1

METODI STATISTICI PER L'IDENTIFICAZIONE DI MODELLI LINEARI

Giorgio Picci‡

Novembre 2006

[1]‡ Dipartimento di Ingegneria dell' Informazione, Universita' di Padova, 35131 Padova, Italy.

Indice

Prefazione **v**

2 INFERENZA STATISTICA E STIMA PARAMETRICA **1**
2.1 Concetti Generali 1
2.2 Teoria Generale della Stima Parametrica 8
2.2.1 Disuguaglianza di Cramèr-Rao 9
2.2.2 Identificabilità 14
2.3 Stima di Massima Verosimiglianza 15

3 STIMA PARAMETRICA SU MODELLI LINEARI **21**
3.1 Modelli Statistici Lineari 21
3.2 La distribuzione χ^2 27
3.3 Il principio dei Minimi Quadrati e il suo significato statistico 32
3.4 Minimi quadrati non lineari 42
3.5 Aspetti numerici dei problemi ai minimi quadrati 43
3.5.1 La Decomposizione ai Valori Singolari (SVD) 48

4 PROBLEMI DI REGRESSIONE MULTIPLA **57**
4.1 Stima della complessità di un modello lineare 57
4.2 Regressione lineare a stadi 59
4.3 Stima della dimensione del modello col criterio FPE 65
4.4 Un algoritmo di regressione lineare a stadi 66

5 MODELLI DINAMICI PER L'IDENTIFICAZIONE **73**
5.1 Introduzione 73
5.2 Modelli statistici lineari per processi del secondo ordine 73
5.3 Modelli parametrici e identificabilità in assenza di retroazione 76
5.4 Alcune classi di modelli e loro parametrizzazione 79
5.5 Identificabilità in presenza di reazione 80
5.5.1 Modelli a errori nelle variabili 80
5.6 Modelli multivariabili 80
5.7 Modelli Gaussiani 80

6 ERGODICITÀ **81**
6.1 Proprietà asintotiche degli stimatori: Consistenza 81

6.2 Processi Ergodici 83
6.3 Ergodicità del secondo ordine 100
6.4 Consistenza dello Stimatore di Massima Verosimiglianza 103

7 TEOREMA DEL LIMITE CENTRALE 109
7.1 Convergenza in legge 109
7.2 Il teorema del limite centrale per d-martingale stazionarie 112
7.3 TLC per variabili dipendenti 115
7.4 Efficienza asintotica 125

8 METODI PEM 127
8.1 Introduzione 127
8.2 Analisi asintotica dello stimatore PEM 129
8.3 La distribuzione asintotica dello stimatore PEM 135
8.4 La matrice d'informazione e il limite di Cramèr-Rao 139
8.5 Modello lineare e stima PEM 141

Bibliografia 143

PREFAZIONE

Prefazione

Capitolo 2

INFERENZA STATISTICA E STIMA PARAMETRICA

2.1 Concetti Generali

Com'è noto, la teoria moderna della Probabilità è una teoria *assiomatica*. Essa lavora su *modelli* probabilistici della realtà fisica, ma non dà alcuna indicazione su come questi modelli possano essere costruiti. Stando alla definizione, per costruire un modello probabilistico di una certa situazione fisica bisogna dare uno *spazio degli esperimenti* Ω (ad esempio l'insieme di tutti i possibili "esiti" di una misura), una "σ-algebra" $\mathcal{A}$ di *eventi* osservabili (i sottoinsiemi "probabilizzabili" di Ω) e una *misura di probabilità* P, definita su $\mathcal{A}$, per cui valgano i noti assiomi.

Mentre è quasi sempre facile (e in ogni caso abbastanza arbitrario) descrivere l'insieme dei possibili risultati di un esperimento relativo a una data situazione fisica per mezzo di un insieme Ω e la classe degli 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.
Si potrebbe dire che la statistica si occupa di assegnare probabilità sulla base dell'evidenza sperimentale. L'assegnazione di P a un dato spazio di esperimenti $\{\Omega, \mathcal{A}\}$ è un processo *induttivo* che richiede cioè una "interpretazione" o meglio una estrapolazione operata in base ai dati sperimentali osservati. Per sua natura, quindi, l'assegnazione di una probabilità *non è mai "sicura"*. Il criterio attraverso il quale si arriva a decidere che una certa P descrive "bene" i risultati di un esperimento non è oggettivabile a priori ma può variare da caso a caso. L'unico vincolo che si pone a priori è che esso sia logicamente consistente con gli assiomi del calcolo della probabilità.
Tipicamente, i dati di un problema di inferenza statistica sono:

- uno spazio di esperimenti $\{\Omega, \mathcal{A}\}$;

- una famiglia $\mathcal{P}$, o più famiglie *disgiunte* $\mathcal{P}_k$, $k = 1, \ldots, N$ (*N finito*), di possibili misure di probabilità P su $\{\Omega, \mathcal{A}\}$;

- il risultato di un esperimento, $\bar{\omega}$, $\bar{\omega} \in \Omega$ ($\bar{\omega}$ è l'osservazione, il dato sperimentale misurato).

I problemi sono essenzialmente di due tipi:

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

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

In entrambi i casi si tratta di costruire (in base a qualche criterio di "fedeltà) una funzione $\bar{\omega} \to \mathcal{P}$ oppure $\bar{\omega} \to \{1, 2, \cdots, k\}$. La distinzione tra stima e verifica di ipotesi è tra numero infinito (stima) o finito (verifica di ipotesi) di alternative possibili.

Un esempio elementare

Supponiamo di lanciare una moneta e sia p = probabilità di avere "testa" (T) e $1-p$ = probabilità di avere "croce" (C). Naturalmente p è incognito. Si vogliono ricavare informazioni su p lanciando la moneta n volte consecutive, supponendo che ogni lancio *"non influenzi" l'esito del successivo*.

Sia Ω = {tutti i possibili esiti di n lanci successivi}. L'insieme Ω contiene tutte le successioni di n simboli "T" e "C". Sia inoltre $\mathcal{A}$ la famiglia di tutti i sottoinsiemi di Ω. Per descrivere "l'indipendenza" dei lanci, si può ragionevolmente prendere la classe di probabilità $\mathcal{P} := \{P_p\}$ su $\{\Omega, \mathcal{A}\}$ definita $\forall \omega \in \Omega$ dalla

$$P_p(\omega) = p^{n(T)} \, (1-p)^{n-n(T)} \quad , \qquad 0 < p < 1 \quad , \tag{2.1}$$

dove $n(T)$ è il numero di simboli "T" nella successione ω. Si vede che P_p è definita univocamente non appena si assegna p $(0 < p < 1)$.

In questo caso la famiglia $\mathcal{P}$ è "parametrica", ovvero

$$\mathcal{P} := \Big\{ P_p \; ; \; 0 < p < 1 \Big\} \quad .$$

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. Viceversa si può pensare di usare l'osservazione $\bar{\omega}$ per validare una convinzione a priori che si ha su p, ad esempio il fatto che $p = 1/2$ (ovvero che T e C sono equiprobabili). In questo secondo caso $\bar{\omega}$ servirà per decidere se P_p appartiene alla classe

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

oppure se P_p sta in

$$\mathcal{P}_1 := \Big\{ P_p \; ; \; p \neq 1/2 \Big\} \quad .$$

Questo è un tipico problema di verifica di ipotesi. $\diamondsuit$

Problemi parametrici

La famiglia $\mathcal{P}$ di possibili misure di probabilità (oppure le N classi $\mathcal{P}_k$, $k = 1, \ldots, N$) costituisce, in un certo senso, l'*informazione a priori* nel problema di inferenza statistica. È ovvio che tanto più ristretta è la classe $\mathcal{P}$, tanto più precisi sono i risultati che ci possiamo aspettare nell'individuazione di P.

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

$$\mathcal{P} = \Big\{ P_\theta \; ; \; \theta \in \Theta \Big\} \quad , \tag{2.2}$$

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 N sottoinsiemi disgiunti (Θ_k, $k = 1, \ldots, N$) di Θ tali per cui $\mathcal{P}_k = \{P_\theta \mid \theta \in \Theta_k\}$, $k = 1, \ldots, N$. In corrispondenza, il problema di verifica di ipotesi diventa quello di decidere, in base ai dati osservati, a quale classe Θ_k appartiene θ.

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

In questo corso ci occuperemo esclusivamente di probabilità *indotte da variabili casuali* o da famiglie (eventualmente infinite) di variabili casuali.[1]

Sia $\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 & \cdots & \mathbf{y}_m \end{bmatrix}'$ una variabile aleatoria m-dimensionale definita su $\{\Omega, \mathcal{A}\}$ (cioè una funzione misurabile da Ω in $\mathbb{R}^m$). Se P è una probabilità definita su $\mathcal{A}$, ricordiamo che la *probabilità indotta da* $\mathbf{y}$, $P_\mathbf{y}$, è la misura di probabilità definita su $\{\mathbb{R}^m, \mathcal{B}^m\}$ ($\mathcal{B}^m :=$ σ-algebra di Borel di $\mathbb{R}^m$) ponendo

$$P_\mathbf{y}(E) := P\Big\{ \omega \mid \mathbf{y}(\omega) \in E \Big\} \quad , \tag{2.3}$$

oppure, più sinteticamente,

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

per ogni evento E di $\mathcal{B}^m$.

È ben noto che $P_\mathbf{y}$ è univocamente individuata dalla sua *funzione distribuzione di probabilità* (o di ripartizione) $F : \mathbb{R}^m \to [0, 1]$

$$F(y) = F(y_1, \ldots, y_m) = P\Big\{ \omega \mid \mathbf{y}_1(\omega) < y_1, \ldots, \mathbf{y}_m(\omega) < y_m \Big\} \tag{2.4}$$

e viceversa. Inoltre, se sul nuovo spazio di probabilità $\{\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P}\}$, in cui

$$\tilde{\Omega} = \mathbb{R}^m \quad , \qquad \tilde{\mathcal{A}} = \mathcal{B}^m \quad , \qquad \tilde{P} = P_\mathbf{y} \quad , \tag{2.5}$$

si definisce la variabile

$$\tilde{\mathbf{y}} : \mathbb{R}^m \to \mathbb{R}^m \quad , \qquad \tilde{\mathbf{y}}_i \left(y_1, \ldots, y_m \right) := y_i \quad , \qquad i = 1, \ldots, m \quad , \tag{2.6}$$

[1] "Variabile casuale" o "aleatoria" verrà abbreviato a "v.c." oppure "v.a." nel seguito. In genere le variabili aleatorie considerate in questo testo avranno valori vettoriali (in $\mathbb{R}^m$). Se $m = 1$ si parla di variabili *scalari* (o reali) mentre per $m > 1$ si usa talvolta la dizione *vettore aleatorio*.

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

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

Ne segue che tutte le probabilità di eventi relativi a $\mathbf{y}$ coincidono con le probabilità degli stessi eventi relativi a $\tilde{\mathbf{y}}$ e pertanto $\tilde{\mathbf{y}}$ e $\mathbf{y}$ possono essere riguardate come la *stessa* variabile casuale.

La rappresentazione (2.5), (2.6) 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 v.c. reale Gaussiana di media μ e varianza σ^2 si intende (o meglio si può sempre intendere) la funzione identità

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

definita sullo spazio di probabilità

$$\{\Omega, \mathcal{A}, P\} = \{\mathbb{R}, \mathcal{B}, P_{\mathbf{y}}\} \quad ,$$

dove $P_{\mathbf{y}}$ è definita dalla relazione

$$P_{\mathbf{y}}(E) = \frac{1}{2\pi\sigma^2} \int_E e^{-\frac{(y-\mu)^2}{2\sigma^2}} \, dy$$

per ogni $E \in \mathcal{B}$. $\{\tilde{\Omega}, \tilde{\mathcal{A}}\}$ si chiama *spazio dei valori campionari* di $\mathbf{y}$.

Si noti che lo spazio $\{\tilde{\Omega}, \tilde{\mathcal{A}}, P_{\mathbf{y}}\}$ 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 $\{\tilde{\Omega}, \tilde{\mathcal{A}}, P_{\mathbf{y}}\}$ non può, ad esempio, essere definita una v.c. (non costante) indipendente da $\mathbf{y}$.

Di norma, nel seguito useremo sempre rappresentazioni sullo spazio dei valori campionari.

Ciò premesso, faremo d'ora in avanti riferimento a problemi di inferenza statistica in cui $\mathcal{P}$ (o $\{\mathcal{P}_k\}$) è una famiglia di misure di probabilità su $\{\mathbb{R}^m, \mathcal{B}^m\}$. Pertanto ogni elemento $P \in \mathcal{P}$ ($\mathcal{P}_k$) individua ed è individuato univocamente da una *funzione di ripartizione* su $\mathbb{R}^m$ ed è quindi equivalente descrivere $\mathcal{P}$ come una famiglia di funzioni di ripartizione, $\mathcal{P} := \{F(\cdot)\}$.

Supporremo inoltre che la "forma" delle F sia nota a priori e che per individuare una F in $\mathcal{P}$ basti assegnare il valore di un parametro p-dimensionale θ. Se Θ è il campo dei valori ammissibili dal parametro, avremo

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

Come si vede, in questo schema i dati sperimentali $(\bar{\omega})$ sono pensati come determinazione 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. Per esempio nel caso di misure affette da errori accidentali, effetto di molte possibili interazioni dell'apparato di misura con disturbi additivi tra loro indipendenti, si prende spesso la famiglia delle distribuzioni Gaussiane m-dimensionali (o come si preferisce talvolta chiamarle "m-variate"), che è caratterizzata, come è noto, dalla funzione densità di probabilità

$$f(y) = (2\pi)^{-m/2} \, |\det \Sigma|^{-1/2} \, \exp -\frac{1}{2} \left\{ (y-\mu)' \, \Sigma^{-1} (y-\mu) \right\} \quad .$$

In questo caso il vettore $\mu \in \mathbb{R}^m$ e la matrice di covarianza $\Sigma \in \mathbb{R}^{m \times m}$ sono dei "parametri" incogniti da determinarsi in base a certi dati disponibili,

$$\{ y_1 \, , \, \ldots \, , \, y_n \} \quad ,$$

dove il vettore $y_t \in \mathbb{R}^m$ è il risultato della t-sima misura di $\mathbf{y}$.

Una questione che si pone allo sperimentatore è come eseguire le misure in modo tale che esse diano la "massima informazione sulla distribuzione (incognita) di $\mathbf{y}$. È chiaro che 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. Teniamo presente che il modello probabilistico che vogliamo costruire (cioè F) deve descrivere proprio (gli effetti di) queste cause d'errore.

Definizione 2.1. *Sia $\mathcal{P} = \{F\}$ 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 $\mathcal{P}$. Si dice allora che $\mathbf{y}_1, \ldots, \mathbf{y}_n$ sono un "campione casuale" di numerosità n relativo alla classe $\mathcal{P}$ (o "estratto" da $\mathcal{P}$).*

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 corsi, 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:

$$F_\theta^n(y_1, \ldots, y_n) = F_\theta(y_1), \ldots, F_\theta(y_n) \quad , \qquad y_t \in \mathbb{R}^m \quad . \tag{2.7}$$

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* (si veda ad esempio [5]).

Notiamo comunque fin da adesso che in molte situazioni concrete 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.

Definizione 2.2. *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.

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

$$\bar{\mathbf{y}}_n = \frac{1}{n} \sum_{t=1}^{n} \mathbf{y}_t \quad ; \tag{2.8}$$

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 $E_0 \bar{\mathbf{y}}_n = E_0 \mathbf{y}$, dove E_0 denota l'operatore di media rispetto alla distribuzione F_{θ_0}. Anticipiamo il fatto notevole, che scende dalla legge dei grandi numeri [2], che il limite

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

esiste *con probabilità 1* e vale $E_0 \mathbf{y} = \int_{\mathbb{R}^m} y \, dF_{\theta_0}(y)$. 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* $E_0 \mathbf{y}$ della distribuzione F_{θ_0}.

La *varianza campionaria*, S^2,

$$S_n^2 := \frac{1}{n} \sum_{t=1}^{n} (\mathbf{y}_t - \bar{\mathbf{y}}_n) \, (\mathbf{y}_t - \bar{\mathbf{y}}_n)' \tag{2.9}$$

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

[2] Questa verrà richiamata nel capitolo 6.

La varianza campionaria di un campione casuale S_n^2 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

$$E_0(\mathbf{y} - E_0\mathbf{y})\ (\mathbf{y} - E_0\mathbf{y})' \quad ,$$

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 13.1

Sia $\mathbf{y} \sim N(\mu, \Sigma\}$ (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 $N(\theta, \Sigma)$: allora

$$V^2 := \frac{1}{n} \sum_{1}^{n} (\mathbf{y}_t - E\mathbf{y})\ (\mathbf{y}_t - E\mathbf{y})'$$

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

Per concludere questo paragrafo introduttivo cercheremo di discutere (assai semplicisticamente) due possibili approcci al problema dell'inferenza statistica.

Il *primo* approccio (che chiameremo "Fisheriano" da R.A. Fisher, uno dei padri fondatori della statistica) postula l'esistenza di un *valore vero*, θ_0, del parametro, in corrispondenza al quale si ha descrizione "esatta" dei dati di misura da parte della d.d.p. F_{θ_0}; θ_0 è ovviamente incognito. Il fatto che il parametro vero sia una grandezza deterministica costante, in linea di principio suscettibile di essere determinata esattamente (ad esempio mediante una serie infinita di esperimenti indipendenti), può essere accettabile se a θ viene dato un significato puramente matematico (ad esempio la varianza di una distribuzione Gaussiana), ma diventa un poco questionabile se il parametro si interpreta invece come il valore di una grandezza fisica che si sta misurando.

Da questa critica prende spunto la *teoria Bayesiana* secondo la quale θ è da riguardarsi *sempre* come una variabile casuale (che denoteremo di norma 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

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

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.

Nel seguito ci si riferirà all'impostazione Fisheriana.

2.2 Teoria Generale della Stima Parametrica

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 2.3. *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 $\mathbf{y}_1, \ldots, \mathbf{y}_n$,*

$$\hat{\theta} = \phi(y_1, \ldots, y_n) \quad , \tag{2.10}$$

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

$$E_{\theta_0} \ \phi(\mathbf{y}_1, \ldots, \mathbf{y}_n) = \theta_0 \quad , \tag{2.11}$$

dove E_{θ_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 (2.11) valga per tutti i possibili valori del parametro. Si arriva così alla definizione seguente,

Definizione 2.4. *Uno stimatore ϕ si dice (uniformemente)* **corretto** *se*

$$E_\theta \ \phi(\mathbf{y}_1, \ldots, \mathbf{y}_n) = \theta \quad , \qquad \forall \theta \in \Theta \quad . \tag{2.12}$$

Un buon stimatore dovrebbe inoltre fornire valori molto concentrati attorno alla media, avere cioè una bassa dispersione. Naturalmente perchè l'dea 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 2.5. *Uno stimatore ϕ si dice (uniformemente) a* **minima varianza** *nella classe $\mathcal{C}$ se la varianza di ϕ*

$$\mathrm{var}_\theta(\phi) := E_\theta(\phi - E_\theta\,\phi)'\,(\phi - E_\theta\,\phi) \tag{2.13}$$

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

$$\mathrm{var}_\theta(\phi) \le \mathrm{var}_\theta(\psi) \quad , \qquad \forall \psi \in \mathcal{C} \quad , \tag{2.14}$$

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

Come abbiamo già visto, per evitare banalità occorre restringere la classe $\mathcal{C}$ a una opportuna sottoclasse di tutte le funzioni misurabili dei dati. Vedremo tra poco che se si prende per $\mathcal{C}$ 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.

2.2.1 Disuguaglianza di Cramèr-Rao

Supponiamo che $\mathbf{x}$ sia un vettore aleatorio r-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 $\mathbf{x}$). Supponiamo che valgano le seguenti ipotesi.

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

A.2) Per ogni statistica ϕ con $E_\theta \phi < \infty$,

$$\frac{\partial}{\partial \theta_i} \int_{\mathbb{R}^r} \phi(x)\; p(x,\theta)\, dx = \int_{\mathbb{R}^r} \phi(x)\; \frac{\partial}{\partial \theta_i}\; p(x,\theta)\, dx$$

per $i = 1, \ldots, p$ e per ogni $\theta \in \Theta$. In particolare

$$\frac{\partial}{\partial \theta_i} \int_{\mathbb{R}^r} p(x,\theta)\, dx = \int_{\mathbb{R}^r} \frac{\partial}{\partial \theta_i}\; p(x,\theta)\, dx.$$

A.3) $$\frac{\partial^2}{\partial \theta_i\, \partial \theta_j} \int_{\mathbb{R}^r} p(x,\theta)\, dx = \int_{\mathbb{R}^r} \frac{\partial^2}{\partial \theta_i\, \partial \theta_j}\; p(x,\theta)\, dx$$

per ogni $i, j = 1, \ldots, p$ e per ogni $\theta \in \Theta$.

Definizione 2.6. *La* matrice di informazione di Fisher $I(\theta)$*, associata alla famiglia parametrica di densità $\{p_\theta\}$ è definita dalla posizione*

$$I(\theta) := \left[E_\theta \left(\frac{\partial \log p(\mathbf{x},\theta)}{\partial \theta_i} \cdot \frac{\partial \log p(\mathbf{x},\theta)}{\partial \theta_j} \right) \right]_{i,j=1,\ldots,p} \tag{2.15}$$

che si può anche scrivere come

$$I(\theta) = \left[-E_\theta \, \frac{\partial^2 \, \log p(\mathbf{x}, \theta)}{\partial \theta_i \, \partial \theta_j} \right]_{i,j=1,\ldots,p} \quad . \tag{2.16}$$

ed è una matrice almeno semidefinita positiva.

Per capire il significato di $I(\theta)$ definiamo il vettore casuale p-dimensionale delle "sensitività" di $\{p_\theta\}$ rispetto al parametro

$$\mathbf{z}_\theta = \frac{\partial \, \log p(\mathbf{x}, \theta)}{\partial \theta_i} \Bigg|_{i,j=1,\ldots,p} \tag{2.17}$$

e notiamo che

$$I(\theta) = E_\theta \, \mathbf{z}_\theta \, \mathbf{z}'_\theta \geq 0 \quad . \tag{2.18}$$

L'equivalenza di (2.16) e (2.15) si ricava notando che $\int p(x, \theta) \, dx \, = \, 1$ (costante rispetto a θ) e che derivando questa uguaglianza membro a membro si trova

$$\int_{\mathbb{R}^r} \frac{\partial p(x, \theta)}{\partial \theta_i} \, dx = 0 \quad , \qquad i = 1, \ldots, p \quad , \tag{2.19}$$

$$\int_{\mathbb{R}^r} \frac{\partial^2 p(x, \theta)}{\partial \theta_i \, \partial \theta_j} \, dx = 0 \quad , \qquad i, j = 1, \ldots, p \quad . \tag{2.20}$$

Dalla (2.19) segue immediatamente che $E_\theta \, \frac{\partial \, \log p}{\partial \theta_i} = 0$ per tutti gli i e quindi

$$E_\theta \, \mathbf{z}_\theta = 0 \tag{2.21}$$

e pertanto $I(\theta)$ è la varianza di $\mathbf{z}_\theta$.

La (2.15) 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 (2.20).

Teorema 2.1 (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*

$$G(\theta) = \left[\frac{\partial g_i(\theta)}{\partial \theta_j} \right]_{\substack{i \, = \, 1, \ldots, q \\ j \, = \, 1, \ldots, p}} \quad . \tag{2.22}$$

Nell'ipotesi che la matrice di Fisher $I(\theta)$ sia invertibile si ha allora:

$$V(\theta) - G(\theta) \; I^{-1}(\theta) \; G'(\theta) \geq 0 \quad , \tag{2.23}$$

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

Prova. Dato che

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

applicando la A.3) si ottiene

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

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

$$E_\theta\ \phi\ \mathbf{z}'_\theta = E_\theta\ \phi\ [\mathbf{z}_\theta^1, \ldots, \mathbf{z}_\theta^p] \quad ,$$

ovvero

$$E_\theta\ \phi\ \mathbf{z}'_\theta = G(\theta) \quad . \tag{2.24}$$

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. $\diamondsuit$

Osservazioni

1. 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 (2.23) diventa

$$V(\theta) - I(\theta)^{-1} \geq 0 \quad . \tag{2.25}$$

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

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

(perché 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 superare il numero positivo $\mathrm{Tr}\ \ I(\theta)^{-1}$,

$$\mathrm{var}_\theta(\phi) \geq \mathrm{Tr}\ \ \Big[I(\theta)^{-1}\Big] \quad , \qquad \forall \theta \quad . \tag{2.26}$$

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

2. Si noti che non è detto che il limite inferiore di Cramèr-Rao sia il 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.

Esempio 13.2

Sia $\mathbf{y} \sim N(\theta, \sigma^2)$ una v.a. scalare con σ^2 nota.
Si ha

$$\log\, p(y,\theta) = C - \frac{1}{2}\,\frac{(y-\theta)^2}{\sigma^2} \quad ,$$

$$\frac{d}{d\theta}\,\log\, p(y,\theta) = \frac{y-\theta}{\sigma^2}$$

e quindi

$$i(\theta) = E_\theta\left(\frac{\mathbf{y}-\theta}{\sigma^2}\right)^2 = \frac{1}{\sigma^4}\cdot\sigma^2 = 1/\sigma^2 \quad .$$

Ne segue che la varianza di ogni stimatore corretto di θ non può essere inferiore alla varianza σ^2 di $\mathbf{y}$.

Supponiamo allora di avere un campione casuale di numerosità n estratto dalla precedente distribuzione Gaussiana. In questo caso si ha a che fare con un vettore aleatorio $\mathbf{x} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ $(r = n)$ e

$$p(y_1, \ldots, y_n\ ,\ \theta) = \prod_{t=1}^{n}\, p(y_t, \theta) \quad .$$

Quindi,

$$\log\, p(y_1, \ldots, y_n\ ,\ \theta) = nC - \frac{1}{2}\,\sum_{t=1}^{n}\frac{(y_t-\theta)^2}{\sigma^2} \quad ,$$

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

e, per l'indipendenza,

$$I(\theta) = E_\theta\left[\frac{d\,\log p}{d\theta}\right]^2 = \frac{1}{\sigma^4}\cdot n\,\sigma^2 = \frac{n}{\sigma^2} \quad .$$

Consideriamo la media campionaria

$$\bar{\mathbf{y}} = \frac{1}{n}\,\sum_{t=1}^{n}\,\mathbf{y}_t$$

che è distribuita come $N(\theta, \sigma^2/n)$. Ovviamente $\bar{\mathbf{y}}$ 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 $\bar{\mathbf{y}}$ 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*). $\diamondsuit$

Esempio 13.3

Sia $\mathbf{y} \sim N(\mu, \theta^2)$, μ nota, e $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ un campione casuale estratto da $N(\mu, \theta^2)$. Consideriamo lo stimatore

$$\bar{S}^2 = \frac{nS^2}{n-1} = \frac{1}{n-1} \sum_{t=1}^{n} (\mathbf{y}_t - \bar{\mathbf{y}})^2 \quad ;$$

vedremo più avanti che $\bar{S}^2$ è uno stimatore corretto di θ^2, di varianza $\frac{2\theta^4}{n-1}$. Il limite di Cramèr-Rao in questo caso vale $2\theta^4/n$ e quindi la varianza di $\bar{S}^2$ è strettamente più grande di $I(\theta)^{-1}$. Si può però dimostrare che uno stimatore corretto di θ^2 non può avere varianza più piccola di $\bar{S}^2$. Da questo esempio segue che $I(\theta)^{-1}$ non è la maggiorazione migliore possibile.

Esercizi

1-1 Mostrare che, se $I(\theta)$ è la matrice di Fisher relativa a $p(\mathbf{y}, \theta)$, quella relativa a un campione casuale di rumorosità n è di $n\, I(\theta)$.
1-2 Mostrare che il limite di Cramèr-Rao per un campione casuale con distribuzione $N(\mu, \theta^2)$ è effettivamente $2\theta^4/n$.

1-3 Mostrare che il limite di Cramèr-Rao per per un campione casuale con distribuzione $N(\theta_1, \theta_2^2)$ è

$$I(\theta)^{-1} = \begin{bmatrix} \theta_2^2/n & 0 \\ 0 & 2\theta_2^4/n \end{bmatrix} \quad .$$

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 *distanza di Kullback* tra due densità f e p il numero

$$I(f, p) := \int_{\mathbb{R}^r} [f \, \log f - f \, \log p] \, dx \; = \int_{\mathbb{R}^r} f \, \log f/p \, dx = E_f \, \log f/p \, ; \qquad (2.27)$$

si dimostra che $I(f, p) \geq 0$ e che $I(f, p) = 0$ solo nel caso in cui $f = p$. $I(f, p)$ può essere preso come misura della deviazione fra le due distribuzioni f e p. $I(f, p)$ non è però una vera metrica perché $I(p, f) \neq I(f, p)$ e non soddisfa la disuguaglianza triangolare.

Supponiamo ora che $f \equiv p(\cdot, \theta_0)$ e $p \equiv p(\cdot, \theta)$, $\theta_0, \theta \in \Theta$. Chiameremo $I(\theta_0, \theta)$ la quantità $I(p(\cdot, \theta_0), p(\cdot, \theta))$. Ponendo $\theta = \theta_0 + \Delta\theta$ si ha

$$I(\theta_0, \theta) = I(\theta_0, \theta_0) + \left.\frac{\partial I}{\partial \theta}\right|_{\theta_0} \Delta\theta + \frac{1}{2} \Delta\theta' \left[\frac{\partial^2 I}{\partial \theta_i \, \partial \theta_j}\right]_{\theta_0} \cdot \Delta\theta + 0(3) \quad .$$

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

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

di modo che,

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

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

Nello stesso modo si verifica poi che

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

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

$$I(\theta_0, \theta) \cong \frac{1}{2}\, \Delta\theta'\; I(\theta_0)\; \Delta\theta \quad ; \tag{2.28}$$

questa relazione dice che, 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.

2.2.2 Identificabilità

Esistono dei casi in cui il campione è "strutturalmente" incapace di dare informazioni utili per la stima di θ. Un esempio (molto banale) potrebbe essere il seguente. Supponiamo che θ sia un parametro bidimensionale (θ_1, θ_2), che $\Theta = \mathbb{R}^2$ e che F_θ dipenda da (θ_1, θ_2) solo attraverso il loro prodotto $\theta_1\theta_2$. Ad esempio, sia $F_\theta \sim N(\theta_1\theta_2, \sigma^2)$ e $\bar\theta = (\bar\theta_1, \bar\theta_2)'$ 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 $F_{\bar\theta}(x) = F_{\hat\theta}(x)$, $\forall x$, e 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 2.7. *Due punti θ_1 e θ_2 in Θ si dicono "indistinguibili" se $F_{\theta_1}(x) = F_{\theta_2}(x)$, $\forall x \in \mathbb{R}^r$.*

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 2.8. *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 2.9. *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 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 modelli di sistemi biologici e fisiologici, e nell'identificazione di sistemi dinamici 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 2.2 (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.*

Dimostrazione. La dimostrazione si può ricondurre alla proprietà della (pseudo)-metrica di Kullback che garantisce $I(\theta_0, \theta) = 0 \Leftrightarrow p(\cdot, \theta_0) = p(\cdot, \theta)$. 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

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

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

2.3 Stima di Massima Verosimiglianza

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

Definizione 2.10. *La* funzione di verosimiglianza *dell'osservazione* x_0 *è la funzione* $L(x_0, \cdot) : \Theta \to R_+$ *(i reali non negativi) definita ponendo*

$$L(x_0, \theta) := p(x_0, \theta) \quad . \tag{2.29}$$

Il "principio della massima verosimiglianza", introdotto da Gauss nel 1856 [11] 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.

È chiaro che, 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 $\forall x_0 \in \mathbb{R}^r$)

$$L\Big(\mathbf{x}, \hat{\theta}(\mathbf{x})\Big) = \max_{\theta \in \Theta} L(\mathbf{x}, \theta) \quad . \tag{2.30}$$

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)

$$\ell(\mathbf{x}, \cdot) = \log L(\mathbf{x}, \cdot) \tag{2.31}$$

si chiama 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*)

$$\frac{\partial \ell}{\partial \theta_k}(\mathbf{x}, \theta) = 0 \quad , \qquad k = 1, \ldots, p \quad , \tag{2.32}$$

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 13.5

Sia $\mathbf{y} \sim N(\theta_1, \theta_2^2)$, 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 (2.32) diventano

$$\frac{\partial \ell}{\partial \theta_1} = \frac{1}{\theta_2^2}\left(\sum_1^n y_i - n\theta_1\right) = 0 \quad ,$$

$$\frac{\partial \ell}{\partial \theta_2^2} = -\frac{n}{2\theta_2^2} + \frac{1}{2\theta_2^4}\sum_1^n (y_i - \theta_1)^2 = 0 \quad .$$

La prima è un'equazione nella sola θ_1 che dà

$$\hat{\theta}_1 = \frac{1}{n}\sum_1^n y_i = \bar{y} \quad .$$

Nella seconda equazione, si noti che

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

Notando che $\sum_1^n (y_i - \bar{y})$ (somma degli scarti dalla media campionaria) è zero e sostituendo inoltre il valore di θ_1 che risolve la prima equazione, si vede che la seconda si riduce a

$$-\frac{n}{2\theta_2^2} + \frac{1}{2\theta_2^4}\sum_1^n (y_i - \bar{y})^2 = 0 \quad ,$$

che porge finalmente

$$\hat{\theta}_2^2 = \frac{1}{n}\sum_1^n (y_i - \bar{y})^2 = s^2 \quad . \tag{2.33}$$

È facile verificare che $(\bar{y}, s^2)$ danno effettivamente un massimo assoluto di $\ell(x, \cdot)$. Si ha così

Proposizione 2.1. *Gli stimatori di M.V. basati su un campione casuale ($\mathbf{y}_1, \ldots, \mathbf{y}_n$) estratto da $N(\theta_1, \theta_2^2)$ sono la media, $\bar{\mathbf{y}}$, e la varianza campionarie, S^2.*

Il risultato continua a valere pari pari nel caso multivariabile. Se $\mathbf{y} \sim N(\theta, \Sigma)$, 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 $\phi = [\bar{\mathbf{y}}, S^2]$ dove $\bar{\mathbf{y}}$ *ed* S^2 *sono la media e la varianza campionarie*.

Proprietà degli stimatori di Massima Verosimiglianza

Gli stimatori di M.V. in generale *non sono corretti*. Per convincersene basta prendere l'esempio (scalare) appena considerato in cui S^2 è lo stimatore di M.V. di θ_2^2. Dalla (2.33) segue che

$$\frac{nS^2}{\theta_2^2} = \sum_1^n \frac{(\mathbf{y}_i - \theta_1)^2}{\theta_2^2} - n\,\frac{(\bar{\mathbf{y}} - \theta_1)^2}{\theta_2^2} \quad . \tag{2.34}$$

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

$$E_\theta\,\frac{nS^2}{\theta_2^2} = n - 1 \quad ,$$

per cui

$$E_\theta\, S^2 = \theta_2^2\,\frac{n-1}{n} \quad . \tag{2.35}$$

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 2.3 (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

$$\tilde{\ell}(x,\gamma) = \ell\Big(x,\, g^{-1}(\gamma)\Big) = \ell(x,\theta)\Big|_{\theta = g^{-1}(\gamma)} \tag{2.36}$$

(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*.

Esempio 13.6

In molte applicazioni si incontrano grandezze che sono per definizione non negative (concentrazioni, densità, quantità prodotte ecc.) e come tali non possono essere descritte per mezzo di d.d.p. Gaussiane. In questi casi si usa spesso modellare le misure con una legge che si chiama *log normale*. (Per alcuni esempi si veda il testo di Cramèr [6, pag. 219–220]). La variabile casuale (scalare) $\mathbf{y}$ è distribuita in modo log-normale se $\mathbf{y} \geq 0$ (c.p.1), e se $\log \mathbf{y} \sim N(\mu, \sigma^2)$ o, più in generale, se per qualche $a \geq 0$, $\log(\mathbf{y} - a) \sim N(\mu, \sigma^2)$. In quest'ultimo caso ovviamente dovrà essere $\mathbf{y} \geq a$ (c.p.1) e si può facilmente controllare che la densità di $\mathbf{y}$ è data dall'espressione

$$\frac{1}{\sigma(y-a)\,2\pi}\ \exp -\frac{1}{2\sigma^2}\ \Big[\log(y-a)-\mu\Big]^2 \quad . \tag{2.37}$$

Supponiamo di osservare una variabile $\mathbf{y}$ distribuita in modo log-normale (con $a = 0$) e di voler trovare le stime di M.V. della sua media e della sua varianza. Ciò che viene immediatamente in mente di fare è di prendere i logaritmi del campione (casuale) $\mathbf{x}_1 = \log \mathbf{y}_1, \ldots, \mathbf{x}_n = \log \mathbf{y}_n$ e da questi ricavare le stime dei parametri θ_1 e θ_2^2 nella distribuzione (Gaussiana) di $\log \mathbf{y}$. Una volta fatto questo le stime della media ξ e della varianza λ^2 dalla distribuzione log-normale si ricaveranno usando le formule (ricordare l'espressione della $E\{e^{t\mathbf{x}}\}$ con $\mathbf{x} \sim N(\theta_1, \theta_2^2)$!)

$$\begin{aligned} \xi &= \exp\left(\theta_1 + \frac{\theta_2^2}{2}\right) \quad , \\ \lambda^2 &= \xi^2\left(e^{\theta_2^2} - 1\right) \quad , \end{aligned} \tag{2.38}$$

che danno appunto media e varianza di $\mathbf{y}$ in funzione della media (θ_1) e varianza (θ_2^2) di $\log \mathbf{y}$. Questo approccio è in effetti validato dal teorema precedente, *purché beninteso si tratti di stime di M.V.*. In conclusione, detto

$$\hat{\theta}_1(x_1, \ldots, x_n) = \bar{x} = \frac{1}{n}\ \sum_{i=1}^{n}\ x_i = \frac{1}{n}\ \sum_{i=1}^{n}\ \log\ y_i$$

lo stimatore di M.V. di θ_1 e

$$\hat{\theta}_2^2(x_1, \ldots, x_n) = S^2 = \frac{1}{n}\ \sum_{i=1}^{n}\ (x_i - \bar{x})^2 = \frac{1}{n}\ \sum_{i=1}^{n}\ (\log y_i - \bar{x})^2$$

lo stimatore di M.V. di θ_2^2, gli stimatori di M.V. di ξ e λ^2, rispettivamente $\hat{\xi}$ e $\hat{\lambda}^2$, sono dati dalle formule

$$\begin{aligned} \hat{\xi}(\mathbf{y}_1, \ldots, \mathbf{y}_n) &= \exp\left(\theta_1 + \frac{\theta_2^2}{2}\right) \quad , \\ \lambda^2(\mathbf{y}_1, \ldots, \mathbf{y}_n) &= \xi^2\left|\exp(\theta_2^2) - 1\right| \quad . \end{aligned} \tag{2.39}$$

□

Un altro caso in cui il principio di invarianza torna assai utile è nella stima della d.d.p. di *stimatori*. In effetti uno stimatore di θ, $\hat{\theta}(\mathbf{y}_1, \ldots, \mathbf{y}_n)$, produce dei numeri in corrispondenza ai dati osservati $(y_1, \ldots, y_n)$ che sono "vicini" in senso probabilistico al parametro vero θ_0. In pratica questa vicinanza si valuta concretamente ad esempio mediante la *varianza* della v.c. $\hat{\theta}$. Sfortunatamente questa varianza è in generale *funzione essa stessa del parametro incognito* θ. È quindi necessario darne *stime* calcolabili in base ai dati. Se queste stime sono di M.V. si può usare il principio di invarianza.

Sia ad esempio $\mathbf{y} \sim N(\mu, \theta_2^2)$ ed $S^2(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ lo stimatore di M.V. del parametro θ_2^2. La varianza di S^2 è

$$\mathrm{var}_\theta\, S^2 = \frac{2\theta_2^2}{n-1} \quad , \tag{2.40}$$

che dipende ancora da θ_2^2. Ne viene che la *stima* di M.V. del parametro $\mathrm{var}_\theta\, S^2$ è

$$\widehat{\mathrm{var}\, S^2} = \frac{2(S^2)^2}{n-1} \quad . \tag{2.41}$$

Se si eccettua il caso di distribuzioni di tipo Gaussiano (che verrà discusso in dettaglio più avanti), le proprietà degli stimatori di M.V. per "piccoli campioni sono poco note. Gli unici risultati generali riguardano il comportamento asintotico (per grandi campioni). Si dimostra che lo stimatore di massima verosimiglianza, è, in ipotesi abbastanza generali, asintoticamente corretto cioè *consistente*, asintiticamente distribuito in modo *Gaussiano* ed *efficiente*, cioè asintoticamente a minima varianza.

È proprio in virtù di queste proprietà che il criterio della massima verosimiglianza è considerato in statistitica il metodo d'elezione per costruire stimatori. Purtroppo in pratica l'operazione di calcolo esplicito dello stimatore si riesce a fare solo in pochissimi casi e bisogna accontentarsi di procedure numeriche per calcolare la stima.

Il Metodo dei Momenti

[DA SCRIVERE]

Capitolo 3
STIMA PARAMETRICA SU MODELLI LINEARI

3.1 Modelli Statistici Lineari

Sia $\mathbf{y}$ un vettore aleatorio n-dimensionale di d.d.p. incognita appartenente alla famiglia parametrica $\{F_\theta \ ; \ \theta \in \Theta\}$.

Noi chiameremo *modello statistico* (o, equivalentemente, *modello probabilstico*) di $\mathbf{y}$ una rappresentazione del tipo

$$\mathbf{y} = f(\theta, \mathbf{w}) \quad , \tag{3.1}$$

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 una relazione del tipo (3.1) che non mediante una famiglia parametrica $\{F_\theta\}$. Una classe tipica di esempi è la seguente.

Il modello della Teoria degli Errori 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

$$\mathbf{y} = s(\theta) + \mathbf{w} \quad , \tag{3.2}$$

dove abbiamo introdotto i due vettori colonna

$$\begin{aligned} s(\theta) &= [s_1(\theta), \ldots, s_n(\theta)]' \quad , \\ \mathbf{w} &= [\mathbf{w}_1, \ldots, \mathbf{w}_n]' \quad . \end{aligned} \tag{3.3}$$

Questo modello del tipo "misura" = "segnale" più "rumore" (Gaussiano) additivo 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 (3.2) la matrice di covarianza del rumore

$$R := E\mathbf{w}\mathbf{w}'$$

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

$$\mathbf{y} = s(\theta) + \sigma\mathbf{w} \quad , \tag{3.4}$$

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

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. Il modello (3.2) può allora essere riscritto come

$$\mathbf{y} = s(\theta) + R^{1/2}\,\mathbf{w} \quad , \tag{3.5}$$

dove ora $R^{1/2} \in R^{n\times n}$ è una radice quadrata della varianza incognita[3] che si assume simmetrica e definita positiva e $\mathbf{w} \sim N(0, I)$. I problemi di stima associati al modello (3.5) sono però molto complicati. Nel seguito considereremo un caso intermedio fra (3.4) e (3.5). Supporremo cioè $\mathbf{w}$ Gaussiano e di covarianza "parzialmente nota", della forma $\sigma^2 R$ con σ^2 *incognita* ed R *nota* e definita positiva.

[3] Questo significa che $R^{1/2}(R^{1/2})' = R$.

Faremo inoltre l'ipotesi che $s(\theta)$ sia "approssimabile" a una funzione lineare di θ, cioè

$$s(\theta) = S\theta \quad , \qquad S \in \mathbb{R}^{n\times p} \quad , \tag{3.6}$$

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

$$\mathbf{y} = S\theta + \sigma\mathbf{w} \quad . \tag{3.7}$$

Stima di M.V. nel modello lineare

Per chiarezza di esposizione conviene a questo punto formulare esplicitamente il problema che ci interessa risolvere.

Problema 3.1. *Trovare le stime di M.V. dei parametri $\theta \in \mathbb{R}^p$ e $\sigma^2 \in \mathbb{R}_+$ nel modello lineare* (3.7)*, 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 $\mathbf{y} \sim N(S\theta, \sigma^2 R)$ e pertanto la funzione di log-verosimiglianza si scrive

$$\begin{aligned}\ell(y,\theta,\sigma^2) &= -\frac{n}{2}\,\log 2\pi - \frac{1}{2}\,\log\left[\det(\sigma^2 R)\right] - \frac{1}{2}\,(y-S\theta)'\,(\sigma^2 R)^{-1}\,(y-S\theta)\\ &= -\frac{n}{2}\,\log 2\pi - \frac{n}{2}\,\log\sigma^2 - \frac{1}{2}\,\log\det R - \frac{1}{2\sigma^2}\,(y-S\theta)'\,R^{-1}\,(y-S\theta)\,,\end{aligned} \tag{3.8}$$

cosicché

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

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

$$\frac{\partial \ell}{\partial \sigma^2} = \frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\,(y-S\theta)'\,R^{-1}\,(y-S\theta) \quad .$$

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

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

e ricordiamo che

$$I(\theta,\sigma) = E_{\theta,\sigma}\begin{bmatrix}\mathbf{z}_\theta\mathbf{z}_\theta' & \mathbf{z}_\theta\mathbf{z}_\sigma\\ \mathbf{z}_\theta'\mathbf{z}_\sigma & \mathbf{z}_\sigma^2\end{bmatrix} \quad . \tag{3.9}$$

Svolgendo i calcoli, si trova

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

Ponendo inoltre

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

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

$$E_{\theta,\sigma}\ \mathbf{z}_\theta \mathbf{z}_\sigma = E_{\theta,\sigma}\left\{\frac{1}{\sigma^2}\ S' R^{-1/2}\ \tilde{\mathbf{y}}\left(-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\ \tilde{\mathbf{y}}'\tilde{\mathbf{y}}\right)\right\}$$

$$= \frac{1}{2\sigma^6}\ S' R^{-1/2}\ E_{\theta,\sigma}\ \tilde{\mathbf{y}}\tilde{\mathbf{y}}'\tilde{\mathbf{y}} = 0 \quad ,$$

dato che i momenti centrali del terz'ordine di una d.d.p. Gaussiana sono nulli. Infine, ponendo

$$Q(y) := (y - S\theta)'\, R^{-1}\,(y - S\theta) \quad ,$$

si vede che

$$E_{\theta,\sigma}\ \mathbf{z}_\sigma^2 = E_{\theta,\sigma}\left\{\frac{1}{2\sigma^2}\ \left[\frac{Q(\mathbf{y})}{2} - n\right]\right\}^2 \quad . \tag{3.10}$$

Come vedremo più avanti, se $\mathbf{y} \sim N(\mu, \Sigma)$, la forma quadratica $(\mathbf{y} - \mu)'\,\Sigma^{-1}\,(\mathbf{y} - \mu)$ ha una distribuzione del tipo χ^2 con un numero di gradi di libertà pari alla dimensione di $\mathbf{y}$. Allora, $\frac{Q(\mathbf{y})}{\sigma^2} \sim \chi^2(n)$ e dalla (3.10) segue

$$E_{\theta,\sigma}\ \mathbf{z}_\sigma^2 = \frac{1}{4\sigma^4}\ \text{Var}\left[\frac{Q(\mathbf{y})}{2}\right] = \frac{n}{2\sigma^4} \quad .$$

Mettendo insieme questi risultati si trova infine

$$I(\theta, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2}\, S' R^{-1} S & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix} \quad . \tag{3.11}$$

Proposizione 3.1. *Nel modello* (3.7) *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 $S' R^{-1} S$ è invertibile e questo avviene allora e solo allora che $S' R^{-1}\, S\theta = 0$ implica $\theta = 0$. Ne segue che lo spazio nullo di S contiene solo il vettore zero. Ovviamente se lo spazio nullo di 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 $\sigma^2 [S' R^{-1} S]^{-1}$. Analogamente quella di uno stimatore corretto di σ^2 non può essere inferiore a $\frac{2\sigma^4}{n}$.

Calcoliamo ora lo stimatore di θ. Dalla $\partial \ell / \partial \theta = 0$, tenendo conto dell'invertibilità di $S' R^{-1} S$ si ricava

$$\hat{\theta}(y) = [S' R^{-1} S]^{-1}\, S' R^{-1} y \quad . \tag{3.12}$$

Inoltre $\hat{\theta}(y)$ fornisce il *massimo assoluto* (rispetto a θ) di $\ell(y, \theta, \sigma)$ dato che la matrice Hessiana

$$\frac{\partial^2 \ell}{\partial \theta_i \, \partial \theta_j} = -\frac{1}{\sigma^2} \, S' R^{-1} S$$

è definita negativa. Quindi $\hat{\theta}(\cdot)$ è lo stimatore di M.V. di θ. Useremo spesso l'abbreviazione

$$\hat{\theta}(y) = Ay \quad , \qquad A := [S' R^{-1} S]^{-1} \, S' R^{-1} \quad . \tag{3.13}$$

Proprietà dello stimatore di M.V. nel modello lineare

Lo stimatore (3.12) del parametro θ ha le seguenti proprietà

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

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

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

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

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

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

$$\hat{\theta}(\mathbf{y}) \sim N \Big(\theta, \sigma^2 \, [S' R^{-1} S]^{-1} \Big) \quad .$$

Queta proprietà segue dalla linearità dello stimatore.

Interpretazione geometrica

Dalla (3.8) è evidente che $\hat{\theta}(y)$ è la funzione che *minimizza*, *rispetto a* θ, *la forma quadratica* $(y - S\theta)' \, R^{-1} (y - S\theta)$, in corrispondenza ad ogni prefissato vettore di osservazioni $y \in \mathbb{R}^n$. Questa forma quadratica si può interpretare come il quadrato di una *distanza* in $\mathbb{R}^n$ *indotta dal prodotto scalare* $\langle x, y \rangle_{R^{-1}} := x' R^{-1} y$, nel senso che

$$(y - S\theta)' \, R^{-1} (y - S\theta) = \|y - S\theta\|^2_{R^{-1}} \quad , \tag{3.14}$$

con ovvio significato dei simboli. Per un dato $y \in \mathbb{R}^n$ minimizzare la distanza (3.14), rispetto a θ, significa *cercare il vettore* $v \in \mathcal{S} := \text{span } (S)$ (lo spazio vettoriale generato dalle colonne di S) *che ha minima distanza da* y.

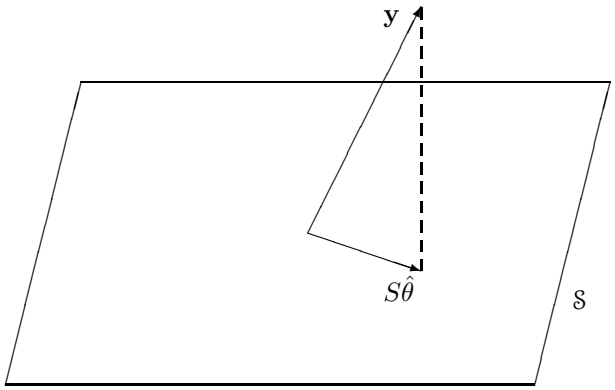

Figura 3.1.1. *Proiezione ortogonale.*

Ne viene che $S\hat{\theta}(y) := S\,Ay$ è la proiezione ortogonale di y sullo spazio $\mathcal{S} = \text{span}\,(S)$. In altri termini, la matrice $P \in \mathbb{R}^{n\times n}$, definita ponendo

$$P = SA \quad , \tag{3.15}$$

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

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

Conviene notare anche che P non è simmetrica come accade nella metrica Euclidea ordinaria, ma piuttosto

$$P' = (SA)' = A'S' = R^{-1}S\,[S'R^{-1}S]^{-1}S' = R^{-1}SAR = R^{-1}PR\,, \tag{3.16}$$

cioè P' è simile a P.

Basandosi sulla classica caratterizzazione geometrica della proiezione ortogonale (Teorema **??** Cap II), si trova allora che l'unico vettore $S\theta$ di $\mathcal{S}$ 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 $\mathcal{S}$ rispetto al prodotto scalare $\langle\cdot,\cdot\rangle_{R^{-1}}$.

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

$$S \perp (y - S\theta) \quad . \tag{3.17}$$

In altre parole

$$S'R^{-1}y - S'R^{-1}\,S\theta = 0 \tag{3.18}$$

e da questa equazione si ricava la nota espressione di $\hat{\theta}(y)$. In altre parole,

Proposizione 3.2. *Nel modello lineare-Gaussiano* (3.7) *lo stimatore a M.V. di* θ *coincide con la funzione dei dati osservati che minimizza la distanza quadratica* (3.14)*. In altre parole,* $\hat{\theta}(y)$ *è lo* stimatore ai minimi quadrati pesati di θ *con matrice peso* R^{-1}.

Riprenderemo la nozione di stimatore ai minimi quadrati più avanti. Occupiamoci ora del calcolo dello stimatore di σ^2. Dalla $\partial\ell/\partial\sigma^2 = 0$ si ricava

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

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 ora ricordare alcune proprietà della distribuzione χ^2.

3.2 La distribuzione χ^2

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

$$P(x \le \mathbf{y} < x + dx) = \frac{1}{2^{n/2}\,\Gamma\left(\frac{n}{2}\right)}\, x^{\left(\frac{n}{2}\right)-1}\, e^{-x/2}\, dx \quad , \tag{3.19}$$

per $x \ge 0$ e zero altrimenti. Nella (3.19) 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

$$\phi(it) := E\, e^{it\,\mathbf{y}} = (1 - 2it)^{-n/2} \quad , \tag{3.20}$$

come si ricava da ordinarie tabelle di trasformate di Fourier. Calcolando i primi momenti si ottiene

$$\begin{aligned} \mu_1 &= n \\ \mu_2 &= 2n \\ \mu_3 &= 8n \\ \mu_4 &= 48n + 12n^2 \qquad \text{ecc...} \end{aligned} \tag{3.21}$$

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

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

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

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

Lemma 3.1 (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*

$$\mathbf{x}_n \xrightarrow{L} \mathbf{x} \qquad \textit{se e solo se} \qquad \phi_n(t) \to \phi(t) \quad , \qquad \forall t \quad . \tag{3.22}$$

Notiamo allora che la f.c., $\phi_n(t)$, di $\mathbf{z}_n$ si può scrivere,

$$\phi_n(t) = E\, e^{it\frac{\mathbf{y}}{\sqrt{2n}}}\, e^{-it\frac{n}{\sqrt{2n}}} = e^{-it\frac{n}{\sqrt{2n}}} \left(1 - \frac{2it}{\sqrt{2n}}\right)^{-n/2}$$

$$= \left(e^{-it\sqrt{\frac{2}{n}}}\right)^{n/2} \left(1 - it\sqrt{\frac{2}{n}}\right)^{-n/2}$$

$$= \left[e^{it\sqrt{\frac{2}{n}}} - it\sqrt{\frac{2}{n}}\, e^{it\sqrt{\frac{2}{n}}}\right]^{-n/2} = \left(1 - \frac{t^2}{n} + \frac{\psi(n)}{n}\right)^{-n/2} \quad ,$$

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 $N(n, 2n)$.

Teorema 3.1. *La somma di N variabili casuali indipendenti $\mathbf{y}_1, \ldots, \mathbf{y}_N$ è distribuita secondo $\chi^2(n)$ se e solo se ciascuna variabile $\mathbf{y}_i$ è $\chi^2(n_i)$. In questo caso si ha*

$$n = \sum_{i=1}^{N} n_i \quad , \tag{3.23}$$

cioè i gradi di libertà si sommano.

Dimostrazione. La prova di questo teorema 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 espressioni del tipo (3.20) si vede in effetti che i gradi di libertà si sommano. □

La distribuzione χ^2 interviene in molte questioni di inferenza statistica. Ricordiamone alcune proprietà, particolarmente importanti.

Proposizione 3.3. *La distribuzione di*

$$n\bar{S}^2/\sigma^2 := \frac{1}{\sigma^2} \sum_{1}^{n} (\mathbf{y}_i - \mu)^2 \quad ,$$

con $\mathbf{y}_i \sim N(\mu, \sigma^2)$ e indipendenti è $\chi^2(n)$.

In effetti basta mostrare che la d.d.p. di $\mathbf{z} := (\mathbf{y} - \mu)^2/\sigma^2$ con $\mathbf{y} \sim N(\mu, \sigma)$ è $\chi^2(1)$ e poi usare il Teorema 3.1. Notiamo che si può scrivere $\mathbf{z} = \mathbf{x}^2$ con $\mathbf{x} \sim N(0, 1)$. 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

$$p_{\mathbf{z}}(z) = \frac{1}{\left|\frac{d}{dx}\, f(x)\right|_{x=f^{-1}(z)}} \left[p_{\mathbf{x}}(\sqrt{z}) + p_{\mathbf{x}}(-\sqrt{z})\right]\ 1(z)$$

$$= \frac{1}{|2\sqrt{z}|}\ \frac{1}{\sqrt{2\pi}}\ [e^{-z/2} + e^{-z/2}]\ 1(z) = \frac{1}{\sqrt{2\pi z}}\ e^{-z/2} \quad , \qquad z \geq 0 \quad ,$$

che è proprio $\chi^2(1)$.

Proposizione 3.4. *La distribuzione della varianza campionaria normalizzata*

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

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

Mostriamo allo scopo il seguente risultato notevole.

Lemma 3.2. *Nelle ipotesi poste, le statistiche* $\bar{\mathbf{y}}$ *ed* S *sono indipendenti.*

Dimostrazione. Per provare il lemma basta far vedere che $\bar{\mathbf{y}}$ e $\mathbf{y}_i - \bar{\mathbf{y}}$ sono scorrelate qualunque sia i. Questo implica che $\bar{\mathbf{y}}$ e $(\mathbf{y}_i - \bar{\mathbf{y}})$, $i = 1, \ldots, n$, sono indipendenti, data l'ipotesi di Gaussianità e quindi l'asserto.

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

$$E\,\tilde{\mathbf{y}}\tilde{\mathbf{y}}_i = \frac{1}{n}\ E\left(\sum_1^n\ \tilde{\mathbf{y}}_k\tilde{\mathbf{y}}_i\right) = \frac{1}{n}\ E(\tilde{\mathbf{y}}_i)^2 = \frac{\sigma^2}{n}$$

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

Usiamo ora la solita identità

$$\sum_1^n\ (\mathbf{y}_i - \mu)^2 = \sum_1^n\ (\mathbf{y}_i - \bar{\mathbf{y}})^2 + n(\bar{\mathbf{y}} - \mu)^2 \tag{3.24}$$

per scrivere

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

dove la somma al secondo membro è di due v.c. *indipendenti*. Sappiamo da A) che $n\bar{S}^2/\sigma^2 \sim \chi^2(n)$ e che $(\bar{\mathbf{y}} - \mu)^2/(\sigma^2/n) \sim \chi^2(1)$ (questo scende ancora dalla proposizione 3.3 con $n = 1$). Per il Teorema 3.1 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 $\mathbf{y}'Q\mathbf{y}$ con $Q = Q'$, che hanno una distribuzione χ^2. Il caso più semplice è il seguente.

Proposizione 3.5. *Se* $\mathbf{y} \sim N(\mu, \Sigma)$ *con* $\mu \in \mathbb{R}^m$ *e* $\Sigma \in \mathbb{R}^{m\times m}$ *definita positiva, allora*

$$(\mathbf{y}-\mu)'\,\Sigma^{-1}(\mathbf{y}-\mu) \sim \chi^2(m) \quad . \tag{3.25}$$

In effetti basta standardizzare $\mathbf{y}$, ponendo

$$\mathbf{z} := \Sigma^{-1/2}(\mathbf{y}-\mu) \quad ;$$

allora $\mathbf{z} = [z_1, \ldots, \mathbf{z}_m]'$ è $N(0, I)$, cioè $\mathbf{z}_1, \ldots, \mathbf{z}_m$ sono *indipendenti* ed $N(0,1)$. Con la posizione fatta si ha poi

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

e quindi l'ultimo membro è $\chi^2(m)$ per il Teorema 3.1.

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

Proposizione 3.6. *Sia* $\mathbf{z} \sim N(0, I)$. *Allora la forma quadratica* $\mathbf{z}'Q\mathbf{z}$ *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* $r = \text{rango}\, Q$.

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 $\mathbb{R}^n$. I suoi autovalori non nulli sono pertanto uguali a uno (in numero di $r = \text{rango}\, Q$). La conclusione segue ancora dal Teorema 3.1.

Caratterizzazione dello stimatore della varianza

Teorema 3.2. *Lo stimatore di M.V. della varianza* σ^2 *nel modello lineare* (3.7) *ha distribuzione di probabilità corrispondente alla*

$$\frac{n\hat\sigma^2(\mathbf{y})}{\sigma^2} \sim \chi^2(n-p) \quad . \tag{3.26}$$

In particolare la sua media e varianza sono date da

$$E_{\theta,\sigma^2}\ \hat\sigma^2(\mathbf{y}) = \sigma^2\ \frac{n-p}{n} \quad , \tag{3.27}$$

$$\text{Var}_{\theta,\sigma^2}\ \hat\sigma^2(\mathbf{y}) = \sigma^4\ \frac{2(n-p)}{n^2} \quad . \tag{3.28}$$

Dimostrazione. Ricordiamo 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 $N(0, I)$. Dalla (3.18) si ricava poi con facili passaggi la

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

dove si è usata la proprietà di similitudine $P' = R^{-1}\,PR$ stabilita nella (3.16). Notiamo ora che la matrice tra parentesi quadre, diciamola Q, è idempotente giacché, sempre per la (3.16), si ha

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

e inoltre il suo rango è $n-p$. Infatti $I-P$ proietta su un sottospazio ortogonale a S e nelle ipotesi correnti $\dim S = p$. Segue allora per la proprietà D) stabilita più sopra che $\mathbf{z}'Q\mathbf{z} \sim \chi^2(n-p)$. □

Osservazione 3.1. *Come si vede dalla (3.27) 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*

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

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* $2\sigma^4/n$.

A conclusione di questa sezione è importante mettere in evidenza ancora una volta il fatto che la stima di M.V. del parametro θ nel modello lineare Gaussiano (3.7) è stata ridotta a un problema di minimizzazione di una distanza quadratica pesata in $\mathbb{R}^n$ fra il vettore dei dati y e una loro descrizione parametrica $y \simeq S\theta$, la matrice peso essendo uguale all'inversa della varianza dal "rumore di osservazione" $\mathbf{w}$. In altre parole, il calcolo di $\hat{\theta}(\mathbf{y})$ e $\hat{\sigma}(\mathbf{y})$ per il modello lineare e gaussiano (3.7) si riduce alla soluzione di un problema di *minimi quadrati pesati*. Come vederemo nel prossimo capitolo, un problema di approssimazione ai minimi quadrati dei dati mediante una funzione lineare di θ ha una soluzione che è sempre una funzione *lineare* nelle osservazioni. Notiamo però che nelle ipotesi di rumore Gaussiano, questo stimatore lineare ha la minima varianza nella classe di tutti gli stimatori corretti di θ, (questa classe potendo includere a priori anche funzioni nonlineari arbitrariamente complicate dei dati). Questa osservazione fornisce un importante legame logico fra quanto è stato esposto in questo capitolo e il metodo di stima ai minimi quadrati, che è assai più primitivo della M.V. (ma di impiego più generale) e che verrà illustrato nel seguito.

3.3 Il principio dei Minimi Quadrati e il suo significato statistico

In generale si dispone raramente di informazione sufficiente per descrivere i dati mediante modelli probabilistici di struttura nota come il modello lineare-Gaussiano (3.7). Si ha un insieme di osservazioni $\{y_1, \ldots, y_n\}$, che noi qui per semplicità supporremo scalari (ma l'estensione al caso vettoriale di quanto verremo esponendo è immediata), che dipendono in genere da una variabile esogena o "di ingresso" $u \in R^q$ i cui valori $\{u_t \ ; \ t = 1, \ldots, n\}$ possono essere misurati (o talvolta scelti) dallo sperimentatore ad ogni serie di misure. Sulla dipendenza di ciascuna misura y_t da u_t (o eventualmente da tutte le $u_1, \ldots, u_n$) si potrebbe anche avere scarsa informazione a priori. In ogni caso, basandosi sulla conosceenza a priori del meccanismo che genera i dati, si assegna a priori una *classe parametrica* di modelli del tipo

$$y_t = f(u_t, \theta, t) \qquad t = 1, \ldots, n \tag{3.29}$$

dove f è una funzione nota e θ un parametro p-dimensionale da determinarsi in modo tale che si abbia la "migliore" descrizione possibile dei dati misurati $(y_1, \ldots, y_n)$ corrispondenti a certi valori assegnati $(u_1, \ldots, u_n)$ alla variabile u negli istanti di misura.

Dettato soprattutto da ragioni di semplicità matematica, come criterio in base al quale si definisce la "migliore" descrizione dei dati si usa spesso la *somma dei quadrati degli scarti tra le misure vere* $\{y_t\}$ *e quelle predette dal modello*

$$\hat{y}_t(\theta) = f(u_t, \theta, t) \quad , \qquad t = 1, \ldots, n \quad , \tag{3.30}$$

in corrispondenza ai valori assegnati $(u_1, \ldots, u_n)$ di u_t e al valore generico θ del parametro.

Si arriva in questo modo a definire una cifra di merito

$$V(\theta) := \sum_1^n \ [y_t - \hat{y}_t(\theta)]^2 = \sum_1^n \ [y_t - f(u_t, \theta, t)]^2 \tag{3.31}$$

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

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

Chiaramente $\hat{\theta}$ dipende da $(y_1, \ldots, y_n)$ e dai valori assegnati $(u_1, \ldots, u_n)$ alla variabile esogena u negli n esperimenti. Scriveremo allora

$$\hat{\theta} = \hat{\theta}(y_1, \ldots, y_n \ , \ u_1, \ldots, u_n) \quad , \tag{3.32}$$

interpretando $\hat{\theta}$ anche come *funzione* dei dati (= misure $\{y_t\}$ più "ingressi" $\{u_t\}$). La funzione $\hat{\theta}$ si chiama *stimatore ai minimi quadrati* di θ e la (3.32) *stima* ai minimi quadrati di θ. Notiamo che queste parole non hanno, per ora, alcun significato statistico. Il criterio dei minimi quadrati (M.Q.) è quindi una 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*.

I problemi di modellizzazione empirica cui abbiamo appena accennato possono venire classificati in vari modi.

1) Se nella (3.29) f non dipende esplicitamente da t si parla di *problemi di regressione*. In questo caso si vuole semplicemente trovare una rappresentazione analitica del legame tra la variabile esogena u (ingresso misurabile) e la variabile dipendente y (uscita).

2) Se nella descrizione del problema non vi sono variabili esogene, u, cioè se f dipende solo da θ e da t, si parla di *curve fitting*.

Una distinzione poco precisa ma importante è tra problemi in cui la struttura del modello è assegnata dalla "fisica" dell'esperimento e problemi in cui la scelta di f è a priori affatto arbitraria. Si parla in questi casi di problemi a struttura fissa o a *scatola grigia* e problemi *a scatola nera*. Esempi del primo tipo sono i seguenti

A) Determinare sperimentalmente l'equazione di stato di un gas. Si sa che la relazione tra le grandezze p, V, T deve essere del tipo

$$pV^\gamma = kT \quad ;$$

prendendo ad esempio $y = p$ e $u = (V, T)$ si può scrivere

$$p = kV^\gamma T$$

e il parametro da determinarsi in base a una serie di esperimenti

$$\begin{bmatrix} (V_1 T_1) & \rightarrow & p_1 \\ \vdots & & \vdots \\ (V_N T_N) & \rightarrow & p_N \end{bmatrix}$$

è il parametro bidimensionale $\theta := (\gamma, k)$.

B) Determinare sperimentalmente i valori dei parametri elettrici (R, L, C) a partire da una registrazione della scarica di una rete di struttura nota. In questo caso si sa che la misura è esprimibile mediante una relazione del tipo

$$y(t) = A_1\, e^{-t/T_1} + A_2\, e^{-t/T_2} + \dots \quad ,$$

dove A_k e T_k sono funzioni note dei parametri elettrici e delle condizioni iniziali.

Notiamo che nei problemi a "scatola grigia i parametri θ nel modello (3.29) hanno un preciso significato fisico e spesso in questi casi lo scopo della stima è di ricavare informazioni su θ più che approssimare in modo ottimo le misure.

Nei problemi **a scatola nera** la fisica che governa l'esperimento è invece poco nota oppure porta a relazioni matematiche troppo complicate e poco affidabili. Si impone allora ai dati un modello di struttura prefissata, in genere la più semplice possibile e meglio trattabile analiticamente (quasi sempre una legge *lineare*). In questi casi però, per costruire modelli che descrivano bene l'andamento della variabile dipendente in un certo campo di condizioni operative (ad esempio modelli usati a scopo previsionale), occorre una scelta molto oculata della struttura e della complessità del modello e, a posteriori, una fase di validazione della struttura scelta.

Un esempio di questa seconda classe di problemi è il seguente.
Si vuole trovare un modello matematico che leghi la produzione per ettaro, y, di una certa coltura alla quantità di fertilizzante, x, e alla quantità di acqua di irrigazione, z. Si hanno dati storici $\{y_t\}$ corrispondenti a certi valori $\{x_t, z_t\}$ per un certo appezzamento di terreno. Per parametrizzare la funzione

$$y = f(x, z)$$

che si vuole determinare, si possono a priori seguire infinite strade, ad esempio supporre f lineare

$$y = \theta_0 + \theta_1 x + \theta_2 z \quad ,$$

o polinomiale

$$y = \theta_0 + \theta_1 x + \theta_2 z + \theta_3 x^2 + \theta_4 xz + \theta_5 z^2 \quad ,$$

oppure prendere

$$y = \theta_0 \; x^{\theta_1} \; z^{\theta_2} \qquad \text{ecc...}$$

Sebbene tutti questi modelli siano riconducibili a problemi di regressione lineari nei parametri (nel terzo caso basta usare i logaritmi) è ovvio che la bontà dell'adattamento ai dati ottenibili nei diversi casi può essere notevolmente diversa. È bene mettere in evidenza fin da ora che la mancanza di informazione a priori sulla struttura "fisica" delle relazioni tra le variabili in gioco si paga sempre in pratica con la necessità di *estensive verifiche a posteriori* sulla significatività dei risultati che si ottengono.

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 [**?**, **?**, **?**, **?**].

Minimi quadrati e serie di Fourier

L'approssimazione ai minimi quadrati è un'idea elementare ma di vastissima portata. Ha ad esempio ispirato la teoria delle serie di funzioni ortonormali. Data una funzione periodica $y(t)$ nell'intervallo $[-T/2, T/2]$ la ridotta n-sima della serie di Fourier di $y(t)$

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

è la combinazione lineare delle funzioni 1, $\sin\frac{2\pi}{T}\,t, \ldots, \sin\frac{2n\pi}{T}\,t$, $\cos\frac{2\pi}{T}\,t$, …, $\cos\frac{2n\pi}{T}\,t$, secondo i coefficienti θ_i, $i = 0, 1, \ldots, 2n$, tale per cui lo scostamento quadratico

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

è minimo. In altri termini i coefficienti di Fourier di f sono stime ai minimi quadrati dei parametri del modello (lineare) $f_n(t,\theta)$ usato per approssimare f. Come vederemo meglio più avanti, l'ortogonalità delle funzioni base usate per la modellizzazione semplifica in modo drammatico la stima dei coefficienti.

Minimi quadrati pesati

In molti casi le misure effettuate non hanno tutte la stessa attendibilità ed è perciò 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,

$$V_Q(\theta) := \sum_1^n \, [y(t) - f(u_t, \theta, t)]^2 \; q_t \quad , \tag{3.33}$$

dove $q_1, \ldots, q_n$ sono numeri positivi, grandi se le misure corrispondenti sono affidabili e piccoli se non lo sono. La (3.33) si può riscrivere come

$$V_Q(\theta) = [y - f(u,\theta)]' \; Q \, [y - f(u,\theta)] = \|y - f(u,\theta)\|_Q^2 \quad , \tag{3.34}$$

dove

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad , \qquad f(u,\theta) = \begin{bmatrix} f(u_1,\theta,1) \\ \vdots \\ f(u_n,\theta,n) \end{bmatrix} \tag{3.35}$$

e $Q = \mathrm{diag}\{q_1, \ldots, q_n\}$. La generalizzazione della (3.34) al caso in cui Q non è diagonale, ma sempre *definita positiva e simmetrica* si presenta spontanea.

La minimizzazione di $V(\theta)$ si può fare esplicitamente nel caso in cui il modello (3.29) è *lineare nei parametri*, cioè quando

$$f(u_t, \theta, t) = \sum_1^p \, \theta_i \, \phi_i(u_t, t) \quad . \tag{3.36}$$

Siccome u_t è una quantità *nota* si può anche ometterla nell'argomento delle ϕ_i (a meno di non voler considerare problemi di scelta ottima o *programmazione ottima* dei valori di u nelle varie misure in modo tale da minimizzare $V(\theta)$ anche rispetto a $(u_1, \ldots, u_n)$).

Scriveremo così la (3.35) nel modo usuale e cioè

$$f(u_t, \theta, t) := s'(t) \; \theta \quad , \tag{3.37}$$

con $s'(t)$ vettore riga p-dimensionale, funzione *nota* dell'indice t. Il blocco delle n misure sia rappresentato dal vettore n-dimensionale y definito in (3.35) e sia S la matrice $n \times p$

$$S = \begin{bmatrix} s'(1) \\ \vdots \\ s'(n) \end{bmatrix} \quad . \tag{3.38}$$

La minimizzazione di $V_Q(\theta)$ diventa allora il problema di minimizzare la forma quadratica in θ,

$$V_Q(\theta) = [y - S\theta]' \, Q[y - S\theta] = \|y - S\theta\|_Q^2 \quad , \tag{3.39}$$

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 $\langle x, y \rangle_Q = x'Qy$) alle colonne di S, ovvero

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

che si può riscrivere come

$$S'Q \; S\theta = S'Qy \quad . \tag{3.40}$$

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

Nel seguito supporremo che sia

$$\operatorname{rango} S = p \leq n \quad . \tag{3.41}$$

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

$$\hat{\theta}(y) = [S'QS]^{-1} \; S'Qy \quad , \tag{3.42}$$

dalla quale si vede che lo stimatore ai M.Q. è *sempre una funzione lineare* delle misure. Nel caso in cui la (3.41) non valga, si può usare la pseudoinversa di $S'QS$, 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

$$P = S \, [SQS]^{-1} \, S'Q \tag{3.43}$$

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

$$\hat{\varepsilon}_t := y_t - s'(t) \, \hat{\theta} \quad , \qquad t = 1, \ldots, n \quad , \tag{3.44}$$

vale

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

Avevamo già incontrato queste formule studiando le proprietà dello stimatore di M.V. di θ nel modello lineare e Gaussiano (3.7). Abbiamo visto che 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.

Osservazione 3.2. *Lo stimatore di M.V. del parametro θ nel modello lineare Gaussiano* (3.7) *è uno stimatore ai M.Q. pesati con matrice Q uguale all'inversa della matrice di varianza del rumore* $\mathbf{w}$.

Questo risultato è per ora solo una curiosa coincidenza. Notiamo che se la distribuzione di $\mathbf{w}$ *non è Gaussiana*, i calcoli fatti alla sezione 3.1 non hanno più valore. In genere lo stimatore di M.V. di θ in un modello lineare in cui $\mathbf{w}$ non è normalmente distribuito è una funzione *non lineare* delle osservazioni e pertanto, vista la (3.42), non può in nessun caso essere uno stimatore ai M.Q..

Siamo così pervenuti alla questione del significato statistico del principio dei M.Q..

Significato statistico della stima ai M.Q.

Ovviamente, per analizzare questo significato bisogna introdurre un minimo di *informazione a priori di tipo probabilistico* sul meccanismo secondo il quale i dati sono effettivamente generati. Si tratta a questo punto di ipotizzare il meno possibile e nello stesso tempo individuare un contesto che permetta di valutare, seppure in modo vago, la *bontà statistica* dello stimatore ai M.Q. (3.42).

Notiamo d'altra parte che il modo in cui lo stimatore ai M.Q. può tenere conto dell'informazione probabilistica che aggiungiamo al modello è abbastanza limitato. L'unico parametro su cui si può giocare è la matrice dei pesi Q.

Ipotesi (Ipotesi sul meccanismo di generazione dei dati). *Supponiamo che le misure $\{y_t\}$ siano generate da un modello lineare del tipo* (3.7)

$$\mathbf{y}_t = s'(t)\,\theta + \sigma \mathbf{w}_t \quad , \qquad t = 1, \ldots, n \quad , \tag{3.45}$$

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

$$E\mathbf{w} = 0 \quad , \qquad \mathrm{Var}\,(\mathbf{w}) = E\mathbf{w}\mathbf{w}' = \sigma^2 R \quad . \tag{3.46}$$

Questo è l'usuale modello lineare di cui ci siamo occupati nel capitolo precedente, con la differenza che ora *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.

Notiamo che l'ipotesi che $\mathbf{w}$ abbia media zero non è affatto essenziale. Conglobando per il momento il parametro σ nell'errore di misura, si assuma che $\bar{\mathbf{w}}_t := \sigma \mathbf{w}_t$ abbia media μ indipendente da t e si ponga $\tilde{\mathbf{w}}_t := \bar{\mathbf{w}}_t - \mu$. Si possono allora riscrivere le misure ponendo

$$\mathbf{y}_t = s'(t)\,\theta + \mu + \tilde{\mathbf{w}}_t \quad , \qquad t = 1, \ldots, n \quad ,$$

dove il vettore aleatorio $\tilde{\mathbf{w}}$ ha media zero e la stessa varianza $\sigma^2 R$ del modello lineare di partenza. Introducendo il nuovo parametro $\theta_{p+1} := \mu$ e aggiungendo una colonna di uno alla matrice S, si ottiene il modello lineare aumentato

$$\mathbf{y} = \begin{bmatrix} & 1 \\ S & \vdots \\ & 1 \end{bmatrix} [\theta_1, \ldots, \theta_p, \theta_{p+1}]' + \tilde{\mathbf{w}} \quad ,$$

in cui $\tilde{\mathbf{w}}$ ha media zero.

Vediamo quali sono le proprietà statistiche dello stimatore ai M.Q. in questo contesto.

Proposizione 3.7. *Qualunque sia $Q > 0$ lo stimatore* (3.42) *è corretto.*

Difatti

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

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, $R = \text{diag}\{r_1, \ldots, r_n\}$, 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

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

Notiamo che il termne $1/\sigma^2$ (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 in (3.33).

La varianza di $\hat{\theta}(\mathbf{y})$ si calcola facilmente a partire dalla (3.46),

$$\text{Var } \hat{\theta}(\mathbf{y}) = [S'QS]^{-1} S'Q\,\sigma^2 R\,QS[S'QS]^{-1} \quad ; \tag{3.47}$$

questa 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 (3.45) 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

$$E\,\phi(\mathbf{y}) = \theta \qquad \text{ovvero} \qquad AS = I \quad , \tag{3.48}$$

Cerchiamo allora in questa classe quella funzione, $\hat{\phi}$, che ha *varianza minima*

$$\text{Var } \hat{\phi}(\mathbf{y}) \leq \text{Var } \phi(\mathbf{y}) \quad . \tag{3.49}$$

La soluzione di questo problema è fornita dal celebre

Teorema 3.3 (di Gauss-Markov). *Il miglior stimatore lineare di θ, per il modello* (3.45) *nel senso appena definito, è lo stimatore di Markov.*

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

$$\sigma^2 \, ARA' \geq \sigma^2 \, (S'R^{-1}S)^{-1} \quad , \tag{3.50}$$

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

$$\hat{A} = [S'R^{-1}S]^{-1} \, S'R^{-1} \quad .$$

è lineare e corretto e la sua varianza è esattamente uguale al secondo membro in (3.50).

Per provare la (3.50) ci si rifà alla disuguaglianza (equivalente alla non-negatività della varianza dell'errore di stima nella teoria della stima lineare Bayesiana [4]),

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

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 (3.48) si ottiene la (3.50). □

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

$$\boldsymbol{\varepsilon}_t(\theta) := \mathbf{y}_t - s'(t) \, \theta \quad , \qquad t = 1, \ldots, n \quad .$$

Notiamo infine che, in accordo con quanto anticipato nelle osservazioni 14.1 e 14.3, *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* (3.45) *è (nota e) Gaussiana*. In genere però, se $\mathbf{w}$ non è normalmente distribuito, la varianza (**??**) può risultare assai più grande della varianza del corrispondente stimatore di M.V. di θ.

A questo punto possiamo concludere la nostra analisi con la seguente affermazione.

Proposizione 3.8. *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), l'espressione* (3.12) *(non fornisce necessariamente lo stimatore a M.V., ma) fornisce comunque lo stimatore che ha minima varianza nella classe degli stimatori* lineari *e corretti*[5] *di θ.*

[4] 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 $\tilde{\mathbf{x}} := \mathbf{x} - \hat{E}\,[\mathbf{x} \mid \mathbf{y}]$.

[5] 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 di σ^2 corrispondente allo stimatore di Markov

Per costruire uno stimatore di σ^2, si può ancora utilizzare la formula

$$\hat{\sigma}^2(y) = \frac{1}{n} \|y - Py\|^2_{R^{-1}} = \frac{1}{n} V_{R^{-1}}(\hat{\theta}) \quad . \tag{3.51}$$

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

$$\begin{aligned} nE\Big(\hat{\sigma}^2(\mathbf{y})\Big) &= E\Big(\mathbf{y}'(I-P)'R^{-1}(I-P)\,\mathbf{y}\Big) \\ &= E\Big(\mathbf{w}'(I-P)'R^{-1}(I-P)\,\mathbf{w}\Big) = E\Big(\mathbf{w}'R^{-1}(I-P)\,\mathbf{w}\Big) \\ &= E\,tr\Big\{\mathbf{w}'R^{-1}(I-P)\,\mathbf{w}\Big\} = E\ \mathrm{Tr}\ \Big\{R^{-1}(I-P)\ \mathbf{w}\mathbf{w}'\Big\} \\ &= E\ \mathrm{Tr}\ \Big\{(I-P)\ (\mathbf{w}\mathbf{w}')\ R^{-1}\Big\} = \mathrm{Tr}\ \Big\{(I-P)\ E(\mathbf{w}\mathbf{w}')\ R^{-1}\Big\} \\ &= \sigma^2\ \mathrm{Tr}\,(I-P) \quad ; \end{aligned} \tag{3.52}$$

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, $\mathrm{Tr}\ P = \dim \mathcal{S} = p$ e quindi

$$E\,\hat{\sigma}^2(\mathbf{y}) = \frac{n-p}{n}\,\sigma^2 \quad . \tag{3.53}$$

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 $c'\theta$ anziché di θ (c' è un vettore riga noto), quello corretto e di varianza minima (sempre nella classe degli stimatori lineari) è semplicemente $c'\hat{\theta}$, 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.

Confronto con l'approccio Bayesiano

Dopo aver passato in rassegna gli aspetti salienti della stima sul modello lineare (3.45) seguendo l'approccio Fisheriano, è interessante fare ora un confronto critico con le formule dell'approccio Bayesiano.

Riscriviamo allora le formule per lo stimatore e la sua varianza d'errore ottenuta seguendo i due approcci (supponiamo ora che σ^2 sia nota e conglobata in R):

$$\text{BAYES} \quad \begin{cases} \hat{x}(\mathbf{y}) &= (P^{-1} + S'R^{-1}S)^{-1}\,S'R^{-1}\mathbf{y} \\ \Lambda &= (P^{-1} + S'R^{-1}S)^{-1} \end{cases} \tag{3.54}$$

$$\begin{matrix}\text{FISHER} \\ \text{(MARKOV)}\end{matrix} \quad \begin{cases} \hat{\theta}(\mathbf{y}) &= [S'R^{-1}S]^{-1}\,S'R^{-1}\mathbf{y} \\ \Sigma &= [S'R^{-1}S]^{-1} \quad . \end{cases} \tag{3.55}$$

Come preannunciato a suo tempo, si vede che quando $P \to \infty$ (ovvero l'informazione a priori su $\mathbf{x}$ diventa sempre più incerta) *le formule Bayesiane coincidono con quelle Fisheriane*. Si vede anche che qualunque sia P si ha sempre

$$\Lambda \leq \Sigma \quad ,$$

dato che

$$P^{-1} + S'R^{-1}S \geq S'R^{-1}S \quad .$$

per cui la stima Bayesiana è sempre "migliore" di quella fatta in assenza di informazione a priori su $\mathbf{x}$ (e ci saremmo sorpresi del contrario).

Quanto sopra per mostrare che le formule del caso Fisheriano si ottengono come caso limite di quelle Bayesiane. È istruttivo tentare anche la via opposta. Si può pensare di dare un'interpretazione Fisheriana alle formule (3.54) introducendo delle misure "fittizie", equivalenti alla conoscenza a priori su $\mathbf{x}$. Supponiamo allora di avere una osservazione addizionale

$$\mathbf{y}_0 = S_0\mathbf{x} + \mathbf{w}_0 \quad , \tag{3.56}$$

effettuata *prima* di quella reale

$$\mathbf{y} = S\mathbf{x} + \mathbf{w} \quad .$$

In questa formula $\mathbf{x}$ è un *parametro* incognito n-dimensionale e i due rumori $\mathbf{w}_0$ e $\mathbf{w}$ sono *scorrelati*.

Scegliamo S_0 e la varianza, R_0, di $\mathbf{w}_0$ in modo tale per cui si abbia

$$P = (S_0' \, R_0^{-1} \, S_0)^{-1} \quad , \tag{3.57}$$

cioè P è la varianza dello stimatore di Markov $\hat{x}(\mathbf{y}_0)$ relativo al modello (3.56).

Dato che $\mathbf{w}_0$ e $\mathbf{w}$ sono scorrelati possiamo scrivere lo stimatore basato sulle due osservazioni sequenziali, $\mathbf{y}_0$ e $\mathbf{y}$, usando le formule del filtro di Kalman. Si ha

$$\begin{aligned} \hat{x}(\mathbf{y}_0, \mathbf{y}) &= \hat{x}(\mathbf{y}_0) + PS'(R + SPS')^{-1} \, [\mathbf{y} - S\hat{x}(\mathbf{y}_0)] \quad , \\ \Sigma &= P - PS'(R + SPS')^{-1} \, SP \quad . \end{aligned} \tag{3.58}$$

La seconda di queste formule è proprio l'espressione originale a suo tempo trovata per la covarianza dell'errore Λ. La prima si riduce all'espressione dello stimatore di Bayes *solo se si pone* $\hat{x}(\mathbf{y}_0) = 0$. (Infatti $I - PS'(R + SPS')^{-1}$ non può annullarsi, come vedremo fra un attimo).

Questo è ragionevole. L'informazione a priori nel caso Bayesiano *non è una misura* (non è costituita da un campione addizionale), ma riguarda solo la *distribuzione* di $\mathbf{x}$.

Le due formulazioni sono perciò da riguardarsi come essenzialmente diverse dal punto di vista statistico. Dal punto di vista algoritmico si hanno le seguenti equivalenze (valide per il modello lineare):

stima di Fisher	=	stima di Bayes con $P \to \infty$
stima di Bayes	=	stima di Fisher con un'osservazione preliminare $\mathbf{y}_0$ e condizione iniziale $\hat{x}(\mathbf{y}_0) = 0$.

L'esempio svolto serve a illustrare alcune considerazioni generali. La prima è relativa all'espressione per la covarianza d'errore Λ, data in termini generali dalla differenza

$$\Lambda = \Sigma_x - \Sigma_{xy}\,\Sigma_y^{-1}\,\Sigma_{yx} \quad , \tag{3.59}$$

che è sempre una matrice almeno semidefinita positiva, dato che si tratta di una varianza. Il termine che si sottrae ($\mathrm{Var}\ \hat{\mathbf{x}}$) rappresenta la riduzione nell'incertezza a priori (Σ_x) che si aveva su $\mathbf{x}$, a cui porta lo stimatore $\hat{\mathbf{x}}$. Si può dire che vale la pena di costruire lo stimatore quando la riduzione di varianza, $\Sigma_{xy}\,\Sigma_y^{-1}\,\Sigma_{yx}$, della varianza a priori di $\mathbf{x}$ che lo stimatore comporta, è significativa. Questo potrebbe sembrare un controsenso perché implica che un buon stimatore debba avere la varianza più grande possibile. In realtà questa è una conseguenza peculiare dell'approccio Bayesiano: $\hat{\mathbf{x}}$ non deve essere il meno disperso possibile attorno alla sua media (questo è il punto di vista Fisheriano), ma bensì il più prossimo possibile a $\mathbf{x}$ come variabile casuale (cfr. il significato di $E\,\|\mathbf{x} - \hat{x}(\mathbf{y})\|^2$). Una conseguenza poco intuitiva di questa diversità di approccio è che lo stimatore Bayesiano $\hat{x}(\mathbf{y})$ *non è mai uno stimatore corretto* (uniformemente). Infatti la media fatta rispetto alla distribuzione condizionata $f(y \mid x)$ (che ora gioca lo stesso ruolo della $p(y, \theta)$ nel caso Fisheriano) di $\hat{x}(\mathbf{y})$ non è x, ma bensì

$$E(\hat{x}(\mathbf{y}) \mid \mathbf{x}) = \Sigma_{xy}\,\Sigma_y^{-1}\,E(\mathbf{x} \mid \mathbf{y}) = \Sigma_{xy}\,\Sigma_y^{-1}\,\Sigma_{yx}\,\Sigma_x^{-1}\,\mathbf{x} := B\mathbf{x}$$

e la matrice B *non può essere l'identità*, dato che dalla (3.59) scende

$$\Lambda\,\Sigma_x^{-1} = I - B$$

e il primo termine *è non nullo* (si rammenti che Σ_x è assunta non singolare). Per il modello lineare si trova in particolare la

$$\Lambda\,\Sigma_x^{-1} = \Lambda\,P^{-1} = I - PS'(R + SPS')^{-1}\,S \neq 0 \quad ,$$

che era stata già richiamata qualche riga più in su.

3.4 Minimi quadrati non lineari

DA SCRIVERE

3.5 Aspetti numerici dei problemi ai minimi quadrati

Condizionamento Numerico delle equazioni normali

Dal punto di vista numerico, la soluzione al calcolatore delle equazioni normali

$$S'QS\,\theta = S'Qy \tag{3.60}$$

può risultare problematica se p è maggiore di $6 \div 7$. Questo perché 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.

In questo paragrafo cercheremo di esporre (senza alcuna pretesa di completezza) alcune idee dell'Algebra Lineare Numerica che possono essere di aiuto quando si ha a che fare con problemi di questo genere. Per una trattazione più approfondita rimandiamo al testo di G. Strang [**?**], a quello di Lawson-Hanson [16] e al testo di Golub e van Loan [12] .

Il primo fatto di cui bisogna tenere conto è che il calcolatore usa un sistema approssimato di rappresentazione dei numeri reali ("floating point arithmetic"). In questo sistema 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 $\alpha = 3,5624179\overline{9}$ ha le rappresentazioni

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

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

$$Ax = b \quad , \tag{3.61}$$

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

Notiamo che 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 3.1. *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*

$$\frac{\|\delta x\|}{\|x\|} \gg \frac{\|\delta\alpha\|}{\|\alpha\|} \quad . \tag{3.62}$$

Esempi

Consideriamo il sistema di equazioni $Ax = b$,

$$\begin{bmatrix} 1 & 1 \\ 1 & 1.0001 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2.0001 \end{bmatrix} \quad ; \tag{3.63}$$

la sua soluzione è $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Se supponiamo di introdurre una perturbazione nel secondo membro, ad esempio

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

si verifica facilmente che la soluzione x diventa

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

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 (3.63) 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} ! $\diamondsuit$

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 3.2. *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 all'approssimazione 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.

Dato che le equazioni normali sono del tipo $Ax = b$, ci occuperemo brevemente del condizionamento numerico di questo problema. Lo schema intuitivo di quanto accade è il seguente.

Figura 5.1

δA e δb sono errori di arrotondamento su A e b, e δx è il corrispondente errore su $x = A^{-1}\, b$. 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*

$$\|x\| = \left|\Sigma\, x_i^2\right|^{1/2} \quad ,$$

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} \quad .$$

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:

$$\|A\|^2 = \sup_{x \neq 0} \frac{x' A' A x}{x' x} \quad . \tag{3.64}$$

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

$$\|A\|^2 = \max_i \ \lambda_i(A'A) \quad . \tag{3.65}$$

Usando la norma di A, si vede facilmente che da $x = A^{-1}\, b$ e $b = Ax$ segue

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

Indichiamo con $c(A)$ il numero $\|A\|\, \|A^{-1}\|$; $c(A)$ è chiamato (indice di) *condizionamento numerico* del problema $Ax = b$ (o della matrice A). Si ha allora

$$\frac{\|\delta x\|}{\|x\|} \leq c(A)\ \frac{\|\delta b\|}{\|b\|} \tag{3.66}$$

e la disuguaglianza è la migliore possibile. Si vede che $c(A)$ è il "coefficiente di amplificazione" degli errori sul termine noto b. Vedremo presto 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

$$\|A\|^2 = \lambda_{\rm MAX}\ (A'A)$$

$$\|A^{-1}\|^2 = \lambda_{\rm MAX}\ (A^{-T}\, A^{-1}) = \lambda_{\rm MAX}\ |(AA')^{-1}| = \frac{1}{\lambda_{\rm MIN}(AA')}$$

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

$$c^2(A) = \frac{\lambda_{\rm MAX}(A'A)}{\lambda_{\rm MIN}(A'A)} \quad ; \tag{3.67}$$

in particolare, se A è simmetrica,

$$c(A) = \left|\frac{\lambda_{\rm MAX}(A)}{\lambda_{\rm MIN}(A)}\right| \quad . \tag{3.68}$$

Da questa formula si vede che il condizionamento numerico di A è una misura di quanto A è "prossima" a essere singolare. Chiaramente le matrici meglio condizionate sono quelle per cui $A'A = I$. In questo caso infatti $\lambda_{\rm MAX}(A'A) = \lambda_{\rm MIN}(A'A) = 1$ e $c(A) = 1$. Queste matrici (*ortogonali*) giocano un ruolo fondamentale nell'analisi numerica.

A titolo di esempio calcoliamo il condizionamento numerico del problema (3.63). La matrice A è simmetrica e si trova subito (approssimativamente)

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

da cui

$$c(A) \cong 4\ 10^4 \quad .$$

Questo valore di $c(A)$ è in accordo con i risultati riportati nell'esempio precedente.

Problema 3.2.

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

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

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

Esaminiamo adesso 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

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

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

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{c(A)\ \frac{\|\delta A\|}{\|A\|}}{1 - c(A)\ \frac{\|\delta A\|}{\|A\|}} \quad . \tag{3.69}$$

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

$$\frac{\|\delta x\|}{\|x\|} \leq c(A)\ \frac{\|\delta A\|}{\|A\|} \quad , \tag{3.70}$$

che è una disuguaglianza dello stesso tipo della (3.66). 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.

Per comprendere come la soluzione delle equazioni normali possa diventare delicata (al crescere della dimensione), supponiamo di voler risolvere il problema $Ax = b$ moltiplicando a sinistra i due membri per A'. Si trova così

$$A'Ax = A'b \quad ;$$

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

$$c(A'A) = \|A'A\|\ \|(A'A)^{-1}\|\ \lambda_{\rm MAX}(A'A)/\lambda_{\rm MIN}(A'A) = c^2(A) \quad .$$

Ne segue che, anche per problemi $Ax = b$ moderatamente ben condizionati, $A'Ax = b$ 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) $A'Ax = A'b$ si perdono esattamente *il doppio* di cifre significative. Questo argomento non è esattamente calzante, dato che con i minimi quadrati si cerca di "risolvere" un sistema lineare

$$y = S\theta \quad , \tag{3.71}$$

che è sempre *incompatibile* perché il numero di equazioni n è sempre molto maggiore di p, ma serve ugualmente a spiegare qualitativamente il fenomeno e, soprattutto, a rendere ragione del successo del metodo di attacco al problema sviluppato dagli analisti numerici. Il punto fondamentale di questo metodo è *dimenticare le equazioni normali* e lavorare direttamente sul sistema (3.71).

3.5.1 La Decomposizione ai Valori Singolari (SVD)

Richiameremo 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 3.4. *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*

$$A = U\Delta V' \tag{3.72}$$

dove Δ ha la struttura quasi diagonale:

$$\Delta = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}, \qquad \Sigma = \operatorname{diag}\{\sigma_1, \ldots, \sigma_n\} \tag{3.73}$$

La matrice $U = [u_1, \ldots, u_m]$ si può costruire prendendo come colonne gli autovettori normalizzati di AA'; analogamente, $V := [v_1, \ldots, v_p]$ si può costruire prendendo come colonne gli autovettori normalizzati di $A'A$. I quadrati dei valori singolari $\{\sigma_1^2 \geq \ldots \geq \sigma_n^2\}$ sono gli autovalori non nulli di AA' (o di $A'A$).

Dimostrazione. Siano $[v_1, \ldots, v_p]$, p autovettori ortonormali di $A'A$ di modo che

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

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

$$AA'(Av_k) = \sigma_k^2(Av_k) \qquad k = 1, \ldots n$$

Si verifica che gli autovettori di AA'

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

sono ortonormali. Infatti

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

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

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

per $k, j \leq n$ e $u_k' A v_j = 0$ altrimenti. Queste relazioni sono equivalenti alla $U'AV = \Delta$ e quindi alla relazione (3.72). □

Possiamo così interpretare la formula (3.67) dicendo che *l'indice di condizionamento numerico di una matrice è il raporto tra il suo massimo e il minimo valore singolare*,

$$c(A) = \frac{\sigma_1(A)}{\sigma_n(A)} \,. \tag{3.74}$$

La SVD fornisce la descrizione più completa che si conosca della struttura di una trasformazione lineare. Dalla (3.72) si ricava, eliminando i prodotti con i blocchi nulli di Δ, la seguente *fattorizzazione a rango pieno di* A

$$A = [u_1, \ldots, u_n]\, \Sigma\, [v_1, \ldots, v_n]' := U_n \Sigma V_n' \tag{3.75}$$

dove U_n, V_n sono le sottomatrici di U, V ottenute eliminando le ultime $m - n$ e $p - n$ colonne. Notiamo che

$$U_n' U_n = I_n = V_n' V_n$$

Ricordiamo che la norma ℓ^2 di una matrice $A \in \mathbb{R}^{m \times p}$ è definita dalla relazione

$$\|A\|_2 := \sup_{\mathbf{x} \neq 0} \frac{\|Ax\|}{\|x\|}$$

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

Corollario 3.1. *Lo spazio immagine e lo spazio nullo di* A *sono dati rispettivamente da:*

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

Inoltre,

$$\|A\|_2 = \|\Sigma\|_2 = \sigma_1, \qquad \|A\|_F^2 = \|\Sigma\|_F^2 = \sigma_1^2 + \ldots + \sigma_n^2 \tag{3.76}$$

La matrice

$$A_k := \sum_{i=1}^{k} \sigma_i u_i\, v_i', \qquad k \leq n \tag{3.77}$$

è la miglior approssimante di rango k *di* A*; infatti*

$$\min_{\mathrm{rango}\,(B)=k} \|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1} \tag{3.78}$$

e inoltre

$$\min_{\mathrm{rango}\,(B)=k} \|A - B\|_F^2 = \|A - A_k\|_F^2 = \sigma_{k+1}^2 + \ldots + \sigma_n^2 \tag{3.79}$$

La dimostrazione di queste proprietà si può trovare ad esempio nel testo [12, p. 584].

Ruolo dell'ortogonalità in Analisi Numerica

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

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

sia minimo. Scriviamo $P_n(x)$ come

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

dove $s'(x) = [1 \quad x \ldots x^n]$. Imponendo il principio di ortogonalità

$$f(x) - \sum_0^n \theta_i \, x^i \perp \text{sp} \left\{ 1 \; ; \; x \ldots x^n \right\} \tag{3.80}$$

e riferendosi al prodotto scalare $\langle f, g \rangle = \int_0^1 f(x)\, g(x)\, dx$, si trovano le equazioni normali per questo problema,

$$\begin{bmatrix} \langle 1,1 \rangle & \langle 1,x \rangle & \ldots & \langle 1,x^n \rangle \\ \vdots & & & \vdots \\ \langle x^n,1 \rangle & \ldots & \ldots & \langle x^n,x^n \rangle \end{bmatrix} \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_n \end{bmatrix} = \begin{bmatrix} \langle 1,f \rangle \\ \vdots \\ \langle x^n,f \rangle \end{bmatrix} \quad . \tag{3.81}$$

A conti fatti, si trova

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

La matrice a primo membro (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 $(1 \; ; \; x \ldots x^n)$ si disponesse di polinomi indipendenti $p_0(x)\, p_1(x) \ldots p_n(x)$ tali che $\langle p_i, p_j \rangle = \delta_{ij}$, l'approssimazione

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

si potrebbe semplicemente ottenere calcolando i prodotti scalari (cfr. la (3.80)) nella

$$\left\langle f - \sum_0^n \theta_i \, p_i(x) \; ; \; p_j \right\rangle = 0 \quad , \qquad j = 0, 1, \ldots, n \quad ,$$

e ricavandone immediatamente

$$\hat{\theta}_j = \langle f, p_j \rangle \quad , \qquad j = 0, 1, \ldots, n \quad . \tag{3.82}$$

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 la serie di Fourier).

Fattorizzazione QR

Supponiamo allora di voler ricavare la stima ai M.Q. di θ partendo da n osservazioni y descritte dal modello

$$y = S\theta + \varepsilon \quad , \tag{3.83}$$

dove $\varepsilon = \varepsilon(\theta)$ è il vettore degli "errori di approssimazione" delle misure, y, attraverso 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, $\langle s_i, s_j \rangle = s_i' s_j = \delta_{ij}$ e si avrebbe $S'S = I$ per cui, in analogia all'esempio appena discusso, la stima $\hat{\theta}$ si ricaverebbe immediatamente,

$$\hat{\theta} = S'y = \begin{bmatrix} \langle s_1, y \rangle \\ \langle s_p, y \rangle \end{bmatrix} \quad . \tag{3.84}$$

(Stiamo qui considerando minimi quadrati "non pesati", ma questa semplificazione non costituisce affatto perdita di generalità). Notiamo che all'ultimo membro di (3.84) compare il vettore delle prime p coordinate di y rispetto a una base ortonormale in $\mathbb{R}^n$ del tipo $\{s_1, s_2, \ldots, s_p, \ldots\}$.

Descriviamo ora il capostipite degli algoritmi usati per risolvere problemi di M.Q.. Il suo nome è *fattorizzazione QR*. L'idea su cui è basato è semplicemente quella di *ortonormalizzare* le colonne di S.

Supponiamo di avere una matrice $n \times p$, $S = [s_1, \ldots, s_p]$, le cui colonne sono indipendenti. Come è noto, l'algoritmo di Gram-Schmidt processa sequenzialmente i vettori $\{s_1, \ldots, s_p\}$ e fornisce altrettanti vettori ortonormali $\{q_1, \ldots, q_p\}$ che sono definiti dalle relazioni

$$\begin{aligned}
&v_1 = s_1 && , \qquad q_1 := v_1 / \|v_1\| \\
&v_2 = s_2 - \langle s_2, q_1 \rangle \, q_1 && , \qquad q_2 := v_2 / \|v_2\| \\
&\quad \vdots && \qquad\quad \vdots \\
&v_k = s_k - \langle s_k q_1 \rangle \, q_1 + \ldots + \langle s_k q_{k-1} \rangle \, q_{k-1} && , \qquad q_k := v_k / \|v_k\| \quad .
\end{aligned} \tag{3.85}$$

Dal punto di vista algebrico, le (3.85) forniscono una fattorizzazione di S di struttura assai particolare. Risolviamo le (3.85) rispetto a $(s_1, \ldots, s_p)$. Si trova

$$\begin{aligned}
s_1 &= \|v_1\| \, q_1 \\
s_2 &= \langle s_2, q_1 \rangle \, q_1 + \|v_2\| \, q_2 \\
&\quad \vdots \\
s_p &= \langle s_p q_1 \rangle \, q_1 + \ldots + \langle s_p, q_{p-1} \rangle \, q_{p-1} + \|v_p\| \, q_p \quad ,
\end{aligned} \tag{3.86}$$

ovvero

$$[s_1, \ldots, s_p] = [q_1, \ldots, q_p] \begin{bmatrix} \|v_1\| & \langle s_2, q_1 \rangle & \ldots & \langle s_p, q_1 \rangle \\ 0 & \|v_2\| & & \\ \vdots & 0 & & \\ \vdots & \vdots & & \\ 0 & 0 & & \|v_p\| \end{bmatrix} \quad ; \tag{3.87}$$

questa relazione si può scrivere simbolicamente come

$$S = \bar{Q}\,\bar{R} \quad , \tag{3.88}$$

dove $\bar{Q}$ è una matrice a *colonne ortonormali*, cioè $\bar{Q}'\bar{Q} = I$ $(p \times p)$ 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

$$Q = \left[\bar{Q} \mid q_{p+1} \ldots q_n\right]$$

$$R = \begin{bmatrix} \bar{R} \\ 0 \end{bmatrix} \quad ,$$

si vede che S si può anche scrivere come

$$S = QR \quad , \tag{3.89}$$

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 (3.85) forniscono un algoritmo ricorsivo per il calcolo di $\bar{Q}$ ed $\bar{R}$. Per ottenere la (3.89) basta aggiungere a $\bar{Q}$ $n - p$ colonne ortonormali (vedremo in seguito che questa operazione si può evitare).

Se moltiplichiamo i due membri della (3.83) per Q' si ottiene allora

$$Q'y = Q\,S\theta + Q'\varepsilon \quad , \tag{3.90}$$

ovvero

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \bar{R} \\ 0 \end{bmatrix} \theta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \quad , \tag{3.91}$$

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

Notiamo subito che

$$\operatorname{span}\{q_1 \ldots q_p\} = \operatorname{span}\{s_1 \ldots s_p\} = \mathcal{S}$$

e pertanto

$$\operatorname{span}\{q_{p+1} \ldots q_n\} = \mathcal{S}^{\perp} \quad .$$

Ne deriva che $\begin{bmatrix} y_1 \\ 0 \end{bmatrix}$ è la proiezione di y su $\mathcal{S}$ (espressa nelle coordinate $\{q_i\}$) e $\begin{bmatrix} 0 \\ y_2 \end{bmatrix}$ è la proiezione di y sul sottospazio $\mathcal{S}^{\perp}$ 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 (3.91) si vede, data l'ortogonalità di Q', che

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

$$= \left\| \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} - \begin{bmatrix} \bar{R}\theta \\ 0 \end{bmatrix} \right\|^2 = \|y_1 - \bar{R}\theta\|^2 + \|y_2\|^2 \quad .$$

Da questa relazione segue immediatamente che

1) $\hat{\theta}$ è soluzione di

$$\bar{R}\theta = y_1 \quad , \tag{3.92}$$

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

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

$$\|\hat{\varepsilon}\|^2 = \|y_2\|^2 \quad . \tag{3.93}$$

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 (3.92) in cui $\bar{R}$ è triangolare.

Notiamo che Q non entra esplicitamente nelle formule (3.92) e (3.93).
Se il modello (3.83) rappresenta delle misure affette da errore ε su cui si ha conoscenza probabilistica a priori, del tipo

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

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

$$\mathrm{Var}\ \hat{\theta} = \sigma^2 \left[S'S\right]^{-1} \quad .$$

Usando la fattorizzazione QR si vede subito che

$$\mathrm{Var}\ \hat{\theta} = \sigma^2 \left(\bar{R}'\bar{R}\right)^{-1} \quad . \tag{3.94}$$

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

$$[S \mid y] \tag{3.95}$$

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

$$\left[\begin{array}{c|c} \bar{R} & y_1 \\ \hline 0 & y_2 \end{array}\right] \quad . \tag{3.96}$$

Giunti a questo punto, ovviamente il grosso del lavoro è stato fatto perché rimane solo da risolvere il sistema (3.92) 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 (3.95) alla forma (3.96).

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 Lawson-Hanson [16]. Qui sotto ne descriveremo uno particolarmente semplice e di uso generale.

Definizione 3.3. *Una matrice di riflessione elementare (o matrice di Householder) è una matrice $n \times n$ del tipo*

$$H(v) = I - 2\,\frac{vv'}{\|v\|^2} \quad , \tag{3.97}$$

dove $v \in \mathbb{R}^n$.

Si verifica subito che

1) $H(v)$ è simmetrica,

2) $H(v)$ è ortogonale,

3) $H^2(v) = I$.

Il nome deriva dal fatto che per un qualunque $x \in \mathbb{R}^n$ l'immagine $H(v)\,x$ di x è il vettore "riflesso" di x rispetto all'iperpiano di $\mathbb{R}^n$ la cui normale è il vettore v. In effetti, posto $u = v/\|v\|$,

$$H(v)\,x = x - 2\,\langle u, x\rangle\,u$$

e si ha la situazione descritta in Figura 5.1.

Le matrici di riflessione possono essere utilizzate per trasformare un generico vettore x di $\mathbb{R}^n$ in un multiplo scalare del vettore $e_1 = [1, 0, \ldots, 0]'$. Dato che $H(v)$ è ortogonale, dovrà necessariamente essere

$$H(v)\,x = I\,\|x\|\,e_1 \quad . \tag{3.98}$$

La scelta di v per arrivare alla (3.98) è suggerita dalla geometria della trasformazione di Figura 5.2.

Figura 5.2

Dato x, si tratta di trovare il piano bisettore $\mathcal{S}$ rispetto al quale $-\|x\|\, e_1$ appare come l'immagine riflessa di x.

Chiaramente la normale, v, dovrà appartenere alla bisettrice dell'angolo piano trasformato dai vettori e_1 e x. Pertanto

$$v = x + \|x\|\, e_1 \tag{3.99}$$

e con questa scelta si può verificare algebricamente che, in effetti,

$$H(v)\, x = -\|x\|\, e_1 \quad .$$

Mediante l'uso di matrici di Householder si può triangolarizzare S in $p-1$ passi e ridursi alla tabella (3.96) partendo dalla tabella (3.95). Denotiamo quest'ultima con il simbolo

$$S_1 = [s_1, \ldots, s_p y] \quad .$$

Prendendo

$$v_1 = s_1 + \|s_1\|\, e_1 \quad ,$$

si ha

$$H(v)\; S_1 = \left[\begin{array}{c|ccc} -\|s_1\| & s'_{12} & \ldots & s'_{1p} \quad y'_1 \\ \hline 0 & s'_{22} & & s'_{2p} \quad y'_2 \\ \vdots & \vdots & & \vdots \quad \vdots \\ 0 & s'_{n2} & & s'_{np} \quad y'_n \end{array}\right] \quad . \tag{3.100}$$

Chiamiamo ora S_2 la matrice $(n-1)\times(p-1)$ nel blocco inferiore a destra e sia s'_2 la sua prima colonna. (L'apice qui non ha ovviamente il significato di trasposizione). Definiamo

$$v_2 := s'_2 + \|s'_2\|\, e'_1 \quad , \qquad v_2 \in \mathbb{R}^{n-1} \quad ,$$

dove e'_1 è il primo vettore della base canonica in $\mathbb{R}^{n-1}$. Evidentemente

$$H(v_2)\; S_2 = \left[\begin{array}{c|ccc} -\|s'_2\| & s''_{23} & \ldots & s''_{2p} \quad y''_2 \\ \hline 0 & s''_{33} & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & s''_{n3} & & s''_{np} \quad y''_n \end{array}\right] \quad . \tag{3.101}$$

È chiaro che trasformando in modo analogo la prima colonna della matrice S_3 $(n-2)\times(p-2)$ e via via $S_4, \ldots, S_{p-1}$ si arriva a una struttura triangolare del tipo (3.96). Si può anche immaginare di operare la trasformazione (3.101) mediante una matrice che è $n\times n$ anziché $(n-1)\times(n-1)$ come la $H(v_2)$. Se poniamo infatti

$$\tilde{H}(v_2) = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & & & \\ \vdots & & & \\ 0 & & & H(v_2) \end{bmatrix}$$

si controlla subito che

$$\tilde{H}(v_2)\,\Big(H(v_1)\,S_1\Big) = \begin{bmatrix} -\|s_1\| & s'_{12} & \dots & s'_{1p} & y'_1 \\ 0 & -\|s_2\| & & s''_{2p} & y''_2 \\ \vdots & 0 & & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & s''_{np} & y''_n \end{bmatrix} \quad ,$$

ovvero la prima riga e la prima colonna di $H(v_1)\,S_1$ *rimangono immutate*. Chiaramente $\tilde{H}(v_2)$ è ancora ortogonale. In questo modo, se si desidera avere a disposizione Q la si può ricavare immediatamente come

$$Q = \tilde{H}(v_{p-1})\;\tilde{H}(v_{p-2})\;\dots\;\tilde{H}(v_2)\;H(v_1) \quad . \tag{3.102}$$

Programmi per l'implementazione di questo algoritmo di fattorizzazione (detto di Householder) si possono trovare nel testo già citato di Lawson-Hanson oppure nella libreria MATLAB [20].

Capitolo 4

PROBLEMI DI REGRESSIONE MULTIPLA

4.1 Stima della complessità di un modello lineare

In molte situazioni concrete il numero di parametri, p, che caratterizza il 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. In termini di modellistica, aumentare p può ad esempio corrispondere all'aggiungere altri modi esponenziali nella descrizione della risposta libera di un sistema lineare, oppure nel considerare l'effetto di variabili di regressione via via meno "importanti" ecc...

Se la numerosità campionaria è fissa (cosa che da ora in avanti supporremo), è abbastanza ovvio che all'aumentare di p si ottiene una descrizione sempre migliore dei dati, nel senso che l'errore quadratico medio

$$\hat{\sigma}^2(y) = \frac{1}{N} \|y - S\hat{\theta}(y)\|^2_{\mathbb{R}^{-1}}$$

diminuisce all'aumentare di p fino a diventare addirittura zero nel caso limite $p = N$. È però abbastanza intuibile che, a parità di misure disponibili, la qualità delle stime ottenute, misurata ad esempio in termini di varianza dei parametri stimati si deteriori 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 renderebbe inservibile il modello (il lettore è invitato a meditare sul fatto che il modello verrà usato poi per descrivere dati *diversi* da quelli usati in fase di stima).

In pratica è quindi necessario procedere per tentativi successivi, aumentando p fino a che il compromesso raggiunto tra bontà del "fit" e dispersione della stima sembra accettabile. Nel contesto della statistica classica "Fisheriana" , il problema della scelta ottima di p può essere visto come 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 di valori plausibili $[p_{\min}, p_{\max}]$ che si può pensare assegnato a priori. Per arrivare a dei criteri di scelta chiari e non troppo complicati, noi inizialmente formuleremo il problema in termini di scelta tra due sole alternative possibili.

Consideriamo due modelli lineari Gaussiani in forma standard

$$\begin{array}{lrcll} M_1: & \mathbf{y} & = & S_1\theta_1 + \boldsymbol{\epsilon} & \theta_1 \in \mathbb{R}^p \\ M_2: & \mathbf{y} & = & [S_1\, S_2] \begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix} + \boldsymbol{\epsilon} & \theta_2 \in \mathbb{R}^k. \end{array} \tag{4.1}$$

In entrambi i casi $\boldsymbol{\epsilon}$ è il vettore aleatorio $\sigma\mathbf{w}$ che assumiamo a media nulla e varianza $\sigma^2 I$ (matrice identità $N \times N$).

Nel modello più "semplice" M_1, supporremo, come sempre, che $\operatorname{rango} S_1 = p$. Nel modello "complicato" M_2, si può supporre senza perdita di generalità che la matrice $S_2 \in \mathbb{R}^{N\times k}$ sia tale che

$$\operatorname{rango}[S_1\, S_2] = p + k \quad . \tag{4.2}$$

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.

Quanto diremo può facilmente essere esteso al caso di varianza di $\mathbf{w}$ diversa dall'identità. Le formule relative al caso generale si possono ricavare da quelle che daremo qui di seguito sostituendo a $\mathbf{y}$ il vettore $L^{-1}\mathbf{y}$ e ad S la matrice $L^{-1}S$, dove L è il fattore (sinistro) di Cholesky della matrice varianza di $\mathbf{w}$.

Ovviamente i due modelli in (4.1) definiscono due diverse famiglie parametriche di misure di probabilità sullo spazio campionario.

Problema 4.1. *Sulla base di una osservazione $\mathbf{y} = y$ dare una regola di decisione "razionale" che scelga quale delle due famiglie ha generato i dati.*

In termini tecnici questo è un problema di verifica di ipotesi "composte". Si dimostra in letteratura [17, 24], che le funzioni di decisione ottimali (nel senso che massimizzano la cosiddetta "potenza" del test) si ottengono considerando il cosiddetto *rapporto di massima verosimigliaza* la cui costruzione richiede preliminarmente la stima (di M.V.) dei parametri dei due modelli.

Osservazione 4.1. Notiamo che il problema può anche essere 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 (4.1) sono da interpretare solo come "approssimazioni" usate per descrivere i dati $\mathbf{y}$. 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. Questa posizione del problema che verrà ripresa in modo più preciso più avanti, conduce alle soluzioni moderne del problema della stima dell'ordine.

4.2 Regressione lineare a stadi

In questo paragrafo 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 $\mathcal{S}$ lo spazio colonne della matrice $S := [S_1\, S_2]$ e con θ il parametro $p+k$ dimensionale $[\theta_1'\, \theta_2']'$ che compare nella (4.1). Naturalmente la stima ai M.Q. (di Markov) di θ è definita dalle solite formule,

$$\begin{aligned} \hat{\theta}(y) &= (S'S)^{-1}\, S'y \\ \text{Var}\ \hat{\theta} &= \sigma^2 (S'S)^{-1} \quad , \end{aligned}$$

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 (4.2) $\mathcal{S}$ si può decomporre in somma diretta

$$\text{span}\,[S] = \text{span}\,[S_1\, S_1] = \mathcal{S}_1 \oplus \mathcal{S}_2 = \text{span}\,[S_1] \oplus \text{span}\,[S_2] \tag{4.3}$$

e questa decomposizione può essere resa *ortogonale* se si introducono i proiettori

$$\begin{array}{lll} P_1 : \mathbb{R}^n \to \mathcal{S}_1 & , & P_1 = S_1 (S_1'\, S_1)^{-1}\, S_1' \quad , \\ R_1 = P_1^{\perp} : \mathbb{R}^n \to \mathcal{S}_1^{\perp} & , & R_1 = I - S_1 (S_1'\, S_1)^{-1}\, S_1' \quad . \end{array} \tag{4.4}$$

usando i quali, in effetti si ha

$$\text{span}\,[S] = \text{span}\,[S_1] \overset{\perp}{\oplus} \text{span}\,[R_1\, S_2]$$

(il simbolo $\overset{\perp}{\oplus}$ sta per somma diretta ortogonale), dato che ovviamente

$$S_2 = P_1 S_2 + R_1 S_2$$

e le colonne di $P_1 S_2$ stanno per definizione in $\mathcal{S}_1$. Pertanto l'ultimo addendo della (4.3) può venire sostituito da $\text{span}\,[R_1\, S_2]$. Sia ora $\hat{y}$ la proiezione ortogonale di y sullo spazio colonne, $\mathcal{S}$, della matrice S. Per l'indipendenza lineare delle colonne di S_1 e S_2 si dovrà poter esprimere in modo unico $\hat{y}$ come $\hat{y}$ nella forma

$$\hat{y} = S_1\, \hat{\theta}_1 + S_2\, \hat{\theta}_2 \quad , \tag{4.5}$$

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 che noi cerchiamo.

Per il principio di ortogonalità dovrà essere $y - \hat{y} \perp \mathcal{S}$ e quindi anche, separatamente,

$$y - \hat{y} \perp \mathcal{S}_1 \quad , \qquad y - \hat{y} \perp R_1\, \mathcal{S}_2 \quad ,$$

che si riscrivono

$$S_1'(y - S_1\hat{\theta}_1 - S_2\hat{\theta}_2) = 0 \quad , \tag{4.6}$$

$$S_2'R_1(y - S_1\hat{\theta}_1 - S_2\hat{\theta}_2) = 0 \quad . \tag{4.7}$$

Queste formule forniscono subito

$$\hat{\theta}_1 = (S_1'S_1)^{-1}\, S_1' \left[y - S_2\hat{\theta}_2 \right] \quad , \tag{4.8}$$

$$\hat{\theta}_2 = (S_2'R_1S_2)^{-1}\, S_2'R_1\, y \quad . \tag{4.9}$$

La prova che $S_2'R_1S_2$ è invertibile si ottiene facilmente se si tiene presente che R_1 è un proiettore. In effetti

$$a'S_2'R_1S_2a = 0 \Rightarrow a'S_2'R_1R_1S_2a = \|R_1S_2a\|^2 = 0$$

e pertanto S_2a deve stare nello spazio nullo di $R_1 = P_1^{\perp}$. Dato che $\mathrm{Ker}\,(P_1^{\perp}) = \mathrm{Im}\,(P_1) = \mathcal{S}_1 = \mathrm{span}\,[S_1]$, segue che $S_2a \in \mathrm{span}\,[S_1]$, ma questo può accadere solo se $a = 0$, dato che le colonne di S_1 ed S_2 sono indipendenti.

Proiezioni oblique

Nella decomposizione (4.5) 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 $\mathcal{S}_1$ lungo $\mathcal{S}_2$ e di y su S_2 lungo $\mathcal{S}_1$.

Dalla formula (4.9) si vede in particolare che i coefficienti, $\hat{\theta}_2$, che esprimono la proiezione obliqua di y su $\mathcal{S}_2$ lungo $\mathcal{S}_1$, si possono calcolare proiettando *ortogonalmente* $y - P_1y = R_1y$ su $(I - P_1)\mathcal{S}_2 = R_1\mathcal{S}_2$ come se si trattasse di un problema di minimi quadrati ordinari. La matrice di proiezione obliqua su S_2 lungo $\mathcal{S}_1$ ha così la rappresentazione

$$P_{2\|1} := S_2(S_2'R_1S_2)^{-1}\, S_2'R_1 \tag{4.10}$$

usando la quale si controlla facilmente che in effetti $P_{2\|1}^2 = P_{2\|1}$, mentre

$$P_{2\|1}'R_1 = R_1P_{2\|1}\,.$$

la quale, visto che R_1 è un proiettore ortogonale e quindi $R_1 = R_1'$, si può riscrivere come $(R_1P_{2\|1})' = P_{2\|1}'R_1' = R_1P_{2\|1}$, i.e. $R_1P_{2\|1}$ è simmetrica (e idempotente) e quindi è essa stessa un *proiettore ortogonale* che, per forza di cose, deve proiettare sul sottospazio $R_1\mathcal{S}_2$, che è il complemento ortogonale di $\mathcal{S}_1$ in $\mathcal{S}$. Infatti:

Proposizione 4.1. *Se P denota la matrice proiezione ortogonale sullo spazio $\mathcal{S}$ e P_1 quella sul sottospazio $\mathcal{S}_1 \subset \mathcal{S}$, $P - P_1$ proietta sul complemento ortogonale $\mathcal{S} \cap \mathcal{S}_1^{\perp}$ e si ha*

$$P - P_1 = R_1P_{2\|1} \tag{4.11}$$

dove $P_{2\|1}$ è il proiettore obliquo definito in (4.10)*.*

Dimostrazione. Usando le formule (4.8) (4.9) si ottiene

$$\begin{aligned} Py = S_1\hat{\theta}_1 + S_2\hat{\theta}_2 &= (S_1'S_1)^{-1}\ S_1'\,y + \left[I - S_1(S_1'S_1)^{-1}\ S_1'\right] S_2(S_2'R_1S_2)^{-1}\ S_2'R_1\,y \\ &= (P_1 + R_1P_{2\|1})\,y \end{aligned}$$

per cui effettivamente si ha $P - P_1 = R_1P_{2\|1}$. La decomposizione $P = P_1 + R_1P_{2\|1}$ è ovviamente ortogonale, stante che $P_1^\top(P - P_1) = P_1R_1P_{2\|1} = 0$. Notiamo che un'affermazione equivalente è la $\mathcal{S} = P_1\mathcal{S} \oplus \mathcal{S} \cap \mathcal{S}_1^\perp$. □

Problema 4.2. 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 $\mathcal{S}_1$ lungo $\mathcal{S}_2$ e arrivare ad una rappresentazione esplicita della decomposizione (4.5), del tipo

$$y = P_{1\|2}y + P_{2\|1}y = S_1(S_1'R_2S_1)^{-1}\ S_1'R_2\,y + S_2(S_2'R_1S_2)^{-1}\ S_2'R_1\,y \qquad (4.12)$$

dove R_2 ha un significato duale a R_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)

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 (4.8) può essere riscritta come

$$\hat{\theta}_1 = \bar{\theta}_1 - (S_1'S_1)^{-1}\ S_1'S_2\hat{\theta}_2 \quad . \qquad (4.13)$$

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 terminie di correzione dovuto all'introduzioine del parametro ulteriore θ_2.

Concentriamoci ora sul calcolo delle varianze degli stimatori. Introduciamo allo

scopo le seguenti notazioni:

$$\bar{\Sigma}_1 := [S_1' S_1]^{-1}$$
$$A_1 := [S_1' S_1]^{-1} S_1'$$
$$\Sigma_2 := [S_2' R_1 S_2]^{-1} \quad ;$$

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 "vero" rispetto a cui si calcola l'aspettazione (e quindi la varianza).

Proposizione 4.2. *Siano $\hat{\theta}_1(\mathbf{y})$ e $\hat{\theta}_2(\mathbf{y})$ gli stimatori di Markov definiti dalle formule* (4.8) *e* (4.9)*. Si ha allora:*

$$\mathrm{Var}_\theta \left\{ \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} \right\} = \sigma^2 \begin{bmatrix} \bar{\Sigma}_1 + A_1 S_2 \Sigma_2 S_2' A_1' & -A_1 S_2 \Sigma_2 \\ -\Sigma_2 S_2' A_1' & \Sigma_2 \end{bmatrix} \quad . \tag{4.14}$$

Dimostrazione. Incominciamo col dimostrare che $\mathrm{Var}_\theta\,[\hat{\theta}_2] = \sigma^2\, \Sigma_2$. Dalla (4.9) si ha

$$\mathrm{Var}_\theta\,[\hat{\theta}_2] = \Sigma_2 S_2' R_1\, \mathrm{Var}_\theta[\mathbf{y}] R_1 S_2 \Sigma_2 = \sigma^2\, \Sigma_2 S_2' R_1 S_2 \Sigma_2 = \sigma^2\, \Sigma_2 \quad ,$$

dato che $\mathrm{Var}_\theta\,[\mathbf{y}] = \sigma^2\, I$ ed R_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:

$$\mathrm{Cov}_\theta \left[\bar{\theta}_1(\mathbf{y}), \hat{\theta}_2(\mathbf{y})\right] = \bar{\Sigma}_1 S_1'\ \mathrm{Var}_\theta\,[\mathbf{y}]\ R_1 S_2 \Sigma_2 = \sigma^2\, \bar{\Sigma}_1 S_1' R_1 S_2 \Sigma_2 = 0 \quad ,$$

perchè $S_1' R_1 = R_1 S_1 = 0$.

Usando ora la (4.13) 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

$$\mathrm{Var}_\theta \left[\bar{\theta}_1(\mathbf{y}) - A_1 S_2 \hat{\theta}_2(\mathbf{y})\right] = \mathrm{Var}_\theta \left[\bar{\theta}_1(\mathbf{y})\right] + A_1 S_2\ \mathrm{Var}_\theta \left[\hat{\theta}_2(\mathbf{y})\right]\ S_2' A_1'$$
$$= \sigma^2 \left[\bar{\Sigma}_1 + A_1 S_2 \Sigma_2 S_2' A_1'\right] \quad .$$

che conclude la dimostrazione della formula (4.14). □

Osservazione 4.2. La formula (4.14) descrive l'effetto dell'aumento del numero di parametri nel modello sulla varianza delle stime e sull' errore quadratico medio residuo. In particolare (4.14) mostra che la "nuova" stima $\hat{\theta}_1$ di θ_1 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

$$\Sigma_1 = \bar{\Sigma}_1 + A_1 S_2\ \Sigma_2\ S_2' A_1'$$

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 accade che le colonne di S_2 sono *ortogonali* a $\mathcal{S}_1$, ovvero se accade che

$$S_1' S_2 = 0 \qquad (S_2' S_1 = 0)$$

le formule si semplificano drammaticamente (dato che $R_1 S_2 = S_2$) e i due stimatori $\hat{\theta}_1$ e $\hat{\theta}_2$ si possono calcolare indipendentemente l'uno dall'altro con le solita formule,

$$\hat{\theta}_i(\mathbf{y}) = (S_i' S_i)^{-1} S_i' \mathbf{y} \quad , \qquad i = 1, 2 \quad .$$

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 nuovoparametro $\beta := R\theta$ 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.. □

Supponiamo allora di voler confrontare l'errore quadratico residuo che si commette descrivendo i dati osservati y mediante un modello delle due classi M_1 ed M_2 definite in (4.1).

Indichiamo con $\bar{\varepsilon}_1(\mathbf{y}) := y - S_1 \bar{\theta}_1(\mathbf{y}) = (I - P_1)\mathbf{y}$ il vettore dei residui usando il modello a p parametri e con $\hat{\varepsilon} = \mathbf{y} - S_1 \hat{\theta}_1(\mathbf{y}) - S_2 \hat{\theta}_2(\mathbf{y}) = (I - P)\mathbf{y}$ quello relativo al modello aumentato e supponiamo inizialmente di non sapere chi sia il modello "vero". Dato che $(I - P)$ e $(P - P_1)$ proiettano su spazi ortogonali (infatti $I - P$ proietta sul complementare $\mathcal{S}^{\perp}$ mentre $(P - P_1)$ proietta sul sottospazio $\mathcal{S} \cap \mathcal{S}_1^{\perp}$, (proposizione 4.1)), si ha,

$$\begin{aligned} \|\bar{\varepsilon}_1\|^2 &= \| \left[(I - P) + (P - P_1)\right] \mathbf{y}\|^2 = \|\hat{\varepsilon}\|^2 + \|(P - P_1)\mathbf{y}\|^2 \\ &= \|\hat{\varepsilon}\|^2 + \|R_1 P_{2\|1} \mathbf{y}\|^2 \, . \end{aligned} \tag{4.15}$$

L'importanza del termine $\|R_1 P_{2\|1}\mathbf{y}\|^2$ all'ultimo membro dipende da quale classe di modelli ha effettivamente generato i dati. Per maggior evidenza possiamo esprimerlo, usando la (4.10), in funzione dello stimatore $\hat{\theta}_2(\mathbf{y})$ come $\hat{\theta}_2(\mathbf{y})' \Sigma_2^{-1} \hat{\theta}_2(\mathbf{y})$, per cui, nel caso in cui il modello che ha effettivamente generato i dati fosse M_1, la stima $\hat{\theta}_2(\mathbf{y})$ descriverebbe solo rumore additivo e si può intuitivamente dedurne che il termine correttivo $\|\bar{\varepsilon}_1\|^2 - \|\hat{\varepsilon}\|^2$ nella (4.15) risulterà mediamente piccolo.

Teorema 4.1. *Assumendo che il modello "vero" sia M_1 (ipotesi H_0), la somma dei quadrati dei residui, $\|\bar{\varepsilon}_1\|^2 = \|R_1\mathbf{y}\|^2$ è uguale a $\|R_1\boldsymbol{\epsilon}\|^2$, e si può esprimere nella forma*

$$\|\bar{\varepsilon}_1\|^2 = \|\hat{\varepsilon}\|^2 + \hat{\theta}_2(\mathbf{y})'\Sigma_2^{-1}\,\hat{\theta}_2(\mathbf{y}) \tag{4.16}$$

dove il termine $\|\hat{\varepsilon}\|^2$ è la somma dei quadrati dei residui che si ottiene usando il modello aumentato M_2 per descrivere i dati. Nell astessa ipotesi (H_0), i due addendi al secondo membro sono indipendenti e hanno distribuzioni di probabilità di tipo χ^2, rispettivamente,

$$\frac{\|\hat{\varepsilon}\|^2}{\sigma^2} \sim \chi^2(N-p-k) \tag{4.17}$$

e

$$\frac{\hat{\theta}_2(\mathbf{y})'\Sigma_2^{-1}\,\hat{\theta}_2(\mathbf{y})}{\sigma^2} \sim \chi^2(k)\,. \tag{4.18}$$

Dimostrazione. La (4.16) era già stata derivata in (4.15). Sostituendo nell'ultima delle (4.15) l'espressione del modello "vero" M_1 si riconosce immediatamente che $R_1P_{2\|1}\mathbf{y} = P'_{2\|1}R_1(S_1\theta_1+\boldsymbol{\epsilon}) = P'_{2\|1}R_1\boldsymbol{\epsilon}$, dove $P'_{2\|1}R_1 = R'_1P_{2\|1}$ è il proiettore ortogonale su $\mathcal{S}\cap\mathcal{S}_1^{\perp}$ che ha per ipotesi rango k. D'altro canto $\hat{\varepsilon} = (I-P)\boldsymbol{\epsilon}$ e quindi $\boldsymbol{\epsilon}'(I-P)'R_1P_{2\|1}\boldsymbol{\epsilon} = 0$, dato che $(I-P)$ proietta sul complemento ortogonale di $\mathcal{S}$.

Ora, come è ben noto, $\frac{1}{\sigma^2}\|\bar{\varepsilon}_1\|^2 \sim \chi^2(N-p)$ e $\frac{1}{\sigma^2}\hat{\theta}_2(\mathbf{y})'\Sigma_2^{-1}\,\hat{\theta}_2(\mathbf{y}) \sim \chi^2(k)$ (Proposizione 3.5). Dato che i due addendi sono indipendenti $\frac{1}{\sigma^2}\|\hat{\varepsilon}\|^2$ deve necessariamente avere distribuzione χ^2 (Teorema 3.1) e il numero di gradi di libertà dev'essere $N-p-k$. □

Per una dimostrazione alternativa si può vedere il [25, p. 73]. Nei testi di statistica si dimostra che date due variabili $\mathbf{z}_1 \sim \chi^2(n_1)$ e $\mathbf{z}_2 \sim \chi^2(n_2)$ indipendenti, il rapporto $\dfrac{\mathbf{z}_1/n_1}{\mathbf{z}_2/n_2}$ ha una distribuzione di probabilità notevole, nota col nome di *distribuzione F* che è tabulata. Nel nostro caso, se il modello vero che ha generato i dati è M_1, il rapporto

$$\varphi(\mathbf{y}) := \frac{N-p-k}{k}\,\frac{\hat{\theta}_2(\mathbf{y})\Sigma_2^{-1}\,\hat{\theta}_2(\mathbf{y})}{\|\hat{\varepsilon}\|^2} = \frac{N-p-k}{k}\,\frac{\|\bar{\varepsilon}_1\|^2-\|\hat{\varepsilon}\|^2}{\|\hat{\varepsilon}\|^2} \tag{4.19}$$

che misura il miglioramento relativo nella descrizione dei dati usando un modello a $p+k$ parametri rispetto alla descrizione con p parametri ha quindi una distribuzione nota, di tipo F.

In effetti, dalle distribuzioni dell'enunciato del teorema 4.1 e si ha, esattamente

$$\varphi(\mathbf{y}) \sim F(k, N-p-k)$$

dove i due argomenti denotano i *gradi di libertà* della distribuzione. Normalmente $N-p$ è molto più grande di k e la distribuzione F si può approssimare molto bene con una χ^2 a k gradi di libertà, per cui in realtà si può usare la relazione,

$$\varphi(\mathbf{y}) \sim \chi^2(k) \qquad \text{(se vale } M_1\text{)} \tag{4.20}$$

Fissata allora la probabilità di commettere un *errore di "prima specie*

$$\alpha := P\{\text{scegliere } M_2 \text{ quando è vero } M_1\}$$

e detto x_α il valore dell'ascissa per cui

$$P_{\chi^2(k)}\{x > x_\alpha\} = 1 - \alpha$$

che si trova sulle tabelle della distribuzione $\chi^2(k)$, si va a vedere se il valore campionario della statistica (4.20) assume valori minori o uguali a x_α. In questo caso si "accetta " l'ipotesi M_1, con la sicurezza del $1 - \alpha\%$ di aver scelto correttamente, beninteso nel caso in cui i dati siano stati davvero generati da M_1. La distribuzione della statistica (4.19) nel caso che il modello vero sia M_2 è complicata e in pratica la probabilità di commettere un "errore di seconda specie"

$$\beta := P\{\text{scegliere } M_1 \text{ quando è vero } M_2\}$$

si stima con simulazioni Monte Carlo. La probabilità $1 - \beta$, di scegliere il modello giusto quando è vero M_2, si chiama anche *potenza del test*.

4.3 Stima della dimensione del modello col criterio FPE

Come abbiamo già osservato 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 (o "calibrazione", come qualche volta è chiamata) ma occorre in realtà valutare la bontà con cui il modello stimato riesce a descrivere dati *futuri*, non usati per l'identificazione del modello. Supponiamo allora di avere a disposizione un vettore di osservazioni $\mathbf{y} := [\mathbf{y}_1' \ \mathbf{y}_2']'$ di dimensione $2N$ e di usare i primi N dati $\mathbf{y}_1$ per l'identificazione di un generico modello lineare standard di dimensione p. Risolviamo così il problema di descrivere i dati $\mathbf{y}_1$ mediante il modello statistico lineare

$$\mathbf{y}_1 = S\theta + \boldsymbol{\epsilon}_1\,, \qquad \mathrm{Var}\,[\boldsymbol{\epsilon}_1] = \sigma^2 I_N \tag{4.21}$$

ottenedo, come è ben noto, il classico stimatore $\hat{\theta}(\mathbf{y}_1) = [S'S]^{-1}S'\mathbf{y}_1$. 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. Naturalmente perchè questa operazione abba 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 in modo equivalente 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. In particolare qui supporremo che le due componenti del vettore $[\mathbf{y}_1' \ \mathbf{y}_2']'$ abbiano stessa media (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 l'errore (in forma vettoriale) di predizione dei dati futuri

$$\boldsymbol{\varepsilon} := \mathbf{y}_2 - S\hat{\theta}(\mathbf{y}_1) \tag{4.22}$$

e calcoliamone la varianza

$$\mathrm{Var}\,[\boldsymbol{\varepsilon}] = \sigma^2 I_N + S[S'S]^{-1}S'\sigma^2 I_N S[S'S]^{-1}S' = \sigma^2\left[I_N + S[S'S]^{-1}S'\right]$$

di modo che

$$\begin{aligned}\frac{1}{N}\,\mathrm{var}\,[\boldsymbol{\varepsilon}] \quad &= \sigma^2\frac{1}{N}\,\mathrm{Tr}\,\left\{I_N + S[S'S]^{-1}S'\right\} = \sigma^2\left\{1 + \mathrm{Tr}\,([S'S]^{-1}S'S)\right\} \\ &= \sigma^2(1+\frac{p}{N})\end{aligned} \tag{4.23}$$

Dalla quale si vede che la varianza scalare dell'errore di predizione dipende linearmente da p. Per usare questo risultato per la stima della dimensione del modello, dobbiamo sostituire alla varianza σ^2, che è un parametro incognito, una sua stima, naturalmente anch'essa basata su un modello a p parametri. Usando lo stimatore corretto della varianza discusso in (3.53)

$$\frac{N}{N-p}\,\hat{\sigma}^2 = \frac{1}{N-p}\|\mathbf{y}_1 - S\hat{\theta}(\mathbf{y}_1)\|^2 = \frac{1}{N-p}\|\hat{\boldsymbol{\epsilon}}_p\|^2$$

dove $\hat{\boldsymbol{\epsilon}}_p$ è il residuo di stima nel modello a p parametri, si arriva così a definire l'indice

$$FPE(p) := \frac{1}{N}\|\hat{\boldsymbol{\epsilon}}_p\|^2\frac{(1+\frac{p}{N})}{(1-\frac{p}{N})} := \hat{\sigma}_p^2\frac{(1+\frac{p}{N})}{(1-\frac{p}{N})} \tag{4.24}$$

che si chiama **errore "finale" di predizione** basato sul modello di dimensione p.

La stima dell'ordine del modello può essere basata sulla minimizzazione di questo indice. Naturalmente per effettuare la minimizzazione 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 ricorsivi del tipo di quello illustrato nel seguente paragrafo.

4.4 Un algoritmo di regressione lineare a stadi

Le formule di aggiornamento della stima (4.13) forniscono un algoritmo di calcolo "a stadi" ("step wise least squares") basato sull'*introduzione sequenziale* (una alla volta) *delle colonne di S nel modello* e sull'aggiornamento della stima corrispondente ad aggiungere ad ogni ciclo una sola nuova variabile di regressione (un solo parametro). Fortunatamente, per questo problema è possibile costruire *algoritmi ricorsivi* (il termine "ricorsivo" è ora relativo all'indice p), che aggiornano ad ogni passo le stime calcolate al passo precedente. Per ovvie ragioni (la dimensione del problema aumenta al crescere di p), non ci si può però aspettare che la complessità di calcolo rimanga costante come avviene per il filtro di Kalman.

Supponiamo di possedere lo stimatore $\theta^k = [\bar{\theta}_1, \ldots, \bar{\theta}_k]'$ ottenuto modellando i dati con il modello lineare a k parametri (in cui abbiamo denotato σw con il simbolo ε)

$$y = S_k\,\theta + \varepsilon \quad , \qquad S_k \in \mathbb{R}^{n\times k} \quad ,$$

e di introdurre una *nuova colonna* (linearmente indipendente), s_{k+1}, in S. Il modello diventa allora a $k+1$ parametri,

$$y = S_{k+1}\,\theta + \varepsilon \quad , \tag{4.25}$$

con

$$S_{k+1} = [S_k \; s_{k+1}] \quad .$$

Usando le formule di aggiornamento per gli stimatori, si trova

$$\hat{\theta}_{k+1} = \frac{1}{s_{k+1} \, R_k \, s_{k+1}} \; s'_{k+1} R_k y = \frac{1}{s'_{k+1} \, R_k \, s_{k+1}} \; s'_{k+1}[y - S'_k \bar{\theta}^k] \tag{4.26}$$

e inoltre

$$\hat{\theta}^k = (S'_k S_k)^{-1} \; S'_k[y - s_{k+1}\hat{\theta}_{k+1}] = \bar{\theta}^k - (S'_k S_k)^{-1} \; S'_k s_{k+1} \hat{\theta}_{k+1} \quad , \tag{4.27}$$

dove si sono usate le notazioni $\hat{\theta}_{k+1}$ e $\hat{\theta}^k$ per indicare lo stimatore di θ_{k+1} e di $[\theta_1, \ldots, \theta_k]'$ relativi al modello aumentato (4.25) ed R_k ha il solito significato di proiettore sul complemento ortogonale dello spazio colonne di S_k,

$$R_k = I - S_k (S'_k S_k)^{-1} \, S'_k \quad . \tag{4.28}$$

Al passo successivo (l'aggiunta della colonna s_{k+2} al modello (4.25)), si aggiorna lo stimatore

$$\bar{\theta}^{k+1} := \begin{bmatrix} \hat{\theta}_k \\ \hat{\theta}_{k+1} \end{bmatrix} \quad , \tag{4.29}$$

con formule esattamente analoghe alle (4.26)–(4.27). Naturalmente il vero problema è quello di fare sequenzialmente anche i calcoli relativi all'aggiornamento dei coefficienti, in particolare dell'inversa $(S'_{k+1} S_{k+1})^{-1}$ a partire da $(S'_k S_k)^{-1}$. Si può pensare di fare questi conti aggiornando la fattorizzazione di Cholesky di $S'_k S_k$, ma se si riflette un momento si vede che questo procedimento non è altro che un metodo di fattorizzazione di S_k come prodotto di una matrice ortogonale Q (che non viene esplicitamente prodotta) e di una triangolare superiore (il fattore *destro* di Cholesky di $S'_k S_k$). In questo modo quindi si introduce implicitamente nel problema una fattorizzazione QR di S. Tanto vale allora cercare di vedere chiaramente come vanno le cose e studiare esplicitamente l'aggiornamento della fattorizzazione QR della matrice S. Tanto per fissare le idee, supponiamo che la fattorizzazione venga calcolata usando matrici elementari di Householder.

Algoritmo a stadi di Golub-Styan

Allo stadio k-esimo ($k \geq 1$) si dispone della matrice ortogonale Q_k, prodotto di k matrici di riflessione elementare, di una matrice triangolare superiore U_k e di un vettore y_k, ottenuto trasformando i dati di misura y attraverso la Q_k, tali che

$$Q_k[S_k \; y] = \left[\begin{array}{c|c} U_k & y_k^1 \\ \hline 0 & y_k^2 \end{array}\right] \quad , \qquad k \text{ righe} \; . \tag{4.30}$$

Chiaramente si ha, con ovvio significato dei simboli,

$$\bar{\theta}^k = U_k^{-1} \; y_k^1 \quad , \qquad \|\varepsilon_k\|^2 = \|y_k^2\|^2 \quad . \tag{4.31}$$

Supponiamo ora di aggiungere a S_k una colonna linearmente indipendente, s_{k+1}, e di disporre sempre della matrice Q_k memorizzata ad esempio mediante i k vettori $v_1, \ldots, v_k$ che definiscono le riflessioni elementari.

Si calcola il prodotto

$$Q_k\ s_{k+1} := \begin{bmatrix} a_{k+1} \\ b_{k+1} \end{bmatrix} \begin{matrix} \} \ k \text{ righe} \\ \} \ n-k \text{ righe} \end{matrix} \tag{4.32}$$

per cui

$$Q_k[S_k\ s_{k+1}\ y] = \begin{bmatrix} U_k & a_{k+1} & y_k^1 \\ 0 & b_{k+1} & y_k^2 \end{bmatrix}$$

e si introduce una nuova riflessione elementare di dimensione $(n-k)\times(n-k)$, H_{k+1}, tale che il vettore $H_{k+1}\ b_{k+1}$ ha tutte le componenti nulle nelle posizioni $k+2,\ldots,n$

$$H_{k+1}\ b_{k+1} = \|b_{k+1}\|\ e_1 \quad . \tag{4.33}$$

H_{k+1} riflette b_{k+1} in $\|b_{k+1}\|\ e_1$ ed è definita da $v_{k+1} = b_{k+1} - \|b_{k+1}\|\ e_1$, dove $e_1 = [1,0,\ldots,0]'$ in $\mathbb{R}^{n-k}$. Definendo allora

$$Q_{k+1} := \begin{bmatrix} I_k & 0 \\ 0 & H_{k+1} \end{bmatrix} \quad , \tag{4.34}$$

$$H_{k+1}\ y_k^2 := z_{k+1} \quad , \tag{4.35}$$

si ha

$$Q_{k+1}[S_k\ s_{k+1}\ y] = \left[\begin{array}{cc|c} U_k & a_{k+1} & y_k^1 \\ & \|b_{k+1}\| & \\ 0 & 0 & z_{k+1} \end{array}\right] \tag{4.36}$$

e questa nuova fattorizzazione permette di ricavare immediatamente le stime $\hat{\theta}^k$ e $\hat{\theta}_{k+1}$ come soluzioni del sistema

$$\begin{bmatrix} U_k & a_{k+1} \\ 0 & \|b_{k+1}\| \end{bmatrix} \begin{bmatrix} \theta_k \\ \hat{\theta}_{k+1} \end{bmatrix} = \begin{bmatrix} y_k^1 \\ z_{k+1}^1 \end{bmatrix} \quad . \tag{4.37}$$

(Al secondo membro di questa equazione si è usato il simbolo z_{k+1}^1 per indicare la prima componente del vettore $(n-k)$-dimensionale z_{k+1}).

Evidentemente la *somma dei quadrati dei residui*, dopo l'introduzione della $(k+1)$-sima variabile di regressione, vale

$$\|\varepsilon_{k+1}\|^2 = \left\| \begin{bmatrix} z_{k+1}^2 \\ \vdots \\ z_{k+1}^{n-k} \end{bmatrix} \right\|^2 = \|z_{k+1}\|^2 - (z_{k+1}^1)^2 = \|H_{k+1}\ y_k^2\|^2 - (z_{k+1}^1)^2$$

ovvero, tenendo conto della seconda relazione in (4.31),

$$\|\varepsilon_{k+1}\|^2 = \|\varepsilon_k\|^2 - (z_{k+1}^1)^2 \quad . \tag{4.38}$$

A questo punto *si può iniziare il passo* $(k+2)$-simo prendendo come dati iniziali

$$U_{k+1} = \begin{bmatrix} U_k & a_{k+1} \\ 0 & \|b_{k+1}\| \end{bmatrix}$$

$$y_{k+1}^1 = \left[(y_k^1)', z_{k+1}^1\right]'$$

$$y_{k+1}^2 = \left[z_{k+1}^2, \ldots, z_{k+1}^{n-k}\right]'$$

e usando la matrice ortogonale Q_{k+1} definita in (4.34) per trasformare il vettore s_{k+2} nel modo analogo a quanto fatto in (4.32) ecc...

Come si vede, questo algoritmo è esattamente l'algoritmo di fattorizzazione di Householder descritto alla fine del capitolo precedente. Per il noto significato geometrico della fattorizzazione QR, si vede subito che il proiettore R_k sullo spazio ortogonale a $\mathcal{S}_k := sp[S_k]$ opera sui vettori s_{k+1} e y semplicemente attraverso le

$$R_k\ s_{k+1} = \begin{bmatrix} 0 \\ b_{k+1} \end{bmatrix} \left[\right\} \ k \text{ righe } \left.\right]$$

ed

$$R_k y = \begin{bmatrix} 0 \\ y_k^2 \end{bmatrix}$$

per cui, ad esempio, la formula (4.26) per $\hat{\theta}_{k+1}$ si può riscrivere come

$$\hat{\theta}_{k+1} = \frac{1}{\|b_{k+1}\|^2}\ b'_{k+1}\ y_k^2 \quad . \tag{4.39}$$

È facile controllare che questa relazione è identica alla

$$\hat{\theta}_{k+1} = \frac{z_{k+1}^1}{\|b_{k+1}\|} \quad , \tag{4.40}$$

che si ricava risolvendo l' ultima equazione in (4.37). La matrice della varianza delle stime si può scrivere immediatamente come

$$\Sigma_{k+1} = \sigma^2 \begin{bmatrix} \Sigma_k + \dfrac{U_k^{-1}\, a_{k+1}\, a'_{k+1}\, U_k^{-T}}{\|b_{k+1}\|^2} & -U_k^{-1}\, a_{k+1}\, \dfrac{1}{\|b_{k+1}\|^2} \\ -\dfrac{1}{\|b_{k+1}\|^2}\, a'_{k+1}\, U_k^{-T} & \dfrac{1}{\|b_{k+1}\|^2} \end{bmatrix} \tag{4.41}$$

semplicemente notando che $\hat{\theta}^k$ può essere espresso nella forma

$$\hat{\theta}^k = \bar{\theta}^k - U_k^{-1}\ a_{k+1}\ \hat{\theta}_{k+1} \quad . \tag{4.42}$$

Questo algoritmo di M.Q. a stadi permette di controllare ad ogni passo quant'è la diminuzione di errore quadratico medio che si ottiene introducendo un ulteriore parametro nel modello e addirittura di confrontare tra loro le diminuzioni corrispondenti all'introduzione di una qualunque colonna addizionale scelta nell'insieme $\{s_{k+1}, s_{k+2}, \ldots, s_p\}$ delle colonne "mancanti" di un modello lineare

$$S\theta = [s_1, \ldots, s_k, \ldots, s_p] \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} \tag{4.43}$$

che si "propone" per descrivere i dati. Chiaramente, il modello (4.43) può risultare inutilmente complicato nel senso che l'introduzione di alcuni fra i parametri $\{\theta_{k+1}, \ldots, \theta_p\}$

può portare a una riduzione relativa dell'errore quadratico medio $\|\varepsilon_k\|^2$ che è piccola o insignificante.

In questo caso conviene controllare quanto si paga in termini di varianza a introdurre il nuovo parametro ed eventualmente decidere di eliminarlo, usando un modello più semplice.

Quest'ultimo ragionamento presenta in realtà un punto debole. In effetti la diminuzione di errore quadratico medio corrispondente all'introduzione di una nuova colonna, s_{k+1}, non dipende solo da s_{k+1} ma ovviamente anche dalle colonne che sono state scelte in precedenza per formare S_k. Un'analisi più soddisfacente del problema richiede strumenti un tantino più raffinati della decomposizione QR .

L'algoritmo di M.Q. a stadi che abbiamo discusso è dovuto a Golub e Styan [13].

Uso della SVD

L'ultima osservazione ci induce a descrivere brevemente un possibile modo, basato sulla decomposizione ai valori singolari, di valutare gli effetti dell' introduzione di un nuovo regressore nel modello. Questo strumento chiarifica di molto l 'analisi del problema della stima della complessità del modello lineare sviluppata nella sezione precedente.

Siano:

$$S_1 = \bar{U}\bar{\Delta}\bar{V}' \qquad S = U\Delta V' \tag{4.44}$$

le SVD delle matrici S_1 e della matrice aumentata $S := [S_1 \;\; S_2]$ dove $\bar{\Delta}$ e Δ hanno la struttura quasi diagonale:

$$\bar{\Delta} = \begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & 0 \end{bmatrix}, \qquad \bar{\Sigma} = \operatorname{diag}\{\bar{\sigma}_1, \ldots, \bar{\sigma}_p\} \tag{4.45}$$

$$\Delta = \begin{bmatrix} \Sigma_1 & 0 & 0 \\ 0 & \Sigma_2 & 0 \\ 0 & 0 & 0 \end{bmatrix}; \tag{4.46}$$

$$\Sigma_1 := \operatorname{diag}\{\sigma_1, \ldots, \sigma_p\} \qquad \Sigma_2 := \operatorname{diag}\{\sigma_{p+1}, \ldots \sigma_{p+k}\}\} \tag{4.47}$$

Da notare che in generale tutti i valori singolari cambiano quando si aggiungono nuove colonne a S_1 e quindi, se $k \geq 1$, si ha $\bar{\Sigma} \neq \Sigma_1$; i.e. $\bar{\sigma}_i \neq \sigma_i\,, i = 1, \ldots, p$.

Cambiando base nel modello lineare aumentato e definendo $\bar{y} := U^\top y$, $\bar{w} := U^\top w$ e $\beta := V^\top \theta$, si ottiene

$$\bar{y} = \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix} \beta + \lambda \bar{w} \tag{4.48}$$

dove, per evitare confusioni abbiamo indicato con λ la deviazione standard del termine d'errore $\boldsymbol{\epsilon}$. In questo modello la varianza di $\bar{\mathbf{w}}$ rimane invariata (uguale a I_N) e le stime dei nuovi parametri si ricavano per ispezione

$$\hat{\boldsymbol{\beta}}_i(\mathbf{y}) = \frac{1}{\sigma_i}\bar{\mathbf{y}}_i; \qquad \operatorname{var}\{\hat{\boldsymbol{\beta}}_i\} = \frac{\lambda^2}{\sigma_i^2} \qquad i = 1, \ldots, p+k\,. \tag{4.49}$$

Da notare che gli stimatori $\hat{\boldsymbol{\beta}}_i(\mathbf{y})\,;\; i = 1, \ldots, p+k$ sono scorrelati (o indipendenti nel caso Gaussiano). Dato che i valori singolari sono ordinati in modo decrescente, si può

dire in generale che le varianze delle stime dei parametri aumentano all'aumentare della complessità del modello. In particolare, se il nuovo valore singolare σ_{p+1} dovuto all' aggiunta di una nuova colonna, s_{p+1}, risulta molto più piccolo di quelli del primo blocco Σ_1, la corrispondente stima del parametro aggiuntivo $\hat{\boldsymbol{\beta}}_{p+1}$ avrà in effetti varianza (molto) maggiore delle altre componenti. Notiamo il seguente fatto notevole:

Il rapporto tra la varianza di $\hat{\boldsymbol{\beta}}_{p+1}$ e quella del primo stimatore $\hat{\boldsymbol{\beta}}_1$ (che è la minima possibile) è il quadrato dell'indice di condizionamento numerico della matrice aumentata.

Questo legame diretto tra il condizionamento numerico della matrice dei regressori e la varianza delle stime dei parametri riguarda allo stesso modo lo stimatore originale $\hat{\boldsymbol{\theta}} = V\hat{\boldsymbol{\beta}}$ di matrice varianza $V \operatorname{Var}\{\hat{\boldsymbol{\beta}}\}V^\top$, la cui varianza scalare è, come si controlla facilmente, la stessa di quella di $\hat{\boldsymbol{\beta}}$. La regola che scatursce da questa osservazione è che bisogna sempre cercare di introdurre nuovi regressori che mantengano il buon condizionamento della matrice S. Al limite, se possibile, introdurre nuovi regressori che siano "quasi ortogonali" alle colonne preesistenti. L'introduzione di regressori che porti ad una matrice aumentata con colonne "quasi dipendenti" [6] è assolutamente da evitare.

Come si vede, c'è un aspetto del problema della regressione a stadi, il problema chiamato in letteratura della *collinearità dei regressori* [26] che non era apparso nell'analisi precedente e va invece attentamente considerato quando si tratta di decidere la complessità di un modello. Oltre a questo c'è ancora il problema di confrontare gli errori residui corrispondenti alle stime dei primi p parametri nelle due situazioni. Per fare questo bisogna confrontare i p valori singolari originali, $\bar{\sigma}_i$ di $\bar{S}_1$ con i primi p valori singolari $\sigma_i\,;\ i = 1, \ldots, p$, della matrice aumentata.

Per risolvere in modo soddisfacente questo problema bisognerebbe introdurre le formule per l' *aggiornamento sequenziale della decomposizione ai valori singolari* corrispondenti all'aggiunta di nuove colonne nella matrice S. Noi però non insisteremo oltre su questo punto. Il prototipo di queste formule e i relativi algoritmi di calcolo sono descritti in [4, **?**].

[6] Notare che questo non significa necessariamente che il nuovo regressore debba essere "quasi dipendente" dalle colonne preesistenti.

Capitolo 5
MODELLI DINAMICI PER L'IDENTIFICAZIONE

5.1 Introduzione

Come si pone il problema dell'identificazione. Approccio statistico

5.2 Modelli statistici lineari per processi del secondo ordine

In questa sezione studieremo una classe di modelli probabilistici di segnali osservati. L'assunto fondamentale è che i dati osservati possano essere descritti come tratti (ovviamente finiti) di traiettorie (in inglese *sample paths*) di processi stocastici del second'ordine[7] debolmente stazionari. Questa ipotesi è molto blanda. Essa verrà comunque discussa in maggior dettaglio nel capitolo dedicato all'ergodicità.

Cosidereremo solo segnali (e processi) a tempo discreto, $t \in \mathbb{Z}$, e sostanzialmente ci occuperemo solo di processi scalari. Senza perdita di generalità assumeremo che tutti i processi in gioco abbiano media nulla.

I processi di questo tipo sono competamente descrivibili per mezzo della funzione di covarianza o, equivalentemente, della distribuzione spettrale di potenza. Come è ben noto, nel caso Gaussiano queste ultime individuano completamente tutte le distribuzioni finito-dimensionali (i.e. la legge di probabilità) del processo. Assumeremo che i segnali osservati siano 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.

In quanto segue supporremo che la variabile $\mathbf{y}$ sia monodimensionale e che sia $\mathbf{y}$ che $\mathbf{u}$ siano descrivibili mediante processi del second'ordine congiuntamente stazionari. Per

[7] Con momenti del secondo ordine finiti.

semplicità assumeremo che anche la variabile di ingresso $\mathbf{u}$ sia monodimensionale anche se l'estensione di quanto verremo dicendo al caso di un ingresso multidimensionale, non presenta difficoltà di sorta.

Dato che $\mathbf{y}$ ed $\mathbf{u}$ sono congiuntamente stazionari, proiettando $\mathbf{y}(t)$ sullo spazio passato $H_t(\mathbf{u})$ si ottiene, come è ben noto, una decomposizione del tipo

$$\mathbf{y}(t) = F(z)\mathbf{u}(t) + \mathbf{v}(t)\,, \qquad t \in \mathbb{Z} \tag{5.1}$$

dove $F(z)$ è una funzione di trasferimento causale (non necessariamente razionale) e $\mathbf{v}$ è un processo stazionario, detto *errore di modellizzazione*, scorrelato dal passato di $\mathbf{u}$, ovvero

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

Daremo qui per noto il concetto di *modello a retroazione* e quello di *retroazione tra processi stocastici*, si veda ad esempio [21, Cap. 7]. Come è noto si può scrivere una decomposizione analoga alla (5.1) per la variabile $\mathbf{u}$,

$$\mathbf{u}(t) = H(z)\mathbf{y}(t) + \mathbf{r}(t)\,, \qquad t \in \mathbb{Z} \tag{5.2}$$

dove $H(z)$ è una funzione di trasferimento causale (non necessariamente razionale). Esistono inoltre (molte) rappresentazioni congiunte (5.1) e (5.2) in cui il processo $\mathbf{r}$ è completamente scorrelato da $\mathbf{v}$. Questi modelli congiunti si chiamano *modelli a retroazione* della coppia $(\mathbf{y},\ \mathbf{u})$. Per la stazionarietà congiunta, il sistema a controreazoone formato dall'interconnessione di (5.1) e (5.2) dev'essere internamente stabile, [21, Cap. 7]. Si dimostra che se (e solo se) $\mathbf{u}$ e $\mathbf{v}$ sono completamente incorrelati, ovvero vale la

$$\mathbb{E}\,\mathbf{v}(t)\mathbf{u}(s) = 0 \qquad t\,, s\, \in \mathbb{Z}. \tag{5.3}$$

si ha *assenza di reazione da* $\mathbf{y}$ *a* $\mathbf{u}$ e in questo caso $H(z) \equiv 0$. In assenza di retroazione la funzione di trasferimento $F(z)$ è ℓ^2-stabile.

In quanto segue assumeremo che $F(z)$ sia strettamente causale, ovvero che $F(\infty) = 0$ e che $\mathbf{v}$ sia un proccesso puramente non deterministico (p.n.d.) che ammette quindi una rappresentazione di innovazione

$$\mathbf{v}(t) = G(z)\mathbf{e}(t)\,, \qquad t \in \mathbb{Z} \tag{5.4}$$

dove $G(z)$ è una funzione di trasferimento (non necessariamente razionale) *a fase minima* che prenderemo sempre normalizzata all'infinito, $G(\infty) = 1$ e 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 $\mathbf{y}(t)$ basato sul passato congiunto di $\mathbf{u}$ e $\mathbf{y}$ all'istante $t-1$. Combinando (5.1) e (5.4) si ottiene

$$\mathbf{y}(t) = F(z)\mathbf{u}(t) + G(z)\mathbf{e}(t)\,, \qquad t \in \mathbb{Z} \tag{5.5}$$

Questo è il modello a cui faremo riferimento in seguito. Naturalmente l'idea di *modello statistico* presuppone che si abbia equivalenza tra la descrizione "esplicita delle variabili in gioco (i processi $\mathbf{y}$ e $\mathbf{u}$) che si ottiene mediante il modello e la descrizione implicita (o "esterna) delle variabili, fatta in generale mediante la loro distribuzione di probabibilità

congiunta, o come nel caso in esame, mediante le statistiche congiunte del secondo ordine. È ovvio che nel caso di presenza di reazione il modello (5.5) da solo non è sufficiente ad individuare univocamente la covarianza o lo spettro congiunti di $\mathbf{y}$ e $\mathbf{u}$ e il modello completo da considerare nel caso di presenza di reazione è quello congiunto comprendente anche la descrizione del canale di retroazione.

Quando c'è reazione, noi supporremo sempre che il processo $\mathbf{u}$ sia generato da una retroazione lineare[8] e *causale*, cioè una retroazione del tipo

$$\mathbf{u}(t) = H(z)\,\mathbf{y}(t) + \mathbf{r}(t) \quad , \tag{5.6}$$

dove $H(z)$ è una funzione (o matrice) di trasferimento non necessariamente razionale, regolare all'infinito e $\mathbf{r}$ un processo stazionario scorrelato da $\mathbf{e}$. Naturalmente, per la stazionarietà congiunta dei processi di ingresso-uscita $[\mathbf{y}\ \mathbf{u}]^\top$, l' interconnessione a retroazione di (5.5) e (5.6) dev'essere *internamente stabile*, ovvero la matrice di trasferimento in catena chiusa, $T(z)$, che trasforma il processo congiunto $[\mathbf{e}\ \mathbf{r}]^\top$ in $[\mathbf{y}\ \mathbf{u}]^\top$ dovrà essere analitica in $\{\,|z| \geq 1\,\}$. Si dimostra allora che c'è un modello (essenzialmente unico) del tipo (5.5) in cui $F(z)$ 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 $\{\,|z| \geq 1\,\}$.

Se queste condizioni sono soddisfatte (beninteso assumendo una reazione causale del tipo (5.6)) si può allora dimostrare (si veda ad esempio [21, cap. 8, proposizione 4.1].) che 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 $t-1$. Il modello (5.5) con reazione, si chiama allora *modello d'innovazione*.

Ciò premesso, passiamo senz'altro a ricordare la struttura del predittore di Wiener (ovvero il predittore stazionario basato su dati disponibili a partire da un istante iniziale infinitamente remoto) per i modello generale di Box Jenkins (5.5) in cui si potrebbe essere in presenza di reazione. Il risultato seguente è standard [21, cap. 7.4].

Teorema 5.1. *Si assuma*[9] *che nel modello* (5.5)*,*

- *$F(z)$ ha almeno un ritardo, ovvero $F(z) = z^{-1}F_1(z)$ dove $F_1(z)$ è limitata all' infinito,*

- *$G(z)$ è normalizzata all'infinto ($G(\infty) = 1$) ed è priva di zeri nella regione $\{|z| \geq 1\}$ del piano complesso,*

- *$G(z)^{-1}F(z)$ è priva di poli nella regione $\{|z| \geq 1\}$ del piano complesso,*

- *$\mathbf{e}(t)$ è scorrelato con la storia passata congiunta $(\mathbf{y}^{t-1}\,,\,\mathbf{u}^{t-1}) := \{\mathbf{y}(s),\,\mathbf{u}(s)\,;\,s < t\}$ dei processi $\mathbf{y}$ e $\mathbf{u}$.*

[8] Nel contesto di descrizioni del second'ordine in cui operiamo la linearità non è un'ipotesi restrittiva.

[9] Come appena spiegato queste ipotesi possono sempre essere verificate se la retroazione è causale.

Allora, 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 descritto dalla

$$\hat{\mathbf{y}}(t \mid t-1) = G(z)^{-1}F_1(z)\mathbf{u}(t-1) + G(z)^{-1}G_1(z)\mathbf{y}(t-1) \tag{5.7}$$

ed $\mathbf{e}(t)$ è l'errore di predizione di un passo (i.e. l'innovazione) di $\mathbf{y}(t)$ dato il passato congiunto $(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})$.

Dimostrazione. Si scriva il modello nella forma

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

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

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

si trova che

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

dove il secondo membro è, per le ipotesi poste, 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 $\mathbf{u}$, questa funzione è proprio il predittore cercato. □

5.3 Modelli parametrici e identificabilità in assenza di retroazione

Per ragioni di semplicità espositiva in questa sezione di norma *assumeremo che non ci sia reazione da $\mathbf{y}$ a $\mathbf{u}$*. In questo caso, ammettendo per il momento che $\mathbf{u}$ abbia densità spettrale $S_{\mathbf{u}}(z)$, la descrizione spettrale di $\mathbf{y}$ indotta dal modello (5.5) si può scrivere esplicitamente come

$$S_{\mathbf{y}}(z) = F(z)S_{\mathbf{u}}(z)F(1/z) + \lambda^2 G(z)G(1/z)\,. \tag{5.8}$$

Come si vede, $S_{\mathbf{y}}(z)$ è parametrizzata dalla densità dell'ingresso, $S_{\mathbf{u}}(z)$, che dipende, come si dice convezionalmente, dalla *condizione sperimentale* e, quando i dati sono raccolti durante il "normale funzionamento" dell'impianto, si pensa normalmente imposta dall'esterno. In particolare, nel caso in cui sia possibile invece progettare l'ingresso nell'esperimento di identificazione, $S_{\mathbf{u}}(z)$ può essere imposta in modo da ottimizzare il risultato dell'esperimento.

In questo caso è bene rimarcare che l'ingresso può spesso essere costituito da combinazioni di segnali "deterministici, ad esempio somme di sinusoidi di frequenze diverse, che non possono essere assimilati a processi con densità spettrale. In questo caso l'espressione (5.8) dovrebbe essere riscritta usando le distribuzioni spettrali, in particolare sotituendo a $S_{\mathbf{u}}(z)$ la relativa distribuzione spettrale di potenza $d\hat{F}_{\mathbf{u}}(z)$.

È ovvio che il problema di inferenza statistica per il modello (5.5) diventa un problema parametrico quando le funzioni di trasferimento $F(z)$ e $G(z)$ *sono funzioni razionali di*

z. In generale si può pensare che i parametri da cui dipendono queste due funzioni razionali siano i coefficienti dei rispettivi polinomi a numeratore e a denominatore, o eventualmente, alcune loro combinazioni algebriche. Si possono ovviamente dare anche casi in cui si ha a priori una conoscenza parziale di questi parametri. Lascieremo per il momento la questione nel vago e useremo la notazione

$$F_\theta(z),\ G_\theta(z) \qquad \theta \in \Theta \subset \mathbb{R}^p$$

per esprimere la dipendenza di $F(z)$ e $G(z)$ da un parametro vettoriale incognito θ, senza specificare esattamente come $F(z)$ e $G(z)$ vi dipendano.

La nozione naturale di identificabiltà da applicare al contesto attuale impone di sostituire modelli probabilstici in senso stretto con modelli per le statistiche del secondo ordine. Per il modello parametrico (5.5)

$$\mathbf{y}(t) = F_\theta(z)\mathbf{u}(t) + G_\theta(z)\mathbf{e}(t)\,, \qquad \theta \in \Theta \subset \mathbb{R}^p \tag{5.9}$$

faremo quindi riferimento alla famiglia parametrica di spettri

$$\begin{aligned} S_{\mathbf{y}}(z;\theta) &= F_\theta(z)S_{\mathbf{u}}(z)F_\theta(1/z) + \lambda^2 G_\theta(z)G_\theta(1/z)\,, \\ S_{\mathbf{yu}}(z;\theta) &= F_\theta(z)S_{\mathbf{u}}(z) \qquad\qquad \theta \in \Theta \subset \mathbb{R}^p \end{aligned} \tag{5.10}$$

con la solita avvertenza di sostituire all'occorrenza densità spettrali con le relative distribuzioni. L'identificabilità del modello (5.5) deve intuitivamente corrispondere a una parametrizzazone non ridondante dello spettro congiunto (5.10). 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 $S_{\mathbf{yu}}$.

Definizione 5.1. *Si assuma assenza di reazione da* $\mathbf{y}$ *a* $\mathbf{u}$*. Il modello* (5.9) *è identificabile (globalmente) nella condizione sperimentale descritta dallo spettro di ingresso* $S_{\mathbf{u}}$ *(o dalla distribuzione spettrale di ingresso* $d\hat{F}_{\mathbf{u}}$*), se la mappa* $\theta \mapsto [S_{\mathbf{y}}(\cdot;\theta)\,, S_{\mathbf{yu}}(\cdot;\theta)\,]$ *è iniettiva in* Θ*, ovvero*

$$[S_{\mathbf{y}}(\cdot;\theta_1)\,, S_{\mathbf{yu}}(\cdot;\theta_1)\,] = [S_{\mathbf{y}}(\cdot;\theta_2)\,, S_{\mathbf{yu}}(\cdot;\theta_2)\,] \;\Rightarrow\; \theta_1 = \theta_2 \tag{5.11}$$

Se si ha iniettività locale in un intorno di θ_0*, si parla di* identificabilità locale *in* θ_0.

Proposizione 5.1. *Nel modello di Box-Jenkins 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*

$$[F_{\theta_1}(z) - F_{\theta_2}(z)]\, S_{\mathbf{u}}(z) = 0 \tag{5.12}$$

$$G_{\theta_1}(z) - G_{\theta_2}(z) = 0 \tag{5.13}$$

per ogni $z = e^{j\omega}$.

Dimostrazione. Dato che

$$F_\theta(z)S_{\mathbf{u}}(z) = S_{\mathbf{yu}}(z;\theta)$$

evidentemente la (5.12) è equivalente all'uguaglianza degli spettri $S_{\mathbf{yu}}(z;\theta_1)$ e $S_{\mathbf{yu}}(z;\theta_2)$. Notiamo poi che se vale la (5.12) si ha $F_{\theta_1}(z)S_{\mathbf{u}}(z)F_{\theta_1}(1/z) = F_{\theta_2}(z)S_{\mathbf{u}}(z)F_{\theta_2}(1/z)$ e quindi segue immediatamente dalla prima delle (5.10) che (5.12) e (5.13) implicano $S_{\mathbf{y}}(z;\theta_1) = S_{\mathbf{y}}(z;\theta_2)$.

Viceversa, facciamo vedere che le (5.11) implicano le (5.12) e (5.13). Per quanto appena visto l'uguaglianza degli spettri incrociati implica la (5.12) e dalla

$$S_{\mathbf{y}}(z;\theta) - F_\theta(z)S_{\mathbf{u}}(z)F_\theta(1/z) = \lambda^2 G_\theta(z)G_\theta(1/z)\,.$$

l'uguaglianza degli spettri dell'uscita implica $G_{\theta_1}(z)G_{\theta_1}(1/z) = G_{\theta_2}(z)G_{\theta_2}(1/z)$. 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. □

Una nozione di identificabilità che spesso in letteratura viene confusa con quella precedente è l'identificabilità *a priori*.

Definizione 5.2. *Il modello* (5.9) *è identificabile* a priori *(globalmente) se la mappa* $\theta \mapsto [F_\theta(\cdot),\ G_\theta(\cdot)]$ *è iniettiva in* Θ*, ovvero*

$$[F_{\theta_1}(\cdot),\ G_{\theta_1}(\cdot)] = [F_{\theta_2}(\cdot),\ G_{\theta_2}(\cdot)] \ \Rightarrow\ \theta_1 = \theta_2 \tag{5.14}$$

Se si ha iniettività locale in un intorno di θ_0*, si parla di* identificabilità a priori locale *in* θ_0.

Per apprezzare la diversità dei due concetti basta analizzare la mappa che descrive la dipendenza dello spettro congiunto dal parametro θ. Questa mappa è in realtà 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. Notiamo subito che l'identificabilità a priori corrisponde all'iniettività della prima componente ed è quindi solo condizione necessaria per l'identificabiltà definita nella definizione 5.1. In effetti se il modello (5.9) non fosse identificabile a priori non sarebbe possibile distinguere i parametri in base all'assegnazione dello spettro congiunto $[S_{\mathbf{y}}(\cdot;\theta)\,, S_{\mathbf{yu}}(\cdot;\theta)]$, comunque venga assegnata la condizione sperimentale.

Rimarchiamo che un modello identificabile a priori potrebbe benissimo non essere identificabile per qualche condizione sperimentale. Il caso più ovvio è quando si ha un ingresso costante per cui $S_{\mathbf{u}}(z) \equiv 0$.

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 apriori è quindi un fatto puramente algebrico. L'identificabilità vera e propria richiede invece (oltre all'identificabilità a priori) che il segnale di ingresso soddisfi ad alcune condizioni note come **persistente eccitazione**. Queste condizioni dipendono dalla struttura della parametrizzazione di $F_\theta(z)$ e verranno discusse più avanti.

5.4 Alcune classi di modelli e loro parametrizzazione

I modelli razionali del tipo (5.5) (detti qualche volta del tipo "Box-Jenkins"), se non vi sono vincoli a priori sui parametri, possono essere parametrizzati mediante i coefficienti dei relatvi polinomi a numeratore e denominatore delle funzioni di trasferimento nel modello parametrico (5.5),

$$F_\theta(z) = \frac{B(z^{-1})}{A(z^{-1})} \qquad G_\theta(z) = \frac{C(z^{-1})}{D(z^{-1})}\,. \tag{5.15}$$

Dato che per convenzione A, C D sono monici e B ha il coefficiente di grado zero (b_0) uguale a zero, ciascun modello può essere descritto mediante

- gli $n = \text{grado}\,(A)$ coefficienti del polinomio $A(z^{-1}) = 1 + \sum_{k=1}^{n} a_k z^{-k}$,
- gli $m = \text{grado}\,(B)$ coefficienti del polinomio $B(z^{-1}) = \sum_{k=1}^{m} b_k z^{-k}$,
- gli $r = \text{grado}\,(C)$ coefficienti del polinomio $C(z^{-1}) = 1 + \sum_{k=1}^{r} c_k z^{-k}$,
- gli $r = \text{grado}\,(D)$ coefficienti del polinomio[10] $D(z^{-1}) = 1 + \sum_{k=1}^{r} d_k z^{-k}$

e quindi con un totale di $p = n+m+2r$ parametri "liberi", più la varianza dell'innovazione λ che conviene considerare separatamente. In realtà il vincolo che il modello (5.5) sia d'innovazione (G a fase minima) si dovrebbe imporre vincolando i coefficienti di C e D a definire polinomi strettamente stabili. Inoltre, se non c'è reazione, anche A dovrebbe essere vincolato a essere strettamente stabile. Questi vincoli definiscono in teoria l'insieme dei parametri ammissibili Θ. Purtroppo la struttura geometrica degli insiemi che definiscono i coefficenti 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

$$\mathcal{A}(z^{-1})\mathbf{y}(t) = \mathcal{B}(z^{-1})\mathbf{u}(t) + \mathcal{C}(z^{-1})\mathbf{e}(t) \tag{5.16}$$

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 (5.16) 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 un modello ARMAX si descrive l'errore di modellizzazione $\mathbf{v}$ 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, l'uso di questa struttura porta in pratica a identificare modelli di ordine più alto del dovuto. Di fatto il modello ARMAX equivalente al Box-Jenkins (5.5) dovrebbe avere la struttura seguente

$$\mathcal{A}(z^{-1}) = A(z^{-1})D(z^{-1}) \quad \mathcal{B}(z^{-1}) = B(z^{-1})D(z^{-1}) \quad \mathcal{C}(z^{-1}) = C(z^{-1})A(z^{-1})$$

[10] Analogamente a quanto fatto sopra si può in generale parametrizzare D assumendo un grado diverso da quello di C. Noi qui assumeremo gradi uguali per non complicare troppo le notazioni. Il lettore può facilmente ripercorrere il ragionamento che segue assumendo $q = \text{grado}\,(D)$ diverso da r.

in cui però i parametri dei polinomi $\mathcal{A}$, $\mathcal{B}$ e $\mathcal{C}$ (in totale $n + r + m + r + r + n = 2n + m + 3r$) non sono più liberi di variare in modo indipendente ma debbono essere vincolati a soddisfare le relazioni algebriche che impongono le relazioni prodotto scritte sopra.

In pratica, nei procedimenti di stima questi vincoli algebrici sono impossibili da rispettare 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

$$\mathcal{A}(z^{-1})\mathbf{y}(t) = \mathcal{B}(z^{-1})\mathbf{u}(t) + \mathbf{e}(t) \tag{5.17}$$

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 (5.17) 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.

5.5 Identificabilità in presenza di reazione

Da scrivere

5.5.1 Modelli a errori nelle variabili

Da scrivere:

5.6 Modelli multivariabili

Vedere il capitolo nel libro: Bittanti (ed.), *Identificazione parametrica* CLUP, Milano.

5.7 Modelli Gaussiani

La funzione di verosimiglianza. Il LImite di Cramèr-Rao per modelli dinamici Gaussiani.

Capitolo 6
ERGODICITÀ

6.1 Proprietà asintotiche degli stimatori: Consistenza

È ovvio che in un qualunque procedimento sensato di inferenza ci si aspetta di ottenere risultati sempre migliori al crescere della numerosità del campione. Lo studio del comportamento di uno stimatore o di un test quando la numerosità campionaria n tende all'infinito serve quindi a dare un'idea delle prestazioni "limite", cioè del massimo che ci si può aspettare dal procedimento di stima (o di verifica di ipotesi) che hanno portato alla scelta dello stimatore o del test. In statistica è talvolta possibile studiare il comportamento limite di uno stimatore quando la numerosità campionaria n tende all'infinito, se si fanno opportune ipotesi sul meccanismo probabilistico che ha generato i dati. Gli strumenti più usati per l'analisi asintotica sono il *teorema ergodico* e il *teorema del limite centrale*. Il primo verrà discusso in questo capitolo.

Definizione 6.1. *Sia* $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ *un campione estratto dalla famiglia* $\{F_\theta \ ; \ \theta \in \Theta\}$ *e sia* θ_0 *il valore vero del parametro. Uno stimatore* $\phi_n(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ *si dice "consistente" se*

$$\lim_{n\to\infty} \ \phi_n(\mathbf{y}_1, \ldots, \mathbf{y}_n) = \theta_0 \quad ; \tag{6.1}$$

se il limite (6.1) *è un limite in probabilità, cioè se la* (6.1) *significa che,* $\forall \varepsilon > 0$*,*

$$\lim_{n\to\infty} \ P_{\theta_0} \left(|\phi_n(\mathbf{y}_1, \ldots, \mathbf{y}_n) - \theta_0| \geq \varepsilon \right) = 0 \quad , \tag{6.2}$$

si parla di consistenza "debole" o semplicemente di consistenza tout-court. Se invece il limite (6.1) *è un limite con probabilità 1 (c.p. 1), si parla di consistenza forte. In questo caso si ha*

$$\lim_{n\to\infty} \ \phi_n(y_1, \ldots, y_n) = \theta_0 \tag{6.3}$$

per tutte le possibili successioni $\{y_1, y_2, \ldots, y_n, \ldots\}$ *di osservazioni in* $(\mathbb{R}^m)^\infty := \mathbb{R}^m \times \mathbb{R}^m \times \ldots$ *(infinite volte), eccettuato al più un insieme di successioni di osservazioni di probabilità zero.*

(Come sia definita la probabilità sullo spazio di tutte le misure "infinitamente lunghe", $(\mathbb{R}^m)^\infty$, è una questione per la quale rimandiamo il lettore ai testi di Teoria della Probabilità). Notiamo che 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.

Una condizione elementare di consistenza

È chiaro che il tipo di consistenza più desiderabile dal punto di vista statistico è quello forte. Questa è d'altra parte difficile da provare in generale. Viceversa, la disuguaglianza di Chebyshev permette di provare la convergenza in probabilità a partire semplicemente dalla convergenza di medie e varianze, usando la classica disuguaglianza,

$$P_\theta\Big(|\phi_n - \theta| \geq \varepsilon\Big) \leq \frac{1}{\varepsilon^2}\, E_\theta\left[(\phi_n - \theta)'\,(\phi_n - \theta)\right] = \frac{1}{\varepsilon^2}\, E_\theta|\phi_n - \theta|^2\,, \qquad (6.4)$$

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 pertanto 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 consistente. Quindi,

Proposizione 6.1. *Se $\phi_n(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ è, per ogni n, uno stimatore corretto e se la sua varianza scalare $\sigma_n^2(\theta)$ tende a zero con n per ogni $\theta \in \Theta$, $\phi_n(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ è consistente.*

Quasi tutti i risultati relativi alla consistenza forte fanno viceversa riferimento alla cosiddetta *Legge forte dei grandi numeri* o più in generale al *Teorema Ergodico* di Birkhoff, di cui ci occuperemo tra un momento.

Esempio 13.4

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 la stessa distribuzione di $\mathbf{y}$ e pertanto la probabilità

$$p\Big(|\bar{\mathbf{y}}_n - \theta_0| \geq \varepsilon\Big)$$

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 $E_0\mathbf{y} = \infty$!).

6.2 Processi Ergodici

Si potrebbe ben affermare che i due teoremi veramente fondamentali della teoria della probabilità sono il *teorema ergodico* e il *teorema del limite centrale*. Questi due teoremi, che in realtà sono noti in varie forme e a vari livelli di generalità, 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 proprio 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.

Studieremo in questa sezione una questione che è intimamente legata al problema generale dell'inferenza statistica cui abbiamo accennato all'inizio del capitolo . In termini generali, il problema è il seguente:

Problema 6.1. *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 cercare di rispondere alla seguente domanda: Che cosa si può dire della legge di probabilità* P *del processo in base alla conoscenza della traiettoria* $\{\bar{y}(t)\}$ *?*

Più avanti introdurremo una classe di processi per cui questo problema ha una soluzione ben definita. Questi processi sono i processi *ergodici*. Per definire l'ergodicità occorre innanzitutto ricordare la nozione di processo stazionario in senso stretto.

Definizione 6.2. *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 immediate e 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*

$$\Sigma(t_1,t_2) := E\left[\mathbf{y}(t_1)-\mu(t_1)\right]\,\left[\mathbf{y}(t_2)-\mu(t_2)\right]'$$

dipende solo dalla distanza temporale $\tau = t_1 - t_2$.

Dato un processo strettamente stazionario $\{\mathbf{y}(t)\}$, si può definire una intera classe di processi, ancora strettamente stazionari, che sono "funzioni di $\{\mathbf{y}(t)\}$", nel modo seguente.

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 delle variabili aleatorie:

$$\mathbf{z} = f(\mathbf{y}(\tau)\ ;\ \tau \in \mathbb{I}) \quad , \tag{6.5}$$

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 semplificare le notazioni, noi supporremo in questo paragrafo che f (e quindi $\mathbf{z}$) prenda solo valori reali. La generalizzazione della teoria a funzioni vettoriali (e matriciali) che si useranno nel seguito è semplice e verrà lasciata al lettore.

Prenderemo in considerazione solo funzioni f a media finita, tali per cui $E|f(\mathbf{y}(\tau)\ |\ \tau \in \mathbb{I})\,| = E\,|\mathbf{z}| < \infty$. Denoteremo inoltre 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).

Sia ora

$$\mathbf{z}(t) := f\Big(\mathbf{y}(t+\tau)\ ;\ \tau \in \mathbb{I}\Big) \quad , \qquad t \in \mathbb{Z} \quad , \tag{6.6}$$

la variabile casuale che si ottiene "traslando" le variabili $\{\mathbf{y}(\tau)\}$ nell'argomento di f di t unità temporali. Chiaramente, al variare di t in $\mathbb{Z}$, la variabile $\mathbf{z}(t)$ descrive ancora un processo stocastico (scalare) $\{\mathbf{z}(t)\}$ che si riconosce immediatamente essere *stazionario in*

senso stretto. Ad esempio, per la stazionarietà di $\{\mathbf{y}(t)\}$, si ha

$$\begin{aligned} P\{\mathbf{z}(t) \in A\} &= P\Big\{ f(\mathbf{y}(t+\tau)\ ;\ \tau \in \mathbb{I}) \in A \Big\} \\ &= P\Big\{ (\mathbf{y}(t+\tau_1), \ldots, \mathbf{y}(t+\tau_N)) \in f^{-1}(A) \Big\} \\ &= P\Big\{ (\mathbf{y}(\tau_1), \ldots, \mathbf{y}(\tau_N)) \in f^{-1}(A) \Big\} \\ &= P\{\mathbf{z} \in A\} \quad , \end{aligned} \tag{6.7}$$

dove con $\tau_1, \ldots, \tau_N$ si sono indicati gli elementi di $\mathbb{I}$ e il simbolo $f^{-1}(A)$ è l'antiimmagine dell'insieme A attraverso f, cioè $f^{-1}(A) := \{x_1, \ldots, x_N \mid f(x_1, \ldots, x_N) \in A\}$. Con un ragionamento analogo si può dimostrare l'invarianza temporale delle distribuzioni congiunte del processo $\{\mathbf{z}(t)\}$ di ordine qualunque.

La definizione astratta di ergodicità

Ricordiamo la definizione dell'operatore di *traslazione temporale* U sulle variabili del processo $\mathbf{y}$, definito tramite la posizione

$$U\mathbf{y}(t) = \mathbf{y}(t+1)$$

L' operatore U può essere esteso per linearità a tutte le variabili $\mathbf{z} \in L^1(\mathbf{y})$ semplicemente ponendo, se $\mathbf{z} = f(\mathbf{y}(\tau) \mid \tau \in \mathbb{I})$,

$$U\mathbf{z} = f(\mathbf{y}(\tau+1) \mid \tau \in \mathbb{I}) = \mathbf{z}(1)$$

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 $L^1(\mathbf{y})$ i quali, per ogni $\mathbf{z} \in L^1(\mathbf{y})$ e $t \in \mathbb{Z}$, traslano in avanti di t unità di tempo la variabile aleatoria $\mathbf{z}$,

$$U^t\mathbf{z} := \mathbf{z}(t) \quad , \tag{6.8}$$

dove $\mathbf{z}(t)$ è definita dalla formula (6.6).

Sia $\{\mathbf{z}_k\}$ una successione convergente in $L^1(\mathbf{y})$. Dato che $E\,|U^t(\mathbf{z}_n - \mathbf{z}_m)| = E\,|\mathbf{z}_n - \mathbf{z}_m|$ si può facilmente vedere che U^t è una trasformazione continua rispetto alla convergenza in media in $L^1(\mathbf{y})$.

Osservazione 6.1. 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 $H(\mathbf{y})$.

In analogia a quanto fatto per lo spazio dei funzionali lineari definiamo qui i sottospazi della storia "passata e "futura di $\mathbf{y}$ all'istante t,

$$L_t^-(\mathbf{y}) := \{\mathbf{z} \mid \mathbf{z} \in L^1(\mathbf{y}),\, \mathbb{I} \subset (-\infty.t]\} \quad L_t^+(\mathbf{y}) := \{\mathbf{z} \mid \mathbf{z} \in L^1(\mathbf{y}),\, \mathbb{I} \subset [\,t, +\infty)\} \tag{6.9}$$

Non è difficile convincersi che $L_t^-(\mathbf{y})$ è un sottospazio chiuso di $L^1(\mathbf{y})$ e 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:

$$L_\infty^-(\mathbf{y}) := \cap_{t \le k} L_t^-(\mathbf{y}) \qquad L_\infty^+(\mathbf{y}) := \cap_{t \ge k} L_t^+(\mathbf{y}) \tag{6.10}$$

In queste relazioni la scelta dell'istante iniziale k è irrilevante dato che le successioni di sottospazi in oggetto sono entrambe monotone.

Definizione 6.3. *La variabile casuale* $\mathbf{z} \in L^1(\mathbf{y})$ *è* invariante *per* U *se*

$$U\mathbf{z} = \mathbf{z} \quad . \tag{6.11}$$

Dalla definizione segue immediatamente che $\mathbf{z}$ è invariante se e solo se

$$\mathbf{z}(t) = U^t \mathbf{z} = \mathbf{z} \quad , \qquad \forall t \in \mathbb{Z} \quad , \tag{6.12}$$

per cui $\mathbf{z} = f(\mathbf{y}(\tau)\ ;\ \tau \in \mathbb{I})$ 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 quindi una costante deterministica, oppure dipende solo dal "comportamento asintotico" di $\{\mathbf{y}(t)\}$ nell'intorno di $\pm\infty$. In altri termini, una variabile invariante (non banale) può solo dipendere dalla "coda" infinitamente futura o infinitamente remota del processo $\{\mathbf{y}(t)\}$. Vale in effetti il seguente risultato

Teorema 6.1. *Le variabili aleatorie invarianti formano un sottospazio (chiuso) di* $L^1(\mathbf{y})$ *e son sempre contenute nei sottospazi passato e futuro remoto, i.e.*

$$L_\infty(\mathbf{y}) \subseteq L_\infty^-(\mathbf{y}) \cap L_\infty^+(\mathbf{y})\,. \tag{6.13}$$

La dimostrazione si può trovare in [23, Lemma 6.1 p.162].

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 generale di quella di processo p.n.d. che si riferisce a sottospazi di $L^2(\mathbf{y})$ generati linearmente dalle variabili del processo. Vedremo più avanti che *un processo p.n.d. in senso stretto è ergodico*.

Esempio 1.1

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. Chiaramente un discorso perfettamente analogo può essere fatto per le variabili

$$\limsup_{t\to\pm\infty}\ \mathbf{y}(t) \quad , \qquad \liminf_{t\to\pm\infty}\ \mathbf{y}(t) \quad .$$

Esempio 1.2

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 facile allora mostrare che $\{\mathbf{y}(t)\}$ è un processo strettamente stazionario. Inoltre, 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.

Definizione 6.4. *Il processo $\{\mathbf{y}(t)\}$ si dice* ergodico *se le uniche variabili aleatorie invarianti per U sono le costanti (deterministiche). In altri termini $\{\mathbf{y}(t)\}$ è ergodico se il sottospazio delle variabili invarianti $L_\infty(\mathbf{y})$ definito in* (6.13) *contiene solo variabili aleatorie costanti (con probabilità 1).*

Come vedremo più avanti i processi ergodici debbono essere molto irregolari. Un classico esempio di processo ergodico è per esempio un processo $\{\mathbf{y}(t)\}$ a variabili i.i.d. (indipendenti e identicamente distribuite). La prova dell'ergodicità di un processo i.i.d. si ottiene facilmente ragionando sulla σ-algebra degli eventi infinitamente futuri o infinitamente remoti del processo (Questi eventi possono avere solo probabilità zero o uno, questa è la *legge dello 0-1* di Kolmogorov). Noi però qui rinunceremo ad approfondire questi aspetti della teoria rimandando alla letteratura, ad es. al testo di Rozanov [23], Esempio 6.1 p. 162.

Una conseguenza interessante dell'ergodicità è che *un processo ergodico non può ammettere limite per $t \to \pm\infty$* a meno che tutte le variabili del processo non si riducano

a delle costanti. 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.

Una catena di Markov è ergodica se e solo se essa ammette un'unica classe ergodica. Solo in questo caso infatti le uniche v.a. invarianti sono costanti (si veda l'Esempio 1.2 presentato più sopra).

Scende dalla definizione che il processo $\{\mathbf{z}(t)\}$ ottenuto traslando una arbitraria funzione $\mathbf{z} = f(\mathbf{y} \in L^1(\mathbf{y})$ di un processo ergodico è ancora ergodico. 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 6.2 (Doob). *Sia $\{\mathbf{y}(t)\}$ un processo i.i.d. a media zero e varianza finita. Si assuma che la successione di numeri reali $\{c_k\}$ sia a quadrato sommabile,*

$$\sum_{-\infty}^{+\infty} c_k^2 < \infty \quad ; \tag{6.14}$$

allora il processo $\{\mathbf{z}(t)\}$ definito dalla

$$\mathbf{z}(t) := \sum_{-\infty}^{+\infty} c_k \, \mathbf{y}(t+k) \tag{6.15}$$

è (strettamente stazionario), a varianza finita ed ergodico.

Una dimostrazione si può trovare nel trattato di Doob [7, p. 460].

Il teorema ergodico

Enunciamo ora uno dei risultati centrali della teoria della probabilità, il cosiddetto *teorema ergodico* di G.D. Birkhoff [3].

Sia $\mathbf{z} = f(\mathbf{y}) \in L^1(\mathbf{y})$ una funzione del processo stazionario $\{\mathbf{y}(t)\}$. Denotiamo, per maggiore evidenza, la traslazione temporale di $\mathbf{z}$ di t unità di tempo, col simbolo

$$\mathbf{z}(t) = U^t \mathbf{z} = U^t f(\mathbf{y}) := f_t(\mathbf{y}) \quad .$$

Teorema 6.3 (Teorema Ergodico). *Sia $\{\mathbf{y}(t)\}$ un processo strettamente stazionario. Il limite*

$$\lim_{T\to\infty} \frac{1}{T+1} \sum_{0}^{T} f_t(\mathbf{y}) \tag{6.16}$$

esiste con probabilità uno per tutte le funzioni $f(\mathbf{y}) \in L^1(\mathbf{y})$ ed è una variabile aleatoria 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.

Per la dimostrazione (che è abbastanza complicata) rimandiamo al testo di Rozanov [23, p. 157].

Mostriamo che nella (6.16) si può passare al limite sotto il segno di aspettazione. Per questo basta in realtà una convergenza molto più debole di quella quasi certa. Nella proposizione seguente assumeremo semplicemente che la convergenza sia in legge ($\xrightarrow{L}$).

Proposizione 6.2. *Sia $\{\mathbf{y}(t)\}$ un processo strettamente stazionario e $\mathbf{z} = f(\mathbf{y}) \in L^1(\mathbf{y})$ una funzione del processo. Se $\bar{\mathbf{z}}_T := \frac{1}{T}\sum_{t=1}^{T} \mathbf{z}(t) \xrightarrow{L} \bar{\mathbf{z}}$ allora, per $T \to \infty$,*

$$\mathbb{E}\,\{\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}(t)\} = \mathbb{E}\, f_t(\mathbf{y}) \to \mathbb{E}\,\bar{\mathbf{z}} \tag{6.17}$$

e quindi $\mathbb{E}\,\mathbf{z}(t) = \mathbb{E}\, f_t(\mathbf{y}) = \mathbb{E}\,\bar{\mathbf{z}}$ per ogni t.

Dimostrazione. Come è noto, per poter inferire la convergenza delle aspettazioni da quella in legge occorre e basta la condizione di *integrabilità uniforme* [2], [**?**, p.17], che nel caso stazionario si scrive

$$\lim_{\alpha\to+\infty} \mathbb{E}\,\{\,|f(\mathbf{y})|\, I_{\{|f(\mathbf{y})|>\alpha\}}\} = 0$$

ed è automaticamente soddisfatta perchè, in base alla disuguaglianza di Chebicheff,

$$\mathbb{E}\,\{\,|f(\mathbf{y})|\, I_{\{|f(\mathbf{y})|>\alpha\}}\} = \mathbb{E}\,\{|f(\mathbf{y})|\,\,|\,|f(\mathbf{y})|>\alpha\}P\{|f(\mathbf{y})|>\alpha\} \le \mathbb{E}\,\{|f(\mathbf{y})|\}\frac{\mathbb{E}\,|f(\mathbf{y})|}{\alpha},$$

che tende a zero per $\alpha \to +\infty$. □

Il corollario seguente viene spesso preso come *definizione di ergodicità*.

Corollario 6.1. *Se e solo se $\{\mathbf{y}(t)\}$ è ergodico si ha*

$$\lim_{T\to\infty}\ \frac{1}{T+1}\ \sum_{0}^{T}\ f_t(\mathbf{y}) = \mathbb{E}\ f(\mathbf{y}) \tag{6.18}$$

con probabilità uno, qualunque sia $f(\mathbf{y}) \in L^1(\mathbf{y})$.

Dimostrazione. Denotiamo con $\bar{\mathbf{z}}_T$ il primo membro in (6.18) e con $\bar{\mathbf{z}}$ la variabile invariante limite per $T \to \infty$. Dato che $\mathbb{E}\ \bar{\mathbf{z}}_T = \mathbb{E}\ f(\mathbf{y})$ per ogni T, le aspettazioni delle medie temporali $\bar{\mathbf{z}}_T$ convergono e si ha $\lim_{T\to\infty}\mathbb{E}\ \bar{\mathbf{z}}_T = \mathbb{E}\ f(\mathbf{y}) = \mathbb{E}\,\bar{\mathbf{z}}$ dove l'ultima eguaglianza scende dalla proposizione precedente. Se il processo è ergodico $\bar{\mathbf{z}}$ è una costante e coincide necessariamente con la sua aspettazione per cui $\bar{\mathbf{z}} = E\,\bar{\mathbf{z}} = E\, f(\mathbf{y})$. Viceversa, è possibile mostrare (ma noi qui non lo faremo) che tutte le variabili invarianti si possono costruire come limiti di Cesàro di sequenze del tipo $\{\bar{\mathbf{z}}_T\}$, al variare di $f(\mathbf{y})$ in $L^1(\mathbf{y})$. Quindi, se vale la (6.18), ogni variabile invariante è una costante deterministica. □

Illustreremo ora alcune applicazioni di questo risultato. In quanto verremo dicendo sarà comodo supporre che lo spazio di probabilità $\{\Omega, \mathcal{A}, P\}$, su cui è definito il processo, sia lo spazio campionario di $\{\mathbf{y}(t)\}$.

La prima applicazione risponde alla questione che abbiamo posto all'inizio del capitolo.

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 $z := I_E(\mathbf{y}(0))$ 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:

$$\begin{aligned}\lim_{T\to\infty} \boldsymbol{\nu}_T(E_1, E_2, k) &= E\left[I_{E_1}(\mathbf{y}(0))\; I_{E_2}(\mathbf{y}(k))\right] \\ &= \int_{E_1}\int_{E_2} d\, F(y_1, y_2\; ;\; k) = P\Big\{\mathbf{y}(t) \in E_1\;,\; \mathbf{y}(t+k) \in E_2\Big\}\end{aligned}$$

per "quasi tutte" le traiettorie del processo.

Una generalizzazione ormai facile porge allora la seguente conclusione.

Corollario 6.2. *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.*

Sull'ipotesi ergodica in statistica

Generalmente in un problema di identificazione (o genericamente, in un problema di inferenza statistica) si dispone di una serie temporale di dati (o misure) $\{\bar{y}(t)\}_{t=0,1,\ldots,T}$ che si cerca di descrivere matematicamente per mezzo di un modello probabilistico. Equivalentemente, si può dire che si vuole descrivere matematicamente la serie temporale osservata come un tratto di una *realizzazione di un processo stocastico* $\{\mathbf{y}(t)\}$. In sostanza si formula il problema di inferenza "imponendo che $\{\bar{y}(t)\}_{t=0,\ldots,T}$ sia un'osservazione di una possibile traiettoria di un processo stocastico $\{\mathbf{y}(t)\}$, nell'intervallo temporale $[0, T]$. Ovviamente la legge di probabilità del processo è incognita ed è proprio ciò che si cerca di determinare in base alle misure a disposizione. C'è da rimarcare qui che nella stragrande maggioranza dei casi pratici le osservazioni sono un dato unico e irripetibile, in altri termini è possibile osservare *una sola traiettoria* ed in base a questa si devono, almeno in linea di principio, inferire le distribuzioni di probabilità del processo $\{\mathbf{y}(t)\}$.

Per quanto visto sopra, perchè il problema di inferenza sia ben posto e abbia un'unica soluzione (per $T \to \infty$,) bisogna che il "modello" $\{\mathbf{y}(t)\}$ dei dati osservati $\{\bar{y}(t)\}_{t=0,1,\ldots,T}$ sia un *processo ergodico*. Si fa perciò l' "ipotesi che i dati osservati siano generati da un processo ergodico. Questa ipotesi, anche se matematicamente necessaria, sembra all'atto pratico arbitraria e assai difficile da verificare. C'è una legittima domanda che l'utente ha in mente in queste situazioni:

Come si può fare a verificare se è ragionevole assumere che certi dati misurati siano stati generati da un processo stocastico stazionario? e come si fa a sapere se questo processo è ergodico?

Qui sotto cercheremo di dare una risposta a queste domande.

Innanzitutto, dato che lo scopo dell'identificazione e, in generale dell'inferenza statistica, è quello di produrre modelli che serviranno per descrivere dati "futuri ed in ogni caso dati *diversi* da quelli usati per la loro calibrazione, alle radici di ogni esperimento di modellizzazione deve esserci il fondato convincimento che

> *I dati futuri continueranno a essere generati dallo stesso "meccanismo fisico che ha prodotto i dati attualmente disponibili.*

Questa, per quanto vaga, è un'ipotesi fondamentale che riguarda la natura dei dati futuri, e postula in sostanza che questi debbano continuare a essere "statisticamente simili a quelli disponibili. Essa è inerente allo stesso scopo della raccolta dei dati ai fini di modellizzazione. Se non valesse, il problema di inferenza non avrebbe senso.

Nel contesto astratto della teoria della probabilità, l' "uniformità statistica a cui abbiamo vagamente accennato, corrisponde al concetto di stazionarietà e, come abbiamo visto, la risolubiltà del problema di identificazione, corrisponde all'ergodicità.

Nel contesto di un esperimento reale si hanno a disposizioni solo dei dati di misura. Per descrivere matematicamente la proprietà di uniformità statistica dei dati futuri, bisogna quindi postulare che l'andamento futuro della serie temporale che si osserva sia quello "tipico delle traiettorie di un processo stazionario. La definizione che segue intende definire proprio questo "andamento tipico.

Sia $z := \{z(t)\}_{t\in\mathbb{Z}_+}$ un segnale a tempo discreto (che per semplicità di trattazione supporremo scalare)

Una *funzione di* z è una funzione a valori reali $f(z) := f(z(t); t \in \mathbb{I})$, $f : \mathbb{R}^I \to \mathbb{R}$ dove $\mathbb{I}$ è un sottointervallo di $\mathbb{Z}_+$, possibilmente infinito. L' *operatore di traslazione* σ sui segnali è definito da $[\sigma z](t) := z(t+1), \quad t \in \mathbb{Z}_+$ di modo che l'applicazione iterata di σ, e.g.

$$[\sigma^t z](s) := z(t+s), \quad t, s \in \mathbb{Z}$$

transforma un segnale z nella sua traslazione di t unità di tempo $z_t := \{z(t+s)\}_{s\in\mathbb{Z}}$. Denotiamo con il simbolo $f_t(z)$ il risultato dell'applicazione di f al segnale traslato $\sigma^t z$, i.e. $f_t(z) := f(z(t+s); s \in I) = f(z_t), t \in\in \mathbb{Z}$.

Definizione 6.5. *Un segnale z si dice*

- Stazionario in senso stretto *se il limite in media (di Cesàro)*

$$\lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} f_t(z) \tag{6.19}$$

  esiste per tutte le funzioni f di una classe sufficientemente ampia, ad es, le funzioni limitate (misurabili).

- Stazionario in senso lato *se il limite esiste almeno per $f(z) = z(0)$ (cosicchè $f_t(z) = z(t)$) e per le funzioni quadratiche*[11] *in z.*

È immediato verificare che se esite il limite (6.19) allora, qualunque sia l'istante $t_0 \geq 0$, esiste anche il limite

$$\lim_{T\to\infty} \frac{1}{T+1} \sum_{t=t_0}^{t_0+T} f_t(z)$$

e coincide con quello in (6.19). Questa proprietà è *l'uniformità statistica dei dati futuri*.

I segnali stazionari in senso lato sono quelli per cui la media e la covarianza camponaria basate su una serie temporale di T dati, convergono quando $T \to \infty$. Questa proprietà è la condizione minima di "uniformità statistica dei dati futuri, necessaria per fare una analisi asintotica dei più semplici algoritmi di identificazione.

La nozione di stazionarietà in senso stretto è introdotta per motivi concettuali. Entrambe si generalizzano ovviamente a segnali a valori vettoriali.

La proposizione seguente mostra che ogni segnale stazionario può essere pensato come una traiettoria "rappresentativa di un processo stazionario. Esiste quindi un modello "a urna da cui si può pensare estratta la triettoria z secondo la legge di probabilità di un processo ergodico. In altri termini è legittimo pensare che z sia generata da un processo stocastico (stazionario) *ergodico*.

Proposizione 6.3. *Dato un segnale stazionario (in senso stretto) z, esiste uno spazio di probabilità $\{\Omega, \mathcal{A}, \mu\}$ e un processo stocastico ergodico definito su questo spazio, $\mathbf{z} := \{ z(t,\omega) \,|\, t \in \mathbb{Z}, \omega \in \Omega \}$, tale che z è una traiettoria rappresentativa di $\mathbf{z}$, i.e.*

$$z(t) = \mathbf{z}(t, \bar{\omega}) \quad t \in \mathbb{Z}$$

[11] i.e. cioè per tutte le f tali che $f(\alpha z) = \alpha^2 f(z)$.

per un qualche evento elementare $\bar{\omega}$ appartenente all'insieme di probabilità uno in cui si ha convergenza delle medie campionarie (6.16)*, come stabilito dal teorema ergodico.*

Dimostrazione. * si prenda $f(z) := I_A(z(0))$ dove I_A è la funzione indicatrice di un insieme di Borel $A \subset \mathbb{R}$ ($I_A(x) = 1$ se $x \in A$ e 0 altrimenti). Definiamo la quantità

$$\nu_T(A) := \frac{1}{T+1} \sum_{t=0}^{T} I_A(z(t)) \tag{6.20}$$

che rappresenta la frequenza relativa delle visite del segnale z all'insieme A. Si vede facilmente che per ogni T la funzione $A \to \nu_T(A)$ è una *misura di probabilità*, cioè una funzione d'insieme contabilmente additiva sugli insiemi di Borel della retta reale. Questo segue dalla relazione $I_{\cup A_k} = \sum I_{A_k}$ valida per ogni sequenza di insiemi disgiunti A_k. Per definizione di segnale stazionario, $\nu_T(A) \to \nu_0(A)$ per $T \to \infty$. Passando al limite, si verifica quindi facilmente che

La funzione d'insieme $A \to \nu_0(A)$ è una misura di probabiltà su $\mathbb{R}$.

Questa misura è poi invariante per traslazione, perchè sostituendo a z il segnale $\sigma^s z$ la somma (6.20) diventa

$$\nu_T^s(A) := \frac{1}{T+1} \sum_{t=0}^{T} I_A(z(t+s)) = \frac{1}{T+1} \sum_{t=-s}^{T-s} I_A(z(t))$$

che converge allo stesso limite.

Più in generale prendiamo

$$f(z) := I_{A_1}(z(\tau_1)) \dots I_{A_n}(z(\tau_n))$$

dove $\tau_1 \dots \tau_n$ sono istanti arbitrari e $A_1 \dots A_n$ insiemi di Borel della retta e consideriamo la frequenza relativa

$$\nu_T(A_1, \tau_1, \dots A_n, \tau_n)) := \frac{1}{T+1} \sum_{t=0}^{T} I_{A_1}(z(t+\tau_1)) \dots I_{A_n}(z(t+\tau_n))$$

di visita all' insieme A_1 all'istante τ_1, seguita da una visita, $\tau_2 - \tau_1$ istanti più tardi all'insieme A_2, etc.. e $\tau_n - \tau_1$ istanti più tardi all'insieme A_n. Per la stazionarietà $\nu_T(A_1, \tau_1, \dots A_n, \tau_n) \to \nu_n(A_1, \tau_1, \dots A_n, \tau_n)$ quando $T \to \infty$ e il limite dipende in realtà solo dalle differenze $\tau_2 - \tau_1$,..., $\tau_n - \tau_1$. Quanto appurato finora per la distribuzione di probabilità $\nu_0(A)$, si generalizza quindi nel seguente risultato.

Lemma 6.1. *Per tutti gli n e per arbitrari tempi $\tau_1 \dots \tau_n$, la funzione d'insieme $(A_1 \times \dots \times A_n) \to \nu_n(A_1, \tau_1, \dots A_n, \tau_n)$ è una misura di probabilità su $\mathbb{R}^n$ che è invariante per traslazione. Più precisamente, la famiglia $\{\nu_k\}_{k \in \mathbb{Z}_+}$ è una famiglia di di distribuzioni di probabilità invarianti per traslazione,* consistente nel senso di Kolmogorov*, nel senso che*

$$\nu_n(A_1, \tau_1, \dots, \mathbb{R}, \tau_n) = \nu_{n-1}(A_1, \tau_1, \dots A_{n-1}, \tau_{n-1})$$

per tutti gli insiemi di Borel $A_1, \ldots, A_{n-1}$ e possibili istanti $\tau_1 \ldots, \tau_n$.

Pertanto, per un famoso teorema di Kolmogorov [**?**], esiste una misura di probabilità μ sullo spazio campionario $\mathbb{R}^{\mathbb{Z}}$ delle sequanze reali, che è l'unica estensione della famiglia di distribuzioni finito-dimensionali $\{\nu_k\}_{k\in\mathbb{Z}_+}$ associate al segnale stazionario z mediante la costruzione illustrata. Questa misura è invariante per l'operatore di traslazione σ agente sui segnali di $\mathbb{R}^{\mathbb{Z}}$. In altre parole, $(\mathbb{R}^{\mathbb{Z}}, \mathcal{Z}, \mu)$ (dove $\mathcal{Z}$ è la σ-algebra degli insiemi di Borel) definisce un *processo stocastico stazionario*, **z**. Inoltre le uniche funzioni di z che sono invarianti per traslazione, $f(\sigma^s z) = f(z)$ per ogni s, sono le costanti (con probabilità uno) □

Siamo così autorizzati se vogliamo, a immaginare che un segnale stazionario sia "estratto da una popolazione di altri possibili segnali secondo una legge di probabilità (sullo spazio campionario) di un processo stazionario. Chiameremo la legge di probabilità costruita nella dimostrazione della proposizione 6.3, la **legge di probabilità vera** del processo.

Osservazione 6.2. *Esistono critiche alla formulazione statistica del problema dell'identificazione. Taluni argomentano che in molti casi in cui si applicano tecniche statistiche, ad esempio in economia e in econometria, si è in presenza di un esperimento individuale per sua natura irripetibile, in cui non esistono "urne da cui si sarebbero potute estrarre altre possibili serie temporali diverse da quelle osservate descriventi lo stesso fenomeno. Questa critica è generata dalla confusione (purtroppo assai diffusa) che porta a confondere i modelli matematici con la realtà fisica. In effetti il modello probabilistico a urna (come del resto le equazioni differenziali in fisica) è soltanto una* entità matematica *astratta e non ha molto senso chiedersi se esista "veramente nella realtà. Il modello deve soltanto essere compatibile con i dati e rispondere alla logica e agli assiomi del calcolo della probabilità.*

Quanto abbiamo esposto per segnali stazionari in senso stretto è evidentemente di interesse prevalentemente concettuale. In pratica spesso si possono fare affermazioni verificabili solo sulle statistiche del primo e secondo ordine dei dati e per questo motivo noi faremo normalmente solo l'ipotesi che i dati siano *stazionari in senso debole*. Inoltre supporremo di norma che le medie campionarie siano state debitamente sottratte dalle osservazioni, per cui saremo d'ora in poi autorizzati a supporre che tutte le osservazioni abbiano *medie zero*. Quindi un segnale m-dimensionale stazionario in senso debole sarà una successione z per la quale il limite

$$\lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} z(t+\tau) z(t)' := \Lambda_0(\tau) \tag{6.21}$$

esiste per tutti i $\tau \in \mathbb{Z}$. A questo proposito vale il seguente risultato.

Proposizione 6.4 (Wiener). *Se esiste il limite* (6.21)*, la funzione $\Lambda_0 := \tau \to \Lambda_0(\tau)$ è una funzione covarianza (i.e. una funzione matriciale di τ, definita positiva).*

La dimostrazione si trova (per il caso a tempo continuo) nel celebre articolo [**?**, **?**]. □

Questa covarianza sarà chiamata la **covarianza vera** del processo. In analogia a quanto visto nella proposizione 6.3 ogni segnale stazionario in senso debole si può pensare come una traiettoria rapresentativa di un *processo (del second'ordine) "vero di covarianza* Λ_0.

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*. Per una discussione più approfondita rinviamo il lettore alla letteratura [10, **?**].

L'essenza del metodo è una tecnica per calcolare "sperimentalmente dei valori attesi 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. Il metodo è fondato sull'osservazione che 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, la cosa è immediata. Basta porre

$$d\,F(x) := \frac{1}{b-a}\, dx$$

e ridefinire opportunamente f. Se I è un intervallo infinito, si può usare la stessa tecnica introducendo un'opportuna densità fittizia (ad esempio ponendo $d\,F(x) = e^{-x^2/2}dx$), che andrà poi "scalata" dalla funzione f.

Ci si riduce quindi al calcolo della media, $E\,f(\mathbf{y})$, 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*, di un algoritmo cioè 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 $[0, 1]$). Con terminologia equivalente diciamo che 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 $F(x)$. (La verifica è banale: $P(\mathbf{y} < x) = P(F^{-1}(z) < x) = P(z < F(x)) = F(x)$, se $\mathbf{z}$ è uniformemente distribuita).

Una facile applicazione del teorema ergodico porge quindi:

$$\lim_{N\to\infty} \frac{1}{N} \sum_1^N f(y_k) = E\,f(\mathbf{y}) = \int_I f(x)\, d\,F(x)$$

e questa formula fornisce un metodo generale per il calcolo approssimato di aspettazioni o di integrali definiti. Si possono anche dare "intervalli di confidenza" che esprimono (in modo probabilistico) l'errore di approssimazione con N finito (si vedano il testi precedentemente citati).

Un problema che si incontra spesso è la difficoltà di calcolare esplicitamente l'inversa della distribuzione di probabilità. In questi casi si può ricorrere ad una densità di probabiltà ausiliaria scelta in modo opportuno. Si supponga ad esempio di dover calcolare l'integrale $\int_{\mathbf{I}} f(x)p(x)dx$ dove $p(x)$ è una densità complicata per cui il metodo descritto prima è difficile da applicare. Si può allora costruire una opportuna densità ausiliaria $q(x)$ diversa da zero e 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, come si suol dire, "campionare" la densità q ma anche 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*.

Il famoso articoli [**?**] e [**?**] hanno portato ad un notevole progresso nel campo della simulazione Montecarlo. L'idea di base è stata la generalizzazione dell' impiego del teorema ergodico da successioni di variabili i.i.d. a catene o processi di Markov (a tempo discreto). 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 [**?**] 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 innavicinabili.

Notiamo per ultimo che la qualificazione "con probabilità 1" che va associata alla formula (6.18) è molto più di effetto psicologico che reale.

Processi p.n.d. e processi dissolventi (mixing)

La definizione di ergodicità di un processo (stazionario) data più sopra è espressa in termini astratti e lascia poco spazio all'intuizione. Cercheremo qui di esprimerla in termini un tantino più concreti.

Siano $f(\mathbf{y})$ e $g(\mathbf{y})$ due funzioni (invarianti nel tempo) del processo le quali, oltre a stare in $L^1(\mathbf{y})$, abbiano momenti del secondo ordine finiti. Questa condizione garantisce ovviamente che, $|E\, f(\mathbf{y})\; g(\mathbf{y})| \leq E\, f^2(\mathbf{y})^{1/2} \cdot E\, g^2(\mathbf{y})^{1/2} < \infty$. Si vede immediatamente che l'ergodicità di $\{\mathbf{y}(t)\}$ implica, in base al teorema ergodico, che

$$\lim_{T\to\infty}\; \frac{1}{T+1}\; \sum_{t=0}^{T}\; f_t(\mathbf{y})\; g(\mathbf{y}) = [E\, f(\mathbf{y})]\;\; g(\mathbf{y}) \tag{6.22}$$

con probabilità uno, qualunque siano f e g nella classe appena descritta.

Sceglieremo f e g nel seguente modo:

$$f\Big(\mathbf{y}(t_1)\ldots\mathbf{y}(t_n)\Big) := I_{A_1}(\mathbf{y}(t_1))\ldots I_{A_n}(\mathbf{y}(t_n))$$

$$g\Big(\mathbf{y}(\tau_1)\ldots\mathbf{y}(\tau_m)\Big) := I_{B_1}(\mathbf{y}(\tau_1))\ldots I_{B_m}(\mathbf{y}(\tau_m)) \quad ,$$

dove $A_1,\ldots,A_n$ e $B_1,\ldots,B_m$ sono sottoinsiemi (di Borel) dalla retta reale (ad esempio intervalli) e I_{A_k}, I_{B_j} sono le relative funzioni indicatrici. Dato che f_t e g sono limitate si può passare al limite sotto il segno di aspettazione in (6.22), ottenendo

$$\lim_{T\to\infty}\; \frac{1}{T+1}\; \sum_{t=0}^{T}\; E\,[f_t(\mathbf{y})\cdot g(\mathbf{y})] = E\, f(\mathbf{y})\cdot E\, g(\mathbf{y}) \quad ,$$

la quale di può scrivere

$$\lim_{T\to\infty}\; \frac{1}{T+1}\; \sum_{t=0}^{T}\; P(A_t\cap B) = P(A)P(B) \quad , \tag{6.23}$$

dove si è usata la notazione

$$A_t := \{\omega\;;\; \mathbf{y}(t+t_1)\in A_1,\ldots,\mathbf{y}(t+t_n)\in A_n\}$$

e analoga per B. Con un'ovvia estensione della terminologia si può dire che A_t è il "traslato" di t unità temporali dall'evento $A = \{\omega\;;\; \mathbf{y}(t_1)\in A_1,\ldots,\mathbf{y}(t_n)\in A_n\}$.

Ricordiamo che per la stazionarietà del processo si ha $P(A_t) = P(A)$ qualunque sia t (dato che l'operatore di traslazione preserva la probabilità degli eventi) per cui al secondo membro di (6.23) si può ugualmente scrivere $P(A_t)P(B)$ al posto di $P(A)P(B)$. È possibile allora interpretare la (6.23) come una condizione di *indipendenza asintotica* delle variabili del processo. Infatti se il processo è ergodico, la probabilità congiunta degli eventi A_t e B converge per $t\to\infty$, nel senso delle medie di Cesàro, al *prodotto* delle probabilità $P(A_t)P(B)$.
Chiaramente questo significa che A_t e B tendono a diventare *indipendenti* quando $t\to+\infty$.

Il tipo di convergenza è però molto debole. In un certo senso, una relazione del tipo

$$\lim_{t\to\pm\infty}\; P(A_t\cap B) = P(A)P(B) \tag{6.24}$$

sarebbe interpretabile in modo molto più diretto. Come vedremo, questa condizione ci riporta alla nozione di processo p.n.d. definita più sopra.

Sul concetto di σ-algebra

In questa sezione è necessario usare il concetto di σ-algebra di eventi in modo un poco più approfondito di quanto abbiamo fatto finora. Vogliamo comunque evitare di infilarci nelle sottigliezze della teoria della probabilità astratta, e cercheremo quindi di dare un'idea "operativa" (anche se non proprio rigorosa) di questo concetto di modo che il lettore possa seguire i ragionamenti che seguono senza troppe sofferenze.

Definizione 6.6. *Un insieme di variabili aleatorie reali, $\mathcal{F}$, in uno spazio di probabilità, è un aggregato* completo *di variabili aleatorie se $\mathcal{F}$, è tale che ogni limite puntuale di funzioni misurabili delle variabili in $\mathcal{F}$ appartiene ancora a $\mathcal{F}$. Diremo equivalentemente che $\mathcal{F}$ è una σ-*algebra*. Le variabili aleatorie in $\mathcal{F}$ sono anche chiamate $\mathcal{F}$-misurabili.*

Il più piccolo aggregato completo che contiene una data famiglia di v.a. reali $\{f_\alpha\,;\,\alpha\in A\}$, si chiama la σ-algebra generata da $\{f_\alpha\,;\,\alpha\in A\}$. *Se la funzione indicatrice, I_E, di un evento E appartiene ad $\mathcal{F}$, diremo convenzionalmente che l'evento E stesso appartiene a $\mathcal{F}$.*

Introduciamo le σ-algebre degli eventi passati, $\mathcal{Y}_t^-$, e futuri, $\mathcal{Y}_t^+$, del processo $\mathbf{y}$ all'istante t, come le σ-algebre generate, rispettivamente dalle variabili passate $\{\mathbf{y}(s)\,;\,s\le t\}$ e future, $\{\mathbf{y}(s)\,;\,s\ge t\}$ all'istante t. Si potrebbe dimostrare che i sottospazi $L_t^-(\mathbf{y})$ e $L_t^+(\mathbf{y})$ contengono esattamente tutte le variabili aleatorie (scalari) che sono integrabili e sono rispettivamente, $\mathcal{Y}_t^-$- e $\mathcal{Y}_t^+$-misurabili.

Conveniamo di dire che l'evento $A=\{\omega\;;\;\mathbf{y}(t_1)\in A_1,\dots,\mathbf{y}(t_n)\in A_n\}$, dove $A_1,\dots,A_n$ sono sottoinsiemi (di Borel) di $\mathbb{R}^m$, è un *evento passato all'istante* t se $t_1,t_2,\dots,t_n\le t$ ed è invece un *evento futuro all'istante* t se, viceversa, $t_1,t_2,\dots,t_n\ge t$. Ne segue che una nozione equivalente di processo p.n.d. è un processo per cui le σ-algebre degli eventi *infinitamente passati* e *infinitamente futuri*

$$\mathcal{Y}_\infty^-:=\cap_{t\le k}\mathcal{Y}_t^-,\qquad \mathcal{Y}_\infty^+:=\cap_{t\ge k}\mathcal{Y}_t^+ \tag{6.25}$$

sono *banali*; i.e. contengono solo funzioni costanti, ovvero solo gli eventi $\{\emptyset,\Omega\}$ (oltre, se si vuole, a tutti gli insiemi di probabilità zero).

Teorema 6.4. *Sia B un qualunque evento relativo al processo $\mathbf{y}$. Per ogni successione di eventi $A_t\in\mathcal{Y}_t^-$, vale la relazione limite*

$$\lim_{t\to-\infty}|\,P(A_t\cap B)-P(A_t)P(B)\,|=0, \tag{6.26}$$

se e solo se $L_\infty^-(\mathbf{y})$ contiene solo v.a. costanti. Dualmente, per ogni successione di eventi futuri, $A_t\in\mathcal{Y}_t^+$, la

$$\lim_{t\to+\infty}|\,P(A_t\cap B)-P(A_t)P(B)\,|=0, \tag{6.27}$$

vale se e solo se $L_\infty^+(\mathbf{y})$ contiene solo v.a. costanti. Le (6.26) (6.27) *valgono contemporaneamente se e solo se il processo $\mathbf{y}$ è p.n.d. (in senso stretto).*

Dimostrazione. È sufficiente occuparci solo della (6.26), dato che l'altra relazione è esattamente duale. Se prendiamo $A_t \equiv A = B \in \mathcal{Y}_{\infty}^{-} \subset \mathcal{Y}_t^{-}$ il passaggio al limite è superfluo e la (6.26) si riduce alla , $| P(A) - P(A)^2 | = 0$ che significa che ogni evento $A \in \mathcal{Y}_{\infty}^{-}$ ha probabilità zero o uno, ovvero la σ-algebra degli eventi infinitamente passati del processo è banale.

Viceversa, data una successione arbitraria di eventi $A_t \in \mathcal{Y}_t^{-}$, consideriamo le variabili aleatorie

$$\boldsymbol{\xi}_t := I_{A_t} - P(A_t), \qquad \boldsymbol{\eta} := I_B - P(B)$$

che appartengono entrambe allo spazio (di Hilbert) $L^2(\mathbf{y})$ e definiamo la proiezione ortogonale di $\boldsymbol{\eta}$ sul sottospazio passato $L_t^{2-}(\mathbf{y}) \subset L_t^{-}(\mathbf{y})$ delle funzioni (causali) del processo $\mathbf{y}$ che hanno momento secondo finito:

$$\boldsymbol{\eta}_t := P\{\boldsymbol{\eta} \mid L_t^{2-}(\mathbf{y})\} = E\{\boldsymbol{\eta} \mid \mathcal{Y}_t^{-}\}$$

Evidentemente, dato che $\boldsymbol{\xi}_t \in L_t^{2-}(\mathbf{y})$, si ha

$$|\langle \boldsymbol{\xi}_t, \boldsymbol{\eta} \rangle| = |\langle \boldsymbol{\xi}_t, \boldsymbol{\eta}_t \rangle| \leq \|\boldsymbol{\xi}_t\|_2 \|\boldsymbol{\eta}_t\|_2$$

dove le norme sono quelle di $L^2(\mathbf{y})$. Sostituendo le espressioni delle variabili al primo membro si trova

$$|P(A_t B) - P(A_t)P(B)| \leq (P(A_t) - P(A_t)^2)\, \|\boldsymbol{\eta}_t\|_2 \leq \|\boldsymbol{\eta}_t\|_2 \,. \qquad (*)$$

Ora, dato che per $t \to -\infty$, $L_t^{2-}(\mathbf{y})$ tende monotonicamente a restringersi ad un sottospazio che contiene solo costanti (equivalentemente, la σ-algebra $\mathcal{Y}_t^{-}$ tende alla σ-algebra banale), si ha

$$\lim_{t \to -\infty} \boldsymbol{\eta}_t = E\, \boldsymbol{\eta} = 0$$

e quindi anche $\lim_{t \to -\infty} |P(A_t B) - P(A_t)P(B)| = 0$. □

Osservazione 6.3. Dato che la maggiorazione nella disuguaglianza (*) non dipende da A_t, il limite è uniforme rispetto alla sequenza $\{A_t\}$ che si considera, il che si può esprimere dicendo che l'estremo superiore rispetto ad A_t di $|P(A_t B) - P(A_t)P(B)|$ tende a zero con t, ovvero

$$\lim_{t \to -\infty} \sup_{A_t \in \mathcal{Y}_t^{-}} | P(A_t \cap B) - P(A_t)P(B) | = 0 \,. \qquad (6.28)$$

Dualmente, la (6.27) può essere rafforzata nella,

$$\lim_{t \to +\infty} \sup_{A_t \in \mathcal{Y}_t^{+}} | P(A_t \cap B) - P(A_t)P(B) | = 0 \,. \qquad (6.29)$$

Notare che le successioni $\{A_t\}$ possono essere qualsiasi.

Corollario 6.3. *Per un processo p.n.d. la quantità*

$$\alpha(\tau) := \sup_{\substack{A_t \in \mathcal{Y}_t^{-} \\ B_t \in \mathcal{Y}_t^{+}}} | P(A_t \cap B_{t+\tau}) - P(A_t)P(B_{t+\tau}) | \qquad (6.30)$$

(è indipendente da t e) tende a zero per $\tau \to +\infty$.

Dimostrazione. Per la stazionarietà si può fissare t in modo arbitrario (ad esempio $t = 0$) e la (6.30) si può riscrivere in modo equivalente con $A_{t-\tau}$ al posto di A_t e B_t al posto di $B_{t+\tau}$. Se $\alpha(\tau)$ non tendesse a zero per $\tau \to +\infty$, ci sarebbe un qualche $\bar{B} \in \mathcal{Y}_t^+$, di probabilità positiva, per cui

$$\sup_{A_{t-\tau} \in \mathcal{Y}_{t-\tau}^-} | P(A_{t-\tau} \cap \bar{B}) - P(A_{t-\tau})P(\bar{B}) | > 0$$

che è in contrasto con la relazione limite (6.26) del teorema 6.4 per cui si avrebbe $L_\infty^-(\mathbf{y})$ non banale e il proceso non potrebbe essere p.n.d. □

La condizione di indipendenza asintotica di passato e futuro di un processo p.n.d. si ritrova descritta in letteratura sotto nomi diversi. Si dice che un processo (stazionario) $\{\mathbf{y}(t)\}$ che soddisfa la (6.28, 6.29) per arbitrarie sequenze di eventi passati e futuri, è *dissolvente* (in inglese si usa l'attributo "*mixing*"). Spiegare l'origine della denominazione ci porterebbe troppo lontano. Per gli approfondimenti del caso rimandiamo alla letteratura.

È ovvio che un processo p.n.d. è ergodico (ma in generale non viceversa). Di fatto, una riscrittura equivalente della condizione (6.24) è

$$\lim_{t \to \infty} \; E\left[f_t(\mathbf{y})\; g(\mathbf{y})\right] = E\, f(\mathbf{y})\, E\, g(\mathbf{y}) \quad ,$$

dove $f(\mathbf{y})$ e $g(\mathbf{y})$ sono arbitrarie funzioni del processo. Se $\mathbf{z} := f(\mathbf{y})$ è una variabile aleatoria invariante si ha $f_t(\mathbf{y}) = f(\mathbf{y})$, $\forall t$. Prendendo $g(\mathbf{y}) = f(\mathbf{y})$ e sostituendo nella formula si ottiene:

$$E(\mathbf{z}^2) = (E\mathbf{z})^2 \quad ,$$

la quale implica immediatamente che $\mathbf{z}$ è una costante deterministica. La condizione (6.24) implica quindi l'ergodicità.

6.3 Ergodicità del secondo ordine

Come vedremo, l'ergodicità giuoca un ruolo essenziale nell'analisi asintotica degli stimatori. Purtroppo però le ipotesi di stazionarietà stretta su cui si basa sono molto forti e praticamente impossibili da verificare nei casi pratici. C'è una nozione di *ergodicità debole* che è sufficiente per analizzare i casi che si presentano più comunemente nell'analisi asintotica degli algoritmi di identificazione di sistemi lineari.

Definizione 6.7. *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*

$$\bar{\mathbf{y}}_T := \frac{1}{T+1} \sum_0^T \mathbf{y}(t) \tag{6.31}$$

e la varianza campionaria del processo

$$\mathbf{S}_T(\tau) := \frac{1}{T+1} \sum_{t=0}^{T} [\mathbf{y}(t+\tau) - \bar{\mathbf{y}}_T] [\mathbf{y}(t) - \bar{\mathbf{y}}_T] \tag{6.32}$$

convergono ai valori veri

$$\lim_{T\to\infty} \bar{\mathbf{y}}_T = \mu \tag{6.33}$$

$$\lim_{T\to\infty} \mathbf{S}_T(\tau) = \Sigma(\tau) \tag{6.34}$$

con probabilità uno.

È ovvio che per un processo ergodico la (6.33) segue direttamente dal teorema di Birkhoff ponendo $f(\mathbf{y}) = \mathbf{y}(0)$ mentre per provare che vale la (6.34) basta scrivere

$$\begin{aligned}(T+1)\,\mathbf{S}_T(\tau) &= \sum_{0}^{T} [\mathbf{y}(t+\tau) - \mu + \mu - \bar{\mathbf{y}}_T]\ [\mathbf{y}(t) - \mu + \mu - \bar{\mathbf{y}}_T]' \\ &= \sum_{0}^{T} [\mathbf{y}(t+\tau) - \mu]\ [\mathbf{y}(t) - \mu]' - (\bar{\mathbf{y}}_T - \mu) \sum_{0}^{T} [\mathbf{y}(t) - \mu]' \\ &\quad - \left(\sum_{0}^{T} [\mathbf{y}(t+\tau) - \mu] \right) (\bar{\mathbf{y}} - \mu)' + (T+1)\,(\bar{\mathbf{y}}_T - \mu)\,(\bar{\mathbf{y}}_T - \mu)'\,.\end{aligned}$$

e notare che gli ultimi tre termini divisi per $T+1$ tendono a zero con probabilità uno per $T \to \infty$.

Nota Qualche volta il limite superiore nella sommatoria che compare in (6.32) è posto uguale a $T - \tau$.
Questo evidentemente quando si devono calcolare "stime" di $\Sigma(\tau)$ in base ad un campione osservato di $T+1$ misure. Si riconosce facilmente che la (6.34) continua a valere anche con questa diversa definizione della varianza campionaria. ♢

In generale un processo può essere ergodico del second'ordine sotto condizioni più deboli dell'ergodicità. Ricordiamo a questo proposito il teorema 6.2 che si può interpretare dicendo che l'uscita 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 (in senso stretto!). Ovviamente il processo di uscita, essendo ergodico in senso stretto, lo è in particolare anche in senso debole. Nel linguaggio della statistica, la media campionaria $\bar{\mathbf{z}}$ e la varianza campionaria S^2 costruite su un campione del processo $\mathbf{z}$, sono stimatori *fortemente consistenti* della media, $\mathbb{E}\,\mathbf{z}$, e della matrice di varianza $\mathbb{E}\,[(\mathbf{z} - E\mathbf{z})\,(\mathbf{z} - E\mathbf{z})']$ dell'uscita del sistema lineare da cui è estratto il campione.

L'indipendenza dei campioni di ingresso può essere rilassata richiedendo che il processo di ingresso sia stazionario del quart'ordine (con momenti del quart'ordine invarianti

per traslazione). Nel testo di Hannan [14, Cap. IV] e in [15, Cap. 4] si dimostra che l'ergodicità del secondo ordine vale per processi generati come uscita di un filtro lineare stabile che ha in ingresso un processo bianco (debolmente stazionario) $\{\mathbf{e}(t)\}$ che soddisfa a delle condizioni supplementari, del tipo

$$\mathbb{E}\,\|\mathbf{e}(t)\|^4 < \infty \tag{6.35}$$

$$\mathbb{E}\,[\mathbf{e}(t) \mid \mathbf{e}(t-1), \mathbf{e}(t-2), \ldots, \mathbf{e}(t_0)] = 0 \qquad t \in \mathbb{Z} \tag{6.36}$$

$$\mathrm{Var}\,[\mathbf{e}(t) \mid \mathbf{e}(t-1), \mathbf{e}(t-2), \ldots, \mathbf{e}(t_0)] = \Lambda \tag{6.37}$$

che sono in genere più deboli della condizione di essere i.i.d..

Un caso limite in cui le nozioni di ergodicità forte e debole coincidono è quello dei processi Gaussiani per i quali si ha la seguente caratterizzazione.

Teorema 6.5. *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 (cioè valgono le (6.33) e (6.34) con probabilità 1).*
3. *La matrice distribuzione spettrale di potenza del processo, $F(e^{i\omega})$, è una funzione continua di ω in $[-\pi, \pi]$.*
4. $\lim_{T\to\infty} \frac{1}{T+1} \sum_0^T \sigma_{ii}^2(\tau) = 0 \qquad i = 1, 2, \ldots, m \quad .$

Di queste condizioni la più utile è probabilmente la (3), dovuta a Maruyama e Grenander (per una trattazione precisa e completa vedere [23, p. 163]), la quale 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*.

Com'è noto, 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 6.2. *Si consideri il processo $\mathbf{z}(t) = \mathbf{x}\cos\omega_0 t + \mathbf{y}\sin\omega_0 t$, $t \in \mathbb{Z}$, dove $\mathbf{z}$ 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

$$\lim_{\tau\to\infty}\ \sigma_{ii}(\tau) = 0 \quad , \qquad i = 1, 2, \ldots, m \quad , \tag{6.38}$$

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 (6.38) è equivalente alla

$$\lim_{\tau\to\infty}\ \Sigma(\tau) = 0 \quad .$$

6.4 Consistenza dello Stimatore di Massima Verosimiglianza

A titolo di applicazione della teoria ergodica, riportiamo qui il teorema fondamentale di consistenza dello stimatore di massima verosimiglianza. L'enunciato si riferisce al caso in cui si disponga di un *campione casuale*, cioè di misure indipendenti e ugualmente distribuite.

Teorema 6.6 (Wald). *Sia $\{p(\cdot,\theta)\ ;\ \theta\in\Theta\}$ una famiglia parametrica di densità di probabilità su $\mathbb{R}^m$ in cui Θ è un dominio (non necessariamente limitato) di $\mathbb{R}^p$. Sia $\hat\theta(\mathbf{y}^n)$ lo stimatore di M.V. del parametro θ basato sul campione casuale $\mathbf{y}^n := \{\mathbf{y}_1,\ldots,\mathbf{y}_n\}$ estratto dalla distribuzione vera $p(\cdot,\theta_0)$, $\theta_0\in\Theta$.*
Assumiamo le seguenti condizioni:

1. *La famiglia $\{p(\cdot,\theta)\ ;\ \theta\in\Theta\}$ è localmente identificabile in θ_0.*
2. *$p(y,\theta)$ è una funzione continua di θ per ogni y.*
3. *$E_\theta\ \log p(\mathbf{y},\theta')$ è finita per ogni θ e θ' in Θ.*
4. *Al crescere di n la successione delle stime $\hat\theta(y^n)$, $n=1,2,\ldots$, si mantiene limitata con probabilità P_{θ_0} uguale a 1 (ovvero, $|\hat\theta(y^n)|\le M$ per quasi tutte le successioni di valori campionari osservati; M dipende in generale dalla particolare successione osservata).*

In queste ipotesi lo stimatore di M.V. è fortemente consistente, ovvero

$$\lim_{n\to\infty}\ \hat\theta(y^n) = \theta_0 \quad ,$$

per tutte le successioni di valori campionari $\{y_1, y_2, \ldots\}$, eccettuato al più un insieme di probabilità P_{θ_0} uguale a zero.

Notiamo che a stretto rigore potrebbero esserci più punti in Θ in cui la verosimiglianza $L(y^n,\theta)$ è massima e quindi per ogni n più di uno stimatore di M.V.. Il teorema asserisce che tutti questi eventuali stimatori si comportano, per n grande, nello stesso modo e possono quindi essere riguardati come funzioni dei dati asintoticamente coincidenti.

Dimostrazione.

Ricordiamo innanzitutto che la distanza di Kullback $I(\theta_0, \theta)$, fra le densità $p(\cdot, \theta_0)$ e $p(\cdot, \theta)$,

$$I(\theta_0, \theta) = E_{\theta_0} \log \frac{p(\mathbf{y}, \theta_0)}{p(\mathbf{y}, \theta)}$$

è *strettamente positiva* per ogni $\theta \neq \theta_0$, proprio grazie all'ipotesi di identificabilità 1). Se si considera ora un intorno sferico sufficientemente piccolo E di θ non contenente θ_0, è possibile mostrare che la distanza di $p(\cdot, \theta_0)$ dall'insieme di densità $\{p(\cdot, \theta) \ ; \ \theta \in E\}$, definita da

$$I(\theta_0, E) := E_{\theta_0} \left\{ \min_{\theta \in E} \log \frac{p(\mathbf{y}, \theta_0)}{p(\mathbf{y}, \theta)} \right\} \quad , \qquad \theta_0 \notin E \quad ,$$

è ancora *strettamente positiva* (e finita). La cosa segue in sostanza dalla continuità di $p(\cdot, \theta)$. Usando semplici proprietà della funzione $\log(\cdot)$ si trova poi

$$\begin{aligned} I(\theta_0, E) &= E_{\theta_0} \left\{ \log \frac{p(\mathbf{y}, \theta_0)}{\max_\theta \ p(\mathbf{y}, \theta)} \right\} \\ &= E_{\theta_0} \left\{ \log p(\mathbf{y}, \theta_0) - \log [\max_\theta \ p(\mathbf{y}, \theta)] \right\} \\ &= E_{\theta_0} \left\{ \log p(\mathbf{y}, \theta_0) - \max_\theta [\log \ p(\mathbf{y}, \theta)] \right\} > 0 \quad . \end{aligned} \tag{6.39}$$

Sia $L(\mathbf{y}^n, \theta)$ la verosimiglianza del campione. Per la legge forte dei grandi numeri,

$$\frac{1}{n} \log L(\mathbf{y}^n, \theta_0) = \frac{1}{n} \sum_1^n \log p(\mathbf{y}_t, \theta_0) \to E_{\theta_0} \log p(\mathbf{y}, \theta_0) \quad ,$$

con probabilità P_{θ_0} uguale a 1 per $n \to \infty$.

Ugualmente la successione di variabili aleatorie

$$\mathbf{z}_t := \max_{\theta \in E} \log p(\mathbf{y}_t, \theta) \quad , \qquad t = 1, 2, \ldots ,$$

è i.i.d. e per la 3) $E_{\theta_0} \mathbf{z}_t < \infty$. Per la legge forte dei grandi numeri si avrà ancora

$$\lim_{n \to \infty} \frac{1}{n} \sum_1^n \mathbf{z}_t = E_{\theta_0} \max_\theta [\log p(\mathbf{y}, \theta_0)] \quad ,$$

con probabilità 1. La disuguaglianza (stretta!) (6.39) fra i limiti porge allora:

$$\frac{1}{n} \left(\log L(\mathbf{y}^n, \theta_0) - \sum_1^n \max_{\theta \in E} \log p(\mathbf{y}_t, \theta) \right) > 0 \quad ,$$

con probabilità 1 per n sufficientemente grande.

Tenendo infine conto del fatto che la somma dei massimi è maggiore o al più uguale al massimo della somma si trova

$$\log L(y^n, \theta_0) > \max_{\theta \in E} \log L(y^n, \theta) \quad , \qquad \theta_0 \notin E \quad , \tag{6.40}$$

per $n \geq \bar{n}$ opportuno,[12] per quasi tutte le successioni di valori campionari $\{y_1, \ldots\}$. Notiamo ora che se un insieme A (chiuso e limitato) è ricoperto dall'unione di un numero finito di intorni sferici $E_1, \ldots, E_N$ in Θ e se $f(\theta)$ è un'arbitraria funzione continua in A, si ha

$$\max_{\theta \in A} f(\theta) \leq \max_{k} \left\{ \max_{\theta \in E_k} f(\theta) \; ; \; k = 1, \ldots, N \right\} \quad ,$$

cosicché la (6.40) vale anche per un *arbitrario* insieme chiuso e limitato A che non contenga θ_0, ovvero

$$\log L(y^n, \theta_0) > \max_{\theta \in A} \log L(y^n, \theta) \quad , \qquad \theta_0 \notin A \quad , \tag{6.41}$$

qualunque sia l'insieme chiuso e limitato $A \subseteq \Theta$, per quasi tutte le successioni campionarie osservate e pur di prendere n sufficientemente grande.

Consideriamo ora lo stimatore di M.V. $\hat{\theta}(\cdot)$. Dato che $L(y^n, \hat{\theta}(y^n)) = \max_{\theta \in \Theta} L(y^n, \theta)$ si avrà in particolare

$$\log L\Big(y^n, \hat{\theta}(y^n)\Big) \geq \log L(y^n, \theta_0) \quad , \qquad \forall n \quad . \tag{6.42}$$

D'altro canto, per l'ipotesi 4) la successione dei punti $\hat{\theta}_n := \hat{\theta}(y^n)$ in cui si raggiunge il massimo di $\log L(y^n, \theta)$ può essere tutta racchiusa in un sottoinsieme chiuso e limitato C di Θ (che dipende in generale dalla successione di dati osservata). Questo naturalmente a meno di casi "sfortunati" che però hanno probabilità zero. Si ha così

$$\log L\Big(y^n, \hat{\theta}(y^n)\Big) = \max_{\theta \in C} \log L(y^n, \theta)$$

per un opportuno insieme C (che è fisso al variare di n). Consideriamo ora l'intorno sferico di θ_0, $S(\varepsilon) := \{\theta \; ; \; |\theta - \theta_0| < \varepsilon\}$ e definiamo l'insieme $A(\varepsilon) := C - S(\varepsilon)$. Ovviamente per ε abbastanza piccolo $A(\varepsilon)$ è ancora chiuso e limitato. Applicando la disuguaglianza (6.41) si ha così

$$\log L(y^n, \theta_0) > \max_{\theta \in A(\varepsilon)} \log L(y^n, \theta)$$

per tutti gli n maggiori o uguali di un opportuno $n(\varepsilon)$. Usando la (6.42) si perviene allora alla

$$\log L\Big(y^n, \hat{\theta}(y^n)\Big) > \max_{\theta \in A(\varepsilon)} \log L(y^n, \theta)$$

che vale, qualunque sia $\varepsilon > 0$ sufficientemente piccolo, pur di prendere $n \geq \bar{n}(\varepsilon)$. Questa disuguaglianza afferma che, per n sufficientemente grande, il punto di massimo, $\hat{\theta}(y^n)$, della funzione di log-verosimiglianza *deve necessariamente trovarsi nella sfera* $S(\varepsilon)$. In formule,

$$\Big|\hat{\theta}(y^n) - \theta_0\Big| < \varepsilon \qquad \text{per} \qquad n \geq \bar{n}(\varepsilon) \quad ,$$

il che equivale a $\lim_{n \to \infty} \hat{\theta}(y^n) = \theta_0$, naturalmente a meno di un insieme eccezionale di successioni campionarie di probabilità zero. □

[12] È bene mettere in guardia il lettore che tanto più vicino a θ_0 si prende E (ovvero quanto più vicina a $p(\cdot, \theta_0)$ è la famiglia $\{p(\cdot, \theta) \; ; \; \theta \in E\}$) tanto più grande dovrà in generale prendersi $\bar{n}$ per assicurarsi la validità della disuguaglianza (6.40). Per $\theta_0 \in E$ essa potrebbe valere soltanto per "$n = \infty$" e col segno $\geq$, ma in questo caso i due termini che si confrontano non sono più definiti.

Osservazioni

La prova della consistenza dello stimatore di M.V. può essere adattata al caso di misure dipendenti sotto opportune ipotesi di dissolvenza dal processo di misura $\{\mathbf{y}\}$. Nella prova si può utilizzare, anziché la legge forte dei grandi numeri, il teorema ergodico di Birkhoff giungendo a un risultato sostanzialmente analogo.

L'ipotesi 4) non è direttamente verificabile sulla base del modello probabilistico ipotizzato per descrivere i dati. Esistono però delle condizioni verificabili sufficienti a garantire che le stime $\hat{\theta}(y^n)$ "non si disperdano troppo". Una tra le più semplici è la $p(\cdot, \theta) = 0$ per $|\theta| \to \infty$, la quale intuitivamente implica che non possano esservi massimi della funzione di verosimiglianza in punti di norma arbitrariamente grande nello spazio dei parametri. Una dimostrazione del teorema di consistenza che usa condizioni di questo tipo può essere trovata nel trattato di Zacks [27, p. 233].

Sul significato della consistenza

Come si vede, l'idea di consistenza è intimamente legata allo schema Fisheriano. Normalmente l'assunzione che i dati siano effettivamente generati da una distribuzione "vera" appartenente proprio alla classe di modelli probabilistici $\{F_\theta \ ; \ \theta \in \Theta\}$ scelta dallo statistico, è non verificabile e, tranne casi molto circoscritti, anche falsa, dato che la famiglia di modelli viene di solito scelta in base a considerazioni di opportunità e di semplicità matematica. Può così sembrare una nozione di scarso significato.

Viceversa, la consistenza anche in questi casi, rimane una nozione di grande interesse pratico e viene anzi riguardata in statistica come una delle proprietà fondamentali per confrontare diverse *metodologie di stima* (= ricette per costruire stimatori, come ad esempio il principio della massima verosimiglianza, il metodo dei minimi quadrati, i metodi a minimizzazione dell'errore di predizione ecc.). Per comprendere la ragione di questo fatto occorre introdurre uno schema di descrizione dei dati osservati un poco più realistico di quello usato finora.

Supponiamo che la famiglia parametrica $\{F_\theta \ ; \ \theta \in \Theta\}$ sia indiciata (oltre che da θ) da un parametro k a valori naturali che chiameremo *complessità* del modello. Scriviamo $\mathcal{F}_k := \{F_\theta \ ; \ \theta \in \Theta_k\}$ e supponiamo che la successione di modelli $\mathcal{F}_1, \mathcal{F}_2, \ldots$ a complessità crescente abbia le seguenti proprietà:

1) $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$, ovvero ciascun elemento $F_\theta^{(k)}$ può essere ottenuto da un opportuno $F_\theta^{(k+1)} \in \mathcal{F}_{k+1}$ scegliendo opportunamente θ in Θ_{k+1}.

2) Sia F_0 la distribuzione *vera* dei dati osservati.[13] Assumeremo che F_0 possa essere approssimata con accuratezza arbitraria da (almeno) un elemento $F_{\theta_0}^{(k)} \in \mathcal{F}_k$ pur di

[13] Per non complicare troppo le notazioni supporremo qui che i dati osservati $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n, \ldots$ siano una successione di vettori aleatori (m-dimensionali) *indipendenti* aventi comune distribuzione di probabilità F_0 (campione casuale). Più in generale bisognerebbe trattare i dati come *processo stocastico* $\{\mathbf{y}_n\}$ e parlare, anziché di distribuzione di probabilità del generico vettore del campione $\mathbf{y}_n$, di *leggi di probabilità* dell'intero processo $\{\mathbf{y}_n\}$. La definizione di modello approssimato che verrà data più oltre e il Teorema di approssimazione 13.3 possono facilmente essere adattati a questo contesto più generale (che, incidentalmente, è di grande interesse per l'identificazione). Le complicazioni di carattere formale renderebbero però la trattazione molto meno trasparente.

prendere k abbastanza grande. In altre parole, per ogni $\varepsilon > 0$, l'estremo superiore

$$\sup_y \left| F_0(y) - F_\theta^{(k)}(y) \right|$$

può essere reso minore di ε pur di prendere k abbastanza grande e di scegliere opportunamente θ in Θ_k.

Nelle applicazioni che incontreremo più avanti la complessità di $\mathcal{F}$ sarà semplicemente la dimensione dello spazio dei parametri. Notiamo che ciascuna famiglia $\mathcal{F}_k$ potrebbe essere chiamata un *modello approssimato* dei dati.

Sia ora fissata una metodologia di stima ovvero una procedura la quale, fissata la classe parametrica $\mathcal{F}$ di d.d.p., permette, per ogni numerosità campionaria n, di calcolare uno stimatore ϕ_n del parametro θ che indica $\mathcal{F}$. (È bene ribadire che non stiamo qui ipotizzando che i dati osservati siano effettivamente distribuiti secondo $\mathcal{F}$). Alla nostra famiglia di modelli approssimati a complessità crescente $\{\mathcal{F}_k\}$ sarà quindi possibile associare una successione di stimatori $\phi_n^{(k)}$ del parametro $\theta \in \Theta_k$, $k = 1, 2, \ldots$, basati su un campione osservato di dati di numerosità n. Qual è la nozione naturale di consistenza per stimatori basati su modelli approssimati? Una definizione che cattura lo spirito dell'idea Fisheriana è la seguente.

Supponiamo di generare *artificialmente* dei dati (per esempio tramite simulazione al calcolatore) distribuiti secondo la legge $F_{\theta_0}^{(k)} \in \mathcal{F}_k$. In questo caso la distribuzione vera appartiene *per costruzione* alla classe di modelli in gioco. *Diremo che $\phi_n^{(k)}$ è intrinsecamente consistente se con dati (simulati) $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$, distribuiti secondo $F_{\theta_0}^{(k)} \in \mathcal{F}_k$, si ha $\lim_{n\to\infty} \phi_n^{(k)}(\mathbf{y}_1, \ldots, \mathbf{y}_n) = \theta_0$ qualunque sia $\theta_0 \in \Theta_k$.* (Il limite è da intendersi in probabilità $P_{\theta_0}^{(k)}$ o con probabilità $P_{\theta_0}^{(k)}$ uguale a 1).

Notiamo che questa definizione è una specie di condizione di non contraddittorietà logica della procedura di stima: perché $\phi_n^{(k)}$ possa chiamarsi a buon diritto stimatore di θ si richiede che, a partire dai dati (artificiali) generati con distribuzione $F_{\theta_0}^{(k)}$ e in presenza della massima possibile informazione campionaria (campione di numerosità infinita), si abbia $\phi_\infty^{(k)}(\mathbf{y}_1, \ldots, \mathbf{y}_n, \ldots) = \theta$, identicamente.

Chiaramente la consistenza intrinseca (riferita naturalmente a una fissata famiglia di modelli) è una proprietà *verificabile* di uno stimatore, dato che essa è indipendente dalla natura della distribuzione vera dei dati. La questione è ora di chiarire che cosa questa proprietà ci permetta di asserire, nel caso in cui si abbia a disposizione un campione (di lunghezza infinita) di dati *reali*, circa la probabilità F_0 secondo cui questi dati sono effettivamente distribuiti.

L'intuizione suggerisce che usando un metodo di stima intrinsecamente consistente e impiegando un modello parametrico approssimato $\mathcal{F}_k$ di complessità sufficientemente alta si debba riuscire a ricostruire F_0 con buona approssimazione. Un'enunciazione formale di questa proprietà può essere data nel modo seguente.

Teorema 6.7. *Siano $\phi_n^{(k)}(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ stimatori intrinsecamente consistenti dal parametro θ*

basati ciascuno sulla famiglia parametrica $\mathcal{F}_k = \{F_\theta^{(k)} \ ; \ \theta \in \Theta_k\}$, $k = 1, 2, \ldots$, *e operanti su un campione casuale di numerosità* n *estratto dalla distribuzione vera* F_0.

Se $\{\mathcal{F}_k\}$ *è una famiglia a complessità crescente di modelli approssimati dei dati osservati, nel senso che soddisfa alle condizioni 1) e 2) viste in precedenza, e se l'applicazione* $\theta \to F_\theta^{(k)}(\cdot)$ *è continua rispetto alla metrica* $\|F_1 - F_2\| := \sup_y |F_1(y) - F_2(y)|$, *allora*

$$\lim_{k\to\infty} \Big(\lim_{n\to\infty} \|F^{(k)}_{\phi_n^{(k)}(\mathbf{y}_1,\ldots,\mathbf{y}_n)} - F_0\| \Big) = 0 \quad .$$

Se la consistenza di $\phi_n^{(k)}$ *è forte, il limite è con probabilità* P_0 *uguale a 1.*

La prova di questo teorema è abbastanza complicata e non verrà riportata qui. Si può trovare in [**?**].

Capitolo 7
IL TEOREMA DEL LIMITE CENTRALE

7.1 Convergenza in legge

Ricordiamo che 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.

Come è ben noto 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 7.1. *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 $\omega \in \mathbb{R}$[14].

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 nel caso di molte statistiche di interesse costruite su processi stazionari.

Lemma 7.1. *Se* $\{\mathbf{x}_n\}$ *è una successione di variabili aleatorie convergente in legge; i.e.* $\mathbf{x}_n \xrightarrow{L} \mathbf{x}$*, che è uniformemente integrabile, in particolare se*

$$\sup_n \|\mathbf{x}_n\|^2 < \infty \tag{7.1}$$

[14] In realtà la condizione di convergenza delle funzioni caratteristiche è anche sufficiente per (e quindi *equivalente* a) la convergenza in distribuzione (teorema di Helly-Bray).

allora tutti i momenti che esistono delle $\mathbf{x}_n$ convergono ai rispettivi momenti della distribuzione limite.

Per la prova vedere [2, p. 32-33]. Si veda a questo proposito anche l'osservazione 7.3 in margine al teorema 7.3.

Il seguente enunciato raccoglie alcune proprietà generali della convergenza in legge che torneranno utili nel seguito di questo capitolo. Per la dimostrazione si veda [9, Cap. 6].

Teorema 7.1 (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**.

Esempio 7.1. *Sia $\mathbf{y}$ un processo ergodico scalare a media μ, varianza σ^2 per cui vale la $\sqrt{N}\bar{\mathbf{y}}_N \xrightarrow{L} \mathcal{N}(\mu, \sigma^2)$ (come vedremo questa è una forma particolare di teorema del limite centrale). Trovare la distribuzione asintotica della statistica*

$$\varphi(\mathbf{y}) := \frac{\sqrt{N}[\bar{\mathbf{y}}_N - \mu]}{\sqrt{s^2_N(\mathbf{y})}}$$

dove $s^2_N(\mathbf{y})$ è la varianza campionaria

$$s^2_N(\mathbf{y}) \;=\; \frac{1}{N}\sum_{t=1}^{N} (\mathbf{y}(t) - \bar{\mathbf{y}}_N)^2$$

Dal risultato derivare la distribuzione asintotica della statistica di Student

$$t(\mathbf{y}) := \frac{[\bar{\mathbf{y}}_N - \mu]}{\sqrt{s^2_N(\mathbf{y})/N - 1}}\,.$$

Soluzione: Per l'ipotesi di ergodocità $s^2_N(\mathbf{y}) \to \sigma^2$ per $N \to \infty$ (con probabilità uno e quindi anche in probabilità). Usando il teorema di Slutsky (punto 3), si vede subito che

$$\varphi(\mathbf{y}) \xrightarrow{L} \mathcal{N}(0, 1)\,.$$

Ricordiamo che se $\mathbf{y}$ fosse Gaussiano e i.i.d., $s_N^2(\mathbf{y}) \sim \chi^2(N-1)$ e quindi la statistica di Student $t(\mathbf{y})$ avrebbe una distribuzione di Student con $N-1$ gradi di libertà. È noto che al tendere di N all'infinito questa distribuzione tende a una normale. Nel nostro caso si può scrivere

$$t(\mathbf{y}) = \sqrt{\frac{N-1}{N}}\varphi(\mathbf{y})$$

e quindi, sempre per il teorema di Slutsky, anche $t(\mathbf{y})$ ha distribuzione asintotica $\mathcal{N}(0,1)$.

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.

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. È facile in questo caso semplice vedere che la varianza di $\bar{\mathbf{y}}_T$ *tende a zero esattamente come* $\dfrac{1}{T}$. Si ha infatti

$$\mathrm{Var}\,(\bar{\mathbf{y}}_T) = \frac{1}{T}\left(\frac{1}{T}\sum_{t=1}^{T}\mathrm{Var}\,(\mathbf{y}(t))\right) = \frac{1}{T}\Sigma\,.$$

In altri termini, per $T \to \infty$, la varianza di $\sqrt{T}\,\bar{\mathbf{y}}_T$ converge alla varianza di $\mathbf{y}(t)$. Come abbiamo accennato nel capitolo precedente dedicato all'ergodicità, in generale una successione di variabili aleatorie di un processo ergodico non può convergere a una variabile che non sia una costante, e in effetti, visto il risultato precedente, $\sqrt{T}\,\bar{\mathbf{y}}_T$ non può quindi convergere ad una variabile non costante in alcuno dei sensi "usuali della teoria della probabilità. Il fatto notevole però è che ciononostante, la successione delle distribuzioni di probabilità delle variabili $\sqrt{T}\,\tilde{\mathbf{y}}_T$ invece converge e converge ad una distribuzione limite che è Gaussiana. Naturalmente questa distribuzione Gaussiana dovrà avere media zero e varianza Σ. La dimostrazione di questo notevole risultato richiede una semplice espansione attorno a $t = 0$ della funzione caratteristica della convoluzione di T distribuzioni di probabilità uguali e si può trovare in quasi tutti i testi di teoria della probabilità.

Questa semplice versione del teorema del limite centrale (TLC) 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 (soprattutto l'analisi asintotica degli stimatori) e anche per motivi di semplicità espositiva noi in questo capitolo considereremo solo successioni $\{\mathbf{y}(t)\}$ che formano un *processo stazionario*. A noi però interesseranno di norma processi le cui variabili sono dipendenti, che sono del resto quelli che si incontrano quasi sempre quando si descrivono segnali di interesse nell'ingegneria.

Se il processo non è a variabili indipendenti occorrono in generale condizioni speciali perchè valga il teorema del limite centrale. Prima di occuparci (per quanto in modo

superficiale) 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..

7.2 Il teorema del limite centrale per d-martingale stazionarie

Iniziamo questa sezione con una introduzione a questa classe di processi. Nella definizione che segue si può fare riferimento alla nozione "operativa" di σ-algebra data nella definizione 6.6.

Definizione 7.1. *Sia* $\{\mathcal{F}_t\,;\, t \in \mathbb{Z}\}$ *una successione crescente di* σ*-algebre, 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, $\mathbf{z}(t)$ *è funzione delle variabili in* $\{\mathcal{F}_t\}$ *(è* $\mathcal{F}_t$*-misurabile), cosa che scriveremo semplicemente come* $\mathbf{z}(t) \in \mathcal{F}_t\,;\, t \in \mathbb{Z}$*,*
- $\mathbf{z}(t+1)$ *è scorrelata da tutte le variabili in* $\mathcal{F}_t$ *ovvero*

$$\mathbb{E}\left\{\mathbf{z}(t+1) \mid \mathcal{F}_t\right\} = 0 \qquad t \in \mathbb{Z}\,.$$

Se la varianza condizionata $\mathbb{E}\left\{\mathbf{z}(t)\mathbf{z}(t)^\top \mid \mathcal{F}_{t-1}\right\}$ *non dipende da* $\mathcal{F}_{t-1}$*, ovvero*

$$\mathbb{E}\left\{\mathbf{z}(t)\mathbf{z}(t)^\top \mid \mathcal{F}_{t-1}\right\} = \Sigma_{\mathbf{z}} < \infty \tag{7.2}$$

diremo che $\mathbf{z}$ *ha* varianza costante*.*

Notiamo incidentalmente che la prima condizione è equivalente alla

$$\mathbb{E}\left\{\mathbf{z}(t) \mid \mathcal{F}_s\right\} = 0 \qquad \forall\, s < t\,. \tag{7.3}$$

Ovviamente una d-martingala ha sempre media zero.

La nozione di d-martingala è più debole di quella di processo i.i.d.; in effetti, se $\{\mathbf{z}(t)\}$ è i.i.d. e prendiamo per $\mathcal{F}_t$ tutte le funzioni misurabili della storia passata $\mathbf{z}^t$, si verifica subito che le due condizioni della definizione sono banalmente soddisfatte. Però per un processo i.i.d. si ha anche $\mathbb{E}\left\{f(\mathbf{z}(t+1)) \mid \mathcal{F}_t\right\} = 0$ per una arbitraria funzione (integrabile) f e questo non è necessariamente vero per una d-martingala. In questo senso diciamo che le d-martingale sono una classe di processi più generale di quella dei processi i.i.d.. Un esempio "canonico" di d-martingala è l'errore di predizione di un passo di un processo $\mathbf{y}$, quando si intende che la predizione è quella ottima non lineare, ovvero è la media condizionata di $\mathbf{y}(t)$, data la storia passata $\mathcal{Y}_t \equiv \{\mathbf{y}^t\}$,

$$\mathbf{z}(t) := \tilde{\mathbf{y}}(t) = \mathbf{y}(t) - \mathbb{E}\left\{\mathbf{y}(t) \mid \mathcal{Y}_{t-1}\right\} \qquad t \in \mathbb{Z}\,.$$

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 definiamo $\mathcal{F}_t$ come (la σ-algebra generata dal) l' aggregato delle funzioni della storia passata $(\mathbf{y}^t,\, \mathbf{u}^t)$ e poniamo

$$\mathbf{z}(t) := \tilde{\mathbf{y}}(t) = \mathbf{y}(t) - \mathbb{E}\left\{\mathbf{y}(t) \mid \mathbf{y}^{t-1},\, \mathbf{u}^{t-1}\right\} \qquad t \in \mathbb{Z}$$

(dove la media condizionata è ancora intesa in senso stretto). È ancora evidente che, con la definizione di $\mathcal{F}_t$ "estesa", il processo $\mathbf{z}$ soddisfa le due condizioni della definizione 7.1.

Una classe ancora più ampia di d-martingale si ottiene considerando processi del tipo

$$\mathbf{z}(t) := \boldsymbol{\varphi}(t)\, \tilde{\mathbf{y}}(t) \qquad \boldsymbol{\varphi}(t) \in \mathcal{F}_{t-1} \tag{7.4}$$

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

$$\mathbb{E}\left\{\mathbf{z}(t) \mid \mathcal{F}_{t-1}\right\} = \boldsymbol{\varphi}(t)\, \mathbb{E}\left\{\tilde{\mathbf{y}}(t) \mid \mathcal{F}_{t-1}\right\} = 0 \qquad t \in \mathbb{Z}\,.$$

Una *martingala* è l'integrale discreto di una d-martingala,

$$\mathbf{x}(t) = \mathbf{x}(0) + \sum_{s=1}^{t} \mathbf{z}(s) \tag{7.5}$$

ed è un processo non stazionario che è la generalizzazione del processo di passeggiata aleatoria.

Il lemma seguente generalizza alle d-martingale la proprietà "additiva" della varianza di somme di variabili aleatorie indipendenti (o scorrelate).

Lemma 7.2. *Per ogni d-martingala* $\mathbf{z}$ *a varianza finita si ha*

$$\mathrm{Var}\left\{\sum_{t=1}^{T} \mathbf{z}(t)\right\} = \sum_{t=1}^{T} \mathrm{Var}\left\{\mathbf{z}(t)\right\} \tag{7.6}$$

Se la d-martigala è stazionaria il secondo membro vale $T\,\Sigma_{\mathbf{z}}$.

Dimostrazione. Facciamo la dimostrazione per il caso scalare. Il caso vettoriale è identico modulo le ovvie complicazioni nelle notazioni. Si ha

$$\begin{aligned}\mathbb{E}\left\{\sum_{t=1}^{T} \mathbf{z}(t)\right\}^2 &= \mathbb{E}\left\{\mathbf{z}(1)^2 + \mathbf{z}(2)^2 + \ldots + \mathbf{z}(T)^2\right\} + \\ &+ \mathbb{E}\left\{2 \sum_{t>s} \mathbf{z}(t)\mathbf{z}(s)\right\} = \\ &= \sum_{t=1}^{T} \mathrm{Var}\left\{\mathbf{z}(t)\right\} + 2 \sum_{t>s} \mathbb{E}\, \mathbf{z}(t)\mathbf{z}(s)\end{aligned}$$

L'ultimo termine è zero giacchè se $t > s$, $\mathbf{z}(s) \in \mathcal{F}_s$, e si può scrivere

$$\mathbb{E}\, \mathbf{z}(t)\mathbf{z}(s) = \mathbb{E}\left\{\mathbb{E}\left[\mathbf{z}(t)\mathbf{z}(s) \mid \mathcal{F}_s\right]\right\} = \mathbb{E}\left\{\mathbb{E}\left[\mathbf{z}(t) \mid \mathcal{F}_s\right] \mathbf{z}(s)\right\} = 0$$

in virtù della proprietà di d-martingala (7.3). □

Sul risultato seguente, formulato inizialmente da P. Levy e J.L. Doob, dimostrato da Billingsley e Ibragimov [2, **?**] e successivamente generalizzato da vari autori, poggia la dimostrazione della normalità asintotica dei metodi di identificazone PEM.

Teorema 7.2. *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*

$$\sqrt{T}\bar{\mathbf{z}}_T \xrightarrow{L} \mathcal{N}(0, \Sigma_{\mathbf{z}}) \tag{7.7}$$

ovvero, la statistica $\sqrt{T}\bar{\mathbf{z}}_T$ *converge in legge alla distribuzione Gaussiana di media zero e varianza* $\Sigma_{\mathbf{z}}$.

Dimostrazione. Seguiremo la traccia di dimostrazione di J.L. Doob [7, p. 383] per il caso scalare e lascieremo al lettore la generalizzazione al caso vettoriale.

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

$$\begin{aligned}
\phi_T(\omega) &= \mathbb{E}\left\{\mathbb{E}\left[e^{i\omega\mathbf{z}(T)} \mid \mathcal{F}_{T-1}\right] e^{i\omega\mathbf{x}(T-1)}\right\} = \\
&= \left[1 - \frac{\sigma^2\omega^2}{2}\right] \mathbb{E}\, e^{i\omega\mathbf{x}(T-1)} + \mathbb{E}\left\{o(\omega^2)e^{i\omega\mathbf{x}(T-1)}\right\} = \\
&= \left[1 - \frac{\sigma^2\omega^2}{2}\right] \phi_{T-1}(\omega) + \bar{o}(\omega^2)
\end{aligned}$$

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 convincersi 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 $\mathcal{N}(0,\, \sigma^2)$. □

Osservazione 7.1. 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.

7.3 La condizione di Lindeberg e il teorema del limite centrale per variabili dipendenti

Per discutere la validità del teorema del limite centrale per processi stazionari di tipo generale (in cui, in particolare, non si richiede l'indipendenza delle variabili del processo) occorre introdurre una condizione che generalizza la classica *condizione di Lindeberg* che, come è noto, fu introdotta per provare la validità del TLC nel caso di processi a variabili indipendenti[15]

Diamo innanzitutto una condizione sufficiente per la convergenza dei momenti secondi delle somme normalizzate $\sqrt{T}\,\bar{\mathbf{y}}_T$ di un processo stazionario (essendo ovvio che, per la stazionarietà, il momento primo, $\mathbb{E}\,\bar{\mathbf{y}}_T = \mu$ non dipende da T).

Teorema 7.3. *Sia* $\mathbf{y}$ *un processo stazionario* [16] *di media* μ *e matrice varianza finita, con distribuzione spettrale assolutamente continua. Se la matrice densità spettrale* $S_{\mathbf{y}}(e^{j\omega})$ *è una funzione continua in* $\omega = 0$*, allora si ha*

$$\Sigma_\infty := \lim_{T\to\infty} \mathrm{Var}\left(\sqrt{T}\,\bar{\mathbf{y}}_T\right) = S_{\mathbf{y}}(e^{j0}). \tag{7.8}$$

Se $\sqrt{T}\,\bar{\mathbf{y}}_T$ *converge in legge a una distribuzione* $\mathcal{D}$ *che ha varianza finita (in particolare se vale il teorema del limite centrale), allora* Σ_∞ *è la covarianza della distribuzione limite* $\mathcal{D}$*.*

Dimostrazione. Supponiamo senza perdita di generalità che il processo $\mathbf{y}$ abbia media zero. Introduciamo il processo stazionario "mediato" $\mathbf{z}_T$ che si ottiene formalmente attraverso un'operazione di filtraggio, simbolicamente descritta dalla

$$\mathbf{z}_T(t) = \frac{1}{\sqrt{T}}\Big(\sum_{k=0}^{T} z^{-k}\Big)\,\mathbf{y}(t), \qquad \mathbf{z}_T = \mathbf{z}_T(0)\,, \qquad t \in \mathbb{Z}$$

Ovviamente, $\mathrm{Var}\,\{\sqrt{T}\,\bar{\mathbf{y}}_T\} = \mathrm{Var}\,\{\mathbf{z}_T(t)\}$ e calcolando la varianza del processo stazionario $\{\mathbf{z}_T(t)\}$, come integrale del suo spettro, si ha

$$\mathrm{Var}\,\{\sqrt{T}\,\bar{\mathbf{y}}_T\} = \mathrm{Var}\,\{\mathbf{z}_T(t)\} = \frac{1}{T}\int_{-\pi}^{\pi} |\sum_{k=0}^{T} e^{-j\omega k}\,|^2 S_{\mathbf{y}}(e^{j\omega})\,\frac{d\omega}{2\pi}\,.$$

[15] Per una storia dell'evoluzione del problema del limite centrale si può consultare il trattato di M. Loève [**?**, Cap. 6]. La condizione di Lindeberg (e il relativo teorema di convergenza del 1922) è menzionata a p. 280.

[16] Qui a rigore basta la stazionarietà in senso debole.

Si tratta di calcolare il limite di questa quantità per $T \to \infty$. È abbastanza facile vedere, calcolando la sommma della serie geometrica finita di ragione $e^{-j\omega}$, che il limite

$$\lim_{T\to\infty} \frac{1}{T} |\sum_{k=0}^{T} e^{-j\omega k}|^2 = \lim_{T\to\infty} \frac{(\sin \omega T/2)^2}{T^2(\sin \omega/2)^2}$$

si comporta come una funzione δ di Dirac (è integrabile, infinta per $\omega = 0$ e zero per $\omega \neq 0$), per cui, in forza della continuità in $\omega = 0$ della densità spettrale, si ha:

$$\lim_{T\to\infty} \mathrm{var}\,\{\mathbf{z}_T\} = \lim_{T\to\infty} \mathrm{var}\,\{\sqrt{T}\bar{\mathbf{y}}_T\} = S_{\mathbf{y}}(e^{j\omega})_{|\omega=0}.$$

Per dimostrare l'ultima affermazione del teorema, ci avvarremo del lemma 7.1. In particolare verifichiamo che le varianze delle variabili $\{\mathbf{z}_T\,;\, T \geq 1\}$ sono uniformemente limitate. Ma questa verifica è immediata perchè, come abbiamo appena visto, $\lim_{T\to\infty} \mathrm{var}\,\{\mathbf{z}_T\}$ esiste ed è finito e quindi la successione $\{\mathrm{var}\,[\mathbf{z}_T]\,;\, T \geq 1\}$ dev'essere limitata. □

Osservazione 7.2. Il teorema appena dimostrato ha una curiosa interpretazione in termini di analisi di Fourier. In effetti, se $\mathbb{E}\,\mathbf{y}(t)\mathbf{y}(s) := \sigma(t-s)$, posto $\Sigma_T := [\sigma(t-s)\,]_{t,s=1}^{T}$, si ha

$$\mathrm{var}\;\mathbf{z}_T = \frac{1}{T}[1\;1\ldots1]\Sigma_T[1\;1\ldots1]^\top = \sigma(0) + 2\sum_{s=1}^{T-1}(1-\frac{s}{T})\sigma(s)$$

che forma una successione monotona crescente tendente alla somma $\sum_{-\infty}^{+\infty}\sigma(s)$. Dall'enunciato del teorema possiamo ricavare quindi l'eguaglianza

$$\sum_{-\infty}^{+\infty}\sigma(s) = S_{\mathbf{y}}(e^{j\omega})_{|\omega=0} \tag{7.9}$$

che vale se $S_{\mathbf{y}}(e^{j\omega})$ è continua in zero. Questa eguaglianza esprime la *convergenza puntuale* della serie di Fourier di $S_{\mathbf{y}}(e^{j\omega})$ in $\omega = 0$. In altri termini, se $S_{\mathbf{y}}(e^{j\omega})$ è continua in zero,

$$\lim_{N\to\infty}\left[\sum_{s=-N}^{+N} e^{-j\omega s}\sigma(s)\right]_{|\omega=0} = S_{\mathbf{y}}(e^{j\omega})_{|\omega=0}\,.$$

Come è noto, senza l'ipotesi di continuità, si può in generale garantire solo la convergenza in L^1 (!) della serie di Fourier di $S_{\mathbf{y}}$ (ma) solo nel senso delle medie di Cesàro [8]. □

Osservazione 7.3. Come lascia intendere l'enunciato del lemma 7.1, la condizione di limitatezza delle varianze

$$\sup_T \mathrm{var}\,\{\sqrt{T}\bar{\mathbf{y}}_T\} < \infty$$

implica *l'integrabilità uniforme delle medie campionarie normalizzate*. Quest'ultima condizione (assumendo medie nulle e variabili scalari) si esprime nella forma seguente,

$$\lim_{\alpha\to\infty}\sup_T \mathbb{E}\,|\sqrt{T}\bar{\mathbf{y}}_T| I_{\{|\sqrt{T}\bar{\mathbf{y}}_T|>\alpha\}} \;=\; 0 \tag{7.10}$$

Vedere e.g. [2, p. 32-33]. Si può così amplificare leggermente il contenuto del teorema 7.3 dicendo che,

Se $\mathbf{y}$ *ha densità spettrale* $S_{\mathbf{y}}(e^{j\omega})$ *continua in* $\omega = 0$*, allora la successione delle medie campionarie normalizzate* $\{\sqrt{T}\bar{\mathbf{y}}_T\}$ *è uniformente integrabile (i.e. vale la* (7.10)*).*

Nelle stesse ipotesi, se $\sqrt{T}\bar{\mathbf{y}}_T$ *converge in legge a una distribuzione* $\mathcal{D}$ *allora tutti i momenti che esistono di* $\sqrt{T}\bar{\mathbf{y}}_T$ *convergono a quelli corrispondenti della distribuzione limite* $\mathcal{D}$*.* □

Per semplificare ancora un poco le notazioni introduciamo la somma normalizzata e centrata $\mathbf{s}_T := \sqrt{T}\,\tilde{\mathbf{y}}_T := \sqrt{T}(\bar{\mathbf{y}}_T - \mu) = \sqrt{T}\sum_{t=1}^{T}(\mathbf{y}(t) - \mu)$, denotiamo la distribuzione di probabilità di questo vettore aleatorio con $F_{\mathbf{s}_T}(x)$ e con $\Phi(x)$ una generica distribuzione Gaussiana m-dimensionale a media nulla. La condizione di Lindeberg (generalizzata) è una condizione necessaria e sufficiente affinchè $F_{\mathbf{s}_T}(x)$ converga a una distribuzione Gaussiana, $\Phi(x)$, quando $T \to \infty$.

Consideriamo un arbitrario intorno sferico $\Gamma_\alpha := \{x \mid \|x\| \leq \alpha, \}$ dell'origine in $\mathbb{R}^m$ e sia I_{Γ_α} la sua funzione indicatrice. Dalla proposizione 7.1 scende che se $F_{\mathbf{s}_T}(x)$ converge a $\Phi(x)$ allora

$$\lim_{T\to\infty} \int_{\Gamma_\alpha} \|x\|^2 \, dF_{\mathbf{s}_T}(x) = \int_{\Gamma_\alpha} \|x\|^2 \, d\Phi(x)$$

dato che $g(x) := \|x\|^2 I_{\Gamma_\alpha}(x)$ è sicuramente una funzione limitata e i suoi punti di discontinuità sono un insieme di probabilità zero per $\Phi(x)$. Inoltre, se il processo ha densità spettrale continua in $\omega = 0$ e c'è convergenza in legge, la varianza (in particolare la varianza scalare) di $\mathbf{s}_T$ deve convergere a quella di $\Phi(x)$, i.e.

$$\lim_{T\to\infty} \int_{\mathbb{R}^m} \|x\|^2 \, dF_{\mathbf{s}_T}(x) = \int_{\mathbb{R}^m} \|x\|^2 \, d\Phi(x)$$

Sia ora $\bar{\Gamma}_\alpha$ la regione esterna (i.e. l'insieme complementare) all'intorno sferico Γ_α. Le due relazioni limite appena scritte implicano ovviamente che

$$\lim_{T\to\infty} \int_{\bar{\Gamma}_\alpha} \|x\|^2 \, dF_{\mathbf{s}_T}(x) = \int_{\bar{\Gamma}_\alpha} \|x\|^2 \, d\Phi(x)\,. \tag{7.11}$$

Notiamo ora che l'integrale della Gaussiana a secondo membro può essere reso piccolo a piacere pur di prendere α abbastanza grande e quindi lo stesso deve valere per il limite del primo integrale. Questa semplice osservazione è in sostanza il contenuto della condizione di Lindeberg.

Teorema 7.4. *Sia* $\mathbf{y}$ *un processo stazionario in senso stretto, con distribuzione spettrale assolutamente continua e matrice densità spettrale* $S_{\mathbf{y}}(e^{j\omega})$ *continua in* $\omega = 0$*. La d.d.p* $F_{\mathbf{s}_T}(x)$ *converge a una Gaussiana se e solo se il limite (che deve esistere per ogni* $\alpha \geq 0$*)*

$$\phi(\alpha) := \lim_{T\to\infty} \int_{\bar{\Gamma}_\alpha} \|x\|^2 \, dF_{\mathbf{s}_T}(x) = \lim_{T\to\infty} \mathbb{E}\left\{\|\mathbf{s}_T\|^2 I_{\{\|\mathbf{s}_T\|>\alpha\}}\right\} \tag{7.12}$$

tende a zero per $\alpha \to \infty$*.*

La discussione precedente era la dimostrazione che questa condizione è necessaria. Per la prova della sufficienza (che è abbastanza complicata) riamandiamo alla letteratura [23].

Verifichiamo che la condizione del teorema è automaticamente soddisfatta nel caso di processi a variabili i.i.d.. Premettiamo allo scopo il lemma seguente.

Lemma 7.3. *Nello spazio Euclideo* $(\mathbb{R}^m)^N$*, dove* $y_k \in \mathbb{R}^m$ *;* $k = 1, 2, \ldots, N$*, vale la seguente relazione di inclusione tra insiemi*

$$\{\|y_1 + y_2 + \ldots + y_N\| \geq \alpha\sqrt{N}\} \subset \bigcup_{k=1}^{N} \{\|y_k\| \geq \alpha\} \tag{7.13}$$

qualunque sia $\alpha > 0$*.*

Dimostrazione. Mostriamo che la relazione vale per $N = 2$. Nel complementare dell'insieme a secondo membro in (7.13) si ha $\|y_k\|^2 < \alpha^2$; $k = 1, 2$ e quindi vale l'implicazione

$$\|y_1 + y_2\|^2 \leq \|y_1\|^2 + \|y_2\|^2 < 2\alpha^2 \;\Leftarrow\; \{\|y_k\|^2 < \alpha^2\,;\, k = 1, 2\}$$

che, passando ai complementari, è equivalente alla

$$\{\|y_1 + y_2\|^2 \geq 2\alpha^2\} \subset \cup_{k=1}^{2} \{\|y_k\|^2 \geq \alpha^2\}\,.$$

Assumendo allora che la (7.13) valga per $N = n$, mostriamo che vale anche per $N = n+1$. Posto $\bar{y} := \sum_{k=1}^{n} y_k$, si ha

$$\|\bar{y} + y_{n+1}\|^2 \leq \|\bar{y}\|^2 + \|y_{n+1}\|^2 < (n+1)\alpha^2 \;\Leftarrow\; \{\|\bar{y}\|^2 < n\alpha^2\} \cap \{\|y_{n+1}\|^2 < \alpha^2\}$$

e quindi anche

$$\{\|y_1+y_2+\ldots+y_{n+1}\|^2 \geq (n+1)\alpha^2\} \subset \{\|y_1+y_2+\ldots+y_n\|^2 \geq n\alpha^2\} \cup \{\|y_{n+1}\|^2 \geq \alpha^2\}$$

ma per l'ipotesi induttiva il secondo membro è contenuto in $\cup_{k=1}^{n+1}\{\|y_k\|^2 \geq \alpha^2\}$, il che conclude la prova. □

Usando l'inclusione (7.13) del lemma precedente si può maggiorare l'integrale in (7.12) con l'espressione

$$\begin{aligned}
&\frac{1}{T}\mathbb{E}\left\{\|\sum_{t=1}^{T}\mathbf{y}(t)\|^2\, I_{\{\|\sum_{t=1}^{T}\mathbf{y}(t)\|>\alpha\sqrt{T}\}}\right\} \leq \frac{1}{T}\mathbb{E}\left\{\|\sum_{t=1}^{T}\mathbf{y}(t)\|^2\, I_{\cup_{t=1}^{T}\{\|\mathbf{y}(t)\|>\alpha\}}\right\} \\
&= \frac{1}{T}\mathbb{E}\left\{\sum_{t,s=1}^{T}\mathbf{y}(t)^\top\mathbf{y}(s)\left[\sum_{t=1}^{T} I_{\{\|\mathbf{y}(t)\|>\alpha\}} - \sum_{t\neq s} I_{\{\|\mathbf{y}(t)\|>\alpha\}} I_{\{\|\mathbf{y}(s)\|>\alpha\}}\right]\right\} \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\{\|\mathbf{y}(t)\|^2\, I_{\{\|\mathbf{y}(t)\|>\alpha\}}\right\} = \mathbb{E}\left\{\|\mathbf{y}(t)\|^2\, I_{\{\|\mathbf{y}(t)\|>\alpha\}}\right\}\,.
\end{aligned}$$

La seconda eguaglianza segue dall'identità $I_{E_1 \cup E_2 \ldots \cup E_T} = I_{E_1} + I_{E_2} + \ldots + I_{E_T} - \prod_{i \neq k} I_{E_i} I_{E_k}$ e la terza dall' indipendenza delle variabili del processo per $t \neq s$. L'ultima espressione (che si può riscrivere come $\int_{\{\|y\|>\alpha\}} \|y\|^2 \, dF_{\mathbf{y}(t)}(y)$ dove $F_{\mathbf{y}(t)}(y)$ è la d.d.p. di una qualunque delle variabili del processo) è indipendente da T e tende manifestamente a zero quando $\alpha \to \infty$.

Il TLC per processi stazionari dissolventi

Presenteremo qui sotto una versione generale del TLC, dovuta essenzialmente a Bernstein e Rosenblatt [1, 23, 22], che non richiede l'indipendenza ed è valida, per somme di variabili di un processo dissolvente.

In particolare, se il processo è p.n.d., si assicura almeno asintoticamente l'indipendenza delle variabili "sufficientemente lontane nel tempo e si può usare nella dimostrazione un argomento simile a quello del caso i.i.d.. Per usare questo tipo di argomenti si impone che il *coefficiente di dissolvenza* (*mixing coefficient* in inglese)

$$\alpha(\tau) := \sup_{A \in \mathcal{Y}_t^-, \, B \in \mathcal{Y}_{t+\tau}^+} | \, P(A \cap B) - P(A)P(B) \, | \,, \qquad \tau > 0, \tag{7.14}$$

che in virtù della stazionarietà del processo, dipende solo da τ, tenda a zero abbastanza velocemente per $\tau \to +\infty$. Serve in effetti che $\alpha(\tau)$ tenda a zero più velocemente di $\frac{1}{\tau}$,

$$\alpha(\tau) = O(\frac{1}{\tau^{1+\epsilon}}) \qquad \epsilon > 0 \tag{7.15}$$

Chiameremo un processo stazionario che soddisfa la condizione (7.15), *fortemente dissolvente* (*strongly mixing* in inglese). Notiamo che per un processo p.n.d., in virtù delle (6.28), (6.29), $\alpha(\tau)$ tende in ogni caso a zero.

Teorema 7.5 (Bernstein, Rosenblatt, Rozanov). *La d.d.p $F_{\mathbf{s}_T}(x)$ di un processo fortemente dissolvente e a varianza finita (soddisfa la condizione di Lindeberg e quindi) converge a una Gaussiana.*

Dimostrazione. La dimostrazione di questo teorema si può trovare nel testo di Rozanov, pp.194-195. □

Nota Bene: La matrice varianza della distribuzione limite $\Phi(x)$, di $\sqrt{T}\,\mathbf{y}_T$ in generale non coincida con quella, $\Sigma = E\,\mathbf{y}(t)\mathbf{y}(t)'$, delle variabili del processo $\mathbf{y}$. Questo vale nel caso particolare in cui il processo è i.i.d. o una d-martingala, ma non è vero in generale. L'espressione da usare per la varianza asintotica è quella data nel teorema 7.3.

Alcune classi di processi per cui vale il TLC

La proprietà di un processo strettamente stazionario $\mathbf{y}$ di essere p.n.d. (in senso stretto), dissolvente etc. sono ovviamente assai difficili da verificare in pratica. Esse giocano un ruolo importante nella teoria perchè vengono automaticamente ereditate da ogni processo $\{\mathbf{z}(t)\}$ ottenuto per traslazione temporale di una qualunque funzione a supporto ($\mathbb{I}$) finito

del processo, $\mathbf{z} = f(\mathbf{y})$[17]. Detta $i = |\mathbb{I}|$ l'estensione dell'insieme $\mathbb{I}$ (la differenza tra il suo massimo e minimo elemento), si può così dedurre che per gli spazi passato e futuro di $\mathbf{z}$ valgono delle inclusioni del tipo,

$$L_t^-(\mathbf{z}) \subset L_{t+i}^-(\mathbf{y}) \qquad L_t^+(\mathbf{z}) \subset L_{t-i}^+(\mathbf{y}) \tag{7.16}$$

qualunque sia t. Equivalentemente, $\mathcal{Z}_t^- \subset \mathcal{Y}_{t+i}^-$, $\mathcal{Z}_t^+ \subset \mathcal{Y}_{t-i}^+$ per le relative σ-algebre e ne segue immediatmante che se $\mathbf{y}$ è p.n.d. (in senso stretto), fortemente dissolvente etc. la stesssa proprietà viene erditata automaticamente dal processo $\{\mathbf{z}(t)\}$. Si ha pertanto il seguente risultato, che può essere visto come un corollario del teorema 7.5.

Teorema 7.6. *Ogni processo generato per traslazione temporale secondo la* (6.6)*, di una funzione a supporto finito $\mathbf{z} = f(\mathbf{y})$ di un processo fortemente dissolvente $\mathbf{y}$ è ancora fortemente dissolvente. Se $\mathbf{y}$ è fortemente dissolvente e $\{\mathbf{z}(t)\}$ ha momenti del secondo ordine finito, i.e. $\mathbf{z} \in L^2(\mathbf{y})$, vale il teorema del limite centrale, nel senso che, detta $F_{\sqrt{T}\,\bar{\mathbf{z}}_T}$ la d.d.p. della variabile $\sqrt{T}\,\bar{\mathbf{z}}_T$, si ha*

$$\lim_{T\to\infty} F_{\sqrt{T}\,\bar{\mathbf{z}}_T}(x) = \Phi(x) \tag{7.17}$$

dove $\Phi(x)$ è una distribuzione Gaussiana.

Esempio 7.2. Assumiamo che il processo scalare $\mathbf{y}$ ammetta momenti di ordine sufficientemente elevato; allora le variabili

$$\frac{1}{T}\sum_1^T \mathbf{y}(t)^2, \qquad \frac{1}{T}\sum_1^T \mathbf{y}(t)\mathbf{y}(t-1), \qquad T = 1, 2, \dots$$

convergono con probabilità uno a $\mu_2 := E\,\mathbf{y}(t)^2$ e a $\sigma(1) := E\,\mathbf{y}(t)\mathbf{y}(t-1)$. Le stesse quantità, normalizzate moltiplicandole per $\sqrt{T}$, sono congiuntamente asintoticamente Gaussiane.

Per il calcolo della media e della varianza asintotiche della distribuzione (Gaussiana) di una funzione non lineare di un processo si può utilizzare il seguente teorema di Cramèr.

Teorema 7.7 (Cramèr). *Sia $\{\mathbf{x}_N\,;\, N = 1, 2, \dots\}$ una successione di vettori aleatori n-dimensionali per cui $\sqrt{N}\,(\mathbf{x}_N - \mu) \xrightarrow{L} N(0, \Sigma)$ 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:*

$$\sqrt{N}\,(g(\mathbf{x}_N) - g(\mu)) \xrightarrow{L} N(0, G(\mu)\Sigma G(\mu)')\,. \tag{7.18}$$

Come applicazione del teorema di Cramèr, consideriamo la successione dei quadrati delle medie campionarie $\{(\bar{\mathbf{y}}_T)^2\}$. Dato che per $T \to \infty$, $\sqrt{T}\,\bar{\mathbf{y}}_T \sim N(0, \sigma^2)$ e $\frac{\partial}{\partial x} g(x) = 2x$, si ha

$$\sqrt{T}\,\left[(\bar{\mathbf{y}}_T)^2 - \mu^2\right] \xrightarrow{L} N(0, 4\mu^2\sigma^2)\,.$$

[17] Quindi $f(\mathbf{y})$ dipende solo da un numero *finito* di variabili del processo $\mathbf{y}$.

Osservazione 7.4. È da tener presente che l'ordine di infinitesimo (la velocità asintotica) con cui la varianza tende a zero può essere in certi casi diversa da $\frac{1}{T}$. Questo accade nell'esempio appena visto se $\mu = 0$. In questo caso si trova infatti che la distribuzione limite del quadrato della media campionaria è *degenere* ($N(0,0)$), in altri termini, il primo membro converge alla costante zero, il che è un chiaro sintomo del fatto che per $T \to \infty$ la varianza della successione converge a zero più rapidamente di $\frac{1}{T}$. Questo fatto si può verificare direttamente notando che

$$\sqrt{T}\left[(\bar{\mathbf{y}}_T)^2 - \mu^2\right] = \left[\sqrt{T}\,(\bar{\mathbf{y}}_T - \mu)\right](\bar{\mathbf{y}}_T + \mu)$$

in cui il primo fattore tra parentesi quadre converge in legge ad una distribuzione Gaussiana, ma il secondo, per il teorema ergodico, tende (quasi certamente) alla costante 2μ che è zero per $\mu = 0$.

Per analizzare casi di questo genere torna utile l'affermazione (1) del teorema di Slutsky 7.1. Esaminiamo l'esempio precedente (con $\mu = 0$) alla luce di questo risultato. Dato che $\sqrt{T}\,\bar{\mathbf{y}}_T \xrightarrow{L} N(0,\sigma^2)$, prendendo $f(x) = x^2$, il teorema di Slutsky asserisce che la d.d.p di $T\,(\bar{\mathbf{y}}_T)^2$ converge alla d.d.p. del quadrato di una variabile Gaussiana $N(0,\sigma^2)$. In altri termini

$$\frac{T\,(\bar{\mathbf{y}}_T)^2}{\sigma^2} \xrightarrow{L} \chi^2(1)$$

e la velocità di convergenza alla distribuzione limite della distribuzione di $(\bar{\mathbf{y}}_T)^2$ è dell'ordine di $\frac{1}{T}$. La varianza tenderà quindi al suo valore asintotico come $\frac{1}{T^2}$.

Esempio 7.3. *Mostrare che la distribuzione asintotica della varianza campionaria $s_T^2(\mathbf{y})$, di un processo scalare i.i.d. con momento del quart'ordine μ_4 finito, è*

$$\sqrt{T}\,(s_T^2(\mathbf{y}) - \sigma^2) \xrightarrow{L} N(0, \mu_4 - \sigma^4)\,.$$

Soluzione:

La varianza campionaria si può esprimere come

$$\begin{aligned} s_T^2(\mathbf{y}) &= \frac{1}{T}\sum_{t=1}^{T}(\mathbf{y}(t) - \bar{\mathbf{y}}_T)^2 = \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}(t)^2 - (\bar{\mathbf{y}}_T)^2 \\ &= \frac{1}{T}\sum_{t=1}^{T}\left[\mathbf{y}(t) - \mu - (\bar{\mathbf{y}}_T - \mu)\right]^2 = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{y}(t) - \mu)^2 - (\bar{\mathbf{y}}_T - \mu)^2 \\ &:= m_2(\mathbf{y}) - m_1(\mathbf{y})^2 \end{aligned}$$

Dato che $\mathbf{y}$ è un processo i.i.d. si ha

$$\sqrt{T}m_2(\mathbf{y}) := \sqrt{T}\frac{1}{T}\sum_{t=1}^{T}(\mathbf{y}(t) - \mu)^2 \xrightarrow{L} \mathcal{N}(\sigma^2,\, \mu_4)$$

dove $\mu_4 := \mathbb{E}\,(\mathbf{y}(t) - \mu)^4$ è il momento centrale del quarto ordine. Analogamente, si ha $\sqrt{T}m_1(\mathbf{y}) \xrightarrow{L} \mathcal{N}(0,\, \sigma^2)$.

Usiamo il teorema di Cramèr (Teorema 7.7). Poniamo

$$s_T^2(\mathbf{y}) = -m_1(\mathbf{y})^2 + m_2(\mathbf{y}) := g(m_1(\mathbf{y}), m_2(\mathbf{y}))$$

e notiamo che

$$\sqrt{T}\left\{\begin{bmatrix} m_1(\mathbf{y}) \\ m_2(\mathbf{y}) \end{bmatrix} - \begin{bmatrix} 0 \\ \sigma^2 \end{bmatrix}\right\} \xrightarrow{L} \mathcal{N}(0, \Sigma)$$

dove

$$\Sigma = \begin{bmatrix} \text{var } \mathbf{y}(t) & \text{cov}\,(\mathbf{y}(t)^2, \mathbf{y}(t)) \\ \text{cov}\,(\mathbf{y}(t)^2, \mathbf{y}(t)) & \text{var } \mathbf{y}(t)^2 \end{bmatrix} = \begin{bmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{bmatrix}$$

Il calcolo del momento terzo μ_3 non serve perchè il gradiente di g rispetto alle due variabili $m_1,\, m_2$ è $g'(m_1, m_2) = [-2m_1,\, 1]$ per cui $g'(0, \sigma^2) = [0,\, 1]$ e

$$g'(0,\sigma^2)\Sigma\left[g'(0,\sigma^2)\right]^\top = \text{var } \mathbf{y}(t)^2 = \mathbb{E}\,\mathbf{y}(t)^4 - (\mathbb{E}\,\mathbf{y}(t)^2)^2 = \mu_4 - \sigma^4$$

per cui in definitiva si ha

$$\sqrt{T}\left[s_T^2(\mathbf{y}) - g(0,\sigma^2)\right] = \sqrt{T}\left[s_T^2(\mathbf{y}) - \sigma^2\right] \xrightarrow{L} \mathcal{N}(0, \mu_4 - \sigma^4)\,.$$

Se la distribuzione di $\mathbf{y}(t)$ è Gaussiana, $\mu_4 = 3\sigma^4$ e la distribuzione limite è $\mathcal{N}(0, 2\sigma^4)$.

Alle volte però si può ottenere il risultato desiderato più semplicemente usando Slutsky.

Esempio 7.4. *Supponiamo che* $\mathbf{y}$ *sia un processo scalare i.i.d. con momento del quart'ordine* μ_4 *finito. Vogliamo ricavare la distribuzione asintotica della statistica di Student dell'esempio 7.1 con il teorema di Cramèr.*

Soluzione: Iniziamo col ricavare la distribuzione asintotica della statistica,

$$\psi(\mathbf{y}) := \frac{1}{\sqrt{N}}\varphi(\mathbf{y}) := \frac{[\bar{\mathbf{y}}_N - \mu]}{\sqrt{s_N^2(\mathbf{y})}}\,.$$

Dall'esercizio precedente ricaviamo subito che

$$\sqrt{N}\left\{\begin{bmatrix} \bar{\mathbf{y}}_N - \mu \\ s_T^2(\mathbf{y}) \end{bmatrix} - \begin{bmatrix} 0 \\ \sigma^2 \end{bmatrix}\right\} \xrightarrow{L} \mathcal{N}(0, \Sigma)$$

dove

$$\Sigma = \begin{bmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{bmatrix}$$

Il momento incrociato μ_3 non interessa. Posto $\psi(\mathbf{y}) = g(\bar{\mathbf{y}}_N - \mu,\, s_T^2(\mathbf{y}))$ con $g(x_1,\, x_2) := \dfrac{x_1}{\sqrt{x_2}}$ si ha,

$$g(\mu_1,\, \mu_2) = \frac{0}{\sigma} = 0, \qquad \frac{\partial g}{\partial x}(x_1,\, x_2) = [\frac{1}{\sqrt{x_2}},\, -\frac{1}{2}x_1 x_2^2\,]$$

dove lo Jacobiano è manifestamente continuo in $(0,\, \sigma^2)$ e $\dfrac{\partial g}{\partial x}(0,\, \sigma^2) = [\,1/\sigma,\, 0\,]$. Per cui, applicando la formula (7.18) del teorema di Cramèr, si trova,

$$\varphi(\mathbf{y}) = \sqrt{N} g(\bar{\mathbf{y}}_N,\, s_T^2(\mathbf{y})) \xrightarrow{L} \mathcal{N}(0,1)$$

che è lo stesso risultato a cui siamo arrivati nell'esempio 7.1. Di fatto nel calcolo che abbiamo fatto la distribuzione asintotica di $s_T^2(\mathbf{y})$ non interviene.

Maggiori informazioni (incluse le dimostrazioni dei teoremi citati in questo paragrafo) si possono trovare nel testo [9].

Sistemi lineari e TLC

Un caso molto interessante per le applicazioni riguarda processi ottenuti come uscita di un filtro lineare sollecitato da un "rumore bianco". Notiamo che l'uscita di un filtro di questo genere dipende dalla storia passata *infinita* del processo di ingresso per cui il teorema 7.6 non è applicabile. Il caso più semplice da analizzare è quello di sistemi lineari tempo-invarianti di dimensione finita, descrivibili mediante le classiche equazioni di stato

$$\mathbf{x}(t+1) = A\mathbf{x}(t) + B\mathbf{e}(t) \tag{7.19}$$

$$\mathbf{y}(t) = C\mathbf{x}(t) \tag{7.20}$$

Teorema 7.8. *Se A è strettamente stabile (autovalori all'interno del cerchio unità) ed* $\mathbf{e}$ *è un processo a varianza finita a cui si applica il teorema del limite centrale, allora il teorema del limite centrale vale anche per il processo di uscita del sistema linere* (7.19)*,* (7.20) *e risulta*

$$\sqrt{T}\,\bar{\mathbf{y}}_T \xrightarrow{L} \mathcal{N}(0, C(I-A)^{-1}BQB^\top(I-A)^{-\top}C^\top) \tag{7.21}$$

dove $Q = Q^\top \geq 0$ è la varianza asintotica di $\sqrt{T}\,\bar{\mathbf{e}}_T$.

Dimostrazione. Se la proprietà dell'enunciato vale per il processo di stato, ovviamente vale anche per quello di uscita. Consideriamo allora la (7.19) e prendiamo la media temporale del primo membro,

$$\sqrt{T}\,\bar{\mathbf{x}}_T^1 := \sqrt{T}\sum_{t=1}^{T}\mathbf{x}(t+1) = A\sqrt{T}\,\bar{\mathbf{x}}_T + B\sqrt{T}\,\bar{\mathbf{e}}_T\,.$$

Ora $\sqrt{T}\,\bar{\mathbf{x}}_T^1$ e $\sqrt{T}\,\bar{\mathbf{x}}_T$ sono processi a varianza finita ed è immediato verificare che sono tra loro asintoticamente equivalenti; i.e. $\sqrt{T}\,\bar{\mathbf{x}}_T^1 - \sqrt{T}\,\bar{\mathbf{x}}_T \to 0$ in probabilità. Quindi, per calcolare la distribuzione asintotica possiamo, in forza del teoreme di Slutsky, sostituire al primo membro $\sqrt{T}\,\bar{\mathbf{x}}_T$ al posto di $\sqrt{T}\,\bar{\mathbf{x}}_T^1$, ottenendo così

$$\sqrt{T}\,\bar{\mathbf{x}}_T = (I-A)^{-1}B\sqrt{T}\,\bar{\mathbf{e}}_T\,.$$

Ne segue che $\sqrt{T}\,\bar{\mathbf{x}}_T$ è asintoticamente normale di varianza asintotica data dalla matrice $(I-A)^{-1}BQB^\top(I-A)^{-\top}$. □

Notiamo in particolare che se $\mathbf{e}$ è i.i.d. allora Q è anche la varianza di $\mathbf{e}(t)$ e la varianza asintotica di $\sqrt{T}\,\bar{\mathbf{y}}_T$ è proprio uguale alla densità spettrale di $\mathbf{y}$ calcolata in $e^{j\omega} = 1$, come stabilito nel teorema 7.3.

Un caso un pò più generale riguarda processi stazionari generati come uscita di un filtro ℓ^2-stabile, non necessariamente razionale, sollecitati da un processo $\mathbf{e}$ che supporremo i.i.d. a media zero e di varianza σ^2 finita. Per semplicità considereremo il caso scalare,

$$\mathbf{y}(t) = \sum_{-\infty}^{t} h(t-s)\,\mathbf{e}(s) \qquad \sum_{0}^{+\infty} h(s)^2 \,<\, \infty\,. \tag{7.22}$$

Notiamo che i sottospazi della storia (strettamente) passata e futura all'istante t, $L_{t-1}^{2-}(\mathbf{e})$ e $L_t^{2+}(\mathbf{e})$ sono *ortogonali* per cui, per ogni intero $\tau > 0$ la proiezione ortogonale di $\mathbf{y}(t)$ su $L_{t-\tau}^{2+}(\mathbf{e})$ si scrive

$$\hat{\mathbf{y}}(t \mid t-\tau) := \mathbb{E}\,[\mathbf{y}(t) \mid L_{t-\tau}^{2+}(\mathbf{e})\,] = \sum_{t-\tau}^{t} h(t-s)\,\mathbf{e}(s)$$

e quindi

$$\mathbf{y}(t) - \hat{\mathbf{y}}(t \mid t-\tau) = \sum_{-\infty}^{t-\tau-1} h(t-s)\,\mathbf{e}(s) \;= \sum_{\tau+1}^{+\infty} h(s)\mathbf{e}(t-s)\,.$$

Da questa relazione si ricava facilmente che

$$\frac{1}{\sqrt{T}}[\sum_{t=1}^{T}\mathbf{y}(t) - \sum_{t=1}^{T}\hat{\mathbf{y}}(t \mid t-\tau)] = \frac{1}{\sqrt{T}}\sum_{\tau+1}^{+\infty} h(s)(\sum_{t=1}^{T}\mathbf{e}(t-s))$$

che tende a zero in media quadratica, per $\tau \to +\infty$ dato che

$$\mathbb{E}\,\{\frac{1}{\sqrt{T}}[\sum_{t=1}^{T}\mathbf{y}(t) - \sum_{t=1}^{T}\hat{\mathbf{y}}(t \mid t-\tau)]\}^2 = \frac{T\sigma^2}{T}\sum_{\tau+1}^{+\infty} h(s)^2 \to 0 \tag{7.23}$$

dato che $h \in \ell^2$. Quindi, per $\tau \to +\infty$ il processo delle medie normalizzate di $\hat{\mathbf{y}}(t \mid t-\tau)$ converge in media quadratica (e quindi anche in probabilità) a $\sqrt{T}\bar{\mathbf{y}}_T$. Dato che $\mathbf{e}$ è fortemente dissolvente e $\hat{\mathbf{y}}(t \mid t-\tau)$ dipende per ogni τ solo da un tratto finito della storia passata di $\mathbf{e}$, scende dal teorema 7.6 che $1/\sqrt{T}\sum_{t=1}^{T}\hat{\mathbf{y}}(t \mid t-\tau)$ è, per ogni τ, asintoticamente Gaussiano.

Purtroppo il successivo passaggio al limite richiede che al tendere di τ all'infinito, le "code" $\sum_{\tau+1}^{+\infty} h(s)^2$ convergano a zero con sufficiente rapidità e per questo serve una condizione più stringente dell'energia finita ($h \in \ell^2$). E. Hannan [**?**] ha dimostrato che è sufficiente assumere

$$\sum_{t=0}^{+\infty} |h(t)| \,<\, \infty \tag{7.24}$$

che è la condizione naturale per la stabilità ingresso-uscita del sistema (7.22).

Teorema 7.9. *Nell' ipotesi* (7.24)*, la successione delle medie normalizzate* $\sqrt{T}\bar{\mathbf{y}}_T$ *del processo definito dall' equazione* (7.22) *converge in legge ad una distribuzione Gaussiana.*

7.4 Efficienza asintotica

Sebbene sia abbastanza chiaro dal punto di vista intuitivo, il concetto di *stimatore asintoticamente efficiente* ovvero di stimatore *asintoticamente a minima varianza* è abbastanza delicato da definire in modo preciso. Una delle difficoltà risiede nel fatto che la varianza di quasi tutti gli stimatori interessanti nelle appplicazioni (che debbono essere *consistenti* ovvero, asintoticamente corretti) 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 quindi *andamenti asintotici* della varianza.

Definizione 7.2. *Sia* $\{\phi_T(\mathbf{y})\,;\, T = 1, 2, \ldots\}$ *una successione di stimatori*[18] *e* $d(T)$ *una funzione di* T *crescente e strettamente posititiva. Diremo che* $\phi_T(\mathbf{y})$ *ha* varianza asintotica Σ *se*

$$\sqrt{d(T)}\,\phi_T(\mathbf{y}) \xrightarrow{L} D(\mu, \Sigma) \tag{7.25}$$

dove $D(\mu, \Sigma)$ *è una d.d.p. di media* μ *e varianza* Σ*, finita e 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 la convergenza in distribuzione implica la convergenza dei momenti per cui la varianza asintotica può anche essere definita come il limite

$$\lim_{T\to\infty} \mathrm{Var}\left[\sqrt{d(T)}\,\phi(\mathbf{y}_T)\right] = \Sigma\,. \tag{7.26}$$

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 asintoticamente corretto (consistente) si può dire che è (asintoticamente) efficiente se la sua varianza asintotica è uguale all'inversa della matrice (asintotica) di Fisher[19]

Come abbiamo avuto modo di vedere, 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

$$\Sigma = I(\theta)^{-1}\,. \tag{7.27}$$

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 $1/T$.

[18] Questa terminologia è un pò pesante e verrà usualmente abbreviata riferendosi semplicemente al generico stimatore $\phi_T(\mathbf{y})$ della sequanza.

[19] Con campioni di numerosità infinita, ci possono essere patologiche eccezioni alla disuguaglianza di Cramèr-Rao. Può infatti accadere che per qualche valore di θ la varianza asintotica di uno stimatore consistente sia strettamente più piccola dell'inversa della matrice asintotica di Fisher. Si dimostra però che questo può accadere solo per un insieme di valori del parametro di misura di Lebesgue nulla. Per maggiori informazioni si vedano [9, **?**].

Capitolo 8

METODI A MINIMIZZAZIONE DELL'ERRORE DI PREDIZIONE

8.1 Introduzione

In questo capitolo parleremo di una tecnica per assegnare, in base ai dati ingresso-uscita misurati su un impianto fisico, un modello appartenente ad una classe parametrica di modelli assegnata a priori. I modelli saranno modelli lineari del tipo discusso nel capitolo 5. In breve, ci occuperemo di *stima parametrica* su modelli dinamici lineari.

Tratteremo sostanzialmente di 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 [18, 19].

Il principio su cui si basano i metodi PEM è molto semplice. 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

$$y^T := \{y(t)\,;\ t = 1, 2, \ldots, T\}\,, \qquad u^T := \{u(t)\,;\ t = 1, 2, \ldots, T\} \tag{8.1}$$

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 $\hat{M}(\theta)$, 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 all'occorrenza 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*:
$$\varepsilon_\theta(t) := y(t) - \hat{y}_\theta(t)\,; \qquad t = 1, 2, \ldots, T$$

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 (5.9), usando formalmente l'espressione per il predittore di Wiener (5.7), si ottiene

$$\begin{aligned}\boldsymbol{\varepsilon}_\theta(t) &= \mathbf{y}(t) - \hat{\mathbf{y}}(t \mid t-1) = \mathbf{y}(t) - G_\theta(z)^{-1}\left[F_\theta(z)\mathbf{u}(t) + (G_\theta(z)-1)\mathbf{y}(t)\right] \\ &= G_\theta(z)^{-1}\left[\mathbf{y}(t) - F_\theta(z)\mathbf{u}(t)\right] \end{aligned} \tag{8.2}$$

che si può interpretare come una rappresentazione del processo $\mathbf{y}$ mediante un modello arbitrario della classe (5.9); i.e.

$$\mathbf{y}(t) = F_\theta(z)\mathbf{u}(t) + G_\theta(z)\boldsymbol{\varepsilon}_\theta(t)\,, \tag{8.3}$$

in cui però l'innovazione è sostituita dall'errore di predizione (che in generale non è bianco).

3. Si minimizza rispetto a θ una cifra di merito che descriva quanto bene (in media) il modello predice il dato successivo. Ad esempio si minimizza *l'errore quadratico medio di predizione*,

$$V_T(\theta) := \frac{1}{T}\sum_{t=1}^{T} \varepsilon_\theta(t)^2 \tag{8.4}$$

o, più in generale, una media degli errori quadratici di predizione pesati da una qualche funzione non negativa β,

$$V_T(\theta) := \frac{1}{T}\sum_{t=1}^{T} \beta(t,T)\varepsilon_\theta(t)^2 \qquad \beta(t,T) > 0 \tag{8.5}$$

che dia 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. 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,

$$\hat{\theta}_T := \operatorname{Arg}\min_\theta V_T(\theta) \tag{8.6}$$

che è appunto la stima PEM del parametro del modello. Naturalmente lo stimatore $\hat{\boldsymbol{\theta}}_T$ 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

$$\hat{\lambda}^2_T := V_T(\hat{\theta}_T) \tag{8.7}$$

dove V_T è dato dall'espressione (8.4).

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 questo capitolo sarà sostanzialmente dedicato all'analisi delle proprietà statistiche dello stimatore PEM.

8.2 Analisi asintotica dello stimatore PEM

Come è facilmente intuibile, lo stimatore PEM risulta essere una funzione complicata dei dati di misura e in effetti, risulta sempre essere una funzione non lineare dei dati. Ne segue che l'unica analisi possibile è quella asintotica, come si suol dire, *per grandi campioni*, ovvero per $T \to \infty$. Ci dovremo quindi accontentare di studiare le proprietà statistiche dello stimatore PEM, in particolare del processo errore di predizione e del relativo predittore $\hat{\boldsymbol{y}}_\theta(t \mid t-1)$, *supponendo che i dati siano segnali stazionari* nel senso spiegato nel capitolo precedente, e quindi modellabili come preocessi stocastici stazionari e di avere a disposizione una successione infinita di dati. Notiamo però che lo schema appena esposto non richiede affatto che la numerosità campionaria tenda all'infinito e può essere applicato anche per "piccoli campioni. In particolare, il predittore basato sul modello $M(\theta)$, che in pratica, in generale dovrà operare su dati finiti, dovrà essere inteso come predittore *non stazionario*, impiegando ad esempio un filtro di Kalman opportunamente inizializzato.

Notiamo anche che l'analisi statistica che faremo non riguarda questa situazione ed è plausibile aspettarsi che altri criteri di stima (ad esempio il criterio della massima verosimiglianza, quando applicabile) possano in questo caso dare risultati migliori.

Il teorema seguente è il primo risultato fondamentale della teoria asintotica della stima PEM.

Teorema 8.1. *Si assuma che*

1. *I dati* (8.1) *sono generati da un processo ergodico del secondo ordine.*
2. *Il modello parametrico $M(\theta)$ è del tipo* (5.9) *e dipende dal parametro in modo razionale;*
3. *La cifra di merito è una funzione quadratica dei dati, ad esempio del tipo* (8.4)*;*
4. *Esiste (almeno) un minimo* (8.6)*; i.e. esiste un insieme compatto $\Theta_0 \subseteq \Theta$ sufficientemente grande che, per $T \to \infty$, $\hat{\theta}_T \in \Theta_0$ con probabilità uno.*

Si ha allora

$$\lim_{T\to\infty} V_T(\theta) = \mathbb{E}_0 \boldsymbol{\varepsilon}_\theta(t)^2 \tag{8.8}$$

dove $\mathbb{E}_0$ denota aspettazione rispetto alla distribuzione del processo vero che ha generato i dati.

I minimizzatori, $\hat{\boldsymbol{\theta}}_T$, di $V_T(\theta)$, convergono tutti con probabilità uno all'insieme dei punti di minimo di $\bar{V}(\theta) := \mathbb{E}_0 \boldsymbol{\varepsilon}_\theta(t)^2$. In effetti, detto $\Delta \subset \Theta_0$ l'insieme dei punti di minimo di $\bar{V}(\theta)$, si ha

$$\lim_{T\to\infty} \hat{\boldsymbol{\theta}}_T \in \Delta \tag{8.9}$$

con probabiltà uno.

Dimostrazione. In effetti la (8.8) segue dalla definizione stessa di ergodicità del secondo ordine. La convergenza dei minimi segue dal fatto che le variabili aleatorie della famiglia $\{\varepsilon_\theta(t)^2\,;\, t \geq 1\}$ sono tutte non negative e quindi uniformemente limitate inferiormente (dalla variabile aleatoria zero) e si può quindi applicare un teorema di Le Cam (vedere la versione "duale" per l'operazione di massimizzazione in [9, teorema 16(a), p. 108]). In queste condizioni si può commutare l'operazione di minimizzazione di $V_T(\theta)$ su un arbitrario insieme compatto con quella di passaggio al limite. Il fatto che i punti di minimo sono (almeno per T grande) contenuti con probabilità uno in un insieme compatto, equivale alla nostra ipotesi che un minimo esista effettivamente per quasi tutte le possibili sequenze di dati di misura. □

Osservazione 8.1. In vece dell'ergodiocità del secondo ordine si può assumere semplicemente stazionarietà del secondo ordine. In questo caso potrebbero anche essere presenti delle componenti armoniche nei segnali in gioco e il processo "vero" corrispondente non sarebbe ergodico del secondo ordine. Questa circostanza in effetti si verifica spesso quando l'ingresso è imposto artificialmente nell'esperimento di identificazione ed è composto da una somma di sinusoidi. In questo caso l'enunciato del teorema continua comunque a valere pur di sopprimere la qualificazione "con probabilità uno".

Il processo vero è stazionario e con momenti del second'ordine finiti per cui può essere decomposto come somma di un predittore lineare a minima varianza d'errore $\hat{\mathbf{y}}_0(t \mid t-1)$, ovvero 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) di $\mathbf{y}(t)$

$$\mathbf{y}(t) = \hat{\mathbf{y}}_0(t \mid t-1) + \mathbf{e}_0(t)\,. \tag{8.10}$$

Usando questa decomposizione e notando che il termine tra parentesi quadre nella decomposizione,

$$\varepsilon_\theta(t) = \mathbf{e}_0(t) + [\hat{\mathbf{y}}_0(t \mid t-1) - \hat{\mathbf{y}}_\theta(t \mid t-1)]$$

è funzione dei dati $(\mathbf{y}^{t-1}\,,\, \mathbf{u}^{t-1})$ e quindi ortogonale ad $\mathbf{e}_0(t)$, si ricava la

$$\begin{aligned}\bar{V}(\theta) &= \operatorname{var}\{\mathbf{e}_0(t)\} + \|\hat{\mathbf{y}}_0(t \mid t-1) - \hat{\mathbf{y}}_\theta(t \mid t-1)\|^2 \\ &= \lambda_0^2 + \|\hat{\mathbf{y}}_0(t \mid t-1) - \hat{\mathbf{y}}_\theta(t \mid t-1)\|^2\end{aligned} \tag{8.11}$$

dove la norma è la norma nello spazio di Hilbert generato linearmente da $(\mathbf{y},\, \mathbf{u})$ (i.e. la varianza). Da questa espressione si vede che *lo stimatore PEM minimizza asintoticamente la distanza tra il predittore"vero" e quello costruito sul modello* $M(\theta)$. Da notare che, 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).

Osserviamo anch eche 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 di modo tale che $\mathbf{e}_0(t)$ è scorrelata da $\hat{\mathbf{y}}_\theta(t \mid t-1)$. Su questo fatto è in effetti basato anche il fondamentale risultato seguente[20].

[20]Qiundi il predittore non può essere *arbitrario* come Ljung e qualche suo ottimista seguace insiste nel propagandare.

Teorema 8.2. *Supponiamo che valgano le ipotesi (1), (2), (3) del teorema precedente e che l'errore di predizione $\varepsilon_\theta(t)$ sia calcolato mediante il predittore lineare a minima varianza $\hat{\mathbf{y}}_\theta(t \mid t-1)$. Supponiamo inoltre che la classe parametrica dei modelli $\mathcal{M} \equiv \{M(\theta);\, \theta \in \Theta\}$ sia identificabile a priori e che il processo $\mathbf{u}$ sia sufficientemente eccitante da garantire l' identificabilità.*

Allora, se il processo (vero) che genera i dati è descritto da un modello che appartiene alla stessa classe parametrica $\mathcal{M}$ dei modelli scelti per l'identificazione, ovvero esiste $\theta_0 \in \Theta$ tale che

$$\mathbf{y} \sim M(\theta_0) \in \mathcal{M}$$

si ha:

$$\lim_{T\to\infty} \hat{\boldsymbol{\theta}}_T = \theta_0 \tag{8.12}$$

con probabilità uno. In altri termini, lo stimatore PEM è consistente.

Dimostrazione. Se $M(\theta_0) \in \mathcal{M}$, si vede subito dalla (8.11) 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 $\{\theta_0\}$. Stabiliamo questo fatto separatamente.

Lemma 8.1. *Se c'è identificabilità*

$$\|\hat{\mathbf{y}}_{\theta_0}(t \mid t-1) - \hat{\mathbf{y}}_\theta(t \mid t-1)\| = 0 \;\Leftrightarrow\; \theta = \theta_0\,; \tag{8.13}$$

i.e. c'è un solo punto di minimo assoluto di $\bar{V}(\theta)$ e $\Delta = \{\theta_0\}$.

Dimostrazione. Dato che il predittore di Wiener è una funzione lineare dei dati si può scrivere simbolicamente

$$\hat{\mathbf{y}}_\theta(t \mid t-1) = L_\theta(z)\mathbf{u}(t-1) + M_\theta(z)\mathbf{y}(t-1)$$

dove $L_\theta(z)$ e $M_\theta(z)$ sono funzioni razionali strettamente stabili. Ne segue che il quadrato della norma in (8.13) si può esprimere nel dominio della frequenza con l'integrale,

$$\|\hat{\mathbf{y}}_{\theta_0}(t \mid t-1) - \hat{\mathbf{y}}_\theta(t \mid t-1)\| = \int_{-\pi}^{\pi} \left[L_{\theta_0}(e^{j\omega}) - L_\theta(e^{j\omega})\; M_{\theta_0}(e^{j\omega}) - M_\theta(e^{j\omega})\right] \cdot \\ \cdot \begin{bmatrix} S_{\mathbf{u}}(e^{j\omega}) & S_{\mathbf{uy}}(e^{j\omega}) \\ S_{\mathbf{yu}}(e^{j\omega}) & S_{\mathbf{y}}(e^{j\omega}) \end{bmatrix} \begin{bmatrix} L_{\theta_0}(e^{j\omega}) - L_\theta(e^{j\omega}) \\ M_{\theta_0}(e^{j\omega}) - M_\theta(e^{j\omega}) \end{bmatrix} \frac{d\omega}{2\pi}$$

where for simplicity we have assumed that $\mathbf{u}$ has a spectral density. If this is not the case one should instead write an analogous expression using spectral distribution functions. Now for all absolute minima of $\bar{V}(\theta)$ the integral above should be zero. However by definition of identifiability, $\mathbf{u}$ is such that the spectral density matrix in the integral is positive definite at least in a set of freqencies ω containing enough points so as to guarantee that $\|\hat{\mathbf{y}}_{\theta_0}(t \mid t-1) - \hat{\mathbf{y}}_\theta(t \mid t-1)\| = 0$ implies

$$L_{\theta_0}(z) \equiv L_\theta(z)\,, \qquad M_{\theta_0}(z) \equiv M_\theta(z)$$

where $\equiv$ means equality for all z.

Confrontando ora le espressioni di $L(z)$ e di $M(z)$ fornite in (5.7) nel teorema 5.1, si controlla facimente che 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 a priori implica infine che questo possa accadere allora e solo allora che $\theta = \theta_0$. □

Il teorema è così dimostrato. □

Problema 8.1. Cosa accade 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 stazionario sinusoidale. Esempio

$$M(\theta): \quad (1 - az^{-1})\mathbf{y}(t) = b\mathbf{u}(t-1) + \mathbf{e}(t)$$

con $\mathbf{u}$ ed $\mathbf{e}$ scorrelati. Calcolare $\bar{V}(\theta)$ (che dipenderà dall'ampiezza del segnale di ingresso) e minimizzarlo rispetto a θ. Verificare se il limite di $\hat{\theta}_T$ dipende dall'ampiezza dell'ingresso. Si ha consistenza? Il processo (deterministico) $\mathbf{u}$ potrebbe anche essere una variabile aleatoria costante di media nulla. Si deve comunque garantire l' identificabilità.

Finora non ci siamo interessati molto alla stima della varianza dell'innovazione. Il seguente risultato rimedia a questa dimenticanza.

Corollario 8.1. *Nelle stesse ipotesi del teorema 8.2, lo stimatore* (8.7) *è uno stimatore consistente della varianza d'innovazione, ovvero*

$$\lim_{T\to\infty} \hat{\lambda}_T^2 = \lambda_0^2 \tag{8.14}$$

con probabilità uno.

Dimostrazione. Usando la (8.2) possiamo scrivere

$$\boldsymbol{\varepsilon}_{\hat{\theta}_T}(t) = G_{\hat{\theta}_T}(z)^{-1}\left[\mathbf{y}(t) - F_{\hat{\theta}_T}(z)\mathbf{u}(t)\right]$$

Nelle ipotesi in cui ci siamo posti, $\hat{\theta}_T$ converge al valore vero θ_0 e quindi passando al limite per $T \to \infty$ nell'espressione precedente è facile dimostrare che l'errore residuo di predizione converge all'innovazione vera $\mathbf{e}_0(t) = \boldsymbol{\varepsilon}_{\theta_0}(t)$. In effetti, dato che $V_T(\hat{\theta}_T)$ è la varianza campionaria di $\boldsymbol{\varepsilon}_{\hat{\theta}_T}(t)$ e che

$$\lim_{T\to\infty} V_T(\hat{\theta}_T) = \mathbb{E}_{\,\theta_0}\boldsymbol{\varepsilon}^2_{\hat{\theta}_0}(t) = \mathbb{E}_{\,0}\mathbf{e}_0^2(t)$$

si vede che anche la varianza di $\boldsymbol{\varepsilon}_{\hat{\theta}_T}(t)$ converge a quella di $\mathbf{e}_0(t)$. □

Il caso di parametrizzazioni indipendenti

Consideriamo un modello del tipo Box-Jenkins (5.9) 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

$$F_\theta(z) \equiv F_\xi(z)\,, \qquad G_\theta(z) \equiv G_\eta(z) \tag{8.15}$$

Ci si chiede se i risultati appena visti possono valere separatamente per le due classi parametriche di funzioni di trasferimento. In un certo senso, ci si chiede se e quando si può parlare di "consistenza parziale". Questa domanda è di interesse in pratica perchè è normalmente più importante identificare correttamente una delle due funzioni di trasferimanto (tipicamente quella della parte "deterministica" $F(z)$) dell'altra.

A questo scopo è sufficiente una nozione di *identificabilità parziale*. Dato che, come vedremo, la risposta al quesito precedente è positiva solo nel caso di processi senza reazione, daremo la definizione solo in questo caso.

Definizione 8.1. *Si assuma che le funzioni di trasferimento $F(z)$ e $G(z)$ nella famiglia* (5.9) *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*

$$S_{\mathbf{yu}}(\cdot;\xi_1) = S_{\mathbf{yu}}(\cdot;\xi_2) \;\Rightarrow\; \xi_1 = \xi_2 \tag{8.16}$$

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 (8.16) è equivalente alla

$$[F_{\xi_1}(z) - F_{\xi_2}(z)]\, S_{\mathbf{u}}(z) \equiv 0 \;\Rightarrow\; \xi_1 = \xi_2\,. \tag{8.17}$$

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 8.3. *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 $\varepsilon_\theta(t)$ 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:

$$\lim_{T\to\infty} \hat{\boldsymbol{\xi}}_T \;=\; \xi_0 \tag{8.18}$$

con probabilità uno. In altri termini, lo stimatore PEM del parametro ξ è consistente.

Dimostrazione. Usando l'espressione (8.2) 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

$$\varepsilon_\theta(t) = G_\eta(z)^{-1}\left[(F_0(z) - F_\xi(z))\mathbf{u}(t) + G_0(z)\mathbf{e}_0(t)\right] := L_{\xi,\,\eta}(z)\mathbf{u}(t) + M_\eta(z)\mathbf{e}_0(t)$$

dove i due addendi nell'ultimo termine sono scorrelati per l'assenza di reazione. Si trova così

$$\bar{V}(\theta) = \bar{V}(\xi,\,\eta) = \text{var}\,\left[L_{\xi,\,\eta}(z)\mathbf{u}(t)\right] + \text{var}\,\left[M_\eta(z)\mathbf{e}_0(t)\right]$$

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 $\hat{\boldsymbol{\theta}}_T$ converge con probabilità uno, sicuramente anche le sue prime componenti, $\hat{\boldsymbol{\xi}}_T$ 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 $\{\xi_0\}$. □

Questo risultato è utile per esaminare 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 $F(z)$. Ad esempio, anche usando modelli molto semplici, come i cosiddetti *modelli a errore di equazione* (*O.E.* = *output error models* in inglese) del tipo

$$\mathbf{y}(t) = F_\theta(z)\mathbf{u}(t) + \mathbf{e}(t) \tag{8.19}$$

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 8.1. *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* $\theta := [a\ b]^T$ *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,

$$\begin{aligned}\text{var } \boldsymbol{\epsilon}_\theta &= \lambda_0^2 + \text{var } [\hat{\mathbf{y}}_0(t \mid t-1) - \hat{\mathbf{y}}_\theta(t \mid t-1)] \\ &= \lambda_0^2 + \text{var } \left\{ \left[\frac{b_0}{1+a_0 z^{-1}} - \frac{b}{1+az^{-1}} \right] \mathbf{u}(t-1) + c_0 \mathbf{e}_0(t-1) \right\} \\ &= \lambda_0^2 + \text{var } \left\{ \left[\frac{b_0}{1+a_0 z^{-1}} - \frac{b}{1+az^{-1}} \right] \mathbf{u}(t-1) \right\} + c_0^2 \lambda_0^2\end{aligned}$$

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 $\theta_0 := [a_0 \;\; b_0]^T$ 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 8.3.

8.3 La distribuzione asintotica dello stimatore PEM

In questa sezione 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 $V_T(\theta)$ si ha in un punto in cui il gradiente si annulla, ovvero

$$\frac{\partial V_T(\theta)}{\partial \theta} := V_T(\theta)' = 0\,.$$

Ora, usando la formula di Taylor arrestata al secondo ordine nel punto $\theta = \theta_0$, si ha

$$V_T(\hat{\boldsymbol{\theta}}_T)' = V_T(\theta_0)' + V_T(\bar{\boldsymbol{\theta}})''(\hat{\boldsymbol{\theta}}_T - \theta_0) = 0 \tag{8.20}$$

dove $V_T(\bar{\boldsymbol{\theta}})''$ è la matrice delle derivate seconde (Hessiana) calcolata in un punto $\bar{\theta}$ dell'intervallo p-dimensionale di estremi θ_0 e $\hat{\boldsymbol{\theta}}_T$, ovvero

$$\theta_0^k \leq \bar{\boldsymbol{\theta}}^k \leq \hat{\boldsymbol{\theta}}_T^k, \qquad k = 1, 2, \ldots, p$$

Supponendo che la matrice Hessiana sia invertibile, dalla (8.20) si può ricavare

$$\hat{\boldsymbol{\theta}}_T - \theta_0 = -\left[\frac{1}{2} V_T(\bar{\boldsymbol{\theta}})'' \right]^{-1} \frac{1}{2} V_T(\theta_0)' \tag{8.21}$$

dove il fattore $\frac{1}{2}$ è stato introdotto per convenienza. Calcoliamo ora il gradiente e la matrice Hessiana usando l'espressione (8.4). Ponendo

$$\boldsymbol{\psi}_\theta(t) := \frac{\partial \hat{\mathbf{y}}_\theta(t \mid t-1)}{\partial \theta}$$

si trova

$$\frac{1}{2}V_T(\theta)' = -\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\psi}_\theta(t)\boldsymbol{\varepsilon}_\theta(t) \tag{8.22}$$

$$\frac{1}{2}V_T(\theta)'' = \frac{1}{T}\sum_{t=1}^{T}\left\{\boldsymbol{\psi}_\theta(t)\boldsymbol{\psi}_\theta(t)^\top - \boldsymbol{\varepsilon}_\theta(t)\left[\frac{\partial^2\boldsymbol{\varepsilon}_\theta(t)}{\partial\theta_i\partial\theta_j}\right]\right\} \tag{8.23}$$

Esaminiamo prima il comportamento asintotico della derivata seconda.

Lemma 8.2. *Nelle ipotesi poste, si ha*

$$\lim_{T\to\infty}\frac{1}{2}V_T(\bar{\boldsymbol{\theta}})'' = \mathbb{E}_{\theta_0}\left\{\boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^\top\right\} \tag{8.24}$$

con probabilità uno.

Dimostrazione. Nelle ipotesi in cui ci siamo messi, $\hat{\boldsymbol{\theta}}_T \to \theta_0$ e quindi anche $\bar{\boldsymbol{\theta}} \to \theta_0$ (con probabiltà uno) e la media temporale in (8.23) tende all 'aspettazione per cui,

$$\frac{1}{2}V_T(\bar{\boldsymbol{\theta}})'' \to \mathbb{E}_{\theta_0}\left\{\boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^\top - \boldsymbol{\varepsilon}_{\theta_0}(t)\left[\frac{\partial^2\boldsymbol{\varepsilon}_\theta(t)}{\partial\theta_i\partial\theta_j}\right]_{|\theta=\theta_0}\right\}$$

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 della (8.23) risultano scorrelati da $\mathbf{e}_0(t)$ e l'ultimo termine ha quindi aspettazione nulla. □

Per quanto riguarda l'altro termine nel prodotto (8.21), si ha

$$\frac{1}{2}V_T(\theta_0)' = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\psi}_{\theta_0}(t)\mathbf{e}_0(t) \tag{8.25}$$

Proposizione 8.1. *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,*

$$\sqrt{T}\frac{1}{2}V_T(\theta_0)' \xrightarrow{L} \mathcal{N}(0, Q) \tag{8.26}$$

Se la varianza condizionata di $\mathbf{e}_0(t)$ è indipendente dai dati $(\mathbf{y}^{t-1},\, \mathbf{u}^{t-1})$, ovvero se

$$\mathbb{E}_0\{\mathbf{e}_0(t)^2 \mid \mathbf{y}^{t-1},\, \mathbf{u}^{t-1}\} = \mathbb{E}_0\{\mathbf{e}_0(t)^2\} = \lambda_0^2\,, \tag{8.27}$$

la matrice varianza asintotica Q è data dalla formula,

$$Q = \lambda_0^2\,\mathbb{E}_0\{\boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^\top\}\,. \tag{8.28}$$

Dimostrazione. Il risultato scende dall'osservazione che $\{\boldsymbol{\psi}_{\theta_0}(t)\mathbf{e}_0(t)\}$ è ancora una d-martingala rispetto alla stessa famiglia di σ-algebre ed è in realtà un corollario immediato del teorema 7.2. L'unica cosa da dimostrare è l'espressione per la varianza asintotica, la quale scende dalla proprietà (**??**), che implica,

$$\mathrm{Var}\left\{\boldsymbol{\psi}_{\theta_0}(t)\mathbf{e}_0(t)\right\} \;=\; \mathbb{E}_0\{\mathbf{e}_0(t)^2\}\,\mathbb{E}_0\{\boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^\top\}\,.$$

□

Notiamo che le condizioni del teorema valgono in particolare se $\mathbf{e}_0$ è un processo i.i.d.. Mettendo assieme il lemma e la proposizione precedenti otteniamo infine il risultato seguente, che è il secondo risultato fondamentale della teoria asintotica della stima PEM.

Teorema 8.4. *Supponiamo che i dati siano ergodici del secondo ordine, il modello vero appartenga alla classe parametrica $\{M(\theta)\}$ e che vi sia identificabilità. Assumiamo inoltre che il processo innovazione $\mathbf{e}_0$ soddisfi alle condizioni descritte nella proposizione 8.1. Allora per lo stimatiore PEM vale il teorema del limite centrale, ovvero*

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \theta_0) \;\xrightarrow{L}\; \mathcal{N}(0, P)\,, \tag{8.29}$$

dove la varianza asintotica P è data dalla formula

$$P = \lambda_0^2\,\left[\mathbb{E}_0\{\boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^\top\}\right]^{-1}\,. \tag{8.30}$$

l'inversa della matrice tra parentesi quadre esiste.

Dimostrazione. Il risultato scende facilmente dalla 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 (8.26). La questione dell'invertibilità verrà chiarita nella prossima sezione. Si veda il corollario 8.2. □

Esempio 8.2. *Vogliamo identificare il sistema "vero":*

$$\mathbf{y}(t) = b_0\mathbf{u}(t-1) + \frac{1}{1+a_0z^{-1}}\mathbf{e}(t)$$

dove $\mathbf{u}$ *ed* $\mathbf{e}$ *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:*

$$M_2(\theta) := \{\mathbf{y}(t) + a\mathbf{y}(t-1) = b_1\mathbf{u}(t-1) + b_2\mathbf{u}(t-2) + \mathbf{e}(t)\,;\, \theta = [a\, b_1\, b_2]^\top\}$$

. 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 $\theta_0 = [a_0\ b_0\ b_0a_0]^\top$. In altri termini, quando $N \to \infty$, $\hat b_{2,N} \to b_0a_0$.

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:

$$\psi_\theta(t)^\top = [\frac{\partial\varepsilon_\theta(t)}{\partial a}\ \frac{\partial\varepsilon_\theta(t)}{\partial b}] = [\,\mathbf{y}(t-1) - b\mathbf{u}(t-2)\ \ -\mathbf{u}(t-1) - a\mathbf{u}(t-2)\,]$$

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:

$$\varepsilon_\theta(t) = \mathbf{y}(t) + a\mathbf{y}(t-1) - b_1\mathbf{u}(t-1) - b_2\mathbf{u}(t-2)$$

Il gradiente è

$$\psi_\theta(t) = [\frac{\partial\varepsilon_\theta(t)}{\partial a}\ \frac{\partial\varepsilon_\theta(t)}{\partial b_1}\ \frac{\partial\varepsilon_\theta(t)}{\partial b_2}] = [\,\mathbf{y}(t-1)\ \ -\mathbf{u}(t-1)\ \ -\mathbf{u}(t-2)\,]$$

Si vede facilmente che $\mathbb{E}_0\mathbf{y}(t-1)^2 = \mathrm{var}\ \mathbf{y}(t) = b_0^2\sigma^2 + \dfrac{\lambda_0^2}{1-a_0^2}$, per cui

$$R := \mathbb{E}_0\psi_\theta(t)\psi_\theta(t)^\top|_{\{\theta=\theta_0\}} = \begin{bmatrix} b_0^2\sigma^2 + \dfrac{\lambda_0^2}{1-a_0^2} & 0 & -b_0\sigma^2 \\ 0 & \sigma^2 & 0 \\ -b_0\sigma^2 & 0 & \sigma^2 \end{bmatrix}$$

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 minore con il primo modello che ha meno parametri.

8.4 La matrice d'informazione e il limite di Cramèr-Rao

In questa sezione deriveremo delle espressioni asintotiche per la matrice di Fisher e il limite di Cramèr-Rao facendo inizialmente riferimento ad un modello probabilistico congiunto delle variabili osservate di struttura generale, del tipo

$$p_\theta(y^T,\, u^T), \qquad \theta \in \Theta$$

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[21]

$$\begin{aligned} p_\theta(y^T,\, u^T) &= p_\theta(y_T,\, u_T \mid y^{T-1},\, u^{T-1}) p_\theta(y^{T-1},\, u^{T-1}) \\ &= p_\theta(y_T, \mid y^{T-1},\, u^{T-1}) p(u_T \mid y^T,\, u^{T-1}) p_\theta(y^{T-1},\, u^{T-1}) \\ &= \ldots \\ &= \prod_{t=1}^{T} p_\theta(y_t \mid y^{t-1},\, u^{t-1}) \prod_{t=1}^{T} p(u_t \mid y^t,\, u^{t-1}) \end{aligned} \tag{8.31}$$

Abbiamo soppresso la dipendenza dal parametro θ nelle probabilità condizionate $p(u_t \mid y^t,\, u^{t-1}),\ t = 1, 2, \ldots$ che descrivono il canale di reazione, dato che non siamo interessati alla sua modellizzazione. Supponiamo che questa famiglia di densità descriva (una famiglia parametrica di) processi *stazionari* per cui nella decomposizione

$$\mathbf{y}(t) = \mathbb{E}_\theta \left[\mathbf{y}(t) \mid \mathbf{y}^{t-1},\, \mathbf{u}^{t-1} \right] + \mathbf{e}(t) := \hat{\mathbf{y}}_\theta(t \mid t-1) + \mathbf{e}(t) \tag{8.32}$$

la densità di probabilità del processo $\mathbf{e}$ *al limite per* $t \to \infty$ *non dipenda dal parametro* θ. Abusando di questo fatto, nella decomposizione (8.32) abbiamo già implicitamente ignorato la dipendenza di $\mathbf{e}$ da θ.

[21] 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.

Questo è quanto accade nel caso di un modello razionale Gaussiano in cui,

$$p_\theta(y_t \mid y^{t-1},\, u^{t-1}) = \frac{1}{\sqrt{2\pi\lambda_\theta^2(t)}}\, \exp -\frac{1}{2}\frac{[y_t - \hat{y}_\theta(t \mid t-1)]^2}{\lambda_\theta^2(t)} \tag{8.33}$$

e, come è ben noto dalla teoria del filtro di Kalman, la varianza dell'innovazione (transitoria) $\lambda_\theta^2(t)$ converge ad una costante λ^2 indipendente dai parametri del modello, quando $t \to \infty$.

Confortati da questo esempio, assumeremo che t sia abbastanza grande da poter scrivere

$$p_\theta(y_t \mid y^{t-1},\, u^{t-1}) = p_{\mathbf{e}}(y_t - \hat{y}_\theta(t \mid t-1)) \tag{8.34}$$

dove $p_{\mathbf{e}}$ denota la densità di probabilità dell'innovazione stazionaria, che non dipende da θ. In altri termini 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 allora il vettore delle sensitività,

$$\mathbf{z}_\theta := \frac{\partial \log p_\theta(\mathbf{y}^T\, \mathbf{u}^T)}{\partial\theta} = \sum_{t=1}^{T} \frac{\partial \log p_\theta(\mathbf{y}(t) \mid \mathbf{y}^{t-1}\, \mathbf{u}^{t-1})}{\partial\theta}$$

Supponendo che $T \to \infty$ e trascurando i primi termini nella somma, si può pensare di essere in regime stazionario per cui si può assumere che valga la rappresentazione (8.34). Senza perdita di generalità, dopo aver eliminato i primi termini transitori, possiamo pensare di ri-inizializzare le somme all'istante $t = 1$, ottenedo così,

$$\mathbf{z}_\theta = \sum_{t=1}^{T} \frac{\partial \log p_{\mathbf{e}}(\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t \mid t-1))}{\partial\theta} = \sum_{t=1}^{T} \frac{\partial \log p_{\mathbf{e}}(\varepsilon)}{\partial\varepsilon}\Big|_{\varepsilon=\mathbf{e}(t)} \frac{\partial\, \hat{\mathbf{y}}_\theta(t \mid t-1))}{\partial\theta}$$

dove il predittore è quello stazionario. Ponendo $\dfrac{\partial \log p_{\mathbf{e}}(\varepsilon)}{\partial\varepsilon}\Big|_{\varepsilon=\mathbf{e}(t)} := \ell'(\mathbf{e}(t))$, si arriva così all'espressione della matrice di Fisher

$$I_T(\theta) = \mathbb{E}_\theta\{\mathbf{z}_\theta \mathbf{z}_\theta^\top\} = \sum_{t,s=1}^{T} \mathbb{E}_\theta\{\ell'(\mathbf{e}(t))\ell'(\mathbf{e}(s))\boldsymbol{\psi}_\theta(t)\boldsymbol{\psi}_\theta(s)^\top\} \tag{8.35}$$

e al seguente risultato.

Teorema 8.5. *Se* $\mathbf{e}$ *è un processo i.i.d., oppure,* $\ell'(\mathbf{e}(t))$ *è una funzione lineare di* $\mathbf{e}(t)$*, il che accade in particolare nel caso di distribuzioni Gaussiane, si ha*

$$I_T(\theta) = \frac{T}{\kappa^2}\mathbb{E}_\theta\{\boldsymbol{\psi}_\theta(t)\boldsymbol{\psi}_\theta(t)^\top\} \tag{8.36}$$

dove $\kappa^2 = \mathbb{E}_\theta[\ell'(\mathbf{e}(t))]^2$*. Nel caso Gaussiano,* $\kappa^2 = \{\mathrm{var}\,\{\mathbf{e}(t)\} = \lambda^2$.

Notiamo in effetti che nel caso Gaussiano, $\dfrac{\partial \log p_{\mathbf{e}}(\varepsilon)}{\partial\varepsilon}\Big|_{\varepsilon=\mathbf{e}(t)} = -\dfrac{\mathbf{e}(t)}{\lambda^2}$.

Ricordando il teorema di Rothenberg 2.2, otteniamo il seguente utile corollario.

Corollario 8.2. *Nelle ipotesi poste, il modello* (8.32) *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 $\frac{1}{T}$. In particolare, nel caso di modelli Gaussiani si ha

$$I_T(\theta)^{-1} \simeq \frac{\lambda^2}{T}\mathbb{E}_\theta\{\boldsymbol{\psi}_\theta(t)\boldsymbol{\psi}_\theta(t)^\top\}^{-1}, \qquad T \to \infty.$$

Possiamo così concludere con una condizione per l' efficienza asintotica dello stiamtore PEM.

Teorema 8.6. *Supponiamo che valgano le stesse ipotesi del teorema 8.4 e che il modello che genera i dati sia Gaussiano. Allora lo stimatore PEM è asintoticamente efficiente; i.e. per $T \to \infty$,*

$$\mathrm{Var}\,\{\hat{\boldsymbol{\theta}}_T\} - I_T(\theta_0)^{-1} \to 0 \tag{8.37}$$

Equivalentemente, per $T \to \infty$, la varianza di $\hat{\boldsymbol{\theta}}_T$ 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.

8.5 Relazione tra stima parametrica sul Modello lineare statico e stima PEM su modelli dinamici lineari.

A questo punto possiamo mettere in luce una notevole relazione che esiste tra la teoria della stima parametrica sul modello lineare-Gaussiano (3.7) che abbiamo presentato nei capitoli 3 e 4 e l'analisi asintotica degli stimatori PEM che stiamo presentando in questo capitolo.

Supponiamo che le osservazioni $\mathbf{y}$ del nostro modello lineare statico siano prodotte da un modello "vero" corrispondente al parametro "vero" θ_0,

$$\mathbf{y} = S\theta_0 + \sigma\mathbf{w} \tag{8.38}$$

e supponiamo anche di aver normalizzato le variabili nel modello lineare moltiplicando a sinistra per l'inverso del fattore di Cholesky L della covarianza di rumore R, in modo da ridurci a un rumore additivo $\sigma\mathbf{w}$ di varianza $\sigma^2 I_N$. In queste condizioni la stima di massima verosimiglianza del parametro θ si può scrivere

$$\hat{\boldsymbol{\theta}}_N = \left[\frac{1}{N}S^\top S\right]^{-1}\frac{1}{N}S^\top\mathbf{y} = \theta_0 + \left[\frac{1}{N}S^\top S\right]^{-1}\frac{1}{N}S^\top\mathbf{w}$$

ovvero

$$\hat{\boldsymbol{\theta}}_N - \theta_0 = Q_N^{-1}\frac{1}{N}S^\top\mathbf{w}, \qquad Q_N := \frac{1}{N}S^\top S \tag{8.39}$$

mentre nelle condizioni che garantiscono la consistenza e la normalità asintotica dello stimatore PEM, usando la (8.21), si può scrivere,

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \theta_0) \simeq \left[\frac{1}{N}\sum_{t=1}^{N} \boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^\top\right]^{-1} \frac{1}{\sqrt{N}}\sum_{t=1}^{N} \boldsymbol{\psi}_{\theta_0}(t)\mathbf{e}_0(t) \tag{8.40}$$

$$\simeq \bar{Q}^{-1}\frac{1}{\sqrt{N}}\sum_{t=1}^{N} \boldsymbol{\psi}_{\theta_0}(t)\mathbf{e}_0(t)\,, \qquad \bar{Q} := E_0\{\boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^\top\} \tag{8.41}$$

dove il simbolo $\simeq$ significa equivalenza asintotica (stesso limite in legge).

Notiamo subito che le (8.39) e la (8.40) coincidono pur di porre:

$$S = \begin{bmatrix} \boldsymbol{\psi}_{\theta_0}(1)^\top \\ \vdots \\ \boldsymbol{\psi}_{\theta_0}(N)^\top \end{bmatrix}, \qquad \sigma\mathbf{w} = \begin{bmatrix} \mathbf{e}_0(1) \\ \vdots \\ \mathbf{e}_0(N) \end{bmatrix}, \tag{8.42}$$

per cui in particolare $\sigma^2 = \lambda_0^2$ e si arriva al seguente importante risultato.

Teorema 8.7. *La distribuzione asintotica dello stimatore PEM del parametro θ per un qualunque modello dinamico lineare che soddisfi le ipotesi di consistenza e normalità asintotica, coincide con la distribuzione (Gaussiana) dello stimatore di massima verosimiglianza del parametro θ nel modello lineare statico* (8.38)*, in cui*

- *si siano fatte le sostituzioni* (8.42)*, supponendo che i vettori $\{\boldsymbol{\psi}_{\theta_0}(t)\,;\, t = 1, 2, \ldots, N\}$ siano quantità deterministiche,*
- *si sia sostituita alla matrice Q_N il suo limite per $N \to \infty$, uguale alla matrice $\bar{Q}$ nella formula* (8.41)*.*

Usando questo risultato si può quindi stabilire una corrispondenza biunivoca tra procedimenti di inferenza statistica "statici" sul modello (8.38) e procedimenti basati su statistiche PEM su un qualunque modello lineare che soddisfa alle ipotesi di consistenza e normalità asintotica.

Bibliografia

[1] S. Bernstein. Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes. *Math. Ann.*, 97:1–59, 1926.

[2] P. Billingsley. *Convergence of probability measures*. Wiley, 1968.

[3] G.D. Birkhoff. Proof of the ergodic hypothesis. *Proc. Nat. Acad. Sciences (USA)*, 17:565–600, 1931.

[4] J. R. Bunch and C.P. Nielsen. Updating the singular value decomposition. *Numer. Math.*, 31:111–129, 1978.

[5] W. G. Cochran. *Sampling Techniques (Third Ed.)*. Wiley, 1977.

[6] Harald Cramèr. *Mathematical Methods of Statistics*. Princeton University Press, 1946.

[7] J. L. Doob. *Stochastic Processes*. Wiley, 1953.

[8] H. Dym and H. P. McKean. *Fourier series and integrals*. Academic Press, 1972. Probability and Mathematical Statistics, No. 14.

[9] T. Ferguson. *A Course in Large Sample Theory*. Chapman and Hall, 1996.

[10] W. Freiberger and U. Grenander. *A short Course in Computational Probability and Statistics*. Springer Verlag, Berlin, 1971.

[11] K.F. Gauss. *Theoria Motus Corporum Coelestium, Liber II of Werke*. Julius Springer, Berlin, 1901.

[12] G.H. Golub and C.R. Van Loan. *Matrix Computation (Third ed.)*. The Johns Hopkins Univ. Studies in the Mathematical Sciences, 1996.

[13] G.H. Golub and G.P.H. Styan. Some aspects of numerical computation for linear models. In *Proceedings of the 7-th annual symposium on the interface of computer science and statistics*, pages 189–192, Iowa State University, 1973.

[14] E. J. Hannan. *Multiple Time Series*. Wiley, 1970.

[15] E.J. Hannan and M. Deistler. *The Statistical Theory of Linear Systems*. Wiley, 1998.

[16] C.L. Lawson and R.J. Hanson. *Solving Least Squares Problems*. Prentice Hall, Englewood Cliffs, 1974.

[17] E. Lehmann. *Testing Statistical Hypotheses (second Ed.)*. Wiley, 1986. reprinted by Springer Verlag.

[18] L. Ljung. *System Identification, Advances and Case Studies*, chapter On the consistency of prediction-error identificaion methods. Academic Press, New York, 1976.

[19] L. Ljung. *System Identification, Theory for the User (second Ed.)*. Prentice Hall, Englewood Cliffs, 1999.

[20] MATLAB. *Using MATLAB Version 6*. The MathWorks Inc., 2002.

[21] G. Picci. *Filtraggio Statistico (Wiener, Levinson, Kalman) e applicazioni*. Libreria Progetto Padova, 1994.

[22] M. Rosenblatt. A central limit theorem and a strong mixing condition. *Proc. Natnl. Acad. Sci. (USA)*, 42:43–47, 1956.

[23] Y. A. Rozanov. *Stationary Random Processes*. Holden-Day, San Francisco, 1967.

[24] H. Scheffè. *The Analysis of Variance*. Wiley, 1953.

[25] T. Söderström and P. Stoica. *System Identification*. Prentice-Hall, 1989.

[26] G.W. Stewart. Collinearity and least squares regression. *Statistical Science*, 2:68–100, 1987.

[27] E. Zacks. *The Theory of Statistical Inference*. Wiley, 1970.