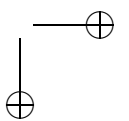
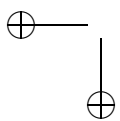
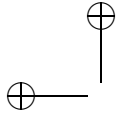


1

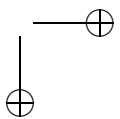
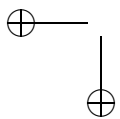
i

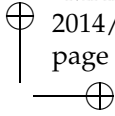
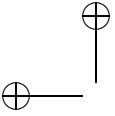




ii

---





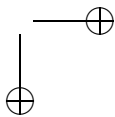
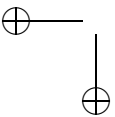
2014/10/  
page i

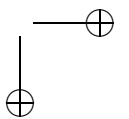
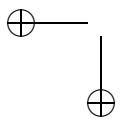
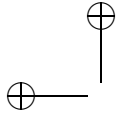
# STATISTICAL METHODS FOR THE IDENTIFICATION OF LINEAR DYNAMICAL SYSTEMS

Giorgio Picci<sup>‡</sup>

October 1, 2014

<sup>1‡</sup> Dipartimento di Ingegneria dell' Informazione, Universita' di Padova, 35131 Padova,  
Italy.



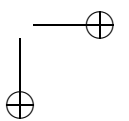
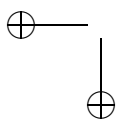
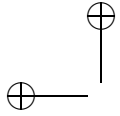


# INDEX

|               |  |            |
|---------------|--|------------|
| <b>Prefae</b> |  | <b>vii</b> |
| 0.1           | Introduction . . . . .   | vii        |
| 0.2           | What is System Identification? . . . . .                             | vii        |
| 0.3           | The essential features of the Identification Problem . . . . .       | viii       |
| 0.4           | Stationary signals and the statistical theory of model building . .  | xi         |
| 0.5           | Related areas . . . . .  | xii        |
| <b>1</b>      | <b>STATISTICAL INFERENCE AND PARAMETER ESTIMATION</b>                | <b>1</b>   |
| 1.1           | Introduction . . . . .   | 1          |
| 1.2           | Classical Theory of Parameter estimation . . . . .                   | 9          |
| 1.2.1         | The Cramèr-Rao Inequality . . . . .                                  | 10         |
| 1.2.2         | Identifiability . . . . .  | 16         |
| 1.3           | Maximum Likelihood . . . . .   | 20         |
| 1.3.1         | Il Metodo dei Momenti . . . . .                                      | 25         |
| <b>2</b>      | <b>LINEAR MODELS</b>   | <b>27</b>  |
| 2.1           | Modelli Statistici Lineari . . . . .                                 | 27         |
| 2.1.1         | Sul processo di Modellizzazione statistica . . . . .                 | 29         |
| 2.2           | Stima di M.V. nel modello lineare . . . . .                          | 32         |
| 2.3           | La distribuzione $\chi^2$ . . . . .                                  | 36         |
| 2.4           | Il principio dei Minimi Quadrati e il suo significato statistico . . | 41         |
| 2.5           | Generalizzazione a misure vettoriali . . . . .                       | 51         |
| 2.6           | Minimi quadrati e modelli non lineari . . . . .                      | 52         |
| 2.7           | Static Neural Networks . . . . .                                     | 53         |
| 2.8           | Universal approximating functions . . . . .                          | 54         |
| 2.9           | Gradient descent and back-propagation . . . . .                      | 54         |
| 2.10          | Aspetti numerici dei problemi ai minimi quadrati . . . . .           | 55         |
| 2.10.1        | La Decomposizione ai Valori Singolari (SVD) . . . . .                | 60         |
| <b>3</b>      | <b>MODELLI DINAMICI PER L'IDENTIFICAZIONE</b>                        | <b>71</b>  |
| 3.1           | Modelli statistici lineari per processi del secondo ordine . . . . . | 71         |
| 3.2           | Modelli parametrici e identificabilità in assenza di retroazione .   | 76         |
| 3.3           | Persistente Eccitazione . . . . .                                    | 81         |
| 3.4           | Alcune classi di modelli e loro parametrizzazione . . . . .          | 84         |

|          |   |            |
|----------|---|------------|
| 3.5      | Identificabilità in presenza di reazione . . . . .                    | 86         |
| 3.5.1    | Modelli a errori nelle variabili . . . . .                            | 89         |
| 3.6      | Modelli multivariabili . . . . .                                      | 90         |
| 3.7      | Modelli Gaussiani . . . . .   | 90         |
| <b>4</b> | <b>ERGODICITÀ</b>   | <b>91</b>  |
| 4.1      | Proprietà asintotiche degli stimatori: Consistenza . . . . .          | 91         |
| 4.2      | Ergodicità: motivazioni fisiche . . . . .                             | 93         |
| 4.3      | Ergodicità: teoria assiomatica . . . . .                              | 96         |
| 4.3.1    | Ergodicità e inferenza statistica . . . . .                           | 103        |
| 4.4      | Processi p.n.d. e processi dissolventi (mixing) . . . . .             | 107        |
| 4.5      | Ergodicità del secondo ordine . . . . .                               | 110        |
| 4.6      | Sull'ipotesi ergodica in statistica . . . . .                         | 113        |
| 4.7      | Consistenza dello Stimatore di Massima Verosimiglianza* . . . . .     | 117        |
| <b>5</b> | <b>CENTRAL LIMIT THEOREMS</b>   | <b>125</b> |
| 5.1      | Convergenza in legge . . . . .  | 125        |
| 5.2      | Il teorema del limite centrale per d-martingale stazionarie . . . . . | 128        |
| 5.3      | TLC per processi stazionari . . . . .                                 | 131        |
| 5.4      | Sistemi lineari e TLC . . . . .                                       | 140        |
| 5.5      | Efficienza asintotica . . . . .                                       | 141        |
| <b>6</b> | <b>METODI PEM</b>   | <b>143</b> |
| 6.1      | Introduzione . . . . .  | 143        |
| 6.2      | Analisi asintotica dello stimatore PEM . . . . .                      | 148        |
| 6.3      | La distribuzione asintotica dello stimatore PEM . . . . .             | 160        |
| 6.4      | La matrice d'informazione e il limite di Cramèr-Rao . . . . .         | 167        |
| 6.5      | Modello lineare e stima PEM . . . . .                                 | 170        |
| 6.6      | Analisi asintotica delle stime di funzioni di trasferimento . . . . . | 171        |
| 6.7      | Aspetti computazionali . . . . .                                      | 175        |
| 6.8      | Algoritmi ricorsivi . . . . .   | 179        |
| <b>7</b> | <b>HYPOTHESIS TESTING</b>   | <b>181</b> |
| 7.1      | Ipotesi semplici . . . . .  | 183        |
| 7.1.1    | Critica all'approccio classico . . . . .                              | 194        |
| 7.2      | Ipotesi composte . . . . .  | 194        |
| 7.2.1    | La distribuzione di Student . . . . .                                 | 197        |
| 7.3      | Tests di ipotesi sul modello lineare . . . . .                        | 201        |
| 7.4      | Calcolo del rapporto di Max verosimiglianza . . . . .                 | 204        |
| 7.4.1    | La distribuzione $F$ . . . . .  | 207        |
| 7.5      | Applicazione all'analisi della varianza . . . . .                     | 208        |
| <b>8</b> | <b>MULTIPLE REGRESSION</b>  | <b>217</b> |
| 8.1      | Stima della complessità di un modello lineare . . . . .               | 217        |
| 8.2      | Regressione lineare a stadi . . . . .                                 | 219        |
| 8.3      | Il test $F$ . . . . .   | 226        |

| INDEX     |  | v          |
|-----------|--|------------|
| 8.4       | Stima della dimensione del modello col criterio FPE . . . . .            | 229        |
| 8.5       | Un algoritmo di regressione lineare a stadi . . . . .                    | 230        |
| <b>9</b>  | <b>ALMOST PERIODIC SIGNALS</b>   | <b>237</b> |
| 9.1       | Rappresentazione di processi puramente deterministici . . . . .          | 238        |
| 9.2       | Metodi non parametrici per la stima di spettri . . . . .                 | 241        |
| 9.3       | Metodi di finestra spettrale . . . . .                                   | 245        |
| 9.4       | Stimatori parametrici di spettro . . . . .                               | 249        |
| 9.5       | Stima di frequenze col metodo PEM . . . . .                              | 249        |
| 9.5.1     | Filtri Notch come modelli ARMA a poli sul cerchio . . . . .              | 251        |
| 9.6       | Stimatori basati sulla correlazione campionaria . . . . .                | 254        |
| 9.6.1     | Espressione matriciale della correlazione . . . . .                      | 254        |
| 9.6.2     | Alcuni metodi basati sulla correlazione campionaria . . . . .            | 258        |
| <b>10</b> | <b>VALIDAZIONE E STIMA DELLA STRUTTURA</b>                               | <b>267</b> |
| 10.1      | Tests di bianchezza dei residui . . . . .                                | 268        |
| 10.1.1    | Il Test del Correlogramma . . . . .                                      | 269        |
| 10.1.2    | Un test di incorrelazione dagli ingressi passati . . . . .               | 271        |
| 10.1.3    | Il test del periodogramma cumulato . . . . .                             | 271        |
| 10.1.4    | Il test $F$ per modelli lineari identificati col metodo<br>PEM . . . . . | 271        |
| 10.2      | Stima dell'ordine . . . . .  | 271        |
| <b>A</b>  | <b>APPENDICE</b>   | <b>273</b> |
| A.1       | Alcuni richiami di teoria della probabilità . . . . .                    | 273        |
| A.1.1     | Sulla nozione di $\sigma$ -algebra . . . . .                             | 273        |
| A.2       | Hilbert space of second-order random variables . . . . .                 | 274        |
| A.3       | Proiezioni ortogonali e media condizionata . . . . .                     | 275        |
| A.3.1     | Integrabilità uniforme . . . . .   | 278        |
|           | <b>BIBLIOGRAPHY</b>  | <b>281</b> |





# PREFACE

## 0.1 Introduction

The importance of System Identification, that is, the automatic construction of mathematical models of dynamical systems from observed data, has grown tremendously in the last decades. Identification techniques have found application in diverse fields like automatic control, econometrics, geophysics, hydrology, structural testing in civil engineering, bioengineering, automotive science, to name just a few principal areas. In particular, recursive identification techniques, have found application in the design and real-time monitoring of industrial processes and in adaptive control and communication systems.

Naturally, the pervasive use of mathematical models in modern science and technology has been afforded and greatly stimulated by the massive diffusion of computers. One could safely say that the enormous progress of microelectronics and computer hardware and the dramatic increase of real-time computing power available after the 1990's are leading to a shift of paradigms in the design of engineering systems. To cope with the growing complexity and the rising demand for sophistication and performance, the design of modern control and communication systems has to be based more than ever before on *quantitative models* of the signals and systems involved. For example, on-line identification algorithms have become a key ingredient in signal processing, where there is a growing demand for modeling procedures which are adapted to the dynamic structure of various types of channels and signals encountered in the applications. Early examples of successful application of this principle have been model-based coding and recognition of audio and video signals. Devices based on these ideas are now part of commercially available communication systems.

## 0.2 What is System Identification?

Consider a physical/economic/biological system, say a paper machine, a power plant or the stock exchange market in Bulgaria. By a "model" of the system we shall mean a mathematical description of the relations existing among certain observed variables of the system. In general the models we are most interested in this book are *dynamical models* describing the temporal evolution of the observed

variables.

The scope of *system identification*, also called *data-based modeling*, is to construct a conceptual framework and algorithms for automatic model building from observed data. The observed variables, which are accessible to measurement, are usually classified as "inputs" or *exogenous* variables, denoted ( $\mathbf{u}$ ), and "outputs" or *explained* variables denoted ( $\mathbf{y}$ ). Normally the variables can only be measured at discrete instants of time  $t$  and collected in a string of data called a *time series* in the statistical literature.

In these lectures we shall mainly discuss the case where the data are collected in one irrepitable experiment. Often no preparation of the experiment is possible (i.e. we cannot choose the experimental conditions or the input function to the system at our will) and one is forced to do the best with the results of that irrepitable experiment.

There may be a variety of different reasons to build models. Here we shall be chiefly interested in model building for the purpose of prediction and control. This means that the identified model should be useful for prediction or control of *future*; i.e. not yet observed, values taken on by the system variables. This fundamental requirement imposes certain restrictions on the type of data which can actually be used for identification.

The allowable models generally belong to a preselected model class, say the class of finite-dimensional linear time-invariant systems of a given order and the identification problem is generally formulated as that of inferring a "best" mathematical model in the model class on the basis of the observed data. So we may say that the identification problem consists of three ingredients: *the data*, *the model class* and *the model selection criterion*. We shall discuss these three ingredients in the next subsection.

### 0.3 The essential features of the Identification Problem

#### The Model class:

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

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

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

If we accept the arguments above it is clear that one essential issue to be addressed for a realistic formulation of the problem is a satisfactory notion of non-rigid, i.e. *approximate*, mathematical modeling of the observed data. The meaning of the word "approximate" should here be understood in the sense that a model should be able to accept as legitimate, data sets (time series) which may possibly differ slightly from each another. Imposing rigid "exact" descriptions of the type  $F(u, y) = 0$  to experimental data has been criticized since the early beginnings of experimental science. Particularly illuminating is Gauss' general philosophical discussion in [19] sect. III, p. 236.

More to the point, there has been a widespread belief in the early years of control science that identification was merely a matter of describing (exactly) the measured data by linear convolution equations of the type

$$y(t) = \sum_{t_0}^t h(t - \tau)u(\tau) \quad (0.3.1)$$

or, equivalently, by matching exactly pointwise harmonic response data with linear transfer function models. Results have always been extremely sensitive even to small perturbations in the data. New incoming data tend to change the model drastically, which means that a model determined in this way has in fact very poor predictive capabilities. The reason is that data obey exactly rigid relations of this kind "with probability zero". If in addition the model class is restricted to be finite-dimensional, which of course is what is really necessary for control applications, imposing the integral equation model (0.3.1) on real data normally leads to disastrous results. This is by now very well-known and documented in the early literature, see e.g. [43, 66, 27, 16]. The fact, expressed in the language of numerical analysis, is that fitting rigid models to discrete data invariably leads to ill-conditioned problems.

Gauss idea of describing data by a *distribution function* is a prime example of thinking in terms of (non-rigid) approximate models<sup>1</sup>. Other alternatives are possible, say using model classes consisting of a rigid "exact" model as a "nominal" object, plus an uncertainty ball around it. In this case, besides a nominal model, the identification procedure is required to provide at least bounds on the magnitude of the relative "uncertainty region" around the nominal model. This type of modeling philosophy is motivated by use in  $H^\infty$  control applications. Here one

<sup>1</sup>A vulgar belief attributes to Gauss the invention of least squares, which is historically not correct. In Gauss' work least squares come out as a solution method for optimally fitting a certain class of *density functions* to the observed data.

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

### The Data:

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

*Mathematical descriptions of uncertain signals* can be quite diverse. Possible choices are stochastic processes or deterministic signals consisting of a nominal path plus an uncertainty bound, sometimes fuzzy things, etc. The crucial difference between theories of model building lies in the background methodology they use for modeling uncertain signals<sup>2</sup>.

### The Methodology:

In these lectures we shall take the mainstream route and model uncertainty with the apparatus of probability theory. In this framework *identification is phrased as a problem of mathematical statistics*.

One could argue that the basic problem of identification is, much more than designing algorithms which fit models to observed data (the easy part), the quantification of the *uncertainty bounds* or the description of the *dynamic errors* which will be incurred when using the model with generic data. Any sensible identification method should provide some mathematical description of how uncertainty is distributed in time or frequency about the nominal identified model. In this respect the probabilistic approach offers a well-established bag of tools. In this setup, at least in the linear wide-sense setting, model uncertainty turns out to be equivalent to *additive random disturbances*; in other words, identifying model uncertainty is equivalent to identifying models for additive stochastic terms affecting a nominal “deterministic” nominal input-output relation. We shall discuss this point in detail in Chapter 3.

---

<sup>2</sup>For this reason we would not like to classify as identification “exact modeling” where the data are certain (i.e. not random) signals assumed to fit exactly some finite set of (linear) relations.

## 0.4 Stationary signals and the statistical theory of model building

Since identification for the purpose of prediction and control makes sense only if you can use the identified model to describe future data, i.e. different data than those employed for its calibration, at the roots of any data-based model building procedure there must be a formalization of the belief that

*future data will continue to be generated by the same "underlying mechanism" that has produced the actual data.*

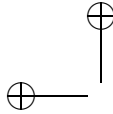
This is a vague but basic assumption on the nature of the data, which are postulated to keep being "statistically the same" in the future. Besides being inherent in the very *purpose* of collecting data for model building this assumption offers the logical background for assessing the *quality* of the identified model, by *asymptotic statistical analysis*, i.e. by assessing the quality of finite-sample models comparing them with the "theoretically best achievable" models which could be identified with data of infinite length. One could probably say that the usefulness of statistics is founded on asymptotic analysis, and that the wide use of statistics and of probabilistic methods in identification is mainly motivated by the large body of effective asymptotic tools which can be applied to assess some basic "quality" features of the estimated model.

Classical statistics traditionally starts by postulating some urn model whereby the data are imagined as being drawn at random from some universe of possible values in a "random trial" where "nature" chooses according to some probability law the current state of the experimental conditions.

It has been argued that this abstract urn model of probability theory looks inadequate to deal with situations like the one we have envisaged, where there is just one irrepitable experiment and there is really no sample space around from which the results of the experiment could possibly have been drawn. The critique has the merit of bringing up an important issue. It should be admitted that in large sectors of the literature the statistical framework is often imposed rather dogmatically. Sometimes statistical procedures are imposed to problems when there really seems to be no physical ground for their applicability. The user is normally left alone wondering if his problem is "stochastic" enough to be authorized to apply algorithm *A*, or his data are instead "deterministic" and he should apply algorithm *B* instead.

In our opinion however, the critique originates from a tendency to confuse physical reality with mathematical modelling. In effect the urn model (the underlying probability space) is just a mathematical device which is not required to have any physical meaning or interpretation and can in principle be used to model anything.

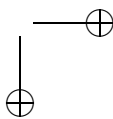
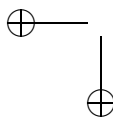
One of the goals of this book is to provide a formal justification for the adoption of the probabilistic description of uncertain data and of statistical techniques in model-building. Formal arguments leading to a probabilistic description of certain types of data can be based on a notion of *stationarity* of the observed signals, a natural condition introduced and discussed in Chapter 4, meant to capture the



idea that future data should be “statistically the same” as past data.

## 0.5 Related areas

There are many areas of statistical modeling which are related to identification such as pattern recognition, cluster analysis and data classification, statistical learning theory, supervised and unsupervised learning, machine learning, support vector machines, neural networks,.. to name just the most widely quoted nicknames see e.g. [50, ?].



## CHAPTER 1

# BACKGROUND ON STATISTICAL INFERENCE AND PARAMETER ESTIMATION

### 1.1 Introduction

Modern Probability Theory is *axiomatic*. It assumes an abstract model of reality consisting of a space of elementary events  $\Omega$  (for example the set of all possible outcomes of a measurement process), a “ $\sigma$ -algebra”  $\mathcal{A}$  of observable *events* (the subsets of  $\Omega$  which are “probabilizable”) and a *probability measure*  $P$ , defined on  $\mathcal{A}$ , obeying a set of well-known axioms.

While it is often rather easy (and in any case quite arbitrary) to describe the set of all possible outcomes of an experiment by a set  $\Omega$  and the class of interesting events by a  $\sigma$ -algebra of subsets of  $\Omega$  (think for example of throwing a dice or of the measurement of the length of a table), except for a very limited number of rather simple situations, the process by which one assigns a probability  $P$  to the space  $\{\Omega, \mathcal{A}\}$ , is a priori not obvious at all.

This process constitutes the subject of Statistics.

One could well say that the scope of Statistics is to assign probabilities on the basis of experimental evidence. This means that assigning a certain measure  $P$  to a given space of experiments  $\{\Omega, \mathcal{A}\}$  is an *inductive process* which requires an interpretation or, better, a rational extrapolation made on certain experimental data. By its very nature, therefore, the assignment of a probability is never *certain*. There are several criteria which may lead to a decision that a certain  $P$  describes “well” the results of an experiment but these criteria may have different purposes and merits and may even not be comparable on an objective basis.

Typically, a statistical inference problem consists of:

- A space of experiments  $\{\Omega, \mathcal{A}\}$ ;
- A family  $\mathcal{P}$ , or a number of disjoint families  $\mathcal{P}_i, i = 1, \dots, k$  ( $k$  finite), of candidate probability measures  $P$  on  $\{\Omega, \mathcal{A}\}$ ;
- The outcome of an experiment,  $\bar{\omega}, \bar{\omega} \in \Omega$  ( $\bar{\omega}$  is the observation; i.e. the measured experimental data).

## 2 CHAPTER 1. STATISTICAL INFERENCE AND PARAMETER ESTIMATION

The inference problems are traditionally classified in two broad categories:

**Estimation:** On the basis of the experimental data  $\bar{\omega}$ , assign an admissible probability measure, i.e. an element  $P = P(\bar{\omega}) \in \mathcal{P}$ .

**Hypothesis Testing:** On the basis of the experimental data  $\bar{\omega}$ , assign  $P$  to one of the subclasses  $\mathcal{P}_i$  (in other words, decide to which subclass  $\mathcal{P}_i$  it belongs to).

In both cases one is asked to construct (based on some inference criterion) a function  $\bar{\omega} \rightarrow \mathcal{P}$ , or  $\bar{\omega} \rightarrow \{1, 2, \dots, k\}$ . The distinction between estimation and hypothesis testing is actually between an infinite versus a finite number of possible alternatives.

### An elementary example

Assume we are tossing a coin and let  $p :=$  probability to observe head, event which will be denoted by the symbol  $T$  and  $1 - p :=$  probability that tail will show instead; event which is denoted by the symbol  $C$ . Naturally,  $p$  is unknown. We want to obtain information on the value of  $p$  by tossing the coin  $N$  consecutive times, assuming that each toss *does not influence the outcome of the other tosses*.

Let  $\Omega = \{\text{all possible outcomes of } N \text{ consecutive tosses}\}$ . The set  $\Omega$  contains all sequences made of  $N$  symbols  $T$  and  $C$  in any possible order. Let  $\mathcal{A}$  be the family of all subsets of  $\Omega$ . It is well-known that this family has the structure of a Boolean Algebra. We translate our assumption that “each toss does not influence the outcome of the other tosses” by defining a class of probability measures which describes each toss as being *independent* of the others. In formulas, this means that our admissible probability measures  $\mathcal{P} := \{P_p\}$  on  $\{\Omega, \mathcal{A}\}$  are defined, for each elementary event  $\omega \in \Omega$  by

$$P_p(\{\omega\}) = p^{n(T)} (1 - p)^{N - n(T)} \quad , \quad 0 < p < 1 \quad , \quad (1.1.1)$$

where  $n(T)$  is the number of symbols  $T$  in the sequence  $\omega$ . Clearly the probability measure  $P_p$  is defined as soon as one assigns a value to  $p$  in the interval  $(0 < p < 1)$ . In this case the family  $\mathcal{P}$  is *parametric*; i.e.

$$\mathcal{P} := \left\{ P_p ; 0 < p < 1 \right\} .$$

Estimating  $P$  is hence the same thing as selecting a plausible value of  $p$  based on the observation of the outcomes of  $N$  successive coin tosses.

Alternatively, one may want to validate some a priori belief on  $p$  for example that  $p = 1/2$  (that is,  $T$  and  $C$  are equiprobable). In this case one deals with an hypothesis testing problem: on the basis of the observation  $\bar{\omega}$  decide whether  $P_p$  belongs to the class

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

or  $P_p$  belongs to the complementary family

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



As we shall see, estimation and hypothesis testing problems are approached by quite different methodologies.  $\diamond$

### Parametric problems

The family of possible probability measures  $\mathcal{P}$  (or the  $k$  classes  $\mathcal{P}_i, i = 1, \dots, k$ ) constitutes the *a priori information* of the statistical inference problem. Very often the choice of  $\mathcal{P}$  is actually dictated by mathematical convenience.

*Parametric* problems are those where  $\mathcal{P}$  has the form

$$\mathcal{P} = \{P_\theta ; \theta \in \Theta\} \quad , \quad (1.1.2)$$

where  $\Theta$  is a subset of a finite dimensional Euclidean space, say  $\Theta \subseteq \mathbb{R}^p$ .

One then speaks of *estimation of the parameter*  $\theta$  or of *testing hypotheses on the parameter*  $\theta$ . In this last case one may as well formulate the problem as deciding if  $\theta$  belongs to one out of  $k$  disjoint subsets  $(\Theta_i, i = 1, \dots, k)$  of  $\Theta$  such that  $\mathcal{P}_i = \{P_\theta | \theta \in \Theta_i\}, i = 1, \dots, k$ .

The coin tossing problem above is parametric. Here  $\Theta$  is the interval  $(0, 1)$ . The two classes  $\Theta_0 = \{1/2\}, \Theta_1 = (0, 1) - \{1/2\}$  parametrize the two alternative hypotheses.

In this course we shall exclusively deal with probabilities induced by random variables or by families (possibly infinite) of random variables. These random variables will in general be vector valued, say  $\mathbb{R}^m$ -valued (often called random vectors). Random variables (or vectors) will be written as column vectors and always be denoted by boldface letters, such as  $\mathbf{x}, \mathbf{y}$  etc. When  $m = 1$  we shall talk about *scalar* random variables. The abbreviation r.v. will sometimes be used.

Let  $\mathbf{y} = [y_1 \cdots y_m]^\top$  be an  $m$ -dimensional random variable defined on the space  $\{\Omega, \mathcal{A}\}$  that is, a measurable function from  $\Omega$  into  $\mathbb{R}^m$ . The **sample space** of  $\mathbf{y}$  is just the space of possible values of  $\mathbf{y}$ , that is  $\mathbb{R}^m$ , together with its Borel  $\sigma$ -algebra  $\mathcal{B}^m$  (the smallest  $\sigma$ -algebra of subsets of  $\mathbb{R}^m$  containing all open intervals). If  $P$  is any probability measure defined on  $\mathcal{A}$ , the *probability induced by  $\mathbf{y}$* ,  $P_{\mathbf{y}}$ , on its sample space  $\{\mathbb{R}^m, \mathcal{B}^m\}$  is defined by the position

$$P_{\mathbf{y}}(E) := P\{\omega | \mathbf{y}(\omega) \in E\}; \quad E \in \mathcal{B}^m \quad (1.1.3)$$

which can be written more economically as,

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

where  $\mathbf{y}^{-1}(E)$  denotes the inverse image of the event  $E \in \mathcal{B}^m$ .

It can be proven that  $P_{\mathbf{y}}$  is uniquely determined by assigning the probability of semi-infinite intervals of  $\mathbb{R}^m$  of the form  $\{y \in \mathbb{R}^m | y_1 \leq \eta_1, \dots, y_m \leq \eta_m\}$  which we shall write symbolically as  $\{y \leq \eta\}$ . That this is so follows from the fact that all sets in the Borel  $\sigma$ -algebra  $\mathcal{B}^m$  are limits of sequences of sets obtained by Boolean operations on such intervals [?].

Hence  $P_{\mathbf{y}}$  is uniquely determined by its *Probability Distribution Function (PDF)*  $F : \mathbb{R}^m \rightarrow [0, 1]$

$$F(\eta) = F(\eta_1, \dots, \eta_m) = P\{\omega | \mathbf{y}_1(\omega) \leq \eta_1, \dots, \mathbf{y}_m(\omega) \leq \eta_m\} \quad (1.1.4)$$

4 CHAPTER 1. STATISTICAL INFERENCE AND PARAMETER ESTIMATION

In fact, consider the probability space  $\{\mathbb{R}^m, \mathcal{B}^m, P_y\}$  and define on it the random variable

$$\tilde{y} : \mathbb{R}^m \rightarrow \mathbb{R}^m, \quad \tilde{y}_i(\eta_1, \dots, \eta_m) := \eta_i, \quad i = 1, \dots, m, \quad (1.1.5)$$

(the identity function) it is easy to check that the PDF of  $\tilde{y}$  namely

$$P_{\tilde{y}}\{\eta \mid \tilde{y}_1(\eta) \leq \eta_1, \dots, \tilde{y}_m(\eta) \leq \eta_m\}$$

is exactly the same as the original PDF,  $F(\eta)$ , of  $y$  defined in (1.1.4). It follows that  $y$  and  $\tilde{y}$  are indistinguishable as the probability of any event  $E \in \mathcal{B}^m$  is computed by integrating the PDF's of  $y$  and  $\tilde{y}$  over  $E$  and hence must coincide. It follows that  $\tilde{y}$  and  $y$  can be regarded as *the same random variable*.

The "canonical" representation (1.1.5) of a random variable is called *representation on its sample space*. It is very handy since it permits to identify  $y$  just by assigning its PDF. This is actually well-known, as one commonly speaks about a "Gaussian random variable" of mean  $\mu$  and variance  $\sigma^2$ , implicitly meaning that the random variable is the identity function on  $\mathbb{R}$

$$y : \mathbb{R} \rightarrow \mathbb{R}, \quad y(y) := y, \quad \forall y \in \mathbb{R},$$

defined on the probability space  $\{\mathbb{R}, \mathcal{B}, P_y\}$  with  $P_y$  defined by

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

for every  $E \in \mathcal{B}$ .

Note that the sample space representation of a random variable is "sewn up" about  $y$  and every random variable defined on the sample space of  $y$ , being a function of the independent variable  $y$  is necessarily a function of  $y$ . Note however that on the sample space of  $y$  there do not exist random variables independent of  $y$ . In the following we shall normally assume that all random variables under study are defined on their sample space. Hence we shall, from now on, only consider inference problems where  $\mathcal{P}$  (or  $\{\mathcal{P}_i\}$ ) is a family of probability measures on  $\{\mathbb{R}^m, \mathcal{B}^m\}$  so that every member  $P \in \mathcal{P}$  ( $P_k$ ) is uniquely defined by a PDF,  $F$  on  $\mathbb{R}^m$ . It will henceforth be equivalent to describe  $\mathcal{P}$  as a family of PDF's, namely  $\mathcal{P} := \{F(\cdot)\}$ .

A *parametric family* of PDF's is therefore

$$\mathcal{P} = \{F_\theta \mid \theta \in \Theta\}, \quad \Theta \subset \mathbb{R}^p$$

where the "functional form" (i.e. the analytic dependence on the independent variable  $y$ ) of each  $F_\theta$  is a priori known and in order to individuate  $F$  in  $\mathcal{P}$  it should be enough to assign the value of a  $p$ -dimensional parameter  $\theta$ . The set  $\Theta$  is the set of *admissible values* of the parameter.

The underlying conceptual scheme is that the experimental data  $y = \{y_1, \dots, y_m\}$  come from  $m$  measurement devices which are modeled as the (in general correlated) components  $y_1, \dots, y_m$  of an  $m$ -dimensional random variable  $y$  having PDF

$F$ . We shall assume to have enough a priori information on the joint PDF  $F$  to choose a parametric family of PDF's to describe the measurement data. Quite often when it is reasonable to assume that the measured quantities are affected by many additive accidental errors, resulting from interactions of the measuring device with the external environment, one may choose the family  $\{F\}$  to be a family of Gaussian  $m$ -dimensional distributions, which is described by the well-known density function of the form

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

In this case the *mean vector*  $\mu \in \mathbb{R}^m$  and the *covariance matrix*  $\Sigma \in \mathbb{R}^{m \times m}$  are the parameters which the family depends on.

### Repeated measurements

In classical statistics one assumes to be able to perform repeated experiments and thereby observe sample values  $y_t$ ,  $t = 1, 2, \dots, y_N$  all coming from experiments governed by the same PDF  $F$ . This scheme can equivalently be described as the observation of a *sequence* of random variables

$$\{\mathbf{y}_1, \dots, \mathbf{y}_N\} ,$$

where each variable  $\mathbf{y}_t$  has the same PDF  $F$ .

Note that the  $\mathbf{y}_t$ 's may in general be correlated. In this respect, a basic question for the experimenter is how he/she should conduct the  $N$  experiments in such a way as to obtain the "maximum information" about the unknown PDF  $F$ . It is clear that in case all measurements were conducted *exactly in the same experimental conditions* the measurement errors would be the same and therefore in the  $N$  experiments one would get  $y_1 = y_2 = \dots = y_N$  and the experimental data obtained in the second, third,..  $N$ -th trial would be completely useless.

For this reason one should try to arrange the sequence of experiments in such a way that the causes of accidental errors should be as different as possible among each other experiment. One should actually keep in mind that the probabilistic model one wants to construct ( $F$ ) should describe precisely the probability distribution induced by these accidental errors.

Assuming that  $F$  has a density  $p(y_1, y_2, \dots, y_N)$ , this problem can be set up mathematically as the maximization of the entropy rate of the joint distribution of the  $N$  random variables  $\mathbf{y}_t$ ;  $t = 1, \dots, N$ , subject to the constraint that all the marginals with respect to  $y_1, y_2, \dots, y_N$  are the same; i.e.

$$\int_{\mathbb{R}^{N-1}} p(y_1, y_2, \dots, y_N) dy_1 dy_2 \dots dy_{k-1} dy_{k+1} \dots dy_N = p(y_k) , \quad k = 1, 2, \dots, N$$

where  $p$  is the fixed density of  $F$ . A result going back to Shannon [?] then states that the optimal  $p$  must be the product

$$p(y_1, y_2, \dots, y_N) = \prod_{k=1}^N p(y_k)$$

6 CHAPTER 1. STATISTICAL INFERENCE AND PARAMETER ESTIMATION

that is, the the  $N$  random experiments should be *independent* (and identically distributed).

**Definition 1.1.** Let  $\mathcal{F}$  be a family of PDF's on  $\mathbb{R}^m$  and let  $\mathbf{y}_1, \dots, \mathbf{y}_N$  be  $m$ -dimensional random variables all having identical PDF  $F \in \mathcal{F}$ , which are mutually independent for any  $F$  in the class  $\mathcal{F}$ . One says that  $\mathbf{y}_1, \dots, \mathbf{y}_N$  are a **random sample** of dimension  $N$  drawn from the class  $\mathcal{F}$ .

In a sense, a random sample provides “maximum information” about the (unknown) distribution function from which it is drawn. In classical statistics, it is very common to assume that the observed data form a random sample. When  $F$  is an element of a parametric family  $\{F_\theta ; \theta \in \Theta\}$  the joint distribution of a random sample can be written for any  $\theta \in \Theta$ , as:

$$F_\theta^N(y_1, \dots, y_N) = F_\theta(y_1), \dots, F_\theta(y_N) \quad , \quad y_t \in \mathbb{R}^m \quad . \quad (1.1.6)$$

Techniques for generating measurements which approximate the ideal situation of a random sample are studied in a branch of statistics called *sampling theory* see e.g. [9].

However the situation of main interest for us is when the data are correlated, that is, the past history at time  $t$ , say  $(y_1, \dots, y_t)$  influences the next sample  $y_{t+1}$ . The main object of interest in this course will in fact be the problem of how to extract from the observed data a mathematical description of this dynamic “influence”. Classical statistics based on random samples will anyway be a foundational background for attacking this more general setup.

**Definition 1.2.** Let  $(\mathbf{y}_1, \dots, \mathbf{y}_N)$  be a sample (not necessarily random) drawn from a PDF  $F$  belonging to a parametric family  $\{F_\theta ; \theta \in \Theta\}$ . A *statistic*, is any (measurable) function  $\phi$ , of  $(\mathbf{y}_1, \dots, \mathbf{y}_N)$ , say

$$\phi : \mathbb{R}^m \times \dots \times \mathbb{R}^m \rightarrow \mathbb{R}^q \quad ,$$

which does not depend on the parameter  $\theta$ .

Being a function of random variables, a statistic is itself a random variable,  $\phi(\mathbf{y}_1, \dots, \mathbf{y}_N)$ , whose PDF can, at least in simple cases, be computed from the joint distribution  $F_\theta^N$  of the sample. Some simple examples will be presented below <sup>3</sup>

The *sample mean*,  $\bar{\mathbf{y}}_N$ ,

$$\bar{\mathbf{y}}_N = \frac{1}{N} \sum_{t=1}^N \mathbf{y}_t \quad ; \quad (1.1.7)$$

is an  $m$ -dimensional statistics. When  $\{\mathbf{y}_t\}$  is a random sample drawn from an unknown “true” PDF  $F_{\theta_0}$  where  $F_{\theta_0} \in \{F_\theta ; \theta \in \Theta\}$ , one has  $\mathbb{E}_{\theta_0} \bar{\mathbf{y}}_N = \mathbb{E}_{\theta_0} \mathbf{y}$ , where  $\mathbb{E}_{\theta_0}$  denotes expectation with respect to  $F_{\theta_0}$ . By the law of large numbers (which will

<sup>3</sup>Often it is of interest to study the behavior of a statistic as a function of  $N$ ; in particular for  $N \rightarrow \infty$ . For this reason a subscript  $N$  is often attached to the symbol, e.g. using a notation like  $\phi_N$ .

be recalled in Chap. 4), the limit

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_1^N \mathbf{y}_t$$

exists with probability one and is equal to  $\mathbb{E}_0 \mathbf{y} = \int_{\mathbb{R}^m} y dF_{\theta_0}(y)$ . In other words, the limit

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_1^N y_t$$

exists for “almost all” possible sample sequences  $\{y_1, y_2, \dots, y_t, \dots\}$  and is actually equal to the mean  $\mathbb{E}_0 \mathbf{y}$  of the true distribution  $F_{\theta_0}$ . This explains the origin of the name.

The *sample variance*,  $\hat{\Sigma}_N^2$ ,

$$\hat{\Sigma}_N^2 := \frac{1}{N} \sum_{t=1}^N (\mathbf{y}_t - \bar{\mathbf{y}}_N) (\mathbf{y}_t - \bar{\mathbf{y}}_N)^\top \tag{1.1.8}$$

is a  $\mathbb{R}_+^{m \times m}$ -valued statistics, in fact a random symmetric positive semidefinite matrix.

For a random sample, this statistics enjoys similar asymptotic properties of  $\bar{\mathbf{y}}_N$ . In fact, if  $\{\mathbf{y}_t\}$  is a random sample drawn from  $F_{\theta_0}$ , the limit

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N (\mathbf{y}_t - \bar{\mathbf{y}}_N) (\mathbf{y}_t - \bar{\mathbf{y}}_N)^\top$$

still exists with probability one, that is for “almost all” possible strings of observations  $\{y_t\}$ , and is equal to  $\mathbb{E}_0(\mathbf{y} - \mathbb{E}_0 \mathbf{y})(\mathbf{y} - \mathbb{E}_0 \mathbf{y})'$  which is the variance matrix of  $\mathbf{y}$  (or of  $F_{\theta_0}$ ).

In the following we shall use the notation  $\mathbf{y} \sim \{F_\theta\}$  to signify that  $\mathbf{y}$  is distributed according to some unknown PDF belonging to the parametric family  $\{F_\theta; \theta \in \Theta\}$ .

The above are just two typical examples of statistics. We just stress that a statistic must be a function of the observed data alone and *cannot depend on the parameter*  $\theta$ .

### Fisher e Bayes

It is common knowledge that there are two main philosophical approaches to statistical inference, the *Classical or Fisherian* and the *Bayesian* approach.

In the Fisherian approach one postulates that the parameter is a deterministic but unknown quantity which by its nature could in principle be determined exactly say, by an infinite series of experiments. This viewpoint can be acceptable when  $\theta$  has an instrumental role in the mathematical description of the experiment, say the mode or the variance of a PDF, or, as we shall often see, the coefficients of

## 8 CHAPTER 1. STATISTICAL INFERENCE AND PARAMETER ESTIMATION

a difference equation describing the dynamics of a random signal. One classically postulates the existence of a “true” value,  $\theta_0$ , of the unknown parameter indexing a “true PDF”,  $F_{\theta_0}$ , which is the PDF according to which the data are truly distributed. Clearly, since any model can only be an approximate description of reality, this postulate may lead to logical contradictions. Nevertheless it formalizes an ideal situation wherein it is possible to assess in a simple way certain basic properties of statistical procedures like unbiasedness, consistency etc. which have an important practical significance.

On the other hand, when the parameter is interpreted as a mathematical variable to describe the possible value taken by physical quantities which are being measured in the experiment, say voltage, mass or length of a physical object, the classical approach may become questionable. Due to the unavoidable interactions with the surrounding environment and due to the limited precision of any measurement device, a physical quantity is never measurable “exactly” and it is in fact doubtful whether it should make any sense at all to assign to it a definite, precise numerical value. Bayesian statistics can be seen as a formalization of this observation. According to the Bayesian viewpoint  $\theta$  should always be regarded as the sample value taken on by some random variable  $\mathbf{x}$ . The statistical model  $\{F_\theta ; \theta \in \Theta\}$  should then be formally converted to a conditional probability distribution

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

which is always possible provided the map  $\theta \rightarrow F_\theta$  is a measurable function of  $\theta$ , which in practice is always the case. What remains open is the question of the probabilistic description of  $\mathbf{x}$ , which is called the *a priori* distribution of the parameter. In some cases this distribution may be known, at least approximately and in this case the Bayesian approach seems to be the natural approach to follow. Bayesian statistics then proceeds, based on “Bayes rule” to formulate statistical inference as a branch of Probability theory. Quite often however the *a priori* distribution of the parameter is not obvious. There is a century long debate about what one should do in this case. The so-called *subjectivistic school* [?] insists that one always has a degree of belief about the possible parameter values and this belief should always be imposed on the problem formulation. We shall not dwell into these ramifications of the Bayesian philosophy.

The Bayesian approach requires the computation of the *a posteriori* probability of the random parameter  $\mathbf{x}$  which is just the conditional probability distribution of  $\mathbf{x}$  given the observations. In the past, this calculation has been a major stumbling block for the practical application of the Bayesian philosophy since the explicit calculation of the posterior distribution can be done only for a very limited class of priors (the so called *conjugate prior distributions*). Nowadays, ultrafast computers and efficient optimization algorithms permit to apply the Bayesian philosophy to a wide class of problems, without worrying about the explicit calculation of the posterior distribution. This has led to an explosion of papers on Bayesian techniques in statistics and some of these new techniques are finding application also in system identification.

In this book we shall mostly deal with inference problems regarding prob-

abilistic models which are to be selected from parametric classes which are often imposed on the data on the basis of mathematical simplicity or convenience. These parameters very seldom have a physical interpretation and very seldom will we have a priori informations about the parameter distribution of these models. For these reasons, in this course we shall normally follow the classical approach.

## 1.2 Classical Theory of Parameter estimation

Consider a sample  $(\mathbf{y}_1, \dots, \mathbf{y}_N)$  drawn from an element of the parametric family  $\{F_\theta; \theta \in \Theta\}$ ,  $\Theta \subseteq \mathbb{R}^p$ .

**Definition 1.3.** An estimator of the parameter  $\theta$  is any statistic  $\phi$  with values in  $\Theta$ . The value taken on by the random variable  $\phi(\mathbf{y}_1, \dots, \mathbf{y}_N)$  corresponding to the sample values  $(y_1, \dots, y_N)$  of  $\mathbf{y}_1, \dots, \mathbf{y}_N$ ,

$$\hat{\theta} = \phi(y_1, \dots, y_N) \quad , \quad (1.2.1)$$

is the **estimate** of  $\theta$ , based on the data  $(y_1, \dots, y_N)$ .

One would of course like that the estimates based on the observed data obtained in an hypothetical series of many measurement experiments, should be "close" to the true parameter value,  $\theta_0$ . One would in particular like that the average estimate corresponding to a large set of experimental measurements, say,  $(y'_1, \dots, y'_N), (y''_1, \dots, y''_N), \dots$ , should be equal to  $\theta_0$ . This condition can be expressed as

$$\mathbb{E}_{\theta_0} \phi(\mathbf{y}_1, \dots, \mathbf{y}_N) = \theta_0 \quad , \quad (1.2.2)$$

where  $\mathbb{E}_{\theta_0}$  is the expectation operator with respect to the true PDF of the observations,  $F_{\theta_0}^N$ . However, since  $\theta_0$  is unknown, this condition cannot be verified. A (quite restrictive) way out is to require that (1.2.2) should hold for all possible values of the parameter, which leads to the following notion,

**Definition 1.4.** An estimator  $\phi$  is said to be **(uniformly) unbiased** if

$$\mathbb{E}_\theta \phi(\mathbf{y}_1, \dots, \mathbf{y}_N) = \theta \quad , \quad \forall \theta \in \Theta \quad . \quad (1.2.3)$$

Another rather natural request is that a good estimator should provide estimates  $\phi(y_1, \dots, y_N)$  which are tightly clustered about their average value. In other words  $\phi$  should have a small variance. Naturally for this condition to make sense one should a priori restrict the class of admissible estimators. For a constant (deterministic) estimator, not depending on the data, would trivially have zero variance but would surely be a useless estimator. The notion introduced in the definition below depends on the specification of a class  $\mathcal{C}$  of admissible estimators.

**Definition 1.5.** The estimator  $\phi$  has **(uniformly) minimum variance** in the class  $\mathcal{C}$  if the variance

$$\text{var}_\theta(\phi) := \mathbb{E}_\theta(\phi - E_\theta \phi)^\top (\phi - E_\theta \phi) \quad (1.2.4)$$

10 CHAPTER 1. STATISTICAL INFERENCE AND PARAMETER ESTIMATION

is the smallest among all estimators belonging to the class  $\mathcal{C}$ , that is

$$\text{var}_\theta(\phi) \leq \text{var}_\theta(\psi) \quad , \quad \forall \psi \in \mathcal{C} \quad , \quad (1.2.5)$$

for all  $\theta \in \Theta$ .

It is obvious that  $\mathcal{C}$  cannot be the class of all functions of the data since a constant deterministic function will have zero variance but obviously be a totally useless estimator.

A natural class of admissible estimators to consider is the class of all unbiased estimators. Note however that for certain classes of parametric PDF's unbiased estimators may not exist.

*Example :* Let  $x$  be a random variable subject to the geometric distribution with parameter of success  $\theta$ , that is, for any natural number  $k$ ,

If  $\phi$  is an unbiased estimator of the parameter  $\theta$ , it must satisfy the unbiasedness equation, that is,

The unique solution of this equation is

Evidently,  $\phi$  is good only when  $\theta$  is very close to 1 or 0, otherwise carries no useful information on  $\theta$ .

Below we will see that if  $\mathcal{C}$  is taken to be the class of all unbiased estimators of  $\theta$ , no such degeneracy is possible. This follows from a celebrated inequality, called the *Cramèr-Rao inequality*.

### 1.2.1 The Cramèr-Rao Inequality

Let  $\mathbf{x}$  be a  $r$ -dimensional random vector with  $\mathbf{x} \sim \{F_\theta ; \theta \in \Theta\}$  ( $\mathbf{x}$  could in particular be a random sample as  $(y_1, \dots, y_N)$ , but the Cramèr-Rao inequality does not require independence of the components of  $\mathbf{x}$ ). We shall assume that the following properties hold;

A.1)  $F_\theta$  admits a density  $p(\cdot, \theta)$  which is twice differentiable with respect to  $\theta$ .

A.2) For every statistics  $\phi$  with  $\mathbb{E}_\theta \phi < \infty$ ,

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

for  $i = 1, \dots, p$  and for every  $\theta \in \Theta$ . In particular,

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

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

for all  $i, j = 1, \dots, p$  and for every  $\theta \in \Theta$ .



1.2. Classical Theory of Parameter estimation

**Definition 1.6.** The Fisher Information Matrix  $I(\theta)$ , of the parametric family of densities  $\{p_\theta\}$  is defined as

$$I(\theta) := \left[ \mathbb{E}_\theta \left( \frac{\partial \log p(\mathbf{x}, \theta)}{\partial \theta_i} \cdot \frac{\partial \log p(\mathbf{x}, \theta)}{\partial \theta_j} \right) \right]_{i,j=1,\dots,p} \quad (1.2.6)$$

which can also be written as

$$I(\theta) = \left[ -E_\theta \frac{\partial^2 \log p(\mathbf{x}, \theta)}{\partial \theta_i \partial \theta_j} \right]_{i,j=1,\dots,p} . \quad (1.2.7)$$

That (1.2.7) and (1.2.6) are equivalent follows by differentiating the identity  $\int p(x, \theta) dx = 1$  (constant with respect to  $\theta$ ) termwise with respect to  $\theta$  getting

$$\int_{\mathbb{R}^r} \frac{\partial p(x, \theta)}{\partial \theta_i} dx = 0 \quad , \quad i = 1, \dots, p \quad , \quad (1.2.8)$$

$$\int_{\mathbb{R}^r} \frac{\partial^2 p(x, \theta)}{\partial \theta_i \partial \theta_j} dx = 0 \quad , \quad i, j = 1, \dots, p \quad . \quad (1.2.9)$$

Equation (1.2.7) then follows from

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

in force of (1.2.9).

In order to understand the meaning of  $I(\theta)$  we shall bring in the  $p$ -dimensional random vector of the sensitivities of  $p(x, \theta)$  with respect to the parameter  $\theta$ ,

$$\mathbf{z}_\theta := \left[ \frac{\partial \log p(\mathbf{x}, \theta)}{\partial \theta_i} \right]_{i=1,\dots,p} \quad (1.2.10)$$

and note that

$$I(\theta) = E_\theta \mathbf{z}_\theta \mathbf{z}_\theta^\top \geq 0 \quad . \quad (1.2.11)$$

where  $\geq 0$  means that the matrix on the left is positive semidefinite. From (1.2.8) it easily follows that  $\mathbb{E}_\theta \frac{\partial \log p}{\partial \theta_i} = 0$  for all  $i$ 's and so

$$E_\theta \mathbf{z}_\theta = 0 \quad (1.2.12)$$

which implies that  $I(\theta)$  is actually the variance of the sensitivity  $\mathbf{z}_\theta$ .

**Theorem 1.1 (The Cramèr-Rao Inequality).** Let  $g$  be a differentiable function from  $\Theta$  to  $\mathbb{R}^q$  and  $\phi$  be an unbiased estimator of  $g(\theta)$ . Let  $V(\theta)$  be the variance matrix of  $\phi$  and  $G(\theta)$  the Jacobian matrix of  $g$ ,

$$G(\theta) = \left[ \frac{\partial g_i(\theta)}{\partial \theta_j} \right]_{\substack{i=1,\dots,q \\ j=1,\dots,p}} . \quad (1.2.13)$$

12 CHAPTER 1. STATISTICAL INFERENCE AND PARAMETER ESTIMATION

Then, if the Fisher matrix  $I(\theta)$  is invertible, one has

$$V(\theta) - G(\theta) I^{-1}(\theta) G'(\theta) \geq 0 \quad , \quad (1.2.14)$$

where  $\geq 0$  means that the matrix on the left is positive semidefinite.

**Proof.** The proof is based on the classical formula for the error variance of the linear Bayesian estimator  $\hat{\phi}(\mathbf{x}) := \mathbb{E}_\theta [\phi(\mathbf{x}) \mid \mathbf{z}_\theta]$  of the vector  $\phi(\mathbf{x})$ , given  $\mathbf{z}_\theta$ , that is

$$\text{Var}_\theta \{ \phi(\mathbf{x}) - \hat{\phi}(\mathbf{x}) \} = \text{Var}_\theta \{ \phi(\mathbf{x}) \} - \text{Cov}_\theta \{ \phi(\mathbf{x}), \mathbf{z}_\theta \} \text{Var}_\theta \{ \mathbf{z}_\theta \}^{-1} \text{Cov}_\theta \{ \phi(\mathbf{x}), \mathbf{z}_\theta \}' \quad . \quad (1.2.15)$$

See for example (2.4.31) or [45, p. 27].

Since  $\phi(\mathbf{x})$  is an unbiased estimator of  $g(\theta)$ ; i.e.

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

by applying property A.3) one gets

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

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

and hence  $\frac{\partial g(\theta)}{\partial \theta_j}$  is the  $j$ -th column of the covariance matrix of  $\phi$  and  $\mathbf{z}_\theta$ ,

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

that is,

$$\mathbb{E}_\theta \phi \mathbf{z}_\theta' = G(\theta) \quad . \quad (1.2.16)$$

The inequality follows since the variance of the random vector  $\phi(\mathbf{x}) - G(\theta) I(\theta)^{-1} \mathbf{z}_\theta$  must be (at least) positive semidefinite.  $\square$

**Remarks**

When  $\phi$  is an unbiased estimator of  $\theta$  (that is if  $g$  is the identity map) one has  $G(\theta) = I$  ( $p \times p$ ) and (1.2.14) becomes

$$V(\theta) - I(\theta)^{-1} \geq 0 \quad . \quad (1.2.17)$$

Since the scalar variance  $\text{var}_\theta(\phi) = \sum_1^p E_\theta(\phi_i - \theta_i)^2$  is the trace of  $V(\theta)$  and

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

(the trace is the sum of the eigenvalues and the eigenvalues of a positive semidefinite matrix are all non-negative) it follows that *the scalar variance of any unbiased estimator of the parameter  $\theta$  cannot be less than the positive number  $\text{Tr} I(\theta)^{-1}$ ,*

$$\text{var}_\theta(\phi) \geq \text{Tr} [I(\theta)^{-1}] \quad , \quad \forall \theta \quad . \quad (1.2.18)$$

1.2. Classical Theory of Parameter estimation

This lower bound only depends on the probabilistic model class  $\{p(\cdot, \theta); \theta \in \Theta\}$  and is *independent of which estimation criterion is used to construct  $\phi$* .

One should however be aware of the fact that the Cramèr-Rao bound is just *one* possible bound for the variance which is not necessarily the tightest possible bound. There are in fact unbiased estimators whose variance is strictly larger than  $\text{Tr} [I(\theta)^{-1}]$  but nevertheless have minimum variance.

**Example 1.1.** Let  $y \sim \mathcal{N}(\theta, \sigma^2)$  be a scalar random variable with a known variance  $\sigma^2$ . Since

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

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

we have

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

Hence the variance of any unbiased estimator of  $\theta$  based on a sample of size one, cannot be smaller than the variance of  $y$ . Assume now we have a random sample of size  $N$  from the same Gaussian distribution. Now we have a random vector  $\mathbf{x} = (y_1, \dots, y_N)$  of dimension  $r = N$  and

$$p(y_1, \dots, y_N, \theta) = \prod_{t=1}^N p(y_t, \theta)$$

and hence

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

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

Since the random variables  $y_1, \dots, y_N$  are independent, it follows that,

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

Let us consider the sample mean

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

which has distribution  $\mathcal{N}(\theta, \sigma^2/N)$ . Since  $\bar{y}_N$  is an unbiased estimator of  $\theta$  with variance  $\sigma^2/N$ , exactly equal to the inverse of the Fisher information, we conclude

14 CHAPTER 1. STATISTICAL INFERENCE AND PARAMETER ESTIMATION

that the sample mean is the best possible estimator of  $\theta$  (of course if the sample distribution is Gaussian). One says that an unbiased estimator whose variance is exactly equal to the inverse of the Fisher information matrix,  $V(\theta) = I(\theta)^{-1}$  is *efficient*.

◇

**Example 1.2.** Let  $\mathbf{y} \sim \mathcal{N}(\mu, \theta^2)$ , where  $\mu$  is known and  $(\mathbf{y}_1, \dots, \mathbf{y}_N)$  is a random sample from  $\mathcal{N}(\mu, \theta^2)$ . Consider the estimator

$$s_N^2 = \frac{N\hat{\sigma}_N^2}{N-1} = \frac{1}{N-1} \sum_{t=1}^N (\mathbf{y}_t - \bar{\mathbf{y}})^2 \quad ;$$

In the next chapter we shall see that  $\frac{Ns_N^2}{\theta^2}$  has a *chi squared* distribution with  $N - 1$  degrees of freedom, which has expectation  $N - 1$  and variance  $2(N - 1)$ . It follows that  $\bar{s}^2$  is an unbiased estimator of  $\theta^2$ , of variance  $\frac{2\theta^4}{N-1}$ . The Cramèr-Rao bound in this case is  $2\theta^4/N$  and hence the variance of  $s_N^2$  is strictly larger than  $I(\theta)^{-1}$ . One can however show [?] that an unbiased estimator of  $\theta^2$  cannot have a smaller variance than that of  $\bar{s}_N^2$ . From this example it follows that  $I(\theta)^{-1}$  is not the best possible lower bound.

◇

**Esercizi**

- 1-1 Let  $I(\theta)$  be the Fisher matrix relative to an arbitrary density  $p(\mathbf{y}, \theta)$ . Show that for a random sample of size  $N$  one has  $I_N(\theta) = N I(\theta)$ .
- 1-2 Show, without using the  $\chi^2$  distribution, that the Cramèr-Rao bound for a random sample from  $\mathcal{N}(\mu, \theta^2)$  of size  $N$  is  $2\theta^4/N$ .
- 1-3 Show that the Cramèr-Rao bound for  $\mathcal{N}(\theta_1, \theta_2^2)$  is

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

**Interpretation of  $I(\theta)$**

In this section we shall define a measure of deviation of two random variables  $\mathbf{x}_1 \sim p(\cdot, \theta_1)$  and  $\mathbf{x}_2 \sim p(\cdot, \theta_2)$  described by the same parametric family of distributions. We shall use this measure to quantify in rather precise terms, the ability of observations extracted from the model, to *discriminate* between different values of the parameter  $\theta$ .

**Definition 1.7.** Let  $f$  and  $p$  be probability densities such that  $p(x) = 0 \Rightarrow f(x) = 0$ . The Kullback-Leibler pseudo-distance between  $f$  and  $p$ , is

$$K(f, p) := \int_{\mathbb{R}^r} [\log f - \log p] f(x) dx = \int_{\mathbb{R}^r} \log f/p f(x) dx = E_f \log f/p; \quad (1.2.19)$$

1.2. Classical Theory of Parameter estimation

It is immediate that  $K(f, p) = 0$  if and only if  $f = p$ . From Jensen inequality:

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

which holds for  $g(x) > 0$  and an arbitrary positive measure  $\mu$ , one gets

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

so that  $K(f, p) \geq 0$ .

For this reason  $K(f, p)$  can be interpreted as a measure of deviation of the probability density  $p$  from a "reference" density  $f$ . Note in fact that  $K(f, p)$  is not symmetric; i.e.  $K(p, f) \neq K(f, p)$  and does not satisfy the triangle inequality. In Information Theory  $K(f, p)$  is called *divergence* and is denoted by the symbol  $D(f||p)$  (here  $p$  is the approximation of  $f$ ). The article in Wikipedia on *Kullback-Leibler divergence* provides a rather complete overview and a bibliography.

Let us assume that the family  $p(\cdot, \theta)$  satisfies the same regularity assumptions listed in Section 1.2.1 and let  $f \equiv p(\cdot, \theta_0)$  and  $p \equiv p(\cdot, \theta)$ ,  $\theta_0, \theta \in \Theta$ . Denoting  $K(p(\cdot, \theta_0), p(\cdot, \theta))$  by  $K(\theta_0, \theta)$  and letting  $\theta = \theta_0 + \Delta\theta$ , one has

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

Since  $K(\theta_0, \theta_0) = 0$  and

$$\left. \frac{\partial K}{\partial \theta_i} \right|_{\theta_0} = - \int_{\mathbb{R}^r} p(x, \theta_0) \frac{\partial \log p(x, \theta)}{\partial \theta_i} dx \quad ,$$

it follows that

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

for all  $i = 1, \dots, p$ .

In the same way one can verify that

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

and hence the first member of this equality is the  $(i, j)$ -th element of the Fisher matrix  $I(\theta_0)$ . Hence, for small variation of the parameter  $\theta$ , it holds

$$K(\theta_0, \theta) \cong \frac{1}{2} \Delta\theta^\top I(\theta_0) \Delta\theta \quad ; \tag{1.2.20}$$

which says that, for small deviations  $\Delta\theta$  of the parameter from the reference value  $\theta_0$ , the Kullback-Leibler distance between  $p(\cdot, \theta)$  and  $p(\cdot, \theta_0)$  is a quadratic form whose weighting matrix is the Fisher matrix  $I(\theta_0)$ . In the next section we will see a remarkable consequence of this fact.

### 1.2.2 Identifiability

There are situations in which the observations are structurally incapable of providing enough information to uniquely locate the value of the parameter  $\theta$  which has generated them. A rather trivial example could be the following. Let  $\theta$  be a two-dimensional parameter  $[\theta_1, \theta_2]^\top$ , ranging on  $\Theta = \mathbb{R}^2$  and let  $F_\theta$  depend on  $(\theta_1, \theta_2)$  only through their product  $\theta_1\theta_2$ ; for example  $F_\theta \sim \mathcal{N}(\theta_1\theta_2, \sigma^2)$ . It is evident that, for any fixed value  $\bar{\theta} = (\bar{\theta}_1, \bar{\theta}_2)^\top$ , the parameters  $\hat{\theta} = (\alpha\bar{\theta}_1, \frac{1}{\alpha}\bar{\theta}_2)^\top$ ,  $\alpha \neq 0$ , define the same PDF; that is  $F_{\hat{\theta}}(x) = F_{\bar{\theta}}(x)$ ,  $\forall x$ . Hence a sample observation extracted from this family, irrespective of its size  $N$ , will never be able to distinguish between  $\bar{\theta}$  and  $\hat{\theta}$ . In this section we shall study this phenomenon in some detail.

**Definition 1.8.** Two parameter values  $\theta'$  and  $\theta''$  in  $\Theta$  are said to be indistinguishable if  $F_{\theta'}(x) = F_{\theta''}(x)$ ,  $\forall x \in \mathbb{R}^r$ . Notation:  $\theta' \simeq \theta''$ .

Evidently  $\simeq$  is an equivalence relation in  $\Theta$ ; in fact it is symmetric, reflexive and transitive. Hence it induces a partition of  $\Theta$  in equivalence classes  $[\theta] := \{\theta' \mid \theta' \simeq \theta\}$  such that  $F_{\theta'} = F_{\theta''}$  if and only if  $\theta'$  and  $\theta''$  belong to the same class  $[\theta]$ . The parameters in the same class are said to be *indistinguishable*.

**Definition 1.9.** The family of PDF's  $\{F_\theta; \theta \in \Theta\}$  (sometimes one says improperly that the parameter  $\theta \in \Theta$ ) is globally identifiable if  $\theta' \simeq \theta''$ , or, equivalently,  $F_{\theta'} = F_{\theta''}$ , implies that  $\theta' = \theta''$  for all  $\theta', \theta''$  in  $\Theta$ .

Hence a family of PDF's  $\{F_\theta; \theta \in \Theta\}$  (or the parameter  $\theta$ ), is globally identifiable if and only if the equivalence classes under indistinguishability reduce to singletons in  $\Theta$ .

For many applications global identifiability is too restrictive. A weaker condition is the following local notion.

**Definition 1.10.** The family of PDF's  $\{F_\theta; \theta \in \Theta\}$  is locally identifiable about  $\theta_0$  if there exists an open neighborhood of  $\theta_0$  which does not contain parameter values  $\theta$  which are indistinguishable from  $\theta_0$  (of course, except  $\theta_0$  itself).

In classical parametric statistics the role of this concept is often overlooked. Identifiability is however an important structural condition of parametric models, especially in modern applications to Identification of dynamic models in Engineering and Econometrics, where the models have often a rather complex parametric structure. Identifiability of linear multi-input multi-output linear systems and the search for identifiable parametrizations thereof has been a major research issue in the past [?, ?, ?, 44]. Identifiability of nonlinear models is still a very active area of research [?].

There is a remarkable relation between (local) identifiability and nonsingularity of the Fisher matrix. This relation is the content of the following Theorem.

**Theorem 1.2 (Rothenberg).** *Let the parametric model  $\{p_\theta; \theta \in \Theta\}$  satisfy the assumptions A.1, A.2, A.3 of Section 1.2.1. Then  $\theta_0$  is locally identifiable if and only if  $I(\theta_0)$  is non-singular.*

*Proof.* [Sketch] The proof is based on the properties of the Kullback-Leibler (pseudo)-metrics which guarantees that  $K(\theta_0, \theta) = 0 \Leftrightarrow p(\cdot, \theta_0) = p(\cdot, \theta)$ . For small deviations  $\Delta\theta$  of the parameter  $\theta$  about the reference value  $\theta_0$ , the Kullback-Leibler distance between the two densities  $p(\cdot, \theta)$  and  $p(\cdot, \theta_0)$  is the quadratic form  $\frac{1}{2} \Delta\theta^\top I(\theta_0) \Delta\theta$ . It follows that in any small enough neighborhood of  $\theta_0$  one can have parameter values  $\theta \neq \theta_0$  for which  $p(\cdot, \theta) = p(\cdot, \theta_0)$  if and only if  $I(\theta_0)$  is singular.  $\square$

In the previous trivial example one has

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

one sees that  $\det I(\theta) = 0, \forall \theta \in \mathbb{R}^2$  and hence the model is never locally identifiable about an arbitrary parameter value  $\theta$ . In fact, the model is globally unidentifiable as all indistinguishability classes contain infinitely many parameter values.

Very often the parametric models used to describe the observations only model the so-called *second order statistics*; that is the mean and variance of the underlying distribution. This is indeed a very common situation in dynamic problems where the observed sample is often a correlated time-series. In this case it is quite common to assume Gaussian distributions even if there is not much evidence for Gaussianity anyway. This assumption can often be dispensed with as it is well-known that the mean and variance identify a Gaussian distribution uniquely. Many concepts in statistics have a *wide-sense* or *second-order* version which does not involve probability distributions but *second order models* consisting of a parametric description of mean and variance. In this sense one can define concepts of *second-order identifiability* either global or local, just referring to the second order statistics instead of the complete probability distribution.

**Esercizi**

**1-4** Show that the parametric model  $F_\theta \sim \mathcal{N}(\theta_1 + \theta_2, \sigma^2)$  is not locally identifiable about any point of  $\mathbb{R}^2$ . In fact this model is globally unidentifiable. Describe the equivalence classes under indistinguishability.

### 1.3 Maximum Likelihood

Let  $\mathbf{x}$  be a random vector taking values in  $\mathbb{R}^r$  (not necessarily a random sample) distributed according to a parametric family of densities  $\{p(\cdot, \theta); \theta \in \Theta\}$  and let  $x_0$  be an observed value of  $\mathbf{x}$ .

**Definition 1.11.** *The Likelihood function of the observation  $x_0$  is the function  $L(x_0, \cdot)$ :*

18 CHAPTER 1. STATISTICAL INFERENCE AND PARAMETER ESTIMATION

$\Theta \rightarrow \mathbb{R}_+$  (the nonnegative reals) defined by

$$L(x_0, \theta) := p(x_0, \theta) \quad . \quad (1.3.1)$$

The “Maximum Likelihood principle”, introduced by Gauss in 1856 [19] and successively popularized by R.A. Fisher, suggests to assume as estimate of  $\theta$ , corresponding to the observation  $x_0$ , the parameter value  $\hat{\theta} \in \Theta$  which maximizes  $L(x_0, \cdot)$

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

implicitly assuming that a maximum exists. The parameter value  $\hat{\theta}$  renders “a posteriori” the observation  $x_0$  the most probable sample according to the family  $\{p(\cdot, \theta) ; \theta \in \Theta\}$ .

Imagine to run many hypothetical experiments each generating a different sample value  $x_0$ . By following the Maximum Likelihood principle one would generate a corresponding family of maximizers  $\hat{\theta}$  each depending on the particular observation. Hence  $\hat{\theta}$  can be also understood as a map  $x_0 \mapsto \hat{\theta}$  from the sample space of the experiment to the parameter space. This map is called the *maximum Likelihood (M.L.) estimator of the parameter  $\theta$* . This estimator,  $\hat{\theta}(\mathbf{x})$ , is a function of the sample and hence is itself a random variable which can in principle be computed by maximizing  $L(\mathbf{x}, \cdot)$  with respect to  $\theta$  (assuming of course that a maximum exists  $\forall x_0 \in \mathbb{R}^r$ ) that is

$$L(\mathbf{x}, \hat{\theta}(\mathbf{x})) = \max_{\theta \in \Theta} p(\mathbf{x}, \theta) \quad . \quad (1.3.2)$$

considering  $\mathbf{x}$  as a free parameter.

To carry on the calculations it is often convenient to maximize the logarithm of  $L(x, \cdot)$  (since  $\log$  is a monotone function of  $L$  it is maximized for the same values of  $\theta$ ). The resulting function of  $\theta$

$$\ell(\mathbf{x}, \cdot) = \log L(\mathbf{x}, \cdot) \quad (1.3.3)$$

is called the *log-likelihood function*. Sometimes, when  $p(\mathbf{x}, \cdot)$  is differentiable with respect to  $\theta$ ,  $\hat{\theta}(\mathbf{x})$  can be computed explicitly by solving a system of  $p$  equations

$$\frac{\partial \ell}{\partial \theta_k}(\mathbf{x}, \theta) = 0 \quad , \quad k = 1, \dots, p \quad , \quad (1.3.4)$$

and then checking which solutions correspond to a maximum of  $\ell(\mathbf{x}, \cdot)$ . In general however one can only solve (1.3.4) numerically and be content with finding a single estimate  $\hat{\theta}$ , given  $x_0$ .

**Example 1.3.** Let  $\mathbf{x} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$  be a random sample of size  $N$  of scalar random variables extracted from the Gaussian distribution  $\mathcal{N}(\theta_1, \theta_2^2)$ . The log-likelihood



function corresponding to the observed sample  $x = (y_1, \dots, y_N)$  is

$$\begin{aligned} \ell(x, \theta) &= \log \left\{ \prod_{i=1}^N \frac{1}{\sqrt{2\pi\theta_2^2}} \exp -\frac{1}{2} \frac{(y_i - \theta_1)^2}{\theta_2^2} \right\} \\ &= -\frac{N}{2} \log 2\pi - \frac{N}{2} \log \theta_2^2 - \frac{1}{2} \sum_1^N \frac{(y_i - \theta_1)^2}{\theta_2^2}. \end{aligned}$$

The equations (1.3.4) become

$$\begin{aligned} \frac{\partial \ell}{\partial \theta_1} &= \frac{1}{\theta_2^2} \left( \sum_1^N y_i - N\theta_1 \right) = 0 \quad , \\ \frac{\partial \ell}{\partial \theta_2^2} &= -\frac{N}{2\theta_2^2} + \frac{1}{2\theta_2^4} \sum_1^N (y_i - \theta_1)^2 = 0 \end{aligned}$$

the first of which is an equation depending only on  $\theta_1$  which yields

$$\hat{\theta}_1 = \frac{1}{N} \sum_1^N y_i = \bar{y}_N \quad . \tag{1.3.5}$$

and substituting this expression in the second equations we easily find

$$\hat{\theta}_2^2 = \frac{1}{N} \sum_1^N (y_i - \bar{y}_N)^2 = \hat{\sigma}_N^2 \quad . \tag{1.3.6}$$

that is, the maximum likelihood estimator of  $\theta_2^2$  is the sample variance. It is immediate to check that the expressions (1.3.5) and (1.3.6) provide an absolute maximum of  $\ell(x, \cdot)$ . Summarizing:

**Proposition 1.1.** *The M.L. estimators of the mean and variance parameters of the Gaussian distribution  $\mathcal{N}(\theta_1, \theta_2^2)$  based on a random sample  $(\mathbf{y}_1, \dots, \mathbf{y}_N)$  are the sample mean and the sample variance.  $\diamond$*

The result holds unchanged in the multivariable case. If  $\mathbf{y} \sim \mathcal{N}(\mu, \Sigma)$ , where  $\mu \in \mathbb{R}^p$  e  $\Sigma \in \mathbb{R}^{p \times p}$  are the unknown mean and variance of  $\mathbf{y}$ , then  $\ell(\mathbf{x}, \mu, \Sigma)$  is maximized by the estimator  $\phi = [\bar{\mathbf{y}}_N, \hat{\Sigma}_N^2]$  where  $\bar{\mathbf{y}}_N$  and  $\hat{\Sigma}_N^2$  are the sample mean and the (matrix-valued) sample variance. See e.g. [62, p. 201].

### Properties of ML estimators

A ML estimator is not necessarily unbiased. For example the ML estimator of  $\theta_2^2$  in the previous example 1.3 is not. This is a consequence of the formula

$$N\hat{\sigma}_N^2(\mathbf{y}) = \sum_1^N (y_i - \theta_1)^2 - N(\bar{\mathbf{y}}_N - \theta_1)^2 \quad . \tag{1.3.7}$$

20 CHAPTER 1. STATISTICAL INFERENCE AND PARAMETER ESTIMATION

which follows from the identity

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

where the term  $\sum_1^N (y_i - \bar{y}_N)$  (sum of the deviations of the sample values from the sample mean), must clearly be zero.

Computing the expectation  $\mathbb{E}_\theta$  of both members in (1.3.7) and recalling that  $\bar{y}_N \sim \mathcal{N}(\theta_1, \theta_2^2/N)$  one finds

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

and hence

$$\mathbb{E}_\theta \hat{\sigma}_N^2 = \theta_2^2 \frac{N - 1}{N} \tag{1.3.8}$$

which shows that  $\hat{\sigma}_N^2$  is biased with a systematic error equal to  $\theta_2^2/N$ .

Biasedness of ML estimators is a consequence of the so-called *invariance principle* which is stated below.

**Theorem 1.3 (Invariance principle).** *Let  $g$  be a function from  $\Theta$  to some multidimensional interval  $\Gamma \subset \mathbb{R}^k$ , ( $k$  finite). If  $\hat{\theta}(\mathbf{x})$  is the M.L. estimator of  $\theta$ , then  $g(\hat{\theta}(\mathbf{x}))$  is the M.L. estimator of  $g(\theta)$ .*

*Proof.* We give a simplified proof assuming that  $g$  is invertible. A complete proof can be found in the original article [?]. Let  $g^{-1}$  be the inverse of  $g$  and define

$$\tilde{\ell}(x, \gamma) = \ell(x, g^{-1}(\gamma)) = \ell(x, \theta) \Big|_{\theta=g^{-1}(\gamma)} \tag{1.3.9}$$

which is just a re-parametrization of the likelihood  $\ell(x, \cdot)$  of  $x$ .

It is now obvious that  $\tilde{\ell}(x, \gamma)$  has a maximum in  $\gamma = \hat{\gamma}(x)$  if and only if  $\ell(x, \theta)$  has a maximum (of the same value) in  $\theta = \hat{\theta}(x)$  and the two maximizing points are related by the transformation  $\theta = g^{-1}(\gamma)$ , that is

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

It follows that the M.L. estimate of  $\gamma$  is  $\hat{\gamma}(x) = g(\hat{\theta}(x))$ .  $\square$

It is then clear that if  $\hat{\theta}$  is an unbiased estimator of  $\theta$ ,  $g(\hat{\theta})$  cannot in general also be an unbiased estimator of  $g(\theta)$  since the operations  $\mathbb{E}_\theta$  and  $g(\cdot)$  do not commute; i.e.

$$\mathbb{E}_\theta g(\hat{\theta}(\mathbf{x})) \neq g(\mathbb{E}_\theta \hat{\theta}(\mathbf{x})) = g(\theta) \quad ,$$

unless  $g$  is linear.

**Esempio 13.6**

[Cramèr]

In many applications 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 [10, pag. 219–220]). La variabile casuale (scalare)  $y$  è distribuita in modo log-normale se  $y \geq 0$  (c.p.1), e se  $\log y \sim \mathcal{N}(\mu, \sigma^2)$  o, più in generale, se per qualche  $a \geq 0$ ,  $\log(y - a) \sim \mathcal{N}(\mu, \sigma^2)$ . In quest'ultimo caso ovviamente dovrà essere  $y \geq a$  (c.p.1) e si può facilmente controllare che la densità di  $y$  è data dall'espressione

$$\frac{1}{\sigma(y-a)2\pi} \exp -\frac{1}{2\sigma^2} \left[ \log(y-a) - \mu \right]^2 . \quad (1.3.10)$$

Supponiamo di osservare una variabile  $y$  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)  $x_1 = \log y_1, \dots, x_N = \log y_N$  e da questi ricavare le stime dei parametri  $\theta_1$  e  $\theta_2^2$  nella distribuzione (Gaussiana) di  $\log y$ . Una volta fatto questo le stime della media  $\xi$  e della varianza  $\lambda^2$  dalla distribuzione log-normale si ricaveranno usando le formule (ricordare l'espressione della  $E\{e^{tx}\}$  con  $x \sim \mathcal{N}(\theta_1, \theta_2^2)$ !)

$$\begin{aligned} \xi &= \exp\left(\theta_1 + \frac{\theta_2^2}{2}\right) , \\ \lambda^2 &= \xi^2 \left(e^{\theta_2^2} - 1\right) , \end{aligned} \quad (1.3.11)$$

che danno appunto media e varianza di  $y$  in funzione della media ( $\theta_1$ ) e varianza ( $\theta_2^2$ ) di  $\log y$ . Questo approccio è in effetti validato dal teorema precedente, purché beninteso si tratti di stime di M.V.. In conclusione, detto

$$\hat{\theta}_1(x_1, \dots, x_N) = \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i = \frac{1}{N} \sum_{i=1}^N \log y_i$$

lo stimatore di M.V. di  $\theta_1$  e

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

lo stimatore di M.V. di  $\theta_2^2$ , gli stimatori di M.V. di  $\xi$  e  $\lambda^2$ , rispettivamente  $\hat{\xi}$  e  $\hat{\lambda}^2$ , sono dati dalle formule

$$\begin{aligned} \hat{\xi}(y_1, \dots, y_N) &= \exp\left(\hat{\theta}_1 + \frac{\hat{\theta}_2^2}{2}\right) , \\ \hat{\lambda}^2(y_1, \dots, y_N) &= \hat{\xi}^2 \left| \exp(\hat{\theta}_2^2) - 1 \right| . \end{aligned} \quad (1.3.12)$$

□

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  $\theta$ ,  $\hat{\theta}(\mathbf{y}_1, \dots, \mathbf{y}_N)$ , produce dei numeri in corrispondenza ai dati osservati  $(y_1, \dots, y_N)$  che sono “vicini” in senso probabilistico al parametro vero  $\theta_0$ . In pratica questa vicinanza si valuta concretamente ad esempio mediante la *varianza* della v.c.  $\hat{\theta}$ . Sfortunatamente questa varianza è in generale *essa stessa funzione del parametro incognito*  $\theta$ . 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  $\mathbf{y} \sim \mathcal{N}(\mu, \theta_2^2)$  e sia  $s_N^2(\mathbf{y}_1, \dots, \mathbf{y}_N)$  lo stimatore di M.V. del parametro  $\theta_2^2$  basata su un campione di numerosità  $N$ . Come vedremo nel prossimo capitolo, la varianza di  $s_N^2$  è

$$\text{var}_\theta s_N^2 = 2(\theta_2^2)^2 \frac{N-1}{N^2} \quad , \quad (1.3.13)$$

che dipende ancora da  $\theta_2^2$ . Ne viene che la formula così com'è è di scarso uso pratico perchè  $\theta_2^2$  è incognito. Possiamo però dare un'espressione calcolabile per la *stima di M.V.* del parametro  $\text{var}_\theta s_N^2$  usando il principio di invarianza, nella forma

$$\widehat{\text{var}} s_N^2 = 2(s_N^2)^2 \frac{N-1}{N^2} \quad . \quad (1.3.14)$$

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. In realtà lo stimatore è *consistente*, asintoticamente 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 statistica 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.

### 1.3.1 Il Metodo dei Momenti

Il *metodo dei momenti* è stato proposto da K. Pearson agli albori della statistica. È un metodo per costruire stimatori di certi parametri di una distribuzione di probabilità nota uguagliando le espressioni dei primi momenti della distribuzione, ad esempio medie e varianze “teoriche”, a quelle dei corrispondenti momenti *campionari*; ad esempio alla media e alla varianza campionarie del campione definite in (1.1.7), (1.1.8). Ovviamente si vede subito che nel caso di distribuzioni Gaussiane *parametrizzate da media e varianza*  $(\theta_1, \theta_2^2)$ , questo principio stipula che gli stimatori di  $(\theta_1, \theta_2^2)$  sono proprio la media e la varianza campionarie del campione.

Quindi nel caso Gaussiano questo metodo di stima è equivalente alla massima verosimiglianza. Però anche nel caso Gaussiano  $(\theta_1, \theta_2^2)$  potrebbero essere date come funzioni note di un altro parametro incognito diverso da media e varianza, ad esempio un unico parametro reale nel caso scalare oppure un parametro vettoriale di dimensione più piccola di quella di  $(\mu, \Sigma)$  nel caso vettoriale. Nel caso di stima di parametri di modelli dinamici Gaussiani il metodo fornisce una procedura più semplice della massima verosimiglianza.

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

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

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

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

mentre il momento secondo è

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

Gli stimatori di  $\alpha$  e  $\beta$  col metodo dei momenti, si ottengono pertanto uguagliando i momenti teorici ai momenti campionari, ovvero risolvendo il sistema di due equazioni:

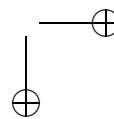
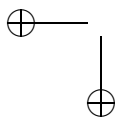
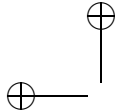
$$\begin{cases} \alpha \beta & = \bar{\mathbf{y}}_N = \frac{1}{N} \sum_{k=1}^N \mathbf{y}_k \\ \beta^2 \alpha (\alpha + 1) & = \bar{\mathbf{m}}_N^2 := \frac{1}{N} \sum_{k=1}^N \mathbf{y}_k^2. \end{cases}$$

ottenedo

$$\hat{\alpha} = \frac{\bar{\mathbf{y}}_N^2}{\bar{\mathbf{m}}_N^2 - \bar{\mathbf{y}}_N^2} \quad \hat{\beta} = \frac{\bar{\mathbf{m}}_N^2 - \bar{\mathbf{y}}_N^2}{\bar{\mathbf{y}}_N}. \quad (1.3.15)$$

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

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



## CHAPTER 2

# PARAMETER ESTIMATION FOR LINEAR MODELS

### 2.1 Linear Statistical Models

Let  $\mathbf{y}$  be an  $N$ -dimensional random vector whose probability distribution is an unknown member of a parametric family  $\{F_\theta ; \theta \in \Theta\}$ .

A *statistical (or probabilistic) model* of  $\mathbf{y}$  is a representation

$$\mathbf{y} = f(\theta, \mathbf{w}) \quad , \quad (2.1.1)$$

where  $f$  is a known function and  $\mathbf{w}$  is a random vector of a known probability distribution, whose probabilistic structure is simpler than that of  $\mathbf{y}$ . Typically one requires  $\mathbf{w}$  to have independent components or to be Gaussian with uncorrelated components.

A statistical model is usually regarded as a description of the physical device which generates the observations. In many applications  $\mathbf{w}$  is a model of the “noise” affecting the observations; the noise being an aggregate description of a multitude of unknown, uncontrollable factors which act on the system so as to make the results of a measurement of  $\mathbf{y}$  impossible to predict exactly; i.e. uncertain. It is a commonly accepted fact that a reasonable mathematical description of this aggregate disturbance factor should be probabilistic although the philosophical grounds for this choice are rather subtle and have been challenged by some [?]. We shall hereafter assume that some probabilistic description of the noise is available. Very often this probabilistic description will be limited to the knowledge of first and second order moments. In any case the *random noise* in the model (2.1.1) will be the source of uncertainty in the relation linking the parameter  $\theta$  which is the primary object of the measurement experiment.

Nonostante 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  $\theta$ , 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 (2.1.1) 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)  $\theta$ .

Ammettiamo che l'incertezza sul risultato di ciascuna misura si possa esprimere come un errore additivo secondo uno schema del tipo

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

dove  $s_k(\theta)$  è la caratteristica "ideale" dello strumento di misura, funzione nota di  $\theta$  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 sommano) 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*  $w_k$ . Supponendo che vi sia assenza di errori sistematici,  $w_k$  può essere ipotizzata a *media nulla*.

Pensiamo allora  $y_k$  come la determinazione di una variabile casuale scalare  $y_k$  e raccogliamo gli  $N$  campioni  $(y_1, \dots, y_N)$  in un vettore colonna  $\mathbf{y}$ . Si può così scrivere sinteticamente

$$\mathbf{y} = s(\theta) + \mathbf{w} \quad , \quad (2.1.2)$$

dove abbiamo introdotto i due vettori colonna

$$\begin{aligned} s(\theta) &= [s_1(\theta), \dots, s_N(\theta)]' \quad , \\ \mathbf{w} &= [\mathbf{w}_1, \dots, \mathbf{w}_N]' \quad . \end{aligned} \quad (2.1.3)$$

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 (2.1.2) la matrice di covarianza del rumore

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

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

$$\mathbf{y} = s(\theta) + \sigma \mathbf{w} \quad , \quad (2.1.4)$$



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

L'altro caso estremo si presenta quando l'intera matrice varianza di  $\mathbf{w}$  è incognita e va quindi considerata tra i parametri incogniti da stimare. Il modello (2.1.2) può allora essere riscritto come

$$\mathbf{y} = s(\theta) + R^{1/2} \mathbf{w} \quad , \quad (2.1.5)$$

dove ora  $R^{1/2} \in R^{N \times N}$  è una radice quadrata della varianza incognita<sup>4</sup> che si assume non singolare e  $\mathbf{w} \sim \mathcal{N}(0, I)$ . I problemi di stima associati al modello (2.1.5) sono però molto complicati. Nel seguito considereremo un caso intermedio fra (2.1.4) e (2.1.5). Supporremo cioè  $\mathbf{w}$  Gaussiano e di covarianza parzialmente nota, della forma  $\sigma^2 R$  con  $\sigma^2$  incognita ed  $R$  nota e definita positiva.

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

$$s(\theta) = S\theta \quad , \quad S \in \mathbb{R}^{N \times p} \quad , \quad (2.1.6)$$

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

$$\mathbf{y} = S\theta + \sigma \mathbf{w} \quad . \quad (2.1.7)$$

**Remark 2.1.** Quando la varianza del rumore  $R$  è nota il modello (2.1.7) può essere facilmente normalizzato a uno in cui  $\text{Var}\{\mathbf{w}\}$  è l'identità. In effetti, dato che  $R$  è simmetrica e definita positiva, essa ammette "radici quadrate"  $R^{1/2} \in R^{N \times N}$  tali che  $R = R^{1/2}(R^{1/2})^\top$ , calcolabili ad esempio mediante una fattorizzazione di Cholesky. Il modello normalizzato si ottiene moltiplicando a sinistra entrambi i membri di (2.1.7) per  $R^{-1/2}$  e ridefinendo opportunamente le variabili come

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

Sebbene questa normalizzazione faciliti i calcoli, in questo capitolo essa non verrà usata perchè nella soluzione dei problemi di stima legati al modello (2.1.7) la matrice  $R$  gioca un ruolo importante di "matrice peso" che ispira i procedimenti empirici di stima ai minimi quadrati che verranno esposti più avanti e sono importanti nelle applicazioni. Con la normalizzazione questo ruolo verrebbe completamente mascherato.

### 2.1.1 Sul processo di Modellizzazione statistica

Nel ricavare il modello lineare (2.1.7) ci siamo per semplicità riferiti sostanzialmente a una classe di problemi di misura che si possono ricondurre allo schema di Gauss. In realtà il modello si presta a molte interpretazioni e generalizzazioni. Una situazione tipica è quella in cui la successione di osservazioni  $\{y_1, \dots, y_N\}$ , che noi qui per semplicità supporremo scalari (ma l'estensione al caso vettoriale di quanto verremo esponendo è immediata), dipende da una *variabile esogena* o "di ingresso"  $u$  i cui campioni  $\{u_t ; t = 1, \dots, N ; u_t \in R^q\}$  possono essere misurati

<sup>4</sup>Questo significa che  $R^{1/2}(R^{1/2})^\top = R$ .

(o talvolta scelti) dallo sperimentatore, in concomitanza o prima dell'acquisizione della misura corrispondente. Sulla dipendenza di ciascuna misura  $y_t$  da  $u_t$  (o eventualmente da una parte o da tutte le  $u_1, \dots, u_N$ ) si potrebbe anche avere scarsa informazione a priori. In ogni caso, basandosi sulla conoscenza a priori del meccanismo che lega i dati "ingresso-uscita" ( $u_t; y_t$ ), si può assegnare a priori una *classe parametrica* di modelli del tipo

$$y_t = f(u_t, \theta, t) + w_t, \quad t = 1, \dots, N \quad (2.1.8)$$

dove  $f$  è una funzione nota e  $\theta$  un parametro  $p$ -dimensionale da determinarsi in modo tale che si abbia la "migliore" descrizione possibile dei dati misurati ( $y_1, \dots, y_N$ ) corrispondenti a certi valori assegnati ( $u_1, \dots, u_N$ ) alla variabile  $u$  negli istanti di misura. Il termine  $w_t$  rappresenta l'*errore di modellizzazione*. In questi casi questo termine non ha un'origine che possa condurre ad una descrizione probabilistica chiara come nel caso della teoria degli errori. L'errore può dipendere dalla inadeguatezza della classe parametrica di funzioni scelta (tipicamente funzioni lineari), dalla presenza di influenze di altre variabili di disturbo non misurabili e quindi non modellate etc. etc. Si *postula* allora per  $w_t$  una descrizione probabilistica, normalmente assumendo che la successione  $\{w_t\}$  sia assimilabile ad una traiettoria di un processo di rumore bianco. Il modello (2.1.8) viene spesso chiamato *modello di regressione*. Nei casi in cui  $f$  dipende linearmente da  $\theta$  il modello si riduce al modello lineare (2.1.6) discusso più sopra.

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 abbastanza 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  $\theta$  nel modello (2.1.8) hanno un preciso significato fisico e spesso in questi casi lo scopo della stima è di ricavare informazioni su  $\theta$  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

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

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.

Osserviamo infine che la definizione (2.1.1) di modello statistico è in realtà abbastanza restrittiva. Essa assume che il vettore delle osservazioni possa essere generato come una funzione nota di un parametro e di un vettore non osservabile di “errori” (o rumore aleatorio) avente certe caratteristiche probabilistiche prefissate. In realtà si possono dare modelli statistici più complessi in cui il vettore delle misure è definito indirettamente facendo intervenire altre variabili non misurabili. Ad esempio, i cosiddetti modelli a *errori nelle variabili* (EIV) sono modelli di regressione in cui sia la variabile “spiegate”  $\{y_t\}$  che le variabili esogene  $\{u_t\}$  sono descritte come se non fossero misurate perfettamente ma fossero entrambe soggette

a errori di misura (o rumore) additivi

$$\hat{y}_t = f(\hat{u}_t; \theta) \tag{2.1.9}$$

$$y_t = \hat{y}_t + w_t \tag{2.1.10}$$

$$u_t = \hat{u}_t + v_t \tag{2.1.11}$$

dove  $\hat{y}_t$ ,  $\hat{u}_t$  sono le uscite e gli ingressi “veri” non osservabili,  $y_t$ ,  $u_t$  le variabili effettivamente misurabili (le osservazioni) e  $w$  e  $v$  sono rumori di misura che si assumono normalmente tra loro scorrelati e scorrelati dai segnali “veri”  $\hat{y}$ ,  $\hat{u}$ . Naturalmente la dipendenza istantanea di  $\hat{y}_t$  da  $\hat{u}_t$  può essere generalizzata assumendo per esempio una dipendenza causale di  $\hat{y}_t$  dalle variabili di regressione precedenti  $\{\hat{u}_s; s \leq t\}$ .

La comprensione di questa classe di modelli non è però ancora soddisfacente e i problemi di stima relativi sono stati affrontati solo in casi molto particolari.

## 2.2 Stima di M.V. nel modello lineare

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

**Problem 2.1.** *Trovare le stime di M.V. dei parametri  $\theta \in \mathbb{R}^p$  e  $\sigma^2 \in \mathbb{R}_+$  nel modello lineare (2.1.7), dove  $S \in \mathbb{R}^{N \times p}$  è una matrice nota e  $w$  è un vettore aleatorio Gaussiano di media zero e varianza nota  $R$ , definita positiva.*

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

$$\begin{aligned} \ell(y, \theta, \sigma^2) &= -\frac{N}{2} \log 2\pi - \frac{1}{2} \log [\det(\sigma^2 R)] - \frac{1}{2} (y - S\theta)^\top (\sigma^2 R)^{-1} (y - S\theta) \\ &= -\frac{N}{2} \log 2\pi - \frac{N}{2} \log \sigma^2 - \frac{1}{2} \log \det R - \frac{1}{2\sigma^2} (y - S\theta)^\top R^{-1} (y - S\theta), \end{aligned} \tag{2.2.1}$$

cosicché

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

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

Inoltre si ha

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

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

$$z_\theta := \frac{\partial \ell(y, \theta, \sigma^2)}{\partial \theta}, \quad z_\sigma := \frac{\partial}{\partial \sigma^2} \ell(y, \theta, \sigma^2),$$

e ricordiamo che

$$I(\theta, \sigma) = E_{\theta, \sigma} \begin{bmatrix} \mathbf{z}_\theta \mathbf{z}_\theta^\top & \mathbf{z}_\theta \mathbf{z}_\sigma^\top \\ \mathbf{z}_\theta^\top \mathbf{z}_\sigma & \mathbf{z}_\sigma^\top \mathbf{z}_\sigma \end{bmatrix}. \quad (2.2.2)$$

Svolgendo i calcoli, si trova

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

Ponendo inoltre

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

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

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

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

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

si vede che

$$E_{\theta, \sigma} \mathbf{z}_\sigma^2 = E_{\theta, \sigma} \left\{ \frac{1}{2\sigma^2} \left[ \frac{\|\tilde{\mathbf{y}}\|^2}{\sigma^2} - N \right] \right\}^2. \quad (2.2.3)$$

Vedremo più avanti che, se  $\mathbf{y} \sim \mathcal{N}(\mu, \Sigma)$ , la forma quadratica  $(\mathbf{y} - \mu)^\top \Sigma^{-1} (\mathbf{y} - \mu)$  ha una distribuzione del tipo  $\chi^2$  con un numero di gradi di libertà pari alla dimensione di  $\mathbf{y}$ . Allora,  $\frac{\|\tilde{\mathbf{y}}\|^2}{\sigma^2} \sim \chi^2(N)$  e dalla (2.3.3) segue

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

Mettendo insieme questi risultati si trova infine

$$I(\theta, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2} S^\top R^{-1} S & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix}. \quad (2.2.4)$$

**Proposition 2.1.** *Nel modello (2.1.7) sia  $N \geq p$ . Allora,  $\theta$  è globalmente identificabile se e solo se  $S$  ha rango  $p$ .*

Difatti  $I(\theta, \sigma^2)$  è non singolare se e solo se  $S^\top R^{-1} S$  è invertibile e questo avviene allora e solo allora che  $S^\top R^{-1} S\theta = 0$  implica  $\theta = 0$ . Ne segue che lo spazio nullo di  $S$  contiene solo il vettore zero. Ovviamente se lo spazio nullo di  $S$  contenesse un  $\xi \neq 0$ ,  $\theta_0$  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  $\theta$  non può essere inferiore a  $\sigma^2[S^\top R^{-1}S]^{-1}$ . Analogamente quella di uno stimatore corretto di  $\sigma^2$  non può essere inferiore a  $\frac{2\sigma^4}{N}$ .

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

$$\hat{\theta}(y) = [S^\top R^{-1}S]^{-1} S^\top R^{-1}y. \quad (2.2.5)$$

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

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

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

$$\hat{\theta}(y) = Ay \quad , \quad A := [S^\top R^{-1}S]^{-1} S^\top R^{-1}. \quad (2.2.6)$$

### Proprietà dello stimatore di M.V. nel modello lineare

Lo stimatore (2.2.5) del parametro  $\theta$  ha le seguenti proprietà:

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

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

dato che manifestamente  $\mathbf{A} S = \mathbf{I}$ . Notare che  $\mathbf{A}$  è una inversa sinistra di  $S$ .

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

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

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

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

Questa proprietà è conseguenza della linearità.

### Interpretazione geometrica

Dalla (2.2.1) è evidente che  $\hat{\theta}(y)$  è la funzione che *minimizza, rispetto a  $\theta$ , la forma quadratica*  $(y - S\theta)^\top R^{-1}(y - S\theta)$ , 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  $\langle x, y \rangle_{R^{-1}} := x^\top R^{-1}y$ , nel senso che

$$(y - S\theta)^\top R^{-1}(y - S\theta) = \|y - S\theta\|_{R^{-1}}^2 \quad , \quad (2.2.7)$$

con ovvio significato dei simboli. Per un dato  $y \in \mathbb{R}^N$  minimizzare la distanza (2.2.7), rispetto a  $\theta$ , significa *cercare il vettore  $v \in S := \text{span}(S)$  (lo spazio vettoriale generato dalle colonne di  $S$ ) che ha minima distanza da  $y$ .*

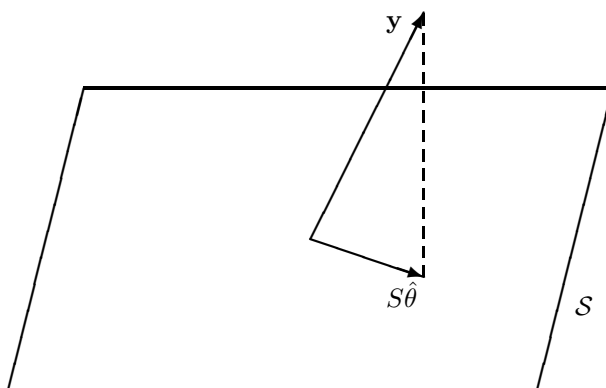


Figure 2.2.1. Proiezione ortogonale.

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

$$P = SA \quad , \quad (2.2.8)$$

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

$$P^\top = (SA)^\top = A^\top S^\top = R^{-1}S[S^\top R^{-1}S]^{-1}S^\top = R^{-1}SAR = R^{-1}PR, \quad (2.2.9)$$

cioè  $P^\top$  è simile a  $P$ .

Basandosi sulla classica caratterizzazione geometrica della proiezione ortogonale (si veda ad esempio [45, Cap II]), si trova allora che l'unico vettore  $S\theta$  di  $S$

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  $S$  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  $\theta$  tale per cui

$$S \perp (y - S\theta). \tag{2.2.10}$$

In altre parole

$$S^T R^{-1} y - S^T R^{-1} S \theta = 0 \tag{2.2.11}$$

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

**Proposition 2.2.** *Nel modello lineare-Gaussiano (2.1.7) lo stimatore a M.V. di  $\theta$  coincide con la funzione dei dati osservati che minimizza la distanza quadratica (2.2.7). In altre parole,  $\hat{\theta}(y)$  è lo stimatore ai minimi quadrati pesati di  $\theta$  con matrice peso  $R^{-1}$ .*

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

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

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  $\hat{\sigma}^2(y)$  è corretto e calcolarne la varianza occorre vedere come è distribuito.

Dobbiamo ora ricordare alcune proprietà della distribuzione  $\chi^2$ .

### 2.3 La distribuzione $\chi^2$

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

$$P(x \leq y < x + dx) = \frac{1}{2^{n/2} \Gamma(\frac{n}{2})} x^{(\frac{n}{2})-1} e^{-x/2} dx, \tag{2.3.1}$$

per  $x \geq 0$  e zero altrimenti. Nella (2.3.1)  $n$  è un numero naturale che si chiama *numero dei gradi di libertà* della distribuzione. La  $\chi^2$  è un caso speciale della distribuzione Gamma; la sua funzione caratteristica (abbreviata a f.c. nel seguito) è

$$\phi(it) := E e^{it y} = (1 - 2it)^{-n/2}, \tag{2.3.2}$$

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

$$\begin{aligned} \mu_1 &= n \\ \mu_2 &= 2n \\ \mu_3 &= 8n \\ \mu_4 &= 48n + 12n^2 \quad \text{ecc...} \end{aligned} \tag{2.3.3}$$



### 2.3. La distribuzione $\chi^2$

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

$$z_n := \frac{y - n}{\sqrt{2n}} ;$$

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

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

**Lemma 2.1 (Helly-Bray).** *Se  $\phi_n(t)$  è la f.c. di  $x_n$  e  $\phi(t)$  è la f.c. di  $x$ , allora*

$$x_n \xrightarrow{L} x \quad \text{se e solo se} \quad \phi_n(t) \rightarrow \phi(t) \quad , \quad \forall t. \quad (2.3.4)$$

Notiamo allora che la f.c.,  $\phi_n(t)$ , di  $z_n$  si può scrivere,

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

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

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

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

La distribuzione  $\chi^2$  interviene in molte questioni di inferenza statistica e qui di seguito ne elencheremo alcune proprietà importanti che stanno alla base del calcolo della distribuzione di probabilità di stimatori che sono forme quadratiche di variabili Gaussiane.

**Proposition 2.3.** *La somma di  $N$  variabili casuali indipendenti  $y_i \sim \chi^2(n_i)$  è distribuita secondo  $\chi^2(n)$  dove*

$$n = \sum_{i=1}^N n_i \quad , \quad (2.3.5)$$

*cioè i gradi di libertà si sommano.*

**Proof.** La prova di questo risultato si basa sulla nota espressione della f.c. della somma  $\sum_1^N \mathbf{y}_i$  di variabili indipendenti come prodotto delle f.c.,  $\phi_i(t)$ , delle  $\mathbf{y}_i$ . Moltiplicando tra loro espressioni del tipo (2.3.2) si vede in effetti che i gradi di libertà si sommano.  $\square$

**Proposition 2.4.** Se  $\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2$  dove  $\mathbf{y}_1$  è indipendente da  $\mathbf{y}_2$  e sia  $\mathbf{y}$  che  $\mathbf{y}_2$  hanno distribuzione  $\chi^2$  con gradi di libertà rispettivamente  $n$  ed  $n_2$  con  $n > n_2$ , allora la variabile  $\mathbf{y}_1$  è distribuita come  $\chi^2(n - n_2)$ .

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

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

Sostituendo in questo rapporto le espressioni (2.3.2) per le rispettive funzioni caratteristiche, si ricava l'asserto.  $\square$

**Proposition 2.5.** La distribuzione di

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

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

**Proof.** In effetti basta mostrare che la d.d.p. di  $\mathbf{z} := (\mathbf{y} - \mu)^2 / \sigma^2$  con  $\mathbf{y} \sim \mathcal{N}(\mu, \sigma)$  è  $\chi^2(1)$  e poi usare la proposizione 2.3. Notiamo che si può scrivere  $\mathbf{z} = \mathbf{x}^2$  con  $\mathbf{x} \sim \mathcal{N}(0, 1)$ . Usando le note regole per il calcolo della distribuzione di una funzione di variabile aleatoria, riferite alla funzione  $z = f(x)$  con  $f(x) = x^2$ , si può calcolare la densità di probabilità di  $\mathbf{z}$  come

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

che è proprio  $\chi^2(1)$ .  $\square$

**Proposition 2.6.** La distribuzione della varianza campionaria normalizzata

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

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

*Proof.* Mostriamo allo scopo il seguente risultato notevole.

**Lemma 2.2.** *Nelle ipotesi poste, le statistiche  $\bar{y}_n$  ed  $s_n^2$  sono indipendenti.*

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

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

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

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

Usiamo ora la solita identità

$$\sum_1^n (y_i - \mu)^2 = \sum_1^n (y_i - \bar{y}_n)^2 + n(\bar{y}_n - \mu)^2 \tag{2.3.6}$$

per scrivere

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

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  $(\bar{y}_n - \mu)^2/(\sigma^2/n) \sim \chi^2(1)$  (questo scende ancora dalla proposizione 2.5 con  $n = 1$ ). Per la proposizione 2.4 il primo addendo al secondo membro deve essere  $\chi^2(n - 1)$ .  $\square$

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

**Proposition 2.7.** *Se  $\mathbf{y} \sim \mathcal{N}(\mu, \Sigma)$  con  $\mu \in \mathbb{R}^m$  e  $\Sigma \in \mathbb{R}^{m \times m}$  definita positiva, allora*

$$(\mathbf{y} - \mu)^\top \Sigma^{-1} (\mathbf{y} - \mu) \sim \chi^2(m). \tag{2.3.7}$$

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

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

e quindi l'ultimo membro è  $\chi^2(m)$  per la proposizione 2.3.

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

**Proposition 2.8.** *Sia  $\mathbf{z} \sim \mathcal{N}(0, I_m)$  e  $Q \in \mathbb{R}^{m \times m}$ . Allora la forma quadratica  $\mathbf{z}^\top Q \mathbf{z}$  è distribuita secondo  $\chi^2$  se e solo se  $Q$  è idempotente, ovvero  $Q = Q^2$ . In questo caso il numero di gradi di libertà è  $r = \text{rango } Q$ .*

**Proof.** La prova di questo risultato è basata su un procedimento di diagonalizzazione di  $Q$ . Dato che  $Q$  è simmetrica (notare che può sempre essere supposta tale) e  $Q = Q^2$ , essa è una matrice di proiezione ortogonale in  $\mathbb{R}^m$ . I suoi autovalori non nulli sono pertanto uguali a uno (in numero di  $r = \text{rango } Q$ ). La decomposizione spettrale della matrice  $Q$  si può così scrivere

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

ovvero

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

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

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

dove il vettore  $r$ -dimensionale  $\mathbf{z}_1 := U_1^\top \mathbf{z}$  è distribuito come  $\mathcal{N}(0, I_r)$ . La conclusione segue ancora dalla proposizione 2.3.  $\square$

### Caratterizzazione dello stimatore della varianza

**Theorem 2.1.** *Lo stimatore di M.V. della varianza  $\sigma^2$  nel modello lineare (2.1.7) ha distribuzione di probabilità corrispondente alla*

$$\frac{N \hat{\sigma}^2(\mathbf{y})}{\sigma^2} \sim \chi^2(N - p). \tag{2.3.8}$$

In particolare la sua media e varianza sono date da

$$E_{\theta, \sigma^2} \hat{\sigma}^2(\mathbf{y}) = \sigma^2 \frac{N - p}{N}, \tag{2.3.9}$$

$$\text{Var}_{\theta, \sigma^2} \hat{\sigma}^2(\mathbf{y}) = \sigma^4 \frac{2(N - p)}{N^2}. \tag{2.3.10}$$

**Proof.** 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

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

### 2.3. La distribuzione $\chi^2$

39

il quale è chiaramente distribuito secondo la  $\mathcal{N}(0, I)$ . Dalla (2.2.11) si ricava poi con facili passaggi la

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

dove si è usata la proprietà di similitudine  $P^\top = R^{-1} P R$  stabilita nella (2.2.9). Notiamo ora che la matrice tra parentesi quadre, diciamola  $Q$ , è idempotente giacché, sempre per la (2.2.9), 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 dalla proposizione 2.8 stabilita più sopra che  $\mathbf{z}^\top Q \mathbf{z} \sim \chi^2(N - p)$ .  $\square$

**Remark 2.2.** Come si vede dalla (2.3.9) 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  $\sigma^2$  la quantità

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

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

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

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  $\theta$  nel modello lineare Gaussiano (2.1.7) 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  $w$ . In altre parole, il calcolo di  $\hat{\theta}(\mathbf{y})$  e  $\hat{\sigma}(\mathbf{y})$  per il modello lineare e gaussiano (2.1.7) si riduce alla soluzione di un problema di *minimi quadrati pesati*. Come vedremo nel prossimo capitolo, un problema di approssimazione ai minimi quadrati dei dati mediante una funzione lineare di  $\theta$  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  $\theta$ , (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.

## 2.4 Il principio dei Minimi Quadrati e il suo significato statistico

In generale si dispone raramente di informazione sufficiente per descrivere l'incertezza nei dati mediante modelli probabilistici di struttura nota come il modello lineare-Gaussiano (2.1.7). Molto spesso, si guarda a modelli del tipo (2.1.8) semplicemente come descrizioni "approssimate" e si cerca di determinare il parametro libero  $\theta$  in modo tale che si abbia la migliore descrizione possibile dei dati misurati  $(y_1, \dots, y_N)$  corrispondenti a certi valori assegnati  $(u_1, \dots, u_N)$  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*

$$\hat{y}_t(\theta) = f(u_t, \theta, t) \quad , \quad t = 1, \dots, N \quad , \quad (2.4.1)$$

in corrispondenza ai valori assegnati  $(u_1, \dots, u_N)$  di  $u_t$  e al valore generico  $\theta$  del parametro.

Si arriva in questo modo a definire una cifra di merito

$$V(\theta) := \sum_1^N [y_t - \hat{y}_t(\theta)]^2 = \sum_1^N [y_t - f(u_t, \theta, t)]^2 \quad (2.4.2)$$

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

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

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

$$\hat{\theta} = \hat{\theta}(y_1, \dots, y_N, u_1, \dots, u_N) \quad , \quad (2.4.3)$$

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  $\theta$  e la (2.4.3) *stima* ai minimi quadrati di  $\theta$ . 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*.

### 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  $y(t)$  nell'intervallo  $[-T/2, T/2]$ , vogliamo trovare una combinazione

lineare delle funzioni  $1, \sin \frac{2\pi}{T} t, \dots, \sin \frac{2n\pi}{T} t, \cos \frac{2\pi}{T} t, \dots, \cos \frac{2n\pi}{T} t$ , secondo i coefficienti  $\theta_i, i = 0, 1, \dots, 2n$  (questa è la ridotta  $n$ -sima della serie di Fourier di  $y(t)$ ) diciamola

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

tale per cui lo scostamento quadratico

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

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

$$V_Q(\theta) := \sum_1^N q_t [y(t) - f(u_t, \theta, t)]^2, \tag{2.4.4}$$

dove  $q_1, \dots, q_N$  sono numeri positivi, grandi se le misure corrispondenti sono affidabili e piccoli se non lo sono. La (2.4.4) si può riscrivere come

$$V_Q(\theta) = [y - f(u, \theta)]^T Q [y - f(u, \theta)] = \|y - f(u, \theta)\|_Q^2, \tag{2.4.5}$$

dove

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \quad f(u, \theta) = \begin{bmatrix} f(u_1, \theta, 1) \\ \vdots \\ f(u_N, \theta, N) \end{bmatrix} \tag{2.4.6}$$

e  $Q = \text{diag}\{q_1, \dots, q_N\}$ . La generalizzazione della (2.4.5) 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 (2.1.8) è *lineare nei parametri*, cioè quando

$$f(u_t, \theta, t) = \sum_1^p \theta_i \phi_i(u_t, t). \tag{2.4.7}$$

Siccome  $u_t$  è una quantità *nota* si può anche ometterla nell'argomento delle  $\phi_i$  (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  $(u_1, \dots, u_N)$ ).

Scriveremo così la (2.4.6) nel modo usuale e cioè

$$f(u_t, \theta, t) := s^\top(t) \theta \quad , \quad (2.4.8)$$

con  $s^\top(t)$  vettore riga  $p$ -dimensionale, funzione *nota* dell'indice  $t$ . Il blocco delle  $N$  misure sia rappresentato dal vettore  $N$ -dimensionale  $y$  definito in (2.4.6) e sia  $S$  la matrice  $N \times p$

$$S = \begin{bmatrix} s^\top(1) \\ \vdots \\ s^\top(N) \end{bmatrix} . \quad (2.4.9)$$

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

$$V_Q(\theta) = [y - S\theta]^\top Q [y - S\theta] = \|y - S\theta\|_Q^2 \quad , \quad (2.4.10)$$

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

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

che si può riscrivere come

$$S^\top Q S \theta = S^\top Q y . \quad (2.4.11)$$

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

Nel seguito supporremo che sia

$$\text{rango } S = p \leq N . \quad (2.4.12)$$

Questa condizione semplifica la trattazione del problema anche se non è essenziale per portare avanti l'analisi. In effetti, se il rango di  $S$  è minore di  $p$ , il problema può essere riparametrizzato usando un numero minore di variabili  $\{\theta_i\}$  e una matrice  $S$  con un numero minore di colonne, di rango pieno. Le equazioni normali si possono allora risolvere ottenendo un'unica soluzione

$$\hat{\theta}(y) = [S^\top Q S]^{-1} S^\top Q y \quad , \quad (2.4.13)$$

dalla quale si vede che lo stimatore ai M.Q. è *sempre una funzione lineare* delle misure. Nel caso in cui la (2.4.12) non valga, si può usare la pseudoinversa di  $S^\top Q S$ , ma in questo caso si perde l'unicità della soluzione.

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

$$P = S [S^\top Q S]^{-1} S^\top Q \quad (2.4.14)$$



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

$$\hat{\varepsilon}_t := y_t - s^\top(t) \hat{\theta} \quad , \quad t = 1, \dots, N \quad , \quad (2.4.15)$$

vale

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

Avevamo già incontrato queste formule studiando le proprietà dello stimatore di M.V. di  $\theta$  nel modello lineare e Gaussiano (2.1.7). Abbiamo visto che in questo caso, il calcolo dello stimatore (di M.V.) di  $\theta$  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.

**Remark 2.3.** *Lo stimatore di M.V. del parametro  $\theta$  nel modello lineare Gaussiano (2.1.7) è uno stimatore ai M.Q. pesati con matrice  $Q$  uguale all'inversa della matrice di varianza del rumore  $w$ .*

Questo risultato è per ora solo una curiosa coincidenza. Notiamo che se la distribuzione di  $w$  non è Gaussiana, i calcoli fatti alla sezione 2.1 non hanno più valore. In genere lo stimatore di M.V. di  $\theta$  in un modello lineare in cui  $w$  non è normalmente distribuito è una funzione *non lineare* delle osservazioni e pertanto, vista la (2.4.13), 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. (2.4.13).

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  $Q$ .

**Hypothesis (Ipotesi sul meccanismo di generazione dei dati).** *Supponiamo che le misure  $\{y_t\}$  siano generate da un modello lineare del tipo (2.1.7)*

$$y_t = s^\top(t) \theta + \sigma w_t \quad , \quad t = 1, \dots, N \quad , \quad (2.4.16)$$

*in cui però gli errori di modellizzazione hanno distribuzione di probabilità non nota. Si sa solo che  $w = [w_1, \dots, w_N]^\top$  è un vettore casuale  $N$ -dimensionale di media nulla e*

varianza  $\sigma^2 R$  con  $R$  nota e definita positiva,

$$E\mathbf{w} = 0 \quad , \quad \text{Var}(\mathbf{w}) = E\mathbf{w}\mathbf{w}^\top = \sigma^2 R. \quad (2.4.17)$$

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  $\sigma$  nell'errore di misura, si assuma che  $\bar{\mathbf{w}}_t := \sigma\mathbf{w}_t$  abbia media  $\mu$  indipendente da  $t$  e si ponga  $\tilde{\mathbf{w}}_t := \bar{\mathbf{w}}_t - \mu$ . Si possono allora riscrivere le misure ponendo

$$\mathbf{y}_t = s^\top(t)\theta + \mu + \tilde{\mathbf{w}}_t \quad , \quad t = 1, \dots, N \quad ,$$

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

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

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

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

**Proposition 2.9.** *Qualunque sia  $Q > 0$  lo stimatore (2.4.13) è corretto.*

Difatti

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

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

Se  $Q = R^{-1}$  lo stimatore ai M.Q. pesati si chiama *stimatore di Markov*. Supponendo per un attimo che  $R$  sia diagonale,  $R = \text{diag}\{r_1, \dots, r_N\}$ , 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

$$q_t = \frac{1}{\text{var } \mathbf{y}_t} = \frac{1}{\sigma^2} \frac{1}{r_t} \quad , \quad t = 1, \dots, N.$$

Notiamo che il termine  $1/\sigma^2$  (incognito) non influisce sulla minimizzazione di  $V_Q(\theta)$  dato che è indipendente da  $t$  e quindi si può portare fuori dal segno di sommatoria in (2.4.4).

La varianza di  $\hat{\theta}(\mathbf{y})$  si calcola facilmente a partire dalla (2.4.17),

$$\text{Var } \hat{\theta}(\mathbf{y}) = [S^T Q S]^{-1} S^T Q \sigma^2 R Q S [S^T Q S]^{-1} ; \quad (2.4.18)$$

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 (2.4.16) 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  $\theta$ , naturalmente nella classe di tutte le possibili funzioni lineari di  $\mathbf{y}$ ,

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

che sono stimatori *corretti* di  $\theta$ , ovvero

$$E \phi(\mathbf{y}) = \theta \quad \text{ovvero} \quad AS = I \quad , \quad (2.4.19)$$

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

$$\text{Var } \hat{\phi}(\mathbf{y}) \leq \text{Var } \phi(\mathbf{y}) . \quad (2.4.20)$$

La soluzione di questo problema è fornita dal celebre

**Theorem 2.2 (di Gauss-Markov).** *Il miglior stimatore lineare di  $\theta$ , per il modello (2.4.16) nel senso appena definito, è lo stimatore di Markov che ha varianza*

$$\text{Var } \hat{\phi}(\mathbf{y}) = \sigma^2 (S^T R^{-1} S)^{-1} . \quad (2.4.21)$$

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

$$\sigma^2 ARA^T \geq \sigma^2 (S^T R^{-1} S)^{-1} \quad , \quad (2.4.22)$$

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

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

è lineare e corretto e la sua varianza è esattamente uguale al secondo membro in (2.4.22).

Per provare la (2.4.22) ci si rifà alla disuguaglianza (equivalente alla non-negatività della varianza dell’errore di stima nella teoria della stima lineare Bayesiana<sup>5</sup>),

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

<sup>5</sup>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  $\tilde{\mathbf{x}} := \mathbf{x} - \hat{E}[\mathbf{x} | \mathbf{y}]$ .

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 (2.4.19) si ottiene la (2.4.22).  $\square$

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

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

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  $\varepsilon_t(\theta)$ , espressi come funzione delle misure e del parametro  $\theta$ ,

$$\varepsilon_t(\theta) := \mathbf{y}_t - \mathbf{s}^\top(t) \theta \quad , \quad t = 1, \dots, N \quad .$$

Notiamo infine che, in accordo con quanto anticipato nelle osservazioni precedenti, lo stimatore di Markov di  $\theta$  coincide con lo stimatore a M.V. nel caso in cui la distribuzione di probabilità di  $\mathbf{w}$  nel modello (2.4.16) è (nota e) Gaussiana. In genere però, se  $\mathbf{w}$  non è normalmente distribuito, la varianza (2.4.21) può risultare assai più grande della varianza del corrispondente stimatore di M.V. di  $\theta$ .

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

**Proposition 2.10.** *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 (2.2.5) (non fornisce necessariamente lo stimatore a M.V., ma) fornisce comunque lo stimatore che ha minima varianza nella classe degli stimatori lineari e corretti<sup>6</sup> di  $\theta$ .*

### Stima di $\sigma^2$ corrispondente allo stimatore di Markov

Per costruire uno stimatore di  $\sigma^2$ , si può ancora utilizzare la formula

$$\hat{\sigma}^2(y) = \frac{1}{N} \|\mathbf{y} - P\mathbf{y}\|_{R^{-1}}^2 = \frac{1}{N} V_{R^{-1}}(\hat{\theta}) \quad . \quad (2.4.23)$$

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  $\chi^2$  in generale. Si può comunque calcolarne la media in modo diretto

$$\begin{aligned} NE(\hat{\sigma}^2(\mathbf{y})) &= E(\mathbf{y}^\top (I - P)^\top R^{-1} (I - P) \mathbf{y}) \\ &= E(\mathbf{w}^\top (I - P)^\top R^{-1} (I - P) \mathbf{w}) = E(\mathbf{w}^\top R^{-1} (I - P) \mathbf{w}) \\ &= E \operatorname{tr}\{\mathbf{w}^\top R^{-1} (I - P) \mathbf{w}\} = E \operatorname{Tr}\{R^{-1} (I - P) \mathbf{w} \mathbf{w}^\top\} \\ &= E \operatorname{Tr}\{(I - P) (\mathbf{w} \mathbf{w}^\top) R^{-1}\} = \operatorname{Tr}\{(I - P) E(\mathbf{w} \mathbf{w}^\top) R^{-1}\} \\ &= \sigma^2 \operatorname{Tr}(I - P) \quad ; \end{aligned} \quad (2.4.24)$$

<sup>6</sup>Nella letteratura anglosassone lo stimatore lineare e corretto a minima varianza si denota con l'acronimo *B.L.U.E.* = Best Linear Unbiased Estimator.

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

$$E \hat{\sigma}^2(\mathbf{y}) = \frac{N - p}{N} \sigma^2. \quad (2.4.25)$$

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

Notiamo che se si vuole costruire uno *stimatore lineare* di  $c^\top \theta$  anziché di  $\theta$  ( $c^\top$  è un vettore riga noto), quello corretto e di varianza minima (sempre nella classe degli stimatori lineari) è semplicemente  $c^\top \hat{\theta}$ , dove  $\hat{\theta}$  è lo stimatore di Markov di  $\theta$ . Nel caso si volesse stimare una funzione *non lineare*,  $c(\theta)$ , di  $\theta$ , 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 (2.4.16) 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  $\sigma^2$  sia nota e conglobata in  $R$ ):

$$\text{BAYES} \quad \begin{cases} \hat{x}(\mathbf{y}) &= (P^{-1} + S^\top R^{-1} S)^{-1} S^\top R^{-1} \mathbf{y} \\ \Lambda &= (P^{-1} + S^\top R^{-1} S)^{-1} \end{cases} \quad (2.4.26)$$

$$\text{FISHER (MARKOV)} \quad \begin{cases} \hat{\theta}(\mathbf{y}) &= [S^\top R^{-1} S]^{-1} S^\top R^{-1} \mathbf{y} \\ \Sigma &= [S^\top R^{-1} S]^{-1}. \end{cases} \quad (2.4.27)$$

Come preannunciato a suo tempo, si vede che quando  $P \rightarrow \infty$  (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

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

per cui la stima Bayesiane 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 (2.4.26) introducendo delle misure fittizie, equivalenti alla conoscenza a priori su  $\mathbf{x}$ . Supponiamo allora di avere una osservazione addizionale

$$\mathbf{y}_0 = S_0 \mathbf{x} + \mathbf{w}_0 \quad , \quad (2.4.28)$$

effettuata *prima* di quella reale

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

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

$$P = (S_0^\top R_0^{-1} S_0)^{-1}, \quad (2.4.29)$$

cioè  $P$  è la varianza dello stimatore di Markov  $\hat{x}(\mathbf{y}_0)$  relativo al modello (2.4.28).

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

$$\begin{aligned} \hat{x}(\mathbf{y}_0, \mathbf{y}) &= \hat{x}(\mathbf{y}_0) + PS^\top (R + SPS^\top)^{-1} [\mathbf{y} - S\hat{x}(\mathbf{y}_0)] , \\ \Sigma &= P - PS^\top (R + SPS^\top)^{-1} SP. \end{aligned} \quad (2.4.30)$$

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

$$\begin{aligned} \text{stima di Fisher} &= \text{stima di Bayes con } P \rightarrow \infty \\ \text{stima di Bayes} &= \text{stima di Fisher con un'osservazione preliminare } \mathbf{y}_0 \\ &\quad \text{e condizione iniziale } \hat{x}(\mathbf{y}_0) = 0. \end{aligned}$$

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

$$\Lambda = \Sigma_x - \Sigma_{xy} \Sigma_y^{-1} \Sigma_{yx}, \quad (2.4.31)$$

che è sempre una matrice almeno semidefinita positiva, dato che si tratta di una varianza. Il termine che si sottrae ( $\text{Var } \hat{x}$ ) rappresenta la riduzione nell'incertezza a priori ( $\Sigma_x$ ) che si aveva su  $\mathbf{x}$ , a cui porta lo stimatore  $\hat{x}$ . 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:  $\hat{x}$  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  $f(y | x)$  (che ora gioca lo stesso ruolo della  $p(y, \theta)$  nel caso Fisheriano) di  $\hat{x}(\mathbf{y})$  non è  $x$ , ma bensì

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

e la matrice  $B$  non può essere l'identità, dato che dalla (2.4.31) scende

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

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

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

che era stata già richiamata qualche riga più in su.

## 2.5 Generalizzazione a misure vettoriali

Supponiamo di avere a che fare con un problema di stima su un modello lineare statico con misure  $\mathbf{y}(t)$  vettoriali; e.g.  $m$ -dimensionali e un parametro *matriciale*  $m \times p$ -dimensionale del tipo

$$\mathbf{y}(t) = \Theta \mathbf{u}(t) + \mathbf{w}(t), \quad t = 1, \dots, N, \quad \Theta \in \mathbb{R}^{m \times p}$$

dove  $\mathbf{u}(t)$  è un ingresso misurato  $r$ -dimensionale e  $\mathbf{w}(t)$  è rumore bianco scalare scorrelato da  $\mathbf{u}(s)$  per ogni  $t, s$ .

Si possono compattare le osservazioni, gli ingressi e le variabili di rumore in matrici a  $N$  colonne, definendo:

$$\mathbf{Y} := [\mathbf{y}(1) \quad \dots \quad \mathbf{y}(N)] \quad \mathbf{U} := [\mathbf{u}(1) \quad \dots \quad \mathbf{u}(N)] \quad \mathbf{W} := [\mathbf{w}(1) \quad \dots \quad \mathbf{w}(N)]$$

e riscrivere il modello in forma matriciale

$$\mathbf{Y} = \Theta \mathbf{U} + \mathbf{W} \tag{2.5.1}$$

associando a questo modello un funzionale quadratico del tipo

$$V(\Theta) = \|\mathbf{Y} - \Theta \mathbf{U}\|_Q^2 = \langle \mathbf{Y} - \Theta \mathbf{U}, \mathbf{Y} - \Theta \mathbf{U} \rangle_Q$$

dove  $\mathbf{Y}, \mathbf{U}$  sono le matrici di dati ingresso-uscita misurati,  $Q$  è una matrice peso  $N \times N$  definita positiva e il prodotto scalare tra matrici (delle stesse dimensioni) è definito dalla

$$\langle \mathbf{X}, \mathbf{Z} \rangle_Q := \text{Trace} \{ \mathbf{X} \mathbf{Q} \mathbf{Z}^T \} = \text{Trace} \{ \mathbf{Z} \mathbf{Q} \mathbf{X}^T \}$$

Si può quindi utilizzare il teorema delle proiezioni e ottenere

$$\hat{\Theta} = \mathbf{Y} \mathbf{Q} \mathbf{U}^T [\mathbf{U} \mathbf{Q} \mathbf{U}^T]^{-1}$$

Da notare che per  $Q = I$ , questa formula riscritta come

$$\hat{\Theta} = \frac{1}{N} \mathbf{Y} \mathbf{U}^\top \left[ \frac{1}{N} \mathbf{U} \mathbf{Q} \mathbf{U}^\top \right]^{-1}$$

“assomiglia molto” a quella ottenibile dal modello lineare (2.5.1) risolvendo l’equazione

$$\mathbb{E} \mathbf{Y} \mathbf{U}^\top = \Theta \mathbb{E} \mathbf{U} \mathbf{U}^\top + \mathbb{E} \mathbf{W} \mathbf{U}^\top.$$

Ad esempio supponiamo di poter osservare le traiettorie di stato, ingresso e uscita di un sistema stocastico lineare stazionario, in generale a più ingressi e a più uscite, senza reazione

$$\begin{bmatrix} \hat{\mathbf{x}}(t+1) \\ \mathbf{y}(t) \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}}(t) \\ \mathbf{u}(t) \end{bmatrix} + \begin{bmatrix} K \\ J \end{bmatrix} \mathbf{w}(t)$$

dove  $\mathbf{w}$  è rumore bianco. Coi dati osservati dall’istante  $t$  in poi si costruiscono le matrici (tutte a  $N + 1$  colonne)

$$\begin{aligned} Y_t &:= [y_t, y_{t+1}, y_{t+2}, \dots, y_{t+N}] \\ X_t &:= [x_t, x_{t+1}, x_{t+2}, \dots, x_{t+N}] \\ U_t &:= [u_t, u_{t+1}, u_{t+2}, \dots, u_{t+N}] \\ X_{t+1} &:= [x_{t+1}, x_{t+2}, \dots, x_{t+N+1}] \end{aligned}$$

Deve allora esistere una traiettoria di rumore bianco  $W_t := [w_t, w_{t+1}, w_{t+2}, \dots, w_{t+N}]$  per cui

$$\begin{bmatrix} X_{t+1} \\ Y_t \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} X_t \\ U_t \end{bmatrix} + \begin{bmatrix} K \\ J \end{bmatrix} W_t$$

## 2.6 Minimi quadrati e modelli non lineari

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

La stima vera e propria si fa con algoritmi iterativi che cercano di aggiornare la stima corrente del parametro in modo da dirigersi verso un minimo dello scarto quadratico medio tra i dati misurati e quelli predetti dal modello (predittore) parametrizzato dalla stima corrente del parametro. Usando tecniche di linearizzazione è



talvolta possibile dare delle valutazioni statistiche approssimate dell' accuratezza delle stime ottenute in questo modo.

Accenniamo al fatto che in molti problemi di regressione a scatola nera la descrizione dei dati mediante modelli parametrizzati linearmente (modelli lineari in  $\theta$ ) 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 imponente 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 nel 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 [11, 47, 61, 63].

## 2.7 Static Neural Networks

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

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

where  $\sigma$  is a "universal" function to be discussed later. In particular,  $\Sigma$  has *no offset* if  $c_0 = 0$ .

For simplicity, we will assume from now on that  $p = 1$ ; (single output nets). Generalizations to the multiple-output case are not hard but complicate the notations. Thus, from now on,  $C \equiv \mathbf{c}^\top$  is a row  $n$ -vector and  $c_0$  is a constant.

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

$$\vec{\sigma}_n(x_1, \dots, x_n) = (\sigma(x_1), \dots, \sigma(x_n))$$

The **behavior** of  $\Sigma$  is defined to be the map

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

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

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

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

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

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

## 2.8 Universal approximating functions

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

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

and the **radial function**

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

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

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

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

**Theorem 2.3 (Wiener 1932).** *In order for the family of shifted versions  $x \mapsto f(x + \tau)$ ;  $\tau \in \mathbb{R}$  of a function  $f \in L^2(\mathbb{R})$ , to be dense in  $L^2(\mathbb{R})$ , it is necessary and sufficient that the Fourier transform  $\hat{f}(j\omega)$  be nonzero almost everywhere.*

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

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

## 2.9 Gradient descent and back-propagation

See Duda Hart p.290-292

## 2.10 Aspetti numerici dei problemi ai minimi quadrati

### Condizionamento Numerico delle equazioni normali

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

$$S^T Q S \theta = S^T Q y \quad (2.10.1)$$

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 [65], a quello di Lawson-Hanson [31] e al testo di Golub e van Loan [20].

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  $\alpha$  viene rappresentato come una coppia  $\alpha = (m, c)$  dove  $m$  è la *mantissa* di  $\alpha$  e  $c$  la sua *caratteristica*. La mantissa è un numero il cui modulo è compreso tra 0,1 e 1 e contiene un *numero fisso*,  $n$ , di cifre significative (ad esempio 6). La caratteristica è l'esponente di 10 tale per cui  $\alpha \cong 10^C$ . Ad esempio  $\alpha = 3,562417\bar{9}$  ha le rappresentazioni

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

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  $f: \mathbb{R}^k \rightarrow \mathbb{R}^p$  definita matematicamente e un vettore  $k$ -dimensionale di "dati"  $\alpha$ . Si vuole calcolare  $x = f(\alpha)$ . Ad esempio, nella soluzione del problema

$$Ax = b \quad , \quad (2.10.2)$$

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,  $\alpha$ , 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  $\alpha$ ).
- 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  $\alpha$  e la seconda dalla "bontà" dell'algoritmo numerico che calcola  $f$ ) tendono sempre a sommarsi.

**Definition 2.1.** Diremo che il problema numerico  $x = f(\alpha)$  è mal condizionato se a piccoli errori percentuali su  $\alpha$  corrispondono grandi errori percentuali su  $x$ . In altri termini, detto  $x = f(\alpha)$  e  $x + \delta x = f(\alpha + \delta\alpha)$  si ha

$$\frac{\|\delta x\|}{\|x\|} \gg \frac{\|\delta\alpha\|}{\|\alpha\|}. \quad (2.10.3)$$

## Esempi

Consideriamo il sistema di equazioni  $Ax = b$ ,

$$\begin{bmatrix} 1 & 1 \\ 1 & 1.0001 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2.0001 \end{bmatrix}; \quad (2.10.4)$$

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

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

si verifica facilmente che la soluzione  $x$  diventa

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

In questo caso  $\|\delta b\|/\|b\| \cong 10^{-4}$ , mentre  $\|\delta x\|/\|x\| = 1/\sqrt{2}$ . Si vede che l'errore  $\delta b$  viene "amplificato" nel calcolo (esatto!) della soluzione del sistema (2.10.4) 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  $10^{22}$ !  $\diamond$

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.

**Definition 2.2.** 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  $\alpha$  (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**

$\delta A$  e  $\delta b$  sono errori di arrotondamento su  $A$  e  $b$ , e  $\delta x$  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

$$\|x\| := \left| \sum_i x_i^2 \right|^{1/2},$$

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

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:

$$\|A\|^2 = \sup_{x \neq 0} \frac{x^T A^T A x}{x^T x}. \quad (2.10.5)$$

Il quoziente al secondo membro è noto come quoziente di Rayleigh ed è uguale all'autovalore massimo di  $A^T A$ ,

$$\|A\|^2 = \max_i \lambda_i(A^T A). \quad (2.10.6)$$

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

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

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

$$\frac{\|\delta x\|}{\|x\|} \leq c(A) \frac{\|\delta b\|}{\|b\|} \quad (2.10.7)$$

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

$$\|A\|^2 = \lambda_{\text{MAX}}(A^T A)$$

$$\|A^{-1}\|^2 = \lambda_{\text{MAX}}(A^{-T} A^{-1}) = \lambda_{\text{MAX}}|(AA^T)^{-1}| = \frac{1}{\lambda_{\text{MIN}}(AA^T)}$$

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

$$c^2(A) = \frac{\lambda_{\text{MAX}}(A^T A)}{\lambda_{\text{MIN}}(A^T A)}; \quad (2.10.8)$$

in particolare, se  $A$  è simmetrica,

$$c(A) = \frac{\lambda_{\text{MAX}}(A)}{\lambda_{\text{MIN}}(A)}. \quad (2.10.9)$$

Da questa formula si vede che se  $A$  è prossima a essere singolare, il condizionamento numerico  $c(A)$  è sicuramente grande. Però una matrice prossima a essere singolare come  $\epsilon I$  con  $\epsilon \rightarrow 0$  ha condizionamento numerico uguale a uno. Chiaramente le matrici meglio condizionate sono quelle per cui  $A^T A = \alpha I$ . In questo caso infatti  $\lambda_{\text{MAX}}(A^T A) = \lambda_{\text{MIN}}(A^T A) = 1$  e  $c(A) = 1$ . Queste sono le matrici *ortogonali*<sup>7</sup>. Esse giocano un ruolo fondamentale nell'analisi numerica.

A titolo di esempio calcoliamo il condizionamento numerico del problema (2.10.4). La matrice  $A$  è simmetrica e si trova subito (approssimativamente)

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

da cui

$$c(A) = 4 \cdot 10^4 .$$

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

**Problem 2.2.**

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

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

*Generalizzare al caso in cui  $A$  non è simmetrica.* ◇

Esaminiamo adesso l'effetto degli errori di arrotondamento su  $A$ . Supponiamo  $\delta b = 0$ . Con semplici calcoli si ricava che la perturbazione  $\delta x$  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,

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{c(A) \frac{\|\delta A\|}{\|A\|}}{1 - c(A) \frac{\|\delta A\|}{\|A\|}} . \tag{2.10.10}$$

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

$$\frac{\|\delta x\|}{\|x\|} \leq c(A) \frac{\|\delta A\|}{\|A\|} \quad , \tag{2.10.11}$$

che è una disuguaglianza dello stesso tipo della (2.10.7). 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  $A^T$ . Si trova così

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

---

<sup>7</sup>Per coerenza, bisognerebbe dire che una matrice per cui  $AA^T = I$  è *ortonormale*.

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

$$c(A^T A) = \|A^T A\| \|(A^T A)^{-1}\| = \lambda_{\text{MAX}}(A^T A) / \lambda_{\text{MIN}}(A^T A) = c^2(A).$$

Ne segue che, anche per problemi  $Ax = b$  moderatamente ben condizionati,  $A^T Ax = b$  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)  $A^T Ax = A^T b$  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 del tipo

$$y = S\theta, \quad (2.10.12)$$

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 (2.10.12).

### 2.10.1 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.

**Theorem 2.4.** 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  $\{\sigma_1 \geq \dots \geq \sigma_n\}$ , detti valori singolari di  $A$ , tali che

$$A = U\Delta V^T \quad (2.10.13)$$

dove  $\Delta$  ha la struttura quasi diagonale:

$$\Delta = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}, \quad \Sigma = \text{diag} \{\sigma_1, \dots, \sigma_n\} \quad (2.10.14)$$

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

**Proof.** Siano  $[v_1, \dots, v_p]$ ,  $p$  autovettori ortonormali di  $A^T A$  di modo che

$$A^T A v_k = \sigma_k^2 v_k \quad k = 1, \dots, n$$

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

$$AA^T (A v_k) = \sigma_k^2 (A v_k) \quad k = 1, \dots, n$$



2.10. Aspetti numerici dei problemi ai minimi quadrati

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

$$u_k := \frac{1}{\sigma_k} Av_k \quad k = 1, \dots, n,$$

sono ortonormali. Infatti

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

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

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

per  $k, j \leq n$  e  $u_k^T Av_j = 0$  altrimenti. Queste relazioni sono equivalenti alla  $U^T AV = \Delta$  e quindi alla relazione (2.10.13).  $\square$

Possiamo così interpretare la formula (2.10.8) dicendo che *l'indice di condizionamento numerico di una matrice è il rapporto tra il suo massimo e il minimo valore singolare,*

$$c(A) = \frac{\sigma_1(A)}{\sigma_n(A)}. \tag{2.10.15}$$

La SVD fornisce la descrizione più completa che si conosca della struttura di una trasformazione lineare. Dalla (2.10.13) si ricava, eliminando i prodotti con i blocchi nulli di  $\Delta$ , la seguente *fattorizzazione a rango pieno di A*

$$A = [u_1, \dots, u_n] \Sigma [v_1, \dots, v_n]^T := U_n \Sigma V_n^T \tag{2.10.16}$$

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

$$U_n^T U_n = I_n = V_n^T V_n. \tag{2.10.17}$$

Ricordiamo che l'usuale norma (detta anche norma 2 o norma  $\ell^2$ ) di una matrice  $A \in \mathbb{R}^{m \times p}$  è definita dalla relazione

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

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.  $\|A\|_F^2 = \sum_{i,j} a_{i,j}^2 = \text{Tr } AA^T = \text{Tr } A^T A$ .

**Corollary 2.1.** *Lo spazio immagine e lo spazio nullo di A sono dati rispettivamente da:*

$$\text{Im}(A) = \text{Im}(U_n), \quad \text{Ker}(A) = \text{Im}([v_{n+1}, \dots, v_p]) \tag{2.10.18}$$

Inoltre,

$$\|A\|_2 = \|\Sigma\|_2 = \sigma_1, \quad \|A\|_F^2 = \|\Sigma\|_F^2 = \sigma_1^2 + \dots + \sigma_n^2 \quad (2.10.19)$$

La matrice

$$A_k := \sum_{i=1}^k \sigma_i u_i v_i^\top, \quad k \leq n \quad (2.10.20)$$

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

$$\min_{\text{rank}(B)=k} \|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1} \quad (2.10.21)$$

e inoltre

$$\min_{\text{rank}(B)=k} \|A - B\|_F^2 = \|A - A_k\|_F^2 = \sigma_{k+1}^2 + \dots + \sigma_n^2 \quad (2.10.22)$$

**Proof.** La (2.10.18) è un'ovvia conseguenza della (2.10.17). In particolare la seconda relazione scende dalla  $\text{Ker}(A) = \{x; V_n^\top x = 0\}$ . La dimostrazione delle altre proprietà si può trovare ad esempio nel testo [20, p. 584].  $\square$

## 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

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

sia minimo. Scriviamo  $P_n(x)$  come

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

dove  $s^\top(x) = [1 \ x \dots \ x^n]$ . Imponendo il principio di ortogonalità

$$f(x) - \sum_0^n \theta_i x^i \perp \text{span} \{1 \ x \dots \ x^n\} \quad (2.10.23)$$

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,

$$\begin{bmatrix} \langle 1, 1 \rangle & \langle 1, x \rangle & \dots & \langle 1, x^n \rangle \\ \vdots & & & \vdots \\ \langle x^n, 1 \rangle & \dots & \dots & \langle x^n, x^n \rangle \end{bmatrix} \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_n \end{bmatrix} = \begin{bmatrix} \langle 1, f \rangle \\ \vdots \\ \langle x^n, f \rangle \end{bmatrix}. \quad (2.10.24)$$

A conti fatti, si trova

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

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

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

si potrebbe semplicemente ottenere calcolando i prodotti scalari (cfr. la (2.10.23)) nella

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

e ricavandone immediatamente

$$\hat{\theta}_j = \langle f, p_j \rangle \quad , \quad j = 0, 1, \dots, n. \tag{2.10.25}$$

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

### Fattorizzazione QR

Supponiamo allora di voler ricavare la stima ai M.Q. di  $\theta$  partendo da  $N$  osservazioni  $y$  descritte dal modello

$$y = S\theta + \varepsilon \quad , \tag{2.10.26}$$

dove  $\varepsilon = \varepsilon(\theta)$  è il vettore degli errori di approssimazione delle misure,  $y$ , attraverso il modello  $S\theta$ . Le  $p$  colonne di  $S = [s_1, \dots, s_p]$  sono linearmente indipendenti, ma in generale non ortonormali. Se lo fossero,  $\langle s_i, s_j \rangle = s_i^\top s_j = \delta_{ij}$  e si avrebbe  $S^\top S = I$  per cui, in analogia all'esempio appena discusso, la stima  $\hat{\theta}$  si ricaverebbe immediatamente,

$$\hat{\theta} = S^\top y = \begin{bmatrix} \langle s_1, y \rangle \\ \vdots \\ \langle s_p, y \rangle \end{bmatrix}. \tag{2.10.27}$$

(Stiamo qui considerando minimi quadrati non pesati, ma questa semplificazione non costituisce affatto perdita di generalità). Notiamo che all'ultimo membro di

(2.10.27) compare il vettore delle prime  $p$  coordinate di  $y$  rispetto a una base ortonormale in  $\mathbb{R}^n$  del tipo  $\{s_1, s_2, \dots, s_p, \dots\}$ .

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

$$\begin{aligned} v_1 &= s_1 & , & & q_1 &:= v_1 / \|v_1\| \\ v_2 &= s_2 - \langle s_2, q_1 \rangle q_1 & , & & q_2 &:= v_2 / \|v_2\| \\ &\vdots & & & & \vdots \\ v_k &= s_k - \langle s_k, q_1 \rangle q_1 + \dots + \langle s_k, q_{k-1} \rangle q_{k-1} & , & & q_k &:= v_k / \|v_k\|. \end{aligned} \tag{2.10.28}$$

Dal punto di vista algebrico, le (2.10.28) forniscono una fattorizzazione di  $S$  di struttura assai particolare. Risolviamo le (2.10.28) rispetto a  $(s_1, \dots, s_p)$ . Si trova

$$\begin{aligned} s_1 &= \|v_1\| q_1 \\ s_2 &= \langle s_2, q_1 \rangle q_1 + \|v_2\| q_2 \\ &\vdots \\ s_p &= \langle s_p, q_1 \rangle q_1 + \dots + \langle s_p, q_{p-1} \rangle q_{p-1} + \|v_p\| q_p \quad , \end{aligned} \tag{2.10.29}$$

ovvero

$$[s_1, \dots, s_p] = [q_1, \dots, q_p] \begin{bmatrix} \|v_1\| & \langle s_2, q_1 \rangle & \dots & \langle s_p, q_1 \rangle \\ 0 & \|v_2\| & & \\ \vdots & 0 & & \\ \vdots & \vdots & & \\ 0 & 0 & & \|v_p\| \end{bmatrix} ; \tag{2.10.30}$$

questa relazione si può scrivere simbolicamente come

$$S = \bar{Q} \bar{R} \quad , \tag{2.10.31}$$

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

$$\begin{aligned} Q &= [\bar{Q} \mid q_{p+1} \dots q_n] \\ R &= \begin{bmatrix} \bar{R} \\ 0 \end{bmatrix} \quad , \end{aligned}$$

si vede che  $S$  si può anche scrivere come

$$S = QR \quad , \quad (2.10.32)$$

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 (2.10.28) forniscono un algoritmo ricorsivo per il calcolo di  $\bar{Q}$  ed  $\bar{R}$ . Per ottenere la (2.10.32) basta aggiungere a  $\bar{Q}$   $n-p$  colonne ortonormali (vedremo in seguito che questa operazione si può evitare).

Se moltiplichiamo i due membri della (2.10.26) per  $Q^\top$  si ottiene allora

$$Q^\top y = Q^\top S\theta + Q^\top \varepsilon \quad , \quad (2.10.33)$$

ovvero

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \bar{R} \\ 0 \end{bmatrix} \theta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \quad , \quad (2.10.34)$$

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

Notiamo subito che

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

e pertanto

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

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 coincide quindi con il *residuo di stima*  $\hat{\varepsilon} = y - Py$ . Il significato di  $\varepsilon_1$  ed  $\varepsilon_2$  verrà discusso più avanti.

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

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

e dalla (2.10.34) si vede, data l'ortogonalità di  $Q^\top$ , che

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

Da questa relazione segue immediatamente che

1)  $\hat{\theta}$  è soluzione di

$$\bar{R}\theta = y_1 \quad , \quad (2.10.35)$$

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

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

$$\|\hat{\varepsilon}\|^2 = \|y_2\|^2. \quad (2.10.36)$$

In altri termini, nel nuovo sistema di coordinate,  $\varepsilon_1$  rappresenta la parte dell'errore di approssimazione  $\varepsilon(\theta)$  che può essere *resa nulla* con la scelta  $\theta = \hat{\theta}$  ( $\hat{\theta}$  è la scelta di  $\theta$  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 (2.10.35) in cui  $\bar{R}$  è triangolare.

Notiamo che  $Q$  non entra esplicitamente nelle formule (2.10.35) e (2.10.36).

Se il modello (2.10.26) rappresenta delle misure affette da errore  $\varepsilon$  su cui si ha conoscenza probabilistica a priori, del tipo

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

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

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

Usando la fattorizzazione QR si vede subito che

$$\text{Var} \hat{\theta} = \sigma^2 (\bar{R}^T \bar{R})^{-1}. \quad (2.10.37)$$

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

$$[S \mid y] \quad (2.10.38)$$

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

$$\left[ \begin{array}{c|c} \bar{R} & y_1 \\ \hline 0 & y_2 \end{array} \right]. \quad (2.10.39)$$

Giunti a questo punto, ovviamente il grosso del lavoro è stato fatto perché rimane solo da risolvere il sistema (2.10.35) 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 (2.10.38) alla forma (2.10.39).

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 [31]. Qui sotto ne descriveremo uno particolarmente semplice e di uso generale.

**Definizione 2.3.** Una matrice di riflessione elementare (o matrice di Householder) è una matrice  $n \times n$  del tipo

$$H(v) = I - 2 \frac{vv^T}{\|v\|^2} \quad , \quad (2.10.40)$$

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  $e_1 = [1, 0, \dots, 0]'$ . Dato che  $H(v)$  è ortogonale, dovrà necessariamente essere

$$H(v)x = I \|x\| e_1. \quad (2.10.41)$$

La scelta di  $v$  per arrivare alla (2.10.41) è suggerita dalla geometria della trasformazione di Figura 5.2.

### Figura 5.2

Dato  $x$ , si tratta di trovare il piano bisettore  $S$  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

$$v = x + \|x\| e_1 \quad (2.10.42)$$

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

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

Mediante l'uso di matrici di Householder si può triangolarizzare  $S$  in  $p - 1$  passi e ridursi alla tabella (2.10.39) partendo dalla tabella (2.10.38). Denotiamo quest'ultima con il simbolo

$$S_1 = [s_1, \dots, s_p y].$$

Prendendo

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

si ha

$$H(v) S_1 = \begin{bmatrix} -\|s_1\| & s'_{12} & \dots & s'_{1p} & y'_1 \\ 0 & \boxed{s'_{22} & & s'_{2p} & y'_2} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & s'_{n2} & & s'_{np} & y'_n \end{bmatrix}. \quad (2.10.43)$$

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

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

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

$$H(v_2) S_2 = \begin{bmatrix} -\|s'_2\| & s''_{23} & \dots & s''_{2p} & y''_2 \\ 0 & \boxed{s''_{33} & & \vdots} \\ \vdots & \vdots & & \vdots \\ 0 & s''_{n3} & & s''_{np} & y''_n \end{bmatrix}. \quad (2.10.44)$$

È chiaro che trasformando in modo analogo la prima colonna della matrice  $S_3$   $(n - 2) \times (p - 2)$  e via via  $S_4, \dots, S_{p-1}$  si arriva a una struttura triangolare del tipo (2.10.39). Si può anche immaginare di operare la trasformazione (2.10.44) mediante una matrice che è  $n \times n$  anziché  $(n - 1) \times (n - 1)$  come la  $H(v_2)$ . Se poniamo infatti

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

si controlla subito che

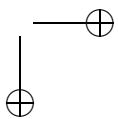
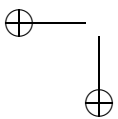
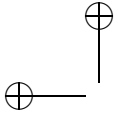
$$\tilde{H}(v_2) (H(v_1) S_1) = \begin{bmatrix} -\|s_1\| & s'_{12} & \dots & s'_{1p} & y'_1 \\ 0 & -\|s_2\| & & s''_{2p} & y''_2 \\ \vdots & 0 & & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & s''_{np} & y''_n \end{bmatrix},$$



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

$$Q = \tilde{H}(v_{p-1}) \tilde{H}(v_{p-2}) \dots \tilde{H}(v_2) H(v_1). \quad (2.10.45)$$

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 [40].



## CHAPTER 3

# MODELLI DINAMICI PER L'IDENTIFICAZIONE

In questo capitolo passeremo in rassegna alcune classi di modelli statistici parametrici che vengono usati spesso nelle applicazioni e studieremo le loro proprietà dal punto di vista della stima parametrica.

### 3.1 Modelli statistici lineari per processi del secondo ordine

Inizieremo questa sezione chiedendoci quali siano i modelli probabilistici più generali che possono descrivere i segnali osservati. Ricordiamo che un *processo stocastico del secondo ordine* è la classe di equivalenza dei processi (definiti sullo stesso spazio di probabilità) che hanno la stessa media e la stessa funzione di covarianza. Un processo del second'ordine è quindi completamente descrivibile assegnando la sua media e la sua funzione di covarianza. Queste due funzioni sono chiamate le *statistiche del secondo ordine* del processo. Come è ben noto, nel caso Gaussiano esse individuano completamente tutte le distribuzioni finito-dimensionali (i.e. la legge di probabilità) del processo.

L'assunto fondamentale è che i dati osservati possano essere assimilati a tratti, ovviamente finiti, di traiettorie (in inglese *sample paths*) di processi stocastici del second'ordine, debolmente stazionari<sup>8</sup>. Questo assunto equivale a dire che ci si limiterà a descrivere i segnali in gioco solo mediante la loro media (costante) e le funzioni di auto e mutua covarianza; i.e. mediante i primi due momenti della loro legge di probabilità congiunta. L'ipotesi di stazionarietà debole che faremo in questo capitolo, è molto blanda e dovrà essere rafforzata per permettere lo studio di certe proprietà asintotiche degli stimatori. Essa verrà comunque discussa in maggior dettaglio nel capitolo dedicato all'ergodicità. Considereremo solo segnali (e processi) a tempo discreto, denoteremo la variabile temporale adimensionale con  $t \in \mathbb{Z}$ , e senza perdita di generalità, assumeremo che tutti i processi in gioco abbiano media nulla. Useremo i concetti e il formalismo della teoria geometrica

<sup>8</sup>Nel seguito la stazionarietà in senso debole verrà chiamata semplicemente stazionarietà.

dei processi del second'ordine (spazi di Hilbert etc.). Per una trattazione esaustiva di questo argomento vedere [54, 44].

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.

Per semplicità supporremo nel seguito che i processi  $\mathbf{y}$  e  $\mathbf{u}$  siano scalari. L'estensione al caso di **ingressi multidimensionali** è di interesse in molte applicazioni ma è facile e verrà lasciata al lettore.

La generalizzazione al caso di *ingressi e uscite multidimensionali* presenta invece difficoltà. In questo caso conviene usare **modelli di stato**.

Faremo l'ipotesi che i dati osservati siano assimilati a tratti finiti di traiettorie (in inglese *sample paths*) di processi stocastici congiuntamente stazionari (in senso debole) e ci limiteremo a cercare *modelli che descrivono solo le statistiche del (primo e) secondo ordine di questi processi*.

È importante realizzare che i processi del second'ordine possono essere descritti da **modelli lineari**. Infatti mediante modelli lineari si possono descrivere completamente la media (normalmente supposta nulla e) le funzioni di auto e mutua covarianza, ovvero gli spettri congiunti dei segnali in gioco; i.e. i primi due momenti della loro legge di probabilità.

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

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

$$\mathbf{y}(t) = F(z)\mathbf{u}(t) + \mathbf{v}(t), \quad t \in \mathbb{Z} \quad (3.1.1)$$

dove  $F(z)$  è una funzione di trasferimento causale (non necessariamente razionale)<sup>9</sup> e  $\mathbf{v}$  è un processo stazionario, detto *errore di modellizzazione*, scorrelato dal passato di  $\mathbf{u}$ , ovvero

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

Notiamo ora che si può scrivere una decomposizione perfettamente analoga alla (3.1.1) anche per la variabile  $\mathbf{u}$ ,

$$\mathbf{u}(t) = H(z)\mathbf{y}(t) + \mathbf{r}(t), \quad t \in \mathbb{Z} \quad (3.1.2)$$

dove  $H(z)$  è una funzione di trasferimento causale (non necessariamente razionale) e il futuro di  $\mathbf{r}$  è scorrelato dalla storia passata di  $\mathbf{y}$ . L'interconnessione dei due modelli (3.1.1) e (3.1.2) produce uno *schema a retroazione* del tipo noto nei controlli automatici.

<sup>9</sup>Nel caso di processi puramente non deterministici (p.n.d), il primo termine in (3.1.1) è il filtro di Wiener causale basato sul passato di  $\mathbf{u}$ . Nel caso generale al termine  $F(z)\mathbf{u}(t)$  può essere dato un significato analogo, per i dettagli matematici si può ad esempio vedere il testo di Rozanov [54].

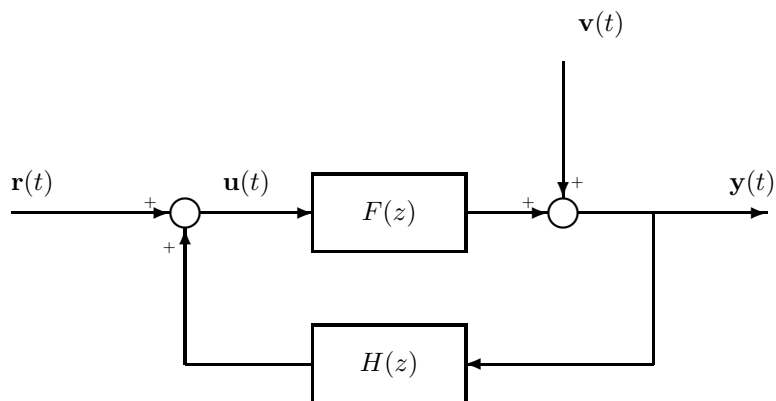


Figure 3.1.1. Modello congiunto dei segnali  $y$  e  $u$ .

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

Una caratteristica poco piacevole che deriva dal modo in cui è stata ricavata la coppia di equazioni (3.1.1) e (3.1.2) è che i segnali  $v$  e  $r$ , che intuitivamente vorremmo essere degli ingressi esogeni generati da meccanismi “esterni” al sistema, risultano essere in generale correlati. Per questo motivo si preferisce descrivere la coppia  $y, u$  per mezzo di schemi a retroazione in cui si impone che gli ingressi  $v$  e  $r$  siano scorrelati. Questi modelli si chiamano *modelli a retroazione* della coppia  $y, u$ . Noi qui daremo per nota la teoria dei modelli a retroazione e in particolare il concetto di *retroazione tra processi stocastici*. La teoria si può trovare esposta ad esempio in [44, Cap. 7].

In un modello a retroazione le funzioni di trasferimento  $F(z)$  e  $H(z)$  sono ancora causali ma non rappresentano necessariamente sistemi lineari stabili. Chi deve essere stabile (più esattamente *internamente stabile*), per garantire la stazionarietà congiunta dei processi  $y, u$ , è il sistema a controreazione formato dall’interconnessione di (3.1.1) e (3.1.2), [44, Cap. 7]. Si dimostra che se (e solo se)  $u$  e  $v$  sono completamente incorrelati, ovvero vale la

$$\mathbb{E} v(t)u(s) = 0 \quad t, s \in \mathbb{Z}. \quad (3.1.3)$$

si ha *assenza di reazione da  $y$  a  $u$* , nel qual caso  $H(z) \equiv 0$ . In assenza di retroazione la funzione di trasferimento  $F(z)$  è analitica in  $\{|z| \geq 1\}$ , ovvero, il corrispondente sistema lineare è  $\ell^2$ -stabile.

Ci restringeremo ora al caso in cui i processi  $y$  e  $u$  sono entrambi *puramente non deterministici* (*p.n.d.*). In questo caso le statistiche del second’ordine congiunte dei due processi sono descritte completamente dalla matrice densità spettrale congiunta

$$S(z) = \begin{bmatrix} S_y(z) & S_{yu}(z) \\ S_{uy}(z) & S_u(z) \end{bmatrix} \quad (3.1.4)$$

che supporremo definita positiva sul cerchio unità. In realtà l'ipotesi che  $\mathbf{u}$  sia p.n.d. non è strettamente necessaria e in pratica è anche poco desiderabile perchè si possono dare casi in cui il segnale di ingresso ha componenti puramente oscillatorie o comunque è un segnale deterministico. Manterremo l'ipotesi per semplicità di esposizione in questa sezione, riservandoci di considerare la presenza di componenti deterministiche a tempo debito.

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

$$\mathbf{v}(t) = G(z)\mathbf{e}(t), \quad t \in \mathbb{Z} \quad (3.1.5)$$

dove  $G(z)$  è una funzione di trasferimento (non necessariamente razionale) a fase minima che prenderemo sempre normalizzata all'infinito,  $G(\infty) = 1$ . Il processo  $\mathbf{e}$  è il *processo innovazione* (non normalizzata), un processo bianco di varianza  $\lambda^2$  che ha il significato di errore di predizione di un passo di  $\mathbf{v}(t)$  basato sulla storia passata del processo all'istante  $t - 1$ . Combinando (3.1.1) e (3.1.5) si ottiene

$$\mathbf{y}(t) = F(z)\mathbf{u}(t) + G(z)\mathbf{e}(t), \quad t \in \mathbb{Z} \quad (3.1.6)$$

Si mostra che, senza perdita di generalità, si può sempre imporre che  $F(z)$  sia strettamente causale, ovvero che  $F(\infty) = 0$ . Questo è il modello generale che descrive la dinamica "in catena diretta" di due processi stazionari del second'ordine 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 probabilità congiunta, o come nel caso in esame, mediante la densità spettrale congiunta. È importante notare che nel caso di presenza di reazione il modello (3.1.6) da solo non individua univocamente nemmeno la covarianza o lo spettro di  $\mathbf{y}$ , dato che esso non specifica come  $\mathbf{u}$  ed  $\mathbf{e}$  sono correlati. Nel caso di presenza di reazione, il modello a retroazione completo da considerare è quello *congiunto*, comprendente anche la descrizione (3.1.2) del canale di retroazione dove ora  $\mathbf{r}$  è un processo stazionario completamente scorrelato da  $\mathbf{v}$  (e quindi anche da  $\mathbf{e}$ ). Possiamo rappresentare anche  $\mathbf{r}$  mediante la sua rappresentazione di innovazione

$$\mathbf{r}(t) = K(z)\mathbf{w}(t) \quad (3.1.7)$$

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

Naturalmente, per la stazionarietà congiunta dei processi di ingresso-uscita  $[\mathbf{y} \ \mathbf{u}]^T$ , l'interconnessione a retroazione di figura dev'essere *internamente stabile*, ovvero la matrice di trasferimento in catena chiusa,

$$T(z) = \begin{bmatrix} G & FK \\ \frac{1-FH}{HG} & \frac{1-FH}{K} \\ \frac{1-FH}{1-FH} & \frac{1-FH}{1-FH} \end{bmatrix} \quad (3.1.8)$$

che trasforma il processo congiunto  $[e \ w]^T$  in  $[y \ u]^T$  dovrà essere analitica in  $\{|z| \geq 1\}$ . In altre parole  $T(z)$  dovrà essere un *fattore spettrale analitico* dello spettro congiunto. Come si è visto in corsi precedenti, anche col vincolo dell'analiticità i fattori spettrali non sono mai unici ed è quindi evidente che ci saranno molte (in generale infinite) rappresentazioni a retroazione della coppia  $(y, u)$ . In effetti queste rappresentazioni sono in corrispondenza biunivoca con la classe dei fattori spettrali quadrati  $T(z)$ , dello spettro congiunto di  $(y, u)$  che sono analitici e "normalizzati a blocchi all'infinito". Si veda [44, Cap. 7] per una discussione approfondita di questi concetti. Si conviene allora di prendere come fattore spettrale rappresentativo quello a fase minima (normalizzato a blocchi all'infinito). A questo fattore spettrale corrisponde un unico *modello a retroazione d'innovazione*. Il risultato seguente è standard, si veda ad esempio [44, cap. 7.4], oppure [62, p. 195], ma data la sua importanza ne daremo ugualmente dimostrazione.

**Theorem 3.1.** *Nel modello a retroazione d'innovazione  $F(z)$  e  $G(z)$  soddisfano alle seguenti condizioni*

1. *C'è almeno un ritardo in  $F$  e  $G$  è normalizzata all'infinito, i.e.  $F(\infty) = 0$  e  $G(\infty) = 1$ .*
2.  *$G(z)^{-1}$  e  $G(z)^{-1}F(z)$  sono analitiche in  $\{|z| \geq 1\}$ .*

*Viceversa, se queste condizioni sono soddisfatte e  $u$  è generato da una reazione causale del tipo (3.1.2) in cui  $r$  è completamente scorrelato da  $e$ , il processo bianco  $e$  è proprio l'innovazione di  $y$ , ovvero  $e(t)$  è l'errore di predizione di un passo di  $y(t)$  basato sul passato congiunto di  $u$  e  $y$  all'istante  $t - 1$ .*

*Posto  $G(z) = 1 + z^{-1}G_1(z)$ , il predittore lineare a minima varianza d'errore di  $y(t)$  basato sulla storia passata congiunta  $(y^{t-1}, u^{t-1})$  è dato dalla formula*

$$\hat{y}(t | t - 1) = G(z)^{-1}F_1(z)u(t - 1) + G(z)^{-1}G_1(z)y(t - 1). \quad (3.1.9)$$

**Proof.** Diamo solo la dimostrazione della seconda parte che è sostanzialmente basata sul ragionamento che porta al predittore di Wiener (3.1.9).

Assumiamo che sia disponibile un modello generale (3.1.6) in cui potrebbe esserci reazione, che soddisfa alle condizioni descritte nell'enunciato.

Facciamo vedere che allora  $e(t)$  è scorrelato dalla storia passata congiunta  $(y^{t-1}, u^{t-1}) := \{y(s), u(s); s < t\}$  dei processi  $y$  e  $u$ . Di fatti, dato che la matrice di trasferimento  $T(z)$  deve rappresentare una trasformazione ingresso uscita  $[e \ w]^T \rightarrow [y \ u]^T$  stabile e causale ( $T(z)$  è analitica in  $\{|z| \geq 1\}$ ), si deve avere

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

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

Scriviamo allora il modello (3.1.6) nella forma

$$y(t) = F(z)u(t) + [G(z) - 1]e(t) + e(t)$$

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

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

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  $\mathbf{u}$  all'istante  $t-1$ , questa funzione è proprio il predittore cercato.  $\square$

Notare che dall'espressione (\*) si ricava il significato della causalità delle funzioni di trasferimento  $G(z)^{-1}$  e  $G(z)^{-1}F(z)$ . Essa serve a garantire che  $\mathbf{e}(t) \in H(\mathbf{y}^t, \mathbf{u}^t)$ .

Quando valgono le condizioni del Teorema 3.1, chiameremo il modello (3.1.6) *modello d'innovazione*.

**Problem 3.1.** Nella dimostrazione non abbiamo usato la rappresentazione d'innovazione (3.1.7) del processo  $\mathbf{r}$ . Ripercorrere il ragionamento e mostrare che il risultato continua a valere anche se  $\mathbf{r}$  non è p.n.d.

### 3.2 Modelli parametrici e identificabilità in assenza di retroazione

È dimostrato in letteratura (e il fatto è stato implicitamente usato nella sezione precedente) che una qualunque coppia di processi stazionari del second'ordine può essere descritta mediante un unico modello a reazione d'innovazione, costituito da una coppia di equazioni simboliche del tipo (3.1.6), (3.1.2), soggetta alle condizioni descritte nell'enunciato del teorema 3.1.

Ne scende che, in assenza di informazioni specifiche sul meccanismo "vero" che ha generato i dati, una classe naturale di modelli dinamici che si può usare per approssimare il modello "vero" dei processi  $\mathbf{y}$  e  $\mathbf{u}$ , è la classe dei modelli lineari che hanno la struttura (3.1.6), assumendo  $F(z)$  e  $G(z)$  *funzioni razionali di  $z$* . Come è ben noto, le funzioni razionali hanno proprietà "universali" di approssimazione di classi di funzioni molto generali. Inoltre la razionalità comporta algoritmi di stima e realizzazione di dimensione finita che permettono un'organizzazione efficiente dei calcoli. Infine, i modelli razionali possono essere raggruppati in classi omogenee in cui, una volta fissati dei *parametri di struttura* tutte le funzioni di trasferimento  $F$  e  $G$  hanno una stessa "complessità" fissata; i.e. dipendono da uno stesso numero finito di parametri scalari (ad esempio, una volta fissati i gradi relativi, i coefficienti dei polinomi a numeratore e denominatore) e il problema di stima che si ha in vista riguarderà un parametro vettoriale di dimensione fissa.

Notiamo che ci si preoccupa di parametrizzare *solo il modello in catena di azione diretta* (3.1.6), dato che per ipotesi, non siamo interessati a identificare un modello per la variabile di ingresso  $\mathbf{u}$ .



In generale si può pensare che i parametri da cui dipendono le due funzioni razionali  $F$  e  $G$  siano i coefficienti dei rispettivi polinomi a numeratore e a denominatore, o eventualmente, alcune loro combinazioni algebriche. Si possono però dare anche casi in cui si ha a priori una conoscenza parziale di questi parametri. Lascieremo per il momento la questione nel vago e faremo riferimento di una classe di modelli parametrici del tipo

$$\mathbf{y}(t) = F_\theta(z)\mathbf{u}(t) + G_\theta(z)\mathbf{e}(t), \quad \theta \in \Theta \subset \mathbb{R}^p \quad (3.2.1)$$

dove  $F_\theta(z)$  e  $G_\theta(z)$  sono funzioni di trasferimento razionali che dipendono da un parametro vettoriale incognito  $\theta$ . Non sarà necessario per il momento specificare esattamente come  $F(z)$  e  $G(z)$  vi dipendano. Postuleremo solo che questa dipendenza da  $\theta$  sia regolare (continua e derivabile quante volte serve).

In questa sezione ci proponiamo di studiare l'identificabilità della classe di modelli (3.2.1) nell'ipotesi 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 densità spettrale di  $\mathbf{y}$  indotta dal modello (3.2.1) si può scrivere esplicitamente come

$$S_{\mathbf{y}}(z) = F(z)S_{\mathbf{u}}(z)F(1/z) + \lambda^2 G(z)G(1/z). \quad (3.2.2)$$

Come si vede,  $S_{\mathbf{y}}(z)$  è parametrizzata dalla densità dell'ingresso,  $S_{\mathbf{u}}(z)$ , e quindi dipende, come si dice convezionalmente, dalla *condizione sperimentale*. Quando i dati sono raccolti durante il "normale funzionamento" dell'impianto, si pensa normalmente che  $S_{\mathbf{u}}$  sia imposta dall'esterno. 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'identificazione.

In quest'ultimo 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 a rigore non ammettono densità spettrale di potenza. In questo caso l'espressione (3.2.2) dovrebbe essere riscritta usando le distribuzioni spettrali, in particolare sostituendo a  $S_{\mathbf{u}}(z)$  la relativa distribuzione spettrale di potenza  $\hat{F}_{\mathbf{u}}(z)$ . Noi useremo a questo scopo la notazione ingegneristica convenzionale in cui una densità può comprendere "funzioni"  $\delta$  di Dirac.

Per poter applicare al contesto attuale la nozione naturale di identificabilità introdotta nel Capitolo 1, sezione 1.2.2, dobbiamo sostituire i modelli probabilistici in senso stretto (distribuzioni di probabilità) con le statistiche del secondo ordine<sup>10</sup>. Per il modello parametrico (3.2.1) faremo quindi riferimento alla famiglia parametrica di spettri

$$\begin{aligned} S_{\mathbf{y}}(z; \theta) &= F_\theta(z)S_{\mathbf{u}}(z)F_\theta(1/z) + \lambda^2 G_\theta(z)G_\theta(1/z), \\ S_{\mathbf{y}\mathbf{u}}(z; \theta) &= F_\theta(z)S_{\mathbf{u}}(z) \quad \theta \in \Theta \subset \mathbb{R}^p \end{aligned} \quad (3.2.3)$$

con la solita avvertenza di sostituire all'occorrenza densità spettrali con le relative distribuzioni. L'identificabilità del modello (3.1.6) deve intuitivamente corrispon-

<sup>10</sup>Dovremmo allora, a rigore parlare di *indistinguibilità e identificabilità del second'ordine, o in senso debole*. Dato che non vale la pena di appesantire ulteriormente la terminologia, non insisteremo oltre su questo punto.

dere a una parametrizzazione non ridondante dello spettro congiunto (3.2.3). 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  $\theta$  sono solo  $S_{\mathbf{y}}$  e lo spettro incrociato  $S_{\mathbf{y}\mathbf{u}}$ .

**Definition 3.1.** *Si assuma assenza di reazione da  $\mathbf{y}$  a  $\mathbf{u}$ . Il modello (3.2.1) 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  $\theta \mapsto [S_{\mathbf{y}}(\cdot; \theta), S_{\mathbf{y}\mathbf{u}}(\cdot; \theta)]$  è iniettiva in  $\Theta$ , ovvero se l'uguaglianza*

$$[S_{\mathbf{y}}(z; \theta_1), S_{\mathbf{y}\mathbf{u}}(z; \theta_1)] = [S_{\mathbf{y}}(z; \theta_2), S_{\mathbf{y}\mathbf{u}}(z; \theta_2)], \quad \forall z \in \mathbb{C} \quad (3.2.4)$$

implica  $\theta_1 = \theta_2$ . Se si ha iniettività locale in un intorno di  $\theta_0$ , si parla di identificabilità locale in  $\theta_0$ .

Dalla definizione si trae facilmente il seguente criterio.

**Proposition 3.1.** *Nel modello (3.2.1), senza reazione, due vettori di parametri  $\theta_1$  e  $\theta_2$  sono indistinguibili nella condizione sperimentale descritta da  $S_{\mathbf{u}}$  (o dalla distribuzione spettrale  $d\hat{F}_{\mathbf{u}}$ ), se e solo se, per ogni  $z = e^{j\omega}$*

$$[F_{\theta_1}(z) - F_{\theta_2}(z)] S_{\mathbf{u}}(z) = 0 \quad (3.2.5)$$

$$G_{\theta_1}(z) - G_{\theta_2}(z) = 0 \quad (3.2.6)$$

Equivalentemente, il modello (3.2.1) è globalment e identificabile se l'unica soluzione delle equazioni (3.2.5), (3.2.6) è  $\theta_1 = \theta_2$ .

**Proof.** Dato che

$$F_{\theta}(z)S_{\mathbf{u}}(z) = S_{\mathbf{y}\mathbf{u}}(z; \theta)$$

evidentemente la (3.2.5) è equivalente all'uguaglianza degli spettri  $S_{\mathbf{y}\mathbf{u}}(z; \theta_1)$  e  $S_{\mathbf{y}\mathbf{u}}(z; \theta_2)$ . Notiamo poi che se vale la (3.2.6) 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 (3.2.3) che (3.2.5) e (3.2.6) implicano  $S_{\mathbf{y}}(z; \theta_1) = S_{\mathbf{y}}(z; \theta_2)$ .

Viceversa, facciamo vedere che le (3.2.4) implicano le (3.2.5) e (3.2.6). Per quanto appena visto l'uguaglianza degli spettri incrociati implica la (3.2.5) 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.  $\square$

Il fatto che la (3.2.5) implichi  $F_{\theta_1}(\cdot) \equiv F_{\theta_2}(\cdot)$  per tutte le funzioni di trasferimento di una certa classe  $\mathcal{F}$ , nel nostro caso una classe parametrica di funzioni razionali  $\mathcal{F} = \{F_{\theta}(\cdot); \theta \in \Theta\}$ , si esprime dicendo che l'ingresso  $\mathbf{u}$  è sufficientemente (o persistentemente) eccitante per la classe  $\mathcal{F}$ .

Una nozione di identificabilità che talvolta in letteratura viene confusa con quella precedente è l'identificabilità *a priori*.

**Definition 3.2.** Il modello (3.2.1) è identificabile a priori (globalmente) se la mappa  $\theta \mapsto [F_\theta(\cdot), G_\theta(\cdot)]$  è iniettiva in  $\Theta$ , ovvero

$$[F_{\theta_1}(z), G_{\theta_1}(z)] = [F_{\theta_2}(z), G_{\theta_2}(z)] \quad \forall z \in \mathbb{C} \Rightarrow \theta_1 = \theta_2. \quad (3.2.7)$$

Se si ha iniettività locale in un intorno di  $\theta_0$ , si parla di identificabilità a priori locale in  $\theta_0$ .

Come si vede, la nozione di identificabilità 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'identificabilità a priori è quindi un fatto puramente algebrico.

Per apprezzare la diversità dei due concetti basta analizzare la mappa che descrive la dipendenza dello spettro congiunto dal parametro  $\theta$ . Questa mappa è composta di due componenti:

$$\theta \mapsto [F_\theta(\cdot), G_\theta(\cdot)] \mapsto [S_{\mathbf{y}}(\cdot; \theta), S_{\mathbf{y}\mathbf{u}}(\cdot; \theta)]$$

Dato che si tratta di una mappa ottenuta per composizione delle due applicazioni  $\theta \mapsto [F_\theta(\cdot), G_\theta(\cdot)]$  e  $[F_\theta(\cdot), G_\theta(\cdot)] \mapsto [S_{\mathbf{y}}(\cdot; \theta), S_{\mathbf{y}\mathbf{u}}(\cdot; \theta)]$  (quest'ultima dipendente dallo spettro dell'ingresso), per l'identificabilità si deve richiedere l'iniettività di entrambi; il che si può riassumere nel modo seguente.

**Proposition 3.2.** L'identificabilità (in assenza di reazione) è equivalente all'identificabilità a priori e alla sufficiente eccitazione dell'ingresso.

È ovvio dalla definizione 3.1, che l'identificabilità a priori è quindi solo condizione *necessaria* per l'identificabilità. In ogni caso, se il modello (3.2.1) non fosse identificabile a priori, non sarebbe nemmeno possibile distinguere i parametri dello spettro congiunto  $[S_{\mathbf{y}}(\cdot; \theta), S_{\mathbf{y}\mathbf{u}}(\cdot; \theta)]$ , per nessuna condizione sperimentale.

Nel caso speciale in cui l'ingresso è rumore bianco, la densità spettrale  $S_{\mathbf{u}}$  è una costante positiva e quindi un *ingresso bianco* è *sufficientemente eccitante* per una qualunque classe di funzioni di trasferimento. Dalle (3.2.5), (3.2.6) si vede immediatamente che in questo caso l'identificabilità a priori è necessaria e sufficiente per l'identificabilità.

Rimarchiamo comunque che in generale un modello identificabile a priori potrebbe benissimo non essere identificabile per qualche condizione sperimentale. Un caso estremo si presenta quando l'ingresso è una funzione costante (al limite nulla) per cui  $S_{\mathbf{u}}(z)$  è zero quasi ovunque. Se viceversa il processo di ingresso ha una componente p.n.d. non nulla (per la definizione di componente p.n.d. di un processo stazionario vedere la decomposizione di Wold, ad esempio in [44, Teor. 5.2]), il suo spettro ha una componente assolutamente continua che è positiva quasi ovunque.

**Proposition 3.3.** *Se il processo di ingresso ha una componente p.n.d. non nulla, è sufficientemente eccitante per qualunque  $\mathcal{F}$  e l'identificabilità è equivalente all'identificabilità a priori.*

Nel caso in cui il segnale di ingresso sia p.d., o addirittura deterministico, l'identificabilità dipende dalla classe parametrica di funzioni di trasferimento  $F_\theta(z)$  che costituiscono il modello. Bisogna richiedere, oltre all'identificabilità a priori, che il segnale di ingresso soddisfi ad alcune condizioni note come condizioni di **persistente eccitazione**. Queste condizioni verranno discusse nella sezione 3.3.

### Identificabilità a priori di modelli razionali

Un'analisi approfondita del concetto di identificabilità a priori porta a qualche sorpresa. In effetti, a meno che il modello non sia parametrizzato linearmente (sia cioè una funzione lineare di  $\theta$ ) si scopre che *l'identificabilità a priori è una proprietà essenzialmente locale*. Consideriamo ad esempio il caso di una funzione di trasferimento scalare razionale del prim'ordine

$$F_\theta(z) = \frac{1 + cz^{-1}}{1 + az^{-1}}; \quad \theta = [a \ c]^\top \quad (3.2.8)$$

dove per semplicità non considereremo i vincoli di stabilità sui parametri  $a$  e  $c$  per cui penseremo  $\theta$  variabile su tutto  $\mathbf{R}^2$ . Ora è ben evidente che

$$F_{\theta_1}(z) = F_{\theta_2}(z) \quad \forall z \Rightarrow \theta_1 = \theta_2$$

eccezion fatta per quei  $\theta$  che appartengono all'insieme

$$\Theta_0 := \{\theta; a = c\} \quad (3.2.9)$$

dato che se  $\theta \in \Theta_0$  il numeratore e il denominatore si cancellano e si ha  $F_\theta(z) = 1$  identicamente per cui, qualunque sia il valore (comune) di  $a$  e  $c$  la funzione  $F_\theta(z)$  è la stessa. In termini un pò più formali, si può dire che la classe parametrica di funzioni di trasferimento (3.2.8) è identificabile localmente in tutti i punti dell'insieme  $\mathbf{R}^2 - \Theta_0$ . In sostanza, è identificabile localmente in tutti i punti dell'insieme dei valori del parametro  $\theta$  per cui *non si hanno cancellazioni tra numeratore e denominatore*.

Questo esempio è rappresentativo del caso generale.

**Proposition 3.4.** *Una famiglia di funzioni razionali parametrizzata attraverso i coefficienti dei polinomi a numeratore e a denominatore, almeno uno dei quali è supposto monico, è localmente identificabile a priori in tutti i valori del parametro che non corrispondono a fattori comuni (e quindi a cancellazioni) tra i due polinomi.*

Dato che l'insieme dei valori del parametro per cui si hanno cancellazioni è sempre un sottoinsieme di misura nulla di  $\Theta$ , si parla di identificabilità *quasi ovunque* o meglio di identificabilità *generica*.

### 3.3 Persistente Eccitazione

La nozione di segnale persistentemente eccitante riguarda segnali “deterministici”, per i quali non è necessariamente data una descrizione statistica.

**Definition 3.3.** Un segnale deterministico  $u = \{u(t); t \in \mathbb{Z}\}$  è detto stazionario del secondo ordine<sup>11</sup> se il limite

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N u(t+\tau)u(t) := r(\tau) \quad (3.3.1)$$

esiste per ogni  $\tau \geq 0$ .

Un segnale stazionario del secondo ordine,  $u$ , è persistentemente eccitante di ordine (almeno)  $n$  se la matrice (di Toeplitz)

$$\mathbf{R}_n := \begin{bmatrix} r(0) & r(1) & \dots & r(n-1) \\ r(1) & r(0) & r(1) & \dots & r(n-2) \\ \vdots & & & \vdots & \\ r(n-1) & r(n-2) & \dots & & r(0) \end{bmatrix} \quad (3.3.2)$$

è definita positiva.

[PROPRIETÀ DI  $r(\tau)$ ]

Notiamo che  $\mathbf{R}_n$  è sempre almeno semidefinita perchè, per un arbitrario polinomio  $p(z^{-1}) := \sum_{k=0}^{n-1} p_k z^{-k}$  nell'operatore di ritardo  $z^{-1}$ , si ha

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N [p(z^{-1})u(t)]^2 = \sum_{k,j=0}^n p_k r(k-j) p_j = p^\top \mathbf{R}_n p \geq 0$$

dove abbiamo denotato col simbolo  $p := [p_0 p_1 \dots p_{n-1}]^\top$  il vettore dei coefficienti di  $p(z^{-1})$ .

Come già notato da Norbert Wiener (nel 1930! [69, 70]), la funzione  $\tau \rightarrow r(\tau)$ ;  $\tau \in \mathbb{Z}$  è quindi una funzione di tipo positivo che si può pertanto assimilare alla funzione di correlazione di un processo stazionario. Ad essa si applica quindi il teorema di Herglotz che stabilisce l'esistenza di una funzione,  $F_u(e^{j\omega})$ , monotona non decrescente sull'intervallo  $[-\pi, \pi]$ , la distribuzione spettrale di potenza di  $u$ , per cui

$$r(\tau) = \int_{-\pi}^{\pi} e^{j\omega\tau} dF_u(e^{j\omega}).$$

Per i segnali deterministici, stazionari del secondo ordine, possiamo quindi parlare di spettro di potenza. Convenzionalmente si usa parlare di “densità spettrale” anche se raramente la  $F_u$  è assolutamente continua. In generale la densità spettrale di un segnale stazionario del secondo ordine contiene impulsi o righe spettrali come si usa dire comunemente.

<sup>11</sup>Questo concetto verrà ripreso nel Capitolo 4.

**Theorem 3.2.** *Un segnale stazionario del second'ordine  $u$  è persistentemente eccitante di ordine esattamente  $n$  se e solo se gli unici punti di crescita della sua distribuzione spettrale  $F_u(e^{j\omega})$  sono  $n$  salti, ovvero la sua densità spettrale consiste esattamente di  $n$  righe spettrali alle frequenze  $\{\omega_1, \omega_2, \dots, \omega_n\}$ <sup>12</sup>, nell'intervallo  $(-\pi, \pi)$ .*

*Proof.* Siano

$$p(z^{-1}) := \sum_{k=0}^{n-1} p_k z^{-k} \quad q(z^{-1}) := \sum_{k=0}^n q_k z^{-k}$$

due polinomi di grado effettivo  $n - 1$  ed  $n$  e si denotino con  $p \in \mathbb{R}^n$  e  $q \in \mathbb{R}^{n+1}$  i rispettivi vettori dei coefficienti. Notiamo che se lo spettro di  $u$  è come descritto nell'enunciato,

$$p^\top \mathbf{R}_n p = \int_{-\pi}^{\pi} |p(e^{j\omega})|^2 dF_u(e^{j\omega}) = \sum_{k=1}^n \sigma_k^2 |p(e^{j\omega_k})|^2 \quad \sigma_k^2 > 0,$$

e se il primo membro di questa espressione fosse uguale a zero il polinomio  $p(z^{-1})$  dovrebbe allora soddisfare alle  $n$  condizioni  $|p(e^{j\omega_k})|^2 = 0; k = 1, 2, \dots, n$ , che sono equivalenti alle

$$p(e^{j\omega_k}) = 0, \quad k = 1, 2, \dots, n \quad (*)$$

condizioni che implicano  $p(z^{-1}) \equiv 0$ , dato che  $p(z^{-1})$  ha grado  $n - 1$ . Notiamo che se si prende un polinomio  $q(z^{-1})$  di grado  $n$ , si può invece avere  $q^\top \mathbf{R}_{n+1} q = 0$  perchè le  $n$  condizioni (\*) possono essere soddisfatte da un polinomio di grado  $n$  non identicamente nullo. Quindi l'ordine di persistente eccitazione è esattamente  $n$ .

Per mostrare che la condizione è anche necessaria dimostriamo che un segnale persistentemente eccitante di ordine esattamente  $n$  è in realtà la somma di  $n$  oscillazioni armoniche di frequenze  $\{\omega_1, \omega_2, \dots, \omega_n\}$ , nell'intervallo  $(-\pi, \pi)$  tra loro diverse. Per ipotesi esiste un polinomio di grado effettivo  $n$  con vettore dei coefficienti non nullo  $q \in \mathbb{R}^n$ , tale che

$$0 = q^\top \mathbf{R}_{n+1} q = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N [q(z^{-1})u(t)]^2$$

Si mostra che il limite a secondo membro è in realtà il quadrato della norma del segnale  $t \rightarrow q(z^{-1})u(t)$ ; i.e. si può scrivere  $\|q(z^{-1})u(\cdot)\|^2$ , in un opportuno spazio di Hilbert di segnali stazionari del second'ordine. Il fatto che questa norma è zero, implica che  $u$  soddisfa l'equazione alle differenze

$$q(z^{-1})u(t) = 0 \quad t \in \mathbb{Z}.$$

Questa equazione alle differenze può solo avere soluzioni limitate, giacchè l'esistenza di un modo esponenziale (crescente o decrescente) nel segnale  $u$  renderebbe infinito qualcuno dei limiti (3.3.1) [CHECK !!]. Ne segue che tutti gli zeri di  $q(z^{-1})$

<sup>12</sup>Dato che per segnali reali  $r(\tau)$  è reale, si tratta in realtà di coppie di frequenze opposte  $\pm\omega_k$ .

debbono trovarsi sul cerchio unitario e debbono avere molteplicità uno. Quindi il segnale  $u$  è la somma di  $n$  oscillazioni armoniche di frequenze  $\{\omega_1, \omega_2, \dots, \omega_n\}$ . Il resto della dimostrazione scende dall'esempio che segue.  $\square$

**Example 3.1 (segnali quasi periodici).** La somma di  $N$  segnali sinusoidali di frequenza diversa

$$u(t) = \sum_{k=1}^N A_k \sin(\omega_k t + \phi_k) \quad \omega_k \neq \omega_j \quad (3.3.3)$$

è un segnale stazionario del secondo ordine, la cui correlazione vale

$$r(\tau) = \sum_{k=1}^N \frac{A_k^2}{2} \cos \omega_k \tau \quad (3.3.4)$$

e il suo spettro consiste di  $2N$  righe (funzioni  $\delta$ ) supportate nei punti  $\{\pm\omega_k\}$ . Il segnale è pertanto persistentemente eccitante (P.E.) di ordine (esattamente)  $2N$ .

Si veda [62, p. 98-109].

**Example 3.2 (Segnali periodici).** Scende dalla teoria della trasformata discreta di Fourier (DFT), che un segnale a tempo discreto, periodico di periodo  $N$ , è la somma di  $N$  componenti sinusoidali di frequenza  $\omega_1 = \frac{2\pi}{N}$ ,  $\omega_2 = 2\frac{2\pi}{N}$ ,  $\dots$ ,  $\omega_N = 2\pi$ . Quindi ogni segnale periodico di periodo  $N$  è P.E. di ordine  $2N$ .

**Example 3.3 (Segnali PRBS).** Un segnale PRBS (*Pseudo Random Binary Sequence*) è un particolare segnale periodico che approssima il rumore bianco ed è spesso usato nelle simulazioni. Si veda [62, p. 124-125].

### Sistemi con ingressi persistentemente eccitanti

È evidente che l'uscita di un sistema lineare stabile con ingresso P. E. di ordine esattamente  $n$  ha uno spettro che contiene al più  $n$  righe. Di fatto, le righe spettrali dell'uscita sono ancora  $n$  a meno che qualche frequenza propria dell'ingresso non coincida con degli zeri della funzione di trasferimento situati sulla circonferenza unità.

**Proposition 3.5.** *Si supponga che il processo  $y$  sia descritto dal modello (3.1.6) (senza reazione) con ingresso un segnale  $u$  persistentemente eccitante di ordine esattamente  $n$ . Se  $S_v(z)$  non si annulla in qualcuna delle  $n$  frequenze proprie di  $u$ , la matrice densità spettrale congiunta (3.1.4) è allora definita positiva nelle  $n$  frequenze proprie dell'ingresso.*

*Proof.* Infatti si ha

$$S(z) = \begin{bmatrix} S_{\mathbf{y}}(z) & S_{\mathbf{y}\mathbf{u}}(z) \\ S_{\mathbf{u}\mathbf{y}}(z) & S_{\mathbf{u}}(z) \end{bmatrix} = \begin{bmatrix} F(z)S_{\mathbf{u}}(z)F(1/z) + \lambda^2 G(z)G(1/z) & F(z)S_{\mathbf{u}}(z) \\ S_{\mathbf{u}}(z)F(1/z) & S_{\mathbf{u}}(z) \end{bmatrix} \quad (3.3.5)$$

$$= \begin{bmatrix} 1 & F(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \lambda^2 G(z)G(1/z) & 0 \\ 0 & S_{\mathbf{u}}(z) \end{bmatrix} \begin{bmatrix} 1 & 0 \\ F(1/z) & 1 \end{bmatrix} \quad (3.3.6)$$

e il determinante di  $S(z)$  è uguale a quello dell'ultimo membro, che è  $\lambda^2 G(z)G(1/z)S_{\mathbf{u}}(z)$ .  
□

### 3.4 Alcune classi di modelli e loro parametrizzazione

I modelli razionali del tipo (3.1.6) (detti qualche volta del tipo "Box-Jenkins"), se non vi sono vincoli a priori sui parametri, possono essere parametrizzati mediante i coefficienti dei relativi polinomi a numeratore e denominatore delle funzioni di trasferimento nel modello parametrico (3.1.6),

$$F_{\theta}(z) = \frac{B(z^{-1})}{A(z^{-1})} \quad G_{\theta}(z) = \frac{C(z^{-1})}{D(z^{-1})}. \quad (3.4.1)$$

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  $n = \deg(A)$  coefficienti del polinomio  $A(z^{-1}) = 1 + \sum_{k=1}^n a_k z^{-k}$ ,
- gli  $m = \deg(B)$  coefficienti del polinomio  $B(z^{-1}) = \sum_{k=1}^m b_k z^{-k}$ ,
- gli  $q = \deg(C)$  coefficienti del polinomio  $C(z^{-1}) = 1 + \sum_{k=1}^q c_k z^{-k}$ ,
- gli  $r = \deg(D)$  coefficienti del polinomio  $D(z^{-1}) = 1 + \sum_{k=1}^r d_k z^{-k}$

e quindi con un totale di  $p = n + m + q + r$  parametri "liberi", più la varianza dell'innovazione  $\lambda$  che conviene considerare separatamente. In realtà il vincolo che il modello (3.1.6) 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  $\Theta$ . Purtroppo la struttura geometrica degli insiemi che definiscono i coefficienti ammissibili è estremamente 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

$$A(z^{-1})\mathbf{y}(t) = B(z^{-1})\mathbf{u}(t) + C(z^{-1})\mathbf{e}(t) \quad (3.4.2)$$



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 (3.4.2) 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 (3.1.6) dovrebbe avere la struttura seguente

$$\mathcal{A}(z^{-1}) = A(z^{-1})D(z^{-1}) \quad \mathcal{B}(z^{-1}) = B(z^{-1})D(z^{-1}) \quad \mathcal{C}(z^{-1}) = C(z^{-1})A(z^{-1})$$

in cui però i parametri dei polinomi  $\mathcal{A}$ ,  $\mathcal{B}$  e  $\mathcal{C}$  (in totale  $n + r + m + r + q + n = 2n + m + 2r + q$ ) non sono più liberi di variare in modo indipendente ma debbono essere vincolati a soddisfare le relazioni algebriche che impongono le relazioni prodotte 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

$$A(z^{-1})\mathbf{y}(t) = B(z^{-1})\mathbf{u}(t) + \mathbf{e}(t) \quad (3.4.3)$$

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 (3.4.3) 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.

**Problem 3.2.** *Determinare l'ordine minimo di persistente eccitazione del segnale di ingresso  $u$  nel modello ARX senza reazione*

$$(1 + a_1z^{-1} + a_2z^{-2} + \dots + a_nz^{-n}) \mathbf{y}(t) = (b_1z^{-1} + b_2z^{-2} + b_3z^{-3}) u(t) + \mathbf{e}(t)$$

per avere identificabilità.

*Soluzione:*

Verifichiamo che un modello ARX senza reazione è identificabile se e solo  $\theta_1 = \theta_2$  è l'unica soluzione delle due equazioni

$$A_{\theta_1}(e^{j\omega}) - A_{\theta_2}(e^{j\omega}) = 0 \quad (3.4.4)$$

$$[B_{\theta_1}(e^{j\omega}) - B_{\theta_2}(e^{j\omega})] S_u(e^{j\omega}) = 0, \quad (3.4.5)$$

dove  $A$  e  $B$  sono i polinomi del modello e  $S_u$  è lo spettro di  $u$  (eventualmente espresso mediante funzioni  $\delta$ ).

In effetti dalla proposizione 3.1 si trae che un modello ARX è globalmente identificabile se e solo se l'unica soluzione delle equazioni

$$\left[ \frac{B_{\theta_1}(z^{-1})}{A_{\theta_1}(z^{-1})} - \frac{B_{\theta_2}(z^{-1})}{A_{\theta_2}(z^{-1})} \right] S_u(z) = 0 \quad (3.4.6)$$

$$\frac{1}{A_{\theta_1}(z^{-1})} - \frac{1}{A_{\theta_2}(z^{-1})} = 0 \quad (3.4.7)$$

è  $\theta_1 = \theta_2$ . Ovviamente la (3.4.7) è equivalente alla (3.4.4) e quindi, accoppiata alla (3.4.6) è equivalente alla (3.4.5).

Quindi nel nostro esempio i parametri  $a_k$  sono senz'altro identificabili qualunque sia  $n$  e qualunque sia lo spettro di  $\mathbf{u}$ , mentre perchè la seconda condizione abbia come unica soluzione  $\theta_1 = \theta_2$  è necessario e sufficiente che i polinomi  $B_{\theta_1}(z^{-1})$  e  $B_{\theta_2}(z^{-1})$  prendano valori uguali per  $m$  valori diversi di frequenza (verificare!). Quindi l'ordine minimo di persistente eccitazione per l'identificabilità di un modello ARX è  $m$ . Nel nostro esempio  $m = 3$  e la condizione di identificabilità impone che lo spettro dell'ingresso abbia almeno tre righe a tre frequenze distinte.

### 3.5 Identificabilità in presenza di reazione

Come abbiamo visto un processo congiunto  $[\mathbf{y}, \mathbf{u}]'$  stazionario e p.n.d. si può sempre descrivere con un modello d'innovazione a retroazione del tipo di figura 3.1.1 dove i processi  $\mathbf{v}$ ,  $\mathbf{r}$  (necessariamente p.n.d.) del modello sono scorrelati, di densità spettrali rispettive  $S_v$ ,  $S_r$  e quindi il modello è completamente descritto dalle quattro funzioni (razionali)  $F(z)$ ,  $H(z)$ ,  $S_v(z)$ ,  $S_r(z)$ . Naturalmente potremmo rappresentare  $S_v$ ,  $S_r$  mediante i rispettivi fattori spettrali canonici, ma queste rappresentazioni non sono naturali nel contesto presente perchè, come abbiamo visto, nel modello d'innovazione a retroazione sia  $G(z)$  che  $K(z)$  potrebbero avere poli instabili.

Sia  $S(z)$  la matrice densità spettrale congiunta del processo  $[\mathbf{y}, \mathbf{u}]'$ . Dato che le medie si assumono sempre nulle, questa è la descrizione probabilistica (del second'ordine) più completa del processo. In questa sezione vogliamo studiare in dettaglio la mappa

$$\mathfrak{R} : (F, H, S_v, S_r) \mapsto S,$$

e in particolare capire quando questa corrispondenza è *iniettiva*, ovvero ad un dato spettro congiunto corrisponde uno e un solo modello a retroazione d'innovazione.

Notiamo che la conoscenza dello spettro congiunto è la massima informazione che possiamo pensare di poter ricavare dall'osservazione del processo  $[y, u]'$ . Questa unicità è quindi un prerequisito fondamentale per l'identificabilità del modello a retroazione, indipendente dalla sua parametrizzazione. Naturalmente il modello a retroazione potrebbe descrivere una situazione reale in cui i dati sono effettive osservazioni ingresso-uscita di un sistema di controllo con una retroazione lineare, ma questo non è a stretto rigore necessario e la retroazione potrebbe essere solo "intrinseca" nel senso spiegato nella sezione 3.1.

Il calcolo dello spettro congiunto dei processi osservabili  $y$ ,  $u$  si può fare facilmente usando l'espressione della matrice di trasferimento (3.1.8), pervenendo così all'espressione

$$S_y = \frac{1}{|1 - HF|^2} S_v + \frac{FF^*}{|1 - HF|^2} S_r \quad (3.5.1)$$

$$S_{yu} = \frac{H^*}{|1 - HF|^2} S_v + \frac{F}{|1 - HF|^2} S_r \quad (3.5.2)$$

$$S_u = \frac{HH^*}{|1 - HF|^2} S_v + \frac{1}{|1 - HF|^2} S_r \quad (3.5.3)$$

dove non abbiamo indicato esplicitamente l'argomento  $e^{j\omega}$  nelle funzioni di trasferimento e l'asterisco indica il complesso coniugato.

Interessiamoci innanzitutto alla relazione tra  $S$  e la funzione di trasferimento in catena aperta  $F(z)$ . Notiamo subito che la retroazione complica la relazione tra gli spettri del segnale di ingresso e quello di uscita. In assenza di reazione si avrebbe semplicemente  $S_{yu} = FS_u$  per cui  $F(e^{j\omega})$  si potrebbe immediatamente ricavare dalla

$$F(e^{j\omega}) = \frac{S_{yu}(e^{j\omega})}{S_u(e^{j\omega})}.$$

Questa è in effetti la formula usata da molti analizzatori di spettro in commercio. In presenza di reazione invece si ha

$$\hat{F}(e^{j\omega}) := \frac{S_{yu}(e^{j\omega})}{S_u(e^{j\omega})} = \frac{F(e^{j\omega}) S_r(e^{j\omega}) + H^*(e^{j\omega}) S_v(e^{j\omega})}{H(e^{j\omega}) H^*(e^{j\omega}) S_v(e^{j\omega}) + S_r(e^{j\omega})} \quad (3.5.4)$$

e si vede che  $\hat{F} = F$  solo nel caso in cui  $H \equiv 0$  (assenza di reazione) o  $S_v \equiv 0$  mentre nel caso in cui  $S_r \equiv 0$  si ha addirittura  $\hat{F} = 1/H$ . In particolare l'assenza di eccitazione dall'ingresso esterno  $r$  (i.e.  $S_r \equiv 0$ ) comporta la *non identificabilità* di  $F$ .

**Lemma 3.1.** *Assumiamo che vi sia reazione (e quindi che  $H(z) \neq 0$ ). Lo spettro congiunto  $S(e^{j\omega})$  è singolare quasi ovunque se e solo se almeno uno dei due spettri  $S_v$ ,  $S_r$  è uguale a zero.*

**Proof.** In effetti usando la rappresentazione di figura 3.1.1 si vede che il processo congiunto si può sempre esprimere come uscita del sistema lineare di funzione di trasferimento (3.1.8) con in ingresso i due processi di innovazione di  $r$  e  $v$ . In altri

termini esiste un fattore spettrale quadrato a fase minima  $W(z)$  di  $S(z)$  normalizzato all'identità all'infinito tale che

$$S(z) = W(z) \begin{bmatrix} \lambda_1^2 & 0 \\ 0 & \lambda_2^2 \end{bmatrix} W(1/z)^\top, \quad W(\infty) = \begin{bmatrix} 1 & 0 \\ H(\infty) & 1 \end{bmatrix}.$$

Dato che  $W(\infty)$  è non singolare, la matrice razionale  $W(z)$  dev'essere non singolare in un intorno di  $\infty$  e quindi quasi dappertutto sul piano complesso. Dalla fattorizzazione precedente si vede quindi che (nell'ipotesi di presenza di reazione,  $H(z) \neq 0$ ),  $S(z)$  è singolare quasi ovunque se e solo se una (o entrambi) delle due varianze  $\lambda_1^2$ ,  $\lambda_2^2$  (e quindi uno o entrambi dei due processi  $\mathbf{r}$ ,  $\mathbf{v}$ ) è uguale a zero.  $\square$

**Proposition 3.6.** *Se e solo se lo spettro congiunto (3.1.4) è non singolare quasi ovunque, le funzioni  $(F, H, S_v, S_r)$  sono individuate univocamente dallo spettro  $S$ ; in altri termini, la mappa  $\mathfrak{X}$  è iniettiva.*

**Proof.** Se lo spettro è non singolare  $\lambda_1^2$  e  $\lambda_2^2$  sono diverse da zero. Siano allora  $\mathbf{e}_1$ ,  $\mathbf{e}_2$  i processi di innovazione di  $\mathbf{v}$  e  $\mathbf{r}$ . Dato che

$$\begin{aligned} \mathbf{e}_1(t) &= G(z)^{-1} [\mathbf{y}(t) - F(z)\mathbf{u}(t)] \\ \mathbf{e}_2(t) &= K(z)^{-1} [\mathbf{u}(t) - H(z)\mathbf{y}(t)] \end{aligned}$$

l'inverso del fattore a fase minima deve avere la struttura

$$W(z)^{-1} = \begin{bmatrix} G(z)^{-1} & G(z)^{-1}F(z) \\ K(z)^{-1}H(z) & K(z)^{-1} \end{bmatrix}$$

e si vede che le funzioni di trasferimento  $F, G, H, K$  sono univocamente determinate da  $W$  e quindi dallo spettro. Dato che  $G$  e  $K$  non risultano necessariamente dei fattori spettrali "canonici" (a fase minima), è più corretto affermare che sono di fatto gli spettri,  $S_r$  ed  $S_v$ , ad essere univocamente individuati.

Viceversa, consideriamo il caso in cui  $S_r$  è uguale a zero e quindi lo spettro congiunto è singolare. Come abbiamo visto più sopra, in questo caso  $F$  non è identificabile e pertanto la condizione è anche necessaria.  $\square$

In realtà le incognite del problema di identificazione in catena chiusa sono solo la funzione di trasferimento in catena di azione diretta,  $F(e^{j\omega})$  e lo spettro dell'errore di modellizzazione relativo,  $S_v(e^{j\omega})$ . Naturalmete in pratica il modello lineare razionale

$$\hat{\mathbf{y}}(t) = F_\theta(z)\mathbf{u}(t)$$

con cui si descrive il legame "deterministico" ingresso-uscita è sempre approssimato e quindi si è sempre in presenza di "errore di modellizzazione". Il termine che abbiamo descritto come termine di "disturbo aleatorio"  $\mathbf{v}$  è in effetti sempre presente e certamente non ha molto senso pensare che sia nullo. Si può così concludere questa discussione affermando che *se (e solo se)  $S_r > 0$ , la conoscenza dello*

spettro congiunto di  $\mathbf{y}$ ,  $\mathbf{u}$ , permette di ricavare in modo univoco  $F$  e  $S_v$ ; in altri termini, come vedremo meglio in seguito, si può affermare che se  $S_r > 0$ , dall'osservazione dei segnali  $\mathbf{y}$ ,  $\mathbf{u}$  (per un tempo teoricamente infinito) si può, in linea di principio, ricavare univocamente il modello in catena di azione diretta del sistema.

La situazione in cui  $\mathbf{r} \simeq 0$ , si descrive dicendo che *il segnale di riferimento  $\mathbf{r}$  non è sufficientemente eccitante*. In pratica la situazione  $\mathbf{r} = 0$  dev'essere pensata come una situazione limite che serve solo a dare delle indicazioni sulla difficoltà di identificare correttamente il sistema in condizioni di insufficiente eccitazione.

**Problem 3.3.** *Discutere l'identificabilità del sistema con retroazione "deterministica",*

$$\begin{aligned} (1 + az^{-1}) \mathbf{y}(t) &= b \mathbf{u}(t) + \mathbf{e}(t) \\ \mathbf{u}(t) &= k \mathbf{y}(t) \end{aligned}$$

in cui  $|a - bk| < 1$ .

*Soluzione:*

Dato che  $S_r \equiv 0$  (non c'è segnale di riferimento esterno) e quindi la retroazione è "deterministica"; i.e.  $\mathbf{u}(t) = H(z) \mathbf{y}(t)$ , lo spettro congiunto è singolare. Quindi ci dobbiamo aspettare che il sistema non sia identificabile. In effetti,  $\mathbf{y}$  è descritto dal modello AR

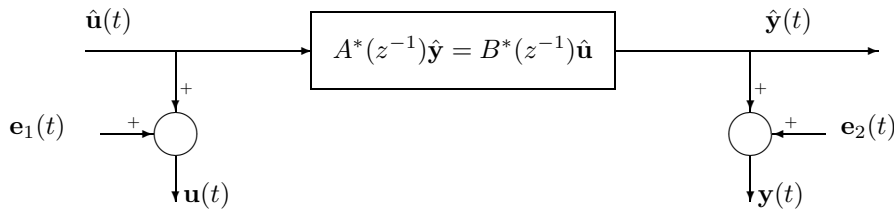
$$[1 + (a - bk)z^{-1}] \mathbf{y}(t) = \mathbf{e}(t)$$

e tutte le coppie di parametri  $[a, b]$  per cui  $a - bk = c$  con  $|c| < 1$  descrivono lo stesso sistema.

□

### 3.5.1 Modelli a errori nelle variabili

Lo schema a blocchi di figura ??escribe un cosiddetto modello a *Errori nelle Variabili (EIV)*



**Figure 3.5.1.** *Schema a blocchi del modello EIV.*

in cui le variabili cosiddette "vere",  $\hat{\mathbf{u}}(t)$  e  $\hat{\mathbf{y}}(t)$  sono legate tra di loro da una equazione alle differenze del tipo

$$A^*(z^{-1}) \hat{\mathbf{y}}(t) = B^*(z^{-1}) \hat{\mathbf{u}}(t)$$

e sono osservate in presenza di rumori additivi  $e_1(t)$  e  $e_2(t)$  (che si assumono spesso bianchi), scorrelati tra di loro ( $e_1 \perp e_2$ ) e con le variabili vere,

$$\begin{cases} \mathbf{u}(t) = \hat{\mathbf{u}}(t) + \mathbf{e}_1(t) & \mathbf{e}_1 \perp \hat{\mathbf{u}} \\ \mathbf{y}(t) = \hat{\mathbf{y}}(t) + \mathbf{e}_2(t) & \mathbf{e}_2 \perp \hat{\mathbf{y}} \end{cases}$$

Nel caso in cui  $e_1(t)$  e  $e_2(t)$  sono bianchi, è facile mostrare che  $\mathbf{y}(t)$  e  $\mathbf{u}(t)$  sono legate da un modello ARMAX. Calcolare i relativi polinomi e il rumore bianco in ingresso al modello.

Discutere se e sotto quali condizioni ci può essere assenza di reazione da  $\mathbf{y}$  ad  $\mathbf{u}$ .

### 3.6 Modelli multivariabili

Vedere il capitolo nel libro: Bittanti (ed.), *Identificazione parametrica* CLUP, Milano.

### 3.7 Modelli Gaussiani

La funzione di verosimiglianza. Il Limite di Cramèr-Rao per modelli dinamici Gaussiani.

## CHAPTER 4

# ERGODICITÀ

Si potrebbe ben affermare che i due risultati veramente fondamentali della teoria della probabilità sono il *teorema ergodico* e il *teorema del limite centrale*. Questi due teoremi, 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 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.

### 4.1 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 viene discusso in questo capitolo.

In analisi asintotica si immagina di avere una successione infinita di osser-

vazioni  $\{y_t\}$ , che si può pensare come una traiettoria di un processo  $\{y(t)\}$  a tempo discreto. Denotiamo con  $F^N(\cdot)$  la distribuzione di probabilità congiunta delle prime  $N$  variabili del processo, che supponiamo essere nota a meno di un certo parametro (di dimensione fissa)  $\theta$ . Sia  $\{\phi_N\}$  una successione di stimatori del parametro  $\theta$  che immaginiamo definiti in base ad un comune criterio di stima (ad esempio la massima verosimiglianza).

**Definition 4.1.** Sia  $(y_1, \dots, y_N)$  un campione estratto dalla distribuzione della famiglia  $\{F_\theta^N; \theta \in \Theta\}$  corrispondente al parametro  $\theta_0$  (il valore vero del parametro). Lo stimatore  $\phi_N(y_1, \dots, y_N)$  si dice **consistente**<sup>13</sup> se

$$\lim_{N \rightarrow \infty} \phi_N(y_1, \dots, y_N) = \theta_0 \quad ; \quad (4.1.1)$$

se il limite (4.1.1) è un limite in probabilità, cioè se la (4.1.1) significa che,  $\forall \varepsilon > 0$ ,

$$\lim_{N \rightarrow \infty} P_{\theta_0} \left( \|\phi_N(y_1, \dots, y_N) - \theta_0\| \geq \varepsilon \right) = 0 \quad , \quad (4.1.2)$$

si parla di consistenza “debole” o semplicemente di consistenza tout-court. Se invece il limite (4.1.1) è un limite con probabilità 1 (c.p. 1), si parla di consistenza forte. In questo caso si ha

$$\lim_{N \rightarrow \infty} \phi_N(y_1, \dots, y_N) = \theta_0 \quad (4.1.3)$$

per tutte le possibili successioni  $\{y_1, y_2, \dots, y_N, \dots\}$  di osservazioni in  $(\mathbb{R}^m)^\infty := \mathbb{R}^m \times \mathbb{R}^m \times \dots$  (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.  $y_1, \dots, y_N, \dots$  sono *realmente* distribuite, cioè la probabilità *vera*, corrispondente al valore “vero”  $\theta_0$  del parametro.

Notiamo anche che la consistenza è una proprietà molto più forte della *correttezza asintotica* che si definisce con la relazione

$$\lim_{N \rightarrow \infty} E_\theta \phi_N(y_1, \dots, y_N) = \theta \quad \forall \theta \in \Theta \quad (4.1.4)$$

che ha scarso significato statistico.

### 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

<sup>13</sup>Qui per *consistenza* si intende “fondatezza logica”; dal latino *consistens*, part. pass. di *consistere* = “stare saldo”.



partire semplicemente dalla convergenza di medie e varianze. Usando la classica disuguaglianza,

$$P_\theta \left( \|\phi_N - \theta\| \geq \varepsilon \right) \leq \frac{1}{\varepsilon^2} E_\theta [(\phi_N - \theta)'(\phi_N - \theta)] = \frac{1}{\varepsilon^2} E_\theta \|\phi_N - \theta\|^2, \quad (4.1.5)$$

dove  $\phi_N := \phi_N(\mathbf{y}_1, \dots, \mathbf{y}_N)$  e  $\|\cdot\|$  indica l'usuale norma euclidea. Notiamo che se  $\phi_N$  è corretto  $E_\theta \phi_N = \theta$  e pertanto l'espressione a secondo membro è la varianza,  $\sigma_N^2(\theta)$ , di  $\phi_N(\mathbf{y}_1, \dots, \mathbf{y}_N)$  divisa per  $\varepsilon^2$ . Si vede che se

$$\lim_{N \rightarrow \infty} \sigma_N^2(\theta) = 0 \quad , \quad \forall \theta \in \Theta \quad ,$$

allora  $\phi_N(\mathbf{y}_1, \dots, \mathbf{y}_N)$  è consistente. Con una lieve generalizzazione si ottiene il seguente criterio.

**Proposition 4.1.** *Se  $\phi_N(\mathbf{y}_1, \dots, \mathbf{y}_N)$  è, uno stimatore asintoticamente corretto e se la sua varianza scalare  $\sigma_N^2(\theta)$  tende a zero con  $N$  per ogni  $\theta \in \Theta$ , allora  $\phi_N(\mathbf{y}_1, \dots, \mathbf{y}_N)$  è consistente.*

Praticamente 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.

**Example 4.1.** Dalle formule (2.3.9) e (2.3.10) si vede che lo stimatore della varianza  $\sigma^2$  nel modello lineare (2.1.7) è asintoticamente corretto e consistente.

**Example 4.2.** Supponiamo che la distribuzione vera di  $y$  (scalare) sia del tipo di Cauchy, ovvero  $dF_{\theta_0}(y) = p(y, \theta_0) dy$  con

$$p(y, \theta) = \frac{1}{\pi} \frac{1}{1 + (y - \theta)^2} \quad , \quad \theta \in R \quad .$$

Sia  $(\mathbf{y}_1, \dots, \mathbf{y}_N)$  un campione casuale e  $\bar{y}_N$  la relativa media campionaria. Usando le funzioni caratteristiche si può vedere che  $\bar{y}_N$  ha, per ogni  $N$ , la stessa distribuzione di  $y$  e pertanto la probabilità

$$P(|\bar{y}_N - \theta_0| \geq \varepsilon)$$

rimane la stessa al variare di  $N$  e non può dunque tendere a zero con  $N$ . Questo implica che  $\bar{y}_N$  non è consistente (tanto meno è fortemente consistente dato che  $E_{\theta_0} \mathbf{y} = \infty!$ ).

## 4.2 Ergodicità: motivazioni fisiche

Le origini della teoria ergodica risalgono alla meccanica statistica di Boltzmann, Maxwell, Gibbs etc. e ai tentativi di dare un fondamento rigoroso alla termodinamica, basato su procedimenti di "media statistica" dei fenomeni microscopici

che essa descrive in modo aggregato. Dal nostro punto di vista un riferimento classico per questo argomento è il libro di Landau-Lifschitz [30]. Una trattazione più moderna, completa e concettualmente soddisfacente si trova nel libro di Martin-Löf [39].

Consideriamo il moto di un sistema microscopico di  $N$  particelle, ciascuna descritta dalle variabili di fase posizione e momento,  $(q_k, p_k)$ , componenti di un vettore  $6N$  dimensionale  $z(t) := [q(t) p(t)]'$ , detto la *fase* del sistema che penseremo evolvere nel tempo in un certo sottoinsieme  $\Omega$  di  $\mathbb{R}^{6N}$  detto *spazio delle fasi*. Negli esempi che interessano  $N$  è tipicamente dell'ordine del numero d Avogadro  $\sim 10^{23}$ . Il moto del sistema è descritto dalle equazioni di Hamilton

$$\dot{q}_k(t) = \frac{\partial}{\partial p_k} H(q, p) \quad (4.2.1)$$

$$\dot{p}_k(t) = -\frac{\partial}{\partial q_k} H(q, p) \quad (4.2.2)$$

dove la funzione Hamiltoniana  $H(q, p) \equiv H(z)$  è l'energia totale del sistema. Queste equazioni, che vanno associate alla fase iniziale  $z(0) \in \Omega$ , determinano l'evoluzione temporale del vettore di fase per tutti i  $t \in \mathbb{R}$ . La soluzione delle equazioni di Hamilton si scrive simbolicamente come

$$z(t) = \Phi(t)z(0)$$

dove l'operatore (in generale non lineare)  $\Phi(t) : \Omega \rightarrow \Omega$ , che si chiama *flusso Hamiltoniano*, ha la proprietà di gruppo

$$\Phi(t+s) = \Phi(t)\Phi(s)$$

che è un modo astratto di scrivere la ben nota proprietà di composizione della soluzione di un'equazione differenziale. Come si può controllare facilmente l'evoluzione del sistema microscopico (4.2.1) soddisfa alla

$$\frac{d}{dt} H(z(t)) = 0$$

cioè mantiene costante l'energia totale del sistema. In altri termini l'evoluzione del sistema microscopico è conservativa.

Dato che i fenomeni in scala microscopica sono estremamente complessi, è naturale cercare di descriverli usando un approccio statistico. Il primo passo per far questo è cercare di modellare lo spazio delle fasi in modo probabilistico. Una distribuzione di probabilità su  $\Omega$  che è invariante per il flusso Hamiltoniano  $\Phi(t)$  del sistema definisce uno *stato di equilibrio termico*. Si dimostra [30, 39] che in uno spazio di fase di dimensione finita, ogni distribuzione di probabilità assolutamente continua che è invariante per il flusso Hamiltoniano ha una densità  $\rho(z)$  della classe di Maxwell-Boltzmann, uguale, a meno di una costante di normalizzazione moltiplicativa a  $\exp[-\frac{1}{2\beta} H(z)]$ , dove il parametro  $\beta > 0$ , ha l'interpretazione di temperatura assoluta. Per una funzione Hamiltoniana di tipo quadratico e uno

spazio di fase Euclideo, queste densità sono quindi di tipo Gaussiano. Per studiare un sistema microscopico in equilibrio termico, è quindi naturale prendere su  $\Omega$  una densità di probabilità invariante del tipo di Maxwell-Boltzmann. In questo modo la fase iniziale  $z(0)$  può essere interpretata come un evento elementare ( $\omega \in \Omega$ ) scelto a caso in uno spazio di probabilità  $\{\Omega, \mathcal{A}, \rho\}$ . Notiamo anche che in equilibrio termico, ogni funzione misurabile  $h : \Omega \rightarrow \mathbb{R}$ , definita sullo spazio delle fasi, può essere identificata con una variabile aleatoria e la relativa evoluzione temporale  $y(t) = h(z(t)) = h(\Phi(t)z(0))$  si può identificare con un *processo stocastico stazionario* definito sullo spazio di probabilità  $\{\Omega, \mathcal{A}, \rho\}$ .

Ora, la termodinamica (di equilibrio) si occupa di studiare le relazioni tra certe funzioni "macroscopicamente osservabili" della fase del sistema, diciamo ancora  $y = h(z)$ , ad es. pressione, volume specifico etc.. Evidentemente, per effetto della dinamica microscopica, ogni osservabile è in realtà funzione del tempo:  $y(t) = h(z(t))$ , l'evoluzione temporale essendo formalmente dovuta al flusso Hamiltoniano microscopico  $z(t) = \Phi(t)z(0)$ . Quando il sistema è in equilibrio termico, si vorrebbe ragionevolmente poter descrivere le osservabili come quantità *costanti nel tempo*, tipicamente prendendone medie temporali su tempi "sufficientemente lunghi" rispetto ai tempi propri della dinamica microscopica del sistema e trascurare le fluttuazioni "statistiche" dovute alla dinamica microscopica. L'idea è insomma di riuscire a costruire una teoria fisica che descriva relazioni tra medie temporali del tipo

$$\bar{y} = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T y(t) dt$$

dove si suppone che  $T \rightarrow \infty$  definisca i tempi "sufficientemente lunghi" cui abbiamo accennato. Notiamo che questo schema ha una prima difficoltà concettuale dovuta al fatto che a priori il limite  $\bar{y}$  potrebbe in genere dipendere dalla fase iniziale. Questo implicherebbe la necessità di dover tener conto della fase iniziale microscopica del sistema, cosa chiaramente impossibile. Inoltre l'uso dei valori medi temporali, pur essendo fisicamente ragionevole, presenta a priori di problemi di calcolo niente affatto banali.

Queste due difficoltà vengono contemporaneamente superate se vale una relazione limite del tipo:

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T y(t) dt = \mathbb{E} h \quad (4.2.3)$$

la quale dice che la media temporale  $\bar{y}$  è uguale alla media probabilistica (o d'insieme)

$$\mathbb{E} h = \int_{\Omega} h(z) \rho(z) dz.$$

Quest'ultima è in linea di principio calcolabile partendo dalla conoscenza del sistema e della distribuzione di equilibrio termico.

Dagli inizi del 900, l'eguaglianza (4.2.3) è stata denominata **ipotesi ergodica**. Essa dev'essere *dimostrata* per ogni classe particolare di sistemi dinamici microscopici per cui si voglia costruire la termodinamica. Di fatto a tutt'oggi l'ipotesi ergodica è stata dimostrata per pochissime classi di sistemi di dimensione finita

[59]. L'opinione di molti studiosi è che essa possa valere in generale solo per sistemi con infiniti gradi di libertà (ovvero come caso limite per  $N \rightarrow \infty$ ).

### 4.3 Ergodicità: teoria assiomatica

Con l'approccio assiomatico alla teoria della probabilità introdotto da Kolmogorov nel 1933, il problema dell'ergodicità è stato semplificato e generalizzato, spogliandolo di tutto il contesto fisico. Si parte semplicemente con un processo stocastico stazionario (in senso stretto!) definito su uno spazio di probabilità astratto  $\{\Omega, \mathcal{A}, P\}$  e ci si chiede sotto quali condizioni possa valere la relazione limite (4.2.3). Per i nostri scopi considereremo in realtà solo processi a tempo discreto.

Converrà innanzitutto ricordare la definizione di processo stazionario in senso stretto.

**Definition 4.2.** 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  $n$ ,

$$F_n(x_1, \dots, x_n, t_1 + \Delta, \dots, t_n + \Delta) = F_n(x_1, \dots, x_n, t_1, \dots, t_n) \quad ,$$

identicamente in  $x_1, \dots, x_n, t_1, \dots, t_n$ , 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*

$$\Sigma(t_1, t_2) := E [\mathbf{y}(t_1) - \mu(t_1)] [\mathbf{y}(t_2) - \mu(t_2)]'$$

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:

$$\mathbf{z} = f(\mathbf{y}(\tau) ; \tau \in \mathbb{I}) \quad , \tag{4.3.1}$$

che sono funzioni "tempo invarianti" del processo. Ad esempio, per un processo scalare  $\{y(t)\}$ , si possono considerare espressioni del tipo

$$\mathbf{z} = y^2(0) + 3y^2(1) y(-1) + \cos y(2) \quad ,$$

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  $E|\mathbf{z}| = E|f(\mathbf{y}(\tau) \mid \tau \in \mathbb{I})| < \infty$ . 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 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

$$\mathbf{z}(t) := f(\mathbf{y}(t + \tau) ; \tau \in \mathbb{I}) \quad , \quad t \in \mathbb{Z} \quad , \quad (4.3.2)$$

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

$$\begin{aligned} P\{\mathbf{z}(t) \in A\} &= P\{f(\mathbf{y}(t + \tau) ; \tau \in \mathbb{I}) \in A\} \\ &= P\{(\mathbf{y}(t + \tau_1), \dots, \mathbf{y}(t + \tau_N)) \in f^{-1}(A)\} \\ &= P\{(\mathbf{y}(\tau_1), \dots, \mathbf{y}(\tau_N)) \in f^{-1}(A)\} \\ &= P\{\mathbf{z} \in A\} \quad , \end{aligned} \quad (4.3.3)$$

dove con  $\tau_1, \dots, \tau_N$  si sono indicati gli elementi di  $\mathbb{I}$  e il simbolo  $f^{-1}(A)$  è l'antiimmagine dell'insieme  $A$  attraverso  $f$ , cioè  $f^{-1}(A) := \{x_1, \dots, x_N \mid f(x_1, \dots, x_N) \in A\}$ . Con un ragionamento analogo si può dimostrare l'invarianza temporale delle distribuzioni congiunte del processo  $\{\mathbf{z}(t)\}$  di ordine qualunque.

**Definition 4.3.** La variabile  $\mathbf{z} = f(\mathbf{y})$  è **invariante** (per traslazione) se  $\mathbf{z}(t) = \mathbf{z}$  per ogni  $t \in \mathbb{Z}$ .

Come vedremo tra poco il concetto di invarianza ha un'importanza fondamentale. Un modo formalmente ortodosso (anche se un pò più complicato) di

definire l'invarianza è attraverso l'introduzione dell'operatore di *traslazione temporale*  $U$ , che agisce trasladando nel tempo le (componenti scalari delle ) variabili del processo  $y$  ed è definito tramite la posizione

$$Uy_k(t) = y_k(t + 1), \quad k = 1, 2, \dots, m.$$

L'operatore  $U$  può essere esteso a tutte le variabili  $z \in L^1(y)$  semplicemente ponendo, se  $z = f(y(\tau) \mid \tau \in \mathbb{I})$ ,

$$Uz = f(y(\tau + 1) \mid \tau \in \mathbb{I}) = z(1).$$

e si controlla facilmente che  $U$  è lineare, invertibile ( $U^{-1}y(t) = y(t - 1)$ ) 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  $L^1(y)$  i quali, per ogni  $t \in \mathbb{Z}$ , trasformano ogni variabile aleatoria  $z \in L^1(y)$  secondo la relazione (di "traslazione" nel tempo)

$$U^t z := z(t) \quad , \quad (4.3.4)$$

dove  $z(t)$  è definita dalla formula (4.3.2). Notiamo che la stazionarietà di  $\{z(t)\}$  (formula (4.3.3)) è equivalente a dire che  $U^t$  è un *operatore che preserva la norma in*  $L^1(y)$ .

Sia  $\{z_k\}$  una successione convergente in  $L^1(y)$ . Dato che  $E|U^t(z_n - z_m)| = E|z_n - z_m|$  si può facilmente vedere che  $U^t$  è una trasformazione continua rispetto alla convergenza (in media) in  $L^1(y)$ .

**Proposition 4.2.** *La variabile casuale  $z \in L^1(y)$  è invariante se e solo se è invariante per l'operatore  $U$ , ovvero*

$$Uz = z \quad . \quad (4.3.5)$$

Dalla definizione segue immediatamente che  $z$  è invariante se e solo se

$$z(t) = U^t z = z \quad , \quad \forall t \in \mathbb{Z} \quad , \quad (4.3.6)$$

il che significa ancora che  $z = f(y(\tau) ; \tau \in \mathbb{I})$  non cambia, comunque si traslino temporalmente le variabili  $y(\tau)$  del processo. Ne segue che  $z$  non dipende affatto dal processo ed è quindi una costante deterministica, oppure dipende solo dal "comportamento asintotico" di  $\{y(t)\}$  nell'intorno di  $\pm\infty$ . Come vedremo meglio nella prossima sezione, una variabile invariante (non banale) può solo dipendere dalla "coda" infinitamente futura o infinitamente remota del processo  $\{y(t)\}$ .

**Lemma 4.1.** *Le variabili aleatorie invarianti formano un sottospazio (chiuso) di  $L^1(y)$ . Denoteremo questo spazio col simbolo  $L_\infty(y)$ .*

**Proof.** In effetti, per la linearità di  $U$ , se  $z_1, z_2 \in L^1(y)$  sono invarianti anche una qualunque loro combinazione lineare soddisfa la condizione (4.3.5). Inoltre, data una successione convergente  $\{z_k\}$  di v.a. invarianti (per le quali  $Uz_k = z_k$ ) segue dalla continuità di  $U$  che anche il loro limite è invariante.  $\square$

Riportiamo ora l'enunciato del fondamentale

**Theorem 4.1 (Teorema Ergodico di Birkhoff).** *Sia  $\{\mathbf{y}(t)\}$  un processo strettamente stazionario. Il limite*

$$\bar{z} := \lim_{T \rightarrow \infty} \frac{1}{T} \sum_1^T f_t(\mathbf{y}) \quad (4.3.7)$$

*esiste con probabilità uno per tutte le funzioni  $f(\mathbf{y}) \in L^1(\mathbf{y})$  ed è una variabile aleatoria che è invariante per l'operatore di traslazione del processo  $\{\mathbf{y}(t)\}$ . Se  $f(\mathbf{y}) \in L^2(\mathbf{y})$  il limite esiste anche in media quadratica.*

La dimostrazione dell'esistenza del limite è abbastanza complicata anche se è stata semplificata considerevolmente rispetto a quella originale di Birkhoff [5]. Rimandiamo il lettore al testo di Rozanov [54, p. 157]. Il fatto che il limite sia una variabile invariante scende dalla continuità di  $U$ . Infatti

$$U^s \bar{z} = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_1^T f_{t+s}(\mathbf{y}) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=s+1}^{T+s} f_t(\mathbf{y})$$

e dato che nell'ultimo membro dell'uguaglianza

$$\begin{aligned} \bar{z} &= \lim_{T \rightarrow \infty} \frac{1}{T+s} \sum_{t=1}^{T+s} f_t(\mathbf{y}) \\ &= \lim_{T \rightarrow \infty} \left[ \frac{1}{T+s} \sum_{t=1}^s f_t(\mathbf{y}) + \frac{1}{T} \frac{T}{T+s} \sum_{t=s+1}^{T+s} f_t(\mathbf{y}) \right] \end{aligned}$$

si ha  $\frac{T}{T+s} \rightarrow 1$  per  $T \rightarrow \infty$ , segue l'asserto. □

Notiamo adesso che si ha

$$\mathbb{E} \left\{ \frac{1}{T} \sum_{t=1}^T f_t(\mathbf{y}) \right\} = \frac{1}{T} \sum_{t=1}^T \mathbb{E} f_t(\mathbf{y}) = \mathbb{E} f(\mathbf{y})$$

dato che  $\mathbf{z}(t) = f_t(\mathbf{y})$  è un processo stazionario. Prendiamo ora l'aspettazione dei due membri nella (4.3.7). Dato che si può passare il limite per  $T \rightarrow \infty$  sotto il segno di aspettazione, si trova

$$\mathbb{E} \bar{z} = \mathbb{E} f(\mathbf{y}). \quad (4.3.8)$$

Diamo allora la seguente definizione.

**Definition 4.4.** *Il processo stazionario  $\mathbf{y}$  è ergodico se tutte le sue variabili invarianti sono costanti deterministiche.*

Il corollario seguente viene spesso preso come *definizione di ergodicità*.

**Corollary 4.1.** *Se e solo se  $\{\mathbf{y}(t)\}$  è ergodico si ha*

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_1^T f_t(\mathbf{y}) = \mathbb{E} f(\mathbf{y}) \quad (4.3.9)$$

con probabilità uno, qualunque sia  $f(\mathbf{y}) \in L^1(\mathbf{y})$ .

**Proof.** Se il processo è ergodico  $\bar{z}$  è una costante e coincide necessariamente con la sua aspettazione per cui scende da (4.3.8) che  $\bar{z} = \mathbb{E} \bar{z} = \mathbb{E} f(\mathbf{y})$ . Viceversa, è possibile mostrare (ma noi qui non lo faremo) che se vale la (4.3.9), ogni variabile invariante è una costante deterministica.  $\square$

### Alcune classi di processi ergodici

A differenza di quanto accade nel contesto fisico, in cui i sistemi dinamici (di dimensione finita) che sono ergodici sembrano essere “pochi”, esistono ampie classi di processi stocastici che sono ergodici. Per studiarne le proprietà avremo bisogno di alcune nozioni preliminari.

Generalizzando la costruzione degli spazi di Hilbert che si introducono nella teoria lineare dei processi a varianza finita, definiamo qui i sottospazi popolati dalle funzioni della storia “passata” e “futura” di  $\mathbf{y}$  all’istante  $t$ ,

$$L_t^-(\mathbf{y}) := \{\mathbf{z} \mid \mathbf{z} \in L^1(\mathbf{y}), \mathbb{I} \subset (-\infty, t]\} \quad L_t^+(\mathbf{y}) := \{\mathbf{z} \mid \mathbf{z} \in L^1(\mathbf{y}), \mathbb{I} \subset [t, +\infty)\} \quad (4.3.10)$$

Non è difficile convincersi che  $L_t^-(\mathbf{y})$  è un sottospazio chiuso di  $L^1(\mathbf{y})$  e che

$$L_{t+s}^-(\mathbf{y}) = U_s L_t^-(\mathbf{y}), \quad t, s \in \mathbb{Z}$$

crece 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:

$$L_\infty^-(\mathbf{y}) := \bigcap_{t \leq k} L_t^-(\mathbf{y}) \quad L_\infty^+(\mathbf{y}) := \bigcap_{t \geq k} L_t^+(\mathbf{y}) \quad (4.3.11)$$

In queste relazioni la scelta dell’istante iniziale  $k$  è irrilevante dato che le successioni di sottospazi in oggetto sono entrambe monotone. Vale allora il seguente risultato.

**Theorem 4.2.** *Il sottospazio delle variabili aleatorie invarianti è sempre contenute nei sottospazi passato e futuro remoto, ovvero*

$$L_\infty(\mathbf{y}) \subseteq L_\infty^-(\mathbf{y}) \cap L_\infty^+(\mathbf{y}). \quad (4.3.12)$$

La dimostrazione si può trovare in [54, 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 stringente di quella di processo p.n.d. (in senso debole) che si riferisce a sottospazi di  $L^2(\mathbf{y})$  generati *linearmente* dalle variabili del processo. Segue immediatamente dal Teorema 4.2 che

**Proposition 4.3.** *Un processo p.n.d. in senso stretto è ergodico.*



**Remark 4.1.** 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  $H(\mathbf{y})$ .

**Example 4.3.** Supponiamo che esista il  $\lim_{t \rightarrow \infty} \mathbf{y}(t)$  (con probabilità 1) e sia  $\mathbf{z}$  la variabile aleatoria

$$\mathbf{z} := \lim_{t \rightarrow \infty} \mathbf{y}(t) \quad .$$

Allora, per la continuità di  $U$ , si ha

$$U\mathbf{z} = \lim_{t \rightarrow \infty} U\mathbf{y}(t) = \lim_{t \rightarrow \infty} \mathbf{y}(t+1) = \mathbf{z}$$

e  $\mathbf{z}$  è invariante. Chiaramente un discorso perfettamente analogo può essere fatto per le variabili

$$\limsup_{t \rightarrow \pm\infty} \mathbf{y}(t) \quad , \quad \liminf_{t \rightarrow \pm\infty} \mathbf{y}(t) \quad .$$

**Example 4.4.** 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  $\pi$ . 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  $A_1, \dots, A_N$ , dove gli insiemi  $A_i$  costituiscono una partizione dello spazio di stato dal processo che qui identificheremo con l’insieme  $\{1, \dots, n\}$  dei primi  $n$  numeri naturali.

Consideriamo variabili casuali aventi la seguente struttura:

$$\mathbf{z} = f(\mathbf{y}(t)) = c_i \quad \text{se} \quad \mathbf{y}(t) \in A_i \quad , \quad i = 1, \dots, N \quad ,$$

ovvero,

$$\mathbf{z} = \sum_1^N c_i I_{A_i}(\mathbf{y}(t)) \quad ,$$

dove  $c_i, i = 1, \dots, N$ , 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.

Come vedremo i processi ergodici debbono essere molto “irregolari”. In effetti, un classico esempio di processo ergodico è un processo  $\{\mathbf{y}(t)\}$  a variabili i.i.d. (indipendenti e identicamente distribuite).

**Proposition 4.4 (Kolmogorov).** *Per un processo a variabili indipendenti lo spazio  $L_\infty(\mathbf{y})$  contiene solo (variabili aleatorie) costanti.*

**Proof.** Dimostriamo che l' affermazione è vera per  $L_{\infty}^{+}(\mathbf{y})$ . (La prova per  $L_{\infty}^{-}(\mathbf{y})$  è analoga.)

Sia  $L_0^n(\mathbf{y})$  il sottospazio di  $L^1(\mathbf{y})$  contenente tutte le funzioni delle variabili  $\{\mathbf{y}(0), \mathbf{y}(1), \dots, \mathbf{y}(n)\}$  che hanno aspettazione finita. Ovviamente ogni variabile aleatoria  $\mathbf{x}$  in  $L_{\infty}^{+}(\mathbf{y})$  deve essere indipendente dalle variabili in  $L_0^n(\mathbf{y})$ , qualunque sia  $n$ ; i.e.

$$\mathbb{E} \mathbf{x} \mathbf{z}^n = \mathbb{E} \mathbf{x} \mathbb{E} \mathbf{z}^n, \quad \mathbf{z}^n \in L_0^n(\mathbf{y})$$

e quindi  $\mathbf{x}$  dovrà essere indipendente da ogni variabile  $\mathbf{z}$  di  $L_0^+(\mathbf{y})$  perchè quest'ultimo è la chiusura (in  $L^1$ ) dello spazio vettoriale generato da tutte le variabili della famiglia  $\{L_0^n(\mathbf{y}); n \geq 0\}$  e ogni  $\mathbf{z} \in L_0^+(\mathbf{y})$ , o appartiene a un qualche  $L_0^n(\mathbf{y})$ , o è il limite (in  $L^1$ ) di qualche sequenza  $\{\mathbf{z}^n\}$ . Insomma, se  $\mathbf{x} \in L_{\infty}^{+}(\mathbf{y})$ ,

$$\mathbb{E} \mathbf{x} \mathbf{z} = \mathbb{E} \mathbf{x} \mathbb{E} \mathbf{z}, \quad \forall \mathbf{z} \in L_0^+(\mathbf{y})$$

D'altro canto  $L_{\infty}^{+}(\mathbf{y})$  è contenuto in  $L_0^+(\mathbf{y})$  e quindi ogni variabile  $\mathbf{x}$  è indipendente da sè stessa

$$\mathbb{E} (\mathbf{x})^2 = \mathbb{E} \mathbf{x} \mathbb{E} \mathbf{x}, \quad \forall \mathbf{x} \in L_{\infty}^{+}(\mathbf{y})$$

il che può essere vero solo se  $\mathbf{x}$  è una costante deterministica.  $\square$

Notiamo che questo risultato (noto come *legge dello 0-1* di Kolmogorov) non richiede la stazionarietà. La dimostrazione "classica" si svolge ragionando sulla  $\sigma$ -algebra degli eventi "infinitamente futuri" o "infinitamente remoti" del processo (si mostra che questi eventi possono avere solo probabilità zero o uno), vedere per esempio [14, pp. 102-103] e [54, Esempio 6.1 p. 162].

Una conseguenza interessante dell'ergodicità è che *un processo ergodico non può ammettere limite per  $t \rightarrow \pm\infty$  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  $\delta$  di Dirac). Qualunque sia il tipo di convergenza (in probabilità, in media o quasi ovunque) secondo la quale  $\mathbf{y}(t) \rightarrow \mathbf{z}$ , 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  $\delta$  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 4.4 presentato più sopra).

Scende dalla definizione che

**Proposition 4.5.** *Il processo  $\{\mathbf{z}(t)\}$  ottenuto per traslazione di una arbitraria funzione  $\mathbf{z} = f(\mathbf{y}) \in L^1(\mathbf{y})$  di un processo ergodico è ancora ergodico.*

**Proof.** 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.  $\square$

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.

**Theorem 4.3 (Doob).** *Sia  $\{\mathbf{e}(t)\}$  un processo i.i.d. a varianza finita. Si assuma che la successione di numeri reali  $\{c_k\}$  sia a quadrato sommabile,*

$$\sum_{-\infty}^{+\infty} c_k^2 < \infty \quad ; \quad (4.3.13)$$

allora il processo  $\{\mathbf{y}(t)\}$  definito dalla

$$\mathbf{y}(t) := \sum_{-\infty}^{+\infty} c_k \mathbf{e}(t+k) \quad (4.3.14)$$

è (strettamente stazionario), a varianza finita ed ergodico.

La condizione (4.3.13) serve a garantire la convergenza della somma. Una dimostrazione si può trovare nel trattato di Doob [14, p. 460]. Notiamo che definendo  $\bar{c}_k := c_{-k}$  la (4.3.14) si riscrive

$$\mathbf{y}(t) := \sum_{-\infty}^{+\infty} \bar{c}_k \mathbf{e}(t-k)$$

che ha l'usuale aspetto di somma di convoluzione. In sostanza, l'uscita di un filtro lineare (non necessariamente causale)  $\ell^2$ -stabile con ingresso rumore bianco (in senso stretto) è un processo ergodico.

Illustreremo ora alcune applicazioni del teorema ergodico. 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  $\{\mathbf{y}(t)\}$ .

### 4.3.1 Ergodicità e inferenza statistica

Studieremo in questa sezione una questione che è intimamente legata al problema generale dell'inferenza statistica cui abbiamo accennato all'inizio del capitolo 1. In termini generali, il problema è il seguente:

**Problem 4.1.** *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  $\{\bar{y}(t)\}$ ?*

Ricordiamo che si chiama *legge di probabilità del processo* la famiglia infinita di distribuzioni di probabilità

$$F_n(x_1, x_2, \dots, x_n; t_1, t_2, \dots, t_n) = P\{\mathbf{y}(t_1) \leq x_1, \mathbf{y}(t_2) \leq x_2, \dots, \mathbf{y}(t_n) \leq x_n\}$$

dove le  $\{x_k\}$  sono variabili reali se il processo è scalare ovvero variabili vettoriali in  $\mathbb{R}^m$  in caso contrario.

Facciamo vedere che esiste una classe di processi per cui questo problema ha una soluzione ben definita, che in un certo senso è sorprendentemente positiva. Questi processi sono per l'appunto i processi *ergodici*.

Sia  $E$  un qualunque sottoinsieme di Borel di  $\mathbb{R}$  (ad esempio un intervallo) e consideriamo la funzione

$$f(\mathbf{y}) := I_E(\mathbf{y}(0)) \quad ,$$

dove  $I_E$  è la funzione indicatrice dell'insieme  $E$ . La variabile casuale

$$\nu_T(E) := \frac{1}{T+1} \sum_{t=0}^T I_E(\mathbf{y}(t))$$

è la frequenza relativa con cui il processo  $\{\mathbf{y}(t)\}$  "visita" l'insieme  $E$ . Se definiamo  $\mathbf{z} := I_E(\mathbf{y}(0))$  e supponiamo che il processo  $\{\mathbf{y}(t)\}$  sia ergodico, per il teorema di Birkhoff il limite

$$\lim_{T \rightarrow \infty} \nu_T(E)$$

esiste con probabilità 1 (ovvero *per tutte le possibili traiettorie del processo, eccettuato al più un insieme di traiettorie di probabilità zero*) e vale:

$$E I_E(\mathbf{y}(0)) = \int_E dF(y) = P(E) \quad ,$$

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  $\nu_T((-\infty, a])$  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

$$\nu_T(E_1, E_2, k) := \frac{1}{T+1} \sum_{t=0}^T I_{E_1}(\mathbf{y}(t)) I_{E_2}(\mathbf{y}(t+k))$$

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:

$$\begin{aligned} \lim_{T \rightarrow \infty} \nu_T(E_1, E_2, k) &= E [I_{E_1}(\mathbf{y}(0)) I_{E_2}(\mathbf{y}(k))] \\ &= \int_{E_1} \int_{E_2} dF(y_1, y_2; k) = P\{\mathbf{y}(t) \in E_1, \mathbf{y}(t+k) \in E_2\} \end{aligned}$$

per “quasi tutte” le traiettorie del processo.

Una generalizzazione ormai facile porge allora la seguente conclusione.

**Theorem 4.1.** *Se il processo  $\{y(t)\}$  è ergodico, la conoscenza di una sola traiettoria è (con probabilità 1) sufficiente a determinare univocamente la legge di probabilità dell'intero processo.*

## 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 [18, 51].

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

$$dF(x) := \frac{1}{b-a} dx$$

e ridefinire opportunamente  $f$ . Se  $I$  è un intervallo infinito, si può usare la stessa tecnica introducendo un'opportuna densità fittizia (ad esempio ponendo  $dF(x) = e^{-x^2/2} dx$ ), che andrà poi “scalata” dalla funzione  $f$ .

Ci si riduce quindi al calcolo della media,  $E f(y)$ , di una funzione (nota) della variabile casuale  $y$  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,  $\{z_1, z_2, \dots\}$  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.  $\{z(t)\}$ , in cui  $z(t)$  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.  $\{y(t)\}$ , nel quale la variabile generica  $y(t)$  è distribuita secondo la  $F(x)$ . (La verifica è banale:  $P(y \leq x) = P(F^{-1}(z) \leq x) = P(z \leq F(x)) = F(x)$ , se  $z$  è uniformemente distribuita).

Una facile applicazione del teorema ergodico porge quindi:

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_1^N f(y_k) = E f(\mathbf{y}) = \int_I f(x) dF(x)$$

e questa formula fornisce un metodo generale per il calcolo approssimato di aspettative 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 i 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 probabilità ausiliaria scelta in modo opportuno. Si supponga ad esempio di dover calcolare l’integrale  $\int_I f(x)p(x)dx$  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  $q(x)$  con supporto (diversa da zero sul) l’intervallo  $I$  ed effettuare il “cambio di misura” descritto dalla formula

$$\int_I f(x)p(x)dx = \int_I f(x)\frac{p(x)}{q(x)}q(x)dx := \int_I g(x)q(x)dx$$

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*. Naturalmente queste tecniche funzionano nel caso di variabili scalari; il campionamento di una densità multivariata è una questione che va affrontata con idee diverse.

I famosi articoli [41] e [25] hanno portato ad un notevole progresso nel campo della simulazione Montecarlo. L’idea di base è stata una tecnica di campionamento che si applica a distribuzioni multivariabili, usando la convergenza di catene di Markov (a tempo discreto) alla distribuzione invariante. Come abbiamo visto, una catena di Markov con un solo insieme di stati ergodici è un processo ergodico per cui vale il teorema di Birkhoff. Ammettendo di voler calcolare l’aspettazione di una funzione del tipo  $\int f(x)\pi(x)dx$  si può generare (i.e. simulare) una catena di Markov ergodica che abbia come distribuzione invariante (necessariamente unica) proprio la  $\pi(x)$ . Per far questo occorre saper costruire una matrice (o, più in generale, un nucleo di probabilità) di transizione che ammetta proprio  $\pi(x)$  come misura invariante. Si dimostra che ci sono in realtà infinite catene ergodiche che hanno  $\pi$  come probabilità invariante e l’articolo [25] descrive un possibile metodo di costruirne una. Fatto questo, si tratta di simulare la catena generando successivamente le variabili di una traiettorie  $\{x(t); t = 1, 2, \dots\}$ . 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 sistemi di calcolo moderni, i metodi

d'elezione per risolvere problemi, come ad esempio il filtraggio non lineare, che solo un decennio fa sembravano innavvicinabili.

Notiamo per ultimo che la qualificazione "con probabilità 1" che va associata alla formula (4.3.9) è molto più di effetto psicologico che reale.

### 4.4 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

$$\lim_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T f_t(\mathbf{y}) g(\mathbf{y}) = [E f(\mathbf{y})] g(\mathbf{y}) \quad (4.4.1)$$

con probabilità uno, qualunque siano  $f$  e  $g$  nella classe appena descritta.

Sceghieremo  $f$  e  $g$  nel seguente modo:

$$f(\mathbf{y}(t_1) \dots \mathbf{y}(t_n)) := I_{A_1}(\mathbf{y}(t_1)) \dots I_{A_n}(\mathbf{y}(t_n))$$

$$g(\mathbf{y}(\tau_1) \dots \mathbf{y}(\tau_m)) := I_{B_1}(\mathbf{y}(\tau_1)) \dots I_{B_m}(\mathbf{y}(\tau_m)) \quad ,$$

dove  $A_1, \dots, A_n$  e  $B_1, \dots, B_m$  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 (4.4.1), ottenendo

$$\lim_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T E [f_t(\mathbf{y}) \cdot g(\mathbf{y})] = E f(\mathbf{y}) \cdot E g(\mathbf{y}) \quad ,$$

la quale si può scrivere

$$\lim_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T P(A_t \cap B) = P(A)P(B) \quad , \quad (4.4.2)$$

dove si è usata la notazione

$$A_t := \{\omega ; \mathbf{y}(t+t_1) \in A_1, \dots, \mathbf{y}(t+t_n) \in A_n\}$$

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  $A = \{\omega ; \mathbf{y}(t_1) \in A_1, \dots, \mathbf{y}(t_n) \in A_n\}$ .

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 (4.4.2) si può ugualmente scrivere  $P(A_t)P(B)$  al posto di  $P(A)P(B)$ . È possibile allora interpretare la (4.4.2) come una condizione di *independenza asintotica* delle variabili del processo. Infatti se il processo è ergodico, la probabilità congiunta degli eventi  $A_t$  e  $B$  converge per  $t \rightarrow \infty$ , 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 *independenti* quando  $t \rightarrow +\infty$ .

Il tipo di convergenza è però molto debole. In un certo senso, una relazione del tipo

$$\lim_{t \rightarrow \pm\infty} P(A_t \cap B) = P(A)P(B) \quad (4.4.3)$$

sarebbe interpretabile in modo molto più diretto. Come vedremo, questa condizione ci riporta alla nozione di processo p.n.d. definita più sopra.

In questa sezione è necessario usare il concetto di  $\sigma$ -algebra di eventi in modo un poco più approfondito di quanto abbiamo fatto finora. Rinviamo allo scopo il lettore all'Appendice A.1.1.

Introduciamo le  $\sigma$ -algebre degli eventi passati,  $\mathcal{Y}_t^-$ , e futuri,  $\mathcal{Y}_t^+$ , del processo  $\mathbf{y}$  all'istante  $t$ , come le  $\sigma$ -algebre generate, rispettivamente dalle variabili passate  $\{\mathbf{y}(s); s \leq t\}$  e future,  $\{\mathbf{y}(s); s \geq t\}$  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  $t_1, t_2, \dots, t_n \leq t$  ed è invece un *evento futuro all'istante  $t$*  se, viceversa,  $t_1, t_2, \dots, t_n \geq t$ . Ne segue che una nozione equivalente di processo p.n.d. è un processo per cui le  $\sigma$ -algebre degli eventi *infinitamente passati* e *infinitamente futuri*

$$\mathcal{Y}_\infty^- := \bigcap_{t \leq k} \mathcal{Y}_t^-, \quad \mathcal{Y}_\infty^+ := \bigcap_{t \geq k} \mathcal{Y}_t^+ \quad (4.4.4)$$

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).

**Theorem 4.4.** *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*

$$\lim_{t \rightarrow -\infty} |P(A_t \cap B) - P(A_t)P(B)| = 0, \quad (4.4.5)$$

*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*

$$\lim_{t \rightarrow +\infty} |P(A_t \cap B) - P(A_t)P(B)| = 0, \quad (4.4.6)$$

*vale se e solo se  $L_\infty^+(\mathbf{y})$  contiene solo v.a. costanti. Le (4.4.5) (4.4.6) valgono contemporaneamente se e solo se il processo  $\mathbf{y}$  è p.n.d. (in senso stretto).*

**Proof.** È sufficiente occuparci solo della (4.4.5), 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 (4.4.5) 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  $\sigma$ -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

$$\xi_t := I_{A_t} - P(A_t), \quad \eta := I_B - P(B)$$

che appartengono entrambe allo spazio (di Hilbert)  $L^2(\mathbf{y})$  e definiamo la proiezione ortogonale di  $\eta$  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:

$$\eta_t := P\{\eta \mid L_t^{2-}(\mathbf{y})\} = E\{\eta \mid \mathcal{Y}_t^-\}$$

Evidentemente, dato che  $\xi_t \in L_t^{2-}(\mathbf{y})$ , si ha

$$|\langle \xi_t, \eta \rangle| = |\langle \xi_t, \eta_t \rangle| \leq \|\xi_t\|_2 \|\eta_t\|_2$$

dove le norme sono quelle di  $L^2(\mathbf{y})$ . Sostituendo le espressioni delle variabili al primo membro si trova

$$|P(A_t B) - P(A_t)P(B)| \leq (P(A_t) - P(A_t)^2) \|\eta_t\|_2 \leq \|\eta_t\|_2. \quad (*)$$

Ora, dato che per  $t \rightarrow -\infty$ ,  $L_t^{2-}(\mathbf{y})$  tende monotonicamente a restringersi ad un sottospazio che contiene solo costanti (equivalentemente, la  $\sigma$ -algebra  $\mathcal{Y}_t^-$  tende alla  $\sigma$ -algebra banale), si ha

$$\lim_{t \rightarrow -\infty} \eta_t = E \eta = 0$$

e quindi anche  $\lim_{t \rightarrow -\infty} |P(A_t B) - P(A_t)P(B)| = 0$ .  $\square$

**Remark 4.2.** 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

$$\lim_{t \rightarrow -\infty} \sup_{A_t \in \mathcal{Y}_t^-} |P(A_t \cap B) - P(A_t)P(B)| = 0. \quad (4.4.7)$$

Dualmente, la (4.4.6) può essere rafforzata nella,

$$\lim_{t \rightarrow +\infty} \sup_{A_t \in \mathcal{Y}_t^+} |P(A_t \cap B) - P(A_t)P(B)| = 0. \quad (4.4.8)$$

Notare che le successioni  $\{A_t\}$  possono essere qualsiasi.

**Corollary 4.2.** Per un processo p.n.d. la quantità

$$\alpha(\tau) := \sup_{\substack{A_t \in \mathcal{Y}_t^- \\ B_t \in \mathcal{Y}_t^+}} |P(A_t \cap B_{t+\tau}) - P(A_t)P(B_{t+\tau})| \quad (4.4.9)$$

(è indipendente da  $t$  e) tende a zero per  $\tau \rightarrow +\infty$ .

**Proof.** Per la stazionarietà si può fissare  $t$  in modo arbitrario (ad esempio  $t = 0$ ) e la (4.4.9) 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  $\tau \rightarrow +\infty$ , 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 (4.4.5) del teorema 4.4 per cui si avrebbe  $L_\infty^-(\mathbf{y})$  non banale e il processo non potrebbe essere p.n.d.  $\square$

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 (4.4.7, 4.4.8) 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 (4.4.3) è

$$\lim_{t \rightarrow \infty} E [f_t(\mathbf{y}) g(\mathbf{y})] = E f(\mathbf{y}) E g(\mathbf{y}) \quad ,$$

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 (4.4.3) implica quindi l'ergodicità.

## 4.5 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 del second'ordine* o *debole* che è sufficiente per studiare i casi che si presentano più comunemente nell'analisi asintotica degli algoritmi di identificazione di sistemi lineari.

**Definition 4.5.** Sia  $\{\mathbf{y}(t)\}$  un processo  $m$ -dimensionale stazionario in senso debole di media  $\mu$  e matrice di covarianza  $\Sigma(\tau)$ . Il processo si dice *ergodico del secondo ordine* (o *debolmente ergodico*) se la media campionaria

$$\bar{\mathbf{y}}_T := \frac{1}{T+1} \sum_0^T \mathbf{y}(t) \tag{4.5.1}$$

e la varianza campionaria del processo

$$\mathbf{S}_T(\tau) := \frac{1}{T+1} \sum_{t=0}^T [\mathbf{y}(t+\tau) - \bar{\mathbf{y}}_T] [\mathbf{y}(t) - \bar{\mathbf{y}}_T] \quad (4.5.2)$$

convergono ai valori veri

$$\lim_{T \rightarrow \infty} \bar{\mathbf{y}}_T = \boldsymbol{\mu} \quad (4.5.3)$$

$$\lim_{T \rightarrow \infty} \mathbf{S}_T(\tau) = \boldsymbol{\Sigma}(\tau) \quad (4.5.4)$$

con probabilità uno.

È ovvio che per un processo ergodico la (4.5.3) segue direttamente dal teorema di Birkhoff ponendo  $f(\mathbf{y}) = \mathbf{y}(0)$  mentre per provare che vale la (4.5.4) basta scrivere

$$\begin{aligned} (T+1) \mathbf{S}_T(\tau) &= \sum_0^T [\mathbf{y}(t+\tau) - \boldsymbol{\mu} + \boldsymbol{\mu} - \bar{\mathbf{y}}_T] [\mathbf{y}(t) - \boldsymbol{\mu} + \boldsymbol{\mu} - \bar{\mathbf{y}}_T]' \\ &= \sum_0^T [\mathbf{y}(t+\tau) - \boldsymbol{\mu}] [\mathbf{y}(t) - \boldsymbol{\mu}]' - (\bar{\mathbf{y}}_T - \boldsymbol{\mu}) \sum_0^T [\mathbf{y}(t) - \boldsymbol{\mu}]' \\ &\quad - \left( \sum_0^T [\mathbf{y}(t+\tau) - \boldsymbol{\mu}] \right) (\bar{\mathbf{y}}_T - \boldsymbol{\mu})' + (T+1) (\bar{\mathbf{y}}_T - \boldsymbol{\mu}) (\bar{\mathbf{y}}_T - \boldsymbol{\mu})'. \end{aligned}$$

e notare che gli ultimi tre termini divisi per  $T+1$  tendono a zero con probabilità uno per  $T \rightarrow \infty$ .

**Nota** Qualche volta il limite superiore nella sommatoria che compare in (4.5.2) è posto uguale a  $T - \tau$ . Si riconosce facilmente che la (4.5.4) continua a valere anche con questa diversa definizione della varianza campionaria.  $\diamond$

In generale un processo può essere ergodico del second'ordine sotto condizioni più deboli dell'ergodicità. In particolare esistono condizioni per l'ergodicità del secondo ordine che riguardano processi generati come uscita di sistemi lineari eccitati da rumore bianco che sono applicabili direttamente ai problemi di analisi asintotica degli algoritmi di identificazione che studieremo nel capitolo 6.

Ricordiamo a questo proposito il teorema 4.3 che si può interpretare dicendo che l'uscita  $\mathbf{y}$  di un sistema lineare tempo-invariante  $\ell^2$ -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, è in particolare anche ergodico in senso debole.

La condizione che il processo di ingresso sia i.i.d. può essere rilassata. Nel testo di Hannan [22, Cap. IV] e in [24, 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 (debolmente stazionario)  $\{e(t)\}$  che è stazionario del quart'ordine (con momenti fino al quart'ordine invarianti per traslazione) e soddisfa alle condizioni,

$$\mathbb{E} \|e(t)\|^4 < \infty \tag{4.5.5}$$

$$\mathbb{E}[e(t) | e^{t-1}] = 0 \quad t \in \mathbb{Z} \tag{4.5.6}$$

$$\text{Var}[e(t) | e^{t-1}] = \Lambda < \infty \tag{4.5.7}$$

dove  $e^{t-1}$  è la sequenza infinita  $\{e(t-1), e(t-2), \dots\}$ . Queste condizioni sono in genere più deboli della condizione di essere i.i.d..

Processi di questo tipo si chiamano *d-martingale* e verranno studiati un pò più in dettaglio nel capitolo 5

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.

**Theorem 4.5.** *Per un processo Gaussiano stazionario  $m$ -dimensionale  $\{y(t)\}$ , le seguenti condizioni sono fra loro equivalenti:*

1. Il processo è ergodico.
2. Il processo è ergodico del secondo ordine (cioè valgono le (4.5.3) e (4.5.4) con probabilità 1).
3. La matrice distribuzione spettrale di potenza del processo,  $F(e^{i\omega})$ , è una funzione continua di  $\omega$  in  $[-\pi, \pi]$ .
4. Si ha incorrelazione asintotica delle variabili del processo, nel senso delle medie di Cesàro,

$$\lim_{T \rightarrow \infty} \frac{1}{T+1} \sum_0^T \sigma_{ii}^2(\tau) = 0 \quad i = 1, 2, \dots, m,$$

dove le  $\sigma_{ii}^2(\tau)$  sono le funzioni covarianza delle componenti  $y_i(t)$ .

Di queste condizioni la più utile è probabilmente la (3), dovuta a Maruyama e Grenander (per una trattazione precisa e completa vedere [54, 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  $\omega_0$ , 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  $\omega_0$ ), per cui il processo avrebbe potenza finita associata alla frequenza  $\omega_0$  (una "riga" spettrale per  $\omega = \omega_0$ ).

**Problem 4.2.** *Si consideri il processo  $z(t) = x \cos \omega_0 t + y \sin \omega_0 t$ ,  $t \in \mathbb{Z}$ , dove  $x$  e  $y$  sono variabili aleatorie scalari Gaussiane di media zero e uguale varianza  $\sigma^2$ , 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

$$\lim_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T \bar{z}(t+\tau) \bar{z}(t) \quad ,$$

con  $\tau$  fissato, e si verifichi che è diverso da  $E[\mathbf{z}(t+\tau) \mathbf{z}(t)]$ . □

Una semplice condizione sufficiente per la validità della (4) è l'incorrelazione asintotica

$$\lim_{\tau \rightarrow \infty} \sigma_{ii}(\tau) = 0 \quad , \quad i = 1, 2, \dots, m \quad , \quad (4.5.8)$$

nel qual caso si riconosce facilmente che la componente  $i$ -sima del processo,  $\{y_i(t)\}$ , non può contenere componenti oscillatorie di ampiezza finita. Questo è, ovviamente, in accordo con la condizione (3).

Notiamo, di passaggio, che la (4.5.8) è equivalente alla

$$\lim_{\tau \rightarrow \infty} \Sigma(\tau) = 0.$$

Segnaliamo infine che

**Corollary 4.3.** *Un processo Gaussiano stazionario e p.n.d in senso debole, lo è anche in senso forte e quindi è in particolare ergodico.*

## 4.6 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 (ad esempio misure ingresso-uscita su un impianto), diciamoli  $\{\bar{y}(t)\}_{t=0,1,\dots,T}$ , 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  $\{\bar{y}(t)\}_{t=0,\dots,T}$  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  $T \rightarrow \infty$ ) è sufficiente che il "modello"  $\{\mathbf{y}(t)\}$  dei dati osservati  $\{\bar{y}(t)\}_{t=0,1,\dots,T}$  sia un *processo ergodico*. Si fa perciò l' "ipotesi" che i dati osservati siano generati da un processo ergodico. Questa ipotesi, anche se matematicamente sensata, 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.

Ricordiamo a questo proposito quanto già detto nella prefazione: per noi lo scopo dell'identificazione e, in generale dell'inferenza statistica, è quello di produrre modelli che serviranno per descrivere dati "futuri", ed quindi dati *diversi* da quelli usati per la loro calibrazione. Alle radici di ogni esperimento di modellizzazione deve esserci quindi 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 appena visto, la risolubilità univoca del problema di identificazione, viene garantita dall'ergodicità.

Nel contesto di un esperimento reale si hanno a disposizione 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 (deterministico) a tempo discreto, che per semplicità di trattazione supporremo scalare. Lascieremo al lettore la generalizzazione al caso di segnali a valori vettoriali.

Una *funzione di  $z$*  è una funzione a valori reali  $f(z) := f(z(t); t \in \mathbb{I})$ ,  $f : \mathbb{R}^I \rightarrow \mathbb{R}$  dove  $\mathbb{I}$  è un sottointervallo di  $\mathbb{Z}_+$ , possibilmente infinito. L'*operatore di traslazione*  $\sigma$  sui segnali è definito da  $[\sigma z](t) := z(t+1)$ ,  $t \in \mathbb{Z}_+$  di modo che l'applicazione iterata di  $\sigma$ , e.g.

$$[\sigma^t z](s) := z(t+s), \quad t, s \in \mathbb{Z}$$

trasforma 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.  $f_t(z) := f(z(t+s); s \in I) = f(z_t)$ ,  $t \in \mathbb{Z}$ .

**Definition 4.6.** *Un segnale  $z$  si dice*

- *Stazionario in senso stretto se il limite in media (di Cesàro)*

$$\lim_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T f_t(z) \quad (4.6.1)$$

esiste per tutte le funzioni  $f$  di una classe sufficientemente ampia, ad esempio, le funzioni limitate misurabili.

- Stazionario in senso lato, o del second'ordine<sup>14</sup> se il limite esiste almeno per  $f(z) = z(0)$  (cosicchè  $f_t(z) = z(t)$ ) e per le correlazioni campionarie; i.e.

$$\lim_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T z(t+\tau)z(t) := r(\tau) \tag{4.6.2}$$

qualunque sia  $\tau \geq 0$ .

È facile verificare che se esiste il limite (4.6.1) allora, qualunque sia l'istante  $t_0 \geq 0$ , il limite

$$\lim_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=t_0}^{t_0+T} f_t(z)$$

esiste e coincide con il limite

$$\lim_{T \rightarrow \infty} \frac{1}{T+t_0+1} \sum_{t=0}^{t_0+T} f_t(z)$$

che è lo stesso definito in (4.6.1). Questa proprietà è l'*uniformità statistica dei dati futuri*.

I segnali stazionari in senso lato sono quelli per cui la media e la covarianza campionaria basate su una serie temporale di  $T$  dati, convergono quando  $T \rightarrow \infty$ . 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 traiettoria  $z$  secondo la legge di probabilità di un processo stazionario.

**Proposition 4.6.** *Dato un segnale stazionario (in senso stretto)  $z$ , esiste uno spazio di probabilità  $\{\Omega, \mathcal{A}, \mu\}$  e un processo stocastico stazionario definito su questo spazio,  $\mathbf{z} := \{z(t, \omega) \mid t \in \mathbb{Z}, \omega \in \Omega\}$ , tale che  $z$  è una traiettoria rappresentativa di  $\mathbf{z}$ , i.e.*

$$z(t) = \mathbf{z}(t, \bar{\omega}) \quad t \in \mathbb{Z}$$

per un qualche evento elementare  $\bar{\omega}$  appartenente all'insieme di probabilità uno in cui si ha convergenza delle medie campionarie (4.3.7), come stabilito dal teorema di Birkhoff 4.1 .

<sup>14</sup>Concetto che abbiamo già introdotto; vedere la definizione 3.3 del capitolo 3.

**Proof.** \* 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

$$\nu_T(A) := \frac{1}{T+1} \sum_{t=0}^T I_A(z(t)) \tag{4.6.3}$$

che rappresenta la frequenza relativa delle visite del segnale  $z$  all'insieme  $A$ . Si verifica facilmente che per ogni  $T$  la funzione  $A \rightarrow \nu_T(A)$  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,  $\nu_T(A) \rightarrow \nu_0(A)$  per  $T \rightarrow \infty$ . Passando al limite, si verifica quindi facilmente che *La funzione d'insieme  $A \rightarrow \nu_0(A)$  è una misura di probabilità su  $\mathbb{R}$ .*

Questa misura è poi invariante per traslazione, perchè sostituendo a  $z$  il segnale  $\sigma^s z$  la somma (4.6.3) 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  $\tau_1$ , 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  $\nu_T(A_1, \tau_1, \dots, A_n, \tau_n) \rightarrow \nu_n(A_1, \tau_1, \dots, A_n, \tau_n)$  quando  $T \rightarrow \infty$  e il limite dipende in realtà solo dalle differenze  $\tau_2 - \tau_1, \dots, \tau_n - \tau_1$ . Quanto appurato finora per la distribuzione di probabilità  $\nu_0(A)$ , si generalizza quindi nel seguente risultato.

**Lemma 4.2.** *Per tutti gli  $n$  e per arbitrari tempi  $\tau_1 \dots \tau_n$ , la funzione d'insieme  $(A_1 \times \dots \times A_n) \rightarrow \nu_n(A_1, \tau_1, \dots, A_n, \tau_n)$  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 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  $A_1, \dots, A_{n-1}$  e possibili istanti  $\tau_1 \dots, \tau_n$ .

Pertanto, per un famoso teorema di Kolmogorov [29], esiste un'unica misura di probabilità  $\mu$  sullo spazio campionario  $\mathbb{R}^{\mathbb{Z}}$  delle sequenze reali, che è l'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  $\sigma$  agente sui segnali di  $\mathbb{R}^Z$ . In altre parole,  $(\mathbb{R}^Z, \mathcal{Z}, \mu)$  (dove  $\mathcal{Z}$  è la  $\sigma$ -algebra degli insiemi di Borel) definisce un *processo stocastico stazionario*,  $z$ .  $\square$

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.

**Definition 4.7.** *Chiameremo la legge di probabilità costruita nella dimostrazione della proposizione 4.6, la **legge di probabilità vera** del processo.*

Quanto abbiamo appena esposto per segnali stazionari in senso stretto è evidentemente di interesse prevalentemente concettuale. In pratica si possono spesso 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 del second'ordine (o in senso debole)* sarà una successione  $\{z(t)\}$  per la quale il limite

$$\lim_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T z(t+\tau)z(t)^\top := \Lambda_0(\tau) \quad (4.6.4)$$

esiste per tutti i  $\tau \in \mathbb{Z}$ . A questo proposito vale il seguente risultato, che si dimostra in modo analogo a quanto visto al capitolo 3.

**Proposition 4.7 (Wiener).** *Se esiste, per ogni  $\tau \geq 0$ , il limite (4.6.4), la funzione  $\Lambda_0 := \tau \rightarrow \Lambda_0(\tau)$  è una funzione matriciale di tipo positivo; i.e. una funzione covarianza.*

Questa covarianza sarà chiamata la **covarianza vera** del processo. In analogia a quanto visto nella proposizione 4.6 ogni segnale stazionario in senso debole si può pensare come una traiettoria rappresentativa di un *processo vero* (del second'ordine) di covarianza  $\Lambda_0$ .

La verifica dell'ergodicità (del second'ordine) del processo "vero" può così essere ricondotta all'enunciato del teorema 4.5. In particolare, condizione necessaria per l'ergodicità è che lo spettro associato alla covarianza vera  $\Lambda_0$  non contenga righe. Evidentemente questa verifica può essere fatta solo asintoticamente.

## 4.7 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.

**Theorem 4.6 (Wald ).** Sia  $\{p(\cdot, \theta) ; \theta \in \Theta\}$  una famiglia parametrica di densità di probabilità su  $\mathbb{R}^m$  in cui  $\Theta$  è un dominio (non necessariamente limitato) di  $\mathbb{R}^p$ . Sia  $\hat{\theta}(\mathbf{y}^n)$  lo stimatore di M.V. del parametro  $\theta$  basato sul campione casuale  $\mathbf{y}^n := \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$  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  $\theta_0$ .
2.  $p(y, \theta)$  è una funzione continua di  $\theta$  per ogni  $y$ .
3.  $E_{\theta} \log p(\mathbf{y}, \theta')$  è finita per ogni  $\theta$  e  $\theta'$  in  $\Theta$ .
4. Al crescere di  $n$  la successione delle stime  $\hat{\theta}(y^n)$ ,  $n = 1, 2, \dots$ , si mantiene limitata con probabilità  $P_{\theta_0}$  uguale a 1 (ovvero,  $|\hat{\theta}(y^n)| \leq M$  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

$$\lim_{n \rightarrow \infty} \hat{\theta}(y^n) = \theta_0 \quad ,$$

per tutte le successioni di valori campionari  $\{y_1, y_2, \dots\}$ , eccettuato al più un insieme di probabilità  $P_{\theta_0}$  uguale a zero.

Notiamo che a stretto rigore potrebbero esserci più punti in  $\Theta$  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.

**Proof.**

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  $\theta$  non contenente  $\theta_0$ , è possibile mostrare che la distanza di  $p(\cdot, \theta_0)$  dall'insieme di densità  $\{p(\cdot, \theta) ; \theta \in E\}$ , definita da

$$I(\theta_0, E) := E_{\theta_0} \left\{ \min_{\theta \in E} \log \frac{p(\mathbf{y}, \theta_0)}{p(\mathbf{y}, \theta)} \right\} \quad , \quad \theta_0 \notin E \quad ,$$

è ancora strettamente positiva (e finita). La cosa segue in sostanza dalla continuità

4.7. Consistenza dello Stimatore di Massima Verosimiglianza\*

di  $p(\cdot, \theta)$ . Usando semplici proprietà della funzione  $\log(\cdot)$  si trova poi

$$\begin{aligned} I(\theta_0, E) &= E_{\theta_0} \left\{ \log \frac{p(\mathbf{y}, \theta_0)}{\max_{\theta} p(\mathbf{y}, \theta)} \right\} \\ &= E_{\theta_0} \left\{ \log p(\mathbf{y}, \theta_0) - \log [\max_{\theta} p(\mathbf{y}, \theta)] \right\} \\ &= E_{\theta_0} \left\{ \log p(\mathbf{y}, \theta_0) - \max_{\theta} [\log p(\mathbf{y}, \theta)] \right\} > 0 \quad . \quad (4.7.1) \end{aligned}$$

Sia  $L(\mathbf{y}^n, \theta)$  la verosimiglianza del campione. Per la legge forte dei grandi numeri,

$$\frac{1}{n} \log L(\mathbf{y}^n, \theta_0) = \frac{1}{n} \sum_1^n \log p(\mathbf{y}_t, \theta_0) \rightarrow E_{\theta_0} \log p(\mathbf{y}, \theta_0) \quad ,$$

con probabilità  $P_{\theta_0}$  uguale a 1 per  $n \rightarrow \infty$ .

Ugualmente la successione di variabili aleatorie

$$\mathbf{z}_t := \max_{\theta \in E} \log p(\mathbf{y}_t, \theta) \quad , \quad t = 1, 2, \dots ,$$

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

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_1^n \mathbf{z}_t = E_{\theta_0} \max_{\theta} [\log p(\mathbf{y}, \theta)] \quad ,$$

con probabilità 1. La disuguaglianza (stretta!) (4.7.1) 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

$$\log L(\mathbf{y}^n, \theta_0) > \max_{\theta \in E} \log L(\mathbf{y}^n, \theta) \quad , \quad \theta_0 \notin E \quad , \quad (4.7.2)$$

per  $n \geq \bar{n}$  opportuno,<sup>15</sup> per quasi tutte le successioni di valori campionari  $\{y_1, \dots\}$ . Notiamo ora che se un insieme  $A$  (chiuso e limitato) è ricoperto dall'unione di un numero finito di intorni sferici  $E_1, \dots, E_N$  in  $\Theta$  e se  $f(\theta)$  è un'arbitraria funzione continua in  $A$ , si ha

$$\max_{\theta \in A} f(\theta) \leq \max_k \left\{ \max_{\theta \in E_k} f(\theta) \right\} \quad , \quad k = 1, \dots, N$$

<sup>15</sup>È bene mettere in guardia il lettore che tanto più vicino a  $\theta_0$  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 (4.7.2). Per  $\theta_0 \in E$  essa potrebbe valere soltanto per "n = ∞" e col segno  $\geq$ , ma in questo caso i due termini che si confrontano non sono più definiti.

cosicché la (4.7.2) vale anche per un *arbitrario* insieme chiuso e limitato  $A$  che non contenga  $\theta_0$ , ovvero

$$\log L(y^n, \theta_0) > \max_{\theta \in A} \log L(y^n, \theta) \quad , \quad \theta_0 \notin A \quad , \quad (4.7.3)$$

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

$$\log L(y^n, \hat{\theta}(y^n)) \geq \log L(y^n, \theta_0) \quad , \quad \forall n \quad . \quad (4.7.4)$$

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  $\Theta$  (che dipende in generale dalla successione di dati osservata). Questo naturalmente a meno di casi "sfortunati" che però hanno probabilità zero. Si ha così

$$\log L(y^n, \hat{\theta}(y^n)) = \max_{\theta \in C} \log L(y^n, \theta)$$

per un opportuno insieme  $C$  (che è fisso al variare di  $n$ ). Consideriamo ora l'intorno sferico di  $\theta_0$ ,  $S(\varepsilon) := \{\theta ; |\theta - \theta_0| < \varepsilon\}$  e definiamo l'insieme  $A(\varepsilon) := C - S(\varepsilon)$ . Ovviamente per  $\varepsilon$  abbastanza piccolo  $A(\varepsilon)$  è ancora chiuso e limitato. Applicando la disuguaglianza (4.7.3) 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 (4.7.4) si perviene allora alla

$$\log L(y^n, \hat{\theta}(y^n)) > \max_{\theta \in A(\varepsilon)} \log L(y^n, \theta)$$

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,

$$|\hat{\theta}(y^n) - \theta_0| < \varepsilon \quad \text{per} \quad n \geq \bar{n}(\varepsilon) \quad ,$$

il che equivale a  $\lim_{n \rightarrow \infty} \hat{\theta}(y^n) = \theta_0$ , naturalmente a meno di un insieme eccezionale di successioni campionarie di probabilità zero.  $\square$

### 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

$\{y\}$ . 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  $|\theta| \rightarrow \infty$ , 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 [72, 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 indicata (oltre che da  $\theta$ ) 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  $\mathcal{F}_1, \mathcal{F}_2, \dots$  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  $\theta$  in  $\Theta_{k+1}$ .
- 2) Sia  $F_0$  la distribuzione *vera* dei dati osservati.<sup>16</sup> Assumeremo che  $F_0$  possa essere approssimata con accuratezza arbitraria da (almeno) un elemento  $F_{\theta_0}^{(k)} \in \mathcal{F}_k$  pur di prendere  $k$  abbastanza grande. In altre parole, per ogni  $\varepsilon > 0$ ,

<sup>16</sup>Per non complicare troppo le notazioni supporremo qui che i dati osservati  $y_1, y_2, \dots, y_n, \dots$  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*  $\{y_n\}$  e parlare, anziché di distribuzione di probabilità del generico vettore del campione  $y_n$ , di *leggi di probabilità* dell'intero processo  $\{y_n\}$ . 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.

l'estremo superiore

$$\sup_y |F_0(y) - F_\theta^{(k)}(y)|$$

può essere reso minore di  $\varepsilon$  pur di prendere  $k$  abbastanza grande e di scegliere opportunamente  $\theta$  in  $\Theta_k$ .

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  $\phi_n$  del parametro  $\theta$  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  $\theta \in \Theta_k$ ,  $k = 1, 2, \dots$ , 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 computer) 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)  $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ , distribuiti secondo  $F_{\theta_0}^{(k)} \in \mathcal{F}_k$ , si ha  $\lim_{n \rightarrow \infty} \phi_n^{(k)}(\mathbf{y}_1, \dots, \mathbf{y}_n) = \theta_0$  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  $\theta$  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  $\phi_\infty^{(k)}(\mathbf{y}_1, \dots, \mathbf{y}_n, \dots) = \theta$ , 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.

**Theorem 4.7.** *Siano  $\phi_n^{(k)}(\mathbf{y}_1, \dots, \mathbf{y}_n)$  stimatori intrinsecamente consistenti dal parametro  $\theta$  basati ciascuno sulla famiglia parametrica  $\mathcal{F}_k = \{F_\theta^{(k)}; \theta \in \Theta_k\}$ ,  $k = 1, 2, \dots$ , e operanti*

#### 4.7. Consistenza dello Stimatore di Massima Verosimiglianza\*

121

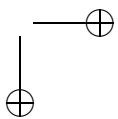
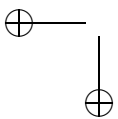
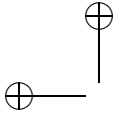
su un campione casuale di numerosità  $n$  estratto dalla distribuzione vera  $F_0$ .

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  $\theta \rightarrow F_\theta^{(k)}(\cdot)$  è continua rispetto alla metrica  $\|F_1 - F_2\| := \sup_y |F_1(y) - F_2(y)|$ , allora

$$\lim_{k \rightarrow \infty} \left( \lim_{n \rightarrow \infty} \|F_{\hat{\phi}_n^{(k)}(\mathbf{y}_1, \dots, \mathbf{y}_n)}^{(k)} - F_0\| \right) = 0 \quad .$$

Se la consistenza di  $\hat{\phi}_n^{(k)}$  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 [?].





## CHAPTER 5

IL TEOREMA DEL LIMITE  
CENTRALE

## 5.1 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)

**Proposition 5.1.** *La successione di v.a.  $\{\mathbf{x}_n\}$  converge in legge a  $\mathbf{x}$  se e solo se  $E g(\mathbf{x}_n) \rightarrow E g(\mathbf{x})$  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  $\omega \in \mathbb{R}^{17}$ .

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 5.1.** *Se  $\{\mathbf{x}_n\}$  è una successione di variabili aleatorie convergente in legge; i.e.*

<sup>17</sup>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).

$\mathbf{x}_n \xrightarrow{L} \mathbf{x}$ , che è uniformemente integrabile, in particolare se

$$\sup_n \|\mathbf{x}_n\|^2 < \infty \tag{5.1.1}$$

allora tutti i momenti che esistono delle  $\mathbf{x}_n$  convergono ai rispettivi momenti della distribuzione limite.

Per la prova vedere [4, p. 32-33]. Si veda a questo proposito anche l'osservazione 5.3 in margine al teorema 5.3.

Il seguente enunciato raccoglie alcune proprietà generali della convergenza in legge che torneranno utili nel seguito di questo capitolo. Per la dimostrazione si veda [17, Cap. 6].

**Theorem 5.1 (Slutsky).** *Si assuma che la sequenza di vettori aleatori  $n$ -dimensionali  $\{\mathbf{x}_N; N = 1, 2, \dots\}$  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  $(\mathbf{x}_N - \mathbf{y}_N) \rightarrow 0$  in probabilità, allora anche  $\mathbf{y}_N$  converge in legge a  $\mathbf{x}$  (ovvero  $\mathbf{y}_N \xrightarrow{L} \mathbf{x}$ ).*
2. *Se  $f : \mathbb{R}^n \rightarrow \mathbb{R}^k$  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  $\{\mathbf{y}_N; N = 1, 2, \dots\}$  converge in legge (o in probabilità) ad una costante  $c$  e se  $f(x) := f(z, y) : \mathbb{R}^{p+m} \rightarrow \mathbb{R}^k$  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  $(\mathbf{x}_N - \mathbf{y}_N) \rightarrow 0$  in probabilità, si dicono **asintoticamente equivalenti**.

**Example 5.1.** *Sia  $\mathbf{y}$  un processo ergodico scalare a media  $\mu$ , varianza  $\sigma^2$  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*

$$\varphi(\mathbf{y}) := \frac{\sqrt{N}[\bar{\mathbf{y}}_N - \mu]}{\sqrt{s_N^2(\mathbf{y})}}$$

dove  $s_N^2(\mathbf{y})$  è la varianza campionaria

$$s_N^2(\mathbf{y}) = \frac{1}{N} \sum_{t=1}^N (\mathbf{y}(t) - \bar{\mathbf{y}}_N)^2$$

Dal risultato derivare la distribuzione asintotica della statistica di Student

$$t(\mathbf{y}) := \frac{[\bar{\mathbf{y}}_N - \mu]}{\sqrt{s_N^2(\mathbf{y})/N - 1}}$$

*Soluzione:* Per l'ipotesi di ergodicità  $s_N^2(\mathbf{y}) \rightarrow \sigma^2$  per  $N \rightarrow \infty$  (con probabilità uno e quindi anche in probabilità). Usando il teorema di Slutsky (punto 3), si vede subito che

$$\varphi(\mathbf{y}) \xrightarrow{L} N(0, 1).$$

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  $N(0, 1)$ .

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  $\mu$  e varianza  $\Sigma$ . 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  $T \rightarrow \infty$ . 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  $\frac{1}{T}$ . Si ha infatti

$$\text{Var}(\bar{\mathbf{y}}_T) = \frac{1}{T} \left( \frac{1}{T} \sum_{t=1}^T \text{Var}(\mathbf{y}(t)) \right) = \frac{1}{T} \Sigma.$$

In altri termini, per  $T \rightarrow \infty$ , 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  $\sqrt{T} \bar{\mathbf{y}}_T$  invece converge e converge ad una distribuzione limite che è Gaussiana. Naturalmente questa distribuzione Gaussiana dovrà avere media zero e varianza  $\Sigma$ . 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..

## 5.2 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  $\sigma$ -algebra data nella definizione ??.

**Definition 5.1.** Sia  $\{\mathcal{F}_t; t \in \mathbb{Z}\}$  una successione crescente di  $\sigma$ -algre, 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

$$\mathbb{E}\{\mathbf{z}(t+1) \mid \mathcal{F}_t\} = \mathbf{0} \quad t \in \mathbb{Z}.$$

Se la varianza condizionata  $\mathbb{E}\{\mathbf{z}(t)\mathbf{z}(t)^\top \mid \mathcal{F}_{t-1}\}$  non dipende da  $\mathcal{F}_{t-1}$ , ovvero

$$\mathbb{E}\{\mathbf{z}(t)\mathbf{z}(t)^\top \mid \mathcal{F}_{t-1}\} = \Sigma_{\mathbf{z}} < \infty \tag{5.2.1}$$

diremo che  $\mathbf{z}$  ha varianza costante.

Notiamo incidentalmente che la prima condizione è equivalente alla

$$\mathbb{E}\{\mathbf{z}(t) \mid \mathcal{F}_s\} = \mathbf{0} \quad \forall s < t. \tag{5.2.2}$$

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  $\mathbb{E}\{f(\mathbf{z}(t+1)) \mid \mathcal{F}_t\} = 0$  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  $y$ , quando si intende che la predizione è quella ottima non lineare, ovvero è la media condizionata di  $y(t)$ , data la storia passata  $\mathcal{Y}_t \equiv \{y^t\}$ ,

$$z(t) := \tilde{y}(t) = y(t) - \mathbb{E}\{y(t) \mid \mathcal{Y}_{t-1}\} \quad t \in \mathbb{Z}.$$

Più in generale si può considerare l’errore di predizione di un passo di  $y$  quando l’informazione disponibile proviene anche dall’osservazione di una variabile “esogena”  $u$ . In questo caso definiamo  $\mathcal{F}_t$  come (la  $\sigma$ -algebra generata dal) l’aggregato delle funzioni della storia passata  $(y^t, u^t)$  e poniamo

$$z(t) := \tilde{y}(t) = y(t) - \mathbb{E}\{y(t) \mid y^{t-1}, u^{t-1}\} \quad t \in \mathbb{Z}$$

(dove la media condizionata è ancora intesa in senso stretto). È ancora evidente che, con la definizione di  $\mathcal{F}_t$  “estesa”, il processo  $z$  soddisfa le due condizioni della definizione 5.1.

Una classe ancora più ampia di d-martingale si ottiene considerando processi del tipo

$$z(t) := \varphi(t) \tilde{y}(t) \quad \varphi(t) \in \mathcal{F}_{t-1} \tag{5.2.3}$$

in cui  $\varphi(t) = \varphi(y^{t-1}, u^{t-1})$  è una funzione (misurabile) della storia passata fino all’istante precedente,  $t - 1$ . In questo caso si ha

$$\mathbb{E}\{z(t) \mid \mathcal{F}_{t-1}\} = \varphi(t) \mathbb{E}\{\tilde{y}(t) \mid \mathcal{F}_{t-1}\} = 0 \quad t \in \mathbb{Z}.$$

Una *martingala* è l’integrale discreto di una d-martingala,

$$x(t) = x(0) + \sum_{s=1}^t z(s) \tag{5.2.4}$$

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 5.2.** *Per ogni d-martingala  $z$  a varianza finita si ha*

$$\text{Var}\left\{\sum_{t=1}^T z(t)\right\} = \sum_{t=1}^T \text{Var}\{z(t)\} \tag{5.2.5}$$

Se la d-martingala è stazionaria il secondo membro vale  $T \Sigma_z$ .

*Proof.* Facciamo la dimostrazione per il caso scalare. Il caso vettoriale è identico

modulo le ovvie complicazioni nelle notazioni. Si ha

$$\begin{aligned} \mathbb{E} \left\{ \sum_{t=1}^T \mathbf{z}(t) \right\}^2 &= \mathbb{E} \{ \mathbf{z}(1)^2 + \mathbf{z}(2)^2 + \dots + \mathbf{z}(T)^2 \} + \\ &+ \mathbb{E} \left\{ 2 \sum_{t>s} \mathbf{z}(t)\mathbf{z}(s) \right\} = \\ &= \sum_{t=1}^T \text{Var} \{ \mathbf{z}(t) \} + 2 \sum_{t>s} \mathbb{E} \mathbf{z}(t)\mathbf{z}(s) \end{aligned}$$

L'ultimo termine è zero giacchè se  $t > s$ ,  $\mathbf{z}(s) \in \mathcal{F}_s$ , e si può scrivere

$$\mathbb{E} \mathbf{z}(t)\mathbf{z}(s) = \mathbb{E} \{ \mathbb{E} [ \mathbf{z}(t)\mathbf{z}(s) \mid \mathcal{F}_s ] \} = \mathbb{E} \{ \mathbb{E} [ \mathbf{z}(t) \mid \mathcal{F}_s ] \mathbf{z}(s) \} = 0$$

in virtù della proprietà di d-martingala (5.2.2).  $\square$

Sul risultato seguente, formulato inizialmente da P. Levy e J.L. Doob, dimostrato da Billingsley e Ibragimov [4, 28] e successivamente generalizzato da vari autori, poggia la dimostrazione della normalità asintotica dei metodi di identificazione PEM.

**Theorem 5.2.** 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

$$\sqrt{T} \bar{\mathbf{z}}_T \xrightarrow{L} \mathcal{N}(0, \Sigma_{\mathbf{z}}) \tag{5.2.6}$$

ovvero, la statistica  $\sqrt{T} \bar{\mathbf{z}}_T$  converge in legge alla distribuzione Gaussiana di media zero e varianza  $\Sigma_{\mathbf{z}}$ .

*Proof.* Seguiremo la traccia di dimostrazione di J.L. Doob [14, p. 383] per il caso scalare e lasceremo 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),  $\sigma^2$ , 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  $\omega$  più rapidamente di  $\omega^2$ . Detta  $\phi_T(\omega)$  la funzione caratteristica della somma  $\mathbf{x}(T) := \sum_{t=1}^T \mathbf{z}(t)$ , si ha

$$\begin{aligned} \phi_T(\omega) &= \mathbb{E} \left\{ \mathbb{E} \left[ e^{i\omega \mathbf{z}(T)} \mid \mathcal{F}_{T-1} \right] e^{i\omega \mathbf{x}(T-1)} \right\} = \\ &= \left[ 1 - \frac{\sigma^2 \omega^2}{2} \right] \mathbb{E} e^{i\omega \mathbf{x}(T-1)} + \mathbb{E} \{ o(\omega^2) e^{i\omega \mathbf{x}(T-1)} \} = \\ &= \left[ 1 - \frac{\sigma^2 \omega^2}{2} \right] \phi_{T-1}(\omega) + \bar{o}(\omega^2) \end{aligned}$$

dove  $\bar{o}(\omega^2)$  è l'aspettazione di una variabile aleatoria in  $\mathcal{F}_{T-1}$  che ha lo stesso modulo di  $o(\omega^2)$  e tende quindi a zero con  $\omega$  più rapidamente di  $\omega^2$ . 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  $\omega$ .

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  $\phi_T$  appena trovata calcolata in  $\omega/\sqrt{T}$ , per cui

$$\phi_T\left(\frac{\omega}{\sqrt{T}}\right) = \left[ 1 - \frac{\sigma^2 \omega^2}{2T} \right]^T + \bar{o}_T\left(\frac{\omega^2}{T}\right)$$

In questa espressione il secondo termine tende a zero per  $T \rightarrow \infty$ , qualunque sia il valore di  $\omega$  fissato, mentre il limite del primo addendo è  $\exp\left\{-\frac{\sigma^2 \omega^2}{2}\right\}$ . Quindi la funzione caratteristica di  $\mathbf{s}(T)$  converge a quella della Gaussiana  $\mathcal{N}(0, \sigma^2)$ .  $\square$

**Remark 5.1.** 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.

### 5.3 Il teorema del limite centrale per processi stazionari

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  $T$ ).

**Theorem 5.3.** *Sia  $\mathbf{y}$  un processo stazionario<sup>18</sup> di media  $\mu$  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*

$$\Sigma_{\infty} := \lim_{T \rightarrow \infty} \text{Var} \left( \sqrt{T} \bar{\mathbf{y}}_T \right) = S_{\mathbf{y}}(e^{j0}). \quad (5.3.1)$$

*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  $\Sigma_{\infty}$  è la covarianza della distribuzione limite  $\mathcal{D}$ .*

<sup>18</sup>Qui a rigore basta la stazionarietà in senso debole.

*Proof.* Supponiamo senza perdita di generalità che il processo  $y$  abbia media zero. Introduciamo il processo stazionario "mediato"  $\mathbf{z}_T$  che si ottiene formalmente attraverso un'operazione di filtraggio, simbolicamente descritta dalla

$$\mathbf{z}_T(t) = \frac{1}{\sqrt{T}} \left( \sum_{k=0}^T z^{-k} \right) \mathbf{y}(t), \quad \mathbf{z}_T = \mathbf{z}_T(0), \quad t \in \mathbb{Z}$$

Ovviamente,  $\text{Var} \{ \sqrt{T} \bar{\mathbf{y}}_T \} = \text{Var} \{ \mathbf{z}_T(t) \}$  e calcolando la varianza del processo stazionario  $\{ \mathbf{z}_T(t) \}$ , come integrale del suo spettro, si ha

$$\text{Var} \{ \sqrt{T} \bar{\mathbf{y}}_T \} = \text{Var} \{ \mathbf{z}_T(t) \} = \frac{1}{T} \int_{-\pi}^{\pi} \left| \sum_{k=0}^T e^{-j\omega k} \right|^2 S_{\mathbf{y}}(e^{j\omega}) \frac{d\omega}{2\pi}.$$

Si tratta di calcolare il limite di questa quantità per  $T \rightarrow \infty$ . È abbastanza facile vedere, calcolando la somma della serie geometrica finita di ragione  $e^{-j\omega}$ , che il limite

$$\lim_{T \rightarrow \infty} \frac{1}{T} \left| \sum_{k=0}^T e^{-j\omega k} \right|^2 = \lim_{T \rightarrow \infty} \frac{(\sin \omega T/2)^2}{T^2 (\sin \omega/2)^2}$$

si comporta come una funzione  $\delta$  di Dirac (è integrabile, infinita per  $\omega = 0$  e zero per  $\omega \neq 0$ ), per cui, in forza della continuità in  $\omega = 0$  della densità spettrale, si ha:

$$\lim_{T \rightarrow \infty} \text{var} \{ \mathbf{z}_T \} = \lim_{T \rightarrow \infty} \text{var} \{ \sqrt{T} \bar{\mathbf{y}}_T \} = S_{\mathbf{y}}(e^{j\omega})|_{\omega=0}.$$

Per dimostrare l'ultima affermazione del teorema, ci avvarremo del lemma 5.1. 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,  $\lim_{T \rightarrow \infty} \text{var} \{ \mathbf{z}_T \}$  esiste ed è finito e quindi la successione  $\{ \text{var} [ \mathbf{z}_T ] ; T \geq 1 \}$  dev'essere limitata.  $\square$

**Remark 5.2.** 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

$$\text{var} \mathbf{z}_T = \frac{1}{T} [1 \ 1 \ \dots \ 1] \Sigma_T [1 \ 1 \ \dots \ 1]^T = \sigma(0) + 2 \sum_{s=1}^{T-1} \left(1 - \frac{s}{T}\right) \sigma(s)$$

che forma una successione monotona crescente tendente alla somma  $\sum_{-\infty}^{+\infty} \sigma(s)$ . Dall'enunciato del teorema possiamo ricavare quindi l'eguaglianza

$$\sum_{-\infty}^{+\infty} \sigma(s) = S_{\mathbf{y}}(e^{j\omega})|_{\omega=0} \tag{5.3.2}$$

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,

$$\lim_{N \rightarrow \infty} \left[ \sum_{s=-N}^{+N} e^{-j\omega s} \sigma(s) \right]_{|\omega=0} = S_y(e^{j\omega})|_{\omega=0}.$$

Come è noto, senza l'ipotesi di continuità, si può in generale garantire solo la convergenza in  $L^1$  (!) della serie di Fourier di  $S_y$  (ma ) solo nel senso delle medie di Cesàro [15].  $\square$

**Remark 5.3.** Come lascia intendere l'enunciato del lemma 5.1, la condizione di limitatezza delle varianze

$$\sup_T \text{var} \{ \sqrt{T} \bar{y}_T \} < \infty$$

implica l'integrabilità uniforme delle medie campionarie normalizzate. Quest'ultima condizione (assumendo medie nulle e variabili scalari) si esprime nella forma seguente,

$$\lim_{\alpha \rightarrow \infty} \sup_T \mathbb{E} |\sqrt{T} \bar{y}_T| I_{\{|\sqrt{T} \bar{y}_T| > \alpha\}} = 0 \tag{5.3.3}$$

Vedere e.g. [4, p. 32-33]. Si può così amplificare leggermente il contenuto del teorema 5.3 dicendo che,

*Se  $y$  ha densità spettrale  $S_y(e^{j\omega})$  continua in  $\omega = 0$ , allora la successione delle medie campionarie normalizzate  $\{\sqrt{T} \bar{y}_T\}$  è uniformemente integrabile (i.e. vale la (5.3.3)).*

*Nelle stesse ipotesi, se  $\sqrt{T} \bar{y}_T$  converge in legge a una distribuzione  $\mathcal{D}$  allora tutti i momenti che esistono di  $\sqrt{T} \bar{y}_T$  convergono a quelli corrispondenti della distribuzione limite  $\mathcal{D}$ .*  $\square$

### La condizione di Lindeberg

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 ma non necessariamente identicamente distribuite<sup>19</sup>

Per semplificare ancora un poco le notazioni introduciamo la somma normalizzata e centrata  $s_T := \sqrt{T} \tilde{y}_T := \sqrt{T}(\bar{y}_T - \mu) = \sqrt{T} \sum_{t=1}^T (y(t) - \mu)$ , denotiamo la distribuzione di probabilità di questo vettore aleatorio con  $F_{s_T}(x)$  e con  $\Phi(x)$  una generica distribuzione Gaussiana  $m$ -dimensionale a media nulla. La condizione di Lindeberg (generalizzata) è una condizione necessaria e sufficiente affinché  $F_{s_T}(x)$  converga a una distribuzione Gaussiana,  $\Phi(x)$ , quando  $T \rightarrow \infty$ .

Consideriamo un arbitrario intorno sferico  $\Gamma_\alpha := \{x \mid \|x\| \leq \alpha, \}$  dell'origine in  $\mathbb{R}^m$  e sia  $I_{\Gamma_\alpha}$  la sua funzione indicatrice. Dalla proposizione 5.1 scende che se

<sup>19</sup>Per una storia dell'evoluzione del problema del limite centrale si può consultare il trattato di M. Loève [38, Cap. 6]. La condizione di Lindeberg (e il relativo teorema di convergenza del 1922) è menzionata a p. 280.

$F_{s_T}(x)$  converge a  $\Phi(x)$  allora

$$\lim_{T \rightarrow \infty} \int_{\Gamma_\alpha} \|x\|^2 dF_{s_T}(x) = \int_{\Gamma_\alpha} \|x\|^2 d\Phi(x)$$

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  $s_T$  deve convergere a quella di  $\Phi(x)$ , i.e.

$$\lim_{T \rightarrow \infty} \int_{\mathbb{R}^m} \|x\|^2 dF_{s_T}(x) = \int_{\mathbb{R}^m} \|x\|^2 d\Phi(x)$$

Sia ora  $\bar{\Gamma}_\alpha$  la regione esterna (i.e. l'insieme complementare) all'intorno sferico  $\Gamma_\alpha$ . Le due relazioni limite appena scritte implicano ovviamente che

$$\lim_{T \rightarrow \infty} \int_{\bar{\Gamma}_\alpha} \|x\|^2 dF_{s_T}(x) = \int_{\bar{\Gamma}_\alpha} \|x\|^2 d\Phi(x). \quad (5.3.4)$$

Notiamo ora che l'integrale della Gaussiana a secondo membro può essere reso piccolo a piacere pur di prendere  $\alpha$  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.

**Theorem 5.4.** *Sia  $y$  un processo stazionario in senso stretto, con distribuzione spettrale assolutamente continua e matrice densità spettrale  $S_y(e^{j\omega})$  continua in  $\omega = 0$ . La d.d.p  $F_{s_T}(x)$  converge a una Gaussiana se e solo se il limite (che deve esistere per ogni  $\alpha \geq 0$ )*

$$\phi(\alpha) := \lim_{T \rightarrow \infty} \int_{\bar{\Gamma}_\alpha} \|x\|^2 dF_{s_T}(x) = \lim_{T \rightarrow \infty} \mathbb{E} \{ \|s_T\|^2 I_{\{\|s_T\| > \alpha\}} \} \quad (5.3.5)$$

tende a zero per  $\alpha \rightarrow \infty$ .

La discussione precedente era la dimostrazione che questa condizione è necessaria. Per la prova della sufficienza (che è abbastanza complicata) riamandiamo alla letteratura [54].

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 5.3.** *Nello spazio Euclideo  $(\mathbb{R}^m)^N$ , dove  $y_k \in \mathbb{R}^m$ ;  $k = 1, 2, \dots, N$ , vale la seguente relazione di inclusione tra insiemi*

$$\{\|y_1 + y_2 + \dots + y_N\| \geq \alpha\sqrt{N}\} \subset \bigcup_{k=1}^N \{\|y_k\| \geq \alpha\} \quad (5.3.6)$$

qualunque sia  $\alpha > 0$ .

*Proof.* Mostriamo che la relazione vale per  $N = 2$ . Nel complementare dell'insieme a secondo membro in (5.3.6) si ha  $\|y_k\|^2 < \alpha^2$ ;  $k = 1, 2$  e quindi vale l'implicazione

$$\|y_1 + y_2\|^2 \leq \|y_1\|^2 + \|y_2\|^2 < 2\alpha^2 \Leftrightarrow \{\|y_k\|^2 < \alpha^2; k = 1, 2\}$$

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 (5.3.6) valga per  $N = n$ , mostriamo che vale anche per  $N = n + 1$ . Posto  $\bar{y} := \sum_{k=1}^n y_k$ , si ha

$$\|\bar{y} + y_{n+1}\|^2 \leq \|\bar{y}\|^2 + \|y_{n+1}\|^2 < (n+1)\alpha^2 \Leftrightarrow \{\|\bar{y}\|^2 < n\alpha^2\} \cap \{\|y_{n+1}\|^2 < \alpha^2\}$$

e quindi anche

$$\{\|y_1 + y_2 + \dots + y_{n+1}\|^2 \geq (n+1)\alpha^2\} \subset \{\|y_1 + y_2 + \dots + y_n\|^2 \geq n\alpha^2\} \cup \{\|y_{n+1}\|^2 \geq \alpha^2\}$$

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.  $\square$

Usando l'inclusione (5.3.6) del lemma precedente si può maggiorare l'integrale in (5.3.5) con l'espressione

$$\begin{aligned} & \frac{1}{T} \mathbb{E} \left\{ \left\| \sum_{t=1}^T \mathbf{y}(t) \right\|^2 I_{\{\|\sum_{t=1}^T \mathbf{y}(t)\| > \alpha\sqrt{T}\}} \right\} \leq \frac{1}{T} \mathbb{E} \left\{ \left\| \sum_{t=1}^T \mathbf{y}(t) \right\|^2 I_{\cup_{t=1}^T \{\|\mathbf{y}(t)\| > \alpha\}} \right\} \\ & = \frac{1}{T} \mathbb{E} \left\{ \sum_{t,s=1}^T \mathbf{y}(t)^\top \mathbf{y}(s) \left[ \sum_{t=1}^T I_{\{\|\mathbf{y}(t)\| > \alpha\}} - \sum_{t \neq s} I_{\{\|\mathbf{y}(t)\| > \alpha\}} I_{\{\|\mathbf{y}(s)\| > \alpha\}} \right] \right\} \\ & = \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left\{ \|\mathbf{y}(t)\|^2 I_{\{\|\mathbf{y}(t)\| > \alpha\}} \right\} = \mathbb{E} \left\{ \|\mathbf{y}(t)\|^2 I_{\{\|\mathbf{y}(t)\| > \alpha\}} \right\}. \end{aligned}$$

La seconda eguaglianza segue dall'identità  $I_{E_1 \cup E_2 \dots \cup E_T} = I_{E_1} + I_{E_2} + \dots + I_{E_T} - \prod_{i \neq k} I_{E_i} I_{E_k}$  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  $\alpha \rightarrow \infty$ .

### Il TLC per processi stazionari dissolventi

Presenteremo qui sotto una versione generale del TLC, dovuta essenzialmente a Bernstein e Rosenblatt [2, 54, 53], che è valida, per somme di variabili di un processo strettamente stazionario dissolvente e che quindi generalizza il TLC per processi i.i.d. al caso di variabili dipendenti. Purtroppo le proprietà di un processo strettamente stazionario  $y$  di essere p.n.d. (in senso stretto), dissolvente etc. sono assai difficili da verificare in pratica.

Come abbiamo visto nella sezione 4.4, per un processo p.n.d. (in senso stretto), si ha, asintoticamente l'indipendenza delle variabili "sufficientemente lontane" nel tempo e si può usare nella dimostrazione del TLC un argomento simile a quello del caso i.i.d.. Per usare questo tipo di argomenti serve però che il *coefficiente di dissolvenza* (*mixing coefficient* in inglese)

$$\alpha(\tau) := \sup_{A \in \mathcal{Y}_t^-, B \in \mathcal{Y}_{t+\tau}^+} |P(A \cap B) - P(A)P(B)|, \quad \tau > 0, \quad (5.3.7)$$

che in virtù della stazionarietà del processo, dipende solo da  $\tau$ , tenda a zero abbastanza velocemente per  $\tau \rightarrow +\infty$ . Serve in effetti che  $\alpha(\tau)$  tenda a zero più velocemente di  $\frac{1}{\tau}$ ,

$$\alpha(\tau) = O\left(\frac{1}{\tau^{1+\epsilon}}\right) \quad \epsilon > 0 \quad (5.3.8)$$

Chiameremo un processo stazionario che soddisfa la condizione (5.3.8), *fortemente dissolvente* (*strongly mixing* in inglese). Notiamo che per un processo p.n.d., in virtù delle (4.4.7), (4.4.8),  $\alpha(\tau)$  tende in ogni caso a zero.

**Theorem 5.5 (Bernstein, Rosenblatt, Rozanov).** *La d.d.p  $F_{s_T}(x)$  di un processo fortemente dissolvente e a varianza finita (soddisfa la condizione di Lindeberg e quindi) converge a una Gaussiana.*

*Proof.* La dimostrazione di questo teorema si può trovare nel testo di Rozanov, pp.194-195.  $\square$

**Nota Bene:** La matrice varianza della distribuzione limite  $\Phi(x)$ , di  $\sqrt{T}y_T$  in generale non coincida con quella,  $\Sigma = E y(t)y(t)'$ , delle variabili del processo  $y$ . 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 5.3.

Sebbene le proprietà di un processo strettamente stazionario  $y$  di essere p.n.d. (in senso stretto), dissolvente etc. siano assai difficili da verificare in pratica, esse giocano un ruolo importante nella teoria perchè *vengono automaticamente ereditate da ogni processo  $\{z(t)\}$  ottenuto per traslazione temporale di una qualunque funzione a supporto  $(\mathbb{I})$  finito del processo,  $z = f(y)$ <sup>20</sup>. 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  $z$  valgono delle inclusioni del tipo,*

$$L_t^-(z) \subset L_{t+i}^-(y) \quad L_t^+(z) \subset L_{t-i}^+(y) \quad (5.3.9)$$

qualunque sia  $t$ . Equivalentemente,  $\mathcal{Z}_t^- \subset \mathcal{Y}_{t+i}^-$ ,  $\mathcal{Z}_t^+ \subset \mathcal{Y}_{t-i}^+$  per le relative  $\sigma$ -algebre e ne segue immediatamente che se  $y$  è p.n.d. (in senso stretto), fortemente dissolvente etc. la stessa proprietà viene ereditata automaticamente dal processo  $\{z(t)\}$ . Si ha pertanto il seguente risultato, che può essere visto come un corollario del teorema 5.5.

<sup>20</sup>Quindi  $f(y)$  dipende solo da un numero *finito* di variabili del processo  $y$ .

**Theorem 5.6.** Ogni processo generato per traslazione temporale secondo la (4.3.2), di una funzione a supporto finito  $\mathbf{z} = f(\mathbf{y})$  di un processo fortemente dissolvante  $\mathbf{y}$  è ancora fortemente dissolvante. Se  $\mathbf{y}$  è fortemente dissolvante 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

$$\lim_{T \rightarrow \infty} F_{\sqrt{T}\bar{\mathbf{z}}_T}(x) = \Phi(x) \tag{5.3.10}$$

dove  $\Phi(x)$  è una distribuzione Gaussiana.

**Example 5.2.** Assumiamo che il processo scalare  $\mathbf{y}$  ammetta momenti di ordine sufficientemente elevato; allora le variabili

$$\frac{1}{T} \sum_1^T \mathbf{y}(t)^2, \quad \frac{1}{T} \sum_1^T \mathbf{y}(t)\mathbf{y}(t-1), \quad T = 1, 2, \dots$$

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.

**Theorem 5.7 (Cramèr).** Sia  $\{\mathbf{x}_N; N = 1, 2, \dots\}$  una successione di vettori aleatori  $n$ -dimensionali per cui  $\sqrt{N}(\mathbf{x}_N - \mu) \xrightarrow{L} \mathcal{N}(0, \Sigma)$  e  $g : \mathbb{R}^n \rightarrow \mathbb{R}^k$  una funzione la cui matrice Jacobiana  $G(x)$  (esiste ed) è continua in un intorno del punto  $x = \mu$ . Allora:

$$\sqrt{N}(g(\mathbf{x}_N) - g(\mu)) \xrightarrow{L} \mathcal{N}(0, G(\mu)\Sigma G(\mu)') \tag{5.3.11}$$

Come applicazione del teorema di Cramèr, consideriamo la successione dei quadrati delle medie campionarie  $\{(\bar{\mathbf{y}}_T)^2\}$ . Dato che per  $T \rightarrow \infty$ ,  $\sqrt{T}\bar{\mathbf{y}}_T \sim \mathcal{N}(0, \sigma^2)$  e  $\frac{\partial}{\partial x}g(x) = 2x$ , si ha

$$\sqrt{T}[(\bar{\mathbf{y}}_T)^2 - \mu^2] \xrightarrow{L} \mathcal{N}(0, 4\mu^2\sigma^2).$$

**Remark 5.4.** 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* ( $\mathcal{N}(0, 0)$ ), in altri termini, il primo membro converge alla costante zero, il che è un chiaro sintomo del fatto che per  $T \rightarrow \infty$  la varianza della successione converge a zero più rapidamente di  $\frac{1}{T}$ . Questo fatto si può verificare direttamente notando che

$$\sqrt{T}[(\bar{\mathbf{y}}_T)^2 - \mu^2] = \left[ \sqrt{T}(\bar{\mathbf{y}}_T - \mu) \right] (\bar{\mathbf{y}}_T + \mu)$$

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  $2\mu$  che è zero per  $\mu = 0$ .

Per analizzare casi di questo genere torna utile l'affermazione (1) del teorema di Slutsky 5.1. Esaminiamo l'esempio precedente (con  $\mu = 0$ ) alla luce di questo risultato. Dato che  $\sqrt{T}\bar{y}_T \xrightarrow{L} \mathcal{N}(0, \sigma^2)$ , prendendo  $f(x) = x^2$ , il teorema di Slutsky asserisce che la d.d.p di  $T(\bar{y}_T)^2$  converge alla d.d.p. del quadrato di una variabile Gaussiana  $\mathcal{N}(0, \sigma^2)$ . In altri termini

$$\frac{T(\bar{y}_T)^2}{\sigma^2} \xrightarrow{L} \chi^2(1)$$

e la velocità di convergenza alla distribuzione limite della distribuzione di  $(\bar{y}_T)^2$  è dell'ordine di  $\frac{1}{T}$ . La varianza tenderà quindi al suo valore asintotico come  $\frac{1}{T^2}$ .

**Example 5.3.** *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  $\mu_4$  finito, è*

$$\sqrt{T}(s_T^2(\mathbf{y}) - \sigma^2) \xrightarrow{L} \mathcal{N}(0, \mu_4 - \sigma^4).$$

*Soluzione:*

La varianza campionaria si può esprimere come

$$\begin{aligned} s_T^2(\mathbf{y}) &= \frac{1}{T} \sum_{t=1}^T (\mathbf{y}(t) - \bar{y}_T)^2 = \frac{1}{T} \sum_{t=1}^T \mathbf{y}(t)^2 - (\bar{y}_T)^2 \\ &= \frac{1}{T} \sum_{t=1}^T [\mathbf{y}(t) - \mu - (\bar{y}_T - \mu)]^2 = \frac{1}{T} \sum_{t=1}^T (\mathbf{y}(t) - \mu)^2 - (\bar{y}_T - \mu)^2 \\ &:= m_2(\mathbf{y}) - m_1(\mathbf{y})^2 \end{aligned}$$

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 5.7). 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  $\mu_3$  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

$$g'(0, \sigma^2) \Sigma [g'(0, \sigma^2)]^\top = \text{var } \mathbf{y}(t)^2 = \mathbb{E} \mathbf{y}(t)^4 - (\mathbb{E} \mathbf{y}(t)^2)^2 = \mu_4 - \sigma^4$$

per cui in definitiva si ha

$$\sqrt{T} [s_T^2(\mathbf{y}) - g(0, \sigma^2)] = \sqrt{T} [s_T^2(\mathbf{y}) - \sigma^2] \xrightarrow{L} \mathcal{N}(0, \mu_4 - \sigma^4).$$

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.

**Example 5.4.** Supponiamo che  $\mathbf{y}$  sia un processo scalare i.i.d. con momento del quart'ordine  $\mu_4$  finito. Vogliamo ricavare la distribuzione asintotica della statistica di Student dell'esempio 5.1 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  $\mu_3$  non interessa. Posto  $\psi(\mathbf{y}) = g(\bar{\mathbf{y}}_N - \mu, s_T^2(\mathbf{y}))$  con  $g(x_1, x_2) := \frac{x_1}{\sqrt{x_2}}$  si ha,

$$g(\mu_1, \mu_2) = \frac{0}{\sigma} = 0, \quad \frac{\partial g}{\partial x}(x_1, x_2) = \left[ \frac{1}{\sqrt{x_2}}, -\frac{1}{2} x_1 x_2^{-3/2} \right]$$

dove lo Jacobiano è manifestamente continuo in  $(0, \sigma^2)$  e  $\frac{\partial g}{\partial x}(0, \sigma^2) = [1/\sigma, 0]$ .

Per cui, applicando la formula (5.3.11) 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 5.1. 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 [17].

## 5.4 Sistemi lineari e TLC

Un caso molto interessante per le applicazioni che abbiamo in vista 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 5.6 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

$$\mathbf{x}(t+1) = A\mathbf{x}(t) + B\mathbf{e}(t) \quad (5.4.1)$$

$$\mathbf{y}(t) = C\mathbf{x}(t) \quad (5.4.2)$$

**Theorem 5.8.** *Se  $A$  è strettamente stabile (autovalori all’interno del cerchio unit ) ed  $\mathbf{e}$    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 lineare (5.4.1), (5.4.2) e risulta*

$$\sqrt{T}\bar{\mathbf{y}}_T \xrightarrow{L} \mathcal{N}(0, C(I-A)^{-1}BQB^\top(I-A)^{-\top}C^\top) \quad (5.4.3)$$

dove  $Q = Q^\top \geq 0$    la varianza asintotica di  $\sqrt{T}\bar{\mathbf{e}}_T$ .

*Proof.* Se la propriet  dell’enunciato vale per il processo di stato, ovviamente vale anche per quello di uscita. Consideriamo allora la (5.4.1) 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   immediato verificare che sono tra loro asintoticamente equivalenti; i.e.  $\sqrt{T}\bar{\mathbf{x}}_T^1 - \sqrt{T}\bar{\mathbf{x}}_T \rightarrow 0$  in probabilit . Quindi, per calcolare la distribuzione asintotica possiamo, in forza del teorema di Slutsky, sostituire al primo membro  $\sqrt{T}\bar{\mathbf{x}}_T$  al posto di  $\sqrt{T}\bar{\mathbf{x}}_T^1$ , ottenendo cos 

$$\sqrt{T}\bar{\mathbf{x}}_T = (I-A)^{-1}B\sqrt{T}\bar{\mathbf{e}}_T.$$

Ne segue che  $\sqrt{T}\bar{\mathbf{x}}_T$    asintoticamente normale di varianza asintotica data dalla matrice  $(I-A)^{-1}BQB^\top(I-A)^{-\top}$ .  $\square$

Notiamo in particolare che se  $\mathbf{e}$    i.i.d. allora  $Q$    anche la varianza di  $\mathbf{e}(t)$  e la varianza asintotica di  $\sqrt{T}\bar{\mathbf{y}}_T$    proprio uguale alla densit  spettrale di  $\mathbf{y}$  calcolata in  $e^{j\omega} = 1$ , come stabilito nel teorema 5.3.

Un caso un p  pi  generale riguarda processi stazionari generati come uscita di un filtro  $\ell^2$ -stabile, non necessariamente razionale, sollecitati da un processo  $\mathbf{e}$  che supporremo i.i.d. a media zero e di varianza  $\sigma^2$  finita. Per semplicit  considereremo il caso scalare,

$$\mathbf{y}(t) = \sum_{-\infty}^t h(t-s)\mathbf{e}(s) \quad \sum_0^{+\infty} h(s)^2 < \infty. \quad (5.4.4)$$



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

$$\hat{\mathbf{y}}(t | t - \tau) := \mathbb{E}[\mathbf{y}(t) | L_{t-\tau}^{2+}(\mathbf{e})] = \sum_{s=t-\tau}^t h(t-s) \mathbf{e}(s)$$

e quindi

$$\mathbf{y}(t) - \hat{\mathbf{y}}(t | t - \tau) = \sum_{s=-\infty}^{t-\tau-1} h(t-s) \mathbf{e}(s) = \sum_{s=\tau+1}^{+\infty} h(s) \mathbf{e}(t-s).$$

Da questa relazione si ricava facilmente che

$$\frac{1}{\sqrt{T}} \left[ \sum_{t=1}^T \mathbf{y}(t) - \sum_{t=1}^T \hat{\mathbf{y}}(t | t - \tau) \right] = \frac{1}{\sqrt{T}} \sum_{s=\tau+1}^{+\infty} h(s) \left( \sum_{t=1}^T \mathbf{e}(t-s) \right)$$

che tende a zero in media quadratica, per  $\tau \rightarrow +\infty$  dato che

$$\mathbb{E} \left\{ \frac{1}{\sqrt{T}} \left[ \sum_{t=1}^T \mathbf{y}(t) - \sum_{t=1}^T \hat{\mathbf{y}}(t | t - \tau) \right] \right\}^2 = \frac{T\sigma^2}{T} \sum_{s=\tau+1}^{+\infty} h(s)^2 \rightarrow 0 \quad (5.4.5)$$

dato che  $h \in \ell^2$ . Quindi, per  $\tau \rightarrow +\infty$  il processo delle medie normalizzate di  $\hat{\mathbf{y}}(t | t - \tau)$  converge in media quadratica (e quindi anche in probabilità) a  $\sqrt{T} \bar{\mathbf{y}}_T$ . Dato che  $\mathbf{e}$  è fortemente dissolvente e  $\hat{\mathbf{y}}(t | t - \tau)$  dipende per ogni  $\tau$  solo da un tratto finito della storia passata di  $\mathbf{e}$ , scende dal teorema 5.6 che  $1/\sqrt{T} \sum_{t=1}^T \hat{\mathbf{y}}(t | t - \tau)$  è, per ogni  $\tau$ , asintoticamente Gaussiano.

Purtroppo il successivo passaggio al limite richiede che al tendere di  $\tau$  all'infinito, le "code"  $\sum_{s=\tau+1}^{+\infty} h(s)^2$  convergano a zero con sufficiente rapidità e per questo serve una condizione più stringente dell'energia finita ( $h \in \ell^2$ ). E. Hannan [23] ha dimostrato che è sufficiente assumere

$$\sum_{t=0}^{+\infty} |h(t)| < \infty \quad (5.4.6)$$

che è la condizione naturale per la stabilità ingresso-uscita del sistema (5.4.4).

**Theorem 5.9.** *Nell' ipotesi (5.4.6), la successione delle medie normalizzate  $\sqrt{T} \bar{\mathbf{y}}_T$  del processo definito dall' equazione (5.4.4) converge in legge ad una distribuzione Gaussiana.*

## 5.5 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 applicazioni (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 5.2.** Sia  $\{\phi_T(\mathbf{y}); T = 1, 2, \dots\}$  una successione di stimatori<sup>21</sup> e  $d(T)$  una funzione di  $T$  crescente e strettamente positiva. Diremo che  $\phi_T(\mathbf{y})$  ha varianza asintotica  $\Sigma$  se

$$\sqrt{d(T)} \phi_T(\mathbf{y}) \xrightarrow{L} D(\mu, \Sigma) \quad (5.5.1)$$

dove  $D(\mu, \Sigma)$  è una d.d.p. di media  $\mu$  e varianza  $\Sigma$ , 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

$$\lim_{T \rightarrow \infty} \text{Var} \left[ \sqrt{d(T)} \phi(\mathbf{y}_T) \right] = \Sigma. \quad (5.5.2)$$

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<sup>22</sup>

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

$$\Sigma = I(\theta)^{-1}. \quad (5.5.3)$$

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  $1/T$ .

<sup>21</sup>Questa terminologia è un pò pesante e verrà usualmente abbreviata riferendosi semplicemente al generico stimatore  $\phi_T(\mathbf{y})$  della sequenza.

<sup>22</sup>Con campioni di numerosità infinita, ci possono essere patologiche eccezioni alla disuguaglianza di Cramèr-Rao. Può infatti accadere che per qualche valore di  $\theta$  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 [17, 33].

## CHAPTER 6

# METODI A MINIMIZZAZIONE DELL'ERRORE DI PREDIZIONE

## 6.1 Introduzione

In questo capitolo tratteremo di identificazione di modelli lineari di tipo ingresso-uscita del tipo discusso nella sezione 3.4 del capitolo 3.

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 [36, 37].

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

$$y^N := \{y(t); t = 1, 2, \dots, N\}, \quad u^N := \{u(t); t = 1, 2, \dots, N\} \quad (6.1.1)$$

si procede come segue:

1. Per un generico valore di  $\theta$ , si costruisce il (miglior, secondo qualche criterio) predittore all'istante  $t - 1$  dell'uscita successiva,  $y(t)$ . Per ogni  $\theta$  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,

$$\hat{M}(\theta) : (y^{t-1}, u^{t-1}) \mapsto \hat{y}_\theta(t | t - 1)$$

La predizione  $\hat{y}_\theta(t | t - 1)$  si può all'occorrenza pensare come funzione dei dati passati (oltre che del parametro  $\theta$ ) e quindi, come una quantità aleatoria (prima di aver misurato i dati). In questo contesto verrà impiegato il simbolo  $\hat{y}_\theta(t | t - 1)$ .

2. Si formano gli *errori di predizione*:

$$\varepsilon_\theta(t) := y(t) - \hat{y}_\theta(t); \quad t = 1, 2, \dots, N$$

che, analogamente a quanto detto per il predittore, possono essere all'occorrenza interpretati come quantità aleatorie, indicate con simboli in grassetto, i.e.  $\varepsilon_\theta(t)$ . Notiamo ad esempio che per la classe di modelli (3.2.1), usando formalmente l'espressione per il predittore di Wiener (3.1.9), si ottiene

$$\begin{aligned}\varepsilon_\theta(t) &= \mathbf{y}(t) - \hat{\mathbf{y}}(t | t-1) = \mathbf{y}(t) - G_\theta(z)^{-1} [F_\theta(z)\mathbf{u}(t) + (G_\theta(z) - 1)\mathbf{y}(t)] \\ &= G_\theta(z)^{-1} [\mathbf{y}(t) - F_\theta(z)\mathbf{u}(t)]\end{aligned}\quad (6.1.2)$$

dalla quale si può ricavare una rappresentazione del processo ("vero") osservato  $\mathbf{y}$  mediante un *modello scelto arbitrariamente nella classe* (3.2.1); i.e.

$$\mathbf{y}(t) = F_\theta(z)\mathbf{u}(t) + G_\theta(z)\varepsilon_\theta(t), \quad (6.1.3)$$

Notare però che in queste rappresentazioni l'innovazione è sostituita dall'errore di predizione (che in generale non è bianco). La stessa idea si applica anche al caso di processi (non necessariamente stazionari) rappresentati mediante modelli di stato con parametri non noti; si veda ad esempio [44, p. 387-388]. L'idea di rappresentazione mediante l'errore di predizione troverà applicazioni importanti nel seguito e lo studente è invitato a meditare sul suo significato.

3. Si minimizza rispetto a  $\theta$  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*,

$$V_N(\theta) := \frac{1}{N} \sum_{t=1}^N \varepsilon_\theta(t)^2 \quad (6.1.4)$$

o, più in generale, una media degli errori quadratici di predizione pesati da un fattore di sconto non negativo  $\beta(N, t)$ ,

$$V_N(\theta) := \frac{1}{N} \sum_{t=1}^N \beta(N, t) \varepsilon_\theta(t)^2 \quad \beta(t, N) > 0 \quad (6.1.5)$$

che per  $N$  piccoli dà peso minore agli errori di predizione compiuti nella fase iniziale dell'algoritmo quando l'influenza di condizioni iniziali stimate in modo approssimato (o incognite) è più deleteria. Per  $N \rightarrow \infty$  il fattore di sconto tende a diventare inutile, e si aggiustano le cose in modo che  $\beta(N, t) \rightarrow 1$ .

Si può anche considerare, invece di  $\varepsilon_\theta$ , un errore di predizione *filtrato* da un opportuno filtro lineare che pesi di più gli errori nella banda di frequenze dove più interessa una identificazione accurata. Infine si può modulare l'errore di predizione attraverso una opportuna funzione non lineare che "saturi" per valori molto grandi di  $\varepsilon_\theta$  e serva a ridurre l'influenza di *outliers* accidentali. In ogni caso, dalla minimizzazione della cifra di merito si ricava una stima di  $\theta$ ,

$$\hat{\theta}_N := \text{Arg} \min_{\theta} V_N(\theta) \quad (6.1.6)$$

che è appunto la stima PEM del parametro del modello. Naturalmente lo stimatore  $\hat{\theta}_N$  che produce la stima come funzione dei dati, viene chiamato *stimatore PEM* del parametro  $\theta$ .

4. Infine si prende come stima della varianza dell'innovazione  $\lambda^2 = \text{var}\{\mathbf{e}(t)\}$ , l'errore quadratico residuo, ovvero

$$\hat{\lambda}_N^2 := V_N(\hat{\theta}_N) \quad (6.1.7)$$

dove  $V_N$  è dato dall'espressione (6.1.4).

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.

### Identificazione PEM di modelli ARX

Consideriamo a titolo di esempio l'identificazione col metodo PEM di modelli ARX (d'innovazione) descritti nel capitolo 3,

$$A(z^{-1})\mathbf{y}(t) = B(z^{-1})\mathbf{u}(t) + \mathbf{e}(t) \quad (6.1.8)$$

assumendo che i gradi  $n$  ed  $m$  dei due polinomi  $A$  e  $B$  siano stati fissati. Definendo

$$\boldsymbol{\varphi}(t)^\top = [-\mathbf{y}(t-1) \quad \dots \quad -\mathbf{y}(t-n) \quad \mathbf{u}(t-1) \quad \dots \quad \mathbf{u}(t-m)]$$

e il vettore dei  $p := n + m$  parametri incogniti

$$\boldsymbol{\theta} := [a_1 \quad \dots \quad a_n \quad b_1 \quad \dots \quad b_m]^\top$$

questo modello si può scrivere in forma di regressione come

$$\mathbf{y}(t) = \boldsymbol{\varphi}(t)^\top \boldsymbol{\theta} + \mathbf{e}(t). \quad (6.1.9)$$

Supponiamo che sia disponibile una serie temporale di dati ingresso-uscita  $\{y(t), u(t); t = 0, 1, 2, \dots, N\}$  che vogliamo descrivere con un modello della classe (6.1.8) usando il metodo PEM. Le cose si semplificano di molto se si assume che la numerosità  $N$  del campione sia abbastanza grande di modo che le condizioni iniziali non abbiano influenza sul predittore di un passo basato sul modello (6.1.8). Ricordiamo che quest'ultimo ha la struttura di un filtro FIR

$$\hat{\mathbf{y}}_\theta(t | t-1) = \boldsymbol{\varphi}(t)^\top \boldsymbol{\theta} \quad (6.1.10)$$

che è manifestamente una *funzione lineare del parametro*  $\boldsymbol{\theta}$ . Inoltre se l'istante iniziale è preso opportunamente, in modo da non coinvolgere campioni iniziali non noti dei processi  $\{y(t)\}$  e  $\{u(t)\}$ , si può ritenere che  $\hat{\mathbf{y}}_\theta(t | t-1)$  sia il predittore a

regime (di Wiener). Definendo i vettori colonna  $\mathbf{y}$  ed  $\mathbf{e}$  con componenti  $y(t)$  e  $e(t)$  per  $t = 1, 2, \dots, N$  e la matrice  $N \times p$

$$\Phi_N := \begin{bmatrix} \varphi(1)^\top \\ \vdots \\ \varphi(N)^\top \end{bmatrix}, \quad (6.1.11)$$

il modello lineare (6.1.9) si può riscrivere  $\mathbf{y} = \Phi_N \theta + \mathbf{e}$  per cui i vettori  $N$ -dimensionali dei predittori e degli errori di predizione corrispondenti hanno l'espressione

$$\hat{\mathbf{y}}_\theta = \Phi_N \theta, \quad \boldsymbol{\varepsilon}_\theta = \mathbf{y} - \Phi_N \theta \quad (6.1.12)$$

e la cifra di merito  $V_N(\theta)$  è essenzialmente il quadrato della norma Euclidea di  $\boldsymbol{\varepsilon}_\theta$ ,

$$V_N(\theta) = \frac{1}{N} \|\mathbf{y} - \Phi_N \theta\|^2.$$

La minimizzazione dell'errore quadratico medio di predizione conduce così alla soluzione di un problema di minimi quadrati. Lo stimatore PEM del parametro  $\theta$  ha la nota espressione

$$\hat{\theta}_N = [\Phi_N^\top \Phi_N]^{-1} \Phi_N^\top \mathbf{y} \quad (6.1.13)$$

che si può anche riscrivere in una forma un pò più esplicita come

$$\hat{\theta}_N = \left[ \sum_{t=1}^N \varphi(t) \varphi(t)^\top \right]^{-1} \sum_{t=1}^N \varphi(t) \mathbf{y}(t). \quad (6.1.14)$$

Quindi l'identificazione di modelli ARX si può ridurre ad un problema di minimi quadrati, che si sa risolvere esplicitamente. Questo fatto è ovviamente un grosso incentivo all'uso di questi modelli.

**Remark 6.1.** Allo scopo di arrivare rapidamente alla formula risolutiva, abbiamo sorvolato su un certo numero di dettagli che l'arguto lettore avrà sicuramente notato. Il primo, è il fatto che per "riempire" le prime righe della matrice  $\Phi_N$  bisogna supporre che i dati a disposizione comprendano anche  $y(-n), \dots, y(-1)$  e  $u(-m), \dots, u(-1)$ . Questo è in realtà un problema risolvibile semplicemente usando i primi  $n$  o  $m$  campioni dei dati per inizializzare i predittori e i rimanenti  $N - n$  (o  $N - m$ ) per la regressione.

Il secondo problema riguarda l'invertibilità della matrice tra parentesi quadre in (6.1.14). Come vedremo più avanti questa invertibilità, almeno per  $N$  grandi, ha a che fare con l'identificabilità del modello e in particolare con la persistente eccitazione del segnale di ingresso. Infine ricordiamo al lettore che la formula (6.1.13) è ottenuta risolvendo le equazioni normali il che è, come spiegato diffusamente nel capitolo 2.10, assolutamente da evitare per ragioni di stabilità numerica.

**Problem 6.1.** Si vuole identificare il modello (6.1.9) minimizzando una somma pesata degli errori di predizione

$$V_N(\theta, \beta) := \frac{1}{N} \sum_{t=1}^N \beta(N, t) \varepsilon_\theta(t)^2$$

dove  $0 < \beta(N, \cdot) \leq 1$  è una funzione peso assegnata (interpretabile come fattore d'oblio per gli errori "vecchi"). Per  $\beta(N, \cdot) \equiv 1$  lo stimatore è fornito dalla nota formula (6.1.14); si chiede di trovare l'espressione di  $\hat{\theta}(N)$  per  $\beta(N, \cdot) \neq 1$ .

Assumiamo il processo  $\mathbf{y}$  sia effettivamente descritto da un modello della classe (6.1.9), che l'ingresso  $\mathbf{u}$  sia stazionario a varianza finita ed  $\mathbf{e}(t)$  sia un processo i.i.d. a media zero che per ogni  $t$  è scorrelato da  $\mathbf{u}^{t-1}$  (e quindi anche da  $\varphi(t)$ <sup>23</sup>), di varianza finita. Secondo voi qual'è una condizione sufficiente su  $\beta(N, \cdot)$  per avere consistenza?

Soluzione:

Definiamo i vettori colonna  $\mathbf{y}$  ed  $\mathbf{e}$  con componenti  $\mathbf{y}(t)$  e  $\mathbf{e}(t)$  per  $t = 0, 1, 2, \dots, N$  e le matrici

$$\Phi_N := \begin{bmatrix} \varphi(1)^\top \\ \vdots \\ \varphi(N)^\top \end{bmatrix}, \quad Q_N := \text{diag} \{ \beta(N, 1), \dots, \beta(N, N) \}$$

il modello lineare si può riscrivere  $\mathbf{y} = \Phi_N \theta + \mathbf{e}$  e la cifra di merito  $V_N(\theta, \beta)$  come il quadrato di una norma pesata,

$$NV_N(\theta, \beta) = \|\mathbf{y} - \Phi_N \theta\|_{Q_N}^2.$$

Usando le formule dei minimi quadrati pesati si trova

$$\hat{\theta}(N) = [\Phi_N^\top Q_N \Phi_N]^{-1} \Phi_N^\top Q_N \mathbf{y}$$

che si riscrive per esteso come

$$\hat{\theta}(N) = \left[ \sum_{t=1}^N \beta(N, t) \varphi(t) \varphi(t)^\top \right]^{-1} \sum_{t=1}^N \beta(N, t) \varphi(t) \mathbf{y}(t).$$

Sostituendo l'espressione di  $\mathbf{y}(t)$  si trova poi

$$\hat{\theta}(N) = \theta + \left[ \frac{1}{N} \sum_{t=1}^N \beta(N, t) \varphi(t) \varphi(t)^\top \right]^{-1} \frac{1}{N} \sum_{t=1}^N \beta(N, t) \varphi(t) \mathbf{e}(t).$$

Dato che  $\mathbb{E} \varphi(t) \varphi(t)^\top < \infty$  si vede che il termine tra parentesi quadre si mantiene limitato (potrebbe in particolare convergere a una costante); si tratta allora di trovare condizioni su  $\beta$  per cui

$$\frac{1}{N} \sum_{t=1}^N \beta(N, t) \varphi(t) \mathbf{e}(t) \rightarrow 0$$

Una condizione sufficiente è che  $\beta(N, t) \rightarrow 1$  per  $N \rightarrow \infty$ , uniformemente in  $t$ . In questo caso per il teorema ergodico si ha

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N \beta(N, t) \varphi(t) \mathbf{e}(t) = \mathbb{E} \varphi(t) \mathbf{e}(t) = 0.$$

Di fatto negli algoritmi ricorsivi si cerca di rendere il fattore d'oblio monotono in  $N$  e  $t$  e convergente a 1 per  $N \rightarrow \infty$ . □

<sup>23</sup>Il lettore dovrebbe a questo punto assicurarsi di sapere perchè questa affermazione è corretta.

## 6.2 Analisi asintotica dello stimatore PEM

Come si è visto per il modello ARX (6.1.8) ed è del resto facilmente intuibile per il caso generale, lo stimatore PEM risulta sempre essere una complicata funzione non lineare dei dati di misura. Ne segue che l'unica analisi statistica possibile dello stimatore PEM è quella asintotica, come si suol dire, *per grandi campioni*, ovvero per  $N \rightarrow \infty$ . In questa sezione inizieremo appunto a studiare le proprietà statistiche dello stimatore PEM quando  $N \rightarrow \infty$ . I nostri dati di misura sono per ipotesi modellabili come traiettorie di processi stocastici stazionari<sup>24</sup> e il presupposto per l'analisi asintotica corrisponderà quindi a supporre di avere a disposizione una traiettoria (semi-)infinita del processo congiunto  $[\mathbf{y} \mathbf{u}]^T$ . In queste condizioni il predittore  $\hat{\mathbf{y}}_\theta(t | t-1)$  e il relativo processo errore di predizione convergono per  $N \rightarrow \infty$  a dei processi stazionari. Ne segue che per l'analisi asintotica potremo pensare che  $\hat{\mathbf{y}}_\theta(t | t-1)$  sia il *predittore di Wiener* basato sul modello  $M(\theta)$  appartenente alla classe parametrica scelta per l'identificazione. Questo fatto, come vedremo, ci faciliterà notevolmente l'analisi.

Notiamo però che la minimizzazione dell'errore quadratico medio di predizione e il calcolo dello stimatore PEM appena esposto non richiedono affatto che la numerosità campionaria tenda all'infinito e possono essere effettuate anche per "piccoli campioni". In questi casi, il predittore basato sul modello  $M(\theta)$ , che opera su dati finiti, dovrà essere inteso come predittore *non stazionario*, e realizzato impiegando ad esempio un filtro di Kalman opportunamente inizializzato. Il problema seguente servirà ad illustrare come si deve procedere in queste circostanze.

**Problem 6.2.** *Volete identificare un modello ARMA (d'innovazione) del tipo*

$$(1 + az^{-1})\mathbf{y}(t) = (1 + cz^{-1})\mathbf{e}(t).$$

*con il metodo PEM. Purtroppo i dati sono pochi (la numerosità campionaria  $N$  è decisamente finita) e non si può usare il predittore di Wiener, ma bisogna usare il predittore di Kalman costruito su un opportuno modello di stato (minimo; i.e. di ordine uno) che descriva il processo  $\mathbf{y}$ .*

*Descrivete per passi un algoritmo iterativo per il calcolo dello stimatore PEM del parametro  $\theta := [a \ c]^T$  basato su un campione di  $N$  osservazioni. In particolare descrivete il passo di aggiornamento del modello (incluse le condizioni iniziali) e del predittore. Non serve addentrarsi troppo nell'algoritmo di ottimizzazione (e in ispecie sul calcolo del gradiente).*

*Soluzione.*

Per calcolare il predittore a memoria finita (non stazionario) serve una realizzazione di stato del modello

$$(1 + az^{-1})\mathbf{y}(t) = (1 + cz^{-1})\mathbf{e}(t)$$

<sup>24</sup>Più in generale potremmo supporre che i dati siano *segnali stazionari del secondo ordine* nel senso definito nella sezione 4.6.



ad esempio la

$$\mathbf{x}(t+1) = -a\mathbf{x}(t) + (c-a)\mathbf{e}(t) \quad (6.2.1)$$

$$\mathbf{y}(t) = \mathbf{x}(t) + \mathbf{e}(t). \quad (6.2.2)$$

dove  $Q = \lambda^2(c-a)^2$ ;  $S = \lambda^2(c-a)$ ;  $R = \lambda^2$ . Dato che il modello deve descrivere un processo stazionario, la varianza di stato iniziale si trova risolvendo l'equazione di Lyapunov  $p = a^2p + \lambda^2(c-a)^2$  che fornisce la condizione iniziale

$$p_0 = \frac{\lambda^2(c-a)^2}{1-a^2} \quad (6.2.3)$$

che si usa per inizializzare l'equazione di Riccati:

$$p(t+1) = a^2p(t) - k(t)^2\lambda(t) + \lambda^2(c-a)^2; \quad (6.2.4)$$

$$k(t) = (ap(t) + \lambda^2(c-a))\lambda(t)^{-1}; \quad (6.2.5)$$

$$\lambda(t) = p(t) + \lambda^2. \quad (6.2.6)$$

Introducendo la varianza normalizzata

$$\pi(t) := \frac{p(t)}{\lambda^2}$$

si può riscrivere tutto nella forma:

$$\pi(t+1) = a^2\pi(t) - k(t)^2\beta(t) + (c-a)^2; \quad (6.2.7)$$

$$k(t) = (a\pi(t) + (c-a))\beta(t)^{-1}; \quad \beta(t) = \frac{\lambda(t)}{\lambda^2} = \pi(t) + 1; \quad (6.2.8)$$

$$\pi(0) = \frac{(c-a)^2}{1-a^2} \quad (6.2.9)$$

che non dipende più dal parametro  $\lambda^2$  ma solo dal parametro bidimensionale  $\theta := [a \ c]^\top$ . Questo è in accordo col fatto che in generale il predittore lineare a minima varianza non dipende dalla varianza del rumore.

Il guadagno  $k(t) \equiv k_\theta(t)$  si usa per calcolare il predittore a memoria finita basato sugli ultimi  $t$  dati:

$$\hat{\mathbf{x}}_\theta(t+1|t) = (a - k_\theta(t))\hat{\mathbf{x}}_\theta(t|t-1) + k_\theta(t)\mathbf{y}(t); \quad \hat{\mathbf{x}}(1|0) = 0 \quad (6.2.10)$$

$$\hat{\mathbf{y}}_\theta(t+1|t) = \hat{\mathbf{x}}_\theta(t+1|t), \quad t = 1, 2, \dots, N. \quad (6.2.11)$$

Usando queste equazioni per un dato vettore di parametri ammissibili  $\theta = [a \ c]^\top$  con  $|a| < 1$ ;  $|c| < 1$ , si definisce una funzione  $\hat{\mathbf{y}}_\theta^N = \text{pred}(\theta, y^N)$  che produce il vettore a  $N$  componenti  $\hat{\mathbf{y}}_\theta^N$  dei predittori di un passo (non stazionari) basati sul modello di parametri  $\theta$  e sui dati  $y^N$ . Si può così formare l'errore quadratico medio di predizione

$$V_N(\theta) = \frac{1}{N}\|\varepsilon_\theta\|^2; \quad \varepsilon_\theta = y^N - \hat{\mathbf{y}}_\theta^N$$

che deve essere minimizzato rispetto a  $\theta$  usando un algoritmo iterativo, ad esempio (ma non necessariamente) un metodo di quasi-Newton del tipo

$$\theta_{k+1} = \theta_k + \left[ \sum_{t=1}^N \psi_{\theta_k}(t) \psi_{\theta_k}(t)^\top \right]^{-1} \sum_{t=1}^N \psi_{\theta_k}(t) \varepsilon_{\theta_k}(t) \quad (6.2.12)$$

$$\lambda_k^2 = V_N(\theta_k) \quad (6.2.13)$$

dove  $\psi_{\theta}(t)$  è il gradiente di  $\hat{y}_{\theta}(t) \equiv \hat{y}_{\theta}(t | t-1)$

$$\psi_{\theta}(t) := \frac{\partial \hat{y}_{\theta}(t)}{\partial \theta} = \frac{\partial \hat{x}_{\theta}(t|t-1)}{\partial \theta}$$

Ad ogni iterazione bisogna riaggiornare il calcolo della stringa dei predittori valutando la funzione

$$\hat{y}_{\theta_k}^N = \text{pred}(\theta_k, y^N); \quad k = 1, 2, \dots$$

Il calcolo del gradiente è un pò complicato e non ce ne occuperemo. □

Il teorema seguente è il primo risultato fondamentale della teoria asintotica della stima PEM.

**Theorem 6.1.** *Si assuma che*

1. I dati (6.1.1) siano generati da un processo stazionario, ergodico del secondo ordine.
2. Il modello parametrico  $M(\theta)$  dipenda dal parametro in modo differenziabile;
3. La cifra di merito sia una funzione quadratica dell'errore di predizione, ad esempio del tipo (6.1.4);
4. Esista (almeno) un minimo (6.1.6) e per  $N \rightarrow \infty$  la sequenza dei minimi rimanga limitata; i.e. esista un insieme compatto  $\Theta_0 \subseteq \Theta$  sufficientemente grande che, per  $N \rightarrow \infty$ ,  $\hat{\theta}_N \in \Theta_0$  con probabilità uno.

Si ha allora

$$\lim_{N \rightarrow \infty} V_N(\theta) = \mathbb{E}_0 \varepsilon_{\theta}(t)^2 \quad (6.2.14)$$

dove  $\mathbb{E}_0$  denota aspettazione rispetto alla distribuzione del processo vero che ha generato i dati. Inoltre, i minimizzatori,  $\hat{\theta}_N$ , di  $V_N(\theta)$ , convergono tutti, con probabilità uno, all'insieme dei punti di minimo di  $\bar{V}(\theta) := \mathbb{E}_0 \varepsilon_{\theta}(t)^2$ . In altri termini, detto  $\Delta \subset \Theta_0$  l'insieme dei punti di minimo di  $\bar{V}(\theta)$ , si ha

$$\lim_{N \rightarrow \infty} \hat{\theta}_N \in \Delta \quad (6.2.15)$$

con probabilità uno.

**Proof.** In effetti la (6.2.14) 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 [17, teorema 16(a), p. 108]). In queste condizioni si può commutare l’operazione di minimizzazione di  $V_N(\theta)$  su un arbitrario insieme compatto con quella di passaggio al limite.  $\square$

L’ipotesi che, per quasi tutte le possibili sequenze di dati di misura, un minimo (finito!),  $\hat{\theta}_N$ , esista effettivamente non è limitativa e si può sempre garantire se la dipendenza dell’errore di predizione dal parametro è di tipo polinomiale o razionale, eventualmente ridefinendo la parametrizzazione in modo opportuno. Il fatto che i punti di minimo siano (almeno per  $N$  grande) contenuti con probabilità uno in un insieme compatto, si può, come vedremo, garantire imponendo l’identificabilità del modello.

**Remark 6.2.** In vece dell’ergodicità 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”.

Se assumiamo che il processo vero sia stazionario (per quanto segue basta che lo sia in senso debole), con momenti del second’ordine finiti e *puramente non deterministico*, come è ben noto esso può essere decomposto nella somma di un predittore lineare a minima varianza d’errore  $\hat{y}_0(t | t - 1)$ , ovvero la proiezione ortogonale di  $y(t)$  sullo spazio di Hilbert generato linearmente dalla storia passata congiunta  $(y^{t-1}, u^{t-1})$ , e dell’errore di predizione di un passo (i.e. l’innovazione) di  $y(t)$

$$y(t) = \hat{y}_0(t | t - 1) + e_0(t). \tag{6.2.16}$$

Usando questa decomposizione e notando che il termine tra parentesi quadre nella decomposizione,

$$\varepsilon_\theta(t) = e_0(t) + [\hat{y}_0(t | t - 1) - \hat{y}_\theta(t | t - 1)]$$

è funzione dei dati  $(y^{t-1}, u^{t-1})$  e quindi ortogonale ad  $e_0(t)$ , si ricava la

$$\begin{aligned} \bar{V}(\theta) &= \text{var} \{e_0(t)\} + \|\hat{y}_0(t | t - 1) - \hat{y}_\theta(t | t - 1)\|^2 \\ &= \lambda_0^2 + \|\hat{y}_0(t | t - 1) - \hat{y}_\theta(t | t - 1)\|^2 \end{aligned} \tag{6.2.17}$$

dove la norma è la norma nello spazio di Hilbert generato linearmente da  $(y, u)$  (i.e. la varianza). Da questa espressione si vede che lo stimatore PEM minimizza asintoticamente la distanza tra il predittore lineare “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  $\hat{y}_0(t | t - 1)$  possono essere molti (anche infiniti).

Osserviamo anche che 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 il che garantisce che  $e_0(t)$  sia scorrelata da  $\hat{y}_\theta(t | t - 1)$ . Su questo fatto è in effetti basato anche il fondamentale risultato seguente<sup>25</sup>.

**Theorem 6.2.** *Supponiamo che valgano le ipotesi (1), (2), (3) del teorema precedente e che il processo (vero) che genera i dati sia descritto da un modello che appartiene alla stessa classe parametrica  $\mathcal{M}$  dei modelli scelti per l'identificazione, ovvero esista  $\theta_0 \in \Theta$  tale che*

$$\mathbf{y} \sim M(\theta_0) \in \mathcal{M}$$

e l'errore di predizione  $\varepsilon_\theta(t)$  sia calcolato mediante il predittore lineare a minima varianza  $\hat{y}_\theta(t | t - 1)$ .

Allora, se la classe parametrica dei modelli  $\mathcal{M} \equiv \{M(\theta); \theta \in \Theta\}$  è identificabile localmente in  $\theta = \theta_0$ , si ha<sup>26</sup>

$$\lim_{N \rightarrow \infty} \hat{\theta}_N = \theta_0 \tag{6.2.18}$$

con probabilità uno. In altri termini, lo stimatore PEM è consistente.

**Proof.** Se  $M(\theta_0) \in \mathcal{M}$ , si vede subito dalla (6.2.17) che il minimo assoluto di  $\bar{V}(\theta)$ , che vale  $\lambda_0^2$ , 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 6.1.** *Se c'è identificabilità*

$$\|\hat{y}_{\theta_0}(t | t - 1) - \hat{y}_\theta(t | t - 1)\| = 0 \Leftrightarrow \theta = \theta_0; \tag{6.2.19}$$

*i.e. c'è un solo punto di minimo assoluto di  $\bar{V}(\theta)$  e  $\Delta = \{\theta_0\}$ .*

**Proof.** Assumiamo per concretezza che vi sia assenza di reazione. Con qualche ritocco la dimostrazione funziona in realtà anche in presenza di reazione.

Supponiamo che l'ingresso  $\mathbf{u}$  sia persistentemente eccitante di ordine (esattamente)  $n \leq \infty$ . Dato che il predittore di Wiener è una funzione lineare dei dati si può scrivere simbolicamente

$$\hat{y}_\theta(t | t - 1) = L_\theta(z)\mathbf{y}(t - 1) + M_\theta(z)\mathbf{u}(t - 1)$$

dove  $L_\theta(z)$  e  $M_\theta(z)$  sono funzioni razionali strettamente stabili. Ne segue che il quadrato della norma in (6.2.19) si può esprimere nel dominio della frequenza con

<sup>25</sup>Quindi il predittore non può essere *arbitrario* come Ljung e qualche suo ottimista seguace insiste nel propagandare.

<sup>26</sup>Nel caso di assenza di reazione si può precisare la condizione richiedendo che il modello sia identificabile a priori in  $\theta = \theta_0$  e che il processo  $\mathbf{u}$  sia sufficientemente eccitante da garantire l'identificabilità.

l'integrale,

$$\|\hat{\mathbf{y}}_{\theta_0}(t | t-1) - \hat{\mathbf{y}}_{\theta}(t | t-1)\| = \int_{-\pi}^{\pi} [L_{\theta_0}(e^{j\omega}) - L_{\theta}(e^{j\omega}) M_{\theta_0}(e^{j\omega}) - M_{\theta}(e^{j\omega})] \cdot \begin{bmatrix} S_{\mathbf{y}}(e^{j\omega}) & S_{\mathbf{y}\mathbf{u}}(e^{j\omega}) \\ S_{\mathbf{u}\mathbf{y}}(e^{j\omega}) & S_{\mathbf{u}}(e^{j\omega}) \end{bmatrix} \begin{bmatrix} L_{\theta_0}(e^{j\omega}) - L_{\theta}(e^{j\omega}) \\ M_{\theta_0}(e^{j\omega}) - M_{\theta}(e^{j\omega}) \end{bmatrix} \frac{d\omega}{2\pi}$$

dove abbiamo usato il simbolo  $S_{\mathbf{u}}$  per lo spettro di  $\mathbf{u}$  con la solita convenzione di descrivere, se necessario, uno spettro a righe mediante funzioni  $\delta$ . Ora, se il primo membro è zero, l'integrando, che è una funzione non negativa della frequenza, deve essere zero per quasi tutti gli  $\omega \in [-\pi, \pi]$ . Questo implica, per la proposizione 3.5 del capitolo 3, che<sup>27</sup>

$$L_{\theta_0}(e^{j\omega_k}) - L_{\theta}(e^{j\omega_k}) = 0, \quad M_{\theta_0}(e^{j\omega_k}) - M_{\theta}(e^{j\omega_k}) = 0$$

almeno nelle  $n$  frequenze  $\omega_1, \omega_2, \dots, \omega_n$  in cui lo spettro di  $\mathbf{u}$  è positivo. Per l'ipotesi di identificabilità l'ordine di persistente eccitazione dell'ingresso è sufficiente a concludere che queste eguaglianze puntuali implicano in realtà l'eguaglianza di  $L_{\theta_0}$  a  $L_{\theta}$  e di  $M_{\theta_0}$  a  $M_{\theta}$  per tutte le frequenze  $\omega \in [-\pi, \pi]$ ; i.e.

$$L_{\theta_0}(z) \equiv L_{\theta}(z), \quad M_{\theta_0}(z) \equiv M_{\theta}(z)$$

dove  $\equiv$  significa uguaglianza per tutti i  $z$ .

Confrontando ora le espressioni di  $L(z)$  e di  $M(z)$  fornite in (3.1.9) nel teorema 3.1, si controlla facilmente che queste ultime relazioni sono equivalenti alle

$$F_{\theta_0}(z) \equiv F_{\theta}(z), \quad G_{\theta_0}(z) \equiv G_{\theta}(z)$$

e quindi a  $M(\theta_0) = M(\theta)$ . L'identificabilità a priori in  $\theta_0$  implica infine che questo possa accadere allora e solo allora che  $\theta = \theta_0$ .  $\square$

Il teorema è così dimostrato.  $\square$

**Problem 6.3.** Si vuole identificare una serie temporale usando un modello ARMA (d'innovazione) di ordine 1,

$$\mathbf{y}(t) + a\mathbf{y}(t-1) = \mathbf{e}(t) + c\mathbf{e}(t-1) \tag{6.2.20}$$

mentre i dati osservati sono in realtà generati da un processo di rumore bianco i.i.d.

$$\mathbf{y}(t) = \mathbf{e}_0(t), \quad \text{var} \{ \mathbf{e}_0(t) \} = \lambda_0^2.$$

Trovare l'insieme dei valori del parametro  $\theta := [a \ c]^T$  a cui converge, quando la numerosità campionaria  $N \rightarrow \infty$ , lo stimatore a minimo errore di predizione  $\hat{\theta}_N$ .

Discutere il risultato; secondo voi

1. Qual'è il valore vero del parametro  $\theta$ ?

<sup>27</sup>Se c'è reazione si può invocare la proposizione 3.6.

2. Lo stimatore PEM  $\hat{\theta}_N$  è consistente?

*Soluzione*

Evidentemente un modello ARMA di ordine 1 dovrà essere una descrizione ridondante (non identificabile) per un processo di rumore bianco. Verifichiamo formalmente questo fatto.

Il processo vero può essere pensato descritto da tutti i modelli ARMA di ordine uno

$$\mathbf{y}(t) + a_0 \mathbf{y}(t-1) = \mathbf{e}(t) + c_0 \mathbf{e}(t-1)$$

in cui numeratore e denominatore sono uguali; i.e.  $c_0 = a_0$ . In sostanza qualunque punto che giace sulla retta  $\{a = c\}$  del piano dei parametri  $\{a, c\}$  è un parametro vero. Ora è immediato verificare che il modello (6.2.20) è identificabile a priori (ricordare che l'identificabilità a priori è una proprietà *locale*) in tutti i punti del piano  $\{a, c\}$ , eccetto proprio in quei punti che giacciono sulla retta  $\{a = c\}$ . Quindi la classe parametrica di modelli (6.2.20) non è identificabile in  $\theta_0$ , qualunque esso sia !.

In ogni caso, il predittore di Wiener (a minima varianza) per un modello ARMA di ordine 1, si può scrivere nella forma

$$\hat{\mathbf{y}}(t | t-1) = \frac{c-a}{1+cz^{-1}} \mathbf{y}(t-1)$$

e l'errore di predizione con dati osservati generati da un processo di rumore bianco i.i.d.  $\mathbf{e}_0(t)$ , è

$$\varepsilon_\theta(t) = \mathbf{e}_0(t) - \frac{c-a}{1+cz^{-1}} \mathbf{e}_0(t-1).$$

Dato che i due termini sono scorrelati la varianza asintotica di  $\varepsilon_\theta(t)$  è

$$E_0(\varepsilon_\theta(t))^2 = \lambda_0^2 \left(1 + \frac{(c-a)^2}{1-c^2}\right)$$

che è evidentemente minima per  $c = a$ , entrambi uguali ad un qualunque valore (di modulo  $< 1$ ).

Quindi lo stimatore a minimo errore di predizione  $\hat{\theta}_N := [\hat{a}_N \hat{c}_N]^\top$  converge, quando la numerosità campionaria  $N \rightarrow \infty$  all'insieme  $\Theta_0 = \{\theta | a = c\}$ . Questo significa solo che  $(\hat{a}_N - \hat{c}_N) \rightarrow 0$ , ma i due stimatori, presi separatamente, in genere non convergono (il che è ovvio se si pensa che la successione di stime si ottiene prendendo, al variare di  $N$ , un punto di minimo di  $V_N(\theta)$  in un insieme di valori estremali del parametro che contiene infiniti punti).

Evidentemente,

1. Anche se il modello vero appartiene alla classe parametrica, il valore vero del parametro non è univocamente determinato. Qualunque  $\theta_0 = [a \ a]^\top$  ( $|a| < 1$ ), è un valore vero.
2. Per definizione lo stimatore  $\hat{\theta}_N$  è consistente se  $\hat{\theta}_N$  converge e converge al valore vero del parametro. Nel nostro caso  $\hat{\theta}_N$  non converge puntualmente e quindi non è consistente. Notare invece che  $(\hat{a}_N - \hat{c}_N)$  è uno stimatore consistente di  $a_0 - c_0 (= 0)$ .

**Problem 6.4.** Cosa accade dell'enunciato del teorema 6.2 se i dati sono semplicemente stazionari del second'ordine? Discutere un semplice esempio in cui il modello ha ordine uno e  $\mathbf{u}(t)$  è un segnale stazionario sinusoidale. Ad esempio si prenda la classe di modelli

$$M(\theta) : (1 - az^{-1})\mathbf{y}(t) = b\mathbf{u}(t - 1) + \mathbf{e}(t)$$

con  $\mathbf{u}(t) = U \sin \omega_0 t$  ed  $\mathbf{e}$  bianco. Calcolare  $\bar{V}(\theta)$  (che dipenderà dall'ampiezza del segnale di ingresso) e minimizzarlo rispetto a  $\theta$ . Verificare se il limite di  $\hat{\theta}_N$  dipende dall'ampiezza dell'ingresso. Si ha consistenza?

Cosa accade se  $\mathbf{u}$ , invece di essere un segnale deterministico stazionario, fosse un processo stazionario puramente deterministico (ad esempio una variabile aleatoria costante di media nulla) scorrelato da  $\mathbf{e}$ .

Si deve comunque garantire l'identificabilità.

Finora non ci siamo interessati molto alla stima della varianza dell'innovazione. Il seguente risultato rimedia a questa disattenzione.

**Corollary 6.1.** Nelle stesse ipotesi del teorema 6.2, lo stimatore (6.1.7) è uno stimatore consistente della varianza d'innovazione, ovvero

$$\lim_{N \rightarrow \infty} \hat{\lambda}_N^2 = \lambda_0^2 \tag{6.2.21}$$

con probabilità uno.

**Proof.** Usando la (6.1.2) possiamo scrivere

$$\varepsilon_{\hat{\theta}_N}(t) = G_{\hat{\theta}_N}(z)^{-1} [\mathbf{y}(t) - F_{\hat{\theta}_N}(z)\mathbf{u}(t)]$$

Nelle ipotesi in cui ci siamo posti,  $\hat{\theta}_N$  converge al valore vero  $\theta_0$  e quindi passando al limite per  $N \rightarrow \infty$  nell'espressione precedente è facile dimostrare che l'errore residuo di predizione converge all'innovazione vera  $\mathbf{e}_0(t) = \varepsilon_{\theta_0}(t)$ . In effetti, dato che  $V_N(\hat{\theta}_N)$  è la varianza campionaria di  $\varepsilon_{\hat{\theta}_N}(t)$  e che

$$\lim_{N \rightarrow \infty} V_N(\hat{\theta}_N) = \mathbb{E}_{\theta_0} \varepsilon_{\theta_0}^2(t) = \mathbb{E}_0 \mathbf{e}_0^2(t)$$

si vede che anche la varianza di  $\varepsilon_{\hat{\theta}_N}(t)$  converge a quella di  $\mathbf{e}_0(t)$ .  $\square$

### Il caso di parametrizzazioni indipendenti

Consideriamo un modello del tipo Box-Jenkins (3.2.1) in cui le funzioni di trasferimento  $F(z)$  e  $G(z)$  sono parametrizzate in modo indipendente. Questo significa che si può decomporre  $\theta$  in due sottovettori indipendenti; i.e.  $\theta = [\xi \ \eta]^T \in \Xi \times E = \Theta$  per cui

$$F_\theta(z) \equiv F_\xi(z), \quad G_\theta(z) \equiv G_\eta(z) \tag{6.2.22}$$

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 trasferimento (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.

**Definition 6.1.** *Si assuma che le funzioni di trasferimento  $F(z)$  e  $G(z)$  nella famiglia (3.2.1) 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 identificabilità (globale) della mappa ingresso-uscita se*

$$S_{\mathbf{y}\mathbf{u}}(\cdot; \xi_1) = S_{\mathbf{y}\mathbf{u}}(\cdot; \xi_2) \Rightarrow \xi_1 = \xi_2 \quad (6.2.23)$$

*Se si ha iniettività locale in un intorno di  $\xi_0$ , si parla di identificabilità locale in  $\xi_0$ .*

Notiamo che per l’assenza di reazione si ha  $S_{\mathbf{y}\mathbf{u}}(z; \xi) = F_{\xi}(z)S_{\mathbf{u}}(z)$ , per cui la (6.2.23) è equivalente alla

$$[F_{\xi_1}(z) - F_{\xi_2}(z)] S_{\mathbf{u}}(z) \equiv 0 \Rightarrow \xi_1 = \xi_2. \quad (6.2.24)$$

Naturalmente nel caso di ingressi con righe spettrali bisognerebbe a rigore riscrivere il primo termine come un integrale rispetto alla distribuzione spettrale  $d\hat{F}_{\mathbf{u}}(z)$ .

**Theorem 6.3.** *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  $\hat{\mathbf{y}}_{\theta}(t | t - 1)$ . Supponiamo inoltre che nella classe parametrica dei modelli  $\mathcal{M} \equiv \{M(\theta); \theta \in \Theta\}$ ,  $F(z)$  e  $G(z)$  siano parametrizzate in modo indipendente come descritto più sopra e che il processo  $\mathbf{u}$  sia sufficientemente eccitante da garantire l’identificabilità della classe di modelli  $\mathcal{F} := \{F_{\xi}(z); \xi \in \Xi\}$ .*

*Allora, se il processo (vero) che genera i dati è descritto da una funzione di trasferimento  $F_0(z)$  che appartiene alla stessa classe parametrica  $\mathcal{F}$  delle funzioni  $F_{\xi}(z)$  scelte per l’identificazione, ovvero esiste  $\xi_0 \in \Xi$  tale che  $F_0(z) \equiv F_{\xi_0}(z)$ , si ha:*

$$\lim_{N \rightarrow \infty} \hat{\xi}_N = \xi_0 \quad (6.2.25)$$

*con probabilità uno. In altri termini, lo stimatore PEM del parametro  $\xi$  è consistente.*

**Proof.** Usando l’espressione (6.1.2) 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

$$\varepsilon_{\theta}(t) = G_{\eta}(z)^{-1} [(F_0(z) - F_{\xi}(z))\mathbf{u}(t) + G_0(z)\mathbf{e}_0(t)] := L_{\xi, \eta}(z)\mathbf{u}(t) + M_{\eta}(z)\mathbf{e}_0(t)$$

dove i due addendi nell’ultimo termine sono scorrelati per l’assenza di reazione. Si trova così

$$\bar{V}(\theta) = \bar{V}(\xi, \eta) = \text{var} [L_{\xi, \eta}(z)\mathbf{u}(t)] + \text{var} [M_{\eta}(z)\mathbf{e}_0(t)]$$



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  $\hat{\theta}_N$  converge con probabilità uno, sicuramente anche le sue prime componenti,  $\hat{\xi}_N$  convergono e convergono necessariamente all'insieme dei punti di  $\Xi$  in cui si ha il minimo del primo addendo. Ma questo insieme è costituito dal solo punto  $\{\xi_0\}$ .  $\square$

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 trasferimento 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

$$\mathbf{y}(t) = F_\theta(z)\mathbf{u}(t) + \mathbf{e}(t) \tag{6.2.26}$$

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.

**Example 6.1.** *Si vuole identificare il sistema "vero"*

$$\mathbf{y}(t) = \frac{b_0}{1 + a_0z^{-1}}\mathbf{u}(t - 1) + \mathbf{e}_0(t) + c_0\mathbf{e}_0(t - 1)$$

dove  $\mathbf{u}$  ed  $\mathbf{e}_0$  sono bianchi, scorrelati, di varianze rispettive  $\sigma^2$  e  $\lambda_0^2$ , usando un modello a errore sull'uscita (O.E.) di ordine 1,

$$\mathbf{y}(t) = \frac{b}{1 + az^{-1}}\mathbf{u}(t - 1) + \mathbf{e}(t).$$

dove  $\mathbf{e}$  è rumore bianco. Trovare l'insieme dei valori del parametro  $\theta := [a \ b]^\top$  a cui converge (quando la numerosità campionaria  $N \rightarrow \infty$ ) 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

$$\hat{\mathbf{y}}_0(t | t - 1) = \frac{b_0}{1 + a_0z^{-1}}\mathbf{u}(t - 1) + c_0\mathbf{e}_0(t - 1)$$

mentre quello per il modello si scrive

$$\hat{\mathbf{y}}_\theta(t | t - 1) = \frac{b}{1 + az^{-1}}\mathbf{u}(t - 1).$$

La varianza dell'errore di predizione si esprime quindi come,

$$\begin{aligned} \text{var } \epsilon_\theta &= \lambda_0^2 + \text{var} [\hat{\mathbf{y}}_0(t | t - 1) - \hat{\mathbf{y}}_\theta(t | t - 1)] \\ &= \lambda_0^2 + \text{var} \left\{ \left[ \frac{b_0}{1 + a_0z^{-1}} - \frac{b}{1 + az^{-1}} \right] \mathbf{u}(t - 1) + c_0\mathbf{e}_0(t - 1) \right\} \\ &= \lambda_0^2 + \text{var} \left\{ \left[ \frac{b_0}{1 + a_0z^{-1}} - \frac{b}{1 + az^{-1}} \right] \mathbf{u}(t - 1) \right\} + c_0^2\lambda_0^2 \end{aligned}$$

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  $\theta$  prendendo

$$\left[ \frac{b_0}{1 + a_0 z^{-1}} - \frac{b}{1 + a z^{-1}} \right] = 0$$

il che accade se e solo se  $a = a_0$  e  $b = b_0$ . In questo senso lo stimatore  $\hat{\theta}_N$  dei parametri della funzione di trasferimento  $F(z)$  è “consistente” i.e. converge ai valori veri  $\theta_0 := [a_0 \ b_0]^\top$  anche se a rigore con questa classe di modelli non si può avere consistenza. Ovviamente a questa conclusione si sarebbe potuti arrivare direttamente in base al teorema 6.3.

### Il “metodo” dei minimi quadrati

Si consideri il modello “vero”

$$\mathbf{y}(t) = F(z) \mathbf{u}(t) + \mathbf{e}_0(t),$$

dove  $F(z)$  è una funzione di trasferimento causale (non necessariamente razionale),  $\{\mathbf{u}(t)\}$  ed  $\{\mathbf{e}_0(t)\}$  sono processi ergodici a media nulla tra loro indipendenti e  $\{\mathbf{e}_0(t)\}$  è bianco.

Si cerca di approssimare la  $F(z)$  con una funzione di trasferimento razionale di grado fissato usando i dati  $\{\mathbf{y}(t)\}$  e  $\{\mathbf{u}(t)\}$  con  $t = 1, \dots, N$ , identificando un modello di tipo “output error”

$$\mathbf{y}(t) = \frac{B(z^{-1})}{A(z^{-1})} \mathbf{u}(t) + \mathbf{e}(t) \tag{6.2.27}$$

dove  $A(z^{-1}) = 1 + a_1 z^{-1} + a_2 z^{-2} + \dots + a_n z^{-n}$  e  $B(z^{-1}) = b_1 z^{-1} + b_2 z^{-2} + \dots + b_m z^{-m}$ . L'identificazione col metodo PEM conduce (come vedremo tra un attimo) ad un problema di stima non lineare perchè i parametri  $\{a_k\}$  compaiono in modo non lineare nel predittore. Per questo motivo qualche volta si usa un metodo empirico che viene genericamente (e impropriamente) chiamato *metodo dei minimi quadrati* che descriviamo qui di seguito.

Usando l'usuale definizione

$$\varphi(t) := [-\mathbf{y}(t-1) \ \dots \ -\mathbf{y}(t-n) \ \mathbf{u}(t-1) \ \dots \ \mathbf{u}(t-m)]^\top$$

e la notazione  $\theta = [a_1 \ a_2 \ \dots \ a_n \ b_1 \ b_2 \ \dots \ b_m]^\top$  per il vettore dei coefficienti, il modello (6.2.27) si può scrivere in forma di regressione

$$\mathbf{y}(t) = \varphi(t)^\top \theta + \mathbf{w}(t) \quad t = 1, \dots, N,$$

dove però  $\mathbf{w}(t) := A(z^{-1})\mathbf{e}(t)$  non è ovviamente rumore bianco. Nonostante ciò si stimano i coefficienti  $\theta$  col metodo dei minimi quadrati ottenendo uno stimatore  $\hat{\theta}_{LS}(N)$  che è descritto dalla nota formula

$$\hat{\theta}_{LS}(N) = \left[ \sum_{t=1}^N \varphi(t) \varphi(t)^\top \right]^{-1} \sum_{t=1}^N \varphi(t) \mathbf{y}(t) \tag{6.2.28}$$

Ci si chiede quando questa espressione fornisca stime consistenti. Assumeremo che  $F(z)$  appartenga alla classe di funzioni razionali definite sopra, per cui si può scrivere

$$\mathbf{y}(t) = \varphi(t)^\top \theta_0 + \mathbf{w}_0(t)$$

con  $\mathbf{w}_0(t) := [\mathbf{e}_0(t) \ \mathbf{e}_0(t-1) \ \dots \ \mathbf{e}_0(t-n)]^\top [1 \ a_{0,1} \ a_{0,2} \ \dots \ a_{0,n}] := \eta(t)^\top \mathbf{a}_0$ , dove il pedice 0 significa parametro "vero". Per verificare se  $\hat{\theta}_{LS}(N)$  è consistente sostituiamo l'espressione precedente nella (6.2.30) ottenendo

$$\hat{\theta}_{LS}(N) = \theta_0 + \left[ \frac{1}{N} \sum_{t=1}^N \varphi(t) \varphi(t)^\top \right]^{-1} \frac{1}{N} \sum_{t=1}^N \varphi(t) \mathbf{w}_0(t). \quad (6.2.29)$$

Passando al limite per  $N \rightarrow \infty$  e usando il teorema ergodico si trova

$$\lim_{N \rightarrow \infty} \hat{\theta}_{LS}(N) = \theta_0 + [\mathbb{E} \varphi(t) \varphi(t)^\top]^{-1} \mathbb{E} \varphi(t) \eta(t)^\top \mathbf{a}_0. \quad (6.2.30)$$

Dato che  $\mathbf{u}$  e  $\mathbf{e}_0$  sono indipendenti, la matrice di correlazione  $\mathbb{E} \varphi(t) \eta(t)^\top$  ha la seguente struttura

$$\mathbb{E} \varphi(t) \eta(t)^\top = \begin{bmatrix} \mathbb{E} \begin{bmatrix} \mathbf{y}(t-1) \\ \vdots \\ \mathbf{y}(t-n) \end{bmatrix} [\mathbf{e}_0(t) \ \mathbf{e}_0(t-1) \ \dots \ \mathbf{e}_0(t-n)] \\ 0 \ 0 \ \dots \ 0 \end{bmatrix}$$

la quale ha la prima colonna di zeri ma, dato che  $\mathbf{y}(t-k)$  e  $\mathbf{e}_0(t-j)$  hanno correlazione diversa da zero se  $k \geq j$  ha una struttura triangolare superiore e c'è sempre un termine diverso da zero nella diagonale secondaria superiore a meno che non sia  $A_0(z^{-1}) \equiv 1$  e il modello vero non si riduca al modello FIR

$$\mathbf{y}(t) = B_0(z^{-1}) \mathbf{u}(t) + \mathbf{e}_0(t).$$

Quindi il cosiddetto "metodo dei minimi quadrati" è consistente solo se il modello è di tipo FIR con rumore bianco additivo.

Lo stimatore PEM, per un modello a errore d'equazione come (6.2.27), è basato sul predittore di Wiener

$$\hat{\mathbf{y}}_\theta(t | t-1) = \frac{B(z^{-1})}{A(z^{-1})} \mathbf{u}(t)$$

e questo predittore non è **mai lineare nei parametri** (cioè esprimibile nella forma  $\varphi(t)^\top \theta$ ), a meno che non sia  $A(z^{-1}) \equiv 1$ . Quindi lo stimatore  $\hat{\theta}_{LS}(N)$  è uno stimatore PEM solo nel caso in cui  $A(z^{-1}) \equiv 1$ .

Lo stimatore PEM di  $\theta$  per un modello a errore d'equazione va calcolato mediante uno degli algoritmi di ottimizzazione iterativa che descriveremo nella sezione 6.7.

### 6.3 La distribuzione asintotica dello stimatore PEM

In questa sezione supporremo sempre di essere nelle condizioni che garantiscono la consistenza dello stimatore PEM, ovvero, supporremo (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 sono modelli per i quali il predittore di un passo è una funzione razionale (e quindi analitica) del parametro  $\theta$  per cui il minimo della cifra di merito  $V_N(\theta)$  si ha in un punto in cui il gradiente si annulla, ovvero

$$\frac{\partial V_N(\theta)}{\partial \theta} := V_N(\theta)' = 0.$$

Ora, usando la formula di Taylor arrestata al secondo ordine nel punto  $\theta = \theta_0$ , si ha

$$V_N(\hat{\theta}_N)' = V_N(\theta_0)' + V_N(\bar{\theta})''(\hat{\theta}_N - \theta_0) = 0 \quad (6.3.1)$$

dove  $V_N(\bar{\theta})''$  è la matrice delle derivate seconde (Hessiana) calcolata in un punto  $\bar{\theta}$  dell'intervallo  $p$ -dimensionale di estremi  $\theta_0$  e  $\hat{\theta}_N$ , ovvero

$$\theta_0^k \leq \bar{\theta}^k \leq \hat{\theta}_N^k, \quad k = 1, 2, \dots, p$$

Supponendo che la matrice Hessiana sia invertibile, dalla (6.3.1) si può ricavare

$$\hat{\theta}_N - \theta_0 = - \left[ \frac{1}{2} V_N(\bar{\theta})'' \right]^{-1} \frac{1}{2} V_N(\theta_0)' \quad (6.3.2)$$

dove il fattore  $\frac{1}{2}$  è stato introdotto per convenienza. Calcoliamo ora il gradiente e la matrice Hessiana usando l'espressione (6.1.4). Ponendo

$$\psi_\theta(t) := \frac{\partial \varepsilon_\theta(t)}{\partial \theta} = - \frac{\partial \hat{y}_\theta(t | t-1)}{\partial \theta}$$

si trova

$$\frac{1}{2} V_N(\theta)' = \frac{1}{N} \sum_{t=1}^N \psi_\theta(t) \varepsilon_\theta(t) \quad (6.3.3)$$

$$\frac{1}{2} V_N(\theta)'' = \frac{1}{N} \sum_{t=1}^N \left\{ \psi_\theta(t) \psi_\theta(t)^\top + \varepsilon_\theta(t) \left[ \frac{\partial^2 \varepsilon_\theta(t)}{\partial \theta_i \partial \theta_j} \right] \right\} \quad (6.3.4)$$

Esaminiamo prima il comportamento asintotico della derivata seconda.

**Lemma 6.2.** *Nelle ipotesi poste, si ha*

$$\lim_{N \rightarrow \infty} \frac{1}{2} V_N(\bar{\theta})'' = \mathbb{E}_{\theta_0} \{ \psi_{\theta_0}(t) \psi_{\theta_0}(t)^\top \} \quad (6.3.5)$$

con probabilità uno.

*Proof.* Nelle ipotesi in cui ci siamo messi,  $\hat{\theta}_N \rightarrow \theta_0$  e quindi anche  $\bar{\theta} \rightarrow \theta_0$  (con probabilità uno) e la media temporale in (??) tende all'aspettazione per cui,

$$\frac{1}{2}V_N(\bar{\theta})'' \rightarrow \mathbb{E}_{\theta_0} \left\{ \psi_{\theta_0}(t)\psi_{\theta_0}(t)^\top + \varepsilon_{\theta_0}(t) \left[ \frac{\partial^2 \varepsilon_{\theta_0}(t)}{\partial \theta_i \partial \theta_j} \right]_{|\theta=\theta_0} \right\}$$

Inoltre, dato che il modello vero appartiene alla classe dei modelli assegnata, si deve avere  $\varepsilon_{\theta_0}(t) = \mathbf{e}_0(t)$ . Infine, dato che sia il gradiente ( $\psi_{\theta_0}(t)$ ), che la derivata seconda di  $\hat{y}_{\theta_0}(t | t-1)$  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 (??) risultano scorrelati da  $\mathbf{e}_0(t)$  e l'ultimo termine ha quindi aspettazione nulla.  $\square$

Per quanto riguarda l'altro termine nel prodotto (6.3.2), si ha

$$\frac{1}{2}V_N(\theta_0)' = \frac{1}{N} \sum_{t=1}^N \psi_{\theta_0}(t)\mathbf{e}_0(t) \tag{6.3.6}$$

**Proposition 6.1.** *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  $\{\psi_{\theta_0}(t)\mathbf{e}_0(t)\}$  è una d-martingala e vale il teorema del limite centrale,*

$$\sqrt{N} \frac{1}{2}V_N(\theta_0)' \xrightarrow{L} \mathcal{N}(0, Q) \tag{6.3.7}$$

Se la varianza condizionata di  $\mathbf{e}_0(t)$  è indipendente dai dati ( $\mathbf{y}^{t-1}$ ,  $\mathbf{u}^{t-1}$ ), ovvero se

$$\mathbb{E}_0\{\mathbf{e}_0(t)^2 | \mathbf{y}^{t-1}, \mathbf{u}^{t-1}\} = \mathbb{E}_0\{\mathbf{e}_0(t)^2\} = \lambda_0^2, \tag{6.3.8}$$

la matrice varianza asintotica  $Q$  è data dalla formula,

$$Q = \lambda_0^2 \mathbb{E}_0\{\psi_{\theta_0}(t)\psi_{\theta_0}(t)^\top\}. \tag{6.3.9}$$

*Proof.* Il risultato scende dall'osservazione che  $\{\psi_{\theta_0}(t)\mathbf{e}_0(t)\}$  è ancora una d-martingala rispetto alla stessa famiglia di  $\sigma$ -algebre ed è in realtà un corollario immediato del teorema del limite centrale per d-martingale 5.2. L'unica cosa da dimostrare è l'espressione per la varianza asintotica, la quale scende dalla proprietà (6.3.8), che implica,

$$\begin{aligned} \text{Var} \{ \psi_{\theta_0}(t)\mathbf{e}_0(t) \} &= \mathbb{E}_0\{ \mathbb{E}_0 [ \mathbf{e}_0(t)^2 | \mathbf{y}^{t-1}, \mathbf{u}^{t-1} ] \psi_{\theta_0}(t)\psi_{\theta_0}(t)^\top \} = \\ &= \mathbb{E}_0\{ \mathbf{e}_0(t)^2 \} \mathbb{E}_0\{ \psi_{\theta_0}(t)\psi_{\theta_0}(t)^\top \}. \end{aligned}$$

$\square$

**Remark 6.3.** Notiamo che le condizioni del teorema valgono in particolare se  $\mathbf{e}_0$  è un processo i.i.d. ma in realtà la condizione di d-martingala è molto più debole.

Essa si può riformulare in questo contesto dicendo che per il processo di osservazione descritto dal modello vero, il predittore non lineare (la media condizionata) di  $\mathbf{y}(t)$  data la storia passata  $(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})$ , è una *funzione lineare dei dati*. Infatti in questo caso l'errore di predizione in senso stretto,  $\mathbf{y}(t) - \mathbb{E}[\mathbf{y}(t) \mid \mathbf{y}^{t-1}, \mathbf{u}^{t-1}]$  (che è una d-martingala) coincide necessariamente con l'usuale l'innovazione (in senso debole)  $\mathbf{e}_0(t)$ . Per modelli che descrivono piccole variazioni dei segnali in gioco è ragionevole pensare che il predittore (in senso stretto) si possa in genere approssimare con una funzione lineare dei dati passati.

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.

**Theorem 6.4.** *Supponiamo che i dati siano ergodici del secondo ordine, il modello vero appartenga alla classe parametrica  $\{M(\theta)\}$  e che vi sia identificabilità (locale) in  $\theta_0$ . Assumiamo inoltre che il processo innovazione  $\mathbf{e}_0$  soddisfi alle condizioni descritte nella proposizione 6.1. Allora per lo stimatore PEM vale il teorema del limite centrale, ovvero*

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \xrightarrow{L} \mathcal{N}(0, P), \quad (6.3.10)$$

dove la varianza asintotica  $P$  è data dalla formula

$$P = \lambda_0^2 [\mathbb{E}_0\{\psi_{\theta_0}(t)\psi_{\theta_0}(t)^\top\}]^{-1}. \quad (6.3.11)$$

l'inversa della matrice tra parentesi quadre esiste.

**Proof.** 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

$$P = [\mathbb{E}_{\theta_0}\{\psi_{\theta_0}(t)\psi_{\theta_0}(t)^\top\}]^{-1} Q [\mathbb{E}_{\theta_0}\{\psi_{\theta_0}(t)\psi_{\theta_0}(t)^\top\}]^{-1}$$

dove  $Q$  è la varianza asintotica del limite (6.3.7). La questione dell'invertibilità verrà chiarita nella prossima sezione. Si veda il corollario 6.2.  $\square$

**Example 6.2.** *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  $\sigma^2$  e  $\lambda_0^2$ , 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:*

$$M_2(\theta) := \{ \mathbf{y}(t) + a\mathbf{y}(t-1) = b_1\mathbf{u}(t-1) + b_2\mathbf{u}(t-2) + \mathbf{e}(t); \theta = [a \ b_1 \ b_2]^\top \}$$

Confrontate i risultati in termini di varianze delle stime.

Soluzione:

Identificazione con il metodo PEM usando modelli del tipo Box-Jenkins: Il modello vero appartiene alla classe  $M_1$  e si ha identificabilità, quindi lo stimatore PEM è consistente e  $\hat{\theta}_N = [\hat{a}_N \ \hat{b}_N]^\top$  converge al parametro vero  $\theta_0 = [a_0 \ b_0]^\top$ .

Identificazione con il metodo PEM usando modelli del tipo ARX: Il modello vero appartiene anche alla classe  $M_2$  e si ha identificabilità, quindi lo stimatore PEM è anch'esso consistente e  $\hat{\theta}_N$  converge al parametro vero che per questo modello è  $\theta_0 = [a_0 \ b_0 \ b_0 a_0]^\top$ . In altri termini, quando  $N \rightarrow \infty$ ,  $\hat{b}_{2,N} \rightarrow b_0 a_0$ .

Calcolo della varianza asintotica dei due stimatori.

- Per i modelli Box-Jenkins:

$$\varepsilon_\theta(t) = (1 + az^{-1})[\mathbf{y}(t) - b\mathbf{u}(t-1)] = (1 + az^{-1})[\mathbf{y}(t) - b\mathbf{u}(t-1)]$$

Il gradiente si calcola facilmente:

$$\psi_\theta(t)^\top = \left[ \frac{\partial \varepsilon_\theta(t)}{\partial a} \quad \frac{\partial \varepsilon_\theta(t)}{\partial b} \right] = [\mathbf{y}(t-1) - b\mathbf{u}(t-2) \quad -\mathbf{u}(t-1) - a\mathbf{u}(t-2)]$$

e si vede subito che  $\mathbf{y}(t-1) - b\mathbf{u}(t-2) = (1 + az^{-1})^{-1}\mathbf{e}(t-1)$ . Quindi le due componenti del gradiente sono scorrelate. Per calcolare la matrice varianza serve:

$$R := \mathbb{E}_0 \psi_\theta(t) \psi_\theta(t)^\top |_{\{\theta=\theta_0\}} = \begin{bmatrix} \frac{\lambda_0^2}{1-a_0^2} & 0 \\ 0 & \sigma^2(1+a_0)^2 \end{bmatrix}$$

che porge,

$$\text{Var} \{ \hat{\theta}_N \} \sim \frac{\lambda_0^2}{N} R^{-1} = \frac{\lambda_0^2}{N} \begin{bmatrix} \frac{1-a_0^2}{\lambda_0^2} & 0 \\ 0 & \frac{1}{\sigma^2(1+a_0)^2} \end{bmatrix}.$$

- 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

$$\psi_\theta(t) = \left[ \frac{\partial \varepsilon_\theta(t)}{\partial a} \quad \frac{\partial \varepsilon_\theta(t)}{\partial b_1} \quad \frac{\partial \varepsilon_\theta(t)}{\partial b_2} \right] = [\mathbf{y}(t-1) \quad -\mathbf{u}(t-1) \quad -\mathbf{u}(t-2)]$$

Si vede facilmente che  $\mathbb{E}_0 \mathbf{y}(t-1)^2 = \text{var } \mathbf{y}(t) = b_0^2 \sigma^2 + \frac{\lambda_0^2}{1-a_0^2}$ , per cui

$$R := \mathbb{E}_0 \psi_{\theta}(t) \psi_{\theta}(t)^{\top} |_{\{\theta=\theta_0\}} = \begin{bmatrix} b_0^2 \sigma^2 + \frac{\lambda_0^2}{1-a_0^2} & 0 & -b_0 \sigma^2 \\ 0 & \sigma^2 & 0 \\ -b_0 \sigma^2 & 0 & \sigma^2 \end{bmatrix}$$

e infine

$$\text{Var } \{\hat{\theta}_N\} \sim \frac{\lambda_0^2}{N} R^{-1} = \frac{\lambda_0^2}{N} \begin{bmatrix} \frac{1-a_0^2}{\lambda_0^2} & 0 & \frac{b_0(1-a_0^2)}{\lambda_0^2} \\ 0 & \frac{1}{\sigma^2} & 0 \\ \frac{b_0(1-a_0^2)}{\lambda_0^2} & 0 & \frac{b_0^2(1-a_0^2)}{\lambda_0^2} + \frac{1}{b_0^2} \end{bmatrix}$$

Si osserva che la varianza asintotica di  $\hat{b}_1$  è minore con il primo modello che ha meno parametri.

**Problem 6.5.** Si identifica con il metodo PEM un modello ARMA (d'innovazione) appartenente alla famiglia

$$(1 + az^{-1})\mathbf{y}(t) = (1 + cz^{-1})\mathbf{e}(t).$$

sapendo che  $\mathbf{e}_0$  è un processo i.i.d. a media zero. Si vuole

1. Calcolare la varianza asintotica,  $\hat{\sigma}_N^2$ , della differenza  $\hat{a}_N - \hat{c}_N$ ,
2. Trovare la distribuzione asintotica del rapporto

$$\mathbf{x}_N := \frac{(\hat{a}_N - \hat{c}_N)^2}{\hat{\sigma}_N^2}$$

3. Usare il risultato precedente per verificare l'ipotesi  $H_0 := \{a_0 = c_0\}$ .

*Soluzione:* È implicito nel testo che il modello vero

$$(1 + a_0 z^{-1})\mathbf{y}(t) = (1 + c_0 z^{-1})\mathbf{e}_0(t)$$

appartiene alla famiglia parametrica che si usa per l'identificazione e quindi, nelle ipotesi poste, lo stimatore PEM è consistente e asintoticamente normale. Detto  $\theta := [a \ c]^{\top}$ , la varianza asintotica di  $\hat{\theta}_N$  è data dalla nota formula

$$\hat{\Sigma}_N = \frac{\lambda_0^2}{N} [\mathbb{E}_{\theta_0} \psi_{\theta_0}(t) \psi_{\theta_0}(t)^{\top}]^{-1}$$

dove

$$\psi_{\theta_0}(t) = \frac{1}{1 + c_0 z^{-1}} \begin{bmatrix} \mathbf{y}(t-1) \\ -\mathbf{e}_0(t-1) \end{bmatrix} = \begin{bmatrix} \frac{1}{1 + a_0 z^{-1}} \mathbf{e}_0(t-1) \\ -\frac{1}{1 + c_0 z^{-1}} \mathbf{e}_0(t-1) \end{bmatrix}$$



da cui si ricava

$$\hat{\Sigma}_N = \frac{\lambda_0^2}{N} \begin{bmatrix} \frac{\lambda_0^2}{1 - a_0^2} & -\frac{\lambda_0^2}{(1 - a_0 c_0)} \\ -\frac{\lambda_0^2}{(1 - a_0 c_0)} & \frac{\lambda_0^2}{1 - c_0^2} \end{bmatrix}^{-1}$$

e quindi

$$\hat{\Sigma}_N = \frac{(1 - a_0 c_0)}{(a_0 - c_0)^2 N} \begin{bmatrix} (1 - a_0^2)(1 - a_0 c_0) & (1 - a_0^2)(1 - c_0^2) \\ (1 - a_0^2)(1 - c_0^2) & (1 - c_0^2)(1 - a_0 c_0) \end{bmatrix}.$$

Ora  $a - c = [1 \ -1] \theta$  e quindi  $\hat{a}_N - \hat{c}_N = [1 \ -1] \hat{\theta}_N$  da cui si ricava immediatamente  $\hat{\sigma}_N^2 = [1 \ -1] \hat{\Sigma}_N [1 \ -1]^T$  per cui

$$\begin{aligned} N \hat{\sigma}_N^2 &= \frac{(1 - a_0 c_0)}{(a_0 - c_0)^2} [1 \ -1] \begin{bmatrix} (1 - a_0^2)(1 - a_0 c_0) & (1 - a_0^2)(1 - c_0^2) \\ (1 - a_0^2)(1 - c_0^2) & (1 - c_0^2)(1 - a_0 c_0) \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \\ &= \frac{(1 - a_0 c_0)}{(a_0 - c_0)^2} (a_0 - c_0)^2 = 1 - a_0 c_0 := \sigma_0^2. \end{aligned}$$

Notiamo che, mentre per  $a_0 = c_0$  la varianza matriciale  $\hat{\Sigma}_N$  diventa infinita ( $\mathbb{E}_{\theta_0} \psi_{\theta_0}(t) \psi_{\theta_0}(t)^T$  diventa singolare), la varianza asintotica di  $\hat{a}_N - \hat{c}_N$  è ben definita anche per  $a_0 = c_0$ .

Dato che asintoticamente  $\sqrt{N}(\hat{a}_N - \hat{c}_N) \sim \mathcal{N}(a_0 - c_0, \sigma_0^2)$  e  $N \hat{\sigma}_N^2 \rightarrow \sigma_0^2$ , se  $a_0 = c_0$ , in base al teorema di Slutsky si ha che

$$\mathbf{x}_N = \frac{(\hat{a}_N - \hat{c}_N)^2}{\hat{\sigma}_N^2} \xrightarrow{L} \chi^2(1)$$

(altrimenti se  $a_0 \neq c_0$ , il rapporto converge ad una distribuzione  $\chi^2(1)$  non centrale).

Si potrebbe mostrare che la statistica  $\mathbf{x}_N$  è proprio quella del test  $F$  (asintotico) per confrontare due modelli ARMA di ordine uno, per i quali

$$\begin{aligned} H_0 : & \quad a_0 = c_0 ; \text{ i.e. } \mathbf{y}(t) = \mathbf{e}_0(t) \\ H_1 : & \quad a_0 \neq c_0 ; \text{ i.e. } \mathbf{y}(t) \neq \mathbf{e}_0(t). \end{aligned}$$

□

**Problem 6.6.** Si vuole identificare il sistema "vero"

$$(1 + a_0 z^{-1})\mathbf{y}(t) = b_0 \mathbf{u}(t - 1) + \mathbf{e}_0(t) \quad |a_0| < 1$$

dove  $\mathbf{u}$  ed  $\mathbf{e}_0$  sono processi i.i.d., scorrelati, di varianze rispettive  $\sigma^2$  e  $\lambda_0^2$ . Per l'identificazione si usano modelli ARMAX di ordine 1,

$$(1 + a z^{-1})\mathbf{y}(t) = b \mathbf{u}(t - 1) + (1 + c z^{-1})\mathbf{e}(t).$$

dove  $\mathbf{e}$  è rumore bianco. Trovare l'insieme dei valori del parametro  $\theta := [a \ b \ c]^\top$  a cui converge (quando la numerosità campionaria  $N \rightarrow \infty$ ) lo stimatore a minimo errore di predizione  $\hat{\theta}_N$ .

Lo stimatore è consistente? Dare un'espressione per la varianza asintotica di  $\hat{c}_N$ .

Soluzione.

È evidente che il sistema "vero" appartiene alla classe parametrica di modelli AR-MAX scelti per l'identificazione (corrisponde infatti alla scelta  $a = a_0$ ,  $b = b_0$ ,  $c = 0$ ). Pertanto, nelle ipotesi poste, (teorema 8.2 delle dispense) lo stimatore PEM è consistente, i.e. converge per  $N \rightarrow \infty$  ai parametri del sistema vero,

$$\hat{\theta}_N := [\hat{a}_N \ \hat{b}_N \ \hat{c}_N]^\top \rightarrow [a_0 \ b_0 \ 0]^\top.$$

In queste condizioni si può calcolare la varianza asintotica di  $\hat{c}_N$  usando la formula 3.11 del capitolo 8 delle dispense

$$P = \lambda_0^2 [\mathbb{E}_0\{\psi_{\theta_0}(t)\psi_{\theta_0}(t)^\top\}]^{-1}.$$

Dalle formule per il gradiente

$$(1+cz^{-1})\psi_a(t) = y(t-1) \quad (1+cz^{-1})\psi_b(t) = -u(t-1) \quad (1+cz^{-1})\psi_c(t) = -\varepsilon_\theta(t-1)$$

per  $\theta = \theta_0$  si ricava

$$\psi_{a_0}(t) = \mathbf{y}(t-1) \quad \psi_{b_0}(t) = -\mathbf{u}(t-1) \quad \psi_{c_0}(t) = -\varepsilon_{\theta_0}(t-1) = -\mathbf{e}_0(t-1)$$

e dato che  $y(t)$  dipende in modo strettamente causale dalla storia passata di  $u$  si ha

$$P^{-1} = \mathbb{E}_0\{\psi_{\theta_0}(t)\psi_{\theta_0}(t)^\top\} = \begin{bmatrix} \sigma_y(0) & 0 & -\lambda_0^2 \\ 0 & \sigma_u^2 & 0 \\ -\lambda_0^2 & 0 & \lambda_0^2 \end{bmatrix}$$

la cui inversa è

$$P = \begin{bmatrix} \frac{1}{\sigma_y(0) - \lambda_0^2} & 0 & \frac{1}{\sigma_y(0) - \lambda_0^2} \\ 0 & \frac{1}{\sigma_u^2} & 0 \\ \frac{1}{\sigma_y(0) - \lambda_0^2} & 0 & \frac{\sigma_y(0)}{\lambda_0^2(\sigma_y(0) - \lambda_0^2)} \end{bmatrix}$$

Ne segue che la varianza asintotica di  $\hat{c}_N$  è  $\frac{\lambda_0^2}{N} P_{3,3} = \frac{1}{N} \frac{\sigma_y(0)}{\sigma_y(0) - \lambda_0^2}$ .

Rimane da calcolare  $\sigma_y(0)$ . Usando il modello

$$\mathbf{y}(t) = -a_0\mathbf{y}(t-1) + b_0\mathbf{u}(t-1) + \mathbf{e}_0(t)$$

si trova facilmente

$$\sigma_y(0) - \lambda_0^2 = \frac{b_0^2\sigma_u^2 + \lambda_0^2}{1 - a_0^2} - \lambda_0^2 = \frac{b_0^2\sigma_u^2 + a_0^2\lambda_0^2}{1 - a_0^2}$$

e quindi, per  $N \rightarrow \infty$ ,

$$\text{var } \hat{c}_N \simeq \frac{1}{N} \frac{b_0^2\sigma_u^2 + \lambda_0^2}{b_0^2\sigma_u^2 + a_0^2\lambda_0^2}.$$

□

## 6.4 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

$$p_{\theta}(y^N, u^N), \quad \theta \in \Theta$$

Useremo i simboli  $y_t, u_t, y^t, u^t, \dots$  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<sup>28</sup>

$$\begin{aligned} p_{\theta}(y^N, u^N) &= p_{\theta}(y_N, u_N | y^{N-1}, u^{N-1}) p_{\theta}(y^{N-1}, u^{N-1}) \\ &= p_{\theta}(y_N, | y^{N-1}, u^{N-1}) p(u_N | y^N, u^{N-1}) p_{\theta}(y^{N-1}, u^{N-1}) \\ &= \dots \\ &= \prod_{t=1}^N p_{\theta}(y_t | y^{t-1}, u^{t-1}) \prod_{t=1}^N p(u_t | y^t, u^{t-1}) \end{aligned} \quad (6.4.1)$$

Abbiamo soppresso la dipendenza dal parametro  $\theta$  nelle probabilità condizionate  $p(u_t | y^t, u^{t-1})$ ,  $t = 1, 2, \dots$  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

$$\mathbf{y}(t) = \mathbb{E}_{\theta} [\mathbf{y}(t) | \mathbf{y}^{t-1}, \mathbf{u}^{t-1}] + \mathbf{e}(t) := \hat{\mathbf{y}}_{\theta}(t | t-1) + \mathbf{e}(t) \quad (6.4.2)$$

al limite per  $t \rightarrow \infty$ ,  $\mathbf{e}$  tende a diventare l'innovazione stazionaria. Assumiamo che al limite *la densità di probabilità di  $\mathbf{e}(t)$  non dipenda dal parametro  $\theta$* . Abusando di questo fatto, nella decomposizione (6.4.2) abbiamo già implicitamente ignorato la dipendenza di  $\mathbf{e}$  da  $\theta$ .

Questo è quanto accade nel caso di un modello razionale con rumore Gaussiano in cui,

$$p_{\theta}(y_t | y^{t-1}, u^{t-1}) = \frac{1}{\sqrt{2\pi\lambda_{\theta}^2(t)}} \exp -\frac{1}{2} \frac{[y_t - \hat{y}_{\theta}(t | t-1)]^2}{\lambda_{\theta}^2(t)} \quad (6.4.3)$$

e, come è ben noto dalla teoria del filtro di Kalman, la varianza dell'innovazione (transitoria)  $\lambda_{\theta}^2(t)$  converge ad una costante  $\lambda^2$  (la varianza dell'innovazione stazionaria) indipendente dal parametro  $\theta$ , quando  $t \rightarrow \infty$ .

<sup>28</sup>Notiamo che c'è una arbitrarietà strutturale nella decomposizione. Avremmo potuto egualmente scrivere i prodotti come  $p_{\theta}(y_t | y^{t-1}, u^t) p(u_t | y^{t-1}, u^{t-1})$  invece di  $p_{\theta}(y_t | y^{t-1}, u^{t-1}) p(u_t | y^t, u^{t-1})$ . La scelta fatta corrisponde ad assegnare il ritardo alla catena di azione diretta.

Confortati da questo esempio, assumeremo che  $t$  sia abbastanza grande da poter scrivere

$$p_\theta(y_t | y^{t-1}, u^{t-1}) = p_e(y_t - \hat{y}_\theta(t | t-1)) \quad (6.4.4)$$

dove  $p_e$  denota la densità di probabilità dell'innovazione stazionaria, che non dipende da  $\theta$ . In altri termini la dipendenza da  $\theta$  della  $p_\theta(y_t | y^{t-1}, u^{t-1})$  si manifesta solo attraverso il predittore (stazionario) di un passo  $\hat{y}_\theta(t | t-1)$ . Calcoliamo allora il vettore delle sensitività,

$$\mathbf{z}_\theta := \frac{\partial \log p_\theta(\mathbf{y}^N \mathbf{u}^N)}{\partial \theta} = \sum_{t=1}^N \frac{\partial \log p_\theta(\mathbf{y}(t) | \mathbf{y}^{t-1} \mathbf{u}^{t-1})}{\partial \theta}$$

Supponendo che  $N \rightarrow \infty$  e trascurando i primi termini nella somma, si può pensare di essere all'incirca in regime stazionario per cui si può assumere che valga la rappresentazione (6.4.4) in cui il predittore è quello stazionario. Senza perdita di generalità, dopo aver eliminato i termini transitori, possiamo pensare di ri-inizializzare le somme all'istante  $t = 1$ , ottenendo così,

$$\mathbf{z}_\theta = \sum_{t=1}^N \frac{\partial \log p_e(\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t | t-1))}{\partial \theta} = - \sum_{t=1}^N \frac{\partial \log p_e(\varepsilon)}{\partial \varepsilon} \Big|_{\varepsilon=\mathbf{e}(t)} \frac{\partial \hat{\mathbf{y}}_\theta(t | t-1)}{\partial \theta}$$

In questa formule il predittore è quello stazionario. Ponendo  $\frac{\partial \log p_e(\varepsilon)}{\partial \varepsilon} \Big|_{\varepsilon=\mathbf{e}(t)} := \ell'(\mathbf{e}(t))$  e

$$\frac{1}{\kappa^2} := \mathbb{E}_\theta[\ell'(\mathbf{e}(t))]^2 \quad (6.4.5)$$

si arriva così all'espressione della matrice di Fisher

$$I_N(\theta) = \mathbb{E}_\theta\{\mathbf{z}_\theta \mathbf{z}_\theta^\top\} = \sum_{t,s=1}^N \mathbb{E}_\theta\{\ell'(\mathbf{e}(t)) \ell'(\mathbf{e}(s)) \psi_\theta(t) \psi_\theta(s)^\top\} \quad (6.4.6)$$

e al seguente risultato.

**Theorem 6.5.** *Se  $\mathbf{e}$  è una  $d$ -martingala rispetto alla famiglia  $(\mathbf{y}^t, \mathbf{u}^t)$  e  $\ell'(\mathbf{e}(t))$  è una funzione lineare di  $\mathbf{e}(t)$ , il che accade in particolare se  $\mathbf{e}$  è un processo Gaussiano, si ha*

$$I_N(\theta) = \frac{N}{\kappa^2} \mathbb{E}_\theta\{\psi_\theta(t) \psi_\theta(t)^\top\} \quad (N \rightarrow \infty) \quad (6.4.7)$$

dove  $\kappa^2$  è definita in (6.4.5). Nel caso Gaussiano,  $\kappa^2 = \text{var}\{\mathbf{e}(t)\} = \lambda^2$ .

*Proof.* In effetti,

$$\begin{aligned} & \sum_{t,s=1}^N \mathbb{E}_\theta \{ \ell'(\mathbf{e}(t)) \ell'(\mathbf{e}(s)) \boldsymbol{\psi}_\theta(t) \boldsymbol{\psi}_\theta(s)^\top \} = \\ & \sum_{t=1}^N \mathbb{E}_\theta \{ \ell'(\mathbf{e}(t))^2 \boldsymbol{\psi}_\theta(t) \boldsymbol{\psi}_\theta(t)^\top \} + \\ & 2 \sum_{t>s}^N \mathbb{E}_\theta \{ \ell'(\mathbf{e}(t)) \ell'(\mathbf{e}(s)) \boldsymbol{\psi}_\theta(t) \boldsymbol{\psi}_\theta(s)^\top \} = \\ & \sum_{t=1}^N \mathbb{E}_\theta \{ \mathbb{E}_\theta [ \ell'(\mathbf{e}(t))^2 \mid \mathbf{y}^{t-1}, \mathbf{u}^{t-1} ] \boldsymbol{\psi}_\theta(t) \boldsymbol{\psi}_\theta(t)^\top \} + \\ & 2 \sum_{t>s}^N \mathbb{E}_\theta \{ \mathbb{E}_\theta [ \ell'(\mathbf{e}(t)) \mid \mathbf{y}^{t-1}, \mathbf{u}^{t-1} ] \ell'(\mathbf{e}(s)) \boldsymbol{\psi}_\theta(t) \boldsymbol{\psi}_\theta(s)^\top \} \end{aligned}$$

e il primo addendo dell'ultima a somma è uguale a (6.4.7), mentre il secondo è zero dato che  $\mathbb{E}_\theta [ \ell'(\mathbf{e}(t)) \mid \mathbf{y}^{t-1}, \mathbf{u}^{t-1} ] = 0$ .

Notiamo poi che se  $\mathbf{e}$  è Gaussiano,  $\left. \frac{\partial \log p_{\mathbf{e}}(\boldsymbol{\varepsilon})}{\partial \boldsymbol{\varepsilon}} \right|_{\boldsymbol{\varepsilon}=\mathbf{e}(t)} = -\frac{\mathbf{e}(t)}{\lambda^2}$  e quindi  $\kappa^2 = \lambda^2$ .

□

Ricordando il teorema di Rothenberg 1.2, otteniamo il seguente utile corollario.

**Corollary 6.2.** *Nelle ipotesi poste, il modello (6.4.2) è localmente identificabile in  $\theta$ , se e solo se la matrice  $\mathbb{E}_\theta \{ \boldsymbol{\psi}_\theta(t) \boldsymbol{\psi}_\theta(t)^\top \}$  è non-singolare.*

Notiamo che l'inversa della matrice di Fisher tende a zero come  $\frac{1}{N}$ . In particolare, nel caso di modelli Gaussiani si ha

$$I_N(\theta)^{-1} = \frac{\lambda^2}{N} \mathbb{E}_\theta \{ \boldsymbol{\psi}_\theta(t) \boldsymbol{\psi}_\theta(t)^\top \}^{-1}, \quad (N \rightarrow \infty).$$

Possiamo così concludere con una condizione per l'efficienza asintotica dello stimatore PEM.

**Theorem 6.6.** *Supponiamo che valgano le stesse ipotesi del teorema 6.4 e che nel modello che genera i dati il processo di innovazione  $\mathbf{e}_0$  sia Gaussiano. Allora lo stimatore PEM è asintoticamente efficiente; i.e. per  $N \rightarrow \infty$ ,*

$$\text{Var} \{ \hat{\boldsymbol{\theta}}_N \} - I_N(\theta_0)^{-1} \rightarrow 0 \tag{6.4.8}$$

Equivalentemente, per  $N \rightarrow \infty$ , la varianza di  $\hat{\boldsymbol{\theta}}_N$  coincide con l'inversa della matrice di Fisher calcolata in  $\theta_0$ .

In conclusione, possiamo affermare che sotto ipotesi ragionevoli sul meccanismo che genera i dati e sulla classe di modelli scelta per l'identificazione, il metodo PEM è asintoticamente ottimale. Naturalmente nulla sappiamo del suo comportamento per piccoli campioni.

### Relazione tra PEM e massima verosimiglianza

Sono asintoticamente equivalenti, vedere [62]. Con la MV si ottengono stime più accurate per piccoli campioni (però serve distribuzione Gaussiana!). Massima Verosimiglianza esatta per modelli AR.

## 6.5 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 (2.1.7) che abbiamo presentato nei capitoli 2 e 8 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"  $\theta_0$ ,

$$\mathbf{y} = S\theta_0 + \sigma\mathbf{w} \quad (6.5.1)$$

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  $\theta$  si può scrivere

$$\hat{\theta}_N = \left[ \frac{1}{N} S^\top S \right]^{-1} \frac{1}{N} S^\top \mathbf{y} = \theta_0 + \left[ \frac{1}{N} S^\top S \right]^{-1} \frac{1}{N} S^\top \mathbf{w}$$

ovvero

$$\hat{\theta}_N - \theta_0 = Q_N^{-1} \frac{1}{N} S^\top \mathbf{w}, \quad Q_N := \frac{1}{N} S^\top S \quad (6.5.2)$$

Nel caso di modelli dinamici del tipo (3.2.1), supponendo che il modello vero appartenga alla classe parametrica con cui si calcola il predittore, l'errore di predizione si può approssimare nel modo seguente,

$$\begin{aligned} \varepsilon_\theta(t) &= \mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t | t-1) = -(\hat{\mathbf{y}}_\theta(t | t-1) - \hat{\mathbf{y}}_{\theta_0}(t | t-1)) + \mathbf{e}_0(t) \\ &\simeq \boldsymbol{\psi}_{\theta_0}^\top(t) (\theta - \theta_0) + \mathbf{e}_0(t), \quad t = 1, 2, \dots, N \end{aligned} \quad (6.5.3)$$

a meno di termini che sono infinitesimi di ordine superiore in  $\|\theta - \theta_0\|$ . Questo è un modello lineare "incrementale" a partire dal quale si può formalmente calcolare uno stimatore ai minimi quadrati della deviazione  $\theta - \theta_0$  del parametro  $\theta$  dal

parametro vero usando la solita formula e ricavarne l'espressione

$$\tilde{\theta}_N - \theta_0 = \left[ \frac{1}{N} \sum_{t=1}^N \psi_{\theta_0}(t) \psi_{\theta_0}(t)^\top \right]^{-1} \frac{1}{N} \sum_{t=1}^N \psi_{\theta_0}(t) \mathbf{e}_0(t). \quad (6.5.4)$$

D'altro canto, nelle condizioni che garantiscono la consistenza e la normalità asintotica dello stimatore PEM, usando la (6.3.2), si può scrivere,

$$\hat{\theta}_N - \theta_0 \simeq \left[ \frac{1}{N} \sum_{t=1}^N \psi_{\theta_0}(t) \psi_{\theta_0}(t)^\top \right]^{-1} \frac{1}{N} \sum_{t=1}^N \psi_{\theta_0}(t) \mathbf{e}_0(t) \quad (6.5.5)$$

$$\simeq \bar{Q}^{-1} \frac{1}{N} \sum_{t=1}^N \psi_{\theta_0}(t) \mathbf{e}_0(t), \quad \bar{Q} := E_0 \{ \psi_{\theta_0}(t) \psi_{\theta_0}(t)^\top \} \quad (6.5.6)$$

dove il simbolo  $\simeq$  significa che i due membri di questa espressione moltiplicati per  $\sqrt{N}$  hanno lo stesso limite in legge. Possiamo pertanto concludere che lo stimatore  $\tilde{\theta}_N$  ricavato dal modello lineare (6.5.3), ha, per  $N \rightarrow \infty$ , lo stesso limite in legge dello stimatore PEM,  $\hat{\theta}_N$ .

Questo ragionamento è una giustificazione euristica del seguente risultato.

**Theorem 6.7.** *La distribuzione asintotica dello stimatore PEM del parametro  $\theta$  per un qualunque modello dinamico lineare che soddisfi le ipotesi di consistenza e normalità asintotica, coincide con la distribuzione limite dello stimatore di massima verosimiglianza del parametro  $\theta$  nel modello lineare statico Gaussiano (6.5.1), in cui si siano fatte le sostituzioni*

$$S = \begin{bmatrix} \psi_{\theta_0}(1)^\top \\ \vdots \\ \psi_{\theta_0}(N)^\top \end{bmatrix}, \quad \sigma \mathbf{w} = \begin{bmatrix} \mathbf{e}_0(1) \\ \vdots \\ \mathbf{e}_0(N) \end{bmatrix}, \quad (6.5.7)$$

e si trattino i vettori  $\{\psi_{\theta_0}(t); t = 1, 2, \dots, N\}$  come quantità deterministiche. Nella distribuzione limite si sostituisce alla matrice  $Q_N$  il suo limite per  $N \rightarrow \infty$ , uguale a  $\bar{Q}$  nella formula (6.5.6).

Usando questo risultato si può quindi stabilire una corrispondenza biunivoca tra procedimenti basati su PEM su un qualunque modello lineare dinamico che soddisfa alle ipotesi di consistenza e normalità asintotica e procedimenti di inferenza statistica sul modello statico (6.5.1).

## 6.6 Analisi asintotica delle stime di funzioni di trasferimento

Finora ci siamo occupati solo delle proprietà asintotiche di stimatori del parametro  $\theta$ , che in realtà è raramente la quantità di interesse diretto nei procedimenti di modellizzazione a partire dai dati. In questa sezione studieremo le proprietà asintotiche

delle corrispondenti funzioni di trasferimento

$$\hat{F}_N(e^{j\omega}) := F_{\hat{\theta}_N}(e^{j\omega}), \quad \hat{G}_N(e^{j\omega}) := G_{\hat{\theta}_N}(e^{j\omega}) \quad (6.6.1)$$

Lo strumento che facilita in gran modo l'analisi asintotica di questi stimatori è il teorema di Cramèr 5.7 del capitolo 5 dal quale discende immediatamente il seguente risultato.

**Theorem 6.8.** *Sia  $W_\theta(e^{j\omega})$  una funzione razionale di  $e^{j\omega}$ , in genere a valori vettoriali in  $\mathbb{C}^r$ , che dipende in modo regolare dal parametro  $\theta$  e*

$$J_\theta(e^{j\omega}) := \left[ \frac{\partial}{\partial \theta_k} W_\theta(e^{j\omega}) \right]_{k=1,2,\dots,p}$$

la matrice Jacobiana. Poniamo  $\hat{J}_N(e^{j\omega}) := J_{\hat{\theta}_N}(e^{j\omega})$ ; allora

$$\lim_{N \rightarrow \infty} \sqrt{N} [W_{\hat{\theta}_N}(e^{j\omega}) - W_{\theta_0}(e^{j\omega})] = \mathcal{N}(0, J_{\theta_0}(e^{j\omega}) P J_{\theta_0}(e^{j\omega})^\top) \quad (6.6.2)$$

in legge, qualunque sia  $\omega \in [-\pi, \pi]$ . La matrice  $P$  è la varianza asintotica di  $\hat{\theta}_N$  definita in (6.3.11).

**Example 6.3.** Si consideri la funzione di trasferimento polinomiale di tipo FIR

$$G_\theta(e^{j\omega}) = \sum_{k=1}^p \theta_k e^{-j\omega k}. \quad (6.6.3)$$

di ordine  $p$  noto. Si vuole stimare la funzione di trasferimento partendo da osservazioni rumorose dell'uscita

$$\mathbf{y}(t) = \varphi(t)^\top \theta + \mathbf{e}(t) \quad \varphi(t) = \begin{bmatrix} \mathbf{u}(t-1) \\ \vdots \\ \mathbf{u}(t-p) \end{bmatrix}, \quad t = 1, 2, \dots, N$$

dove  $\theta = [\theta_1 \ \dots \ \theta_p]^\top$  ed  $\mathbf{e}(t)$  è un processo bianco che per ogni  $t$  è scorrelato da  $\mathbf{u}^{t-1}$  (e quindi anche da  $\varphi(t)$ ), di varianza  $\lambda^2$ . Si identifica il modello usando un metodo PEM.

Discutere la consistenza dello stimatore  $\hat{G}_N(e^{j\omega}) := G_{\hat{\theta}_N}(e^{j\omega})$  e dare un'espressione per la sua varianza asintotica in funzione della frequenza.

Discutere in particolare il caso in cui  $\mathbf{u}$  è rumore bianco.

### Distribuzione asintotica di stimatori non consistenti

Consideriamo il seguente problema.

Si vuole identificare il sistema vero con rumore  $\mathbf{e}_0$  i.i.d.

$$\mathbf{y}(t) = b_0 \mathbf{u}(t-1) + \mathbf{e}_0(t), \quad \text{var} \{ \mathbf{e}_0 \} = \lambda_0^2 \quad (6.6.4)$$



sottoposto ad un ingresso descritto dal modello

$$\mathbf{u}(t) = \rho \mathbf{u}(t-1) + \mathbf{w}(t)$$

in cui  $|\rho| < 1$  e  $\mathbf{w}$  è un rumore i.i.d. di varianza  $\sigma^2$  indipendente da  $\mathbf{e}_0$ , usando una classe di modelli del tipo

$$\mathbf{y}(t) = b\mathbf{u}(t-2) + \mathbf{e}(t). \quad (6.6.5)$$

in cui si assume che  $\mathbf{e}(t)$  sia indipendente dalla storia passata  $\mathbf{u}^{t-1}$ . Questa condizione è compatibile con la presenza di una retroazione causale da  $\mathbf{y}$  ad  $\mathbf{u}$  ed è comunque sufficiente perchè il predittore di Wiener di  $\mathbf{y}(t)$  sia  $b\mathbf{u}(t-2)$ .

Si chiede di calcolare la distribuzione asintotica dello stimatore PEM  $\hat{b}_N$ .

Ci sono due possibili strade per risolvere questo problema.

*Primo approccio.* Usando il modello per  $\mathbf{u}$ , si può riscrivere il modello vero come

$$\mathbf{y}(t) = \rho b_0 \mathbf{u}(t-2) + b_0 \mathbf{w}(t-1) + \mathbf{e}_0(t) \quad (6.6.6)$$

in cui, per l'indipendenza dei segnali in gioco, il processo:

$$\tilde{\mathbf{w}}(t) := b_0 \mathbf{w}(t-1) + \mathbf{e}_0(t) \quad (6.6.7)$$

è ancora rumore bianco i.i.d. di varianza

$$\tilde{\lambda}_0^2 = b_0^2 \sigma_w^2 + \lambda_0^2.$$

Da notare che per la stabilità asintotica del modello dell'ingresso, si ha  $\mathbf{u}(t) \in \mathbf{H}_t(\mathbf{w})$  e quindi  $\mathbf{w}(t)$  è indipendente dalla storia passata  $\mathbf{u}^{t-1}$ . Quindi anche  $\tilde{\mathbf{w}}(t)$  è indipendente dalla storia passata  $\mathbf{u}^{t-1}$  e il modello vero appartiene alla classe dei modelli (6.6.5). Ne segue che lo stimatore PEM è consistente e asintoticamente normale.

La sua varianza asintotica si calcola con la solita formula (8.3.11). Il gradiente dell'errore di predizione si calcola immediatamente

$$\psi_0(t) = \mathbf{u}(t-2)$$

e la varianza asintotica ha l'espressione

$$\text{var} \{ \hat{b}_N \} \rightarrow \frac{\tilde{\lambda}_0^2}{N \sigma_u^2} = \frac{1 - \rho^2}{N} \left[ 1 + \frac{\lambda_0^2}{\sigma_w^2} \right]. \quad (6.6.8)$$

Come si vede al tendere di  $\rho$  a 1 (quindi all'aumentare della banda, ovvero all'aumentare della potenza statistica,  $\sigma_u^2$ , dell'ingresso) la varianza asintotica della stima tende a zero.

*Secondo approccio.* Senza tener conto della dinamica dell'ingresso, si può a priori pensare che il modello vero non appartenga necessariamente alla classe dei modelli (6.6.5). Si procede allora al calcolo della varianza dell'errore di predizione, il quale risulta avere l'espressione

$$\varepsilon_\theta(t) = (\rho b_0 - b)\mathbf{u}(t-2) + b_0 \mathbf{w}(t-1) + \mathbf{e}_0(t)$$

e, per l'indipendenza dei segnali in gioco, si trova

$$\text{var} [\varepsilon_\theta(t)] = (\rho b_0 - b)^2 \sigma_{\mathbf{u}}^2 + b_0^2 \sigma_{\mathbf{w}}^2 + \lambda_0^2$$

che è minima per  $b = \rho b_0$ , per cui il limite a cui tende lo stimatore PEM, è

$$\lim_{N \rightarrow \infty} \hat{b}_N = \rho b_0$$

(che è diverso da  $b_0$ ) e non si ha consistenza.

Ora, il predittore di Wiener basato sul modello (6.6.5) è una funzione lineare di  $b$ , per cui lo stimatore PEM è lo stimatore ai minimi quadrati

$$\hat{b}_N = \left[ \frac{1}{N} \sum_{t=1}^N \mathbf{u}(t-2)^2 \right]^{-1} \frac{1}{N} \sum_{t=1}^N \mathbf{u}(t-2) \mathbf{y}(t)$$

e sostituendo in questa espressione il modello vero (6.6.6) si trova

$$\hat{b}_N = \rho b_0 + \left[ \frac{1}{N} \sum_{t=1}^N \mathbf{u}(t-2)^2 \right]^{-1} \frac{1}{N} \sum_{t=1}^N \mathbf{u}(t-2) (b_0 \mathbf{w}(t-1) + \mathbf{e}_0(t)).$$

In questa espressione compare il termine  $\tilde{\mathbf{w}}(t) = b_0 \mathbf{w}(t-1) + \mathbf{e}_0(t)$  che è rumore bianco i.i.d. di varianza  $\tilde{\lambda}_0^2$  e per le ipotesi di indipendenza fatte, si vede che  $\mathbf{u}(t-2) \tilde{\mathbf{w}}(t)$  è una d-martingala di varianza

$$\sigma_0^2 = \mathbb{E} \{ \mathbf{u}(t-2)^2 \tilde{\mathbf{w}}(t)^2 \} = \mathbb{E} \{ \mathbf{u}(t-2)^2 \} \mathbb{E} \{ \tilde{\mathbf{w}}(t)^2 \} = \sigma_{\mathbf{u}}^2 \tilde{\lambda}_0^2 = \frac{\sigma_{\mathbf{w}}^2}{1 - \rho^2} \tilde{\lambda}_0^2.$$

Per il teorema del limite centrale si ha ( $L$  lim è il limite in legge),

$$L \lim \sqrt{N} \frac{1}{N} \sum_{t=1}^N \mathbf{u}(t-2) \tilde{\mathbf{w}}(t) = \mathcal{N}(0, \sigma_0^2)$$

e applicando il teorema di Slutsky si può trovare la distribuzione asintotica di  $\hat{b}_N$ .

$$\begin{aligned} L \lim_{N \rightarrow \infty} \frac{1}{\sqrt{N}} (\hat{b}_N - \rho b_0) &= L \lim_{N \rightarrow \infty} \frac{1}{\sigma_{\mathbf{u}}^2} \frac{1}{\sqrt{N}} \sum_{t=1}^N \mathbf{u}(t-2) \tilde{\mathbf{w}}(t) \\ &= \mathcal{N}(0, \frac{\sigma_0^2}{\sigma_{\mathbf{u}}^2}) = \mathcal{N}(0, \frac{\tilde{\lambda}_0^2}{\sigma_{\mathbf{u}}^2}) \end{aligned} \quad (6.6.9)$$

per cui la varianza asintotica dello stimatore è

$$\text{var} \{ \hat{b}_N \} \sim \frac{1}{N} \frac{\tilde{\lambda}_0^2}{\sigma_{\mathbf{u}}^2} = \frac{1 - \rho^2}{N} \left[ 1 + \frac{\lambda_0^2}{\sigma_{\mathbf{w}}^2} \right]$$

che è la stessa formula trovata col primo approccio.

## Formula di Whittle

per la verosimiglianza e per il limite di C-R. (DA SCRIVERE)

## 6.7 Aspetti computazionali

Trattiamo solo problemi in cui  $N$  è grande, interessa il comportamento asintotico e si può quindi approssimare il predittore con quello di Wiener. Altrimenti si massimizza numericamente la verosimiglianza esatta (iterazione sul filtro di Kalman).

### Algoritmi di Quasi-Newton per la minimizzazione locale

La minimizzazione dell'errore quadratico medio di predizione (6.1.4) si può fare esplicitamente (ed esattamente) solo per modelli a predittore lineare nei parametri, tipicamente per modelli ARX. In questo caso si tratta di risolvere un problema ai minimi quadrati per il quale esistono routines stabili e ben collaudate. Ricordiamo che la formula esplicita per lo stimatore (6.1.13) ha interesse prevalentemente teorico e non dovrebbe mai essere usata per calcolare  $\hat{\theta}_N$ .

Se si usano modelli di tipo generale e quindi modelli per i quali il predittore ha memoria infinita (non ha la struttura di un filtro FIR, come per i modelli ARX) alcuni parametri del modello determinano i modi del sistema e compaiono implicitamente in modo non lineare nella risposta impulsiva del predittore. In questi casi il predittore è funzione non lineare dei parametri del modello e la minimizzazione del funzionale (6.1.4) si può fare solo per via numerica mediante algoritmi iterativi di ottimizzazione. In questo paragrafo illustreremo una classe di algoritmi che sono particolarmente adatti alla specifica struttura (quadratica) dell'errore quadratico medio di predizione. Il capostipite di questi algoritmi è l'*algoritmo di Newton* (o Newton-Rapson) che descriveremo qui sotto.

Sia  $V(\theta) : \mathbb{R}^p \rightarrow \mathbb{R}$  una funzione di classe  $C^2$ . Per minimizzare  $V(\theta)$  rispetto a  $\theta$  possiamo pensare di approssimare la funzione localmente attorno ad un punto  $\bar{\theta}$  mediante una funzione quadratica,

$$V(\theta) \simeq V(\bar{\theta}) + (\theta - \bar{\theta})^\top V'(\bar{\theta}) + \frac{1}{2}(\theta - \bar{\theta})^\top V''(\bar{\theta})(\theta - \bar{\theta}) \quad (6.7.1)$$

dove

$$V'(\bar{\theta}) := \frac{\partial}{\partial \theta} V(\theta) \Big|_{\theta=\bar{\theta}}, \quad V''(\bar{\theta}) = \left[ \frac{\partial^2}{\partial \theta_i \partial \theta_j} V(\theta) \right] \Big|_{\theta=\bar{\theta}}$$

sono il gradiente (che scriviamo come un vettore colonna) e la matrice Hessiana di  $V(\theta)$  calcolati in  $\bar{\theta}$ . La minimizzazione della funzione quadratica al secondo membro di (6.7.1) conduce all'equazione

$$V''(\bar{\theta})(\theta - \bar{\theta}) = -V'(\bar{\theta})$$

le quale, nell'ipotesi che  $V''(\bar{\theta})$  sia non singolare, fornisce l'espressione del punto di minimo "approssimato",

$$\hat{\theta} = \bar{\theta} - [V''(\bar{\theta})]^{-1} V'(\bar{\theta}). \quad (6.7.2)$$

L'idea base dell'algoritmo è di usare questa equazione in modo iterativo prendendo  $\bar{\theta} = \theta_k$  e  $\theta_{k+1} = \hat{\theta}$  e iterando su  $k$ . Come si vede il punto di minimo "approssimato",  $\hat{\theta}$  si calcola partendo da  $\bar{\theta}$  e muovendosi nella direzione del gradiente negativo (quindi una direzione di discesa) "filtrata" attraverso l'inversa della matrice Hessiana. Se  $\bar{\theta}$  è vicino ad un punto di minimo si può supporre che per continuità  $V''(\bar{\theta})$  sia definita positiva e quindi che la direzione del gradiente "filtrata" (che viene chiamata *direzione di Newton*), sia ancora una direzione di discesa. Se però si è lontani da un minimo  $V''(\bar{\theta})$  potrebbe essere indefinita oppure addirittura definita negativa e la direzione di Newton potrebbe puntare verso un punto di crescita della funzione anzichè un punto di decrescita. Come si vede questo procedimento, benchè porti al minimo di una funzione quadratica in una sola iterazione, ha dei grossi difetti se applicato a funzioni non quadratiche.

Per rimediare a questi difetti si introducono delle approssimazioni della matrice Hessiana che abbiano la proprietà di essere sempre almeno semidefinite positive. Queste approssimazioni danno origine ad una famiglia di algoritmi che si chiamano di *Quasi-Newton*. Ne illustreremo uno particolarmente adatto alla nostra funzione costo.

Usando le espressioni per le derivate (6.3.3),(6.3.4) e pensando di essere in una situazione limite ottimistica in cui  $\theta_k$  è vicino al valore vero  $\theta_0$ , si può pensare di trascurare il termine con le derivate seconde di  $\varepsilon_\theta(t)$  nell'espressione della matrice Hessiana, ricavandone così un'approssimazione

$$H_\theta(N) := \frac{1}{N} \sum_{t=1}^N \psi_\theta(t) \psi_\theta(t)^\top \simeq \frac{1}{2} V_N(\theta)'' \quad (6.7.3)$$

che è sempre almeno semidefinita positiva (e non richiede il calcolo delle derivate seconde di  $\varepsilon_\theta(t)$ ). Inoltre, se si è nelle condizioni in cui vale il teorema del limite centrale 6.4, l'espressione  $H_\theta(N)$  converge con probabilità uno per  $N \rightarrow \infty$  a una matrice invertibile (in effetti a  $\lambda_0^2$  volte la matrice di Fisher).

Si perviene così ad un classe di algoritmi che ha la seguente struttura.

**Algorithm 6.1.**

Data la stringa dei dati di ingresso  $[\mathbf{y}^N, \mathbf{u}^N] = [\mathbf{y}(1) \dots \mathbf{y}(N), \mathbf{u}(1) \dots \mathbf{u}(N)]$ , la classe di modelli (6.1.3), un valore iniziale  $\theta_0$  e la stima  $\theta_k$  alla  $k$ -sima iterazione,

1. Si calcola la stringa degli errori di predizione  $\varepsilon_{\theta_k} = [\varepsilon_{\theta_k}(1) \dots \varepsilon_{\theta_k}(N)]^\top$  risolvendo l'equazione

$$\varepsilon_{\theta_k}(t) = G_{\theta_k}(z)^{-1} [\mathbf{y}(t) - F_{\theta_k}(z) \mathbf{u}(t)] \quad (6.7.4)$$

prendendo condizioni iniziali arbitrarie (ad esempio nulle).

2. Si calcola la stringa dei gradienti  $\Psi_{\theta_k} = [\psi_{\theta_k}(1) \dots \psi_{\theta_k}(N)] \in \mathbb{R}^{p \times N}$ , dell'errore di predizione del modello  $M(\theta_k)$ .
3. Si calcola la matrice pseudo-Hessiana

$$H_{\theta_k} := \sum_{t=1}^N \psi_{\theta_k}(t) \psi_{\theta_k}(t)^\top = \Psi_{\theta_k} \Psi_{\theta_k}^\top$$

e la sua inversa  $P_{\theta_k} := H_{\theta_k}^{-1}$ .

4. Si aggiorna  $\theta_k$  mediante la,

$$\theta_{k+1} - \theta_k = -H_{\theta_k}^{-1} \Psi_{\theta_k} \varepsilon_{\theta_k} \quad (6.7.5)$$

5. Si torna al passo 1) ponendo  $\theta_k = \theta_{k+1}$ .

Questo schema di principio richiede alcuni commenti.

1. Dato che il calcolo di  $\varepsilon_{\theta_k}$  e del gradiente  $\psi_{\theta_k}$  sono basati su predittori di Wiener in cui le condizioni iniziali (incognite) vengono scelte in modo arbitrario, i dati iniziali vengono "maltrattati" dall'algoritmo e per questo motivo sarebbe opportuno minimizzare un funzionale di costo scontato del tipo (6.1.5).
2. L'approssimazione della matrice Hessiana può risultare singolare o mal condizionata specialmente nelle iterazioni iniziali. Con molti parametri il funzionale costo potrebbe comunque risultare poco sensibile alla variazione di qualche parametro e anche l'Hessiano esatto potrebbe risultare mal condizionato. Per rimediare a questo mal condizionamento si può usare una tecnica di *regolarizzazione* e definire l'inversa "regolarizzata" come

$$P_{\theta_k} := [H_{\theta_k} + \delta(\theta_k) I]^{-1} \quad (6.7.6)$$

dove  $\delta(\theta)$  è una opportuna funzione scalare positiva, ad esempio una costante o un termine proporzionale al quadrato di una norma pesata di  $\theta$  del tipo  $\theta^T Q \theta$  con  $Q = Q^T > 0$ . Questo termine si può interpretare in un contesto Bayesiano come una varianza a priori del parametro  $\theta$  [3, 68]. Per quanto visto nel capitolo 2 questo termine introduce un errore sistematico nella stima e deve essere scelto opportunamente piccolo.

3. Il calcolo dell'errore di predizione richiede che il polinomio numeratore della funzione di trasferimento  $G_{\theta_k}(z)$ ,  $C_{\theta_k}(z^{-1})$ , supponiamolo di grado  $q$ , e quello a denominatore  $A_{\theta_k}(z^{-1})$  (di grado  $n$ ) siano *polinomi stabili* nel senso che  $z^q C_{\theta_k}(z^{-1}) = 0 \Rightarrow |z| < 1$  e  $z^n A_{\theta_k}(z^{-1}) = 0 \Rightarrow |z| < 1$  (a meno di improbabili cancellazioni con il denominatore  $D_{\theta_k}(z^{-1})$  di  $G_{\theta_k}$ ). La stessa cosa accade per il calcolo del gradiente.

Questi vincoli di stabilità sono stati sistematicamente trascurati nella minimizzazione dell'errore di predizione, che avrebbe a rigore dovuto essere una minimizzazione *vincolata* dalle condizioni di stabilità appunto, di  $C_{\theta_k}(z^{-1})$  e  $A_{\theta_k}(z^{-1})$ .

4. Il calcolo del gradiente si può fare riferendosi alla parametrizzazione standard del modello di Box-Jenkins descritta dai polinomi (3.4.1). Le quattro

componenti in cui viene naturalmente partizionato il gradiente,

$$\begin{aligned} \psi_a(t) &:= \left[ \frac{\partial \varepsilon_\theta(t)}{\partial a_i} \right]_{i=1, \dots, n} ; & \psi_b(t) &:= \left[ \frac{\partial \varepsilon_\theta(t)}{\partial b_i} \right]_{i=1, \dots, m} ; \\ \psi_c(t) &:= \left[ \frac{\partial \varepsilon_\theta(t)}{\partial c_i} \right]_{i=1, \dots, q} ; & \psi_d(t) &:= \left[ \frac{\partial \varepsilon_\theta(t)}{\partial d_i} \right]_{i=1, \dots, r} ; \end{aligned}$$

si possono calcolare derivando membro a membro rispetto ai parametri il modello (6.1.3) riscritto nella forma

$$C_{theta}(z^{-1})A_\theta(z^{-1})\varepsilon_\theta(t) = A_\theta(z^{-1})D_\theta(z^{-1})\mathbf{y}(t) - B_\theta(z^{-1})D_\theta(z^{-1})\mathbf{u}(t). \quad (6.7.7)$$

Calcolando le derivate rispetto alle prime componenti  $a_1, b_1, c_1, d_1$ , si trovano le equazioni alle differenze

$$\begin{aligned} C_\theta(z^{-1})A_\theta(z^{-1})\psi_{a_1}(t) &:= -C_\theta(z^{-1})\varepsilon_\theta(t-1) + D_\theta(z^{-1})\mathbf{y}(t-1); \\ C_\theta(z^{-1})A_\theta(z^{-1})\psi_{b_1}(t) &:= -D_\theta(z^{-1})\mathbf{u}(t-1); \\ C_\theta(z^{-1})A_\theta(z^{-1})\psi_{c_1}(t) &:= -A_\theta(z^{-1})\varepsilon_\theta(t-1); \\ C_\theta(z^{-1})A_\theta(z^{-1})\psi_{d_1}(t) &:= A_\theta(z^{-1})\mathbf{y}(t-1) - B_\theta(z^{-1})\mathbf{u}(t-1); \end{aligned}$$

che vengono di norma associate a condizioni iniziali nulle. In questo caso le componenti di indice maggiore di uno si possono ottenere semplicemente per traslazione; ad esempio

$$\begin{aligned} \psi_{a_k}(t) &= \psi_{a_1}(t-k), \quad k = 1, 2, \dots, n, & \psi_{b_k}(t) &= \psi_{b_1}(t-k), \quad k = 1, 2, \dots, m, \\ \psi_{c_k}(t) &= \psi_{c_1}(t-k), \quad k = 1, 2, \dots, q, & \psi_{d_k}(t) &= \psi_{d_1}(t-k), \quad k = 1, 2, \dots, r. \end{aligned}$$

Da notare che la stabilità di  $C_\theta(z^{-1})A_\theta(z^{-1})$  è essenziale per poter portare avanti i calcoli.

5. L'algoritmo (6.7.5) richiede il calcolo dell'inversa  $H_{\theta_k}^{-1}$  ad ogni iterazione. Questo calcolo si può organizzare in forma ricorsiva (in  $t$ ). Ricordando che  $P_{\theta_k}(t) := H_{\theta_k}^{-1}(t)$ , e

$$H_{\theta_k}(t+1) := \sum_{s=1}^t \psi_{\theta_k}(s)\psi_{\theta_k}(s)^\top + \psi_{\theta_k}(t+1)\psi_{\theta_k}(t+1)^\top$$

usando il lemma di inversione di matrice si trova la relazione ricorsiva in  $t$ ,

$$P_{\theta_k}(t+1) = P_{\theta_k}(t) - P_{\theta_k}(t)\psi_{\theta_k}(t+1) \frac{1}{1 + \psi_{\theta_k}(t+1)^\top P_{\theta_k}(t)\psi_{\theta_k}(t+1)} \psi_{\theta_k}(t+1)^\top P_{\theta_k}(t)$$

che è una ricorsione "alla Riccati". La ricorsione è valida anche quando è presente il termine di regolarizzazione che si può interpretare come una varianza (normalizzata) a priori assegnata come condizione iniziale all'istante  $t = 0$ . Naturalmente in pratica conviene sempre usare una versione simmetrizzata di questa equazione [44].

6. Notiamo che l'algoritmo di Newton si può interpretare come un metodo iterativo per risolvere le equazioni normali di un problema ai minimi quadrati non lineare. La formula (6.7.5) si può infatti interpretare come la risoluzione iterativa di una successione di problemi di stima ai minimi quadrati su dei modelli lineari "incrementali", del tipo (6.5.3), ciascuno definito dalle equazioni

$$\varepsilon_{\theta_k}(t) = \mathbf{y}(t) - \hat{\mathbf{y}}_{\theta_k}(t | t-1) = \boldsymbol{\psi}_{\theta_k}(t) (\boldsymbol{\theta} - \boldsymbol{\theta}_k) + \mathbf{e}_0(t), \quad t = 1, 2, \dots, N$$

ovvero, in notazione vettoriale

$$\varepsilon_{\theta_k} = \boldsymbol{\Psi}_{\theta_k} (\boldsymbol{\theta} - \boldsymbol{\theta}_k) + \mathbf{e}_0. \quad (6.7.8)$$

Questi problemi possono essere risolti *senza formare le equazioni normali* con tecniche del tipo fattorizzazione QR, come visto nella sezione 2.10 del capitolo 2.

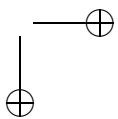
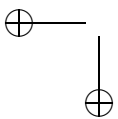
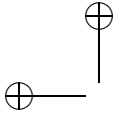
7. Rimane comunque il fatto che qualunque algoritmo di minimizzazione può unicamente portare nella prossimità di *minimi relativi* che potrebbero in linea di principio essere alquanto lontani dal minimo assoluto. Per questo motivo l'inizializzazione; i.e. la scelta di  $\boldsymbol{\theta}_0$ , è spesso un passo importante e può essere consigliabile far girare l'algoritmo partendo da diverse stime iniziali.

## 6.8 Algoritmi ricorsivi

Motivazione: Algoritmi adattativi. Filtro di Kalman condizionato per modelli ARX. Regressione lineare ricorsiva. L'equazione di Riccati e la varianza normalizzata. Comportamento asintotico. Fattore d'oblio. Modelli ARMAX: Algoritmi di Newton ricorsivi approssimati: RPEM, PLR, etc. Filtro di Kalman esteso. Problemi di convergenza.

### Modelli a predittore lineare nei parametri

Equazioni di Yule-Walker. Stabilità e legame con l'algoritmo di Levinson.





## CHAPTER 7

## VERIFICA DI IPOTESI

Sia  $y$  un vettore aleatorio di dimensione  $n$  e  $y \sim F_\theta$ , con  $\theta$  parametro incognito in  $\Theta \subset \mathbb{R}^p$ . Siano date  $M + 1$  regioni disgiunte  $\Theta_0, \Theta_1, \dots, \Theta_M$  di  $\Theta$ , non necessariamente esaustive per  $\Theta$ . Chiameremo  $H_k$  la classe  $\{F_\theta, \theta \in \Theta_k\}$ ,  $k = 0, 1, \dots, M$ . Il problema della verifica delle  $M + 1$  ipotesi

$$H_0, H_1, \dots, H_M$$

è quello di decidere in base ai dati osservati,  $y$ , se la distribuzione  $F_\theta$  (incognita) di  $y$  appartiene o no ad una delle classi  $\{H_k\}$ . Si tratta cioè di trovare una funzione di decisione

$$\phi : \mathbb{R}^n \rightarrow \{0, 1, \dots, M\}$$

che assegni ad ogni possibile risultato  $y$  dell'osservazione  $y$  una ed una sola delle classi  $H_k$ . Si decide in sostanza che una delle  $H_k$  è "vera" in base all'osservazione  $y = y$ .

Naturalmente questa decisione si fa con certi criteri ottimali e inoltre, come sempre in statistica, la decisione è essa stessa incerta, ovvero l'assegnazione  $y \mapsto H_k$  avviene sempre con una certa probabilità d'errore.

Anche per la verifica d'ipotesi si possono seguire due approcci, quello Fisheriano e quello Bayesiano. Ripetiamo che è inutile discutere su quale sia quello "giusto". I due approcci si riferiscono a situazioni in cui l'informazione a priori su  $\theta$  è radicalmente diversa. Nel caso fosse assegnata una distribuzione di probabilità "a priori" sulle  $M + 1$  a classi, sarebbe sbagliato non tenerne conto. In queste note seguiremo l'approccio Fisheriano in cui *nulla si sa a priori sull'appartenenza di  $F_\theta$  alle classi  $\{H_k\}$* .

**Definition 7.1.** *Un'ipotesi  $H_k$  si dice semplice (o composta) se  $\Theta_k$  contiene un solo punto di  $\Theta$  (oppure se ne contiene più d'uno).*

Ad esempio, se  $y$  è il vettore dei risultati dell'osservazione successiva di  $M$  lanci indipendenti di una moneta ( $y = 0$  se esce croce,  $y = 1$  se esce testa),  $\theta = p$ ,

$p \in (0, 1)$ ; l'ipotesi  $H_0$  corrispondente a  $p = \frac{1}{2}$  è *semplice*, mentre l'ipotesi  $H_1 : \{p; p > \frac{1}{2}\}$  è *composta*.

Supporremo, d'ora in avanti, di avere solo *due* ipotesi,  $H_0$  e  $H_1$ .

**Definition 7.2.** Si definisce *test dell'ipotesi  $H_0$  "contro" l'ipotesi alternativa  $H_1$  una statistica*

$$\phi : \mathbb{R}^n \rightarrow \{0, 1\}.$$

Diremo anche che la  $\phi$  definisce il test.

La teoria è formulata in modo tale che  $H_0$  (ipotesi nulla) ha un ruolo "privilegiato". Se  $\phi(y) = 0$  si dice che si *accetta*  $H_0$ , se  $\phi(y) = 1$  si dice che si *rifiuta*  $H_0$ . Naturalmente queste decisioni si prendono con una certa probabilità d'errore e la teoria si occupa di trovare statistiche per cui questa probabilità d'errore sia la più piccola possibile.

Ovviamente, assegnare una statistica del test, significa dire chi è la regione di  $\mathbb{R}^n$  in cui essa vale 0, oppure 1. È tradizione definire  $\phi$  dando la regione,  $\mathcal{C} \subseteq \mathbb{R}^n$  in cui  $\phi(y) = 1$ :

**Definition 7.3.** Si chiama *regione critica  $\mathcal{C}$  del test  $\phi$  la regione di rifiuto di  $H_0$  cioè*

$$\mathcal{C} = \{y; \phi(y) = 1\}. \quad (7.0.1)$$

Ne segue che un test è assegnato dando la sua regione critica  $\mathcal{C}$ . In genere  $\mathcal{C}$  è un sottoinsieme dello spazio campionario definito da un sistema di disuguaglianze del tipo  $\{y; \psi_k(y) \leq c_k, k = 1, \dots, m\}$  dove  $\psi$  è una certa statistica, in generale a valori vettoriali in  $\mathbb{R}^m$ . Noi penseremo ogni statistica  $\psi$  la cui regione critica è la stessa di  $\phi$ , definire *lo stesso test*.

Nel decidere se accettare o no  $H_0$  si possono ovviamente commettere degli errori. Il caso ideale sarebbe trovare una statistica  $\phi$  che *discrimina esattamente* le due famiglie di probabilità  $H_0$  e  $H_1$ , cioè una  $\phi$  per cui

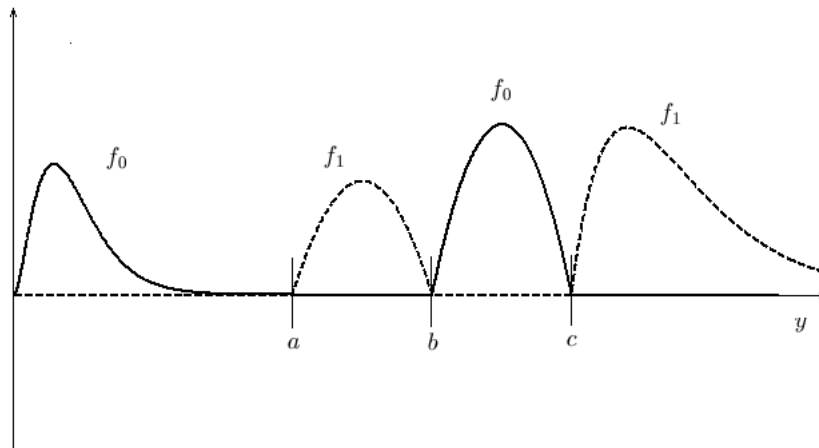
$$\begin{cases} \phi(\mathbf{y}) = 0 & \text{se } \mathbf{y} \sim H_0 \\ \phi(\mathbf{y}) = 1 & \text{se } \mathbf{y} \sim H_1 \end{cases}, \quad (7.0.2)$$

con ovvio significato delle notazioni. Questo accade solo in situazioni degeneri. Supponiamo per esempio che  $H_0$  e  $H_1$  siano entrambi *semplici* e che  $F_{\theta_0}$  e  $F_{\theta_1}$  ammettano densità  $f_0(y)$  e  $f_1(y)$ .

**Lemma 7.1.** Si ha *discriminazione perfetta tra  $f_0$  e  $f_1$  se e solo se  $f_0$  e  $f_1$  sono ortogonali; i.e.*

$$\int_{\mathbb{R}^n} f_0(x)f_1(x) dx = 0.$$

In particolare  $f_1$  è *strettamente positiva* negli insiemi in cui  $f_0$  si annulla e, viceversa, dove  $f_1$  si annulla  $f_0$  è *strettamente positiva*.



Sia infatti  $\mathcal{C}$  la regione in cui  $f_1(y) > 0$

$$\mathcal{C} = \{y; f_1(y) > 0\}.$$

Chiaramente, se  $y \sim f_0$  allora  $y$  appartiene a  $\mathcal{C}$  con probabilità zero, cioè

$$\mathbb{P}_0(y \in \mathcal{C}) = 0$$

$$\mathbb{P}_1(y \in \mathcal{C}) = 1$$

Quindi basta prendere  $\mathcal{C}$  come regione critica e si ha un test perfetto.

In generale si ha invece la situazione descritta dalla seguente tabella

|                     | Ipotesi vera |       |
|---------------------|--------------|-------|
|                     | $H_0$        | $H_1$ |
| Decisione per $H_0$ | O.K.         | II    |
| Decisione per $H_1$ | I            | O.K.  |

(7.0.3)

- Se è vera  $H_0$  ma  $y \in \mathcal{C}$  e quindi  $(\phi(y) = 1)$  si decide di rifiutarla si ha un cosiddetto errore di *prima specie*.
- Se  $H_0$  è falsa (cioè  $y \sim H_1$ ) ma accade che  $y$  appartenga al complementare  $\bar{\mathcal{C}}$  di  $\mathcal{C}$  e quindi si decide che vale  $H_0$ , si ha un errore di *seconda specie*.

Il calcolo delle probabilità d'errore è essenziale per valutare il comportamento di un test. Questo calcolo si presenta particolarmente facile se  $H_0$  e  $H_1$  sono ipotesi semplici.

## 7.1 Ipotesi semplici

Siano  $H_0 = \{F_0\}$  ed  $H_1 = \{F_1\}$  due ipotesi *semplici*. Allora la probabilità,  $\alpha$ , di commettere un errore di prima specie e quella,  $\beta$ , di commettere un errore di

secondo specie, sono

$$\alpha = \int_{\mathcal{C}} dF_0(y) = \mathbb{P}_0(\mathcal{C}) \tag{7.1.1}$$

$$\beta = \int_{\mathcal{C}^c} dF_1(y) = \mathbb{P}_1(\mathbb{R}^n \setminus \mathcal{C}) = 1 - \mathbb{P}_1(\mathcal{C}) \tag{7.1.2}$$

La probabilità

$$1 - \beta = \mathbb{P}_1(\mathcal{C}) = \mathbb{P}(y \in \mathcal{C}) \tag{7.1.3}$$

che, quando vale  $H_1$ , i valori osservati cadano nella regione critica (cioè la probabilità di rifiutare  $H_0$  quando  $H_0$  è falsa) si chiama *potenza* (o potere discriminante) del test.

Notiamo subito che se  $H_1$  non è semplice ovvero  $\Theta_1$  contiene più di un valore del parametro, la *potenza è una funzione di  $\theta$* .

La terminologia che si usa nelle comunicazioni elettriche è leggermente diversa. Si fa riferimento ad un problema di radar in cui  $H_0$  e  $H_1$  rappresentano rispettivamente *assenza e presenza di bersaglio*. Si definiscono

$\mathbb{P}_F$ : probabilità di “falso allarme” (dire che c’è il bersaglio quando non c’è);

$\mathbb{P}_M$ : probabilità di “perdere il bersaglio” (dire che non c’è, cioè accettare  $H_0$ , quando invece il bersaglio c’è;  $M$  sta per “miss”, perdita);

$\mathbb{P}_D$ : probabilità di discriminazione (dire che c’è il bersaglio quando in effetti è presente).

Ovviamente

$$\mathbb{P}_F = \alpha, \quad \mathbb{P}_M = \beta, \quad \mathbb{P}_D = 1 - \beta$$

cioè la probabilità di discriminazione coincide con la potenza.

Un test “ottimo” per discriminare tra  $H_0$  e  $H_1$  sarebbe evidentemente quello per cui  $\alpha$  e  $\beta$  sono i più piccoli possibile. In realtà si tratta di due obiettivi contrastanti. Siano ad esempio  $H_0 \sim \mathcal{N}(\mu_0, \sigma^2)$  ed  $H_1 \sim \mathcal{N}(\mu_1, \sigma^2)$  con  $\mu_0 < \mu_1$ , due valori assegnati. Supponiamo  $n = 1$  e di voler costruire una regione critica  $\mathcal{C}$  della forma

$$\mathcal{C} := \{y; y \geq c\}.$$

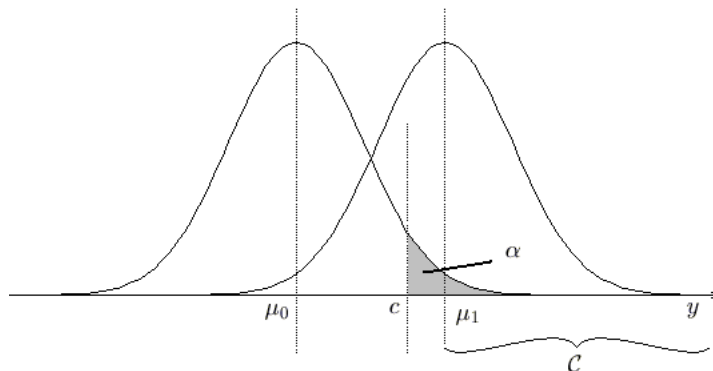
Dalla figura si vede che si può scegliere  $c$  in modo da avere  $\alpha$  piccolo quando si vuole,

$$\alpha = \frac{1}{\sqrt{2\pi}\sigma} \int_c^{+\infty} e^{-\frac{1}{2} \frac{(y-\mu_0)^2}{\sigma^2}} dy.$$

Più grande si prende  $c$ , però, più aumenta la probabilità  $\beta$  di classificare incorrettamente  $H_1$  come  $H_0$ . Infatti

$$\beta = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^c e^{-\frac{1}{2} \frac{(y-\mu_1)^2}{\sigma^2}} dy$$

cresce all’aumentare di  $c$ .



Il procedimento classico è allora quello di fissare  $\alpha$  e di cercare la regione critica  $C$  che dà il  $\beta$  più piccolo possibile (ovvero il massimo valore della potenza  $1 - \beta$ ). Una siffatta regione critica si chiama “la migliore regione critica di misura  $\alpha$ ” (in inglese B.C.R. = best critical region of size  $\alpha$ ). Sussiste a questo proposito il fondamentale:

**Lemma 7.2 (di Neyman – Pearson).** La miglior regione critica di misura  $\alpha$  per verificare l’ipotesi  $H_0 = \{f_0\}$  contro l’alternativa (semplice)  $H_1 = \{f_1\}$  è l’insieme dei punti nello spazio campionario  $\mathbb{R}^n$  per cui

$$C = \{y; \Lambda(y) \geq k\} \tag{7.1.4}$$

dove

$$\Lambda(y) = \frac{f_1(y)}{f_0(y)} \tag{7.1.5}$$

e la costante  $k$  è scelta in modo tale da aversi

$$\int_C f_0(y) dy = \alpha. \tag{7.1.6}$$

Ci sono varie versioni di questo Lemma che trattano del caso in cui  $H_0$  e  $H_1$  non si possono dare mediante densità e si preoccupano dell’eventualità in cui la frontiera di  $C$  ha probabilità positiva. Rimandiamo al classico testo di Lehmann [32] per una trattazione completa.

Notiamo che per le (7.1.4) e (7.1.5),  $C$  contiene i punti  $y$  in cui  $f_0(y) = 0$ , come intuitivamente ci si aspetta.

**Proof.** Sia  $I_C(y)$  la funzione indicatrice della regione critica  $C$ . Dobbiamo massimizzare, rispetto a  $C \subseteq \mathbb{R}^n$

$$1 - \beta = \int_C f_1(y) dy = \int_{\mathbb{R}^n} I_C(y) \frac{f_1(y)}{f_0(y)} f_0(y) dy = \mathbb{E}_0[I_C \Lambda] \tag{7.1.7}$$

con il vincolo

$$\alpha = \mathbb{E}_0[I_C]. \tag{7.1.8}$$

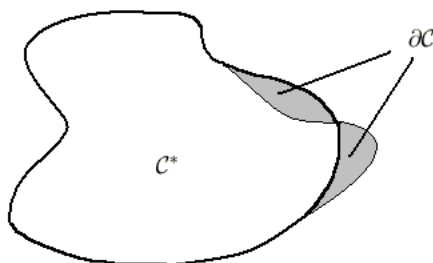
Questo è un cosiddetto problema di “frontiera libera” del calcolo delle variazioni.

Introducendo il moltiplicatore di Lagrange  $\lambda$ , si tratta di massimizzare, rispetto a  $I_C$

$$J(C) := \mathbb{E}_0[I_C \Lambda] - \lambda \{ \mathbb{E}_0[I_C] - \alpha \}.$$

Sia  $C^*$  la miglior regione critica. Se perturbiamo  $C^*$  di  $\delta C$  (infinitesima) dovremo avere  $\delta J(C^*) = 0$ , ovvero

$$\mathbb{E}_0[I_{\delta C}(\Lambda - \lambda)] = \int_{\delta C} [\Lambda(y) - \lambda] f_0(y) dy = 0, \forall \delta C,$$



da cui si vede ( $f_0(y) \geq 0$ ) che nei punti della frontiera di  $C^*$  dev'essere

$$\Lambda(y) = \lambda \quad y \in \partial C^*.$$

In particolare, il moltiplicatore  $\lambda$  dev'essere positivo visto che  $\Lambda(y) \geq 0$ . Riscrivendo  $J(C)$  come

$$J(C) = \mathbb{E}_0[I_C(\Lambda - \lambda)] + \lambda \alpha$$

si vede che  $C^*$  dà un massimo di  $J(C)$  solo se in  $C^*$  si prendono i punti per cui  $\Lambda(y) - \lambda \geq 0$ .  $\square$

**Example 7.1.** Sia  $H_0 = \{\mathcal{N}(\mu_0, \sigma^2)\}$ ,  $H_1 = \{\mathcal{N}(\mu_1, \sigma^2)\}$  con la stessa varianza  $\sigma^2$  e medie  $\mu_0 \neq \mu_1$  assegnate. Sia  $\mathbf{y}_1, \dots, \mathbf{y}_N$  un campione casuale di numerosità  $N$ . Allora

$$\Lambda(y_1, \dots, y_N) = \exp - \frac{1}{2\sigma^2} \left\{ \sum_1^N (y_i - \mu_1)^2 - \sum_1^N (y_i - \mu_0)^2 \right\}.$$

Usando la classica decomposizione  $\sum_1^N (y_t - \mu)^2 = \sum_1^N (y_t - \bar{y}_N)^2 + N(\bar{y}_N - \mu)^2$  si

ottiene

$$\begin{aligned}\Lambda(y) &= \exp -\frac{1}{2\sigma^2} \{N(\bar{y}_N - \mu_1)^2 - N(\bar{y}_N - \mu_0)^2\} \\ &= \exp -\frac{N}{2\sigma^2} [(\bar{y}_N - \mu_0)^2 + 2(\bar{y}_N - \mu_0)(\mu_0 - \mu_1) + (\mu_0 - \mu_1)^2 - (\bar{y}_N - \mu_0)^2] \\ &= \exp -\frac{N}{2\sigma^2} [2(\bar{y}_N - \mu_0)(\mu_0 - \mu_1) + (\mu_0 - \mu_1)^2].\end{aligned}$$

La disuguaglianza  $\Lambda(y) \geq k$  che definisce la regione critica si può anche riscrivere  $\log \Lambda(y) \geq \log k$  ovvero,

$$\frac{N}{2\sigma^2} [2\bar{y}_N(\mu_1 - \mu_0) + \mu_1^2 - \mu_0^2] \geq \log k$$

la quale, se  $\mu_1 > \mu_0$  è equivalente alla

$$\bar{y}_N \geq \frac{1}{2}(\mu_1 + \mu_0) + \frac{\sigma^2}{N(\mu_1 - \mu_0)} \log k \quad (7.1.9)$$

Se viceversa fosse  $\mu_0 > \mu_1$  la regione critica sarebbe definita dalla

$$\bar{y}_N \leq \frac{1}{2}(\mu_1 + \mu_0) - \frac{\sigma^2}{N(\mu_0 - \mu_1)} \log k. \quad (7.1.10)$$

Indichiamo il secondo membro di queste disuguaglianze con  $c_1$  o  $c_2$ ; la miglior regione critica è quindi definita dalle

$$\mathcal{C}_1 : \{y; \bar{y}_N \geq c_1\} \quad \text{se } \mu_0 < \mu_1 \quad (7.1.11)$$

$$\mathcal{C}_2 : \{y; \bar{y}_N \leq c_2\} \quad \text{se } \mu_1 < \mu_0. \quad (7.1.12)$$

Notiamo che siamo riusciti ad esprimere la regione critica per mezzo della statistica  $\bar{y}_N$ . Quindi, assegnato  $\alpha$ , e supponendo ad esempio  $\mu_1 > \mu_0$  si può ragionare come segue.

Sotto  $H_0$ ,  $\bar{y}_N \sim \mathcal{N}(\mu_0, \frac{\sigma^2}{N})$ . Normalizzando,  $\frac{\sqrt{N}}{\sigma}(\bar{y} - \mu_0) \sim \mathcal{N}(0, 1)$ , per cui esprimendo  $c_1$  come

$$c_1 = a_1 \frac{\sigma}{\sqrt{N}} + \mu_0 \quad (7.1.13)$$

si ha

$$\mathcal{C}_1 = \{y \mid \frac{\sqrt{N}}{\sigma}(\bar{y} - \mu_0) \geq a_1\}$$

e, fissato  $\alpha$ , si può andare sulle tavole della  $\mathcal{N}(0, 1)$  e trovare  $a_1$  in modo tale che

$$\mathbb{P}_0(\mathcal{C}_1) = \mathbb{P}_0\left(\frac{\sqrt{N}}{\sigma}(\bar{y} - \mu_0) \geq a_1\right) = \alpha.$$

Usando la (7.1.13) si ottiene  $c_1$  e quindi  $\mathcal{C}_1$ .

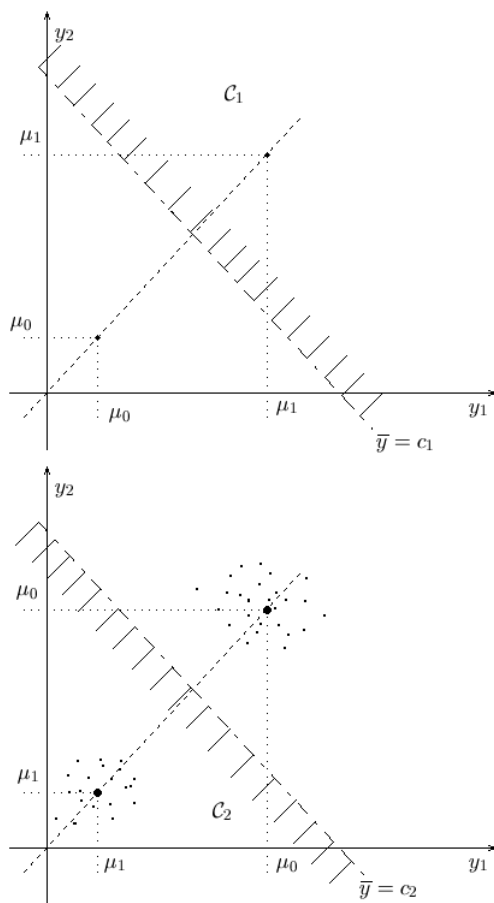


Figure 7.1.1. Regioni critiche per l'esempio 7.1

Questo esempio può essere usato per descrivere il ricevitore ottimo in un sistema di comunicazione digitale in cui la sorgente emette un segnale binario del tipo di figura 7.1.2,

Il periodo  $T$  del segnale trasmesso è esattamente  $n$  volte il periodo di campionamento  $T_c$  del ricevitore. Il ricevitore fornisce (dopo demodulazione) un segnale che è la somma di quello di figura più un rumore bianco Gaussiano  $w(t)$  di media zero e varianza  $T_c\sigma^2$  (nota). Questo segnale viene campionato con periodo di  $T_c$  secondi, in sincronismo con la sorgente, ottenendo così, ogni  $T$  secondi,  $n$  campioni descrivibili con lo schema,

$$\begin{cases} y_t = \mu_0 + w_t, & t = 1, \dots, n \text{ sotto } H_0 \\ y_t = \mu_1 + w_t, & t = 1, \dots, n \text{ sotto } H_1 \end{cases} \quad (7.1.14)$$

Alla fine di ogni periodo di durata  $T$  bisogna decidere in base alle  $n$  misure rice-



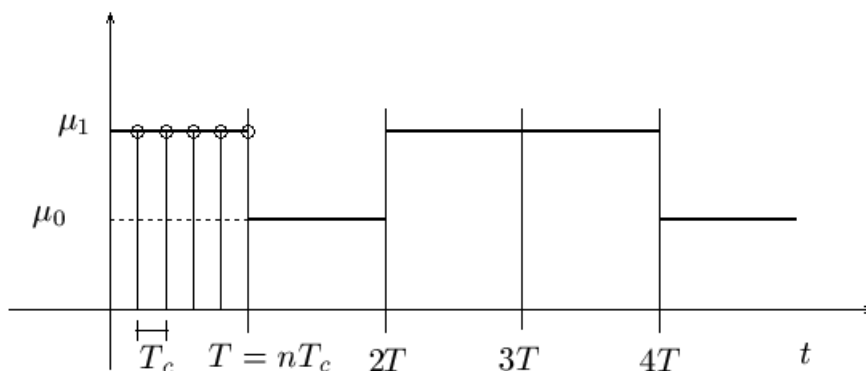


Figure 7.1.2. Segnale binario campionato

vute,  $y_1, \dots, y_n$ , se il segnale trasmesso era  $\mu_0$  o  $\mu_1$ . Lo schema del ricevitore che risulta dalla soluzione 7.1.12 è quello di figura 7.1.3

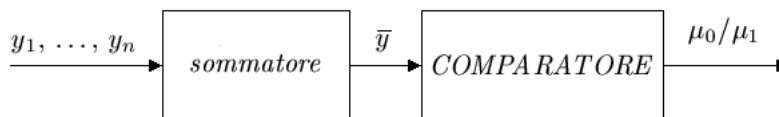


Figure 7.1.3. Ricevitore ottimo per la trasmissione numerica.

Questo ricevitore è progettato in modo tale che la probabilità di decidere erroneamente  $\mu_1$  quando il segnale effettivamente trasmesso è  $\mu_0$  è uguale al valore prefissato  $\alpha$ .

È altrettanto importante però conoscere la potenza  $1 - \beta = \mathbb{P}_D$ . Questa si può in generale calcolare solo a posteriori. In questo caso si ha

$$1 - \beta = \mathbb{P}_1\{y; \bar{y} \geq c_1\}$$

dove  $c_1$  ora è fissato. Basta allora notare che

$$\mathcal{C}_1 = \left\{y; \frac{\sqrt{n}}{\sigma}(\bar{y} - \mu_1) \geq \frac{\sqrt{n}}{\sigma}(c_1 - \mu_1)\right\}$$

dove  $\frac{\sqrt{n}}{\sigma}(\bar{y} - \mu_1) \sim \mathcal{N}(0, 1)$ , sotto  $H_1$ .

In fase di progetto si hanno dei limiti inferiori su  $1 - \beta$  in base ai quali si progetta il ricevitore. A questo scopo si può servirsi di grafici che danno  $1 - \beta = \mathbb{P}_D$  in funzione di  $\alpha$  e del rapporto

$$d = \frac{\sqrt{n}|\mu_1 - \mu_0|}{\sigma}$$

Questi grafici vengono chiamati ROC (*Receiver Operating Characteristic*) e si possono trovare in letteratura, ad esempio nel libro di Van Trees [67] a pag. 38. Essi sono del tipo di figura 7.1.4

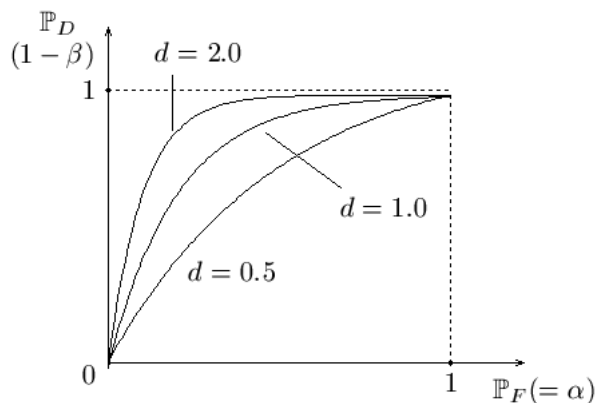


Figure 7.1.4. Caratteristiche del ricevitore (ROC).

**Example 7.2.** Siano  $H_0 = \mathcal{N}(0, \sigma_0^2)$ ,  $H_1 = \mathcal{N}(0, \sigma_1^2)$  due ipotesi da verificare in base all'osservazione di un campione casuale di numerosità  $N$ . Ovviamente

$$f_i(y_1 \dots y_N) = \frac{1}{(\sqrt{2\pi}\sigma_i)^N} \exp -\frac{1}{2} \frac{\sum_1^N y_t^2}{\sigma_i^2} \quad i = 0, 1$$

e

$$\Lambda(y) = \left(\frac{\sigma_0}{\sigma_1}\right)^N \exp \left\{ -\frac{1}{2} \left( \frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2} \right) \sum_1^N y_t^2 \right\}.$$

per cui la regione critica è definita dalla disuguaglianza

$$\frac{1}{2} \frac{\sigma_1^2 - \sigma_0^2}{\sigma_0^2 \sigma_1^2} \sum_1^N y_t^2 \geq N \log \frac{\sigma_1}{\sigma_0} + \log k.$$

Suponiamo che  $\sigma_1^2 > \sigma_0^2$  allora la disuguaglianza  $\Lambda(y) \geq k$  è equivalente alla

$$\sum_1^N y_t^2 \geq \frac{\sigma_0^1 \sigma_2^1}{\sigma_1^2 - \sigma_0^2} [N \log \frac{\sigma_1^2}{\sigma_0^2} + 2 \log k]$$

dove il secondo membro è  $> 0$ .

Per calcolare  $c^2$  nella formula che dà la regione critica

$$\mathcal{C}^* = \left\{ y; \sum_1^N y_t^2 \geq c^2 \right\}$$

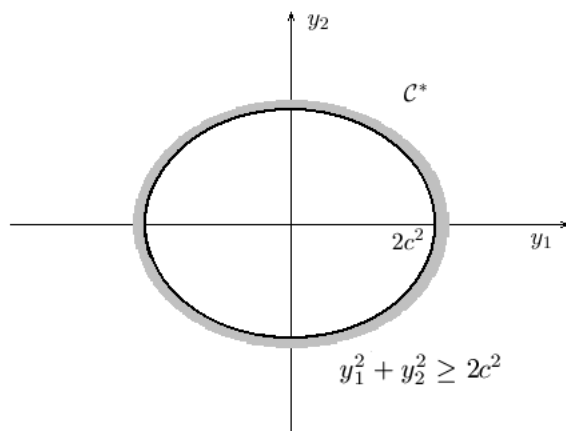


Figure 7.1.5. Regione critica per l'esempio 7.2

basta ricordare che, sotto  $H_0$

$$\frac{\sum_1^N \mathbf{y}_t^2}{\sigma_0^2} \sim \chi^2(N)$$

quindi, fissato  $\alpha$ , si va sulle tabelle di  $\chi^2(N)$  e si trova il valore  $a$  per cui  $\mathbb{P}\{\frac{\sum_1^N \mathbf{y}_t^2}{\sigma_0^2} \geq a\} = \alpha$ , e ovviamente si pone  $c^2 = \frac{a}{\sigma_0^2}$ . Per calcolare la potenza, ricordiamo che  $\frac{\sum_1^N \mathbf{y}_t^2}{\sigma_1^2} \sim \chi^2(N)$  e si va a vedere chi è

$$\mathbb{P}_1 \left[ \frac{\sum_1^N \mathbf{y}_t^2}{\sigma_1^2} \geq \frac{c^2}{\sigma_1^2} \right] = \mathbb{P} \left[ \frac{\sum_1^N \mathbf{y}_t^2}{\sigma_1^2} \geq \frac{a}{\sigma_0^2 \sigma_1^2} \right] = 1 - \beta.$$

Nello studio dei sistemi di comunicazione si ha spesso a che fare con problemi di discriminazione di segnale che si formulano in modo naturale come problemi di verifica d'ipotesi. Supponiamo qui che  $t$  sia la variabile temporale continua. Tipicamente si deve decidere se vale una delle due ipotesi

$$\begin{cases} H_1 : \mathbf{y}(t) = s(t) + \mathbf{w}(t) & 0 \leq t \leq T \\ H_0 : \mathbf{y}(t) = \mathbf{w}(t) & 0 \leq t \leq T \end{cases} \quad (7.1.15)$$

dove  $s(t)$  è il segnale utile, che può essere di forma completamente nota, nota a meno del valore di certi parametri,  $s(t) = s(t, \theta)$ , oppure ignota. Nell'ultimo caso si ha in genere una descrizione probabilistica di  $s(t)$ . Il rumore  $w(t)$  è noto probabilisticamente (tipicamente è rumore bianco Gaussiano).

**Example 7.3 (Il ricevitore a correlazione).** Supporremo  $s(t)$  sia una *funzione nota* e  $w(t)$  rumore *bianco Gaussiano* di media zero e varianza  $\sigma^2$  nota.

Per risolvere il problema useremo un'idea di U. Grenander [?]. Prendiamo un sistema di funzioni ortonormali in  $[0, T]$

$$\phi_1(t), \phi_2(t), \dots, \phi_n(t), \dots$$

con

$$\int_0^T \phi_i(t)\phi_j(t) dt = \delta_{ij}.$$

Come  $\phi_1(t)$  possiamo sempre scegliere la funzione

$$\phi_1(t) = \frac{s(t)}{\|s(\cdot)\|} = \frac{s(t)}{\left[\int_0^T s^2(t) dt\right]^{\frac{1}{2}}} \quad (7.1.16)$$

dove il denominatore è la radice quadrata dell'energia,  $E$ , del segnale. Calcolando la correlazione temporale di  $\mathbf{y}(t)$  con  $\phi_i(t)$  si trova

$$\mathbf{y}_i := \langle \mathbf{y}, \phi_i \rangle = \int_0^T \mathbf{y}(t)\phi_i(t) dt \quad i = 1, 2, \dots$$

per cui se vale  $H_1$  si ha

$$\begin{aligned} \mathbf{y}_1 &= \frac{1}{\sqrt{E}} \int_0^T s(t)s(t) dt + \int_0^T \mathbf{w}(t)\frac{s(t)}{E} dt \\ &= \sqrt{E} + \mathbf{w}_1 \end{aligned}$$

e

$$\mathbf{y}_i = \int_0^T \mathbf{w}(t)\phi_i(t) dt \quad i = 2, 3, \dots$$

dato che  $s(t)$  e  $\phi_i(t)$  sono ortogonali per  $i \geq 2$ .

Se vale  $H_0$  invece si ha

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{w}_1, \\ \mathbf{y}_i &= \mathbf{w}_i \quad i = 1, 2, \dots \end{aligned}$$

Notiamo che le variabili casuali  $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_i, \dots\}$  sono indipendenti e Gaussiane. Infatti

$$\begin{aligned} \mathbb{E}\mathbf{w}_i\mathbf{w}_j &= \mathbb{E} \int_0^T \mathbf{w}(t)\phi_i(t) dt \int_0^T \mathbf{w}(\tau)\phi_j(\tau) d\tau \\ &= \int_0^T \int_0^T \phi_i(t)\phi_j(\tau)\mathbb{E}[\mathbf{w}(t)\mathbf{w}(\tau)] dt d\tau \\ &= \int_0^T \int_0^T \phi_i(t)\phi_j(\tau)\sigma^2\delta(t-\tau) dt d\tau \\ &= \sigma^2 \int_0^T \phi_i(t)\phi_j(\tau) dt = \sigma^2\delta_{ij}. \end{aligned}$$

Ne segue che il problema può essere riformulato come segue.

Sotto  $H_1$

$$\begin{cases} \mathbf{y}_1 = \sqrt{E} + \mathbf{w}_1 \\ \mathbf{y}_i = \mathbf{w}_i \quad i = 2, 3, \dots \end{cases} \quad (7.1.17)$$

Sotto  $H_0$

$$y_i = \mathbf{w}_i, \quad i = 1, 2, 3, \dots \quad (7.1.18)$$

dove il "processo" discreto  $\{\mathbf{w}_i\}$  è *bianco*, Gaussiano di media zero e varianza  $\sigma^2$ . In altri termini le "osservazioni"  $\{\mathbf{y}_i\}$  sono una famiglia di variabili Gaussiane indipendenti sotto entrambe le ipotesi, in particolare,

$$\begin{cases} \mathbf{y}_1 \sim \mathcal{N}(\sqrt{E}, \sigma^2) & \text{sotto } H_1 \\ \mathbf{y}_1 \sim \mathcal{N}(0, \sigma^2) & \text{sotto } H_0 \end{cases} \quad (7.1.19)$$

e, se  $i \geq 2$ ,

$$\mathbf{y}_i \sim \mathcal{N}(0, \sigma^2) \quad (7.1.20)$$

sotto *entrambe* le ipotesi.

Calcolando  $\Lambda(\mathbf{y}_1 \dots \mathbf{y}_n)$  per un qualunque  $n$  finito si vede subito che per l'indipendenza e per la (7.1.20), si ha

$$\Lambda(\mathbf{y}_1 \dots \mathbf{y}_n) = \Lambda(\mathbf{y}_1) \quad (7.1.21)$$

ovvero *la decisione ottima è funzione solo del valore assunto da  $\mathbf{y}_1$* . Facendo i conti,

$$\begin{aligned} \Lambda(\mathbf{y}_1) &= \exp -\frac{1}{2\sigma^2} [(\mathbf{y}_1 - \sqrt{E})^2 - \mathbf{y}_1^2] \\ &= \exp -\frac{1}{2\sigma^2} [-2\mathbf{y}_1\sqrt{E} + E] \\ &= \exp \frac{1}{\sigma^2} \left[ \int_0^T \mathbf{y}(t)s(t) dt - \frac{1}{2} \int_0^T s^2(t) dt \right]. \end{aligned} \quad (7.1.22)$$

Questa formula si ritrova in numerose varietà di problemi di discriminazione. È nota col nome di "*Likelihood Ratio formula*".

La regione critica del test si ottiene imponendo che

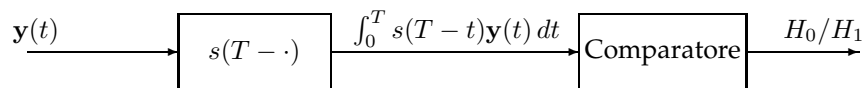
$$\log \Lambda(\mathbf{y}_1) \geq k$$

ovvvero

$$\mathbf{y}_1 = \int_0^T y(t)s(t) dt \geq \frac{1}{2} \int_0^T s^2(t) dt + \sigma^2 k = \frac{1}{2}E + \sigma^2 k = c.$$

Ricordiamo che  $\mathbf{y}_1 \sim \mathcal{N}(\sqrt{E}, \sigma^2)$ , sotto  $H_0$ , per cui  $c$  si può calcolare dalla  $c = \frac{a}{\sigma} - \sqrt{E}$ , dove  $a$  è il punto della curva di  $\mathcal{N}(0, 1)$  corrispondente a probabilità della coda superiore pari ad  $\alpha$ .

Lo schema del ricevitore ottimo è riportato in figura 7.1.6



**Figure 7.1.6.** *Struttura del ricevitore a correlazione*

Questo schema viene anche chiamato “ricevitore a correlazione” o a “filtro adattato”. Il secondo nome deriva dal fatto che  $y_1$  si può pensare come l’uscita all’istante  $T$  del filtro lineare di funzione di trasferimento

$$h(t) = s(T - t)$$

(che non è causale).

Sul problema della discriminazione di segnali sono stati scritti molti libri. Oltre al classico testo di van Trees già citato, segnaliamo specialmente [34].

### 7.1.1 Critica all’approccio classico

Come abbiamo già detto, nella teoria statistica classica dei tests di ipotesi  $H_0$  e  $H_1$  giocano un ruolo non simmetrico; in particolare  $H_0$  gioca un ruolo in un certo senso privilegiato, dato che fissando  $\alpha$  ci si tutela solo sulla probabilità di commettere un errore di prima specie (rifiutando  $H_0$  quando è vera). Questa libertà di scegliere  $\alpha$  può condurre a paradossi. Accade normalmente che tutelandosi in modo molto conservativo per evitare un possibile rifiuto di  $H_0$ , ovvero scegliendo  $\alpha$  molto piccolo, si finisce coll’acceptare  $H_0$  quando invece la scelta di  $H_1$  sarebbe più ragionevole. Sebbene il Lemma di Neyman-Pearson garantisca che la probabilità dell’errore di seconda specie  $\beta$  venga minimizzata dalla procedura, in genere  $\alpha$  non dice nulla (o quasi) sulla correttezza della scelta di  $H_0$  quando in realtà vale  $H_1$ . Occorrerebbe a questo scopo confrontare  $\alpha$  con la probabilità dell’errore di seconda specie,  $\beta$ , di accettare  $H_0$  quando invece vale  $H_1$ . Nei problemi ingegneristici, come quelli illustrati sopra, spesso le due ipotesi giocano spesso un ruolo completamente simmetrico e sarebbe forse più ragionevole considerare procedimenti che garantiscono uguali probabilità d’errore  $\alpha$  e  $\beta$ . Da questo punto di vista l’approccio Bayesiano, in cui si postula una distribuzione a priori  $\{p_0, p_1\}$  che misura la verosimiglianza delle due ipotesi, è meno criticabile.

## 7.2 Ipotesi composte

Molto spesso  $H_1$  (e/o  $H_0$ ) è un’ipotesi *composta*:

$$H_i = \{f(\cdot, \theta); \theta \in \Theta_i\} \quad i = 0, 1$$

dove  $\Theta_i$  è un sottoinsieme di  $\Theta$ . Questo accade ad esempio nel problema di discriminazione di segnali se  $s(t)$  è funzione di uno o più parametri incogniti. In

questa situazione nel rapporto

$$\Lambda(y, \theta) = \frac{f(y, \theta_1)}{f(y, \theta_0)} \quad \theta_1 \in \Theta_1, \theta_0 \in \Theta_0 \quad (7.2.1)$$

in cui abbiamo provvisoriamente introdotto il parametro composto,  $\theta = (\theta_1, \theta_0)$  dove  $\theta_1 \in \Theta_1, \theta_0 \in \Theta_0$ , è abbastanza naturale sostituire al parametro  $(\theta_1, \theta_0)$  una sua stima. Questo significa, sostituire a  $\theta$ , uno *stimatore*  $\hat{\theta}$ , funzione di  $y$ , che tenga conto delle diverse regioni ammissibili del parametro sotto le due ipotesi. Le cose funzionano bene, a patto di prendere  $\hat{\theta}$  uguale allo *stimatore di massima verosimiglianza*, nel senso che se a  $\theta_i$  nel rapporto (7.2.1) si sostituisce la statistica  $\hat{\theta}_i, i = 0, 1$ , che massimizza  $f(y, \theta_i)$  nell'insieme  $\Theta_i, i = 0, 1$ , il test ha certe proprietà ottimali quali la consistenza, l'asintotica normalità, etc.<sup>29</sup> uguali a quelle di cui gode lo stimatore di M.V..

**Definition 7.4.** Sia  $H_0 = \{f(\cdot, \theta); \theta \in \Theta_0\}$  e  $H_1 = \{f(\cdot, \theta); \theta \in \Theta_1\}$ . Si chiama rapporto di (massima) verosimiglianza la *quantità*

$$L(y) := \frac{f(y, \hat{\theta}_1(y))}{f(y, \hat{\theta}_0(y))} \quad (7.2.2)$$

dove le statistiche  $\hat{\theta}_i, i = 0, 1$ , massimizzano nei rispettivi domini  $\Theta_i$  la funzione di verosimiglianza  $f(y, \cdot)$ , ovvero,

$$\hat{\theta}_1(y) := \text{Arg} \max_{\{\theta \in \Theta_1\}} f(y, \theta) \quad (7.2.3)$$

$$\hat{\theta}_0(y) := \text{Arg} \max_{\{\theta \in \Theta_0\}} f(y, \theta). \quad (7.2.4)$$

Si può allora pensare di definire la regione critica del test assumendo,

$$C := \{y; L(y) \geq k\} \quad (7.2.5)$$

Naturalmente per fissare la costante  $k$  bisogna usare una procedura un poco più complicata dato che ora  $\alpha = \alpha(\theta)$  è funzione di  $\theta \in \Theta_0$ . Notiamo però che

$$\alpha(\theta) = \int_C f(y, \theta) dy \quad \theta \in \Theta_0$$

e, se prendiamo il max rispetto a  $\theta \in \Theta_0$  nei due membri e supponiamo che il max dell'integrale sia l'integrale del max rispetto a  $\theta$ , si ha

$$\alpha_0 = \max_{\theta \in \Theta_0} \alpha(\theta) = \int_C f(y, \hat{\theta}_0(y)) dy.$$

<sup>29</sup>Non possiamo qui addentrarci nella definizione di queste proprietà, per cui rimandiamo il lettore alla bibliografia.

Chiamiamo  $\hat{f}_0(y)$  la densità  $f(y, \hat{\theta}_0(y))$ . Allora se si prende  $k$  in modo tale che

$$\int_C \hat{f}_0(y) dy = \int_{\{L(y) \geq k\}} \hat{f}_0(y) dy = \int_k^\infty \hat{p}_0(l) dl = \alpha_0,$$

dove  $\hat{p}_0(l)$  è la distribuzione della v.a.  $L = L(\mathbf{y})$  con  $\mathbf{y} \sim \hat{f}_0(y)$ , si commette un errore di prima specie  $\leq \alpha_0$ .

In realtà nella grande maggioranza dei casi pratici, sotto  $H_0$ ,  $L(\mathbf{y})$  è distribuita in modo indipendente da  $\theta$ , cioè la d.d.p.  $p_0(l)$  di  $L(\mathbf{y})$ , con  $\mathbf{y} \sim \{f(y, \theta), \theta \in \Theta_0\}$ , non dipende da  $\theta$ . In questo caso  $p_0(\cdot)$  si può calcolare come se la d.d.p. di  $\mathbf{y}$  corrispondesse ad un qualunque valore  $\bar{\theta}$  di  $\Theta_0$ , in particolare a quello,  $\hat{\theta}_0$ , che dà il massimo di  $f(y, \theta)$  in  $\Theta_0$ . Ne segue che  $\hat{p}_0 = p_0$  e quindi, fissato  $\alpha$ , se si prende la regione critica  $C = \{y; L(y) \geq k_\alpha\}$  dove  $k_\alpha$  è determinata dalla

$$\int_{k_\alpha}^\infty p_0(l) dl = \alpha \tag{7.2.6}$$

si ha la probabilità  $\alpha$  di commettere un errore di prima specie qualunque sia  $\theta \in \Theta_0$ .

In generale invece la distribuzione di  $L(\mathbf{y})$  sotto  $H_1$  dipende da  $\theta \in \Theta_1$ , per cui la potenza del test

$$[1 - \beta](\theta) = \int_{k_\alpha}^\infty p_1(l, \theta) dl \tag{7.2.7}$$

( $p_1(\cdot, \theta)$  è la d.d.p. di  $L(y)$  con  $\mathbf{y} \sim \{f(\cdot, \theta); \theta \in \Theta_1\}$ ) è funzione di  $\theta \in \Theta_1$ . Notiamo che, se nella (7.2.7) si prende  $\theta \in \Theta_0$  si ha

$$p_1(l, \theta) \equiv p_0(l)$$

e  $[1 - \beta](\theta) = \alpha$ , indipendentemente da  $\theta$ .

Spesso si usa definire la potenza del test semplicemente come la probabilità di rifiutare  $H_0$ . In quest'ottica, la potenza è una funzione di  $\theta$  definita  $\forall \theta \in \Theta$ .

Per  $\theta \in \Theta_0$  (cioè sotto  $H_0$ ) si ottiene l'errore di prima specie  $\alpha$  (o  $\alpha(\theta)$  in generale). Un tipico andamento della funzione potenza è illustrato in figura ??.

**Example 7.4.** Sia  $\mathbf{y} \sim \mathcal{N}(\theta_1, \theta_2^2)$  e  $H_0 : \theta_1 = \mu, \theta_2^2 > 0$ , dove  $\mu$  è un valore fissato. Vogliamo verificare l'ipotesi  $H_0$ , cioè che la media di  $\mathbf{y}$  sia proprio il valore assegnato a  $\mu$ , contro tutte le possibili alternative, sulla base di  $N$  osservazione indipendenti estratte da  $\mathcal{N}(\theta_1, \theta_2^2)$ . Ovviamente si ha

$$\Theta_0 = \{\theta; \theta_1 = \mu, \theta_2^2 > 0\}$$

$$\Theta_1 = \{\theta; \theta_1 \neq \mu, \theta_2^2 > 0\}$$

e quindi  $\Theta_1$  corrisponde al semipiano aperto  $\{\theta_1, \theta_2^2 > 0\}$  privo della semiretta  $\theta_1 = \mu$ . La funzione di verosimiglianza

$$f(y_1, \dots, y_N, \theta) = (2\pi\theta_2^2)^{-\frac{N}{2}} \exp \left\{ -\frac{1}{2\theta_2^2} \sum_1^N (y_t - \theta_1)^2 \right\} \tag{7.2.8}$$



va massimizzata separatamente su  $\Theta_0$  e su  $\Theta_1$ . Su  $\Theta_0$ , questo corrisponde a calcolare lo stimatore di M.V. per  $\theta_2^2$  quando la media è nota e vale  $\mu$ . Quindi

$$(\hat{\theta}_2^2)_0(y) = \bar{s}_N^2(y) = \frac{1}{N} \sum_1^N (y_t - \mu)^2 \quad (7.2.9)$$

per cui

$$f(y, \hat{\theta}_0(y)) = [2\pi \bar{s}_N^2(y)]^{-\frac{N}{2}} \exp\left(-\frac{N}{2}\right). \quad (7.2.10)$$

Per massimizzare  $f(y, \theta)$  su  $\Theta_1$  si può massimizzare su  $\Theta = \{\theta_1, \theta_2^2 > 0\}$  e controllare poi che i valori ottenuti di  $\hat{\theta}_1$  non stanno sulla retta  $\theta_1 = \mu$ . Si ha allora

$$\begin{aligned} (\hat{\theta}_1)_1(y) &= \bar{y}_N = \frac{1}{N} \sum_1^N y_t \\ (\hat{\theta}_2^2)_1(y) &= s_N^2(y) = \frac{1}{N} \sum_1^N (y_t - \bar{y}_N)^2. \end{aligned} \quad (7.2.11)$$

Ovviamente  $\bar{y}_N = \mu$  con probabilità zero  $\forall \theta \in \Theta$ , per cui queste stime di M.V. sono anche il massimo di  $f(y, \theta)$  su  $\Theta_1$ . A conti fatti,

$$f(y, \hat{\theta}_1(y)) = [2\pi s_N^2(y)]^{-\frac{N}{2}} \exp\left(-\frac{N}{2}\right) \quad (7.2.12)$$

per cui, facendo il rapporto, si trova

$$\begin{aligned} L(y) &= \left[ \frac{s_N^2(y)}{\bar{s}_N^2(y)} \right]^{-\frac{N}{2}} = \left[ \frac{s_N^2(y) + (\bar{y}_N - \mu)^2}{s_N^2(y)} \right]^{\frac{N}{2}} \\ &= \left[ 1 + \frac{(\bar{y}_N - \mu)^2}{s_N^2(y)} \right]^{\frac{N}{2}}. \end{aligned} \quad (7.2.13)$$

Definiamo ora la variabile casuale

$$\mathbf{t} := \frac{\mathbf{y} - \mu}{\sqrt{\frac{s_N^2(\mathbf{y})}{N-1}}} \quad (7.2.14)$$

detta  $t$  di Student, che sotto l'ipotesi  $H_0$  ha una distribuzione notevole che appare spesso in statistica.

### 7.2.1 La distribuzione di Student

Siano  $\mathbf{y} \sim \mathcal{N}(0, 1)$  e  $\mathbf{x} \sim \chi^2(n)$ , variabili indipendenti. Allora il rapporto

$$\mathbf{t} := \frac{\mathbf{y}}{\sqrt{\mathbf{x}/n}} \quad (7.2.15)$$

è distribuito secondo la densità di probabilità

$$p_n(t) = \frac{1}{\sqrt{n} B(1/2, n/2)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \quad t \in \mathbb{R} \quad (7.2.16)$$

che si chiama *distribuzione di Student a n gradi di libertà* e si denota col simbolo  $S(n)$ . Nella (7.2.16)  $B$  è la funzione Beta di Eulero, definita dalla

$$B(p, q) := \int_0^1 x^{p-1}(1-x)^{q-1} dx = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$$

dove la funzione  $\Gamma$  è la nota generalizzazione del fattoriale. Come è noto, per  $n$  intero maggiore di 1 la funzione Gamma vale  $\Gamma(n) = (n-1)!$ .

La dimostrazione della (7.2.16) si trova ad esempio in [26, p. 135]. La distribuzione di Student è una delle distribuzioni notevoli della statistica classica. Essa si trova tabulata in varie forme in letteratura. Per  $n = 1$  essa si riduce alla distribuzione di Cauchy

$$S(1) \equiv \frac{1}{\pi(1+t^2)}$$

e quindi si vede che alcuni momenti di  $S(n)$  possono non esistere. In effetti si dimostra che  $S(n)$  possiede momenti fino fino all'  $n - 1$ -simo compreso e che questi valgono

$$\begin{aligned} \mu_r &= 0 \quad \text{se } r \text{ è dispari e } r < n \\ \mu_r &= \frac{\Gamma(\frac{1}{2}n - r)\Gamma(r + \frac{1}{2})}{\Gamma(\frac{1}{2}n)\Gamma(\frac{1}{2})} \quad \text{se } r \text{ è pari e } 2r < n. \end{aligned}$$

Si può poi mostrare direttamente che per  $n \rightarrow \infty$  la  $S(n)$  converge ad una normale  $\mathcal{N}(0, 1)$ .

Da quanto esposto sopra si vede che la statistica  $t$  in (7.2.14) è distribuita secondo la legge di Student,  $S(N - 1)$ , ad  $N - 1$  gradi di libertà. Dato che  $L(y)$  si può scrivere come

$$L(y) = \left[1 + \frac{1}{N-1}t^2\right]^{\frac{N}{2}} \quad (7.2.17)$$

si vede che  $L(y)$  dipende da  $y$  solo attraverso la statistica  $t$ . Notiamo che la regione critica  $C := \{y; L(y) \geq k\}$  può scriversi equivalentemente come

$$C := \left\{y; |t(y)| \geq +\sqrt{(N-1)(k^{\frac{2}{N}} - 1)}\right\} \quad (7.2.18)$$

ovvero

$$C := \{y; |t(y)| \geq c\}, \quad c > 0.$$

Dalla (7.2.17) si vede che sotto  $H_0$ ,  $L(y)$  ha distribuzione indipendente dal parametro libero,  $\theta_2^2$  e quindi la probabilità,  $\alpha$ , di commettere un errore di prima specie, non dipende da  $\theta_2^2$ . Fissato  $\alpha$  si può infatti trovare  $c_\alpha$  tale che

$$\int_{c_\alpha}^\infty p_{N-1}(t) dt = \frac{\alpha}{2} \quad (7.2.19)$$

e questo valore  $c_\alpha$  definisce una regione critica tale che la probabilità di rifiutare (quando è vera) l'ipotesi  $\theta_1 = \mu$ , è la stessa qualunque sia il valore della varianza incognita  $\theta_2^2$ . Invece, sotto  $H_1$ , la v.a.  $t$  non è più distribuita come  $\mathcal{S}(N-1)$  ed in generale la distribuzione di  $t$  quando  $\mathbf{y} \sim \mathcal{N}(\theta_1, \theta_2^2)$  dipende da  $\theta_1$  e da  $\theta_2^2$  attraverso il cosiddetto "parametro di non centralità"

$$\delta = \frac{\sqrt{N}}{\theta_2}(\theta_1 - \mu).$$

Il calcolo della potenza si presenta difficoltoso (si possono usare tavole della distribuzione  $\mathcal{S}$  "non centrale").  $\diamond$

### Osservazioni

Vale la pena di estrapolare dall'esempio appena discusso alcune considerazioni di carattere generale. La prima è la seguente.

In molti casi l'ipotesi da verificare,  $H_0$ , si presenta assegnando un valore fissato per alcune delle componenti ( $\theta_1, \dots, \theta_p$ ) del parametro  $p$ -dimensionale  $\theta$ . Se supponiamo che queste componenti siano le prime  $k$  ( $k \geq 1$ ) e scriviamo  $\theta$  come

$$\theta = \begin{bmatrix} \beta \\ \eta \end{bmatrix} \quad \beta \in \mathbb{R}^k, \eta \in \mathbb{R}^{p-k} \quad (7.2.20)$$

allora possiamo supporre che  $H_0$  sia della forma

$$H_0 := \{\theta; \beta = \beta_0\} \quad , \quad (7.2.21)$$

con  $\beta_0$  un vettore fissato di  $\mathbb{R}^k$ . In questi casi  $H_1$  è l'ipotesi alternativa

$$H_1 := \{\theta; \beta \neq \beta_0\} \quad . \quad (7.2.22)$$

Ne viene che  $\Theta_0$ , o si riduce ad un punto ( $k = p$ ) oppure sta in un sottospazio di  $\mathbb{R}^p$  di dimensione inferiore a  $p$  per cui la massimizzazione di  $f(y, \theta)$  su  $\Theta_1$  dà (con probabilità uno) lo stesso risultato della massimizzazione di  $f(y, \theta)$  sull'intero spazio dei parametri  $\Theta$ , questo a meno di situazioni patologiche (che hanno probabilità zero) in cui  $f(y, \theta)$  è massimizzata da valori  $\hat{\theta}_i$  dei parametri che non dipendono dai dati osservati. In generale, quindi

$$\max_{\theta \in \Theta_1} f(y, \theta) = \max_{\theta \in \Theta} f(y, \theta) \quad (7.2.23)$$

e cioè  $\hat{\theta}_1(\mathbf{y})$  è l'ordinario stimatore di M.V.,  $\hat{\theta}(\mathbf{y})$ , di  $\theta$ . Per questa ragione, quando  $H_0$  e  $H_1$  sono nella forma (7.2.21), (7.2.22) si può scrivere

$$L(\mathbf{y}) = \frac{f(\mathbf{y}, \hat{\theta}(\mathbf{y}))}{f(\mathbf{y}, \beta_0, \hat{\eta}(\mathbf{y}))} \quad (7.2.24)$$

dove  $\hat{\eta}(\mathbf{y})$  è lo stimatore ("condizionato") di M.V., ovvero,  $\eta(\mathbf{y})$  massimizza  $f(\mathbf{y}, \beta_0, \eta)$  rispetto ad  $\eta$  nella regione ammissibile  $\Theta_0$ , per l'ipotesi  $H_0$ .

Notiamo allora che, siccome  $\Theta_0 \subset \Theta$ , si ha sempre

$$f(y, \hat{\theta}(y)) = \max_{\theta \in \Theta} f(y, \theta) \geq \max_{\theta \in \Theta_0} f(y, \theta) = f(y, \beta_0, \hat{\eta}(y)) \quad (7.2.25)$$

e quindi  $L(y) \geq 1 \forall y \in \mathbb{R}^n$  (notare che nella (7.2.18) bisogna supporre infatti  $k \geq 1$ ). Intuitivamente, tanto più grande è  $L(y)$ , tanto più “verosimile” è l’ipotesi  $H_1$ , dato che  $f(y, \hat{\theta}_1(y)) \gg f(y, \hat{\theta}_0(y))$ , e quindi si accetta  $H_1$  nella regione  $\{L(y) \geq k\}$ .

Nel seguito supporremo che  $H_0$  e  $H_1$  siano ipotesi del tipo (7.2.21), (7.2.22). Le connessioni con la stima di M.V., in particolare con il teorema di Wald (Teorema 4.6) che vedremo nel Capitolo 4, permettono di dimostrare dei risultati molto precisi ed utili quando  $N$  è grande (per “grandi campioni” come si dice comunemente). Anche qui bisognerà supporre che le osservazioni  $(y_1 \dots y_N)$  siano un campione casuale.

**Theorem 7.1 (Wald).** *Si consideri la partizione (7.2.20) e sia  $H_0$  definita come in (7.2.21), allora, sotto  $H_0$ ,*

1. la statistica

$$l(\mathbf{y}) := 2 \log L(\mathbf{y}) \quad (7.2.26)$$

converge quando  $N \rightarrow \infty$  con probabilità 1 verso

$$Q(\mathbf{y}) = [\hat{\beta}(\mathbf{y}) - \beta_0]^\top I^{-1}(\beta_0) [\hat{\beta}(\mathbf{y}) - \beta_0] \quad (7.2.27)$$

dove  $\hat{\beta}(\mathbf{y})$  è il vettore delle prime  $k$ -componenti dello stimatore a M.V. di  $\theta$  e  $I(\beta)$  è la sottomatrice di informazione di Fisher, di dimensione  $k \times k$ , corrispondente al parametro  $\beta$ .

2. Sotto  $H_0$ ,  $\sqrt{N} \hat{\beta}_k(\mathbf{y}) \xrightarrow{L} \mathcal{N}(\beta_0, I^{-1}(\beta_0))$  e quindi, asintoticamente

$$Q(\mathbf{y}) \sim \chi^2(k). \quad (7.2.28)$$

3. Sotto  $H_1$ , ponendo  $\theta_0 = (\beta_0, \eta)$  si ha

$$Q(\mathbf{y}) \sim \chi^2(k, \lambda)$$

( $\chi^2$  non centrale) dove il parametro di non centralità  $\lambda$  è dato dalla formula

$$\lambda = [\theta - \theta_0]^\top I^{-1}(\theta_0) [\theta - \theta_0]. \quad (7.2.29)$$

4. Inoltre il test di max verosimiglianza è consistente nel senso che se  $C_N$  è una regione critica di misura  $\alpha$  fissata, basato su  $N$  campioni, allora,

$$\lim_{N \rightarrow \infty} \int_{C_N} f(y_1, \dots, y_N, \theta) dy^N = 1 \quad \forall \theta \in \Theta_1 \quad (7.2.30)$$

dove  $C_N = \{y_1, \dots, y_N; L(y_1, \dots, y_N) \geq k_\alpha\}$ .

Maggiori dettagli si possono trovare sul [60] vol. II pag. 240 e segg.

### 7.3 Tests di ipotesi sul modello lineare

Riscriviamo qui il modello lineare standardizzato  $N$ -dimensionale

$$y = S\theta + \sigma w, \quad w \sim \mathcal{N}(0, I_N) \quad (7.3.1)$$

ottenuto eventualmente col procedimento di normalizzazione illustrato all'inizio del capitolo 2 (osservazione 2.1).

Si chiamano *ipotesi lineari* quelle esprimibili attraverso funzioni lineari di  $\theta$ , nella fattispecie,  $H_0$  è un'ipotesi lineare, se per qualche  $H \in \mathbb{R}^{k \times p}$ ,  $k \leq p$  e  $\beta_0$  un vettore assegnato in  $\mathbb{R}^k$ , si può scrivere

$$H_0 := \{\theta; H\theta = \beta_0\}. \quad (7.3.2)$$

È chiaro che vale la pena di considerare solo ipotesi espresse mediante matrici  $H$  di rango pieno,  $k < p$ . Alla verifica di ipotesi lineari può essere ricondotta la diagnostica del modello lineare, che comprende l'insieme di verifiche a posteriori sulla significatività delle stime puntuali di  $\theta$ , ottenute con i metodi della M.V. (o dei M.Q.) su un insieme assegnato di osservazioni. Tipici esempi sono:

1. *Ipotesi di adeguatezza del modello lineare*

$$H_0 := \{\theta = 0\} \quad (7.3.3)$$

(in questo caso  $H = I$ ,  $\beta_0 = 0$ ). Si tratta di verificare se esiste effettivamente un accoppiamento tra segnale e misure (almeno se ne esiste uno di tipo *lineare nei parametri*, come quello ipotizzato dal modello lineare). Se si accetta  $H_0$  con livello di significatività  $\alpha$ , si decide che con certezza statistica  $1 - \alpha$ , le misure  $y$  sono essenzialmente costituite da rumore.

2. *Ipotesi sul numero di parametri significativi*

$$H_0 := \{\theta_{k+1} = \theta_{k+2} = \dots = \theta_p = 0\} \quad (7.3.4)$$

Si tratta di verificare se il modello è *sovraparametrizzato*. In generale se si confrontano tra di loro modelli lineari con un numero diverso di parametri ci si può sempre ridurre a verificare un'ipotesi del tipo (7.3.4). Si voglia ad esempio decidere quale dei due modelli di regressione

$$y_t = \theta_0 + \theta_1 u_t + \epsilon_t \quad (7.3.5)$$

$$y_t = \theta_0 + \theta_1 u_t + \theta_2 u_t^2 + \epsilon_t \quad (7.3.6)$$

$t = 1, \dots, N$ , si adatta meglio alle osservazioni  $(u_1, \dots, u_N)$  e  $(y_1, \dots, y_N)$ . Il modello (7.3.5) corrisponde evidentemente all'ipotesi  $H_0 : \theta_2 = 0$ .

3. *Ipotesi sulla significatività delle stime puntuali.*

Chiamiamo *regione di confidenza* (sotto l'ipotesi  $H_0$ ) di misura  $1 - \alpha$ , la regione complementare a quella critica (di misura  $\alpha$ ). Se  $\hat{\theta}(\bar{y})$  è la stima di  $\theta$

corrispondente all'osservazione  $\bar{y}$ , si può pensare di validare la stima verificando l'ipotesi

$$H_0 : \left\{ \theta = \hat{\theta}(\bar{y}) \right\} \quad . \quad (7.3.7)$$

In questo caso cercare la regione critica del test significa sostanzialmente cercare una regione di confidenza per  $\hat{\theta}(\bar{y})$  di coefficiente  $\alpha$  uguale al livello di significatività del test.

Consideriamo dunque il modello lineare (7.3.1) ed esaminiamo l'effetto del vincolo lineare  $H\theta = \beta_0$  (che vale solo sotto l'ipotesi  $H_0$ ) sulla stima di  $\theta$ .

Ricordiamo che, data la normalità delle osservazioni, il metodo di M.V. si riduce a quello dei M.Q. (non pesati ovvero con  $R^{-1} = I$ , nel nostro caso).

Sotto  $H_0$ , lo stimatore  $\hat{\theta}_0$  si trova minimizzando la distanza di  $y$  dal sottospazio colonne  $\mathcal{S}$ , di  $S$ , tenendo conto del vincolo (7.3.2); in altri termini, la combinazione lineare delle colonne di  $S$  che minimizza  $\|y - S\theta\|^2$  non può più essere fatta con coefficienti  $\theta \in \mathbb{R}^p$  arbitrari, ma deve invece essere costruita con coefficienti  $(\theta_1, \dots, \theta_p)$  che soddisfano l'equazione  $H\theta = \beta_0$ .

$$\hat{\theta}_0(y) = \text{Arg} \min_{\theta \in \{\theta; H\theta = \beta_0\}} \|y - S\theta\|^2 \quad . \quad (7.3.8)$$

Se  $H$  ha rango  $k$ , la condizione  $H\theta = \beta_0$  ( $k$  equazioni lineari in  $\theta$ ) fornisce solo  $p - k$  "parametri liberi" tra le  $p$  componenti di  $\theta$  e quindi solo  $p - k$  combinazioni lineari indipendenti delle colonne di  $S$ . Questo significa che il minimo di  $\|y - S\theta\|^2$  soggetto a  $H\theta = \beta_0$  si trova proiettando  $y$  non più su tutto  $\mathcal{S}$  ma bensì su un opportuno sottospazio affine  $p - k$  dimensionale,  $\mathcal{H}$ , di  $\mathcal{S}$  definito dalla

$$\mathcal{H} := \text{span} \{S\theta; H\theta = \beta_0\} \quad (7.3.9)$$

Tutto questo vale ovviamente nell'ipotesi che il vincolo (7.3.2) sussista effettivamente, cioè sotto  $H_0$ . Se neghiamo  $H_0$ , diciamo in sostanza che il vincolo non sussiste più e quindi (sotto  $H_1$ ) la stima di M.V.  $\hat{\theta}_1(y) = \hat{\theta}(y)$  si trova col metodo usuale, cioè proiettando  $y$  su  $\mathcal{S}$ . Si veda la figura 7.3.1

Dato che  $\mathcal{S} \supset \mathcal{H}$ , la distanza del punto estremo del vettore  $y$  da  $\mathcal{S}$  dev'essere minore di quella da  $\mathcal{H}$ ; si vede quindi abbastanza chiaramente che la somma dei quadrati dei residui nelle due situazioni è diversa. Se vale  $H_1$  si ha un' "errore di approssimazione" di  $y$  mediante  $S\hat{\theta}$  che è sempre minore del corrispondente errore  $\|y - S\hat{\theta}_0(y)\|^2$  sotto  $H_0$ . Definiamo allora la somma dei quadrati dei residui sotto le due ipotesi

$$\begin{aligned} H_0 : \quad R_0^2(y) &= \|y - S\hat{\theta}_0(y)\|^2 \\ H_1 : \quad R_1^2(y) &= \|y - S\hat{\theta}(y)\|^2 \end{aligned} \quad (7.3.10)$$

Tra  $R_0^2$  ed  $R_1^2$  sussiste una semplice relazione abbastanza ovvia se si guarda alla geometria del problema.

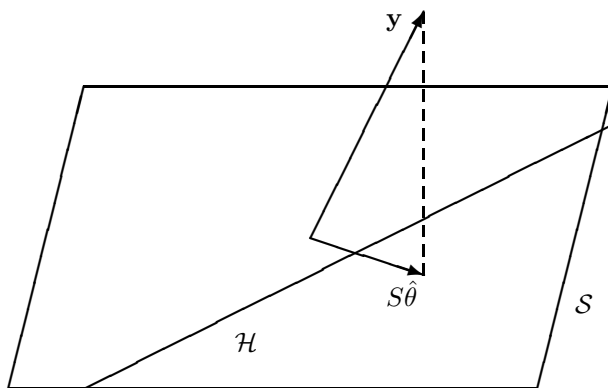


Figure 7.3.1. Proiezione ortogonale sul sottospazio  $\mathcal{H}$ .

**Lemma 7.3.**  $S\hat{\theta}_0(y)$  è la proiezione ortogonale di  $S\hat{\theta}(y)$  su  $\mathcal{H}$  e quindi

$$R_0^2(y) = R_1^2(y) + \|S\hat{\theta}(y) - S\hat{\theta}_0(y)\|^2 \tag{7.3.11}$$

*Proof.* Introduciamo per comodità i simboli

$$\hat{\mu}_0 := S\hat{\theta}_0(y) \quad , \quad \hat{\mu}_1 := S\hat{\theta}(y) . \tag{7.3.12}$$

Per provare la prima affermazione ricordiamo che  $y - \hat{\mu}_1$  è ortogonale a  $\mathcal{S}$  e quindi in particolare a  $\mathcal{H}$ . Inoltre  $y - \hat{\mu}_0$  è ortogonale a  $\mathcal{H} \subset \mathcal{S}$  (principio di ortogonalità). Per la linearità del prodotto scalare,  $\langle y - \hat{\mu}_1 - (y - \hat{\mu}_0), \mathcal{H} \rangle = 0 \Rightarrow \langle \hat{\mu}_1 - \hat{\mu}_0, \mathcal{H} \rangle = 0$ .

Allora

$$R_0^2 = \|y - \hat{\mu}_1 + \hat{\mu}_1 - \hat{\mu}_0\|^2 = \|y - \hat{\mu}_1\|^2 + \|\hat{\mu}_1 - \hat{\mu}_0\|^2 + 2 \langle y - \hat{\mu}_1, \hat{\mu}_1 - \hat{\mu}_0 \rangle \tag{7.3.13}$$

ma il prodotto scalare è nullo perchè  $\hat{\mu}_0$  e  $\hat{\mu}_1 \in \mathcal{S}$  (e quindi anche la loro differenza) e  $y - \hat{\mu}_1$  è ortogonale a  $\mathcal{S}$  ( $\hat{\mu}_1$  è la proiezione ortogonale su  $\mathcal{S}$ !). Quindi la (7.3.13) si riduce a

$$R_0^2 = R_1^2 + \|\hat{\mu}_1 - \hat{\mu}_0\|^2$$

(teorema di Pitagora) che è proprio la (7.3.13).  $\square$

Notiamo che se  $H_0$  è vera gli scarti  $\|\hat{\mu}_1 - \hat{\mu}_0\|^2$  (che dipendono dal campione osservato  $y$ ) saranno in media piccoli dato che la stima,  $S\hat{\theta}(y)$ , anche se calcolata senza tener conto della (7.3.2) tende per  $n$  grande ad essere molto vicina a  $S\theta_0$  ( $\theta_0$  è il valore vero) e  $S\theta_0$  sta in  $\mathcal{H}$  per ipotesi.

Viceversa se  $H_0$  è falsa  $S\hat{\theta}(y)$  rimane fuori dal sottospazio  $\mathcal{H}$  anche se  $N \rightarrow \infty$ . Se ne ricava che il rapporto

$$\frac{\|\hat{\mu}_1 - \hat{\mu}_0\|^2}{R_1^2} = \frac{R_0^2 - R_1^2}{R_1^2} \tag{7.3.14}$$

è *piccolo* se vale  $H_0$  e *grande* se vale  $H_1$ .

Vedremo che questo intuitivo criterio di verifica di  $H_0$  è in sostanza quello fornito dal test di massima verosimiglianza.

## 7.4 Calcolo del rapporto di Max verosimiglianza

Come abbiamo visto la densità  $f(y, \theta, \sigma^2)$  del vettore aleatorio  $y$ , si può scrivere come

$$f(y, \theta, \sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp -\frac{1}{2\sigma^2} \|y - S\theta\|^2 \quad (7.4.1)$$

e, dalla discussione precedente segue che *sotto*  $H_1$ , lo stimatore  $(\hat{\theta}_1, \hat{\sigma}_1^2)$  di  $(\theta, \sigma^2)$ , può essere calcolato massimizzando su tutto lo spazio dei parametri. Quindi  $\hat{\theta}_1$  e  $\hat{\sigma}_1^2$  sono gli ordinari stimatori di M.V. di  $\theta$  e  $\sigma^2$  calcolati al capitolo 2,

$$\hat{\theta}_1 = \text{Arg} \min_{\theta} \|y - S\theta\|^2 \quad \hat{\sigma}_1^2 = \frac{1}{N} \|y - S\hat{\theta}_1\|^2 = \frac{1}{N} R_1^2(y) \quad (7.4.2)$$

sostituendo (7.4.2) in (7.4.1) si trova

$$f(y, \hat{\theta}_1(y), \hat{\sigma}_1^2(y)) = \left[ 2\pi \frac{R_1^2(y)}{N} \right]^{-\frac{N}{2}} \exp -\frac{N}{2}. \quad (7.4.3)$$

*Sotto*  $H_0$ , lo stimatore  $\hat{\theta}_0$  risolve il problema di minimo vincolato (7.3.8) per cui possiamo rifarci alle considerazioni esposte al paragrafo precedente. Lo stimatore della varianza  $\sigma^2$ , si trova eseguendo la massimizzazione della funzione di verosimiglianza soggetta al vincolo  $H\theta = \beta_0$  dopo aver effettuato la massimizzazione vincolate rispetto a  $\theta$ . Con un procedimento ormai familiare, si trova

$$\hat{\sigma}_0^2(y) = \frac{1}{N} \|\bar{y} - S\hat{\theta}_0(y)\|^2 = \frac{1}{N} R_0^2(y) \quad (7.4.4)$$

Sostituendo nella funzione densità, si trova così

$$f(y, \hat{\theta}_0(y), \hat{\sigma}_0^2(y)) = \left[ 2\pi \frac{R_0^2(y)}{N} \right]^{-N/2} \exp(-N/2)$$

per cui,

$$L(y) = \left[ \frac{R_0^2(y)}{R_1^2(y)} \right]^{N/2} = \left[ \frac{R_0^2(y) - R_1^2(y)}{R_1^2(y)} + 1 \right]^{N/2} \quad (7.4.5)$$

e quindi  $L(y)$  è una funzione biunivoca del rapporto (7.3.14). Abbiamo così scoperto che il rapporto (7.3.14) è proprio la statistica prescritta dal test di massima verosimiglianza.

Dobbiamo ora vedere come effettivamente si può calcolare la differenza  $R_0^2(y) - R_1^2(y)$  e poi studiare la distribuzione del rapporto  $R_0^2(y) - R_1^2(y) / R_1^2(y)$ . Finora abbiamo studiato il problema di minimo vincolato (7.3.8) solo da un punto



di vista geometrico qualitativo senza però ottenere una soluzione esplicita. Consideriamo allo scopo la matrice  $H$  che definisce l'ipotesi  $H_0$  e definiamo un parametro  $k$ -dimensionale  $\beta$  ponendo

$$\beta := H\theta. \tag{7.4.6}$$

Lo stimatore di M.V. di  $\beta$  è

$$\hat{\beta}(\mathbf{y}) = H\hat{\theta}(\mathbf{y}) = H[S^\top S]^{-1}S^\top \mathbf{y} \tag{7.4.7}$$

e la sua varianza normalizzata è

$$\frac{1}{\sigma^2} \text{Var}(\hat{\beta}(\mathbf{y})) = H[S^\top S]^{-1}H^\top := D. \tag{7.4.8}$$

Notiamo che  $H_0$  afferma che il valore di  $\beta$  è la quantità prefissata  $\beta_0$ ; cioè  $H_0 := \{\beta = \beta_0\}$ . Quindi se riusciamo a trasformare il problema (7.3.8) in uno in cui le prime variabili  $(\theta_1 \dots \theta_k)$  sono sostituite da  $(\beta_1 \dots \beta_k)$  il vincolo  $H\theta = \beta_0$  si riduce semplicemente ad aver imposti valori noti e prefissati  $(\beta_{01}, \dots, \beta_{0k})$  alle prime  $k$  variabili e il problema diventa di minimo non vincolato nelle rimanenti  $p - k$ . Cerchiamo allora un cambiamento di base in  $\mathbb{R}^N$  nel quale le prime  $k$  equazioni del modello  $\mathbf{y} = S\theta + \sigma\mathbf{w}$  diventino del tipo  $\mathbf{z} = H\theta + \sigma\mathbf{e}$ , con  $\mathbf{z}$  vettore  $k$ -dimensionale. Questo si traduce nel cercare una matrice  $Q$  tale per cui:

$$QS\theta = H\theta, \quad \forall \theta \in \mathbb{R}^p$$

ovvero  $QS = H$ , dove ovviamente  $Q$  dovrà essere di dimensione  $k \times N$ . Il problema ha sicuramente soluzione (non necessariamente unica), dato che le righe di  $H$  sono  $k \leq p$  vettori di  $\mathbb{R}^p$  linearmente indipendenti che quindi stanno sempre dentro lo spazio righe di  $S$ , dato che quest'ultimo per ipotesi è  $\mathbb{R}^p$ . Proviamo a trovare una soluzione della forma  $Q = CS^\top$ , con  $C \in \mathbb{R}^{k \times p}$ . Dovrà essere  $CS^\top S = H$  e quindi,

$$C = H(S^\top S)^{-1} \quad Q = H(S^\top S)^{-1}S^\top. \tag{7.4.9}$$

Nella nuova base allora,

$$H(S^\top S)^{-1}S^\top \mathbf{y} = H\theta + \sigma H(S^\top S)^{-1}S^\top \mathbf{w} := \beta + \sigma\mathbf{e}, \tag{7.4.10}$$

da cui si ricavano il seguente utile risultato.

**Lemma 7.4.** Sotto  $H_0$  lo stimatore di M.V. del parametro  $\beta$  è dato dalla formula

$$\hat{\beta}(\mathbf{y}) = H\hat{\theta}(\mathbf{y}) = \beta_0 + \sigma\mathbf{e}, \quad \mathbf{e} \sim \mathcal{N}(0, D) \tag{7.4.11}$$

dove  $\mathbf{e} = H(S^\top S)^{-1}S^\top \mathbf{w}$ , che ha matrice varianza  $D$  definita in (7.4.8).

**Lemma 7.5.** Si ha

$$\|S\hat{\theta}(\mathbf{y}) - S\hat{\theta}_0(\mathbf{y})\|^2 = \|\hat{\beta}(\mathbf{y}) - \beta_0\|_{D^{-1}}^2 \tag{7.4.12}$$

Il primo membro è la differenza  $\|\hat{\mu}_1 - \hat{\mu}_0\|^2$  definita in (7.3.12).

**Proof.** Dobbiamo finalmente risolvere il problema di minimo vincolato (7.3.8). Allo scopo, introduciamo il moltiplicatore di Lagrange  $\lambda \in \mathbb{R}^k$  e consideriamo il problema

$$\min_{\theta} \{ \|y - S\theta\|^2 + \lambda^\top (H\theta - \beta_0) \}.$$

Calcolando il gradiente rispetto a  $\theta$  della funzione Lagrangiana si trova la condizione

$$-2S^\top(y - S\theta) + H^\top\lambda = 0$$

da cui si ricava per l'estremale l'espressione

$$\hat{\theta}_0(y) = [S^\top S]^{-1} S^\top y - \frac{1}{2} [S^\top S]^{-1} H^\top \lambda, \quad (*)$$

che dipende da  $\lambda$ . Il moltiplicatore si ricava dalla condizione di vincolo che deve valere per  $\hat{\theta}_0(y)$  e possiamo riscrivere come  $H\hat{\theta}_0(y) = \beta_0$ . Questa è equivalente alla

$$\frac{1}{2} H [S^\top S]^{-1} H^\top \lambda = H [S^\top S]^{-1} S^\top y - \beta_0$$

ovvero alla

$$\frac{1}{2} D \lambda = \hat{\beta}(y) - \beta_0.$$

Sostituendo in (\*) si trova

$$\hat{\theta}_0(y) = \hat{\theta}(y) - [S^\top S]^{-1} H^\top D^{-1} [\hat{\beta}(y) - \beta_0]$$

ovvero

$$S [\hat{\theta}(y) - \hat{\theta}_0(y)] = S [S^\top S]^{-1} H^\top D^{-1} [\hat{\beta}(y) - \beta_0]$$

che conduce immediatamente a quanto si voleva provare.  $\square$

**Theorem 7.2.** *La decomposizione (7.3.11) si può riscrivere nella forma*

$$R_0^2(\mathbf{y}) = \|\hat{\beta}(\mathbf{y}) - \beta_0\|_{D^{-1}}^2 + R_1^2(\mathbf{y}) \quad (7.4.13)$$

e i due membri della somma a secondo membro  $\|\hat{\beta}(\mathbf{y}) - \beta_0\|_{D^{-1}}^2$  e  $R_1^2(\mathbf{y})$ , sono variabili aleatorie indipendenti sotto entrambe le ipotesi.

**Proof.** Ricordiamo che  $\hat{\beta}(\mathbf{y}) = H\hat{\theta}(\mathbf{y})$  e  $R_1^2(\mathbf{y}) = \|\mathbf{y} - S\hat{\theta}(\mathbf{y})\|^2$ ; pertanto basta far vedere che  $\hat{\theta}(\mathbf{y})$  e  $\mathbf{y} - S\hat{\theta}(\mathbf{y})$  sono indipendenti. Usando il proiettore  $P = S(S^\top S)^{-1} S^\top$ , si trova

$$\begin{aligned} \text{Cov} [\hat{\theta}(\mathbf{y}), (\mathbf{y} - P\mathbf{y})] &= \mathbb{E} [\hat{\theta}(\mathbf{y}) (\mathbf{y} - S\hat{\theta}(\mathbf{y}))^\top] = \sigma(S^\top S)^{-1} S^\top \mathbb{E}(\mathbf{y}\mathbf{y}^\top) (I - P)^\top \\ &= \sigma^2 (S^\top S)^{-1} S^\top (I - P) \\ &= \sigma^2 [(S^\top S)^{-1} S^\top - (S^\top S)^{-1} S^\top S (S^\top S)^{-1} S^\top] = 0. \end{aligned}$$

□

Notiamo adesso che sotto  $H_0$  si ha  $\hat{\beta}(\mathbf{y}) \sim \mathcal{N}(\beta_0, \sigma^2 D)$  (Lemma 7.4), e quindi

$$\frac{R_0^2(\mathbf{y}) - R_1^2(\mathbf{y})}{\sigma^2} = \frac{1}{\sigma^2} \|\hat{\beta}(\mathbf{y}) - \beta_0\|_{D^{-1}}^2 \sim \chi^2(k) \quad (7.4.14)$$

Si può dimostrare che sotto  $H_1$  la (7.4.14) ha una distribuzione  $\chi^2$  non centrale,  $\chi^2(k, \delta)$ ;  $\delta$  essendo il parametro di non centralità<sup>30</sup>:

$$\delta = \frac{1}{\sigma^2} \|\hat{H}_0 \theta - \beta_0\|_{D^{-1}}^2.$$

Abbiamo poi ricordato piú volte che la somma dei residui "non vincolata"  $R_1^2(\mathbf{y})$  è distribuita come  $\chi^2(N - p)$ :

$$\frac{R_1^2(\mathbf{y})}{\sigma^2} \approx \chi^2(N - p). \quad (7.4.15)$$

Possiamo adesso occuparci della distribuzione di probabilità del rapporto (7.4.5).

### 7.4.1 La distribuzione $F$

Siano  $\mathbf{x}_1 \sim \chi^2(n_1)$  e  $\mathbf{x}_2 \sim \chi^2(n_2)$ , variabili indipendenti. Allora il rapporto

$$\mathbf{z} := \frac{\mathbf{x}_1/n_1}{\mathbf{x}_2/n_2} \quad (7.4.16)$$

è distribuito secondo la densità di probabilità

$$p_{n_1, n_2}(z) = \left[ \frac{\Gamma(\frac{n_1 + n_2}{2})}{\Gamma(\frac{n_1}{2}) + \Gamma(\frac{n_2}{2})} \right] \left( \frac{n_1}{n_2} \right)^{\frac{n_1}{2}} \frac{z^{\frac{n_1}{2} - 1}}{\left( 1 + \frac{n_1}{n_2} z \right)^{\frac{n_1 + n_2}{2}}} \quad z \in \mathbb{R}_+ \quad (7.4.17)$$

che si chiama *distribuzione  $F$  di Snedecor* a  $n_1$  e  $n_2$  gradi di libertà e si denota col simbolo  $\mathcal{F}(n_1, n_2)$ .

Per la dimostrazione di questa formula si può ancora fare riferimento al testo di Hogg and Craig [26]. Dopo la Gaussiana, la distribuzione  $F$  è forse una delle distribuzioni più importanti della statistica classica. Essa si trova tabulata in varie forme in letteratura. Per  $n_1 = 1$  essa è la distribuzione di probabilità di  $t^2$ , la variabile di Student a  $n_2$  gradi di libertà, elevata al quadrato.

La media  $\mu_1$  e la moda,  $m$ , di  $\mathcal{F}(n_1, n_2)$  esistono se  $n_1$  e  $n_2$  sono strettamente maggiori di 1 e valgono:

$$\mu_1 = \frac{n_2}{n_2 - 2}, \quad m = \frac{n_2(n_1 - 2)}{n_1(n_2 + 2)}$$

<sup>30</sup>Per la definizione e i dettagli il riferimento standard è il classico libro di Scheffè [56]

Si dimostra che

$$L - \lim_{n_2 \rightarrow \infty} n_1 \mathbf{z} = \chi^2(n_1), \tag{7.4.18}$$

il limite essendo in distribuzione (cf. Il capitolo 5). Inoltre se  $\mathbf{z} \sim \mathcal{F}(n_1, n_2)$  e  $a := a(n_1, n_2)$  è definito dalla

$$\mathbb{P}(\mathbf{z} \geq a) = \alpha$$

il valore di  $b$  per cui

$$\mathbb{P}(\mathbf{z} \leq b) = \alpha$$

è uguale al reciproco di  $a$  calcolato in base alla distribuzione  $\mathcal{F}(n_1, n_2)$  in cui i gradi di libertà sono scambiati; i.e.

$$b(n_1, n_2) = a(n_2, n_1).$$

Vedere ad esempio il sito web <http://econtools.com/jevons/java/Graphics2D/FDist.html>

**Theorem 7.3.** Sotto  $H_0$ , il rapporto

$$\mathbf{z} := \frac{(N - p)}{k} \frac{\|\hat{\beta}(\mathbf{y}) - \beta_0\|_{D^{-1}}^2}{R_1^2(\mathbf{y})} \tag{7.4.19}$$

è distribuito secondo la distribuzione  $F$  di Snedecor,  $\mathcal{F}(k, N - k)$ . La regione critica del test si può quindi scrivere

$$C := \{y; \mathbf{z} \geq k_\alpha\} \tag{7.4.20}$$

dove  $k_\alpha$  corrisponde al valore assegnato di  $\alpha$  secondo la distribuzione  $F$ .

Il test basato sul rapporto (7.4.19) si chiama **test F** ed è di impiego molto generale.

Qualche volta il calcolo di  $R_0^2$  è semplice e conviene calcolare il numeratore di  $F$  senza passare attraverso l'espressione dello stimatore  $\hat{\beta}$ .

## 7.5 Applicazione all'analisi della varianza

Supponiamo di avere  $p$  campioni di numerosità  $N_1, \dots, N_p$  estratti da  $p$  popolazioni  $\mathcal{N}(\mu_i, \sigma^2)$  con  $i = 1, \dots, p$ , in cui le medie  $\mu_1, \dots, \mu_p$  e la varianza (che è la stessa) sono ignote. Ad esempio potrebbe trattarsi di  $N_1, \dots, N_p$  misure di una resistenza fatte su  $p$  ponti di Wheatstone fisicamente diversi, ma aventi le stesse caratteristiche di precisione. Indicando con  $\mathbf{y}_i$  l' $i$ -esimo blocco di misure e con  $\theta = [\mu_1, \dots, \mu_p]^\top$  il vettore  $p$  dimensionale delle medie incognite, si può usare il modello lineare:

$$\mathbf{y} := \begin{bmatrix} \mathbf{y}_1 \\ \cdot \\ \cdot \\ \mathbf{y}_p \end{bmatrix} = \begin{bmatrix} e_{N_1} & 0 & 0 & 0 \\ 0 & \cdot & 0 & 0 \\ 0 & 0 & \cdot & 0 \\ 0 & 0 & 0 & e_{N_p} \end{bmatrix} \theta + \sigma \mathbf{w} \tag{7.5.1}$$

dove  $e_{N_i} = [1 \dots 1]^\top \in \mathbb{R}^{N_i}$  e  $\mathbf{w} \sim \mathcal{N}(0, I_N)$ , è rumore bianco  $N_1 + \dots + N_p = N$  dimensionale.

Si vuole verificare l'ipotesi:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_p \quad (7.5.2)$$

contro l'alternativa che le medie siano diverse. Denotiamo con  $\{y_{it}; i = 1, \dots, p; t = 1, \dots, N_i\}$  le  $p$  stringhe di osservazioni ottenute nei  $p$  procedimenti di misura e calcoliamo la somma dei residui sotto  $H_1$  e sotto  $H_0$ . Quando la (7.5.2) non vale (i.e. sotto  $H_1$ ), si ha:

$$R_1^2(y) = \min_{\mu_1, \dots, \mu_p} \sum_{i=1}^p \sum_{t=1}^{N_i} (y_{it} - \mu_i)^2 = \sum_{i=1}^p \min_{\mu_i} \sum_{t=1}^{N_i} (y_{it} - \mu_i)^2 \quad (7.5.3)$$

e siccome il valore di  $\mu_i$  che dà il minimo di ciascuna somma è:  $\hat{\mu}_i = \bar{y}_{N_i} = \frac{1}{N_i} \sum_{t=1}^{N_i} y_{it}$ , si ottiene

$$R_1^2(y) = \sum_{i=1}^p \sum_{t=1}^{N_i} (y_{it} - \bar{y}_{N_i})^2 = \sum_{i=1}^p s_{N_i}^2,$$

che è distribuita come  $\sigma^2 \chi^2(N - p)$ . Sotto  $H_0$ , la media è la stessa e quindi

$$R_0^2(y) = \min_{\mu} \sum_{i=1}^p \sum_{t=1}^{N_i} (y_{it} - \mu)^2 = \sum_{i=1}^p \sum_{t=1}^{N_i} (y_{it} - \bar{y}_N)^2,$$

che sappiamo essere distribuita come  $\sigma^2 \chi^2(N - 1)$ . La differenza  $R_0^2 - R_1^2$  si può calcolare usando l'identità

$$\sum_{t=1}^{N_i} (y_{it} - \bar{y}_N)^2 = \sum_{t=1}^{N_i} [y_{it} - \bar{y}_{N_i} + (\bar{y}_{N_i} - \bar{y}_N)]^2 = \sum_{t=1}^{N_i} (y_{it} - \bar{y}_{N_i})^2 + N_i (\bar{y}_{N_i} - \bar{y}_N)^2$$

che permette di scrivere

$$R_0^2 - R_1^2 = \sum_1^p N_i (\bar{y}_{N_i} - \bar{y}_N)^2 \quad (7.5.4)$$

che è una somma pesata delle deviazioni delle medie per i singoli gruppi, dalla media totale  $\bar{y}_N$ . Notiamo che in questo problema  $k = p - 1$  dato che la (7.5.2) si può riscrivere

$$\mu_1 - \mu_2 = 0; \dots; \mu_1 - \mu_p = 0,$$

e queste sono  $p - 1$  equazioni indipendenti.

**Example 7.5.** Supponiamo di avere tre serie di misure indipendenti estratte da tre distribuzioni Gaussiane di ugual varianza con statistiche descritte nella tabella 7.1 seguente e di voler verificare l'ipotesi ( $H_0$ ) che le medie delle tre distribuzioni siano uguali.

| Serie | $N_i$ | $\sum_t y_{it}$ | $\bar{y}_{N_i}$ |
|-------|-------|-----------------|-----------------|
| 1     | 83    | 11.227          | 135.87          |
| 2     | 51    | 7.049           | 138.22          |
| 3     | 8     | 1.102           | 137.75          |

Table 7.1.

Con semplici calcoli si trova  $R_0^2 = 4616.64$  e  $R_0^2 - R_1^2 = 238.59$ . Nel nostro caso  $k = p - 1 = 2$  e  $N - p = 142 - 3 = 139$  per cui

$$F = \frac{139}{2} \frac{238.59}{4616.64 + 238.59} = 3.79.$$

Andando nella tabella di  $F(2, 139)$  con vari valori di  $\alpha$  si trovano i valori critici:

| $\alpha$   | 0.10 | 0.05 | 0.025 | 0.01 |
|------------|------|------|-------|------|
| $k_\alpha$ | 2.30 | 3.00 | 3.70  | 4.65 |

per cui ci si trova nella regione critica a meno di non scegliere una probabilità d'errore di prima specie molto piccola. A questo punto ci si chiede se le medie

$\mu_i$  delle tre popolazioni sono significativamente diverse oppure no. La risposta è che le medie si possono considerare diverse (si rifiuta  $H_0$ ), se non si richiede a questa affermazione una certezza statistica troppo elevata. Diciamo che *con probabilità leggermente maggiore del 97,5 per cento le medie sono da considerarsi diverse*. Non c'è invece evidenza sufficiente per fare la stessa affermazione con certezza statistica del 99 per cento. Se si volesse una certezza statistica del 99 per cento di non commettere errori rifiutando l'ipotesi, ci si troverebbe nella regione di accettazione e quindi si dovrebbe decidere di accettare  $H_0$ . In questo caso però la probabilità  $\alpha$  non direbbe nulla (o quasi) sulla correttezza della scelta quando vale  $H_1$ . Occorrerebbe a questo scopo effettuare il calcolo della probabilità dell'errore di seconda specie,  $\beta$ , di accettare  $H_0$  quando invece vale  $H_1$ , il che è normalmente complicato perchè  $H_1$  è un'ipotesi composta. La stessa difficoltà si incontra ovviamente per il calcolo della potenza del test. È bene tener presente che di norma,  $\beta$  cresce al diminuire di  $\alpha$  (vedere ad esempio la figure 7.1) per cui la decisione di accettare  $H_0$  con  $\alpha$  molto piccoli si rivela in generale priva di senso, dato che questo può comportare valori elevati di  $\beta$ , anche prossimi ad  $1 - \alpha$ , vedere [60].  $\diamond$

Sotto  $H_1$  il rapporto  $F$  definito in (7.4.19) non è più distribuito come  $F(k, N - p)$ , ma bensì come una  $F$  non centrale dipendente dal cosiddetto parametro di non cen-

tralità  $\lambda$ , dato dalla:

$$\lambda^2 = \lambda^2(\theta, \sigma^2) = \frac{1}{\sigma^2} \|\beta - \beta_0\|_{D^{-1}}^2 = \frac{1}{\sigma^2} \|H\theta - \beta_0\|_{D^{-1}}^2. \quad (7.5.5)$$

La  $F$  non centrale può essere approssimata con una centrale. Una approssimazione sufficiente in molti casi si ottiene dalla relazione (che ovviamente va intesa tra variabili casuali)

$$F(n_1, n_2, \lambda) \cong \frac{n_1 + \lambda}{n_1} F(n_1^*, n_2) \quad (7.5.6)$$

con  $n_1^*$  dato da:

$$n_1^* = (n_1 + \lambda)^2 / (n_1 + 2\lambda). \quad (7.5.7)$$

In questo modo la potenza si può calcolare usando le tavole di  $F(n_1^*, n_2)$ . Chiaramente  $n_1^*$  non è più un intero e si usa l'intero più prossimo (ovvero si interpola). La potenza corrispondente a  $\lambda = \lambda(\theta, \sigma^2)$  è allora:

$$1 - \beta(\theta, \sigma^2) = \int_{\frac{n_1 + \lambda}{n_1 + 2\lambda} a_\alpha}^{\infty} dF(n_1^*, n_2) \quad (7.5.8)$$

si vede che ponendo  $\lambda = 0$  (ovvero  $\beta = \beta_0$ ) in questa formula si ottiene  $\alpha$ . Bisogna però ricorrere a tavole complete della distribuzione  $F$  (in cui  $F$  è calcolata per valori grandi di  $\alpha$ ).

**Example 7.6 (Rao pag.227).** Si vuole stimare la capacità craniale  $C$  come funzione di 3 dimensioni lineari,  $L, B, H$ , mediante una formole del tipo

$$C \cong \alpha L^{\theta_1} B^{\theta_2} H^{\theta_3} \quad (7.5.9)$$

disponendo di una serie di 86 misure di  $C, L, B, H$ . Passando ai logaritmi e definendo:

$$y = \log C, \quad x_1 = \log L, \quad x_2 = \log B, \quad x_3 = \log H, \quad \theta_0 : \log \alpha,$$

la (9.92) si riscrive

$$y \simeq \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$

ovvero

$$y_t = \theta_0 + \theta_1 x_{1t} + \theta_2 x_{2t} + \theta_3 x_{3t} + \epsilon_t \quad (7.5.10)$$

$t = 1, \dots, 86$ . Supporremo gli errori  $\epsilon_t$  Gaussiani indipendenti a media nulla e di uguale varianza incognita  $\sigma^2$ . Il problema è dunque ridotto ad un problema di regressione lineare. Siano

$$\bar{x}_i = \frac{1}{86} \sum_1^{86} x_{it}, \quad i = 1, 2, 3$$

le medie campionarie delle variabili  $x_i, i = 1, 2, 3$ . Conviene sottrarre alla (7.5.10) l'equazione per le medie campionarie

$$\bar{y} = \theta_0 + \theta_1 \bar{x}_1 + \theta_2 \bar{x}_2 + \theta_3 \bar{x}_3 + \bar{\epsilon}$$

ricavando

$$y_t - \bar{y} = \sum_1^3 \theta_i (x_{it} - \bar{x}_i) + (\epsilon_t - \bar{\epsilon}). \quad (7.5.11)$$

In questo modo si riducono i parametri a 3 e una volta ottenuta la stima  $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ ;  $\hat{\theta}_0$  si ricava dalla

$$\hat{\theta}_0 = \bar{y} - (\hat{\theta}_1 \bar{x}_1 + \hat{\theta}_2 \bar{x}_2 + \hat{\theta}_3 \bar{x}_3) \quad (7.5.12)$$

(Lo studente verifichi che se nel modello lineare  $y = S\theta + \epsilon$  la prima colonna è tutta di 1, lo stimatore della prima componente di  $\theta$  ha proprio l'espressione (7.5.13)). Riscriviamo la (7.5.11) in forma vettoriale

$$\Delta \mathbf{y} = S\theta + \sigma \mathbf{w}, \quad \mathbf{w} \sim \mathcal{N}(0, I).$$

Le medie campionarie tratte dalle 86 misure sono:

$$\bar{y} = 3.17; \quad \bar{x}_1 = 2.275; \quad \bar{x}_2 = 2.15; \quad \bar{x}_3 = 2.11$$

Inoltre si ha

$$S^T S = \begin{bmatrix} 0.0187 & 0.0085 & 0.0068 \\ 0.0085 & 0.029 & 0.0088 \\ 0.0068 & 0.0088 & 0.029 \end{bmatrix} \quad S^T \Delta y = \begin{bmatrix} 0.030 \\ 0.044 \\ 0.036 \end{bmatrix}$$

Calcolando l'inversa, si trova

$$[S^T S]^{-1} = \begin{bmatrix} 64.21 & -15.57 & -10.49 \\ -15.57 & 41.71 & -9.00 \\ -10.49 & -9.00 & 39.88 \end{bmatrix}$$

da cui  $\hat{\theta} = [S^T S]^{-1} S^T \Delta y$ , ha i valori numerici

$$\hat{\theta}_1 = 0.88, \quad \hat{\theta}_2 = 1.04, \quad \hat{\theta}_3 = 0.73$$

e dalla (7.5.13)

$$\hat{\theta}_0 = -2.618$$

per cui la formula stimata è

$$C = 0.00241 L^{0.88} B^{1.04} H^{0.73}.$$

La somma dei quadrati dei residui si calcola con la formula

$$R_1^2 = \|\Delta y\|^2 - \|S\hat{\theta}\|^2 = \|\Delta y\|^2 - \langle S\hat{\theta}, \Delta y \rangle = \|\Delta y\|^2 - \hat{\theta}^T S^T \Delta y \quad (7.5.13)$$

ovvero

$$R_1^2 = \sum_{t=1}^{86} (y_t - \bar{y})^2 - (\hat{\theta}_1 0.030 + \hat{\theta}_2 0.044 + \hat{\theta}_3 0.036) \quad (7.5.14)$$

$$= 0.127 - 0.099 = 0.028 \quad (7.5.15)$$



Per ottenere una stima corretta della varianza usiamo la formula:

$$\hat{\sigma}^2 = \frac{R_1^2}{N-4} = \frac{0.028}{82} = 0.00034$$

notiamo che il numero di parametri  $\theta_i$  incogniti è sempre 4 anche se abbiamo usato il trucco di ridurre la dimensione del problema a 3. La matrice di varianze e covarianze di  $(\theta_1 \theta_2 \theta_3)$  si ottiene moltiplicando  $[S^T S]^{-1}$  per  $\sigma^2$ . Ad esempio, si trova

$$\text{var } \hat{\theta}_1 = 64.21 \times 3.4 \cdot 10^{-4} \cong 220 \cdot 10^{-4} = 0.022.$$

L'espressione per la varianza di  $\hat{\theta}_0$  si può ricavare dalla (7.5.13). Lasciamo i facili calcoli al lettore.

Questi calcoli completano la fase di stima del modello. Occorre adesso passare alla fase di validazione della stima.

## A. Test di adeguatezza del modello

Verifichiamo l'ipotesi

$$H_0 : \theta_1 = \theta_2 = \theta_3 = 0$$

(Non ha molto senso verificare anche  $\theta_0 = 0$ , cioè  $\alpha = 1$  nella (7.5.9), dato che questo corrisponderebbe a verificare  $C = 1 + \text{"rumore"}$ . Quest'ultima è evidentemente l'ipotesi di non accoppiamento tra le variabili L, B, H e la variabile C che si vuole descrivere).

Per verificare  $H_0$  col test F basta procurarsi  $R_0^2 - R_1^2 = \|\hat{\beta} - \beta_0\|_{D^{-1}}^2 = \|\hat{\beta}\|_{D^{-1}}^2$ , dato che in questo caso  $\beta = [\theta_1 \theta_2 \theta_3]^T \equiv \theta$  e  $\beta_0 = 0$ . Allora

$$D = [S^T S]^{-1}$$

e quindi

$$\|\hat{\beta}\|_{D^{-1}}^2 = \|\hat{\theta}\|_{D^{-1}}^2 = \hat{\theta}^T S^T S \hat{\theta} = \hat{\theta}^T S^T \Delta y$$

nell'ultimo passaggio si è sfruttata l'ortogonalità  $S\hat{\theta} \perp \Delta y - S\theta$  per cui  $\langle S\hat{\theta}, S\hat{\theta} \rangle = \langle S\hat{\theta}, \Delta y \rangle$ . Ne segue che  $R_0^2 - R_1^2$  è semplicemente l'ultimo addendo nella (7.5.13). A titolo di verifica notiamo che sotto  $H_0$ ,

$$R_0^2 = \min_{\theta_0} \|\Delta y - S\theta\|^2 = \|\Delta y\|^2$$

che è proprio il primo addendo in (7.5.13).

Ne consegue che:

$$F_A = \frac{N-p}{k} \frac{R_0^2 - R_1^2}{R_1^2} = \frac{86-4}{3} \frac{0.099}{0.028} = 97.4$$

Sulla tabella di  $F(3, 82)$  si trovano i seguenti valori critici  $k_\alpha$

il che porta sempre a rifiutare l'ipotesi (A), anche se si prende  $\alpha$  estremamente piccolo.

|            |      |      |       |      |       |
|------------|------|------|-------|------|-------|
| $\alpha$   | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 |
| $k_\alpha$ | 2.15 | 2.70 | 3.90  | 4.00 | 4.60  |

## B. uguaglianza degli esponenti

Verifichiamo l'ipotesi

$$H_0 = \theta_1 = \theta_2 = \theta_3$$

che in questo caso equivale a  $H\theta = 0$  ovvero

$$\beta := \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \theta = 0$$

e si potrebbe usare la formula  $\|\hat{\beta}\|_{D^{-1}}^2 = R_0^2 - R_1^2$ , come fatto in precedenza. Però è più conveniente calcolare direttamente  $R_0^2$ . Ponendo  $\theta_1 = \theta_2 = \theta_3 = \eta$  si ha

$$\theta = \mathbf{1}\eta; \quad \mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

e, dato che  $\Delta y = (S\mathbf{1})\eta + \sigma w$ , le equazioni normali diventano

$$(\mathbf{1}^\top S^\top S \mathbf{1})\eta = \mathbf{1}^\top S^\top \Delta y.$$

Per calcolare  $\hat{\eta}$  basta quindi sommare gli elementi di  $S^\top S$  e di  $S^\top \Delta y$ , ottenendo

$$0.125\eta = 0.11 \Rightarrow \hat{\eta} = 0.887$$

per cui

$$R_0^2 = \|\Delta y - S\mathbf{1}\hat{\eta}\|^2 = \|\Delta y\|^2 - \hat{\eta}(\mathbf{1}^\top S^\top \Delta y)$$

e usando la (7.5.13),

$$R_0^2 - R_1^2 = 0.099 - \hat{\eta} \cdot 0.11 = 0.099 - 0.098 = 0.001.$$

Si trova così

$$F_B = \frac{N-p}{k} \frac{0.001}{0.028} = \frac{84}{2} \frac{0.001}{0.028} = 1.4$$

Per  $\alpha = 0.10$ , il punto critico di  $F(2, 82)$  è 2.77 per cui siamo nella regione di accettazione. Non c'è quindi evidenza sperimentale per considerare  $\theta_1$ ,  $\theta_2$  e  $\theta_3$  tra loro diversi.

## C. Esponenti tutti uguali a 1

Vogliamo ora considerare l'ipotesi

$$H_0 = \theta_1 = \theta_2 = \theta_3 = 1$$

equivalente a

$$I\theta = \mathbf{1}.$$

Si ha allora

$$R_0^2 - R_1^2 = \|\hat{\theta} - \mathbf{1}\|_{[S^\top S]}^2 = (\hat{\theta} - \mathbf{1})^\top (S^\top S)(\hat{\theta} - \mathbf{1}) = 0.0026$$

In questo caso  $k = 3$  e  $F_C = 2.5$ . La tabella dei valori critici della distribuzione è

|  |      |      |       |
|--|------|------|-------|
|  | 0.10 | 0.05 | 0.025 |
|  | 2.15 | 2.70 | 3.30  |

L'ipotesi si accetta con  $\alpha = 0.05$  e si rifiuta con  $\alpha = 0.10$ . Il caso è un poco dubbio.

### D. La somma degli esponenti è uguale a 3

Vogliamo verificare se

$$\theta_1 + \theta_2 + \theta_3 = 3$$

Detto  $\beta := \theta_1 + \theta_2 + \theta_3$ , si ha  $\beta_0 = 3$  e

$$\hat{\beta} = \hat{\theta}_1 + \hat{\theta}_2 + \hat{\theta}_3 = 2.65$$

mentre  $D$  è semplicemente la somma degli elementi di  $[S^\top S]^{-1}$ ; i.e.  $D = 75.7$ , per cui

$$\|\hat{\beta} - 3\|^2/D = \frac{(2.65 - 3)^2}{75 \cdot 7} = \frac{(0.35)^2}{75.7}$$

e si trova  $F_D = 4.72$ . I valori critici di  $F(1, 82)$  sono

|                 |      |      |       |      |
|-----------------|------|------|-------|------|
| $\alpha$        | 0.10 | 0.05 | 0.025 | 0.01 |
| $\alpha_\alpha$ | 2.77 | 3.96 | 5.20  | 6.95 |

e con  $\alpha = 0.05$   $F_D$  è nella regione critica. È ragionevole rifiutare  $H_0$ .

### E. L'esponente di $H$ è uguale a zero

Vediamo se la variabile  $H$  (quella che ha l'esponente più piccolo) è significativa. Saggiamo l'ipotesi:

$$\theta_3 = 0$$

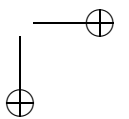
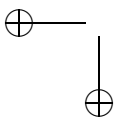
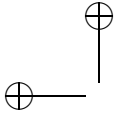
In questo caso  $\hat{\beta} = \hat{\theta}_3 = 0.73$ ,  $D = 39.88$  e si calcola subito

$$\|\hat{\beta} - \beta_0\|_{D^{-1}}^2 = \frac{(0.73)^2}{39.9} = 0.0135$$

per cui

$$F_E = \frac{0.0135}{3.4 \cdot 10^{-4}} = 39.72$$

Dalla tabella dei valori critici di  $F(1, 82)$  si vede che  $F_E$  cade nella regione critica anche per valori molto piccoli di  $\alpha$ . Pertanto si rifiuta l'ipotesi  $\theta_3 = 0$ .  $\diamond$



## CHAPTER 8

# MULTIPLE REGRESSION AND COMPLEXITY ESTIMATION

## 8.1 Stima della complessità di un modello lineare

In molte circostanze che si incontrano in pratica 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

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

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 deteriora all'aumentare di  $p$ . Al limite, per  $p$  molto grande, il "fit" perfetto ottenuto usando un elevatissimo numero di parametri è in pratica di nessuna utilità dato che la grande varianza delle stime 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 alternative possibili.

Consideriamo due modelli lineari Gaussiani in forma standard

$$\begin{aligned} M_1 : \quad \mathbf{y} &= S_1 \theta_1 + \epsilon & \theta_1 &\in \mathbb{R}^p \\ M_2 : \quad \mathbf{y} &= [S_1 \ S_2] \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \epsilon & \theta_2 &\in \mathbb{R}^k. \end{aligned} \quad (8.1.1)$$

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

Nel modello più “semplice”  $M_1$ , supporremo, come sempre, che  $\text{rango } S_1 = p$ . 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

$$\text{rango } [S_1 \ S_2] = p + k \quad . \quad (8.1.2)$$

Se ciò non accade, il modello può facilmente essere riparametrizzato eliminando le colonne di  $S_2$  che sono linearmente dipendenti e ridefinendo opportunamente  $\theta_2$ .

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 (8.1.1) definiscono due diverse famiglie parametriche di misure di probabilità sullo spazio campionario.

**Problem 8.1.** 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”. Come abbiamo visto e come è spiegato in letteratura, ad esempio in [32, 55], le funzioni di decisione ottimali (che massimizzano asintoticamente la potenza del test) si ottengono considerando il cosiddetto *rapporto di massima verosimiglianza* la cui costruzione richiede preliminarmente la stima (di M.V.) dei parametri dei due modelli. Per questo motivo inizieremo lo studio del problema 8.1 mettendo in relazione gli stimatori di M.V. dei parametri nei due modelli  $M_1$  e  $M_2$ .

**Remark 8.1.** 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 (8.1.1) 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.

## 8.2 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  $\theta_1$  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  $\theta$  il parametro  $p + k$  dimensionale  $[\theta_1^\top \ \theta_2^\top]^\top$  che compare nella (8.1.1). Naturalmente la stima ai M.Q. (di Markov) di  $\theta$  è definita dalle solite formule,

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

$$\text{Var } \hat{\theta} = \sigma^2 (S^\top S)^{-1} \quad ,$$

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  $\theta_1$  relativa al modello di ordine  $p$  per effetto dell'aggiunta dei  $k$  ulteriori parametri.

Per la (8.1.2)  $\mathcal{S}$  si può decomporre in somma diretta

$$\text{span } [S] = \text{span } [S_1 \ S_2] = \mathcal{S}_1 \oplus \mathcal{S}_2 = \text{span } [S_1] \oplus \text{span } [S_2] \quad (8.2.1)$$

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

$$P_1 : \mathbb{R}^N \rightarrow \mathcal{S}_1 \quad , \quad P_1 = S_1 (S_1^\top S_1)^{-1} S_1^\top \quad , \quad (8.2.2)$$

$$P_1^\perp : \mathbb{R}^N \rightarrow \mathcal{S}_1^\perp \quad , \quad P_1^\perp = I - S_1 (S_1^\top S_1)^{-1} S_1^\top \quad .$$

Per semplificare le notazioni in seguito denoteremo con  $Q_1$  la matrice  $P_1^\perp$ . Dato che  $P_1 + Q_1 = I$ , si ha

$$S_2 = P_1 S_2 + Q_1 S_2$$

e siccome le colonne di  $P_1 S_2$  stanno per definizione in  $\mathcal{S}_1$ , l'ultimo addendo della (8.2.1) può venire sostituito da  $\text{span } [Q_1 \ S_2]$ . Quindi

$$\text{span } [S] = \text{span } [S_1] \overset{\perp}{\oplus} \text{span } [Q_1 \ S_2] \quad (8.2.3)$$

dove il simbolo  $\overset{\perp}{\oplus}$  sta per somma diretta ortogonale. 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  $\hat{y}$  nella forma

$$\hat{y} = S_1 \hat{\theta}_1 + S_2 \hat{\theta}_2 \quad , \quad (8.2.4)$$

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  $\theta_1$  e  $\theta_2$  nel modello a  $p + k$  parametri.

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

$$y - \hat{y} \perp S_1 \quad , \quad y - \hat{y} \perp Q_1 S_2 \quad ,$$

che si riscrivono

$$S_1^\top (y - S_1 \hat{\theta}_1 - S_2 \hat{\theta}_2) = 0 \quad , \quad (8.2.5)$$

$$S_2^\top Q_1 (y - S_1 \hat{\theta}_1 - S_2 \hat{\theta}_2) = 0 \quad . \quad (8.2.6)$$

Queste formule forniscono subito

$$\hat{\theta}_1 = (S_1^\top S_1)^{-1} S_1^\top [y - S_2 \hat{\theta}_2] \quad , \quad (8.2.7)$$

$$\hat{\theta}_2 = (S_2^\top Q_1 S_2)^{-1} S_2^\top Q_1 y \quad . \quad (8.2.8)$$

La prova che  $S_2^\top Q_1 S_2$  è invertibile si ottiene facilmente se si tiene presente che  $Q_1$  è un proiettore. In effetti

$$a^\top S_2^\top Q_1 S_2 a = 0 \Rightarrow a^\top S_2^\top Q_1^\top Q_1 S_2 a = \|Q_1 S_2 a\|^2 = 0$$

e pertanto  $S_2 a$  deve stare nello spazio nullo di  $Q_1 = P_1^\perp$ . Dato che  $\text{Ker}(P_1^\perp) = \text{Im}(P_1) = S_1 = \text{span}[S_1]$ , segue che  $S_2 a \in \text{span}[S_1]$ , ma questo può accadere solo se  $a = 0$ , dato che le colonne di  $S_1$  ed  $S_2$  sono indipendenti.

Se indichiamo con il simbolo  $\bar{\theta}_1$  la stima di  $\theta_1$  ottenuta descrivendo i dati con un modello lineare a  $p$  parametri del tipo  $M_1$ , la (8.2.7) può essere riscritta come

$$\hat{\theta}_1 = \bar{\theta}_1 - (S_1^\top S_1)^{-1} S_1^\top S_2 \hat{\theta}_2 \quad . \quad (8.2.9)$$

che esprime la stima di  $\theta_1$  ottenuta con il modello lineare a  $p+k$  parametri, come la somma di  $\bar{\theta}_1$  e di un termine di correzione dovuto all'introduzione del parametro ulteriore  $\theta_2$ .

### Interpretazione geometrica: Proiezioni oblique

Nella decomposizione (8.2.4) 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  $S_1$  lungo  $S_2$  e di  $y$  su  $S_2$  lungo  $S_1$ .

Dalla formula (8.2.8) si vede in particolare che  $\hat{\theta}_2$  si può ricavare dalla relazione di ortogonalità

$$Q_1 y - S_2 \hat{\theta}_2 \perp Q_1 S_2$$

di modo che la proiezione obliqua di  $y$  su  $S_2$  lungo  $S_1$ , si può *calcolare* facendo prima la proiezione *ortogonale* di  $Q_1 y = y - P_1 y$  sul sottospazio  $(I - P_1)S_2 = Q_1 S_2$  (che è calcolabile risolvendo un problema di minimi quadrati ordinari) e poi moltiplicando per  $S_2$  il parametro  $\hat{\theta}_2$  trovato in questo modo<sup>31</sup>. La matrice di *proiezione obliqua su  $S_2$  lungo  $S_1$*  ha così la rappresentazione

$$P_{2\parallel 1} := S_2 (S_2^\top Q_1 S_2)^{-1} S_2^\top Q_1 \quad (8.2.10)$$

<sup>31</sup>Per evitare interpretazioni errate notiamo che  $S_2 \hat{\theta}_2$  non può essere la proiezione ortogonale di  $Q_1 y = y - P_1 y$  sul sottospazio  $Q_1 S_2$ . In effetti quest'ultimo non è nemmeno un sottospazio di  $S_2$ .



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

$$P_{2\parallel 1}^\top Q_1 = Q_1 P_{2\parallel 1}.$$

la quale, visto che  $Q_1$  è un proiettore ortogonale e quindi  $Q_1 = Q_1^\top$ , si può riscrivere come  $(Q_1 P_{2\parallel 1})^\top = P_{2\parallel 1}^\top Q_1^\top = Q_1 P_{2\parallel 1}$ , i.e.  $Q_1 P_{2\parallel 1}$  è simmetrica (e idempotente) e quindi è essa stessa un *proiettore ortogonale* che, per forza di cose, deve proiettare sul sottospazio  $Q_1 S_2$ , che è il complemento ortogonale di  $S_1$  in  $S$ . Infatti:

**Proposition 8.1.** *Sia  $P$  la matrice proiezione ortogonale da  $\mathbb{R}^N$  sullo spazio  $S$  e  $P_1$  quella sul sottospazio  $S_1 \subset S$ . Allora  $P - P_1$  è il proiettore ortogonale che proietta sul complemento ortogonale  $S \cap S_1^\perp$  e che ha la rappresentazione*

$$P - P_1 = Q_1 P_{2\parallel 1} \tag{8.2.11}$$

dove  $P_{2\parallel 1}$  è il proiettore obliquo definito in (8.2.10).

**Proof.** Basta dimostrare la (8.2.11). Usando le formule (8.2.2) e (8.2.7) si ottiene

$$\begin{aligned} \hat{y} = Py &= S_1(S_1^\top S_1)^{-1} S_1^\top y - S_1(S_1^\top S_1)^{-1} S_1^\top S_2 \hat{\theta}_2(y) + S_2 \hat{\theta}_2(y) \\ &= P_1 y + [I - S_1(S_1^\top S_1)^{-1} S_1^\top] S_2 \hat{\theta}_2(y) \\ &= (P_1 + Q_1 P_{2\parallel 1}) y \end{aligned}$$

per cui effettivamente si ha  $P - P_1 = Q_1 P_{2\parallel 1}$ . La decomposizione  $P = P_1 + Q_1 P_{2\parallel 1}$  è ovviamente ortogonale, stante che  $P_1^\top (P - P_1) = P_1 Q_1 P_{2\parallel 1} = 0$ . Notiamo che un'affermazione equivalente è la  $S = P_1 S \oplus S \cap S_1^\perp$ .  $\square$

**Problem 8.2.** Verificare che  $P_{2\parallel 1}$  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  $S_1$  lungo  $S_2$  e arrivare ad una rappresentazione esplicita della decomposizione (8.2.4), del tipo

$$y = P_{1\parallel 2} y + P_{2\parallel 1} y = S_1(S_1^\top Q_2 S_1)^{-1} S_1^\top Q_2 y + S_2(S_2^\top Q_1 S_2)^{-1} S_2^\top Q_1 y \tag{8.2.12}$$

dove  $Q_2$  ha un significato duale a  $Q_1$ . Questa espressione è forse più semplice della decomposizione ortogonale che abbiamo illustrato sopra ma è meno comoda da usare perchè non è ortogonale.

Figura 5.3 (proiezione obliqua)

### Confronto delle varianze

Concentriamoci ora sul calcolo delle varianze degli stimatori. Introduciamo allo scopo le seguenti notazioni:

$$\begin{aligned}\bar{\Sigma}_1 &:= [S_1^\top S_1]^{-1} \\ A_1 &:= [S_1^\top S_1]^{-1} S_1^\top \\ \Sigma_2 &:= [S_2^\top Q_1 S_2]^{-1} \quad ;\end{aligned}$$

ovviamente,  $\bar{\theta}_1 = A_1 y$  e  $\text{Var}_{\theta_1} \bar{\theta}_1 = \sigma^2 \bar{\Sigma}_1$ . 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).

**Proposition 8.2.** Siano  $\hat{\theta}_1(\mathbf{y})$  e  $\hat{\theta}_2(\mathbf{y})$  gli stimatori di Markov definiti dalle formule (8.2.7) e (8.2.8). Si ha allora:

$$\text{Var}_{\theta} \begin{Bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{Bmatrix} = \sigma^2 \begin{bmatrix} \bar{\Sigma}_1 + A_1 S_2 \Sigma_2 S_2^\top A_1^\top & -A_1 S_2 \Sigma_2 \\ -\Sigma_2 S_2^\top A_1^\top & \Sigma_2 \end{bmatrix} . \quad (8.2.13)$$

**Proof.** Incominciamo col dimostrare che  $\text{Var}_{\theta} [\hat{\theta}_2] = \sigma^2 \Sigma_2$ . Dalla (8.2.8) si ha

$$\text{Var}_{\theta} [\hat{\theta}_2] = \Sigma_2 S_2^\top Q_1 \text{Var}_{\theta} [\mathbf{y}] Q_1 S_2 \Sigma_2 = \sigma^2 \Sigma_2 S_2^\top Q_1 S_2 \Sigma_2 = \sigma^2 \Sigma_2 \quad ,$$

dato che  $\text{Var}_{\theta} [\mathbf{y}] = \sigma^2 I$  ed  $Q_1$  è idempotente.

Mostriamo ora che i due stimatori  $\hat{\theta}_1(\mathbf{y})$  e  $\hat{\theta}_2(\mathbf{y})$  sono scorrelati. Si ha infatti:

$$\text{Cov}_{\theta} [\hat{\theta}_1(\mathbf{y}), \hat{\theta}_2(\mathbf{y})] = \bar{\Sigma}_1 S_1^\top \text{Var}_{\theta} [\mathbf{y}] Q_1 S_2 \Sigma_2 = \sigma^2 \bar{\Sigma}_1 S_1^\top Q_1 S_2 \Sigma_2 = 0 \quad ,$$

perchè  $S_1^\top Q_1 = Q_1 S_1 = 0$ .

Usando ora la (8.2.9) si trova

$$\text{Cov}_{\theta} [\hat{\theta}_1(\mathbf{y}), \hat{\theta}_2(\mathbf{y})] = -A_1 S_2 \text{Var}_{\theta} [\hat{\theta}_2] = -\sigma^2 A_1 S_2 \Sigma_2 \quad .$$

Calcoliamo infine  $\text{Var}_{\theta} [\hat{\theta}_1(\mathbf{y})]$ . Dato che  $\bar{\theta}_1(\mathbf{y})$  e  $\hat{\theta}_2(\mathbf{y})$  sono scorrelati, si ha

$$\begin{aligned}\text{Var}_{\theta} [\hat{\theta}_1(\mathbf{y}) - A_1 S_2 \hat{\theta}_2(\mathbf{y})] &= \text{Var}_{\theta} [\bar{\theta}_1(\mathbf{y})] + A_1 S_2 \text{Var}_{\theta} [\hat{\theta}_2(\mathbf{y})] S_2^\top A_1^\top \\ &= \sigma^2 [\bar{\Sigma}_1 + A_1 S_2 \Sigma_2 S_2^\top A_1^\top] \quad .\end{aligned}$$

che conclude la dimostrazione della formula (8.2.13).  $\square$

**Remark 8.2.** La formula (8.2.13) descrive l'effetto dell'aumento del numero di parametri nel modello sulla varianza delle stime e sull'errore quadratico medio residuo. In particolare (8.2.13) mostra che la "nuova" stima  $\hat{\theta}_1$  di  $\theta_1$  è generalmente "peggiore" della prima in termini di varianza. La varianza,  $\Sigma_1$ , di  $\hat{\theta}_1$  è in effetti *più grande* di quella di  $\bar{\theta}_1$ , essendo

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

e il termine che si somma a  $\bar{\Sigma}_1$  è in generale non nullo.

Purtroppo però la varianza delle stime del parametro  $\theta$  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  $S_1$ , ovvero se accade che

$$S_1^\top S_2 = 0 \quad (S_2^\top S_1 = 0)$$

le formule si semplificano drammaticamente (dato che  $Q_1 S_2 = S_2$ ) e i due stimatori  $\hat{\theta}_1$  e  $\hat{\theta}_2$  si possono calcolare indipendentemente l'uno dall'altro con le solite formule,

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

In particolare si trova  $\hat{\theta}_1 = \bar{\theta}_1$  e quindi anche  $\Sigma_1 = \bar{\Sigma}_1$ . Per comprendere questo fenomeno (che a prima vista può sembrare sconcertante) basta pensare che ci sono molte parametrizzazioni del modello "ideale"  $S\theta$  che sono assolutamente equivalenti agli effetti di descrivere i dati  $y$ . Per esempio, introducendo una fattorizzazione  $QR$  di  $S$ , vede facilmente che si può sempre fattorizzare  $S$  come prodotto di una matrice a colonne ortogonali (le prime  $p+k$  colonne di  $Q$ ) per una matrice quadrata  $R \in \mathbb{R}^{p+k \times p+k}$  non singolare (a struttura triangolare inferiore). Definendo il nuovo parametro  $\beta := R\theta$  si può riparametrizzare il modello in modo tale che le colonne di  $S$  siano ortogonali. In questo caso la varianza di  $\hat{\beta}_1$  non aumenta aumentando la parametrizzazione del modello con  $k$  nuovi parametri. La morale della storia è che la varianza delle stime dei parametri *dipende dal sistema di coordinate scelto per rappresentare il modello* (in breve, "dalla base"). I confronti dovrebbero essere quindi fatti solo tra quantità che sono *invarianti per cambio di base*. Quantità di questo genere sono ad esempio gli errori residui di modellizzazione.  $\square$

**Problem 8.3.** Si sa che dei dati osservati  $\{y(t)\}$  possono essere descritti mediante il seguente modello lineare,

$$\mathbf{y}(t) = a + bt + \mathbf{e}(t), \quad t = 1, \dots, N$$

dove  $a, b$  sono parametri incogniti e  $\{\mathbf{e}(t)\}$  è rumore bianco Gaussiano (a media nulla) di varianza  $\sigma^2$  incognita. In realtà si è interessati alla stima del solo coefficiente angolare  $b$ . Si propone allora il seguente modello alternativo che descrive la "derivata discreta" dei dati. Detto  $\mathbf{z}(t) := \mathbf{y}(t) - \mathbf{y}(t-1)$  si scrive:

$$\mathbf{z}(t) = b + \mathbf{w}(t), \quad t = 1, \dots, N$$

dove  $\mathbf{w}(t) = \mathbf{e}(t) - \mathbf{e}(t - 1)$ .

Scrivere i due modelli lineari corrispondenti in forma vettoriale e individuarne le matrici  $S$  e la matrice varianza  $\sigma^2 R$  dell'errore di misura. Confrontare le stime (di M.V.) del parametro  $b$  ottenute usando i due modelli e dire quale delle due ha varianza minore. Spiegare intuitivamente il risultato.

Soluzione : Il modello lineare,

$$y(t) = a + bt + \mathbf{e}(t), \quad t = 1, \dots, N$$

si può riscrivere in forma vettoriale come:

$$\mathbf{y} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ \vdots & \vdots \\ 1 & N \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \mathbf{e} := [s_1 \quad s_2] \begin{bmatrix} a \\ b \end{bmatrix} + \mathbf{e}$$

per cui la matrice varianza della stima (di M.V.) del parametro  $\theta := [a \ b]^\top$  è

$$\text{Var} \{ \hat{\theta}_N \} = \sigma^2 \begin{bmatrix} s_1^\top s_1 & s_1^\top s_2 \\ s_2^\top s_1 & s_2^\top s_2 \end{bmatrix}^{-1}$$

Ne segue che la varianza della seconda componente  $\hat{a}_N$  è

$$\begin{aligned} \text{var} \{ \hat{a}_N \} &= \sigma^2 \frac{s_1^\top s_1}{s_1^\top s_1 s_2^\top s_2 - (s_1^\top s_2)^2} = \sigma^2 \frac{N}{N s_2^\top s_2 - (\sum_{k=1}^N k)^2} \\ &= \sigma^2 \frac{1}{\sum_{k=1}^N k^2 - 1/N (\sum_{k=1}^N k)^2} \end{aligned}$$

Il modello alternativo in forma vettoriale si scrive:

$$\mathbf{z} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} b + \mathbf{w} := s_1 b + \mathbf{w}$$

dove, posto  $\mathbf{e} := [\mathbf{e}(1) \ \mathbf{e}(2) \ \dots \ \mathbf{e}(N)]^\top$ , si ha

$$\mathbf{w} := \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ \dots & & & \ddots & \dots \\ 0 & \dots & -1 & 1 \end{bmatrix} \mathbf{e} := L\mathbf{e}$$

stante che  $\mathbf{e}(0)$  non è disponibile. Da questa relazione la matrice varianza dell'errore di misura si scrive  $\sigma^2 R = \sigma^2 L L^\top$  e la varianza della stima è quindi

$$\text{var} \{ \hat{b}_N \} = [s_1^\top (\sigma^2 R)^{-1} s_1]^{-1} = \sigma^2 / \|L^{-1} s_1\|^2$$

8.2. Regressione lineare a stadi

Ora è facile calcolare l'inversa di  $L$  e da questa ottenere

$$L^{-1}s_1 = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 1 & \dots & 0 \\ \dots & & & \ddots & \dots \\ 1 & 1 & \dots & & 1 \end{bmatrix} s_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \\ \vdots \\ N \end{bmatrix}$$

per cui col secondo modello si ottiene

$$\text{var} \{\hat{b}_N\} = \sigma^2 / \sum_{k=1}^N k^2$$

che è minore di quella dello stimatore nel modello a due parametri. Il motivo di questa differenza sta nel fatto che il secondo modello è parametrizzato in modo più parsimonioso del primo.  $\square$

**Problem 8.4.** Si consideri il modello lineare

$$y(t) = a + bt + e(t), \quad t = 1, \dots, N$$

dove  $a, b$  sono parametri incogniti e  $\{e(t)\}$  è rumore bianco Gaussiano (a media nulla) di varianza  $\sigma^2$ . Si è interessati alla stima del solo coefficiente angolare  $b$ .

Trovare lo stimatore  $\bar{a}_N$  di  $a$ , che si otterrebbe se non ci fosse il regressore  $s_2b$  (i.e.  $b = 0$ ).

Potrebbe sembrare logico definire un vettore  $z$  di osservazioni "centrate" togliendo a ciascuna componente di  $y$  la stima dell'offset  $a$  (calcolata in precedenza),  $\bar{a}_N$ , e postulare per  $z := y - s_1\bar{a}_N$  un modello senza termine costante del tipo

$$z = s_2b + w$$

dove  $w$  è ancora rumore bianco. Trovare lo stimatore di M.V.  $\hat{b}_N$  di  $b$  basato su questo modello e dare delle condizioni necessarie e sufficienti per la sua correttezza (ovviamente sapendo che il modello vero è quello con entrambi i regressori).

Soluzione : Il modello lineare,

$$y(t) = a + bt + e(t), \quad t = 1, \dots, N$$

si può riscrivere in forma vettoriale come:

$$y = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ \vdots & \vdots \\ 1 & N \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + e := [s_1 \quad s_2] \begin{bmatrix} a \\ b \end{bmatrix} + e$$

Senza il regressore  $s_2b$ , la stima  $\bar{a}_N$  è la media campionaria delle osservazioni,

$$\bar{a}_N = \frac{1}{s_1^\top s_1} s_1^\top y := \bar{y}_N$$

ma se è presente anche il secondo regressore si ha

$$\bar{a}_N = \frac{1}{s_1^\top s_1} s_1^\top \left\{ \begin{bmatrix} s_1 & s_2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \mathbf{w} \right\} = a + \frac{1}{s_1^\top s_1} s_1^\top s_2 b + \frac{1}{s_1^\top s_1} s_1^\top \mathbf{w}$$

e quindi, la media campionaria  $\bar{a}_N$  è uno stimatore corretto di  $a$  se e solo se  $s_1^\top s_2 = 0$ .

Usando il modello senza termine costante del tipo  $\mathbf{z} = s_2 b + \mathbf{w}$ , lo stimatore di  $b$  si esprimerebbe nella forma

$$\hat{b}_N = \frac{1}{s_2^\top s_2} s_2^\top (\mathbf{y} - s_1 \bar{a}_N)$$

per cui

$$\mathbb{E}_\theta \hat{b}_N = \frac{1}{s_2^\top s_2} s_2^\top \mathbb{E}_\theta (\mathbf{y} - s_1 \bar{a}_N) = \frac{1}{s_2^\top s_2} s_2^\top [(s_1(a - \mathbb{E}_\theta \hat{a}_N) + s_2 b)]$$

per cui  $\hat{b}_N$  è uno stimatore corretto se lo è  $\bar{a}_N$ ; i.e. se  $s_1^\top s_2 = 0$ .

Alternativamente, con qualche passaggio si vede che

$$\hat{b}_N = \frac{1}{s_2^\top s_2} s_2^\top \left[ I - s_1 \frac{1}{s_1^\top s_1} s_1^\top \right] (s_2 b + \mathbf{w}).$$

e quindi  $\hat{b}_N$  è corretto se e solo se

$$\left[ I - s_1 \frac{1}{s_1^\top s_1} s_1^\top \right] s_2 = s_2$$

dove il termine tra parentesi quadre è il proiettore ortogonale sul complemento ortogonale di  $\text{span}\{s_1\}$ . Quindi  $\hat{b}_N$  è corretto se e solo se  $s_2$  appartiene al complemento ortogonale dello spazio  $\text{span}\{s_1\}$ ; in sostanza se e solo se  $s_1^\top s_2 = 0$ .  $\square$

### 8.3 Il test $F$

Il problema di verifica d'ipotesi che ci interessa in questo capitolo è un problema di verifica di ipotesi lineari, sostanzialmente dello stesso tipo visto nelle sezioni 7.3 e 7.4 del capitolo precedente. Cercheremo qui di dare una forma particolarmente intuitiva e trasparente al rapporto di verosimiglianza (7.4.5), che abbiamo in realtà già calcolato in generale nella sezione 7.4.

Dobbiamo confrontare l'errore quadratico residuo che si commette descrivendo i dati osservati  $y$  mediante un modello delle due classi  $M_1$  (ipotesi  $H_0$ ) ed  $M_2$  (ipotesi  $H_1$ ) definite in (8.1.1).

Indichiamo con  $\bar{\varepsilon}(y) := y - S_1 \bar{\theta}_1(y) = (I - P_1)y$  il vettore dei residui usando il modello a  $p$  parametri e con  $\hat{\varepsilon} = y - S_1 \hat{\theta}_1(y) - S_2 \hat{\theta}_2(y) = (I - P)y$  quello relativo al modello aumentato e supponiamo inizialmente di non sapere chi sia il modello "vero". Ricordiamo che  $(I - P)$  e  $(P - P_1)$  proiettano su spazi ortogonali

(Proposizione 8.1); infatti  $I - P$  proietta sul complementare  $\mathcal{S}^\perp$  mentre  $(P - P_1)$  proietta sul sottospazio  $\mathcal{S} \cap \mathcal{S}_1^\perp$ , per cui possiamo scrivere,

$$\begin{aligned} R_0(\mathbf{y})^2 = \|\hat{\varepsilon}\|^2 &= \|(I - P) + (P - P_1)\mathbf{y}\|^2 = \|\hat{\varepsilon}\|^2 + \|(P - P_1)\mathbf{y}\|^2 \\ &= \|\hat{\varepsilon}\|^2 + \|Q_1 P_{2||1}\mathbf{y}\|^2 = R_1(\mathbf{y})^2 + \|Q_1 P_{2||1}\mathbf{y}\|^2. \end{aligned} \quad (8.3.1)$$

Il termine  $\|Q_1 P_{2||1}\mathbf{y}\|^2$  all'ultimo membro si può esprimere, usando la (8.2.10), in funzione dello stimatore  $\hat{\theta}_2(\mathbf{y})$  come

$$\|Q_1 P_{2||1}\mathbf{y}\|^2 = \hat{\theta}_2(\mathbf{y})^\top \Sigma_2^{-1} \hat{\theta}_2(\mathbf{y})$$

per cui  $R_0(\mathbf{y})^2$  si può riscrivere nella forma

$$R_0(\mathbf{y})^2 = R_1(\mathbf{y})^2 + \hat{\theta}_2(\mathbf{y})^\top \Sigma_2^{-1} \hat{\theta}_2(\mathbf{y}) = R_1(\mathbf{y})^2 + \|\hat{\theta}_2(\mathbf{y})\|_{\Sigma_2^{-1}}^2 \quad (8.3.2)$$

dove  $\sigma^2 \Sigma_2$  è la varianza dello stimatore  $\hat{\theta}_2(\mathbf{y})$ .

L'importanza del termine correttivo in questa espressione dipende da quale classe di modelli ha effettivamente generato i dati. Nel caso in cui il modello che ha effettivamente generato i dati fosse quello a soli  $p$  parametri,  $M_1$ , i regressori addizionali  $S_2 \theta_2$  e lo stimatore  $\hat{\theta}_2(\mathbf{y})$  descriverebbero solo rumore bianco additivo e si può intuitivamente dedurre che in questo caso  $S_2 \hat{\theta}_2$  risulterà mediamente piccolo. Il termine  $\|\hat{\theta}_2(\mathbf{y})\|_{\Sigma_2^{-1}}$  nella (8.3.1) risulterà in particolare piccolo rispetto all'errore quadratico complessivo  $R_1(\mathbf{y})^2$ .

**Theorem 8.1.** *Se vale l'ipotesi  $H_0$ , i due addendi al secondo membro della (8.3.2) sono indipendenti e hanno entrambi distribuzioni di probabilità di tipo  $\chi^2$ ; rispettivamente,*

$$\frac{\|\hat{\varepsilon}\|^2}{\sigma^2} \sim \chi^2(N - p - k) \quad (8.3.3)$$

e

$$\frac{\hat{\theta}_2(\mathbf{y})^\top \Sigma_2^{-1} \hat{\theta}_2(\mathbf{y})}{\sigma^2} \sim \chi^2(k) \quad (8.3.4)$$

per cui il rapporto

$$\varphi(\mathbf{y}) := \frac{N - p - k}{k} \frac{\|\hat{\theta}_2(\mathbf{y})\|_{\Sigma_2^{-1}}^2}{R_1(\mathbf{y})^2} \quad (8.3.5)$$

ha distribuzione  $\mathcal{F}(k, N - p - k)$ .

**Proof.** Assumendo che il modello "vero" sia  $M_1$  (ipotesi  $H_0$ ), la somma dei quadrati dei residui,  $\|\hat{\varepsilon}\|^2 = \|(I - P_1)\mathbf{y}\|^2$  è uguale a  $\|Q_1 \epsilon\|^2$ . Sostituendo nell'ultima delle (8.3.1) l'espressione del modello "vero"  $M_1$  e usando la relazione  $P_{2||1}^\top Q_1 = Q_1^\top P_{2||1} = Q_1 P_{2||1}$ , si riconosce immediatamente che  $Q_1 P_{2||1}\mathbf{y} = P_{2||1}^\top Q_1 (S_1 \theta_1 + \epsilon) = P_{2||1}^\top Q_1 \epsilon$ . D'altro canto  $\hat{\varepsilon} = (I - P)\epsilon$  e quindi  $\epsilon^\top (I - P)^\top Q_1 P_{2||1} \epsilon = 0$ , dato che  $(I - P)$  proietta sul complemento ortogonale di  $\mathcal{S}$ . Questo fatto implica incorrelazione di  $\hat{\varepsilon}$  e  $Q_1 P_{2||1} \epsilon$  e quindi l'indipendenza dei due termini al secondo membro in (8.3.1).

Infine, come è ben noto (proposizione 2.8),  $\frac{1}{\sigma^2} \|\hat{\varepsilon}\|^2 \sim \chi^2(N - (p + k))$ , indipendentemente da quale modello ha generato i dati, e, come abbiamo visto a suo tempo (proposizione 2.7),  $\frac{1}{\sigma^2} \hat{\theta}_2(\mathbf{y})^\top \Sigma_2^{-1} \hat{\theta}_2(\mathbf{y}) \sim \chi^2(k)$  perchè la media di  $\hat{\theta}_2(\mathbf{y})$  è zero sotto  $H_0$ . In effetti, dalle distribuzioni dell'enunciato del teorema si ha, esattamente

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

dove i due argomenti denotano i *gradi di libertà* della distribuzione  $F$ .  $\square$

Per una dimostrazione alternativa si può vedere il [62, p. 73].

**Problem 8.5.** Usando l'indipendenza degli addendi nella relazione (8.3.2), dare una dimostrazione alternativa del fatto che sotto  $H_0$ ,

$$\frac{1}{\sigma^2} \|\hat{\varepsilon}\|^2 \sim \chi^2(N - (p + k)).$$

*Soluzione* Sotto  $H_0$ , è chiaro che  $\frac{1}{\sigma^2} \|\hat{\varepsilon}\|^2 \sim \chi^2(N - p)$ . Inoltre abbiamo visto che nelle stesse ipotesi  $\frac{1}{\sigma^2} \hat{\theta}_2(\mathbf{y})^\top \Sigma_2^{-1} \hat{\theta}_2(\mathbf{y}) \sim \chi^2(k)$ . Dato che i due addendi sono indipendenti,  $\frac{1}{\sigma^2} \|\hat{\varepsilon}\|^2$  deve necessariamente avere distribuzione  $\chi^2$  (Teorema 2.3) e il numero di gradi di libertà dev'essere  $N - p - k$ .  $\diamond$

Normalmente  $N - p$  è molto più grande di  $k$  e la distribuzione  $F$  si può approssimare molto bene con una  $\chi^2$  a  $k$  gradi di libertà, nel senso che per  $N \rightarrow \infty$ , vale la relazione,

$$k \varphi(\mathbf{y}) \sim \chi^2(k) \quad (\text{se vale } M_1) \quad (8.3.6)$$

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  $x_\alpha$  il valore dell'ascissa per cui

$$P_{\chi^2(k)}\{k \varphi(\mathbf{y}) > x_\alpha\} = \alpha$$

che si trova sulle tabelle della distribuzione  $\chi^2(k)$ , si va a vedere se il valore campionario della statistica  $k\varphi(\mathbf{y})$  assume valori maggiori o uguali a  $x_\alpha$ . In questo caso si rifiuta l'ipotesi del modello "semplice"  $M_1$ , con probabilità  $\alpha$  di commettere un errore, beninteso nel caso in cui i dati siano stati davvero generati da  $M_1$ . La distribuzione della statistica (8.3.5) 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 può stimare con simulazioni Monte Carlo oppure con il metodo approssimato basato sulla  $F$  non centrale descritto al capitolo precedente. Ricordiamo che la probabilità  $1 - \beta$ , di scegliere il modello giusto quando è vero  $M_2$  è la *potenza del test*.



## 8.4 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 due vettori di osservazioni  $\mathbf{y} := [\mathbf{y}_1^\top \mathbf{y}_2^\top]^\top$  che per semplicità assumeremo di uguale dimensione  $N$  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

$$\mathbf{y}_1 = S\theta + \epsilon_1, \quad \text{Var}[\epsilon_1] = \sigma^2 I_N \quad (8.4.1)$$

ottenedo, come è ben noto, il classico stimatore  $\hat{\theta}(\mathbf{y}_1) = [S^\top S]^{-1} S^\top \mathbf{y}_1$ . Vogliamo ora valutare la "bontà statistica" del modello stimato,  $S\hat{\theta}(\mathbf{y}_1)$  per descrivere i dati  $\mathbf{y}_2$  che abbiamo tenuto da parte. Naturalmente perchè questa operazione abbia senso dobbiamo supporre che i dati nei successivi  $N$  campioni siano stati "generati dallo stesso meccanismo" che ha generato  $\mathbf{y}_1$ , il che si può esprimere in modo equivalente dicendo che le d.d.p. (o almeno le statistiche del primo 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  $[\mathbf{y}_1^\top \mathbf{y}_2^\top]^\top$  abbiano lo stesso vettore di media  $\mu$  (che potrebbe essere qualunque) e che la varianza complessiva di  $\mathbf{y}$  sia  $\sigma^2 I_{2N}$ . In questo modo  $\mathbf{y}_1$  e  $\mathbf{y}_2$  risultano scorrelati.

Consideriamo allora il vettore errore *finale* di predizione dei dati futuri

$$\epsilon := \mathbf{y}_2 - S\hat{\theta}(\mathbf{y}_1) \quad (8.4.2)$$

che ha media  $\mu - S[S^\top S]^{-1} S^\top \mu$  per cui sottraendo la media e calcolando la varianza di  $\epsilon$  si trova

$$\text{Var}[\epsilon] = \sigma^2 I_N + S[S^\top S]^{-1} S^\top \sigma^2 I_N S[S^\top S]^{-1} S^\top = \sigma^2 [I_N + S[S^\top S]^{-1} S^\top].$$

Come misura dell'errore finale di predizione prendiamo la varianza scalare normalizzata che ha l'espressione

$$\begin{aligned} \frac{1}{N} \text{var}[\epsilon] &= \sigma^2 \frac{1}{N} \text{Tr} \{ I_N + S[S^\top S]^{-1} S^\top \} = \sigma^2 \{ 1 + \text{Tr}([S^\top S]^{-1} S^\top S) \} \\ &= \sigma^2 \left( 1 + \frac{p}{N} \right) \end{aligned} \quad (8.4.3)$$

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  $\sigma^2$ , 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 (2.4.25)

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

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

$$FPE(p) := \frac{1}{N} \|\hat{\epsilon}_p\|^2 \frac{(1 + \frac{p}{N})}{(1 - \frac{p}{N})} := \hat{\sigma}_p^2 \frac{(1 + \frac{p}{N})}{(1 - \frac{p}{N})} \quad (8.4.4)$$

che viene anch'esso chiamato **errore finale di predizione** basato su un modello di dimensione  $p$ .

La stima dell'ordine del modello può essere basata sulla minimizzazione di questo indice. 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.

## 8.5 Un algoritmo di regressione lineare a stadi

Le formule di aggiornamento della stima (8.2.9) 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  $\theta^k = [\theta_1, \dots, \theta_k]^\top$  ottenuto modellando i dati con il modello lineare a  $k$  parametri (in cui abbiamo denotato  $\sigma w$  con il simbolo  $\varepsilon$ )

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

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

$$y = S_{k+1} \theta + \varepsilon \quad , \quad (8.5.1)$$

con

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

Usando le formule di aggiornamento per gli stimatori, si trova

$$\hat{\theta}_{k+1} = \frac{1}{s_{k+1}^\top R_k s_{k+1}} s_{k+1}^\top R_k y = \frac{1}{s_{k+1}^\top R_k s_{k+1}} s_{k+1}^\top [y - S_k^\top \bar{\theta}^k] \quad (8.5.2)$$

e inoltre

$$\hat{\theta}^k = (S_k^\top S_k)^{-1} S_k^\top [y - s_{k+1} \hat{\theta}_{k+1}] = \bar{\theta}^k - (S_k^\top S_k)^{-1} S_k^\top s_{k+1} \hat{\theta}_{k+1} \quad , \quad (8.5.3)$$

dove si sono usate le notazioni  $\hat{\theta}_{k+1}$  e  $\hat{\theta}^k$  per indicare lo stimatore di  $\theta_{k+1}$  e di  $[\theta_1, \dots, \theta_k]^\top$  relativi al modello aumentato (8.5.1) ed  $R_k$  ha il solito significato di proiettore sul complemento ortogonale dello spazio colonne di  $S_k$ ,

$$R_k = I - S_k(S_k^\top S_k)^{-1} S_k^\top \quad . \quad (8.5.4)$$

Al passo successivo (l'aggiunta della colonna  $s_{k+2}$  al modello (8.5.1)), si aggiorna lo stimatore

$$\bar{\theta}^{k+1} := \begin{bmatrix} \hat{\theta}_k \\ \hat{\theta}_{k+1} \end{bmatrix} \quad , \quad (8.5.5)$$

con formule esattamente analoghe alle (8.5.2)–(8.5.3). Naturalmente il vero problema è quello di fare sequenzialmente anche i calcoli relativi all'aggiornamento dei coefficienti, in particolare dell'inversa  $(S_{k+1}^\top S_{k+1})^{-1}$  a partire da  $(S_k^\top S_k)^{-1}$ . Si può pensare di fare questi conti aggiornando la fattorizzazione di Cholesky di  $S_k^\top S_k$ , 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  $S_k^\top S_k$ ). 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

$$Q_k[S_k \ y] = \left[ \begin{array}{c|c} U_k & y_k^1 \\ \hline 0 & y_k^2 \end{array} \right] \quad , \quad k \text{ righe} \quad . \quad (8.5.6)$$

Chiaramente si ha, con ovvio significato dei simboli,

$$\bar{\theta}^k = U_k^{-1} y_k^1 \quad , \quad \|\varepsilon_k\|^2 = \|y_k^2\|^2 \quad . \quad (8.5.7)$$

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  $v_1, \dots, v_k$  che definiscono le riflessioni elementari.

Si calcola il prodotto

$$Q_k \ s_{k+1} := \left. \begin{array}{l} a_{k+1} \\ b_{k+1} \end{array} \right\} \begin{array}{l} k \text{ righe} \\ n - k \text{ righe} \end{array} \quad (8.5.8)$$

per cui

$$Q_k[S_k \ s_{k+1} \ y] = \left[ \begin{array}{cc|c} U_k & a_{k+1} & y_k^1 \\ \hline 0 & b_{k+1} & y_k^2 \end{array} \right]$$

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  $k + 2, \dots, n$

$$H_{k+1} b_{k+1} = \|b_{k+1}\| e_1 \quad . \quad (8.5.9)$$

$H_{k+1}$  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  $e_1 = [1, 0, \dots, 0]^T$  in  $\mathbb{R}^{n-k}$ . Definendo allora

$$Q_{k+1} := \begin{bmatrix} I_k & 0 \\ 0 & H_{k+1} \end{bmatrix} \quad , \quad (8.5.10)$$

$$H_{k+1} y_k^2 := z_{k+1} \quad , \quad (8.5.11)$$

si ha

$$Q_{k+1} [S_k \ s_{k+1} \ y] = \left[ \begin{array}{cc|c} U_k & a_{k+1} & y_k^1 \\ \hline 0 & 0 & z_{k+1} \end{array} \right] \quad (8.5.12)$$

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

$$\begin{bmatrix} U_k & a_{k+1} \\ 0 & \|b_{k+1}\| \end{bmatrix} \begin{bmatrix} \theta_k \\ \hat{\theta}_{k+1} \end{bmatrix} = \begin{bmatrix} y_k^1 \\ z_{k+1}^1 \end{bmatrix} \quad . \quad (8.5.13)$$

(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 (8.5.7),

$$\|\varepsilon_{k+1}\|^2 = \|\varepsilon_k\|^2 - (z_{k+1}^1)^2 \quad . \quad (8.5.14)$$

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

$$\begin{aligned} U_{k+1} &= \begin{bmatrix} U_k & a_{k+1} \\ 0 & \|b_{k+1}\| \end{bmatrix} \\ y_{k+1}^1 &= [(y_k^1)^T, z_{k+1}^1]^T \\ y_{k+1}^2 &= [z_{k+1}^2, \dots, z_{k+1}^{n-k}]^T \end{aligned}$$

e usando la matrice ortogonale  $Q_{k+1}$  definita in (8.5.10) per trasformare il vettore  $s_{k+2}$  nel modo analogo a quanto fatto in (8.5.8) 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  $S_k := sp[S_k]$  opera sui vettori  $s_{k+1}$  e  $y$  semplicemente attraverso le

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

ed

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

per cui, ad esempio, la formula (8.5.2) per  $\hat{\theta}_{k+1}$  si può riscrivere come

$$\hat{\theta}_{k+1} = \frac{1}{\|b_{k+1}\|^2} b_{k+1}^\top y_k^2 \quad . \quad (8.5.15)$$

È facile controllare che questa relazione è identica alla

$$\hat{\theta}_{k+1} = \frac{z_{k+1}^1}{\|b_{k+1}\|} \quad , \quad (8.5.16)$$

che si ricava risolvendo l'ultima equazione in (8.5.13). La matrice della varianza delle stime si può scrivere immediatamente come

$$\Sigma_{k+1} = \sigma^2 \begin{bmatrix} \Sigma_k + \frac{U_k^{-1} a_{k+1} a_{k+1}^\top U_k^{-T}}{\|b_{k+1}\|^2} & -U_k^{-1} a_{k+1} \frac{1}{\|b_{k+1}\|^2} \\ -\frac{1}{\|b_{k+1}\|^2} a_{k+1}^\top U_k^{-T} & \frac{1}{\|b_{k+1}\|^2} \end{bmatrix} \quad (8.5.17)$$

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

$$\hat{\theta}^k = \bar{\theta}^k - U_k^{-1} a_{k+1} \hat{\theta}_{k+1} \quad . \quad (8.5.18)$$

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  $\{s_{k+1}, s_{k+2}, \dots, s_p\}$  delle colonne "mancanti" di un modello lineare

$$S\theta = [s_1, \dots, s_k, \dots, s_p] \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} \quad (8.5.19)$$

che si "propone" per descrivere i dati. Chiaramente, il modello (8.5.19) può risultare inutilmente complicato nel senso che l'introduzione di alcuni fra i parametri  $\{\theta_{k+1}, \dots, \theta_p\}$  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 [21].

### Uso della SVD

L'ultima osservazione ci induce a descrivere brevemente un possibile modo, basato sulla decomposizione ai valori singolari, che serve a 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:

$$S_1 = \bar{U} \bar{\Delta} \bar{V}^T \quad S = U \Delta V^T \quad (8.5.20)$$

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

$$\bar{\Delta} = \begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & 0 \end{bmatrix}, \quad \bar{\Sigma} = \text{diag} \{ \bar{\sigma}_1, \dots, \bar{\sigma}_p \} \quad (8.5.21)$$

$$\Delta = \begin{bmatrix} \Sigma_1 & 0 & 0 \\ 0 & \Sigma_2 & 0 \\ 0 & 0 & 0 \end{bmatrix}; \quad (8.5.22)$$

$$\Sigma_1 := \text{diag} \{ \sigma_1, \dots, \sigma_p \} \quad \Sigma_2 := \text{diag} \{ \sigma_{p+1}, \dots, \sigma_{p+k} \} \quad (8.5.23)$$

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.  $\bar{\sigma}_i \neq \sigma_i, i = 1, \dots, p$ .

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

$$\bar{y} = \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix} \beta + \lambda \bar{w} \quad (8.5.24)$$

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

$$\hat{\beta}_i(\mathbf{y}) = \frac{1}{\sigma_i} \bar{y}_i; \quad \text{var} \{ \hat{\beta}_i \} = \frac{\lambda^2}{\sigma_i^2} \quad i = 1, \dots, p+k. \quad (8.5.25)$$

Da notare che gli stimatori  $\hat{\beta}_i(\mathbf{y}); i = 1, \dots, p+k$  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  $\sigma_{p+1}$  dovuto all'aggiunta di una nuova colonna,  $s_{p+1}$ , risulta molto più piccolo di quelli del primo blocco  $\Sigma_1$ , la corrispondente stima del parametro aggiuntivo  $\hat{\beta}_{p+1}$  avrà in effetti varianza (molto) maggiore delle altre componenti. Notiamo il seguente fatto notevole:

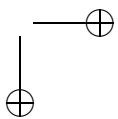
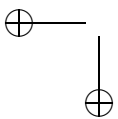
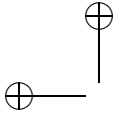
*Il rapporto tra la varianza di  $\hat{\beta}_{p+1}$  e quella del primo stimatore  $\hat{\beta}_1$  (che è la minima possibile) è il quadrato dell'indice di condizionamento numerico della matrice aumentata.*

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  $\hat{\theta} = V\hat{\beta}$  di matrice varianza  $V \text{Var}\{\hat{\beta}\}V^T$ , la cui varianza scalare è, come si controlla facilmente, la stessa di quella di  $\hat{\beta}$ . La regola che scaturisce 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"<sup>32</sup> 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* [64] 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  $\sigma_i$ ;  $i = 1, \dots, p$ , 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 [7, 8].

<sup>32</sup>Notare che questo non significa necessariamente che il nuovo regressore debba essere "quasi dipendente" dalle colonne preesistenti.





## CHAPTER 9

# IDENTIFICAZIONE DI SEGNALI QUASI PERIODICI IN RUMORE ADDITIVO

In questo capitolo assumeremo che il segnale da analizzare sia la somma di oscillazioni sinusoidali di ampiezza e frequenza incognita e di una componente puramente non deterministica (a spettro continuo). Discuteremo la modellizzazione della componente periodica e affronteremo successivamente il problema della stima del segnale e in special modo quello della stima delle frequenze delle sue componenti armoniche.

Il problema della stima di frequenze è un problema di stima non lineare che si presenta in numerosissime applicazioni. Per questo motivo è stato molto studiato in letteratura. Probabilmente tra le prime investigazioni "storiche" possiamo annoverare il problema della stima del periodo di rivoluzione del sole agli inizi del novecento [57, 71], il problema della stima della frequenza dei cicli macroeconomici etc etc.. Naturalmente la stima di frequenze è importante anche in numerosissime applicazioni ingegneristiche.

I metodi più antichi per la stima di frequenze si rifanno semplicemente al calcolo della trasformata di Fourier del segnale. Il modulo quadro della trasformata è una stima dello spettro di potenza che dovrebbe presentare dei picchi ben marcati in corrispondenza alle frequenze di componenti armoniche. Purtroppo però, se il segnale è rumoroso, la stima dello spettro ottenuta con la trasformata di Fourier è una stima pessima dal punto di vista statistico. Questo fatto è stato mostrato chiaramente da Bartlett, [1] che nel 1950 ha prodotto delle famose espressioni della varianza asintotica dello stimatore di spettro ottenuto con la trasformata di Fourier.

Naturalmente uno dei problemi di questo metodo è il fatto che la stima dello spettro è vista come un problema di stima *non parametrica*, cioè la stima di una *funzione* della frequenza, fatta con dati che sono irrimediabilmente *finiti*. Anche se nella letteratura degli anni cinquanta si è cercato di rimediare a queste difficoltà introducendo vari artifici cosiddetti di "finestratura temporale", i metodi basati sull'analisi di Fourier (periodogramma) vengono normalmente usati solo per una verifica "visiva" dei risultati.

I metodi più affidabili sono metodi di stima *parametrica*. In letteratura ne sono stati proposti molti, ma apparentemente non esiste un'analisi che permetta

di stabilire quale metodo sia migliore di altri e in quali circostanze. Una delle difficoltà dell'utente è in effetti di riuscire a orientarsi per capire quale dei metodi presenti in letteratura si adatti meglio al problema in esame.

Ci si è riproposto in questo capitolo di fare una rassegna critica dei metodi principali di stima di frequenze presentati in letteratura. La rassegna riguarda anche l'analisi di certi fondamenti "teorici" dei metodi discussi, come la convergenza di certe statistiche (varianza e spettro campionari) di segnali con componenti periodiche, al tendere della numerosità campionaria all'infinito. Questi presupposti non sono banali perchè, come è ben noto, un segnale con componenti periodiche *non è ergodico* e il limite della covarianza campionaria anche se normalmente esiste, dipende dalla particolare traiettoria del segnale osservato. Questi dettagli sono di norma totalmente ignorati in letteratura.

## 9.1 Rappresentazione di processi puramente deterministici

Per le definizioni di processi stazionari puramente non deterministici (in seguito p.n.d) e puramente deterministici (in seguito p.d.) rimandiamo alla letteratura ad es. [45]. Per quanto servirà in questo testo un processo p.d. sarà semplicemente un processo (in generale complesso) che è somma di  $\nu$  componenti armoniche elementari, del tipo

$$\mathbf{z}(t) = \sum_{k=1}^{\nu} \mathbf{z}_k e^{i\omega_k t}, \quad t \in \mathbb{Z} \quad (9.1.1)$$

dove le  $\omega_k \in [-\pi, \pi]$  sono pulsazioni reali che si possono senza perdita di generalità supporre diverse tra loro e le  $\mathbf{z}_k$ ,  $k = 1, \dots, \nu$  sono variabili aleatorie (complesse) a varianza finita<sup>33</sup>. La stazionarietà, implica in particolare che le correlazioni delle varie componenti armoniche

$$\mathbb{E} [\mathbf{z}_k \bar{\mathbf{z}}_h] e^{i\omega_k t - i\omega_h s} \quad k, h = 1, 2, \dots, \nu$$

debbano dipendere da  $t - s$ , il che può accadere solo per  $k = h$  mentre per  $k \neq h$  si deve necessariamente avere  $\mathbb{E} [\mathbf{z}_k \bar{\mathbf{z}}_h] = 0$ . Si vede così che le  $\{\mathbf{z}_k\}$  debbono essere tra loro scorrelate. In queste ipotesi, calcolando la funzione di correlazione di  $\mathbf{z}$  si trova

$$r(t, s) = \mathbb{E} \mathbf{z}(t) \bar{\mathbf{z}}(s) = \sum_{k=1}^{\nu} \sigma_k^2 e^{i\omega_k(t-s)}, \quad \sigma_k^2 = \mathbb{E} |\mathbf{z}_k|^2$$

da cui  $r(t, s) = r(t-s)$  e quindi  $\mathbf{z}$  è effettivamente un processo stazionario (in senso debole).

Per un processo *reale* si ha  $\mathbf{z}(t) = \bar{\mathbf{z}}(t)$  (il complesso coniugato) per ogni  $t$  e quindi un segnale p.d. reale può essere convenientemente espresso come la somma di  $2\nu$  componenti oscillatorie scorrelate, del tipo

$$\mathbf{z}(t) = \frac{\mathbf{z}(t) + \bar{\mathbf{z}}(t)}{2} = \sum_{k=-\nu}^{\nu} \frac{1}{2} \mathbf{z}_k e^{i\omega_k t}, \quad \omega_{-k} = -\omega_k \quad \mathbf{z}_{-k} = \bar{\mathbf{z}}_k \quad (9.1.2)$$

<sup>33</sup>In questo capitolo non assumeremo che i segnali in gioco abbiano media nulla.

dove le  $\{z_k\}$  sono variabili aleatorie tra loro scorrelate e  $\bar{z}_k$  denota la variabile complesso-coniugata. Possiamo interpretare il termine (aggiuntivo) corrispondente a  $k = 0$  nella somma come un' eventuale componente continua a frequenza zero ( $\omega_0 = 0$ ) del segnale.

Scrivendo  $z_k = x_k + iy_k$ , per indice negativo ( $-k$ ) si ha  $z_{-k} = x_k - iy_k$ ; e quindi l'incorrelazione dei coefficienti a indice diverso implica che

$$\mathbb{E} \{ (x_k + iy_k) \overline{(x_k - iy_k)} \} = \mathbb{E} \{ (x_k^2 - y_k^2) + 2ix_k y_k \} = 0$$

da cui scende che  $\mathbb{E} x_k^2 = \mathbb{E} y_k^2$  e  $\mathbb{E} x_k y_k = 0$ . Ora, ogni componente armonica elementare del processo (9.1.2) si può scrivere in forma reale come

$$z_k(t) := \frac{1}{2} \{ z_k e^{i\omega_k t} + \bar{z}_k e^{-i\omega_k t} \} = x_k \cos \omega_k t - y_k \sin \omega_k t \quad k = 1, \dots, \nu. \quad (9.1.3)$$

Sia  $\sigma_k^2 := \mathbb{E} |z_k(t)|^2$  la potenza statistica della componente  $k$ -sima. Con un facile calcolo si trova

$$\sigma_k^2 = \mathbb{E} |z_k(t)|^2 = \mathbb{E} |z_k(0)|^2 = \mathbb{E} x_k^2 = \mathbb{E} y_k^2$$

con la convenzione che l'eventuale componente continua ( $\omega_0 = 0$ ) ha potenza  $\sigma_0^2$ .

La rappresentazione (9.1.3) ha una forma equivalente in termini di modello di stato che ci tornerà utile più avanti. Ogni componente armonica elementare ha una realizzazione di stato del tipo

$$\begin{bmatrix} x_k(t+1) \\ y_k(t+1) \end{bmatrix} = \begin{bmatrix} \cos \omega_k & -\sin \omega_k \\ \sin \omega_k & \cos \omega_k \end{bmatrix} \begin{bmatrix} x_k(t) \\ y_k(t) \end{bmatrix} \quad (9.1.4)$$

$$z_k(t) = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x_k(t) \\ y_k(t) \end{bmatrix} \quad (9.1.5)$$

con condizioni iniziali aleatorie scorrelate  $x_k(0) = x_k$ ,  $y_k(0) = y_k$  di ugual varianza  $\sigma_k^2$ . Giustapponendo queste  $\nu$  rappresentazioni elementari si ottiene un modello di stato complessivo per il segnale  $z$  che scriviamo simbolicamente nella forma

$$s(t+1) = A s(t) \quad (9.1.6)$$

$$z(t) = c^\top s(t) \quad (9.1.7)$$

dove  $s(t)$  è il vettore di stato di dimensione  $2\nu$  ottenuto incolonnando i vettori di stato elementari in (9.1.5) per  $k = 1, 2, \dots, \nu$ . La matrice  $A$  ha una struttura diagonale a blocchi  $A = \text{diag} \{A_1, \dots, A_\nu\}$  in cui i blocchi  $A_k$ , di dimensione  $2 \times 2$ , descrivono la dinamica della componente elementare di frequenza  $\omega_k$ . Notare che le  $A_k$  sono tutte matrici ortogonali.

La varianza di stato del modello (9.1.4) è una matrice diagonale

$$P = \mathbb{E} s(0)s(0)^\top = \text{diag} \{ \sigma_0^2 I_2, \dots, \sigma_\nu^2 I_2 \} \quad (9.1.8)$$

e, usando una formula ben nota, si trova l'espressione per la funzione di covarianza:

$$\sigma(\tau) = c^\top A^\tau P c = \sum_{k=0}^{\nu} \sigma_k^2 \cos \omega_k \tau \quad (9.1.9)$$

da cui lo spettro di  $\mathbf{z}$  (che dev'essere una funzione pari) si può scrivere nella forma

$$\phi(\omega) = \sum_{k=-\nu}^{\nu} \frac{1}{2} \sigma_k^2 \delta(\omega - \omega_k), \quad \sigma_{-k}^2 = \sigma_k^2. \quad (9.1.10)$$

**Remark 9.1.** Il modello (9.1.6) è stato ricavato in una base speciale assumendo di avere frequenze diverse nella rappresentazione (9.1.1). Ci si chiede se in un qualunque modello di stato di un processo scalare p.d. possano esserci autovalori multipli di  $A$ . Se si richiede l'osservabilità del modello, la risposta, è no. Questo fatto si può facilmente verificare usando il criterio di Hautus. Con lo stesso criterio si vede facilmente che le rappresentazioni di ordine 2 per le eventuali componenti a frequenze  $\omega_0 = 0$  e  $\omega_k = \pm\pi$  sono ridondanti. Per queste componenti la rappresentazione minima ha ovviamente dimensione uno.

A questo modello di stato corrisponde una descrizione "ingresso-uscita" del tipo

$$A(z^{-1})\mathbf{z}(t) = 0 \quad A(z^{-1}) = \prod_{k=1}^{\nu} (1 - 2\cos\omega_k z^{-1} + z^{-2}) \quad (9.1.11)$$

dove  $A(z^{-1}) = z^{-n} \det(zI - A)$  è il polinomio caratteristico della matrice  $A$ . Questa descrizione ingresso-uscita è *minima* nel senso che non esiste polinomio di grado più basso di  $A(z^{-1})$  che annulla il segnale  $\mathbf{z}(t)$ ; i.e. (9.1.11) è l'equazione alle differenze di ordine minimo possibile che descrive  $\mathbf{z}(t)$ . Notiamo che, introducendo le condizioni iniziali (aleatorie) la Z-trasformata della soluzione  $\mathbf{z}$  si può esprimere come una funzione razionale

$$\mathbf{z}(t) = \frac{N(z^{-1})}{A(z^{-1})}$$

dove  $N(z^{-1})$  è un polinomio a coefficienti aleatori determinati dalle condizioni iniziali.

### Segnali quasi periodici in rumore bianco

Supponiamo ora che il segnale osservato  $\mathbf{y}(t)$  sia somma di oscillazioni sinusoidali di ampiezza e frequenza incognita e di una componente di rumore che supponiamo puramente non deterministica (a spettro continuo),

$$\mathbf{y}(t) = \mathbf{z}(t) + \mathbf{e}(t) \quad (9.1.12)$$

dove  $\mathbf{z}(t)$  è un segnale p.d. reale del tipo analizzato nella sezione precedente e  $\mathbf{e}(t)$  è un segnale a spettro continuo. In questo capitolo supporremo che  $\mathbf{e}(t)$  sia *rumore bianco* di varianza incognita  $\sigma^2$ . Il caso in cui il rumore additivo ha struttura più complessa verrà studiato nei capitoli successivi. In questa ipotesi, usando (9.1.11) e combinando con l'espressione (9.1.12) possiamo descrivere il segnale mediante un modello ingresso-uscita del tipo

$$A(z^{-1})\mathbf{y}(t) = A(z^{-1})\mathbf{e}(t). \quad (9.1.13)$$

Da notare che la cancellazione del fattore comune  $A(z^{-1})$  nei due termini non è lecita perchè il sistema parte da *condizioni iniziali non nulle* al tempo zero. La risposta del sistema (9.1.13) è in realtà costituita di due termini: il primo uguale all'evoluzione libera corrispondente a condizioni iniziali aleatorie e ingresso nullo, che determina la parte p.d.  $z$ , descritta dal modello di tipo equazione alle differenze (9.1.11) visto al paragrafo precedente. Il secondo termine è la risposta forzata del sistema (a partire da condizioni iniziali nulle) che si riduce semplicemente al rumore bianco  $e$ , visto che a causa delle condizioni iniziali nulle si possono cancellare numeratore e denominatore nella funzione di trasferimento descritta dal modello (9.1.13).

Questo modello *parametrico*, che sembra sia stato usato per la prima volta da Nehorai [42], costituirà il punto di partenza per l'identificazione delle frequenze incognite del segnale con metodi parametrici, in particolare col metodo PEM.

## 9.2 Metodi non parametrici per la stima di spettri

Nel seguito useremo l'acronimo inglese PSD (Power Spectral Density) per abbreviare la dizione *Densità Spettrale di Potenza*. Faremo qui una breve rassegna di metodi per la stima della PSD di un segnale stazionario. Due ottimi riferimenti per la trattazione di questo argomento da un punto di vista ingegneristico sono il testo di Porat [48] (in particolare il capitolo 4) e il libro di Roberts e Mullis [52].

Il più antico e semplice approccio al problema della stima spettrale e quindi, in particolare, alla stima della frequenza di eventuali componenti periodiche del segnale, è l'analisi di Fourier. Come abbiamo già detto la stima dello spettro può essere vista come un problema di stima *non parametrica*, nel senso che l'oggetto da stimare è una *funzione* (della frequenza) e non un parametro di dimensione fissa, indipendente dalla numerosità campionaria. Notiamo che i metodi non parametrici sono assolutamente generali perchè stimano lo spettro solo sulla base di campioni del segnale in esame, senza usare alcuna informazione a priori sulla struttura del segnale.

Consideriamo un processo stocastico scalare  $\{y(t); t = 0, \pm 1, \pm 2, \dots\}$  stazionario (in senso lato) e a varianza finita. Come è noto la funzione (sequenza) di correlazione di  $y$  è definita dalla relazione

$$r(k) = E\{y(t+k)\bar{y}(t)\}$$

dove la barra denota il complesso coniugato. Si mostra che  $r(k) = \bar{r}(-k)$  e  $r(0) \geq |r(k)|$  per ogni  $k$ . A rigore, solo i processi per cui  $r(k)$  ammette trasformata di Fourier sono descrivibili per mezzo di una PSD che è una vera e propria funzione di  $\omega$ . In particolare, la correlazione dei processi che hanno componenti periodiche di potenza finita (che danno origine a *righe spettrali*), che sono quelli di principale interesse in questo capitolo e nei seguenti, non è a stretto rigore F-trasformabile e si deve usare una nozione generalizzata di spettro in cui compaiono funzioni  $\delta$  di Dirac. In questo senso per noi la PSD del processo  $y$  sarà la trasformata di Fourier

generalizzata (nel senso delle distribuzioni) della sequenza di correlazione:

$$\phi(\omega) = \sum_{k=-\infty}^{+\infty} r(k)e^{-i\omega k} \quad (9.2.1)$$

Quando  $\phi(\omega)$  è una funzione vera e propria, essa è positiva in  $[-\pi, \pi]$ . In ogni caso il suo integrale è la potenza statistica  $\mathbb{E} \mathbf{y}(t)^2$ , del processo. Se il processo  $\mathbf{y}$  è reale,  $\phi(\omega) = \phi(-\omega)$  ( $\phi$  è una funzione pari di  $\omega$ ). Ricordiamo che se al posto della correlazione si prende la sequenza delle covarianze  $\sigma(\tau)$ , si ha una relazione analoga che descrive lo spettro di potenza denotato comunemente con  $S(\omega)$ , che non ha riga spettrale in  $\omega = 0$  (componente continua).

Nel caso di nostro interesse i dati a disposizione sono una sequenza finita di campioni  $\{y(1), \dots, y(N)\}$  del processo. Il **Periodogramma**  $\hat{\phi}_N(\omega)$ , è la stima dello spettro che si ottiene prendendo la norma quadrato della trasformata discreta di Fourier (DFT) della sequenza

$$\hat{\phi}_N(\omega) = \frac{1}{N} \left| \sum_{t=1}^N y(t)e^{-i\omega t} \right|^2 \quad (9.2.2)$$

Il motivo della normalizzazione (divisione per  $N$ ) si può far risalire formalmente alla definizione della DFT, ma è forse più intuitivo basarsi sulla classica relazione tra periodogramma e correlazione campionaria calcolata sui campioni  $\{y(1), \dots, y(N)\}$ . Quest'ultima è definita dalla

$$\hat{r}(k) = \frac{1}{N} \sum_{t=1}^{N-k} y(t+k)\bar{y}(t), \quad k = 0, 1, \dots, N-1 \quad (9.2.3)$$

e da  $\hat{r}(k) = \overline{\hat{r}(-k)}$  per  $k < 0$ . Notare che  $\hat{r}(k)$  è una sequenza definita solo per  $|k| \leq N-1$ . A rigore questo stimatore non è corretto (è *biased* in inglese) ma quello che si otterrebbe dividendo per  $N-k$  anzichè  $N$  oltre ad avere varianza maggiore (specie per  $k$  vicino ad  $N$ ) non gode della proprietà notevole descritta qui sotto.

**Proposition 9.1.** *Si ha*

$$\hat{\phi}_N(\omega) = \sum_{k=-(N-1)}^{+(N-1)} \hat{r}(k)e^{-i\omega k} \quad (9.2.4)$$

dove  $\hat{r}(k)$  rappresenta la correlazione campionaria definita dalla (9.2.3).

**Proof.** Si ha

$$\begin{aligned} \hat{\phi}_N(\omega) &= \frac{1}{N} \sum_{t=1}^N \sum_{s=1}^N y(t)\bar{y}(s)e^{-i\omega(t-s)} \\ &= \sum_{k=-(N-1)}^{N-1} \left\{ \frac{1}{N} \sum_{s=1}^{N-k} y(s+k)\bar{y}(s) \right\} e^{-i\omega k}. \end{aligned}$$

□

Il periodogramma coincide quindi con la stima dello spettro che si ottiene calcolando la Fourier-trasformata discreta (DFT) della sequenza *finita* di correlazioni campionarie  $\hat{r}(k)$ . Un risultato completamente analogo vale per il periodogramma ottenuto dalle covarianze campionarie, eliminando dal segnale la componente continua e prendendo le deviazioni  $y(t) - \bar{y}_N$  al posto dei dati originali.

L'analisi delle proprietà statistiche di  $\hat{\phi}_N(\omega)$  è molto importante perchè dimostra la scarsa qualità del periodogramma come stimatore di PSD; due misure spesso utilizzate per caratterizzare la performance dello stimatore sono l'errore sistematico (*bias*) ovvero la deviazione della stima dal suo valore atteso e la sua *varianza*.

Dai risultati noti del capitolo 1 oppure direttamente dalla relazione (9.2.3) si ricava che la media della correlazione campionaria è,  $\mathbb{E} \hat{r}(k) = (1 - \frac{k}{N})r(k)$  per  $k \geq 0$  e si trova subito che il periodogramma non è uno stimatore corretto perchè,

$$\begin{aligned} \mathbb{E} [\hat{\phi}_N(\omega)] &= \sum_{k=-(N-1)}^{+(N-1)} \mathbb{E} \hat{r}(k) e^{-i\omega k} = \sum_{k=-(N-1)}^{+(N-1)} \left(1 - \frac{|k|}{N}\right) r(k) e^{-i\omega k} \\ &:= \sum_{k=-\infty}^{+\infty} w_N(k) r(k) e^{-i\omega k} \end{aligned}$$

dove la funzione triangolare  $w_N$  è chiamata *finestra di Bartlett*. Nel dominio della frequenza questa relazione si scrive

$$\mathbb{E} [\hat{\phi}_N(\omega)] = \int_{-\pi}^{\pi} \phi(\lambda) W_N(\omega - \lambda) \frac{d\lambda}{2\pi} \tag{9.2.5}$$

dove  $W_N(\omega)$  è la trasformata di Fourier di  $w_N$  data da

$$W_N(\omega) = \frac{1}{N} \left[ \frac{\sin(\frac{\omega N}{2})}{\sin(\frac{\omega}{2})} \right]^2$$

Affinchè  $\mathbb{E} [\hat{\phi}_N(\omega)]$  sia il più vicino possibile a  $\phi(\omega)$ ,  $W_N(\omega)$  dovrebbe essere una buona approssimazione dell'impulso di Dirac ma questo non accade per piccoli valori di  $N$  poichè la larghezza a 3 dB del lobo principale di  $W_N(\omega)$  è  $\frac{2\pi}{N}$  ( $\frac{1}{N}$  in frequenza). Il lobo principale della funzione  $W_N(\omega)$  causa il problema dello *smearing* per cui due picchi in  $\phi(\omega)$  separati in frequenza per meno di  $\frac{1}{N}$  non sono risolvibili; per questo  $\frac{1}{N}$  è detto *limite di risoluzione spettrale del periodogramma*. Questo fatto d'altro canto è ovvio se si interpreta la trasformata (9.2.4) come una trasformata *finita*, nel qual caso la frequenza  $\omega$  della trasformata dev'essere interpretata come una variabile *discreta*,  $\omega \equiv \omega_k = \frac{2\pi}{N}k$ ;  $k = 1, 2, \dots, N$ .

Segue dalla formula (9.2.5) che quando  $N \rightarrow \infty$ :

$$\lim_{N \rightarrow +\infty} \mathbb{E} [\hat{\phi}_N(\omega)] = \phi(\omega)$$

per cui il periodogramma è uno stimatore spettrale *asintoticamente corretto* (*asymptotically unbiased in inglese*).

Il seguente risultato (vedere [49, p. 7] per la dimostrazione) è un pò meno generale ma va nella stessa direzione.

**Proposition 9.2.** *Per processi che soddisfano alla condizione*

$$\lim_{N \rightarrow +\infty} \frac{1}{N} \sum_{k=-N}^N |k| |r(k)| = 0$$

la PSD  $\phi(\omega)$  si può calcolare come limite

$$\phi(\omega) = \lim_{N \rightarrow +\infty} \mathbb{E} \left\{ \frac{1}{N} \left| \sum_{t=1}^N \mathbf{y}(t) e^{-i\omega t} \right|^2 \right\} \quad (9.2.6)$$

Da questa relazione si vede ancora che il periodogramma è una stima asintoticamente corretta della PSD.

In realtà la correttezza asintotica non è una proprietà molto utile e si vorrebbe che il limite, senza il segno di aspettazione, fosse uguale alla densità teorica  $\phi(\omega)$ . Questa proprietà dello stimatore, che è la ben nota proprietà di *consistenza*, è invece la proprietà desiderabile nelle applicazioni. Come vedremo tra poco, purtroppo il periodogramma *non è uno stimatore consistente*.

**Theorem 9.1.** *La varianza asintotica di  $\hat{\phi}_N(\omega)$  è data dalla relazione seguente:*

$$\lim_{N \rightarrow +\infty} E [\hat{\phi}_N(\omega) - \phi(\omega)]^2 = \begin{cases} 2\phi^2(\omega) & \omega = 0 \text{ oppure } \pm \pi \\ \phi^2(\omega) & 0 < \omega < \pi \end{cases} \quad (9.2.7)$$

mentre per  $\omega_1 \neq \omega_2$  si ha

$$\lim_{N \rightarrow +\infty} E \left\{ [\hat{\phi}_N(\omega_1) - \phi(\omega_1)][\hat{\phi}_N(\omega_2) - \phi(\omega_2)] \right\} = 0 \quad (9.2.8)$$

Da questo risultato ricaviamo che i valori del periodogramma  $\hat{\phi}_N(\omega)$  sono variabili aleatorie le cui deviazioni standard sono (anche per  $N \rightarrow \infty$  !) uguali ai corrispondenti valori del PSD. Perciò il periodogramma è, come già accennato in precedenza, uno stimatore spettrale inconsistente (in media quadratica e in probabilità e quindi anche con probabilità uno) che oscilla attorno al suo valore medio asintotico (il valore vero  $\phi(\omega)$ ) con una varianza che si mantiene molto elevata anche per  $N \rightarrow +\infty$ . Inoltre le variabili aleatorie  $\{\hat{\phi}_N(\omega); 0 \leq \omega \leq \pi\}$  sono asintoticamente scorrelate. In sostanza il periodogramma si comporta asintoticamente come un *rumore bianco* in frequenza.



Nella figura 9.2.1 è riportato lo spettro vero (in nero) del processo descritto dal modello ARMA

$$y(t) - 1.5y(t-1) + 0.7y(t-2) = e(t) - e(t-1) + 0.2e(t-2) \quad (9.2.9)$$

mentre in blu è riportato il periodogramma con  $N = 64$  e in verde quello con  $N = 512$ .

**Figure 9.2.1.** Spettro vero e periodogramma per l'esempio (9.2.9)

Quanto detto finora costituisce la principale limitazione del periodogramma come stimatore spettrale. Negli anni cinquanta, per cercare di migliorare le proprietà statistiche del periodogramma sono stati introdotti molti metodi di *finestrazione temporale*, alcuni dei quali sono presentati nel prossimo paragrafo. In genere questi metodi diminuiscono la varianza dello spettro stimato ma ne aumentano però il valore di bias.

### 9.3 Metodi di finestrazione spettrale

L'analisi statistica della sezione precedente dà indicazioni per cercare di migliorare la stima dello spettro, per mezzo di opportune tecniche di finestrazione temporale. Qui esamineremo solo alcune di queste tecniche.

#### Metodo di Blackman-Tukey

Lo stimatore spettrale di Blackman-Tukey è dato da:

$$\hat{\phi}_{BT}(\omega) = \sum_{k=-(M-1)}^{+(M-1)} w(k) \hat{r}(k) e^{-i\omega k} \quad (9.3.1)$$

dove  $\{w(k)\}$  è una finestra temporale, funzione pari ( $w(k) = w(-k)$ ), con  $w(0) = 1$ ,  $w(k) = 0$  per  $|k| \geq M$  con  $M < N$ . Nel caso in cui  $w(k)$  sia una finestra rettangolare ( $w(k) = 1$ ) si ottiene semplicemente una versione troncata di  $\hat{\phi}_N(\omega)$ .

Esprimendo la (9.3.1) in frequenza si trova ancora la relazione di convoluzione analoga alla (9.2.5),

$$\hat{\phi}_{BT}(\omega) = \hat{\phi}_N(\omega) \star W(\omega) = \int_{-\pi}^{\pi} \hat{\phi}_N(\lambda) W_B(\omega - \lambda) \frac{d\lambda}{2\pi} \quad (9.3.2)$$

Poichè per la maggior parte delle finestre spettrali,  $W(\omega)$ , hanno un picco dominante in  $\omega = 0$  allora da (9.3.2), lo stimatore spettrale di Blackman-Tukey corrisponde a una *media pesata "localmente"* del periodogramma. La media pesata in

(9.3.2), nell'intorno della frequenza  $\omega$ , fa in modo che diminuiscano le elevate variazioni del periodogramma attorno al valor medio di PSD; però se da una parte decresce la varianza dall'altra si verifica l'effetto indesiderato di riduzione della risoluzione spettrale (o aumento del bias). Infatti un'analisi statistica di  $\hat{\phi}_{BT}(\omega)$  analoga a quella di  $\hat{\phi}_N(\omega)$  dimostra come la risoluzione dello stimatore di Blackman-Tukey sia dell'ordine di  $1/M$  mentre la varianza è dell'ordine di  $M/N$  volte quella del periodogramma. Perciò la scelta della lunghezza  $M$  della finestra si deve basare su un compromesso tra risoluzione spettrale e varianza della stima; non è possibile ottenere una diminuzione simultanea di entrambe le quantità.

Una volta fissato  $M$  è importante effettuare una scelta del tipo di finestra perchè questa va a influire sulle prestazioni della stima; anche in questo caso è necessario stabilire un compromesso tra uno stretto lobo centrale (per ridurre lo smearing) e una diminuzione dei lobi secondari (per ridurre il leakage). Non si può ridurre allo stesso tempo l'energia nel lobo principale e in quelli secondari. Possiamo concludere perciò la discussione affermando che una volta scelto  $M$  lo stimatore  $\hat{\phi}_{BT}(\omega)$  è asintoticamente biased ma la varianza tende a zero al tendere di  $N$  all'infinito.

### Metodo di Bartlett

Il metodo di Bartlett per cercare di ridurre la varianza del periodogramma suddivide gli  $N$  campioni a disposizione in  $L = N/M$  sottocampioni formati da  $M$  campioni ciascuno e poi media i periodogrammi ottenuti dai sottocampioni per ogni valore di  $\omega$ .

Siano

$$y_j(t) = y((j-1)M + t), \quad t = 1, \dots, M, j = 1, \dots, L$$

i campioni del  $j$ -esimo sottocampione e sia

$$\hat{\phi}_j(\omega) = \frac{1}{M} \left| \sum_{t=1}^M y_j(t) e^{-i\omega t} \right|^2 \quad (9.3.3)$$

il corrispondente periodogramma.

Lo stimatore spettrale di Bartlett è allora dato da

$$\hat{\phi}_B(\omega) = \frac{1}{L} \sum_{j=1}^L \hat{\phi}_j(\omega) \quad (9.3.4)$$

Analizziamo le proprietà dello stimatore mettendolo in relazione con quello di Blackman-Tukey;

Riscriviamo (9.3.3) come

$$\hat{\phi}_j(\omega) = \sum_{k=-(M-1)}^{(M-1)} \hat{r}_j(k) e^{-i\omega k} \quad (9.3.5)$$

dove  $\{\hat{r}_j(k)\}$  è la sequenza di covarianza corrispondente al  $j$ -esimo sottocampione. Inserendo (9.3.5) in (9.3.4) si ottiene

$$\hat{\phi}_B(\omega) = \sum_{k=-(M-1)}^{(M-1)} \left[ \frac{1}{L} \sum_{j=1}^L \hat{r}_j(k) \right] e^{-i\omega k} = \sum_{k=-(M-1)}^{(M-1)} \hat{r}_B(k) e^{-i\omega k}$$

Lo stimatore di Bartlett è perciò simile nella forma allo stimatore di Blackman-Tukey con una finestra rettangolare di lunghezza  $M$ ; poichè il lobo principale di questa finestra è più stretto rispetto ad altri tipi di finestre ne segue che, nella classe degli stimatori di Blackman-Tukey, lo stimatore di Bartlett presenta uno smearing minore e quindi una migliore risoluzione ma aumenta il leakage e la varianza. Da questa discussione e dalle proprietà dello stimatore di Blackman-Tukey precedentemente illustrate si dimostra che, rispetto al metodo del periodogramma, il metodo di Bartlett presenta una riduzione della risoluzione e una diminuzione della varianza di un fattore  $L$ .

### Metodo di Welch

Il metodo di Welch si basa su quello di Bartlett ma presenta due differenze: i segmenti di campioni possono sovrapporsi e ogni segmento viene finestrato prima del calcolo del periodogramma.

Sia

$$y_j(t) = y((j-1)K + t), \quad t = 1, \dots, M, j = 1, \dots, S$$

il  $j$ -esimo segmento;  $(j-1)K$  è il valore iniziale della  $j$ -esima sequenza di campioni. Se  $K = M$  le sequenze non si sovrappongono e perciò siamo nel caso dello stimatore di Bartlett; tuttavia è consigliato usare  $K = M/2$ , si ottengono così  $S \approx 2M/N$  segmenti.

Il periodogramma finestrato corrispondente a  $y_j(t)$  è dato da

$$\hat{\phi}_j(\omega) = \frac{1}{MP} \left| \sum_{t=1}^M v(t) y_j(t) e^{-i\omega t} \right|^2 \quad (9.3.6)$$

dove  $P$  denota la potenza della finestra temporale  $\{v(t)\}$ :

$$P = \frac{1}{M} \sum_{t=1}^M |v(t)|^2$$

Lo stimatore di Welch si trova mediando i periodogrammi finestrati in (9.3.6):

$$\hat{\phi}_W(\omega) = \frac{1}{S} \sum_{j=1}^S \hat{\phi}_j(\omega)$$

In questo modo si ottiene una riduzione della varianza rispetto allo stimatore di Bartlett anche se non particolarmente significativa; lo stimatore di Welch può essere calcolato efficientemente tramite FFT ed è uno dei metodi di stima spettrale più frequentemente usati. Infine si dimostra che esso può essere approssimato allo stimatore di Blackman-Tukey che teoricamente è il metodo preferito.

### Metodo di Daniell

L'idea che sta alla base del metodo di Daniell, per ridurre l'elevata varianza del periodogramma, è di mediare quest'ultimo su piccoli intervalli centrati sulla frequenza  $\omega$ . Lo stimatore di Daniell, che si può implementare con una FFT, è quindi dato da:

$$\hat{\phi}_D(\omega_k) = \frac{1}{2J+1} \sum_{j=(k-J)}^{(k+J)} \hat{\phi}_N(\omega)$$

dove

$$\omega_k = \frac{2\pi}{\tilde{N}}k, \quad k = 0, \dots, \tilde{N} - 1$$

con  $\tilde{N} \gg N$  per ottenere un campionamento più preciso di  $\hat{\phi}_N(\omega)$ ; il parametro  $J$  invece dovrebbe essere scelto sufficientemente piccolo in modo da garantire che  $\phi(\omega)$  sia costante sull'intervallo

$$\left[ \omega - \frac{2\pi}{\tilde{N}}J, \quad \omega + \frac{2\pi}{\tilde{N}}J \right] \quad (9.3.7)$$

Introducendo  $\beta = 2J/\tilde{N}$  possiamo scrivere (9.3.7) nella seguente forma

$$[\omega - \pi\beta, \omega + \pi\beta]$$

e ottenere la versione continua dello stimatore di Daniell:

$$\hat{\phi}_D(\omega) = \frac{1}{2\pi\beta} \int_{-\omega-\pi\beta}^{\omega+\pi\beta} \hat{\phi}(\zeta) d\zeta \quad (9.3.8)$$

da cui si può intuire come aumentando  $\beta$  decresca la varianza e anche la risoluzione (aumentando così il bias). Confrontando infatti (9.3.8) con (9.3.2) ricaviamo che lo stimatore di Daniell è un caso particolare della classe di Blackman-Tukey degli stimatori, con la finestra spettrale rettangolare:

$$W(\omega) = \begin{cases} 1/\beta, & \omega \in [\omega - \pi\beta, \omega + \pi\beta] \\ 0, & \text{altrimenti} \end{cases}$$

mostrando perciò la diminuzione in precedenza accennata di risoluzione e varianza di un fattore  $M = 1/\beta$  rispetto al metodo del periodogramma.

Concludiamo questa sezione dicendo che i vari metodi di stima dello spettro basati sul periodogramma che sono stati finora presentati tentano tutti di ridurre la varianza del periodogramma a spese però di una riduzione della risoluzione. Si tratta comunque di metodi che sono essenzialmente forme speciali dell'approccio di Blackman-Tukey.

## 9.4 Stimatori parametrici di spettro

Questo argomento è esposto molto chiaramente nei libri [52, Cap. 11] e [48, Cap 6] per cui noi qui faremo solo una breve sintesi dei risultati principali.

Gli estimatori parametrici di spettro sono in teoria estimatori di densità spettrale e assumono quindi che il processo che genera i dati sia p.n.d.. In realtà, dato che si riducono all'identificazione di modelli AR, essi sono semplici e facili da calcolare e pertanto vengono usati anche per la stima di componenti quasi periodiche, semplicemente cercando di individuare le frequenze a cui si presentano dei picchi particolarmente acuti della funzione PSD.

Supponiamo di avere stime di  $n+1$  campioni successivi della funzione di correlazione o di covarianza di un processo stazionario. Vogliamo trovare una stima dello spettro del processo (supposto p.n.d.), che sia la più generale possibile, nel senso che non richieda altre informazioni sul processo oltre all'assegnazione dei suoi primi  $n+1$  momenti secondi.

Per risolvere questo problema ci si può rifare al *principio della massima entropia*: vedere ad esempio [45, p. 235-240]. Si dimostra che il processo stazionario con densità spettrale a massima entropia, che ha i primi  $n+1$  campioni di covarianza assegnati, è un processo autoregressivo di ordine  $n$ . Ne scende che lo stimatore spettrale a massima entropia si ottiene costruendo il modello AR che è individuato dalle  $n+1$  stime di covarianza assegnate. Questo stimatore si può calcolare partendo dalla matrice delle covarianze stimate, che deve essere definita positiva, risolvendo un sistema di equazioni lineari, cosiddetto di *Yule-Walker*. In teoria la matrice di covarianza di un processo stazionario dovrebbe aver struttura di Toeplitz, ma se si usano covarianze campionarie la struttura di Toeplitz si perde. Essa si recupera solo asintoticamente per  $N \rightarrow \infty$ . Bisognerebbe usare estimatori di covarianza più sofisticati che impongano la struttura di Toeplitz anche con  $N$  finito; per questo si può far riferimento all'articolo [12]. In questo caso le stime si possono calcolare in modo efficiente e ricorsivo in  $n$ , con l'algoritmo di Levinson.

Se  $N$  è piccolo si ricorre alla cosiddetta tecnica di Burg, [?], che allo stato è un procedimento semi-empirico che però dà in genere buoni risultati.

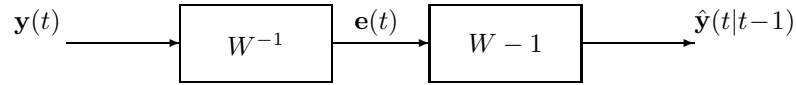
## 9.5 Stima di frequenze col metodo PEM

In questa sezione descriveremo un metodo di stima di frequenze di un segnale quasi periodico immerso in rumore bianco del tipo (9.1.12), basato sul principio della minimizzazione dell'errore di predizione (PEM). Il metodo è stato proposto originariamente da Nehorai [42] nel 1985 e successivamente è stato riesaminato e affinato da vari autori.

Come è ben noto [45] il predittore di Wiener per un processo descritto da un modello di innovazione razionale (di tipo ARMA) di funzione di trasferimento stabile e a fase minima,  $W(z)$ , normalizzata all'infinito ( $W(\infty) = 1$ ), è dato dalla formula

$$\hat{y}(t | t-1) = W(z)^{-1}[W(z) - 1]y(t) \quad (9.5.1)$$

vedere la figura 9.5.1.



**Figure 9.5.1.** *Struttura del predittore di un passo.*

L'errore di predizione associato a questo predittore

$$\mathbf{e}(t) = \mathbf{y}(t) - \hat{\mathbf{y}}(t | t - 1)$$

ha, come è ben noto, varianza minima nella classe di tutti i predittori lineari basati sulla storia passata di  $\mathbf{y}$ . Il processo  $\mathbf{e}$  è bianco e *causalmente equivalente* a  $\mathbf{y}$  nel senso che

$$\mathbf{H}_t^-(\mathbf{y}) \equiv \mathbf{H}_t^-(\mathbf{e})$$

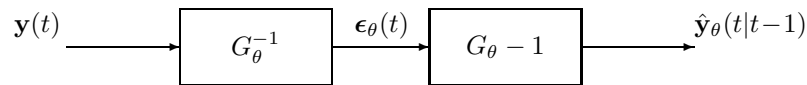
dove  $\mathbf{H}_t^-$  indica lo spazio di Hilbert generato dalla storia passata fino all'istante  $t$ . Quando il modello "vero" del processo non è noto, si fissa, basandosi sull'informazione disponibile a priori, una classe parametrica di funzioni di trasferimento a fase minima  $\{G_\theta(z); \theta \in \Theta\}$  e si calcola l'errore di predizione del modello  $G_\theta$  che è definito come la differenza

$$\epsilon_\theta(t) := \mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t | t - 1) \quad (9.5.2)$$

dove, in accordo con la (9.5.1),

$$\hat{\mathbf{y}}_\theta(t | t - 1) = G_\theta(z)^{-1}[G_\theta(z) - 1]\mathbf{y}(t) \quad (9.5.3)$$

è il predittore "approssimato" costruito in base al modello  $G_\theta$ .



**Figure 9.5.2.** *Struttura del predittore approssimato.*

Notiamo che, sostituendo la (9.5.3) nella (9.5.2) l'errore di predizione si può esprimere in forma simbolica con la semplice formula

$$\epsilon_\theta(t) = G_\theta^{-1}(z)y(t) \quad (9.5.4)$$

in altri termini l'errore di predizione associato al modello  $G_\theta$  si ottiene semplicemente filtrando il processo con la funzione di trasferimento inversa  $G_\theta^{-1}$ . Notiamo anche che questa operazione è stabile perchè  $G_\theta^{-1}$  ha tutti i poli all'interno del cerchio unitario.

### 9.5.1 Filtri Notch come modelli ARMA a poli sul cerchio

Il nostro intento è di generalizzare questa procedura al caso di processi con una componente p.d., ai quali in senso stretto non sarebbe applicabile per una serie di ovvi motivi.

Proviamo innanzitutto a definire una classe parametrica di modelli che descriva un processo  $y$ , che è la somma di  $n$  oscillazioni aleatorie più un rumore bianco additivo. Facendo riferimento a quanto detto nel paragrafo 9.1, useremo a questo scopo la classe di modelli (9.1.13), che sono modelli ARMA a poli tutti sulla circonferenza unitaria. Notiamo che il polinomio  $A(z^{-1})$  ha un'espressione del tipo

$$A(z^{-1}) = \prod_{k=1}^n (1 - 2\cos\omega_k z^{-1} + z^{-2}) = \tag{9.5.5}$$

$$1 + a_1 z^{-1} + a_2 z^{-2} + \dots + a_n z^{-n} + \dots + a_2 z^{-2n+2} + a_1 z^{-2n+1} + 1 z^{-2n} \tag{9.5.6}$$

e può essere parametrizzato direttamente negli  $n$  coefficienti incogniti

$$\theta := [a_1, a_2, \dots, a_n]^\top$$

che sono funzioni delle frequenze  $\omega_1, \omega_2, \dots, \omega_n$  attraverso i rispettivi coseni [42].

Per poter calcolare effettivamente l'errore di predizione associato a questa classe di modelli (formula (9.5.4)) è necessario che gli zeri delle funzione di trasferimento candidate (che in teoria dovrebbero essere sul cerchio e coincidere con i poli) siano stabili; i.e. strettamente dentro il cerchio unità. Per cercare di soddisfare a queste due esigenze si fissano gli zeri nei  $2n$  punti  $z_k = \rho e^{\pm i\omega_k}$  dove il parametro  $0 < \rho < 1$  è scelto prossimo a 1.

Per l'identificazione si usa l'algoritmo di Gauss-Newton nella forma standard dei metodi PEM per modelli ARMA

$$\theta_{k+1} = \theta_k + \left[ \sum_{t=1}^N \psi_{\theta_k}(t) \psi_{\theta_k}(t)^\top \right]^{-1} \sum_{t=1}^N \psi_{\theta_k}(t) \varepsilon_{\theta_k}(t) \tag{9.5.7}$$

dove  $\psi_{\theta}(t)$  è il gradiente di  $\varepsilon_{\theta}(t)$  cambiato di segno

$$\psi_{\theta}(t) := - \frac{\partial \varepsilon_{\theta}(t)}{\partial \theta}$$

e  $\varepsilon_{\theta}(t)$  si calcola con la formula (9.5.4), che nel nostro caso si può scrivere

$$A_{\theta}(\rho z^{-1}) \varepsilon_{\theta}(t) = A_{\theta}(z^{-1}) \mathbf{y}(t) \tag{9.5.8}$$

Questa equazione alle differenze può essere riscritta in modo da facilitare l'aggiornamento ricorsivo dei parametri, nella forma seguente

$$\varepsilon_{\theta}(t) = \mathbf{y}(t) + \mathbf{y}(t - 2n) - \rho^{2n} \varepsilon_{\theta}(t - 2n) - \varphi(t)^\top \theta \tag{9.5.9}$$

dove  $\varphi_\theta(t) = [\varphi_1(t) \varphi_2(t) \dots \varphi_n(t)]^\top$  è definita dalle

$$\varphi_i(t) = \begin{cases} -\mathbf{y}(t-i) - \mathbf{y}(t-2n+i) + \rho^i \varepsilon_\theta(t-i) + \rho^{2n-i} \varepsilon_\theta(t-2n+i) \\ \text{per } 1 \leq i \leq n-1 \end{cases} \quad (9.5.10)$$

$$\varphi_i(t) = -\mathbf{y}(t-n) + \rho^n \varepsilon_\theta(t-n) \quad \text{per } i = n. \quad (9.5.11)$$

Usando la (9.5.8) per il calcolo del gradiente si trova la relazione (vedere [42] per i dettagli)

$$\psi_\theta(t) = \frac{1}{A_\theta(\rho z^{-1})} \varphi_\theta(t) \quad (9.5.12)$$

che permette, una volta ottenuto il parametro corrente  $\theta = \theta_k$ , di calcolare il gradiente  $\psi_{\theta_k}(t)$  allo stadio  $k$ -simo. In sostanza l'algoritmo di ottimizzazione è composto dei seguenti passi

### Algorithm 9.1.

Data la stringa dei dati di ingresso  $\mathbf{y} = [\mathbf{y}(1) \dots \mathbf{y}(N)]^\top$ , e la stima  $\theta_k$  alla  $k$ -sima iterazione,

1. Si calcola la stringa degli errori di predizione  $\varepsilon_{\theta_k} = [\varepsilon_{\theta_k}(1) \dots \varepsilon_{\theta_k}(N)]^\top$  risolvendo l'equazione alle differenze (9.5.8). Questo si fa usando lo schema esplicito (9.5.9) che richiede il calcolo dell'array  $\varphi_{\theta_k} = [\varphi_{\theta_k}(1) \dots \varphi_{\theta_k}(N)]$
2. Si calcola il gradiente  $\Psi_{\theta_k} = [\psi_{\theta_k}(1) \dots \psi_{\theta_k}(N)]$ , usando l'equazione alle differenze (9.5.12) in cui  $A_\theta(\rho z^{-1}) = A_{\theta_k}(\rho z^{-1})$ .
3. si calcola la matrice pseudo-Hessiana

$$H_{\theta_k} := \sum_{t=1}^N \psi_{\theta_k}(t) \psi_{\theta_k}(t)^\top = \Psi_{\theta_k} \Psi_{\theta_k}^\top$$

e la sua inversa  $P_{\theta_k} := H_{\theta_k}^{-1}$ . Questo calcolo si potrebbe anche organizzare in forma ricorsiva come visto nell'algoritmo generale PEM del capitolo precedente.

4. Si aggiorna  $\theta_k$  usando la (9.5.7),

$$\theta_{k+1} = \theta_k + P_{\theta_k} \Psi_{\theta_k} \varepsilon_{\theta_k}$$

5. Si torna al passo 1) ponendo  $\theta_k = \theta_{k+1}$ .

Questo problema di stima è in genere mal condizionato. Nella funzione obiettivo si osserva un minimo molto pronunciato con una regione di attrazione che è tanto più piccola quanto più grande (prossimo a 1) si prende  $\rho$  [42, 6]. L'inizializzazione è quindi importante. Si può partire da delle stime iniziali ottenute da un periodogramma oppure per identificazione di un opportuno modello AR.

Fuori dalla regione di attrazione il gradiente è "piccolo" e la matrice Hessiana è mal condizionata per cui si possono avere esempi di convergenza estremamente



lenta o di accumulo di errori di arrotondamento (con molte iterazioni). Per ovviare a questo problema si sceglie  $\rho$  variabile con il passo di iterazione. Per  $k$  piccoli, quando le stime sono molto incerte, si prende  $\rho$  "piccolo" e poi lo si fa crescere con legge esponenziale, ad esempio

$$\rho(k+1) = \rho_0 \rho(k) + (1 - \rho_0) \rho(\infty) \quad (9.5.13)$$

dove  $\rho(\infty)$  è il valore a regime desiderato (ad esempio 0.995, [42, p. 987]) e  $\rho_0$  è la costante di tempo che determina il tasso di crescita di  $\rho(k)$ . Nehorai suggerisce di prendere  $\rho_0 \simeq 0.99$ .

Questa versione dell'algoritmo non è "ricorsiva" come in [42] ma usa tutti i dati disponibili in "batch". Questo aggrava un pò i calcoli ma, in linea di principio, dovrebbe portare a prestazioni migliori. In ogni caso nel calcolo dell'errore di predizione e del gradiente i dati iniziali sono sempre male utilizzati e sarebbe opportuno usare un fattore d'oblio  $\lambda(t)$  aggiornato con una relazione simile alla (9.5.13).

### Considerazioni sulla stabilità dell'algoritmo

Come è facile intuire, con stime iniziali poco affidabili, il polinomio  $A_{\theta_k}(z^{-1})$  potrebbe risultare instabile con conseguenze disastrose sul calcolo iterativo dell'errore di predizione e del gradiente. C'è però da notare che il polinomio che determina al passo  $k$ -simo la dinamica dell'errore di predizione e del gradiente (9.5.8), (9.5.12), non è  $A_{\theta_k}(z^{-1})$  ma bensì il polinomio "scalato"  $A_{\theta_k}(\rho_k z^{-1})$  in cui il fattore  $\rho_k < 1$  ha un effetto stabilizzante e può riportare i poli a modulo leggermente maggiore di uno dentro il cerchio unitario. Questo è probabilmente il motivo per cui, a quanto afferma Nehorai in [42], non si osserva praticamente mai il fenomeno dell'instabilità.

### Considerazioni sul Bias

Come notato in [6], l'introduzione del fattore di scala  $\rho$  nel modello (9.1.13) porta ad una descrizione, a stretto rigore "non corretta" del segnale che si vuole identificare. Si può calcolare esplicitamente l'errore asintotico (*bias*) che si commette nella stima PEM del termine  $\cos \omega_k$  utilizzando il modello "scalato". Nel modello con una sola sinusoide gli autori di [6] trovano

$$\cos \hat{\omega} = \frac{(1 + \rho^2) \cos \omega}{2\rho}$$

che con  $\rho = \rho(\infty) = .99$  diventa  $1.00005 \cos \omega$ . Come si vede si tratta di errori tollerabili.

### Difetti e possibili generalizzazioni

In molte situazioni pratiche il modello "sinusoidi in rumore bianco" potrebbe essere poco realistico e il termine d'errore e sarebbe più accuratamente descrivibile

come rumore colorato. Se il rumore additivo non è bianco la modellizzazione mediante un processo ARMA come in (9.1.13) non è più valida e i presupposti del metodo vengono a cadere. Inoltre il mal condizionamento rende l'algoritmo delicato e possono essere necessari aggiustamenti *ad hoc* (regolarizzazione etc.) per i casi problematici.

## 9.6 Stimatori basati sulla correlazione campionaria

Come abbiamo visto, l'ottimizzazione per la minimizzazione dell'errore quadratico medio di predizione deve essere effettuata con algoritmi iterativi di discesa che portano inevitabilmente a *minimi locali*. Per questo motivo sono stati proposti in letteratura dei metodi che non sono basati sulla minimizzazione di una cifra di merito ma sono classificabili come varianti dei *metodi dei momenti* e sono sostanzialmente basati su operazioni di fattorizzazione (eventualmente di tipo SVD) della matrice correlazione (o covarianza) campionaria del segnale di uscita. Questi metodi hanno il grosso vantaggio di essere robusti e numericamente stabili e di fornire il risultato in "un colpo", senza bisogno di iterazioni.

### 9.6.1 Espressione matriciale della correlazione

Vogliamo dare una formula semplice e compatta per la funzione di correlazione di un processo p.d. descritto da un modello di stato del tipo (9.1.6). Allo scopo possiamo anche pensare che  $\mathbf{z}(t)$  sia a valori vettoriali (ad esempio in  $\mathbb{R}^m$ ) e si ottenga mediante una matrice stato-uscita generale  $C \in \mathbb{R}^{m \times n}$ ; ( $n = 2\nu$ ) invece di una matrice riga come  $c^\top$  in (9.1.6). Dalla rappresentazione di stato si ottiene facilmente la seguente espressione,

$$R(\tau) := \mathbb{E} \mathbf{z}(t + \tau) \mathbf{z}(t)^\top = \mathbb{E} \mathbf{z}(\tau) \mathbf{z}(0)^\top = C A^\tau P C^\top \quad \tau \geq 0 \quad (9.6.1)$$

dove  $P$  è la matrice varianza dello stato iniziale:  $P := \mathbb{E} \mathbf{s}(0) \mathbf{s}(0)^\top$ . Nella particolare base in cui è rappresentato il modello (9.1.6),  $P$  è diagonale a blocchi con blocchi diagonali  $2 \times 2$  del tipo

$$P_k = \begin{bmatrix} \sigma_k^2 & 0 \\ 0 & \sigma_k^2 \end{bmatrix} \quad k = 1, 2, \dots, n.$$

Introducendo  $\Phi_+(z) = C(zI - A)^{-1} \bar{G}$  con  $\bar{G} = PC^\top$  lo spettro si può così formalmente decomporre nella forma

$$\Phi(z) = \Phi_+(z) + \Phi_+(1/z)^\top = C(zI - A)^{-1} \bar{G} + \bar{G}^\top (z^{-1}I - A^\top)^{-1} C^\top. \quad (9.6.2)$$

Da notare che la matrice  $A$  ha solo autovalori a coppie complesso-coniugate sulla circonferenza unità dove, a rigore, la trasformata  $Z$  della funzione covarianza non converge. Quindi questa espressione ha solo un valore formale.

Nel seguito sarà importante avere delle espressioni esplicite per la matrice di covarianza di vettori costruiti con traslazioni temporali di  $\mathbf{y}(t)$ . Notiamo in

particolare che da (9.1.6) si ottiene

$$\begin{aligned} \mathbf{y}^m &:= \begin{bmatrix} \mathbf{y}(t) \\ \mathbf{y}(t+1) \\ \vdots \\ \mathbf{y}(t+m-1) \end{bmatrix} = \begin{bmatrix} c^\top \\ c^\top A \\ \vdots \\ c^\top A^{m-1} \end{bmatrix} \mathbf{s}(t) + \begin{bmatrix} \mathbf{e}(t) \\ \mathbf{e}(t+1) \\ \vdots \\ \mathbf{e}(t+m-1) \end{bmatrix} \\ &:= \Omega_m \mathbf{s}(t) + \mathbf{e}^m \end{aligned} \quad (9.6.3)$$

dove  $\Omega_m$  (che scriveremo semplicemente  $\Omega$  quando non c'è pericolo di confusione) è la *matrice di osservabilità* del sistema (9.1.6) che ha la seguente struttura,

$$\Omega_m = \begin{bmatrix} 1 & 0 & \dots & 1 & 0 \\ \cos \omega_1 & -\sin \omega_1 & \dots & \cos \omega_\nu & -\sin \omega_\nu \\ \dots & \dots & \dots & \dots & \dots \\ \cos(m-1)\omega_1 & -\sin(m-1)\omega_1 & \dots & \cos(m-1)\omega_\nu & -\sin(m-1)\omega_\nu \end{bmatrix} \quad (9.6.4)$$

Nell'ipotesi che le frequenze  $\omega_k$  siano tutte diverse tra loro, se  $m \geq 2\nu$  la matrice ha manifestamente rango  $n = 2\nu$  e il sistema è osservabile. Calcolando la covarianza del vettore  $\mathbf{y}^m$ , nell'ipotesi che  $\mathbf{e}$  sia rumore bianco, si trova

$$R := E\{\mathbf{y}^m (\mathbf{y}^m)^\top\} = \Omega P \Omega^\top + \sigma^2 I_{2\nu} \quad (9.6.5)$$

dove

$$P = E\{\mathbf{s}(t)\mathbf{s}(t)^\top\} = \text{diag}\{P_1, \dots, P_\nu\}$$

in cui  $P_k$  è la varianza di stato del modello elementare di indice  $k$ , (9.1.4), che ha la forma

$$P_k = \begin{bmatrix} \sigma_k^2 & 0 \\ 0 & \sigma_k^2 \end{bmatrix}$$

Si può verificare che questa matrice risolve l'equazione di Lyapunov  $P_k = A_k P_k A_k^\top$ .

Nel seguito useremo anche una formula per la covarianza di due vettori, costruiti analogamente a  $\mathbf{y}^m$  in (9.6.3) ma relativi a due intervalli disgiunti. Tenedo conto del fatto che  $\mathbf{s}(t-k) = (A^\top)^k \mathbf{s}(t)$ , si trova

$$\Gamma := E \left\{ \begin{bmatrix} \mathbf{y}(t+1) \\ \mathbf{y}(t+2) \\ \vdots \\ \mathbf{y}(t+m) \end{bmatrix} [\mathbf{y}(t) \quad \mathbf{y}(t-1) \quad \dots \quad \mathbf{y}(t-l)] \right\} \quad (9.6.6)$$

$$= \begin{bmatrix} r(1) & r(2) & \dots & r(l+1) \\ r(2) & r(3) & \dots & r(l+2) \\ \dots & \dots & \dots & \dots \\ r(m) & r(m+1) & \dots & r(l+m) \end{bmatrix} \quad (9.6.7)$$

$$= E\{\Omega_m \mathbf{s}(t) \mathbf{s}(t)^\top \Omega_l^\top\} = \Omega_m P \Omega_l^\top \quad (9.6.8)$$

dove

$$\Omega_m := \begin{bmatrix} c^\top A \\ \vdots \\ c^\top A^m \end{bmatrix}, \quad \Omega_l^\top := [c \quad Ac \quad \dots \quad A^l c].$$

Da notare che se  $m$  e  $l$  sono maggiori di  $n = 2\nu$  entrambe queste matrici hanno rango  $n$  e quindi anche  $\Gamma$  ha rango  $n$ . Come si vede  $\Gamma$  è una matrice di Hankel, similmente a quanto accade per le matrici di covarianza costruite con le correlazioni dell'uscita di un sistema stocastico lineare che rappresenta un processo p.n.d. Di fatto questa struttura sta alla base dei metodi cosiddetti "a sottospazi" che verranno illustrati in seguito.

### Limite di covarianze campionarie per segnali p.d.

L'analisi dei metodi di stima che si usano in teoria dell'identificazione sono in genere basati sulla *ergodicità del secondo ordine* dei segnali in gioco. Questa proprietà si esprime dicendo che il limite delle covarianze campionarie

$$\hat{R}(\tau) := \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N y(t+\tau)y^\top(t)$$

calcolato per una singola traiettoria  $y$  del processo  $\mathbf{y}$ , è uguale all'aspettazione  $R(\tau) = E\mathbf{y}(t+\tau)\mathbf{y}(t)^\top$ , indipendentemente dalla traiettoria scelta. Sfortunatamente questa proprietà *non vale* per processi p.d.. Quindi il presupposto (comunemente dato per scontato in letteratura) che si possa sostituire la covarianza vera con il limite di quella campionaria anche quando sono in gioco segnali con componente p.d., non è affatto ovvio e va giustificato. In questo paragrafo esamineremo in dettaglio la questione.

Supponiamo che il processo  $\mathbf{y}(t)$  abbia la forma (9.1.12) e si possa quindi esprimere come l'uscita di un modello di stato:

$$\mathbf{x}(t+1) = A\mathbf{x}(t) \tag{9.6.9}$$

$$\mathbf{y}(t) = c\mathbf{x}(t) + \mathbf{e}(t) \tag{9.6.10}$$

in cui possiamo anche supporre che  $\mathbf{e}$  sia un processo p.n.d. generale, scorrelato da  $\mathbf{x}$ . Ovviamente  $c\mathbf{x}$  è la componente p.d. di  $\mathbf{y}$ . Come già visto nei nei paragrafi precedenti, si può supporre che la matrice  $A$  abbia una struttura diagonale a blocchi  $A = \text{diag}\{A_1, \dots, A_\nu\}$  in cui i blocchi  $A_k$  sono matrici oscillatorie di dimensione  $2 \times 2$ , mentre

$$c = [c_1^\top \quad c_2^\top \quad \dots \quad c_\nu^\top]$$

dove ciascun blocco riga  $c_k$  ha la forma  $c_k^\top = [1 \quad 0]$ . Si può quindi esprimere  $\mathbf{y}(t)$  nel seguente modo:

$$\mathbf{y}(t) = \sum_{k=1}^{\nu} c_k^\top \mathbf{x}_k(t) + \mathbf{e}(t) = \sum_{k=1}^{\nu} c_k^\top A_k^t \mathbf{x}_k(0) + \mathbf{e}(t)$$

con condizioni iniziali aleatorie  $\mathbf{x}_k(0)$ ,  $k = 1, 2, \dots, \nu$  tra loro scorrelate. Consideriamo ora per semplicità di notazione solo la componente  $k$ -sima della parte p.d.:

$$\mathbf{y}_k(t) := \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \cos \omega_k t & -\sin \omega_k t \\ \sin \omega_k t & \cos \omega_k t \end{bmatrix} \begin{bmatrix} \mathbf{x}_{k1}(0) \\ \mathbf{x}_{k2}(0) \end{bmatrix} \quad (9.6.11)$$

$$= \mathbf{x}_{k1}(0) \cos \omega_k t - \mathbf{x}_{k2}(0) \sin \omega_k t \quad (9.6.12)$$

$$:= \mathbf{a}_k \sin(\omega_k t + \varphi_k) \quad (9.6.13)$$

da cui si vede che la componente p.d. del segnale si può esprimere come somma di  $\nu$  componenti sinusoidali ciascuna delle quali con ampiezza  $\mathbf{a}_k$ , frequenza  $\omega_k$  e fase  $\varphi_k$ . L'ampiezza aleatoria  $\mathbf{a}_k$  è legata al vettore delle condizioni iniziali  $\mathbf{x}_k(0)$  dalla relazione:

$$\mathbf{a}_k^2 = \mathbf{x}_{k1}^2(0) + \mathbf{x}_{k2}^2(0).$$

Supponiamo di osservare una singola traiettoria del processo  $\mathbf{y}$  e studiamo il limite della covarianza campionaria quando la numerosità campionaria  $T$  tende all'infinito. Come si è già visto ([62, p. 108]) il limite per  $T \rightarrow \infty$  della correlazione campionaria di una somma di segnali sinusoidali (deterministici) del tipo  $y_k(t) = a_k \sin(\omega_k t + \varphi_k)$ ,  $k = 1, 2, \dots, \nu$  con  $\omega_k \neq \omega_j$ , è:

$$\hat{r}(\tau) = \sum_{k=1}^{\nu} \frac{a_k^2}{2} \cos \omega_k \tau \quad (9.6.14)$$

mentre la correlazione "vera" è

$$r(\tau) = \sum_{k=1}^{\nu} \sigma_k^2 \cos \omega_k \tau \quad \sigma_k^2 = 1/2 E \mathbf{a}_k^2 = E \mathbf{x}_{k1}^2(0) = E \mathbf{x}_{k2}^2(0). \quad (9.6.15)$$

Inoltre, si può dimostrare che la correlazione campionaria tra un segnale sinusoidale e una traiettoria  $\{e(t)\}$  di un processo p.n.d. ergodico a media nulla è sempre uguale a zero

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N y_k(t + \tau) e(t) = 0$$

(si veda [62]) per cui il limite della covarianza campionaria del segnale complessivo  $\mathbf{y}$  ha la forma

$$\hat{r}(\tau) = \sum_{k=1}^{\nu} \frac{a_k^2}{2} \cos \omega_k \tau + r_e(\tau) \quad (9.6.16)$$

dove, per l'ergodicità di  $\mathbf{e}$ ,  $r_e(\tau) = E \mathbf{e}(t + \tau) \mathbf{e}(t)^\top$  è la covarianza "vera" di  $\mathbf{e}$ .

Notiamo adesso che considerando ancora una volta solamente il blocco  $k$ -

simo si può rappresentare (il limite del-) la correlazione campionaria come

$$\hat{r}_k(\tau) = \frac{a_k^2}{2} \cos \omega_k \tau = \tag{9.6.17}$$

$$= [1 \ 0] \begin{bmatrix} \cos \omega_k \tau & \sin \omega_k \tau \\ -\sin \omega_k \tau & \cos \omega_k \tau \end{bmatrix} \begin{bmatrix} \frac{x_{k1}^2 + x_{k2}^2}{2} & 0 \\ 0 & \frac{x_{k1}^2 + x_{k2}^2}{2} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{9.6.18}$$

$$= c_k A_k^T \hat{P}_k c_k^T \tag{9.6.19}$$

Nel caso generale di segnali che sono somma di  $\nu$  componenti sinusoidali e di una componente p.n.d. la correlazione campionaria ha pertanto la seguente espressione

$$\hat{r}(\tau) = \sum_{k=1}^n \frac{a_k^2}{2} \cos \omega_k \tau + r_e(\tau) = \tag{9.6.20}$$

$$= c A^T \hat{P} c^T + r_e(\tau) \tag{9.6.21}$$

in cui le matrici  $A$  e  $P$  hanno strutture diagonali a blocchi. Si ottiene così un'espressione formalmente analoga alla (9.6.14) in cui però la matrice  $P$  è sostituita da una matrice diagonale a blocchi  $\hat{P}$  che dipende dalla particolare traiettoria del segnale. Notiamo che  $\hat{P}$  è simmetrica e non singolare con probabilità uno.

### 9.6.2 Alcuni metodi basati sulla correlazione campionaria

Passeremo in rassegna alcuni di questi metodi che sono sempre basati sull'ipotesi di osservare la somma di un segnale quasi periodico con sovrapposto del rumore bianco. Se il rumore additivo non è bianco si incorre in difficoltà e sembra che il metodo d'elezione da usare in queste circostanze sia l'*identificazione a sottospazi* che studieremo nel capitolo ?? . I metodi di stima che passeremo in rassegna più sotto si possono vedere come rudimentali precursori dell'identificazione a sottospazi.

#### Metodo di Yule-Walker esteso

Il metodo di Yule-Walker esteso, per il quale si usa l'acronimo inglese HOYW (High-Order Yule-Walker), si basa sul seguente ragionamento. Dato che in genere non si conosce il numero di componenti armoniche del segnale, si può sempre supporre che esso sia descritto da un modello ARMA del tipo (9.1.13), di ordine  $l$  sufficientemente elevato, superiore all'ordine  $n = 2\nu$  del modello minimo "vero":

$$\mathbf{y}(t) + b_1 \mathbf{y}(t-1) + \dots + b_l \mathbf{y}(t-l) = \mathbf{e}(t) + b_1 \mathbf{e}(t-1) + \dots + b_l \mathbf{e}(t-l) \tag{9.6.22}$$

che si può scrivere anche come

$$B(z^{-1})\mathbf{y}(t) = B(z^{-1})\mathbf{e}(t) \tag{9.6.23}$$

dove il polinomio  $B(z^{-1})$  si può pensare ottenuto moltiplicando  $A(z^{-1})$  in (9.1.13) per un certo altro polinomio  $\bar{A}(z^{-1})$  di grado pari a  $(l - n)$ .

Riscriviamo l'equazione (9.6.22) nella seguente forma piú concisa

$$[\mathbf{y}(t) \quad \mathbf{y}(t-1) \quad \dots \quad \mathbf{y}(t-l)] \begin{bmatrix} 1 \\ b \end{bmatrix} = \mathbf{e}(t) + \dots + b_l \mathbf{e}(t-l) \quad (9.6.24)$$

premultiplichiamo per  $[\mathbf{y}(t-l-1) \dots \mathbf{y}(t-l-m)]^T$ , dove  $m$  é un intero positivo che sarà specificato in seguito e calcoliamo le matrici covarianza a primo e secondo membro. Tenedo conto del fatto che per  $k > 0$  si ha  $E[\mathbf{y}(t-k)\mathbf{e}(t)] = 0$  e sostituendo  $E[\mathbf{y}(t-k)\mathbf{y}(t-j)] = r(j-k) = r(k-j)$ , si ottiene la seguente relazione:

$$\begin{bmatrix} r(l+1) & r(l) & \dots & r(1) \\ r(l+2) & r(l+1) & \dots & r(2) \\ \dots & \dots & \dots & \dots \\ r(l+m) & r(l+m-1) & \dots & r(m) \end{bmatrix} \begin{bmatrix} 1 \\ b \end{bmatrix} := \Gamma^c \begin{bmatrix} 1 \\ b \end{bmatrix} = 0. \quad (9.6.25)$$

dove la matrice  $\Gamma^c$  é la matrice  $\Gamma$  definita in (9.6.7) con le colonne in ordine opposto. La (9.6.25) scritta in forma scalare, è simile al classico sistema di equazioni di Yule-Walker [62, pp. 288]:

$$r(k) + \sum_{i=1}^l b_i r(k-i) = 0, \quad k = l+1, \dots, l+m.$$

Notiamo che scambiando l'ordine dei coefficienti, i.e. ponendo  $\bar{b}_k := b_{l-k}$ , il sistema di equazioni lineari (9.6.25) si potrebbe scrivere nella forma equivalente

$$\Gamma \begin{bmatrix} \bar{b} \\ 1 \end{bmatrix} = 0.$$

In ogni caso il vettore dei parametri soddisfa il sistema di equazioni:

$$\begin{bmatrix} r(l) & \dots & r(1) \\ \vdots & \dots & \vdots \\ r(l+m-1) & \dots & r(m) \end{bmatrix} \mathbf{b} = - \begin{bmatrix} r(l+1) \\ \vdots \\ r(l+m) \end{bmatrix} \quad (9.6.26)$$

Il metodo di HOYW per stimare le frequenze del segnale in esame utilizza il risultato precedente sostituendo però le covarianze teoriche  $\{r(k)\}$  con le covarianze campionarie  $\{\hat{r}(k)\}_{k=1}^{l+m}$  ricavate a partire dai campioni del segnale a disposizione.

Ovviamente a causa degli errori di stima in  $\{\hat{r}(k)\}$  invece di una uguaglianza si ha in realtà solo una relazione approssimata del tipo:

$$\begin{bmatrix} \hat{r}(l) & \dots & \hat{r}(1) \\ \vdots & \dots & \vdots \\ \hat{r}(l+m-1) & \dots & \hat{r}(m) \end{bmatrix} \hat{\mathbf{b}} \approx - \begin{bmatrix} \hat{r}(l+1) \\ \vdots \\ \hat{r}(l+m) \end{bmatrix} \quad (9.6.27)$$

Il passo fondamentale del metodo consiste nel risolvere (9.6.27) in  $\hat{\mathbf{b}}$  con il metodo ai minimi quadrati. A questo proposito bisogna però fare alcune precisazioni.

Denotiamo con  $\hat{\Gamma}$  la matrice  $m \times l$  delle covarianze campionarie in (9.6.27). Anche se per  $m, l \geq n$ , il rango di  $\Gamma$  è in teoria uguale ad  $n$ , quello di  $\hat{\Gamma}$  sarà sempre pieno (uguale al minimo tra  $m$  ed  $l$ ).

Per un processo del tipo (9.1.12), con  $N$  componenti oscillatorie e sovrapposto un rumore bianco, il limite per  $N \rightarrow \infty$  di  $\{\hat{r}(k)\}$  esiste ma non è necessariamente uguale a  $\{r(k)\}$ . Da quanto visto nel paragrafo precedente (formula (9.6.21)) si vede comunque che  $\hat{\Gamma}$  al limite tende ad una matrice di rango  $n$ , il che permetterebbe di determinare il numero di componenti armoniche.

Dato che  $\text{rank } \hat{\Gamma} \simeq n$  il sistema (9.6.27) è mal condizionato dal punto di vista numerico. Infatti si può dimostrare che ogni metodo ai minimi quadrati che stima  $\hat{b}$  direttamente da (9.6.27) ha una scarsa accuratezza. Per far fronte a queste difficoltà si usa la *Decomposizione ai Valori Singolari (SVD)* (si veda in appendice sezione ??) della matrice  $\hat{\Gamma}$ .

Sia

$$\hat{\Gamma} = U \Sigma V^T = [U_1 \quad U_2] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} \quad (9.6.28)$$

la SVD di  $\hat{\Gamma}$ , dove  $U, V$  sono matrici ortogonali di dimensioni rispettivamente  $m \times m$  e  $l \times l$  e  $\Sigma$  è una matrice diagonale  $m \times l$  a elementi positivi (i *valori singolari*) ordinati in senso decrescente.

Si dimostra che la matrice  $\hat{\Gamma}_n$ , ottenuta scartando in (9.6.28) la sottomatrice  $\Sigma_2$  che contiene i valori singolari a indice maggiore di  $n$ ,

$$\hat{\Gamma}_n := U_1 \Sigma_1 V_1^T$$

è la migliore approssimazione di rango  $n$  di  $\hat{\Gamma}$  in una varietà di possibili norme. Usando  $\hat{\Gamma}_n$  in (9.6.27) al posto di  $\hat{\Gamma}$  si ottiene il sistema di equazioni di HOYW di rango troncato:

$$\hat{\Gamma}_n \hat{b} \approx - \begin{bmatrix} \hat{r}(l+1) \\ \vdots \\ \hat{r}(l+m) \end{bmatrix} \quad (9.6.29)$$

che può essere risolta con un metodo ai minimi quadrati ottenendo:

$$\hat{b} = -V_1 \Sigma_1^{-1} U_1^T \begin{bmatrix} \hat{r}(l+1) \\ \vdots \\ \hat{r}(l+m) \end{bmatrix} \quad (9.6.30)$$

dove  $V_1 \Sigma_1^{-1} U_1^T$  è la pseudoinversa di  $\hat{\Gamma}_n$ . Una volta ottenuta la stima di  $\hat{b}$ , si considera il polinomio

$$1 + \sum_{k=1}^l \hat{b}_k z^{-k}$$

e le stime delle frequenze  $\{\hat{\omega}_k\}$  del segnale si fanno coincidere con le posizioni angolari delle  $n$  radici del polinomio che si trovano più vicine al cerchio di raggio



unitario. Si assume così che le "radici del segnale", ovvero le radici di  $A(z)$ , siano sempre piú vicine al cerchio di raggio unitario delle "radici del rumore" o di  $\tilde{A}(z)$ .

Per la stazionarietà, quando  $N \rightarrow \infty$  tutte le radici di  $B(z)$  debbono trovarsi all'interno del cerchio unitario (chiuso) ma quando si ha a disposizione un numero di campioni finito questa proprietà non può sempre essere garantita e di conseguenza il metodo HOYW produce delle stime di frequenza distorte. Questo é un problema comune a tutti i metodi che stimano le frequenze dalle radici di un polinomio di grado superiore a  $n$ , come vedremo in seguito.

Per concludere discutiamo brevemente la scelta dei parametri  $m$  e  $l$ . In pratica é consigliabile usare  $l \approx m$  e scegliere i valori di questi interi in modo in modo che  $l + m$  sia all'incirca pari a  $1/3$  del numero di campioni osservati.

### Metodi MUSIC e di Pisarenko

Il metodo MUSIC (*MUltiple SIgnal Classification*) e quello di Pisarenko, che é un caso speciale del primo, come sarà spiegato in seguito, si basano sul modello di covarianza introdotto in (9.6.5) con  $m > 2\nu$  e in particolare sulla matrice di covarianza  $R$  che per comodità riscriviamo qui sotto:

$$R = \Omega P \Omega^\top + \sigma^2 I \tag{9.6.31}$$

Poiché la matrice  $\Omega P \Omega^\top$  ha rango  $2\nu$ , essa possiede  $2\nu$  autovalori, che denoteremo  $\{\tilde{\lambda}_k, k = 1, 2, \dots, 2\nu\}$ , strettamente positivi e i rimanenti  $(m - 2\nu)$  tutti uguali a zero. Denotiamo con  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$  gli autovalori di  $R$  e siano inoltre  $\{s_1, \dots, s_{2\nu}\}$  gli autovettori ortonormali associati a  $\{\lambda_1, \dots, \lambda_{2\nu}\}$  e  $\{g_1, \dots, g_{m-2\nu}\}$  quelli corrispondenti a  $\{\lambda_{2\nu+1}, \dots, \lambda_m\}$ .

Vale il seguente utile risultato.

**Lemma 9.1.** *Gli autovalori di  $R$  sono dati dalla relazione*

$$\lambda_k = \tilde{\lambda}_k + \sigma^2 \quad (k = 1, \dots, m)$$

dove  $\{\tilde{\lambda}_k\}_{k=1}^m$  sono gli autovalori di  $\Omega P \Omega^\top$  listati in ordine non crescente. L'insieme degli autovalori di  $R$  può così essere suddiviso in due sottoinsiemi:

$$\begin{cases} \lambda_k > \sigma^2 & k = 1, \dots, 2\nu \\ \lambda_k = \sigma^2 & k = 2\nu + 1, \dots, m \end{cases} \tag{9.6.32}$$

*Dimostrazione.* Per il teorema spettrale esiste una matrice ortogonale (di autovettori)  $T \in \mathbb{R}^{m \times m}$  tale che

$$\Omega P \Omega^\top = T \tilde{\Lambda} T^{-1}$$

dove  $\tilde{\Lambda} := \text{diag}\{\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_m\}$ . Il risultato scende dall'osservazione che il termine  $\sigma^2 I$  è diagonale nella stessa base, dato che si può scrivere  $\sigma^2 I = T \sigma^2 I T^{-1}$ .  $\square$

Gli autovettori associati a ognuno di questi sottoinsiemi possiedono alcune proprietà, qui di seguito riportate, che vengono usate nell'algoritmo per la stima

di frequenze. Siano

$$S = [s_1, \dots, s_{2\nu}] \quad (m \times n), \quad G = [g_1, \dots, g_{m-2\nu}] \quad (m \times (m - 2\nu)) \quad (9.6.33)$$

le matrici formate dagli autovettori ortonormali associati agli autovalori di  $R$ . Dalla definizione di  $R$  e da (9.6.32) si ottiene:

$$RG = G \begin{bmatrix} \lambda_{2\nu+1} & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{bmatrix} = \sigma^2 G = \Omega P \Omega^\top G + \sigma^2 G \quad (9.6.34)$$

L'ultima uguaglianza implica che  $\Omega P \Omega^\top G = 0$  ovvero, dato che la matrice  $AP$  ha rango di colonna pieno,

$$\Omega^\top G = 0 \quad (9.6.35)$$

cioè le colonne  $\{g_k\}$  di  $G$  appartengono allo spazio nullo di  $\Omega^\top$ .

L'equazione (9.6.35) sta all'abace dei metodi MUSIC. Scritta esplicitamente ha il seguente aspetto

$$\begin{bmatrix} 1 & \cos \omega_1 & \cos 2\omega_1 & \dots & \cos(m-1)\omega_1 \\ 0 & -\sin \omega_1 & \sin 2\omega_1 & \dots & \sin(m-1)\omega_1 \\ \dots & & & & \dots \\ \dots & & & & \dots \\ 1 & \cos \omega_n & \cos 2\omega_n & \dots & \cos(m-1)\omega_n \\ 0 & -\sin \omega_n & \sin 2\omega_n & \dots & \sin(m-1)\omega_n \end{bmatrix} [g_1 \quad \dots \quad g_{m-2\nu}] = 0$$

Notiamo adesso che questo sistema consiste di  $n$  blocchi di coppie di equazioni reali della forma

$$\begin{bmatrix} 1 & \cos \omega_k & \cos 2\omega_k & \dots & \cos(m-1)\omega_k \\ 0 & -\sin \omega_k & \sin 2\omega_k & \dots & \sin(m-1)\omega_k \end{bmatrix} g_j = 0 \quad k = 1, \dots, n \quad j = 1, \dots, m-2\nu$$

che si possono scrivere in forma complessa moltiplicando a sinistra per  $[1 \ i]$  nella forma,

$$[1 \ e^{i\omega_k} \ e^{i2\omega_k} \ \dots \ e^{i(m-1)\omega_k}] g_j = 0 \quad k = 1, \dots, n \quad j = 1, \dots, m-2\nu \quad (9.6.36)$$

che si può interpretare dicendo che

**Proposition 9.3.** *Gli  $m - 2\nu$  polinomi di grado  $m - 1$*

$$a_j(z) := [1 \ z \ z^2 \ \dots \ z^{m-1}] g_j = 0 \quad j = 1, \dots, m - 2\nu$$

*si annullano tutti nei punti  $z = e^{\pm i\omega_k}$ .*

Quindi le frequenze incognite  $\pm\omega_k$  si possono (in teoria) trovare calcolando gli zeri di modulo unitario di ciascun polinomio  $a_j(z)$ . Il fatto che si possa costruire un numero arbitrario  $(m - 2\nu)$  di polinomi  $a_j(z)$  può essere usato per migliorare

la stima delle frequenze incognite. Il metodo di Pisarenko [46] che è stato il primo metodo di questo tipo proposto in letteratura inizialmente costruiva solo un vettore  $g$  e quindi un solo polinomio  $a(z)$ .

In realtà, dato che

- la covarianza  $R$  dovrà essere stimata in base ai dati osservati e quindi sarà sempre affetta da rumore e quindi di rango pieno,
- le equazioni (9.6.36) sono quindi solo delle uguaglianze approssimate,
- il modello “vero” del segnale potrebbe essere più complesso di quello ipotizzato; in particolare potrebbe contenere rumore additivo arbitrario (non bianco) il che comporta autovalori del “rumore”  $\{\lambda_{2\nu+1}, \dots, \lambda_m\}$  tra loro diversi,
- il modello “vero” potrebbe poi avere un numero più grande di frequenze nella componente p.d. e quindi potrebbero esserci più zeri (approssimativamente) a modulo unitario degli  $n$  del modello ipotizzato,

Queste difficoltà richiedono una trattazione statistica del problema. In particolare è necessaria una regola per decidere quali sono gli autovalori “piccoli” ( $\{\lambda_{2\nu+1}, \dots, \lambda_m\}$ ) e approssimativamente uguali alla varianza  $\sigma^2$ .

È in particolare necessaria una trattazione statistica del sistema di equazioni  $a_j(z) = 0$ ,  $j = 1, 2, \dots, m - 2\nu$ . Questo problema è risolto in letteratura con vari “trucchi”, i più noti dei quali, *pseudospectrum MUSIC* e *Root MUSIC*, vengono descritti qui sotto.

Il metodo dello Pseudospectrum determina le stime di frequenza considerando la posizione dei  $\nu$  picchi più alti della funzione:

$$\frac{1}{a(z)^T g g^T a(z)}$$

sul cerchio unitario  $\{|z| = 1\}$ . Il metodo RootMusic determina invece le stime di frequenza come le posizioni angolari delle  $\nu$  coppie di radici più vicine al cerchio di raggio unitario dell'equazione

$$a(z)^T g g^T a(z) = 0.$$

## Il metodo ESPRIT

Questo metodo è basato su una proprietà della matrice di osservabilità di un sistema lineare che in letteratura è chiamata qualche volta *shift-invariance*. Ritroveremo questa idea più in dettaglio quando studieremo i metodi a sottospazi.

Sia

$$\Omega := \begin{bmatrix} C \\ CA \\ \vdots \\ CA^m \end{bmatrix}$$

la matrice di osservabilità “estesa” ( $m \geq n$ ) di un sistema lineare di ordine  $n$

$$\begin{cases} \mathbf{x}(t+1) = A \mathbf{x}(t) \\ \mathbf{y}(t) = C \mathbf{x}(t) \end{cases}$$

che assumeremo osservabile. Introducendo le matrici “traslate”

$$\uparrow \Omega := \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{m-1} \end{bmatrix} \quad \downarrow \Omega := \begin{bmatrix} CA \\ CA^2 \\ \vdots \\ CA^m \end{bmatrix} \quad (9.6.37)$$

evidentemente si ha

$$(\downarrow \Omega) = (\uparrow \Omega) A \quad (9.6.38)$$

e, dato che  $\uparrow \Omega$  ha le colonne linearmente indipendenti (per l’osservabilità), esiste un’unica matrice  $A \in \mathbb{R}^{n \times n}$  che soddisfa l’equazione (9.6.38). In altri termini, la matrice  $A$  è univocamente determinata dalla matrice di osservabilità estesa del sistema. Su questa idea è basato il metodo cosiddetto ESPRIT per la stima delle  $N$  frequenze incognite del segnale.

Consideriamo il nostro segnale (9.1.12) con rumore additivo bianco, la cui dinamica è descritta dal sistema lineare osservabile (9.1.6) di ordine  $2\nu$ . Un modo numericamente affidabile per calcolare una base di vettori per lo spazio immagine della matrice di osservabilità è di ricondursi all’espressione (9.6.5) della covarianza di  $\mathbf{y}^m$ .

**Lemma 9.2.** *La matrice di autovettori  $S$  definita in (9.6.33) è una base per lo spazio immagine di  $\Omega$ , in formule*

$$\text{Im} \{S\} = \text{Im} \{\Omega\} \quad (9.6.39)$$

*Dimostrazione.* Abbiamo sostanzialmente già visto che  $\Omega^\top G = 0$  e quindi le colonne della matrice  $G$  sono una base per lo spazio nullo di  $\Omega^\top$ , ovvero  $\text{Ker} \{\Omega^\top\} = \text{Im} \{G\}$  e quindi, dato che per una arbitraria matrice reale  $M$  vale la  $\text{Ker} \{M^\top\} = \text{Im} \{M\}^\perp$ , prendendo il complemento ortogonale in  $\mathbb{R}^m$ , si trova

$$\text{Im} \{\Omega\} = ((\text{Im} \{\Omega\})^\perp)^\perp = (\text{Im} \{G\})^\perp = \text{Im} \{S\}$$

dato che  $\text{Im} \{G\} = \text{Im} \{S\}^\perp$ . □

Notiamo adesso che, data l’uguaglianza (9.6.39), deve esistere una matrice invertibile  $T \in \mathbb{R}^{n \times n}$  tale che  $S = \Omega T$ . Ne viene che la relazione (9.6.38) vale anche per le matrici traslate  $\uparrow S$  e  $\downarrow S$  definite in modo analogo alle  $\downarrow \Omega$ ,  $\uparrow \Omega$ . Insomma,

$$(\downarrow S) = (\uparrow S) \hat{A} \quad (9.6.40)$$

dove  $\hat{A} := T^{-1}AT$ .

In conclusione, risolvendo il sistema (9.6.40)(che è in generale sovradeterminato) si può stimare una matrice simile alla  $A$  e quindi ricavare le frequenze incognite dagli autovalori, che in teoria dovrebbero stare tutti sul cerchio unitario. Un

possibile metodo di soluzione è mediante i minimi quadrati. La soluzione calcolata con le equazioni normali (di solo uso concettuale) è

$$\hat{A} = [(\uparrow S)^\top (\uparrow S)]^{-1} (\uparrow S)^\top (\downarrow S).$$

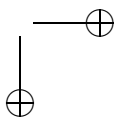
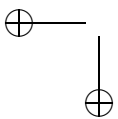
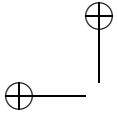
Una volta stimata  $A$ , le frequenze di oscillazione si trovano prendendo gli  $n$  autovalori di  $\hat{A}$  più vicini al cerchio unitario.

### Considerazioni sulla bontà statistica dei metodi a correlazione

In generale si può affermare che se la correlazione campionaria converge con probabilità uno a quella vera (il che accade in particolare con segnali ergodici del secondo ordine) le stime coi metodi basati sulla correlazione campionaria convergono ai parametri "veri", quelli che soddisfano le relative relazioni limite che coinvolgono la covarianza vera. Sotto queste ipotesi i metodi basati sulla correlazione campionaria producono quindi in generale stimatori consistenti.

In realtà i segnali quasi periodici non sono ergodici (nemmeno del secondo ordine) e la correlazione limite per  $N \rightarrow \infty$  dipende dall'ampiezza delle componenti periodiche. La linearità dei modelli fa sì che i parametri "veri" del modello in gioco, al limite soddisfino delle relazioni lineari che sono le stesse del caso di segnali ergodici. Quindi, anche se la covarianza limite è a stretto rigore aleatoria la relazione lineare è la stessa e il ragionamento continua a valere.

Viceversa è molto più difficile dare delle espressioni utili per la distribuzione e la varianza asintotica di questi stimatori. La letteratura (specialmente quella ingegneristica) sorvola su questi punti.



## CHAPTER 10

VALIDAZIONE E STIMA  
DELLA STRUTTURA

Finora abbiamo trattato l'identificazione di modelli di sistemi stocastici lineari assumendo sempre che gli ordini dei vari polinomi che definiscono la classe dei modelli in gioco siano noti. In questo contesto, come abbiamo visto, l'identificazione si riduce ad un puro problema di stima parametrica. Di fatto gli ordini, che chiameremo per semplicità *la struttura* della classe di modelli su cui si va a fare l'identificazione, non è quasi mai nota e una parte importante del problema dell'identificazione è proprio quella di identificare la struttura. In questo capitolo tratteremo appunto di alcune tecniche per risolvere questo problema.

Dovremo in seguito parametrizzare la classe di modelli che consideriamo non solo mediante il parametro  $\theta$  ma anche mediante un *multi-indice di struttura* (tipicamente gli ordini dei polinomi nei modelli a scatola nera) che denoteremo col simbolo  $\nu$ . Ad esempio per modelli del tipo Box-Jenkins, il multi-indice di struttura sarà  $\nu = (n, m, q, r)$ , in corrispondenza al quale il parametro avrà dimensione  $p_\nu = n + m + q + r$ , dipendente dalla struttura. Indicheremo in generale la classe di modelli (di innovazione) di struttura fissata  $\nu$  col simbolo

$$\mathcal{M}_\nu = \{M_\nu(\theta); \theta \in \Theta_\nu\}; \quad \Theta_\nu \subseteq \mathbb{R}^{p_\nu}$$

Qualche volta diremo anche che la classe  $\mathcal{M}_\nu$  ha *complessità*  $\nu$ , finita. Notare che in generale un processo stazionario è descritto da un modello non razionale, che, in un certo senso, ha complessità infinita, per cui l'esistenza effettiva di un modello "vero" di complessità finita è in generale una finzione matematica la quale, anche se talvolta utile, va trattata come tale.

L'idea fondamentale per l'identificazione della struttura è quella di confrontare tra loro modelli identificati di struttura diversa. Supponendo di voler considerare solo un numero finito di possibili strutture da confrontare, il problema si può naturalmente formulare come un problema di verifica di ipotesi (multiple). Nel contesto Fisheriano, si formula spesso il problema come la verifica di un'ipotesi privilegiata ( $H_0$ ) che corrisponde alla struttura vera,  $\nu_0$ , del modello vero. Questo naturalmente è da interpretare *cum grano salis* come abbiamo a suo tempo fatto per la nozione di consistenza di uno stimatore parametrico. Purtroppo in generale non

è possibile individuare statistiche ottimali per questi tests e si ricorre a certi indici di accuratezza che quantificano quanto bene un modello descrive i dati. Questi indici sono tipicamente la bianchezza dell'errore residuo di predizione, la correlazione tra residui e ingresso, l'errore finale di predizione e altri che descriveremo in dettaglio più avanti.

## 10.1 Tests di bianchezza dei residui

Riprendiamo la rappresentazione (6.1.3) di un processo stazionario  $\mathbf{y}$  mediante un modello generico di una classe parametrica, pilotato dal relativo errore di predizione. Riscriviamo questa rappresentazione introducendo esplicitamente l'indice di struttura, nella forma

$$\mathbf{y}(t) = F_{\nu, \theta}(z)\mathbf{u}(t) + G_{\nu, \theta}(z)\varepsilon_{\nu, \theta}(t), \quad \theta \in \Theta_{\nu}. \quad (10.1.1)$$

Ricordiamo che in questa rappresentazione  $\varepsilon_{\nu, \theta}(t)$  è il processo errore di predizione associato al predittore lineare a minima varianza costruito a partire dal modello. Il risultato seguente sta alla base di un importante criterio di validazione.

**Proposition 10.1.** *Se per qualche  $(\nu_0, \theta_0)$  l'errore di predizione  $\varepsilon_{\nu_0, \theta_0}(t)$  è rumore bianco, allora il processo congiunto  $\{\mathbf{y}, \mathbf{u}\}$  è descritto dal modello vero*

$$\mathbf{y}(t) = F_{\nu_0, \theta_0}(z)\mathbf{u}(t) + G_{\nu_0, \theta_0}(z)\mathbf{e}_0(t), \quad (10.1.2)$$

dove  $\mathbf{e}_0(t) \equiv \varepsilon_{\nu_0, \theta_0}(t)$ . Viceversa, se il processo è descritto da un modello (vero) di complessità finita come il (10.1.2), allora l'errore di predizione del modello coincide con l'innovazione  $\mathbf{e}_0$ .

**Proof.** Di fatto, dato che il modello d'innovazione, e in particolare, il processo d'innovazione della coppia  $\{\mathbf{y}, \mathbf{u}\}$  sono unici, se  $\varepsilon_{\nu_0, \theta_0}(t)$  è rumore bianco, allora il modello (10.1.1) corrispondente è il modello d'innovazione di  $\{\mathbf{y}, \mathbf{u}\}$ . Viceversa, se il processo è descritto dal modello vero (10.1.2), allora  $\mathbf{e}_0(t)$  è anche l'errore di predizione corrispondente ai parametri  $\nu_0, \theta_0$ .  $\square$

L'enunciato del Teorema di consistenza 6.2 si può interpretare nel presente contesto nel modo seguente.

**Corollary 10.1.** *Se il processo (vero) che genera i dati è ergodico del secondo ordine ed è descritto da un modello che appartiene alla classe parametrica  $\mathcal{M}_{\nu_0}$  e la classe parametrica dei modelli  $\mathcal{M}_{\nu_0}$  è identificabile localmente in  $\theta = \theta_0$ , allora lo stimatore PEM  $\hat{\theta}_N$  è consistente e converge per  $N \rightarrow \infty$ , al parametro vero  $\theta_0$  con probabilità uno.*

Supponiamo ora che il modello  $M_{\nu}(\theta)$  sia stato identificato con il metodo PEM sulla base di dati osservati di numerosità  $N$  e denotiamo con  $\varepsilon_{\nu, \hat{\theta}_N}(t)$  l'errore residuo di predizione di un modello di struttura  $\nu$  basato su un campione di numerosità  $N$ .



**Proposition 10.2.** *Se  $\hat{\theta}_N$  è uno stimatore consistente di  $\theta$  per il modello (10.1.1) e se il limite, che esiste con probabilità uno,*

$$\lim_{N \rightarrow \infty} \varepsilon_{\nu, \hat{\theta}_N}(t) := \varepsilon_{\nu, \theta_0}(t) \tag{10.1.3}$$

*è rumore bianco, allora  $\nu \equiv \nu_0$  è la struttura vera e  $\varepsilon_{\nu, \theta_0}(t) = e_0(t)$ .*

**Proof.** Infatti, per la continuità dell'errore di predizione rispetto al parametro  $\theta$  del modello, il processo limite  $\varepsilon_{\nu, \theta_0}(t)$  è l'errore di predizione del modello  $M_{\nu}(\theta_0)$  e allora per concludere basta ricordare l'enunciato della proposizione 10.1.  $\square$

### 10.1.1 Il Test del Correlogramma

Questo test usa la covarianza campionaria dell'errore residuo di predizione del modello stimato. Denotando per semplicità  $\varepsilon_{\nu, \hat{\theta}_N}(t)$  come  $\hat{\varepsilon}(t)$ , si costruisce la covarianza campionaria

$$\hat{\lambda}(\tau) := \frac{1}{N} \sum_{t=\tau}^N \hat{\varepsilon}(t) \hat{\varepsilon}(t - \tau) \tag{10.1.4}$$

dove si prende  $0 \leq \tau \leq \tau_{Max}$ . Il valore di  $\tau_{Max}$  è scelto opportunamente piccolo, tipicamente pari ad  $1/20 \sim 1/50$  di  $N$  per evitare effetti ai bordi della varianza della stima, [?].

L'ipotesi  $H_0$  da testare è che  $\nu = \nu_0$ . Naturalmente manteniamo ferme le ipotesi che assicurano la consistenza dello stimatore PEM (Corollario 10.1). Sotto  $H_0$  si ha allora

$$\lim_{N \rightarrow \infty} \hat{\lambda}(\tau) = 0, \quad \text{for } \tau \neq 0 \tag{10.1.5}$$

$$\lim_{N \rightarrow \infty} \hat{\lambda}(0) = \lambda_0^2, \quad \text{for } \tau = 0. \tag{10.1.6}$$

con probabilità uno.

**Proposition 10.3.** *Si consideri la statistica a valori vettoriali*

$$\hat{\mathbf{i}} := \frac{1}{N} \sum_{t=m}^N \hat{\varepsilon}(t) \begin{bmatrix} \hat{\varepsilon}(t-1) \\ \vdots \\ \hat{\varepsilon}(t-m) \end{bmatrix} = \begin{bmatrix} \hat{\lambda}(1) \\ \vdots \\ \hat{\lambda}(m) \end{bmatrix} \tag{10.1.7}$$

e la sua versione normalizzata  $\hat{\mathbf{r}} := \frac{1}{\hat{\lambda}(0)} \hat{\mathbf{i}}$ . Si ha

$$N \hat{\mathbf{r}}^\top \hat{\mathbf{r}} = N \frac{1}{\hat{\lambda}(0)} \hat{\mathbf{i}}^\top \hat{\mathbf{i}} \xrightarrow{L} \chi^2(m) \tag{10.1.8}$$

**Proof.** Notiamo che per il modello vero i prodotti  $\mathbf{e}_0(t)\mathbf{e}_0(t-k)$ ;  $k = 1, \dots, m$  sono d-martingale stazionarie per cui vale il teorema del limite centrale,

$$\sqrt{N} \frac{1}{N} \sum_{t=m}^N \mathbf{e}_0(t) \begin{bmatrix} \mathbf{e}_0(t-1) \\ \vdots \\ \mathbf{e}_0(t-m) \end{bmatrix} \xrightarrow{L} \mathcal{N}(0, P).$$

Dato che

$$\mathbb{E}(\mathbf{e}_0(t)\mathbf{e}_0(t-k))^2 = \lambda_0^4, \quad \mathbb{E}(\mathbf{e}_0(t)\mathbf{e}_0(t-k))(\mathbf{e}_0(s)\mathbf{e}_0(s-k)) = 0 \quad \text{per } t \neq s$$

si trova

$$P = \text{Var} \left\{ \mathbf{e}_0(t) \begin{bmatrix} \mathbf{e}_0(t-1) \\ \vdots \\ \mathbf{e}_0(t-m) \end{bmatrix} \right\} = \lambda_0^4 I_m. \quad (10.1.9)$$

Ora, sotto l'ipotesi  $H_0$  si ha

$$\hat{\mathbf{e}}(t)\hat{\mathbf{e}}(t-k) \rightarrow \mathbf{e}_0(t)\mathbf{e}_0(t-k)$$

con probabilità uno. Si ha così

$$\sqrt{N} \left\{ \frac{1}{N} \sum_{t=m}^N \hat{\mathbf{e}}(t) \begin{bmatrix} \hat{\mathbf{e}}(t-1) \\ \vdots \\ \hat{\mathbf{e}}(t-m) \end{bmatrix} - \frac{1}{N} \sum_{t=m}^N \mathbf{e}_0(t) \begin{bmatrix} \mathbf{e}_0(t-1) \\ \vdots \\ \mathbf{e}_0(t-m) \end{bmatrix} \right\} \rightarrow 0$$

in legge, dato che la differenza dentro parentesi tende a zero con probabilità uno e il secondo termine moltiplicato per  $\sqrt{N}$  converge in legge per cui lo stesso deve accadere anche al primo termine [CHECK !].

Per il primo enunciato del teorema di Slutsky (Teorema 5.1), la statistica (10.1.7) converge allora in legge alla distribuzione Gaussiana  $\mathcal{N}(0, \lambda_0^4 I_m)$ .

Dato che  $\lambda_0^2$  è incognita conviene normalizzare la (10.1.7), introducendo il vettore dei coefficienti di correlazione campionari  $\hat{r}(k) = \hat{\lambda}(k)/\hat{\lambda}(0)$ . Usando ancora la (10.1.6) e il teorema di Slutsky si vede che per  $N \rightarrow \infty$ ,

$$\hat{r}(k) \xrightarrow{L} \mathcal{N}(0, 1)$$

e quindi anche

$$\hat{\mathbf{r}} := \frac{1}{\hat{\lambda}(0)} \hat{\mathbf{I}} \xrightarrow{L} \mathcal{N}(0, I_m). \quad (10.1.10)$$

Questo risultato implica che

$$N \hat{\mathbf{r}}^\top \hat{\mathbf{r}} = N \frac{1}{\hat{\lambda}(0)} \hat{\mathbf{I}}^\top \hat{\mathbf{I}} \xrightarrow{L} \chi^2(m)$$

che è quanto si voleva dimostrare.  $\square$

La statistica:  $N$  volte la norma del vettore dei coefficienti di correlazione campionari,  $N \hat{\mathbf{r}}^\top \hat{\mathbf{r}}$  può quindi essere usata come statistica di test. Si rifiuta l'ipotesi che  $\nu = \nu_0$  con probabilità d'errore di prima specie  $\alpha$  se i valori della statistica sono più grandi di  $k_\alpha$  dove  $P_{\chi^2(m)}\{x \geq k_\alpha\} = \alpha$ . Il valore di  $m$  si prende normalmente pari a  $m = 5 \sim 10$ .

### 10.1.2 Un test di incorrelazione dagli ingressi passati

[DA SCRIVERE]

### 10.1.3 Il test del periodogramma cumulato

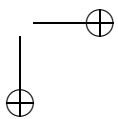
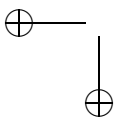
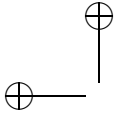
[DA SCRIVERE]

### 10.1.4 Il test $F$ per modelli lineari identificati col metodo PEM

[DA SCRIVERE]

## 10.2 Stima dell'ordine

[DA SCRIVERE]



## APPENDIX A

## Appendix

## A.1 Alcuni richiami di teoria della probabilità

A.1.1 Sulla nozione di  $\sigma$ -algebra

Come è ben noto il concetto fondamentale nella teoria della probabilità è quello di *spazio di probabilità*  $\{\Omega, \mathcal{A}, P\}$  in cui  $\Omega$  è un insieme astratto (lo spazio degli eventi elementari),  $\mathcal{A}$  è una  $\sigma$ -algebra (Booleana) di sottoinsiemi di  $\Omega$  (gli *eventi*) e  $P$  è una misura di probabilità su  $\mathcal{A}$ . Ricordiamo che la naturale struttura di algebre Booleana rispetto alle operazioni insiemistiche di intersezione unione e complementazione, della famiglia di tutti sottoinsiemi di  $\Omega$ , che sarebbe il naturale modello matematico per tutte le proposizioni logiche che si possono costruire mediante eventi elementari di  $\Omega$ , è in generale troppo ricca e deve essere ristretta richiedendo che  $\mathcal{A}$  sia chiusa rispetto alle operazioni di unione (o intersezione) *numerabile* di insiemi. Il che definisce appunto il concetto di  $\sigma$ -algebra. La  $\sigma$ -algebra “naturale” negli spazi Euclidei  $\mathbb{R}$  o  $\mathbb{C}$  è quella indotta (o “generata”) dagli insiemi aperti. Si chiama  *$\sigma$ -algebra di Borel* e si indica col simbolo  $\mathcal{B}$ .

Una variabile aleatoria (reale o complessa)  $x$  è una funzione definita su  $\Omega$  che è *misurabile* rispetto a  $\mathcal{A}$ , o in breve,  *$\mathcal{A}$ -misurabile*, il che significa che le antiimmagini degli insiemi di Borel appartengono a  $\mathcal{A}$ ; i.e.

$$x^{-1}(E) \in \mathcal{A}, \quad \forall E \in \mathcal{B}.$$

La  $\sigma$ -algebra *indotta* (o *generata*) dalla variabile  $x$  è la famiglia di tutti i sottoinsiemi di  $\Omega$  che sono antiimmagini di insiemi di Borel

$$\mathcal{X} := \{x^{-1}(E); E \in \mathcal{B}\}.$$

Si dimostra facilmente che se (e solo se)  $x$  è misurabile,  $\mathcal{X}$  è in effetti una sub- $\sigma$ -algebra di  $\mathcal{A}$ . Questo concetto si può generalizzare alla  $\sigma$ -algebra indotta (o generata) da una famiglia (numerabile o non) di variabili aleatorie  $\{x_\alpha; \alpha \in A\}$ , definendo  $\mathcal{X}$  come la più piccola  $\sigma$ -algebra contenente la famiglia di eventi  $\{x_\alpha^{-1}(E); E \in \mathcal{B}\}$ .

$\mathcal{B}, \alpha \in A$ . Si usa spesso la notazione

$$\mathcal{X} = \sigma \{ \mathbf{x}_\alpha ; \alpha \in A \}.$$

Da notare che ogni  $\sigma$ -algebra si può pensare indotta da una famiglia di variabili aleatorie reali, ad esempio dalla famiglia delle funzioni indicatrici  $\{I_A ; A \in \mathcal{A}\}$ , dove

$$I_A(\omega) = \begin{cases} = 1 & \text{se } \omega \in A \\ = 0 & \text{se } \omega \notin A \end{cases}$$

Nelle applicazioni, si ha normalmente a che fare con una  $\sigma$ -algebra  $\mathcal{A}$  che è stata indotta da una certa famiglia di variabili aleatorie, ad esempio da certe variabili di un processo stocastico. In questo caso, dire che una variabile aleatoria  $\mathbf{x}$  è  $\mathcal{A}$ -misurabile ha un significato molto concreto; significa semplicemente dire che  $\mathbf{x}$  è *funzione delle variabili che hanno generato  $\mathcal{A}$* . In termini più precisi, vale il risultato seguente.

**Theorem A.1.** *Se  $\mathcal{A}$  è generata da una famiglia numerabile di variabili aleatorie; i.e.*

$$\mathcal{A} = \sigma \{ \mathbf{y}_k ; k \in \mathbb{Z} \}$$

per ogni variabile  $\mathcal{A}$ -misurabile  $\mathbf{x}$ , esiste una funzione  $\mathcal{B}^\infty$ -misurabile,  $\varphi : \mathbb{R}^{\mathbb{Z}} \rightarrow \mathbb{R}$ , tale che

$$\mathbf{x} = \varphi(\{ \mathbf{y}_k ; k \in \mathbb{Z} \}) \tag{A.1.1}$$

Per la prova si può vedere il testo di Shiryaev [58, p. 174].

Data una  $\sigma$ -algebra  $\mathcal{F}$ , generata da una famiglia di variabili aleatorie reali  $\{f_\alpha ; \alpha \in A\}$  possiamo quindi pensare alle variabili  $\mathcal{F}$ -misurabili semplicemente come funzioni delle funzioni appartenenti all'aggregato  $\{f_\alpha ; \alpha \in A\}$ . Questa idea "operativa" del concetto di misurabilità ripetuto ad una  $\sigma$ -algebra è spesso utile per interpretare nozioni astratte della teoria della probabilità.

## A.2 Hilbert space of second-order random variables

A random variable is just a measurable function defined on some underlying probability space  $\{\Omega, \mathcal{A}, P\}$ . The symbol  $\mathbb{E} \{ \mathbf{x} \} := \int_\Omega \mathbf{x} dP$  denotes mathematical expectation of the random variable  $\mathbf{x}$ . Random variables which have finite second moment,  $\mathbb{E} \{ |\mathbf{x}|^2 \} < \infty$ , are commonly called *second order* random variables.

The set of real or complex-valued second-order random variables  $\mathbf{x}$  defined on the same probability space  $\{\Omega, \mathcal{A}, P\}$  is obviously a linear vector space under the usual operations of sum and multiplication by real (or complex) numbers. This vector space comes naturally equipped with an inner product

$$\langle \mathbf{x}, \mathbf{z} \rangle = \mathbb{E} \mathbf{x} \bar{\mathbf{z}},$$

which is just the correlation of the random variables  $\mathbf{x}, \mathbf{z}$ . Note that the norm  $\| \mathbf{x} \| = \langle \mathbf{x}, \mathbf{x} \rangle^{1/2}$  induced by this inner product (the square root of the second moment of

$x$ ) is positive, i.e.  $\|x\| = 0 \Leftrightarrow x = 0$ , only if we agree to identify random variables which are equal almost surely, i.e. differ on a set of probability zero. Consider the set of equivalence classes of second-order random variables  $x$  with respect to almost sure equality. This set, once equipped with the inner product  $\langle \cdot, \cdot \rangle$ , becomes an inner product space, denoted  $L^2(\Omega, \mathcal{A}, P)$ . Convergence with respect to the norm of this space is called *convergence in mean square*. It is a very well-known fact that  $L^2(\Omega, \mathcal{A}, P)$  is actually closed with respect to convergence in mean square and is therefore a Hilbert space.

### Notations and conventions

In this book the term *subspace* of a Hilbert space  $\mathbf{H}$ , will in general mean *closed* subspace. For finite-dimensional vectors,  $|v|$  will denote Euclidean norm (or absolute value in the scalar case).

The *sum* of two linear vector spaces  $\mathbf{X} + \mathbf{Y}$ , is, by definition, the linear vector space  $\{x + y \mid x \in \mathbf{X}, y \in \mathbf{Y}\}$ . Even when  $\mathbf{X}$  and  $\mathbf{Y}$  are both (closed) subspaces, this linear manifold may fail to be closed. The (*closed*) *vector sum* of  $\mathbf{X}$  and  $\mathbf{Y}$ , denoted  $\mathbf{X} \vee \mathbf{Y}$ , is the closure of  $\mathbf{X} + \mathbf{Y}$ .

In this book, the symbols  $+$ ,  $\vee$ ,  $\dot{+}$  and  $\oplus$  will denote sum, (closed) vector sum, direct sum, and *orthogonal* direct sum of subspaces. The symbol  $\mathbf{X}^\perp$  denotes the orthogonal complement of the subspace  $\mathbf{X}$  with respect to some predefined ambient space. The *linear vector space generated by a family of elements*  $\{x_\alpha\}_{\alpha \in \mathbb{A}} \subset \mathbf{H}$ , denoted  $\text{span}\{x_\alpha \mid \alpha \in \mathbb{A}\}$ , is the vector space whose elements are all finite linear combinations of the *generators*  $\{x_\alpha\}$ . The *subspace generated by the family*  $\{x_\alpha\}_{\alpha \in \mathbb{A}}$  is the closure of this linear vector space and is denoted by  $\overline{\text{span}}\{x_\alpha \mid \alpha \in \mathbb{A}\}$ .

### A.3 Proiezioni ortogonali e media condizionata

Si consideri il seguente problema: una variabile aleatoria del second'ordine  $x$ , i cui valori non sono accessibili all'ossrvazione, può essere misurata indirettamente mediante un qualche strumento di misura. Lo strumento produce una succesione di osservazioni che modelliamo come una traiettoria di un processo stocastico reale  $y = \{y(t) \mid t \in \mathbb{T}\}$ , definito sullo stesso spazio di probabilità di  $x$ . Per il momento l'insieme temporale  $\mathbb{T}$  potrebbe essere qualunque, i reali gli interi, o un sottointervallo finito o infinito dei medesimi.

Dalla traiettoria osservata vogliamo ottenere la "migliore ricostruzione possibile" (in qualche senso da definirsi) del valore campionario incognito  $x$  di  $\mathbf{x}$ . Questo significa che vogliamo trovare una funzione delle osservazioni (uno *stimate*),  $\varphi(y)$ , che produce "in media" il più piccolo errore di stima,  $x - \varphi(y)$ . Dato che sia  $x$  che le componenti di  $y$  si assumono con momenti del second'ordine finiti, possiamo pensarli come elementi dello spazio di Hilbert  $L^2(\Omega, \mathcal{A}, P)$  e richiedere che anche  $\varphi(y)$  abbia momenti del second'ordine finiti e quindi appartenga allo stesso spazio.

Ora, le funzioni a quadrato sommabile del processo  $y$  formano un sottospazio vettoriale chiuso di  $L^2(\Omega, \mathcal{A}, P)$  di funzioni misurabili rispetto alla  $\sigma$ -algebra  $\mathcal{Y} \subset$

$\mathcal{A}$ , generata dalle variabili del processo  $\mathbf{y}$ , che indichiamo con  $L^2(\Omega, \mathcal{Y}, P)$ <sup>34</sup>. Ogni stimatore ammissibile  $\varphi(\mathbf{y})$  sarà un elemento del sottospazio  $L^2(\Omega, \mathcal{Y}, P)$  dello spazio di Hilbert  $L^2(\Omega, \mathcal{A}, P)$ .

È quindi naturale formulare il nostro problema nel modo seguente: trovare una variabile aleatoria  $\mathbf{z} = \varphi(\mathbf{y})$  in  $L^2(\Omega, \mathcal{Y}, P)$ , per cui l'errore di stima  $\mathbf{x} - \mathbf{z}$  ha la più piccola possibile norma  $L^2$ ; in altri termini risolve il seguente problema di ottimo

$$\min_{\mathbf{z} \in L^2(\Omega, \mathcal{Y}, P)} \|\mathbf{x} - \mathbf{z}\| \tag{A.3.1}$$

dove  $\|\mathbf{x} - \mathbf{z}\|^2 = \mathbb{E}\{|\mathbf{x} - \mathbf{z}|^2\}$ . È immediato convincersi che una condizione necessaria perché  $\mathbf{z}$  sia una soluzione del problema è che  $\mathbb{E}\mathbf{z} = \mathbb{E}\mathbf{x}$ . Ne segue che  $\mathbb{E}\{|\mathbf{x} - \mathbf{z}|^2\}$  è interpretabile come la varianza dell'errore di stima e il nostro problema si può descrivere come la ricerca dello stimatore, funzione dei dati  $\mathbf{y}$ , che ha la *minima varianza d'errore*.

È ben noto che questo problema ha un'unica soluzione. Ricordiamo a questo proposito il seguente risultato.

**Lemma A.1.** *Sia  $\mathbf{Y}$  un sottospazio chiuso di uno spazio di Hilbert space  $\mathbf{H}$ . Dato un elemento  $\mathbf{x} \in \mathbf{H}$ , l'elemento  $\mathbf{z} \in \mathbf{Y}$  che ha la minima distanza da  $\mathbf{x}$ ; i.e. minimizza  $\|\mathbf{x} - \mathbf{z}\|$  è unico ed è la proiezione ortogonale di  $\mathbf{x}$  su  $\mathbf{Y}$ .*

*Condizione necessaria e sufficiente affinché  $\mathbf{z}$  sia la proiezione ortogonale di  $\mathbf{x}$  su  $\mathbf{Y}$  è che  $\mathbf{x} - \mathbf{z} \perp \mathbf{Y}$ , equivalentemente, che per un arbitrario sistema di generatori  $\{\mathbf{y}_\alpha; \alpha \in A\}$  di  $\mathbf{Y}$  valga la*

$$\langle \mathbf{x} - \mathbf{z}, \mathbf{y}_\alpha \rangle = 0, \quad \alpha \in A \tag{A.3.2}$$

(*principio di ortogonalità*).

Il miglior stimatore di  $\mathbf{x}$  nel senso appena descritto è quindi la *proiezione ortogonale* di  $\mathbf{x}$  sul sottospazio  $L^2(\Omega, \mathcal{Y}, P)$ , delle funzioni a quadrato sommabile dei dati di misura  $f(\mathbf{y})$ . Questa proiezione si denota

$$\mathbf{z} = \mathbb{E}[\mathbf{x} | \mathcal{Y}] \equiv \mathbb{E}[\mathbf{x} | \mathcal{Y}]$$

ed è chiamata *media (o aspettazione) condizionata* della variabile  $\mathbf{x}$ , dato l'aggregato  $\mathcal{Y} = \{\mathbf{y}(t) | t \in \mathbb{T}\}$ , o, equivalentemente, data la  $\sigma$ -algebra  $\mathcal{Y}$  generata da  $\mathbf{y}$ .

Se si prendono come generatori di  $L^2(\Omega, \mathcal{Y}, P)$  le funzioni indicatrici  $\{I_A, A \in \mathcal{Y}\}$ , possiamo riscrivere la relazione (A.3.2) come

$$\mathbb{E}\{\mathbf{x}I_A\} = \mathbb{E}\{\mathbf{z}I_A\}, \quad \forall A \in \mathcal{Y}$$

che è la definizione usuale che si trova nei testi di probabilità.

Sfortunatamente, ad eccezione di un numero limitato di casi (incluso il caso Gaussiano) la media condizionate è praticamente impossibile da calcolare e si ricorre alla più maneggevole approssimazione lineare che è individuata dai soli momenti congiunti del primo e secondo ordine. Vedere ad es. [44].

<sup>34</sup>È importante notare che la misura di probabilità  $P$  è definita su una  $\sigma$ -algebra più piccola di  $\mathcal{A}$  e andrebbe a rigore denotata con un simbolo diverso, ad esempio  $P_{\mathcal{Y}}$ . Noi per semplicità useremo ancora la stessa notazione  $P$ .



### Wide-sense theory and linear estimation

It will be convenient to subtract off the expected values from all random quantities involved (which will henceforth assumed to have zero-mean). Let

$$\mathbf{H}(y) := \text{span} \{y_k \mid k = 1, \dots, m\}$$

be the (finite dimensional) subspace of  $L^2(\Omega, \mathcal{A}, P)$  linearly generated by the components of  $\mathbf{y}$ . The *best linear estimator* of  $\mathbf{x}$  based on (or given)  $\mathbf{y}$ , is the  $n$ -dimensional random vector  $\hat{\mathbf{x}}$ , whose components  $\hat{x}_k \in \mathbf{H}(y)$ ,  $k = 1, \dots, n$ , individually solve the minimum problems

$$\min_{z_k \in \mathbf{H}(y)} \|\mathbf{x}_k - z_k\| \quad k = 1, \dots, n, \tag{A.3.3}$$

In view of Lemma A.1,  $\hat{x}_k$  is just the orthogonal projection of  $\mathbf{x}_k$  onto  $\mathbf{H}(y)$ . According to our previous conventions, we shall denote this projection by the symbols

$$\mathbb{E} [\mathbf{x}_k \mid \mathbf{H}(y)] \quad \text{or} \quad \mathbb{E}^{\mathbf{H}(y)} \mathbf{x}_k$$

The notation  $\mathbb{E} [\mathbf{x} \mid \mathbf{Y}]$  will be used also when  $\mathbf{x}$  is vector-valued. The symbol will then just denote the vector with components  $\mathbb{E} [x_k \mid \mathbf{Y}]$ ,  $k = 1, \dots, n$ . When the projection is expressed in terms of some specific set of generators say  $\mathbf{y} = \{y_\alpha\}$  (i.e.  $\mathbf{Y} = \overline{\text{span}}\{y_\alpha\}$ ), we shall denote it  $\mathbb{E} [\mathbf{x} \mid \mathbf{y}]$ .

**Proposition A.1.** *Let  $\mathbf{x}$  and  $\mathbf{y}$  be zero-mean second-order random vectors of dimensions  $n$  and  $m$  respectively with covariance matrix*

$$\Sigma = \mathbb{E} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}^\top = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix} \tag{A.3.4}$$

*Then the orthogonal projection (minimum variance linear estimator) of  $x$  onto the linear subspace spanned by the components of  $y$  is given by*

$$\mathbb{E} [x \mid y] = \Sigma_{xy} \Sigma_y^\dagger y \tag{A.3.5}$$

*where  $\dagger$  denotes the Moore-Penrose pseudoinverse.<sup>35</sup> The (residual) error vector has covariance matrix,*

$$\Lambda := \text{Var} (x - \mathbb{E} [x \mid y]) = \Sigma_x - \Sigma_{xy} \Sigma_y^\dagger \Sigma_{yx}. \tag{A.3.6}$$

*This is the smallest error covariance matrix obtainable in the class of all linear functions of the data, i.e.  $\Lambda \leq \text{Var} (x - Ay)$  for any matrix  $A \in \mathbb{R}^{n \times m}$ , where the inequality is understood in the sense of the positive semidefinite ordering among symmetric matrices.*

**Proof.** Writing the vector  $z$  as  $z = Ay$ , and invoking the orthogonality condition (A.3.2) for each component  $x_k$ , we obtain

$$\mathbb{E} \{(x - Ay)y'\} = 0 \quad (n \times m)$$

<sup>35</sup>See Section ?? in the appendix for the definition of the Moore-Penrose pseudoinverse.

which is equivalent to  $\Sigma_{xy} - A\Sigma_y = 0$ . If  $\Sigma_y$  is non-singular the pseudoinverse is a true inverse, and (A.3.5) is proven. The case when  $\Sigma_y$  is singular is discussed in [35].  $\square$

### A.3.1 Integrabilità uniforme

L'integrabilità uniforme (vedere ad esempio [58, p. 188],[4, p. 32], [13, p.17]) è una condizione che garantisce che le variabili di un processo aleatorio non prendano valori "troppo grandi" per  $t \rightarrow \infty$ , in modo tale che le aspettative delle variabili rimangano uniformemente limitate per  $t \rightarrow \infty$ .

**Definition A.1.** Una famiglia di variabili aleatorie  $\{\mathbf{x}_n; n = 0, 1, \dots\}$  è uniformemente integrabile se, per  $c \rightarrow +\infty$ ,

$$\sup_n \int_{\{|\mathbf{x}_n| > c\}} |\mathbf{x}_n|(\omega) P(d\omega) \rightarrow 0 \tag{A.3.7}$$

o, in notazione equivalente

$$\lim_{c \rightarrow +\infty} \left\{ \sup_n \mathbb{E} I_{\{|\mathbf{x}_n| > c\}} |\mathbf{x}_n| \right\} \rightarrow 0. \tag{A.3.8}$$

Si dimostra che se le  $\{\mathbf{x}_n\}$  sono uniformemente integrabili, si ha

$$\sup_n \mathbb{E} |\mathbf{x}_n| < \infty$$

vedere [58, pp. 189-190]. Questa condizione non è in realtà sufficiente per l'integrabilità uniforme, vedere ad es [4, p. 32] e la proposizione A.2 più sotto. Il risultato seguente chiarisce il ruolo chiave dell'integrabilità uniforme in problemi di convergenza di successioni di v.a.

**Theorem A.2.** Si assuma che  $\mathbf{x}_n \xrightarrow{L} \mathbf{x}$ . Se le  $\{\mathbf{x}_n\}$  sono uniformemente integrabili, allora

$$\mathbb{E} \mathbf{x}_n \rightarrow \mathbb{E} \mathbf{x} \tag{A.3.9}$$

Se sia le  $\mathbf{x}_n$  che  $\mathbf{x}$  sono non-negative e hanno aspettazione finita, allora (A.3.9) implica che le  $\{\mathbf{x}_n\}$  sono uniformemente integrabili.

Per il teorema di Slutsky, la convergenza in legge implica  $\mathbf{x}_n^k \xrightarrow{L} \mathbf{x}^k$  per ogni  $k \geq 0$ , per cui,

**Corollary A.1.** Se le  $\{\mathbf{x}_n\}$  sono uniformemente integrabili, la convergenza in legge  $\mathbf{x}_n \xrightarrow{L} \mathbf{x}$  implica la convergenza di tutti i momenti che esistono; i.e.

$$\mathbb{E} \mathbf{x}_n^k \rightarrow \mathbb{E} \mathbf{x}^k \tag{A.3.10}$$

Il teorema A.2 vale in particolare se le  $\mathbf{x}_n$  e la  $\mathbf{x}$  sono definite sullo stesso spazio di probabilità e  $\mathbf{x}_n \rightarrow \mathbf{x}$  con probabilità uno (oppure in probabilità).

**Corollary A.2.** *Se le  $\{\mathbf{x}_n\}$  sono uniformemente integrabili e  $\mathbf{x}_n \rightarrow \mathbf{x}$  con probabilità uno, allora  $\mathbb{E} \mathbf{x} < \infty$  e*

$$\mathbb{E} \mathbf{x}_n \rightarrow \mathbb{E} \mathbf{x}, \quad n \rightarrow \infty \tag{A.3.11}$$

$$\mathbb{E} |\mathbf{x}_n - \mathbf{x}| \rightarrow 0, \quad n \rightarrow \infty \tag{A.3.12}$$

e quindi, se  $c'$  è integrabilità uniforme, la convergenza con probabilità uno implica la convergenza in  $L^1$ .

Notare che, se per  $n$  grandi,  $|\mathbf{x}_n| \leq \mathbf{z}$  con  $\mathbb{E} |\mathbf{z}| < \infty$ , la famiglia  $\{\mathbf{x}_n; n = 0, 1, \dots\}$  è uniformemente integrabile ([4, p. 32]). Quindi il teorema A.2 contiene in particolare il *teorema della convergenza dominata di Lebesgue*.

**Proposition A.2.** *Un processo stazionario in senso stretto  $\{\mathbf{y}(t)\}$  per cui  $\mathbb{E} |\mathbf{y}(0)|^2 < \infty$ , è uniformemente integrabile.*

*Proof.* Infatti

$$\begin{aligned} \sup_{t \geq 0} \int_{\{|\mathbf{y}(t)| > c\}} |\mathbf{y}(t, \omega)| P(d\omega) &= \int_{\Omega} I_{\{|\mathbf{y}(0)| > c\}} |\mathbf{y}(0, \omega)| P(d\omega) \\ &\leq [\mathbb{E} I_{\{|\mathbf{y}(0)| > c\}}]^{1/2} [\mathbb{E} |\mathbf{y}(0)|^2]^{1/2} \end{aligned}$$

per la disuguaglianza di Schwartz. L'ultimo membro tende a zero per  $c \rightarrow +\infty$ , dato che

$$\mathbb{E} I_{\{|\mathbf{y}(0)| > c\}} = P\{|\mathbf{y}(0)| > c\} \leq \frac{\mathbb{E} |\mathbf{y}(0)|^2}{c^2} \rightarrow 0$$

per la disuguaglianza di Chebichev.  $\square$

### Applicazione al teorema ergodico

Facciamo vedere che vale la relazione limite (??) del capitolo 4. Per questo basta in realtà una convergenza molto più debole di quella quasi certa del teorema ergodico. Nella proposizione seguente assumeremo semplicemente che la convergenza sia in legge ( $\xrightarrow{L}$ ).

**Proposition A.3.** *Sia  $\{\mathbf{y}(t)\}$  un processo strettamente stazionario,  $\mathbf{z} = f(\mathbf{y}) \in L^2(\mathbf{y})$  una funzione del processo e  $\mathbf{z}(t) = f_t(\mathbf{y})$  il processo delle traslazioni temporali. Se*

$$\bar{\mathbf{z}}_T := \frac{1}{T} \sum_{t=1}^T \mathbf{z}(t) \xrightarrow{L} \bar{\mathbf{z}}$$

allora, per  $T \rightarrow \infty$ ,

$$\mathbb{E} \left\{ \frac{1}{T} \sum_{t=1}^T \mathbf{z}(t) \right\} = \mathbb{E} \mathbf{z}(t) \rightarrow \mathbb{E} \bar{\mathbf{z}}$$

e quindi  $\mathbb{E} \bar{\mathbf{z}} = \mathbb{E} \mathbf{z}(t)$  per ogni  $t$ .

**Proof.** Come abbiamo visto, per poter inferire la convergenza delle aspettative da quella in legge occorre e basta la condizione di *integrabilità uniforme* del processo  $\{\mathbf{z}(t)\}$ , che nel nostro caso è garantita dal fatto che  $\mathbf{z} = f(\mathbf{y}) \in L^2(\mathbf{y})$ . L'ultima affermazione scende dal fatto che l'aspettazione della media temporale non dipende da  $T$ .  $\square$

# BIBLIOGRAPHY

- [1] M. S. Bartlett. Periodogram analysis and continuous spectra. *Biometrika*, 37:1–16, 1950.
- [2] S. Bernstein. Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes. *Math. Ann.*, 97:1–59, 1926.
- [3] M. Bertero. Linear inverse and ill-posed problems. In *Advances in Electronics and Electron Physics*, volume 75, pages 1–120. Academic Press, 1989.
- [4] P. Billingsley. *Convergence of probability measures*. Wiley, 1968.
- [5] G.D. Birkhoff. Proof of the ergodic theorem. *Proc. Nat. Acad. Sciences (USA)*, 17:565–600, 1931.
- [6] S. Bittanti, M. Campi, and S. Savaresi. Unbiased estimation of a sinusoid in colored noise via adaptive notch filters. *Automatica*, 33:209–215, 1997.
- [7] J. R. Bunch and C.P. Nielsen. Updating the singular value decomposition. *Numer. Math.*, 31:111–129, 1978.
- [8] J. R. Bunch, C.P. Nielsen, and D. Sorensen. Rank one modification of the symmetric eigenvalue problem. *Numer. Math.*, 31:31–48, 1978.
- [9] W. G. Cochran. *Sampling Techniques (Third Ed.)*. Wiley, 1977.
- [10] Harald Cramèr. *Mathematical Methods of Statistics*. Princeton University Press, 1946.
- [11] G. Cybenko. Approximation by sigmoidal functions. *Mathematics of Control Signals and Systems*, pages 303–314, 1989.
- [12] A.P. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.
- [13] A. Van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge U.K., 1998.
- [14] J. L. Doob. *Stochastic Processes*. Wiley, 1953.
- [15] H. Dym and H. P. McKean. *Fourier series and integrals*. Academic Press, 1972. Probability and Mathematical Statistics, No. 14.

- [16] M. P. Ekstrom. A spectral characterization of the ill-conditioning in numerical deconvolution. *IEEE Trans, Audio Electroacustics*, AU-21:344–348, 1973.
- [17] T. Ferguson. *A Course in Large Sample Theory*. Chapman and Hall, 1996.
- [18] W. Freiberger and U. Grenander. *A short Course in Computational Probability and Statistics*. Springer Verlag, Berlin, 1971.
- [19] K.F. Gauss. *Theoria Motus Corporum Coelestium*. Julius Springer, Berlin, 1901.
- [20] G.H. Golub and C.R. Van Loan. *Matrix Computation (Third ed.)*. The Johns Hopkins Univ. Studies in the Mathematical Sciences, 1996.
- [21] 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.
- [22] E. J. Hannan. *Multiple Time Series*. Wiley, 1970.
- [23] E.J. Hannan. Da trovare. 16:??, 1979.
- [24] E.J. Hannan and M. Deistler. *The Statistical Theory of Linear Systems*. Wiley, 1998.
- [25] W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.
- [26] R. V. Hogg and A. T. Craig. *Introduction to Mathematical Statistics 3d edition*. Macmillan Co. New York, 1969.
- [27] B. R. Hunt. A theorem on the difficulty of numerical deconvolution. *IEEE Trans, Audio Electroacustics*, AU-20:., 1972.
- [28] I.A. Ibragimov. A central limit theorem for a class of dependent random variables. *Theory of Probability and Applications*, 7:349–382, 1963.
- [29] A.N. Kolmogorov. *Foundations of Probability Theory*. Chelsea, 1956. Translation of the 1933: Grundbegriffe der Wahrscheinlichkeitsrechnung.
- [30] L. D. Landau and E. M. Lifschits. *Statistical Physics (Trad italiana : Fisica Statistica)*. Editori Riuniti, 1978.
- [31] C.L. Lawson and R.J. Hanson. *Solving Least Squares Problems*. Prentice Hall, Englewood Cliffs, 1974.
- [32] E. Lehmann. *Testing Statistical Hypotheses (second Ed.)*. Wiley, 1986. reprinted by Springer Verlag.
- [33] E. Lehmann and R. Casella. *Theory of Point Estimation*. Springer Texts in Statistics, 1998.

- [34] Bernard C. Levy. *Principles of signal detection and parameter estimation*. Springer Verlag, 2008.
- [35] R. Liptser and A.N. Shiriyayev. *Statistics of Random Processes, Voll I, II*. Springer Verlag, 1977.
- [36] L. Ljung. On the consistency of prediction-error identification methods. In R.K. Mehra and D. G. Lainiotis, editors, *System Identification: Advances and Case Studies*. Academic Press, New York, 1976.
- [37] L. Ljung. *System Identification, Theory for the User (second Ed.)*. Prentice Hall, Englewood Cliffs, 1999.
- [38] M. Loeve. *Probability Theory*. Van Nostrand Reinhold, 1963. Reprinted by Springer Verlag in 1986.
- [39] Anders Martin-Löf. *Statistical mechanics and the foundations of thermodynamics*. Springer Verlag, Lecture Notes in Physics, 1979.
- [40] MATLAB. *Using MATLAB Version 6*. The MathWorks Inc., 2002.
- [41] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journ. Chem. Phys.*, 21:1087–1092, 1953.
- [42] A. Nehorai. A minimal parameter adaptive notch filter with constrained poles and zeros. *IEEE Trans. Acust. Speech Sign. Process.*, ASSP-33:983–996, 1985.
- [43] D. L. Phillips. A technique for the numerical solution of certain integral equations of the first kind. *Journal of the Assoc. Comput. Mach.*, 9:97–101, 1962.
- [44] G. Picci. *Filtraggio Statistico (Wiener, Levinson, Kalman) e applicazioni*. Libreria Progetto Padova, 1994.
- [45] Giorgio Picci. *Filtraggio Statistico (Wiener, Levinson, Kalman) e Applicazioni*. Ed. Libreria Progetto, Padova, 2007.
- [46] V. F. Pisarenko. The retrieval of harmonics from a covariance function. *Geophys. Journ Royal Astron. Soc.*, 33:347–366, 1973.
- [47] T. Poggio and F. Girosi. Networks for approximation and learning. *Proc. IEEE*, 78(9):1481–1497, 1990.
- [48] B. Porat. *Digital Processing of Random Signals*. Prentice Hall, Englewood Cliffs, N.J., 1994.
- [49] P. Stoica and R. L. Moses. *Spectral Analysis of Signals*. Prentice-Hall, New Jersey, 2005.
- [50] David G. Stork Richard O. Duda, Peter E. Hart. *Pattern Classification, 2nd Edition*. Wiley, 2000.

- [51] C.P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, 1999. Springer Texts in Statistics.
- [52] Richard A. Roberts and Clifford T. Mullis. *Digital signal processing*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1987.
- [53] M. Rosenblatt. A central limit theorem and a strong mixing condition. *Proc. Natnl. Acad. Sci. (USA)*, 42:43–47, 1956.
- [54] Y. A. Rozanov. *Stationary Random Processes*. Holden-Day, San Francisco, 1967.
- [55] H. Scheffè. *The Analysis of Variance*. Wiley, 1953.
- [56] H. Scheffè'. *The Analysis of Variance*. Wiley, New York, 1959.
- [57] A. Schuster. On lunar and solar periodicities of earthquakes. *Proc. Royal Soc.*, 61:455–465, 1897.
- [58] A. N. Shiryaev. *Probability (Second edition)*. Springer Verlag, New York, 1989.
- [59] Ya. G. Sinai. *Introduction to Ergodic Theory*. Princeton University Press, 1976.
- [60] A. Stuart Sir Maurice Kendall and J.K. Ord. *The advanced theory of statistics : Voll. 1,2,3*. Griffin, High Wycombe U.K., 1983.
- [61] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P. Glorennec, H. Hjalmarsson, and A. Juditsky. Nonlinear black-box modeling in system identification: a unified overview. *Automatica*, 31(12):1691–1724, 1995.
- [62] T. Söderström and P. Stoica. *System Identification*. Prentice-Hall, 1989.
- [63] E. Sontag. *Essays on Control*, chapter Neural Networks for Control. Birkhauser, Basel, Switzerland, 1993.
- [64] G.W. Stewart. Collinearity and least squares regression. *Statistical Science*, 2:68–100, 1987.
- [65] G. Strang. *Linear Algebra and its Applications*. Academic Press, New York, 1976.
- [66] S. Twomey. The application of numerical filtering to the solution of integral equations of the first kind encountered in indirect sensing measurements. *Journal of the Franklin Institute*, 279:95–109, 1965.
- [67] H. T. van Trees. *Detection Estimation and Modulation Theory Vol. I*. Wiley, 1976.
- [68] G. Wahba. *Spline methods for observational data*. SIAM CBMS-NSF series, Philadelphia, 1990.
- [69] N. Wiener. Generalized harmonic analysis. *Acta Mathematica*, 55:117–258, 1930.
- [70] N. Wiener. *The Fourier integral and certain of its applications*. Cambridge U.P., 1933.



- 
- [71] G. U. Yule. On a method of investigating periodicities in disturbed series with special references to Wolfer's sunspot numbers. *Phil. Trans. Royal Soc. A*, 226:267–298, 1927.
- [72] E. Zacks. *The Theory of Statistical Inference*. Wiley, 1970.