in \mathbb{R}^{N\times p}$ è una matrice nota e $\mathbf{w}$ è un vettore aleatorio Gaussiano di media zero e varianza nota R, definita positiva.

Per risolvere questo problema notiamo innanzitutto che $\mathbf{y} \sim \mathcal{N}(S\theta, \sigma^2 R)$ e pertanto la funzione di log-verosimiglianza si scrive

$$\begin{aligned}\ell(y,\theta,\sigma^2) &= -\frac{N}{2}\log 2\pi - \frac{1}{2}\log\left[\det(\sigma^2 R)\right] - \frac{1}{2}(y-S\theta)^\top(\sigma^2 R)^{-1}(y-S\theta)\\ &= -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\sigma^2 - \frac{1}{2}\log\det R - \frac{1}{2\sigma^2}(y-S\theta)^\top R^{-1}(y-S\theta),\end{aligned}$$

cosicché

$$\frac{\partial \ell}{\partial \theta} = \frac{1}{\sigma^2} S^\top R^{-1} (y - S\theta)$$

(ricordare che il gradiente rispetto a x di $f^\top(x) A f(x)$ è $2 \frac{\partial f}{\partial x} A f(x)$, se lo si esprime come vettore colonna).
Inoltre si ha

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} (y - S\theta)^\top R^{-1} (y - S\theta).$$

Calcoliamo la matrice di Fisher $I(\theta, \sigma^2)$.

$$\mathbf{z}_\theta := \frac{\partial \ell(\mathbf{y}, \theta, \sigma^2)}{\partial \theta} \quad , \qquad \mathbf{z}_\sigma := \frac{\partial}{\partial \sigma^2} \ell(\mathbf{y}, \theta, \sigma^2) \quad ,$$

e ricordiamo che

$$I(\theta, \sigma) = E_{\theta,\sigma} \begin{bmatrix} \mathbf{z}_\theta \mathbf{z}_\theta^\top & \mathbf{z}_\theta \mathbf{z}_\sigma \\ \mathbf{z}_\theta^\top \mathbf{z}_\sigma & \mathbf{z}_\sigma^2 \end{bmatrix}.$$

Svolgendo i calcoli, si trova

$$\begin{aligned} E\,\mathbf{z}_\theta \mathbf{z}_\theta^\top &= \frac{1}{\sigma^4}\, S^\top R^{-1}\, E_{\theta,\sigma}\{(\mathbf{y}-S\theta)\,(\mathbf{y}-S\theta)^\top\}\, R^{-1} S \\ &= \frac{1}{\sigma^4}\, S^\top R^{-1}\, \sigma^2 R\, R^{-1} S = \frac{1}{\sigma^2}\, S^\top R^{-1} S. \end{aligned}$$

Ponendo inoltre

$$\tilde{\mathbf{y}} := R^{-1/2}\,(\mathbf{y}-S\theta)$$

si riconosce che $\tilde{\mathbf{y}} \sim \mathcal{N}(0,\sigma^2 I)$ e

$$\begin{aligned} E_{\theta,\sigma}\, \mathbf{z}_\theta \mathbf{z}_\sigma &= E_{\theta,\sigma}\left\{ \frac{1}{\sigma^2}\, S^\top R^{-1/2}\, \tilde{\mathbf{y}} \left(-\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\, \tilde{\mathbf{y}}^\top \tilde{\mathbf{y}} \right) \right\} \\ &= \frac{1}{2\sigma^6}\, S^\top R^{-1/2}\, E_{\theta,\sigma}\, \tilde{\mathbf{y}} \tilde{\mathbf{y}} \tilde{\mathbf{y}}^\top \tilde{\mathbf{y}} = 0 \quad , \end{aligned}$$

dato che $\tilde{\mathbf{y}}$ ha media zero e i momenti centrali del terz'ordine di una d.d.p. Gaussiana sono nulli. Infine, dato che

$$\|\tilde{\mathbf{y}}\|^2 = (\mathbf{y}-S\theta)^\top R^{-1}\,(\mathbf{y}-S\theta) \quad ,$$

si vede che

$$E_{\theta,\sigma}\, \mathbf{z}_\sigma^2 = E_{\theta,\sigma} \left\{ \frac{1}{2\sigma^2} \left[\frac{\|\tilde{\mathbf{y}}\|^2}{\sigma^2} - N \right] \right\}^2 .$$

Vedremo più avanti che , se $\mathbf{y} \sim \mathcal{N}(\mu, \Sigma)$, la forma quadratica $(\mathbf{y}-\mu)^\top \Sigma^{-1}(\mathbf{y}-\mu)$ ha una distribuzione del tipo χ^2 con un numero di gradi di libertà pari alla dimensione di $\mathbf{y}$. Allora, $\frac{\|\tilde{\mathbf{y}}\|^2}{\sigma^2} \sim \chi^2(N)$ e la sua varianza è uguale a $2N$. Segue

$$E_{\theta,\sigma}\, \mathbf{z}_\sigma^2 = \frac{1}{4\sigma^4} \operatorname{Var} \left[\frac{\|\tilde{\mathbf{y}}\|^2}{\sigma^2} \right] = \frac{N}{2\sigma^4} .$$

Mettendo insieme questi risultati si trova infine

$$I(\theta, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2} S^\top R^{-1} S & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix} .$$

Proposizione 2 *Nel modello lineare sia $N \geq p$. Allora, θ è globalmente identificabile se e solo se S ha rango p.*

Difatti $I(\theta, \sigma^2)$ è non singolare se e solo se $S^\top R^{-1} S$ è invertibile e questo avviene allora e solo allora che $S^\top R^{-1} S\theta = 0$ implica $\theta = 0$. Ovviamente se $\mathrm{Ker}\, S$ contenesse un $\xi \neq 0$, θ_0 e $\theta_0 + \xi$ sarebbero indistinguibili.

D'ora in avanti supporremo sempre le p colonne di S *linearmente indipendenti* (rango $S = p$). Ciò equivale all'esistenza dell'inversa $I^{-1}(\theta, \sigma^2)$ e la minima varianza di uno stimatore corretto di θ non può essere inferiore a $\sigma^2 [S^\top R^{-1} S]^{-1}$. Analogamente quella di uno stimatore corretto di σ^2 non può essere inferiore a $\frac{2\sigma^4}{N}$.

LO STIMATORE DI θ

$$\frac{\partial \ell}{\partial \theta} = \frac{1}{\sigma^2} S^\top R^{-1} (y - S\theta) = 0 \quad \Rightarrow \quad \hat{\theta}(y) = [S^\top R^{-1} S]^{-1} S^\top R^{-1} y$$

INTERPRETAZIONE GEOMETRICA

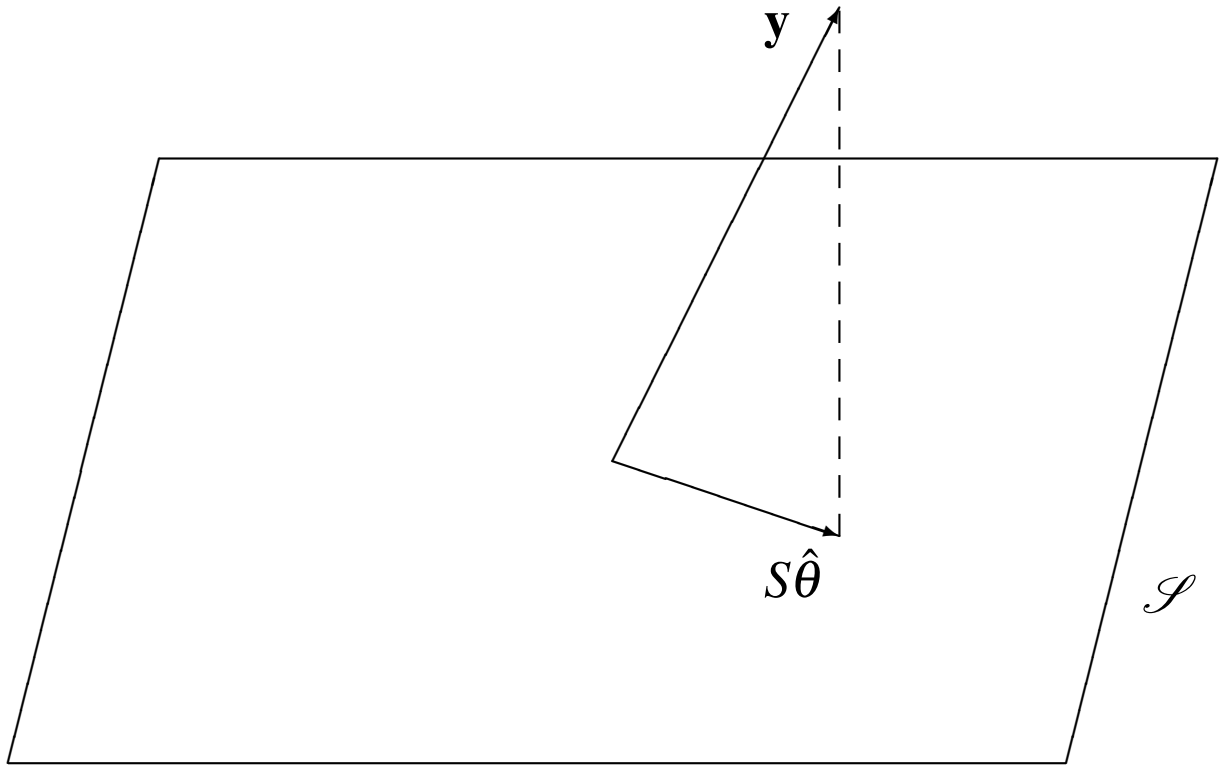
$$\hat{\theta}(y) = \min_{\theta} (y - S\theta)^\top R^{-1} (y - S\theta)$$

Quadrato della *distanza* in $\mathbb{R}^N$ *indotta dal prodotto scalare*

$$\langle x, y \rangle_{R^{-1}} := x^\top R^{-1} y$$

$$(y - S\theta)^\top R^{-1} (y - S\theta) = \|y - S\theta\|^2_{R^{-1}} \quad ,$$

Dato $y \in \mathbb{R}^N$ minimizzare la distanza rispetto a θ, significa *cercare il vettore* $v \in \mathscr{S} := \text{span}\,(S)$ (lo spazio vettoriale generato dalle colonne di S) *che ha minima distanza da* y.



Proiezione ortogonale.

PROIEZIONE ORTOGONALE

$S\hat{\theta}(y) := SAy$ è la proiezione ortogonale di y sullo spazio $\mathscr{S} = \mathrm{span}(S)$. In altri termini, la matrice $P \in \mathbb{R}^{N\times N}$, definita ponendo

$$P = SA \quad ,$$

è il proiettore ortogonale (rispetto al prodotto scalare $\langle\cdot,\cdot\rangle_{R^{-1}}$) *di* $\mathbb{R}^N$ *su* $\mathscr{S}$. Difatti P è idempotente ($P = P^2$), essendo

$$SA \cdot SA = S \cdot I \cdot A = SA$$

ma non è simmetrica come accade nella metrica Euclidea ordinaria, ma piuttosto

$$P^\top = (SA)^\top = A^\top S^\top = R^{-1}S\left[S^\top R^{-1} S\right]^{-1} S^\top = R^{-1}SAR = R^{-1}PR, \qquad (15)$$

cioè $P^\top$ è simile a P.

Classica caratterizzazione geometrica della proiezione ortogonale:

l'unico vettore $S\theta$ di $\mathscr{S}$ che ha distanza minima da $y \in \mathbb{R}^N$, secondo la metrica $\|\cdot\|_{R^{-1}}$, è quello per cui l'errore, $y - S\theta$, è ortogonale a $\mathscr{S}$ rispetto al prodotto scalare $\langle \cdot, \cdot \rangle_{R^{-1}}$.

Nella nostra ipotesi le colonne di S sono linearmente indipendenti: $\hat{\theta}(y)$ è l'unico vettore θ tale per cui

$$S \perp_{R^{-1}} (y - S\theta).$$

In altre parole

$$S^\top R^{-1} y - S^\top R^{-1} S\theta = 0$$

Proposizione 3 *Nel modello lineare-Gaussiano lo stimatore a M.V. di θ coincide con la funzione dei dati osservati che minimizza la distanza quadratica . In altre parole, $\hat{\theta}(y)$ è lo* **stimatore ai minimi quadrati pesati** *di θ con matrice peso R^{-1}.*

PROPRIETÀ DI $\hat{\theta}(\mathbf{y})$

1. $\hat{\theta}(\mathbf{y})$ *è uno stimatore corretto*. Infatti

$$E_{\theta,\sigma} A\mathbf{y} = A\,S\theta = \theta \quad ,$$

dato che $AS = I$, A è una inversa sinistra di S.

2. $\hat{\theta}(\mathbf{y})$ *ha varianza* $\sigma^2[S^\top R^{-1} S]^{-1}$ *coincidente con quella data dal limite* di Cramèr-Rao. Pertanto $\hat{\theta}(\mathbf{y})$ è uno stimatore a *minima varianza*. Infatti

$$\begin{aligned} & E_{\theta,\sigma}(A\mathbf{y}-\theta)\,(A\mathbf{y}-\theta)^\top \\ &= E_{\theta,\sigma}\,(AS\theta + A(\sigma\mathbf{w}) - \theta)\,(AS\theta + A(\sigma\mathbf{w}) - \theta)^\top \\ &= E_{\theta,\sigma}\,A(\sigma\mathbf{w})\,(\sigma\mathbf{w})^\top A^\top = \sigma^2 ARA^\top \\ &= \sigma^2[S^\top R^{-1} S]^{-1} S^\top R^{-1} R R^{-1} S[S^\top R^{-1} S]^{-1} = \sigma^2[S^\top R^{-1} S]^{-1}. \end{aligned}$$

3. Lo stimatore $\hat{\theta}(\mathbf{y})$ *è normalmente distribuito*, i.e.

$$\hat{\theta}(\mathbf{y}) \sim N\left(\theta, \sigma^2 \left[S^\top R^{-1} S\right]^{-1}\right) .$$

Queta proprietà è conseguenza della linearità.

LO STIMATORE DI σ^2

Dalla $\partial \ell / \partial \sigma^2 = 0$ si ricava

$$\hat{\sigma}^2(y) = \frac{1}{N} \left(y - S\hat{\theta}(y)\right)^\top R^{-1} \left(y - S\hat{\theta}(y)\right)^\top = \frac{1}{N} \|y - Py\|^2_{R^{-1}},$$

cioè $\hat{\sigma}^2(y)$ è il quadrato della norma dell'errore di approssimazione di y mediante il vettore $Py = S\hat{\theta}(y)$, divisa per N. Per vedere se $\hat{\sigma}^2(\mathbf{y})$ è corretto e calcolarne la varianza occorre vedere come è distribuito.

Dobbiamo ricordare alcune proprietà della distribuzione χ^2.

LA DISTRIBUZIONE χ^2

Si dice che la variabile scalare $\mathbf{y}$ è distribuita secondo $\chi^2(n)$ se

$$P(x \leq \mathbf{y} < x+dx) = \frac{1}{2^{n/2}\,\Gamma\left(\frac{n}{2}\right)}\, x^{\left(\frac{n}{2}\right)-1}\, e^{-x/2}\, dx \quad ,$$

per $x \geq 0$ e zero altrimenti n è un numero naturale che si chiama *numero dei gradi di libertà* della distribuzione. La χ^2 è un caso speciale della distribuzione Gamma; la sua funzione caratteristica (abbreviata a f.c. nel seguito) è

$$\phi(it) := E\, e^{it\,\mathbf{y}} = (1-2it)^{-n/2} \quad ,$$

basta tener presente che nella trasformata di Fourier il fattore esponenziale è $e^{-i\omega y}$. Usando questa espressione si possono ricavare facilmente delle formule per i momenti della distribuzione. I primi momenti *centrali* sono

$$\begin{aligned} \mu_1 &= n \\ \mu_2 &= 2n \\ \mu_3 &= 8n \\ \mu_4 &= 48n + 12n^2 \qquad \text{ecc...} \end{aligned}$$

Consideriamo una v.c. $\mathbf{y} \sim \chi^2(n)$ e introduciamo la variabile standardizzata

$$\mathbf{z}_n := \frac{\mathbf{y} - n}{\sqrt{2n}} \quad ;$$

che ha media zero e varianza 1 (per ogni n). Notiamo che $\mathbf{z}_n$ non è più distribuita secondo χ^2 (ricordare che l'unica d.d.p. la cui forma funzionale si conserva per trasformazioni lineari è la *Gaussiana*!).

Il limite in distribuzione, $L - \lim_{n\to\infty} \mathbf{z}_n$, è una variabile Gaussiana standardizzata $\mathcal{N}(0,1)$. Ricordiamo a questo proposito il seguente risultato (che daremo per noto).

Proposizione 4 (Helly-Bray) *Se $\phi_n(t)$ è la f.c. di $\mathbf{x}_n$ e $\phi(t)$ è la f.c. di $\mathbf{x}$, allora*

$$\mathbf{x}_n \xrightarrow{L} \mathbf{x} \qquad \textit{se e solo se} \qquad \phi_n(t) \to \phi(t) \quad , \qquad \forall t \, .$$

La f.c., $\phi_n(t)$, di $\mathbf{z}_n$ si può scrivere,

$$\begin{aligned}
\phi_n(t) &= E\, e^{it\frac{\mathbf{y}}{\sqrt{2n}}}\, e^{-it\frac{n}{\sqrt{2n}}} = e^{-it\frac{n}{\sqrt{2n}}} \left(1 - \frac{2it}{\sqrt{2n}}\right)^{-n/2} \\
&= \left(e^{-it\sqrt{\frac{2}{n}}}\right)^{n/2} \left(1 - it\sqrt{\frac{2}{n}}\right)^{-n/2} \\
&= \left[e^{it\sqrt{\frac{2}{n}}} - it\sqrt{\frac{2}{n}}\, e^{it\sqrt{\frac{2}{n}}}\right]^{-n/2} = \left(1 - \frac{t^2}{n} + \frac{\psi(n)}{n}\right)^{-n/2} \quad ,
\end{aligned}$$

dove $\lim_{n\to\infty} \psi(n) = 0$. Passando al $\lim_{n\to\infty} \phi_n(t)$ si ha, per una nota formula dell'analisi,

$$\phi(t) = \lim_{n\to\infty} (1 - t^2/n)^{n/2} = e^{-t^2/2} \quad ,$$

che è proprio la f.c. di una variabile Gaussiana standardizzata. In sostanza per n grandi una variabile $\chi^2(n)$ si comporta come una Gaussiana $\mathcal{N}(n, 2n)$.

La distribuzione χ^2 interviene in molte questioni di inferenza statistica e qui di seguito ne elencheremo alcune proprietà importanti che stanno alla

base del calcolo della distribuzione di probabilità di stimatori che sono forme quadratiche di variabili Gaussiane.

Proposizione 5 *La somma di N variabili casuali indipendenti $\mathbf{y}_i \sim \chi^2(n_i)$ è distribuita secondo $\chi^2(n)$ dove*

$$n = \sum_{i=1}^{N} n_i \quad ,$$

cioè i gradi di libertà si sommano.

Prova: La prova di questo risultato si basa sulla nota espressione della f.c. della somma $\sum_1^N \mathbf{y}_i$ di variabili indipendenti come prodotto delle f.c., $\phi_i(t)$, delle $\mathbf{y}_i$. Moltiplicando tra loro le funzioni caratteristiche si vede in effetti che i gradi di libertà si sommano. □

Proposizione 6 *Se $\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2$ dove $\mathbf{y}_1$ è indipendente da $\mathbf{y}_2$ e sia $\mathbf{y}$ che $\mathbf{y}_2$ hanno distribuzione χ^2 con gradi di libertà rispettivamente n ed n_2 con $n > n_2$, allora la variabile $\mathbf{y}_1$ è distribuita come $\chi^2(n - n_2)$.*

Prova: Ovviamente per l'indipendenza, la f.c. di $\mathbf{y}$ è $\phi = \phi_1 \phi_2$ e quindi

$$\phi_1 = \frac{\phi}{\phi_2}$$

Sostituendo in questo rapporto le espressioni per le rispettive funzioni caratteristiche, si ricava l'asserto. □

Proposizione 7 *La distribuzione di* $\frac{1}{\sigma^2} \sum_1^n (\mathbf{y}_i - \mu)^2$ *, con* $\mathbf{y}_i \sim \mathcal{N}(\mu, \sigma^2)$ *e indipendenti è* $\chi^2(n)$*.*

Prova: In effetti basta mostrare che la d.d.p. di $\mathbf{z} := (\mathbf{y} - \mu)^2/\sigma^2$ con $\mathbf{y} \sim \mathcal{N}(\mu, \sigma)$ è $\chi^2(1)$ e poi usare la proposizione 5. Notiamo che si può scrivere $\mathbf{z} = \mathbf{x}^2$ con $\mathbf{x} \sim \mathcal{N}(0,1)$. Usando le note regole per il calcolo della distribuzione di una funzione di variabile aleatoria, riferite alla funzione

$z = f(x)$ con $f(x) = x^2$, si può calcolare la densità di probabilità di $\mathbf{z}$ come

$$\begin{aligned} p_{\mathbf{z}}(z) &= \frac{1}{\left|\frac{d}{dx} f(x)\Big|_{x=f^{-1}(z)}\right|} \left[p_{\mathbf{x}}(\sqrt{z}) + p_{\mathbf{x}}(-\sqrt{z})\right] 1(z) \\ &= \frac{1}{|2\sqrt{z}|} \frac{1}{\sqrt{2\pi}} \left[e^{-z/2} + e^{-z/2}\right] 1(z) = \frac{1}{\sqrt{2\pi z}} e^{-z/2} \quad , \qquad z \geq 0 \quad , \end{aligned}$$

che è proprio $\chi^2(1)$. □

Proposizione 8 *La distribuzione della varianza campionaria normalizzata*

$$\frac{n\hat{\sigma}_n^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_1^n (\mathbf{y}_i - \bar{\mathbf{y}}_n)^2 \quad ,$$

con $\mathbf{y}_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$, *indipendenti, è* $\chi^2(n-1)$.

Prova: Mostriamo allo scopo il seguente risultato notevole.

Lemma 1 *Nelle ipotesi poste, le statistiche* $\bar{\mathbf{y}}_n$ *ed* $\hat{\sigma}_n^2$ *sono indipendenti.*

Prova: Basta far vedere che $\bar{\mathbf{y}}_n$ e $\mathbf{y}_i - \bar{\mathbf{y}}_n$ sono scorrelate qualunque sia i. Questo implica che $\bar{\mathbf{y}}_n$ e $(\mathbf{y}_i - \bar{\mathbf{y}}_n)$, $i = 1, \ldots, n$, sono indipendenti, data l'ipotesi di Gaussianità e quindi l'asserto.

Definendo $\tilde{\mathbf{y}}_i = \mathbf{y}_i - \mu$ e $\tilde{\mathbf{y}} = \bar{\mathbf{y}}_n - \mu$ si ha $\mathbf{y}_i - \bar{\mathbf{y}}_n = \tilde{\mathbf{y}}_i - \tilde{\mathbf{y}}$ ed $E\,\bar{\mathbf{y}}_n(\mathbf{y}_i - \bar{\mathbf{y}}_n) = E\,\tilde{\mathbf{y}}(\tilde{\mathbf{y}}_i - \tilde{\mathbf{y}}) = E(\tilde{\mathbf{y}}\tilde{\mathbf{y}}_i) - E(\tilde{\mathbf{y}})^2$. Per l'indipendenza delle variabili $\mathbf{y}_i$,

$$E\,\tilde{\mathbf{y}}\tilde{\mathbf{y}}_i = \frac{1}{n} E\left(\sum_{k=1}^{n} \tilde{\mathbf{y}}_k \tilde{\mathbf{y}}_i\right) = \frac{1}{n} E(\tilde{\mathbf{y}}_i)^2 = \frac{\sigma^2}{n}$$

e quindi confrontando con l'espressione $E(\tilde{\mathbf{y}})^2 = \sigma^2/n$, si ottiene la conclusione. □

Usiamo ora la solita identità

$$\sum_{1}^{n} (\mathbf{y}_i - \mu)^2 = \sum_{1}^{n} (\mathbf{y}_i - \bar{\mathbf{y}}_n)^2 + n(\bar{\mathbf{y}}_n - \mu)^2$$

per scrivere

$$\sum_1^n \frac{(\mathbf{y}_i - \mu)^2}{\sigma^2} = \sum_1^n \frac{(\mathbf{y}_i - \bar{\mathbf{y}}_n)^2}{\sigma^2} + n \frac{(\bar{\mathbf{y}}_n - \mu)^2}{\sigma^2}$$

dove la somma al secondo membro è di due v.c. *indipendenti*. Sappiamo da A) che il primo membro$\sim \chi^2(n)$ e che $(\bar{\mathbf{y}}_n - \mu)^2/(\sigma^2/n) \sim \chi^2(1)$ (questo scende ancora dalla proposizione 7 con $n = 1$). Per la proposizione 6 il primo addendo al secondo membro deve essere $\chi^2(n-1)$. □

Tutte le considerazioni fin qui fatte sono relative al caso scalare. Se $\mathbf{y}$ è un vettore aleatorio m-dimensionale ci si interessa della struttura delle forme quadratiche del tipo $\mathbf{y}^\top Q\mathbf{y}$ con $Q = Q^\top$, che hanno una distribuzione χ^2. Il caso più semplice è il seguente.

Proposizione 9 *Se* $\mathbf{y} \sim \mathcal{N}(\mu, \Sigma)$ *con* $\mu \in \mathbb{R}^m$ *e* $\Sigma \in \mathbb{R}^{m\times m}$ *definita positiva, allora*

$$(\mathbf{y} - \mu)^\top \Sigma^{-1} (\mathbf{y} - \mu) \sim \chi^2(m)\,.$$

In effetti basta standardizzare $\mathbf{y}$, ponendo $\mathbf{z} := \Sigma^{-1/2}(\mathbf{y} - \mu)$; allora $\mathbf{z} = [z_1, \ldots, \mathbf{z}_m]^\top$ è $\mathcal{N}(0, I)$, cioè $\mathbf{z}_1, \ldots, \mathbf{z}_m$ sono *indipendenti* ed $\mathcal{N}(0,1)$. Con la posizione fatta si ha poi

$$(\mathbf{y} - \mu)^\top \Sigma^{-1} (\mathbf{y} - \mu) = \mathbf{z}^\top \mathbf{z} = \sum_1^m \mathbf{z}_i^2$$

e quindi l'ultimo membro è $\chi^2(m)$ per la proposizione 5.

Una caratterizzazione meno banale e di uso molto frequente è la seguente.

Proposizione 10 *Sia $\mathbf{z} \sim \mathcal{N}(0, I_m)$ e $Q \in \mathbb{R}^{m \times m}$. Allora la forma quadratica $\mathbf{z}^\top Q \mathbf{z}$ è distribuita secondo χ^2 se e solo se Q è idempotente, ovvero $Q = Q^2$. In questo caso il numero di gradi di libertà è $r = \operatorname{rango} Q$.*

Prova: La prova di questo risultato è basata su un procedimento di diagonalizzazione di Q. Dato che Q è simmetrica (notare che può sempre essere supposta tale) e $Q = Q^2$, essa è una matrice di proiezione ortogonale

in $\mathbb{R}^m$. I suoi autovalori non nulli sono pertanto uguali a uno (in numero di $r = \text{rango}\, Q$). La decomposizione spettrale della matrice Q si può così scrivere

$$Q = U \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} U^\top, \qquad UU^\top = U^\top U = I_m$$

ovvero

$$Q = U_1 U_1^\top,$$

dove U_1 è la matrice $m \times r$ formata dalle prime r colonne (ortonormali) di U. Si ha perciò

$$\mathbf{z}^\top Q \mathbf{z} = \mathbf{z}_1^\top \mathbf{z}_1$$

dove il vettore r-dimensionale $\mathbf{z}_1 := U_1^\top \mathbf{z}$ è distribuito come $\mathcal{N}(0, I_r)$. La conclusione segue ancora dalla proposizione 5. □

DISTRIBUZIONE DI $\hat{\sigma}^2(\mathbf{y})$

Teorema 4 *Lo stimatore di M.V. della varianza σ^2 nel modello lineare ha distribuzione di probabilità che corrisponde alla*

$$\frac{N\hat{\sigma}^2(\mathbf{y})}{\sigma^2} \sim \chi^2(N-p)\,. \tag{16}$$

In particolare la sua media e varianza sono date da

$$E_{\theta,\sigma^2}\,\hat{\sigma}^2(\mathbf{y}) = \sigma^2\,\frac{N-p}{N} \quad , \tag{17}$$

$$\mathrm{Var}_{\theta,\sigma^2}\,\hat{\sigma}^2(\mathbf{y}) = \sigma^4\,\frac{2(N-p)}{N^2}\,. \tag{18}$$

Dato che $\mathbf{y} - P\mathbf{y} = (\mathbf{y} - S\theta) - P(\mathbf{y} - S\theta) = \sigma(I-P)\,\mathbf{w}$. Definiamo allora il vettore casuale

$$\mathbf{z} := R^{-1/2}\mathbf{w} \quad ,$$

il quale è chiaramente distribuito secondo la $\mathcal{N}(0, I)$. Dalla () si ricava poi con facili passaggi la

$$\frac{N\hat{\sigma}^2(\mathbf{y})}{\sigma^2} = \mathbf{w}^\top (I-P)^\top R^{-1}(I-P)\,\mathbf{w} = \mathbf{z}^\top \left[R^{-1/2}(I-P)\,R^{1/2}\right]\mathbf{z} \quad ,$$

dove si è usata la proprietà di similitudine $P^\top = R^{-1}PR$ stabilita nella (15). Notiamo ora che la matrice tra parentesi quadre, diciamola Q, è **simmetrica e idempotente** giacché,

$$Q^2 = R^{-1/2}(I-P)^2 R^{1/2} = R^{-1/2}(I-P)\,R^{1/2} = Q$$

e il suo rango è $N-p$. Infatti $I-P$ proietta su un sottospazio ortogonale a S e nelle nostre ipotesi $\dim S = p$. Segue allora dalla proposizione 10 stabilita più sopra che $\mathbf{z}^\top Q\mathbf{z} \sim \chi^2(N-p)$.

Come si vede dalla (17) lo stimatore $\hat{\sigma}^2(\mathbf{y})$ *non è corretto*. L'errore sistematico che si commette, uguale a $-\sigma^2\, p/N$, tende però a zero al crescere della numerosità campionaria. Si noti che l'errore sistematico può facilmente essere eliminato assumendo come stimatore di σ^2 la quantità

$$s^2(\mathbf{y}) := \frac{1}{N-p}\,\|\mathbf{y} - S\hat{\theta}(\mathbf{y})\|^2_{R^{-1}}\,.$$

Questa correzione si paga però con una *varianza maggiore*. Infatti dalla $(N-p)\,s^2(\mathbf{y})/\sigma^2 \sim \chi^2(N-p)$ segue facilmente che

$$\mathrm{Var}_{\theta,\sigma^2}\, s^2(\mathbf{y}) = \frac{2\sigma^4}{N-p} \quad ,$$

che è strettamente più grande di $2\sigma^4\,(N-p)/N^2$. Notiamo per inciso che la varianza di $\hat{\sigma}^2(\mathbf{y})$ è più piccola del limite inferiore di Cramer-Rao, pari a $2\sigma^4/N$.

APPROSSIMAZIONE AI MINIMI QUADRATI

Tecnica deterministica per descrivere (approssimativamente) dati misurati mediane un modello parametrico

Dati misurati $(y_1,\ldots,y_N)$ funzione incognita di certi valori assegnati o misurati $(u_1,\ldots,u_N)$ della variabile dipendente u negli istanti di misura.

Classe parametrica di modelli:

$$\hat{y}_t(\theta) = f(u_t,\theta,t) \quad , \qquad t = 1,\ldots,N, \qquad \theta \in \Theta \subseteq \mathbb{R}^p$$

criterio di approssimazione **quadratico**

$$V(\theta) := \sum_1^N [y_t - \hat{y}_t(\theta)]^2 = \sum_1^N [y_t - f(u_t,\theta,t)]^2$$

Il modello che meglio descrive i dati osservati è quello corrispondente al valore $\hat{\theta}$, di θ per cui $V(\hat{\theta})$ è *minimo*:

$$V(\hat{\theta}) = \min_{\theta\in\Theta} V(\theta).$$

Semplice regola empirica per costruire modelli parametrici di dati osservati e può in linea di principio essere usato per descrivere dati mediante *modelli di struttura affatto arbitraria*.

Ovviamente $\hat{\theta}$ dipende da $(y_1, \ldots, y_N)$ e dai valori assegnati $(u_1, \ldots, u_N)$ alla variabile esogena u negli N esperimenti.

$$\hat{\theta} = \hat{\theta}(y_1, \ldots, y_N \, ; u_1, \ldots, u_N) \quad ,$$

La funzione $\hat{\theta}$ viene anche chiama *stimatore ai minimi quadrati* di θ ma queste parole non hanno alcun significato statistico.

Spesso le misure effettuate non hanno tutte la stessa attendibilità ed è ragionevole dare peso *minore* agli errori di predizione corrispondenti a misure cattive. Questo porta all'introduzione dei cosiddetti *minimi quadrati pesati*, definendo il criterio quadratico pesato,

$$V_Q(\theta) := \sum_1^N q_t \left[y(t) - f(u_t, \theta, t)\right]^2 \ ,$$

dove $q_1, \ldots, q_N$ sono numeri positivi, grandi se le misure corrispondenti sono affidabili e piccoli se non lo sono. Si può riscrivere come

$$V_Q(\theta) = \left[y - f(u,\theta)\right]^\top \ Q \left[y - f(u,\theta)\right] = \|y - f(u,\theta)\|_Q^2 \quad ,$$

dove

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \quad , \qquad f(u,\theta) = \begin{bmatrix} f(u_1,\theta,1) \\ \vdots \\ f(u_N,\theta,N) \end{bmatrix} \tag{19}$$

e $Q = \mathrm{diag}\{q_1, \ldots, q_N\}$ o, più in generale, con Q non diagonale, ma sempre *definita positiva e simmetrica*.

MINIMI QUADRATI E MODELLI LINEARI

La minimizzazione di $V(\theta)$ si può fare esplicitamente nel caso in cui il modello è *lineare nei parametri*, cioè quando

$$f(u_t, \theta, t) = \sum_{1}^{p} s_i(u_t, t)\, \theta_i \,.$$

Siccome u_t è una quantità *nota* si può omettere nell'argomento e scrivere

$$f(u_t, \theta, t) := s^\top(t)\ \theta \quad ,$$

con $s^\top(t)$ vettore riga p-dimensionale, funzione *nota* dell'indice t. Il blocco delle N misure sia rappresentato dal vettore N-dimensionale y e S la matrice $N \times p$

$$S = \begin{bmatrix} s^\top(1) \\ \vdots \\ s^\top(N) \end{bmatrix} \,.$$

Il problema di descrivere i dati $(y_1, \ldots, y_N)$ mediante il modello lineare deterministico $y = S\theta$ diventa allora quello di minimizzare la forma quadratica

in θ,

$$V_Q(\theta) = [y - S\theta]^\top Q[y - S\theta] = \|y - S\theta\|_Q^2 \quad , \tag{20}$$

che abbiamo già risolto nella sezione precedente. Invocando il teorema della proiezione, il valore di θ che minimizza $V_Q(\theta)$ sarà quello per cui l'errore $y - S\theta$ è ortogonale (secondo la metrica definita dal prodotto scalare $\langle x, y \rangle_Q = x^\top Q y$) alle colonne di S, ovvero

$$S^\top Q(y - S\theta) = 0 \quad ,$$

che si può riscrivere come

$$S^\top Q\, S\theta = S^\top Q y\,. \tag{21}$$

ritrovando così le famose *equazioni normali* dei minimi quadrati.

Nel seguito supporremo che sia

$$\text{rango}\, S = p \leq N\,. \tag{22}$$

Questa condizione semplifica la trattazione del problema anche se non è essenziale per portare avanti l'analisi. In effetti, se il rango di S è minore

di p, il problema può essere riparametrizzato usando un numero minore di variabili $\{\theta_i\}$ e una matrice S con un numero minore di colonne, di rango pieno. Le equazioni normali si possono allora risolvere ottenendo un'unica soluzione

$$\hat{\theta}(y) = [S^\top Q S]^{-1} S^\top Q y \quad , \tag{23}$$

dalla quale si vede che lo stimatore ai M.Q. è *sempre una funzione lineare* delle misure. Nel caso in cui la (22) non valga, si può usare la pseudoinversa di $S^\top Q S$, ma in questo caso si perde l'unicità della soluzione.

Indichiamo con P il proiettore Q-ortogonale da $\mathbb{R}^N$ sul sottospazio generato dalle colonne di S. Come già visto, questo proiettore si può scrivere

$$P = S\,[S^\top Q S]^{-1} S^\top Q \tag{24}$$

e la somma pesata (secondo Q) dei quadrati degli *errori di predizione* corrispondenti a $\theta = \hat{\theta}$ (detti *residui di stima*)

$$\hat{\varepsilon}_t := y_t - s^\top(t)\,\hat{\theta} \quad , \qquad t = 1, \ldots, N \quad , \tag{25}$$

vale

$$\begin{aligned} V_Q(\hat{\theta}) &= \hat{\varepsilon}^\top Q\hat{\varepsilon} = \|y - Py\|_Q^2 = y^\top (I-P)^\top Q(I-P)\, y \\ &= y^\top Q(I-P)\, y = y^\top Qy - y^\top QPy = \|y\|_Q^2 - y^\top QP^2 y \\ &= \|y\|_Q^2 - y^\top P^\top QPy = \|y\|_Q^2 - \|Py\|_Q^2 = \|y\|_Q^2 - \|S\hat{\theta}(y)\|_Q^2 . \end{aligned}$$

(teorema di Pitagora) Avevamo già incontrato queste formule studiando le proprietà dello stimatore di M.V. di θ nel modello lineare e Gaussiano. In questo caso, il calcolo dello stimatore (di M.V.) di θ si riduceva a un problema ai minimi quadrati con matrice peso $Q = R^{-1}$, l'inversa della covarianza del rumore. Vale la pena di registrare esplicitamente questo fatto.

NOTA BENE: Lo stimatore di M.V. del parametro θ nel modello lineare Gaussiano (14) è uno stimatore ai M.Q. pesati con matrice Q uguale all'inversa della matrice di varianza del rumore $\mathbf{w}$.

MINIMI QUADRATI E MODELLI SATISTICI LINEARI

[Ipotesi sul meccanismo di generazione dei dati] Supponiamo che le misure $\{y_t\}$ siano generate da un modello lineare del tipo

$$\mathbf{y}_t = s^\top(t)\,\theta + \sigma \mathbf{w}_t \quad , \qquad t = 1, \ldots, N \quad , \tag{26}$$

in cui però gli errori di modellizzazione hanno distribuzione di probabilità non nota. Si sa solo che $\mathbf{w} = [\mathbf{w}_1, \ldots, \mathbf{w}_N]^\top$ è un vettore casuale N-dimensionale di media nulla e varianza $\sigma^2 R$ con R nota e definita positiva,

$$E\mathbf{w} = 0 \quad , \qquad \mathrm{Var}\,(\mathbf{w}) = E\mathbf{w}\mathbf{w}^\top = \sigma^2 R\,. \tag{27}$$

Nulla viene ipotizzato sulla distribuzione di probabilità di $\mathbf{w}$. In altre parole, l'informazione a priori sul modo in cui le misure sono generate è in questo caso molto più vaga.

Che proprietà statistiche ha lo stimatore ai M.Q. ?

Proposizione 11 *Qualunque sia $Q > 0$ lo stimatore ai MQ pesati è corretto.*

Difatti

$$\hat{\theta}(\mathbf{y}) = [S^\top Q S]^{-1} S^\top Q [S\theta + \bar{\mathbf{w}}] = \theta + [S^\top Q S]^{-1} S^\top Q \bar{\mathbf{w}} \quad ,$$

dato che $E\bar{\mathbf{w}} = 0$, $E\,\hat{\theta}(\mathbf{y}) = \theta$.

Se $Q = R^{-1}$ lo stimatore ai M.Q. pesati si chiama **stimatore di Markov**. Supponendo per un attimo che R sia diagonale, $R = \mathrm{diag}\{r_1, \ldots, r_N\}$, è evidente che la scelta della matrice peso Q più naturale in accordo con l'interpretazione che le abbiamo dato in termini di affidabilità delle misure è ovviamente quella di prenderla anch'essa diagonale con elementi

$$q_t = \frac{1}{\mathrm{var}\,\mathbf{y}_t} = \frac{1}{\sigma^2}\,\frac{1}{r_t}, \qquad t = 1, \ldots, N.$$

Notiamo che il termine $1/\sigma^2$ (incognito) non influisce sulla minimizzazione di $V_Q(\theta)$ dato che è indipendente da t e quindi si può portare fuori dal segno di sommatoria.

La varianza di $\hat{\theta}(\mathbf{y})$ vale

$$\mathrm{Var}\,\hat{\theta}(\mathbf{y}) = [S^\top Q S]^{-1} S^\top Q\, \sigma^2 R\, Q S [S^\top Q S]^{-1} \quad ;$$

l' espressione dipende ovviamente dalla matrice peso Q. È importante cercare la matrice dei pesi in corrispondenza alla quale la varianza di $\hat{\theta}$ è *minima*. (Qui usiamo come al solito "minima" nel senso dell'ordinamento fra matrici: $A \geq B$ se $A - B$ è semidefinita positiva).

Come abbiamo già detto, il principio dei M.Q. (pesati o no) applicato al modello lineare (26) può fornire solo stimatori che sono *funzioni lineari* delle osservazioni y. Ci si può allora chiedere in quali condizioni questo principio fornisce almeno *il miglior stimatore lineare* di θ, naturalmente nella classe di tutte le possibili funzioni lineari di $\mathbf{y}$,

$$\phi(\mathbf{y}) = A\mathbf{y} \quad , \qquad A \in R^{p \times N} \quad ,$$

che sono stimatori *corretti* di θ, ovvero

$$E\,\phi(\mathbf{y}) = \theta \qquad \text{ovvero} \qquad AS = I \quad .$$

Cerchiamo allora in questa classe quella funzione, $\hat{\phi}$, che ha *varianza minima*.
La soluzione di questo problema è fornita dal celebre

Teorema 5 (di Gauss-Markov) *Il miglior stimatore lineare di θ, per il modello (26) nel senso appena definito, è lo stimatore di Markov, che ha varianza*

$$\operatorname{Var} \hat{\phi}(\mathbf{y}) = \sigma^2 (S^\top R^{-1} S)^{-1} . \tag{28}$$

Si tratta di far vedere che la varianza di $A\mathbf{y}$, con A soddisfacente il vincolo $AS = I$, soddisfa alla disuguaglianza

$$\sigma^2 ARA^\top \geq \sigma^2 (S^\top R^{-1} S)^{-1} \quad , \tag{29}$$

che si può interpretare come un limite inferiore di Cramèr-Rao per stimatori lineari e corretti di θ. In effetti lo stimatore di Markov, definito dalla

$$\hat{A} = [S^\top R^{-1} S]^{-1} S^\top R^{-1} .$$

è lineare e corretto e la sua varianza è esattamente uguale al secondo membro in (29).

Per provare la (29) ci si rifà alla disuguaglianza (equivalente alla non-negatività della varianza dell'errore di stima nella teoria della stima lineare Bayesiana *),

$$ARA^\top \geq ARC^\top (CRC^\top)^{-1} CRA^\top \quad ,$$

valida per una arbitraria matrice di rango pieno $C \in R^{p \times n}$. Si verifica facilmente che scegliendo $C = \hat{A}$ e usando la $\hat{A}S = I$, si ottiene la (29). □

Se il processo d'errore $\{\mathbf{w}_t\}$ è stazionario e scorrelato, cioè

$$E(\mathbf{w}_t \mathbf{w}_s) = \sigma^2 \, \delta_{t,s} \quad , \qquad \forall t, s \quad ,$$

allora lo stimatore di Markov coincide con lo stimatore ordinario ai M.Q., quello che si ottiene minimizzando la somma dei quadrati degli errori di

*Sia $\mathbf{n}$ un vettore aleatorio a componenti ortonormali, $\mathbf{x} := AR^{1/2}\mathbf{n}$ e $\mathbf{y} := CR^{1/2}\mathbf{n}$. Si scriva l'espressione per la varianza dell'errore di stima $\tilde{\mathbf{x}} := \mathbf{x} - \hat{E}\,[\mathbf{x} \mid \mathbf{y}]$.

modellizzazione $\boldsymbol{\varepsilon}_t(\theta)$, espressi come funzione delle misure e del parametro θ,

$$\boldsymbol{\varepsilon}_t(\theta) := \mathbf{y}_t - s^\top(t)\,\theta \quad , \qquad t = 1, \ldots, N\,.$$

Notiamo infine che, in accordo con quanto anticipato nelle osservazioni precedenti, *lo stimatore di Markov di θ coincide con lo stimatore a M.V. nel caso in cui la distribuzione di probabilità di $\mathbf{w}$ nel modello (26) è (nota e) Gaussiana*. In genere però, se $\mathbf{w}$ non è normalmente distribuito, la varianza (28) può risultare assai più grande della varianza del corrispondente stimatore di M.V. di θ.

Proposizione 12 *Se nel modello lineare $\mathbf{y} = S\theta + \sigma\mathbf{w}$ è nota la matrice R, ma la distribuzione di probabilità è incognita (oppure non è Gaussiana), lo stimatore $\hat{\theta}(\mathbf{y}) = [S^\top R^{-1} S]^{-1} S^\top R^{-1}\,\mathbf{y}$ è comunque quello che ha minima varianza nella classe degli stimatori* **lineari e corretti**[†] *di θ.*

[†]Nella letteratura anglosassone lo stimatore lineare e corretto a minima varianza si denota con l'acronimo *B.L.U.E.* =Best Linear Unbiased Estimator.

STIMA DELLA VARIANZA

Stima di σ^2 corrispondente allo stimatore di Markov Per costruire uno stimatore di σ^2, si può ancora utilizzare la formula

$$\hat{\sigma}^2(y) = \frac{1}{N} \|y - Py\|^2_{R^{-1}} = \frac{1}{N} V_{R^{-1}}(\hat{\theta}) \,. \tag{30}$$

Interpretazione: $\hat{\sigma}(y)$ è lo scarto quadratico medio pesato con matrice $Q = R^{-1}$. È ovvio però che $\hat{\sigma}^2(y)$ non ha più distribuzione di tipo χ^2 in generale. Si può comunque calcolarne la media in modo diretto

$$\begin{aligned} NE\left(\hat{\sigma}^2(\mathbf{y})\right) &= E\left(\mathbf{y}^\top (I-P)^\top R^{-1}(I-P)\,\mathbf{y}\right) \\ &= E\left(\mathbf{w}^\top (I-P)^\top R^{-1}(I-P)\,\mathbf{w}\right) = E\left(\mathbf{w}^\top R^{-1}(I-P)\,\mathbf{w}\right) \\ &= E\,tr\{\mathbf{w}^\top R^{-1}(I-P)\,\mathbf{w}\} = E\,\mathrm{Tr}\{R^{-1}(I-P)\,\mathbf{w}\mathbf{w}^\top\} \\ &= E\,\mathrm{Tr}\{(I-P)\,(\mathbf{w}\mathbf{w}^\top)\,R^{-1}\} = \mathrm{Tr}\{(I-P)\,E(\mathbf{w}\mathbf{w}^\top)\,R^{-1}\} \\ &= \sigma^2\,\mathrm{Tr}(I-P) \quad ; \end{aligned} \tag{31}$$

nel primo passaggio si è usata l'identità $(I-P)\,\mathbf{y} = (I-P)\,S\theta + (I-P)\,\mathbf{w} = (I-P)\,\mathbf{w}$. Inoltre, come è noto, $\mathrm{Tr}\,P = \dim \mathscr{S} = p$ e quindi

$$E\,\hat{\sigma}^2(\mathbf{y}) = \frac{N-p}{N}\,\sigma^2\,. \tag{32}$$

Ne viene che $\frac{N}{N-p}\,\hat{\sigma}^2$ è uno stimatore corretto di σ^2.

Notiamo che se si vuole costruire uno *stimatore lineare* di $c^\top\theta$ anziché di θ ($c^\top$ è un vettore riga noto), quello corretto e di varianza minima (sempre nella classe degli stimatori lineari) è semplicemente $c^\top\hat{\theta}$, dove $\hat{\theta}$ è lo stimatore di Markov di θ. Nel caso si volesse stimare una funzione *non lineare*, $c(\theta)$, di θ, il procedimento, che continuerebbe a valere per la stima di M.V., non è più valido. Per poter calcolare uno stimatore non lineare servirebbero in generale tutti i momenti di $\hat{\theta}(\mathbf{y})$, mentre il modello ne fornisce solo due.

MINIMI QUADRATI E MODELLI NON LINEARI

Nei modelli di fenomeni fisici o biologici i parametri hanno un significato fisico o biologico e sono l'oggetto principale del procedimento di stima. Purtroppo essi appaiono spesso in modo non lineare. Problemi di stima di parametri in modelli non lineari richiedono preliminarmentre un'*analisi di identificabilità* del modello. Quando si cerca di determinare il valore di un parametro incognito da misure ingresso-uscita è essenziale sapere *a priori* se il problema è ben posto, almeno in condizioni ideali di assenza di disturbi o errori di misura. Perchè il problema sia ben posto occorre che la corrispondenza tra parametro e coppie ingresso-uscita (che soddisfano le equazioni del sistema) sia *iniettiva* altrimenti valori diversi del parametro potrebbero dar luogo allo stesso comportamento ingresso-uscita e essere indistinguibili. Questa verifica di *identificabilità a priori* è spesso difficile e allo stato non esistono metodologie generali per farla.

La stima vera e propria si fa con algoritmi iterativi che cercano di aggiornare la stima corrente del parametro in modo da dirigersi verso un

minimo dello scarto quadratico medio tra i dati misurati e quelli predetti dal modello (predittore) parametrizzato dalla stima corrente del parametro. Usando tecniche di linearizzazione è talvolta possibile dare delle valutazioni statistiche approssimate dell' accuratezza delle stime ottenute in questo modo.

Accenniamo al fatto che in molti problemi di regressione a scatola nera la descrizione dei dati mediante modelli parametrizzati linearmente (modelli lineari in θ) può risultare inadeguata. Negli ultimi decenni si è dedicato un enorme sforzo di ricerca per studiare le proprietà di approssimazione di classi parametriche di modelli (intrinsecamente non lineari nei parametri) chiamate **Reti Neurali**. Dato che anche una breve descrizione di queste classi di modelli ci porterebbe fuori dal tema principale di queste note, dobbiamo rimandare il lettore alla letteratura, non senza però avvertirlo che su questo argomento esiste una mole imponenete di materiale scritto da personaggi dalle dubbie credenziali scientifiche, che spesso si richiamano a fumose motivazioni neuro-biologiche che si rifanno a dei modelli primitivi

del “neurone” introdotti nl 1945 da McCulloch e Pitts che sono stati successivamente dimostrati essere grossolanamente irrealistici dal punto di vista fisiologico. Ciononostante è invalso in questo settore l’uso di un linguaggio di tipo mistico-biologico che poco o niente ha a che fare col soggetto e apparentemente serve unicamente a fare “audience”. Consigliamo di riferirsi agli articoli originali [**?**, **?**, **?**, **?**].

STATIC NEURAL NETWORKS

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be any function, and let $m; n; p$ be positive integers. A *single hidden layer net* with m inputs, p outputs, n hidden units, and activation function σ is specified by a pair of matrices $B\,C$ and a pair of vectors τ, c_0, where B and C are respectively real matrices of sizes $n \times m$ and $p \times n$, and τ and c_0 are respectively real vectors of size n and p. We denote such a net by a 5-tuple

$$\Sigma = \{B,\, C,\, \tau,\, c_0,\, \sigma\}$$

where σ is a "universal" function to be discussed later. In particular, Σ has *no offset* if $c_0 = 0$.
For simplicity, we will assume from now on that $p = 1$; (single output nets). Generalizations to the multiple-output case are not hard but complicate the notations.
Thus, from now on, $C \equiv \mathbf{c}^\top$ is a row n-vector and c_0 is a constant.

Let $\vec{\sigma}_n : \mathbb{R}^n \to \mathbb{R}^n$ indicate the application of σ to each coordinate of an n-vector:

$$\vec{\sigma}_n(x_1; \ldots; x_n = (\sigma(x_1); \ldots; \sigma(x_n))$$

The **behavior** of Σ is defined to be the map

$$\mathbf{\Sigma} : \mathbb{R}^n \to \mathbb{R} : \quad u \mapsto \mathbf{c}^\top \vec{\sigma}_n(Bu + \tau) + c_0 \,.$$

In other words, the behavior of a network is a composition of the type

$$f \circ \vec{\sigma}_n \circ g,$$

where $f(x) = \mathbf{c}^\top x + c_0$ and $g(u) = Bu + \tau$ are affine maps. Given two networks Σ and $\hat{\Sigma}$, we say that they are *input/output equivalent* and denote

$$\Sigma \sim \hat{\Sigma}$$

if $\mathbf{\Sigma} \equiv \hat{\mathbf{\Sigma}}$ (equality of functions). The identifiability question for (single layer) neural networks , then, is: when does $\mathbf{\Sigma} \equiv \hat{\mathbf{\Sigma}}$ imply $\Sigma = \hat{\Sigma}$?

"UNIVERSAL" APPROXIMATING FUNCTIONS

Although the hidden layer always involves replicas of the same nonlinear function σ, in the literature there are different choices of σ which nevertheless seem to work roughly the same. Two popular classes of functions are the so-called **sigmoid**

$$\sigma(x) = \frac{1}{1+e^{-ax}}, \qquad a > 0$$

and the **radial function**

$$\sigma(x) = \phi(-a|x|^2), \qquad a > 0.$$

The success of neural networks as approximation devices probably stems from the ability that linear combinations of shifted activation function the type

$$\sum_k c_k \sigma(x - \tau_k)$$

have to approximate arbitrary nonlinear functions. In this respect an old result of Wiener states this fact in rigorous terms as follows.

Teorema 6 (Wiener 1932) *In order for the family of shifted versions* $x \mapsto f(x+\tau)\,;\tau \in \mathbb{R}$ *of a function* $f \in L^2(\mathbb{R})$*, to be dense in* $L^2(\mathbb{R})$*, it is necessary and sufficient that the Fourier transform* $\hat{f}(j\omega)$ *be nonzero almost everywhere.*

In other words, an arbitrary $g \in L^2(\mathbb{R})$ can be approximated arbitrarily closely (in $L^2(\mathbb{R})$) by a linear combination of shifts $\sum_k c_k f(x+\tau_k),$ of the function f, if and only if $\hat{f}(j\omega)$ is nonzero almost everywhere.

Although polynomial functions such as $\sum_k a_k x^k$ are obviously not L^2 functions, nevertheless, their generalized Fourier transform in the sense of distributions, is a sum of derivatives of Dirac δ functions whose support is concentrated at the zero frequency. It may then be guessed that these functions should have poor approximation properties in the sense defined above. In fact it is trivial to check that the span of shifted polynomials of degree n in x is still a polynomial of (at most) the same degree. Hence the popular approximation by "Taylor series"-like linearly parametrized models, turns out to be a very bad activation function.

GRADIENT DESCENT AND BACK-PROPAGATION

See Duda Hart p.290-292

ASPETTI NUMERICI

La soluzione al calcolatore delle equazioni normali

$$S^\top Q S\,\theta = S^\top Q y$$

può risultare problematica se p è "grande". Gli errori di arrotondamento possono venire esaltati e amplificati di molti ordini di grandezza durante il procedimento di calcolo a meno di non seguire delle avvertenze particolari.

Il calcolatore usa un sistema approssimato di rappresentazione dei numeri reali ("floating point arithmetic"). Un numero reale α viene rappresentato come una coppia $\alpha = (m, c)$ dove m è la *mantissa* di α e c la sua *caratteristica*. La mantissa è un numero il cui modulo è compreso tra 0,1 e 1 e contiene un *numero fisso*, n, di cifre significative (ad esempio 6). La caratteristica è l'esponente di 10 tale per cui $\alpha \cong 10^C$. Ad esempio $\alpha = 3,562417\bar{9}$ ha le rappresentazioni

$$f\ell(\alpha) = 0.356242 \qquad 10^1 \qquad \text{se } n = 6$$
$$f\ell(\alpha) = 0.35624 \qquad 10^1 \qquad \text{se } n = 5 \quad \text{ecc.}$$

Gli errori che risultano da questa approssimazione si chiamano errori di *arrotondamento*.

Molti problemi numerici possono essere descritti nel modo seguente: si ha una funzione $f : \mathbb{R}^k \to \mathbb{R}^p$ definita matematicamente e un vettore k-dimensionale di "dati" α. Si vuole calcolare $x = f(\alpha)$. Ad esempio, nella soluzione del problema

$$Ax = b \quad , \tag{33}$$

i dati sono $\alpha = (A, b)$ e la funzione f è definita da $f(\alpha) = A^{-1}b$.

Bisogna tenere presenti due aspetti del problema.

A) I dati, α, vengono rappresentati con un'aritmetica finita nel calcolatore e sono pertanto affetti da errori di arrotondamento, $\delta\alpha$ (nel calcolatore viene immagazzinato $\alpha + \delta\alpha$, *non* α).

B) Non è in generale possibile implementare algoritmi che calcolano *esattamente* la funzione f. In generale bisogna (o è più conveniente per varie ragioni) ricorrere ad approssimazioni. In pratica f viene calcolata in modo approssimato; l'algoritmo che si programma fornisce una approssimazione, $g(\cdot)$, di $f(\cdot)$.

Queste due cause d'errore, se pur distinte (la prima dipende dal numero di cifre significative che si usano nella rappresentazione di α e la seconda dalla "bontà" dell'algoritmo numerico che calcola f) tendono sempre a sommarsi.

Definizione 11 *Diremo che il problema numerico* $x = f(\alpha)$ *è* **mal condizionato** *se a piccoli errori percentuali su* α *corrispondono grandi errori percentuali su* x*. In altri termini, detto* $x = f(\alpha)$ *e* $x + \delta x = f(\alpha + \delta\alpha)$ *si ha*

$$\frac{\|\delta x\|}{\|x\|} \gg \frac{\|\delta\alpha\|}{\|\alpha\|}. \tag{34}$$

Esempio

$$\begin{bmatrix} 1 & 1 \\ 1 & 1.0001 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2.0001 \end{bmatrix} \quad ;$$

la cui soluzione è $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Se introduciamo una perturbazione nel secondo membro, ad esempio

$$b + \delta b = \begin{bmatrix} 2 \\ 2.0002 \end{bmatrix} \quad ,$$

la soluzione x diventa

$$x + \delta x = \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$

In questo caso $\|\delta b\|/\|b\| \cong 10^{-4}$, mentre $\|\delta x\|/\|x\| = 1/\sqrt{2}$. Si vede che l'errore δb viene "amplificato" nel calcolo (esatto!) della soluzione del sistema di molti ordini di grandezza. Nel libro di Wilkinson, *The Algebraic Eigenvalue Problem* (Oxford U.P. 1963), è mostrato che il fattore di amplificazione nella soluzione di

$$\begin{bmatrix} 0,501 & -1 & 0 & & \\ 0 & 0,502 & -1 & & \\ & & \ddots & -1 & \\ & & & 0,600 \end{bmatrix} x = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

è dell'ordine di 10^{22} !

Il fatto che un problema numerico sia mal condizionato è una sua *caratteristica intrinseca* che non può essere modificata dall'algoritmo con cui si calcola effettivamente $x = f(\alpha)$. Per contro anche un problema ben condizionato può essere "rovinato" dall'uso di algoritmi inadatti, che esaltano gli errori di round-off ecc...

Intuitivamente, un “buon” algoritmo dal punto di vista numerico perturba f in modo tale da non peggiorare di molto gli errori in x dovuti all’aritmetica finita del calcolatore.

Definizione 12 *Un algoritmo g, per il problema $x = f(\alpha)$, è* **numericamente stabile** *se per ogni $\alpha \in \mathbb{R}^k$ c’è una perturbazione $\delta\alpha$, di α (percentualmente) dello stesso ordine degli errori di arrotondamento, tale che $f(\alpha+\delta\alpha)$ e $g(\alpha)$ differiscono (percentualmente) di una quantità dello stesso ordine di $f(\alpha+\delta\alpha)-f(\alpha)$.*

In altre parole, gli errori introdotti da un algoritmo numericamente stabile *possono sempre essere imputati ad errori di arrotondamento con cui si rappresentano i dati*. Per dimostrare che un algoritmo g è numericamente stabile bisogna far vedere quindi che la soluzione reale $y = g(\alpha)$ si può ottenere come soluzione teorica di un problema con dati perturbati (cioè $y = f(\alpha+\delta\alpha)$) in cui la perturbazione $\|\delta\alpha\|/\|\alpha\|$ è dello stesso ordine di quella introdotta dall’arrotondamento.

Chiaramente nessun algoritmo, per quanto stabile esso sia, è in grado di fornire soluzioni accurate di un problema mal condizionato. C'è però la garanzia che un algoritmo stabile non "rovina" un problema ben condizionato.

CONDIZIONAMENTO NUMERICO

Le equazioni normali sono del tipo $Ax = b$. Supponiamo per il momento che A possa essere immagazzinata esattamente dal calcolatore ($\delta A = 0$). Il problema è di caratterizzare il legame tra gli errori relativi $\|\delta b\|/\|b\|$ e $\|\delta x\|/\|x\|$. Useremo sempre norme *Euclidee*

$$\begin{aligned} \|x\| &:= |\sum_i x_i^2|^{1/2} \quad , \\ \|A\| &= \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} . \end{aligned}$$

Dall'ultima definizione si vede che $\|A\|$ è il più piccolo numero $k > 0$ per cui $\|Ax\| \leq k\|x\|$. $\|A\|$ si può calcolare come segue:

$$\|A\|^2 = \sup_{x \neq 0} \frac{x^\top A^\top A x}{x^\top x} . \tag{35}$$

Il quoziente al secondo membro è noto come *quoziente di Rayleigh* ed è *uguale all' autovalore massimo di* $A^\top A$,

$$\|A\|^2 = \max_i \lambda_i(A^\top A) . \tag{36}$$

Di fatto, dato che $A^\top A$ è simmetrica, la matrice degli autovettori normalizzati

$$U := \begin{bmatrix} u_1 & \dots & u_n \end{bmatrix}, \qquad UU^\top = U^\top U = I$$

diagonalizza $A^\top A$

$$U^\top A^\top A U = \operatorname{diag}\{\lambda_1, \dots, \lambda_n\} \quad \lambda_1 \geq \dots \geq \lambda_n \geq 0$$

e cambiando base: $y := U^\top x$ ovvero $x^\top = y^\top U$,

$$\frac{x^\top A^\top A x}{x^\top x} = \frac{y^\top \operatorname{diag}\{\lambda_1, \dots, \lambda_n\} y}{y^\top y} = \frac{\lambda_1 y_1^2 + \dots + \lambda_n y_n^2}{y_1^2 + \dots + y_n^2} \leq \frac{\lambda_1 \|y\|^2}{\|y\|^2}$$

Il massimo si ottiene per $y_2 = y_3 = \dots = y_n = 0$ e vale λ_1.

Usando la norma di A, si vede facilmente che da $x = A^{-1}b$ e $b = Ax$ seguono le $\|\delta x\| \leq \|A^{-1}\|\|\delta b\|$ e $\|x\| \geq \|A\|^{-1}\|b\|$, per cui

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\|\,\|A^{-1}\|\,\frac{\|\delta b\|}{\|b\|}$$

Il numero $c(A) := \|A\|\,\|A^{-1}\|$ è il coefficiente di amplificazione degli errori sul termine noto b. È chiamato (indice di) **condizionamento numerico** del problema $Ax = b$ (o della matrice A). Vedremo che $c(A)$ descrive completamente il condizionamento numerico del problema $Ax = b$. Da $I = AA^{-1}$ scende che

$$1 = \|I\| \leq \|A\|\,\|A^{-1}\| = c(A)$$

e perciò $c(A)$ è effettivamente, sempre, un coefficiente di *amplificazione*. Ricordando che

$$\begin{aligned} \|A\|^2 &= \lambda_{\mathrm{MAX}}\,(A^\top A) \\ \|A^{-1}\|^2 &= \lambda_{\mathrm{MAX}}\,(A^{-\top}A^{-1}) = \lambda_{\mathrm{MAX}}\,(AA^\top)^{-1} = \frac{1}{\lambda_{\mathrm{MIN}}(AA^\top)} \end{aligned}$$

e tenendo conto che $A^\top A$ e $AA^\top$ hanno gli stessi autovalori (se $AA^\top a = \lambda_0 a$ allora $A^\top A(A^\top a) = \lambda_0(A^\top a)$ e λ_0 è anche un autovalore di $A^\top A$ con

autovettore $A^\top a$) si vede che

$$c^2(A) = \frac{\lambda_{\mathrm{MAX}}(A^\top A)}{\lambda_{\mathrm{MIN}}(A^\top A)} \quad ; \tag{37}$$

in particolare, se A è simmetrica,

$$c(A) = \frac{\lambda_{\mathrm{MAX}}(A)}{\lambda_{\mathrm{MIN}}(A)}. \tag{38}$$

Da questa formula si vede che se A è prossima a essere singolare, il condizionamento numerico $c(A)$ puo' essere grande. Però una matrice prossima a essere singolare come εI con $\varepsilon \to 0$ ha condizionameno numerico uguale a uno. Chiaramente le matrici meglio condizionate sono quelle per cui $A^\top A = \alpha I$. In questo caso infatti $\lambda_{\mathrm{MAX}}(A^\top A) = \lambda_{\mathrm{MIN}}(A^\top A) = 1$ e $c(A) = 1$. Queste sono le matrici *ortogonali**. Esse giocano un ruolo fondamentale nell'analisi numerica.

A titolo di esempio calcoliamo il condizionamento numerico del problema

*Per coerenza, bisognerebbe dire che una matrice per cui $AA^\top = I$ è *ortonormale*.

2×2 precedente. La matrice A è simmetrica e si trova (approssimativamente)

$$\lambda_{\mathrm{MAX}} = 2 \quad ; \quad \lambda_{\mathrm{MIN}} = 10^{-4}/2 \quad ,$$

da cui $c(A) = 4\ 10^4$ che è in accordo con i risultati riportati nell'esempio precedente.

Esercizio:

Dimostrare che se A è simmetrica e b è parallelo all'autovettore di A corrispondente a λ_{MIN}, mentre δb è parallelo all'autovettore di A corrispondente a λ_{MAX}, si ha *esattamente*

$$\frac{\|\delta x\|}{\|x\|} = c(A)\, \frac{\|\delta b\|}{\|b\|} .$$

Generalizzare al caso in cui A non è simmetrica. $\diamondsuit$

Esaminiamo l'effetto degli errori di arrotondamento su A. Supponiamo $\delta b = 0$. Con semplici calcoli si ricava che la perturbazione δx nella soluzione di $(A + \delta A)\,\bar{x} = b$ soddisfa (al prim'ordine)

$$\delta A \; \delta x = b \quad ,$$

dove $x = x + \delta x$ e $Ax = b$. Se ne ricava,

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{c(A)\,\frac{\|\delta A\|}{\|A\|}}{1 - c(A)\,\frac{\|\delta A\|}{\|A\|}} .$$

Se $c(A)\,\|\delta A\|/\|A\|$ è trascurabile rispetto a 1,

$$\frac{\|\delta x\|}{\|x\|} \leq c(A)\,\frac{\|\delta A\|}{\|A\|} \quad , \tag{39}$$

che è una disuguaglianza dello stesso tipo della precedente. Si vede che il fattore di amplificazione $c(A)$ descrive il condizionamento del problema sia rispetto a errori sul termine noto b che sulla matrice A.

LA DECOMPOSIZIONE AI VALORI SINGOLARI (SVD)

Richiamiamo un risultato di algebra delle matrici che, nonostante sia estremamente utile, spesso non viene insegnato nei corsi di base. Si tratta della cosiddetta *Decomposizione ai Valori Singolari* (SVD) di una matrice.

Teorema 7 *Sia* $A \in \mathbb{R}^{m\times p}$ *una matrice di rango* $n \leq \min(m,p)$. *Esistono due matrici ortogonali* $U \in \mathbb{R}^{m\times m}$ *e* $V \in \mathbb{R}^{p\times p}$ *e una successione ordinata di numeri reali positivi* $\{\sigma_1 \geq \ldots \geq \sigma_n\}$, *detti* valori singolari *di* A, *tali che*

$$A = U\Delta V^\top \tag{40}$$

dove Δ *ha la struttura quasi diagonale:*

$$\Delta = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}, \qquad \Sigma = \mathrm{diag}\{\sigma_1, \ldots, \sigma_n\} \tag{41}$$

La matrice $U = [u_1, \ldots, u_m]$ *si può costruire prendendo come colonne gli autovettori normalizzati di* $AA^\top$; *analogamente,* $V := \left[v_1, \ldots, v_p\right]$ *si può costruire prendendo come colonne gli autovettori normalizzati di* $A^\top A$. *I quadrati*

dei valori singolari $\{\sigma_1^2 \geq \ldots \geq \sigma_n^2\}$ *sono gli autovalori non nulli di* $AA^\top$ *(o di* $A^\top A$*).*

Prova: Siano $\left[v_1, \ldots, v_p\right]$, p autovettori ortonormali di $A^\top A$ di modo che

$$A^\top A v_k = \sigma_k^2 v_k \qquad k = 1, \ldots n$$

e $A^\top A v_k = 0$ per $k > n$. Notare che gli ultimi $p - n$ autovettori possono essere scelti in modo sostanzialmente arbitrario. Moltiplicando a sinistra per A si ottiene

$$AA^\top (A v_k) = \sigma_k^2 (A v_k) \qquad k = 1, \ldots n$$

Si verifica che gli autovettori di $AA^\top$ normalizzati tramite la

$$u_k := \frac{1}{\sigma_k} A v_k \qquad k = 1, \ldots n\,,$$

sono ortonormali. Infatti

$$\langle u_k, u_j \rangle = \frac{v_k^\top A^\top A v_j}{\sigma_k \sigma_j} = \frac{\sigma_j^2}{\sigma_k \sigma_j} \langle v_k, v_j \rangle = \frac{\sigma_j^2}{\sigma_k \sigma_j} \delta_{kj}$$

Completiamo ora la famiglia $\{u_1, \ldots, u_n\}$ con altri $m-n$ (auto)vettori nello spazio nullo di $AA^\top$ in modo da ottenere una base ortonormale in $\mathbb{R}^m$. Un semplice calcolo fornisce

$$u_k^\top A v_j = \frac{v_k^\top A^\top A v_j}{\sigma_k} = \frac{\sigma_j^2}{\sigma_k}\langle v_k, v_j\rangle = \frac{\sigma_j^2}{\sigma_k}\delta_{kj}$$

per $k, j \leq n$ e $u_k^\top A v_j = 0$ altrimenti. Queste relazioni sono equivalenti alla $U^\top A V = \Delta$ e quindi alla relazione (40). □

Possiamo così dire che *l'indice di condizionamento numerico di una matrice è il raporto tra il suo massimo e il minimo valore singolare*,

$$c(A) = \frac{\sigma_1(A)}{\sigma_n(A)}\,. \tag{42}$$

La SVD fornisce la descrizione più completa della struttura di una trasformazione lineare. Dalla (40) si ricava, eliminando i prodotti con i blocchi nulli di Δ, la seguente *fattorizzazione a rango pieno di* A

$$A = [u_1, \ldots, u_n]\,\Sigma\,[v_1, \ldots, v_n]^\top := U_n \Sigma V_n^\top \tag{43}$$

dove U_n, V_n sono le sottomatrici di U, V ottenute eliminando le ultime $m-n$ e $p-n$ colonne. Notiamo che le n colonne di U_n e le n righe di $V_n^\top$ sono ancora ortonormali, i.e.

$$U_n^\top U_n = I_n = V_n^\top V_n \,.$$

La cosiddetta *norma di Frobenius* $\|A\|_F$ è la radice quadrata della somma dei quadrati degli elementi, i.e. $\|A\|_F^2 = \sum_{i,j} a_{i,j}^2 = \mathrm{Tr} AA^\top = \mathrm{Tr} A^\top A$.

Corollario 1 *Lo spazio immagine e lo spazio nullo di* A *sono :*

$$\mathrm{Im}\,(A) = \mathrm{Im}\,(U_n), \qquad \mathrm{Ker}\,(A) = \mathrm{Im}\,(\left[v_{n+1}, \ldots, v_p\right])$$

Inoltre,

$$\|A\|_2 = \|\Sigma\|_2 = \sigma_1, \qquad \|A\|_F^2 = \|\Sigma\|_F^2 = \sigma_1^2 + \ldots + \sigma_n^2$$

La matrice

$$A_k := \sum_{i=1}^{k} \sigma_i u_i v_i^\top, \qquad k \leq n$$

è la miglior approssimante di rango k di A; infatti

$$\min_{B\,;\,\mathrm{rank}\,(B)=k} \|A-B\|_2 = \|A-A_k\|_2 = \sigma_{k+1} \tag{44}$$

e inoltre

$$\min_{B\,;\,\mathrm{rank}\,(B)=k} \|A-B\|_F^2 = \|A-A_k\|_F^2 = \sigma_{k+1}^2 + \ldots + \sigma_n^2 \tag{45}$$

La dimostrazione di queste proprietà si può trovare ad esempio nel [Golub-Van Loan] p. 584.

CONDIZIONAMENTO NUMERICO DEI M.Q.

Supponiamo A quadrata di voler risolvere il problema $Ax = b$ moltiplicando a sinistra i due membri per $A^\top$. Si trova così

$$A^\top Ax = A^\top b \quad ;$$

ora il condizionamento numerico di questo problema non è più quello di A, **ma bensì quello di** $A^\top A$. Evidentemente,

$$c(A^\top A) = \|A^\top A\| \, \|(A^\top A)^{-1}\| = \lambda_{\mathrm{MAX}}(A^\top A)/\lambda_{\mathrm{MIN}}(A^\top A) = c^2(A)\,.$$

Ne segue che, anche per problemi $Ax = b$ moderatamente ben condizionati, $A^\top Ax = b$ può risultare assai mal condizionato. Se $c(A) \cong 10^c$, il naturale c dà il numero di cifre significative che si perdono nella soluzione di $Ax = b$. Siccome $c(A)^2 = 10^{2c}$, risolvendo il problema (apparentemente identico) $A^\top Ax = A^\top b$ si perdono esattamente **il doppio** di cifre significative.

L' argomento non è esattamente calzante, dato che con i minimi quadrati si cerca di risolvere un sistema lineare del tipo

$$y = S\theta \quad , \qquad (*)$$

che è sempre *incompatibile* perché il numero di equazioni N è sempre molto maggiore di p, ma serve ugualmente a spiegare qualitativamente il fenomeno. In realtà si dimostra che nella formula per $c(S)$ entra la la pseudoinversa invece che A^{-1}.

Il metodo di attacco al problema dei MQ sviluppato dagli analisti numerici è di lavorare direttamente sul sistema $(*)$ e **dimenticare le equazioni normali !**

RUOLO DELL' ORTOGONALITÀ

Sia data una funzione $f(x)$ sull'intervallo [0,1] e supponiamo di voler trovare il polinomio $P_n(x)$ di grado fissato, n, che approssima meglio $f(x)$ nel senso dei minimi quadrati. Si vuole trovare cioè $\hat{P}_n(x)$ tale che

$$\int_0^1 |f(x) - P_n(s)|^2 dx$$

sia minimo. Scriviamo $P_n(x)$ come

$$P_n(x) = \theta_0 1 + \theta_1 x + \ldots + \theta_n x^n = [1\, x \ldots x^n] \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_n \end{bmatrix} := s^\top(x)\, \theta\,,$$

dove $s^\top(x) = [1\, x \ldots x^n]$. Riferendosi al prodotto scalare $\langle f, g \rangle = \int_0^1 f(x)\, g(x)\, dx$ e imponendo il principio di ortogonalità

$$f(x) - \sum_0^n \theta_i\, x^i \perp \text{span}\,\{1\, x \ldots x^n\}$$

si trovano le equazioni normali per questo problema,

$$\begin{bmatrix} \langle 1,1\rangle & \langle 1,x\rangle & \dots & \langle 1,x^n\rangle \\ \vdots & & & \vdots \\ \langle x^n,1\rangle & \dots & \dots & \langle x^n,x^n\rangle \end{bmatrix} \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_n \end{bmatrix} = \begin{bmatrix} \langle 1,f\rangle \\ \vdots \\ \langle x^n,f\rangle \end{bmatrix}.$$

La matrice simmetrica a primo membro è l'analoga di $S^\top S$.

A conti fatti, si trova

$$\begin{bmatrix} 1 & 1/2 & \dots & \frac{1}{n+1} \\ 1/2 & 1/3 & & \frac{1}{n+2} \\ \vdots & & & \\ \frac{1}{n+1} & \frac{1}{n+2} & \dots & \frac{1}{2n+1} \end{bmatrix} \theta = \begin{bmatrix} \langle 1, f \rangle \\ \vdots \\ \langle x^n, f \rangle \end{bmatrix}.$$

La matrice a primo membro (detta matrice di Hilbert) è terribilmente mal condizionata. Per $n = 10$ il suo condizionamento numerico è all'incirca 10^{13}. Questo sembra rendere l'approssimazione polinomiale un problema impossibile anche per valori non troppo elevati di n. In realtà si sa bene che si tratta di un problema di routine in analisi numerica. L'idea chiave per la sua soluzione è quella di usare *polinomi ortogonali*. Se invece di avere $(1\, x \dots x^n)$ si disponesse di polinomi indipendenti $p_0(x)\, p_1(x) \dots p_n(x)$ tali che $\langle p_i, p_j \rangle = \delta_{ij}$, l'approssimazione ai minimi quadrati

$$f(x) \cong \sum_0^n \theta_i \, p_i((x)$$

si potrebbe semplicemente ottenere calcolando i prodotti scalari

$$\langle f - \sum_{0}^{n} \theta_i \, p_i(x) \, ; \, p_j \rangle = 0 \qquad j = 0, 1, \ldots, n \, ,$$

e ricavandone immediatamente

$$\hat{\theta}_j = \langle f, p_j \rangle \quad , \qquad j = 0, 1, \ldots, n \, .$$

Questo è il metodo che si usa in pratica e che sta, ad esempio, a fondamento dei vari metodi di sviluppo in serie di funzioni ortonormali (come, in particolare la serie di Fourier).

M.Q. E SERIE DI FOURIER

Data una funzione $y(t)$ nell'intervallo $[-T/2, T/2]$, vogliamo trovare una combinazione lineare delle funzioni 1, $\sin\frac{2\pi}{T}t, \ldots, \sin\frac{2n\pi}{T}t$, $\cos\frac{2\pi}{T}t$, $\ldots$, $\cos\frac{2n\pi}{T}t$, secondo i coefficienti θ_i, $i = 0, 1, \ldots, 2n$ (questa è la ridotta n-sima della serie di Fourier di $y(t)$) diciamola

$$\begin{aligned} f_n(t,\theta) \;\; := \;\; & \theta_0 + \theta_1 \, \sin\frac{2\pi}{T}t + \theta_2 \, \cos\frac{2\pi}{T}t + \ldots \\ & + \theta_{2n-1} \, \sin\frac{2n\pi}{T}t + \theta_{2n} \, \cos\frac{2n\pi}{T}t \end{aligned}$$

tale da minimizzare lo scostamento quadratico

$$V(\theta) = \int_{-T/2}^{T/2} |y(t) - f_n(t,\theta)|^2 \, dt$$

Le funzioni in gioco si pensano come vettori nello spazio di funzioni su $[-T/2, T/2]$ dotato del prodotto scalare $\langle f, g \rangle = \int_{-T/2}^{T/2} f(t)\, g(t) \frac{dt}{T}$

Il valore del parametro $2n+1$-dimensionale θ che minimizza questo funzionale è proprio il vettore dei primi $2n+1$ coefficienti di Fourier di y. In altri termini i coefficienti di Fourier di y sono stime ai minimi quadrati dei parametri del modello (lineare) $f_n(t,\theta) \simeq S\theta$ usato per approssimare y.

L'ortogonalità delle funzioni base usate per la modellizzazione (le colonne della matrice S!) semplifica in modo drammatico la stima dei coefficienti.

LA FATTORIZZAZIONE QR

Consideriamo per semplicità minimi quadrati non pesati, ma questa semplificazione non costituisce perdita di generalità. Vogliamo calcolare la stima ai M.Q. di θ partendo da N osservazioni y descritte dal modello

$$y = S\theta + \varepsilon \quad ,$$

dove $\varepsilon = \sigma w$ è il vettore degli errori di approssimazione delle misure, y con il modello $S\theta$.

Le p colonne di $S = [s_1, \ldots, s_p]$ sono linearmente indipendenti, ma in generale non ortonormali. Se lo fossero, $\langle s_i, s_j \rangle = s_i^\top s_j = \delta_{ij}$ e si avrebbe $S^\top S = I$ per cui la stima $\hat{\theta}$ si ricaverebbe immediatamente,

$$\hat{\theta} = S^\top y = \begin{bmatrix} \langle s_1, y \rangle \\ \ldots \\ \langle s_p, y \rangle \end{bmatrix} .$$

$\hat{\theta}$ è il vettore delle prime p coordinate di y rispetto alla base ortonormale $\{s_1, s_2, \ldots, s_p, \ldots\}$ in $\mathbb{R}^N$.

L'idea della fattorizzazione QR è semplicemente quella di *ortonormalizzare* le colonne di S.

Come è noto, l'algoritmo di Gram-Schmidt processa sequenzialmente i vettori colonna di $S = [s_1, \ldots, s_p]$ e fornisce altrettanti vettori ortonormali $\{q_1, \ldots, q_p\}$ definiti dalle relazioni

$$
\begin{aligned}
v_1 &= s_1 && , && q_1 := v_1/\|v_1\| \\
v_2 &= s_2 - \langle s_2, q_1 \rangle\, q_1 && , && q_2 := v_2/\|v_2\| \\
&\vdots && && \vdots \\
v_k &= s_k - \langle s_k q_1 \rangle\, q_1 + \ldots + \langle s_k q_{k-1} \rangle\, q_{k-1} && , && q_k := v_k/\|v_k\| \,.
\end{aligned}
$$

Questa porta ad una fattorizzazione di S di struttura assai particolare. Risolviamo rispetto a $(s_1, \ldots, s_p)$:

$$
\begin{aligned}
s_1 &= \|v_1\|\, q_1 \\
s_2 &= \langle s_2, q_1 \rangle\, q_1 + \|v_2\|\, q_2 \\
&\vdots \\
s_p &= \langle s_p q_1 \rangle\, q_1 + \ldots + \langle s_p, q_{p-1} \rangle\, q_{p-1} + \|v_p\|\, q_p \quad ,
\end{aligned}
$$

ovvero

$$[s_1,\ldots,s_p] = [q_1,\ldots,q_p] \begin{bmatrix} \|v_1\| & \langle s_2,q_1\rangle & \ldots & \langle s_p,q_1\rangle \\ 0 & \|v_2\| & & \\ \vdots & 0 & & \\ \vdots & \vdots & & \\ 0 & 0 & & \|v_p\| \end{bmatrix} \quad ;$$

che si può scrivere simbolicamente come

$$S = \bar{Q}\bar{R} \quad , \tag{46}$$

dove $\bar{Q}$ è una matrice a *colonne ortonormali*, cioè $\bar{Q}^\top \bar{Q} = I \ (p \times p)$ e $\bar{R}$ è *triangolare superiormente*. Se completiamo la base $\{q_1,\ldots,q_p\}$ con $N-p$ vettori $\{q_{p+1},\ldots,q_N\}$ in modo da ottenere una base ortonormale per $\mathbb{R}^N$ e introduciamo le matrici

$$Q := \left[\bar{Q} \mid q_{p+1}\cdots q_N\right], \qquad R := \begin{bmatrix} \bar{R} \\ 0 \end{bmatrix} \quad ,$$

S si può scrivere come

$$S = QR \quad , \tag{47}$$

cioè S *viene fattorizzata come il prodotto di una matrice ortogonale e una triangolare superiormente*.

Questa è la famosa *fattorizzazione* QR di S. Le equazioni (**??**) forniscono un algoritmo ricorsivo per il calcolo di $\bar{Q}$ ed $\bar{R}$. Per ottenere la (47) basta aggiungere a $\bar{Q}$ a destra $N-p$ colonne ortonormali (In realtà questa operazione non è necessaria).

Se moltiplichiamo i due membri della $y = S\theta + \varepsilon$ per $Q^\top$ si ottiene

$$Q^\top y = Q^\top S\theta + Q^\top \varepsilon \quad ,$$

ovvero

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \bar{R} \\ 0 \end{bmatrix} \theta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \quad , \tag{48}$$

dove y_1 e y_2 sono i vettori delle componenti di y rispetto a $(q_1 \dots q_p)$ e $(q_{p+1} \cdots q_N)$.

Notiamo subito che

$$\text{span}\{q_1 \dots q_p\} = \text{span}\{s_1 \dots s_p\} = \mathscr{S}$$

e pertanto

$$\text{span}\{q_{p+1} \dots q_N\} = \mathscr{S}^{\perp}.$$

Ne deriva che $\begin{bmatrix} y_1 \\ 0 \end{bmatrix}$ è la proiezione di y su $\mathscr{S}$ (espressa nelle coordinate $\{q_i\}$) e $\begin{bmatrix} 0 \\ y_2 \end{bmatrix}$ è la proiezione di y sul sottospazio $\mathscr{S}^{\perp}$ e concide quindi con il *residuo di stima* $\hat{\varepsilon} = y - Py$. Il significato di ε_1 ed ε_2 verrà discusso più avanti.

Ora, il principio (deterministico) dei minimi quadrati consiste nel cercare il valore di θ che minimizza la norma dell'errore di approssimazione $\varepsilon = \varepsilon(\theta)$,

$$\|\varepsilon(\theta)\|^2 = \|y - S\theta\|^2$$

e dalla (48) si vede, data l'ortogonalità di $Q^\top$, che

$$\begin{aligned} \|\varepsilon(\theta)\|^2 &= \|Q^\top \varepsilon(\theta)\|^2 = \|Q^\top y - Q^\top S\theta\|^2 \\ &= \left\| \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} - \begin{bmatrix} \bar{R}\theta \\ 0 \end{bmatrix} \right\|^2 = \|y_1 - \bar{R}\theta\|^2 + \|y_2\|^2 . \end{aligned}$$

Da questa relazione segue immediatamente che

1) $\hat{\theta}$ è soluzione di

$$\bar{R}\theta = y_1 \quad , \tag{49}$$

con $\bar{R}$ matrice triangolare superiormente.

2) Il *residuo* $\hat{\varepsilon} = \varepsilon(\hat{\theta})$ ha norma pari a

$$\|\hat{\varepsilon}\|^2 = \|y_2\|^2 . \tag{50}$$

In altri termini, nel nuovo sistema di coordinate, ε_1 rappresenta la parte dell'errore di approssimazione $\varepsilon(\theta)$ che può essere *resa nulla* con la scelta

$\theta = \hat{\theta}$ ($\theta = \hat{\theta}$ è la scelta di θ con cui si riesce a descrivere *esattamente*, tramite il modello $S\theta$, le prime p misure, y_1). In conclusione, se si ortonormalizzano le colonne di S la soluzione del problema ai M.Q. si riduce a risolvere un'equazione algebrica come la (49) in cui $\bar{R}$ è triangolare.

Notiamo che Q non entra esplicitamente nelle formule (49) e (50).

Se il modello (**??**) rappresenta delle misure affette da errore ε su cui si ha conoscenza probabilistica a priori, del tipo

$$\varepsilon = \sigma \mathbf{w} \quad , \qquad E\mathbf{w} = 0 \quad , \qquad \operatorname{cov}(\mathbf{w}) = I \quad ,$$

allora la soluzione del problema ai M.Q. fornisce lo stimatore di Markov per θ. In questo caso interessa calcolare la matrice di covarianza dello stimatore

$$\operatorname{Var} \hat{\theta} = \sigma^2 [S^\top S]^{-1} .$$

Usando la fattorizzazione QR si vede subito che

$$\operatorname{Var} \hat{\theta} = \sigma^2 (\bar{R}^\top \bar{R})^{-1} . \tag{51}$$

Si vede che anche in questo caso la conoscenza esplicita di Q non è richiesta. In pratica si parte dalla tabella

$$[S \mid y] \tag{52}$$

e si cerca di ridurla, attraverso trasformazioni ortogonali, alla forma

$$\left[\begin{array}{c|c} \bar{R} & y_1 \\ \hline 0 & y_2 \end{array}\right] . \tag{53}$$

Giunti a questo punto, ovviamente il grosso del lavoro è stato fatto perché rimane solo da risolvere il sistema (49) che è triangolare e per di più di sole p equazioni in p incognite. Ciò che distingue i vari algoritmi è il procedimento di ortonormalizzazione, o meglio il procedimento di riduzione della tabella (52) alla forma (53).

Si potrebbe usare Gram-Schmidt, ma esistono molti altri algoritmi con caratteristiche di stabilità molto migliori e basso carico computazionale. Per una descrizione esaustiva rimandiamo al classico testo di Lawson-Hanson.

MODELLI DINAMICI PER L'IDENTIFICAZIONE

Ricordiamo che un *processo stocastico del secondo ordine* è una classe di equivalenza di processi che hanno la stessa media e la stessa funzione di covarianza ma non necessariamente la stessa legge di probabilità. Un processo del second'ordine è completamente descrivibile assegnando la sua media e la sua funzione di covarianza. Queste sono le *statististiche del secondo ordine* del processo. Se il processo è Gaussiano esse individuano completamente tutte le distribuzioni finito-dimensionali (i.e. la legge di probabilità) del processo.

Nel seguito la stazionarietà in senso debole verrà chiamata semplicemente stazionarietà. L'ipotesi di stazionarietà debole dovrà essere rafforzata per lo studio delle proprietà asintotiche degli stimatori. (Ad es. l'ergodicità richiede stazionarietà stretta.)
Cosidereremo solo segnali a tempo discreto, denoteremo la variabile temporale adimensionale con $t \in \mathbb{Z}$, e senza perdita di generalità, assumeremo che tutti i processi in gioco abbiano media nulla. Useremo i concetti e il formalismo della teoria geometrica dei processi del second'ordine (spazi di Hilbert etc.).

I segnali osservabili sono di due tipi:

- Variabili di uscita (denotate col simbolo $\mathbf{y}$): variabili di cui si vuole ricercare la descrizione statistica.

- Variabili esogene o di ingresso (denotate col simbolo $\mathbf{u}$) : variabili la cui descrizione statistica non interessa ma che influenzano le variabili di uscita $(\mathbf{y})$ e servono a spiegarne l'andamento temporale.

Per semplicità supp. che i processi $\mathbf{y}$ e $\mathbf{u}$ siano scalari. L'estensione al caso di **ingressi multidimensionali** è facile. Di interesse in molte applicazioni. La generalizzazione al caso di *ingressi e uscite multidimensionali* presenta difficoltà. In questo caso conviene usare **modelli di stato**.

MODELLI STATISTICI LINEARI PER PROCESSI DEL SECONDO ORDINE

Faremo l'ipotesi che i dati osservati siano assimilati a tratti finiti di traiettorie (in inglese *sample paths*) di processi stocastici debolmente stazionari. Ci limiteremo a cercare *modelli che descrivono solo le statistiche del (primo e)secondo ordine di questi processi*.
È importante realizzare che i processi del second'ordine possono essere descritti da **modelli lineari**. Questi modelli lineari descrivono completamene la media (normalmente supposta nulla e) le funzioni di auto e mutua covarianza, ovvero gli spettri congiunti dei segnali in gioco. i.e. i primi due momenti della loro legge di probabilità.

Esempio: la rappresentazione di Wold per processi p.n.d. è un modello lineare. Questa rappresentazione non dipende dalla distribuzione di probabilità ma solo dalle statistiche del secondo ordine.

RETROAZIONE TRA PROCESSI

Supp. $\mathbf{y}$ ed $\mathbf{u}$ congiuntamente stazionari. Proiettando $\mathbf{y}(t)$ sullo spazio passato $H_t(\mathbf{u})$ si ottiene una decomposizione del tipo

$$\mathbf{y}(t) = F(z)\mathbf{u}(t) + \mathbf{v}(t)\,, \qquad t \in \mathbb{Z} \tag{54}$$

dove $F(z)$ è una funzione di trasferimento causale (non necessariamente razionale).

Per processi p.n.d il primo termine in (54) è il filtro di Wiener causale basato sul passato di $\mathbf{u}$ e $\mathbf{v}$ è un processo stazionario, *l'errore di modellizzazione*, scorrelato dal passato di $\mathbf{u}$

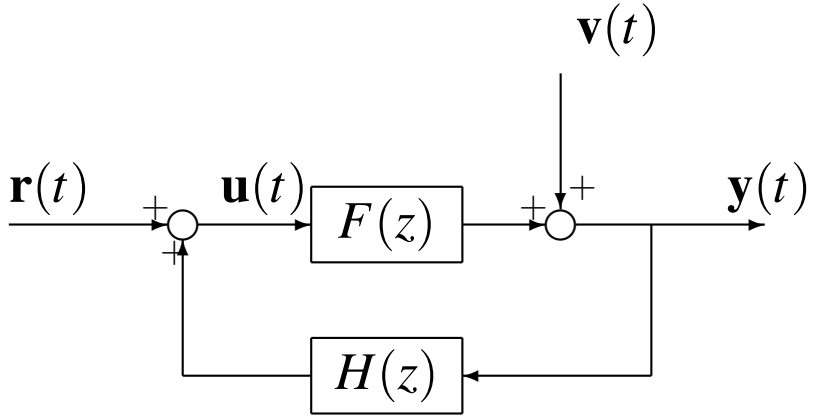
$$\mathbb{E}\,\mathbf{v}(t)\mathbf{u}(s) = 0 \qquad t \geq s\,.$$

Si può scrivere una decomposizione perfettamente analoga alla (54) anche per la variabile $\mathbf{u}$,

$$\mathbf{u}(t) = H(z)\mathbf{y}(t) + \mathbf{r}(t)\,, \qquad t \in \mathbb{Z}$$

dove $H(z)$ è una funzione di trasferimento causale (non necessariamente razionale) e il futuro di $\mathbf{r}$ è scorrelato dalla storia passata di $\mathbf{y}$.

L'interconnessione produce uno *schema a retroazione* del tipo noto nei controlli automatici.



Modello congiunto dei segnali $\mathbf{y}$ e $\mathbf{u}$.

Notare che la retroazione è un fatto intrinseco che deriva dalla correlazione mutua tra i due processi e può non corrispondere necessariamente a schemi fisici "visibili" di interconnessione a retroazione tra i due segnali.

Intuitivamente vorremmo i segnali $\mathbf{v}$ e $\mathbf{r}$ essere degli ingressi esogeni generati da meccanismi "esterni" al sistema.

Ma questi $\mathbf{v}$ e $\mathbf{r}$ risultano essere in generale correlati. Si preferisce descrivere la coppia $\mathbf{y}, \mathbf{u}$ per mezzo di schemi a retroazione in cui **si impone che gli ingressi $\mathbf{v}$ e $\mathbf{r}$ siano scorrelati**. Questi modelli si chiamano **modelli a retroazione** della coppia $\mathbf{y}, \mathbf{u}$.

Teoria dei modelli a retroazione: Esistono sempre modelli a retroazione per $(\mathbf{y}, \mathbf{u})$ in cui

◦ $\mathbf{v}$ e $\mathbf{r}$ sono scorrelati,

◦ $F(z)$ e $H(z)$ sono funzioni di trasferimento causali. Una delle due si può prendere strettamente causale (con un ritardo),

◦ Il sistema in catena chiusa con matrice di trasferimento

$$T(z) = \begin{bmatrix} \dfrac{1}{1-FH} & \dfrac{F}{1-FH} \\ \dfrac{H}{1-FH} & \dfrac{1}{1-FH} \end{bmatrix}$$

che trasforma il processo congiunto $[\mathbf{v}\ \mathbf{r}]^\top$ in $[\mathbf{y}\ \mathbf{u}]^\top$ è internamente stabile. Questo garantisce la stazionarietà congiunta di $(\mathbf{y}, \mathbf{u})$.

Si ha *assenza di reazione da* $\mathbf{y}$ *a* $\mathbf{u}$ se $H(z) \equiv 0$

Teorema 8 **Se e solo se $\mathbf{u}$ e $\mathbf{v}$ sono completamente incorrelati, ovvero vale la**

$$\mathbb{E}\,\mathbf{v}(t)\mathbf{u}(s) = 0 \qquad t\,, s \in \mathbb{Z}. \tag{55}$$

si ha *assenza di reazione da* $\mathbf{y}$ *a* $\mathbf{u}$ *, nel qual caso* $H(z) \equiv 0$.

In un modello a retroazione le funzioni di trasferimento $F(z)$ e $H(z)$ sono causali ma non necessariamente stabili. Chi deve essere stabile (più esattamente *internamente stabile*), per garantire la stazionarietà congiunta dei processi $\mathbf{y}$, $\mathbf{u}$, è il **sistema a controreazione complessivo**.

Ci restringeremo ora al caso in cui i processi $\mathbf{y}$ e $\mathbf{u}$ sono entrambi p.n.d. con matrice densità spettrale congiunta

$$S(z) = \begin{bmatrix} S_{\mathbf{y}}(z) & S_{\mathbf{yu}}(z) \\ S_{\mathbf{uy}}(z) & S_{\mathbf{u}}(z) \end{bmatrix} \tag{56}$$

che supporremo definita positiva sul cerchio unità.
NB: In pratica $\mathbf{u}$ può avere componenti puramente oscillatorie o essere un segnale puramente deterministico. Considereremo la presenza di componenti deterministiche a tempo debito.

Nelle ipotesi in cui ci siamo posti, anche $\mathbf{v}$ è un processo puramente non deterministico che ammette quindi una rappresentazione di innovazione

$$\mathbf{v}(t) = G(z)\mathbf{e}(t)\,, \qquad t \in \mathbb{Z}$$

dove $G(z)$ è una funzione di trasferimento (non necessariamente razionale) *a fase minima* che prenderemo sempre **normalizzata all'infinito**, $G(\infty) = 1$. Il processo $\mathbf{e}$ è il *processo innovazione* (non normalizzata), un processo bianco di varianza λ^2 che ha il significato di errore di predizione di un passo di $\mathbf{v}(t)$ basato sulla storia passata del processo all'istante $t-1$. Combinando si ottiene

$$\mathbf{y}(t) = F(z)\mathbf{u}(t) + G(z)\mathbf{e}(t)\,, \qquad t \in \mathbb{Z} \qquad\qquad (*)$$

Senza perdita di generalità, si può sempre imporre che $F(z)$ sia *strettamente causale*, ovvero che $F(\infty) = 0$. Questo è il modello generale che descrive la dinamica "in catena diretta" di due processi stazionari del second' ordine.

NB : nel caso di presenza di reazione il modello $(*)$ da solo non individua univocamente nemmeno la covarianza o lo spettro di $\mathbf{y}$, dato

che non specifica come $\mathbf{u}$ ed $\mathbf{e}$ sono correlati. Se c'è reazione, il modello da considerare è quello *congiunto*, comprendente anche la descrizione del canale di retroazione

$$\mathbf{u}(t) = H(z)\mathbf{y}(t) + \mathbf{r}(t)\,, \qquad t \in \mathbb{Z}$$

dove ora $\mathbf{r}$ è un processo stazionario completamente scorrelato da $\mathbf{v}$ (e quindi anche da $\mathbf{e}$). Possiamo rappresentare anche $\mathbf{r}$ mediante la sua rappresentazione di innovazione

$$\mathbf{r}(t) = K(z)\mathbf{w}(t)$$

dove $K(z)$ è una funzione di trasferimento (non necessariamente razionale) *a fase minima* normalizzata all'infinito, $K(\infty) = 1$ e il processo $\mathbf{w}$ è il *processo innovazione* (non normalizzata) di $\mathbf{r}(t)$.

Per la stazionarietà congiunta dei processi $[\mathbf{y}\,\mathbf{u}]^\top$, l' interconnessione a retroazione di figura dev'essere *internamente stabile*, ovvero la matrice di

trasferimento in catena chiusa,

$$W(z) = \begin{bmatrix} \dfrac{G(z)}{1-F(z)H(z)} & \dfrac{F(z)K(z)}{1-F(z)H(z)} \\ \dfrac{H(z)G(z)}{1-F(z)H(z)} & \dfrac{K(z)}{1-F(z)H(z)} \end{bmatrix} = T(z) \begin{bmatrix} G(z) & 0 \\ 0 & K(z) \end{bmatrix} \tag{57}$$

che trasforma il processo bianco $[\mathbf{e}\ \mathbf{w}]^\top$ in $[\mathbf{y}\ \mathbf{u}]^\top$ dovrà essere analitica in $\{|z| \geq 1\}$.
$W(z)$ è un *fattore spettrale quadrato e analitico* dello spettro congiunto $S(z)$.

I fattori spettrali non sono mai unici quindi ci sono molte rappresentazioni a retroazione della coppia $(\mathbf{y}, \mathbf{u})$. In effetti queste rappresentazioni sono in corrispondenza biunivoca con la classe dei fattori spettrali quadrati $W(z)$, dello spettro congiunto di $(\mathbf{y}, \mathbf{u})$ che sono analitici e "normalizzati a blocchi all'infinito"

$$S(z) = W(z) \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}, \qquad W(\infty) = \begin{bmatrix} 1 & 0 \\ * & 1 \end{bmatrix}$$

Supponiamo che $W(z)$ sia un fattore spettrale quadrato normalizzato a blocchi di $S(z)$. Imponendo la struttura a retroazione si ricavano facilmente le relazioni

$$F = W_{12}W_{22}^{-1}, \qquad G = W_{11} - W_{12}W_{22}^{-1}W_{21}$$
$$H = W_{21}W_{11}^{-1}, \qquad K = W_{22} - W_{21}W_{11}^{-1}W_{12}$$

che forniscono quattro funzioni razionali causali mediante le quali si possono formalmente rappresentare $\mathbf{y}(t)$ e $\mathbf{u}(t)$ tramite un modello a retroazione. In effetti $F(\infty) = 0$ e quindi $F(z)$ è strettamente causale, $H(\infty) = W_{21}(\infty)$ è finita, $G(\infty) = 1$ e $K(\infty) = 1$. Notare che se si parte da un fattore spettrale $W(z)$ analitico in $\{|z| \geq 1\}$ tutti i quattro blocchi $W_{i,j}(z)$ debbono essere analitici e quindi il modello a retroazione definito dalle quattro funzioni di trasferimento è *internamente stabile*. In genere G e K non sono però a fase minima.

Vogliamo il *modello a retroazione d'innovazione*. Come si fa a riconoscerlo?

Teorema 9 *Nel modello a retroazione d'innovazione* $F(z)$ *e* $G(z)$ *soddisfano alle seguenti condizioni*

1. *C'è almeno un ritardo in* F *e* G *è normalizzata all'infinito, i.e.* $F(\infty) = 0$ *e* $G(\infty) = 1$.

2. $G(z)^{-1}$ *e* $G(z)^{-1}F(z)$ *sono analitiche in* $\{\, |z| \geq 1 \,\}$.

Viceversa, se queste condizioni sono soddisfatte e $\mathbf{u}$ *è generato da una reazione causale del tipo* $\mathbf{u}(t) = H(z)\mathbf{y}(t) + \mathbf{r}(t)$ *in cui* $\mathbf{r}$ *è completamente scorrelato da* $\mathbf{e}$*, il processo bianco* $\mathbf{e}$ *è proprio l'innovazione di* $\mathbf{y}$*, ovvero* $\mathbf{e}(t)$ *è l' errore di predizione di un passo di* $\mathbf{y}(t)$ *basato sul passato congiunto di* $\mathbf{u}$ *e* $\mathbf{y}$ *all'istante* $t-1$.

Teorema 10 *Dato* $\mathbf{y} = F(z)\mathbf{u} + G(z)\mathbf{e}$ *(in cui potrebbe esserci reazione), che soddisfa alle condizioni dell'enunciato. Posto* $G(z) = 1 + z^{-1}G_1(z)$*, il predittore lineare a minima varianza d'errore di* $\mathbf{y}(t)$ *basato sulla storia passata congiunta* $(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})$ *è dato dalla formula*

$$\hat{\mathbf{y}}(t \mid t-1) = G(z)^{-1}F_1(z)\mathbf{u}(t-1) + G(z)^{-1}G_1(z)\mathbf{y}(t-1)\,. \tag{58}$$

Prova : Assumiamo che sia disponibile un modello che soddisfa alle condizioni del teorema 9. Dalla

$$\mathbf{e}(t) = G(z)^{-1}\left[\mathbf{y}(t) - F(z)\mathbf{u}(t)\right] \tag{59}$$

Per la causalità delle funzioni $G(z)^{-1}$ e $G(z)^{-1}F(z)$, si ha $\mathbf{e}(t) \in H(\mathbf{y}^t, \mathbf{u}^t)$.

Mostriamo che $\mathbf{e}(t)$ è scorrelato dal passato $(\mathbf{y}^{t-1}, \mathbf{u}^{t-1}) := \{\mathbf{y}(s), \mathbf{u}(s)\,; s < t\}$ dei processi $\mathbf{y}$ e $\mathbf{u}$. Dato che la matrice di trasferimento $W(z)$ deve rappresentare una trasformazione ingresso-uscita $[\mathbf{e}\ \mathbf{w}]^\top \to [\mathbf{y}\ \mathbf{u}]^\top$ stabile e causale ($W(z)$ è analitica in $\{|z| \geq 1\}$), si deve avere

$$H_t(\mathbf{y}, \mathbf{u}) \subset H_t(\mathbf{e}, \mathbf{w})\,.$$

Pertanto, dato che $\mathbf{e}$ è bianco, e i processi bianchi $\mathbf{e}$ e $\mathbf{w}$ sono completamente scorrelati, per cui $\mathbf{e}(t+1) \perp H_t(\mathbf{e}, \mathbf{w})$, si ha $\mathbf{e}(t+1) \perp H_t(\mathbf{y}, \mathbf{u})$ per ogni t.

Scriviamo allora il modello come soma ortogonale

$$\mathbf{y}(t) = \{F(z)\mathbf{u}(t) + [G(z) - 1]\,\mathbf{e}(t)\} + \mathbf{e}(t)$$

dove la somma dei primi due termini è funzione dei dati passati $(\mathbf{u}^{t-1}, \mathbf{e}^{t-1})$, dato che sia $F(z)$ che $G(z) - 1$ hanno (almeno) un ritardo. Sostituendo l'espressione

$$\mathbf{e}(t) = G(z)^{-1}\,[\mathbf{y}(t) - F(z)\mathbf{u}(t)] \qquad (\dagger)$$

si trova

$$F(z)\mathbf{u}(t) + [G(z) - 1]\,\mathbf{e}(t) = G(z)^{-1}F_1(z)\mathbf{u}(t-1) + G(z)^{-1}G_1(z)\mathbf{y}(t-1)$$

dove il secondo membro è una funzione causale dei dati $(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})$. Dato che $\mathbf{e}(t)$ è scorrelato col passato dei processi $\mathbf{y}$ e $\mathbf{u}$ all'istante $t-1$, questa funzione è proprio il predittore. □

Il teorema 9 è un corollario. Dall'espressione $(\dagger)$ si ricava il significato della causalità delle funzioni di trasferimento $G(z)^{-1}$ e $G(z)^{-1}F(z)$. Essa serve a garantire che $\mathbf{e}(t) \in H(\mathbf{y}^t, \mathbf{u}^t)$.

NB: *Nella dimostrazione abbiamo usato* $\mathbf{w}$ *e quindi la rappresentazione d' innovazione del processo* $\mathbf{r}$*. Verificare che il risultatato continua a valere anche se* $\mathbf{r}$ *non è p.n.d.*

In realtà nel modello *completo* d'innovazione il rumore bianco nel modello del segnale $\mathbf{u}$ dev'essere l'errore di predizione di un passo, ovvero $\mathbf{w} = \mathbf{e_u}$ dove

$$\mathbf{e_u}(t) = \mathbf{u}(t) - \hat{\mathbf{u}}(t \mid t-1) = \mathbf{u}(t) - \hat{\mathbb{E}}[\mathbf{u}(t) \mid \mathbf{y}^t, \mathbf{u}^{t-1}]$$

Lo studente derivi condizioni su H e K analoghe a quelle del teorema 9, affinchè questo sia il caso.

MODELLI PARAMETRICI RAZIONALI

Una coppia di processi stazionari del second'ordine p.n.d. può sempre essere descritta mediante un modello a reazione d'innovazione, costituito da una coppia di equazioni simboliche

$$\mathbf{y}(t) = F(z)\mathbf{u}(t) + G(z)\mathbf{e}(t)$$
$$\mathbf{u}(t) = H(z)\mathbf{y}(t) + K(z)\mathbf{w}(t)$$

dove F e G sono soggette alle condizioni dell'enunciato del teorema 9 ma possono essere funzioni irrazionali.

Una classe naturale di modelli dinamici che si può usare per approssimare il modello "vero" dei processi $\mathbf{y}$ e $\mathbf{u}$, è la classe dei modelli lineari che hanno questa struttura ma dove $F(z)$ **e** $G(z)$ **sono funzioni razionali di** z.

Le funzioni razionali hanno proprietà "universali" di approssimazione di classi di funzioni molto generali. Inoltre la razionalità comporta algoritmi di stima e realizzazione di dimensione finita che permettono un'organizzazione

efficiente dei calcoli. Infine, i modelli razionali possono essere raggruppati in classi omogenee in cui, una volta fissati dei *parametri di struttura (gradi dei polinomi)* le funzioni di trasferimento F e G della classe hanno la stessa "complessità" fissata; i.e. dipendono dallo stesso numero finito di parametri scalari e il problema di stima che si ha in vista riguarderà un parametro vettoriale di dimensione fissa.

Notiamo che ci si preoccupa di parametrizzare **solo il modello in catena di azione diretta** dato che per ipotesi, non siamo interessati a identificare un modello per la variabile di ingresso $\mathbf{u}$.

In generale si può pensare che i parametri da cui dipendono le due funzioni razionali F e G siano i coefficienti dei rispettivi polinomi a numeratore e a denominatore, o eventualmente, alcune loro combinazioni algebriche.

LA CLASSE DI MODELLI

Classe di modelli parametrici (di innovazione) a struttura fissata:

$$\mathbf{y}(t) = F_\theta(z)\mathbf{u}(t) + G_\theta(z)\mathbf{e}(t)\,, \qquad \theta \in \Theta \subset \mathbb{R}^p \qquad (\dagger)$$

dove $F_\theta(z)$ e $G_\theta(z)$ sono funzioni di trasferimento razionali di gradi fissati, che dipendono da un parametro vettoriale incognito θ. Non specifichiamo per il momento come $F(z)$ e $G(z)$ dipendono da θ. Postuleremo solo che questa dipendenza da θ sia regolare (continua e derivabile quante volte serve).

In realtà la classe di modelli di innovazione ha restrizioni sui parametri ammissibili (stabilità) che ignoriamo. Inoltre dipende anche dal parametro $\lambda^2 = \mathrm{var}\{\mathbf{e}(t)\}$ che non mettiamo esplicitamente in evidenza.

IDENTIFICABILITÀ IN ASSENZA DI REAZIONE

Ammettiamo per il momento che $\mathbf{u}$ abbia densità spettrale $S_{\mathbf{u}}(z)$. La densità spettrale di $\mathbf{y}$ indotta dal modello è

$$S_{\mathbf{y}}(z) = F(z)S_{\mathbf{u}}(z)F(1/z) + \lambda^2 G(z)G(1/z)\,. \qquad \mathrm{var}\mathbf{e}(t) = \lambda^2 \qquad (60)$$

$S_{\mathbf{y}}(z)$ è parametrizzata dalla densità dell'ingresso, $S_{\mathbf{u}}(z)$, e quindi dipende, dalla *condizione sperimentale*. Quando i dati sono raccolti durante il "normale funzionamento" dell'impianto, $S_{\mathbf{u}}$ è imposta dall'esterno.

Se è possibile invece progettare l'ingresso nell'esperimento di identificazione, $S_{\mathbf{u}}(z)$ può essere imposta in modo da ottimizzare il risultato dell'identificazione.

In questo caso l'ingresso può essere costituito da combinazioni di segnali "deterministici", ad es. somme di sinusoidi di frequenze diverse, che a rigore non ammettono densità spettrale di potenza. L'espressione (60) dovrebbe essere riscritta usando le *distribuzioni spettrali*, sostituendo a $S_{\mathbf{u}}(z)$ la relativa distribuzione spettrale di potenza $\hat{F}_{\mathbf{u}}(z)$. Per semplicità useremo densità con "funzioni" δ di Dirac.

La nozione di identificabiltà introdotta in statistica rigurda modelli probabilistici. Noi dobbiamo sostituire i modelli probabilstici in senso stretto (distribuzioni di probabilità) con le statistiche del secondo ordine*.
Al posto del modello probabilistico considereremo **la densità spettrale di potenza congiunta.** Il nostro modello è quindi la famiglia parametrica di spettri

$$\begin{aligned} S_{\mathbf{y}}(z;\theta) &= F_\theta(z)S_{\mathbf{u}}(z)F_\theta(1/z)+\lambda^2 G_\theta(z)G_\theta(1/z)\,, \\ S_{\mathbf{yu}}(z;\theta) &= F_\theta(z)S_{\mathbf{u}}(z) \qquad\qquad \theta\in\Theta\subset\mathbb{R}^p \end{aligned}$$

L'identificabilità del modello $(\dagger)$ deve corrispondere a una parametrizzazone non ridondante dello spettro congiunto.
Notiamo che, dato che non interessa modellare $\mathbf{u}$, lo spettro $S_{\mathbf{u}}(z)$ non è parametrizzato e quindi gli elementi dello spettro congiunto che dipendono da θ sono solo $S_{\mathbf{y}}$ e lo spettro incrociato $S_{\mathbf{yu}}$.

*Dovremmo allora, a rigore parlare di indistinguibilità e identificabilità del second'ordine, o in senso debole.

Definizione 13 *Si assuma assenza di reazione da* $\mathbf{y}$ *a* $\mathbf{u}$*. Il modello* $(\dagger)$ *è identificabile (globalmente) nella condizione sperimentale descritta dallo spettro di ingresso* $S_{\mathbf{u}}$*, se la mappa* $\theta \mapsto [S_{\mathbf{y}}(\cdot;\theta), S_{\mathbf{yu}}(\cdot;\theta)]$ *è* **iniettiva** *in* Θ*, ovvero se l'uguaglianza*

$$[S_{\mathbf{y}}(z;\theta_1), S_{\mathbf{yu}}(z;\theta_1)] = [S_{\mathbf{y}}(z;\theta_2), S_{\mathbf{yu}}(z;\theta_2)], \qquad \forall z \in \mathbb{C}$$

implica $\theta_1 = \theta_2$*. Se si ha iniettività locale in un intorno di* θ_0*, si parla di* identificabilità locale *in* θ_0*.*

Dalla definizione si trae il seguente criterio.

Proposizione 13 *Nel modello* $(\dagger)$*, senza reazione, due vettori di parametri* θ_1 *e* θ_2 *sono indistinguibili nella condizione sperimentale descritta da* $S_{\mathbf{u}}$ *(o dalla distribuzione spettrale* $d\hat{F}_{\mathbf{u}}$*), se e solo se, per ogni* $z = e^{j\omega}$

$$\left[F_{\theta_1}(z) - F_{\theta_2}(z)\right] S_{\mathbf{u}}(z) = 0, \tag{61}$$

$$G_{\theta_1}(z) - G_{\theta_2}(z) = 0 \tag{62}$$

Equivalentemente, il modello $(\dagger)$ *è globalmente identificabile se l'unica soluzione delle equazioni* (61), (62) *è* $\theta_1 = \theta_2$*.*

Prova: $\Rightarrow$ Dato che

$$F_\theta(z)S_{\mathbf{u}}(z) = S_{\mathbf{yu}}(z;\theta)$$

evidentemente la (61) implica l'uguaglianza degli spettri $S_{\mathbf{yu}}(z;\theta_1)$ e $S_{\mathbf{yu}}(z;\theta_2)$. Notiamo poi che

$$F_\theta(z)S_{\mathbf{u}}(z)F_\theta(1/z) = S_{\mathbf{yu}}(z;\theta)S_{\mathbf{u}}^{-1}(z)S_{\mathbf{uy}}(z;\theta)\,, \qquad (\ddagger)$$

per cui $S_{\mathbf{yu}}(z;\theta_1) \equiv S_{\mathbf{yu}}(z;\theta_2)$ implica $F_{\theta_1}(z)S_{\mathbf{u}}(z)F_{\theta_1}(1/z) \equiv F_{\theta_2}(z)S_{\mathbf{u}}(z)F_{\theta_2}(1/z)$. Quindi, se vale la (62) si ha $S_{\mathbf{y}}(z;\theta_1) = S_{\mathbf{y}}(z;\theta_2)$.

$\Leftarrow$ Viceversa, l'uguaglianza degli spettri incrociati implica (61). Dalla $(\ddagger)$ e

$$S_{\mathbf{y}}(z;\theta) - F_\theta(z)S_{\mathbf{u}}(z)F_\theta(1/z) = \lambda^2 G_\theta(z)G_\theta(1/z)\,.$$

le uguaglianze degli spettri incrociati e di uscita implicano $G_{\theta_1}(z)G_{\theta_1}(1/z) = G_{\theta_2}(z)G_{\theta_2}(1/z)$. Dato che si conviene di prendere sempre $G_\theta(z)$ a fase minima e normalizzata (e quindi univocamente determinata dal prodotto $G_\theta(z)G_\theta(1/z)$) segue l'asserto. $\square$

SUFFICIENTE (O PERSISTENTE) ECCITAZIONE

Se la condizione

$$\left[F_{\theta_1}(z) - F_{\theta_2}(z)\right] S_{\mathbf{u}}(z) = 0$$

implica $F_{\theta_1}(\cdot) \equiv F_{\theta_2}(\cdot)$ per tutte le funzioni di trasferimento di una certa classe $\mathcal{F}$, si dice che **l'ingresso $\mathbf{u}$ è sufficientemente (o persistentemente) eccitante per la classe $\mathcal{F}$**.

Nel nostro caso $\mathcal{F}$ è una classe parametrica di funzioni razionali $\mathcal{F} = \{F_\theta(\cdot)\,;\theta \in \Theta\}$.

Nel caso speciale in cui l'ingresso $\mathbf{u}$ **è rumore bianco**, la densità spettrale $S_{\mathbf{u}}$ è una costante positiva e quindi **un ingresso bianco è sufficientemente eccitante per una qualunque classe di funzioni di trasferimento**. Dalle (61), (62) si vede immediatamente che in questo caso l'identificabiltà a priori è necessaria e sufficiente per l'identificabiltà.

NB: un modello identificabile a priori potrebbe benissimo non essere identificabile per qualche condizione sperimentale. Un caso estremo si presenta quando l' ingresso è una funzione costante (al limite nulla) per cui $S_{\mathbf{u}}(z)$ è zero quasi ovunque.
Se viceversa il processo di ingresso ha una componente p.n.d. non nulla (ricordare la decomposizione di Wold !), il suo spettro ha una componente assolutamente continua che è positiva quasi ovunque.

Proposizione 14 *Se il processo di ingresso ha una componente p.n.d. non nulla, è sufficientemente eccitante per qualunque $\mathcal{F}$ e l'identificabilità è equivalente all'identificabilità a priori.*

IDENTIFICABILITÀ A PRIORI

Una nozione di identificabilità che talvolta in letteratura viene confusa con quella precedente è l'identificabilità *a priori*.

Definizione 14 *Il modello* $(\dagger)$ *è identificabile* **a priori** *(globalmente) se la mappa* $\theta \mapsto [F_\theta(\cdot),\, G_\theta(\cdot)]$ *è iniettiva in* Θ*, ovvero*

$$[F_{\theta_1}(z),\, G_{\theta_1}(z)] = [F_{\theta_2}(z),\, G_{\theta_2}(z)] \quad \forall z \in \mathbb{C} \Rightarrow \theta_1 = \theta_2\,.$$

Se si ha iniettività locale in un intorno di θ_0*, si parla di* identificabilità a priori locale *in* θ_0.

Come si vede, la nozione di identificabiltà a priori non ha nulla di "probabilistico" e riguarda solo il modo in cui sono parametrizzate le funzioni di trasferimento $F_\theta(z),\, G_\theta(z)$. La verifica dell'identificabiltà a priori è quindi un fatto puramente algebrico.

Per apprezzare la diversità dei due concetti basta analizzare la mappa che descrive la dipendenza dello spettro congiunto dal parametro θ. Questa mappa è composta di due componenti:

$$\theta \mapsto [F_{\theta}(\cdot),\, G_{\theta}(\cdot)] \mapsto [S_{\mathbf{y}}(\cdot;\theta)\,, S_{\mathbf{yu}}(\cdot;\theta)]$$

Dato che si tratta di una mappa ottenuta per composizione delle due applicazioni $\theta \mapsto [F_{\theta}(\cdot),\, G_{\theta}(\cdot)]$ e $[F_{\theta}(\cdot),\, G_{\theta}(\cdot)] \mapsto [S_{\mathbf{y}}(\cdot;\theta)\,, S_{\mathbf{yu}}(\cdot;\theta)]$ (quest'ultima dipendente dallo spettro dell'ingresso), per l'identificabilità si deve richiedere l'iniettività di entrambi; il che si può riassumere nel modo seguente.

Proposizione 15 *L'identificabilità (in assenza di reazione) è equivalente all'identificabilità a priori e alla sufficiente eccitazione dell'ingresso.*

È ovvio dalla definizione 13, che l'identificabilità a priori è solo condizione *necessaria* per l'identificabiltà. In ogni caso, se il modello $(\dagger)$ non fosse identificabile a priori, non sarebbe possibile distinguere i parametri dello spettro congiunto $[S_{\mathbf{y}}(\cdot;\theta)\,, S_{\mathbf{yu}}(\cdot;\theta)]$, per nessuna condizione sperimentale.

Nel caso in cui il segnale di ingresso è p.d., o addirittura deterministico, l'identificabilità dipende dalla classe parametrica di funzioni di trasferimento $F_{\boldsymbol{\theta}}(z)$ che costituiscono il modello. Bisogna richiedere, oltre all'identificabilità a priori, che il segnale di ingresso soddisfi a condizioni di **persistente eccitazione**. Queste condizioni verranno discusse nelle prossime slides.

IDENTIFICABILITÀ A PRIORI DI MODELLI RAZIONALI

Un' analisi approfondita del concetto di identificabiltà a priori porta a qualche sorpresa. A meno che il modello non sia parametrizzato linearmente (sia cioè una funzione lineare di θ) si scopre che *l' identificabiltà a priori è una proprietà essenzialmente locale*. Consideriamo ad esempio una funzione di trasferimento scalare razionale del prim'ordine

$$F_\theta(z) = \frac{1+cz^{-1}}{1+az^{-1}}; \qquad \theta = [a\ c]^\top \tag{63}$$

dove per semplicità non considereremo i vincoli di stabilità sui parametri a e c per cui penseremo θ variabile su tutto $\mathbf{R}^2$. Ora è ben evidente che

$$F_{\theta_1}(z) = F_{\theta_2}(z) \quad \forall z \ \Rightarrow\ \theta_1 = \theta_2$$

eccezion fatta per quei θ che appartengono all'insieme:

$$\Theta_0 := \{\theta\,;\, a = c\}$$

dato che se $\theta \in \Theta_0$ il numeratore e il denominatore si cancellano e si ha $F_\theta(z) = 1$ identicamente per cui, qualunque siano $a = c \in \Theta_0$ la funzione $F_\theta(z)$ è $= 1$!.
La classe parametrica di funzioni di trasferimento (63) è identificabile localmente in tutti i punti dell'insieme $\mathbb{R}^2 - \Theta_0$. In sostanza, è identificabile localmente in tutti i punti dell'insieme dei valori del parametro θ per cui *non si hanno cancellazioni tra numeratore e denominatore*.

Questo esempio è rappresentativo del caso generale.

Proposizione 16 *Una famiglia di funzioni razionali parametrizzata attraverso i coefficienti dei polinomi a numeratore e a denominatore, almeno uno dei quali è supposto monico, è localmente identificabile a priori in tutti i valori del parametro che non corrispondono a fattori comuni (e quindi a cancellazioni) tra i due polinomi.*

Dato che l'insieme dei valori del parametro per cui si hanno cancellazioni è sempre un sottoinsieme di misura nulla di Θ, si parla di identificabilità *quasi ovunque* o meglio di identificabilità *generica*.

PERSISTENTE ECCITAZIONE

La nozione di segnale persitentemente eccitante riguarda segnali "deterministici", per i quali non è necesssariamente data una descrizione statistica.

Definizione 15 *Un segnale deterministico* $u = \{u(t)\,; t \in \mathbb{Z}\}$ *è detto* stazionario del secondo ordine *se il limite*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} u(t+\tau)u(t) := r(\tau)$$

esiste per ogni $\tau \geq 0$.

Un segnale stazionario del secondo ordine, u*, è* persitentemente eccitante di ordine (almeno) n *se la matrice (di Toeplitz)*

$$\mathbf{R}_n := \begin{bmatrix} r(0) & r(1) & \dots & & r(n-1) \\ r(1) & r(0) & r(1) & \dots & r(n-2) \\ \vdots & & & \vdots & \\ r(n-1) & r(n-2) & \dots & & r(0) \end{bmatrix}, \qquad \textit{è definita positiva.}$$

PROPRIETÀ DI $r(\tau)$

Notiamo che $\mathbf{R}_n$ è sempre almeno semidefinita perchè, per un arbitrario polinomio $p(z^{-1}) := \sum_{k=0}^{n-1} p_k z^{-k}$ nell' operatore di ritardo z^{-1}, si ha

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} [p(z^{-1})u(t)]^2 = \sum_{k,j=0}^{n} p_k\, r(k-j)\, p_j = p^\top \mathbf{R}_n\, p \geq 0$$

dove abbiamo denotato col simbolo $p := [p_0\, p_1 \ldots p_{n-1}]^\top$ il vettore dei coefficienti di $p(z^{-1})$.

Come già notato da Norbert Wiener (nel 1930!), la funzione $\tau \to r(\tau)\,;\, \tau \in \mathbb{Z}$ è quindi una funzione *di tipo positivo* che si può pertanto assimilare alla funzione di correlazione di un processo stazionario. Ad essa si applica quindi il teorema di Herglotz che stabilisce l'esistenza di una funzione monotona non decrescente sull'intervallo $[-\pi, \pi]$, $F_u(e^{j\omega})$, la **distribuzione spettrale di potenza** di u, per cui

$$r(\tau) = \int_{-\pi}^{\pi} e^{j\omega\tau}\, dF_u(e^{j\omega})\,.$$

Per i segnali deterministici, stazionari del secondo ordine, possiamo quindi parlare di *spettro di potenza*. Convenzionalmente si usa parlare di "densità spettrale" anche se raramente la F_u è assolutamente continua. In generale la densità spettrale di un segnale stazionario del secondo ordine contiene impulsi o *righe spettrali* come si usa dire comunemente.

Teorema 11 *Un segnale stazionario del second'ordine u è persistentemente eccitante di ordine esattamente n se e solo se gli unici punti di crescita della sua distribuzione spettrale $F_u(e^{j\omega})$ sono n salti, ovvero la sua densità spettrale consiste esattamente di n righe spettrali alle frequenze $\{\omega_1, \omega_2, \ldots, \omega_n\}^*$, nell'intervallo $(-\pi, \pi)$.*

Prova: Siano

$$p(z^{-1}) := \sum_{k=0}^{n-1} p_k z^{-k} \qquad q(z^{-1}) := \sum_{k=0}^{n} q_k z^{-k}$$

*Dato che per segnali reali $r(\tau)$ è reale, si tratta in realtà di coppie di frequenze opposte $\pm\omega_k$.

due polinomi di grado effettivo $n-1$ ed n e si denotino con $p \in \mathbb{R}^n$ e $q \in \mathbb{R}^{n+1}$ i rispettivi vettori dei coefficienti. Notiamo che se lo spettro di u è come descritto nell'enunciato,

$$p^\top \mathbf{R}_n \, p = \int_{-\pi}^{\pi} |p(e^{j\omega})|^2 \, dF_u(e^{j\omega}) = \sum_{k=1}^{n} \sigma_k^2 |p(e^{j\omega_k})|^2 \qquad \sigma_k^2 > 0 \,,$$

e se il primo membro di questa espressione fosse uguale a zero il polinomio $p(z^{-1})$ dovrebbe allora soddisfare alle n condizioni $|p(e^{j\omega_k})|^2 = 0; k = 1,2,\ldots,n$, che sono equivalenti alle

$$p(e^{j\omega_k}) = 0 \,, \qquad k = 1,2,\ldots,n \qquad (*)$$

condizioni che implicano $p(z^{-1}) \equiv 0$, dato che $p(z^{-1})$ ha grado $n-1$. Notiamo che se si prende un polinomio $q(z^{-1})$ di grado n, si può invece avere $q^\top \mathbf{R}_{n+1} \, q = 0$ perchè le n condizioni $(*)$ possono essere soddisfatte da un polinomio di grado n non identicamente nullo. Quindi l'ordine di persistente eccitazione è esattamente n.

Per mostrare che la condizione è anche necessaria dimostriamo che un segnale persistentemente eccitante di ordine esattamente n è in realtà la

somma di n oscillazioni armoniche di frequanze $\{\omega_1, \omega_2, \ldots, \omega_n\}$, nell'intervallo $(-\pi, \pi)$ tra loro diverse. Per ipotesi esiste un polinomio di grado effettivo n con vettore dei coefficienti non nullo $q \in \mathbb{R}^n$, tale che

$$0 = q^\top \mathbf{R}_{n+1}\, q = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} [q(z^{-1})u(t)]^2$$

Si mostra che il limite a secondo membro è in realtà il quadrato della norma del segnale $t \to q(z^{-1})u(t)$; i.e. si può scrivere $\|q(z^{-1})u(\cdot)\|^2$, in un opportuno spazio di Hilbert di segnali stazionari del second'ordine. Il fatto che questa norma è zero, implica che u soddisfa l'equazione alle differenze

$$q(z^{-1})u(t) = 0 \qquad t \in \mathbb{Z}\,.$$

Questa equazione alle differenze può solo avere soluzioni limitate, giacchè l'esistenza di un modo esponenziale (crescente o decrescente) nel segnale u renderebbe infinito qualcuno dei limiti (**??**). Ne segue che tutti gli zerio di $q(z^{-1})$ debbono trovarsi sul cerchio unitario e debbono avere molteplicità uno. Quindi il segnale u è la somma di n oscillazioni armoniche di frequanze $\{\omega_1, \omega_2, \ldots, \omega_n\}$. Il resto della dimostrazione scende dall' esempio che segue. □

SEGNALI QUASI PERIODICI

Proposizione 17 *La somma di N segnali sinusoidali di frequenza diversa*

$$u(t) = \sum_{k=1}^{N} A_k \sin(\omega_k t + \phi_k) \qquad \omega_k \neq \omega_j \quad , \qquad 0, \omega_k, 2\pi$$

è un segnale stazionario del secondo ordine, la cui correlazione vale

$$r(\tau) = \sum_{k=1}^{N} \frac{A_k^2}{2} \cos \omega_k \tau$$

e il suo spettro consiste di $2N$ righe (funzioni δ) supportatate nei punti $\{\pm\omega_k\}$. Il segnale è pertanto persistentemente eccitante (P.E.) di ordine (esattamente) $2N$.

Si veda [Söderström-Stoica p. 98-109]

SEGNALI PERIODICI

Scende dalla teoria della trasformata discreta di Fourier (DFT), che un segnale a tempo discreto, periodico di periodo N, è la somma di N componenti sinusoidali di frequenza $\omega_1 = \frac{2\pi}{N}, \omega_2 = 2\frac{2\pi}{N}, \ldots, \omega_N = 2\pi$. Quindi ogni segnale periodico di periodo N è P.E. di ordine $2N$.

Esempio 1 (Segnali PRBS) Un segnale PRBS (*Pseudo Random Binary Sequence*) è un particolare segnale periodico che approssima il rumore bianco ed è spesso usato nelle simulazioni. *Si veda [Söderström-Stoica p. 124]*

SISTEMI CON INGRESSI P.E.

L'uscita di un sistema lineare stabile con ingresso P. E. di ordine esattamente n ha uno spettro che contiene al più n righe a meno che qualche frequenza propria dell'ingresso non coincida con degli zeri di $G(z)$ situati sulla circonferenza unità.

Proposizione 18 *Si supponga che il processo* $\mathbf{y}$ *sia descritto dal modello* $(\dagger)$ *senza reazione, con ingresso* $\mathbf{u}$ *persitentemente eccitante di ordine esattamente* n*. Se* $S_{\mathbf{v}}(z)$ *non si annulla in qualcuna delle* n *frequenze proprie di* $\mathbf{u}$*, la matrice densità spettrale congiunta* (56) *è allora definita positiva nelle* n *frequenze proprie dell'ingresso.*

Prova: Infatti si ha

$$S(z) = \begin{bmatrix} S_{\mathbf{y}}(z) & S_{\mathbf{yu}}(z) \\ S_{\mathbf{uy}}(z) & S_{\mathbf{u}}(z) \end{bmatrix} = \begin{bmatrix} F(z)S_{\mathbf{u}}(z)F(1/z)+\lambda^2 G(z)G(1/z) & F(z)S_{\mathbf{u}}(z) \\ S_{\mathbf{u}}(z)F(1/z) & S_{\mathbf{u}}(z) \end{bmatrix}$$

$$= \begin{bmatrix} 1 & F(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \lambda^2 G(z)G(1/z) & 0 \\ 0 & S_{\mathbf{u}}(z) \end{bmatrix} \begin{bmatrix} 1 & 0 \\ F(1/z) & 1 \end{bmatrix}$$

e il determinante di $S(z)$ è uguale a $\lambda^2 G(z)G(1/z)S_{\mathbf{u}}(z)$. □

ALCUNE CLASSI DI MODELLI RAZIONALI

I modelli razionali del tipo $(\dagger)$ (detti di "Box-Jenkins"), possono essere parametrizzati mediante i coefficienti dei relatvi polinomi a numeratore e denominatore,

$$F_\theta(z) = \frac{B(z^{-1})}{A(z^{-1})} \qquad G_\theta(z) = \frac{C(z^{-1})}{D(z^{-1})}.$$

A, CD **sono monici e** B **ha il coefficiente di grado zero** (b_0) **uguale a zero**. I parametri "liberi" sono

- $n = \mathrm{Deg}(A)$ coefficienti del polinomio $A(z^{-1}) = 1 + \sum_{k=1}^{n} a_k z^{-k}$,
- $m = \mathrm{Deg}(B)$ coefficienti del polinomio $B(z^{-1}) = \sum_{k=1}^{m} b_k z^{-k}$,
- $q = \mathrm{Deg}(C)$ coefficienti del polinomio $C(z^{-1}) = 1 + \sum_{k=1}^{q} c_k z^{-k}$,
- $r = \mathrm{Deg}(D)$ coefficienti del polinomio $D(z^{-1}) = 1 + \sum_{k=1}^{r} d_k z^{-k}$

con un totale di $p = n + m + q + r$, più la varianza dell'innovazione λ^2 che conviene considerare separatamente.

In realtà il vincolo che il modello $(\dagger)$ sia d'innovazione si dovrebbe imporre vincolando i coefficienti di C, A e D.
Se non c'è reazione, C A e D dovrebbe essere vincolati a essere strettamente stabili. Questi vincoli definiscono in teoria l'insieme dei parametri ammissibili Θ.
Purtroppo la struttura geometrica degli insiemi che definiscono i coefficienti ammissibili è estremamante complicata e di fatto non è nemmeno nota, se il grado del polinomio è maggiore di quattro; per cui in pratica la stabilità viene imposta a posteriori.

In pratica si considerano spesso delle sottoclassi particolari di modelli razionali. La più diffusa è la famiglia dei **modelli ARMAX**

$$\mathcal{A}(z^{-1})\mathbf{y}(t) = \mathcal{B}(z^{-1})\mathbf{u}(t) + \mathcal{C}(z^{-1})\mathbf{e}(t)\,, \qquad \text{(ARMAX)}$$

in cui si prendono $\mathcal{A}$ e $\mathcal{C}$ monici e $\mathcal{B}$ con il coefficiente di grado zero uguale a zero. Questi modelli possono essere parametrizzati mediante i coefficienti dei tre polinomi $\mathcal{A}$, $\mathcal{B}$ e $\mathcal{C}$.

Notiamo che il modello Box-Jenkins equivalente a (ARMAX) ha

$$A(z^{-1}) = \mathcal{A}(z^{-1}), \quad B(z^{-1}) = \mathcal{B}(z^{-1}), \quad C(z^{-1}) = \mathcal{C}(z^{-1}), \quad D(z^{-1}) = \mathcal{A}(z^{-1})$$

e quindi usando ARMAX si descrive $\mathbf{v}$ con una funzione di trasferimento G che ha *gli stessi poli di* $F(z)$. Se non vi sono motivi "fisici" per pensare che questo possa essere veramente il caso, *questa struttura porta in pratica a identificare modelli di ordine più alto del dovuto.*

Ogni ARMAX equivalente al Box-Jenkins deve avere la struttura:

$$\mathcal{A}(z^{-1}) = A(z^{-1})D(z^{-1}) \quad \mathcal{B}(z^{-1}) = B(z^{-1})D(z^{-1}) \quad \mathcal{C}(z^{-1}) = C(z^{-1})A(z^{-1})$$

in cui però i parametri dei polinomi $\mathcal{A}$, $\mathcal{B}$ e $\mathcal{C}$ (in totale $n+r+m+r+q+n = 2n+m+2r+q$) *non sono più liberi di variare in modo indipendente ma debbono essere vincolati a soddisfare le relazioni algebriche dei prodotti scritte sopra.*

In pratica, nei procedimenti di stima questi vincoli algebrici sono impossibili da imporre e quindi le cancellazioni tra $\mathcal{A}$ e $\mathcal{B}$, $\mathcal{A}$ e $\mathcal{C}$ e $\mathcal{B}$ e $\mathcal{C}$ che dovrebbero ristabilire gli ordini corretti nel modello Box-Jenkins equivalente non avvengono mai. Di conseguenza *l'uso di modelli ARMAX porta in generale a sovrastimare gli ordini dei polinomi* e a stime delle funzioni di trasferimento F e G in cui ci sono delle "quasi cancellazioni" polo-zero.

Una sottoclasse estremamente popolare dei modelli ARMAX è quella dei modelli ARX, che sono del tipo

$$\mathcal{A}(z^{-1})\mathbf{y}(t) = \mathcal{B}(z^{-1})\mathbf{u}(t) + \mathbf{e}(t) \tag{64}$$

in cui si prendono $\mathcal{A}$ monico e $\mathcal{B}$ con il coefficiente di grado zero uguale a zero. Il polinomio $\mathcal{C}$ è preso uguale a 1. Anche questi modelli possono essere parametrizzati mediante i coefficienti di $\mathcal{A}$ e $\mathcal{B}$.

Notiamo che il modello Box-Jenkins equivalente a (64) ha

$$A(z^{-1}) = \mathcal{A}(z^{-1}) \quad B(z^{-1}) = \mathcal{B}(z^{-1}) \quad C(z^{-1}) = 1 \quad D(z^{-1}) = \mathcal{A}(z^{-1})$$

e quindi usando un modello ARX si descrive l'errore di modellizzazione $\mathbf{v}$ con un modello *puramente autoregressivo che ha gli stessi poli di* $F(z)$. Se non vi sono motivi "fisici" per pensare che questo possa essere veramente il caso, l'uso di questa struttura porta in pratica a stimare modelli di ordine molto più alto del dovuto e (come vedremo) può portare a stime distorte.

Problema 2 *Determinare l'ordine minimo di persistente eccitazione del segnale di ingresso u nel modello ARX senza reazione*

$$\left(1+a_1z^{-1}+a_2z^{-2}+\ldots+a_nz^{-n}\right)\mathbf{y}(t)=\left(b_1z^{-1}+b_2z^{-2}+b_3z^{-3}\right)u(t)+\mathbf{e}(t)$$

per avere identificabiltà.

Sia $\theta=[a_1\ldots\ a_n\ b_1\ b_2\ b_3]^\top$. Il modello è globalmente identificabile se e solo se l'unica soluzione delle equazioni

$$\left[\frac{B_{\theta_1}(z^{-1})}{A_{\theta_1}(z^{-1})}-\frac{B_{\theta_2}(z^{-1})}{A_{\theta_2}(z^{-1})}\right]S_{\mathbf{u}}(z)=0$$

$$\frac{1}{A_{\theta_1}(z^{-1})}-\frac{1}{A_{\theta_2}(z^{-1})}=0$$

è $\theta_1=\theta_2$. Equivalente alle due equazioni

$$\begin{aligned}\left[B_{\theta_1}(e^{j\omega})-B_{\theta_2}(e^{j\omega})\right]S_u(e^{j\omega}) &= 0,\\ A_{\theta_1}(e^{j\omega})-A_{\theta_2}(e^{j\omega}) &= 0\end{aligned}$$

dove S_u è lo spettro di u (eventualmente espresso mediante funzioni δ).

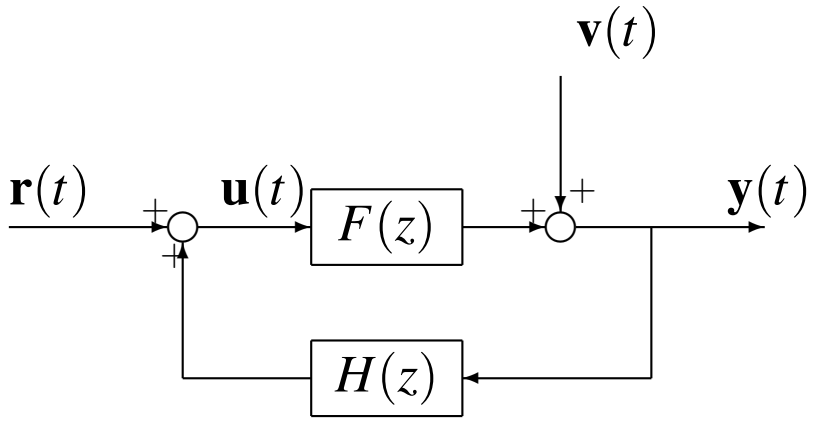
Quindi se $A_\theta(z^{-1})$ è monico i parametri a_k sono senz'altro tutti identificabili qualunque sia n e qualunque sia lo spettro di $\mathbf{u}$. Mentre

$$\left[B_{\theta_1}(e^{j\omega}) - B_{\theta_2}(e^{j\omega})\right] S_u(e^{j\omega}) = 0$$

ha come unica soluzione $\theta_1 = \theta_2$ se e solo se i polinomi $B_{\theta_1}(z^{-1})$ e $B_{\theta_2}(z^{-1})$ hanno valori uguali per almeno m valori diversi di frequenza.
Quindi nel nostro esempio $m = 3$ e la condizione di identificabilità impone che lo **spettro dell'ingresso abbia almeno tre righe a tre frequenze distinte**.
In generale l'ordine minimo di persistente eccitazione per l'identificabilità di un modello ARX è m.

IDENTIFICABILITÀ CON REAZIONE

Come abbiamo visto un processo congiunto $[\mathbf{y}, \mathbf{u}]'$ stazionario e p.n.d. si può sempre descrivere con un modello d'innovazione a retroazione del tipo



dove $\mathbf{v}, \mathbf{r}$ (necessariamente p.n.d.) sono scorrelati, di densità spettrali rispettive $S_{\mathbf{v}}, S_{\mathbf{r}}$.
Il modello è completamente descritto dalle quattro funzioni (razionali)

$$F(z), H(z), S_{\mathbf{v}}(z), S_{\mathbf{r}}(z)$$

$S_{\mathbf{v}}, S_{\mathbf{r}}$ potrebbero essere descritte dai rispettivi fattori spettrali canonici, ma in un modello d'innovazione a retroazione $G(z)$ $K(z)$ sono diverse e possono avere poli instabili.

Naturalmente il modello a retroazione potrebbe descrivere una situazione reale in cui i dati sono effettive osservazioni ingresso-uscita di un sistema di controllo con una retroazione lineare, ma questo non è necessario e la retroazione potrebbe essere solo "intrinseca".

Sia $S(z)$ la matrice densità spettrale congiunta del processo $[\mathbf{y}\,,\,\mathbf{u}]'$. Vogliamo capire quando **ad un dato spettro (congiunto) corrisponde uno e un solo modello a retroazione.**

Questa unicità è un prerequisito fondamentale per l'identificabilità del modello a retroazione, *indipendente dalla parametrizzazione*.

Dalla matrice di trasferimento del sistema in catena chiusa

$$\begin{bmatrix} \mathbf{y}(t) \\ \mathbf{u}(t) \end{bmatrix} = \begin{bmatrix} \dfrac{1}{1-FH} & \dfrac{F}{1-FH} \\ \dfrac{H}{1-FH} & \dfrac{1}{1-FH} \end{bmatrix} \begin{bmatrix} \mathbf{v}(t) \\ \mathbf{r}(t) \end{bmatrix}$$

si calcola lo spettro congiunto

$$\begin{aligned} S_{\mathbf{y}} &= \frac{1}{|1-HF|^2} S_{\mathbf{v}} + \frac{FF^*}{|1-HF|^2} S_{\mathbf{r}} \\ S_{\mathbf{yu}} &= \frac{H^*}{|1-HF|^2} S_{\mathbf{v}} + \frac{F}{|1-HF|^2} S_{\mathbf{r}} \\ S_{\mathbf{u}} &= \frac{HH^*}{|1-HF|^2} S_{\mathbf{v}} + \frac{1}{|1-HF|^2} S_{\mathbf{r}} \end{aligned}$$

Le funzioni di trasferimento dipendono da $e^{j\omega}$ e l'asterisco indica il complesso coniugato.

Problema: Lo spettro congiunto S determina univocamente la funzione di trasferimento in catena aperta $F(z)$?

In assenza di reazione si avrebbe semplicemente $S_{\mathbf{yu}}(e^{j\omega}) = F(e^{j\omega})S_{\mathbf{u}}(e^{j\omega})$ per cui $F(e^{j\omega})$ si ricava immediatamente dalla

$$F(e^{j\omega}) = \frac{S_{\mathbf{yu}}(e^{j\omega})}{S_{\mathbf{u}}(e^{j\omega})}.$$

Questa è la formula usata da molti analizzatori di spettro in commercio. In presenza di reazione invece si ha

$$\hat{F}(e^{j\omega}) := \frac{S_{\mathbf{yu}}(e^{j\omega})}{S_{\mathbf{u}}(e^{j\omega})} = \frac{F(e^{j\omega})\,S_{\mathbf{r}}(e^{j\omega}) + H^*(e^{j\omega})\,S_{\mathbf{v}}(e^{j\omega})}{H(e^{j\omega})H^*(e^{j\omega})\,S_{\mathbf{v}}(e^{j\omega}) + S_{\mathbf{r}}(e^{j\omega})} \tag{65}$$

si vede che $\hat{F} = F$ solo nel caso in cui $H \equiv 0$ (assenza di reazione) o $S_{\mathbf{v}} \equiv 0$

nel caso in cui $S_{\mathbf{r}} \equiv 0$ si ha addirittura $\hat{F} = 1/H$.

L'assenza di eccitazione dall'ingresso esterno $\mathbf{r}$ (i.e. $S_{\mathbf{r}} \equiv 0$) comporta la *non identificabilità di* F.

Lemma 2 *Assumiamo che vi sia reazione (e quindi che $H(z) \neq 0$). Lo spettro congiunto $S(e^{j\omega})$ è singolare quasi ovunque se e solo se almeno uno dei due spettri $S_{\mathbf{v}}, S_{\mathbf{r}}$ è uguale a zero.*

Prova: Il processo congiunto si può sempre esprimere come uscita del sistema lineare di funzione di trasferimento $W(z)$ con in ingresso i due processi di innovazione di $\mathbf{r}$ e $\mathbf{v}$. Esiste un fattore spettrale quadrato a fase minima $W(z)$ di $S(z)$ normalizzato all'identità all'infinito tale che

$$S(z) = W(z) \begin{bmatrix} \lambda_1^2 & 0 \\ 0 & \lambda_2^2 \end{bmatrix} W(1/z)^\top, \qquad W(\infty) = I\,.$$

Dato che $W(\infty) = I$, la matrice razionale $W(z)$ dev'essere non singolare in un intorno di ∞ e quindi quasi dappertutto sul piano complesso. Dalla fattorizzazione si vede quindi che (nell'ipotesi di presenza di reazione, $H(z) \neq 0$), $S(z)$ è singolare quasi ovunque se e solo se una (o entrambi) delle due varianze λ_1^2, λ_2^2 (e quindi uno o entrambi dei due processi $\mathbf{r}, \mathbf{v}$) è uguale a zero. □

Proposizione 19 *Se e solo se lo spettro congiunto* (56) *è non singolare quasi ovunque, le funzioni* $(F, H, S_{\mathbf{v}}, S_{\mathbf{r}})$ *sono individuate univocamente dallo spettro* $\mathcal{S}$*; in altri termini, la mappa* $\mathfrak{R}$ *è iniettiva.*

Prova: Se lo spettro è non singolare λ_1^2 e λ_2^2 sono diverse da zero. Siano allora $\mathbf{e}_1, \mathbf{e}_2$ i processi di innovazione di $\mathbf{y}$ e $\mathbf{u}$. Dato che

$$\begin{aligned} \mathbf{e}_1(t) &= G(z)^{-1}\left[\mathbf{y}(t) - F(z)\mathbf{u}(t)\right] \\ \mathbf{e}_2(t) &= K(z)^{-1}\left[\mathbf{u}(t) - H(z)\mathbf{y}(t)\right] \end{aligned}$$

l'inverso del fattore a fase minima deve avere la struttura

$$W(z)^{-1} = \begin{bmatrix} G(z)^{-1} & G(z)^{-1}F(z) \\ K(z)^{-1}H(z) & K(z)^{-1} \end{bmatrix}$$

e si vede che le funzioni di trasferimento F, G, H, K sono univocamente determinate da W e quindi dallo spettro. Dato che G e K non risultano necessariamente dei fattori spettrali "canonici" (a fase minima), è più corretto affermare che sono di fatto gli spettri, $S_{\mathbf{r}}$ ed $S_{\mathbf{v}}$, ad essere univocamente individuati.

Viceversa, consideriamo il caso in cui $S_{\mathbf{r}}$ è uguale a zero e quindi lo spettro congiunto è singolare. Come abbiamo visto più sopra, in questo caso F non è identificabile e pertanto la condizione è anche necessaria. □

In realtà le incognite del problema di identificazione in catena chiusa sono solo la funzione di trasferimento in catena di azione diretta, $F(e^{j\omega})$ e lo spettro dell'errore di modellizzazione relativo, $S_{\mathbf{v}}(e^{j\omega})$. Naturalmete in pratica il modello lineare razionale

$$\hat{\mathbf{y}}(t) = F_{\theta}(z)\mathbf{u}(t)$$

con cui si descrive il legame "deterministico" ingresso-uscita è sempre approssimato e quindi si è sempre in presenza di "errore di modellizzazione $\mathbf{v}$. $\mathbf{v}$ è in effetti sempre presente e certamente non ha molto senso pensare che sia nullo.

Si può così concludere affermando che *se (e solo se) $S_{\mathbf{r}} > 0$, la conoscenza dello spettro congiunto di $\mathbf{y}, \mathbf{u}$, permette di ricavare in modo univoco F e $S_{\mathbf{v}}$;* in altri termini, se $S_{\mathbf{r}} > 0$, dall'osservazione dei segnali $\mathbf{y}, \mathbf{u}$ (per un tempo

teoricamente infinito) si può, in linea di principio, ricavare univocamente il modello in catena di azione diretta del sistema.

La situazione in cui $\mathbf{r} \simeq 0$, si descrive dicendo che *il segnale di riferimento* $\mathbf{r}$ *non è sufficientemente eccitante*. In pratica la situazione $\mathbf{r} = 0$ dev'essere pensata come una situazione limite che serve solo a dare delle indicazioni sulla difficoltà di identificare correttamente il sistema in condizioni di insufficiente eccitazione.

Problema 3 *Discutere l'identificabilità del sistema con retroazione "deterministica",*

$$\begin{aligned}(1+az^{-1})\,\mathbf{y}(t) &= b\,u(t-1)+\mathbf{e}(t)\\ \mathbf{u}(t) &= k\mathbf{y}(t)\end{aligned}$$

in cui $|a-bk|<1$.

Soluzione:
Dato che $S_{\mathbf{r}}\equiv 0$ (non c'è segnale di riferimento esterno) e quindi la retroazione è "deterministica"; i.e. $\mathbf{u}(t)=H(z)\,\mathbf{y}(t)$, lo spettro congiunto è singolare. Quindi ci dobbiamo aspettare che il sistema non sia identificabile. In effetti, $\mathbf{y}$ è descritto dal modello AR

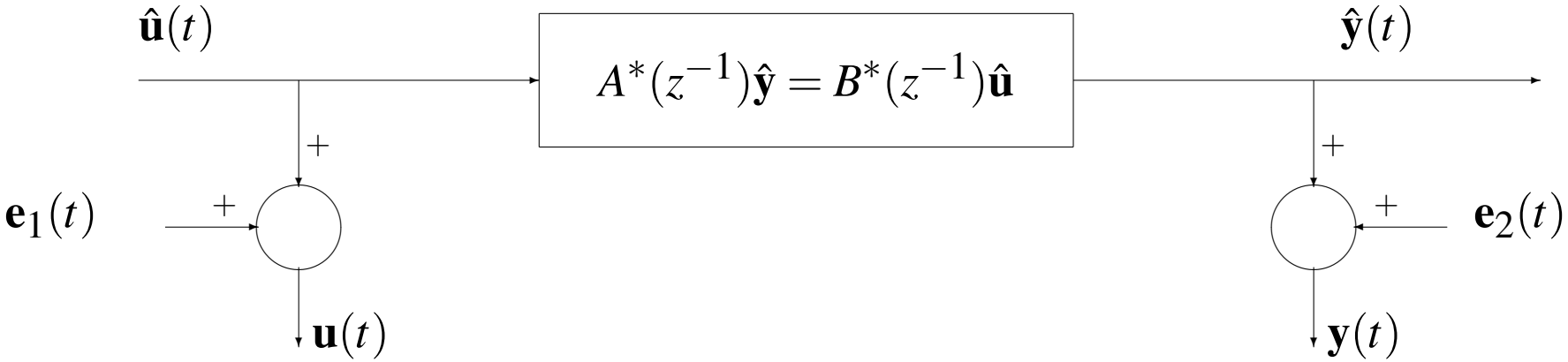
$$[1+(a-bk)z^{-1}]\,\mathbf{y}(t)=\mathbf{e}(t)$$

e tutte le coppie di parametri $[a\,,b]$ per cui $a-bk=c$ con $|c|<1$ descrivono lo stesso sistema.

□

MODELLI A ERRORI NELLE VARIABILI

Lo schema a blocchi di figura descrive un cosiddetto modello a *Errori nelle Variabili (EIV)*



in cui le variabili cosiddette “vere”, $\hat{\mathbf{u}}(t)$ e $\hat{\mathbf{y}}(t)$ sono legate tra di loro da una equazione alle differenze del tipo

$$A^*(z^{-1})\,\hat{\mathbf{y}}(t) = B^*(z^{-1})\,\hat{\mathbf{u}}(t)$$

e sono osservate in presenza di rumori additivi $\mathbf{e}_1(t)$ e $\mathbf{e}_2(t)$ (che si assumono spesso bianchi), scorrelati tra di loro ($\mathbf{e}_1 \perp \mathbf{e}_2$) e con le variabili vere,

$$\begin{cases} \mathbf{u}(t) = \hat{\mathbf{u}}(t) + \mathbf{e}_1(t) & \mathbf{e}_1 \perp \hat{\mathbf{u}} \\ \mathbf{y}(t) = \hat{\mathbf{y}}(t) + \mathbf{e}_2(t) & \mathbf{e}_2 \perp \hat{\mathbf{y}} \end{cases}.$$

Nel caso in cui $\mathbf{e}_1(t)$ e $\mathbf{e}_2(t)$sono bianchi, è facile mostrare che $\mathbf{y}(t)$ e $\mathbf{u}(t)$ sono legate da un modello ARMAX. Calcolare i relativi polinomi e il rumore bianco in ingresso al modello.

Discutere se e sotto quali condizioni ci può essere assenza di reazione da $\mathbf{y}$ ad $\mathbf{u}$.

MINIMIZZAZIONE DELL'ERRORE DI PREDIZIONE

Tratteremo di identificazione di modelli lineari di tipo ingresso-uscita del tipo discusso precedentemente .

Metodi basati sulla **minimizzazione dell'errore di predizione** (in inglese *PEM = Prediction Error Methods*). Questi metodi hanno costituito per lungo tempo il cavallo di battaglia dell'identificazione. Il merito di averne proposto e propagandato capillarmente l'uso va senz'altro ascritto a Lennart Ljung.

Principio su cui si basano i metodi PEM :
Dato un modello $M(\theta)$ appartenente ad una assegnata classe parametrica $\mathcal{M} \equiv \{M(\theta)\,;\,\theta \in \Theta\}$ e una sequenza di dati ingresso-uscita

$$y^N := \{y(t)\,;\,t = 1,2,\ldots,N\}\,, \qquad u^N := \{u(t)\,;\,t = 1,2,\ldots,N\}$$

si procede come segue:

1. Per un generico valore di θ, si costruisce il (miglior, secondo qualche criterio) predittore all'istante $t-1$ dell'uscita successiva, $y(t)$. Per ogni θ fissato, questo predittore è una funzione (deterministica) dei dati passati, denotata col simbolo $\hat{y}_\theta(t \mid t-1)$ (oppure $\hat{M}(\theta)$), che produce la (miglior) predizione di $y(t)$ effettuabile in base al modello selezionato ed ai dati misurati,

$$\hat{M}(\theta) : (y^{t-1},\, u^{t-1}) \mapsto \hat{y}_\theta(t \mid t-1)$$

La predizione $\hat{y}_\theta(t \mid t-1)$ si può pensare come funzione dei dati passati (oltre che del parametro θ) e quindi, come una quantità aleatoria (prima di aver misurato i dati). In questo contesto verrà impiegato il simbolo $\hat{\mathbf{y}}_\theta(t \mid t-1)$.

2. Si formano gli *errori di predizione*:

$$\varepsilon_\theta(t) := y(t) - \hat{y}_\theta(t)\,; \qquad t = 1,2,\ldots,N$$

che, analogamente a quanto detto per il predittore, possono essere all'occorrenza interpretati come quantità aleatorie, indicate con simboli in grassetto, i.e. $\boldsymbol{\varepsilon}_\theta(t)$.

Notiamo ad esempio che per la classe di modelli $(\dagger)$, usando formalmente l'espressione per il predittore di Wiener, si ottiene

$$\begin{aligned} \boldsymbol{\varepsilon}_{\theta}(t) &= \mathbf{y}(t) - \hat{\mathbf{y}}(t \mid t-1) = \mathbf{y}(t) - G_{\theta}(z)^{-1}\left[F_{\theta}(z)\mathbf{u}(t) + (G_{\theta}(z)-1)\mathbf{y}(t)\right] \\ &= G_{\theta}(z)^{-1}\left[\mathbf{y}(t) - F_{\theta}(z)\mathbf{u}(t)\right] \end{aligned} \tag{66}$$

dalla quale si può ricavare una rappresentazione del processo ("vero") osservato $\mathbf{y}$ mediante un *modello scelto arbitrariamente nella classe* $\mathcal{M}$; i.e.

$$\mathbf{y}(t) = F_{\theta}(z)\mathbf{u}(t) + G_{\theta}(z)\boldsymbol{\varepsilon}_{\theta}(t)\,, \tag{67}$$

Notare però che in queste rappresentazioni l'innovazione è sostituita dall'errore di predizione (che in generale non è bianco).
L'idea di rappresentazione mediante l'errore di predizione troverà applicazioni importanti nel seguito e lo studente è invitato a meditare sul suo significato.

3. Si minimizza rispetto a θ una cifra di merito **l'errore quadratico medio di predizione** che descriva quanto bene (in media) il modello predice il dato successivo:

$$V_N(\theta) := \frac{1}{N}\sum_{t=1}^{N} \varepsilon_\theta(t)^2$$

o, più in generale, una media degli errori quadratici di predizione pesati da un fattore di sconto non negativo $\beta(N,t)$,

$$V_N(\theta) := \frac{1}{N}\sum_{t=1}^{N} \beta(N,t)\varepsilon_\theta(t)^2 \qquad \beta(t,N) > 0 \tag{68}$$

che per N piccoli dà peso minore agli errori di predizione compiuti nella fase iniziale dell'algoritmo quando l'influenza di condizioni iniziali stimate in modo approssimato (o incognite) è più deleteria. Per $N \to \infty$ il fattore di sconto tende a diventare inutile, e si aggiustano le cose in modo che $\beta(N,t) \to 1$.

Si può anche considerare, invece di ε_θ, un errore di predizione *filtrato* da un opportuno filtro lineare che pesi di più gli errori nella banda di frequanze dove più interessa una identificazione accurata. Infine si può modulare l'errore di predizione attarverso una opportuna funzione non lineare che "saturi" per valori molto grandi di ε_θ e serva a ridurre l'influenza di *outliers* accidentali. In ogni caso, dalla minimizzazione della cifra di merito si ricava una stima di θ,

$$\hat{\theta}_N := \operatorname{Arg}\min_{\theta} V_N(\theta) \tag{69}$$

che è appunto la stima PEM del parametro del modello. Naturalmente lo stimatore $\hat{\boldsymbol{\theta}}_N$ che produce la stima come funzione dei dati, viene chiamato *stimatore PEM* del parametro θ.

4. Infine si prende come stima della varianza dell'innovazione $\lambda^2 = \operatorname{var}\{\mathbf{e}(t)\}$, *l'errore quadratico residuo*, ovvero

$$\hat{\lambda}_N^2 := V_N(\hat{\theta}_N) \tag{70}$$

dove V_N è dato da una delle espressioni precedenti.

OSSERVAZIONE:

Per quanto questa procedura possa apparire intuitivamente sensata, l'unica giustificazione valida per la sua adozione nei procedimenti di identificazione sta nelle sue proprietà statistiche. Per questo motivo la teoria sarà sostanzialmente dedicata all'analisi delle proprietà statistiche dello stimatore PEM.

IDENTIFICAZIONE PEM DI MODELLI ARX

Consideriamo modelli ARX (d'innovazione) parametrizzati a "scatola nera",

$$A(z^{-1})\mathbf{y}(t) = B(z^{-1})\mathbf{u}(t) + \mathbf{e}(t) \qquad \text{(ARX)}$$

assumendo che i gradi n ed m dei due polinomi A e B siano stati fissati. Definendo

$$\boldsymbol{\varphi}(t)^\top = \begin{bmatrix} -\mathbf{y}(t-1) & \dots & -\mathbf{y}(t-n) & \mathbf{u}(t-1) & \dots & \mathbf{u}(t-m) \end{bmatrix}$$

e il vettore dei $p := n+m$ parametri incogniti

$$\theta := \begin{bmatrix} a_1 & \dots & a_n & b_1 & \dots & b_m \end{bmatrix}^\top$$

questo modello si può scrivere in forma di regressione come

$$\mathbf{y}(t) = \boldsymbol{\varphi}(t)^\top \theta + \mathbf{e}(t)\,.$$

Supponiamo che sia disponibile una serie temporale di dati ingresso-uscita $\{y(t)\,, u(t)\,; t = t_0, t_0-1, \dots, 0, 1, 2, \dots, N\}$ che vogliamo descrivere con un modello della classe (ARX) usando il metodo PEM.

Dato che il modello è di innovazione $\mathbf{e}(t) \perp \mathbf{y}^{t-1}, \mathbf{u}^{t-1}$ e $\boldsymbol{\varphi}(t)^\top \theta$ è il predittore a minima varianza di $\mathbf{y}(t)$ che ha la struttura di un filtro FIR

$$\hat{\mathbf{y}}_\theta(t \mid t-1) = \boldsymbol{\varphi}(t)^\top \theta$$

che è manifestamente una *funzione lineare del parametro* θ.
L'istante iniziale t_0 è preso opportunamente, in modo che $\boldsymbol{\varphi}(t)$ non coinvolga campioni iniziali non noti dei processi $\{\mathbf{y}(t)\}$ e $\{\mathbf{u}(t)\}$.
$\hat{\mathbf{y}}_\theta(t \mid t-1)$ coincide con il predittore a regime (di Wiener).
Definendo i vettori colonna $\mathbf{y}$ ed $\mathbf{e}$ con componenti $\mathbf{y}(t)$ e $\mathbf{e}(t)$ per $t = 1, 2, \ldots, N$ e la matrice $N \times p$

$$\Phi_N := \begin{bmatrix} \varphi(1)^\top \\ \vdots \\ \varphi(N)^\top \end{bmatrix}, \tag{71}$$

il modello lineare (ARX) si può riscrivere

$$\mathbf{y} = \Phi_N \theta + \mathbf{e}$$

Il vettore N-dimensionali dei predittori e degli errori di predizione corrispondenti

$$\hat{\mathbf{y}}_\theta = \Phi_N \theta\,, \qquad \boldsymbol{\varepsilon}_\theta = \mathbf{y} - \Phi_N \theta$$

La cifra di merito $V_N(\theta)$ è il quadrato della norma Euclidea di $\boldsymbol{\varepsilon}_\theta$,

$$V_N(\theta) = \frac{1}{N} \|\mathbf{y} - \Phi_N \theta\|^2 .$$

La minimizzazione dell'errore quadratico medio di predizione porta a un **problema di minimi quadrati**. Lo stimatore PEM del parametro θ ha la nota espressione

$$\hat{\theta}_N = \left[\Phi_N^\top \Phi_N\right]^{-1} \Phi_N^\top \mathbf{y}$$

che si può anche riscrivere

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

Quindi **l'identificazione di modelli ARX si può ridurre ad un problema di minimi quadrati**, che si sa risolvere esplicitamente. Questo fatto è ovviamente un grosso incentivo all'uso di questi modelli.

Osservazioni

• Facciamo l'ipotesi che i dati siano distribuiti secondo un **modello "vero"** della classe (ARX) con gradi n, m (struttura) fissati. Quali sono le proprietà statistiche di questo stimatore?

• $\hat{\theta}_N$ è adesso una funzione NON LINEARE dei dati. Non sappiamo se è corretto e non sappiamo calcolarne la varianza. Potremo solo studiare cosa accade per $N \to \infty$.

• Anche se i processi $\mathbf{y}$ e $\mathbf{u}$ fossero Gaussiani $\hat{\theta}_N$ ha distribuzone impossibile da calcolare.

L'identificazione con dati finiti non si sa analizzare!

Problema 4 *Si vuole identificare il modello (ARX) minimizzando una somma pesata degli errori di predizione*

$$V_N(\theta,\beta) := \frac{1}{N}\sum_{t=1}^{N} \beta(N,t)\,\varepsilon_\theta(t)^2$$

dove $0 < \beta(N,\cdot) \leq 1$ *è una funzione peso assegnata (interpretabile come fattore d'oblio per gli errori "vecchi"). Si chiede di trovare l'espressione di* $\hat{\theta}(N)$*.*

Soluzione:
Definiamo i vettori colonna $\mathbf{y}$ ed $\mathbf{e}$ con componenti $\mathbf{y}(t)$ e $\mathbf{e}(t)$ per $t = 0,1,2,\ldots,N$ e le matrici

$$\Phi_N := \begin{bmatrix} \varphi(1)^\top \\ \vdots \\ \varphi(N)^\top \end{bmatrix}, \qquad Q_N := \operatorname{diag}\{\beta(N,1),\ldots,\beta(N,N)\}$$

il modello lineare si può riscrivere $\mathbf{y} = \Phi_N\,\theta + \mathbf{e}$ e la cifra di merito $V_N(\theta,\beta)$ come il quadrato di una norma *pesata*,

$$NV_N(\theta,\beta) = \|\mathbf{y} - \Phi_N\,\theta\|^2_{Q_N}.$$

Usando le formule dei minimi quadrati pesati si trova

$$\hat{\theta}(N) = \left[\Phi_N^\top Q_N \Phi_N\right]^{-1} \Phi_N^\top Q_N \mathbf{y}$$

che si riscrive per esteso come

$$\hat{\theta}(N) = \left[\sum_{t=1}^{N} \beta(N,t)\varphi(t)\varphi(t)^\top\right]^{-1} \sum_{t=1}^{N} \beta(N,t)\varphi(t)\mathbf{y}(t)\,.$$

Sostituendo l'espressione di $\mathbf{y}(t)$ si trova poi

$$\hat{\theta}(N) = \theta + \left[\frac{1}{N}\sum_{t=1}^{N} \beta(N,t)\varphi(t)\varphi(t)^\top\right]^{-1} \frac{1}{N}\sum_{t=1}^{N} \beta(N,t)\varphi(t)\mathbf{e}(t)\,.$$

Dato che $\mathbb{E}\,\varphi(t)\varphi(t)^\top < \infty$ si vede che il termine tra parentesi quadre si mantiene limitato (potrebbe in particolare convergere a una costante); si tratta allora di trovare condizioni su β per cui

$$\frac{1}{N}\sum_{t=1}^{N} \beta(N,t)\varphi(t)\mathbf{e}(t) \to 0$$

Una condizione sufficiente è che $\beta(N,t) \to 1$ per $N \to \infty$, uniformemente in t. In questo caso per il teorema ergodico si ha

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \beta(N,t)\varphi(t)\mathbf{e}(t) = \mathbb{E}\,\varphi(t)\mathbf{e}(t) = 0\,.$$

Di fatto negli algoritmi ricorsivi si cerca di rendere il fattore d'oblio monotono in N e t e convergente a 1 per $N \to \infty$. □

CALCOLO DEL PREDITTORE

La minimizzazione dell'errore quadratico medio di predizione richiede il calcolo del predittore $\hat{\mathbf{y}}_\theta(t \mid t-1)$ basato sul modello $M(\theta)$ e sui dati disponibili.
Nel caso di modelli (ARX) tutto è semplice perchè il predittore coincide col predittore di Wiener che dipende in realtà solo da una finestra finita di dati passati.
Se si identifica un modello più generale (ad es ARMAX) questo non è più vero. Per calcolare il **predittore con dati finiti** si deve intendere un predittore *non stazionario*, realizzato impiegando un **filtro di Kalman** basato su un opportuno modello di stato.

Problema 5 *Volete identificare un modello ARMA (d'innovazione) del tipo*

$$(1 + az^{-1})\mathbf{y}(t) = (1 + cz^{-1})\mathbf{e}(t)\,.$$

con il metodo PEM. Purtroppo i dati sono pochi (la numerosità campionaria N è decisamente finita) e non si può usare il predittore di Wiener, ma bisogna usare il predittore di Kalman costruito su un opportuno modello di stato (minimo; i.e. di ordine uno) che descriva il processo $\mathbf{y}$.

Descrivete per passi un algoritmo iterativo per il calcolo dello stimatore PEM del parametro $\theta := [a\ c]^\top$ basato su un campione di N osservazioni. In particolare descrivete il passo di aggiornamento del modello (incluse le condizioni iniziali) e del predittore. Non serve addentrarsi troppo nell'algoritmo di ottimizzazione (e in ispecie sul calcolo del gradiente).

Soluzione:

Per calcolare il predittore a memoria finita (non stazionario) serve una realizzazione di stato del modello

$$(1+az^{-1})\mathbf{y}(t) = (1+cz^{-1})\mathbf{e}(t)$$

ad esempio la

$$\begin{aligned} \mathbf{x}(t+1) &= -a\mathbf{x}(t) + (c-a)\mathbf{e}(t) \\ \mathbf{y}(t) &= \mathbf{x}(t) + \mathbf{e}(t)\,. \end{aligned}$$

dove

$$Q = \lambda^2(c-a)^2; \quad S = \lambda^2(c-a); \quad R = \lambda^2\,.$$

Dato che il modello deve descrivere un processo stazionario, la varianza di stato iniziale si trova risolvendo Lyapunov: $p = a^2p + \lambda^2(c-a)^2$ che fornisce la condizione iniziale

$$p_0 = \frac{\lambda^2(c-a)^2}{1-a^2}$$

che si usa per inizializzare l'equazione di Riccati e per il calcolo del guadagno

$$\begin{aligned} p(t+1) &= a^2 p(t) - k(t)^2 \lambda(t) + \lambda^2 (c-a)^2; \\ k(t) &= (a p(t) + \lambda^2 (c-a)) \lambda(t)^{-1}; \\ \lambda(t) &= p(t) + \lambda^2. \end{aligned}$$

Introducendo la varianza normalizzata

$$\pi(t) := \frac{p(t)}{\lambda^2}$$

si può riscrivere tutto nella forma:

$$\begin{aligned} \pi(t+1) &= a^2 \pi(t) - k(t)^2 \beta(t) + (c-a)^2; \\ k(t) &= (a\pi(t) + (c-a)) \beta(t)^{-1}; \quad \beta(t) = \frac{\lambda(t)}{\lambda^2} = \pi(t) + 1; \\ \pi(0) &= \frac{(c-a)^2}{1-a^2} \end{aligned}$$

che non dipende più dal parametro λ^2 ma solo dal parametro bidimensionale $\theta := [a\ c]^\top$. Questo è in accordo col fatto che in generale il predittore lineare a minima varianza non dipende dalla varianza del rumore.

Il guadagno $k(t) \equiv k_\theta(t)$ si usa per calcolare il predittore a memoria finita basato sugli ultimi t dati:

$$\hat{\mathbf{x}}_\theta(t+1|t) = (a - k_\theta(t))\hat{\mathbf{x}}_\theta(t|t-1) + k_\theta(t)\mathbf{y}(t)\,; \qquad \hat{\mathbf{x}}(1|0) = 0 \tag{72}$$

$$\hat{\mathbf{y}}_\theta(t+1|t) = \hat{\mathbf{x}}_\theta(t+1|t)\,, \qquad t = 1,2,\ldots,N. \tag{73}$$

Usando queste equazioni per un vettore **fissato** di parametri ammissibili $\theta = [a\,c]^\top$ con $|a| < 1\,; |c| < 1$, si definisce una funzione $\hat{y}_\theta^N = \mathrm{pred}(\theta, y^N)$ che produce il vettore a N componenti $\hat{y}_\theta^N$ dei predittori di un passo (non stazionari) basati sul modello di parametri θ e sui dati y^N. Si può così formare l'errore quadratico medio di predizione

$$V_N(\theta) = \frac{1}{N}\|\varepsilon_\theta\|^2\,; \qquad \varepsilon_\theta = y^N - \hat{y}_\theta^N$$

che si miimizza rispetto a θ usando un algoritmo iterativo, ad esempio un metodo di quasi-Newton del tipo

$$\theta_{k+1} = \theta_k + \left[\sum_{t=1}^{N} \psi_{\theta_k}(t)\psi_{\theta_k}(t)^\top\right]^{-1} \sum_{t=1}^{N} \psi_{\theta_k}(t)\varepsilon_{\theta_k}(t) \tag{74}$$

$$\lambda_k^2 = V_N(\theta_k) \tag{75}$$

dove $\psi_\theta(t)$ è il gradiente di $\hat{y}_\theta(t) \equiv \hat{y}_\theta(t \mid t-1)$

$$\psi_\theta(t) := \frac{\partial \hat{y}_\theta(t)}{\partial \theta} = \frac{\partial \hat{x}_\theta(t|t-1)}{\partial \theta}$$

Ad ogni iterazione bisogna riaggiornare il calcolo della stringa dei predittori valutando la funzione

$$\hat{y}^N_{\theta_k} = \mathrm{pred}(\theta_k, y^N)\,; k = 1, 2, \ldots.$$

Il calcolo del gradiente è un pò complicato e non ce ne occuperemo. □

CONSISTENZA

In un procedimento sensato di inferenza ci si aspetta di ottenere risultati sempre migliori al crescere della numerosità del campione.

Si immagina di avere una successione infinita di osservazioni $\{y_t\}$, che si può pensare come una traiettoria di un processo $\{\mathbf{y}(t)\}$ a tempo discreto. Denotiamo con $F_\theta^N(\cdot)$ la distribuzione di probabilità congiunta delle prime N variabili del processo, che supponiamo essere nota a meno di un certo parametro (di dimensione fissa) θ. Sia $\{\phi_N\}$ una successione di stimatori del parametro θ che immaginiamo definiti in base allo stesso criterio di stima (ad esempio la massima verosimiglianza).

Definizione 16 *Sia* $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ *un campione estratto dalla distribuzione* $F_{\theta_0}^N$ *appartenete alla famiglia* $\{F_\theta^N ; \theta \in \Theta\}$ *(*θ_0 *è il valore* vero *del parametro). Lo stimatore* $\phi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ *si dice* **consistente** *se*

$$\lim_{N\to\infty} \phi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N) = \theta_0 \quad ;$$

Se il limite è un *limite con probabilità 1* (c.p. 1), si parla di consistenza forte. In questo caso si ha

$$\lim_{N\to\infty} \phi_N(y_1,\ldots,y_N) = \theta_0$$

per "quasi tutte" le possibili successioni $\{y_1, y_2, \ldots, y_N, \ldots\}$ di osservazioni. Supponiamo per semplicità le misure scalari ($m = 1$). Lo spazio campionario è $\mathbb{R}^\infty := \mathbb{R} \times \mathbb{R} \times \ldots$ (infinite volte). La probabilità è definita sullo spazio di tutte le misure "infinitamente lunghe", $\mathbb{R}^\infty$.

Definizione 17 *Se il limite è un limite in probabilità, cioè se* $\forall \varepsilon > 0$,

$$\lim_{N\to\infty} P_{\theta_0}\left(\|\phi_N(\mathbf{y}_1,\ldots,\mathbf{y}_N) - \theta_0\| \geq \varepsilon\right) = 0 \quad ,$$

si parla di consistenza "debole".

In ogni caso la probabilità a cui si fa riferimento nella definizione è quella secondo cui le v.c. $\mathbf{y}_1,\ldots,\mathbf{y}_N,\ldots$ sono *realmente* distribuite, cioè la probabilità *vera*, corrispondente al valore "vero" θ_0 del parametro.

N.B. la consistenza è una proprietà molto più forte della **correttezza asintotica** :

$$\lim_{N\to\infty} E_\theta\, \phi_N(\mathbf{y}_1,\ldots,\mathbf{y}_N) = \theta \qquad \forall\theta \in \Theta$$

che ha scarso significato statistico.

In generale risultati relativi alla consistenza forte si dimostrano usando il **Teorema Ergodico** di Birkhoff del quale a cosiddetta *Legge (forte) dei grandi numeri* è un caso particolare.

La disuguaglianza di Chebyshev permette di provare la convergenza in probabilità a partire semplicemente dalla convergenza di medie e varianze. Usando la classica disuguaglianza,

$$P_\theta\Big(\|\phi_N - \theta\| \geq \varepsilon\Big) \leq \frac{1}{\varepsilon^2}\, E_\theta\left[(\phi_N - \theta)^\top (\phi_N - \theta)\right] = \frac{1}{\varepsilon^2}\, E_\theta \|\phi_N - \theta\|^2,$$

dove $\phi_N := \phi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ e $\|\cdot\|$ indica l'usuale norma euclidea. Notiamo che se ϕ_N è corretto $E_\theta\, \phi_N = \theta$ e l'espressione a secondo membro è la varianza, $\sigma_N^2(\theta)$, di $\phi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ divisa per ε^2. Si vede che se

$$\lim_{N\to\infty} \sigma_N^2(\theta) = 0 \quad , \qquad \forall \theta \in \Theta \quad ,$$

allora $\phi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ è (debolmente) consistente.

Proposizione 20 *Se $\phi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ è, uno stimatore asintoticamente corretto e se la sua varianza scalare $\sigma_N^2(\theta)$ tende a zero con N per ogni $\theta \in \Theta$, allora $\phi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ è debolmente consistente.*

Esempio 2 Dalle formule (17) e (18) si vede che lo stimatore della varianza σ^2 nel modello lineare (14) è asintoticamente corretto e consistente.

Esempio 3 Supponiamo che la distribuzione vera di $\mathbf{y}$ (scalare) sia del tipo di Cauchy, ovvero $dF_{\theta_0}(y) = p(y,\theta_0)\,dy$ con

$$p(y,\theta) = \frac{1}{\pi}\,\frac{1}{1+(y-\theta)^2} \quad , \qquad \theta \in R \quad .$$

Sia $(\mathbf{y}_1,\ldots,\mathbf{y}_N)$ un campione casuale e $\bar{\mathbf{y}}_N$ la relativa media campionaria. Usando le funzioni caratteristiche si può vedere che $\bar{\mathbf{y}}_N$ ha, per ogni N, la stessa distribuzione di $\mathbf{y}$ e pertanto la probabilità

$$P(|\bar{\mathbf{y}}_N - \theta_0| \geq \varepsilon)$$

rimane la stessa al variare di N e non può dunque tendere a zero con N. Questo implica che $\bar{\mathbf{y}}_N$ non è consistente (tanto meno è fortemente consistente dato che $E_{\theta_0}\mathbf{y} = \infty$!).

ERGODICITÀ

Si potrebbe ben affermare che i due risultati veramente fondamentali della teoria della probabilità sono il *teorema ergodico* e il *teorema del limite centrale*. Questi due teoremi sono praticamente gli unici due risultati della teoria (che è assiomatica, come tutte le teorie matematiche) che permettono di stabilire un legame col mondo empirico e su di essi si basa la verifica e l'analisi delle proprietà dei procedimenti di inferenza statistica.

Questi teoremi permettono di formulare previsioni "sperimentalmente verificabili" su certe classi di esperimenti aleatori (anche se un pò idealizzati) e su procedimenti di inferenza basati sui risultati di questi esperimenti.

Sia il teorema ergodico che il teorema del limite centrale sono teoremi limite che si riferiscono al caso in cui il numero di osservazioni su cui si basa la costruzione di una certa statistica o di un certo procedimento di inferenza, tende all' infinito.

PROCESSI STAZIONARI IN SENSO STRETTO

IPOTESI: Le misure formano un processo stocastico stazionario **in senso stretto !**

Definizione 18 *Il processo* $\{\mathbf{y}(t)\}$ *è stazionario (in senso stretto) se tutte le sue distribuzioni di ordine finito sono invarianti per traslazione temporale, ovvero si ha, per ogni* n*,*

$$F_n(x_1,\ldots,x_n,t_1+\Delta,\ldots,t_n+\Delta) = F_n(x_1,\ldots,x_n,t_1,\ldots,t_n) \quad ,$$

identicamente in $x_1,\ldots,x_n$*,* $t_1,\ldots,t_n$*, qualunque sia* $\Delta \in \mathbb{Z}$*.*

Conseguenze ben note della definizione sono:

- la distribuzione di probabilità del primo ordine $F(x,t)$ di un processo stazionario $\{\mathbf{y}(t)\}$ non dipende da t; ovvero le variabili, $\mathbf{y}(t)$, $t \in \mathbb{Z}$, sono tutte *identicamente distribuite*;

- la distribuzione congiunta (del second'ordine) $F_2(x_1, x_2, t_1, t_2)$ delle variabili $\mathbf{y}(t_1)$, $\mathbf{y}(t_2)$, dipende solo dallo scostamento temporale $\tau = t_1 - t_2$ e non dall'origine dei tempi (o dalla "data") a cui ci si riferisce.
  In particolare la *media* del processo, $\mu(t) := E\,\mathbf{y}(t)$, è costante nel tempo, uguale ad un certo vettore fisso $\mu \in \mathbb{R}^m$ e la *matrice di covarianza*

$$\Sigma(t_1, t_2) := E\left[\mathbf{y}(t_1) - \mu(t_1)\right]\,\left[\mathbf{y}(t_2) - \mu(t_2)\right]^\top$$

  dipende solo dalla distanza temporale $\tau = t_1 - t_2$.

LO SPAZIO DELLE STATISTICHE TEMPO-INVARIANTI

Sia $\mathbb{I}$ un sottoinsieme qualunque, finito o infinito, di $\mathbb{Z}$ e consideriamo funzioni f (misurabili) *che non dipendono esplicitamente dal tempo*, delle variabili $\{\mathbf{y}(\tau)\,;\,\tau \in \mathbb{I}\}$. Si definiscono così le variabili aleatorie:

$$\mathbf{z} = f(\mathbf{y}(\tau)\,;\,\tau \in \mathbb{I}) \quad ,$$

che sono funzioni "tempo invarianti" del processo.
Ad esempio, per un processo scalare $\{\mathbf{y}(t)\}$, si possono considerare espressioni del tipo

$$\mathbf{z} = \mathbf{y}^2(0) + 3\mathbf{y}^2(1)\,\mathbf{y}(-1) + \cos\mathbf{y}(2) \quad ,$$

oppure

$$\mathbf{z} = \sum_{-\infty}^{+\infty} c_i\,\mathbf{y}(i) \quad ,$$

dove i c_i sono numeri reali e la serie si suppone convergente.

Per semplicità supporremo che f (e quindi $\mathbf{z}$) prenda solo valori reali. La generalizzazione a funzioni vettoriali (e matriciali) è immediata.

Consideriamo solo funzioni f tali per cui

$$E\,|\mathbf{z}| = E\,|f(\mathbf{y}(\tau) \mid \tau \in \mathbb{I})| < \infty.$$

e denotiamo con $L^1(\mathbf{y})$ lo spazio vettoriale popolato dalle funzioni del processo $\{\mathbf{y}(t)\}$ che soddisfano a questa condizione. Chiaramente $L^1(\mathbf{y})$ è uno spazio vettoriale reale e si può mostrare che con l'introduzione della norma

$$\|\,\mathbf{z}\,\| := E\,|\,\mathbf{z}\,|$$

$L^1(\mathbf{y})$ diventa uno spazio di Banach (quindi completo). Analogamente, si può definire $L^2(\mathbf{y})$ come lo lo spazio vettoriale delle funzioni del processo $\{\mathbf{y}(t)\}$ per cui $E\,|\mathbf{z}|^2 < \infty$. Quest'ultimo è in realtà uno spazio di Hilbert rispetto al solito prodotto scalare tra variabili aleatorie. Per la disuguaglianza di Schwartz, ogni variabile aleatoria che ha momento del second'ordine finito ha necessariamente anche media finita, per cui $L^1(\mathbf{y}) \supset L^2(\mathbf{y})$ (come spazi vettoriali).

Dato un processo strettamente stazionario $\{\mathbf{y}(t)\}$, si può definire una intera classe di processi, ancora strettamente stazionari, che sono "funzioni di $\{\mathbf{y}(t)\}$", ponendo

$$\mathbf{z}(t) := f(\mathbf{y}(t+\tau)\,;\,\tau \in \mathbb{I}) \quad , \qquad t \in \mathbb{Z} \quad ,$$

la variabile casuale $\mathbf{z}(t)$ si ottiene "traslando" le variabili $\{\mathbf{y}(\tau)\}$ nell'argomento di f di t unità temporali. Notazione:

$$f(\mathbf{y}(t+\tau)\,;\,\tau \in \mathbb{I}) := f_t(\mathbf{y})$$

Al variare di t in $\mathbb{Z}$, la variabile $\mathbf{z}(t)$ descrive ancora un processo stocastico (scalare) $\{\mathbf{z}(t)\}$ *stazionario in senso stretto*.
Ad esempio, per la stazionarietà di $\{\mathbf{y}(t)\}$, si ha

$$\begin{aligned} P\{\mathbf{z}(t) \in A\} &= P\Big\{ f(\mathbf{y}(t+\tau)\,;\,\tau \in \mathbb{I}) \in A \Big\} \\ &= P\Big\{ (\mathbf{y}(t+\tau_1), \ldots, \mathbf{y}(t+\tau_N)) \in f^{-1}(A) \Big\} \\ &= P\Big\{ (\mathbf{y}(\tau_1), \ldots, \mathbf{y}(\tau_N)) \in f^{-1}(A) \Big\} \\ &= P\{\mathbf{z} \in A\}\,, \end{aligned}$$

Definizione 19 *La variabile* $\mathbf{z} = f(\mathbf{y})$ *è* **invariante** *(per traslazione) se* $\mathbf{z}(t) = \mathbf{z}$ *per ogni* $t \in \mathbb{Z}$.

Un modo equivalente (anche se un pò formalistico) di definire l'invarianza è attraverso l'introduzione dell'operatore di *traslazione temporale* U, che agisce traslando nel tempo le variabili del processo $\mathbf{y}$ ed è definito tramite la posizione

$$U\mathbf{y}_k(t) = \mathbf{y}_k(t+1)\,, \qquad k = 1, 2, \ldots, m\,.$$

L' operatore U può essere esteso a tutte le variabili $\mathbf{z} \in L^1(\mathbf{y})$ semplicemente ponendo, se $\mathbf{z} = f(\mathbf{y}(\tau) \mid \tau \in \mathbb{I})$,

$$U\mathbf{z} = f(\mathbf{y}(\tau+1) \mid \tau \in \mathbb{I}) = \mathbf{z}(1)\,.$$

e si vede che U è lineare, invertibile ($U^{-1}\mathbf{y}(t) = \mathbf{y}(t-1)$) e può essere iterato più volte dando luogo ad una famiglia di trasformazioni lineari $\{U^t\}_{t\in\mathbb{Z}}$ (operatori di traslazione temporale) su $L^1(\mathbf{y})$.

Ogni variabile aleatoria $\mathbf{z} \in L^1(\mathbf{y})$ può essere traslata nel tempo

$$U^t \mathbf{z} := \mathbf{z}(t) \quad ,$$

dove $\mathbf{z}(t) := f_t(\mathbf{y})$. Per la stazionarietà di $\{\mathbf{z}(t)\}$ U^t è un *operatore che preserva la norma in* $L^1(\mathbf{y})$.:

$$\|\mathbf{z}(t)\| = \mathbb{E}|f_t(\mathbf{y})| = \mathbb{E}|f_0(\mathbf{y})| = \|\mathbf{z}\|$$

Sia $\{\mathbf{z}_k\}$ una successione convergente in $L^1(\mathbf{y})$. Dato che

$$\mathbb{E}\,|U^t(\mathbf{z}_n - \mathbf{z}_m)| = \mathbb{E}\,|\mathbf{z}_n - \mathbf{z}_m|$$

si vede che U^t *è una trasformazione continua rispetto alla convergenza (in media) in* $L^1(\mathbf{y})$.

Dalla definizione:
La variabile casuale $\mathbf{z} \in L^1(\mathbf{y})$ è *invariante* se e solo se è invariante per l'operatore U, ovvero

$$U\mathbf{z} = \mathbf{z} \quad .$$

Equivalente a dire che $\mathbf{z}$ è invariante se e solo se

$$\mathbf{z}(t) = U^t\mathbf{z} = \mathbf{z} \quad , \qquad \forall t \in \mathbb{Z} \quad , \tag{76}$$

il che significa ancora che $\mathbf{z} = f(\mathbf{y}(\tau)\,;\tau \in \mathbb{I})$ non cambia, comunque si traslino temporalmente le variabili $\mathbf{y}(\tau)$ del processo. Ne segue che o $\mathbf{z}$ non dipende affatto dal processo ed è una costante deterministica, oppure dipende solo dal "comportamento asintotico" di $\{\mathbf{y}(t)\}$ nell'intorno di $\pm\infty$.

Esempio 4 Supponiamo che esista il $\lim_{t\to\infty}\mathbf{y}(t)$ (con probabilità 1) e sia $\mathbf{z}$ la variabile aleatoria

$$\mathbf{z} := \lim_{t\to\infty}\mathbf{y}(t) \quad .$$

Allora, per la continuità di U, si ha

$$U\mathbf{z} = \lim_{t\to\infty} U\,\mathbf{y}(t) = \lim_{t\to\infty}\mathbf{y}(t+1) = \mathbf{z}$$

e $\mathbf{z}$ è invariante. Sono invarianti anche le variabili $\limsup_{t\to\pm\infty}\mathbf{y}(t)$ e $\liminf_{t\to\pm\infty}\mathbf{y}(t)$ Un discorso analogo può essere fatto per nozioni di limite più generali, ad esempio nel senso delle medie di Cesàro ma questo discorso verrà ripreso più avanti.

Una variabile invariante (non banale) può solo dipendere dalla "coda" infinitamente futura o infinitamente remota del processo $\{\mathbf{y}(t)\}$.

Proposizione 21 *Le variabili aleatorie invarianti formano un sottospazio (chiuso) di* $L^1(\mathbf{y})$ *che denotiamo col simbolo* $L_\infty(\mathbf{y})$.

Prova: In effetti, per la linearità di U, se $\mathbf{z}_1, \mathbf{z}_2 \in L^1(\mathbf{y})$ sono invarianti lo è anche una qualunque loro combinazione lineare.
Inoltre, data una successione convergente $\{\mathbf{z}_k\}$ di v.a. invarianti (per le quali $U\mathbf{z}_k = \mathbf{z}_k$) segue dalla continuità di U che anche il loro limite è invariante. □

Esempio: $\mathbf{y}(t) = \mathbf{a}\cos(\omega t + \boldsymbol{\theta})$ con $\mathbb{E}|\mathbf{a}| < \infty$, $\mathbf{a}, \boldsymbol{\theta}$ indipendenti e $\boldsymbol{\theta}$ uniformemente distribuita su $[-\pi, \pi]$. Questo processo è stazionario in senso stretto. Lo spazio $L^1(\mathbf{y})$ è costituito da tutte le funzioni $f(\mathbf{a}, \boldsymbol{\theta})$ che hanno aspettazione finita. Dato che

$$U_\tau \mathbf{y}(t) = \mathbf{a}\cos[\omega t + (\boldsymbol{\theta} + \omega\tau)]$$

l'operatore di translazione agisce sul processo semplicemente cambiando la sua fase $\boldsymbol{\theta} \mapsto \boldsymbol{\theta} + \omega\tau$. Quindi per ogni variabile di $L^1(\mathbf{y})$ si ha

$$U_\tau f(\mathbf{a}, \boldsymbol{\theta}) = f[\mathbf{a}, (\boldsymbol{\theta} + \omega\tau)]$$

Le variabili invarianti per il processo sono quindi le statistiche che non dipendono dalla fase; i.e. $\mathbf{z} = f(\mathbf{a})$.

IL CASO GAUSSIANO

Supponiamo che $\mathbf{y}(t) = \mathbf{x}\cos(\omega t) + \mathbf{y}\sin(\omega t)$ con $\mathbf{x}, \mathbf{y}$ entrambi Gaussiane indipendenti di media zero e varianza uguale. Si mostra (vedere ad es. il testo di Papoulis) che le due variabili

$$\mathbf{a} = \sqrt{\mathbf{x}^2 + \mathbf{y}^2}, \qquad \boldsymbol{\theta} = \arctan\frac{\mathbf{y}}{\mathbf{x}}$$

sono indipendenti, la prima ha distribuzione Maxwelliana e la seconda uniforme in $[-\pi, \pi]$ e quindi il processo $\mathbf{y}(t)$ si può anche scrivere nella forma della slide precedente.

È facile mostrare che questo processo è stazionario in senso stretto; basta mostrare che la sua covarianza dipende solo dalla differenza degli argomenti temporali.

IL TEOREMA ERGODICO

Teorema 12 (Teorema Ergodico di Birkhoff) *Sia* $\{\mathbf{y}(t)\}$ *un processo strettamente stazionario. Il limite*

$$\bar{\mathbf{z}} := \lim_{T\to\infty} \frac{1}{T} \sum_{1}^{T} f_t(\mathbf{y}) \tag{77}$$

esiste con probabilità uno per tutte le funzioni $f(\mathbf{y}) \in L^1(\mathbf{y})$ *ed è una variabile aleatoria che è invariante per l'operatore di traslazione del processo* $\{\mathbf{y}(t)\}$. *Se* $f(\mathbf{y}) \in L^2(\mathbf{y})$ *il limite esiste anche in media quadratica.*

Il fatto che il limite sia una variabile invariante scende dalla continutà di U. Infatti

$$U^s\bar{\mathbf{z}} = \lim_{T\to\infty} \frac{1}{T} \sum_{1}^{T} f_{t+s}(\mathbf{y}) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=s+1}^{T+s} f_t(\mathbf{y})$$

e dato che

$$\begin{aligned}\bar{\mathbf{z}} &= \lim_{T\to\infty}\frac{1}{T+s}\sum_{t=1}^{T+s} f_t(\mathbf{y}) \\ &= \lim_{T\to\infty}\left[\frac{1}{T+s}\sum_{t=1}^{s} f_t(\mathbf{y}) + \frac{1}{T}\frac{T}{T+s}\sum_{t=s+1}^{T+s} f_t(\mathbf{y})\right]\end{aligned}$$

e nell'ultimo membro dell'uguaglianza si ha $\frac{T}{T+s}\to 1$ per $T\to\infty$, segue che $\bar{\mathbf{z}} = U^s\bar{\mathbf{z}}$. □

Notiamo adesso che si ha

$$\mathbb{E}\{\frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{y})\} = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\, f_t(\mathbf{y}) = \mathbb{E}\, f(\mathbf{y})$$

dato che $\mathbf{z}(t) = f_t(\mathbf{y})$ è un processo stazionario. Prendiamo ora l'aspettazione dei due mebri nella (77). Dato che si può passare il limite per $T\to\infty$ sotto il segno di aspettazione , si trova

$$\mathbb{E}\,\bar{\mathbf{z}} = \mathbb{E}\, f(\mathbf{y})\,. \tag{78}$$

Diamo allora la seguente definizione.

Definizione 20 *Il processo stazionario* $\mathbf{y}$ *è* **ergodico** *se tutte le sue variabili invarianti sono costanti deterministiche.*

Il corollario seguente viene spesso preso come *definizione di ergodicità*.

Corollario 2 *Se e solo se* $\{\mathbf{y}(t)\}$ *è ergodico si ha*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{1}^{T} f_t(\mathbf{y}) = \mathbb{E}\, f(\mathbf{y}) \tag{79}$$

con probabilità uno, qualunque sia $f(\mathbf{y}) \in L^1(\mathbf{y})$.

Prova: Se il processo è ergodico $\bar{\mathbf{z}}$ è una costante e coincide necessariamente con la sua aspettazione per cui scende da (78) che $\bar{\mathbf{z}} = \mathbb{E}\,\bar{\mathbf{z}} = \mathbb{E}\, f(\mathbf{y})$. Viceversa, è possibile mostrare (ma noi qui non lo faremo) che se vale la (79), ogni variabile invariante è una costante deterministica. □

CLASSI DI PROCESSI ERGODICI

Generalizziamo gli spazi di Hilbert che si introducono nella teoria lineare dei processi a varianza finita. Definiamo i sottospazi popolati dalle funzioni della storia "passata" e "futura" di $\mathbf{y}$ all'istante t,

$$L_t^-(\mathbf{y}) := \{\mathbf{z} \mid \mathbf{z} \in L^1(\mathbf{y}), \mathbb{I} \subset (-\infty, t]\} \quad L_t^+(\mathbf{y}) := \{\mathbf{z} \mid \mathbf{z} \in L^1(\mathbf{y}), \mathbb{I} \subset [t, +\infty)\}$$

Il passato $L_t^-(\mathbf{y})$ e il futuro $L_t^+(\mathbf{y})$ all'istante t sono sottospazi chiusi di $L^1(\mathbf{y})$. È ovvio che

$$L_{t+s}^-(\mathbf{y}) = U_s L_t^-(\mathbf{y}), \qquad t, s \in \mathbb{Z}$$

cresce monotonicamente con t. Analogamente il sottospazio della storia futura si propaga nel tempo in modo stazionario ed è decrescente al crescere di t. *Il passato e il futuro remoto* di $L^1(\mathbf{y})$ sono i sottospazi:

$$L_\infty^-(\mathbf{y}) := \cap_{t \le k} L_t^-(\mathbf{y}) \qquad L_\infty^+(\mathbf{y}) := \cap_{t \ge k} L_t^+(\mathbf{y})$$

La scelta dell'istante iniziale k è irrilevante dato che le successioni di sottospazi in oggetto sono entrambe monotone.

Teorema 13 *Il sottospazio delle variabili aleatorie invarianti è sempre contenute nei sottospazi passato e futuro remoto, ovvero*

$$L_\infty(\mathbf{y}) \subseteq L_\infty^-(\mathbf{y}) \cap L_\infty^+(\mathbf{y})\,. \tag{80}$$

Un processo per cui $L_\infty^-(\mathbf{y})$ e $L_\infty^+(\mathbf{y})$ contengono solo variabili aleatorie costanti (con probabilità uno) si chiama **puramente non deterministico (p.n.d.) in senso stretto**. Questa nozione è molto più stringente di quella di processo p.n.d. (in senso debole) che si riferisce a sottospazi di $L^2(\mathbf{y})$ generati *linearmente* dalle variabili del processo. Segue che

Proposizione 22 *Un processo p.n.d. in senso stretto è ergodico.*

Notiamo qui che lo spazio dei **funzionali lineari** *del processo* $\mathbf{y}$*, che viene denotato col simbolo* $H(\mathbf{y})$*, è un sottospazio molto "sottile" di* $L^2(\mathbf{y}) \subset L^1(\mathbf{y})$ *e che l'operatore di traslazione relativo (che viene normalmente denotato con lo stesso simbolo* U*) si può pensare come la restrizione di* U *al sottospazio* $H(\mathbf{y})$*.*

LEGGE DEI GRANDI NUMERI

D'ora in avanti supponiamo sempre che $\mathbb{E}\,|\mathbf{y}(t)| < \infty$.

Teorema 14 (Kolmogorov) *Per un processo a variabili indipendenti lo spazio* $L_\infty(\mathbf{y})$ *contiene solo (variabili aleatorie) costanti.*

Prova: Dimostriamo che l' affermazione è vera per $L_\infty^+(\mathbf{y})$. (La prova per $L_\infty^-(\mathbf{y})$ è analoga.)

Sia $L_0^n(\mathbf{y})$ il sottospazio di $L^1(\mathbf{y})$ contenente tutte le funzioni delle variabili $\{\mathbf{y}(0), \mathbf{y}(1), \ldots, \mathbf{y}(n)\}$ che hanno aspettazione finita. Ovviamente ogni variabile aleatoria $\mathbf{x}$ in $L_\infty^+(\mathbf{y})$ deve essere indipendente dalle variabili in $L_0^n(\mathbf{y})$, qualunque sia n; i.e.

$$\mathbb{E}\,\mathbf{x}\mathbf{z}^n = \mathbb{E}\,\mathbf{x}\,\mathbb{E}\,\mathbf{z}^n\,, \qquad \mathbf{z}^n \in L_0^n(\mathbf{y})$$

e quindi $\mathbf{x}$ dovrà essere indipendente da ogni variabile $\mathbf{z}$ di $L_0^+(\mathbf{y})$ perchè quest'ultimo è la chiusura (in L^1) dello spazio vettoriale generato da tutte

le variabili della famiglia $\{L_0^n(\mathbf{y}); n \geq 0\}$ e ogni $\mathbf{z} \in L_0^+(\mathbf{y})$, o appartiene a un qualche $L_0^n(\mathbf{y})$, o è il limite (in L^1) di qualche sequenza $\{\mathbf{z}^n\}$. Insomma, se $\mathbf{x} \in L_\infty^+(\mathbf{y})$,

$$\mathbb{E}\,\mathbf{xz} = \mathbb{E}\,\mathbf{x}\,\mathbb{E}\,\mathbf{z}, \qquad \forall \mathbf{z} \in L_0^+(\mathbf{y})$$

D'altro canto $L_\infty^+(\mathbf{y})$ è contenuto in $L_0^+(\mathbf{y})$ e quindi ogni variabile $\mathbf{x}$ è indipendente da sè stessa

$$\mathbb{E}\,(\mathbf{x})^2 = \mathbb{E}\,\mathbf{x}\,\mathbb{E}\,\mathbf{x}, \qquad \forall \mathbf{x} \in L_\infty^+(\mathbf{y})$$

il che può essere vero solo se $\mathbf{x}$ è una costante deterministica. □

Notiamo che questo risultato (noto come *legge dello 0-1* di Kolmogorov) non richiede la stazionarietà. La dimostrazione “classica” si svolge ragionando sulla σ-algebra degli eventi ”infinitamente futuri” o ”infinitamente remoti” del processo (si mostra che questi eventi possono avere solo probabilità zero o uno).

I processi ergodici debbono essere molto ”irregolari”. Un classico esempio di processo ergodico è un processo $\{\mathbf{y}(t)\}$ a variabili i.i.d. (rumore bianco in senso stretto).

Corollario 3 *Ogni processo i.i.d. è ergodico.*

Proposizione 23 *Un processo ergodico non può ammettere limite per* $t \to \pm\infty$ *a meno che tutte le traiettorie del processo coincidano (con probabilità 1).*

In effetti il limite, diciamolo $\mathbf{z}$, sarebbe una v.c. costante per l'ergodicità e quindi distribuita in modo degenere (come la funzione δ di Dirac). Qualunque sia il tipo di convergenza (in probabilità, in media o quasi ovunque) secondo la quale $\mathbf{y}(t) \to \mathbf{z}$, ne seguirebbe necessariamente che le distribuzioni delle variabili $\mathbf{y}(t)$ tendono a quella di $\mathbf{z}$. Ma le $\{\mathbf{y}(t)\}$ hanno, per ogni t, la stessa distribuzione e questa può "convergere" alla distribuzione δ solo se essa stessa è degenere.

Proposizione 24 *Il processo* $\{\mathbf{z}(t)\}$ *ottenuto per traslazione di una arbitraria funzione* $\mathbf{z} = f(\mathbf{y}) \in L^1(\mathbf{y})$ *di un processo ergodico è ancora ergodico.*

Prova: Di fatto il sottospazio delle variabili invarianti $L_\infty(\mathbf{z})$ è contenuto in $L_\infty(\mathbf{y})$ e quindi se quest'ultimo è triviale lo è anche il primo. □

Questo vale in particolare se $\mathbf{y}$ è un *processo i.i.d.*, che ha le variabili indipendenti e identicamente distribuite. Da questa considerazione si può individuare una classe di processi ergodici che torna utile nelle applicazioni all'identificazione.

Teorema 15 (Doob) *Sia* $\{\mathbf{e}(t)\}$ *un processo i.i.d. a varianza finita. Si assuma che la successione di numeri reali* $\{c_k\}$ *sia a quadrato sommabile,*

$$\sum_{-\infty}^{+\infty} c_k^2 < \infty \quad ; \tag{81}$$

allora il processo $\{\mathbf{y}(t)\}$ *definito dalla*

$$\mathbf{y}(t) := \sum_{-\infty}^{+\infty} c_k \, \mathbf{e}(t+k) \tag{82}$$

è (strettamente stazionario), a varianza finita ed ergodico.

La condizione (81) serve a garantire la convergenza della somma. Notiamo che definendo $\bar{c}_k := c_{-k}$ la (82) si riscrive

$$\mathbf{y}(t) := \sum_{-\infty}^{+\infty} \bar{c}_k \, \mathbf{e}(t-k)$$

che ha l'usuale aspetto di somma di convoluzione. In sostanza, **l'uscita di un filtro lineare (non necessariamente causale) ℓ^2-stabile con ingresso rumore bianco (in senso stretto) è un processo ergodico.**

ERGODICITÀ E INFERENZA STATISTICA

Studieremo una questione che è intimamente legata al problema generale dell'inferenza statistica. In termini generali, il problema è il seguente:

Problema 6 *Supponiamo che sia disponibile una serie infinita di dati* $\{\bar{y}(t) \mid t \in \mathbb{Z}\}$ *che penseremo essere una traiettoria di un processo stocastico* $\mathbf{y}$*. Supponiamo cioè che* $\{\bar{y}(t)\}_{t\in\mathbb{Z}} = \{\mathbf{y}(t,\bar{\omega})\}_{t\in\mathbb{Z}}$ *per qualche* $\bar{\omega} \in \Omega$*. Vogliamo rispondere alla seguente domanda: Che cosa si può dire della legge di probabilità* P *del processo in base alla conoscenza della sola traiettoria* $\{\bar{y}(t)\}$ *?*

Ricordiamo che si chiama *legge di probabilità del processo* la famiglia infinita di distribuzioni di probabiltà

$$F_n(x_1,x_2,\ldots,x_n;t_1,t_2,\ldots,t_n) = P\{\mathbf{y}(t_1) \le x_1, \mathbf{y}(t_2) \le x_2, \ldots, \mathbf{y}(t_n) \le x_n\}$$

dove le $\{x_k\}$ sono variabili reali se il processo è scalare ovvero variabili vettoriali in $\mathbb{R}^m$ in caso contrario.

Facciamo vedere che esiste una classe di processi per cui questo problema ha una soluzione ben definita, che in un certo senso è sorprendentemente positiva. Questi processi sono per l'appunto i processi *ergodici*.

Sia E un qualunque sottoinsieme di Borel di $\mathbb{R}$ (ad esempio un intervallo) e consideriamo la funzione

$$f(\mathbf{y}) := I_E(\mathbf{y}(0)) \quad ,$$

dove I_E è la funzione indicatrice dell'insieme E. La variabile casuale

$$\boldsymbol{\nu}_T(E) := \frac{1}{T+1} \sum_{t=0}^{T} I_E(\mathbf{y}(t))$$

è la frequenza relativa con cui il processo $\{\mathbf{y}(t)\}$ "visita" l'insieme E. Se definiamo $\mathbf{z} := I_E(\mathbf{y}(0))$ e supponiamo che il processo $\{\mathbf{y}(t)\}$ sia ergodico, per il teorema di Birkhoff il limite

$$\lim_{T\to\infty} \boldsymbol{\nu}_T(E)$$

esiste con probabilità 1 (ovvero *per tutte le possibili traiettorie del processo, eccettuato al più un insieme di traiettorie di probabilità zero*) e vale:

$$E\, I_E(\mathbf{y}(0)) = \int_E d\, F(y) = P(E) \quad ,$$

ovvero è uguale proprio alla probabilità che $\mathbf{y}(t) \in E$ (che ovviamente non dipenda da t). Se si prende $E = (-\infty, a]$, la quantità $\boldsymbol{\nu}_T((-\infty, a])$ che, si badi bene, *è calcolata osservando una sola traiettoria del processo*, è, al crescere di T, una approssimazione sempre più accurata del (e al limite è esattamente uguale al) valore della funzione distribuzione di probabilità di $\{\mathbf{y}(t)\}$ nel punto a.

Se si considerano ora due insiemi E_1, E_2 e si definisce

$$f(\mathbf{y}) = I_{E_1}(\mathbf{y}(0))\, I_{E_2}(\mathbf{y}(k)) \quad ,$$

la variabile casuale

$$\boldsymbol{\nu}_T(E_1, E_2, k) := \frac{1}{T+1} \sum_{t=0}^{T} I_{E_1}(\mathbf{y}(t))\, I_{E_2}(\mathbf{y}(t+k))$$

ha ancora il significato di frequenza relativa con cui una traiettoria del processo visita prima l'insieme E_1 e k istanti dopo l'insieme E_2. Se $\{\mathbf{y}(t)\}$ è

ergodico, si ha allora:

$$\begin{aligned}\lim_{T\to\infty} \boldsymbol{\nu}_T(E_1,E_2,k) &= E\left[I_{E_1}(\mathbf{y}(0))\, I_{E_2}(\mathbf{y}(k))\right] \\ &= \int_{E_1}\int_{E_2} dF(y_1,y_2\,;k) = P\Big\{\mathbf{y}(t)\in E_1\,;\,\mathbf{y}(t+k)\in E_2\Big\}\end{aligned}$$

per “quasi tutte” le traiettorie del processo.

Una generalizzazione ormai facile porge allora la seguente conclusione.

Teorema 16 *Se il processo $\{\mathbf{y}(t)\}$ è ergodico, la conoscenza di una sola traiettoria è (con probabilità 1) sufficiente a determinare univocamente la legge di probabilità dell’intero processo.*

Problema: Sia $\{\mathbf{y}(t)\}$ un processo i.i.d. di media μ e varianza σ^2, che si può supporre nota. Si vuole stimare μ usando la media campionaria

$$\hat{\boldsymbol{\mu}}_N = \frac{1}{N}\sum_{k=1}^{N} \mathbf{y}(k)$$

Quanto grande dev'essere N per avere $|\hat{\boldsymbol{\mu}}_N - \mu| \leq 10^{-3}\sigma$ con probabilità del 97.5 %?

Il problema si può risolvere con la disuguaglianza di Chebychev:

$$P\{|\hat{\boldsymbol{\mu}}_N - \mu| \geq 10^{-3}\sigma\} \leq \frac{\mathrm{var}\{\hat{\boldsymbol{\mu}}_N\}}{10^{-6}\sigma^2} = \frac{1}{10^{-6}N}$$

per cui

$$P\{|\hat{\boldsymbol{\mu}}_N - \mu| < 10^{-3}\sigma\} = 1 - P\{|\hat{\boldsymbol{\mu}}_N - \mu| \geq 10^{-3}\sigma\} \geq 1 - \frac{1}{10^{-6}N}$$

Per cui si ha $|\hat{\boldsymbol{\mu}}_N - \mu| < 10^{-3}\sigma$ con probabilità almeno del 97.5 % se $\frac{1}{10^{-6}N} < 0.025$, ovvero $N > 40 \times 10^6$.

IL METODO DI MONTECARLO

Come seconda applicazione del teorema ergodico menzioneremo qui una tecnica di simulazione particolarmente usata in statistica: il cosiddetto *metodo di Montecarlo*.

L'essenza del metodo è una tecnica per calcolare "sperimentalmente" dei valori attesi, tipicamente medie e varianze di stimatori, impossibili da calcolare esplicitamente, usando il teorema ergodico. Più in generale, questa tecnica consente di approssimare integrali arbitrari del tipo

$$\int_I f(x)\,dx \quad ,$$

dove I è un intervallo finito o infinito. L'integrale può sempre essere scritto come l'aspettazione di una opportuna funzione di variabile casuale, trasformando la misura dx rispetto alla quale si deve fare l'integrazione in una misura di probabilità. Se $I = [a,b]$ è un intervallo finito, basta porre

$$dF(x) := \frac{1}{b-a}\,dx$$

e riscalare opportunamente f. Se I è un intervallo infinito, si può introdurre un'opportuna densità fittizia (ad esempio ponendo $dF(x) = e^{-x^2/2}dx$), che andrà poi "scalata" dalla funzione f.

Ci si riduce quindi al calcolo della media, $\mathbb{E}\, f(\mathbf{y})$, di una funzione (nota) della variabile casuale $\mathbf{y}$ che ha distribuzione di probabilità $F(x)$.

Supponiamo di disporre di un *generatore di numeri (pseudo)-casuali*, cioè di un algoritmo che fornisce successioni numeriche, $\{z_1, z_2, \ldots\}$ assimilabili ad una serie di misure *indipendenti* ed *ugualmente distribuite* (usualmente in modo uniforme nell'intervallo $[0, 1]$). Il generatore fornisce successioni assimilabili alle traiettorie di un processo i.i.d. $\{\mathbf{z}(t)\}$, in cui $\mathbf{z}(t)$ ha distribuzione uniforme nell'intervallo $[0, 1]$.

Trasformando ciascun dato z_k secondo la relazione

$$y_k := F^{-1}(z_k)$$

si ottiene allora una successione $\{y_k\}$ che si può pensare generata dal processo i.i.d. $\{\mathbf{y}(t)\}$, nel quale la variabile generica $\mathbf{y}(t)$ è distribuita secondo la $F(x)$.

$$P(\mathbf{y} \leq x) = P(F^{-1}(z) \leq x) = P(z \leq F(x)) = F(x)$$

, se $\mathbf{z}$ è uniformemente distribuita.

Il teorema ergodico porge quindi:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{1}^{N} f(y_k) = E\, f(\mathbf{y}) = \int_I f(x)\, dF(x)$$

e questa formula fornisce un metodo generale per il calcolo approssimato di aspettazioni o di integrali definiti. Quanto rapida è la convergenza? Usando Chebycheff

$$P\{|\frac{1}{N} \sum_{1}^{N} f(\mathbf{y}_k) - E\, f(\mathbf{y})| > \varepsilon\} \leq \frac{\mathrm{var}\{\frac{1}{N}\sum_{1}^{N} f(\mathbf{y}_k)\}}{\varepsilon^2}$$

dato che le v.a. $f(\mathbf{y}_k)$ sono indipendenti

$$\mathrm{var}\{\frac{1}{N}\sum_{1}^{N} f(\mathbf{y}_k)\} = \frac{1}{N}\mathrm{var}\{f(\mathbf{y}_k)\} := \frac{1}{N}\sigma_f^2 \,.$$

Usando il TLC si possono anche dare intervalli di confidenza in funzione di N.

Spesso è impossibile calcolare esplicitamente l'inversa della distribuzione di probabilità. In questi casi si può ricorrere ad una densità di probabiltà ausiliaria scelta in modo opportuno. Se $p(x)$ è una densità complicata per cui il metodo descritto prima è difficile da applicare, per calcolare l'integrale

$$\int_{\mathbf{I}} f(x)p(x)dx$$

si può costruire una opportuna densità ausiliaria $q(x)$ con supporto l'intervallo $\mathbf{I}$ ed effettuare il "cambio di misura" descritto dalla formula

$$\int_{\mathbf{I}} f(x)p(x)dx = \int_{\mathbf{I}} f(x)\frac{p(x)}{q(x)}q(x)dx := \int_{\mathbf{I}} g(x)q(x)dx$$

in cui formalmente appare una funzione da integrare diversa ma una densità di probabilità q facile da simulare. Si può così, cercare di scegliere q in modo tale che i punti in cui g è grande (e contribuisce di più all'integrale) abbiano probabilità più alta e vengano quindi generati più "spesso" nella simulazione in modo da rendere il processo più efficiente. Tecniche di questo genere si chiamano di *importance sampling*. Naturalmente queste tecniche funzionano nel caso di variabili scalari; il campionamento di una densità multivariata è una questione che va affrontata con idee diverse.

CATENE DI MARKOV ERGODICHE

Sia $\{\mathbf{y}(t)\}$ una catena di Markov finita con matrice di transizione M. Sia $\pi = M\pi$ una distribuzione invariante per M e supponiamo che $\mathbf{y}(0)$ sia distribuita secondo la π. È possibile allora mostrare che $\{\mathbf{y}(t)\}$ è un processo strettamente stazionario.
Dato che una distribuzione invariante assegna probabilità zero agli stati transitori, possiamo senz'altro supporre che la catena (non abbia stati transitori e) consista di N classi ergodiche $A_1, \ldots, A_N$, dove gli insiemi A_i costituiscono una partizione dello spazio di stato dal processo che qui identificheremo con l'insieme $\{1, \ldots, n\}$ dei primi n numeri naturali.

Consideriamo variabili casuali aventi la seguente struttura:

$$\mathbf{z} = f(\mathbf{y}(t)) = c_i \qquad \text{se} \qquad \mathbf{y}(t) \in A_i \quad , \quad i = 1, \ldots, N \quad ,$$

ovvero,

$$\mathbf{z} = \sum_{1}^{N} c_i \, I_{A_i}(\mathbf{y}(t)) \quad ,$$

dove c_i, $i = 1, \ldots, N$, sono numeri reali arbitrari e $I_{A_i}(\mathbf{y})$ è la funzione indicatrice dell'insieme A_i.

È facile constatare che $\mathbf{z}$ è una variabile invariante. Infatti, se $\mathbf{y}(t, \omega) \in A_i$ per qualche t, allora $\mathbf{y}(t, \omega) \in A_i$ per ogni $t \in \mathbb{Z}_+$ e $I_{A_i}(\mathbf{y}(t)) = I_{A_i}(\mathbf{y}(\tau))$, $\forall t, \tau$. Evidentemente, se e solo se $N = 1$, $\mathbf{z}$ si riduce ad una costante deterministica.

Si può mostrare che tutte le variabili invarianti per la catena hanno questa struttura.

MARKOV CHAIN MONTE CARLO

L'idea di base è una tecnica di campionamento che si applica a distribuzoni multivariabili, usando la convergenza di catene di Markov (a tempo discreto) alla distribuzione invariante. Come abbiamo visto, una catena di Markov con un solo insieme di stati ergodici è un processo ergodico per cui vale il teorema di Birkhoff.

Ammettendo di voler calcolare l'aspettazione di una funzione del tipo $\int f(x)\pi(x)dx$ si può generare (i.e. simulare) una catena di Markov ergodica che abbia come distribuzione invariante (necessariamente unica) proprio la $\pi(x)$. Per far questo occorre saper costruire una matrice (o, più in generale, un nucleo di probabilità) di transizione che ammetta proprio $\pi(x)$ come misura invariante. Si dimostra che ci sono in realtà infinite catene ergodiche che hanno π come probabilità invariante e l'articolo [Hastings-1970] descrive un possibile metodo di costruirne una. Fatto questo, si tratta di simulare la catena generando successivamente le variabili di una traiettorie $\{x(t)\,;t=1,2,\ldots\}$. Quando $\mathbf{x}(t)$ è arrivato a convergere al processo

stazionario distribuito secondo $\pi(x)$ si può usare il teorema ergodico nel modo usuale.

Concludiamo questa brevissima carrellata menzionando appena la gran mole di lavoro di ricerca che si sta portando avanti in questi anni su questi metodi che stanno diventando, grazie ai progressi dei sitemi di calcolo moderni, i metodi d'elezione per risolvere problemi, come ad esempio il filtraggio non lineare, che solo un decennio fa sembravano innavvicinabili.

Notiamo per ultimo che la qualificazione "con probabilità 1" che va associata alla formula (79) è molto più di effetto psicologico che reale.

ERGODICITÀ DEL SECONDO ORDINE

L'ergodicità e la stazionarietà stretta su cui si basa sono ipotesi molto forti e praticamente impossibili da verificare. C'è una nozione di *ergodicità del second'ordine* o *debole* che è sufficiente per studiare i casi che si presentano nell'analisi asintotica degli algoritmi di identificazione di sistemi lineari.

Definizione 21 *Sia* $\{\mathbf{y}(t)\}$ *un processo* m*-dimensionale stazionario in senso debole di media* μ *e matrice di covarianza* $\Sigma(\tau)$*. Il processo si dice* ergodico del secondo ordine *(o* debolmente ergodico*) se la media campionaria*

$$\bar{\mathbf{y}}_T := \frac{1}{T+1} \sum_{0}^{T} \mathbf{y}(t)$$

e la varianza campionaria del processo

$$\hat{\mathbf{\Sigma}}_T(\tau) := \frac{1}{T+1} \sum_{t=0}^{T} [\mathbf{y}(t+\tau) - \bar{\mathbf{y}}_T] [\mathbf{y}(t) - \bar{\mathbf{y}}_T]^\top$$

convergono con probabilità uno a delle costanti.

Proposizione 25 *Le costanti della definizione debbono essere i valori veri, ovvero:*

$$\begin{aligned} \lim_{T\to\infty} \bar{\mathbf{y}}_T &= \mu \\ \lim_{T\to\infty} \hat{\mathbf{\Sigma}}_T(\tau) &= \Sigma(\tau)\,. \end{aligned}$$

con probabilità uno.

Prova: Già dimostrato per la media. La proposizione si dimostra facendo vedere che $\hat{\mathbf{\Sigma}}_T(\tau)$ e $1/T+1\sum_0^T [\mathbf{y}(t+\tau)-\mu]\,[\mathbf{y}(t)-\mu]^\top$ hanno lo stesso limite (costante), dato che l'aspettazione di quest'ultimo termine è proprio

$\Sigma(\tau)$. Di fatto

$$
\begin{aligned}
(T+1)\,\hat{\mathbf{\Sigma}}_T(\tau) &= \sum_0^T \left[\mathbf{y}(t+\tau)-\mu+\mu-\bar{\mathbf{y}}_T\right]\left[\mathbf{y}(t)-\mu+\mu-\bar{\mathbf{y}}_T\right]^\top \\
&= \sum_0^T \left[\mathbf{y}(t+\tau)-\mu\right]\left[\mathbf{y}(t)-\mu\right]^\top - (\bar{\mathbf{y}}_T-\mu)\sum_0^T \left[\mathbf{y}(t)-\mu\right]^\top \\
&- \left(\sum_0^T \left[\mathbf{y}(t+\tau)-\mu\right]\right)(\bar{\mathbf{y}}-\mu)^\top + (T+1)\,(\bar{\mathbf{y}}_T-\mu)\,(\bar{\mathbf{y}}_T-\mu)^\top,
\end{aligned}
$$

gli ultimi tre termini divisi per $T+1$ tendono a zero con probabilità uno per $T \to \infty$. □

Corollario 4 *Se $f(\mathbf{y})$ è una funzione lineare di un processo ergodico del second'ordine $\mathbf{y}$ allora*

$$
\lim_{T\to\infty} \frac{1}{T+1}\sum_{t=0}^{T} f_t(\mathbf{y})^2 = \mathbb{E}\, f(\mathbf{y})^2 .
$$

Nota: Qualche volta il limite superiore nella sommatoria è posto uguale a $T - \tau$. Si riconosce facilmente che le due relazioni sono equivalenti.
Un processo può essere ergodico del second'ordine sotto condizioni più deboli dell'ergodicità. Condizioni utili per l'ergodicità del secondo ordine: processi generati come uscita di sistemi lineari eccitati da rumore bianco.

Ricordiamo che **l'uscita $\mathbf{y}$ di un sistema lineare tempo-invariante ℓ^2-stabile (i.e. la cui risposta impulsiva è a quadrato sommabile) che ha in ingresso un processo i.i.d. a varianza finita, è un processo ergodico**. (Ovviamente il processo di uscita, essendo ergodico in senso stretto, è in particolare anche ergodico in senso debole.)

La condizione che il processo di ingresso sia i.i.d. può essere rilassata. Hannan ha dimostrato che l'ergodicità del secondo ordine vale per processi generati come uscita di un filtro lineare stabile che ha in ingresso un processo (debolmente stazionario) $\{\mathbf{e}(t)\}$ che ha momenti fino al quart'ordine invarianti per traslazione e soddisfa alle condizioni,

$$\mathbb{E}\,\|\mathbf{e}(t)\|^4 < \infty \tag{83}$$

$$\mathbb{E}\,[\mathbf{e}(t) \mid \mathbf{e}^{t-1}] = 0 \qquad t \in \mathbb{Z} \tag{84}$$

$$\mathrm{Var}\,[\mathbf{e}(t) \mid \mathbf{e}^{t-1}] = \Lambda < \infty \tag{85}$$

dove $\mathbf{e}^{t-1}$ è la sequenza inifinita $\{\mathbf{e}(t-1), \mathbf{e}(t-2), \ldots,\}$. Queste condizioni sono in genere più deboli della condizione di essere i.i.d..

Processi di questo tipo si chiamano *d-martingale* e veranno studiati più avanti.

Per processi Gaussiani ergodicità forte e debole coincidono:

Teorema 17 *Per un processo Gaussiano stazionario m-dimensionale $\{\mathbf{y}(t)\}$, le seguenti condizioni sono fra loro equivalenti:*

1. *Il processo è ergodico.*

2. *Il processo è ergodico del secondo ordine.*

3. *La matrice distribuzione spettrale di potenza del processo, $F(e^{i\omega})$, è una funzione continua di ω in $[-\pi, \pi]$.*

4. *Si ha incorrelazione asintotica delle variabili del processo, nel senso delle medie di Cesàro,*

$$\lim_{T\to\infty} \frac{1}{T+1} \sum_{0}^{T} \sigma_{ii}^2(\tau) = 0 \qquad i = 1, 2, \ldots, m,$$

dove le $\sigma_{ii}^2(\tau)$ sono le funzioni covarianza delle componenti $\mathbf{y}_i(t)$.

Di queste condizioni la più utile è probabilmente la (3), che dice che un processo Gaussiano è ergodico se e solo se il suo spettro non ha righe, ovvero, se e solo se il processo non ha *componenti oscillatorie di ampiezza finita*.

Infatti, le uniche discontinuità della funzione distribuzione spettrale (che è monotona e limitata) possono essere salti di ampiezza finita. Scrivendo per semplicità $F(\omega)$ al posto di $F(e^{i\omega})$ e supponendo che la F abbia una discontinuità in ω_0, si ha:

$$F(\omega_0+) - F(\omega_0) := \Delta F(\omega_0) \neq 0$$

($F(\omega_0+)$ sta per il limite destro di F in ω_0), per cui il processo avrebbe potenza finita associata alla frequenza ω_0 (una "riga" spettrale per $\omega = \omega_0$).

Problema 7 *Si consideri il processo $\mathbf{z}(t) = \mathbf{x}\cos\omega_0 t + \mathbf{y}\sin\omega_0 t$, $t \in \mathbb{Z}$, dove $\mathbf{x}$ e $\mathbf{y}$ sono variabili aleatorie scalari Gaussiane di media zero e uguale varianza σ^2, fra loro scorrelate.*

– Mostrare che $\{\mathbf{z}(t)\}$ è stazionario di covarianza $\sigma(\tau) = \sigma^2 \cos\omega_0\tau$ e media zero.
– Se $\bar{x}, \bar{y}$ sono valori campionari di $\mathbf{x}$ e $\mathbf{y}$, si calcoli il limite

$$\lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \bar{z}(t+\tau)\, \bar{z}(t) \quad ,$$

con τ fissato, e si verifichi che è diverso da $E[\mathbf{z}(t+\tau)\,\mathbf{z}(t)]$. □

Una semplice condizione sufficiente per la validità della (4) è l'incorrelazione asintotica

$$\lim_{\tau\to\infty} \sigma_{ii}(\tau) = 0 \quad , \qquad i = 1, 2, \ldots, m \quad , \tag{86}$$

nel qual caso si riconosce facilmente che la componente i-sima del processo, $\{\mathbf{y}_i(t)\}$, non può contenere componenti oscillatorie di ampiezza finita. Questo è, ovviamente, in accordo con la condizione (3).
Notiamo, di passaggio, che la (86) è equivalente alla

$$\lim_{\tau\to\infty} \Sigma(\tau) = 0\,.$$

Segnaliamo infine che

Corollario 5 *Un processo Gaussiano stazionario e p.n.d in senso debole, lo è anche in senso forte e quindi è in particolare ergodico.*

ANALISI ASINTOTICA DELLO STIMATORE PEM

Primo risultato fondamentale della teoria asintotica della stima PEM.

Teorema 18 *Si assuma che*
◦ *I dati* (y^N, u^N) *sono generati da un processo stazionario, ergodico del secondo ordine;*
◦ *La cifra di merito è l'errore quadratico medio di predizione;*
Allora

$$\lim_{N\to\infty} V_N(\theta) = \mathbb{E}_0\, \boldsymbol{\varepsilon}_\theta(t)^2$$

dove $\mathbb{E}_0$ *denota aspettazione rispetto alla distribuzione del processo vero che ha generato i dati.*

Prova Per il teorema ergodico

$$\lim_{N\to\infty} V_N(\theta) = \lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} \varepsilon_\theta(t)^2$$

dove $\varepsilon_\theta(t)^2$ è funzione quadratica dei dati. □

Teorema 19 *Si assuma che*

◦ *Il predittore* $\hat{\mathbf{y}}_\theta(t \mid t-1)$ *basato sul modello* $M(\theta)$ *sia una funzione regolare del parametro (continua e differenziabile);*

◦ $\forall N$ *esiste (almeno) un minimo globale,* $\hat{\theta}_N$*, di* $V_N(\theta)$ *e per* $N \to \infty$ *si possano scambiare l'operazione di limite e la minimizzazione.*

Allora i minimizzatori, $\hat{\boldsymbol{\theta}}_N$*, di* $V_N(\theta)$*, convergono tutti, con probabilità uno, all'insieme dei punti di minimo di* $\bar{V}(\theta) := \mathbb{E}_0 \boldsymbol{\varepsilon}_{\boldsymbol{\theta}}(t)^2$.

In altri termini, detto $\Delta \subset \Theta_0$ l'insieme dei punti di minimo di $\bar{V}(\theta)$, si ha

$$\lim_{N\to\infty} \hat{\boldsymbol{\theta}}_N \in \Delta$$

con probabiltà uno.

Cenno di Prova : La convergenza dei minimi si può dimostrare usando il fatto che le variabili aleatorie della famiglia $\{\boldsymbol{\varepsilon}_{\boldsymbol{\theta}}(t)^2; t \geq 1\}$ sono tutte non negative e quindi uniformemente limitate inferiormente (dalla variabile aleatoria zero) e si può quindi applicare un teorema di L. Le Cam che permette di commutare l'operazione di minimizzazione di $V_N(\theta)$ su un arbitrario insieme compatto con quella di passaggio al limite. □

L' ipotesi che, per quasi tutte le possibili sequenze di dati di misura, esista un minimo (finito!), $\hat{\boldsymbol{\theta}}_N$, non è limitativa e si può garantire se la dipendenza dell'errore di predizione dal parametro è di tipo polinomiale o razionale, eventualmente restringendo la parametrizzazione in modo opportuno. Il fatto che i punti di minimo siano (almeno per N grande) contenuti con probabilità uno in un insieme compatto, si può garantire imponendo l'identificabiltà del modello.

La convergenza continua a valere (pur di sopprimere la qualificazione "con probabilità uno") se invece dell'ergodiocità del secondo ordine si assume semplicemente stazionarietà del secondo ordine.
In questo caso potrebbero essere presenti delle componenti armoniche nei segnali in gioco e il processo "vero" corrispondente non sarebbe ergodico del secondo ordine. Questa circostanza si verifica spesso quando l'ingresso è imposto artificialmente nell'esperimento di identificazione ed è composto da una somma di sinusoidi.

ANALISI ASINTOTICA DELLO STIMATORE PEM

Assumiamo che esista un **processo vero** $\mathbf{y}(t)$ stazionario (in senso debole), con momenti del second'ordine finiti e *puramente non deterministico*.
Allora $\mathbf{y}(t)$ può essere decomposto nella somma di un predittore lineare a minima varianza d'errore $\hat{\mathbf{y}}_0(t \mid t-1)$, (la proiezione ortogonale di $\mathbf{y}(t)$ sullo spazio di Hilbert generato linearmente dalla storia passata congiunta $(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})$), e dell'errore di predizione di un passo (i.e. l'innovazione)

$$\mathbf{y}(t) = \hat{\mathbf{y}}_0(t \mid t-1) + \mathbf{e}_0(t)\,. \tag{87}$$

Il termine tra parentesi quadre nella decomposizione,

$$\boldsymbol{\varepsilon}_\theta(t) = \mathbf{e}_0(t) + [\hat{\mathbf{y}}_0(t \mid t-1) - \hat{\mathbf{y}}_\theta(t \mid t-1)]$$

è funzione dei dati $(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})$ e quindi ortogonale ad $\mathbf{e}_0(t)$, per cui

$$\begin{aligned} \bar{V}(\theta) &= \operatorname{var}\{\mathbf{e}_0(t)\} + \|\hat{\mathbf{y}}_0(t \mid t-1) - \hat{\mathbf{y}}_\theta(t \mid t-1)\|^2 \\ &= \lambda_0^2 + \|\hat{\mathbf{y}}_0(t \mid t-1) - \hat{\mathbf{y}}_\theta(t \mid t-1)\|^2 \end{aligned}$$

la norma è la norma nello spazio di Hilbert generato linearmente da $(\mathbf{y}, \mathbf{u})$ (i.e. la varianza).

QUINDI: lo stimatore PEM minimizza asintoticamente la distanza tra il predittore lineare "vero" e quello costruito sul modello $M(\theta)$.

Osservazioni:

Senza condizioni ulteriori sul nostro problema, i predittori (in realtà i modelli) a distanza minima da $\hat{\mathbf{y}}_0(t \mid t-1)$ possono essere molti (anche infiniti).

L' interpretazione di $\bar{V}(\theta)$ come distanza L^2 tra predittori è basata in modo cruciale sul fatto che i predittori che si costruiscono, sono predittori (lineari) a minima varianza d'errore ($\mathbf{e}_0(t)$ è scorrelata da $\hat{\mathbf{y}}_\theta(t \mid t-1)$). Su questo fatto è in effetti basato anche il fondamentale risultato seguente*.

*Quindi il predittore non può essere *arbitrario* come Ljung e qualche suo ottimista seguace insiste nel propagandare.

Teorema 20 *Supponiamo che valgano le ipotesi dei teoremi precedenti e che il processo (vero) che genera i dati sia descritto da un modello che appartiene alla stessa classe parametrica* $\mathcal{M}$ *dei modelli scelti per l'identificazione, ovvero esista* $\theta_0 \in \Theta$ *tale che*

$$\mathbf{y} \sim M(\theta_0) \in \mathcal{M}$$

e l'errore di predizione $\boldsymbol{\varepsilon}_\theta(t)$ *sia calcolato mediante il predittore lineare a minima varianza* $\hat{\mathbf{y}}_\theta(t \mid t-1)$.

Allora, se la classe parametrica dei modelli $\mathcal{M} \equiv \{M(\theta);\, \theta \in \Theta\}$ *è identificabile localmente in* $\theta = \theta_0$*, si ha*[*]

$$\lim_{N\to\infty} \hat{\boldsymbol{\theta}}_N = \theta_0 \tag{88}$$

con probabilità uno. In altri termini, lo stimatore PEM è consistente.

[*]Nel caso di assenza di reazione si può precisare la condizione richiedendo che il modello sia identificabile a priori in $\theta = \theta_0$ e che il processo $\mathbf{u}$ sia sufficientemente eccitante da garantire l' identificabilità.

Prova Se $M(\theta_0) \in \mathcal{M}$, si vede subito che il minimo assoluto di $\bar{V}(\theta)$, che vale λ_0^2, si può raggiungere per $\theta = \theta_0$. Quindi $\theta_0 \in \Delta$. Si tratta di dimostrare che sotto le ipotesi di identificabilità l'insieme dei punti di minimo si riduce al solo $\{\theta_0\}$.

Lemma 3 *Se c'è identificabilità*

$$\|\hat{\mathbf{y}}_{\theta_0}(t \mid t-1) - \hat{\mathbf{y}}_{\theta}(t \mid t-1)\| = 0 \;\Leftrightarrow\; \theta = \theta_0\,; \tag{89}$$

i.e. c'è un solo punto di minimo assoluto di $\bar{V}(\theta)$ *e* $\Delta = \{\theta_0\}$.

Prova: Dato che il predittore di Wiener è una funzione lineare dei dati si può scrivere simbolicamente

$$\hat{\mathbf{y}}_{\theta}(t \mid t-1) = L_{\theta}(z)\mathbf{y}(t-1) + M_{\theta}(z)\mathbf{u}(t-1)$$

dove $L_{\theta}(z) = G_{\theta}^{-1}(z)G_{1,\theta}(z)$ e $M_{\theta}(z) = G_{\theta}^{-1}(z)F_{1,\theta}(z)$ sono funzioni razionali strettamente stabili.

Ne segue che il quadrato della norma in (89) si può esprimere nel dominio della frequenza con l'integrale,

$$\|\hat{\mathbf{y}}_{\theta_0}(t \mid t-1) - \hat{\mathbf{y}}_{\theta}(t \mid t-1)\|^2 = \int_{-\pi}^{\pi} \left[L_{\theta_0}(e^{j\omega}) - L_{\theta}(e^{j\omega})\; M_{\theta_0}(e^{j\omega}) - M_{\theta}(e^{j\omega}) \right] \cdot$$
$$\cdot \begin{bmatrix} S_{\mathbf{y}}(e^{j\omega}) & S_{\mathbf{yu}}(e^{j\omega}) \\ S_{\mathbf{uy}}(e^{j\omega}) & S_{\mathbf{u}}(e^{j\omega}) \end{bmatrix} \begin{bmatrix} L_{\theta_0}(e^{-j\omega}) - L_{\theta}(e^{-j\omega}) \\ M_{\theta_0}(e^{-j\omega}) - M_{\theta}(e^{-j\omega}) \end{bmatrix} \frac{d\omega}{2\pi}$$

$S_{\mathbf{u}}$ è lo spettro di $\mathbf{u}$ con la solita convenzione di descrivere, se necessario, uno spettro a righe mediante funzioni δ. Ora, se si ha identificabilità lo spettro congiunto dev'essere non singolare quasi ovunque in $|z| = 1$ *. Se il primo membro è zero, l'integrando, che è una funzione non negativa della frequenza, deve essere zero per quasi tutti gli $\omega \in [-\pi, \pi]$. Questo implica che

$$L_{\theta_0}(e^{j\omega_k}) - L_{\theta}(e^{j\omega_k}) = 0\,, \qquad M_{\theta_0}(e^{j\omega_k}) - M_{\theta}(e^{j\omega_k}) = 0$$

almeno nelle n frequenze $\omega_1, \omega_2, \ldots, \omega_n$ in cui lo spettro di $\mathbf{u}$ è positivo. Per l'ipotesi di identificabiltà l'ordine di persistente eccitazione dell'ingresso

*Cf. la proposizione 19.

è sufficiente a concludere che queste eguaglianze puntuali implicano in realtà l'eguaglianza di L_{θ_0} a L_θ e di M_{θ_0} a M_θ per tutte le frequenze $\omega \in [-\pi, \pi]$; i.e.

$$L_{\theta_0}(z) \equiv L_\theta(z)\,, \qquad M_{\theta_0}(z) \equiv M_\theta(z)$$

dove $\equiv$ significa uguaglianza per tutti i z.

Confrontando ora le espressioni di $L(z)$ e di $M(z)$, queste ultime relazioni sono equivalenti alle

$$F_{\theta_0}(z) \equiv F_\theta(z)\,, \qquad G_{\theta_0}(z) \equiv G_\theta(z)$$

e quindi a $M(\theta_0) = M(\theta)$. L'dentificabilità locale in θ_0 implica infine che questo possa accadere allora e solo allora che $\theta = \theta_0$. □

Il teorema è così dimostrato. □

Problema 8 *Si vuole identificare un sistema senza reazione descritto dal modello vero*

$$\mathbf{y}(t) + a_0 \mathbf{y}(t-1) = b_0 \mathbf{u}(t-1) + \mathbf{e}_0(t) + c_0 \mathbf{e}_0(t-1)$$

dove $\mathbf{e}_0$ *è bianco di varianza* λ_0^2*, usando un modello ARX della classe*

$$\mathbf{y}(t) + a\,\mathbf{y}(t-1) = b\mathbf{u}(t-1) + \mathbf{e}(t) \tag{90}$$

Descrivere l'insieme dei valori del parametro $\theta := [a\ b]^\top$ *a cui converge lo stimatore a minimo errore di predizione* $\hat{\theta}_N$ *quando la numerosità campionaria* $N \to \infty$*.*

Si assuma che i dati siano ergodici del secondo ordine.

Soluzione:
Scriviamo il predittore vero nella forma implicita

$$\hat{\mathbf{y}}_0(t \mid t-1) = -a_0\mathbf{y}(t-1) + b_0\mathbf{u}(t-1) + c_0\mathbf{e}_0(t-1)$$

e l'errore di predizione

$$\begin{aligned}\boldsymbol{\varepsilon}_\theta(t) &= \mathbf{e}_0(t) + \hat{\mathbf{y}}_0(t \mid t-1) - \hat{\mathbf{y}}_\theta(t \mid t-1) = \\ &\quad \mathbf{e}_0(t) + (a-a_0)\mathbf{y}(t-1) + (b_0-b)\mathbf{u}(t-1) + c_0\mathbf{e}_0(t-1)\end{aligned}$$

la cui varianza, tenendo conto dell'ortogonalità $\mathbf{u} \perp \mathbf{e}_0$, risulta

$$\begin{aligned}\bar{V}(\theta) &= \mathbb{E}_0\boldsymbol{\varepsilon}_\theta^2(t) = (1+c_0^2)\lambda_0^2 + \\ &\quad + (a-a_0)^2\sigma_{\mathbf{y}}(0) + (b_0-b)^2\sigma_{\mathbf{u}}(0) + 2(a-a_0)(b_0-b)\sigma_{\mathbf{yu}}(0) + 2(a-a_0)c_0\lambda_0^2\end{aligned}$$

dove $\sigma_{\mathbf{y}}(\tau)$, $\sigma_{\mathbf{u}}(\tau)$, $\sigma_{\mathbf{yu}}(\tau)$ sono le auto e mutue covarianze dei segnali di uscita e ingresso, che ovviamene non dipendono da θ ma solo dai parametri veri.

Uguagliando a zero le derivate di $\bar{V}(\theta)$ rispetto ad a e b si trova:

$$\begin{aligned} 2(a-a_0)\sigma_{\mathbf{y}}(0)+2(b_0-b)\sigma_{\mathbf{yu}}(0)+2c_0\lambda_0^2 &= 0 \\ -2(b_0-b)\sigma_{\mathbf{u}}(0)-2(a-a_0)\sigma_{\mathbf{yu}}(0) &= 0 \end{aligned}$$

Risolvendo la seconda e successivamente sostituendo nella prima si ottengono le relazioni

$$\hat{b} = b_0 + (\hat{a}-a_0)\frac{\sigma_{\mathbf{yu}}(0)}{\sigma_{\mathbf{u}}(0)}; \qquad \hat{a} = a_0 - \frac{c_0\lambda_0^2}{\sigma_{\mathbf{y}}(0) - \dfrac{\sigma_{\mathbf{yu}}(0)^2}{\sigma_{\mathbf{u}}(0)}}$$

da cui si vede che

1. $\hat{a}_N$ può essere uno stimatore consistente solo se $c_0 = 0$ (ovvero il modello vero appartiene alla classe)

2. quindi $\hat{b}_N$ può essere uno stimatore consistente solo se $\sigma_{\mathbf{yu}}(0) = 0$, il che accade ad esempio se $\mathbf{u}$ è un processo bianco (lo studente dimostri questa affermazione !).

Problema 9 *Si vuole identificare una serie temporale usando un modello ARMA (d'innovazine) di ordine 1,*

$$\mathbf{y}(t) + a\,\mathbf{y}(t-1) = \mathbf{e}(t) + c\,\mathbf{e}(t-1) \tag{91}$$

mentre i dati osservati sono in realtà generati da un processo di rumore bianco i.i.d.

$$\mathbf{y}(t) = \mathbf{e}_0(t), \qquad \mathrm{var}\{\mathbf{e}_0(t)\} = \lambda_0^2 .$$

Trovare l'insieme dei valori del parametro $\theta := [a\,c]^\top$ *a cui converge, quando la numerosità campionaria* $N \to \infty$*, lo stimatore a minimo errore di predizione* $\hat{\theta}_N$*.*

Discutere il risultato; secondo voi :

- *Qual'è il valore vero del parametro* θ*?*
- *Lo stimatore PEM* $\hat{\theta}_N$ *è consistente?*

Soluzione:
Evidentemente un modello ARMA di ordine 1 dovrà essere una descrizione ridondante (non identificabile) per un processo di rumore bianco. Verifichiamo formalmente questo fatto.

Il processo vero appartiene alla classe parametrica; può essere descritto da tutti i modelli ARMA di ordine uno

$$\mathbf{y}(t) + a_0\mathbf{y}(t-1) = \mathbf{e}(t) + c_0\mathbf{e}(t-1)$$

in cui numeratore e denominatore sono uguali; i.e. $c_0 = a_0$. Qualunque punto che giace sulla retta $\{a = c\}$ del piano dei parametri $\{a, c\}$ è un parametro vero. Abbiamo già visto che il modello (91) è identificabile a priori (ricordare che l'identificabilità a priori è una proprietà *locale*) in tutti i punti del piano $\{a, c\}$, eccetto proprio in quei punti che giacciono sulla retta $\{a = c\}$. Quindi la classe parametrica di modelli (91) *non è identificabile in* θ_0*, qualunque esso sia !*.

In ogni caso, il predittore di Wiener (a minima varianza) per un modello ARMA di ordine 1, si può scrivere nella forma

$$\hat{\mathbf{y}}(t \mid t-1) = \frac{c-a}{1+cz^{-1}}\,\mathbf{y}(t-1)$$

e l'errore di predizione con dati osservati generati da un processo di rumore bianco i.i.d. $\mathbf{e}_0(t)$, è

$$\varepsilon_\theta(t) = \mathbf{e}_0(t) - \frac{c-a}{1+cz^{-1}}\,\mathbf{e}_0(t-1)\,.$$

Dato che i due termini sono scorrelati la varianza asintotica di $\boldsymbol{\varepsilon}_\theta(t)$ è

$$E_0(\boldsymbol{\varepsilon}_\theta(t))^2 = \lambda_0^2\,(1+\frac{(c-a)^2}{1-c^2})$$

che è evidentemente minima per $c=a$, entrambi uguali ad un qualunque valore (di modulo <1).

Quindi lo stimatore a minimo errore di predizione $\hat{\theta}_N := [\hat{a}_N\ \hat{c}_N]^\top$ converge, quando la numerosità campionaria $N \to \infty$ all'insieme $\Delta = \{\theta \mid a=c\}$.

Questo significa solo che $(\hat{a}_N - \hat{c}_N) \to 0$, ma i due stimatori, presi separatamente, in genere non convergono (il che è ovvio se si pensa che la successione di stime si ottiene prendendo, al variare di N, un punto di minimo di $V_N(\theta)$ in un insieme di valori estremali del parametro che contiene infiniti punti).

Conclusione:

◦ Anche se il modello vero appartiene alla classe parametrica, il valore vero del parametro non è univocamente determinato. Qualunque $\theta_0 = [a\ a]^\top$ ($|a| < 1$), è un valore vero.
◦ Per definizione lo stimatore $\hat{\theta}_N$ è *consistente* se $\hat{\theta}_N$ converge e converge al valore vero del parametro. Nel nostro caso $\hat{\theta}_N$ non converge puntualmente e quindi non è consistente.

Notare invece che $(\hat{a}_N - \hat{c}_N)$ è uno stimatore consistente di $a_0 - c_0 (= 0)$.

Problema 10 *Cosa accade dell'enunciato del teorema 20 se i dati sono semplicemente stazionari del second'ordine? Discutere un semplice esempio in cui il modello ha ordine uno e* $\mathbf{u}(t)$ *è un segnale stazionario sinusoidale. Ad esempio si prenda la classe di modelli*

$$M(\theta): \quad (1-az^{-1})\mathbf{y}(t) = b\mathbf{u}(t-1) + \mathbf{e}(t)$$

con $\mathbf{u}(t) = U\sin\omega_0 t$ *ed* $\mathbf{e}$ *bianco. Calcolare* $\bar{V}(\theta)$ *(che dipenderà dall'ampiezza del segnale di ingresso) e minimizzarlo rispetto a* θ*. Verificare se il limite di* $\hat{\theta}_N$ *dipende dall'ampiezza dell'ingresso. Si ha consistenza?*

Cosa accade se $\mathbf{u}$*, invece di essere un segnale deterministico stazionario, fosse un processo stazionario puramente deterministico (ad esempio una variabile aleatoria costante di media nulla) scorrelato da* $\mathbf{e}$*.*

Si deve comunque garantire l' identificabilità.

Finora non ci siamo interessati alla stima della varianza dell'innovazione. Il seguente risultato completa il teorema di consistenza.

Corollario 6 *Nelle stesse ipotesi del teorema 20, l'errore quadratico medio residuo* (70) *è uno stimatore consistente della varianza d'innovazione, ovvero*

$$\lim_{N\to\infty} \hat{\lambda}_N^2 = \lambda_0^2 \tag{92}$$

con probabilità uno.

Prova: Usando la (66) possiamo scrivere

$$\boldsymbol{\varepsilon}_{\hat{\theta}_N}(t) = G_{\hat{\theta}_N}(z)^{-1}\left[\mathbf{y}(t) - F_{\hat{\theta}_N}(z)\mathbf{u}(t)\right]$$

Nelle ipotesi in cui ci siamo posti, $\hat{\theta}_N$ converge al valore vero θ_0 e quindi passando al limite per $N \to \infty$ nell'espressione precedente si vede che l'errore residuo di predizione converge all'innovazione vera $\mathbf{e}_0(t) = \boldsymbol{\varepsilon}_{\theta_0}(t)$. In effetti, dato che $V_N(\hat{\theta}_N)$ è la varianza campionaria di $\boldsymbol{\varepsilon}_{\hat{\theta}_N}(t)$ e che

$$\lim_{N\to\infty} V_N(\hat{\theta}_N) = \mathbb{E}_{\theta_0}\boldsymbol{\varepsilon}_{\theta_0}^2(t) = \mathbb{E}_0\mathbf{e}_0^2(t)$$

si vede che anche la varianza di $\boldsymbol{\varepsilon}_{\hat{\theta}_N}(t)$ converge a quella di $\mathbf{e}_0(t)$. □

Il caso di parametrizzazioni indipendenti

Consideriamo un modello del tipo Box-Jenkins in cui le funzioni di trasferimento $F(z)$ e $G(z)$ *sono parametrizzate in modo indipendente*. Questo significa che si può decomporre θ in due sottovettori indipendenti; i.e. $\theta = [\xi\ \eta]^\top \in \Xi \times E = \Theta$ per cui

$$F_\theta(z) \equiv F_\xi(z)\,, \qquad G_\theta(z) \equiv G_\eta(z) \tag{93}$$

Ci si chiede se i risultati appena visti possono valere separatamente per le due classi parametriche di funzioni di trasferimento. Si può parlare di "consistenza parziale" ? Domanda è di interesse in pratica perchè è normalmente più importante identificare correttamente una delle due funzioni di trasferimanto (tipicamente quella della parte "deterministica" $F(z)$).

La risposta al quesito precedente è positiva solo nel caso di processi senza reazione. Discuteremo solo questo caso.

Definizione 22 *Si assuma che le funzioni di trasferimento $F(z)$ e $G(z)$ nella famiglia siano parametrizzate in modo indipendente e che vi sia assenza di reazione da $\mathbf{y}$ a $\mathbf{u}$. Diremo che, nella condizione sperimentale descritta da $S_{\mathbf{u}}(z)$ (o dalla distribuzione spettrale $d\hat{F}_{\mathbf{u}}(z)$), si ha identificabiltà (globale) della mappa ingresso-uscita se*

$$S_{\mathbf{yu}}(\cdot;\xi_1) = S_{\mathbf{yu}}(\cdot;\xi_2) \Rightarrow \xi_1 = \xi_2$$

Se si ha iniettività locale in un intorno di ξ_0, si parla di identificabilità locale *in ξ_0.*

Notiamo che per l'assenza di reazione si ha $S_{\mathbf{yu}}(z;\xi) = F_{\xi}(z)S_{\mathbf{u}}(z)$, per cui la condizione è equivalente alla

$$\left[F_{\xi_1}(z) - F_{\xi_2}(z)\right] S_{\mathbf{u}}(z) \equiv 0 \Rightarrow \xi_1 = \xi_2\,.$$

Naturalmente nel caso di ingressi con righe spettrali bisognerebbe a rigore riscrivere il primo termine come un integrale rispetto alla distribuzione spettrale $d\hat{F}_{\mathbf{u}}(z)$.

Teorema 21 *Supponiamo che i dati siano generati da processi ergodici del secondo ordine e che non vi sia reazione da* $\mathbf{y}$ *ad* $\mathbf{u}$*. L'errore di predizione* $\boldsymbol{\varepsilon}_\theta(t)$ *sia calcolato mediante il predittore lineare a minima varianza* $\hat{\mathbf{y}}_\theta(t \mid t-1)$*. Supponiamo inoltre che nella classe parametrica dei modelli* $\mathcal{M} \equiv \{M(\theta);\, \theta \in \Theta\}$*,* $F(z)$ *e* $G(z)$ *siano parametrizzate in modo indipendente come descritto più sopra e che il processo* $\mathbf{u}$ *sia sufficientemente eccitante da garantire l' identificabilità della classe di modelli* $\mathcal{F} := \{F_\xi(z)\,;\, \xi \in \Xi\}$*.*

Allora, se il processo (vero) che genera i dati è descritto da una funzione di trasferimento $F_0(z)$ *che appartiene alla stessa classe parametrica* $\mathcal{F}$ *delle funzioni* $F_\xi(z)$ *scelte per l'identificazione, ovvero esiste* $\xi_0 \in \Xi$ *tale che* $F_0(z) \equiv F_{\xi_0}(z)$*, si ha:*

$$\lim_{N\to\infty} \hat{\boldsymbol{\xi}}_N = \xi_0 \tag{94}$$

con probabilità uno. In altri termini, lo stimatore PEM del parametro ξ è consistente*.*

Prova: Usando l'espressione (66) e sostituendo al posto di $\mathbf{y}$ la sua rappresentazione mediante il modello vero $\mathbf{y} = F_0(z)\mathbf{u}(t) + G_0(z)\mathbf{e}_0(t)$, si trova

$$\boldsymbol{\varepsilon}_\theta(t) = G_\eta(z)^{-1}\left[(F_0(z) - F_\xi(z))\mathbf{u}(t) + G_0(z)\mathbf{e}_0(t)\right] := L_{\xi,\eta}(z)\mathbf{u}(t) + M_\eta(z)\mathbf{e}_0(t)$$

dove i due addendi nell'ultimo termine sono scorrelati per l'assenza di reazione. Si trova così

$$\bar{V}(\theta) = \bar{V}(\xi, \eta) = \mathrm{var}\left[L_{\xi,\eta}(z)\mathbf{u}(t)\right] + \mathrm{var}\left[M_\eta(z)\mathbf{e}_0(t)\right]$$

Ora, dato che $F_0(z) = F_{\xi_0}(z)$ e si ha identificabilità parziale, il primo termine ha un unico minimo (zero) per $\xi = \xi_0$. Dato che $\hat{\boldsymbol{\theta}}_N$ converge con probabilità uno, sicuramente anche le sue prime componenti, $\hat{\boldsymbol{\xi}}_N$ convergono e convergono necessariamente all'insieme dei punti di Ξ in cui si ha il minimo del primo addendo. Ma questo insieme è costituito dal solo punto $\{\xi_0\}$. □

Cosa accade in pratica quando si usano modelli in cui si ha scarsa conoscenza a priori dello spettro dell'errore di modellizzazione? Anche se il modello per questo processo è grossolanamente errato, si può comunque avere consistenza per la stima della funzione di trasferiemento ingresso-uscita $F(z)$.

MODELLI A ERRORE DI EQUAZIONE

Si chiamano *Output Error models (O.E.)* in inglese; sono del tipo

$$\mathbf{y}(t) = F_{\theta}(z)\mathbf{u}(t) + \mathbf{e}(t) \tag{95}$$

dove $\mathbf{e}$ è bianco e scorrelato da $\mathbf{u}$. Si possono avere stime consistenti di $F_0(z)$ anche se la vera $G_0(z)$ è molto diversa da 1.

Esempio 5 *Si vuole identificare il sistema "vero"*

$$\mathbf{y}(t) = \frac{b_0}{1 + a_0 z - 1}\mathbf{u}(t-1) + \mathbf{e}_0(t) + c_0\mathbf{e}_0(t-1)$$

dove $\mathbf{u}$ *ed* $\mathbf{e}_0$ *sono bianchi, scorrelati, di varianze rispettive* σ^2 *e* λ_0^2*, usando un modello a errore sull'uscita (O.E.) di ordine 1,*

$$\mathbf{y}(t) = \frac{b}{1 + az^{-1}}\mathbf{u}(t-1) + \mathbf{e}(t)\,.$$

dove $\mathbf{e}$ *è rumore bianco. Trovare l'insieme dei valori del parametro* $\theta := [a\ b]^\top$ *a cui converge (quando la numerosità campionaria* $N \to \infty$*) lo stimatore a minimo errore di predizione* $\hat{\theta}_N$*.*

Soluzione:

Il predittore per il sistema "vero" si può scrivere senza esprimere $\mathbf{e}_0$ in funzione dei dati ingresso-uscita, come

$$\hat{\mathbf{y}}_0(t \mid t-1) = \frac{b_0}{1+a_0 z^{-1}}\mathbf{u}(t-1) + c_0\mathbf{e}_0(t-1)$$

mentre quello per il modello si scrive

$$\hat{\mathbf{y}}_\theta(t \mid t-1) = \frac{b}{1+az^{-1}}\mathbf{u}(t-1)\,.$$

La varianza dell'errore di predizione si esprime quindi come,

$$\begin{aligned}
\mathrm{var}\boldsymbol{\varepsilon}_\theta &= \lambda_0^2 + \mathrm{var}\left[\hat{\mathbf{y}}_0(t \mid t-1) - \hat{\mathbf{y}}_\theta(t \mid t-1)\right] \\
&= \lambda_0^2 + \mathrm{var}\left\{\left[\frac{b_0}{1+a_0 z^{-1}} - \frac{b}{1+az^{-1}}\right]\mathbf{u}(t-1) + c_0\mathbf{e}_0(t-1)\right\} \\
&= \lambda_0^2 + \mathrm{var}\left\{\left[\frac{b_0}{1+a_0 z^{-1}} - \frac{b}{1+az^{-1}}\right]\mathbf{u}(t-1)\right\} + c_0^2\lambda_0^2
\end{aligned}$$

dato che se $\mathbf{u}$ ed $\mathbf{e}_0$ sono scorrelati, lo sono anche funzioni lineari arbitrarie della storia dei due processi.

Questa varianza si può minimizzare facilmente rispetto a θ prendendo

$$\left[\frac{b_0}{1+a_0 z^{-1}} - \frac{b}{1+az^{-1}}\right] = 0$$

il che accade se e solo se $a = a_0$ e $b = b_0$. In questo senso lo stimatore $\hat{\theta}_N$ dei parametri della funzione di trasferimento $F(z)$ è "consistente" i.e. converge ai valori veri $\theta_0 := [a_0\, b_0]^\top$ anche se a rigore con questa classe di modelli non si può avere consistenza. Ovviamente a questa conclusione si sarebbe potuti arrivare direttamente in base al teorema 21.

Il "metodo" dei minimi quadrati

Si consideri il modello "vero"

$$\mathbf{y}(t) = F(z)\,\mathbf{u}(t) + \mathbf{e}_0(t),$$

dove $F(z)$ è una funzione di trasferimento causale (non necessariamente razionale), $\{\mathbf{u}(t)\}$ ed $\{\mathbf{e}_0(t)\}$ sono processi ergodici a media nulla tra loro indipendenti e $\{\mathbf{e}_0(t)\}$ è bianco.

Si cerca di approssimare la $F(z)$ con una funzione di trasferimento razionale di grado fissato usando i dati $\{\mathbf{y}(t)\}$ e $\{\mathbf{u}(t)\}$ con $t = 1,\ldots,N$, identificando un modello di tipo "output error"

$$\mathbf{y}(t) = \frac{B(z^{-1})}{A(z^{-1})}\mathbf{u}(t) + \mathbf{e}(t) \tag{96}$$

dove $A(z^{-1}) = 1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_n z^{-n}$ e $B(z^{-1}) = b_1 z^{-1} + b_2 z^{-2} + \ldots + b_m z^{-m}$. L'identificazione col metodo PEM conduce (come vedremo tra un attimo) ad un problema di stima non lineare perchè i parametri $\{a_k\}$ compaiono in modo non lineare nel predittore. Per questo motivo qualche volta

si usa un metodo empirico che viene genericamente (e impropriamente) chiamato *metodo dei minimi quadrati* che descriviamo qui di seguito.

Definendo :

$$\varphi(t) := [-\mathbf{y}(t-1) \ \ldots -\mathbf{y}(t-n) \ \ \mathbf{u}(t-1) \ \ldots \ \mathbf{u}(t-m)]^\top$$

e $\theta := [a_1 \ a_2 \ \ldots a_n \ \ b_1 \ b_2 \ \ldots \ b_m]^\top$ il vettore dei coefficienti, il modello (96) si può scrivere in forma di regressione

$$\mathbf{y}(t) = \varphi(t)^\top \theta + \mathbf{w}(t) \qquad t = 1, \ldots, N,$$

dove però $\mathbf{w}(t) := A(z^{-1})\mathbf{e}(t)$ non è ovviamente rumore bianco. Nonostante ciò si stimano i coefficienti θ col metodo dei minimi quadrati ottenendo uno stimatore $\hat{\theta}_{LS}(N)$ che è dato dalla nota formula

$$\hat{\theta}_{LS}(N) = \left[\sum_{t=1}^{N} \varphi(t)\varphi(t)^\top\right]^{-1} \sum_{t=1}^{N} \varphi(t)\,\mathbf{y}(t) \tag{97}$$

Ci si chiede quando questa espressione fornisca stime consistenti. Assumeremo che $F(z)$ appartenga alla classe di funzioni razionali definite sopra, per cui si può scrivere

$$\mathbf{y}(t) = \varphi(t)^\top \theta_0 + \mathbf{w}_0(t)$$

con $\mathbf{w}_0(t) := [\,\mathbf{e}_0(t)\,\mathbf{e}_0(t-1)\,\ldots,\mathbf{e}_0(t-n)\,]^\top [\,1\,a_{0,1}\,a_{0,2}\,\ldots a_{0,n}\,] := \eta(t)^\top \mathbf{a}_0$, dove il pedice $_0$ significa parametro "vero". Per verificare se $\hat{\theta}_{LS}(N)$ è consistente sostituiamo l'espressione precedente nella (99) ottenedo

$$\hat{\theta}_{LS}(N) = \theta_0 + \left[\frac{1}{N}\sum_{t=1}^{N} \varphi(t)\varphi(t)^\top\right]^{-1} \frac{1}{N}\sum_{t=1}^{N} \varphi(t)\,\mathbf{w}_0(t)\,. \tag{98}$$

Passando al limite per $N \to \infty$ e usando il teorema ergodico si trova

$$\lim_{N\to\infty} \hat{\theta}_{LS}(N) = \theta_0 + \left[\mathbb{E}\,\varphi(t)\varphi(t)^\top\right]^{-1} \mathbb{E}\,\varphi(t)\,\eta(t)^\top \mathbf{a}_0\,. \tag{99}$$

Dato che $\mathbf{u}$ e $\mathbf{e}_0$ sono indipendenti, la matrice di correlazione $\mathbb{E}\,\varphi(t)\,\eta(t)^\top$ ha la seguente struttura

$$\mathbb{E}\,\varphi(t)\,\eta(t)^\top = \begin{bmatrix} \mathbb{E}\begin{bmatrix}\mathbf{y}(t-1)\\ \vdots \\ \mathbf{y}(t-n)\end{bmatrix} [\,\mathbf{e}_0(t)\,\mathbf{e}_0(t-1)\,\ldots,\mathbf{e}_0(t-n)\,] \\ 0\,0\,\ldots\,0 \end{bmatrix}$$

la quale ha la prima colonna di zeri ma, dato che $\mathbf{y}(t-k)$ e $\mathbf{e}_0(t-j)$ hanno correlazione diversa da zero se $k \geq j$ ha una struttura triagolare superiore

e c'è sempre un termine diverso da zero nella diagonale secondaria superiore a meno che non sia $A_0(z^{-1}) \equiv 1$ e il modello vero non si riduca al modello FIR

$$\mathbf{y}(t) = B_0(z^{-1})\,\mathbf{u}(t) + \mathbf{e}_0(t)\,.$$

Quindi il cosiddetto "metodo dei minimi quadrati" è consistente solo se il modello è di tipo FIR con rumore bianco additivo.

Lo stimatore PEM, per un modello a errore d'equazione come (96), è basato sul predittore di Wiener

$$\hat{\mathbf{y}}_\theta(t \mid t-1) = \frac{B(z^{-1})}{A(z^{-1})}\mathbf{u}(t)$$

e questo predittore non è **mai lineare nei parametri** (cioè esprimibile nella forma $\varphi(t)^\top \theta$), a meno che non sia $A(z^{-1}) \equiv 1$. Quindi lo stimatore $\hat{\theta}_{LS}(N)$ è uno stimatore PEM solo nel caso in cui $A(z^{-1}) \equiv 1$.

Lo stimatore PEM di θ per un modello a errore d'equazione va calcolato mediante algoritmi di ottimizzazione iterativa non-lineare.

Analisi nel dominio della frequenza

Notazione: per semplicità scriviamo $H(e^{j\omega}) \equiv H(\omega)$. Supponiamo che la classe di modelli sia generale, Box-Jenkins. Facciamo l'ipotesi che non ci sia reazione $(\mathbf{u} \perp \mathbf{e})$.

$$\mathbf{y}(t) = F_0(z)\mathbf{u}(t) + G_0(z)\mathbf{e}(t)\,, \qquad M(\theta) \;\equiv\; \mathbf{y}(t) = F_\theta(z)\mathbf{u}(t) + G_\theta(z)\mathbf{e}(t)\,.$$

Modello vero non necessariamente nella classe. Lo spettro di potenza dell'errore di predizione

$$\begin{aligned}\varepsilon_\theta(t) &= G_\theta^{-1}(z)[\mathbf{y}(t) - F_\theta(z)\mathbf{u}(t)] = \\ &= G_\theta^{-1}(z)[(F_0(z) - F_\theta(z))u(t) + G_0(z)\mathbf{e}_0(t)]\,,\end{aligned}$$

risulta

$$S_\varepsilon(\omega) = |F_0(\omega) - F_\theta(\omega)|^2 \frac{S_\mathbf{u}(\omega)}{|G_\theta(\omega)|^2} + \frac{S_\mathbf{v}(\omega)}{|G_\theta(\omega)|^2}$$

Nell'identificazione PEM si cerca quindi di minimizzare la quantità

$$\bar{V}(\theta) = \operatorname{var} \varepsilon_\theta(t) = \int_{-\pi}^{\pi} |F_0(\omega) - F_\theta(\omega)|^2 \frac{S_\mathbf{u}(\omega)}{|G_\theta(\omega)|^2} \frac{d\omega}{2\pi} + \lambda_0^2 \int_{-\pi}^{\pi} \frac{|G_0(\omega)|^2}{|G_\theta(\omega)|^2} \frac{d\omega}{2\pi}$$

Identificazione di serie temporali.

In questo caso $\mathbf{u} \equiv 0$ e

$$\bar{V}(\theta) = \lambda_0^2 \int_{-\pi}^{\pi} \frac{|G_0(\omega)|^2}{|G_\theta(\omega)|^2} \frac{d\omega}{2\pi}$$

Il minimo di questa espressione si ha se e solo se $G_\theta = G_0$ (entrambi fattori spettrali a fase minima normalizzati a 1 all'infinito) e vale λ_0^2. Infatti:

$$\bar{V}(\theta) = \lambda_0^2 + \|\hat{\mathbf{y}}_0(t \mid t-1) - \hat{\mathbf{y}}_\theta(t \mid t-1)\|^2$$

Approssimazione "sperimentale" di modelli complessi

Vogliamo identificare un sistema dalla dinamica complessa (i.e. con funzione di trasferimento lineare ma di ordine elevato):

$$y(t) = F_0(z)u(t)\,,$$

con errore di modellizzazione (rumore sulle misure dell'uscita) trascurabile. Siamo interessati ad effettuare un'identificazione con un modello di ordine basso (*experimental model reduction*). Per questo motivo aggiungiamo una componente stocastica al modello approssimato che descriva in modo statistico l'errore di modellizzazione.

Consideriamo una classe di modelli Box–Jenkins

$$\mathbf{y}(t) = F_\theta(z)\mathbf{u}(t) + G_\theta(z)\mathbf{e}(t)\,.$$

in cui l'ordine di $F_\theta(z)$ è fissato, minore di quello di $F_0(z)$.

L'errore di predizione del modello approssimato è

$$\boldsymbol{\varepsilon}_\theta(t) = G_\theta^{-1}(z)[\mathbf{y}(t) - F_\theta(z)\mathbf{u}(t)] = G_\theta^{-1}(z)[(F_0(z) - F_\theta(z))\mathbf{u}(t)]\,,$$

e il suo spettro risulta

$$S_{\boldsymbol{\varepsilon}}(\omega) = \frac{|F_0(\omega) - F_\theta(\omega)|^2 S_{\mathbf{u}}(\omega)}{|G_\theta(\omega)|^2}\,.$$

Nell'identificazione PEM si cerca di minimizzare la varianza asintotica

$$\bar{V}(\theta) = \operatorname{var} \varepsilon_\theta(t) = \int_{-\pi}^{\pi} |F_0(\omega) - F_\theta(\omega)|^2 Q(\omega) d\omega\,,$$

che si puo' interpretare come una distanza pesata in frequenza con funzione peso

$$Q(\omega) = \frac{S_{\mathbf{u}}(\omega)}{|G_\theta(\omega)|^2}\,.$$

Se si può scegliere lo spettro dell'ingresso $S_{\mathbf{u}}(\omega)$ si può ottimizzare l'approssimazione in una banda di frequenze desiderata. Il ruolo di G_θ è più sottile.

Esempio 1: Identificazione con modelli OE

Si effettua l'identificazione con una classe di modelli che ha $G_\theta = 1$ ovvero Output–Error, ad esempio del tipo

$$\mathbf{y}(t) = \frac{b_1 z^{-1} + b_2 z^{-2}}{1 + f_1 z^{-1} + f_2 z^{-2}} \mathbf{u}(t) + \mathbf{e}(t)\,,$$

Se l'ingresso è bianco, o comunque a larga banda, $Q(\omega)$ è all'incirca costante e si ottiene un'approssimazione che pesa in ugual modo gli errori di approssimazione su tutta una banda estesa di frequenze. Il risultato è generalmente mediocre.

Vedere il diagramma di Bode nell'esercitazione di Martedi' prossimo.

Esempi0 2: Identificazione con modelli ARX

Si ripeta la procedura, utilizzando questa volta un modello di tipo ARX, ad esempio

$$(1+a_1z^{-1}+a_2z^{-2})\mathbf{y}(t) = (b_1z^{-1}+b_2z^{-2})\mathbf{u}(t)+\mathbf{e}(t)$$

In questo caso $Q(\omega) = S_{\mathbf{u}}(\omega)|A_\theta(\omega)|^2$. Se $\mathbf{u}$ è bianco o a larga banda, questo è un filtro FIR generalmente di tipo passa alto e l'approssimazione alle basse frequenze risulta peggiore di prima.

La funzione $Q(\omega)$ pesa maggiormente determinate frequenze a scapito di altre, Per ottenere una migliore approssimazione di F_0 in una banda di frequenze voluta conviene filtrare l'errore di predizione con un filtro passa–banda:

$$\varepsilon_\theta^L(t) = L(z)\varepsilon_\theta(t)\,.$$

In questo modo il problema PEM diventa

$$\theta^* = \arg\min_\theta \operatorname{var}\varepsilon_\theta^L(t) = \arg\min_\theta \int_{-\pi}^{\pi} |F_0(\omega)-F_\theta(\omega)|^2 Q^L(\omega)d\omega\,,$$

Dove:

$$Q^L(\omega) = \frac{|L(\omega)|^2}{|G_\theta(\omega)|^2} S_u(\omega)\,.$$

Come si vede il filtraggio equivale a identificare il sistema mediante un modello Box-Jenkins con

$$G^L(z) := \frac{G_\theta(z)}{L(z)}$$

Cf. Lennart Ljung. *System Identification. Theory for the user*. Capp. 8.5 e 13.1. Ci sono vari esempi di simulazioni.

ASPETTI COMPUTAZIONALI

Trattiamo solo problemi in cui N è grande, interessa il comportamento asintotico e si può quindi approssimare il predittore con quello di Wiener. Altrimenti si massimizza numericamente la verosimiglianza esatta (iterazione sul filtro di Kalman). *Algoritmi di Quasi-Newton per la minimizzazione locale*

La minimizzazione dell'errore quadratico medio di predizione si può fare esplicitamente (ed esattamente) solo per modelli a predittore lineare nei parametri, tipicamente per modelli ARX. In questo caso si tratta di risolvere un problemi ai minimi quadrati per il quale esistono rutines stabili e ben collaudate.

Se si usano modelli di tipo generale e quindi modelli per i quali il predittore ha memoria infinita (non ha la struttua di un filtro FIR, come per i modelli ARX) alcuni parametri del modello determinano i modi del sistema e compaiono implicitamente in modo non lineare nella risposta impulsiva del predittore. In questi casi il predittore è funzione non lineare dei parametri

del modello e la minimizzazione di $V_N(\theta)$ si può fare solo per via numerica mediante algoritmi iterativi di ottimizzazione.
Illustreremo una classe di algoritmi particolarmente adatti alla specifica struttura (quadratica) dell'errore quadratico medio di predizione. Il capostipite di questi algoritmi è *l'algoritmo di Newton* (o Newton-Rapson) che descriveremo qui sotto.

Sia $V(\theta) : \mathbb{R}^p \to \mathbb{R}$ una funzione di classe C^2. Per minimizzare $V(\theta)$ rispetto a θ possiamo pensare di approssimare la funzione localmente attorno ad un punto $\bar{\theta}$ mediante una funzione quadratica,

$$V(\theta) \simeq V(\bar{\theta}) + (\theta - \bar{\theta})^\top V'(\bar{\theta}) + \frac{1}{2}(\theta - \bar{\theta})^\top V''(\bar{\theta})\,(\theta - \bar{\theta}) \tag{100}$$

dove

$$V'(\bar{\theta}) := \frac{\partial}{\partial \theta} V(\theta)_{|\,\theta=\bar{\theta}}\,, \qquad V''(\bar{\theta}) = \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} V(\theta)\right]_{|\,\theta=\bar{\theta}}$$

sono il gradiente (che scriviamo come un vettore colonna) e la matrice Hessiana di $V(\theta)$ calcolati in $\bar{\theta}$. La minimizzazione della funzione quadrat-

ica al secondo membro di (100) conduce all'equazione

$$V''(\bar{\theta})\,(\theta - \bar{\theta}) = -V'(\bar{\theta})$$

le quale, nell'ipotesi che $V''(\bar{\theta})$ sia non singolare, fornisce l'espressione del punto di minimo "approssimato",

$$\hat{\theta} = \bar{\theta} - \left[V''(\bar{\theta})\right]^{-1} V'(\bar{\theta})\,. \tag{101}$$

L'idea base dell'algoritmo è di usare questa equazione in modo iterativo prendendo $\bar{\theta} = \theta_k$ e $\theta_{k+1} = \hat{\theta}$ e iterando su k. Come si vede il punto di minimo "approssimato", $\hat{\theta}$ si calcola partendo da $\bar{\theta}$ e muovendosi nella direzione del gradiente negativo (quindi una direzione di discesa) "filtrata" attraverso l'inversa della matrice Hessiana. Se $\bar{\theta}$ è vicino ad un punto di minimo si può supporre che per continuità $V''(\bar{\theta})$ sia definita positiva e quindi che la direzione del gradiente "filtrata" (che viene chiamate *direzione di Newton)*, sia ancora una direzione di discesa. Se però si è lontani da un minimo $V''(\bar{\theta})$ potrebbe essere indefinita oppure addirittura definita negativa e la direzione di Newton potrebbe puntare verso un punto di crescita

della funzione anzichè un punto di decrescita. Come si vede questo poced-imento, benchè porti al minimo di una funzione quadratica in una sola iter-azione, ha dei grossi difetti se applicato a funzioni non quadratiche.

Per rimediare a questi difetti si introducono delle approssimazioni della ma-trice Hessiana che abbiano la proprietà di essere sempre almeno semidef-inite positive. Queste approssimazioni danno origine ad una famiglia di algoritmi che si chiamano di *Quasi-Newton*. Ne illustreremo uno partico-larmente adatto alla nostra funzione costo.

Usando le espressioni per le derivate (115),(116) e pensando di essere in una situazione limite ottimistica in cui θ_k è vicino al valore vero θ_0, si può pensare di trascurare il termine con le derivate seconde di $\varepsilon_\theta(t)$ nell'espressione della matrice Hessiana, ricavandone così un'approssimazione

$$H_\theta(N) := \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{\psi}_\theta(t)\boldsymbol{\psi}_\theta(t)^\top \simeq \frac{1}{2}V_N(\theta)'' \tag{102}$$

che è sempre almeno semidefinita positiva (e non richiede il calcolo delle derivate seconde di $\varepsilon_\theta(t)$). Inoltre, se si è nelle condizioni in cui vale il teorema del limite centrale, l'espressione $H_\theta(N)$ converge con probabilità uno per $N \to \infty$ a una matrice invertibile (in effetti a λ_0^2 volte la matrice di Fisher).

Si perviene così ad un classe di algoritmi che ha la seguente struttura.

Data la stringa dei dati di ingresso $[\mathbf{y}^N, \mathbf{u}^N] = [\mathbf{y}(1)\dots\mathbf{y}(N), \mathbf{u}(1)\dots\mathbf{u}(N)]$, la classe di modelli (67), un valore iniziale θ_0 e la stima θ_k alla $k-$sima iterazione,

1. Si calcola la stringa degli errori di predizione $\boldsymbol{\varepsilon}_{\theta_k} = [\boldsymbol{\varepsilon}_{\theta_k}(1)\dots\boldsymbol{\varepsilon}_{\theta_k}(N)]^\top$ risolvendo l'equazione

$$\boldsymbol{\varepsilon}_{\theta_k}(t) = G_{\theta_k}(z)^{-1}\left[\mathbf{y}(t) - F_{\theta_k}(z)\mathbf{u}(t)\right] \tag{103}$$

   prendendo condizioni iniziali arbitrarie (ad esempio nulle).

2. Si calcola la stringa dei gradienti $\boldsymbol{\Psi}_{\theta_k} = [\boldsymbol{\psi}_{\theta_k}(1) \dots \boldsymbol{\psi}_{\theta_k}(N)] \in \mathbb{R}^{p \times N}$, dell'errore di predizione del modello $M(\theta_k)$.

3. Si calcola la matrice pseudo-Hessiana

$$H_{\theta_k} := \sum_{t=1}^{N} \boldsymbol{\psi}_{\theta_k}(t) \boldsymbol{\psi}_{\theta_k}(t)^\top = \boldsymbol{\Psi}_{\theta_k} \boldsymbol{\Psi}_{\theta_k}^\top$$

   e la sua inversa $P_{\theta_k} := H_{\theta_k}^{-1}$.

4. Si aggiorna θ_k mediante la,

$$\theta_{k+1} - \theta_k = -H_{\theta_k}^{-1} \boldsymbol{\Psi}_{\theta_k} \boldsymbol{\varepsilon}_{\theta_k} \tag{104}$$

5. Si torna al passo 1) ponendo $\theta_k = \theta_{k+1}$.

Questo schema di principio richiede alcuni commenti.

1. Dato che il calcolo di $\boldsymbol{\varepsilon}_{\theta_k}$ e del gradiente $\boldsymbol{\psi}_{\theta_k}$ sono basati su predittori di Wiener in cui le condizioni iniziali (incognite) vengono scelte in modo arbitrario, i dati iniziali vengono "maltrattati" dall'algoritmo e per questo motivo sarebbe opportuno minimizzare un funzionale di costo scontato del tipo (68).

2. L' approssimazione della matrice Hessiana può risultare singolare o mal condizionata specialmente nelle iterazioni iniziali. Con molti parametri il funzionale costo potrebbe comunque risultare poco sensibile alla variazione di qualche parametro e anche l'Hessiano esatto potrebbe risultare mal condizionato. Per rimediare a questo mal condizionamento si può usare una tecnica di *regolarizzazione* e definire l'inversa "regolarizzata" come

$$P_{\theta_k} := \left[H_{\theta_k} + \delta(\theta_k)\, I\right]^{-1} \tag{105}$$

dove $\delta(\theta)$ è una opportuna funzione scalare positiva, ad esempio una costante o un termine proporzionale al quadrato di una norma pesata

di θ del tipo $\theta^\top Q\theta$ con $Q = Q^\top > 0$. Questo termine si può interpretare in un contesto Bayesiano come una varianza a priori del parametro θ. Per quanto visto all'inizio, questo termine introduce un errore sistematico nella stima e deve essere scelto opportunamente piccolo.

3. Il calcolo dell'errore di predizione richiede che il polinomio numeratore della funzione di trasferimento $G_{\theta_k}(z)$, $C_{\theta_k}(z^{-1})$, supponiamolo di grado q, e quello a denominatore $A_{\theta_k}(z^{-1})$ (di grado n) siano *polinomi stabili* nel senso che $z^q C_{\theta_k}(z^{-1}) = 0 \Rightarrow |z| < 1$ e $z^n A_{\theta_k}(z^{-1}) = 0 \Rightarrow |z| < 1$ (a meno di improbabili cancellazioni con il denominatore $D_{\theta_k}(z^{-1})$ di G_{θ_k}). La stessa cosa accade per il calcolo del gradiente.

   Questi vincoli di stabiltà sono stati sistematicamente trascurati nella minimizzazione dell'errore di predizione, che avrebbe a rigore dovuto essere una minimizzazione *vincolata* dalle condizioni di stabilità appunto, di $C_{\theta_k}(z^{-1})$ e $A_{\theta_k}(z^{-1})$.

4. Il calcolo del gradiente si può fare riferendosi alla parametrizzazione standard del modello di Box-Jenkins descritta dai polinomi (**??**). Le quattro componenti in cui viene naturalmente partizionato il gradiente,

$$\begin{aligned}
\boldsymbol{\psi}_a(t) &:= \left[\frac{\partial \boldsymbol{\varepsilon}_\theta(t)}{\partial a_i}\right]_{i=1,\ldots,n}; \quad \boldsymbol{\psi}_b(t) := \left[\frac{\partial \boldsymbol{\varepsilon}_\theta(t)}{\partial b_i}\right]_{i=1,\ldots,m}; \\
\boldsymbol{\psi}_c(t) &:= \left[\frac{\partial \boldsymbol{\varepsilon}_\theta(t)}{\partial c_i}\right]_{i=1,\ldots,q}; \quad \boldsymbol{\psi}_d(t) := \left[\frac{\partial \boldsymbol{\varepsilon}_\theta(t)}{\partial d_i}\right]_{i=1,\ldots,r};
\end{aligned}$$

si possono calcolare derivando membro a membro rispetto ai parametri il modello (67) riscritto nella forma

$$C_\theta(z^{-1})A_\theta(z^{-1})\boldsymbol{\varepsilon}_\theta(t) = A_\theta(z^{-1})D_\theta(z^{-1})\mathbf{y}(t) - B_\theta(z^{-1})D_\theta(z^{-1})\mathbf{u}(t)\,.$$

Calcolando le derivate rispetto alle prime componenti a_1, b_1, c_1, d_1, si trovano le equazioni alle differenze

$$\begin{aligned}
C_\theta(z^{-1})A_\theta(z^{-1})\boldsymbol{\psi}_{a_1}(t) &:= -C_\theta(z^{-1})\boldsymbol{\varepsilon}_\theta(t-1) + D_\theta(z^{-1})\mathbf{y}(t-1)\,; \\
C_\theta(z^{-1})A_\theta(z^{-1})\boldsymbol{\psi}_{b_1}(t) &:= -D_\theta(z^{-1})\mathbf{u}(t-1)\,; \\
C_\theta(z^{-1})A_\theta(z^{-1})\boldsymbol{\psi}_{c_1}(t) &:= -A_\theta(z^{-1})\boldsymbol{\varepsilon}_\theta(t-1)\,; \\
C_\theta(z^{-1})A_\theta(z^{-1})\boldsymbol{\psi}_{d_1}(t) &:= A_\theta(z^{-1})\mathbf{y}(t-1) - B_\theta(z^{-1})\mathbf{u}(t-1)\,;
\end{aligned}$$

che vengono di norma associate a condizioni iniziali nulle. In questo caso le componenti di indice maggiore di uno si possono ottenere semplicemente per traslazione; ad esempio

$$\boldsymbol{\psi}_{a_k}(t) = \boldsymbol{\psi}_{a_1}(t-k)\,, \quad k = 1,2,\ldots,n, \quad \boldsymbol{\psi}_{b_k}(t) = \boldsymbol{\psi}_{b_1}(t-k)\,, \quad k = 1,2,\ldots,m,$$

$$\boldsymbol{\psi}_{c_k}(t) = \boldsymbol{\psi}_{c_1}(t-k)\,, \quad k = 1,2,\ldots,q, \quad \boldsymbol{\psi}_{d_k}(t) = \boldsymbol{\psi}_{d_1}(t-k)\,, \quad k = 1,2,\ldots,r\,.$$

Da notare che la stabilità di $C_\theta(z^{-1})A_\theta(z^{-1})$ è essenziale per poter portare avanti i calcoli.

5. L'algoritmo (104) richiede il calcolo dell'inversa $H_{\theta_k}^{-1}$ ad ogni iterazione. Questo calcolo si può organizzare in forma ricorsiva (in t). Ricordando che $P_{\theta_k}(t) := H_{\theta_k}^{-1}(t)$, e

$$H_{\theta_k}(t+1) := \sum_{s=1}^{t} \boldsymbol{\psi}_{\theta_k}(s)\boldsymbol{\psi}_{\theta_k}(s)^\top + \boldsymbol{\psi}_{\theta_k}(t+1)\boldsymbol{\psi}_{\theta_k}(t+1)^\top$$

usando il lemma di inversione di matrice si trova la relazione ricorsiva in t,

$$P_{\theta_k}(t+1) = P_{\theta_k}(t) - P_{\theta_k}(t)\,\psi_{\theta_k}(t+1)\frac{1}{1+\boldsymbol{\psi}_{\theta_k}(t+1)^\top P_{\theta_k}(t)\,\boldsymbol{\psi}_{\theta_k}(t+1)}\boldsymbol{\psi}_{\theta_k}(t+1)^\top P_{\theta_k}(t)$$

che è una ricorsione "alla Riccati". La ricorsione è valida anche quando è presente il termine di regolarizzazione che si può interpretare come una varianza (normalizzata) a priori assegnata come condizione iniziale all'istante $t = 0$. Naturalmente in pratica conviene sempre usare una versione simmetrizzata di questa equazione, vedi [Stima e Filtraggio!].

6. L'algoritmo di Newton si può interpretare come un metodo iterativo per risolvere le equazioni normali di un un problema ai minimi quadrati non lineare. La formula (104) si può interpretare come la risoluzione iterativa di una successione di problemi di stima ai minimi quadrati su dei modelli lineari "incrementali", ciascuno definito dalle equazioni

$$\boldsymbol{\varepsilon}_{\theta-\theta_k}(t) = \mathbf{y}(t) - \hat{\mathbf{y}}_{\theta_k}(t \mid t-1) = -\boldsymbol{\psi}_{\theta_k}^\top(t)\,(\theta-\theta_k) + \mathbf{e}(t), \quad t = 1,2,\ldots,N$$

ovvero, in notazione vettoriale

$$\boldsymbol{\varepsilon}_{\theta-\theta_k} = -\boldsymbol{\Psi}_{\theta_k}(\theta - \theta_k) + \mathbf{e}\,. \tag{106}$$

Questi problemi possono essere risolti *senza formare le equazioni normali* con tecniche del tipo fattorizzazione QR, come visto.

7. Rimane comunque il fatto che qualunque algoritmo di minimizzazione può unicamente portare nella prossimità di *minimi relativi* che potrebbero in linea di principio essere alquanto lontani dal minimo assoluto. Per questo motivo l'inizializzazione; i.e. la scelta di θ_0, è spesso un passo importante e può essere consigliabile far girare l'algoritmo partendo da diverse stime iniziali.

LA DISTRIBUZIONE ASINTOTICA DELLO STIMATORE PEM

Per descrivere la dispersione di uno stimatore attorno al suo valore limite (supponiamo che esita) e eventualmente verificarne l'ottimalità servirebbe conoscere la sua varianza asintotica. Purtroppo questo concetto è delicato da definire perchè è facile andare incontro a banalità. Ad esempio uno stimatore consistente ha per definizione varianza asintotica uguale a zero e questa affermazione ci dice poco o nulla su come si comporta la varianza per $N \to \infty$.

La strada normalmente seguita è quella di caratterizzare la distribuzione limite dello stimatore

CONVERGENZA IN LEGGE

Definition 1 *Una successione di v.a.* $\{\mathbf{x}_n\}$ *(in generale a valori vettoriali),* converge in legge (o in distribuzione) *a* $\mathbf{x}$*, notazione:* $\mathbf{x}_n \xrightarrow{L} \mathbf{x}$*, se la successione delle d.d.p.* $\{F_n\}$*, delle variabili* $\{\mathbf{x}_n\}$ *converge alla d.d.p.* F *di* $\mathbf{x}$*, in tutti i punti* x *in cui* F *è continua.*

La convergenza in legge è estremamente debole. Essa è implicata dalla convergenza in probabilità e quindi anche dalla convergenza in media e dalla convergenza quasi ovunque. Vale il seguente risultato (che riportiamo qui per comodità del lettore)

Proposizione 26 *La successione di v.a.* $\{\mathbf{x}_n\}$ *converge in legge a* $\mathbf{x}$ *se e solo se* $E\, g(\mathbf{x}_n) \to E\, g(\mathbf{x})$ *per tutte le funzioni* g *limitate che sono continue in un insieme di probabilità uno per la d.d.p. di* $\mathbf{x}$*.*

Una conseguenza di questo risultato è che la convergenza in legge implica quella delle funzioni caratteristiche $\phi_n(i\omega) := E\, e^{i\omega \mathbf{x}_n}$ alla $\phi(i\omega) := E\, e^{i\omega \mathbf{x}}$, per ogni $\omega \in \mathbb{R}^*$.

*In realtà la condizione di convergenza delle funzioni caratteristiche è anche sufficiente

Come è ben noto i momenti di una distribuzione di probabilità sono le derivate della funzione caratteristica calcolate in $\omega = 0$. Ovviamente, dalla convergenza delle $\phi_n(i\omega)$ non segue necessariamente quella delle derivate in $\omega = 0$, per cui **la convergenza in legge non implica necessariamente la convergenza dei momenti** (ovviamente quelli che esistono). Quindi in generale medie, varianze, etc., della successione $\{\mathbf{x}_n\}$, non convergono necessariamente a media, varianza etc. del limite.

L'implicazione però vale per successioni che soddisfano la condizione seguente.

Lemma 4 *Se $\{\mathbf{x}_n\}$ è una successione di variabili aleatorie convergente in legge; i.e. $\mathbf{x}_n \xrightarrow{L} \mathbf{x}$, per cui*

$$\sup_n \|\mathbf{x}_n\|^2 < \infty \tag{107}$$

allora tutti i momenti che esistono delle $\mathbf{x}_n$ convergono ai rispettivi momenti della distribuzione limite.

per (e quindi *equivalente* a) la convergenza in distribuzione (teorema di Helly-Bray).

TEOREMA DI SLUTSKY

Teorema 22 (Slutsky) *Si assuma che la sequenza di vettori aleatori n-dimensionali $\{\mathbf{x}_N; N = 1,2,\ldots\}$ converga in legge a $\mathbf{x}$ (ovvero $\mathbf{x}_N \xrightarrow{L} \mathbf{x}$). Allora:*

1. *Se $\{\mathbf{y}_N\}$ è una successione di vettori aleatori per cui $(\mathbf{x}_N - \mathbf{y}_N) \to 0$ in probabilità, allora anche $\mathbf{y}_N$ converge in legge a $\mathbf{x}$ (ovvero $\mathbf{y}_N \xrightarrow{L} \mathbf{x}$).*

2. *Se $f : \mathbb{R}^n \to \mathbb{R}^k$ è una funzione continua, allora $f(\mathbf{x}_N) \xrightarrow{L} f(\mathbf{x})$.*

3. *In particolare, se $\mathbf{x}_N = [\mathbf{z}'_N \, \mathbf{y}'_N]'$ dove la sequenza di vettori aleatori m-dimensionali $\{\mathbf{y}_N; N = 1,2,\ldots\}$ converge in legge (o in probabilità) ad una costante c e se $f(x) := f(z,y) : \mathbb{R}^{p+m} \to \mathbb{R}^k$ è una funzione continua nei due argomenti, allora $f(\mathbf{z}_N, \mathbf{y}_N) \xrightarrow{L} f(\mathbf{z}, c)$.*

Due successione di vettori aleatori $\{\mathbf{x}_N\}$ e $\{\mathbf{y}_N\}$ per cui $(\mathbf{x}_N - \mathbf{y}_N) \to 0$ in probabilità, si dicono **asintoticamente equivalenti**.

IL TEOREMA DEL LIMITE CENTRALE

Il *teorema del limite centrale (TLC)* fu scoperto da De Moivre e Laplace per variabili discrete alla fine del settecento e successivamente esteso da Gauss al caso di variabili continue indipendenti. La versione "classica" riguarda la convergenza della distribuzione di somme di variabili aleatorie *indipendenti e identicamente distribuite (i.i.d)* ad una distribuzione Gaussiana.

Cosa si intende per distribuzione limite? Consideriamo un processo $\mathbf{y}$ a variabili i.i.d. (rumore bianco "in senso stretto") di media μ e varianza Σ. Per il teorema ergodico, la media campionaria $\bar{\mathbf{y}}_T = \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}(t)$ converge alla media $\mu = E\,\mathbf{y}(t)$ con probabilità uno per $T \to \infty$. Ovviamente la varianza di $\bar{\mathbf{y}}_T$ deve quindi convergere a zero e la **distribuzione limite è quindi degenere**.

Notiamo che la varianza di $\bar{\mathbf{y}}_T$ *tende a zero esattamente come* $\frac{1}{T}$:

$$\mathrm{Var}\,(\bar{\mathbf{y}}_T) = \frac{1}{T}\left(\frac{1}{T}\sum_{t=1}^{T}\mathrm{Var}\,(\mathbf{y}(t))\right) = \frac{1}{T}\Sigma\,.$$

In altri termini, per $T \to \infty$, la varianza di $\sqrt{T}\,\bar{\mathbf{y}}_T$ ha un limite finito: la varianza di $\mathbf{y}(t)$. La distribuzione limite (se esiste) non può essere degenere.

Teorema (TLC classico): la distribuzioni di probabilità delle variabili $\sqrt{T}\,\bar{\mathbf{y}}_T$ converge ad una distribuzione Gaussiana di media μ e varianza Σ.

Prova: espansione attorno a $\omega = 0$ della funzione caratteristica della convoluzione di T distribuzioni di probabilità uguali... (si può trovare in quasi tutti i testi di teoria della probabilità). □

N.B.: Non serve dimostrare che la varianza converge....

Esempio 6 *Sia* $\mathbf{y}$ *un processo i.i.d. scalare a media* μ*, varianza* σ^2 *per cui vale la* $\sqrt{N}\bar{\mathbf{y}}_N \xrightarrow{L} \mathcal{N}(\mu, \sigma^2)$ *(teorema del limite centrale). Trovare la distribuzione asintotica della statistica*

$$\varphi_N(\mathbf{y}) := \frac{\sqrt{N}\bar{\mathbf{y}}_N - \mu}{\sqrt{\hat{\sigma}_N^2(\mathbf{y})}}$$

dove $\hat{\sigma}_N^2(\mathbf{y})$ *è la varianza campionaria*

$$\hat{\sigma}_N^2(\mathbf{y}) = \frac{1}{N}\sum_{t=1}^{N}(\mathbf{y}(t) - \bar{\mathbf{y}}_N)^2$$

Soluzione: Per l'ergodocità $\hat{\sigma}_N^2(\mathbf{y}) \to \sigma^2$ per $N \to \infty$ (con probabilità uno e quindi anche in probabilità). *Usando il teorema di Slutsky (punto 3), si vede subito che*

$$\varphi_N(\mathbf{y}) \xrightarrow{L} \mathcal{N}(0,1)\,.$$

La versione del teorema del limite centrale (TLC) per processi i.i.d. è stata generalizzata in letteratura al caso di successioni di variabili (o vettori) aleatori indipendenti che non hanno necessariamente la stessa distribuzione di probabilità (ad esempio la varianza di $\mathbf{y}(t)$ può in generale dipendere da t).

Per le applicazioni che abbiamo in vista (analisi asintotica degli stimatori) noi considereremo solo processi stazionari $\{\mathbf{y}(t)\}$. A noi però interessano processi le cui variabili sono dipendenti, quelli che si incontrano quando si descrivono segnali di interesse nell'ingegneria.

Se il processo non è a variabili indipendenti occorrono delle condizioni perchè valga il teorema del limite centrale. Prima di occuparci del caso generale, conviene discutere un caso notevole di processi stocastici, detti **d-martingale**, in cui la dimostrazione del TLC è sostanzialmente analoga a quella del caso i.i.d..

ASPETTAZIONE CONDIZIONATA

Consideriamo lo spazio di Hilbert $L^2(\mathbf{y})$, di tutte le statistiche (non-lineari) di un processo $\mathbf{y}$ che hanno momento del secondo ordine finito, i.e. funzioni $f(\mathbf{y}(s)\,; s \in \mathbb{Z})$ con $\mathbb{E}\, f(\mathbf{y})^2 < \infty$.

Sia $\mathbf{x}$ una variabile con momento del secondo ordine finito. Vogliamo trovare la miglior statistica dei dati in $L^2(\mathbf{y})$ che approssima $\mathbf{x}$ in modo ottimo nel senso che l'errore quadratico

$$\mathbb{E}\,[\mathbf{x} - f(\mathbf{y})]^2, \qquad f(\mathbf{y}) \in L^2(\mathbf{y})$$

è minimo. La soluzione è la proiezione ortogonale di $\mathbf{x}$ su $L^2(\mathbf{y})$.

La proiezione ortogonale di $\mathbf{x}$ su $L^2(\mathbf{y})$ è la media condizionata $\mathbb{E}\,(\mathbf{x} \mid \mathbf{y})$

Principio di ortogonalità

$$\mathbf{x} - \mathbb{E}\,(\mathbf{x} \mid \mathbf{y}) \perp L^2(\mathbf{y}) \tag{108}$$

ESEMPI DI D-MARTINGALE

La condizione di d-martingala è più debole di quella di processo i.i.d.; a metà tra processo a variabili scorrelate e i.i.d.
Un esempio "canonico" di d-martingala è l'errore di predizione di un passo di un processo stazionario $\mathbf{y}$, quando si intende che la predizione è quella ottima non lineare, ovvero è la **media condizionata** di $\mathbf{y}(t)$, data la storia passata $L_t^2(\mathbf{y}) := \{ f(\mathbf{y}^t) \mid \mathbb{E} f(\mathbf{y}^t)^2 < \infty \}$,

$$\mathbf{z}(t) := \tilde{\mathbf{y}}(t) = \mathbf{y}(t) - \mathbb{E}\{\mathbf{y}(t) \mid L_{t-1}^2(\mathbf{y})\} \qquad t \in \mathbb{Z}.$$

Prendiamo $L^2(\mathbf{y})$ invece di $L^1(\mathbf{y})$ perchè $L^2(\mathbf{y})$ è una spazio di Hilbert e si può definire la proiezione ortogonale.
$\mathbf{z}$ ha le proprietà seguenti:
$\mathbf{z}(t) \in L_t^2(\mathbf{y})$,
$\mathbf{z}(t+1)$ è scorrelata da tutte le variabili in $L_t^2(\mathbf{y})$ ovvero

$$\mathbb{E}\{\mathbf{z}(t+1) \mid L_t^2(\mathbf{y})\} = 0 \qquad t \in \mathbb{Z}.$$

Dato che $\mathbb{E}\{\mathbf{z}(t) \mid L^2_{t-1}(\mathbf{y})\} = 0$

$$\mathbb{E}\mathbf{z}(t) = \mathbb{E}\{\mathbb{E}[\mathbf{z}(t) \mid L^2_{t-1}(\mathbf{y})]\} = 0$$

e quindi $\mathbb{E}\{\mathbf{z}(t)\mathbf{z}(t)^\top) = \mathrm{Var}(\mathbf{z}(t))$
Se la varianza condizionata $\mathbb{E}\{\mathbf{z}(t)\mathbf{z}(t)^\top \mid L^2_{t-1}(\mathbf{y})\}$ non dipende da $\mathbf{y}$, ovvero

$$\mathbb{E}\{\mathbf{z}(t)\mathbf{z}(t)^\top \mid L^2_{t-1}(\mathbf{y})\} = \Sigma_{\mathbf{z}} < \infty$$

allora la d-martingala ha anche varianza costante. Diciamo che è *stazionaria*.

Più in generale si può considerare l'errore di predizione di un passo di $\mathbf{y}$ quando l'informazione disponibile proviene anche dall'osservazione di una variabile "esogena" $\mathbf{u}$. In questo caso si condiziona rispetto all' aggregato delle funzioni (a momento secondo finito) della storia passata $(\mathbf{y}^t, \mathbf{u}^t)$

$$\mathbf{z}(t) := \tilde{\mathbf{y}}(t) = \mathbf{y}(t) - \mathbb{E}\{\mathbf{y}(t) \mid \mathbf{y}^{t-1}, \mathbf{u}^{t-1}\} \qquad t \in \mathbb{Z}$$

(dove la media condizionata è ancora intesa in senso stretto). È evidente che, il processo $\mathbf{z}$ soddisfa ancora le due condizioni della definizione pur di sostituire $L_t^2(\mathbf{y})$ con $L_t^2(\mathbf{y}, \mathbf{u})$.

Una classe ancora più ampia di d-martingale si ottiene considerando processi del tipo

$$\mathbf{z}(t) := \boldsymbol{\varphi}(t)\,\tilde{\mathbf{y}}(t) \qquad \boldsymbol{\varphi}(t) \in L_{t-1}^2(\mathbf{y}, \mathbf{u})$$

in cui $\boldsymbol{\varphi}(t) = \boldsymbol{\varphi}(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})$ è una funzione (misurabile) della storia passata fino all'istante precedente, $t-1$. In questo caso si ha

$$\mathbb{E}\{\mathbf{z}(t) \mid L_{t-1}^2(\mathbf{y}, \mathbf{u})\} = \boldsymbol{\varphi}(t)\,\mathbb{E}\{\tilde{\mathbf{y}}(t) \mid L_{t-1}^2(\mathbf{y}, \mathbf{u})\} = 0 \qquad t \in \mathbb{Z}.$$

Una *martingala* è l'integrale discreto di una d-martingala,

$$\mathbf{x}(t) = \mathbf{x}(0) + \sum_{s=1}^{t} \mathbf{z}(s) \tag{109}$$

ed è un processo non stazionario che è la generalizzazione del processo di passeggiata aleatoria.

Notazione Gli esempi si possono generalizzare pensando che invece di $L_t^2(\mathbf{y}, \mathbf{u})\}$ si possono considerare spazi più grando contenenti ad es. anche variabili di rumore non osservabili. Pensiamo in astratto di avere una famiglia crescente di statistiche $\{\mathcal{F}_t\}$ (tutte a momento secondo finito) senza dover specificare esattamente quali sono i processi da cui dipendono.

LE D-MARTINGALE

Definizione 23 *Sia* $\{\mathcal{F}_t\,; t \in \mathbb{Z}\}$ *una successione crescente di sottospazi di* L^2*, i.e.* $\mathcal{F}_t \subset \mathcal{F}_{t+1}$*. Il processo stocastico* $\{\mathbf{z}(t)\,; t \in \mathbb{Z}\}$ *è una* martingala differenza*, o brevemente, una* d-martingala *rispetto alla famiglia* $\{\mathcal{F}_t\}$*, se,*

◦ *Per ogni* t*,* $\mathbf{z}(t) \in \mathcal{F}_t\,; t \in \mathbb{Z}$*,*

◦ $\mathbf{z}(t+1)$ *è scorrelata da tutte le variabili in* $\mathcal{F}_t$ *ovvero*

$$\mathbb{E}\{\mathbf{z}(t+1) \mid \mathcal{F}_t\} = 0 \qquad t \in \mathbb{Z}\,.$$

Notiamo che la seconda condizione è equivalente alla

$$\mathbb{E}\{\mathbf{z}(t) \mid \mathcal{F}_s\} = \mathbb{E}\{\mathbb{E}\left[\mathbf{z}(t) \mid \mathcal{F}_{t-1}\right] \mid \mathcal{F}_s\} = 0 \qquad \forall s < t\,.$$

In particolare una d-martingala ha sempre media zero.
Se la famiglia $\{\mathcal{F}_t\}$ *è stazionaria* $\mathbf{z}$ *è un processo stazionario e la varianza condizionata* $\mathbb{E}\{\mathbf{z}(t)\mathbf{z}(t)^\top \mid \mathcal{F}_{t-1}\}$ *non dipende da* $\mathcal{F}_{t-1}$*, ovvero*

$$\mathbb{E}\{\mathbf{z}(t)\mathbf{z}(t)^\top \mid \mathcal{F}_{t-1}\} = \Sigma_{\mathbf{z}} < \infty$$

Il lemma seguente generalizza alle d-martingale la proprietà "additiva" della varianza di somme di variabili aleatorie indipendenti (o scorrelate).

Lemma 5 *Per ogni d-martingala* $\mathbf{z}$ *a varianza finita si ha*

$$\mathrm{Var}\{\sum_{t=1}^{T}\mathbf{z}(t)\} = \sum_{t=1}^{T}\mathrm{Var}\{\mathbf{z}(t)\} \tag{110}$$

Se la d-martigala è stazionaria il secondo membro vale $T\,\Sigma_{\mathbf{z}}$.

Prova: Facciamo la dimostrazione per il caso scalare. Il caso vettoriale è identico modulo le ovvie complicazioni nelle notazioni. Si ha

$$\begin{aligned}
\mathbb{E}\{\sum_{t=1}^{T}\mathbf{z}(t)\}^2 &= \mathbb{E}\{\mathbf{z}(1)^2+\mathbf{z}(2)^2+\ldots+\mathbf{z}(T)^2\}+ \\
&+ \mathbb{E}\{2\sum_{t>s}\mathbf{z}(t)\mathbf{z}(s)\} = \\
&= \sum_{t=1}^{T}\mathrm{Var}\{\mathbf{z}(t)\}+2\sum_{t>s}\mathbb{E}\,\mathbf{z}(t)\mathbf{z}(s)
\end{aligned}$$

L'ultimo termine è zero giacchè se $t > s$, $\mathbf{z}(s) \in \mathcal{F}_s$, e si può scrivere

$$\mathbb{E}\mathbf{z}(t)\mathbf{z}(s) = \mathbb{E}\{\mathbb{E}\left[\mathbf{z}(t)\mathbf{z}(s) \mid \mathcal{F}_s\right]\} = \mathbb{E}\{\mathbb{E}\left[\mathbf{z}(t) \mid \mathcal{F}_s\right]\mathbf{z}(s)\} = 0$$

in virtù della proprietà di d-martingala. □

IL TLC PER LE D-MARTINGALE

Su questo risultato, formulato inizialmente da P. Levy e J.L. Doob, dimostrato da Billingsley e Ibragimov e successivamente generalizzato da vari autori, poggia la dimostrazione della normalità asintotica dei metodi di identificazone PEM.

Teorema 23 *Sia* $\{\mathbf{z}(t)\}$ *una d-martingala stazionaria a varianza finita,* $\Sigma_{\mathbf{z}} = \mathbb{E}\mathbf{z}(t)\mathbf{z}(t)^{\top}$*. Si ha allora*

$$\sqrt{T}\bar{\mathbf{z}}_T \xrightarrow{L} \mathcal{N}(0, \Sigma_{\mathbf{z}}) \tag{111}$$

ovvero, la statistica $\sqrt{T}\bar{\mathbf{z}}_T$ *converge in legge alla distribuzione Gaussiana di media zero e varianza* $\Sigma_{\mathbf{z}}$*.*

Prova (caso scalare): Notiamo che la funzione caratteristica *condizionata* di ciascuna variabile $\mathbf{z}(t)$ ammette derivata seconda in zero uguale alla varianza (condizionata), σ^2, di $\mathbf{z}(t)$ e pertanto si può scrivere

$$\mathbb{E}\left[e^{i\omega\mathbf{z}(t)} \mid \mathcal{F}_{t-1}\right] = \mathbb{E}\left[1 + i\omega\mathbf{z}(t) - \frac{\omega^2}{2}\mathbf{z}(t)^2 + \boldsymbol{\eta}(\omega, \mathbf{z}(t)) \mid \mathcal{F}_{t-1}\right] = 1 - \frac{\sigma^2\omega^2}{2} + o(\omega^2)$$

dove $o(\omega^2)$ è una variabile aleatoria in $\mathcal{F}_{t-1}$ che tende a zero con ω più rapidamente di ω^2. Detta $\phi_T(\omega)$ la funzione caratteristica della somma $\mathbf{x}(T) := \sum_{t=1}^{T} \mathbf{z}(t)$, si ha

$$\begin{aligned}
\phi_T(\omega) &= \mathbb{E}\left\{ \mathbb{E}\left[e^{i\omega \mathbf{z}(T)} \mid \mathcal{F}_{T-1} \right] e^{i\omega \mathbf{x}(T-1)} \right\} = \\
&= \left[1 - \frac{\sigma^2 \omega^2}{2} \right] \mathbb{E}\{ e^{i\omega \mathbf{x}(T-1)} \} + \mathbb{E}\{ o(\omega^2) e^{i\omega \mathbf{x}(T-1)} \} = \\
&= \left[1 - \frac{\sigma^2 \omega^2}{2} \right] \phi_{T-1}(\omega) + \bar{o}(\omega^2)
\end{aligned}$$

dove $\bar{o}(\omega^2)$ è l'aspettazione di una variabile aleatoria in $\mathcal{F}_{T-1}$ che ha lo stesso modulo di $o(\omega^2)$ e tende quindi a zero con ω più rapidamente di ω^2. Risolvendo l'equazione alle differenze si trova

$$\phi_T(\omega) = \left[1 - \frac{\sigma^2 \omega^2}{2} \right]^T + \bar{o}_T(\omega^2)$$

dove $\bar{o}_T(\omega^2)$ è ancora un infinitesimo di ordine superiore al secondo in ω.

Ora, è immediato che la funzione caratteristica di $\mathbf{s}(T) := \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbf{z}(t)$ è

la ϕ_T appena trovata calcolata in $\omega/\sqrt{T}$, per cui

$$\phi_T(\frac{\omega}{\sqrt{T}}) = \left[1 - \frac{\sigma^2}{2}\frac{\omega^2}{T}\right]^T + \bar{o}_T(\frac{\omega^2}{T})$$

In questa espressione il secondo termine tende a zero per $T \to \infty$, qualunque sia il valore di ω fissato, mentre il limite del primo addendo è $\exp\{-\frac{\sigma^2\omega^2}{2}\}$. Quindi la funzione caratteristica di $\mathbf{s}(T)$ converge a quella della Gaussiana $\mathcal{N}(0, \sigma^2)$. □

Dobbiamo per onestà avvisare il lettore del fatto che questa dimostrazione non è completamente rigorosa. Infatti abbiamo sorvolato su alcune questioni tecniche che riguardano la prova del fatto che i termini d'errore "integrati" $\bar{o}(\omega^2)$ *e* $\bar{o}_T(\omega^2)$ *sono ben definiti e tendono effettivamente a zero. Purtroppo nelle dimostrazioni rigorose che si trovano in letteratura l'argomento intuitivo che abbiamo usato è molto poco riconoscibile. Ci accontenteremo pertanto della "dimostrazione" che abbiamo dato.*

EFFICIENZA ASINTOTICA

Il concetto di *stimatore asintoticamente efficiente* ovvero di stimatore *asintoticamente a minima varianza* è abbastanza delicato da definire in modo preciso. La difficoltà risiede nel fatto che la varianza di quasi tutti gli stimatori interessanti nelle appplicazioni (che sono *consistenti*) deve tendere a zero al crescere della numerosità campionaria ed è evidente che, interpretando in modo letterale la nozione di "varianza asintotica", si trovano delle banalità. Le quantità che si devono confrontare sono *andamenti asintotici* della varianza.

Definizione 24 *Sia* $\{\phi_T(\mathbf{y}); T = 1,2,\ldots\}$ *una successione di stimatori e* $d(T)$ *una funzione di* T *crescente e strettamente posititiva. Diremo che* $\phi_T(\mathbf{y})$ *ha* varianza asintotica Σ *se*

$$\sqrt{d(T)}\,\phi_T(\mathbf{y}) \xrightarrow{L} D(\mu,\Sigma)$$

dove $D(\mu,\Sigma)$ *è una d.d.p. di media* μ *e varianza* Σ*, finita e* strettamente definita positiva*.*

In sostanza per T "grandi" la varianza della distribuzione di $\phi_T(\mathbf{y})$ si può approssimare con l'espressione $\frac{1}{d(T)}\Sigma$.

Da notare che la condizione di positività $\Sigma > 0$ nella definizione è essenziale perchè esclude che ci possano essere combinazioni lineari di componenti di $\phi_T(\mathbf{y})$ che hanno varianza asintotica nulla, il che significa che l'ordine di infinitesimo della varianza di queste combinazioni è diverso da $O(\frac{1}{d(T)})$.

Ricordiamo anche che SE la convergenza in distribuzione implica la convergenza dei momenti, la varianza asintotica può anche essere definita come il limite

$$\lim_{T\to\infty} \text{Var}\left[\sqrt{d(T)}\,\phi(\mathbf{y}_T)\right] = \Sigma\,. \tag{112}$$

Diremo allora che lo stimatore $\phi_T(\mathbf{y})$ è asintoticamente *efficiente* se la sua varianza asintotica è la più piccola possibile. In particolare, se lo stimatore è consistente si può dire che è asintoticamente efficiente se la sua varianza asintotica è uguale all'inversa della matrice (asintotica) di Fisher.

Come vedremo, la varianza di un'ampia classe di stimatori tende a zero come $1/T$. In questi casi, si può mostrare che la matrice di Fisher di un campione di numerosità T è proporzionale a T e quindi (assumendo identificabilità locale) la sua inversa ha la forma $\frac{1}{T} I(\theta)^{-1}$. In questi casi l'efficienza si esprime attraverso l'uguaglianza

$$\Sigma = I(\theta)^{-1}.$$

In molti testi di statistica, l'efficienza è in realtà definita solo per stimatori che hanno una distribuzione limite Gaussiana con velocità di convergenza $1/d(T)$ proporzionale a $1/T$.

IL TLC PER LO LO STIMATORE PEM

Supporremo sempre di essere nelle condizioni che garantiscono la consistenza dello stimatore PEM, ovvero, supporemo (almeno) che i dati siano ergodici del secondo ordine, il modello vero appartenga alla classe parametrica $\{M(\theta)\}$ e che vi sia identificabilità. Ricordiamo anche che i modelli che consideriamo son modelli per i quali il predittore di un passo è una funzione razionale (e quindi analitica) del parametro θ per cui il minimo della cifra di merito $V_N(\theta)$ si ha in un punto in cui il gradiente si annulla, ovvero

$$\frac{\partial V_N(\theta)}{\partial \theta} := V_N(\theta)' = 0\,.$$

Usando la formula di Taylor arrestata al secondo ordine nel punto $\theta = \theta_0$, si ha

$$V_N(\hat{\boldsymbol{\theta}}_N)' = V_N(\theta_0)' + V_N(\bar{\boldsymbol{\theta}})''(\hat{\boldsymbol{\theta}}_N - \theta_0) = 0 \tag{113}$$

dove $V_N(\bar{\boldsymbol{\theta}})''$ è la matrice delle derivate seconde (Hessiana) calcolata in un punto $\bar{\theta}$ dell'intervallo p-dimensionale di estremi θ_0 e $\hat{\boldsymbol{\theta}}_N$, ovvero

$$\theta_0^k \leq \bar{\boldsymbol{\theta}}^k \leq \hat{\boldsymbol{\theta}}_N^k, \qquad k = 1, 2, \ldots, p$$

Supponendo che la matrice Hessiana sia invertibile, dalla (113) si può ricavare

$$\hat{\boldsymbol{\theta}}_N - \theta_0 = -\left[\frac{1}{2} V_N(\bar{\boldsymbol{\theta}})''\right]^{-1} \frac{1}{2} V_N(\theta_0)' \tag{114}$$

dove il fattore $\frac{1}{2}$ è stato introdotto per convenienza.

Calcoliamo ora il gradiente e la matrice Hessiana usando l'espressione di $V_N(\theta)$. Ponendo

$$\boldsymbol{\psi}_\theta(t) := \frac{\partial \boldsymbol{\varepsilon}_\theta(t)}{\partial \theta} = -\frac{\partial \hat{\mathbf{y}}_\theta(t \mid t-1)}{\partial \theta}$$

si trova

$$\frac{1}{2} V_N(\theta)' = \frac{1}{N} \sum_{t=1}^{N} \boldsymbol{\psi}_\theta(t) \boldsymbol{\varepsilon}_\theta(t) \tag{115}$$

$$\frac{1}{2} V_N(\theta)'' = \frac{1}{N} \sum_{t=1}^{N} \left\{ \boldsymbol{\psi}_\theta(t) \boldsymbol{\psi}_\theta(t)^\top + \boldsymbol{\varepsilon}_\theta(t) \left[\frac{\partial^2 \boldsymbol{\varepsilon}_\theta(t)}{\partial \theta_i \partial \theta_j} \right] \right\} \tag{116}$$

Esaminiamo prima il comportamento asintotico della derivata seconda.

Lemma 6 *Nelle ipotesi poste, si ha*

$$\lim_{N \to \infty} \frac{1}{2} V_N(\bar{\boldsymbol{\theta}})'' = \mathbb{E}_{\theta_0} \left\{ \boldsymbol{\psi}_{\theta_0}(t) \boldsymbol{\psi}_{\theta_0}(t)^\top \right\} \tag{117}$$

con probabilità uno.

Prova : Nelle ipotesi in cui ci siamo messi, $\hat{\boldsymbol{\theta}}_N \to \theta_0$ e quindi anche $\bar{\boldsymbol{\theta}} \to \theta_0$ (con probabiltà uno) e la media temporale in (116) tende all 'aspettazione per cui,

$$\frac{1}{2}V_N(\bar{\boldsymbol{\theta}})'' \to \mathbb{E}_{\theta_0}\left\{\boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^\top + \boldsymbol{\varepsilon}_{\theta_0}(t)\left[\frac{\partial^2 \boldsymbol{\varepsilon}_\theta(t)}{\partial\theta_i\partial\theta_j}\right]_{|\theta=\theta_0}\right\}$$

Inoltre, dato che il modello vero appartiene alla classe dei modelli assegnata, si deve avere $\boldsymbol{\varepsilon}_{\theta_0}(t) = \mathbf{e}_0(t)$. Infine, dato che sia il gradiente $(\boldsymbol{\psi}_\theta(t))$, che la derivata seconda di $\hat{\mathbf{y}}_\theta(t \mid t-1)$ sono necessariamente funzioni (lineari) solo dei dati passati $(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})$, tutti gli elementi nella matrice delle derivate seconde a secondo membro risultano scorrelati da $\mathbf{e}_0(t)$ e l'ultimo termine ha quindi aspettazione nulla. □

Per quanto riguarda l'altro termine nel prodotto (114), si ha

$$\frac{1}{2}V_N(\theta_0)' = \frac{1}{N}\sum_{t=1}^{N}\boldsymbol{\psi}_{\theta_0}(t)\mathbf{e}_0(t) \tag{118}$$

Proposizione 27 *Se il processo innovazione nel modello vero, $\mathbf{e}_0$, è una d-martingala stazionaria rispetto alla famiglia crescente generata dai dati passati $(\mathbf{y}^t, \mathbf{u}^t)$ e ha varianza finita, allora anche il processo $\{\boldsymbol{\psi}_{\theta_0}(t)\mathbf{e}_0(t)\}$ è una d-martingala e vale il teorema del limite centrale,*

$$\sqrt{N}\frac{1}{2}V_N(\theta_0)' \xrightarrow{L} \mathcal{N}(0, Q) \tag{119}$$

Se la varianza condizionata di $\mathbf{e}_0(t)$ è indipendente dai dati $(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})$, ovvero se

$$\mathbb{E}_0\{\mathbf{e}_0(t)^2 \mid \mathbf{y}^{t-1}, \mathbf{u}^{t-1}\} = \mathbb{E}_0\{\mathbf{e}_0(t)^2\} = \lambda_0^2, \tag{120}$$

la matrice varianza asintotica Q è data dalla formula,

$$Q = \lambda_0^2\, \mathbb{E}_0\{\boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^\top\}. \tag{121}$$

Prova : Il risultato scende dall'osservazione che $\{\boldsymbol{\psi}_{\theta_0}(t)\mathbf{e}_0(t)\}$ è ancora una d-martingala rispetto allo stesso flusso informativo $\{\mathbf{y}^t, \mathbf{u}^t\}$ ed è in realtà un corollario immediato del teorema del limite centrale per d-martingale

23. L'unica cosa da dimostrare è l'espressione per la varianza asintotica, la quale scende dalla proprietà (120), che implica,

$$\begin{aligned}\mathrm{Var}\{\boldsymbol{\psi}_{\theta_0}(t)\mathbf{e}_0(t)\} = & \quad \mathbb{E}_0\{\mathbb{E}_0\left[\mathbf{e}_0(t)^2 \mid \mathbf{y}^{t-1}, \mathbf{u}^{t-1}\right]\}\boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^\top\} = \\ = & \quad \mathbb{E}_0\{\mathbf{e}_0(t)^2\}\,\mathbb{E}_0\{\boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^\top\}.\end{aligned}$$

□

Osservazione 1 *Notiamo che le condizioni del teorema valgono in particolare se* $\mathbf{e}_0$ *è un processo i.i.d. ma in realtà la condizione di d-martingala è molto più debole. Essa si può riformulare in questo contesto dicendo che per il processo di osservazione descritto dal modello vero,* **il predittore non lineare (la media condizionata) di** $\mathbf{y}(t)$ **data la storia passata** $(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})$**, è una funzione lineare dei dati***.*

Infatti in questo caso l'errore di predizione in senso stretto, $\mathbf{y}(t) - \mathbb{E}[\mathbf{y}(t) \mid \mathbf{y}^{t-1}, \mathbf{u}^{t-1}]$ *(che è una d-martingala) coincide necessariamente con l'usuale l'innovazione (in senso debole)* $\mathbf{e}_0(t)$*. Per modelli che descrivono piccole variazioni dei segnali in gioco è ragionevole pensare che il predittore (in senso stretto) si possa in genere approssimare con una funzione lineare dei dati passati.*

LA DISTRIBUZIONE ASINTOTICA DELLO STIMATORE PEM

Questo è il secondo risultato fondamentale della teoria asintotica della stima PEM.

Teorema 24 *Supponiamo che i dati siano ergodici del secondo ordine, il modello vero appartenga alla classe parametrica* $\{M(\theta)\}$ *e che vi sia identificabilità (locale) in* θ_0*. Assumiamo che il processo innovazione* $\mathbf{e}_0$ *sia una d-martingala stazionaria a varianza finita. Allora per lo stimatore PEM vale il teorema del limite centrale,*

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \theta_0) \stackrel{L}{\to} \mathcal{N}(0, P)\,, \tag{122}$$

dove la varianza asintotica P *è data dalla formula*

$$P = \lambda_0^2 \left[\mathbb{E}_0\{\boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^\top\}\right]^{-1} \tag{123}$$

e l'inversa della matrice tra parentesi quadre esiste.

Prova : Il risultato scende dalla (114). Per $N \to \infty$

$$\hat{\boldsymbol{\theta}}_N - \theta_0 \simeq - \left[\frac{1}{N} \sum_{t=1}^{N} \boldsymbol{\psi}_{\bar{\boldsymbol{\theta}}}(t) \boldsymbol{\psi}_{\bar{\boldsymbol{\theta}}}(t)^\top \right]^{-1} \frac{1}{N} \sum_{t=1}^{N} \boldsymbol{\psi}_{\theta_0}(t) \boldsymbol{\varepsilon}_{\theta_0}(t)$$

si usa la terza affermazione del teorema di Slutsky. L'espressione per la varianza si ottiene notando che la distribuzione limite ha come matrice varianza

$$P = \left[\mathbb{E}_{\theta_0} \left\{ \boldsymbol{\psi}_{\theta_0}(t) \boldsymbol{\psi}_{\theta_0}(t)^\top \right\} \right]^{-1} Q \left[\mathbb{E}_{\theta_0} \left\{ \boldsymbol{\psi}_{\theta_0}(t) \boldsymbol{\psi}_{\theta_0}(t)^\top \right\} \right]^{-1}$$

dove Q è la varianza asintotica del limite (119). La questione dell'invertibilità verrà chiarita più avanti. □

Esempio 7 *Vogliamo identificare il sistema "vero":*

$$\mathbf{y}(t) = b_0\mathbf{u}(t-1) + \frac{1}{1+a_0 z^{-1}}\mathbf{e}_0(t)\,, \qquad (*)$$

dove $\mathbf{u}$ *ed* $\mathbf{e}_0$ *sono rumori bianchi scorrelati di media zero e varianze* σ^2 *e* λ_0^2*, usando due possibili classi di modelli:*

- *Modelli di tipo Box-Jenkins della forma:*

$$M_1(\theta) := \{\mathbf{y}(t) = b\mathbf{u}(t-1) + \frac{1}{1+az^{-1}}\mathbf{e}(t)\,;\, \theta = [a\,b]^\top\}$$

- *Modelli ARX:*

$$M_2(\theta) := \{\mathbf{y}(t) + a\mathbf{y}(t-1) = b_1\mathbf{u}(t-1) + b_2\mathbf{u}(t-2) + \mathbf{e}(t)\,;\, \theta = [a\ b_1\ b_2]^\top\}$$

entrambi senza reazione. Confrontatre i risultati in termini di varianze delle stime.

Soluzione:

Identificazione con il metodo PEM usando modelli del tipo Box-Jenkins: Il modello vero appartiene alla classe M_1 e si ha identificabilità, quindi lo stimatore PEM è consistente e $\hat{\theta}_N = [\hat{a}_N \, \hat{b}_N]^\top$ converge al parametro vero $\theta_0 = [a_0 \, b_0]^\top$.

Identificazione con il metodo PEM usando modelli del tipo ARX: Il modello vero appartiene anche alla classe M_2 e si ha identificabilità, quindi lo stimatore PEM è anch'esso consistente e $\hat{\theta}_N$ converge al parametro vero che per questo modello è $\theta_0 = [a_0 \;\; b_0 \;\; b_0 a_0]^\top$. In altri termini, quando $N \to \infty$, $\hat{b}_{2,N} \to b_0 a_0$.

Calcolo della varianza asintotica dei due stimatori.

- Per i modelli Box-Jenkins:

$$\varepsilon_{\theta}(t) = (1 + az^{-1})[\mathbf{y}(t) - b\mathbf{u}(t-1)] = (1 + az^{-1})[\mathbf{y}(t) - b\mathbf{u}(t-1)]$$

Il gradiente si calcola facilmente:

$$\psi_\theta(t)^\top = [\frac{\partial \varepsilon_\theta(t)}{\partial a} \; \frac{\partial \varepsilon_\theta(t)}{\partial b}] = [\mathbf{y}(t-1) - b\mathbf{u}(t-2) \quad -\mathbf{u}(t-1) - a\mathbf{u}(t-2)]$$

e si vede subito che $\mathbf{y}(t-1) - b\mathbf{u}(t-2) = (1+az^{-1})^{-1}\mathbf{e}(t-1)$. Quindi le due componenti del gradiente sono scorrelate. Per calcolare la matrice varianza serve:

$$R := \mathbb{E}_0 \psi_\theta(t)\psi_\theta(t)^\top|_{\{\theta=\theta_0\}} = \begin{bmatrix} \dfrac{\lambda_0^2}{1-a_0^2} & 0 \\ 0 & \sigma^2(1+a_0^2) \end{bmatrix}$$

che porge,

$$\mathrm{Var}\{\hat{\theta}_N\} \sim \frac{\lambda_0^2}{N} R^{-1} = \frac{\lambda_0^2}{N} \begin{bmatrix} \dfrac{1-a_0^2}{\lambda_0^2} & 0 \\ 0 & \dfrac{1}{\sigma^2(1+a_0^2)} \end{bmatrix}.$$

Per i modelli ARX:

$$\varepsilon_\theta(t) = \mathbf{y}(t) + a\mathbf{y}(t-1) - b_1\mathbf{u}(t-1) - b_2\mathbf{u}(t-2)$$

e il gradiente è

$$\psi_\theta(t) = [\frac{\partial \varepsilon_\theta(t)}{\partial a} \; \frac{\partial \varepsilon_\theta(t)}{\partial b_1} \; \frac{\partial \varepsilon_\theta(t)}{\partial b_2}] = [\mathbf{y}(t-1) \;\; -\mathbf{u}(t-1) \;\; -\mathbf{u}(t-2)]$$

Calcolo della varianza di $\mathbf{y}(t-1)$ usando il modello Box-Jenkins $(*)$

$$\mathbb{E}_0\mathbf{y}(t-1)^2 = \operatorname{var}\mathbf{y}(t) = b_0^2\sigma^2 + \frac{\lambda_0^2}{1-a_0^2}$$

per cui

$$R := \mathbb{E}_0\psi_\theta(t)\psi_\theta(t)^\top|_{\{\theta=\theta_0\}} = \begin{bmatrix} \dfrac{b_0^2\sigma^2(1-a_0^2)+\lambda_0^2}{1-a_0^2} & 0 & -b_0\sigma^2 \\ 0 & \sigma^2 & 0 \\ -b_0\sigma^2 & 0 & \sigma^2 \end{bmatrix}$$

e infine

$$\mathrm{Var}\{\hat{\theta}_N\} \sim \frac{\lambda_0^2}{N} R^{-1} = \frac{\lambda_0^2}{N} \begin{bmatrix} \frac{1-a_0^2}{\lambda_0^2} & 0 & \frac{b_0(1-a_0^2)}{\lambda_0^2} \\ 0 & \frac{1}{\sigma^2} & 0 \\ \frac{b_0(1-a_0^2)}{\lambda_0^2} & 0 & \frac{b_0^2(1-a_0^2)}{\lambda_0^2} + \frac{1}{b_0^2} \end{bmatrix}$$

Si osserva che la varianza asintotica di $\hat{b}_1$ è $\frac{\lambda_0^2}{N}\frac{1}{\sigma^2}$ che è **maggiore** di quella stimata con il primo modello: $\frac{\lambda_0^2}{N}\frac{1}{\sigma^2(1+a_0^2)}$ che ha meno parametri.

SUL CALCOLO DELLA VARIANZA DI PROCESSI A SPETTRO RAZIONALE

Questi processi si possono sempre rappresentare con un modello di stato

$$\mathbf{x}(t+1) = A\mathbf{x}(t) + B\mathbf{w}(t)$$
$$\mathbf{y}(t) = C\mathbf{x}(t) + D\mathbf{w}(t)$$

dove A è asintoticamente stabile e $\mathbf{w}(t)$ è bianco, diciamo di varianza W. La varianza di $\mathbf{y}$ si calcola risolvendo l'equazione di Lyapunov

$$\Sigma = A\Sigma A^\top + BWB^\top$$

e poi calcolando $\Sigma_{\mathbf{y}} = C\Sigma C^\top + DWD^\top$

Questo metodo di calcolo si può usare anche per calcolare la covarianza di due processi $\mathbf{y}_1$, $\mathbf{y}_2$ se i modelli di stato hanno in ingresso **lo stesso rumore bianco** ad esempio:

$$\mathbf{x}_i(t+1) = A_i\mathbf{x}_i(t) + B_i\mathbf{w}(t)$$
$$\mathbf{y}_i(t) = C_i\mathbf{x}_i(t) + D_i\mathbf{w}(t), \qquad i = 1,2$$

dove le A_i sono asintoticamente stabili. Moltiplicando $\mathbf{x}_1(t+1)$ per $\mathbf{x}_2(t+1)^\top$ e tendo conto che per la stazionarietà:

$$\mathbb{E}\mathbf{x}_1(t+1)\mathbf{x}_2(t+1)^\top = \Sigma_{1,2} = \mathbb{E}\mathbf{x}_1(t)\mathbf{x}_2(t)^\top$$

si trova

$$\Sigma_{1,2} = A_1\Sigma_{1,2}A_2^\top + B_1 W B_2^\top$$

e infine $\Sigma_{\mathbf{y}_1,\mathbf{y}_2} = C_1\Sigma_{1,2}C_2^\top + D_1 W D_2^\top$

ATTENZIONE: occorre verificare che $\mathbf{x}(t)$ e $\mathbf{w}(t)$ siano scorrelati!

quindi occorre che A oppure A_1 e A_2 siano asintoticamente stabili.

Non conviene mai integrare spettri!!!!

Esercizio: Calcolo della varianza di $\mathbf{y}$ per il modello ARMAX dell'esercizio precedente

$$\mathbf{y}(t) = -a\mathbf{y}(t-1) + b_1\mathbf{u}(t-1) + b_2\mathbf{u}(t-2) + \mathbf{e}(t), \qquad \mathbf{u} \perp \mathbf{e}$$

ovvero:

$$\mathbf{y}(t) = \frac{b_1 z^{-1} + b_2 z^{-2}}{1 + az^{-1}}\mathbf{u}(t) + \frac{1}{1 + az^{-1}}\mathbf{e}(t)$$

dove i due addendi sono completamente scorrelati. Bisogna procurarsi una realizzazione di stato di $F(z)$, ad es :

$$\begin{bmatrix}\mathbf{x}_1(t+1)\\ \mathbf{x}_2(t+1)\end{bmatrix} = \begin{bmatrix}0 & 0\\ 1 & a\end{bmatrix}\begin{bmatrix}\mathbf{x}_1(t)\\ \mathbf{x}_2(t)\end{bmatrix} + \begin{bmatrix}b_2\\ b_1\end{bmatrix}\mathbf{u}(t)$$

$$\mathbf{y}(t) = \begin{bmatrix}0 & 1\end{bmatrix}\mathbf{x}(t)$$

Lo studente risolva Lyapunov e calcoli la varianza di $\mathbf{y}$. Successivamente verifichi l'esattezza del calcolo ponendo $a = a_0$, $b_1 = b_0$ e $b_2 = b_0 a_0$ e confrontando con l'espressione ottenuta con il modello di Box-Jenkins.

Problema 11 *Si identifica con il metodo PEM un modello ARMA (d'innovazione) appartenente alla famiglia*

$$(1+az^{-1})\mathbf{y}(t) = (1+cz^{-1})\mathbf{e}(t)\,.$$

sapendo che $\mathbf{e}_0$ *è un processo i.i.d. a media zero. Si vuole*

1. *Calcolare la varianza asintotica,* $\hat{\sigma}_N^2$ *della differenza* $\hat{a}_N - \hat{c}_N$*,*

2. *Trovare la distribuzione asintotica del rapporto*

$$\mathbf{x}_N := \frac{(\hat{a}_N - \hat{c}_N)^2}{\hat{\sigma}_N^2}$$

Soluzione : È implicito nel testo che il modello vero

$$(1+a_0z^{-1})\mathbf{y}(t) = (1+c_0z^{-1})\mathbf{e}_0(t)$$

appartiene alla famiglia parametrica che si usa per l'identificazione e quindi, nelle ipotesi poste, lo stimatore PEM è consistente e asintoticamente normale. Detto $\theta := [a\ c]^\top$, la varianza asintotica di $\hat{\theta}_N$ è data dalla nota formula

$$\hat{\Sigma}_N = \frac{\lambda_0^2}{N}\left[\mathbb{E}_{\theta_0}\psi_{\theta_0}(t)\psi_{\theta_0}(t)^\top\right]^{-1}$$

dove

$$\psi_{\theta_0}(t) = \frac{1}{1+c_0 z^{-1}}\begin{bmatrix}\mathbf{y}(t-1)\\ -\mathbf{e}_0(t-1)\end{bmatrix} = \begin{bmatrix}\dfrac{1}{1+a_0 z^{-1}}\mathbf{e}_0(t-1)\\ -\dfrac{1}{1+c_0 z^{-1}}\mathbf{e}_0(t-1)\end{bmatrix}$$

da cui si ricava

$$\hat{\Sigma}_N = \frac{\lambda_0^2}{N}\begin{bmatrix}\dfrac{\lambda_0^2}{1-a_0^2} & -\dfrac{\lambda_0^2}{(1-a_0c_0)}\\ -\dfrac{\lambda_0^2}{(1-a_0c_0)} & \dfrac{\lambda_0^2}{1-c_0^2}\end{bmatrix}^{-1}$$

e quindi

$$\hat{\Sigma}_N = \frac{(1-a_0c_0)}{N\,(a_0-c_0)^2}\begin{bmatrix}(1-a_0^2)(1-a_0c_0) & (1-a_0^2)(1-c_0^2)\\ (1-a_0^2)(1-c_0^2) & (1-c_0^2)(1-a_0c_0)\end{bmatrix}.$$

Ora $a-c=[1\ -1]\,\theta$ e quindi $\hat{a}_N-\hat{c}_N=[1\ -1]\,\hat{\theta}_N$ da cui si ricava immediatamente $\hat{\sigma}^2_N=[1\ -1]\,\hat{\Sigma}_N\,[1\ -1]^\top$ per cui

$$\begin{aligned}N\hat{\sigma}^2_N &= \frac{(1-a_0c_0)}{(a_0-c_0)^2}\,[1\ -1]\begin{bmatrix}(1-a_0^2)(1-a_0c_0) & (1-a_0^2)(1-c_0^2)\\ (1-a_0^2)(1-c_0^2) & (1-c_0^2)(1-a_0c_0)\end{bmatrix}\begin{bmatrix}1\\-1\end{bmatrix}\\ &= \frac{(1-a_0c_0)}{(a_0-c_0)^2}\,(a_0-c_0)^2 = 1-a_0c_0 := \sigma_0^2\,.\end{aligned}$$

Notiamo che, mentre per $a_0=c_0$ la varianza matriciale $\hat{\Sigma}_N$ diventa infinita ($\mathbb{E}_{\theta_0}\psi_{\theta_0}(t)\psi_{\theta_0}(t)^\top$ diventa singolare), la varianza asintotica di $\hat{a}_N-\hat{c}_N$ si può definire anche per $a_0=c_0$.

Difatti se $a_0 - c_0 \neq 0$ si ha $L - \lim \sqrt{N}(\hat{a}_N - \hat{c}_N) = \mathcal{N}(a_0 - c_0, \sigma_0^2)$ ma ovviamente la famiglia di distribuzioni $\mathcal{N}(\mu, \sigma_0^2)$ converge in legge quando $\mu \to 0$ alla normale centrata $\mathcal{N}(0, \sigma_0^2)$.
Quindi se $\hat{a}_N - \hat{c}_N \to 0$ (con probabilità uno), $L - \lim \sqrt{N}(\hat{a}_N - \hat{c}_N) = \mathcal{N}(0, \sigma_0^2)$.

Dato allora che, in ogni caso $\sqrt{N}(\hat{a}_N - \hat{c}_N) \to \mathcal{N}(a_0 - c_0, \sigma_0^2)$ e $N\hat{\sigma}_N^2 \to \sigma_0^2$, se $a_0 = c_0$, in base al teorema di Slutsky si ha che

$$\mathbf{x}_N = \frac{(\hat{a}_N - \hat{c}_N)^2}{\hat{\sigma}_N^2} \xrightarrow{L} \chi^2(1)$$

(altrimenti se $a_0 \neq c_0$,il rapporto converge ad una distribuzione $\chi^2(1)$ non centrale).

Problema 12 *Si vuole identificare il sistema "vero"*

$$(1+a_0 z^{-1})\mathbf{y}(t) = b_0 \mathbf{u}(t-1) + \mathbf{e}_0(t) \qquad |a_0| < 1$$

dove $\mathbf{u}$ *ed* $\mathbf{e}_0$ *sono processi i.i.d., scorrelati, di varianze rispettive* σ^2 *e* λ_0^2*. Per l'identificazione si usano modelli ARMAX di ordine 1,*

$$(1+az^{-1})\mathbf{y}(t) = b\mathbf{u}(t-1) + (1+cz^{-1})\mathbf{e}(t)\,.$$

dove $\mathbf{e}$ *è rumore bianco. Trovare l'insieme dei valori del parametro* $\theta := [a\ b\ c]^\top$ *a cui converge per* $N \to \infty$*, lo stimatore a minimo errore di predizione* $\hat{\theta}_N$*.*
Dare un'espressione per la varianza asintotica di $\hat{c}_N$*.*

Soluzione.
È evidente che il sistema "vero" appartiene alla classe parametrica di modelli ARMAX scelti per l'identificazione (corrisponde infatti alla scelta $a = a_0,\ b = b_0,\ c = 0$). Pertanto, nelle ipotesi fatte lo stimatore PEM è consistente, i.e. converge per $N \to \infty$ ai parametri del sistema vero,

$$\hat{\theta}_N := [\hat{a}_N\ \hat{b}_N\ \hat{c}_N]^\top \to [a_0\ b_0\ 0]^\top .$$

In queste condizioni si può calcolare la varianza asintotica di $\hat{c}_N$, usando la formula

$$P = \lambda_0^2 \left[\mathbb{E}_0 \{ \boldsymbol{\psi}_{\theta_0}(t) \boldsymbol{\psi}_{\theta_0}(t)^\top \} \right]^{-1} .$$

Dalle formule per il gradiente

$$(1+cz^{-1})\boldsymbol{\psi}_a(t) = y(t-1) \quad (1+cz^{-1})\boldsymbol{\psi}_b(t) = -u(t-1) \quad (1+cz^{-1})\boldsymbol{\psi}_c(t) = -\varepsilon_{\theta}(t-1)$$

per $\theta = \theta_0$ si ricava

$$\boldsymbol{\psi}_{a_0}(t) = \mathbf{y}(t-1) \quad \boldsymbol{\psi}_{b_0}(t) = -\mathbf{u}(t-1) \quad \boldsymbol{\psi}_{c_0}(t) = -\boldsymbol{\varepsilon}_{\theta_0}(t-1) = -\mathbf{e}_0(t-1)$$

e dato che $y(t)$ dipende in modo strettamente causale dalla storia passata di u si ha

$$P^{-1} = \mathbb{E}_0\{\boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^\top\} = \begin{bmatrix} \sigma_{\mathbf{y}}(0) & 0 & -\lambda_0^2 \\ 0 & \sigma_{\mathbf{u}}^2 & 0 \\ -\lambda_0^2 & 0 & \lambda_0^2 \end{bmatrix}$$

la cui inversa è

$$P = \begin{bmatrix} \frac{1}{\sigma_{\mathbf{y}}(0)-\lambda_0^2} & 0 & \frac{1}{\sigma_{\mathbf{y}}(0)-\lambda_0^2} \\ 0 & \frac{1}{\sigma_{\mathbf{u}}^2} & 0 \\ \frac{1}{\sigma_{\mathbf{y}}(0)-\lambda_0^2} & 0 & \frac{\sigma_{\mathbf{y}}(0)}{\lambda_0^2(\sigma_{\mathbf{y}}(0)-\lambda_0^2)} \end{bmatrix}$$

Ne segue che la varianza asintotica di $\hat{c}_N$ è $\frac{\lambda_0^2}{N}P_{3,3} = \frac{1}{N}\frac{\sigma_{\mathbf{y}}(0)}{\sigma_{\mathbf{y}}(0)-\lambda_0^2}$.

Rimane da calcolare $\sigma_{\mathbf{y}}(0)$. Usando il modello

$$\mathbf{y}(t) = -a_0\mathbf{y}(t-1) + b_0\mathbf{u}(t-1) + \mathbf{e}_0(t)$$

si trova facilmente

$$\sigma_{\mathbf{y}}(0) - \lambda_0^2 = \frac{b_0^2\sigma_{\mathbf{u}}^2 + \lambda_0^2}{1-a_0^2} - \lambda_0^2 = \frac{b_0^2\sigma_{\mathbf{u}}^2 + a_0^2\lambda_0^2}{1-a_0^2}$$

e quindi, per $N \to \infty$,

$$\mathrm{var}\hat{c}_N \;\simeq\; \frac{1}{N}\frac{b_0^2\sigma_{\mathbf{u}}^2 + \lambda_0^2}{b_0^2\sigma_{\mathbf{u}}^2 + a_0^2\lambda_0^2} = \frac{1}{N}\frac{1+\dfrac{\lambda_0^2}{b_0^2\sigma_{\mathbf{u}}^2}}{1+a_0^2\dfrac{\lambda_0^2}{b_0^2\sigma_{\mathbf{u}}^2}}$$

Da questa espressione si vede che la varianza di $\hat{c}_N$ è massima per $a_0 = 0$ e minima per $|a_0| \to 1$.

LA MATRICE ASINTOTICA DI FISHER E IL LIMITE DI C-R

Facciamo riferimento ad un modello probabilistico congiunto delle variabili osservate di struttura generale, del tipo

$$p_\theta(y^N, u^N), \qquad \theta \in \Theta$$

Useremo i simboli $y_t, u_t, y^t, u^t, \ldots$ come variabili correnti nelle densità di probabilità di variabili aleatorie che sarebbero normalmente denotate con le stesse lettere in carattere grassetto (esempio $p_{\mathbf{y}(t)}(x) \equiv p(y_t)$). Con questa convenzione, usando ripetutamente le note regole delle probabilità condizionate si ottiene

$$\begin{aligned} p_\theta(y^N, u^N) &= p_\theta(y_N, u_N \mid y^{N-1}, u^{N-1}) p_\theta(y^{N-1}, u^{N-1}) \\ &= p_\theta(y_N, \mid y^{N-1}, u^{N-1}) p(u_N \mid y^N, u^{N-1}) p_\theta(y^{N-1}, u^{N-1}) \\ &= \ldots \\ &= \prod_{t=1}^{N} p_\theta(y_t \mid y^{t-1}, u^{t-1}) \prod_{t=1}^{N} p(u_t \mid y^t, u^{t-1}) \end{aligned}$$

N.B: Le probabilità condizionate $p(u_t \mid y^t, u^{t-1})$, $t = 1, 2, \ldots$ che descrivono il canale di reazione, non dipendono dal parametro θ (non siamo interessati alla sua modellizzazione).

Notiamo che c'è una arbitrarietà strutturale nella decomposizione. Avremmo potuto egualmente scrivere i prodotti come

$$p_\theta(y_t \mid y^{t-1}, u^t)\, p(u_t \mid y^{t-1}, u^{t-1})$$

invece di $p_\theta(y_t \mid y^{t-1}, u^{t-1})\, p(u_t \mid y^t, u^{t-1})$. La scelta fatta corrisponde ad assegnare il ritardo alla catena di azione diretta.

Supponiamo che questa famiglia di densità descriva (una famiglia parametrica di) processi *stazionari* per cui nella decomposizione

$$\mathbf{y}(t) = \mathbb{E}_\theta \left[\mathbf{y}(t) \mid \mathbf{y}^{t-1}, \mathbf{u}^{t-1} \right] + \mathbf{e}(t) := \hat{\mathbf{y}}_\theta(t \mid t-1) + \mathbf{e}(t)$$

al limite per $t \to \infty$, $\mathbf{e}$ tende a diventare l'innovazione stazionaria. Assumiamo che al limite **la densità di probabilità di $\mathbf{e}(t)$ non dipenda dal parametro θ**.

Questo è quanto accade nel caso di un modello razionale con rumore Gaussiano,

$$p_\theta(y_t \mid y^{t-1}, u^{t-1}) = \frac{1}{\sqrt{2\pi\lambda_\theta^2(t)}} \exp -\frac{1}{2}\frac{[y_t - \hat{y}_\theta(t \mid t-1)]^2}{\lambda_\theta^2(t)}$$

e, come è ben noto dalla teoria del filtro di Kalman, la varianza dell'innovazione (transitoria) $\lambda_\theta^2(t)$ converge per $t \to \infty$ ad una costante λ^2 (la varianza dell'innovazione stazionaria) indipendente dai parametri del sistema, θ.

Assumeremo di poter scrivere per $t \to \infty$,

$$p_\theta(y_t \mid y^{t-1}, u^{t-1}) = p_{\mathbf{e}}(y_t - \hat{y}_\theta(t \mid t-1)) \qquad (\#)$$

dove $p_{\mathbf{e}}$ è la densità di probabilità dell'innovazione stazionaria, che non dipende da θ. La dipendenza da θ della $p_\theta(y_t \mid y^{t-1}, u^{t-1})$ si manifesta solo attarverso il predittore (stazionario) di un passo $\hat{y}_\theta(t \mid t-1)$.

Calcoliamo il vettore delle sensitività,

$$\mathbf{z}_\theta := \frac{\partial \log p_\theta(\mathbf{y}^N \mathbf{u}^N)}{\partial \theta} = \sum_{t=1}^{N} \frac{\partial \log p_\theta(\mathbf{y}(t) \mid \mathbf{y}^{t-1} \mathbf{u}^{t-1})}{\partial \theta}$$

Supponendo che $N \to \infty$ e trascurando i primi termini nella somma, si può pensare di essere all'incirca in regime stazionario per cui si può assumere che valga la rappresentazione $(\#)$ in cui il predittore è quello stazionario. Dopo aver eliminato i termini transitori, possiamo ri-inizializzare le somme all'istante $t = 1$, ottenendo,

$$\mathbf{z}_\theta = \sum_{t=1}^{N} \frac{\partial \log p_{\mathbf{e}}(\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t \mid t-1))}{\partial \theta} = -\sum_{t=1}^{N} \frac{\partial \log p_{\mathbf{e}}(\varepsilon)}{\partial \varepsilon}\Big|_{\varepsilon=\mathbf{e}(t)} \frac{\partial \hat{\mathbf{y}}_\theta(t \mid t-1))}{\partial \theta}$$

dove il predittore è quello stazionario. Poniamo

$$\frac{\partial \log p_{\mathbf{e}}(\varepsilon)}{\partial \varepsilon}\Big|_{\varepsilon=\mathbf{e}(t)} := \ell'(\mathbf{e}(t)), \qquad \frac{1}{\kappa^2} := \mathbb{E}_\theta[\ell'(\mathbf{e}(t))]^2 \tag{124}$$

Si arriva così all'espressione della matrice di Fisher

$$I_N(\theta) = \mathbb{E}_\theta\{\mathbf{z}_\theta \mathbf{z}_\theta^\top\} = \sum_{t,s=1}^{N} \mathbb{E}_\theta\{\ell'(\mathbf{e}(t))\ell'(\mathbf{e}(s))\boldsymbol{\psi}_\theta(t)\boldsymbol{\psi}_\theta(s)^\top\}$$

e al seguente risultato.

Teorema 25 *Se* $\mathbf{e}$ *è una d-martingala rispetto alla famiglia* $(\mathbf{y}^t, \mathbf{u}^t)$ *e* $\ell'(\mathbf{e}(t))$ *è una funzione lineare di* $\mathbf{e}(t)$*, il che accade in particolare se* $\mathbf{e}$ *è un processo Gaussiano, si ha*

$$I_N(\theta) = \frac{N}{\kappa^2}\mathbb{E}_\theta\{\boldsymbol{\psi}_\theta(t)\boldsymbol{\psi}_\theta(t)^\top\} \qquad (N \to \infty) \tag{125}$$

dove κ^2 *è definita in* (124)*. Nel caso Gaussiano,* $\kappa^2 = \mathrm{var}\{\mathbf{e}(t)\} = \lambda^2$*.*

Prova:

$$\sum_{t,s=1}^{N} \mathbb{E}_\theta\{\ell'(\mathbf{e}(t))\ell'(\mathbf{e}(s))\boldsymbol{\psi}_\theta(t)\boldsymbol{\psi}_\theta(s)^\top\} =$$

$$\sum_{t=1}^{N} \mathbb{E}_\theta\{\ell'(\mathbf{e}(t))^2\boldsymbol{\psi}_\theta(t)\boldsymbol{\psi}_\theta(t)^\top\} +$$

$$2\sum_{t>s}^{N} \mathbb{E}_\theta\{\ell'(\mathbf{e}(t))\ell'(\mathbf{e}(s))\boldsymbol{\psi}_\theta(t)\boldsymbol{\psi}_\theta(s)^\top\} =$$

$$\sum_{t=1}^{N} \mathbb{E}_\theta\{\mathbb{E}_\theta\left[\ell'(\mathbf{e}(t))^2 \mid \mathbf{y}^{t-1}, \mathbf{u}^{t-1}\right]\boldsymbol{\psi}_\theta(t)\boldsymbol{\psi}_\theta(t)^\top\} +$$

$$2\sum_{t>s}^{N} \mathbb{E}_\theta\{\mathbb{E}_\theta\left[\ell'(\mathbf{e}(t)) \mid \mathbf{y}^{t-1}, \mathbf{u}^{t-1}\right]\ell'(\mathbf{e}(s))\boldsymbol{\psi}_\theta(t)\boldsymbol{\psi}_\theta(s)^\top\}$$

e il primo addendo dell'ultima a somma è uguale a (125), mentre il secondo è zero dato che $E_\theta\left[\ell'(\mathbf{e}(t)) \mid \mathbf{y}^{t-1}, \mathbf{u}^{t-1}\right] = 0$.

Notiamo poi che se $\mathbf{e}$ è Gaussiano, $\frac{\partial \log p_{\mathbf{e}}(\varepsilon)}{\partial \varepsilon}\Big|_{\varepsilon=\mathbf{e}(t)} = -\frac{\mathbf{e}(t)}{\lambda^2}$ e quindi $\kappa^2 = \lambda^2$.
□

Ricordando il teorema di Rothenberg, otteniamo

Corollario 7 *Nelle ipotesi poste, il modello $\{p_\theta(y^N, u^N)\}$ è localmente identificabile in θ, se e solo se la matrice $\mathbb{E}_\theta\{\boldsymbol{\psi}_\theta(t)\boldsymbol{\psi}_\theta(t)^\top\}$ è non-singolare.*

Notiamo che l'inversa della matrice di Fisher tende a zero come $\frac{1}{N}$. In particolare, nel caso di modelli Gaussiani si ha

$$I_N(\theta)^{-1} = \frac{\lambda^2}{N}\mathbb{E}_\theta\{\boldsymbol{\psi}_\theta(t)\boldsymbol{\psi}_\theta(t)^\top\}^{-1}, \qquad (N \to \infty).$$

Notare che qui si parla di identificabilità in senso stretto. L'enunciato continua a valere anche se l'identificabilità si intende nel senso elle statistiche del secondo ordine.

Teorema 26 **Supponiamo che valgano le stesse ipotesi del teorema 24 e che nel modello che genera i dati il processo di innovazione $\mathbf{e}_0$ sia Gaussiano. Allora lo stimatore PEM è asintoticamente efficiente; i.e. per $N \to \infty$,**

$$\mathrm{Var}\{\hat{\boldsymbol{\theta}}_N\} - I_N(\theta_0)^{-1} \to 0$$

Equivalentemente, per $N \to \infty$, la varianza di $\hat{\boldsymbol{\theta}}_N$ coincide con l'inversa della matrice di Fisher calcolata in θ_0.

In conclusione, possiamo affermare che sotto ipotesi ragionevoli sul meccanismo che genera i dati e sulla classe di modelli scelta per l'identificazione, **il metodo PEM è asintoticamente ottimale**. Naturalmente nulla sappiamo del suo comportamento per piccoli campioni.

RELAZIONE TRA STIMA PEM E MAX VEROSIMIGLIANZA

Tradizionalmente si ricorre alla stima di massima versimiglianza quando i dati sono pochi e l'approssiamzione asintotica ($N \to \infty$) non si può fare. Se la densità condizionata di $\mathbf{y}(t)$ dato il passato stretto $\{\mathbf{y}^{t-1}, \mathbf{u}^{t-1}\}$ è Gaussiana, ovvero

$$p_\theta(y_t \mid y^{t-1}, u^{t-1}) = \frac{1}{\sqrt{2\pi\lambda_\theta^2(t)}} \exp -\frac{1}{2}\frac{[y_t - \hat{y}_\theta(t \mid t-1)]^2}{\lambda_\theta^2(t)}, \qquad t = 1, 2, \ldots, N$$

si può scrivere esplicitamente la verosimiglianza. Dato che il canale di reazione non è parametrizzato, in realtà basta quella condizionata

$$L(\theta, y^N, u^N) = \prod_{t=1}^{N} p_\theta(y_t \mid y^{t-1}, u^{t-1})$$

Questa espressione richiede il calcolo del predittore transitorio $\hat{y}_\theta(t \mid t-1)$ e della varianza dell'innnovazione transitoria $\lambda_\theta^2(t)$ per ogni t. Questo si può fare, una volta fissato un valore di θ, implementando un **filtro di Kalman**.

Tranne per il caso di modelli ARX, non c'è speranza di trovare soluzioni esplicite e si ricorre a metodi di ottimizzazione iterativi.
Dato che i modelli sono di innovazione:

$$\lim_{N\to\infty} \lambda_\theta^2(t) = \lambda^2$$

(il limite dipende solo dal modello e non dai dati!) per $N \to \infty$ il problema di ottimo diventa lo stesso del metodo PEM con predittore stazionario. Quindi

Se l'innovazione è Gaussiana il metodo PEM è asintoticamente equivalente alla massima verosimiglianza.

RELAZIONE CON IL MODELLO LINEARE STATICO

Esiste una notevole relazione tra la teoria della stima parametrica sul modello statico lineare-Gaussiano che abbiamo visto all'inizio e l'analisi asintotica degli stimatori PEM.
Supponiamo che $\mathbf{y}$ sia descritto da un modello "vero" corrispondente al parametro "vero" θ_0,

$$\mathbf{y} = S\theta_0 + \mathbf{w}, \qquad \mathrm{Var}[\mathbf{w}] = \sigma^2 I_N$$

($\mathbf{y}$ è normalizzata moltiplicando a sinistra per l'inverso del fattore di Cholesky L della covarianza di rumore R), in modo da ridurci a un rumore $\sigma\mathbf{w}$ di varianza $\sigma^2 I_N$. La stima di massima verosimiglianza di θ è

$$\hat{\boldsymbol{\theta}}_N = \left[\frac{1}{N}S^\top S\right]^{-1}\frac{1}{N}S^\top \mathbf{y} = \theta_0 + \left[\frac{1}{N}S^\top S\right]^{-1}\frac{1}{N}S^\top \mathbf{w}$$

ovvero

$$\hat{\boldsymbol{\theta}}_N - \theta_0 = Q_N^{-1}\frac{1}{N}S^\top \mathbf{w}, \qquad Q_N := \frac{1}{N}S^\top S$$

Nel caso di modelli dinamici, se il modello vero appartiene alla classe parametrica con cui si calcola il predittore, l'errore di predizione si può approssimare col modello lineare "incrementale"

$$\begin{aligned}\boldsymbol{\varepsilon}_\theta(t) &= \mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t \mid t-1) = -(\hat{\mathbf{y}}_\theta(t \mid t-1) - \hat{\mathbf{y}}_{\theta_0}(t \mid t-1)) + \mathbf{e}_0(t) \\ &\simeq \boldsymbol{\psi}_{\theta_0}^\top(t)\,(\theta - \theta_0) + \mathbf{e}_0(t)\,, \quad t = 1, 2, \ldots, N\end{aligned}$$

a meno di termini che sono infinitesimi di ordine superiore in $\|\theta - \theta_0\|$. Introduciamo uno stimatore *incrementale* ai minimi quadrati di $\theta - \theta_0$

$$\tilde{\boldsymbol{\theta}}_N - \theta_0 = \left[\frac{1}{N}\sum_{t=1}^{N} \boldsymbol{\psi}_{\theta_0}(t)\,\boldsymbol{\psi}_{\theta_0}(t)^\top\right]^{-1} \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{\psi}_{\theta_0}(t)\,\boldsymbol{\varepsilon}_\theta(t)$$

sostituendo si trova

$$\tilde{\boldsymbol{\theta}}_N - \theta_0 = \theta - \theta_0 + \left[\frac{1}{N}\sum_{t=1}^{N} \boldsymbol{\psi}_{\theta_0}(t)\,\boldsymbol{\psi}_{\theta_0}(t)^\top\right]^{-1} \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{\psi}_{\theta_0}(t)\,\mathbf{e}_0(t)\,. \qquad (\dagger\dagger)$$

dove θ è un valore qualunque "vicino" a θ_0. Possiamo anche porre $\theta = \theta_0$.

Nelle condizioni che garantiscono la consistenza e la normalità asintotica dello stimatore PEM, si può scrivere,

$$\hat{\boldsymbol{\theta}}_N - \theta_0 \simeq \left[\frac{1}{N}\sum_{t=1}^{N} \boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^\top\right]^{-1} \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{\psi}_{\theta_0}(t)\mathbf{e}_0(t)$$

dove il simbolo $\simeq$ significa che i due membri di questa espressione moltiplicati per $\sqrt{N}$ hanno lo stesso limite in legge. Possiamo pertanto concludere (per Slutsky) che:

Se $\theta \to \theta_0$, **lo stimatore ai minimi quadrati $\tilde{\boldsymbol{\theta}}_N$ ricavato dal modello lineare incrementale, ha, per $N \to \infty$, lo stesso limite in legge dello stimatore PEM, $\hat{\boldsymbol{\theta}}_N$.**

Proposizione 28 *La distribuzione asintotica dello stimatore PEM del parametro θ per un modello dinamico lineare che soddisfi le ipotesi di consistenza e normalità asintotica, coincide con la distribuzione limite dello stimatore di massima verosimiglianza del parametro θ nel modello lineare statico Gaussiano , in cui si sono fatte le sostituzioni*

$$S = \begin{bmatrix} \boldsymbol{\psi}_{\theta_0}(1)^\top \\ \vdots \\ \boldsymbol{\psi}_{\theta_0}(N)^\top \end{bmatrix}, \qquad \mathbf{w} = \begin{bmatrix} \mathbf{e}_0(1) \\ \vdots \\ \mathbf{e}_0(N) \end{bmatrix},$$

e si trattano i vettori $\{\boldsymbol{\psi}_{\theta_0}(t); t = 1, 2, \ldots, N\}$ come quantità **deterministiche**.

L'espressione limite per la varianza si ottiene sostituendo alla matrice Q_N il suo limite per $N \to \infty$, uguale a $\bar{Q} := E_0\{\boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^\top\}$.

DISTRIBUZIONE ASINTOTICA DI STIMATORI NON CONSISTENTI

Consideriamo il seguente problema:
Si vuole identificare il sistema vero con rumore $\mathbf{e}_0$ i.i.d.

$$\mathbf{y}(t) = b_0\mathbf{u}(t-1) + \mathbf{e}_0(t), \qquad \text{var}\{\mathbf{e}_0\} = \lambda_0^2$$

sottoposto ad un ingresso descritto dal modello

$$\mathbf{u}(t) = \rho\mathbf{u}(t-1) + \mathbf{w}(t)$$

in cui $|\rho| < 1$ e $\mathbf{w}$ è un rumore i.i.d. di varianza σ^2 indipendente da $\mathbf{e}_0$, usando una classe di modelli del tipo

$$\mathbf{y}(t) = b\mathbf{u}(t-1) + \mathbf{e}(t)\,. \qquad (**)$$

in cui si assume che $\mathbf{e}(t)$ sia indipendente dalla storia passata $\mathbf{u}^{t-1}$.

Si chiede di calcolare la distribuzione asintotica dello stimatore PEM $\hat{b}_N$.

Si può ben dire che il modello vero ($b = b_0$) appartiene alla classe parametrica (**). Questa proprietà è indipendente dalla dinamica propria dell'ingesso.

Ne segue che lo stimatore PEM è consistente e asintoticamente normale.

La sua varianza asintotica si calcola con la solita formula. il predittore di Wiener di $\mathbf{y}(t)$ è $b\mathbf{u}(t-1)$ e il gradiente dell'errore di predizione si calcola immediatamente

$$\psi_0(t) = -\mathbf{u}(t-1)$$

Quindi, detta $\sigma_{\mathbf{u}}(\tau)$ la covarianza di $\mathbf{u}$, la varianza asintotica ha l'espressione

$$\mathrm{var}\{\hat{b}_N\} \;\to\; \frac{\lambda_0^2}{N\,\sigma_{\mathbf{u}}(0)} = \frac{1-\rho^2}{N}\,\frac{\lambda_0^2}{\sigma_{\mathbf{w}}^2}\,.$$

Come si vede al tendere di ρ a 1 (quindi all'aumentare della banda, ovvero all'aumentare della potenza statistica, $\sigma_{\mathbf{u}}(0)$, dell'ingresso) la varianza asintotica della stima tende a zero.

A quale classe appartiene il modello vero?

Usando il modello per $\mathbf{u}$, si può riscrivere il modello vero come

$$\mathbf{y}(t) = \rho b_0 \mathbf{u}(t-2) + b_0 \mathbf{w}(t-1) + \mathbf{e}_0(t)$$

in cui, per l'indipendenza dei segnali in gioco, il processo:

$$\tilde{\mathbf{e}}_0(t) := b_0 \mathbf{w}(t-1) + \mathbf{e}_0(t)$$

è ancora rumore bianco i.i.d. di varianza $\tilde{\lambda}_0^2 = b_0^2 \sigma_{\mathbf{w}}^2 + \lambda_0^2$. Da notare che per la stabilità asintotica del modello dell'ingresso, si ha $\mathbf{u}(t) \in \mathbf{H}_t(\mathbf{w})$ e quindi $\mathbf{w}(t)$ è indipendente dalla storia passata $\mathbf{u}^{t-1}$. Quindi anche $\tilde{\mathbf{e}}_0(t)$ è indipendente dalla storia passata $\mathbf{u}^{t-1}$ ed è quindi l'innovazione. Questo modello vero non appartiene più alla classe di prima ma alla:

$$\mathbf{y}(t) = b\mathbf{u}(t-2) + \mathbf{e}(t)\,. \qquad (***)$$

Combinandolo con la dinamica dell'ingresso, il modello vero adesso appartiene alla classe dei modelli $(***)$. Calcoliamo la varianza dell'errore di predizione, il quale risulta avere l'espressione

$$\varepsilon_\theta(t) = (\rho b_0 - b)\mathbf{u}(t-2) + b_0\mathbf{w}(t-1) + \mathbf{e}_0(t)$$

e, per l'indipendenza dei segnali in gioco, si trova

$$\mathrm{var}[\,\varepsilon_\theta(t)\,] = (\rho b_0 - b)^2\,\sigma_{\mathbf{u}}(0) + b_0^2\,\sigma_{\mathbf{w}}^2 + \lambda_0^2$$

che è minima per $b = \rho\, b_0$, per cui il limite a cui tende lo stimatore PEM, è

$$\lim_{N\to\infty} \hat{b}_N = \rho\, b_0$$

dato che ci interessa stimare il parametro vero b_0, potremmo dire che non si ha consistenza. In realtà combinando con la dinamica dell'ingresso abbiamo introdotto nel modello un parametro "spurio" ρ che descrive $\mathbf{u}$. Noi vogliamo stimare solo i parametri del modello vero **indipendentemente dalla dinamica di $\mathbf{u}$**.

Il predittore di Wiener di $\mathbf{y}(t)$ col modello $(***)$ è $b\mathbf{u}(t-2)$ e il gradiente dell'errore di predizione si calcola immediatamente

$$\psi_0(t) = -\mathbf{u}(t-2)$$

Quindi, detta $\sigma_{\mathbf{u}}(\tau)$ la covarianza di $\mathbf{u}$, la varianza asintotica di $\hat{b}_N$ si può calcolare con la formula del teorema 24 e vale

$$\mathrm{var}\{\hat{b}_N\} \;\to\; \frac{\tilde{\lambda}_0^2}{N\,\sigma_{\mathbf{u}}(0)} = \frac{1-\rho^2}{N}\,[\,b_0^2 + \frac{\lambda_0^2}{\sigma_{\mathbf{w}}^2}\,]\;.$$

La varianza asintotica adesso è maggiore di quella ottenuta col modello $(**)$.

Verifica: il predittore di Wiener basato sul modello $(***)$ è una funzione lineare di b, per cui lo stimatore PEM è lo stimatore ai minimi quadrati

$$\hat{b}_N = \left[\frac{1}{N}\sum_{t=1}^{N} \mathbf{u}(t-2)^2\right]^{-1} \frac{1}{N}\sum_{t=1}^{N} \mathbf{u}(t-2)\,\mathbf{y}(t)$$

e sostituendo in questa espressione il modello vero si trova

$$\hat{b}_N = \rho b_0 + \left[\frac{1}{N}\sum_{t=1}^{N} \mathbf{u}(t-2)^2\right]^{-1} \frac{1}{N}\sum_{t=1}^{N} \mathbf{u}(t-2)\,(b_0\mathbf{w}(t-1) + \mathbf{e}_0(t))\,.$$

In questa espressione compare il termine $\tilde{\mathbf{w}}(t) = b_0\mathbf{w}(t-1) + \mathbf{e}_0(t)$ che è rumore bianco i.i.d. di varianza $\tilde{\lambda}_0^2$ e per le ipotesi di indipendenza fatte, si vede che $\mathbf{u}(t-2)\,\tilde{\mathbf{w}}(t)$ è una d-martingala di varianza

$$\sigma_0^2 = \mathbb{E}\{\mathbf{u}(t-2)^2\,\tilde{\mathbf{w}}(t)^2\} = \mathbb{E}\{\mathbf{u}(t-2)^2\}\,\mathbb{E}\{\tilde{\mathbf{e}}_0(t)^2\} = \sigma_{\mathbf{u}}(0)\,\tilde{\lambda}_0^2 = \frac{\sigma_{\mathbf{w}}^2}{1-\rho^2}\,\tilde{\lambda}_0^2\,.$$

Per il teorema del limite centrale si ha ($L-\lim$ è il limite in legge),

$$L-\lim \sqrt{N}\,\frac{1}{N}\sum_{t=1}^{N} \mathbf{u}(t-2)\tilde{\mathbf{w}}(t) = \mathcal{N}(0\,,\,\sigma_0^2)$$

Dato che

$$\lim \frac{1}{N}\sum_{t=1}^{N} \mathbf{u}(t-2)^2 = \sigma_{\mathbf{u}}(0)$$

applicando il teorema di Slutsky si può trovare la distribuzione asintotica di $\hat{b}_N$.

$$L \lim_{N\to\infty} \frac{1}{\sqrt{N}}(\hat{b}_N - \rho\, b_0) = L \lim_{N\to\infty} \frac{1}{\sigma_{\mathbf{u}}(0)} \frac{1}{\sqrt{N}} \sum_{t=1}^{N} \mathbf{u}(t-2)\tilde{\mathbf{w}}(t)$$

$$= \mathcal{N}(0, \frac{\sigma_0^2}{\sigma_{\mathbf{u}}(0)^2}) \quad = \quad \mathcal{N}(0, \frac{\tilde{\lambda}_0^2}{\sigma_{\mathbf{u}}(0)})$$

per cui la varianza asintotica dello stimatore è

$$\mathrm{var}\{\hat{b}_N\} \sim \frac{1}{N}\frac{\tilde{\lambda}_0^2}{\sigma_{\mathbf{u}}(0)} = \frac{1-\rho^2}{N}[b_0^2 + \frac{\lambda_0^2}{\sigma_{\mathbf{w}}^2}]$$

che è la stessa formula trovata col Teorema 24.

STIMA PEM DI FUNZIONI DI TRASFERIMENTO

Finora ci siamo occupati solo delle proprietà asintotiche di stimatori del parametro θ, che in realtà è raramente la quantità di interesse diretto nei procedimenti di modellizzazione a partire dai dati. Studieremo le proprietà asintotiche delle stime di funzioni di trasferimento

$$\hat{F}_N(e^{j\omega}) := F_{\hat{\theta}_N}(e^{j\omega}), \qquad \hat{G}_N(e^{j\omega}) := G_{\hat{\theta}_N}(e^{j\omega})$$

Lo strumento che facilita in gran modo l'analisi asintotica di questi stimatori è il teorema di Cramèr

Teorema 27 (Cramèr) *Sia* $\{\mathbf{x}_N; N = 1,2,\ldots\}$ *una successione di vettori aleatori* n*-dimensionali per cui* $\sqrt{N}(\mathbf{x}_N - \mu) \xrightarrow{L} \mathcal{N}(0,\Sigma)$ *e* $g : \mathbb{R}^n \to \mathbb{R}^k$ *una funzione la cui matrice Jacobiana* $G(x)$ *(esiste ed) è continua in un intorno del punto* $x = \mu$*. Allora:*

$$\sqrt{N}(g(\mathbf{x}_N) - g(\mu)) \xrightarrow{L} \mathcal{N}(0, G(\mu)\Sigma G(\mu)'). \tag{126}$$

Applicazione : consideriamo la successione dei quadrati delle medie campionarie $\{(\bar{\mathbf{y}}_T)^2\}$. Dato che per $T \to \infty$, $\sqrt{T}\,(\bar{\mathbf{y}}_T - \mu) \sim \mathcal{N}(0, \sigma^2)$ e $\frac{\partial}{\partial x} g(x) = 2x$, si ha

$$\sqrt{T}\left[(\bar{\mathbf{y}}_T)^2 - \mu^2\right] \xrightarrow{L} \mathcal{N}(0, 4\mu^2\sigma^2)\,.$$

Notare però che se $\mu = 0$ la distribuzione limite è degenere (ha varianza zero) e la relazione non dice nulla sulla distribuzione asintotica. Dato che

$$\sqrt{T}\bar{\mathbf{y}}_T \xrightarrow{L} \mathcal{N}(0, \sigma^2)$$

e quindi $\dfrac{\sqrt{T}\bar{\mathbf{y}}_T}{\sigma} \xrightarrow{L} \mathcal{N}(0,1)$, applicando Slutsky si ha

$$\left(\frac{\sqrt{T}\bar{\mathbf{y}}_T}{\sigma}\right)^2 \xrightarrow{L} \chi^2(1)$$

Pertanto la varianza asintotica di $(\bar{\mathbf{y}}_T)^2$ è

$$\frac{\sigma^2}{T^2}\mathrm{var}\{\,\chi^2(1)\,\} = \frac{2\sigma^4}{T^2}$$

che tende a zero come $1/T^2$.

Dal Teorema di Cramèr discende immediatamente il seguente risultato.

Teorema 28 *Sia $W_\theta(e^{j\omega})$ una funzione razionale di $e^{j\omega}$, in genere a valori vettoriali in $\mathbb{C}^r$, che dipende in modo regolare dal parametro θ e*

$$J_\theta(e^{j\omega}) := [\frac{\partial}{\partial \theta_k} W_\theta(e^{j\omega})]_{k=1,2,\ldots,p}$$

la matrice Jacobiana. Allora

$$\lim_{N\to\infty} \sqrt{N}\,[W_{\hat{\theta}_N}(e^{j\omega}) - W_{\theta_0}(e^{j\omega})] = \mathscr{N}\left(0, J_{\theta_0}(e^{j\omega}) P J_{\theta_0}(e^{j\omega})^\top\right) \qquad (127)$$

in legge, qualunque sia $\omega \in [-\pi, \pi]$. La matrice P è la varianza asintotica di $\hat{\theta}_N$.

Esempio 8 Si consideri la funzione di trasferimento polinomiale di tipo FIR

$$G_\theta(e^{j\omega}) = \sum_{k=1}^{p} \theta_k e^{-j\omega k}. \tag{128}$$

di ordine p noto. Si vuole stimare la funzione di trasferimento partendo da osservazioni rumorose dell'uscita

$$\mathbf{y}(t) = \varphi(t)^\top \theta + \mathbf{e}(t) \qquad \varphi(t) = \begin{bmatrix} \mathbf{u}(t-1) \\ \vdots \\ \mathbf{u}(t-p) \end{bmatrix}, \qquad t = 1, 2, \ldots, N$$

dove $\theta = \begin{bmatrix} \theta_1 & \ldots & \theta_p \end{bmatrix}^\top$ ed $\mathbf{e}(t)$ è un processo bianco che per ogni t è scorrelato da $\mathbf{u}^{t-1}$ (e quindi anche da $\varphi(t)$), di varianza λ^2. Si identifica il modello usando un metodo PEM.

Discutere la consistenza dello stimatore $\hat{G}_N(e^{j\omega}) := G_{\hat{\theta}_N}(e^{j\omega})$ e dare un'espressione per la la sua varianza asintotica in funzione della frequenza.

Discutere in particolare il caso in cui $\mathbf{u}$ è rumore bianco.

STIMA DI FREQUENZE

La stima di componenti periodiche (e delle relative frequenze) di segnali aleatori è un problema molto importante che ha generato libri e migliaia di articoli.

Le applicazioni possono essere le più svariate, dalla rivelazione di guasti in macchine rotanti, alla cancellazione di rumori periodici, all'analisi di spettri di varia natura alla modellizzazione della marea astronomica etc etc.

Dato che la frequenza entra nelle funzioni seno e/o coseno, il problema di stima di frequenza è non lineare.

SEGNALI QUASI PERIODICI

Un processo QP è somma di ν componenti armoniche elementari (in generale complesse),

$$\mathbf{z}(t) = \sum_{k=1}^{\nu} \mathbf{z}_k e^{j\omega_k t}, \qquad t \in \mathbb{Z}$$

dove $\omega_k \in [-\pi, \pi]$ sono pulsazioni che si possono supporre diverse tra loro e le $\mathbf{z}_k$, $k = 1, \ldots, \nu$ sono variabili aleatorie (complesse) a media zero e varianza finita.

Stazionarietà: Le correlazioni delle varie componenti armoniche

$$\mathbb{E}\left[\mathbf{z}_k \bar{\mathbf{z}}_h\right] e^{j\omega_k t - j\omega_h s} \qquad k, h = 1, 2, \ldots, \nu$$

debbono dipendere da $t - s$, il che può accadere solo per $k = h$ mentre per $k \neq h$ si deve necessariamente avere $\mathbb{E}\left[\mathbf{z}_k \bar{\mathbf{z}}_h\right] = 0 \Rightarrow$ Le $\{\mathbf{z}_k\}$ debbono essere tra loro scorrelate. Allora:

$$\sigma_{\mathbf{z}}(t, s) = \mathbb{E}\,\mathbf{z}(t)\bar{\mathbf{z}}(s) = \sum_{k=1}^{\nu} \sigma_k^2 e^{j\omega_k (t-s)}, \qquad \sigma_k^2 = \mathbb{E}\,|\mathbf{z}_k|^2$$

da cui $\sigma_{\mathbf{z}}(t,s) = \sigma_{\mathbf{z}}(t-s)$ e $\mathbf{z}$ è effettivamente un processo stazionario (in senso debole).

Se il processo è *reale* $\mathbf{z}(t) = \mathbb{R}e\{\mathbf{z}(t)\} = \dfrac{\mathbf{z}(t)+\bar{\mathbf{z}}(t)}{2}$ e le componenti armoniche debbono essere a coppie complesse coniugate

$$\mathbf{z}(t) = \sum_{k=-\nu}^{\nu} \frac{1}{2}\mathbf{z}_k e^{j\omega_k t}, \qquad \omega_{-k} = -\omega_k \qquad \mathbf{z}_{-k} = \bar{\mathbf{z}}_k$$

dove le $\{\mathbf{z}_k\}$ sono variabili aleatorie tra loro scorrelate. Il termine corrispondente a $k = 0$ è un' eventuale componente continua a frequenza zero $\omega_0 = 0$ (frequenza zero è reale).
Scrivendo $\mathbf{z}_k = \mathbf{x}_k + i\mathbf{y}_k$, per indice negativo $(-k)$ si ha $\mathbf{z}_{-k} = \mathbf{x}_k - i\mathbf{y}_k$; e quindi l'incorrelazione dei coefficienti a indice diverso implica che

$$\mathbb{E}\,\mathbf{z}_k\bar{\mathbf{z}}_{-k} = \mathbb{E}\{(\mathbf{x}_k + i\mathbf{y}_k)\overline{(\mathbf{x}_k - i\mathbf{y}_k)}\} = \mathbb{E}\{(\mathbf{x}_k^2 - \mathbf{y}_k^2) + 2i\mathbf{x}_k\mathbf{y}_k\} = 0$$

da cui

$$\mathbb{E}\,\mathbf{x}_k^2 = \mathbb{E}\,\mathbf{y}_k^2, \qquad \mathbb{E}\,\mathbf{x}_k\mathbf{y}_k = 0\,.$$

Ogni componente armonica elementare si può scrivere in forma reale come

$$\mathbf{z}_k(t) := \frac{1}{2}\{\mathbf{z}_k e^{j\omega_k t} + \bar{\mathbf{z}}_k e^{-j\omega_k t}\} = \mathbf{x}_k \cos \omega_k t - \mathbf{y}_k \sin \omega_k t \qquad k = 1, \ldots, \nu.$$

Il segnale ha potenza statistica $\sigma_k^2 := \mathbb{E}\, \mathbf{z}_k(t)^2 = \mathbb{E}\, \mathbf{z}_k(0)^2 = \mathbb{E}\, \mathbf{x}_k^2 = \mathbb{E}\, \mathbf{y}_k^2$ e ha una realizzazione di stato del tipo

$$\begin{bmatrix} \mathbf{x}_k(t+1) \\ \mathbf{y}_k(t+1) \end{bmatrix} = \begin{bmatrix} \cos \omega_k & -\sin \omega_k \\ \sin \omega_k & \cos \omega_k \end{bmatrix} \begin{bmatrix} \mathbf{x}_k(t) \\ \mathbf{y}_k(t) \end{bmatrix}$$

$$\mathbf{z}_k(t) = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}_k(t) \\ \mathbf{y}_k(t) \end{bmatrix}$$

con condizioni iniziali aleatorie scorrelate $\mathbf{x}_k(0) = \mathbf{x}_k$, $\mathbf{y}_k(0) = \mathbf{y}_k$ di ugual varianza σ_k^2.
Il modello di stato complessivo per il segnale $\mathbf{z}$ ha la forma

$$\mathbf{s}(t+1) = A\,\mathbf{s}(t) \tag{129}$$

$$\mathbf{z}(t) = c^\top \mathbf{s}(t) \tag{130}$$

$\mathbf{s}(t)$ è il vettore di stato di dimensione $2\nu+1$ ottenuto incolonnando i vettori di stato elementari per $k = 0, 1, 2, \ldots, \nu$.

La $\mathbf{z}_0(t) \equiv \mathbf{z}_0(0) \equiv \mathbf{z}_0$ è una componente continua costante. Togliendola, la matrice A ha una struttura diagonale a blocchi $A = \mathrm{diag}\{A_1, \ldots A_\nu\}$ in cui tutti i blocchi A_k, di dimensione 2×2, sono matrici ortogonali. La varianza di stato del modello è una matrice diagonale

$$P = \mathbb{E}\mathbf{s}(0)\mathbf{s}(0)^\top = \mathrm{diag}\{\sigma_1^2 I_2, \ldots, \sigma_\nu^2 I_2\}$$

e la funzione di covarianza di $\mathbf{z}_k(t) = \mathbf{x}_k \cos \omega_k t - \mathbf{y}_k \sin \omega_k t$ è :

$$\mathbb{E}\mathbf{z}_k(t+\tau)\mathbf{z}_k(t) = \sigma_k^2 \cos \omega_k \tau$$

(usare $\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$) per cui

$$\sigma_{\mathbf{z}}(\tau) = c^\top A^\tau P c = \sum_{k=1}^{\nu} \sigma_k^2 \cos \omega_k \tau$$

da cui lo spettro di $\mathbf{z}$ (che dev'essere una funzione pari) si può scrivere nella forma

$$\phi(\omega) = \sum_{k=-\nu}^{\nu} \frac{1}{2}\sigma_k^2 \, \delta(\omega - \omega_k), \quad \sigma_{-k}^2 = \sigma_k^2 \, .$$

Osservazione 2 Ci si chiede se in un qualunque modello di stato di un processo scalare QP possano esserci autovalori multipli di A. La risposta, è no se si richiede l'osservabilità del modello. Questo fatto si può verificare usando il criterio di Hautus. Con lo stesso criterio si vede facilmente che le rappresentazioni di ordine 2 per le eventuali componenti a frequenze $\omega_k = \pm\pi$ sono ridondanti. Per queste componenti la rappresentazione minima ha ovviamente dimensione uno.

Al modello di stato corrisponde una descrizione "ingresso-uscita" del tipo

$$A(z^{-1})\mathbf{z}(t) = 0 \qquad A(z^{-1}) = \prod_{k=1}^{\nu} (1 - 2\cos\omega_k z^{-1} + z^{-2}) \tag{131}$$

dove $A(z^{-1}) = z^{-n}\det(zI - A)$ è il polinomio caratteristico della matrice A. SE $\omega_k \neq \pm\pi$, questa descrizione ingresso-uscita è *minima* nel senso che non esiste polinomio di grado più basso di $A(z^{-1})$ che annulla il segnale $\mathbf{z}(t)$; i.e. l'equazione alle differenze ha l'ordine minimo possibile per descrivere $\mathbf{z}(t)$. Se ci sono zeri in $z = \pm 1$ bisogna moltiplicare $A(z^{-1})$ per i corrispondenti fattori $1 \pm z^{-1}$.

Notiamo che, introducendo le condizioni iniziali (aleatorie) la Z-trasformata della soluzione $\mathbf{z}$ si può esprimere come una funzione razionale

$$\mathbf{z}(t) = \frac{N(z^{-1})}{A(z^{-1})}$$

dove $N(z^{-1})$ è un polinomio a coefficienti aleatori determinati dalle condizioni iniziali.

SEGNALI QUASI PERIODICI IN RUMORE BIANCO

Supponiamo che il segnale osservato $\mathbf{y}(t)$ sia

$$\mathbf{y}(t) = \mathbf{z}(t) + \mathbf{e}(t)$$

dove $\mathbf{z}(t)$ è un segnale QP reale del tipo analizzato nella sezione precedente e $\mathbf{e}(t)$ è un segnale a spettro continuo. Spesso si assume che $\mathbf{e}(t)$ sia *rumore bianco* di varianza incognita σ^2.
In questa ipotesi, possiamo descrivere il segnale mediante un modello ingresso-uscita del tipo

$$A(z^{-1})\mathbf{y}(t) = A(z^{-1})\mathbf{e}(t)\,. \tag{132}$$

NB: la cancellazione del fattore comune $A(z^{-1})$ nei due termini non è lecita perchè il sistema (132) parte da **condizioni iniziali non nulle** al tempo zero.

STIMA DI FREQUENZE COL METODO PEM

Descriveremo un metodo di stima di frequenze di un segnale QP immerso in rumore bianco basato su PEM. Il metodo è stato proposto da Nehorai nel 1985 e successivamente è stato riesaminato e affinato da vari autori.

Identificazione di modelli ARMA: si fissa, basandosi sull' informazione disponibile a priori, una classe parametrica di funzioni di trasferimento a fase minima $\{G_\theta(z)\,;\,\theta\in\Theta\}$ e si calcola *l'errore di predizione del modello* G_θ che è definito come la differenza

$$\boldsymbol{\varepsilon}_\theta(t) := \mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t\,|\,t-1) \qquad \hat{\mathbf{y}}_\theta(t\,|\,t-1) = [G_\theta(z)-1]\,G_\theta(z)^{-1}\mathbf{y}(t)$$

è il predittore "approssimato" costruito in base al modello G_θ.

L'errore di predizione associato al modello G_θ si ottiene semplicemente filtrando il processo con la funzione di trasferiemnto inversa G_θ^{-1}

$$\boldsymbol{\varepsilon}_\theta(t) = G_\theta^{-1}(z)\mathbf{y}(t)$$

Notiamo che questa operazione è possibile se G_θ non ha zeri sul cerchio unitario.

FILTRI NOTCH

Si vuole generalizzare la procedura PEM al caso di processi con una componente QP in rumore bianco, ai quali in senso stretto non sarebbe applicabile.

Il processo $\mathbf{y}$ è descritto dalla classe di modelli (132), a poli e zeri tutti sulla circonferenza unitaria. Il polinomio $A(z^{-1})$ è un polinomio **simmetrico**

$$A(z^{-1}) = \prod_{k=1}^{n} (1 - 2\cos\omega_k z^{-1} + z^{-2}) =$$

$$1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_n z^{-n} + \ldots + a_2 z^{-2n+2} + a_1 z^{-2n+1} + 1 z^{-2n} =$$

$$z^{-n}[\,1 z^{n} + a_1 z^{n-1} + a_2 z^{n-2} + \ldots + a_n + \ldots + a_2 z^{-n+2} + a_1 z^{-n+1} + 1 z^{-n}]$$

che dipende solo da n parametri. Può essere parametrizzato mediante n coefficienti incogniti $\theta := [a_1, a_2, \ldots, a_n]^\top$ che dipendono dai coseni delle frequenze $\omega_1, \omega_2, \ldots, \omega_n$.

Per calcolare l'errore di predizione associato a questa classe di modelli gli zeri delle funzione di trasferimento debbono essere stabili; i.e. strettamente dentro il cerchio unità. Si approssima il polinomio a numeratore di $G_\theta(z)$ spostando gli zeri nei $2n$ punti $z_k = \rho e^{\pm j\omega_k}$ dove il parametro $0 < \rho < 1$ è scelto prossimo a 1.
Se $z^{2n}A(z^{-1})$ ha uno zero in α allora $z^{2n}A(\rho z^{-1})$ ha uno zero in $\rho\alpha$.
L'errore di predizione si calcola col filtro approssimato:

$$\boldsymbol{\varepsilon}_\theta(t) \simeq \frac{A(z^{-1})}{A(\rho z^{-1})}\mathbf{y}(t)$$

che ha poli (di modulo ρ) interni al cerchio unità. L'equazione alle differenze corrispondente è

$$A_\theta(\rho z^{-1})\varepsilon_\theta(t) = A_\theta(z^{-1})\mathbf{y}(t)\,.$$

Questa equazione alle differenze può essere riscritta

$$(1+\rho a_1 z^{-1}+\ldots+\rho^n a_n z^{-n}+\ldots+\rho^{2n-1}a_1 z^{-2n+1}+\rho^{2n}z^{-2n})\varepsilon_\theta(t) = \\ (1+a_1 z^{-1}+\ldots+a_n z^{-n}+\ldots+a_1 z^{-2n+1}+z^{-2n}))\mathbf{y}(t)$$

ovvero

$$\varepsilon_\theta(t) = \mathbf{y}(t)+\mathbf{y}(t-2n)-\rho^{2n}\varepsilon_\theta(t-2n)-\varphi(t)^\top\theta \qquad (133)$$

dove $\varphi_\theta(t) = [\varphi_1(t)\,\varphi_2(t)\ldots\varphi_n(t)]^\top$ è definita dalle

$$\begin{aligned} \varphi_i(t) &= \begin{cases} -\mathbf{y}(t-i)-\mathbf{y}(t-2n+i)+\rho^i\varepsilon_\theta(t-i)+\rho^{2n-i}\varepsilon_\theta(t-2n+i) \\ \text{per} \quad 1\le i\le n-1 \end{cases} \\ \varphi_i(t) &= -\mathbf{y}(t-n)+\rho^n\varepsilon_\theta(t-n) \qquad \text{per} \quad i=n\,. \end{aligned} \qquad (134)$$

Per il calcolo del gradiente di $\varepsilon_\theta(t)$ cambiato di segno

$$\psi_\theta(t) := -\frac{\partial\varepsilon_\theta(t)}{\partial\theta}$$

si trova la relazione (vedere il lavoro di Nehorai (1985) per i dettagli)

$$\psi_\theta(t) = \frac{1}{A_\theta(\rho z^{-1})}\varphi_\theta(t) \qquad (135)$$

ALGORITMO DI OTTIMIZZAZIONE

Si usa l'algoritmo di Gauss-Newton nella forma standard dei metodi PEM per modelli ARMA

$$\theta_{k+1} = \theta_k + \left[\sum_{t=1}^{N} \psi_{\theta_k}(t)\psi_{\theta_k}(t)^\top\right]^{-1} \sum_{t=1}^{N} \psi_{\theta_k}(t)\varepsilon_{\theta_k}(t) \qquad (136)$$

dove $\psi_{\theta_k}(t)$ è il gradiente che si aggiorna iterativamente sostituendo in (135) allo stadio $k-$simo al parametro corrente $\theta = \theta_k$.

Data la stringa dei dati di ingresso $\mathbf{y} = [\mathbf{y}(1) \dots \mathbf{y}(N)]^\top$, e la stima θ_k alla $k-$sima iterazione,

1. Si calcola la stringa degli errori di predizione $\varepsilon_{\theta_k} = [\varepsilon_{\theta_k}(1) \dots \varepsilon_{\theta_k}(N)]^\top$ risolvendo l'equazione alle differenze (133). Lo schema esplicito richiede il calcolo dell'array $\varphi_{\theta_k} = [\varphi_{\theta_k}(1) \dots \varphi_{\theta_k}(N)]$

2. Si calcola il gradiente $\Psi_{\theta_k} = [\psi_{\theta_k}(1) \dots \psi_{\theta_k}(N)]$, usando l'equazione alle differenze (135) in cui $A_\theta(\rho z^{-1}) = A_{\theta_k}(\rho z^{-1})$.

3. si calcola la matrice pseudo-Hessiana

$$H_{\theta_k} := \sum_{t=1}^{N} \psi_{\theta_k}(t) \psi_{\theta_k}(t)^\top = \Psi_{\theta_k} \Psi_{\theta_k}^\top$$

   e la sua inversa $P_{\theta_k} := H_{\theta_k}^{-1}$. Questo calcolo si potrebbe anche organizzare in forma ricorsiva come visto nell'algoritmo generale PEM del capitolo precedente.

4. Si aggiorna θ_k usando la (136),

$$\theta_{k+1} = \theta_k + P_{\theta_k} \Psi_{\theta_k} \varepsilon_{\theta_k}$$

5. Si torna al passo 1) ponendo $\theta_k = \theta_{k+1}$.

Questo problema di stima è in genere mal condizionato. Nella funzione obiettivo si osserva un minimo molto pronunciato con una regione di attrazione che è tanto più piccola quanto più grande (prossimo a 1) si prende ρ. L'inizializzazione è quindi importante. Si può partire da delle stime iniziali ottenute da un periodogramma oppure per identificazione di un opportuno modello AR.

Fuori dalla regione di attrazione il gradiente è "piccolo" e la matrice Hessiana è mal condizionata per cui si possono avere esempi di convergenza estremamente lenta o di accumulo di errori di arrotondamento (con molte iterazioni). Per ovviare a questo problema si sceglie ρ variabile con il passo

di iterazione. Per k piccoli, quando le stime sono molto incerte, si prende ρ "piccolo" e poi lo si fa crescere con legge esponenziale, ad esempio

$$\rho(k+1) = \rho_0 \rho(k) + (1-\rho_0)\rho(\infty) \tag{137}$$

dove $\rho(\infty)$ è il valore a regime desiderato (Nehorai suggerisce 0.995) e ρ_0 è la costante di tempo che determina il tasso di crescita di $\rho(k)$. Nehorai suggerisce di prendere $\rho_0 \simeq 0.99$ e un valore iniziale $\rho(1) = 0.8$.

Questa versione dell'algoritmo non è "ricorsiva" come quella dell'articolo ma usa tutti i dati disponibili in "batch". Questo aggrava un pò i calcoli ma, in linea di principio, dovrebbe portare a prestazioni migliori. In ogni caso nel calcolo dell'errore di predizione e del gradiente i dati iniziali sono sempre male utilizzati e sarebbe opportuno usare un *fattore d'oblio* $\lambda(t)$ aggiornato con una relazione simile alla (137).

Considerazioni sulla stabilità dell'algoritmo

Come è facile intuire, con stime iniziali poco affidabili, il polinomio $A_{\theta_k}(z^{-1})$ potrebbe risultare instabile con conseguenze disastrose sul calcolo iterativo dell'errore di predizione e del gradiente. C'è però da notare che il polinomio che determina al passo k-simo la dinamica dell'errore di predizione e del gradiente (135), non è $A_{\theta_k}(z^{-1})$ ma bensì il polinomio "scalato" $A_{\theta_k}(\rho_k z^{-1})$ in cui il fattore $\rho_k < 1$ ha un effetto stabilizzante e può riportare i poli a modulo leggermente maggiore di uno dentro il cerchio unitario. Questo è probabilmente il motivo per cui, a quanto afferma Nehorai, non si osserva praticamente mai il fenomeno dell' instabilità.

Considerazioni sul Bias

L'introduzione del fattore di scala ρ nel modello (132) porta ad una descrizione, a stretto rigore "non corretta" del segnale che si vuole identificare. Si può calcolare esplicitamente l'errore asintotico (*bias*) che si commette nella stima PEM del termine $\cos\omega_k$ utilizzando il modello "scalato". Nel modello con una sola sinusoide gli autori di [**?**] trovano

$$\hat{\cos}\,\omega = \frac{(1+\rho^2)\cos\omega}{2\rho}$$

che con $\rho = \rho(\infty) = .99$ diventa $1.00005 \cos \omega$. Come si vede si tratta di errori tollerabili.

In molte situazioni pratiche il modello “sinusoidi in rumore bianco” potrebbe essere poco realistico e il termine d’errore $\mathbf{e}$ sarebbe più accuratamente descrivibile come rumore colorato. Se il rumore additivo non è bianco la modellizzazione mediante un processo ARMA come in (132) non è piu‘ valida e i presupposti del metodo vengono a cadere. Inoltre il mal condizionamento rende l’algoritmo delicato e possono essere necessari aggiustamenti *ad hoc* (regolarizzazione etc.) per i casi problematici.

ALGORITMI RICORSIVI

Motivazione: Identificazione in tempo reale per algoritmi adattativi.

Modello ARX

$$\mathbf{y}(t) = \varphi^\top(t)\theta + \mathbf{e}(t), \qquad t = 1,2,\ldots,$$

Stima con dati disponibili all'istante t:

$$\hat{\theta}(t) = \left[\sum_{s=1}^{t} \varphi(s)\varphi(s)^\top\right]^{-1} \sum_{s=1}^{t} \varphi(s)\mathbf{y}(s)$$

arriva il dato successivo: $(\mathbf{y}(t+1), \mathbf{u}(t))$; formiamo $\varphi(t+1)$ "shiftando" i dati

$$\hat{\theta}(t+1) = \left[\sum_{s=1}^{t} \varphi(s)\varphi(s)^\top + \varphi(t+1)\varphi(t+1)^\top\right]^{-1} \left[\sum_{s=1}^{t} \varphi(s)\mathbf{y}(s) + \varphi(t+1)\mathbf{y}(t+1)\right]$$

Vogliamo esprimere $\hat{\theta}(t+1)$ in funzione di $\hat{\theta}(t)$ e dei nuovi dati entranti.

Poniamo

$$P(t) := \left[\sum_{s=1}^{t} \varphi(s)\varphi(s)^\top\right]^{-1} \Rightarrow \hat{\theta}(t) = P(t)\sum_{s=1}^{t} \varphi(s)\mathbf{y}(s)$$

allora dal LEMMA DI INVERSIONE DI MATRICE:

$$[A+BCD]^{-1} = A^{-1} - A^{-1}B\,[C^{-1}+DA^{-1}B]^{-1}DA^{-1}$$

applicato a $P(t+1)^{-1} = P(t)^{-1} + \varphi(t+1)\varphi(t+1)^\top$ ponendo $A = P(t)^{-1}$, $B = \varphi(t+1) = D^\top$, e $C = 1$, si trova

$$\begin{aligned} P(t+1) &= \left[P(t)^{-1} + \varphi(t+1)\varphi(t+1)^\top\right]^{-1} \\ &= P(t) - P(t)\varphi(t+1)\left[1 + \varphi(t+1)^\top P(t)\varphi(t+1)\right]^{-1}\varphi(t+1)^\top P(t) \end{aligned}$$

Il termine tra parentesi quadre è uno scalare !
Sostituiamo nell'espressione di $\hat{\theta}(t+1)$:

$$\hat{\theta}(t+1) = P(t+1)\left[\sum_{s=1}^{t} \varphi(s)\mathbf{y}(s) + \varphi(t+1)\mathbf{y}(t+1)\right]$$

L'ALGORITMO RICORSIVO PER MODELLI ARX

Si trova dopo qualche passaggio

$$\hat{\theta}(t+1) = \hat{\theta}(t) + k(t)\left[\mathbf{y}(t+1) - \varphi(t+1)^{\top}\hat{\theta}(t)\right]$$

$$k(t) = P(t)\varphi(t+1)\frac{1}{1+\varphi(t+1)^{\top}P(t)\varphi(t+1)}$$

$$P(t+1) = P(t) - P(t)\varphi(t+1)\left[1+\varphi(t+1)^{\top}P(t)\varphi(t+1)\right]^{-1}\varphi(t+1)^{\top}P(t)\,.$$

Che sono le equazioni del Filtro di Kalman per il modello

$$\begin{cases} \boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t)\,, \qquad \theta(0) = \theta_0 \\ \mathbf{y}(t) = \varphi(t)^{\top}\theta(t) + \mathbf{e}(t) \end{cases}$$

$\varphi(t+1)^{\top}\hat{\theta}(t)$ è il predittore di $\mathbf{y}(t+1)$ basato sull'informazione disponibile all'istante t. Da notare che la varianza del rumore di misura λ_0^2 (che è incognita) non entra nelle equazioni. Al posto della varianza di rumore λ_0^2 c'è un 1.

L'ALGORITMO RICORSIVO PER MODELLI ARX

Definiamo la matrice

$$\Sigma(t) : \frac{\lambda_0^2}{t}\left[\frac{1}{t}\sum_{s=1}^{t}\varphi(s)\varphi(s)^\top\right]^{-1} = \lambda_0^2 P(t)$$

che è uno stimatore consistente della varianza asintotica di $\hat{\boldsymbol{\theta}}(t)$. Soddisfa l'equazione di Riccati del KF con λ_0^2 al posto dell'1:

$$\Sigma(t+1) = \Sigma(t) - \Sigma(t)\varphi(t+1)\left[\lambda_0^2 + \varphi(t+1)^\top\Sigma(t)\varphi(t+1)\right]^{-1}\varphi(t+1)^\top\Sigma(t)$$

Come si inizializza l'algoritmo? In teoria dovremmo aspettare fino all'istante t_0 in cui $\sum_{s=1}^{t_0}\varphi(s)\varphi(s)^\top$ diventa invertibile e poi calcolare $P(t_0)$ e $\hat{\theta}(t_0)$.

In realtà si prende $P(0) = \alpha I_p$, $\alpha > 0$ e $\hat{\theta}(0) = 0$.

COMPORTAMENTO ASINTOTICO

In realtà sappiamo già come si comporta lo stimatore. Con ipotesi di ergodicità del secondo ordine

$$\frac{1}{t}\sum_{s=1}^{t}\varphi(s)\varphi(s)^{\top} \rightarrow \mathbb{E}_0\varphi(s)\varphi(s)^{\top} := \bar{\Sigma}$$

Quindi: $P(t)$ e $\Sigma(t)$ tendono a zero per $t \rightarrow \infty$.

Si può mostrare che anche $k(t) \rightarrow 0$ per $t \rightarrow \infty$. Il filtro asintoticamente"si spegne" (ha imparato il valore vero θ_0).
Quindi il filtro non può inseguire parametri variabili nel tempo!
Per correr dietro a parametri variabili (lentamente) nel tempo bisogna mettere del rumore di modello! Bisogna descrivere il problema con l'approccio Bayesiano!

Prova che $k(t) \to 0$

L'equazione di Riccati

$$P(t+1) = P(t) - P(t)\varphi(t+1)\left[1 + \varphi(t+1)^\top P(t)\varphi(t+1)\right]^{-1}\varphi(t+1)^\top P(t)$$

si può riscrivere

$$P(t) - P(t+1) = k(t)\left[1 + \varphi(t+1)^\top P(t)\varphi(t+1)\right]k(t)^\top \geq k(t)k(t)^\top$$

Dato che $P(t) \to 0$ per $t \to \infty$ anche $k(t)k(t)^\top$ deve tendere a zero.

ALGORITMI RICORSIVI APPROSSIMATI

Per modelli a predittore non lineare nei parametri non esistono formule ricorsive esatte come per il modello ARX.
Si fanno approssimazioni; si trovano formule ricorsive che però non si sa più se convergono o no e se convergono al valore vero.
Ricordiamo l'algoritmo di quasi-Newton (104)

$$\theta_{k+1} - \theta_k = -H_{\theta_k}^{-1} \boldsymbol{\Psi}_{\theta_k} \boldsymbol{\varepsilon}_{\theta_k}$$

ovvero,

$$\theta_{k+1} - \theta_k = -\left[\sum_{s=1}^{N} \boldsymbol{\psi}_{\theta_k}(s) \boldsymbol{\psi}_{\theta_k}(s)^\top\right]^{-1} \sum_{s=1}^{N} \boldsymbol{\psi}_{\theta_k}(s) \boldsymbol{\varepsilon}_{\theta_k}(s)\,.$$

supponendo di avere dati fino a $N = t$, all'iterazione k-sima si ha

$$\hat{\boldsymbol{\theta}}_{k+1}(t) - \hat{\boldsymbol{\theta}}_k = -\left[\sum_{s=1}^{t} \boldsymbol{\psi}_{\hat{\boldsymbol{\theta}}_k}(s) \boldsymbol{\psi}_{\hat{\boldsymbol{\theta}}_k}(s)^\top\right]^{-1} \sum_{s=1}^{t} \boldsymbol{\psi}_{\hat{\boldsymbol{\theta}}_k}(s) \boldsymbol{\varepsilon}_{\hat{\boldsymbol{\theta}}_k}(s)\,.$$

Cerchiamo di ricavarne un algoritmo PEM ricorsivo approssimato. Si fa coincidere l'indice di iterazione col tempo: $k \equiv t$ e $\hat{\boldsymbol{\theta}}_k \equiv \hat{\boldsymbol{\theta}}(t)$

$$\hat{\boldsymbol{\theta}}(t+1) - \hat{\boldsymbol{\theta}}(t) = -\left[\sum_{s=1}^{t} \boldsymbol{\psi}_{\hat{\boldsymbol{\theta}}(t)}(s)\boldsymbol{\psi}_{\hat{\boldsymbol{\theta}}(t)}(s)^\top\right]^{-1} \sum_{s=1}^{t} \boldsymbol{\psi}_{\hat{\boldsymbol{\theta}}(t)}(s)\,\boldsymbol{\varepsilon}_{\hat{\boldsymbol{\theta}}(t)}(s)\,.$$

Cambiamo segno al gradiente

$$\hat{\boldsymbol{\varepsilon}}(t) \simeq \boldsymbol{\varepsilon}_{\hat{\boldsymbol{\theta}}(t)}(t)\,, \qquad \hat{\boldsymbol{\psi}}(t) := -\boldsymbol{\psi}_{\hat{\boldsymbol{\theta}}(t)}(t)$$

in modo da ottenere un algoritmo formalmente simile a quello per i modelli ARX

$$\hat{\boldsymbol{\theta}}(t+1) = \hat{\boldsymbol{\theta}}(t) + k(t)\left[\mathbf{y}(t+1) - \hat{\boldsymbol{\psi}}(t+1)^\top \hat{\boldsymbol{\theta}}(t)\right] \tag{138}$$

$$k(t) = P(t)\hat{\boldsymbol{\psi}}(t+1)\frac{1}{1+\hat{\boldsymbol{\psi}}(t+1)^\top P(t)\hat{\boldsymbol{\psi}}(t+1)} \tag{139}$$

$$P(t+1) = P(t) - P(t)\hat{\boldsymbol{\psi}}(t+1)\left[1+\hat{\boldsymbol{\psi}}(t+1)^\top P(t)\hat{\boldsymbol{\psi}}(t+1)\right]^{-1}\hat{\boldsymbol{\psi}}(t+1)^\top P(t)\,. \tag{140}$$

Come si calcolano $\boldsymbol{\varepsilon}_{\hat{\boldsymbol{\theta}}(t)}$, $\boldsymbol{\psi}_{\hat{\boldsymbol{\theta}}(t)}(t)$? Con $\theta = \theta_k$ fissato, ad ogni iterazione si risolveva un'equazione alle differenze usando tutti i dati di ingresso-uscita fino all'istante t. Per un modello B-J: Sostituire $\theta_k \to \hat{\theta}(t)$ nell'equazione

$$C_{\theta_k}(z^{-1})A_{\theta_k}(z^{-1})\boldsymbol{\varepsilon}_{\theta}(t) = A_{\theta_k}(z^{-1})D_{\theta_k}(z^{-1})\mathbf{y}(t) - B_{\theta_k}(z^{-1})D_{\theta_k}(z^{-1})\mathbf{u}(t)\,.$$

e risolvere di norma associata a condizioni iniziali nulle. Analogo discorso per il gradiente.

Quindi per calcolare $\boldsymbol{\varepsilon}_{\hat{\boldsymbol{\theta}}(t)}$ occorrerebbe una memoria che aumenta con t !
Vogliamo un algoritmo a memoria finita costane nel tempo !.

Inoltre con la definizione data il termine $\hat{\boldsymbol{\psi}}(t+1)\hat{\boldsymbol{\theta}}(t)$ **dipende da** $\hat{\boldsymbol{\theta}}(t+1)$!
Quindi per avere un algoritmo causale bisogna ridefinire $\hat{\boldsymbol{\psi}}(t)$ usando la stima del parametro all'istante precedente

$$\hat{\boldsymbol{\psi}}(t) \simeq -\boldsymbol{\psi}_{\hat{\boldsymbol{\theta}}(t-1)}(t)$$

PEM RICORSIVO PER MODELLI ARMAX

Si vogliono **algoritmi a memoria costante!** Sostituendo $\theta \equiv \hat{\theta}(t)$ in

$$C_\theta(z^{-1})\varepsilon_\theta(t) = A_\theta(z^{-1})y(t) - B_\theta(z^{-1})u(t) \qquad (\dagger)$$

si ottiene

$$C_{\hat{\theta}(t)}(z^{-1})\hat{\boldsymbol{\varepsilon}}(t) = A_{\hat{\theta}(t)}(z^{-1})y(t) - B_{\hat{\theta}(t)}(z^{-1})u(t)$$

ovvero

$$\hat{\boldsymbol{\varepsilon}}(t) = y(t) + \sum_{k=1}^{n} \hat{a}_k y(t-k) - \sum_{k=1}^{m} \hat{b}_k u(t-k) - \sum_{k=1}^{n} \hat{c}_k \hat{\boldsymbol{\varepsilon}}(t-k) := \mathbf{y}(t) - \hat{\boldsymbol{\varphi}}(t)^\top \hat{\boldsymbol{\theta}}(t)$$

dove

$$\hat{\boldsymbol{\varphi}}(t)^\top = \begin{bmatrix} -y(t-1) & \dots & -y(t-n) & u(t-1) & \dots & u(t-m) & \hat{\boldsymbol{\varepsilon}}(t-1) & \dots & \hat{\boldsymbol{\varepsilon}}(t-n) \end{bmatrix}$$

L'espressione in blu di $\hat{\boldsymbol{\varepsilon}}(t)$ ricorda più un errore di filtraggio. L'errore di predizione dovrebbe essere

$$\hat{\boldsymbol{\varepsilon}}(t) = y(t) - \hat{\boldsymbol{\varphi}}(t)^\top \hat{\theta}(t-1)$$

È conveniente usare un simbolo diverso. Usiamo $\tilde{\boldsymbol{\varepsilon}}(t-k)$ (che dipende da $\hat{\theta}(t-k)$) per l'errore di filtraggio per cui riscriveremo

$$\hat{\boldsymbol{\varphi}}(t)^{\top} = \begin{bmatrix} -y(t-1) & \dots & -y(t-n) & u(t-1) & \dots & u(t-m) & \tilde{\boldsymbol{\varepsilon}}(t-1) & \dots & \tilde{\boldsymbol{\varepsilon}}(t-n) \end{bmatrix}$$

Questo vettore dipende dai dati $(y^{t-1},\, u^{t-1})$ e dalle stime $\hat{\theta}(t-\ k\ );k = 1,2,\dots,n$ in istanti precedenti t. Si calcola l 'errore di filtraggio all'istante t con la

$$\tilde{\boldsymbol{\varepsilon}}(t) = \mathbf{y}(t) - \hat{\boldsymbol{\varphi}}(t)^{\top}\hat{\boldsymbol{\theta}}(t)$$

che serve per aggiornare $\hat{\boldsymbol{\varphi}}(t) \to \hat{\boldsymbol{\varphi}}(t+1)$ all'istante $t+1$.
L'equazione alle differenze per $\boldsymbol{\psi}_{\hat{\boldsymbol{\theta}}(t)}(t)$ si trova derivando $(\dagger)$ e sostituendo $\hat{\theta}(t)$ a θ :

$$C_{\hat{\theta}(t)}(z^{-1})\boldsymbol{\psi}_{\hat{\boldsymbol{\theta}}(t)}{}^{\top}(t) =$$
$$\begin{bmatrix} y(t-1) & \dots & y(t-n) & -u(t-1) & \dots & -u(t-m) & -\tilde{\boldsymbol{\varepsilon}}(t-1) & \dots & -\tilde{\boldsymbol{\varepsilon}}(t-n) \end{bmatrix}$$

come si vede questo $\boldsymbol{\psi}_{\hat{\boldsymbol{\theta}}(t)}(t)$ dipende dagli $\hat{c}_k(t)$ (e quindi da $\hat{\theta}(t)$).

Il gradiente "causale" $\hat{\boldsymbol{\psi}}(t+1)$ si calcola con la stessa formula in cui a destra gli argomenti temporali sono shiftati di un colpo:

$$C_{\hat{\theta}(t)}(z^{-1})\hat{\boldsymbol{\psi}}(t+1)^{\top}(t) = \hat{\boldsymbol{\varphi}}(t+1)^{\top} =$$
$$\begin{bmatrix} -y(t) & \dots & -y(t-n+1) & u(t) & \dots & u(t-m+1) & \tilde{\boldsymbol{\varepsilon}}(t) & \dots & \tilde{\boldsymbol{\varepsilon}}(t-n+1) \end{bmatrix} \quad (141)$$

Notare che anche $\tilde{\boldsymbol{\varepsilon}}(t)$ si calcola usando la stima corrente $\hat{\boldsymbol{\theta}}(t)$ nota all'istante t

Per aggiornare ricorsivamente la stima $\hat{\boldsymbol{\theta}}(t)$ a memoria costante con la (138) bisognerebbe quindi tenere in memoria, oltre a $\hat{\boldsymbol{\theta}}(t)$, i dati necessari per aggiornare la transizione $\hat{\boldsymbol{\varphi}}(t) \to \hat{\boldsymbol{\varphi}}(t+1)$ che sono:

· la stringa finita degli ultimi $n+m$ dati ingresso-uscita fino all'istante t,

· gli n campioni passati $\tilde{\boldsymbol{\varepsilon}}(t-1), \dots, \tilde{\boldsymbol{\varepsilon}}(t-n)$,

Inoltre:

· i $3n$ campioni passati delle prime componenti del gradiente $\hat{\boldsymbol{\psi}}(t), \dots, \hat{\boldsymbol{\psi}}(t-n+1)$, necessari per aggiornare la soluzione di (141).

L' algoritmo descritto qualche volta viene chiamato RAML, (Approximate recursive Maximum likelihood). La sua inizializzazione non è una faccenda banale che viene spesso risolta in modo brutale prendendo un $\theta(0)$ "opportuno" e ponendo tutte le altre condizioni iniziali uguali a zero. Ci sono poi vari modi di approssimare in modo causale le condizioni iniziali correnti ad esempio prendendo:

$$\tilde{\boldsymbol{\varepsilon}}(t-k) = z^{-k}\tilde{\boldsymbol{\varepsilon}}(t)\,, \qquad \hat{\boldsymbol{\psi}}(t-k) = z^{-k}\hat{\boldsymbol{\psi}}(t)$$

Si trovano in letteratura molte approssimazioni allo schema descritto [vedere il libro di Ljung e il Södeström-Stoica].

Dato che questi algoritmi non sono mai una realizzazione esatta dello stimatore PEM, non si può dire nulla sulle loro proprietà statistiche nè sulla loro convergenza. Esistono però condizioni sufficienti per garantire la convergenza ad un minimo locale ... [vedere ancora il libro di Ljung].

IL FATTORE D'OBLIO

Questi algoritmi "approssimati" maltrattano i dati per t piccolo!!! Occore un fattore di sconto per non fare pesare troppo l'influenza delle stime iniziali. Si minimizza un criterio scontato

$$V_t(\theta,\beta) := \sum_{s=1}^{t} \beta(t,s)\,\varepsilon_\theta(s)^2$$

in cui β è una funzione monotona crescente in s con $\beta(t,t)=1$. Si prende $\beta(t,s)$ con andamento simile a $\lambda^{t-s}, 0<\lambda<1$ ma con coefficiente tempo-variante che fornisca un andamento meno ripido: soluzione dell' equazione " all'indietro':

$$\beta(t,s-1)=\beta(t,s)\lambda(s); \qquad s<t; \qquad \beta(t,t)=1 \text{ per } s=t$$

che fornisce

$$\beta(t,t-1)=1\cdot\lambda(t),\ \beta(t,t-2)=\lambda(t)\lambda(t-1),\ldots, \quad \beta(t,s)=\prod_{k=s+1}^{t}\lambda(k)$$

Il coefficiente λ si fa partire da un valore iniziale $\lambda(0)$ prossimo a uno e si fa convergere ad un valore molto prossimo a 1, e.g. $1-\lambda_0 > \lambda(0)$ con un andamento descritto dalla

$$\lambda(k) = \lambda_0 \lambda(k-1) + (1-\lambda_0), \qquad \lambda_0 \simeq 0.99 \quad \lambda(0) \simeq 0.95$$

Questi coefficienti sono stati aggiustati "sperimentalmente" su molti esempi "tipici".
Si potrebbe anche definire β mediante un'equazione duale "in avanti"

$$\beta(t+1,s) = \lambda(t)\beta(t,s); \qquad t > s; \qquad \beta(s,s) = 1 \text{ per } t = s$$

che fornisce

$$\beta(s+1,s) = \lambda(s),\ \beta(s+2,s) = \lambda(s+1)\lambda(s), \ldots, \quad \beta(t,s) = \prod_{k=s}^{t-1} \lambda(k)$$

l'ultima espressione vale per $t > s$; ovviamente si tiene conto che $\beta(t,t) = 1$.

ALGORITMI CON FATTORE D'OBLIO

L'ottimizzazione con fattore di sconto **per modelli ARX** è trattata nel Problema 4. Usando il lemma d'inversione si arriverebbe ad una equazione ricorsiva per $P(t)$ del tipo

$$P(t+1) = \frac{1}{\lambda(t)} \left[P(t) - P(t)\varphi(t+1) \left[\lambda(t) + \varphi(t+1)^\top P(t) \varphi(t+1) \right]^{-1} \varphi(t+1)^\top P(t) \right]$$

e all'espressione del guadagno con fattore d'oblio

$$k(t) = P(t)\varphi(t+1) \frac{1}{\lambda(t) + \varphi(t+1)^\top P(t) \varphi(t+1)}.$$

Naturalmente per modelli ARX il fattore d'oblio in realtà non serve. Per modelli a predittore non lineare si deve quindi sostituire a $\varphi(t)$ in queste espressioni il gradiente $\hat{\psi}(t)$ calcolato nella stima corrente di θ (il gradiente $\hat{\psi}(t)$ va a sostituire $\varphi(t)$).

STIMA DELLA COMPLESSITÀ DI UN MODELLO LINEARE

In molti casi il numero di parametri, p nel modello lineare $\mathbf{y} = S\theta + \sigma\mathbf{w}$ non è un dato del problema assegnato a priori, ma piuttosto un parametro che deve essere variato per confrontare l'adeguatezza di modelli più o meno complicati a descrivere i dati di misura.
Nella statistica classica Fisheriana, il problema della scelta ottima di p è un *problema di verifica d'ipotesi*: in base ai dati osservati decidere se il "modello vero" che li ha generati ha complessità p pari ad uno dei numeri naturali compresi in un certo intervallo $[p_{\min}, p_{\max}]$.

Però l'uso di test statistici richiede di fissare a priori una probabilità d'errore α in modo essenzialmente arbitrario, con risultanti ambiguità.

Se la numerosità campionaria è fissa (cosa che da ora in avanti supporremo), è ovvio che all'aumentare di p si ottiene una descrizione sempre migliore dei dati, nel senso che l'errore quadratico medio

$$\hat{\sigma}^2(\mathbf{y}) = \frac{1}{N} \|y - S\hat{\theta}(\mathbf{y})\|^2_{\mathbb{R}^{-1}}$$

diminuisce all'aumentare di p **fino a diventare addirittura zero nel caso limite** $p = N$. Però, a parità di misure disponibili, la qualità delle stime ottenute, misurata ad esempio dalla varianza dei parametri stimati si deteriora all'aumentare di p. Al limite, per p molto grande, il "fit" perfetto ottenuto usando un elevatissimo numero di parametri è in pratica di nessuna utilità dato che la grande varianza delle stime rende inservibile il modello (il modello verrà usato poi per descrivere dati *diversi* da quelli usati in fase di stima).

Notiamo che il problema è adesso inquadrato in un'ottica diversa da quella Fisheriana, senza cioè assumere che esista necessariamente un modello vero di dimensione finita che ha generato i dati. In questo caso i modelli di diversa complessità sono da interpretare solo come **approssimazioni** usate per descrivere dei dati $\mathbf{y}$ che potrebbero anche non avere una descrizione "vera" del tipo ipotizzato ma richiedere anche un numero infinito di parametri. In questo caso (che è quello realistico) non esiste un modello stimato di dimensione finita che possa descrivere i dati in modo consistente. C'è quindi sempre un errore di modellizzazione (*bias*) che è diverso da zero anche con dati infiniti.

È quindi necessario procedere per tentativi successivi, aumentando p fino a che il compromesso raggiunto tra bontà del "fit" e varianza della stima sembra accettabile. Questo è il celebre

bias versus variance dilemma

CONFRONTO DI PARAMETRIZZAZIONI

Vogliamo confrontare le varianze degli stimatori di due diverse parametrizzazioni usate per descrivere gli stessi dati. Inizialmente formuleremo il problema in termini di confronto tra due sole alternative possibili. Consideriamo due modelli lineari statici in forma standard

$$\begin{aligned} M_1: \quad & \mathbf{y} = S_1\theta_1 + \sigma\mathbf{w} && \theta_1 \in \mathbb{R}^p \\ M_2: \quad & \mathbf{y} = [S_1\, S_2]\begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix} + \sigma\mathbf{w} && \theta_2 \in \mathbb{R}^k. \end{aligned} \tag{142}$$

In entrambi i casi $\mathbf{w}$ è a media nulla e varianza I_N (matrice identità $N \times N$).

Nel modello più "semplice" M_1, supporremo che $\text{rango}\, S_1 = p$.
Nel modello "complicato" M_2, senza perdita di generalità assumiamo che

$$\text{rango}\,[S_1\, S_2] = p + k \quad . \tag{143}$$

Se ciò non accade, il modello può facilmente essere riparametrizzato eliminando le colonne di S_2 che sono linearmente dipendenti e ridefinendo opportunamente θ_2.

REGRESSIONE LINEARE A STADI

Cercheremo di derivare delle formule per le stime dei parametri e per la varianza del modello M_2 che esprimano queste quantità come correzioni apportate alla stima e alla varianza del parametro θ_1 nel modello M_1. Questo procedimento va sotto il nome di *regressione (ai M.Q.) a stadi*.

Indichiamo con $\mathscr{S}$ lo spazio colonne della matrice $S := [S_1\, S_2]$ e con θ il parametro $p+k$ dimensionale $[\theta_1^\top\, \theta_2^\top]^\top$ che compare nella (142). Naturalmente la stima ai M.Q. (di Markov) di θ è definita dalle solite formule,

$$\begin{aligned} \hat{\theta}(y) &= (S^\top S)^{-1} S^\top y \\ \operatorname{Var} \hat{\theta} &= \sigma^2 (S^\top S)^{-1} \quad , \end{aligned}$$

nelle quali però le matrici da invertire sono ora di dimensione $(p+k)\times(p+k)$. Vogliamo mettere in evidenza come si modifica la stima di θ_1 relativa al modello di ordine p per effetto dell'aggiunta dei k ulteriori parametri.

Per la (143) $\mathscr{S}$ si può decomporre in somma diretta

$$\text{span}\,[S] = \text{span}\,[S_1\,S_2] = \mathscr{S}_1 \oplus \mathscr{S}_2 = \text{span}\,[S_1] \oplus \text{span}\,[S_2] \tag{144}$$

e questa decomposizione può essere resa *ortogonale* se si introducono i due proiettori complementari

$$\begin{array}{lll} P_1 : \mathbb{R}^N \to \mathscr{S}_1 & , & P_1 = S_1 (S_1^\top S_1)^{-1} S_1^\top \quad , \\ P_1^\perp : \mathbb{R}^N \to \mathscr{S}_1^\perp & , & P_1^\perp = I - S_1 (S_1^\top S_1)^{-1} S_1^\top \quad . \end{array}$$

Per semplificare le notazioni in seguito denoteremo con Q_1 la matrice $P_1^\perp$. Dato che $P_1 + Q_1 = I$, si ha

$$S_2 = P_1 S_2 + Q_1 S_2$$

e siccome le colonne di $P_1 S_2$ stanno per definizione in $\mathscr{S}_1$, l'ultimo addendo della (144) può venire sostituito da $\text{span}\,[Q_1\,S_2]$. Quindi

$$\text{span}\,[S] = \text{span}\,[S_1] \overset{\perp}{\oplus} \text{span}\,[Q_1\,S_2] \tag{145}$$

dove il simbolo $\overset{\perp}{\oplus}$ sta per somma diretta ortogonale.

Sia $\hat{y}$ la proiezione ortogonale di y sullo spazio colonne, $\mathscr{S}$, della matrice S. Per l'indipendenza lineare delle colonne di S_1 e S_2 si può esprimere in modo unico $\hat{y}$ nella forma

$$\hat{y} = S_1 \hat{\theta}_1 + S_2 \hat{\theta}_2 \quad , \tag{146}$$

dove $\hat{\theta}_1$ e $\hat{\theta}_2$ sono vettori che rappresentano i corrispondenti coefficienti nelle combinazioni lineari delle colonne di S_1 ed S_2. Ovviamente $\hat{\theta}_1$ e $\hat{\theta}_2$ sono proprio le stime dei parametri θ_1 e θ_2 nel modello a $p+k$ parametri.

Per il principio di ortogonalità dovrà essere $y - \hat{y} \perp \mathscr{S}$ e quindi anche, separatamente,

$$y - \hat{y} \perp \mathscr{S}_1 \quad , \qquad y - \hat{y} \perp Q_1 \mathscr{S}_2 \quad ,$$

che si riscrivono

$$S_1^\top (y - S_1 \hat{\theta}_1 - S_2 \hat{\theta}_2) = 0 \quad ,$$
$$S_2^\top Q_1 (y - S_1 \hat{\theta}_1 - S_2 \hat{\theta}_2) = 0 \quad .$$

Queste formule forniscono

$$\hat{\theta}_1 = (S_1^\top S_1)^{-1} S_1^\top \left[y - S_2 \hat{\theta}_2 \right] \quad , \tag{147}$$

$$\hat{\theta}_2 = (S_2^\top Q_1 S_2)^{-1} S_2^\top Q_1 y \, . \tag{148}$$

Fatto : $S_2^\top Q_1 S_2$ è invertibile.
Questo si mostra facilmente ricordando che Q_1 è un proiettore. In effetti

$$a^\top S_2^\top Q_1 S_2 a = 0 \Rightarrow a^\top S_2^\top Q_1^\top Q_1 S_2 a = \|Q_1 S_2 a\|^2 = 0$$

e pertanto $S_2 a$ deve stare nello spazio nullo di $Q_1 = P_1^\perp$. Dato che $\mathrm{Ker}\,(P_1^\perp) = \Im(P_1) = \mathscr{S}_1 = \mathrm{span}\,[S_1]$, segue che $S_2 a \in \mathrm{span}\,[S_1]$, ma questo può accadere solo se $a = 0$, dato che le colonne di S_1 ed S_2 sono indipendenti. $\diamond$

Se indichiamo con il simbolo $\bar{\theta}_1$ la stima di θ_1 ottenuta descrivendo i dati con un modello lineare a p parametri del tipo M_1, la (147) può essere riscritta come

$$\hat{\theta}_1 = \bar{\theta}_1 - (S_1^\top S_1)^{-1} S_1^\top S_2 \hat{\theta}_2 \quad . \tag{149}$$

che esprime la stima di θ_1 ottenuta con il modello lineare a $p+k$ parametri, come la somma di $\bar{\theta}_1$ e di un termine di correzione dovuto all'introduzione del parametro ulteriore θ_2.

INTERPRETAZIONE GEOMETRICA

Nella decomposizione (146) i due addendi $S_1\hat{\theta}_1$ e $S_2\hat{\theta}_2$ hanno il significato geometrico di *proiezioni oblique* rispettivamente di y su $\mathscr{S}_1$ lungo $\mathscr{S}_2$ e di y su S_2 lungo $\mathscr{S}_1$.

Dalla formula (148) si vede in particolare che $\hat{\theta}_2$ si può ricavare dalla relazione di ortogonalità

$$Q_1 y - S_2\theta_2 \perp Q_1 S_2$$

di modo che la proiezione obliqua di y su $\mathscr{S}_2$ lungo $\mathscr{S}_1$, si può *calcolare* facendo prima la proiezione *ortogonale* di $Q_1 y = y - P_1 y$ sul sottospazio $(I - P_1)\mathscr{S}_2 = Q_1\mathscr{S}_2$ (che è calcolabile risolvendo un problema di minimi quadrati ordinari) e poi moltiplicando per S_2 il parametro $\hat{\theta}_2$ trovato in questo modo *. La matrice di *proiezione obliqua su* S_2 *lungo* $\mathscr{S}_1$ ha così la rappresentazione

$$P_{2\|1} := S_2(S_2^\top Q_1 S_2)^{-1} S_2^\top Q_1 \tag{150}$$

*Per evitare interpretazioni errate notiamo che $S_2\hat{\theta}_2$ *non può essere la proezione ortogonale di di* $Q_1 y = y - P_1 y$ *sul sottospazio* $Q_1\mathscr{S}_2$. In effetti quest'ultimo non è nemmeno un sottospazio di $\mathscr{S}_2$.

usando la quale si controlla facilmente che in effetti $P_{2\|1}^2 = P_{2\|1}$, mentre

$$P_{2\|1}^\top Q_1 = Q_1 P_{2\|1}\,.$$

la quale, visto che Q_1 è un proiettore ortogonale e quindi $Q_1 = Q_1^\top$, si può riscrivere come $(Q_1 P_{2\|1})^\top = P_{2\|1}^\top Q_1^\top = Q_1 P_{2\|1}$, i.e. $Q_1 P_{2\|1}$ è simmetrica (e idempotente) e quindi è essa stessa un *proiettore ortogonale* che, per forza di cose, deve proiettare sul sottospazio $Q_1\mathscr{S}_2$, che è il complemento ortogonale di $\mathscr{S}_1$ in $\mathscr{S}$. Infatti:

Proposizione 29 *Sia P la matrice proiezione ortogonale da $\mathbb{R}^N$ sullo spazio $\mathscr{S}$ e P_1 quella sul sottospazio $\mathscr{S}_1 \subset \mathscr{S}$. Allora $P - P_1$ è il proiettore ortogonale che proietta sul complemento ortogonale $\mathscr{S} \cap \mathscr{S}_1^\perp$ e che ha la rappresentazione*

$$P - P_1 = Q_1 P_{2\|1} \tag{151}$$

dove $P_{2\|1}$ è il proiettore obliquo definito in (150)*.*

Prova Basta dimostrare la (151). Usando le formule (**??**) e (147) si ottiene

$$\begin{aligned}\hat{y} = Py &= S_1(S_1^\top S_1)^{-1} S_1^\top y - S_1(S_1^\top S_1)^{-1} S_1^\top S_2 \hat{\theta}_2(y) + S_2\hat{\theta}_2(y) \\ &= P_1 y + \left[I - S_1(S_1^\top S_1)^{-1} S_1^\top\right] S_2\hat{\theta}_2(y) \\ &= (P_1 + Q_1 P_{2\|1})\, y\end{aligned}$$

per cui effettivamente si ha $P - P_1 = Q_1 P_{2\|1}$. La decomposizione $P = P_1 + Q_1 P_{2\|1}$ è ovviamente ortogonale, stante che $P_1^\top (P - P_1) = P_1 Q_1 P_{2\|1} = 0$. Notiamo che un'affermazione equivalente è la $\mathscr{S} = P_1\mathscr{S} \oplus \mathscr{S} \cap \mathscr{S}_1^\perp$. □

Problema 13 Verificare che $P_{2\|1}$ è idempotente, il suo nucleo è S_1 e la sua immagine è lo spazio colonne di S_2.

Si può dare una rappresentazione del tutto analoga della proiezione obliqua di y su $\mathscr{S}_1$ lungo $\mathscr{S}_2$ e arrivare ad una rappresentazione esplicita della decomposizione (146), del tipo

$$y = P_{1\|2} y + P_{2\|1} y = S_1(S_1^\top Q_2 S_1)^{-1} S_1^\top Q_2\, y + S_2(S_2^\top Q_1 S_2)^{-1} S_2^\top Q_1\, y \quad (152)$$

dove Q_2 ha un significato duale a Q_1. Questa espressione è forse più semplice della decomposizione ortogonale che abbiamo illustrato sopra ma è meno comoda da usare perchè non è ortogonale.

Figura 5.3 (proiezione obliqua)

CONFRONTO DELLE VARIANZE

Concentriamoci ora sul calcolo delle varianze degli stimatori. Introduciamo allo scopo le seguenti notazioni:

$$\begin{aligned} \bar{\Sigma}_1 &:= [S_1^\top S_1]^{-1} \\ A_1 &:= [S_1^\top S_1]^{-1} S_1^\top \\ \Sigma_2 &:= [S_2^\top Q_1 S_2]^{-1} \quad ; \end{aligned}$$

ovviamente, $\bar{\theta}_1 = A_1 y$ e $\mathrm{Var}_{\theta_1} \bar{\theta}_1 = \sigma^2 \bar{\Sigma}_1$. Nel seguito i pedici $_{\theta_1}$ e $_\theta$ staranno ad indicare il modello rispetto a cui si calcola l'aspettazione (e quindi la varianza).

Proposizione 30 *Siano $\hat{\theta}_1(\mathbf{y})$ e $\hat{\theta}_2(\mathbf{y})$ gli stimatori di Markov definiti dalle formule (147) e (148). Si ha allora:*

$$\mathrm{Var}_\theta \left\{ \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} \right\} = \sigma^2 \begin{bmatrix} \bar{\Sigma}_1 + A_1 S_2 \Sigma_2 S_2^\top A_1^\top & -A_1 S_2 \Sigma_2 \\ -\Sigma_2 S_2^\top A_1^\top & \Sigma_2 \end{bmatrix} \quad . \qquad (153)$$

Prova: Incominciamo col dimostrare che $\mathrm{Var}_\theta[\hat{\theta}_2] = \sigma^2\Sigma_2$. Dalla (148) si ha

$$\mathrm{Var}_\theta[\hat{\theta}_2] = \Sigma_2 S_2^\top Q_1 \mathrm{Var}_\theta[\mathbf{y}] Q_1 S_2 \Sigma_2 = \sigma^2 \Sigma_2 S_2^\top Q_1 S_2 \Sigma_2 = \sigma^2 \Sigma_2 \quad ,$$

dato che $\mathrm{Var}_\theta[\mathbf{y}] = \sigma^2 I$ ed Q_1 è idempotente.

Mostriamo ora che *i due stimatori* $\bar{\theta}_1(\mathbf{y})$ *e* $\hat{\theta}_2(\mathbf{y})$ *sono scorrelati*. Si ha infatti:

$$\mathrm{Cov}_\theta\left[\bar{\theta}_1(\mathbf{y}), \hat{\theta}_2(\mathbf{y})\right] = \bar{\Sigma}_1 S_1^\top \mathrm{Var}_\theta[\mathbf{y}] Q_1 S_2 \Sigma_2 = \sigma^2 \bar{\Sigma}_1 S_1^\top Q_1 S_2 \Sigma_2 = 0 \quad ,$$

perchè $S_1^\top Q_1 = Q_1 S_1 = 0$.

Usando ora la (149) si trova

$$\mathrm{Cov}_\theta\left[\hat{\theta}_1(\mathbf{y}), \hat{\theta}_2(\mathbf{y})\right] = -A_1 S_2 \mathrm{Var}_\theta[\hat{\theta}_2] = -\sigma^2 A_1 S_2 \Sigma_2 \quad .$$

Calcoliamo infine $\mathrm{Var}_\theta[\hat{\theta}_1(\mathbf{y})]$. Dato che $\bar{\theta}_1(\mathbf{y})$ e $\hat{\theta}_2(\mathbf{y})$ sono scorrelati, si ha

$$\begin{aligned} \mathrm{Var}_\theta\left[\bar{\theta}_1(\mathbf{y}) - A_1 S_2 \hat{\theta}_2(\mathbf{y})\right] &= \mathrm{Var}_\theta\left[\bar{\theta}_1(\mathbf{y})\right] + A_1 S_2 \mathrm{Var}_\theta\left[\hat{\theta}_2(\mathbf{y})\right] S_2^\top A_1^\top \\ &= \sigma^2 [\bar{\Sigma}_1 + A_1 S_2 \Sigma_2 S_2^\top A_1^\top] \quad . \end{aligned}$$

che conclude la dimostrazione della formula (153). □

La formula (153) descrive l'effetto dell'aumento del numero di parametri nel modello sulla varianza delle stime. In particolare mostra che la stima $\hat{\theta}_1$ di θ_1 nel modello maggiorato è generalmente "peggiore" della prima in termini di varianza. La varianza, Σ_1, di $\hat{\theta}_1$ è in effetti *più grande* di quella di $\bar{\theta}_1$, essendo

$$\Sigma_1 = \bar{\Sigma}_1 + A_1 S_2 \, \Sigma_2 \, S_2^\top A_1^\top$$

e il termine che si somma a $\bar{\Sigma}_1$ è in generale non nullo.

Purtroppo però la varianza delle stime del parametro θ non è un criterio"oggettivo" per arrivare alla scelta dell'ordine del modello. Infatti, se le colonne di S_2 sono *ortogonali* a $\mathscr{S}_1$, ovvero

$$S_1^\top S_2 = 0 \qquad (S_2^\top S_1 = 0)$$

le formule si semplificano (dato che $Q_1 S_2 = S_2$) e i due stimatori $\hat{\theta}_1$ e $\hat{\theta}_2$ si possono calcolare indipendentemente l'uno dall'altro con le solita formule,

$$\hat{\theta}_i(\mathbf{y}) = (S_i^\top S_i)^{-1} S_i^\top \mathbf{y} \quad , \qquad i = 1,2 \quad .$$

In particolare si trova $\hat{\theta}_1 = \bar{\theta}_1$ e quindi anche $\Sigma_1 = \bar{\Sigma}_1$.

Per comprendere questo fenomeno (che a prima vista può sembrare sconcertante) basta pensare che ci sono molte parametrizzazioni del modello "ideale" $S\theta$ che sono assolutamente equivalenti agli effetti di descrivere i dati y. Per esempio, introducendo una fattorizzazione QR di S, vede facilmente che si può sempre fattorizzare S come prodotto di una matrice a colonne ortogonali (le prime $p+k$ colonne di Q) per una matrice quadrata $R \in \mathbb{R}^{p+k \times p+k}$ non singolare (a struttura triangolare inferiore). Definendo il nuovo parametro $\beta := R\theta$ si può riparametrizzare il modello in modo tale che le colonne di S siano ortogonali. In questo caso la varianza di $\hat{\beta}_1$ non aumenta aumentando la parametrizzazione del modello con k nuovi parametri.

La morale della storia è che la varianza delle stime dei parametri *dipende dal sistema di coordinate scelto per rappresentare il modello* (in breve, "dalla base"). I confronti dovrebbero essere quindi fatti solo tra quantità che sono *invarianti per cambio di base*. Quantità di questo genere sono ad esempio **gli errori residui di modellizzazione**. □

LA “CROSS VALIDATION”

Dato che i modelli servono in ultima analisi a costruire predittori per dati “futuri” (non ancora osservati) si può allora porre un problema di scelta del modello che fornisce l’*approssimazione ottima dei dati* (non di un ipotetico modello vero). Si sceglierà così quel modello che dà *la migliore predizione dei dati futuri*. Beninteso l’errore di predizione dovrà qui tener conto anche dell’ **incertezza introdotta nel modello usato per la predizione dal fatto che esso usa necessariamente un parametro stimato** che è, esso stesso, una variabile aleatoria.

La bontà di un modello stimato non si può giudicare solo dall’accuratezza con cui esso esegue il *fit* dei dati usati per l’identificazione ma dalla bontà con cui il modello stimato riesce a descrivere dati *futuri*, non usati per l’identificazione .

Questa posizione del problema che verrà ripresa in modo più preciso più avanti, conduce alle soluzioni moderne del problema della stima dell’ordine.

IL CRITERIO FPE

Supponiamo di avere a disposizione due vettori di osservazioni $\mathbf{y} := [\mathbf{y}_1^\top \; \mathbf{y}_2^\top]^\top$ che per semplicità assumeremo di ugual dimensione N e di usare i primi N dati $\mathbf{y}_1$ per l'identificazione di un generico modello lineare standard di dimensione p. I dati $\mathbf{y}_1$ sono descritti dal modello lineare

$$\mathbf{y}_1 = S\theta + \mathbf{w}_1\,, \qquad \mathrm{Var}\,[\mathbf{w}_1] = \sigma^2 I_N \tag{154}$$

ottenedo il classico stimatore $\hat{\theta}(\mathbf{y}_1) = [S^\top S]^{-1} S^\top \mathbf{y}_1$.
Vogliamo ora valutare la "bontà statistica" del modello stimato, $S\hat{\theta}(\mathbf{y}_1)$ per descrivere i dati $\mathbf{y}_2$ che abbiamo tenuto da parte.

Perchè questa operazione abbia senso dobbiamo supporre che i dati nei successivi N campioni siano stati *generati dallo stesso meccanismo che ha generato* $\mathbf{y}_1$, il che si può esprimere dicendo che che le d.d.p. (o almeno le statistiche del priimo e secondo ordine) di $\mathbf{y}_1$ e $\mathbf{y}_2$ debbono essere le stesse.

Supporremo che le due componenti del vettore $[\mathbf{y}_1^\top \ \mathbf{y}_2^\top]^\top$ abbiano lo stesso vettore di media μ (che potrebbe essere qualunque) e che la varianza complessiva di $\mathbf{y}$ sia $\sigma^2 I_{2N}$. In questo modo $\mathbf{y}_1$ e $\mathbf{y}_2$ risultano scorrelati.

Consideriamo allora il cosiddetto *errore finale di predizione* dei dati futuri (vettoriale)

$$\boldsymbol{\varepsilon} := \mathbf{y}_2 - S\hat{\theta}(\mathbf{y}_1) \tag{155}$$

che ha media $\mu - S[S^\top S]^{-1}S^\top \mu$ per cui sottraendo la media e calcolando la varianza di $\boldsymbol{\varepsilon}$ si trova

$$\mathrm{Var}\,[\boldsymbol{\varepsilon}] = \sigma^2 I_N + S[S^\top S]^{-1}S^\top \sigma^2 I_N S[S^\top S]^{-1}S^\top = \sigma^2 \left[I_N + S[S^\top S]^{-1}S^\top \right].$$

Come misura dell'errore finale di predizione prendiamo la varianza scalare normalizzata che ha l'espressione

$$\begin{aligned} \frac{1}{N}\mathrm{var}[\boldsymbol{\varepsilon}] \ &= \sigma^2 \frac{1}{N}\mathrm{Tr}\left\{ I_N + S[S^\top S]^{-1}S^\top \right\} = \sigma^2 \left\{ 1 + \frac{1}{N}\mathrm{Tr}([S^\top S]^{-1}S^\top S) \right\} \\ &= \sigma^2 (1 + \frac{p}{N}) \end{aligned} \tag{156}$$

Si vede che la varianza dell'errore finale di predizione **cresce linearmente con** p. La varianza σ^2, è un parametro incognito, e dobbiamo usare una sua stima, naturalmente anch'essa basata su un modello a p parametri. Usiamo lo stimatore corretto della varianza

$$\frac{N}{N-p}\hat{\sigma}_p^2 = \frac{1}{N-p}\|\mathbf{y}_1 - S\hat{\theta}(\mathbf{y}_1)\|^2 = \frac{1}{N-p}\|\hat{\boldsymbol{\varepsilon}}_p\|^2$$

dove $\hat{\boldsymbol{\varepsilon}}_p$ è il residuo di stima nel modello a p parametri. si arriva a definire l'indice

$$FPE(p) := \frac{1}{N}\|\hat{\boldsymbol{\varepsilon}}_p\|^2 \frac{(1+\frac{p}{N})}{(1-\frac{p}{N})} := \hat{\sigma}_p^2 \frac{(1+\frac{p}{N})}{(1-\frac{p}{N})} \tag{157}$$

che viene anch'esso chiamato **errore finale di predizione** basato su un modello di dimensione p.

La stima dell'ordine del modello può essere basata sulla minimizzazione di questo indice. Occorre preliminarmente identificare un certo numero di modelli di ordine crescente in un intervallo di valori plausibili di p e calcolare il relativo errore residuo quadratico medio. I calcoli si possono organizzare in modo efficiente usando algoritmi M.Q. a stadi.

IL CRITERIO FPE PER MODELLI DINAMICI

Supponiamo per il momento di avere una classe di modelli dinamici lineari a p parametri liberi e che lo stimatore PEM $\hat{\boldsymbol{\theta}}_N$, del parametro θ soddisfi le ipotesi di consistenza e normalità asintotica. Abbiamo dati $\{\mathbf{y}^{N_1}, \mathbf{u}^{N_1}\}$ sull' intervallo $[1,\, N_1]$, con cui costruiamo lo stimatore PEM $\hat{\boldsymbol{\theta}}_1$ e vogliamo descrivere col modello $M(\hat{\boldsymbol{\theta}}_1)$ dei dati $\{\mathbf{y}^{N_2}, \mathbf{u}^{N_2}\}$ su un intervallo "lontano" $[t_0+1,\, t_0+N_2]$.
Per l'ergodicità possiamo supporre che i due sets di dati siano approssimativamente **indipendenti**.

Calcoliamo la varianza dell'errore **finale** (asintotico) di predizione usando lo stimatore $\hat{\boldsymbol{\theta}}_1$ supponendo $t_0 \to \infty$.

$$W_{N_1}(p) := \mathbb{E}\left[\boldsymbol{\varepsilon}^2(t, \hat{\boldsymbol{\theta}}_1)\right], \qquad t > t_0 >> N_1$$

in cui $\boldsymbol{\varepsilon}^2(t, \hat{\boldsymbol{\theta}}_1)$ dipende dai dati in $[1,\, N_1]$ attraverso lo stimatore $\hat{\boldsymbol{\theta}}_1 = \hat{\boldsymbol{\theta}}_1(\mathbf{y}^{N_1}, \mathbf{u}^{N_1})$ e dai dati $\{\mathbf{y}^{N_2}, \mathbf{u}^{N_2}\}$ usati per calcolare il predittore.

Per $N_1 \to \infty$ si ha $\hat{\boldsymbol{\theta}}_1 \to \theta_0$ e si può approssimare

$$W_{N_1}(p) \simeq \mathbb{E}\left[\boldsymbol{\varepsilon}(t,\theta_0) + \frac{\partial \boldsymbol{\varepsilon}(t,\theta)}{\partial \theta}_{|\theta=\theta_0}(\hat{\boldsymbol{\theta}}_1 - \theta_0)\right]^2 = \mathbb{E}\left[\mathbf{e}_0(t) + \boldsymbol{\psi}_{\theta_0}^{\top}(t)(\hat{\boldsymbol{\theta}}_1 - \theta_0)\right]^2$$
$$= \lambda_0^2 + \mathbb{E}\left\{\mathrm{Tr}\left[\boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^{\top}\ (\hat{\boldsymbol{\theta}}_1 - \theta_0)(\hat{\boldsymbol{\theta}}_1 - \theta_0)^{\top}\right]\right\}.$$

Scambiamo le operazioni di Traccia (che è lineare) e di aspettazione. Per l'indipendenza l'aspettazione è il prodotto di due aspettazioni. Quella rispetto ai dati in $[1,\, N_1\,]$, per $N_1 \to \infty$, è

$$\mathbb{E}\,(\hat{\boldsymbol{\theta}}_1 - \theta_0)(\hat{\boldsymbol{\theta}}_1 - \theta_0)^{\top} \simeq \frac{\lambda_0^2}{N_1}\left\{\mathbb{E}\,\boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^{\top}\right\}^{-1}$$

e quindi si trova

$$W_{N_1}(p) \simeq \lambda_0^2(1 + \frac{p}{N_1}),$$

Prendendo come stimatore corretto di λ_0^2, l'errore quadratico medio residuo

diviso per $N_1 - p$ si trova la formula asintotica (valida per $N_1 \to \infty$)

$$W_{N_1}(p) \simeq V_{N_1}(\hat{\boldsymbol{\theta}}_1)\frac{(1+\frac{p}{N_1})}{(1-\frac{p}{N_1})} \qquad (FPE)$$

Si vede che al crescere di p l'errore quadratico medio di predizione $V_{N_1}(\hat{\boldsymbol{\theta}}_1)$ diminuisce ma il termine moltiplicativo aumenta (all'incirca linermente) con p. Naturalmente per usare questo criterio di stima bisogna identificare una serie di modelli a diversa complessità p e andare poi a scegliere quello a minimo FPE.

IL CRITERIO AIC E L' MDL

Dato che normalmente $N >> p$ s ipuò approssimare

$$\frac{(1+\frac{p}{N})}{(1-\frac{p}{N})} \simeq (1+\frac{2p}{N})$$

e prendendo i logaritmi, si trova il **criterio AIC di Akaike:**

$$AIC := \log V_N(\hat{\boldsymbol{\theta}}_N) + \frac{2p}{N}.$$

Si dimostra che sia questo criterio che il FPE tendono a sovrastimare la dimensione p (non sono stimatori consistenti). Uno stimatore consistente di p è il valore di p che minimizza il **criterio MDL (Minimum Description Length) di Rissanan**

$$MDL(p) := \log V_N(\hat{\boldsymbol{\theta}}_N) + \frac{\beta(p,N)}{N}$$

dove β è una funzione che asintoticamente cresce come $\log N$. Si prende normalmente $\beta(p,N) = p \log N$.